University of Sheffield

JADE@Sheffield

How JADE is making an impact

to research at the University of Sheffield

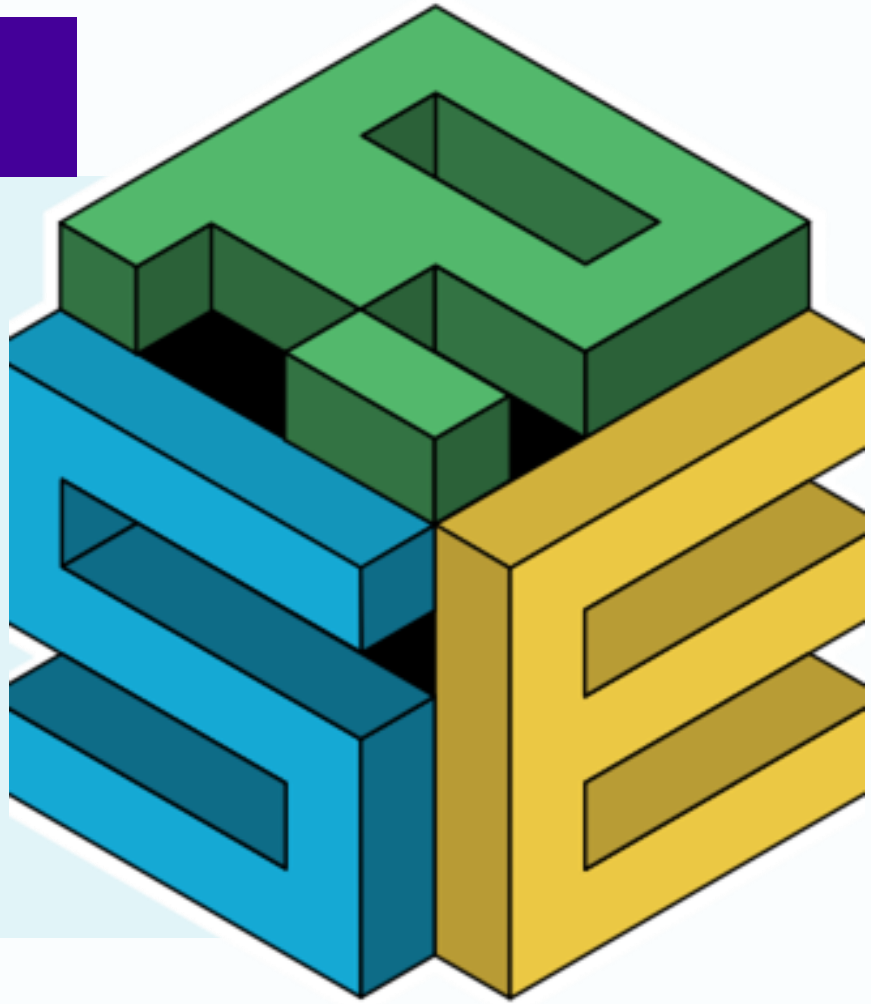
Twin Karmakharm 29/09/2023

Twin Karmakharm

Senior Research Software Engineer

RSE Contact and Support for JADE at University of Sheffield

https://rse.shef.ac.uk



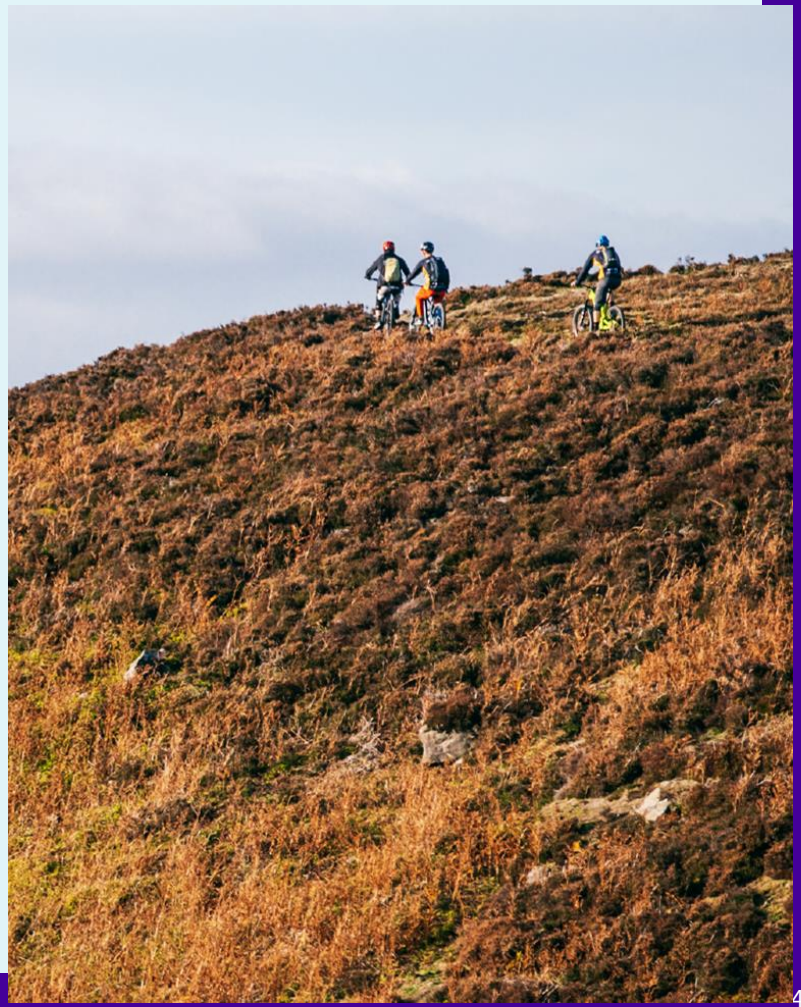
Overview

- JADE@Sheffield
- Case study 1:
  DNA Damage Competes with Sequence to Pin a Plectoneme
- Case study 2:
  NLP Research at Nikos' Lab
- Case study 3:
  Planar cell polarity components folding prediction
- Conclusion

University of Sheffield

JADE@Sheffield

University of Sheffield

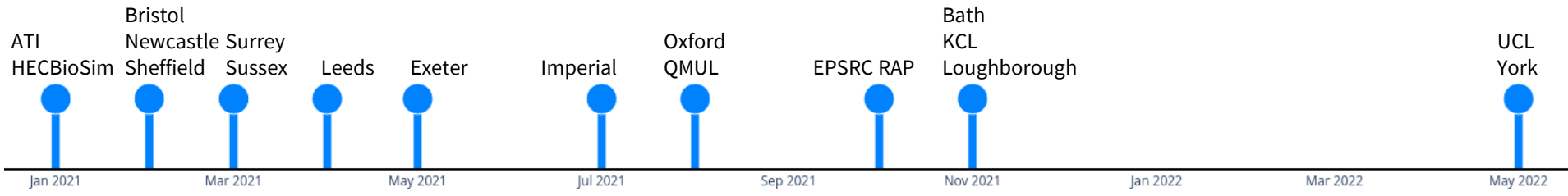


Some context:
Sheffield's Generally Accessible GPUs (Local)

