Ontology development and machine learning for automating materials science research and materials discovery

Fabio Le Piane[1−2], **Tommaso Forni**[1−3], **Matteo Baldoni**[1], **Francesco Mercuri**[1]
1. DAIMON Team, ISMN-CNR, Via P. Gobetti 101, 40129 Bologna, Italy
2. Department of Computer Science and Engineering, Alma Mater Studiorum - Università di Bologna, Mura Anteo Zamboni 7, 40126 Bologna, Italy
3. Department of Control and Computer Engineering, Politecnico di Torino
fabio.lepiane@ismn.cnr.it, matteo.baldoni@ismn.cnr.it, tommaso.forni@ismn.cnr.it, francesco.mercuri@cnr.it

Abstract

Industrial progress is strongly connected to materials development. New functional materials, specifically tailored to solve specific tasks, are often needed for enabling new technologies and industry processes and to develop new products. Materials science has a long history in enforcing high performance computing paradigms. Traditional approaches to computational materials science, however, are quickly reaching their limits both in throughput and in predictive power. New data-driven computational paradigms, including machine learning and deep learning, offer now opportunities both in replacing and supporting computational simulations. The widespread adoption of data-driven technologies in materials science is often spoiled by fragmentation and non-standardisation of materials data and computing approaches. Pairing the advances in data-driven computing technologies with efforts in knowledge engineering in the field of advanced materials can lead to substantial improvements, accelerating the development process and improving the quality of results. Here, we present some preliminary result in this direction, showing how the application of machine learning to materials science paired to a domain ontology specifically designed for molecular materials enabled us to accelerate the research on materials and materials discovery within a broad application domain.

1 Introduction

Materials science is known to be a crucial pillar in the development of both technological and industrial applications. Research in materials science has been boosted, in recent years, by advanced and cutting-edge computational tools. However, traditional computational chemistry and materials science methods lack the needed efficiency and predictivity in tackling complex problems related to materials, as for example in the development of materials for advanced applications in technology.

To overcome such limitations, in recent years the application of data-driven technologies to materials science research activities, ranging from machine learning to knowledge organization and ontology development, have emerged[Ong, 2019; Khatib e De Jong, 2020; Hong *et al.*, 2020; Butler *et al.*, 2018; Schmidt *et al.*, 2019; Mueller *et al.*, 2016; Ashino, 2010; Sanchez-Lengeling e Aspuru-Guzik, 2018; Cheung *et al.*,]. The latter, in particular, offers the possibility to efficiently gather unified information within a specific domain. The application of domain ontologies for materials can support researchers in the development of standardized environments and for unifying research information and knowledge. This approach can also assist integration of research knowledge in different sub-topics, including merging data originating from empirical and computational experiments.

In this work, we discuss the steps required to pair formal representations of knowledge in the domain of materials science and technologies to computational workflows for materials design and property prediction assisted by machine learning. The approach is demonstrated in a real application scenario in the development of molecular materials, which constitute a relevant class of materials for advanced applications.

2 Ontology development

Ontologies have proved to be a fundamental building block in knowledge organization. In our work, we developed a domain ontology targeted to the organization of both chemical and physical knowledge about materials and to organization and standardization of actual research workflows, including computational and empirical work. We initially focused our work on molecular materials, that is, materials based on molecular systems (molecules, polymers, etc.) as their constituting units. The focus on molecular materials allows us to limit the scope of the domain, yet targeting systems of great relevance in the field of advanced materials for applications. We started gathering information from domain experts, interviewing them in order to understand the underlying structures of their specific work in the context of molecular materials development and applications.

This led us to the development of MAMBO, the Materials And Molecules Basic Ontology[Le Piane *et al.*,]. The main application scenarios are:

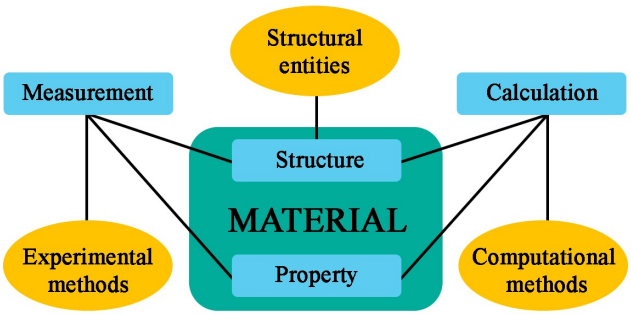

Figura 1: Core concepts of MAMBO. The central concept of `Material` is shown in teal, `Structure`, `Property`, `Measurement` and `Calculation` classes are shown in cyan. The yellow boxes are placeholders for the respective hierarchies.

