

Scuola Internazionale Superiore di Studi Avanzati

Physics area

Data-driven coordinate discovery and causal inference in molecular simulations

Ph.D. Course in Physics and Chemistry of Biological Systems

Supervisor:
Prof. Ali Hassanali

Candidate:
Debarshi Banerjee

Academic Year 2025/2026

This page intentionally left blank.

Acknowledgements

"Hope is a good thing, maybe the best of things, and no good thing ever dies."

— The Shawshank Redemption (1994)

Where do I even start with these acknowledgements? I don't think I could thank everyone who has played a crucial role in my life that has led to this point without this becoming its own separate thesis. I am also absolutely terrified I will forget someone – if that should happen, please chalk it up to poor memory and incompetence.

I am incredibly grateful to both Ali and Alessandro for being the most wonderful supervisors I could have ever asked for. Both academically and personally, they have always been tremendously supportive and I have learnt a lot from them. Perhaps the biggest thing they taught me is that doing great science is not incompatible with being a great human being. They also helped me overcome one of my biggest weaknesses: my tendency to get lost in the weeds when given scientific independence. Somehow, they helped me find my way without ever restricting that independence. I sometimes take them for granted a little, but whenever I hear horror stories from friends and colleagues about their supervisors, I remember how lucky I am to have worked with and been supervised by *both* Ali and Ale.

My undergraduate supervisors, Prof. Deb Shankar Ray and Dr. Rahul Sharma (RS), introduced me to research and instilled in me a deep love of statistical physics, especially its application to physico-chemical problems. During the disastrous period that was the Covid-19 pandemic, they kept encouraging me to remain in academia and continue with research. At the time, I was deeply disillusioned and unsure whether I had the resources or the luxury to pursue this "passion" project instead of getting a "real" job. RS, in particular, played an outsized role in keeping my desire to do research alive during the darkest nights, when I had given up hope. Thank you both. Even the preliminary steps towards this thesis would not have been possible without your guidance and support over all these years.

I would be remiss if I didn't thank Ivan Girotto, my master's co-supervisor and the person single-handedly responsible for my coming to ICTP in the first place. On a rainy July afternoon in Kolkata, at the height of the pandemic, I had very poor internet connectivity during my interview for the MHPC at ICTP. Continuing seemed untenable,

and I apologised for wasting the committee's time and left, giving up what seemed like my only option at the time. Ivan emailed me and offered me a second chance, which I gladly accepted. For the second interview, I stayed with a friend who had better WiFi. It would not be an exaggeration to say that Ivan's kindness, followed by his constant support when I came to ICTP and later applied for the PhD at SISSA, changed the course of my life.

Over the years, I have had the privilege of working with some brilliant colleagues, many of whom I have come to call friends. This illustrious list includes, but is not limited to, Colin, Vittorio, Gonza, Solana, Alex, Marta, Leon, Diganta, Natalia, Cesare, Giulia, Eddy, Martina, Khatereh, and Uriel. I have learnt many scientific lessons from them, and even more non-scientific ones. I owe particular thanks to Solana for teaching me how to make "less-wordy" slides, Colin for helping me run my first-ever MD simulations, Gonza for showing me how to plot orbitals in VMD, and Vittorio for patiently answering the million (possibly repetitive) questions I barrage him with. To everyone in the Hassanali and Laio groups whom I have had the pleasure of working or interacting with: thank you. I could scarcely have asked for better colleagues.

My friends in Trieste have made my time here more memorable and wonderful than I could ever have dreamed. There was the Brownies club with Mini and Nour; the Gossip Club with the aforementioned protagonists alongside Idris, Kyrell, Moha, Sohaib, and others; the *Dopolavoro* hangouts with Colin and Mehmet; the *Grande Muraglia* dinners with Natalia; the kebabs, gelatos, and very brief stint of boxing with Solana; and the mental health walks with Ludo. Trieste would not have been the same without you guys.

With some of these friends, I also share a particular "political" bond, rooted in the simple belief that all human beings, no matter where they are born, how they look, or whom they love, should be treated equally and not discriminated against. It is a shame that, in today's world, what should be a matter of basic human rights has instead become a political issue. To my fellow believers in fighting for a better tomorrow – Nour, Mini, Ludo, Moha, Solana, Idris, Bellou – you have, and will continue to have, my utmost support and respect.

There are so many other wonderful people I have met along the way and while I feel that trying to remember them all is an exercise in futility, I shall attempt to test my memory. My time at SISSA and ICTP has been enriched by friends with whom I have shared an apparently infinite supply of jokes. If my PhD were measured in the laughs I have shared with my friends, I could die a happy (albeit poor, and possibly still ungraduated) man. Contributing to these laughs, apart from the names mentioned above, are Fedo, Matte, Astrid, Aldo, Ronald, and Antonio. You have all made my life brighter by your very presence. Thank you all for the memories. I shall carry you all in my heart into the next chapter of my life, and I hope to find friends half as wonderful as you in the years to come (*if* they are to come).

Trieste has the unfortunate habit of bringing people into your life only to have them depart soon after. This is probably true of academia in general, but let me indulge

my nostalgic side here for a moment. Over the years, I have said goodbye to many wonderful people I met here, some of whom I knew only briefly and others for much longer. I think in particular of Kai, Mehmet, Colin, Debora, and Marwa, who have all left a lasting impression on my life. I didn't know half of you as well as I would have liked (apologies to Tolkien for the shameless plagiarism), but you all brightened my days, and your departures from Trieste – and, in some cases, from my life – left a vacancy in my heart.

At some point, karaoke became our favourite Friday-night activity. To the "regular" karaoke gang – Matte, Natalia, Solange, Maurizio, Ronald, Marcello – even though I guess most of us haven't gone in months: thank you for the music and dancing that filled my heart, and for (some of) your beautiful voices, which occasionally hurt my ears. I should also single out Solange here for valiantly attempting to teach me Italian (and having what can be generously described as *mild* success).

I also want to remember Sam, who is sadly no longer with us. He brought joy to everyone around him, and I often wish I had known him better and spent more time with him. Sam, we miss you, and you will forever remain in our hearts.

There are many people who make our everyday lives at ICTP and SISSA better, whether by helping us navigate bureaucracy or providing our regular dose of caffeine. At ICTP, I would particularly like to thank Francesca for helping me with all the bureaucracy when I first came to Trieste (remember the Covid green pass rules, and how vaccinations in one country weren't accepted in another?) and for always patiently answering every query I threw at her and calming my "bureaucracy paranoia". I would also like to thank Alessandro for the many ping-pong games we shared, and Sabina for helping me collect clothes for refugees and transport them to Linea d'Ombra. Thank you all.

Asja, my office neighbour, and Nicola are among the gems of the CMSP section at ICTP. I have shared fascinating conversations with them over the years, both academic and personal. Thank you both for all the advice, jokes, and discussions!

As I prepare to leave Trieste, I also think back to the friends I first met during the MHPC. Marco, Sera, Zainab – thank you for welcoming me so warmly and for helping me, both practically and emotionally, through my first months in Trieste and beyond. I admire you all enormously for the people you are.

Kuntal and Triasha, thank you for the support you gave me during Covid, and for being the best friends I could have asked for during our undergraduate days. Although I haven't seen either of you in person for a long time, I am grateful that we have stayed in touch. Debankur da was also a great source of support during those dark days. We often stayed on Discord until 6 AM (local time), discussing all sorts of existential subjects while playing chess, and he kept encouraging me to continue with academia and research. I should also mention Anwesha, who has put up with my constant mockery of her and her career choices, but who was always there for me during some very dark periods of my life.

Wrishik deserves a special mention for 12+ years of friendship. He is the friend

from earlier in this treatise (in case someone has bothered to read it in its entirety) who selflessly lent me his house (and WiFi) for the ICTP interview (during Covid, I might add). We have spent many nights watching TV series or football together, and many more talking about science, politics, and whatnot. He exemplifies the time-honoured Bengali tradition of adda. I am eternally grateful for his friendship, and to his parents for the hospitality, care, and affection they have shown me over the years.

And finally, my family. There are scarcely words that could express the depth of my gratitude or do justice to their sacrifices. I am incredibly grateful to my late grandparents – Dadu and Didun. And last, and definitely *most* importantly, the constant support and infinite sacrifices of my beloved Frodo and Mom are **the** reason I have been able to make it this far. Thank you all. I love you more than I can ever express.

বহু দিন ধরে বহু ক্রোশ দূরে
বহু ব্যয় করি বহু দেশ ঘুরে
দেখিতে গিয়েছি পর্বতমালা,
দেখিতে গিয়েছি সিন্ধু।
দেখা হয় নাই চক্ষু মেলিয়া
ঘর হতে শুধু দুই পা ফেলিয়া
একটি ধানের শিষের উপরে
একটি শিশিরবিন্দু।

This page intentionally left blank.

Abstract

Molecular simulations generate high-dimensional trajectories in which the mechanistically relevant features are buried across widely separated time scales ranging from collective structural motions that govern, for example, protein fluctuations to electronic transitions induced by light absorption in photochemistry. While such simulations capture the underlying atomic and electronic dynamics in fine detail, the cause-and-effect relationships that drive these types of processes remain poorly understood and emerge only through targeted analysis.

This thesis develops and applies data-driven, information-theoretic tools to this extraction problem, using distances between molecular configurations and the information they carry as the central objects of analysis. The work spans two complementary directions: the excited-state photophysics of biomolecules that display novel photochemistry, and the causal structure between different degrees of freedom that can be extracted from equilibrium molecular dynamics.

We first study the intrinsic, non-aromatic fluorescence of L-cysteine crystals, showing through a synergy of experiments and excited-state simulations, that a change in the solvent used during crystallization from light to heavy water, alters the crystal packing and hydrogen-bonding network. This in turn, constrains the thiol and amine motions responsible for non-radiative decay, and thereby enhances emission.

Motivated by the difficulty of identifying such decay-driving motions, we then introduce an unsupervised machine learning framework based on the Differentiable Information Imbalance that ranks nuclear degrees of freedom by their ability to predict electronic observables, and extracts compact, interpretable coordinates that explain the non-radiative decay processes across a wide range of photochemical systems.

In the second part of the thesis, we turn to causal inference in molecular systems. Using the aforementioned Information Imbalance method, we establish the conditions under which directed putative causal links emerge between collective variables at equilibrium, relating them to asymmetries in information content and to a separation of relaxation time scales. We further elucidate how one can quantify genuine causal links using interventional experiments.

Finally, we apply this causal inference framework to the collective dynamics of folded proteins, uncovering directed causal relationships between the leading principal components (PCs) of two proteins, Ubiquitin and NTL9. In both of these cases we observe that the leading PC has a causal link towards the lower-variance PCs, al-

though the nature of this link and its correlation with spatial delocalization of the PCs varies between the two systems considered. Together, these studies show that using our causal inference measure built on distances rather than on probability densities, one can infer interpretable, and highly non-trivial causal relationships in molecular systems.

This page intentionally left blank.

Contents

List of publications xiv

Synopsis xvi

1 Introduction **1**

2 Theoretical Background and Methods **11**

2.1 Classical Molecular Dynamics 11

2.2 Hybrid Quantum Mechanics/Molecular Mechanics 14

2.3 Non-Adiabatic Molecular Dynamics and Trajectory Surface Hopping . . 16

2.4 Coulomb Matrix Descriptors 19

2.5 Information Imbalance 19

2.6 Transfer Entropy 22

2.7 Imbalance Gain 23

3 Crystallization of L-Cysteine in Heavy Water Induces Intrinsic Fluorescence **25**

3.1 Introduction 25

3.2 Methods 27

3.2.1 Classical MD simulations 27

3.2.2 Ground-state QM/MM simulations 28

3.2.3 Non-adiabatic MD simulations 28

3.3 Results 29

3.3.1 Experimental Measurements Reveal Enhanced Intrinsic Fluorescence of Cysteine in D_2O vs H_2O 29

3.3.2 Supramolecular Packing of Cysteine Crystals are Different in D_2O vs H_2O 31

3.3.3 Molecular Mechanisms and Electronic Origins Behind Intrinsic Fluorescence of Cysteine in Heavy Water 32

3.3.4 Inhibiting C–N Distortions in Cysteine Luminescence 37

3.3.5 Cysteine in Heavy Water Exhibits Synergistic Singlet and Triplet Emission 38

3.4 Discussion 39

4 Machine learning the non-radiative decay modes in photochemical processes 41
4.1 Introduction . . . 42
4.2 Methods . . . 44
4.2.1 Chemical Systems . . . 45
4.2.2 Protocol for Analysis . . . 45
4.3 Results . . . 49
4.3.1 Furan . . . 49
4.3.2 Methaniminium Cation . . . 51
4.3.3 L-glutamine . . . 52
4.3.4 L-pyroglutamine-ammonium . . . 54
4.3.5 Molecular Motors . . . 55
4.3.6 Non-Radiative Decay Pathways Along Principal Components . . . 57
4.4 Discussion . . . 59

5 Towards a robust approach to infer causality from molecular dynamics simulations 63
5.1 Introduction . . . 63
5.2 Methods . . . 65
5.2.1 Langevin dynamics of the model systems . . . 65
5.2.2 Molecular dynamics simulations . . . 66
5.3 Results . . . 68
5.3.1 Unidirectional information transfer between collective variables in a molecular system . . . 68
5.3.2 Emergence of causal links in model systems at equilibrium . . . 72
5.3.3 Asymmetries and the response to external interventions . . . 75
5.4 Discussion . . . 77

6 Investigating causality between principal components in protein dynamics 79
6.1 Introduction . . . 79
6.2 Methods . . . 82
6.2.1 Trajectory analysis and convergence . . . 82
6.2.2 Imbalance Gain analysis of principal components . . . 83
6.3 Results . . . 83
6.3.1 Ubiquitin . . . 83
6.3.2 NTL9 . . . 86
6.4 Discussion . . . 88

7 Conclusions and perspectives 92

Bibliography 100

Appendices

A Supplementary Information: Photoluminescence of L-Cysteine crystals **139**
A.1 Experimental Methods . . . 139
A.1.1 L-Cysteine crystallization . . . 139
A.1.2 Confocal Microscopy . . . 139
A.1.3 Quantum Yield . . . 139
A.1.4 Fluorescence lifetime microscopy (FLIM) . . . 140
A.1.5 Mass Spectrometry . . . 140
A.1.6 Single-crystal X-Ray Diffraction . . . 140
A.1.7 Data collection details for the Cysteine crystal . . . 141
A.1.8 Cys-D Crystal Structure Details . . . 142
A.1.9 Cys-H Crystal Structure Details . . . 143
A.2 Validation of Computational Methods . . . 144
A.3 Details of excitation and non-radiative decay modes . . . 145
A.4 Supplementary Figures . . . 146

B Supplementary Information: Machine learning non-radiative decay modes **153**
B.1 Additional information on the studied systems . . . 154
B.1.1 Cysteine crystals . . . 162
B.2 Benchmarking and sensitivity analysis . . . 163
B.2.1 Convergence with the number of trajectories . . . 163
B.2.2 Analysis restricted to the vicinity of the conical intersection . . . 165
B.2.3 Sensitivity to the hyperparameters of the method . . . 166

C Supplementary Information: Inferring causality from molecular dynamics **169**
C.1 Transfer Entropy . . . 169
C.2 Asymmetric transfer entropy in the Markov system . . . 174
C.3 Unobserved states in the Markov system . . . 176
C.4 Supplementary figures . . . 177

D Supplementary Information: Causality between principal components **182**
D.1 Brüschweiler collectivity index . . . 182
D.2 NTL9 results for another folded trajectory fragment . . . 183
D.3 Supplementary Figures . . . 184

This page intentionally left blank.

List of publications

The results presented in this thesis are based on the following four publications and preprints:

1. **D. Banerjee**, S. Chibh, O. S. Tiwari, G. Díaz Mirón, M. Monti, H. R. Yakir, S. Pawar, D. Fixler, L. J. W. Shimon, E. Gazit, and A. Hassanali, *Crystallization of L-Cysteine in Heavy Water Induces Intrinsic Fluorescence*, Angew. Chem. Int. Ed. **64** (2025), e202505331. DOI: 10.1002/anie.202505331. [1]

2. **D. Banerjee**, G. Díaz Mirón, A. Rodriguez, and A. Hassanali, *Machine learning the non-radiative decay modes in photochemical processes*, arXiv:2605.08381 (2026). DOI: 10.48550/arXiv.2605.08381. [2]

3. V. Del Tatto[†], **D. Banerjee**[†], A. Hassanali, and A. Laio, *Towards a robust approach to infer causality from molecular dynamics simulations*, J. Chem. Phys. **162** (2025), 244105. DOI: 10.1063/5.0267926. [3]
   [†]These authors contributed equally to this work.

4. **D. Banerjee**, A. Hassanali, and A. Laio, *Investigating causality between principal components in protein dynamics*, J. Chem. Phys. **165** (2026), 094122. DOI: 10.1063/5.0349449. [4]

Other publications

During the course of this doctoral work, the author also contributed to the following publications and pre-prints, which do not form part of this thesis:

1. **D. Banerjee**, K. Azizi, C. K. Egan, E. D. Donkor, C. Malosso, S. Di Pino, G. Díaz Mirón, M. Stella, G. Sormani, G. N. Hozana, *et al.*, *Aqueous solution chemistry in silico and the role of data-driven approaches*, Chem. Phys. Rev. (2024). DOI: 10.1063/5.0207567. [5]

2. G. Díaz Mirón, C. R. Lien-Medrano, **D. Banerjee**, U. N. Morzan, M. A. Sentef, R. Gebauer, and A. Hassanali, *Exploring the Mechanisms behind Non-aromatic Fluorescence with the Density Functional Tight Binding Method*, J. Chem. Theory Comput. (2024). DOI: 10.1021/acs.jctc.4c00125. [6]

3. G. Díaz Mirón, C. R. Lien-Medrano, **D. Banerjee**, M. Monti, B. Aradi, M. A. Sentef, T. A. Niehaus, and A. Hassanali, *Non-adiabatic couplings in surface hopping with tight binding density functional theory: The case of molecular motors*, J. Chem. Theory Comput. (2024). DOI: 10.1021/acs.jctc.4c01263. [7]

4. S. Di Pino, **D. Banerjee**, M. Monti, G. Díaz Mirón, G. Cassone, and A. Hassanali, *Deconstructing the origins of interfacial catalysis: Why electric fields are inseparable from solvation*, J. Chem. Phys. (2025). DOI: 10.1063/5.0288327. [8]

5. M. Monti, L. Cimmino, G. Díaz Mirón, C. Diaferia, **D. Banerjee**, M. Stella, L. Vitagliano, A. Accardo, and A. Hassanali, *Investigating the Role of pH and Counterions in the Intrinsic Fluorescence of Solid-State L-Lysine*, J. Phys. Chem. B (2025). DOI: 10.1021/acs.jpcb.5c05756. [9]

6. V. Athiyarath, E. Naranjo, D. Dave, O. G. Ridwan, D. A. A. Rodriguez, Q. Zhu, M. Monti, G. Díaz Mirón, **D. Banerjee**, A. Hassanali, *et al.*, *Water-mediated reconfigurable topology and mechanics in porous peptide materials*, Matter (2026). DOI: 10.1016/j.matt.2026.102669. [10]

7. L. Huet, V. Del Tatto, **D. Banerjee**, A. Laio, and A. A. Hassanali, *Causality in Liquid Water as a Hallmark of Emergent Glassy Dynamics*, arXiv:2604.19491 (2026). DOI: 10.48550/arXiv.2604.19491. [11]

8. F. Tarasi, G. Díaz Mirón, **D. Banerjee**, D. Scherlis, *Hydrogen Detachment in Diprotic Aromatics: O–H versus N–H Photodissociation in m-Aminophenol*, J. Phys. Chem. Letters (2026). DOI: 10.1021/acs.jpclett.6c02112. [12]

Synopsis

This thesis develops and applies data-driven, information-theoretic tools for extracting interpretable, low-dimensional and causal descriptions of molecular systems from the high-dimensional data produced by their simulation. The work is organised around four studies, presented as self-contained chapters, which together span the excited-state photophysics of small bio-molecules and the causal analysis of equilibrium molecular dynamics simulations. A recurring methodological thread is the use of distances between configurations, and the information they carry, as a robust proxy for ranking, selecting, and weighting the degrees of freedom that matter.

In Chapter 3, we investigate the intrinsic, non-aromatic fluorescence of crystals of the amino acid L-cysteine, and its sensitivity to the use of light water (H_2O) vs heavy water (D_2O) for crystallisation . Combining spectroscopic experiments with excited-state non-adiabatic molecular dynamics, we show that crystals grown in heavy water adopt a distinct packing and hydrogen-bonding network relative to those grown in light water, that this constrains the thiol and amine motions responsible for non-radiative decay, and that the heavy-water polymorph consequently emits light more strongly. We also propose the presence of potential intersystem crossing and triplet emission in this system due to evidence of strong singlet-triplet coupling.

In Chapter 4, we address the general problem of identifying, from non-adiabatic dynamics simulations, the nuclear coordinates that drive a molecule towards a conical intersection and that are primarily implicated in non-radiative decay processes. We introduce an unsupervised machine learning framework based on the Differentiable Information Imbalance [13, 14], which ranks nuclear degrees of freedom by how well they predict electronic observables such as energy gaps and oscillator strengths, and extracts compact, physically interpretable reduced coordinates. Applied to five chemically diverse systems, the method recovers known mechanistic coordinates and reveals a systematic distinction between observables controlled by a few localised motions and those depending on more collective rearrangements.

In Chapter 5, we study the conditions under which directed causal links can be inferred between collective variables of a molecular system at equilibrium. Using the causal inference approach developed with the Information Imbalance method [13, 15],

we relate the emergence of unidirectional causality to asymmetries in information content and to a separation of relaxation time scales, and we clarify the difference between the necessary conditions provided by bivariate measures and the stronger statements requiring interventions or causal sufficiency. We demonstrate how these causal links can emerge in molecular dynamics simulations with the example of the amino acid tryptophan.

In Chapter 6, we apply this causal inference framework to the collective dynamics of the folded proteins, ubiquitin and NTL9. By treating principal components that describe the conformational fluctuations as candidate causal variables that form a complete orthogonal basis, we uncover directed relationships among the leading components, endowing the standard reduced description of protein motion with an internal, hierarchical, directed causal structure.

Finally, Chapter 7 summarises the main results and discusses perspectives, including how we can assess causal links in a more robust multivariate fashion, the design of computational interventions for causal inference in molecular dynamics, and the extension of the information-theoretic feature-selection framework to larger and condensed-phase photochemical systems.

1 Introduction

Visible light emission is widespread in nature, appearing in phenomena as varied as sunlight, aurorae, flames, electrical discharges, fluorescent minerals, and bioluminescent organisms [16, 17, 18]. Although the physical mechanisms responsible for these phenomena differ, emitted light provides a direct signature of how matter behaves upon interaction with light. Its colour, intensity, and temporal behaviour can therefore reveal information about the structure and environment of the emitting system at the molecular scale [19, 20].

This sensitivity has made light emission an important tool in spectroscopy, microscopy, biological imaging, and chemical sensing [19, 21, 22]. Fluorescent molecules can serve as probes of local polarity, electrostatic interactions, molecular conformations, and changes in their surrounding environment [23, 24]. Molecular light emission is consequently useful not only as a phenomenon to be explained, but also as an experimental observable through which otherwise inaccessible molecular processes can be studied [19].

Figure 1.1 shows the remarkable diversity of fluorescence colours exhibited by proteins under ultraviolet illumination. This seemingly simple observation conceals a wealth of underlying physics and chemistry. What determines whether a protein fluoresces at all, and what controls the wavelength and intensity of the emitted light? Answering these questions requires identifying the chemical groups that absorb and emit light, the excited electronic states involved, and the pathways through which these states evolve before fluorescence can occur.

Yet understanding the origin of fluorescence is only the first step. Once a molecule is promoted to an excited electronic state, it no longer exists in equilibrium with its environment. Instead, the excitation initiates a cascade of structural and dynamical responses, particularly in the surrounding solvent. In biomolecular systems, where water is the dominant environment, these coupled relaxation processes are collectively referred to as solvation dynamics. Their complexity immediately raises further questions: which molecular coordinates respond most strongly to electronic excitation, which play only a minor role, and how can these responses be quantified? More gener-

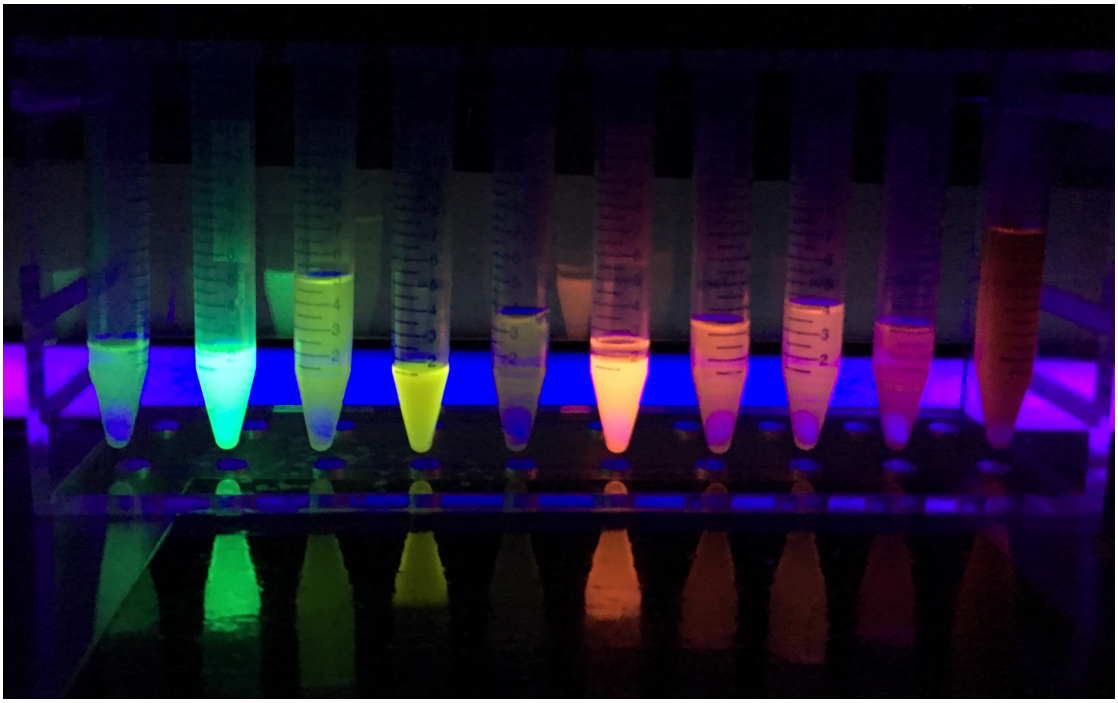


Figure 1.1: *An image showing different proteins that produce different fluorescent colours when exposed to ultraviolet light. Licensed under CC BY-SA 4.0. Source:* https://commons.wikimedia.org/wiki/File:Fluorescence_from_Fluorescent_Proteins.jpg

ally, how does one disentangle cause from effect in such highly collective, many-body dynamical processes?

One can identify which molecular coordinates are most strongly correlated with an observed response. Such correlations provide valuable clues towards a mechanistic interpretation, but they do not reveal directionality: does a particular degree of freedom initiate the dynamics, or is it itself a consequence of other molecular processes? Answering this question requires not only temporal information but also rigorous frameworks for establishing causality. Developing such an understanding, beginning with the photochemical origins of fluorescence and culminating in the identification of causal relationships in complex molecular dynamics, is the overarching goal of this thesis.

We begin by examining the molecular mechanisms underlying the fluorescence shown in Figure 1.1. The microscopic origin of molecular light emission remains an active area of research because an electronically excited molecule can evolve through several competing radiative and non-radiative pathways. Which pathway dominates depends on the electronic structure of the molecule, its nuclear motion, and its interactions with the surrounding environment. This complexity is commonly organised within the conceptual framework of a *Jablonski diagram*, as shown in Figure 1.2. The schematic shows the relevant potential-energy surfaces (PES): the ground state (S_0) in blue, the first excited singlet state (S_1) in red, and the first excited triplet state (T_1) in black. The horizontal axis represents a nuclear or reaction coordinate along which the surfaces are drawn, whereas the vertical axis represents their energy.

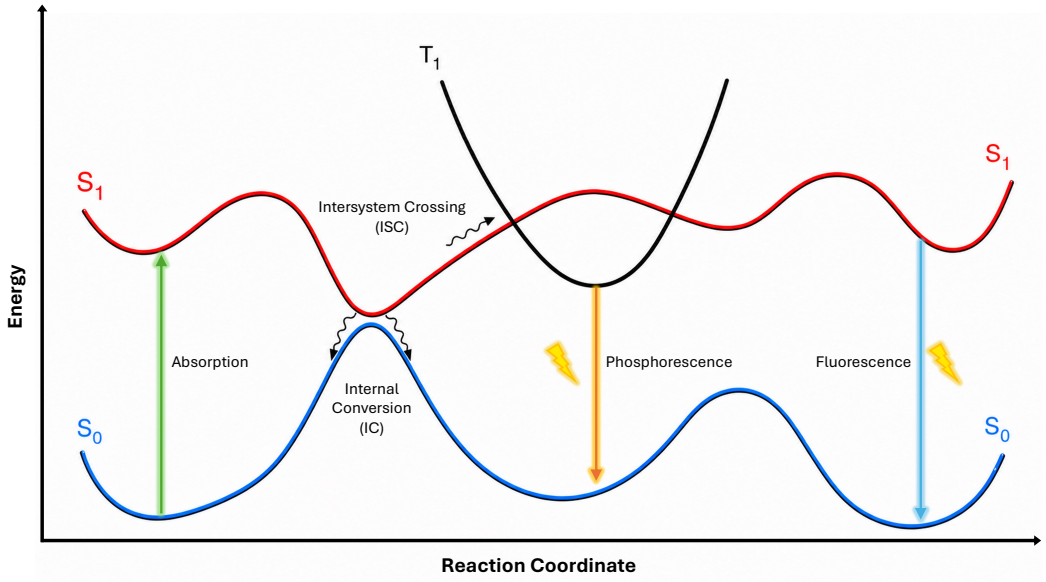


Figure 1.2: *Jablonski diagram showing common photophysical processes between the ground state S_0, the first excited singlet state S_1, and the first excited triplet state T_1. The processes shown here include absorption, as well as radiative processes like fluorescence and phosphorescence (represented with solid lines), as well as non-radiative processes like internal conversion and intersystem crossing (represented with wavy lines).*

After absorbing a photon, a molecule is promoted to an electronically excited singlet state – this process is called absorption and shown with a green line. Once the molecule undergoes absorption, its electronic distribution changes before the nuclei have time to respond. The system is therefore placed on an excited-state potential-energy surface at approximately its ground-state geometry (the Franck-Condon region). Nuclear motion then carries it away from the Franck-Condon region as energy is redistributed among vibrational degrees of freedom and exchanged with the surrounding environment. During this relaxation, the excited population can return to a lower electronic state (in this schematic, the ground state) through either a non-radiative process called internal conversion (IC) which is represented by a wavy line, or through the emission of light (fluorescence) which is represented by a blue line on the far right of the image.

One way for the system to undergo IC is when a *conical intersection* is encountered. A conical intersection is often represented as one point where two surfaces meet. For a polyatomic molecule, the degenerate geometries generally form a continuous seam. A nonlinear molecule with N atoms has $3N - 6$ internal degrees of freedom, so a generic seam between two same-spin states, in the absence of additional symmetry constraints, has dimension $3N - 8$. The motion from the Franck-Condon region to the conical intersection can involve different sets of coordinates. Different trajectories may also approach different regions of the same seam through distinct pathways [25, 26].

Another option is for the system to transition to a different electronic excited state with a different spin multiplicity (for example, the triplet state, T_1). This non-radiative process is called intersystem crossing (ISC) and it is also represented by a wavy line.

While singlet-triplet transitions are usually forbidden, under certain circumstances such as high spin-orbit coupling (SOC) they can occur more freely. From this triplet state the system could also relax back to the ground state via emission of light, which is called phosphorescence (represented by an orange line in the centre of the image). Typically phosphorescence is much weaker in intensity than fluorescence, and it also occurs over much longer timescales (milliseconds vs nanoseconds). In the closed-shell biomolecular systems that we primarily study in this thesis, fluorescence becomes observable when an emissive singlet state remains populated long enough for radiative decay to compete with the available non-radiative pathways [19].

A natural question when studying such photochemical phenomena concerns their molecular origins and the internal degrees of freedom responsible for the different photochemical pathways. Typically, in biomolecular fluorescence, there is an understanding that the emission of light originates primarily from aromatic or conjugated moieties which possess delocalised π electrons, relatively low-lying excited states, and transitions with appreciable intensities/oscillator strengths. Their often rigid structures can also restrict some of the large-amplitude distortions that promote rapid internal conversion. Intrinsic protein fluorescence is consequently attributed mainly to the three amino acids with aromatic side chains: tryptophan, tyrosine, and phenylalanine. In water at neutral pH, all three amino acids absorb at energies between 4.5 and 5.0 eV (approximately 280–250 nm), with tryptophan absorbing near the lower end of this range. Their emission peaks occur around 3.5 eV (348 nm) for tryptophan, 4.1 eV (303 nm) for tyrosine, and 4.4 eV (282 nm) for phenylalanine [27]. Tryptophan normally makes the largest contribution because it absorbs strongly at commonly used excitation wavelengths, emits at the longest wavelength, and has both a higher absorbance and a substantially higher quantum yield than the other two aromatic amino acids.

Peptide bonds and non-aromatic side chains also absorb light, although their strongest absorption generally lies deeper in the ultraviolet at energies $\sim$ 5–6 eV (200–250 nm). Low-lying states associated with lone pairs, including many $n\pi^*$ states, are usually weakly allowed and often have efficient routes to non-radiative decay. Biomolecules without aromatic residues or conjugated cofactors have therefore traditionally been expected to show little intrinsic emission in the near-ultraviolet or visible range.

Experiments over the past two decades have challenged this chemical intuition. Visible or near-visible emission has been reported from crystals and concentrated solutions of non-aromatic amino acids, short peptides, peptide dendrimers, oligosaccharides, amyloid-like fibrils, and proteins devoid of aromatic residues [28, 29, 30, 31, 32, 33, 34, 35]. This broad class of observations has been described using the umbrella term non-traditional intrinsic luminescence (NTIL) [33, 34]. Non-aromatic fluorescence (NAF) [30, 36, 37, 35, 38] is a special case of NTIL in which luminescence occurs despite the absence of conventional chromophores, such as conjugated moieties or aromatic groups. Luminescence can also arise primarily from aggregation or clustering; these phenomena are referred to as aggregation-induced emission (AIE) and cluster-

triggered emission (CTE), respectively [39, 40, 41, 42, 43, 44, 45]. AIE and CTE do not require the absence of a conventional chromophore like NAF does. They nevertheless describe related cases in which groups that are normally weakly absorbing or non-emissive acquire different optical properties within a particular molecular environment.

Several systems satisfy combinations of these criteria. Amyloid fibrils lacking aromatic residues undergo weak absorption at 3.3–3.6 eV (340–380 nm) and display structure-dependent blue or green emission around 2.3–2.8 eV (440–530 nm), while amino acid crystals can show substantial changes in their optical response when their packing changes [46, 30, 32, 35]. Aggregation and crystallisation are common routes to non-aromatic emission, but they are not strict requirements. Related absorption, and in some cases weak emission, has also been reported in highly charged proteins and concentrated solutions [47, 48].

Early work suggested that hydrogen bonding between peptide groups could support electronic delocalisation through the backbone and shift electronic transitions into an experimentally accessible range [28]. Our group has pioneered a variety of work in elucidating the mechanisms behind NAF.

First, studies of amyloid fibrils drew attention to proton motion within dense hydrogen-bond networks. The suppression of emission under strongly acidic or basic conditions, together with electronic-structure calculations, indicated that proton transfer and proton delocalisation can alter excitation energies in particular cross-β arrangements [30]. In subsequent work, short hydrogen bonds have also been connected with longer excited-state survival, where they tend to hinder distortions that otherwise facilitate internal conversion, thereby reducing access to efficient non-radiative decay pathways [37].

Other work done in the group emphasises the key role that carbonyl groups, which are ubiquitous in biomolecules, play in NAF. Calculations on amyloid-like peptide models showed that amide deplanarisation and carbonyl elongation can stabilise localised $n\pi^*$ states and change their oscillator strengths [36]. These distortions can convert electronic states that are otherwise too high in energy to emit in the visible range. Carbonyl elongation has also been identified as an important coordinate for non-radiative relaxation. Strong local interactions, such as short hydrogen bonds, can restrict this motion and keep the excited system away from configurations where the excited and ground states approach degeneracy [38]. This *carbonyl-lock* mechanism explains the non-aromatic fluorescence emission across a wide class of systems, but alternative mechanisms have not been excluded [35, 38]. The literature therefore supports a set of recurring molecular ingredients rather than a universal non-aromatic fluorophore.

The different proposals can be organised around two requirements for emission. The system must first possess an excited state in the relevant energy range with a non-zero probability of radiative decay. It must then remain on that state long enough to emit. Supramolecular organisation can influence both conditions. Intermolecular con-

tacts, hydrogen bonds, and local electric fields can stabilise states that are too high in energy or emit too weakly in an isolated molecule. The same environment can restrict structural distortions such as bond stretches, torsions, proton transfers, and deplanarisation motions that would otherwise promote rapid internal conversion. The origin of NAF therefore depends on both the electronic character of the accessible states and the nuclear pathways governing non-radiative decay away from these states.

Molecular crystals provide a useful setting in which to examine the relationship between structural features and their optical properties, and also to compare experimental results with theoretical simulations. Their packing can often be determined with atomic resolution, and different polymorphs allow the same molecular composition to be compared under different geometrical constraints. Changes in molecular conformation or intermolecular arrangement can produce substantial changes in solid-state emission, including changes in colour, lifetime, and the relative importance of singlet and triplet pathways [49, 50, 51, 52]. In **Chapter 3**, we investigate L-cysteine crystals grown experimentally from H_2O and D_2O. The latter appear to exhibit NAF, whereas the former do not. This study follows earlier work conducted by our group and further examines the molecular motions responsible for NAF. Using non-adiabatic molecular dynamics (NAMD) simulations, we explore how the network of intermolecular contacts can alter both the electronic states and the non-radiative decay pathways available after excitation. This work broadens our understanding of the hydrogen-bonding networks that stabilise non-aromatic fluorescence.

In many studies, including the work on cysteine in **Chapter 3**, the relevant coordinates that are responsible for non-radiative decay have been identified through visual inspection supported by chemical intuition. This approach can be effective when a small number of trajectories display an obvious bond rupture or torsion. It becomes less reliable when the motion is distributed over several coordinates, when multiple channels coexist, or when the electronically important displacement is small. It is also difficult to apply consistently across chemically different systems. Chemical intuition remains useful, but high-dimensional trajectory ensembles contain more coupled motions than can usually be assessed by inspection alone.

As illustrated in Figure 1.2, non-radiative processes like IC and ISC also depend on the nuclear positions. But it is not obvious how in a high-dimensional molecular system one can determine the appropriate degrees of freedom along which a system can decay non-radiatively. Put simply, the question becomes: how do we determine which coordinates are needed to draw a schematic like Figure 1.2, where the probability of IC becomes higher when the two potential-energy surfaces approach one another? Principal Component Analysis (PCA) has also been widely applied to identify collective nuclear motions associated with internal conversion or intersystem crossing pathways. However, because PCA ranks modes according to structural variance, it may emphasise large-amplitude or low-frequency motions that are not directly relevant to non-adiabatic transitions, while neglecting smaller-amplitude motions that lead to electronic transitions [53, 54, 55, 56, 57, 58].

The goal is to determine the coordinates that lead to the *conical-intersection seam*, the region where two adiabatic potential-energy surfaces become (approximately) degenerate and the Born-Oppenheimer separation of electronic and nuclear motion breaks down as the system can transition from one electronic state to another non-radiatively [59, 60, 61]. A small electronic energy gap is an indication that a trajectory is approaching a region where a transition or a *hop* from one electronic state to another can become likely. The probability of changing electronic state also depends on the coupling between the states, and the available kinetic energy. Oscillator strength describes another aspect of the dynamics. It depends on both the energy separation and the squared transition-dipole magnitude, and it indicates whether a configuration, if the system should emit light, is comparatively bright or dark. The coordinates associated with gap closure need not be the coordinates associated with a change in optical activity.

Chapter 4 approaches this task using the Information Imbalance (II) measure and its differentiable formulation (DII) [13, 14]. We propose an algorithm using the DII that identifies which internal degrees of freedom — bonds, angles, and dihedrals — are most informative about an electronic observable such as the energy gap or oscillator strength. This approach returns chemically interpretable coordinates that are primarily responsible for the non-radiative decay pathways that are explored in NAMD simulations, as well as those coordinates that help explain high- or low-intensity emission (or oscillator strengths).

Although the selected coordinates highlight plausible components of the underlying mechanism, their association with the observed dynamics does not establish whether they initiate the process or merely respond to it. Resolving this ambiguity requires time-resolved evidence to determine the sequence of events, together with controlled perturbations capable of demonstrating causality. This distinction marks the transition from the photochemical characterisation presented in the first part of this thesis to the causal analysis developed in the second.

Most analyses of molecular trajectories use observational data. The principle of *Granger causality*, which is used frequently in this thesis, defines a directed predictive relationship from X to Y: X Granger-causes Y when the past of X improves prediction of the future of Y beyond the information already contained in the past of Y [62]. Another popular causal inference method, based on conditional mutual information and inspired by the same principle as Granger causality, is Transfer Entropy. Transfer Entropy (TE) measures the reduction in uncertainty about the future of Y obtained by including the present of X in addition to the present value of Y [63]. These approaches detect forms of directed predictability, but their interpretation as causal effects requires further assumptions.

Ordinary time-lagged cross-correlation functions do not provide this directional information between pairs of variables. For equilibrium MD simulations, detailed balance ensures that pairwise cross-correlations are symmetric under exchange of variables. Cross-correlation can therefore reveal dynamical coupling, but it cannot iden-

tify a driver and a response. TE conditions the future target on both its own present value and the present value of a candidate source. It therefore probes a three-variable dependence that is absent from a two-point cross-correlation function. Granger causality, transfer entropy, and related information measures have been applied to molecular simulations of residue communication, protein allostery, conformational transitions, and reaction-coordinate selection [64, 65, 66, 67, 68, 69, 70, 71, 72, 73]. These studies report directional predictive relationships in molecular trajectories. They also illustrate the challenge of using transfer entropy as a causal measure, since this requires estimating high-dimensional probability distributions that can often be computationally intractable.

Directional interpretations are widespread in descriptions of aqueous molecular processes. In Marcus theory, solvent reorganisation governs the barrier crossing during electron transfer, with translational and rotational solvent motions contributing to the reorganisation energy. Likewise, electron and proton transfer are often described as being facilitated by solvent motions or electric-field fluctuations that modulate the reactive coordinate [74, 75, 76]. Solvent fluctuations have also been proposed to enable protein conformational changes [77, 78, 79] and to promote hydrophobic collapse through de-wetting transitions [80, 81, 82].

Across these examples, directional influence is inferred primarily from mechanistic models or a separation of characteristic timescales. Although such interpretations have been highly successful in explaining experimental observations, they do not provide a direct means of extracting directionality from molecular trajectories. Because, as noted earlier, conventional time-correlation functions are symmetric under equilibrium dynamics satisfying detailed balance, they cannot distinguish the dynamically leading variable from the responding one. This limitation raises a fundamental question: how can cause-and-effect relationships in molecular systems be identified quantitatively and rigorously from molecular trajectories?

Time-Dependent Fluorescence Stokes Shift (TDFSS) experiments provide a useful setting in which to formulate this distinction and connect the causal questions to the photochemical results from the first part of the thesis. These experiments monitor the time-resolved red shift of a fluorescent probe following optical excitation. The altered excited-state charge distribution induces relaxation of the surrounding solvent, protein, and probe, collectively stabilising the excited state [84, 85, 86, 87]. As illustrated in Figure 1.3, tryptophan is widely used as an intrinsic probe because excitation substantially changes the electrostatics of its aromatic indole group which acts as the chromophore. Tryptophan-based TDFSS has therefore been used to examine hydration and environmental relaxation over timescales extending from femtoseconds to nanoseconds [88, 89, 90, 91].

Assigning these relaxation timescales to specific molecular processes remains difficult because solvent and protein motions are strongly coupled and often overlap. Correlations can identify motions associated with the spectral shift but cannot determine their direction of influence. A mechanistic interpretation of TDFSS therefore requires

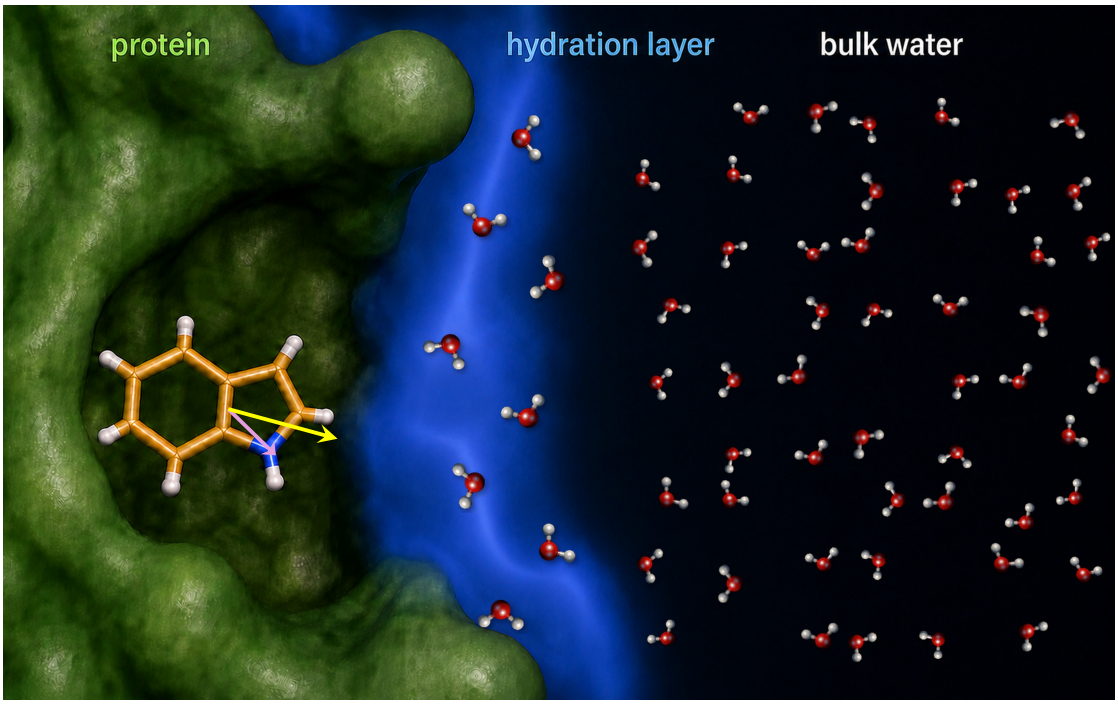


Figure 1.3: *A schematic of how photo-excitation in proteins usually works: the indole group of tryptophan, which is primarily responsible for the photochemistry, is shown (carbon atoms in orange, hydrogen in white, and nitrogen in blue) and the rest of the protein is represented in green. The hydration layer, that is the water molecules (oxygen atoms in red, and hydrogen in white) that make up the solvent region in the immediate vicinity of the protein is shown in blue, whereas the rest of the bulk region is black. The dipole of the indole ring prior to photo-excitation is shown with a pink arrow, whereas the increased dipole after photo-excitation (and its new direction) is shown with a yellow arrow. Image adapted from Ref. [83] and upscaled to higher resolution using AI tools.*

causal inference to distinguish whether protein motion drives solvent reorganisation, solvent dynamics drive the protein response, or whether both arise from a common underlying process.

Chapter 5 develops a causal measure suitable for equilibrium MD simulations. We use the Imbalance Gain (IG) approach developed by Del Tatto et al. [15], compare its results with those obtained using transfer entropy, and develop heuristics for inferring and justifying causal links in molecular systems. Because IG, which is based on the II measure, uses ranks of distances rather than explicitly estimating a joint probability density, it remains practical for multidimensional variables and avoids many of the computational challenges associated with transfer entropy. In this chapter, we investigate causal links in model Langevin systems and in a solvated tryptophan molecule. In the latter case we observe causal links between different structural variables (such as dihedrals), as well as energetic quantities (such as protein–solvent interaction energies).

There are a variety of open questions in the molecular sciences that can be better addressed through the lens of a "cause" – "effect" framing. This is not limited to the questions that naturally arise in terms of protein–solvent interactions and which microscopic interactions are primarily responsible for the relaxation of the system in the con-

text of the TDFSS experiments mentioned above. Earlier we investigated causal links between dihedral angles of a tryptophan molecule. This leads us to a natural question of looking at causal links between different structural collective variables (CVs). A popular class of CVs for biomolecular systems, especially proteins, is principal components (PCs) that describe conformational fluctuations very well. PCA is commonly used to project an atomistic protein trajectory onto an orthogonal basis set of directions ranked according to the amount of structural variance they explain [92, 93]. A natural question that arises is if one dihedral angle can have a causal link to another, can one PC also have a causal link to another? If so, what does this tell us about the nature of information transfer in proteins?

Chapter 6 attempts to address this question by applying the same causal framework to collective motion in folded proteins. While PCs are decorrelated at the start, they are not necessarily dynamically independent. Different PCs can retain memory on different timescales and may predict one another at finite lag times. The analysis of the PCs obtained from long-timescale MD of two proteins in this chapter tests whether the PCs display directional causal relationships that are not otherwise represented by just their corresponding explained variance. Pairwise IG estimates identify statistically meaningful causal couplings among selected PCs in both proteins.

In summary, the chapters of this thesis are linked by the distinction between identifying relevant coordinates and assigning direction to their relationships. In **Chapter 2** we introduce the key theoretical concepts and tools that have been used in this thesis. **Chapter 3** studies how crystal packing modifies the electronic states and decay pathways of L-cysteine, including the predicted participation of triplet states. **Chapter 4** develops an information-theoretic procedure for identifying nuclear coordinates associated with energy-gap closure and changes in oscillator strength during non-adiabatic dynamics. **Chapter 5** examines the conditions under which information-transfer asymmetries in equilibrium molecular systems can be given a causal interpretation. **Chapter 6** applies this analysis to the principal components of two folded-state proteins: ubiquitin and NTL9. **Chapter 7** discusses the broader implications and the directions that follow from these results. Across these studies, distance-based measures provide a common way to compare molecular representations, select informative coordinates, and examine directional predictive relationships without relying on explicit high-dimensional probability densities.

2

Theoretical Background and Methods

This chapter collects the theoretical and methodological background that is used repeatedly throughout the thesis. It is organised in two parts. The first concerns the simulation methods that generate the trajectories analysed in the following chapters: classical molecular dynamics (Section 2.1), hybrid quantum mechanics/molecular mechanics (Section 2.2), and non-adiabatic molecular dynamics within the trajectory surface hopping framework (Section 2.3). Then there is a brief overview of the Coulomb Matrix descriptor framework (Section 2.4) which we use in Chapter 4. The second concerns the information-theoretic and distance-based tools used to analyse those trajectories: the Information Imbalance and its differentiable formulation (Section 2.5), the Transfer Entropy (Section 2.6), and the Imbalance Gain (Section 2.7).

2.1 Classical Molecular Dynamics

Molecular Dynamics (MD) is a computational method for studying the physical motion of atoms as governed by their mutual interactions. In principle, these interactions are described most accurately by a quantum-mechanical treatment, in which the nuclei and electrons are represented by wave functions obtained by solving the corresponding Schrödinger equation [94]. The need to treat the electronic degrees of freedom explicitly depends on the phenomena being investigated. Processes involving changes in electronic structure, such as bond breaking and formation, charge transfer, or electronic excitation, generally require a quantum-mechanical description. By contrast, many structural, thermodynamic, and dynamical properties can be studied without resolving the electronic motion explicitly. For such applications, the effect of the electrons can be incorporated into a parametrized potential-energy function. This approximation enables simulations of much larger systems and longer timescales than are generally accessible with first-principles methods.

Molecular dynamics methods propagate the atomic nuclei through time by numerically integrating their equations of motion. *Classical MD* specifically indicates that the nuclei are treated as classical point particles obeying Newton's equations. It does not, by itself, specify how the forces acting on them are evaluated. Within the Born–Oppenheimer approximation [95], the electronic and nuclear motions are separated because electrons respond much more rapidly than nuclei. Consequently, the nuclei can be regarded as moving on a potential-energy surface determined by the electronic structure. In first-principles or *ab-initio* MD (AIMD), the electronic energy and the corresponding forces are calculated on the fly using an electronic-structure method. In empirical force-field MD, which is the approach considered here, the electronic degrees of freedom are not treated explicitly. Instead, the potential-energy surface is approximated by an analytical force field whose parameters are obtained by fitting to quantum-chemical calculations, experimental data, or a combination of both. The resulting forces are used to propagate the nuclear coordinates, and the acceleration $\mathbf{a}_i(t)$ of atom i at time t is given by

$$\mathbf{a}_i(t) = \ddot{\mathbf{r}}_i(t) = \frac{d\mathbf{v}_i(t)}{dt} = \frac{\mathbf{F}_i(t)}{m_i} \tag{2.1}$$

where $\mathbf{r}_i$, $\mathbf{v}_i$, and m_i are, respectively, the position, velocity, and mass of atom i, and $\mathbf{F}_i$ is the force acting on it. This force is obtained from the force field $U(\{\mathbf{r}_i\})$ as its negative gradient with respect to the atomic coordinates,

$$\mathbf{F}_i = -\frac{\partial U(\{\mathbf{r}_i\})}{\partial \mathbf{r}_i}, \quad i = 1, 2, 3, \ldots, N \tag{2.2}$$

where N is the number of atoms in the system.

Once the force field, and hence the force on each atom, is known, the trajectory of every atom is obtained by integrating Equation 2.1 numerically, using schemes such as the Verlet or Velocity-Verlet algorithms [96]. The general strategy is to discretize time into small steps Δt and to advance Equation 2.1 from t to $t + \Delta t$. The magnitude of the time lag Δt controls the numerical accuracy of the resulting positions and velocities. In the Verlet algorithm, the position of atom i at $t + \Delta t$ is computed from its positions at t and $t - \Delta t$ together with its acceleration at t:

$$\mathbf{r}_i(t + \Delta t) = 2\mathbf{r}_i(t) - \mathbf{r}_i(t - \Delta t) + \mathbf{a}_i(t)\Delta t^2 \tag{2.3}$$

Here the velocities are not propagated explicitly, instead, they can be recovered at a given time from a finite difference of the positions at $t + \Delta t$ and $t - \Delta t$:

$$\mathbf{v}_i(t) = \frac{\mathbf{r}_i(t + \Delta t) - \mathbf{r}_i(t - \Delta t)}{2\Delta t} \tag{2.4}$$

When explicit positions and velocities are required at every step, the velocity-Verlet algorithm is preferred. It advances the position and velocity of each atom from t to $t + \Delta t$ as:

$$\mathbf{r}_i(t + \Delta t) = \mathbf{r}_i(t) + \mathbf{v}_i(t)\Delta t + \frac{\mathbf{a}_i(t)\Delta t^2}{2} \tag{2.5}$$

$$\mathbf{v}_i(t + \Delta t) = \mathbf{v}_i(t) + \left[\frac{\mathbf{a}_i(t) + \mathbf{a}_i(t + \Delta t)}{2}\right]\Delta t \tag{2.6}$$

As formulated above, the algorithm samples atomic trajectories from the microcanonical (NVE) ensemble, in which the number of particles (N), the volume (V), and the total energy (E) are held fixed. However in many cases, one wants to sample different thermodynamic ensembles to make meaningful connections with experiment. This is usually carried out at constant temperature (T) and pressure (P), and additional algorithms are needed to control these quantities and thereby sample from the canonical (NVT) or isothermal–isobaric (NPT) ensembles. Temperature is commonly regulated by thermostats such as the Andersen [97], Nosé–Hoover [98], or stochastic velocity-rescaling (SVR) [99] algorithms, which exchange energy with the system by periodically modifying the atomic velocities. Pressure is controlled by barostats such as the Berendsen [100], Parrinello–Rahman [101], or stochastic cell-rescaling (SCR) [102] methods, which act by systematically rescaling the system volume.

In an empirical, non-reactive framework, the force field is conventionally partitioned into bonded and non-bonded contributions,

$$U_{\text{total}} = U_{\text{bonded}} + U_{\text{non-bonded}} \tag{2.7}$$

where $U_{\text{total}} = U(\{\mathbf{r}_i\})$. The bonded term describes intramolecular degrees of freedom such as bond lengths, angles, and dihedrals. In rigid-molecule models these are held at their equilibrium values throughout the simulation using constraint algorithms such as SHAKE [103], SETTLE [104], or RATTLE [105]. The non-bonded term accounts for the intermolecular electrostatic and weak van der Waals interactions, which are assumed to be pairwise additive. The electrostatic contribution is modeled by a Coulomb potential and the van der Waals contribution by the "6–12" Lennard-Jones potential, giving

$$U_{\text{total}} = U_{\text{bonded}} + \sum_{ij}\left[\frac{q_i q_j}{r_{ij}} + 4\epsilon_{ij}\left(\frac{\sigma_{ij}^{12}}{r_{ij}^{12}} - \frac{\sigma_{ij}^{6}}{r_{ij}^{6}}\right)\right] \tag{2.8}$$

In Equation 2.8, the first term in the sum is the Coulomb interaction between the partial charges q_i and q_j, and the second is the Lennard-Jones potential describing the van der Waals interaction. The two contributions differ markedly in range. The Lennard-Jones term decays rapidly as r_{ij}^{-6} and is therefore short-ranged, so it is evaluated efficiently by applying a spherical cutoff beyond which the interaction is neglected, often supplemented by a long-range dispersion correction. The Coulomb term, in contrast, decays only as r_{ij}^{-1} and therefore it is long-ranged and quite computationally expensive to estimate as a direct pairwise evaluation scales unfavourably as $O(N^2)$. It is therefore handled by Ewald summation [106], which splits the elec-

trostatic sum into a short-range part evaluated in real space and a smooth long-range part evaluated in reciprocal space. In practice this is done with the Particle Mesh Ewald (PME) method [107, 108], which interpolates the charges onto a grid and computes the reciprocal-space sum via the fast Fourier transform, reducing the cost to $O(N \log N)$.

2.2 Hybrid Quantum Mechanics/Molecular Mechanics

Whenever the process of interest involves a change in electronic structure while the surrounding environment can still be described classically, the two descriptions above must be combined. Hybrid quantum mechanics/molecular mechanics (QM/MM) [109, 110] provides this combination, and the electrostatic embedding variant [111] is the one used in this thesis.

While classical force fields describe large biomolecular systems efficiently, their fixed functional form cannot capture inherently electronic processes such as bond breaking and formation, charge transfer, or electronic excitation. A fully quantum-mechanical treatment would describe such processes accurately but is computationally prohibitive for systems containing tens of thousands of atoms. Hybrid quantum mechanics/molecular mechanics (QM/MM) methods, first introduced by Warshel and Levitt [109], reconcile these two requirements by partitioning the system into two regions: a small subsystem where the chemistry of interest takes place (such as, a reactive site, an active center, or a chromophore) is treated quantum-mechanically, while the surrounding environment (protein, solvent, membrane) is described by a classical force field [111]. This strategy has since become the standard approach for modeling reactivity and electronic processes in biomolecular systems [112, 113, 114].

In the additive QM/MM scheme, the total energy of the system is written as the sum of three contributions,

$$E_{\mathrm{QM/MM}} = E_{\mathrm{QM}} + E_{\mathrm{MM}} + E_{\mathrm{int}} \tag{2.9}$$

where E_{QM} is the energy of the QM subsystem obtained from the electronic-structure calculation, E_{MM} is the force-field energy of the MM subsystem, and E_{int} is the coupling energy describing the interaction between the two regions. This coupling term is conventionally decomposed into bonded, van der Waals, and electrostatic contributions,

$$E_{\mathrm{int}} = E_{\mathrm{bonded}} + E_{\mathrm{vdW}} + E_{\mathrm{elec}} \tag{2.10}$$

The bonded term arises only when covalent bonds cross the QM/MM boundary, in which case a capping scheme (such as link hydrogen atoms) is used to saturate the dangling valences. This term, alongside the van der Waals terms are generally evaluated at the MM level, since E_{vdW} can be described by the Lennard-Jones potential of the force field. The electrostatic term E_{elec} is the most important coupling contribution and the one that most distinguishes the various QM/MM flavours.

The manner in which the electrostatic QM–MM interaction is handled defines the embedding scheme [111, 113]. In *mechanical embedding*, the QM calculation is performed essentially in the gas phase: the MM charges exert no influence on the QM electron density, and the QM–MM electrostatic interaction is evaluated classically using a fixed set of point charges assigned to the QM atoms. This is simple but neglects the polarization of the QM region by its environment. In *electrostatic embedding*, which is the scheme adopted here, the MM point charges are incorporated directly into the electronic Hamiltonian of the QM region, so that the QM density polarizes self-consistently in response to the surrounding charge distribution. A third, more elaborate level, *polarized embedding*, additionally allows the MM region to respond to the QM charge density, however it is not considered here.

In electrostatic embedding, each MM atom I, carrying a point charge q_I at position $\mathbf{R}_I$, enters the QM electronic Hamiltonian as an additional external one-electron operator [112, 113],

$$\hat{H}_{\mathrm{QM}}^{\mathrm{eff}} = \hat{H}_{\mathrm{QM}}^{0} - \sum_{i \in \mathrm{QM}} \sum_{I \in \mathrm{MM}} \frac{q_I}{|\mathbf{r}_i - \mathbf{R}_I|} + \sum_{A \in \mathrm{QM}} \sum_{I \in \mathrm{MM}} \frac{Z_A\, q_I}{|\mathbf{R}_A - \mathbf{R}_I|} \tag{2.11}$$

where $\hat{H}_{\mathrm{QM}}^{0}$ is the Hamiltonian of the isolated QM subsystem, $\mathbf{r}_i$ are the coordinates of the QM electrons, and $\mathbf{R}_A$ and Z_A are the positions and nuclear charges of the QM atoms (atomic units are used throughout). Solving the resulting self-consistent-field problem yields a QM electron density $\rho(\mathbf{r})$ that is polarized by the MM environment, and the electrostatic coupling energy takes the form

$$E_{\mathrm{elec}} = -\sum_{I \in \mathrm{MM}} q_I \int \frac{\rho(\mathbf{r})}{|\mathbf{r} - \mathbf{R}_I|}\, d\mathbf{r} + \sum_{A \in \mathrm{QM}} \sum_{I \in \mathrm{MM}} \frac{Z_A\, q_I}{|\mathbf{R}_A - \mathbf{R}_I|} \tag{2.12}$$

The first term is the interaction of the QM electron density with the MM point charges, and the second is the interaction of the QM nuclei with those same charges. Because the MM charges are present throughout the SCF procedure, this scheme captures the electrostatic influence of the environment on the QM wavefunction, which is its main advantage over mechanical embedding.

As a general consideration, in electrostatic-embedding QM/MM methods, representing the MM charge distribution by bare point charges can lead to excessive polarization of the QM electron density, particularly near the QM/MM boundary. In severe cases, electron density may artificially leak, or "spill out," toward nearby MM charges. Depending on the QM/MM implementation, this artifact can be reduced by replacing the point charges with smeared charge distributions, such as Gaussian distributions, which ensure that the QM–MM Coulomb interaction remains finite at short distances as an electron approaches an MM atom.

2.3 Non-Adiabatic Molecular Dynamics and Trajectory Surface Hopping

Excited-state relaxation requires going beyond dynamics on a single potential energy surface. The trajectory surface hopping (TSH) technique, and in particular the SHARC approach [115, 116] (Surface Hopping including ARbitrary Couplings) — an extension of Tully's fewest switches surface hopping (FSSH) [117] — allows the nuclei to be propagated classically on quantum-chemical potential energy surfaces calculated on the fly [118, 119, 120, 121]. The basic theory is reviewed below.

Mixed quantum–classical dynamics

The Born–Oppenheimer approximation, which underlies most of our chemical intuition, breaks down whenever two or more electronic states approach each other in energy: the nuclear wave packet then splits over several coupled potential energy surfaces (PESs), and non-adiabatic transitions between electronic states become possible [118]. A full quantum-dynamical treatment of this branching wave packet is only feasible for small systems, since the cost scales exponentially with the number of nuclear degrees of freedom. TSH circumvents this bottleneck through a mixed quantum–classical partitioning: the electrons are treated quantum-mechanically, while the nuclei are propagated as an ensemble of independent classical trajectories, each evolving on a single electronic PES at any given time [118, 121]. The statistical character of the true wavepacket dynamics, including its branching into different electronic states and reaction channels, is recovered by allowing the trajectories to *hop* stochastically between surfaces and by averaging observables over the whole ensemble [117, 118].

Following Tully [117], the nuclei are assumed to move along classical trajectories $\mathbf{R}(t)$, and the time-dependent electronic wave function can be expanded in a basis of electronic states, where $\{\psi_K\}$ represents the electronic wave function for state K, that depend parametrically on the nuclear coordinates,

$$\Psi(\mathbf{r}, t; \mathbf{R}) = \sum_K c_K(t)\, \psi_K(\mathbf{r}; \mathbf{R}(t)) \tag{2.13}$$

Here, $\mathbf{r}$ represents the electronic coordinates and $c_K(t)$ are the complex expansion coefficients. Inserting this ansatz into the time-dependent electronic Schrödinger equation yields the equation of motion for the coefficients [118, 121],

$$\frac{\mathrm{d}c_J}{\mathrm{d}t} = -\sum_K c_K \left(\frac{\mathrm{i}}{\hbar} H_{JK} + \sigma_{JK}^{\mathrm{NAC}} \right), \tag{2.14}$$

with the electronic Hamiltonian matrix elements $H_{JK} = \langle \psi_J | \hat{H}_e | \psi_K \rangle$ and the scalar

(time-derivative) non-adiabatic coupling

$$\sigma_{JK}^{\mathrm{NAC}} = \langle \psi_J | \frac{\partial}{\partial t} | \psi_K \rangle = \mathbf{d}_{JK} \cdot \mathbf{v}, \qquad \mathbf{d}_{JK} = \langle \psi_J | \nabla | \psi_K \rangle , \tag{2.15}$$

where $\mathbf{d}_{JK}$ is the non-adiabatic coupling (NAC) vector and $\mathbf{v}$ the classical nuclear velocity. Equation (2.14) is the semiclassical time-dependent Schrödinger equation, and the coefficients c_J define the density matrix ρ. The diagonal elements of the density matrix, $\rho_{KK} = |c_K|^2$, give the electronic populations, and the off-diagonal elements $\rho_{JK} = c_J c_K^*$ the electronic coherences.

While the electronic coefficients evolve according to Eq. (2.14), the nuclei are propagated by Newton's equations on the single Born-Oppenheimer PES of the currently occupied (*active*) state L,

$$M_\alpha \frac{\mathrm{d}^2 \mathbf{R}_\alpha}{\mathrm{d}t^2} = -\nabla_\alpha H_{LL}, \tag{2.16}$$

where M_α and $\mathbf{R}_\alpha$ are the mass and position of nucleus α, and in an adiabatic basis H_{LL} is simply the adiabatic energy E_L of the active state [121]. The electronic and nuclear problems are thus coupled: the electronic Hamiltonian depends parametrically on $\mathbf{R}(t)$, while the nuclear force is dictated by the active electronic state.

The Fewest-Switches Surface Hopping (FSSH) criterion

The defining feature of a TSH method is the prescription used to decide when a trajectory switches between different potential-energy surfaces. In Tully's fewest-switches algorithm [117], which is the most common formalism of TSH and also the variant used in this work, the hopping probabilities are chosen so that the fraction of trajectories running on each state reproduces the quantum populations $|c_K|^2$ while the *number* of hops is kept to a minimum. This leads to the instantaneous probability of hopping from the active state L to another state J during a time step Δt [118, 121],

$$P_{L\to J} = \max\left[0,\ \frac{2\,\Delta t}{\rho_{LL}} \left(\hbar^{-1} \operatorname{Im}\left(H_{LJ}\,\rho_{JL}\right) - \operatorname{Re}\left(\sigma_{LJ}^{\mathrm{NAC}}\,\rho_{JL}\right)\right)\right], \tag{2.17}$$

which, in an adiabatic basis where the Hamiltonian is diagonal, reduces to $P_{L\to J} = \max[\,0,\ -2\Delta t\ \rho_{LL}^{-1} \operatorname{Re}(\sigma_{LJ}^{\mathrm{NAC}} \rho_{JL})]$. A hop is thus invoked only when the population of the active state decreases, so that the fewest-switches probability must be evaluated at every time step along the trajectory. The stochastic decision is then made by drawing a uniform random number $r_t \in [0, 1]$ and switching to the state J that satisfies [118, 121]

$$\sum_{K=1}^{J-1} P_{L\to K} < r_t \le \sum_{K=1}^{J} P_{L\to K}. \tag{2.18}$$

Two further ingredients are required to make the algorithm physically consistent. First, because the new active state after a hop generally lies at a different potential energy, the nuclear velocities must be rescaled to conserve the total energy across the hop.

The rescaling is usually applied along the direction of the NAC vector $\mathbf{d}_{LJ}$ (motivated by the Pechukas force) or, when NAC vectors are unavailable, along the velocity vector [118, 121]. If the kinetic energy available in the relevant direction is insufficient to compensate the energy gap, the hop is classically forbidden and rejected, and this is referred to as a *frustrated hop* [118]. Second, because each trajectory follows the gradient of a single active state whereas the coefficients in Eq. (2.14) keep evolving coherently, the independent-trajectory approximation renders TSH overcoherent: the electronic coherences ρ_{JK} do not decay after the wave packet has left the coupling region, spoiling the internal consistency between the average populations and the fraction of trajectories in each state [118, 121]. An *a posteriori* decoherence correction is therefore applied at every step, most commonly the energy-based decay-of-mixing scheme of Granucci and Persico [122], in which the coefficients of the inactive states are damped as $c_K^{\text{new}} = c_K\, e^{-\Delta t/\tau_{KL}}$ with a decoherence time $\tau_{KL} = \hbar\,|E_K - E_L|^{-1}(C + \varepsilon/E_{\text{kin}})$, and the active-state L has its coefficient rescaled as $c_L^{\text{new}} = \frac{c_L}{|c_L|}\left[\,1 - \sum_K |c_K^{\text{new}}|^2\,\right]^{1/2}$.

Within TSH, the evolution and branching of the nuclear wave-packet density are approximated statistically rather than represented explicitly. An ensemble of independent mixed quantum–classical trajectories is initialized by sampling a distribution of nuclear positions and momenta, while stochastic surface hops distribute the trajectories among different electronic states and reaction channels during the dynamics. The resulting trajectory fractions and ensemble averages provide approximate electronic-state populations, branching ratios, and other observables accessible within the TSH framework [117, 118, 121].

The SHARC extension: surface hopping with arbitrary couplings

Standard FSSH as described above is formulated in the adiabatic basis and accounts only for the non-adiabatic couplings σ_{JK}^{NAC} that connect electronic states of the *same* spin multiplicity, i.e. it describes internal conversion but not intersystem crossing. The SHARC approach [115, 116] (Surface Hopping including ARbitrary Couplings) generalises the method so that *any* type of state-to-state coupling, most importantly the spin–orbit couplings (SOCs) that drive intersystem crossing, can be treated on the same footing. This is essential for the systems studied here, where singlet and triplet manifolds are coupled by SOCs.

The central idea of SHARC is to add these couplings directly to the electronic Hamiltonian and then perform the dynamics in the basis that diagonalises it. The total electronic Hamiltonian is written as Ref. [116]

$$\hat{H}^{\text{total}} = \hat{H}^{\text{MCH}} + \hat{H}^{\text{additional}}, \tag{2.19}$$

where $\hat{H}^{\text{MCH}}$ is the molecular Coulomb Hamiltonian (the commonly named "MCH" representation), and $\hat{H}^{\text{additional}}$ collects the extra couplings such as the SOC operator, or electric field-dipole moment couplings. In the MCH representation these ad-

ditional couplings appear as off-diagonal elements of the Hamiltonian matrix $\mathbf{H}^{\mathrm{MCH}}$. The SHARC dynamics is instead run on the eigenstates of the full Hamiltonian, obtained by a unitary transformation $\mathbf{U}$ that diagonalises the MCH matrix [116],

$$\mathbf{H}^{\mathrm{diag}} = \mathbf{U}^{\dagger}\,\mathbf{H}^{\mathrm{MCH}}\,\mathbf{U}. \tag{2.20}$$

This diagonalization gives us the approximate eigenenergies and eigenstates of the total Hamiltonian. The set of eigenstates of this total Hamiltonian, $\hat{H}^{\mathrm{total}}$, are referred to as the "diagonal" representation. In practice, both of these representations are of prime relevance in the SHARC approach: the MCH representation, in which the electronic-structure quantities are computed, and the diagonal representation, in which the nuclear propagation and the fewest-switches surface hopping are carried out.

Apart from this transformation, the SHARC dynamics retains the full TSH machinery described above (performed in the diagonal representation): the semiclassical propagation of the coefficients (Eq. (2.14)), the fewest-switches hopping criterion (Eq. (2.17)), kinetic-energy rescaling after a hop, rejection of frustrated hops, decoherence correction, and ensemble averaging. This unified treatment of internal conversion and intersystem crossing is what makes the approach applicable to systems in which singlet and triplet manifolds are coupled.

2.4 Coulomb Matrix Descriptors

The Coulomb matrix [123] (CM) is a descriptor that is computed between all atoms for the trajectories in a given electronic state. The Coulomb matrix is defined as:

$$\mathrm{CM}_{ij} = \begin{cases} 0.5 Z_i^{2.4} & \text{for } i = j \\ \frac{Z_i Z_j}{r_{ij}} & \text{for } i \neq j \end{cases} \tag{2.21}$$

The diagonal elements can be seen as the interaction of an atom with itself and are essentially a polynomial fit of the atomic energies to the nuclear charge Z_i. The off-diagonal elements represent the Coulomb repulsion between nuclei i and j.

