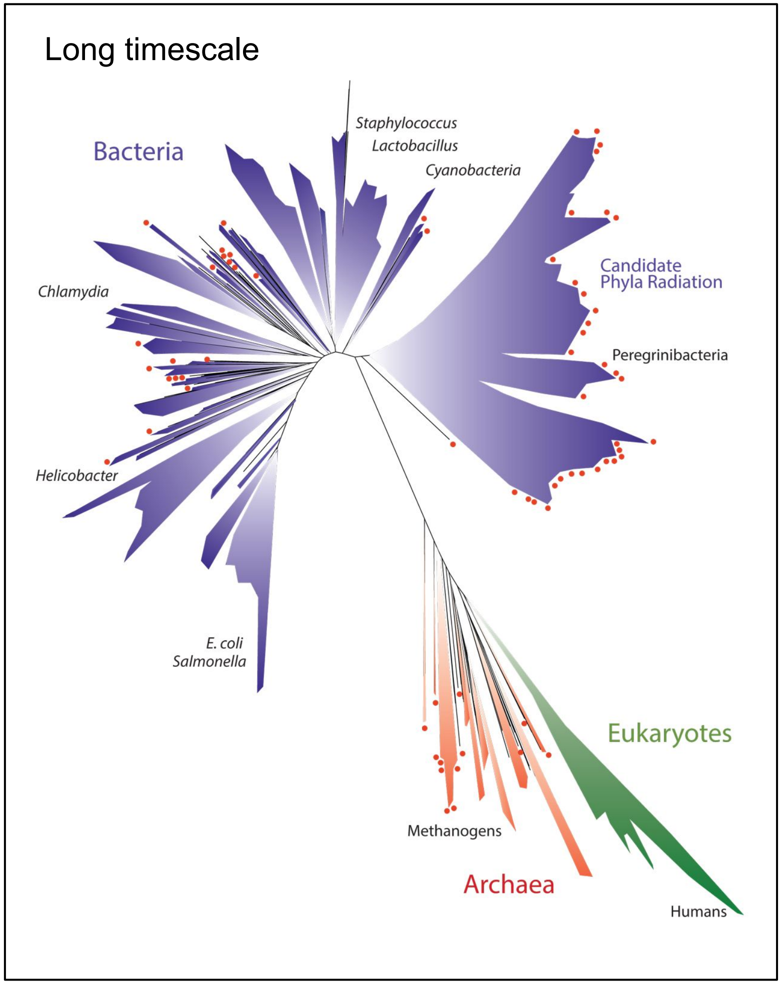
Machine learning methods for phylogenomics

Tom Williams
Royal Society University Research Fellow/Senior Lecturer
School of Biological Sciences, Bristol

Jean Golding Institute Showcase 2018

@tweethinking / tom.a.williams@bristol.ac.uk

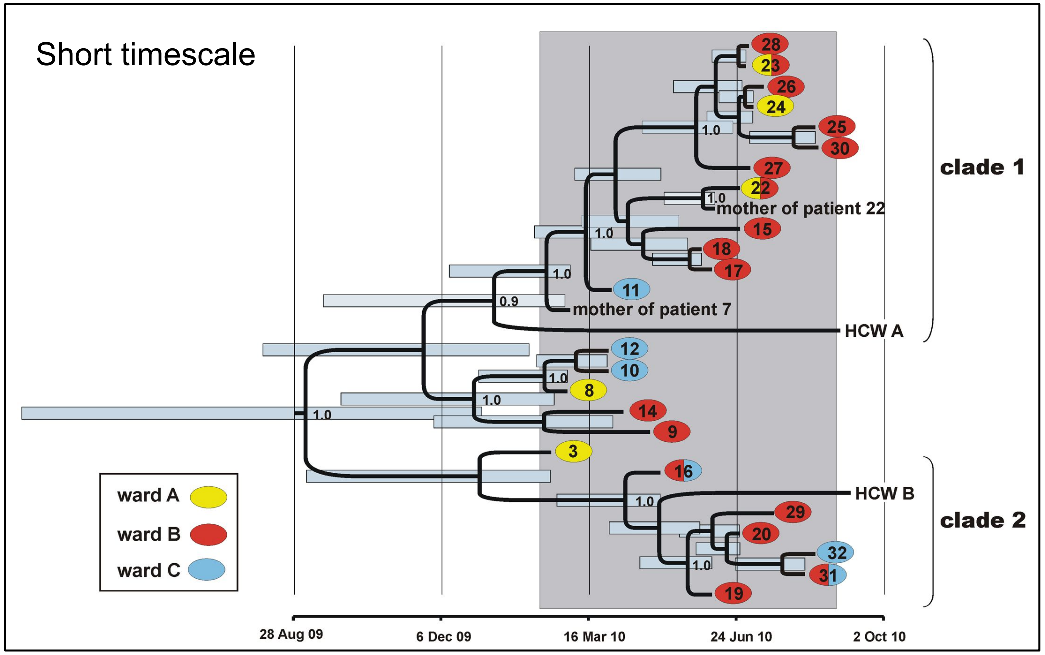
Phylogenomics: learning about evolution from whole-genome data



