Predicting physical properties of straight chain hydrocarbons using machine learning

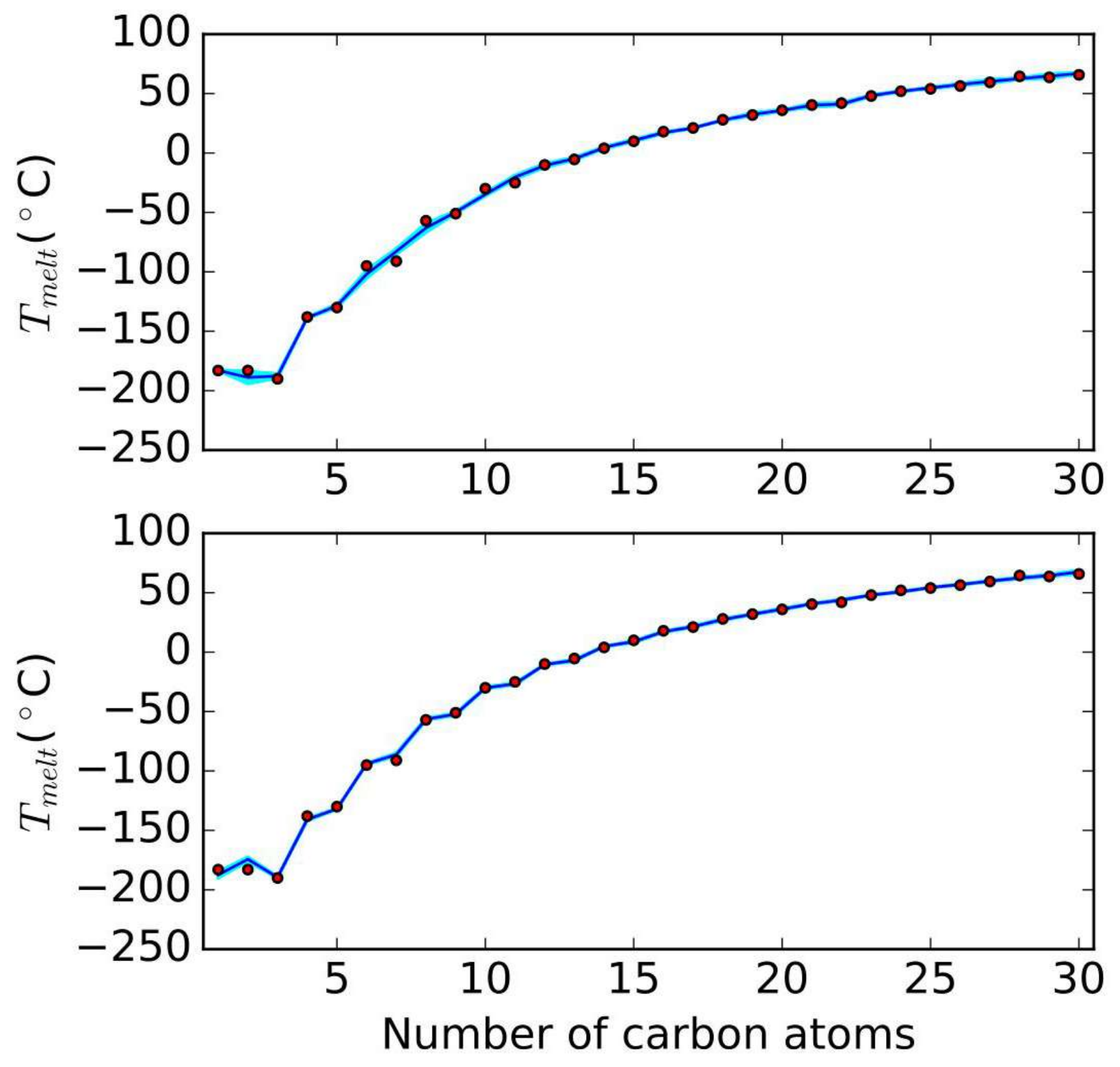
Pavao Santak, Department of Physics (TCM),ps727@cam.ac.uk

Introduction

Many real life physical systems are data driven. However, most of data sets describing those systems are highly fragmented. Historically, drawing inferences from fragmented data sets has been difficult. A new approach based on neural networks and developed by Dr.Gareth Conduit enables us to draw inferences from fragmented data sets. We apply this new approach to predict the physical properties of straigh chain hydrocarbons.

