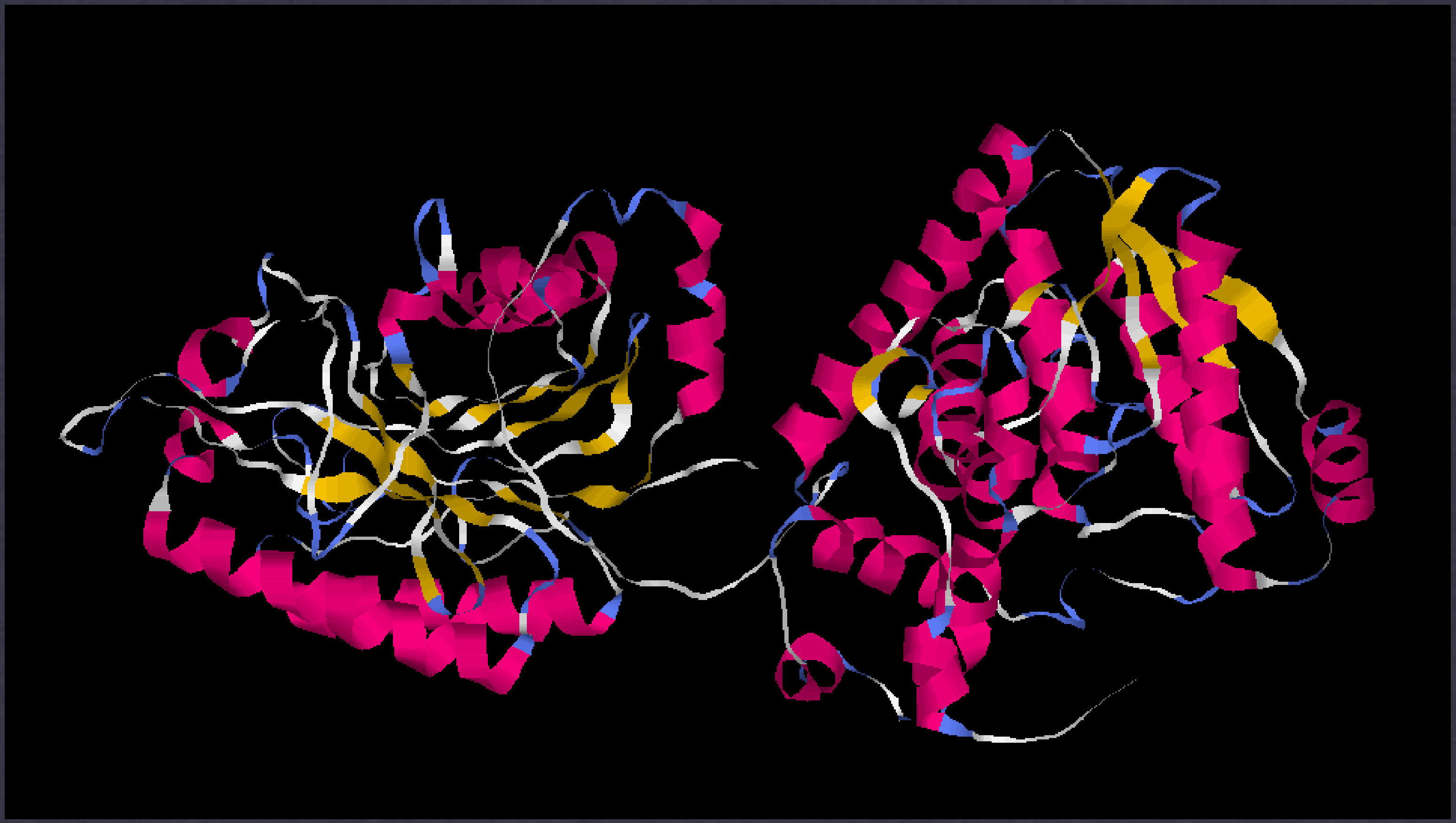


Discovering the Mechanisms of Life

Computer Science In Practice ~
Dr.Gianluca Pollastri

Outline

- The Take-Home Message
- What are proteins?
- Learning stuff about proteins
- Using computers to learn
- Examples of what we have learned
- Conclusions and questions

Take Home Message

- COMPUTERS CAN BE USED TO DISCOVER. MORE: COMPUTERS CAN DISCOVER
- A massive amount of information about how life works, out there
- Information that is raw, messy and not well understood:
  - It’s like having an encyclopedia with everything in it, but we don’t know the language it’s written in.
  - We can use intelligent computer programs to decipher the language.

Introduction: the numbers

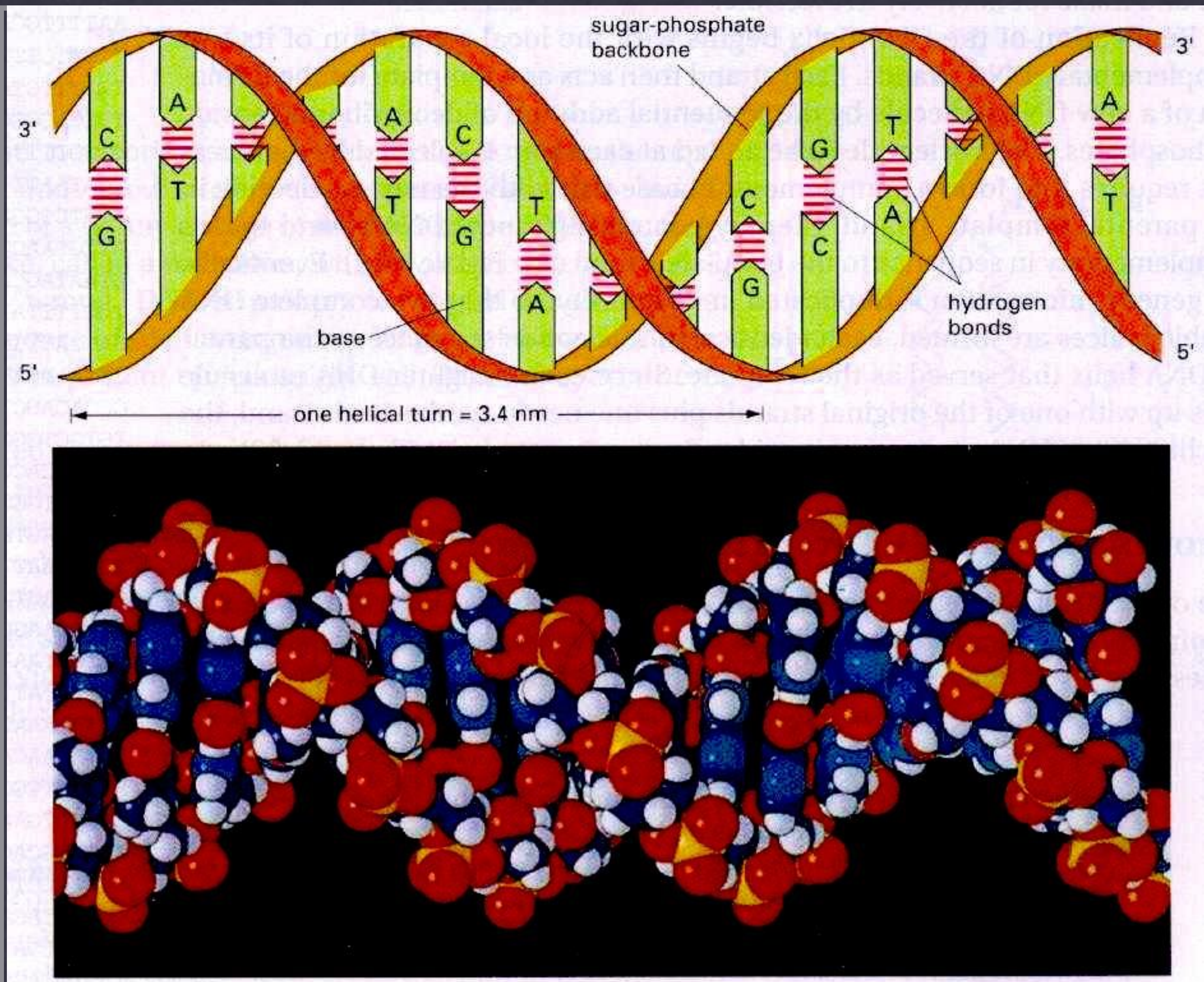
- First rough draft of the human DNA published in 2000
- Now we know DNA sequences for thousands of organisms
- A total of 4,743,666,948,635 letters of DNA, freely available.

