Machine learning methods for quaternary structure validation

Or, how I learned to stop worrying and love the bond

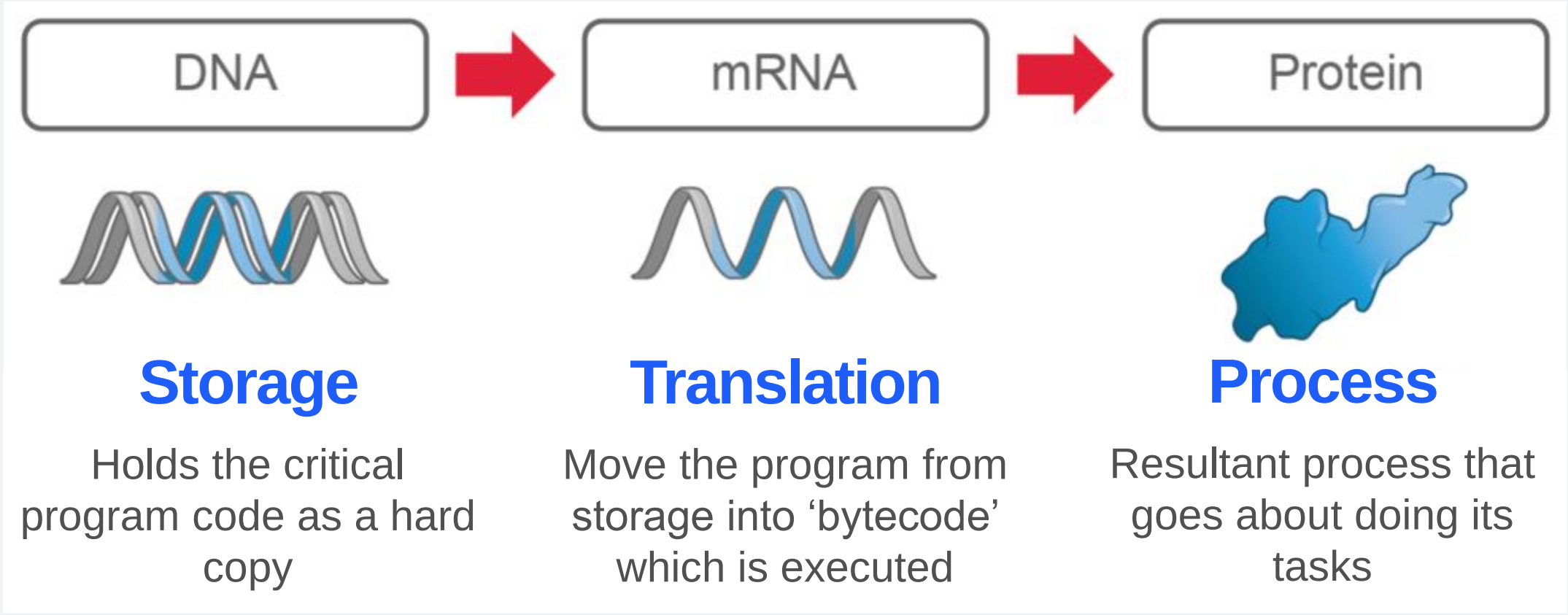
Nick Whyatt, STFC-UKRI

17/09/2024

CCP-EM

Science and Technology Facilities Council

A computational analogy...

