Protein language models

Chris Dallago (cdallago@nvidia.com) @ BeVAS / EPFL Apr 18, 2023

RostLab

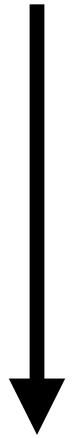
NVIDIA

Agenda

- Why predict proteins?
- Data-driven feature extraction approaches
- LLMs for protein representations
  - P.S.: LLMs for gene representations
- LLMs for protein design

What we want

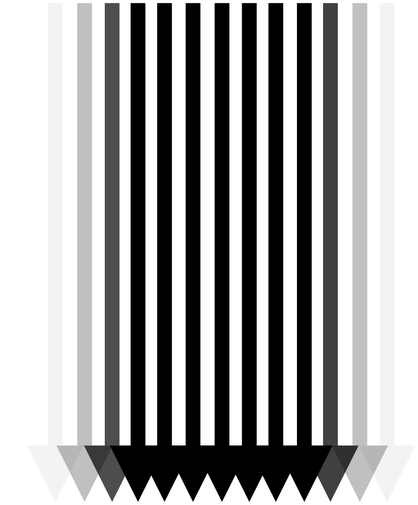
MALLHSARVLSGVASAFHPGLAAAASARASSWWAHVEMGPPDPILGVTEAYKRDTNSKKMNLGVG



Mitochondrial

What we want

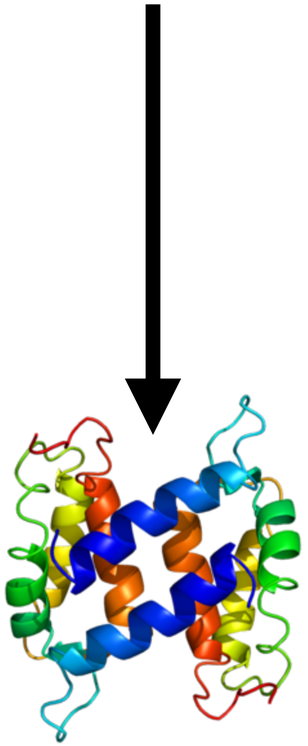
MALLHSARVLSGVASAFHPGLAAAASARASSWWAHVEMGPPDPILGVTEAYKRDTNSKKMNLGVG



Alpha helical

What we want

MALLHSARVLSGVASAFHPGLAAAASARASSWWAHVEMGPPDPILGVTEAYKRDTNSKKMNLGVG



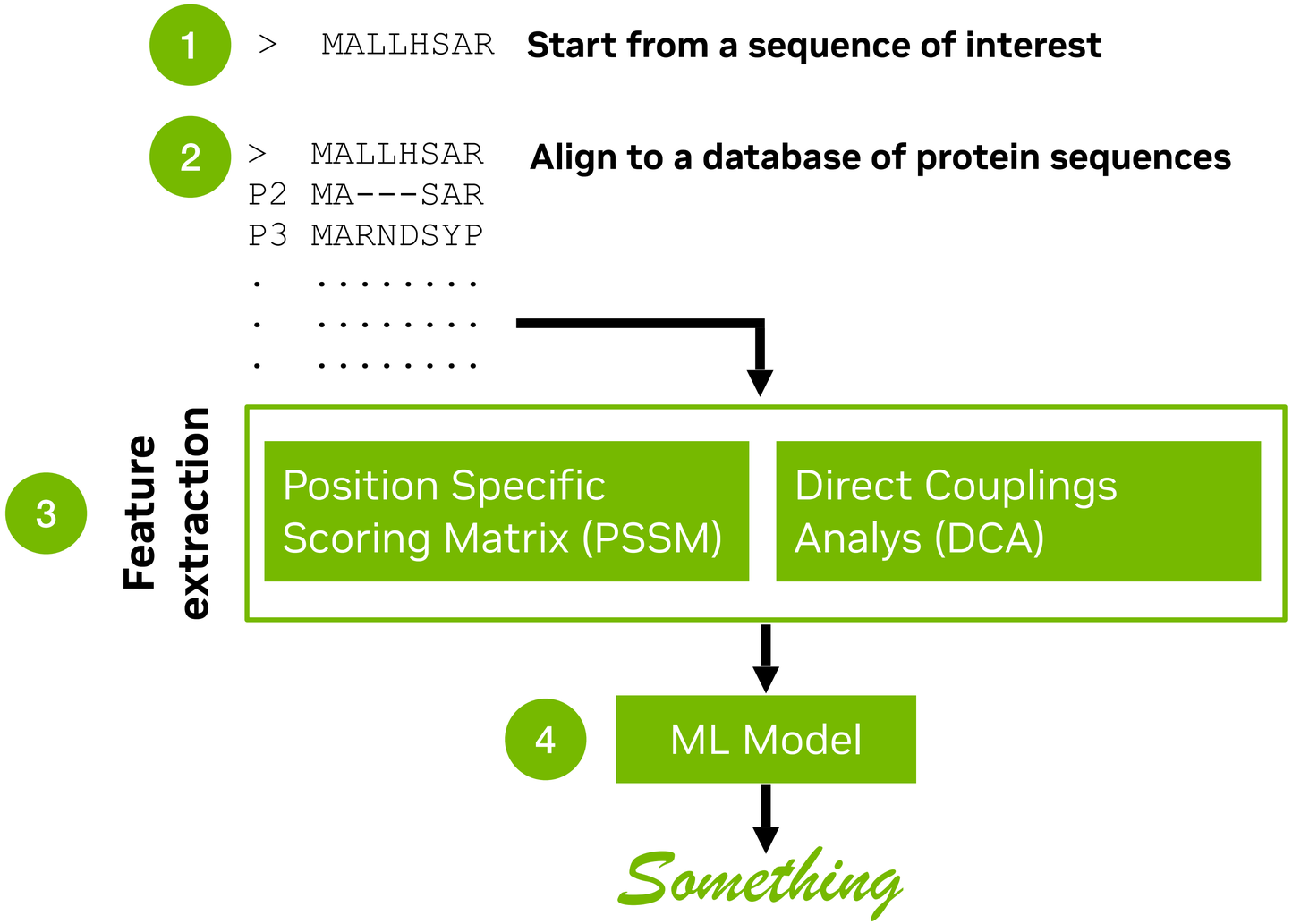
What we want

MALLHSARVLSGVASAFHPGLAAAASARASSWWAHVEMGPPDPILGVTEAYKRDTNSKKMNLGVG

Something

How we did it

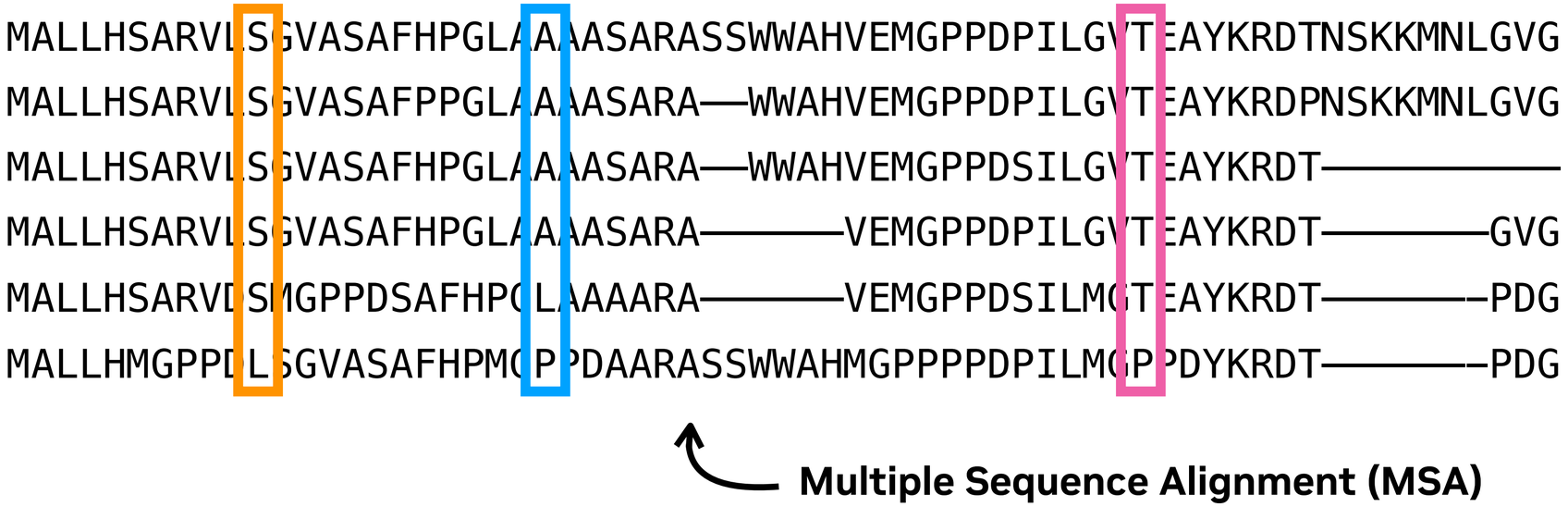
Extract features from alignment and use ML



3

Feature extraction

Position-Specific Scoring Matrix (PSSM)



	M	A	L	L	…
A	0.0	0.4	0.1	0.1	…
M	0.9	0.2	0.1	0.1	…
G	0.0	0.0	0.1	0.0	…
…	…	…	…	…	…

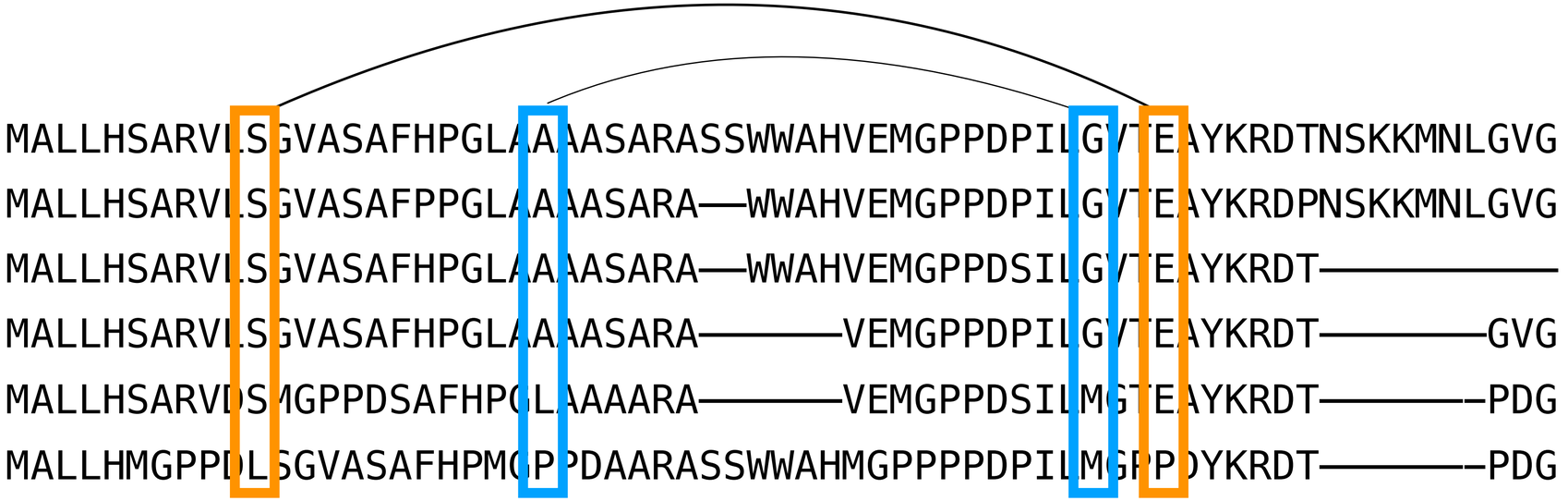
Matrix sequence representation:

PSSM = $L \times 21$

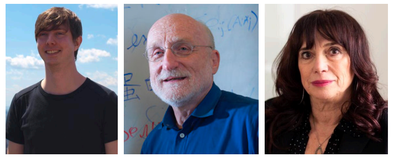
3

Feature extraction

Direct Couplings Analysis (DCA)



For each pair of residues R1 and R2, what's the likelihood that R1 changes to X if R2 changes to Y?



The EVcouplings Python framework for coevolutionary sequence analysis
Thomas A Hopf, Anna G Green, Benjamin Schubert, Sophia Mersmann, Charlotta PI Schärfe, John B Ingraham, Agnes Toth-Petroczy, Kelly Brock, Adam J Riesselman, Perry Palmedo, Chan Kang, Robert Sheridan, Eli J Draizen, Christian Dallago, Chris Sander, Debora S Marks
2019

How we did it

1 `> MALLHSAR` **Start from a sequence of interest**

2
```
>  MALLHSAR
P2 MA---SAR
P3 MARNDSYP
```
Align to a database of protein sequences

De-novo

3 **Feature extraction**

- Position Specific Scoring Matrix (PSSM)
- Direct Couplings Analys (DCA)

4 ML Model

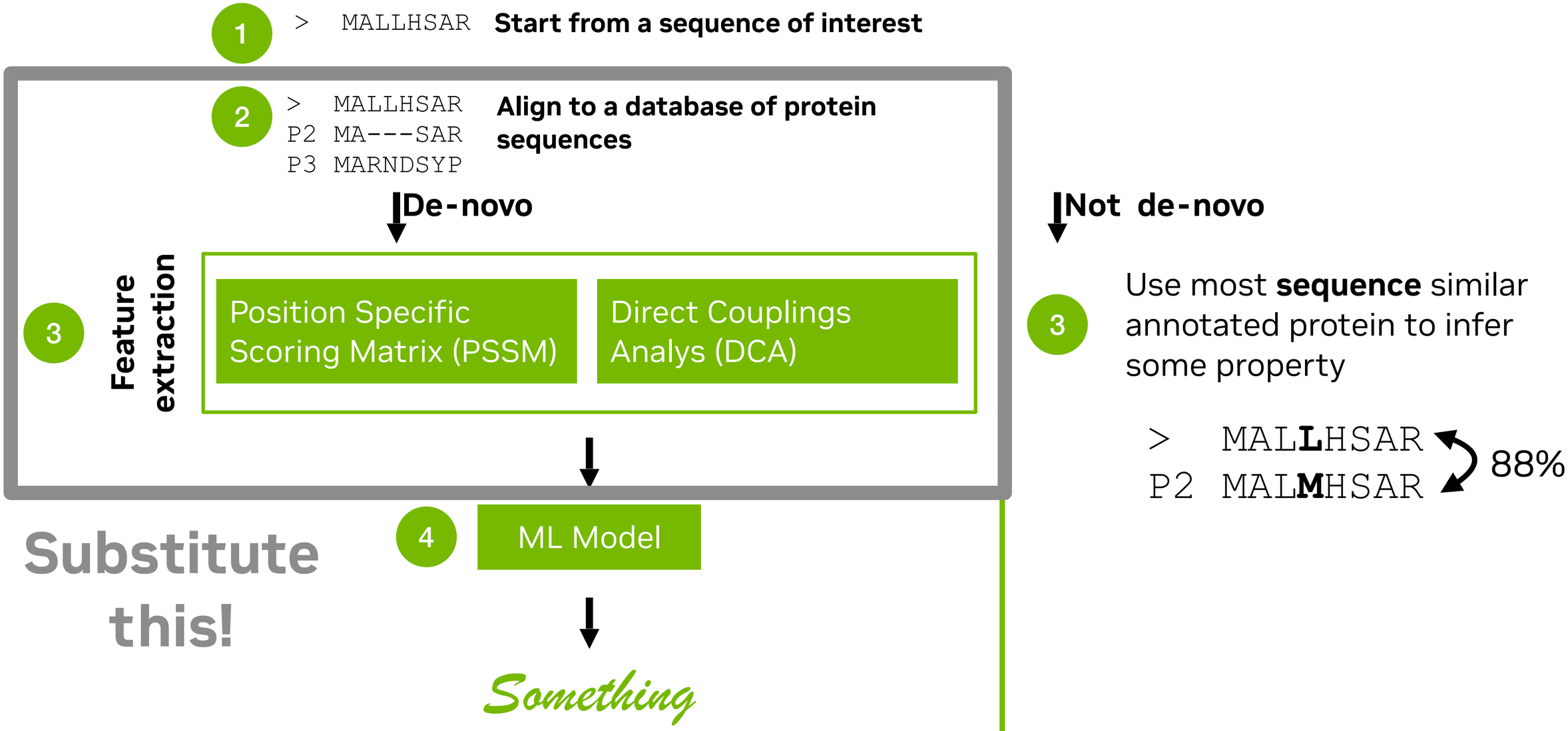
Something

Not de-novo

3 Use most **sequence** similar annotated protein to infer some property

```
>  MALLHSAR
P2 MALMHSAR
```
88%

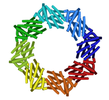
How we did it


1
> MALLHSAR
Start from a sequence of interest
2
> MALLHSAR
P2 MA---SAR
P3 MARNDSYP
Align to a database of protein sequences
De-novo
Not de-novo
3
Feature extraction
Position Specific Scoring Matrix (PSSM)
Direct Couplings Analys (DCA)
3
Use most sequence similar annotated protein to infer some property
> MALLHSAR
P2 MALMHSAR
88%
4
ML Model
Something
Substitute this!

Existing feature extraction approaches

	Compute quickly*	Available for all proteins	Captures residue context	Captures complexity
BLOSUM	✅	✅	⛔	⛔
PSSM	⛔	⚠️	✅	⚠️
DCA	⛔	⛔	✅	✅

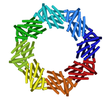
* at inference



Can we find a better representation?

