

Characterisation and Identification of STEC O157:H7 Pathogenicity using Machine Learning

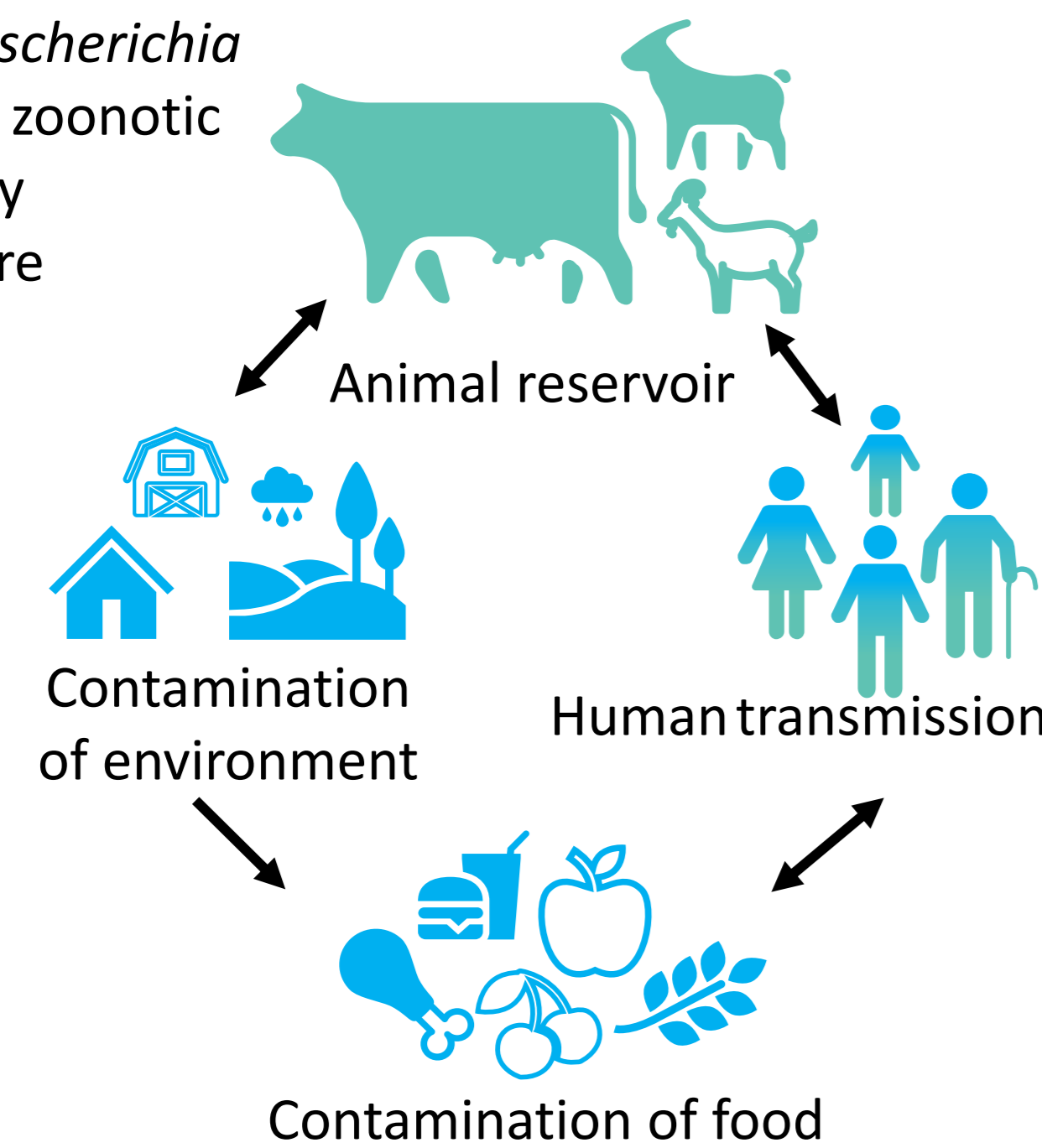
Suniya Khatun¹, David Greig^{2,3,5}, Lauren Cowley⁴, Claire Jenkins^{2,3}, Timothy J. Dallman^{2,3}

¹Institute of Structural and Molecular Biology, Division of Biosciences, University College London, London, WC1E 6BT, United Kingdom. ²National Infection Service, Public Health England, 61 Colindale Avenue, London, NW9 5EQ, United Kingdom.