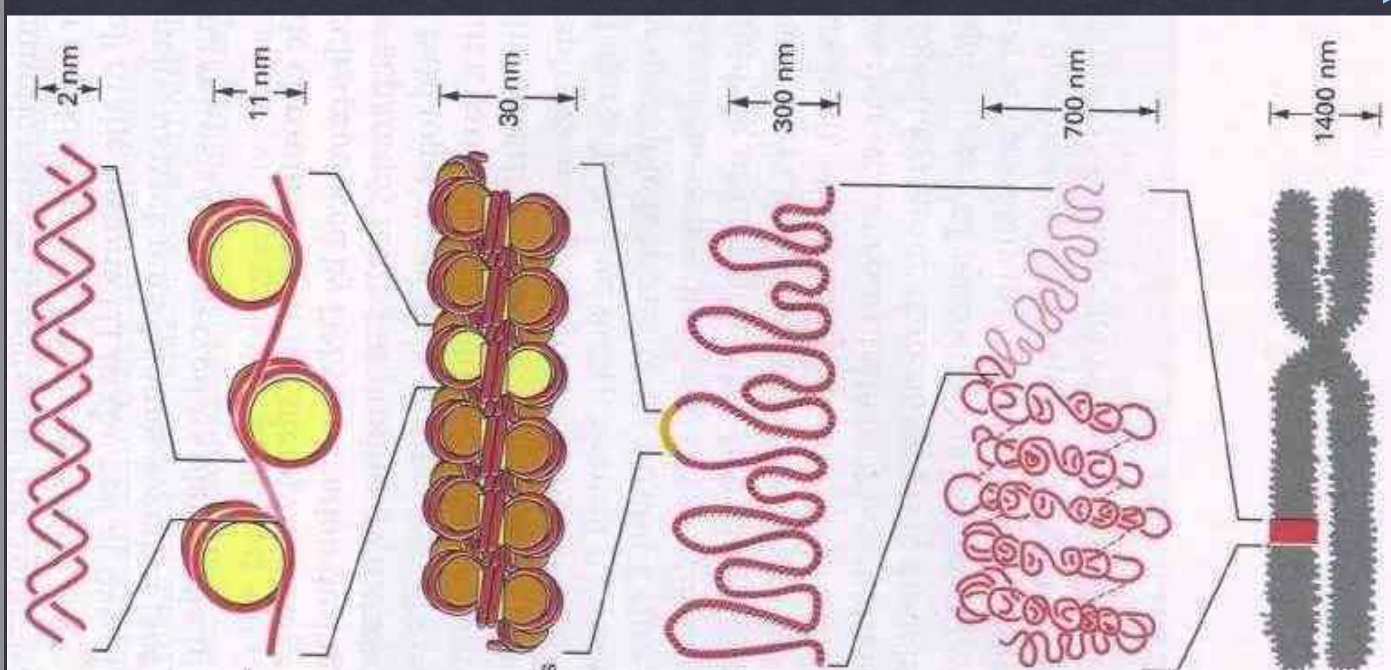
The numbers (2)

- 4,743,666,948,635 letters.
- 4 million copies of War and Peace
- A 140 storey skyscraper full of books.
- (last year it was 100)

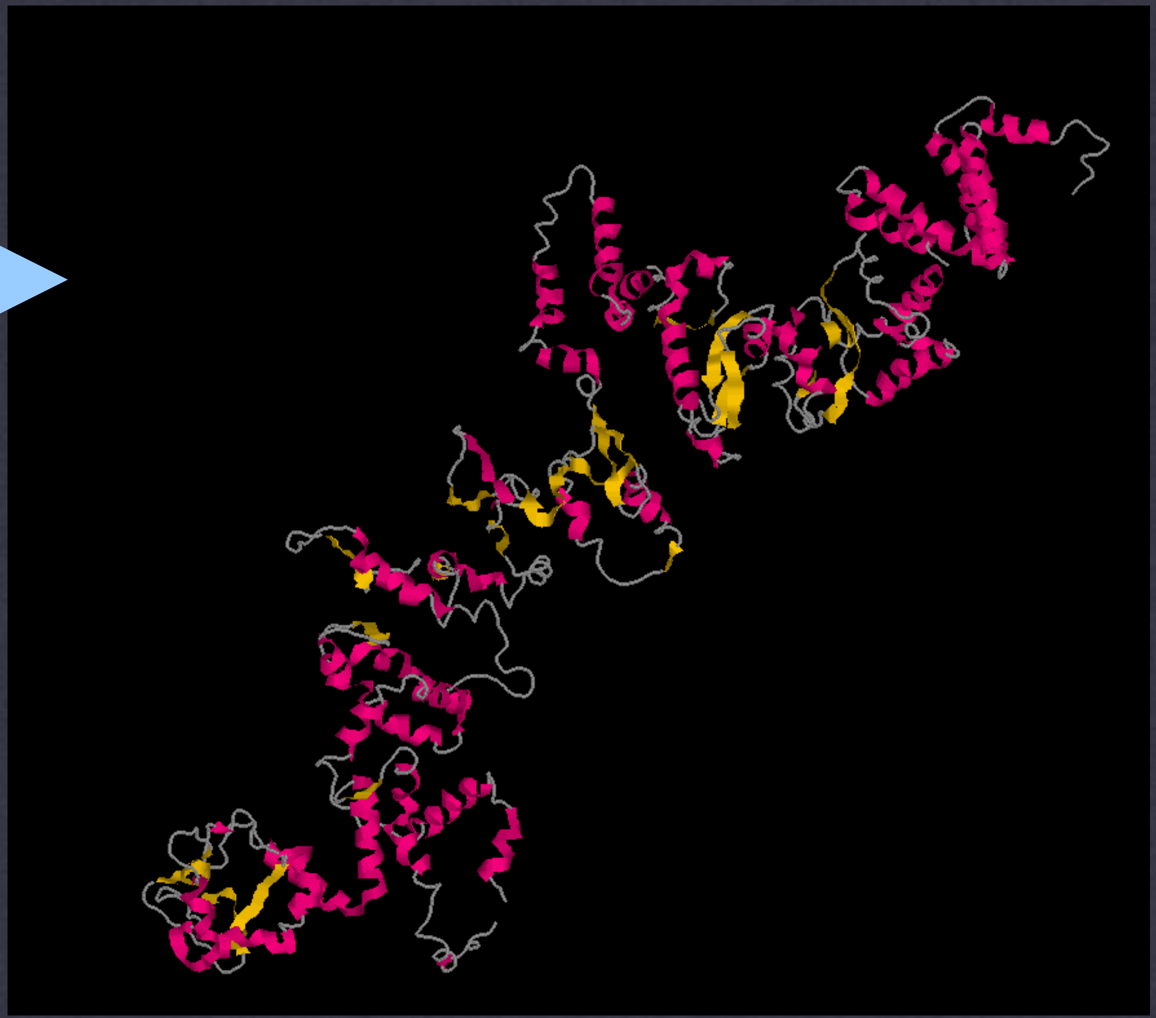
The numbers (3)