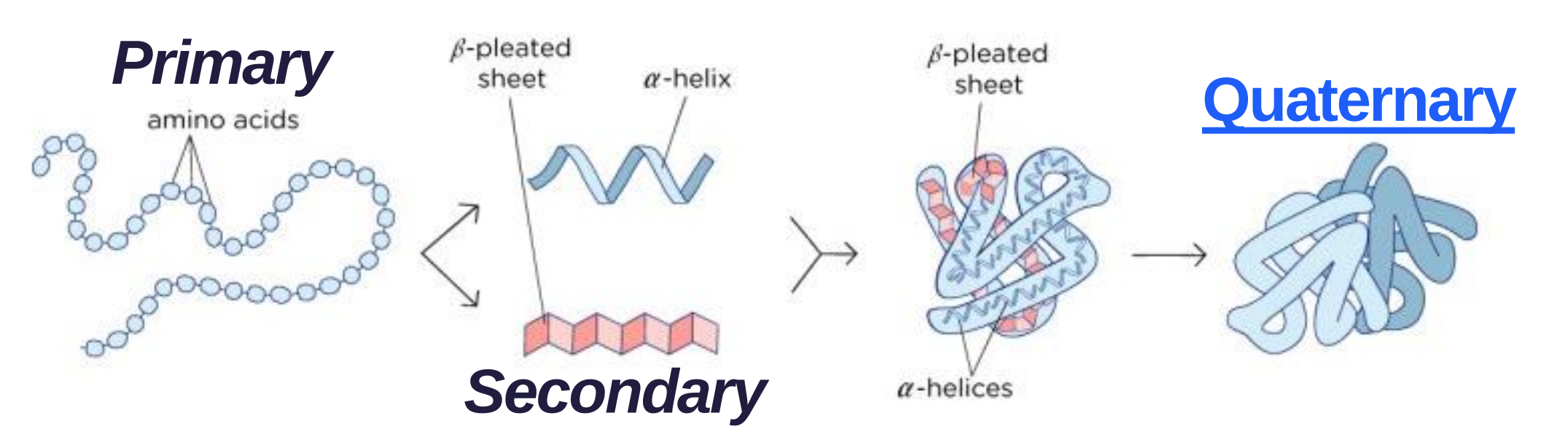


Science and Technology Facilities Council

Quaternary Structure of Proteins

Protein structure is defined by amino acids...

... but classical statistics on amino acid sequences cannot predict structure



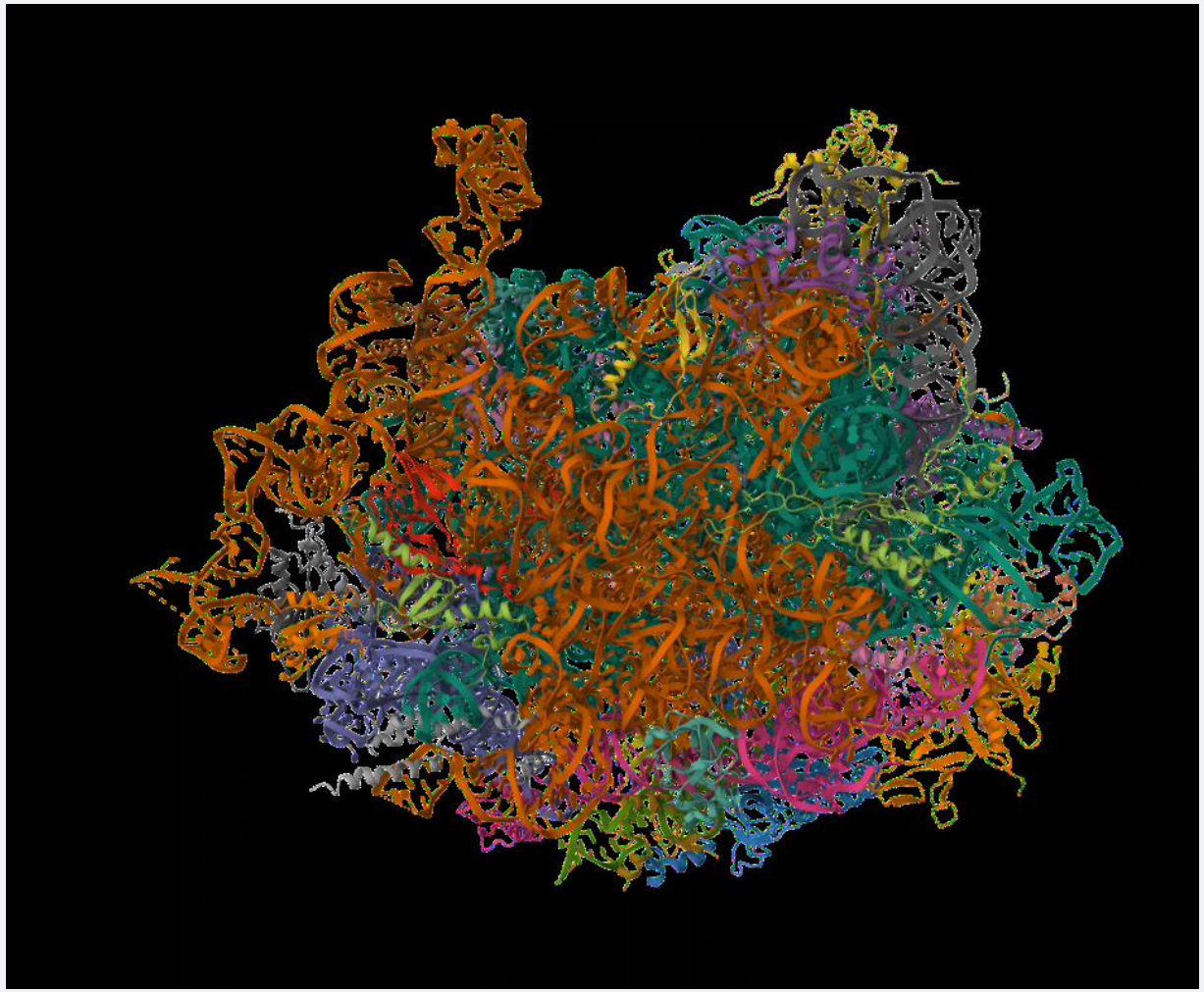
Science and Technology Facilities Council

cryo-EM can determine molecular structure

(to ± 3 angstrom)