- ShARC - 16 Nvidia K80 (Will be decommissioned 30th Nov)
- Bessemer - 4 Nvidia V100 (since 2019)
- Stanage - 64 Nvidia A100, 12 Nvidia H100 (2023)
- V100 Available as VMs
- Required buying dedicated workstation or dedicated server nodes for additional GPUs
  - e.g. DGX-1 (8xNvidia P100) and 40x Nvidia V100 for Computer Science

JADE 2 First usage

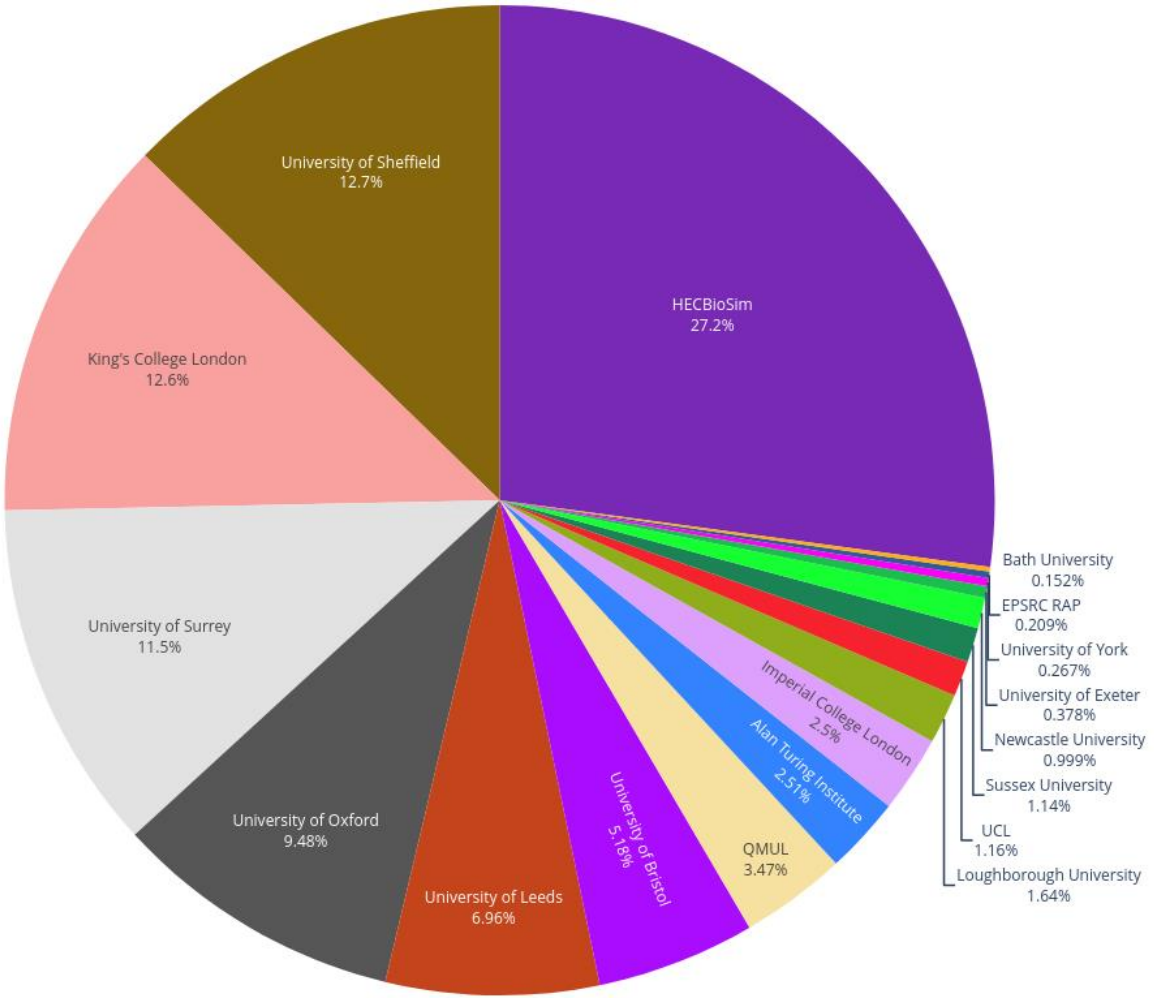
- Delayed start due to COVID
- Staggered access to the system
- Sheffield was one of the first to join (existing JADE 1 partner)



