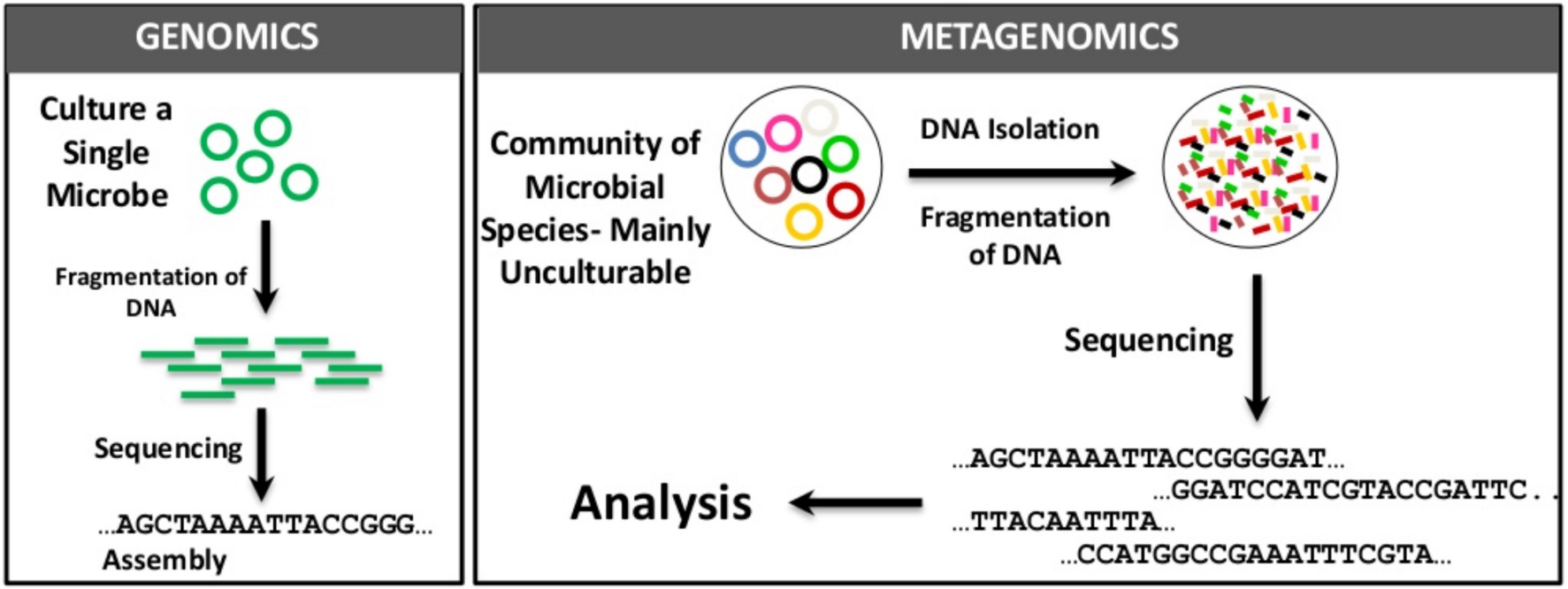
Recap of day 1

- What does it mean that two proteins are homologous?
- What is homology inference?
- What is a P-value? What is an E-value?
- How can we still find homologies between eukaryotic, bacterial and archaeal proteins, given the many mutations per amino acid since the Last Common Ancestor of all life)?
- Give another name for amino acid substitution matrix that would make sense.
- How are substitution matrix scores computed?
- How are sequence profile scores computed?
- What is iterative profile search? What tools exist?
- What are protein domains? What is their relevance?
- Why are some parts of proteins disordered (unstructured)?
- What are the key ideas of the algorithm to compute the best-scoring alignment between two sequences?

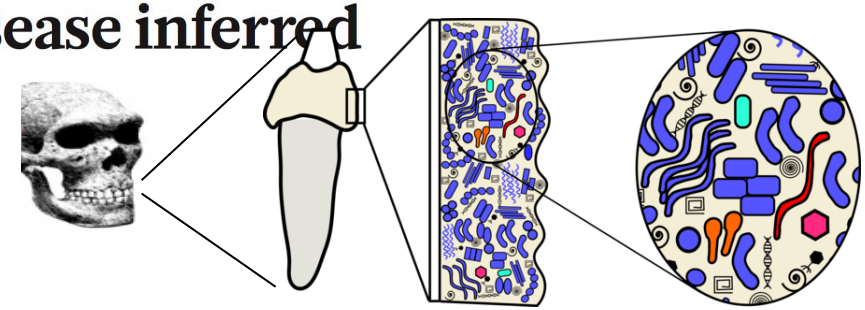
With metagenomics we can study the ~99% uncultivable microbes by sequencing their DNA directly from environment



Metagenomics age of enlightenment

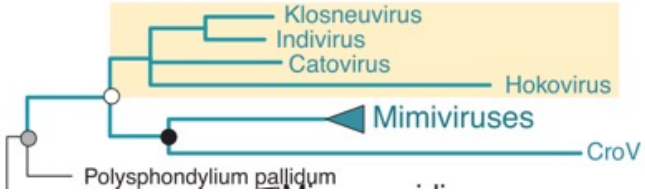
Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus

Nature 2017, Apr 20



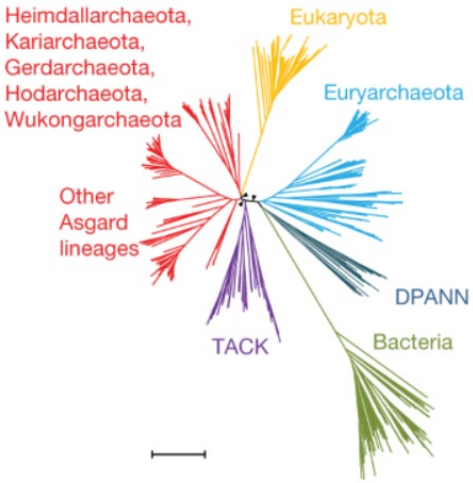
Giant viruses with an expanded complement of translation system components

Science 2017, Apr 7



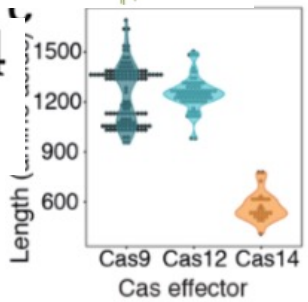
Expanded diversity of Asgard archaea and their relationships with eukaryotes

Nature 2021, Apr 7



Programmed DNA destruction by miniature CRISPR-Cas14 enzymes

Science 2018, Nov 16



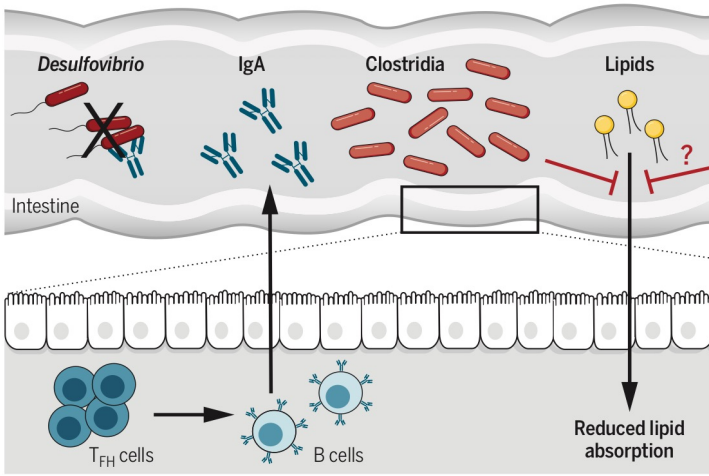
Metagenomics age of enlightenment

Neonatal selection by Toll-like receptor 5 influences long-term gut microbiota composition

Nature 2018, Aug 23

Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth

Nature 2019, Oct03



T cell-mediated regulation of the microbiota protects against obesity

Science 2019, Jul 26

The microbiota regulate neuronal function and fear extinction learning

Nature 2019, Oct 23

Potential role of indolelactate and butyrate in multiple sclerosis revealed by integrated microbiome-metabolome analysis

Cell Rep Med 2021, Apr 20

Metagenomics age of enlightenment

Neonatal selection by Toll-like receptor 5 influences long-term gut microbiota composition

Nature 2018, Aug 23

Stunted microbiota and opportunistic pathogen colonizat

Nature 2019

T cell-m microbi

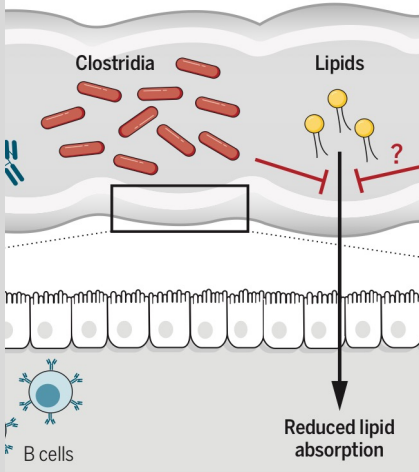
Science 20

The micr learning

Nature 2019, Oct 23

Applications:

- **Human health (gut, skin, ...)**
- **Ecology & climate**
- **Enzymes for biotechnology**
- **New drugs and natural compounds**
- **Evolution, tree of life**
- **...**

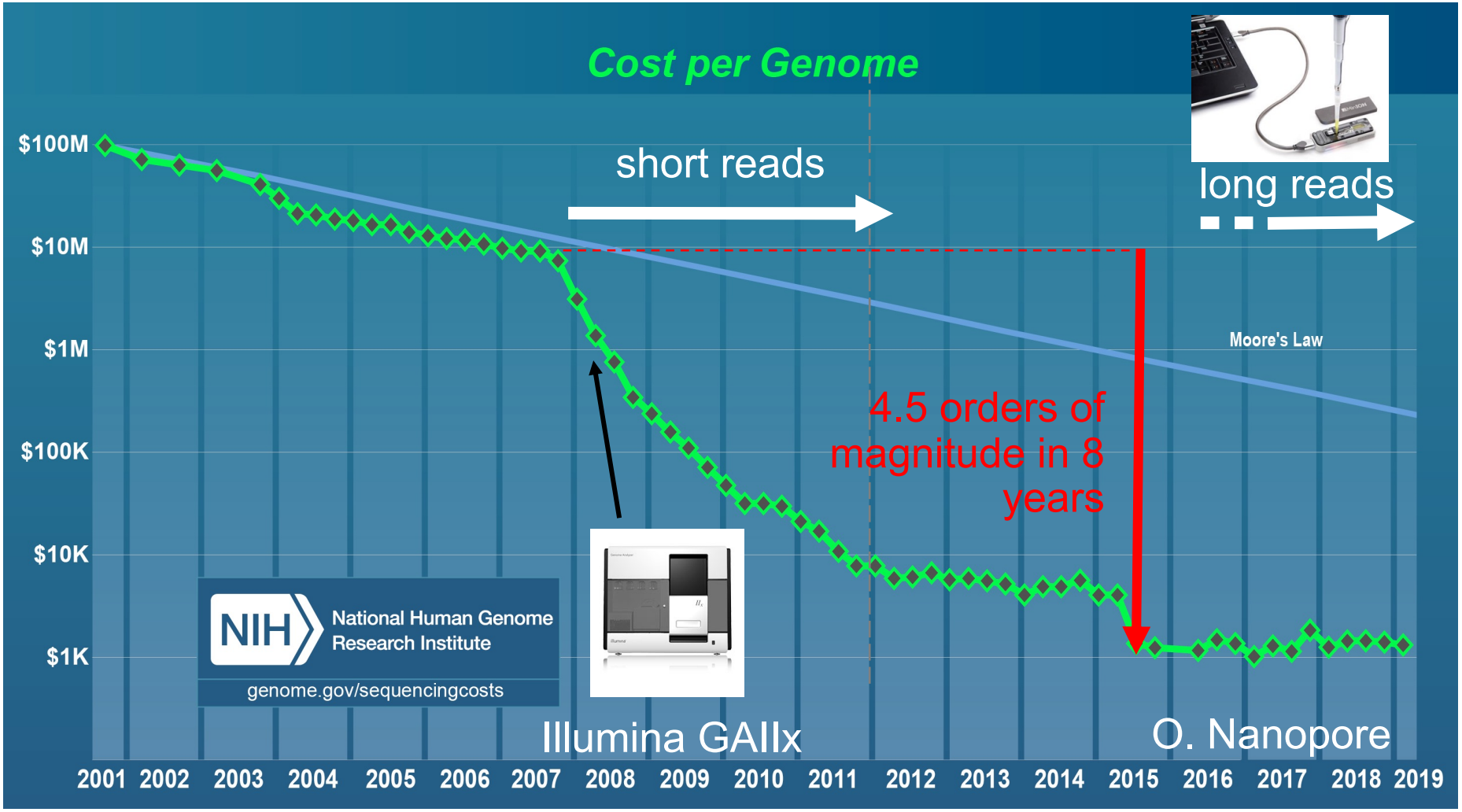


xtinction

Potential role of indolelactate and butyrate in multiple sclerosis revealed by integrated microbiome-metabolome analysis