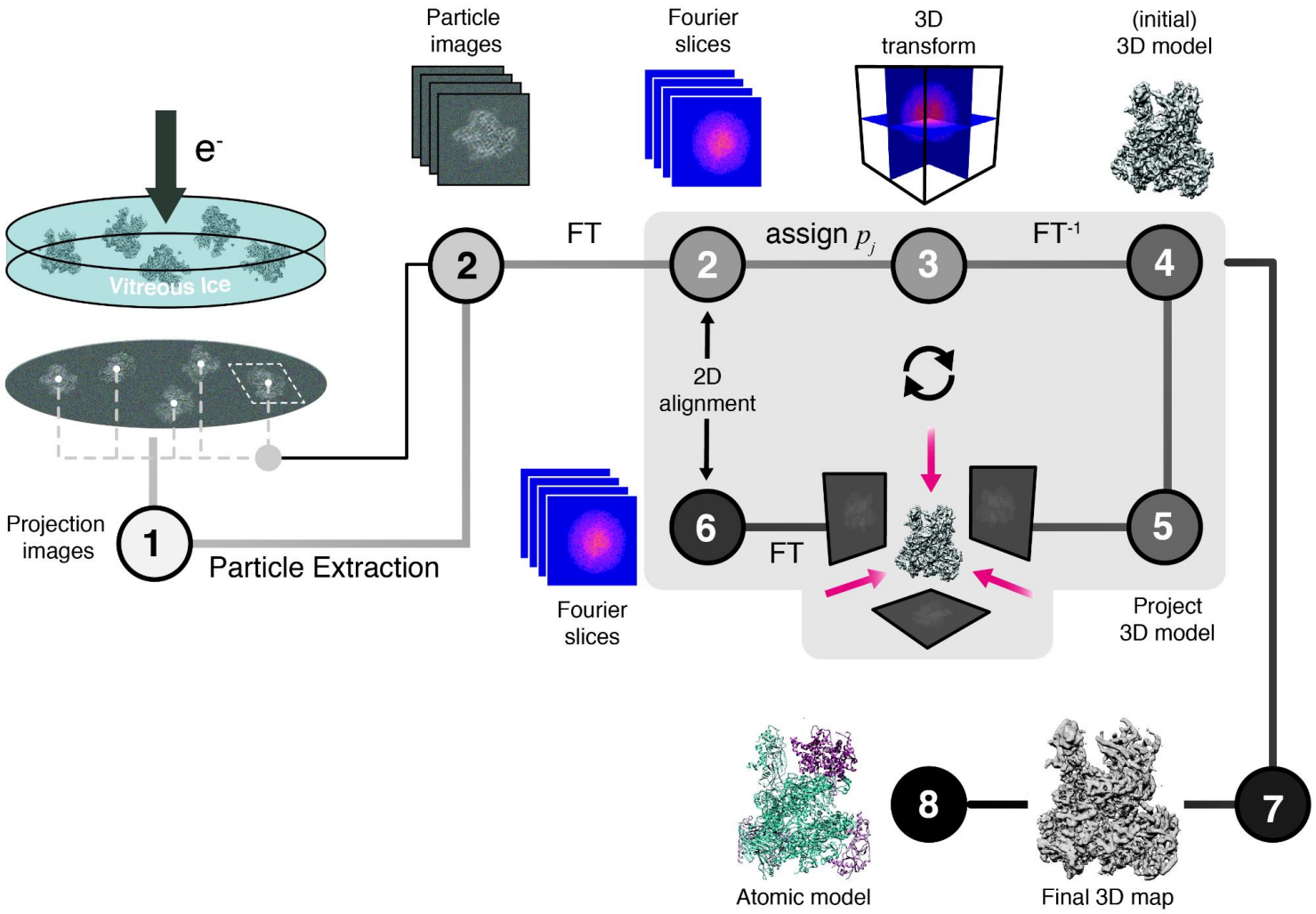
Structure defines function

- Increasingly accurate molecular structure determination gives us new insight into their mechanisms
- Resolution is not a global attribute – *areas of importance* are often **significantly lower resolution**
- Complex method: error accumulates quickly, and can propagate throughout the model



Science and Technology Facilities Council

The cryo-EM single particle workflow



Jakobi. Europhysics News. (2020)