- A 140 storey skyscraper full of books.
- Next year it will be ~190.
- All written in alien language.
- If we decipher it we have the keys to the meaning of life.





MAKKKDLTTDNEIFVAQKLAEEELNTNEINEPLERLDFKSFDNNKELLDYQQQALINAFRMLVAYFRDFKENKKEFYAFY
QKHYSFAHCDFAKKKLNPLLKSHFKVENHCVSFENFINRLAFYMATGSGKTIVIIKLVELLSVAIRMGLIPKKNIMFFSA
NENLIQQFEKEIEKYNRNKDYFKQIDFKSLKSVTHKDFYRAPKDSVIKQITLFYYRADLMNDEESKENLLNYKDYWDNGE
NYVILDEAHKGNKSESKRQAIFSLLSLKGFLFNFSATFTEESDLITAVYNLSVGEWVKLGYGKESVLLKKNNLNAFKELK
DLNDREKEIALLKALLLLGMQKRYKTEGYFYDPLMLVFTHSVNVKNSDAEIFFKTLARVIENDDGSDFLKAKEDLLEELK
NPEFLFSDDKDKDYKVKVFKEGLKSMDFKGLKEEVFYANNGHIEVIINPKNNQEIAFKLNTSDKVFCLIRIGDITEWICE
KLKSVKVVSKNLSFKEESYFSQIDKSSINILVGSRTFDTGWDSTRPSVILFLNIGLDDDAKKLVKQSFGRGVRIESVKNQ
RQRLAYLDIDGAIKKALKPNAAMLETLFVIPTNYASLEAILKFQKESENKGENRGSWREIKLEKTPIKHALFVPCYRKEQ
TSILELPESASFKMSEKNFKDLKEYFNLMSEKHFILKHEIYDPKDYMQLKKMTQEAHFNKVSTWHYKDLDYMISEIKGKL
YPNQKVPKDEFNALDSEKIVHFKRIKVKADKKEELVKTIQEVKEYAPLDKETLIKKIAQGEIDPYDTEKHKQNKTFKVGG
AELLKLKEHYYTPLIKAKNCDWLKHVVKVESESDFLEELLKITETLQENYDFWAFSKIDEHLDNLFIPYFNNAAERKFFP
DFIFWLEKGGTQIICFIDPKGSKHTDYEHKADAYQLFKDKIFNPKDNPNLKIKVVLKFYGDKDEVADGYRDYWIKKGKLE
DFFLKQLA



DNA makes proteins..

..and proteins go do the jobs of life

Proteins

- Once you take out the water, you are mostly made of proteins
- Most of your tissues are proteins, hormones are proteins; proteins transfer most of the messages around your body
- Proteins, like DNA, are sequences of letters (amino acids)
- But unlike DNA, once the letters are all together, they fold over forming very interesting, unique structures; that determine how they work

Protein shapes

Shapes encode functions

Protein function

- Explains life: proteins *are* life.
- Once we know how life works we can do something about it..
  - Repair faulty mechanisms (e.g. genetic diseases)
  - Prevent/fight diseases by designing new drugs
  - Improve quality of life, slow/stop ageing
  - Improve crops, breeds, etc.

Modelling protein structures..

- What is known:
  - ~170,000,000 sequences (1$ each)
  - ~160,000 structures (100,000$ each)
- Gap widening. How do we bridge it?
- Figuring out the structures by looking at the strings?
- Hard, but Nature does it all the time.

Physics?

