SOUND.AI

CALL 2 – 2023-24

APPLICATION FORM – v. 27/09/2023

PROJECT PROPOSAL

Title of the Doctoral Research Project:

Enhancing Biofuel Potential: Advanced Characterization of Lipids from Algal Biomass Using Artificial Intelligence

Thesis supervisor and supervisory team:

Main supervisor – Project Leader	
Surname	Miteva
First name	Tsveta
Doctoral school of affiliation	ED388
HDR	NO
Research unit	Laboratoire de Chimie Physique - Matière et Rayonnement - UMR7614 - CNRS/Sorbonne Université - Richard TAÏEB Team B1 SPEED - Spectroscopy, electronic structure and dynamics - Stéphane CARNIATO
Professional address	Sorbonne University, Pierre and Marie Curie Campus 4 Place Jussieu, 75005 Paris
Email	tsveta.miteva@sorbonne-universite.fr
Phone	+33(0)1 44 27 66 33
Doctoral students currently supervised by the thesis director	0

Co-supervisor (if applicable)	
Surname	Friha
First name	Hela
Doctoral school of affiliation	Gustave Eiffel University, COSYS/IMSE, 5 Bd Descartes 77454, Champs sur Marne, France. ED 532: École Doctorale Sciences, Ingénierie et Environnement (in the process of joining the ED 532 doctoral school)
HDR	NO
Research unit	Département Composants-Systèmes COSYS – Dir.: N. Hautière Laboratoire Instrumentation, Modélisation, Simulation et Expérimentation (IMSE) – Dir.: R. Linguerri Research associate (in progress) at Physico-chimie, nanoscience, et technologies quantiques – Resp. M. Hochlaf
Professional address	IPSA - École d'ingénieurs Aéronautique et Spatiale Paris 63 Bd de Brandebourg Bis, 94200 Ivry-sur-Seine
Email	hela.friha@ipsa.fr
Phone	+33(0)1 84 07 15 00
Doctoral students currently supervised by the thesis director	- None - Regular supervision of master students (M2)

Co-supervisor (if applicable)	
Surname	Hochlaf
First name	Majdi
Doctoral school of affiliation	Gustave Eiffel University, COSYS/IMSE, 5 Bd Descartes 77454, Champs sur Marne, France. ED 532: École Doctorale Sciences, Ingénierie et Environnement
HDR	YES
Research unit	Département Composants-Systèmes COSYS – Dir.: N. Hautière Laboratoire Instrumentation, Modélisation, Simulation et Expérimentation (IMSE) – Dir.: R. Linguerri Research team Physico-chimie, nanoscience, et technologies quantiques – Resp. M. Hochlaf
Professional address	5 boulevard Descartes Champs-sur-Marne 77454 Marne-la-Vallée cedex 2
Email	majdi.hochlaf@univ-eiffel.fr
Phone	+33(0)1 84 07 15 00
Doctoral students currently supervised by the thesis director	A. Gloriod (ED 532)

DETAILED DOCTORAL RESEARCH PROJECT

SUMMARY

Given the urgent need to reduce greenhouse gas emissions, particularly CO_2, alternative approaches are being explored, such as capturing CO_2 from the air and converting it into value-added products (VAPs). Microalgae are more effective than terrestrial plants in bio-fixing CO_2 by assimilating it into algal biomass, which contains valuable components like lipids, sugars, proteins, and pigments. This biomass is a valuable resource for renewable bioenergy and can find applications in various industries. However, the high equipment costs limit the application of VAPs as biofuels, resulting in higher prices compared to fossil fuels. One solution is optimizing experimental photo-bioreactors through detailed characterization of algal biomass using Infrared (IR) and Raman spectroscopies. To address this challenge, this project aims to establish a theoretical protocol using machine learning (ML) algorithms to characterize algal biomass by predicting band origins, intensities, and vibrational frequencies of IR and Raman spectra. To accomplish this goal, a large dataset of molecular spectra of biofuel organic species will be generated using advanced conformational sampling methods and standard quantum-chemical approaches like density functional theory (DFT). The obtained theoretical model will be applied to predict the optimal experimental conditions for microalgal cultivation which maximize the production of lipids by representative microalgae. The outcomes of this project will contribute to the development of non-invasive methods for mapping VAPs in vivo, but also to the analysis of mixtures of compounds in planetary atmospheres, interstellar media, and combustion products.

