PROGRAMME INTITUTS ET INITIATIVES

Appel à projet – campagne 2021

Proposition de projet de recherche doctoral (PRD)

SCAI - Sorbonne Center of Artificial Intelligence

Intitulé du projet de recherche doctoral (PRD): Machine learning methods for computational studies in origins of life

Directeur.rice de thèse porteur.euse du projet (titulaire d'une HDR) :

NOM : **SAITTA** Prénom : **A. Marco**
Titre : Professeur des Universités ou
e-mail : marco.saitta@sorbonne-université.fr
Adresse professionnelle : Campus PMC, 23-24/309
(site, adresse, bât., bureau)

Unité de Recherche :
Intitulé : IMPMC
Code *(ex. UMR xxxx)* : UMR 7590

École Doctorale de rattachement de l'équipe (future école doctorale du.de la doctorant.e) : **ED397-Physique Chimie des Matériaux**

Doctorant.e.s actuellement encadré.e.s par la.e directeur.rice de thèse (préciser le nombre de doctorant.e.s, leur année de 1^e^ inscription et la quotité d'encadrement) : 3 (2018, 2019, 2020), tous à 50%, quotité totale 1,5

--

Co-encadrant.e :

NOM : **VUILLEUMIER** Prénom : **Rodolphe**
Titre : Professeur des Universités ou HDR ☒
e-mail : rodolphe.vuilleumier@ens.fr

Unité de Recherche :
Intitulé : PASTEUR
Code *(ex. UMR xxxx)* : UMR 8640

École Doctorale de rattachement : **ED388-ChimiePhysiqueChimieAnalytique ParisCentre**
Ou si ED non Alliance SU :

Doctorant.e.s actuellement encadré.e.s par la.e co-directeur.rice de thèse (préciser le nombre de doctorant.e.s, leur année de 1^{e} inscription et la quotité d'encadrement) : 4 (2017, 2019, 2020, 2020), dont 3 à 50%, quotité totale 2,5

Co-encadrant.e :

NOM :	**Pietrucci**	Prénom :	**Fabio**
Titre :	Maître de Conférences des Universités ou	HDR	☒

e-mail : fabio.pietrucci@sorbonne-universite.fr

Unité de Recherche :
Intitulé : Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie
Code *(ex. UMR xxxx)* : UMR 7590

École Doctorale de rattachement : **ED397-Physique Chimie des Matériaux**
Ou si ED non Alliance SU :

Doctorant.e.s actuellement encadré.e.s par la.e co-directeur.rice de thèse (préciser le nombre de doctorant.e.s, leur année de 1^{e} inscription et la quotité d'encadrement) :3 (2018, 2019, 2020), tous à 50%, quotité totale 1,5

Cotutelle internationale : ☒ Non ☐ Oui, précisez Pays et Université :

Selon vous, ce projet est-il susceptible d'intéresser une autre Initiative ou un autre Institut ?
☐ Non ☒ Oui, précisez IMat - Institut des Matériaux

Description du projet de recherche doctoral *(en français ou en anglais) :*

Ce texte sera diffusé en ligne : il ne doit pas excéder 3 pages et est écrit en interligne simple.

Détailler le contexte, l'objectif scientifique, la justification de l'approche scientifique ainsi que l'adéquation à l'initiative/l'Institut.

Le cas échéant, préciser le rôle de chaque encadrant ainsi que les compétences scientifiques apportées. Indiquer les publications/productions des encadrants en lien avec le projet.
Préciser le profil d'étudiant(e) recherché.

Scientific context

Research in origins of life aims at finding answers to the formidably complex problem of the emergence of life from the modern versions of Charles Darwin's celebrated "primordial soup". One of the most challenging problems in very interdisciplinary research is the emergence of RNA, widely considered a crucial step in the evolution processes leading towards living organisms. The aim of this project is to provide a new understanding of this problem [1,2]: How did RNA, and its elementary constituents, spontaneously form and then polymerize, despite being very unstable to hydrolysis [3,4] under realistic prebiotic conditions? Those simple yet complex questions require a quantitative understanding of bio/geochemical reactions occurring over a broad range of time and length scales as well as thermodynamic and chemical conditions. From this point of view, it is generally considered that ab initio computational prebiotic chemistry can hardly provide quantitative answers, as its typical size- and time scales are too small to fully take into account this complexity.