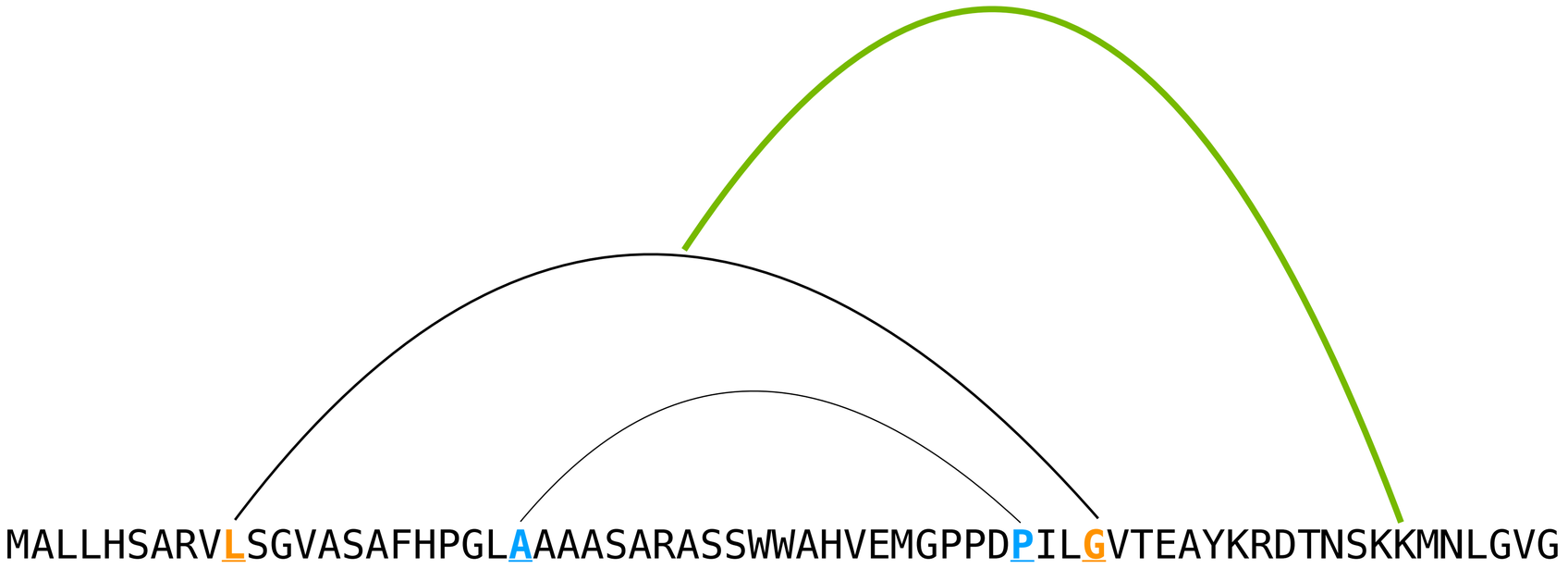
	Compute quickly*	Available for all proteins	Captures residue context	Captures complexity
BLOSUM	✅	✅	⛔	⛔
PSSM	⛔	⚠️	✅	⚠️
DCA	⛔	⛔	✅	✅
Better Rep™	✅	✅	✅	✅

*** at inference**

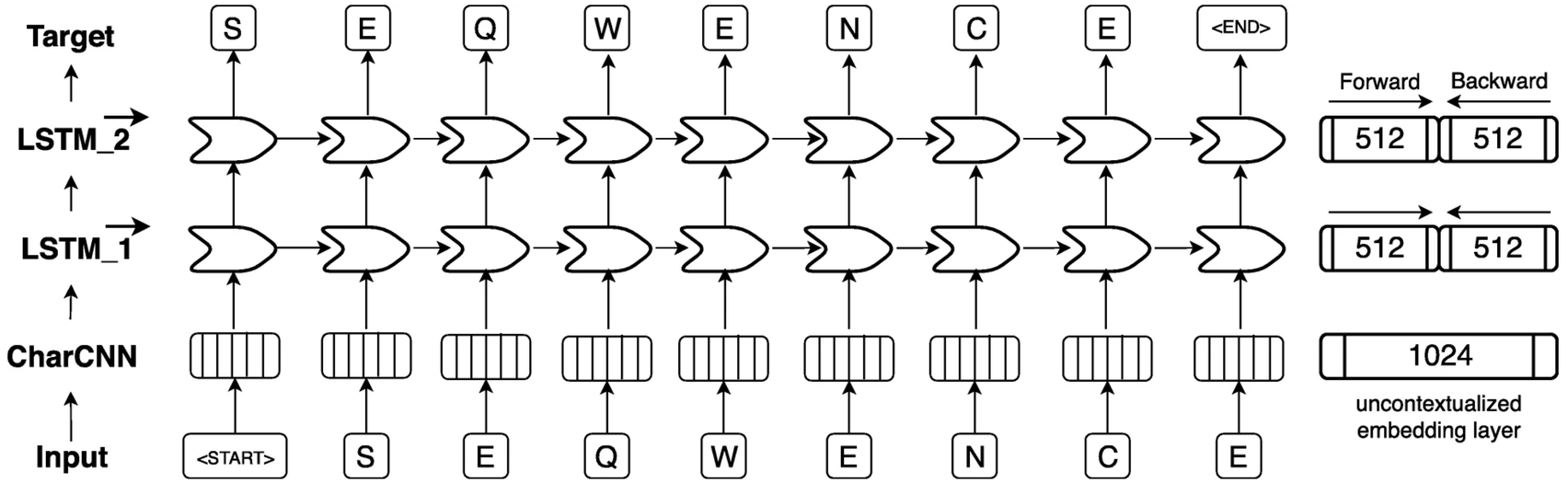


Available for all + informative

Should be: single sequence + capture higher order



Using LSTMs



Objective: Predict next amino acid (AA) in protein

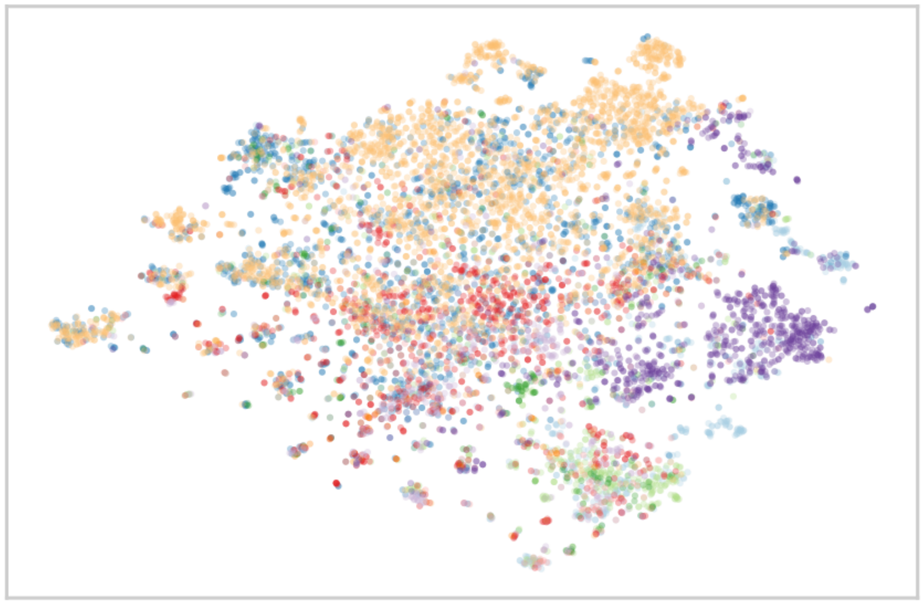
- Bidirectional LSTM: go from left to right + right to left (independently; only loss is summed)
- Train on: UniRef50 (a redundancy reduced dataset of sequences) —> Why?
  - All sequence data == too much data to train on! (= computational limitations)
  - Highly redundant sequences don't add new information!

Modeling aspects of the language of life through transfer-learning protein sequences
Michael Heinzinger, Ahmed Elnaggar, Yu Wang, Christian Dallago, Dmitrii Nechaev, Florian Matthes, Burkhard Rost
2019

Unsupervised: if you squint your eyes...

Function

(c) Subcellular Localization



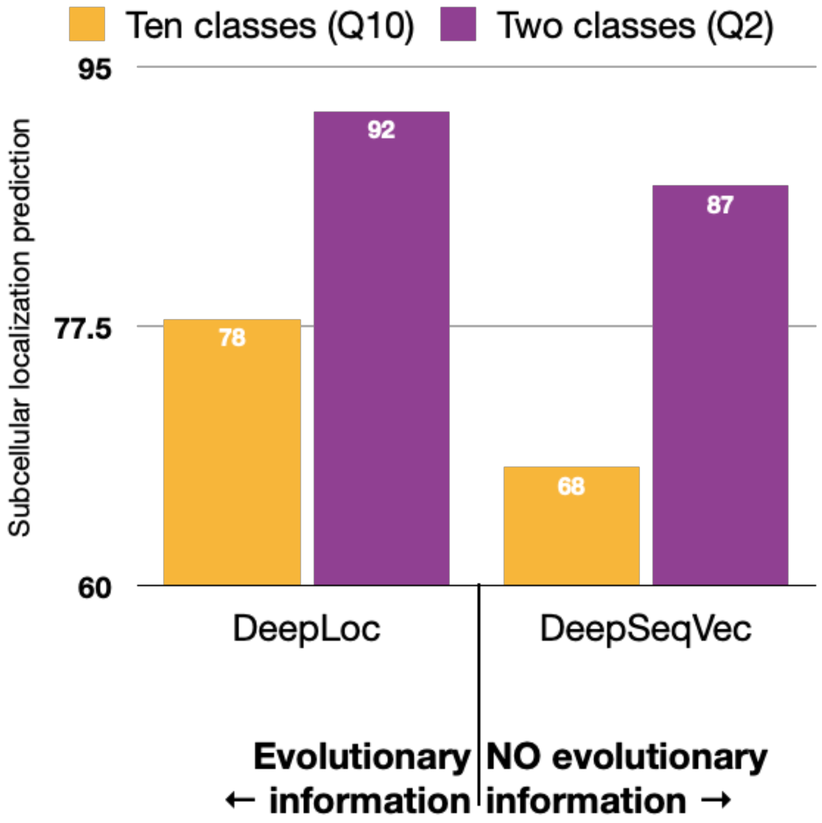
Structure

(b) SCOPe

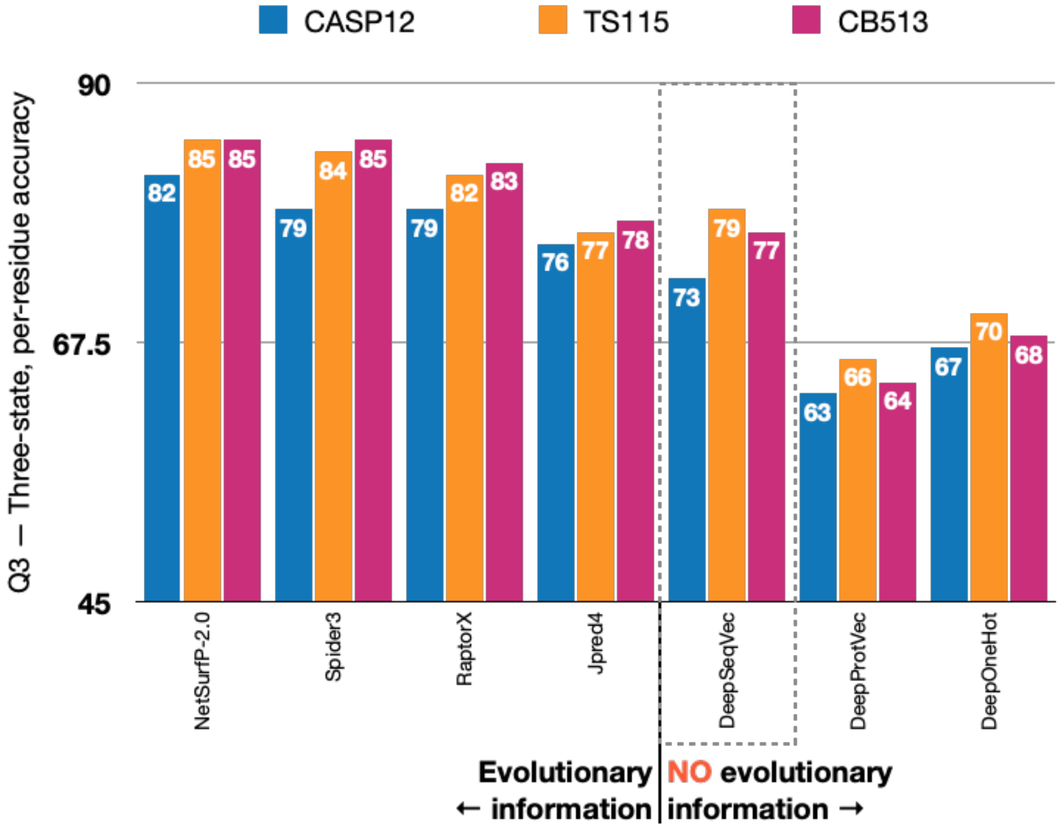
Modeling aspects of the language of life through transfer-learning protein sequences
Michael Heinzinger, Ahmed Elnaggar, Yu Wang, Christian Dallago, Dmitrii Nechaev, Florian Matthes, Burkhard Rost
2019

Supervised: not top, but “good enough” & fast!

Function



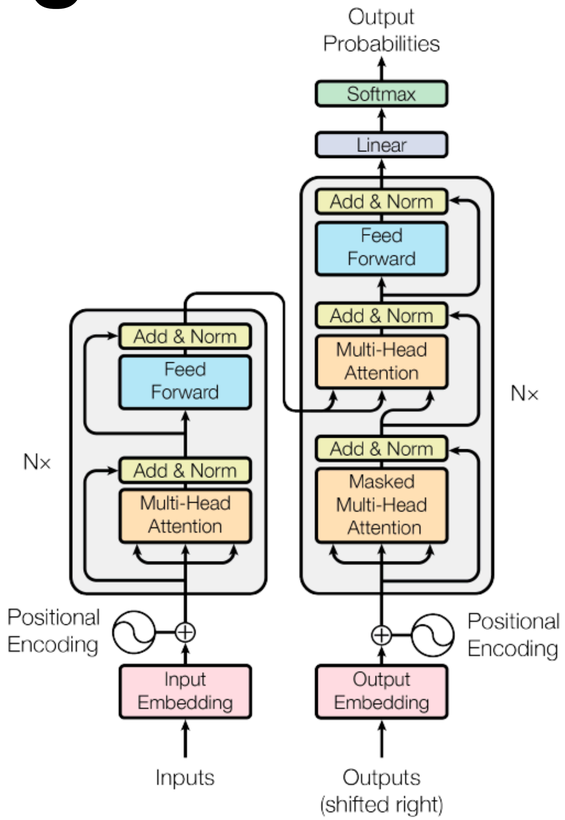
Structure



Modeling aspects of the language of life through transfer-learning protein sequences
Michael Heinzinger, Ahmed Elnaggar, Yu Wang, Christian Dallago, Dmitrii Nechaev, Florian Matthes, Burkhard Rost
2019


How can we improve?

Using Transformers



Objective: predict residue based on other residues and position in protein (MLM)

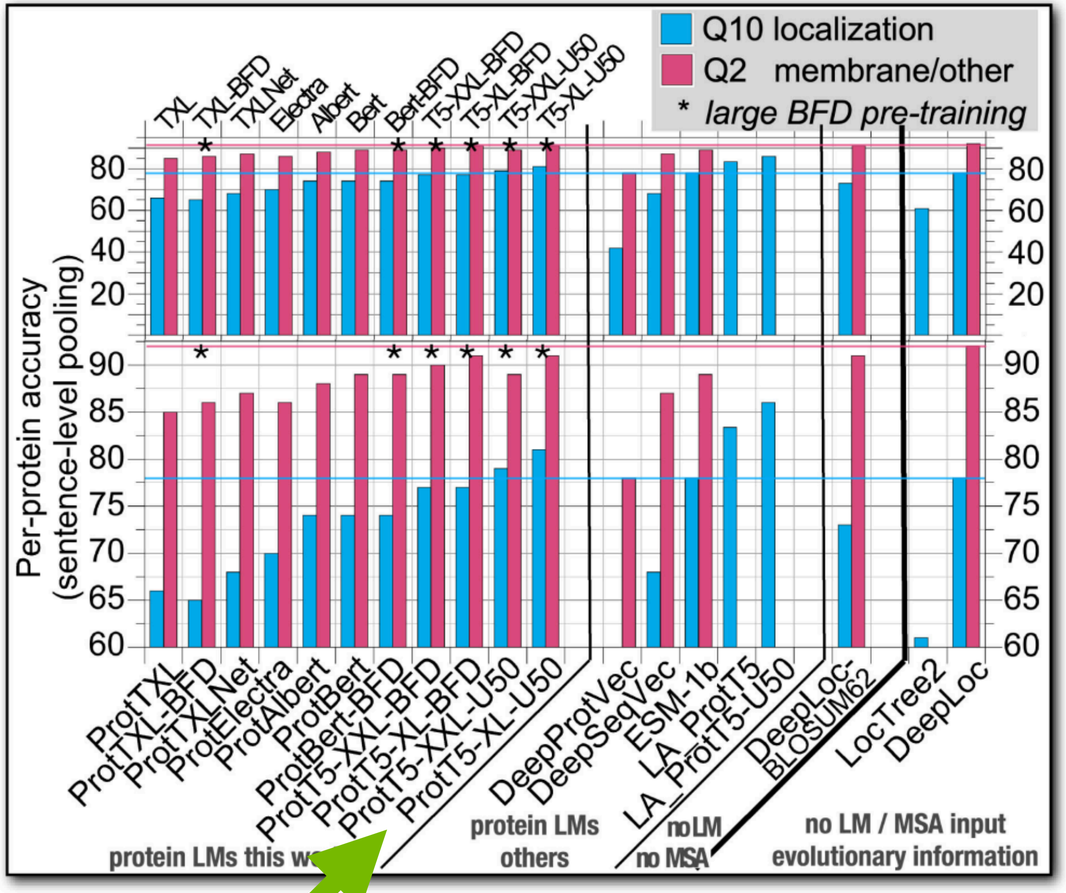
- Train very large models (LLMs)
- Train on large datasets (BFD)

→ **Need high performance computing infrastructure & optimised frameworks**

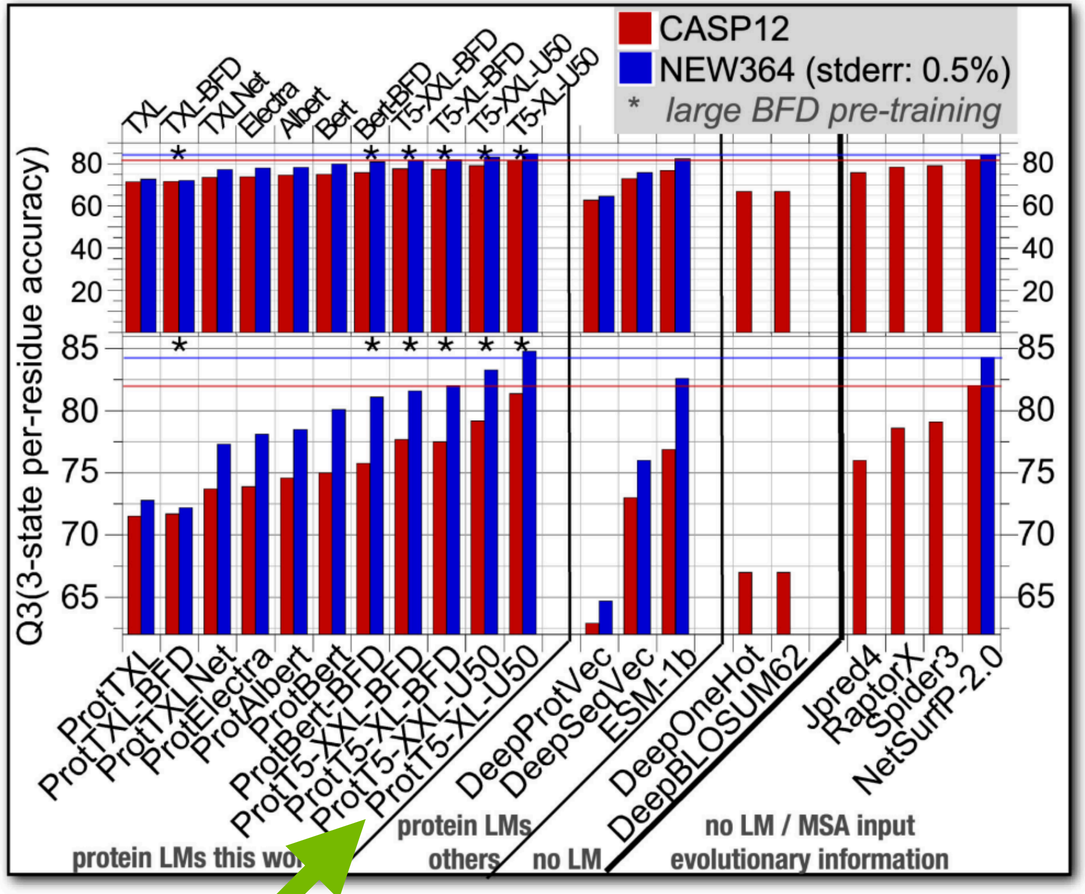
ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing
Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Debsindhu Bhowmik, Burkhard Rost
2020

Accuracy reaching top!