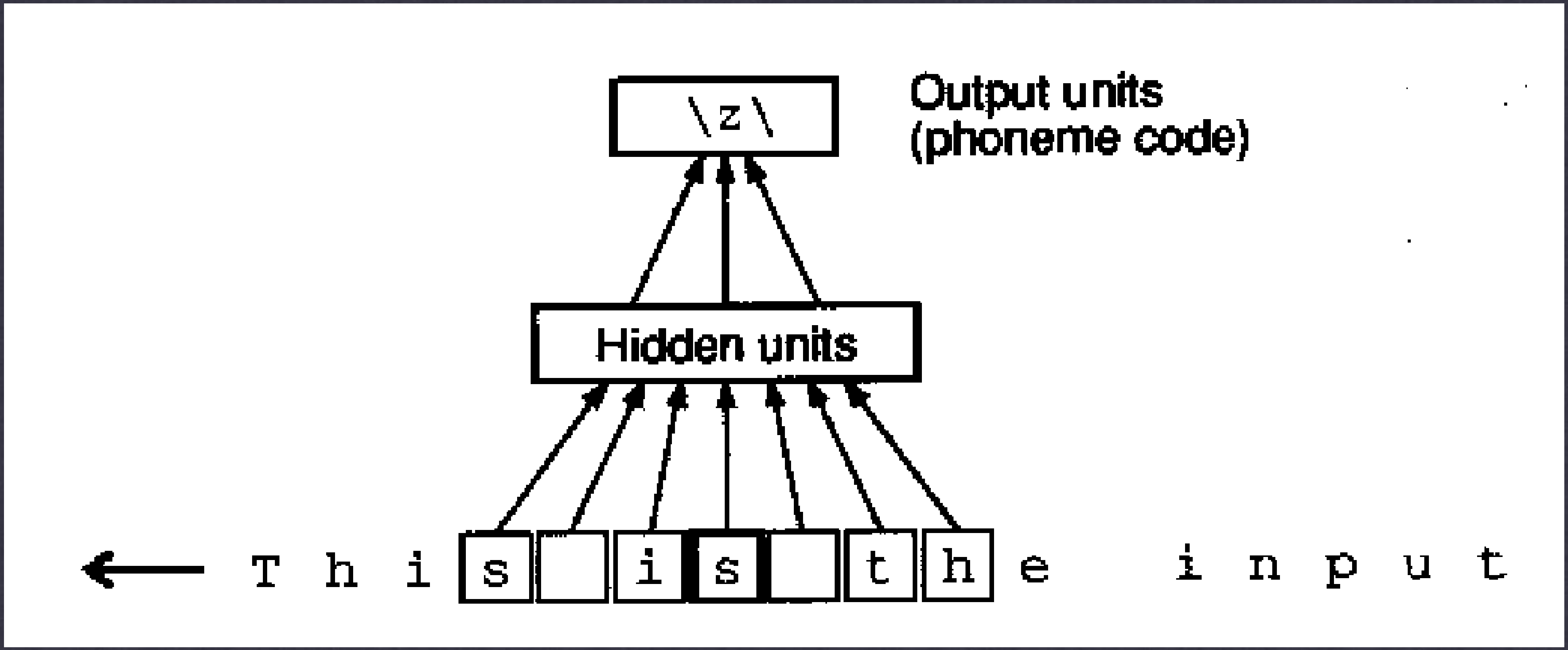
- Couldn’t we just solve the problem by physics?
- Folding@home. Hundreds of thousands of PS3 world-wide donate time to do this
- Result: 1 tiny protein every few months
- Too hard!

Machines that learn

- Computer programs that learn how to do stuff from examples; e.g., Artificial Neural Networks
- One well known scenario is: you have some examples of inputs (what is going on) and outputs (what you should do in response to it)
- They have been used to recognise faces, objects, speech, etc

Examples (1)

- Teaching a machine how to read aloud
- In English this is not trivial, as the sound of a letter depends on the surrounding letters (context)
- Examples to learn from: text (input) and corresponding sounds (output)
- Once the computer program has learned, it can be used to read



Sejnowski and Rosenberg, "NETtalk, a parallel network that learns to read aloud", Cognitive Science, 14, 179-211 (1986)

Reading English aloud

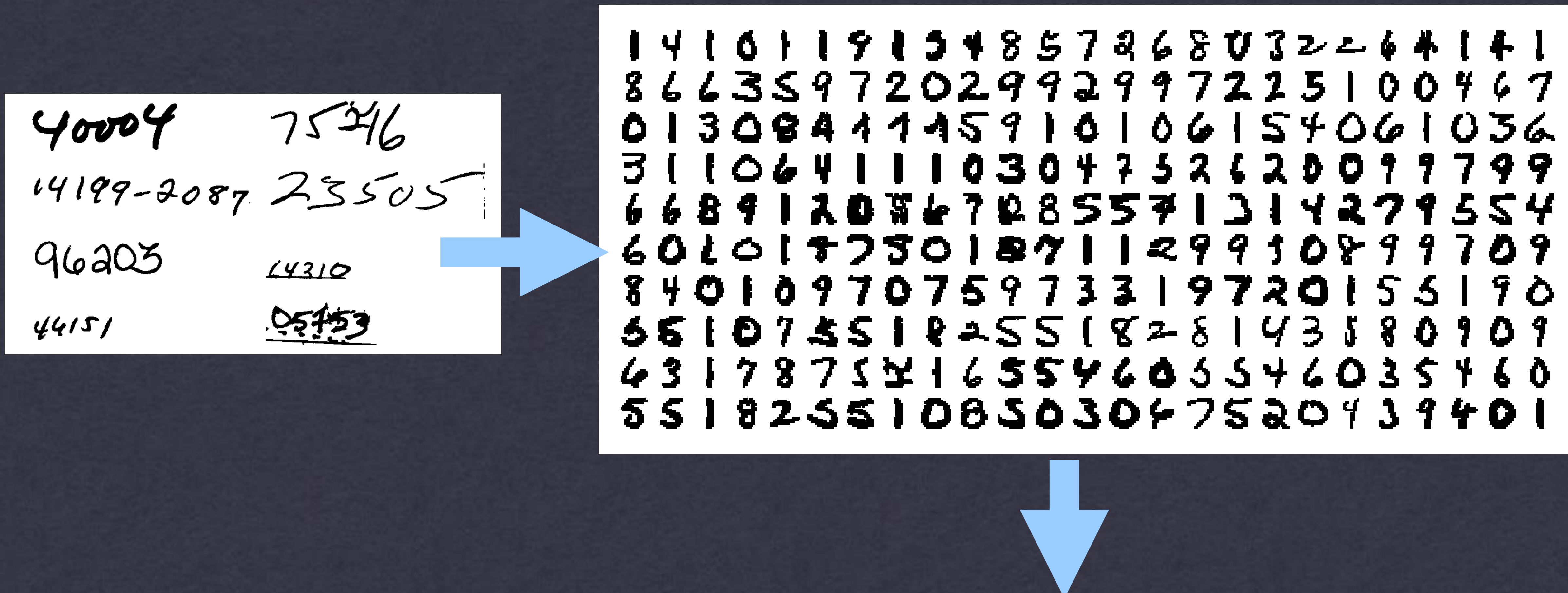
We do not give rules, we let our program learn them

Examples (2)

- Recognising (possibly hand-written) digits, to sort letters in a post office
- Not trivial as: people write in all sorts of different ways; even typed characters come in a dazzling array of different fonts
- Examples to learn from: images of digits (input) and corresponding values (output)
- Once the computer program has learned, it can be used to sort letters

Handwritten digit recognition

We do not give rules, we let our program learn them



0, 1, 2, 3, 4, 5, 6, 7, 8, 9

Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard and L. D. Jackel: Handwritten digit recognition with a back-propagation network, in Touretzky, David (Eds), Advances in Neural Information Processing Systems 2 (NIPS*89), Morgan Kaufman, Denver, CO, 1990

Machine Learning for videogame playing!

