Ruben STAUB | Theoretical Chemist

+33 6 11 37 73 60 • ruben.staub@ens-lyon.fr
perso.ens-lyon.fr/ruben.staub/ • RubenStaub

Theoretical chemist with a strong taste for Computer Science, I am passionate about designing models and computational tools for solving chemical challenges.

Education

Diplomas

- **PhD in Theoretical Chemistry** — **ENS de Lyon**
  Ongoing, Development of computational tools for simulating reactive interfaces and applications to heterogeneous catalysis — *2017–2020*
- **Master's degree in Chemistry & Physics** — **ENS de Lyon**
  Science of Matter program (open), 78+60 ECTS credits, avg. grade: 15,6/20 — *2014–2015/2016–2017*
- **Master's degree in Chemoinformatics** — **University of Strasbourg**
  Double diploma, top of my class, avg. grade: 17.4/20 — *2016–2017*
- **Bachelor's degree in Computer Science** — **ENS de Lyon**
  Admission in L3, avg. grade: 14,6/20 — *2015–2016*

Internships

- **Development of Energy Decomposition Analysis tools for catalysis** — **M2 Chemistry internship**
  Laboratoire de Chimie, ENS de Lyon — *February 2017–July 2017*

  I crafted a new theory unifying the Absolutely Localized Molecular Orbital formalism with mixed-states theory, and I implemented it as a patch for the open-source simulation software CP2K (Fortran 2003).
  - Supervisor: Stephan Steinmann

- **Graph isomorphism detection for Chemoinformatics** — **L3 Computer Science internship**
  Laboratoire DAVID, University of Versailles — *June 2016–July 2016*

  I designed and implemented various optimisations for the isomorphism detection on planar maps, into a C-based open source project for the combinatorial generation of molecular cages.
  - Supervisor: Yann Strozecki

- **Molecular simulation of a metal-organic interface** — **M1 Chemistry internship**
  Physical-Chemistry Institute, University of Zurich — *May 2015–July 2015*

  I performed a theoretical study on the chemisorption of 2-azidoethanethiol onto a Au(111) surface, using CP2K.
  - Supervisor: Marcella Iannuzzi

- **Development of analysis tool for Metabolomic** — **L3 Chemistry internship**
  CRMN, Lyon — *June 2014–July 2014*

  I developed a Python script for analysing multidimensional NMR spectra by performing peak attribution using a metabolomic database.
  - Supervisor: Bénédicte Éléna-Hermann

Skills

- **Languages:** French (native tongue) ; English ; German and Spanish (extremely limited) ; learning Japanese.
- **Main programming languages:** Python3 ; Bash.
  Have also worked with: HTML/CSS ; PHP/MySQL ; C/C++ ; Javascript ; Fortran ; ...
- **Soft skills:** Scientific communication ; Teamwork ; Geographical mobility ; Interdisciplinarity and high versatility ; First aid and CPR certification ; ...

Extra-curricular projects

- **Web services:** I love tackling new challenges through hand-crafted web applications ranging from game theory for daily decisions to embedded domotics solutions (RPi + old smartphone), and including various grading tools, internal bibliography sharing platform for my lab, ...
  I have developed in collaboration with mathematics teachers an online platform for e-learning through quizzes, based on models and algorithms that I designed to optimize the learning curve of students and provide teachers with pedagogy insights (through basic machine learning and statistical inferences)
- **Competitions :** I am always looking for more challenges: International Chemistry Olympiads ; ByteCode ; CodeJam ; BattleDev ; Google Hash Code.
- **Scientific and technological popularisation:** Active participation to the annual French Science Festival ; Volunteer at Paris Ubuntu-parties (Linux popularisation, assistance, installations) ; Weekly presentation of Bash/Python features during lab group meetings ; Helping colleagues through Bash/Python scripts ; Teaching Quantum mechanics and Numerical toolbox and programming for Physics and Chemistry undergraduates ; Animating annual initiations for school/lab newcomers to Bash/Linux, LaTeX, Python, ...

Research experience

Publications and presentations

- **Energy Decomposition Analysis for Metal Surface–Adsorbate Interactions by Block Localized Wave Functions** — **JCTC**
  Ruben Staub, Marcella Iannuzzi, Rustam Z. Khaliullin and Stephan N. Steinmann — *November 2018*

  Mean-field generalisation of the Absolutely Localized Molecular Orbital framework to mixed-states. This allows for a rigorous analysis of charge transfer involving metallic surfaces, at the DFT level. Application to energy decomposition analysis (EDA) of various adsorbates on Pt(111). More recently, this tool was also applied to gain insights from EDA of water adlayers on noble metal surfaces.
  - Oral presentations: Congress SCF 2018, Montpellier (France) ; CP2K Day, Lyon (France) ; ROAD4CAT Symphonium, Lyon (France)
  - Article link: https://doi.org/10.1021/acs.jctc.8b00957
  - Application to water adlayers on noble metal surfaces: https://doi.org/10.26434/chemrxiv.12263117.v1

- **Parameter-free coordination numbers for solutions and interfaces** — **J. Chem. Phys.**
  Ruben Staub and Stephan N. Steinmann — *January 2020*

  Adaptation of the Solid-Angle Nearest-Neighbors algorithm to tackle local anisotropy, through a simple yet efficient correction term. This work provides with highly adaptive coordination numbers, particularly relevant for interfaces. Application to coordination analysis of cations near a graphite anode and coordination-based model Hamiltonian construction for Au-Cu alloys nanoparticles.
  - Oral presentation: CTTC VIII, Krakow (Poland)
  - Poster presentation: JTMS 2019, Paris (France)
  - Article link: https://doi.org/10.1063/1.5135696

- **Efficient Recursive Least Squares for Rank-Deficient Matrices** — **Appl. Math. Comput.**
  Ruben Staub and Stephan N. Steinmann — *Pending submission*

  Derivation of a simple formula for updating the Moore-Penrose pseudoinverse of a matrix after adding a row. Using an efficient subspace separation, this reformulation allows for the computation of the least squares solution x to the equation $Ax = b$, where A is an $n \times m$ matrix of rank k, in $O(nmk)$. A single update step being computed in $O(mk)$, such algorithm is particularly relevant for updating a model Hamiltonian with a large feature space (underdetermined) and input data redundancy (overdetermination of relevant features).