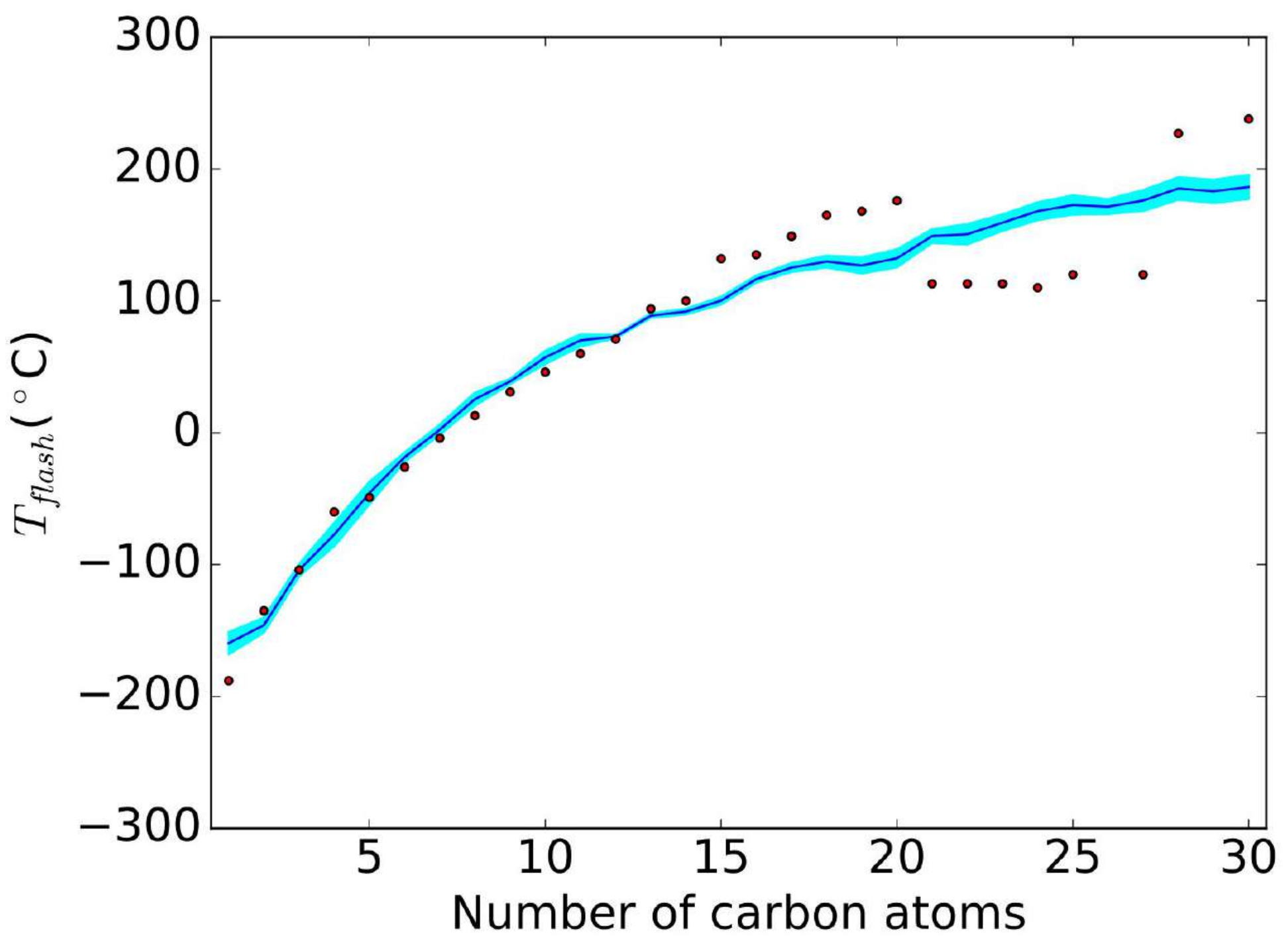
Capturing simple correlations

By identifying appropriate molecular descriptors, our neural network models are able to accurately reproduce the melting point curve. We access the accuracy of our models through cross validation, for which we get R^2=0.9992. Without identifying appropriate molecular descriptors, accuracy of our neural network models is still very high, (R^2=0.9987), but the models failed to capture the step behavior in melting point.



