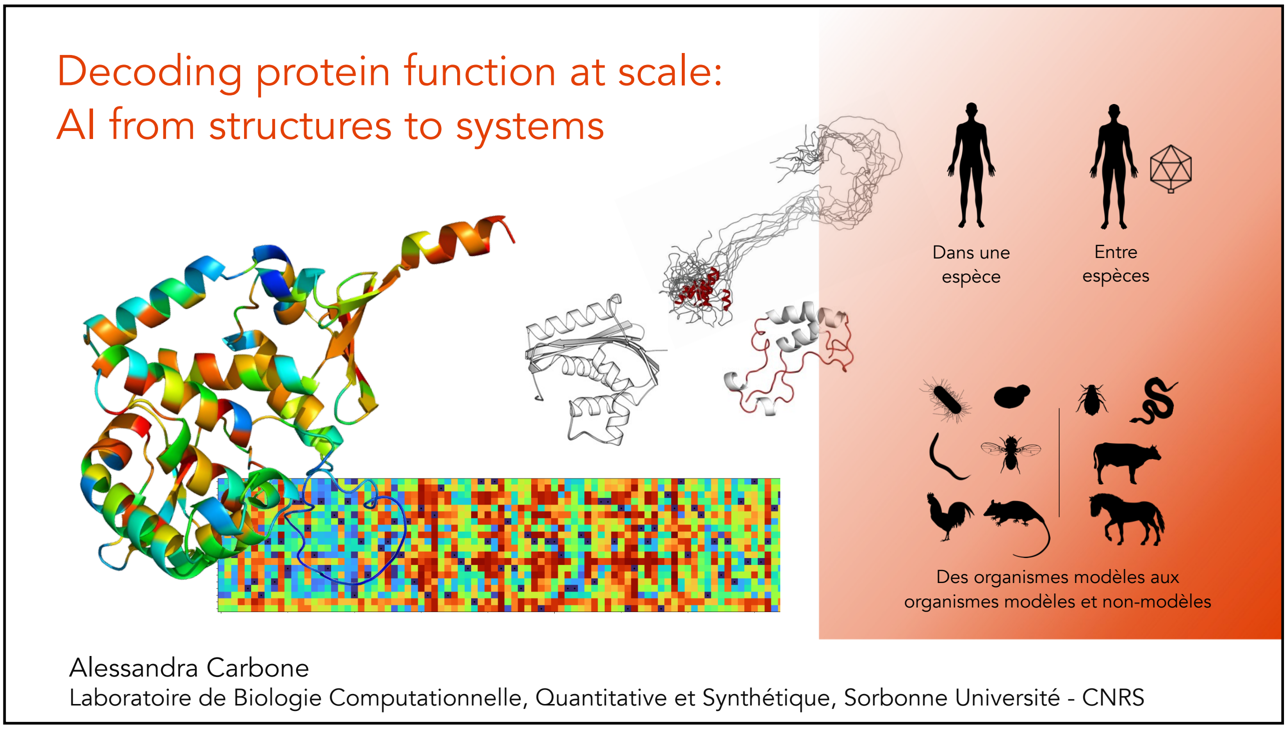


Decoding protein function at scale: AI from structures to systems

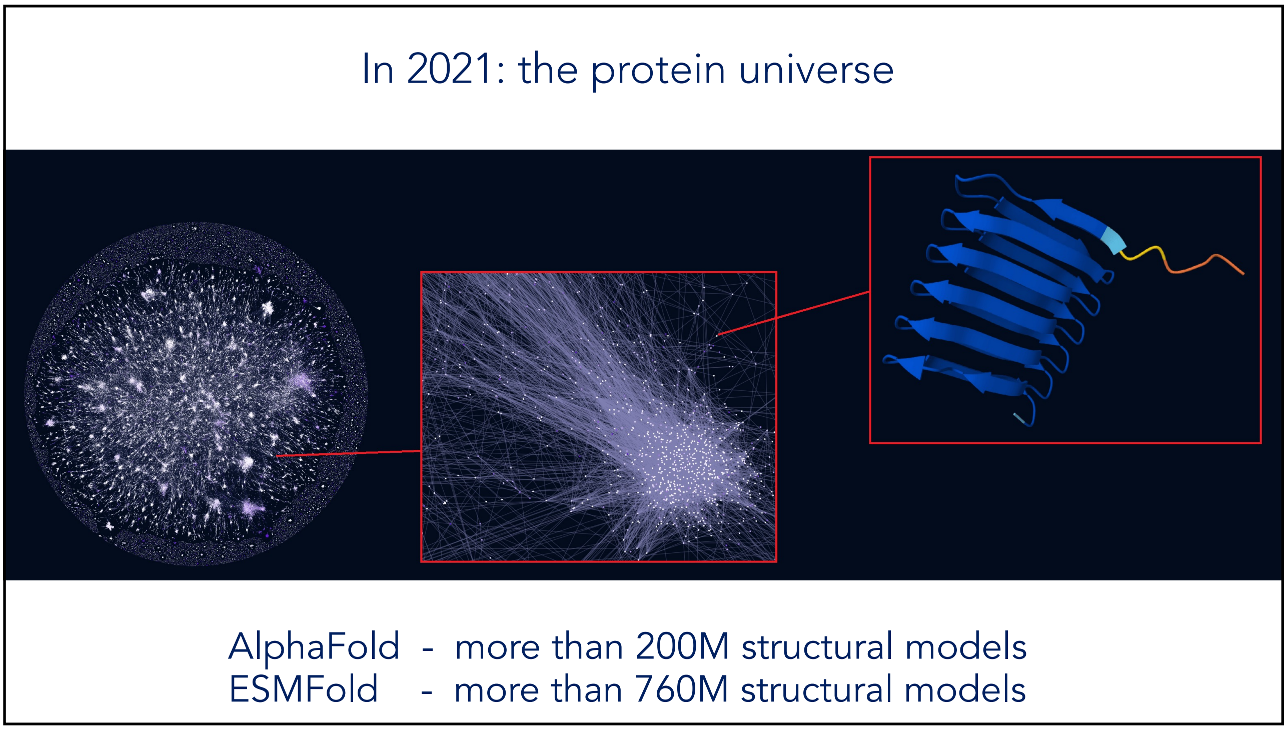
Dans une espèce

Entre espèces

Des organismes modèles aux organismes modèles et non-modèles

Alessandra Carbone
Laboratoire de Biologie Computationnelle, Quantitative et Synthétique, Sorbonne Université - CNRS