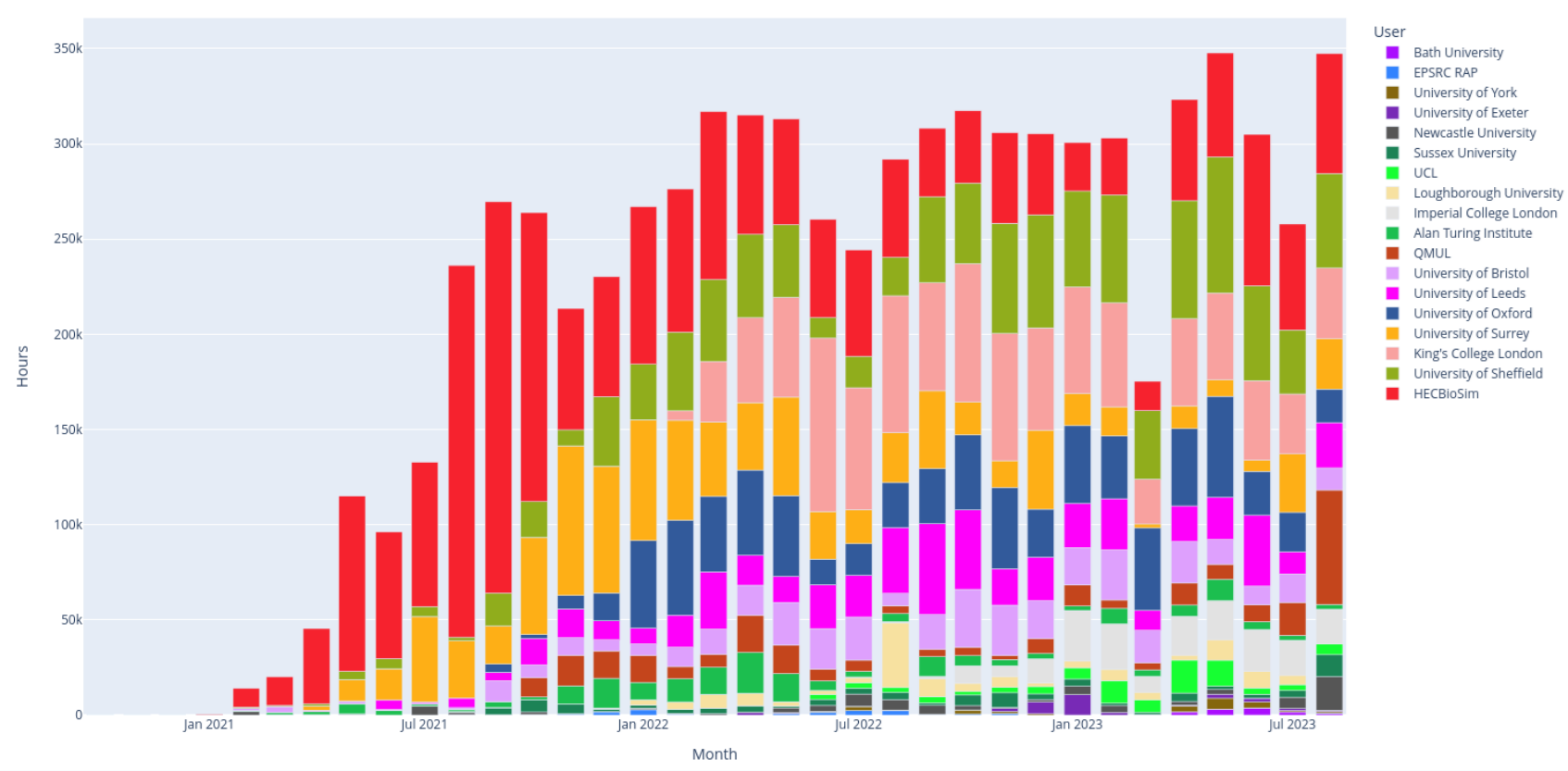
Month of first usage

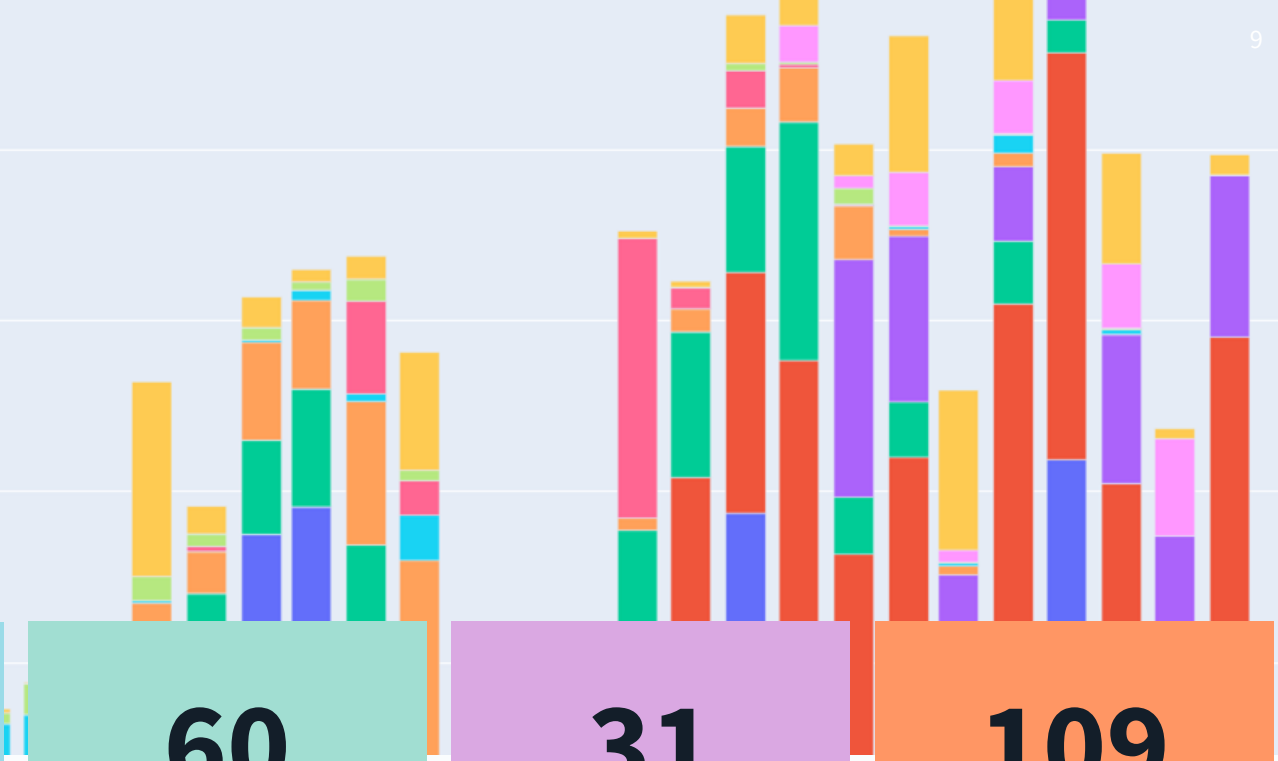
JADE 2 Usage by partners

1. HECBioSim
2. **Sheffield**
3. KCL
4. Surrey
5. Oxford
6. Leeds
7. Bristol
8. QMUL
9. ATI
10. Imperial
11. Loughborough
12. UCL
13. Sussex
14. Newcastle
15. Exeter
16. York
17. EPSRC RAP
18. Bath



Usage over time: JADE 2 GPU Hours/month





Sheffield's usage of JADE 2

60

total projects

31

active users
(past 6 months)