2.5 Information Imbalance

The *information imbalance* (II) is a statistical test, introduced by Glielmo *et al.*, [13] which quantifies the relationship between different distance measures defined over the same dataset. It is used in two distinct ways in this thesis. In the first, one asks whether a set of structural coordinates can explain the variation of an electronic observable such as an energy gap or an oscillator strength: since vanishing energy gaps correspond to the approach to a conical intersection, structural variables that encode the energy-gap landscape identify the nuclear motions responsible for reaching the crossing of two potential energy surfaces, and the same construction applied to oscillator strengths

identifies the coordinates that govern the emergence of bright and dark states. In the second, the II is used as the building block of a measure of directional information transfer between collective variables, the Imbalance Gain of Section 2.7. Both uses rest on the same formalism, which is reviewed here.

Information imbalance provides a quantitative measure for comparing the information content of two distance metrics, d_A and d_B, defined over the same set of data points. Typically d_A is built on a representation whose informativeness is being tested — a structural representation such as internal coordinates or Coulomb matrix elements (Section 2.4), or the present state of a collective variable — while d_B is built on the target quantity, such as an electronic observable or the future state of another variable.

As an example, distances in the space of an electronic observable can be defined between two configurations i and j as $d_B^{ij} = \left\|\Delta E^i - \Delta E^j\right\|$, where ΔE^i and ΔE^j denote the electronic energy gaps associated with configurations i and j, respectively, and $\|\cdot\|$ denotes the Euclidean norm. Distances d_A^{ij} are defined analogously using a chosen set of structural coordinates, or Coulomb Matrix elements.

Given a distance metric d_A, one can define r_A^{ij} as the *rank* of point j among all points in the dataset when ordered by their distance from point i according to d_A. Similarly, r_B^{ij} denotes the rank of point j with respect to point i according to the distance metric d_B. Intuitively, the representation A is considered informative about representation B if configurations that are close neighbors according to d_A also remain close neighbors according to d_B. In other words, d_A is informative with respect to d_B when the neighborhood structure defined by d_A preserves the neighborhood structure defined by d_B.

The II from d_A to d_B is defined as:

$$\Delta(d_A \to d_B) := \frac{2}{N}\langle r_B \mid r_A \leq k\rangle = \frac{2}{N^2\, k} \sum_{i,j:\, r_A^{ij} \leq k} r_B^{ij}\,, \tag{2.22}$$

where N is the total number of points in the data set and k is the number of nearest neighbor points that are considered. Eq. 2.22 gives us the II, represented as $\Delta(d_A \to d_B)$, as a number between 0 and 1. $\Delta(d_A \to d_B) = 0$ occurs when all points that are nearest neighbors according to d_A remain nearest neighbors in d_B, whereas $\Delta(d_A \to d_B) = 1$ occurs when the nearest neighbors according to d_A are randomly distributed in d_B. The former case is when d_A is maximally informative with respect to d_B, and the latter is when it is minimally informative. While the Information Imbalance was originally introduced with $k = 1$ [13], it was observed in Ref. [15] that using larger values of k — up to a small fraction of the dataset size (e.g., $k/N = 5\%$) — yields more stable and less noisy estimates, particularly when the distance spaces are constructed from high-dimensional variables. The value of k used in each application is stated in the corresponding chapter.

A recent development extends the original II framework by introducing a differentiable formulation [14]. In this approach, the individual features that define a distance metric are multiplied by learnable weights, which are optimized through gradient-

based methods. By allowing the weights to adapt during optimization, this formulation naturally accommodates heterogeneous features that may differ substantially in their physical units or numerical scales. Consequently, it avoids the need for manual pre-processing or ad hoc rescaling of the input variables.

This approach, termed the *Differentiable Information Imbalance* (DII), enables the automatic identification of the most informative subset of structural descriptors for a given dataset. By learning the optimal weighting of features that best preserves the neighborhood structure of a target representation, DII effectively determines which coordinates contribute most strongly to defining the relevant distance metric. The DII can be formally written as

$$\mathrm{DII}(d_A(\boldsymbol{w}) \to d_B) := \frac{2}{N^2} \sum_{i,j=1}^{N} c_{ij}(\lambda, d_A(\boldsymbol{w})) r_B^{ij} \tag{2.23}$$

where,

$$c_{ij}(\lambda, d_A(\boldsymbol{w})) := \frac{e^{-d_A^{ij}(\boldsymbol{w})/\lambda}}{\sum_{m \neq i} e^{-d_A^{im}(\boldsymbol{w})/\lambda}} \tag{2.24}$$

In this expression, $\boldsymbol{w}$ denotes a set of variational parameters that are optimized via gradient descent to assign different relative weights to the features defining the distance d_A. The hyperparameter λ is selected based on characteristic length scales of the dataset, typically determined from the average and minimum nearest-neighbor distances. Through this optimization procedure, the method simultaneously identifies which features in d_A contain the most predictive information with respect to d_B and determines the optimal combination in which these features should contribute to the distance metric. In the limiting case $\lambda \to 0$, the DII formulation (Eq. 2.23) reduces to a continuous analogue of the standard Information Imbalance defined in Eq. 2.22.

In realistic applications where the II has been employed across diverse problems, there is no universal threshold for what constitutes an informative value; however, empirical benchmarks can be established. In practice, II values in the range of 0.1–0.4 are typically considered highly informative and represent near-optimal performance, particularly in molecular systems where noise and imperfect correlations between descriptor spaces (e.g., structural features) and target observables (e.g., energy gaps or oscillator strengths) are unavoidable. Values between 0.4–0.7 still indicate a moderate degree of informativeness, while values in the range of 0.7–0.9 correspond to weak correlations. Values approaching 1.0, by contrast, indicate an absence of meaningful information content [124, 125, 126, 14, 8].

When the DII is used for feature selection, the space of possible feature combinations must be explored without incurring the combinatorial growth associated with testing all possible subsets. Throughout this thesis this is done with the forward greedy selection strategy introduced in Ref. [13], combined with the DII implementation provided in the `DADApy` Python library [127]. The sequence of dimensionality-reduction steps in which this selection is embedded is described, where relevant, in the corre-

sponding chapter.

2.6 Transfer Entropy

The Transfer Entropy, in its bivariate formulation, quantifies how the future state of a time-dependent random variable (RV) Y_t can be better predicted given the knowledge of the current states of both Y_t and a second variable X_t, rather than using only the present state of Y_t [63, 128]. We use the following definition of Transfer Entropy in direction $X \to Y$:

$$\mathrm{TE}_{X\to Y}(\tau) := H(Y_\tau \mid Y_0) - H(Y_\tau \mid X_0,\, Y_0)\,, \tag{2.25}$$

where τ is a discrete and positive time lag and $H(\cdot \mid \cdot)$ denotes a conditional entropy. For discrete RVs, both terms can be written as conditional Shannon entropies,

$$H(Y_\tau \mid Y_0) = -\sum_{y_\tau,\, y_0} P(y_\tau,\, y_0) \log P(y_\tau \mid y_0)\,, \tag{2.26a}$$

$$H(Y_\tau \mid X_0, Y_0) = -\sum_{y_\tau,\, x_0, y_0} P(y_\tau,\, x_0, y_0) \log P(y_\tau \mid x_0, y_0), \tag{2.26b}$$

where $P(\cdot,\, \cdot)$ and $P(\cdot \mid \cdot)$ denote the joint and conditional probability distributions of their arguments, respectively. The sums run over all possible realizations of the RVs Y_τ, X_0 and Y_0, denoted by lowercase letters. For continuous-valued RVs, Eqs. (2.26) can be generalized to *differential* entropies, which are obtained by replacing sums with integrals and discrete distributions with probability densities. Throughout this thesis, Shannon and differential entropies are computed over *stationary* distributions generated by *equilibrium* processes. In this setting, the choice of the initial time $t = 0$ is arbitrary.

Intuitively, $\mathrm{TE}_{X\to Y}(\tau)$ quantifies how much knowing X at time 0 helps to reduce the uncertainty about Y at time τ, beyond what is already explained by the past of Y itself. Technically, the inequality $\mathrm{TE}_{X\to Y}(\tau) > 0$ is equivalent to the conditional dependence relationship $X_0 \not\perp\!\!\!\perp Y_\tau \mid Y_0$ (read: Y_τ depends on X_0 given Y_0), which allows stating the existence of a causal link (direct or indirect) from X_0 to Y_τ, if X and Y are not affected by any common driver Z (details are provided in Appendix C.1). If the same measure in the opposite direction, $\mathrm{TE}_{Y\to X}$, is equal to zero after lag τ, we say that the transfer of information from X to Y is (effectively) unidirectional after that time lag.

Computing Eqs. (2.26) in practice, given the time-series realizations of X and Y, can be challenging. When X and Y are one-dimensional variables, the joint and conditional probabilities in Eqs. (2.26) can be estimated by trivial maximum likelihood estimators [129], which involve counting the occurrence of all joint events (in the discrete case) or constructing histograms (in the continuous case). However, when X and Y are multi-dimensional, classical estimation techniques often break down due to the curse of dimensionality, which severely hinders the accurate estimation of probability distributions — particularly in the continuous setting. This issue frequently

arises in molecular systems, where the CVs describing the mesoscopic state are often intrinsically high-dimensional (for example, all the internal dihedrals of a protein molecule). In such scenarios, nearest-neighbor-based entropy estimators [130, 131, 132, 133] represent a computationally efficient alternative. Where Transfer Entropy is computed in this thesis, the nearest-neighbor estimator proposed in Ref. [133] is used for continuous-valued variables, and a simple maximum likelihood estimator [129] for discrete ones. Technical details on these estimators are provided in Appendix C.1.

2.7 Imbalance Gain

As an alternative strategy to mitigate the curse of dimensionality, Ref. [15] introduced the Imbalance Gain (IG), a distance-based measure inspired by Transfer Entropy that does not require the explicit estimation of probability densities. It probes conditional independence by a suitable rank statistics, and builds directly on the Information Imbalance of Section 2.5.

In the same spirit of Transfer Entropy and Granger Causality, in Ref. [15] it was proposed to use Eq. (2.22) to verify whether the prediction of d_{Y_τ} can be improved by using a "mixed" distance space including both variables X_0 and Y_0, rather than Y_0 alone. Specifically, the condition $\mathrm{TE}_{X\to Y} > 0$ was translated into the following inequality:

$$\min_{\alpha} \Delta(d_{\alpha X_0, Y_0} \to d_{Y_\tau}) < \Delta(d_{Y_0} \to d_{Y_\tau})\,. \tag{2.27}$$

Here, $d_{(\cdot)}$ denotes a Euclidean distance built over the subscript variables. For example, notation $d_{\alpha X_0, Y_0}$, where α is a multiplicative scalar factor, identifies the following distance between points i and j:

$$d^{ij}_{\alpha X_0, Y_0} = \left(\alpha^2 \|X_0^i - X_0^j\|^2 + \|Y_0^i - Y_0^j\|^2\right)^{1/2}\,, \tag{2.28}$$

where $\|\cdot\|$ is the Euclidean (L^2) norm. The parameter α controls the relative importance of X_0 and Y_0 in the optimal distance space that one can construct to predict Y_τ. Indices i and j denote independent initial conditions ($t = 0$), which can be sampled from a single trajectory if the process is stationary.

Eq. (2.27) can be written equivalently as $\mathrm{IG}_{X\to Y}(\tau) > 0$, by defining the Imbalance Gain in direction $X \to Y$ as [15]

$$\mathrm{IG}_{X\to Y}(\tau) := \frac{\Delta(d_{Y_0} \to d_{Y_\tau}) - \min_\alpha \Delta(d_{\alpha X_0, Y_0} \to d_{Y_\tau})}{\Delta(d_{Y_0} \to d_{Y_\tau})}\,. \tag{2.29}$$

If $\mathrm{IG} = 0$, adding the information about X_0 does not improve the prediction of Y_τ, since the best possible predictor is Y_0 alone. A positive value of $\mathrm{IG}_{X\to Y}(\tau)$ indicates instead that the inclusion of X_0 improves the prediction of the future state of Y, providing evidence for a putative causal link at lag τ. As already discussed, the observation of $\mathrm{IG} > 0$ between two variables does not imply that a causal link between the two ex-

ists, since both variables could be caused by a third unobserved variable. However, if one finds that IG = 0 (or small, in practical applications) one can conclude that the causal link does not exist [15]. Since the IG is estimated using distances between configurations, it naturally accommodates high-dimensional descriptors and avoids the computational problems that are encountered when one attempts to infer causal links from Transfer Entropy, which requires estimating conditional probability distributions in high-dimensional spaces [63, 134].

Because the Information Imbalance is computed over a set of points, the estimate presupposes access to multiple replicas of the trajectory. If the available data consist of a single trajectory, an ensemble of replicas is constructed by dividing the trajectory into non-overlapping chunks. The range and resolution of the grid over which α is optimised, and the number of neighbours k, are specified per application in the corresponding chapter.

We also quantify a putative causal link between two variables from the integral of the IG over time (IG Total):

$$\mathrm{IGT}_{X\to Y} = \int_0^\infty \mathrm{IG}_{X\to Y}(\tau')d\tau' \tag{2.30}$$

We note that previous studies using data from equilibrium molecular dynamics simulations have also considered asymmetries in the time-lagged two-body cross correlation functions $\langle X_0 Y_\tau\rangle$ to infer causal links [72, 66]. However such correlation functions are invariant under the exchange of X and Y: $\langle X_0 Y_\tau\rangle = \langle X_{-\tau} Y_0\rangle = \langle X_\tau Y_0\rangle$ (the first equality holds under the assumption of stationarity, while the second follows from time-reversibility). Therefore, asymmetries in these correlation functions, if observed, can only be due to statistical errors, violations of time-reversibility induced by the integrator, and/or by the thermostat/barostat.

3

Crystallization of L-Cysteine in Heavy Water Induces Intrinsic Fluorescence

Developing non-invasive techniques that can probe how solvents modulate the nucleation pathways of bioorganic molecules in solution remains an active and open area of research. Herein, we investigate the crystallization of the amino acid L-Cysteine and show that both the structure of the crystal and its intrinsic fluorescence can be drastically altered by solvent isotope effects. Crystals formed in heavy water exhibit markedly different intermolecular packing as well as strikingly different monomer conformations compared to those in light water. Remarkably, these differences in the supramolecular packing result in significantly elevated intrinsic fluorescence in the crystal that is formed in heavy water. Using a combination of experimental techniques and advanced electronic structure approaches, we elucidate the molecular interactions within the crystals that govern both the electronic origins and the intensity of their emission. These findings demonstrate how tuning the solvent by changing its isotope leads to the emergence of design principles for new intrinsic fluorophores that could serve as novel sensing probes for biomedical applications.

3.1 Introduction

The formation of ordered supramolecular assemblies by various metabolites including amino acids, has been extensively studied in recent years. These simple and biocompatible building blocks can assemble into ordered architectures with unique chemical, physical and optical properties [135, 136, 137, 138, 139]. One of the most intriguing observations within this context, is non-traditional intrinsic luminescence (NTIL) [33, 34], a phenomenon where non-aromatic or more generally, non-conjugated groups under physico-chemical confinement exhibit visible emission. Curiously, this phenomenon appears to contradict the conventional chemical paradigm that fluorescence

in chemical and biological systems arises from aromatic or, more generally, conjugated groups [19].

NTIL is typically associated with hetero-atomic sub-luminophores (HASLs), which are chemical moieties rich in lone-pair electrons or polar bonds, such as amines, amides, ethers, hydroxyls, esters, nitriles, imines, carbonyls (C=O) or sulfur/phosphorus groups. The key principle behind NTIL is that when these HASLs are forced into close proximity and their molecular motion is restricted due to chemical-crosslinking [140, 141] and various forms of confinement [142, 143] due to, for example, non-covalent interactions such as hydrogen-bonding. This unexpected non-aromatic fluorescence (NAF) has now been observed in a variety of biological and synthetic supramolecular assemblies constituting amino acids [32, 144, 37, 10], peptides [28, 145], peptide-based dendrimers [146, 147], oligosaccharides [31], amyloid fibrils [29, 30, 148, 149, 150, 46], and other proteins [151, 47, 40, 48], all of which lack aromatic building blocks. In the last decade, there have been several proposals on the electronic and spectroscopic origins of this phenomenon. This includes the delocalization of electrons across the peptide backbone structure [28], the existence of short hydrogen-bonds that permit proton transfer [30, 37], charge transfer excitations between polar amino acids [47], and constrained carbonyl groups leading to fluorescence [152, 36, 35, 38]. The manifestation of these various effects in different types of aggregates or assemblies [39, 40, 41, 42, 43, 44, 45] with different chemical properties, remains an open area of study from both experimental and theoretical perspectives.

Organic crystals offer excellent model systems for exploring the microscopic origins of NAF since one can determine the key interactions governing this phenomenon in a more controlled setting. It is widely recognized that molecular packing plays a crucial role in organic solid-state emission [52], where even minor changes in molecular arrangement can lead to substantial variations in photophysical properties [49, 50, 51, 153]. For example, different crystal polymorphs of the same compound can display variations in emission wavelengths or engage in fundamentally different photophysical processes, including thermally activated delayed fluorescence (TADF), and persistent room temperature phosphorescence (p-RTP) [154, 155, 156, 157, 158, 159]. Interestingly, very recent work has shown that changing the chemistry and physical interactions associated with amino acids in crystals through the introduction of sugars, can significantly enhance luminescence [160].

Our search for novel compounds exhibiting NAF, has led us to investigate how crystal polymorphism can affect optical properties in L-Cysteine (Cys). The thiol chemistry associated with Cys is essential for stabilizing tertiary and quaternary protein structures [161, 162, 163]. Furthermore, the thiol group also makes Cys extremely sensitive to redox chemistry – a factor that has been linked to the medical pathology of various diseases associated with protein aggregation [164, 165, 166, 167, 168]. Within the context of fluorescence spectroscopy, Cys has also been suggested to function as a quencher of fluorescence of aromatic and heterocyclic compounds [169, 170, 171, 172].

In this work, we present a joint experimental and theoretical study expanding our

understanding of non-aromatic fluorescence focusing on Cysteine crystals grown in H_2O (Cys-H) and D_2O (Cys-D). Over the last decades, there has been a growing appreciation that biophysical processes [173] such as protein folding and aggregation can be subtly modulated using either light (H_2O) or heavy water (D_2O). Surprisingly, these two crystals have starkly different crystal structures as well as optical properties. Interestingly, Cys-D exhibits enhanced fluorescence compared to Cys-H. This enhancement originates from subtle differences in the molecular packing and hydrogen-bonding networks of Cys crystals formed in light versus heavy water. By deploying excited-state non-adiabatic molecular dynamics simulations, we unravel the nature of the electronic states that trigger non-radiative decay pathways involving the amine (C−N) and thiol (S−H) groups differently for Cys structures crystallized in light and heavy water. Our findings form part of a broader effort elucidated in previous studies showing that single amino acids or polypeptide chains in isolation do not display these anomalous optical properties. However, crystals (including L-Cysteine) appear to display blue fluorescence emission, a feature which is strongly modulated by the nature of the interactions that stabilize the crystal or aggregate [174, 33, 32, 34].. We also show the importance of the coupling of both singlet and triplet state emission for Cys-D indicating a much more complex origin in the underlying photoluminescence than is evident at first sight. These insights could guide the rational design of bioorganic fluorophores with tunable emission, opening new avenues for imaging, sensing, and photonics in biological systems.

3.2 Methods

The theoretical background underlying the simulations reported in this chapter — classical molecular dynamics, hybrid QM/MM with electrostatic embedding, and non-adiabatic molecular dynamics within the trajectory surface hopping and SHARC formalisms — is presented in Chapter 2. This section describes only the setup specific to the Cys-H and Cys-D crystals.

3.2.1 Classical MD simulations

The system was first prepared by extracting the crystal structures and building an appropriately large supercell (6×9×5 for Cys-H and 9×6×5 for Cys-D). The Generalized Amber Force Field (GAFF) [175] was used for all MM parameters. First, the system was equilibrated at $300\,\mathrm{K}$ in the NVT ensemble for $10\,\mathrm{ns}$ with a time-step of $2\,\mathrm{fs}$ and with the Stochastic Velocity Rescaling (SVR) thermostat [99] with $\tau = 1.0\,\mathrm{ps}$. Next, a $200\,\mathrm{ns}$ NVT simulation was performed with the same thermostat parameters and a constraint on all bonds involving hydrogen atoms. From this simulation, 40 independent frames were extracted every $5\,\mathrm{ns}$. These frames are used as input for the QM/MM simulations that follow. All the classical simulations were running using AMBER 2023 software [176, 177].

3.2.2 Ground-state QM/MM simulations

To get an accurate idea of the optical properties of these crystals, we need to consider the effect of the environment. Since it is computationally prohibitive to model the dynamics of the entire crystal with a fully quantum-mechanical approach, we use a QM/MM [109, 110] setup with an electrostatic embedding scheme [111], as described in Section 2.2.

First, we take the 40 independent frames we extracted from the classical MD, and then select the relevant QM dimers (see Fig. A.4**a**, A.4**b**). Then we proceed with the equilibration of the temperature of the selected QM dimer region at $300\,\mathrm{K}$. This is done using ORCA 5.0.4 [178, 179, 180, 181, 182] for the QM region, together with AMBER 2023 [176, 177] for the MM. We work at the DFT level of theory with the CAM-B3LYP functional [183] and the Karlsruhe def2-SVP basis set [184, 185]. To speedup calculations, the RIJ-COSX [186] approximation is used to estimate the HF exchange and Coulomb integrals. The QM region is equilibrated in the NVT ensemble at $300\,\mathrm{K}$, employing the SVR thermostat [99] with a very tight coupling ($\tau = 0.01\,\mathrm{ps}$), with no bond constraints, and a time-step of $0.5\,\mathrm{fs}$. After $1\,\mathrm{ps}$, we verify temperature equilibration and then run it for another $5\,\mathrm{ps}$ with a more typical thermostat coupling ($\tau = 0.5\,\mathrm{ps}$). The final frames (nuclear positions and velocities) of these equilibrated trajectories are taken as the initial conditions for the NAMD QM/MM simulations that follow. These frames are also used to perform absorption spectra calculations which give us an insight into the nature and excitation energies of relevant transitions.

3.2.3 Non-adiabatic MD simulations

To study the excited state dynamics and photophysical relaxation of these two systems, we use the Trajectory Surface Hopping (TSH) technique, in particular, the SHARC approach [115, 116], which is an extension of Tully's Fewest Switches Surface Hopping (FSSH) [117] method. This class of methods allow the nuclei to be propagated classically on quantum-chemical potential energy surfaces calculated on-the-fly [118, 119, 120, 121].

As discussed in Section 2.3, the advantage of the SHARC formalism is the ease of incorporating both time-dependent non-adiabatic coupling (TD-NAC) and spin-orbit coupling (SOC) on the same footing. We include the first 5 singlet states ($S_0 - S_4$) and 5 triplet states ($T_1 - T_5$) in the calculation. Notably, in SHARC, the individual M_s components of the triplet states are explicitly included, resulting in a total of 20 electronic states considered in the simulations.

All trajectories were started from S_1 after simulating vertical excitation, and then evolved with a time-step of $0.5\,\mathrm{fs}$ in the NVE ensemble using the velocity verlet integration scheme [96, 187]. The functional and basis set used for NAMD are the same as in the ground-state QM/MM dynamics and excitation energies and oscillator strengths were calculated with TDDFT [188] using the Tamm-Dancoff Approximation (TDA) [189] to overcome well-know problems with triplet instabilities [190, 191]. The

usage of TDA in TDDFT has also been shown to improve its behavior at/near conical intersections for a variety of systems [192, 193, 194].

The time-dependent non-adiabatic coupling is computed using the numerical differentiation scheme suggested by Hammes-Schiffer and Tully [195] which use the overlap integrals between the wavefunctions at different time-steps. In addition, we employed the energy-based decoherence correction proposed by Granucci et al. [122] with a decoherence parameter of 0.1 Hartree. For the integration of the electronic equation of motion, 25 sub-steps were chosen. The trajectories were propagated for $2\,\mathrm{ps}$, with the criteria for a hop to the ground state set at an energy difference between the active state and S_0 at $0.2\,\mathrm{eV}$ in accordance with other studies using TDDFT [196, 197, 198]. In our approach, we adjust the kinetic energy of the system by rescaling the velocity after a surface hop to conserve energy, while no action is taken in case a hop is frustrated due to insufficient kinetic energy.

The TSH simulations were done using a framework of SHARC 3.0.1 [199] (TSH implementation), ORCA 5.0.4 [178, 179, 180, 181, 182] (QM calculations), and TINKER 6.3.3 [200] (MM calculations).

3.3 Results

3.3.1 Experimental Measurements Reveal Enhanced Intrinsic Fluorescence of Cysteine in D_2O vs H_2O

We prepared crystals of L-Cys in D_2O and H_2O using the heating and cooling method [201]. Specifically, L-Cys was dissolved in H_2O and D_2O at $90\,^{\circ}\mathrm{C}$ until the solution became clear and transparent. Subsequently, the amino acid solutions were cooled to room temperature to allow crystals to form.

Following crystallization, confocal microscopy was used to investigate the optical properties of the crystals. The crystals were illuminated by laser light, and their fluorescence was recorded. An excitation wavelength of $405\,\mathrm{nm}$ was used to capture images of fluorescence and brightfield. It was evident from the images taken of the crystals formed in H_2O that the fluorescence was very low, as seen in Fig. 3.1**a**. Crystals formed in D_2O however, showed significantly higher fluorescence than crystals formed in H_2O (Fig. 3.1**b**). Confocal microscopy was also performed on multiple crystals together and the same trend was observed with crystals formed in D_2O displaying higher fluorescence (Fig. 3.1**c** and Fig. 3.1**d**).

To obtain a more quantitative comparison of the fluorescence intensity arising from the crystals formed in H_2O versus D_2O, the emission spectra were extracted from the confocal microscope and the fluorescence signals were quantified. The crystals prepared in D_2O produced more enhanced fluorescence signals that were approximately an order of magnitude higher than those prepared in H_2O (Fig. 3.1**e**). Note that these features were reproduced on three independent measurements on multiple crystals and the standard error of these measurements is shown in Fig. 3.1**e**.

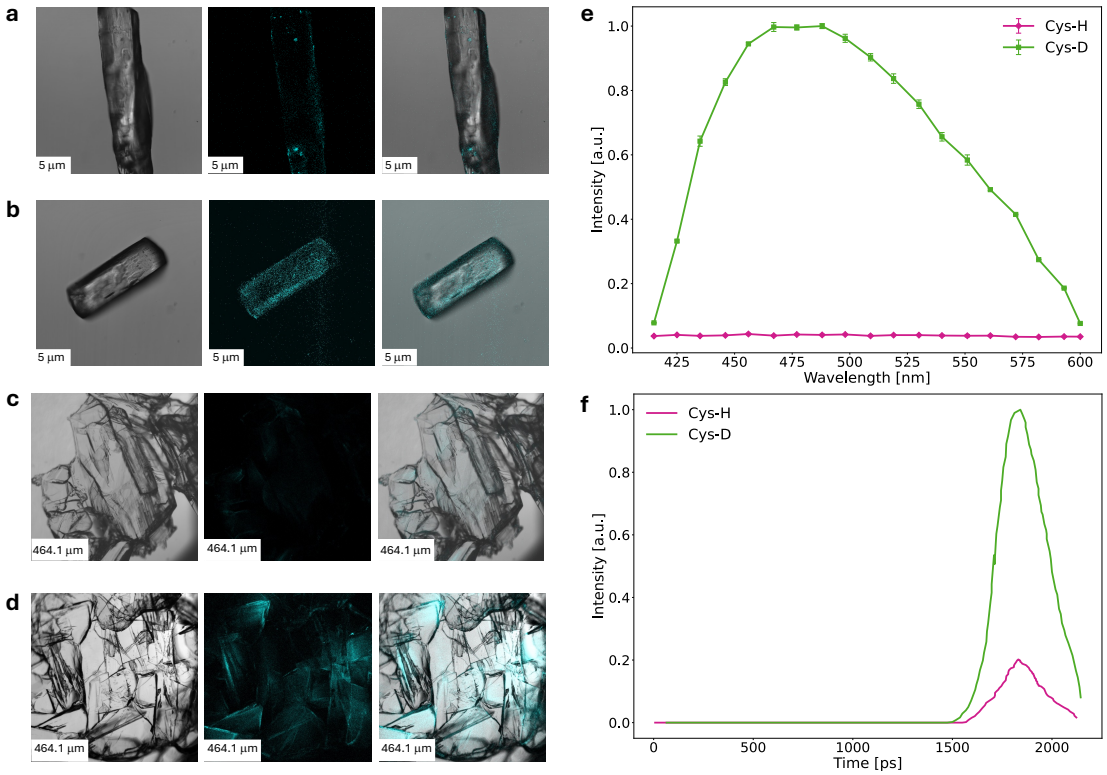


Figure 3.1: *In panels **a** and **b**, the confocal microscopic images of single crystals of Cys-H and Cys-D that are excited at a wavelength of* $405\,\mathrm{nm}$ *are shown, respectively. In panels **c** and **d**, the confocal microscopic images of aggregate of crystals of Cys-H and Cys-D that are excited at a wavelength of 405 nm are shown, respectively. Each of these four panels consist of brightfield, fluorescence, and merged image of the crystals (from left to right). In panel **e**, the emission spectra of Cys-H (pink) and Cys-D (green) captured with a confocal microscope is shown. The standard error over 3 independent measurements are highlighted with vertical bars as well. In panel **f**, the FLIM spectra after excitation at 480 nm of crystal aggregates captured with a time-correlated single-photon counting (TCSPC) system for Cys-H (pink) and Cys-D (green) is shown.*

Although the preceding results indicate that the optical properties of the crystals in H_2O and D_2O are different, the setup of the confocal microscope has a limited sensitivity for resolving the fluorescence intensity. We thus turned to using a time-correlated single-photon counting (TCSPC) system to conduct fluorescence lifetime imaging microscopy (FLIM). Cys crystals were uniformly distributed on a microscope slide prior to imaging. The FLIM system, equipped with a TCSPC card and two detectors, was used to capture images and record the fluorescence lifetimes of the crystals in both H_2O and D_2O under identical conditions. Using FLIM, we also find that the crystals in D_2O display enhanced fluorescence compared to those in H_2O (Fig. 3.1**f**). In addition, the absolute quantum yield was determined for both Cys-H and Cys-D at room temperature on multiple crystals, the former being $0(\pm 0.05)\%$ (that is, there was no measurable quantum yield within the error bars of the instrument), and the latter at $8(\pm 0.02)\%$ (see Fig. A.1 for details).

3.3.2 Supramolecular Packing of Cysteine Crystals are Different in D_2O vs H_2O

Cysteine has two well-known crystal polymorphs when crystallized in H_2O at both ambient and low temperatures. One of them is a monoclinic crystal [202, 203], while the other is orthorhombic [204, 205, 206]. The striking differences between the optical properties of Cys-H and Cys-D indicate that there must be some important structural differences between the two crystals. Therefore, to understand the structural features contributing to the differences in the optical properties between the two crystals in H_2O and D_2O, we first investigated their crystal packing. We resolved the latter structure by performing single-crystal X-ray diffraction (SC-XRD) (see Section A.1 for more details), while the former was taken from Ref. [203]. Crystal data collection and refinement parameters are shown in Appendix Table A.1. To the best of our knowledge, this is the first time the structure of Cys has been crystallized and characterized using D_2O as a solvent. Neither Cys-H nor Cys-D contains any crystallographic light/heavy water molecules. One interesting feature to note is that the Cys-D structure bears a striking similarity to the orthorhombic structures that have been observed in Refs. [204, 205, 206, 207], indicating an underlying crystal polymorphism. We have independently verified that single-crystal X-ray diffraction analysis of Cysteine in H2O yields a monoclinic structure consistent with previous reports in the literature (see Appendix Table A.2 for more details).

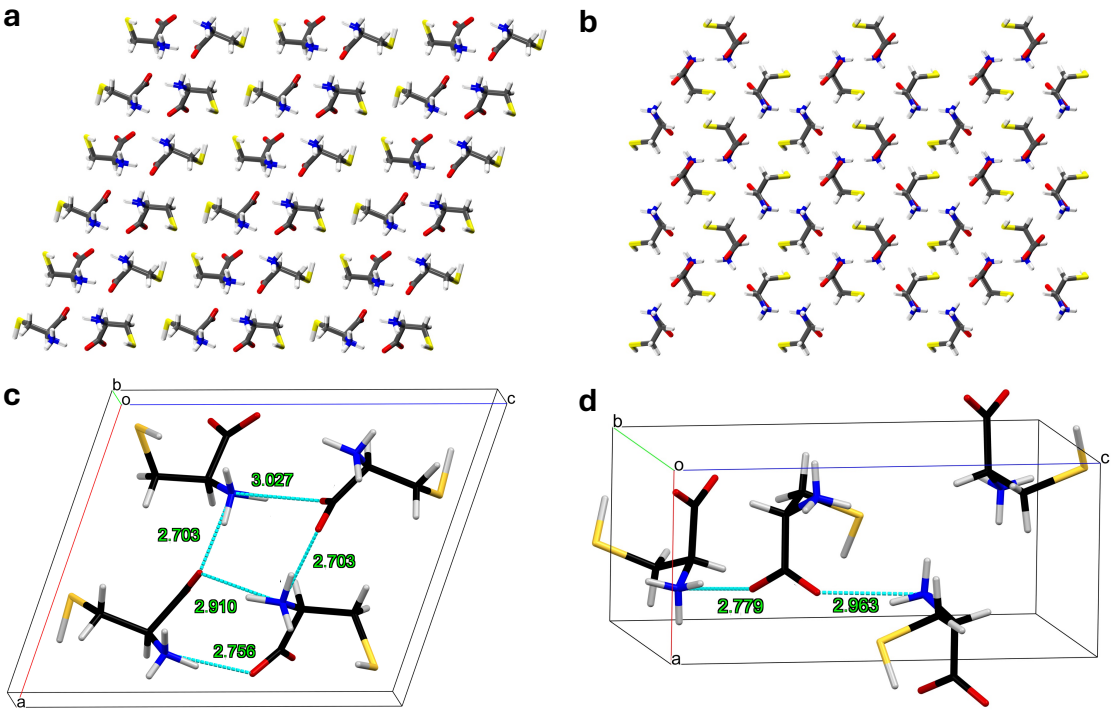


Figure 3.2: *Panels **a** and **b** show an overview of the crystal packing in Cys-H [203] and Cys-D respectively. Panel **c** shows the monoclinic unit cell of Cys-H, while **d** shows the orthorhombic unit cell of Cys-D, each containing 4 monomers, with some typical hydrogen bonds and their lengths between O (red) and N (blue) atoms highlighted.*

The two different structures thus formed by crystallizing cysteine in H_2O (Cys-H) and D_2O (Cys-D) are shown in Fig. 3.2**a** (monoclinic) and Fig. 3.2**b** (orthorhombic). Both crystals consist of 4 monomers in the unit cell with subtle differences in the packing arrangement, specific interactions involving the hydrogen bonds (HBs), and the conformations of the cysteine molecules. Specifically, the N-C termini HBs typically range between 2.70–3.02 Å in Cys-H, (Fig. 3.2**c**) while in Cys-D (Fig. 3.2**d**) they range between 2.76–2.96 Å. Differences along other dihedral angles involving the $O-C-C_\alpha-C_\beta$ are also observed (see Fig. A.2) which ultimately play an important role in affecting the fate of cysteine on the excited state.

While the XRD experiments provide insight into the different crystal structures of Cys-H and Cys-D, they do not reveal the positions of hydrogen atoms, as these cannot be resolved. However, given the subtle differences between Cys-H and Cys-D, it is important to establish whether these crystallographic differences arise from a hydrogen isotope effect or an environmental solvent effect. Therefore, electrospray ionization mass spectrometry (ESI-MS) in positive ion mode was performed on both crystals and the results are shown in Fig. A.3. The mass spectrum of Cys-H (Fig. A.3**a**) shows an intense m/z peak at 122.1, which is consistent with the fragment of the protonated molecule ($[cySH+H]^+$) [208]. The same m/z peak is observed at 122.4 for Cys-D (Fig. A.3**b**), thus confirming the presence of hydrogen atoms in both crystals. Other characteristic peaks of cysteine, i.e. at m/z 152 and 241, can be also observed in the spectra, as well as the ion peak corresponding to sodiated cystine ($[cySScy+Na]^+$) at m/z 263. The detected ion peaks provide further validation that the differences in the physical properties in the two structures do not originate from isotope differences - all non-carbon, non-nitrogen, and non-sulfur atoms are hydrogen in both crystals.

3.3.3 Molecular Mechanisms and Electronic Origins Behind Intrinsic Fluorescence of Cysteine in Heavy Water

The preceding experimental results show a striking effect of the solvent (H_2O vs D_2O) on the resulting cysteine crystal structures. To understand the differences in the optical properties of Cys-H and Cys-D, we turned to conducting mixed quantum/classical molecular (QM/MM) electronic structure calculations that probe the absorption, emission, and non-radiative decay mechanisms. Time-Dependent Density Functional Theory (TDDFT) with the CAM-B3LYP functional was used for all our electronic structure calculations (see Section 3.2 for more details). We carved out two monomers from each unit cell forming a dimer which was treated at QM level (see Fig. A.4). Note that the criterion for choosing these QM dimers was based on making sure that the highest number of HBs were formed between the N and C termini. This was then embedded in a periodic environment consisting of MM cysteine molecules. Given the high computational cost of these simulations, we strike a careful balance between efficiency and accuracy, guided by previous approaches introduced by some of us [38, 6]. In Section A.2, we present validation tests using more accurate electronic structure methods

and a larger QM region to justify our approach.

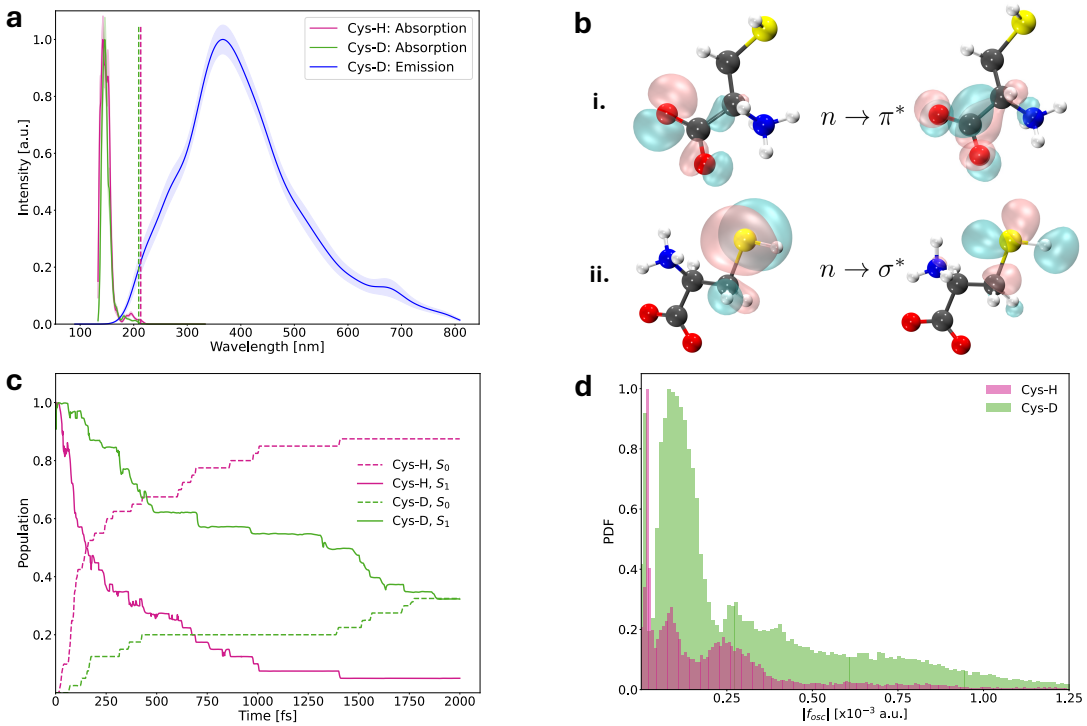


Figure 3.3: *In panel **a**, the theoretically computed absorption spectra for Cys-H (pink) and Cys-D (green) and fluorescence emission spectrum (blue) for Cys-D is shown. The shaded blue region denotes the standard error in theoretical estimation since it is an average over 40 independent trajectories. The dashed vertical lines (green - 209 nm, pink - 215 nm) represents the wavelength corresponding to the $S_0 \rightarrow S_1$ transition in our electronic structure calculations for Cys-D and Cys-H, respectively. Panel **b** shows the 2 types of $S_0 \rightarrow S_1$ transitions that occur commonly. In (**b.i**), we see an $n \rightarrow \pi^*$ transition that occurs on the carbonyl (CO) group, whereas in (**b.ii**) we see an $n \rightarrow \sigma^*$ transition that is present in the SH moiety. Panel **c** shows the evolution of the average population of the singlet electronic states, where S_1 and S_0 are denoted by solid and dashed lines respectively. Results corresponding to Cys-D and Cys-H are shown in green and pink, respectively. In panel **d**, the oscillator strengths obtained from the TSH simulations, corresponding to $S_1 \rightarrow S_0$ emission are shown. Results corresponding to Cys-D and Cys-H are shown in green and pink, respectively.*

We first determined the absorption spectra for both systems using TDDFT, based on a sample of 40 independent configurations from the ground state equilibrium simulations (see Section 3.2 for further details). Figure 3.3**a** shows the absorption spectra for Cys-H and Cys-D. In both cases, the lowest excitation energies are found to be around 210 nm (5.9 eV) (see Fig. A.5**a** for a zoomed-in version). To gain further insight into the nature of the low-lying electronic transitions, we examined the molecular orbitals involved in the $S_0 \rightarrow S_1$ excitations. Figure 3.3**b** illustrates the molecular orbitals associated with two typical excitations observed in both systems: i) n $\rightarrow \pi^*$ transition involving the C-terminus carboxylate groups and ii) n $\rightarrow \sigma^*$ transition localized on the S−H (thiol) group.

Although both of these transitions can characterize the $S_0 \rightarrow S_1$ excitation, there are differences in their occurrence in both systems. In Cys-H, the initial excitation is more or less equally probable to be of either type, while for Cys-D, it is much more likely

to be of type i) than type ii) (see Appendix Table A.3 for details). The differences in crystal packing between Cys-H and Cys-D seem to significantly influence the nature of electronic transitions in the Franck-Condon region. Consequently, the variations observed in fluorescence are expected to become more pronounced as the system evolves in its electronically excited states.

Analysis of the 40 simulated trajectories for each system revealed distinct non-radiative decay pathways for Cys-H and Cys-D. In Cys-H, 90% of trajectories decayed non-radiatively to the ground electronic state, while in Cys-D, this pathway was observed in only 33% of trajectories. This difference corroborates the experimentally observed enhancement of emission intensity upon the change of packing from the crystal structure of Cys-H to that of Cys-D. Furthermore, Cys-D displays sizable intersystem crossing (ISC), with an additional 33% of trajectories populating the triplet states. In contrast, ISC was less prevalent in Cys-H, accounting for only 7.5% of the simulated trajectories. At the end of our $2\,\mathrm{ps}$ NAMD simulations, only 5% of trajectories remained in S_1 for Cys-H, while 33% did so for Cys-D. For the latter, we can use the trajectories that remain on S_1 to estimate the fluorescence emission spectrum which is shown by the blue curve in Figure 3.3**a**. This emission spectrum is blue-shifted compared to the experimental spectra by about $80\,\mathrm{nm}$ $(0.5\,\mathrm{eV})$ which is within the theoretical error bars of the electronic structure methods we are employing. Later on (see Section A.2 and Fig. A.8), we present benchmarks to determine theoretical error bars for our choice of electronic structure method and discuss their implications for comparisons with experimental results.

We begin first by examining the behavior of the excited state dynamics involving the singlet states. Figure 3.3**c** shows the average electronic populations of the S_0 and S_1 states for both Cys-H and Cys-D. Here we clearly observe that in Cys-H, the S_1 population reduces significantly within the first $250\,\mathrm{fs}$, and thereafter continues decaying further, while for Cys-D, the population decay occurs over a significantly longer timescale. The corresponding curves for the S_0 state are also shown mirroring the behavior of the S_1. These theoretical predictions are thus fully consistent with the experimental observation of photoluminescence only in Cys-D.

The lower probability of non-radiative decay in Cys-D, on its own, is not sufficient to lead to non-aromatic fluorescence. Specifically, we also need the excited state on which the system exists to be *bright*. This is characterized in terms of the oscillator strength between S_1 and S_0. In Figure 3.3**d** we see the probability distribution function of the oscillator strengths for the $S_1 \rightarrow S_0$ transition, and we can see clearly that Cys-D has a higher probability of having a larger oscillator strength than Cys-H, thereby contributing to a *bright* photoluminescence which is, again, in good agreement with the experimental results.

Having established the differences in the energetic and electronic properties on the excited state of Cys-H and Cys-D we next dissect the vibrational modes that lead to non-radiative decay in both systems. In recent works [30, 36, 37, 38], research by some of us has shown that specific vibrational modes such as the carbonyl stretch, proton

transfer, and deplanarization play a critical role in modulating the excited-state lifetimes of glutamine, peptide chains, and amyloids. In particular, by constraining/locking these modes the barriers to accessing conical intersections (CIs) can be raised. As a result, this pathway is either delayed or completely inhibited, reducing the rate of the non-radiative decay and thereby increasing the likelihood of photon emission. The excitations involving the n-π^* and n-σ^* shown earlier in Figure 3.3**b**, prompted us to examine vibrational coordinates associated with these regions of the molecules. Specifically, in the case of the n-π^* transition, a natural expectation would be the lengthening of the C=O bond while for the n-σ^*, there would be an increase in the S−H bond length. Appendix Table A.4 summarizes the statistics associated with the four main vibrational modes we identified that play a role in the photochemistry, details of which are now discussed.

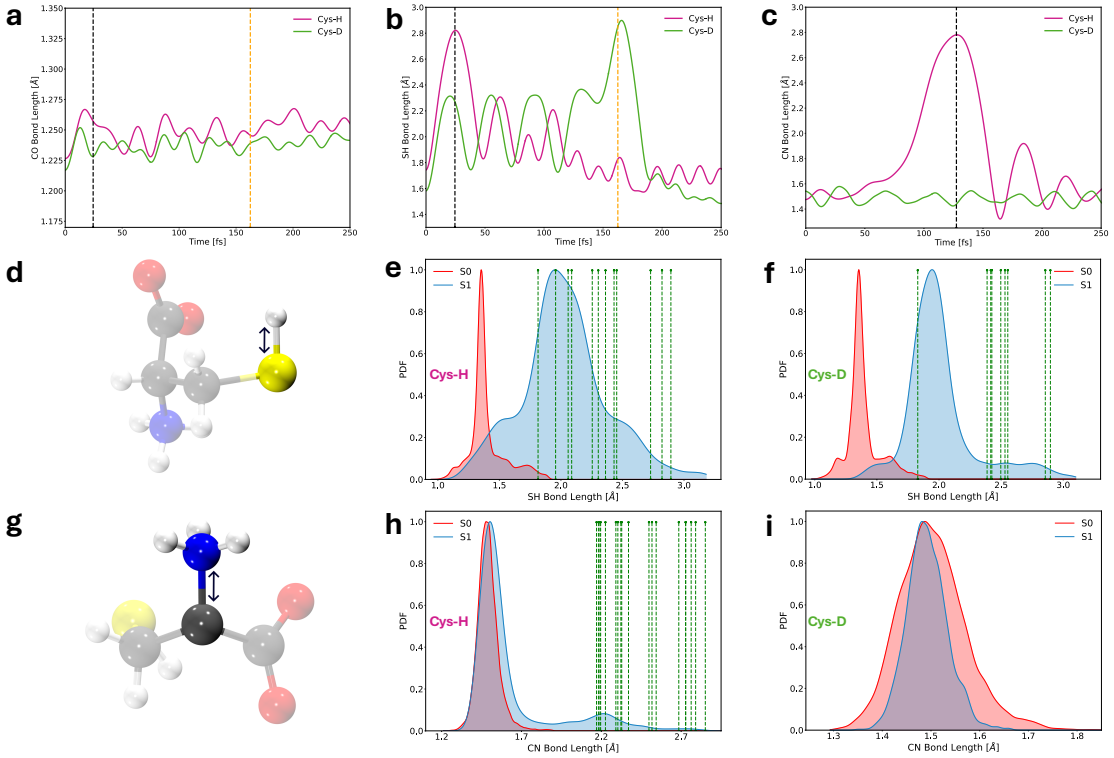


Figure 3.4: *Panels **a**,**b**,**c** show the time-series evolution of CO, SH, CN bond lengths across model trajectories. The bond lengths for Cys-D and Cys-H are shown in green and pink respectively, while the time of encountering the conical intersection is shown with orange and black dashed vertical lines, respectively. The 2 major non-radiative decay modes are the stretching of the S−H and C−N bonds, depicted in panels **d** and **g**. The Probability Distribution Function (PDF) of the S−H bond for Cys-H and Cys-D are respectively shown in panels **e** and **f**, where the PDF corresponding to S_0 is shown in red and the one corresponding to S_1 is shown in blue. The dashed green lines depict the bond lengths at which the conical intersection is reached, and non-radiative decay occurs from $S_1 \rightarrow S_0$. Similarly, in panels **h** and **i**, the PDF of the C−N bond for Cys-H and Cys-D are shown.*

A visual inspection of some of the excited-state trajectories indicated that most of the activity involves two vibrational modes, the S−H and C−N bond lengths. Although the initial excitation originates in an n-π^* transition localized on the carbonyl group, the C=O bond length in all our trajectories displays only moderate fluctuations ranging

between 1.1–1.4 Å and it therefore does not appear to be a key player in the deactivation mechanism. On the other hand, for the excitations involving n-σ^* near the sulfide group, the S−H bond undergoes large geometrical changes which ultimately lead to non-radiative decays in both Cys-H and Cys-D.

Investigating a sample trajectory for both Cys-H and Cys-D that undergoes non-radiative decay, we show the time-series of the C=O bond (Fig. 3.4**a**) and observe no significant distortions that coincide with the system encountering the CI and decaying non-radiatively to the ground state. Instead, in the time series of the S−H bond (Fig. 3.4**b**) we observe a large extension of this bond at the point the system decays to S_0. For Cys-H, 12 out of the 35 decays to the ground state occur through this mechanism, whereas for Cys-D, 9 out of 13 non-radiative decays to the ground state follow this pathway.

From the preceding analysis, the majority of the non-radiative decays for Cys-H remain unexplained and this led to us looking for alternative decay modes. In Fig. 3.4**c** we show a different trajectory where the C−N stretches to a large value (2.8 Å) in concert with a decay of the system to the ground state. In fact, this turns out to be the dominant mode of non-radiative decay to the ground state for Cys-H, with 19 out of the 35 decays following this mechanism. Remarkably, this mode of decay appears unique to Cys-H and does not occur in Cys-D. In fact, the subtle changes in the underlying electronic structure are ultimately rooted in the differences of the O−C−C_α−C_β dihedral angles between Cys-H and Cys-D. Specifically, looking at the probability distribution of this dihedral while the system is in S_1 (see Fig. A.7**b**), we see that a planar geometry (0°, 180°, 360°) is far more probable in Cys-H and effectively not sampled in Cys-D. This means that the extended O−C−C_α−C_β planarization is likely to lead to an enhancement of resonance within the molecule for the crystal formed in heavy water.

To conclude the discussion of decay mechanisms (which are summarized in Appendix Table A.4), the remaining 4 out of 35 decays for Cys-H were through the stretching of the C−S bond, often involving a dissociation of the same. For Cys-D, there was 1 out of the 13 decays involving a C−S stretch. The remaining 3 decays for Cys-D were through a stretch of the C−COO^- bond, but these decays, which followed ISC, will be elaborated on in Section 3.3.5.

To better quantify the differences in the ground and excited state distributions along some of these coordinates and how they affect the access to the CI, we show probability distributions of the S−H and C−N modes in the ground and excited state for Cys-H and Cys-D in Figure 3.4. Looking at the S−H mode (visualized in Fig. 3.4**d**), we clearly see that in both Cys-H (Fig. 3.4**e**) and Cys-D (Fig. 3.4**f**) the probability distribution of the S−H bond when the system is in S_1 is peaked around 2.0 Å, while in S_0 the equilibrium bond lengths are usually between 1.2–1.5 Å. The bond lengths at which the system encounters the CI are marked through green vertical lines, and they range between 2–3 Å. The position of the CI appears to be closer to the peak of the distribution of the S_1 state in Cys-H compared to Cys-D, consistent with the idea of a less constrained mode in the former as observed in an earlier study for the carbonyl stretch [38].

Looking at the probability distribution of the C−N mode (visualized in Fig. 3.4**g**) in both Cys-H (Fig. 3.4**h**) and Cys-D (Fig. 3.4**i**), we clearly see that the bond can stretch to larger values in S_1 only in Cys-H, while in Cys-D it remains confined between 1.4–1.6 Å. This leads to the system decaying through the conical intersection along the C−N mode only in Cys-H, and the enhanced optical activity of Cys-D can be primarily attributed to the absence of this mode of decay. Furthermore, the large overlap in the distribution of the C−N mode in Cys-D creates a situation that makes accessing the CIs along this mode more difficult. We will elaborate more on the C−N mode and its uniqueness in Cys-H in the following Section 3.3.4.

3.3.4 Inhibiting C−N Distortions in Cysteine Luminescence

We have shown that the key modes leading to the non-radiative decay are the stretch of the S−H and C−N bonds. Extending the S−H bond leads to easy access of the conical intersection in both Cys-H and Cys-D. On the other hand, the C−N mode is found exclusively in Cys-H. The differences between the two systems originate from a combination of subtle characteristics induced by the conformation of the cysteine molecules in each crystal and their local environment.

We first examined the evolution of the electronic structure of the S_1 state during a typical decay event. The vast majority of trajectories undergoing a non-radiative decay via the C−N stretch, start out with an initial $n - \pi^*$ transition on the C=O at the point of photo excitation (Figure A.7**a.i**). Examining the transitions across the trajectory, we notice that the electron density delocalizes along the $C-C_\alpha-C_\beta$ chain, and this simultaneously gives rise to a partial $n - \sigma^*$ on the $C_\alpha-N$ bond. These changes make the $C_\alpha-N$ bond weaker allowing it to stretch more easily (Figure A.7**a.ii**).

These subtle variations in the underlying electronic structure ultimately stem from differences in the $O-C-C_\alpha-C_\beta$ dihedral angles between Cys-H and Cys-D. Specifically, examining the probability distribution of this dihedral in the S1 state (Figure A.7**b**) reveals that a planar geometry (0°, 180°, 360°) is significantly more probable in Cys-H, while it is scarcely sampled in Cys-D. This suggests that the extended $O-C-C\alpha-C_\beta$ planarization in Cys-H likely enhances resonance within the molecule.

Another feature that plays an important role is local structural differences between the two crystals. In particular, there appear to be a larger number of oxygen atoms around the N-terminus in Cys-H compared to Cys-D (see the N-O running coordination number function in Figure A.7**c**)). This negative electrostatic potential around the N-terminus serves as a trigger to facilitate dissociation of the C-N bond. This feature is also reflected in potential energy surface scans along the C-N bond which show a barrier of over 1 eV for dissociation in the S_1 state in Cys-D compared to Cys-H (Figure A.7**d**).

3.3.5 Cysteine in Heavy Water Exhibits Synergistic Singlet and Triplet Emission

As alluded to earlier, in Cys-D, we observe that the photochemistry is also heavily influenced by intersystem crossing. Fig. 3.5**a** shows the time-evolution of the population of the triplet state for both Cys-H and Cys-D where one clearly sees that the significant increase in the triplet population appears to be a unique feature of Cys-D, where about 35% trajectories remain in a triplet state at the end of our $2\,\mathrm{ps}$ dynamics, while 7.5% of trajectories for Cys-H do the same. In fact, 17 out of 40 trajectories undergo ISC for Cys-D, and 3 of these decay to the ground state from the triplet states. While no clear vibrational mode could be associated to the $S_1 \rightarrow T_1$ ISC transitions, the $T_1 \rightarrow S_0$ transitions can be associated with a clear stretch of the $C-COO^-$ bond. This also appears exclusively in Cys-D, and the few trajectories that do undergo ISC for Cys-H do not decay to the ground state via a further ISC.

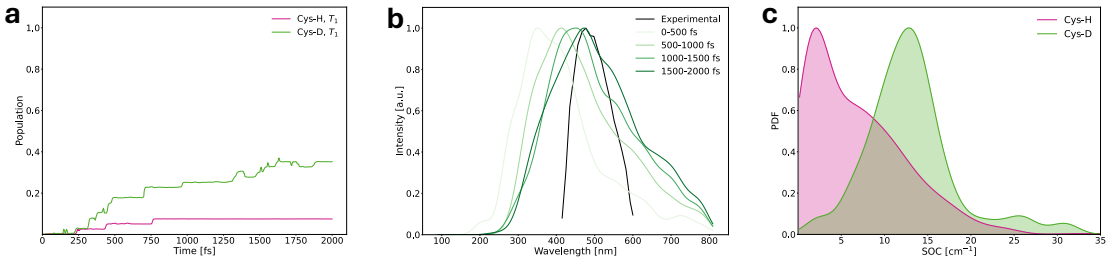


Figure 3.5: *Panel **a** shows the cumulative triplet state population of Cys-H (pink) and Cys-D (green). In panel **b** we see the time-dependent emission spectra from the triplet state for Cys-D, in blocks of* $500\,\mathrm{fs}$*, with the shades of green getting darker for successive blocks of time, while the experimental emission spectra is shown in black. Panel **c** shows a PDF of the SOC computed from 1000 random frames for both Cys-H (pink) and Cys-D (green).*

Our simulations indicate that the ISC in cysteine crystals occurs on a sub-picosecond timescale. Several previous theoretical and experimental studies in a wide-variety of systems have shown [209, 210, 211, 212, 213, 214] that intersystem crossing can occur on similar timescales. This suggests that the observed luminescence observed on the nanosecond timescale in the experiments may arise from a mix of both singlet and triplet emission. In order to provide a theoretical prediction of the possible emission spectra arising from the lowest lying triplet state, the $T_1 \rightarrow S_0$ energy gap, the corresponding transition-dipole moment, and oscillator strength need to be determined. We do this on a subset of configurations, selecting 100 random frames for each $50\,\mathrm{nm}$ block in the emission energies between $200\,\mathrm{nm}$ to $800\,\mathrm{nm}$. This gives us mean oscillator strengths for each range of energies which we then use to compute the time-dependent emission spectra in blocks of $500\,\mathrm{fs}$ (Fig. 3.5**b**). Here we observe that the peak of the emission energies appears to converge, perhaps fortuitously, to the peak of the experimental emission spectra. On the other hand, if we look at the time-dependent emission spectra from S_1, it converges to approximately $400\,\mathrm{nm}$ which is an $80\,\mathrm{nm}$ $(0.5\,\mathrm{eV})$ blue-shift compared to the experimental results (see Fig. A.5**b**). It is important to note here that the oscillator strengths for triplet emission $(\sim 10^{-6})$ are about three orders of

magnitude smaller than those corresponding to the singlet emission ($\sim 10^{-3}$). Consequently we would expect any emission arising from the triplet states to be much weaker than the singlet emission.

Since the vast majority of the ISC occurs from $S_1 \rightarrow T_1$, we look at the likely values of the spin-orbit coupling (SOC) for the two systems in Fig. 3.5**c**. Having computed the PDF of the SOC over 1000 randomly selected frames while the system is in S_1, we see a clear difference in the two systems, with Cys-D (green) much more likely to have higher values of the SOC than Cys-H (pink), which is consistent with the behavior we observe in our simulations. The significant ISC observed in Cys-D can be further quantified, by visualizing for all trajectories, the cumulative probability of ISC while the system is in S_1. In Fig. A.6**a**, we clearly see that trajectories of Cys-D have a much higher cumulative probability of ISC than Cys-H. As we have already noted above, Cys-D is more likely to have a higher SOC than Cys-H, and this contributes to the increased ISC probability for the former. Part of the reason might also be the fact that Cys-H undergoes non-radiative decay to S_0 so quickly that the system has no opportunity to experience ISC. Another way of understanding the preferential ISC experienced by Cys-D is by looking at the distribution of energy gaps between S_1 and the closest triplet state [215] (which is usually T_1) in Fig. A.6**b**. In Cys-H, ISC from the singlet excited state is less probable due to the fact that there is a larger gap to the energetically closest triplet state. The distribution has a larger peak near 0 for Cys-D than for Cys-H, which coincides with more frequent occurrences of nearly-degenerate S_1 and T_1 states during Cys-D trajectories.

3.4 Discussion

In this joint experimental and theoretical work, we have taken an important step to understanding new governing principles underlying the phenomenon of intrinsic fluorescence in bioorganic crystals. Cysteine molecules crystallized in light and heavy water display rather striking differences in their crystallographic packing and optical properties. Using excited-state simulations, we rationalize the electronic and vibrational origins of the enhanced fluorescence observed in the crystal that is formed in heavy water.

Recent studies have highlighted that the careful incorporation of heavy atoms, such as sulfur, can improve the efficiency of thermally activated delayed fluorescence (TADF) photocatalysts [216]. Our findings indicate that Cys-D, with its relatively high SOC, could open new possibilities for the design of such materials. In addition, prior research [174] has shown that crystallised non-aromatic amino acids (L-Lysine, L-Serine, L-Isoleucine) can also display persistent room temperature phosphorescence (p-RTP) in addition to fluorescence. These previous findings reinforce the theoretical predictions made in the current work that synergistic singlet and triplet emission may be observed in L-Cysteine and that this is sensitive to differences in the molecular packing between the two crystals formed in light and heavy water.

The ability to probe in a time-resolved manner nucleation mechanisms of amino acids or proteins in solution remains an open challenge. The intrinsic fluorescence of amino-acids and how they potentially change as a function of clustering or aggregation offers a non-invasive manner to study both thermodynamics and kinetic pathways of crystallization without the introduction of external fluorescent probes.

In addition, solvent isotope effects in biophysical processes such as altering the secondary structure and aggregation of proteins have been observed in a wide class of biomolecular systems. Here we have shown that the simple effect of moving from light to heavy water alters the crystallization pathways and ultimately also the optical properties of the two crystals. It should be underscored that the solvent isotope effects we show in this work on the emission arise exclusively from the manner in which H_2O and D_2O alter the stability of the thermodynamically stable crystal packing of Cysteine molecules in solution. Since both crystals are devoid of any crystallographic water molecules (H_2O or D_2O) and the mass-spectroscopy shows that both crystals consist of only protons (H not D), the effect on fluorescence emission is strongly modulated by the differences in the nature of the electronic states and non-radiative decay pathways as observed in our simulations, which originate from the distinct features in crystal packing. Our results set the stage for exploring the complexity of self-assembly combining both solvent-isotope effects along with changes in intrinsic fluorescence.

In summary, this study establishes a fundamental principle of non-aromatic fluorescence which is a clear testament to how relatively subtle structural packing changes can lead to massively different optical properties for crystal polymorphs of amino acids. Based on this molecular picture, we may be able to rationally design new generations of non-aromatic amino-acid based optical devices of unique photonic and electronic properties as well as intrinsic biocompatibility.

4

Machine learning the non-radiative decay modes in photochemical processes

Non-radiative decay in photoexcited molecular systems is driven by nuclear motion toward conical intersections (CIs), where electronic states become degenerate and nonadiabatic transitions occur. Identifying the nuclear degrees of freedom responsible for CI access from nonadiabatic molecular dynamics (NAMD) simulations remains challenging because the underlying motions are high-dimensional and collective. Here, we introduce an unsupervised, information-theoretic framework based on Differentiable Information Imbalance (DII) to identify the nuclear coordinates governing CI access directly from trajectory surface hopping (TSH) simulations. By quantifying correlations between structural descriptors and electronic observables, including energy gaps and oscillator strengths, the method ranks nuclear degrees of freedom by predictive relevance. A multi-step protocol then extracts low-dimensional, physically interpretable modes associated with non-radiative decay. We apply the framework to the methaniminium cation, furan, L-glutamine, L-pyroglutamine-ammonium, and a photoactive molecular motor. Across all systems, the method recovers known mechanistic coordinates while revealing the relative importance of competing modes when multiple structural distortions contribute to CI access. The analysis also reveals a systematic distinction between observables: energy gaps are typically governed by a small number of localized coordinates, whereas oscillator strengths depend on more collective and distributed structural rearrangements. Overall, the DII-based framework combines predictive power with direct interpretability, providing a general and scalable route for extracting mechanistic insight from high-dimensional NAMD data and constructing reduced-dimensional models of excited-state dynamics.

4.1 Introduction

Photoinduced processes in molecular systems are frequently governed by chemistry associated with ultrafast transitions between electronic states. In polyatomic molecules, these transitions are typically mediated by conical intersections (CIs) [59, 217, 61], which represent points of degeneracy between adiabatic potential energy surfaces. At such geometries, the Born–Oppenheimer approximation breaks down and nuclear and electronic motions become strongly coupled, enabling efficient non-radiative relaxation pathways. Conical intersections therefore play a central role in photochemical reactivity, controlling excited-state lifetimes and branching ratios in systems ranging from small organic chromophores to biological molecules and photoactive materials.

A widely used theoretical framework for investigating these processes is trajectory surface hopping (TSH) [117, 115, 116, 218], which allows for simulating non-adiabatic molecular dynamics (NAMD) in high-dimensional systems in a computationally efficient manner [118, 121]. Modern implementations allow the generation of large ensembles of excited-state trajectories that capture complex relaxation mechanisms and competing decay channels. Despite this progress, extracting mechanistic insight from such simulations remains challenging [219]. In particular, it is often unclear which nuclear degrees of freedom actively drive the approach to the conical intersection within the high-dimensional configuration space sampled by the trajectories.

From a theoretical perspective, the topology of a conical intersection is locally characterized by the branching space, spanned by the gradient difference and non-adiabatic coupling vectors (NACVs), which lift the degeneracy between the intersecting electronic states [220, 221, 60, 222]. However, this description is strictly local and typically obtained at optimized minimum-energy conical intersection (MECI) geometries. In dynamical simulations, trajectories may approach the intersection along collective nuclear distortions that are not trivially related to these local vectors, and the relevant coordinates may involve nonlinear combinations of atomic motions distributed across the molecule [25, 26]. As a consequence, identifying the dominant degrees of freedom governing access to CIs from a large ensemble of trajectories remains a nontrivial problem. This challenge becomes even more pronounced for larger systems where the number of nuclear degrees of freedom grows rapidly and the relevant motions are often distributed across many coupled coordinates [223, 224, 55, 57, 225].

In recent years, machine learning techniques have become increasingly embedded in both the execution and analysis of molecular simulations, in the ground and excited states, as well as in the prediction of molecular excited-state properties [226, 227, 228, 229, 230, 231, 232]. The behavior of molecular systems is generally governed by fluctuations across high-dimensional free-energy landscapes, [233] which complicates the task of extracting clear mechanistic interpretations from simulation data. This challenge has motivated substantial efforts toward developing data-driven methodologies capable of characterizing, interrogating, and rationalizing the complex dynamical behavior observed in molecular simulations. In the context of nonadiabatic molecular

dynamics (NAMD), the selection of such coordinates has often relied on chemical intuition and visual inspection of trajectories, where key distortions are inferred by examining representative geometries or displacement patterns associated with electronic transitions.

A variety of different approaches have been proposed to identify the nuclear degrees of freedom governing excited-state dynamics and nonradiative decay. Early efforts largely relied on normal mode analysis or principal component analysis (PCA) to extract dominant motions from molecular dynamics trajectories, which were then used to construct reduced-dimensionality models such as linear vibronic coupling (LVC) Hamiltonians or to guide quantum dynamics simulations [234, 53, 55, 235]. However, normal mode analysis is intrinsically a local, harmonic description defined around a reference minimum and is therefore most appropriate for small-amplitude motions in relatively rigid systems. In strongly anharmonic regimes, such as those involving large structural rearrangements or dissociation, the normal mode picture breaks down and may fail to capture the relevant nuclear distortions driving the dynamics. PCA has also been widely applied to trajectory surface hopping simulations to identify collective nuclear motions associated with internal conversion or intersystem crossing pathways [53, 54, 56, 57, 58]. However, because PCA ranks modes according to structural variance, it may emphasize large-amplitude or low-frequency motions that are not directly relevant to nonadiabatic transitions, while neglecting smaller-amplitude motions that lead to electronic transitions. This limitation has been explicitly observed in several systems, where dominant principal components correspond to collective motions which may not necessarily lead to non-adiabatic transitions [55].

To overcome these limitations, nonlinear dimensionality reduction techniques, such as diffusion maps and related manifold learning approaches, have been employed to extract low-dimensional reaction coordinates directly from nonadiabatic dynamics simulations [26, 236, 237]. While these methods can capture nonlinear correlations in the data, the resulting collective variables are often difficult to interpret physically and require additional, sometimes ad hoc, procedures to map them back onto chemically meaningful internal coordinates [223, 224]. Alternative strategies based on clustering combined with PCA or multidimensional scaling have also been used to identify distinct dynamical channels, with mechanistic insight obtained by comparing representative structures across clusters [224, 57]. In parallel, information-theoretic approaches have been introduced to quantify the relationship between structural descriptors and electronic observables, for example by computing mutual information between internal coordinates or atom-centered features and quantities such as energy gaps or nonadiabatic couplings [238, 239, 240, 241, 242]. While these methods provide a more direct link between nuclear motion and electronic response, they typically operate at the level of individual descriptors or atoms, and do not directly yield collective modes suitable for reduced-dimensionality dynamical models. These considerations highlight the need for approaches that can identify collective nuclear coordinates directly from dynamical data, while explicitly incorporating their relevance to electronic structure

changes and nonadiabatic processes.

In this work we address this problem using the information imbalance framework introduced by Glielmo and co-workers [13, 14], which provides an information-theoretic approach to quantify the predictive relationship between different representations of a system. The method evaluates how well the neighborhood structure defined by one set of variables can reproduce that defined by another, thereby enabling the identification of coordinates that retain the most relevant information about a target observable [124, 125, 8]. In the context of non-adiabatic dynamics, this allows us to determine which nuclear degrees of freedom best encode the approach to the conical intersection as reflected in quantities such as the electronic energy gap or hopping events along trajectory surface hopping simulations.

By applying this framework to ensembles of TSH trajectories, we identify low-dimensional collective coordinates governing CI access directly from dynamical data, without relying on optimized intersection geometries or predefined reaction coordinates. We first demonstrate the approach on well-characterized model systems, including furan [243, 244] and the methaniminium cation [25, 245, 246, 247], where the extracted motions can be directly compared with established mechanistic interpretations. We then extend the analysis to two amino-acid systems, L-glutamine and L-pyroglutamine-ammonium, which are central to the phenomenon of non-aromatic fluorescence studied extensively in our group (NAF) [32, 144, 37, 30, 36, 35, 38, 1, 9] and exhibit more complex photophysical behavior. NAF refers to photoinduced light emission from chemical systems that lack conventional aromatic or extended π-conjugated chromophores [174, 33, 34, 35, 160]. In the specific cases considered here, experiments showed that L-glutamine is not optically active, but upon heating in water and complexation with an ammonium ion it undergoes a chemical transformation to form L-pyroglutamine-ammonium, a supramolecular assembly stabilized by short hydrogen bonds and displaying enhanced fluorescence [37]. Finally, we consider a well-studied class of photoactive molecular motors [248, 249, 250, 251, 252], illustrating the ability of the method to uncover mechanistically relevant degrees of freedom in larger and structurally diverse systems.

4.2 Methods

In this work, we employ an unsupervised machine learning approach based on the Information Imbalance (II) method to identify the nuclear degrees of freedom most relevant for understanding the mechanism associated with non-radiative decay. The II and its differentiable formulation (DII), together with the Coulomb matrix descriptors used to represent molecular configurations, are introduced in Sections 2.5 and 2.4. Here we describe the chemical systems analysed and the protocol we build on top of these tools.

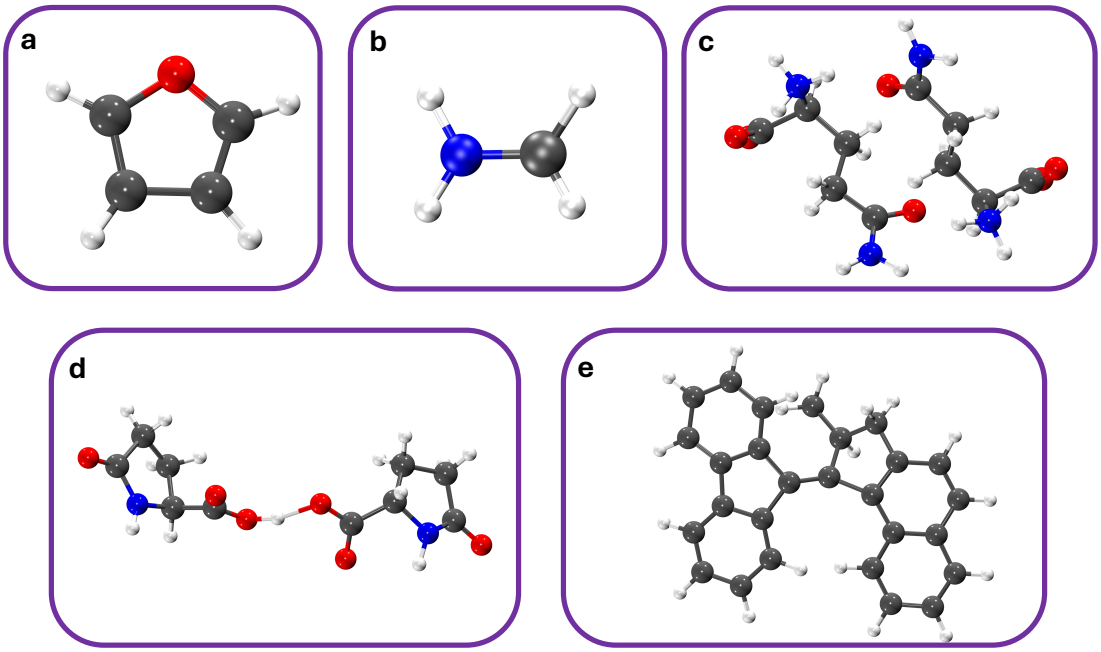


Figure 4.1: *Molecular systems investigated in this work: (a) Furan, (b) Methaniminium cation, (c) L-glutamine, (d) L-pyroglutamine–ammonium, and (e) Molecular Motor: 9-(2,4,7-trimethyl-2,3-dihydro-1H-inden-1-ylidene)-9H-fluorene.*

4.2.1 Chemical Systems

There are 5 systems under study, and these trajectories were published in Refs. [25, 245, 6, 7]: Furan, Methaniminium Cation, L-glutamine (L-gln), L-pyroglutamine-ammonium (L-pyro), and Molecular Motor: 9-(2,4,7-trimethyl-2,3-dihydro-1H-inden-1-ylidene)-9H-fluorene. Fig. 4.1 shows the chemical structures of the 5 systems studied in this work.