1



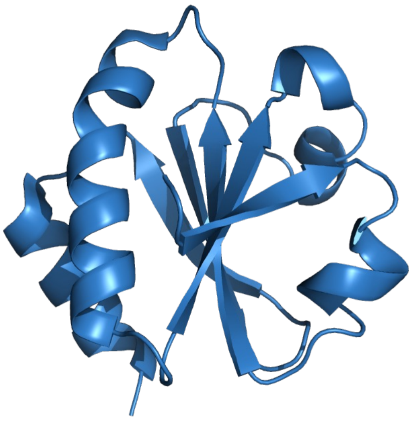
In 2021: the protein universe

AlphaFold - more than 200M structural models
ESMFold - more than 760M structural models

2

AlphaFold 2 solved the protein structure prediction problem

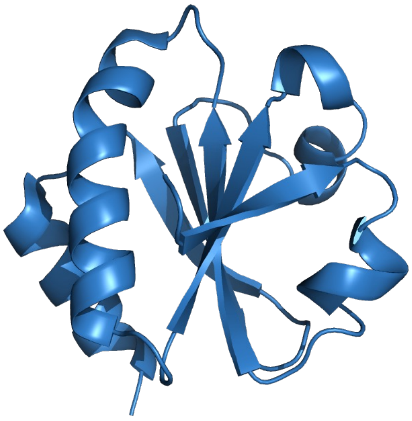
MTLRKLLTGELLTLASRQQLIDWMEADKVGGP
LLRSALPAGWFIADKSGAGERGSRGIPEDRNR
GLAASCWFIADKCFCVLALLTLAKLEKDFRLLGL
CRQQLIDWGELLTLASADKVGGPLLRSALPLEK
DFRLLGAGWFIADKSGAMETLRKLLTGERGSS
CWFIADKCLAKLCFCVLALLRGIPEDRNRGLAA



3

AlphaFold 2 solved the protein structure prediction problem

MTLRKLLTGELLTLASRQQLIDWMEADKVGGP
LLRSALPAGWFIADKSGAGERGSRGIPEDRNR
GLAASCWFIADKCFCVLALLTLAKLEKDFRLLGL
CRQQLIDWGELLTLASADKVGGPLLRSALPLEK
DFRLLGAGWFIADKSGAMETLRKLLTGERGSS
CWFIADKCLAKLCFCVLALLRGIPEDRNRGLAA

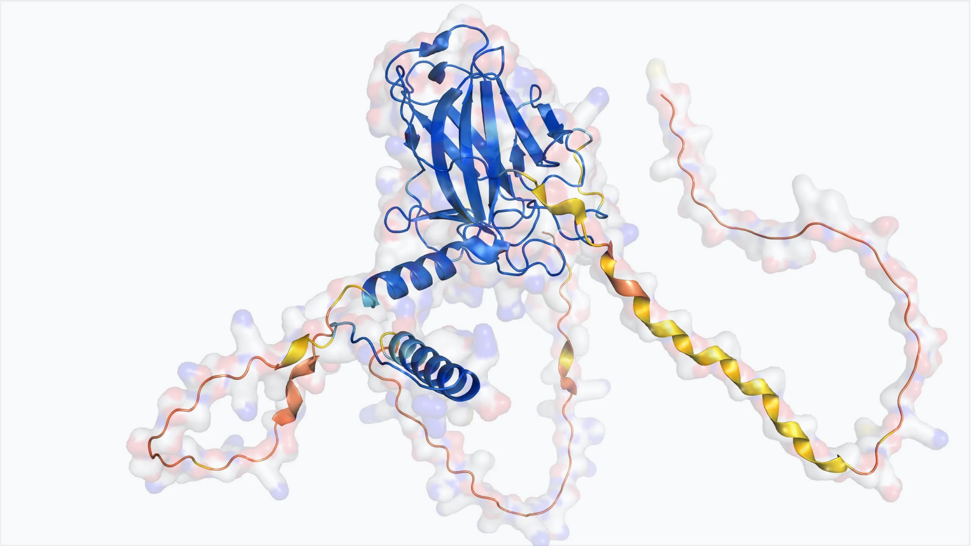


Proteins are complex microscopic machines that drive every process in a living cell.

Their 3D structure largely defines the protein’s function.
Knowledge of their shape is critical for drug discovery and understanding disease.

4

AlphaFold 2 solved the protein structure prediction problem



p53 is a cellular tumor antigen related to diseases such as cancer

5

This was the first major proof that AI/DL can be a powerful tool to advance biological science

6

This was the first major proof that AI/DL can be a powerful tool to advance biological science

