Large Language Models for Material Design

Jérémie Cabessa[1] and Marie-Pierre Gaigeot[2]

[1]Laboratoire DAVID, Université Versailles-Saint-Quentin – Université Paris-Saclay, 78000 Versailles
[2]Laboratoire LAMBE, UMR8587, Université d'Evry Val d'Essonne – Université Paris-Saclay, 91025 Evry

jeremie.cabessa@uvsq.fr and mariepierre.gaigeot@univ-evry.fr

Abstract: Large language models (LLMs), like GPT4, ChatGPT, Llama 2 and MISTRAL have revolutionized the field of Machine Learning. Recently, LLMs have been successfully adapted in the context of chemistry, due to the efficient representation of molecules by means of SMILES and SELFIES languages. Molecular/material design, which consists in the discovery of new molecules and materials, is a field of central importance, with multiple socio-technological implications, for which vibrational spectroscopy represents an essential experimental technique. This project proposes a bi-disciplinary approach to material design from the perspectives of LLMs and vibrational spectroscopy. In a first step, the prediction of vibrational spectra from molecular SMILES/SELFIES will be studied. In a second step, the challenging inverse problem of predicting SMILES/SELFIES from corresponding spectra will be considered – a direction which has never been achieved in the literature. Overall, this project constitutes a progress in the domains of molecular reconstruction in material design and inverse problems in machine learning.

Keywords: Machine Learning, Deep Learning; Chemoinformatics; Large Language Models (LLMs); Graph Neural Networks (GNNs); Inverse Problems; Molecular/Material Design; Vibrational Spectroscopy; Molecular Structures.

Résumé: Les grands modèles de langage (LLMs), tels que GPT4, ChatGPT, Llama 2 et MISTRAL, ont révolutionné le domaine de l'apprentissage automatique. Récemment, les LLMs ont été adaptés avec succès dans le domaine de la chimie théorique, en raison de la représentation efficace des molécules par le biais des langages SMILES et SELFIES. Par ailleurs, la conception de matériaux et de molécules, qui consiste en la découverte de nouveaux composés moléculaires, est un domaine d'importance majeure, aux implications socio-technologiques multiples, et pour lequel la spectroscopie vibrationnelle représente une technique expérimentale essentielle. Ce projet propose une approche bi-disciplinaire du design moléculaire par le biais des LLMs et de la spectroscopie vibrationnelle. Premièrement, la prédiction de spectres vibrationnels à partir de SMILES/SELFIES moléculaires sera étudiée. Deuxièmement, le problème inverse, plus complexe, de la prédiction de SMILES/SELFIES à partir de spectres vibrationnels sera considéré, une direction qui n'a à ce jour jamais été réalisée. Plus généralement, ce projet constitue un progrès dans les domaines de la reconstruction moléculaire en design de matériaux, et des problèmes inverses en apprentissage automatique.

Mots-clés: Apprentissage automatique, Apprentissage profond; Chémoinformatique; Grands modèles de langages (LLMs); Réseaux de neurones graphiques (GNN); Problèmes inverses; Design moléculaire/matériel; Spectroscopie vibrationnelle; Structures moléculaires.

1. Project Description

1.1. Research context

The discovery of new molecules and materials is of central importance with multiple technological and societal implications. In this context, *inverse molecular/material design* concerns the prediction or the generation of stable and synthesizable compounds which comply with desired properties and functionalities [1, 2]. In the chemical-physical community, classical methods for inverse molecular design usually involve exploring the potential energy surface (PES) of the molecular system to discover stable conformations, which is computationally prohibitive. However, in the last decade, various machine learning techniques have been applied to this task [2], including recurrent neural networks (RNNs), variational autoencoders (VAEs), generative adversarial networks (GANs), reinforcement learning (RL), or hybrid methods [1, 4], invertible neural networks [5], conditional generative neural networks [6], or other kinds of autoencoder-based neural networks [7]. Within this approach, the generation of compounds is mainly achieved via AI generative models for which the desired properties and functionalities are provided as inputs in an encoded way.