Function



Structure

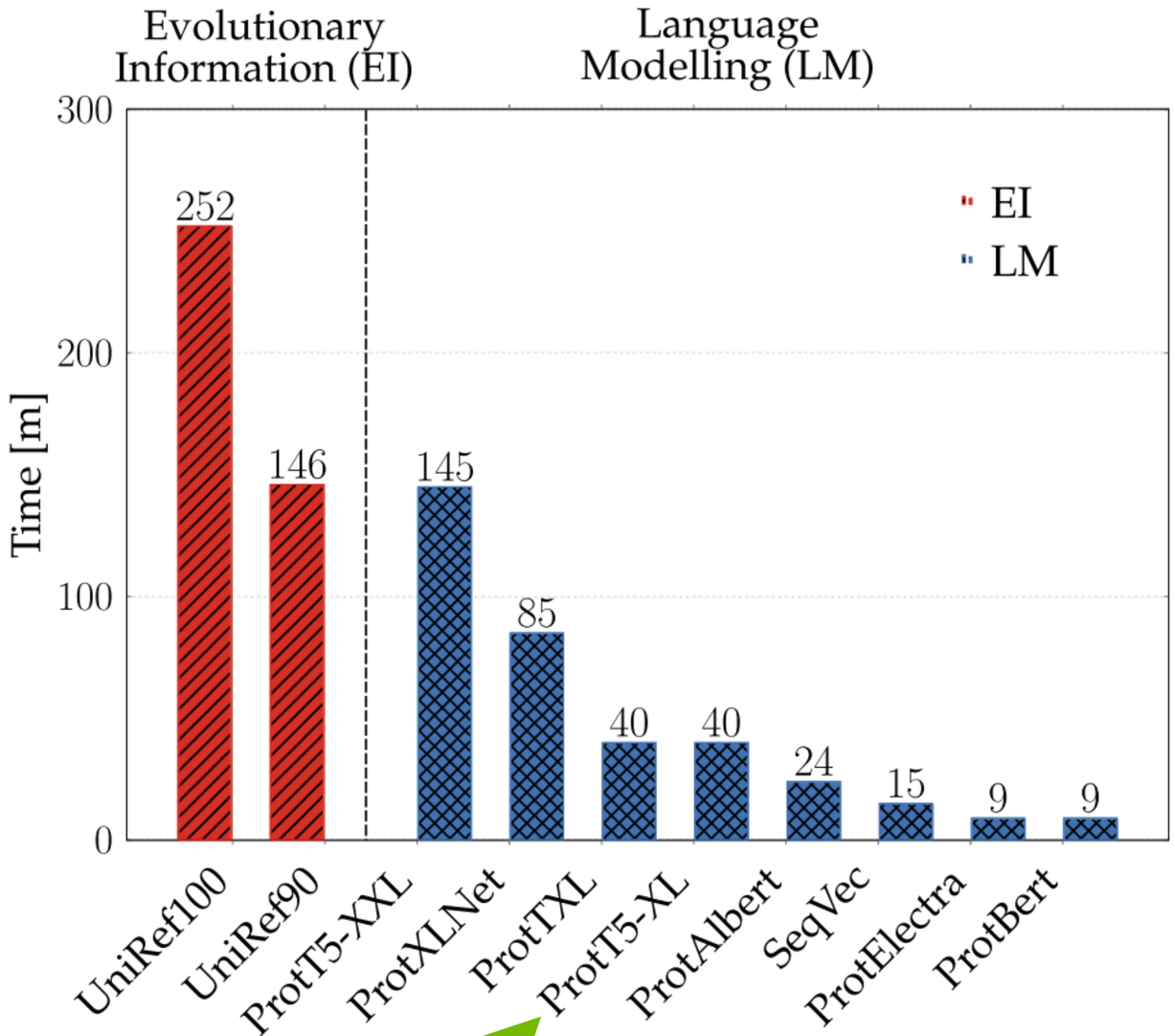


ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Debsindhu Bhowmik, Burkhard Rost

2020

Best performing model very fast



ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing
Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Debsindhu Bhowmik, Burkhard Rost
2020

What we want

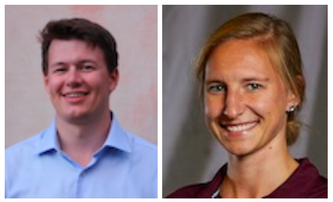
MALLHSARVLSGVASAFHPGLAAAASARASSWWAHVEMGPPDPILGVTEAYKRDTNSKKMNLGVG

Something

Make science accessible!

bioembeddings.com

Learned Embeddings from Deep Learning to Visualize and Predict Protein Sets
Dallago, C., Schütze, K., Heinzinger, M., Olenyi, T., Littmann, M., Lu, A. X., Yang, K. K., Min, S., Yoon, S., Morton, J. T., & Rost, B.
2021

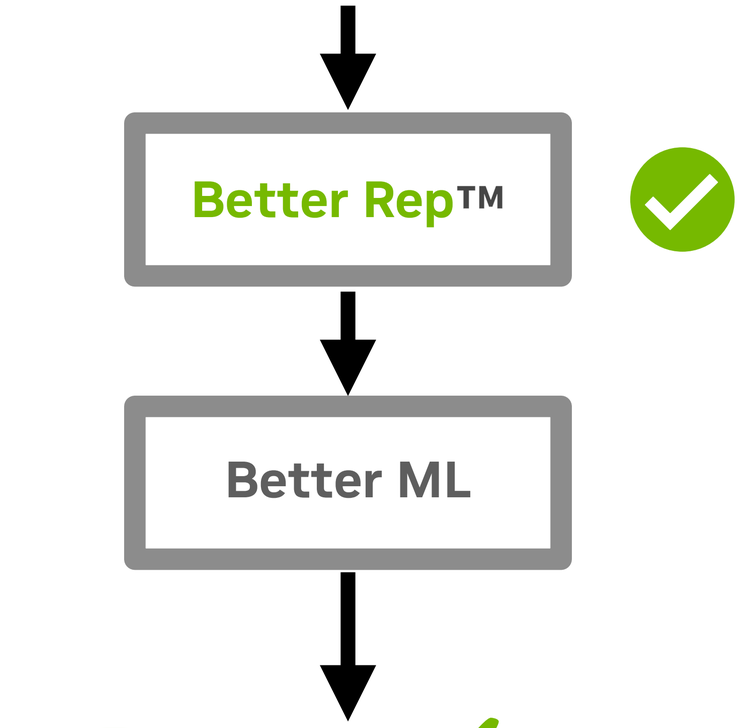


embed.predictprotein.org

LambdaPP: Fast and accessible protein-specific phenotype predictions
Olenyi T, Marquet C, Heinzinger M, Kroeger B, Nikolova T, Saendig P, Bernhofer M, Schuetze K, Littmann M, Mirdita M, Steinegger M., Dallago C. & Rost, B.
2022

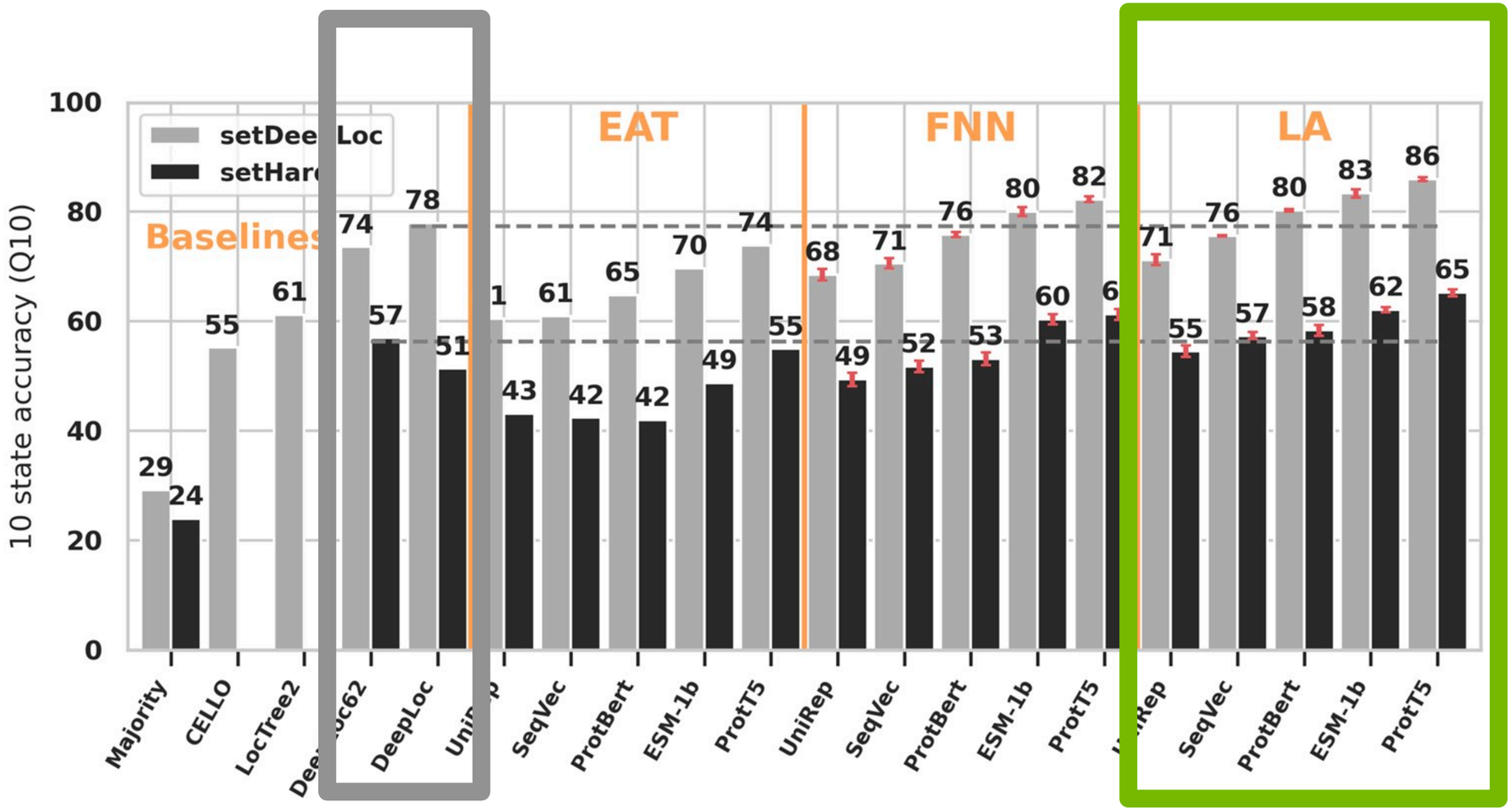
What we want

MALLHSARVLSGVASAFHPGLAAAASARASSWWAHVEMGPPDPILGVTEAYKRDTNSKKMNLGVG



Something

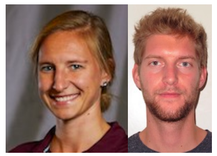
Light Attention to predict Subcellular Location



Light Attention Predicts Protein Location from the Language of Life
Hannes Stärk, Christian Dallago, Michael Heinzinger, Burkhard Rost
2021

Do it for many tasks...

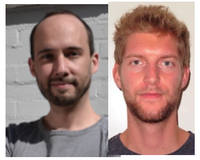
Protein Function ◀ ▶ Protein Structure



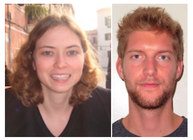
Embeddings from protein language models predict conservation and variant effects
Céline Marquet, Michael Heinzinger, Tobias Olenyi, Christian Dallago, Michael Bernhofer, Kyra Erckert, Burkhard Rost
2021 —> Human genetics

ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing
Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Debsindhu Bhowmik, Burkhard Rost **2019 —> IEEE Trans. on Patt. An. And Mach. Int.**

Mbed: transmembrane proteins predicted through language model embeddings.
Bernhofer, Michael, and Burkhard Rost
2022 —> BMC bioinformatics



Protein language-model embeddings for fast, accurate, and alignment-free protein structure prediction.
Weißenow, Konstantin, Michael Heinzinger, and Burkhard Rost.
2022 —> Structure



Embeddings from deep learning transfer GO annotations beyond homology.
Littmann, Maria, Michael Heinzinger, Christian Dallago, Tobias Olenyi, and Burkhard Rost.
2021 —> In Review

Protein embeddings and deep learning predict binding residues for various ligand classes
Maria Littmann, Michael Heinzinger, Christian Dallago, Konstantin Weissenow & Burkhard Rost
2021 —> Scientific reports

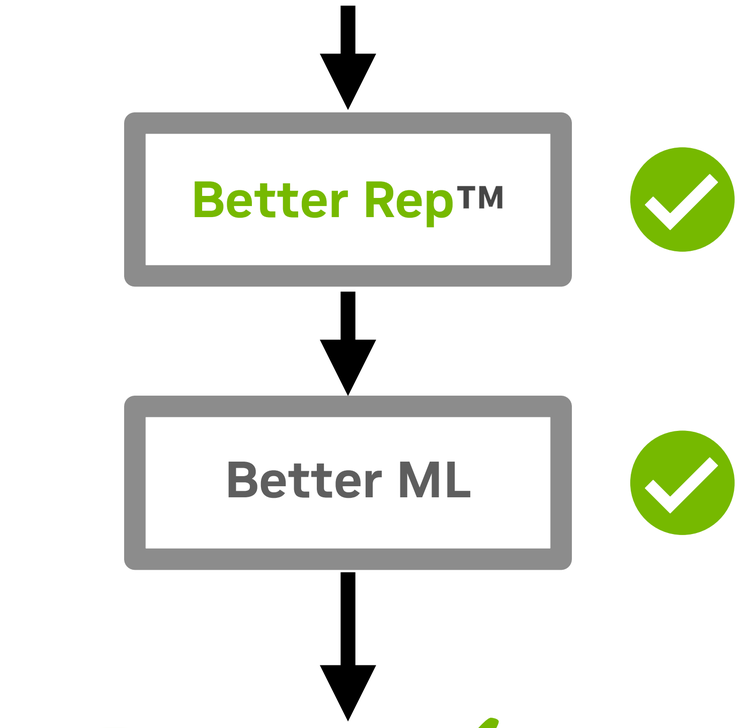
Contrastive learning on protein embeddings enlightens midnight zone.
Heinzinger, Michael, Maria Littmann, Ian Sillitoe, Nicola Bordin, Christine Orengo, and Burkhard Rost.
2022 —>NAR genomics and bioinformatics

Light Attention Predicts Protein Location from the Language of Life
Hannes Stärk, Christian Dallago, Michael Heinzinger, Burkhard Rost
2021 —> Bioinformatics Advances

SETH predicts nuances of residue disorder from protein embeddings.
Ilzhoefer, Dagmar, Michael Heinzinger, and Burkhard Rost.
2022 —> In Review

What we want

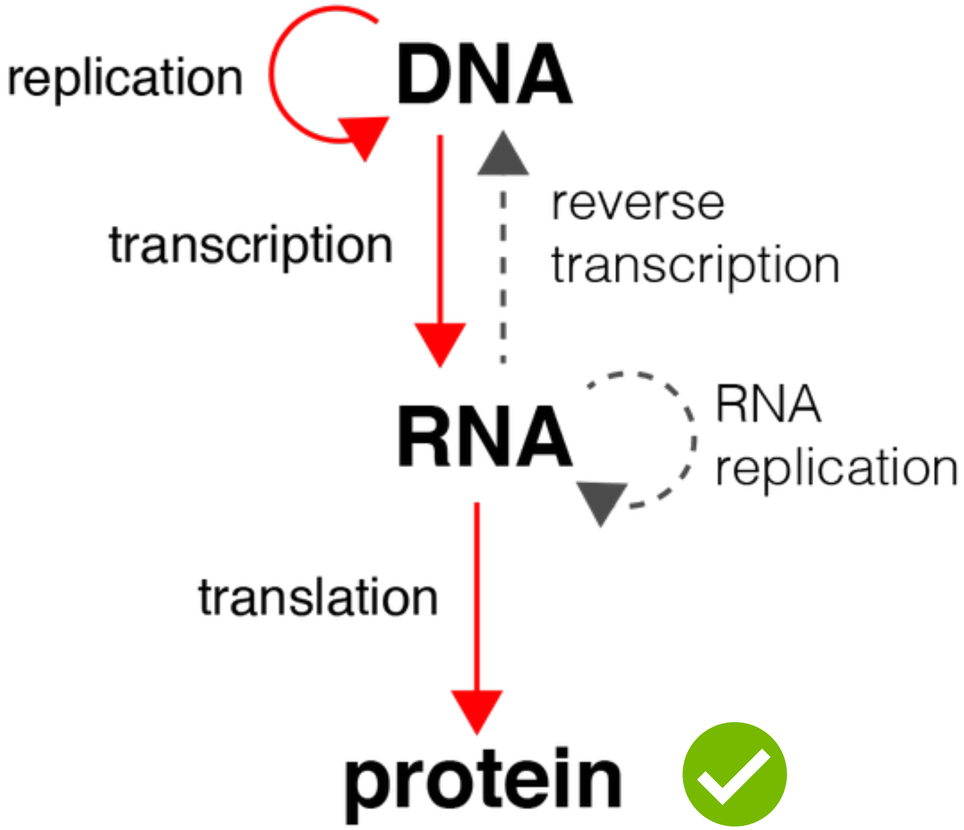
MALLHSARVLSGVASAFHPGLAAAASARASSWWAHVEMGPPDPILGVTEAYKRDTNSKKMNLGVG




Geneworld

The central dogma of mol biology

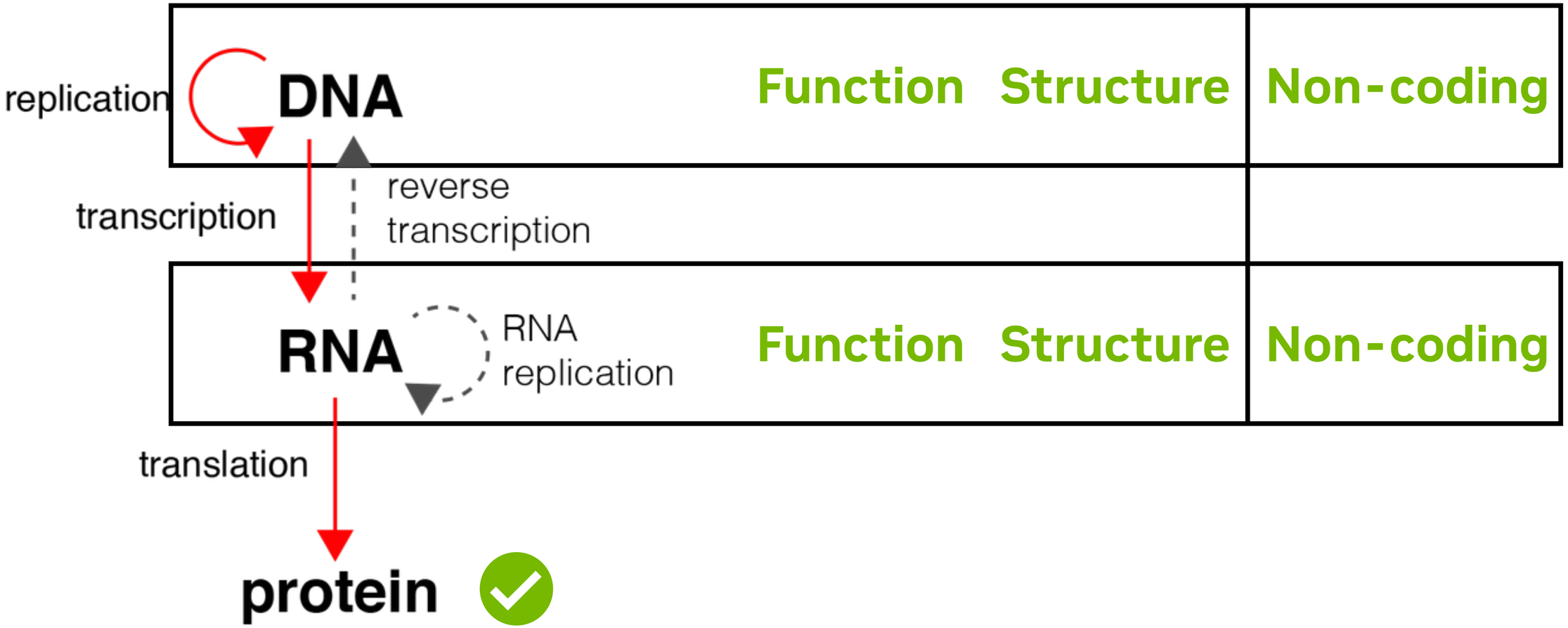
Protein LMs can do cool things: what about upstream?



https://www.quora.com/What-is-the-central-dogma-of-molecular-biology-Is-it-true.

