3iA Côte d'Azur
Interdisciplinary Institute for Artificial Intelligence

Postdoctoral research topic

- Title of the proposed topic: **Decoding the energy landscape: understanding and redesigning the emergent properties of a kinetic transition network**
- Research axis of the 3IA: AI for Computa:onal Biology and Bio-Inspired AI
- **Supervisors:**
  - **David J. Wales, University of Cambridge, dw34@cam.ac.uk**
  - **Frederic Cazals. Inria, Frederic.Cazals@inria.fr**
- The laboratory and/or research group: the post-doc will be located at Inria Sophia Antipolis, with regular visits to Cambridge.

Apply by sending an email directly to the supervisors.
The application will include:

- Letter of recommendation of the supervisor indicated above
- Curriculum vitæ including the list of the scientific publications
- Motivation letter
- Letter of recommendation of the thesis supervisor

- Description of the topic:

Decoding the energy landscape: understanding and redesigning the emergent properties of a kinetic transition network

CONTEXT. Describing a potential energy surface in terms of local minima and the transition states that connect them provides a conceptual and computational framework for understanding and predicting observable properties [1]. Landscapes involving competing morphologies support multiple potential energy funnels, which may exhibit characteristic heat capacity features and relaxation time scales. These connections between the organisation of the landscape and structure, dynamics and thermodynamics are universal, and can be extended to the loss landscapes associated with machine learning [2].

Databases of minima and transition states constitute a kinetic transition network (KTN), and there are three principal tools required to understand how observable properties are encoded, namely (1) global optimisation for structure prediction; (2) enhanced sampling of densities of states for thermodynamic properties; (3) rare event techniques to extract global dynamics. This project will advance the state-of-the-art by developing new theory and computational tools for analysis of mechanism and rates in large, ill-conditioned transition networks. In particular, we will gain insight by comparing the organisation of landscapes in biophysics and machine learning, and correlating this structure with the emergent properties.

WORKPLAN

The geometry optimisation techniques employed in basin-hopping global optimisation and characterisation of transition states and pathways between local minima are relatively mature, and these tools will be exploited to construct new databases for the SARS-CoV-2 spike protein and the RNA genome of this virus, and for the loss function associated with neural networks for benchmark problems in machine learning. To extract observable dynamical properties, and the analogues defined for a loss function landscape [2], we will exploit and develop some new tools to predict kinetically relevant paths from the infinite number of routes through a KTN, and calculate phenomenological rates.

The new theory and associated computational tools for analysis of complex kinetic transition networks [3[has three principal components: (1) convergence of observable kinetic properties; (2) calculating observables; (3) coarse-graining the network. The key observables that we wish to calculate for comparison with experiment are moments of the first passage time between the initial and final states of interest. For biomolecules these states are usually the denatured ensemble and the functional form corresponding to the native state. One key observable encoded in the landscape is the first passage time. We must therefore sample the underlying KTN sufficiently to converge the probability distribution function of the passage time. We have derived expressions for pairwise and pathwise measures of the sensitivity to new network connections, which have proved to be very effective in guiding sampling and network convergence [3,6]. A key objective is to optimise this procedure for larger systems and extend it using Bayesian techniques, in combination with the coarse-graining and kinetic path sampling methods described below.

Kinetic path sampling [4] (kPS) uses graph transformation to simplify the description of an escape trajectory from a trapping energy basin. This procedure permits exact and efficient sampling of Markov chains, including higher moments of the passage time. We will combine kPS with the m distinct paths (mDP) algorithm [5] to determine whether the networks support parallel pathways. The mDP method uses a scalable path deviation algorithm to identify the m most kinetically relevant paths in a transition network, where each path is distinguished by a distinct rate-limiting edge.

Methods to reduce the network dimensionality, while preserving the key observables, produce essential gains in efficiency for large networks and highly metastable systems. We will analyse two new coarse-graining methods, namely a moment-based approach based on spectral analysis of the relaxation modes [6], and partial graph transformation. The moment-based approach filters the relaxation modes to preserve the first and second moments of the passage time [6]. In contrast, partial graph transformation lumps together and prunes nodes from the network. Implementing this methodology in the PATHSAMPLE program is a key objective, which will enable us to extract mean first passage times and associated rates in large, poorly conditioned networks associated with broken ergodicity and rare events.

In a complementary approach, we will also explore a method relying on a hierarchical representation of energy landscapes based on topological persistence, a framework providing a nested representation based on barriers [7]. In theory, the derivation of the stationary distribution of a Markov chain can be performed with matrix inversion, at a cost that is cubic in the number of states. This cubic complexity being prohibitive for the systems we aim to study, in a manner akin to importance sampling, we will explore novel algorithms iterating in tandem the calculation of the stationary distribution for a coarse-grained system, and the refinement of this model when novel states and transitions are considered. Practically, we will combine the hierarchical representation of landscapes provided in the Structural Bioinformatics Library [8], and state-of-the-art numerical methods for Markov chains provided in MarmoteCore [9]. In a different context, we have indeed shown recently that such calculations could be performed to convergence within minutes for Markov models with tens of thousands of nodes [10]. We also anticipate that this approach will yield a novel strategy to compare energy landscapes, an important problem for which few effective solutions have been developed [11].

Having developed an efficient and robust framework for understanding the properties of landscapes that feature metastability we will consider how these features emerge from the underlying intermolecular potential or the structure of a neural network. This insight will provide the foundations for our long-term ambition to design target properties through mutations of a molecular system or the architecture of a machine learning problem.

REFERENCES

[1] D. J. Wales, Energy Landscapes, Cambridge University Press, Cambridge (2003).
[2] P. C. Verpoort, A. A. Lee, D. J. Wales. Proc. National Acad. Sci. USA.117, 21857 (2020)
[3] T. D. Swinburne and D. J. Wales, J. Chem. Theory Comput. 16, 26612679 (2020).
[4] D. J. Sharpe and D. J. Wales, J. Chem. Phys. J Chem Phys 153, 024121 (2020).
[5] D. J. Sharpe and D. J. Wales, J. Chem. Phys. 151, 124101 (2019).
[6] D. Kannan, T. D. Swinburne, D. J. Sharpe and D. J. Wales, J. Chem. Phys. 153, 134115 (2020).
[7] F. Cazals, et al. Journal of computational chemistry 36 (2015): 1213-1231.
[8] Cazals, F. and T. Dreyfus. Bioinformatics 33 (2017): 997-1004.
[9] Jean-Marie, Alain. "marmoteCore: a Markov modeling platform." Proceedings of the 11th EAI International Conference on Performance Evaluation Methodologies and Tools. 2017.
[10] A. Sales-de-Queiroz et al, Gene prioritization based on random walks with restarts and absorbing states, to define gene sets regulating drug pharmacodynamics from single-cell analyses, submitted, 2021.
[11] J. M. Carr. et al. J. Chem. Phys. 144 (2016): 054109.