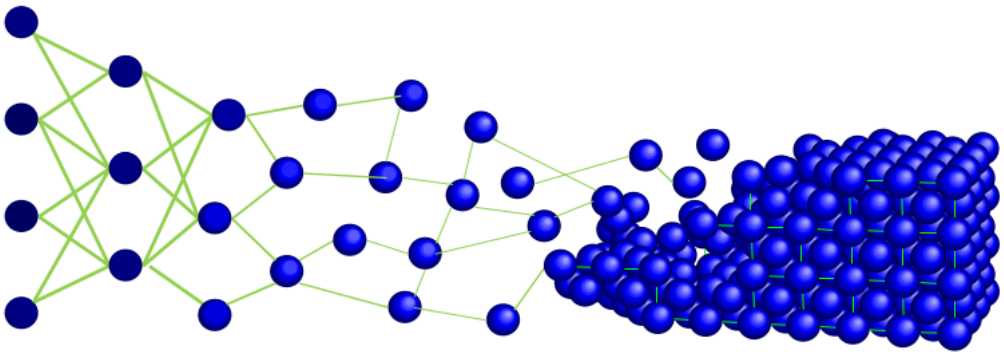
Workshop « L'intelligence artificielle pour la chimie des matériaux », ICMPE, Thiais, 25 Septembre 2018

Recueil des résumés

https://ai4mater-sci.sciencesconf.org



1. Présentations orales

9:40-10:10

Application of Machine Learning for Materials Science

Nataliya Sokolovska[1]

[1]Sorbonne Université, Paris

The number of applications of data mining and machine learning for chemistry and materials science increases rapidly. There is a hope that recent developments in machine learning will accelerate the progress in materials science, that they will help to generate novel materials, and to explore their properties. We will consider three big families of machine learning methods: supervised learning, unsupervised learning, and reinforcement learning, and the state-of-the-art prediction methods such as linear, non-linear, and deep classifiers. We will discuss feature engineering including feature construction and feature selection. I will provide a brief overview of a machine learning pipeline: data (feature) representation, training a model, and testing it in collaboration with human experts.

10:10-10:30

Machine Learning and High-throughput Computational Screening

Ambroise van Roekeghem[1]

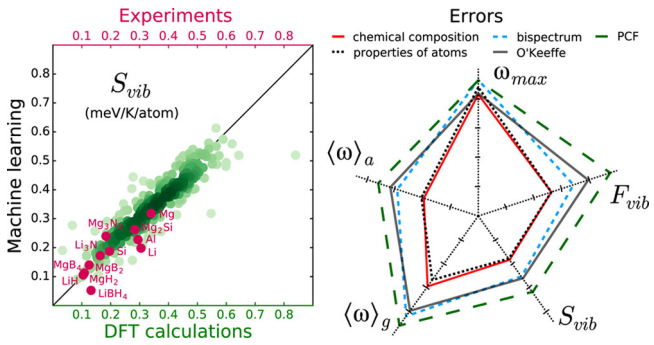
[1]CEA, LITEN, 17 Rue des Martyrs, 38054 Grenoble, France

The new field of high-throughput ab-initio materials modeling has been raising considerable interest in the last decade. This is due to the conjunction of two recent developments: the increasing availability of computational data, and the achievements obtained by machine-learning methods based on such large datasets. While the search for new, better energy materials has been a long-standing issue, this new route promises to accelerate drastically the discovery of new materials. Several experimental confirmations have already demonstrated the potential of those *in silico* predictions.

High-throughput computing has been originally pushed forward by groups in the U.S., giving birth to large databases such as the Materials Project or AFLOWlib. Europe is now catching up, with groups at the state of the art in different countries and large European projects such as the NOMAD and AiiDA repositories. At present, several challenges remain to be solved. For example, to improve the prediction of the (meta-)stability of compounds,

notably at finite temperature; to accelerate more advanced electronic structure methods so that they can be used high-throughput; or more practically, to tighten the links with experiments and industrial developments.