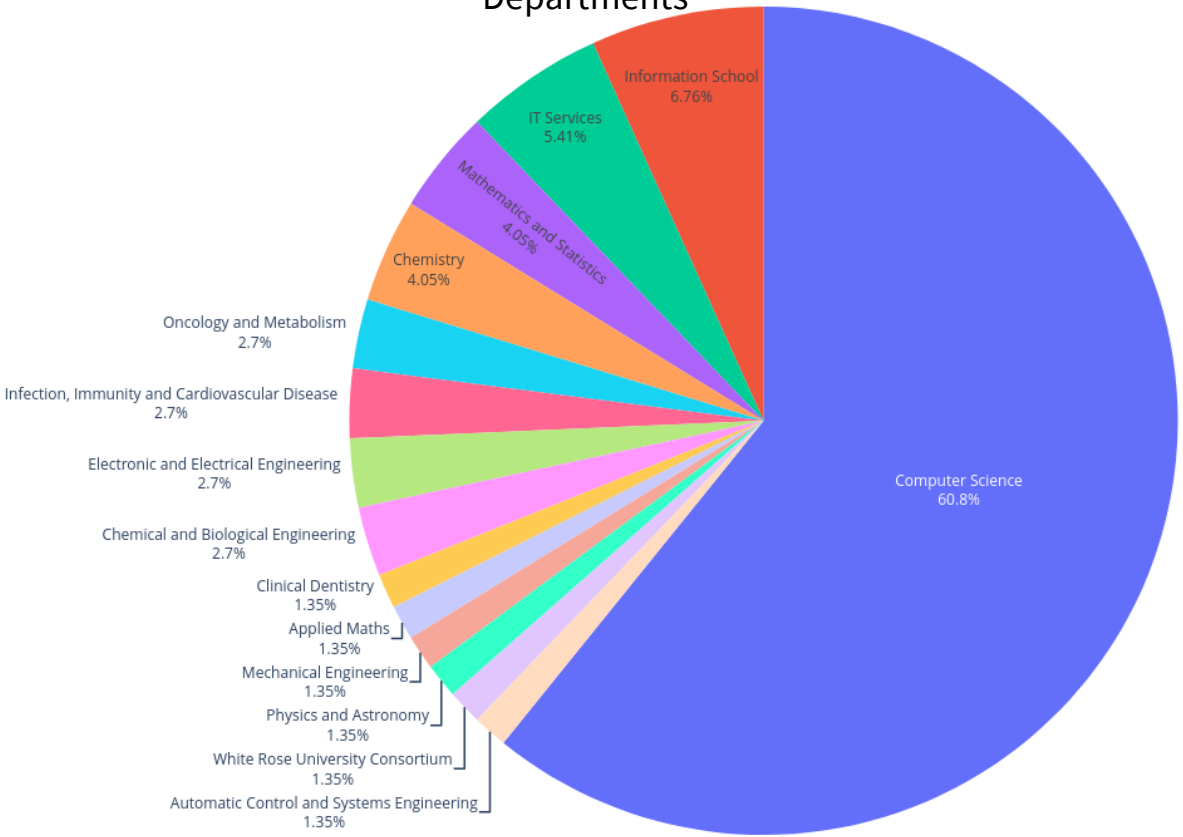
109

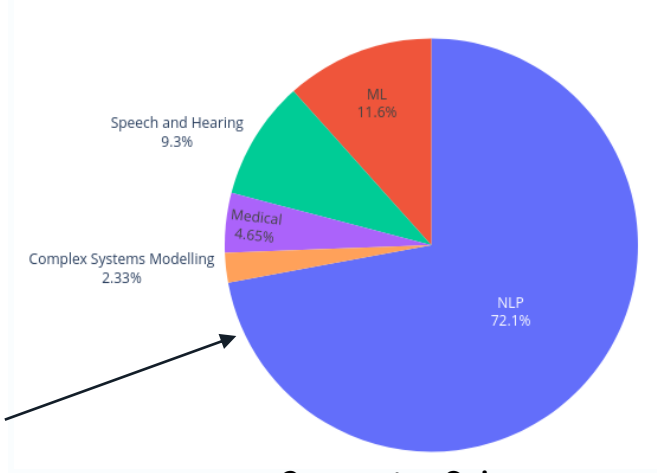
years of GPU time

(957,172 GPU hours)

