Path Reweighting Methods for underdamped Langevin Dynamics for Molecular Systems

Inaugural-Dissertation
to obtain the academic degree
Doctor rerum naturalium (Dr. rer. nat.)

submitted to the Department of Biology, Chemistry, Pharmacy
of Freie Universität Berlin

by

Stefanie Kieninger

December, 2022

I hereby declare that this dissertation was written and prepared by me independently and that no sources and aids other than those indicated have been used. Intellectual property of other authors has been marked accordingly. I also declare that I have not applied for an examination procedure at any other institution or that I have not submitted the dissertation in this or any other form to any other faculty as a dissertation.

The research presented in this dissertation was carried out during the period from November 2018 to December 2022 at the institute of chemistry and biochemistry at Freie Universität Berlin in the group of Prof. Dr. Bettina G. Keller.

First Reviewer:
Prof. Dr. Bettina G. Keller
Department of Biology, Chemistry, Pharmacy
Physical and Theoretical Chemistry
Freie Universität Berlin
Arnimallee 22, 14195 Berlin

Second Reviewer:
PD Dr. Marcus Weber
Zuse Institute Berlin (ZIB)
Takustraße 7, 14195 Berlin

Date of Disputation: 04.04.2023

Acknowledgments

The past four years have been a great enrichment for me both as a scientist and as a colleague and friend. I have been fortunate to meet many excellent people who helped me to constantly broaden my horizon.

First and foremost, I would like to thank Prof. Dr. Bettina G. Keller for the opportunity to work on such an interesting and challenging topic. Your ability to illustrate complex problems in a simple and effective manner, and your way to lead a productive group, which distinguishes itself by mutually respectful and beneficial interactions as well as friendship, impressed me deeply. You helped me to become the scientist I am today and I am grateful to have had you as a female role model in a field where I have often been the only woman.

Next, I would like to acknowledge my second supervisor PD Dr. Marcus Weber who readily shared his expertise and was always available for feedback on mathematical questions.

I would like to express my deepest gratitude to the theoretical chemistry department and Bettina Keller's group in particular. You people created a positive and supportive work atmosphere that has inspired me and left a lasting impression. The extensive coffee breaks, interesting discussions, group barbecues and other (sometimes socially awkward) moments which we shared as a group and as friends will always remain in my memory.

A special thanks goes to Dr. Luca Donati, who always took the time to engage in scientific discussions and readily shared his expertise to help me grow as a scientist.

Additionally, I would like to acknowledge my collaboration partner Minh Nguyen Trung and his supervisor Prof. Dr. Dorothea Fiedler for the fruitful collaboration during which I learned many new things about biochemical experiments.

Another special thanks goes to PD Dr. Dirk Andrae. Your fascination for mathematics and willingness to pass on knowledge to a new generation of students deeply inspired me. Being a tutor for your mathematics courses was a major building block of my education and certainly paved the path to the science I do today.

Last but certainly not least, I thank my partner Bekka. I am deeply grateful for your unconditional support, warmth and loyalty. You have been my companion and best friend for so many years and I am absolutely excited to start a new chapter of our book together in Denmark.

Contents

List of Publications

Part A: Dynamical Quantities via Molecular Dynamics Simulations

A1: "GROMACS Stochastic Dynamics and BAOAB are equivalent configurational sampling algorithms"
S. Kieninger, B. G. Keller
J. Chem. Theory Comput., **2022**, *18*, 5792–5798.
DOI: 10.1021/acs.jctc.2c00585

A2: "Dynamical reweighting methods for Markov models"
*S. Kieninger, *L. Donati, B. G. Keller
Curr. Opin. Struct. Biol., **2020**, *61*, 124–131.
DOI: 10.1016/j.sbi.2019.12.018
* Co-first author

A3: "Path probability ratios for Langevin dynamics - Exact and approximate"
S. Kieninger, B. G. Keller
J. Chem. Phys., **2021**, *154*, 094102.
DOI: 10.1063/5.0038408

Part B: Dynamical Quantities via Experimental Data

B1: "Stable Isotopomers of *myo*-Inositol Uncover a Complex MINPP1-Dependent Inositol Phosphate Network"
M. Nguyen Trung, S. Kieninger, Z. Fandi, D. Qiu, G. Liu, N. K. Mehendale, A. Saiardi, H. Jessen, B. G. Keller, D. Fiedler
ACS Cent. Sci., **2022**, *8*, 1683-1694.
DOI: 10.1021/acscentsci.2c01032

Contribution to the publications: A description of my contribution to the listed publications is summarized in the preface of the respective sections.

List of Abbreviations

BIRD	Bilinear Rotation Decoupling
HMQC	Heteronuclear Multiple Quantum Coherence
NMR	Nuclear Magnetic Resonance
CE-MS	Capillary Electrophoresis - Mass Spectrometry
MD	Molecular Dynamics
ODE	Ordinary Differential Equation
PDE	Partial Differential Equation
SDE	Stochastic Differential Equation
FP	Fokker-Planck
MSM	Markov State Model
SI	Supporting Information
$X(t)$	random variable
$W(t)$	Wiener process
q	position
p	momentum
η	random number
Δt	time step
m	mass
k_B	Boltzmann constant
T	temperature
ξ	collision rate
$V(q)$	potential energy function
$U(x)$	bias
$\boldsymbol{\omega}$	path
ω_k	state at iteration step k
Ω	state space
Γ	configuration space
$\mathbf{M}$	mass matrix
E^{kin}	kinetic energy
N_{dof}	number of degrees of freedom
$\Delta\eta$	random number difference
τ	lag time
$C(\tau)$	time-correlation function
$\mathbf{C}(\tau)$	time-correlation/count matrix
$\mathbf{T}(\tau)$	transition probability matrix
g	configuration space reweighting factor
M	path space reweighting factor

Abstract

Knowledge about the dynamical properties of biomolecules is essential to understand their function in biological processes. This thesis approaches the task to compute dynamical properties with two different strategies. Part **A** focuses on Molecular Dynamics (MD) simulations combined with path reweighting. Three of the most widely used underdamped Langevin integrators for MD simulations are the splitting methods BAOAB and BAOA which are available in the MD packages OpenMM and AMBER and the Gromacs Stochastic Dynamics (GSD) integrator implemented in GROMACS. We found that all three integrators are equivalent configurational sampling algorithms and thus yield configurational properties at equivalent accuracy. MD simulations with stochastic integrators such as Langevin integrators offer the possibility to reweight estimated dynamical properties using path reweighting. With path reweighting we can for example recover the original dynamics from MD simulation that have been conducted with enhanced sampling methods. The key component of path reweighting is the path reweighting factor M which strongly depends on the chosen integrator. We derive M_L for underdamped Langevin dynamics propagated by a variant of the Langevin Leapfrog integrator. Additionally, we present two strategies which can be used as blueprints to straightforwardly derive M_L for other Langevin integrators. The previously reported path reweighting factor matches the Euler-Maruyama integrator for overdamped Langevin dynamics and was used as standard reweighting factor even though the MD simulation was conducted with an underdamped Langevin integrator. We prove that this path reweighting factors differs from the exact M_L only by $\mathcal{O}(\xi^4\Delta t^4)$ and thus yields highly accurate dynamical reweighting results (Δt is the integration time step, and ξ is the collision rate.).
Part **B** of this thesis combines experimental and theoretical approaches to investigate Multiple Inositol Polyphosphate Phosphatase 1 (MINPP1)-mediated inositol polyphosphate (InsP) networks. We use ^{13}C-labeling experiments combined with nuclear magnetic resonance spectroscopy (NMR) to uncover a novel branch of InsP dephosphorylation in human cells. Additionally, we extract the corresponding reaction rates using a Markovian kinetic scheme as theoretical model to describe the network.

Zusammenfassung

Wissen über die dynamischen Eigenschaften von Biomolekülen ist für das Verständnis ihrer Funktion in biologischen Prozessen unerlässlich. Diese Arbeit geht die Berechnung dynamischer Eigenschaften auf zwei verschiedene Arten an. Teil **A** konzentriert sich auf moleküldynamische (MD) Simulationen in Kombination mit Pfadumgewichtung. Drei der am weitesten verbreiteten Langevin-Integratoren für MD-Simulationen sind die Splittingmethoden BAOAB und BAOA, die in den MD-Paketen OpenMM und AMBER verfügbar sind, sowie der Gromacs Stochastic Dynamics (GSD)-Integrator, der in GROMACS implementiert ist. Wir zeigen, dass alle drei Integratoren äquivalente Konfigurations-Sampling-Algorithmen sind und somit Konfigurationseigenschaften mit gleicher Genauigkeit liefern. MD-Simulationen mit stochastischen Integratoren wie z. B. Langevin-Integratoren bieten die Möglichkeit, dynamische Eigenschaften mit Hilfe von Pfadumgewichtungsmethoden umzugewichten. Mit Pfadumgewichtungsmethoden können wir zum Beispiel die ursprüngliche Dynamik aus MD-Simulationen wiederherstellen, welche mit Enhanced-Sampling-Methoden durchgeführt wurden. Die Schlüsselkomponente von Pfadumgewichtungsmethoden ist der Pfadumgewichtungsfaktor M, der stark von dem gewählten Integrator abhängt. Wir leiten M_L für eine angevin-Dynamik her, die mittels einer Variante des Langevin-Leapfrog-Integrators propagiert wird. Zusätzlich stellen wir zwei Strategien vor, die als Blaupausen zum Herleiten der M_L anderer Langevin-Integratoren verwendet werden können. Der zuvor berichtete Pfadumgewichtungsfaktor entspricht dem Euler-Maruyama-Integrator für überdämpfte Langevin-Dynamik und wurde als Standardpfadumgewichtungsfaktor verwendet, obwohl die jeweilige MD-Simulation mit einem Langevin-Integrator durchgeführt wurde. Wir beweisen, dass sich dieser Pfadumgewichtungsfaktor vom exakten M_L nur um $\mathcal{O}(\xi^4 \Delta t^4)$ unterscheidet und somit hochpräzise umgewichtete dynamische Größen liefert (Δt ist der Integrationszeitschritt und ξ ist die Kollisionsrate.).

Teil **B** dieser Arbeit kombiniert experimentelle und theoretische Ansätze um Inositol Polyphosphat (InsP) Netzwerke, welche von Multiple Inositol Polyphosphate Phosphatase 1 (MINPP1) vermittel wurden, zu untersuchen. Wir verwenden ^{13}C-Markierungsexperimente in Kombination mit Kernspinresonanzspektroskopie (NMR), um einen neuen Zweig der InsP-Deposphorylierung in menschlichen Zellen aufzudecken. Darüber hinaus extrahieren wir die entsprechenden Reaktionsraten unter Verwendung eines Markov'schen kinetischen Schemas als theoretisches Modell zur Beschreibung des Netzwerks.

1 Introduction

Motivation Everywhere in nature, dynamical processes occur on timescales of various magnitudes. Our solar system formed and evolved over billions of years,[1] continents on earth have taken several million years to form,[2] and weather phenomena can last for minutes to centuries.[3] Much shorter time scales can be observed on a microscopic scale. The dynamics of biomolecules, for example, span multiple orders of magnitude,[4–7] from fast bond vibrations[8,9] to slow protein folding events.[10–16] Biomolecules are fundamental building blocks of living organisms and their dynamics are a critical element of their function.[17] To investigate the behavior of biomolecules, we can evaluate dynamical quantities such as binding rates, time scales of conformational changes between metastable states, relaxation times and reaction rates. To measure dynamical quantities, we can either deploy computer simulations or perform experiments. In laboratory experiments, we usually measure the properties of bulk matter. One of the advantages of laboratory experiments is that we can, for example, observe biomolecules in environments as close as possible to their corresponding living organism or we can isolate them in biochemical experiments. However, the experimental techniques to handle biomolecules and measure their properties with quantification methods usually pose a major challenge. In computer simulations such as Molecular Dynamics (MD) or Quantum Mechanics (QM) simulations, we usually describe the system on a molecular level. A big advantage of computer simulations is that we can observe microscopic quantities that would be extremely complex or even impossible to observe in laboratory experiments. Unfortunately, depending on the size of the system and the level of theory, computer simulations can be computationally expensive and very time consuming.
This thesis explores both simulation-based and experimental strategies to determine dynamical quantities. Part **A** focuses on MD simulations and the Girsanov reweighting method which can be used to shed light onto dynamical processes characterized by large energy barriers. Part **B** combines experimental and theoretical approaches in order to study the complex dephosphorylation network of two different inositol polyphosphates.

A: MD simulations MD simulations are computer simulations that describe the system of interest on a molecular level. The molecules are represented as a set of particles, whose inter- and intramolecular interactions are approximated by a potential energy function and a corresponding parameter set referred to as a force field. Solving the equation of motion yields an integration scheme, also called integrator, that can be used by an MD simulation program to approximate the true solution of the given equation of motion as a time-discretized trajectory.[18–20] The trajectories can then be analyzed to give insight into the different dynamical processes of the system.[21–25]
This thesis focuses on Langevin dynamics approximated by Langevin integrators[26], which belong to the class of stochastic integrators and are widely deployed to correctly sample the canonical ensemble.[27] There exist two regimes for Langevin dynamics: the under-

damped regime governed by the Langevin equation of motion and the overdamped regime that corresponds to its high friction limit. The most commonly used integrator for overdamped Langevin dynamics is the Euler-Maruyama integrator[28,29], however, for underdamped Langevin dynamics a huge variety of different integrators[19,30–45] is available. As a consequence, we are immediately confronted with the questions: What are the similarities and differences between the underdamped Langevin integrators? Which integrator is the best choice for a specific application? During the last decade, these questions have been addressed and similarities between integrators under specific conditions have been reported.[30,31,33,37,40,46–50] Part **A1** of this thesis picks up at this point and argues that two of the most widely used Langevin integrators are equivalent configurational sampling algorithms.

Potential reweighting Due to large energy barriers, some molecular transitions have time scales well beyond what MD simulations are capable of covering. To overcome this limitation we can deploy enhanced sampling techniques that either raise the temperature of the system[51–55] or introduce a bias to the potential energy function[56–60] in order to provide more energy to cross the barriers. Unfortunately, enhanced sampling techniques alter the dynamics because they add energy to the system as schematically shown in fig. 1.1, left. To recover the original dynamical information from the biased simulations we have to apply dynamical reweighting techniques, such as potential reweighting or temperature reweighting. This thesis focuses on potential reweighting and for further information on temperature reweighting, the reader is referred to Refs. [61–63]. Another application of potential reweighting is found in the context of force field optimization. Fig. 1.1, right schematically shows how potential reweighting can help to study the influence of force field parameters on the dynamics of the system with minimal computational effort.
In recent years, a number of potential reweighting techniques have been proposed that are based on different formulations of molecular transitions. Part **A2** of this thesis collects all of these methods in a review paper and groups them into four main categories based on the framework from which the method has been developed and its underlying assumptions. Path reweighting represents one of these categories and is a method that reweights path ensemble averages like time-lagged correlation functions from short paths generated by a biased simulation. For successful reweighting in path space, we need the mathematical expression of the path reweighting factor M, which is the ratio of the probability to observe a path in the unbiased potential and the probability to observe the same path in the biased potential. Unfortunately, this expression strongly depends on the integrator that was used to generate the biased paths and has to be derived from the corresponding integrator equations. For overdamped Langevin dynamics, the path reweighting factor is connected to the Onsager-Machlup function[64–66] and has been known for several decades. Path reweighting factors for underdamped Langevin dynamics integrators have not yet been reported. Additionally, we face the problem that M is very expensive to calculate because it requires knowledge of the configurational state and the corresponding unbiased forces at each and every simulation step. This quickly leads to performance issues and memory problems for high-dimensional systems. As

a consequence, potential reweighting has been limited to diffusion in low-dimensional energy landscapes[67–70] and to short trajectories of alanine dipeptide.[71] The Girsanov reweighting method[72–75] for Markov State Models[76–81] (MSMs) proposed a solution to this problem. The method expresses M in terms of the random number η_k that was drawn by the integrator at each simulation step k and the random number difference $\Delta\eta_k$ that relates the biased trajectory to the unbiased forces. This way, only the biasing force instead of the unbiased forces is required at each simulation step. Donati *et al.* additionally implemented Girsanov reweighting on-the-fly and were the first to report the application of potential reweighting to a long trajectory of a molecular system.[73] Later on, Donati *et al.* successfully combined the enhanced sampling technique metadynamics[58–60] with Girsanov reweighting to compute mean first hitting times in alanine dipeptide and to estimate the implied timescale associated with the opening and closing of a β-hairpin as well as the corresponding conformational states.[74,82] In both works, the authors use underdamped Langevin dynamics to generate the biased trajectories and achieved excellent results with an reweighting factor M_{approx} that approximates $\Delta\eta_k$ with the expression derived from overdamped Langevin dynamics. However, at the time it was not quiet clear why this approximation yielded such accurate results. Part **A3** of this thesis provides an answer to this question including the corresponding mathematical proof and, for the first time in literature, reports the path reweighting factor for an underdamped Langevin integrator. Additionally, **A3** reports a simple scheme that can be used to derive the path reweighting factor for any other underdamped Langevin integrator and provides detailed insight into the relationship between biasing force, biased forces and random numbers. Please note that this relationship has already been indicated in Refs. [72] and [73].

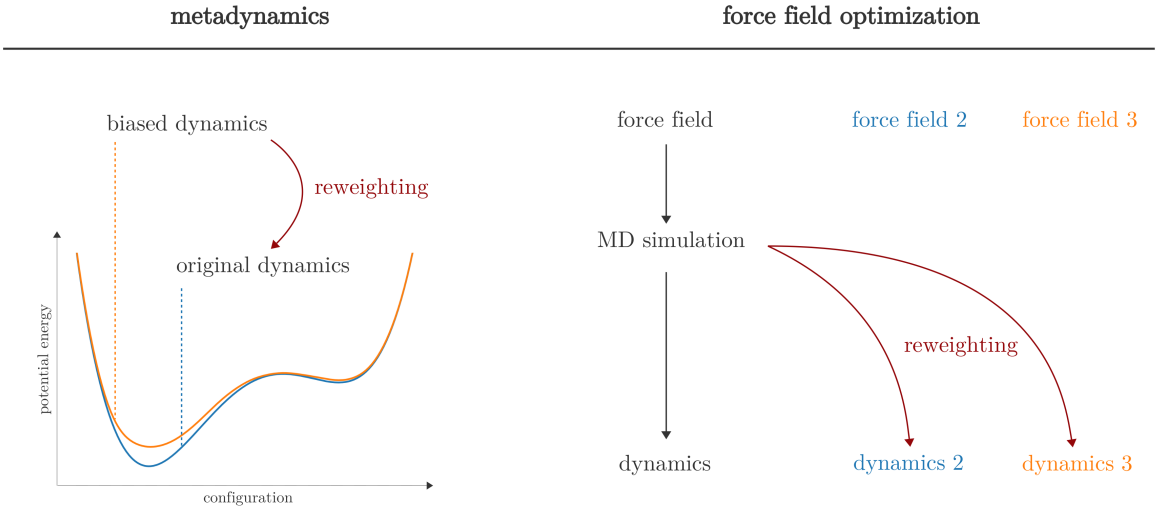


Figure 1.1:
Schematic representation of two different potential reweighting applications: in combination with metadynamics (left) and as a tool for force field optimizations (right).

B: Experimental data and kinetic models The combination of theoretical and experimental approaches can be a very powerful tool to investigate the function of biomolecules in natural processes like metabolism or signal transduction pathways. Dynamical quantities are usually not directly accessible with analytical methods and we have to measure correlations or time series of quantities that contain dynamical information about the system. Unfortunately, it can be very difficult to extract this kind of information from experimental data because it is often noisy and different species are difficult to distinguish. However, if the measured data set is well-resolved and interpretable, theoretical models can be used to extract dynamical quantities. A very simple but powerful theoretical model to extract reaction rates from experimentally measured time series of concentrations is the kinetic scheme.[83,84]

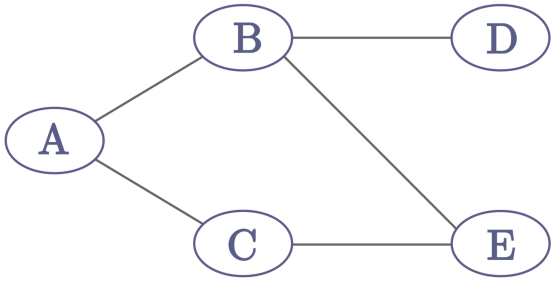


Figure 1.2:
Schematic representation of a kinetic scheme with 5 states.

A kinetic scheme as depicted in Fig. 1.2 can be understood as a network of distinct states where the connections between the states represent the reaction step from one state to the other. If the connections are constant with respect to time, the kinetic scheme is called Markovian. The mathematical representation of a kinetic scheme is called a master equation and describes the dynamics in terms of the probability to occupy each of the states at time t. As an example, we can interpret Fig. 1.2 as the reaction network of an enzyme that converts substrate A via two intermediates B and C to the products D and E along two distinct pathways. In this context, we can understand the different species $A, B, \ldots, E$ as distinct states and link the concentration of a species at time t to the probability to find the system in the corresponding state. Additionally, we can interpret the line between two states as the (time-constant) reaction rate for the forward and the backward reaction. The kinetic scheme can be used to either calculate the time evolution of the concentrations from the reaction rates or to extract the reaction rates from the time evolution of the concentrations.
A real life example of such a kinetic network is the inositol polyphosphate (InsP) metabolism in eukaryotes. InsPs are small, water-soluble molecules that play a key role in fundamental physiological processes.[85–87] All InsPs have a *myo*-inositol scaffold but vary in their phosphorylation pattern.[88–91] Studies showed, that inositol-1,3,4,5,6-pentakisphosphate ($InsP_5$[2OH])

and inositol hexakisphosphate ($InsP_6$) are precursors for the biosynthesis of inositol pyrophosphates (PP-InsPs) which have recently drawn increasing attention due to their involvement in central signaling processes.[92] Moreover, $InsP_6$ has been reported to be a structural cofactor in proteins and protein complexes and as a "molecular glue" for protein-protein interactions.[93–96] It is known that different InsPs can be converted into each other via phosphorylation or dephosphorylation pathways catalyzed by kinases or phosphatases, respectively. The kinase-mediated phosphorylation pathways have been studied fairly well, however, the dephosphorylation pathways are largely unresolved.[97] The only known phosphatase in the human genome able to dephosphorylate $InsP_6$ is MinPP1 (Multiple Inositol Polyphosphate Phosphatase 1). MINPP1 has recently been linked to a neurodegenerative disease that severely impacts cognitive functions and life expectancy.[98,99] To gain a deeper insight into the functions of InsPs and the role of MINPP1 in human health and diseases it can be very beneficial to study the MINPP1 mediated dephosphorylation pathways of different InsPs.
Part **B1** of this thesis studies the MINPP1 mediated dephosphorylation of $InsP_5$[2OH] and $InsP_6$ in collaboration with the Fiedler group (Leibniz-Forschungsinstitut für Molekulare Pharmakologie and Humboldt-Universität zu Berlin). We combine advanced biochemical ^{13}C-labeling experiments with BIRD-{^{1}H-^{13}C}HMQC-NMR measurements to extract the concentration time series of different InsP metabolites and deploy a kinetic scheme to extract reaction rates from the experimental data set.

1.1 Structure and research questions

This thesis is divided into two parts: **A** dynamical quantities via MD simulations and **B** dynamical quantities via experimental data.

Section 2 provides a short explanation of the mathematical background and the methods used to address the research questions in both parts. We explain the basic concepts of stochastic calculus, introduce underdamped and overdamped Langevin dynamics and establish the connection between the corresponding stochastic differential equations and their Fokker-Planck equations. Furthermore, we introduce path probabilities and path expected values and summarize the concept of path reweighting. Finally, we briefly recall the theoretical background for Markov State Models.

Section 3 presents all publications that have been produced as a part of this thesis. Additionally, we give a short summary of each paper in the individual prefaces.

A1 In this publication we focus on the underdamped Langevin integrators BAOAB, BAOA and GSD. We address the research questions:
1) What are the similarities and differences between the Langevin integrators?
2) Which integrator is the best choice for a certain application?

A2 In this publication we summarize and categorize state-of-the art potential reweighting techniques. We address the question:
1) Which potential reweighting methods exist?
2) What are their similarities and differences?

A3 In this publication, we focus on the path reweighting for overdamped and underdamped Langevin dynamics. We address the research questions:
1) How can we derive path reweighting factors for underdamped Langevin dynamics integrators?
2) Why does reweighting with M_{approx} yield excellent reweighting results for underdamped Langevin dynamics?
3) What is the relationship between the different representations of the reweighting factor?

SI Supporting Information for part A
This section presents non-published additional information for part **A**. We repeat the numerical accuracy studies reported in part **A1** with more underdamped Langevin integrators and extend the study to dynamical quantities. Additionally, we apply the strategy introduced in part **A3** to derive

the path reweighting factor M for the ABOBA integrator.

B1 This publication combines experimental and theoretical approaches to shed light onto the MINPP1 mediated dephosphorylation pathways of $InsP_5[2OH]$ and $InsP_6$, respectively. We address the research questions:
1) Do lower phosphorylated InsPs play a role in the InsP metabolism in mammalian cells?
2) What are the distinct MINPP1 mediated dephosphorylation pathways of $InsP_5[2OH]$ and $InsP_6$?

Section 4 draws the conclusion of this thesis and provides an outlook on how research could be proceeded from here.

2 Theory

2.1 Stochastic calculus

In Molecular Dynamics (MD) simulations, the dynamics are governed by an equation of motion which describes the movement of each particle in the system according to classical mechanics. To perform simulations in the NVT ensemble (number of particles N, volume V, temperature T) and hence control the temperature, we can deploy deterministic thermostats such as the Nosé-Hoover thermostat or stochastic thermostats such as the Andersen[100] or the Langevin thermostat.[27] This thesis focuses on Langevin thermostats which include stochastic forces to couple the system to an external heat bath. Consequently, we have to work with stochastic processes, Stochastic Differential Equations (SDEs) and methods from stochastic calculus. This section briefly introduces the Wiener process and stochastic integration and uses both to derive the solution of the Ornstein-Uhlenbeck process. All content within this section is based on Refs. [19, 29, 101, 102]. The mathematics is presented for the one-dimensional case but can also be extended to N dimensions (see Refs. [19, 29, 101, 102]).

2.1.1 Stochastic processes

A stochastic process is a mathematical model that can be used to describe systems that appear to vary in a random manner. They have a variety of applications for example in biology,[103] chemistry,[83] physics[83,104] or signal processing.[105] A stochastic process can be described by a random variable $X(t)$ whose values $x_0, x_1, x_2, \dots$ at times $t_0, t_1, t_2, \dots$ can be measured. The time evolution of $X(t)$ can be represented by an SDE. Additionally, we assume the existence of a set of joint probability densities $P(x_0, t_0; x_1, t_1; \dots)$ which describe the system completely. This means, that we can either study the system from a probabilistic point of view via diffusion equations and Fokker-Planck equations (discussed in sec. 2.3) or from a trajectory point of view via the SDE (discussed sec. 2.2). Note that sec. 2.3.1 discusses the connection between FP-equations and SDEs in detail. In general, there exists a variety of stochastic processes with different properties such as Bernoulli, Wiener or Poisson processes.[29,101] However, this thesis only focuses on Wiener processes (sec. 2.1.3).

2.1.2 The Gaussian distribution

The Gaussian distribution, also called normal distribution or bell curve, is a type of continuous probability distribution. In physics, metrology and social science it is used to represent the distribution of real-valued random variables. One of its most prominent applications is

the description of Brownian motion. The importance of Gaussian distributions in the previously mentioned fields is partly due to the central limit theorem.[106] The theorem states that the properly normalized sum of many independent random variables converges towards a Gaussian distribution as the number of random variables increases. This is the case even if the random variables themselves are not Gaussian.

Consider the Gaussian random variable X. The corresponding probability density function is the Gaussian

$$X \sim P(x) = \sqrt{\frac{1}{2\pi\alpha^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\alpha^2}\right) \tag{2.1}$$

with μ being the mean and α^2 representing the variance of $P(x)$. A Gaussian with mean μ and variance α^2 can equivalently be represented by the notation $\mathcal{N}(\mu, \alpha^2)$.
The moments of X are given as

$$\text{first moment} \qquad \langle X \rangle = \mu \tag{2.2}$$

$$\text{second moment} \qquad \left\langle X^2 \right\rangle = \mu^2 + \alpha^2 \,. \tag{2.3}$$

The first moment $\langle X \rangle$ is always equal to the mean μ. If the mean is zero $\mu = 0$ then the second moment is always equal to the variance

$$\mathrm{Var}(X) = \left\langle (X-\mu)^2 \right\rangle = \langle X^2 \rangle - \langle X \rangle^2 = \langle X^2 \rangle = \alpha^2 \,. \tag{2.4}$$

Another general property of the variance is that $\mathrm{Var}(kX) = k^2 \mathrm{Var}(X)$ holds for $k \in \mathbb{R}$. Hence, we can describe any random Variable $X \sim \mathcal{N}(\mu, \alpha^2)$ by the scaled and shifted random variable Z

$$X = \alpha Z + \mu \quad \text{with} \quad Z \sim \mathcal{N}(0, 1) \,, \tag{2.5}$$

where Z is distributed according to the standard Gaussian with zero mean and unit variance. Gaussians with unit variance are more convenient to handle and with the relation in eq. 2.5 we can simplify any given Gaussian.

2.1.3 The Wiener process

A Wiener process $W(t)$ is a real-valued continuous-time stochastic process with stationary independent increments which is Markovian and has normally distributed increments $W(t) - W(s)$. There are many areas in which the Wiener process occurs frequently, e.g. physics, economics, finance, evolutionary biology, mathematics and engineering. In physics

it is used to describe diffusion processes like Brownian motion or other types of diffusion that can be represented by a Fokker-Planck (FP) or Langevin dynamics. A Wiener process can be characterized in many different ways. It can for example be constructed as the scaling limit of a random walk which will not be discussed in detail in this thesis.

The derivation we want to show here uses the probabilistic description of the stochastic process $X(t)$. In this context, we introduce the conditional probability density $P(x,t|x_0,t_0)$ which describes the probability to find the system in state x at time t given it was in x_0 at t_0. Additionally, we assume that $X(t)$ is Markovian. Then, the time evolution of $P(x,t|x_0,t_0)$ is given by the diffusion equation

$$\frac{\partial}{\partial t}P(x,t|x_0,t_0) = \frac{1}{2}\frac{\partial^2}{\partial x^2}P(x,t|x_0,t_0)\,, \tag{2.6}$$

which is a second order partial differential equation (PDE). We consider the boundary conditions

$$\lim_{x\to\pm\infty} P(x,t|x_0,t_0) = 0 \tag{2.7}$$

and the initial condition P_0 at $t = t_0$ is given as a Dirac δ-function at x_0

$$P(x,t_0|x_0,t_0) = P_0 = \delta(x - x_0)\,. \tag{2.8}$$

To solve eq. 2.6 with this initial condition we transfer the problem to Fourier space. In Fourier space, the second order PDE in eq. 2.6 reduces to a first order ordinary differential equation (ODE) which is easy to solve. We define the Fourier transform $\widehat{P}(s,t|s_0,t_0)$ of the probability density function $P(x,t|x_0,t_0)$ as

$$\widehat{P}(s,t|s_0,t_0) = \int_{-\infty}^{\infty} \mathrm{d}x\, \exp(isx)\, P(x,t|x_0,t_0) \qquad \text{Fourier transform} \tag{2.9}$$

$$P(x,t|x_0,t_0) = \int_{-\infty}^{\infty} \frac{\mathrm{d}s}{2\pi}\, \exp(-isx)\, \widehat{P}(s,t|s_0,t_0) \qquad \text{inverse Fourier transform}\,. \tag{2.10}$$

Hence, the Fourier transform of the FP-equation (eq. 2.6) reads

$$\frac{\partial}{\partial t}\widehat{P}(s,t|s_0,t_0) = -\frac{1}{2}s^2\widehat{P}(s,t|s_0,t_0)\,. \tag{2.11}$$

and the corresponding solution is given as

$$\begin{aligned}\widehat{P}(s,t|s_0,t_0) &= \exp\left(-\frac{1}{2}(t-t_0)s^2\right)\,\widehat{P}_0 \\ &= \exp\left(-\frac{1}{2}(t-t_0)s^2\right)\,\exp(isx_0) \\ &= \exp\left(-\left[\frac{1}{2}(t-t_0)s^2 + ix_0 s\right]\right)\,, \end{aligned} \tag{2.12}$$

where $\widehat{P}_0$ is the Fourier transform of the initial condition in eq. 2.8

$$\widehat{P}_0 = \exp(isx_0)\,. \tag{2.13}$$

We perform the inverse Fourier transformation of the solution in eq. 2.12 and get the expression

$$P(x,t|x_0,t_0) = \sqrt{\frac{1}{2\pi(t-t_0)}}\,\exp\left(-\frac{1}{2}\frac{(x-x_0)^2}{(t-t_0)}\right)\,. \tag{2.14}$$

$P(x,t|x_0,t_0)$ is a Gaussian with the properties

$$\text{1st moment (mean)} \qquad \langle X(t)\rangle = x_0 \tag{2.15}$$

$$\text{2nd moment} \qquad \left\langle X^2(t)\right\rangle = x_0^2 + (t-t_0) \tag{2.16}$$

$$\text{variance} \qquad \left\langle [X(t)-x_0]^2\right\rangle = t - t_0\,. \tag{2.17}$$

Eq. 2.14 clearly shows, that the initially sharp probability density function P_0 gets spread out over time.

Let's now focus on the Wiener process $W(t)$. Per definition, the element at time $t_0 = 0$ is given as $W(0) = 0$ and from eqs. 2.15 - 2.17 follows that

$$\text{mean} \qquad \langle W(t)\rangle = W(0) = 0 \tag{2.18}$$

$$\text{variance} \qquad \left\langle [W(t)-W(0)]^2\right\rangle = \left\langle W(t)^2\right\rangle = t\,. \tag{2.19}$$

The increments of W are independent, meaning that all future increments $W(t+\Delta t) - W(t)$ with $\Delta t \geq 0$ are independent of the past values $W(s)$ with $s \leq t$. Furthermore, the increments are distributed according to a Gaussian $W(t+\Delta t) - W(t) \sim \mathcal{N}(0, \Delta t)$ with zero mean and Δt variance. The properties of $W(t)$ are summarized in table 1.

Table 1: Properties of the Wiener process $W(t)$.

1.	$W(0) = 0$
2.	W has independent increments
3.	W has Gaussian increments $W(t + \Delta t) - W(t) \sim \mathcal{N}(0, \Delta t)$
4.	$W(t)$ is continuous in t

2.1.4 Stochastic integration

In the previous section, we introduced the Wiener process $W(t)$ as a time-continuous stochastic process. In this section, we want to extend the methods of calculus to such a stochastic process.
Before we move to stochastic calculus, let's recall the definition of a Riemann integral. Consider the time-continuous real-valued integrable function $g(t)$ on the interval $[0, \tau]$. The integral with respect to time is given by the Riemann integral

$$\begin{aligned} \int_0^\tau g(t)\,\mathrm{d}t &= \lim_{n\to\infty} \sum_{i=0}^{n-1} g(\chi_i)\delta t \\ &= \lim_{n\to\infty} \sum_{i=0}^{n-1} g(\chi_i)[t_{i+1} - t_i]\,, \end{aligned} \tag{2.20}$$

which is defined as the limit of a Riemann sum (right hand side). To construct the Riemann sum, we divide the interval $[0, \tau]$ into n sub-intervals $[t_i, t_{i+1}]$ with $i = 0, 1, \ldots, n-1$ and $t_0 = 0, t_n = \tau$. The time increment $\delta t = [t_{i+1} - t_i] = \tau/n$ is constant for all sub-intervals. Each element of the Riemann sum represents the area of a rectangle with width δt and height $g(\chi_i)$, where χ_i is chosen from the i-th sub-interval $t_i \leq \chi_i \leq t_{i+1}$. In this context, the choice of the χ_i is arbitrary and it is irrelevant if we take the left or right endpoint, the midpoint or any other point from the respective sub-interval.

To approach stochastic calculus let's start with an explicit example. Similar to eq. 2.20, we can define the integral as the limit of a sum

$$\int_0^\tau W(t)\,\mathrm{d}W(t) = \lim_{n\to\infty} \sum_{i=0}^{n-1} W(\chi_i)\left[W(t_{i+1}) - W(t_i)\right]\,. \tag{2.21}$$

Again, we divide the interval $[0, \tau]$ into n sub-intervals $[t_i, t_{i+1}]$ with $i = 0, 1, \ldots, n-1$ and $t_0 = 0, t_n = \tau$. The term $[W(t_{i+1}) - W(t_i)]$ represents the increments of the Wiener process whose properties are shown in table 1. However, in contrast to the Riemann integral in

eq. 2.20, the choice of the intermediate point χ_i with $t_i \leq \chi_i \leq t_{i+1}$ is not arbitrary.
To emphasize this fact, let's explicitly calculate the right hand side of eq. 2.21. Using the left point rule $\chi_i = t_i$ and the abbreviations $W(t_i) = W_i$ and $W(t_{i+1}) = W_{i+1}$ we get

$$\int_0^\tau W(t)\,\mathrm{d}W(t) = \lim_{n\to\infty} \sum_{i=0}^{n-1} W_i \left[W_{i+1} - W_i\right] . \tag{2.22}$$

With the identity

$$\begin{aligned} W_{i+1}^2 - W_i^2 &= (W_{i+1} - W_i)^2 + 2W_i(W_{i+1} - W_i) \\ \Leftrightarrow \quad W_i(W_{i+1} - W_i) &= \frac{1}{2}\Big((W_{i+1}^2 - W_i^2) - (W_{i+1} - W_i)^2 \Big) \end{aligned} \tag{2.23}$$

we can replace the summand as

$$\begin{aligned} \int_0^\tau W(t)\,\mathrm{d}W(t) &= \lim_{n\to\infty} \frac{1}{2} \sum_{i=0}^{n-1} \left[(W_{i+1}^2 - W_i^2) - (W_{i+1} - W_i)^2 \right] \\ &= \lim_{n\to\infty} \frac{1}{2} \left[\sum_{i=0}^{n-1} (W_{i+1}^2 - W_i^2) - \sum_{i=0}^{n-1} (W_{i+1} - W_i)^2 \right] . \end{aligned} \tag{2.24}$$

The first sum in eq. 2.24 can easily be computed as

$$\sum_{i=0}^{n-1} (W_{i+1}^2 - W_i^2) = W_n^2 - W_0^2 = W^2(t_n) - W^2(0) = W^2(\tau) . \tag{2.25}$$

According to the third point in table 1, the limit of the second sum in eq. 2.24 converges to

$$\lim_{n\to\infty} \sum_{i=0}^{n-1} (W_{i+1} - W_i)^2 = t_n = \tau . \tag{2.26}$$

Consequently, the solution of the stochastic integral in eq. 2.21 with the left point rule $\chi_i = t_i$ is given as

$$\int_0^\tau W(t)\,\mathrm{d}W(t) = \frac{1}{2}\left(W^2(\tau) - \tau\right) \tag{2.27}$$

Using the left point rule is also known as Itô calculus and eq. 2.27 is called an Itô stochastic integral. Approaching eq. 2.21 with the midpoint rule $\chi_i = t_{i+\frac{1}{2}}$ is called Stratonovich calculus and a Stratonovich stochastic integral is denoted as $\int_0^\tau \circ\, \mathrm{d}W(t)$. For the example in

eq. 2.21 we get in the Stratonovich case

$$\begin{aligned}\int_0^\tau W(t) \circ \mathrm{d}W(t) &= \lim_{n\to\infty} \sum_{i=0}^{n-1} W(t_{i+\frac{1}{2}}) \left[W(t_{i+1}) - W(t_i)\right] \\ &= \lim_{n\to\infty} \sum_{i=0}^{n-1} \frac{W(t_{i+1}) + W(t_i)}{2} \left[W(t_{i+1}) - W(t_i)\right] \\ &= \lim_{n\to\infty} \frac{1}{2} \sum_{i=0}^{n-1} \left[W^2(t_{i+1}) - W^2(t_i)\right] \\ &= \frac{1}{2} W^2(\tau)\,. \end{aligned} \tag{2.28}$$

It is apparent that Itô calculus (eq. 2.27) and Stratonovich calculus (eq. 2.28) differ from each other. From now on and throughout this thesis and all included publications, we solely use Itô calculus to compute stochastic integrals. For a more detailed discussion on the similarities and differences between Itô and Stratonovich calculus the reader is referred to Refs. [29, 101].

2.1.5 Properties of Itô stochastic integrals

Consider a smooth deterministic function $g(t)$ and a Wiener process $W(t)$. The Itô stochastic integral defined as

$$Y(\tau) \stackrel{\text{def}}{=} \int_0^\tau g(t)\, \mathrm{d}W(t) \tag{2.29}$$

represents a random variable $Y(\tau)$ which is distributed according to a Gaussian with the moments

$$\text{mean} \qquad \langle Y(\tau) \rangle = 0 \tag{2.30}$$

$$\text{variance} \qquad \left\langle Y^2(\tau) \right\rangle = \int_0^\tau g^2(t)\, \mathrm{d}t\,. \tag{2.31}$$

Eq. 2.31 represents the variance for a random variable that is given as an Itô stochastic integral and the equality holds due to the Itô isometry.[29] Additionally, we want to point out that eq. 2.29 can be seen as the solution of the stochastic differential equation

$$\mathrm{d}Y(\tau) = g(\tau)\, \mathrm{d}W(\tau) \tag{2.32}$$

given the initial condition $Y(0) = 0$.

Finally, let's consider the time interval $[0, \Delta t]$ and $g(t) = 1$ for all t. Via eq. 2.29 we can

define the random variable $X(\Delta t)$ as

$$X(\Delta t) = \int_0^{\Delta t} \mathrm{d}W(t)\,, \tag{2.33}$$

where $X(\Delta t)$ is the solution of the SDE

$$\mathrm{d}X(\Delta t) = \mathrm{d}W(\tau) \qquad \text{with} \quad X(0) = 0\,. \tag{2.34}$$

The first two moments of $X(\Delta t)$ are given as

$$\text{mean} \qquad \langle X(\Delta t)\rangle = 0 \tag{2.35}$$

$$\text{variance} \qquad \left\langle X^2(\Delta t)\right\rangle = \int_0^{\Delta t} \mathrm{d}t = \Delta t\,. \tag{2.36}$$

In general, eq. 2.33 computes the increments of the Wiener process and thus it is not surprising that eqs. 2.35 and 2.36 are equivalent to eqs. 2.18 and 2.19. As a consequence, $X(\Delta t)$ is distributed according to

$$X(\Delta t) \sim \mathcal{N}(0, \Delta t) = \sqrt{\Delta t}\mathcal{N}(0, 1)\,, \tag{2.37}$$

where the equality is explained in eq. 2.5.

2.1.6 Itô's formula

Itô's formula is used for changing variables in the stochastic case and can be interpreted as the stochastic counterpart to the chain rule. The formula can be used to convert an SDE, e.g. a Langevin equation of motion, to the corresponding Fokker-Planck equation (see section 2.3.1).

To derive the formula, we consider an arbitrary function $f(x, t)$ which is a twice differentiable scalar function. With the Taylor expansion for bivariate functions we can express $\mathrm{d}f(x, t)$ as

$$\mathrm{d}f(x,t) = \frac{\partial f(x,t)}{\partial t}\,\mathrm{d}t + \frac{\partial f(x,t)}{\partial x}\,\mathrm{d}x + \frac{1}{2}\frac{\partial^2 f(x,t)}{\partial x^2}\,(\mathrm{d}x)^2 + \cdots\,. \tag{2.38}$$

Suppose x is a stochastic process that obeys the SDE

$$\mathrm{d}x = A(x,t)\,\mathrm{d}t + B(x,t)\,\mathrm{d}W(t)\,, \tag{2.39}$$

with drift $A(x,t) \in \mathbb{R}$, Wiener process $W(t)$ and diffusion $B(x,t) \in \mathbb{R}$. Substituting $\mathrm{d}x$ in eq. 2.38 with eq. 2.39 yields

$$\begin{aligned}
\mathrm{d}f(x,t) &= \frac{\partial f(x,t)}{\partial t}\,\mathrm{d}t + \frac{\partial f(x,t)}{\partial x}\Big(A(x,t)\,\mathrm{d}t + B(x,t)\mathrm{d}W(t)\Big) \\
&\quad + \frac{1}{2}\frac{\partial^2 f(x,t)}{\partial x^2}\Big(A^2(x,t)\,(\mathrm{d}t)^2 + 2A(x,t)B(x,t)\,\mathrm{d}t\,\mathrm{d}W(t) + B^2(x,t)(\mathrm{d}W(t))^2\Big) + \cdots \\
&= \frac{\partial f(x,t)}{\partial t}\,\mathrm{d}t + \frac{\partial f(x,t)}{\partial x}A(x,t)\,\mathrm{d}t + \frac{\partial f(x,t)}{\partial x}B(x,t)\mathrm{d}W(t) \\
&\quad + \frac{1}{2}\frac{\partial^2 f(x,t)}{\partial x^2}\Big(A^2(x,t)\,(\mathrm{d}t)^2 + 2A(x,t)B(x,t)\,\mathrm{d}t\,\mathrm{d}W(t) + B^2(x,t)\mathrm{d}t\Big) + \cdots\,,
\end{aligned} \tag{2.40}$$

where we used $(\mathrm{d}W(t))^2 = \mathrm{d}t$.[101] The terms that scale with $(\mathrm{d}t)^2$ and $\mathrm{d}t\mathrm{d}W(t)$ are neglectable because they are of higher order than the last term and thus quickly vanish in the limit $\mathrm{d}t \to 0$

$$\mathrm{d}f(x,t) = \frac{\partial f(x,t)}{\partial t}\,\mathrm{d}t + \frac{\partial f(x,t)}{\partial x}A(x,t)\,\mathrm{d}t + \frac{\partial f(x,t)}{\partial x}B(x,t)\mathrm{d}W(t) + \frac{1}{2}\frac{\partial^2 f(x,t)}{\partial x^2}B^2(x,t)\mathrm{d}t\,. \tag{2.41}$$

Rearranging yields Itô's formula

$$\mathrm{d}f(x,t) = \left(\frac{\partial f(x,t)}{\partial t} + \frac{\partial f(x,t)}{\partial x}A(x,t) + \frac{1}{2}\frac{\partial^2 f(x,t)}{\partial x^2}B^2(x,t)\right)\mathrm{d}t + \frac{\partial f(x,t)}{\partial x}B(x,t)\,\mathrm{d}W(t)\,. \tag{2.42}$$

The term with the second derivative indicates that changing variables in the Itô stochastic case cannot be described with ordinary calculus.

2.1.7 The Ornstein-Uhlenbeck process

The Ornstein-Uhlenbeck process is a model to describe the momentum p of a Brownian particle with mass m under the influence of friction. Besides the friction force, the model also includes a random force (stochastic force), which is represented by a Wiener process $W(t)$. The corresponding SDE is given as

$$\begin{aligned}
\mathrm{d}q(t) &= \frac{1}{m}p(t)\,\mathrm{d}t \\
\mathrm{d}p(t) &= -\xi p(t)\mathrm{d}t + \sigma\,\mathrm{d}W(t)\,, \qquad \sigma = \sqrt{2k_B T\xi m} > 0
\end{aligned} \tag{2.43}$$

with collision rate ξ, Boltzmann constant k_B and temperature T. With the previously defined Itô stochastic integral we are able to derive an analytic solution for the Ornstein-Uhlenbeck process. First, we introduce the integration factor $\exp(\xi t)$ which has the property

$$\mathrm{d}\Big(p(t)\exp(\xi t)\Big) = \exp(\xi t)\,\mathrm{d}p(t) + \xi p(t)\exp(\xi t)\,\mathrm{d}t \tag{2.44}$$

and multiply eq. 2.43 by $\exp(\xi t)$

$$\exp(\xi t)\mathrm{d}p(t) = -\xi p(t)\exp(\xi t)\mathrm{d}t + \sigma\exp(\xi t)\,\mathrm{d}W(t)\,. \tag{2.45}$$

We can replace the term on the left hand side and the first term on the right hand side by eq. 2.44

$$\begin{aligned} &\mathrm{d}\Big(p(t)\exp(\xi t)\Big) = \sigma\exp(\xi t)\,\mathrm{d}W(t) \\ \Leftrightarrow \qquad &\exp(\xi t)p(t) = p(0) + \sigma\int_0^t \exp(\xi s)\,\mathrm{d}W(s) \end{aligned} \tag{2.46}$$

and integrate both sides to obtain

$$p(t) = \exp(-\xi t)p(0) + \sigma\exp(-\xi t)\int_0^t \exp(\xi s)\,\mathrm{d}W(s)\,, \tag{2.47}$$

where $p(0)$ represents the initial condition. The remaining integral is an Itô stochastic integral which defines the random variable $Y(t)$ (via eq. 2.29) with mean $\langle Y(t)\rangle = 0$ and variance

$$\big\langle Y^2(t)\big\rangle = \int_0^t \exp(2\xi s)\,\mathrm{d}s = \frac{\exp(2\xi t)-1}{2\xi}\,. \tag{2.48}$$

Because of eq. 2.5, we can represent $Y(t)$ by the scaled random variable $\eta(t)$

$$\int_0^t \exp(\xi s)\,\mathrm{d}W(s) = Y(t) = \sqrt{\frac{\exp(2\xi t)-1}{2\xi}}\,\eta(t) \qquad \text{with} \quad \eta(t)\sim\mathcal{N}(0,1)\,. \tag{2.49}$$

Inserting eq. 2.49 into eq. 2.47 yields

$$p(t) = \exp(-\xi t)p(0) + \sigma\exp(-\xi t)\sqrt{\frac{\exp(2\xi t)-1}{2\xi}}\,\eta(t)\,. \tag{2.50}$$

We insert $\sigma = \sqrt{2k_B T \xi m}$ and get the solution of the Ornstein-Uhlenbeck as

$$p(t) = \exp(-\xi t)p(0) + \sqrt{k_B T m \left(1 - \exp(-2\xi t)\right)}\,\eta(t), \quad \eta(t) \sim \mathcal{N}(0,1)\,. \tag{2.51}$$

with initial momentum $p(0)$.

