Report on the outcomes of a Short-Term Scientific Mission[1]

Action number: CA20129

Grantee name: Louey Charradi

Details of the STSM

Title: Machine Learning based methodology towards the determination of the abundances of C60+ in astrophysical media

Start and end date: 15/05/2025 to 13/07/2025

Description of the work carried out during the STSM

Description of the activities carried out during the STSM. Any deviations from the initial working plan shall also be described in this section.

The STSM focused on developing a comprehensive computational framework for characterizing the C60+-He complex, a system of significant astrophysical importance due to C60+'s role as a carrier of diffuse interstellar bands (DIBs) observed at wavelengths 9577 Å and 9632 Å. The research encompassed three main components: systematic quantum chemical method evaluation, potential energy surface (PES) generation, and advanced machine learning model development with rigorous validation.

An exhaustive comparison of quantum chemical methods was conducted to identify the optimal computational approach for this challenging 61-atom system. Semi-empirical methods were systematically tested: GFN-FF yielded -95.1 cm^{-1} binding energy at ~6.3 Å but lacked quantum mechanical effects; GFN1-xTB achieved -83.2 cm^{-1} with basic D3 dispersion; GFN2-xTB provided -129.8 cm^{-1} with advanced D4 dispersion model. Advanced DFT methods using PBE0 functional encountered severe convergence issues: aug-cc-pVDZ runs failed due to basis set linear dependence and incorrect overlap matrices, while cc-pVDZ produced unrealistic energies (-1.3E+07 a.u.) from KS-SCF instabilities. Wavefunction-based methods faced similar challenges: MP2 succeeded with cc-pVDZ (-2279.420 a.u., -13.445 cm^{-1} interaction energy) but failed with aug-cc-pVDZ due to RHF-SCF convergence failure and cc-pVTZ due to numerical instability. CCSD, CCSD(T), MP2-F12, and CCSD(T)-F12 , but such computations were too large and did not run even in larger computational clusters is addition to coupled-cluster amplitude convergence issues and F12 correction divergence. This extensive debugging process, requiring meticulous analysis of convergence criteria and basis set dependencies, ultimately validated GFN2-xTB as the optimal balance of accuracy and computational feasibility. I used several packages such as CREST 3.0.2 (Pracht, P.; Grimme, S.; Bannwarth, C.; et al.), MOLPRO 2015 (Werner, Knowles,

[1] This report is submitted by the grantee to the Action MC for approval and for claiming payment of the awarded grant. The Grant Awarding Coordinator coordinates the evaluation of this report on behalf of the Action MC and instructs the GH for payment of the Grant.

Manby, et al.) and GAUSSIAN(16 (Frisch et al). Anyway, these benchmarks allowed validating the use of GFN2-xTB for the generation of the interaction potential between He and C_{60}^{+}.

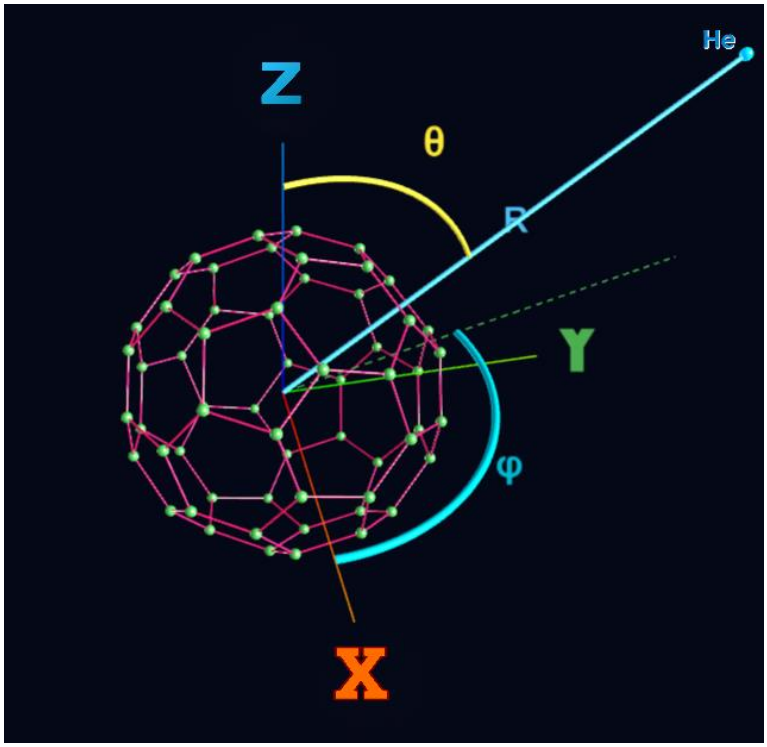

***Figure 1:**: Jacobi coordinate system for the C_{60}^{+}-He*

Using CREST conformational sampling with GFN2-xTB, approximately 10,000 PES data points were systematically generated across intermolecular distances (R: 4.5-50.0 Å) and full angular coverage (Theta: 0-180°, Phi: 0-360°). The dataset captured the complex interplay of dispersion forces, electrostatic interactions, and quantum mechanical effects, with C_{60}^{+}'s icosahedral symmetry creating characteristic energy patterns corresponding to helium approach toward molecular faces, edges, and vertices.