All the systems studied in this work were simulated in prior work done by some of the authors using TSH at the time-dependent density functional tight binding (TD-DFTB) level of theory and their non-radiative decay mechanisms were validated against higher levels of electronic structure theory, either time-dependent density functional theory (TD-DFT) or multireference configuration interaction with single and double excitations (MR-CISD) depending on the system. TD-DFT is a widely used electronic-structure approach for computing excited-state properties, including excitation energies, oscillator strengths, and excited-state gradients. TD-DFTB is a semi-empirical approximation to TD-DFT that retains the same general framework while substantially reducing the computational cost, making it well suited for generating large ensembles of NAMD trajectories.

4.2.2 Protocol for Analysis

The main steps of the protocol are summarized in Figure 4.2. Below we outline important technical aspects of each step. The procedure begins with input data coming from NAMD trajectories such as the positions of all particles as a function of time. In

addition, in NAMD simulations, one typically has the evolution of other physical quantities associated with the optical properties such as the electronic energy gap and the oscillator strength.

1. Extract the Coulomb Matrix ($\mathrm{CM}_{ij}^{S_l}$) for each frame of an ensemble of NAMD trajectories in a given electronic excited state, S_l, for each pair of atoms i and j. Accumulate the results over all the frames to make a probability distribution of the elements of the Coulomb Matrix P^{S_l}. We also compute the Coulomb Matrix for all the frames in the state the system hops to ($\mathrm{CM}^{S_{l-1}}$), and the corresponding probability distribution of all pairwise Coulomb Matrix values, $P(\mathrm{CM}_{ij}|S_l)$ and $P(\mathrm{CM}_{ij}|S_{l-1})$. In general the system can hop between any two electronic states ($S_l \rightarrow S_{l-1}$). Specifically for the CI that we will focus on in this work, we look at transitions from $S_1 \rightarrow S_0$.
2. Compute the Jensen-Shannon divergence (JSD), which is a symmetrized version of the Kullback–Leibler Divergence (KLD) [253], between $P(\mathrm{CM}_{ij}|S_l)$ and $P(\mathrm{CM}_{ij}|S_{l-1})$.
   The KLD, which is a measure of similarity between 2 probability distributions R and Q, is defined as
   $$D_{\mathrm{KL}}(R||Q) = \sum_{x \in X} R(x) \log \frac{R(x)}{Q(x)}$$
   Instead, the JSD is defined so as to symmetrize the KLD
   $$D_{\mathrm{JS}} = 0.5(D_{\mathrm{KL}}(R||M) + D_{\mathrm{KL}}(Q||M)), \text{where } M = 0.5(R + Q)$$
   The outcome of determining the JSD is that it gives us "hotspots" of *pairwise atomic interactions* (i.e., Coulomb Matrix elements) that change significantly from one state to another. This allows us to do a first dimensionality reduction step selecting only these "hotspot" atoms . We do this hotspot selection based on a threshold value of the JSD that we select such that the hotspots are those values where the JSD > threshold. These selected Coulomb Matrix hotspots are labeled as $\mathrm{CM}_{\mathrm{hotspot}}^{S_l}$. threshold is an adjustable parameter.
3. Next, we compute the $\mathrm{DII}(\mathrm{CM}_{\mathrm{hotspot}}^{S_l} \rightarrow \Delta E_{l,l-1})$, for all atoms i, j belonging to the "hotspots", assuming we are analysing the NAMD trajectory where the system jumps from S_l to S_{l-1}. In addition, where the oscillator strength (f^{osc}) is available, we also compute $\mathrm{DII}(\mathrm{CM}_{\mathrm{hotspot}}^{S_l} \rightarrow f_{l,l-1}^{osc})$.
   We filter the results of these DII calculations to identify a small set of Coulomb-matrix elements, and therefore the corresponding atom pairs, whose structural variations are most informative of changes in the electronic energy gaps. The filtering is done by selecting the `nfeatures` (default value selected in this work is 10) values with the least DII. We denote these *filtered* hotspots from the Coulomb Matrix as $\mathrm{CM}_{\mathrm{f}}^{S_l}$.
   These coordinates describe nuclear motions that modulate the relative electronic-

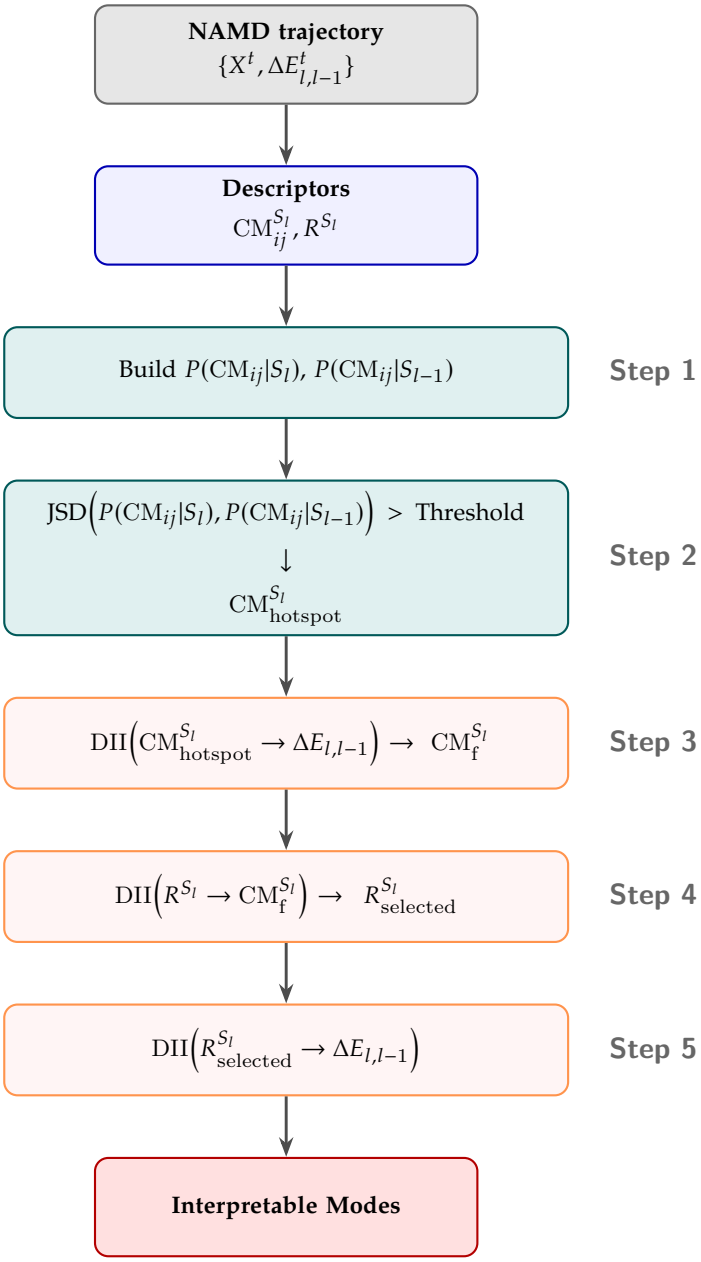


Figure 4.2: *Workflow for identifying state-relevant internal coordinates from trajectory data. Starting from NAD trajectories, Coulomb matrix descriptors are constructed and compared across electronic states using the Jensen–Shannon divergence to identify structurally relevant atom pairs ("hotspots"). These are evaluated via DII against the energy gap to construct a filtered descriptor set* $\mathrm{CM_f}$*. Internal coordinates are then selected based on their predictive power with respect to* $\mathrm{CM_f}$ *and used for final feature selection against* $\Delta E_{l,l-1}$*. Boxes highlighted in teal are steps using the descriptors and statistical analysis methods, and in orange are the steps where the DII is used.*

state energies during excited-state relaxation. When such motions lead to closure of the relevant energy gap, they provide a dynamical signature of approach toward a CI region. Oscillator strengths are then analyzed as a complementary electronic target, allowing us to determine which coordinates govern the intensity of the emission.

4. While the Coulomb-matrix analysis identifies the atom pairs that are most informative of the target electronic observable, these descriptors are not always the most chemically interpretable representation of the underlying motion. We therefore map the selected atoms onto an internal-coordinate representation, which allows the relevant structural changes to be expressed in terms of bonds, angles, and dihedrals.

   Next, for all frames belonging to a given electronic state S_l, we generate the Z-matrix representation $\mathbf{R}^{S_l}$ using a combination of the quantum chemistry software ORCA [178, 179, 180, 181] and an in-house python script, which contains the corresponding internal coordinates, including bonds, angles, and dihedrals. The Z-matrix therefore serves as an initial set of chemically interpretable coordinates rather than an exclusive definition of the candidate feature space. Supplementing it with coordinates involving the atoms identified through the Coulomb-matrix analysis allows relevant motions absent from the initial Z-matrix to enter the subsequent feature selection. Then we compute $\mathrm{DII}(\mathbf{R}_k^{S_l} \rightarrow \mathrm{CM}_\mathrm{f}^{S_l})$, i.e., we compute the DII for every internal coordinate ($\mathbf{R}_k^{S_l}$) to the Coulomb Matrix hotspots that best explain the energy gaps/oscillator strengths. This allows us to select a subset of internal coordinates ($\mathbf{R}_{\mathrm{selected}}^{S_l}$). As in the previous step, here also we do the feature selection such that a specified `nfeatures` (set to 10 by default) of the most informative coordinates are selected as the $\mathbf{R}_{\mathrm{selected}}^{S_l}$.
5. We subsequently do a final round of DII calculations using the selected internal coordinates to determine which ones best predict the energy gaps/oscillator strengths - $\mathrm{DII}(\mathbf{R}_{\mathrm{selected}}^{S_l} \rightarrow \Delta E_{l,l-1})$ and $\mathrm{DII}(\mathbf{R}_{\mathrm{selected}}^{S_l} \rightarrow f_{l,l-1}^{osc})$. This gives us a set of physically interpretable modes that are most informative on the energy gaps and oscillator strengths.

Uncertainties in the DII curves were estimated using 50 independent bootstrap resamples. At each number of selected features, the reported DII and its uncertainty correspond to the mean and standard deviation of the bootstrap estimates, respectively. These uncertainties characterize variability within the available dataset and are considered when interpreting differences between curves and improvements upon adding features.

One motivation for adopting this multistep procedure, rather than directly computing the DII between all internal coordinates and the energy gaps, is to reduce the dimensionality of the candidate feature space. While in principle DII can directly infer the modes from a much larger and highly redundant set of internal coordinates, it needs a number of samples that grows exponentially with the number of features

to perform feature selection in this space. This would be computationally intractable, particularly for larger systems.

The Coulomb matrix provides descriptors invariant under global translations and rotations. This representation is not required by DII itself and alternative structural descriptors can also be used. Furthermore, the descriptors and target observable need not share the same physical units, since the optimized feature weights accommodate differences in their numerical scales. Constructing the Coulomb matrix scales as $O(N^2)$ per configuration, where N is the number of atoms. The subsequent DII cost also depends on the number of sampled configurations and candidate features. In the present applications, the largest system (the molecular motors) contains 47 atoms, and non-adiabatic trajectory generation was the dominant computational expense. For larger assemblies, alternative initial descriptors (such as the internal coordinates themselves), as we have demonstrated in previous applications of the DII [8], or a more restrictive JSD threshold could reduce the downstream feature-selection cost. The latter does not remove the initial descriptor-construction cost and must be checked to ensure that informative coordinates are retained. Neither was necessary for the applications considered here, however.

4.3 Results

In this section, we apply the workflow to a variety of systems of diverse photochemistry. We begin with two benchmark systems that have been extensively studied in photochemistry using NAMD simulations, furan and methaniminium cation, to test whether the selected coordinates recover established non-radiative decay mechanisms. We then analyze two amino-acid systems, extensively studied in our group in the context of non-aromatic fluorescence (NAF) [32, 144, 37, 30, 36, 35, 38, 1], L-glutamine (L-gln) and L-pyroglutamine-ammonium (L-pyro), where we examine descriptors for both the energy gap and the oscillator strength, f^{osc}. Finally, we consider an overcrowded alkene molecular motor, a light-driven molecular machine in which photoexcitation is converted into directional rotation around a central double-bond axle [248, 249, 250]. Because motor function depends on the non-radiative relaxation from the excited state to the ground state, this system provides a chemically motivated test case for identifying the coordinates that govern both non-radiative decay and bright-to-dark state conversion [251, 252, 7].

4.3.1 Furan

Furan is a canonical benchmark for non-adiabatic dynamics and was investigated in a previous study from our group [7] using TSH at the TD-DFTB level of theory. Our analysis focuses on the S_1/S_0 non-radiative decay channel. Fig. 4.3**a** (purple curve) reports DII($\text{CM}_{ij}^{S_1} \rightarrow \Delta E_{1,0}$) as a function of the number of selected Coulomb-matrix (CM) features. This decreases gradually, from about 0.89 to about 0.63 over the first

eight features, indicating that no single atom pair dominates the gap fluctuations. The most informative CM pairs are shown as the inset in Fig. 4.3**a**: one {C,C} and two {C,O} pairs, emphasizing that the relevant structural signal is distributed over the heteroatom-containing ring framework.

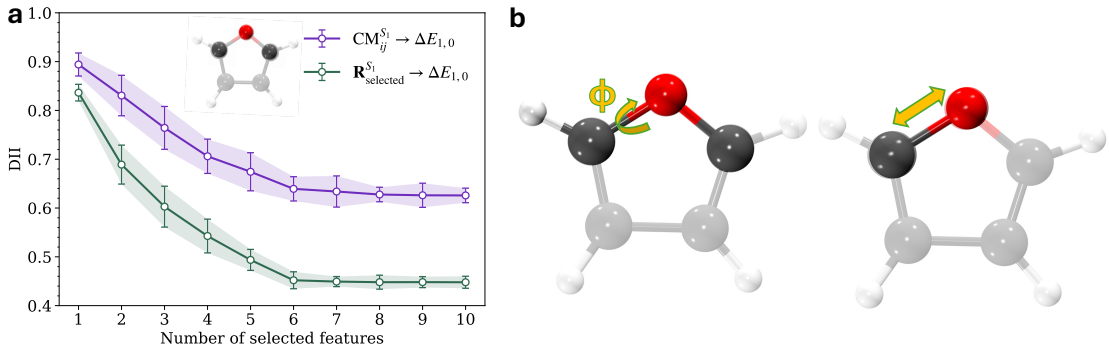


Figure 4.3: (*a*) *DII between Coulomb-matrix descriptors and the* $S_1 \rightarrow S_0$ *energy gap as a function of the number of selected features* (*purple*) *and DII between internal coordinates and* $S_1 \rightarrow S_0$ *energy gap as a function of the number of selected features* (*green*). *Shaded areas show error bars computed from 50 independent bootstrap resamples. The inset shows the most informative atom pairs* ("*hotspots*") *identified from Coulomb matrix analysis are highlighted: the O atom and the two adjacent C atoms.* (*b*) *The most informative structural modes extracted from the DII analysis: the* $C-O-C-H$ *dihedral and a* $C-O$ *stretching coordinate.*

Moving to the chemically interpretable internal coordinates (green curve in Fig. 4.3**a**), we observe that the ranked internal-coordinate representation lowers DII from about 0.84 to about 0.45 only after including six modes, again pointing to a collective decay coordinate. The leading interpretable modes are shown in Fig. 4.3**b**: a C–O–C–H dihedral and a C–O stretching coordinate. Fig. B.1 shows the next-ranked contributions, including the second C–O stretch, the C–O–C angle, and a C–C distance; an explicit ring-puckering coordinate also appears among the top six selected features. Altogether, these modes describe coupled out-of-plane deformation (the C–O–C–H dihedral and C–O–C angle are most notably correlated to the ring puckering) and ring-opening (the C–O and C–C stretches), consistent with prior theoretical studies that associate furan's non-radiative decay with puckering and ring-opening pathways in a multidimensional photochemical landscape [243, 244, 7].

The method introduced earlier to learn the relevant coordinates leading to the conical intersection involve the use of several different hyper-parameters. These include the JSD threshold, the λ parameter in Equation 2.23, the number of trajectories used in the analysis (size of dataset) and finally, the choice from where the data are sampled relative to the position of the conical intersection. For the case of Furan, we have conducted a thorough sensitivity analysis over all these different variables. In summary, the DII curves are rather insensitive to the choice of the JSD threshold. In the case of the λ parameter which governs the width of the softmax function in the DII, as long as this is not made unphysically large, the DII curves are also largely consistent. On the other hand, changing the size of the data set or altering the region of sampling, appears to change the absolute magnitude of the DII. However, the dominant features

that contribute to the DII are broadly preserved. Further details are present in the Appendix B.2.

4.3.2 Methaniminium Cation

Methaniminium cation ($CH_2NH_2^+$) serves as a minimal model for protonated Schiff bases which are critical in the photophysics of retinal [254] and has been investigated with several electronic-structure methods [25, 245, 246, 247]. We analyze TSH trajectories computed with TD-DFTB in a previous work published by our group [7].

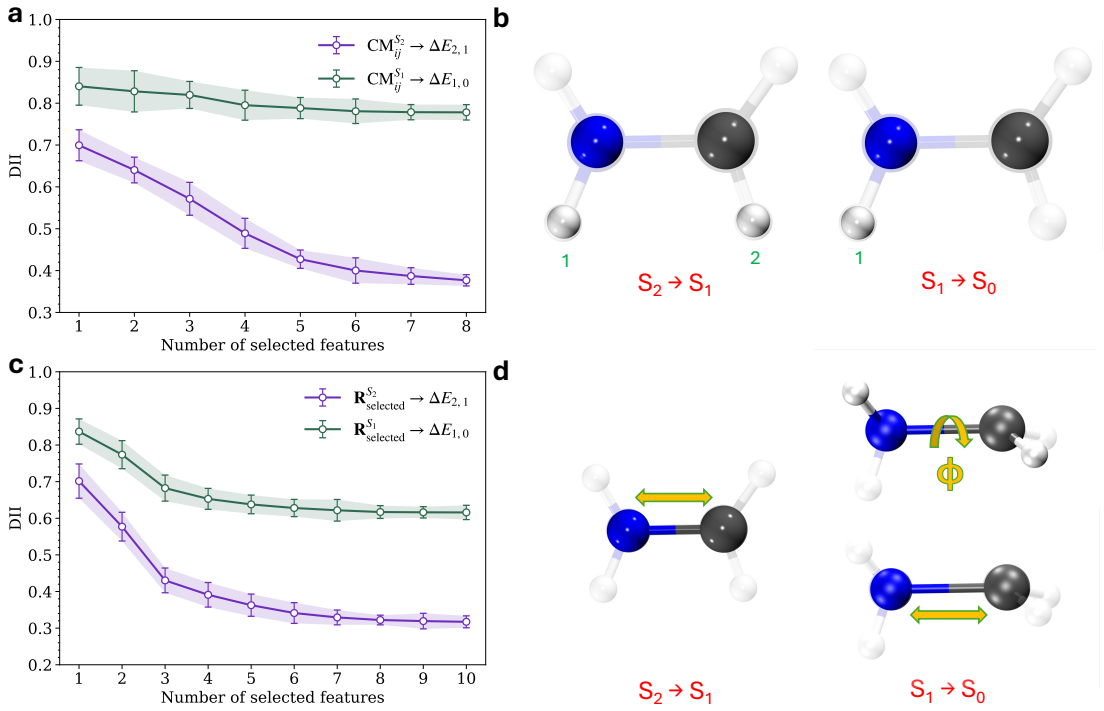


Figure 4.4: *(a) DII between Coulomb-matrix descriptors and energy gaps as a function of the number of selected features. Purple corresponds to the $S_2 \rightarrow S_1$ and green corresponds to the $S_1 \rightarrow S_0$ relaxation channels respectively. Shaded areas show error bars computed from 50 independent bootstrap resamples. (b) Most informative atom pairs ("hotspots") identified from Coulomb matrix analysis are highlighted. The left and right parts of the panel corresponds to the $S_2 \rightarrow S_1$ and $S_1 \rightarrow S_0$ channels respectively. (c) DII between internal coordinates and energy gaps as a function of the number of selected features. Purple corresponds to the $S_2 \rightarrow S_1$ and green corresponds to the $S_1 \rightarrow S_0$ relaxation channels respectively. Similar to panel (a), shaded areas show error bars computed from 50 independent bootstrap resamples. (d) The most informative structural modes extracted from the DII analysis: The $S_2 \rightarrow S_1$ transition is governed primarily by C–N stretching (left), while the $S_1 \rightarrow S_0$ channel involves the $H(C)-C-N-H(N)$ torsion and $C-N$ stretching (right).*

The photochemistry of $CH_2NH_2^+$ primarily involves studying the excitation to the S_2 state and the subsequent non-radiative decays to S_1 and then S_0. The photochemistry of this system is well known to be described by these three states [25, 245]. We thus analyze two non-radiative channels: $S_2 \rightarrow S_1$ and $S_1 \rightarrow S_0$. In Fig. 4.4**a** we report DII($CM_{ij}^{S_2} \rightarrow \Delta E_{2,1}$) (purple) and DII($CM_{ij}^{S_1} \rightarrow \Delta E_{1,0}$) (green), i.e., how well selected Coulomb-matrix (CM) atom pairs explain each energy gap. The $S_2 \rightarrow S_1$ curve decreases substantially faster, indicating that a small number of pairwise descriptors

already captures much of the relevant information for $\Delta E_{2,1}$, whereas $\Delta E_{1,0}$ requires a larger set of pairwise descriptors and the features that are selected are less informative on the energy gap. Fig. 4.4**b** identifies the dominant pairs: for $S_2 \rightarrow S_1$, {C,N}, {N,H2}, and {C,H1}; for $S_1 \rightarrow S_0$, {C,N} remains important, with {N,H1} and {C,H1} completing the top three. These are the aforementioned "hotspots" identified in Step 2 of the analysis protocol.

We then translate these pair-level signatures into chemically interpretable internal coordinates according to Steps 4 and 5 in the analysis protocol. In Fig. 4.4**c**, only 2–3 features are sufficient to approach the minimum DII for $S_2 \rightarrow S_1$, while the $S_1 \rightarrow S_0$ curve decreases more slowly and saturates at a higher DII. The top coordinates are shown in Fig. 4.4**d**. For $S_2 \rightarrow S_1$, the leading mode is clearly C−N stretching, followed by a torsion around the C−N bond (Fig. B.3**a**). This agrees with earlier work identifying the C−N stretch as the dominant coordinate for the S_2/S_1 crossing. In contrast, for $S_1 \rightarrow S_0$, the most informative coordinate is the H(C)−C−N−H(N) dihedral, with C−N stretching as a secondary contributor and N−CH_2 pyramidalization appearing as the next relevant degree of freedom (Fig. B.3**b**).

The single-feature DII values in Fig. 4.4**c** (about 0.70 for S_2/S_1 and about 0.85 for S_1/S_0) further support this asymmetry: one dominant coordinate explains the upper-state decay significantly better than it explains the lower-state decay. As additional coordinates are included, the stronger drop for $S_2 \rightarrow S_1$ and the more modest improvement for $S_1 \rightarrow S_0$ indicate that the latter process is intrinsically more collective. This is consistent with previous mechanistic analyses reporting that access to the S_1/S_0 seam involves coupled torsion, C−N stretching, and pyramidalization-type distortions of the CH_2 group [25, 245].. In summary, these findings demonstrate that our protocol provides a semi-automated manner for diagnosing relevant photochemical decay pathways in a well studied system involving non-adiabatic hops associated with three electronic states.

This system demonstrates that the protocol can identify informative coordinates for two different state pairs, S_2/S_1 and S_1/S_0, by analyzing their energy gaps separately. The preceding analysis demonstrates that we are able to resolve mechanisms and modes associated with two different conical intersections. Extension to more complex multistate systems requires adequate sampling of each relevant transition and pathway. Separate analyses of identifiable trajectory subsets could help distinguish competing channels, but the resolving power of this approach in systems with densely spaced electronic states remains to be assessed.

4.3.3 L-glutamine

The preceding two systems demonstrate that our protocol can identify non-radiative decay modes in well-studied prototypical molecules with different numbers of electronic states and active nuclear degrees of freedom. We next turn to amino-acid-based systems associated with NAF, beginning with L-glutamine (L-gln). Experimentally,

L-gln is non-fluorescent, consistent with efficient non-radiative relaxation back to the ground state on a short timescale. Motivated by this connection between structural constraints, and different non-radiative decay mechanisms and fluorescence properties, we analyze not only the energy gap $\Delta E_{1,0}$ but also the oscillator strength $f_{1,0}^{\text{osc}}$ along the trajectories. This allows us to determine if the oscillator strength can be explained accurately by simple structural coordinates.

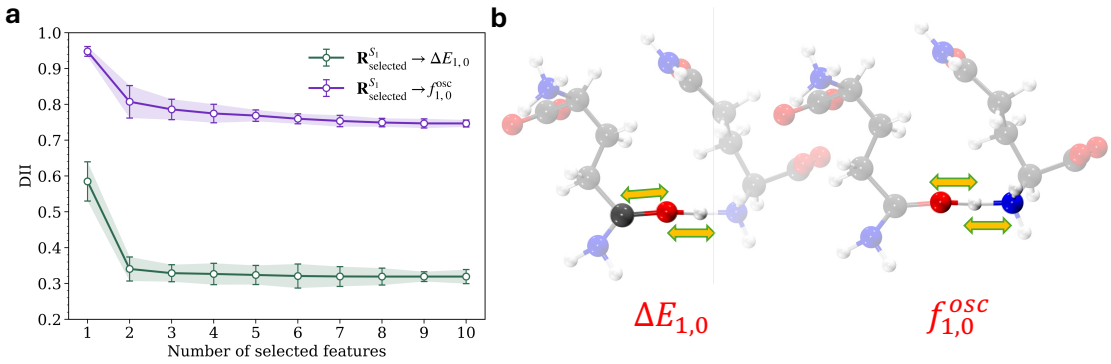


Figure 4.5: *(a) DII between internal coordinates and energy gap/oscillator strength as a function of the number of selected features. Purple corresponds to the $f_{1,0}^{\text{osc}}$ and green corresponds to the $\Delta E_{1,0}$ respectively. Shaded areas show error bars computed from 50 independent bootstrap resamples. (b) The most informative structural modes extracted from the DII analysis: The $\Delta E_{1,0}$ is governed primarily by the carbonyl C−O stretch and an O−H distance (left), while the $f_{1,0}^{\text{osc}}$ is described by the O−H and H−N stretching (right).*

In Fig. B.4**a** we compare DII trends for two targets, $\Delta E_{1,0}$ (green) and $f_{1,0}^{\text{osc}}$ (purple), using selected Coulomb-matrix (CM) elements. The contrast is striking: for the energy gap, DII drops from about 0.58 to about 0.32 with only two selected features and then plateaus, whereas for oscillator strength it remains much higher (about 0.95 to about 0.72 over 10 features). Fig. B.4**b** shows the corresponding key atom pairs: for $\Delta E_{1,0}$, the dominant signal involves the carbonyl atoms plus one neighboring NH_3^+ hydrogen; for $f_{1,0}^{\text{osc}}$, the carbonyl oxygen and all three NH_3^+ hydrogens are most implicated.

The same hierarchy appears with chemically interpretable internal coordinates. In Fig. 4.5**a** we display the DII results for both $\Delta E_{1,0}$ (green) and $f_{1,0}^{\text{osc}}$ (purple) using the internal coordinates. Fig. 4.5**b** shows that the top two descriptors are the carbonyl C−O stretch and an O−H distance associated with proton transfer; together they already recover the low-DII regime (~ 0.32). In fact, the single feature associated with the C−O stretch already yields a DII of ~ 0.58, indicating that this coordinate alone captures a substantial fraction of the structural information relevant to the energy gap. For the oscillator strength, the best descriptors are proton-transfer-related distances (O−H and H−N), but the DII stays near ~ 0.76, indicating weaker direct structural predictability than for the energy gap. This is consistent with previous L-gln analyses, where non-radiative decay is dominated by localized carbonyl/proton-transfer motion [38, 6].

The analysis of the oscillator strength is particularly noteworthy, as it highlights a distinct behavior compared to the energy gap. Specifically, the DII values remain significantly higher, indicating that the oscillator strength is governed by more com-

plex relationships than those captured by the simple structural modes considered here. Nevertheless, moderate correlations are still observed. Importantly, the modes that are most informative for describing the energy gap are not necessarily identical to those relevant for f^{osc}.

4.3.4 L-pyroglutamine-ammonium

In contrast to L-gln, L-pyroglutamine-ammonium (L-pyro) corresponds to the fluorescent product formed after chemical transformation of L-gln upon heating in water, as shown experimentally in previous work [37]. Subsequent nonadiabatic dynamics studies further showed that the short-hydrogen bonds in L-pyro suppress access to the dominant non-radiative decay pathway by restricting the carbonyl-elongation mode and associated proton-transfer motion [38, 6]. As a result, only about 20% of trajectories in the present ensemble undergo non-radiative relaxation to the ground state. This makes L-pyro a challenging case of a system with a limited amount of data because the analysis must recover meaningful decay coordinates from a limited number of reactive events.

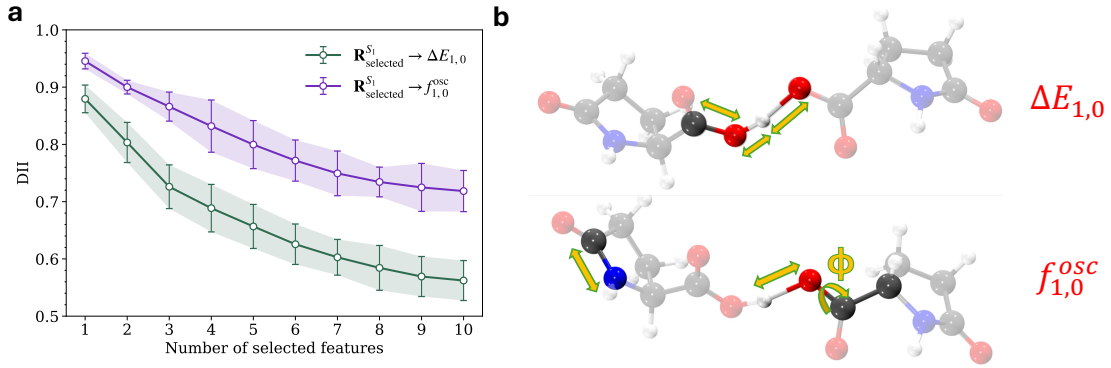


Figure 4.6: *(a) DII between internal coordinates and energy gap/oscillator strength as a function of the number of selected features. Purple corresponds to the $f^{\mathrm{osc}}_{1,0}$ and green corresponds to the $\Delta E_{1,0}$ respectively. Shaded areas show error bars computed from 50 independent bootstrap resamples. (b) The most informative structural modes extracted from the DII analysis: The $\Delta E_{1,0}$ is governed primarily by the carbonyl C−O stretch and the proton transfer involving the two O−H distances (top), while the $f^{\mathrm{osc}}_{1,0}$ is described by a O−H stretch, a C−C−O−H dihedral, and a C−N ring stretch (bottom).*

In Fig. B.6**a** we analyze DII trends for both $\Delta E_{1,0}$ and $f^{\mathrm{osc}}_{1,0}$ using selected Coulomb-matrix (CM) features. Both curves decrease gradually, indicating pronounced collectivity. The energy gap remains more predictable than oscillator strength (DII approaches ~ 0.50 for $\Delta E_{1,0}$ versus ~ 0.66 for $f^{\mathrm{osc}}_{1,0}$), again showing that pairwise structural descriptors are less effective for optical observables. Fig. B.6**b** highlights atoms from the top five CM pairs; these are concentrated around carbonyl C−O groups, the inter-dimer hydrogen-bond region, and the carbon atom adjacent to the carbonyl. Similar modes were identified by work done in our group in Ref. [38].

The internal-coordinate analysis in Fig. 4.6**a** gives a consistent picture. For $\Delta E_{1,0}$, the leading coordinates are the carbonyl C−O stretch and the proton-transfer coordinate involving the two O−H bonds (top panel of Fig. 4.6**b**). For $f^{\mathrm{osc}}_{1,0}$, the most relevant

features shift toward proton-transfer and nearby torsional motions: an O−H coordinate, a C−C−O−H dihedral, and again a C−N ring stretch (bottom panel of Fig. 4.6**b**). The slow DII saturation that we see for both $\Delta E_{1,0}$ and $f_{1,0}^{\text{osc}}$) in Fig. 4.6**a** confirms that both observables are governed by coupled distortions rather than by a single dominant mode.

As in the case of L-gln, the modes governing f^{osc} only partially overlap with those correlated with the energy gap. In particular, the O−H stretching mode associated with proton transfer is common to both. In contrast, the C−N ring stretch and the C−C−O−H dihedral are uniquely relevant for f^{osc} and do not appear among the dominant modes describing the energy gap. This further reinforces the non-trivial observation that modes strongly correlated with the oscillator strength are not necessarily those that most effectively explain the energy gap.

We note that the oscillator strength ($f_{1,0}^{\text{osc}}$) depends on both the energy gap ($\Delta E_{1,0}$) and the transition dipole moment (TDM). Given the partial overlap between the coordinates selected for the energy gap and those for the oscillator strength, it is natural to question whether this overlap arises primarily from the explicit dependence of f^{osc} on ΔE. To assess this, we examined DII($\mathbf{R}^{S_1}_{\text{selected}} \rightarrow f_{1,0}^{\text{osc}}(\text{TDM})$), where $f_{1,0}^{\text{osc}}(\text{TDM})$ denotes the oscillator strength constructed by retaining only the TDM contribution, excluding the energy gap term. For both L-gln and L-pyro, the resulting DII values are nearly identical to those obtained using the full oscillator strength (see Fig. B.5 and Fig. B.7). In particular, the absolute magnitudes differ by less than 0.02–0.03, the trends across the DII curves are also consistent, and the dominant physical modes remain unchanged relative to those identified from DII($\mathbf{R}^{S_1}_{\text{selected}} \rightarrow f_{1,0}^{\text{osc}}$). We will see in the next section however that this is not necessarily true in general.

4.3.5 Molecular Motors

Next, we analyze the overcrowded molecular motor [255], a light-activated system where photoexcitation initiates rotation around a central double-bond region [250].

This motor is also a useful test case because the key early-time photophysics combines two coupled observables: non-radiative relaxation and a bright-to-dark transition in S_1 (drop in oscillator strength) before ground-state recovery [248]. Previous studies implicate central-bond torsion, bond-stretching and pyramidalization in this process, [251, 252, 7] but disagree on which motion is the main driver [248, 252, 249]. Our goal here is to identify the most informative coordinate directly from trajectory data.

In Fig. B.9**a**, DII computed from selected Coulomb-matrix (CM) features decreases rapidly for both targets, reaching about ~ 0.27 for $\Delta E_{1,0}$ and ~ 0.35 for $f_{1,0}^{\text{osc}}$. The most informative CM pairs are highlighted in Fig. B.9**b**: for the energy gap, {C1,C3}, {C1,C5}, and {C2,C6}; for oscillator strength, {C2,C4}, {C1,C3}, and {C1,C5}. These pairs mainly track atoms whose relative geometry changes strongly along the central-bond rotation pathway.

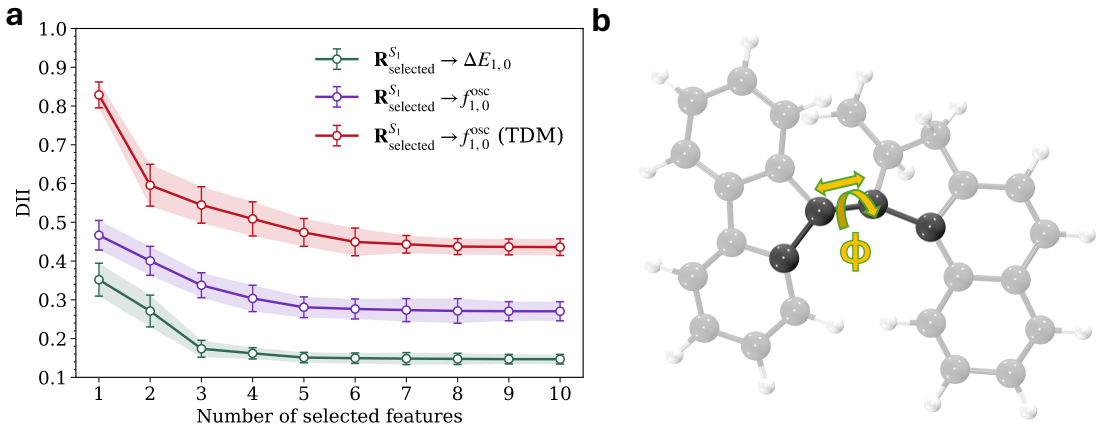


Figure 4.7: *(a) DII between internal coordinates and energy gap/oscillator strength as a function of the number of selected features. Purple corresponds to the $f^{\text{osc}}_{1,0}$, red corresponds to $f^{\text{osc}}_{1,0}$(TDM), and green corresponds to the $\Delta E_{1,0}$ respectively. Shaded areas show error bars computed from 50 independent bootstrap resamples. (b) The most informative structural modes extracted from the DII analysis: both the $\Delta E_{1,0}$ and the $f^{\text{osc}}_{1,0}$ is governed primarily by the central torsion dihedral and the central double bond stretch that are highlighted.*

We subsequently map this signal to interpretable internal coordinates. As shown in Fig. 4.7**a** with only three coordinates, DII already reaches very low values (about 0.17 for $\Delta E_{1,0}$ (green) and about 0.33 for $f^{\text{osc}}_{1,0}$ (purple)), indicating that a compact coordinate set captures most of the relevant dynamics. The leading coordinates are the central torsional dihedral and the associated double-bond stretch (Fig. 4.7**b**), with an additional smaller contribution from a second nearby dihedral (see Fig. B.10).

Quantitatively, the primary dihedral alone yields DII values of approximately 0.35 for $\Delta E_{1,0}$ and 0.46 for $f^{\text{osc}}_{1,0}$, making it the single most informative coordinate for both non-radiative decay and bright-to-dark conversion. As in previous cases, however, the correlation between structural coordinates and the oscillator strength remains weaker, as reflected by the higher DII value. While the central dihedral captures the behavior of the energy gap nearly optimally, its correlation with the oscillator strength is comparatively less precise. The next most relevant contributions arise from the central double-bond stretch and an additional dihedral around the same region (see Fig. B.10), although their influence is clearly less significant. In contrast to the L-gln and L-pyro systems, the modes governing the energy gap and oscillator strength in this system coincide. In both cases, the central torsional dihedral and the associated double-bond stretch emerge as the most informative coordinates, representing a notable scenario in which structural modes exhibit a relatively strong correlation with the oscillator strength. Taken together, these results support a mechanism in which torsional motion about the central bond is the most informative coordinate.

This system also highlights a practical strength of the protocol. The CM-filtering step alone does not explicitly return all atoms of the most informative dihedral, yet the subsequent internal-coordinate ranking resolves the dominant mechanistic mode unambiguously. In other words, the multi-step DII procedure can disentangle coupled photochemical coordinates even when the first feature-selection stage is not directly

interpretable in terms of a single geometrical mode.

The strong overlap between the modes governing the energy gap and the oscillator strength that we observed in the previous section motivated a more detailed analysis of the contribution arising solely from the transition dipole moment (TDM), i.e., $f_{1,0}^{\mathrm{osc}}(\mathrm{TDM})$. As before, we compute $\mathrm{DII}(\mathbf{R}^{S_1}_{\mathrm{selected}} \rightarrow f_{1,0}^{\mathrm{osc}}(\mathrm{TDM}))$, shown as the red curve in Fig. 4.7**a**. In contrast to the behavior observed for L-gln and L-pyro, the DII in this case saturates at a significantly higher value (~ 0.44) compared to that obtained using the full oscillator strength (purple curve), which reaches ~ 0.27. Similarly, the top three coordinates yield a DII of only ~ 0.54 for $f_{1,0}^{\mathrm{osc}}(\mathrm{TDM})$, as opposed to ~ 0.33 for the full $f_{1,0}^{\mathrm{osc}}$. Notably, these top three coordinates are identical to those identified previously, and appear in the same order; however, their ability to describe the TDM-only contribution is markedly reduced relative to their performance for the full oscillator strength. This behavior contrasts quite significantly with the L-gln and L-pyro systems, where $f_{1,0}^{\mathrm{osc}}(\mathrm{TDM})$ closely mirrors the behavior of $f_{1,0}^{\mathrm{osc}}$. The present results therefore indicate that, in this molecular motor system, the energy gap plays a more dominant role in determining the oscillator strength than in the amino acid systems considered earlier. More generally, repeating our protocol with other types of electronic observables coming from population analysis in quantum chemistry [178, 179, 180, 181, 256] where the electron density is partitioned into atomic contributions, could be interesting to explore in the future within this context.

4.3.6 Non-Radiative Decay Pathways Along Principal Components

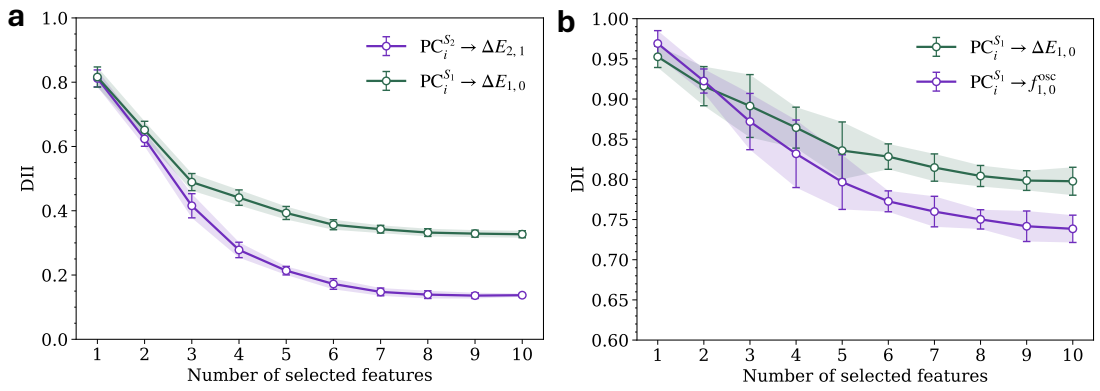


Figure 4.8: (*a*) *Methaniminium Cation: DII between the leading PCs and energy gaps as a function of the number of selected features. Purple corresponds to the* $S_2 \rightarrow S_1$ *and green corresponds to the* $S_1 \rightarrow S_0$ *relaxation channels respectively.* (*b*) *L-gln: DII between the leading PCs and energy gap/oscillator strength as a function of the number of selected features. Purple corresponds to the* $f_{1,0}^{\mathrm{osc}}$ *and green corresponds to the* $\Delta E_{1,0}$ *respectively. Shaded areas for both panels show error bars computed from 50 independent bootstrap resamples.*

The internal-coordinate analysis presented above starts from a broad structural representation and systematically selects the coordinates that are most informative of the target electronic observables. In some of the systems we studied (L-gln and Molecular Motors), this procedure leads to a rapid decrease of the DII, indicating that a

small number of localized coordinates is sufficient to capture most of the relevant structural information. However, in other systems (Furan and L-pyro), the DII decreases more gradually as additional coordinates are included. This behavior suggests that the changes in the energy gap may depend on a more distributed nuclear distortion rather than on a small set of localized bonds, angles, or dihedrals.

To assess this possibility, we complement the internal-coordinate analysis with a principal component (PC) based representation of the trajectory ensemble. Principal components provide a linear, variance-maximizing basis for nuclear motion and therefore offer a compact description of correlated, delocalized fluctuations in configuration space [257]. We therefore compute the DII between PC coordinates and the relevant electronic observables across all systems, using PCA not as an alternative to the chemically interpretable coordinate selection protocol, but as a reference representation for evaluating the extent to which the observables are controlled by collective nuclear motion. Compared to the final step of the protocol where we use internal coordinates, when using PCs we instead compute $\mathrm{DII}(\mathrm{PC}_i^{S_l} \rightarrow \Delta E_{l,l-1})$ and $\mathrm{DII}(\mathrm{PC}_i^{S_l} \rightarrow f_{l,l-1}^{\mathrm{osc}})$ for each PC_i, to determine the PCs that contribute most to the energy gap and the oscillator strength respectively.

For the methaniminium cation, projecting the dynamics onto PCs substantially enhances predictability. In particular, $\mathrm{DII}(\mathrm{PC}_i^{S_2} \rightarrow \Delta E_{2,1})$ and $\mathrm{DII}(\mathrm{PC}_i^{S_1} \rightarrow \Delta E_{1,0})$ decrease rapidly with the number of retained components (Fig. 4.8**a**) , with the first ~10 PCs nearly saturating the predictability of the S_2/S_1 gap. A similar trend is observed for $\Delta E_{1,0}$, where the DII is reduced to ~ 0.32, significantly below the ~ 0.60 obtained using the top internal coordinates. This demonstrates that collective combinations of motions encode a large fraction of the information relevant to nonadiabatic relaxation, consistent with previous observations that dimensionality reduction techniques can recover dominant reaction pathways directly from trajectory data without prior mechanistic assumptions [223, 225]. Similar trends are observed in Furan (see caption of Fig. B.2 for a more detailed summary).

In contrast, the amino-acid systems reveal a qualitatively different regime. As shown in Fig. 4.8**b**, for L-gln the PC representation performs significantly worse for the energy gap $\Delta E_{1,0}$, with the DII saturating near ~ 0.80, compared to ~ 0.32 obtained using only two interpretable internal coordinates. This clearly indicates that the dominant relaxation coordinate is highly localized, and that its signal is diluted when embedded into delocalized PC directions. For the oscillator strength $f_{1,0}^{\mathrm{osc}}$, PCs provide only a marginal improvement (~ 0.73 versus ~ 0.76), suggesting that this observable does not benefit strongly from a purely variance-based representation. This breakdown of PCA highlights a known limitation: while PCs optimally capture structural variance, they do not necessarily align with the physically relevant directions governing electronic transitions. In L-pyro (Fig. B.8), PCs noticeably improve the mapping to the oscillator strength (DII decreases from ~ 0.72 to ~ 0.63), however, the improvement for the energy gap is minimal (~ 0.56 to ~ 0.54).

Finally, for the molecular motors (Fig. B.11), PC-based descriptors achieve low DII

values (~ 0.29 for the energy gap and ~ 0.33 for oscillator strength), confirming that collective motions are highly informative in these larger, more flexible systems. These values however, remain slightly higher than those obtained using targeted internal coordinates, again pointing to the presence of specific mechanistic directions that are not optimally aligned with variance-maximizing modes.

Taken together, these results establish a clear dichotomy. Principal components are highly effective at compressing high-dimensional trajectory data and can capture substantial predictive information, particularly in systems where the dynamics is intrinsically collective. This is consistent with their widespread use in analyzing nonadiabatic dynamics, where they can reveal dominant motion patterns without prior knowledge [56, 57]. However, because PCs are constructed solely based on structural variance, they may fail to isolate the specific nuclear distortions that control electronic transitions, especially when these are localized or weakly represented in the global variance [55]. In contrast, the DII-selected internal coordinates retain direct physical interpretability while achieving comparable, or in several cases superior predictive power, highlighting their suitability for mechanistic analysis of nonadiabatic processes.

4.4 Discussion

Across all systems studied here, the information-imbalance workflow provides a consistent strategy to move from high-dimensional trajectory data to a compact, physically interpretable set of decay coordinates. The main practical outcome is that we can rank candidate descriptors by information content and then map those rankings to mechanistic modes, rather than imposing a reaction coordinate *a priori*. This is especially useful for excited-state dynamics, where multiple coupled distortions can coexist and where local CI descriptors are not always straightforward to transfer to dynamical ensembles.

The physical reliability of the modes identified by our protocol is naturally conditioned on the quality of the excited-state trajectories and the underlying electronic-structure description. Linear-response TDDFT and TDDFTB have known limitations in describing the multireference character and topology of the S_1/S_0 intersection. Our analysis considers structural evolution along the approach to the intersection, including configurations with larger energy gaps, rather than exclusively the immediate degeneracy region. For Furan [243, 244] and the Methaniminium cation [25, 245, 246, 247], the selected modes agree with established mechanisms obtained using different electronic-structure methods. For the amino-acid systems, earlier multireference calculations and conical-intersection optimization provide complementary support for the carbonyl-elongation and proton-transfer mechanisms recovered here [38].

To assess the sensitivity to the electronic-structure method directly, we repeated the L-glutamine analysis using TD-DFT trajectories from Ref. [38]. Both TD-DFT and TD-DFTB recover the same leading structural modes, including the carbonyl C–O stretch and the proton-transfer-related O–H distance (Figure B.12). The numerical DII curves

nevertheless show differences: TDDFTB reaches approximately 0.34 with two selected features and subsequently plateaus near 0.32, whereas TDDFT exhibits a more gradual decrease, reaching approximately 0.37 with ten features. Thus, the same physical coordinates remain informative across the two levels of theory, although their combined ability to describe the energy gap differs. Together with the earlier multireference evidence, this comparison supports the physical interpretation of the selected modes. More generally, the framework provides a means of comparing structural signatures across different simulation methods.

Several system-level trends emerge from the results. First, the method distinguishes localized from collective decay channels in a chemically transparent way. Methaniminium shows a relatively localized dominant coordinate for S_2/S_1 but a more collective S_1/S_0 channel, while furan displays broad collectivity with coupled puckering and ring-opening coordinates. In L-gln, the key energy-gap signal remains strongly localized (carbonyl stretching/proton-transfer coordinates), whereas in L-pyro the same observables are more distributed and require a broader coordinate set although the dominant modes are still the same carbonyl stretching and proton-transfer coordinates. For molecular motors, the analysis identifies torsion around the central bond as the most informative mode governing both non-radiative decay and bright-to-dark conversion.

The molecular-motor case highlights how the workflow can clarify competing mechanistic interpretations in strongly coupled photochemical systems. Prior work has debated whether bright-state quenching is primarily associated with pyramidalization or with central-bond torsional/elongation coordinates [248, 252, 249]. Here, the DII analysis identifies the central torsional coordinate as the most informative descriptor for both $\Delta E_{1,0}$ and $f_{1,0}^{\mathrm{osc}}$, while the central bond stretch and pyramidalization appear as secondary, coupled responses. Thus, the analysis supports a picture in which torsional motion around the central bond acts as the most informative coordinate that explains both energy-gap closure and the change of the oscillator strength. Equally important, this mode is recovered even when the initial CM-filtering stage does not directly isolate all dihedral atoms, highlighting the added mechanistic value of the full multi-step II workflow for strongly coupled photochemical pathways.

Second, the analysis reveals a systematic asymmetry between electronic targets: the energy gap is generally easier to explain than the oscillator strength using simple interpretable structural descriptors. This distinction is physically expected because the oscillator strength contains both an energetic and a transition-dipole moment contribution. To disentangle these two effects, we additionally analyzed the transition-dipole contribution alone. This decomposition shows an interesting duality: in the case of L-gln and L-pyro we observe that both the extent to which the internal coordinates describe f^{osc} and $f^{\mathrm{osc}}(\mathrm{TDM})$ (as reflected by the value of the DII), as well as the dominant coordinates involved in the description remain consistent. Thus, in these systems, the transition dipole itself is controlled by structural changes that are less localized than the modes that govern the energy gap.

On the other hand, the molecular motor provides a contrasting case: the same central torsional coordinate governs both ΔE and f^{osc}, and the TDM-only analysis indicates that the bright-to-dark conversion is also organized by the same coupled torsional/bond-stretching motion. However, in this case, the explainability of $f^{\mathrm{osc}}(\mathrm{TDM})$ is substantially reduced relative to f^{osc}. This suggests that the stronger structural predictability of f^{osc} arises primarily from the explicit contribution of ΔE. Thus, while the same structural features also influence the TDM contribution, their correlation with $f^{\mathrm{osc}}(\mathrm{TDM})$ is significantly weaker than with the full oscillator strength. Overall, these results show that the coordinate controlling energy gap closure is not generally identical to the coordinate controlling optical activity, although the two can coincide in systems where the same structural distortion controls both non-radiative decay and electronic-state character.

This approach is also data-efficient. The Furan ensemble-size tests (Fig B.14) demonstrate that the same leading decay coordinates can be recovered using 20 rather than 200 trajectories, although the absolute DII values remain ensemble-size dependent. This supports the feasibility of identifying informative modes from reduced trajectory datasets. In L-pyroglutamine-ammonium, the analysis also recovers physically meaningful coordinates despite only approximately 20% of trajectories undergoing non-radiative relaxation. The amount of data required nevertheless depends on the system and the sampling of its relevant decay pathways. This is relevant for realistic NAMD simulations, where computational cost can restrict analysis to a few thousand representative frames (often only a subset of all generated configurations). Even in these data-limited settings, the internal-coordinate ranking remains robust enough to extract dominant channels.

Principal component (PC) descriptors are useful for identifying collective structural distortions in NAMD simulations that explain the energy gap. However, they are not universally optimal for predicting electronic observables. In cases where the relevant photochemical process is localized, explicit internal coordinates can provide superior predictive performance. Furthermore, the most informative PCs are not necessarily those that capture the largest structural variance, and their relevance must therefore be evaluated explicitly. In this sense, PC analysis should be viewed as a diagnostic tool for detecting collective motions that correlate with electronic observables (such as energy gaps or oscillator strengths), rather than as a substitute for physically interpretable internal coordinates. The II-based approach can also be used to determine the most informative principal components in predicting the energy gap which could then in turn be used to create low-dimensional representations of systems where the nuclear motion is also described quantum mechanically, such as in multiconfigurational time-dependent Hartree (MCTDH) with linear vibronic coupling (LVC) models. This is an extension of the approach done previously in Ref. [55] where normal modes were selected based on a linear fit to energy gaps.

Another interesting class of systems where our techniques could be applied is on photodissociation reactions. In Appendix B.1.1 we show that the protocol devel-

oped here recovers the relevant dissociation coordinates from the previously published QM/MM trajectories of crystals of the amino acid Cysteine (Chapter 3). The method identifies both the relevant coordinates of the dissociation of the proton from the sulphur group (S−H bond) and the cleavage of the C−N bond.

In conclusion, the II-based protocol offers a unified and scalable route to identify non-radiative decay modes directly from trajectory ensembles, while retaining mechanistic interpretability across systems ranging from small benchmarks to complex fluorescent materials and molecular motors.

5

Towards a robust approach to infer causality from molecular dynamics simulations

The ability to distinguish between correlation and causation of variables in molecular systems remains an interesting and open area of investigation. In this work, we probe causality in a molecular system using two independent computational methods that infer the causal direction through the language of information transfer. Specifically, we demonstrate that a molecular dynamics simulation involving a single Tryptophan in liquid water displays asymmetric information transfer between specific collective variables, such as solute and solvent coordinates. Analyzing a discrete Markov-state and Langevin dynamics on a 2D free energy surface, we show that the same kind of asymmetries can emerge even in extremely simple systems, undergoing equilibrium and time-reversible dynamics. We use these model systems to rationalize the unidirectional information transfer in the molecular system in terms of asymmetries in the underlying free energy landscape and/or relaxation dynamics of the relevant coordinates. Finally, we propose a computational experiment that allows one to decide if an asymmetric information transfer between two variables corresponds to a genuine causal link.

5.1 Introduction

One of the most intriguing foundational questions in molecular science concerns how cause-and-effect relationships, plainly observed in our mesoscopic and macroscopic world, emerge from dynamic equations that are time-reversible at the microscopic scale. Measuring causality in a system described by classical equations of motion is highly non-trivial, and has been the object of intense investigation [64, 65, 66, 67, 258,

68, 259, 69, 70, 71, 260, 72, 261, 262]. For this purpose, molecular dynamics (MD) simulations offer the possibility to interrogate specific microscopic degrees of freedom that are often unattainable or very challenging to observe and manipulate in experimental approaches. In order to infer causality among Collective Variables (CVs) of interest, such as inter-residue distances [71] or dihedral angles [72, 258] in small proteins, various studies have previously analyzed MD simulations using Granger causality (GC) [62, 263], Transfer Entropy (TE) [63, 134] or time-lagged two-body cross-correlation functions (CCFs) [72, 66].

In other contexts, such as medical studies [264], sociology [265], and epidemiology [266], causal questions have been addressed for decades through the lens of causal inference [267, 268, 269]. This field provides a rigorous statistical framework to answer counterfactual questions such as "Had the value of variable X been different, would the value of Y have been different as well?". The common strategy to address this question from observational time series, namely in the absence of ad-hoc manipulations of the putative causal variable X, is measuring conditional dependencies between pairs of variables at different times [270]. Informally, if a variable Y at time τ depends on a variable X at time zero for all possible conditioning sets including the observed variables up to time τ, one infers that X causes Y. Crucially, this conclusion can be drawn only if no unobserved common cause of X and Y exists, or if all common causes of X and Y are included in the search space of conditioning sets. A common cause of two variables is typically referred to as "confounder", and the hypothesis that all confounders are observed is referred to as "causal sufficiency". While the most general algorithms to infer the causal graph rely on iterative conditional independence tests for each pair of variables [271, 272, 273], an alternative approach in the case of time series is to compute the Transfer Entropy (TE), which is equivalent to carry out a single conditional independence test for each pair of variables [270]. In the following, we will refer to any measure of conditional (in)dependence in observational time series data, such as TE, as *information transfer*.

If the existence of unobserved common drivers cannot be ruled out, one may assess the existence of a causal relationship $X \rightarrow Y$ by measuring the average or distributional changes in Y resulting from two or more manipulations, or interventions, over X. A (hard) intervention on $X(0)$, denoted as $do(X(0) = x)$, is an ideal experiment where the value of X at time zero is set to x independently of the value of any other variable, observed or not, that is not caused by $X(0)$. Given two independent interventions $do\,(X(0) = x)$ and $do\,(X(0) = x')$, the causal effect of $X(0)$ on $Y(\tau)$ can be measured from the difference between the post-interventional distributions $p(Y(\tau)|do(X(0) = x))$ and $p(Y(\tau)|do(X(0) = x'))$. Importantly, this interventional approach not only allows to formulate causal statements when unobserved common drivers are present, but also provides a direct quantification of causal effects [274]. Pearl's "do-calculus" [267] provides tools to compute post-interventional distributions from observational data, under the assumption of causal sufficiency.

In this work we investigate the emergence of *strongly asymmetric* information trans-

fers, which indicate candidate *unidirectional* causal relationships, in molecular systems where the microscopic interactions are bidirectional due to Newton's third law. In particular:

- Using a very simple molecular system, a Tryptophan (TRP) molecule solvated in water, we show that unidirectional information transfers between one-dimensional CVs can be inferred by measuring the Transfer Entropy [64, 65, 66, 67, 68, 259, 261, 69, 70, 262, 258], or using an approach introduced recently [15] when the CVs are high-dimensional (Sec. 5.3.1).
- We show that such unidirectional information transfer can be present even in model systems which rigorously obey a time-reversible dynamics with stationary probability measure, such as a discrete-time Markov process, for which the Transfer Entropy can be computed analytically, and a Langevin dynamics on a two-dimensional potential energy surface (Sec. 5.3.2).
- We show that *if all the variables are observed*, these asymmetries allow predicting the effect of suitable interventional experiments, for example a $do(X)$, in the language of causal inference. However, the presence of a unidirectional information transfer is not a sufficient condition to decide if a causal relationship exists: in a Langevin model with three variables *in which only two are observed*, we measure an asymmetric Transfer Entropy that does not correspond to a causal relationship. We show that this can be revealed by appropriate interventional experiments (Sec. 5.3.3).

5.2 Methods

We search for unidirectional information transfer between pairs of variables by estimating the Transfer Entropy and, for high-dimensional collective variables, by using the Imbalance Gain, an approach developed recently [15]. These measures are defined in Chapter 2: the Transfer Entropy and its estimators in Section 2.6, the Information Imbalance in Section 2.5, and the Imbalance Gain in Section 2.7. This section describes the model and molecular systems to which they are applied.

5.2.1 Langevin dynamics of the model systems

The free energy landscapes depicted in Figs. 5.4**d** and 5.4**g** were sampled with an overdamped Langevin dynamics,

$$\dot{X}(t) = -\frac{1}{\gamma_X}\frac{\partial V(X,Y)}{\partial X} + \sqrt{\frac{2k_BT}{\gamma_X}}\,\eta_X(t)\,, \tag{5.1a}$$

$$\dot{Y}(t) = -\frac{1}{\gamma_Y}\frac{\partial V(X,Y)}{\partial Y} + \sqrt{\frac{2k_BT}{\gamma_Y}}\,\eta_Y(t)\,, \tag{5.1b}$$

where $V(X,Y)$ is the free energy of the system, γ_X (γ_Y) is the friction coefficient for X (Y), T is the temperature, k_B denotes the Boltzmann constant and $\eta_X(t)$, $\eta_Y(t)$ are independent noise terms satisfying $\langle \eta_X(t)\,\eta_X(t')\rangle = 2\gamma_X\,k_B\,T\delta(t-t')$ (similarly for $\eta_Y(t)$). In both cases, we set $k_B\,T = 1$. In Fig. 5.4**d**, the friction coefficients were both set to 1, while in Fig. 5.4**g** we set $\gamma_X = 1$ and $\gamma_Y = 0.1$, enforcing a slower diffusion in the X direction.

We integrated Eqs. 5.1 with a simple Euler scheme for $N = 2500$ independent initial conditions $(X(0), Y(0))$, with $X(0) \sim \mathcal{N}(0,1)$ and $Y(0) \sim \mathcal{N}(0,1)$. A time step $dt = 0.01$ was employed and the state was saved every 10 steps. From each trajectory, the initial 1000 time points were excluded from the analysis.

5.2.2 Molecular dynamics simulations

We performed molecular dynamics simulation of a Tryptophan molecule in water solution. Tryptophan (TRP) is a fluorescent amino acid which serves as a model molecular system to investigate the coupling and driving of different solvent and solute degrees of freedom. It is also an ideal starting point to model the effects of an external perturbation on a molecular system, particularly the protein-solvent interactions. In our case, this external perturbation is the photo-excitation of TRP. The indole group, which is the chromophore in TRP, has two closely spaced electronic excited states, $^1\mathrm{L_a}$ and $^1\mathrm{L_b}$. The lowest excited singlet state of the indole in the gas phase and non-polar solvents is $^1\mathrm{L_b}$ state, while in polar media (such as water), the $^1\mathrm{L_a}$ state is lowest energy singlet state [275, 276, 277, 278, 279]. The indole moiety can be modelled in a "classical" fashion simulating vertical electronic excitation using a protocol based on the work of Sobolewski and Domcke [278]. The ground state topology charges generated by the force field are left unaltered, and a topology to simulate the system in the excited state is generated by applying the Mulliken charge differences between $^1\mathrm{L_a}$ and $\mathrm{S_0}$ from the CASSCF calculations in Ref. [278]. The ultrafast internal conversion between $^1\mathrm{L_b}$ and $^1\mathrm{L_a}$ which occurs in a few femtoseconds in polar solvents is not modelled in this approach. This protocol is well-established and has been used in the past for modelling the indole moiety of TRP in the excited state and studying a variety of physical interactions in good agreement with experiments [280, 83, 281]. The veracity of this protocol can be further justified by analysing the dipole moment of the indole group, since this is a distinctive feature of the difference between the ground state and the 2 accessible excited states. Indeed, our ground state (GS) and excited state (ES) configuration has a dipole moment of 2.4 D and 4.8 D respectively, which is consistent with the values from Ref. [278].

The Zwitterionic form of TRP was chosen where the N-terminus was protonated as -$NH3^+$ and the C-terminus was deprotonated as -COO^-. The AMBER ff19SB protein force field [282] was used, and as per the recommendation of the AMBER developers, it was paired with the non-polarizable four-point rigid Optimal Point Charge (OPC) water model [283, 284]. Since we are interested in highly accurate protein-

solvent interactions, the OPC model was an obvious choice due to the faithful reproduction of various experimental properties of bulk water, such as density, dielectric constant, heat of vaporization, self-diffusion coefficient and viscosity [285, 286, 287]. GROMACS 2023.2 [288, 289, 290, 291, 292, 293, 294, 295] was used to perform the simulations where the system was first equilibrated in the NVT ensemble at 300 K for 10-ns using the Stochastic Velocity Rescaling thermostat [296]. Then, we run a constant pressure (NPT) simulation for a further 10-ns using the Stochastic Cell Rescaling barostat [102] at 1 atm and 300 K. Eventually, we carried out a production NVT run of 1-µs at 300 K using the Leapfrog integrator with a 2-fs time-step. Buffered Verlet lists were used to keep track of atomic neighbors. The long-range electrostatic interactions were calculated using the smooth Particle Mesh Ewald (PME) algorithm [107, 108]. A shifted Lennard-Jones potential with a cutoff value of 1.2 nm and long range dispersion corrections for energy and pressure was used. The system was solvated with 960 OPC water molecules in a cubic box of size 3.5 nm with periodic boundary conditions. The geometry of the water molecules was constrained by the SETTLE algorithm [104] while for the TRP, bonds with hydrogen atoms were constrained using LINCS [297].

The final dataset for the Imbalance Gain (IG) analysis is constructed by sampling frames every 1-ns from the 1-µs trajectories in the GS and ES separately, and using these as the starting points for 1000 short 500-ps simulations in the NVE ensemble. This ensemble is used instead of the more common NVT or NPT settings in order to emulate the conditions of the dynamical systems described in Section 5.3.2. Indeed, the dynamics in the microcanonical ensemble is rigorously time-reversible, as the Hamilton's equations satisfy a well-known time-invariance property under a change of sign of the momenta. Since a practical implementation of this dynamics is likely to violate time-reversibility for numerical and/or algorithmic reasons, as evidenced by a non-zero energy drift, we fine-tuned the simulation parameters in order to keep this drift below -5.95×10^{-5} kJ/(mol·ps) per atom. All rotations and translations were removed from the TRP from these NVE simulations, and the resultant trajectories were used for the IG and TE analysis.

Coordination numbers

The relevant coordination numbers that were calculated and are part of the IG analysis were done using a switching function of the form:

$$\sum_i \sum_j s_{ij} = \frac{1.0}{1.0 + \exp((d_{ij} - d_0)/\sigma)}, \tag{5.2}$$

where σ controls the width of the switching function, and d_0 is a cutoff distance for 2 atoms i and j. In this work, $d_0 = 0.15$ was chosen.

5.3 Results

5.3.1 Unidirectional information transfer between collective variables in a molecular system

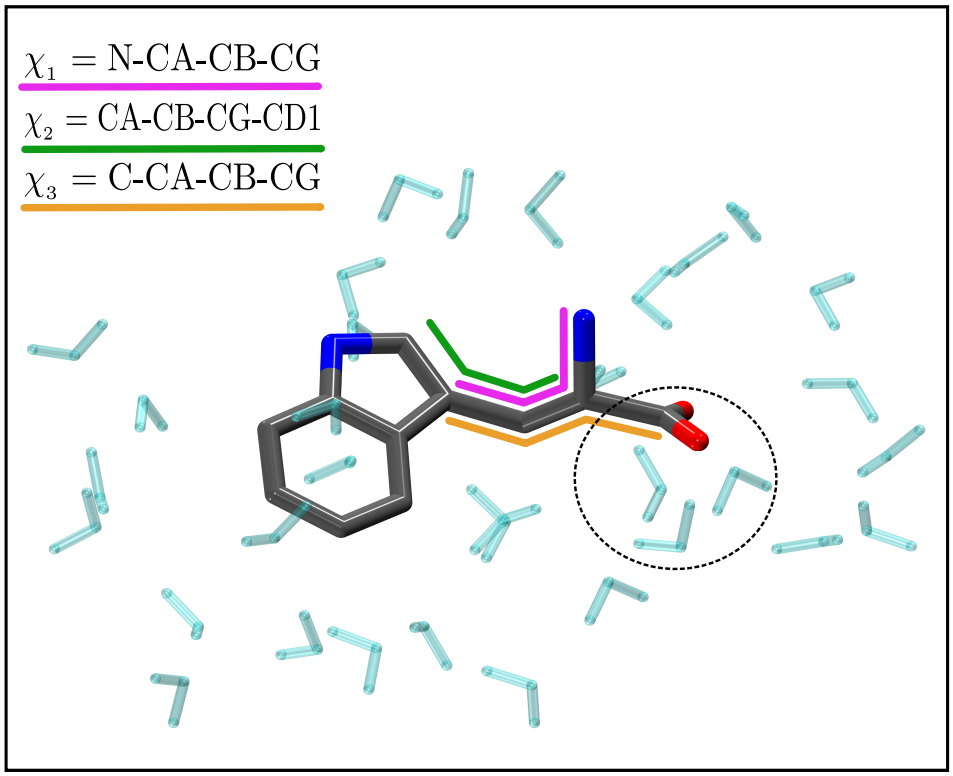


Figure 5.1: *A TRP molecule in "licorice" representation with some surrounding water molecules. The three key dihedrals, χ_1, χ_2 and χ_3, are shown in bold lines colored in pink, green, and orange respectively. The dotted circle depicts a key region for measuring coordination numbers and protein-solvent interactions.*

We first show that asymmetries in the Imbalance Gain and in the Transfer Entropy can be observed in a molecular system undergoing equilibrium and time-reversible dynamics. These asymmetries denote unidirectional information transfer between specific collective variables and, as we will see below, candidate causal links. Asymmetries in the TE have already been reported in several previous studies using molecular dynamics simulations (see for example Refs [64, 65, 66, 67, 68, 259, 261, 69, 70, 262, 258]). TE is however limited to constructing probability distributions in low-dimensions, whereas the IG was introduced precisely to overcome this limitation.

We focus on the molecular dynamics simulations of TRP in water described in Section 5.2.2. TRP is a naturally occurring fluorophore whose optical properties have been extensively studied to probe solvation dynamics - the response of protein and water coordinates following photoexcitation [298, 280, 87]. We conducted microsecond-long equilibrium molecular dynamics (MD) simulations of TRP in water on both the ground and excited electronic states (GS and ES, respectively) in order to uncover unidirectional dependencies between specific solute and solvent coordinates. Our model for

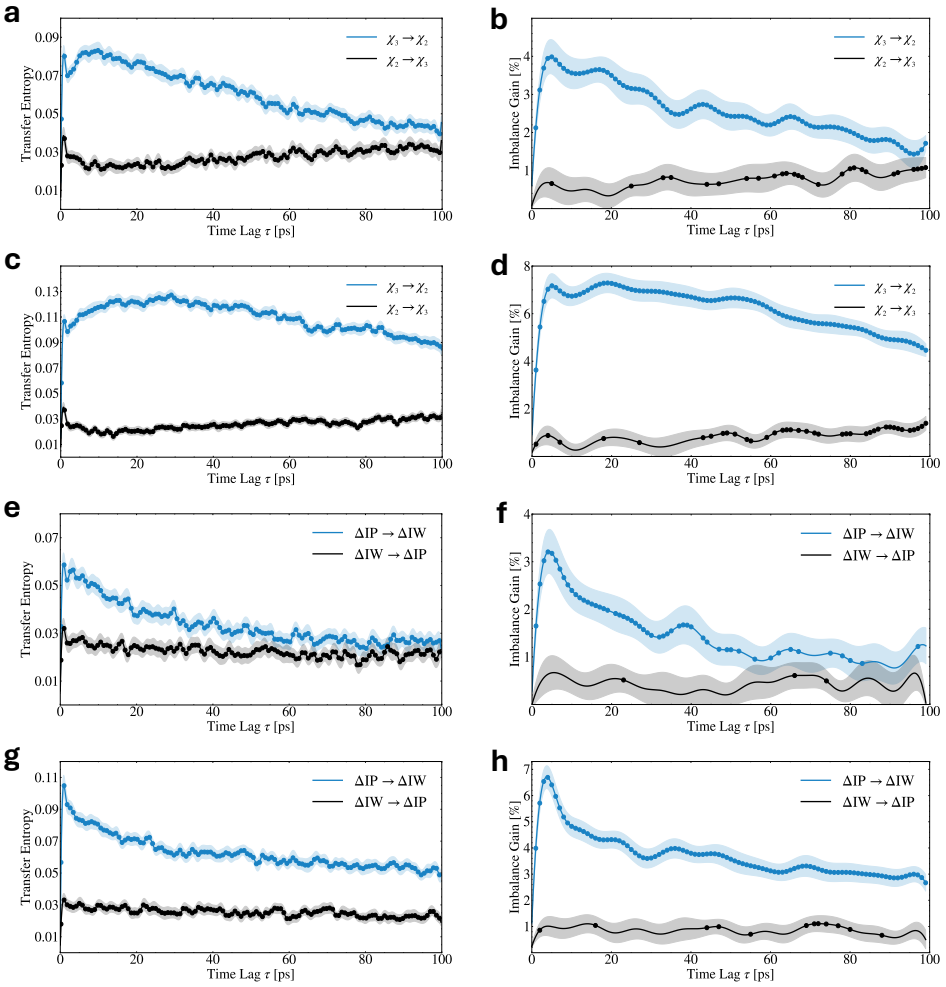


Figure 5.2: ***a*** *and* ***b****: TE and IG curves for the CVs* χ_3 *and* χ_2 *in the GS.* ***c*** *and* ***d****: TE and IG curves for the same CVs in the ES.* ***e*** *and* ***f****: TE and IG curves for the CVs* $\Delta\mathrm{IP}$ *and* $\Delta\mathrm{IW}$ *in the GS.* ***g*** *and* ***h****: TE and IG curves for the same CVs in the ES. Shaded regions denote error bars over 14 independent estimates, and the bold points denote values that are determined to be significantly different from 0 using a t-test* ($p < 0.001$)*. The IG was computed using* $k = 5$ *neighbors.*

the photoexcitation mirrors previous studies that involve adjusting the point charges in the indole group to capture the change in the magnitude and direction of the dipole moment [280, 83, 281] (see Section 5.2.2 for more details). From these simulations, we examined the relationships between a wide variety of variables that probe the coupling between the conformational changes of the TRP and the response of the surrounding water molecules. In Fig. 5.1, we show some of the relevant collective variables that we investigate. Primarily, we look at a few key dihedrals around the indole moiety. The three dihedrals we look at are χ_1, which shows the linkage of the N-terminus to the indole group, χ_2, which describes the orientation of the indole with respect to C_α, and χ_3, which informs on the linkage of the C-terminus to the indole moiety. In addition, we also consider solvation coordinates by looking at coordination numbers between the water molecules and the atoms in the system most likely involved in interactions with the water molecules: the N-terminus (n_{NT}), the C-terminus (n_{CT}), the two carbonyl Oxygens of the C-terminus (n_{O1}, n_{O2}), and the N-H group of the indole (n_{NE1}).