Biological data can be generated computationally

7

Foundational models for

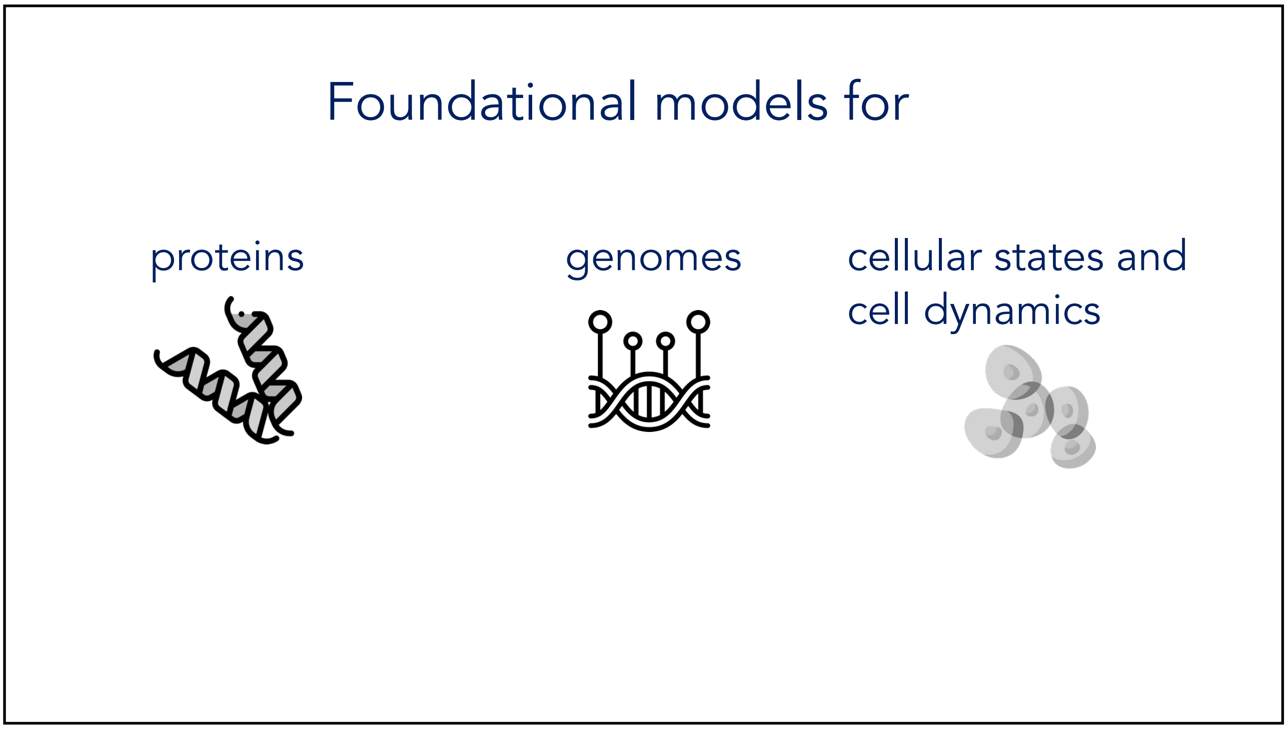
proteins

Pretrained on raw/ unlabelled sequences, they reveal intricate biological patterns

These patterns generalize across tasks, can be reused across biological problems

Sometimes they enable predictions without additional training

8



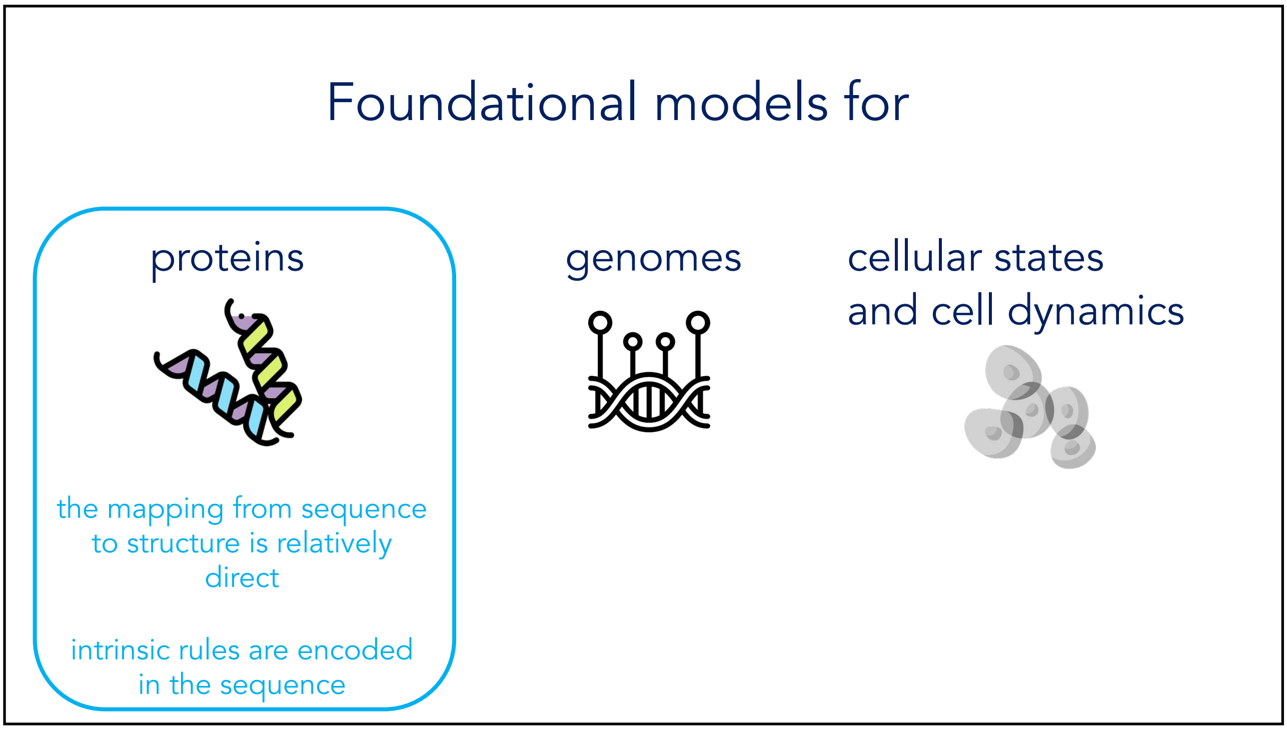
Foundational models for

proteins

genomes

cellular states and cell dynamics

9



Foundational models for

proteins

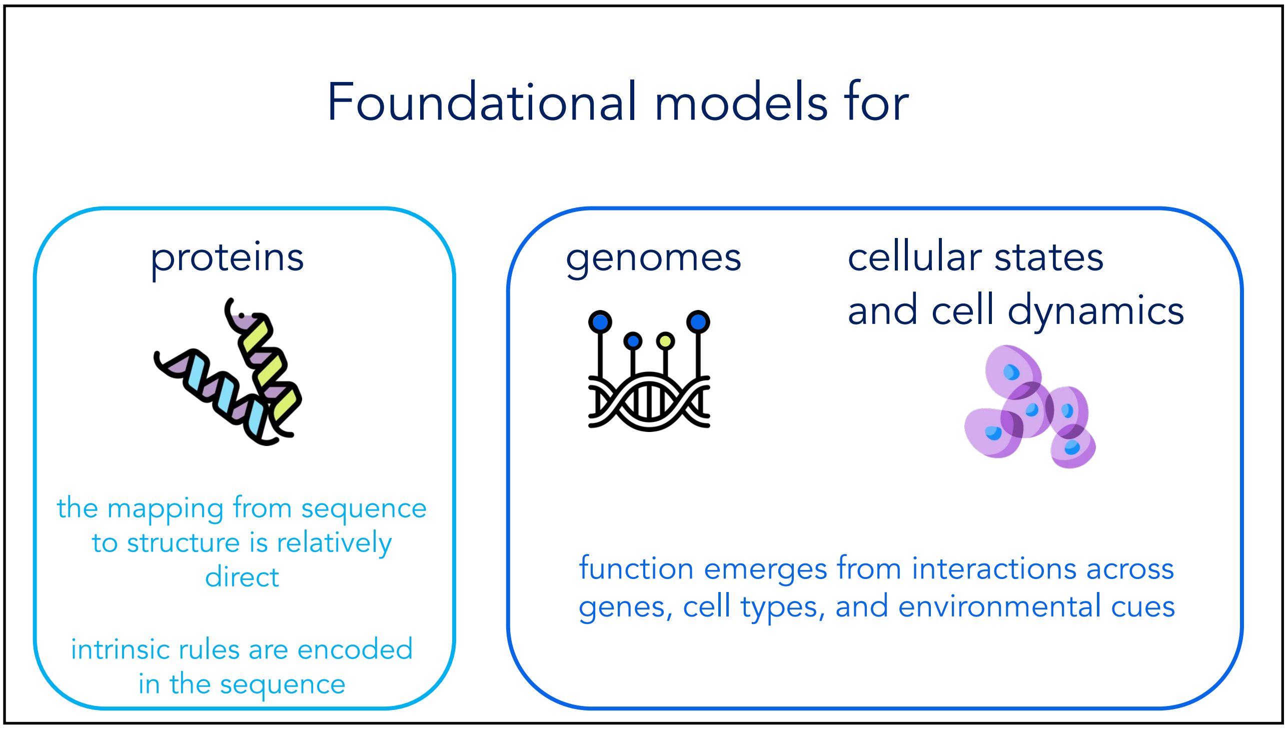
the mapping from sequence to structure is relatively direct