TITLE

Enhancing Biofuel Potential: Advanced Characterization of Lipids from Algal Biomass Using Artificial Intelligence

MAIN SUPERVISOR

Dr Tsveta MITEVA will be the coordinator of the project. She is a CNRS researcher at LCP-MR since 2018. Dr Miteva obtained her doctoral degree in theoretical chemistry in 2015 at the University of Heidelberg. She has been working on electronic decay processes in the gas and liquid phase since 2010. Since 2021, the coordinator has been actively working on the data-driven hyperspectral unmixing in painted works of art and since 2023, on the theoretical characterization of algal biomass for biofuels. In 2023, Dr Miteva has supervised two master students on the latter project together with Dr Friha. For the last four years, the coordinator has consecrated a large part of her time on training in machine learning (ML) and deep learning (DL), which are nowadays becoming widely used techniques in multiple fields, including theoretical chemistry and spectroscopy. She successfully completed a training program on ML and Artificial Intelligence proposed by Sorbonne Center of AI (Dec, 2020 – Jul, 2021). To date, the coordinator has 38 papers in international peer reviewed journals, out of which one in Nature Communications, three in Physical Review Letters, and one in the Journal of Physical Chemistry Letters. She has co-authored 4 publications on applications of ML in Quantum Chemistry[1,2]. Since the beginning of 2023, Dr Miteva has had a fruitful collaboration with Dr Friha and Prof Hochlaf, the results of which have been presented on an international conference on Biofuels 2023 in London, UK. Dr Miteva will actively supervise the doctoral student.

Allowing for the realization of the coordinator's research ideas, will significantly boost her scientific career by promoting her independence in research and project management. Her expertise on ML and DL will be an asset for the successful accomplishment of the project's goals. Eventually, bringing a novel topic to LCP-MR and opening up new collaborations with IPSA and Université Gustave Eiffel will allow the coordinator to consolidate her own career pathway, create a new research team, acquire valuable experience in scientific project management and it will open up the path for her Habilitation.

DESCRIPTION OF THE PROJECT

The project will be led by Dr. Miteva, a CNRS researcher with expertise in quantum chemical calculations and the application of Machine Learning (ML) algorithms in quantum chemistry and cultural heritage. The project will be carried out by a doctoral researcher under the supervision of Dr. Miteva, in collaboration with Dr. Friha (Institute of Polytechnic Science and Aeronautics (IPSA)) and Prof. Hochlaf (U. Gustave Eiffel). The experimental part of the project will be carried out under the supervision of Dr. Friha at Laboratoire Bioproduction (IPSA). This project is an interdisciplinary one aiming for the development of ***resources for a sustainable planet***.

Introduction and state-of-the-art

Given the urgent need to reduce greenhouse gas emissions, particularly CO_2, which is responsible for over 50% of the anthropogenic greenhouse effect[3], alternative approaches are being explored. These approaches involve capturing CO_2 from the air and converting it into value-added products (VAPs). Microalgae have been shown to be more effective than terrestrial plants in bio-fixing CO_2[4,5]. When exposed to light, microalgae assimilate CO_2 and convert it into algal biomass, which contains valuable components such as lipids, sugars, proteins, and pigments[6]. This biomass is considered a valuable resource for renewable bioenergy[7,8] and can be utilized across various industries[9]. Despite the potential of VAPs as biofuels, their application is limited due to the high equipment costs, resulting in higher biofuel prices compared to fossil fuels[10]. One possible solution is optimizing experimental photo-bioreactors, which requires a detailed characterization of algal biomass components. Infrared (IR) and Raman spectroscopies have proven to be effective in characterizing biofuel composition[11]. Such experiments are currently conducted by researchers at IPSA, with whom a close collaboration is envisaged. ***One of the main challenges in finding the optimal growth conditions in experimental photobioreactors is deciphering the spectral contributions of individual organic compounds in the mixture, given their relatively low concentration.*** The sheer number and complexity of the molecular species of interest make it virtually impossible to compute the spectra of all molecular conformers.

Project goals

The first goal of this project is to establish a theoretical protocol using ML algorithms for detailed algal biomass characterization. This protocol aims to predict the band origins, intensities, and vibrational frequencies of IR and Raman spectra. To train the model, a large dataset of molecular spectra of biofuel organic species is needed. Multiple conformers of molecules present in the biomass will be generated using state-of-the-art methods for conformational sampling[12], and their Raman/IR spectra will be computed with standard quantum-chemical approaches such as density functional theory (DFT).

The second goal of this project is to find the optimal experimental conditions which maximize the lipid production from microalgae. At this stage, the theoretical protocol developed by the doctoral student will be employed to analyze Raman and Infrared spectra of cultures of the prototype strain under a range of experimental conditions, and the optimal experimental conditions maximizing the production of lipids for biofuels will be identified.

Project workflow

The project's workflow consists of two parts.

