PRINCIPAL ACADEMIC TUTOR
Eliodoro Chiavazzo, Energia (PoliTO)

ACADEMIC TUTOR
Luca Bergamasco, Energia (PoliTO)

EXTERNAL INSTITUTIONS
IBM - ENI - DOW Benelux - Uppsala University - EPFL

EXTERNAL TUTOR
Cristiano Malossi, IBM
Tom Verbrugge, DOW

TEAM MEMBERS

Lorenzo Chiavarini,
Mechanical Engineering
PoliTO

Massimo Bini,
Mathematical Engineering
PoliTO

Francesca Mignacco,
Physics of Complex Systems, PoliTO

Francesco Mori,
Physics of Complex Systems, PoliTO

Veronica Piazza,
Chemical Engineering
PoliMI

Emilia Rosselli Del Turco,
Integrated Product Design
PoliMI

Francesco Signorato,
Mechanical Engineering
PoliTO

Silvio Trespi,
Chemical Engineering
PoliMI

iMAT:

Digitalizing, democratising and empowering materials development via Artificial Intelligence

Executive summary

The development of new materials has been identified by the European Material Modelling Council (EMMC) [1] as one of the main innovation drivers for the European industry. This concerns also a focus on sustainability of energy storage solutions, where the usage of optimised materials can bring important results. However, decision making in R&D departments regarding material discovery requires considerable investments in terms of time and money. Material Modelling has been used to screen materials and focus the company's efforts, but the required skills are usually not available. The figure of Translator has therefore been identified by the EMMC to bridge academic and industrial world but, due to the broad knowledge required for this purpose, the Translator cannot be an expert on each field therefore an aid is needed in the form of easily accessible information [2]. Due to the scarcity and incompleteness of existing databases, we decided to exploit the enormous amount of scientific literature as source of data to provide insights. These, information is however unstructured and need to be extracted and transformed in structured data such as databases and graphs. This process needs to be almost fully automatic and for this purpose Natural Language Processing (NPL) techniques can be applied. NPL is a branch of Artificial Intelligence used to process and analyse human language data. Its application to scientific literature is however challenging due to several limiting factors such as scientific language and data format. Even if some studies have been performed in the field, no fully comprehensive process pipeline from data recollection to extracted data structuring is available on the market. We decided to develop this pipeline on a real case study, the usage of zeolites in thermal storage, but constructing it in a way not to require any in-depth programming expertise, also to assure the adaptability of the process to different case studies. To do so we exploited also the collaboration with IBM Research [3], including several their tools. Our model is now able to autonomously recognise useful data in zeolites literature. We can then interrogate these data to have highlights about their distribution over relevant axis and emerging trends. Facing a new research topic, it is now possible to adapt this process and screen the existing literature to have important insights on how to better address efforts and resources. The pivotal difference and main value in our process is that it adopts and applies a set of simple and easily available tools, requiring no hard skills, and it leads the user throughout the whole process. It becomes then an important asset for future researches as the starting point for the construction of a new model, based on different topics and literature, which can be applied by users who are not familiar with Artificial Intelligence and NPL techniques.