The central dogma of mol biology

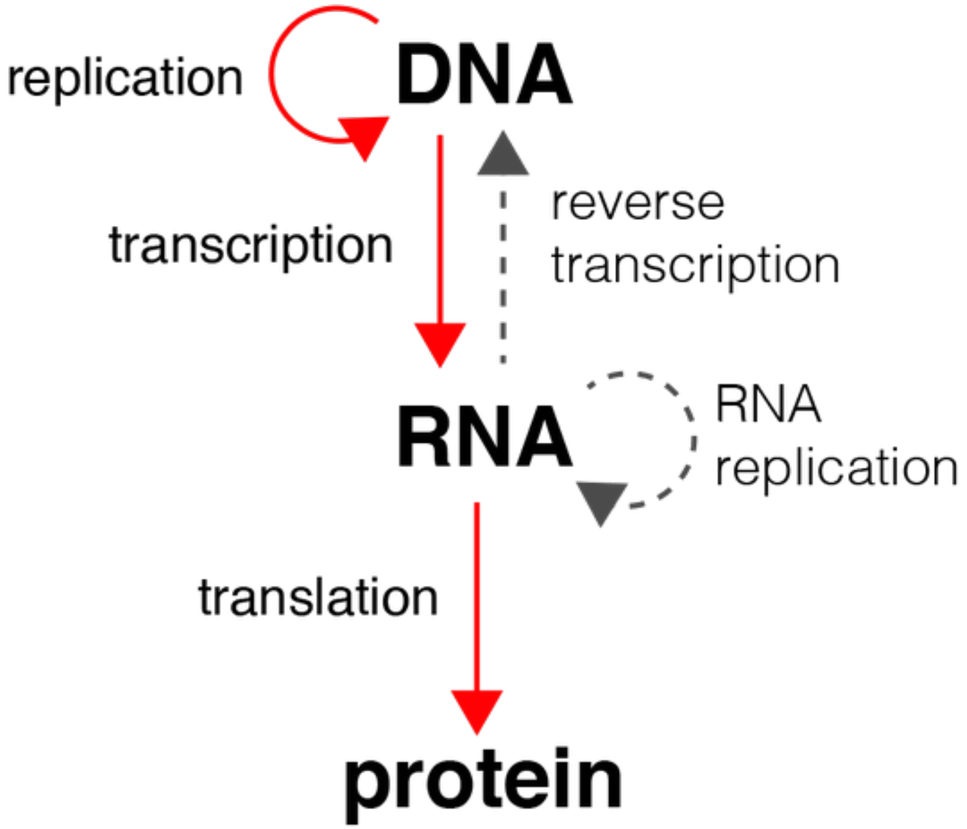
Nucleotide sequences do all sorts of cool things



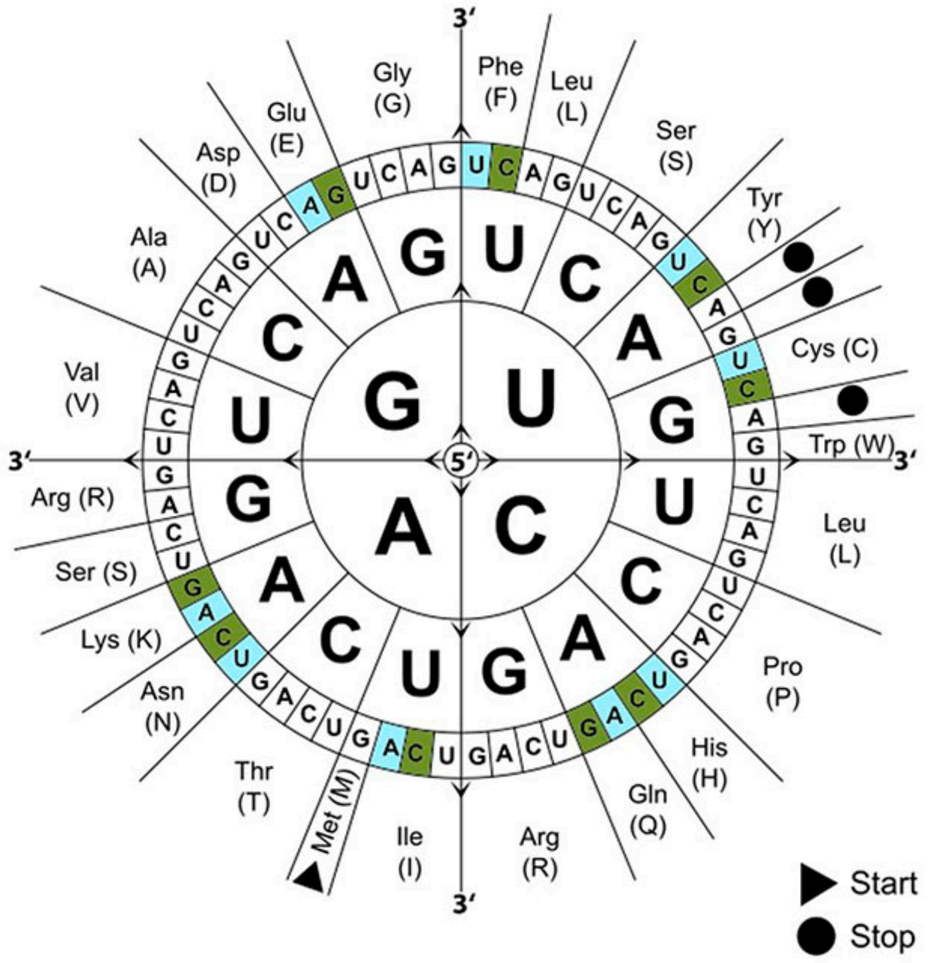
https://www.quora.com/What-is-the-central-dogma-of-molecular-biology-Is-it-true.

The central dogma of mol biology

Even protein-centric upstream views give more info



https://www.quora.com/What-is-the-central-dogma-of-molecular-biology-Is-it-true.



Amino acid	Codon	*P. patens*	*A. thaliana*	*S. cerevisiae*	*S. pombe*	*H. Sapiens*
Cys	UGC	+	+	-	+	+
	UGU	-	-	+	-	-
Glu	GAA	-	-	+	-	-
	GAG	+	+	-	+	+
Phe	UUC	+	+	+	+	+
	UUU	-	-	-	-	-
His	CAC	+	+	+	+	+
	CAU	-	-	-	-	-
Ile	AUA	-	-	/	-	-
	AUC	+	+	/	-	+
	AUU					
Lys	AAA	-	-	-	-	-
	AAG	+	+	+	+	+
Asn	AAC	+	+	+	+	+
	AAU	-	-	-	-	-
Gln	CAA	-	-	-	+	-
	CAG	+	+	+	-	+
Tyr	UAC	+	+	+	+	+
	UAU	-	-	-	-	-

https://www.researchgate.net/publication/320729847_Combination_of_the_Endogenous_lhcsr1_Promoter_and_Codon_Usage_Optimization_Boosts_Protein_Expression_in_the_Moss_Physcomitrella_patens

What’s a genomic sequence?

Protein centric view: data explodes

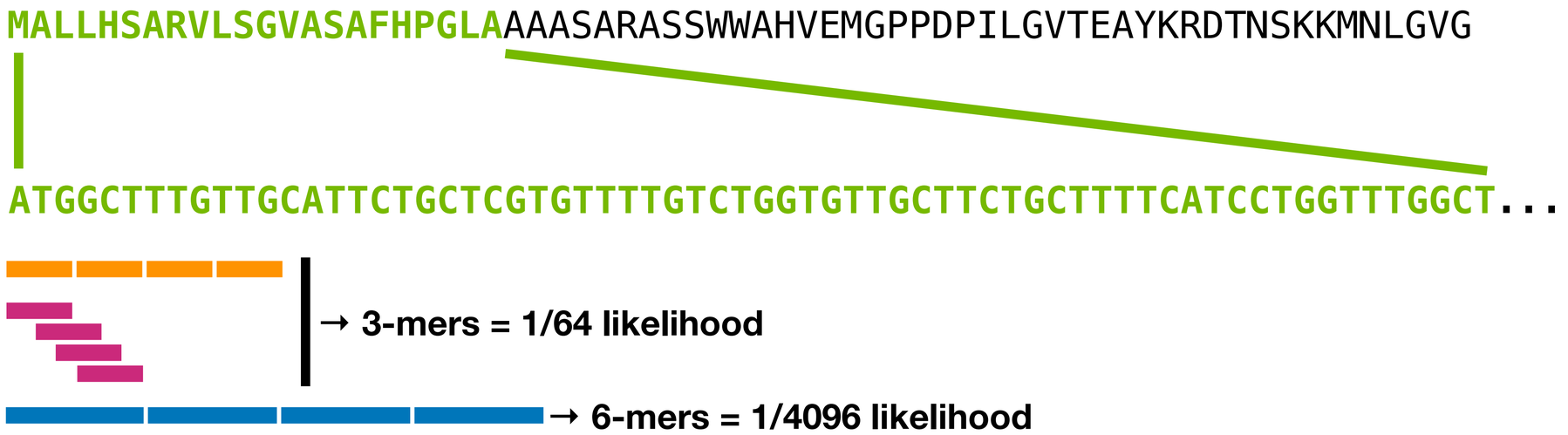
Length explosion

What’s a genomic sequence?

Codon view, more coarse or detailed?

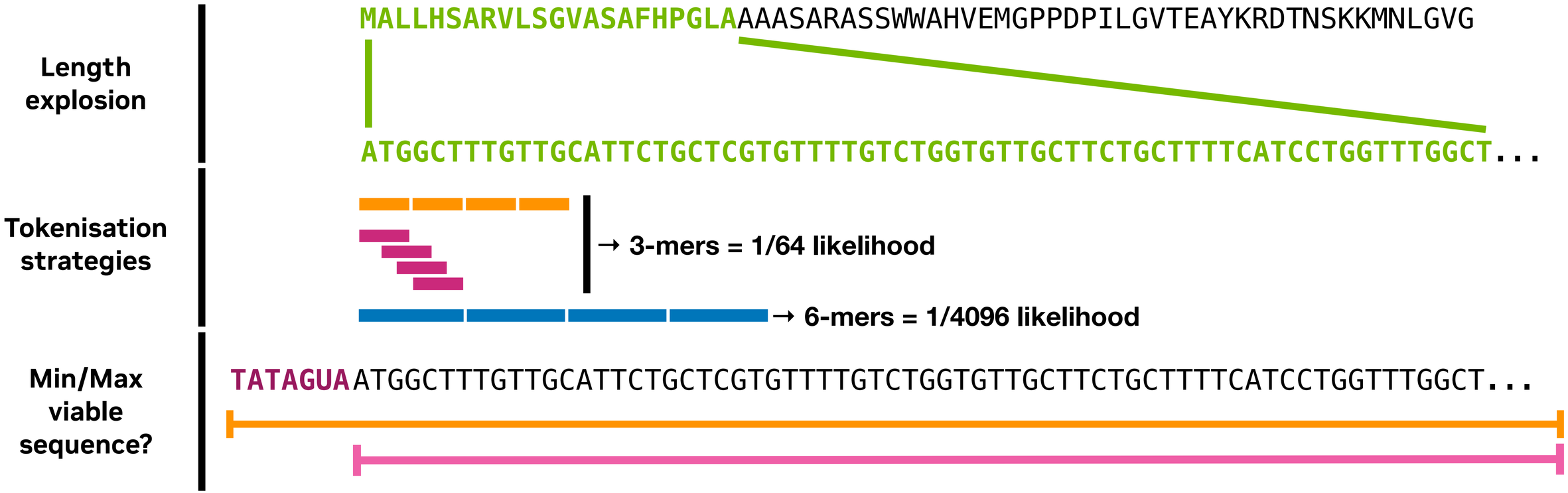
Length explosion

Tokenisation strategies



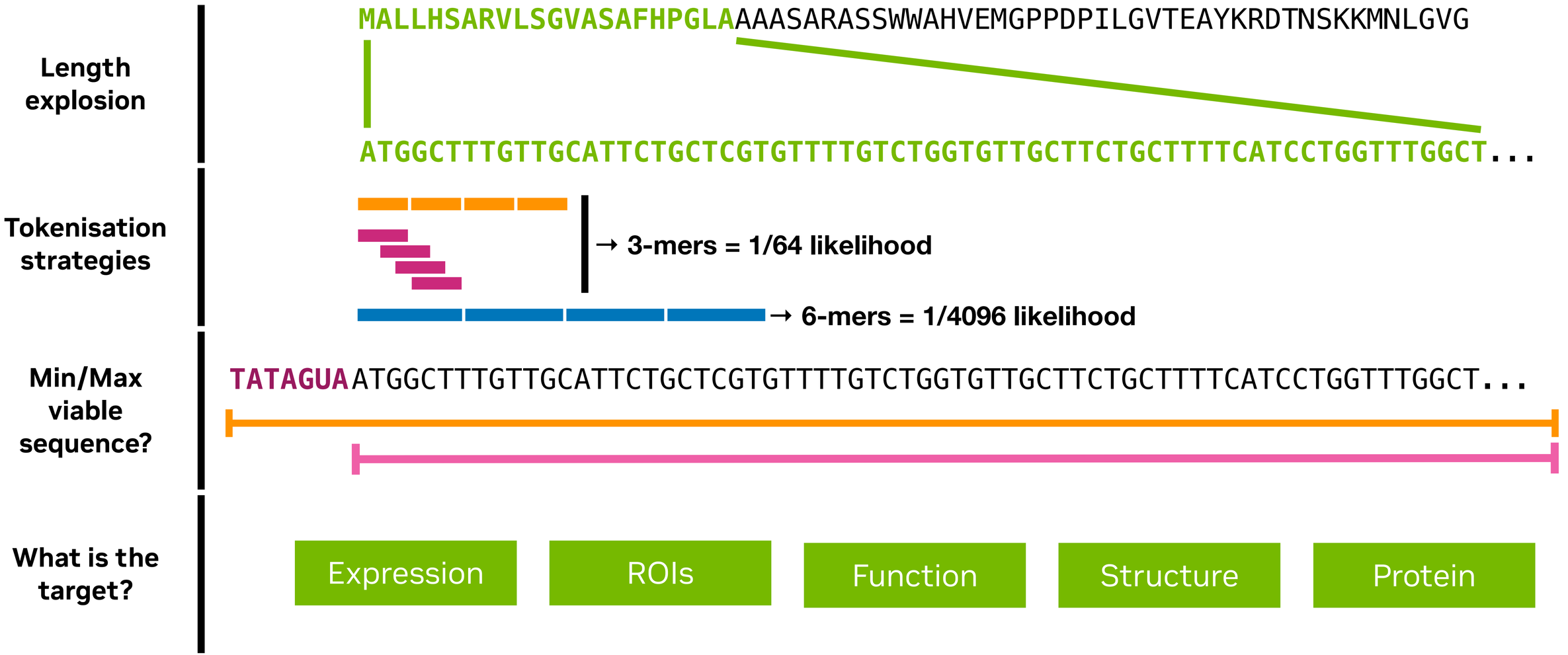
What's a genomic sequence?

Sequence = gene?



What's a genomic sequence?

Predict or generate?