2.2 Langevin dynamics

As previously mentioned, we can use stochastic thermostats such as the Langevin thermostat to perform MD simulations at a constant temperature. Langevin thermostats, also called Langevin dynamics, include stochastic processes that are based on the Wiener process. Consequently, the Langevin equation of motion is given as a stochastic differential equation (SDE). Additionally, there exists a high friction limit of Langevin dynamics which is called overdamped Langevin dynamics. To ensure a clear distinction between the two, we refer to Langevin dynamics as underdamped Langevin dynamics from now on. This section briefly introduces underdamped and overdamped Langevin dynamics for a one-dimensional system as well as for a system with N particles. Furthermore, we present a method to derive numerical schemes, also called integrators, that can be used to approximate the solution of the underdamped Langevin equation of motion as a time-discretized path. All content within this section is based on Refs. [19, 28, 30, 31, 101, 106, 107].

2.2.1 Underdamped Langevin dynamics

Underdamped Langevin dynamics is a mathematical model that can be used to describe the dynamics of molecular systems immersed in a fluid. The model includes friction and random forces that mimic the viscous aspect of the fluid, where the random forces arise from the numerous collisions with the surrounding solvent molecules. Additionally, a drift force governed by a potential energy function $V(x)$ can be included. The random force couples the system to an external heat bath and thus enables an exchange of energy between the system and the heat bath. Underdamped Langevin dynamics models the system in the NVT ensemble (canonical ensemble) where particle number N, volume V and temperature T are constant. Consequently, underdamped Langevin dynamics can be used as a thermostat to control the temperature of the system. Furthermore, underdamped Langevin dynamics is Markovian, reversible and ergodic.

The equation of motion for underdamped one-dimensional Langevin dynamics with time-constant friction is

$$\begin{aligned} \mathrm{d}q(t) &= \frac{1}{m} p(t)\,\mathrm{d}t \\ \mathrm{d}p(t) &= \underbrace{-\nabla_q V(q(t))\,\mathrm{d}t}_{\text{drift term}} \underbrace{-\xi p(t)\,\mathrm{d}t}_{\text{friction term}} + \underbrace{\sigma\,\mathrm{d}W(t)}_{\text{random term}}\,, \qquad \sigma = \sqrt{2k_B T \xi m}\,, \end{aligned} \tag{2.52}$$

with mass m, position $q(t)$ and momentum $p(t)$ at time t, collision rate ξ, temperature T, Boltzmann constant k_B, potential energy function $V(q(t))$, gradient $\nabla_q = \partial/\partial q$ and Wiener process $W(t)$ with properties as summarized in table 1. The solution $\omega(t) = (q(t), p(t)) \in \Omega$ of eq. 2.52, also called trajectory, fully represents the state of the system at time t, with $\Omega \subset \mathbb{R}^2$ denoting the state space. Eq. 2.52 represents an SDE as defined in eq. 2.39 with $x(t) = \omega(t)$, $A(\omega(t), t) = -(\nabla_q V(q(t)) + \xi p(t))$ and $B(\omega(t), t) = \sigma > 0$. Note, that the SDE which describes the Ornstein-Uhlenbeck process (eq. 2.43) represents an underdamped Langevin dynamics with $\nabla_q V(q(t)) = 0$ for all $q(t)$.

For a system with N particles that can move in three-dimensional Euclidean space the Langevin equation of motion is given as

$$\begin{aligned} \mathrm{d}\mathbf{q}(t) &= \mathbf{M}^{-1}\mathbf{p}(t)\,\mathrm{d}t \\ \mathrm{d}\mathbf{p}(t) &= -\nabla_{\mathbf{q}} V(\mathbf{q}(t))\,\mathrm{d}t - \xi\mathbf{p}(t)\,\mathrm{d}t + \hat{\sigma}\mathbf{M}^{\frac{1}{2}}\,\mathrm{d}\mathbf{W}(t)\,, \qquad \hat{\sigma} = \sqrt{2k_B T \xi}\,, \end{aligned} \tag{2.53}$$

with $\mathbf{q}, \mathbf{p}, \mathbf{W} \in \mathbb{R}^{3N}$, mass matrix $\mathbf{M} = \mathrm{diag}\{m_1, m_1, m_1, \ldots, m_N, m_N, m_N\} \in \mathbb{R}^{3N\times 3N}$ and gradient $\nabla_{\mathbf{q}}$. The solution of eq. 2.53 is the state-space vector $\boldsymbol{\omega}(t) = (\mathbf{q}(t), \mathbf{p}(t)) \in \Omega$ with $\Omega \subset \mathbb{R}^{6N}$.

2.2.2 The scaling factor of the random term

The random term in a underdamped Langevin equation of motion (eq. 2.52) is represented by a Wiener process which is scaled by the constant $\sigma > 0$. From a mathematical point of view, the choice of σ is arbitrary. However, if the underdamped Langevin equation of motion is supposed to accurately represent the dynamics of the system in the sense of classical mechanics, σ has to be defined accordingly. In classical mechanics, the energy of a system is evenly distributed among all degrees of freedom in thermal equilibrium.[106] In this context, the equipartition theorem

$$\left\langle E^{\mathrm{kin}} \right\rangle = \frac{1}{2} N_{\mathrm{dof}} k_B T \tag{2.54}$$

relates the temperature T of a system with N_{dof} degrees of freedom to its average kinetic energy $\langle E^{\text{kin}} \rangle$. Consequently, for each degree of freedom the average squared momentum in the long-time limit $\langle p^2 \rangle$ is

$$\left\langle p^2 \right\rangle = k_B T m \ . \tag{2.55}$$

Under this condition, the momenta have to be distributed according to the Maxwell-Boltzmann distribution

$$P(p) = \sqrt{\frac{1}{2\pi k_B T m}} \exp\left(-\frac{p^2}{2 k_B T m} \right) . \tag{2.56}$$

To understand why σ is set to $\sigma = \sqrt{2 k_B t \xi m}$, we start at eq. 2.52. The drift term can be neglected because it only depends on the positions $q(t)$ which leads us to eq. 2.43. The corresponding solution for the momenta $p(t)$ with $\sigma > 0$ is given in eq. 2.50 and rearranging yields

$$p(t) = \exp(-\xi t) p(0) + \sigma \sqrt{\frac{1 - \exp(-2\xi t)}{2\xi}} \eta(t) \quad \text{with} \quad \eta(t) \sim \mathcal{N}(0,1) \, . \tag{2.57}$$

In order to compare eq.2.57 to the equipartition theorem we compute $\langle p^2(t) \rangle$ where we average over $\eta(t)$

$$\begin{aligned} \left\langle p^2(t) \right\rangle = & \ \exp(-2\xi t) p^2(0) + 2 \exp(-\xi t) p(0) \sigma \sqrt{\frac{1 - \exp(-2\xi t)}{2\xi}} \left\langle \eta(t) \right\rangle \\ & + \sigma^2 \left(\frac{1 - \exp(-2\xi t)}{2\xi} \right) \left\langle \eta^2(t) \right\rangle \\ = & \ \exp(-2\xi t) p^2(0) + \sigma^2 \left(\frac{1 - \exp(-2\xi t)}{2\xi} \right) . \end{aligned} \tag{2.58}$$

where we used the mean $\langle \eta(t) \rangle = 1$ and the variance $\langle \eta^2(t) \rangle = 1$. Finally, we can compute the long-time limit $t \to \infty$ and define the result to be equal to the equipartition theorem

$$\lim_{t \to \infty} \left\langle p^2(t) \right\rangle = \frac{\sigma^2}{2\xi} \stackrel{\text{def}}{=} k_B T m \tag{2.59}$$

which yields $\sigma = \sqrt{2 k_B T \xi m}$. Eq. 2.59 is also known as the fluctuation-dissipation relation.

2.2.3 Langevin splitting schemes

Note, that parts of this section were taken from the supporting information of paper **A1** presented in sec. 3.1 where we summarized the concept of splitting methods for underdamped Langevin dynamics.

The solution of the underdamped Langevin equation of motion shown in eq. 2.52 is the time-continuous path or trajectory $\omega(t) = (q(t), p(t))$ where $q(t)$ denotes the position and $p(t)$ denotes the momentum at time t. Unfortunately, for most systems it is impossible to derive an analytic expression for $\omega(t)$. However, we can discretize eq. 2.52 in time and derive numerical integration schemes which are also called integrators. These integrators can be used to generate a time-discretized approximation $\omega_\tau = (\omega_0, \omega_1, \dots \omega_n)$ of $\omega(t)$ at resolution Δt where $\tau = n \cdot \Delta t$ denotes the path length. The initial state $w_0 = (q_0, p_0)$ has to be specified *a priori* as a starting point for the integrator.
There exist various different methods to derive numerical integration schemes for problems as shown in eq. 2.52 and hence, a huge variety of underdamped Langevin integrators[19,30–45] has been reported. Here, we focus on splitting methods which split the vector field in the Langevin equation of motion into three parts labeled A, B and O

$$\mathrm{d}\begin{pmatrix} q(t) \\ p(t) \end{pmatrix} = \underbrace{\begin{pmatrix} \frac{p(t)}{m} \\ 0 \end{pmatrix} \mathrm{d}t}_{A} + \underbrace{\begin{pmatrix} 0 \\ -\nabla_q V(q(t)) \end{pmatrix} \mathrm{d}t}_{B} + \underbrace{\begin{pmatrix} 0 \\ -\xi p(t) \mathrm{d}t + \sigma \mathrm{d}W(t) \end{pmatrix}}_{O} . \tag{2.60}$$

with $\sigma = \sqrt{2 k_B T \xi m}$. Part A and B yield PDEs while part O represents the SDE of the Ornstein-Uhlenbeck process (eq. 2.43).

$$A: \quad \mathrm{d}q(t) = \frac{p(t)}{m} \mathrm{d}t \tag{2.61a}$$

$$B: \quad \mathrm{d}p(t) = -\nabla_q V(q(t)) \, \mathrm{d}t \tag{2.61b}$$

$$O: \quad \mathrm{d}p(t) = -\xi p(t) \, \mathrm{d}t + \sigma \, \mathrm{d}W(t) \, . \tag{2.61c}$$

With the initial condition $(q_k, p_k)^\top$, each of the three parts can be solved separately to yield the update operators $\mathcal{A}, \mathcal{B}$ and $\mathcal{O}$ which act on the discrete state $(q_k, p_k)^\top$ at iteration step k

$$\mathcal{A}\begin{pmatrix} q_k \\ p_k \end{pmatrix} = \begin{pmatrix} q_k + \frac{\Delta t}{m} p_k \\ p_k \end{pmatrix} \tag{2.62a}$$

$$\mathcal{B}\begin{pmatrix} q_k \\ p_k \end{pmatrix} = \begin{pmatrix} q_k \\ p_k - \Delta t \, \nabla_q V(q_k) \end{pmatrix} \tag{2.62b}$$

$$\mathcal{O}\begin{pmatrix} q_k \\ p_k \end{pmatrix} = \begin{pmatrix} q_k \\ e^{-\xi \Delta t} p_k + \sqrt{k_B T m (1 - e^{-2\xi \Delta t})}\, \eta_k \end{pmatrix} \quad \text{with} \quad \eta_k \sim \mathcal{N}(0,1)\,. \tag{2.62c}$$

Eq. 2.62c is the solution of the Ornstein-Uhlenbeck process as reported in eq. 2.51 and η_k denotes a random number which is drawn from a standard Gaussian distribution. Update operator $\mathcal{A}$ is a deterministic update in position space. Update operators $\mathcal{B}$ and $\mathcal{O}$ represent updates in momentum space with $\mathcal{B}$ being deterministic and $\mathcal{O}$ being stochastic. Eqs. 2.62a - 2.62c can be used to derive a variety of integrators by applying different sequences of $\mathcal{A}, \mathcal{B}$ and $\mathcal{O}$ to perform a full time step update $(q_k, p_k)^\top \to (q_{k+1}, p_{k+1})^\top$.

The ABO integrator

To illustrate the concept let's consider the ABO integrator. The name of the integrator denotes the update sequence in a left-to-right fashion. Thus, we first apply the update operator $\mathcal{A}$, then $\mathcal{B}$ and lastly $\mathcal{O}$ to perform a full time step update

$$\begin{pmatrix} q_{k+1} \\ p_{k+1} \end{pmatrix} = \mathcal{O}\mathcal{B}\mathcal{A}\begin{pmatrix} q_k \\ p_k \end{pmatrix}. \tag{2.63}$$

When we work with update operators as in eq. 2.63, the operator sequence is given in a right-to-left fashion because the operator to the very right is the first to act on the state of the system. With eqs. 2.62a - 2.62c the integrator equations for the ABO integrator are

$$q_{k+1} = q_k + \frac{\Delta t}{m} p_k \tag{2.64a}$$

$$p_{k+1/2} = p_k - \Delta t \, \nabla_q V(q_{k+1}) \tag{2.64b}$$

$$p_{k+1} = e^{-\xi \Delta t} p_{k+1/2} + \sqrt{k_B T m (1 - e^{-2\xi \Delta t})}\, \eta_k\,. \tag{2.64c}$$

Please note, that the subscript 1/2 is solely used to enumerate the intermediate steps for the momentum update and have no relation to intermediate physical time. That is, read $p_{k+1/2}$ as "1 of 2 intermediate steps completed". Other possible splitting sequences are AOB, BAO, BOA, OAB and OBA. All integrators of this type are called first order accurate, because the corresponding error of the splitting scheme is $\mathcal{O}(\Delta t^2)$.[19,30,31,108]

We can also construct update sequences in which one or more operators occur twice. In this case, the respective update operator is carried out for half a time step $\frac{\Delta t}{2}$, and we denote the corresponding operators with a prime

$$\mathcal{A}' \begin{pmatrix} q_k \\ p_k \end{pmatrix} = \begin{pmatrix} q_k + \frac{\Delta t}{2m} p_k \\ p_k \end{pmatrix} \tag{2.65a}$$

$$\mathcal{B}' \begin{pmatrix} q_k \\ p_k \end{pmatrix} = \begin{pmatrix} q_k \\ p_k - \frac{\Delta t}{2} \nabla_q V(q_k) \end{pmatrix} \tag{2.65b}$$

$$\mathcal{O}' \begin{pmatrix} q_k \\ p_k \end{pmatrix} = \begin{pmatrix} q_k \\ e^{-\xi \frac{\Delta t}{2}} p_k + \sqrt{k_B T m (1 - e^{-\xi \Delta t})}\, \eta_k \end{pmatrix} \quad \text{with} \quad \eta_k \sim \mathcal{N}(0,1)\,. \tag{2.65c}$$

Eqs. 2.65a - 2.65c are for example relevant for the symmetric integrators ABOBA, BAOAB, AOBOA, BOAOB, OBABO and OABAO or for the non-symmetric integrator BAOA.[19,30,31] Note, that OBABO is also called also called Bussi-Parrinello thermostat.[34]

The ABOBA integrator

To illustrate the concept of half time step updates, let's consider the ABOBA integrator. Here, we have the two consecutive half time step updates A' and B', then the full time step update O and finally the two consecutive half time step updates B' and A'

$$\begin{pmatrix} q_{k+1} \\ p_{k+1} \end{pmatrix} = \mathcal{A}'\mathcal{B}'\mathcal{O}\mathcal{B}'\mathcal{A}' \begin{pmatrix} q_k \\ p_k \end{pmatrix}. \tag{2.66}$$

Note, that the prime notation is omitted in the integrator name. With eqs. 2.62a - 2.62c and 2.65a - 2.65c the integrator equations for the ABOBA integrator are given as

$$q_{k+1/2} = q_k + \frac{\Delta t}{2m} p_k \tag{2.67a}$$

$$p_{k+1/3} = p_k - \frac{\Delta t}{2} \nabla_q V(q_{k+1/2}) \tag{2.67b}$$

$$p_{k+2/3} = e^{-\xi \Delta t} p_{k+1/3} + \sqrt{k_B T m (1 - e^{-2\xi \Delta t})}\, \eta_k \quad \text{with} \quad \eta_k \sim \mathcal{N}(0,1) \tag{2.67c}$$

$$p_{k+1} = p_{k+2/3} - \frac{\Delta t}{2} \nabla_q V(q_{k+1/2}) \tag{2.67d}$$

$$q_{k+1} = q_{k+1/2} + \frac{\Delta t}{2m} p_{k+1}\,. \tag{2.67e}$$

Again, the subscript $1/2, 1/3$ and $2/3$ are solely used to enumerate the intermediate steps for the position and momentum updates and have no relation to intermediate physical time. This kind of splitting is called Strang splitting[109] which is second order accurate. The error of strang splitting is of order $\mathcal{O}(\Delta t^3)$.[19,30,31,108] The integrator equations for BAOAB, AOBOA, BOAOB, OBABO and OABAO and BAOA are summarized in Appendix A.2.

Finally, we want to mention that the splitting methods presented in this section can also be used to approximate the solution of high-dimensional underdamped Langevin dynamics as defined in eq. 2.53. In this case, the full time step position and momentum updates of each degree of freedom are defined by the respective integrator equations. For example, to update the state $(\mathbf{x}_k, \mathbf{p}_k)^\top \to (\mathbf{x}_{k+1}, \mathbf{p}_{k+1})^\top$ with $\mathbf{q}_k, \mathbf{p}_k \in \mathbb{R}^{3N}$ using the ABOBA integrator, each degree of freedom $(q_k^{(i)}, p_k^{(i)})^\top$ is updated according to eqs. 2.67a - 2.67e, where $i = 1, 2, \ldots, 3N$.

2.2.4 Overdamped Langevin dynamics

Overdamped Langevin dynamics, also called Brownian dynamics is the high friction limit $\frac{\mathrm{d}p}{\mathrm{d}t} \ll \xi p$ of eq. 2.52 and thus a special case of Langevin dynamics. The assumption is that due to high friction the acceleration can be neglected and we can set $\mathrm{d}p = 0$ in eq. 2.52

$$m\,\mathrm{d}q(t) = p(t)\,\mathrm{d}t \tag{2.68}$$

$$0 = -\nabla_q V(q(t))\,\mathrm{d}t - \xi p(t)\mathrm{d}t + \sqrt{2k_B T \xi m}\,\mathrm{d}W(t)\,. \tag{2.69}$$

We merge eqs. 2.68 and 2.69, rearrange and get the SDE for overdamped Langevin dynamics

$$\mathrm{d}q(t) = -\frac{\nabla_q V(q(t))}{\xi m}\,\mathrm{d}t + \widehat{\sigma}\,\mathrm{d}W(t)\,, \qquad \widehat{\sigma} = \sqrt{\frac{2k_B T}{\xi m}}\,, \tag{2.70}$$

with mass m, time t, position $q(t)$, momentum $p(t)$, collision rate ξ, temperature T, Boltzmann constant k_B, potential energy function $V(q(t))$, gradient $\nabla_q = \partial/\partial q$ and Wiener process $W(t)$ with properties as summarized in table 1. The solution $\omega(t) = q(t) \in \Gamma$ of eq. 2.70 fully describes the state of the system at time t where $\Gamma \subset \mathbb{R}$ denotes the configuration space. Overdamped Langevin dynamics is Markovian, reversible and ergodic.

For a system with N particles that evolves in three-dimensional Euclidean space according to overdamped Langevin dynamics we get

$$\mathrm{d}\mathbf{q}(t) = -\frac{\nabla_{\mathbf{q}} V(\mathbf{q}(t))}{\xi}\mathbf{M}^{-1}\,\mathrm{d}t + \bar{\sigma}\mathbf{M}^{-\frac{1}{2}}\,\mathrm{d}W(t)\,, \qquad \bar{\sigma} = \sqrt{\frac{2k_B T}{\xi}}\,, \tag{2.71}$$

with $\mathbf{q}, \mathbf{W} \in \mathbb{R}^{3N}$, mass matrix $\mathbf{M} = \mathrm{diag}\{m_1, m_1, m_1, \ldots, m_N, m_N, m_N\} \in \mathbb{R}^{3N\times 3N}$ and gradient $\nabla_{\mathbf{q}}$. The solution of eq. 2.71 is $\mathbf{q}(t) \in \Gamma$ with $\Gamma \subset \mathbb{R}^{3N}$.

2.2.5 The Euler-Maruyama method

The Euler-Maruyama method is the extension of the Euler method to SDEs and can for example be used to solve the equation of motion for overdamped Langevin dynamics. The Euler-Maruyama integrator approximates the true solution $\omega(t) = q(t)$ of eq. 2.70 as the time discretized path $\omega_\tau = (\omega_0, \omega_1, \ldots, \omega_n) = (q_0, q_1, \ldots, q_n)$ with resolution Δt where $\tau = n \cdot \Delta t$ denotes the path length. The initial position q_0 must be specified *a priori* as a starting point for the integrator.
The Euler-Maruyama integrator is given as

$$q_{k+1} = q_k - \frac{\nabla_q V(q_k)}{\xi}\,\Delta t + \sqrt{\frac{2k_BT}{\xi m}}\,\eta_k \quad \text{with} \quad \eta_k \sim \mathcal{N}(0,1)\,, \tag{2.72}$$

where the random number η_k at iteration step k is drawn from a standard Gaussian.

For a system with N particles that evolves according to eq. 2.71 in three-dimensional Euclidean space, the Euler-Maruyama integrator is given as

$$q_{k+1}^{(i)} = q_k^{(i)} - \frac{\nabla_q V(q_k^{(i)})}{\xi}\,\Delta t + \sqrt{\frac{2k_BT}{\xi m^{(i)}}}\,\eta_k^{(i)} \quad \text{with} \quad \eta_k^{(i)} \sim \mathcal{N}(0,1)\,, \tag{2.73}$$

where $m^{(i)}, q_k^{(i)}$ and $\eta_k^{(i)}$ denote the mass, position and momentum of the i-th degree of freedom with $\mathbf{q}_k \in \mathbb{R}^{3N}$ and $\boldsymbol{\eta}_k \in \mathbb{R}^{3N}$.

2.3 The Fokker-Planck equation

Consider the stochastic process $X(t)$ whose time evolution can be described with an SDE, e.g. as a Langevin dynamics, that includes a drift, diffusion and friction. The corresponding Fokker-Planck (FP) equation is a partial differential equation (PDE) which governs $X(t)$ as the time evolution of the probability density function $P(x, t|x_0, t_0)$ associated with $X(t)$. In other words, we can approach a given diffusion process via two different points of view and either study an SDE (Langevin picture) or a PDE (FP-picture). Depending on the research question, one of the two approaches may be more effective than the other. However, it is reasonable to study both points of view to grasp the full picture of the nature of diffusion processes.
This section briefly establishes the connection between SDEs (Langevin dynamics) and FP-equations for a one-dimensional system and introduces the FP-equation for overdamped Langevin dynamics and the Ornstein-Uhlenbeck process. For the sake of completeness, we show the high dimensional versions of the respective FP-equations but refer to Refs. [29, 101]

for a detailed derivation. All content within this section is based on Refs. [19, 29, 101, 102].

2.3.1 From stochastic differential equations to Fokker-Planck equations

Consider the random process $X(t)$ defined by the SDE in eq. 2.39

$$\mathrm{d}x = A(x,t)\,\mathrm{d}t + \sigma\,\mathrm{d}W(t)\,, \tag{2.74}$$

with drift $A(x,t)$ and constant diffusion term $B(x,t) = \sigma > 0$. Additionally, we assume that $X(t)$ has the conditional probability density function $P(x,t|x_0,t_0)$. With Itô's formula (eq. 2.42) we can write

$$\mathrm{d}f(x,t) = \left(\frac{\partial f(x,t)}{\partial t} + \frac{\partial f(x,t)}{\partial x}A(x,t) + \frac{\sigma^2}{2}\frac{\partial^2 f(x,t)}{\partial x^2}\right)\mathrm{d}t + \sigma\frac{\partial f(x,t)}{\partial x}\,\mathrm{d}W(t)\,, \tag{2.75}$$

for the arbitrary twice differentiable function $f(x,t)$ and Wiener process $W(t)$. Taking the expectation yields

$$\left\langle \mathrm{d}f(x,t)\right\rangle = \left(\left\langle\frac{\partial f(x,t)}{\partial t}\right\rangle + \left\langle\frac{\partial f(x,t)}{\partial x}A(x,t)\right\rangle + \frac{\sigma^2}{2}\left\langle\frac{\partial^2 f(x,t)}{\partial x^2}\right\rangle\right)\mathrm{d}t\,, \tag{2.76}$$

where we used we used that $\partial f/\partial x$ and $\mathrm{d}W(t)$ are uncorrelated[29,101]

$$\left\langle\sigma\frac{\partial f(x,t)}{\partial x}\,\mathrm{d}W(t)\right\rangle = \sigma\left\langle\frac{\partial f(x,t)}{\partial x}\right\rangle\left\langle\mathrm{d}W(t)\right\rangle = 0\,. \tag{2.77}$$

We divide eq. 2.76 by $\mathrm{d}t$ and get

$$\begin{aligned}\frac{\langle\mathrm{d}f(x,t)\rangle}{\mathrm{d}t} &= \frac{\mathrm{d}}{\mathrm{d}t}\left\langle f(x,t)\right\rangle \\ &= \left\langle\frac{\partial f(x,t)}{\partial t}\right\rangle + \left\langle\frac{\partial f(x,t)}{\partial x}A(x,t)\right\rangle + \frac{\sigma^2}{2}\left\langle\frac{\partial^2 f(x,t)}{\partial x^2}\right\rangle\,.\end{aligned} \tag{2.78}$$

With the boundary conditions

$$\begin{aligned}\lim_{x\to\pm\infty} P(x,t|x_0,t_0) &= 0 \\ \lim_{x\to\pm\infty}\frac{\partial P(x,t|x_0,t_0)}{\partial x} &= 0\end{aligned} \tag{2.79}$$

we can compute the expected values in eq. 2.78 as

$$\frac{\mathrm{d}}{\mathrm{d}t}\int \mathrm{d}x\, f(x,t)P(x,t|x_0,t_0) = \int \mathrm{d}x\, \frac{\partial f(x,t)}{\partial t}P(x,t|x_0,t_0) + \int \mathrm{d}x\, \frac{\partial f(x,t)}{\partial x}A(x,t)P(x,t|x_0,t_0) + \frac{\sigma^2}{2}\int \mathrm{d}x\, \frac{\partial^2 f(x,t)}{\partial x^2}P(x,t|x_0,t_0)\,, \tag{2.80}$$

where we abbreviated $\int \mathrm{d}x = \int_{-\infty}^{\infty} \mathrm{d}x$. On the other hand, the product rule gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\int \mathrm{d}x\, f(x,t)P(x,t|x_0,t_0) = \int \mathrm{d}x\, \frac{\partial f(x,t)}{\partial t}P(x,t|x_0,t_0) + \int \mathrm{d}x\, f(x,t)\frac{\partial P(x,t|x_0,t_0)}{\partial t} \tag{2.81}$$

and setting eq. 2.81 equal to eq. 2.80 yields

$$\begin{aligned}
&\int \mathrm{d}x\, f(x,t)\frac{\partial P(x,t|x_0,t_0)}{\partial t} \\
&= \int \mathrm{d}x\, \frac{\partial f(x,t)}{\partial x}A(x,t)P(x,t|x_0,t_0) + \frac{\sigma^2}{2}\int \mathrm{d}x\, \frac{\partial^2 f(x,t)}{\partial x^2}P(x,t|x_0,t_0) && (2.82)\\
&= -\int \mathrm{d}x\, f(x,t)\frac{\partial}{\partial x}\Big(A(x,t)P(x,t|x_0,t_0)\Big) + \frac{\sigma^2}{2}\int \mathrm{d}x\, f(x,t)\frac{\partial^2}{\partial x^2}P(x,t|x_0,t_0)\,. && (2.83)
\end{aligned}$$

In the last step, we integrated by parts as shown in detail in appendix A.1. Eq. 2.81 has to hold for all integrable functions $f(x,t)$ and thus

$$\frac{\partial P(x,t|x_0,t_0)}{\partial t} = -\frac{\partial}{\partial x}\Big(A(x,t)P(x,t|x_0,t_0)\Big) + \frac{\sigma^2}{2}\frac{\partial^2}{\partial x^2}P(x,t|x_0,t_0)\,, \tag{2.84}$$

which is the FP-equation that corresponds to the SDE in eq. 2.74.

Equivalently, we can derive the FP-equation for a system with N particles

$$\frac{\partial}{\partial t}P(\mathbf{x},t|\widetilde{\mathbf{x}},t_0) = -\sum_{i=1}^{3N}\frac{\partial}{\partial x_i}\Big(A(\mathbf{x},t)P(\mathbf{x},t|\widetilde{\mathbf{x}},t_0)\Big) + \frac{\sigma^2}{2}\sum_{i=1}^{3N}\sum_{j=1}^{3N}\frac{\partial^2}{\partial x_i \partial x_j}P(\mathbf{x},t|\widetilde{\mathbf{x}},t_0)\,, \tag{2.85}$$

where $P(\mathbf{x},t|\widetilde{\mathbf{x}},t_0)$ denotes the conditional probability density function to find the system in state $\mathbf{x} = (x_1, x_2, \ldots, x_{3N})^\top$ at time t given it was in $\widetilde{\mathbf{x}}$ at time t_0. For a detailed derivation of eq. 2.85 the reader is referred to Ref. [101].

2.3.2 The Smoluchowski equation

Consider the SDE for overdamped Langevin dynamics (Brownian dynamics) defined in eq. 2.74 with $x = q$, $A(x,t) = -\nabla_q V(q)/(\xi m)$ and $\sigma = \widehat{\sigma}$. Consequently, we can use eq. 2.84 to derive the FP-equation that corresponds to overdamped Langevin dynamics

$$\frac{\partial}{\partial t}P(q,t|q_0,t_0) = \frac{\partial}{\partial q}\Big(\frac{\nabla_q V(q(t))}{\xi m}\,P(q,t|q_0,t_0)\Big) + \frac{\widehat{\sigma}^2}{2}\frac{\partial^2}{\partial q^2}P(q,t|q_0,t_0) \tag{2.86}$$

with $\widehat{\sigma} = \sqrt{(2k_BT)/(\xi m)}$ and the conditional probability density function $P(q,t|q_0,t_0)$ to find the system at position q at time t given it was at q_0 at t_0. Eq. 2.86 is also called Smoluchowski equation.

For a system with N particles that moves in three-dimensional Euclidean space, the Smoluchowski equation is given as

$$\frac{\partial}{\partial t}P(\mathbf{q},t|\widetilde{\mathbf{q}},t_0) = \sum_{i=1}^{3N}\frac{\partial}{\partial q_i}\Big(\frac{\nabla_q V(\mathbf{q}(t))}{\xi}\mathbf{M}^{-1}\,P(\mathbf{q},t|\widetilde{\mathbf{q}},t_0)\Big) + \frac{\widehat{\sigma}^2}{2}\sum_{i=1}^{3N}\sum_{j=1}^{3N}\frac{\partial^2}{\partial q_i\partial q_j}P(\mathbf{q},t|\widetilde{\mathbf{q}},t_0)\,, \tag{2.87}$$

where $P(\mathbf{q},t|\widetilde{\mathbf{q}},t_0)$ denotes the conditional probability density function to find the system at position $\mathbf{q} = (q_1, q_2, \ldots, q_{3N})^\top$ at time t given it was in $\widetilde{\mathbf{q}}$ at time t_0 and $\widehat{\sigma}$ as defined above.

2.3.3 The Fokker-Planck equation of the Ornstein-Uhlenbeck process

The SDE in eq. 2.43 describes the Ornstein-Uhlenbeck process in one dimension. Replacing the variable $x = p(t)$ and $A(p,t) = -\xi p(t)$ in eq. 2.84 yields the FP-equation for the Ornstein-Uhlenbeck process

$$\frac{\partial}{\partial t}P(p,t|p_0,t_0) = \frac{\partial}{\partial p}\Big(\xi p\cdot P(p,t|p_0,t_0)\Big) + \frac{\sigma^2}{2}\frac{\partial^2}{\partial p^2}\,P(p,t|p_0,t_0)\,, \tag{2.88}$$

with $\sigma = \sqrt{2k_BT\xi m}$ and $P(p,t|p_0,t_0)$ being the conditional probability to find the particle with momentum p at time t given it had the momentum p_0 at time t_0. The solution of eq. 2.88

is the Gaussian

$$P(p,t|p_0,t_0) = \sqrt{\frac{\xi}{\pi\sigma^2\big(1-\exp(-2\xi(t-t_0))\big)}} \cdot \exp\left(-\frac{\xi}{\sigma^2}\frac{\Big(p-p_0\exp(-\xi(t-t_0))\Big)^2}{1-\exp(-2\xi(t-t_0))}\right). \tag{2.89}$$

that converges to

$$\lim_{t\to\infty} P(p,t|p_0,t_0) = \sqrt{\frac{\xi}{\pi\sigma^2}}\exp\left(-\frac{\xi}{\sigma^2}p^2\right) \tag{2.90}$$

in the long-time limit. With $\sigma = \sqrt{2k_BT\xi m}$, eq. 2.90 represents the Maxwell-Boltzmann distribution defined in eq. 2.56. For a more-dimensional treatment of the Ornstein-Uhlenbeck process, the reader is referred to Refs. [29, 101].

2.4 Path probabilities and path expected values

Stochastic integrators, such as underdamped or overdamped Langevin integrators, generate time-discretized paths with a probability between zero and one. This probability is called time-discretized path probability. As illustrated in fig. 2.1a, there are two different routes which can be taken to derive the corresponding mathematical expression given that we start at the equation of motion. Following route 1, we discretize the equation of motion in time and construct numerical integrators (MD algorithms) as explained in sec. 2.2.3. Based on the integrator equations, we can subsequently derive an expression for the time-discretized path probability by averaging-out the stochastic part. Depending on the chosen integrator, this approach yields slightly different expressions for the path probability. Paper **A3** in sec. 3.3 demonstrates the second step of route 1 using a simplified version of the underdamped Langevin integrator developed by Izaguirre *et al.*[39] as an example.
Following route 2, we stay in the time-continuous picture and formulate the time-continuous path probability as a path integral. The path integral can subsequently be time-discretized to yield an expression for the time-discretized path probability. However, the path integral formalism is not trivial and requires fundamental knowledge about the Wiener measure. For a detailed discussion on this topic, the reader is referred to Refs. [102, 103, 110].
This section introduces the concept of time-discretized path probabilities for underdamped Langevin dynamics and briefly discusses path ensemble averages.

Consider the time-discretized path $\boldsymbol{\omega} = (\omega_0, \omega_1, \ldots, \omega_n) = ((q_0,p_0),(q_1,p_1),\ldots,(q_n,p_n))$ generated by underdamped Langevin dynamics (eq. 2.52) with time step Δt. Here, n denotes

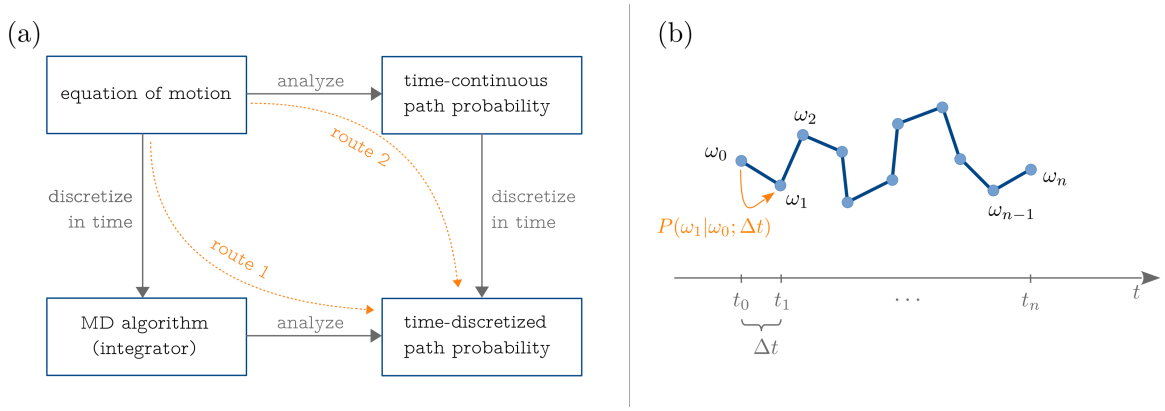


Figure 2.1:
(a) Schematic representation of how time-discretized path probabilities can be derived from an equation of motion.
(b) Schematic representation of a time-discretized path $\boldsymbol{\omega} = (\omega_0, \omega_1, \ldots, \omega_n)$ of length n with $\omega_k \in \mathbb{R}^2$. The time t_k is given as $t_k = k\Delta t$ with $k = 0, 1, \ldots, n$. $P(\omega_1|\omega_0; \Delta t)$ represents the single-step probability to observe the step $\omega_0 \to \omega_1$ in time Δt.

the number of time steps and $\omega_k \in \Omega$ represents the state at iteration step k with state space $\Omega \in \mathbb{R}^2$. A schematic representation of $\boldsymbol{\omega}$ is shown in fig. 2.1b. The dynamics is governed by the potential energy function $V(q)$ and the path $\boldsymbol{\omega} \in \mathcal{S}$ of length $n+1$ is an element of the path space $\mathcal{S} \in \Omega^{n+1}$. The equilibrium distribution $P(\omega)$ of the state ω is given as

$$P_\pi(\omega) = P_\pi(q,p) = \frac{1}{Z} \exp\left(-\frac{V(q)}{k_B T}\right) \cdot \sqrt{\frac{1}{2\pi k_B T m}} \exp\left(-\frac{p^2}{2k_B T m}\right), \tag{2.91}$$

with partition function $Z = \int_{-\infty}^{\infty} \mathrm{d}q \exp(-V(q)/(k_B T))$, temperature T, Boltzmann constant k_B and mass m. $P_\pi(\omega)$ is associate with the state space probability measure

$$\pi(A) = \int_A \mathrm{d}\omega P_\pi(\omega) \qquad \forall A \subset \Omega\,, \tag{2.92}$$

where $\pi(A)$ describes the probability to find the system in subset A of state space Ω. The probability $P(\boldsymbol{\omega})$ to observe the particular path $\boldsymbol{\omega}$ is

$$\begin{aligned} P(\boldsymbol{\omega}) &= P(\omega_0, \omega_1, \ldots, \omega_n) \\ &= P_\pi(\omega_0)\, P_\mu(\omega_1, \ldots, \omega_n|\omega_0) \\ &= P_\pi(\omega_0) \cdot \prod_{k=1}^{n-1} P(\omega_{k+1}|\omega_k; \Delta t)\,, \end{aligned} \tag{2.93}$$

with $P_\pi(\omega_0)$ being the equilibrium distribution of the initial state ω_0 as defined in eq. 2.91. $P(\omega_1, \ldots, \omega_n|\omega_0)$ is the conditional probability to observe the particular path $\boldsymbol{\omega}$ conditioned to the initial state ω_0 and $P(\omega_{k+1}|\omega_k; \Delta t)$ denotes the single-step probability to observe the

step $\omega_k \to \omega_{k+1}$ in time Δt. The equality in eq. 2.93 holds because $\boldsymbol{\omega}$ is Markovian. Note, that the functional form of the single-step probability depends on the chosen integrator and on the potential energy function. The conditional path probability $P_\mu(\omega_1, \ldots, \omega_n|\omega_0)$ is associated with the path probability measure

$$\mu(\mathcal{A}) = \int_{A_1} \mathrm{d}\omega_1 \int_{A_2} \mathrm{d}\omega_2 \cdots \int_{A_n} \mathrm{d}\omega_n \, P_\mu(\omega_1, \ldots, \omega_n|\omega_0) \,, \tag{2.94}$$

where $\mu(\mathcal{A})$ describes the probability to find the system in subset $\mathcal{A} = A_1 \times A_2 \times \cdots \times A_n \subset \Omega^n$ with $A_i \subset \Omega$.

Next, we introduce the path observable $o : \mathcal{S} \to \mathbb{R}$ and the corresponding path expected value

$$\langle o \rangle = \int_\Omega \mathrm{d}\omega_0 \int_\Omega \mathrm{d}\omega_1 \cdots \int_\Omega \mathrm{d}\omega_n \, o(\boldsymbol{\omega}) \, P(\boldsymbol{\omega}) \tag{2.95a}$$

$$= \lim_{s\to\infty} \frac{1}{s} \sum_{i=1}^{s} o(\boldsymbol{\omega}^{(i)}) \,. \tag{2.95b}$$

which can be estimated from a set of s paths $\{\boldsymbol{\omega}^{(1)}, \boldsymbol{\omega}^{(2)}, \ldots, \boldsymbol{\omega}^{(s)}\}$ generated at the same conditions as described above. The equality in eq. 2.95b holds because underdamped Langevin dynamics is ergodic.

Another path expected value is the time-correlation function $C(\tau)$ of the integrable function $f : \Omega \to \mathbb{R}$ which maps a discrete state ω_i of the path $\boldsymbol{\omega}$ to a real number

$$\begin{aligned} C(\tau) &= \int_\Omega \mathrm{d}\omega_0 \int_\Omega \mathrm{d}\omega_1 \cdots \int_\Omega \mathrm{d}\omega_n \, f(\omega_0) \, f(\omega_n) P(\boldsymbol{\omega}) \\ &= \lim_{s\to\infty} \frac{1}{s} \sum_{i=1}^{s} f(\omega_0^{(i)}) \, f(\omega_n^{(i)}) \end{aligned} \tag{2.96}$$

where $\tau = n\Delta t$ denotes the lag time. Since underdamped Langevin dynamics is ergodic, $C(\tau)$ can also be estimated from a set of paths a set of s paths $\{\boldsymbol{\omega}^{(1)}, \boldsymbol{\omega}^{(2)}, \ldots, \boldsymbol{\omega}^{(s)}\}$ generated at the same conditions as described above with $\boldsymbol{\omega}^{(i)} = (\omega_0^{(i)}, \omega_1^{(i)}, \ldots, \omega_n^{(i)})$.

2.5 Path reweighting

Path reweighting is a technique that can be used to compute dynamical quantities of unbiased dynamics from a set of path generated at a biased dynamics. In the case of a stochastic dynamics, such as underdamped Langevin dynamics, the Girsanov theorem[111] can be exploited to realize reweighting in path space. Hence, path reweighting is sometimes called Girsanov

reweighting.[73–75] In this section, we briefly introduce the concept of path reweighting for underdamped Langevin dynamics in one-dimensional Euclidean space. Transferring the concept to a high-dimensional space is straightforward.[73,74] Here, the biased dynamics is called simulation system and the unbiased dynamics are called target system. All content of this section is based on Refs. [72–74]

Consider a particle that moves according to underdamped Langevin dynamics (eq. 2.52) in one-dimensional Euclidean space. In the simulation system, the particle moves in the simulation potential $V(q)$. In the target system, the particle moves in a perturbed version of the simulation potential which we call target potential

$$\widetilde{V}(q) = V(q) + U(q)\,, \tag{2.97}$$

with $U(q)$ denoting the bias. Additionally, consider the time-discretized path $\boldsymbol{\omega} = (\omega_0, \omega_1, \ldots, \omega_n)$ with $\omega_k \in \Omega$, where $\Omega \subset \mathbb{R}^2$ denotes the state space. The probability $P(\boldsymbol{\omega})$ to observe this specific $\boldsymbol{\omega}$ at the simulation potential is defined in eq. 2.93. Similarly, the path probability $\widetilde{P}(\boldsymbol{\omega})$ to observe $\boldsymbol{\omega}$ at the target potential is given as

$$\widetilde{P}(\boldsymbol{\omega}) = \widetilde{P_\pi}(\omega_0)\, \widetilde{P_\mu}(\omega_1, \ldots, \omega_n|\omega_0)\,. \tag{2.98}$$

Here, the equilibrium distribution $\widetilde{P_\pi}(\omega_0)$

$$\widetilde{P_\pi}(\omega) = \widetilde{P_\pi}(q,p) = \frac{1}{\widetilde{Z}} \exp\left(-\frac{\widetilde{V}(q)}{k_B T}\right) \cdot \sqrt{\frac{1}{2\pi k_B T m}} \exp\left(-\frac{p^2}{2k_B T m}\right)\,, \tag{2.99}$$

with partition function $\widetilde{Z} = \int_{-\infty}^{\infty} \mathrm{d}q \exp(-\widetilde{V}(q)/(k_B T))$ is associated with the state space probability measure at the target potential

$$\widetilde{\pi}(A) = \int_A \mathrm{d}\omega \widetilde{P_\pi}(\omega) \qquad \forall A \subset \Omega\,. \tag{2.100}$$

The conditional probability $\widetilde{P_\mu}(\omega_1, \ldots, \omega_n|\omega_0)$ is associated with the path probability measure at the target potential

$$\widetilde{\mu}(\mathcal{A}) = \int_{A_1} \mathrm{d}\omega_1 \int_{A_2} \mathrm{d}\omega_2 \cdots \int_{A_n} \mathrm{d}\omega_n\, \widetilde{P_\mu}(\omega_1, \ldots, \omega_n|\omega_0)\,. \tag{2.101}$$

The core concept of path reweighting is, to compare the probability $\widetilde{P}(\boldsymbol{\omega})$ to observe $\boldsymbol{\omega}$ at the target potential to the probability $P(\boldsymbol{\omega})$ to observe the exact same path at the simulation

potential

$$\begin{aligned} W(\boldsymbol{\omega}) &= \frac{\widetilde{P}(\boldsymbol{\omega})}{P(\boldsymbol{\omega})} \\ &= \frac{\widetilde{P_\pi}(\omega_0)}{P_\pi(\omega_0)} \cdot \frac{\widetilde{P_\mu}(\omega_1, \omega_2, \ldots, \omega_n|\omega_0)}{P_\mu(\omega_1, \omega_2, \ldots, \omega_n|\omega_0)} \\ &= g(\omega_0) \cdot M(\omega_1, \ldots, \omega_n|\omega_0)\,. \end{aligned} \tag{2.102}$$

The first term in eq. 2.102 is the state space reweighting factor g. It is defined as the likelihood ratio between the probability measures $\widetilde{\pi}$ and π

$$g(\omega_0) = \frac{\mathrm{d}\widetilde{\pi}}{\mathrm{d}\pi} = \frac{\widetilde{P_\pi}(\omega_0)}{P_\pi(\omega_0)} = \frac{Z}{\widetilde{Z}} \exp\left(-\frac{U(q_0)}{k_B T}\right) \tag{2.103}$$

and requires the condition of absolute continuity

$$\widetilde{\pi}(A) = 0 \Rightarrow \pi(A) = 0, \qquad \forall A \subset \Omega \tag{2.104}$$

between the state probability measures $\widetilde{\pi}$ and π (eq. 2.92).[29] M is called path space reweighting factor or just path reweighting factor

$$M(\omega_1, \ldots, \omega_n|\omega_0) = \frac{\widetilde{P_\mu}(\omega_1, \omega_2, \ldots, \omega_n|\omega_0)}{P_\mu(\omega_1, \omega_2, \ldots, \omega_n|\omega_0)} \tag{2.105}$$

and its functional form depends on the chosen integrator. Deriving M for Langevin integrators is a fundamental part of this thesis and and we refer to sec. 3.3 and sec. 3.4. For Langevin dynamics, the existence of M is guaranteed by the Girsanov theorem[111] given the absolute continuity

$$\widetilde{\mu}(\mathcal{A}) = 0 \Rightarrow \mu(\mathcal{A}) = 0, \qquad \forall \mathcal{A} \subset \Omega^n \tag{2.106}$$

between the path probability measures $\widetilde{\mu}$ and μ (eq. 2.94) is fulfilled.[29] The product of g and M (eq. 2.102) can be interpreted as the weight $W(\boldsymbol{\omega})$ the path $\boldsymbol{\omega}$ has with respect to the target potential. If the weight is larger than one, $W > 1$, then ω has a higher probability at the target potential than at the simulation potential. This means, its contribution towards a path expected value is more significant at the target potential than at the simulation potential. The opposite is true if $W < 1$ and if $W = 1$ then $\boldsymbol{\omega}$ has equal probability in both potentials.