³NIHR HPRU in Gastrointestinal Infections, PHE, London, UK. ⁴University of Bath, Biology and Biochemistry, Bath, BA2 7AY, United Kingdom. ⁵The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, UK.

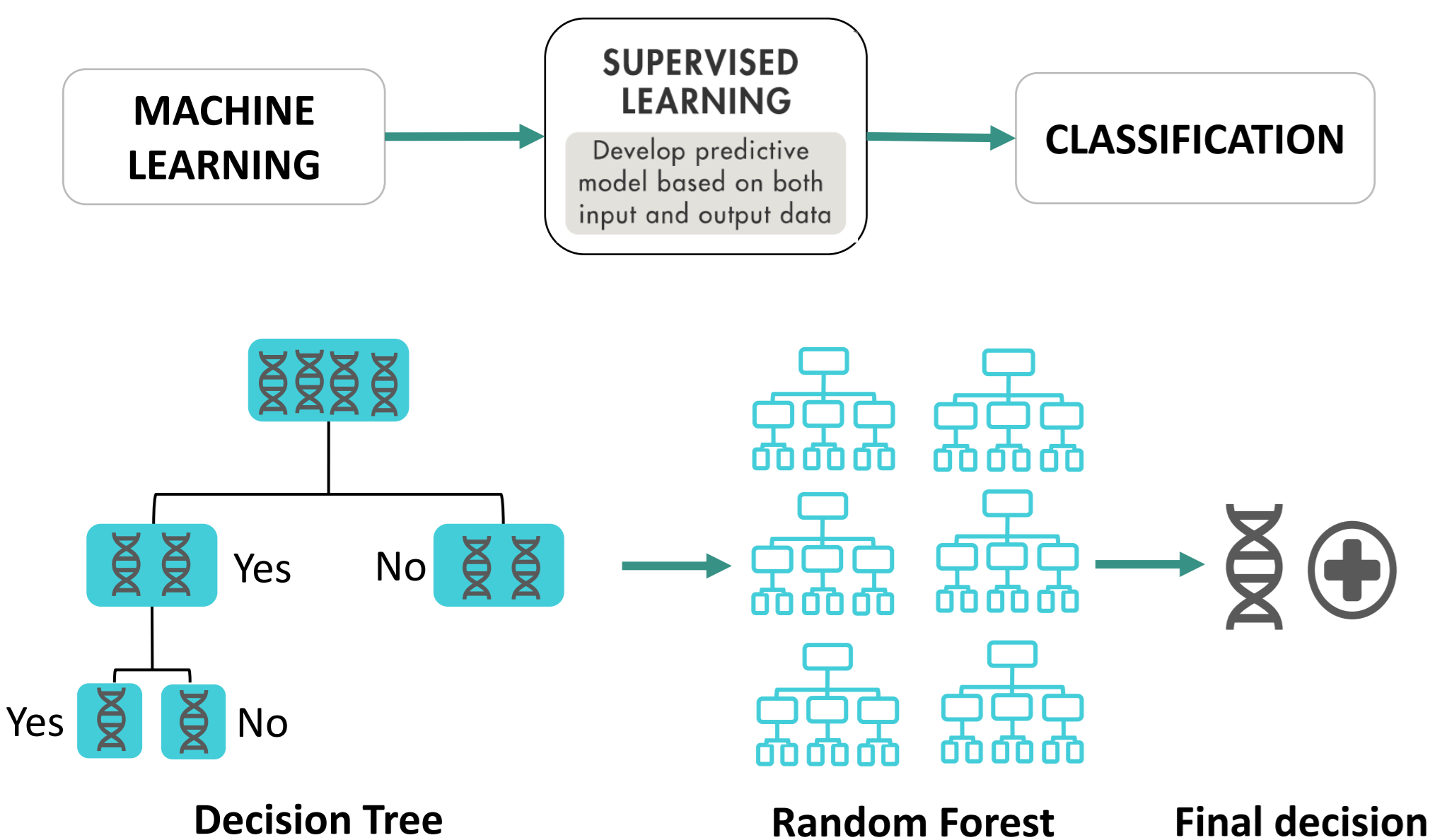
Introduction

Shiga toxin-producing *Escherichia coli* O157:H7 (STEC) is a zoonotic pathogen that is globally dispersed, causing severe gastroenteritis when transmitted from ruminants to humans.