The Nucleotide Transformer

Goal: general representation model for genomes

Dataset	Multi-species reference genomes (model organisms + more) [Sup. Tab. 3 & 4]
Tokenisation strategy	6-mers = 1/4096 likelihood (effectively 4104 with special tokens)
“Sequence” definition	Any 6000 nucleotide sequence (= 3000 tokens) [memory bound; overlapping 50 nuc]
What do we predict?	Interesting sites: promoters, enhancers, etc. + effect of mutation

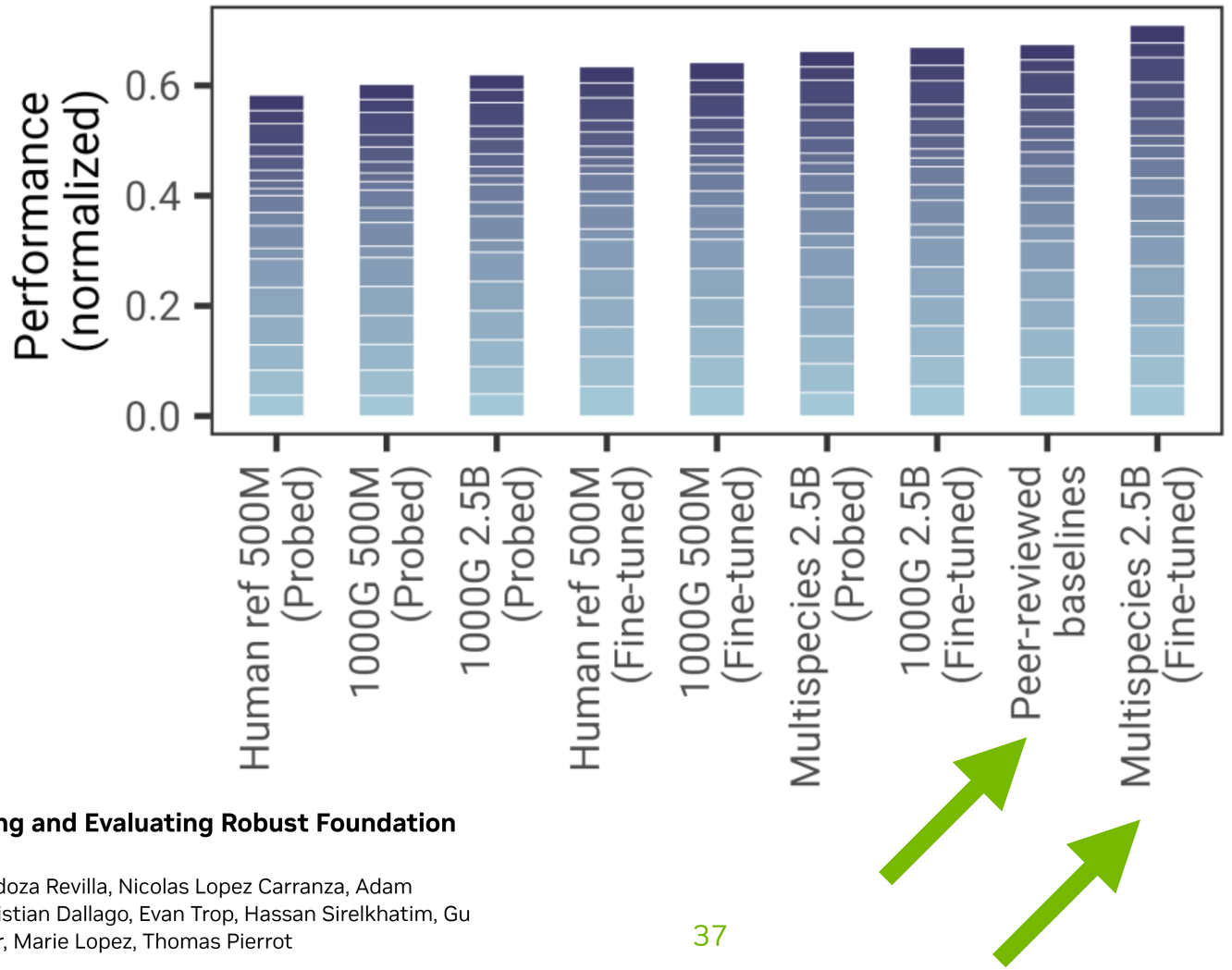
The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Hassan Sirelkhatim, Gu illaume Richard, Marcin Skwark, Karim Beguir, Marie Lopez, Thomas Pierrot

2023 BioRxiv

The Nucleotide Transformer

Goal: general representation model for genomes



Downstream features:

- Promoter regions
- Enhancer regions
- Splice sites

The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Hassan Sirelkhatim, Gu illaume Richard, Marcin Skwark, Karim Beguir, Marie Lopez, Thomas Pierrot

2023 BioRxiv

Genome-scale language models

Goal: learn genes —> generate genomes

Dataset	>110 million unique prokaryotic gene sequences from BV-BRC (bv-brc.org)
Tokenisation strategy	Codons, i.e. 3-mers = 1/64 likelihood
“Sequence” definition	Genes (language modelling) that come together into genomes (diffusion model)
What do we predict?	Generate new genomes → predict viral escape

GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics

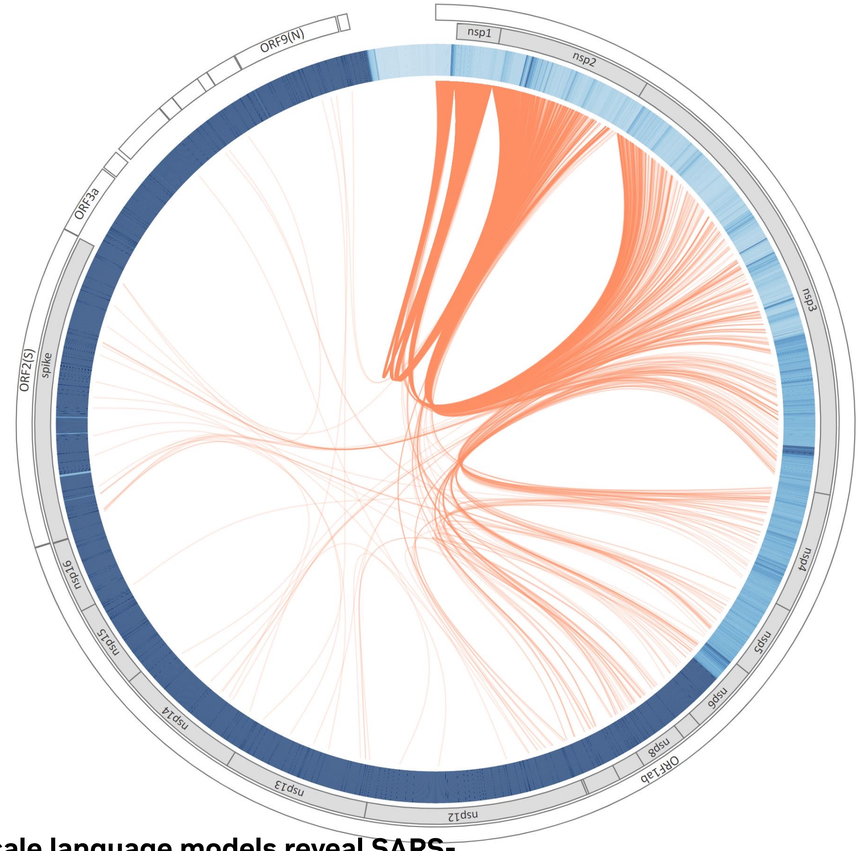
Maxim Zvyagin, Alexander Brace, Kyle Hippe, Yuntian Deng, Bin Zhang, Cindy Orozco Bohorquez, Austin Clyde, Bharat Kale, Danilo Perez-Rivera, Heng Ma, Carla M. Mann, Michael Irvin, J. Gregory Pauloski, Logan Ward, Valerie Hayot Sasson, Murali Emani, Sam Foreman, Zhen Xie, Diangen Lin, Maulik Shukla, Weili Nie, Josh Romero, Christian Dallago, Arash Vahdat, Chaowei Xiao, Thomas Gibbs, Ian Foster, James J. Davis, Michael E. Papka, Thomas Brettin, Rick Stevens, Anima Anandkumar, Venkatram Vishwanath, Arvind Ramanathan

2022 BioRxiv + Gordon Bell special prize

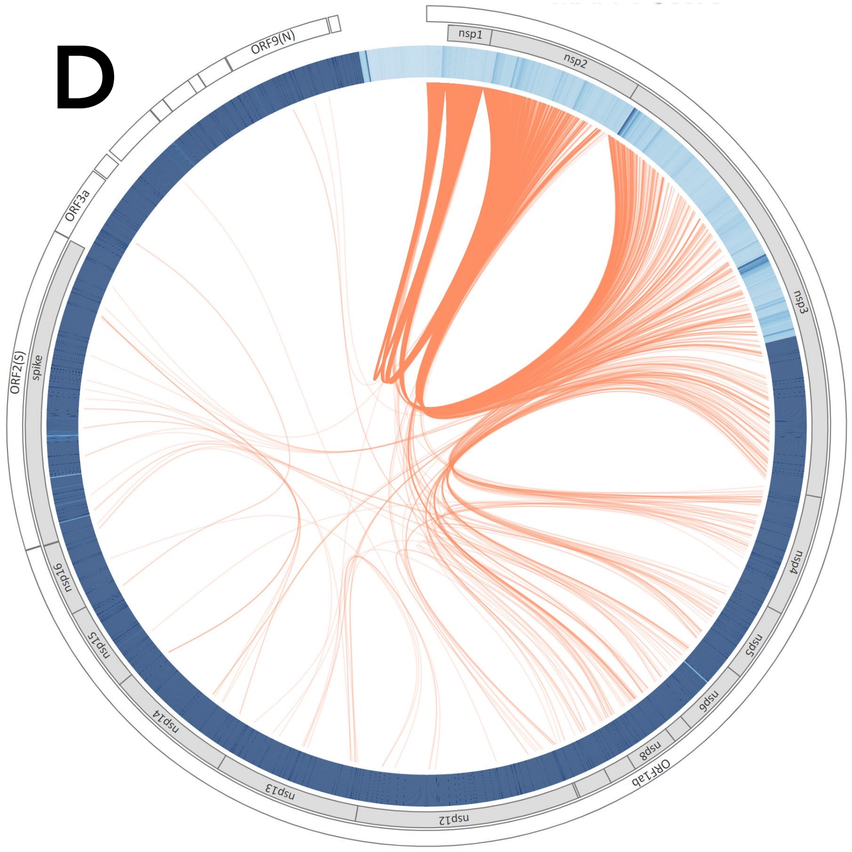
Genome-scale language models

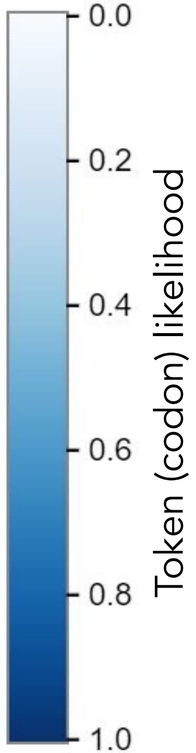
Goal: learn genes —> generate genomes

C



D





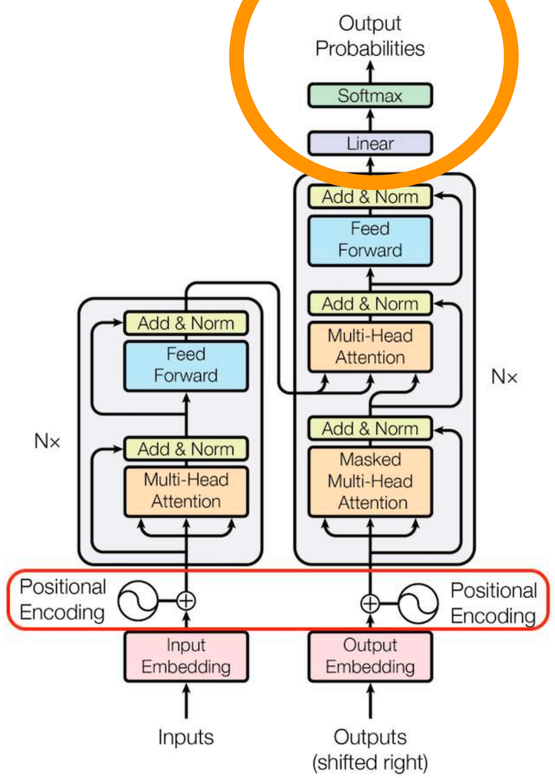
GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics

Maxim Zvyagin, Alexander Brace, Kyle Hippe, Yuntian Deng, Bin Zhang, Cindy Orozco Bohorquez, Austin Clyde, Bharat Kale, Danilo Perez-Rivera, Heng Ma, Carla M. Mann, Michael Irvin, J. Gregory Pauloski, Logan Ward, Valerie Hayot Sasson, Murali Emani, Sam Foreman, Zhen Xie, Diangen Lin, Maulik Shukla, Weili Nie, Josh Romero, Christian Dallago, Arash Vahdat, Chaowei Xiao, Thomas Gibbs, Ian Foster, James J. Davis, Michael E. Papka, Thomas Brettin, Rick Stevens, Anima Anandkumar, Venkatram Vishwanath, Arvind Ramanathan

2022 BioRxiv + Gordon Bell special prize


VESPA & FLIP

Path 1: what can we do right now?

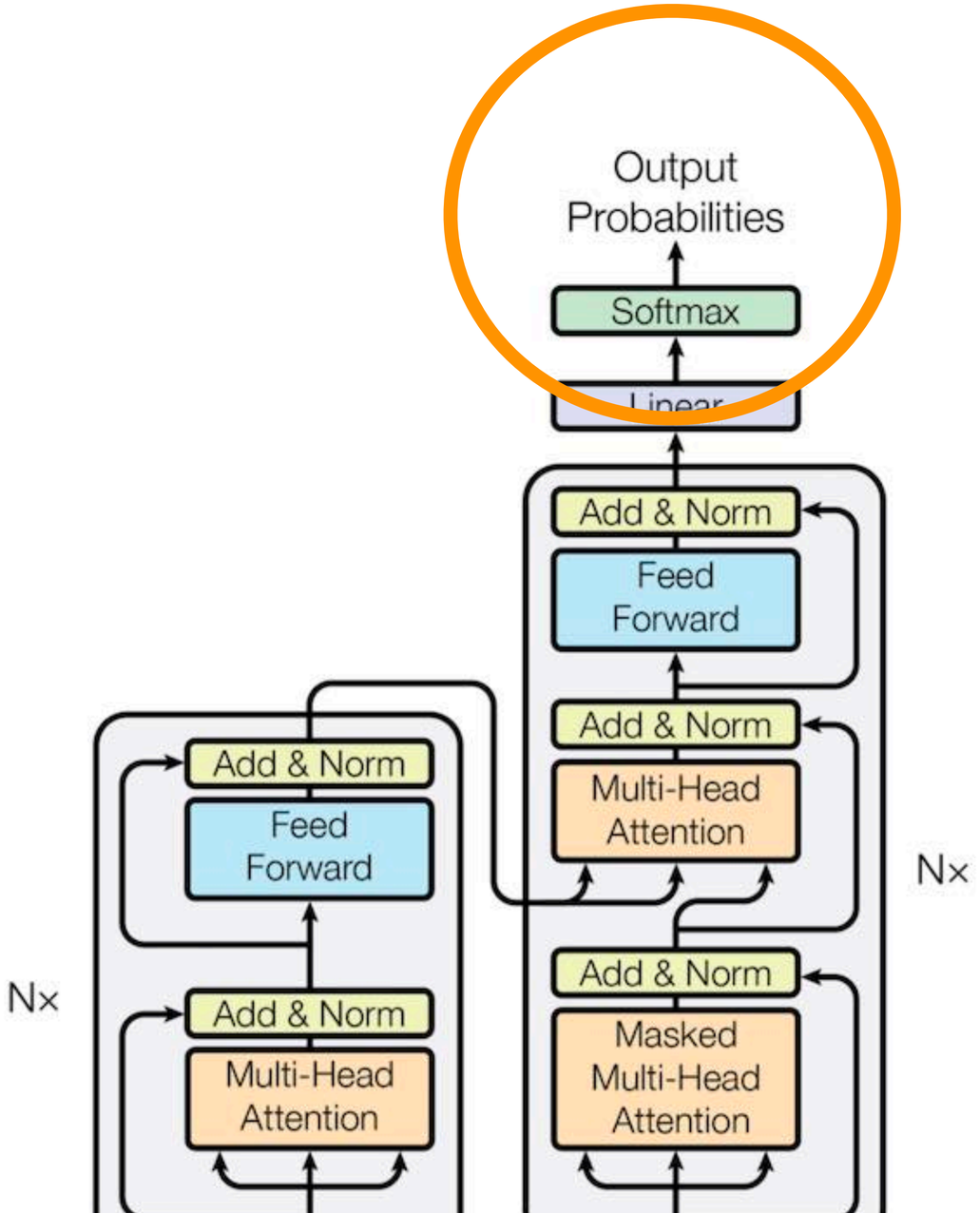


- Objective: Predict residue based on other residues and position in protein (MLM)

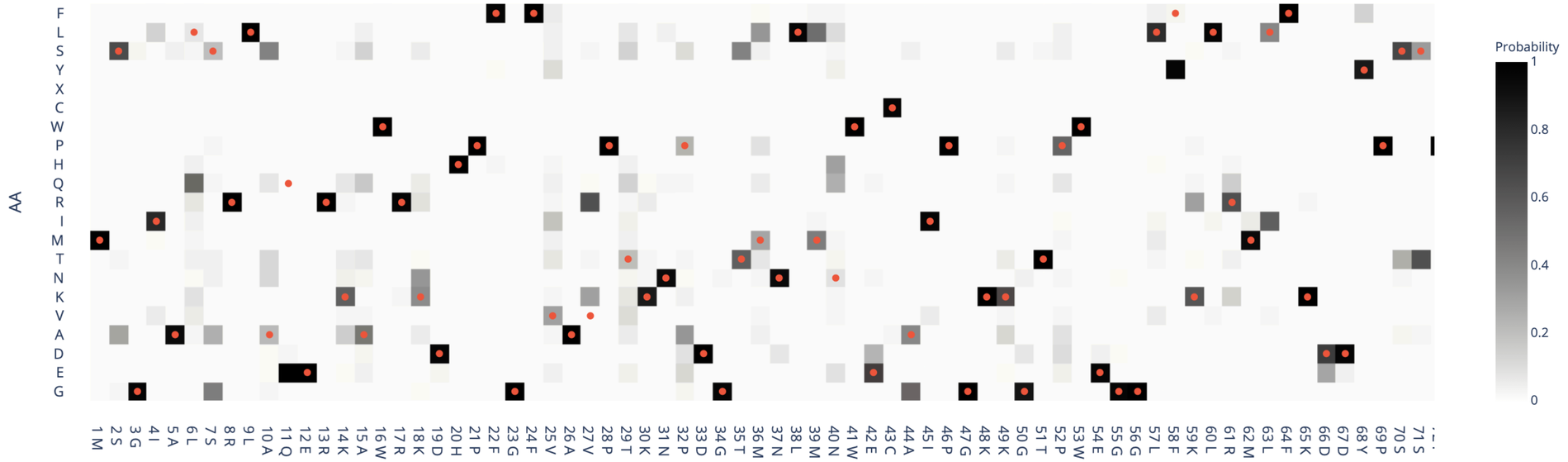
- Train on: BFD (the biggest sequence database available) —> Why?
  - Transformers have big parameter spaces = bigger datasets are beneficial

ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing
Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Debsindhu Bhowmik, Burkhard Rost
2020

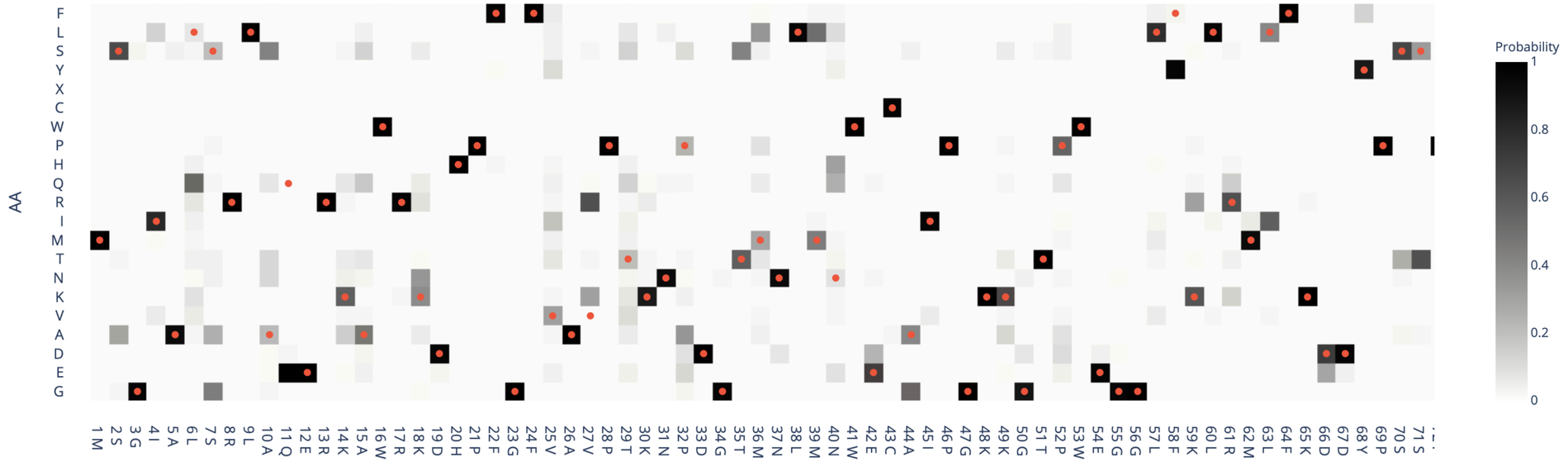
Path 1: what can we do right now?