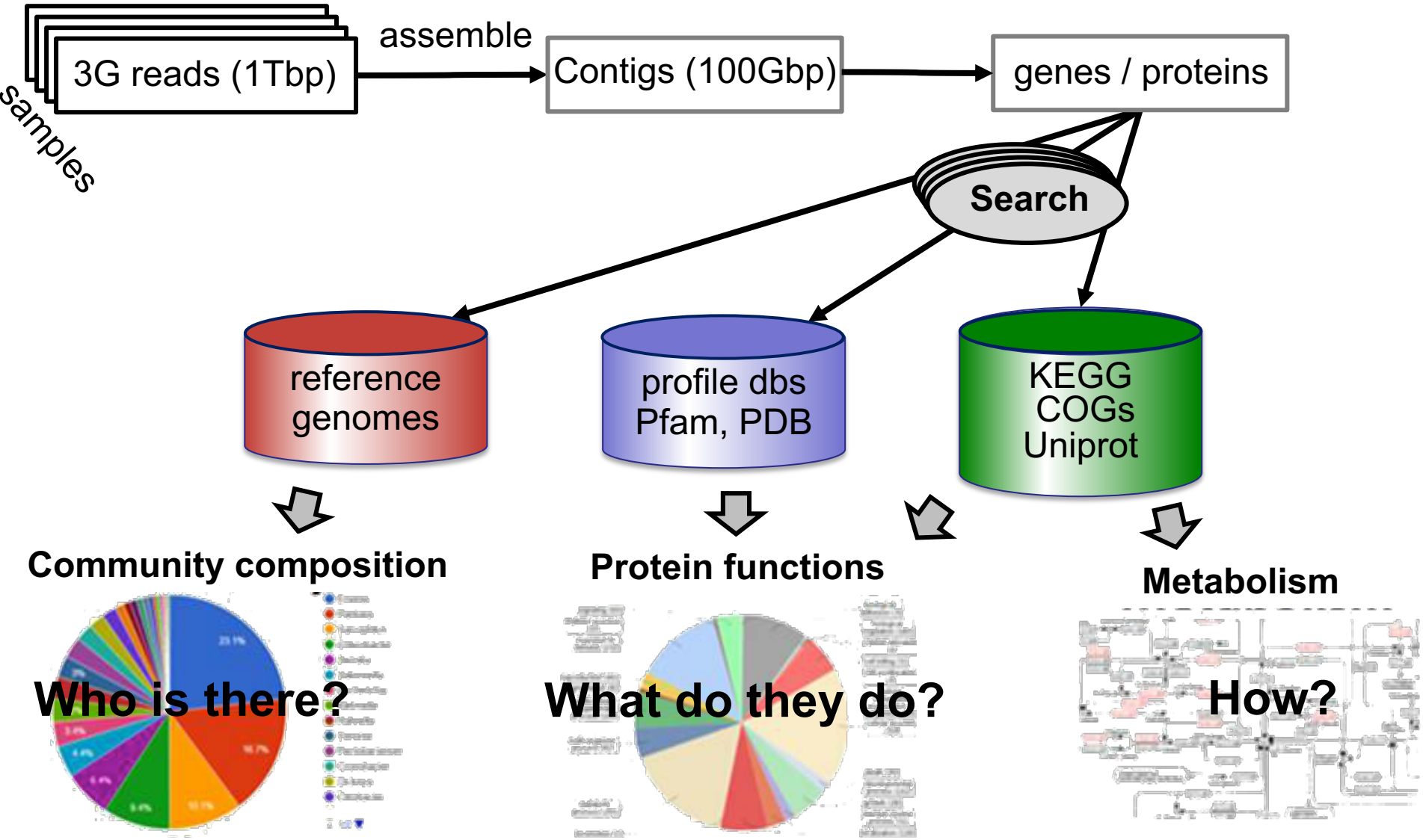
Cell Rep Med 2021, Apr 20

Metagenomics is driven by fast-decreasing sequencing costs

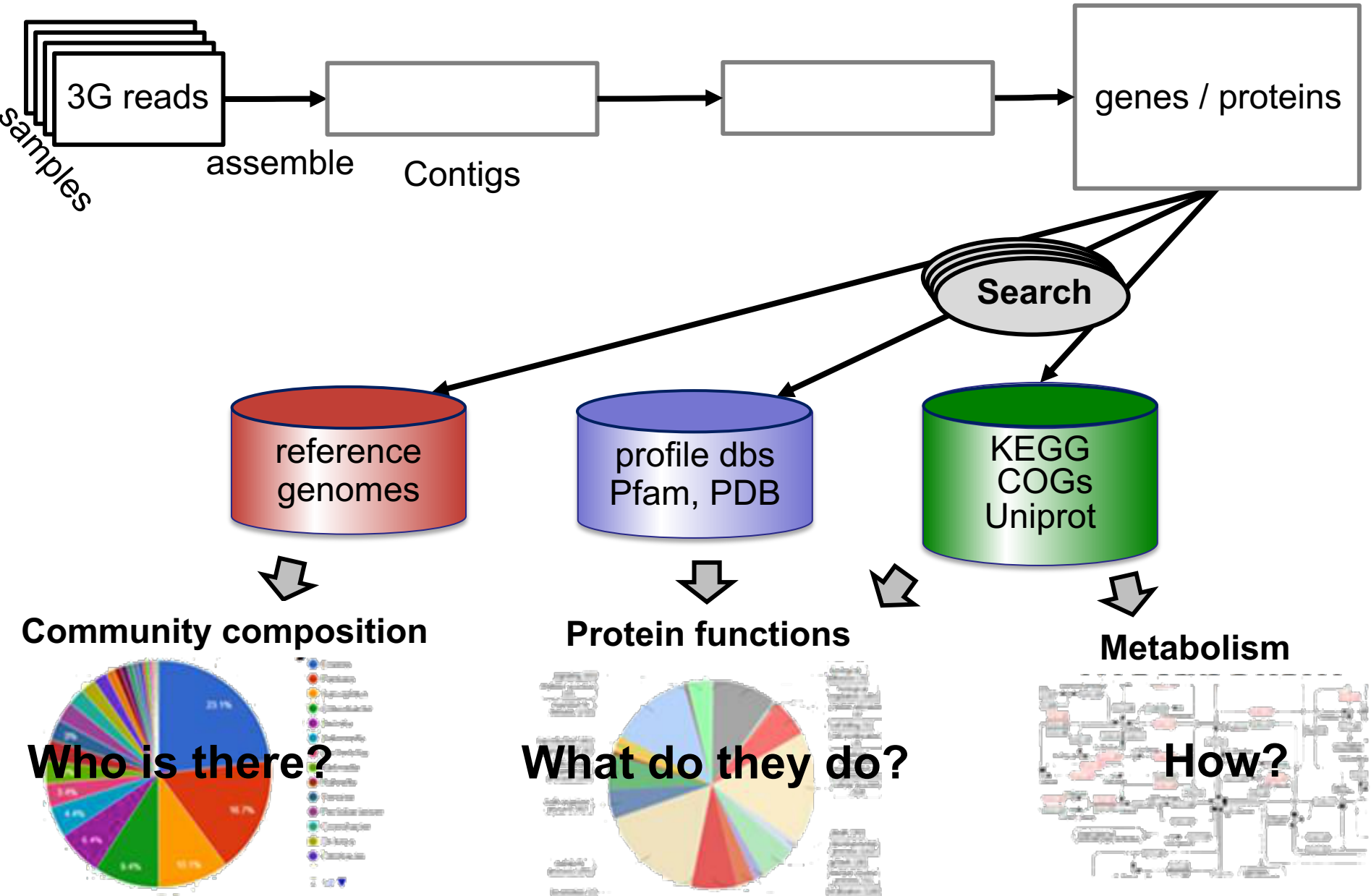


- **Costs for computing by far exceed sequencing costs**
- **Bottleneck: sequence searches**

Shotgun metagenomics data analysis



Shotgun metagenomics data analysis



MICROBIOLOGY

nature

Metagenomics

Philip Hugenholtz and Gene W. Tyson

Vol 455|25 September 2008

What other bottlenecks are there?

The gap between characterized and hypothetical proteins identified in metagenomes is widening at an alarming rate. Next to computational resources, uncharacterized gene products are likely to be the biggest bottleneck for the foreseeable future. This means that our under-

Often, 50%-90% of ORFs remain unannotated: no function, no taxon

MMseqs2

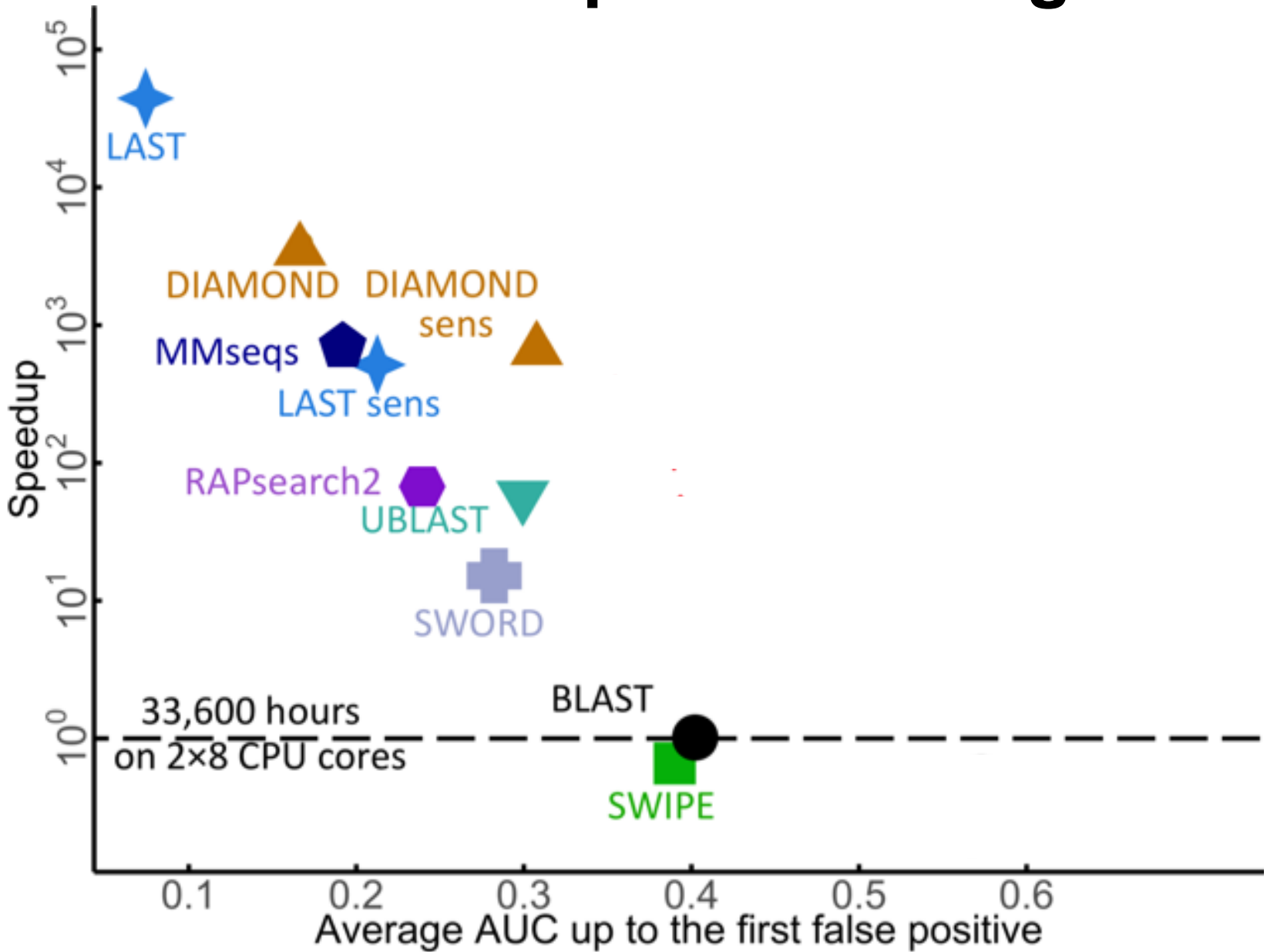
Ultrafast and sensitive sequence and profile searches

Martin Steinegger

with Milot Mirdita, Eli Levy Karin, Clovis Galiez, Ruoshi Zhang

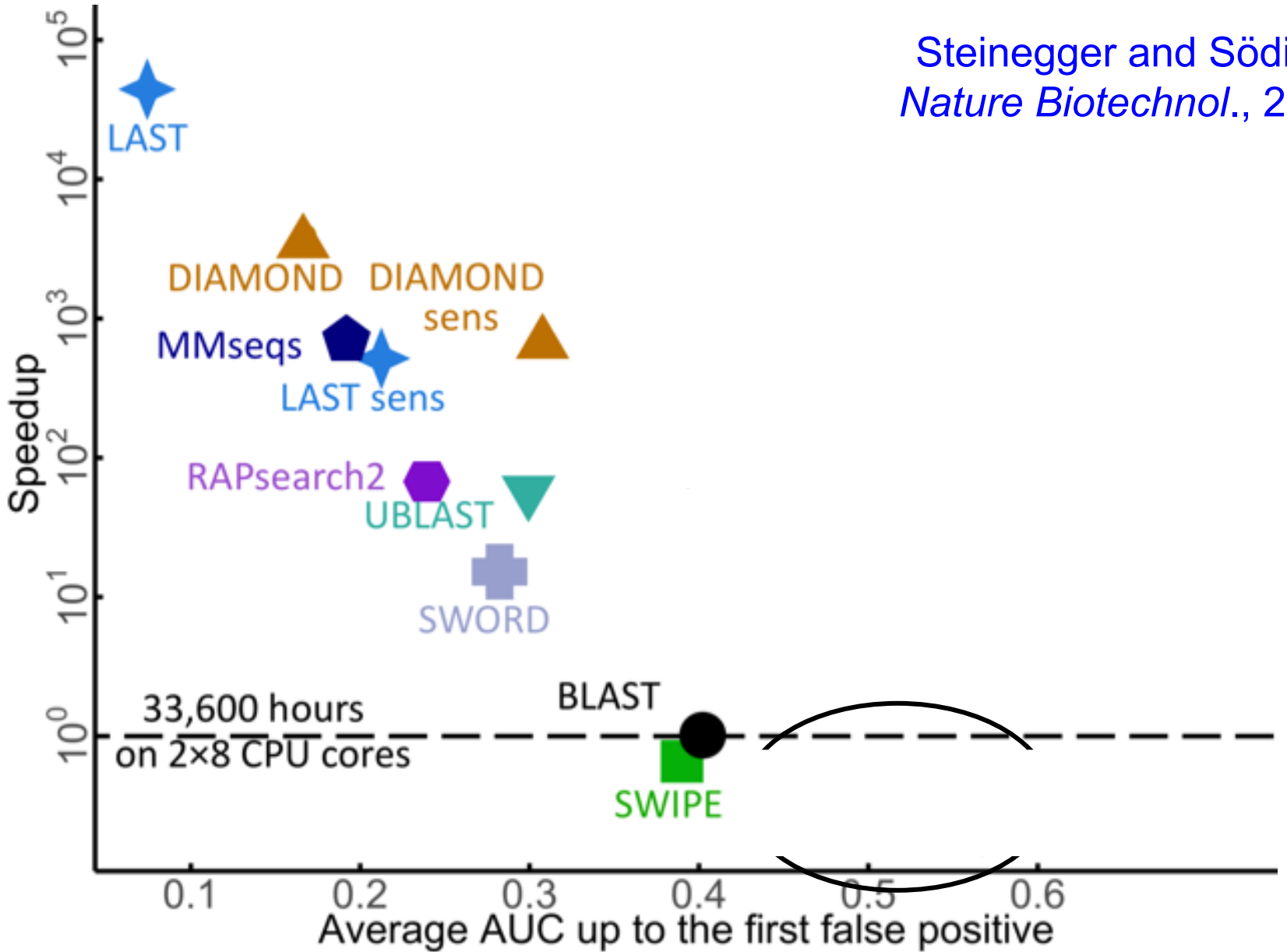
5 minutes ☺

Faster but less sensitive search tools have been developed for metagenomics



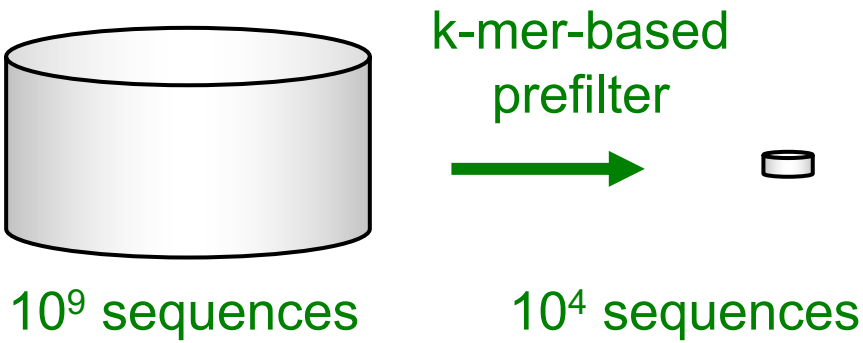
MMseqs profile searches 300 times faster and more sensitive than PSI-BLAST

Steinegger and Söding, *Nature Biotechnol*., 2017.