Is there any **association** between STEC **k-mer sequence** and the observed **clinical outcome**?

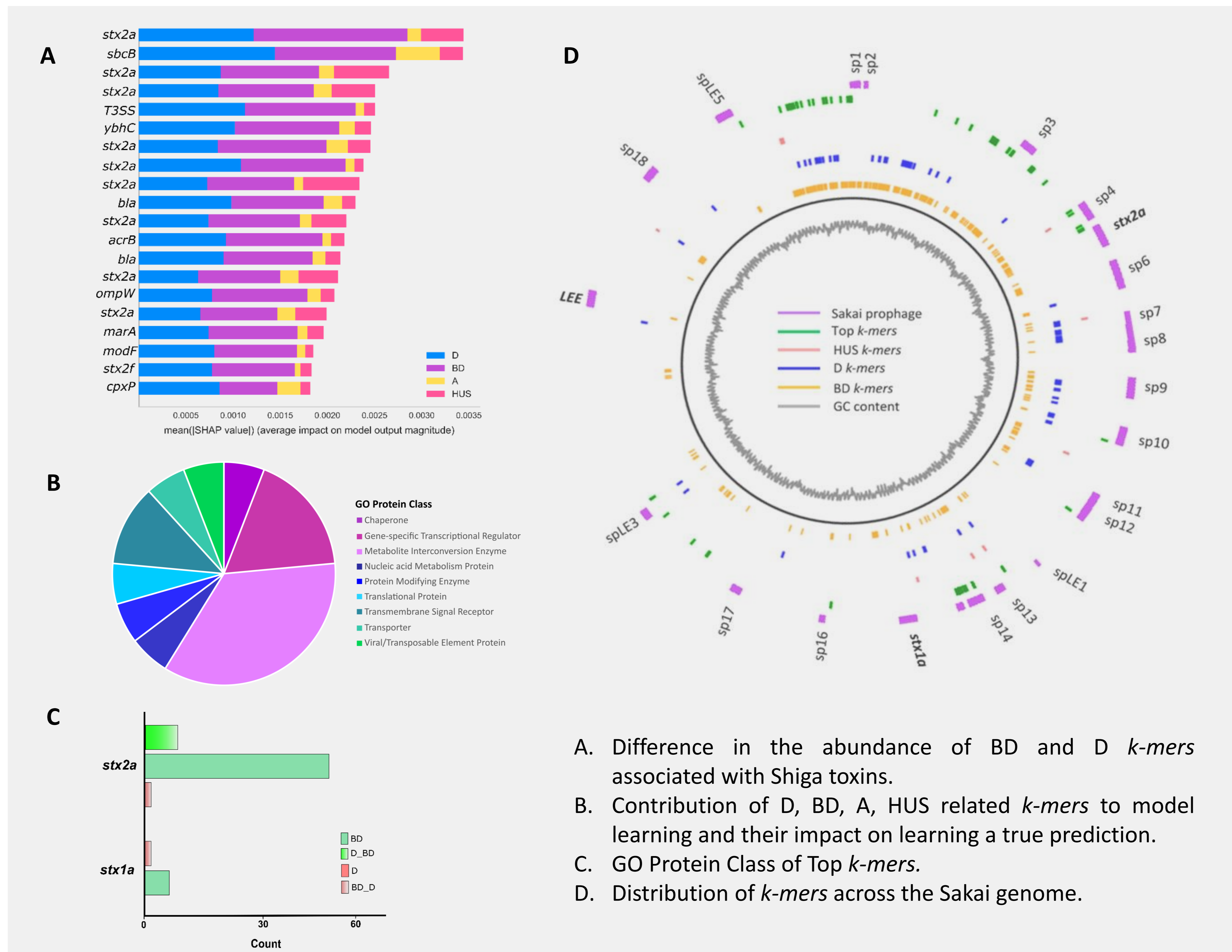
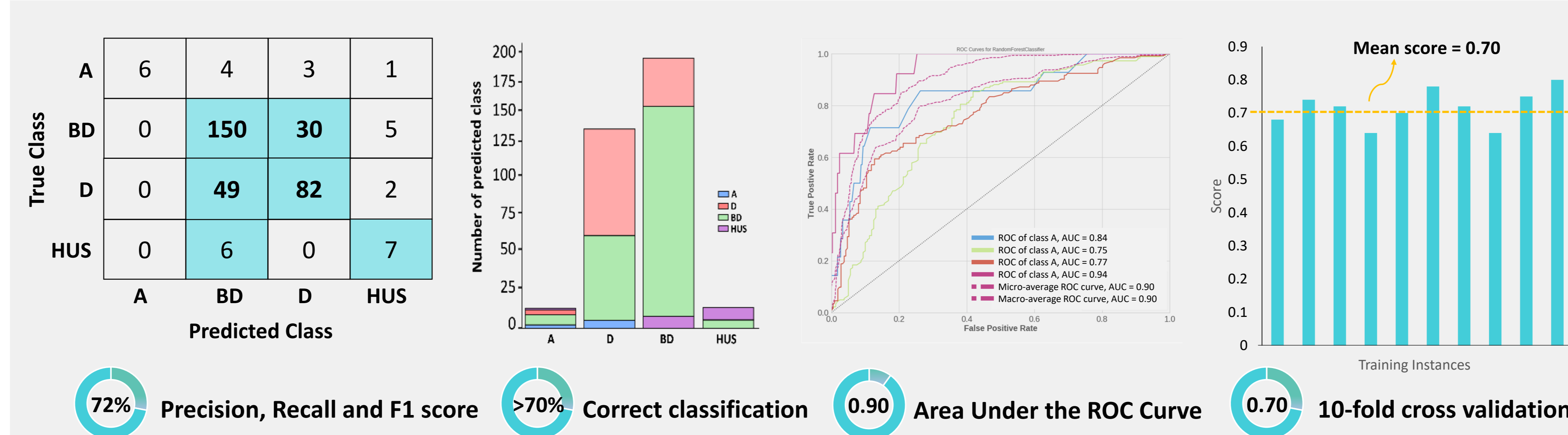
Method