Key Words

MATERIALS
RESEARCH
MACHINE LEARNING
TRANSLATOR

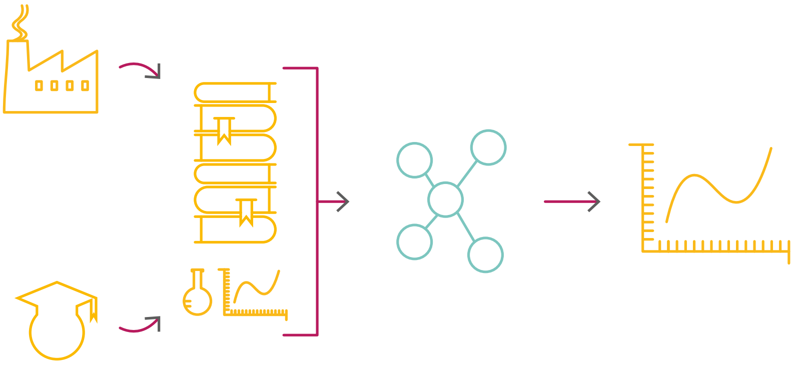
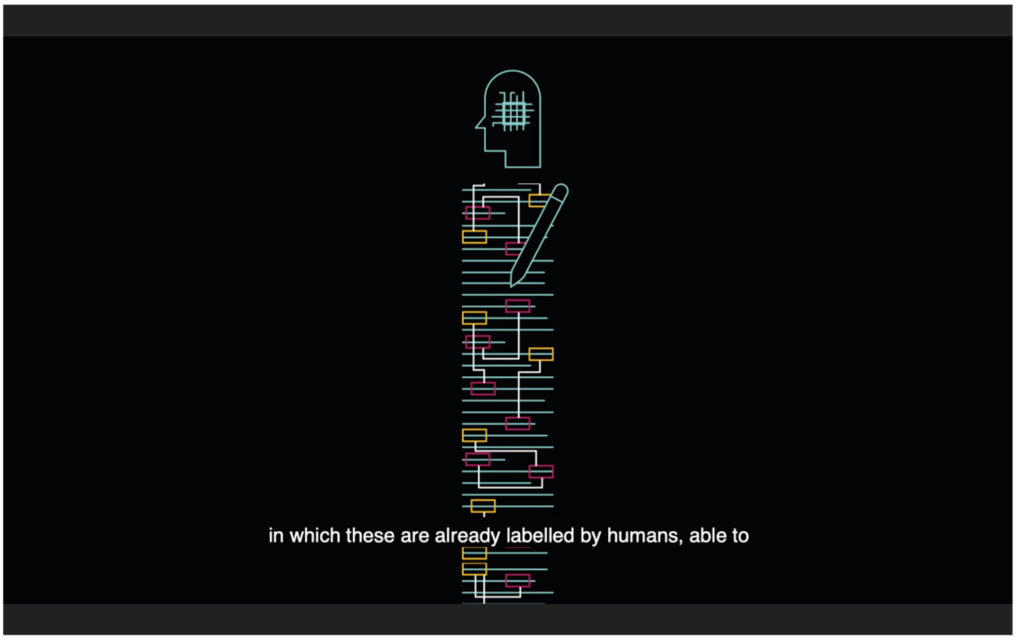
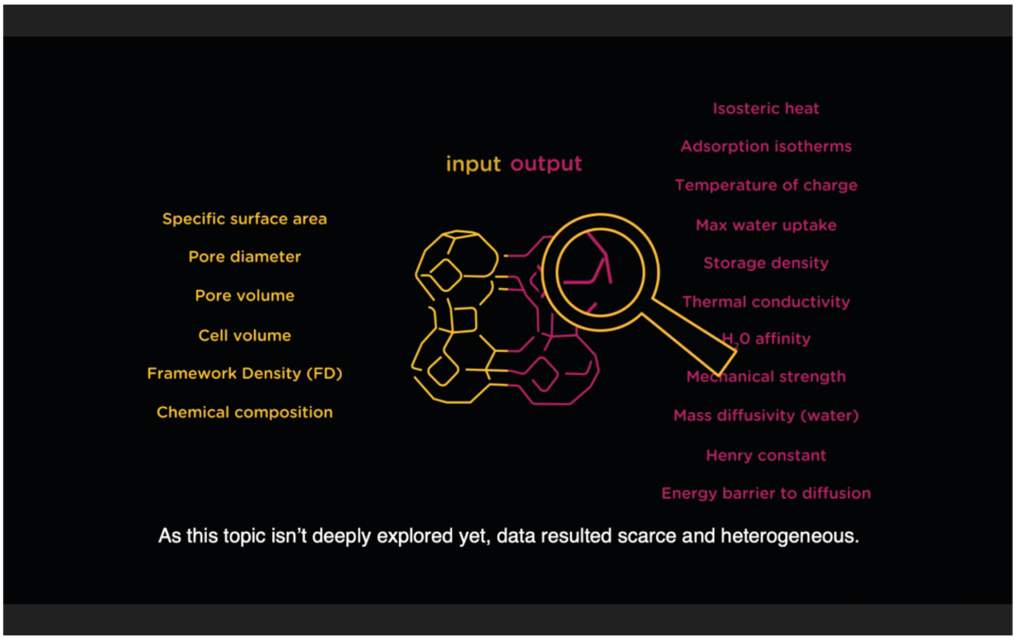


Fig. 1 - Concept of the process



in which these are already labelled by humans, able to



input output

Specific surface area

Pore diameter

Pore volume

Cell volume

Framework Density (FD)

Chemical composition

Isosteric heat

Adsorption isotherms

Temperature of charge

Max water uptake

Storage density

Thermal conductivity

H_2O affinity

Mechanical strength

Mass diffusivity (water)

Henry constant

Energy barrier to diffusion

As this topic isn’t deeply explored yet, data resulted scarce and heterogeneous.