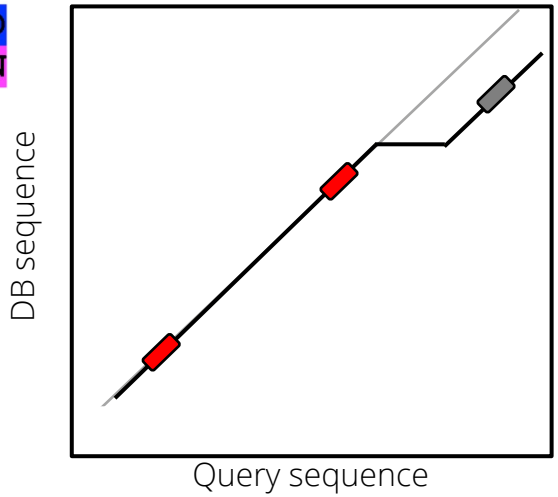
Fast and sensitive prefilter is most critical part for search performance

Reduces search space 10^5-fold while losing few true positives



- **Key ideas for prefilter in MMseqs**
  - Match *long* & *similar* *k*-mers 

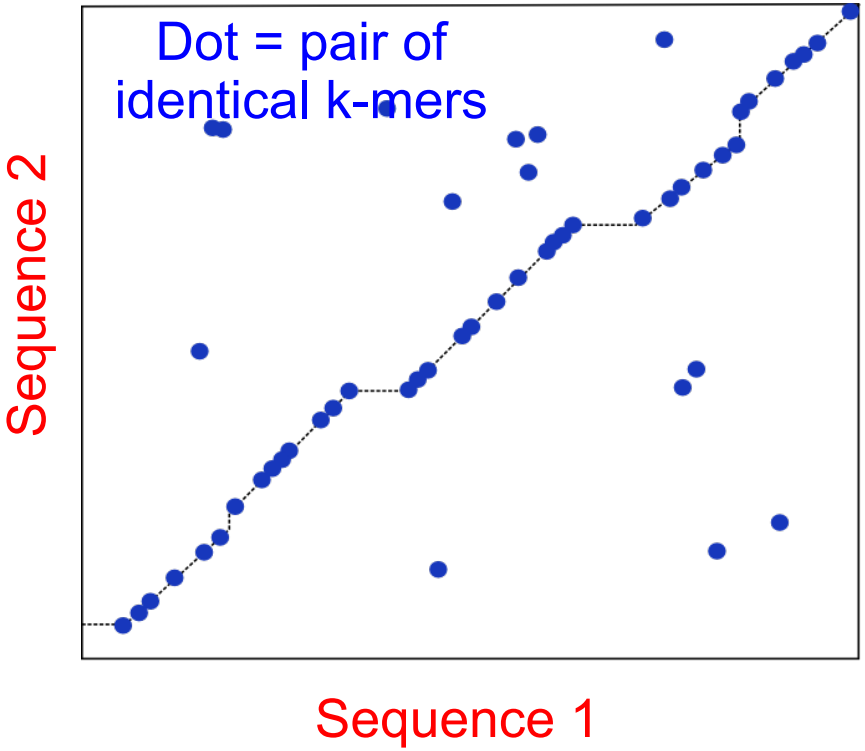
  - Two *k*-mer matches  without gap in-between
  - Sequence *profiles*!
  - No random memory access in innermost loop



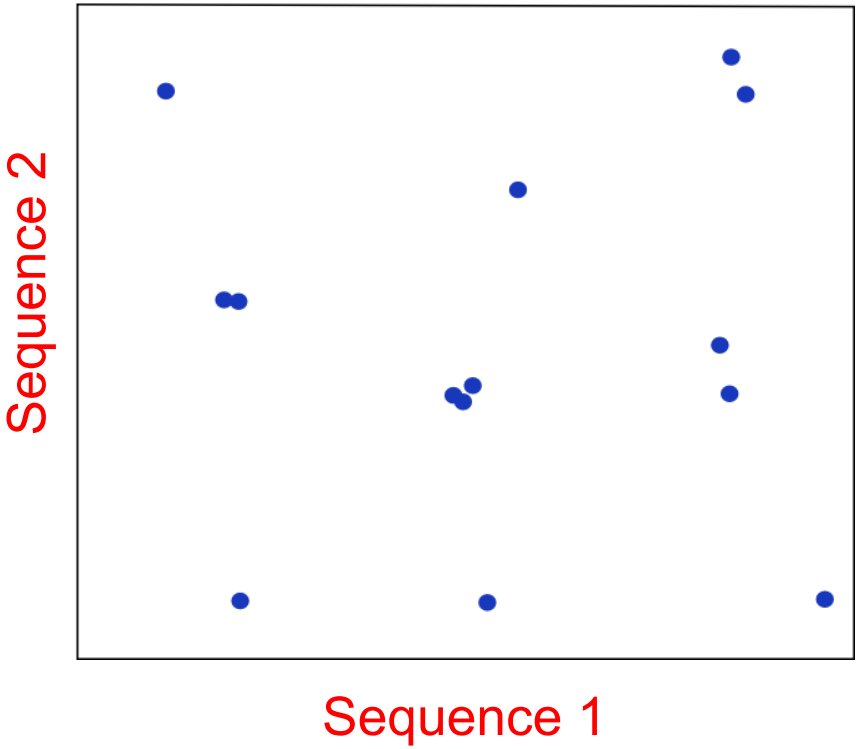
Conventional alignment-free comparison: count identical k-mers

Sequence 1 ... VRLS ... PLCW ... YAGD ...
Sequence 2 ... VRLS ... PLCW ... YAGD ...

Homologous proteins

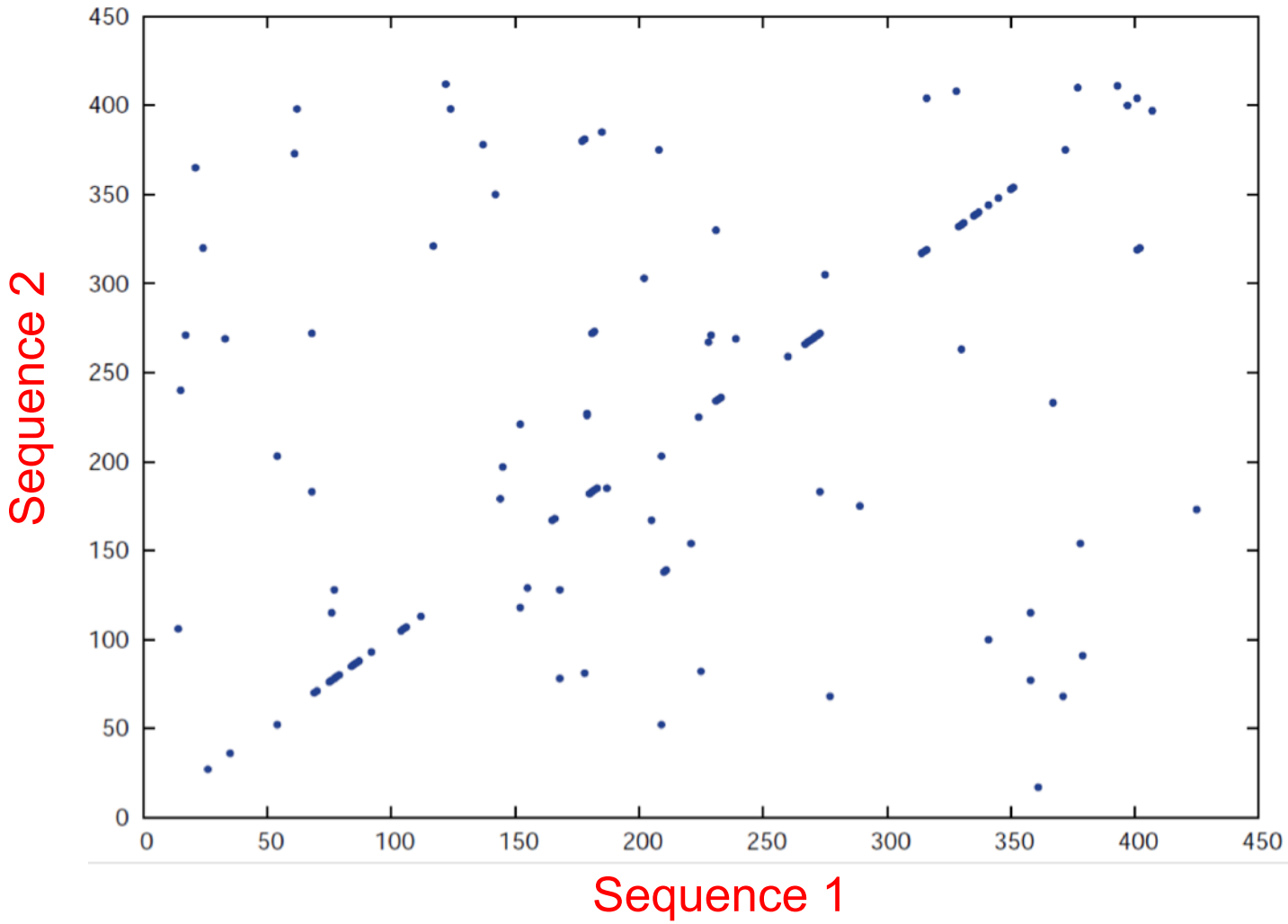


Unrelated proteins

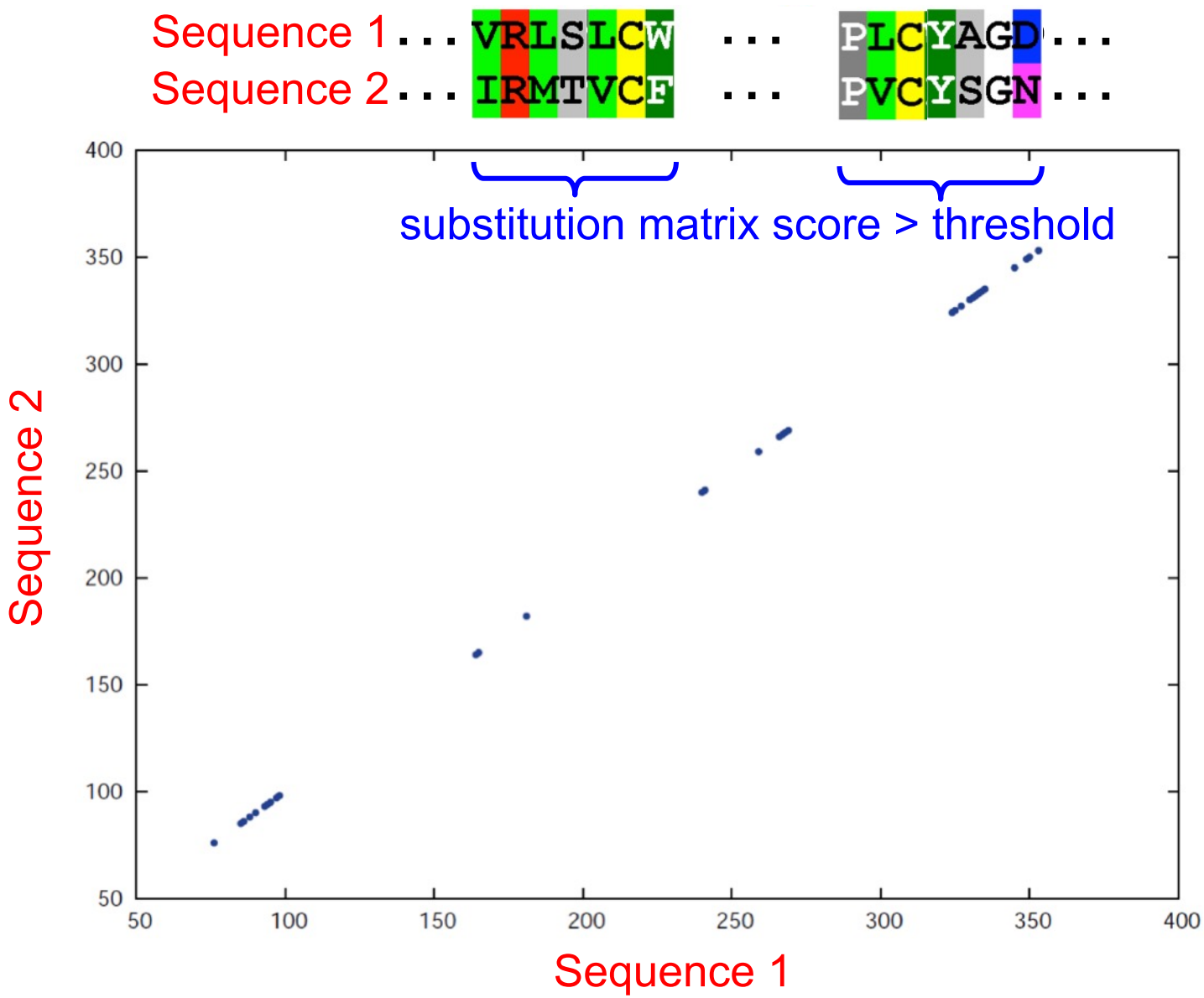


Most 3-mer matches occur by chance

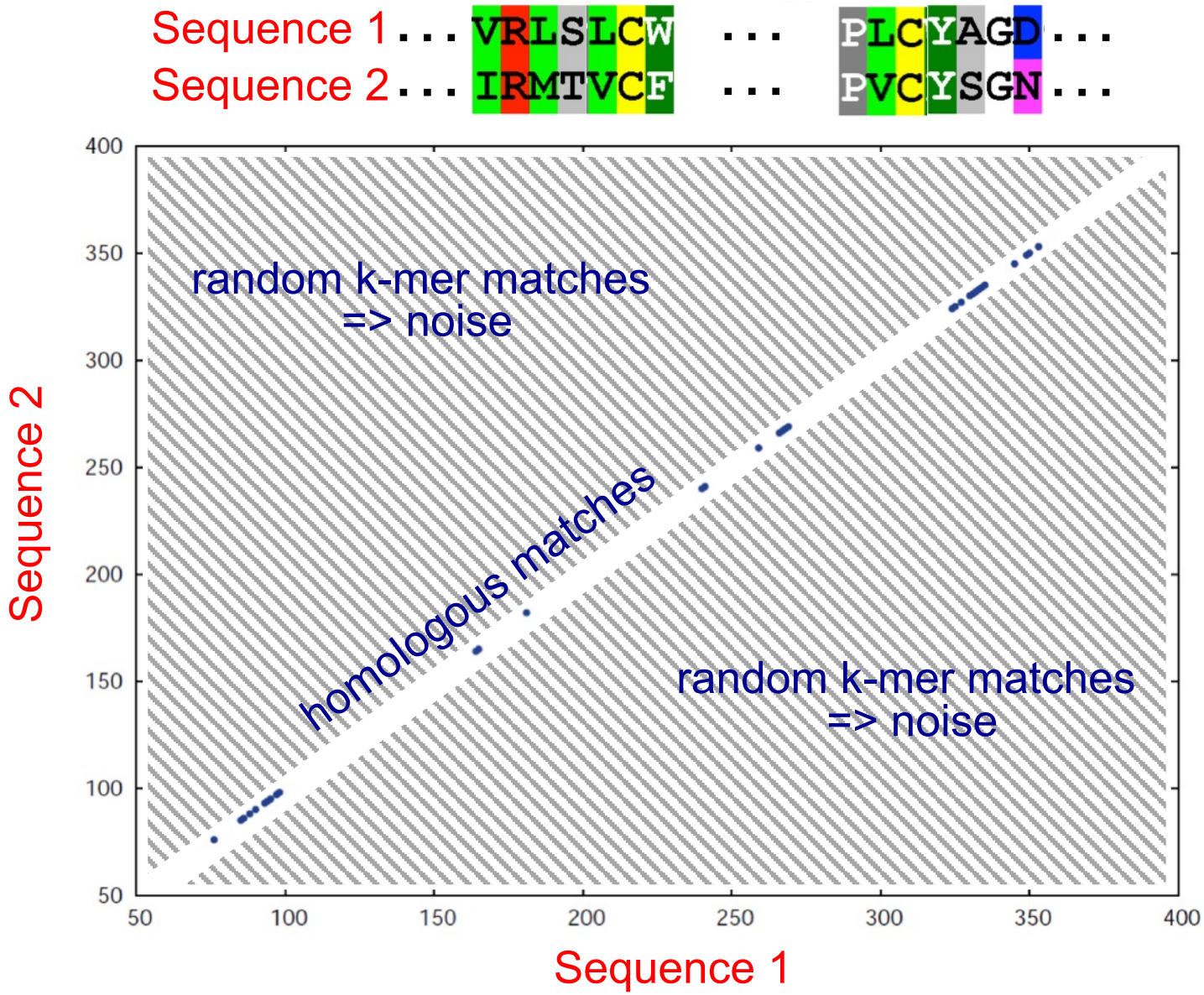
Sequence 1 ...RLS ... PLC ... YAG ...
Sequence 2 ...RLS ... PLC ... YAG ...