intrinsic rules are encoded in the sequence

genomes

cellular states and cell dynamics

10



Foundational models for

proteins

the mapping from sequence to structure is relatively direct

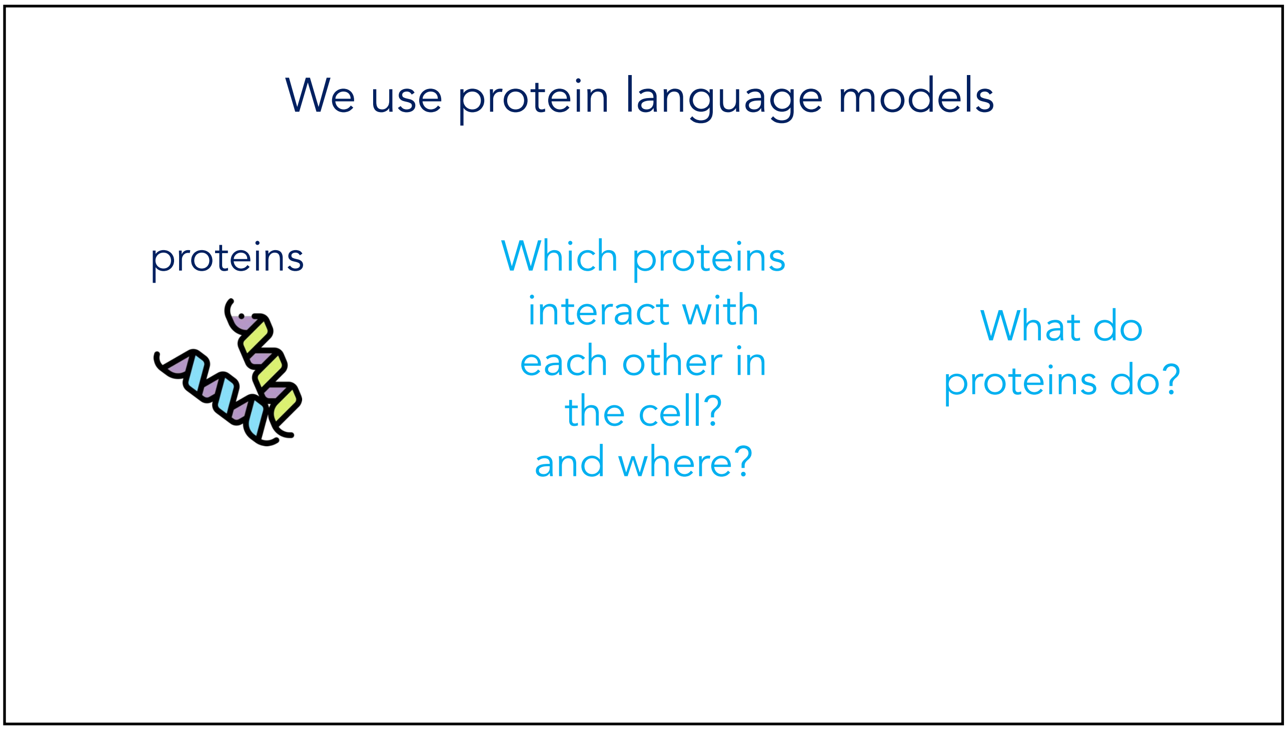
intrinsic rules are encoded in the sequence

genomes

cellular states and cell dynamics

function emerges from interactions across genes, cell types, and environmental cues

11



We use protein language models

proteins

Which proteins interact with each other in the cell? and where?

What do proteins do?

12

We use protein language models

proteins

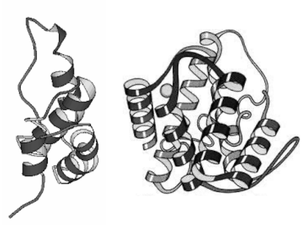
Which proteins interact with each other in the cell?
and where?

What do proteins do?

13

2006-2013

Molecular modeling



on structures

14

2006-2013

Molecular modeling

on structures

2021-

Deep learning

MTLRKLLTGELLTLASR
QQLIDWMEADKVGG
PLLRSALPAGWFIADK
SGAGERGSRGIPEDRN
RGLAASCWFIADKCFC
VLALLTLAKLEKDFRLL
GLC...

MRQQLIDWGELLTLAS
ADKVGGPLLRSALPLE
KDFRLLGAGWFIADKS
GAMETLRKLLTGERGS
SCWFIADKCLAKLCFC
VLALLRGIPEDRNRGL
AAT...