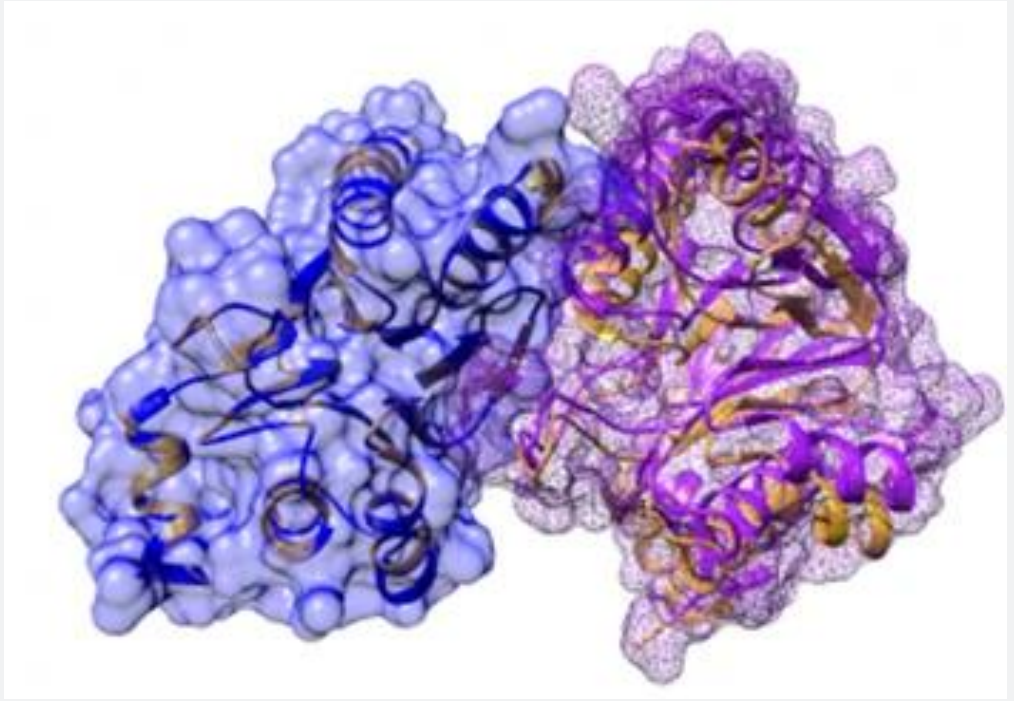
Science and Technology Facilities Council

TU Delft

Interfaces between chains propagate error

Here's some ways they go wrong!

- fitted models are **usually built sequentially**, i.e. one at a time
- **segmentation techniques are not accurate enough** to identify boundaries between the subunits;
- building the **model of only one protomer** and **applying symmetry operations**; and
- integrating models of subunits built in **maps reconstructed by refinement focused on certain segment(s)** of the macromolecule



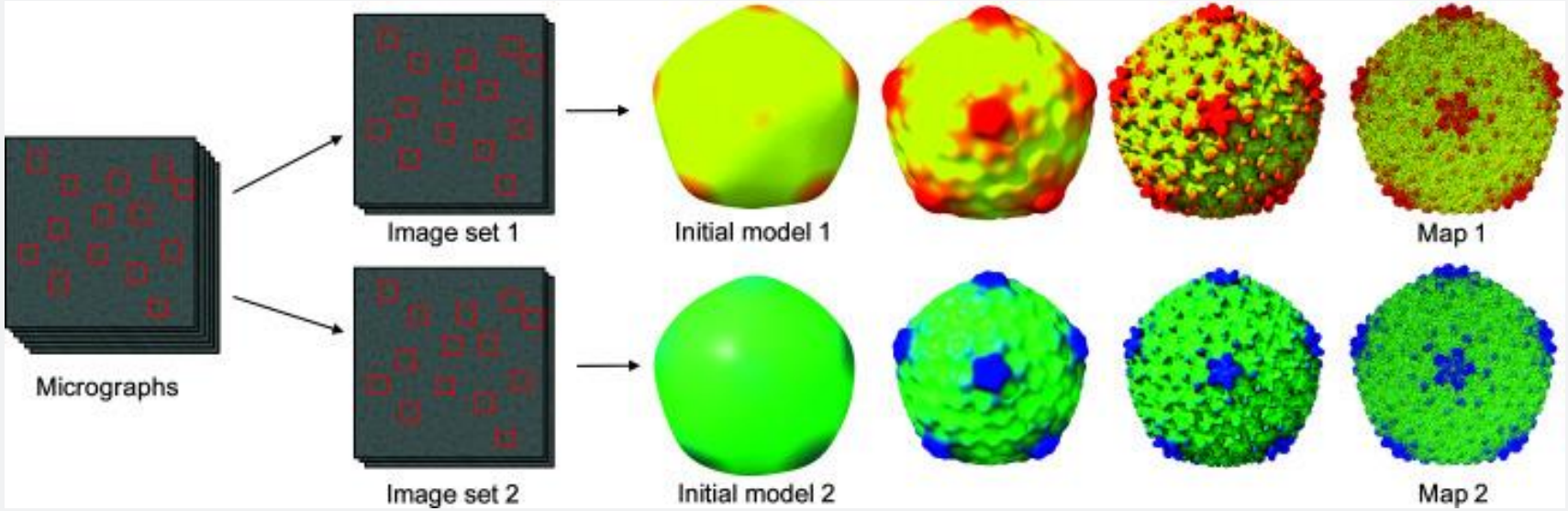
Malhotra, S., Joseph, A.P., Thiyagalingam, J. et al. Assessment of protein–protein interfaces in cryo-EM derived assemblies. Nat Commun 12, 3399 (2021). https://doi.org/10.1038/s41467-021-23692-x