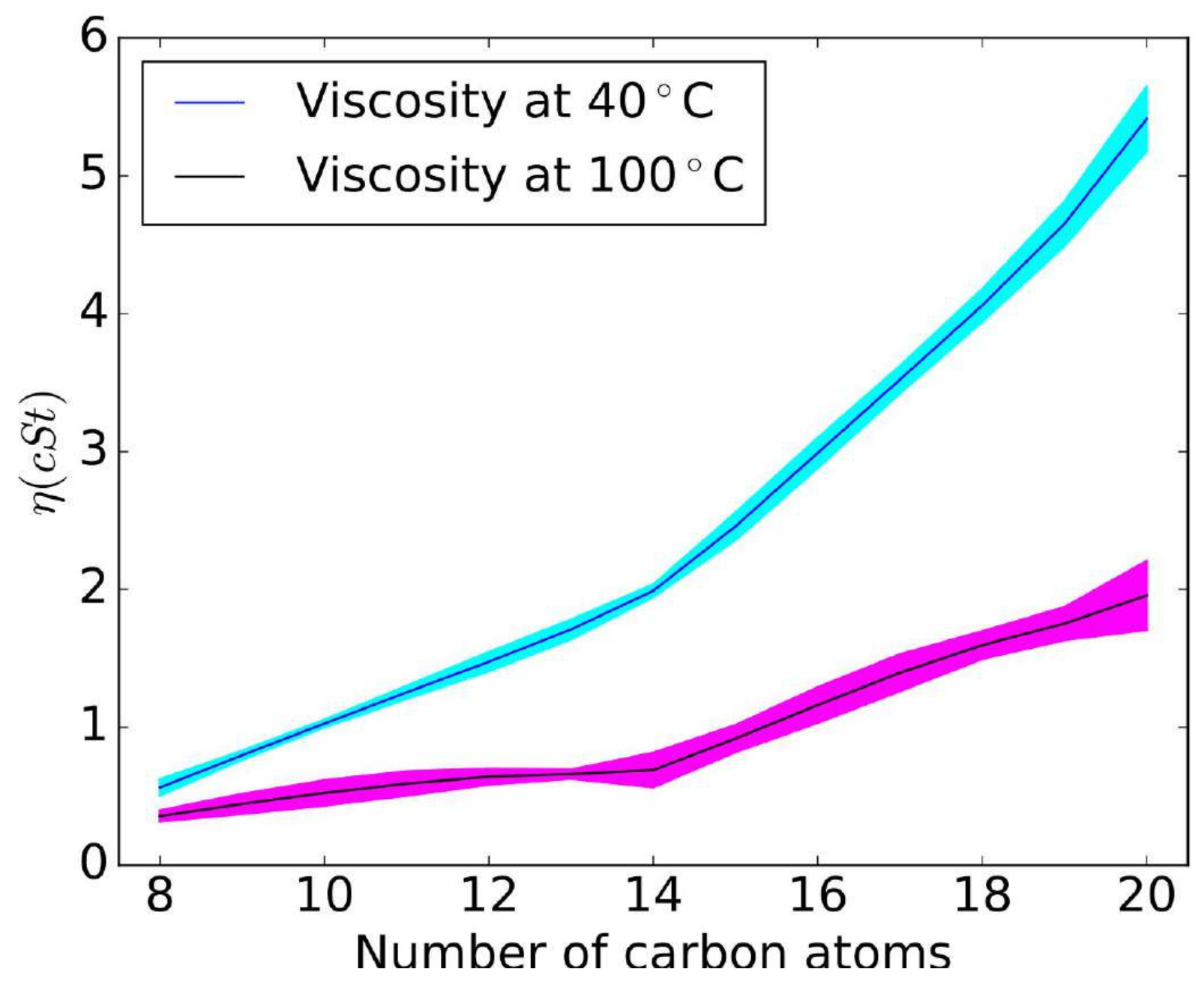
Detecting erroneous entries

Our neural networks are able to detect erroneous data entries. We were able to detect erroneous data entries for flash point. R^2 value we obtained is again high (0.9100), but we see that many of experimental data entries are more than one error bar away from our predictions.



Predicting kinematic viscosity

Our neural networks can learn from incomplete data. We were able to predict kinematic viscosity from incomplete density and dynamic viscosity data.. We collected data for dynamic viscosity and density as a function of temperature and pressure and built neural network models whose accuracy we confirmed through cross validation. Using predicted values for dynamic viscosity and density, we predicted kinematic viscosity and its uncertainty at 1 atmospheric pressure and 40°C and 100°C, which are temperatures of



We predict physical properties of hydrocarbons using machine learning. We show that we can capture the simplest of correlations by introducing appropriate molecular descriptors as input features in our neural network models. In addition, we show that our neural network models are capable of detecting erroneous data entries. Finally, we work with a fragmented data set for viscosity and density as a function of temperature and pressure and show that our neural network models can fill in the gaps in our knowledge.

Pavao Santak acknowledges financial support from BP plc and thanks Dr.Gareth Conduit for useful discussions .