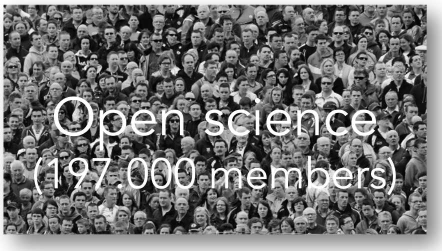
on sequences

15

2006-2013

Modecular modeling

3 years



with World Community Grid (IBM)

2021-

Deep learning

1 week

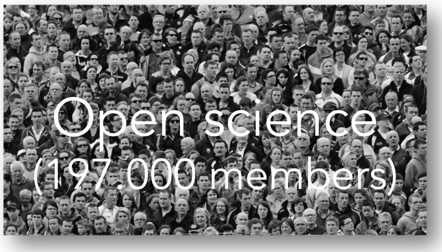
With a cluster of GPUs at Sorbonne Université

16

2006-2013

Modecular modeling

3 years



with World Community Grid (IBM)

2 000 x 2 000

2021-

Deep learning

1 week

With a cluster of GPUs at Sorbonne Université

20 000 x 20 000

17

AI has shifted the limits

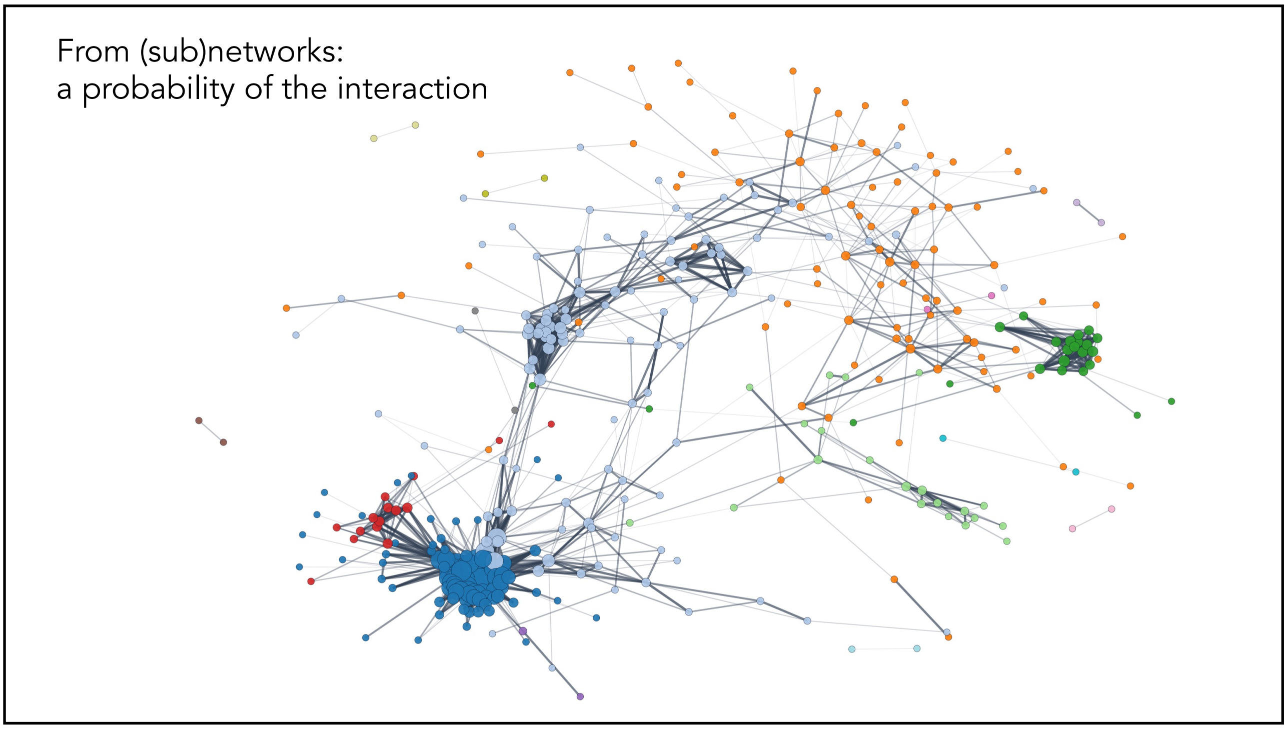
We can now build:

- personalised PPI network
- cross-species PPI networks

⬆

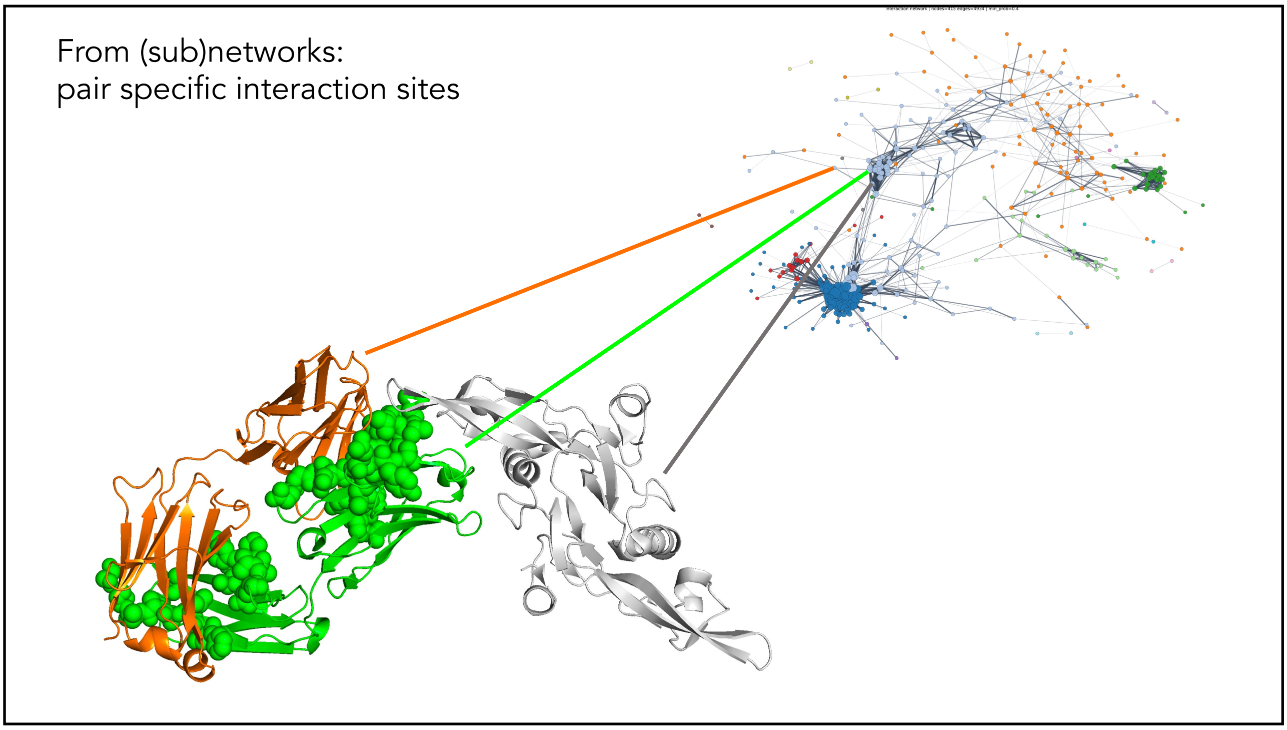
20 000 x 20 000

18



From (sub)networks:
a probability of the interaction

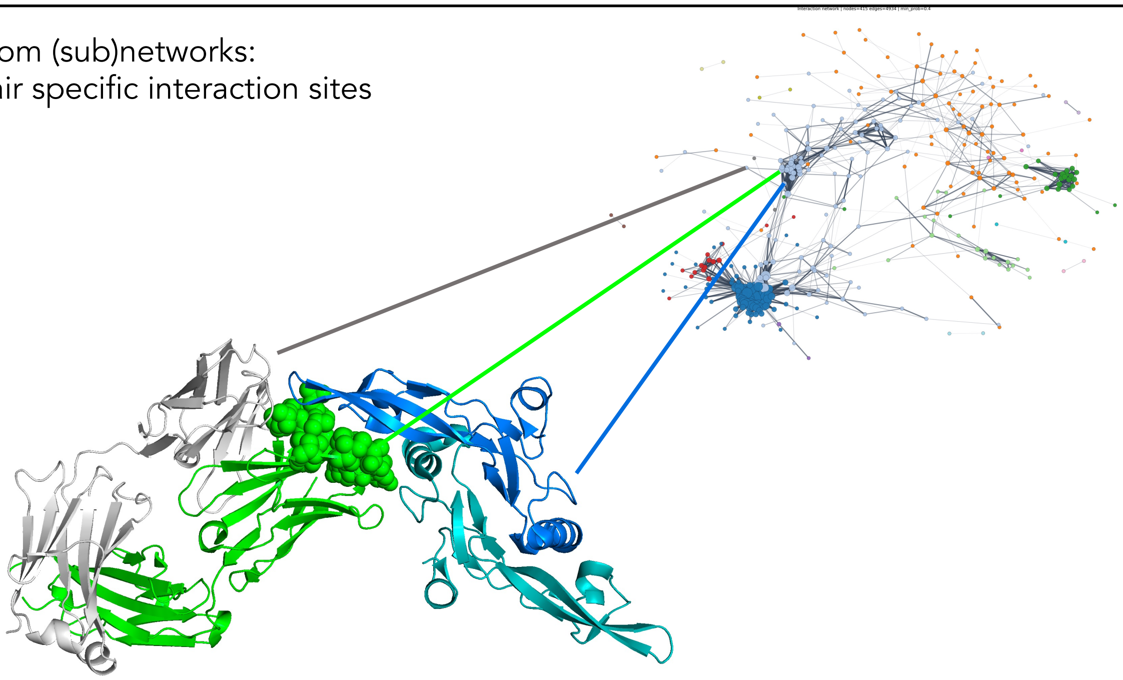
19



From (sub)networks:
pair specific interaction sites

20

From (sub)networks:
pair specific interaction sites



21

AI has shifted the limits

AI enabled

- scaling in PPI reconstruction
- improved interaction prediction
- improved binding-site specificity

22

We use protein language models

proteins

Which proteins interact with each other in the cell?
and where?

What do proteins do?

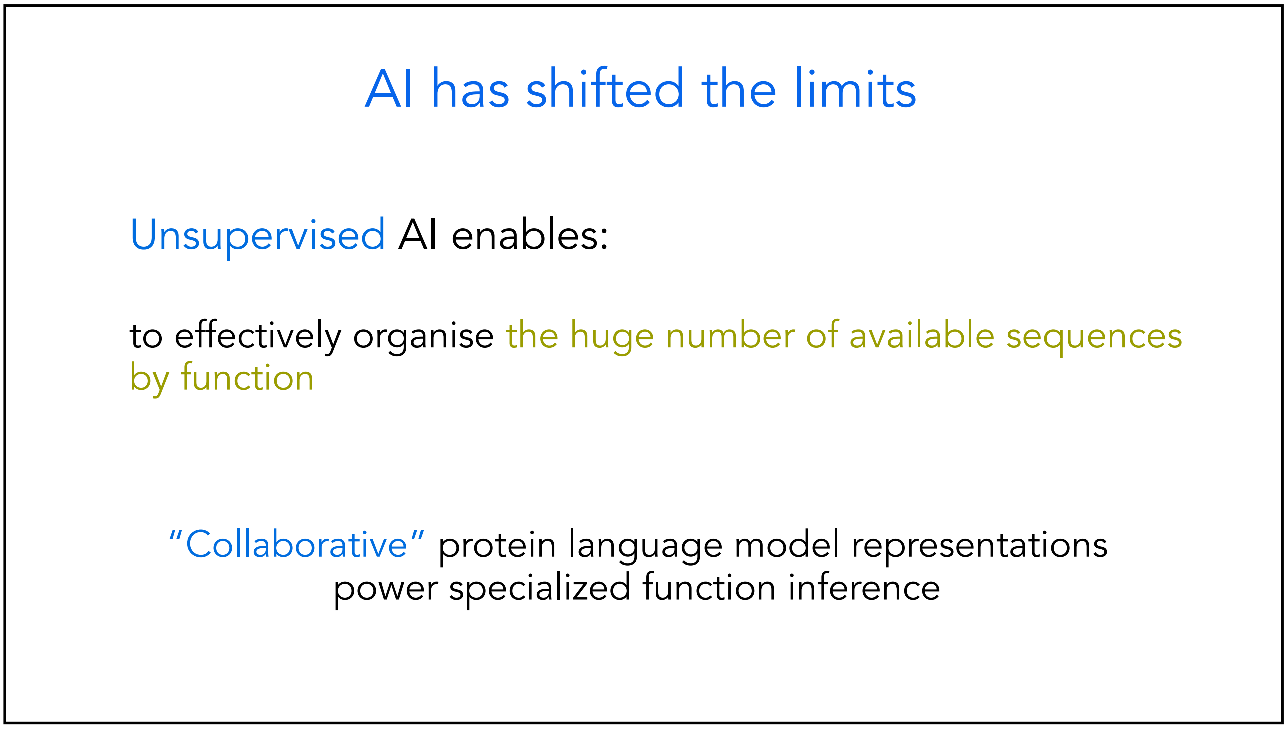
23

What do proteins do?