Vibrational spectroscopy represents a major experimental technique for 3D-structure characterization. Thanks to a one-to-one relationship between spectral fingerprints and 3D molecular structures [8–13], molecular design processes can naturally be approached from the perspective of vibrational spectroscopy. In this context, the problem of going from a given structure to a corresponding spectrum has been addressed either via ab initio/force field calculations [12, 14] or by means of machine learning techniques [15–17]. In the latter case, graph-theoretical tools can be used to encode key structural descriptors of molecular systems, and deep neural networks are trained to predict the spectrum characteristics from these encoded features [15, 16]. A more modern approach gets rid of the encoding scheme and use graph neural networks (GNNs) directly on the molecular graphs (2D graphs) to efficiently predict their vibrational spectra [17].

The inverse spectrum-to-structure mapping is more challenging, but also more impactful. Finding a unique 3D structure, or a small set of 2D graphs, responsible for a given spectrum inside the huge chemical space of atom positions is intractable by brute-force methods. In machine learning, various methods targeted at discovering molecules with desired functionalities have been investigated (variational autoencoders, generative neural networks, reinforcement learning, recurrent neural networks, and hybrid approaches) [1, 18]. For instance, a conditional generative neural network for the generation of 3D molecular structures with specified chemical and structural properties (e.g., motifs, gap values, low, intermediate and high relative atomic energy) has been proposed [6]. Also, DeepMind has achieved a ground-breaking solution to the 50-year-old 'protein-folding problem' with AlphaFold, a neural network-based model for protein 3D structure prediction from its amino acid sequence [19]. In relation with our framework, a few studies have been pursued to predict protein secondary structures from their vibrational spectra [13, 18].

Besides, in the late 80's, the SMILES (Simplified Molecular Input Line Entry System) language has been introduced as an efficient string representation of chemical structures, particularly well suited to computational processing [20]. More recently, the SELFIES (Self-Referencing Embedded Strings) molecular string representation has been proposed to overcome lack of robustness of SMILES [21]. These considerations coupled with the huge breakthroughs achieved by *large language models (LLMs)* (GPT4, ChatGPT, Llama, MISTRAL, etc.) [22] naturally led to the development of LLMs specifically designed for the processing of chemical languages. In fact, numerous LLMs have been pre-trained in a self-supervised way on massive datasets of SMILES and SELFIES. For instance, ChemBERTa, MolBERT, SMILES-BERT, ChemBERTa-2 and ChemGPT have been pre-trained on datasets of various sizes, using masked language modelling (MLM) and/or multi-task regression (MTR) tasks in order to learn enriched molecular fingerprints from their string representations [23–25]. Afterwards, the models have been fine-tuned on a multitude of classification and regression downstream tasks to evaluate the quality of the fingerprints.

Chemistry-based LLMs show promising results on many of these downstream tasks. LLMs have also been successfully applied to solve molecular graph specific problems by being paired with graph neural networks (text-rich graph or text-paired graph approaches) [26].

To the best of our knowledge, an LLM-based approach addressing the bi-directional relationship between 2D graphs and/or 3D molecular structures with vibrational spectra has not been undertaken.

1.2. This PhD

In this project, we address the bi-directional relationship between molecular structures and their vibrational spectra from the perspective of large language models (LLMs). Our goal is to develop adapted LLMs capable of predicting vibrational spectra from molecular SMILES/SELFIES (see Figure 1), and conversely, capable of predicting or generating molecular SMILES/SELFIES from corresponding vibrational spectra (see Figure 2).

In a first stage, the direction of predicting vibrational spectra from SMILES/SELFIES will be considered (see Figure 1, dashed box). The first step would be to fine-tune chemical-based LLMs (ChemBERTa, ChemGPT) on datasets of smiles and spectra already implemented in the literature and using customized loss functions previously described [16, 17]. In a second step, we aim to improve the spectra predictions with enriched LLMs. To that end, additional molecular information will be provided to the LLMs (for instance, gas/liquid phase, temperature, concentration, etc.) – either in a prompted way or in a numerical way. The accuracy of the predicted spectra will be compared to the spectra reported in the literature.