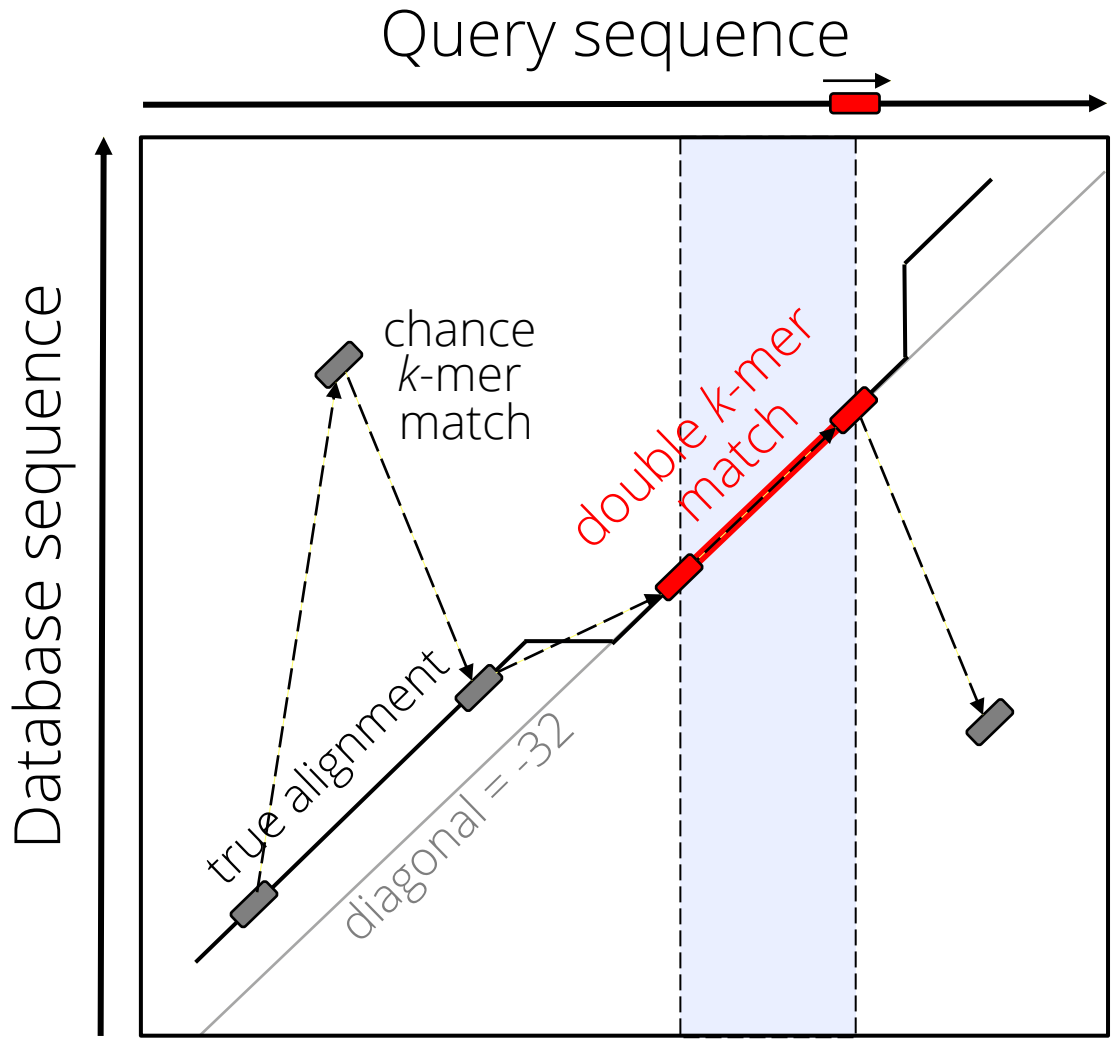
MMseqs: sum scores of similar 7-mers



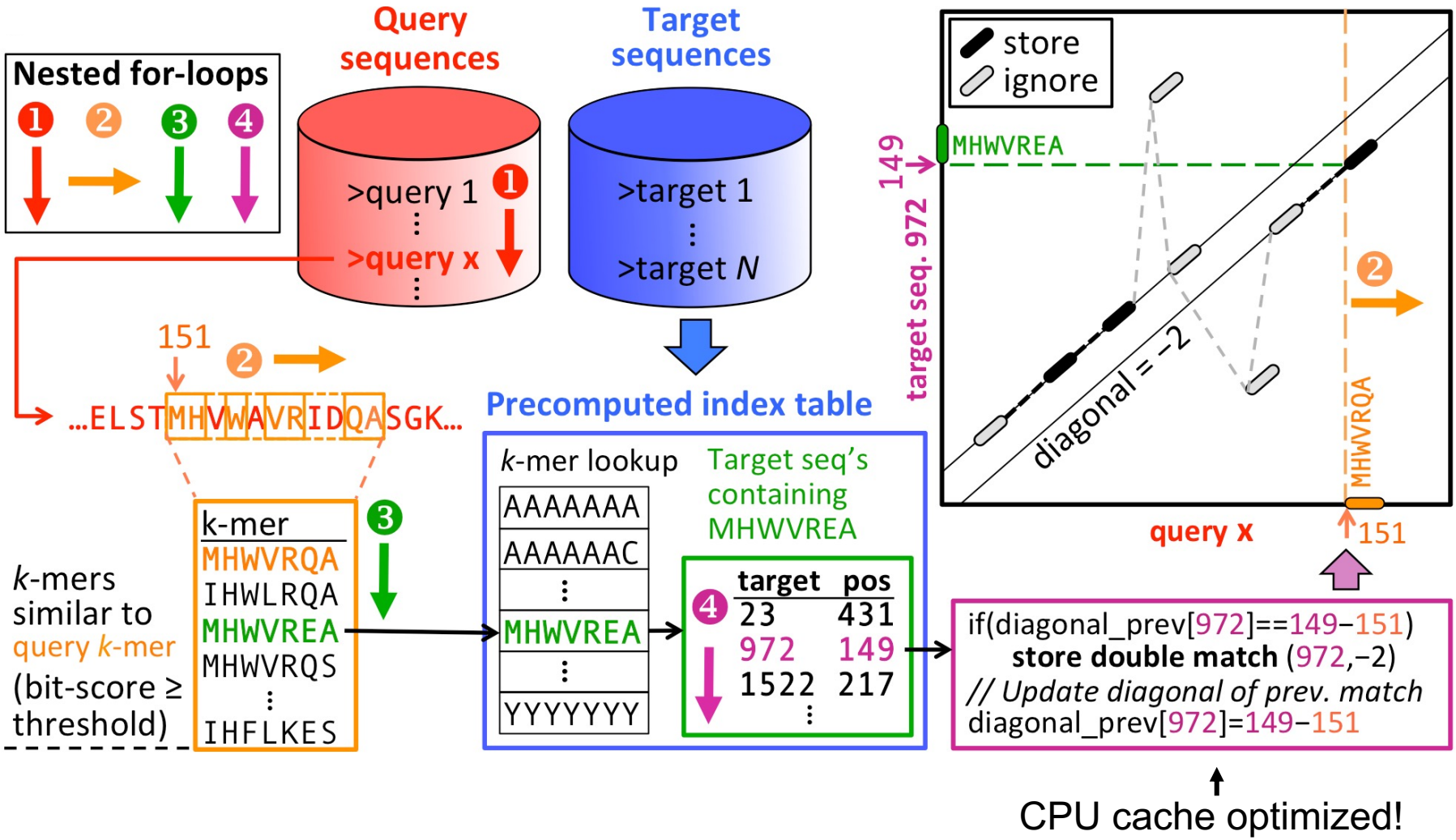
But: how to suppress the many random matches in hatched part of matrix?



Find db sequences with 2 consecutive *k*-mer matches on same diagonal



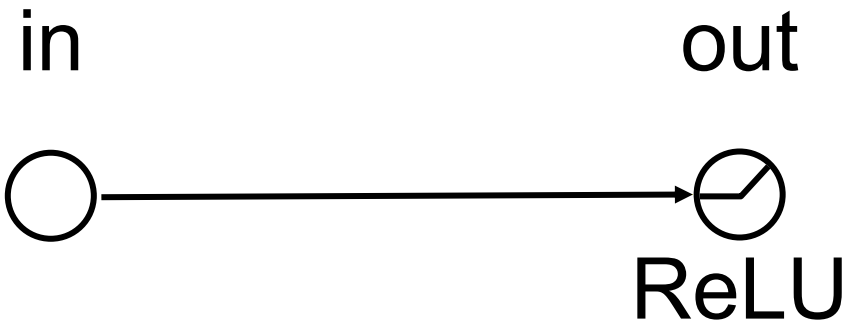
Find 2 consecutive *k*-mer matches on same diagonal



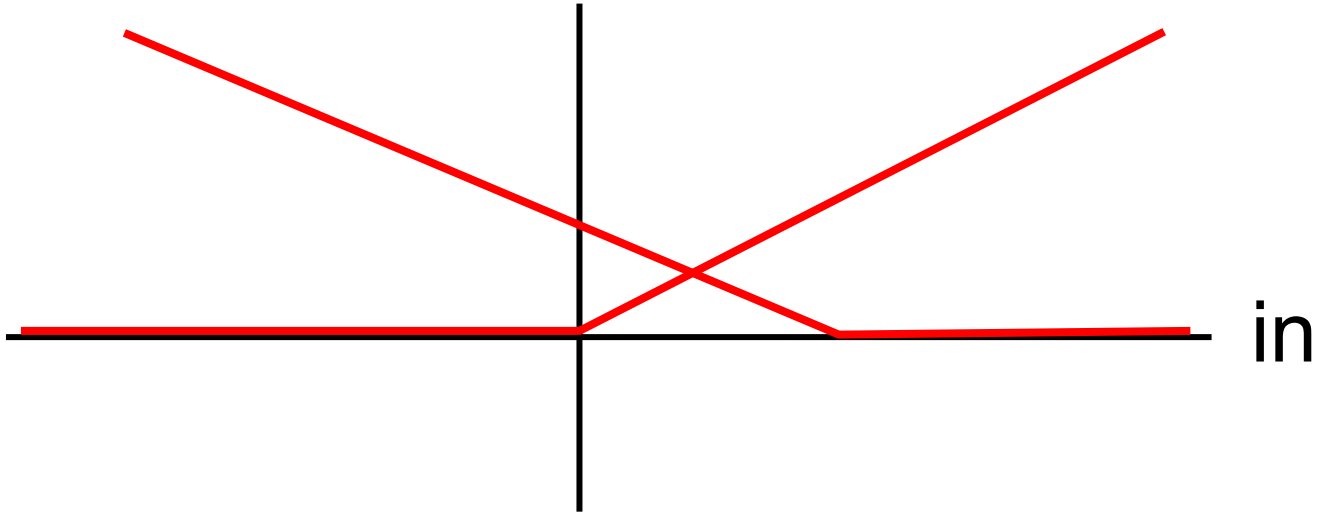
Executive summary

Whizz tour into deep learning

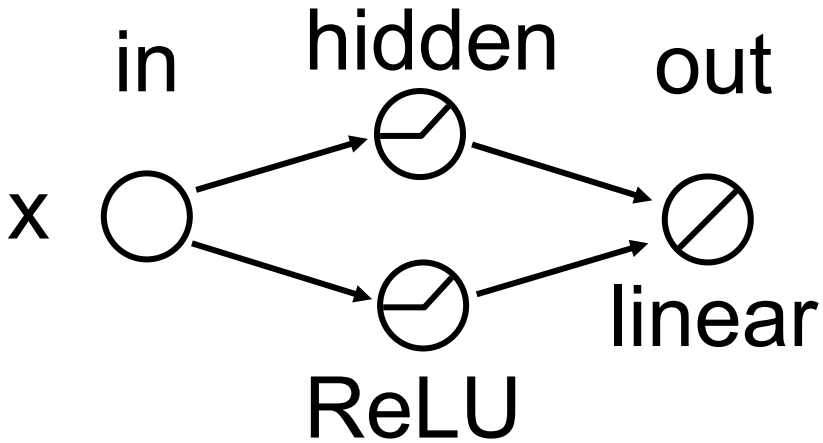
A rectifying linear unit (ReLU)
– the basic nonlinear unit of neural networks



out = max(0, a*in + b)