In this presentation, I will introduce the field of high-throughput computational screening, show a few examples of how simple machine learning techniques are currently used in this field, and discuss some of the challenges ahead.



REF: F. Legrain et. al, *How Chemical Composition Alone Can Predict Vibrational Free Energies and Entropies of Solids*, Chem. Mater. **29**, 6220 (2017)

Computer-assisted design of complex metallic alloys

10:30-10:50

Edern Menou[1], Emmanuel Bertrand[2], Jérémy Rame[1], Clara Desgranges[1], Gérard Ramstein[3], Franck Tancret[2]

[1]SAFRAN Tech – SAFRAN (FRANCE) – France

[2]Institut des Matériaux Jean Rouxel (IMN), Université de Nantes, CNRS : UMR6502 – France

[3]Laboratoire d'Informatique de Nantes-Atlantique, Université de Nantes, CNRS : UMR6241 – France

The design of metallic alloys incorporating many elements such as nickel-based superalloys and high-entropy alloys is not trivial. The high number of alloying elements not only results in a very large space of producible alloys, but also makes any physical metallurgy-based modelling of physicochemical properties hard to formulate or validate due to the numerous interacting parameters. In order to solve these issues, a computer-assisted design method is proposed. It circumvent the difficulty of finding promising alloys amongst the huge set of all possibilities by using either genetic algorithms or a carefully bounded systematic grid search approach. Several data-intensive models are employed in order to guide this search for high performance alloys. On the

one hand, the method of computer-coupling of phase diagrams and thermochemistry (CALPHAD) enables the determination of thermochemical features of an alloy given its composition (e.g., constitution, high-temperature stability, weldability). On the other hand, Gaussian process regression is used to model and provide estimates of various thermomechanical properties as a function of temperature and alloy composition (e.g. yield stress, tensile strength, creep life). The combination of these tools has been exploited for designing high performance alloys of different classes, including, for gas turbines, nickel-based superalloys for disks and blades. The approach has also been applied to the design of high entropy alloys. In each use case, alloys are found whose combination of properties surpasses that of existing alternatives.

11:05-11:35

Generalized Stochastic simulation algorithm for Artificial Chemistry

Hedi Soula[1]

[1]Sorbonne Université, Paris

Artificial chemistries (AC) are useful tools and a simple shortcut for the study of artificial life. In many works, ACs are quite straightforward or simplistic or highly unrealistic (or all combined) but in several works AC are extremely complex. Among them, we focus of Hutton Artificial Chemistry HuAC where reactions act on the nodes of a graph (so-called the atoms) where the connected components composed the actual molecules of the environment. The main works from Hutton are based on a 2D simulator (squirm) with auto-replication and several other properties. This paper proposes a computation framework and software that cancel the need for 2d space simulation in the HuAC while keeping a lot of the features of this chemistry. It relies on the Stochastic Simulation Algorithm that has been here adapted to work on graph structure. In order to test it, we simulated Huttons auto-replication which relies heavily on strong spatial inter- actions in a spaceless environment. In addition, due to the increase in performance, we develop some preliminary work on Random Chemical Worlds where reactions are randomly selected. We showed on simple metrics that the fraction of reactions among all possible is a general parameter that acts on the system similarly to a phase transition.

Prédire une géométrie moléculaire convergée par des modèles d'apprentissage automatique