Science and Technology Facilities Council

We can prevent bad models with metrics

We have different metrics for different targets

- **Global** measurements: cross-correlation coefficient, mutual information
- **Local** metrics target specific areas of poor model fit: local mutual information, TEMPy local scores, segment based mander's overlap coefficient, segment based cross-correlation, Q-scores, EMRinger...
- **Geometric** models, such as MolProbity and CaBLAM
- But **none** of these metrics specifically target **quaternary structure**



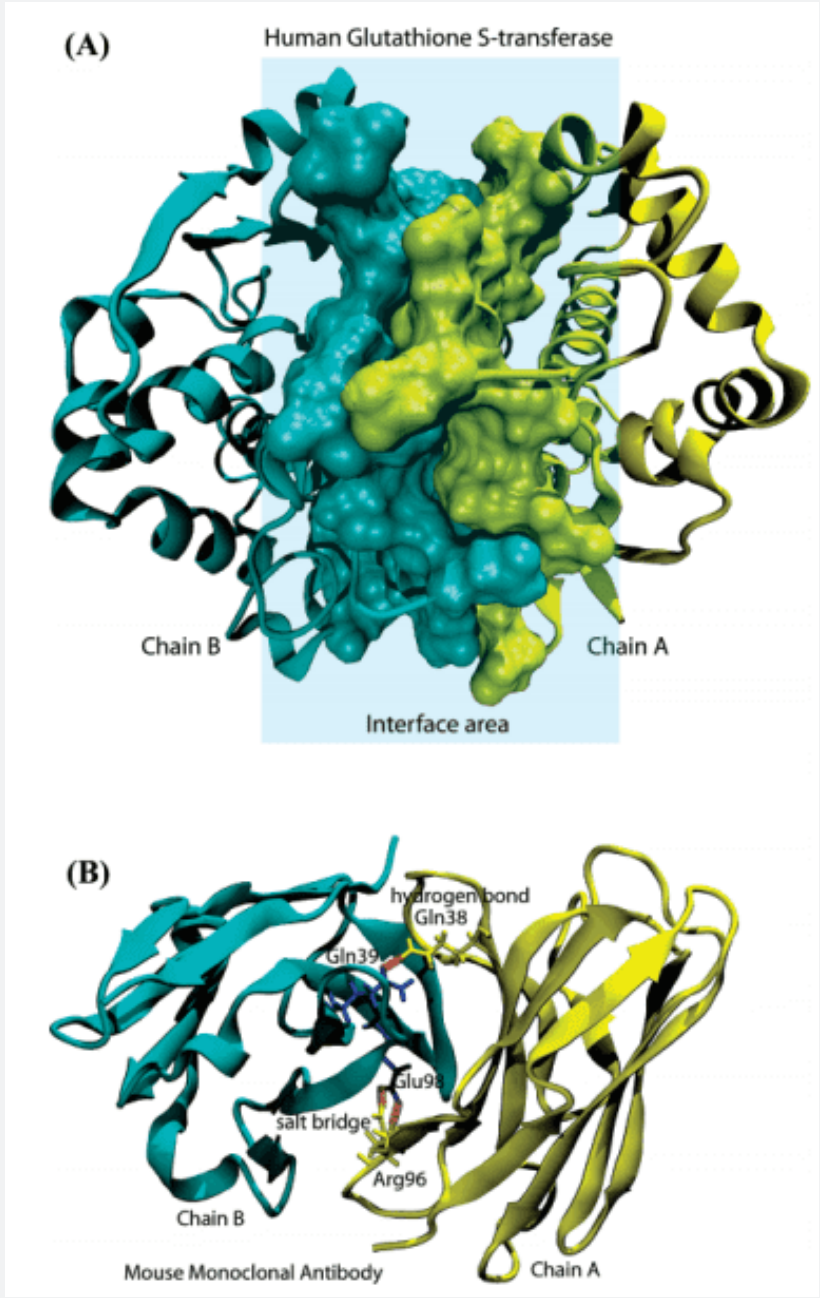
Science and Technology Facilities Council