Moreover, in the documents themselves synonyms were used to refer to the same concept.

Project description written by the Principal Academic Tutor

The rapid evolution of contemporary technologies requires continuous development of novel and/or improved materials. In this context, materials modelling techniques and numerical simulations can serve as powerful means to accelerate industrial innovation, alleviating time consumption and costs of tailored experimental procedures. Thanks to the always increasing computational power available, simulations for material screening have the possibility to become everyday more intensive - see e.g. high-throughput simulations. Notwithstanding, these techniques require proper modelling expertise, proper infrastructure and a considerable commitment. Apart from the modelling and simulation part, the generated (big) data outcome must then be properly interpreted and treated, to extract the most relevant information (mining) for the purpose. This is one of the reasons why, at least in the context of materials modelling - yet not only - development of proper data analysis and mining tools are assuming every day more and more importance, both in academia and industrial sector. These tools are more and more often landing towards the emerging field of artificial intelligence, term which now encompasses quite a large number of techniques and applications – e.g. machine learning. Apart from pure (simulation) data mining, these techniques have recently been applied to screen (i.e. mine) the information available in scientific literature in the form of text, using Natural Language Processing (NLP) techniques combined with machine learning. This approach represents an alternative approach to the generation and mining of big data from high-throughput numerical simulations and aims to take advantage of the huge amount of information which is already available in scientific papers.

This project focuses on materials analysis in the abovementioned context about efficient data handling and its continuously growing importance for industrial innovation (e.g. industrial digitalization). This attitude perfectly fits the role of the Translators, which have been recently defined by the European Materials Modelling Council (EMMC, https://emmc.info) as professional profiles able to bridge industrial needs and state-of-the-art academic research to drive industrial innovation. The project has been finally shaped to focus on the application of state-of-the-art artificial intelligence (IBM Watson) for automatic literature screening and materials data extraction via NLP/ML to speed up properties' assessment. The application case has been chosen to focus on the properties of materials for thermo-chemical heat storage (e.g. water adsorption on different zeolites), which is an active field of research due to its relevance to many industrial sectors, such as e.g. automotive and civil. The project consists of four main types of activities: (1) LEARN the basic technical concepts required to work in the context of data (NLP/ML in this case) for materials screening; (2) DEVELOP a proper machinery to implement the idea; (3) ANALYZE the outcome and assess the perspective impact; (4) CONVINCE the audience that the idea is valuable using proper communication channels. This project has benefited of the cooperation of the European Materials Modelling Council and of different academic and industrial partners to provide support and feedback on specific tasks.

Team description by skill

As Alta Scuola Politecnica is a program specifically designed to foster a multidisciplinary collaboration, our skillset was very various and required a precise organization.

The main competences present in our group were related on one hand to Chemistry and Material Science (Material Modelling sub-team) and on the other hand to computer science and digital innovation (AI and Machine Learning sub-team). This formal division was meant to explore the interactions between the two main topics of the project: Artificial Intelligence and Material Modelling. The former group was assigned with the task of framing the research context within the structure required to handle it with a machine learning logic. The latter one was required to explore the existing programming tools which would be pertinent in building the actual model and develop new solutions. The distinction was, however, blurry since the cooperation of all members was required for several tasks. For instance, a precise knowledge of the chemical properties of materials is required to build and train a supervised machine learning model.

The Communication Coordinator, thanks to her competencies in design and communication, covered the outreach and communication strategies, gathering and organising the outcomes from the rest of the group. The Team Controller was pivotal in keeping a connection within the different subunits of the team and sticking to the project schedule.