JADE 2 Users by department

Departments





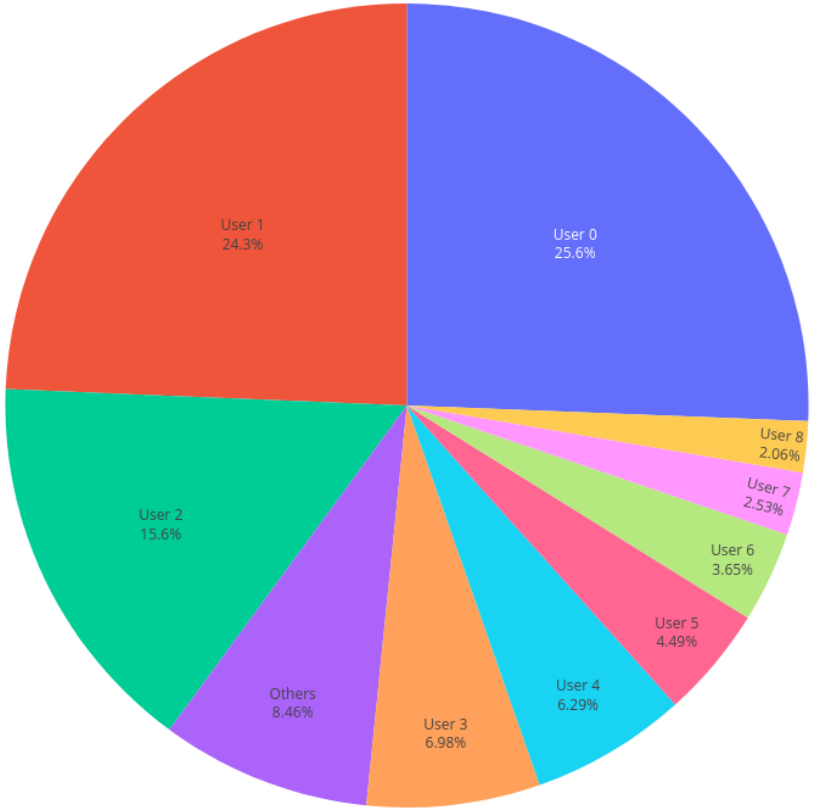
Computer Science Research Groups

JADE 2 Percentage of GPU hours by top 9 users + others

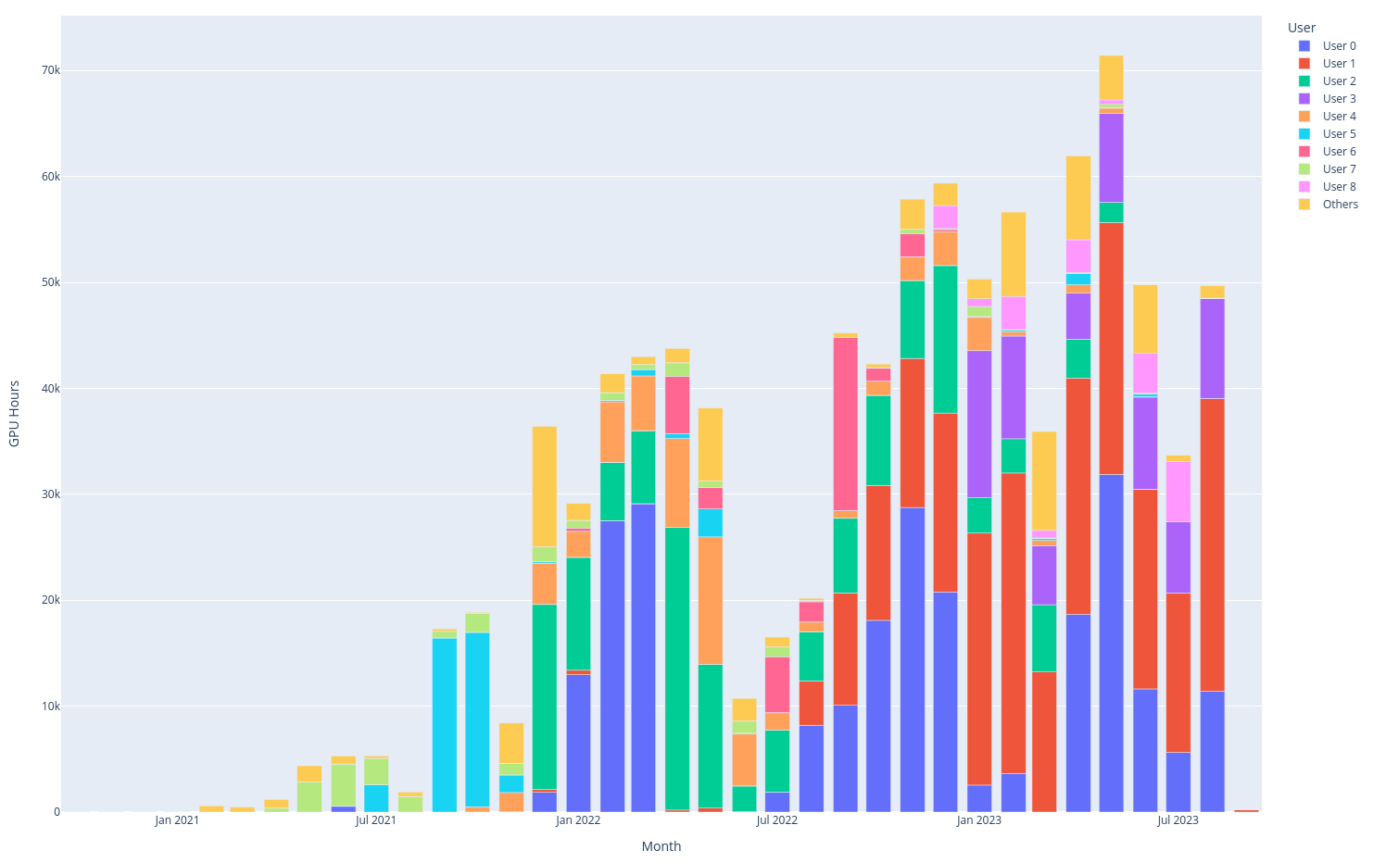
Top 3 users are from:

- Natural Language Processing Group (Computer Science)
- Chemistry

Each used more than all other users (outside of top 10) combined



JADE 2 GPU hours/month/user (top 9+others)



DNA Damage Competes with Sequence to Pin a Plectoneme

Victoria E. Hill (Sheffield), Agnes Noy (York) and Timothy D. Craggs (Sheffield)

CRAGGS LAB
SINGLE MOLECULE BIOPHYSICS

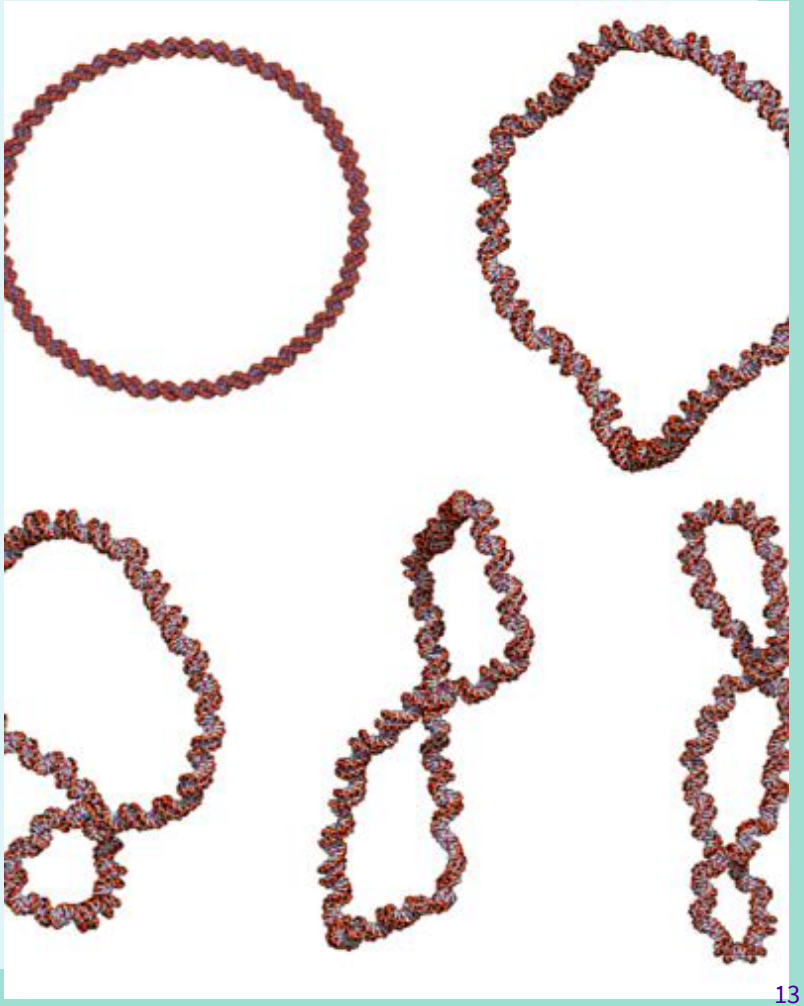
White Rose
Mechanistic Biology DTP
Universities of Leeds, Sheffield & York