This doctoral project is based on these very premises, as we wish to tackle this grand challenge, i.e. the formation and stability of RNA in abiotic conditions, by using a novel combination of state-of-the-art free-energy/chemical space sampling with recent breakthroughs in the development of machine learning potentials. The latter will be designed to combine the accuracy of ab initio electronic structure methods with the efficiency of simple force fields. In order to access the required time- and length scales of the simulations, high-dimensional neural networks will be employed to compute the energies and forces of the atomic configurations. This type of machine learning potential, which we are already developing in-house, thanks to a fruitful collaboration with Jörg Behler (Göttingen) - founder of this field in computational matter - is able to speed up simulations by several orders of magnitude, while maintaining the accuracy of the underlying electronic structure calculations and the ability to make and break bonds.

Our method

Our group, after a major breakthrough in the first computer simulation of the historical Miller experiment [5], has made significant recent advances on the topological description and the ab initio computational modeling of realistic condensed-phase chemical reactions, explicitly taking into account key parameters such as temperature, pressure, pH and addressing inhomogeneous and/or discontinuous systems [6,7]. We have successfully applied this approach on a variety of "elementary" prebiotic chemistry problems: the high-energy chemistry of formamide [8] and formaldehyde [9], the synthesis of small sugars [10], a study of the chemical network of liquid methanol [11], some specific elementary synthesis steps of RNA nucleotides [12-14], and the hydrothermal decomposition of amino acids in meteoritic parent bodies [15], this latter work being solicited by and done in collaboration with scientists at the NASA Goddard Space Flight center. On the other hand, progress in this field through ab initio calculations can only be incremental, due to their high intrinsic computational cost: the development of machine-learning / neural network potentials for atomistic simulations can provide the framework to dramatically upscale the high-accuracy exploration of complex prebiotic chemical networks. Our expertise in the generation of extensive trajectories sampling the phase space of complex chemical systems – together with the structure-energy databases we accumulated in the last years - will be a strategic advantage within the machine learning challenge.

In our HDNNP approach the short-range contribution Es of the total energy Etot is expressed by a sum of atomic energy contributions Ei. The atomic environments are defined by a cutoff radius Rc, which is typically between 6 and 10 Å. The positions of all neighboring atoms inside the resulting cutoff spheres are described by many-body atom-centered symmetry functions Gi [16], which ensure the mandatory rotational, translational and permutational invariances of the potential-energy surface. A vector G of symmetry functions forming a structural fingerprint of the environment of each atom serves as input for atomic feed-forward neural networks, yielding the atomic energy contributions. In addition, the long-range electrostatic energy Ee can be included based on environment-dependent atomic charges, which are expressed by a second set of atomic NNs [17]. The total energy is then the sum of a "short-range" energy Es and the electrostatic energy Ee. The analytic form of the atomic NNs contains a large number of weight parameters a and b, which are optimized iteratively using a reference set of energies and forces obtained from electronic structure calculations. Once the optimum set of parameters has been found, which accurately reproduces these energies and forces, the HDNNP can be used to predict energies and forces of similar structures with about the same accuracy at a fraction of the computational costs. As a rule of thumb about 500 atoms can be computed per second and CPU-core. The method has been implemented in the software package RuNNer, which is freely available as open-source software under the GPL3 license and is continuously further developed [18], including by our group. Thus, all software required to carry out the project is readily available.

Work Plan

We plan to start with important constituents of the AMP nucleotide, specifically adenine (only containing H/C/N atoms), and ribose (only containing H/C/O atoms). The separate study of the degradation of adenine and ribose, by using our established approach, will provide meaningful reaction channels for the inverse reactions, i.e. the ones starting with HCN and formaldehyde, respectively. Importantly, this will make the construction of (initially separate) HDNNPs more manageable ensuring a quick start of the project. In parallel, we will study the degradation of the full RNA nucleotide, which will allow to discover new reaction channels, possibly different from the direct condensation of the nucleobase and the sugar, and thus in line with recent experimental studies [19,20]. The individual HDNNPs will then be merged and improved based on the combined data sets provided by the degradation AIMD trajectories, and will then be used to study, with our free-energy approach, the synthesis reactions, starting with the decomposition products found in the previous steps. The significant gains in length- and time-scales that will be obtained from the HDNNPs-based FES calculations with respect to their fully ab initio counterparts will allow to efficiently and quantitatively sample multiple reaction channels and environmental conditions. The role of the PhD student will be to train the neural networks by generating new trajectories and by exploiting the existing ones, and then to carry out the neural network calculations to determine the free energy of these chemical reactions.

The objectives of this project are challenging, as they involve different disciplines (artificial intelligence, physics, chemistry, but also a biology/Earth science perspective), but promise to be successful by combining the “traditional” and “recent” expertise of the research team, which is a leader in the field of computational prebiotic chemistry and has recently invested in the development of machine-learning-based potentials. This combination will provide precisely the kind of synergy that is needed for such an ambitious research project to succeed, and is a unique opportunity in order to achieve a better understanding of the environments relevant to the origins of life. Quite naturally this project falls perfectly within the perimeter of the SCAI.

