Kromek Group plc NETPark, Thomas Wright Way, Sedgefield, UK www.kromek.com

Predicting Pathogenic Properties From Single Nanopore Reads

1) A Read-by-Read Early Warning System

- Monkeypox, Langya Henipavirus, and other recent zoonoses have shown how animal diseases can quickly mutate to become human-pathogenic, with COVID causing disease on a societal scale.
- Nanopore based real-time environmental sequencing enables the development of early warning systems, but new interpretation tools are required to capitalise on these advances.
- ***We aim to quickly and accurately predict properties of reads from known and unknown pathogens in a sample-independent manner. Our system infers a range of easily interpretable characteristics directly on a read-by-read basis.***

2) A Neural Network Approach

- We have developed neural net architectures which take in noisy nanopore reads (.fastq), and predict key threat characteristics associated with the genome of origin including:
  - Pathogenicity to humans and other vertebrates
  - Likely symptoms and tissue tropisms
  - Interactions with animal hosts
  - Bio-safety level requirements

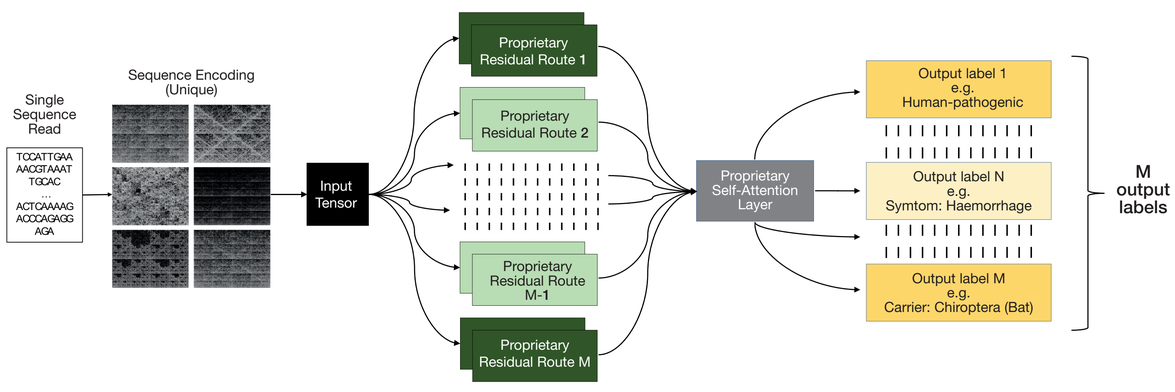

Figure 1: *The novel architecture of **Deep Trait Inference** for **Genomic Reads** (**DeepTIGR™**)*

3) Training Data Campaigns

- Building on DeepSimulator [1], we can create simulated reads from public and proprietary reference genomes. These genomes are chosen to encapsulate a range of viral pathogens, as well as benign counterexamples.
- To label these reads, we are compiling an extensive database of characteristics associated with >450 viruses, prokarya, eukarya.
  - Many of these characteristics, including host tissue tropisms and animal host associations, are not available in databases. Instead they require verification from the literature in virology and epidemiology.
  - Of the 23,562 labels now associated with our data, 6214 have required and undergone literature verification (see **purple** in Fig. 2).
- By training our models with this growing corpus of characteristics for a range of dangerous pathogens, our neural net is uniquely able to identify these hazardous traits in known and emerging species.

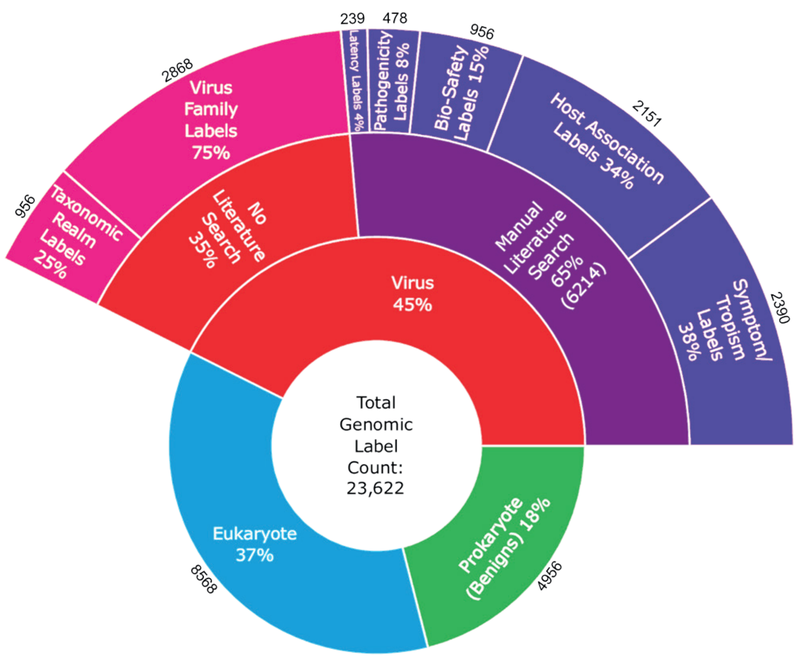

Figure 2: *Our database contains several hundred genomes. For each genome, 49 labels are assigned. While many labels are obvious from taxonomy, many pathogen-associated labels require manual literature verification. Here we show the proportion of labels associated with each taxonomic realm, with those which require this verification in **purple**.*

DeepTIGR NEEDS YOU

- If the work shown here is of interest to you, our system is currently available as an API at ***deeptigr.kromek.com***.
- If you would like to try ***DeepTIGR*** on your nanopore read data, please contact **deeptigr@kromek.com** and we would be delighted to provide you with user account details and instructions.
- ***We will not save your data*** unless you explicitly give us permission to do so.

