To William Demant Fonden

Project plan for external research stay in MLG, Cambridge

To Whom it May Concern,

For my external research stay I have ideas for combining deep generative models (DGM), uncertainty estimation, multi-task learning and active learning (AL) for molecular exploration and detection through Raman spectra as a molecular fingerprint. This is mainly inspired by the work of my host-supervisor, Assistant Professor, José Miguel Hernández-Lobato, on ChemVAE, GVAE, BNN-LV, PBP, the recent work on Deep Learning Spectroscopy by my supervisor, Associate Prof. Mikkel N. Schmidt and my own work w. Prof. Ole Winther developing a Gaussian-mixture-VAE for gene expressions and exploration of cell types in latent space. These ideas lead to Miguel, my supervisors and me agreeing on and planning a 4 months research project in the Machine Learning Group, University of Cambridge, which now needs additional external funding from you.

4 June 2019
MV

The project for this 4 months external stay is a part of my PhD project in Active Deep Learning for Nano-Sensor Systems, which is described in detail below, and will therefore be focused on developing deep learning models for Raman spectroscopy, which is the most commonly used technique for detection of molecules. Therefore I will here motivate Raman spectroscopy and explain why a data-driven approach will benefit the research and industry in pharmaceutical engineering.

The first research question is: Can determine and detect the fingerprint of the building blocks of nature itself in a smarter more data efficient way? The hypothesis is: We can use deep neural networks for learning the underlying process behind how light is scattered by linking the Raman spectra and various representations of molecular structure. Learning how to generate the fingerprint of molecules, will make it possible to detect the molecules, their concentrations and properties as medicine contained in micro-containers by non-invasive methods, e.g. lasers.

Raman scattering stems from driven molecular vibration coming from symmetric and asymmetric stretching of bonds between atoms in molecules, when the incident laser hits the molecule. The intensity is below 0.01% of the direct Raleigh scattered light, so it needs filtering and the signal-to-noise-ratio is low, which also argues for using Surface-Enhanced Raman Spectroscopy (SERS) or Coherent anti-Stokes Raman scattering spectroscopy (CARS). SERS leads to a very complex behaviour and makes it hard to simulate with DFT. The frequency of the Raman scattered light (Raman bands) will depend on the strength of the atomic bonds and atomic masses and can be modelled in the time domain with an ordi-

DTU Compute
Department of Applied Mathematics
Computer Science
www.compute.dtu.dk

Building 321, Room 106
Richard Petersens Pl,
2800 Kgs. Lyngby, Denmark

Dir. +45 61603013
Tel.. +45 45 25 30 31 and
maxvo@dtu.dk

nary differential equation like Hooke's law and Newton's 2nd law of motion. Through quantum mechanical equations the differential functional theory (DFT) can be used to simulate Raman spectra, but for complicated molecular compounds and environment like SERS, the simulation rarely fits with real world measurements and will often be re-adjusted manually to fit this. This suggests using a data-driven approach by generating spectra through deep graph neural networks, which are commonly used for linking molecular properties to the molecular structure.
Here is the representations of molecules and what we need for predicting molecular properties and Raman spectra. We need both molecular structure, atomic mass and band strength to determine vibrational modes and differential equations, so here is a list of these commonly used measures:

- SMILES (Text representation of molecules)
- Coulomb matrix (Energy interaction based on distances between all atoms)
- Bag of bonds
- Histograms
- Radial distribution functions
- Chemical environment
- ACSF

We can find these resources at:

- https://www.researchgate.net/post/Free_Database_with_Raman_spectra
- https://www.ccdc.cam.ac.uk/solutions/csd-system/components/csd/
- https://serc.carleton.edu/research_education/crystallography/xldatabases.html
- http://oqmd.org/
- http://quantum-machine.org/datasets/

At the IDUN Center of Excellence, researchers can as well provide huge amounts of data for our project through their ongoing experiments with Raman spectroscopy on micro-container drug-delivery. With computational resources at both University of Cambridge and DTU Compute, we can scale up to big data and thereby the representational strength of our deep learning models tremendously.

For this project the goal is for me to be the main author on two publications in high-impact conferences or journals and will be conducted in collaboration with:

Host-supervisor:	José Miguel Hernández-Lobato - Assistant Professor
Principal supervisor:	Mikkel Nørgaard Schmidt - Associate Professor
Co-supervisor:	Tommy Sonne Alstrøm - Senior Researcher

Best regards,

Maximillian Fornitz Vording
PhD Student
Section for Cognitive Systems