- **Reinforcement sampling for model Hamiltonian construction**
  Ruben Staub and Stephan N. Steinmann — *Ongoing project*

  Introducing and adapting reinforcement learning techniques to model fitting in Chemistry. This project tackles the automatic and iterative generation of input dataset for model Hamiltonians handling many-body effects. A typical model Hamiltonian training set is composed of hundreds of structures hand-picked by a chemist, using its expertise and chemical intuition to select a diverse set of relevant inputs. Treating the generation of a relevant input as a strategy-based game, a novel UCT-based approach is designed, effectively replacing chemical intuition with reinforcement learning. A carefully crafted pre-exploration step is developed and added to further optimize the well-known UCT framework in this context. This novel extension exploits the partially trained model itself to provide with domain knowledge while maximizing input usefulness for the model, effectively preventing local minima trapping.
  - Oral presentation at an interdisciplinary thematic meeting on Machine Learning, Lyon (France)
  - Poster presentation at the Annual Meeting of Chemistry Graduate School 2019, Lyon (France)

Past and future research project

Early in my research, I was interested in studying the chemical bond from a very fundamental point of view: Chemical bonding can be described using the DFT framework as electronic density reorganisation between chemical entities. This reorganisation can be naturally decomposed into polarisation (reorganisation within the defined entities) and charge-transfer (reorganisation in-between the entities). This fundamental decomposition can be performed the most rigorously using the ALMO (Absolutely Localized Molecular Orbitals) formalism. However, the ALMO formalism becomes completely intractable when combined with mixed-states theory, which is required for describing metals. Therefore, I unified these theories into a mean-field approximation called S-ALMO, allowing for a fundamental density-based description of most chemical bonds.

In practice, however, bonds are commonly defined by geometrical considerations instead of density-based analyses. As a consequence, I was interested in a fundamental definition of bonding from a purely topological point of view. Most coordination algorithms rely on predefined cutoffs to define chemically relevant bonds. Only a few algorithms try to capture such information through the local topological environment and therefore define parameter-free bonds solely based on topological considerations, the state-of-the-art being the SANN (Solid Angle based Nearest Neighbours) algorithm. Based on an intuitive and powerful idea (an atom's nearest neighbours must cover its field of view), SANN provides with locally adaptive parameter-free coordination numbers. However, I showed that such algorithm is only valid to describe locally isotropic structures (i.e. mainly bulk of closed packed materials). Nonetheless, I developed a natural extension of this algorithm, called ASANN, able to tackle simple local anisotropy while remaining parameter-free. This method effectively provides with a fundamental topology-based definition of bonds applicable to close packed bulks, liquids and interfaces.

Even though bonds are topologically defined, they contain chemically relevant information, providing remarkable simple topological descriptors for describing energy. I was therefore interested in studying bonds as energy descriptors. The most simple yet rigorous framework linking bonds to energy is probably the cluster expansion based model Hamiltonian, especially in the context of adsorption energy for heterogeneous catalysis. In this framework, the total adsorption energy of an occupied surface is simply the sum of the adsorption energy for each adsorbate taken alone, plus lateral interactions. Lateral interactions are defined as corrective terms for 2-body, 3-body, ... In practice, such expansion can be stopped after the second or third order correction terms, leading to a finite linear formula between multiplicities of topological patterns and total adsorption energy (i.e. a model Hamiltonian). Fitting such model Hamiltonian provide with a fundamental framework for defining the energy contributions of topological patterns.
Such fitting is usually performed on a training set composed of hundreds of structures hand-picked by a chemist. For the model Hamiltonian to be relevant for describing a given reaction (e.g. though Kinetic Monte-Carlo simulations), the chemist uses his own expertise and chemical intuition to select a diverse set of chemically relevant structures. In practice, this selection is rarely optimal and certainly not reproducible/transferable. Treating the generation of a relevant input as a strategy-based game, I designed a novel active learning scheme for model Hamiltonians, based on a UCT (Upper Confidence Tree). This approach effectively replaces chemical intuition with reproducible reinforcement learning techniques. I also developed a domain knowledge extension on top of the UCT framework, in order to further optimize the training set construction and reduce the total required computational time.

For the future, I intend to continue working on high(er) level descriptions, shifting to more advanced Chemoinformatics, while ideally keeping an insightful quantum point of view.
This means working with more complex models, combining high level descriptions and state-of-the-art machine learning methods. I am nonetheless dedicated to never be a passive user. Instead, I care about keeping a deep understanding of underlying models and assumptions, going back to mathematical derivations if needed. In other words, doing research with a strong methodological development flavour.
I also intend to apply the much lesser known reinforcement techniques whenever applicable, especially through active learning schemes in non-enumerable sampling spaces.
In summary, I plan to adapt state-of-the-art machine learning and algorithmic methods to solve chemical challenges.

References contact information

- Stephan Steinmann
  - PhD supervisor
  - +33 4 72 72 81 55 • stephan.steinmann@ens-lyon.fr
- Carine Michel
  - Head of Theoretical Chemistry lab
  - +33 4 72 72 88 47 • carine.michel@ens-lyon.fr