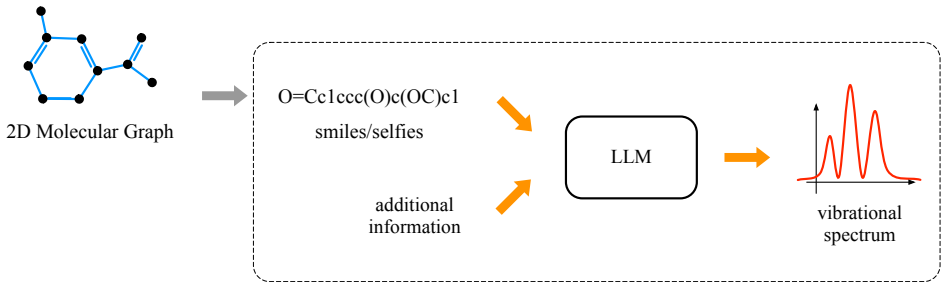

Figure 1: Prediction of the vibrational spectrum of a molecular compound from its SMILES/SELFIES representation, using an LLM (dashed box). The initial step represented by the grey arrow can be achieved using python libraries.

In a second stage, the inverse problem of predicting or generating molecular SMILES/SELFIES from vibrational spectra will be considered (see Figure 2, dashed box). This case is more complex, since the chemical-based LLMs are designed to take SMILES/SELFIES as input as opposed to the required vibrational spectra. Hence, the LLMs will be adapted to handle these alternative inputs (red arrow in Figure 2), which represents a challenge. In contrast, the output capabilities of the LLMs (orange arrow in Figure 2) poses no particular problem, since LLMs are designed to generate SMILES/SELFIES, as requested. Architecturally-wise, the models could even be coupled with GNNs in a text-rich graph or text-paired graph mode for better predictive capabilities [26]. The adapted LLMs will possibly need to be pre-trained in this new framework, before being fine-tuned on our dataset of spectra and smiles.

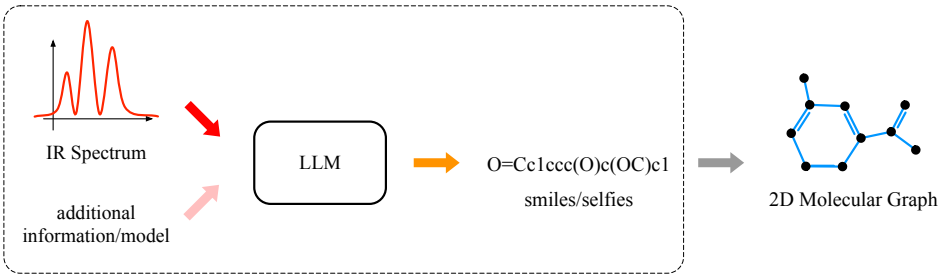

Figure 2: Prediction/generation of a molecular SMILES/SELFIES from its vibrational spectrum with an LLM (dashed box).

Our interdisciplinary project might bring several added values to the fields machine learning and theoretical chemistry. In machine learning, this project fits within the context of graph prediction and inverse problems in general. In theoretical chemistry, it contributes to the important research efforts on inverse molecular design.

1.3. Work plan

The project is divided into the following work packages:

WP1 Knowledge acquisition (3–6 months). This work package concerns the bi-disciplinary knowledge acquisition: LLMs, LLMs for chemistry, GNNs and inverse problems, on the one hand, as well as molecular structures and graphs, vibrational spectroscopy and inverse molecular design, on the other hand.

WP2 Stage 1: SMILES/SELFIES-to-spectra (6–9 months). This work package concerns the implementation, adaptation, and fine-tuning of ChemBERTa, ChemGPT, or related models on datasets of smiles and spectra. The obtained models will predict molecular spectra from given SMILES/SELFIES.

WP3 Stage 2: spectra-to-SMILES/SELFIES (18–21 months). This stage constitues the challenge of the PhD. This work package concerns the implementation, enrichment, and fine-tuning of spectro-adapted LLMs on datasets of spectra and smiles. The obtained models will predict SMILES/SELFIES from given molecular spectra. The LLMs will require suitable architectural adaptations and pre-training adjustments to fit our goals.

WP4 Manuscript (6 months). This work package concerns the writing of the PhD manuscript as well as related papers.

2. Collaboration and Supervision Details

2.1. Co-supervision

The PhD will be co-supervised by **Prof. Jérémie Cabessa (DAVID, UVSQ - Paris-Saclay)** and **Prof. Marie-Pierre Gaigeot (LAMBE, UEVE - Paris-Saclay)**.