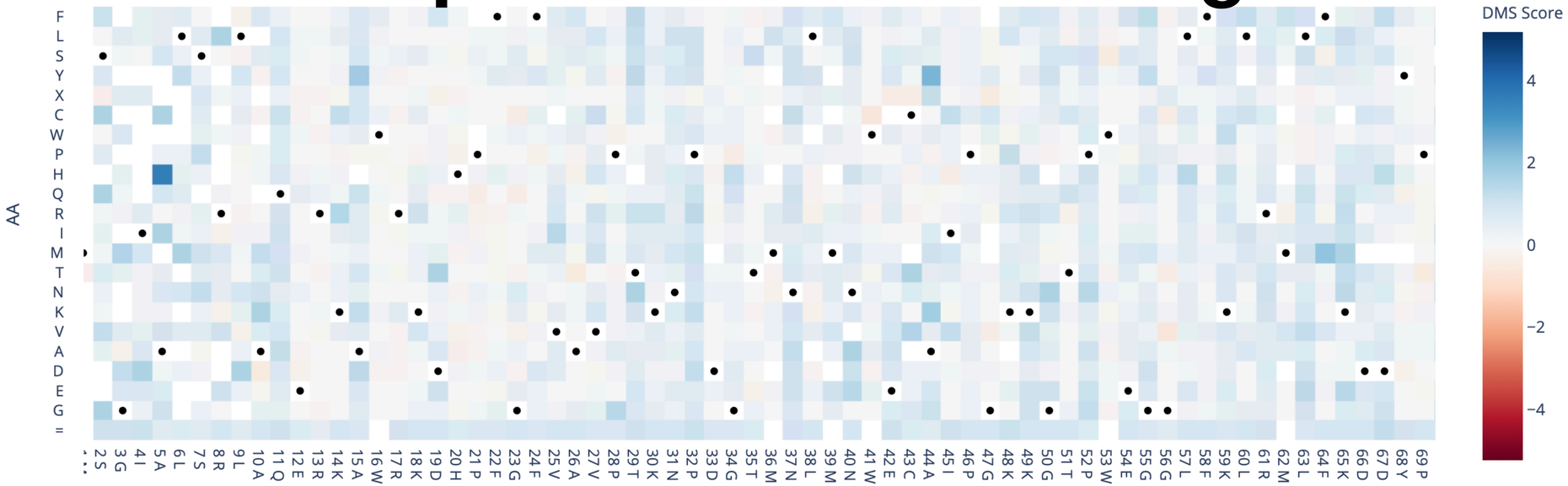
Residue probabilities from ProtBERT



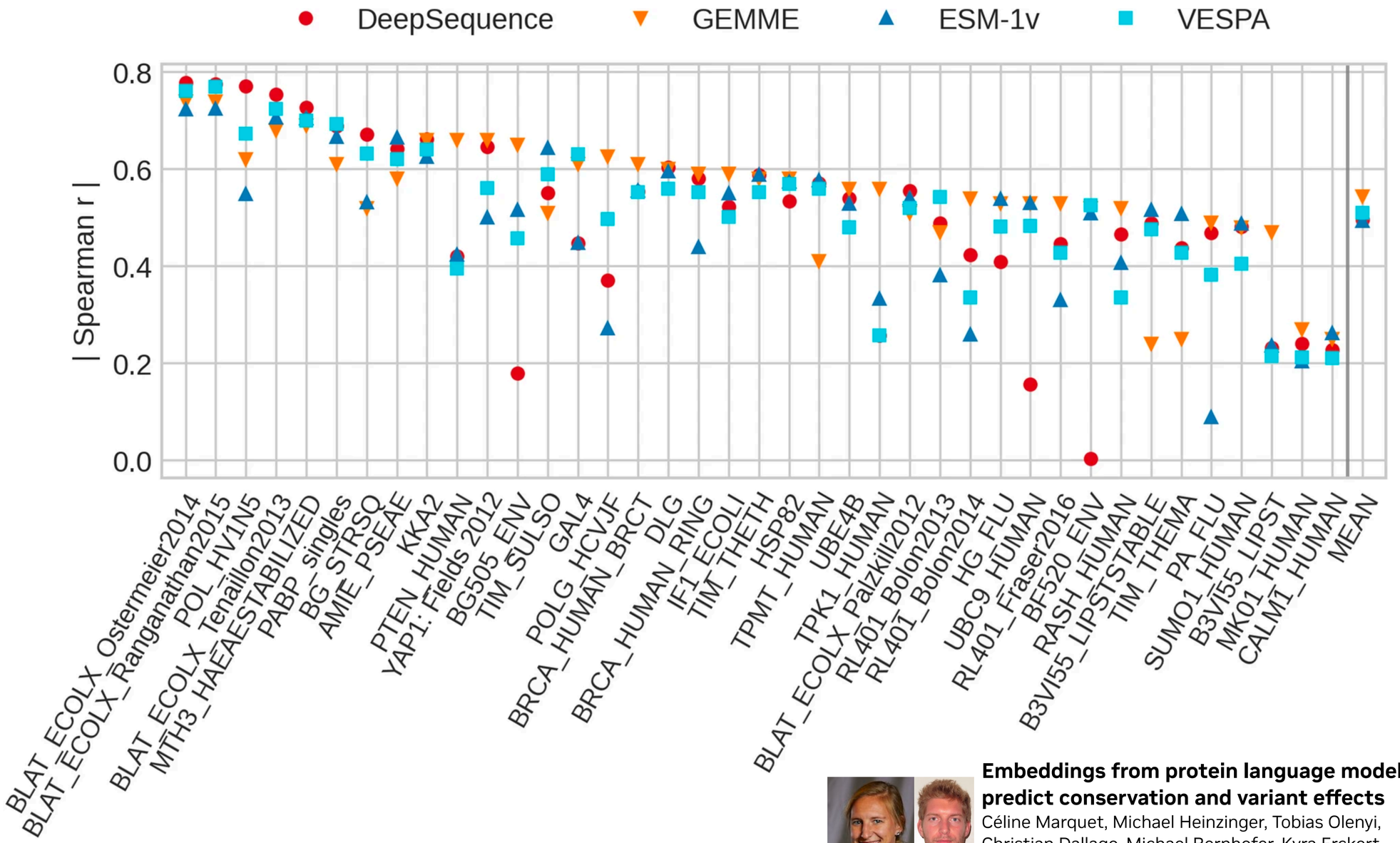
Residue probabilities from ProtBERT



Vs. DeepMutational scanning data



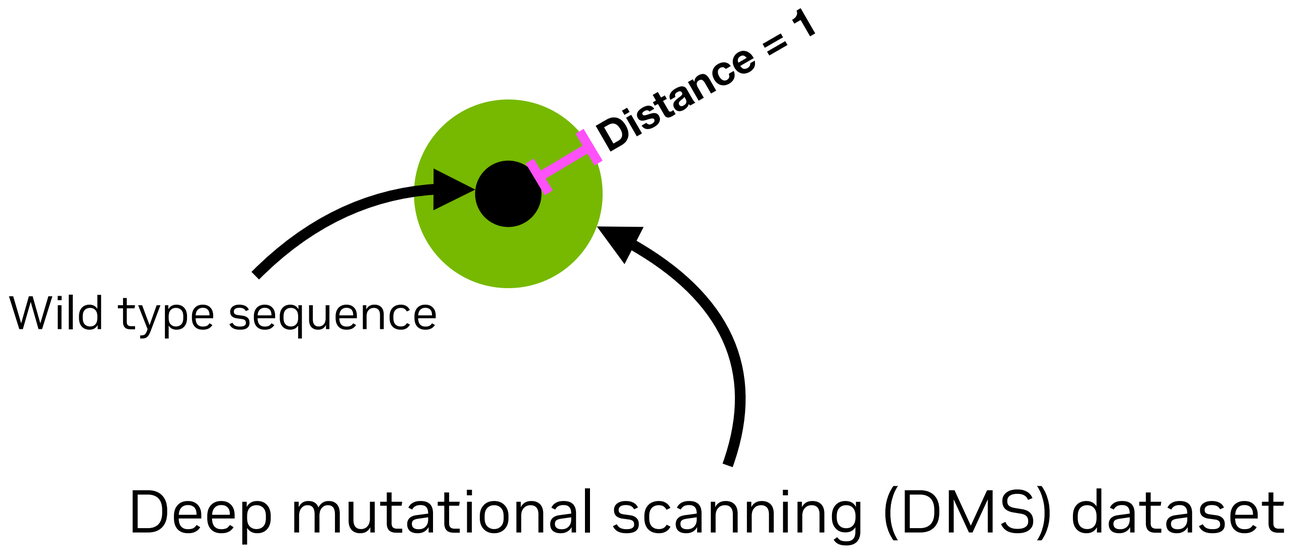
VESPA - predicting fitness from protein LMs



Embeddings from protein language models predict conservation and variant effects
Céline Marquet, Michael Heinzinger, Tobias Olenyi, Christian Dallago, Michael Bernhofer, Kyra Erckert, Burkhard Rost
2021 —> Human genetics

Build an evaluation framework

Can we go beyond single mutations?

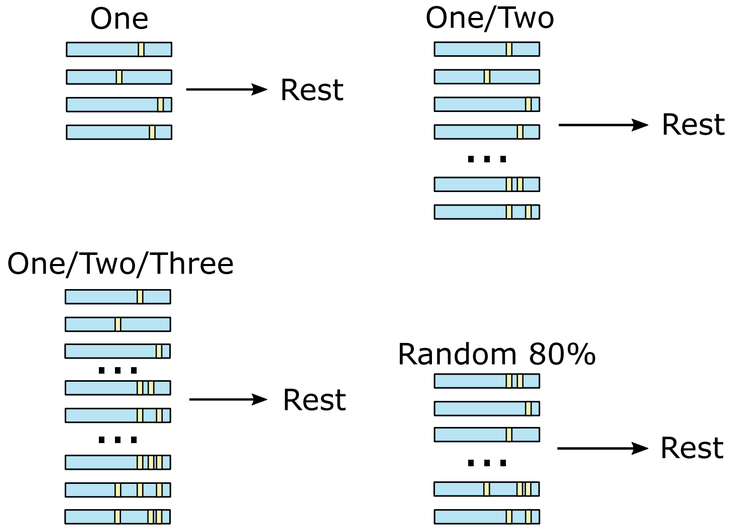


FLIP: Benchmark tasks in fitness landscape inference for proteins
Christian Dallago, Jody Mou, Kadina E Johnston, Bruce Wittmann, Nick Bhattacharya, Samuel Goldman, Ali Madani, Kevin K Yang
2021 —> NeurIPS

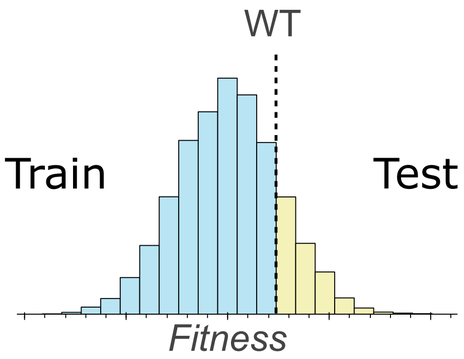
Build an evaluation framework

What are practically meaningful data splits?

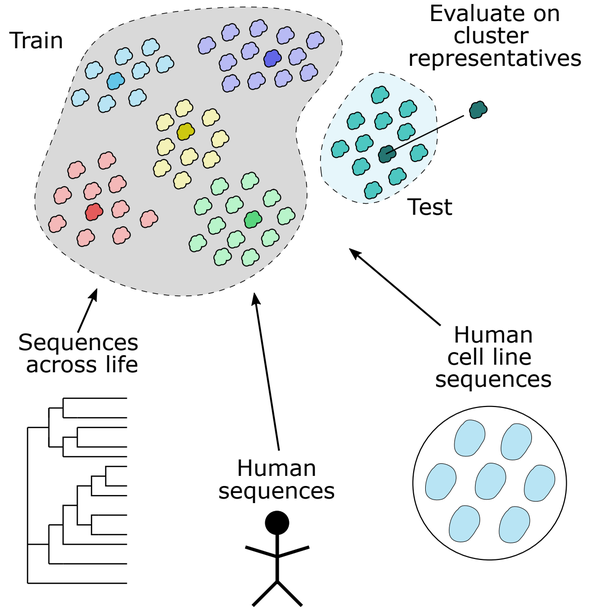
Predict more mutations from fewer mutations!



Predict higher activity!



Classic: predict sequence different by 20% sequence identity

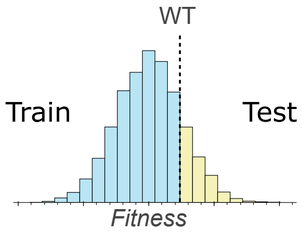


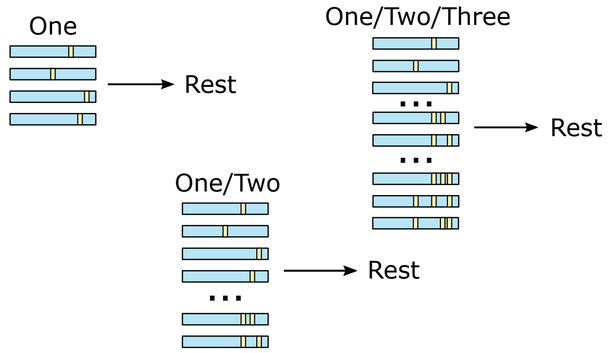
FLIP: Benchmark tasks in fitness landscape inference for proteins
Christian Dallago, Jody Mou, Kadina E Johnston, Bruce Wittmann, Nick Bhattacharya, Samuel Goldman, Ali Madani, Kevin K Yang
2021 —> NeurIPS

Build an evaluation framework

Table 4: GB1 baselines (metric: Spearman correlation)

Model	1-vs-rest	2-vs-rest	3-vs-rest	low-vs-high
ESM-1b (per AA)	0.28	0.55	0.79	0.59
ESM-1b (mean)	0.32	0.36	0.54	0.13
ESM-1b (mut mean)	-0.08	0.19	0.49	0.45
ESM-1v (per AA)	0.28	0.28	0.82	0.51
ESM-1v (mean)	0.32	0.32	0.77	0.10
ESM-1v (mut mean)	0.19	0.19	0.80	0.49
ESM-untrained (per AA)	0.06	0.06	0.48	0.23
ESM-untrained (mean)	0.05	0.05	0.46	0.10
ESM-untrained (mut mean)	0.21	0.21	0.57	0.13
Ridge	0.28	0.59	0.76	0.34
CNN	0.17	0.32	0.83	0.51
Levenshtein	0.17	0.16	-0.04	-0.10
BLOSUM62	0.15	0.14	0.01	-0.13





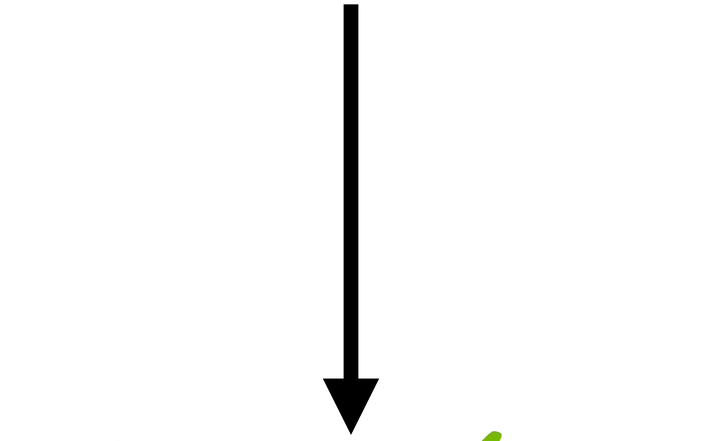
More training sequences

FLIP: Benchmark tasks in fitness landscape inference for proteins
Christian Dallago, Jody Mou, Kadina E Johnston, Bruce Wittmann, Nick Bhattacharya, Samuel Goldman, Ali Madani, Kevin K Yang
2021 —> NeurIPS


Protein Design (PGP)