In addition to these structural quantities, we also examined variables that probe the changes in the interaction energy between the TRP and the environment arising from a photoexcitation. This can then be partitioned separately into contributions coming from the interactions of the chromophore (the indole moiety) with water molecules (IW) and with the peptide chain (IP). We simulated a vertical excitation of the system by replacing the ground-state charges with that of the excited state. This gives us the "interaction energy gaps" collective variables as:

$$\Delta\,\mathrm{IP} = \mathrm{IP}_{S_1} - \mathrm{IP}_{S_0}\,, \tag{5.3a}$$

$$\Delta\,\mathrm{IW} = \mathrm{IW}_{S_1} - \mathrm{IW}_{S_0}\,. \tag{5.3b}$$

The quantities $\Delta\,\mathrm{IP}$ and $\Delta\,\mathrm{IW}$ give us the contribution of the indole-peptide and indole-water interaction energies, respectively, to the total energy gap between the ground and excited state. These energy gap variables are crucial ingredients that go into understanding the microscopic origins behind Time-Dependent Fluorescence Stokes Shift experiments [299].

In Fig. 5.2 we show pairs of collective variables displaying approximately unidirectional information transfer in the TE and IG, on the timescale of the first 100 ps. Figs. 5.2**a** and 5.2**b** show the TE and IG between the χ_3 and χ_2 dihedral angles in the GS, respectively, while Figs. 5.2**c** and 5.2**d** present the same measures in the ES. For both sets of simulations, we observe a large IG in direction $\chi_3 \rightarrow \chi_2$, while the IG in the opposite direction is negligible. Similar behavior is also observed in the case of the TE. In the ES, the TE and IG from χ_3 to χ_2 decay on a much longer timescale compared to the GS; in addition, we observe that the unidirectional information transfer is more marked in the ES compared to the GS. The same behavior also holds for the CVs χ_1 and χ_2 (see Fig. C.6). This suggests that the timescales associated with the flow of information between different modes is altered in the GS versus ES.

In Fig. 5.2 (panels **e**, **f**, **g**, **h**), we perform the same analysis for the energetic variables Δ IP and Δ IW. These variables probe the total change in the electrostatic interaction energy between chromophore and peptide backbone (IP) or chromophore and water (IW), as a consequence of the excitation (see Section 5.2.2). In Figs. 5.2**e** and 5.2**f** we show the TE and IG for the energetic variables in the GS, while in Figs. 5.2**g** and 5.2**h** we report the same measures in the ES. In this case, we observe a unidirectional information transfer from Δ IP to Δ IW. Similarly to the case of dihedrals, for the energetic variables, the relaxation of both the TE and IG is significantly slowed down in the ES. It should be noted that for one-dimensional CVs the TE and the IG provide fully consistent results.

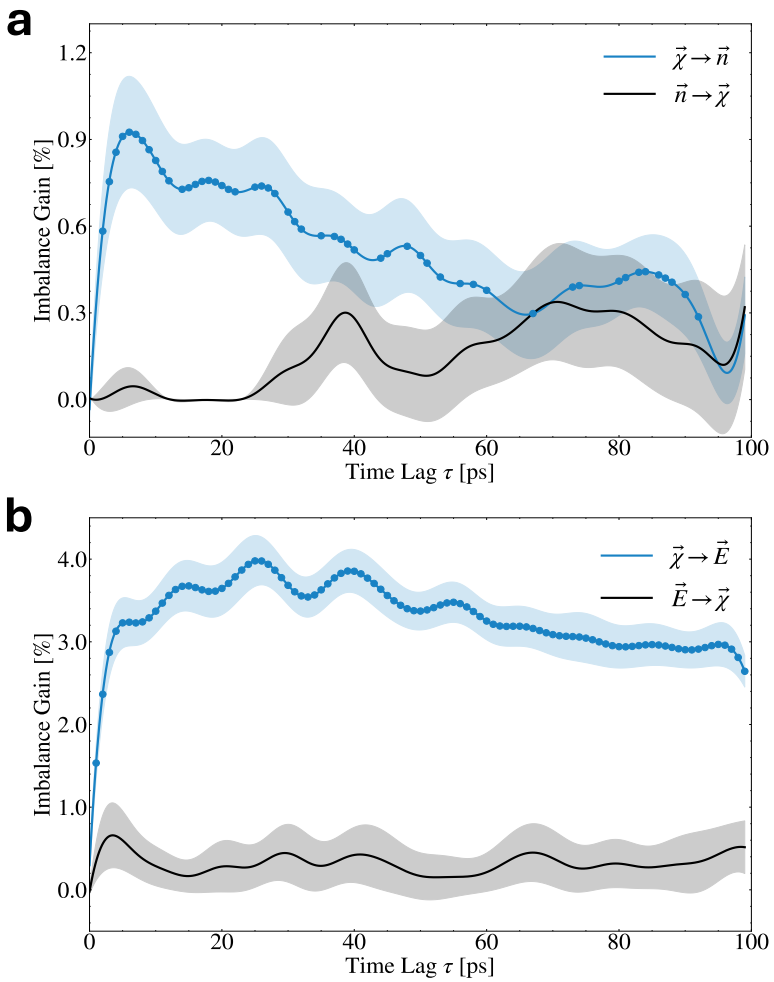


Figure 5.3: ***a****: IG for* $\vec{\chi} \to \vec{n}$ *(blue) and* $\vec{n} \to \vec{\chi}$ *(black).* ***b****: IG for* $\vec{\chi} \to \vec{E}$ *(blue) and* $\vec{E} \to \vec{\chi}$ *(black). Shaded regions denote error bars over 14 independent estimates, and bold points denote values that are significantly different from 0 according to a t-test* ($p < 0.001$).

Next, we extend the analysis to multidimensional CVs, involving solvent coordinates and the energetic variables mentioned before. Conducting this type of analysis using TE is difficult due to the need to construct high-dimensional probability distributions. Fig. 5.3**a** shows the Imbalance Gain between two multidimensional CVs, $\vec{\chi}$ and $\vec{n}$, which represent a collection of dihedrals and coordination numbers, respectively.

The dihedral vector, $\vec{\chi} = (\chi_1, \chi_2, \chi_3)$, is composed of the 3 dihedrals discussed above and shown in Fig. 5.1, while $\vec{n} = (n_{\mathrm{CT}}, n_{\mathrm{O1}}, n_{\mathrm{O2}}, n_{\mathrm{NT}}, n_{\mathrm{NE1}})$ includes the coordination numbers of the water oxygens with the C-terminus (CT), the carbonyl O atoms (O1 and O2), the N-terminus (NT), and the indole N-H (NE1). Fig. 5.3**a** shows the emergence of a unidirectional information transfer from $\vec{\chi}$ to $\vec{n}$. Finally, we computed the IG between $\vec{\chi}$ and $\vec{E}$, where $\vec{E} = (\Delta\mathrm{IP}, \Delta\mathrm{IW})$, which also unveils a clear asymmetry (Fig. 5.3**b**).

5.3.2 Emergence of causal links in model systems at equilibrium

To interpret the results of the previous section from the molecular dynamics simulations of Tryptophan, we conducted a similar analysis on simple model systems: a discrete-time Markov process (Fig. 5.4**a**) and two Langevin dynamics on different free energy surfaces (FES) (Figs. 5.4**d** and 5.4**g**). In all these systems the dynamics satisfies detailed balance.

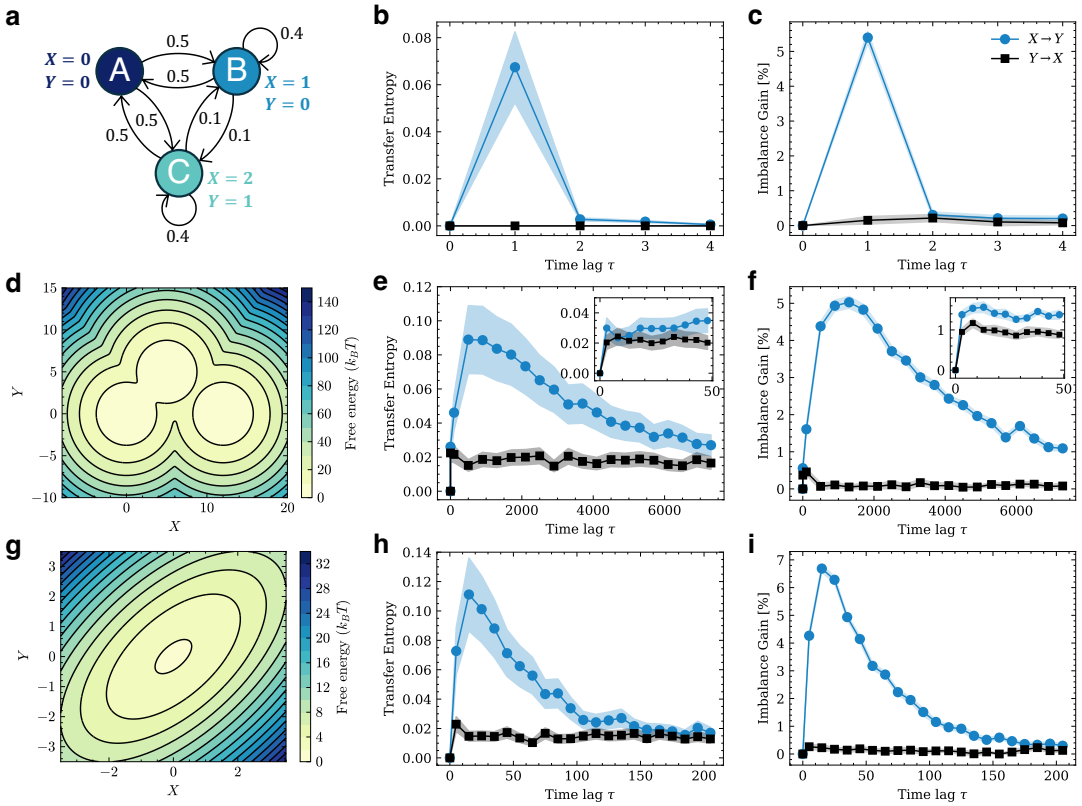


Figure 5.4: ***a****: Markov diagram for a three-state system, with transition probabilities reported next to the arrows. X and Y are discrete dynamical variables whose values in each state are specified close to corresponding node.* ***b*** *and* ***c****: TE and IG from the dynamics generated by* ***a****, as a function of the time lag* τ*.* ***d****: Asymmetric FES used for an overdamped Langevin dynamics with identical friction coefficients for X and Y.* ***e*** *and* ***f****: TE and IG from the Langevin dynamics run over* ***d****.* ***g****: Symmetric FES used for an overdamped Langevin dynamics with different friction coefficients for X and Y* ($\gamma_X = 1$, $\gamma_Y = 0.1$)*.* ***h*** *and* ***i****: TE and IG from the Langevin dynamics run over* ***g****. Shaded regions denote error bars over 20 independent estimates, each based on* $N = 2000$ *trajectories. The IG was computed using* $k = 25$ *neighbors.*

In the Markov system of Fig. 5.4**a**, the three states A, B and C are uniquely identified by variable X, which assumes different values (0, 1 or 2) in each state. The second variable, Y, can only distinguish states A and B ($Y = 0$) from state C ($Y = 1$), but not A and B from each other. Therefore, Y contains information that is redundant once the value of X is known. For such a system, it is possible to show both analytically (Section C.2) and numerically (Fig. 5.4**b**) that the TE is non-zero in direction $X \to Y$, while it is exactly zero in the reverse direction. This finding is reproduced by the IG as a function of the time lag τ (Fig. 5.4**c**), which is significantly different from zero only in direction $X \to Y$, and only for the time lag $\tau = 1$. We note that in this system, the actual causal link from X to Y is instantaneous, as Y_t is a deterministic function of X_t. Although TE and IG cannot directly test whether causal links are instantaneous, they can detect their presence at larger time lags (see Section C.1).

As a second example, we consider an overdamped Langevin dynamics (see Section 5.2.1) carried out over the FES of Fig. 5.4**d**, which can be seen as a continuous version of the previous Markov system. Again, variable X carries more information than Y about the true state of the system, as the three minima can be distinguished by projecting the free energy along X, while only two minima can be identified by projecting along Y. This is sufficient to observe a TE unbalanced in direction $X \to Y$ (Figs. 5.4**e** and 5.4**f**), when time lags of order of the transition times are considered. The information transfer in direction $Y \to X$ is instead close to zero according to both measures, although the TE in this direction suggests the presence of a bias towards nonzero values. This is compatible with the observations of Ref. [15], where it was shown that this measure (computed with the same estimator [133]) is prone to false positives in the direction without causal coupling. This problem is strongly alleviated by the Imbalance Gain (Fig. 5.4**f**). For smaller time lags, the thermal fluctuations within a single minimum play a role, and Y still carries information about the state of the system that is not included in X. This is reflected by comparable information transfers in the two directions for very small τ (insets of Figs. 5.4**e** and 5.4**f**).

As a third example, we consider a Langevin dynamics in the FES of Fig. 5.4**g**, which is symmetric under the exchange of X and Y. If the Langevin dynamics is generated using the same friction coefficient for both the variables, no information transfer asymmetry appears between X and Y (see Fig. C.5). In contrast, using a smaller friction coefficient for Y leads to the emergence of a clear unidirectional flow from X to Y (Figs. 5.4**h** and 5.4**i**).

Despite the differences, all previous examples describe a scenario where variable X is already maximally predictive with respect to its own future, and variable Y can only add redundant information on the future of X. In contrast, the uncertainty over the future of Y can be reduced if the current state of X is known.

Importantly, in the three examples such a "predictivity asymmetry" emerges from different mechanisms. In the first example, X provides a complete description of the system's state, while Y describes the system with a certain level of degeneracy. While X resolves the degeneracy of Y by distinguishing states that are identical according

to Y, the opposite is not true. In the second example, the system is two-dimensional, but the relaxation time within each minimum is much shorter than the transition times between the minima, so that the only information still retrievable at long time scales is the knowledge of the minimum in which the system is trapped. Therefore, Y carries non-redundant information that allows improving the prediction of X only for small time lags, but not in the far future, where all relevant information is already contained in X. In this case, X and Y are CVs retaining independent information of the true state, with X being more informative than Y in the long time-scale regime. In these first two examples, the asymmetric information transfer is rooted in the different information content that the variables retain about the true system's state.

In the third example, the symmetry of the FES implies that X and Y retain the same level of information of the system's state at a given time. However, such information levels become significantly different if referred to the *future* state of the system, as a consequence of the different relaxation times: while the description provided by X at time zero can still be used as a good proxy of the system's state at time τ, the same does not apply for Y if Y has already equilibrated. Therefore, this information transfer asymmetry is due to adiabatic separation: variable X moves slowly, leaving to Y the time to relax. In this condition, all the information on the long time-scale dynamics is provided by X alone.

The results described in Sec. 5.3.1 can be explained in light of the mechanisms just identified. In Figs. 5.5**a** and 5.5**b** we plot the FES as a function of the two dihedral angles (χ_3 and χ_2), for the GS and ES respectively. In the FES in Fig. 5.5**a**, more minima can be discerned by χ_3 than by χ_2. More precisely, the marginal free energies of the two angles (see Fig. C.7) show three minima for χ_3 and two minima for χ_2, with a higher barrier for χ_3 ($\sim 6 - 9\,k_BT$) than for χ_2 ($\sim 4\,k_BT$). This indicates that χ_3 resolves the "degeneracies" of χ_2 more effectively than vice versa, or equivalently, that χ_3 serves as a better CV than χ_2. In the ES (Fig. 5.5**b**), some of the minima along χ_3 (specifically those in $(\chi_2, \chi_3) = (-2, 1)$, $(\chi_2, \chi_3) = (1.5, -1.25)$ and $(\chi_2, \chi_3) = (-1.75, 3)$) become more pronounced. As shown in Figs. 5.2**c** and 5.2**d**, this leads to a more pronounced information transfer in direction $\chi_3 \rightarrow \chi_2$. The slower decay of the TE and IG curves is determined by the deeper FES minima, which make χ_3 a slower mode in the ES than in the GS.

To further rationalize the asymmetries observed in Fig. 5.2, we turned to the FES between the two energetic variables, illustrated in Figs. 5.5**c** and 5.5**d**. In sharp contrast to the FES involving the dihedrals, these distributions show a single broad minimum, which is only slightly asymmetric in the two variables. However, the two CVs display decorrelation times that are different for Δ IP (~ 7 ps in the GS, ~ 11 ps in the ES) and Δ IW (~ 1 ps in the GS, ~ 2 ps in the ES), making Δ IP a slower CV than Δ IW (see Fig. C.8). This leads us to conclude that the asymmetric information transfer between the indole-peptide energetics and the indole-solvent energetics is a molecular example of the scenario shown earlier in Fig. 5.4 (panels **g**,**h**,**e**).

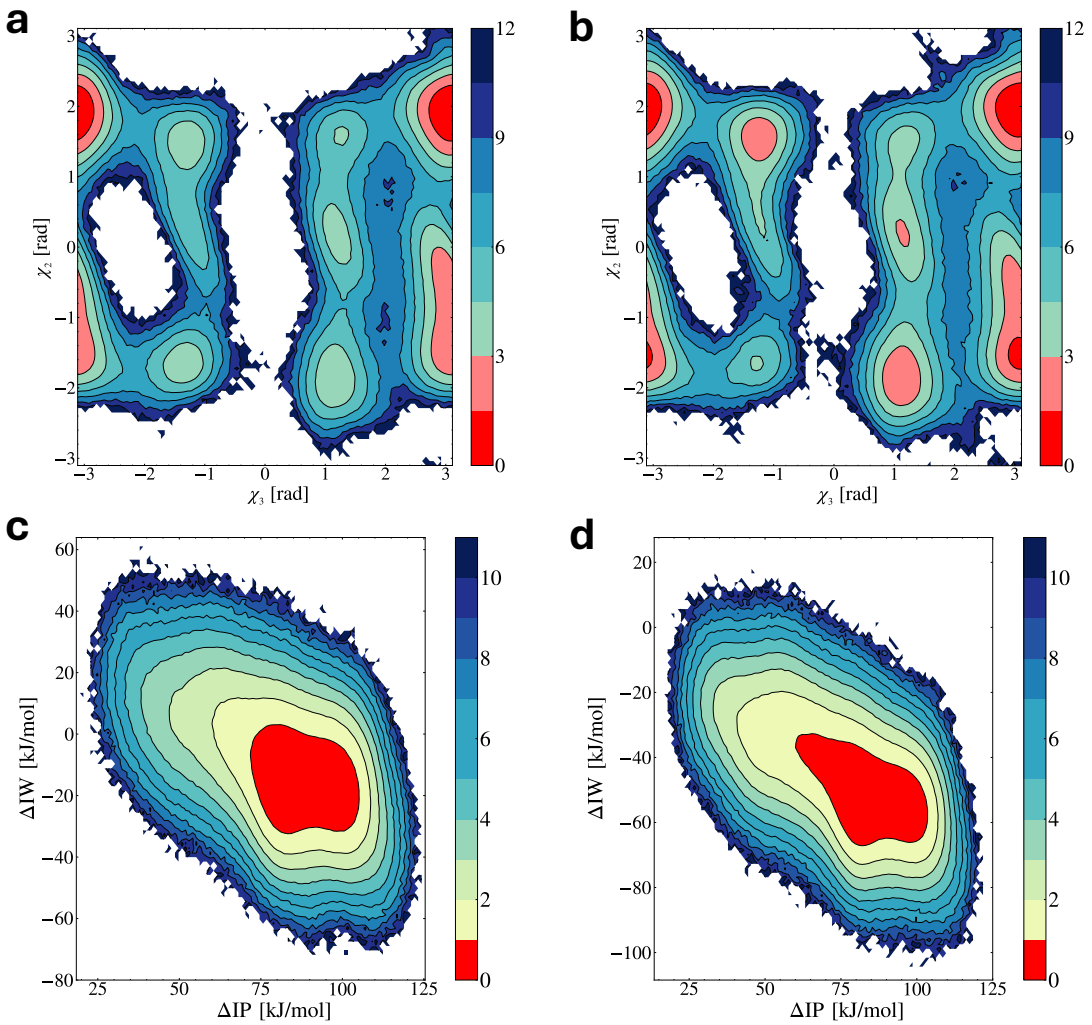


Figure 5.5: ***a*** *and* ***b****: the FES for* χ_3 *and* χ_2 *in the GS and ES respectively.* ***c*** *and* ***d****: the FES for* ΔIP *and* ΔIW *in the GS and ES respectively.*

5.3.3 Asymmetries and the response to external interventions

In this section we will show that observing a unidirectional information transfer between two variables is a *necessary, but not sufficient* condition for the existence of a genuine causal link. Using the tools of causal inference, we analyze the behavior of the system in response to an active intervention, which we apply by setting a variable to a specific value, without changing any other variable that is not a direct cause of the manipulated one. In Langevin models, a "hard" intervention $do(X_0 = x)$ can be thought to as an ideal experiment where the "natural" state of the system at time $t = 0$, (X_0, Y_0), is instantaneously set to (x, Y_0). After the external manipulation, the system is left free to evolve according to its unperturbed dynamics. Interventions provide an intuitive framework to speak about causality: we can state that X causes Y if and only if an intervention on X_0 leads to a measurable effect on Y_τ for some $\tau > 0$. This effect can be quantified, for example, by measuring the Kullback-Leibler (KL) divergence between the distributions of Y_τ under two different interventions on X_0.

In Fig. 5.6 we show the effect of interventions on variables X and Y in two different

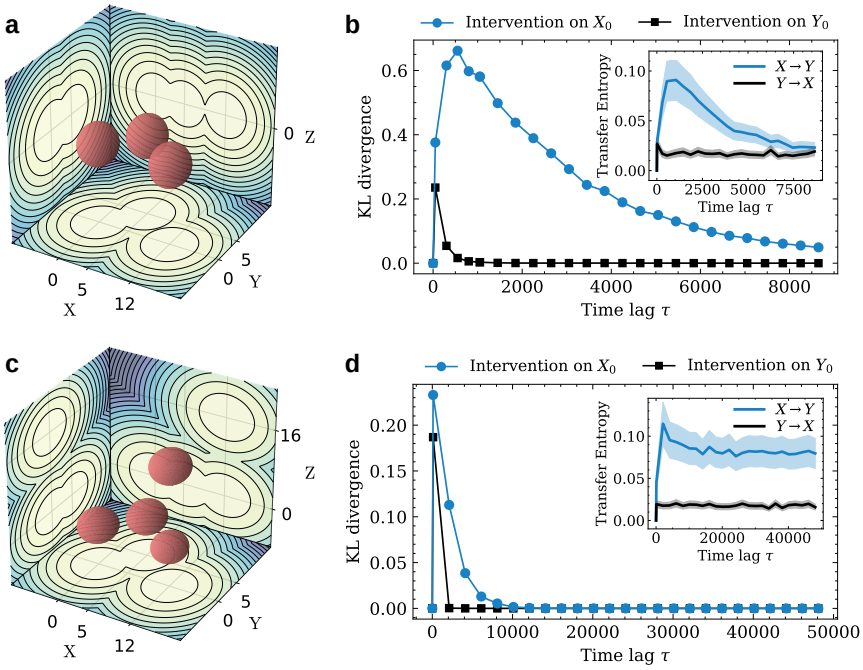


Figure 5.6: *Effect of different interventions on two three-dimensional Langevin system with the same free energy projection in the XY-plane.* ***a*** *and* ***c****: Red spheres represent the free energy isosurface for* $F = 13\,k_B\,T$*. The three planes show the free energy projections for all pairs variables.* ***b*** *and* ***d****: KL divergences between post-interventional distributions, intervening on* X_0 *(blue curves) and on* Y_0 *(black curves), as a function of the lag* τ *between the intervention and the time at which its effect is observed. Specifically, the two curves display* $D_{KL}\left[p(Y_\tau \mid X_0 = 0) \| p(Y_\tau \mid X_0 = 12)\right]$ *and* $D_{KL}\left[p(X_\tau \mid Y_0 = 0) \| p(X_\tau \mid Y_0 = 5)\right]$*, respectively.*

three-dimensional model systems whose two-dimensional free energy as a function of X and Y is exactly identical to the free energy in Fig. 5.4**d**. We treat the third variable, Z, as if it were unobserved, computing only bivariate Transfer Entropies between X and Y.

In the first system (top row) the Z variable has the same distribution in the three minima, and the TE (inset in panel **b**) is qualitatively equivalent to the two-dimensional case shown in Fig. 5.4. In panel **b** we show the effect of $do(X_0)$ and $do(Y_0)$ experiments (blue and black curves, respectively), using as interventional values the positions of the furthest minima seen by each variable ($X_0 = 0,\ 12$ and $Y_0 = 0,\ 5$). The effect on Y_τ of the interventions on X_0 is still visible for large time lags, whereas the interventions on Y_0 have no effect on X_τ after a time scale comparable to the relaxation time within the minima. Therefore, in this example, the Transfer Entropy provides qualitatively the same information that one would infer by performing an external manipulation of the system.

In the second system (bottom row of Fig. 5.4), Z does not distinguish the three minima seen by X, but reveals a fourth minimum, hidden for X, which features a higher free energy barrier than all "visible" barriers. The bivariate Transfer Entropy (inset of panel **d**) shows qualitatively the same asymmetry observed in the previous example, suggesting a unidirectional causal link $X \to Y$. However, intervening on X_0 affects Y_τ

on significantly shorter time scales than those deducible from Transfer Entropy (blue curve in panel **d**). Thus, using Transfer Entropy to infer the existence of a causal effect of X on Y would lead to the wrong conclusion: over longer time-scales, the unobserved variable Z behaves as a common driver of X and Y. In Fig. C.9 we support this statement by showing that an intervention on Z_0 results in a long-term effect on both X_τ and Y_τ, while it has no effect when applied to the system of Fig. 5.6**a**.

5.4 Discussion

In this work, we investigated the emergence of information transfer asymmetries in systems obeying equilibrium and time-reversible dynamics, and whether these asymmetries can be interpreted as causal links. We measured information transfers by using the Transfer Entropy and the Imbalance Gain. Crucially, both these observables efficaciously probe three-body dependencies, involving the present state of both the putative driver and the driven variables, and the future state of the latter. Standard time-lagged two-body cross-correlation functions (CCFs), which have been used to infer causal links from MD simulations [72, 66], cannot report on the directional flow of information between variables in stationary and time-reversible systems, as they are symmetric by construction. We illustrate this property by computing the CCFs between some of the relevant collective variables for TRP (see Figs. C.10 and C.11).

Consistent with previous studies [64, 65, 66, 67, 258, 68, 259, 69, 70, 71, 260, 72, 261, 262], we observed empirically that information transfer asymmetries can emerge even in a simple but realistic molecular system, namely a solvated TRP molecule. The choice of this system is motivated by Fluorescence Stokes Shift experiments, where TRP can be electronically excited by absorbing UV photons and used to probe solvation dynamics [88, 89, 90, 91]. Using model systems, we identified two mechanisms that explain the emergence of such asymmetries: (i) the asymmetry in the information content of different CVs, namely the capacity of one CV to describe states and transitions hidden to the others, and (ii) the discrepancy in their relaxation times. In particular, we found that the most informative CVs and the slowest CVs act as "sources" of information towards other CVs. Remarkably, all the asymmetries observed in the TRP system can be explained according to either one mechanism or the other. We also note that the first mechanism implies the second, as a CV that identifies more free energy minima can only relax on longer time-scales than a CV for which some minima are hidden. These findings provide enhanced insight into earlier studies [71, 73], indicating that molecular descriptors selected by Granger Causality [71] or Transfer Entropy [73] can accurately characterize transition states. The asymmetry of information flow also opens up interesting perspectives on how to measure the chemical physics of coupling between protein and water degrees of freedom [300, 301, 302]. For the case of TRP, we observe that there is unidirectional flow of information from protein coordinates such as the dihedrals to the solvent. It would be interesting to understand the extent to which this directionality changes for tryptophan embedded in different chemical environments

in proteins.

Information transfer asymmetries inferred on equilibrium dynamics are typically associated to causal relationships [65, 69, 70, 262]. In this work, we have explicitly shown that such asymmetries are only a sufficient condition for inferring causal links, as unobserved CVs - namely, those not considered in the analysis - may act as common drivers of observed CVs. Specifically, if Z identifies a higher free energy barrier than those observed by X and Y, Z behaves as a common driver of X and Y on time scales comparable with the transition time to the hidden minimum.

Our findings suggest two possible routes for discovering causal relationships in molecular systems: either using a set of CVs that can be safely assumed to be "causally sufficient" – that is, unaffected by unobserved common drivers, or performing explicit interventional experiments on CVs of interest. The first approach is viable by considering a large pool of CVs, such as all key dihedrals of a molecule [71], and estimating information transfers in a multivariate fashion (see Section C.1). This approach is unavoidably affected by the curse of dimensionality as the number of CVs increases, although methodologies designed for high-dimensional settings, such as the IG and its extensions [303], promise to alleviate this issue.

The second approach necessitates the design of interventional experiments on molecular CVs. This approach sounds natural in a simulation setting, in which one can perform arbitrary manipulations on the system, but poses some practical challenges. In particular, the interventional experiments that we carried out on the model systems (Sec. 5.3.3) were applied to "orthogonal" CVs X and Y, such that an instantaneous variation of X at time $t = 0$ does not affect Y at the same time. This may be the case for CVs that describe spatially separated regions of a molecular system, such as distant sites within a protein. However, CVs of interest can also depend on a common subset of degrees of freedom that generate instantaneous dependencies. In this scenario, setting $X(0)$ to an arbitrary value ($do(X(0) = x)$) may be practically impossible without changing also $Y(0)$. As an example, in our TRP system, the dihedrals χ_1, χ_2 and χ_3 depend on common atomic positions affected by rigid constraints, and in turn, not all arbitrary choices of such angles are possible. Moreover, "hard" interventions such as those applied in this work appear challenging in molecular dynamics simulations, as they would require an instantaneous modification of several degrees of freedom, making it necessary to develop appropriate protocols. The design of suitable interventional experiments on molecular CVs will be the subject of future work.

6

Investigating causality between principal components in protein dynamics

Principal component analysis (PCA) is widely used to characterize collective protein motions from molecular dynamics (MD) simulations. While PCA identifies the dominant modes of structural fluctuation, it does not reveal whether different principal components (PCs) causally influence each other. Here, we investigate this question using a recently introduced causal-discovery framework [15], which allows to infer putative causal asymmetries between high-dimensional time series. We apply this approach to long-timescale MD trajectories of two proteins. By analyzing relationships among PCs, we construct directed networks describing how PCs influence one another across time scales. These directional relationships, whose existence is a necessary condition for the presence of a causal link, are not captured by conventional covariance-based analyses and provide information that is complementary to PCA and Time-lagged Independent Component Analysis (TICA). Our results suggest that our causal inference approach can uncover previously hidden aspects of the dynamical organization of protein motions and offer a new perspective on this very popular class of collective variables.

6.1 Introduction

Understanding how collective motions in proteins are organized across multiple time scales remains a central challenge in molecular biophysics [304, 305, 306, 307, 308, 309, 310, 311, 312]. Molecular dynamics (MD) simulations now routinely generate trajectories of sufficient length and resolution to characterize fluctuations ranging from fast local motions to rare conformational transitions [313, 314, 315, 316, 317, 318, 319, 320, 321], enabling the investigation of both thermodynamic and kinetic properties. In particular, equilibrium trajectories of small proteins extending to biologically relevant

time scales are now available [322, 315, 323], providing unprecedented opportunities to study the relationship between structure, dynamics, and function.

One of the most widely used approaches for analyzing protein dynamics in the folded state is Principal Component Analysis (PCA) [324, 325, 326, 327, 328, 329, 330, 331, 332, 92, 333, 93]. In its standard formulation, PCA is performed by computing the covariance matrix of the positions of a selected subset of protein atoms (for example, backbone α carbon atoms) after removing overall translational and rotational motions. The principal components (PCs) are the eigenvectors of this covariance matrix, while the corresponding eigenvalues quantify the variance along each component [334, 329, 331, 335, 336, 92, 93]. When only a small number of eigenvalues account for a large fraction of the total variance, protein motions can be represented in a reduced-dimensional basis spanned by the leading PCs. In this basis, the covariance matrix is diagonal by construction, meaning that fluctuations along different PCs are uncorrelated at zero time lag. Owing to its ability to provide a compact description of collective structural fluctuations and to identify large-amplitude motions, PCA has become a standard tool for characterizing conformational landscapes and functionally relevant motions in proteins [337, 335, 338, 339, 340, 341, 342, 332, 343, 344, 92, 345, 346, 347, 93].

When the focus shifts from structural characterization to dynamical behavior, Time-lagged Independent Component Analysis (TICA) provides a natural extension of PCA [348, 349, 350, 351]. While principal components are uncorrelated at zero lag time, correlations generally arise between them when the system is observed at a finite time delay, τ. $C(\tau)$ denotes the time-lagged covariance matrix between the centered and aligned atomic coordinates at times 0 and τ. In the PCA basis, the corresponding lagged covariance matrix is given by $\Gamma(\tau) = C^{-1}(0)C(\tau)$ [348, 350, 352, 353]. This matrix would remain diagonal for all values of τ only in the idealized case of purely harmonic dynamics. For realistic protein motions $\Gamma(\tau)$ is generally non-diagonal, reflecting dynamical couplings between different PCs. The eigenvectors of $\Gamma(\tau)$, commonly referred to as time-lagged independent components (ICs), identify collective coordinates associated with the slowest dynamical processes. As a result, they are widely employed to characterize long-timescale conformational changes, construct kinetic models, and define collective variables for enhanced-sampling simulations [350, 351, 354, 352, 355, 356, 357].

These developments raise a natural question: can collective structural fluctuations identified by PCA exert asymmetric dynamical influences on one another? More specifically, can principal components extracted from protein MD trajectories *cause* each other? [267, 358, 270] In the language of causal inference, stating that, for example, PC_2 is a putative cause of PC_4 means that knowledge of the present state of PC_2 improves the prediction of the future evolution of PC_4 beyond what can be inferred from the history of PC_4 alone [62]. A natural first attempt to address this question would be to examine the time-lagged covariance matrix between PCs, $\Gamma(\tau)$. If, for instance, the off-diagonal element $\Gamma_{24}(\tau)$ is large, one may conclude that PC_2 and PC_4 are dynamically coupled. However, dynamical coupling does not necessarily imply causal

influence. In particular, causal relationships are generally directional: it may happen that PC_2 influences the future evolution of PC_4, while the converse is not true. Such asymmetries, if present, cannot be inferred from $\Gamma(\tau)$ alone. For reversible equilibrium dynamics, the transfer operator is self-adjoint with respect to the equilibrium measure. Consequently, TICA identifies slow decorrelating coordinates but cannot assign a directional driver–response interpretation to couplings between pairs of PCs.

Addressing this question requires tools that go beyond covariance or cross-correlation based analyses. Several approaches have been proposed to infer causality from molecular simulations. Early studies employed multivariate autoregressive models and analyses inspired by Granger causality to identify directed relationships between coordinates or residue fluctuations [64, 65, 69, 71]. An information-theoretic approach, dubbed transfer entropy [63, 134], has also been used to probe directional information transfer and allosteric communication in proteins [66, 67, 68]. More broadly, causal inference for time series has developed into rich theoretical frameworks aimed at predicting information asymmetries, conditional dependencies, and structural causal graphs [62, 267, 270, 269]. In practice, however, applying these approaches to atomistic MD remains challenging because the statistical errors of causal estimators are typically large when the causal links are weak.

To address these difficulties, we recently introduced the Imbalance Gain (IG), a statistical test designed to identify putative causal couplings in time-dependent data [13, 15, 3]. The IG can be viewed as a proxy for the transfer entropy [63, 134] and can also be efficiently estimated for noisy and high-dimensional data [15]. It is also related to Granger causality and related methods but, unlike those approaches, it does not require fitting an explicit dynamical model. Recent studies have shown that putative causal couplings identified through the IG can robustly reveal genuine causal relationships even in equilibrium molecular dynamics, despite the time-reversible nature of the microscopic equations of motion and the stationarity of the underlying distribution [3]. The same framework has also uncovered non-trivial causal relationships involving translational and rotational degrees of freedom in liquid water [11]. It is important to note that the IG and other similar approaches infer the presence of *putative* causal links from observational data. The presence of an asymmetric dynamical coupling between two variables is a *necessary* condition of the presence of a causal link between the two variables. However, this condition is not sufficient, since the two variables could be simultaneously caused by a third unobserved variable. Revealing the presence of confounders and building a genuine causal graph [267, 270, 269] in molecular modeling remains an open challenge.

In the present work, we use the IG framework to investigate whether the principal components describing collective protein motions exhibit putative causal links. We analyze two long-timescale trajectories generated by D. E. Shaw Research: ubiquitin (PDB ID: 1UBQ) [323] and NTL9 (PDB ID: 2HBA) [315]. NTL9 is a fast-folding protein whose trajectory samples large conformational changes and multiple folding and unfolding events [315], whereas ubiquitin represents a prototypical native-state sys-

tem whose equilibrium fluctuations span timescales from picoseconds to milliseconds and involve subtle long-range collective motions [323]. By comparing these two systems, we assess whether putative causal links among PCs emerge, and whether such links are controlled by variance, relaxation timescale, or spatial collectivity. We find that both proteins exhibit a causal structure, but with distinct physical character: in ubiquitin, the directionality of the putative causal link originates from a highly collective, protein-wide mode and affects only the next few leading components, whereas in NTL9, a localized slow mode centered on the β-turn directionally couples to several faster modes, including more delocalized principal components.

6.2 Methods

The Information Imbalance and the Imbalance Gain used throughout this chapter are defined in Chapter 2, in Sections 2.5 and 2.7 respectively. This section describes the trajectories analysed and the parameters entering the analysis.

6.2.1 Trajectory analysis and convergence

For both systems, the use of long molecular dynamics trajectories is essential to ensure that the principal components are statistically converged and representative of the underlying folded-state dynamics. Convergence was assessed by dividing each selected trajectory into two halves, independently computing the principal components for each half, and evaluating their consistency through the cosine similarity between the corresponding eigenvectors. Refer to Appendix Figs. D.2 and D.4 for the results of this calculation for Ubiquitin and NTL9 respectively. The observation that the same dominant PCs are recovered from both halves provides a stringent check that the PCA is not affected by insufficient sampling.

Ubiquitin

The ubiquitin trajectory analyzed in this work corresponds to the $1\ \mathrm{ms}$ simulation reported in Ref. [323]. In that study, the RMSD was computed over residues 2–71, thereby excluding the flexible C-terminal tail. We therefore restricted our analysis to the same residue range. Moreover, the trajectory exhibits a conformational transition at approximately $0.65\ \mathrm{ms}$, as reflected in the RMSD profile shown in Fig. 1 of Ref. [323]. To avoid mixing distinct conformational regimes, we limited the present analysis to the first $0.65\ \mathrm{ms}$ of the simulation.

NTL9

The NTL9 protein was simulated for a total of $2.9\ \mathrm{ms}$ in Ref. [315], distributed across four independent simulations. Full details of these simulations are provided in the Supporting Information of Ref. [315]. Among these trajectories, the first two have

lengths of 1.112 ms and 1.073 ms, respectively, and each contains two long folded-state segments exceeding 200 μs. Since our objective was to analyze principal components representative of the folded-state dynamics, we selected two such folded trajectory segments for the present analysis. The results reported below were obtained from the segment spanning 830–1060 μs of the 1.073 ms trajectory, whereas the consistency check reported in Section D.2 was performed using the segment spanning 290–580 μs of the 1.112 ms trajectory.

6.2.2 Imbalance Gain analysis of principal components

In the present work, we apply this framework to PCs obtained from protein dynamics. Treating individual PCs as time-dependent variables, we evaluate the lag-dependent IG between pairs of components to assess whether specific modes systematically improve the prediction of others. This allows us to infer the presence of putative causal links in principal component space. To simplify the visualization and the discussion, we restrict our analysis to the top five PCs.

Principal components are obtained from the C_α atoms of the trajectory fragments described above. In evaluating the Information Imbalance of Eq. (2.22), the number of neighbours is set to $k = 50$, which is about 2% of the number of configurations N. Because the Imbalance Gain is computed over an ensemble of replicas, each trajectory is divided into non-overlapping chunks, which provide the independent initial conditions entering the distances. The optimal value of α that maximises the IG for a given time lag τ is determined by searching over a grid of 500 evenly spaced values for $\alpha \in [0, 10]$.

6.3 Results

6.3.1 Ubiquitin

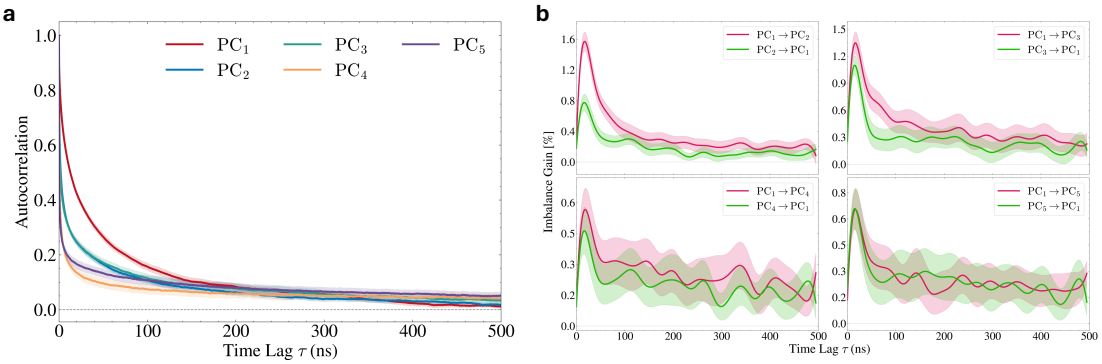


Figure 6.1: *Ubiquitin: (a) The autocorrelation functions of the top 5 PCs, $PC_{i=1,5}$, which are colored red, blue, green, orange, and purple respectively, with shaded areas denoting error bars. (b) The pairwise IG curves over time between $PC_1 \Leftrightarrow PC_{2,3,4,5}$. The IG values are on the Y-axis, and time lag (τ) in nanoseconds is on the X-axis. Pink lines denote the direction $PC_1 \rightarrow PC_{2,3,4,5}$, whereas green lines denote the reverse direction. Shaded areas denote error bars computed over 20 independent realizations.*

In the first application, we analyze the native-state dynamics of ubiquitin using the $1\ \mathrm{ms}$ molecular dynamics trajectory reported in Ref. [323]. As discussed in that work, the trajectory exhibits a mild conformational transition at approximately $0.65\ \mathrm{ms}$. Since our goal is to characterize causal relationships between principal components within a single native-state basin, we restricted the analysis to the first $0.65\ \mathrm{ms}$ of the simulation. This choice avoids mixing fluctuations associated with distinct conformational substates and ensures that the resulting PCs describe a well-defined region of the native-state ensemble. The eigenvalue spectrum and the explained variance of the PCs are shown in Fig. D.1.

We first characterize the intrinsic relaxation timescales of the leading modes. In Fig. 6.1**a**, we report the autocorrelation functions of the first five principal components. All five modes decorrelate within approximately $500\ \mathrm{ns}$, which is substantially shorter than the analyzed trajectory segment. We determine the relaxation times from the integral of the autocorrelation function. This indicates that the selected portion of the simulation is sufficiently long to sample the dominant fluctuations associated with these PCs. Among the modes considered, PC_1 is the slowest, while PC_2 and PC_3 decorrelate on comparable, shorter timescales.

Previous studies on equilibrium MD simulations have used apparent asymmetries in time-lagged two-body cross-correlation functions, $\mathbb{E}[X_0 Y_\tau]$, to infer causal relationships [72, 66]. However, for equilibrium dynamics satisfying stationarity and time reversibility (as is the case for MD simulations), these correlation functions are invariant under exchange of the two variables. In particular, $\mathbb{E}[X_0 Y_\tau] = \mathbb{E}[X_{-\tau} Y_0] = \mathbb{E}[X_\tau Y_0]$, where the first equality follows from stationarity and the second from time-reversibility. Therefore, any observed asymmetry in such correlation functions cannot be interpreted as evidence of directional causality. Instead, it must arise from statistical errors, numerical violations of time reversibility introduced by the integration scheme, or from perturbations associated with thermostats and barostats. The IG (or transfer entropy) approach is instead asymmetric and directional, and can be viewed conceptually as a many-body cross-correlation function where the information of the caused variable at the present time breaks the symmetry otherwise seen in two-body cross-correlation functions [15, 3].

Therefore, we next examine putative causal links between the leading modes using the Imbalance Gain, instead of using two-body cross-correlation functions between the PCs. In Fig. 6.1**b**, we show the pairwise IG curves between PC_1 and $\mathrm{PC}_{2,3,4,5}$. The strongest asymmetry is observed for the pairs $\mathrm{PC}_1 \Leftrightarrow \mathrm{PC}_2$ and $\mathrm{PC}_1 \Leftrightarrow \mathrm{PC}_3$, where the dominant direction of information transfer is $\mathrm{PC}_1 \rightarrow \mathrm{PC}_{2,3}$. This indicates that knowledge of PC_1 improves the prediction of the future evolution of PC_2 and PC_3 more than the reverse. The asymmetry becomes weaker for PC_4, where the driving effect from PC_1 remains detectable but is considerably more modest. For PC_5, the IG signal is substantially reduced and becomes only weakly discernible. Consistently, when lower-variance PCs are considered, the directional signal progressively decreases and eventually approaches zero, as shown in the additional pairwise IG curves reported in

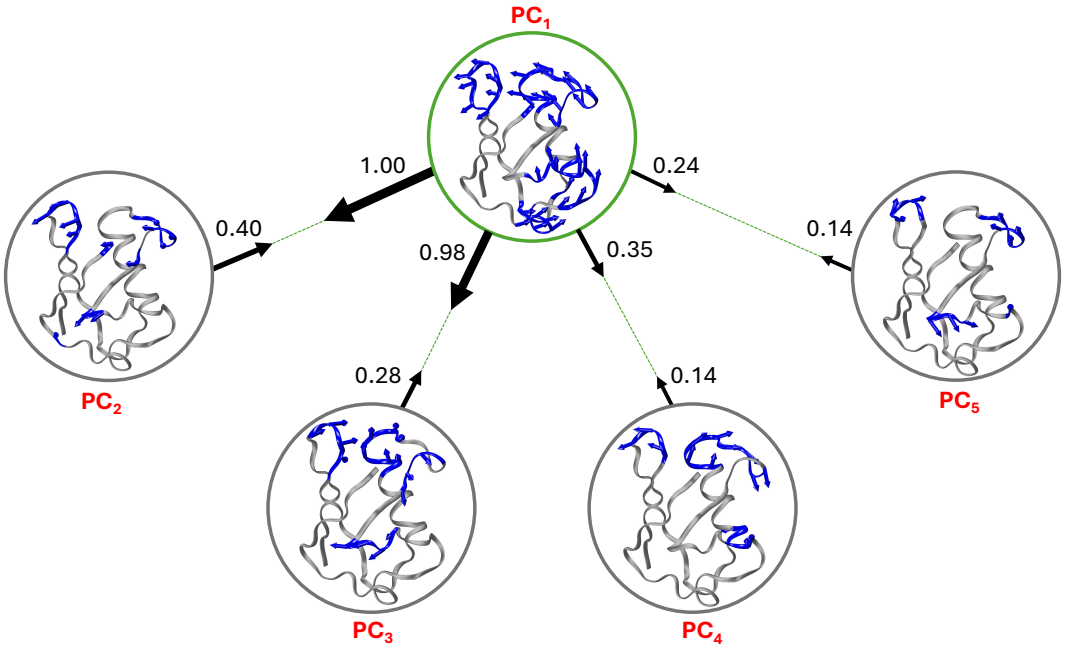


Figure 6.2: *Ubiquitin: Putative causal links between* PC_1 *and the remaining leading principal components. Nodes represent the first five PCs, and directed edges indicate the normalized integrated Imbalance Gain,* IGT*, defined in Eq. 2.30. For each pair, the edge length and width, as well as the numerical label report the relative strength of the directional information transfer, with larger values corresponding to stronger causal links. The strongest links are observed from* PC_1 *toward* PC_2 *and* PC_3*, whereas the links involving* PC_4 *and* PC_5 *are substantially weaker. For each PC, the residues with the largest contributions to the corresponding eigenvector are highlighted in blue on the protein structure, as quantified using the Brüschweiler collectivity index. The arrows in blue show the corresponding directions of the eigenvector for the highlighted residues, while the rest of the protein is shown in gray.*

the appendix (see Fig. D.7). Overall, these results suggest that the dynamical influence of PC_1 is primarily exerted on the other dominant modes and progressively fades for less important PCs. This behavior is consistent with our empirical observations from both Langevin-based models as well as simple molecular systems, in which the slowest collective variable typically appears as a putative causal driver [15, 3].

The putative causal links between pairs of PCs are summarized in Fig. 6.2 using the normalized IGT values defined in Eq. 2.30. The graph highlights the dominant links between PC_1 and $\mathrm{PC}_{2,3,4,5}$. In agreement with the time-resolved IG curves, the strongest directed links correspond to $\mathrm{PC}_1 \to \mathrm{PC}_2$ and $\mathrm{PC}_1 \to \mathrm{PC}_3$, confirming that the leading mode acts as the main driver of the next two dominant modes. A weaker directional asymmetry is also observed for the $\mathrm{PC}_1 \Leftrightarrow \mathrm{PC}_{4,5}$ links, which remain biased toward the $\mathrm{PC}_1 \to \mathrm{PC}_{4,5}$ direction. However, the magnitude of these links is markedly smaller than for PC_2 and PC_3, indicating that the causal influence of PC_1 rapidly weakens as one moves toward lower-variance PCs. As we will see in the next section, however, this is not necessarily a universal rule. In addition, Fig. D.9 shows that PC_2 exhibits appreciable asymmetric IG with both PC_3 and PC_4. In these cases, the dominant direction of information transfer is $\mathrm{PC}_2 \to \mathrm{PC}_{3,4}$, while the corresponding reverse links are noticeably weaker. By contrast, the IG between PC_2 and PC_5 is much smaller and

only weakly directional. The same trend persists for lower-variance PCs, for which the causal signal rapidly decreases and eventually becomes negligible.

In Fig. 6.2, we also highlight the residues that contribute most strongly to each principal component. These contributions were quantified using the Brüschweiler collectivity index, which quantifies the degree to which the collective motions are delocalized in the protein [359, 360]. Details are provided in Fig. D.5. This analysis shows that PC_1 is the most delocalized mode among those considered, with a collectivity index of approximately 0.8, indicating that residues distributed across the entire protein contribute substantially to this collective fluctuation. By contrast, PC_2–PC_5 are more localized, with collectivity indices in the range 0.4–0.6. The dominant role of PC_1 would suggest a scenario in which more collective protein motions have a causal influence on more spatially localized fluctuations. In the following section we will show that this scenario does not hold in general.

6.3.2 NTL9

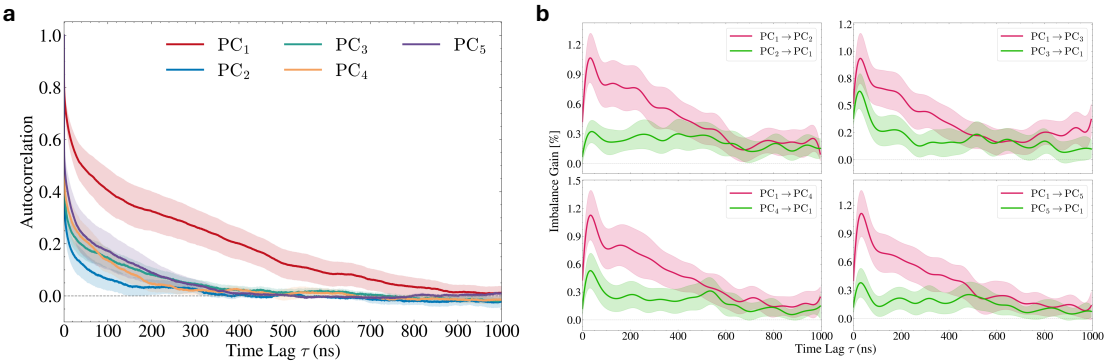


Figure 6.3: *NTL9: (a) The autocorrelation functions of the top 5 PCs, $\mathrm{PC}_{i=1,5}$, which are colored red, blue, green, orange, and purple respectively, with shaded areas denoting error bars. (b) The pairwise IG curves over time between $\mathrm{PC}_1 \Leftrightarrow \mathrm{PC}_{2,3,4,5}$. The IG values are on the Y-axis, and time lag (τ) in nanoseconds is on the X-axis. Pink lines denote the direction $\mathrm{PC}_1 \rightarrow \mathrm{PC}_{2,3,4,5}$, whereas green lines denote the reverse direction. Shaded areas denote error bars computed over 20 independent realizations.*

In the second application, we analyze folded-state segments of the NTL9 trajectory. In Ref. [315], NTL9 was simulated for approximately 2.9 ms, during which the protein undergoes multiple reversible folding–unfolding transitions. Within these simulations, the protein remains folded for several extended intervals; in particular, we identified four independent folded segments longer than 200 μs. The results discussed in the following are obtained by analyzing one representative folded segment. The same analysis performed on the remaining folded segments yields qualitatively consistent results, as reported in the appendix (see Section D.2 and Figs. D.13 – D.17). This allows us to focus specifically on causal relationships between PCs within the folded-state dynamics, rather than on the much larger structural changes associated with folding and unfolding. The eigenvalue spectrum and the explained variance of the PCs are shown in Fig. D.3.

Fig. 6.3**a** shows the autocorrelation functions of the first five principal components. In contrast to ubiquitin, where the leading modes decorrelate on more comparable sub-microsecond timescales, NTL9 displays a clearer separation between the slowest mode and the remaining PCs. In particular, PC_1 decorrelates on a timescale of approximately $1\ \mu\mathrm{s}$, whereas $\mathrm{PC}_{2,3,4,5}$ relax substantially faster, with characteristic timescales of roughly $200\ \mathrm{ns}$. This separation indicates that PC_1 captures a distinct slow fluctuation within the folded-state ensemble of NTL9, while the remaining leading PCs describe faster modes embedded within the same folded basin.

We then examine whether this separation of timescales is reflected in the directionality of information transfer between the leading PCs. In Fig. 6.3**b**, we report the pairwise IG curves between PC_1 and $\mathrm{PC}_{2,3,4,5}$. In all cases, the curves show a pronounced asymmetry, with the $\mathrm{PC}_1 \rightarrow \mathrm{PC}_{2,3,4,5}$ direction consistently dominating over the reverse direction. Thus, knowledge of the slow coordinate PC_1 substantially improves the prediction of the future evolution of the faster leading PCs, whereas the converse effect is much weaker. Compared with ubiquitin, where the directional influence of PC_1 is strongest for PC_2 and PC_3 but rapidly decreases for lower-variance modes, the NTL9 folded segment exhibits a more systematic and persistent directional coupling from PC_1 toward all the other PCs considered (see Fig. D.8).

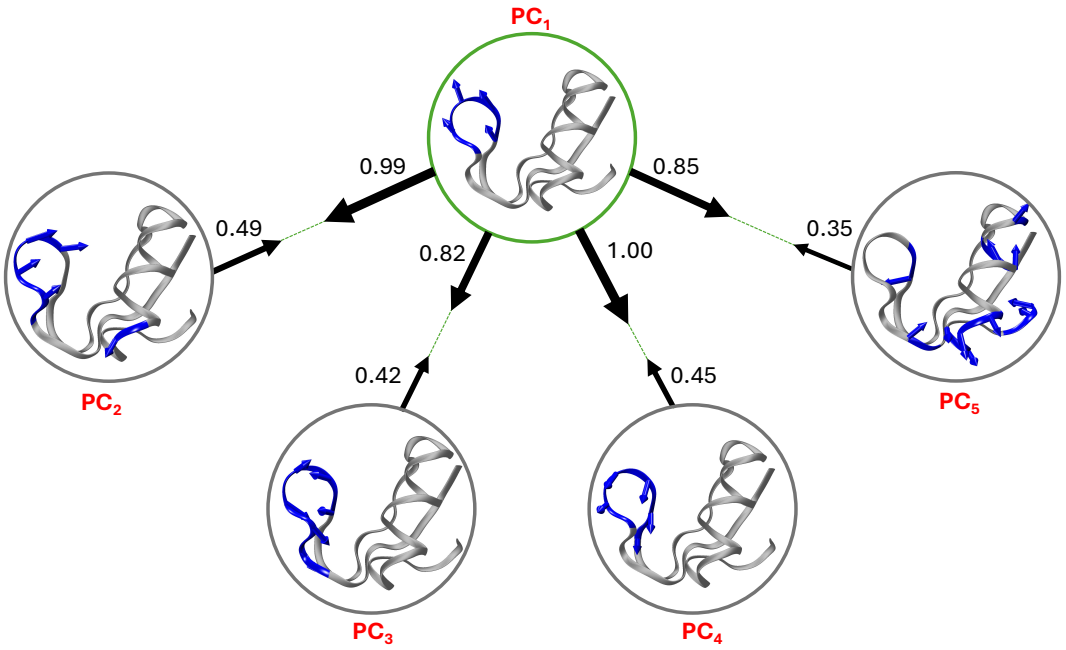


Figure 6.4: *NTL9: Putative causal links between* PC_1 *and the remaining leading principal components. Nodes represent the first five PCs, and directed edges indicate the normalized integrated Imbalance Gain,* IGT, *defined in Eq. 2.30. For each pair, the edge length and width, as well as the numerical label report the relative strength of the directional information transfer, with larger values corresponding to stronger causal links.* PC_1 *displays strong directional influence toward all other leading PCs considered, identifying it as a dominant slow driver within the folded-state dynamics of NTL9. For each PC, the residues with the largest contributions to the corresponding eigenvector are highlighted in blue on the protein structure, as quantified using the Brüschweiler collectivity index. The arrows in blue show the corresponding directions of the eigenvector for the highlighted residues, while the rest of the protein is shown in gray.*

This trend is summarized in Fig. 6.4, where we report the putative causal links constructed from the normalized IGT values defined in Eq. 2.30. The graph confirms the behavior observed in the time-resolved IG curves: all links involving PC_1 are strongly asymmetric and predominantly oriented from PC_1 toward $\mathrm{PC}_{2,3,4,5}$. By contrast, when PC_2 is taken as the reference mode, the links with $\mathrm{PC}_{3,4,5}$ are largely reciprocal and of comparable magnitude, as shown in Fig. D.10. Taken together, these results identify PC_1 as a dominant slow mode that directionally conditions the dynamics of several faster principal components within the folded state of NTL9.

We again use the Brüschweiler collectivity index [359, 360] to identify the residues that contribute most strongly to each principal component; these residues are highlighted in blue in Fig. 6.4. Additional details for this particular protein are provided in Fig. D.6. This analysis reveals a trend that differs markedly from ubiquitin. In ubiquitin, PC_1 is the most delocalized mode, with contributions distributed across the protein. By contrast, in NTL9, PC_1 is comparatively localized, with a collectivity index of approximately 0.4. The collectivity indices of $\mathrm{PC}_{2,3,4,5}$ are 0.52, 0.38, 0.29, and 0.65, respectively. Structurally, PC_1 is primarily localized around the β-turn region of the protein, and PC_2, PC_3, and PC_4 also involve motions broadly concentrated in this region, with additional secondary contributions. This suggests that the strong information transfer from PC_1 to these modes reflects a local hierarchy of motions, in which a slow fluctuation centered around the β-turn organizes faster fluctuations in the same structural region.

A particularly interesting case is PC_5, which is substantially more delocalized than the other higher-order modes considered here. Despite this difference in spatial character, the $\mathrm{PC}_1 \rightarrow \mathrm{PC}_5$ link remains strongly asymmetric. This observation indicates that directional information transfer between PCs is not necessarily determined by localization or delocalization. In the case of NTL9, a localized slow fluctuation can carry predictive information about a more delocalized mode involving other parts of the protein. This result highlights the nontrivial character of information transfer in the PC representation and suggests that the present framework may provide a useful route for identifying dynamical couplings between motions occurring in distinct regions of a protein, with potential relevance for future studies of allosteric communication.

6.4 Discussion

In this work, we have investigated the presence of putative causal links between principal components extracted from long equilibrium protein trajectories. Principal component analysis is most often used as a dimensionality-reduction tool, where the leading modes are interpreted in terms of variance, collectivity, and structural deformation. Here, we have taken a complementary perspective and asked whether the present value of one PC improves the prediction of the future evolution of another PC. By using the Imbalance Gain approach [15], we obtain a time-lagged and directional measure of information transfer between collective modes. Importantly, the directional information

transfer identified by our approach cannot be identified by the TICA analysis, since the time-delayed covariance matrix between PCs is symmetric if the dynamics satisfies detailed balance, as is the case for molecular dynamics.

The two systems studied here display qualitatively distinct patterns of directional coupling. In ubiquitin, the leading principal component, which is also the slowest one, is a highly delocalized, protein-wide mode and exhibits the strongest information transfer toward PC_2 and PC_3. The influence of PC_1 becomes progressively weaker and more symmetric for higher-order PCs, indicating that in this case the dominant collective fluctuation primarily influences a few leading modes. This behavior is consistent with a picture in which a broad native-state fluctuation provides a slowly varying dynamical context for more localized motions, while lower-variance modes become increasingly weakly coupled to the dominant PC.

In NTL9, PC_1 is clearly the slowest mode and displays a pronounced directional influence on all the other leading PCs considered. Unlike ubiquitin, however, this dominant mode is not the most delocalized one. Instead, PC_1 is localized around the β-turn region, while some of the modes receiving information from it, in particular PC_5, are more spatially extended. This result shows that directional information transfer between PCs is not simply determined by the collectivity or variance of the corresponding modes. A localized slow fluctuation can carry predictive information about faster and more delocalized motions, suggesting that the causal organization of PC space can reveal dynamical relationships that are not apparent from the PCA spectrum or collectivity indices alone.

The relationship between the free-energy landscape and causal coupling remains an important open question. While the FES of the PCs of Ubiquitin and NTL9 do not reveal distinct metastable substates within the folded basin (see Figs. D.11 and D.12), they exhibit different degrees of asymmetry and confinement. In particular, in Ubiquitin , PC_1 displays a slightly more rugged profile, whereas the remaining PCs are characterized by relatively smooth, valley-like free-energy landscapes. Instead, in NTL9 the free energy rises more steeply along PC_1, with no signs of ruggedness. As shown previously [3], asymmetries in the FES can bias dynamical couplings between collective variables. However, these free energy landscapes show that for the systems analysed in this work the landscape by itself is not a sufficient predictor of the asymmetries. Future studies that condition causality analyses on different metastable states, particularly along folding–unfolding landscapes, may provide further insight into how state-dependent relaxation processes govern causal relationships in biomolecular dynamics.

It is important to remark that PCA and the IG between PCs probe two different aspects. PCA identifies directions of large-amplitude structural variance, whereas IG asks whether fluctuations along one such direction improve the prediction of another at later times. The first criterion is geometric and instantaneous, while the second is dynamical and time-lagged. Therefore, the most informative causal driver need not be the PC with the largest variance or the most spatially delocalized eigenvector. In

ubiquitin, these properties approximately coincide: the most collective mode is also the dominant source of directional information transfer. In NTL9, they separate: the slowest and most predictive mode is relatively localized, yet it influences both localized and more distributed fluctuations. This distinction highlights the value of studying causal structure in PC space rather than relying exclusively on variance, relaxation times, or spatial collectivity. In the systems analyzed so far, we observe several cases in which the IGT values are broadly symmetric between pairs of PCs (e.g., between $\mathrm{PC}_2 \Leftrightarrow \mathrm{PC}_{3,4,5}$ in NTL9, see Fig. D.10). However, we have not yet identified a case in which the dominant causal direction is inverted relative to the ordering of explained variance; that is, we do not observe lower-variance PCs acting as stronger causal drivers of higher-variance PCs.

The present results suggest several natural extensions. One direction is to move from pairwise causal graphs toward multivariate formulations in which the influence of a PC is assessed while accounting for the simultaneous effects of other modes. The approach introduced in Ref. [303] might allow us to perform this analysis, with the caveat that the statistical uncertainty of the inferred collective links should be carefully controlled. Performing this analysis would help distinguish whether the observed links correspond to direct dynamical couplings or are mediated by intermediate PCs. A second direction is to connect PC-level causality to residue-level or structural mechanisms, for example by combining IG with mode localization, contact maps, or residue-wise dynamical descriptors. This would be particularly useful for systems such as NTL9, where a localized slow fluctuation appears to influence more extended protein motions.

Our results suggest an interpretation of the directional coupling in terms of the residues that contribute to each PC. In simple terms, there are certain residue fluctuations which can have an asymmetric influence on other residue fluctuations. Our analysis should be considered a first step towards an approach for identifying those residues. The causal asymmetry observed between PCs also suggests a route toward experimentally testable interventions in protein dynamics. Because each PC can be decomposed into residue-level contributions, residues with large contributions to a putative driver PC constitute candidate sites for localized perturbations whose effects may subsequently emerge in the motion described by a driven PC. In this sense, the inferred directionality predicts how perturbing one collective mode may alter the future evolution of another and identifies the timescale over which that response becomes detectable. This perspective could be applied to proteins with known allosteric regulation, ligand binding, or conformational switching, where directional couplings between collective modes may provide a compact description of how local perturbations propagate through the protein [361, 362, 363, 307, 364, 365, 366, 367, 66, 368, 369, 370]. Applied without prior assumptions about known allosteric sites or pathways, the approach could help identifying protein regions that may initiate an allosteric response and characterizing how the resulting dynamical influence propagates through the structure. Similarly, an optical probe placed on specific sites on the protein surface

could provide an experimentally accessible measure of a particular asymmetric influence between different protein regions. Monitoring its response after the photoexcitation of a chromophore would offer a way to test the predicted directional coupling and our approach could guide the selection of perturbation and observation sites in future experiments.

7

Conclusions and perspectives

High-dimensional molecular simulations are easier to interpret mechanistically when their information can be expressed in terms of a small number of physically meaningful variables. This thesis approaches this problem from two complementary directions. The first asks which nuclear motions and intermolecular arrangements are most informative about photoinduced chemical processes. The second asks whether directional relationships can be assigned to the resulting variables and under what conditions those relationships can be interpreted causally. Across both themes tackled in the thesis, distances between configurations provide a practical way to compare information without requiring the construction of a high-dimensional probability density.

The main results of the thesis can be summarised as follows:

- In **Chapter 3**, we use non-adiabatic molecular dynamics (NAMD) simulations to investigate the enhanced non-aromatic fluorescence (NAF) observed experimentally in L-cysteine crystals grown in heavy water (D_2O). Our key finding is that the unique crystallographic packing and resulting hydrogen-bond network in the heavy-water crystal give rise to two complementary effects: they suppress non-radiative decay channels involving specific vibrational degrees of freedom and enhance spin–orbit coupling, thereby promoting intersystem crossing.
- In **Chapter 4**, we develop an information-theoretic feature selection approach using the differentiable information imbalance (DII) to determine which nuclear degrees of freedom explain energy gaps (and therefore non-radiative decay mechanisms) and oscillator strengths from NAMD simulations. We validate this approach on some well-known systems (methaniminium cation and furan) exhibiting varied photochemistry, on systems exhibiting NAF, and then uncover new insights into the non-radiative mechanisms in photoactive molecular motors.
- In **Chapter 5**, we apply the *Imbalance Gain* approach developed by Del Tatto et al. in Ref. [15] within the context of equilibrium simulations satisfying detailed bal-

ance. We uncover some conditions for observing causal relationships in molecular systems, and we devise a protocol for performing interventional experiments to validate these putative causal links. In addition, we explore the emergence of these causal links between structural and energetic variables in MD simulations of a solvated tryptophan molecule.

- In **Chapter 6**, we further apply the causal methodology from the previous chapter to study putative causal relationships between principal components (PCs) of protein MD simulations. PCs are often used to describe protein conformations, and we uncover non-trivial causal relationships between these PCs in the case of two different proteins with millisecond-long MD trajectories: NTL9 and ubiquitin. We find different causal structures between the leading PCs that are not easily explained by the variance of the PCs or their spatial collectivity.

Each chapter of the thesis has opened up several different directions and future perspectives which we will now elaborate on.

For the study presented in **Chapter 3**, and for the field of non-aromatic fluorescence more generally, a central question is how broadly the physical principles inferred from cysteine can be transferred to other systems. Crystal packing controls access to non-radiative channels in this system, but the same S−H and C−N motions need not control relaxation in other amino acids or peptide assemblies. Previous work by our group has identified several mechanisms associated with NAF, including carbonyl stretching, short hydrogen bonds, and proton transfer. Although none appears to be universal, together they could provide a set of design motifs for developing proteins with enhanced intrinsic luminescence using generative protein-design tools [371, 372, 373, 374, 375, 376, 377, 378].

A further limitation of the present theoretical approach concerns the substantial discrepancy, often amounting to 2–3 eV, between the calculated absorption energies and the photon energies corresponding to the experimental excitation wavelengths [379, 380, 381, 382, 383, 384, 385]. Such discrepancies may arise from an incomplete representation of the extended molecular environment, including long-range electrostatic interactions, polarisation effects, and the structural fluctuations present at finite temperature. Indeed, in **Chapter 3** and Appendix A, we demonstrate that systematically increasing the size of the QM region shifts the calculated absorption features towards the experimentally observed energies. This result highlights the importance of treating a sufficiently large portion of the crystal at the quantum-mechanical level and ultimately motivates simulations in which the entire periodic system is described in a quantum-mechanical fashion. However, performing *ab initio* molecular dynamics (AIMD) for large periodic systems over the timescales required for adequate configurational sampling would be computationally prohibitive. Machine-learning interatomic potentials (MLIPs) provide an efficient approximation to the potential-energy surface by learning the relationship between atomic configurations and their corresponding energies and forces from a set of reference electronic-structure calculations [386, 387, 388, 389,

390, 391, 392, 393, 394, 395, 396]. Once trained, they can replace repeated electronic-structure evaluations during molecular dynamics, substantially reducing the computational cost while retaining accuracy close to that of the underlying reference method within the sampled configurational space. For the present problem, MLIPs trained on periodic electronic-structure calculations could therefore enable longer trajectories and larger simulation cells than are accessible with AIMD. Representative configurations obtained from these trajectories could then be subjected to higher-level excited-state calculations, providing a more complete account of both the extended crystal environment and its finite-temperature fluctuations in the predicted absorption spectra.

In addition, the mechanisms underlying the different crystallisation pathways of cysteine in H_2O and D_2O remain an open question. Crystals grown in the two solvents adopt markedly different packing arrangements: monoclinic and orthorhombic, respectively[202, 204, 205, 206, 203, 207]. However, as already noted, closely related orthorhombic cysteine crystals can also be grown in H_2O under different conditions. The differences in the final optical properties are not attributed directly to NQEs because the resulting crystals contain protons rather than deuterons. Nevertheless, crystal nucleation may be affected by the slower diffusion of D_2O relative to H_2O, by isotope-dependent NQEs arising from differences in zero-point energy, or by both effects. The distinct nucleation pathways may therefore have either thermodynamic or kinetic origins. Isotopic substitution in the solvent changes the zero-point energy fluctuations, thereby affecting the structure, dynamics, and thermodynamic properties of the hydrogen-bond network in water [397, 398, 173]. Path-integral molecular dynamics (PIMD), which incorporates NQEs through a ring-polymer representation of the nuclei, could therefore be used to quantify isotope-dependent changes in the relative free energies of cysteine polymorphs and their solvated pre-nucleation structures. Such calculations require multiple replicas of the system and are consequently very expensive when the energies and forces are evaluated directly using electronic-structure methods. Machine-learning interatomic potentials (MLIPs) trained on electronic-structure data [386, 387, 388, 389, 390, 391, 392, 393, 394, 395, 396] could provide the required energies and forces at substantially lower computational cost, enabling PIMD simulations of larger aqueous systems and more extensive configurational sampling [399, 400]. Combined with free-energy calculations and enhanced-sampling methods, this approach could quantify isotope-dependent changes in the relative stability of cysteine polymorphs and pre-nucleation structures, thereby helping to clarify the origins of the different crystallisation outcomes in H_2O and D_2O.

On the computational side, although simulating these systems over nanosecond or longer timescales is unrealistic using current methods, such timescales may become accessible using excited-state MLIPs [401, 402, 403, 404, 405, 406, 407]. Westermayr, Dral, Malosso, and others have made progress in this direction; however, extending MLIP-based approaches to problems of this complexity remains challenging. Another avenue where MLIPs could help with a limitation of our approach is our use of QM/MM-based methods. While these approaches have been shown to work very well in a variety of

systems, they omit some physical effects that could be captured within a fully periodic treatment.

Another priority is to clarify the possible role of triplet states in the observed emission. The NAMD simulations predict substantial intersystem crossing in Cys-D but cover only the first two picoseconds, whereas the measured luminescence occurs on nanosecond timescales and triplet-state emission is generally expected on still longer timescales. Moreover, the calculated oscillator strengths for triplet emission are approximately three orders of magnitude smaller than those for singlet emission, although the experimentally observed emission peaks coincide with the predicted triplet-emission energies. Further experiments will therefore be required to isolate any triplet-state contribution and determine whether the system exhibits phosphorescence.

Next, we move on to the work done in **Chapter 4**. Larger chromophores, molecular aggregates, and chromophores embedded in heterogeneous environments will require descriptors that include both intramolecular deformation and changes due to the response of the environmental bath. Intermolecular contacts, local electric fields, hydrogen-bond geometries, and solvent or protein coordinates may need to be considered alongside bond lengths, angles, and torsions. This is especially relevant for condensed-phase photochemistry, where a coordinate localised on the chromophore may be insufficient even when the electronic transition is formally local.

Beyond energy gaps and oscillator strengths, non-adiabatic coupling vectors (NACVs) could be used as targets to determine which structural features best explain the excited-state behaviour. Because the hopping observed in TSH also depends strongly on the quality of the electronic-structure method, the analysis could be applied to trajectories calculated using multireference methods. The DII-selected coordinates could then be projected onto the branching space, which is spanned by the gradient-difference and non-adiabatic-coupling vectors. Comparing the resulting overlaps would test whether the method recovers the physically relevant directions. This would extend the approach from identifying modes sampled by trajectories undergoing non-radiative decay towards locating conical-intersection regions without prespecifying the relevant coordinates.

Another possible validation of our method would be to construct reduced vibronic Hamiltonians (such as in MCTDH with LVC) from the selected modes for quantum dynamical treatments. One could then test whether dynamics performed in this space reproduce state populations, branching ratios, lifetimes, and spectra. A coordinate that predicts an energy gap over an existing ensemble is not automatically reliable outside that ensemble, so the reduced description should be tested against pathways and observables that were not used to select it.