UKRI Biotechnology and Biological Sciences Research Council

EPSRC
Engineering and Physical Sciences Research Council

HECBioSim

University of Sheffield



Background

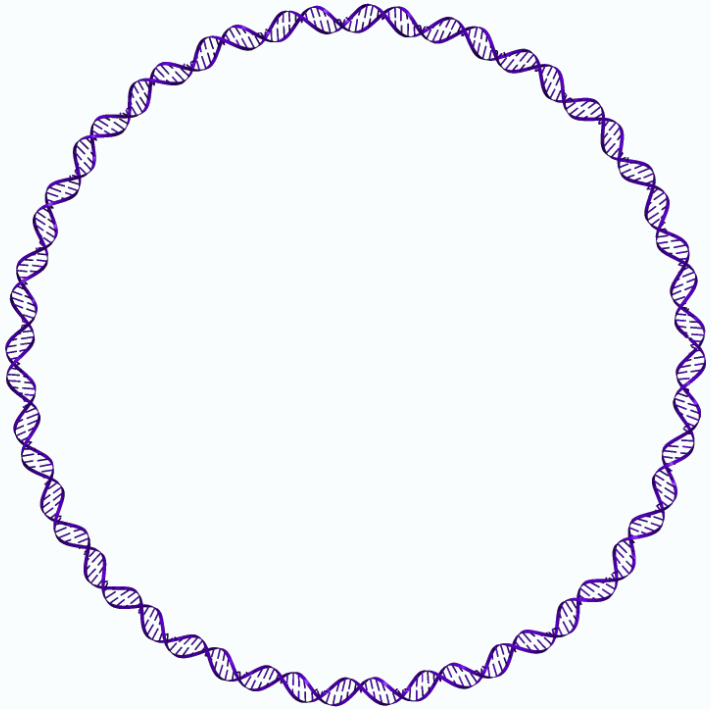
- Genomic DNA is maintained in a supercoiled state which facilitates the formation of plectonemes.
- DNA Damage is repaired with varying efficiency across the genome.
  - This may be linked to the accessibility of the damage to the relevant repair proteins.
- Hypothesis: Are damages located to the tip of plectonemes?
  - DNA repair proteins are known to bind to bent DNA

Formation of plectonemic structures (figure of eight conformations) due to supercoiling. Starting from a flat planar minicircle (top left), different shapes are formed due to varying strain and damage (top right and bottom images).

https://doi.org/10.1016/j.bpj.2022.11.1281

Simulation

- Coarse-grained models
  - Lower computational requirement
  - Less accurate and a range of DNA damage types cannot be studied due to a lack of atomic detail
  - A single atom can elicit flexibility changes propagated over the length of the DNA
- All-atom simulation
  - 339 base pair minicircle, three different sequences
  - Insertion of the same DNA damage types into the different sequences



https://doi.org/10.1016/j.bpj.2022.11.1281

Simulation

- Simulated using Amber
- The simulations are run in implicit solvent to remove computational cost in increase dynamics per calculation
  - Discrete water and ion atoms are omitted
  - Removes friction
  - Speeds up the conformational dynamics of the DNA
- Still takes 2 weeks for each simulation!

https://doi.org/10.1016/j.bpj.2022.11.1281

Testimonials

CRAGGS LAB
SINGLE MOLECULE BIOPHYSICS

“To our knowledge, these are the first all-atom simulations of DNA damage within supercoiled DNA, made possible through the use of JADE 2. This really allows us to study the effect of different damage types and of sequence with a high level of detail.”

NLP Research at Nikos' Lab

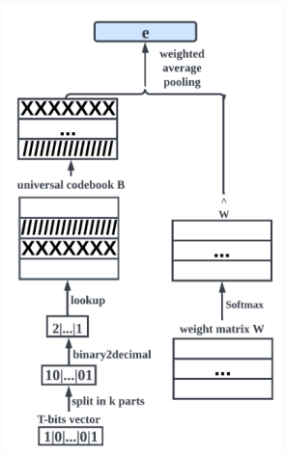
Huiyin Xue, Ahmed Alajrami, Danae Sánchez Villegas, Nikolaos Aletras

University of Sheffield

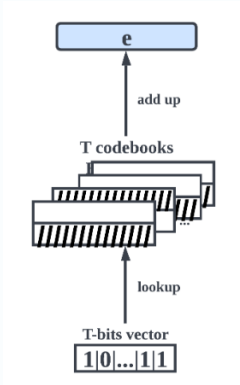
HashFormers: Towards Vocabulary-independent Pre-trained Transformers

- Background
  - The pursuit of developing larger pre-training large language models comes with large computational requirements.
    - Hindering researchers with limited access to computing resources to participate in advancing the field.
    - Has direct environmental implications such as large carbon emissions, conflicting with the principles of Green artificial intelligence development.
- Methodology
  - Research applies memory-efficient hash embeddings that could support unlimited vocabulary, enabling training with reduced computational resource
    - Introduction of three hashing approaches Pooling, Additive and Projection
- Results
  - Reduction from millions of embedding parameters to 99K with minor effect on model performance (from 124.6M total parameters to 86.1M)

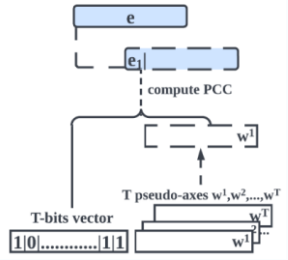
Three different hashing approaches



HASHFORMER-Pool



HASHFORMER-Add



HASHFORMER-Proj

Huiyin Xue and Nikolaos Aletras http://dx.doi.org/10.18653/v1/2022.emnlp-main.536

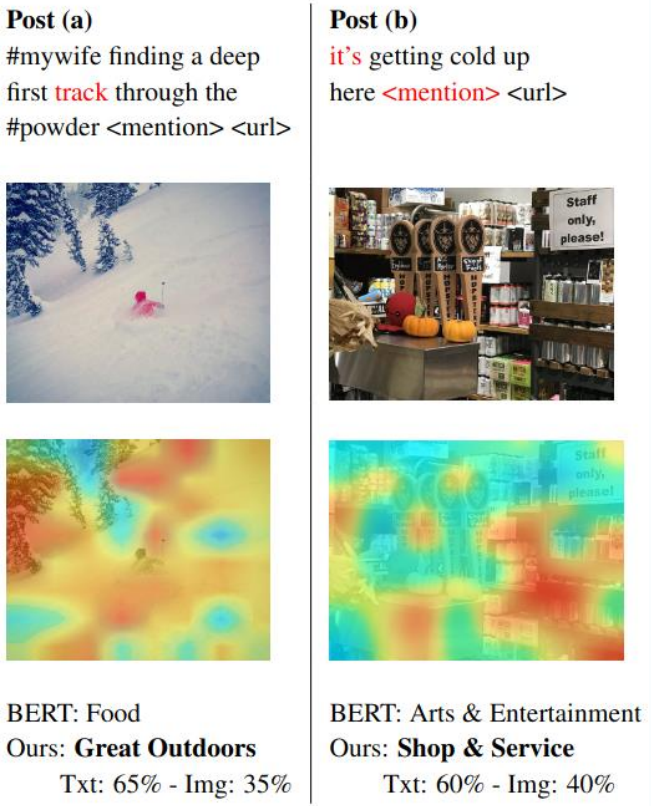
How does the pre-training objective affect what large language models learn about linguistic properties?

