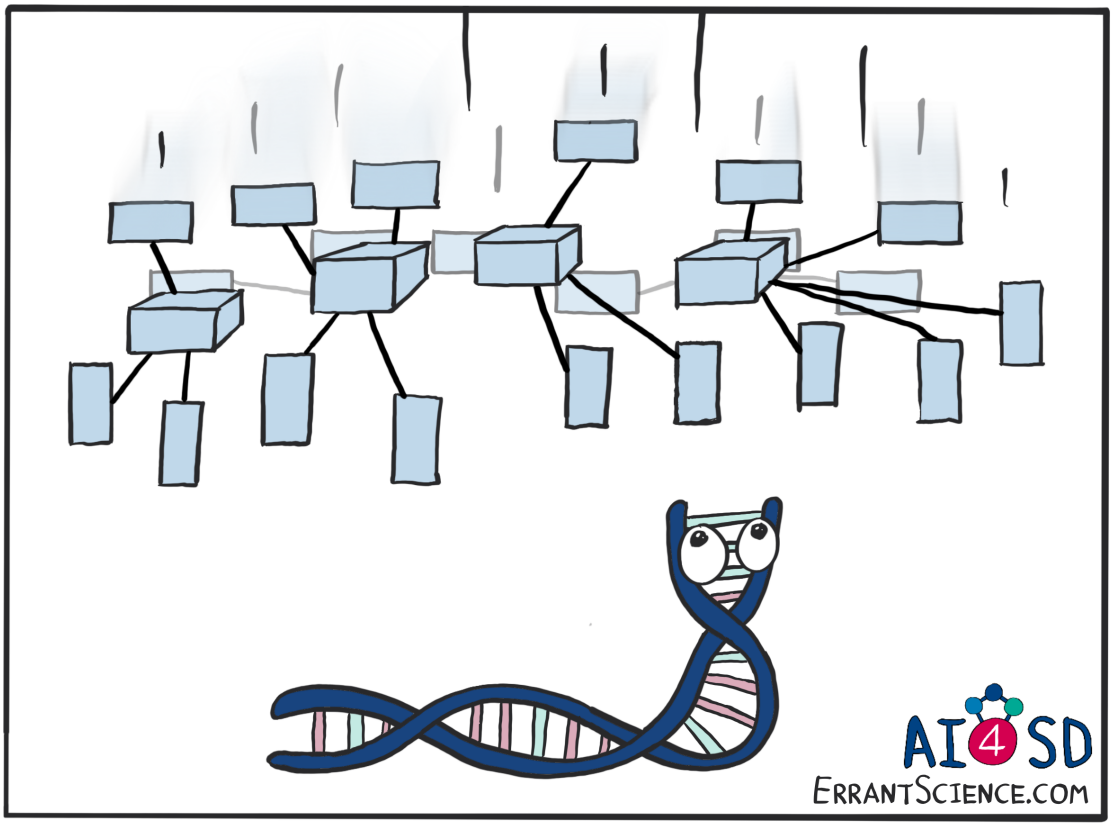
AI 4 SCIENTIFIC DISCOVERY

AI 4 Science Discovery Network+

Application of Capsule Net for automated DNA sequencing using tunnelling spectroscopy Final Report

Project Dates: 01/07/2019 - 31/12/2019

University of Birmingham, School of Chemistry



Professor Tim Albrecht & Dr Anton Vladyka
University of Birmingham

Report Date: 22/06/2020

Application of Capsule Net for automated DNA sequencing using tunnelling spectroscopy
AI4SD-Project-Series:Report-1_Albrecht_Final
Report Date: 22/06/2020
DOI: 10.5258/SOTON/P0036

Network: Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery
This Network+ is EPSRC Funded under Grant No: EP/S000356/1

Principal Investigator: *Professor Jeremy Frey*
Co-Investigator: *Professor Mahesan Niranjan*
Network+ Coordinator: *Dr Samantha Kanza*

Contents

1 Project Details

Title	Application of Capsule Net for automated DNA sequencing using tunnelling spectroscopy
Funding reference	AI4SD-FundingCall1_016
Lead Institution	University of Birmingham
Project Dates	01/07/2019 - 31/12/2019
Website	N/A
Keywords	Machine Learning; single-molecule detection; Capsule Nets; Deep Learning

