

Water as an active component of allosteric communication in protein phosphatases

Master 2 research project

Scientific context

Allosteric regulation allows a perturbation in one region of a protein to modify the structure, dynamics, or activity of a distant functional site. Molecular dynamics (MD) simulations provide access to the conformational ensembles underlying these processes.

We analyze long-range communication using **MDiGest**, a Python framework that extracts dynamical couplings from MD trajectories and represents proteins as communication networks [2]. In these networks, residues are nodes and their dynamical or interaction-based couplings describe how distant regions communicate.

Most communication models include only protein degrees of freedom. However, protein activation also reorganizes cavities, electrostatic interactions, hydrogen bonds, and hydration. Recent work has shown that water networks can change substantially between functional conformations and can be analyzed explicitly using graph-based approaches such as WatCon [1].

The central question of this project is:

Does hydration contain information about catalytic activation that is not already encoded in the protein conformational state?

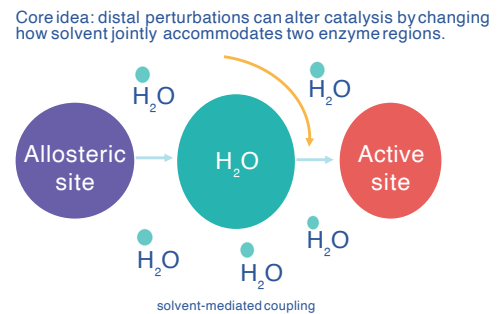


Figure 1: Central hypothesis: solvent may contribute directly to communication between regulatory and catalytic regions.

We will test this by separating the global progress of activation from the local organization of the catalytic site.

Model systems

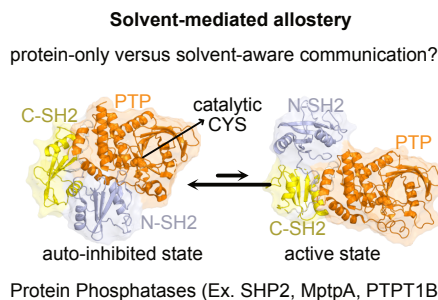


Figure 2: SHP2 switches between autoinhibited and active conformations.

We will study protein phosphatases for which inactive and active conformations are experimentally characterized.

SHP2 is particularly attractive because activation involves a large rearrangement from an autoinhibited state, in which the N-SH2 domain blocks the catalytic PTP domain, to an open active state [4]. This provides a clear global activation coordinate describing domain opening and a separate local catalytic coordinate describing organization of the active site.

MtpA provides a complementary system in which MD and NMR have already identified communication between a distal allosteric site and the catalytic acid loop [5]. PTP1B provides a methodological benchmark because hydration changes accompanying catalytic-loop opening and closure have already been characterized [1, 6].

Comparison of structurally distinct phosphatases will test whether solvent involvement is specific to one system or reflects a more general feature of phosphatase allostery.

What will we test?

The project separates three quantities.

Global activation coordinate. A coordinate q describes progression between inactive and active protein conformations. In SHP2, for example, this can describe opening of the N-SH2/PTP interface.

Catalytic coordinate. A second coordinate c describes local active-site organization using structural quantities known to distinguish catalytically competent from less competent conformations.

Hydration variables. A set of variables W describes the solvent environment, including occupancy of recurrent hydration sites, water-mediated contacts, hydrogen-bonding patterns, and local water-network organization.

We will generate ensembles at several points along the activation coordinate. Within each ensemble, the global coordinate will be weakly restrained while hydration and local catalytic degrees of freedom remain free to fluctuate.

The central quantitative test is:

At the same global activation state, does knowing the hydration state provide additional information about the catalytic state?

We will quantify this using **conditional mutual information**:

$$I(W; c \mid q).$$

If this quantity is close to zero, hydration does not provide additional information about catalytic organization once the protein activation state is known.

If it is reproducibly larger than zero, hydration contains information about catalytic organization that is not reducible to the global protein conformation.

Project workflow

1. Generate ensembles along activation

We will start from experimentally determined inactive and active structures.

Enhanced sampling will be used when necessary to identify representative intermediate conformations. Structures from inactive, intermediate, and active regions will then be used to initiate independent unbiased or weakly restrained MD simulations.

The restraint will keep the system close to a chosen global activation coordinate while allowing hydration, catalytic loops, side chains, and other local degrees of freedom to fluctuate.

2. Characterize hydration

For each ensemble, we will identify recurrent hydration sites, water occupancy, hydrogen-bonding patterns, and water-mediated protein contacts.

Because individual water molecules continuously exchange with bulk solvent, hydration will be represented by **persistent spatial sites** rather than by the identity of individual water molecules.

WatCon [1] will provide a starting framework for identifying and characterizing these sites.

Hydration will initially be described using continuous, physically interpretable variables rather than by forcing the data into a predefined number of discrete water states.