Jérémie Cabessa, is a Professor in computer science at the laboratoire DAVID, Université Versailles Saint-Quentin-en-Yvelines – Paris-Saclay. His research activities mainly focus on *artificial neural networks* from both a theoretical and an applied perspective, which includes neural computation, computational complexity of neural networks, deep learning, and natural language processing (NLP), and bio-inspired computing.

Marie-Pierre Gaigeot is a Professor in Physics and Chemistry at LAMBE - UMR8587, Université d'Evry Val d'Essonne – Paris Saclay and senior member of the IUF. Her domains of expertise focus on *theoretical and computational chemistry* and *theoretical vibrational spectroscopy*, and involve, among others, molecular dynamics simulations and direct modeling of vibrational anharmonic signals. Prof. Gaigeot has already been involved in interdisciplinary projects at the interface of chemistry, graph theory and artificial intelligence (AI), in particular with Prof. Dominique Barth from DAVID, UVSQ - Paris-Saclay.

2.2. Collaboration

The collaboration between Prof. Cabessa and Prof. Gaigeot represents an ideal fit and clear added-value to the proposed project, since their respective research expertises cover the two main aspects of the project: (1) deep learning, neural networks, large language models (LLMs) and graph neural networks (GNNs); and (2) theoretical chemistry and vibrational spectroscopy, molecular structure identification, modeling of vibrational spectroscopies.

More specifically, the contributions of the co-supervisors can be described as follows: The role of Prof. Cabessa concerns the supervision of the computer scientist part of the project. This includes the analysis and formalization of the tasks as well as the development, implementation and pre-training/fine-tuning of the LLMs. The role of Prof. Gaigeot consists in supervising the chemo-informatics component of the project. This involves the identification of relevant molecular spectral and structural features necessary to the achievement of good learning performance, the conception of all processes with respect to their chemical aspects, the participation in the elaboration and implementation of the models, as well as the creation of the molecular datasets.

3. Evaluation criteria

In terms of publications, the outcome of the project is estimated to 2 or 3 conference papers in NN/AI/ML good or top-tier conferences (IJCNN, ECAI, IJCAI, AAAI, NIPS) as well as 1 or 2 journal papers in a good bio-informatics or chemistry journal (PloS Computational Biology, JACS, etc.). More generally, a solution to WP3 would constitute a significant advance in theoretical chemistry regarding the impactful domain of inverse molecular/material design. It would also enable a leap in the field of inverse problems in machine learning.

4. Required skills

The required skills for the cadidate are: strong background in ML and neural networks; very good programming skills, preferably in Python; experience with neural network libraries like PyTorch or Keras is strongly recom-

mended; prior knowledge in chemo-/bio-informatics would be a plus.

5. Implementation details

The candidate will be hosted by the laboratoire DAVID at UVSQ-Paris Saclay: https://www.david.uvsq.fr/. An office space, a computer and an access to GPU computing resources will be granted. Funding to attend summer schools, conferences or workshops is available to the candidate. The monthly salary is in line with the official salary scale from the public sector. Salary can be increased upon additional teaching duties.

6. References

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

[2] Alex Zunger. Inverse design in search of materials with target functionalities. *Nature Reviews Chemistry*, 2(4):0121, 2018.

[3] Patrick Reiser, Marlen Neubert, André Eberhard, Luca Torresi, Chen Zhou, Chen Shao, Houssam Metni, Clint van Hoesel, Henrik Schopmans, Timo Sommer, and Pascal Friederich. Graph neural networks for materials science and chemistry. *Communications Materials*, 3(1):93, Nov 2022.

[4] Daniel C. Elton, Zois Boukouvalas, Mark D. Fuge, and Peter W. Chung. Deep learning for molecular design—a review of the state of the art. *Mol. Syst. Des. Eng.*, 4:828–849, 2019.

[5] Victor Fung, Jiaxin Zhang, Guoxiang Hu, P. Ganesh, and Bobby G. Sumpter. Inverse design of two-dimensional materials with invertible neural networks. *npj Computational Materials*, 7(1):200, Dec 2021.