In path reweighting, eq. 2.102 is used to express the path expected value $\widetilde{\langle o \rangle}$ at the target potential in terms of the path probability $P(\boldsymbol{\omega})$ at the simulation potential

$$\begin{aligned}\widetilde{\langle o \rangle} &= \int_\Omega \mathrm{d}\omega_0 \int_\Omega \mathrm{d}\omega_1 \cdots \int_\Omega \mathrm{d}\omega_n \, o(\boldsymbol{\omega}) \, \widetilde{P}(\boldsymbol{\omega}) \\ &= \int_\Omega \mathrm{d}\omega_0 \int_\Omega \mathrm{d}\omega_1 \cdots \int_\Omega \mathrm{d}\omega_n \, o(\boldsymbol{\omega}) \, W(\boldsymbol{\omega}) \, P(\boldsymbol{\omega}) \\ &= \int_\Omega \mathrm{d}\omega_0 \int_\Omega \mathrm{d}\omega_1 \cdots \int_\Omega \mathrm{d}\omega_n \, o(\boldsymbol{\omega}) \, g(\omega_0) M(\omega_1, \ldots, \omega_n | \omega_0) \, P(\boldsymbol{\omega}) \, . \end{aligned} \tag{2.107}$$

Consequently, we can compute $\widetilde{\langle o \rangle}$ at the target potential from a set of paths $S^{\mathrm{sim}} = \{\boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(s)}\}$ generated at the simulation potential

$$\begin{aligned}\widetilde{\langle o \rangle} &= \lim_{s\to\infty} \frac{1}{s} \sum_{i=1}^{s} W(\boldsymbol{\omega}^{(i)}) o(\boldsymbol{\omega}^{(i)}) \\ &= \lim_{s\to\infty} \frac{1}{s} \sum_{i=1}^{s} g(\omega_0^{(i)}) \, M(\omega_1^{(i)}, \ldots, \omega_n^{(i)} | \omega_0^{(i)}) \, o(\boldsymbol{\omega}^{(i)}) \, , \end{aligned} \tag{2.108}$$

with $\boldsymbol{\omega}^{(i)} = (\omega_0^{(i)}, \omega_1^{(i)}, \ldots, \omega_n^{(i)})$.

Equivalently, we can compute the time-correlation function $\widetilde{C}(\tau)$ at the target potential

$$\begin{aligned}\widetilde{C}(\tau) &= \lim_{s\to\infty} \frac{1}{s} \sum_{i=1}^{s} W(\boldsymbol{\omega}^{(i)}) f(\omega_0^{(i)}) f(\omega_n^{(i)}) \\ &= \lim_{s\to\infty} \frac{1}{s} \sum_{i=1}^{s} g(\omega_0^{(i)}) \, M(\omega_1^{(i)}, \ldots, \omega_n^{(i)} | \omega_0^{(i)}) \, f(\omega_0^{(i)}) f(\omega_n^{(i)}) \, , \end{aligned} \tag{2.109}$$

from a set of paths S^{sim} generated at the simulation potential with time step Δt. $\tau = n\Delta t$ denotes the lag time. Eqs. 2.108 and 2.109 clearly show, that each path contributes to the path expected value with a certain weight. Paths that are improbable in the target system have a small weight and their contribution to the path expected value is rather insignificant. This observation is a direct consequence of violating the absolute continuity requirements in eqs. 2.104 and 2.106.

2.6 Markov State Models

Molecular systems are high dimensional and the corresponding dynamics can be very complex. Markov State Models[72–75] (MSMs) are a tool to evaluate Molecular Dynamics (MD) data in order to build a low-dimensional model which captures the slow dynamics of the system. MSMs can be build from MD trajectories that were generated by a dynamics which ensures

Markovianity, ergodicity and microreversibility. An MSM discretizes the configuration space Γ of a system into y disjoint subsets $B_1, B_2, \ldots, B_y$ with $\bigcup_{i=1}^{y} B_i = \Gamma$ and approximates the trajectory as a Markov jump process between these discrete states. The key component of an MSM is the time-correlation matrix $\mathbf{C}(\tau) \in \mathbb{R}^{y \times y}$ whose elements c_{ij} are given as time-correlation functions. Eq. 2.96 defines the time-correlation function for the arbitrary function $f : \Omega \to \mathbb{R}$. In the case of MSMs, this arbitrary function in replaced by the indicator function

$$\mathbb{1}_{B_i}(\omega) = \begin{cases} 1, & \text{if } \omega \in B_i \\ 0, & \text{else} \end{cases} \tag{2.110}$$

and the elements c_{ij} of the time-correlation matrix $\mathbf{C}(\tau)$ with lag time $\tau = n\Delta t$ can be computed from a set of s paths $\{\boldsymbol{\omega}^{(1)}, \boldsymbol{\omega}^{(2)}, \ldots, \boldsymbol{\omega}^{(s)}\}$ as

$$c_{ij}(\tau) = \lim_{s \to \infty} \frac{1}{s} \sum_{i=1}^{s} \mathbb{1}_{B_i}(\omega_0^{(i)})\, \mathbb{1}_{B_j}(\omega_n^{(i)})\,, \tag{2.111}$$

where $\boldsymbol{\omega}^{(i)} = (\omega_0^{(i)}, \omega_1^{(i)}, \ldots, \omega_n^{(i)})$ denotes the i-th path. The set of paths can for example be extracted from a single long ergodic MD trajectory.[112] In this case, the sum in eq. 2.111 can be interpreted as counting the transitions from B_i to B_j within time τ along the trajectory. Row-normalizing $\mathbf{C}(\tau)$ yields the transition probability matrix $\mathbf{T}(\tau)$ whose elements $t_{ij}(\tau)$ describe the probability to observe a jump from subset B_i to subset B_j in time τ

$$\mathbf{t}_{ij}(\tau) = \frac{c_{ij}(\tau)}{\sum_{j=1}^{y} c_{ij}(\tau)}\,. \tag{2.112}$$

The left an right eigenvectors $\mathbf{l}_i$ and $\mathbf{r}_i$ of the transition probability matrix can be computed by solving the eigenvalue problems

$$\mathbf{T}(\tau)\mathbf{r_i} = \lambda_i(\tau)\mathbf{r}_i \tag{2.113a}$$

$$\mathbf{l}_i^{\top}\mathbf{T}(\tau) = \lambda_i(\tau)\mathbf{l}_i^{\top}\,, \tag{2.113b}$$

where λ_i denotes the corresponding eigenvalues. The approximation quality of the MSM can be evaluated by checking whether the implied time scales

$$t_i = -\frac{\tau}{\ln(\lambda_i(\tau))} \tag{2.114}$$

are constant $\forall \tau > 0$. The dominant left and right eigenvectors characterize the MSM and represent the slow dynamical processes of the system. The corresponding eigenvalues provide information about the timescales at which the respective process occurs.

Finally, we want to point out that path reweighting can be used to build the MSM of a target dynamics from a set of paths generated at a simulation dynamics. Since the construction of MSMs is based on computing time-correlation functions, the reweighting concept explained in sec. 2.5 can straight-forwardly be applied to reweight MSMs.

3 Publications

3.1 Paper A1

"GROMACS Stochastic Dynamics and BAOAB are equivalent configurational sampling algorithms"

S. Kieninger, B. G. Keller
J. Chem. Theory Comput., **2022**, *18*, 5792–5798

DOI: 10.1021/acs.jctc.2c00585
URL: `https://doi.org/10.1021/acs.jctc.2c00585`
Computational scripts available on Github: `https://github.com/bkellerlab/GSD_BAOA_BAOAB`
Preprint: `https://doi.org/10.48550/arXiv.2204.02105`

Contributions
Stefanie Kieninger and Bettina G. Keller conceived the project and wrote the manuscript. B.G.K. sketched the proof that the GROMACS Stochastic Dynamics integrator is equivalent to the BAOA integrator. S.K. summarized the mathematical and physical context in the "Theory" section, wrote the supporting information, summarized the comparison to the BAOAB integrator and created all figures in the manuscript and the supporting information. S.K. conducted all computational work. Both authors contributed to the final version of the manuscript.

Summary

We can use Molecular Dynamics (MD) simulations to investigate stationary and dynamical quantities of a given system, which are usually calculated as averages with respect to the underlying equilibrium distributions. In an MD program, the dynamics are governed by an equation of motion which describes the movement of each particle in the system according to classical physics. To solve the equation of motion, the MD program uses a numerical algorithm, also called an integrator, that generates a time-discretized trajectory at time step resolution. A trajectory is a time series that contains the positions and the momenta of each particle in the system at every time step Δt.[18,19,29]
Langevin integrators are widely used to perform MD simulations that sample the canonical ensemble (NVT; constant temperature T, particle number N and volume V).[27] They include a stochastic force that couples the system to an external heat bath which is the reason why they are also called Langevin thermostats. The corresponding equilibrium distributions are the configurational Boltzmann distribution in position space and the Maxwell-Boltzmann distribution in momentum space.
There exist different approaches to solve the Langevin equation of motion and a huge variety of different Langevin integrators have been reported.[19,30–45] It is also known, that the accuracy of the sampled equilibrium distributions varies with the chosen integrator and the size of the time step.[30,31,33,37,40,46–50] Due to the large number of different Langevin integrators, we are confronted with two questions:

1) What are the similarities and differences between the Langevin integrators?
2) Which integrator is the best choice for a certain application?

In this publication, we aim to answer question 1 and partially answer question 2 for the widely used Langevin integrators: BAOAB[30,31], GROMACS Stochastic Dynamics (GSD)[35] and BAOA[36]. Both the BAOAB and the BAOA integrator emerge from an approach that splits the vector field in the Langevin equation of motion into three parts labeled A, B and O. Part A and B are ordinary differential equations (ODEs) that describes a deterministic motion in position and momentum space, respectively. Part O represents a stochastic differential equation called the Ornstein-Uhlenbeck process which describes the motion in momentum space that is due to the friction and random force. All three parts can be solved separately yielding the update operators $\mathcal{A}, \mathcal{B}$ and $\mathcal{O}$. Combining the update operators consecutively yields Langevin integrators whose names represents the applied sequence of update operators in a left-to-right fashion.[19,30,31]
The BAOAB integrator is frequently used in atomistic MD simulations and is implemented in the toolkit OpenMMTools[113] for the MD package OpenMM.[114] It has been shown analytically as numerically, that BAOAB accurately reproduces the configurational Boltzmann distribution even at large time steps[30,31,33,47,115] and is frequently used in atomistic MD

simulations.[116,117] The BAOA integrator, also called LFMiddle[37], has recently been implemented in the MD packages OpenMM and AMBER[118]. BAOAB and BAOA are closely related, as has already been reported in literature.[30,36,37,50] Both algorithms sample the same positions and only differ in the momenta by a shift of $\frac{\Delta t}{2}$. In contrast to BAOAB and BAOA, the GSD integrator has been derived by extending the leapfrog algorithm for deterministic dynamics by an impulsive application of friction.[35] GSD has been implemented as the standard Langevin algorithm in GROMACS[119] and has been treated separately from BAOAB and BAOA in literature.

In this publication, we show analytically and numerically that GSD and BAOA are equivalent algorithms. We additionally visualize this result with numerical experiments in a one-dimensional model system, a cubic water box and an ideal gas. Since BAOA and BAOAB sample the same positions, it immediately follows that we can transfer BAOAB's superior configurational properties to GSD. Likewise, any analysis or benchmark of the configurational accuracy obtained for one of the three integrators equally applies to the other two integrators. These results also have practical implications on path reweighting methods.[73–75,120] The mathematical expression of the path reweighting factor strongly depends on the integrator equations, meaning that we expect the same expression for GSD, BAOA and BAOAB. Furthermore, the numerical studies in this publication imply that the BAOA/GSD integrator samples the Maxwell-Boltzmann distribution with higher accuracy and thus yields more accurate kinetic averages. Similar observations are mentioned in Ref. [37] and the documentations of OpenMM[121] and AMBER 2021.[122] However, we explicitly want to point out that the accuracy of the Boltzmann and Maxwell-Boltzmann distribution that we can observe for BAOA/GSD might not extend to correlations between positions and momenta.

The full publication is available at `https://doi.org/10.1021/acs.jctc.2c00585`.

3.2 Paper A2

"Dynamical reweighting methods for Markov models"

*S. Kieninger, *L. Donati, B. G. Keller
Curr. Opin. Struct. Biol., **2020**, *61*, 124–131.

DOI: 10.1016/j.sbi.2019.12.018
URL: `https://doi.org/10.1016/j.sbi.2019.12.018`
Preprint: `https://doi.org/10.48550/arXiv.1910.07894`
* Co-first author

Contributions
Stefanie Kieninger and Luca Donati are co-first authors of this publication. All three authors contributed to the literature research, categorizing the reweighting methods and summarizing the applications and formulating the assumptions of the respective methods. S.K. focused on writing the sections "Markov state models" and "Path reweighting", L.D. concentrated on writing the sections "Reweighting by rescaling the flux" and "Reweighting Kramers rate theory" and Bettina G. Keller focused on writing the section "Reweighting by formulating a likelihood function" and created Figure 1. All three authors contributed to the final version of the manuscript.

Summary

The conformational dynamics of biomolecules are essential for their function in living organisms.[5–7,123] Biomolecules have a high dimensional potential energy surface with various local minima that correspond to metastable conformations, separated by barriers of different heights. In principal, we can use Markov State Models[76–81] (MSMs) constructed from Molecular Dynamics (MD) simulations to capture the slow dynamics of a given system. The dominant eigenspace of the MSM describes the transitions between the long-lived conformational states and the corresponding eigenvalues can be used to calculate the relaxation time scales for the equilibrations across the energy barriers. However, high energy barriers can make it very challenging or even impossible to investigate the dynamics of long-lived conformations of biomolecules with unbiased MD simulations.
To overcome these difficulties, we can combine MD simulations based on a stochastic thermostat with enhanced sampling methods, such as umbrella sampling[56,57] or metadynamics.[58–60] The enhanced sampling techniques introduce a bias to the potential energy to facilitate the exploration of state space. Unfortunately, the bias alters the dynamics of a system and we have to apply dynamical reweighting methods to recover the unbiased dynamics.

In this publication, we focus on potential reweighting methods and answer the questions:

1) Which potential reweighting methods exist?
2) What are their similarities and differences?

Based on the framework the method has been derived from, we classify state-of-the-art potential reweighting methods into four categories, explain the underlying assumptions and summarize the applications that have been reported so far.
In the first category, we collect the reweighting methods[124–126] that are designed to reweight Kramers rate theory,[127,128] which describes the reaction rate k_{AB} between two long-lived conformations A and B along a single reaction coordinate. A very successful method in this category is the combination of reweighting and infrequent metadynamics[126] which has been used to investigate several protein-ligand unbinding processes.[129–131]
The other three categories include methods which can be used to reweight MSMs. To construct an MSM, we discretize the state space into disjoint subsets S_i and either compute the rate matrix $\mathbf{K}$ whose elements k_{ij} represent the rates between subsets S_i and S_j, or we count the number of transitions c_{ij} from S_i to S_j in a given trajectory and use these counts to calculate the transition probability matrix $\mathbf{P}$, whose elements p_{ij} describe the probability to go from S_i to S_j. Note, that the matrices $\mathbf{P}$ and $\mathbf{K}$ are strictly related and therefore share the same eigenspace.
Methods[24,132] in the category "Reweighting by rescaling the flux" describe the dynamics in terms of the probability density function that spreads over state space as time evolves

and reweight the elements of the rate matrix **K**. The reweighting is based on rescaling the geometrical averages of the stationary weights of adjacent subsets S_i and S_j. The methods have been used to study membrane permeabilities of a series of drug molecules[24] and to successfully predict the effects of mutations on the folding kinetics of proteins.[133]
The category "Reweighting by formulating a likelihood function" collects methods[134–137] that consider MD simulations at different biasing potentials. They formulate a likelihood function for each potential which depends on the elements p_{ij} of the transition matrix **P** and on the transition counts c_{ij} observed in the MD simulation at the respective biasing potential. The likelihood is then maximized by varying the elements p_{ij} to obtain a statistically optimal MSM at the corresponding biasing potential. The reweighting is based on combining all likelihood functions into an overall likelihood function, such that data from all potentials can be used to optimize **P** at a specific potential. The methods have been used to study the complete binding equilibrium of a small inhibitor molecules to proteins.[135,137]
The category "Path reweighting" collects methods[67–69,71–74,82] which use a long trajectory generated at the biased potential and cut it into short path snippets. The snippets are then used to reweight the transition counts c_{ij} from which we can calculate **P**. The reweighting is based on calculating the weights which the simulated paths would have in the unbiased potential. The methods have been used to compute mean first hitting times in alanine dipeptide[82] and to estimate the time scales associated with the opening and closing of a β-hairpin peptide.[74]

Since this publication does not include dynamical reweighting methods like temperature reweighting or reweighting for path sampling strategies, we want to refer to Refs. [61–63] and Refs. [138, 139], respectively. Moreover, we want to mention Ref. [140] as an excellent work that has been published after this publication has been released and can be classified in the category "Reweighting by rescaling the flux". Lastly, we want to point to a recent review[75] that focuses on two methods from the categories "Reweighting by rescaling the flux" and "Path reweighting".

The full publication is available at `https://doi.org/10.1016/j.sbi.2019.12.018`.

3.3 Paper A3

"Path probability ratios for Langevin dynamics - Exact and approximate"

S. Kieninger, B. G. Keller
J. Chem. Phys., **2021**, *154*, 094102.

DOI: 10.1063/5.0038408
URL: `https://doi.org/10.1063/5.0038408`
Python3 scripts available as supplementary material.
Preprint: `https://doi.org/10.48550/arXiv.2011.12849`

Contributions
Stefanie Kieninger and Bettina G. Keller conceived the project and wrote the manuscript. S.K. derived the reweighting factor for the ISP scheme for Langevin dynamics and proved that, under certain conditions, this reweighting factor is equivalent to the reweighting factor derived from overdamped Langevin dynamics. S. K. structured, formulated and summarized the shown mathematics and the "Theory". S. K. worked out the differences between the different reweighting factors and provided the corresponding overview. S. K. conducted all computational work. B.G.K. created Figure 5.A and S.K. created all other Figures. Both authors contributed to the final version of the manuscript.

Summary

In part **A2** of this thesis, we introduced Markov State Models[76–81] (MSMs) and potential reweighting techniques[75,120] to recover unbiased MSMs from Molecular Dynamics (MD) trajectories generated with enhanced sampling techniques. MSMs are a useful tool to capture the slow dynamics of a system along a small number of relevant coordinates. The core piece of an MSM is the count matrix $\mathbf{C}$, whose elements are represented by time-lagged correlation functions.
Enhanced sampling techniques, such as umbrella sampling[56,57] or metadynamics,[58–60] introduce a bias to the potential energy in order to facilitate the exploration of state space during the MD simulation. To construct the unbiased MSM from the biased trajectory, we can for example use path reweighting techniques[67–69,71–74,82] to "remove" the impact of the bias on elements of the count matrix $\mathbf{C}$.
In path reweighting, we split a trajectory generated at the biased potential into short paths ω and calculate the statistical weight W that each path would have in the unbiased potential. The weights and the respective biased paths are then used to calculate the unbiased count matrix $\mathbf{C}$, and subsequently the unbiased MSM. The statistical weight $W = g \cdot M$ depends on the potential energy and is composed of the weight in state space g (state space reweighting factor) and the weight in path space M (path space reweighting factor). One of the difficulties in path reweighting is to find a mathematical expression of the path reweighting factor, because M strongly depends on the integrator that was used to generate the biased paths.

In part **A1** of this thesis, we introduced Langevin integrators[19,30–45] as stochastic integrators that can be used to perform MD simulations in the canonical ensemble.[27] In the context of path reweighting, Langevin integrators are beneficial because we can use the Girsanov theorem[29,111] as a basis to derive an expression for M via the conditional path probability p. The path reweighting factor for overdamped Langevin dynamics has been known for several decades[64–66], whereas M for underdamped Langevin integrators has not been reported prior to this publication. As a solution, path reweighting methods used an approximate path reweighting factor M_{approx} to reweight underdamped Langevin dynamics.[72–74] Surprisingly, this strategy yielded excellent results although the derivation of M_{approx} is based on overdamped Langevin dynamics.

This publication reports the reweighting factor M_L for a variant of the Langevin leapfrog integrator developed by Izaguirre, Sweet, and Pande (ISP integrator).[39] With M_L, we are able to perform exact path reweighting for an underdamped Langevin integrator for the first time. Additionally, this publication aims to answer several questions:

1) How can we derive path reweighting factors for underdamped Langevin dynamics integrators?

2) Why does reweighting with M_{approx} yield excellent reweighting results for underdamped Langevin dynamics?

3) What is the relationship between the different representations of the reweighting factor?

In the context of question 1, we present two different strategies to derive M_L for the ISP integrator. In one strategy, we derive the expression of the conditional path probability p in terms of the path ω and the corresponding force by integrating out the random number dependency. M_L is then defined as the ratio of the probability a given path would have in the unbiased potential and the probability the same path would have in the biased potential. In the other strategy, we ask what random number η is needed to generate ω in the biased potential and what random number is needed to generate the same ω in the unbiased potential. We can then derive an expression for M_L via the random number difference $\Delta\eta = \widetilde{\eta} - \eta$. This second strategy is conceptually easy and straight forward to apply which is why we proposes the strategy as a blueprint to derive M_L for other Langevin integrators.
Having the expression for M_L provides the basis to tackle question 2. In this context, we prove analytically that M_{approx} is an excellent approximation to M_L given that the time step Δt and the collision rate fulfill the condition $\xi\Delta t < 1$. To understand why $M_{\text{approx}} \approx M_L$ is true, we show that the integrator choice only seems to have a minor effect on the random number difference $\Delta\eta$. Additionally, we demonstrate the same result numerically by reweighting the MSM of a one-dimensional model potential and the MSM on the torsion angle in butane.
In the context of question 3, we explain that the path reweighting factor for underdamped Langevin dynamics can be expressed in terms of the sampled path $M_L(\omega)$ or in terms of the random number sequence, that was used to generate this path $M_L(\eta)$. The same applies to overdamped Langevin dynamics and we get $M_o(\omega)$ and $M_o(\eta)$, with $M_o(\eta) = M_{\text{approx}}$. We show that $M_L(\omega)$ and $M_L(\eta)$ are connected via η as defined in the ISP integrator equation and $M_o(\omega)$ and $M_o(\eta)$ are connected via η as defined in the Euler-Maruyama integrator[28,29] equation.
For a path generated with ISP we demonstrate analytically and numerically that $M_o(\eta) = M_L(\eta) = M_L(\omega) \neq M_o(\omega)$. Please note that Refs. [72] and [73] already include an indirect answer to question 3.

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

Path probability ratios for Langevin dynamics—Exact and approximate

Cite as: J. Chem. Phys. 154, 094102 (2021); doi: 10.1063/5.0038408
Submitted: 22 November 2020 • Accepted: 9 February 2021 •
Published Online: 1 March 2021

S. Kieninger and B. G. Keller[a]

AFFILIATIONS

Department of Biology, Chemistry, Pharmacy, Freie Universität Berlin, Arnimallee 22, D-14195 Berlin, Germany

Note: This paper is part of the JCP Special Collection in Honor of Women in Chemical Physics and Physical Chemistry.
[a]Author to whom correspondence should be addressed: bettina.keller@fu-berlin.de

ABSTRACT

Path reweighting is a principally exact method to estimate dynamic properties from biased simulations—provided that the path probability ratio matches the stochastic integrator used in the simulation. Previously reported path probability ratios match the Euler–Maruyama scheme for overdamped Langevin dynamics. Since molecular dynamics simulations use Langevin dynamics rather than overdamped Langevin dynamics, this severely impedes the application of path reweighting methods. Here, we derive the path probability ratio M_L for Langevin dynamics propagated by a variant of the Langevin Leapfrog integrator. This new path probability ratio allows for exact reweighting of Langevin dynamics propagated by this integrator. We also show that a previously derived approximate path probability ratio M_{approx} differs from the exact M_L only by $\mathcal{O}(\xi^4\Delta t^4)$ and thus yields highly accurate dynamic reweighting results. (Δt is the integration time step, and ξ is the collision rate.) The results are tested, and the efficiency of path reweighting is explored using butane as an example.

I. INTRODUCTION

Molecular dynamics are astonishingly complex and occur in a wide range of length and timescales.[1–3] To elucidate the mechanisms by which different parts of a molecular system interact and how macroscopic properties arise from these interactions, molecular dynamics (MD) simulations have become an indispensable tool.[4–9] Because the timescales covered by MD simulations are often orders of magnitude lower than the slowest timescale of the system, a wide variety of enhanced sampling techniques have been developed, which distort the dynamics of the simulation such that rare molecular transitions occur more frequently. This can be achieved by raising the temperature or by adding a bias to the potential energy function.[10,11] How to extract the correct values of dynamical properties (mean-first passage times, residence times, binding rates, or transition probabilities) from these accelerated dynamics is an open question and a very active field of research.

The goal of dynamical reweighting methods is to estimate dynamical properties of the system at a target state $\widetilde{S}$ from a trajectory generated at simulation state S. S could correspond to a higher temperature or to a biased potential. Starting points for the derivation of dynamical reweighting methods are Kramers rate theory,[12–15] the likelihood function for estimating the transition probabilities from MD trajectories,[16–19] or a discretization of the Fokker–Planck equation.[7,20–22] The methods differ in the ease of use and the severity of the assumptions they make.[23]

A principally exact formalism to reweight dynamic properties is path reweighting methods, which have been reported already early in Refs. 24–28. In path reweighting methods, the trajectory generated at state S is split into short paths ω. Then, the path probability $\widetilde{P}_L(\omega;\Delta t|(x_0,v_0))$ of a given ω at the target state $\widetilde{S}$ is calculated by reweighting the path probability $P_L(\omega;\Delta t|(x_0,v_0))$ of ω at the simulation state S,

$$\widetilde{P}_L(\omega;\Delta t|(x_0,v_0)) \approx M\cdot P_L(\omega;\Delta t|(x_0,v_0)). \tag{1}$$

(x_0, v_0) is the initial state of the path ω, and Δt is the integration time step. $M(\omega)$ is the path probability ratio or reweighting factor. Equation (1) is exact if the path probability ratio $M = \widetilde{P}_L(\omega;\Delta t|(x_0,v_0))/P_L(\omega;\Delta t|(x_0,v_0))$ is derived from the numerical integration scheme used to generate ω. The mathematical basis

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

for path reweighting methods is the Girsanov theorem,[29,30] or else, they can be derived from the Onsager–Machlup action.[24–27,31] A prerequisite for path reweighting is that a stochastic integrator is used in the MD simulation, e.g., a Langevin thermostat.

However, it has been challenging to apply path reweighting to simulations of large molecular systems. For example, the variance of the reweighting estimators increases rapidly with increasing path length such that for long paths, reweighting becomes inefficient compared to direct simulation of the target state. Combining path reweighting techniques with Markov state models (MSMs) alleviates this problem.[32–35] In MSMs,[36–42] the dynamics of the system is represented by transitions between discrete states in the conformational space of the molecular system, where the lag time τ of the transition is much shorter than the slow timescales of the system. Thus, only short paths of length τ are needed to estimate and reweight the transition probabilities.

Second, a number of technical difficulties arise. The path probability ratio M decreases exponentially with the path length τ such that the standard numerical accuracy is quickly exceeded. This problem can be solved by using high precision arithmetic libraries.[35] To calculate the path probability ratio M, one needs to know the trajectory and the random numbers of the stochastic integrator at every integration time step. Writing this information to disk at every integration time step is not a workable option. We, therefore, proposed to calculate the path reweighting factor "on-the-fly" during the simulation and to write out intermediate results at regular intervals, e.g., whenever the positions are written to disk. The additional storage requirements and computational costs for the "on-the-fly"-calculations are negligible compared to the overall cost of the simulation.[34,35] Having solved the technical challenges, we tested the path reweighting method on several peptides using path lengths of up to $\tau = 600$ ps.[34,35] Applications to larger systems and longer path lengths are likely within reach.

Yet, the equation for the path probability ratio M poses a barrier to a more widespread use of path reweighting methods. Because M is derived from the stochastic integration scheme used to simulate the system, one cannot readily apply a path probability ratio derived for one integration scheme to a simulation generated by another integration scheme.

In temperature reweighting, i.e., when simulation and target state differ in the temperature, only the random term of the stochastic integrator is affected by the change in temperature. Path probability ratios for temperature reweighting have been constructed by rescaling the normal distributions of the random or noise terms of the stochastic integration scheme.[32,43]

In potential reweighting, i.e., when simulation and target state differ in the potential energy function, one needs to account for changes in the drift terms of the stochastic integration scheme. The path probability ratio M_o for the Euler–Maruyama scheme for overdamped Langevin dynamics has been reported multiple times.[24–26,33] However, the dynamics of large molecular systems is better reproduced by Langevin dynamics, and MD programs implement a wide variety of Langevin integration schemes.[44–53] The time-continuous Onsager–Machlup action for Langevin dynamics has been reported,[27] but to the best of our knowledge, path probability ratios for Langevin integration schemes M_L have not yet been reported. Thus, exact path reweighting for Langevin dynamics has not been possible, so far.

In Refs. 34 and 35, we demonstrated that path reweighting can be applied to biased simulations of large molecular systems, nonetheless. We used an approximate path probability ratio M_{approx} that is based on the path probability ratio for the Euler–Maruyama scheme but uses the random numbers that are generated during the Langevin MD simulation. We tested M_{approx} extensively, and for low-dimensional model systems and for molecular systems, this approximate path probability ratio yielded very accurate results. In these two publications, we used a variant of the Langevin Leapfrog integration scheme developed by Izaguirre, Sweet, and Pande[49] to propagate the system. Both the Langevin Leapfrog integration scheme and its variant are implemented in OpenMM[54] (see Appendix A). We will abbreviate the variant by the "ISP scheme."

In this contribution, we derive the path probability ratio M_L for Langevin dynamics propagated by a variant of the Langevin Leapfrog integrator.[49] M_L allows for exact reweighting of Langevin dynamics (Sec. IV). We analyze why M_{approx} is an excellent approximation to M_L (Sec. VI), and we discuss whether there are scenarios in which M_o is a viable approximation to M_L (Sec. V). The general framework of the path reweighting equations and the corresponding equations for the Euler–Maruyama scheme are summarized in Secs. II and III. Section VIII reports the computational details.

II. PATH REWEIGHTING

The path probability $P(\omega; \Delta t|(x_0, v_0))$ is the probability to generate a time-discretized path $\omega = (x_0, x_1, \ldots, x_n)$ starting in a predefined initial state (x_0, v_0) at the simulation potential $V(x)$. The notation emphasizes that the probability is conditioned on an initial state (x_0, v_0) and that the path has been generated with a fixed time step Δt, whereas ω is the argument of the function. In short, $P(\omega; \Delta t|(x_0, v_0))$ maps a path in position space to a probability. Its functional form depends on the integration scheme used to generate ω and the potential energy function.

The path probability ratio is the ratio between the probability $\widetilde{P}(\omega; \Delta t|(x_0, v_0))$ to generate a path ω at a target potential,

$$\widetilde{V}(x) = V(x) + U(x), \tag{2}$$

and the probability $P(\omega; \Delta t|(x_0, v_0))$ to generate the same path ω at the simulation potential $V(x)$,

$$M(\omega; \Delta t|(x_0, v_0)) = \frac{\widetilde{P}(\omega; \Delta t|(x_0, v_0))}{P(\omega; \Delta t|(x_0, v_0))}. \tag{3}$$

The potential energy function $U(x)$ is usually called perturbation or bias.

In integration schemes for stochastic dynamics, random numbers are used to propagate the system. If a single random number is drawn per integration step, then the probability to generate ω is equal to the probability $P(\eta)$ to generate the corresponding random number sequence $\eta = (\eta_0, \eta_1, \ldots, \eta_{n-1})$,

$$P(\omega; \Delta t|(x_0, v_0)) = P(\eta), \tag{4}$$

where ω and η are linked by the equations for the integration scheme. Since the random numbers η_k are drawn from a Gaussian

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

TABLE I. References to the equations for the properties introduced in Sec. II.

		Overdamped Langevin	Langevin
Equation of motion		Eq. (11)	Eq. (19)
Integration scheme		Eq. (12)	Eqs. (20) and (21)
Path probability	$P(\omega; \Delta t \vert (x_0, v_0))$	Eq. (13)	Eq. (22)
Path probability ratio	$M(\omega; \Delta t \vert (x_0, v_0))$	Eq. (14)	Eq. (23)
Random number	η_k	Eq. (15)	Eq. (24)
Random number difference	$\Delta\eta_k$	Eq. (17)	Eq. (26)
Random number probability ratio	$M(\omega, \eta; \Delta t \vert (x_0, v_0))$	Eq. (18)	Eq. (27)

distribution with zero mean and unit variance, the functional form of $P(\eta)$ is

$$P(\eta) = N \exp\left(-\frac{1}{2}\sum_{k=0}^{n-1}\eta_k^2\right), \quad N = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}}. \tag{5}$$

$P(\eta)$ is a function that maps a random number sequence to a probability. One can interpret Eq. (4) as a change in variables from ω to η, where the change is defined by the equations for the integration scheme.

Suppose that η is the random number sequence needed to generate ω at a simulation potential $V(x)$. To generate the same path at a target potential $\widetilde{V}(x)$, one would need a different random number sequence $\widetilde{\eta} = (\widetilde{\eta}_0, \widetilde{\eta}_1, \ldots, \widetilde{\eta}_{n-1})$, with

$$\widetilde{\eta}_k = \eta_k + \Delta\eta_k. \tag{6}$$

$\Delta\eta_k$ is the random number difference, and it depends on the integration scheme and the difference between the two potentials. The random number probability ratio is the ratio between the probability of drawing η and the probability of drawing $\widetilde{\eta}_k$,

$$\begin{aligned}\frac{P(\widetilde{\eta})}{P(\eta)} &= \frac{N \exp\left(-\frac{1}{2}\sum_{k=0}^{n-1}(\eta_k + \Delta\eta_k)^2\right)}{N \exp\left(-\frac{1}{2}\sum_{k=0}^{n-1}\eta_k^2\right)} \\ &= \exp\left(-\sum_{k=0}^{n-1}\eta_k \cdot \Delta\eta_k\right) \cdot \exp\left(-\frac{1}{2}\sum_{k=0}^{n-1}(\Delta\eta_k)^2\right).\end{aligned} \tag{7}$$

Mathematically, the following has happened in the previous paragraph. The path ω remained unchanged. The functional form of the path probability has changed as $\widetilde{P}(\omega; \Delta t \vert (x_0, v_0))$ because the potential energy enters the equations for the integration scheme. Likewise, the change in variables from ω to $\widetilde{\eta}$ has changed. The functional form of the random number probability remains the same [Eq. (5)]. The analogon to Eq. (4) for the target potential is

$$\widetilde{P}(\omega; \Delta t \vert (x_0, v_0)) = P(\widetilde{\eta}), \tag{8}$$

where ω and $\widetilde{\eta}$ are linked by the equations for the integration scheme using $\widetilde{V}(x)$. Given the two changes in variables for the simulation and the target potential, the path probability ratio [Eq. (3)] and the random number probability ratio [Eq. (7)] are equal. Note that Eq. (3) is a ratio of two different functions that have the same argument ω, whereas Eq. (7) is the ratio of the same function with different arguments η and $\widetilde{\eta}$.

Equation (7) is of little practical use because $\widetilde{\eta}$ is not available from a simulation at the simulation state. However, the random number difference $\Delta\eta_k$ can be expressed as a function of ω, and the random number probability ratio can thus be expressed as a function of ω and η,

$$M(\omega, \eta; \Delta t \vert (x_0, v_0)) = \frac{P(\widetilde{\eta})}{P(\eta)}. \tag{9}$$

For a path ω and the corresponding random number sequence η that was used to generate this path, we will use the following equality:

$$M(\omega, \eta; \Delta t \vert (x_0, v_0)) = M(\omega; \Delta t \vert (x_0, v_0)). \tag{10}$$

The functional form and the value of the properties introduced in this section depend strongly on the integration scheme. In Sec. III, we summarize the equations for the Euler–Maruyama scheme for overdamped Langevin dynamics. In Sec. IV, we derive the corresponding equations for the ISP integration scheme for Langevin dynamics (see Table I). Throughout this manuscript, properties associated with Langevin dynamics are subscripted with L, and properties associated with overdamped Langevin dynamics are subscripted with o.

III. OVERDAMPED LANGEVIN DYNAMICS

A. Equation of motion and integration scheme

Consider a one particle system that moves in a one-dimensional position space with temperature T and potential energy function V. The overdamped Langevin equation of motion is

$$\dot{x}(t) = -\frac{\nabla V(x(t))}{\xi m} + \sqrt{\frac{2k_BT}{\xi m}}\,\eta(t), \tag{11}$$

with particle mass m, position x, velocity $v = \dot{x}$, and Boltzmann constant k_B. $x(t) \in \Omega_o$ is the state of the system at time t, where $\Omega_o \subset \mathbb{R}$ is the state space of the system. The collision rate ξ (in units of s^{-1}) models the interaction with the thermal bath. $\eta(t) \in \mathbb{R}$ describes an uncorrelated Gaussian white noise with unit variance centered at zero, which is scaled by the volatility $\sqrt{\frac{2k_BT}{\xi m}}$.

A numerical algorithm to calculate an approximate solution to Eq. (11) is the Euler–Maruyama integration scheme,[30,55]

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

$$x_{k+1} = x_k - \frac{\nabla V(x_k)}{\xi m}\Delta t + \sqrt{\frac{2k_BT}{\xi m}}\sqrt{\Delta t}\,\eta_{o,k}, \tag{12}$$

where Δt is the time step, x_k is the position, and $\eta_{o,k}$ is the random number at iteration k. The random numbers are drawn from a Gaussian distribution with zero mean and unit variance. For $k = 0, \ldots, n-1$, Eq. (12) yields a time-discretized overdamped Langevin path $\omega_o = (x_0, x_1, \ldots, x_n)$, which starts at the pre-defined initial position x_0. Note that while the state of the system at iteration k is defined by the position x_k, the progress to x_{k+1} depends on x_k and on the value of the random number $\eta_{o,k}$. The random number sequence that was used to generate a specific ω_o is denoted by $\eta_o = (\eta_{o,0}, \ldots, \eta_{o,n-1})$.

B. Path probability and path probability ratio

The probability to observe a path ω_o generated by the Euler–Maruyama scheme [Eq. (12)] is[28,34,56,57]

$$P_o(\omega_o;\Delta t|x_0) = \left[\sqrt{\frac{\xi m}{4\pi k_BT\Delta t}}\right]^n \cdot \exp\left(-\frac{\xi m}{4k_BT\Delta t}\sum_{k=0}^{n-1}\left(x_{k+1} - x_k + \frac{\Delta t}{\xi m}\nabla V(x_k)\right)^2\right). \tag{13}$$

For the Euler–Maruyama scheme, the path probability $P_o(\omega_o;\Delta t|x_0)$ does not depend on the initial velocity; hence, we dropped v_0 in the notation. However, it does depend on the potential energy function $V(x)$ that has been used in Eq. (12) to generate the path ω_o.

The path probability that the same path ω_o has been generated at a target potential $\widetilde{V}(x)$ [Eq. (2)] is $\widetilde{P}_o(\omega_o;\Delta t|x_0)$, which is obtained by replacing the potential $V(x)$ with $\widetilde{V}(x)$ in Eq. (13). The ratio between the two path probabilities is

$$M_o(\omega_o;\Delta t|x_0) = \frac{\widetilde{P}_o(\omega_o;\Delta t|x_0)}{P_o(\omega_o;\Delta t|x_0)} = \exp\left(-\frac{\sum_{k=0}^{n-1}(x_{k+1}-x_k)\left(\nabla\widetilde{V}(x_k) - \nabla V(x_k)\right)}{2k_BT}\right) \times \exp\left(-\frac{\sum_{k=0}^{n-1}\left(\nabla\widetilde{V}^2(x_k) - \nabla V^2(x_k)\right)\Delta t}{4k_BT\xi m}\right). \tag{14}$$

Equation (14) is a function of the path ω_o and does not depend on the random number sequence η_o explicitly. It is equivalent to Eq. (B4) in Ref. 34.

C. Random numbers and random number probability ratio

Given ω_o, the sequence of random numbers η_o that was used to generate ω_o at the simulation potential $V(x)$ can be back-calculated by rearranging Eq. (12) for $\eta_{o,k}$,

$$\eta_{o,k} = \sqrt{\frac{\xi m}{2k_BT\Delta t}}\left(x_{k+1} - x_k + \frac{\nabla V(x_k)}{\xi m}\Delta t\right). \tag{15}$$

We remark that the path probability [Eq. (13)] can formally be derived by inserting Eq. (15) into Eq. (5). Since Eq. (15) defines a coordinate transformation from x_k to $\eta_{o,k}$, one needs to normalize with respect to the new coordinates in order to obtain the correct normalization constant. The random number sequence $\widetilde{\eta}_o$ needed to generate ω_o at a target potential $\widetilde{V}(x)$ is calculated by inserting Eq. (2) into Eq. (15),

$$\widetilde{\eta}_{o,k} = \sqrt{\frac{\xi m}{2k_BT\Delta t}}\left(x_{k+1} - x_k + \frac{\nabla V(x_k)}{\xi m}\Delta t\right) + \sqrt{\frac{\Delta t}{2k_BT\xi m}}\nabla U(x_k) = \eta_{o,k} + \Delta\eta_{o,k}. \tag{16}$$

Equation (15) defines the change in variables from ω to η_o for the Euler–Maruyama scheme at the simulation potential. Likewise, Eq. (16) defines the change in variables from ω to $\widetilde{\eta}_o$ at the target potential. The random number difference is

$$\Delta\eta_{o,k} = \sqrt{\frac{\Delta t}{2k_BT\xi m}}\nabla U(x_k). \tag{17}$$

It depends on the perturbation $U(x)$, but not on the simulation potential $V(x)$. Inserting $\Delta\eta_{o,k}$ [Eq. (17)] into Eq. (7) yields the random number probability ratio,

$$M_o(\omega_o,\eta_o;\Delta t|x_0) = \exp\left(-\sum_{k=0}^{n-1}\sqrt{\frac{\Delta t}{2k_BT\xi m}}\nabla U(x_k)\cdot\eta_{o,k}\right) \cdot \exp\left(-\frac{1}{2}\sum_{k=0}^{n-1}\frac{\Delta t}{2k_BT\xi m}\left(\nabla U(x_k)\right)^2\right). \tag{18}$$

Because of Eq. (10), Eqs. (14) and (18) are equal. However, the two probability ratios use different time-series and different information on the system to evaluate the path probability ratio. To evaluate Eq. (14), one needs the path ω_o, the simulation potential $V(x)$, and the target potential $\widetilde{V}(x)$. To evaluate Eq. (18), one needs the path ω_o, the random number sequence for the simulation potential η_o, and the perturbation $U(x)$. Because $U(x)$ often only affects a few coordinates of the systems, i.e., it is low-dimensional, Eq. (18) is computationally more efficient. Besides the force calculation $-\nabla V(x)$ needed to generate the path ω_o, it requires an additional force calculation $-\nabla U(x)$ only along the coordinates that are affected by the perturbation. By contrast, Eq. (14) requires an additional force calculation on the entire system $-\nabla\widetilde{V}(x)$.

IV. LANGEVIN DYNAMICS

A. Equation of motion and integration scheme

Consider a one particle system that moves in a one-dimensional position space with temperature T and potential energy function V. The Langevin equation of motion is

$$m\ddot{x}(t) = -\nabla V(x(t)) - \xi m\dot{x}(t) + \sqrt{2k_BT\xi m}\,\eta(t), \tag{19}$$

with particle mass m, position x, velocity $v = \dot{x}$, acceleration $a = \ddot{x}$, and Boltzmann constant k_B. The state of the system at time t is deter-

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

mined by the position and the velocity $(x(t), \dot{x}(t)) \in \Omega_L$, where $\Omega_L \subset \mathbb{R}^2$ is the state space of the system. The collision rate ξ (in units of s^{-1}) models the interaction with the thermal bath. $\eta \in \mathbb{R}$ describes an uncorrelated Gaussian white noise with unit variance centered at 0, which is scaled by the volatility $\sqrt{2k_BT\xi m}$.

A numerical algorithm to calculate an approximate solution to Eq. (19) is the ISP scheme,[49]

$$x_{k+1} = x_k + \exp(-\xi\,\Delta t)\, v_k \Delta t - [1 - \exp(-\xi\,\Delta t)]\, \frac{\nabla V(x_k)}{\xi m} \Delta t + \sqrt{\frac{k_B T}{m}\,[1 - \exp(-2\xi\,\Delta t)]}\, \eta_{L,k}\, \Delta t, \tag{20}$$

$$v_{k+1} = \frac{x_{k+1} - x_k}{\Delta t}, \tag{21}$$

where Δt is the time step, x_k is the position, v_k is the velocity, and $\eta_{L,k}$ is the random number at iteration k (see Appendix A). The random numbers are drawn from a Gaussian distribution with zero mean and unit variance. For $k = 0, \ldots, n-1$, Eqs. (20) and (21) yield a time-discretized Langevin path $\omega_L = ((x_0, v_0), (x_1, v_1), \ldots, (x_n, v_n))$, which starts at the pre-defined initial state (x_0, v_0). Note that while the state of the system at iteration k is defined by the tuple $(x_k, v_k) \in \Omega_L$, the progress to (x_{k+1}, v_{k+1}) depends on (x_k, v_k) and on the value of the random number $\eta_{L,k}$. The random number sequence that was used to generate a specific ω_L is denoted by $\eta_L = (\eta_{L,0}, \ldots, \eta_{L,n-1})$.

The position x_{k+1} is treated as a random variable because it directly depends on a random number [Eq. (20)], while the velocity v_{k+1} is calculated from the new position x_{k+1} and the preceding position x_k. Because the velocity v_k in Eq. (20) is determined by the positions x_k and x_{k-1} [Eq. (21)], it carries a small memory effect into the time-evolution of x.

B. Path probability and path probability ratio

The probability to generate a path ω_L by the ISP scheme [Eqs. (20) and (21)] at the simulation potential $V(x)$ is

$$P_L(\omega_L; \Delta t|(x_0, v_0)) = \left[\prod_{k=0}^{n-1} \delta\left(v_{k+1} - \frac{x_{k+1} - x_k}{\Delta t}\right)\right] \cdot \left[\sqrt{\frac{m}{2\pi k_B T \Delta t^2 (1 - \exp(-2\xi\Delta t))}}\right]^n \times \exp\left(-\sum_{k=0}^{n-1} \frac{m\left(x_{k+1} - x_k - \exp(-\xi\Delta t) v_k \Delta t + (1 - \exp(-\xi\Delta t))\frac{\nabla V(x_k)}{\xi m}\Delta t\right)^2}{2k_B T (1 - \exp(-2\xi\Delta t))\Delta t^2}\right). \tag{22}$$

The derivation of Eq. (22) is shown in Appendixes B and C. Appendix B explains the strategy for the derivation, and Appendix C shows how to solve the integrals that appear in the derivation.

The path probability $\widetilde{P}_L(\omega_L; \Delta t|(x_0, v_0))$ to generate a path ω_L by the ISP scheme at the target potential is obtained by inserting $\widetilde{V}(x)$ [Eq. (2)] into Eq. (22). The path probability ratio for overdamped Langevin dynamics is

$$\begin{aligned} M_L(\omega_L; \Delta t|(x_0, v_0)) &= \frac{\widetilde{P}_L(\omega_L; \Delta t|(x_0, v_0))}{P_L(\omega_L; \Delta t|(x_0, v_0))} \\ &= \exp\left(-\frac{\sum_{k=0}^{n-1}(x_{k+1} - x_k)\left(\nabla\widetilde{V}(x_k) - \nabla V(x_k)\right)}{k_B T \xi (1 + \exp(-\xi\Delta t))\Delta t}\right) \\ &\quad \cdot \exp\left(\frac{\sum_{k=0}^{n-1} v_k\left(\nabla\widetilde{V}(x_k) - \nabla V(x_k)\right)}{k_B T \xi (1 + \exp(\xi\Delta t))}\right) \\ &\quad \cdot \exp\left(-\frac{\exp(\xi\Delta t) - 1}{\exp(\xi\Delta t) + 1} \cdot \frac{\sum_{k=0}^{n-1}\left(\nabla\widetilde{V}^2(x_k) - \nabla V^2(x_k)\right)}{2k_B T \xi^2 m}\right). \end{aligned} \tag{23}$$

Analogous to Eq. (14), Eq. (23) is a function of the path ω_L and does not depend on the random number sequence η_L.