The second half of this thesis, comprising **Chapter 5** and **Chapter 6**, develops and applies causal approaches to molecular systems. Several associated challenges warrant further investigation and form the subject of ongoing work and will be addressed in future publications.

While our causality analysis in terms of the IG suggests putative causal links, one

of the consistent challenges we have encountered is how to physically interpret the magnitude of the IG. In simple terms, what does an IG of 5% mean? The same issue applies to transfer entropy where the result is reported in bits or nats. What does a TE of 0.2 bits mean physically?

Throughout this thesis, the IG has been reported as a percentage, reflecting its definition as a relative difference. In a dataset of 100 trajectories, an $\mathrm{IG}(\tau) = 1\%$ from **A** to **B** indicates that including **A** in the prediction criterion improves the proximity to a reference realisation by one neighbour order, on average, compared to when **A** is excluded $(\alpha = 0)$. Unlike quantities such as correlation coefficients, the IG does not admit a universal scale on which one can define *small* or *large* values. Instead, its characteristic magnitude depends on the dimensionality and complexity of the underlying dynamical system. A variety of applications of the IG illustrate this trend: values approaching 80% are observed for coupled Rössler dynamical systems [15], approximately 7% for two-dimensional Langevin dynamics [3], 3–6% for atomistic molecular simulations of an amino acid in water [3] and 1–2% [4] between principal components of the dynamics of a folded protein. As the number of interacting degrees of freedom increases, the directional influence becomes distributed across an increasing number of coupled degrees of freedom, reducing the pairwise IG associated with any single descriptor.

An IG analysis supports three physically meaningful interpretations. First, and most importantly, an IG value that is statistically consistent with zero indicates the absence of detectable coupling between the corresponding observables; this interpretation remains valid when the IG is estimated pairwise. Second, changes in the IG for a given set of observables can be meaningfully compared across different physical, dynamical, or thermodynamic conditions. Third, under a fixed set of conditions, differences in the IG values obtained for distinct pairs of observables provide a meaningful measure of their relative coupling strengths. The accompanying causality analysis also suggests two useful heuristics for interpreting collective variables (CVs). The first relates the symmetry or asymmetry of the inferred coupling to the information captured by the CVs in the corresponding free-energy surface (FES): two CVs that resolve the relevant features of the FES to a similar extent may exhibit a relatively symmetric relationship, whereas an asymmetric relationship may indicate that one CV is more informative than the other. The second concerns the decorrelation times of the CVs, since a slowly decorrelating variable retains information for longer and may consequently exhibit a stronger apparent directional influence than a rapidly decorrelating variable. Although these heuristics are consistent with the behaviour observed in several cases, they do not hold universally and should therefore be regarded as empirical diagnostics rather than formal criteria for establishing causality.

Pairwise Imbalance Gain (IG) is useful for screening relationships and identifying asymmetric predictive structure, but it cannot distinguish a direct link from one induced by a common driver or an unobserved confounder. A central challenge is that of *causal sufficiency*: whether all relevant causes have been included in the analysis. A multivariate analysis can determine whether the history of one coordinate contin-

ues to improve prediction after relevant alternative variables have been included [15, 303]. Extending the pairwise IG approach to the multivariate DII-based formulation introduced in Ref. [303] is therefore an important next step. Such a formulation requires us to ensure "causal sufficiency", at least in an approximate sense, since we should be able to condition on every possible causal driver to make sure that no unobserved confounders are responsible for spurious causal links. It could also remove the need to choose the IG variational parameter α manually by learning the optimal causal weights directly. This would make it much more convenient to combine heterogeneous variables with wildly different magnitudes and periodicities without worrying about scaling them.

Causal analysis of molecular trajectories can provide a basis for designing targeted perturbations of selected collective variables (CVs) in MD simulations. By testing how the dynamics respond when a candidate CV is modified, these simulations may help distinguish variables that actively govern a process from those that merely fluctuate alongside it, thereby identifying potential control points for subsequent experiments. Protein allostery provides a representative example in which directional couplings between collective modes may provide a compact description of how local perturbations propagate through the protein [361, 363, 307, 364, 365, 366, 367, 368, 369, 370]. Abruptly fixing or assigning a CV to an arbitrary value, however, may generate physically inaccessible configurations or violate the geometrical constraints linking it to other molecular degrees of freedom. *Softer* interventions, such as restrained ensembles, biased potentials, or controlled modifications of the forces acting on selected coordinates, offer a more physically realistic alternative and can be implemented naturally within molecular dynamics simulations [408].

The issues that have been highlighted above pertaining to causal sufficiency, pairwise versus multivariate estimates, and interventions apply to the work done in **Chapter 6** as well. The analysis of PCs in ubiquitin and NTL9 suggests a practical setting in which these ideas can be developed. PCs form an orthogonal basis, but they are not dynamically independent. The pairwise IG asymmetries suggest a dynamical organisation that is absent from the variance spectrum alone. The next question is how much of that organisation survives a change of basis and how it maps back to specific residues, contacts, and secondary-structure elements. Conditional analysis can test whether a link between two components is direct or transmitted through another mode. Molecular simulations yield free-energy surfaces with different minima and possible transition paths between them. A conditional IG analysis could be restricted to different minima or different transition states to allow one to understand how causal links change depending on the initial condition one explores. Residue-wise decompositions can also identify the structural regions that carry the influence. Nonlinear coordinates may be required when a conformational transition cannot be represented adequately by a linear combination of Cartesian displacements. Proteins with known allosteric regulation offer an especially useful test because the putative source and response can be spatially separated, which makes both interpretation and interventional

experiments more tractable [361, 368, 345].

The concept of a *driver–response* relationship is ubiquitous in physical chemistry. We are investigating the causal links and CVs that govern protein folding and unfolding, interfacial photochemical reactions and their associated electric fields, and solvent relaxation following photoexcitation in biomolecules. The final topic concerns the theoretical modelling of TDFSS experiments, which were introduced in **Chapter 1** as a motivation for studying causal relationships in soft-matter systems.

The causal framework developed in this thesis provides an opportune moment to revisit TDFSS through the language of causality. To briefly recap, the measured relaxation follows a sudden change in the charge distribution of a fluorophore, but the signal reports on the combined response of the fluorophore, the surrounding protein, and the solvent. Decomposing the energy gap into protein and water contributions is informative, but it does not by itself establish which component drives the other. Nor does agreement between computing the Stokes shift from an equilibrium correlation function and from the non-equilibrium relaxation prove that the same directional organisation is present in both ensembles. Photoexcitation is a controlled perturbation (or one might consider it an *intervention*) to the chromophore's electronic state.

The analysis of tryptophan in water presented in **Chapter 5** provides a proof of principle that information-theoretic approaches can resolve directionality within coupled chromophore–solvent fluctuations, revealing a directional relationship from the chromophore energetics to the chromophore–solvent interactions under equilibrium conditions. The broader significance of this result lies in the possibility of extending the same framework to the long-standing question of directionality in coupled protein–water motions: namely, under what conditions protein fluctuations drive the surrounding hydration environment, when water instead drives protein motion, and how this balance depends on the local chemical and structural environment. Extending the analysis to explicitly non-equilibrium trajectories initiated by photoexcitation would connect the inferred directional relationships more closely to experimental relaxation measurements and determine whether the relationships identified from stationary equilibrium fluctuations persist when the system is driven. Finally, comparing these non-equilibrium responses with predictions obtained from equilibrium energy-gap fluctuations would provide a direct test of linear-response theory, establishing when the equilibrium ensemble adequately represents photoinduced relaxation and when excitation accesses configurations or pathways that are not captured by spontaneous equilibrium fluctuations [409, 410]. Together, these developments would establish a general framework for identifying the microscopic directionality underlying coupled solute, solvent, and biomolecular dynamics.

Taken together, this thesis establishes a unified information-theoretic framework for identifying the molecular coordinates that govern complex dynamics and for resolving the directional relationships between them. In doing so, it lays the groundwork for a deeper mechanistic understanding of molecular processes from excited-state photochemistry to biomolecular dynamics.

This page intentionally left blank.

Bibliography

[1] Debarshi Banerjee et al. “Crystallization of L-Cysteine in Heavy Water Induces Intrinsic Fluorescence”. en. In: *Angewandte Chemie International Edition* 64.29 (2025), e202505331. DOI: 10.1002/anie.202505331.

[2] Debarshi Banerjee et al. “Machine learning the non-radiative decay modes in photochemical processes”. In: *arXiv preprint arXiv:2605.08381* (2026).

[3] Vittorio Del Tatto et al. “Towards a robust approach to infer causality from molecular dynamics simulations”. en. In: *The Journal of Chemical Physics* 162.24 (2025), p. 244105. DOI: 10.1063/5.0267926.

[4] Debarshi Banerjee, Ali Hassanali, and Alessandro Laio. “Investigating causality between principal components in protein dynamics”. In: *The Journal of Chemical Physics* 165.9 (2026), p. 094122. DOI: 10.1063/5.0349449.

[5] Debarshi Banerjee et al. “Aqueous solution chemistry in silico and the role of data-driven approaches”. In: *Chemical Physics Reviews* 5.2 (2024). DOI: 10.1063/5.0207567.

[6] Gonzalo Diáz Miroń et al. “Exploring the Mechanisms behind Non-aromatic Fluorescence with the Density Functional Tight Binding Method”. In: *Journal of Chemical Theory and Computation* 20.9 (2024), pp. 3864–3878. https://doi.org/10.1021/acs.jctc.4c00125.

[7] Gonzalo Diáz Miroń et al. “Non-adiabatic couplings in surface hopping with tight binding density functional theory: The case of molecular motors”. In: *Journal of Chemical Theory and Computation* 20.23 (2024), pp. 10602–10614. https://doi.org/10.1021/acs.jctc.4c01263.

[8] Solana Di Pino et al. “Deconstructing the origins of interfacial catalysis: Why electric fields are inseparable from solvation”. en. In: *The Journal of Chemical Physics* 163.18 (2025), p. 184505. DOI: 10.1063/5.0288327.

[9] Marta Monti et al. “Investigating the Role of pH and Counterions in the Intrinsic Fluorescence of Solid-State l -Lysine”. en. In: *The Journal of Physical Chemistry B* 129.51 (2025), pp. 13105–13116. DOI: 10.1021/acs.jpcb.5c05756.

[10] Vignesh Athiyarath et al. “Water-mediated reconfigurable topology and mechanics in porous peptide materials”. In: *Matter* 9.4 (2026). DOI: 10.1016/j.matt.2026.102669.

[11] Leon Huet et al. “Causality in Liquid Water as a Hallmark of Emergent Glassy Dynamics”. In: *arXiv* arXiv:2604.19491 (2026). arXiv:2604.19491. DOI: 10.48550/arXiv.2604.19491.

[12] Facundo Tarasi et al. “Hydrogen Detachment in Diprotic Aromatics: O–H versus N–H Photodissociation in m-Aminophenol”. In: *The Journal of Physical Chemistry Letters* (2026). DOI: 10.1021/acs.jpclett.6c02112.

[13] Aldo Glielmo et al. “Ranking the information content of distance measures”. In: *PNAS Nexus* 1.2 (2022), pgac039. DOI: 10.1093/pnasnexus/pgac039.

[14] Romina Wild et al. “Automatic feature selection and weighting in molecular systems using Differentiable Information Imbalance”. en. In: *Nature Communications* 16.1 (2025), p. 270. DOI: 10.1038/s41467-024-55449-7.

[15] Vittorio Del Tatto et al. “Robust inference of causality in high-dimensional dynamical processes from the Information Imbalance of distance ranks”. In: *Proceedings of the National Academy of Sciences* 121.19 (2024), e2317256121. DOI: 10.1073/pnas.2317256121.

[16] Cornelis R Ronda. *Luminescence: from theory to applications*. John Wiley & Sons, 2007.

[17] Stephen B Mende. “Observing the magnetosphere through global auroral imaging: 2. Observing techniques”. In: *Journal of Geophysical Research: Space Physics* 121.10 (2016), pp. 10–638.

[18] Steven HD Haddock, Mark A Moline, and James F Case. “Bioluminescence in the sea”. In: *Annual review of marine science* 2 (2010), pp. 443–493.

[19] Joseph R Lakowicz. *Principles of fluorescence spectroscopy*. Springer, 2006.

[20] Bernard Valeur and Mário Nuno Berberan-Santos. *Molecular fluorescence: principles and applications*. John Wiley & Sons, 2013.

[21] A Prasanna De Silva et al. “Signaling recognition events with fluorescent sensors and switches”. In: *Chemical reviews* 97.5 (1997), pp. 1515–1566.

[22] Roger Y Tsien. “The green fluorescent protein”. In: *Annual review of biochemistry* 67.1 (1998), pp. 509–544.

[23] Robert B Macgregor Jr and Gregorio Weber. “Fluorophores in polar media: spectral effects of the Langevin distribution of electrostatic interactions”. In: *Annals of the New York Academy of Sciences* 366.1 (1981), pp. 140–154.

[24] Franklyn G Prendergast et al. "Synthesis, spectral properties, and use of 6-acryloyl-2-dimethylaminonaphthalene (Acrylodan). A thiol-selective, polarity-sensitive fluorescent probe." In: *Journal of Biological Chemistry* 258.12 (1983), pp. 7541–7544.

[25] Mario Barbatti, Adélia J. A. Aquino, and Hans Lischka. "Ultrafast two-step process in the non-adiabatic relaxation of the CH2 molecule". en. In: *Molecular Physics* 104.5–7 (2006), pp. 1053–1060. DOI: 10.1080/00268970500417945.

[26] Aaron M. Virshup, Jiahao Chen, and Todd J. Martínez. "Nonlinear dimensionality reduction for nonadiabatic dynamics: The influence of conical intersection topography on population transfer rates". en. In: *The Journal of Chemical Physics* 137.22 (2012), 22A519. DOI: 10.1063/1.4742066.

[27] Amar BT Ghisaidoobe and Sang J Chung. "Intrinsic tryptophan fluorescence in the detection and analysis of proteins: a focus on Förster resonance energy transfer techniques". In: *International journal of molecular sciences* 15.12 (2014), pp. 22518–22538.

[28] Anshuman Shukla et al. "A novel UV laser-induced visible blue radiation from protein crystals and aggregates: scattering artifacts or fluorescence transitions of peptide electrons delocalized through hydrogen bonding?" In: *Archives of Biochemistry and Biophysics* 428.2 (2004), pp. 144–153.

[29] Loretta Laureana Del Mercato et al. "Charge transport and intrinsic fluorescence in amyloid-like fibrils". In: *Proceedings of the National Academy of Sciences* 104.46 (2007), pp. 18019–18024.

[30] Dorothea Pinotsi et al. "Proton transfer and structure-specific fluorescence in hydrogen bond-rich protein structures". In: *Journal of the American Chemical Society* 138.9 (2016), pp. 3046–3057.

[31] Yang Yu et al. "Oligosaccharides self-assemble and show intrinsic optical properties". In: *Journal of the American Chemical Society* 141.12 (2019), pp. 4833–4838.

[32] Zohar A. Arnon et al. "On-off transition and ultrafast decay of amino acid luminescence driven by modulation of supramolecular packing". In: *iScience* 24.7 (2021), p. 102695. DOI: 10.1016/j.isci.2021.102695.

[33] Donald A Tomalia et al. "Non-traditional intrinsic luminescence: inexplicable blue fluorescence observed for dendrimers, macromolecules and small molecular structures lacking traditional/conventional luminophores". In: *Progress in Polymer Science* 90 (2019), pp. 35–117.

[34] Saixing Tang et al. "Nonconventional luminophores: characteristics, advancements and perspectives". In: *Chemical Society Reviews* 50.22 (2021), pp. 12616–12655.

[35] Uriel N. Morzan et al. "Non-aromatic fluorescence in biological matter: the exception or the rule?" In: *The Journal of Physical Chemistry B* 126.38 (2022), pp. 7203–7211.

[36] Luca Grisanti et al. "Toward Understanding Optical Properties of Amyloids: A Reaction Path and Nonadiabatic Dynamics Study". en. In: *Journal of the American Chemical Society* 142.42 (2020), pp. 18042–18049. DOI: 10.1021/jacs.0c07134.

[37] Amberley D. Stephens et al. "Short hydrogen bonds enhance nonaromatic protein-related fluorescence". en. In: *Proceedings of the National Academy of Sciences* 118.21 (2021), e2020389118. DOI: 10.1073/pnas.2020389118.

[38] Gonzalo Díaz Mirón et al. "The carbonyl-lock mechanism underlying non-aromatic fluorescence in biological matter". en. In: *Nature Communications* 14.1 (2023), p. 7325. DOI: 10.1038/s41467-023-42874-3.

[39] Yuning Hong, Jacky WY Lam, and Ben Zhong Tang. "Aggregation-induced emission: phenomenon, mechanism and applications". In: *Chemical Communications* 29 (2009), pp. 4332–4353.

[40] Qian Wang et al. "Reevaluating Protein Photoluminescence: Remarkable Visible Luminescence upon Concentration and Insight into the Emission Mechanism". en. In: *Angewandte Chemie International Edition* 58.36 (2019), pp. 12667–12673. DOI: 10.1002/anie.201906226.

[41] Jie Yang, Manman Fang, and Zhen Li. "Organic luminescent materials: the concentration on aggregates from aggregation-induced emission". In: *Aggregate* 1.1 (2020), pp. 6–18.

[42] Pai Liu et al. "MDM2-Associated Clusterization-Triggered Emission and Apoptosis Induction Effectuated by a Theranostic Spiropolymer". In: *Angewandte Chemie International Edition* 132.22 (2020), pp. 8513–8517.

[43] Zheng Zhao et al. "Aggregation-induced emission: new vistas at the aggregate level". In: *Angewandte Chemie International Edition* 59.25 (2020), pp. 9888–9907.

[44] Shuyuan Zheng et al. "Accessing tunable afterglows from highly twisted nonaromatic organic AIEgens via effective through-space conjugation". In: *Angewandte Chemie International Edition* 59.25 (2020), pp. 10018–10022.

[45] Yangbin Xie et al. "Cluster-triggered excitation-dependent phosphorescence emission in polymorphic arylborinic acid". In: *Inorganic Chemistry Communications* 168 (2024), p. 112895.

[46] Dorothea Pinotsi et al. "A label-free, quantitative assay of amyloid fibril growth based on intrinsic fluorescence". In: *ChemBioChem* 14.7 (2013), p. 846.

[47] Saumya Prasad et al. “Near UV-Visible electronic absorption originating from charged amino acids in a monomeric protein”. In: *Chemical Science* 8.8 (2017), pp. 5416–5433.

[48] Amrendra Kumar et al. “Weak intrinsic luminescence in monomeric proteins arising from charge recombination”. In: *Biophysical Journal* 118.3 (2020), 468a.

[49] Sunil Varughese. “Non-covalent routes to tune the optical properties of molecular materials”. In: *Journal of Materials Chemistry C* 2.18 (2014), pp. 3499–3516.

[50] Damir A Safin et al. “Polymorphism driven optical properties of an anil dye”. In: *CrystEngComm* 18.38 (2016), pp. 7249–7259.

[51] Jie Yang et al. “The influence of the molecular packing on the room temperature phosphorescence of purely organic luminogens”. In: *Nature communications* 9.1 (2018), p. 840.

[52] Zhongfu An et al. “Stabilizing triplet excited states for ultralong organic phosphorescence”. In: *Nature Materials* 14.7 (2015), pp. 685–690.

[53] G. Capano et al. “Photophysics of a copper phenanthroline elucidated by trajectory and wavepacket-based quantum dynamics: a synergetic approach”. en. In: *Physical Chemistry Chemical Physics* 19.30 (2017), pp. 19590–19600. DOI: 10.1039/C7CP00436B.

[54] Andrew J. Atkins and Leticia González. “Trajectory Surface-Hopping Dynamics Including Intersystem Crossing in [Ru(bpy)3]2+”. en. In: *The Journal of Physical Chemistry Letters* 8.16 (2017), pp. 3840–3845. DOI: 10.1021/acs.jpclett.7b01479.

[55] Sebastian Mai and Leticia González. “Identification of important normal modes in nonadiabatic dynamics simulations by coherence, correlation, and frequency analyses”. en. In: *The Journal of Chemical Physics* 151.24 (2019), p. 244115. DOI: 10.1063/1.5129335.

[56] Jiawei Peng et al. “Analysis of bath motion in MM-SQC dynamics via dimensionality reduction approach: Principal component analysis”. en. In: *The Journal of Chemical Physics* 154.9 (2021), p. 094122. DOI: 10.1063/5.0039743.

[57] Yifei Zhu et al. “The principal component analysis of the ring deformation in the nonadiabatic surface hopping dynamics”. en. In: *Physical Chemistry Chemical Physics* 24.39 (2022), pp. 24362–24382. DOI: 10.1039/D2CP03323B.

[58] Pavel S. Rukin et al. “Complementing Adiabatic and Nonadiabatic Methods To Understand Internal Conversion Dynamics in Porphyrin Derivatives”. en. In: *Journal of Chemical Theory and Computation* 20.24 (2024), pp. 10759–10769. DOI: 10.1021/acs.jctc.4c00698.

[59] David R. Yarkony. "Diabolical conical intersections". en. In: *Reviews of Modern Physics* 68.4 (1996), pp. 985–1013. DOI: 10.1103/RevModPhys.68.985.

[60] Wolfgang Domcke and David R. Yarkony. "Role of Conical Intersections in Molecular Spectroscopy and Photoinduced Chemical Dynamics". en. In: *Annual Review of Physical Chemistry* 63.1 (2012), pp. 325–352. DOI: 10.1146/annurev-physchem-032210-103522.

[61] Spiridoula Matsika. "Electronic Structure Methods for the Description of Nonadiabatic Effects and Conical Intersections". en. In: *Chemical Reviews* 121.15 (2021), pp. 9407–9449. DOI: 10.1021/acs.chemrev.1c00074.

[62] C. W. J. Granger. "Investigating Causal Relations by Econometric Models and Cross-spectral Methods". In: *Econometrica* 37.3 (1969), pp. 424–438. DOI: 10.2307/1912791. (Visited on 11/07/2022).

[63] Thomas Schreiber. "Measuring Information Transfer". In: *Phys. Rev. Lett.* 85 (2 2000), pp. 461–464. DOI: 10.1103/PhysRevLett.85.461.

[64] A. Gorecki, J. Trylska, and B. Lesyng. "Causal relations in molecular dynamics from the multi-variate autoregressive model". In: *Europhysics Letters* 75.3 (2006), p. 503. DOI: 10.1209/epl/i2006-10129-2.

[65] Hiqmet Kamberaj and Arjan van der Vaart. "Extracting the Causality of Correlated Motions from Molecular Dynamics Simulations". In: *Biophysical Journal* 97.6 (2009), pp. 1747–1755. DOI: https://doi.org/10.1016/j.bpj.2009.07.019.

[66] Aysima Hacisuleyman and Burak Erman. "Entropy Transfer between Residue Pairs and Allostery in Proteins: Quantifying Allosteric Communication in Ubiquitin". In: *PLOS Computational Biology* 13.1 (2017), pp. 1–23. DOI: 10.1371/journal.pcbi.1005319.

[67] Aysima Hacisuleyman and Burak Erman. "Causality, transfer entropy, and allosteric communication landscapes in proteins with harmonic interactions". In: *Proteins: Structure, Function, and Bioinformatics* 85.6 (2017), pp. 1056–1064. DOI: https://doi.org/10.1002/prot.25272.

[68] Tim Hempel, Nuria Plattner, and Frank Noé. "Coupling of Conformational Switches in Calcium Sensor Unraveled with Local Markov Models and Transfer Entropy". In: *Journal of Chemical Theory and Computation* 16.4 (2020), pp. 2584–2593. DOI: 10.1021/acs.jctc.0c00043.

[69] Yifei Qi and Wonpil Im. "Quantification of Drive-Response Relationships Between Residues During Protein Folding". In: *Journal of Chemical Theory and Computation* 9.8 (2013), pp. 3799–3805. DOI: 10.1021/ct4002784.

[70] Liqun Zhang, Thomas Centa, and Matthias Buck. "Structure and Dynamics Analysis on Plexin-B1 Rho GTPase Binding Domain as a Monomer and Dimer". In: *The Journal of Physical Chemistry B* 118.26 (2014), pp. 7302–7311. DOI: 10.1021/jp503668k.

[71] Marcin Sobieraj and Piotr Setny. "Granger Causality Analysis of Chignolin Folding". In: *Journal of Chemical Theory and Computation* 18.3 (2022), pp. 1936–1944. DOI: 10.1021/acs.jctc.1c00945.

[72] Sutapa Dutta, Mahua Ghosh, and Jaydeb Chakrabarti. "Spatio-temporal coordination among functional residues in protein". In: *Scientific Reports* 7 (2017), p. 40439. DOI: 10.1038/srep40439.

[73] Juan R. Perilla and Thomas B. Woolf. "Towards the prediction of order parameters from molecular dynamics simulations in proteins". In: *The Journal of Chemical Physics* 136.16 (2012), p. 164101. DOI: 10.1063/1.3702447.

[74] Patricia Saura et al. "Electric fields control water-gated proton transfer in cytochrome c oxidase". In: *Proceedings of the National Academy of Sciences* 119.38 (2022). Publisher: Proceedings of the National Academy of Sciences, e2207761119. DOI: 10.1073/pnas.2207761119. (Visited on 03/16/2026).

[75] Ali Hassanali et al. "Proton transfer through the water gossamer". In: *Proceedings of the National Academy of Sciences* 110.34 (2013). Publisher: Proceedings of the National Academy of Sciences, pp. 13723–13728. DOI: 10.1073/pnas.1306642110. (Visited on 01/14/2026).

[76] Bradley J. Siwick and Huib J. Bakker. "On the Role of Water in Intermolecular Proton-Transfer Reactions". In: *J. Am. Chem. Soc.* 129.44 (2007). Publisher: American Chemical Society, pp. 13412–13420. DOI: 10.1021/ja069265p. (Visited on 03/16/2026).

[77] Young Min Rhee and Vijay S. Pande. "Solvent Viscosity Dependence of the Protein Folding Dynamics". In: *J. Phys. Chem. B* 112.19 (2008). Publisher: American Chemical Society, pp. 6221–6227. DOI: 10.1021/jp076301d. (Visited on 02/27/2026).

[78] H. Frauenfelder et al. "Protein folding is slaved to solvent motions". In: *Proceedings of the National Academy of Sciences* 103.42 (2006). Publisher: Proceedings of the National Academy of Sciences, pp. 15469–15472. DOI: 10.1073/pnas.0607168103. (Visited on 02/27/2026).

[79] Del Lucent, V. Vishal, and Vijay S. Pande. "Protein folding under confinement: A role for solvent". In: *Proceedings of the National Academy of Sciences* 104.25 (2007). Publisher: Proceedings of the National Academy of Sciences, pp. 10430–10434. DOI: 10.1073/pnas.0608256104. (Visited on 02/27/2026).

[80] Ruhong Zhou et al. "Hydrophobic Collapse in Multidomain Protein Folding". In: *Science* 305.5690 (2004). Publisher: American Association for the Advancement of Science, pp. 1605–1609. DOI: 10.1126/science.1101176. (Visited on 02/19/2026).

[81] Pu Liu et al. "Observation of a dewetting transition in the collapse of the melittin tetramer". en. In: *Nature* 437.7055 (2005). Publisher: Nature Publishing Group, pp. 159–162. DOI: 10.1038/nature03926. (Visited on 02/19/2026).

[82] Manoj V. Athawale et al. "Effects of lengthscales and attractions on the collapse of hydrophobic polymers in water". In: *Proceedings of the National Academy of Sciences* 104.3 (2007), pp. 733–738. DOI: 10.1073/pnas.0605139104.

[83] Tanping Li et al. "Hydration Dynamics and Time Scales of Coupled Water–Protein Fluctuations". English. In: *Journal of the American Chemical Society* 129.11 (2007), pp. 3376–3382. DOI: 10.1021/ja0685957.

[84] Mark Maroncelli and Graham R. Fleming. "Computer simulation of the dynamics of aqueous solvation". English. In: *The Journal of Chemical Physics* 89.8 (1988), pp. 5044–5069. DOI: 10.1063/1.455649.

[85] Emily A. Carter and James T. Hynes. "Solvation dynamics for an ion pair in a polar solvent: Time-dependent fluorescence and photochemical charge transfer". English. In: *The Journal of Chemical Physics* 94.9 (1991), pp. 5961–5979. DOI: 10.1063/1.460431.

[86] Peter J. Rossky and John D. Simon. "Dynamics of chemical processes in polar solvents". English. In: *Nature* 370.6487 (1994), pp. 263–269. DOI: 10.1038/370263a0.

[87] Samir Kumar Pal et al. "Biological Water: Femtosecond Dynamics of Macromolecular Hydration". In: *The Journal of Physical Chemistry B* 106.48 (2002), pp. 12376–12395. DOI: 10.1021/jp0213506.

[88] Michel Vincent et al. "Nanosecond Fluorescence Dynamic Stokes Shift of Tryptophan in a Protein Matrix". English. In: *The Journal of Physical Chemistry B* 104.47 (2000), pp. 11286–11295. DOI: 10.1021/jp000638x.

[89] Jorge Peon, Samir Kumar Pal, and Ahmed H. Zewail. "Hydration at the surface of the protein Monellin: Dynamics with femtosecond resolution". English. In: *Proceedings of the National Academy of Sciences* 99.17 (2002), pp. 10964–10969. DOI: 10.1073/pnas.162366099.

[90] Jianhua Xu et al. "Ultrafast Fluorescence Dynamics of Tryptophan in the Proteins Monellin and IIA Glc". English. In: *Journal of the American Chemical Society* 128.4 (2006), pp. 1214–1221. DOI: 10.1021/ja055746h.

[91] Jianhua Xu et al. "Picosecond Fluorescence Dynamics of Tryptophan and 5-Fluorotryptophan in Monellin: Slow Water–Protein Relaxation Unmasked". English. In: *The Journal of Physical Chemistry B* 119.11 (2015), pp. 4230–4239. DOI: 10.1021/acs.jpcb.5b01651.

[92] Charles C David and Donald J Jacobs. "Principal component analysis: a method for determining the essential dynamics of proteins". In: *Protein dynamics: Methods and protocols*. Springer, 2013, pp. 193–226.

[93] Akio Kitao. "Principal component analysis and related methods for investigating the dynamics of biological macromolecules". In: *j* 5.2 (2022), pp. 298–317.

[94] Paul Adrien Maurice Dirac. *The principles of quantum mechanics*. 27. Oxford university press, 1981.

[95] Max Born and W Heisenberg. "Zur quantentheorie der molekeln". In: *Original Scientific Papers Wissenschaftliche Originalarbeiten* (1985), pp. 216–246.

[96] Loup Verlet. "Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules". In: *Physical Review* 159.1 (1967), p. 98.

[97] Hans C Andersen. "Molecular dynamics simulations at constant pressure and/or temperature". In: *The Journal of chemical physics* 72.4 (1980), pp. 2384–2393.

[98] Nosé Shuichi. "Constant temperature molecular dynamics methods". In: *Progress of Theoretical Physics Supplement* 103 (1991), pp. 1–46.

[99] Giovanni Bussi, Davide Donadio, and Michele Parrinello. "Canonical sampling through velocity rescaling". In: *The Journal of Chemical Physics* 126.1 (2007).

[100] Herman JC Berendsen et al. "Molecular dynamics with coupling to an external bath". In: *The Journal of chemical physics* 81.8 (1984), pp. 3684–3690.

[101] Michele Parrinello and Aneesur Rahman. "Polymorphic transitions in single crystals: A new molecular dynamics method". In: *Journal of Applied physics* 52.12 (1981), pp. 7182–7190.

[102] Mattia Bernetti and Giovanni Bussi. "Pressure control using stochastic cell rescaling". English. In: *The Journal of Chemical Physics* 153.11 (2020), p. 114107. DOI: 10.1063/5.0020514.

[103] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman JC Berendsen. "Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes". In: *Journal of computational physics* 23.3 (1977), pp. 327–341.

[104] Shuichi Miyamoto and Peter A. Kollman. "Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models". English. In: *Journal of*

Computational Chemistry 13.8 (1992), pp. 952–962. DOI: 10.1002/jcc.540130805.

[105] Hans C Andersen. "Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations". In: *Journal of computational Physics* 52.1 (1983), pp. 24–34.

[106] PJAP Ewald. "Evaluation of optical and electrostatic lattice potentials". In: *Ann. Phys.* 64 (1921), pp. 253–287.

[107] Tom Darden, Darrin York, and Lee Pedersen. "Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems". English. In: *The Journal of Chemical Physics* 98.12 (1993), pp. 10089–10092. DOI: 10.1063/1.464397.

[108] Ulrich Essmann et al. "A smooth particle mesh Ewald method". English. In: *The Journal of Chemical Physics* 103.19 (1995), pp. 8577–8593. DOI: 10.1063/1.470117.

[109] Arieh Warshel and Michael Levitt. "Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme". In: *Journal of Molecular Biology* 103.2 (1976), pp. 227–249.

[110] Elizabeth Brunk and Ursula Rothlisberger. "Mixed quantum mechanical/-molecular mechanical molecular dynamics simulations of biological systems in ground and electronically excited states". In: *Chemical Reviews* 115.12 (2015), pp. 6217–6263.

[111] Dirk Bakowies and Walter Thiel. "Hybrid models for combined quantum mechanical and molecular mechanical approaches". In: *The Journal of Physical Chemistry* 100.25 (1996), pp. 10580–10594.

[112] Martin J Field, Paul A Bash, and Martin Karplus. "A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations". In: *Journal of computational chemistry* 11.6 (1990), pp. 700–733.

[113] Hans Martin Senn and Walter Thiel. "QM/MM Methods for Biomolecular Systems". en. In: *Angewandte Chemie International Edition* 48.7 (2009), pp. 1198–1229. DOI: 10.1002/anie.200802019.

[114] Garima Jindal and Arieh Warshel. "Exploring the Dependence of QM/MM Calculations of Enzyme Catalysis on the Size of the QM Region". en. In: *The Journal of Physical Chemistry B* 120.37 (2016), pp. 9913–9921. DOI: 10.1021/acs.jpcb.6b07203.

[115] Martin Richter et al. "SHARC: ab initio molecular dynamics with surface hopping in the adiabatic representation including arbitrary couplings". In: *Journal of Chemical Theory and Computation* 7.5 (2011), pp. 1253–1258.

[116] Sebastian Mai, Philipp Marquetand, and Leticia González. "Nonadiabatic dynamics: The SHARC approach". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 8.6 (2018), e1370.

[117] John C Tully. "Molecular dynamics with electronic transitions". In: *The Journal of Chemical Physics* 93.2 (1990), pp. 1061–1071.

[118] Mario Barbatti. "Nonadiabatic dynamics with trajectory surface hopping method". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1.4 (2011), pp. 620–633.

[119] João Pedro Malhado, Michael J Bearpark, and James T Hynes. "Non-adiabatic dynamics close to conical intersections and the surface hopping perspective". In: *Frontiers in chemistry* 2 (2014), p. 97.

[120] Maurizio Persico and Giovanni Granucci. "An overview of nonadiabatic dynamics simulations methods, with focus on the direct approach versus the fitting of potential energy surfaces". In: *Theoretical Chemistry Accounts* 133 (2014), pp. 1–28.

[121] Rachel Crespo-Otero and Mario Barbatti. "Recent advances and perspectives on nonadiabatic mixed quantum–classical dynamics". In: *Chemical Reviews* 118.15 (2018), pp. 7026–7068.

[122] Giovanni Granucci and Maurizio Persico. "Critical appraisal of the fewest switches algorithm for surface hopping". In: *The Journal of Chemical Physics* 126.13 (2007).

[123] Matthias Rupp et al. "Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning". en. In: *Physical Review Letters* 108.5 (2012), p. 058301. DOI: 10.1103/PhysRevLett.108.058301.

[124] Edward Danquah Donkor, Alessandro Laio, and Ali Hassanali. "Do Machine-Learning Atomic Descriptors and Order Parameters Tell the Same Story? The Case of Liquid Water". en. In: *Journal of Chemical Theory and Computation* 19.14 (2023), pp. 4596–4605. DOI: 10.1021/acs.jctc.2c01205.

[125] Edward Danquah Donkor et al. "Beyond Local Structures in Critical Supercooled Water through Unsupervised Learning". en. In: *The Journal of Physical Chemistry Letters* 15.15 (2024), pp. 3996–4005. DOI: 10.1021/acs.jpclett.4c00383.

[126] Romina Wild et al. "Maximally informative feature selection using Information Imbalance: Application to COVID-19 severity prediction". en. In: *Scientific Reports* 14.1 (2024), p. 10744. DOI: 10.1038/s41598-024-61334-6.

[127] Aldo Glielmo et al. "DADApy: Distance-based analysis of data-manifolds in Python". In: *Patterns* 3.10 (2022).

[128] Milan Palu et al. "Synchronization as adjustment of information rates: Detection from bivariate time series". In: *Phys. Rev. E* 63 (4 2001), p. 046211. DOI: 10.1103/PhysRevE.63.046211.

[129] Liam Paninski. "Estimation of Entropy and Mutual Information". In: *Neural Computation* 15.6 (2003), pp. 1191–1253. DOI: 10.1162/089976603321780272.

[130] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. "Estimating mutual information". In: *Phys. Rev. E* 69 (6 2004), p. 066138. DOI: 10.1103/PhysRevE.69.066138.

[131] Brian C. Ross. "Mutual Information between Discrete and Continuous Data Sets". In: *PLOS ONE* 9.2 (2014), pp. 1–5. DOI: 10.1371/journal.pone.0087357.

[132] Damiano Lombardi and Sanjay Pant. "Nonparametric k-nearest-neighbor entropy estimator". In: *Phys. Rev. E* 93 (1 2016), p. 013310. DOI: 10.1103/PhysRevE.93.013310.

[133] Octavio César Mesner and Cosma Rohilla Shalizi. "Conditional Mutual Information Estimation for Mixed, Discrete and Continuous Data". In: *IEEE Transactions on Information Theory* 67.1 (2021), pp. 464–484. DOI: 10.1109/TIT.2020.3024886.

[134] Katerina Hlaváčková-Schindler et al. "Causality detection based on information-theoretic approaches in time series analysis". In: *Physics Reports* 441.1 (2007), pp. 1–46. DOI: https://doi.org/10.1016/j.physrep.2006.12.004.

[135] Sachchidanand Soaham Gupta et al. "Amino acid derived biopolymers: Recent advances and biomedical applications". In: *International Journal of Biological Macromolecules* 188 (2021), pp. 542–567.

[136] Santu Bera et al. "Self-assembly of aromatic amino acid enantiomers into supramolecular materials of high rigidity". In: *ACS Nano* 14.2 (2020), pp. 1694–1706.

[137] Houman Kholafazad Kordasht et al. "Poly (amino acids) towards sensing: Recent progress and challenges". In: *TrAC Trends in Analytical Chemistry* 140 (2021), p. 116279.

[138] Keyu Geng et al. "Covalent organic frameworks: design, synthesis, and functions". In: *Chemical Reviews* 120.16 (2020), pp. 8814–8933.

[139] Xue-Qing Zhao et al. "Fabrication of amino acid-based supramolecular hydrogel with silver ions for improved antibacterial properties". In: *Materials Letters* 300 (2021), p. 130161.

[140] Bin Fei et al. "Enhanced fluorescence and thermal sensitivity of polyethylenimine modified by Michael addition". In: *Polymer* 51.8 (2010), pp. 1845–1852.

[141] Shoujun Zhu et al. “Non-conjugated polymer dots with crosslink-enhanced emission in the absence of fluorophore units”. In: *Angewandte Chemie International Edition* 54.49 (2015), pp. 14626–14637.

[142] Maria J Jasmine, Manniledam Kavitha, and Edamana Prasad. “Effect of solvent-controlled aggregation on the intrinsic emission properties of PAMAM dendrimers”. In: *Journal of luminescence* 129.5 (2009), pp. 506–513.

[143] Maria J Jasmine and Edamana Prasad. “Fractal growth of PAMAM dendrimer aggregates and its impact on the intrinsic emission properties”. In: *The journal of physical chemistry B* 114.23 (2010), pp. 7735–7742.

[144] Amrendra Kumar et al. “Role of charged amino acids in sullying the fluorescence of tryptophan or conjugated Dansyl probe in monomeric proteins”. In: *Biophysical Journal* 121.3 (2022), 416a.

[145] Xiaoyu Wang et al. “The Intrinsic Fluorescence of Peptide Self-Assemblies Across pH Levels”. en. In: *Angewandte Chemie International Edition* 64.9 (2025), e202420567. DOI: 10.1002/anie.202420567.

[146] Wan In Lee, Yoonjung Bae, and Allen J Bard. “Strong blue photoluminescence and ECL from OH-terminated PAMAM dendrimers in the absence of gold nanoparticles”. In: *Journal of the American Chemical Society* 126.27 (2004), pp. 8358–8359.

[147] Dongjun Wang and Toyoko Imae. “Fluorescence emission from dendrimers and its pH dependence”. In: *Journal of the American Chemical Society* 126.41 (2004), pp. 13204–13205.

[148] Patrik K Johansson and Patrick Koelsch. “Label-free imaging of amyloids using their intrinsic linear and nonlinear optical properties”. In: *Biomedical Optics Express* 8.2 (2017), pp. 743–756.

[149] Chyi Wei Chung et al. “Label-free characterization of amyloids and alpha-synuclein polymorphs by exploiting their intrinsic fluorescence property”. In: *Analytical Chemistry* 94.13 (2022), pp. 5367–5374.

[150] Jonathan Pansieri et al. “Ultraviolet–visible–near-infrared optical properties of amyloid fibrils shed light on amyloidogenesis”. en. In: *Nature Photonics* 13.7 (2019), pp. 473–479. DOI: 10.1038/s41566-019-0422-6.

[151] Arpan Bhattacharya et al. “Direct evidence of intrinsic blue fluorescence from oligomeric interfaces of human serum albumin”. In: *Langmuir* 33.40 (2017), pp. 10606–10615.

[152] Chamani Niyangoda et al. “Carbonyl-based blue autofluorescence of proteins and amino acids”. In: *PLoS One* 12.5 (2017), e0176983.

[153] Mengmeng Li et al. "Impact of polymorphism on the optoelectronic properties of a low-bandgap semiconducting polymer". In: *Nature communications* 10.1 (2019), p. 2867.

[154] Guoqing Zhang et al. "Polymorphism and reversible mechanochromic luminescence for solid-state difluoroboron avobenzone". In: *Journal of the American Chemical Society* 132.7 (2010), pp. 2160–2162.

[155] Seong-Jun Yoon and SooYoung Park. "Polymorphic and mechanochromic luminescence modulation in the highly emissive dicyanodistyrylbenzene crystal: secondary bonding interaction in molecular stacking assembly". In: *Journal of Materials Chemistry* 21.23 (2011), pp. 8338–8346.

[156] Xinggui Gu et al. "Polymorphism-dependent emission for Di (p-methoxylphenyl) dibenzofulvene and analogues: optical waveguide/amplified spontaneous emission behaviors". In: *Advanced Functional Materials* 22.23 (2012), pp. 4862–4872.

[157] Chiara Botta et al. "Polymorphism-dependent aggregation induced emission of a push–pull dye and its multi-stimuli responsive behavior". In: *Journal of Materials Chemistry C* 4.14 (2016), pp. 2979–2989.

[158] Marcelo Echeverri et al. "Untangling the mechanochromic properties of benzothiadiazole-based luminescent polymorphs through supramolecular organic framework topology". In: *Journal of the American Chemical Society* 142.40 (2020), pp. 17147–17155.

[159] Ziqi Deng et al. "Revealing the excited-state mechanisms of the polymorphs of a hot exciton material". In: *Nature Communications* 16.1 (2025), p. 258.

[160] Qiang Zhang et al. "Synergistic Photoluminescence Enhancement in Nonaromatic Amino Acids and Sugars via Glycosylation". en. In: *Advanced Functional Materials* (2025), p. 2423603. DOI: 10.1002/adfm.202423603.

[161] Thomas E Creighton. "Disulphide bonds and protein stability". In: *BioEssays* 8.2-3 (1988), pp. 57–63.

[162] Ervin Welker et al. "Intramolecular VersusIntermolecular Disulfide Bonds in Prion Proteins". In: *Journal of Biological Chemistry* 277.36 (2002), pp. 33477–33481.

[163] Wen Chen et al. "Intramolecular disulfide bond between catalytic cysteines in an intein precursor". In: *Journal of the American Chemical Society* 134.5 (2012), pp. 2500–2503.

[164] M Puka-Sundvall et al. "Neurotoxicity of cysteine: interaction with glutamate". In: *Brain Research* 705.1-2 (1995), pp. 65–70.

[165] R Janaky et al. "Mechanisms of L-cysteine neurotoxicity". In: *Neurochemical Research* 25 (2000), pp. 1397–1405.

[166] Rodney P Guttmann and Tamara J Powell. "Redox regulation of cysteine-dependent enzymes in neurodegeneration". In: *International Journal of Cell Biology* 2012.1 (2012), p. 703164.

[167] Nathaniel Schafheimer et al. "Tyrosine/cysteine cluster sensitizing human γD-crystallin to ultraviolet radiation-induced photoaggregation in vitro". In: *Biochemistry* 53.6 (2014), pp. 979–990.

[168] Cristiana Valle and Maria Teresa Carrì. "Cysteine modifications in the pathogenesis of ALS". In: *Frontiers in Molecular Neuroscience* 10 (2017), p. 5.

[169] Robert F Steiner and Edward P Kirby. "Interaction of the ground and excited states of indole derivatives with electron scavengers". In: *The Journal of Physical Chemistry* 73.12 (1969), pp. 4130–4135.

[170] Danni L Harris and Bruce S Hudson. "Photophysics of tryptophan in bacteriophage T4 lysozymes". In: *Biochemistry* 29.22 (1990), pp. 5276–5285.

[171] Yu Chen and Mary D Barkley. "Toward understanding tryptophan fluorescence in proteins". In: *Biochemistry* 37.28 (1998), pp. 9976–9982.

[172] Weihong Qiu et al. "Ultrafast quenching of tryptophan fluorescence in proteins: Interresidue and intrahelical electron transfer". In: *Chemical Physics* 350.1-3 (2008), pp. 154–164.

[173] Giulia Giubertoni, Mischa Bonn, and Sander Woutersen. "D2O as an imperfect replacement for H2O: Problem or opportunity for protein research?" In: *The Journal of Physical Chemistry B* 127.38 (2023), pp. 8086–8094.

[174] Xiaohong Chen et al. "Prevalent intrinsic emission from nonaromatic amino acids and poly (amino acids)". In: *Science China Chemistry* 61 (2018), pp. 351–359.

[175] Junmei Wang et al. "Development and testing of a general amber force field". In: *Journal of Computational Chemistry* 25.9 (2004), pp. 1157–1174.

[176] David A Case et al. "The Amber biomolecular simulation programs". In: *Journal of Computational Chemistry* 26.16 (2005), pp. 1668–1688.

[177] David A. Case et al. "AmberTools". en. In: *Journal of Chemical Information and Modeling* 63.20 (2023), pp. 6183–6191. DOI: 10.1021/acs.jcim.3c01153.

[178] Frank Neese. "The ORCA program system". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 2.1 (2012), pp. 73–78.

[179] Frank Neese. "Software update: the ORCA program system, version 4.0". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 8.1 (2018), e1327.

[180] Frank Neese et al. “The ORCA quantum chemistry program package”. In: *The Journal of Chemical Physics* 152.22 (2020).

[181] Frank Neese. “Software update: The ORCA program system–Version 5.0”. In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 12.5 (2022), e1606.

[182] Frank Neese. “The SHARK integral generation and digestion system”. In: *Journal of Computational Chemistry* 44.3 (2023), pp. 381–396.

[183] Takeshi Yanai, David P Tew, and Nicholas C Handy. “A new hybrid exchange–correlation functional using the Coulomb-attenuating method (CAM-B3LYP)”. In: *Chemical Physics Letters* 393.1-3 (2004), pp. 51–57.

[184] Florian Weigend and Reinhart Ahlrichs. “Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy”. In: *Physical Chemistry Chemical Physics* 7.18 (2005), pp. 3297–3305.

[185] Arnim Hellweg and Dmitrij Rappoport. “Development of new auxiliary basis functions of the Karlsruhe segmented contracted basis sets including diffuse basis functions (def2-SVPD, def2-TZVPPD, and def2-QVPPD) for RI-MP2 and RI-CC calculations”. In: *Physical Chemistry Chemical Physics* 17.2 (2015), pp. 1010–1017.

[186] Frank Neese et al. “Efficient, approximate and parallel Hartree–Fock and hybrid DFT calculations. A ‘chain-of-spheres’ algorithm for the Hartree–Fock exchange”. In: *Chemical Physics* 356.1-3 (2009), pp. 98–109.

[187] William C Swope et al. “A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters”. In: *The Journal of Chemical Physics* 76.1 (1982), pp. 637–649.

[188] Mark E Casida. “Time-dependent density functional response theory for molecules”. In: *Recent Advances In Density Functional Methods: (Part I)*. World Scientific, 1995, pp. 155–192.

[189] So Hirata and Martin Head-Gordon. “Time-dependent density functional theory within the Tamm–Dancoff approximation”. In: *Chemical Physics Letters* 314.3-4 (1999), pp. 291–299.

[190] Michael JG Peach, Matthew J Williamson, and David J Tozer. “Influence of triplet instabilities in TDDFT”. In: *Journal of Chemical Theory and Computation* 7.11 (2011), pp. 3578–3585.

[191] Michael JG Peach and David J Tozer. "Overcoming low orbital overlap and triplet instability problems in TDDFT". In: *The Journal of Physical Chemistry A* 116.39 (2012), pp. 9783–9789.

[192] Enrico Tapavicza et al. "Mixed time-dependent density-functional theory/classical trajectory surface hopping study of oxirane photochemistry". In: *The Journal of Chemical Physics* 129.12 (2008).

[193] Yang Yang et al. "Conical intersections from particle–particle random phase and Tamm–Dancoff approximations". In: *The Journal of Physical Chemistry Letters* 7.13 (2016), pp. 2407–2411.

[194] Spiridoula Matsika. "Electronic structure methods for the description of nonadiabatic effects and conical intersections". In: *Chemical Reviews* 121.15 (2021), pp. 9407–9449.

[195] Sharon Hammes-Schiffer and John C Tully. "Proton transfer in solution: Molecular dynamics with quantum transitions". In: *The Journal of Chemical Physics* 101.6 (1994), pp. 4657–4667.

[196] Felix Plasser et al. "Surface hopping dynamics with correlated single-reference methods: 9H-adenine as a case study". In: *Journal of Chemical Theory and Computation* 10.4 (2014), pp. 1395–1405.

[197] Jin Wen, Sebastian Mai, and Leticia González. "Excited-state dynamics simulations of a light-driven molecular motor in solution". In: *The Journal of Physical Chemistry A* 127.45 (2023), pp. 9520–9529.

[198] Lea M Ibele et al. "Excimer intermediates en route to long-lived charge-transfer states in single-stranded adenine DNA as revealed by nonadiabatic dynamics". In: *The Journal of Physical Chemistry Letters* 11.18 (2020), pp. 7483–7488.

[199] Sebastian Mai et al. *SHARC3.0: Surface Hopping Including Arbitrary Couplings – Program Package for Non-Adiabatic Dynamics*. Zenodo. Version v3.0.0. 2023. DOI: 10.5281/zenodo.7828641.

[200] Joshua A. Rackers et al. "Tinker 8: Software Tools for Molecular Design". en. In: *Journal of Chemical Theory and Computation* 14.10 (2018), pp. 5273–5289. DOI: 10.1021/acs.jctc.8b00529.

[201] Om Shanker Tiwari et al. "Entropically-Driven Co-assembly of l-Histidine and l-Phenylalanine to Form Supramolecular Materials". In: *ACS Nano* 17.4 (2023), pp. 3506–3517.

[202] M.T. Harding and H. Long. "The crystal and molecular structure of L-cysteine". In: *Acta Crystallogr. B Struct. Chem.* 24.8 (1968), pp. 1096–1102.

[203] CH Görbitz and BJORN Dalhus. "l-Cysteine, monoclinic form, redetermination at 120K". In: *Acta Crystallographica Section C: Crystal Structure Communications* 52.7 (1996), pp. 1756–1759.

[204] KA Kerr and JP Ashmore. "Structure and conformation of orthorhombic L-cysteine". In: *Acta Crystallogr. B Struct. Chem.* 29.10 (1973), pp. 2124–2127.

[205] KA Kerr, J PATRICK Ashmore, and THOMAS F Koetzle. "A neutron diffraction study of L-cysteine". In: *Acta Crystallogr. B Struct. Chem.* 31.8 (1975), pp. 2022–2026.

[206] Stephen A Moggach, Stewart J Clark, and Simon Parsons. "l-Cysteine-I at 30 K". In: *Acta Crystallographica Section E: Structure Reports Online* 61.8 (2005), o2739–o2742.

[207] Stephen A Moggach et al. "High-pressure polymorphism in L-cysteine: the crystal structures of L-cysteine-III and L-cysteine-IV". In: *Acta Crystallographica Section B: Structural Science* 62.2 (2006), pp. 296–309.

[208] CAA Ghumman et al. "TOF-SIMS study of cystine and cholesterol stones". In: *Journal of Mass Spectrometry* 47.5 (2012), pp. 547–551. DOI: https://doi.org/10.1002/jms.2972.

[209] Niels H Damrauer et al. "Femtosecond dynamics of excited-state evolution in [Ru (bpy) 3] 2+". In: *Science* 275.5296 (1997), pp. 54–57.

[210] Sangwoon Yoon et al. "Direct observation of the ultrafast intersystem crossing in tris(2,2′-bipyridine) ruthenium(II) using femtosecond stimulated Raman spectroscopy". In: *Molecular Physics* 104.8 (2006), pp. 1275–1282.

[211] Jimena S Zugazagoitia, César Xavier Almora-Díaz, and Jorge Peon. "Ultrafast intersystem crossing in 1-nitronaphthalene. An experimental and computational study". In: *The Journal of Physical Chemistry A* 112.3 (2008), pp. 358–365.

[212] Andrew J Atkins and Leticia González. "Trajectory surface-hopping dynamics including intersystem crossing in [Ru (bpy) 3] 2+". In: *The Journal of Physical Chemistry Letters* 8.16 (2017), pp. 3840–3845.

[213] Kieran M Farrell et al. "Heavy-atom-substituted nucleobases in photodynamic applications: substitution of sulfur with selenium in 6-thioguanine induces a remarkable increase in the rate of triplet decay in 6-selenoguanine". In: *Journal of the American Chemical Society* 140.36 (2018), pp. 11214–11218.

[214] Danielle C Teles-Ferreira et al. "Ultrafast excited-state decay mechanisms of 6-thioguanine followed by sub-20 fs UV transient absorption spectroscopy". In: *Molecules* 27.4 (2022), p. 1200.

[215] Sebastian Mai et al. "The influence of the electronic structure method on intersystem crossing dynamics. The case of thioformaldehyde". In: *Journal of Chemical Theory and Computation* 15.6 (2019), pp. 3470–3480.

[216] Ryoga Hojo et al. "Imidazophenothiazine-Based Thermally Activated Delayed Fluorescence Materials with Ultra-Long-Lived Excited States for Energy Transfer Photocatalysis". In: *Journal of the American Chemical Society* 145.33 (2023), pp. 18366–18381.

[217] Benjamin G. Levine and Todd J. Martínez. "Isomerization Through Conical Intersections". en. In: *Annual Review of Physical Chemistry* 58.1 (2007), pp. 613–634. DOI: 10.1146/annurev.physchem.57.032905.104612.

[218] Mario Barbatti et al. "Newton-X Platform: New Software Developments for Surface Hopping and Nuclear Ensembles". en. In: *Journal of Chemical Theory and Computation* 18.11 (2022), pp. 6851–6865. DOI: 10.1021/acs.jctc.2c00804.

[219] Andreas Dreuw. "Why Computational Photochemistry Is Challenging and Will Probably Remain So: A Quantum Chemist's Perspective". en. In: *Advanced Science* (2026), e21012. DOI: 10.1002/advs.202521012.

[220] Gregory J. Atchity, Sotirios S. Xantheas, and Klaus Ruedenberg. "Potential energy surfaces near intersections". en. In: *The Journal of Chemical Physics* 95.3 (1991), pp. 1862–1876. DOI: 10.1063/1.461036.

[221] David R. Yarkony. "Conical Intersections: The New Conventional Wisdom". en. In: *The Journal of Physical Chemistry A* 105.26 (2001), pp. 6277–6293. DOI: 10.1021/jp003731u.

[222] Samer Gozem et al. "Shape of Multireference, Equation-of-Motion Coupled-Cluster, and Density Functional Theory Potential Energy Surfaces at a Conical Intersection". en. In: *Journal of Chemical Theory and Computation* 10.8 (2014), pp. 3074–3084. DOI: 10.1021/ct500154k.

[223] Xusong Li et al. "Analysis of the Geometrical Evolution in On-the-Fly Surface-Hopping Nonadiabatic Dynamics with Machine Learning Dimensionality Reduction Approaches: Classical Multidimensional Scaling and Isometric Feature Mapping". en. In: *Journal of Chemical Theory and Computation* 13.10 (2017), pp. 4611–4623. DOI: 10.1021/acs.jctc.7b00394.

[224] Xusong Li et al. "Analysis of trajectory similarity and configuration similarity in on-the-fly surface-hopping simulation on multi-channel nonadiabatic photoisomerization dynamics". en. In: *The Journal of Chemical Physics* 149.24 (2018), p. 244104. DOI: 10.1063/1.5048049.

[225] Yifei Zhu et al. "Unsupervised Machine Learning in the Analysis of Nonadiabatic Molecular Dynamics Simulation". en. In: *The Journal of Physical Chemistry Letters* 15.38 (2024), pp. 9601–9619. DOI: 10.1021/acs.jpclett.4c01751.

[226] Nicholas E. Jackson et al. "Introduction to Machine Learning for Molecular Simulation". In: *Journal of Chemical Theory and Computation* 19.14 (2023). PMID: 37489106, pp. 4335–4337. DOI: 10.1021/acs.jctc.3c00735.

[227] Julia Westermayr and Philipp Marquetand. "Machine Learning for Electronically Excited States of Molecules". en. In: *Chemical Reviews* 121.16 (2021), pp. 9873–9926. DOI: 10.1021/acs.chemrev.0c00749.

[228] Edoardo Cignoni et al. "Electronic Excited States from Physically Constrained Machine Learning". en. In: *ACS Central Science* 10.3 (2024), pp. 637–648. DOI: 10.1021/acscentsci.3c01480.

[229] Patrizia Mazzeo et al. "Electrostatic embedding machine learning for ground and excited state molecular dynamics of solvated molecules". In: *Digital Discovery* 3.12 (2024), pp. 2560–2571. DOI: 10.1039/D4DD00295D.

[230] Max Pinheiro et al. "ULaMDyn: enhancing excited-state dynamics analysis through streamlined unsupervised learning". en. In: *Digital Discovery* 4.3 (2025), pp. 666–682. DOI: 10.1039/D4DD00374H.

[231] Chris John et al. "Machine-Learning Framework for Excitation Energies of Chromophores in Polarizable Environments". en. In: *Journal of Chemical Information and Modeling* 66.4 (2026), pp. 2030–2040. DOI: 10.1021/acs.jcim.5c02424.

[232] Patrizia Mazzeo, Lorenzo Cupellini, and Benedetta Mennucci. "Multiscale Machine Learning Prediction of Infrared Spectra of Solvated Molecules". en. In: *Journal of Chemical Theory and Computation* 22.4 (2026), pp. 1883–1895. DOI: 10.1021/acs.jctc.5c01959.

[233] Aldo Glielmo et al. "Unsupervised Learning Methods for Molecular Simulation Data". In: *Chemical Reviews* 121.16 (2021). PMID: 33945269, pp. 9722–9758. DOI: 10.1021/acs.chemrev.0c01195.

[234] Julius P. P. Zauleck et al. "Two New Methods To Generate Internal Coordinates for Molecular Wave Packet Dynamics in Reduced Dimensions". en. In: *Journal of Chemical Theory and Computation* 12.12 (2016), pp. 5698–5708. DOI: 10.1021/acs.jctc.6b00800.

[235] Vincent Delmas et al. "Automated Selection of Nuclear Coordinates for Reduced Dimensionality Nonadiabatic Dynamics". en. In: *Journal of Chemical Theory and Computation* 21.13 (2025), pp. 6611–6621. DOI: 10.1021/acs.jctc.5c00110.

[236] Andrey K. Belyaev et al. "Nonadiabatic nuclear dynamics of the ammonia cation studied by surface hopping classical trajectory calculations". en. In: *The Journal of Chemical Physics* 142.10 (2015), p. 104307. DOI: 10.1063/1.4913962.

[237] Gareth W. Richings and Scott Habershon. "Analyzing Grid-Based Direct Quantum Molecular Dynamics Using Non-Linear Dimensionality Reduction". en. In: *Molecules* 26.24 (2021), p. 7418. DOI: 10.3390/molecules26247418.

[238] Pedram Tavadze et al. "A Machine-Driven Hunt for Global Reaction Coordinates of Azobenzene Photoisomerization". en. In: *Journal of the American Chemical Society* 140.1 (2018), pp. 285–290. DOI: 10.1021/jacs.7b10030.

[239] Wei Bin How et al. "Significance of the Chemical Environment of an Element in Nonadiabatic Molecular Dynamics: Feature Selection and Dimensionality Reduction with Machine Learning". en. In: *The Journal of Physical Chemistry Letters* 12.50 (2021), pp. 12026–12032. DOI: 10.1021/acs.jpclett.1c03469.

[240] Wei Bin How et al. "Dimensionality reduction in machine learning for nonadiabatic molecular dynamics: Effectiveness of elemental sublattices in lead halide perovskites". en. In: *The Journal of Chemical Physics* 156.5 (2022), p. 054110. DOI: 10.1063/5.0078473.

[241] Guoqing Zhou, Weibin Chu, and Oleg V. Prezhdo. "Structural Deformation Controls Charge Losses in MAPbI3: Unsupervised Machine Learning of Nonadiabatic Molecular Dynamics". en. In: *ACS Energy Letters* 5.6 (2020), pp. 1930–1938. DOI: 10.1021/acsenergylett.0c00899.

[242] Spencer M. Mangan et al. "Dependence between Structural and Electronic Properties of CsPbI3: Unsupervised Machine Learning of Nonadiabatic Molecular Dynamics". en. In: *The Journal of Physical Chemistry Letters* 12.35 (2021), pp. 8672–8678. DOI: 10.1021/acs.jpclett.1c02361.

[243] Takao Fuji et al. "Ultrafast photodynamics of furan". en. In: *The Journal of Chemical Physics* 133.23 (2010), p. 234303. DOI: 10.1063/1.3518441.

[244] Sven Oesterling et al. "Substituent effects on the relaxation dynamics of furan, furfural and β-furfural: a combined theoretical and experimental approach". en. In: *Physical Chemistry Chemical Physics* 19.3 (2017), pp. 2025–2035. DOI: 10.1039/C6CP06240G.

[245] Mario Barbatti et al. "The on-the-fly surface-hopping program system Newton-X: Application to ab initio simulation of the nonadiabatic photodynamics of benchmark systems". en. In: *Journal of Photochemistry and Photobiology A: Chemistry* 190.2–3 (2007), pp. 228–240. DOI: 10.1016/j.jphotochem.2006.12.008.

[246] Daniel Hollas et al. "Nonadiabatic Ab Initio Molecular Dynamics with the Floating Occupation Molecular Orbital-Complete Active Space Configuration

Interaction Method". en. In: *Journal of Chemical Theory and Computation* 14.1 (2018), pp. 339–350. DOI: 10.1021/acs.jctc.7b00958.

[247] Jiří Suchan, Jiří Janoš, and Petr Slavíček. "Pragmatic Approach to Photodynamics: Mixed Landau–Zener Surface Hopping with Intersystem Crossing". en. In: *Journal of Chemical Theory and Computation* 16.9 (2020), pp. 5809–5820. DOI: 10.1021/acs.jctc.0c00512.

[248] Jamie Conyard et al. "Ultrafast dynamics in the power stroke of a molecular rotary motor". en. In: *Nature Chemistry* 4.7 (2012), pp. 547–551. DOI: 10.1038/nchem.1343.

[249] Andranik Kazaryan et al. "Surface Hopping Excited-State Dynamics Study of the Photoisomerization of a Light-Driven Fluorene Molecular Rotary Motor". en. In: *Journal of Chemical Theory and Computation* 7.7 (2011), pp. 2189–2199. DOI: 10.1021/ct200199w.

[250] Víctor García-López, Dongdong Liu, and James M. Tour. "Light-Activated Organic Molecular Motors and Their Applications". en. In: *Chemical Reviews* 120.1 (2020), pp. 79–124. DOI: 10.1021/acs.chemrev.9b00221.

[251] Xiaojuan Pang et al. ""Watching" the Dark State in Ultrafast Nonadiabatic Photoisomerization Process of a Light-Driven Molecular Rotary Motor". en. In: *The Journal of Physical Chemistry A* 121.6 (2017), pp. 1240–1249. DOI: 10.1021/acs.jpca.6b12253.

[252] Jin Wen, Sebastian Mai, and Leticia González. "Excited-State Dynamics Simulations of a Light-Driven Molecular Motor in Solution". en. In: *The Journal of Physical Chemistry A* 127.45 (2023), pp. 9520–9529. DOI: 10.1021/acs.jpca.3c05841.

[253] Solomon Kullback and Richard A Leibler. "On information and sufficiency". In: *The annals of mathematical statistics* 22.1 (1951), pp. 79–86.

[254] Xuchun Yang et al. "Quantum–classical simulations of rhodopsin reveal excited-state population splitting and its effects on quantum efficiency". In: *Nature chemistry* 14.4 (2022), pp. 441–449.

[255] Richard A. van Delden et al. "Exploring the boundaries of a light-driven molecular motor design: new sterically overcrowded alkenes with preferred direction of rotation". en. In: *Organic & Biomolecular Chemistry* 2.10 (2004), pp. 1531–1541. DOI: 10.1039/B402222J.

[256] MJ ea Frisch et al. *Gaussian 16*. 2016.

[257] Michael Buchner, Branka M. Ladanyi, and Richard M. Stratt. "The short-time dynamics of molecular liquids. Instantaneous-normal-mode theory". en. In:

The Journal of Chemical Physics 97.11 (1992), pp. 8522–8535. DOI: 10.1063/1.463370.

[258] Nuray Sogunmez and Ebru Demet Akten. "Information Transfer in Active States of Human β_2-Adrenergic Receptor via Inter-Rotameric Motions of Loop Regions". English. In: *Applied Sciences* 12.17 (2022), p. 8530. DOI: 10.3390/app12178530.

[259] Daniel Barr et al. "Importance of Domain Closure for the Autoactivation of ERK2". In: *Biochemistry* 50.37 (2011), pp. 8038–8048. DOI: 10.1021/bi200503a.

[260] Jingxuan Zhu et al. "Neural relational inference to learn long-range allosteric interactions in proteins from molecular dynamics simulations". In: *Nature Communications* 13.1 (2022), pp. 1–16. DOI: 10.1038/s41467-022-29331-.

[261] Juan R. Perilla, Daniel J. Leahy, and Thomas B. Woolf. "Molecular dynamics simulations of transitions for ECD epidermal growth factor receptors show key differences between human and drosophila forms of the receptors". In: *Proteins: Structure, Function, and Bioinformatics* 81.7 (2013), pp. 1113–1126. DOI: https://doi.org/10.1002/prot.24257.

[262] Sunhwan Jo, Yifei Qi, and Wonpil Im. "Preferred conformations of N-glycan core pentasaccharide in solution and in glycoproteins". In: *Glycobiology* 26.1 (2015), pp. 19–29. DOI: 10.1093/glycob/cwv083.

[263] Ali Shojaie and Emily B. Fox. "Granger Causality: A Review and Recent Advances". In: *Annual Review of Statistics and Its Application* 9.Volume 9, 2022 (2022), pp. 289–319. DOI: https://doi.org/10.1146/annurev-statistics-040120-010930.

[264] Xing Wu et al. "Causal inference in the medical domain: a survey". In: *Applied Intelligence* 54 (2024), pp. 1–24. DOI: 10.1007/s10489-024-05338-9.

[265] Markus Gangl. "Causal Inference in Sociological Research". In: *Annual Review of Sociology* 36.Volume 36, 2010 (2010), pp. 21–47. DOI: https://doi.org/10.1146/annurev.soc.012809.102702.

[266] Kenneth J. Rothman and Sander Greenland. "Causation and Causal Inference in Epidemiology". In: *American Journal of Public Health* 95.S1 (2005), S144–S150. DOI: 10.2105/AJPH.2004.059204.

[267] Judea Pearl. *Causality: Models, Reasoning and Inference*. 2nd. USA: Cambridge University Press, 2009. DOI: 10.1017/CBO9780511803161.

[268] P. Spirtes and K. Zhang. "Causal discovery and inference: concepts and recent methodological advances". In: *Appl. Inform.* 3.3 (2016), pp. 1–28. DOI: 10.1186/s40535-016-0018-x.

[269] Jakob Runge et al. "Causal inference for time series". In: *Nature Reviews Earth & Environment* 4 (2023). DOI: 10.1038/s43017-023-00431-y.

[270] J. Runge. "Causal network reconstruction from time series: From theoretical assumptions to practical estimation". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 28.7 (2018), p. 075310.

[271] Peter Spirtes and Clark Glymour. "An Algorithm for Fast Recovery of Sparse Causal Graphs". In: *Social Science Computer Review* 9.1 (1991), pp. 62–72. DOI: 10.1177/089443939100900106.

[272] TS Verma and Judea Pearl. "Equivalence and Synthesis of Causal Models". In: *Probabilistic and Causal Inference: The Works of Judea Pearl*. 1st ed. New York, NY, USA: Association for Computing Machinery, 2022, pp. 221–236. ISBN: 9781450395861. https://doi.org/10.1145/3501714.3501732.

[273] Peter Spirtes et al. *Causation, Prediction, and Search*. Mit Press: Cambridge, 1993.

[274] Jakob Runge et al. "Detecting and quantifying causal associations in large nonlinear time series datasets". In: *Science Advances* 5.11 (2019), eaau4996. DOI: 10.1126/sciadv.aau4996.

[275] Pill-Soon Song and William E. Kurtin. "Photochemistry of the model phototropic system involving flavines and indoles. III. A spectroscopic study of the polarized luminescence of indoles". English. In: *Journal of the American Chemical Society* 91.17 (1969), pp. 4892–4906. DOI: 10.1021/ja01045a052.

[276] Bo Albinsson et al. "Near-ultraviolet electronic transitions of the tryptophan chromophore: linear dichroism, fluorescence anisotropy, and magnetic circular dichroism spectra of some indole derivatives". English. In: *The Journal of Physical Chemistry* 93.18 (1989), pp. 6646–6654. DOI: 10.1021/j100355a016.

[277] Luis Serrano-Andrés and Björn O. Roos. "Theoretical Study of the Absorption and Emission Spectra of Indole in the Gas Phase and in a Solvent". English. In: *Journal of the American Chemical Society* 118.1 (1996), pp. 185–195. DOI: 10.1021/ja952035i.

[278] Andrzej L. Sobolewski and Wolfgang Domcke. "Ab initio investigations on the photophysics of indole". English. In: *Chemical Physics Letters* 315.3–4 (1999), pp. 293–298. DOI: 10.1016/S0009-2614(99)01249-X.

[279] C. Dedonder-Lardeux et al. "External electric field effect on the lowest excited states of indole: ab initio and molecular dynamics study". English. In: *Physical Chemistry Chemical Physics* 5.22 (2003), p. 5118. DOI: 10.1039/b308866a.

[280] Ali A. Hassanali et al. "A Molecular Dynamics Study of Lys-Trp-Lys: Structure and Dynamics in Solution Following Photoexcitation". English. In: *The Journal of Physical Chemistry B* 110.21 (2006), pp. 10497–10508. DOI: 10.1021/jp0601926.

[281] Khatereh Azizi et al. "Examining the origins of observed terahertz modes from an optically pumped atomistic model protein in aqueous solution". English. In: *PNAS Nexus* 2.8 (2023). Ed. by P Fratzl, pgad257. DOI: 10.1093/pnasnexus/pgad257.

[282] Chuan Tian et al. "ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution". In: *Journal of Chemical Theory and Computation* 16.1 (2020), pp. 528–552. DOI: 10.1021/acs.jctc.9b00591.

[283] Saeed Izadi, Ramu Anandakrishnan, and Alexey V. Onufriev. "Building Water Models: A Different Approach". English. In: *The Journal of Physical Chemistry Letters* 5.21 (2014), pp. 3863–3871. DOI: 10.1021/jz501780a.

[284] Saeed Izadi and Alexey V. Onufriev. "Accuracy limit of rigid 3-point water models". English. In: *The Journal of Chemical Physics* 145.7 (2016), p. 074501. DOI: 10.1063/1.4960175.

[285] Yeyue Xiong, Parviz Seifpanahi Shabane, and Alexey V. Onufriev. "Melting Points of OPC and OPC3 Water Models". English. In: *ACS Omega* 5.39 (2020), pp. 25087–25094. DOI: 10.1021/acsomega.0c02638.

[286] Carmelo Tempra, O. H. Samuli Ollila, and Matti Javanainen. "Accurate Simulations of Lipid Monolayers Require a Water Model with Correct Surface Tension". English. In: *Journal of Chemical Theory and Computation* 18.3 (2022), pp. 1862–1869. DOI: 10.1021/acs.jctc.1c00951.

[287] Tadashi Ando. "Shear viscosity of OPC and OPC3 water models". English. In: *The Journal of Chemical Physics* 159.10 (2023), p. 101102. DOI: 10.1063/5.0161476.

[288] H.J.C. Berendsen, D. Van Der Spoel, and R. Van Drunen. "GROMACS: A message-passing parallel molecular dynamics implementation". English. In: *Computer Physics Communications* 91.1–3 (1995), pp. 43–56. DOI: 10.1016/0010-4655(95)00042-E.

[289] Erik Lindahl, Berk Hess, and David Van Der Spoel. "GROMACS 3.0: a package for molecular simulation and trajectory analysis". English. In: *Journal of Molecular Modeling* 7.8 (2001), pp. 306–317. DOI: 10.1007/s008940100045.