Goal

The main goal of the iMat project is to explore the possibilities of application to exploit emerging Artificial Intelligence algorithms to speed up research in the field of Material Modelling. Indeed, the EMMC also identifies the development of new materials as one of the main innovation drivers for the European industry, also from the point of view of sustainability. Therefore, great attention has been directed towards the screening of existing and hypothetical materials to have anticipating insights about the most promising directions and focus a company's energies on a few attempts. In this context, the professional figure of the Translator has been identified with the aim of supporting and bridging between academic research and business necessities. However, in many cases Translators may not have a complete knowledge of the problem they are working on. An AI tool able to read and understand a large amount of scientific articles would be crucial to strengthen the role of the translator. Indeed, such an algorithm would speed up bibliographic research and allow to summarise a large amount of unstructured information into a simple structured database. Therefore,, at the beginning of this project we wanted to build a completely-automatized tool for the text-mining of quantitative information from scientific articles.

Understanding the problem

Materials have marked the evolution of mankind since prehistoric times, naming the ages based on the dominant materials. Despite the fact that everything around us is made of materials, for almost the entire course of human history, the discovery of new materials was pursued as a sort of mystic art. Up to less a century ago this process involved a tremendous amount of trial-and-error and expensive testing. The traditional material discovery process is a tortuous path consisting of several steps. Generally, the time required from the formulation of the idea of a new material to its deployment is remarkably long: first of all, it requires skilled technicians to perform long and expensive experiments, that are necessary to get meaningful data. Secondly, the tests are carried out mainly with the aim of incrementally improving existing chemistries and already-well-established materials rather than with the idea of investigating completely new materials to assess their potential. However, modern advances in the understanding of the intrinsic physics provide a comprehensive framework to supervise the materials discovery process, laying the foundations of a new discipline called computational materials science. Materials modelling allows to optimize the scheduling of experimental activities, hence limiting expensive trial-and-error procedures. However, it requires highly specialized professional figures and high-performance-computing facilities. Hence, money and time consuming hard work is usually necessary to foster the research for innovative materials and this is an important limiting factor, especially for Small Medium Enterprises (SMEs) [4]. A completely different approach envisages the employment of Artificial Intelligence (AI) techniques to sift through the already huge amount of scientific articles and reports, assuming that the dataset under analysis embeds the intrinsic physics. Hence, the data-driven approach could gather relevant insights and propose innovative molecular structures to be further investigated. However, the major part of useful data appears in the scientific literature in an unstructured form and a well-established methodology to arrange them in a database for data-mining purposes still does not exist. EMMC has identified the professional figure of the Translator to push materials innovation. Since the Translator does not have a specific knowledge on each case study he has to face, it is necessary for him to easily gather the various pieces of information found in the scientific literature. Therefore, the development of a procedure to extract and organize data is of pivotal importance. A specific case study that is particularly relevant in the field of materials modeling has been investigated during this project. To reach a low-carbon emission system, energy production from renewable sources is strongly promoted nowadays, but their intermittent nature threatens to limit their applicability. To tackle this issue, the development of an efficient Thermal Energy Storage System (TESS) is of fundamental importance[5]. In particular, the use of zeolites in sorption-based TESS represents an important current research topic. Hence, chemical and energy industrial realities could be strongly interested in a process that would allow them to screen the state-of-the-art knowledge and that would manifest the best route to be followed for material innovation.

Exploring the opportunities

The present work opens a wide range of opportunities to be exploited. For instance it allows to offer external institutions a tool based on Natural Language Processing to extract numerical data from scientific literature. Furthermore, regarding the EMMC objective to bridge research and industry, it corroborates the role of the Translator by complementing his work with a tool that can foster materials modelling. In addition the support and the supervision of IBM guarantees to explore the potentialities of IBM Watson tools [6] and to gain precious insights on cutting edge Artificial Intelligence tools. Moreover, the case study of Thermal Energy Storage allows to tackle one of the most promising challenge that the present world has to tackle: renewable energy. Finally the innovative process that is developed can be formalized in the publication of a scientific paper.

