cecam Centre Européen de Calcul Atomique et Moléculaire	**Workshop Scientific Report** Please do not repeat the program (unless there were last-minute changes) or the initial description - we already have this material.
Title	Machine learning in atomistic simulations
Organizers	Gareth Tribello Jörg Behler

Scope of the workshop (one-two paragraphs)

Atomistic simulations are a standard tool in chemistry and physics as they allow us to use accurate models, often based on quantum mechanics, to describe a variety of chemical and physical processes. For instance, using these tools important insight can be gained into phase transitions in materials, nanotechnological devices and even the biomolecules that make up the human body.

One of the main challenges in modern simulation is to take the data that emerges from our simulations, which is inevitably high-dimensional as it describes the motions of thousands of atoms, and to derive simpler theories that describe the underlying physical principles. This is necessary in part because human beings cannot understand high-dimensional data. More importantly, however, the most accurate theories for describing atoms and molecules are computationally very expensive, and they can only be used to perform relatively short simulations of small systems. Recently, a key paradigm in the simulation community has thus been to use the relatively limited data that one can extract from short simulations using expensive calculations to develop new, cheaper models that can be used to understand phenomena that take place over timescales and length scales that are considerably longer. This approach can be very time consuming, and to address this problem a recent development has been the use of machine learning algorithms to perform these sorts of calculations. Prior to this conference we identified three areas where this approach seemed to be particularly promising:

a) The use of dimensionality reduction to understand high-dimensional trajectory data
b) The use of clustering and dimensionality reduction algorithms in developing better methods for enhanced sampling
c) The use of neural networks and gaussian processes in interpolating potential energy surfaces calculated using an accurate quantum mechanical reference method

During the three day program of the conference we thus endeavored to discuss topic (a) on day 1 (10th Sept. 2012), (b) on day 2 (11th Sept. 2012) and (c) on day 3 (12th Sept. 2012). Furthermore, where possible, we endeavored to incorporate researchers from chemistry and physics who are experts on the particular problems encountered in these fields together with applied maths researchers more involved in the development of the methods that have begun to be used in chemistry and physics.

Main **outcomes** of key presentations (one page)

In what follows the outcomes of for each of the themes of the conference are discussed separately.

Dimensionality reduction for trajectories

The talks in the conference showed that machine learning can generate useful tools for visualizing trajectories and that often these approaches can generate physically-meaningful, low-dimensional models for the high-dimensionality trajectory data. These tools either divide the trajectory frames into a relatively small number of disjoint sets (clusters) or they produce a low-dimensionality map that describes how the points are distributed across phase space. Some particularly interesting approaches that were presented are diffusion maps, sketch-map and tree-based clustering. In addition, the value of collaborating with statisticians was illustrated by the talk given by Albert Sheih who showed how the quality of any given clustering could be tested and how Bayesian statistics could be used for clustering. Lastly, the talks by Ilia Horenko and Michael Bronstein demonstrated how the latest ideas from applied maths could be used to refine the methodologies we are using.

Enhanced sampling

During the second day of the conference the focus was on algorithms that can be used to accelerate the rate at which phase space is explored during molecular dynamics simulations. Some of the tools presented under this heading work by introducing a bias potential that flattens the underlying free energy landscape so that kinetic traps are removed. Others work by exploring a space of transition pathways or by using a large number of short trajectories. This second class of methods is advantageous because, unlike methods based on a bias, a small number of collective variables are not required. That said the conference demonstrated that progress is being made in developing biassing methods that can work with large number of collective variables. In particular, Gaussian Mixture Umbrella Sampling, reconnaissance metadynamics and path metadynamics all generate high-dimensional bias potentials. Furthermore, impressive applications to complex systems were presented for all these methods. We also heard about new developments to biasing algorithms such as adaptive hills for metaydnamics, field-metadynamics and the use of non-linear dimensionality reduction coordinates as CVs.