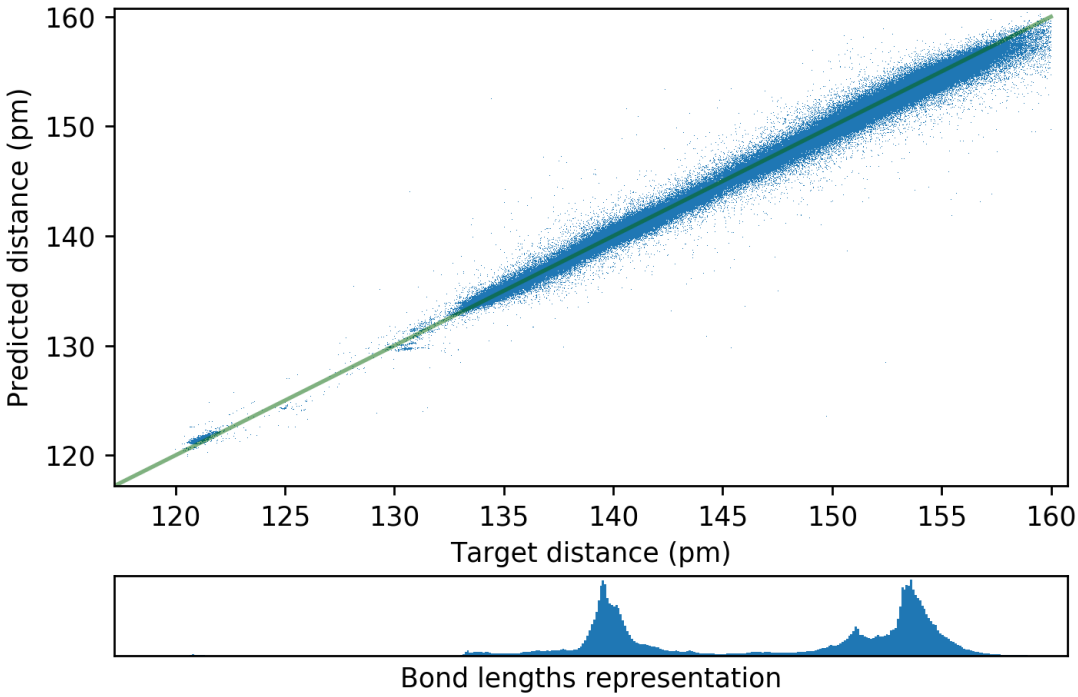
11:35-11:55

Thomas Cauchy[1], Benoit Da Mota[2]

[1]Université d'Angers, MOLTECH-Anjou,

[2]Université d'Angers, LERIA

Molecular chemistry is defined as the science of molecular (discrete) entities and its researchers form the largest community in Chemistry. Today, hundred of millions of molecules are known. In their vast majority, they are composed of less than a hundred of nuclei and less than a thousand of electrons. The chemical properties of such molecules depend of the electrons localization. Such properties can only be computed by approximate methods. Since the democratization of scientific computing, calculations in chemistry has become an essential part of every research. Without any collaborative information system, tons of redundant results are generated and then deleted each year. Such discarded information could have been employed for others studies. There is a blatant lack of curated data in open access associated with artificial intelligence or statistical models for the exploration and exploitation of these data. We present QuChempedia, a quantum (molecular) chemistry collaborative encyclopedia and its first artificial intelligence application: prediction of converged molecular geometry with neural networks.



11:55-12:15

Generative Adversarial Networks for Finding New Crystal Structures

Asma Atmna[1,2], Asma Nouira[1], Jean-Claude Crivello[1], Nataliya Sokolovska[2]

[1]ICMPE-CNRS, Thiais

[2]Sorbonne Université, Paris

Our motivation is to propose an efficient approach to generate novel stable metal hydrides for hydrogen storage. This combinatorial problem is handled in practice through combinatorial DFT calculations. Therefore, it can take many hours of human experts to construct and evaluate new data. Unsupervised learning methods such as Generative Adversarial Networks (GANs) can be used efficiently to produce new data, and have shown promising results in image processing applications. In this talk, we illustrate how GANs can be used to generate new chemically stable crystallographic structures with increased domain complexity. We present a model inspired by a cross-domain GAN (DiscoGAN) and test our approach on two pseudo–binaries systems: (Pd,Ni)–H and (Mg,Ti)–H.