[1] S. A. Benner, H.-J. Kim, M. A. Carrigan, Asphalt, Water, and the Prebiotic Synthesis of Ribose, Ribonucleosides, and RNA, Acc. Chem. Res. 45 (2012) 2025.

[2] D. Kopetzki, M. Antonietti, Hydrothermal formose reaction, New J. Chem. 35 (2011) 1787.

[3] N. El-Murr, M.-C. Maurel, M. Rihova, J. Vergne, G. Hervé, M. Kato, K. Kawamura, Behavior of a hammerhead ribozyme in aqueous solution at medium to high temperatures. Naturwissenschaften 99 (2012) 731.

[4] L. Da Silva, M.-C. Maurel, D. Deamer, Salt-Promoted Synthesis of RNA-like Molecules in Simulated Hydrothermal Conditions, J. Mol. Evol. 80 (2015) 86.

[5] A. M. Saitta, F. Saija, Miller experiments in atomistic computer simulations, PNAS 111 (2014) 13768.

[6] A. Pérez-Villa, F. Pietrucci, A. M. Saitta, Prebiotic chemistry and origins of life research with atomistic computer simulations, Phys. Life Rev. Adv. Article, DOI: 10.1016/j.plrev.2018.09.004 (2018).

[7] F. Pietrucci, A. M. Saitta, Formamide reaction network in gas phase and solution via a unified theoretical approach: towards a reconciliation of different prebiotic scenarios, PNAS 112 (2015) 15030.

[8] M. Ferus, F. Pietrucci, A. M. Saitta et al., Formation of nucleobases in a Miller-Urey reducing atmosphere, PNAS 114 (2017) 4306.

[9] M. Ferus et al., Prebiotic synthesis initiated in formaldehyde by laser plasma simulating high-velocity impacts, Astronomy & Astrophys. 626 (2019) A52.

[10] G. Cassone, J. Sponer, J. E. Sponer, F. Pietrucci, A. M. Saitta, F. Saija, Synthesis of (D)-erythrose from glycolaldehyde aqueous solutions under electric field, Chem. Comm. 54 (2018) 3211.

[11] G. Cassone, F. Pietrucci, F. Saija, F. Guyot, A. M. Saitta, One-step electric-field driven methane and formaldehyde synthesis from liquid methanol, Chem. Sci. 8 (2017) 2329.

[12] A. Pérez-Villa, A. M. Saitta et al., Synthesis of RNA nucleotides in plausible prebiotic conditions from ab initio computer simulations, J. Phys. Chem. Lett. 7 (2018) 4981.

[13] G. Cassone, J. Sponer, F. Saija, E. Di Mauro, A. M. Saitta, J. E. Sponer, Stability of 2 ',3 ' and 3 ',5 ' cyclic nucleotides in formamide and in water, Phys. Chem. Chem. Phys. 19 (2017) 1817.

[14] J. E. Sponer et al., Prebiotic synthesis of nucleic acids and their building blocks at the atomic level - merging models and mechanisms from advanced computations and experiments, Phys. Chem. Chem. Phys. 18 (2016) 20047.

[15] F. Pietrucci, J. C. Aponte, A. Perez-Villa, J. E. Elsila, J. P. Dworkin, A. M. Saitta, Hydrothermal decomposition of amino acids and origins of prebiotic meteoritic organic compounds, ACS Earth Space Chem 6 (2018) 588.

[16] T. Morawietz, A. Singraber, C. Dellago, J. Behler, How van der Waals interactions determine the unique properties of water, PNAS 113 (2016) 8368.

[17] N. Artrith, T. Morawietz, J. Behler, High-dimensional neural-network potentials for multicomponent systems: Applications to zinc oxide, Phys. Rev. B 83 (2011) 153101.

[18] RuNNer – A program for constructing high-dimensional neural network potential-energy surfaces, J. Behler, Universität Göttingen, 2020.

Homepage: http://www.uni-goettingen.de/de/560580.html

[19] S. Islam, D.-K. Bucar, M. W. Powner, Prebiotic selection and assembly of proteinogenic amino acids and natural nucleotides from complex mixtures, Nature Chem. 9 (2017) 584.

[20] S. Becker et al., Unified prebiotically plausible synthesis of pyrimidine and purine RNA ribonucleotides, Science 366 (2019) 76.

SORBONNE UNIVERSITÉ

Merci d’enregistrer votre fichier au format PDF et de le nommer :
«ACRONYME de l’initiative/institut – AAP 2021 – NOM Porteur.euse Projet »

Fichier envoyer simultanément par e-mail à l’ED de rattachement et au programme :
cd_instituts_et_initiatives@listes.upmc.fr avant le 20 février.