Classical method: PI-score

(Protein-protein interface assessment)

The task:

- Given two reasonably conventional connected protein chains, **assess if their interface is suitable or not** as a single float ‘score’
- Essentially, a **binary classification task**
- These ‘chains’ are connected atomic structures in a static, 3D representation
- We assume an interface to exist if atoms are closer than is electrostatically possible otherwise – typically a cutoff of 4-7 angstrom, solved classically
- Input is a cube of size N^3 centred on the interface



Science and Technology Facilities Council

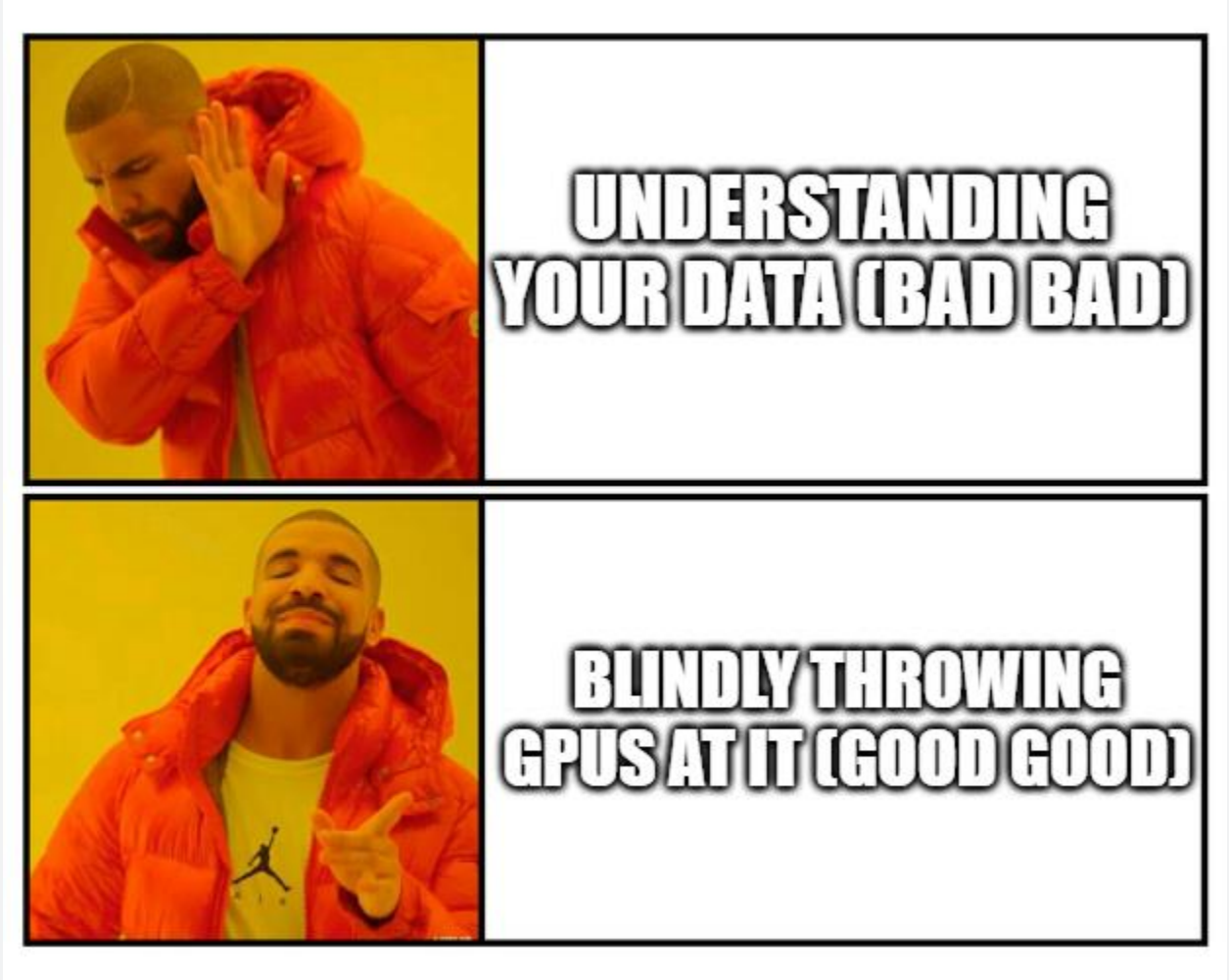
Classical method: PI-score

- Requires calculation of twelve features of *varying complexity*...
- ... and an additional augmented dataset, which requires a complex search...
- ... taking **days to produce** all models with features using current scripts!
- Uses **variety of third-party tools**, many of which are in a *poor state of maintenance*
- **Hamstrung by docking algorithm** needing constant complex directory changes (???)
- Entirely **incompatible with modern file format** (mmCIF (though everything is, lol))