- Background
  - The project investigates how the pre-training objective affect what large language models learn about linguistic properties.
  - Several pre-training objectives, such as masked language modeling (MLM), have been proposed to pre-train language models (e.g. BERT) with the aim of learning better language representations.
- Methodology
  - Pre-train BERT with two linguistically motivated objectives and three non-linguistically motivated ones. We then probe for linguistic characteristics encoded in the representation of the resulting models.
- Results
  - Strong evidence that there are only small differences in probing performance between the representations learned by the two different types of objectives.

Ahmed Alajrami and Nikolaos Aletras http://dx.doi.org/10.18653/v1/2022.acl-short.16

Point-of-Interest Type Prediction using Text and Images

- Background
  - Inferring the type of a place from where a social media post was shared.
    - Useful for studies in computational social science including sociolinguistics, geosemiotics, and cultural geography
    - Has applications in geosocial networking technologies such as recommendation and visualization systems.
- Methodology
  - Studies the visual and text content in eight different types of places, such as restaurants, offices, or the outdoors.
  - Comparison between text-only, image-only and multi-modal models.
- Results
  - Significant improvements over text-only approach.



Danae Sánchez Villegas and Nikolaos Aletras https://doi.org/10.48550/arXiv.2109.00602

Analyzing Online Political Advertisements

- First computational study on online political ads with the aim to:
  - infer the political ideology of an ad sponsor
  - identify whether the sponsor is an official political party or a third-party organization
- Created two large datasets for the binary classification tasks of
  - Ideology: Conservative/Liberal - 5,548 samples, 242 unique sponsors
  - Sponsor type: Political party/Third party - 15,116 samples, 665 unique sponsors
- Collected from Google transparency report platform, ads published between May 31, 2018 up to October 11, 2020 158,599 ads
- Various multi-modal classifiers trained using the data
  - BERT+EfficientNet using image, text and descriptive caption (Densecap API) showed the best performance

Danae Sánchez Villegas, Saeid Mokaram, and Nikolaos Aletras https://doi.org/10.48550/arXiv.2105.04047

Testimonials

https://nikosaletras.com/

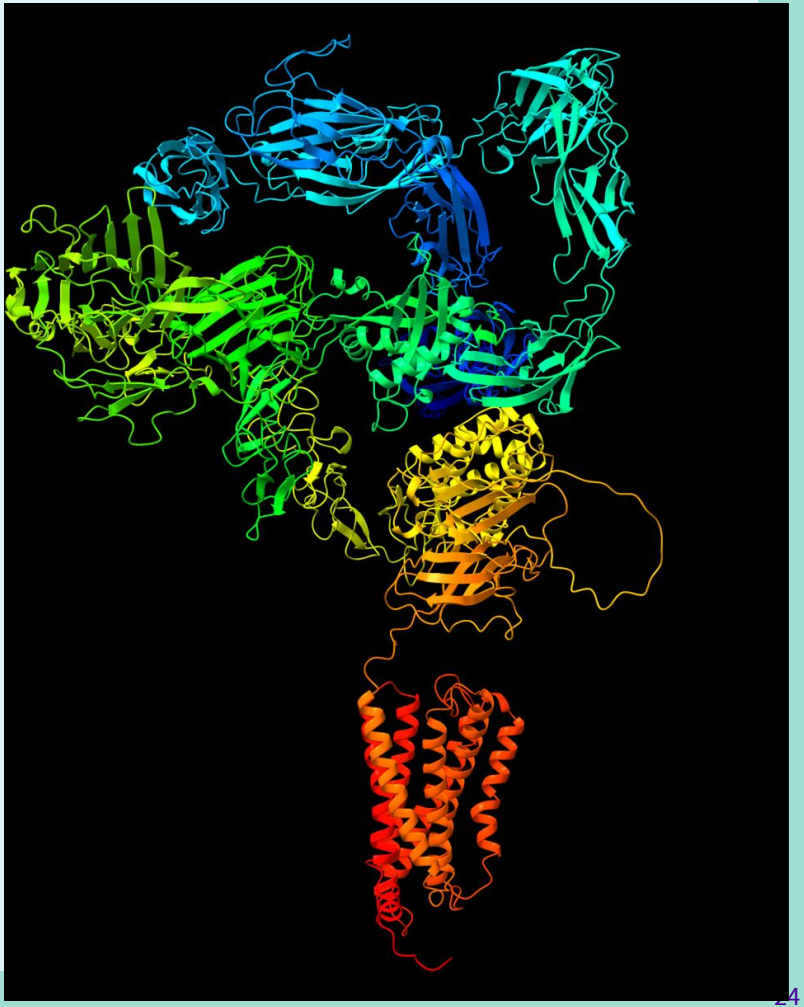
“JADE made it possible to continue my research on efficient language models, which usually requires pre-training large models on a large corpus” - Huiyin Xue

“JADE has a huge impact on my research. It empowers me to conduct large-scale experiments with unprecedented ease and efficiency.” - Ahmed Alajrami

Planar cell polarity components folding prediction

Ian Groves, David Strutt

University of Sheffield



Developmental biology

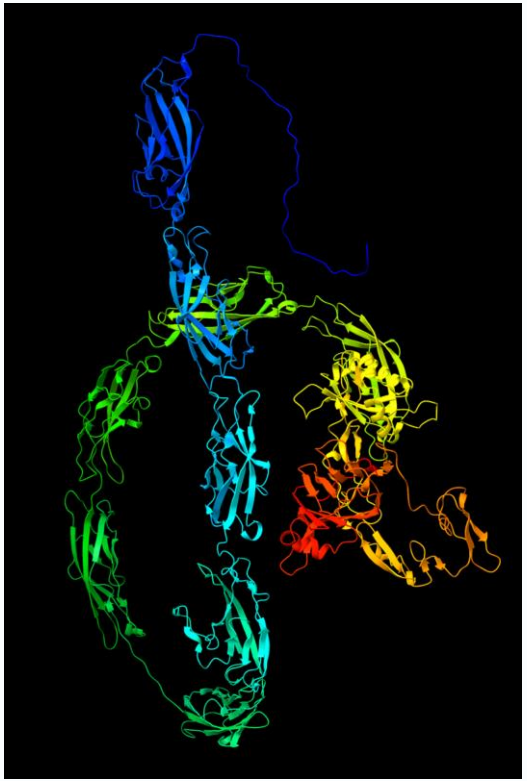
- A field focusing on how embryos develop.
- Increasingly interested in describing the development of cells, tissues, and organs at systems levels
  - Involves the use of mathematical and computational models to provide insight to the intricate developmental processes.