* The first part of the project involves 3 work packages (WP) which will be accomplished by the doctoral student under the supervision of Dr Miteva during the first 2 years of the doctoral dissertation. Frequent discussions with Prof. Hochlaf at this stage will ensure its smooth workflow and the successful accomplishment of the project's goals. Occasional discussions with Dr Friha are envisaged to target the most relevant molecular systems to consider.

First, we will create a large dataset of molecular structures and their respective Raman and IR spectra using advanced *ab initio* methodologies in close collaboration with Prof. Hochlaf. Subsequently, this

dataset will be employed to train a ML algorithm for predicting spectra based on molecular structure. The trained model will then be utilized to predict spectra for molecular species whose spectra are unknown. Details of each work package can be found below.

WP I. 1 Generation of conformers for selected molecules in algal biomass

The dataset needed to train the model will be built from the conformers of the 6 most abundant saturated fatty acids present in algal lipid extracts – octanoic, decanoic, dodecanoic, myristic, palmitic and stearic acid[13]. Multiple conformers of these representative molecules will be generated using CREST[12]. The latter tool is based on semiempirical tight-binding methods combined with a meta-dynamics driven search algorithm and was specifically designed for conformational sampling. Two sets of conformers will be generated – in the presence and absence of solvent, the effect of the solvent on the generated conformer ensemble will be studied at this stage, we will use water as a solvent, as it is present in the algal biomass.

WP I. 2 Creation of a large dataset of vibrational frequencies and Raman/IR intensities of the conformers

The size of the generated dataset will be about 10k conformers per molecular species. Clearly, computing the spectra of such a large number of molecules using standard quantum chemistry approaches is unfeasible. Instead, the IR/Raman spectra of a cleaver selection of the aforementioned 60k conformers will be subsequently computed with these approaches, specifically, DFT/B3LYP. The effect of the solvent on the spectra will be evaluated by first computing the Raman and IR spectra of the lowest-energy conformers of each fatty acid in the presence and absence of solvent. The polarizable continuum model[14] will be used to simulate the solvation effect. At this stage, an atomic basis set optimization will be carried out by comparing the calculated spectra with experimental ones. The comparison is an important step in the preparation of the dataset as it provides an estimate of the errors already present in the training data. Moreover, finding an optimal atomic basis set will allow for a rapid generation of training data.

WP I. 3 Development of an ML model for the prediction of vibrational frequencies and Raman/IR intensities

The above generated database of conformers and their respective IR and Raman spectra, will then be used as a training dataset. The input geometry will be represented in the form of a Coulomb matrix[15]. A supervised machine learning approach using neural networks (NNs) will be adopted. At this stage, we plan to test several NN architectures (by varying the hidden layers, activation functions, training epochs) in order to find the model which best describes the input data. It should be noted that since the Coulomb matrix was first introduced, the representation of molecular structures in the form of a Coulomb matrix has become a standard descriptor in ML algorithms used in computational chemistry[16].

The successful completion of WP I.1-3 will result in the creation of a machine learning model capable of predicting Raman/IR spectra for fatty acids.

* The second part of the project involves a single work package which will be accomplished by the doctoral student under the supervision of Dr Miteva and Dr Friha during the 3rd year of the doctoral thesis.

WP II Application of the ML model to optimizing the experimental conditions maximizing lipid production in the photobioreactor

Research has revealed that the green microalga *Botryococcus Brauni*, when cultivated in wastewater, exhibits an increased capacity to produce lipids and liquid hydrocarbons compared to other strains[11,17]. This characteristic renders this microalga the prototype strain for our comprehensive study.

During this stage of the project, Dr Friha will initiate a culture of the prototype strain and perform spectroscopic analysis of the cultivated biomass using IR and Raman spectroscopies. The spectroscopic data collected by Dr Friha will then be analyzed by the doctoral researcher using the theoretical

methodology developed in WP I. This approach will provide detailed information on the chemical composition of molecular species present in the reaction medium and their respective proportions, and will shed light on the experimental conditions which maximize lipid production.

The results obtained in this project will be a first step towards understanding algal biomass composition and the optimal experimental conditions for lipid production by microalgae. This will allow for a wide flexibility in analyzing mixtures of compounds, by, for example simulating the spectra under certain experimental conditions (solvent, temperature). Moreover, using transfer learning, the theoretical model can be applied to the study of other molecules present in the algal biomass. The proposed project therefore represents a crucial step in going beyond the currently available algorithms for the simulation of spectra of chemical compounds mixtures and will open new horizons in other fields, such as in the analysis of the Earth's atmosphere, other planetary atmospheres, interstellar media, combustion products, but also in fields such as cultural heritage and biopharmaceutics.