- Retrieving structured information from databases on molecular materials. We want to enable semantic searches on multi-scale modelling and characterization data including information about the basic chemico-physical entities constituting parts of the active systems (for example molecules, polymers, etc.), aggregated systems up to full-scale and macroscopic systems.
- Defining complex research workflows for systems based on molecular materials. The target is to model all the steps of a complex workflow which addresses a specific scientific/technological problem in the framework of molecular materials, being that computational or empirical. For example, the modelled workflows can lead to the implementation of efficient computational approaches for the high-throughput screening of properties of a given class of molecular systems.

Moreover, the data obtained by simulations can further be used to implement predictive data-driven models, for example for designing novel materials.
Unsurprisingly, MAMBO revolves around the `Material` class, which is strongly related to the concepts of `Material Structure` and `Material Property`, which in turn are also connected to `Measurement` and `Calculation`, which are the classes that represent the empirical and computational sub-domains, respectively. This shallow structure is then linked to deeper hierarchies in the ontology. The relatively simple structure of MAMBO core helps to quickly grasp the fundamental ideas of the discipline and of the ontology itself. A figure representing main MAMBO concepts and their relations is shown in Fig. 1.

3 Property prediction using machine learning

In recent years, machine learning methods have applied with success to investigate the properties of molecular materials. The vast majority of these studies are focused on the properties of individual molecules, targeting the correlation between molecular structure and resulting properties. The properties of several technological materials constituted by molecular aggregates, however, depend on both molecular structure and on aggregation morphology, as for example in the case of nanoscale materials[Baldoni *et al.*, 2018]. Computational methods for predicting the properties of molecular materials must therefore integrate the properties of individual molecules with information about aggregation morphology, which, in turn, can be related to materials fabrication and processing. The definition of a scalable modelling paradigm able to simulate and predict the properties of molecular materials as a function of molecular structure and aggregation/fabrication conditions can potentially enable high-throughput development of novel materials for technological applications.
For that reasons, we designed and implemented a computational workflow for the simulation of the properties of molecular materials integrated with a machine learning scheme for enhancing the computational workload. The workflow is based on a multi-scale top-down approach, in which target properties are defined from the application to the molecular scale. The workflow is implemented through top-down hierarchical data structures, which connects the properties of molecular materials at the nanoscale to the atomistic/electronic scale. Modelling data are generated by applying domain-specific simulation protocols, and machine learning approaches are used to enable the scale reduction, providing a local mapping at a lower scale of the properties of large molecular aggregates, reducing the overall computational load.
Gathering huge amounts of data is not always feasible in this domain, and we decided to rely on simpler but more data-efficient model, leaving the application of deep learning methods to future works. We developed an effective set of features to describe molecular systems, and in particular the relative configuration of molecules constituting aggregates: using the actual coordinates of atoms in molecules, we are able to determine their mutual position in the 3D space, described as a combination of their mutual orientation and distance. We implemented and tested a wide range of feature vectors aimed at the description of the mutual position of molecules in molecular aggregates. A particularly efficient feature set is composed of two rotation matrices, describing the relative orientation and position in 3D space of two molecules, respectively, and the scalar distance between the centroids of the two molecules. This representation proved to be quite effective for the description of the intermolecular configurations in aggregates[Le Piane *et al.*, 2020]. By using a dataset with around 2000 entries to train a kernel ridge regressor, we achieved an accurate prediction of properties related to the relative configuration of molecular pairs in aggregates. This result is remarkable even due to the skew of the distribution of the actual data. In this set of experiments, we considered properties related to the electronic structure of molecular pairs, which can be used to predict the potential efficiency of molecular materials as semiconductors. In figure 2 the correlation between the actual value and the predicted value of our target property is shown. Here, you can also appreciate the left-skew of the truth's distribution. The difficulty to uniformly map the feature space in the training dataset is another problem to overcome in order to effectively apply machine learning to this domain, and this is going to be the main focus of our future work.