What we want

MALLHSARVLSGVASAFHPGLAAAASARASSWWAHVEMGPPDPILGVTEAYKRDTNSKKMNLGVG

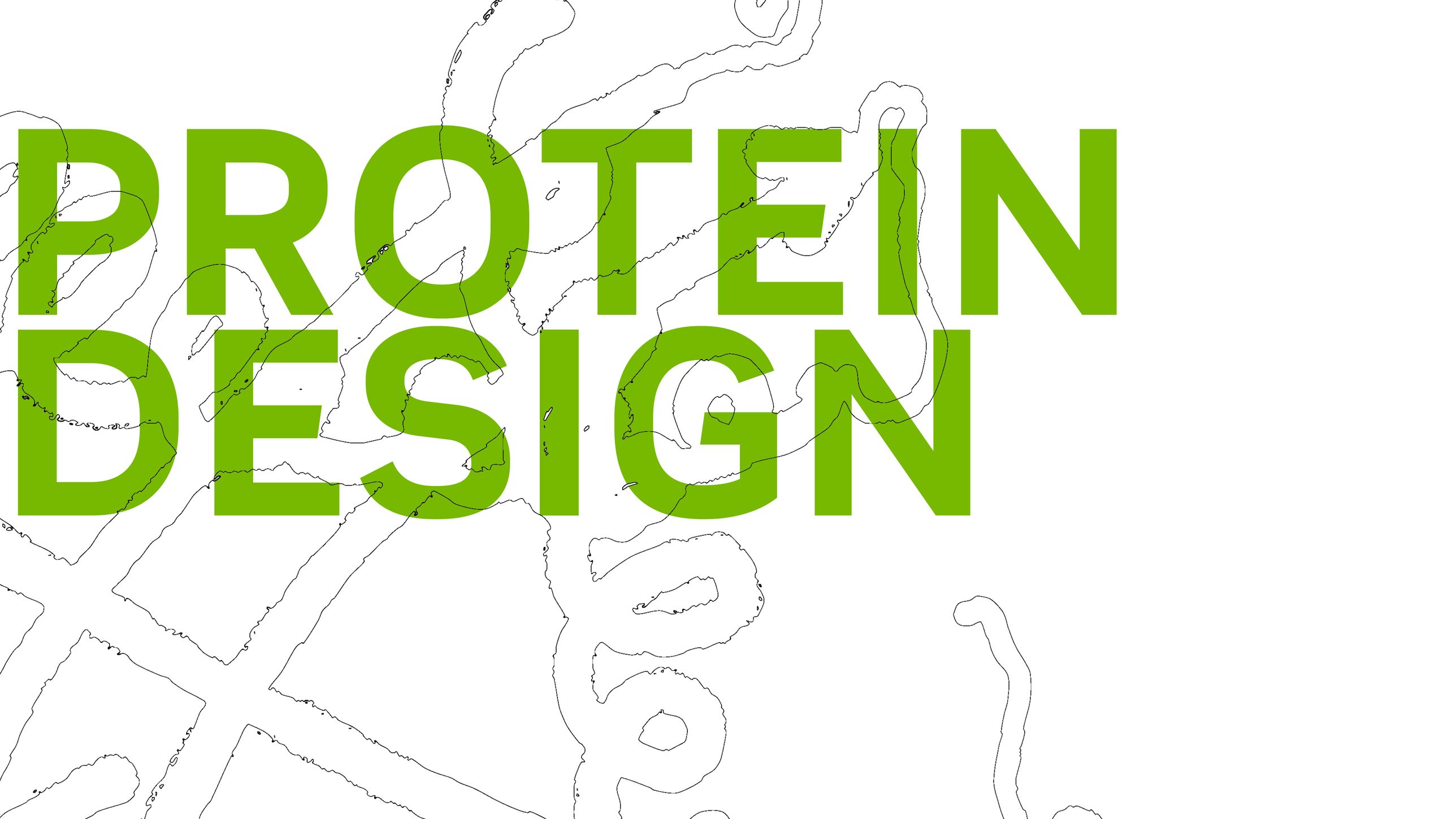


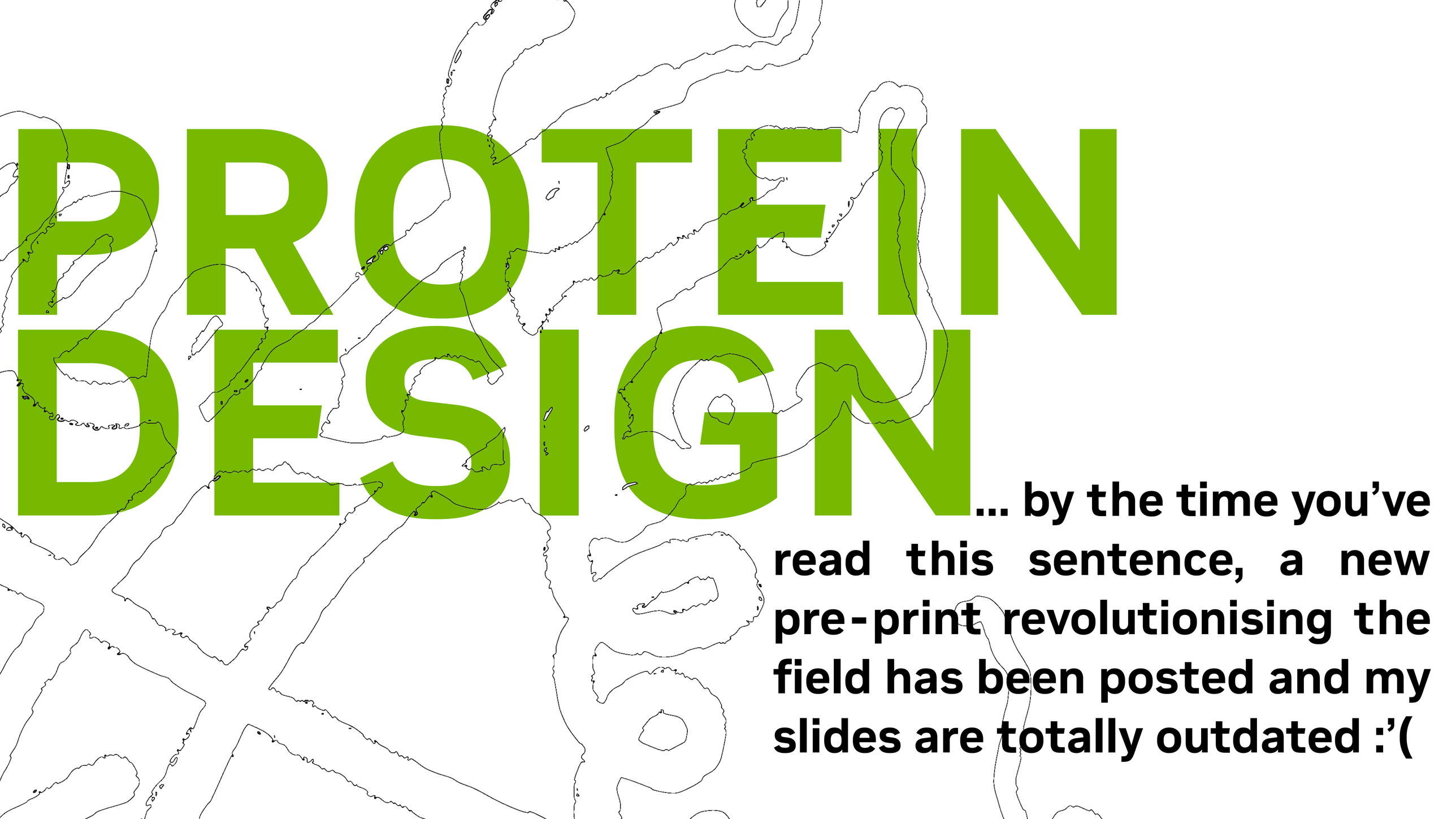
Something

What we want

MALLHSARVLSGVASAFHPGLAAAASARASSWWAHVEMGPPDPILGVTEAYKRDTNSKKMNLGVG

Something

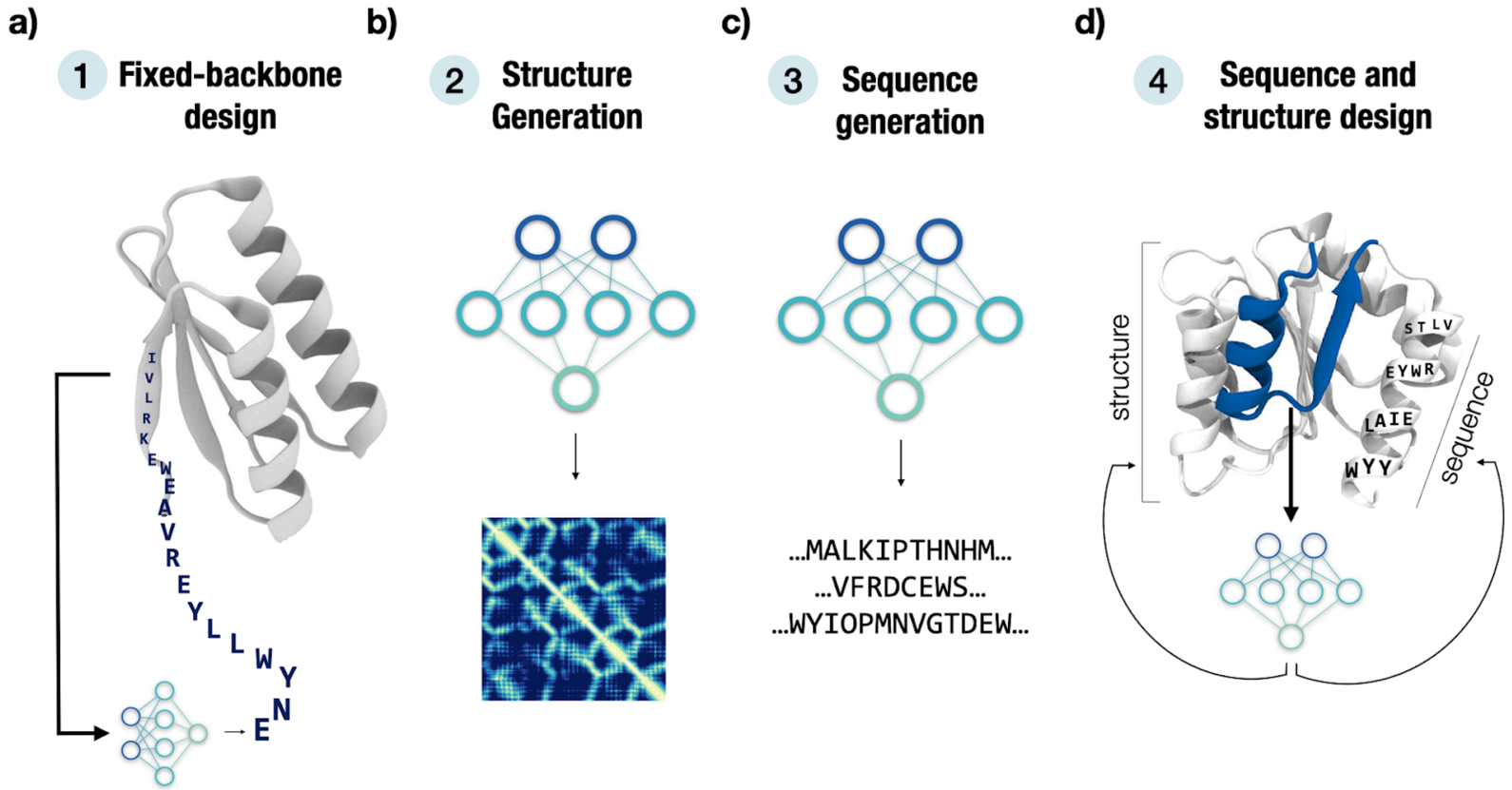

PROTEIN
DESIGN



PROTEIN DESIGN

... by the time you've read this sentence, a new pre-print revolutionising the field has been posted and my slides are totally outdated :'(

Protein design objectives

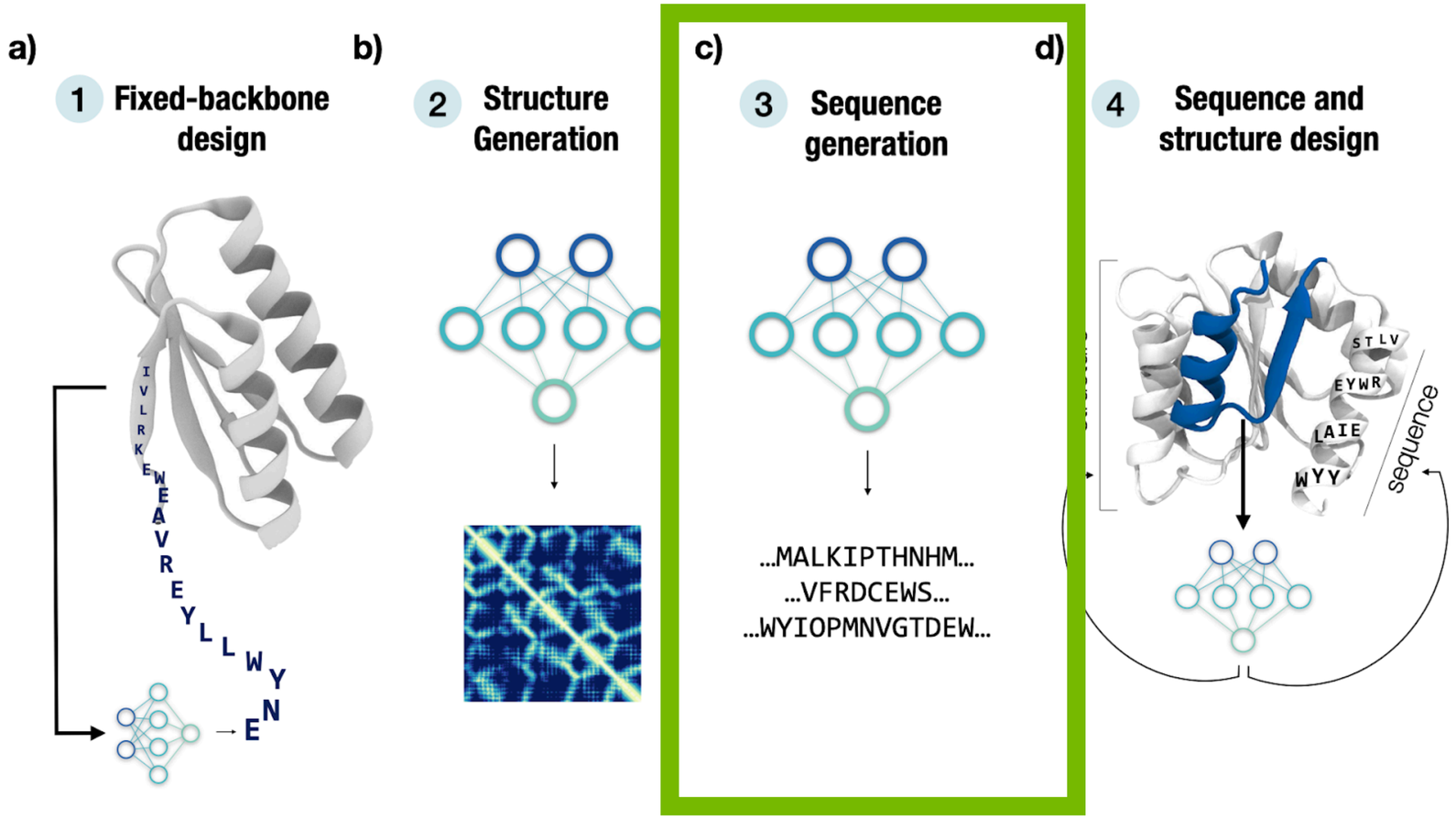


From sequence to function through structure: deep learning for protein design
Noelia Ferruz, Michael Heinzinger, Mehmet Akdel, Alexander Goncearenco, Luca Naef, Christian Dallago
2022 BioRxiv

Scientists are using AI to dream up revolutionary new proteins
Ewen Callaway
2022 Nature

Protein design objectives

Let’s focus on sequence generation



From sequence to function through structure: deep learning for protein design
Noelia Ferruz, Michael Heinzinger, Mehmet Akdel, Alexander Goncearenco, Luca Naef, Christian Dallago
2022 BioRxiv

Scientists are using AI to dream up revolutionary new proteins
Ewen Callaway
2022 Nature

Offline protein design pipeline

1 ProtGPT2

Generate protein sequences using a generative LLM

2 ProtT5

Predict features from the protein sequences using oracle LLM

3 LambdaPP

Select for proteins with interesting features

From sequence to function through structure: deep learning for protein design
Noelia Ferruz, Michael Heinzinger, Mehmet Akdel, Alexander Goncearenco, Luca Naef, Christian Dallago
2022 BioRxiv

Offline protein design pipeline

Way faster than wetlab experiments!

1 ProtGPT2

Generate protein sequences using a generative LLM

2 ProtT5

Predict features from the protein sequences using oracle LLM

3 LambdaPP

Select for proteins with interesting features

Done in ~3h for 100’000 sequences

From sequence to function through structure: deep learning for protein design
Noelia Ferruz, Michael Heinzinger, Mehmet Akdel, Alexander Goncearenco, Luca Naef, Christian Dallago
2022 BioRxiv

Offline protein design pipeline

Protein sequence generation

1 ProtGPT2

Generate protein sequences using a generative LLM

- ProtGPT2 is a decoder-only architecture of 36 layers and 738 million parameters
- Generate unconditionally (i.e., without priors on family, function or structure)
- Temperature value fixed to get diverse yet not non-sense sequences

Generated library
100’000 sequences

From sequence to function through structure: deep learning for protein design
Noelia Ferruz, Michael Heinzinger, Mehmet Akdel, Alexander Goncearenco, Luca Naef, Christian Dallago
2022 BioRxiv

Offline protein design pipeline

Protein prediction from sequence

2 ProtT5

Predict features from the protein sequences using oracle LLM

- Disorder
- Conservation
- Single peptides
- Subcellular location
- Secondary structure
- CATH structure class
- Topology (membrane structure)
- Binding ability (metal, nucleotides, small molecules)
- Gene Ontology (molecular function, biological process, cellular compartment)
- … and more

Generated library
100’000 sequences

UniRef90 sampled
100’000 sequences

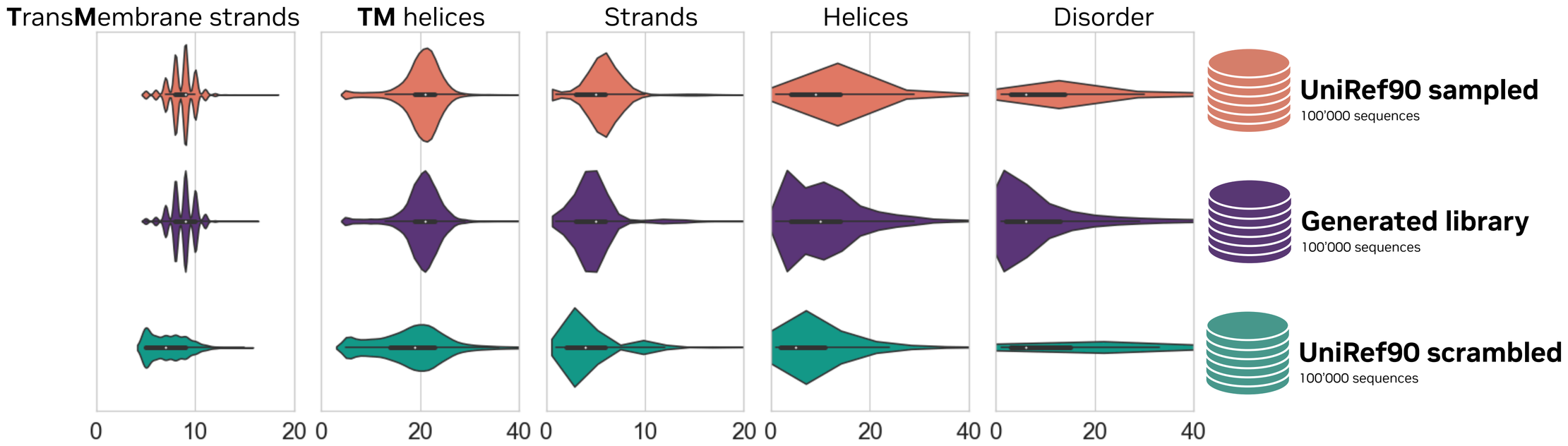
UniRef90 scrambled
100’000 sequences

From sequence to function through structure: deep learning for protein design
Noelia Ferruz, Michael Heinzinger, Mehmet Akdel, Alexander Goncearenco, Luca Naef, Christian Dallago
2022 BioRxiv

Offline protein design pipeline

Assessment: generated vs. natural

What is the distribution of stretch (consecutive amino acids involved in some structural motif) lengths?