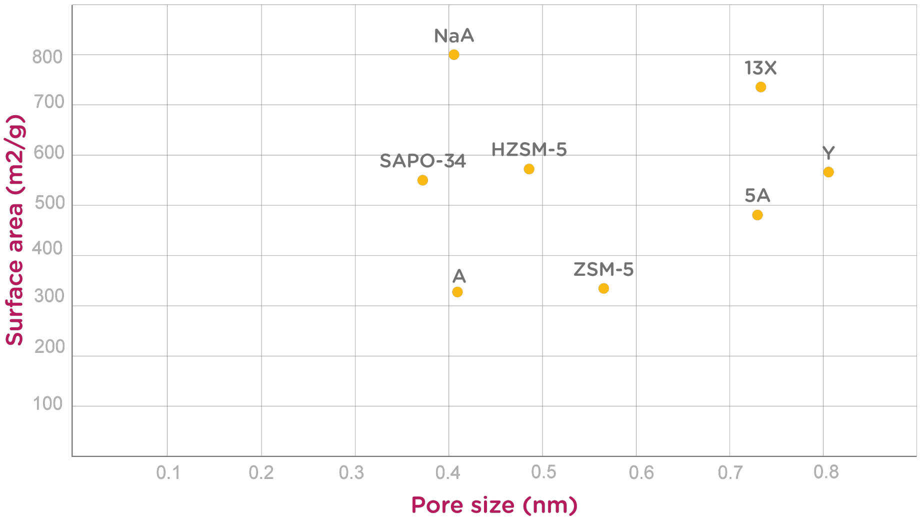


Fig. 2 - Extrapolated Ashby map comparing different zeolites over two relevant properties in TESS

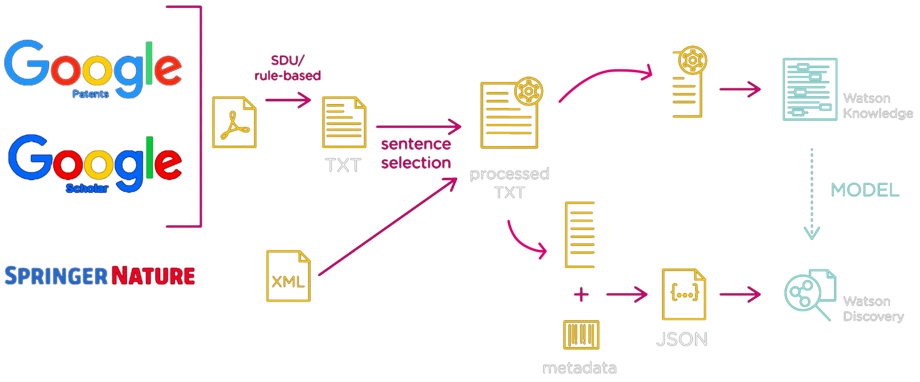


Fig. 3 - Preprocessing steps and model construction

Generating a solution

Our solution consists of a multistep process that exploits Artificial Intelligence and Natural Language processing to extrapolate structured data from unstructured information contained in scientific literature. Every year, around 2.5 millions of scientific articles are published and the development of an efficient automatic procedure of information extraction would make a major breakthrough in research, enhancing the connection between industry and academia.

Our results demonstrate that text-mining methods based on machine learning can be exploited to translate a verbal unsystematic input into a well-organised quantitative output that can be easily visualised. We applied AI tools provided by IBM Watson to convert scientific papers from PDF format to plain text and to implement a model that analyses them sentence-by-sentence. Our model categorises words according to their meaning and identifies relevant relations between them. In this way, we were able to automatically recognise patterns in the text and extract quantitative information related to our case study: zeolitic materials. We concentrated on a list of zeolites and properties of interest and we collected the corresponding numerical values. The data were eventually used to create Ashby maps, an efficient visualisation technique that can support decision-making.

Main bibliographic references

[1] Council, E.M.M.:TheEMMCRoadMap2018forMaterialsModellingandInformatics, 2018
[2] Council, E.M.M.:EMMCTranslatorsGuide.https://emmc.info/wp-content/uploads/2017/12/TranslatorsGuide.pdf,2017.
[3] IBM research.http://www.research.ibm.com/.
[4] Commission, E.:SMEPerformanceReview.https://ec.europa.eu/growth/smes/business-friendly-environment/performance-review_en,2018.
[5] Dincer I.,R.A.: Thermal Energy Storage: Systems and Applications. Wiley,2011.
[6] IBM Watson.https://www.ibm.com/watson/.