Reasonable performance with **86% validation accuracy!**
But still **struggles with cryo-EM targets**, as only trained on X-ray crystallography – dataset has just under **4,000 interfaces**

Is there a better way?

It’s time to machine learn



Science and Technology Facilities Council

(D)PI-score

Enhance the dataset

- Around **10,000 X-ray structures** under **2.5 angstrom precision**
- **2,000ish EM structures** under **3 angstrom**
- Calculate which have valid **PPIs**
- Find internal interface similarities in *iALIGN*
- Generate docked models with *ZDOCK* on **sufficiently dissimilar interfaces**
- **PD2** ‘near-native’ (green), **ND** wildly inaccurate (pink)
- Remove some structures without sufficiently long chains or too many chains (30+)

Whyatt N., unpublished work (2024)

Science and Technology Facilities Council

(D)PI-score

Dive into the data...

- A .PDB or PDBx/mmCIF file is an extremely complicated mess of **atomic 3D coordinates, charges, atomic labels, residue labels, chain labels,** and a LOT of meta information
- To distil this down to the essentials, we **take the atomic positions** of all the atoms of four key elements: *Oxygen, Nitrogen, Carbon, Sulphur*
- We separate each element as a feature, where each feature is a list of x, y, z atomic coordinates for an atom of that type
- We form grids of size N^3, where N=**32 angstroms**
- We **centre the grid on the mean coordinate of a given interface** (maintaining coherency with the structure file)

Science and Technology Facilities Council

Whyatt N., unpublished work (2024)

(D)PI-score

Parallelise *everything*

- By creating docked models of our new set of around **12,000 interfaces**, this would take us minimum *two weeks* of continuous processing...
- In parallel? Roughly **10 times speedup** – constrained through spurious file creation due to docking algorithm
- Interface similarity assessments – **linear speedup**, plus a little extra due to optimisations in file writes
- All tied to automatic, easy(ish) scripts to use
- Also combines batch structure downloader

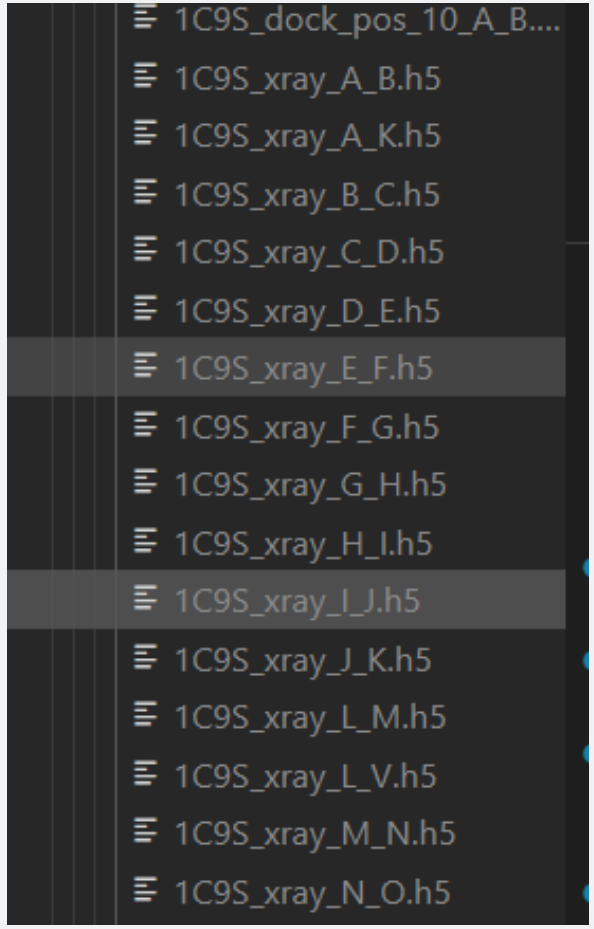
Whyatt N., unpublished work (2024)

Science and Technology Facilities Council

(D)PI-score

Network Design

- Various methods considered: graph, Euclidean, graph into CNN...
- Current iteration is a plain old 3D CNN with heavy use of residuals – deep network over wide
- Data augmentation is critical for robustness and accuracy – rotation, cropping, etc
- Dataset has been reduced very effectively – multiple layers to compare atoms to their counterparts close and far
- Ship of Theseus approach – gradually replace dataloader, loss, dataset, etc...