From sequence to function through structure: deep learning for protein design
Noelia Ferruz, Michael Heinzinger, Mehmet Akdel, Alexander Goncearenco, Luca Naef, Christian Dallago
2022 BioRxiv

Offline protein design pipeline

Select sequences of interest

3 LambdaPP

Select for proteins with interesting features

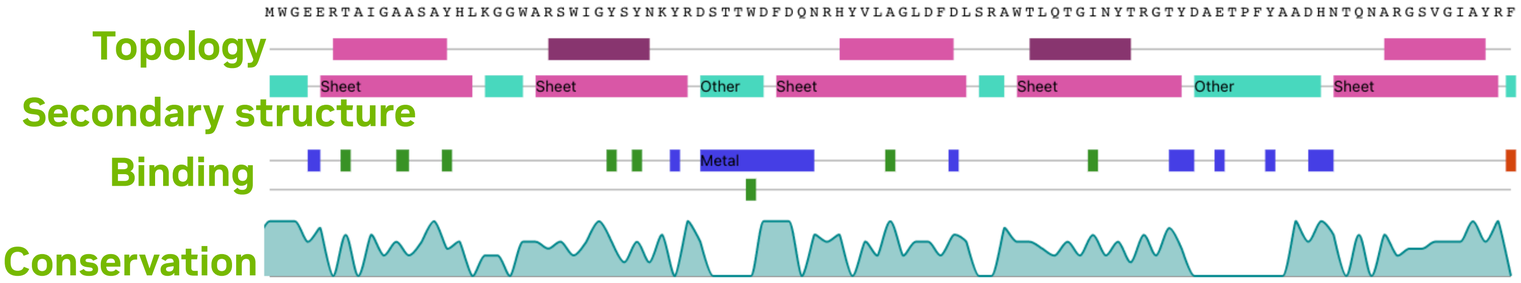
```
filtering_order = ['length', 'transmembrane_strand_percent', 'small_percent']

for sequence in data.query(
    '''
    length < 100 and \
    transmembrane_strand_count > 0 and \
    small_count > 0
    '''
).sort_values(filtering_order, ascending=False)
```

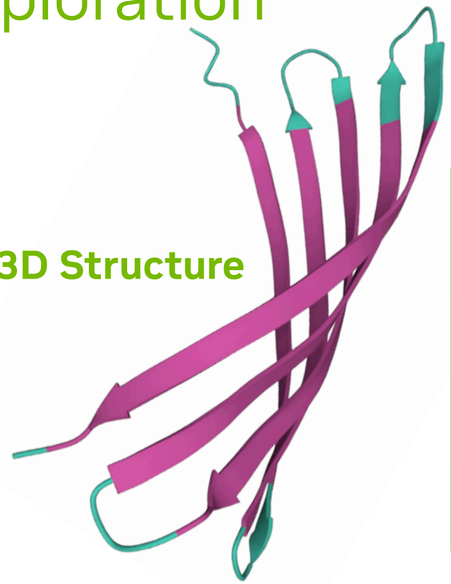
Header: >seq86311, L=99, ppl=61.625
Sequence length: 99
Transmembrane strand content: 47.47%
Small-molecule binding content: 8.08%
View on LambdaPP: https://embed.predictprotein.org/#/
MWGEERTAIGAASAYHLKGGWARSWIGYSYNKYRDSTTWDFDQNRHYVLAGLDFDLSRAWT
LQTGINYTRGTYDAETPFYAADHNTQNARGSVGIAYRF

From sequence to function through structure: deep learning for protein design
Noelia Ferruz, Michael Heinzinger, Mehmet Akdel, Alexander Goncearenco, Luca Naef, Christian Dallago
2022 BioRxiv

LambdaPP

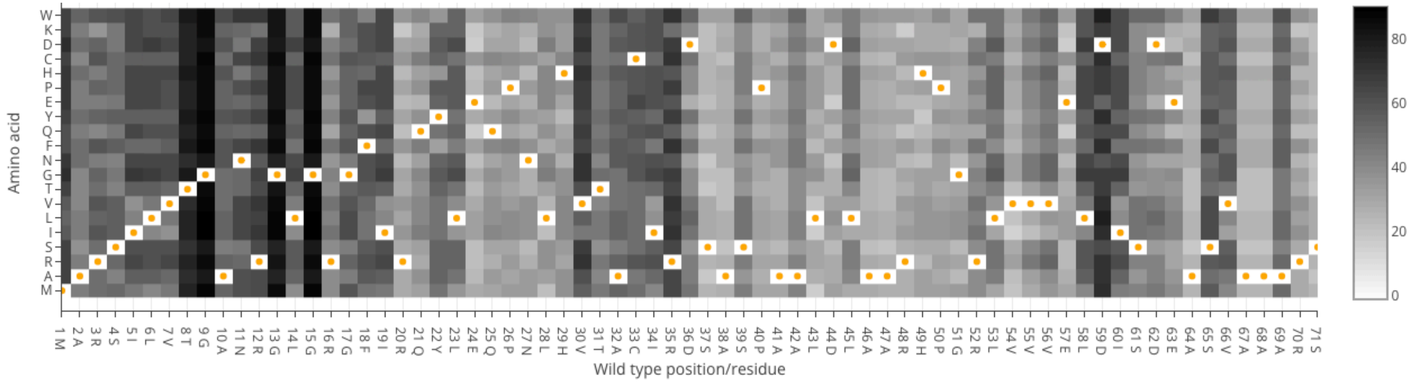
Visual protein predictions exploration





Subcellular location

Single Amino Acid Variant Effect



Biological process (BPO)		
Reference Seq.	**GO Name**	**Reliability Index**
Q9L7R3	transmembrane transport	0.31
Q9L7R3	oligosaccharide transport	0.31
Q9L7R3	ion transport	0.31

Molecular function (MFO)		
Reference Seq.	**GO Name**	**Reliability Index**
Q9L7R3	porin activity	0.31

Gene Ontology

Cellular Component (CCO)		
Reference Seq.	**GO Name**	**Reliability Index**
Q47534	cell outer membrane	0.33

https://embed.predictprotein.org/#/
MWGEERTAIGAASAYHLKGGWARSWIGYSYNKYRDSTTWDFDQNRHYV
LAGLDFDLSRAWTLQTGINYTRGTYDAETPFYAADHNTQNARGSVGIAYRF

LambdaPP: Fast and accessible protein-specific phenotype predictions
Olenyi, Marquet et al., Dallago & Rost.
2022 BioRxiv

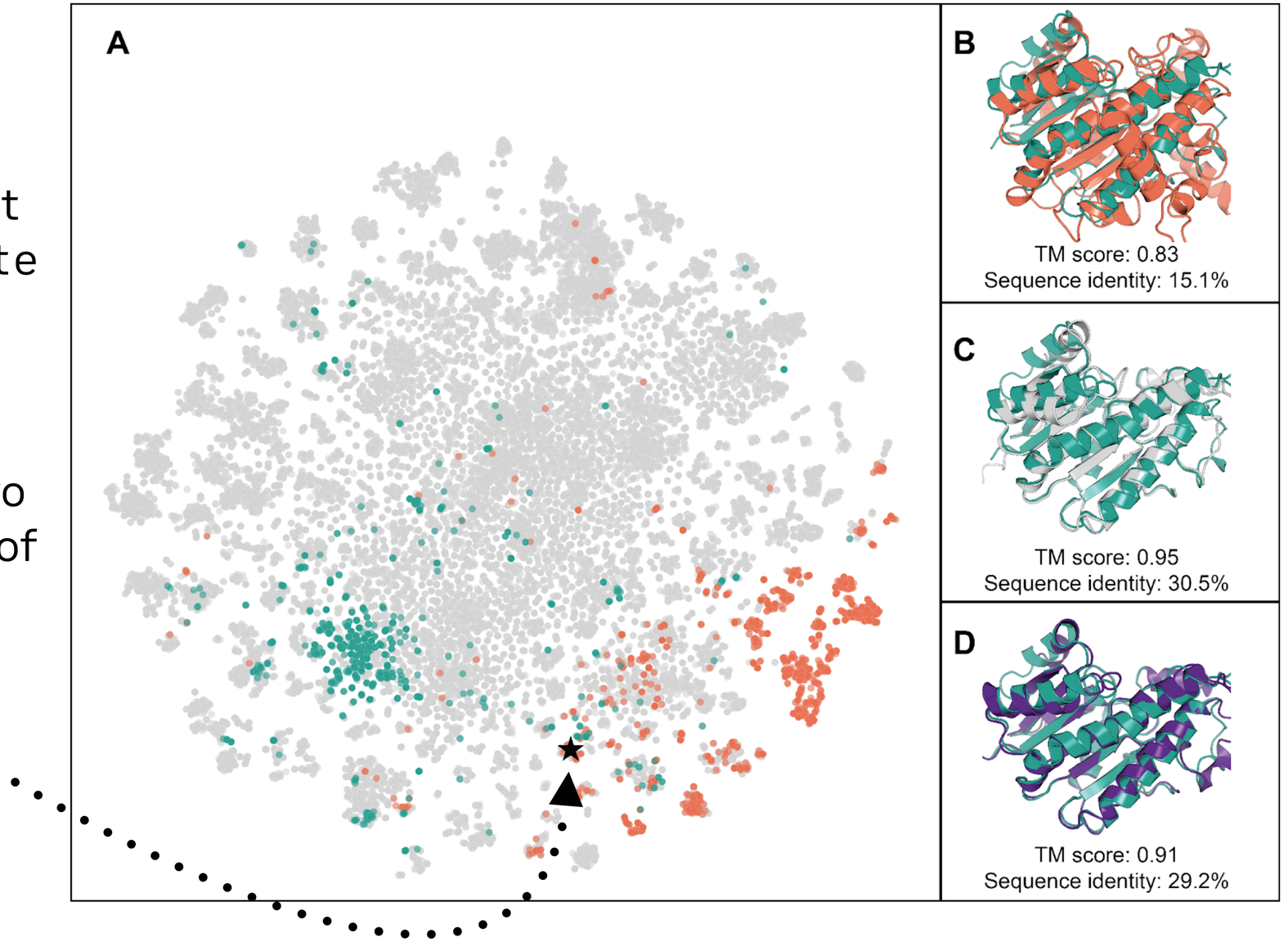
NVIDIA

Found a sequence diverse BGC sequence

generated sequences that match “secondary metabolite biosynthetic process”.

sequences from two databases known to be part of biosynthetic gene clusters.

★ selected generated sequence.



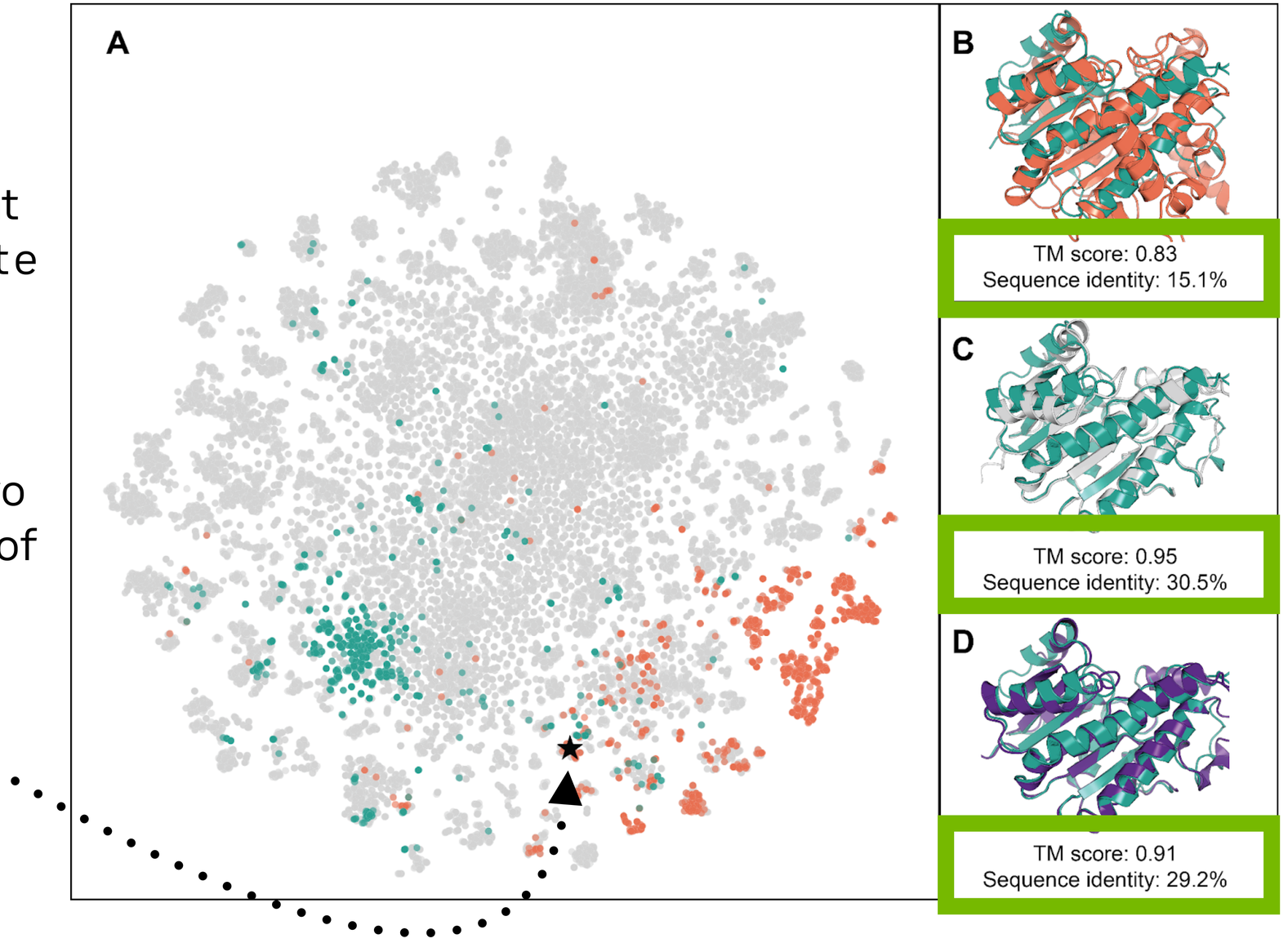
From sequence to function through structure: deep learning for protein design
Noelia Ferruz, Michael Heinzinger, Mehmet Akdel, Alexander Goncearenco, Luca Naef, Christian Dallago
2022 BioRxiv

Found a sequence diverse BGC sequence

generated sequences that match “secondary metabolite biosynthetic process”.

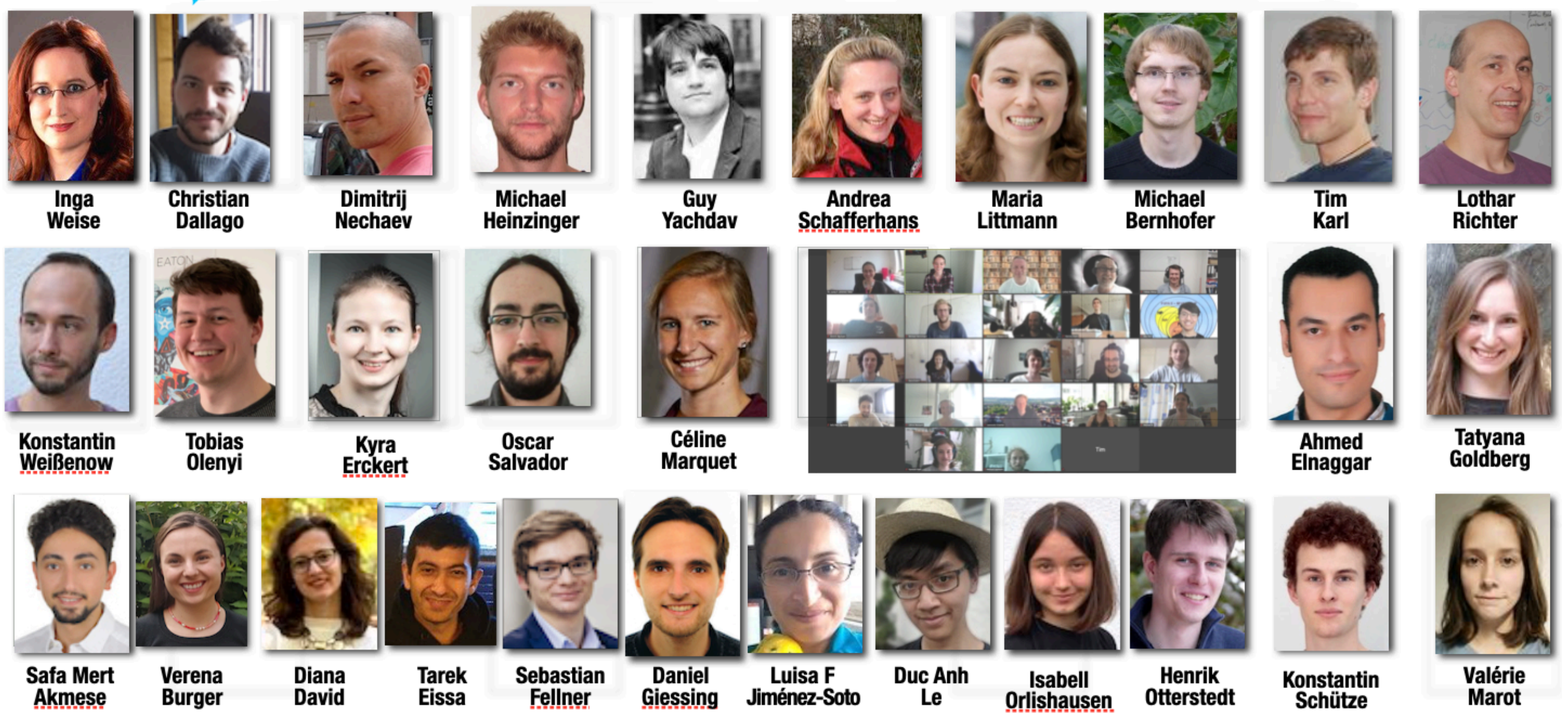
sequences from two databases known to be part of biosynthetic gene clusters.

★ selected generated sequence.



From sequence to function through structure: deep learning for protein design
Noelia Ferruz, Michael Heinzinger, Mehmet Akdel, Alexander Goncearenco, Luca Naef, Christian Dallago
2022 BioRxiv

www.rostlab.org



Burkhard Rost

Chris Dallago

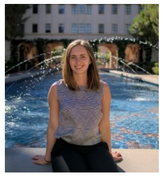
Chris Sander @ Harvard

Michael Heinzinger

Thank you!

Kevin K. Yang @ MSRNE

Jody Mou @ MIT



Kadina E. Johnston @ Caltech

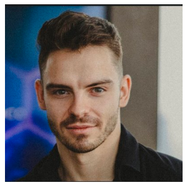
Bruce Wittmann @ Caltech

Nick Bhattacharya @ UC Berkeley

Samuel Goldman @ MIT

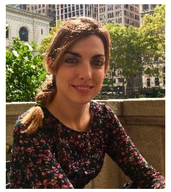
Ali Madani @ Profluent Bio

Alexander Goncearenco @ VantAI



Luca Naef @ VantAI

Mehmet Akdel @ VantAI



Noelia Ferruz @ Uni Girona

Yana Bromberg @ Rutgers

Arvind Ramanathan @ Argonne