13:35-14:35

Machine Learning Like a Physicist

Michele Ceriotti[1]

[1]COSMO, EPFL, Lausanne, Suisse

Statistical regression techniques have become very fashionable as a tool to predict the properties of systems at the atomic scale, sidestepping much of the computational cost of first-principles simulations and making it possible to perform simulations that require thorough statistical sampling without compromising on the accuracy of the electronic structure model. In this talk I will argue how data-driven modelling can be rooted in a mathematically rigorous and physically-motivated framework, and how this is beneficial to the accuracy and the transferability of the model. I will also highlight how machine learning - despite amounting essentially at data interpolation - can provide important physical insights on the behavior of complex systems, on the synthesizability and on the structure-property relations of materials. I will give examples concerning all sorts of atomistic systems, from semiconductors to molecular crystals [1], and properties as diverse as drug-protein interactions[2], dielectric response of aqueous systems[3] and NMR chemical shielding in the solid state[4].

1 F. Musil, S. De, J. Yang, J. E. J. E. Campbell, G. M. G. M. Day, and M. Ceriotti, Chem. Sci. 9 (2018) 1289

2 A. P. A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. J. R. Kermode, G. Csányi, and M. Ceriotti, Sci. Adv. 3, (2017) e1701816

3 A. Grisafi, D. M. Wilkins, G. Csányi, and M. Ceriotti, Phys. Rev. Lett. 120 (2018) 36002

4 `http://shiftml.org`

How many materials are left to discover? An exploration of quaternary space

14:35-14:55

Michael Sluydts[1,2], Michiel Larmuseau[1,2], Titus Crepain[1], Karel Dumon[1], Kurt Lejaeghere[1,3], Stefaan Cottenier[1,2]

[1]Center for Molecular Modeling, Ghent University – Belgique
[2]Department of Electrical Energy, Metals, Mechanical Constructions & Systems, Ghent University – Belgique
[3]Department of Applied Physics, Ghent University – Belgique

The frontier of materials science is shifting evermore towards the development of 'exotic' functional materials, which display an unfamiliar combination of properties. The underlying behavior giving rise to these materials' properties is often too complex to predict purely from their crystal structure. New exotic materials are thus largely developed bymimicking existing materials, inevitably introducing bias.

While truly new exotic materials are likely to exist in unknown regions of materials space, it is unlikely we will find them through biased exploration. At the same time, random exploration is unsustainable given the time required to synthesize and characterize new materials. Several questions thus arise. How can we explore the vast materials space intelligently, yet without bias? And perhaps most importantly: how many materials are left to discover?

We investigate this fundamental question by creating a database of hypothetical crystals in quaternary space, where experimental exploration is limited. By employing highthroughput ab initio methods, we are able to predict various properties of these unknown materials, including their stability. Furthermore, applying machine learning during the screening procedure yields a ten-fold speedup over brute-force exploration. This yields a relatively unbiased, yet fast exploration method.

By comparing the discovery rate, composition and structure of the new materials with that of experimentally known quaternary phases, an estimation can be made of how many materials are yet to be discovered within this region of materials space.

14:55-15:15

Evaluating a linear machine learning force field for aluminium

Maarten Cools-Ceuppens[1], Toon Verstraelen[2]

[1]Center for Molecular Modeling, Ghent University – Belgique

[2]Center for Molecular Modeling, Ghent University (CMM) – Technologiepark 903 BE-9052 Zwijnaarde, Belgique

Molecular dynamics is a well-established tool to study structural and dynamical properties of a broad class of materials. The potential energy surface (PES) and its derivatives (the forces acting on the nuclei) are the key ingredients in those simulations. Ideally, one should use ab-initio techniques, which approximate the exact many-body Schrödinger equation, to construct the PES. In practice this is not feasible for extended systems or long simulation times, due to the increasing computational effort. For this very reason, machine learning force fields are being developed. They allow a fast evaluation of the forces and energies of a molecular system at the accuracy of ab-initio techniques, which enables us to simulate multiple new molecular systems that were not accessible up till now.