C. Random numbers and random number probability ratio

Given ω_L, the sequence of random numbers η_L, which was used to generate ω_L at the simulation potential $V(x)$, can be back-calculated by rearranging Eq. (20) for $\eta_{L,k}$,

$$\eta_{L,k} = \sqrt{\frac{m}{k_B T (1 - \exp(-2\xi\Delta t))\Delta t^2}} \times \left(x_{k+1} - x_k - \exp(-\xi\Delta t) v_k \Delta t + (1 - \exp(-\xi\Delta t))\frac{\nabla V(x_k)}{\xi m}\Delta t\right). \tag{24}$$

The random number sequence $\widetilde{\eta}_L$ needed to generate ω_L at a target potential $\widetilde{V}(x)$ is calculated by inserting Eq. (2) into Eq. (24),

$$\begin{aligned} \widetilde{\eta}_{L,k} &= \sqrt{\frac{m}{k_B T (1 - \exp(-2\xi\Delta t))\Delta t^2}} \times \left(x_{k+1} - x_k - \exp(-\xi\Delta t)(x_k - x_{k-1}) + (1 - \exp(-\xi\Delta t))\frac{\nabla V(x_k)}{\xi m}\Delta t\right) \\ &\quad + \sqrt{\frac{1}{k_B T \xi^2 m}} \cdot \frac{1 - \exp(-\xi\Delta t)}{\sqrt{1 - \exp(-2\xi\Delta t)}} \nabla U(x_k) \\ &= \eta_{L,k} + \Delta\eta_{L,k}. \end{aligned} \tag{25}$$

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

Equation (24) defines the change in variables from ω to η_L for the ISP scheme at the simulation potential. Likewise, Eq. (25) defines the change in variables from ω to $\widetilde{\eta}_L$ at the target potential. The random number difference is

$$\Delta\eta_{L,k} = \sqrt{\frac{1}{k_B T \xi^2 m}} \cdot \frac{1-\exp(-\xi\Delta t)}{\sqrt{1-\exp(-2\xi\Delta t)}} \nabla U(x_k). \tag{26}$$

Again, the random number difference depends on the perturbation potential $U(x)$, but not on the simulation potential $V(x)$. Inserting $\Delta\eta_{L,k}$ [Eq. (26)] into Eq. (7) yields the random number probability ratio,

$$\begin{aligned} &M_L(\omega_L, \eta_L; \Delta t|(x_0, v_0)) \\ &= \exp\left(-\frac{1-\exp(-\xi\Delta t)}{\sqrt{1-\exp(-2\xi\Delta t)}} \cdot \frac{\sum_{k=0}^{n-1} \nabla U(x_k)\, \eta_{L,k}}{\sqrt{k_B T \xi^2 m}} \right) \\ &\times \exp\left(-\frac{(1-\exp(-\xi\Delta t))^2}{1-\exp(-2\xi\Delta t)} \cdot \frac{\sum_{k=0}^{n-1} \nabla U^2(x_k)}{2 k_B T \xi^2 m} \right). \end{aligned} \tag{27}$$

Analogous to the path probability ratio for overdamped Langevin dynamics, $M_L(\omega_L; \Delta t|(x_0, v_o))$ [Eq. (23)] and $M_L(\omega_L, \eta_L; \Delta t|(x_0, v_0))$ [Eq. (27)] yield the same path probability ratio for a given path ω_L that has been generated using the random number sequence η_L, but they use different arguments. Again, the path probability from random numbers $M_L(\omega_L, \eta_L; \Delta t|(x_0, v_0))$ requires an additional force calculation $-\nabla U(x)$ only along the coordinates that are affected by the perturbation, making it computationally more efficient than $M_L(\omega_L; \Delta t|(x_0, v_0))$ in most cases.

V. COMPARING LANGEVIN AND OVERDAMPED LANGEVIN DYNAMICS

A. Test system

Our test system is a one-dimensional one particle system at the simulation potential $V(x)$ (Fig. 1, orange line) and at the target potential $\widetilde{V}(x)$ (Fig. 1, black line). The trajectories generated at $V(x)$ will be reweighted to the target potential $\widetilde{V}(x)$. The black lines in Fig. 4(b) represent the first three dominant MSM eigenfunctions[40] associated with the target potential. The implied timescales[37] are $t_0 = \infty$, $t_1 = 20.5$ s, and $t_2 = 6.0$ s, which are shown as black lines in Fig. 4(c). Computational details are reported in Sec. VIII.

B. From random numbers η to paths ω_o and ω_L

Given a random number sequence $\eta = (\eta_0, \ldots, \eta_{n-1})$ and a starting state (x_0, v_0), one can use the Euler–Maruyama scheme to generate an overdamped Langevin path ω_o, or else, one can use the ISP scheme to generate a Langevin path ω_L. We discuss briefly how the difference between ω_o and ω_L depends on the combined parameter $\xi\Delta t$, which can be interpreted as the number of collisions per time step.

In the limit of high friction $\xi m\dot{x} \gg m\ddot{x}$, the Langevin dynamics [Eq. (19)] approaches the overdamped Langevin dynamics

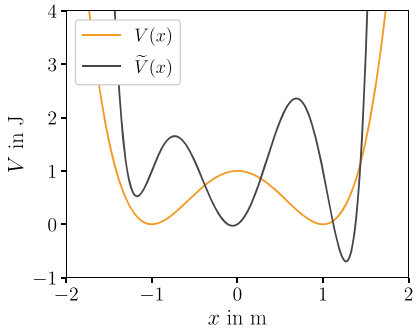


FIG. 1. Simulation potential $V(x)$ (orange) and target potential $\widetilde{V}(x)$ (black).

[Eq. (11)]. More specifically, in Eq. (19), set $m\ddot{x} = 0$, and rearranging yields Eq. (11). However, even though the equation of motion for Langevin dynamics converges to the equation of motion for overdamped Langevin dynamics, the ISP scheme [Eqs. (20) and (21)] does not converge to the Euler–Maruyama scheme [Eq. (12)] in the limit of high friction. By "high friction," we denote the range of collision rates ξ for which $e^{-\xi\Delta t} \approx 0$ in Eq. (20), but $\left|\frac{\nabla V}{\xi m}\right| > 0$. (As reference, $e^{-0.1} = 0.904$, $e^{-1} = 0.368$, and $e^{-5} = 0.007$.) If $e^{-\xi\Delta t} \approx 0$, then also $e^{-2\xi\Delta t} \approx 0$, and Eq. (20) becomes

$$x_{k+1} \approx x_k - \frac{\nabla V(x_k)}{\xi m}\Delta t + \sqrt{\frac{k_B T}{m}}\, \eta_{L,k}\, \Delta t. \tag{28}$$

The first two terms on the right-hand side are identical to the Euler–Maruyama scheme [Eq. (12)], but the random number term differs from the Euler–Maruyama scheme. Thus, even in the limit of high friction, the two algorithms yield different paths for a given random number sequence η. The difference between a Langevin path ω_L and an overdamped Langevin path ω_o can be scaled by the combined parameter $\xi\Delta t$. For some value $\xi\Delta t > 1$, the difference between the two paths becomes minimal before increasing again, but for no value of $\xi\Delta t$, the two paths fully coincide.

When Langevin integration schemes are used as a thermostat in MD simulations, the optimal friction coefficient should reproduce the expected temperature fluctuations and therefore depends on the system and the simulation box.[58] Reported collision rates[49,50,59] (while keeping the time step at $\Delta t = 0.002$ ps) range from 0.1 ps^{-1} to ~100 ps^{-1}, corresponding to $\xi\Delta t = 0.0002$ to $\xi\Delta t = 0.2$. However, even for a large collision rate of 100 ps^{-1}, $e^{-\xi\Delta t} = e^{-0.2} = 0.819 \not\approx 0$. For these two reasons—MD simulations are not conducted in the high-friction regime, and even in the high-friction regime, ω_o differs from ω_L—a simulation with the ISP scheme yields a materially different path ensemble than a simulation with the Euler–Maruyama scheme.

C. From a path ω to random numbers η_o and η_L

In Sec. V B, we showed that given a random number sequence η, the path generated by the Euler–Maruyama integration scheme

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

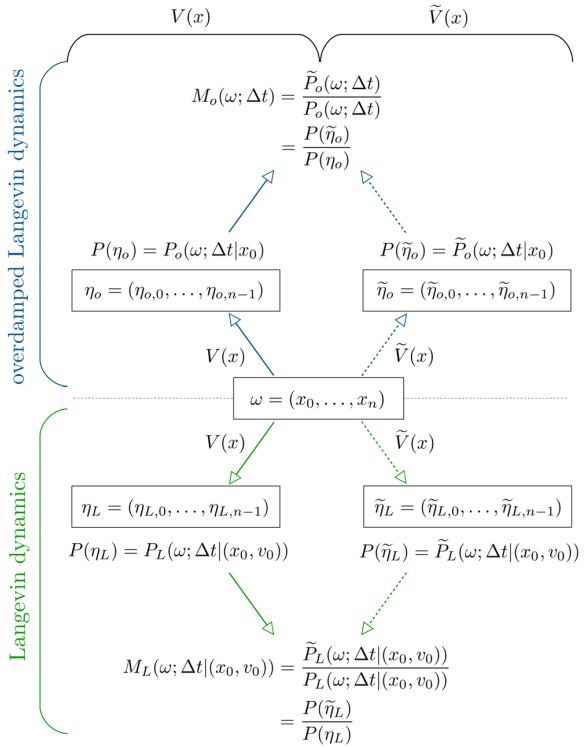


FIG. 2. Overview of path probabilities and path probability ratios for a sample path $\omega = (x_0, \ldots x_n)$.

for overdamped Langevin dynamics differs from the path generated by the ISP integration scheme for Langevin dynamics. More relevant for path reweighting is the reverse situation: Given a sample path $\omega = (x_0, \ldots, x_n)$ in position space and the parameters of the dynamics (m, V, T, ξ, k_B, and Δt), how does the random number sequence η_o needed to generate ω with the Euler–Maruyama scheme [Eq. (12)] differ from the random number sequence η_L needed to generate the same ω with the ISP scheme [Eqs. (20) and (21)]? An equivalent question is as follows: How does the path probability that ω has been generated by the Euler–Maruyama scheme differ from the path probability that ω has been generated by the ISP scheme? How does this difference affect the path probability ratios between the simulation and a target potential? Figure 2 gives an overview of the quantities we will compare. Note that we dropped the index o or L from the path ω because ω is a given dataset, which will be analyzed using various approaches to calculate the path probabilities.

First, we need to discuss whether such a comparison between the ISP scheme and the Euler–Maruyama scheme is even possible. From an algorithmic view point, this is clearly possible because both integrators [Eqs. (12) and (20)] use a single random number per integration time step. The path probabilities are then equal to the probability of the different random number sequences η_L and η_o needed to generate ω. From a physical view point, the answer is not as clear because overdamped Langevin dynamics evolves in position space (x_k), whereas Langevin dynamics evolves in phase space (x_k, v_k). The velocity v_k enters the integration scheme [Eq. (20)] and the path probability [Eq. (22)]. However, v_k is fully determined by the current position x_k and the previous position x_{k-1} [Eq. (21)]. Thus, if the initial velocity v_0 is known, the position trajectory is enough to evaluate the path probability [Eq. (22)], and the comparison to overdamped Langevin dynamics is possible.

We consider the test system described in Sec. V A at the simulation potential $V(x)$ (double-well potential) simulated by the ISP scheme for Langevin dynamics. With $\xi = 50\ \mathrm{s}^{-1}$ and $\Delta t = 0.01$ s, we have $e^{-\xi\Delta t} = e^{-0.5} = 0.607 \not\approx 0$, meaning that the system is not in the high-friction limit. Figure 3(a) additionally shows that with these parameters $\mathcal{O}(\xi m\dot{x}) \approx \mathcal{O}(m\ddot{x})$ and also according to the criterion for the stochastic differential equation, the system is not in the high-friction limit.

Figure 3(b) shows a sample path $\omega = (x_0, x_1, \ldots, x_{10})$. Figure 3(c) shows the random numbers η_o needed to generate ω with the Euler–Maruyama scheme [blue solid line, calculated using Eq. (15)] and the random numbers η_L needed to generate ω with the ISP scheme [green solid line, calculated using Eq. (24)]. As expected for the low-friction regime, these two random number sequences differ markedly.

Consequently, the path probabilities differ. Figure 3(d) shows the unnormalized path probability for generating ω with the Euler–Maruyama scheme (blue solid line),

$$P_o(\omega; \Delta t|x_0) \sim \exp\left(-\frac{\xi m}{4k_B T\Delta t}\sum_{k=0}^{n-1}\left(x_{k+1} - x_k + \frac{\Delta t}{\xi m}\nabla V(x_k)\right)^2\right), \tag{29}$$

and for generating ω with the ISP scheme (green solid line),

$$P_L(\omega; \Delta t|(x_0, v_0)) \sim \exp\left(-\sum_{k=0}^{n-1}\frac{m\left(x_{k+1} - x_k - \exp(-\xi\Delta t)v_k\Delta t + (1 - \exp(-\xi\Delta t))\frac{\nabla V(x_k)}{\xi m}\Delta t\right)^2}{2k_B T(1 - \exp(-2\xi\Delta t))\Delta t^2}\right), \tag{30}$$

where we omitted those factors from Eqs. (13) and (22) that cancel in the path probability ratio. We checked that the path probabilities are consistent with $P(\eta_o)$ and $P(\eta_L)$. The two path probabilities diverge from the first simulation step on. After ten integration time steps, they differ by two orders of magnitude. Clearly, $P_L(\omega; \Delta t|(x_0, v_0))$ cannot be used as an approximation for $P_o(\omega; \Delta t|x_0)$.

However, an interesting observation arises when we consider reweighting ω to the target potential $\widetilde{V}(x)$ (triple-well potential).

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

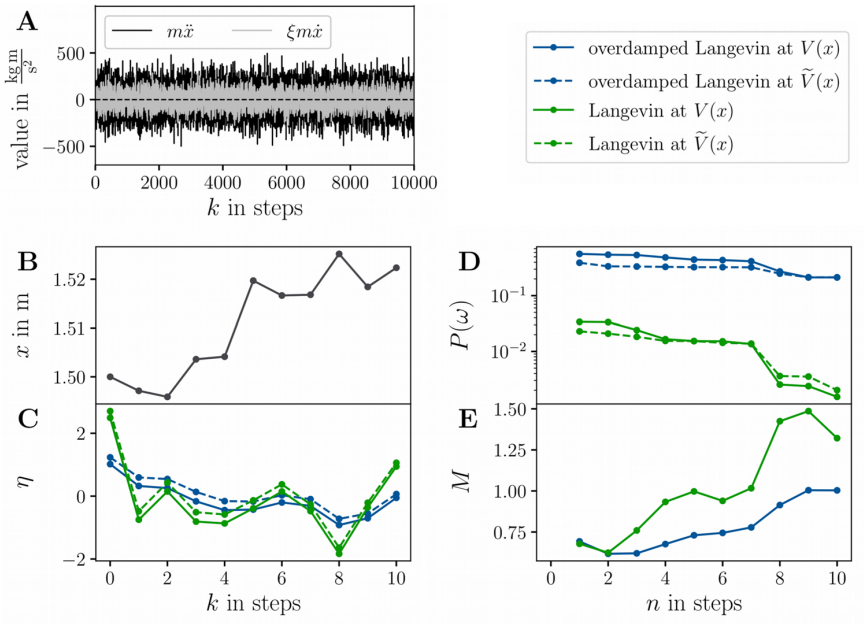


FIG. 3. (a) The acceleration term $m\ddot{x}$ and the friction $\xi m\dot{x}$ for the test system at $V(x)$. (b) Example path ω of length $n = 10$. (c) Random number sequences η_L (green solid), η_o (blue solid), $\widetilde{\eta}_L$ (green dashed), and $\widetilde{\eta}_o$ (blue dashed) that correspond to ω. (d) Path probabilities $P_L(\omega; \Delta t|(x_0, v_0))$ (green solid), $P(\omega; \Delta t|x_0)$ (blue solid), $\widetilde{P}_L(\omega; \Delta t|(x_0, v_0))$ (green dashed), and $\widetilde{P}_o(\omega; \Delta t|x_0)$ (blue dashed). (e) Path probability ratios: $M_L(\omega, \Delta t|(x_0, v_0))$ (green) and $M_o(\omega; \Delta t|x_0)$ (blue).

Figure 3(c) shows the random numbers $\widetilde{\eta}_o$ needed to generate ω with the Euler–Maruyama scheme at $\widetilde{V}(x)$ [blue dashed line, calculated using Eq. (16)] and the random numbers $\widetilde{\eta}_L$ needed to generate ω with the ISP scheme at $\widetilde{V}(x)$ [green dashed line, calculated using Eq. (25)]. The corresponding unnormalized path probabilities $\sim \widetilde{P}_o(\omega; \Delta t|x_0)$ and $\sim \widetilde{P}_L(\omega; \Delta t|(x_0, v_0))$ are shown as dashed lines in Fig. 3(d). Strikingly, a change in the integration scheme from the Euler–Maruyama scheme to ISP has a much stronger influence on the random numbers and the path probability than the modification of the potential energy function. Figure 3(e) shows the path probability ratios, i.e., the ratio between the dashed and the solid lines in Fig. 3(d), for the Euler–Maruyama scheme $M_o = M_o(\omega; \Delta t|x_0) = M_o(\omega, \eta_o; \Delta t|x_0)$ (blue line) and the ISP scheme $M_L = M_L(\omega; \Delta t|(x_0, v_0)) = M_L(\omega, \eta_L; \Delta t|(x_0, v_0))$ (green line). Because, within an integration scheme, the path probability does not change drastically when going from the simulation potential $V(x)$ to the target potential $\widetilde{V}(x)$, both path probability ratios remain at ≈1 throughout the path and follow similar curves, that is, the path probability ratios for Langevin and overdamped Langevin dynamics are much more similar than the underlying path probabilities.

D. Path reweighting

We return to the scenario described in the Introduction and ask the following: are the two path probability ratios similar enough that we can use M_o as an approximation to M_L in Eq. (1)? Figure 4(a) compares different ways to calculate the path probability $\widetilde{P}_L(\omega; \Delta t|(x_0, v_0))$, i.e., the probability with which an example path ω would have been generated at the target potential $\widetilde{V}(x)$. The black line is the reference solution calculated by inserting $\widetilde{V}(x)$ into Eq. (22). It is identical to the green dashed line in Fig. 3(d). The green line in Fig. 4(a) shows the reweighted path probability, where we used the exact path probability ratio for the ISP scheme, $M_L(\omega; \Delta t|(x_0, v_0))$ [Eq. (23)], in Eq. (1). As expected, this reweighted path probability coincides with the directly calculated path probability. The blue line shows the reweighted path probability, where we used the path probability ratio for the Euler–Maruyama scheme, $M_o(\omega; \Delta t|x_0)$ [Eq. (14)], as an approximation to M_L in Eq. (1). The path probability deviates from the reference solution, but overall follows a similar curve.

Figure 4(a) merely serves to illustrate the concepts. With only ten steps, the example path ω is far too short to judge the accuracy of the two path probability ratios for reweighting dynamic properties. We, therefore, constructed MSMs for the target potential $\widetilde{V}(x)$. The reference solution has been generated from simulations at the target potential $\widetilde{V}(x)$ using the ISP scheme. The dominant MSM eigenfunctions and associated implied timescales are shown as black lines in Figs. 4(b) and 4(c). Next, we ran simulations at the simulation potential $V(x)$ using the ISP scheme and constructed a reweighted MSM using the exact reweighting factor $M_L(\omega; \Delta t|(x_0, v_0))$ [Eq. (23)]. The dominant MSM eigenfunctions are shown as green lines in Fig. 4(b). They exactly match the reference solution. The reweighted implied timescales are shown as green lines in Fig. 4(c) and are in good agreement with the reference solution. Finally, we used the simulation at $V(x)$ to construct a reweighted MSM using the reweighting factor for the Euler–Maruyama scheme $M_o(\omega; \Delta t|x_0)$ [Eq. (14)]. The dominant MSM eigenfunctions are shown as blue lines in Fig. 4(b). The eigenfunctions differ considerably from the reference solution. Most notably,

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

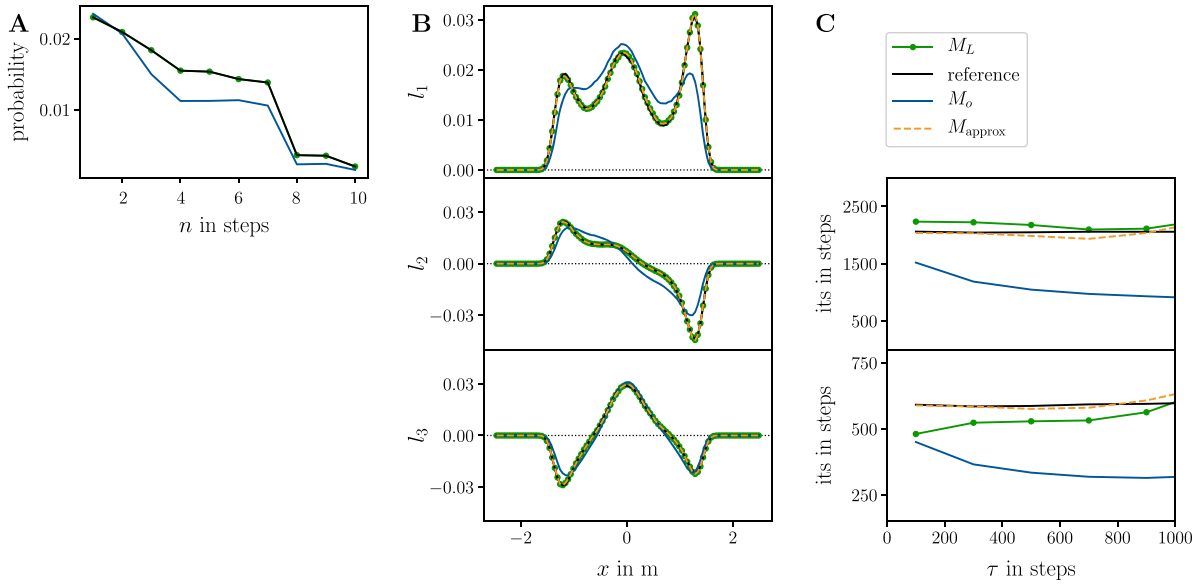


FIG. 4. (a) Reference and reweighted path probabilities for ω for Langevin dynamics. (b) Reference and reweighted first three dominant MSM left eigenfunctions l_1, l_2, and l_3 associated with $\widetilde{V}(x)$ for Langevin dynamics. (c) Reference and reweighted implied timescales corresponding to l_2 and l_3.

the stationary distribution is not reproduced correctly [blue line in the upper panel of Fig. 4(b)]. The left peak is reduced to a shoulder of the central peak, and the relative heights of central peak and the right peak do not match those of the reference solution. Likewise, the implied timescales [blue line in Fig. 4(c)] are severely underestimated. This indicates that using the path probability ratio for overdamped Langevin dynamics, $M_o(\omega;\, \Delta t|x_0)$, to reweight Langevin trajectories does not yield acceptable results.

VI. APPROXIMATE PATH PROBABILITY RATIO

A. Derivation and numerical results

With the results from Sec. IV, the exact random number probability ratio $M_L(\omega,\, \eta_L;\, \Delta t|(x_0,\, v_0))$ [Eq. (7)] for the ISP scheme is straightforward to evaluate from a simulation at $V(x)$: the random number sequence $\eta = \eta_L$ can be recorded during the simulation, and the random number difference $\Delta\eta = \Delta\eta_L$ is given by Eq. (26). Inserting η_L and $\Delta\eta_L$ into Eq. (7) yields $M_L(\omega,\, \eta_L;\, \Delta t|(x_0,\, v_0))$. However, $\Delta\eta_{L,k}$ in Eq. (26) is specific to the ISP scheme. If one uses a different Langevin integration scheme to simulate the dynamics at $V(x)$, one needs to adapt Eq. (26) via the strategy outlined in Sec. IV.

Fortunately, the random number difference for overdamped Langevin dynamics $\Delta\eta_{o,k}$ [Eq. (17)] is approximately equal to $\Delta\eta_{L,k}$ for any given perturbation $U(x)$. Figure 3(c) already suggests that. In Appendix D, we show that the difference between $\Delta\eta^2_{L,k}$ and $\Delta\eta^2_{o,k}$ is, in fact, only of $\mathcal{O}(\xi^4\Delta t^4)$ so that for $\xi\Delta t < 1$, we can assume with high accuracy that

$$\Delta\eta_{L,k} \approx \Delta\eta_{o,k},$$
$$\sqrt{\frac{1}{k_BT\xi^2 m}}\,\frac{1-\exp(-\xi\Delta t)}{\sqrt{1-\exp(-2\xi\Delta t)}}\cdot\nabla U(x_k) \approx \sqrt{\frac{\Delta t}{2k_BT\xi m}}\cdot\nabla U(x_k). \tag{31}$$

The difference between $\Delta\eta_{L,k}$ and $\Delta\eta_{o,k}$ is determined by the prefactors in front of $\nabla U(x_k)$ in Eq. (31), which are shown as a function of $\xi\Delta t$ in Fig. 5(b). For $\xi\Delta t < 1$, the two curves are virtually identical.

With the approximation in Eq. (31), we can derive an approximate random number probability ratio, by using the recorded η_L, but substituting $\Delta\eta_{L,k}$ [Eq. (26)] by $\Delta\eta_{o,k}$ [Eq. (17)] in Eq. (7), we obtain

$$\begin{aligned} M_L(\omega,\eta_L;\Delta t|(x_0,v_0)) &\approx M_{\text{approx}}(\omega,\eta_L;\Delta t|x_0) \\ &= \exp\left(-\sum_{k=0}^{n-1}\sqrt{\frac{\Delta t}{2k_BT\xi m}}\nabla U(x_k)\cdot\eta_{L,k}\right) \\ &\quad\cdot\exp\left(-\frac{1}{2}\sum_{k=0}^{n-1}\frac{\Delta t}{2k_BT\xi m}(\nabla U(x_k))^2\right). \end{aligned} \tag{32}$$

Equation (32) has the same functional form as the random number probability ratio for the Euler–Maruyama scheme $M_o(\omega,\, \eta_o;\, \Delta t|x_0)$ [Eq. (18)], but it uses η_L, the random numbers generated during the ISP simulation, instead of η_o. Equation (32) is the approximation

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

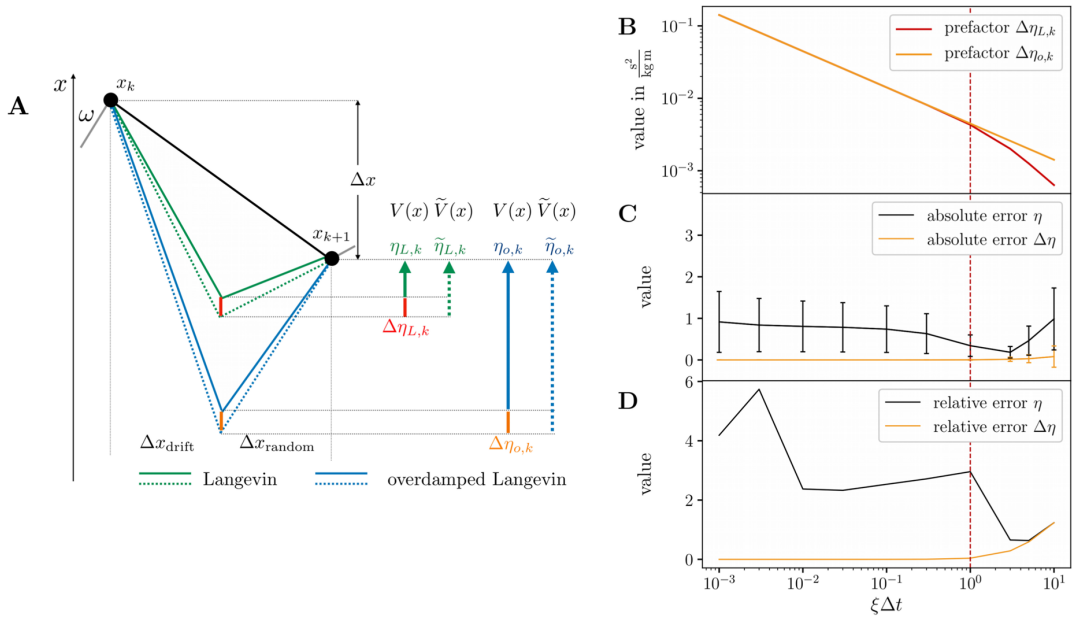


FIG. 5. (a) Sketch of a step $x_k \to x_{k+1}$ and the quantities of influence for Langevin and overdamped Langevin dynamics. (b) Prefactors of $\Delta\eta_{L,k}$ and $\Delta\eta_{o,k}$ as a function of $\xi\Delta t$. (c) Absolute difference (absolute error) between the random numbers $\langle|\eta_{o,k} - \eta_{L,k}|\rangle$ and the random number differences $\langle|\Delta\eta_{o,k} - \Delta\eta_{L,k}|\rangle$ as a function of $\xi\Delta t$. (d) Relative difference (relative error) between the random numbers $\langle|(\eta_{o,k} - \eta_{L,k})/\eta_{L,k}|\rangle$ and the random number differences $\langle|(\Delta\eta_{o,k} - \Delta\eta_{L,k})/\Delta\eta_{L,k}|\rangle$ as a function of $\xi\Delta t$.

that we used in Refs. 34 and 35 because we had not yet derived $M_L(\omega, \eta_L; \Delta t|(x_0, v_0))$ [Eqs. (23) and (27)].

Figure 4 demonstrates the accuracy of the approximate random number probability ratio $M_{\text{approx}}(\omega, \eta_L; \Delta t|x_0)$ [Eq. (32)] for our test system. The orange dashed line in Fig. 4(a) shows the reweighted path probability for the short example path, where we used $M_{\text{approx}}(\omega, \eta_L; \Delta t|x_0)$ [Eq. (32)], in Eq. (1). It exactly matches the reference solution (black line).

Next, we constructed a reweighted MSM for the target potential $\widetilde{V}(x)$ based on our simulations at the simulation potential $V(x)$ using $M_{\text{approx}}(\omega, \eta_L; \Delta t|x_0)$ [Eq. (32)] to reweight the transition counts. The dominant MSM eigenfunctions of the reweighted MSM are shown as orange dashed lines in Fig. 4(b). They exactly match the reference solution. The reweighted implied timescales are shown as orange dashed lines in Fig. 4(c) and seem to match the reference solution even better than the ones calculated using the exact path probability ratio [green line in Fig. 4(c)]. However, the difference between the orange dashed line and the green line is likely within statistical uncertainty. In summary, $M_{\text{approx}}(\omega, \eta_L; \Delta t|x_0)$ is a highly accurate approximation to $M_L(\omega, \eta_L; \Delta t|x_0)$ for $\xi\Delta t < 1$. Using $M_{\text{approx}}(\omega, \eta_L; \Delta t|x_0)$ instead of $M_L(\omega, \eta_L; \Delta t|x_0)$ could even have the following advantages: (*i*) the implementation is less error-prone because the functional form of M_{approx} is simpler than that of M_L and (*ii*) M_{approx} might be numerically more stable because the calculation of exponential function on the left-hand side of Eq. (31) is avoided.

B. Intuition

We discuss why $M_{\text{approx}}(\omega, \eta_L; \Delta t|x_0)$ is a better approximation to $M_L(\omega, \eta_L; \Delta t|x_0)$ than $M_o(\omega; \Delta t|x_0) = M_o(\omega, \eta_o; \Delta t|x_0)$. Figure 5(a) shows one integration time step of a stochastic integration scheme from x_k to x_{k+1} (black line). From k to $k+1$, the system has progressed by $\Delta x = x_{k+1} - x_k$. In the ISP scheme, this progress is composed of a progress

$$\Delta x_{\text{drift},L} = \exp(-\xi\,\Delta t)\, v_k \Delta t - [1 - \exp(-\xi\,\Delta t)]\, \frac{\nabla V(x_k)}{\xi m} \Delta t \qquad (33)$$

due to the drift force and the velocity of the system [second and third terms on the right-hand side of Eq. (20)] and a progress

$$\Delta x_{\text{random},L} = \sqrt{\frac{k_B T}{m} [1 - \exp(-2\xi\,\Delta t)]}\, \eta_{L,k}\, \Delta t \qquad (34)$$

due to the random force [fourth term on the right-hand side of Eq. (20)] such that $\Delta x = \Delta x_{\text{drift},L} + \Delta x_{\text{random},L}$. $\Delta x_{\text{drift},L}$ and $\Delta x_{\text{random},L}$ are illustrated as green solid lines in Fig. 5(a). The probability of generating the step $x_k \to x_{k+1}$ is determined by $\Delta x_{\text{random},L}$, which is proportional to the random number $\eta_{L,k}$ (green solid arrow).

With a different potential energy function $\widetilde{V}(x)$ at x_k, the displacement due to the drift force differs from the original $\Delta x_{\text{drift},L}$. To achieve the same overall displacement Δx, $\Delta x_{\text{random},L}$ needs to

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

be adjusted (green dotted line). The corresponding random number $\widetilde{\eta}_{L,k}$ is shown as a green dotted arrow, and the difference between the two random numbers $\Delta\eta_{L,k}$ is shown as a red line. In path reweighting, one constructs $\widetilde{\eta}_{L,k}$ by adding $\Delta\eta_{L,k}$ to $\eta_{L,k}$,

$$\widetilde{\eta}_{L,k} = \eta_{L,k} + \Delta\eta_{L,k} \tag{35}$$

[analogous to Eq. (6)], which then yields the general form of the random number probability ratio in Eq. (7).

An analogous analysis applies to the Euler-Maruyama scheme, where the progress due to the drift force is

$$\Delta x_{\mathrm{drift},o} = -\frac{\nabla V(x_k)}{\xi m}\,\Delta t \tag{36}$$

[second term on the right-hand side of Eq. (12)], and the progress due to the random force is

$$\Delta x_{\mathrm{random},o} = \sqrt{\frac{2k_BT}{\xi m}}\,\sqrt{\Delta t}\,\eta_{o,k} \tag{37}$$

[third term on the right-hand side of Eq. (12)]. In Fig. 5(a), $\Delta x_{\mathrm{drift},o}$ and $\Delta x_{\mathrm{random},o}$ are illustrated as blue solid lines, and the random number is represented as a blue solid arrow. With a different potential energy function $\widetilde{V}(x)$ at x_k, the progress due to the drift force differs from the original $\Delta x_{\mathrm{drift},o}$. To achieve the same overall progress Δx, $\Delta x_{\mathrm{random},o}$ needs to be adjusted (blue dotted line). The corresponding random number $\widetilde{\eta}_{o,k}$ is shown as a blue dotted arrow, and the difference between the two random numbers $\Delta\eta_{o,k}$ is shown as an orange line.

In Sec. VI A, we have shown that $\Delta\eta_{L,k} \approx \Delta\eta_{o,k}$ (for $\xi\Delta t < 1$). Thus, approximating $\Delta\eta_{L,k}$ by $\Delta\eta_{o,k}$ in Eq. (35), or, visually, approximating the red line by the orange line in Fig. 5(a), is valid. However, the displacement due to the drift $\Delta x_{\mathrm{drift},o}$ in the Euler–Maruyama scheme can differ strongly from $\Delta x_{\mathrm{drift},L}$ in the ISP scheme, and consequently, the random numbers needed to generate the same overall progress Δx differ

$$\eta_{L,k} \not\approx \eta_{o,k} \tag{38}$$

[blue solid and green solid arrow in Fig. 5(a)]. Consequently, approximating $\eta_{L,k}$ by $\eta_{o,k}$ in Eq. (35), or visually approximating the green solid arrow by the blue solid arrow in Fig. 5(a), is not valid.

The exact random number probability ratio $M_L(\omega, \eta_L; \Delta t|(x_0, v_0))$ [Eq. (27)] uses the exact η_L recorded during the simulation and the exact $\Delta\eta_L$ [Eq. (26)]. It therefore yields results that exactly match the reference solutions (green lines in Fig. 4). $M_{\mathrm{approx}}(\omega, \eta_L; \Delta t|x_0)$ uses the exact η_L recorded during the simulation but approximates $\Delta\eta_{L,k}$ by $\Delta\eta_{o,k}$. This introduces only a small error but still yields excellent reweighting results in our test system (orange dashed lines in Fig. 4). However, in $M_o(\omega; \Delta t|x_0) = M_o(\omega, \eta_o; \Delta t|x_0)$, one additionally approximates η_L by η_o. The difference between η_L and η_o is much larger than the difference between $\Delta\eta_L$ and $\Delta\eta_o$, and this additional approximation leads to the distorted reweighting results we observed as the blue lines in Fig. 4.

The proportions in Fig. 5(a) are not exaggerated. The black line in Fig. 5(c) shows the average absolute difference between the random numbers $\langle|\eta_{o,k} - \eta_{L,k}|\rangle$ as a function of $\xi\Delta t$. Visually, this is the difference between the green solid arrow and the blue solid arrow in Fig. 5(a). The orange line in Fig. 5(c) shows the average absolute difference between the random number differences $\langle|\Delta\eta_{o,k} - \Delta\eta_{L,k}|\rangle$, i.e., the difference between the orange and the red line in Fig. 5(a). The graph has been calculated by averaging over a path with 10^6 time steps. The standard deviations are shown as vertical bars. $\langle|\Delta\eta_{o,k} - \Delta\eta_{L,k}|\rangle$ is close to 0 for all values of $\xi\Delta t$, whereas there is a substantial difference between η_L and η_o. $\langle|\eta_{o,k} - \eta_{L,k}|\rangle$ has a minimum at $\xi\Delta t \approx 2$ because the difference between the Euler–Maruyama scheme and the ISP scheme is minimal for $\xi\Delta t \approx 2$ (see Sec. V B). Figure 5(d) shows the corresponding average relative errors. For $\xi\Delta t > 1$, $\langle|(\eta_o - \eta_L)/\eta_L|\rangle$ (black line) decreases in accordance with the decrease in the absolute difference $\langle|(\eta_o - \eta_L)|\rangle$ and $\langle|(\Delta\eta_o - \Delta\eta_L)/\Delta\eta_L|\rangle$ (orange line) increases, reflecting the fact that the approximation [Eq. (31)] does not hold for $\xi\Delta t > 1$. However, for $\xi\Delta t < 1$, the region in which MD simulations are conducted, the relative error for the random numbers is much larger than the relative error for the random number difference. This reinforces that the random numbers $\eta_{L,k}$ should not be approximated in the path probability ratio but, instead, should be recorded from the simulation at $V(x)$. By contrast, the random number difference $\Delta\eta_{L,k}$ can reliably be approximated by Eq. (31).

VII. MOLECULAR EXAMPLE: BUTANE

The slowest degree of freedom in butane is the torsion around the C_2–C_3 bond, which exhibits three metastable states: the trans-conformation at $\phi = \pi$ and the two gauche-conformations at $\phi = \pm\frac{1}{3}\pi$. Consequently, butane has three dominant MSM eigenvectors, where l_1 corresponds to the stationary density and l_2 and l_3 represent slow transitions along ϕ [Fig. 6(a)]. Because the two gauche-conformations are equally populated, l_2 and l_3 are degenerate [Fig. 6(b)]. We simulated butane in implicit water at three different temperatures, $T = 300$ K, $T = 200$ K, and $T = 150$ K, using direct and biased simulations. As we lower the temperature, we expect that the relative population of the trans-conformation increases, but that otherwise, the overall shape and sign-structure of the eigenvectors remain unchanged.

At $T = 300$ K and $T = 200$ K, the reweighting results using $M_{\mathrm{approx}}(\omega, \eta_L; \Delta t|x_0)$ [Eq. (32), orange dashed line] or $M_L(\omega, \eta_L; \Delta t|(x_0, v_0))$ [Eq. (23), green solid line] match the MSM obtained by direct simulation. In particular, the eigenvectors are reproduced with very high precision. By contrast, the reweighted results using $M_o(\omega; \Delta t|x_0)$ [Eq. (14), blue line] deviate considerably from the reference MSMs obtained by direct simulations. The stationary distribution l_1 is not reproduced correctly, which then leads to further errors in the dominant eigenvectors l_2 and l_3. The associated implied timescales are underestimated. Moreover, for $T = 200$ K and $T = 150$ K, the use of $M_o(\omega; \Delta t|x_0)$ yielded numerically instable transition matrices for lag times of $\tau > 100$ ps. This demonstrates that path reweighting with an appropriate path probability ratio, such as $M_{\mathrm{approx}}(\omega, \eta_L; \Delta t|x_0)$ or $M_L(\omega, \eta_L; \Delta t|(x_0, v_0))$, yields accurate results. However, $M_o(\omega; \Delta t|x_0)$ should not be used as an approximation for the exact path probability ratio $M_L(\omega, \eta_L; \Delta t|(x_0, v_0))$.

Note that reweighting results using the approximate probability ratio $M_{\mathrm{approx}}(\omega, \eta_L; \Delta t|x_0)$ are virtually indistinguishable from the results using the exact probability ratio $M_L(\omega, \eta_L; \Delta t|(x_0, v_0))$ for all three temperatures. This confirms our analysis that $M_{\mathrm{approx}}(\omega, \eta_L; \Delta t|x_0)$ can be used as highly accurate approximation to $M_L(\omega, \eta_L; \Delta t|(x_0, v_0))$.

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

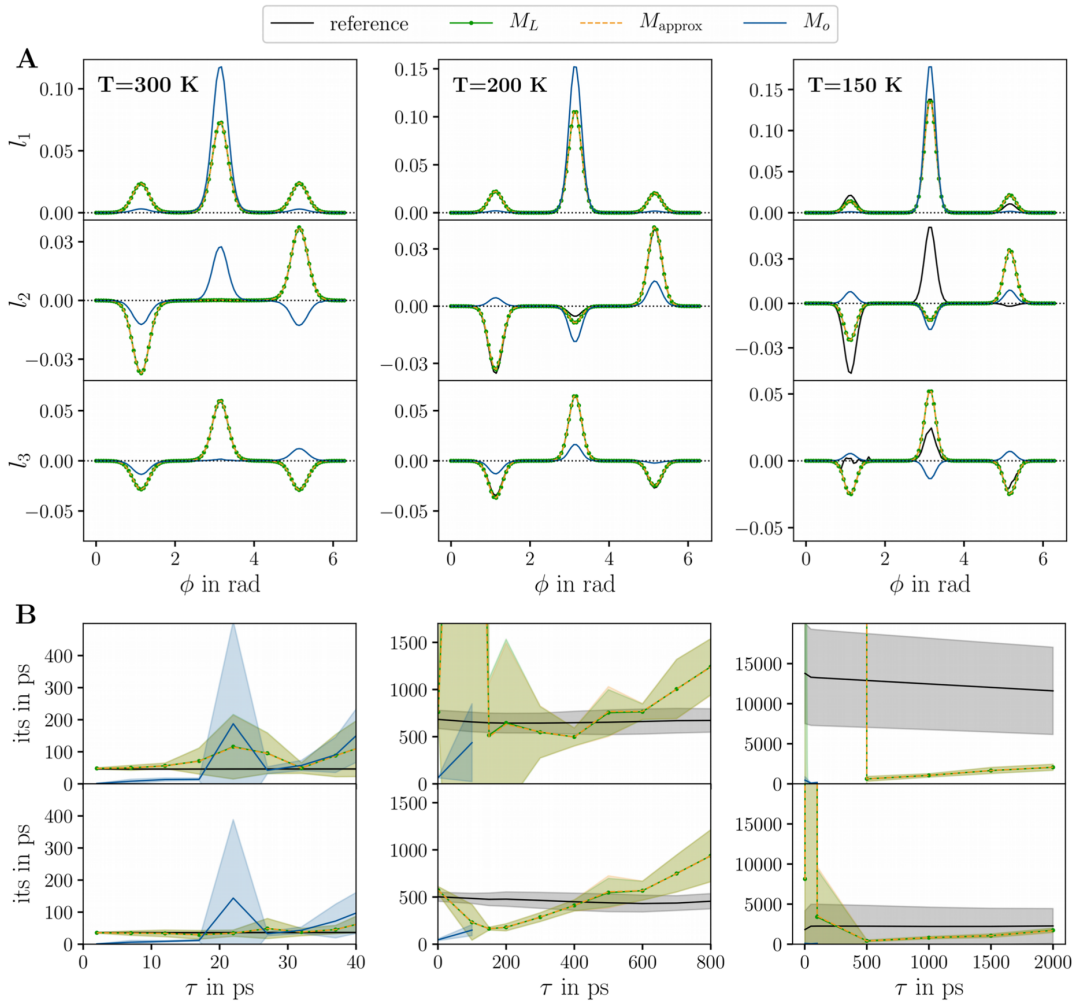


FIG. 6. Dynamics of the torsion angle in butane at T = 300 K, 200 K, and 150 K. (a) Dominant left eigenfunctions l_1, l_2, and l_3 of the MSM along the torsion angle ϕ, obtained by evaluating direct simulations at the target potential, as well as by reweighting biased simulations. (b) Implied timescales corresponding to l_2 and l_3 in (a). Solid lines: mean and shaded area: standard deviation. Standard deviations for the eigenvectors are too small to be shown.

The variation of the temperature from 300 K to 200 K and 150 K illustrates under which circumstances path reweighting is an efficient method. At T = 300 K, many transitions across the torsion angle barriers are observed in the direct simulation. Path reweighting and direct simulation yield identical results. However, path reweighting has a larger statistical uncertainty. At T = 200 K, fewer transitions are observed in the direct simulations, which results in an increased statistical uncertainty in the direct MSM. Finally, at T = 150 K, the transitions in the direct simulation are insufficient to correctly sample the stationary density. The MSM of the direct simulation predicts a higher population for the gauche-conformation at $\phi = +\frac{1}{3}\pi$ than for the gauche-conformation at $\phi = -\frac{1}{3}\pi$, which is clearly a sampling error. This error in the stationary density then leads to vastly incorrect estimates for l_2 and l_3. Additionally, the direct MSM predicts that the degeneracy is lifted. By contrast, the results of the reweighted MSM are in line with what we expect: the gauche-conformations are equally populated, the overall shapes of the dominant eigenvectors correspond to those of the eigenvectors at higher temperatures, and l_2 and l_3 are degenerate. In conclusion, path reweighting in combination with enhanced sampling techniques is a promising tool in situations, where the stationary density cannot be sampled accurately by direct simulation.

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

VIII. METHODS

A. Simulations of the test system

The test system is a one-dimensional one particle system with mass $m = 1$ kg and $k_BT = 2.494$ J (corresponding to $k_B = 0.008\,314$ J/K and $T = 300$ K). The simulation potential (orange line in Fig. 1) is

$$V(x) = (x^2 - 1)^2, \tag{39}$$

and the target potential (black line in Fig. 1) is

$$\widetilde{V}(x) = 4\left(x^3 - \frac{3}{2}x\right)^2 - x^3 + x. \tag{40}$$

For the results in Figs. 3–5, we simulated the system using the ISP scheme [Eqs. (20) and (21)] with a time step of $\Delta t = 0.01$ s. The initial conditions were $x_0 = 1.50$ m, $v_0 = 0$ m/s. The number of time steps N_t, the collision rate ξ, and the potential energy function used are summarized in Table II.

In Fig. 3(a), we computed the acceleration $\ddot{x} = a$ as $a_{k+1} = \frac{v_{k+1} - v_k}{\Delta t}$. Figure 3(b) displays the first ten steps of the simulation as an example path ω, and all quantities displayed in Figs. 3(c)–3(e) are calculated from this short path.

The absolute and relative differences of the random numbers in Fig. 5 were calculated as

$$\langle|\eta_{o,k} - \eta_{L,k}|\rangle = \frac{1}{N_t - 1}\sum_{k=0}^{N_t-1}|\eta_{o,k} - \eta_{L,k}| \tag{41}$$

and

$$\left\langle\left|\frac{\eta_{o,k} - \eta_{L,k}}{\eta_{L,k}}\right|\right\rangle = \frac{1}{N_t - 1}\sum_{k=0}^{N_t-1}\left|\frac{\eta_{o,k} - \eta_{L,k}}{\eta_{L,k}}\right|. \tag{42}$$

Analogous equations were used for $\langle|\Delta\eta_{o,k} - \Delta\eta_{L,k}|\rangle$ and $\langle|(\Delta\eta_{o,k} - \Delta\eta_{L,k})/\Delta\eta_{L,k}|\rangle$. $\eta_{L,k}$ was recorded during the simulation. We used Eq. (26) to calculate $\Delta\eta_{L,k}$, Eq. (15) to calculate $\eta_{o,k}$, and Eq. (17) to calculate $\Delta\eta_{o,k}$.

The reference MSM in Figs. 4(b) and 4(c) has been constructed from the simulation at the target potential $\widetilde{V}(x)$. The state space has been discretized using a regular grid of 100 microstates $(S_1, \ldots, S_{100})$ in the range $-1.7 \le x \le 1.6$. Transition counts between microstates were calculated as

$$c_{ij}(\tau) = \frac{1}{N_t - \tau}\sum_{k=0}^{N_t-\tau}\chi_i(x_k)\chi_j(x_{k+\tau}), \tag{43}$$

TABLE II. Simulation parameters.

Figures	N_t	ξ	Potential
3(a)	10^5	50 s^{-1}	$V(x)$
4(b) and 4(c)	10^7	50 s^{-1}	$V(x)$
4(b) and 4(c)	10^7	50 s^{-1}	$\widetilde{V}(x)$
5(c) and 5(d)	10^7	0.1 s^{-1}–1000 s^{-1}	$V(x)$

with

$$\chi_i(x) = \begin{cases} 1 & \text{if } x \in S_i \\ 0 & \text{else,} \end{cases} \tag{44}$$

where x_k is the trajectory and lag time $\tau = 200$ steps. The resulting count matrix $\mathbf{C}(\tau)$ was symmetrized as $\mathbf{C}(\tau) + \mathbf{C}^\top(\tau)$ to enforce detailed balance and row-normalized to obtain the MSM transition matrix $\mathbf{T}(\tau)$. The dominant MSM eigenvectors l_i and associated eigenvalues $\lambda_i(\tau)$ were calculated from $\mathbf{T}(\tau)$ using a standard eigenvalue solver, and the implied timescales were calculated as $t_i = -\tau/\ln(\lambda_i(\tau))$.

The reweighted MSMs in Figs. 4(b) and 4(c) have been constructed from the simulation at the simulation potential $V(x)$ using the same grid and lag time as for the reference MSM. Transition counts between microstates were counted and reweighted as[34,35]

$$\widetilde{c}_{ij}(\tau) = \frac{1}{N_t - \tau}\sum_{k=0}^{N_t-\tau} W((x_k, x_{k+1}, \ldots, x_{k+\tau}); \Delta t|(x_k, v_k))\,\chi_i(x_k)\chi_j(x_{k+\tau}). \tag{45}$$

The weight W is defined as

$$W((x_k, x_{k+1}, \ldots, x_{k+\tau}); \Delta t|(x_k, v_k)) = g(x_k)\cdot M((x_k, x_{k+1}, \ldots, x_{k+\tau}); \Delta t|(x_k, v_k)), \tag{46}$$

with M being the path probability ratio [Eq. (3)] and g being

$$g(x_k) = \exp\left(-\frac{U(x_k)}{k_BT}\right), \tag{47}$$

where the perturbation U is defined in Eq. (2). The remaining procedure was analogous to the reference MSM.