The Computational Model

- Creating effective computational models requires a detailed understanding of the underlying biology.
  - One way in which we can provide more detail to these models is through determining the structure of proteins that are key to developmental processes - as the structure of proteins can inform the interactions they have with other proteins.
  - However, traditional ways of deriving protein structure from a sequence of amino acids are labour intensive and slow, as amino acid chains fold into complex 3-D shapes which are not immediately obvious from the linear sequence.

Alphafold

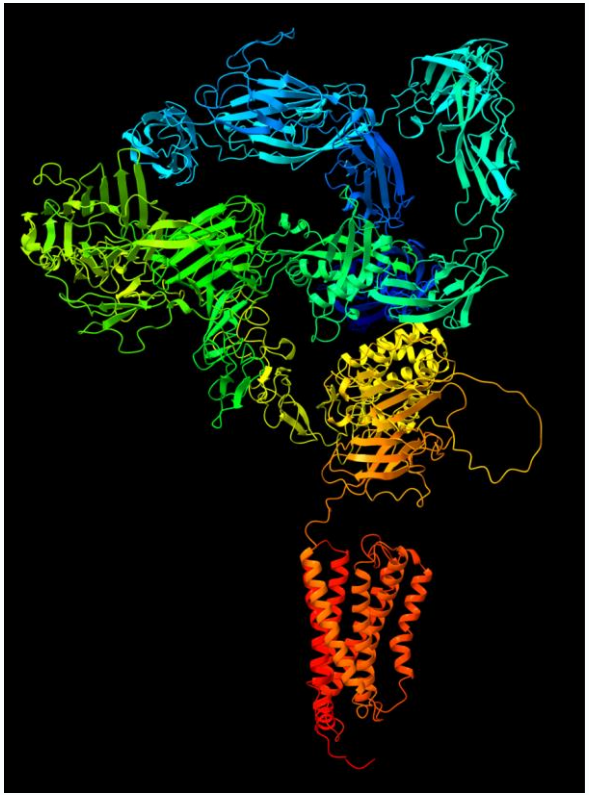
- A deep learning model which predicts the 3-D structure of a protein.
- However, running alphafold is computationally intensive and the labs at Biosciences Sheffield have run into resource limits on the various GUI implementations of Alphafold.
- Now installed on JADE, including the full set of protein databases that it uses.
    - Available to all JADE users! (https://docs.jade.ac.uk)



Part of cadherin Celsr1

Alphafold

- Allow the labs full control over the inference, and to predict much more complex proteins from longer amino acid sequences.
- Plans to explore competing models e.g. Meta’s ESMFold



Part of cadherin Celsr1

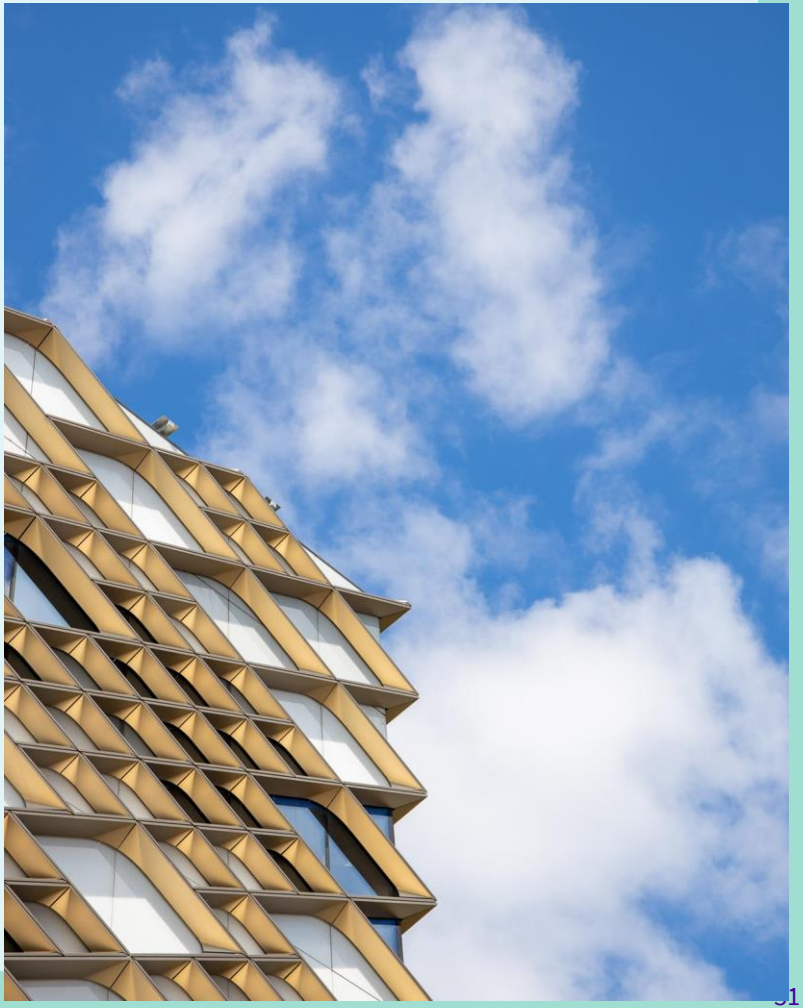
University of Sheffield

Ongoing issues

A few issues/complaints…

- Long waiting times at peak times.
- Having to suspend work for Jade maintenance.
- Lack of debug-only nodes.
- No internet connection on compute nodes.
- Having to pre-download/cache datasets or model weights using the log-in node, can’t schedule this workflow as a job.
- Not enough GPU memory for training LLMs.
- Attend the breakout session in the afternoon to discuss!

Conclusion



University of Sheffield

Conclusion

- JADE has been a tremendous resource for Sheffield
  - 60 projects, 31 active users, 109 years of compute, 20+ publications
- JADE 3 and beyond?
  - Hardware - More of everything!
    - More nodes dedicated to interactive sessions/debugging
    - Non-GPU nodes with internet connectivity that can run jobs
  - More community building!
  - Environmental considerations
    - Improving energy efficiency
    - Better monitoring of GPU usage
    - Promoting awareness e.g. carbon estimate for every job?

Acknowledgements

Thank you to the researchers and for the contributions from the RSE Team, IT Services and Data Analytics Service.

Any questions?

t.karmakharm@sheffield.ac.uk