Most state-of-the-art machine learning force fields (deep neural networks, kernel ridge regression ...) split the total energy into atomic contributions. Each of those atomic energies are learned based on atomic descriptors or features, which are invariant under rotations, translations and permutations of equivalent atoms. This ensures the conservation of translational and rotational momentum. Consequently, these atomic feature vectors serve as the input in neural networks or kernel based methods. For example, SchNet [1] (a deep neural network) expands the inter-atomic distances into a Gaussian basis while SOAP [2], (using Gaussian Aproximation Potentials), makes use of Spherical Harmonics to describe angular features and Gaussians to describe the radial part.

Here, we show that a simple ridge regression model (i.e. the atomic energy is a linear function of its atomic features) can contend with state-of-the-art machine learning force fields. As a test case, a force field for aluminium is constructed. The training dataset [3] consist of about 11000 different configurations (bulk, surfaces, clusters, dislocations ...) generated using multiple md-trajectories. Unlike the typical image-classification datasets, md-trajectories are highly correlated training sets. For this reason, some md-trajectories are excluded from this dataset and put together in an external validation set. Next, all the other data is shuffled and split in a training (90%) and test set (10%).

The simple linear model achieves mean absolute errors (MAE) around 0.03 eV / A, on the training and test set, outperforming conventional Embedded

Atom Models (0.08 - 0.20 eV / A). In comparison, deep neural nets like SchNet can halve our errors. However, they have external validation errors which are higher than the ones for our linear model. The same happens when the number of features increases, confirming the observations made in a recent paper [4], where selecting the most important features improve the training accuracy while combating overfitting. Hence, by simplifying our model and by decreasing the number of features, our force field is capable to generalize well, even when using highly time-correlated md-trajectories as trainig data.

1 K. T. Schütt *et al. J. Chem. Phys.* **148** 241722 (2018)

2 A. P. Bartók *et al. Phys. Rev. B* **87** 184115 (2013)

3 V. Botu *et al. J. Phys. Chem. C* **121** 511-522 (2017)

4 G. Imbalzano *et al. J. Chem. Phys.* **148** 241730 (2018)

Chemical space modeling and visualization with generative topographic maps

15:35-15:55

Gilles Marcou[1], Dragos Horvath[1], Alexandre Varnek[1]

[1]Laboratory of Chemoinformatics (UMR7140), CMC, Université de Strasbourg, Strasbourg, France

The Generative Topographic Mapping (GTM) algorithm is an unsupervised method to map high dimensional data to a two-dimensional representation [1]. In the process, the GTM builds a probabilistic model of the data that can be exploited for data characterization, comparison or classification and regression model building. The presentation aims to be an introduction to this methodology and is illustrated with several cases.

- QSAR and GTM study of electrolytic solvents: the viscosity, ionic conductivity and oxydation potential of carbonate and sulfate compounds was modeled (Figure 1). The most promising ones were investigated experimentaly as ingredients for the formulation of new electrolytes.

- The analysis of a chemical library of over 2M compounds from 36 chemical providers, plus the NCI [1b]. Through the use of an innovative methodology, the Generative Topographic Map (GTM), it is possible to get a detailed view and understanding of the intimate structure of the embedding chemical space, compare the chemical libraries of commercial providers and rationalize decisions for compound purchasing.