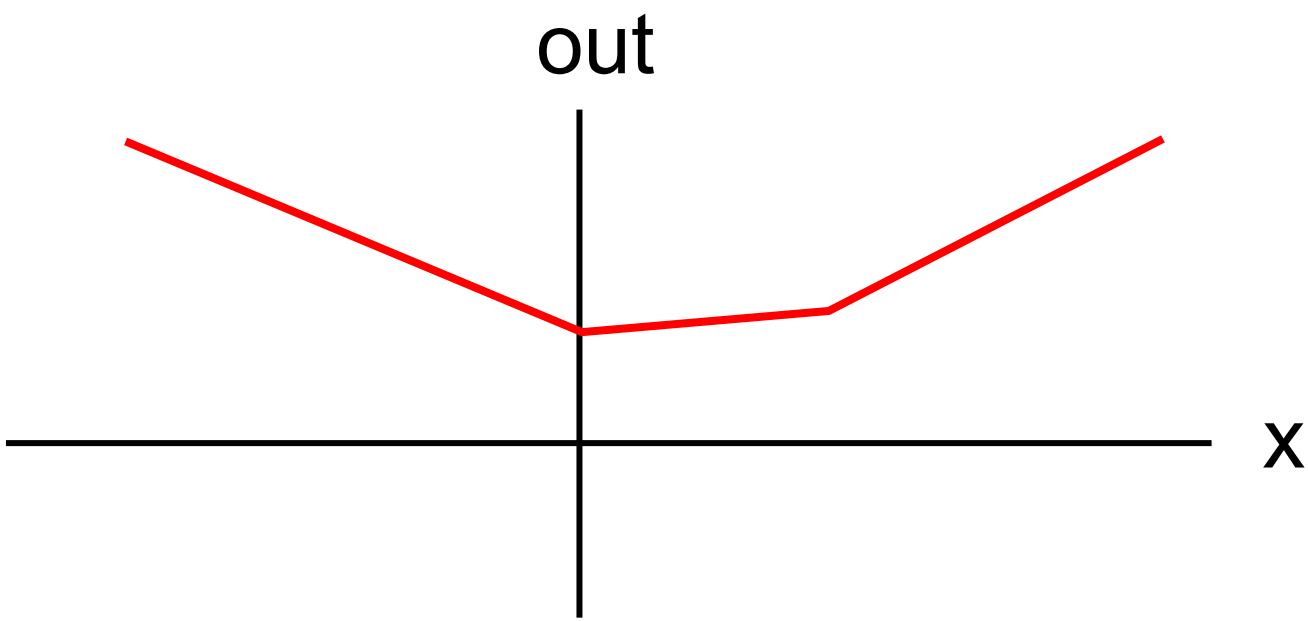


Two rectifying linear units combined linearly

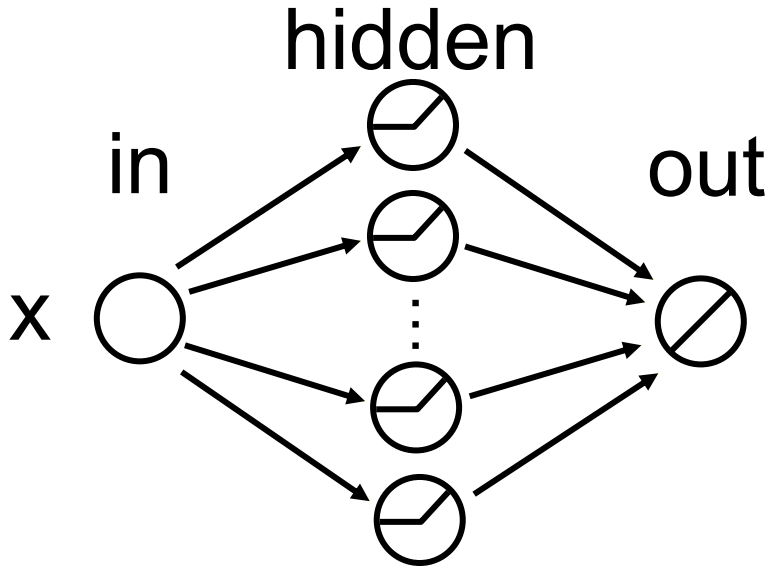


$$\mathrm{ReLU}_i = \max(0, w_i * x + b_i)$$

$$\mathrm{Out} = \sum_j w_j' \, \mathrm{ReLU}_j + b_j$$

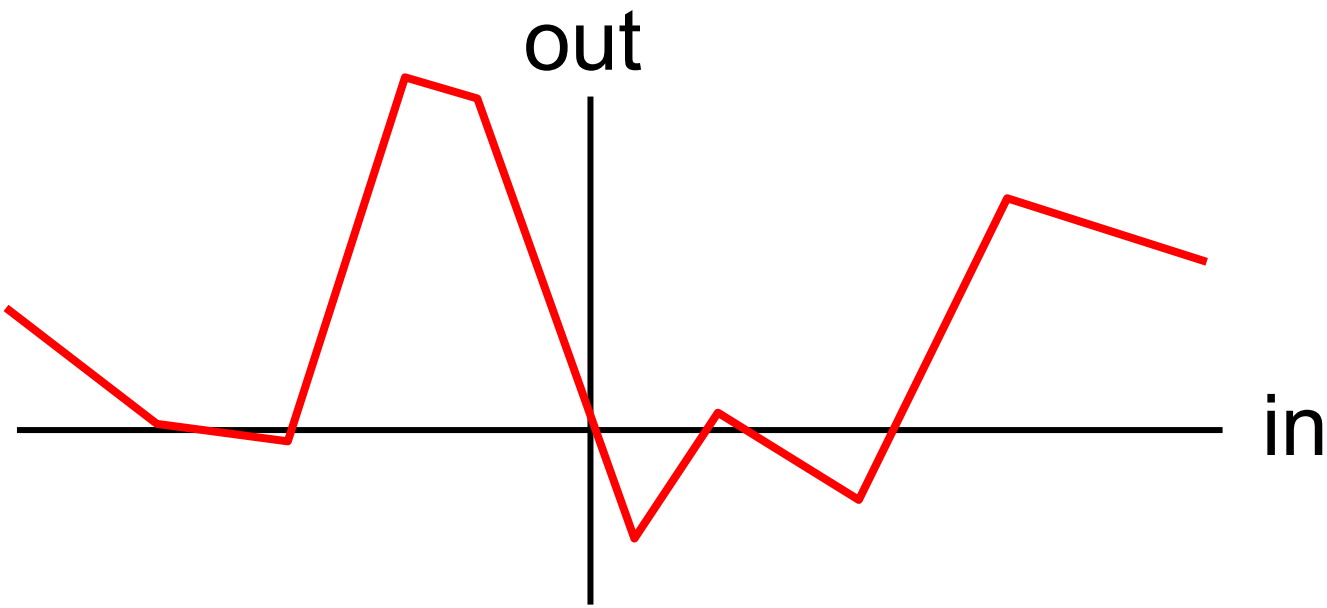


Ten rectifying linear units combined linearly

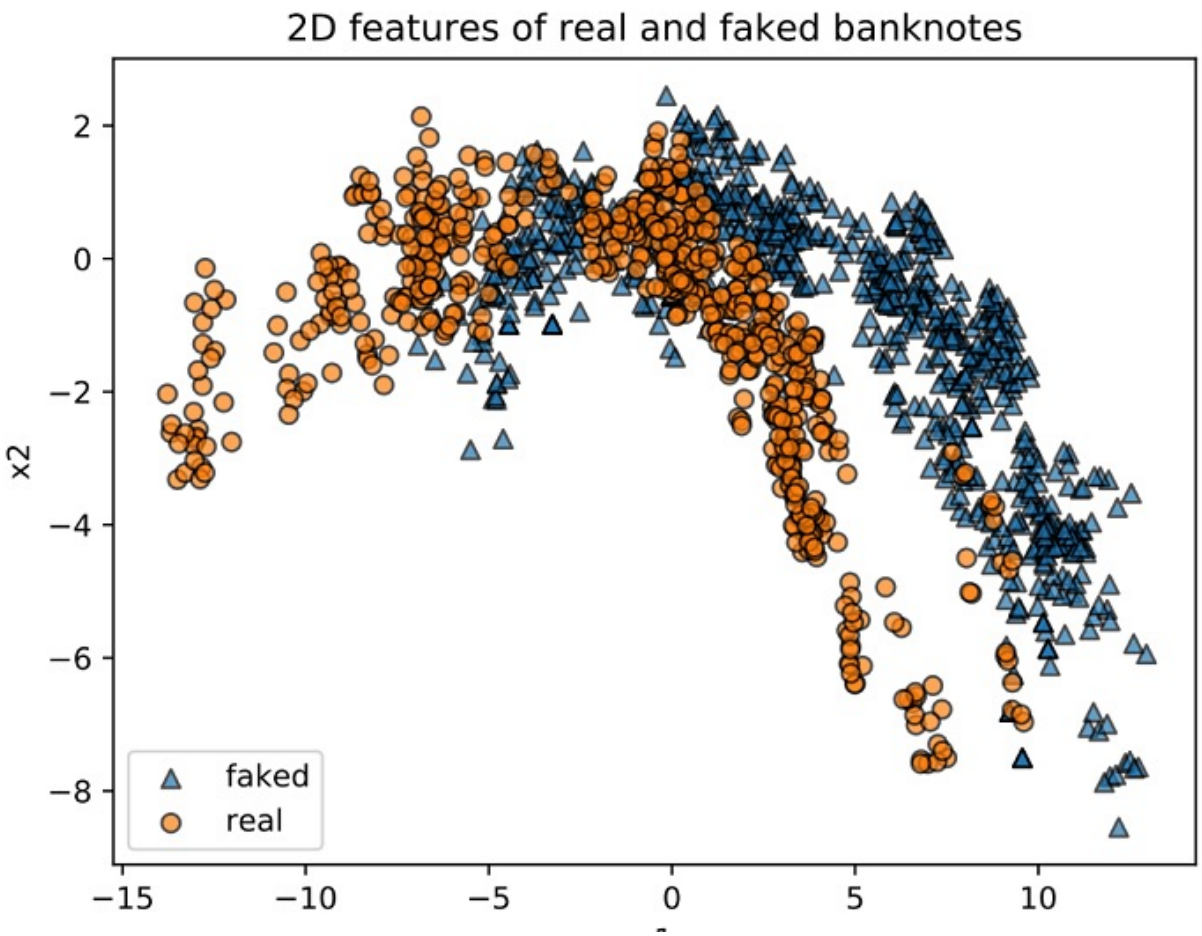


$$\mathrm{ReLU}_i = \max(0, w_i * x + b_i)$$

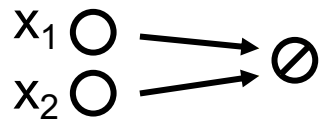
$$\mathrm{Out} = \sum_j w_j' \, \mathrm{ReLU}_j + b_j'$$



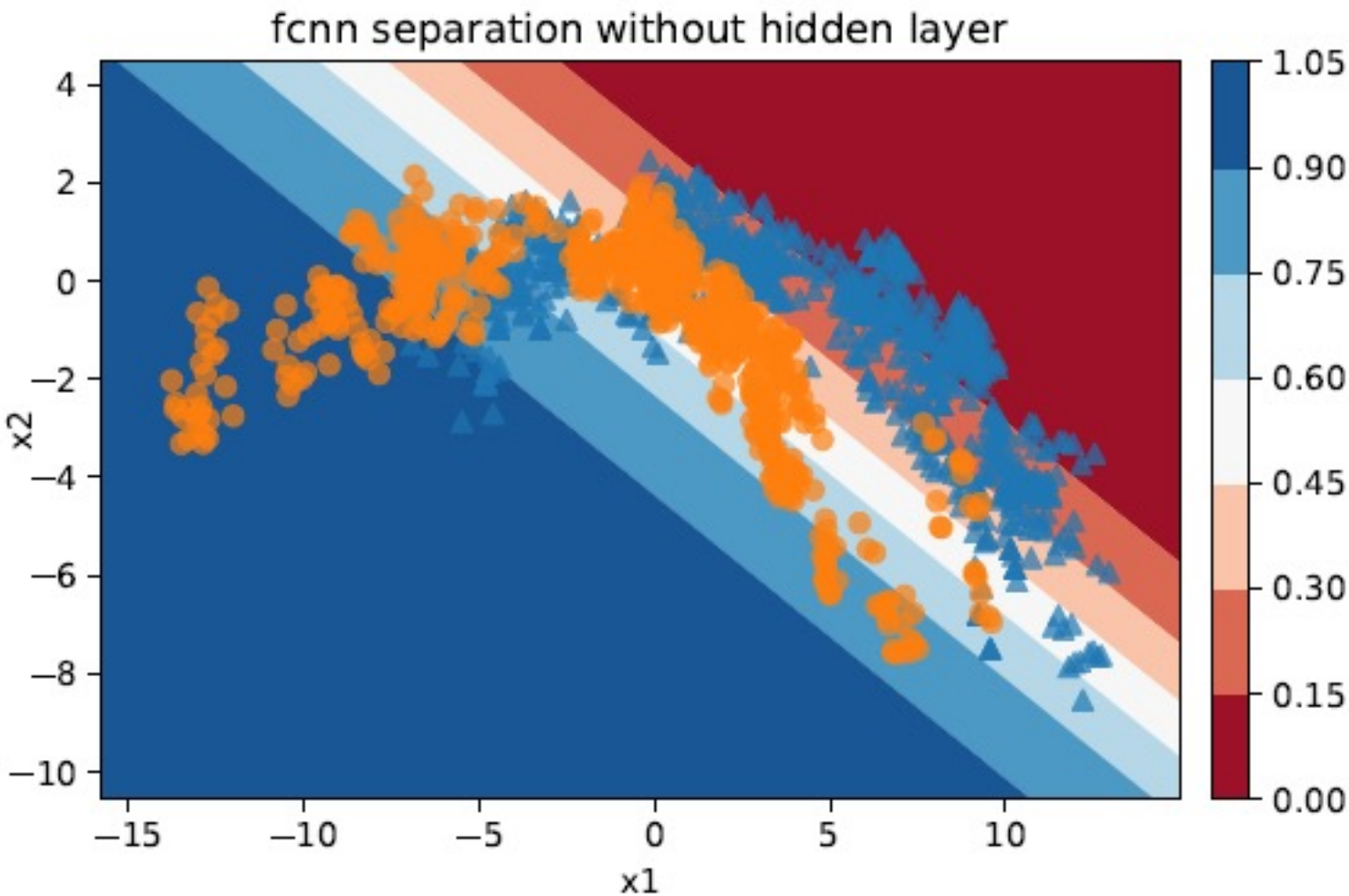
Dense neural networks can approximate any multivariate function arbitrarily well



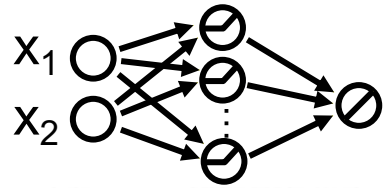
Dense neural networks can approximate any multivariate function arbitrarily well



$$\text{out} = w_1\, x_1 + w_2\, x_2 + b_j$$

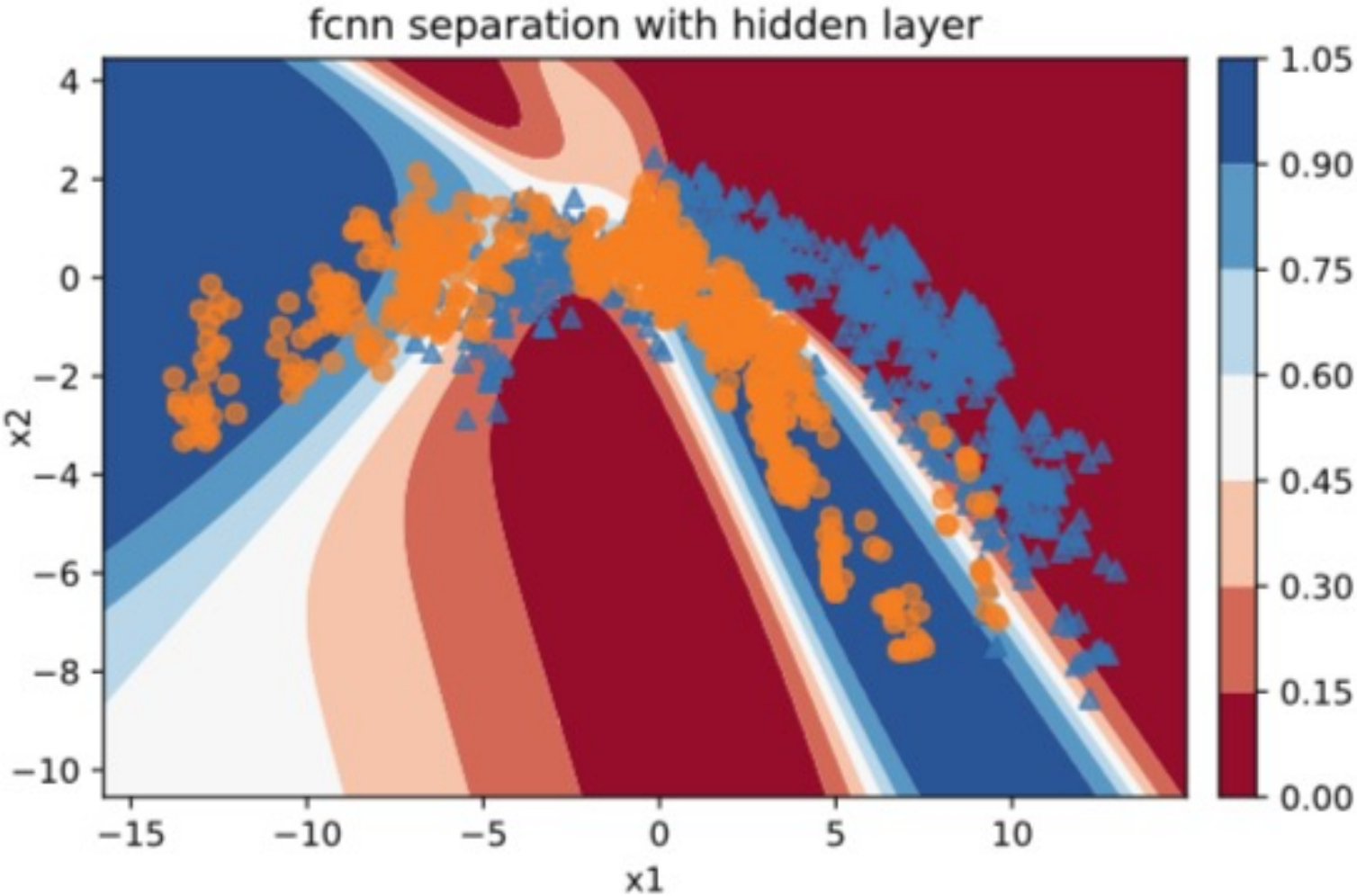


Dense neural networks can approximate any multivariate function arbitrarily well



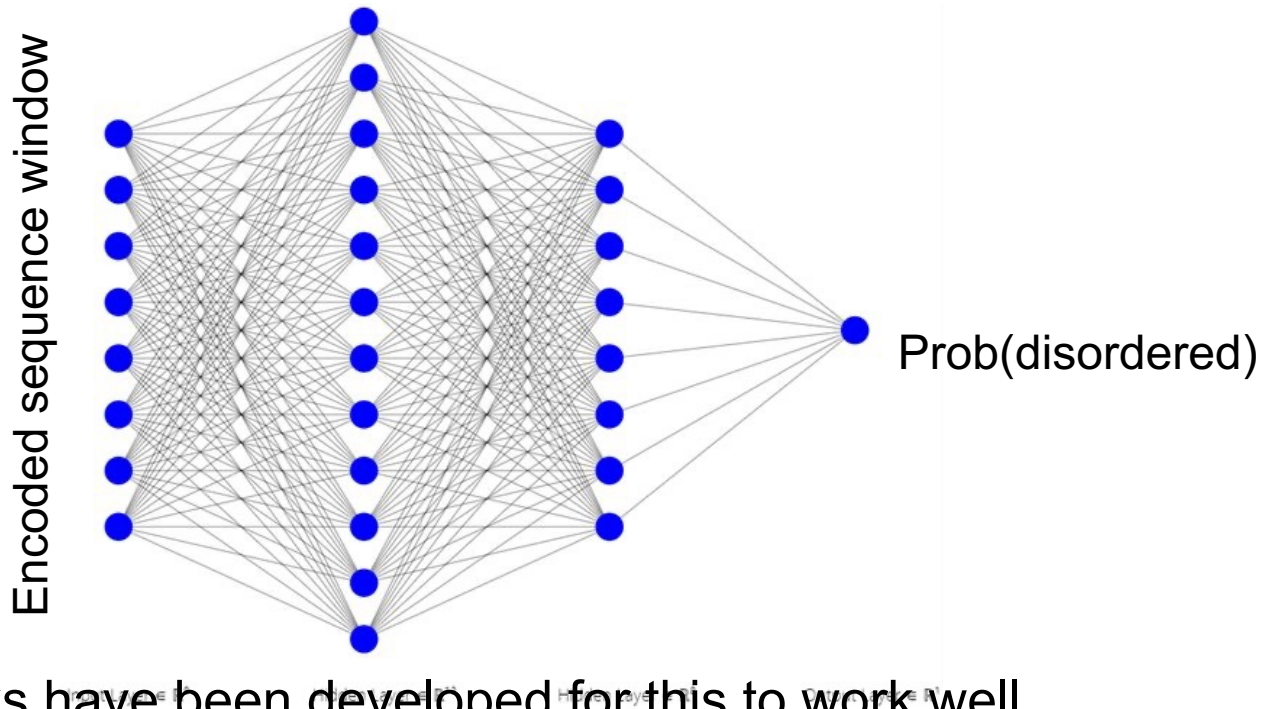
$ReLU_i = \max(0, w_i * in + b_i)$

$Out = \sum_j w_j' \, ReLU_j + b_j'$



Neural networks can be trained with training data to learn any multivariate function

(somewhat well)



Many technical tricks have been developed for this to work well.