4) In Silico Read Results

- ***DeepTIGR*** works on in silico reads from un-trained genomes, at high F1-score.
- Across all 49 characteristics for 12 un-trained pathogenic species and 16 benign species, our model achieves balanced F1 scores in excess of 93%.
- In Fig. 3 we show confidence distributions from ***DeepTIGR*** for some key characteristics for a subsample of viral and non-viral species. We see close model agreement with ground truth values.

*Figure 3: Sample results for un-trained virus species. Histograms show **DeepTIGR** confidences for 1k in silico reads from each species (row) and for each characteristic (column). **Green** backgrounds indicate positive (true) ground truths. **Red** backgrounds indicate negative (false) ground truths, i.e., a **green** box and peak at **1** is a good result, while a **red** box with a peak at **0** is a good result.*
***Blue** backgrounds indicate unknown ground truths.*
*Rightmost column shows predicted biosafety levels, with ground truth as the **red** line.*

5) Nanopore Read Results

We have begun to optimise DeepTIGR to real nanopore reads, by predicting characteristics for a few sequenced samples of high purity, sourced locally and from the NCBI Sequence Record Archive (SRA). In Fig. 4 we show divergences of the DeepTIGR predictions from the ground truth for four species samples of high purity. The maximum divergence for COVID is larger than for the benign species, owing to its richer set of pathogenic characteristics. Despite this, for a fixed confidence threshold of 0.5, we recover a balanced per-read, per-trait median accuracy of 96.9% on these nanopore data.

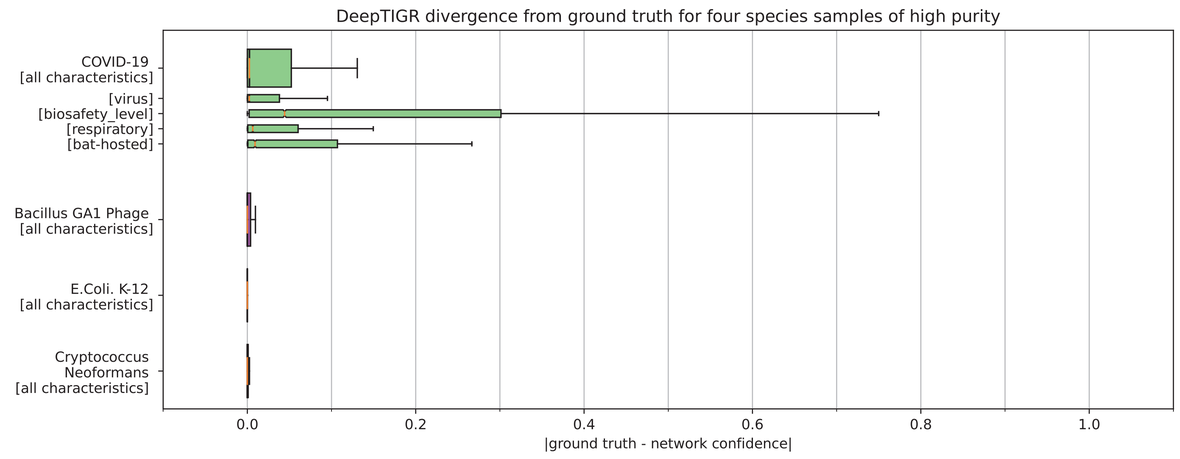

*Figure 4: Box-and-whisker plots showing the divergence from the ground truth values for **DeepTIGR**-inferred confidences for four nanopore species samples of high purity. The y-axis labels indicate the sampled species. In addition to the averages over all characteristics for our four species, we also show individual plots for four salient traits for COVID-19.*

6) In-Silico vs Nanopore Noise Comparisons

- We are currently comparing noise statistics between our in-silico ***DeepSimulator*** reads and a range of nanopore reads drawn from various studies and libraries prepared with a range of methods.
- This is enabling us to optimize our read simulation procedure to better reflect a range of real-world read noise profiles.

*Figure 5: CIGAR match length proportions for nanopore Zaire ebolavirus reads (**blue dashed line**), nanopore SARS-CoV-2 reads (**red dashed line**), and in silico reads from DeepSimulator (solid lines) when the nanopore signal cutoff frequency is varied. This is one of several parameters which can be optimized for better training data.*

References
[1] Yu Li, Sheng Wang, Chongwei Bi, Zhaowen Qiu, Mo Li, Xin Gao, DeepSimulator1.5: a more powerful, quicker and lighter simulator for Nanopore sequencing, Bioinformatics, Volume 36, Issue 8, 2020, pp. 2578–2580

Acknowledgements
We gratefully acknowledge the support of the UK Defence Science and Technology Laboratory under contract no. DSTLX1000160308.