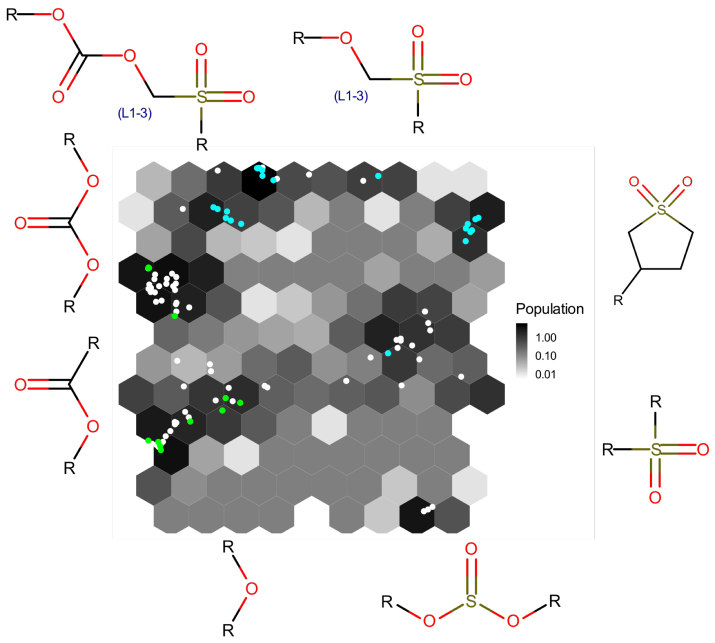


Figure 1: The density of the chemical space of electrolytic solvent ingredients. Dark areas are highly populated regions and light areas are low populated regions. The projections of the compounds on the map are represented by dots. The colored dots are the location of the newly synthesized electrolytic ingredients.

- The emergence of "Universal Maps" [2] [3]. These maps are trained over the ChEMBL and are able produce reasonable predictions over an ensemble of biological properties. It is shown how these maps are combined to improve their utility.

References:

1 (a) Bishop, C. M.; Svensén, M.; Williams, C. K., GTM: The generative topographic mapping. Neural computation 1998, 10 (1), 215-234; (b) Kireeva, N.; Baskin, I.; Gaspar, H.; Horvath, D.; Marcou, G.; Varnek, A., Generative Topographic Mapping (GTM): Universal Tool for Data Visualization, Structure-Activity Modeling and Dataset Comparison. Molecular informatics 2012, 31 (3-4), 301-312

2 Sidorov, P.; Viira, B.; Davioud-Charvet, E.; Maran, U.; Marcou, G.; Horvath, D.; Varnek, A., QSAR modeling and chemical space analysis of antimalarial compounds. Journal of computer-aided molecular design 2017, 31 (5), 441-451.

3 Sidorov, P.; Gaspar, H.; Marcou, G.; Varnek, A.; Horvath, D., Mappability of drug-like

Accurate deep neural network potential for predicting properties of solids

15:55-16:15

Anton Bochkarev[1], Ambroise Van Roekeghem[1], Natalio Mingo[1]

[1]CEA, LITEN, 17 Rue des Martyrs, 38054 Grenoble, France

Density Functional Theory is a very versatile tool which allows to compute multiple properties of materials. Nowadays, it is routinely applied for predicting, e.g., binding and cohesive energies of molecules and solid, electronic band structures, vibrational properties at 0K. Nevertheless, as the computational complexity of the DFT calculations scales non-linearly with the system size,the applications are often limited to the systems containing at maximum a few hundreds of atoms. It is therefore difficult to apply DFT for studying, e.g., solids with defects, properties of the materials at finite temperature, dynamical effects. To overcome these issues, the "classical" interatomic potentials are applied. Usually, these potentials approximate the energy of interaction between atoms by some kind of analytical function with the parameters which are adjusted to match experimentally known properties. The main disadvantage of these potentials is the lack of accuracy and transferability. The machine-learning technics provide the way to produce the interatomic potentials which are addressing deficiencies of the "classical" interatomic potentials while staying computationally efficient. We present an interatomic machine-learning potential trained on DFT calculations using artificial neural networks (ANN). Our algorithm simultaneously trains on the DFT data for energies and atomic forces. This leads to the more efficient utilization of the DFT data as well as an increased accuracy of the potential. Our machine-learning potential is also universal in terms of the size of the system and its chemical composition. We demonstrate its versatility and accuracy via computing various properties of solids and compare the results with direct DFT calculations.