B. Butane—Direct simulations

We performed all-atom MD simulations of n-butane in implicit water using the OpenMM 7.4.1[54] simulation package. The GAFF (Generalized Amber Force Field)[60] was used to model butane, and the GBSA (Generalized Born Surface Area) model[61] was used to model implicit water. Interactions beyond 1 nm were truncated. The trajectory was propagated according to the ISP integration scheme for a $3N$-dimensional system,

$$x^i_{k+1} = x^i_k + \exp(-\xi\,\Delta t)\,v^i_k\Delta t - [1 - \exp(-\xi\,\Delta t)]\,\frac{\nabla_i V(\mathbf{x}_k)}{\xi m_i}\Delta t + \sqrt{\frac{k_BT}{m_i}[1 - \exp(-2\xi\,\Delta t)]}\,\eta^i_{L,k}\,\Delta t, \tag{48}$$

$$v^i_{k+1} = \frac{x^i_{k+1} - x^i_k}{\Delta t}, \tag{49}$$

with $i = 1, 2, \ldots, 3N$ and N being the number of atoms. x^i_k, v^i_k, and η^i_k are the position, velocity, and random number along dimension i at iteration step k, m_i is the mass of dimension i, and $\nabla_i V(\mathbf{x}_k)$ denotes the gradient of $V(\mathbf{x}_k)$ along dimension i measured at the position $\mathbf{x}_k$, with $\mathbf{x} \in \mathbb{R}^{3N}$. We implemented the ISP integration scheme using

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

the simtk.openmm.openmm.CustomIntegrator[62] class of OpenMM. The collision rate was $\xi = 10\ \mathrm{ps}^{-1}$. The simulation time step was $\Delta t = 0.002$ ps. Positions were written to disk every `txout` = 50 steps = 0.1 ps. We generated three trajectories with 500 ns each at $T = 300$ K, $T = 200$ K, and $T = 150$ K. These direct simulations correspond to simulations at the target potential $\widetilde{V}(\mathbf{x})$.

For the analysis, we cut each trajectory into five pieces of length 100 ns. For each 100-ns-trajectory, we constructed a MSM following the procedure outlined in Sec. VIII A. As state space we chose the C_2–C_3 dihedral angle ϕ, which we discretized using a regular grid of 100 microstates in the range $0 \le \phi \le 2\pi$. This resulted in five MSMs for each temperature. Figure 6 shows the mean and the standard deviation of the first three left MSM eigenvectors (evaluated at lag time $\tau = 1$ ps) and the mean and the standard deviations of the associated implied timescale.

C. Butane—Path reweighting

We biased the simulations along the C_2–C_3 dihedral angle ϕ. To generate the bias potential $U(\phi)$, we constructed a histogram of the free-energy function $\widetilde{F}(\phi)$,

$$\widetilde{F}(\phi) = -k_B T \ln(\widetilde{p}(\phi)), \tag{50}$$

where $\widetilde{p}(\phi)$ is the stationary density along ϕ as measured from the 500 ns direct simulations at $T = 300$ K. Fitting the histogram with a third order Fourier series yielded

$$\begin{aligned}\widetilde{F}_{300\,\mathrm{K}}(\phi) = {} & 8.985 + 3.122\cos(\omega\phi) + 0.959\cos(2\omega\phi) \\ & + 7.742\cos(3\omega\phi) + 0.095\sin(\omega\phi) \\ & + 0.047\sin(2\omega\phi) + 0.002\sin(3\omega\phi),\end{aligned} \tag{51}$$

with $\omega = 0.989$. The same procedure for the simulation at $T = 200$ K yielded

$$\begin{aligned}\widetilde{F}_{200\,\mathrm{K}}(\phi) = {} & 8.311 + 2.847\cos(\omega\phi) + 0.841\cos(2\omega\phi) \\ & + 7.697\cos(3\omega\phi) + 0.046\sin(\omega\phi) \\ & + 0.026\sin(2\omega\phi) + 0.004\sin(3\omega\phi),\end{aligned} \tag{52}$$

with $\omega = 0.989$. $\widetilde{F}_{300\,\mathrm{K}}(\phi)$ and $\widetilde{F}_{200\,\mathrm{K}}(\phi)$ are almost identical. The simulation at $T = 150$ K did not yield a converged stationary density, and thus, no free-energy function was constructed for this temperature, and instead, $\widetilde{F}_{300\,\mathrm{K}}(\phi)$ was used.

The biased simulations were carried out with the potential

$$V_\alpha(\mathbf{x}) = \widetilde{V}(\mathbf{x}) - \alpha \cdot \widetilde{F}(\phi(\mathbf{x})), \tag{53}$$

where $\widetilde{V}(\mathbf{x})$ is the target potential and $\alpha \in [0, 1]$ specifies the bias strength. $V_\alpha(\mathbf{x})$ corresponds to the "simulation potential" within the terminology of this paper; thus,

$$U(\phi(\mathbf{x})) = \alpha \cdot \widetilde{F}(\phi(\mathbf{x})). \tag{54}$$

α was set to 0.1 in all biased simulations, corresponding to "10% of the full metadynamics potential." We carried out biased simulations at three temperatures $T = 300$ K, $T = 200$ K, and $T = 150$ K, with bias potentials $U_{300\,\mathrm{K}}(\phi) = 0.1 \cdot \widetilde{F}_{300\,\mathrm{K}}(\phi)$, $U_{200\,\mathrm{K}}(\phi) = 0.1 \cdot \widetilde{F}_{200\,\mathrm{K}}(\phi)$, and $U_{150\,\mathrm{K}}(\phi) = 0.1 \cdot \widetilde{F}_{300\,\mathrm{K}}(\phi)$. All other simulation parameters were as described in Sec. VIII B.

The path probability ratios for the biased simulations were calculated on-the-fly[34,35] and were written to disk at the same frequency `txout` as the positions. For the approximate path probability ratio M_{approx}, we calculated

$$\begin{aligned}\mathbb{M}_{\mathrm{approx}}(b) = \sum_{i=1}^{3N} \sum_{k=(b-1)\cdot\mathrm{txout}}^{b\cdot\mathrm{txout}-1} \Bigg(& -\sqrt{\frac{\Delta t}{2k_B T \xi m_i}} \nabla_i U(\mathbf{x}_k)\, \eta_{L,k}^i \\ & - \frac{\Delta t}{4k_B T \xi m_i} \left(\nabla_i U(\mathbf{x}_k)\right)^2 \Bigg)\end{aligned} \tag{55}$$

and constructed the complete path probability ratio as

$$M_{\mathrm{approx}}(\boldsymbol{\omega}, \boldsymbol{\eta}_L; \Delta t | \mathbf{x}_0) = \exp\left(\sum_{b=1}^{A} \mathbb{M}_{\mathrm{approx}}(b) \right) \tag{56}$$

during the construction of the MSM, where $A \in \mathbb{N}$ such that $\tau = A \cdot \mathtt{txout} \cdot \Delta t$.

For the Langevin path probability ratio M_L, we calculated the terms

$$\mathbb{M}_{L,1}(b) = \sum_{i=1}^{3N} \sum_{k=(b-1)\cdot\mathrm{txout}}^{b\cdot\mathrm{txout}-1} (x_{k+1}^i - x_k^i)\left(\nabla_i \widetilde{V}(\mathbf{x}_k) - \nabla_i V(\mathbf{x}_k)\right), \tag{57}$$

$$\mathbb{M}_{L,2}(b) = \sum_{i=1}^{3N} \sum_{k=(b-1)\cdot\mathrm{txout}}^{b\cdot\mathrm{txout}-1} v_k^i \left(\nabla_i \widetilde{V}(\mathbf{x}_k) - \nabla_i V(\mathbf{x}_k)\right), \tag{58}$$

$$\mathbb{M}_{L,3}(b) = \sum_{i=1}^{3N} \sum_{k=(b-1)\cdot\mathrm{txout}}^{b\cdot\mathrm{txout}-1} \frac{\left(\left(\nabla_i \widetilde{V}(\mathbf{x}_k)\right)^2 - \left(\nabla_i V(\mathbf{x}_k)\right)^2\right)}{m_i} \tag{59}$$

and constructed the complete path probability ratio as

$$\begin{aligned}& M_L(\boldsymbol{\omega}, \Delta t | \mathbf{x}_0) \\ & = \exp\left[\sum_{b=1}^{A} \left(- \frac{\mathbb{M}_{L,1}(b)}{k_B T \xi (1 + \exp(-\xi \Delta t)) \Delta t} \right.\right. \\ & \quad \left.\left. + \frac{\mathbb{M}_{L,2}(b)}{k_B T \xi (1 + \exp(\xi \Delta t))} - \frac{\exp(\xi \Delta t) - 1}{\exp(\xi \Delta t) + 1} \frac{\mathbb{M}_{L,3}(b)}{2 k_B T \xi^2} \right) \right]\end{aligned} \tag{60}$$

during the construction of the MSM, where $A \in \mathbb{N}$ such that $\tau = A \cdot \mathtt{txout} \cdot \Delta t$.

For the overdamped Langevin path probability ratio M_o, we calculated the terms

$$\mathbb{M}_{o,1}(b) = \sum_{i=1}^{3N} \sum_{k=(b-1)\cdot\mathrm{txout}}^{b\cdot\mathrm{txout}-1} (x_{k+1}^i - x_k^i)\left(\nabla_i \widetilde{V}(\mathbf{x}_k) - \nabla_i V(\mathbf{x}_k)\right), \tag{61}$$

$$\mathbb{M}_{o,2}(b) = \sum_{i=1}^{3N} \sum_{k=(b-1)\cdot\mathrm{txout}}^{b\cdot\mathrm{txout}-1} \frac{\left(\left(\nabla_i \widetilde{V}(\mathbf{x}_k)\right)^2 - \left(\nabla_i V(\mathbf{x}_k)\right)^2\right)}{m_i} \tag{62}$$

and constructed the complete path probability ratio as

$$M_o(\boldsymbol{\omega}, \Delta t | \mathbf{x}_0) = \exp\left[\sum_{b=1}^{A} \left(- \frac{\mathbb{M}_{o,1}(b)}{2 k_B T} - \frac{\mathbb{M}_{o,2}(b)\, \Delta t}{4 k_B T \xi} \right) \right] \tag{63}$$

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

during the construction of the MSM, where $A \in \mathbb{N}$ such that $\tau = A \cdot \texttt{txout} \cdot \Delta t$.

For the analysis, we cut each trajectory into five pieces of length 100 ns. For each 100-ns-trajectory, we constructed a MSM following the procedure outlined in Sec. VIII A. As state space we chose the C_2–C_3 dihedral angle ϕ, which we discretized using a regular grid of 100 microstates in the range $0 \le \phi \le 2\pi$. Transition counts between microstates were counted and reweighted as described in Eq. (45) with $x_k = \phi_k$ and

$$g(\phi_k) = \exp\left(-\frac{U(\phi_k)}{k_B T}\right) = \exp\left(-\frac{0.1 \cdot \widetilde{F}(\phi_k)}{k_B T}\right), \tag{64}$$

where ϕ_k is the first entry in the path of length τ. This resulted in five reweighted MSMs for each temperature. Figure 6 shows the mean and the standard deviation of the first three left MSM eigenvectors (evaluated at lag time $\tau = 1$ ps) and the mean and the standard deviations of the associated implied timescale.

Example scripts for simulation and analysis are included as the supplementary material.

IX. CONCLUSION AND OUTLOOK

We have presented two strategies to derive the path probability ratio M_L for the ISP scheme. In the first strategy, the correctly normalized path probability is derived by integrating out the random number η_k from the one-step transition probability. In the second strategy, the equations for the ISP scheme are solved for η_k, and the resulting transformation is used as a change in variables on the Gaussian probability density of the random numbers. This yields an unnormalized path probability. The path probability ratio M_L is then calculated as the ratio between the path probability at the target potential $\widetilde{P}_L(\omega_L; \Delta t|(x_0, v_0))$ and the path probability at the simulation potential $P_L(\omega_L; \Delta t|(x_0, v_0))$.

With M_L, we are now able to perform exact path reweighting for trajectories generated by the ISP integration scheme. Moreover, the two strategies serve as a blueprint for deriving path probability ratios for other Langevin integration schemes, which use Gaussian white noise.[44–47,49–53] Thus, path reweighting can now readily be applied to MD simulation conducted at the NVT ensemble thermostatted with a stochastic thermostat.

We compared the approximate path probability ratio M_{approx} that we used in earlier publications[34,35] to the exact path probability ratio M_L, both analytically and numerically. We showed that the two expressions only differ by $\mathcal{O}(\xi^4 \Delta t^4)$. Thus, M_{approx} is an excellent approximation to M_L for Langevin MD simulations. To understand why the approximation is so good, we showed that the random number η_k needed to generate a given step $x_k \to x_{k+1}$ is highly dependent on the integration scheme. However, $\Delta\eta_k$, the difference between the random number $\widetilde{\eta}_k$ at $\widetilde{V}(x)$ and the random number η_k at $V(x)$, has about the same value in the ISP scheme and in the Euler–Maruyama scheme.

In M_{approx}, one uses the random numbers directly recorded during the simulation at $V(x)$, which does not introduce any error and approximates $\Delta\eta_k$ by the expression from the Euler–Maruyama scheme $\Delta\eta_{o,k}$ to construct $\widetilde{\eta}_k$.

We have chosen the ISP algorithm for the present analysis in order to be consistent with our previous work.[34,35] However, the same strategy can be used to derive the path probability ratio for other Langevin integrators.[44–47,49–53] Specifically, solve the integrator equations for the random number η_k; from there, derive an expression for $\Delta\eta_k$, record η_k during the simulation at $V(x)$ and calculate $\Delta\eta_k$ on the fly, and insert η_k and $\Delta\eta_k$ into Eq. (7). For a large application of path reweighting, using a modern Langevin integrator is likely worthwhile, such as the the BAOAB method[51] (or alternatively the VRORV method[53]). This method is exceptionally efficient at sampling the configurational stationary distribution, which allows for increasing the time step.[51,53]

It is tempting to speculate that $\Delta\eta_k$ for other Langevin integration schemes could also have about the same value as $\Delta\eta_{o,k}$ for the Euler–Maruyama scheme. This would open up a route to a general approximate path probability ratio $M_{\approx}$ and would eliminate the problem that the path probability needs to be adapted for each integration scheme. On the other hand, the structure of the ISP scheme is closer to that of the Euler–Maruyama scheme than most other Langevin integrators. Whether the approximate path probability can indeed be generalized to these integrators is, therefore, not yet obvious and needs to be checked carefully.

Our one-dimensional test system and our molecular system showed that the accuracy of the reweighting sensitively depends on an accurate representation of η_k in the path probability ratio. For example, reweighting a Langevin path by the path probability ratio for the Euler–Maruyama scheme yielded very distorted results. Neither the MSM eigenvectors nor the implied timescales were reproduced correctly. It is, however, possible that the distortion is less severe in the limit of infinite sampling of the combined space of molecular states and random numbers (probably less relevant to actual applications) or if the dynamics is projected onto a reaction coordinate before the reweighted dynamical properties are evaluated (probably very relevant to actual applications).

We used path reweighting to reweight MSMs. The dynamical property that is reweighted to estimate a transition probability is a correlation function. It is important to point out that correlation functions are a combination of path ensemble averages, where the path is conditioned on a particular initial state (x_0, v_0) and a phase-space ensemble average for the initial states. Thus, the total reweighting factor for MSMs is combined of the path probability ratio M for the path ensemble average and the Boltzmann probability ratio for the phase-space ensemble average $g(x)$ [Eq. (47)].[27,32–34] Even though the reweighting of the path ensemble average can be made exact, by averaging over the initial states within a microstate, one assumes local equilibrium within this microstate.[23] Beyond local equilibrium, the formalism has been extended to reweighting transition probabilities from non-equilibrium steady-state simulations.[63]

When is the combination of enhanced sampling and path reweighting more efficient than a direct simulation? This depends on the uncertainty of the transition counts estimated from a direct simulation [Eq. (43)] compared to the uncertainty of the reweighted transition counts [Eq. (45)]. The molecular example demonstrated that path reweighting is particularly useful if the stationary density cannot be sampled accurately by direct simulation with the available computer resources. Furthermore, the efficiency of path reweighting increases if the number of transitions at the enhanced sampling simulation is large compared to the direct simulation and if the weights $W = g \cdot M$ are not too small. The path probability ratio M decreases

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

with the path length τ and with the dimensionality of the bias potential U. The path length is kept short by combining path reweighting with MSMs and can be further limited by using advanced MSM discretization techniques.[64–66] The bottleneck for the dimensionality U already occurs at the stage of sampling because most enhanced sampling techniques[10] are limited to very low-dimensional biases in practice. Note that increasing the dimensionality of the overall system does not lower the efficiency of the path reweighting. The question of how strong the bias should be is more difficult to answer. Strong biases increase the transitions in the biased simulation but reduce both g and M. In Ref. 35, we empirically found that a bias of ca. 10% of the full metadynamics biasing potential yielded optimal results, but this will likely depend on the system. Here, we have restricted ourselves to systems with low barriers in the order of k_BT so that we could generate reference solutions by direct simulation. However, we believe that path reweighting is most useful for systems with large barriers that cannot be sampled by direct simulation. An example is the β-hairpin folding equilibrium in Ref. 35.

Path reweighting is closely related to path sampling techniques, in particular path sampling techniques that aim at optimizing the path action.[67–70] The combination of enhanced sampling, path sampling, and path reweighting might change the way we explore the molecular state space and investigate rare events.

SUPPLEMENTARY MATERIAL

See the supplementary material for an example OpenMM script and the corresponding Python3 scripts to construct a reweighted MSM.

DEDICATION

This paper is dedicated to Dr. Irina V. Gopich, a master of stochastic processes. Her work has influenced the way scientists in the field think about the dynamics of molecules—in simulation and in experiment.

ACKNOWLEDGMENTS

The authors would like to thank Luca Donati and Marcus Weber for helpful comments on this manuscript. This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 2008—390540038—UniSysCat and through grant CRC 1114 "Scaling Cascades in Complex Systems," Project Number 235221301, Project B05" Origin of scaling cascades in protein dynamics."

APPENDIX A: LANGEVIN LEAPFROG AND THE ISP SCHEME

Izaguirre, Sweet, and Pande developed the following Langevin Leapfrog algorithm:

$$v_{k+\frac{1}{2}} = \exp\left(-\xi\frac{\Delta t}{2}\right)v_k - \left[1-\exp\left(-\xi\frac{\Delta t}{2}\right)\right]\frac{\nabla V(x_k)}{\xi m} + \sqrt{\frac{k_BT}{m}\left[1-\exp(-\xi\Delta t)\right]}\,\eta_k, \tag{A1}$$

$$x_{k+1} = x_k + v_{k+\frac{1}{2}}\Delta t, \tag{A2}$$

$$v_{k+1} = \exp\left(-\xi\frac{\Delta t}{2}\right)v_{k+\frac{1}{2}} - \left[1-\exp\left(-\xi\frac{\Delta t}{2}\right)\right]\frac{\nabla V(x_{k+1})}{\xi m} + \sqrt{\frac{k_BT}{m}\left[1-\exp(-\xi\Delta t)\right]}\,\eta_{k+1} \tag{A3}$$

[Eqs. (14)–(16) in Ref. 49]. First, the velocity $v_{k+\frac{1}{2}}$ is updated by a half step using v_k, x_k, and a random number η_k [Eq. (A1)]. Then, the position update to x_{k+1} is computed from x_k, assuming a constant velocity $v_{k+\frac{1}{2}}$ in the interval $[k, k+1]$ [Eq. (A2)]. Finally, the remaining half step of the velocities to v_{k+1} is computed using x_{k+1}, $v_{k+\frac{1}{2}}$, and a new random number η_{k+1} [Eq. (A3)].

This Langevin Leapfrog algorithm has been converted to the following full-step scheme in the C^{++} CpuLangevinDynamics class of OpenMM:[71]

$$v_{k+1} = \exp(-\xi\Delta t)v_k - \left[1-\exp(-\xi\Delta t)\right]\frac{\nabla V(x_k)}{\xi m} + \sqrt{\frac{k_BT}{m}\left[1-\exp(-2\xi\Delta t)\right]}\,\eta_k, \tag{A4}$$

$$x_{k+1} = x_k + v_{k+1}\Delta t, \tag{A5}$$

where the velocities are propagated by a full step [i.e., $\Delta t/2$ in Eq. (A1) is replaced by Δt, and Δt in Eq. (A1) is replaced by $2\Delta t$] and the position update is based on v_k rather than on $v_{k+\frac{1}{2}}$. The second half step for the velocities [Eq. (A3)] is omitted. This integration scheme only uses a single random number per iteration. Equations (A4) and (A5) are the integration scheme we used in Refs. 34 and 35. To distinguish it from the original Langevin Leapfrog scheme [Eqs. (A1)–(A3)], we will refer to Eqs. (A4) and (A5) as the "ISP scheme."

To be able to analyze the path probability as a function of the positions, we rearrange Eqs. (A4) and (A5) such that we first update the positions using a stochastic step [replace v_{k+1} in Eq. (A5) by Eq. (A4)] and then update the velocity as finite difference [rearrange Eq. (A5) with respect to v_{k+1}]. This yields Eqs. (20) and (21).

APPENDIX B: PATH PROBABILITY FOR LANGEVIN DYNAMICS

We derive the closed-form expression for $P_L(\omega_L; \Delta t|(x_0, v_0))$ in Eq. (22) from the integration scheme [Eqs. (20) and (21)] by following the approach in Ref. 57. As a first step, we derive a closed-form expression for the one-step probability $P_L(x_{k+1}, v_{k+1}; \Delta t|(x_k, v_k))$ of observing a step $(x_k, v_k) \to (x_{k+1}, v_{k+1})$. According to Eqs. (20) and (21), the tuple (x_{k+1}, v_{k+1}) at iteration step $k+1$ is entirely determined by the tuple (x_k, v_k) at iteration step k if additionally the random number η_k is known. Thus, $P_L(x_{k+1}, v_{k+1}; \Delta t|(x_k, v_k, \eta_k))$, i.e., the one-step probability with fixed random number η_k, is a Dirac delta function centered at (x_{k+1}, v_{k+1}). Our strategy is to derive a closed-form expression for this Dirac delta function

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

using Eqs. (20) and (21) and to integrate out the dependency on η_k. In this Appendix, we omit the index L in $\eta_{L,k}$ to simplify the notation.

We reformulate the two-dimensional probability $P_L(x_{k+1}, v_{k+1}; \Delta t|(x_k, v_k, \eta_k))$ as a product of two one-dimensional probabilities,

$$P_L(x_{k+1}, v_{k+1}; \Delta t|(x_k, v_k, \eta_k)) = P_L(v_{k+1}; \Delta t|(x_{k+1}, x_k, v_k, \eta_k)) \cdot P_L(x_{k+1}; \Delta t|(x_k, v_k, \eta_k)) \quad \text{(B1)}$$

using the rule $P(A, B|C) = P(A|B, C) \cdot P(B|C)$, with $A = v_{k+1}$, $B = x_{k+1}$, and $C = (x_k, v_k, \eta_k)$. This rule is the extension of the conditional probability $P(A, B) = P(A|B) \cdot P(B)$ to an additional condition C. The first factor is a Dirac delta function constrained to Eq. (21),

$$P_L(v_{k+1}; \Delta t|(x_{k+1}, x_k, v_k, \eta_k)) = P_L(v_{k+1}; \Delta t|(x_{k+1}, x_k)) = \delta\left(v_{k+1} - \frac{x_{k+1} - x_k}{\Delta t}\right), \quad \text{(B2)}$$

where the first equality emphasizes that v_{k+1} does not depend on η_k or v_k in Eq. (21). Note that the probability of the velocity v_{k+1} [Eq. (B2)] does not depend on a random number, which mirrors our previous observation that v_{k+1} is not treated as a random variable in Eq. (21). The second factor in Eq. (B1) is a Dirac delta function constrained to Eq. (20),

$$P_L(x_{k+1}; \Delta t|(x_k, v_k, \eta_k)) = \delta\Big(x_{k+1} - x_k - \exp(-\xi\,\Delta t)\, v_k \Delta t + [1 - \exp(-\xi\,\Delta t)]\, \frac{\nabla V(x_k)}{\xi m} \Delta t - \sqrt{\frac{k_B T}{m}\,[1 - \exp(-2\xi\,\Delta t)]}\, \eta_k \Delta t\Big). \quad \text{(B3)}$$

Reinserting the two factors into Eq. (B1) yields the desired closed-form expression for $P_L(x_{k+1}, v_{k+1}; \Delta t|(x_k, v_k; \eta_k))$. Since we know that the random numbers η_k are drawn from a Gaussian distribution $P(\eta_k)$ with zero mean and unit variance

$$P(\eta_k) = N^{-1} \exp\left(-\frac{\eta_k^2}{2}\right), \qquad N = \sqrt{2\pi}, \quad \text{(B4)}$$

we can average out the random number dependency in Eq. (B1) to obtain the one-step probability,

$$\begin{aligned} &P_L(x_{k+1}, v_{k+1}; \Delta t|(x_k, v_k)) \\ &= \int_{-\infty}^{\infty} \mathrm{d}\eta_k P(\eta_k)\, P_L(x_{k+1}, v_{k+1}; \Delta t|(x_k, v_k, \eta_k)) \\ &= \delta\left(v_{k+1} - \frac{x_{k+1} - x_k}{\Delta t}\right) \\ &\quad \cdot \int_{-\infty}^{\infty} \mathrm{d}\eta_k P_\eta(\eta_k)\, P_L(x_{k+1}; \Delta t|(x_k, v_k, \eta_k)). \end{aligned} \quad \text{(B5)}$$

The challenge lies in solving the integral in this equation. The solution, which is detailed in Appendix C, yields the closed-form expression for the one-step probability,

$$P_L(x_{k+1}, v_{k+1}; \Delta t|(x_k, v_k)) = \delta\left(v_{k+1} - \frac{x_{k+1} - x_k}{\Delta t}\right) \cdot \sqrt{\frac{m}{2\pi k_B T \Delta t^2 (1 - \exp(-2\xi\Delta t))}} \times \exp\left(-\frac{m\left(x_{k+1} - x_k - \exp(-\xi\Delta t) v_k \Delta t + (1 - \exp(-\xi\Delta t))\frac{\nabla V(x_k)}{\xi m}\Delta t\right)^2}{2k_B T(1 - \exp(-2\xi\Delta t))\Delta t^2}\right). \quad \text{(B6)}$$

Applying the Chapman–Kolmogorov equation[72] recursively to the one-step probability yields the closed-form expression for the path probability $P_L(\omega_L; \Delta t|(x_0, v_0))$, shown in Eq. (22).

APPENDIX C: SOLVING THE DOUBLE INTEGRAL

We compute the integral

$$P_L(x_{k+1}; \Delta t|(x_k, v_k)) = \int_{-\infty}^{\infty} \mathrm{d}\eta_k\, P(\eta_k)\, P_L(x_{k+1}; \Delta t|(x_k, v_k, \eta_k)) \quad \text{(C1)}$$

from Eq. (B5). First, we replace $P(\eta_k)$ according to Eq. (B4). Second, we substitute $P_L(x_{k+1}; \Delta t|(x_k, v_k, \eta_k))$, which is a δ-function [Eq. (B3)], with its Fourier transform

$$\delta(z - z') = \int_{-\infty}^{+\infty} \frac{\mathrm{d}w}{2\pi} \exp\big(iw(z - z')\big), \quad \text{(C2)}$$

where $z = x_{k+1}$ and z' is equal to the right-hand side of Eq. (20). This yields a double integral whose outer integral is with respect to w,

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

while the inner integral is with respect to η_k,

$$\begin{aligned} &P_L(x_{k+1};\Delta t|(x_k,v_k)) \\ &= \int_{-\infty}^{+\infty}\frac{\mathrm{d}w}{2\pi}\int_{-\infty}^{+\infty}\frac{\mathrm{d}\eta_k}{N}\exp\left(-\frac{\eta_k^2}{2}\right) \\ &\quad\times\exp\left(iw\left[x_{k+1}-x_k-\exp(-\xi\Delta t)\,v_k\Delta t\right.\right. \\ &\quad+[1-\exp(-\xi\Delta t)]\,\frac{\nabla V(x_k)}{\xi m}\Delta t \\ &\quad\left.\left.-\sqrt{\frac{k_BT}{m}\,[1-\exp(-2\xi\Delta t)]}\,\eta_k\Delta t\right]\right) \\ &= \int_{-\infty}^{+\infty}\frac{\mathrm{d}w}{2\pi}\exp(iwB)\int_{-\infty}^{+\infty}\frac{\mathrm{d}\eta_k}{N}\exp\left(-\frac{\eta_k^2}{2}-iwR\,\eta_k\right), \end{aligned} \tag{C3}$$

where we moved all terms that do not depend on η_k out of the inner integral and defined the abbreviations,

$$\begin{aligned} B &= \left[x_{k+1}-x_k-\exp(-\xi\Delta t)\,v_k\Delta t\right. \\ &\quad\left.+[1-\exp(-\xi\Delta t)]\,\frac{\nabla V(x_k)}{\xi m}\Delta t\right], \\ R &= \Delta t\sqrt{\frac{k_BT}{m}\,[1-\exp(-2\xi\Delta t)]}. \end{aligned} \tag{C4}$$

Both integrals in Eq. (C3) can be solved with the completing-the-square technique for Gaussian integrals. The goal of this technique is to expand and rearrange the inner integral such that we can use the analytic solution

$$\int_{-\infty}^{\infty}\mathrm{d}x\,\exp\left(-a(x\pm b)^2\right)=\sqrt{\frac{\pi}{a}}\qquad\text{for } a,b\in\mathbb{R}. \tag{C5}$$

This can be achieved by a systematic step-to-step procedure that can be applied to all Gaussian integrals of this type,

$$\begin{aligned} &\int_{-\infty}^{+\infty}\frac{\mathrm{d}\eta_k}{N}\exp\left(-\frac{\eta_k^2}{2}-iwR\,\eta_k\right) \\ &= \int_{-\infty}^{+\infty}\frac{\mathrm{d}\eta_k}{N}\exp\left(-\frac{1}{2}\left[\eta_k^2+2iwR\,\eta_k\overbrace{+i^2w^2R^2-i^2w^2R^2}^{=0}\right]\right) \\ &= \exp\left(-\frac{w^2R^2}{2}\right)\int_{-\infty}^{+\infty}\frac{\mathrm{d}\eta_k}{N}\exp\left(-\frac{1}{2}(\eta_k+iwR)^2\right) \\ &= \exp\left(-\frac{w^2R^2}{2}\right)\frac{1}{N}\sqrt{2\pi} \\ &= \exp\left(-\frac{w^2R^2}{2}\right). \end{aligned} \tag{C6}$$

In the first line, we isolate η_k^2 by factoring out $-\frac{1}{2}$ and complete the first binomial formula by adding a zero. Then, we separate the exponent into the binomial formula and the term $\exp\left(-\frac{w^2R^2}{2}\right)$, which can be moved in front of the integral because it does not depend on η_k. In the third line, we solve the remaining integral using Eq. (C5), which can be further simplified by inserting the normalization constant of the Gaussian distribution: $N=\sqrt{2\pi}$.

Inserting Eq. (C6) into Eq. (C3) yields the outer integral

$$\int_{-\infty}^{+\infty}\frac{\mathrm{d}w}{2\pi}\exp(iwB)\exp\left(-\frac{w^2R^2}{2}\right)=\int_{-\infty}^{+\infty}\frac{\mathrm{d}w}{2\pi}\exp\left(-\frac{w^2R^2}{2}+iwB\right),$$

which is solved using the same procedure,

$$\begin{aligned} &\int_{-\infty}^{+\infty}\frac{\mathrm{d}w}{2\pi}\exp\left(-\frac{w^2R^2}{2}+iwB\right) \\ &= \int_{-\infty}^{+\infty}\frac{\mathrm{d}w}{2\pi}\exp\left(-\frac{R^2}{2}\left[w^2+\frac{2iwB}{R^2}+\overbrace{\frac{i^2B^2}{R^4}-\frac{i^2B^2}{R^4}}^{=0}\right]\right) \\ &= \exp\left(-\frac{B^2}{2R^2}\right)\int_{-\infty}^{\infty}\frac{\mathrm{d}w}{2\pi}\exp\left(-\frac{R^2}{2}\left(w+\frac{iB}{R^2}\right)^2\right) \\ &= \exp\left(-\frac{B^2}{2R^2}\right)\frac{1}{2\pi}\sqrt{\frac{2\pi}{R^2}} \\ &= \sqrt{\frac{1}{2\pi R^2}}\exp\left(-\frac{B^2}{2R^2}\right). \end{aligned} \tag{C7}$$

Inserting the expressions for the constants R and B [Eq. (C4)] yields

$$\begin{aligned} &P_L(x_{k+1};\Delta t|(x_k,v_k)) \\ &= \sqrt{\frac{m}{2\pi k_BT\Delta t^2(1-\exp(-2\xi\Delta t))}} \\ &\quad\times\exp\left(-\frac{m\left(x_{k+1}-x_k-\exp(-\xi\Delta t)v_k\Delta t+(1-\exp(-\xi\Delta t))\frac{\nabla V(x_k)}{\xi m}\Delta t\right)^2}{2k_BT(1-\exp(-2\xi\Delta t))\Delta t^2}\right). \end{aligned} \tag{C8}$$

This is inserted into Eq. (B5) to yield Eq. (B6).

APPENDIX D: PROOF OF EQ. (31)

$$\begin{aligned} \frac{\left(1-e^{-x}\right)^2}{x\cdot\left(1-e^{-2x}\right)} &= \frac{1}{2}-\frac{x^2}{24}+\frac{x^4}{240}\pm\mathcal{O}(x^5), \\ \left(1-e^{-x}\right)^2 &= \frac{x}{2}\cdot\left(1-e^{-2x}\right)-\frac{x^2}{24}\cdot x\cdot\left(1-e^{-2x}\right) \\ &\quad\pm\mathcal{O}(x^4)\cdot x\cdot\left(1-e^{-2x}\right), \\ \left(1-e^{-x}\right)^2 &= \frac{x}{2}\cdot\left(1-e^{-2x}\right)-\mathcal{O}(x^4). \end{aligned} \tag{D1}$$

The first line shows the Taylor expansion of the expression on the right-hand side. To obtain the second line, we multiplied by $x\cdot\left(1-e^{-2x}\right)$. In the third line, we used the fact that the leading term of the Taylor expansion of $x\cdot\left(1-e^{-2x}\right)$ is $2x^2$, thus yielding an error of $\mathcal{O}(x^4)$. Substituting $x=\xi\Delta t$ yields

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

$$\left(1-e^{-\xi\Delta t}\right)^2 = \frac{\xi\Delta t}{2}\cdot\left(1-e^{-2\xi\Delta t}\right) - \mathcal{O}(\xi^4\Delta t^4),$$
$$\left(1-e^{-\xi\Delta t}\right)^2 \approx \frac{\xi\Delta t}{2}\cdot\left(1-e^{-2\xi\Delta t}\right), \tag{D2}$$

and multiplying by $\frac{1}{k_B T\,\xi^2\, m\left(1-e^{-2\xi\Delta t}\right)}\left(\nabla U(x_k)\right)^2$ yields

$$\frac{1}{k_B T\xi^2 m}\frac{\left(1-e^{-\xi\Delta t}\right)^2}{1-e^{-2\xi\Delta t}}\left(\nabla U(x_k)\right)^2 \approx \frac{\Delta t}{2k_B T\xi m}\left(\nabla U(x_k)\right)^2, \tag{D3}$$
$$\Delta\eta_{L,k}^2 \approx \Delta\eta_{o,k}^2.$$

Thus, the difference between $\Delta\eta_{L,k}^2$ [Eq. (26)] and $\Delta\eta_{o,k}^2$ [Eq. (17)] is of order $\mathcal{O}(\xi^4\Delta t^4)$. Equation (D3) is Eq. (31) squared. □

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] I. V. Gopich, "Multisite reversible association in membranes and solutions: From non-Markovian to Markovian kinetics," J. Chem. Phys. **152**, 104101 (2020).
[2] T. J. Lane, D. Shukla, K. A. Beauchamp, and V. S. Pande, "To milliseconds and beyond: Challenges in the simulation of protein folding," Curr. Opin. Struct. Biol. **23**, 58 (2013).
[3] R. O. Dror, R. M. Dirks, J. P. Grossman, H. Xu, and D. E. Shaw, "Biomolecular simulation: A computational microscope for molecular biology," Annu. Rev. Biophys. **41**, 429 (2012).
[4] E. P. Barros, L. Casalino, Z. Gaieb, A. C. Dommer, Y. Wang, L. Fallon, L. Raguette, K. Belfon, C. Simmerling, and R. E. Amaro, "The flexibility of ACE2 in the context of SARS-CoV-2 infection," Biophys. J. **120**, 1 (2020).
[5] T. J. Harpole and L. Delemotte, "Conformational landscapes of membrane proteins delineated by enhanced sampling molecular dynamics simulations," Biochim. Biophys. Acta **1860**, 909 (2018).
[6] Z. Cournia, B. Allen, and W. Sherman, "Relative binding free energy calculations in drug discovery: Recent advances and practical considerations," J. Chem. Inf. Model. **57**, 2911 (2017).
[7] M. Badaoui, A. Kells, C. Molteni, C. J. Dickson, V. Hornak, and E. Rosta, "Calculating kinetic rates and membrane permeability from biased simulations," J. Phys. Chem. B **122**, 11571 (2018).
[8] J.-O. Joswig, J. Anders, H. Zhang, C. Rademacher, and B. G. Keller, "Molecular mechanism of the pH-dependent calcium affinity in Langerin," bioRxiv:986851 (2020).
[9] A. S. J. S. Mey, B. Allen, H. E. B. Macdonald, J. D. Chodera, M. Kuhn, J. Michel, D. L. Mobley, L. N. Naden, S. Prasad, A. Rizzi, J. Scheen, M. R. Shirts, G. Tresadern, and H. Xu, "Best practices for alchemical free energy calculations," Living J. Comput. Mol. Sci. Living J. Comp. Mol. Sci. ASAP Version, pages **2**, 1, available at https://www.livecomsjournal.org/article/18378-best-practices-for-alchemical-free-energy-calculations-article-v1-0.
[10] M. Tuckerman, *Monte Carlo Statistical Mechanics: Theory and Molecular Simulation* (Oxford University Press, Inc., New York, 2010), pp. 300–304.
[11] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*, 1st ed. (Academic Press, San Diego, San Francisco, NY, Boston, London, Sydney, Tokyo, 2002).
[12] C. A. F. de Oliveira, D. Hamelberg, and J. A. McCammon, "Estimating kinetic rates from accelerated molecular dynamics simulations: Alanine dipeptide in explicit solvent as a case study," J. Chem. Phys. **127**, 175105 (2007).
[13] P. Tiwary and M. Parrinello, "From metadynamics to dynamics," Phys. Rev. Lett. **111**, 230602 (2013).
[14] O. Valsson, P. Tiwary, and M. Parrinello, "Enhancing important fluctuations: Rare events and metadynamics from a conceptual viewpoint," Annu. Rev. Phys. Chem. **67**, 159 (2016).
[15] R. Casasnovas, V. Limongelli, P. Tiwary, P. Carloni, and M. Parrinello, "Unbinding kinetics of a p38 MAP kinase type II inhibitor from metadynamics simulations," J. Am. Chem. Soc. **139**, 4780 (2017).
[16] H. Wu, A. S. J. S. Mey, E. Rosta, and F. Noé, "Statistically optimal analysis of state-discretized trajectory data from multiple thermodynamic states," J. Chem. Phys. **141**, 214106 (2014).
[17] A. S. J. S. Mey, H. Wu, and F. Noé, "xTRAM: Estimating equilibrium expectations from time-correlated simulation data at multiple thermodynamic states," Phys. Rev. X **4**, 041018 (2014).
[18] H. Wu, F. Paul, C. Wehmeyer, and F. Noé, "Multiensemble Markov models of molecular thermodynamics and kinetics," Proc. Natl. Acad. Sci. U. S. A. **113**, E3221 (2016).
[19] L. S. Stelzl, A. Kells, E. Rosta, and G. Hummer, "Dynamic histogram analysis to determine free energies and rates from biased simulations," J. Chem. Theory Comput. **13**, 6328 (2017).
[20] D. J. Bicout and A. Szabo, "Electron transfer reaction dynamics in non-Debye solvents," J. Chem. Phys. **109**, 2325 (1998).
[21] E. Rosta and G. Hummer, "Free energies from dynamic weighted histogram analysis using unbiased markov state model," J. Chem. Theory Comput. **11**, 276 (2014).
[22] L. Donati, M. Heida, B. G. Keller, and M. Weber, "Estimation of the infinitesimal generator by square-root approximation," J. Phys.: Condens. Matter **30**, 425201 (2018).
[23] S. Kieninger, L. Donati, and B. G. Keller, "Dynamical reweighting methods for Markov models," Curr. Opin. Struct. Biol. **61**, 124 (2020).
[24] D. M. Zuckerman and T. B. Woolf, "Dynamic reaction paths and rates through importance-sampled stochastic dynamics," J. Chem. Phys. **111**, 9475 (1999).
[25] T. B. Woolf, "Path corrected functionals of stochastic trajectories: Towards relative free energy and reaction coordinate calculations," Chem. Phys. Lett. **289**, 433 (1998).
[26] D. M. Zuckerman and T. B. Woolf, "Efficient dynamic importance sampling of rare events in one dimension," Phys. Rev. E **63**, 016702 (2000).
[27] C. Xing and I. Andricioaei, "On the calculation of time correlation functions by potential scaling," J. Chem. Phys. **124**, 034110 (2006).
[28] A. B. Adib, "Stochastic actions for diffusive dynamics: Reweighting, sampling, and minimization," J. Phys. Chem. B **112**, 5910 (2008).
[29] I. V. Girsanov, "On transforming a certain class of stochastic processes by absolutely continuous substitution of measures," Theory Probab. Appl. **5**, 285 (1960).
[30] B. Øksendal, *Stochastic Differential Equations: An Introduction with Applications*, 6th ed. (Springer Verlag, Berlin, 2003).
[31] L. Onsager and S. Machlup, "Fluctuations and irreversible processes," Phys. Rev. **91**, 1505 (1953).
[32] J.-H. Prinz, J. D. Chodera, V. S. Pande, W. C. Swope, J. C. Smith, and F. Noé, "Optimal use of data in parallel tempering simulations for the construction of discrete-state Markov models of biomolecular dynamics," J. Chem. Phys. **134**, 244108 (2011).
[33] C. Schütte, A. Nielsen, and M. Weber, "Markov state models and molecular alchemy," Mol. Phys. **113**, 69 (2015).
[34] L. Donati, C. Hartmann, and B. G. Keller, "Girsanov reweighting for path ensembles and Markov state models," J. Chem. Phys. **146**, 244112 (2017).
[35] L. Donati and B. G. Keller, "Girsanov reweighting for metadynamics simulations," J. Chem. Phys. **149**, 072335 (2018).
[36] W. Huisinga, C. Schütte, and A. M. Stuart, "Extracting macroscopic stochastic dynamics: Model problems," Commun. Pure Appl. Math. **56**, 234 (2003).
[37] W. C. Swope, J. W. Pitera, and F. Suits, "Describing protein folding kinetics by molecular dynamics simulations. 1. Theory," J. Phys. Chem. B **108**, 6571 (2004).
[38] N.-V. Buchete and G. Hummer, "Coarse master equations for peptide folding dynamics," J. Phys. Chem. B **112**, 6057 (2008).
[39] B. Keller, X. Daura, and W. F. van Gunsteren, "Comparing geometric and kinetic cluster algorithms for molecular simulation data," J. Chem. Phys. **132**, 074110 (2010).

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

[40]J.-H. Prinz, H. Wu, M. Sarich, B. G. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, "Markov models of molecular kinetics: Generation and validation," J. Chem. Phys. **134**, 174105 (2011).
[41]J.-H. Prinz, B. Keller, and F. Noé, "Probing molecular kinetics with Markov models: Metastable states, transition pathways and spectroscopic observables," Phys. Chem. Chem. Phys. **13**, 16912 (2011).
[42]B. E. Husic and V. S. Pande, "Markov state models: From an art to a science," J. Am. Chem. Soc. **140**, 2386 (2018).
[43]J. D. Chodera, W. C. Swope, F. Noé, J.-H. Prinz, M. R. Shirts, and V. S. Pande, "Dynamical reweighting: Improved estimates of dynamical properties from simulations at multiple temperatures," J. Chem. Phys. **134**, 244107 (2011).
[44]W. F. van Gunsteren and H. J. C. Berendsen, "Algorithms for Brownian dynamics," Mol. Phys. **45**, 637 (1981).
[45]A. Brünger, C. L. Brooks, and M. Karplus, "Stochastic boundary conditions for molecular dynamics simulations of ST2 water," Chem. Phys. Lett. **105**, 495 (1984).
[46]G. Stoltz, "Path sampling with stochastic dynamics: Some new algorithms," J. Comput. Phys. **225**, 491 (2007).
[47]G. Bussi and M. Parrinello, "Accurate sampling using Langevin dynamics," Phys. Rev. E **75**, 056707 (2007).
[48]M. Ceriotti, G. Bussi, and M. Parrinello, "Langevin equation with colored noise for constant-temperature molecular dynamics simulations," Phys. Rev. Lett. **102**, 020601 (2009).
[49]J. A. Izaguirre, C. R. Sweet, and V. Pande, "Multiscale dynamics of macromolecules using normal mode Langevin," Pac. Symp. Biocomput. **15**, 240 (2010).
[50]N. Goga, A. J. Rzepiela, A. H. de Vries, S. J. Marrink, and H. J. C. Berendsen, "Efficient algorithms for Langevin and DPD dynamics." J. Chem. Theory Comput. **8**, 3637 (2012).
[51]B. Leimkuhler and C. Matthews, "Robust and efficient configurational molecular sampling via Langevin dynamics," J. Chem. Phys. **138**, 174102 (2013).
[52]D. A. Sivak, J. D. Chodera, and G. E. Crooks, "Time step rescaling recovers continuous-time dynamical properties for discrete-time Langevin integration of nonequilibrium systems," J. Phys. Chem. B **118**, 6466 (2014).
[53]J. Fass, D. Sivak, G. Crooks, K. Beauchamp, B. Leimkuhler, and J. Chodera, "Quantifying configuration-sampling error in Langevin simulations of complex molecular systems," Entropy **20**, 318 (2018).
[54]P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, and V. S. Pande, "OpenMM 7: Rapid development of high performance algorithms for molecular dynamics," PLoS Comput. Biol. **13**, 1 (2017).
[55]N. Bou-Rabee, "Time integrators for molecular dynamics," Entropy **16**, 138 (2014).
[56]C. C. Chow and M. A. Buice, "Path integral methods for stochastic differential equations," J. Math. Neurosci. **5**, 1 (2015).
[57]P. C. Bressloff, *Stochastic Processes in Cell Biology*, 1st ed. (Springer, New York, 2014).
[58]P. H. Hünenberger, "Thermostat algorithms for molecular dynamics simulations," in *Advanced Computer Simulation* (Springer, Berlin, Heidelberg, 2005).
[59]J. E. Basconi and M. R. Shirts, "Effects of temperature control algorithms on transport properties and kinetics in molecular dynamics simulations," J. Chem. Theory Comput. **9**, 2887 (2013).
[60]J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, "Development and testing of a general amber force field," J. Comput. Chem. **25**, 1157 (2004).
[61]A. Onufriev, D. Bashford, and D. A. Case, "Exploring protein native states and large-scale conformational changes with a modified generalized born model," J. Comput. Chem. **55**, 383 (2004).
[62]See http://docs.openmm.org/latest/api-python/generated/simtk.openmm.openmm.CustomIntegrator.html for information about the CustomIntegrator Class of the simulation package OpenMM; accessed 25 January 2021.
[63]M. Bause, T. Wittenstein, K. Kremer, and T. Bereau, "Microscopic reweighting for nonequilibrium steady-state dynamics," Phys. Rev. E **100**, 060103 (2019).
[64]G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé, "Identification of slow molecular order parameters for Markov model construction," J. Chem. Phys. **139**, 015102 (2013).
[65]F. Nüske, B. G. Keller, G. Pérez-Hernández, A. S. J. S. Mey, and F. Noé, "Variational approach to molecular kinetics," J. Chem. Theory Comput. **10**, 1739–1752 (2014).
[66]O. Lemke and B. G. Keller, "Density-based cluster algorithms for the identification of core sets," J. Chem. Phys. **145**, 164104 (2016).
[67]L. T. Chong, A. S. Saglam, and D. M. Zuckerman, "Path-sampling strategies for simulating rare events in biomolecular systems," Curr. Opin. Struct. Biol. **43**, 88–94 (2017).
[68]G. Grazioli and I. Andricioaei, "Advances in milestoning. I. Enhanced sampling via wind-assisted reweighted milestoning (WARM)," J. Chem. Phys. **149**, 084103 (2018).
[69]P. D. Dixit, J. Wagoner, C. Weistuch, S. Pressé, K. Ghosh, and K. A. Dill, "Perspective: Maximum caliber is a general variational principle for dynamical systems," J. Chem. Phys. **148**, 010901 (2018).
[70]E. K. Peter, J.-E. Shea, and A. Schug, "CORE-MD, a path correlated molecular dynamics simulation method," J. Chem. Phys. **153**, 084114 (2020).
[71]See https://github.com/openmm/openmm/blob/master/platforms/cpu/src/CpuLangevinDynamics.cpp for information about the CpuLangevinDynamics Class of the simulation package OpenMM accessed 15 November 2020.
[72]C. W. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*, 2nd ed. (Springer Verlag, Berlin, Heidelberg, 1983).