500 000
experimentally functionally characterised sequences

2.4 billion
non-redundant protein sequences available

24



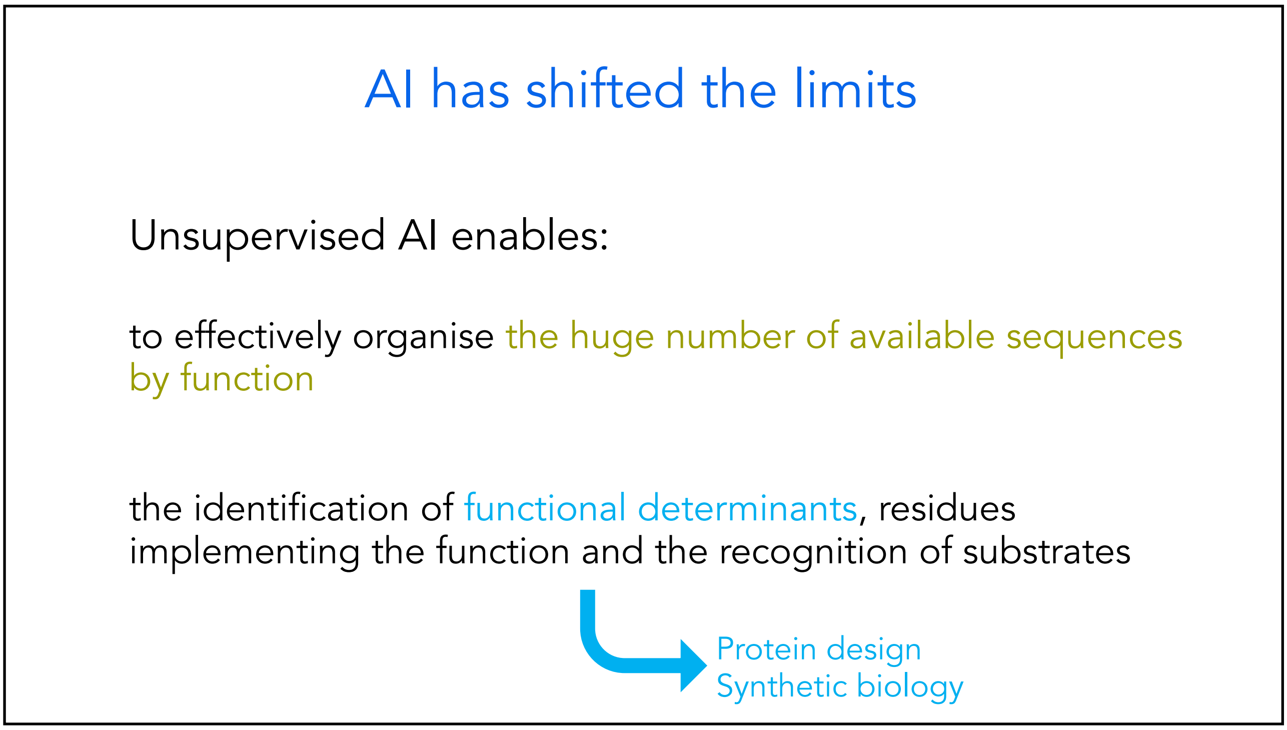
AI has shifted the limits

Unsupervised AI enables:

to effectively organise the huge number of available sequences by function

"Collaborative" protein language model representations power specialized function inference

25



AI has shifted the limits

Unsupervised AI enables:

to effectively organise the huge number of available sequences by function

the identification of functional determinants, residues implementing the function and the recognition of substrates

Protein design
Synthetic biology

26

AI has shifted the limits

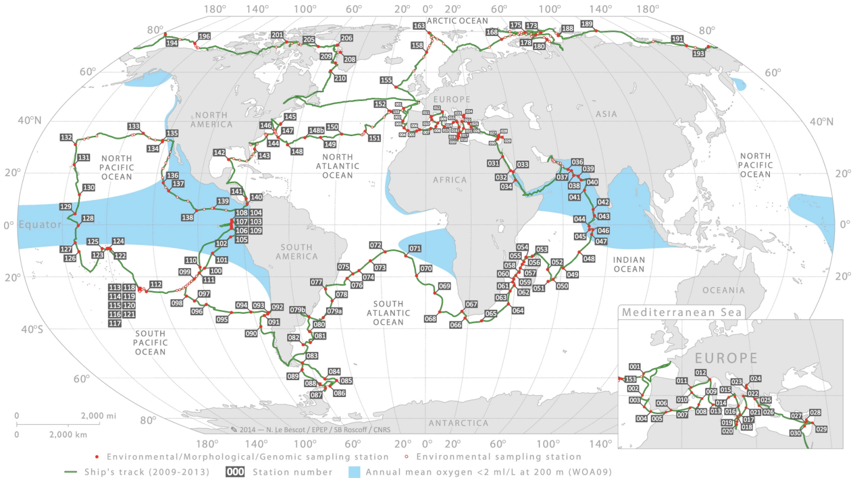
Unsupervised AI enables:

to effectively organise the huge number of available sequences by function

a large-scale mapping of protein functions in the environment

27

Tara Ocean project

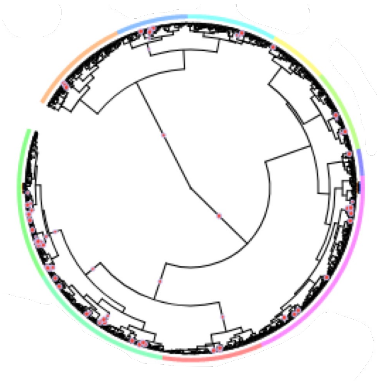


more than 200 stations

[2] Pesant et al. 2015.