16:15-16:35

Modélisation prédictive des relations propriétés-composition dans les verres d'oxydes

Damien Perret[1], Alexandra Garcin[1]

[1]Commissariat à l'Energie Atomique et aux Energies Alternatives – DE2D/SEVT/LDMC – CEA Marcoule – 30207 Bagnols-sur-Cèze – France

Depuis une cinquante d'années, le verre est utilisé comme matériau de confinement des déchets nucléaires de haute et moyenne activité. Le besoin de modéliser et de prédire via une approche statistique et empirique les propriétés des verres nucléaires en fonction de leur composition a débuté dans les années 90. Les études antérieures consistaient en une évaluation paramétrique des propriétés des verres en fonction de la composition, en faisant varier un ou deux composants autour d'une composition de référence. Cette méthode nécessite d'effectuer un grand nombre de formulations et ne permet d'expliquer la variation de propriété qu'autour d'une composition de référence. Le besoin de connaître la variation des propriétés d'intérêt en tout point du domaine de composition s'est avéré rapidement nécessaire et cependant incompatible avec une approche paramétrique simple d'une part, et avec un grand nombre d'éléments chimiques à faire varier en même temps, d'autre part. La complexité de la composition du verre nucléaire rend impossible l'utilisation d'outils théoriques (DFT, dynamique moléculaire, contraintes topologiques) et nécessite le développement de modèles statistiques empiriques. Parmi les propriétés d'intérêt à modéliser, on peut citer la durabilité chimique du verre, la température de transition vitreuse, ou encore la viscosité de la fonte verrière.

Depuis les années 2000, l'augmentation significative de la puissance des outils informatiques a permis l'utilisation d'algorithmes performants dans les méthodes de data mining. Par exemple, il est aujourd'hui possible de modéliser efficacement la température de transition vitreuse d'un verre de composition complexe à l'aide de réseaux de neurones. La viscosité de la fonte verrière est une propriété plus difficile à modéliser, du fait de sa très grande variabilité sur les échelles de température et de composition. Cette communication vise à présenter une méthode récemment développée, qui associe les techniques de plan d'expériences, de régression multilinéaire et de réseaux de neurones. L'outil développé utilise les données de formulation verrière générées au CEA ces 30 dernières années ainsi qu'un grand nombre de données collectées dans la littérature. Il permet de prédire la température de transition vitreuse ainsi que la viscosité de la fonte verrière à différentes températures.

2. Présentations par affiche

Hyperbolic Support Vector Machine P1

Aya El Dakdouki[1,2]

[1]Université des Sciences et Technologies de Lille - Lille I
[2]INRIA Nancy, Université Henri Poincaré - Nancy I

Dans l'apprentissage automatique, les machines à vecteurs de support (SVM) sont un ensemble de techniques d'apprentissage supervisé destinées à résoudre des problèmes de discrimination et de régression. Les SVM sont une généralisation des classifieurs linéaires. Dans ce papier, j'introduirai un nouveau classifieur multi-classe à marge basé sur des classes de fonctions à valeurs vectorielles, dont chaque fonction composante est associée à une catégorie. Il sagit d'une machine à noyau dont les surfaces de séparation sont hyperboliques et il généralise les SVM. Ensuite, j'établirai ses propriétés statistiques, parmi lesquelles la Fisher consistance et je montrerai les classes de fonctions composantes sont des classes Glivenko-Cantelli uniformes (GC) ceci en établissant un majorant de la complexité de Rademacher. Cela donne un risque garanti de ce classifieur.