3. Test whether hydration informs catalytic organization

Within each restrained ensemble, we will calculate the catalytic coordinate c and the hydration variables W for the same trajectory frames.

We will then quantify whether hydration contains information about catalytic organization after the global activation coordinate has been accounted for using conditional mutual information.

The analysis will be repeated at several positions along the activation coordinate to determine where the solvent contribution is strongest.

Permutation controls will be used to establish the significance of the measured information by disrupting the relationship between hydration and catalytic organization while preserving their individual distributions.

4. Determine the molecular origin of the hydration–catalysis relationship

The statistical test establishes whether hydration contains additional information about catalytic organization. We will then use network analysis to determine how this coupling is organized molecularly.

For the same ensembles, we will construct:

- a conventional protein-only MDiGest communication network;
- a solvent-aware network containing both protein residues and persistent hydration sites.

MDiGest will be extended to incorporate protein–water and water–water dynamical couplings.

Communication between known regulatory and catalytic regions will be quantified using current-flow network analysis, which integrates multiple routes through the molecular network [3].

We will determine whether hydration features associated with different catalytic geometries participate in reproducible communication between regulatory and catalytic regions, and how this communication changes along activation.

Expected outcomes

The project will provide a quantitative test of whether hydration contributes information about catalytic activation beyond the global protein conformational state.

If conditional mutual information between hydration and catalytic organization is negligible once the activation coordinate is controlled, solvent rearrangement is likely to be largely a consequence of protein conformational change.

If hydration retains significant information about catalytic organization at fixed activation state, this will identify solvent as an additional degree of freedom in the activation mechanism.

If the same hydration features also participate in reproducible solvent-mediated communication between regulatory and catalytic regions, the network analysis will provide a molecular explanation for this coupling.

Comparison across different phosphatases will test the generality of the mechanism.

Perspective

This Master’s project is part of a broader research program developed in collaboration with Yale University and the Flatiron Institute in New York.

A subsequent PhD project will extend the analysis to genuine transition-path ensembles to determine how solvent reorganization evolves during activation and how it relates to transition kinetics and the free-energy landscape.

A central extension will be the implementation and application of **Joint Solvation Interaction (JSI)** calculations [7]. JSI provides a thermodynamic framework for quantifying whether distant protein regions are coupled through their shared solvent environment. This will connect the information-theoretic and network analyses developed during the M2 with a quantitative free-energy description of solvent-mediated allostery.

Student profile

The project is suitable for a Master 2 student with training in **physics, biophysics, computational biology**, or a related field.

Relevant coursework in thermodynamics, statistical mechanics, molecular biophysics, or molecular simulation is desirable.

Previous experience with Python is strongly preferred.

Experience with molecular dynamics, GROMACS, MDAnalysis, graph theory, or information theory is useful but not required. Training will be provided in molecular simulation, trajectory analysis, hydration analysis, scientific programming, and network approaches to protein dynamics.

Methods

Molecular dynamics; GROMACS; enhanced and restrained sampling; Python; MDAnalysis; WatCon; MDiGest; hydration-site analysis; conditional mutual information; graph theory; current-flow analysis.

References

- [1] A.-L. R. Brownless, T. Harrison-Rawn, and S. C. L. Kamerlin, “WatCon: A Python Tool for Analysis of Conserved Water Networks Across Protein Families,” *JACS Au* **5**, 6379–6395 (2025). doi:10.1021/jacsau.5c00447.
- [2] F. Maschietto, B. Allen, G. W. Kyro, and V. S. Batista, “MDiGest: A Python package for describing allostery from molecular dynamics simulations,” *J. Chem. Phys.* **158**, 215103 (2023). doi:10.1063/5.0140453.
- [3] W. M. Botello-Smith and Y. Luo, “Robust Determination of Protein Allosteric Signaling Pathways,” *J. Chem. Theory Comput.* **15**, 2116–2126 (2019). doi:10.1021/acs.jctc.8b01197.
- [4] R. A. P. Pádua, Y. Sun, I. Marko, W. Pitsawong, J. B. Stiller, R. Otten, and D. Kern, “Mechanism of activating mutations and allosteric drug inhibition of the phosphatase SHP2,” *Nat. Commun.* **9**, 4507 (2018). doi:10.1038/s41467-018-06814-w.
- [5] F. Maschietto, E. Zavala, B. Allen, J. P. Loria, and V. S. Batista, “MptpA kinetics enhanced by allosteric control of an active conformation,” *J. Mol. Biol.* **434**, 167540 (2022). doi:10.1016/j.jmb.2022.167540.
- [6] C. Wiesmann, K. J. Barr, J. Kung, et al., “Allosteric inhibition of protein tyrosine phosphatase 1B,” *Nat. Struct. Mol. Biol.* **11**, 730–737 (2004). doi:10.1038/nsmb803.
- [7] A. Hassanali and C. K. Egan, “The Joint Solvation Interaction,” *Entropy* **26**, 749 (2024). doi:10.3390/e26090749.