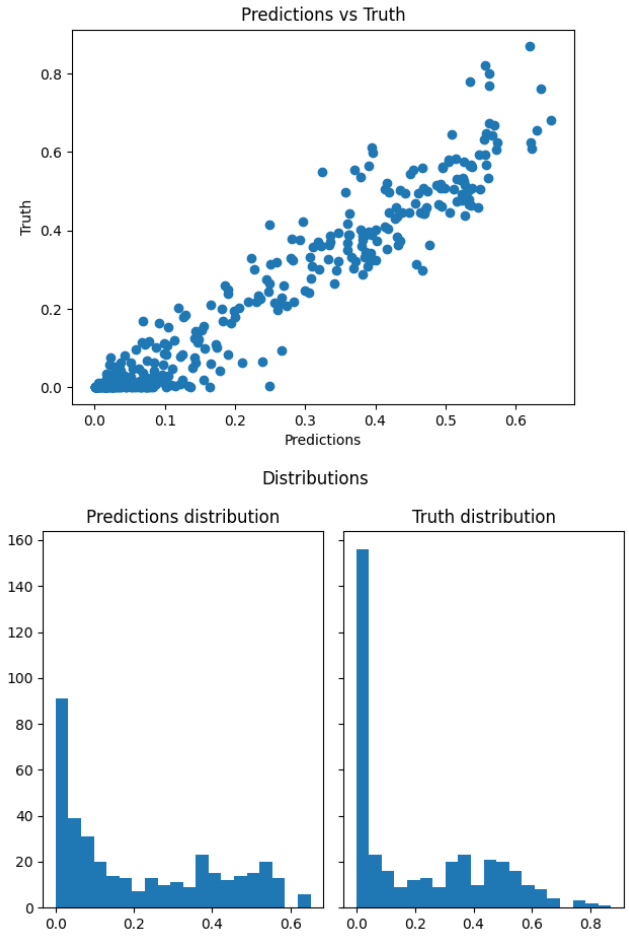

Figura 2: Correlation between predicted electronic coupling and true electronic coupling in molecular aggregates of a target semiconductor material (see text). Top panel: correlation between the predicted values (on the x-axis) and ground truth (on the y-axis). Bottom panel: distribution histogram of predicted values (left) and truth values (right). While both graphs shows how the ML algorithm can actually predict materials properties with sufficient accuracy, it is also evident how the strong left-skew of the truth distribution is making it harder for the model to fully capture the structure of the problem.

4 Wrapping the two together: automating data-driven techniques in materials science

The combination between a domain ontology focusing on knowledge on materials and computational materials science workflows can potentially accelerate research and the discovery of new materials. This approach can also be extended to similar but yet different problems. Namely, a specialized domain ontology is used to support a consistent description of computational workflow, which enables high-throughput generation of data and subsequent application to machine learning models for predictions. This approach can be used for example to set up an automated workflow for evaluating structure/property relationships in molecular materials. First, the steps needed to link a molecule/property relationship to a predictive machine learning model, in term of a computing workflow, are evaluated. In this process, MAMBO classes and structures to represent tasks and data structures are used. In principle, all steps should be properly mapped by concepts and relationships defined within MAMBO. This approach allows us to reuse large parts of this work to different yet related problems and scenarios. While the featurization process could be different case-by-case, our approach is not based on the specific molecule at hand. In other words, if the underlying structure of the problem is similar to another one already analyzed we should be able to approach it in the same way. Moreover, selected features could be easily integrated with others. In the specific case of descriptors for molecular aggregates, if materials properties can be related to distance and mutual orientation of molecular units, MAMBO classes and relations can be applied to identify a lightweight yet effective representation of the target structure/property relationships.
To assess the applicability of our approach, we identified a new target molecule, the basic constituent of a molecular material, and a new target property. We applied the structure of MAMBO to define a consistent computational workflow, which was used to generate data and to build predictive machine learning models. We considered the relationship between the local structure of molecular materials in aggregates and a target property related to the electronic structure of the system. A preliminary result of the approach proposed is shown in Figure 3: here, we used a nearly identical features set and machine learning stack as shown before (Fig. 2) to predict another property for a different molecule. While the two properties are somewhat related, we were able to predict the new property without changing our approach. Moreover, since the molecule at hand has a more simple structure than the one used in the first experiment, we have been able to obtain an even better fitting with fewer data, which is another promising aspect of our approach. Work is in progress to extend the domain of the ontology, covering more general aspects of research on molecular materials, and to apply more efficient and accurate machine learning models. Future work will also be focused on the development and implementation of generic automated frameworks for the integration and analysis of data on materials, for the development of workflows to support computational and experimental research on materials and on the realization of data-centric predictive systems for materials design and discovery.