Most important:

- **Back-propagation** = efficient way to compute partial derivatives of outputs with respect to each of the neural network weights (given the training data)
- Stochastic gradient descent
- Automatic differentiation

How disorder prediction works

```
Query    ...DPLLIAETLRQAAMLVFHAGYGVPVGYHFLMATLDYTCHLDHLGVS...
```

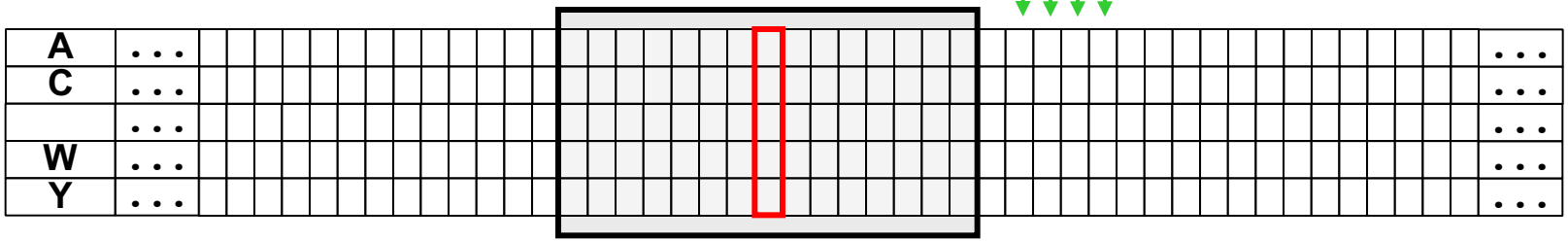
PSI-BLAST / HHblits search

homologs

```
Query    ...DPLLIAETLRQAAMLVFHAGYGVPVGYHFLMATLDYTCHLDHLGVS...
AfsA     ...DPMQIAETMRQVGLHLAHAEFDVPLGHHFIMWDMS-VSRVEHLGVT...
JadW1    ...DPMLVAETIPETSMLVAHAELGVPLDEQFVMWDLS-SADSEALTVD...
BarX     ...DPLLASETIRQVGTLLSHAEFGVSFGDQFLMWDLH--VRPEQAGVA...
FarX     ...DPLMCAETIRQIAYLLGHAEFAVPFGHQFVLSSLR--ANVEHLGVT...
```

calculate profile

amino acid frequencies



use **neural network** to predict disorder from windows

The neural network has learnt pattern ↔ disorder/order relations

```
predict ...OOOOOOOODDDDDDDDDDDDDDDDDDDDDDDDDDDDDOOOOOOOOOOOOOOO...
```

Best methods reach per-residue accuracy ~ 80%, but what is disorder really?

How secondary structure prediction works

```
Query    ...DPLLIAETLRQAAMLVFHAGYGVPVGYHFLMATLDYTCHLDHLGVS...
```

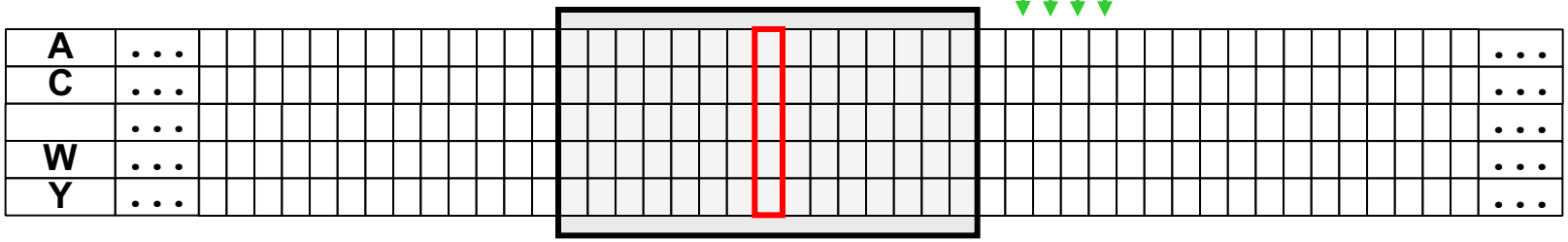
PSI-BLAST / HHblits search

homologs

```
Query    ...DPLLIAETLRQAAMLVFHAGYGVPVGYHFLMATLDYTCHLDHLGVS...
AfsA     ...DPMQIAETMRQVGLHLAHAEFDVPLGHHFIMWDMS-VSRVEHLGVT...
JadW1    ...DPMLVAETIPETSMLVAHAELGVPLDEQFVMWDLS-SADSEALTVD...
BarX     ...DPLLASETIRQVGTLLSHAEFGVSFGDQFLMWDLH--VRPEQAGVA...
FarX     ...DPLMCAETIRQIAYLLGHAEFAVPFGHQFVLSSLR--ANVEHLGVT...
```

calculate profile

amino acid frequencies



use **neural network** to predict SS from windows

The neural network has learnt pattern ↔ SS state relations

```
ss_pred ...CCCHHHHHHHHHHHHHHHHHHHCCCCCCCCEEEEEEEEEEEECHHHCC...
```

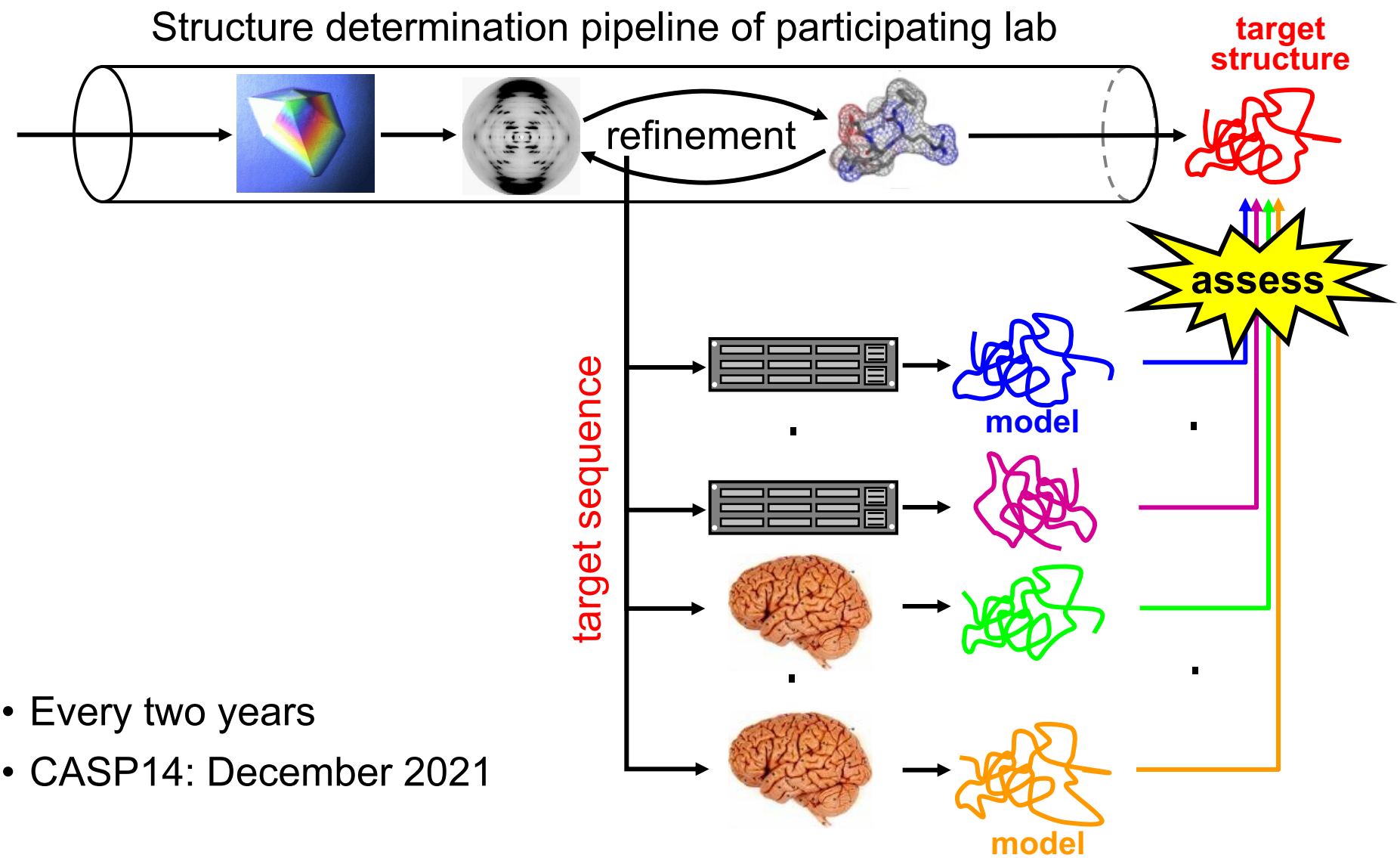
Best methods reached per-residue accuracy up to ~ 85%

5 minutes ☺

Protein structure prediction with AlphaFold

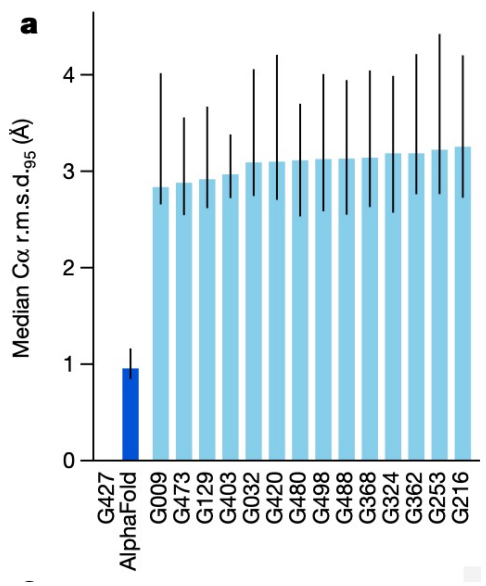
Critical Assessment of Structure Prediction

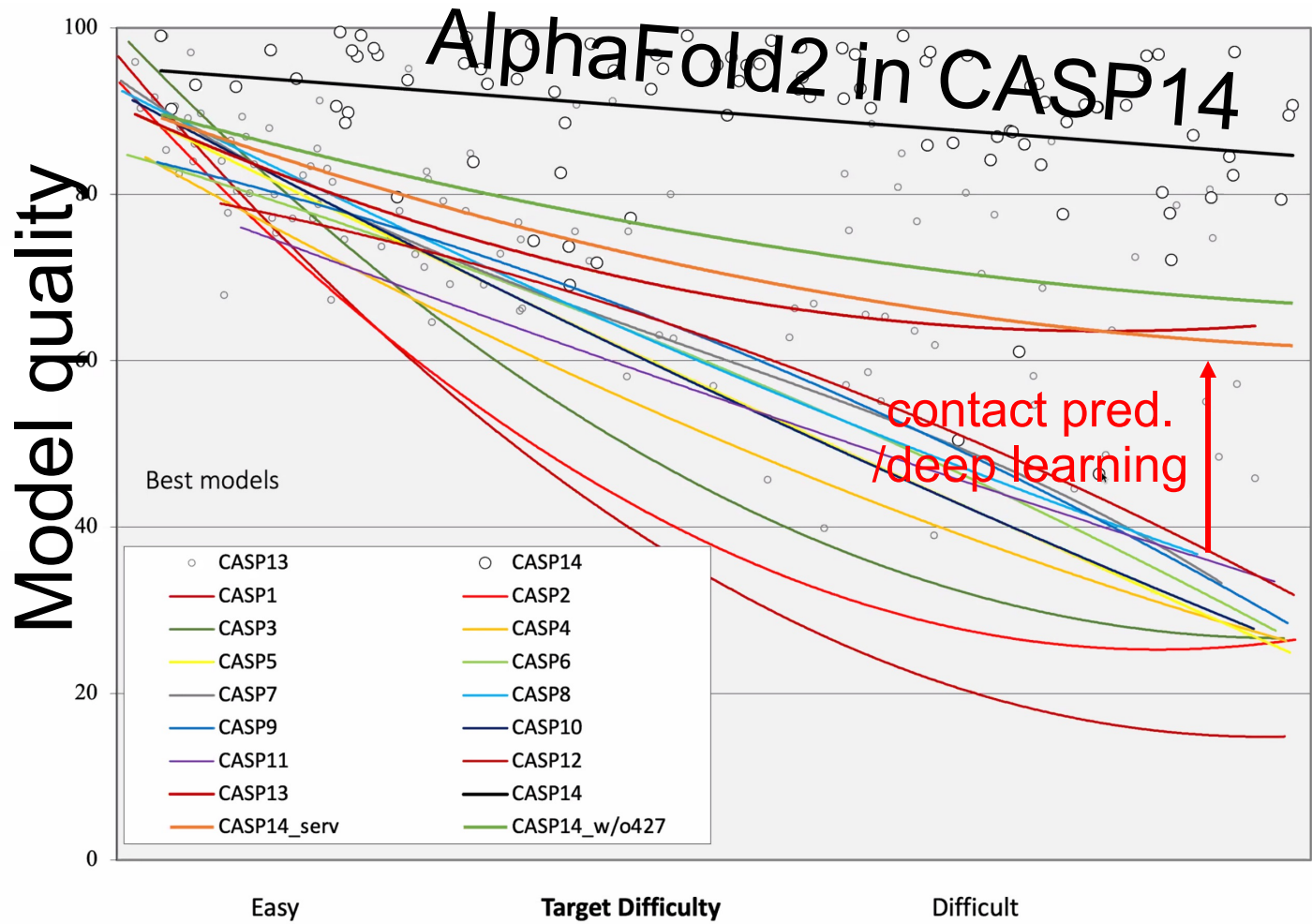
a community-wide blind structure prediction benchmark