3.4 SI: Supporting Information for part A

All content of this section has not been published in any form prior to this thesis.

3.4.1 SI: Introduction

In part **A1** of this thesis, we introduced four underdamped Langevin integrators BAOAB, ABOBA,[30,31] BAOA[36] and an integrator developed by Goga *et al.*[35] which we call GROMACS Stochastic Dynamics (GSD). We argued that GSD and BAOAB are equivalent configurational sampling algorithms and showed that GSD and BAOA are equivalent algorithms in both configurational and momentum space. Further, we discussed the numerical accuracy for configurational as well as kinetic properties for all four integrators in the case of a one-dimensional model system and a water box at near-ambient conditions. As explained in sec. 2.2.3, the BAOAB and ABOBA integrators can be derived by Strang splitting[109] the vector field of the underdamped Langevin equation of motion into three parts.[19,30,31] This strategy yields four other symmetric integrators AOBOA, BOAOB, OBABO and OABAO, where OBABO is also called Bussi-Parrinello thermostat.[34] Here, we repeat the numerical accuracy study shown in part **A1** for the AOBOA, BOAOB, OBABO and OABAO integrators and extend the study to dynamical quantities by investigating the performance of all seven integrators in the context of computing escape rates.

Furthermore, we want to follow up on part **A3** of this thesis, where we extended path reweighting to underdamped Langevin dynamics and derived the exact path reweighting factor for a variant of the underdamped Langevin integrator developed by Izaguirre, Sweet and Pande.[39] In the same publication, we suggested a strategy to derive the random number difference $\Delta\eta$ and subsequently the path reweighting factor M for other underdamped Langevin integrators. Here, we use this strategy to formally derive M for the ABOBA integrator.

3.4.2 SI: Numerical accuracy

Numerical accuracy: Statistical properties

Part **A1** of this thesis (sec. 3.1) tests the numerical accuracy of ABOBA, BAOAB and BAOA/GSD for a one-dimensional potential where the test concept is based on Ref. [31]. In this SI, we extend the test to BOAOB, OBABO, AOBOA and OABAO and present the results in fig. 3.1. In part **A1**, we generated extraordinary long trajectories to compute the results for ABOBA, BAOAB and BAOA/GSD. In fig. 3.1, however, we compute the equilibrium distributions and the temperatures from shorter trajectories yielding identical accuracy but much shorter simulation times.

This means that we conducted new simulations for ABOBA, BAOAB and BAOA/GSD in

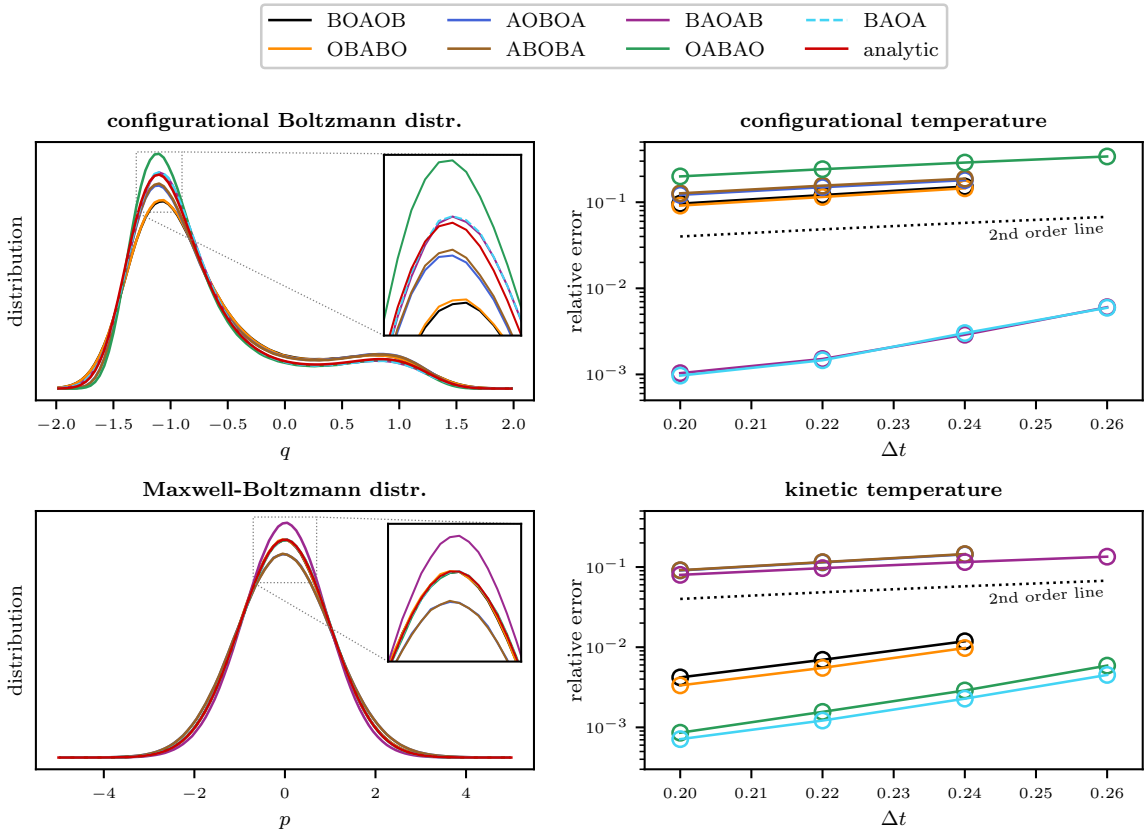


Figure 3.1:
Left column: Configurational Boltzmann distributions with the inset magnifying the region around the deepest well (top) and Maxwell-Boltzmann distributions with the inset magnifying the region around the mean momentum (bottom) for the one-dimensional potential and time step $\Delta t = 0.25$. The analytic distributions are shown in red. **Right column:** Relative error in the average configurational temperature (top) and average kinetic temperature (bottom) for the one-dimensional potential. The equation of the second order line is $\varepsilon^{\text{2nd}} = \exp(\Delta t^2)$.

fig. 3.1 and did not reuse the results published in Ref. [141] (part **A1**). The corresponding computational details are summarized in sec. 3.4.4.

The left column in fig. 3.1 compares the sampled equilibrium distributions in position and momentum space for each integrator and a rather large time step $\Delta t = 0.25$ to the analytic Boltzmann (top) and the analytic Maxwell-Boltzmann distribution (bottom). As expected, the most accurate sampling in position space is achieved by BAOAB and BAOA while all other integrators yield results that deviate to varying degrees from the analytical Boltzmann distribution. AOBOA and ABOBA generate the same configurational distribution which slightly underestimates the distribution in the steepest well. The results for BOAOB and OBABO are equivalent and their distribution significantly underestimates the steepest well. OABAO significantly overestimates the distribution in the steepest well. In momentum space, BOAOB, BAOA, OBABO and OABAO generate indistinguishable equilibrium distributions

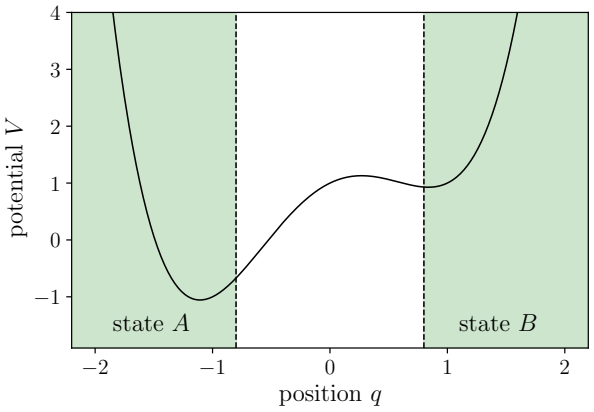


Figure 3.2:
One-dimensional potential with configuration space discretization into two states A and B.

that are a very accurate representation of the corresponding Maxwell-Boltzmann distribution. The results for AOBOA and ABOBA are equivalent. Their velocity distribution underestimates the analytic distribution whereas BAOAB overestimates the analytic distribution.
As a measure for the accuracy with which each integrator reproduces the target temperature we show the relative error (eq. 3.18) as a function of Δt for the same system. The average temperature can either be computed as an average with respect to the Boltzmann distribution, which is called configurational temperature T_{conf}, or as an average with respect to the Maxwell-Boltzmann distribution which is referred to as kinetic temperature T_{kin}. All integrators yield second order accuracy, meaning we expect an relative error of $\varepsilon^{\mathrm{2nd}} = \exp(\Delta t^2)$ (dotted line) for both temperatures. In accordance with the observed accuracy of the sampled Boltzmann distributions, BAOAB and BAOA yield the most accurate configurational temperatures, whereas all other integrators produce significantly less accurate results with more than 10 % discrepancy to the reference temperature for all Δt. Furthermore, groups of integrators that generate equivalent Boltzmann distributions also yield equivalent configurational temperatures. In accordance with the observed accuracy of the sampled Maxwell-Boltzmann distributions, OBABO, BAOAB, OABAO and BAOA yield accurate kinetic temperatures with less than 1 % discrepancy where OABAO and BAOA yield the most accurate results even at large Δt. In comparison, BAOAB, AOBOA and ABOBA compute kinetic temperatures with more than 10 % discrepancy to the reference temperature. Additionally, the results in the right column in Fig. 3.1 show that BAOA, BAOAB and OABAO remain numerically stable even at large time steps $\Delta t > 0.24$.

Numerical accuracy: Dynamical properties

In fig. 3.3, we consider the same one-dimensional potential as in fig. 3.1 and test how accurate the seven integrators compute escape rates at rather large $\Delta t > 0.2$. In this context, we divide the position space into two distinct states A and B, that are spatially separated from each other as depicted in fig. 3.2.

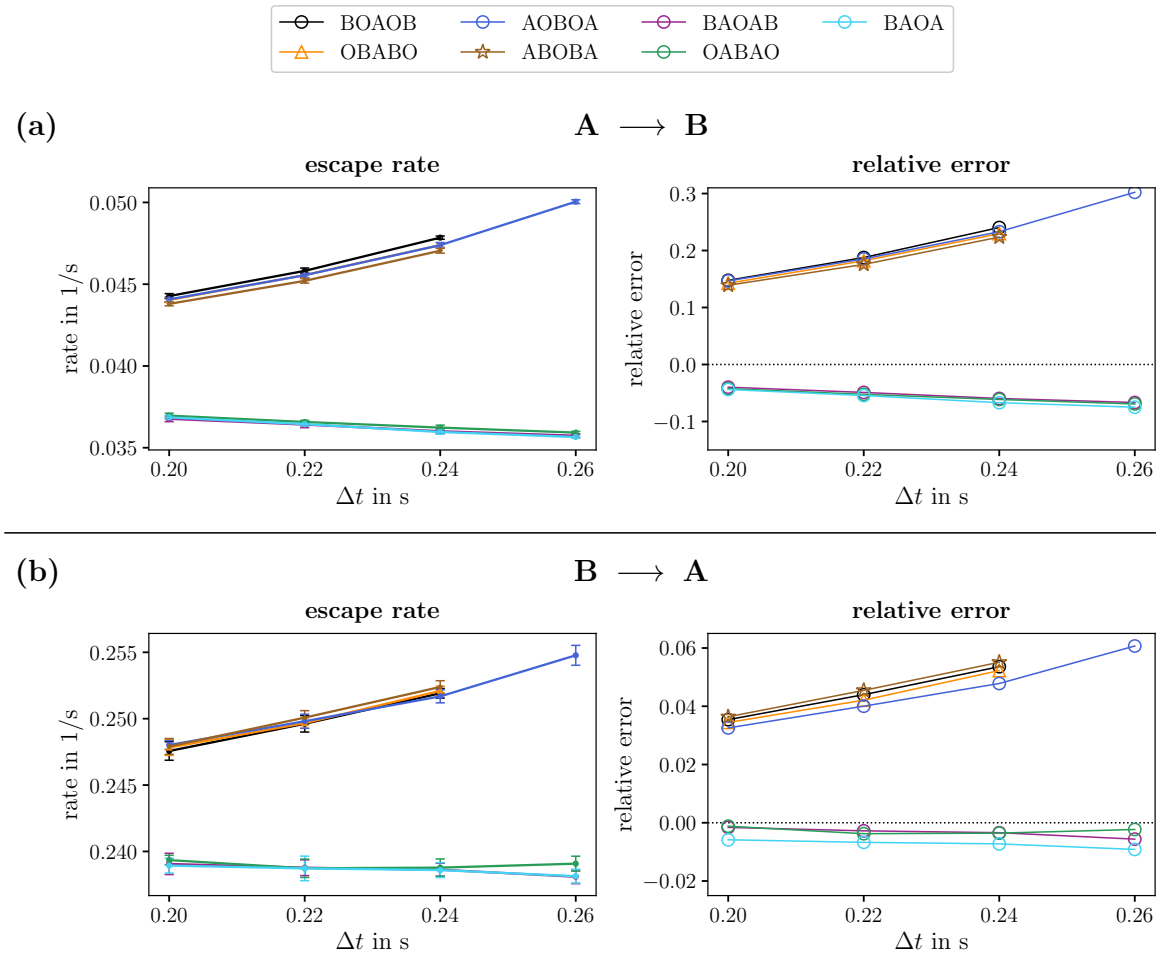


Figure 3.3:
(a) Transition from state A to B: Escape rates (left) where the markers represent the mean and the error bars the standard deviation computed from 10 independent long trajectories per integrator and time step combination. Relative error of the escape rate as a function of Δt with respect to the escape rate computed at the suitable small time step $\Delta t = 0.001$ for each integrator individually.
(b) Transition from state B to A as described above.

The escape rates $r_{A\to B}$ and $r_{B\to A}$ describe the average rates at which the system transitions from state A to B or from state B to A, respectively. Since the particle has to cross a much larger barrier when going from $A \to B$, we expect to observe a significantly smaller rate than for the transition $B \to A$. The left column in fig. 3.3 shows the average escape rates as a function of Δt for each integrator where the markers represent the mean and the error bars the standard deviation computed from 10 independent trajectories per integrator and time step combination. As expected, the escape rates for the transition $A \to B$ are approximately one order of magnitude smaller than the escape rates for the transition $B \to A$. The right column in fig. 3.3 shows the relative error $\varepsilon(\Delta t) = (r(\Delta t) - r_{\text{ref}})/r_{\text{ref}}$ of the escape rates $r(\Delta t)$ for each integrator. In addition to the strength of the deviation, this definition of $\varepsilon(\Delta t)$ also allows a statement about whether an integrator over- or underestimates the escape rate at

the given Δt compared to the reference value. Since all integrators yield a trajectory that converges to the true solution for $\Delta t \to 0$, we computed the escape rate each integrator yields at a suitable small time step $\Delta t = 0.001$ as the respective reference value r_{ref}. In contrast to the statistical quantities examined in fig. 3.1, the results for the escape rates indicate that the accuracy of dynamical quantities is more affected by large time steps. For both escape rates, BOAOB, OBABO, AOBOA, and ABOBA yield very similar results that overestimate the respective rate with increasing discrepancy along the time axis. BAOAB, BAOA and OABAO yield almost identical results that underestimate the respective rate with approximately constant discrepancy along the time axis. The discrepancy between the results of the integrators overestimating and the integrators underestimating the rate is approximately 0.007 1/s for both rates. In total, BAOAB, BAOA and OABAO yield the most accurate rates for both transitions and all time steps.

We suggest that the observed splitting of the results into a set of integrators that overestimates and a set of integrators that underestimates the escape rate could be due to the respective splitting sequences. In the case of BAOAB, BAOA and OABAO, the momentum updates B and O are interspersed by a position update A, whereas all integrators that overestimate the escape rate perform the combined momentum update "BO", "OB", "BOB" or "OBO".

The results shown in fig. 3.3 were computed at the stability limit of the integrators with respect to the time step. For model systems like the one we considered here, we usually run the MD simulations at time steps much smaller than the respective stability limit of the integrator. Consequently, the differences in the results of the seven integrators becomes negligible. In MD simulations for molecular systems however, we typically choose a time step of 2 fs which is indeed close to the stability limit of the underdamped Langevin integrators as shown by Leimkuhler and Matthews.[31] This means that the underdamped Langevin integrators might yield dynamical quantities with for example a difference up to 10 % as indicated by fig. 3.3. However, in complex systems the accuracy with which rates can be computed is limited and literature usually reports rates on a logarithmic scale which ranges over multiple orders of magnitude.[142–144] In this context, the differences between the underdamped Langevin integrators is negligible, especially compared to the uncertainties introduced by the force field or insufficient sampling.

3.4.3 SI: Path reweighting factor of the ABOBA integrator

Part **A3** of this thesis suggests a strategy to derive the random number difference $\Delta\eta_k$ at iteration step k for an underdamped Langevin integrator which can then be used to derive the path reweighting factor M. A part of this strategy is to solve the underdamped Langevin integrator equation for the random number η_k which can get quiet cumbersome and confusing. Here, we derive the random number difference $\Delta\eta_k$ for the ABOBA integrator in

a mathematical more formal way.
As explained in sec. 2.2.3, the vector field of the Langevin equation of motion can be split into three parts A, B and O where every part can be solved separately. The corresponding update operators for a full time step update from state $(q_k, p_k)^\top \in \Omega$ to $(q_{k+1}, p_{k+1})^\top \in \Omega$ with state space $\Omega \in \mathbb{R}^2$ are given as

$$\mathcal{A}\begin{pmatrix} q_k \\ p_k \end{pmatrix} = \begin{pmatrix} q_k + a p_k \\ p_k \end{pmatrix} \tag{3.1a}$$

$$\mathcal{B}\begin{pmatrix} q_k \\ p_k \end{pmatrix} = \begin{pmatrix} q_k \\ p_k + b(q_k) \end{pmatrix} \tag{3.1b}$$

$$\mathcal{O}\begin{pmatrix} q_k \\ p_k \end{pmatrix} = \begin{pmatrix} q_k \\ d\, p_k + f\, \eta_k \end{pmatrix} \tag{3.1c}$$

where we used the abbreviations

$$a = \frac{\Delta t}{m} \tag{3.2a}$$

$$b(q_k) = -\Delta t \nabla_q V(q_k) \tag{3.2b}$$

$$d = e^{-\xi \Delta t} \tag{3.2c}$$

$$f = \sqrt{k_B T m (1 - e^{-2\xi \Delta t})}\,. \tag{3.2d}$$

The corresponding update operators for a half time step update are

$$\mathcal{A}'\begin{pmatrix} q_k \\ p_k \end{pmatrix} = \begin{pmatrix} q_k + a' p_k \\ p_k \end{pmatrix} \tag{3.3a}$$

$$\mathcal{B}'\begin{pmatrix} q_k \\ p_k \end{pmatrix} = \begin{pmatrix} q_k \\ p_k + b'(q_k) \end{pmatrix} \tag{3.3b}$$

$$\mathcal{O}'\begin{pmatrix} q_k \\ p_k \end{pmatrix} = \begin{pmatrix} q_k \\ d'\, p_k + f'\, \eta_k \end{pmatrix} \tag{3.3c}$$

where the abbreviations are denoted with a prime

$$a' = \frac{\Delta t}{2m} \tag{3.4a}$$

$$b'(q_k) = -\frac{\Delta t}{2} \nabla_q V(q_k) \tag{3.4b}$$

$$d' = e^{-\xi \frac{\Delta t}{2}} \tag{3.4c}$$

$$f' = \sqrt{k_B T m (1 - e^{-\xi \Delta t})}\,. \tag{3.4d}$$

For a detailed explanation of this splitting method, we refer to sec. 2.2.3. In the case of ABOBA, the update operator $\mathcal{U}_{ABOBA}$ which defines the update $(q_k, p_k)^\top \to (q_{k+1}, p_{k+1})^\top$

can be calculated as

$$\begin{aligned}
\mathcal{U}_{ABOBA}\begin{pmatrix} q_k \\ p_k \end{pmatrix} &= \mathcal{A}'\mathcal{B}'\mathcal{O}\mathcal{B}'\mathcal{A}'\begin{pmatrix} q_k \\ p_k \end{pmatrix} \\
&= \mathcal{A}'\mathcal{B}'\mathcal{O}\mathcal{B}'\begin{pmatrix} q_k + a'p_k \\ p_k \end{pmatrix} \\
&= \mathcal{A}'\mathcal{B}'\mathcal{O}\begin{pmatrix} q_k + a'p_k \\ p_k + b'(q_k + a'p_k) \end{pmatrix} \\
&= \mathcal{A}'\mathcal{B}'\begin{pmatrix} q_k + a'p_k \\ dp_k + db'(q_k + a'p_k) + f\eta_k \end{pmatrix} \\
&= \mathcal{A}'\begin{pmatrix} q_k + a'p_k \\ dp_k + db'(q_k + a'p_k) + f\eta_k + b'(q_k + a'p_k) \end{pmatrix} \\
&= \begin{pmatrix} q_k + a'p_k + a'[dp_k + db'(q_k + a'p_k) + f\eta_k + b'(q_k + a'p_k)] \\ dp_k + db'(q_k + a'p_k) + f\eta_k + b'(q_k + a'p_k) \end{pmatrix} \quad (3.5)
\end{aligned}$$

Next, we consider a target system governed by the target potential $\widetilde{V}(q) = V(q)+U(q)$, where we call $V(q)$ simulation potential and $U(q)$ bias. To generate the same update $(q_k, p_k)^\top \to (q_{k+1}, p_{k+1})^\top$ as in eq. 3.5 at the target potential, we use a different random number $\widetilde{\eta}_k$ per iteration step in order to account for the change in the potential. The corresponding update operator $\widetilde{\mathcal{U}}_{ABOBA}$ is given as

$$\widetilde{\mathcal{U}}_{ABOBA}\begin{pmatrix} q_k \\ p_k \end{pmatrix} = \begin{pmatrix} q_k + a'p_k + a'[dp_k + d\widetilde{b'}(q_k + a'p_k) + f\widetilde{\eta}_k + \widetilde{b'}(q_k + a'p_k)] \\ dp_k + d\widetilde{b'}(q_k + a'p_k) + f\widetilde{\eta}_k + \widetilde{b'}(q_k + a'p_k) \end{pmatrix} \quad (3.6)$$

with $\widetilde{b'}(q_k)$ being defined in eq. 3.4b where we inserted the target potential $\widetilde{V}(q)$. We require that both update operator yield the same update, i.e. the path remains unchanged,

$$\begin{pmatrix} q_{k+1} \\ p_{k+1} \end{pmatrix} = \mathcal{U}_{ABOBA}\begin{pmatrix} q_k \\ p_k \end{pmatrix} = \widetilde{\mathcal{U}}_{ABOBA}\begin{pmatrix} q_k \\ p_k \end{pmatrix}. \quad (3.7)$$

Thus, we need to solve

$$\mathcal{U}_{ABOBA}\begin{pmatrix} q_k \\ p_k \end{pmatrix} - \widetilde{\mathcal{U}}_{ABOBA}\begin{pmatrix} q_k \\ p_k \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad (3.8)$$

for $\Delta\eta_k = \widetilde{\eta}_k - \eta_k$, i.e. we determine the change in the random number that yields the same path even though the potential has changed. Inserting eqs. 3.5 and 3.7 into eq. 3.8 yields

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} q_k + a'p_k + a'(dp_k + db'(q_k + a'p_k) + f\eta_k + b'(q_k + a'p_k)) \\ dp_k + db'(q_k + a'p_k) + f\eta_k + b'(q_k + a'p_k) \end{pmatrix} -$$

$$
\begin{aligned}
& \begin{pmatrix} q_k + a'p_k + a'(dp_k + d\widetilde{b}'(q_k + a'p_k) + f\widetilde{\eta}_k + \widetilde{b}'(q_k + a'p_k)) \\ dp_k + d\widetilde{b}'(q_k + a'p_k) + f\widetilde{\eta}_k + \widetilde{b}'(q_k + a'p_k) \end{pmatrix} \\
= & \begin{pmatrix} a'[d \cdot b'(q_k + a'p_k) + f\eta_k + b'(q_k + a'p_k)] \\ db'(q_k + a'p_k) + f\eta_k + b'(q_k + a'p_k) \end{pmatrix} - \\
& \begin{pmatrix} a'[d \cdot \widetilde{b}'(q_k + a'p_k) + f\widetilde{\eta}_k + \widetilde{b}'(q_k + a'p_k)] \\ d\widetilde{b}'(q_k + a'p_k) + f\widetilde{\eta}_k + \widetilde{b}'(q_k + a'p_k) \end{pmatrix} \\
= & \begin{pmatrix} a' \\ 1 \end{pmatrix} [d \cdot b'(q_k + a'p_k) + f\eta_k + b'(q_k + a'p_k)] - \\
& \begin{pmatrix} a' \\ 1 \end{pmatrix} [d \cdot \widetilde{b}'(q_k + a'p_k) + f\widetilde{\eta}_k + \widetilde{b}'(q_k + a'p_k)] \\
= & \begin{pmatrix} a' \\ 1 \end{pmatrix} [d \cdot b'(q_k + a'p_k) + f\eta_k + b'(q_k + a'p_k) - d \cdot \widetilde{b}'(q_k + a'p_k) - f\widetilde{\eta}_k - \widetilde{b}'(q_k + a'p_k)] .
\end{aligned}
$$

We use $\widetilde{\eta}_k = \eta_k + \Delta\eta_k$ and $U(q) = \widetilde{V}(q) - V(q)$ to get

$$
\begin{aligned}
\begin{pmatrix} 0 \\ 0 \end{pmatrix} &= \begin{pmatrix} a' \\ 1 \end{pmatrix} [(d+1) \cdot b'(q_k + a'p_k) - (d+1) \cdot \widetilde{b}'(q_k + a'p_k) + f(\eta_k - \widetilde{\eta}_k)] \\
&= \begin{pmatrix} a' \\ 1 \end{pmatrix} [(d+1) \cdot \frac{\Delta t}{2} \nabla_q U(q_{k+1/2}) - f\Delta\eta_k] . \qquad (3.9)
\end{aligned}
$$

where we substituted $q_{k+1/2} = q_k + a'p_k$ (eq. 2.67a). Solving eq. 3.9 for $\Delta\eta_k$ yields the random number difference

$$
\begin{aligned}
\Delta\eta_k &= \frac{(d+1)}{f} \frac{\Delta t}{2} \nabla_q U(q_{k+1/2}) \\
&= \frac{1 + e^{-\xi\Delta t}}{\sqrt{k_B T m (1 - e^{-2\xi\Delta t})}} \frac{\Delta t}{2} \nabla_q U(q_{k+1/2}) \qquad (3.10)
\end{aligned}
$$

According to eqs. 7 and 9 in Ref. [145] (part **A3**) the exact path reweighting factor for the ABOBA integrator is then given as

$$
M(\boldsymbol{\omega}|\omega_0) = \exp\left(- \sum_{k=0}^{n-1} \left[\eta_k \cdot \Delta\eta_k + \frac{1}{2} (\Delta\eta_k)^2 \right] \right) \qquad (3.11)
$$

with $\Delta\eta_k$ as defined in eq. 3.10.

In principal, the same strategy can be used to derive the path reweighting factor for other Langevin integrators. However, we strongly emphasize that for some integrators, the random number difference might not simultaneously be able to account for both, the changes in configuration as well as momentum space. This means, that some integrators can only generate an identical configuration sequence $\mathbf{q} = (q_0, q_1, \ldots, q_n)$ or an identical momentum sequence

$\mathbf{p} = (p_0, p_1, \ldots, p_n)$ at both potentials but not an identical path $\boldsymbol{\omega}$ in state space. To adapt the presented strategy to such cases is the task of a future work.

3.4.4 SI: Methods

For all numerical studies, we implemented the ABOBA, BAOAB, BAOA, AOBOA, BOAOB, OBABO, OABAO integrator equations as defined in appendix A.2 in Python 3. Following Ref. [31], we chose the tilted double well potential

$$V(q) = (q^2 - 1)^2 + q, \qquad q \in \mathbb{R} \tag{3.12}$$

(fig. 3.2) to test how accurate the Langevin integrators generate the equilibrium distributions, reproduce the configurational and kinetic temperature (fig. 3.1) and to compute the escape rates (fig. 3.3) at large time steps. The potential defined in eq. 3.12 has two minima at $q \approx -1$ and $q \approx 1$ and a maximum at $q \approx 0$. We set the following simulation parameters for all simulations: Boltzmann constant $k_B = 1$, collision rate $\xi = 1$, mass $m = 1$, temperature $T = 1$, initial position $q_0 = 0$ and initial velocity $v_0 = 0$.

Fig. 3.1, left column: We generated a trajectory with 10^7 iterations at time step $\Delta t = 0.25$ for all seven integrators ABOBA, BAOAB, BAOA, AOBOA, BOAOB, OBABO, OABAO. We computed the distributions as normed histograms where we divided the interval $[-2, 2]$ for the Boltzmann distribution and the interval $[-5, 5]$ for the Maxwell-Boltzmann distributions into 100 equidistant bins. As a reference, we compute the configurational Boltzmann distribution

$$\phi(q) = \frac{\exp\left(-\frac{1}{k_B T} V(q)\right)}{\int\limits_{-\infty}^{\infty} \exp(-\frac{1}{k_B T} V(q)) \mathrm{d}q}, \tag{3.13}$$

and the analytical Maxwell-Boltzmann distribution

$$\rho(p) = \sqrt{\frac{1}{2 k_B T m \pi}} \exp\left(-\frac{1}{2 k_B T m} p^2\right). \tag{3.14}$$

Fig. 3.1, right column: We generated 500 independent trajectories of length $n = 10^7$ iterations for each of the seven integrators at different time steps $\Delta t = 0.20, 0.22, 0.24, 0.26$. We evaluated the average configurational temperature

$$T_{\mathrm{conf}} = \frac{\langle q \cdot \nabla_q V(q) \rangle}{k_B} = \frac{\lim\limits_{n \to \infty} \frac{1}{n} \sum\limits_{i=0}^{n} q_i \cdot \nabla_q V(q_i)}{k_B} \tag{3.15}$$

and the average kinetic temperature

$$T_{\text{kin}} = \frac{\left\langle \frac{p^2}{m} \right\rangle}{k_B} = \frac{\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n} \frac{p_i^2}{m}}{k_B} \tag{3.16}$$

from each trajectory. This resulted in 500 independent values for T_{conf} and 500 independent values for T_{kin} from which we computed the average

$$T_{\text{av}}(\Delta t) = \frac{\sum_{i=1}^{500} T_{\text{conf/kin}}^{(i)}}{500}\,, \tag{3.17}$$

for each integrator and time step combination. The relative errors were computed as

$$\varepsilon(\Delta t) = \frac{|T_{\text{ref}} - T_{\text{av}}(\Delta t)|}{T_{\text{ref}}} \tag{3.18}$$

with reference temperature $T_{\text{ref}} = 1$.

Fig. 3.3: To compute escape rates, we discretized the position space into two distinct states A and B, where the particle is in state A if $q < -0.8$ and in state B if $q > 0.8$ (see Fig. 3.2). Accordingly, the escape time $t^i_{A\to B}$ represents the time that evolves between the first time the particle enters state A and the first time it subsequently enters state B. We defined the escape time $t^i_{B\to A}$ as the time that evolves between the first time the particle enters state B and the first time it subsequently enters state A. Please note that these definitions include recrossing events, e.g. when the particle leaves state A and returns to state A without reaching state B.
We used different values for the time step $\Delta t = 0.20, 0.22, 0.24, 0.26$ and all seven underdamped Langevin integrators mentioned above. For each time step and integrator combination, we conducted 10 independent MD simulations of length 10^7 iterations and extracted the escape time $t^i_{A\to B}$ of all n escape events $A \to B$ and the escape time $t^i_{B\to A}$ of all m escape events $B \to A$ that occurred along a trajectory. We computed the average escape rates

$$r_{A\to B} = \frac{1}{\text{mean}(t^i_{A\to B})} = \frac{n}{\sum_{i=1}^{n} t^i_{A\to B}} \tag{3.19}$$

$$r_{B\to A} = \frac{1}{\text{mean}(t^i_{B\to A})} = \frac{m}{\sum_{i=1}^{m} t^i_{B\to A}}\,. \tag{3.20}$$

per trajectory, yielding 10 values for $r_{A\to B}$ and 10 values for $r_{B\to A}$ per integrator and time step combination. In the left column of fig. 3.3, we show the mean $r(\Delta t)$ of the respective 10 rates as markers and the corresponding standard deviation as error bars for each integrator

as a function of Δt.

To get a reference escape rate, we generated 10 independent trajectories of length 10^8 iterations at time step $\Delta t = 0.001$ for each integrator. Following the protocol described above, we computed the reference escape rate r_{ref} for each transition as the mean of the 10 escape rates $r_{A\to B}$ and the mean of the 10 escape rates $r_{B\to A}$. The relative error $\varepsilon(\Delta t)$ is shown in the right column of fig. 3.3 and was computed for each integrator individually as

$$\varepsilon(\Delta t) = \frac{r(\Delta t) - r_{\text{ref}}}{r_{\text{ref}}} \,. \tag{3.21}$$

3.5 Paper B1

"Stable Isotopomers of *myo*-Inositol Uncover a Complex MINPP1-Dependent Inositol Phosphate Network"

M. Nguyen Trung, S. Kieninger, Z. Fandi, D. Qiu, G. Liu, N. K. Mehendale, A. Saiardi, H. Jessen, B. G. Keller, D. Fiedler

ACS Cent. Sci., **2022**, *8*, 1683-1694.
DOI: 10.1021/acscentsci.2c01032
URL: `https://doi.org/10.1021/acscentsci.2c01032`
Computational scripts available on Github:
`https://github.com/bkellerlab/MINPP1_reactionNetwork`
xlsx-file of quantification tables for cell extracts available at:
`https://doi.org/10.1021/acscentsci.2c01032`

Contributions
Minh Nguyen Trung and Dorothea Fiedler conceived the project with input from Adolfo Saiardi. M.N.T. and D.F. wrote the "Supporting Information" and the major part of the manuscript. All computational analysis was conducted by Stefanie Kieninger. S.K. also wrote the "Supporting information: Numerical Analysis" and contributed the section "MINPP1 exhibits different kinetic properties towards $InsP_5$[2OH] and $InsP_6$" and Figure 6.a+b to the manuscript. M.N.T. performed most biochemical and *in vitro* experiments, NMR measurements, organic syntheses, non-computational interpretation of the experimental data and prepared all CE-MS samples. Zeinab Fandi contributed to the initial testing of the MINPP1 expression and reactivity optimizations under the supervision of M.N.T and D.F.. Danye Qiu and Guizhen Liu measured and analyzed all CE-MS samples under supervision of Henning Jessen. Neelay K. Mehendale contributed to the biochemical experiments that were required during the revision process. All authors contributed to the final version of the manuscript.

Summary

Part **B** of this thesis nicely illustrates that the combination of experimental and theoretical approaches is a powerful strategy to study complex processes in nature. One such a complex process is the inositol polyphosphate metabolism in eukaryotes.
Inositol polyphosphates (InsPs) are small, water-soluble molecules that vary in the phosphorylation pattern of their *myo*-inositol scaffold. They act as secondary messengers in signal transduction pathways and are key to fundamental physiological processes.[85–87] Inositol-1,3,4,5,6-pentakisphosphate ($InsP_5$[2OH]) and inositol hexakisphosphate (InP_6) are reported to be the most abundant InsPs in mammalian cells.[93–96,146,147] Both are highly phosphorylated InsPs and can be converted into lower phosphorylated InsPs via dephosphorylation pathways.[97]
The dephosphorylation is mediated by Multiple Inositol Polyphosphate Phosphatase 1 (MINPP1), which is the only known enzyme in the human genome capable to dephosphorylate $InsP_6$.[148–150] Since MINPP1 was recently linked to a genetic disorder which affects cognitive functions and life expectancy of humans,[98,99] it is of great interest to shed light onto MINPP1 mediated processes. Depending on the phosphorylation level of the InsP, MINPP1 was shown to dephosphorylate different positions with different affinities and kinetics.[151,152] In the case of $InsP_6$, MINPP1 is annotated as a 3-phosphatase, meaning it predominantly removes the phosphoryl group at the 3-position.[153]
To date, literature only reports *in vitro* experiments which showed MINPP1 mediated dephosphorylation of $InsP_6$ to only sparsely annotated InsPs with three or four phosphoryl groups.[90,91,154–156] However, the distinct dephosphorylation pathways with the relevant InsP intermediates and the corresponding kinetics are still unknown and the role of lower phosphorylated InsPs remains unclear. Due to the limitations of current analytical tools, it is particularly challenging to analyze mixtures of InsP metabolites and to provide answers to the previous questions.

This publication presents a strategy to distinguish between different InsPs and simultaneously measure the time evolution of the respective InsP concentrations via BIRD-{^{1}H-^{13}C}HMQC-NMR measurements. In this context, we synthesize[157] fully and asymmetrically ^{13}C-labeled isotopomers of *myo*-inositol and InsPs and utilize them in both, biochemical and cellular metabolic labeling experiments. We demonstrate that the NMR signals of different InsPs cluster in a systematic manner, where the asymmetric labeling experiments help to resolve enantiomers. We use this perception in a methodological way to identify the different InsPs signals and answer the following questions:

1) Do lower phosphorylated InsPs play a role in the InsP metabolism in mammalian cells?

2) What are the distinct MINPP1 mediated dephosphorylation pathways of $InsP_5$[2OH] and $InsP_6$?

In the context of question 1, we report inositol 2-monophosphate (Ins(2)P) and inositol 2,3-bisphosphate (Ins(2,3)P_2) as major mammalian metabolites, as has previously been noted.[91,158] Furthermore, CE-MS experiments with cells lacking MINPP1 could not detect any InsP(2,3)P_2 and InsP(2) metabolites. We conclude that the formation of InsP(2,3)P_2 and InsP(2) depends on MINPP1 and we assume that they are generated directly from $InsP_6$. To validate this assumption, we aim to answer question 2.

In biochemical experiments, we investigated the *in vitro* activity of MINPP1 against fully ^{13}C-labeled $InsP_5$[2OH] and $InsP_6$, respectively. We performed 2D NMR measurements to extract the progress curves of the main intermediates over a period of 69 hours and built a kinetic network assumption for each dephosphorylation pathway. The progress curves describe the time evolution of the concentration of each InsP at a suitable time resolution and thus constitute a data set which is very well suited for computing dynamical quantities, such as reaction rates. Since MINPP1 is a phosphatase, we can describe the dephosphorylation network as a set of consecutive, irreversible, one-step reactions between the individual InsPs and use a Markovian kinetic scheme to extract the reaction rate of each dephosphorylation step. In this context, Markovian means that the reaction rates are time constant. We formulated the corresponding Master equation and determined the reaction rate matrix with a numerical optimization routine that minimized the error between the theoretically predicted and the experimentally measured progress curves of each InsP. Our experimental as well as numerical results confirm MINPP1's annotation as a 3-phosphatase and the two highest reaction rates correspond to the reported canonical MINPP1 activity towards $InsP_5$[2OH].[151,152] In the case of $InsP_6$, the experiments confirmed the formation of Ins(2,3)P_2 and $InsP_2$ as products of the MINPP1 mediated dephosphorylation pathway. Experiments with asymmetrically ^{13}C-labeled $InsP_6$ revealed that both Ins(2,3)P_2 and Ins(1,2)P_2 are formed with an excess of Ins(2,3)P_2.

Finally, we showed numerically as well as experimentally that $InsP_6$ could act as an inhibitor for the dephosphorylation of the MINPP1-generated intermediates.

ACS central science

http://pubs.acs.org/journal/acscii Research Article

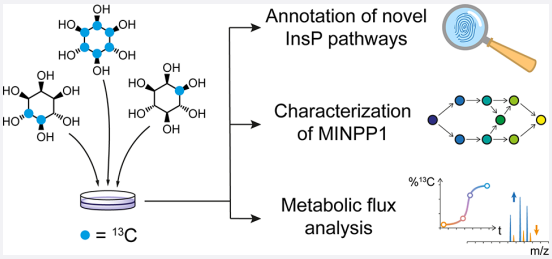
Stable Isotopomers of *myo*-Inositol Uncover a Complex MINPP1-Dependent Inositol Phosphate Network

Minh Nguyen Trung, Stefanie Kieninger, Zeinab Fandi, Danye Qiu, Guizhen Liu, Neelay K. Mehendale, Adolfo Saiardi, Henning Jessen, Bettina Keller, and Dorothea Fiedler*

Cite This: *ACS Cent. Sci.* 2022, 8, 1683−1694 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The water-soluble inositol phosphates (InsPs) represent a functionally diverse group of small-molecule messengers involved in a myriad of cellular processes. Despite their centrality, our understanding of human InsP metabolism is incomplete because the available analytical toolset to characterize and quantify InsPs in complex samples is limited. Here, we have synthesized and applied symmetrically and unsymmetrically ^{13}C-labeled *myo*-inositol and inositol phosphates. These probes were utilized in combination with nuclear magnetic resonance spectroscopy (NMR) and capillary electrophoresis mass spectrometry (CE-MS) to investigate InsP metabolism in human cells. The labeling strategy provided detailed structural information via NMR—down to individual enantiomers—which overcomes a crucial blind spot in the analysis of InsPs. We uncovered a novel branch of InsP dephosphorylation in human cells which is dependent on MINPP1, a phytase-like enzyme contributing to cellular homeostasis. Detailed characterization of MINPP1 activity in vitro and in cells showcased the unique reactivity of this phosphatase. Our results demonstrate that metabolic labeling with stable isotopomers in conjunction with NMR spectroscopy and CE-MS constitutes a powerful tool to annotate InsP networks in a variety of biological contexts.



INTRODUCTION

Myo-inositol polyphosphates (InsPs) are ubiquitous, water-soluble small molecules found in all eukaryotes. InsPs are involved in a wide spectrum of biological functions as they are key to fundamental physiological processes. A well-characterized example is inositol-1,4,5-trisphosphate ($Ins(1,4,5)P_3$) as a Ca^{2+} release factor. More recently, InsPs were shown to regulate the activity of class I histone deacetylases as well as Bruton's tyrosine kinase (Btk), which implies a wider role for InsPs in transcriptional regulation and in governing intracellular signal transduction.[1−3]

The InsPs vary greatly with respect to their phosphorylation patterns, and over 20 different InsPs are currently thought to be part of mammalian InsP metabolism.[4−7] The most abundant InsPs in mammalian cells are inositol-1,3,4,5,6-pentakisphosphate ($InsP_5[2OH]$) and inositol hexakisphosphate (also called phytic acid, $InsP_6$), with cellular concentrations ranging from the lower micromolar range to >100 μM in human cells and even in the sub-millimolar range in slime molds.[8,9] $InsP_5[2OH]$ and $InsP_6$ are precursors for the biosynthesis of inositol pyrophosphates (PP-InsPs), which have recently drawn increasing attention due to their dense phosphorylation patterns and their involvement in central signaling processes.[10] $InsP_6$ is also found in a growing number of proteins and protein complexes as a structural cofactor or as a "molecular glue" for protein−protein interactions.[11−14]

While the kinase-mediated pathways of InsP biosynthesis are fairly well studied, there is limited information on dephosphorylation of InsPs in mammalian cells,[15] especially with respect to the higher phosphorylated members. To date, MINPP1 (Multiple Inositol Polyphosphate Phosphatase 1) is the only recognized enzyme in the human genome capable of dephosphorylating $InsP_6$.[16,17] MINPP1 is related to phytases, a highly conserved group of enzymes in many other organisms that can dephosphorylate various InsPs.[18] MINPP1 has been shown to play a role in apoptosis, ER-related stress, and bone and cartilage tissue formation.[17,19] Recently, MINPP1 was connected to a genetic disorder: patients with loss-of-function mutations in MINPP1 exhibit pontocerebellar hypoplasia (PCH), a neurodegenerative disease severely impacting cognitive functions and life expectancy.[20,21] Therefore, it is important to understand the molecular mechanisms of MINPP1-governed functions in healthy and diseased states.

Although MINPP1 is annotated as a 3-phosphatase, i.e., it predominantly removes the phosphoryl group at the 3-position

Received: September 2, 2022
Published: December 5, 2022

https://doi.org/10.1021/acscentsci.2c01032
ACS Cent. Sci. 2022, 8, 1683−1694

ACS Central Science http://pubs.acs.org/journal/acscii Research Article

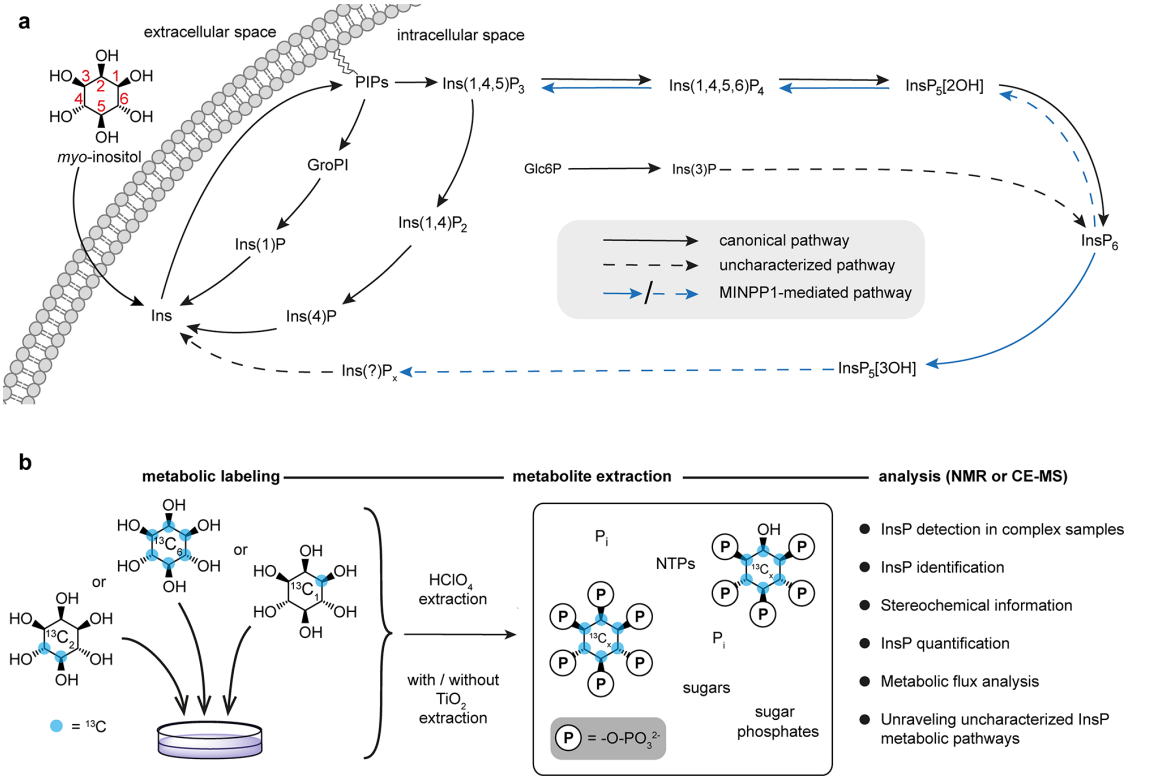


Figure 1. Probing InsP metabolism with *myo*-inositol isotopomers. (a) Simplified overview of InsP metabolism with MINPP1-mediated processes highlighted. It is assumed that MINPP1 dephosphorylates $InsP_6$ and various other InsPs down to sparsely annotated $InsP_3$ isomers. PIPs, phosphatidylinositol phosphates; GroPI, glycerophosphoinositol; Glc6P, glucose-6-phosphate; MINPP1, multiple inositol polyphosphate phosphatase1; Ins(*X*,*Y*)P_z, *myo*-inositol with *z* phosphoryl groups at positions *X*,*Y*; $InsP_5$[*X*OH], inositol pentakisphosphate with a hydroxyl group at position *X*. IUPAC numbering convention of the positions on the inositol scaffold is shown in red. (b) Workflow for the analysis of cellular InsP pools through metabolic labeling: human cells are grown in medium devoid of nonlabeled *myo*-inositol but supplemented with an isotopomer of *myo*-inositol ([$^{13}C_6$]Ins, 4,5[$^{13}C_2$]*myo*-inositol, 1[$^{13}C_1$]*myo*-inositol, or 3[$^{13}C_1$]*myo*-inositol) which are incorporated into the cellular InsP pool. Metabolites are then extracted, resulting in a complex sample containing all water-soluble biomolecules, such as nucleotide triphosphates (NTPs), inorganic phosphate (P_i), and the labeled InsPs. This mixture can be analyzed via NMR or CE-MS exploiting NMR activity and mass difference of the ^{13}C label.

of $InsP_6$,[22] MINPP1 is also able to dephosphorylate several InsPs at different positions with varying affinities and kinetics.[23,24] The current assumption is that MINPP1 dephosphorylates $InsP_6$ to hitherto only sparsely annotated $InsP_{4/3}$ species (Figure 1a).[6,7,25−27] However, this activity has only been demonstrated in vitro and in intact cells overexpressing a cytosolic variant of MINPP1. Whether this activity is relevant in vivo and which InsP intermediates are exactly involved is still not clear.[25] Furthermore, there is no consensus how MINPP1 accesses its InsP substrates. While early studies suggest MINPP1 to be localized to the ER,[28,29] others have also shown alternative localizations into the Golgi, in lysosomes, or even secreted in exosomes.[30,31]

Probing and quantifying InsP metabolites and their interconversion is still a challenging task due to the limitations of current analytical tools. Many established methods for the detection and analysis of InsPs rely on some form of physicochemical separation of different InsPs from a complex mixture. The most common methods are strong-anion exchange chromatography (SAX-HPLC)-based fractionation in combination with radiolabeling and scintillation counting, high-density polyacrylamide electrophoresis with cationic staining, or, more recently, capillary electrophoresis coupled to mass spectrometry (CE-MS).[26,32−37] Most of these methods are sensitive and powerful for the analysis of highly phosphorylated InsPs, but the separation and detection of lower InsPs (i.e., $InsP_1$, $InsP_2$, and $InsP_3$ species) in a mixture with isobaric sugar phosphates remain difficult. Our group recently established a metabolic labeling strategy using isotopically labeled [$^{13}C_6$]*myo*-inositol. Analysis of the extracted metabolites by 2D nuclear magnetic resonance (NMR) spectroscopy enabled the quantification of higher phosphorylated InsPs ($InsP_6$, $InsP_5$[2OH], and PP-InsPs; Figure 1b) without the need for analytical separation.[38] In addition, 2D-NMR measurements provide important information on the InsP phosphorylation patterns and should be able to detect the whole range of InsP metabolites, including the lower phosphorylated species.