Potential fitting

The construction of reliable interatomic potentials is a mandatory requirement if the results from atomistic simulations are to be meaningful. Consequently, a lot of effort has been put in the development of a huge number of different approaches to represent potential-energy surfaces. Still the reliability and numerical accuracy for most conventional potentials based on physical approximations is not satisfying. A lot of progress has been made in recent years in applying machine learning techniques such as artificial neural networks or Gaussian approximation potentials. In these methods a database of structures and corresponding QM energies are used to generate an interpolating potential. A number of progress reports have shown that the range of chemical problems that can be addressed using this approach has been increased significantly in recent years. Marco Bernasconi's talk which discussed how this approach could be used to examine phase-change materials was particularly impressive.

Report on selected discussions (one page)
eg. Were there interesting hints for new research? for new developments? for collaborations?

We did not place a formal discussion session in the conference, but there were extended and very fruitful informal discussions after the talks, in the coffee breaks and at the poster session. While it is not possible to give a comprehensive overview of all these discussion, in the following sections several highlights will be listed.

Dimensionality reduction for trajectories

Many of the trajectory analysis algorithms that were presented at the conference rest on assumptions about the nature of the input data. This is not necessarily a problem if these algorithms are just being used to generate attractive visualization of the high-dimensionality data. However, the validity of these assumptions needs to be further tested when these algorithms are used to extract physical quantities. In addition, more work needs to be done to convince the rest of the community that these complex forms of analysis are worthwhile. In this regard the fact that many open source analysis tools have been developed is a good first step. Lastly, the talk by Dellago suggested that there are many problems outside the remit of biochemistry, which is where the use of these methods is most established, where these algorithms could play a useful role.

Enhanced sampling

The conference suggested that the best way to improve the current crop of enhanced sampling is to use multiple methods at once. For instance, collective variables for methods such as metadynamics can be generated using data obtained from transition path sampling. Some particularly interesting examples of this are the path metadynamics method presented by Bernd Ensing and the CV likelihood methods presented by Baron Peters. Further refinement of these ideas is essential if we are ever to use these ideas to examine both kinetics and thermodynamics.

Potential fitting

The present workshop was the first time that a large number of researchers working in this field met in one placed. This ensured that the discussion was very stimulating for all participants in the workshop. Comparing the different approaches but also mainly research interest-guided applications gave a lot of insight into possible future research directions. All machine-learning-derived potentials are very accurate and unbiased. However, there are still a number of different mathematical approaches being used in the field. The two main techniques are based on systematic but truncated many-body expansions, which are very well suited for molecular systems, and potentials based on local structural features, that are also applicable to systems with strong many-body character such as solids and in particular metals. The most valuable outcome was probably the possibility to discuss and compare a number of technical aspects, since many machine learning-based potentials face similar conceptual challenges.

To what extent were the **objectives** of the workshop achieved (strong points, weak points)? (one paragraph at least)

We believe the workshop was very successful and we got a lot of very positive feedback from the participants. Our aim to bring together experts from very different fields which typically do not meet since they are working in different scientific communities was very fruitful, because there are many common problems and solutions that can be transferred from one community to another. Furthermore, everyone was very impressed by the wealth of different methods that are being applied to these difficult problems. In addition, a great deal of effort seems to have been put into distributing methods to the community, which is superb. One slight disappointment was that we were not able to attract many researchers from maths department. In addition, the workshop was rather technical so we imagine that it would have been difficult for someone not already working in this area to learn how best to apply what was discussed to their own particular problems. That said, at this early stage of development of the field, a technical conference was probably what was required as it allowed for an appraisal of the relative merits and demerits of the various approaches that are being tried.

Do you have suggestions for new workshops/tutorials/conferences on the topic?

During the conference a number of the attendees approached us and suggested that we should repeat this conference again in a few years. The fields examined during this conference are growing rapidly and there is a lot of interest from the wider community. This second fact was evidenced by the large number of attendees we had in our workshop. We were, in fact, oversubscribed and a number of applicants could not participate because of the limited number of places.

During the workshop we found out that the researchers developing many of the tools discussed during the workshop are putting a lot of effort into sharing their research with the wider community. There was much talk of websites, open-source code repositories and so on. Given that the methods being used in this field are not generally taught during physics or chemistry degrees there may well be merit in organizing a tutorial where the developers of these tools can teach other researcher how to use the tools that they have developed and shared with the community.