2 Project Team

2.1 Principal Investigator

Name and Title: Tim Albrecht Professor of Physical Chemistry
Association: University of Birmingham, School of Chemistry
Work Email: t.albrecht@bham.ac.uk
Website Link: albrechtlab.com

2.2 Co-Investigators

Name and Title: Dr Eduardo Alonso, Reader and Head of Department
Association: City, University of London, School of Mathematics, Computer Science & Engineering, Department of Computer Science
Work Email: e.alonso@city.ac.uk
Website Link: https://www.city.ac.uk/people/academics/eduardo-alonso

2.3 Other Researchers & Collaborators

Dr Anton Vladyka is the postdoctoral research associate currently working on the project. He has replaced Dr Joseph Hamill, who was originally named on the application, as Joseph continued his work on another project taking place in the group.

Mr. Christropher Weaver has completed his Master project in the Albrecht during the 2018/19 academic year with a project on the detection of DNA nucleotides using Scanning Tunnelling Microscopy in solution. He was able to generate the first experimental proof-of-concept data, which formed the basis for the current proposal. As Chris has a strong background in AI, having completed the Chemistry with a Year in Computer Sciences MSci degree in Birmingham, he was also able to repeat some of the work we had published in our 2017 Nanotechnology paper on Deep Learning in Single-Molecule Science. He will join the group again in October, as PhD student, with the aim to systematically investigate various experimental parameters to optimise the detection of DNA nucleotides via tunnelling. He will then also benefit from the foundation in AI techniques that has been set up in the current project. As a consequence, however, we decided to put greater focus on the AI aspects of the project at this early stage and then intensify our effort around the generation of experimental data towards the end of the project. Chris will then also be able to continue the development of AI-based approaches, given his background.

New collaborations have been enabled by the AI3SD project, albeit after it formally ended, in particular with colleagues in Computer Science and the Data Science Imstitute at the University Birmingham (Prof. Iain Styles, Dr Hyung Jin Chang). This includes joint MSc project (Ertugrul Hadzhaoglu) in the area of AI-enhanced sensing. TA is now also Affiliate Member of the AI Research Centre at City, University of London, and he has founded the "AIChem" Interest Group in the School of Chemistry, which is a platform for knowledge exchange in the area of AI in Chemistry.

3 Publicity Summary

In this project, we take the next step for quantum tunnelling-based biosensing and sequencing to become promising contender for label-free 'next-next' generation sequencing of biopolymers. This could be achieved by combining state-of-the-art surface chemistry, nanoscience and chemical sensing with Capsule Nets (CN) as a novel Deep Learning methodology, to maximise the extraction of information from the tunnelling data. Our results to date show a moderate improvement in detection accuracy when using Capsule Nets, compared to Convolutional Neural Networks. CapsNets were however superior when dealing with incomplete data, for example when recognising events at the edges of the recording window, which could partly offset the increased computational cost. Ongoing work beyond this project now aims to further explore this aspect and support our findings with increasing amounts of experimental data.

In addition, the project has allowed us to explore a new approach to data classification of single-molecule charge transport data using Transfer Learning and Image Recognition Networks (IRN). In this context, the feature extractor of IRNs trained on millions of (unrelated) image data is used to recognise characteristics in conductance-distance or current-time data, which are then clustered and interpreted. This means that no charge transport data are required for training the feature extractor, removing the need for large, application-specific training data.

4 Executive Summary

The aim of this project is to explore the limits of quantum-based sensing for individual DNA nucleotides A, T, C and G in combination with Capsule Nets ("CapsNets"), which is a realistic goal for a 4-month feasibility study. If its performance is comparable to or even exceeds current commercial sequencing techniques at this early stage, then this makes a strong case for further, significant investment towards the implementation of a complete sequencing platform on the medium- and long-term. So far, our AI-based studies have mainly involved simulated data, in order to better understand the operation and performance of CapsNets under different conditions. However, work is continuing at a larger scale with an expanding set of experimental data, in order to validate the approach.

5 Aims and Objectives

In the proposed project, we aimed to take quantum tunnelling-based biosensing and sequencing to a level that would make the most promising contender for label-free 'next-next' generation sequencing of biopolymers. This was going to be achieved by combining state-of-the-art surface chemistry, nanoscience and chemical sensing with Capsule Nets (CN) as a novel Deep Learning methodology, to maximise the extraction of information from the tunnelling data. In a wider context, we also set out to explore new applications of Machine Learning in Chemistry and the Natural Sciences.