Here, we combined fully ^{13}C-labeled and asymmetrically ^{13}C-labeled isotopomers of *myo*-inositol and InsPs in both biochemical and cellular metabolic labeling experiments. Making use of their inherent properties (position-specific NMR activity and different molecular masses), we uncovered an uncharacterized branch of human InsP metabolism.

https://doi.org/10.1021/acscentsci.2c01032
ACS Cent. Sci. 2022, 8, 1683−1694

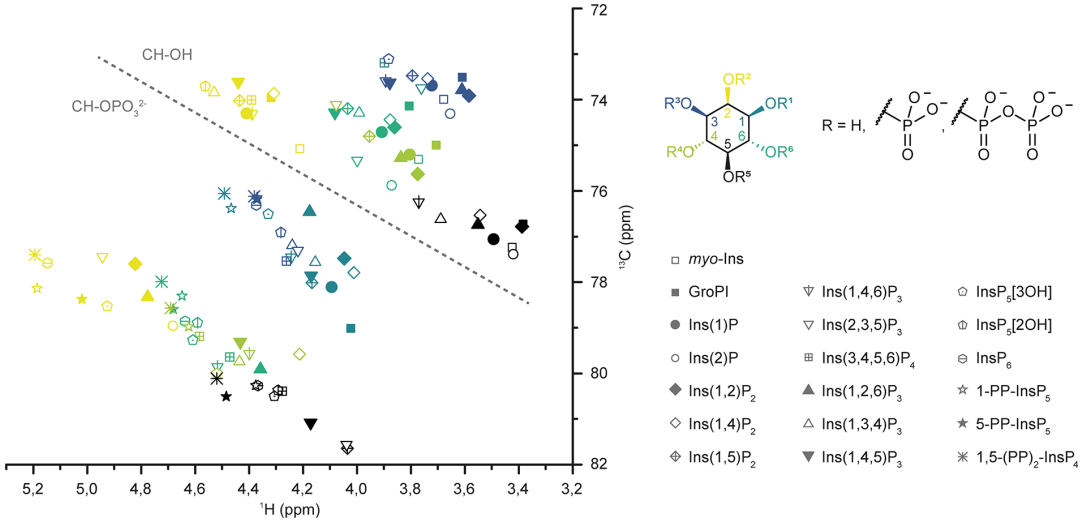


Figure 2. HMQC signals of InsPs with different phosphorylation patterns cluster systematically. Collection of BIRD-{^{1}H,^{13}C}HMQC NMR data of various InsP standards in metabolic extract buffer conditions (saturated $KClO_4$ in D_2O, pH* 6.0). HMQC signals of different InsPs are represented with symbols, while the position on the inositol ring is color-coded. HMQC signals cluster together depending on phosphorylation status (dotted line) and position on the inositol ring. CH groups bearing the pyrophosphate moiety of PP-InsPs or the 1-glyceryl phosphate group of GroPI cluster with the phosphorylated CH groups and were treated accordingly for creating bagplots (Figure S1).

$Ins(2,3)P_2$ and Ins(2)P were identified as major InsPs species in human cells, and their levels are dependent on MINPP1 activity toward $InsP_6$ in vitro and in cellula. Through in vitro characterization, computational kinetic modeling, and metabolic flux via CE-MS analysis, we dissect the complex reactivity of MINPP1. We envision that this combined application of *myo*-inositol isotopomers in NMR and CE-MS experiments will help unravel complex InsP networks in different biological contexts in the future.

RESULTS

InsP Phosphorylation Patterns Are Well Resolved by BIRD-{^{1}H,^{13}C}HMQC NMR Spectra. The analysis of complex mixtures of InsP metabolites still constitutes a significant analytical challenge. To identify inositol-derived signals in biological samples via NMR in a methodical way, BIRD-{^{1}H,^{13}C}HMQC-NMR spectra of 19 different InsPs and PP-InsPs (commercially available or synthesized) were recorded and assigned. The collective data of these spectra illustrate that the NMR signals of InsPs cluster in a systematic manner (Figures 2 and S1). NMR signals corresponding to methine groups adjacent to a nonphosphorylated hydroxyl substituent (**CH**−OH) are separated from methine group signals with a phosphate substituent (**CH**−O−PO_3^{2-}), which are collectively shifted downfield in both ^{1}H and ^{13}C dimensions. Within these two groups, clusters for the different positions on the *myo*-inositol ring are apparent. The 2- and 5-positions form clusters of their own, while positions 1 and 3 as well as positions 4 and 6 are intertwined due to the symmetry plane of the *myo*-inositol ring. These combined spectra illustrate that a complete set of NMR signals of an InsP can be used to determine the phosphorylation pattern, and thus the identity, of a given InsP. In the case of chiral InsPs, their NMR spectra cannot be used for a definitive assignment but can narrow the identity down to a pair of enantiomers. For distinguishing two InsP enantiomers, a desymmetrization strategy has to be employed, such as unsymmetrical isotopic labeling of the *myo*-inositol ring with ^{13}C, as will be discussed below.

Ins(2,3)P_2 and Ins(2)P Are Major Mammalian Metabolites. We next performed metabolic labeling of human cell lines (HEK293, HCT116, HT29, H1Hela, H1975) with [$^{13}C_6$]*myo*-inositol (Figure 1b).[38] In brief, cells were grown in a custom medium based on DMEM which contains no natural [^{12}C]*myo*-inositol but is instead supplemented with [$^{13}C_6$]*myo*-inositol or an isotopomer of choice (see below). After the cells incorporated the ^{13}C label into their InsP pool to equilibrium (over 2 passages), cells were harvested and their water-soluble metabolites extracted and analyzed by BIRD-{^{1}H,^{13}C}HMQC-NMR. This NMR experiment detects ^{13}CH groups selectively over nonlabeled CH groups, making it particularly suitable for measuring the ^{13}C-labeled InsP pool within a complex background. The information from Figure 2 allowed us to annotate all detectable ^{13}C-labeled species from such extracts. Quantification was performed through relative integration of the signal corresponding to the 2-position against an internal standard and back-calculated to packed cell volumes. The annotation of the different InsPs in an HCT116 metabolic extract is shown exemplarily in Figure 3a (for full annotation see Figure S2). The same set of InsP species was observed in all other cell lines as well (Figures 3b and S3): the major labeled species include $InsP_6$, $InsP_5$[2OH], 1/3-glycerophospho-*myo*-inositol (1/3-GroPI), inositol 1- or 3-monophosphate (Ins(1/3)P), inositol 1,2- or 2,3-bisphosphate (Ins(1/3,2)P_2), inositol 2-monophosphate (Ins(2)P), and *myo*-inositol. All of these metabolite assignments were validated through spike-in experiments with commercially available InsP standards into ^{13}C-labeled metabolic extracts (Figure S4). Interestingly, labeling of *Schizosaccharomyces pombe* (Figure S5) revealed a somewhat different metabolite composition.

In order to differentiate the possible enantiomers in the mammalian InsP pool, we synthesized asymmetrically ^{13}C-labeled *myo*-inositols following our previously published

https://doi.org/10.1021/acscentsci.2c01032
ACS Cent. Sci. 2022, 8, 1683−1694

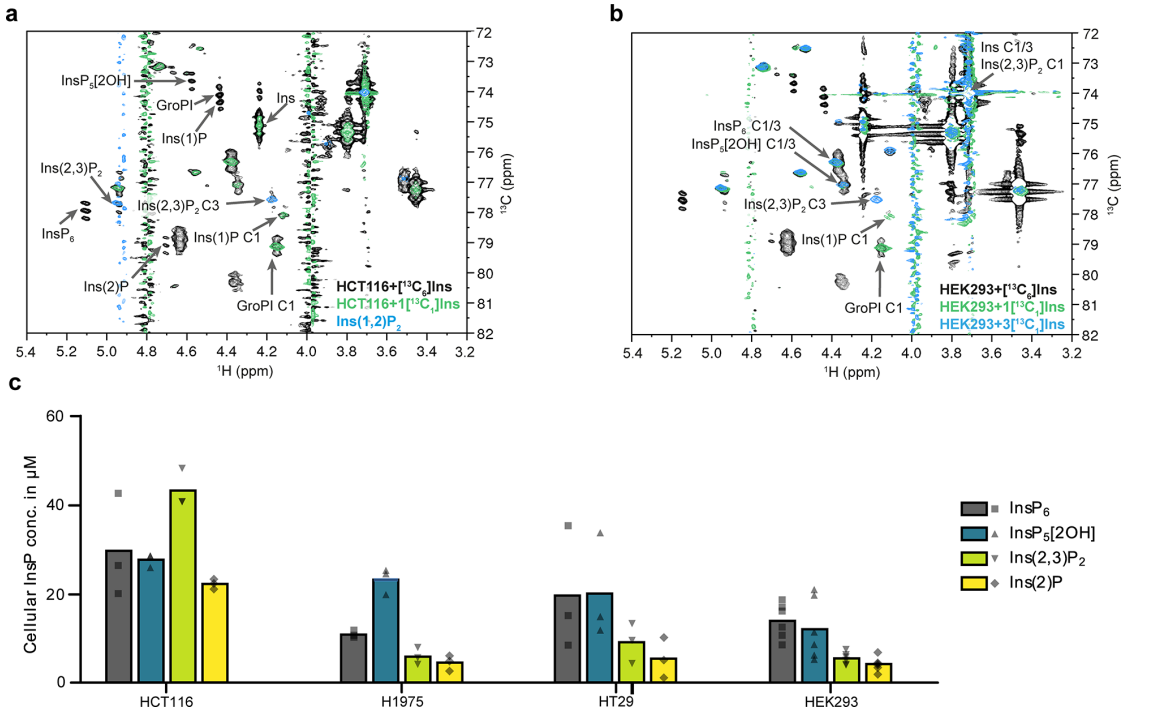


Figure 3. Identification and quantification of major InsPs in human cells. (a) Overlay of BIRD-{^{1}H,^{13}C}HMQC-NMR spectra of metabolic extracts from HCT116 cells which were labeled with either [$^{13}C_6$]*myo*-inositol (black spectrum) or 1[$^{13}C_1$]*myo*-inositol (green) and a reference spectrum of Ins(1,2)P_2 (blue). Annotation of identified InsPs was limited to the most important signals for clarity. Complete annotation is provided in Figure S2. (b) Overlay of BIRD-{^{1}H,^{13}C}HMQC-NMR spectra of metabolic extracts from HEK293 cells which were labeled with [$^{13}C_6$]*myo*-inositol (black), 1[$^{13}C_1$]*myo*-inositol (green), or 3[$^{13}C_1$]*myo*-inositol (blue). Annotation was limited to C1/3 positions for clarity. 1[$^{13}C_1$]*myo*-inositol-labeled 1 positions of GroPI and Ins(1)P confirm their enantiomeric identity. In contrast, the phosphorylated 3 position of Ins(2,3)P_2 is confirmed by labeling with 3[$^{13}C_1$]*myo*-inositol. (c) Scatter dot plot of quantified InsPs from metabolic extracts of various cells (HCT116, n = 3; H1975, n = 3; HT29, n = 3; HEK293, n = 6, biological replicates) with bars representing the means.

protocol.[38] Using the singly labeled isotopomer 1[$^{13}C_1$]*myo*-inositol and doubly labeled 4,5[$^{13}C_2$]Ins, respectively, we repeated the metabolic labeling in HEK293 and HCT116 cells. Focusing on the 1[$^{13}C_1$]*myo*-inositol labeling, the resulting spectra (Figures 3a, 3b, and S6a) show that the signals that correspond to the phosphorylated 1/3-positions of 1/3-GroPI and Ins(1/3)P are labeled, i.e., the enantiomers present in mammalian cells are 1-GroPI and Ins(1)P. The phosphorylated 1/3-position of Ins(1/3,2)P_2 is not labeled, which identifies Ins(2,3)P_2 as the prevalent enantiomer, an observation that was reproducible in both cell lines. To confirm this conclusion, HEK293 cells were also labeled with 3[$^{13}C_1$]*myo*-inositol. Now, the phosphorylated position of the putative $InsP_2$ remains labeled, unambiguously identifying Ins(2,3)P_2 as the main $InsP_2$ enantiomer present in human cell lines (Figure 3b).

GroPI and Ins(1)P are established products of cellular phosphatidylinositide turnover;[7,39] their detection was therefore anticipated. The presence of Ins(2,3)P_2 and Ins(2)P in the micromolar range (especially in HCT116 cells, see Figure 3c) was an unexpected observation. Ins(2,3)P_2 and Ins(2)P have not been associated with any established InsP-related pathway so far. Although Ins(2)P and Ins(1/3,2)P_2 were detected in 1995 by Mitchell and colleagues,[7,40] these metabolites received little attention and have been neglected since then.

Overall, the structural information contained in the HMQC-NMR spectra could be used to assign all detectable ^{13}C-labeled species in mammalian cells, and in combination with the asymmetrical inositol isotopomers, enantiomers could be resolved spectroscopically. This analysis uncovered high amounts of previously poorly characterized lower InsPs, which were not easily accessible with other analytical methods.

Formation of Ins(2,3)P_2 and Ins(2)P Is Dependent on MINPP1. In the biosynthetic pathway toward $InsP_6$ there are no InsP intermediates that are phosphorylated at the 2-position. The 2-phosphoryl group of $InsP_6$ is installed only in the last step, in which IPPK (inositol pentakisphosphate 2-kinase) converts $InsP_5$[2OH] to $InsP_6$. Ins(2,3)P_2 and Ins(2)P may therefore be generated downstream of $InsP_6$. A central InsP phosphatase is the mammalian phytase-like enzyme MINPP1, the only recognized $InsP_6$ phosphatase. To investigate possible relationships between Ins(2,3)P_2, Ins(2)P, and MINPP1, we turned our attention to cells lacking MINPP1. *MINPP1*$^{-/-}$ HEK293 cells were labeled with [$^{13}C_6$]*myo*-inositol, and the metabolites were analyzed by NMR (Figure 4). The *MINPP1*$^{-/-}$ cells exhibited slightly elevated $InsP_6$ levels and accumulated one new InsP species, which was assigned as $InsP_5$[3OH] or its enantiomer $InsP_5$[1OH] (Figure S4e). $InsP_5$[1/3OH] was not present in any investigated WT cell line. Labeling of *MINPP1*$^{-/-}$ HEK293 cells with the asymmetric isotopomers 1[$^{13}C_1$]*myo*-

https://doi.org/10.1021/acscentsci.2c01032
ACS Cent. Sci. 2022, 8, 1683−1694

ACS Central Science http://pubs.acs.org/journal/acscii Research Article

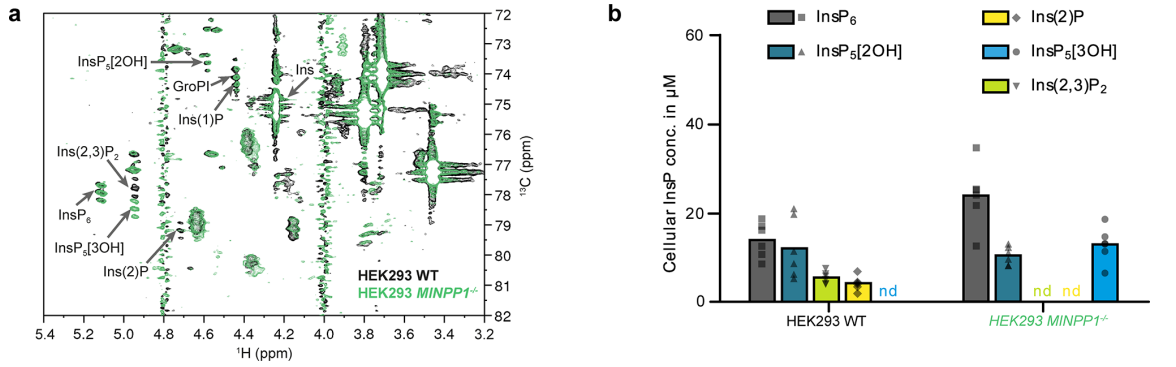


Figure 4. Identification of InsPs in HEK293 and *MINPP1*$^{-/-}$ HEK293 cells. (a) Overlay of [$^{13}C_6$]*myo*-inositol-labeled HEK293 (black) and *MINPP1*$^{-/-}$ HEK293 cells (green). Ins(2,3)P_2 and Ins(2)P are not observable in *MINPP1*$^{-/-}$ cells; instead, InsP_5[3OH] accumulates. (b) Scatter dot plot of quantified InsPs from these cell lines (WT, n = 6 same data as in Figure 3c for illustrative purposes; *MINPP1*$^{-/-}$, n = 6, biological replicates). Bars represent the means, nd = not detected. Enantiomer-specific identification of InsP_5[3OH] is shown in Figure S6.

inositol, 3[$^{13}C_1$] *myo*-inositol, and 4,5[$^{13}C_2$]*myo*-inositol unambiguously identified the InsP_5 in question as InsP_5[3OH] (Figure S6b, S6d, and S6e).

Strikingly, another change observed in the *MINPP1*$^{-/-}$ cell extracts was the complete absence of Ins(2,3)P_2 and Ins(2)P, establishing a connection between MINPP1 and these lower phosphorylated InsPs. The lack of an undefined InsP_2 species was also noted in a previous analysis of the same cell line using a radiolabeling approach.[21] Taking into consideration the only sparsely annotated intermediates and products of MINPP1-mediated dephosphorylation of InsP_6, it seemed possible that MINPP1 could generate Ins(2,3)P_2 and Ins(2)P directly from InsP_6.

MINPP1 Dephosphorylates InsP_5[2OH] and InsP_6 via Fully Distinct Pathways. To validate this hypothesis, we next sought to investigate the in vitro activity of MINPP1 against different InsPs. The expression and purification of recombinant MINPP1 in *E. coli* was optimized to isolate protein yields compatible with biochemical reactions on an NMR scale (Figures S7 and S8). Next, MINPP1 was incubated with fully $^{13}C_6$-labeled InsP_5[2OH], and the reaction was monitored using 2D NMR measurements. In the first experiments we chose a substrate concentration of 50 μM, which is in the middle to upper range of physiological concentrations (Figure S9).[7,34,41] To enable the detection and assignment of all intermediates, we subsequently increased the substrate concentration to 175 μM, which did not alter the overall outcome (Figure 5a). The structures of the intermediates were identified using the information from Figure 2 and additional cross-correlation NMR and spike-in experiments where necessary. In agreement with the annotation of MINPP1 as a 3-phosphatase, the first major intermediates for InsP_5[2OH] dephosphorylation are Ins(1,4,5,6)P_4 and subsequently Ins(1,4,5)P_3. MINPP1 therefore directly reverses the phosphorylation reactions catalyzed by IPMK (inositol phosphate multikinase).[24,42,43] Ins(1,4,5)P_3 is subsequently converted slowly to a mixture of different Ins$P_{1/2}$s (Figure 5b; a full scheme with all minor intermediates is shown in Figure S10).

We then proceeded to probe MINPP1-mediated dephosphorylation of InsP_6. In contrast to InsP_5[2OH] as a substrate, we observed a complex mixture of intermediates (Figure 5c). In addition, the overall conversion of InsP_6 was visibly slower. The two major reaction paths are depicted in Figure 5d (complete scheme in Figure S11): One dephosphorylation sequence proceeds via InsP_5[3OH] and Ins(1,2,6)P_3 as intermediates, and a second pathway generates Ins(1,2,3)P_3 as an intermediate via an unidentified (due to low abundance) InsP_5 isomer. Importantly, Ins(1/3,2)P_2 and Ins(2)P were observed as the final products of the dephosphorylation of InsP_6, validating that MINPP1 is capable of generating these InsPs directly from InsP_6.

To assess which enantiomers were formed during MINPP1-mediated dephosphorylation of InsP_6, we synthesized 1[$^{13}C_1$]-InsP_6.[38] The InsP_5, InsP_4, and InsP_3 intermediates which are produced by MINPP1 from 1[$^{13}C_1$]InsP_6 are enantiopure, as no dephosphorylation of the 1-position was observed (detailed explanation is given in Figures S12a and S12b). Surprisingly though, a mixture of Ins(1,2)P_2 and Ins(2,3)P_2 was formed during the later stages of the reaction (Figures S12c and S12d). The rather high ratio of Ins(2,3)P_2 to Ins(1,2)P_2 suggests that Ins(1,2)P_2 is formed exclusively via Ins(1,2,6)P_3, and Ins(1,2,3)P_3 is selectively converted to Ins(2,3)P_2. Both InsP_2s are, in turn, dephosphorylated to Ins(2)P. Our in vitro assessment of MINPP1 activity thus confirms the notion that MINPP1 can directly generate the novel cellular InsP species from InsP_6.

Another interesting observation, which runs counter to assumptions on MINPP1 activity,[6,25,44−46] is that the dephosphorylation sequences for InsP_6 and InsP_5[2OH] do not share any overlap (compare Figures S10, S11, S20, S26, and S27) because MINPP1 seems to be incapable of removing the phosphoryl group at the 2-position. Likely, the charged phosphoryl group on the only axial position of the *myo*-inositol scaffold plays a role in positioning the InsPs inside MINPP1's catalytic pocket.[47]

MINPP1 Exhibits Different Kinetic Properties toward InsP_5[2OH] and InsP_6. To characterize the kinetic properties of MINPP1, we next numerically determined the reaction rates of the dephosphorylation steps from the respective experimental data based on a time-independent rate model. We formulated the kinetics of the reaction network as a Master equation and approximated the corresponding rate matrix with a least-squares method that iteratively optimized the rates with respect to the scaled experimental data.[48,49] The reaction rates for the MINPP1 reaction starting with InsP_5[2OH] as a

https://doi.org/10.1021/acscentsci.2c01032
ACS Cent. Sci. 2022, 8, 1683−1694

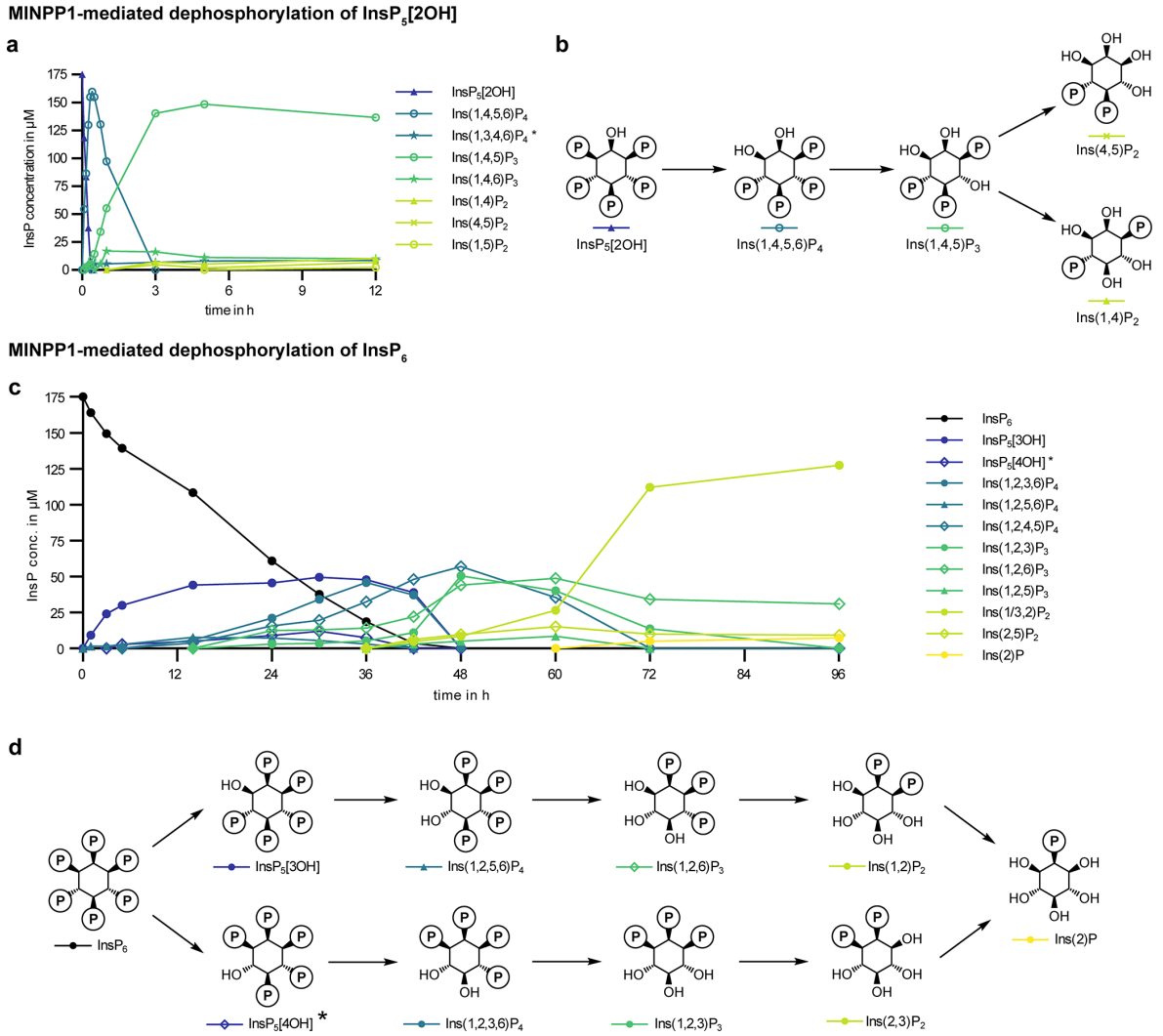
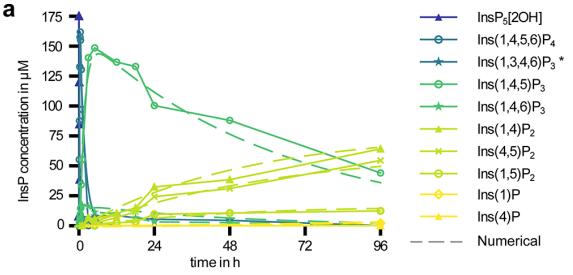


Figure 5. Dephosphorylation of $InsP_5[2OH]$ and $InsP_6$ by MINPP1 in vitro. (a) Progress curves of MINPP1 reaction with 175 μM $[^{13}C_6]InsP_5[2OH]$ showing the first 12 h of the reaction (for full scope of progress curves and progress curves at 50 μM substrate concentration see SI). Progress curves shown here are representative of two replicates. (b) Simplified reaction scheme of the MINPP1-mediated dephosphorylation of $InsP_5[2OH]$. Complete reaction scheme that includes all minor intermediates is in Figure S10. (c) Progress curves of MINPP1 reaction with 175 μM $[^{13}C_6]InsP_6$ with simplified reaction scheme depicting the two main reaction paths. Progress curves shown here are representative of two replicates. Corresponding progress curve for 50 μM substrate concentration in Figure S13. (d) Simplified reaction scheme for the dephosphorylation of $InsP_6$. Complete reaction scheme that includes all intermediates is in Figure S11. Note that the two enantiomers $Ins(1,2)P_2$ and $Ins(2,3)P_2$ are quantified together. (*) Structure of these InsPs could not be assigned with certainty due to low abundance and interference of more abundant signals.

substrate are shown in Figure 6b and were calculated from the experimental data (Figure 5a) and the corresponding network (Figure S10). The calculated rates predict progress curves (Figures 6a and S25) that are in good agreement with the experimental data, which supports the assumption of time-independent rates and thus the absence of inhibition processes. The highest reaction rate (k_20 equaling 330 nmol/(min mg enzyme)) also corresponds to the canonical MINPP1 activity toward $InsP_5[2OH]$ in the literature.[23]

However, in the case of $InsP_6$, the computational analysis of the experimental data (Figure 5c) with the network assumption depicted in Figure S11 yielded poor results; only the consumption of $InsP_6$ could be numerically analyzed with a rate of 9.3×10^{-4} min^{-1} (see SI). The poor fits indicate that the rates in the $InsP_6$ dephosphorylation network might not be time independent but are instead affected by inhibition processes that implicitly introduce a time dependence. Because of its relative stability and slow dephosphorylation, it seemed possible that $InsP_6$ could act as an inhibitor for the dephosphorylation of the MINPP1-generated intermediates.[23] This notion is further reinforced by the fact that the conversion of the intermediates progressed notably faster with lower $InsP_6$ starting concentrations (Figure S13). To test this, $[^{13}C_6]$-$InsP_5[2OH]$ was incubated with MINPP1 in the presence of

https://doi.org/10.1021/acscentsci.2c01032
ACS Cent. Sci. 2022, 8, 1683−1694



b

	Substrate	Product	Rate constants [min^{-1}]
k_20	InsP$_5$[2OH]	Ins(1,4,5,6)P$_4$	$9.8 \cdot 10^{-2}$
k_32	Ins(1,4,5,6)P$_4$	Ins(1,4,5)P$_3$	$9.3 \cdot 10^{-3}$
k_42*	Ins(1,4,5,6)P$_4$	Ins(1,4,6)P$_3$	$1.0 \cdot 10^{-3}$
k_53	Ins(1,4,5)P$_3$	Ins(1,4)P$_2$	$1.1 \cdot 10^{-4}$
k_63	Ins(1,4,5)P$_3$	Ins(1,5)P$_2$	$3.7 \cdot 10^{-5}$
k_73*	Ins(1,4,5)P$_3$	Ins(4,5)P$_2$	$1.0 \cdot 10^{-4}$
k_54	Ins(1,4,6)P$_3$	Ins(1,4)P$_2$	$3.5 \cdot 10^{-4}$
k_85	Ins(1,4)P$_2$	Ins(1)P	$4.0 \cdot 10^{-6}$
k_95	Ins(1,4)P$_2$	Ins(4)P	$1.2 \cdot 10^{-5}$
k_86	Ins(1,5)P$_2$	Ins(1)P	$4.5 \cdot 10^{-5}$
k_97	Ins(4,5)P$_2$	Ins(4)P	$1.0 \cdot 10^{-6}$

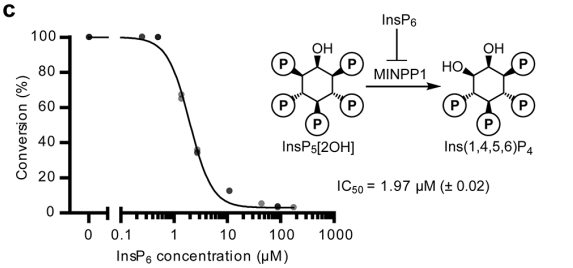


Figure 6. Numerical assessment of MINPP1 reaction rates. (a) Experimental and numerically approximated progress curves of MINPP1 dephosphorylation reactions with 175 μM InsP$_5$[2OH]. Solid lines represent the experimental data (same data as in Figure 5a). Dashed lines represent the progress curves predicted by the numerically determined reaction rates. (b) Numerically determined reaction rates representative of two replicates. Reaction rates marked with an asterisk (*) are subject to constraints. SI also includes attempted numerical approximation of the MINPP1 reaction with InsP$_6$. (c) Demonstration that InsP$_6$ can inhibit dephosphorylation of InsP$_5$[2OH] (175 μM) by MINPP1 (0.5 μM) with high potency. IC_{50} value is reported with standard error of $\log_{10} IC_{50}$ in brackets.

different amounts of [$^{13}C_6$]InsP$_6$. Indeed, a clear inhibitory effect of InsP$_6$ on the dephosphorylation of InsP$_5$[2OH] by MINPP1 was observed with an apparent IC_{50} value of 2 μM (Figure 6c). With changing substrate concentrations, the IC_{50} value also changed as predicted by the Cheng–Prusoff equation, indicating that this inhibition is likely competitive (Figure S14).[50]

Ins(2,3)P$_2$ and InsP$_5$[3OH] Are Biosynthetically Derived from InsP$_6$ In Cells. With the biochemical confirmation that MINPP1 can generate InsP$_5$[3OH], Ins(2,3)P$_2$, and Ins(2)P in vitro, we sought to perform metabolic flux analysis to confirm this reaction sequence in living cells. HEK293 or *MINPP1*$^{-/-}$ HEK293 cells were labeled with [$^{13}C_6$]*myo*-inositol to equilibrium and subsequently exposed to medium containing 4,5[$^{13}C_2$]*myo*-inositol for various periods of time before harvesting (Figure 7a). These two isotopomers were chosen to enable analysis by CE-MS: A mass difference of at least 2 Da allows the distinction of the differently labeled InsPs but also the differentiation of Ins(2,3)P$_2$ from other highly abundant, nonlabeled sugar bisphosphates. Following cell lysis, InsP mixtures were extracted with TiO_2 beads and analyzed via CE-MS to monitor the incorporation of the $^{13}C_2$-isotopomers and the decrease of the $^{13}C_6$-isotopomers simultaneously.

CE-MS analysis readily detected the expected [$^{13}C_6$]- and [$^{13}C_2$]InsP species. In addition, all samples contained around 3% of nonlabeled InsPs ($^{12}C_6$), which presumably stems from inositol neogenesis from glucose-6-phosphate.[51] The metabolic flux analysis (Figure 7b) indicates that exogenous *myo*-inositol is incorporated first into the pool of InsP$_5$[2OH], then into InsP$_6$, and last into Ins(2,3)P$_2$ (whose chemical identity was also confirmed with standards in CE-MS measurements, Figure S15). This incorporation sequence supports the hypothesis that Ins(2,3)P$_2$ is indeed derived from InsP$_6$ in human cells and is not an intermediate in the biosynthesis of InsP$_5$[2OH] or InsP$_6$ (Figure 7d).

In *MINPP1*$^{-/-}$ HEK293 cells, no Ins(2,3)P$_2$ was observed above the limit of detection, although the sensitivity of CE-MS is superior to NMR. Thus, CE-MS analysis confirms that generation of Ins(2,3)P$_2$ is dependent on MINPP1. Similarly, in the biosynthetic sequence, InsP$_5$[3OH] is generated after InsP$_6$ (Figure 7c and 7d), hinting at an unidentified 3-phosphatase activity acting on InsP$_6$, which has been suggested in the past.[16] Nevertheless, InsP$_5$[3OH] was not detectable in HEK293 WT cells.

DISCUSSION

We have expanded the detection and identification of complex InsP mixtures using different isotopomers of *myo*-inositol, InsP$_5$[2OH], and InsP$_6$ in both cellular and biochemical settings. Detection via NMR spectroscopy provided important structural information, enabling the assignment of previously poorly characterized InsPs. Application of asymmetrically labeled 1[$^{13}C_1$]*myo*-inositol, 3[$^{13}C_1$]*myo*-inositol, and 4,5-[$^{13}C_2$]*myo*-inositol readily facilitated the distinction of enantiomers in a complex sample, which has remained an analytical challenge to this day. InsP isotopomers with different masses also proved to be useful tools when used in combination with CE-MS analysis, as the higher sensitivity of this technique allows for detailed metabolic flux analyses.

Taking advantage of our labeled *myo*-inositol isotopomers and InsPs, we uncovered a branch of human InsP metabolism mediated by MINPP1, which was confirmed through in-depth characterization of MINPP1's reactivity in vitro and in cellula. The in vitro data illustrated that InsP$_5$[2OH] is the preferred substrate for MINPP1, compared to InsP$_6$. Under identical reaction conditions, InsP$_5$[2OH] was depleted with an apparent reaction rate that is 2 orders of magnitude higher than the rate for InsP$_6$ (9.8×10^{-2} versus 9.3×10^{-4} min^{-1} or ~330 versus ~3 nmol/(min mg enzyme), respectively). These activities are in line with previous kinetic analyses of mammalian MINPP1 (211 and 12 nmol/(min mg enzyme), respectively).[23] The subsequent slow dephosphorylation of Ins(1,4,5)P$_3$ in vitro (rate constant of 10^{-4} min^{-1} or 0.3 nmol/(min mg)) is likely not biologically significant as there are several other Ins(1,4,5)P$_3$ dephosphorylating enzymes with 4–5 magnitudes higher activity (5300–25 000 nmol/(min mg)).[54] Interestingly, depletion of cellular MINPP1 did not significantly alter InsP$_5$[2OH] levels, suggesting that other

https://doi.org/10.1021/acscentsci.2c01032
ACS Cent. Sci. 2022, 8, 1683–1694

ACS Central Science http://pubs.acs.org/journal/acscii Research Article

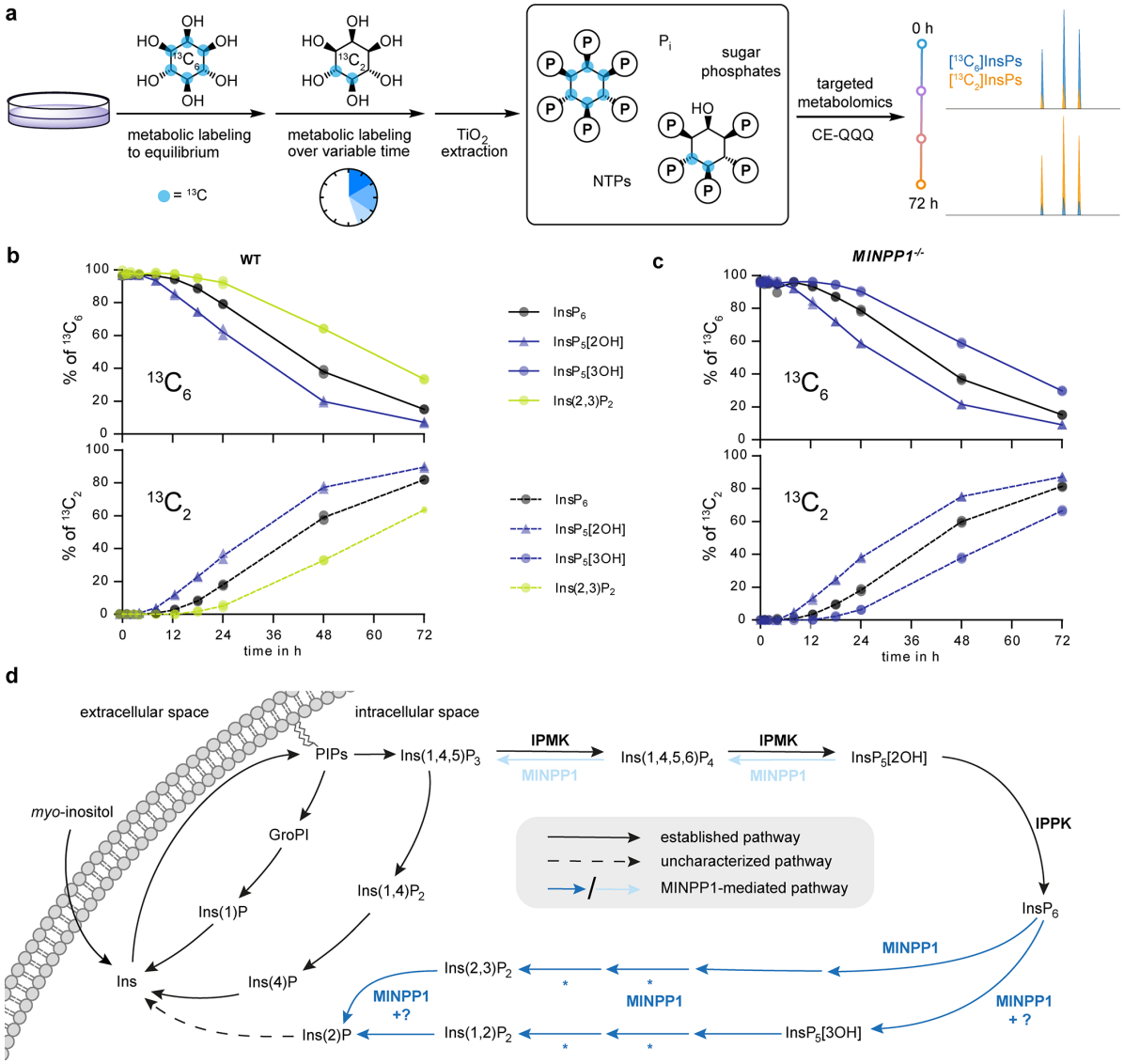


Figure 7. Metabolic flux analysis via time-dependent isotopic exchange of InsPs in HEK293 and *MINPP1*$^{-/-}$ HEK293 cells. (a) General workflow of the metabolic flux analysis. (b and c) Ratios of 6-fold ^{13}C-labeled and doubly ^{13}C-labeled InsPs HEK293 (b) and *MINPP1*$^{-/-}$ HEK293 (c) cells in TiO_2-extracted cell lysates. Data of two biological replicates are plotted individually, and means are connected with lines. All extracts contained a constant ~3% of nonlabeled InsP species, likely stemming from de novo inositol synthesis (Figure S16). Ins(2,3)P_2 in *MINPP1*$^{-/-}$ cells and InsP$_5$[3OH] in WT cells were below the limit of detection. (d) Updated overview of MINPP1-mediated InsP metabolism in human cells. As shown in this work, MINPP1 can dephosphorylate both InsP$_6$ (blue arrows) and InsP$_5$[2OH] (light blue arrows) via through two distinct, nonoverlapping metabolic pathways. Question mark hints toward unidentified phosphatase activities, which might explain the accumulation of InsP$_5$[3OH] observed in *MINPP1*$^{-/-}$ cells or how Ins(2,3)P_2 accumulates selectively in cells while both enantiomers are generated in vitro. Asterisks indicate that we cannot rule out the existence of additional phosphatases that might assist MINPP1-mediated dephosphorylation of InsP$_6$.

enzymes are able to dephosphorylate InsP$_5$[2OH] in a cellular setting.[52,53]

In contrast to the straightforward reaction paths for InsP$_5$[2OH] dephosphorylation by MINPP1, the dephosphorylation of InsP$_6$ occurs via an intricate network of intermediates. The first observable intermediates can be attributed to the 3-phosphatase activity of MINPP1; however, a significant portion of InsP$_6$ must initially be dephosphorylated at a different position because the symmetrical Ins(1,2,3)P_3 accumulates as an intermediate. Despite this complicated dephosphorylation network, the InsP$_6$ dephosphorylation sequence converges to two final compounds, Ins(1/3,2)P_2 and Ins(2)P in vitro. In all human cells we tested, Ins(2,3)P_2 and Ins(2)P were present at notable concentrations and constitute a hitherto uncharacterized part of mammalian InsP metabolism. It was somewhat surprising that Ins(2,3)P_2 is the predominant InsP$_2$ species within cells, given that MINPP1 is annotated as a 3-phosphatase. Our in vitro data demonstrate

https://doi.org/10.1021/acscentsci.2c01032
ACS Cent. Sci. 2022, 8, 1683−1694

that MINPP1 is capable of producing both enantiomers, Ins(1,2)P_2 and Ins(2,3)P_2, via the aforementioned dephosphorylation pathways from InsP_6. It thus seems feasible that Ins(1,2)P_2 can also be generated by MINPP1 in cells but may be depleted faster to Ins(2)P by either MINPP1 (which could be modified in its activity through post-translational modifications or different isoforms[31]) or a separate phosphatase altogether.

Remarkably, the many different dephosphorylation products of InsP_6 do not overlap with any intermediates of InsP_5[2OH] dephosphorylation, because MINPP1 appears incapable of removing the phosphoryl group at the 2-position of the inositol ring (Figure 7d). While MINPP1 converts InsP_5[2OH] to its biosynthetic precursors Ins(1,3,4,5)P_4 and Ins(1,4,5)P_3 in vitro, InsP_6 on the other hand is exclusively dephosphorylated to metabolites, which keep the phosphoryl group at the 2-position. This data is in stark contrast to the common assumption that MINPP1 would convert InsP_6 to InsP_5[2OH], as is often depicted in overview schemes on InsP metabolism.[6,25,44−46] It was shown in the past that the phosphoryl group at the 2-position of the *myo*-inositol ring (the only axial position) can play an important role for proper recognition of InsPs by protein binding partners.[55,56] Our data further corroborates the importance of the phosphorylation status of the 2-position (and thus IPPK activity) because it appears that InsPs may be "sorted" into the known and reversible InsP network (when InsPs contain a free hydroxyl group at the 2-position) or InsPs enter the slower, and potentially irreversible, MINPP1-mediated circuit where they remain phosphorylated at the 2-position.

While this sorting could be accomplished solely by the preferred dephosphorylation by MINPP1, the accessibility to the two different substrates InsP_5[2OH] and InsP_6 likely also plays a role. We found that the dephosphorylation of InsP_5[2OH] was strongly inhibited by low concentrations of InsP_6 in vitro (Figures 6c and S15). In the cellular context, this potent inhibitory effect of the abundant InsP_6 metabolite raises the question if, and how, MINPP1 can dephosphorylate InsP_5[2OH] at all. MINPP1 would need to access localized pools of said InsPs that are tightly regulated to either avoid or make use of the inhibitory effect. Interestingly, MINPP1 is thought to predominantly localize to the ER,[28] so how it accesses cytosolic (and presumably nuclear) InsPs is a question that has yet to be answered. While some studies have shown that MINPP1 (isoforms) might also be localized in cellular compartments other than the ER (Figure S17),[30] or could be even secreted,[31] tools to measure intracellular concentrations of different InsPs with spatial resolution are currently not available. An intriguing avenue for regulation could be that MINPP1 remains localized to intracellular organelles (ER or lysosomes) into which InsPs are controllably translocated and then dephosphorylated. This dephosphorylation could potentially proceed all the way to *myo*-inositol—with the aid of additional phosphatases—which could then be released through inositol transporters like SLC2A13 (HMIT). HMIT is known to be localized in intracellular membranes due to its ER-retention sequence and internalization sequence.[30,31,57−59]

Using asymmetrically isotope-labeled *myo*-inositol, it was possible to assign the uncharacterized InsP_5 isomer that accumulates in *MINPP1*$^{-/-}$ cells as InsP_5[3OH]. This accumulation appears counterintuitive, since MINPP1 is currently the only known enzyme in the human genome capable of generating InsP_5[3OH]. Nevertheless, Chi et al. also observed a residual 3-phosphatase activity in *MINPP1*$^{-/-}$ mice.[16] An analogous activity in human cells could be responsible for producing InsP_5[3OH] from InsP_6, as illustrated by our CE-MS-based metabolic flux analysis. Elucidating the identity of this 3-phosphatase will be of interest in the future as it constitutes an additional point of regulation within the InsP network. Furthermore, two recently reported cell lines with elevated intracellular phosphate levels were shown to contain a nonannotated InsP_5 isomer (which we assume is also InsP_5[1/3OH] based on the SAX-HPLC elution profiles).[51,60] Once the absolute configuration of these InsP_5 isomers has been determined, and ideally the enzymatic activities responsible for generating these isomers, the impact of cellular phosphate homeostasis on InsP signaling could be further explored.