Long timescale

We want to understand:

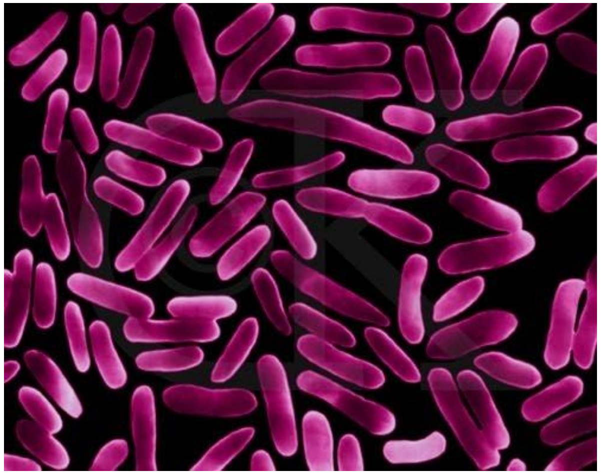
- Relationships among lifeforms
- Biodiversity
- Functional differences
- Transmission



Short timescale

Hug et al. (2016) *Nat Eco Evo;* Nubel et al. (2013) *PLoS ONE*; Williams et al. (2017) *PNAS*

Evolution is complicated!

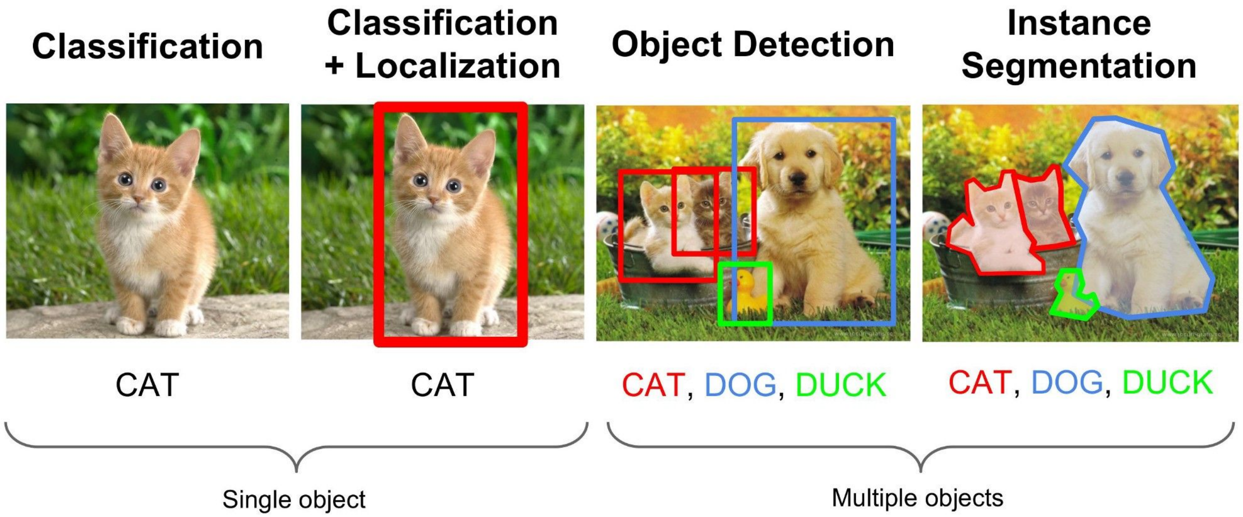


≠

Machine learning is good at complicated

Data-driven phylogenomics: can we have a computer learn these complex patterns and use them to make predictions?