6 Methodology

6.1 Scientific Methodology

The project has employed state-of-the-art experiments in (electro)chemical sensing, surface electrochemistry, state-of-the-art electronics and cutting-edge Deep Learning methodologies. It is highly interdisciplinary and only the proposed multi-faceted approach provides a means for a step change towards 'next-next' generation biopolymer sequencing in the future. As target analytes, we plan to use the four DNA nucleotides A, T, C and G (as mono-phosphates). Other analogues such as methylated cytosine and oxidized bases, which play important roles in disease diagnostics, could be included in follow-up projects. The proposed project is formally split into two, fully integrated work packages (WP):

A) Tunnelling detection of target analytes (TA, AV). Target DNA nucleotides will be dissolved in aqueous (phosphate) buffers or organic solvents such as mesitylene at concentrations in the low microM range. Too low concentrations result in very long recording times, while too high concentrations can make impossible to detect 'stochastic', individual binding events (crowding). Using tunnelling current-time recordings at given tip/substrate distances in an STM configuration, several 1000s of individual bindings events will be detected for each analyte (typical event frequency $\sim$ 1 Hz) and then analysed using the AI techniques in part B). Experimental parameters that can be varied to optimise the detection performance include the surface coating of the substrate and tip, the applied potential, the tip/substrate bias voltage and the surface chemistry (via molecular adsorbates), as appropriate. For training, the solution will only contain a single type of nucleotides, but mixtures will be employed during the actual prediction tasks.

B) Development and validation of DL-enhanced biosensing (EA, AV, TA). For the analysis of the data obtained in A), we will use CNs to learn representations of the tunnelling data that are most discriminative to classify the respective target analytes. The CN will be trained via dynamic routing to minimise classification error, involving a large training dataset of tunnelling signals from known biomolecules (cf. A)). For new unseen data, the CN will produce a classification that identifies the biomolecule from the tunnelling signal. Alternative DL architectures and heuristics will also be explored in accordance to the type, quantity and quality of data produced in part A), namely, combinations of CNs and Recurrent Neural Networks (Long Short-Term Memory networks in particular) which are ideally suited to capture the dynamics of sequential data as well as autoencoders in the case of unlabelled data. Special attention will be given to the optimization of hyper-parameters including number of layers and units per layer, activation functions (ReLU and variations), and regulation techniques (e.g., subsampling).

6.2 AI Methodology

CNNs are a Deep Learning variant of Artificial Neural Networks which have become very popular in image recognition and classification and are based on the application of convolutional filters to inputs. Using the backpropagation algorithm, CNNs learn filters to extract specific features from an input image. Once trained, feature representation in the network typically increases in abstraction for each successive layer. While the use of receptive fields maintain spatial information through the layers, much contextual information from previous layers is lost through max pooling. This leaves CNNs vulnerable to adversarial examples which may contain a set of required component features, but with other incompatible characteristics. This could be detrimental for the data we have generated in this project, because tunnelling data can be very sensitive to environmental noise (mechanical, electric).

To overcome this drawback, we planned to use CNs instead: A key innovation of CNs is a routing mechanism known as "routing by agreement". This routing of activation signals (coupling coefficient) from child to parent capsules is calculated iteratively using similarity between capsule outputs: children with similar properties to the parent capsule will see their coupling coefficient iteratively increase, while connections between dissimilar parents and children will weaken. In other words, for every new sample presented to the model, each layer will perform some reconciliation between bottom-up sensory input and the top-down expectation. This mechanism has clear applicability in the context of predictive coding, which posits that inputs are being compared to prior expectations at every level of the system, generating model adaption.

We have used CNs to learn representations of the tunnelling data that are most discriminative to classify the respective target analytes (here: DNA nucleotides A, T, G and C). The CN was trained via back-propagation to minimise classification error, involving a large training dataset of tunnelling signals from known biomolecules. For new unseen data, the CN has produced a classification that identified the biomolecule from the tunnelling signal. Depending on the type, quantity and quality of the data generated in WP1, we may also explore alternative Deep Learning architectures and heuristics. This includes Wave Nets, a generative model that is fully probabilistic and autoregressive and uses dilated causal convolutions to model complex joint probability distributions over raw data with very high sample rates; Recurrent Neural Networks (Long Short-Term Memory networks in particular) which are ideally suited to capture the dynamics of sequential data as well as autoencoders in the case of unlabelled data. Special attention was given to the optimization of hyper-parameters including the number of layers and units per layer, activation functions, and regulation techniques.