[290] David Van Der Spoel et al. "GROMACS: Fast, flexible, and free". English. In: *Journal of Computational Chemistry* 26.16 (2005), pp. 1701–1718. DOI: 10.1002/jcc.20291.

[291] Berk Hess et al. "GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation". English. In: *Journal of Chemical Theory and Computation* 4.3 (2008), pp. 435–447. DOI: 10.1021/ct700301q.

[292] Sander Pronk et al. "GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit". English. In: *Bioinformatics* 29.7 (2013), pp. 845–854. DOI: 10.1093/bioinformatics/btt055.

[293] Mark James Abraham et al. "GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers". English. In: *SoftwareX* 1–2 (2015), pp. 19–25. DOI: 10.1016/j.softx.2015.06.001.

[294] Szilárd Páll et al. "Tackling Exascale Software Challenges in Molecular Dynamics Simulations with GROMACS". English. In: *Solving Software Challenges for Exascale*. Ed. by Stefano Markidis and Erwin Laure. Vol. 8759. Cham: Springer International Publishing, 2015, pp. 3–27. DOI: 10.1007/978-3-319-15976-8_1.

[295] Mark Abraham et al. *GROMACS 2023.2 Source code*. 2023. DOI: 10.5281/ZENODO.8134397.

[296] Giovanni Bussi, Davide Donadio, and Michele Parrinello. "Canonical sampling through velocity rescaling". English. In: *The Journal of Chemical Physics* 126.1 (2007), p. 014101. DOI: 10.1063/1.2408420.

[297] Berk Hess et al. "LINCS: A linear constraint solver for molecular simulations". English. In: *Journal of Computational Chemistry* 18.12 (1997), pp. 1463–1472. DOI: 10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H.

[298] Lennart Nilsson and Bertil Halle. "Molecular origin of time-dependent fluorescence shifts in proteins". English. In: *Proceedings of the National Academy of Sciences* 102.39 (2005), pp. 13867–13872. DOI: 10.1073/pnas.0504181102.

[299] Bertil Halle and Lennart Nilsson. "Does the Dynamic Stokes Shift Report on Slow Protein Hydration Dynamics?" English. In: *The Journal of Physical Chemistry B* 113.24 (2009), pp. 8210–8213. DOI: 10.1021/jp9027589.

[300] P. W. Fenimore et al. "Bulk-solvent and hydration-shell fluctuations, similar to α- and β-fluctuations in glasses, control protein motions and functions". English. In: *Proceedings of the National Academy of Sciences* 101.40 (2004), pp. 14408–14413. DOI: 10.1073/pnas.0405573101.

[301] H. Frauenfelder et al. "Protein folding is slaved to solvent motions". English. In: *Proceedings of the National Academy of Sciences* 103.42 (2006), pp. 15469–15472. DOI: 10.1073/pnas.0607168103.

[302] Luyuan Zhang et al. "Mapping hydration dynamics around a protein surface". English. In: *Proceedings of the National Academy of Sciences* 104.47 (2007), pp. 18461–18466. DOI: 10.1073/pnas.0707647104.

[303] Matteo Allione, Vittorio Del Tatto, and Alessandro Laio. *Linear scaling causal discovery from high-dimensional time series by dynamical community detection*. 2025. https://arxiv.org/abs/2501.10886.

[304] Martin Karplus and J Andrew McCammon. "Molecular dynamics simulations of biomolecules". In: *Nature structural biology* 9.9 (2002), pp. 646–652.

[305] Katherine Henzler-Wildman and Dorothee Kern. "Dynamic personalities of proteins". In: *Nature* 450.7172 (2007), pp. 964–972.

[306] David D Boehr, Ruth Nussinov, and Peter E Wright. "The role of dynamic conformational ensembles in biomolecular recognition". In: *Nature chemical biology* 5.11 (2009), pp. 789–796.

[307] Jingjing Guo and Huan-Xiang Zhou. "Protein allostery and conformational dynamics". In: *Chemical reviews* 116.11 (2016), pp. 6503–6515.

[308] Kwangho Nam and Magnus Wolf-Watz. "Protein dynamics: The future is bright and complicated!" In: *Structural Dynamics* 10.1 (2023).

[309] Rita Grandori. "Protein structure and dynamics in the era of integrative structural biology". In: *Frontiers in Biophysics* 1 (2023), p. 1219843.

[310] Junichi Ono et al. "Conformational dynamics in proteins: entangled slow fluctuations and nonequilibrium reaction events". In: *The Journal of Physical Chemistry B* 128.1 (2023), pp. 20–32.

[311] Gilad Haran. "How fast dynamics affect slow function in protein machines: A single-molecule perspective". In: *Biophysical Journal* 123.3 (2024), 321a.

[312] CE Hatton and P Mehrabi. "Exploring the dynamics of allostery through multidimensional crystallography". In: *Biophysical Reviews* 16.5 (2024), pp. 563–570.

[313] David E Shaw et al. "Millisecond-scale molecular dynamics simulations on Anton". In: *Proceedings of the conference on high performance computing networking, storage and analysis*. 2009, pp. 1–11.

[314] John L Klepeis et al. "Long-timescale molecular dynamics simulations of protein structure and function". In: *Current opinion in structural biology* 19.2 (2009), pp. 120–127.

[315] Kresten Lindorff-Larsen et al. "How Fast-Folding Proteins Fold". en. In: *Science* 334.6055 (2011), pp. 517–520. DOI: 10.1126/science.1208351.

[316] Ron O Dror et al. "Biomolecular simulation: a computational microscope for molecular biology". In: *Annual review of biophysics* 41 (2012), pp. 429–452.

[317] Thomas J Lane et al. "To milliseconds and beyond: challenges in the simulation of protein folding". In: *Current opinion in structural biology* 23.1 (2013), pp. 58–65.

[318] Scott A Hollingsworth and Ron O Dror. "Molecular dynamics simulation for all". In: *Neuron* 99.6 (2018), pp. 1129–1143.

[319] David J Huggins et al. "Biomolecular simulations: From dynamics and mechanisms to computational assays of biological activity". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 9.3 (2019), e1393.

[320] David E Shaw et al. "Anton 3: twenty microseconds of molecular dynamics simulation before lunch". In: *Proceedings of the international conference for high performance computing, networking, storage and analysis*. 2021, pp. 1–11.

[321] Agastya P Bhati et al. "Long time scale ensemble methods in molecular dynamics: Ligand–protein interactions and allostery in SARS-CoV-2 targets". In: *Journal of chemical theory and computation* 19.11 (2023), pp. 3359–3378.

[322] David E Shaw et al. "Atomic-level characterization of the structural dynamics of proteins". In: *Science* 330.6002 (2010), pp. 341–346.

[323] Kresten Lindorff-Larsen et al. "Picosecond to Millisecond Structural Dynamics in Human Ubiquitin". en. In: *The Journal of Physical Chemistry B* 120.33 (2016), pp. 8313–8320. DOI: 10.1021/acs.jpcb.6b02024.

[324] Akio Kitao, Fumio Hirata, and Nobuhiro Gō. "The effects of solvent on the conformation and the collective motions of protein: normal mode analysis and molecular dynamics simulations of melittin in water and in vacuum". In: *Chemical physics* 158.2-3 (1991), pp. 447–472.

[325] Toshiko Ichiye and Martin Karplus. "Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations". In: *Proteins: Structure, Function, and Bioinformatics* 11.3 (1991), pp. 205–217.

[326] Angel E García. "Large-amplitude nonlinear motions in proteins". In: *Physical review letters* 68.17 (1992), p. 2696.

[327] Andrea Amadei, Antonius BM Linssen, and Herman JC Berendsen. "Essential dynamics of proteins". In: *Proteins: Structure, Function, and Bioinformatics* 17.4 (1993), pp. 412–425.

[328] Steven Hayward, Akio Kitao, and Nobuhiro Gō. "Harmonicity and anharmonicity in protein dynamics: a normal mode analysis and principal component analysis". In: *Proteins: Structure, Function, and Bioinformatics* 23.2 (1995), pp. 177–186.

[329] Steven Hayward and Nobuhiro Go. "Collective variable description of native protein dynamics". In: *Annual review of physical chemistry* 46.1 (1995), pp. 223–250.

[330] Manel A Balsera et al. "Principal component analysis and long time protein dynamics". In: *The Journal of Physical Chemistry* 100.7 (1996), pp. 2567–2572.

[331] Akio Kitao and Nobuhiro Go. "Investigating protein dynamics in collective coordinate space". In: *Current opinion in structural biology* 9.2 (1999), pp. 164–169.

[332] Gia G Maisuradze, Adam Liwo, and Harold A Scheraga. "Principal component analysis for protein folding dynamics". In: *Journal of molecular biology* 385.1 (2009), pp. 312–329.

[333] Rodrigo Cossio-Pérez, Juliana Palma, and Gustavo Pierdominici-Sottile. "Consistent principal component modes from molecular dynamics simulations of proteins". In: *Journal of Chemical Information and Modeling* 57.4 (2017), pp. 826–834.

[334] Michael Buchner, Branka M Ladanyi, and Richard M Stratt. "The short-time dynamics of molecular liquids. Instantaneous-normal-mode theory". In: *The Journal of chemical physics* 97.11 (1992), pp. 8522–8535.

[335] Lennard M Horstink et al. "Functionally important correlated motions in the single-stranded DNA-binding protein encoded by filamentous phage Pf3". In: *Journal of molecular biology* 287.3 (1999), pp. 569–577.

[336] Herman JC Berendsen and Steven Hayward. "Collective protein dynamics in relation to function". In: *Current opinion in structural biology* 10.2 (2000), pp. 165–169.

[337] P-L Chau et al. "Functional concerted motions in the bovine serum retinol-binding protein". In: *Journal of computer-aided molecular design* 13.1 (1999), pp. 11–20.

[338] Yuguang Mu, Phuong H Nguyen, and Gerhard Stock. "Energy landscape of a small peptide revealed by dihedral angle principal component analysis". In: *Proteins: Structure, Function, and Bioinformatics* 58.1 (2005), pp. 45–52.

[339] Francesco Capozzi et al. "Essential dynamics of helices provide a functional classification of EF-hand proteins". In: *Journal of proteome research* 6.11 (2007), pp. 4245–4255.

[340] Alexandros Altis et al. "Dihedral angle principal component analysis of molecular dynamics simulations". In: *The Journal of chemical physics* 126.24 (2007).

[341] Francesco Pontiggia, Andrea Zen, and Cristian Micheletti. "Small-and large-scale conformational changes of adenylate kinase: a molecular dynamics study of the subdomain motion and mechanics". In: *Biophysical journal* 95.12 (2008), pp. 5901–5912.

[342] Alexandros Altis et al. "Construction of the free energy landscape of biomolecules via dihedral angle principal component analysis". In: *The Journal of chemical physics* 128.24 (2008).

[343] Giulia Morra et al. "Corresponding functional dynamics across the Hsp90 Chaperone family: insights from a multiscale analysis of MD simulations". In: *PLoS computational biology* 8.3 (2012), e1002433.

[344] Gia G Maisuradze et al. "Local vs global motions in protein folding". In: *Journal of chemical theory and computation* 9.7 (2013), pp. 2907–2921.

[345] Florian Sittel, Abhinav Jain, and Gerhard Stock. "Principal component analysis of molecular dynamics: On the use of Cartesian vs. internal coordinates". In: *The Journal of Chemical Physics* 141.1 (2014).

[346] Matthias Ernst, Florian Sittel, and Gerhard Stock. "Contact-and distance-based principal component analysis of protein dynamics". In: *The Journal of chemical physics* 143.24 (2015).

[347] Matthias Post, Steffen Wolf, and Gerhard Stock. "Principal component analysis of nonequilibrium molecular dynamics simulations". In: *The Journal of chemical physics* 150.20 (2019).

[348] Lutz Molgedey and Heinz Georg Schuster. "Separation of a mixture of independent signals using time delayed correlations". In: *Physical review letters* 72.23 (1994), p. 3634.

[349] Yusuke Naritomi and Sotaro Fuchigami. "Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: the case of domain motions". In: *The Journal of chemical physics* 134.6 (2011).

[350] Guillermo Pérez-Hernández et al. "Identification of slow molecular order parameters for Markov model construction". In: *The Journal of chemical physics* 139.1 (2013).

[351] Christian R Schwantes and Vijay S Pande. "Improvements in Markov state model construction reveal many non-native interactions in the folding of NTL9". In: *Journal of chemical theory and computation* 9.4 (2013), pp. 2000–2009.

[352] Frank Noé and Cecilia Clementi. "Kinetic distance and kinetic maps from molecular dynamics simulation". In: *Journal of chemical theory and computation* 11.10 (2015), pp. 5002–5011.

[353] Guillermo Pérez-Hernández and Frank Noé. "Hierarchical time-lagged independent component analysis: computing slow modes and reaction coordinates for large molecular systems". In: *Journal of chemical theory and computation* 12.12 (2016), pp. 6118–6129.

[354] Feliks Nuske et al. "Variational approach to molecular kinetics". In: *Journal of chemical theory and computation* 10.4 (2014), pp. 1739–1752.

[355] Mohammad M. Sultan and Vijay S Pande. "tICA-metadynamics: accelerating metadynamics by using kinetically selected collective variables". In: *Journal of chemical theory and computation* 13.6 (2017), pp. 2440–2447.

[356] Steffen Schultze and Helmut Grubmüller. "Time-lagged independent component analysis of random walks and protein dynamics". In: *Journal of Chemical Theory and Computation* 17.9 (2021), pp. 5766–5776.

[357] Luigi Bonati, GiovanniMaria Piccini, and Michele Parrinello. "Deep learning the slow modes for rare events sampling". In: *Proceedings of the National Academy of Sciences* 118.44 (2021), e2113533118.

[358] Peter Spirtes. *Causation, Prediction and Search*. MIT press, 2000.

[359] Rafael Brüschweiler. "Collective protein dynamics and nuclear spin relaxation". en. In: *The Journal of Chemical Physics* 102.8 (1995), pp. 3396–3403. DOI: 10.1063/1.469213.

[360] JJ Prompers, SF Lienin, and R Brüschweiler. "Collective reorientational motion and nuclear spin relaxation in proteins". In: *Biocomputing 2001*. World Scientific, 2000, pp. 79–88.

[361] Ivan Rivalta et al. "Allosteric pathways in imidazole glycerol phosphate synthase". In: *Proceedings of the National Academy of Sciences* 109.22 (2012), E1428–E1436.

[362] Austin D Vogt and Enrico Di Cera. "Conformational selection or induced fit? A critical appraisal of the kinetic mechanism". In: *Biochemistry* 51.30 (2012), pp. 5894–5902.

[363] Adam T VanWart et al. "Exploring residue component contributions to dynamical network models of allostery". In: *Journal of chemical theory and computation* 8.8 (2012), pp. 2949–2961.

[364] Samuel Hertig, Naomi R Latorraca, and Ron O Dror. "Revealing atomic-level mechanisms of protein allostery with molecular dynamics simulations". In: *PLoS computational biology* 12.6 (2016), e1004746.

[365] S Bowerman and J Wereszczynski. "Detecting allosteric networks using molecular dynamics simulation". In: *Methods in enzymology* 578 (2016), pp. 429–447.

[366] Ivan Rivalta et al. "Allosteric communication disrupted by a small molecule binding to the imidazole glycerol phosphate synthase protein–protein interface". In: *Biochemistry* 55.47 (2016), pp. 6484–6494.

[367] Jian Wang et al. "Mapping allosteric communications within individual proteins". In: *Nature communications* 11.1 (2020), p. 3862.

[368] Shoshana J Wodak et al. "Allostery in its many disguises: from theory to applications". In: *Structure* 27.4 (2019), pp. 566–578.

[369] Adele Hardie et al. "Deconstructing allostery by computational assessment of the binding determinants of allosteric PTP1B modulators". In: *Communications Chemistry* 6.1 (2023), p. 125.

[370] Mattia Bernetti et al. "Probing allosteric communication with combined molecular dynamics simulations and network analysis". In: *Current Opinion in Structural Biology* 86 (2024), p. 102820.

[371] John Jumper et al. "Highly accurate protein structure prediction with AlphaFold". In: *nature* 596.7873 (2021), pp. 583–589.

[372] Justas Dauparas et al. "Robust deep learning–based protein sequence design using ProteinMPNN". In: *Science* 378.6615 (2022), pp. 49–56.

[373] Zeming Lin et al. "Evolutionary-scale prediction of atomic-level protein structure with a language model". In: *Science* 379.6637 (2023), pp. 1123–1130.

[374] Joseph L Watson et al. "De novo design of protein structure and function with RFdiffusion". In: *Nature* 620.7976 (2023), pp. 1089–1100.

[375] Thomas Hayes et al. "Simulating 500 million years of evolution with a language model". In: *BioRxiv* (2024), pp. 2024–07.

[376] Josh Abramson et al. "Accurate structure prediction of biomolecular interactions with AlphaFold 3". In: *Nature* 630.8016 (2024), pp. 493–500.

[377] Kiera H Sumida et al. "Improving protein expression, stability, and function with ProteinMPNN". In: *Journal of the American Chemical Society* 146.3 (2024), pp. 2054–2061.

[378] Woody Ahern et al. "Atom-level enzyme active site scaffolding using RFdiffusion2". In: *Nature Methods* 23.1 (2026), pp. 96–105.

[379] Yu Kay Law and Ali A Hassanali. "Role of quantum vibrations on the structural, electronic, and optical properties of 9-methylguanine". In: *The Journal of Physical Chemistry A* 119.44 (2015), pp. 10816–10827.

[380] Subrahmanyam Sappati et al. "Nuclear quantum effects in a HIV/cancer inhibitor: The case of ellipticine". In: *The Journal of Chemical Physics* 145.20 (2016).

[381] Francesco Ambrosio, Giacomo Miceli, and Alfredo Pasquarello. "Structural, dynamical, and electronic properties of liquid water: A hybrid functional study". In: *The Journal of Physical Chemistry B* 120.30 (2016), pp. 7456–7470.

[382] YK Law and AA Hassanali. "The importance of nuclear quantum effects in spectral line broadening of optical spectra and electrostatic properties in aromatic chromophores". In: *The Journal of Chemical Physics* 148.10 (2018).

[383] Unmesh Mondal et al. "Effect of Quantum Delocalization on Temperature Dependent Double Proton Transfer in Molecular Crystals of Terephthalic Acid". In: *The Journal of Physical Chemistry B* 127.23 (2023), pp. 5263–5272.

[384] Margaret L Berrens et al. "Nuclear quantum effects on the electronic structure of water and ice". In: *The Journal of Physical Chemistry Letters* 15.26 (2024).

[385] Germaine Neza Hozana, Gonzalo Díaz Mirón, and Ali Hassanali. "Data-Driven Discovery of the Origins of UV Absorption in the Alpha-3C Protein". In: *The Journal of Physical Chemistry B* 129.19 (2025), pp. 4728–4737.

[386] Jörg Behler and Michele Parrinello. "Generalized neural-network representation of high-dimensional potential-energy surfaces". In: *Physical review letters* 98.14 (2007), p. 146401.

[387] Kristof Schütt et al. "Schnet: A continuous-filter convolutional neural network for modeling quantum interactions". In: *Advances in neural information processing systems* 30 (2017).

[388] Kristof T Schütt et al. "Schnet–a deep learning architecture for molecules and materials". In: *The Journal of chemical physics* 148.24 (2018).

[389] Linfeng Zhang et al. "Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics". In: *Physical review letters* 120.14 (2018).

[390] R Zubatyuk et al. "Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network." In: *Science Advances* 5.8 (2019), eaav6490–eaav6490.

[391] Kristof Schütt, Oliver Unke, and Michael Gastegger. "Equivariant message passing for the prediction of tensorial properties and molecular spectra". In: *International conference on machine learning*. PMLR. 2021, pp. 9377–9388.

[392] Ilyes Batatia et al. "MACE: Higher order equivariant message passing neural networks for fast and accurate force fields". In: *Advances in neural information processing systems* 35 (2022), pp. 11423–11436.

[393] Thorben Frank, Oliver Thorsten Unke, and Klaus Robert Muller. "So3krates: Equivariant attention for interactions on arbitrary length-scales in molecular systems". In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. 2022. https://openreview.net/forum?id=tlUnxtAmcJq.

[394] Simon Batzner et al. "E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials". en. In: *Nature Communications* 13.1 (2022), p. 2453. DOI: 10.1038/s41467-022-29939-5.

[395] Arslan Mazitov et al. “PET-MAD as a lightweight universal interatomic potential for advanced materials modeling”. In: *Nature Communications* 16.1 (2025), p. 10653.

[396] Dylan M Anstine, Roman Zubatyuk, and Olexandr Isayev. “AIMNet2: a neural network potential to meet your neutral, charged, organic, and elemental-organic needs”. In: *Chemical Science* 16.23 (2025), pp. 10228–10244.

[397] Michele Ceriotti et al. “Nuclear quantum effects in water and aqueous systems: Experiment, theory, and current challenges”. In: *Chemical Reviews* 116.13 (2016), pp. 7529–7550.

[398] Thomas E Markland and Michele Ceriotti. “Nuclear quantum effects enter the mainstream”. In: *Nature Reviews Chemistry* 2.3 (2018), p. 0109.

[399] Venkat Kapil et al. “First-principles spectroscopy of aqueous interfaces using machine-learned electronic and quantum nuclear effects”. In: *Faraday Discussions* 249 (2024), pp. 50–68.

[400] Iryna Zaporozhets et al. “Accurate nuclear quantum statistics on machine-learned classical effective potentials”. In: *The Journal of Chemical Physics* 161.13 (2024).

[401] Julia Westermayr, Michael Gastegger, and Philipp Marquetand. “Combining SchNet and SHARC: The SchNarc machine learning approach for excited-state dynamics”. In: *The journal of physical chemistry letters* 11.10 (2020), pp. 3828–3834.

[402] Julia Westermayr and Philipp Marquetand. “Deep learning for UV absorption spectra with SchNarc: First steps toward transferability in chemical compound space”. In: *The Journal of Chemical Physics* 153.15 (2020).

[403] Sascha Mausenberger et al. “Spainn: equivariant message passing for excited-state nonadiabatic molecular dynamics”. In: *Chemical Science* 15.38 (2024), pp. 15880–15890.

[404] Mikołaj Martyka et al. “Charting electronic-state manifolds across molecules with multi-state learning and gap-driven dynamics via efficient and robust active learning”. In: *npj computational materials* 11.1 (2025), p. 132.

[405] Rhyan Barrett, Christoph Ortner, and Julia Westermayr. “Transferable Machine Learning Potential X-MACE for Excited States using Integrated DeepSets”. In: *arXiv preprint arXiv:2502.12870* (2025).

[406] Mikołaj Martyka et al. “OMNI-P2x universal neural network potential for excited-state simulations”. en. In: *Nature Communications* 17.1 (2026), p. 4949. DOI: 10.1038/s41467-026-71380-5.

[407] Cesare Malosso et al. “Transferable machine learning of excited-state dynamics with extremal pooling”. In: *arXiv preprint arXiv:2606.16859* (2026).

[408] Frederick Eberhardt and Richard Scheines. “Interventions and causal inference”. In: *Philosophy of science* 74.5 (2007), pp. 981–995.

[409] Herbert B. Callen and Theodore A. Welton. “Irreversibility and Generalized Noise”. English. In: *Physical Review* 83.1 (1951), pp. 34–40. DOI: 10.1103/PhysRev.83.34.

[410] R Kubo. “The fluctuation-dissipation theorem”. In: *Reports on Progress in Physics* 29.1 (1966), pp. 255–284. DOI: 10.1088/0034-4885/29/1/306.

[411] George M Sheldrick. “SHELXT–Integrated space-group and crystal-structure determination”. In: *Acta Crystallographica Section A: Foundations and Advances* 71.1 (2015), pp. 3–8.

[412] George M Sheldrick. “Crystal structure refinement with SHELXL”. In: *Acta Crystallographica Section C: Structural Chemistry* 71.1 (2015), pp. 3–8.

[413] Oleg V Dolomanov et al. “OLEX2: a complete structure solution, refinement and analysis program”. In: *Journal of Applied Crystallography* 42.2 (2009), pp. 339–341.

[414] Adele D Laurent and Denis Jacquemin. “TD-DFT benchmarks: a review”. In: *International Journal of Quantum Chemistry* 113.17 (2013), pp. 2019–2039.

[415] Jochen Schirmer. “Beyond the random-phase approximation: A new approximation scheme for the polarization propagator”. In: *Physical Review A* 26.5 (1982), p. 2395.

[416] Alexander B Trofimov and Jochen Schirmer. “An efficient polarization propagator approach to valence electron excitation spectra”. In: *Journal of Physics B: Atomic, Molecular and Optical Physics* 28.12 (1995), p. 2299.

[417] Thomas V Papineau, Denis Jacquemin, and Morgane Vacher. “Which electronic structure method to choose in trajectory surface hopping dynamics simulations? Azomethane as a case study”. In: *The Journal of Physical Chemistry Letters* 15.2 (2024), pp. 636–643.

[418] Milan Palu š and Martin Vejmelka. “Directionality of coupling from bivariate time series: How to avoid false causalities and missed connections”. In: *Phys. Rev. E* 75 (5 2007), p. 056211. DOI: 10.1103/PhysRevE.75.056211.

[419] Young-Il Moon, Balaji Rajagopalan, and Upmanu Lall. “Estimation of mutual information using kernel density estimators”. In: *Phys. Rev. E* 52 (3 1995), pp. 2318–2321. DOI: 10.1103/PhysRevE.52.2318.

[420] R. Steuer et al. "The mutual information: Detecting and evaluating dependencies between variables". In: *Bioinformatics* 18.suppl_2 (2002), S231–S240. DOI: 10.1093/bioinformatics/18.suppl_2.S231.

[421] Stefan Frenzel and Bernd Pompe. "Partial Mutual Information for Coupling Analysis of Multivariate Time Series". In: *Phys. Rev. Lett.* 99 (20 2007), p. 204101. DOI: 10.1103/PhysRevLett.99.204101.

[422] N. N. Leonenko L. F. Kozachenko. "Sample Estimate of the Entropy of a Random Vector". In: *Probl. Peredachi Inf.* 23 (2 1987), pp. 9–16. https://zbmath.org/?q=an:0633.62005.

[423] Daniel Chicharro and Stefano Panzeri. "Algorithms of causal inference for the analysis of effective connectivity among brain regions". In: *Frontiers in Neuroinformatics* 8 (2014). DOI: 10.3389/fninf.2014.00064.

This page intentionally left blank.

Appendices

This page intentionally left blank.

A

Supplementary Information: Crystallization of L-Cysteine in Heavy Water Induces Intrinsic Fluorescence

A.1 Experimental Methods

A.1.1 L-Cysteine crystallization

A solution of L-Cysteine (200 mg) in $1\,\mathrm{mL}$ of H_2O and D_2O was prepared. To obtain a clear, transparent solution, the samples were heated to $90\,^{\circ}\mathrm{C}$. After heating, the solutions were gradually cooled to room temperature to facilitate self-assembly and crystallization.

A.1.2 Confocal Microscopy

A Leica SP8 Lightning confocal microscope with Leica Application Suite X (LAS X) software was used to capture confocal images of crystals. A $405\,\mathrm{nm}$ laser was used to excite the samples. A wavelength range of $415\,\mathrm{nm}$ to $600\,\mathrm{nm}$ was set for emission. Image analysis was performed using a Leica SP8 confocal microscope to extract the emission graph.

A.1.3 Quantum Yield

Fluorescence measurements were carried out with a Horiba Scientific Fluoromax-4 spectrofluorometer. Absolute fluorescence quantum yield measurements were performed using a Quanta-Phi integrating sphere connected to the Fluoromax-4.

A.1.4 Fluorescence lifetime microscopy (FLIM)

All fluorescence lifetime (FLT) measurements were performed using a two-channel laser scanning confocal microscope (DCS 120, Becker & Hickl GmbH, Berlin, Germany). The full width at half maximum (FWHM) of the excitation pulse for this system is of the order of 10–100 ps. For this study, 256×256 pixel sample areas were excited with $50\,\mathrm{MHz}$ $473\,\mathrm{nm}$ laser pulses and detected using a $495\,\mathrm{nm}$ LP and a 620/60 nm HQ filter and 0.5–2.0 mm pinhole. Each sample was measured at least 3 times for $240\,\mathrm{s}$ each measurement. Fluorescence emission spectra were performed using a Cary Eclipse Fluorescence Spectrophotometer (Varian, USA), and were obtained using an excitation of $480\,\mathrm{nm}$ and a gain of $400\,\mathrm{V}$.

A.1.5 Mass Spectrometry

The samples for mass spectrometry were prepared by dissolving Cys at a concentration of $200\,\mathrm{mg\,mL^{-1}}$ in D_2O and H_2O separately. This solution was heated to $90\,^{\circ}\mathrm{C}$ and then allowed to cool gradually. Mass spectrometry analysis was carried out using an Acquity UPLC system connected to a TQD XEVO triple quadrupole ESI source mass spectrometer (Waters, Milford, MA, USA). The measurements were performed in positive ionization mode using ESI-MS.

A.1.6 Single-crystal X-Ray Diffraction

Preparation of crystal data collection Crystal structures of the Cys molecule were obtained from the Single crystal X-ray diffraction (SCXRD) technique. The following method was used to obtain the crystal structure:

Crystals used for data collection were grown using the slow solvent evaporation method. The dry Cys amino acid powder was first dissolved in D_2O at concentrations of $200\,\mathrm{mg\,mL^{-1}}$. Then, the samples were heated at $90\,^{\circ}\mathrm{C}$ for 3 hrs followed by a vortex to dissolve them completely and allowed to cool down gradually. The formation of a needle-like crystal of Cys took place after a few days by slow evaporation of the D_2O solvent. For data collection, crystals were coated in Paratone oil (Hampton Research), mounted on a MiTeGen cryo-loop, and flash-frozen in liquid nitrogen. Single crystal diffraction data were collected at $120\,\mathrm{K}$ on a Rigaku XtaLAB diffractometer system with Dectris Pilatus 3R 200K-A detector using CuKα radiation at $\lambda = 1.541\,84\,\text{Å}$.

Crystal data processing and structural refinement The diffraction data were collected and processed using the CrysAlisPro 1.171.42.67a suite of programs (RigakuOD 2022). The structures were solved by direct methods using SHELXT-2016/4 and refined by full-matrix least-squares against F2 with SHELXL-2016/4. Atoms were refined independently and anisotropically, with the exception of hydrogen atoms, which were located in the electron density and refined isotropically. Crystal data collection and refinement parameters are shown in Supplementary Table A.1, and the complete data can

be found in the cif file as Supporting Information. The crystallographic data have been deposited in the Cambridge Crystallographic Data Centre (CCDC) with ID 2358399 for Cys.

Accession Codes The CCDC database under accession code 2358399 contains the supplementary crystallographic data of the Cys crystals in D_2O for this paper.

A.1.7 Data collection details for the Cysteine crystal

A Cys crystal suitable for X-ray diffraction was immersed in Paratone–N oil and mounted in a MiTeGen loop and flash-frozen in liquid nitrogen at $120\,\mathrm{K}$. Data were collected with a Rigaku XtaLabPro with microfocus sealed tube source and a Dectris Pilatus 3R 200K-A detector using CuKα radiation ($1.541\,84\,\text{Å}$). Data were collected as ω scans using CrysAlisPro 1.171.42.67a suite of programs (Rigaku OD, 2022). Data were integrated and reduced in CrysAlisPro. The crystal structure was solved by using direct methods (SHELXT-2016/4) [411] and refined by full-matrix least-squares methods against F^2 (SHELXL-2016/4) [412] as implemented in Olex2 [413]. All the non-hydrogen atoms were refined with anisotropic displacement parameters. The hydrogen atoms were located in the electron density and refined isotropically. The Mercury 2020.3.0 software was used for molecular graphics.

A.1.8 Cys-D Crystal Structure Details

Table A.1: *Crystallographic data collection and refinement statistics data for the crystal of Cysteine in D_2O at* 120 K.

Compound	Cysteine (D_2O)
CCDC number	2358399
Crystal description	Colorless Needle
Diffractometer	Rigaku XtaLabPro
Temperature (K)	120
Chemical Formula	$C_6H_{14}N_2O_4S_2$
Formula weight (g/mol)	242.31
Radiation	CuKα ($\lambda = 1.541\,84$ Å)
z	2
Crystal System	Orthorhombic
Bond-length (C-C) (Å)	0.0035
Space group	$P2_12_12_1$
a (Å)	5.41839(11)
b (Å)	8.13325(17)
c (Å)	12.0157(2)
α (°)	90
β (°)	90
γ (°)	90
Volume (Å^3)	529.520(18)
ρ (g/cm^3)	1.520
μ (mm^{-1})	4.552
$F(000)$	256.0
Crystal size/mm^3	0.355 × 0.077 × 0.073
2Θ range for data collection/°	13.144 to 158.22
Index ranges	$-3 \leq h \leq 6$, $-10 \leq k \leq 10$, $-14 \leq l \leq 15$
Reflections collected	2622
Independent reflections	1128 [$R_{\text{int}} = 0.0283$, $R_{\text{sigma}} = 0.0311$]
Data/restraints/parameters	1128/0/92
Goodness-of-fit on F^2	1.120
Final R indexes [I≥ 2σ (I)]	$R_1 = 0.0336$, w$R_2 = 0.0914$
Final R indexes [all data]	$R_1 = 0.0337$, w$R_2 = 0.0914$
Largest diff. peak/hole / e Å^{-3}	0.28/-0.43
Flack parameter	0.02(2)

A.1.9 Cys-H Crystal Structure Details

Table A.2: *Crystallographic data collection and refinement statistics data for the crystal of Cysteine in H_2O at* 120 K.

Compound	Cysteine (H_2O)
CCDC number	2444375
Crystal description	White crystalline plate
Diffractometer	Rigaku XtaLabPro
Temperature (K)	120 K
Chemical Formula	$C_3H_7NO_2S$
Formula weight (g/mol)	121.16
Radiation	CuKα ($\lambda = 1.541\,84$ Å)
z	4
Crystal System	Monoclinic
Bond-length (C-C) (Å)	0.0060
Space group	$P2_1$
a (Å)	9.4370(3)
b (Å)	5.21640(10)
c (Å)	11.3434(3)
α (°)	90
β (°)	109.040(3)
γ (°)	90
Volume (Å^3)	527.85(3)
ρ (g/cm^3)	1.525
μ (mm^{-1})	4.566
$F(000)$	256.0
Crystal size/mm^3	0.2 × 0.04 × 0.01
2Θ range for data collection/°	8.246 to 149.3
Index ranges	$-11 \leq h \leq 11$, $-6 \leq k \leq 6$, $-14 \leq l \leq 14$
Reflections collected	9099
Independent reflections	2127 [$R_{\text{int}} = 0.0442$, $R_{\text{sigma}} = 0.0356$]
Data/restraints/parameters	2127/1/164
Goodness-of-fit on F^2	0.765
Final R indexes [I≥ 2σ(I)]	$R_1 = 0.0442$, w$R_2 = 0.1127$
Final R indexes [all data]	$R_1 = 0.0461$, w$R_2 = 0.1146$
Largest diff. peak/hole / e Å^{-3}	0.60/-0.39
Flack parameter	0.021(17)

A.2 Validation of Computational Methods

One important aspect to highlight is that our calculated excitation energy, corresponding to the transition to the first excited state, exhibits a blue shift of 195 nm $(2.84\,\mathrm{eV})$ compared to the experimental value. A potential source of this discrepancy is the choice of functional and basis set [414]. However, we validated our approach by testing different combinations of basis sets and functionals for the TDDFT calculations, as well as the algebraic diagrammatic construction to second-order (ADC(2)) method [415, 416]. As we show here, our def2-SVP/CAM-B3LYP approach is consistent with a variety of other methods including def2-SVP/wB97X-D3 (Figure A.8**a**), def2-QZVPP/CAM-B3LYP (Figure A.8**b**), def2-SVP/ADC(2) (Figure A.8**c**). In all of these cases, the lowest excitation energies corresponding to $S_0 \rightarrow S_1$ transitions are relatively similar for both systems. The nature of the transitions also largely remain the same, and are dominated by the $n \rightarrow \pi^*$ on the C–O and $n \rightarrow \sigma^*$ on the S–H. In prior work done in our group [38], we have observed that TDDFT level of theory provides a great cost/accuracy tradeoff while still reproducing the essential features obtained by using higher-level, albeit much costlier, multireference methods.

Another possible source of error is the size of the QM subsystem, the choice of which involves a delicate balance of computational feasibility and accuracy. Previous work done by our group [38, 6] has employed QM dimers for a variety of amino acid crystals and found that these reasonably faithfully reproduce the key experimentally observed behavior. In the case of cysteine, there still remains a noticeable blue-shift of the absorption spectra and the lowest excitation energy. In order to understand how sensitive the absorption spectra are to the size of the QM region, we performed some additional benchmarks. Indeed, Figure A.8**d** demonstrates that increasing the QM region from a dimer to a heptamer (which contains 98 atoms) results in a red-shift of the spectra and subsequently reduces the discrepancy between the theory and experiment to $\sim 2.1\,\mathrm{eV}$ while preserving the same physical origin for the transition to the first excited state.

Finally, we wanted to verify the impact of both a larger QM region (Figure A.9**a**), as well as a different electronic structure method (ADC(2)) near the conical intersection (Figure A.9**b**). ADC(2) is widely considered the most accurate single-reference method for describing conical intersections [196, 417]. As previously mentioned, in our TDDFT/TSH setup, we set the criterion for a hop to the ground state when the energy gap falls below $0.2\,\mathrm{eV}$. In the validation using a larger QM region, we see that the energy gaps between the first excited state and the ground state are preserved remarkably well when using pentamers instead of dimers as the QM region. In the second test, we applied ADC(2) using the same QM dimer as in the TDDFT case. While some differences emerge compared to TDDFT, the overall energy gaps and conical intersection remain largely unchanged.

A.3 Details of excitation and non-radiative decay modes

Table A.3: *A summary of the nature of $S_0 \rightarrow S_1$ excitations for Cys-H and Cys-D.*

Nature of $S_0 \rightarrow S_1$ excitation	Cys-H	Cys-D
$n \rightarrow \pi*$ (C−O)	23	29
$n \rightarrow \sigma*$ (S−H)	17	11

Table A.4: *A summary of the different modes of non-radiative decay for Cys-H and Cys-D.*

Non-radiative decay mode	Cys-H	Cys-D
S−H	12	9
C_α−N	19	-
C_β−S	4	1
C_α−COO^-	-	3

Table A.5: *Summary of Quantum Mechanical (QM) calculation parameters in Fig. A.8*

QM subsystem	**QM method**	**Functional**	**Basis Set**	**Figure**
Dimer	TD-DFT	CAM-B3LYP	def2-SVP	3.3**a**
Dimer	TD-DFT	**wB97X-D3**	def2-SVP	A.8**a**
Dimer	TD-DFT	CAM-B3LYP	**def2-QZVPP**	A.8**b**
Dimer	**ADC(2)**	N/A	def2-SVP	A.8**c**
Heptamer	TD-DFT	CAM-B3LYP	def2-SVP	A.8**d**

A.4 Supplementary Figures

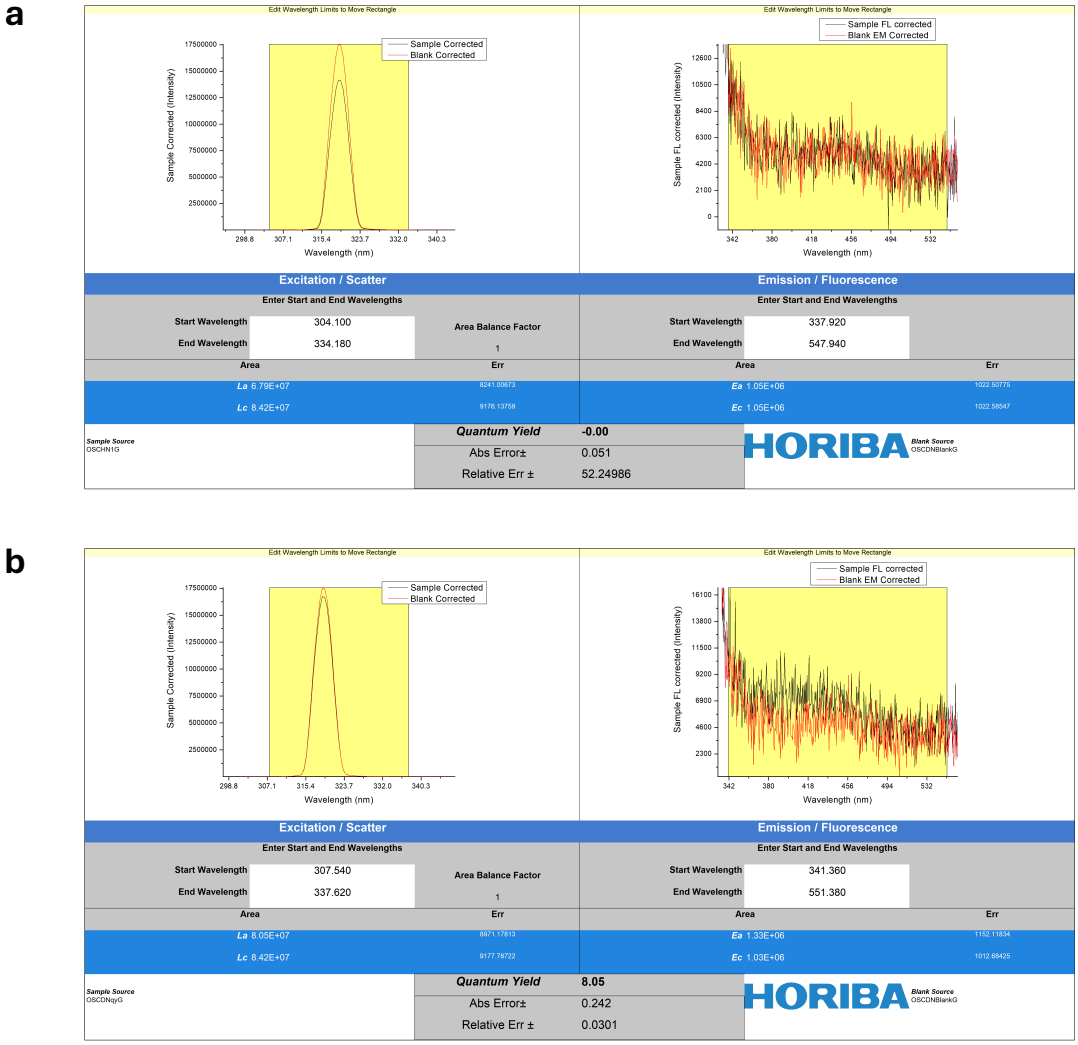


Figure A.1: *Quantum yield of Cys-H (**a**) and Cys-D (**b**) measured under excitation of 320 nm for multiple crystals.*

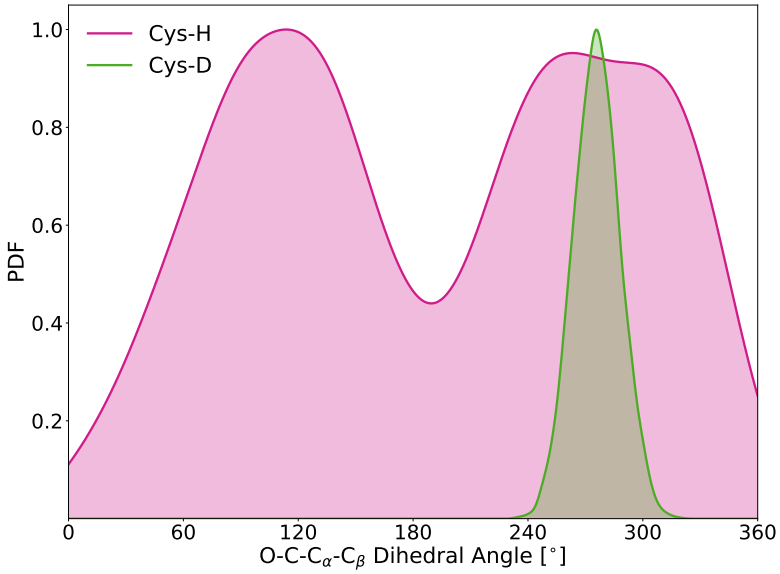


Figure A.2: *A distribution of the $O-C-C_\alpha-C_\beta$ dihedral angle for both Cys-H (pink) and Cys-D (green) from finite-temperature classical NVT simulations in the ground state.*

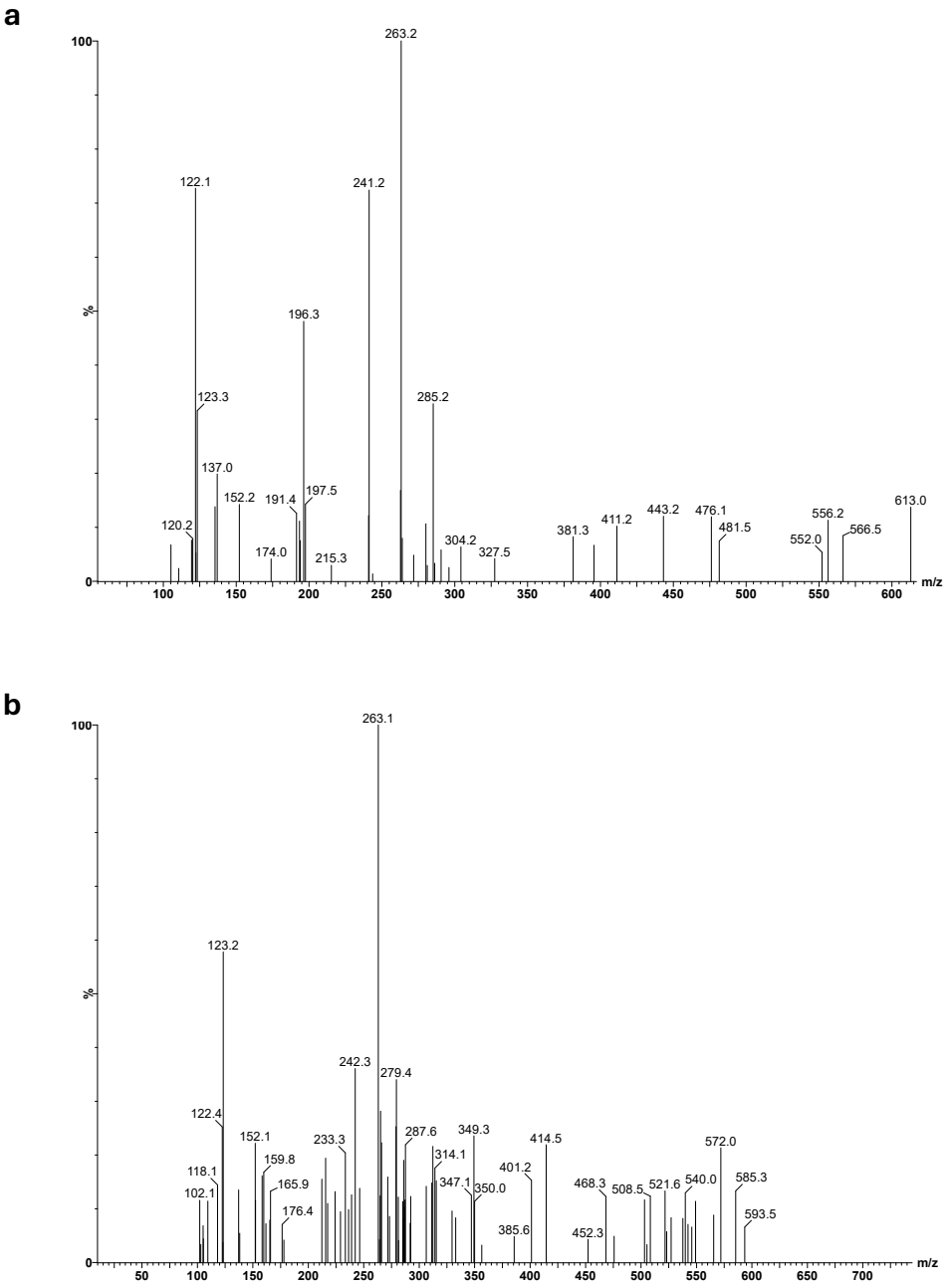


Figure A.3: *Mass spectroscopy for Cys-H (**a**) and Cys-D (**b**) in positive-ion mode.*

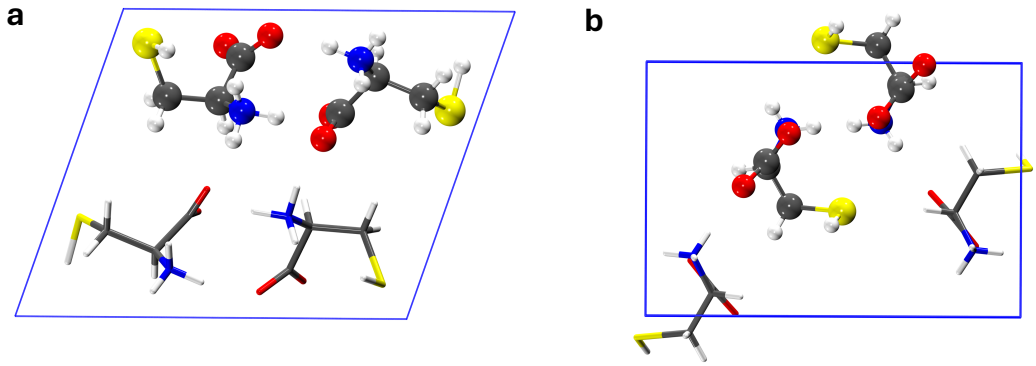


Figure A.4: *Single unit cell of Cys-H (panel **a**) and Cys-D (panel **b**) with the dimers chosen as the relevant QM subsystem in the QM/MM setup employed highlighted in "CPK" representation.*

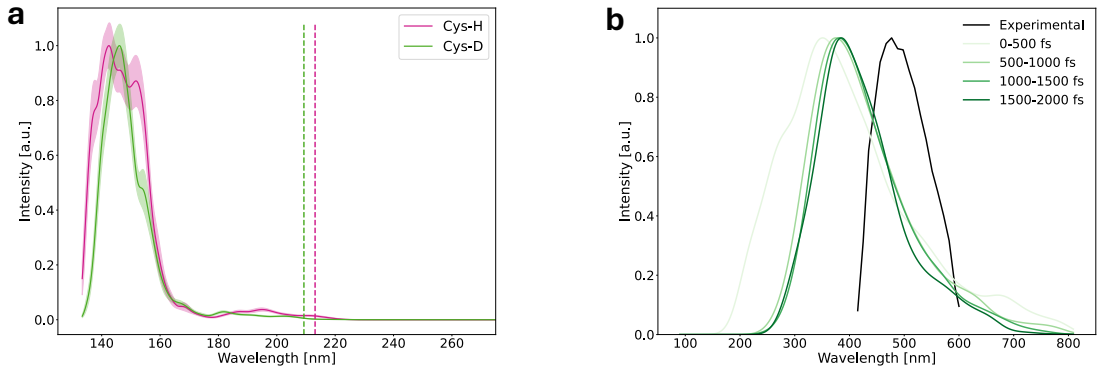


Figure A.5: *Panel **a** shows the theoretically computed absorption spectra for Cys-H and Cys-D in pink and green, respectively. The shaded region denotes error bars over 40 independent estimates. The dashed vertical lines denote positions of the lowest energy excitations* $(S_0 \longrightarrow S_1)$ *which are* $215\,\mathrm{nm}$ *and* $209\,\mathrm{nm}$ *for Cys-H and Cys-D, respectively. Panel **b** shows the time-dependent emission spectra from* S_1 *for Cys-D, in blocks of 500 fs, with the shades of green getting darker for successive blocks of time, while the experimental emission spectra is shown in black.*

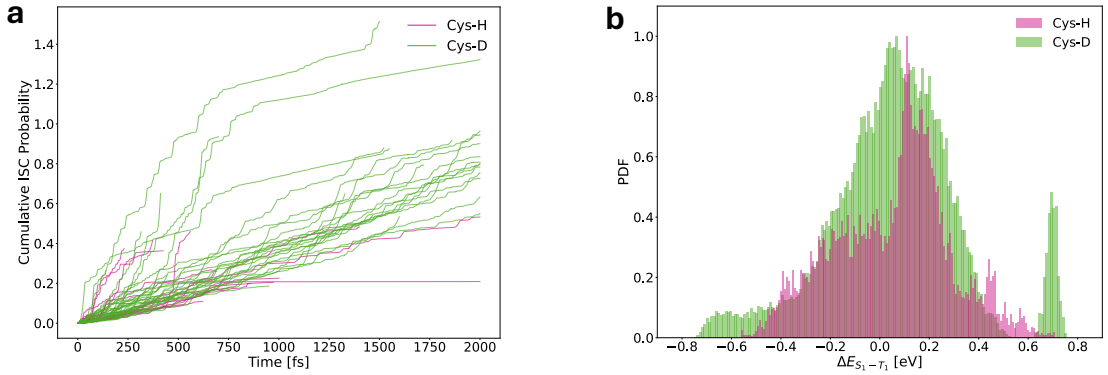


Figure A.6: *Panel **a** shows the time-evolution of the cumulative ISC probability from* S_1 *to any of the triplet states for each NAMD trajectory. Panel **b** shows a distribution of the energy gaps between* S_1 *and* T_x *(which denotes the energetically closest triplet state, usually* T_1*). The data for Cys-H and Cys-D is shown in pink and green, respectively.*

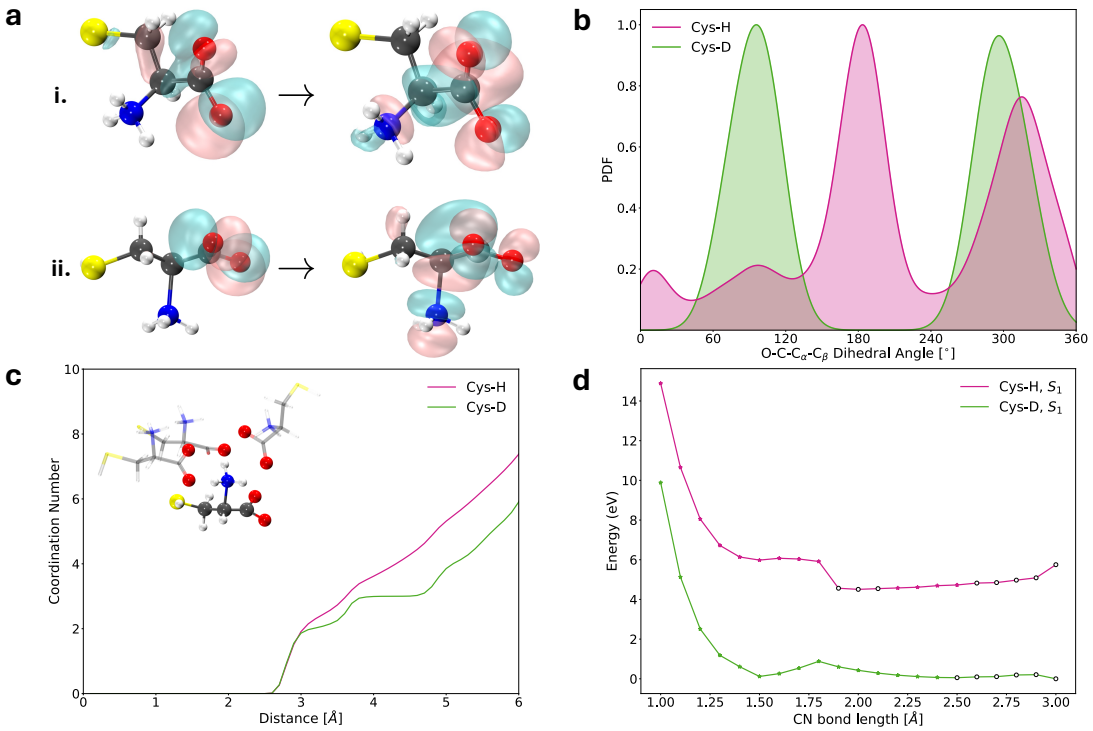


Figure A.7: *Panel **a.i** shows the occupied (left) and virtual (right) orbitals for the $S_0 \rightarrow S_1$ transition at the starting frame of a reference NAMD trajectory where Cys-H decays non-radiatively to S_0 using the C-N stretching mode. Panel **a.ii** shows the same orbitals, for the same NAMD trajectory, at the frame of encountering the conical intersection.*
*Panel **b** shows a distribution of the $O-C-C_\alpha-C_\beta$ dihedral angle for both Cys-H (pink) and Cys-D (green) while the system is in S_1.*
*Panel **c** displays the running coordination number of Nitrogen-Oxygen for Cys-H (pink) and Cys-D (green). The inset of the figure shows a sample frame from an NAMD trajectory where the it is clear that the NH_3^+ stretches into a "cavity" of O atoms.*
*Panel **d** shows the excited state relaxed potential energy scan for different values of the $N-C_\alpha$ bond with the lines showing the energies for the optimized structures in S_1 for each bond length scan value, for Cys-H (pink) and Cys-D (green). The small circles represent points where the geometry optimization has not converged even after 1500 optimization cycles, which are usually those very close to the conical intersection. For clarity of visualization, the pink curve corresponding to Cys-H is shifted up by $5\,\mathrm{eV}$.*

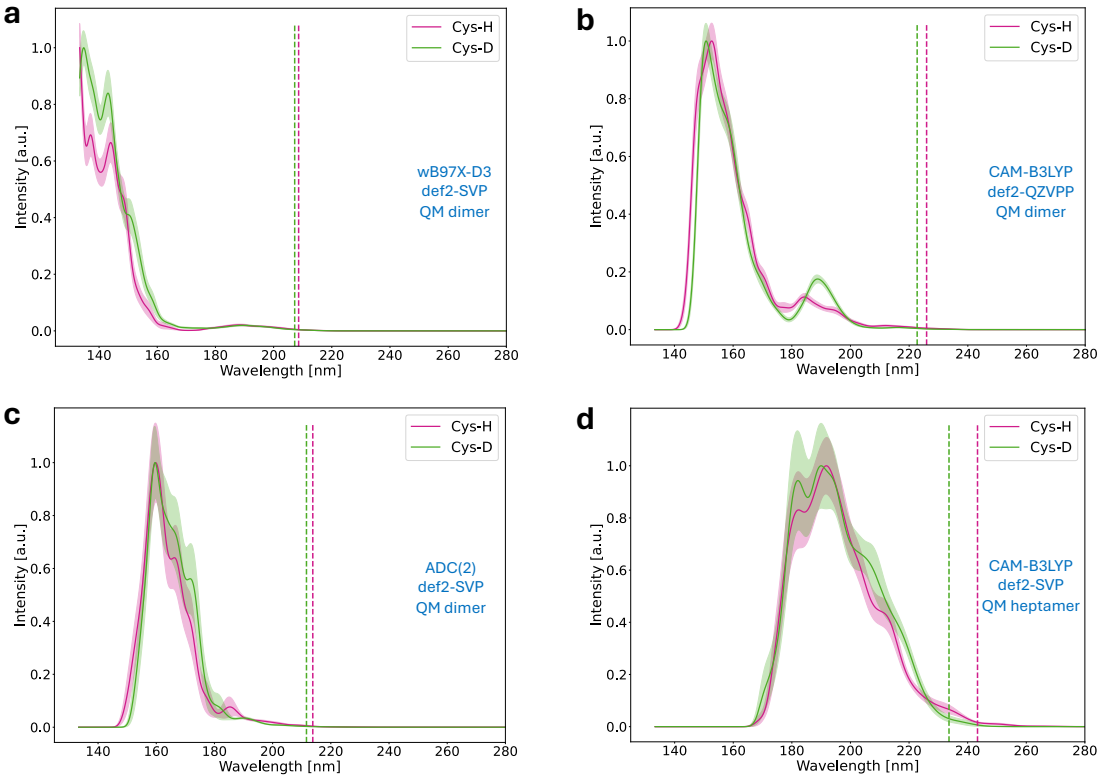


Figure A.8: *This figure shows the results of a variety of theoretical absorption spectra calculations to test the accuracy of our choice of electronic structure and QM/MM setup. In all panels the shaded curves in pink and green correspond to Cys-H and Cys-D respectively, while the dashed vertical lines correspond to the lowest energy excitations ($S_0 \rightarrow S_1$), following the same colorscheme.*
These calculations are benchmarked against the primary level of theory used for absorption spectra and NAMD calculations in this work (CAM-B3LYP/def2-SVP).
The table below summarizes the different cases explored here, each involving a change to one element (functional, basis set, QM subsystem, or electronic structure method) relative to the primary level of theory.
In all cases the calculations are done using ORCA 5.0.4.

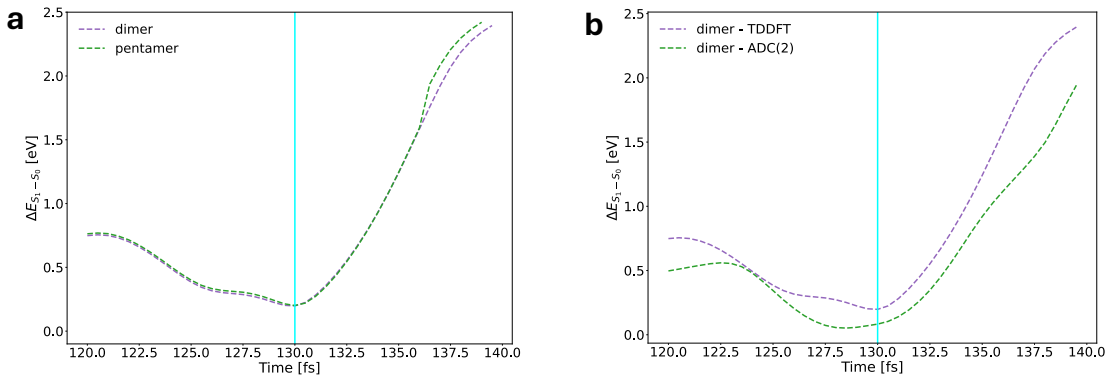


Figure A.9: *Panel **a** shows the energy gap between S_1 and S_0 for a Cys-H trajectory decaying via the $C-N$ mode, near the conical intersection, where the original energy gap considering dimers in the QM region is shown with purple dashed lines, and the recomputed energy gap expanding the QM region to a pentamer is shown in green dashed lines. These are both computed using TDDFT.*
*Panel **b** shows the same energy gap for the QM dimer as in **a**, computed at both TDDFT (purple) and ADC(2) (green) levels of theory using ORCA 5.0.4.*
Here it should be noted that while attempting the validation of the pentamer energies using ADC(2), as implemented in ORCA 5.0.4, we ran into issues pertaining to computational cost.
We reiterate here that since TDDFT has issues describing the CI between S_1 and S_0, we choose a cutoff value of $0.2\,\mathrm{eV}$ to determine a hop to the ground state.

B

Supplementary Information: Machine learning the non-radiative decay modes in photochemical processes

B.1 Additional information on the studied systems

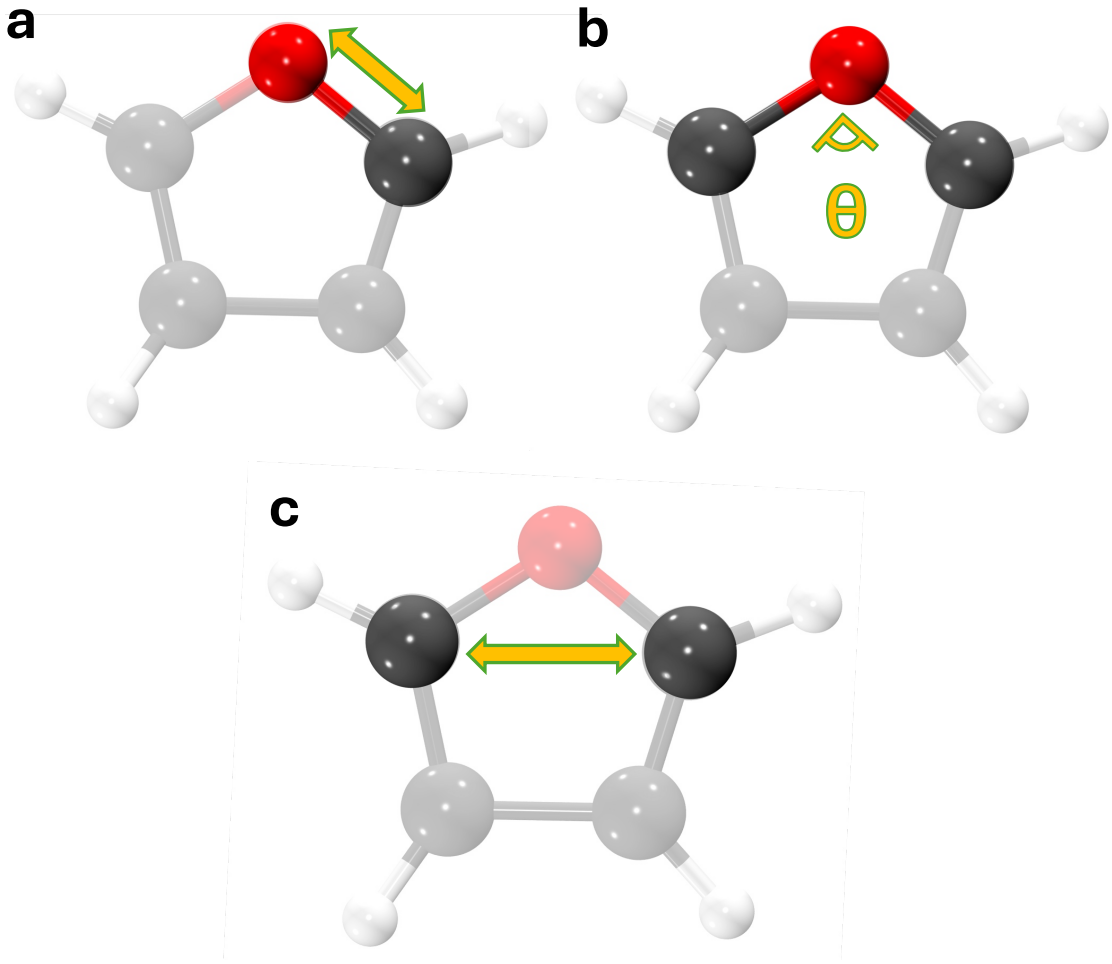


Figure B.1: ***Furan:*** *Additional internal-coordinate modes for furan that explain the* $S_1 \rightarrow S_0$ *non-radiative relaxation:* (a) $C-O$ *stretch.* (b) $C-O-C$ *angle.* (c) $C-C$ *separation that corresponds to the ring-opening mode.*

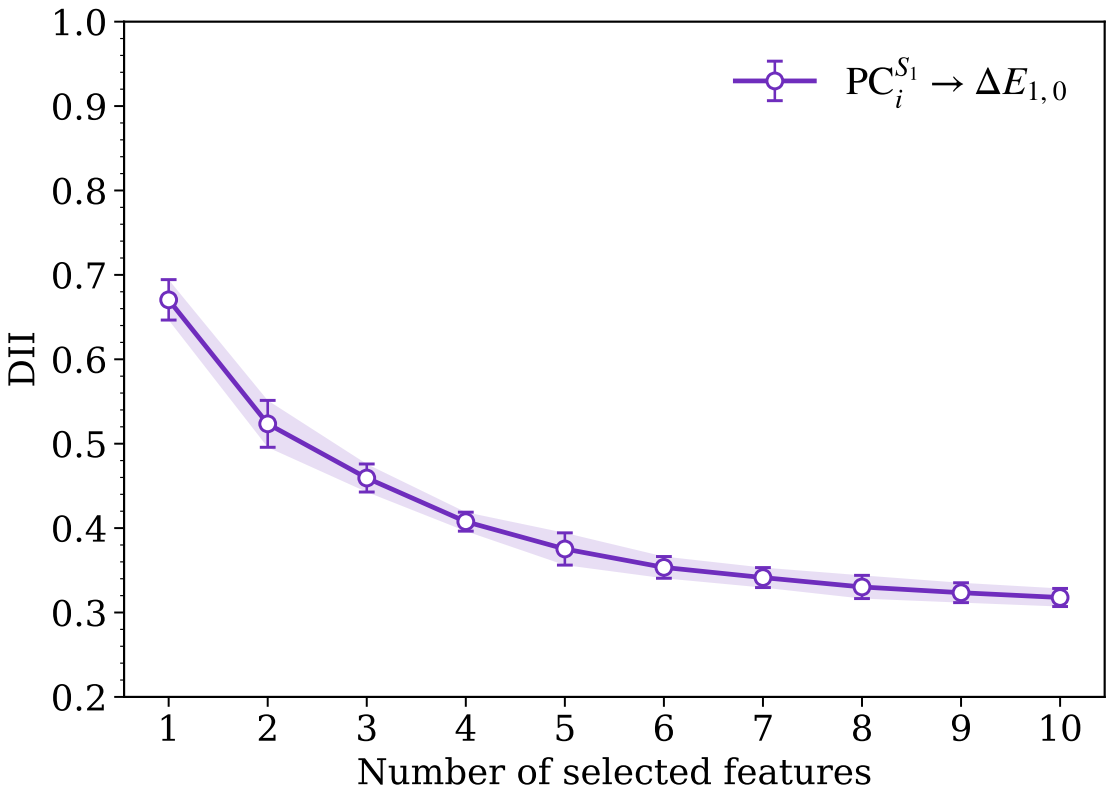


Figure B.2: ***Furan:*** *DII between the leading PCs and the energy gap* ($\Delta E_{1,0}$) *as a function of the number of selected features. Shaded areas show error bars computed from 50 independent bootstrap resamples. The inclusion of PCs continues to lower the DII to* $\sim$ 0.32*. This indicates that, even when a select number of chemically intuitive coordinates already captures part of the dynamics, additional predictive information is contained in more delocalized collective modes. Such behavior reflects the inherently high-dimensional and strongly coupled nature of nonadiabatic dynamics, where multiple nuclear degrees of freedom contribute simultaneously to the evolution toward conical intersections.*

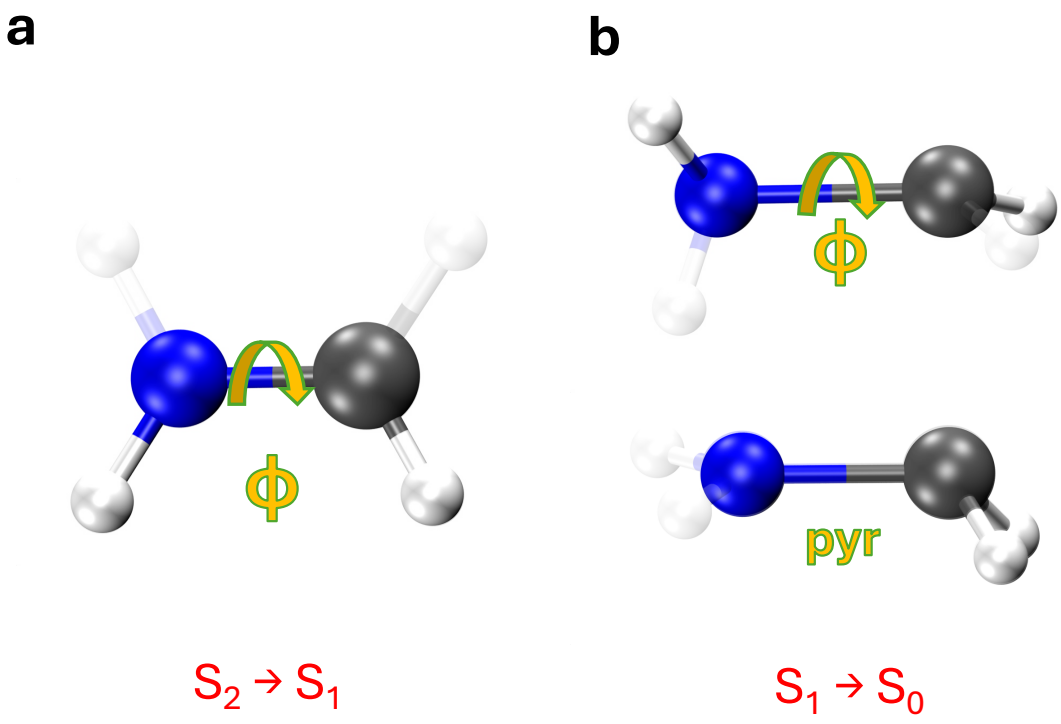


Figure B.3: ***Methaniminium cation:*** *Additional internal-coordinate modes for the methaniminium cation: (a) Mode relevant for the $S_2 \longrightarrow S_1$ transition, which involves a torsion around the $C-N$ bond. (b) Modes relevant for the $S_1 \longrightarrow S_0$ transition, which involve a torsion around the $C-N$ bond as well as the pyramidalization of the CH_2 group.*

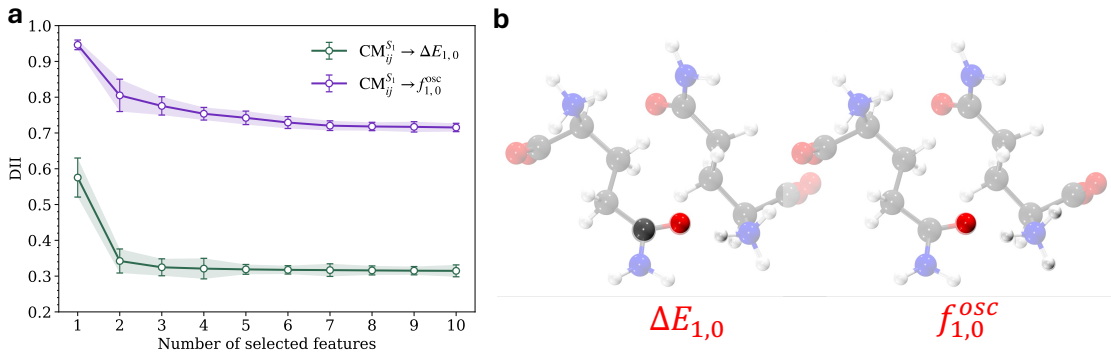


Figure B.4: ***L-glutamine:*** *(a) DII between Coulomb-matrix descriptors and the energy gap/oscillator strength as a function of the number of selected features. Purple and green correspond to $f_{1,0}^{\mathrm{osc}}$ and $\Delta E_{1,0}$, respectively. Shaded areas show error bars computed from 50 independent bootstrap resamples. (b) Most informative atom pairs ("hotspots") identified from the Coulomb-matrix analysis. The left and right parts of the panel correspond to $\Delta E_{1,0}$ and $f_{1,0}^{\mathrm{osc}}$, respectively.*