Funding

Additional funding of 50k€ required for the successful accomplishment of the project will be requested from IPSA within the framework of the Mac-Bioré project (https://www.supbiotech.fr/2021/06/mac-biore-bioreacteur-laboratoire-experimental-recherche/), in which Dr Friha participates. Indeed, IPSA has already participated in co-funding PhD students by contributing by to 50% of the total funding.

Candidate profile

Education: A Master's degree in physical chemistry and applications, with proficiency in Python programming is required. Knowledge in Machine Learning and Deep Learning methods is considered advantageous.

Desired qualities: Precision, autonomy, adaptability, and initiative.

PROPOSED INDUSTRIAL AND/OR INTERNATIONAL SECONDMENTS

During the doctoral thesis, the PhD student will be seconded to IPSA, where he/she will attend the experiments carried out by Dr Friha. The PhD student will participate in the experimental development and cultivation of the microalga, as well as in the spectral analysis of the obtained algal biomass.

REFERENCES

(1) Gao, J. W.; Miteva, T.; Wu, Y.; Wang, J. G.; Dubois, A.; Sisourat, N. *Phys. Rev. A* **2021**, *103* (3), L030803 https://doi.org/10.1103/PhysRevA.103.L030803.

(2) Casier, B.; Carniato, S.; Miteva, T.; Capron, N.; Sisourat, N. *J. Chem. Phys.* **2020**, *152* (23), 234103 https://doi.org/10.1063/5.0009264.

(3) *Global Emissions*. Center for Climate and Energy Solutions. https://www.c2es.org/content/international-emissions/ (accessed 2023-06-28).

(4) Dahiya, A. *Bioenergy: Biomass to Biofuels*; Academic Press, 2014.

(5) Tebbani, S.; Lopes, F.; Filali, R.; Dumur, D.; Pareau, D. CO_2 Biofixation by Microalgae.

(6) Daneshvar, E.; Wicker, R. J.; Show, P.-L.; Bhatnagar, A. *Chem. Eng. J.* **2022**, *427*, 130884 https://doi.org/10.1016/j.cej.2021.130884.

(7) Podkuiko, L.; Kasemets, M.-L.; Kikas, T.; Lips, I. *Environ. Clim. Technol.* **2020**, *24* (3), 188–200 https://doi.org/10.2478/rtuect-2020-0096.

(8) Liu, W.; Fu, D.; Pan, T.; Singh, R. P. *Water* **2021**, *13* (22), 3190 https://doi.org/10.3390/w13223190.

(9) Enamala, M. K.; Enamala, S.; Chavali, M.; Donepudi, J.; Yadavalli, R.; Kolapalli, B.; Aradhyula, T. V.; Velpuri, J.; Kuppam, C. *Renew. Sustain. Energy Rev.* **2018**, *94*, 49–68 https://doi.org/10.1016/j.rser.2018.05.012.

(10) Rafa, N.; Ahmed, S. F.; Badruddin, I. A.; Mofijur, M.; Kamangar, S. *Front. Energy Res.* **2021**, *9*, 749968. https://doi.org/10.3389/fenrg.2021.749968.

(11) Chun, H. J.; Waqued, S.; Thapa, H. R.; Han, A.; Yakovlev, V. V.; Laane, J.; Devarenne, T. P. *J. Mol. Struct.* **2017**, *1129*, 216–221 https://doi.org/10.1016/j.molstruc.2016.09.081.

(12) Pracht, P.; Bohle, F.; Grimme, S. *Phys. Chem. Chem. Phys.* **2020**, *22* (14), 7169–7192. https://doi.org/10.1039/C9CP06869D.

(13) Wu, H.; Volponi, J. V.; Oliver, A. E.; Parikh, A. N.; Simmons, B. A.; Singh, S. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108* (9), 3809–3814 https://doi.org/10.1073/pnas.1009043108.

(14) Barone, V.; Cossi, M.; Tomasi, J. *J. Comp. Chem.* **1998**, *19* (4), 404–417 https://doi.org/10.1002/(SICI)1096-987X.

(15) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. *Phys. Rev. Lett.* **2012**, *108* (5), 058301 https://doi.org/10.1103/PhysRevLett.108.058301.

(16) Elton, D. C.; Boukouvalas, Z.; Butrico, M. S.; Fuge, M. D.; Chung, P. W. *Sci Rep* **2018**, *8* (1), 9059 https://doi.org/10.1038/s41598-018-27344-x.

(17) Weiss, T. L.; Chun, H. J.; Okada, S.; Vitha, S.; Holzenburg, A.; Laane, J.; Devarenne, T. P. *J. Biol. Chem.* **2010**, *285* (42), 32458–32466 https://doi.org/10.1074/jbc.M110.157230.