Machine learning has proven useful for analysis of complex data in computer vision.



Ouaknine (2018, medium.com)

Machine learning problems in biology

Data-driven phylogenomics: can we have a computer learn these complex patterns and use them to make predictions?

Relevant problems (in classification and prediction):

- Is this organism is a pathogen?
- Will this mutation cause disease?
- What is the optimal growth temperature of this organism?
- What is the evolutionary tree relating these organisms?

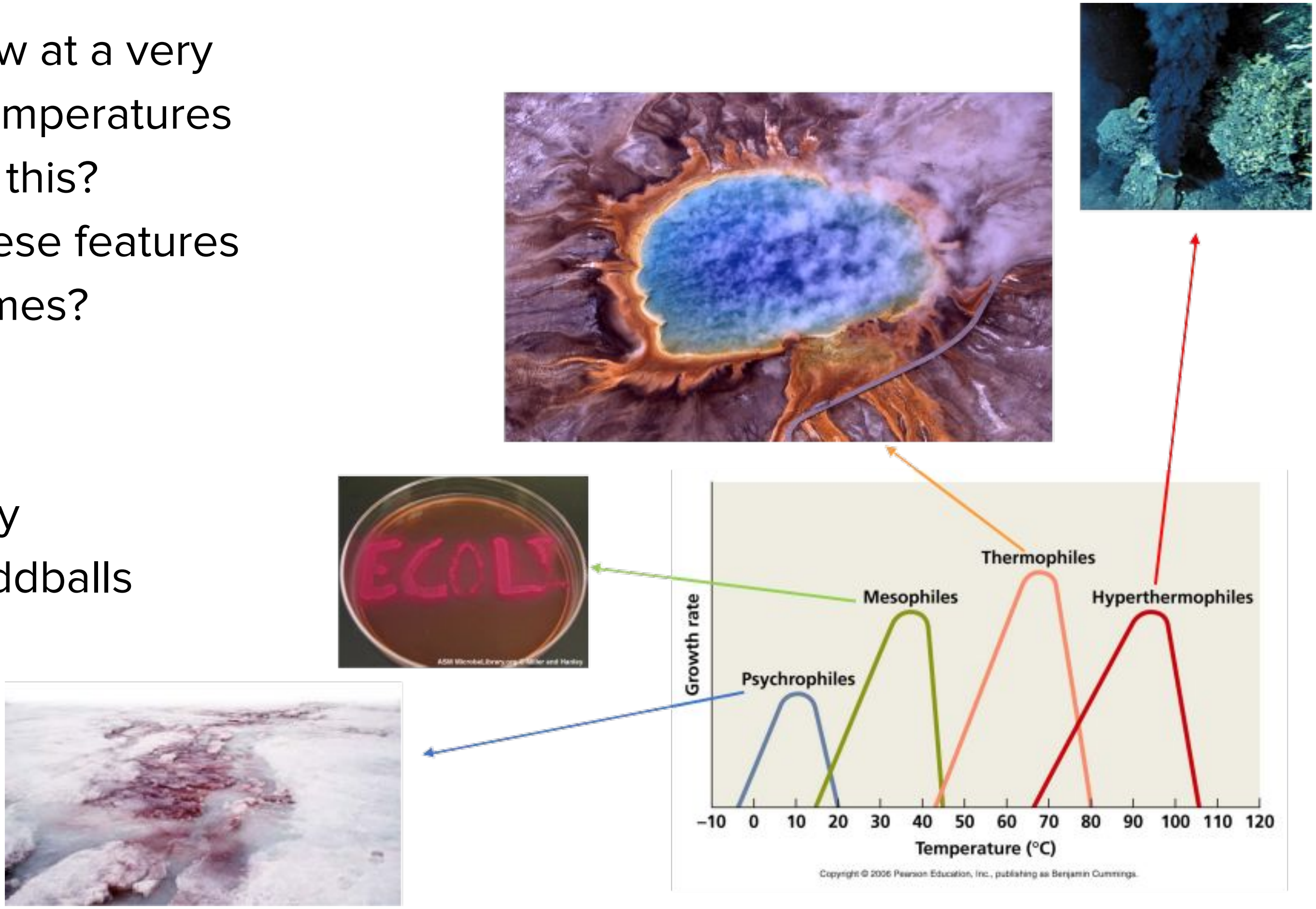
We obtained seedcorn funding from the JGI to begin this project!