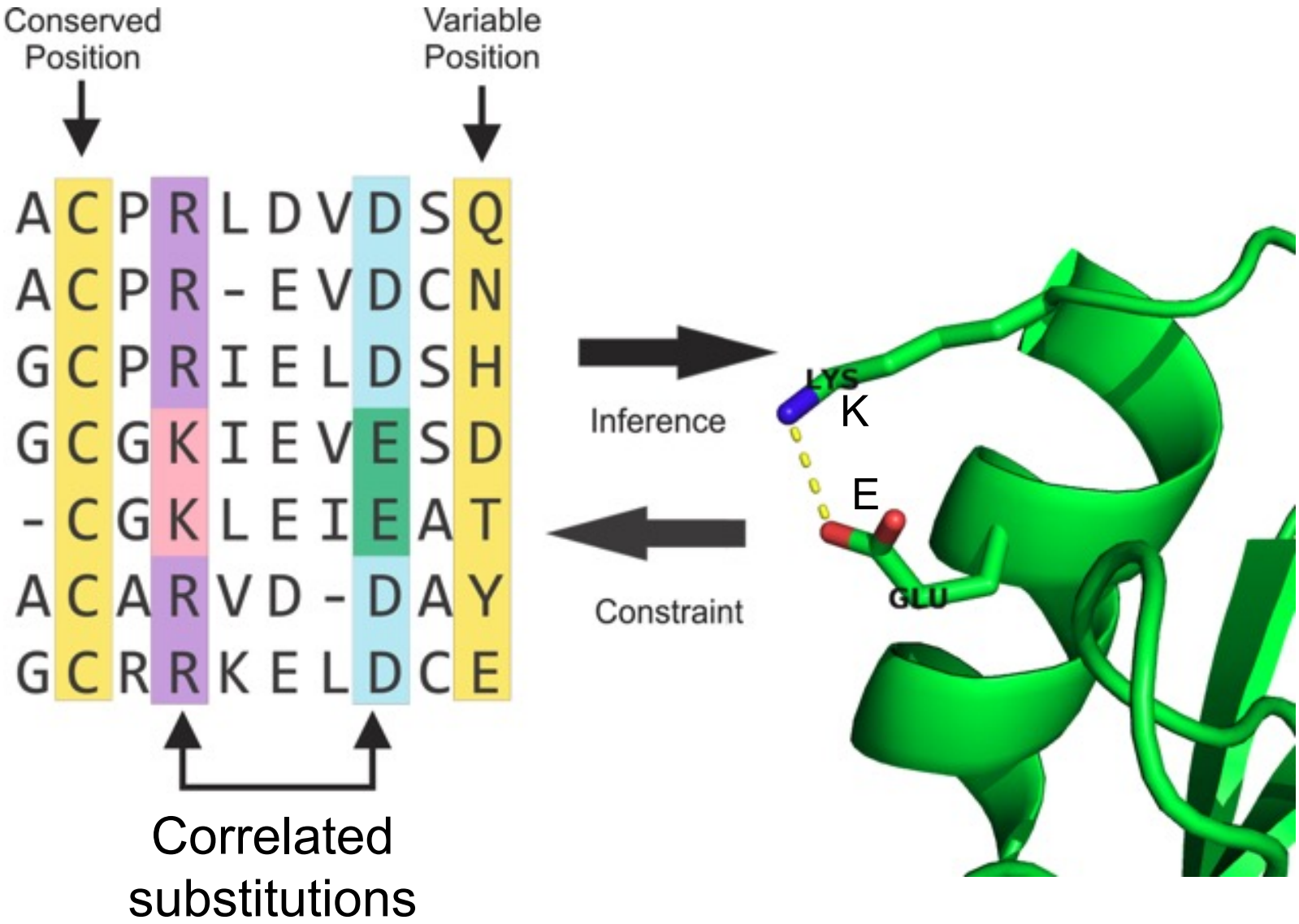
- Every two years
- CASP14: December 2021

Big leap in recent protein structure prediction benchmark CASP14 (Dec 2020)

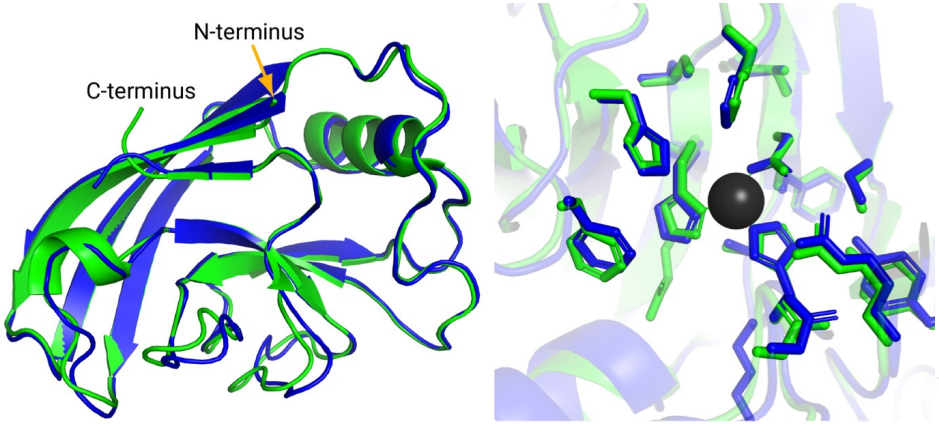




Correlated substitutions in multiple sequence alignments predict residue-residue contacts



AlphaFold is transformative for protein bioinfo., structural biology & biotechnology



Article

Highly accurate protein structure prediction with AlphaFold

https://doi.org/10.1038/s41586-021-03819-2
Received: 11 May 2021
Accepted: 12 July 2021
Published online: 15 July 2021
Open access
Check for updates

John Jumper[1,4✉], Richard Evans[1,4], Alexander Pritzel[1,4], Tim Green[1,4], Michael Figurnov[1,4], Olaf Ronneberger[1,4], Kathryn Tunyasuvunakool[1,4], Russ Bates[1,4], Augustin Žídek[1,4], Anna Potapenko[1,4], Alex Bridgland[1,4], Clemens Meyer[1,4], Simon A. A. Kohl[1,4], Andrew J. Ballard[1,4], Andrew Cowie[1,4], Bernardino Romera-Paredes[1,4], Stanislav Nikolov[1,4], Rishub Jain[1,4], Jonas Adler[1], Trevor Back[1], Stig Petersen[1], David Reiman[1], Ellen Clancy[1], Michal Zielinski[1], Martin Steinegger[2,3], Michalina Pacholska[1], Tamas Berghammer[1], Sebastian Bodenstein[1], David Silver[1], Oriol Vinyals[1], Andrew W. Senior[1], Koray Kavukcuoglu[1], Pushmeet Kohli[1] & Demis Hassabis[1,4✉]

DEEPMIND'S AI PREDICTS STRUCTURES FOR A VAST TROVE OF PROTEINS

AlphaFold neural network produced 'transformative' database of more than 350,000 structures.

Article

Highly accurate protein structure prediction for the human proteome

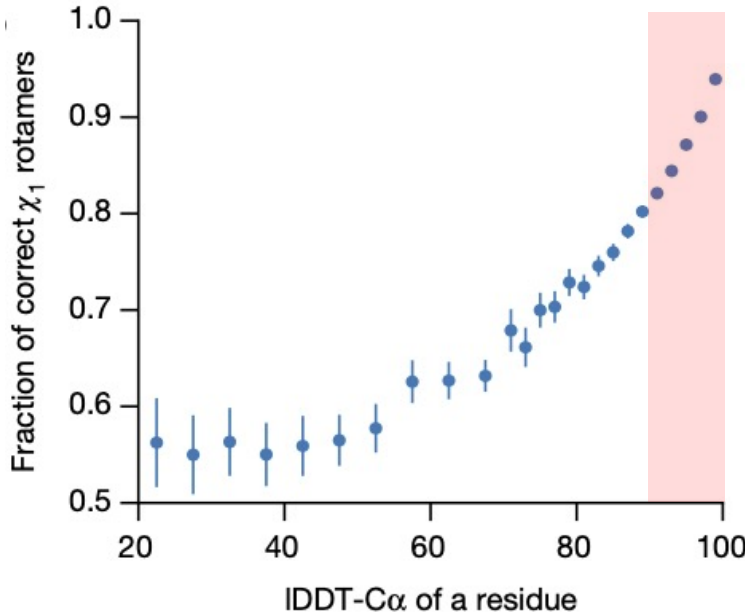
https://doi.org/10.1038/s41586-021-03828-1
Received: 11 May 2021
Accepted: 16 July 2021
Published online: 22 July 2021
Open access

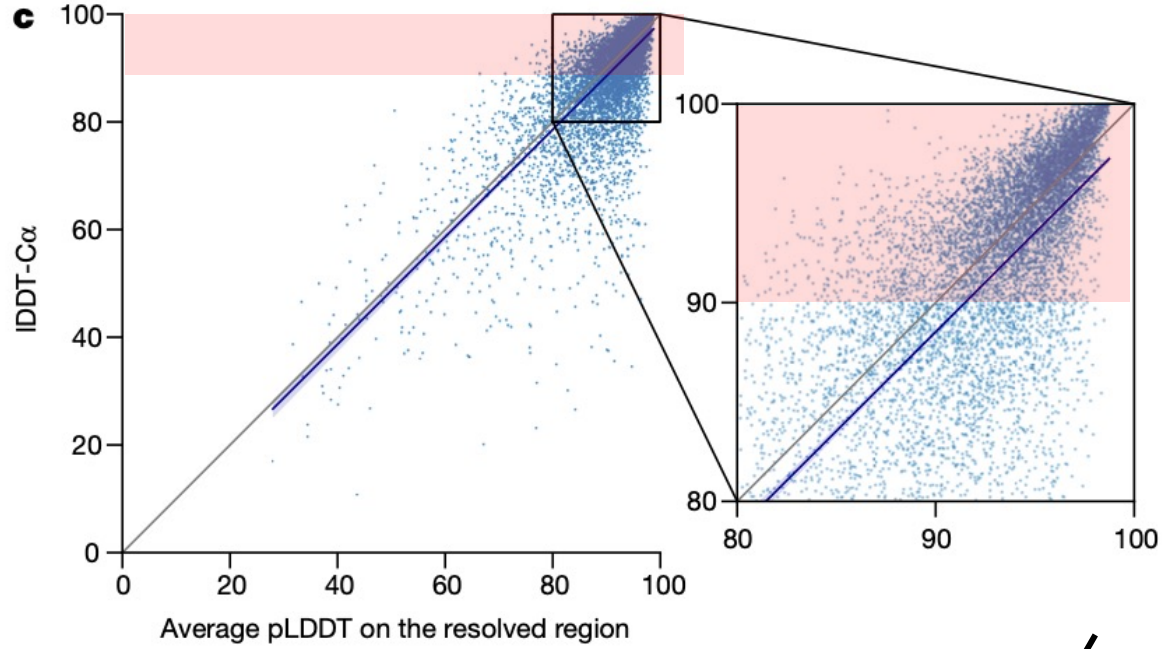
Kathryn Tunyasuvunakool[1✉], Jonas Adler[1], Zachary Wu[1], Tim Green[1], Michal Zielinski[1], Augustin Žídek[1], Alex Bridgland[1], Andrew Cowie[1], Clemens Meyer[1], Agata Laydon[1], Sameer Velankar[2], Gerard J. Kleywegt[2], Alex Bateman[2], Richard Evans[1], Alexander Pritzel[1], Michael Figurnov[1], Olaf Ronneberger[1], Russ Bates[1], Simon A. A. Kohl[1], Anna Potapenko[1], Andrew J. Ballard[1], Bernardino Romera-Paredes[1], Stanislav Nikolov[1], Rishub Jain[1], Ellen Clancy[1], David Reiman[1], Stig Petersen[1], Andrew W. Senior[1], Koray Kavukcuoglu[1], Ewan Birney[2], Pushmeet Kohli[1], John Jumper[1,3✉] & Demis Hassabis[1,3✉]

By the end of this year, EMBL EBI will hold structural models of 130 million proteins

"Everything that relies on a protein sequences we can now do with protein structures" (Mohammed AlQuraishi , Columbia U.)

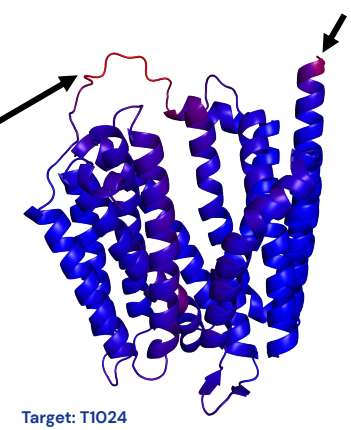
Most predictions by AlphaFold are crystallography-grade
… and the prediction of local quality is excellent



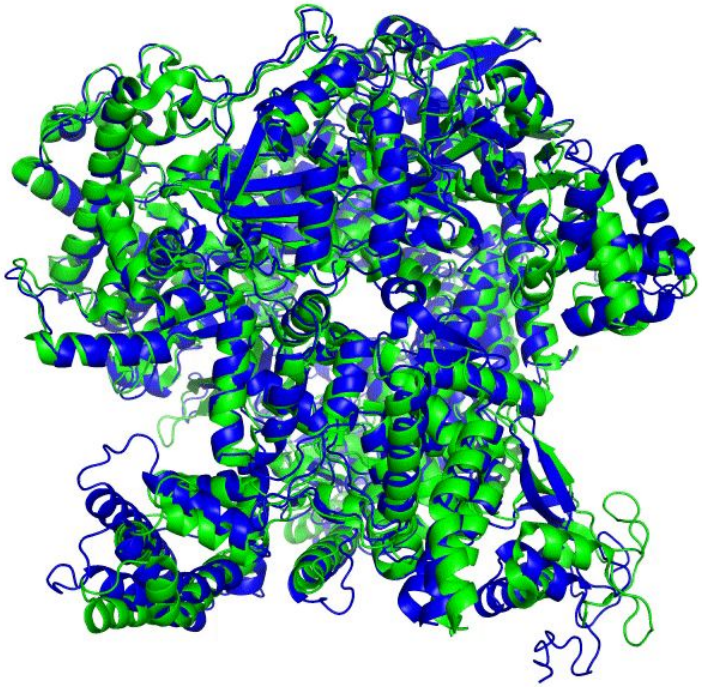


crystallography-grade quality

low quality



Protein example: T1044 (RNA Polymerase)



- Folding as a single long chain
- Long-chain-trained model trained after the submission