[6] Niklas W. A. Gebauer, Michael Gastegger, Stefaan S. P. Hessmann, Klaus-Robert Müller, and Kristof T. Schütt. Inverse design of 3d molecular structures with conditional generative neural networks. *Nature Communications*, 13(1):973, 2022.

[7] Lei Huang, Hengtong Zhang, Tingyang Xu, and Ka-Chun Wong. Mdm: Molecular diffusion model for 3d molecule generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4):5105–5112, Jun. 2023.

[8] Kim Greis, Carla Kirschbaum, Gert von Helden, and Kevin Pagel. Gas-phase infrared spectroscopy of glycans and glycoconjugates. *Current Opinion in Structural Biology*, 72:194–202, 2022.

[9] Anouk M. Rijs and Jos Oomens. *Gas-Phase IR Spectroscopy and Structure of Biological Molecules*. Springer Cham, 2015.

[10] Maike Lettow, Márk'o Grabarics, Eike Mucha, Daniel A. Thomas, Lukasz Polewski, Joanna Freyse, Jörg Rademann, Gerard Meijer, Gert von Helden, and Kevin Pagel. Ir action spectroscopy of glycosaminoglycan oligosaccharides. *Analytical and Bioanalytical Chemistry*, 412(3):533–537, 2020.

[11] Irina Dyukova, Eduardo Carrascosa, Robert P Pellegrinelli, and Thomas R Rizzo. Combining cryogenic infrared spectroscopy with selective enzymatic cleavage for determining glycan primary structure. *Anal Chem*, 92(2):1658–1662, January 2020.

[12] Marie-Pierre Gaigeot and Riccardo Spezia. Theoretical methods for vibrational spectroscopy and collision induced dissociation in the gas phase. *Top Curr Chem*, 364:99–151, 2015.

[13] Shouning Yang, Qiaoling Zhang, Huayan Yang, Haimei Shi, Aichun Dong, Li Wang, and Shaoning Yu. Progress in infrared spectroscopy as an efficient tool for predicting protein secondary structure. *International Journal of Biological Macromolecules*, 206:175–187, 2022.

[14] Marie-Pierre Gaigeot. Some opinions on md-based vibrational spectroscopy of gas phase molecules and their assembly: An overview of what has been achieved and where to go. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 260:119864, 2021.

[15] Ch. Affolter and J.T. Clerc. Prediction of infrared spectra from chemical structures of organic compounds using neural networks. *Chemometrics and Intelligent Laboratory Systems*, 21(2):151–157, 1993.

[16] Sheng Ye, Kai Zhong, Jinxiao Zhang, Wei Hu, Jonathan D. Hirst, Guozhen Zhang, Shaul Mukamel, and Jun Jiang. A machine learning protocol for predicting protein infrared spectra. *Journal of the American Chemical Society*, 142(45):19071–19077, 2020. PMID: 33126795.

[17] Charles McGill, Michael Forsuelo, Yanfei Guan, and William H. Green. Predicting infrared spectra with message passing neural networks. *Journal of Chemical Information and Modeling*, 61(6):2594–2609, 2021. PMID: 34048221.

[18] Markus C. Hemmer and Johann Gasteiger. Prediction of three-dimensional molecular structures using information from infrared spectra. *Analytica Chimica Acta*, 420(2):145–154, 2000.

[19] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, Aug 2021.

[20] David Weininger. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, Feb 1988.

[21] Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-referencing embedded strings (selfies): A 100string representation. *Machine Learning: Science and Technology*, 1(4):045024, oct 2020.

[22] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models. *CoRR*, abs/2303.18223, 2023.

[23] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *CoRR*, abs/2010.09885, 2020.

[24] Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta-2: Towards chemical foundation models. *CoRR*, abs/2209.01712, 2022.

[25] Nathan C. Frey, Ryan Soklaski, Simon Axelrod, Siddharth Samsi, Rafael Gómez-Bombarelli, Connor W. Coley, and Vijay Gadepally. Neural scaling of deep chemical models. *Nature Machine Intelligence*, 5(11):1297–1305, Nov 2023.

[26] Bowen Jin, Gang Liu, Chi Han, Meng Jiang, Heng Ji, and Jiawei Han. Large language models on graphs: A comprehensive survey. *CoRR*, abs/2312.02783, 2023.