Machine learning to predict growth temperatures

- Prokaryotes grow at a very wide range of temperatures
- How do they do this?
- Can we learn these features from their genomes?

Applications:

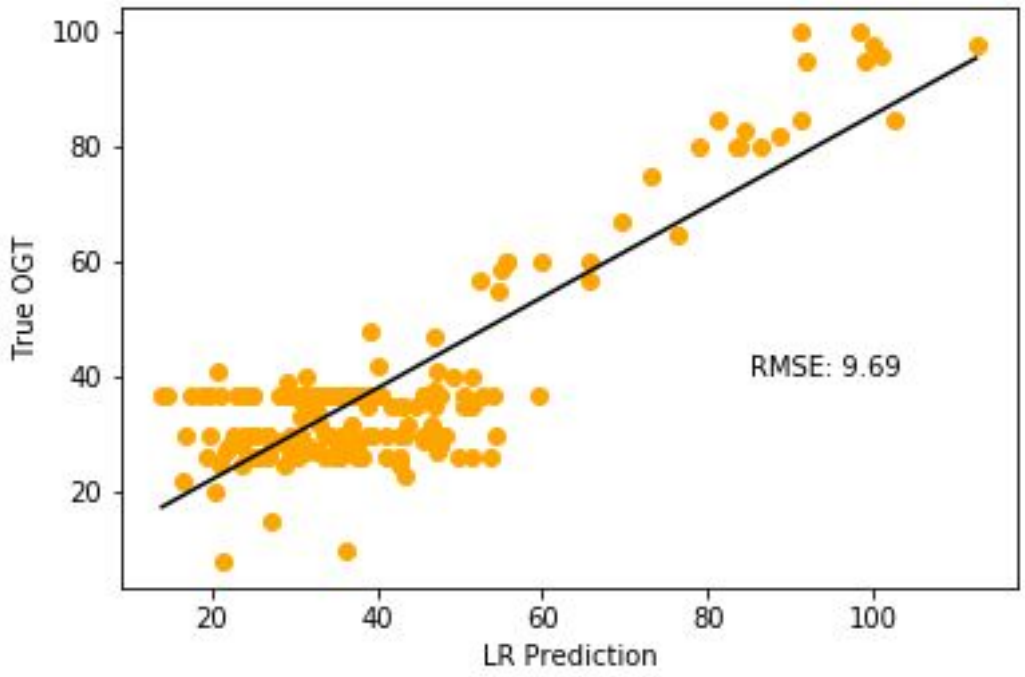
- Synthetic biology
- Lab culture of oddballs



Copyright © 2006 Pearson Education, Inc., publishing as Benjamin Cummings.

Ed Moody, Mark Beaumont (Biological Sciences); James Fearn, Colin Campbell (Engineering Maths)