Whyatt N., unpublished work (2024)

Science and Technology Facilities Council

(D)PI-score

Tentative evaluation

- Our network scores an **87.6%** (+/- 1.9%) on 5-fold validation, an improvement of **1.6%**
  - ... but it can do so **consistently on a wider domain** (cryo-EM and X-ray data, as opposed to just X-ray)
  - it can do so an **order of magnitude faster** – O(1 second) versus 2-3 minutes, per interface (slower/running individual scripts for tasks)
  - it is easily **retrainable for new tasks** (biological vs crystal contacts) or updated with new structures
  - it can **run on conventional hardware**, and is very easy to set up :)
- ... and currently doesn't calculate any features save atomic labelling – next step, optimising with more (easy to calculate) features

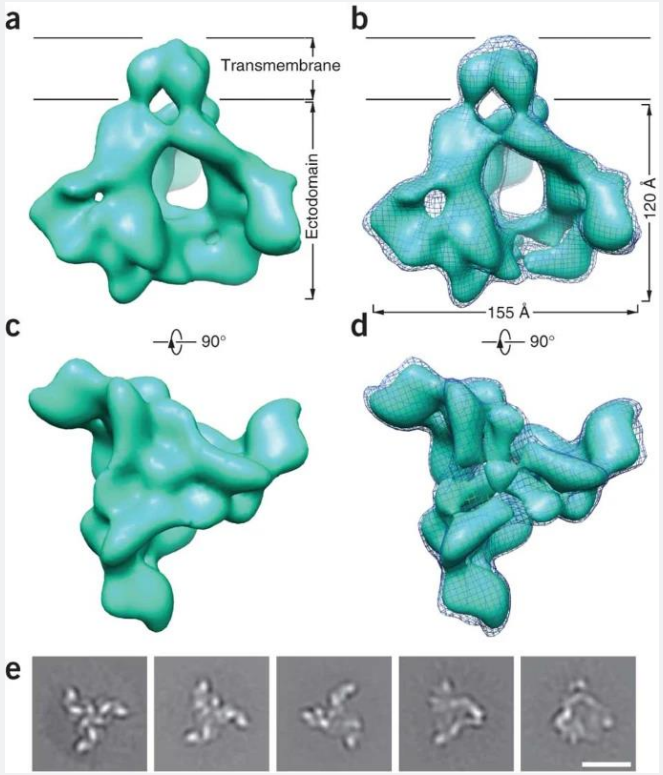
Whyatt N., unpublished work (2024)

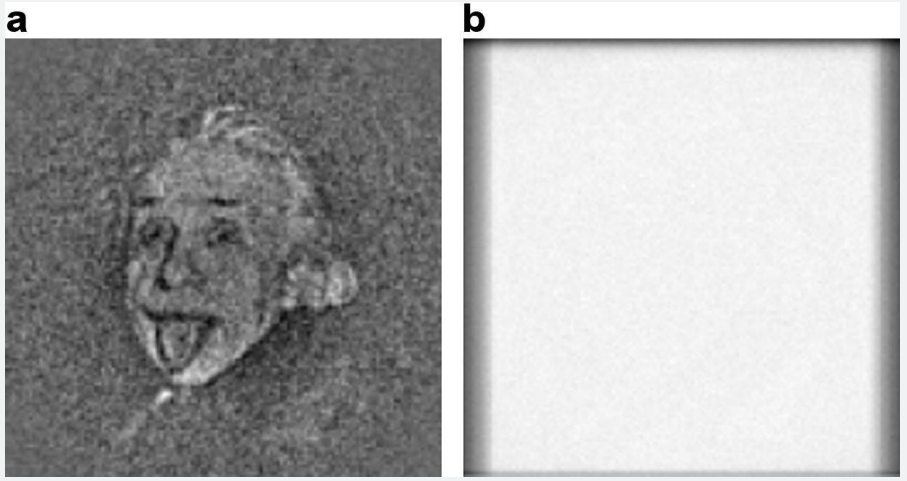
Thank you!

Case study: Mao et. al. 2012

Alternatively, ‘einstein from noise’

- “... in which the experimenter honestly believes they have recorded images of their particles, whereas in reality, most if not all of their data consist of pure noise.”
- “Selection of particles using cross-correlation methods can then lead to 3D maps that resemble the model used in the initial selection and provide the illusion of progress.” (Henderson, 2013)
- But the model was an **HIV membrane trimer**...





Science and Technology Facilities Council