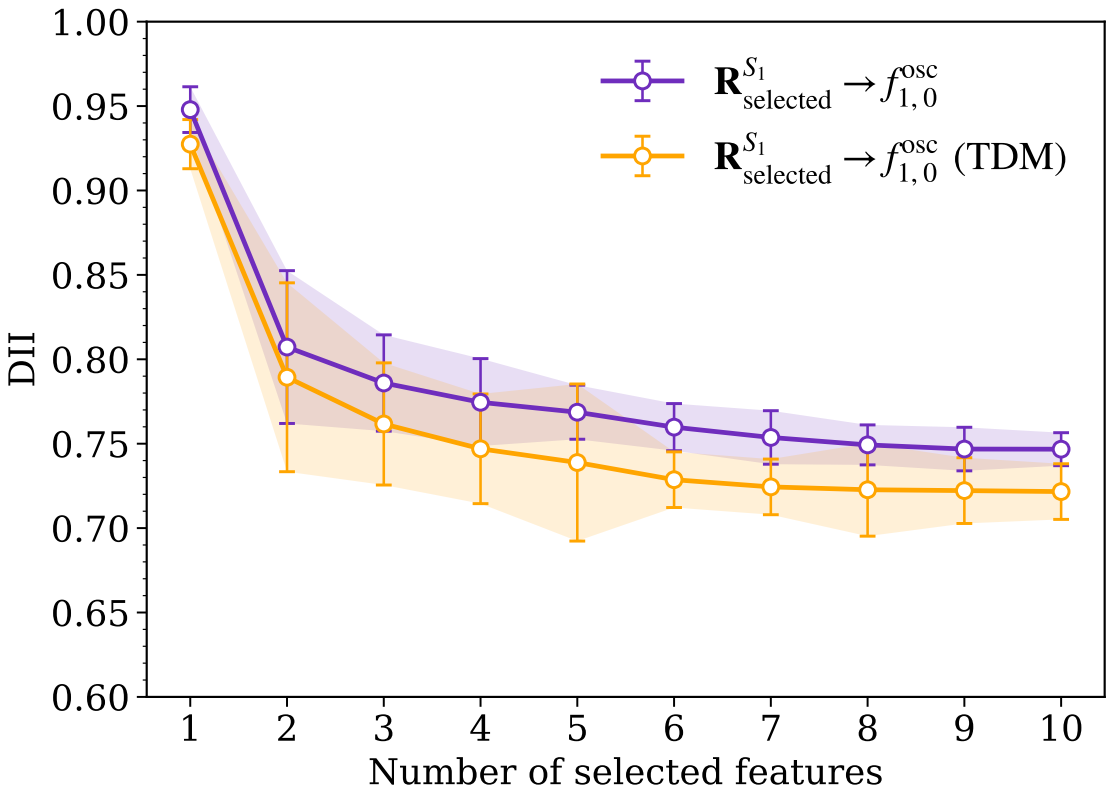


Figure B.5: ***L-glutamine:*** *DII between internal coordinates and the oscillator strength ($f^{\text{osc}}_{1,0}$), and between internal coordinates and the part of the oscillator strength arising from the transition dipole moment only ($f^{\text{osc}}_{1,0}$(TDM)), as a function of the number of selected features (purple and orange, respectively). Shaded areas show error bars computed from 50 independent bootstrap resamples.*

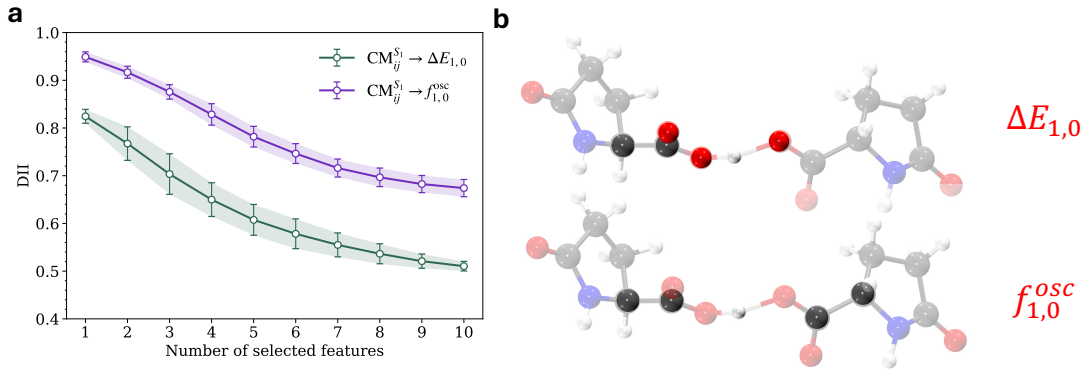


Figure B.6: ***L-pyro:*** *(a) DII between Coulomb-matrix descriptors and the energy gap/oscillator strength as a function of the number of selected features. Purple and green correspond to $f_{1,0}^{\mathrm{osc}}$ and $\Delta E_{1,0}$, respectively. Shaded areas show error bars computed from 50 independent bootstrap resamples. (b) Most informative atom pairs ("hotspots") identified from the Coulomb-matrix analysis. The top and bottom parts of the panel correspond to $\Delta E_{1,0}$ and $f_{1,0}^{\mathrm{osc}}$, respectively.*

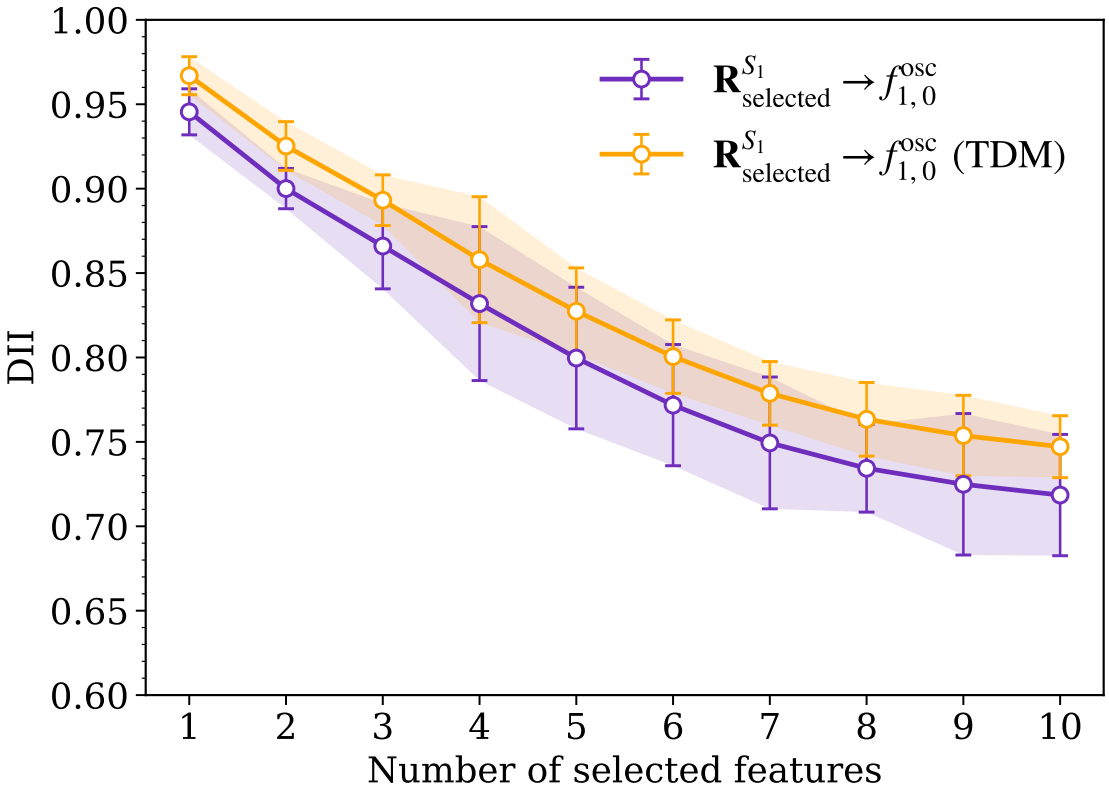


Figure B.7: ***L-pyro:*** *DII between internal coordinates and the oscillator strength ($f_{1,0}^{\mathrm{osc}}$), and between internal coordinates and the part of the oscillator strength arising from the transition dipole moment only ($f_{1,0}^{\mathrm{osc}}$(TDM)), as a function of the number of selected features (purple and orange, respectively). Shaded areas show error bars computed from 50 independent bootstrap resamples.*

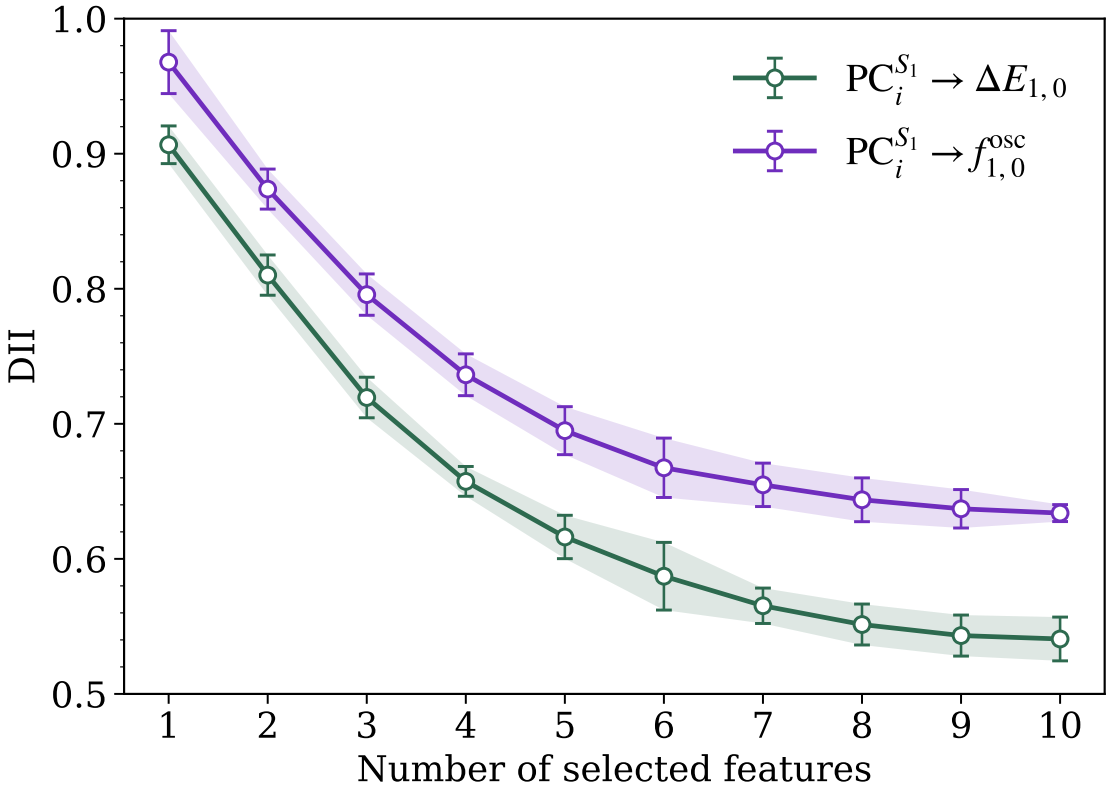


Figure B.8: ***L-pyro:*** *DII between the leading PCs and the energy gap/oscillator strength as a function of the number of selected features. Purple and green correspond to $f_{1,0}^{\mathrm{osc}}$ and $\Delta E_{1,0}$, respectively. Shaded areas show error bars computed from 50 independent bootstrap resamples.*

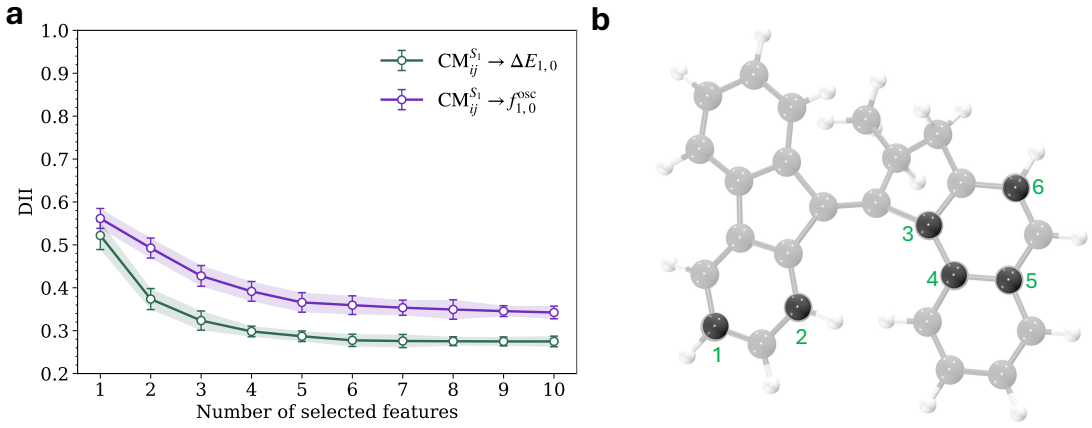


Figure B.9: ***Molecular motor:*** *(a) DII between Coulomb-matrix descriptors and the energy gap/oscillator strength as a function of the number of selected features. Purple, red, and green correspond to $f_{1,0}^{\mathrm{osc}}$, $f_{1,0}^{\mathrm{osc}}(\mathrm{TDM})$, and $\Delta E_{1,0}$, respectively. Shaded areas show error bars computed from 50 independent bootstrap resamples. (b) Most informative atom pairs ("hotspots") identified from the Coulomb-matrix analysis. The atoms involved in the hotspots for both $\Delta E_{1,0}$ and $f_{1,0}^{\mathrm{osc}}$ are highlighted together and numbered to distinguish the different C atoms.*

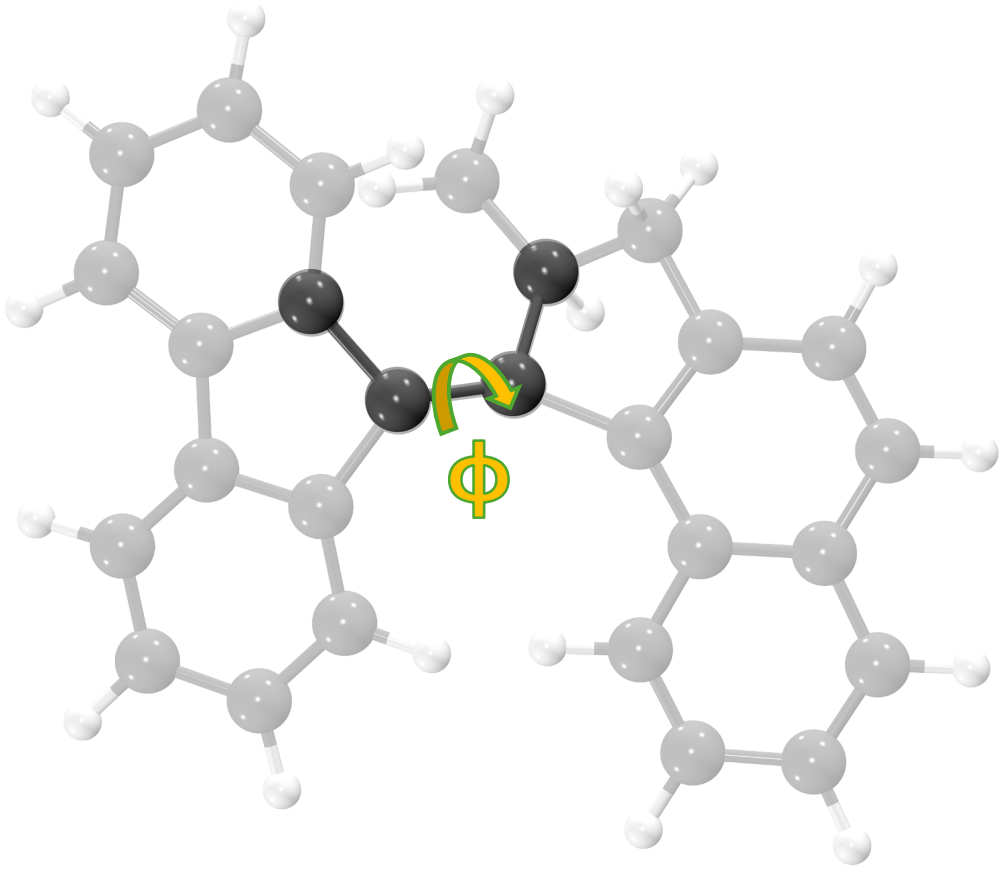


Figure B.10: ***Molecular motor:*** *Additional internal-coordinate mode, involving a secondary torsion around the central dihedral, that explains both the* $S_1 \rightarrow S_0$ *energy gap* ($\Delta E_{1,0}$) *and the oscillator strength* ($f_{1,0}^{\mathrm{osc}}$).

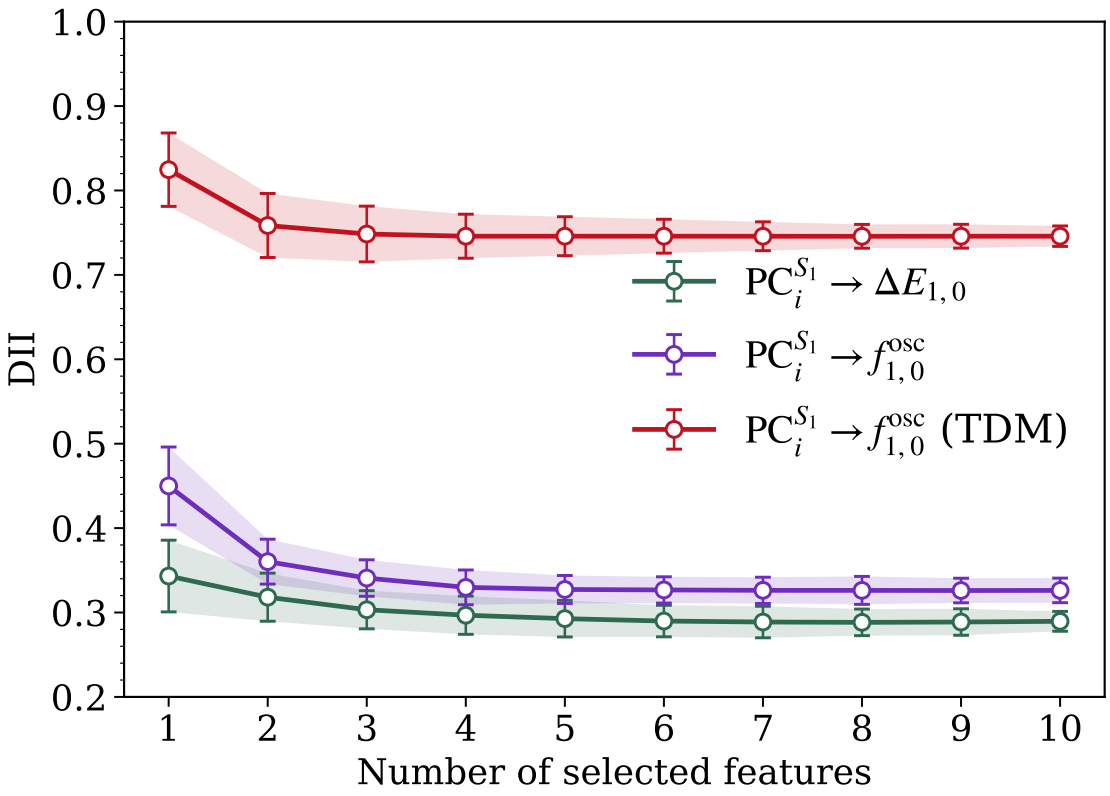


Figure B.11: ***Molecular motor:*** *DII between the leading PCs and the energy gap/oscillator strength as a function of the number of selected features. Purple and green correspond to $f_{1,0}^{\mathrm{osc}}$ and $\Delta E_{1,0}$, respectively. Shaded areas show error bars computed from 50 independent bootstrap resamples.*

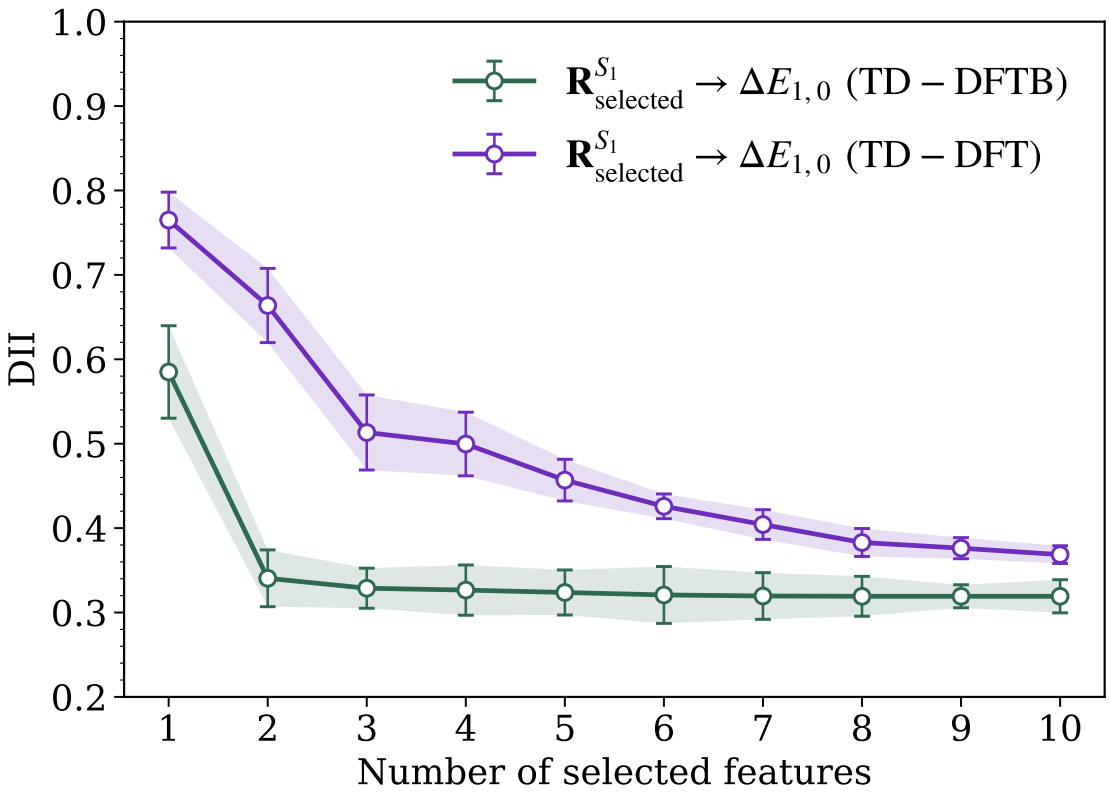


Figure B.12: ***L-glutamine:*** *Comparison of the DII analysis using trajectories generated with TD-DFTB and TD-DFT. The DII between the selected S_1 internal coordinates ($\mathbf{R}^{S_1}_{\text{selected}}$) and the S_1–S_0 energy gap ($\Delta E_{1,0}$) is shown as a function of the number of selected features. Both analyses recover the same leading structural modes, including the carbonyl C−O stretch and the proton-transfer-related O−H distance. The TD-DFTB curve (green) decreases rapidly with the first two features, whereas the TD-DFT curve (purple) decreases more gradually. Shaded areas show error bars computed from 50 independent bootstrap resamples. With ten selected features, the DII reaches approximately 0.32 and 0.37 for TD-DFTB and TD-DFT, respectively, demonstrating consistent mode identification alongside differences in quantitative informativeness.*

B.1.1 Cysteine crystals

We applied our approach to NAMD simulations generated in a previous study [1] on crystals of the amino acid cysteine, motivated again by the observation of non-aromatic fluorescence. In that study, which combined simulations and experiments, it was shown that the non-radiative decay of cysteine crystallised in light water (Cys-H) and heavy water (Cys-D) proceeds through two photodissociative pathways: the dissociation of a proton from the sulphur atom, and the cleavage of the C−N bond. The C−N pathway is suppressed in Cys-D, which leads to the fluorescence observed in its crystals. The nonadiabatic simulations in the original study used a QM/MM approach with two cysteine monomers in the QM region. Initial excitations were approximately equally distributed between these monomers, so each type of informative coordinate appears twice in the selected feature set (Figure B.13).

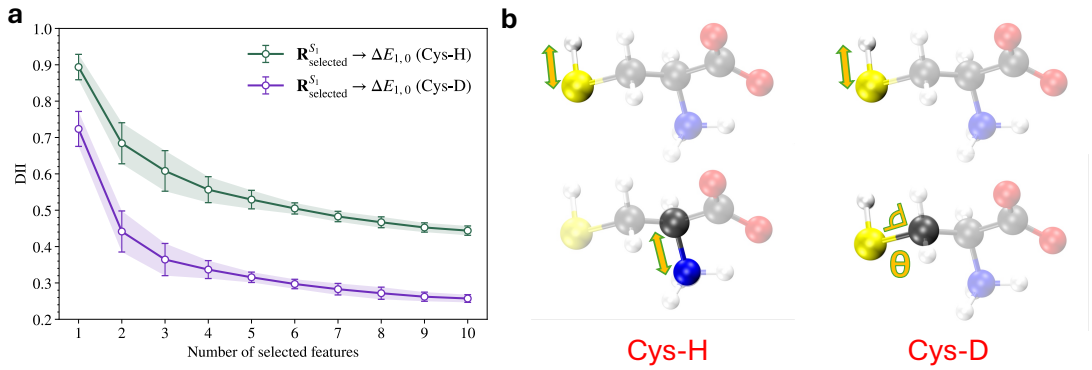


Figure B.13: ***Cysteine:*** *(a) DII between internal coordinates and the energy gap as a function of the number of selected features. Purple and green correspond to cysteine crystallised in D_2O (Cys-D) and in H_2O (Cys-H), respectively. Shaded areas show error bars computed from 50 independent bootstrap resamples. (b) The most informative structural modes extracted from the DII analysis. For Cys-H, the four leading features comprise the $S-H$ and $C-N$ distances on both monomers and yield a DII of approximately 0.55. For Cys-D, the four leading features comprise the two $S-H$ distances and the two $C-S-H$ angles, yielding a DII of approximately 0.30. For clarity, one representative coordinate of each type is illustrated. The subsequent decrease in DII upon including further coordinates is more pronounced for Cys-H, indicating that additional structural information improves the description of its energy gap.*

B.2 Benchmarking and sensitivity analysis

B.2.1 Convergence with the number of trajectories

An important practical consideration is the sensitivity of the analysis to the size and phase-space coverage of the trajectory ensemble. We therefore repeated the complete DII analysis for furan using ensembles containing 20, 50, 75, 100, 150, and 200 trajectories, while retaining the same descriptor construction, hyperparameters, and feature-selection procedure. As shown in Figure B.14, all ensemble sizes produce a rapid decrease in DII over the first five to six selected coordinates, followed by a plateau. For the complete ensemble of 200 trajectories, the DII decreases from approximately 0.82 with one selected feature to approximately 0.45 with ten features. The corresponding ten-feature values for the smaller subsets lie between 0.25 and 0.35. Thus, the absolute DII values retain a finite-ensemble dependence and should not be interpreted as numerically converged across the different ensemble sizes. In particular, the lower values obtained for some of the smaller ensembles do not imply improved predictive performance; they result from overfitting when too many features are fitted on too few frames.

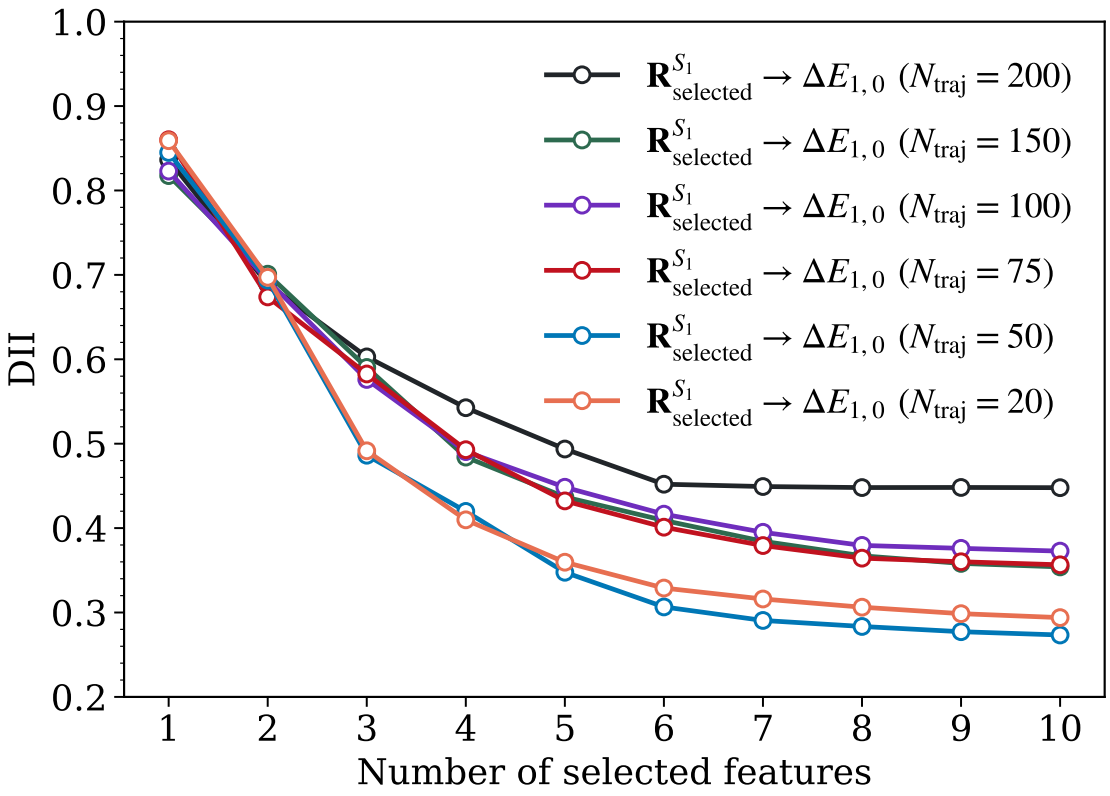


Figure B.14: ***Furan:*** *Sensitivity of the DII feature-selection analysis to the trajectory-ensemble size. The DII between the selected S_1 internal coordinates ($\mathbf{R}^{S_1}_{\text{selected}}$) and the S_1–S_0 energy gap ($\Delta E_{1,0}$) is shown as a function of the number of selected features for ensembles containing $N_{\text{traj}} = 20, 50, 75, 100, 150$, and 200 trajectories. All ensembles were analyzed using the same descriptor construction, hyperparameters, and feature-selection procedure. The curves exhibit a similar rapid decrease over the first five to six features, followed by a plateau. Although the absolute plateau value depends on the ensemble size, all ensembles recover the same five leading coordinates: the C–O–C–H dihedral, two C–O stretching coordinates, the C–O–C angle, and the C–C separation. These coordinates consistently describe the coupled out-of-plane puckering and ring-opening distortions associated with the $S_1 \rightarrow S_0$ relaxation pathway. Only the ordering of closely correlated coordinates changes between ensemble sizes. Error bars are omitted to avoid visual crowding; the uncertainty of each estimate is approximately 0.01–0.04.*

Nevertheless, the physical modes selected by the analysis are stable with respect to the trajectory-ensemble size. All six ensembles recover the same five leading coordinates: the C–O–C–H dihedral and the leading C–O stretching coordinate, together with the second C–O stretch, the C–O–C angle, and the C–C separation shown in Figure B.1(a)–(c). The C–O–C–H dihedral and C–O–C angle describe the out-of-plane, puckering-related deformation of the ring, whereas the two C–O stretching coordinates and the C–C separation describe the associated ring-opening distortion. Only the ordering of closely correlated coordinates at the level of the best pairs or triplets changes between ensemble sizes; the composition of the top-five feature set and its physical interpretation remain unchanged. The coupled puckering and ring-opening mechanism inferred from the complete ensemble is therefore recovered even with 20 trajectories, corresponding to 10% of the available trajectory set, although the absolute DII values remain sample-size dependent. These tests show that the method is highly data efficient in recovering the physically relevant coordinates.

B.2.2 Analysis restricted to the vicinity of the conical intersection

To assess whether the selected coordinates remain informative close to non-radiative transitions, we repeated the analysis using only configurations within ±20 fs of the hopping event. The restricted-window analysis recovers the same five leading coordinates: the C−O−C−H dihedral, the two C−O stretches, the C−O−C angle, and the C−C separation. The DII decreases most strongly over the first five to six features and reaches approximately 0.34 with ten features, compared with approximately 0.45 for the all-frames analysis (Figure B.15). Thus, the coupled puckering and ring-opening signatures persist close to the hopping event, supporting their relevance beyond differences between the early and late portions of the trajectories.

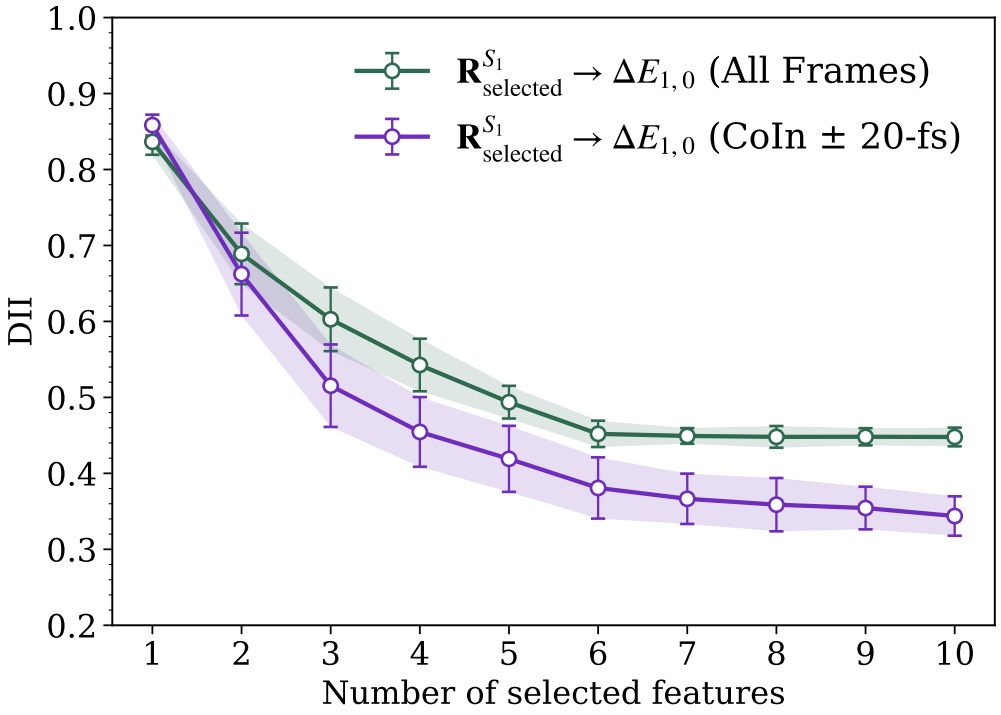


Figure B.15: ***Furan:*** *Sensitivity of the DII analysis to the time window around the hopping event. The DII between the selected S_1 internal coordinates ($\mathbf{R}^{S_1}_{\text{selected}}$) and the S_1–S_0 energy gap ($\Delta E_{1,0}$) is shown as a function of the number of selected features, using all available frames (green) or configurations within ±20 fs of the hopping event (purple). Shaded areas show error bars computed from 50 independent bootstrap resamples. Both analyses recover the same five leading coordinates associated with coupled puckering and ring-opening distortions: the C−O−C−H dihedral, the two C−O stretches, the C−O−C angle, and the C−C separation. Restricting the time window lowers the ten-feature DII from approximately 0.45 to 0.34 while preserving the main feature-selection trend.*

As for furan, we tested the sensitivity of the L-glutamine energy-gap analysis by restricting the configurations used to construct the initial dataset to those within ±20 fs of the hopping event. This restricted analysis (Figure B.16) also recovers the carbonyl C−O stretch and the proton-transfer-related O−H distance as the leading coordinates, and retains the sharp initial decrease in DII. Carbonyl elongation and proton transfer therefore remain informative close to the hopping event. Together with the corresponding analysis for furan, this comparison supports the stability of the physical interpre-

tation across the sampled time windows, while showing that the absolute DII values depend on the configurations included. This need not hold in general and should be examined case by case.

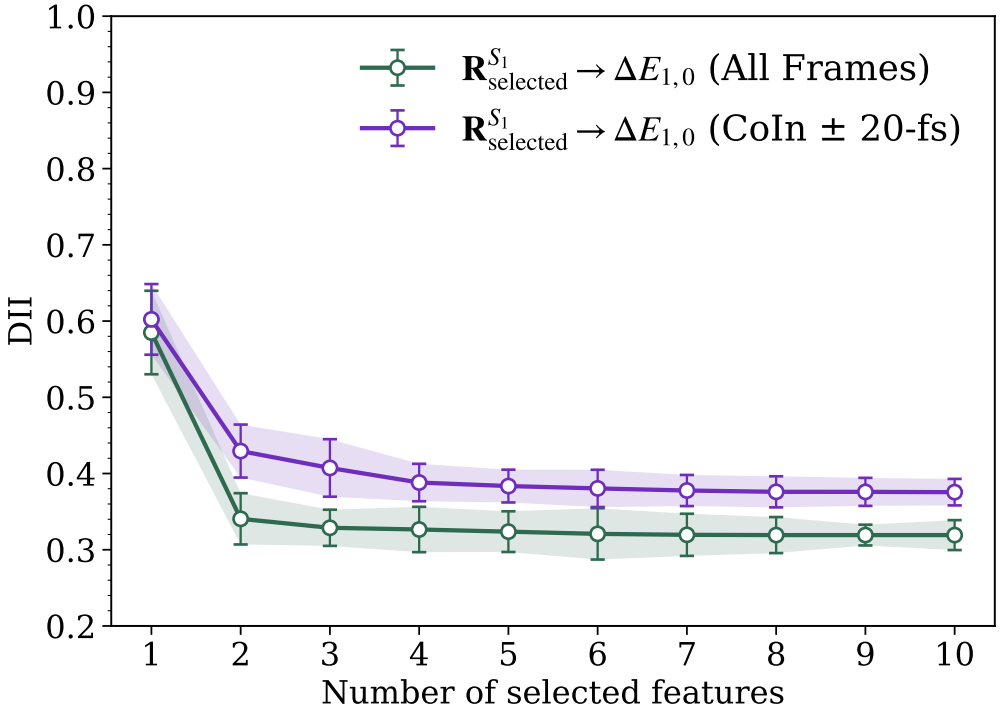


Figure B.16: ***L-glutamine:*** *Sensitivity of the DII analysis to the time window around the hopping event. The DII between the selected S_1 internal coordinates ($\mathbf{R}^{S_1}_{\text{selected}}$) and the S_1–S_0 energy gap ($\Delta E_{1,0}$) is shown as a function of the number of selected features, using all available frames (green) or configurations within ±20 fs of the hopping event (purple). Shaded areas show error bars computed from 50 independent bootstrap resamples. Both analyses recover the carbonyl $C-O$ stretch and the proton-transfer-related $O-H$ distance as the leading coordinates and exhibit a sharp initial decrease in DII followed by smaller improvements.*

B.2.3 Sensitivity to the hyperparameters of the method

We also examined the sensitivity of the furan results to the JSD hotspot threshold and to the softmax length scale λ in the DII. For JSD cutoffs of 0.09, 0.10, 0.12, 0.15, and 0.20, the DII curves are closely aligned and plateau at approximately 0.44–0.45 (Figure B.17). All tested cutoffs recover the same five leading coordinates, corresponding to the coupled puckering and ring-opening mechanism.

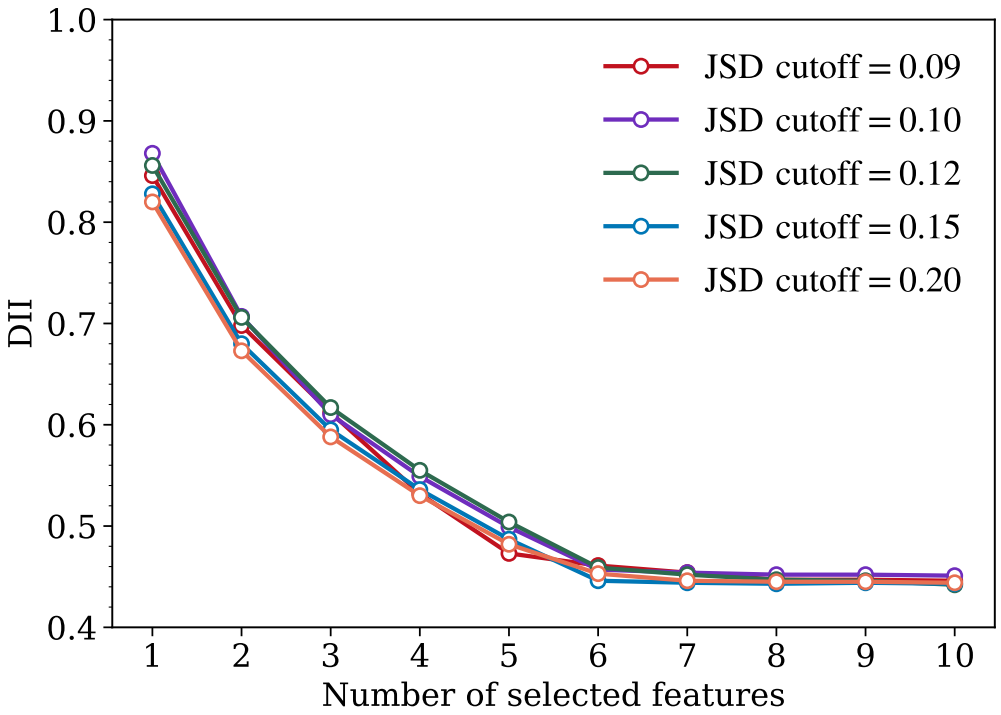


Figure B.17: ***Furan:*** *Sensitivity of the DII feature-selection analysis to the JSD hotspot threshold. The DII between the selected S_1 internal coordinates and the S_1–S_0 energy gap ($\Delta E_{1,0}$) is shown as a function of the number of selected features for JSD cutoffs of 0.09, 0.10, 0.12, 0.15, and 0.20. All curves exhibit similar trends and plateau at approximately 0.44–0.45. Every tested cutoff recovers the same five leading coordinates: the C−O−C−H dihedral, the two C−O stretches, the C−O−C angle, and the C−C separation, demonstrating stable identification of the dominant puckering and ring-opening signatures over the threshold range examined.*

Repeating the analysis with fixed λ values of 0.02, 0.1, 0.2, and 0.5 yields ten-feature DII values of approximately 0.40, 0.45, 0.46, and 0.70, respectively (Figure B.18). Although the numerical DII depends on the smoothing scale, the same five leading coordinates persist at every setting, with changes in their ordering. Beyond these five features, the selected coordinates differ, most notably at $\lambda = 0.5$. Thus, the dominant puckering and ring-opening signatures are stable across the tested settings, while the absolute DII values and secondary features show greater sensitivity.

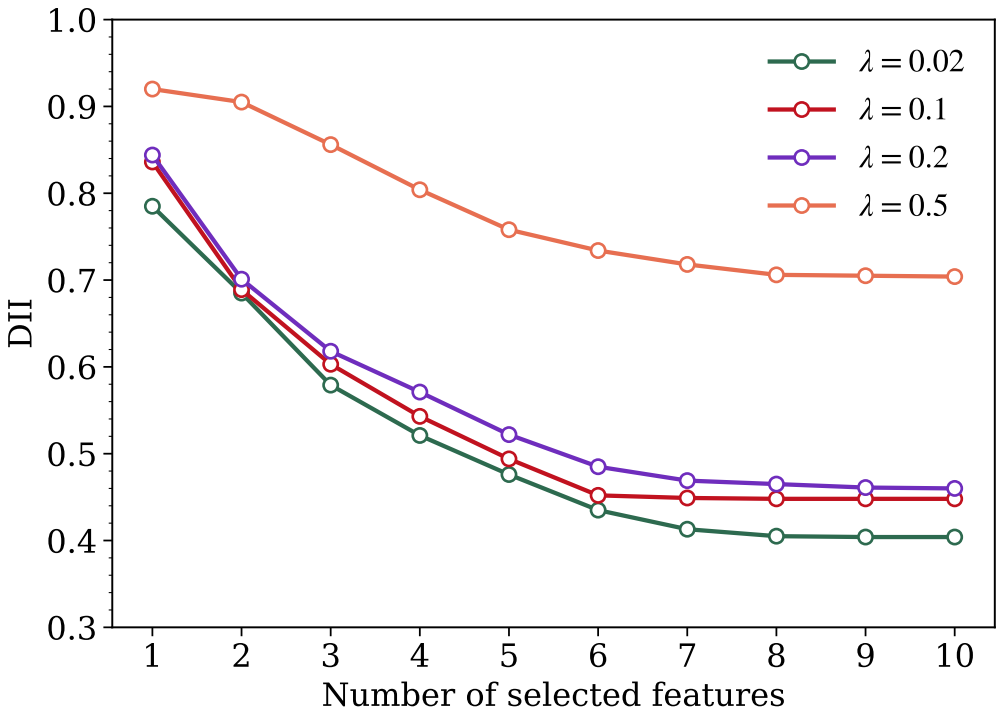


Figure B.18: ***Furan:*** *Sensitivity of the DII feature-selection analysis to the softmax length scale λ. The DII between the selected S_1 internal coordinates ($\mathbf{R}^{S_1}_{\text{selected}}$) and the S_1–S_0 energy gap ($\Delta E_{1,0}$) is shown as a function of the number of selected features for fixed λ values of 0.02, 0.1, 0.2, and 0.5. The ten-feature DII values are approximately 0.40, 0.45, 0.46, and 0.70, respectively. The same five leading coordinates are recovered at every setting, although their ordering changes: the C−O−C−H dihedral, the two C−O stretches, the C−O−C angle, and the C−C separation. Beyond the first five features, the selected coordinates differ at $\lambda = 0.5$. The leading physical signatures therefore persist across the tested settings, while the absolute DII values and secondary feature selections depend more strongly on the smoothing scale.*

C

Supplementary Information: Towards a robust approach to infer causality from molecular dynamics simulations

C.1 Transfer Entropy

Bivariate case

In this section we compare alternative definitions of the bivariate Transfer Entropy. The standard definition provided by Schreiber [63] reads

$$\mathrm{TE}_{X\to Y} = I\left(X_t^-; Y_{t+1} \mid Y_t^-\right) \tag{C.1a}$$

$$= H(Y_{t+1} \mid Y_t^-) - H(Y_{t+1} \mid X_t^-, Y_t^-)\,, \tag{C.1b}$$

where $I(\cdot,\cdot|\cdot)$ denotes a *conditional mutual information*, equivalent to the difference of entropies in Eq. (C.1b). Here, X_t^- and Y_t^- are time windows built over the past of X and Y:

$$X_t^- = (X_t, X_{t-1}, ..., X_{t-k+1})\,, \tag{C.2a}$$

$$Y_t^- = (Y_t, Y_{t-1}, ..., Y_{t-l+1})\,. \tag{C.2b}$$

Equations (C.2) assume that the dynamics of X and Y can be approximately described as Markov processes of order k and l, respectively. Notice that this definition inherently assumes $\tau = 1$.

Palu š et al. [418] provided a formulation with an explicit dependence on the time

lag between the present and the future:

$$\mathrm{TE}_{X\to Y}(\tau) = I\left(X_t^-; Y_{t+\tau} \mid Y_t^-\right) \tag{C.3a}$$
$$= H(Y_{t+\tau} \mid Y_t^-) - H(Y_{t+1} \mid X_t^-, Y_t^-)\,. \tag{C.3b}$$

As was observed in related works [15, 303], the choice of $\tau > 1$ can be beneficial for analyzing time series generated by continuous dynamics, since conditional dependencies may manifest more clearly at larger time scales compared to those of the true interactions. The definition provided in Eq. 2.25, which we used throughout this work, is obtained from Eq. (C.3) by replacing X_t^- and Y_t^- with the single frames X_t and Y_t, respectively.

Estimator for discrete RVs

By exploiting the relationship that links joint and conditional entropies, $H(X \mid Y) = H(X, Y) - H(Y)$, the TE defined in Eq. 2.25 can be developed as the following sum of Shannon entropies:

$$\mathrm{TE}_{X\to Y}(\tau) := H(Y_\tau \mid Y_0) - H(Y_\tau \mid X_0,\, Y_0)\,, \tag{C.4a}$$
$$= -H(Y_\tau, X_0, Y_0) + H(X_0, Y_0) + H(Y_\tau, Y_0) - H(Y_0)\,. \tag{C.4b}$$

In the discrete case, each term in Eq. (C.4b) takes the form

$$H(\mathbf{Z}) = -\sum_{\mathbf{z}} P(\mathbf{z}) \log P(\mathbf{z})\,, \tag{C.5}$$

where $\mathbf{Z}$ is a multi-dimensional RV (only one-dimensional when $\mathbf{Z} = Y_0$), $P(\mathbf{z})$ denotes the joint probability of its components (which is a trivial one-dimensional probability mass function when $\mathbf{Z}$ is one-dimensional), and the sum runs over all possible realizations of $\mathbf{Z}$.

The maximum likelihood or "plug-in" estimator for the joint entropy of Eq. (C.5) is simply obtained by replacing $P(\mathbf{z})$ with its empirical frequency $n_\mathbf{z}/N$, where $n_\mathbf{z}$ is the number of occurrences of the joint event $\mathbf{z}$, and N is the total number of observations of the random variable $\mathbf{Z}$ [129]:

$$H(\mathbf{Z}) \simeq -\sum_{\mathbf{z}} \frac{n_\mathbf{z}}{N} \log \frac{n_\mathbf{z}}{N}\,. \tag{C.6}$$

In turn, the TE can be computed by estimating all terms in Eq. (C.4b) in this manner. We employed this simple estimator, as implemented in the method `information_mutual_conditional` of the Python library `pyitlib` (version 0.2.3), to compute the TE for the Markov system of Fig. 5.4.

Estimator for continuous RVs

The decomposition of the TE in Eq. (C.4b) still holds in the continuous case, by replacing the Shannon entropy of Eq. (C.5) with the differential entropy

$$H(\mathbf{Z}) = -\int d\mathbf{z}\, p(\mathbf{z}) \log p(\mathbf{z})\,, \tag{C.7}$$

where $p(\mathbf{z})$ is the joint probability density of the components of $\mathbf{Z}$ and the integration domain is the support of $p(\mathbf{z})$. In principle, differential entropies can be computed by using histograms, namely by partitioning the support of $p(\mathbf{z})$ into bins of finite size and approximating the integral of Eq. (C.7) with a sum which has the same form of Eq. (C.6). Other density-based approaches exploit kernel density estimators [419, 420]. In this work, we considered a density-free estimator based on k-nearest neighbor statistics [133], which we briefly review in the following. We refer to Refs. [130, 421, 133] for a comprehensive overview of nearest-neighbor estimators for information-theoretic quantities.

According to the approach first proposed in Ref [422], an estimator for the differential entropy in Eq. (C.7), assuming $\mathbf{Z} \in \mathbb{R}^d$ and given the observations $\{\mathbf{z}_i\}_{i=1}^N$, is given by

$$H(\mathbf{Z}) \simeq \psi(N) - \psi(k) + \log c_d + \frac{d}{N}\sum_{i=1}^{N} \log \varepsilon(i)\,. \tag{C.8}$$

Here, $\psi(x)$ is the digamma function ($\psi(x) = \frac{d\log\Gamma(x)}{dx}$), c_d is the volume of the d-dimensional unit sphere, which depends on the specific L^p distance employed, and $\varepsilon(i)$ is twice the distance from $\mathbf{z}_i$ to its k-th neighbour, in the same distance space. This expression can be derived (see for example Ref. [130]) by assuming a uniform local density around each point z_i, up to the k-th nearest neighbour.

The conditional mutual information estimator employed in this work [133] builds on Eq. (C.8) to approximate each term in the decomposition of Eq. (C.4b). Specifically, this estimator makes use of an input parameter k to pre-compute a fixed radius $\epsilon(i)$ for each point i, instead of using a fixed number of nearest neighbors. Such a fixed radius is defined as the L^∞ distance from $\mathbf{z}_i$ and its k-the nearest neighbor in the space $\mathbf{Z} = (Y_\tau, X_0, Y_0)$. The use of a point-dependent number of neighbors and L^∞ norms allows for a perfect cancellation of the last two terms in Eq. (C.8) among the different entropy terms, bringing to

$$TE_{X\to Y}(\tau) \simeq \sum_{i=1}^{N} \big(\psi(k) - \psi(n_{X_0Y_0,i}) - \psi(n_{Y_\tau Y_0,i}) + \psi(n_{Y_0,i})\big) \tag{C.9}$$

in the absence of distances ties. Here, $n_{X_0Y_0,i}$, $n_{Y_\tau Y_0,i}$ and $n_{Y_0,i}$ are the number of neighbors of $\mathbf{z}_i$ within the radius $\epsilon(i)$, in the spaces of the subscript variables. The estimator is implemented in the Python library `knncmi`, which we used by setting $k = 7$ for all calculations, as in Ref. [133]. For computing the Transfer Entropy between dihedral

angles, the distances employed by such a nearest-neighbor estimator were computed by enforcing the correct periodicity in $[0, 2\pi)$, namely by replacing the standard L^∞ distance between points $\mathbf{z}_i$, $\mathbf{z}_j \in \mathbb{R}^d$,

$$\|\mathbf{z}_i - \mathbf{z}_j\|_\infty = \max_\alpha |z_i^\alpha - z_j^\alpha| \,, \tag{C.10}$$

with the periodic distance

$$\|\mathbf{z}_i - \mathbf{z}_j\|_{\infty,2\pi} = \max_\alpha \left(\min_{k\in\mathbb{Z}} |z_i^\alpha - z_j^\alpha - 2\pi k| \right) . \tag{C.11}$$

Connection to Structural Causal Models

We use the language of Structural Causal Models (SCMs) [423] to draw a connection between the conditional dependence relationships assessed by Transfer Entropy and the formulation of causal statements. A discrete-time SCM for two time-dependent variables X_t and Y_t can be written as:

$$X_t := f(pa(X_t), \eta_t^X) \,, \tag{C.12a}$$
$$Y_t := g(pa(Y_t), \eta_t^Y) \,, \tag{C.12b}$$

where f and g are generic and possibly nonlinear functions which describe the details of the interactions, $pa(X_t)$ ($pa(Y_t)$) is a set containing all direct causes of X_t (Y_t), called its "parents", while η_t^X and η_t^Y are noise terms which account for the effect of unobserved variables. In this model, causal sufficiency is satisfied by requiring that η_t^X and η_t^Y are independent random variables, as any dependence would imply the presence of a common (and unobserved) cause of X_t and Y_t. Under fairly general assumptions [270], the structure of Eqs. (C.12) is uniquely identified by a set of conditional independence relationships. For example, condition $X_{t-\tau} \notin pa(Y_{t+1})$ (namely, $X_{t-\tau}$ is *not* a direct cause of Y_{t+1}) corresponds to the conditional independence relationship $X_{t-\tau} \perp\!\!\!\perp Y_{t+1} \mid pa(Y_{t+1})$. In order to directly assess such a condition from time series data, one has to either infer the conditioning set $pa(Y_{t+1})$ using iterative approaches [423, 270], or use a larger set that includes $pa(Y_{t+1})$ without breaking the conditional independence above. Assuming that the SCM does not include instantaneous causal links, this can be achieved by using as conditioning set the entire past of the "target" variable, $X_t^- \cup Y_t^-$, of which $pa(Y_{t+1})$ is necessarily a subset. As shown in Eqs. (C.2), X_t^- and Y_t^- can be constructed in practice up to the maximum interaction lag that is believed to be present in the unknown dynamics. Following this strategy, the tested relationship $X_{t-\tau} \perp\!\!\!\perp Y_{t+1} \mid pa(Y_{t+1})$ can be replaced by

$$X_{t-\tau} \perp\!\!\!\perp Y_{t+1} \mid \left(X_t^- \cup Y_t^-\right) \setminus X_{t-\tau} \,, \tag{C.13}$$

where $\left(X_t^- \cup Y_t^-\right) \setminus X_{t-\tau}$ denotes the past of the series without $X_{t-\tau}$. If this condition is satisfied, then $X_{t-\tau} \notin pa(Y_{t+1})$; otherwise, one concludes that $X_{t-\tau}$ is a direct cause of

Y_{t+1}.

Transfer Entropy can be viewed as a method for testing conditional independencies of the form given in Eq. (C.13), by considering multiple time lags simultaneously. Specifically, Schreiber's definition (see Eq. (C.1)) tests the conditional independence $X_t^- \perp\!\!\!\perp Y_{t+1} \mid Y_t^-$, which is obtained from Eq. (C.13) by replacing $X_{t-\tau}$ with X_t^-. In this case, $\mathrm{TE}_{X\to Y} > 0$ allows deducing that a causal link from a component of X_t^- to Y_{t+1} exists, but not the specific interaction lag τ. For example, the conditional *dependence* $X_t^- \not\perp\!\!\!\perp Y_{t+1} \mid Y_t^-$ may originate from direct links $X_{t-\tau} \to Y_{t+1}$, for any $X_{t-\tau} \in X_t^-$. Similarly, the conditional independence tested by the Transfer Entropy of Ref. [418] (see Eq. (C.3)) is $X_t^- \perp\!\!\!\perp Y_{t+\tau} \mid Y_t^-$. Also in this case, condition $\mathrm{TE}_{X\to Y} > 0$ allows inferring the existence of a lag-unspecific causal link from X to Y. Finally, the Transfer Entropy employed in this work (see Eq. 2.25) tests the conditional independence $X_t \perp\!\!\!\perp Y_{t+\tau} \mid Y_t$.

Importantly, instantaneous links such as that from X to Y in the three-state Markov system (Fig. 5.4**a**) can still be detected by measuring Transfer Entropy, although not identified as instantaneous. In that system, for example, the direct and instantaneous link $X_t \to Y_t$ generates a conditional dependence $X_t \not\perp\!\!\!\perp Y_{t+1}|Y_t$ through the indirect causal link $X_t \to X_{t+1} \to Y_{t+1}$, which results in $\mathrm{TE}_{X\to Y}(\tau = 1) > 0$ (see Fig. 5.4**b** and Section C.2).

Multivariate Transfer Entropy

The multivariate generalization of the Transfer Entropy used in this work is:

$$\mathrm{TE}_{X\to Y|Z}(\tau) := I(X_t; Y_{t+\tau} \mid Y_t, Z_t)\,, \tag{C.14}$$

where Z_t is a (possibly multidimensional) variable containing putative common drivers of X and Y. Eq. (C.14) tests the conditional independence relationship $X_t \perp\!\!\!\perp Y_{t+\tau} \mid \{Y_t, Z_t\}$, which allows stating that X_t is not a direct or indirect cause of $Y_{t+\tau}$. Similarly to the bivariate case, such a conclusion can be drawn by assuming that there are no unobserved common drivers of X and Y (namely, variables causing both X and Y which are not included in Z). The fact that this measure does not allow distinguishing between direct and indirect causal links, namely mediated by Z, directly comes from the fact that it does not infer the specific interaction lag, as in the bivariate case. As was discussed in Ref. [303], the use of a single frame at time t in place of a time window does not affect the interpretation of the measure up to the assumption that the maximum time lag of direct links is 1, which is satisfied by all time series employed in this work.

In Fig. C.1 we show a comparison between the bivariate $\mathrm{TE}_{X\to Y}$ (panel **a**) and the multivariate $\mathrm{TE}_{X\to Y|Z}$ (panel **a**), for the three-dimensional Langevin system shown in Fig. 5.6**c**. The bivariate TE erroneously suggests the presence of a link from X to Y in the large time-scale regime, while its multivariate version allows deducing that Z is a common driver of X and Y. This conclusion can be drawn by observing that $\mathrm{TE}_{Z\to X|Y} > 0$ and $\mathrm{TE}_{Z\to Y|X} > 0$ (Fig. C.2). The multivariate Transfer Entropies in Figs. C.1 and C.2

were computed with the same nearest-neighbor estimator employed for the bivariate case (see Eq. (C.9)), which can be straightforwardly generalized to

$$TE_{X\to Y}(\tau) \simeq \sum_{i=1}^{N} \left(\psi(k) - \psi(n_{X_0Y_0Z_0,i}) - \psi(n_{Y_\tau Y_0 Z_0,i}) + \psi(n_{Y_0Z_0,i})\right). \tag{C.15}$$

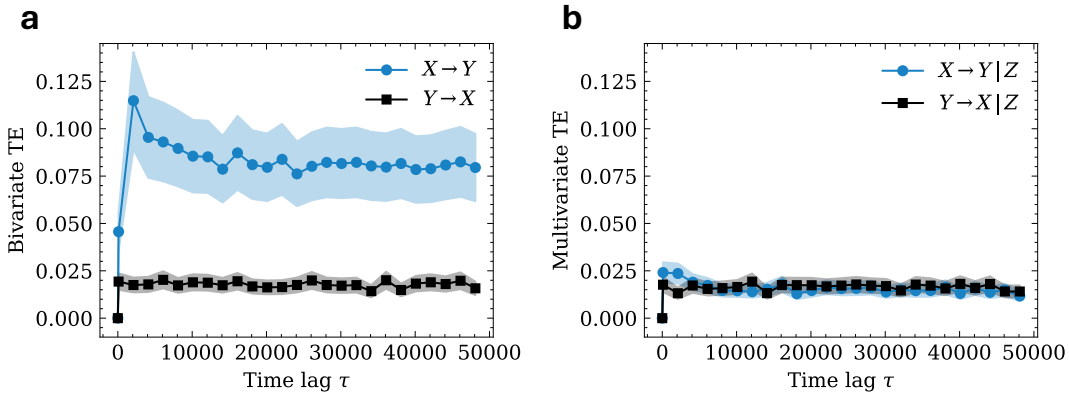


Figure C.1: ***a**: Bivariate TE from X to Y (blue) and from Y to X (black) as a function of the lag τ, for the Langevin systems of Fig. 5.6**c**. **b**: Multivariate $TE_{X\to Y|Z}$ (blue) and $TE_{Y\to X|Z}$ (black) for the same Langevin system.*

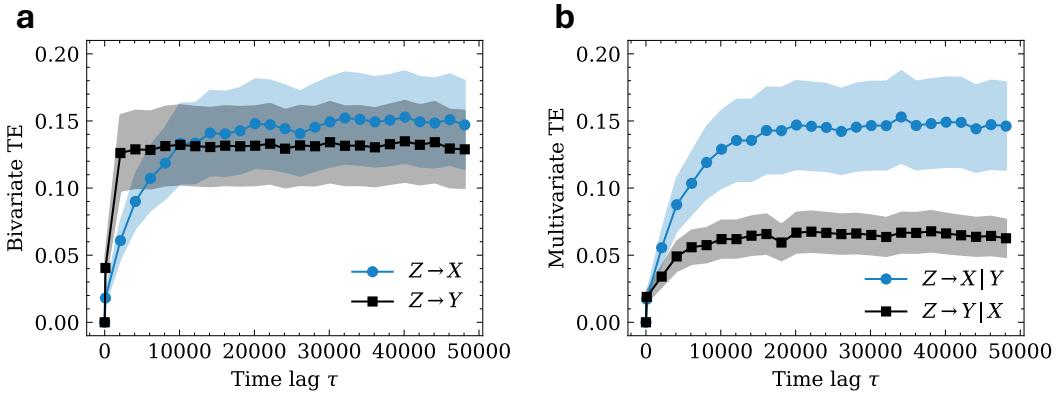


Figure C.2: ***a**: Bivariate TE from Z to X (blue) and from Z to Y (black) as a function of the lag τ, for the Langevin systems of Fig. 5.6**c**. **b**: Multivariate $TE_{Z\to X|Y}$ (blue) and $TE_{Z\to Y|X}$ (black) for the same Langevin system.*

C.2 Asymmetric transfer entropy in the Markov system

We show in this Section that $\mathrm{TE}_{X\to Y}(\tau) > 0$ and $\mathrm{TE}_{Y\to X}(\tau) = 0$ for the Markov system depicted in Fig. 5.4, focusing on the case $\tau = 1$.

Denoting with s_t the state of the system at time t ($s_t \in \{A, B, C\}$), it is straightfor-

ward to show that the detailed balance condition

$$p(s_t)\, p(s_{t+1}|s_t) = p(s_{t+1})\, p(s_t|s_{t+1}) \tag{C.16}$$

is satisfied by the equilibrium probability distribution $p(A) = p(B) = p(C) = 1/3$. The Markov system can be equivalently represented using the variables X and Y alone as in Fig. C.3.

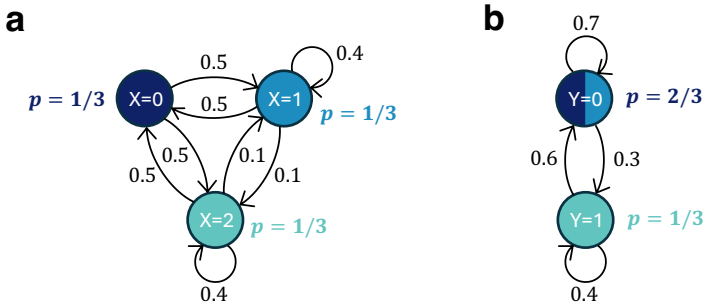


Figure C.3: *Marginal Markov diagrams for the variables X (panel **a**) and Y (panel **b**), derived from the Markov diagram of Fig. 5.4**a**. Transition probabilities are reported in black next to the arrows; probabilities of the equilibrium distribution are written next to the nodes with the same color of the corresponding state. While the diagram for X is equivalent to the original Markov system, in the diagram for Y two states are combined into a single node.*

Starting from direction $X \to Y$, we recall that the definition of TE employed in this work (see Eq. 2.25), written for $\tau = 1$, reads:

$$\mathrm{TE}_{X\to Y}(\tau = 1) = H(Y_{t+1} \mid Y_t) - H(Y_{t+1} \mid X_t, Y_t)\,, \tag{C.17}$$

where

$$H(Y_{t+1} \mid Y_t) = -\sum_{y_t, y_{t+1}} p(y_{t+1} \mid y_t)\, p(y_t) \log p(y_{t+1} \mid y_t)\,, \tag{C.18a}$$

$$H(Y_{t+1} \mid X_t, Y_t) = -\sum_{x_t, y_t, y_{t+1}} p(y_{t+1} \mid x_t, y_t)\, p(x_t, y_t) \log p(y_{t+1} \mid x_t, y_t)\,. \tag{C.18b}$$

Here and in the following, we use capital letters (e.g., X_t and Y_t) to denote RVs, and lowercase letters (e.g., x_t and y_t) to denote their possible realizations. By substituting the probabilities appearing in Fig. C.3 in these equations one finds:

$$H(Y_{t+1} \mid Y_t) = -\frac{2}{3}\Big(0.3\log(0.3) + 0.7\log(0.7)\Big) - \frac{1}{3}\Big(0.4\log(0.4) + 0.6\log(0.6)\Big)\,, \tag{C.19a}$$

$$H(Y_{t+1} \mid X_t, Y_t) = -\frac{1}{3}\Big(2\times 0.5\log(0.5) + 0.1\log(0.1) + 0.9\log(0.9) + 0.4\log(0.4) + 0.6\log(0.6)\Big)\,. \tag{C.19b}$$

Since $H(Y_{t+1} \mid Y_t) > H(Y_{t+1} \mid X_t, Y_t)$, the Transfer Entropy in direction $X \to Y$ is positive ($\mathrm{TE}_{X\to Y}(\tau = 1) \simeq 0.07$).

Similarly, the Transfer Entropy in direction $Y \to X$ can be written as:

$$\mathrm{TE}_{Y\to X}(\tau = 1) = H(X_{t+1} \mid X_t) - H(X_{t+1} \mid X_t, Y_t)\,, \tag{C.20}$$

where

$$H(X_{t+1} \mid X_t) = -\sum_{x_t, x_{t+1}} p(x_{t+1} \mid x_t)\, p(x_t) \log p(x_{t+1} \mid x_t) \tag{C.21a}$$

$$H(X_{t+1} \mid X_t, Y_t) = -\sum_{x_t, y_t, x_{t+1}} p(x_{t+1} \mid x_t, y_t)\, p(x_t, y_t) \log p(x_{t+1} \mid x_t, y_t)\,. \tag{C.21b}$$

In this case, all non-zero probabilities appearing in Eq. (C.21b) actually depend on $X_{t+1} = x_{t+1}$ and $X_t = x_t$ only, as the state of the system at time t is fully determined once X_t is known. This implies $H(X_{t+1} \mid X_t) = H(X_{t+1} \mid X_t, Y_t)$ and, as a consequence, $\mathrm{TE}_{Y\to X}(\tau = 1) = 0$. We note that if all transition probabilities were set to 0.5, deleting the self-loops of nodes B and C in the original Markov system, an incidental independence on X_t would appear in the non-zero probabilities of Eq. (C.18b) also in direction $X \to Y$, resulting in $\mathrm{TE}_{X\to Y}(\tau = 1) = 0$.

C.3 Unobserved states in the Markov system

In this section, we consider a Markov system similar to that introduced in Fig. 5.4**a**, but with an additional state D and a third variable Z which is more informative than X (Fig. C.4, panel **a**). Here, Z can distinguish all the states of the system (panel **b**), X has a degeneracy in states C and D (panel **c**), and Y is degenerate both in C, D and in A, B (panel **d**). In the following we show that the bivariate Transfer Entropy satisfies $\mathrm{TE}_{X\to Y}(\tau = 1) > 0$, while the multivariate transfer entropy (see Section C.1) including Z in the conditioning set leads to $\mathrm{TE}_{X\to Y|Z}(\tau = 1) = 0$.

The equilibrium probability distribution satisfying detailed balance for this Markov-state system is $p(A) = p(B) = 1/3$, $p(C) = 1/4$ and $p(D) = 1/12$. We note that this system is observationally equivalent to that of Fig. 5.4**a** when Z is unobserved, since the two systems share the same transition probabilities between the states distinguished by X and Y.

Since the system is observationally equivalent to the three-state system of Fig. 5.4**a**, the bivariate Transfer Entropy in direction $X \to Y$ is the same computed in Section C.2 ($\mathrm{TE}_{X\to Y}(\tau = 1) \simeq 0.07$).

As in the bivariate case, the multivariate Transfer Entropy from X to Y given Z can be decomposed as the difference of two Shannon entropies,

$$\mathrm{TE}_{X\to Y|Z} = H(Y_{t+1} \mid Y_t, Z_t) - H(Y_{t+1} \mid X_t, Y_t, Z_t)\,, \tag{C.22}$$

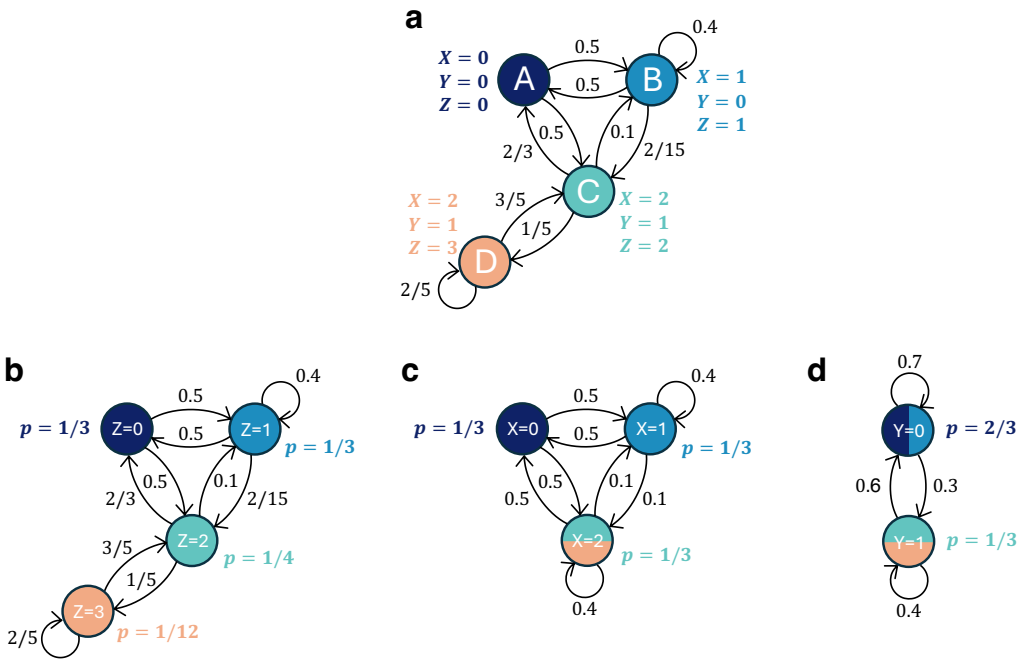


Figure C.4: ***a****: Markov system where each state is defined by the triplet* (X, Y, Z)*. Transition probabilities are reported next to the corresponding arrows.* ***b****: Marginal Markov diagram for* Z*.* ***c****: Marginal Markov diagram for* X*.* ***d****: Marginal Markov diagram for* Y*. In panels* ***b****,* ***c*** *and* ***d****, the equilibrium probabilities are reported next to the corresponding states.*

where

$$H(Y_{t+1} \mid Y_t, Z_t) = -\sum_{y_t, z_t, y_{t+1}} p(y_{t+1} \mid y_t, z_t)\, p(y_t \mid z_t)\, p(z_t) \log p(y_{t+1} \mid y_t)\,, \tag{C.23a}$$

$$H(Y_{t+1} \mid X_t, Y_t, Z_t) = -\sum_{x_t, y_t, z_t, y_{t+1}} p(y_{t+1} \mid x_t, y_t, z_t)\, p(y_t \mid x_t, z_t) \times p(x_t, z_t) \log p(y_{t+1} \mid x_t, y_t, z_t)\,. \tag{C.23b}$$

In all the non-zero probabilities appearing in Eq. (C.23b), the dependence on $X_t = x_t$ can be removed, as X_t is a deterministic function of Z_t. As a consequence, $H(Y_{t+1} \mid Y_t, Z_t) = H(Y_{t+1} \mid X_t, Y_t, Z_t)$ and $\mathrm{TE}_{X \to Y|Z}(\tau = 1) = 0$.