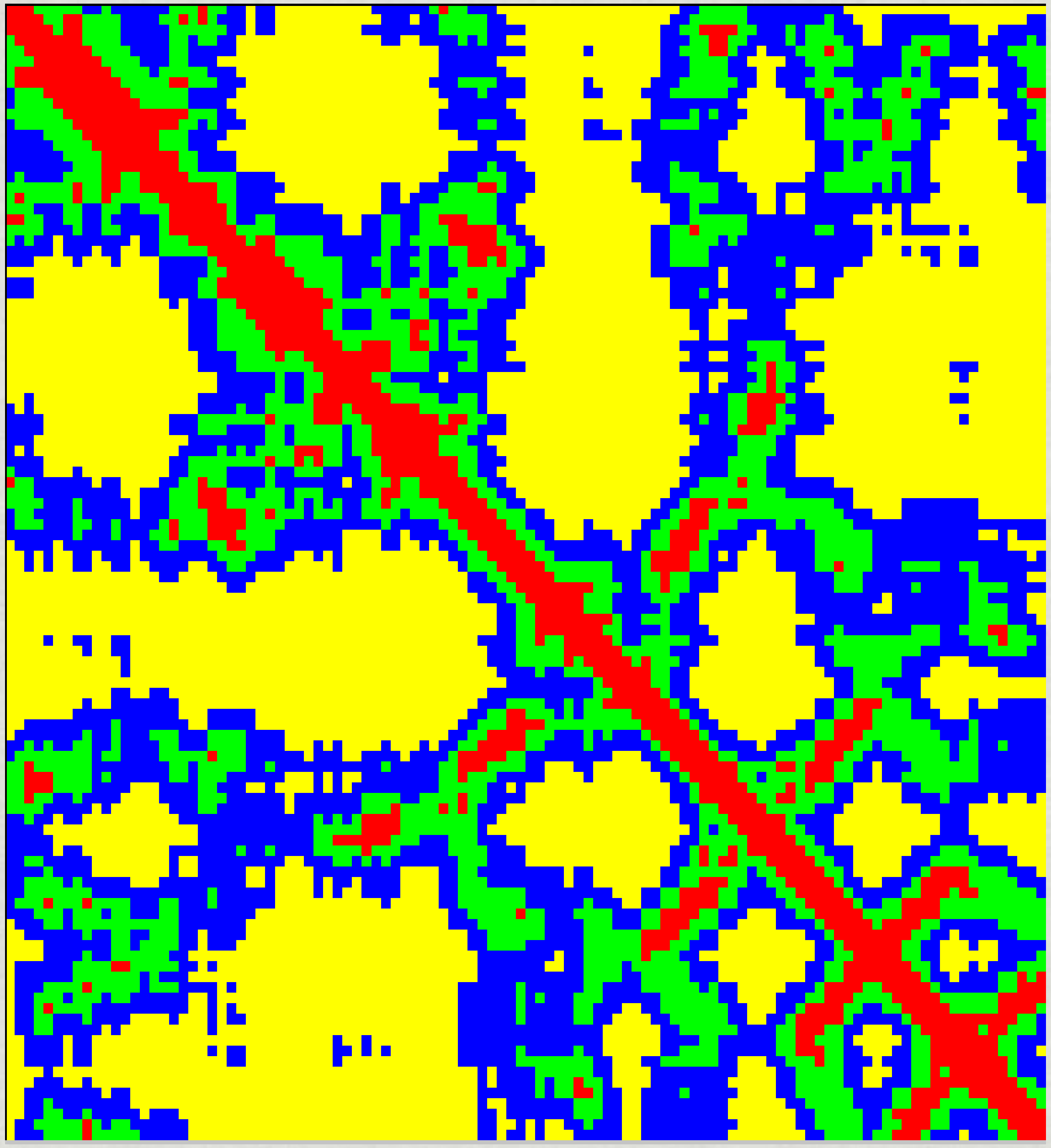
AlphaGo

- A computer program that cracked arguably the hardest board game.
- It is now out of reach of human players.
- Impressively, the latest (and best) version, learned entirely from self play.

Machine learning and proteins

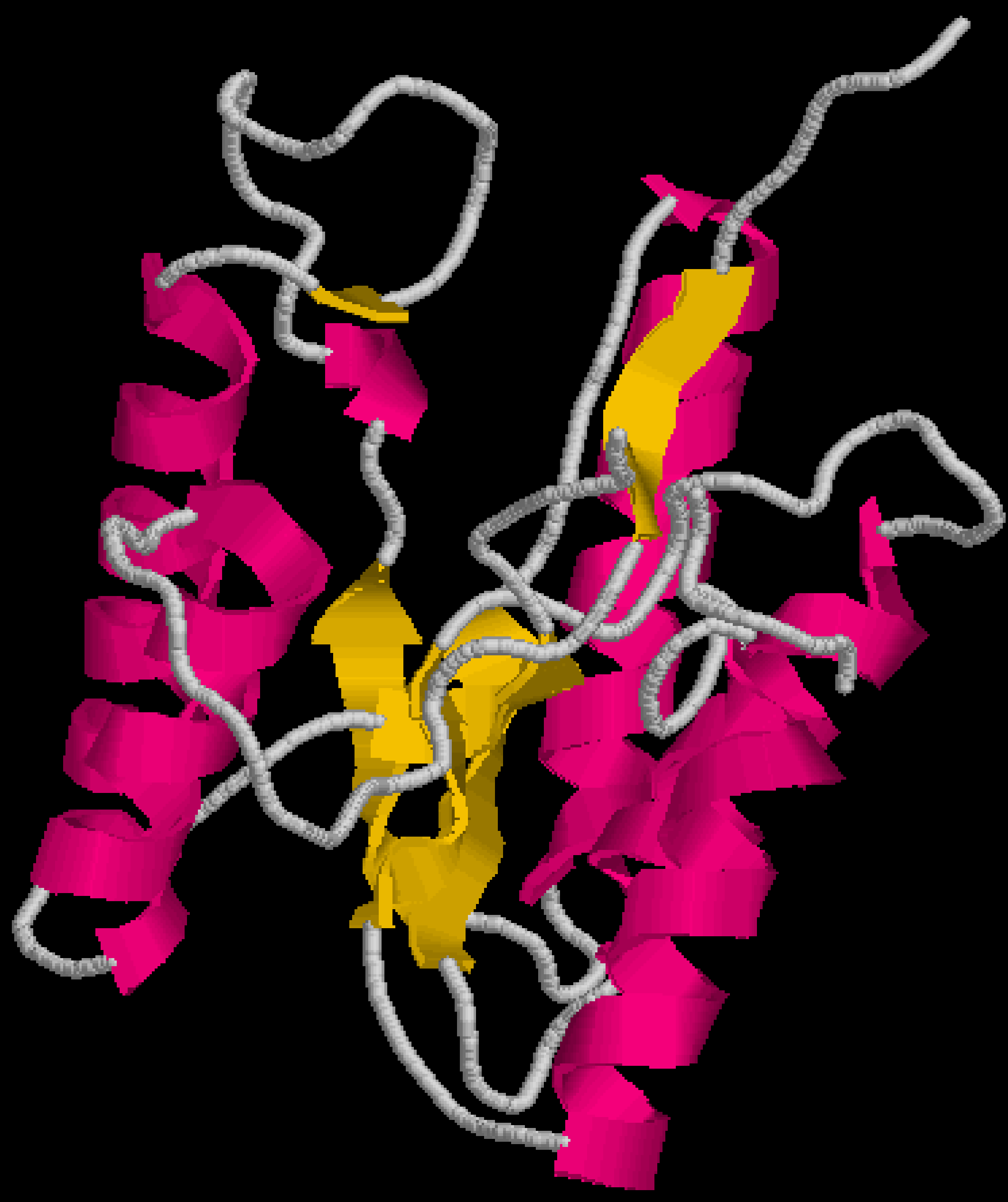
- Can be adapted to the problem of figuring out the shape from the string
- What is going on: the string
- What we should do about it: the shape
- Feed the computer program with enough examples of string/shape, go have a holiday, then come back and check if it has learned