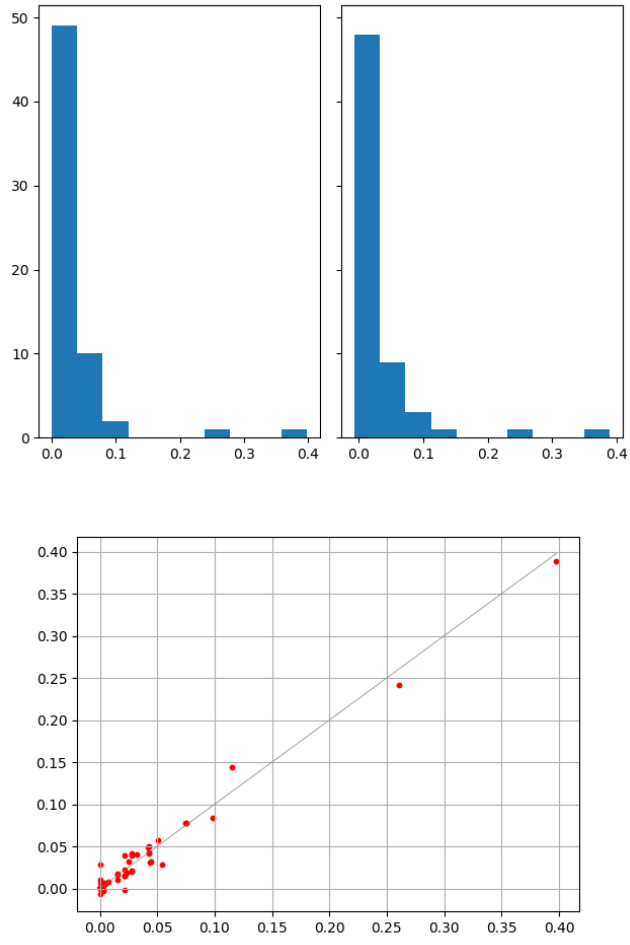

Figura 3: Correlation plot and distribution histogram for the second experiment. It is visible how, even with a very limited amount of data (tens of molecular pairs) we have been able to obtain a model able to understand the relation between the molecules relative position and the new property analyzed. In this case, we've also been able to do this despite an analogous skew in the data, probably due to the simpler structure of the molecule and a more linear relation between the features and the target property

Riferimenti bibliografici

[Ashino, 2010] Toshihiro Ashino. Materials Ontology: An Infrastructure for Exchanging Materials Information and Knowledge. *Data Science Journal*, 9(0), jun 2010.

[Baldoni *et al.*, 2018] Matteo Baldoni, Andrea Lorenzoni, Alessandro Pecchia, e Francesco Mercuri. Spatial and orientational dependence of electron transfer parameters in aggregates of iridium-containing host materials for OLEDs: coupling constrained density functional theory with molecular dynamics. *Physical Chemistry Chemical Physics*, 20(45):28393–28399, nov 2018.

[Butler *et al.*, 2018] Keith T. Butler, Daniel W. Davies, Hugh Cartwright, Olexandr Isayev, e Aron Walsh. Machine learning for molecular and materials science. *Nature 2018 559:7715*, 559(7715):547–555, jul 2018.

[Cheung *et al.*,] Kwok Cheung, John Drennan, e Jane Hunter. Towards an Ontology for Data-driven Discovery of New Materials Introduction and Objectives.

[Hong *et al.*, 2020] Yang Hong, Bo Hou, Hengle Jiang, e Jingchao Zhang. Machine learning and artificial neural network accelerated computational discoveries in materials science. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 10(3):e1450, may 2020.

[Khatib e De Jong, 2020] Muammar El Khatib e Wibe A De Jong. ML4Chem: A Machine Learning Package for Chemistry and Materials Science. mar 2020.

[Le Piane *et al.*,] Fabio Le Piane, Matteo Baldoni, Mauro Gaspari, e Francesco Mercuri. Introducing MAMBO: Materials And Molecules Basic Ontology.

[Le Piane *et al.*, 2020] Fabio Le Piane, Matteo Baldoni, e Francesco Mercuri. Predicting the properties of molecular materials: multiscale simulation workflows meet machine learning. jul 2020.

[Mueller *et al.*, 2016] Tim Mueller, Aaron Gilad Kusne, e Rampi Ramprasad. Machine Learning in Materials Science. *Reviews in Computational Chemistry*, 29:186–273, may 2016.

[Ong, 2019] Shyue Ping Ong. Accelerating materials science with high-throughput computations and machine learning. *Computational Materials Science*, 161:143–150, 2019.

[Sanchez-Lengeling e Aspuru-Guzik, 2018] Benjamin Sanchez-Lengeling e Alan Aspuru-Guzik. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 361(6400):360–365, jul 2018.

[Schmidt *et al.*, 2019] Jonathan Schmidt, Mário R.G. Marques, Silvana Botti, e Miguel A.L. Marques. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials 2019 5:1*, 5(1):1–36, aug 2019.