28

Unsupervised reconstruction of functional spaces/trees



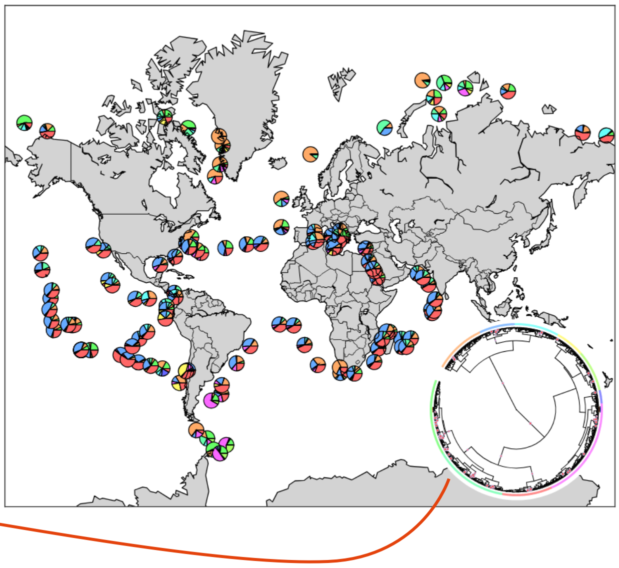
Clusters/subtrees describe functional diversity

🟩 🟩 🟥 🟪 🟦 🟩 🟨 🟦 🟦 🟧

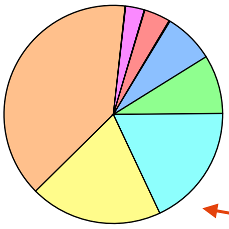
functional clusters ⟷ environmental parameters:
temperature
salinity
physico-chemical

for each protein family

29



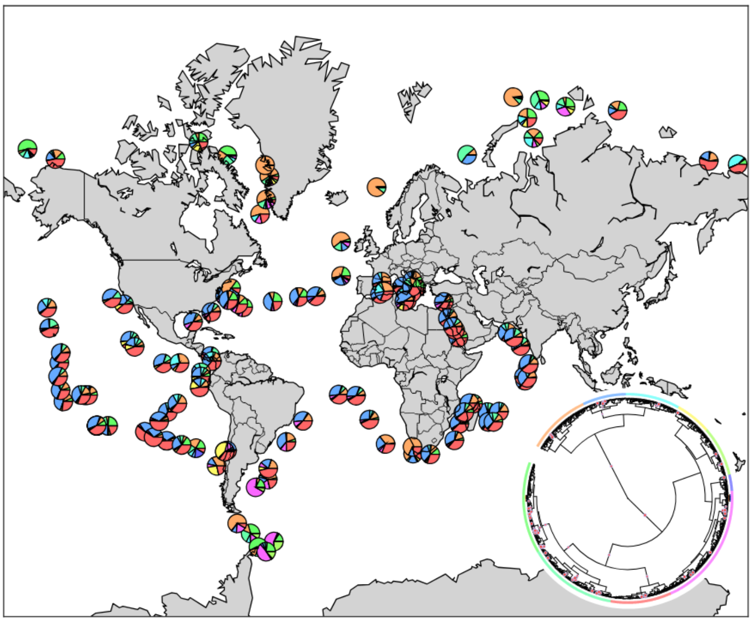
For each station:
a map of the functional abundance of cold-shock proteins (MetaG)



Mapping the abundance of functions to geographical locations

30

A planetary map of cold-shock proteins in the environment

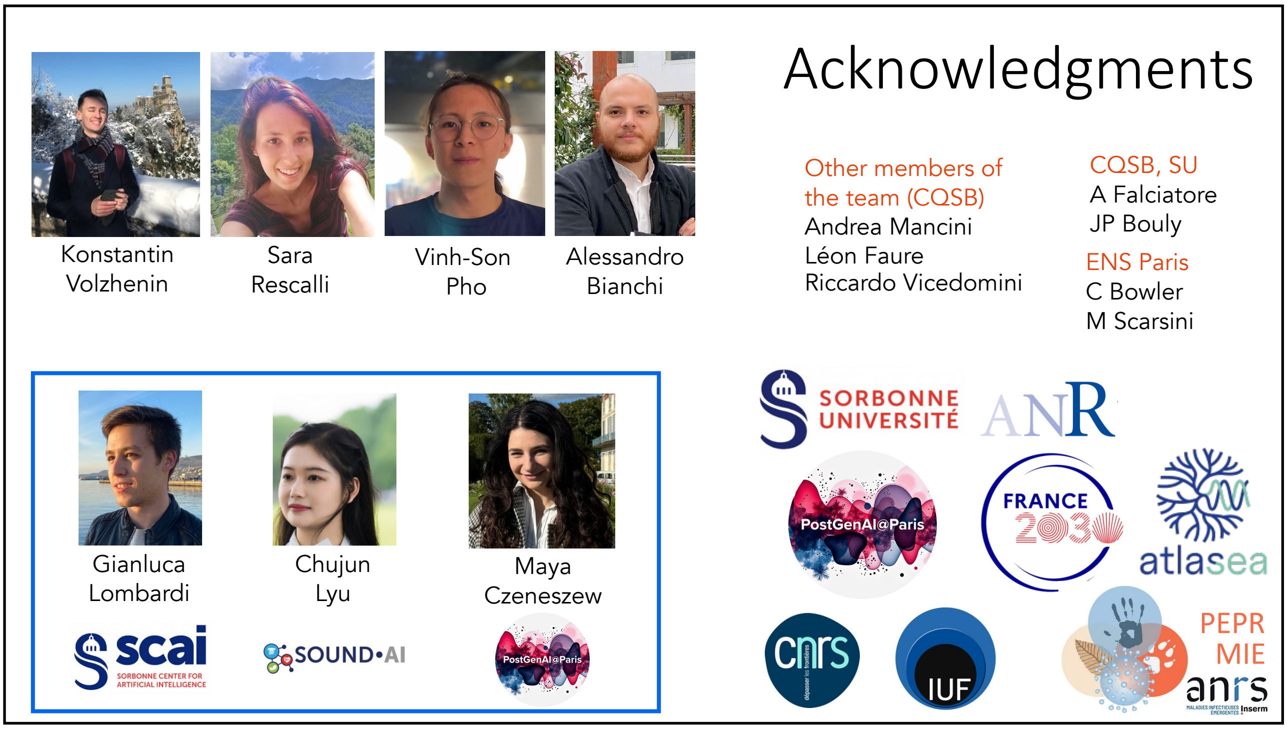


31

Deep learning enables today to

identify interactions and interaction sites at genome scale
and
explore protein functions in both the environment and synthetic space!

32



Acknowledgments

Konstantin Volzhenin

Sara Rescalli

Vinh-Son Pho

Alessandro Bianchi

Other members of the team (CQSB)
Andrea Mancini
Léon Faure
Riccardo Vicedomini

CQSB, SU
A Falciatore
JP Bouly

ENS Paris
C Bowler
M Scarsini

Gianluca Lombardi

Chujun Lyu

Maya Czeneszew

PEPR MIE

33