Afterwards, A CatBoost model was built using 43 physics-informed features derived from intermolecular interaction principles. Distance-based terms included van der Waals (r^{-6}), electrostatics (r^{-1}), Pauli repulsion (r^{-12}), and Lennard-Jones potentials (σ = 3.0 Å, ε = 1.0 eV). Angular features captured icosahedral symmetry via trigonometric terms, spherical harmonics (Y_1^0–Y_2^2), and anisotropy descriptors. Composite features combined radial-angular couplings, energy scaling, and regime indicators. Optuna tuning over 50 trials identified optimal hyperparameters: 4,276 iterations, 0.031 learning rate, and depth 7

REFERENCES:

- *Pracht, P.; Grimme, S.; Bannwarth, C.; et al., "CREST—A program for the exploration of low-energy molecular chemical space." Journal of Chemical Physics, **160**, 114110 (2024).*
- *Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., et al. Gaussian 16, Revision B.01, Gaussian, Inc., Wallingford CT, 2016.*
- *Werner, H.-J., Knowles, P. J., Manby, F. R., et al. "The Molpro quantum chemistry package." Journal of Chemical Physics 152, 144107 (2020)45.*

Description of the STSM main achievements and planned follow-up activities

Description and assessment of whether the STSM achieved its planned goals and expected outcomes, including specific contribution to Action objective and deliverables, or publications resulting from the STSM. Agreed plans for future follow-up collaborations shall also be described in this section.

The STSM achieved exceptional results, successfully establishing a robust computational methodology for astrophysically relevant large molecular complexes. The CatBoost model demonstrated outstanding performance with R^2 = 0.9999, MAE = 8.36 cm^{-1}, and RMSE = 38.44 cm^{-1} across the entire energy range. Regional performance analysis revealed excellent accuracy across all interaction regimes: Pauli repulsion region (MAE: 27.4 cm^{-1} despite energies reaching 20,000+ cm^{-1}), equilibrium region (MAE: 1.8 cm^{-1}), and long-range dispersion region (MAE: 1.3 cm^{-1}).

Surface cut analyses at R = 6 Å and Theta = 40° validated the model's ability to capture both radial and angular dependencies, accurately reproducing the icosahedral symmetry patterns and energy contours of the actual quantum mechanical surface. Feature importance analysis identified intermolecular distance (R), angular anisotropy (anisotropy_phi), and combined interaction terms (energy_scale) as primary predictive factors.

The work makes significant contributions to MultIChem COST Action objectives, specifically addressing WG1's focus on fundamental understanding of complex molecular systems through atomistic-level theoretical and multiscale computational modelling. The developed methodology demonstrates: (1) advanced multiscale modelling approaches combining quantum chemistry with machine learning for complex molecular interactions, (2) validated computational protocols that can be adapted for irradiation-driven chemistry studies, and (3) successful integration of physics-informed AI techniques with traditional quantum mechanical methods, contributing to WG1's task of developing general multiscale methodologies.

The comprehensive dataset of ~10,000 interaction energies and the validated ML model contribute to WG1's data handling objectives, providing a framework for constructing molecular interaction databases. The methodology's transferability to other fullerene-noble gas systems aligns with WG1's emphasis on exploring complex molecular systems and their interactions.

Supporting WG4's dissemination and training objectives, the research outcomes will be communicated through: manuscript preparation for high-impact journals focusing on computational methodology and astrophysical applications, presentation at MultIChem annual conferences, and contribution to training materials for Early Career Investigators. The integration of traditional quantum chemistry with modern ML approaches provides excellent educational content for Training Schools. For instance, I will be attending the NanoSpace AI in Astrochemistry Training School this August at Aalto University in Helsinki, which offers a unique opportunity to explore these interdisciplinary methods through lectures, hands-on exercises, and collaborative discussions.

Planned follow-up activities include: extension to other C_{60}^+-H_2 complex to build systematic interaction databases, integration with experimental spectroscopic data to validate computational predictions, development of molecular dynamics simulations using the trained ML potential, and collaboration with WG1 members to adapt the methodology for radiation-induced molecular transformations. We thus working in implementing these PESs in MOLSCAT code in order to compute the collision cross sections and deduce the collisional rates, which are needed to better estimate the abundance of C_{60}^+ in interstellar medium. A collaborative framework has been established with Prof. Hochlaf's group for continued research supporting both computational methodology development and experimental validation studies within the MultIChem network.