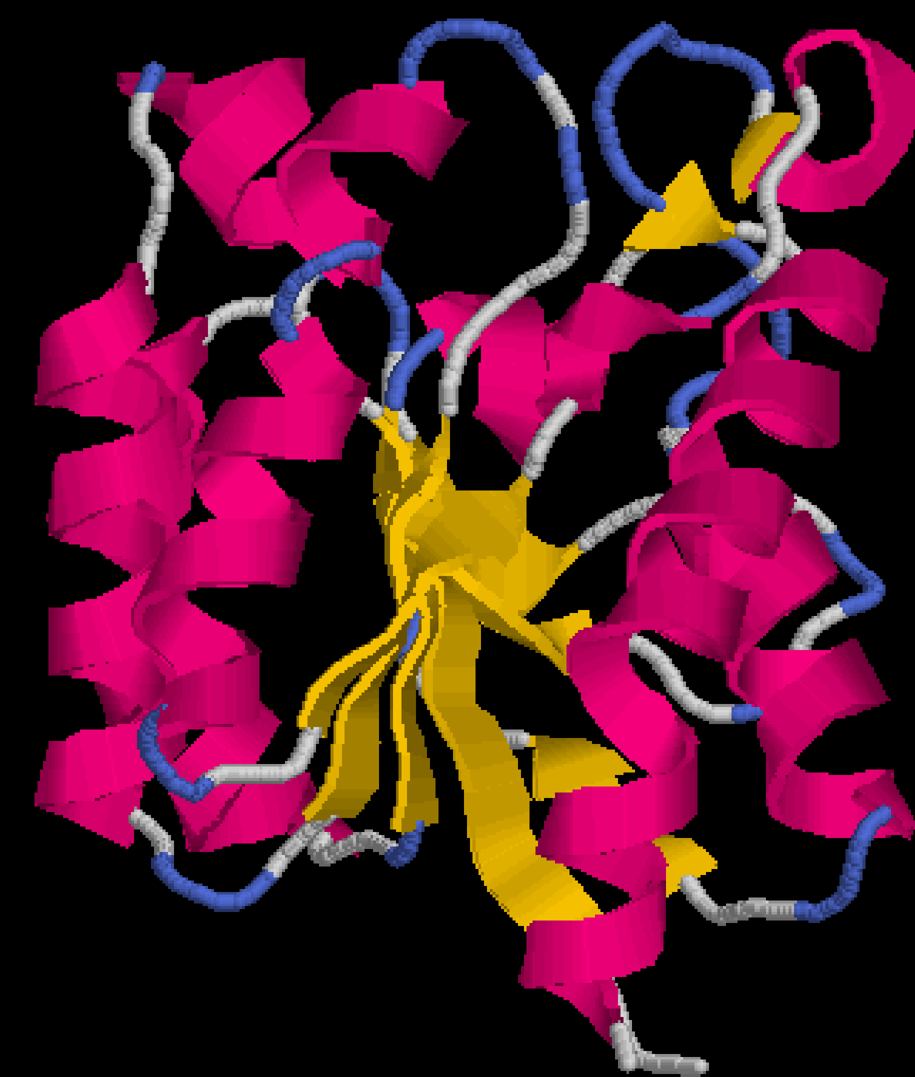
Ignoring translation and rotation: distance maps



Complexity?

- We need to train a machine learning program on known protein examples:
  - 1 training takes 2-3 months on 1 state-of-the-art computer;
  - a full system costs ~50,000 hours (that's years, but one can use many computers at the same time)
- Once the system has learned, a prediction doesn't take much time (minutes)