- 1 Data
- 2 Train
- 3 Evaluate
- 4 Model

- 1148 STEC O157:H7 isolated from human cases in 2017 and 2018.
- DNA sequence of each isolate were sub-sequenced into k length of **k-mers** (9-100).
- 1 million randomly selected **k-mers** were used to train the Random Forest classifier.
- k-mers copy number was **normalised**.
- Symptoms classed into: **Asymptomatic (A), Diarrhoea (D), Bloody Diarrhoea (BD), Hemolytic uremic syndrome (HUS)**
- Recursive feature elimination with **10-fold cross validation** performed.

Results



- Difference in the abundance of BD and D **k-mers** associated with Shiga toxins.
- Contribution of D, BD, A, HUS related **k-mers** to model learning and their impact on learning a true prediction.
- GO Protein Class of Top **k-mers**.
- Distribution of **k-mers** across the Sakai genome.

Conclusion

- 1 Association between STEC genome and clinical outcome learnt using Random Forest classifier. Association found to be linked to specific **k-mer** profile.
- 2 BD specific **k-mers** and **k-mers** associated with Shiga toxins had greater impact on models learning.
- 3 Pathogenicity highly dependent on the presence of specific types of Shiga toxin.
- 4 Severity of clinical outcome is strongly correlated to the lineage of STEC O157:H7 isolate.

Acknowledgement

The research was funded by BBSRC in collaboration with the National Institute for Health Research Protection Research Unit (NIHR HPRU) in Gastrointestinal Infections Unit Public Health England (PHE). The views expressed are those of the author(s) and not necessarily the NIHR, the Department of Health and Social Care or Public Health England.

Reference

1. Byrne L, Jenkins C, Launders N, Elson R, Adak GK. *Epidemiol. Infect.* 2015;**143**:3475–3487
2. Cowley LA, et al. *Microb Genom.* 2016;**2**;9; doi:10.1099/mgen.0.000084.
3. Breiman L, Friedman J, Stone C, Olshen R. Classification and regression trees. 1st ed. Taylor & Francis; 1984
4. Inward CD, Milford DV, Taylor CM. *Pediatric Nephrology*, 1993; **7**, 771–772.
5. Dallman JT, et al. *Microb Genom.* 2015;**1**;3