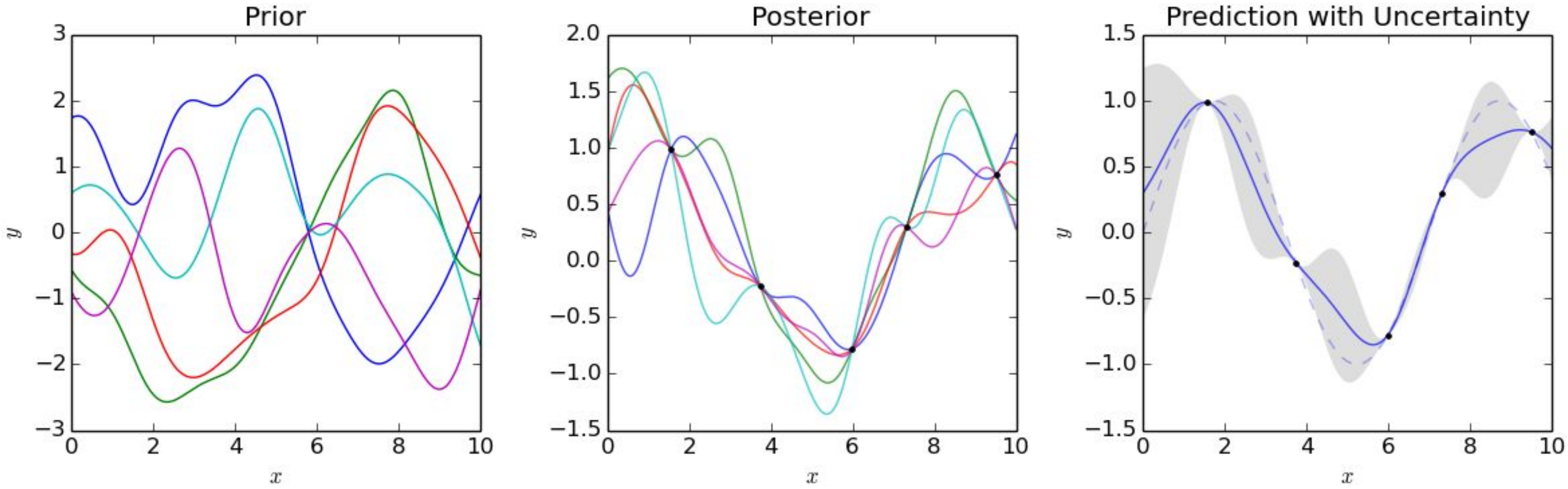
Linear regression is somewhat useful



- Amino acid usage correlates with optimal growth temperature
- Previous work identified **I, V, Y, W, R, E, L** as associated with high growth temperatures
- But, overfits data (phylogenetic signal!)

Zeldovich et al. (2007) *PLOS Comput Biol*

Modelling more complex relationships with ML

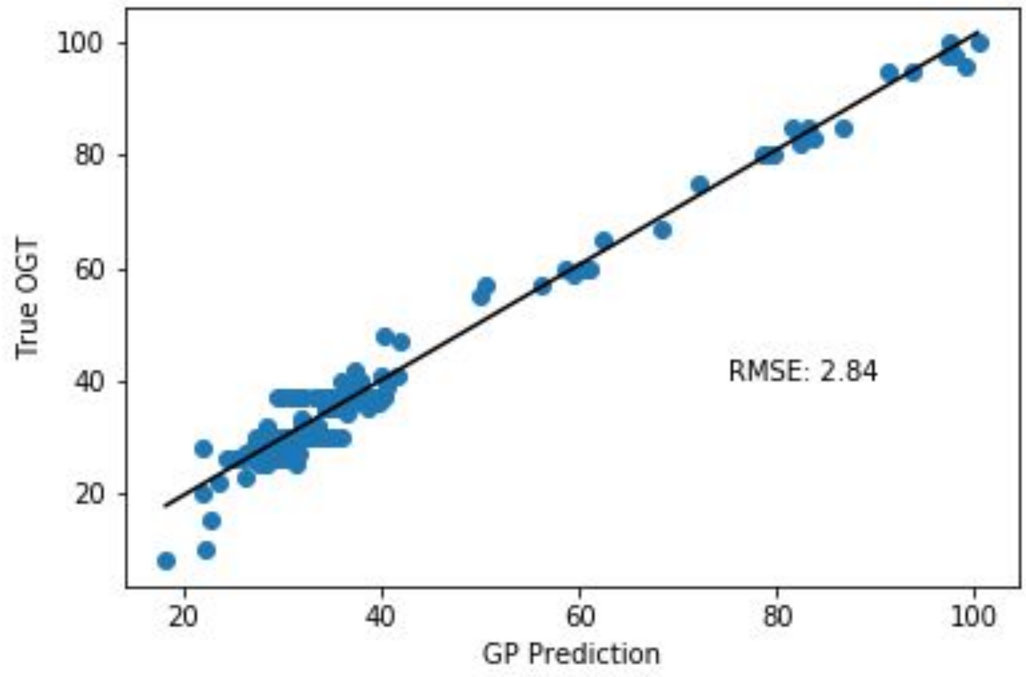


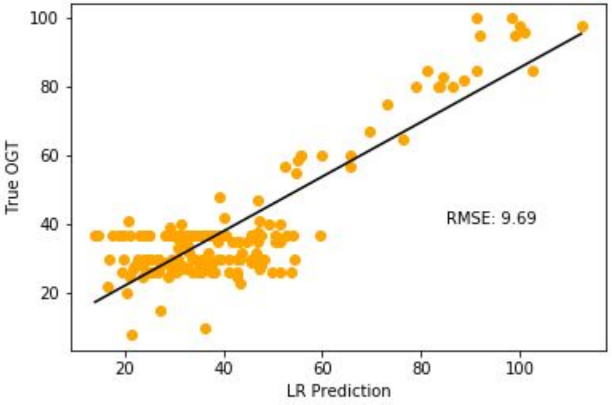
Gaussian process model:

- Sample many different possible relationships between *x* and *y* compatible with the training data
- Make predictions averaging over the distribution of functions and their probabilities

Rasmussen and Williams (2005)

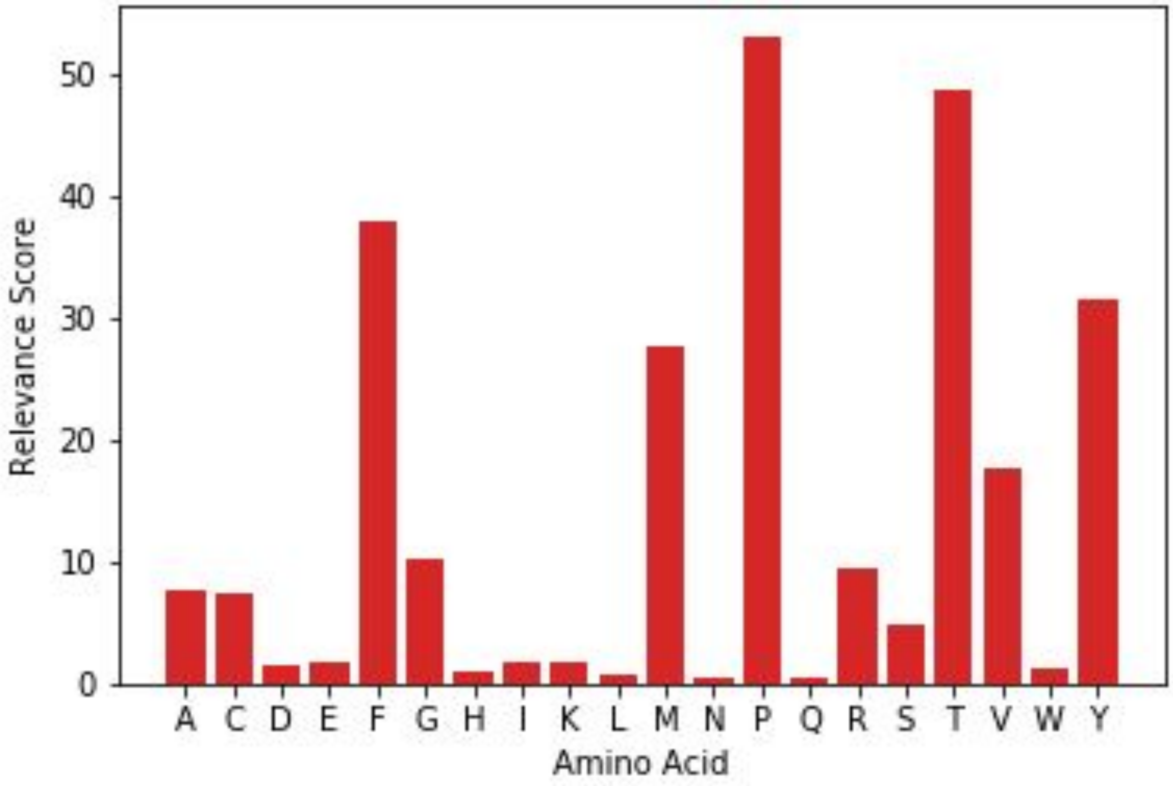
Gaussian process model beats parametric approaches





- Predicts better across the range
- Seems not to overfit (less dataset-dependent)

Which amino acids co-vary with growth temperature?



- Overlap with, but distinct from, published predictions using linear regression **(not IVYWREL)**
- Estimate magnitudes of contribution (which are most important?)
- Proline is an interesting one!

Conclusions and future work

- ML techniques good for making predictions in evolutionary biology.
- Apply Gaussian process model to prediction of cultivation temperatures for uncultivated microbes
- (Try to) make thermostable variants of proteins with targeted changes

Acknowledgements:

- Ed Moody, James Fearn, Mark Beaumont, Colin Campbell
- Jean Golding Institute