C.4 Supplementary figures

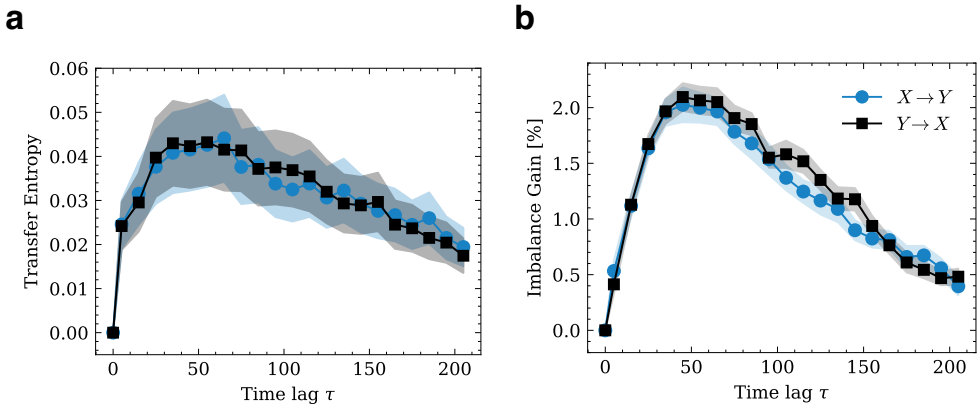


Figure C.5: *Transfer Entropy (panel **a**) and Imbalance Gain (panel **b**) between two variables X and Y described by the same free energy of Fig. 5.4**g**, following a Langevin dynamics with identical friction coefficients* ($\gamma_X = \gamma_Y = 1$).

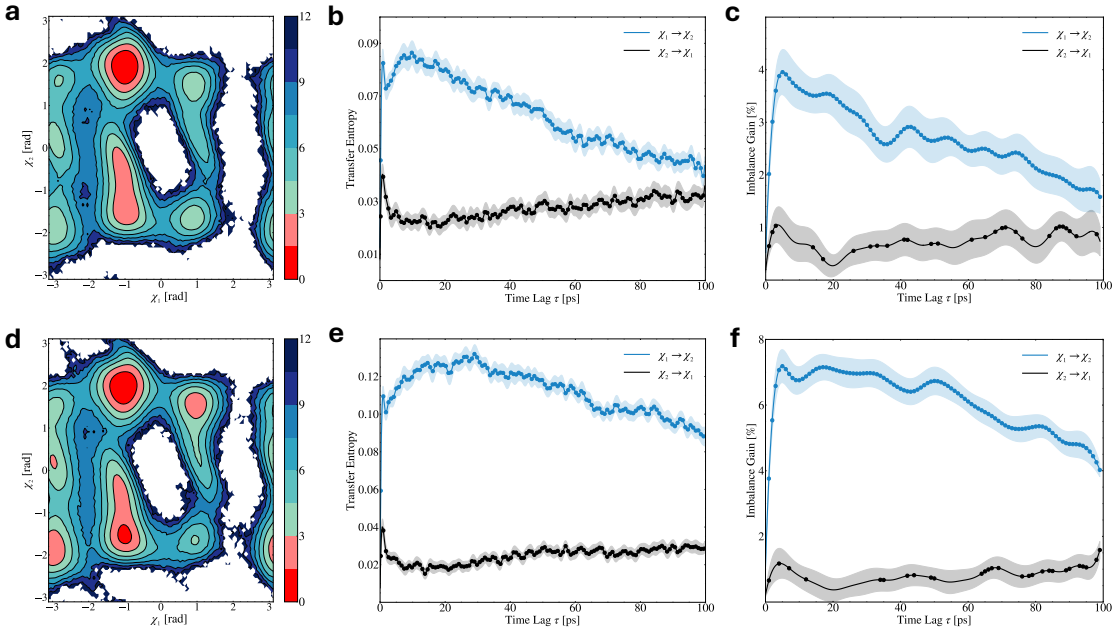


Figure C.6: ***a** and **d**: FES between χ_1 and χ_2 in the GS and ES, respectively. **b** and **c**: TE and IG curves for the CVs χ_1 and χ_2 in the GS. **e** and **f**: TE and IG curves for the same CVs in the ES. Shaded regions denote error bars over 14 independent estimates, and bold points denote values that are significantly different from 0 according to a t-test* ($p < 0.001$).

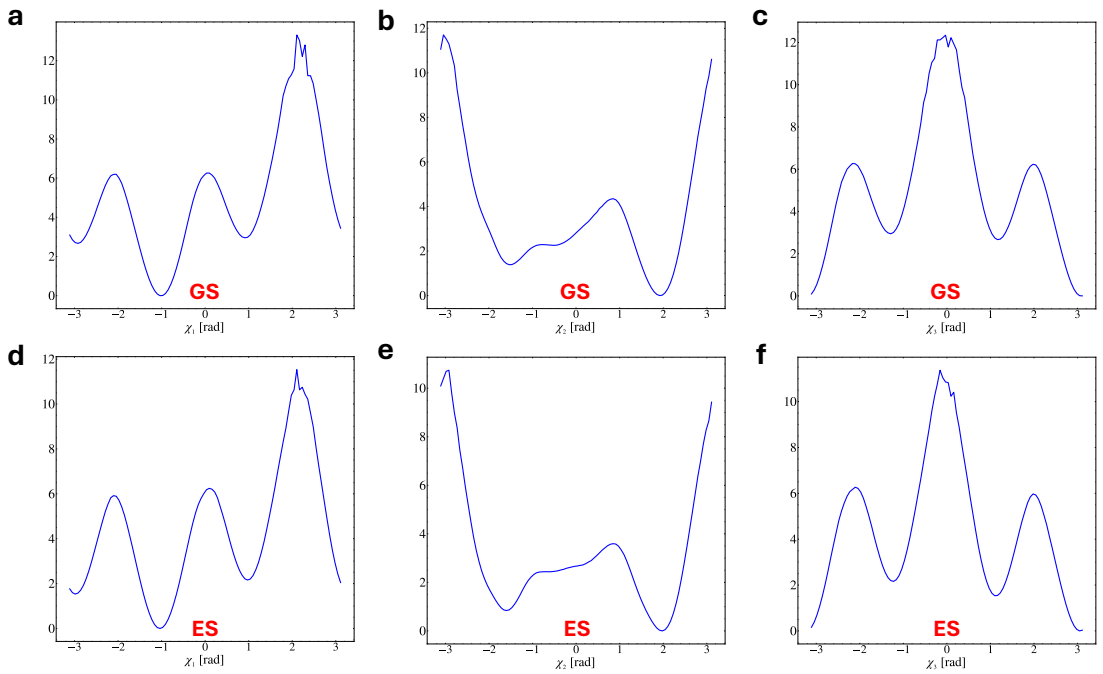


Figure C.7: ***a***,***b***,***c****: one-dimensional free energies of the three dihedrals under investigation in this paper - χ_1, χ_2, χ_3, in the GS.* ***d***,***e***,***f****: one-dimensional free energies of χ_1, χ_2, χ_3, in the ES.*

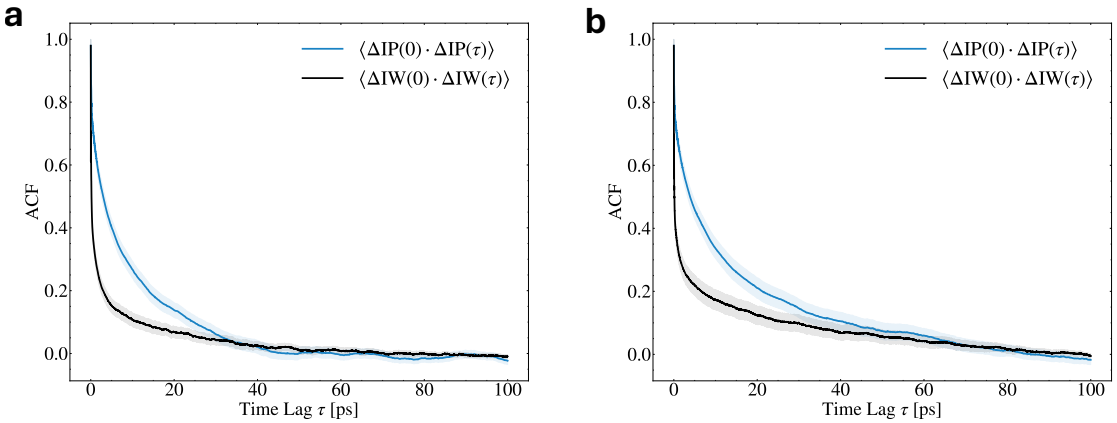


Figure C.8: ***a*** *and* ***b****: The autocorrelation functions (ACF) of Δ IP (blue) and Δ IW (black) in the GS and ES, respectively. Each curve was fit to a single exponential function of the form $\exp(-\frac{t}{\tau})$, where τ represents the decorrelation time of the corresponding variable. The decorrelation times for Δ IP and Δ IW in the GS (****a****) are 7.1 ps and 1.0 ps, respectively. In the ES (****b****), these values change to 10.7 ps for Δ IP and 1.6 ps for Δ IW.*

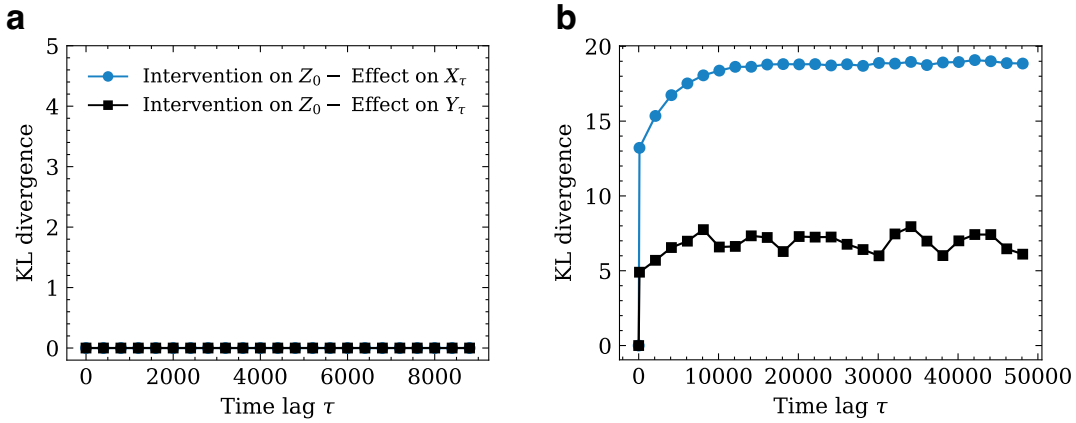


Figure C.9: ***a****: Effect of interventions on Z_0 for the three-dimensional Langevin systems shown in Fig. 5.6**a**. **b**: Effect of the same interventions for the three-dimensional Langevin system of Fig. 5.6**c**. The blue and black curves show the KL divergences measuring the effect of the interventions on X_τ and Y_τ, respectively. Specifically, the two curves display $D_{KL}\left[p(X_\tau \mid Z_0 = 0) \,\|\, p(X_\tau \mid Z_0 = 16)\right]$ and $D_{KL}\left[p(Y_\tau \mid Z_0 = 0) \,\|\, p(Y_\tau \mid Z_0 = 16)\right]$. The interventional values for Z_0 were set to the Z-coordinates of the furthest free energy minima in Fig. 5.6**c**.*

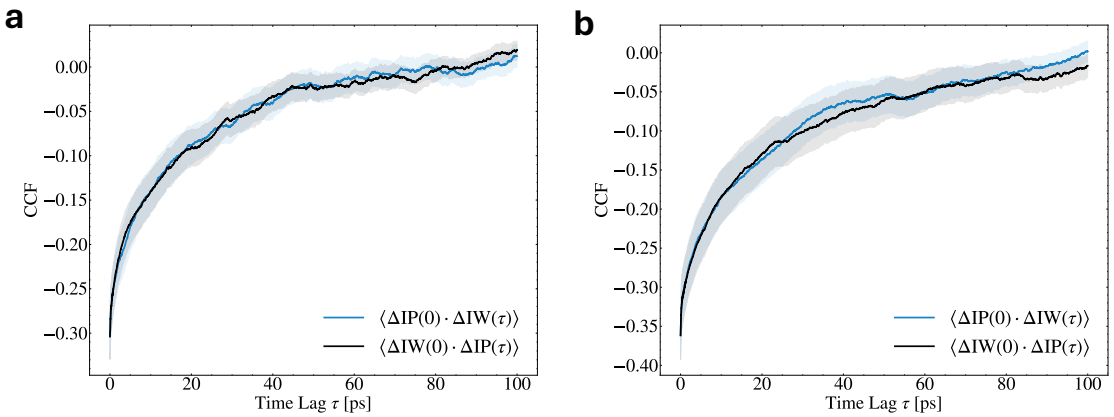


Figure C.10: ***a*** *and* ***b****: The cross-correlation functions between ΔIP and ΔIW in the GS and ES, respectively. The case where the time lagged variable is ΔIW is shown in blue, and where the time lagged variable is ΔIP is shown in black.*

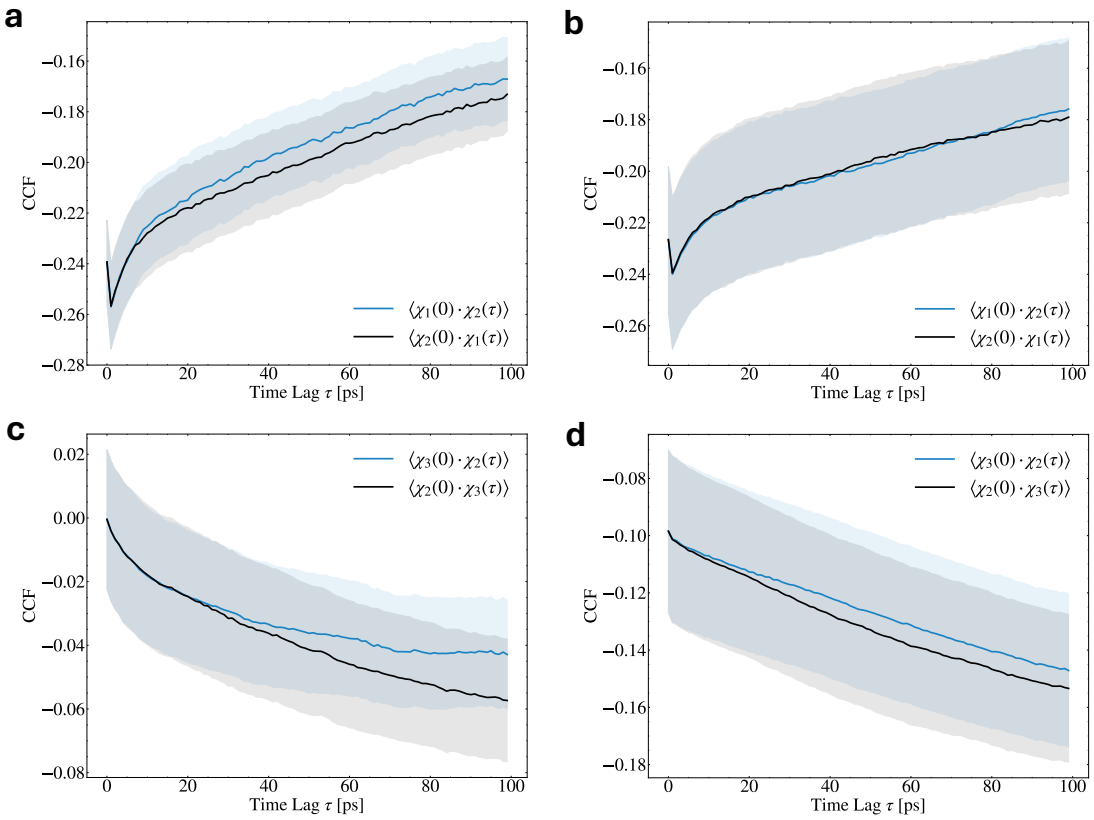


Figure C.11: ***a*** *and* ***b****: Cross-Correlation Functions (CCF) between* χ_1 *and* χ_2 *in the GS and ES, respectively.* ***c*** *and* ***d****: CCFs between* χ_3 *and* χ_2 *in the GS and ES, respectively. To account for the periodicity of such variables, CCFs of the form* $\langle \chi_i(0) \cdot \chi_j(\tau) \rangle$ *were computed as* $\frac{1}{2}\langle |\chi_i(0)|^2_{2\pi} + |\chi_j(\tau)|^2_{2\pi} - |\chi_i(0) - \chi_j(\tau)|^2_{2\pi} \rangle$*, where* $|\chi|_{2\pi} = \min_k |\chi - 2\pi k|$.

D

Supplementary Information: Investigating causality between principal components in protein dynamics

D.1 Brüschweiler collectivity index

To quantify the spatial extent of each principal component (PC), we computed the Brüschweiler collectivity index [359, 360]. In the present analysis, PCA was performed using only the C_α atoms. Therefore, each C_α atom was associated with a single residue, and the collectivity index can be interpreted as a residue-level measure of how broadly a given PC mode is distributed over the protein.

For a protein containing N C_α atoms, the eigenvector associated with PC_k is written as

$$\mathbf{q}_k = \left(q_{k,1x}, q_{k,1y}, q_{k,1z}, q_{k,2x}, q_{k,2y}, q_{k,2z}, \ldots, q_{k,Nx}, q_{k,Ny}, q_{k,Nz}\right),$$

where $(q_{k,ix}, q_{k,iy}, q_{k,iz})$ denotes the three-dimensional displacement component of residue i along mode k. The squared amplitude associated with residue i was computed as

$$a_{k,i} = q_{k,ix}^2 + q_{k,iy}^2 + q_{k,iz}^2.$$

The normalized residue-wise contribution to mode k was then defined as

$$p_{k,i} = \frac{a_{k,i}}{\sum_{j=1}^{N} a_{k,j}}, \qquad \sum_{i=1}^{N} p_{k,i} = 1.$$

Thus, $p_{k,i}$ represents the fraction of the total squared amplitude of PC_k carried by residue i.

The Brüschweiler collectivity index of mode k was calculated as

$$\kappa_k = \frac{1}{N} \exp\left[-\sum_{i=1}^{N} p_{k,i} \log p_{k,i} \right].$$

The index ranges approximately from $1/N$ to 1. A value close to $1/N$ indicates a highly localized mode, in which the displacement is concentrated on a single residue or a small number of residues. Conversely, a value close to 1 indicates a collective mode, in which the displacement amplitude is broadly and approximately uniformly distributed over the protein. In this work, larger values of κ_k were therefore interpreted as more delocalized, collective PC modes, whereas smaller values indicated more localized modes.

D.2 NTL9 results for another folded trajectory fragment

To assess the robustness of the causal relationships inferred for NTL9, we repeated the analysis on an independent folded-state trajectory fragment. Specifically, we considered the segment spanning 290–580 μs of the 1.112 ms trajectory, during which the protein remains in the folded state. This analysis provides an additional consistency check for the results reported in Chapter 6, which were obtained from a distinct folded segment of the NTL9 trajectory.

The autocorrelation functions of the first five principal components are shown in Fig. D.13. They display the same qualitative behavior observed for the trajectory fragment analyzed in the main text: PC_1 is the slowest relaxing mode, while the remaining leading PCs decay on shorter time scales. This indicates that the separation between the dominant slow mode and the higher-order PCs is not specific to a single folded-state segment.

We next computed the pairwise IG between PC_1 and $\mathrm{PC}_{2,3,4,5}$, shown in Fig. D.14. The resulting IG curves again reveal a pronounced directional asymmetry, with the dominant information transfer oriented from PC_1 toward the other leading PCs. This behavior is summarized by the causal graph in Fig. D.15, constructed from the normalized integrated information transfer, IGT. The graph confirms that PC_1 acts as the dominant source of directional information transfer in this folded trajectory fragment, consistent with the conclusions drawn from the analysis in the main text.

Finally, Fig. D.16 reports the corresponding pairwise IG estimates between PC_2 and $\mathrm{PC}_{3,4,5}$. In contrast to the strongly asymmetric links involving PC_1, these couplings are broadly reciprocal. Overall, the analysis of this independent folded-state fragment supports the robustness of the main conclusion: within the folded-state dynamics of NTL9, PC_1 behaves as a slow mode that directionally conditions several faster principal components, whereas the causal links among the lower-variance PCs are weaker and more symmetric.

Further characterization of the leading PCs using the Brüschweiler collectivity in-

dex, reported in Fig. D.17, shows that, as in the trajectory fragment discussed in the main text, the dominant modes remain relatively localized and are mainly concentrated around the β-turn region. This indicates that the spatial character of the leading PCs is conserved across the two folded segments. The main difference is that, in the present fragment, the first substantially delocalized mode that is strongly driven by PC_1 is PC_6, rather than PC_5 as observed in the analysis in the main text.

D.3 Supplementary Figures

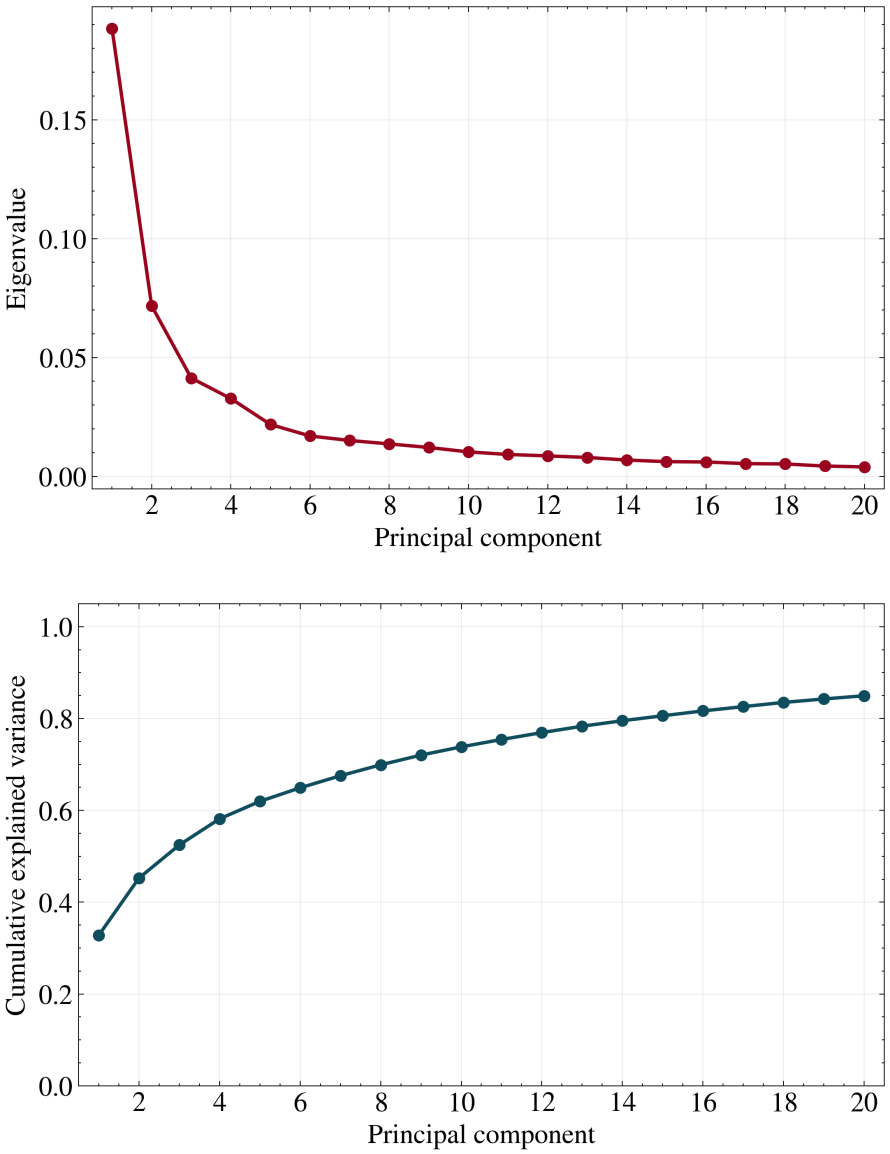


Figure D.1: *PCA eigenvalue spectrum and explained variance for the Ubiquitin trajectory.*

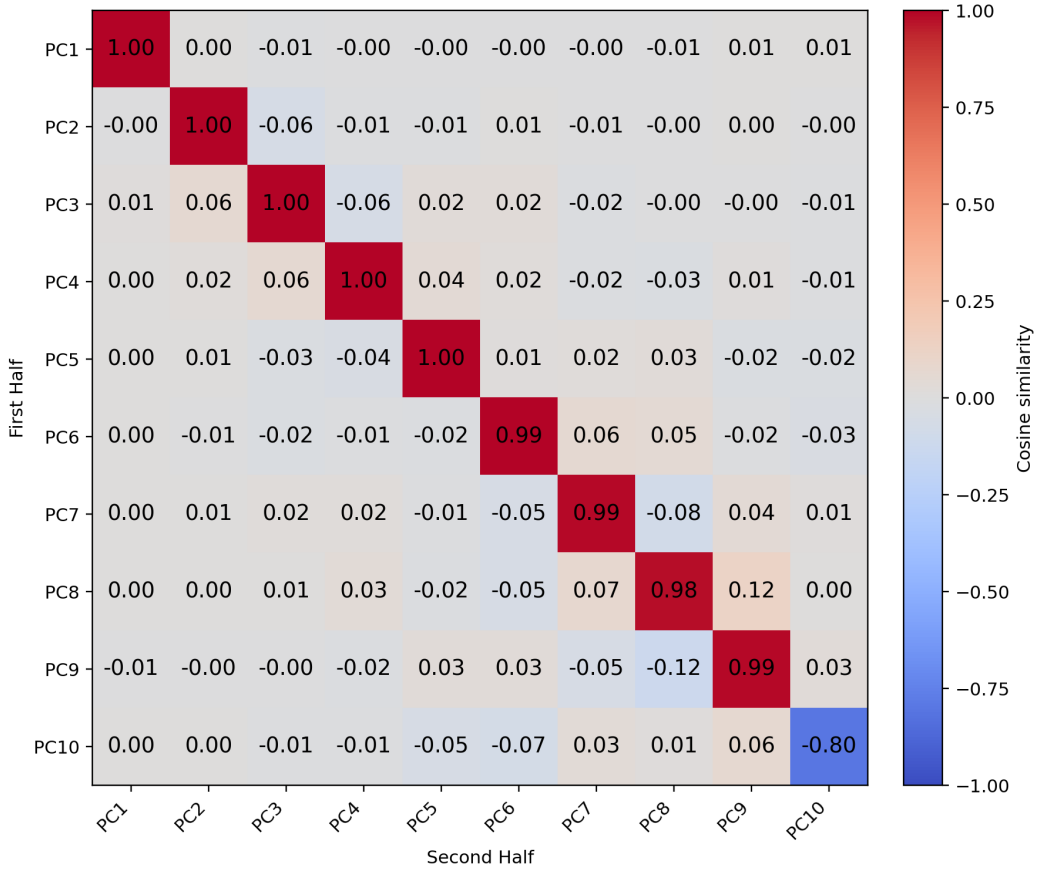


Figure D.2: *Ubiquitin: Cosine similarity between the first ten PCs computed independently from the two halves of the analyzed ubiquitin trajectory. The near-diagonal structure indicates that the dominant PCs are consistently recovered in both halves, confirming that the PCA is converged and that the leading modes retain the same character across the trajectory.*

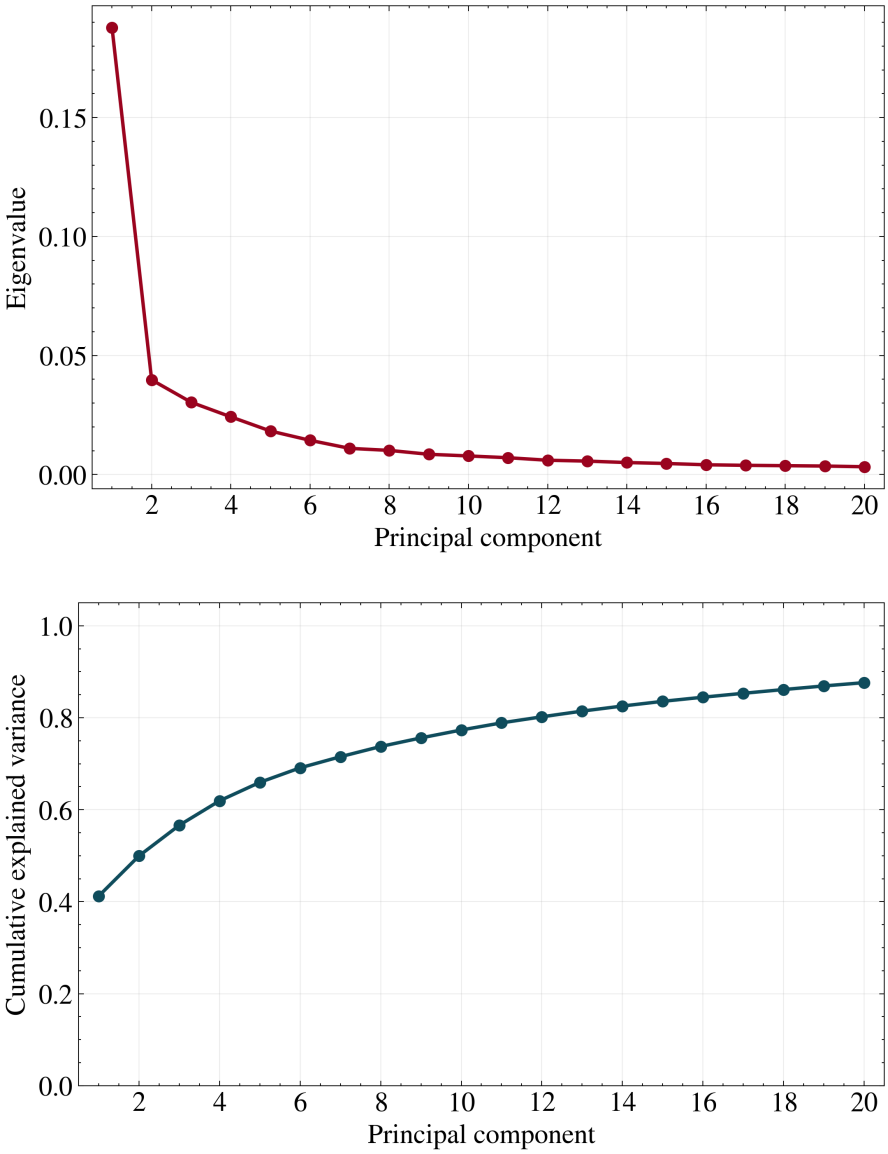


Figure D.3: *PCA eigenvalue spectrum and explained variance for the NTL9 trajectory.*

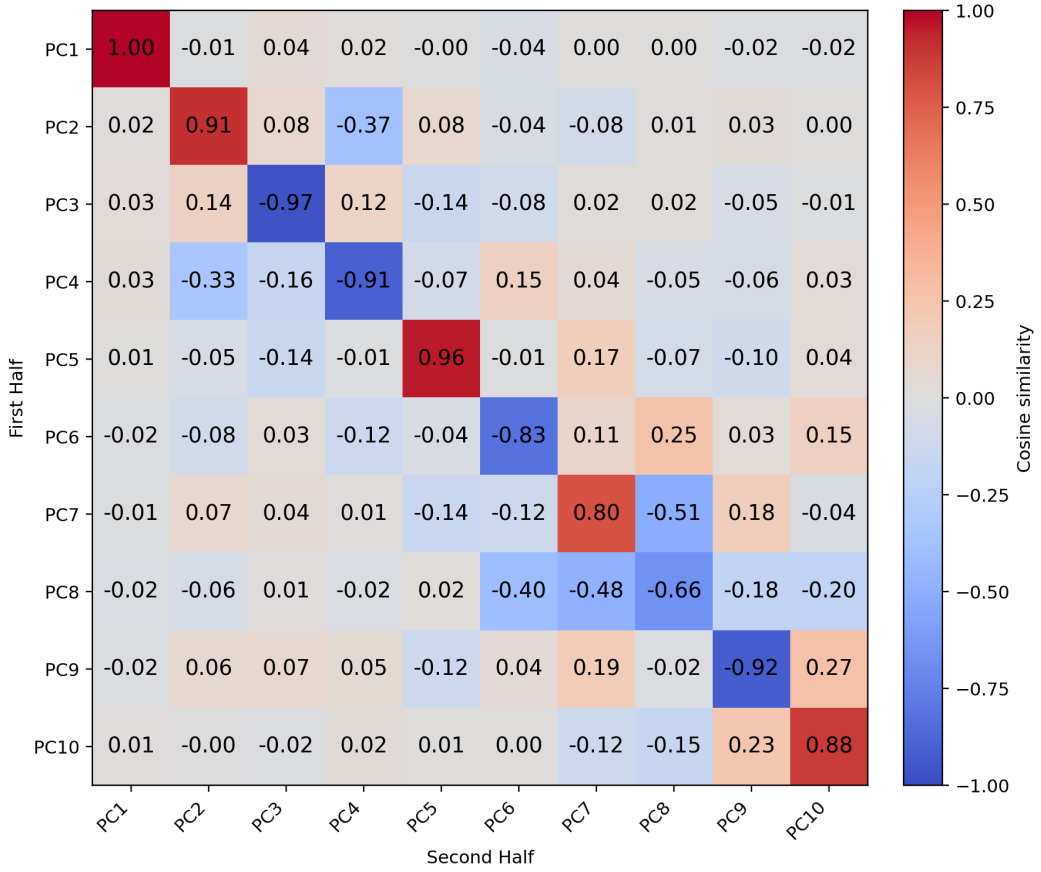


Figure D.4: *NTL9: Cosine similarity between the first ten PCs computed independently from the two halves of the analyzed NTL9 trajectory. The near-diagonal structure, especially for the top 5 PCs, indicates that the dominant PCs are consistently recovered in both halves, confirming that the PCA is converged and that the leading modes retain the same character across the trajectory.*

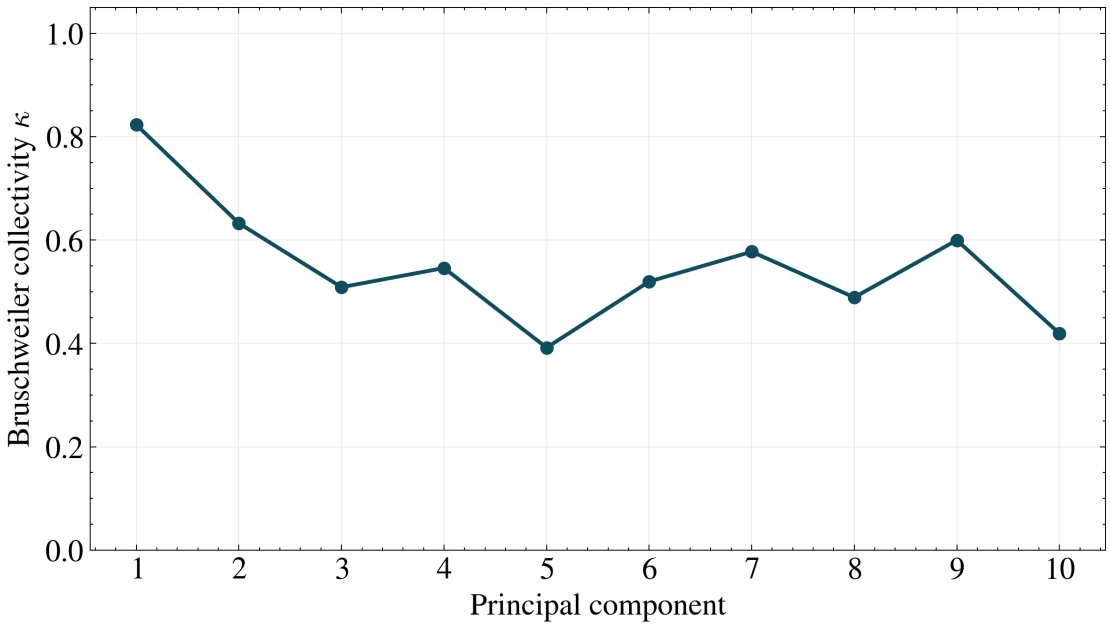


Figure D.5: *Ubiquitin: Brüschweiler collectivity index for the leading PCs. The collectivity index quantifies the spatial extent of each mode.*

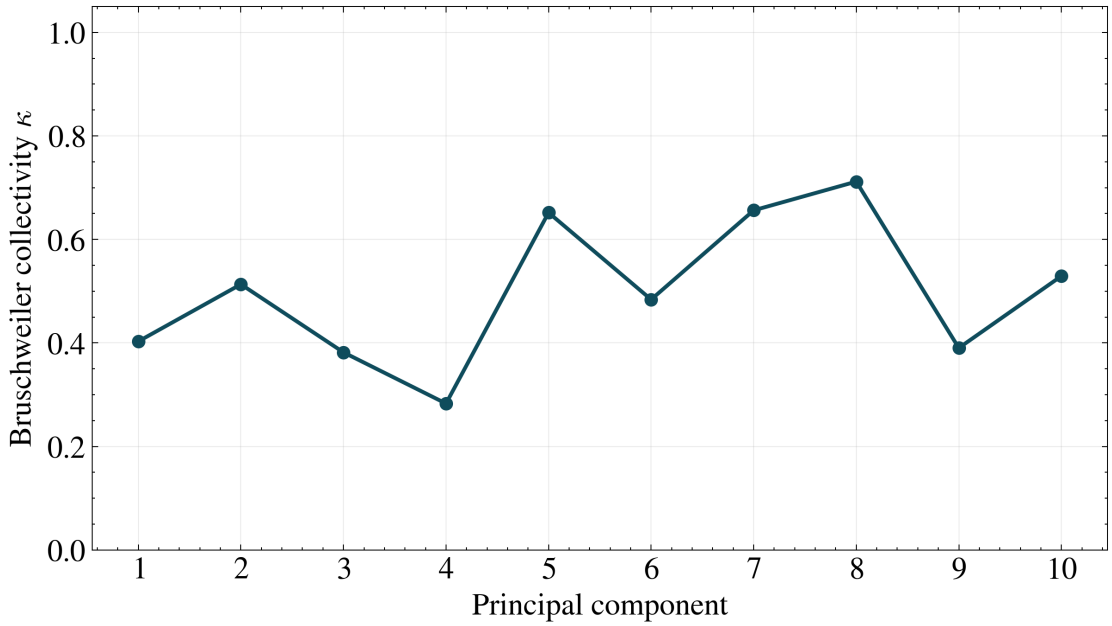


Figure D.6: *NTL9: Brüschweiler collectivity index for the leading PCs. The collectivity index quantifies the spatial extent of each mode.*

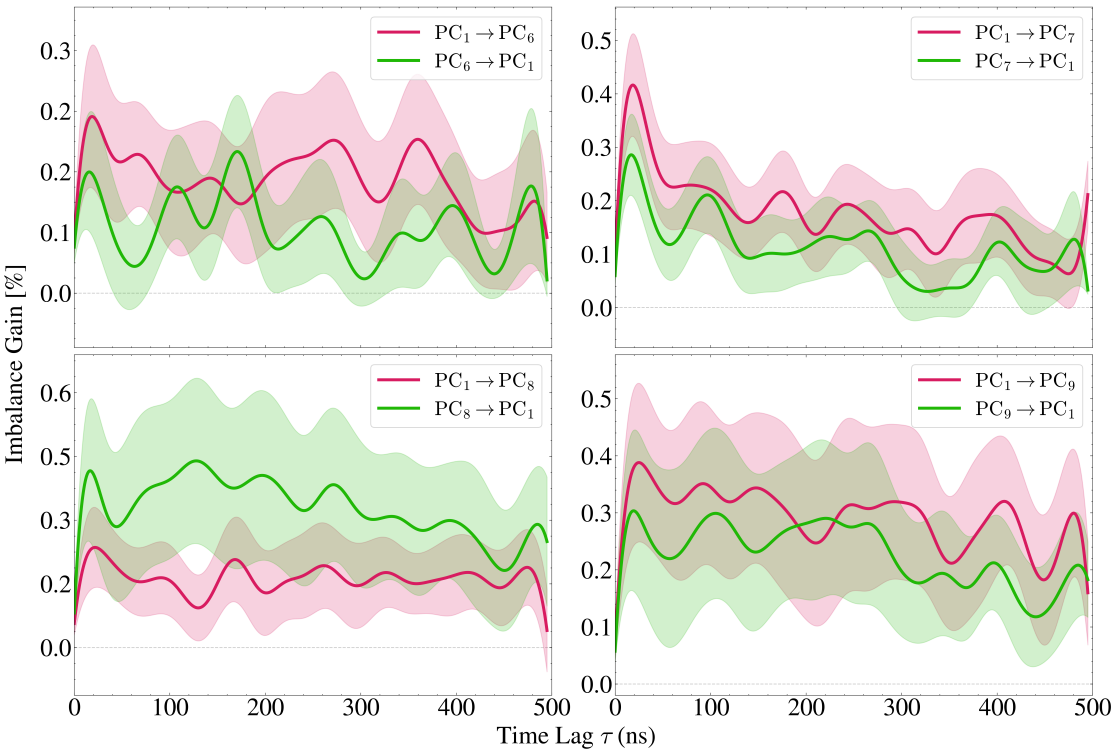


Figure D.7: *Ubiquitin: Pairwise IG curves between* $\mathrm{PC}_1 \Leftrightarrow \mathrm{PC}_{6,7,8,9}$. *The IG values are on the Y-axis, and time lag* (τ) *in nanoseconds is on the X-axis. Pink lines denote the direction* $\mathrm{PC}_1 \to \mathrm{PC}_{6,7,8,9}$, *whereas green lines denote the reverse direction. Shaded areas denote error bars computed over 20 independent realizations.*

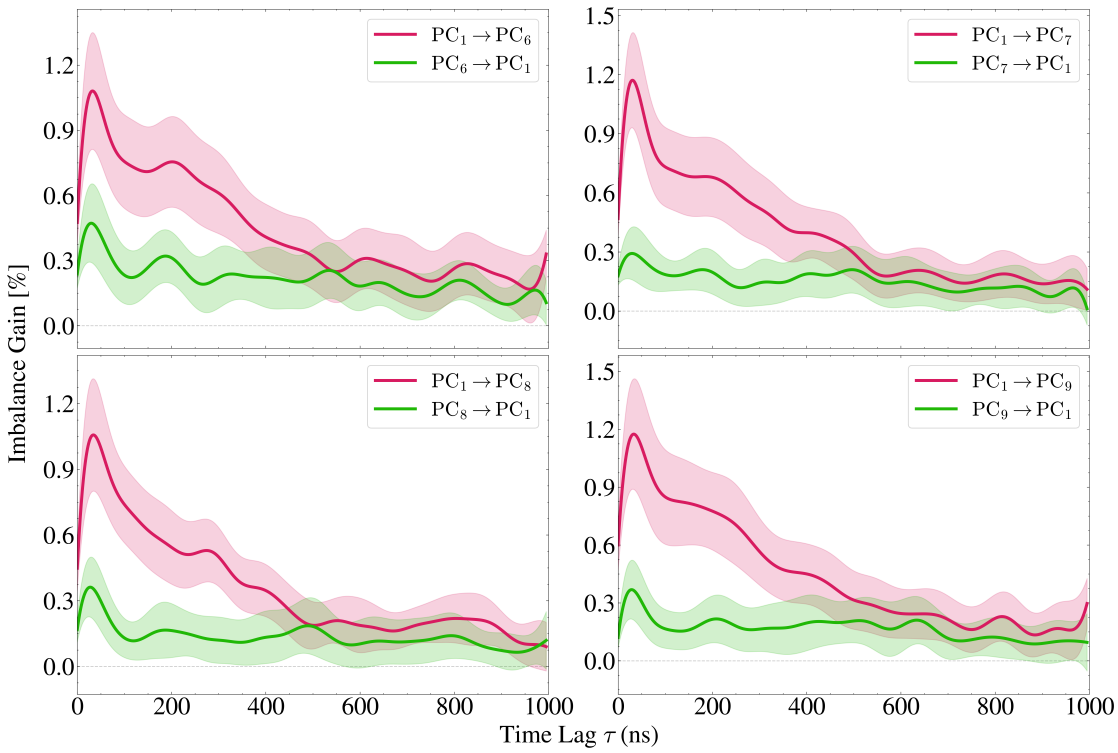


Figure D.8: *NTL9: Pairwise IG curves between* $\mathrm{PC}_1 \Leftrightarrow \mathrm{PC}_{6,7,8,9}$. *The IG values are on the Y-axis, and time lag* (τ) *in nanoseconds is on the X-axis. Pink lines denote the direction* $\mathrm{PC}_1 \rightarrow \mathrm{PC}_{6,7,8,9}$, *whereas green lines denote the reverse direction. Shaded areas denote error bars computed over 20 independent realizations.*

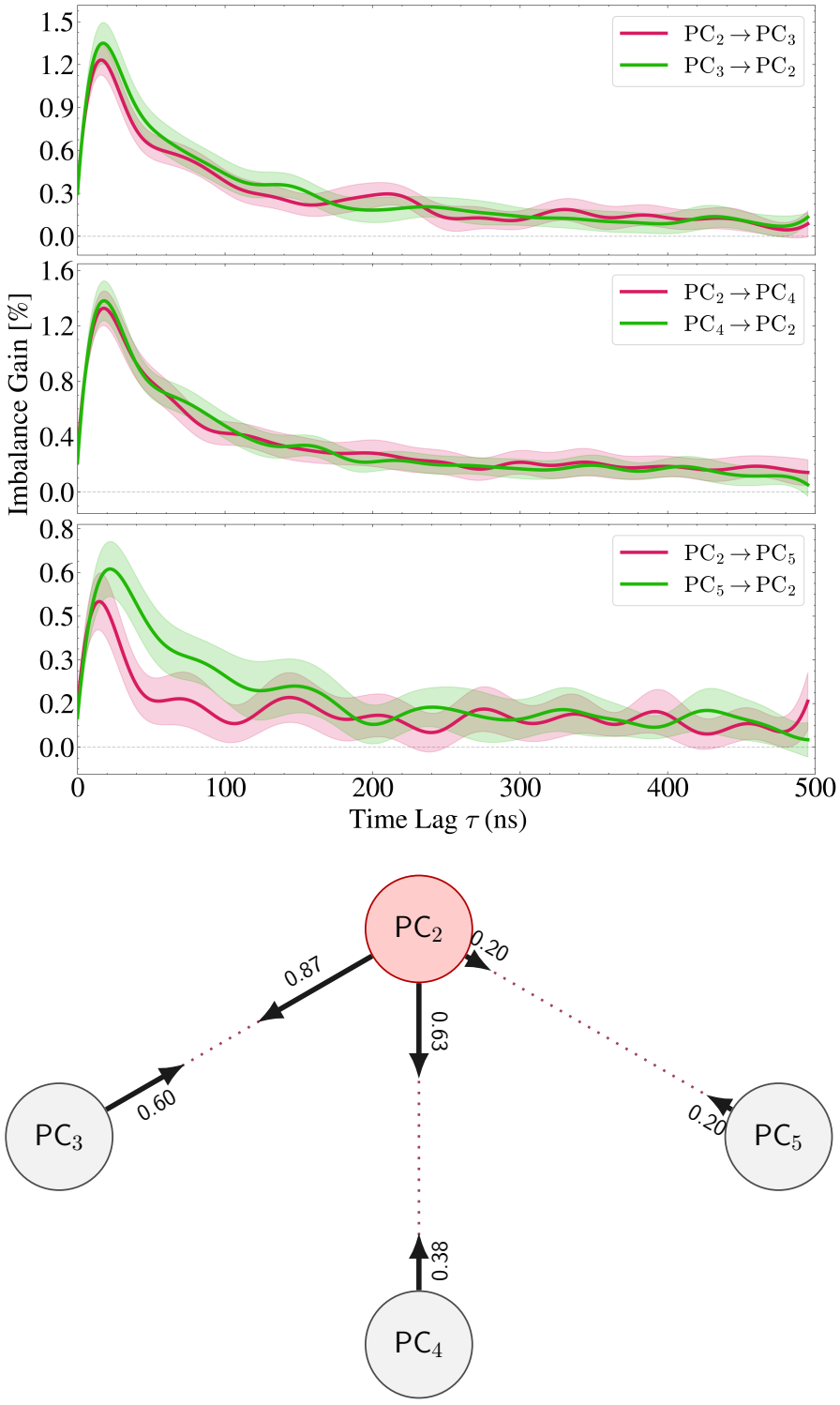


Figure D.9: *Ubiquitin: The top panel shows the pairwise IG curves between* $\mathrm{PC}_2 \Leftrightarrow \mathrm{PC}_{3,4,5}$. *The IG values are on the Y-axis, and time lag* (τ) *in nanoseconds is on the X-axis. Pink lines denote the direction* $\mathrm{PC}_2 \rightarrow \mathrm{PC}_{3,4,5}$, *whereas green lines denote the reverse direction. Shaded areas denote error bars computed over 20 independent realizations. The bottom panel shows the putative causal links summarizing the normalized IGT between* $\mathrm{PC}_2 \Leftrightarrow \mathrm{PC}_{3,4,5}$. *Arrow lengths are proportional to the magnitude of the IGT, and the corresponding IGT values are reported next to each arrow.* PC_2 *exhibits moderately asymmetric causal links with* PC_3 *and* PC_4, *with weaker information transfer in the reverse directions. In contrast, the link between* PC_2 *and* PC_5 *is weak and approximately bidirectional.*

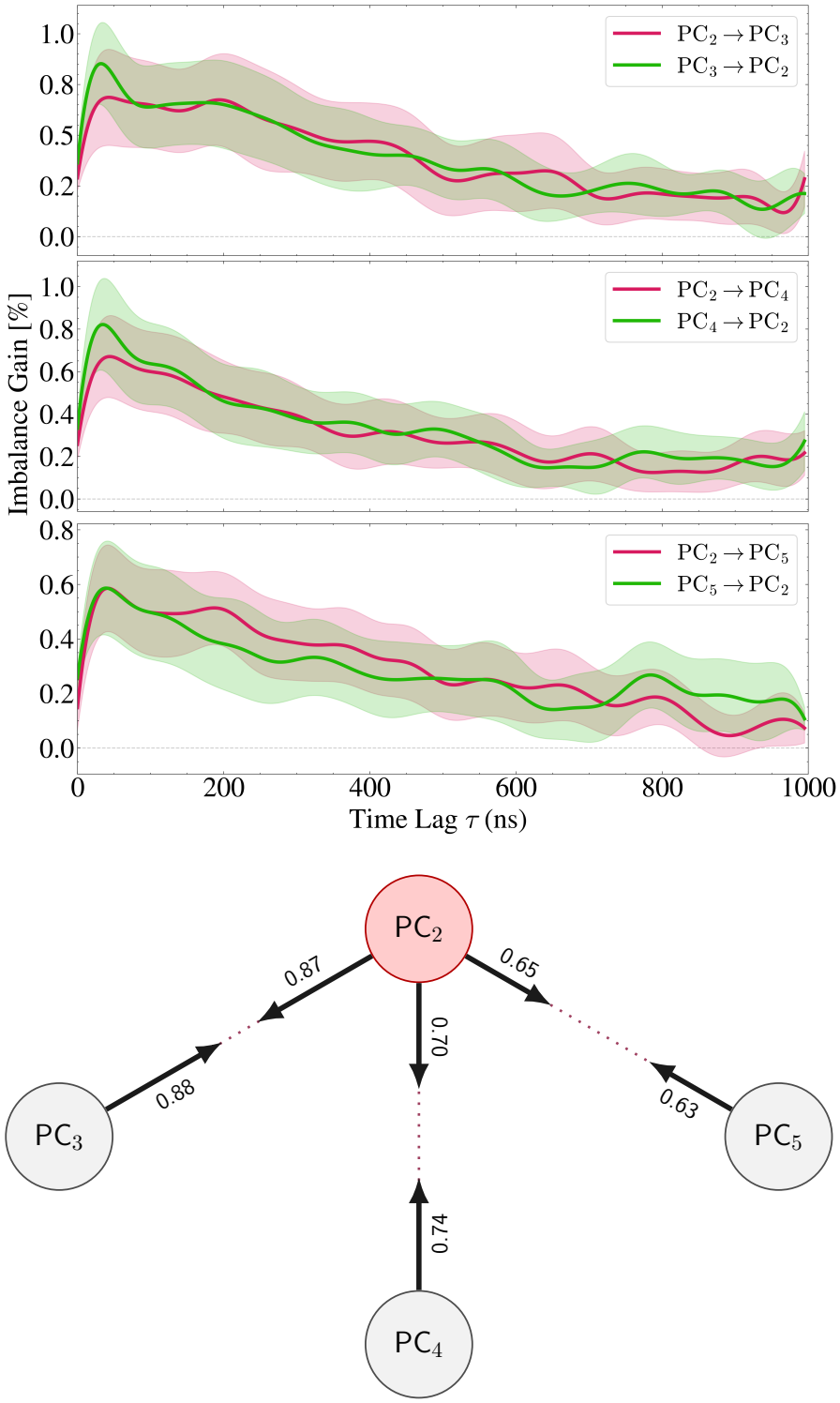


Figure D.10: *NTL9: The top panel shows the pairwise IG curves between* $\mathrm{PC}_2 \Leftrightarrow \mathrm{PC}_{3,4,5}$. *The IG values are on the Y-axis, and time lag* (τ) *in nanoseconds is on the X-axis. Pink lines denote the direction* $\mathrm{PC}_2 \rightarrow \mathrm{PC}_{3,4,5}$, *whereas green lines denote the reverse direction. Shaded areas denote error bars computed over 20 independent realizations. The bottom panel shows the putative causal links summarizing the normalized IGT between* $\mathrm{PC}_2 \Leftrightarrow \mathrm{PC}_{3,4,5}$. *Arrow lengths are proportional to the magnitude of the IGT, and the corresponding IGT values are reported next to each arrow.* PC_2 *exhibits bidirectional, symmetric causal links with* $\mathrm{PC}_{3,4,5}$.

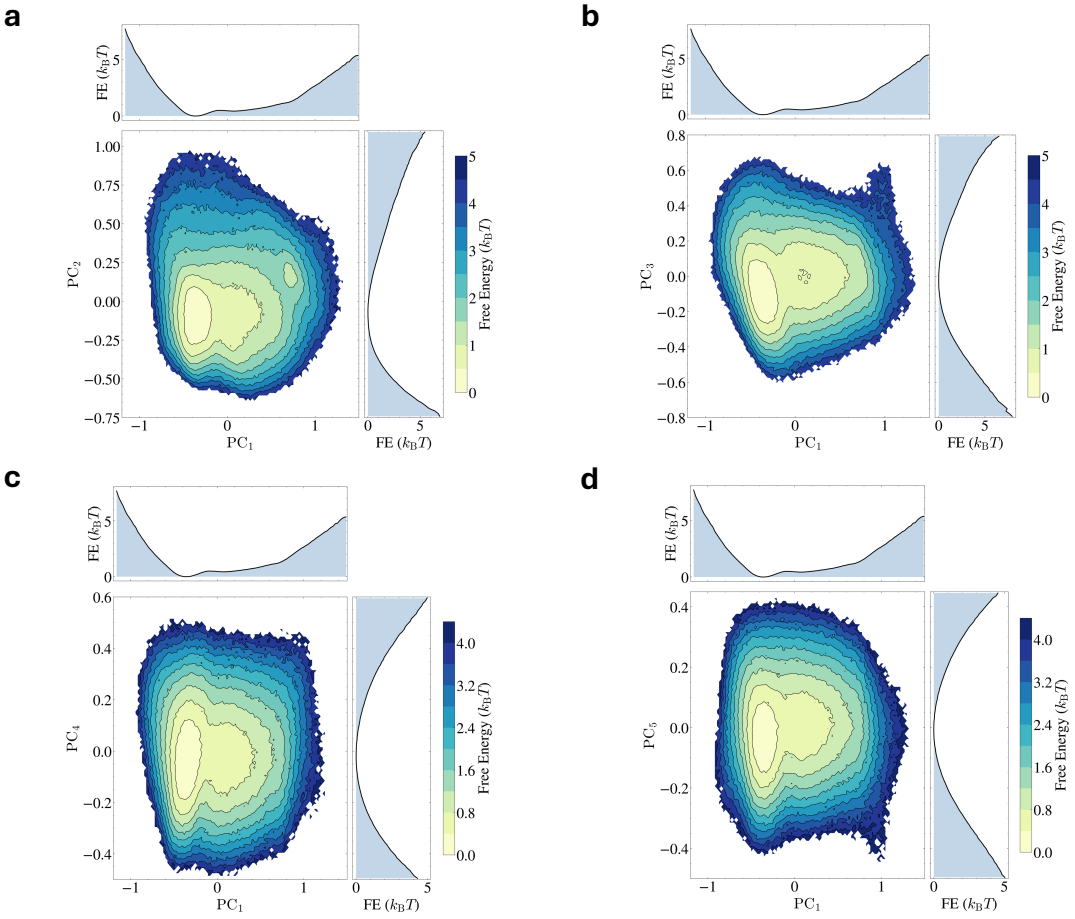


Figure D.11: *Ubiquitin: Free Energy Surface (FES) between* PC_1 *and* $\mathrm{PC}_{2,3,4,5}$ *shown in panels a,b,c,d respectively.*

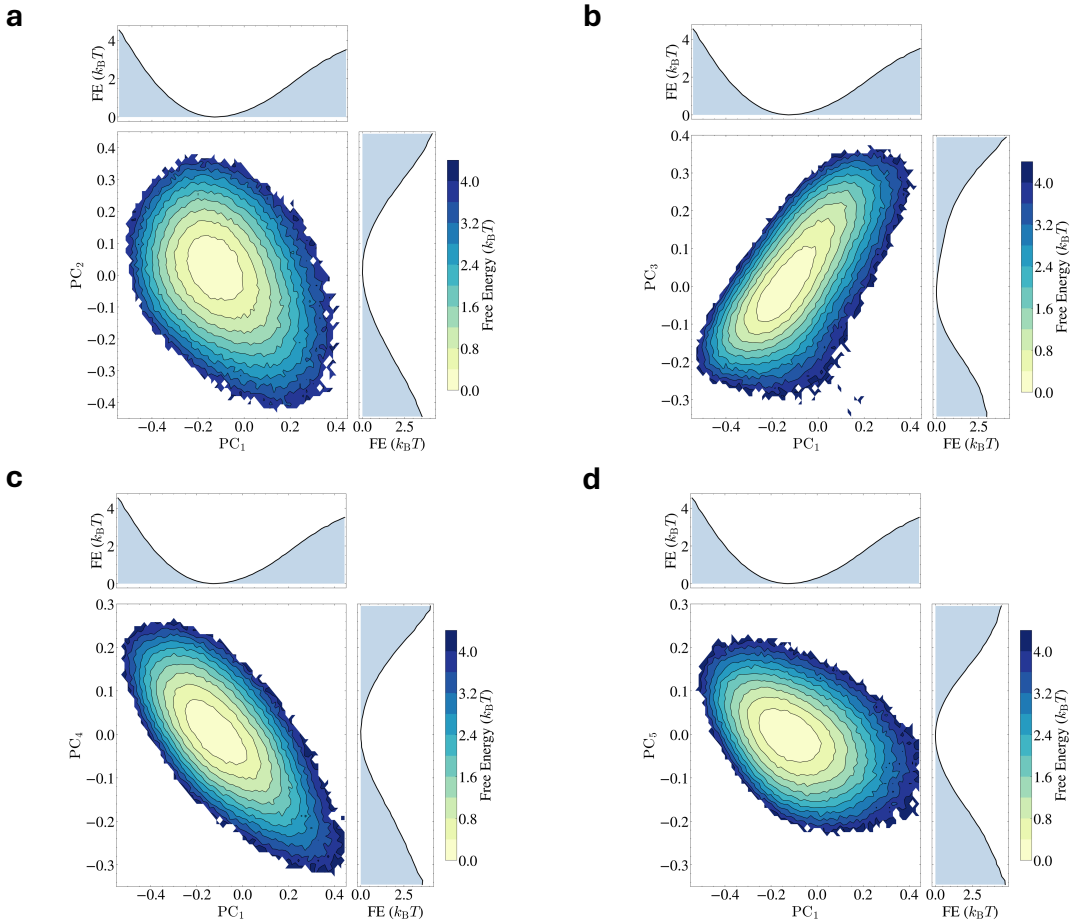


Figure D.12: *NTL9: Free Energy Surface (FES) between* PC_1 *and* $\mathrm{PC}_{2,3,4,5}$ *shown in panels a,b,c,d respectively.*

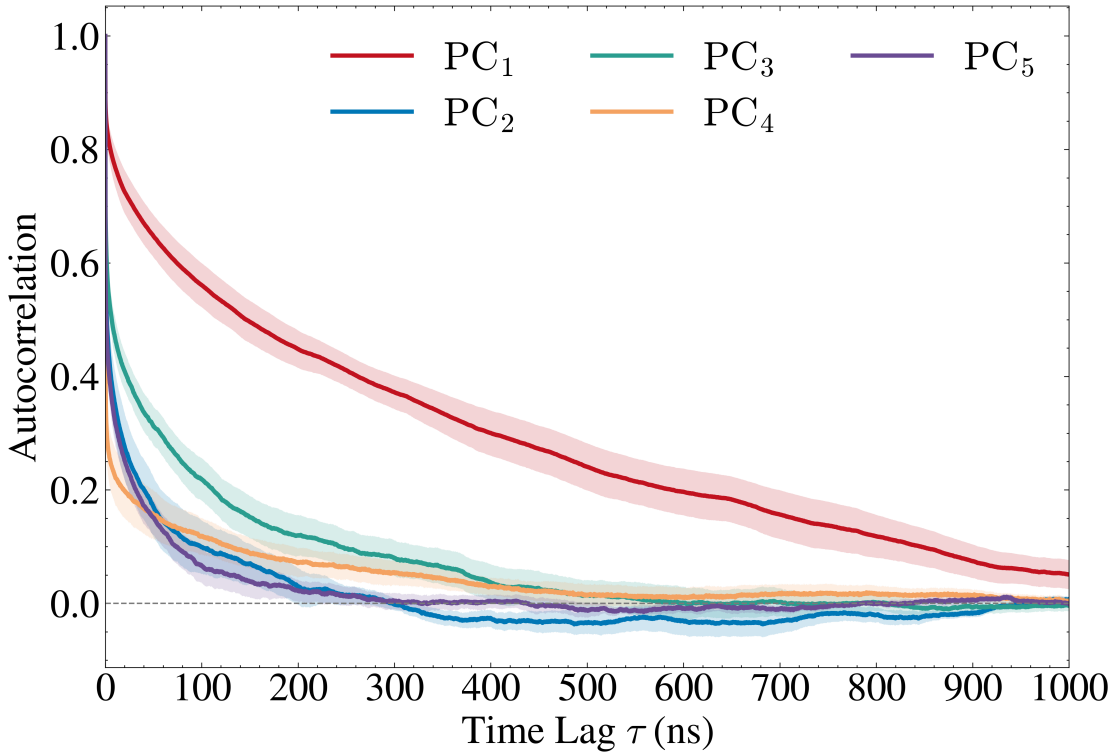


Figure D.13: *The autocorrelation functions of the top 5 PCs,* $\mathrm{PC}_{i=1,5}$*, which are colored red, blue, green, orange, and purple respectively, with shaded areas denoting error bars.*

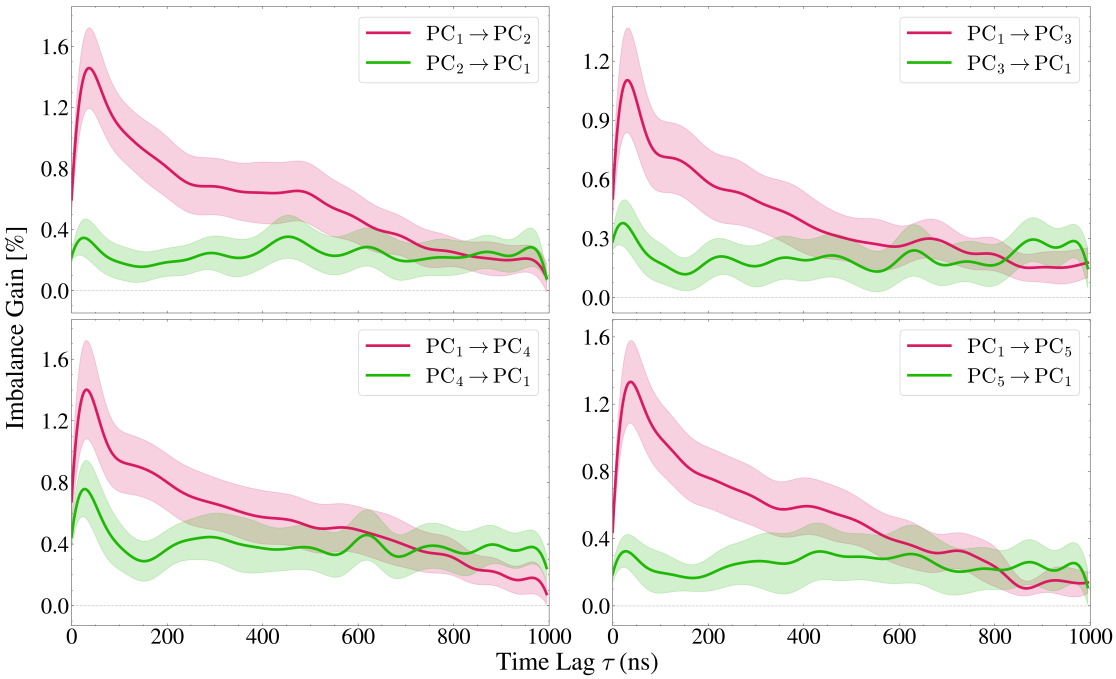


Figure D.14: *The pairwise IG curves over time between* $\mathrm{PC}_1 \Leftrightarrow \mathrm{PC}_{2,3,4,5}$*. The IG values are on the Y-axis, and time lag* (τ) *in nanoseconds is on the X-axis. Pink lines denote the direction* $\mathrm{PC}_1 \rightarrow \mathrm{PC}_{2,3,4,5}$*, whereas green lines denote the reverse direction. Shaded areas denote error bars computed over 20 independent realizations.*

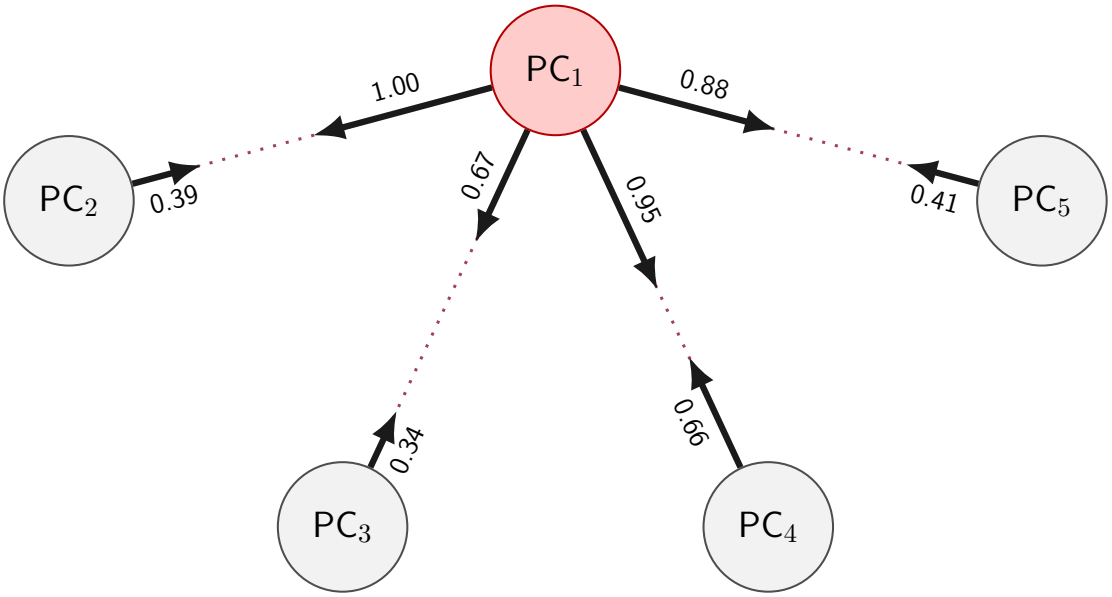


Figure D.15: *Putative causal links between* PC_1 *and the remaining leading principal components. Nodes represent the first five PCs, and directed edges indicate the normalized integrated information transfer,* IGT*, defined in Eq. 2.30. For each pair, the edge length, as well as the numerical label report the relative strength of the directional information transfer, with larger values corresponding to stronger causal links.*

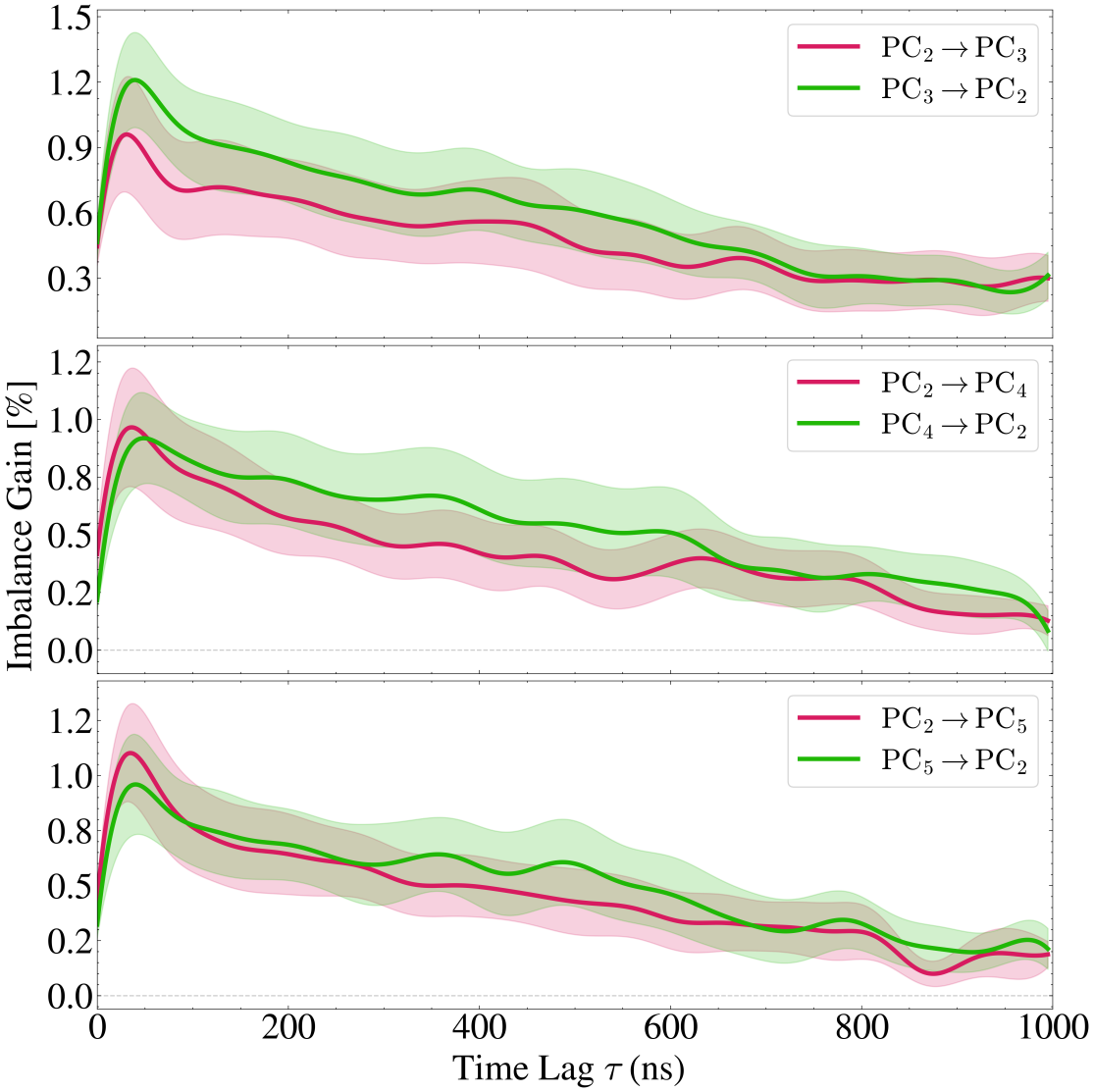


Figure D.16: *The pairwise IG curves over time between* $PC_2 \Leftrightarrow PC_{3,4,5}$. *The IG values are on the Y-axis, and time lag* (τ) *in nanoseconds is on the X-axis. Pink lines denote the direction* $PC_2 \rightarrow PC_{3,4,5}$, *whereas green lines denote the reverse direction. Shaded areas denote error bars computed over 20 independent realizations.*

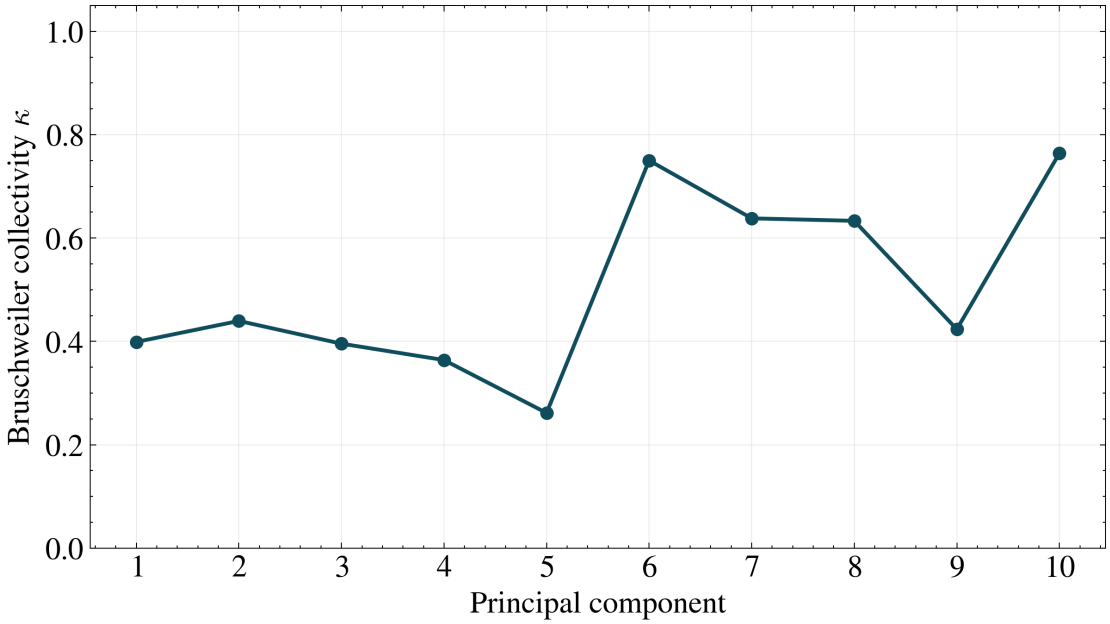


Figure D.17: *Brüschweiler collectivity index for the leading PCs. The collectivity index quantifies the spatial extent of each mode.*