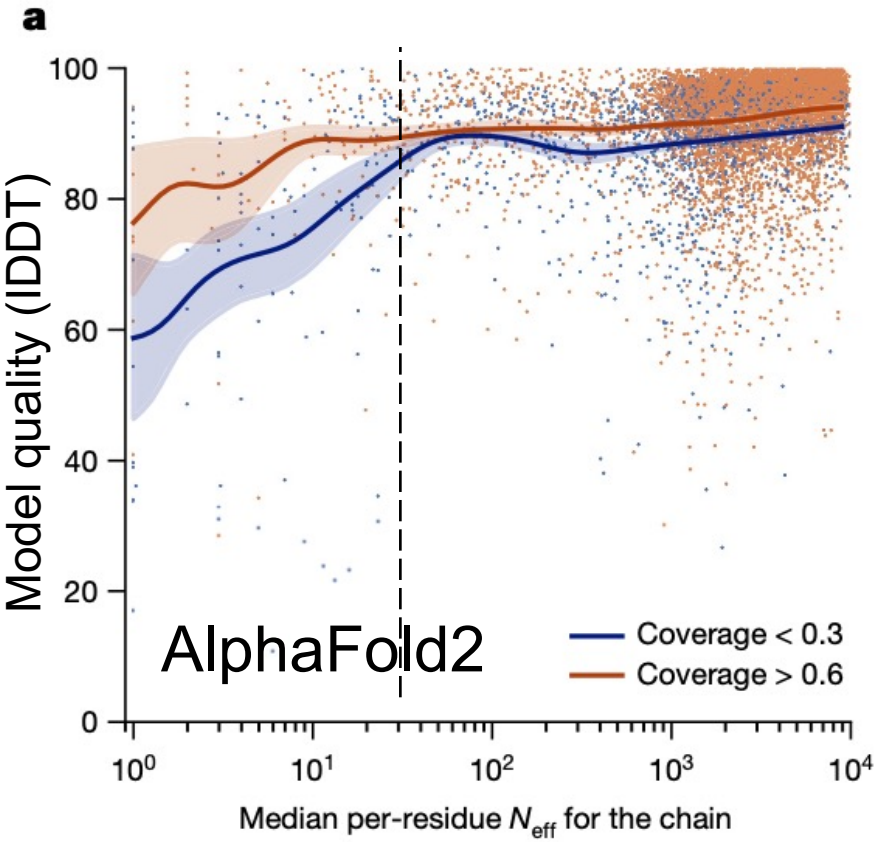
Individual domains

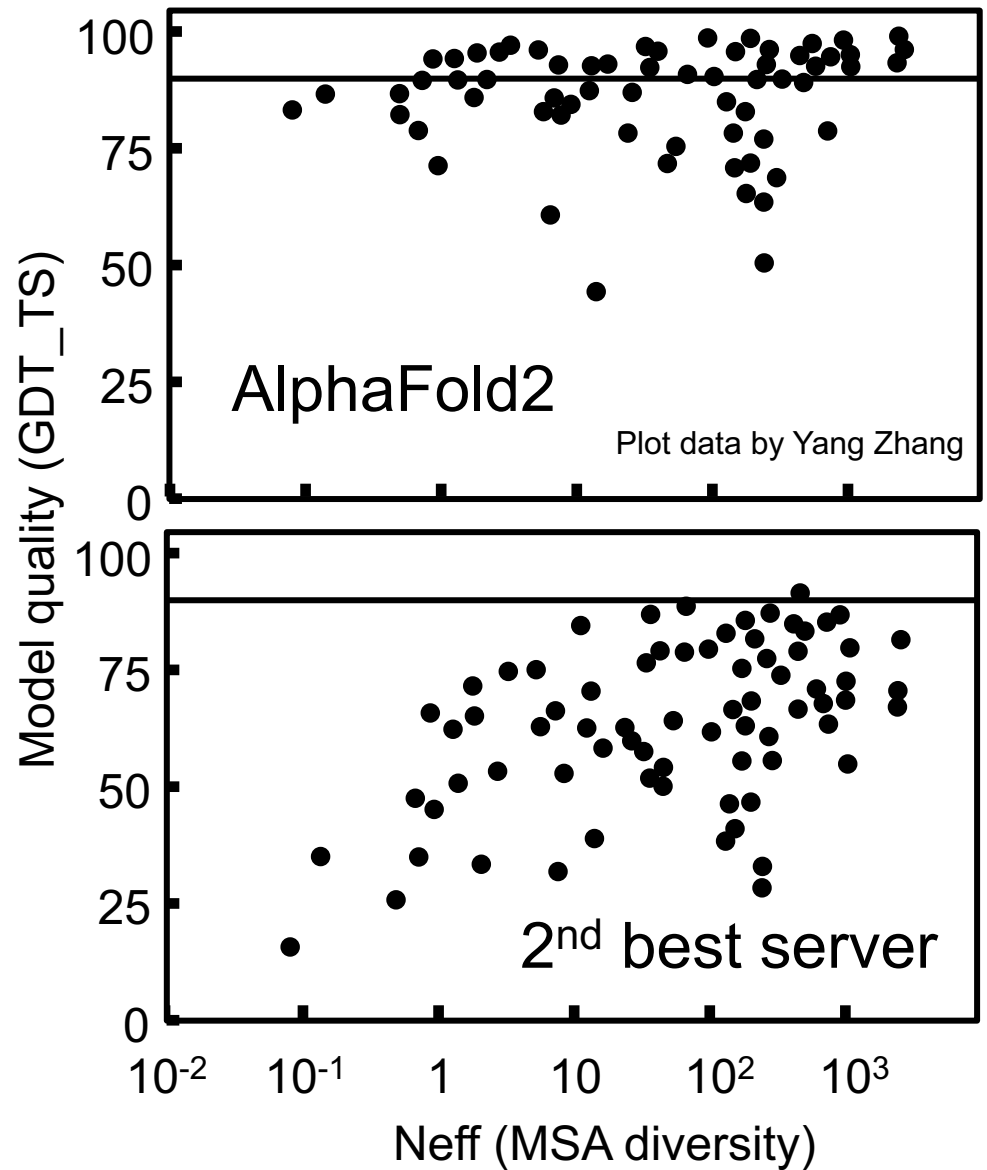
T1041 T1042 T1043

6VR4: Leiman, P.G., et al. Virion-packaged DNA-dependent RNA polymerase of crAss-like phage phi14:2 (CASP target). (To be published.)

Ground truth
Prediction

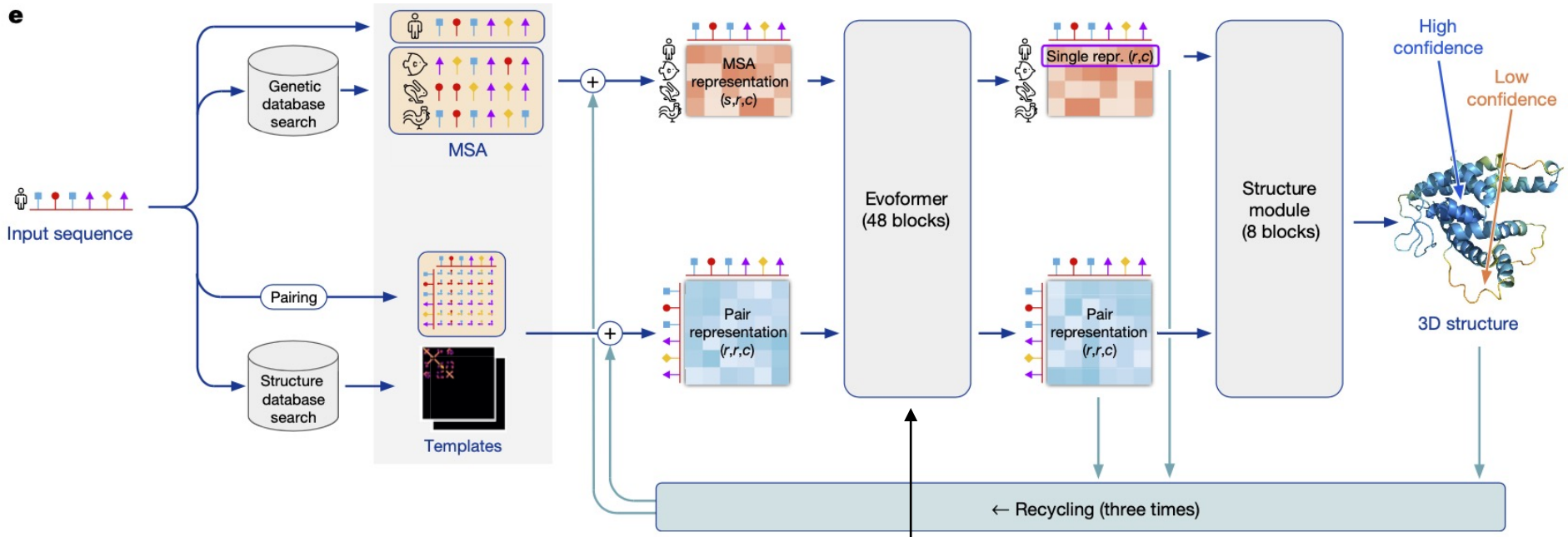
AlphaFold2 can predict accurate models with only 30 sequences in the MSA (and others cannot)





Does AlphaFold really only use correlated mutations in a better way?
It does not look that way!

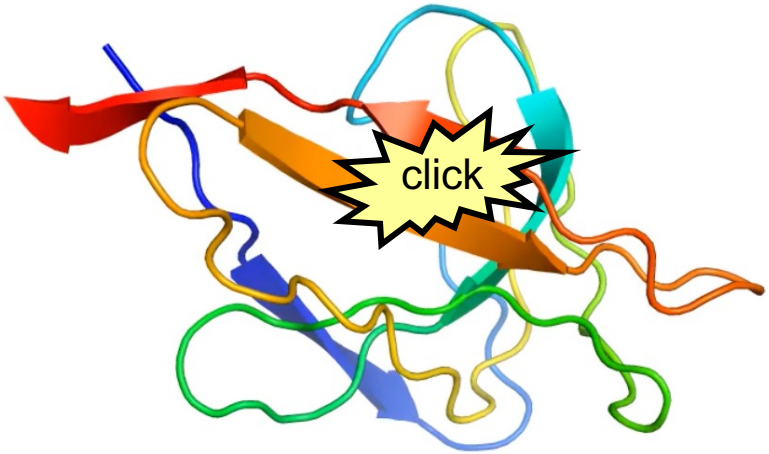
AlphaFold evolves abstract representation of MSA and of residues pairs which improve each other step by step (by “attention”)

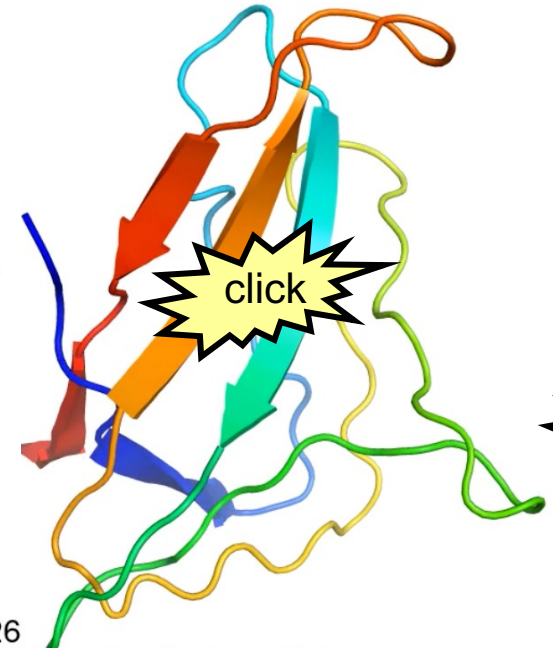


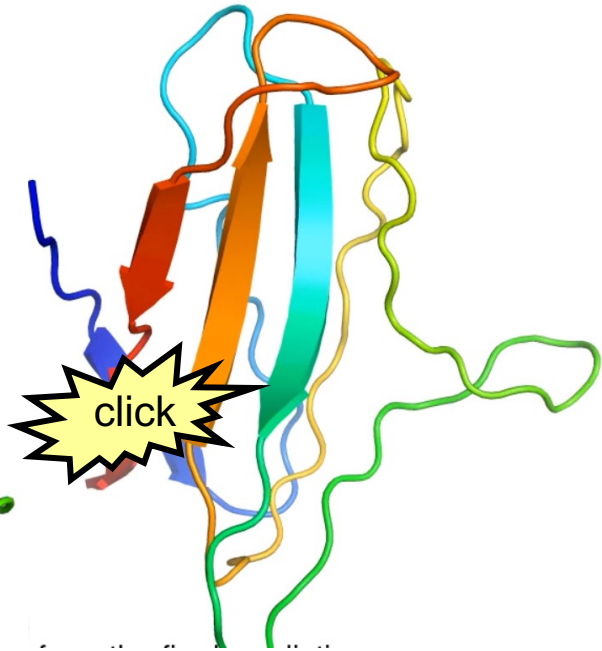
Let’s have a look at a movie of how the predicted structure evolves along the 48 evoformer blocks

AlphaFold learns from structure-MSA pairs what local sequences are compatible with which interacting local backbone geometries – because the local sidechain packing works

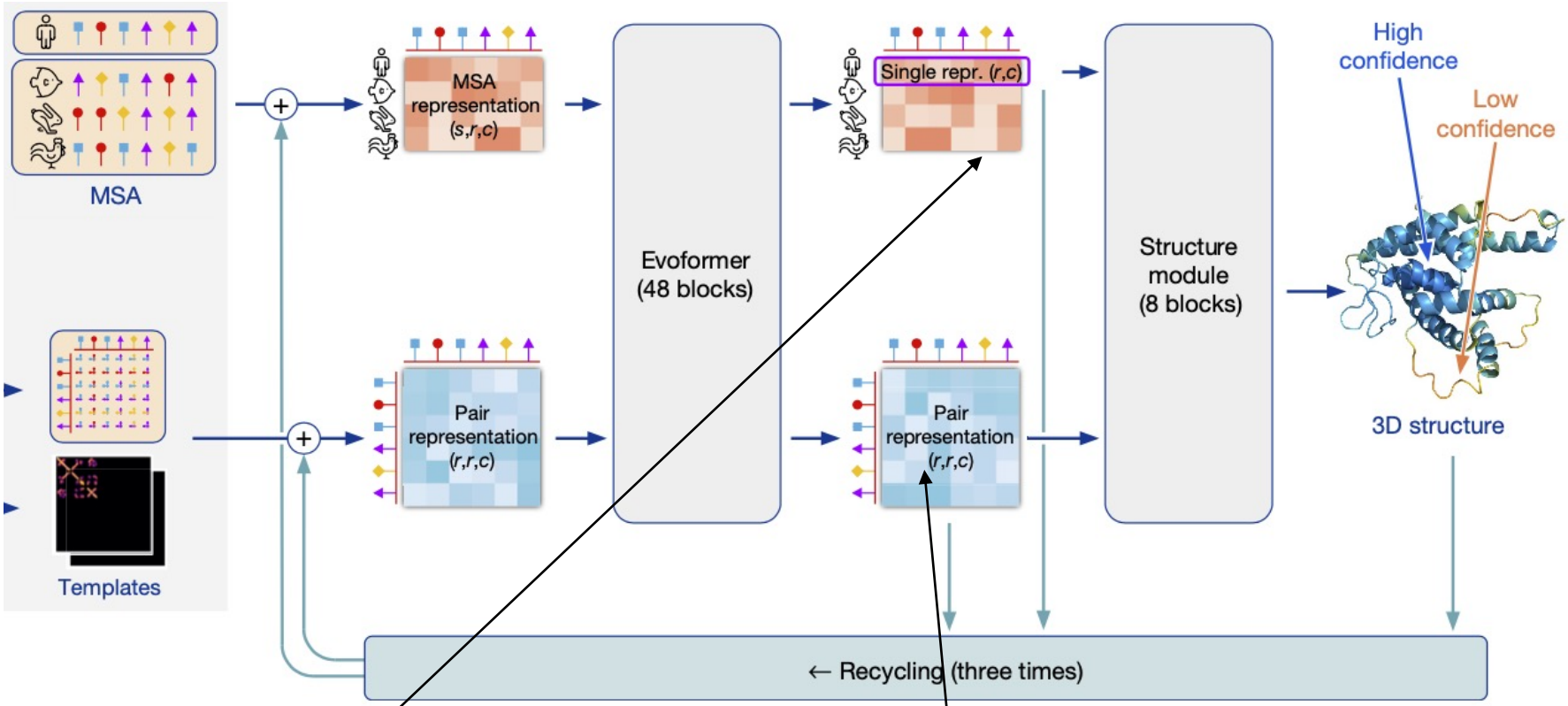
(this is my personal take – but I am quite convinced)



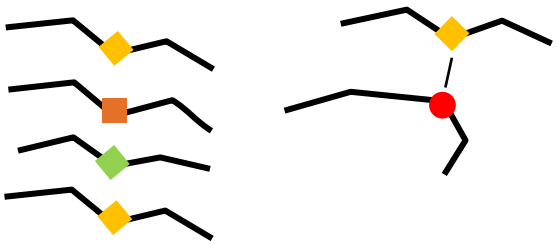




AlphaFold evolves abstract representation of MSA and of residues pairs which improve each other step by step (by “attention”)



This comun holds an abstract representation of the local geometry of these sequences



This cell holds an abstract representation of the local geometry of how the two residues and the neighboring ones interact

What does that mean for biology?

- “This will change medicine. It will change research. It will change bioengineering. It will change everything.”
  Andrei Lupas, MPI Developmental Biology Tübingen.
  See https://www.nature.com/articles/d41586-020-03348-4

- “AF2 is profoundly transformative because it may do for structure what DNA sequencing did for genomics.“
  Mohammed AlQureishi, Harvard
  https://moalquraishi.wordpress.com/2020/12/08/alphafold2-casp14-it-feels-like-ones-child-has-left-home/

- Next tasks expected to be tackled by deep learning:
  protein complexes, protein conformations, protein dynamics, RNA structure, protein-DNA, ligand binding,
  protein design!, ligand design!

sufficient training data?

Thanks for your participation!

See you back at 13:30h ☺