In relation to the Transfer Learning work (feature recognition) the following approach has been applied. To realize automatic recognition of the nucleotides from the I(t) signal, a Capsule Network was built based on the 'Dynamic Routing Between Capsules' paper (G. Hinton et al., 2017) and implemented in Python using the Pytorch framework. As an input of the network, the slices of the original trace with 500 data point were used. The structure of CN is mimicking the architecture of CNN for nucleotide detection (Albrecht, T. et al. Nanotechnology 28, 2017). The encoder of the network includes two capsule layers which inter-capsule routing. The goal of the routing is to make hierarchical connections between features extracted at previous layers of the network.

7 Interim Results

With simulated current-time data, a CNN and a Capsule Net were shown to yield similar results.

- Because of the possibility to extract spatial information of the features in data, the Capsule Net did not require pre-processing of the data. As a consequence, Capsule Net is more steady towards edge effects (when the only event in the trace occurring in the very beginning or very end of the trace). Secondly, Capsule Nets could be a promising technique for the classification of mixed data, where individual trace contains events from different nucleotides, and analysing the capsule states every event can be related to its class (to be studied additionally).
- Because of larger number of computing operations, Capsule Nets appear to be more computationally expensive for training, but being trained, the prediction is almost as fast as using CNN with similar architecture.
- In related work, we could show that Transfer Learning can enhance the performance of Autoencoders in unsupervised classification tasks for single-molecule charge transport data.

This work has been summarised in a recent pre-print and is currently under consideration in the IOP journal Machine Learning: Science and Technology (see below).

8 Outputs

Publications:

- Anton Vladyka, Tim Albrecht et al., "A Comparison of Capsule Net and Deep Convolution Neural Networks during the Single-Molecule Detection of DNA Nucleotides" (in preparation)
- Anton Vladyka, Tim Albrecht, "Unsupervised Classification of Single-Molecule Data with Autoencoders and Transfer Learning", arXiv preprint arXiv:2004.01239 (https://arxiv.org/abs/2004.01239); and manuscript under consideration at MLST (IOP).

Talks:

- Tim Albrecht, AI3SD Science Network Discovery+ meeting 2019, oral presentation (http://www.ai3sd.org/conference2019/agenda)
- Upcoming: Online seminar on Molecular Electronics, organised by researchers at the University of Liverpool (date TBC)

9 Conclusions

- Capsule Nets can outperform CNNs in the recognition of tunnelling events in some circumstances, especially when detection events occur at the egdes of the recording window and are thus incomplete. Hence, CapsNets seem to be more robust in this regard.
- When corrected for this effect, the performance of CapsNets and CNNs was however comparable and in the region of 90-95%, similar to our previous work reported in Nanotechnology in 2019.
- We also explored new applications of Machine Learning in single-molecule science, namely in the unsupervised classification of single-molecule charge transport data. Interestingly, Image Recognition Networks, previously trained on unrelated image data, were able to extract features from the data and identify previously unknown sub-populations. This is important, because this finding relaxes the requirement of using large amounts of task-specific data for training, which might not always be available.

10 Next Steps

In terms of future funding applications and our longer term plans, our primary target would be EPSRC, but given the rapid dynamics of the sequencing market potentially also venture capital funding, where TA already has existing links. In this context, it will be key to demonstrate superior performance based on experimental data and also compared to currently established sequencing technologies, in order to demonstrate that a step change is indeed within reach. Further experimental work to test the performance of the approach is currently ongoing, as part of a PhD project.

11 References

Please see main text above.

12 Data & Software Links

Our Transfer Learning work is currently under consideration at a peer-reviewed journal, but a pre-print is also available here: arXiv preprint arXiv:2004.01239. This pre-print also includes a step-by-step guide to using Image Recognition Networks for feature extraction in charge transport data (see Supporting Information).

The Capsule Nets used in this project was implemented in Python using pyTorch framework for deep learning, based on publicly available implementation https://github.com/gram-ai/capsule-networks. The generator of simulated nucleotide tunneling was also implemented in Python. The Matlab version is available upon request (TA) and was also used here: T Albrecht, G Slabaugh, E Alonso, SMMR Al-Arif, Nanotechnology 2019 28 (42), 423001 (https://iopscience.iop.org/article/10.1088/1361-6528/aa8334)