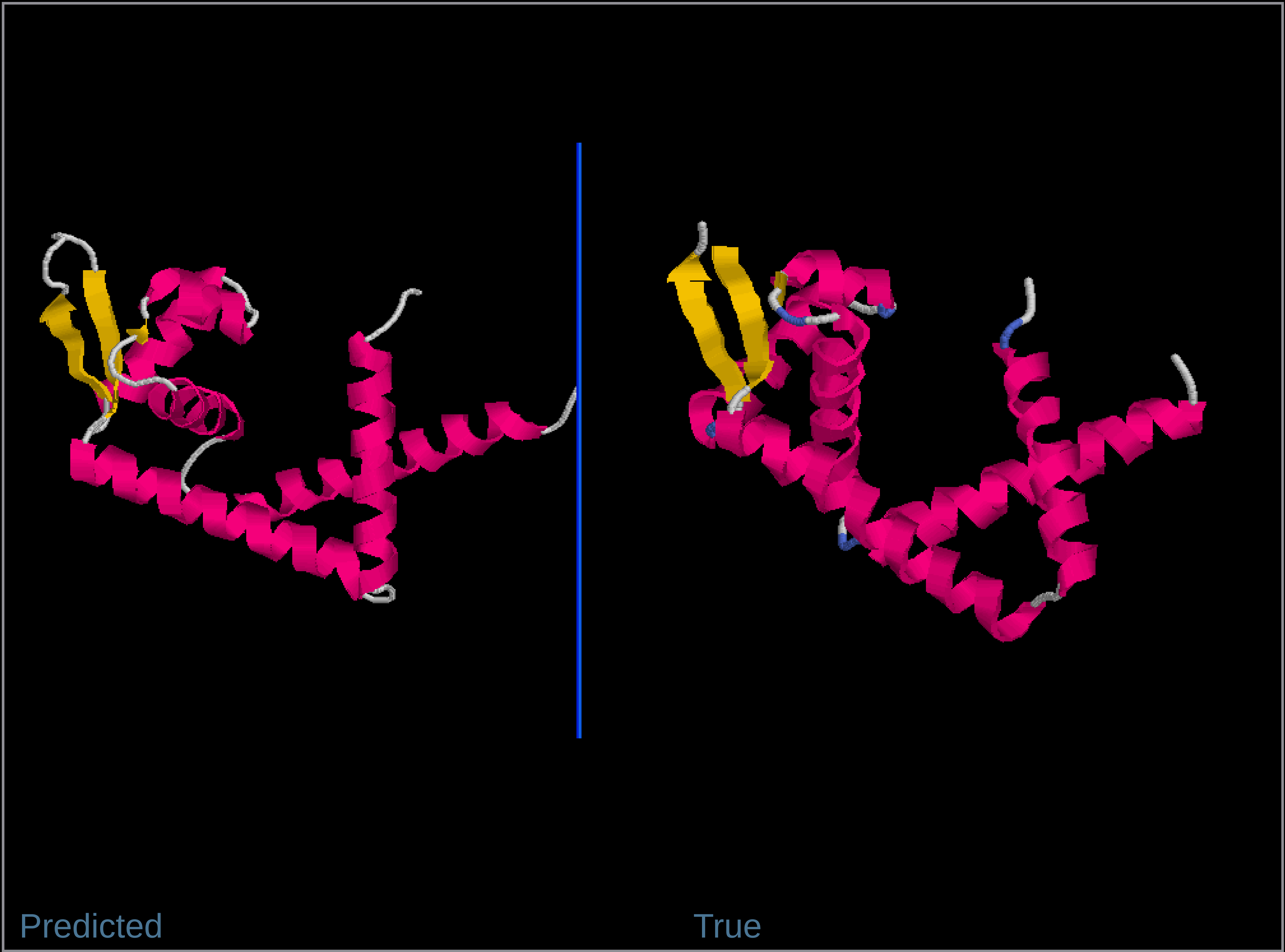
Predicted

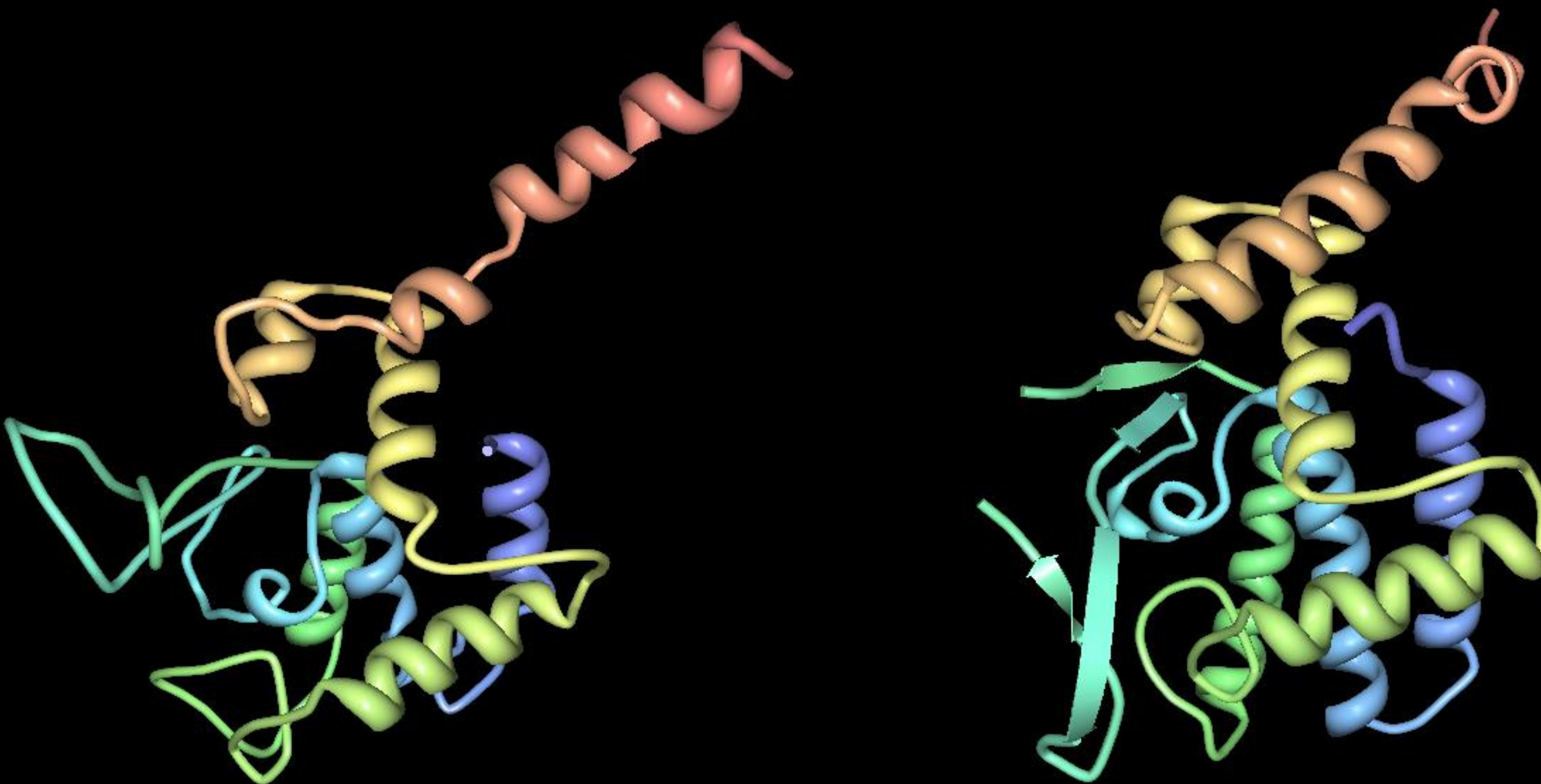
True

Predicted

True


Predicted
True

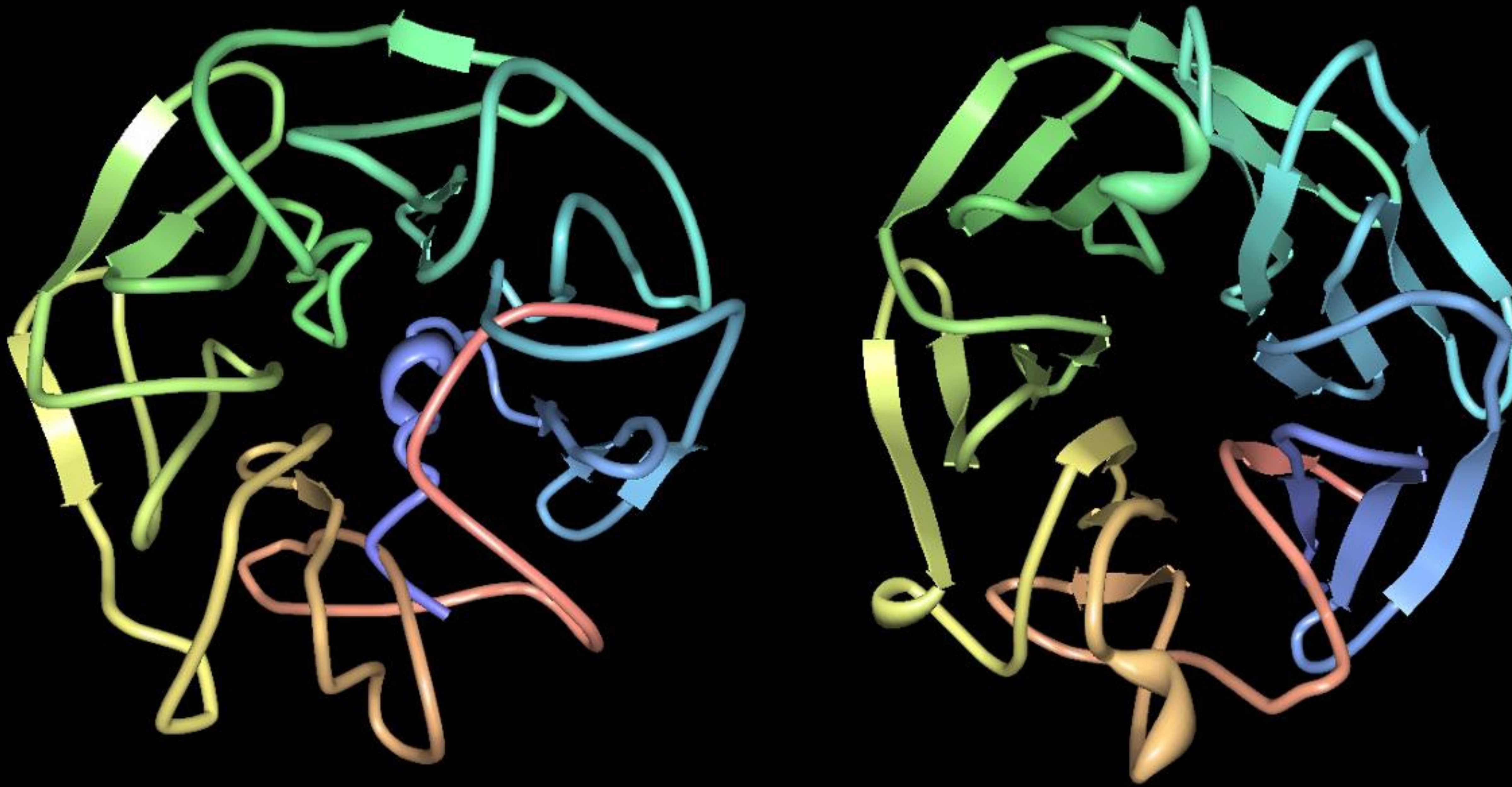

Predicted
True



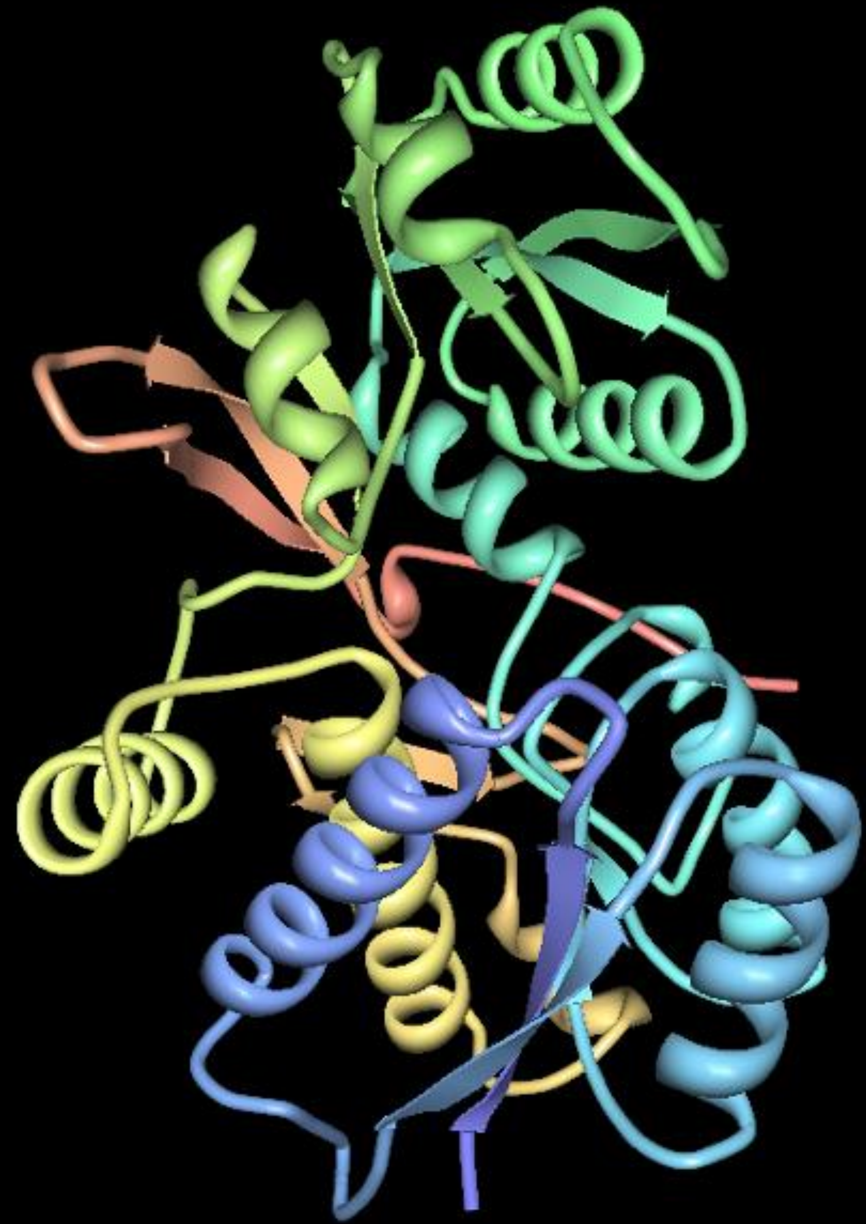
CASP Target: T0623, PDB ID: 3NKH
21 Blast templates, SeqID: 0.18

CASP Target: T0613, PDB ID: 3OBI
82 Blast templates, SeqID: 0.49

CASP Target: T0548, PDB ID: 3NNQ
no Blast templates, SeqID: 0.04



CASP Target: T0558, PDB ID: 3NO2
no Blast templates, SeqID: 0.09



CASP Target: T0528, PDB ID: 3N0X
55 Blast templates, SeqID: 0.21

Mission accomplished?

- Closer now than we were 10 or 20 years ago
- 20 years ago we could produce good solutions in ~40% of cases
- Nowadays this is close to 80%
- The hard core of the problem is shrinking
- Databases with millions of good predictions are freely available
- COMPUTERS HAVE MADE HUGE DISCOVERIES!

Hard core

- It is shrinking, but not trending to zero
- For instance synthetic proteins:
  - They do not exist in nature
  - May be engineered to substitute other proteins, to block proteins, to do some jobs more efficiently, to function as materials, etc.
  - We already know that they will never be easy to deal with
- Hundreds of research groups world-wide are still working on the problem. No shrinking!

Take Home Message

- COMPUTERS CAN BE USED TO DISCOVER. MORE: COMPUTERS CAN DISCOVER
- A massive amount of information about how life works, out there
- This information is raw, messy, we know very little about it:
  - It’s like having an encyclopedia with everything in it, but we don’t know the language it’s written in.
  - We can use intelligent computer programs to decipher the language.