Towards better efficiency of interatomic linear machine learning potentials P2

Alexandra Goryaeva[1], Jean-Bernard Maillet[2], Mihai-Cosmin Marinica

[1]DEN - Service de Recherches de Métallurgie Physique – Commissariat à l'énergie atomique et aux énergies alternatives (CEA) Paris-Saclay – France
[2]DAM, DIF – Commissariat à l'énergie atomique et aux énergies alternatives (CEA) Paris-Saclay – France

In the field of materials science, machine learning potentials have achieved maturity and became worthwhile alternative to conventional interatomic potentials. In this work we profile some characteristics of linear machine learning methods. Being numerically fast and easy to implement, these methods offer many advantages and appear to be very attractive for large length and time scale calculations.

However, we emphasize that in order to be accurate on some target properties these methods eventually yield overfitting. This feature is rather independent of training database and descriptor accuracy. At the same time, the

major weakness of these potentials, i.e. lower accuracy with respect to the kernel potentials, proves to be their strength: within the confidence limits of the potential fitting, one can rely on less accurate but faster descriptors in order to boost the numerical efficiency. Here, we propose a hybrid type of atomic descriptor that combines the original forms of radial and spectral descriptors. Flexibility in choice of mixing proportions between the two descriptors ensures a user defined control over accuracy / numerical efficiency of the resulting hybrid descriptor form. These observations open many avenues in the field of linear machine learning potentials that up to now are preferentially coupled with more robust and computationally expensive spectral descriptors.

P3

Towards a more compact representations of microstructures using deep learning

Michiel Larmuseau[1,2,3,4], Michael Sluydts[3,4], Tom Dhaene[1], Stefaan Cottenier[2,3,4]

[1]SUMO Lab, Ghent University – Belgique
[2]OCAS – Belgique
[3]Center for Molecular Modeling, Ghent University – Belgique
[4]Department of Electrical Energy, Metals, Mechanical Constructions and Systems, Ghent University – Belgique

For decades, metallurgists have relied on intuitive physical features of the microstructures such as the grain size to establish a quantitative link with the properties of the material. However, for complex, multiphase steels features such as the grain size cannot always be discerned and hence these traditional methods fall short. In order to analyse these types of microstructures, methods from computer vision have been successfully applied as an alternative. Inspired by the surge of deep learning in many scientific fields, we investigate the potential of deep learning in extract relevant features from microstructures. To this end, a supervised technique called "Triplet Networks" is used. This method allows to represent each microstructure in a low dimensional space where the distance between microstructures belonging to the same class is minimized. We obtain promising results for both optical and SEM images for challenging microstructure recognition tasks using very compact representations. We compare our deep learning results with other computer vision methods.

Hyperbolic Support Vector Machine

P4

Felix Musil[1], Federico Paruzzo[1], Albert Hofstetter[1], Sandip De[1], Michele Ceriotti[1], Lyndon Emsley[1]

[1]COSMO, EPFL, Lausanne, Suisse

The calculation of chemical shifts in solids has enabled methods to determine crystal structures in powders. The dependence of chemical shifts on local atomic environments sets them among the most powerful tools for structure elucidation of powdered solids or amorphous materials. Unfortunately, this dependency comes with the cost of high accuracy first-principle calculations to qualitatively predict chemical shifts in solids. Machine learning methods have recently emerged as a way to overcome the need for explicit high accuracy first-principle calculations. However, the vast chemical and combinatorial space spanned by molecular solids, together with the strong dependency of chemical shifts of atoms on their environment, poses a huge challenge for any machine learning method. Here we propose a machine learning method based on local environments to accurately predict chemical shifts of different molecular solids and of different polymorphs within DFT accuracy (RMSE of 0.49 ppm (1H), 4.3ppm (13C), 13.3 ppm (15N), and 17.7 ppm (17O) with R2 of 0.97 for 1H, 0.99 for 13C, 0.99 for 15N, and 0.99 for 17O). We also demonstrate that the trained model is able to correctly determine, based on the match between experimentally-measured and ML-predicted shifts, structures of cocaine and the drug 4-[4-(2-adamantylcarbamoyl)-5-tert-butylpyrazol-1-yl]benzoic acid in a chemical shift based NMR crystallography approach.