The physiological role of the herein described dephosphorylation pathway for InsP_6 and its intermediates has yet to be explored. The InsPs produced by MINPP1 could be part of a recycling system converting InsP_6 back to Ins(2)P, which might be converted to *myo*-inositol by an inositol monophosphatase (although the lithium-sensitive human enzymes IMPA1/2 are not known to act on Ins(2)P[4,61]). As MINPP1 is a homologue of phytases, which take part in inositol recycling/scavenging, this possibility does not seem far fetched.[18] We cannot exclude the existence of other unknown phosphatases that contribute to this dephosphorylation pathway; however, the accumulation of InsP_5[3OH] in *MINPP1*$^{-/-}$ cells suggests that MINPP1 is obligatory for the dephosphorylation of InsP_5[3OH]. Furthermore, the complete absence of Ins(2,3)P_2 in *MINPP1*$^{-/-}$ cells indicates that MINPP1 must carry out the key dephosphorylation of InsP_6 on the path toward Ins(2,3)P_2. In addition, it remains to be investigated which enzymes can utilize the herein identified Ins(1/3,2)P_2 as substrates. Whether any of the InsP_6-derived MINPP1 products have signaling functions themselves is also an open question. It is possible that some MINPP1-generated InsPs (or the lack thereof) could be important contributing factors in MINPP1-regulated processes, i.e., ER stress, endochondral ossification, and neuronal function.[17,19,21] For example, it would be interesting to investigate if the hyperaccumulation of InsP_5[3OH] or the absence of Ins(2,3)P_2 and Ins(2)P is partially responsible for causing PCH in patients with MINPP1 loss-of-function mutations.[20,21] Ucuncu et al. proposed that hyperaccumulation of InsP_6 in neuronal cells of PCH patients might be a mechanistic cause of this disease by chelating iron ions.[21] In fact, all InsP species which possess the 1,2,3-phosphorylated motif might be capable of binding iron ions.[62] In contrast to their reported 3−4-fold increase of [^{3}H]InsP_6 levels (normalized against total tritiated PIPs) in *MINPP1*$^{-/-}$ HEK293 cells compared to WT cells, we only observed a slight increase using the same cell line but normalizing against packed cell volume. This discrepancy points toward several interesting possibilities: (a) PIP levels, or the incorporation of exogenous *myo*-inositol, could be (indirectly) influenced by MINPP1 activity, (b) radioactivity-induced cell stress could have an effect on MINPP1 expression,[17] or (c) knockout of MINPP1 changes the cell shape/volume. To differentiate between these possibilities, different quantification methods (e.g., normalization against total protein or DNA concentration) should be compared in the future, and the composition of PIP isotopomers during metabolic labeling experiments could be probed with mass spectrometry-based methods.[63−65]

https://doi.org/10.1021/acscentsci.2c01032
ACS Cent. Sci. 2022, 8, 1683−1694

As a next step, the combination of inositol isotopomers, NMR and CE-MS which we used in this study could be useful to probe InsP metabolism in a variety of biological contexts. For example, it could be investigated how the InsP pool changes during ER-related stress, during which MINPP1 is upregulated, and how this might correlate with the onset of apoptosis.[17] Another interesting application would be to determine the fate of inositol (phosphates) in pathogenic parasites such as *T. cruzi*, in which InsP metabolism is essential for the developmental cycle.[66] The question if or how InsP metabolism of the host cell and the parasite influences each other might lead to new therapeutic avenues for these parasitoses. The dissection of InsP degradation in an extracellular context, namely, how InsPs contained in food are converted by digestive processes or the gut microbiome and if the resulting metabolites might have beneficial or detrimental effects on health, is also a fascinating question.[47,67,68] With the tools and methods reported here, these topics now become addressable.

METHODS

All experimental and computational methods are described in the Supporting Information.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acscentsci.2c01032.

Quantification of cellular insotiol phosphates (XLSX)

Supplementary figures, methods for all in vitro and cellular experiments, and NMR spectra (PDF)

Methods and additional figures for the computational part of this work (PDF)

Transparent Peer Review report available (PDF)

AUTHOR INFORMATION

Corresponding Author

Dorothea Fiedler − *Leibniz-Forschungsinstitut für Molekulare Pharmakologie, 13125 Berlin, Germany; Institut für Chemie, Humboldt-Universität zu Berlin, 12489 Berlin, Germany;* orcid.org/0000-0002-0798-946X; Email: fiedler@fmp-berlin.de

Authors

Minh Nguyen Trung − *Leibniz-Forschungsinstitut für Molekulare Pharmakologie, 13125 Berlin, Germany; Institut für Chemie, Humboldt-Universität zu Berlin, 12489 Berlin, Germany*

Stefanie Kieninger − *Institut für Chemie und Biochemie, Freie Universität Berlin, 14195 Berlin, Germany;* orcid.org/0000-0002-7013-8537

Zeinab Fandi − *Leibniz-Forschungsinstitut für Molekulare Pharmakologie, 13125 Berlin, Germany*

Danye Qiu − *Institut für Organische Chemie, Albert-Ludwigs-Universität Freiburg, 79104 Freiburg, Germany*

Guizhen Liu − *Institut für Organische Chemie, Albert-Ludwigs-Universität Freiburg, 79104 Freiburg, Germany*

Neelay K. Mehendale − *Leibniz-Forschungsinstitut für Molekulare Pharmakologie, 13125 Berlin, Germany*

Adolfo Saiardi − *MRC Laboratory for Molecular Cell Biology, University College London, WC1E 6BT London, United Kingdom;* orcid.org/0000-0002-4351-0081

Henning Jessen − *Institut für Organische Chemie, Albert-Ludwigs-Universität Freiburg, 79104 Freiburg, Germany;* orcid.org/0000-0002-1025-9484

Bettina Keller − *Institut für Chemie und Biochemie, Freie Universität Berlin, 14195 Berlin, Germany*

Complete contact information is available at: https://pubs.acs.org/10.1021/acscentsci.2c01032

Author Contributions

M.N.T. performed most biochemical and in vitro experiments, data analyses, and organic syntheses. S.K. performed all kinetic modeling analyses under the supervision of B.K. Z.F. contributed to initial testing of MINPP1 expression and reactivity optimization. D.Q. and G.L. measured all CE-MS samples under the supervision of H.J. N.M. performed the cellular fractions and Western blots. D.F. and M.N.T. conceived the project with input from A.S. D.F. and M.N.T. supervised the experimental work by Z.F. M.N.T. and D.F. prepared the initial draft of the paper with S.K. contributing the kinetic modeling part, and all authors contributed to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

M.N.T. and S.K. were funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (EXC 2008-390540038) UniSysCat. S.K. was partially funded by Deutsche Forschungsgemeinschaft (DFG) through grant CRC 1114 "Scaling Cascades in Complex Systems", Project Number 235221301, Project B05 "Origin of scaling cascades in protein dynamics". We thank Peter Schmieder for providing valuable guidance with NMR topics, Lena von Oertzen and Kathrin Motzny for their assistance in cell culture matters, Robert Puschmann for initial synthesis of *myo*-inositols, and all members of the Fiedler lab for proofreading.

REFERENCES

(1) Watson, P. J.; Fairall, L.; Santos, G. M.; Schwabe, J. W. R. Structure of HDAC3 Bound to Co-Repressor and Inositol Tetraphosphate. *Nature* **2012**, *481* (7381), 335−340.

(2) Wang, Q.; Vogan, E. M.; Nocka, L. M.; Rosen, C. E.; Zorn, J. A.; Harrison, S. C.; Kuriyan, J. Autoinhibition of Bruton's Tyrosine Kinase (Btk) and Activation by Soluble Inositol Hexakisphosphate. *Elife* **2015**, *4*, e06074.

(3) Blind, R. D. Structural Analyses of Inositol Phosphate Second Messengers Bound to Signaling Effector Proteins. *Adv. Biol. Regul.* **2020**, *75*, 100667.

(4) Kanehisa, M. Toward Understanding the Origin and Evolution of Cellular Organisms. *Protein Sci.* **2019**, *28* (11), 1947−1951.

(5) Irvine, R. F.; Schell, M. J. Back in the Water: The Return of the Inositol Phosphates. *Nat. Rev. Mol. Cell Biol.* **2001**, *2* (5), 327−338.

(6) Chatree, S.; Thongmaen, N.; Tantivejkul, K.; Sitticharoon, C.; Vucenik, I. Role of Inositols and Inositol Phosphates in Energy Metabolism. *Molecules* **2020**, *25* (21), 5079.

(7) Barker, C. J.; Wright, J.; Hughes, P. J.; Kirk, C. J.; Michell, R. H. Complex Changes in Cellular Inositol Phosphate Complement Accompany Transit through the Cell Cycle. *Biochem. J.* **2004**, *380* (2), 465−473.

(8) Qiu, D.; Wilson, M. S.; Eisenbeis, V. B.; Harmel, R. K.; Riemer, E.; Haas, T. M.; Wittwer, C.; Jork, N.; Gu, C.; Shears, S. B.; Schaaf, G.; Kammerer, B.; Fiedler, D.; Saiardi, A.; Jessen, H. J. Analysis of

Inositol Phosphate Metabolism by Capillary Electrophoresis Electrospray Ionization Mass Spectrometry. *Nat. Commun.* **2020**, *11*, 1−12.
(9) Letcher, A. J.; Schell, M. J.; Irvine, R. F. Do Mammals Make All Their Own Inositol Hexakisphosphate? *Biochem. J.* **2008**, *416* (2), 263−270.
(10) Nguyen Trung, M.; Furkert, D.; Fiedler, D. Versatile Signaling Mechanisms of Inositol Pyrophosphates. *Curr. Opin. Chem. Biol.* **2022**, *70*, 102177.
(11) Dick, R. A.; Zadrozny, K. K.; Xu, C.; Schur, F. K. M.; Lyddon, T. D.; Ricana, C. L.; Wagner, J. M.; Perilla, J. R.; Ganser-Pornillos, B. K.; Johnson, M. C.; Pornillos, O.; Vogt, V. M. Inositol Phosphates Are Assembly Co-Factors for HIV-1. *Nature* **2018**, *560* (7719), 509−512.
(12) Watson, P. J.; Millard, C. J.; Riley, A. M.; Robertson, N. S.; Wright, L. C.; Godage, H. Y.; Cowley, S. M.; Jamieson, A. G.; Potter, B. V. L.; Schwabe, J. W. R. Insights into the Activation Mechanism of Class I HDAC Complexes by Inositol Phosphates. *Nat. Commun. 2016 71* **2016**, *7* (1), 1−13.
(13) Lin, H.; Yan, Y.; Luo, Y.; So, W. Y.; Wei, X.; Zhang, X.; Yang, X.; Zhang, J.; Su, Y.; Yang, X.; Zhang, B.; Zhang, K.; Jiang, N.; Chow, B. K. C.; Han, W.; Wang, F.; Rao, F. IP6-Assisted CSN-COP1 Competition Regulates a CRL4-ETV5 Proteolytic Checkpoint to Safeguard Glucose-Induced Insulin Secretion. *Nat. Commun. 2021 121* **2021**, *12* (1), 1−13.
(14) Lin, H.; Zhang, X.; Liu, L.; Fu, Q.; Zang, C.; Ding, Y.; Su, Y.; Xu, Z.; He, S.; Yang, X.; Wei, X.; Mao, H.; Cui, Y.; Wei, Y.; Zhou, C.; Du, L.; Huang, N.; Zheng, N.; Wang, T.; Rao, F. Basis for Metabolite-Dependent Cullin-RING Ligase Deneddylation by the COP9 Signalosome. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117* (8), 4117−4124.
(15) Köhn, M. Turn and Face the Strange: A New View on Phosphatases. *ACS Cent. Sci.* **2020**, *6* (4), 467−477.
(16) Chi, H.; Yang, X.; Kingsley, P. D.; O'Keefe, R. J.; Puzas, J. E.; Rosier, R. N.; Shears, S. B.; Reynolds, P. R. Targeted Deletion of Minpp1 Provides New Insight into the Activity of Multiple Inositol Polyphosphate Phosphatase In Vivo. *Mol. Cell. Biol.* **2000**, *20* (17), 6496−6507.
(17) Kilaparty, S. P.; Agarwal, R.; Singh, P.; Kannan, K.; Ali, N. Endoplasmic Reticulum Stress-Induced Apoptosis Accompanies Enhanced Expression of Multiple Inositol Polyphosphate Phosphatase 1 (Minpp1): A Possible Role for Minpp1 in Cellular Stress Response. *Cell Stress Chaperones* **2016**, *21* (4), 593−608.
(18) Kilaparty, S. P.; Singh, A.; Baltosser, W. H.; Ali, N. Computational Analysis Reveals a Successive Adaptation of Multiple Inositol Polyphosphate Phosphatase 1 in Higher Organisms through Evolution. *Evol. Bioinform. Online* **2014**, *10*, 239−250.
(19) Caffrey, J. J.; Hidaka, K.; Matsuda, M.; Hirata, M.; Shears, S. B. The Human and Rat Forms of Multiple Inositol Polyphosphate Phosphatase: Functional Homology with a Histidine Acid Phosphatase up-Regulated during Endochondral Ossification. *FEBS Lett.* **1999**, *442* (1), 99−104.
(20) Appelhof, B.; Wagner, M.; Hoefele, J.; Heinze, A.; Roser, T.; Koch-Hogrebe, M.; Roosendaal, S. D.; Dehghani, M.; Mehrjardi, M. Y. V.; Torti, E.; Houlden, H.; Maroofian, R.; Rajabi, F.; Sticht, H.; Baas, F.; Wieczorek, D.; Jamra, R. A. Pontocerebellar Hypoplasia Due to Bi-Allelic Variants in MINPP1. *Eur. J. Hum. Genet.* **2021**, *29*, 411.
(21) Ucuncu, E.; Rajamani, K.; Wilson, M. S. C.; Medina-Cano, D.; Altin, N.; David, P.; Barcia, G.; Lefort, N.; Banal, C.; Vasilache-Dangles, M. T.; Pitelet, G.; Lorino, E.; Rabasse, N.; Bieth, E.; Zaki, M. S.; Topcu, M.; Sonmez, F. M.; Musaev, D.; Stanley, V.; Bole-Feysot, C.; Nitschké, P.; Munnich, A.; Bahi-Buisson, N.; Fossoud, C.; Giuliano, F.; Colleaux, L.; Burglen, L.; Gleeson, J. G.; Boddaert, N.; Saiardi, A.; Cantagrel, V. MINPP1 Prevents Intracellular Accumulation of the Chelator Inositol Hexakisphosphate and Is Mutated in Pontocerebellar Hypoplasia. *Nat. Commun.* **2020**, *11* (1), 6087.
(22) Gene ID: 9562, Homo sapiens, MINPP1 multiple inositol-polyphosphate phosphatase 1. Gene [Internet]. National Library of Medicine (US), National Center for Biotechnology Information: Bethesda, MD; https://www.ncbi.nlm.nih.gov/gene/9562 (accessed 2022−08−26).
(23) Nogimori, K.; Hughes, P. J.; Glennon, M. C.; Hodgson, M. E.; Putney, J. W.; Shears, S. B. Purification of an Inositol (1,3,4,5)-Tetrakisphosphate 3-Phosphatase Activity from Rat Liver and the Evaluation of Its Substrate Specificity. *J. Biol. Chem.* **1991**, *266* (25), 16499−16506.
(24) Craxton, A.; Caffrey, J. J.; Burkhart, W.; Safrany, T. S.; Shears, B. S. Molecular Cloning and Expression of a Rat Hepatic Multiple Inositol Polyphosphate Phosphatase. *Biochem. J.* **1997**, *328* (1), 75−81.
(25) Yu, J.; Leibiger, B.; Yang, S. N.; Caffery, J. J.; Shears, S. B.; Leibiger, I. B.; Barker, C. J.; Berggren, P. O. Cytosolic Multiple Inositol Polyphosphate Phosphatase in the Regulation of Cytoplasmic Free Ca^{2+} Concentration. *J. Biol. Chem.* **2003**, *278* (47), 46210−46218.
(26) Barker, C. J.; Illies, C.; Berggren, P.-O. HPLC Separation of Inositol Polyphosphates. In *Inositol Phosphates and Lipids*; Barker, C. J., Ed.; Humana Press: New York, 2010; pp 21−46 DOI: 10.1007/978-1-60327-175-2_2.
(27) Shears, S. B. A Short Historical Perspective of Methods in Inositol Phosphate Research. In *Inositol Phosphates*; Miller, G., Ed.; Springer US: New York, 2020; pp 1−28 DOI: 10.1007/978-1-0716-0167-9_1.
(28) Ali, N.; Craxton, A.; Shears, S. B. Hepatic Ins(1,3,4,5)P_4 3-Phosphatase Is Compartmentalized inside Endoplasmic Reticulum. *J. Biol. Chem.* **1993**, *268* (9), 6161−6167.
(29) Craxton, A.; Ali, N.; Shears, S. B. Comparison of the Activities of a Multiple Inositol Polyphosphate Phosphatase Obtained from Several Sources: A Search for Heterogeneity in This Enzyme. *Biochem. J.* **1995**, *305* (2), 491−498.
(30) Windhorst, S.; Lin, H.; Blechner, C.; Fanick, W.; Brandt, L.; Brehm, M. A.; Mayr, G. W. Tumour Cells Can Employ Extracellular Ins(1,2,3,4,5,6)P_6 and Multiple Inositol-Polyphosphate Phosphatase 1 (MINPP1) Dephosphorylation to Improve Their Proliferation. *Biochem. J.* **2013**, *450* (1), 115−125.
(31) Zubair, M.; Hamzah, R.; Griffin, R.; Ali, N. Identification and Functional Characterization of Multiple Inositol Polyphosphate Phosphatase1 (Minpp1) Isoform-2 in Exosomes with Potential to Modulate Tumor Microenvironment. *PLoS One* **2022**, *17*, e0264451.
(32) Brown, N. W.; Marmelstein, A. M.; Fiedler, D. Chemical Tools for Interrogating Inositol Pyrophosphate Structure and Function. *Chem. Soc. Rev.* **2016**, *45*, 6311−6326.
(33) Wilson, M. S. C.; Bulley, S. J.; Pisani, F.; Irvine, R. F.; Saiardi, A. A Novel Method for the Purification of Inositol Phosphates from Biological Samples Reveals That No Phytate Is Present in Human Plasma or Urine. *Open Biol.* **2015**, *5* (3), 150014.
(34) Qiu, D.; Eisenbeis, V. B.; Saiardi, A.; Jessen, H. J. Absolute Quantitation of Inositol Pyrophosphates by Capillary Electrophoresis Electrospray Ionization Mass Spectrometry. *J. Vis. Exp.* **2021**, *2021* (174), 1−13.
(35) Losito, O.; Szijgyarto, Z.; Resnick, A. C.; Saiardi, A. Inositol Pyrophosphates and Their Unique Metabolic Complexity: Analysis by Gel Electrophoresis. *PLoS One* **2009**, *4* (5), e5580.
(36) Wilson, M. S. C.; Saiardi, A. Importance of Radioactive Labelling to Elucidate Inositol Polyphosphate Signalling. *Top. Curr. Chem.* **2017**, *375* (1), 1−21.
(37) Mayr, G. W. A Novel Metal-Dye Detection System Permits Picomolar-Range h.p.l.c. Analysis of Inositol Polyphosphates from Non-Radioactively Labelled Cell or Tissue Specimens. *Biochem. J.* **1988**, *254*, 585−591.
(38) Harmel, R. K.; Puschmann, R.; Nguyen Trung, M.; Saiardi, A.; Schmieder, P.; Fiedler, D. Harnessing ^{13}C-Labeled Myo-Inositol to Interrogate Inositol Phosphate Messengers by NMR. *Chem. Sci.* **2019**, *10* (20), 5267−5274.
(39) Corda, D.; Zizza, P.; Varone, A.; Filippi, B. M.; Mariggiò, S. The Glycerophosphoinositols: Cellular Metabolism and Biological Functions. *Cell. Mol. Life Sci.* **2009**, *66* (21), 3449−3467.
(40) Barker, C. J.; French, P. J.; Moore, A. J.; Nilsson, T.; Berggren, P. O.; Bunce, C. M.; Kirk, C. J.; Michell, R. H. Inositol 1,2,3-Trisphosphate and Inositol 1,2- and/or 2,3-Bisphosphate Are Normal

Constituents of Mammalian Cells. *Biochem. J.* **1995**, *306* (2), 557−564.
(41) Mountford, J. C.; Bunce, C. M.; French, P. J.; Michell, R. H.; Brown, G. Intracellular Concentrations of Inositol, Glycerophosphoinositol and Inositol Pentakisphosphate Increase during Haemopoietic Cell Differentiation. *BBA - Mol. Cell Res.* **1994**, *1222* (1), 101−108.
(42) Chang, S. C.; Miller, A. L.; Feng, Y.; Wente, S. R.; Majerus, P. W. The Human Homolog of the Rat Inositol Phosphate Multikinase Is an Inositol 1,3,4,6-Tetrakisphosphate 5-Kinase. *J. Biol. Chem.* **2002**, *277* (46), 43836−43843.
(43) Saiardi, A.; Nagata, E.; Luo, H. R.; Sawa, A.; Luo, X.; Snowman, A. M.; Snyder, S. H. Mammalian Inositol Polyphosphate Multikinase Synthesizes Inositol 1,4,5-Trisphosphate and an Inositol Pyrophosphate. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98* (5), 2306−2311.
(44) Gu, C.; Liu, J.; Liu, X.; Zhang, H.; Luo, J.; Wang, H.; Locasale, J. W.; Shears, S. B. Metabolic Supervision by PPIP5K, an Inositol Pyrophosphate Kinase/Phosphatase, Controls Proliferation of the HCT116 Tumor Cell Line. *Proc. Natl. Acad. Sci. U. S. A.* **2021**, *118* (10), 1−8.
(45) Thomas, M. P.; Potter, B. V. L. The Enzymes of Human Diphosphoinositol Polyphosphate Metabolism. *FEBS J.* **2014**, *281* (1), 14−33.
(46) Li, Q.; Shortreed, M.; Wenger, C.; Frey, B.; Schaffer, L.; Scalf, M.; Smith, L. Global Post-Translational Modification Discovery. *J. Proteome Res.* **2017**, *16* (4), 1383−1390.
(47) Acquistapace, I. M.; Thompson, E. J.; Kühn, I.; Bedford, M. R.; Brearley, C. A.; Hemmings, A. M. Insights to the Structural Basis for the Stereospecificity of the Escherichia Coli Phytase, AppA. *Int. J. Mol. Sci.* **2022**, *23* (11), 6346.
(48) Gardiner, C. W. *Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences*, 2nd ed.; Springer Berlin Heidelberg, 1985.
(49) Van Kampen, N. G. *Stochastic Processes*, 1st ed.; North-Holland Publishing Co.: Amsterdam, New York, Oxford, 1981.
(50) Yung-Chi, C.; Prusoff, W. H. Relationship between the Inhibition Constant (KI) and the Concentration of Inhibitor Which Causes 50 per Cent Inhibition (I50) of an Enzymatic Reaction. *Biochem. Pharmacol.* **1973**, *22* (23), 3099−3108.
(51) Desfougères, Y.; Wilson, M. S. C.; Laha, D.; Miller, G. J.; Saiardi, A. ITPK1Mediates the Lipid-Independent Synthesis of Inositol Phosphates Controlled by Metabolism. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116* (49), 24551−24561.
(52) Chamberlain, P. P.; Qian, X.; Stiles, A. R.; Cho, J.; Jones, D. H.; Lesley, S. A.; Grabau, E. A.; Shears, S. B.; Spraggon, G. Integration of Inositol Phosphate Signaling Pathways via Human ITPK1. *J. Biol. Chem.* **2007**, *282* (38), 28117−28125.
(53) Caffrey, J. J.; Darden, T.; Wenk, M. R.; Shears, S. B. Expanding Coincident Signaling by PTEN through Its Inositol 1,3,4,5,6-Pentakisphosphate 3-Phosphatase Activity. *FEBS Lett.* **2001**, *499* (1−2), 6−10.
(54) Zhang, X.; Jefferson, A. B.; Auethavekiat, V.; Majerus, P. W. The Protein Deficient in Lowe Syndrome Is a Phosphatidylinositol-4,5-Bisphosphate 5-Phosphatase. *Proc. Natl. Acad. Sci. U. S. A.* **1995**, *92* (11), 4853−4856.
(55) Furkert, D.; Hostachy, S.; Nadler-Holly, M.; Fiedler, D. Triplexed Affinity Reagents to Sample the Mammalian Inositol Pyrophosphate Interactome. *Cell Chem. Biol.* **2020**, *27* (8), 1097−1108.
(56) Wild, R.; Gerasimaite, R.; Jung, J.-Y.; Truffault, V.; Pavlovic, I.; Schmidt, A.; Saiardi, A.; Jessen, H. J.; Poirier, Y.; Hothorn, M.; Mayer, A. Control of Eukaryotic Phosphate Homeostasis by Inositol Polyphosphate Sensor Domains. *Science* **2016**, *352* (6288), 986−990.
(57) Su, X. B.; Ko, A.-L. A.; Saiardi, A. Regulations of Myo-Inositol Homeostasis: Mechanisms, Implications, and Perspectives. *Adv. Biol. Regul.* **2022**, 100921.
(58) Uldry, M.; Steiner, P.; Zurich, M.-G.; Béguin, P.; Hirling, H.; Dolci, W.; Thorens, B. Regulated Exocytosis of an H^+/Myo-Inositol Symporter at Synapses and Growth Cones. *EMBO J.* **2004**, *23* (3), 531−540.
(59) Daniel, E. Di; Kew, J. N.; Maycox, P. R. Investigation of the H^+-Myo-Inositol Transporter (HMIT) as a Neuronal Regulator of Phosphoinositide Signalling. *Biochem. Soc. Trans.* **2009**, *37* (5), 1139−1143.
(60) López-Sánchez, U.; Tury, S.; Nicolas, G.; Wilson, M. S.; Jurici, S.; Ayrignac, X.; Courgnaud, V.; Saiardi, A.; Sitbon, M.; Battini, J. L. Interplay between Primary Familial Brain Calcification-Associated SLC20A2 and XPR1 Phosphate Transporters Requires Inositol Polyphosphates for Control of Cellular Phosphate Homeostasis. *J. Biol. Chem.* **2020**, *295* (28), 9366−9378.
(61) ENZYME: 3.1.3.25. KEGG [Internet]; https://www.genome.jp/entry/3.1.3.25 (accessed 2022−08−26).
(62) Veiga, N.; Torres, J.; Mansell, D.; Freeman, S.; Domínguez, S.; Barker, C. J.; Díaz, A.; Kremer, C. Chelatable Iron Pool": Inositol 1,2,3-Trisphosphate Fulfils the Conditions Required to Be a Safe Cellular Iron Ligand. *J. Biol. Inorg. Chem.* **2009**, *14* (1), 51−59.
(63) Kim, Y.; Shanta, S. R.; Zhou, L.-H.; Kim, K. P. Mass Spectrometry Based Cellular Phosphoinositides Profiling and Phospholipid Analysis: A Brief Review. *Exp. Mol. Med.* **2010**, *42* (1), 1.
(64) Pettitt, T. R.; Dove, S. K.; Lubben, A.; Calaminus, S. D. J.; Wakelam, M. J. O. Analysis of Intact Phosphoinositides in Biological Samples. *J. Lipid Res.* **2006**, *47* (7), 1588−1596.
(65) Milne, S. B.; Ivanova, P. T.; DeCamp, D.; Hsueh, R. C.; Brown, H. A. A Targeted Mass Spectrometric Analysis of Phosphatidylinositol Phosphate Species. *J. Lipid Res.* **2005**, *46* (8), 1796−1802.
(66) Mantilla, B. S.; Amaral, L. D. D.; Jessen, H. J.; Docampo, R. The Inositol Pyrophosphate Biosynthetic Pathway of Trypanosoma Cruzi. *ACS Chem. Biol.* **2021**, *16* (2), 283−292.
(67) Sakamoto, K.; Vucenik, I.; Shamsuddin, A. M. [^{3}H]Phytic Acid (Inositol Hexaphosphate) Is Absorbed and Distributed to Various Tissues in Rats. *J. Nutr.* **1993**, *123* (4), 713−720.
(68) Bui, T. P. N.; Mannerås-Holm, L.; Puschmann, R.; Wu, H.; Troise, A. D.; Nijsse, B.; Boeren, S.; Bäckhed, F.; Fiedler, D.; deVos, W. M. Conversion of Dietary Inositol into Propionate and Acetate by Commensal Anaerostipes Associates with Host Health. *Nat. Commun.* **2021**, *12* (1), 1−16.

https://doi.org/10.1021/acscentsci.2c01032
ACS Cent. Sci. 2022, 8, 1683−1694

Supporting Information

Stable isotopomers of *myo*-inositol uncover a complex MINPP1-dependent inositol phosphate network

Minh Nguyen Trung[1,2], Stefanie Kieninger[3], Zeinab Fandi[1], Danye Qiu[4], Guizhen Liu[4], Neelay K. Mehendale[1], Adolfo Saiardi[5], Henning Jessen[4], Bettina Keller[3], Dorothea Fiedler[1,2*]

Affiliations

[1] Leibniz-Forschungsinstitut für Molekulare Pharmakologie, Robert-Rössle-Straße 10, 13125 Berlin, Germany

[2] Institut für Chemie, Humboldt-Universität zu Berlin, Brook-Taylor-Straße 2, 12489 Berlin, Germany

[3] Institut für Chemie und Biochemie, Freie Universität Berlin, Arnimallee 22, 14195 Berlin, Germany

[4] Institut für Organische Chemie, Albert-Ludwigs-Universität Freiburg, Albertstraße 21, 79104 Freiburg, Germany

[5] MRC Laboratory for Molecular Cell Biology, University College London, WC1E 6BT London, UK

* corresponding author: fiedler@fmp-berlin.de

Table of Contents

Abbreviations

ACN	acetonitrile
BIRD	bilinear rotation decoupling
BIRD-HMQC	HMQC with BIRD pulse
BPG	2,3-bisphoshpoglycerate
CD	circular dichroism spectroscopy
CE-MS	capillary electrophoresis electrospray mass spectrometry
4,5-DCI	4,5-dicyanoimidazole
DCl	deuterium chloride
DCM	dichloromethane
DMEM	Dulbecco's Modified Eagle Medium
DMSO	dimethylsulfoxide
DTT	dithiothreitol
EDTA	ethylenediaminetetraacetic acid
FBS	fetal bovine serum
GndHCl	guanidium hydrochloride
GroPI	glycerophosphoinositol
HMQC	heteronuclear multiple-quantum correlation
Ins	*myo*-inositol
InsPx	inositol phosphate
IPS	inositol phosphate synthase
IPTG	isopropyl *β*-D-1-thiogalactopyranoside
MINPP1	multiple inositol polyphosphate phosphatase 1
MWCO	molecular weight cut-off
NAD^+	nicotinamide adenine dinucleotide (oxidized form)
NaOD	sodium deuteroxide
NMR	nuclear magnetic resonance (spectroscopy)
OD_{600}	optical density (at 600 nm)
ORF	open reading frame
PCV	packed cell volume
ppm	parts per million
rt	room temperature
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
TB	terrific broth
TMPBr	tetramethylphosphonium bromide

Supporting Figures and Tables

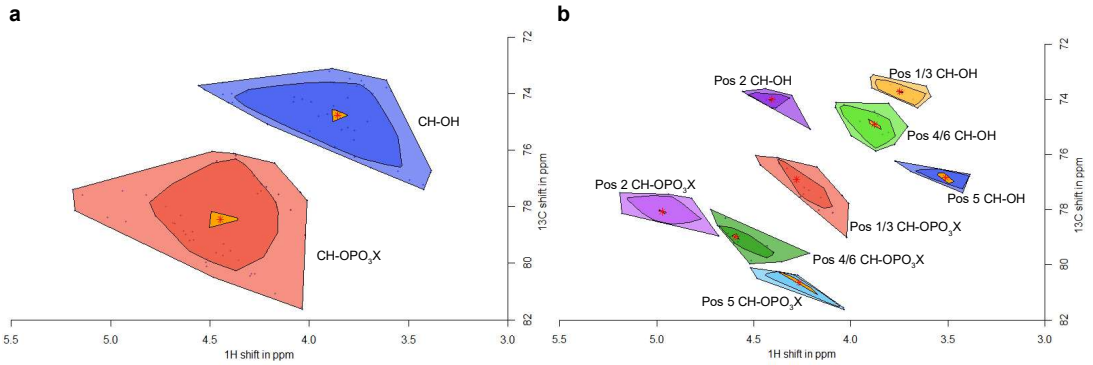


Figure S1: The bagplots illustrate the clustering depending on phosphororylation state (**a**) (blue: OH groups, red: phosphorylated groups) and position on the inositol ring (**b**). A fence factor of 6 was used to include all data points in the respective bags, the underlying data is the same as shown in Figure 2.

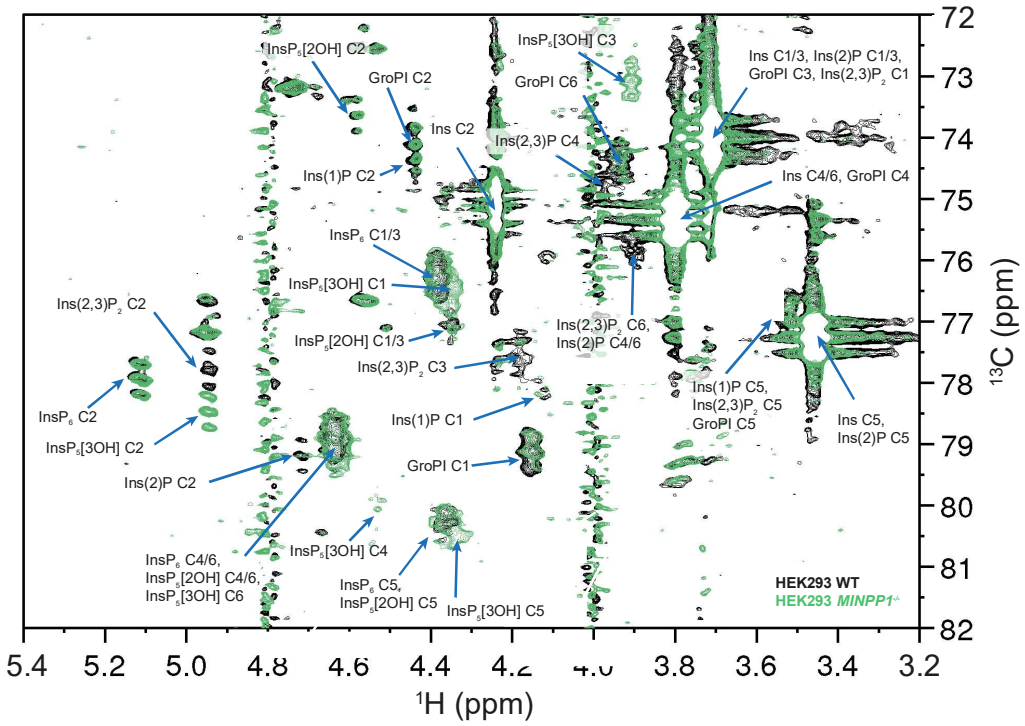


Figure S2: Complete annotation of [$^{13}C_6$]Ins-labeled HEK293 WT (black spectrum) and *MINPP1*$^{-/-}$ (green spectrum) metabolic extracts.

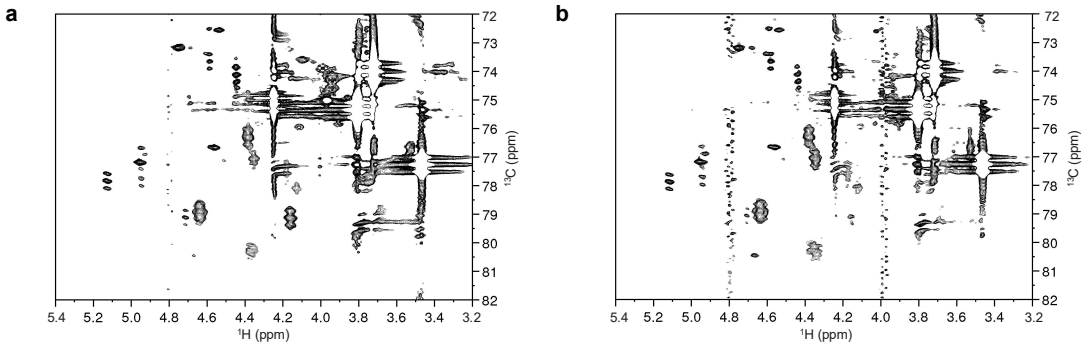


Figure S3: Additional NMR spectra of metabolic extracts from immortalized human wild-type cells. (**a**): HT29, (**b**): H1975. All labeled wild-type cells lines contain the same set of InsPs with varying concentrations.

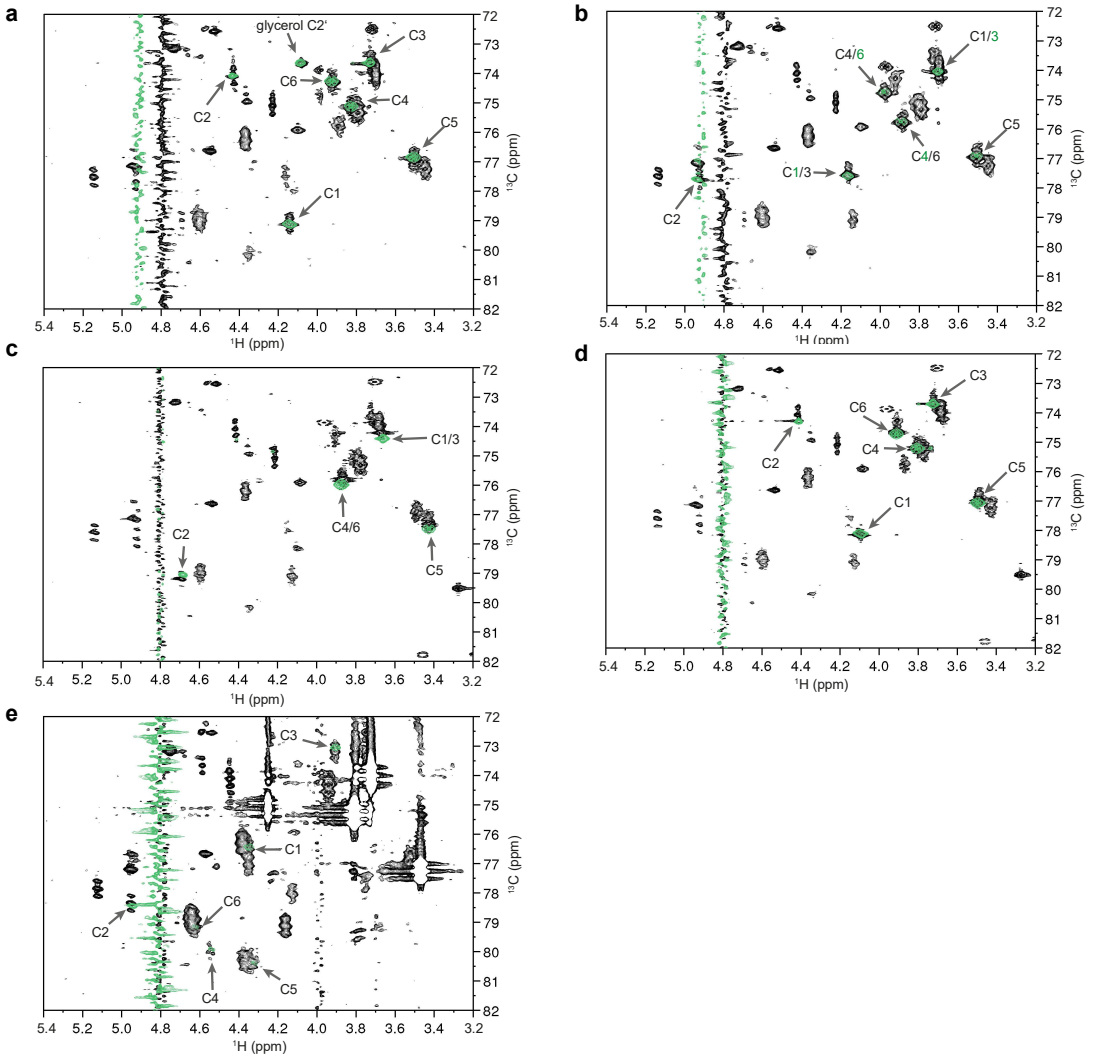


Figure S4: NMR spectra of [$^{13}C_6$]Ins-labeled H1Hela WT metabolic extracts spiked with InsP standards (black) overlayed with spectra of the same standard in saturated $KClO_4$ solution in D_2O, pH* = 6.0 (green). a: GroPI; b: Ins(1,2)P_2; c: Ins(2)P, d: Ins(1)P. e: NMR spectrum of [$^{13}C_6$]Ins-labeled HEK293 *MINPP1*$^{-/-}$ metabolic extract (black) overlayed with InsP_5[3OH] (green). The corresponding positions on the inositol ring are annotated with arrows. For Ins(1,2)P_2 the annotations for the spike-in standard are written in green while the annotation for the other enantiomer Ins(2,3)P_2, which is the species present in mammalian cells, are written in black. Note that in a and b the solvent signal is shifted between the extract and the InsP standards due to different sample temperatures during NMR measurement.

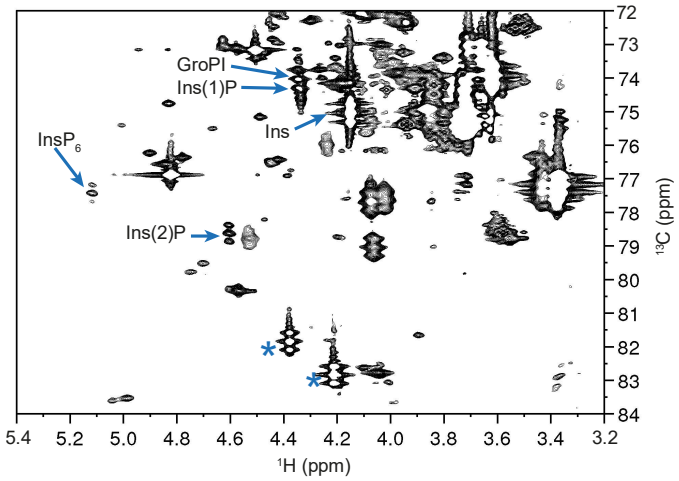


Figure S5: HMQC spectrum of metabolically [$^{13}C_6$]*myo*-inostiol-labeled *S. pombe*. The labeling protocol has already been published elsewhere.[1] While several InsPs were observed that overlap with mammalian InsP species ($InsP_6$, Ins(2)P, GroPI and Ins(1)P, annotation was limited to the 2-position for clarity), *S. pombe* extracts contain multiple high-intensity triplet signals that do not match any clusters established in Figure 2. Therefore, these *myo*-inositol-derived species likely do not represent *myo*-inositol phosphates. The exact identity of these metabolites will be addressed in future work.

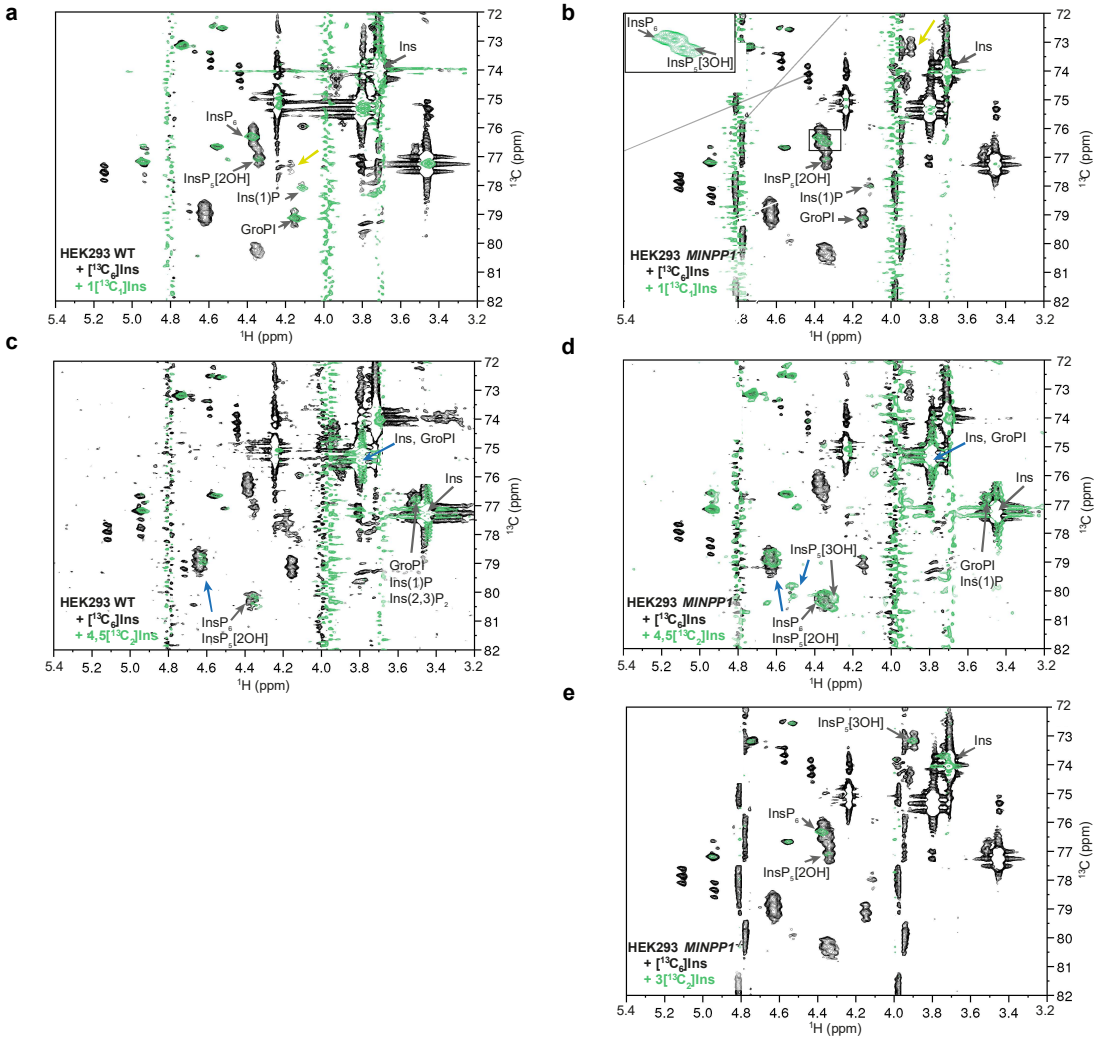


Figure S6: HMQC spectra of HEK cell lines which were metabolically labeled with asymmetrical isotopomers of *myo*-inositol (green) overlayed on the respective spectra of [$^{13}C_6$]*myo*-inositol-labeled cells (black). Annotations were limited to labeled positions. (a): 1[$^{13}C_1$]*myo*-inositol-labeled HEK293 WT cells (same data as in Fig. 3b). This labeling experiment illustrates that the signal for the phosphorylated position of Ins(1/3,2)P_2 (yellow arrow) is not labeled, thus excluding Ins(1,2)P_2 as a possible enantiomer. Note that non-labeled positions of *myo*-inositol are still visible due to the large relative abundance of *myo*-inositol in the extract. (b): 1[$^{13}C_1$]*myo*-inositol-labeled HEK293 *MINPP1*$^{-/-}$ cells. Here the signal for the 1-position of $InsP_5$[1/3OH] is clearly still phosphorylated (insert showing the magnified region), while the dephosphorylated 1/3-position (yellow arrow) is not labeled, indicating that the enantiomer present cannot be $InsP_5$[1OH], but must be $InsP_5$[3OH]. (c): 4,5[$^{13}C_2$]*myo*-inositol-labeled HEK293 WT cells. For 4,5[$^{13}C_2$]*myo*-inositol-labeled spectra the 4-positions are marked with a blue arrow and 5-positions with a black arrow. Note that the InsP signals now show a characteristic doublet pattern due to ^{13}C-^{13}C coupling. (d): 4,5[$^{13}C_2$]*myo*-inositol-labeled HEK293 WT *MINPP1*$^{-/-}$ cells. Here, the signal for the 4-position of $InsP_5$[3OH] is shifted away from the signals of the 4/6-positions of $InsP_6$/$InsP_5$[2OH], which is consistent with the observed shifts for the $InsP_5$[3OH] standard, but not $InsP_5$[1OH] (see also Figure S4e). (e):

3[$^{13}C_1$]*myo*-inositol-labeled HEK293 *MINPP1*$^{-/-}$ cells. The dephosphorylated position of $InsP_5$[1/3OH] is labeled, consistent with the enantiomer $InsP_5$[3OH].

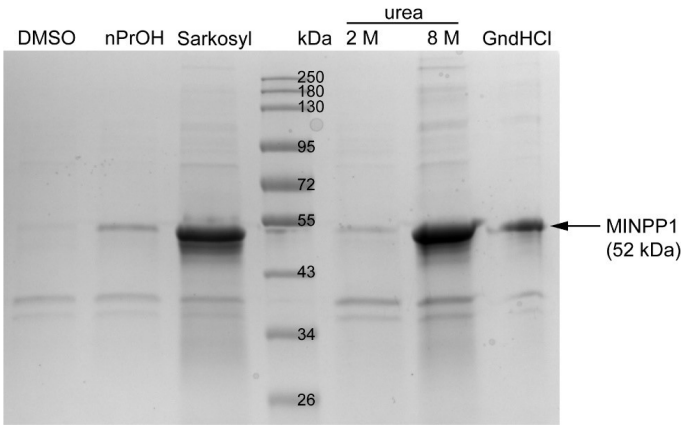


Figure S7: Testing resolubilization buffers for MINPP1 purification from inclusion bodies. MINPP1 was expressed according to the procedure described in the Experimental section. One part of the cell debris pellet obtained after lysis was washed only once with DI water, weighted, resuspended in little water and distributed into six 15 mL tubes (110 mg of wet pellet per tube). 4 mL of each resolubilization buffer (Experimental section under Cloning and production of MINPP1) were added to each tube, and incubated for 16 h at 4 °C on a reciprocal shaker. The tubes were centrifuged (30 min, 3000 g, 4 °C). 10 µL of supernatant were each diluted with 60 µL deionized water, 30 µL SDS running buffer, 40 µL Lämmli-buffer (incl. *β*-mercaptothanol) and all samples except for the guanidinium hydrochloride-based sample were boiled for 5 min at 90 °C. 30 µL of each sample were loaded on an SDS-PAGE gel, 150 V were applied until the loading marker completely ran into the gel. The wells were then flushed with SDS running buffer to remove excess guanidium hydrochloride to prevent gel distortions. Then the SDS-PAGE was continued (150 V, 45 min) and stained using colloidal Coomassie.

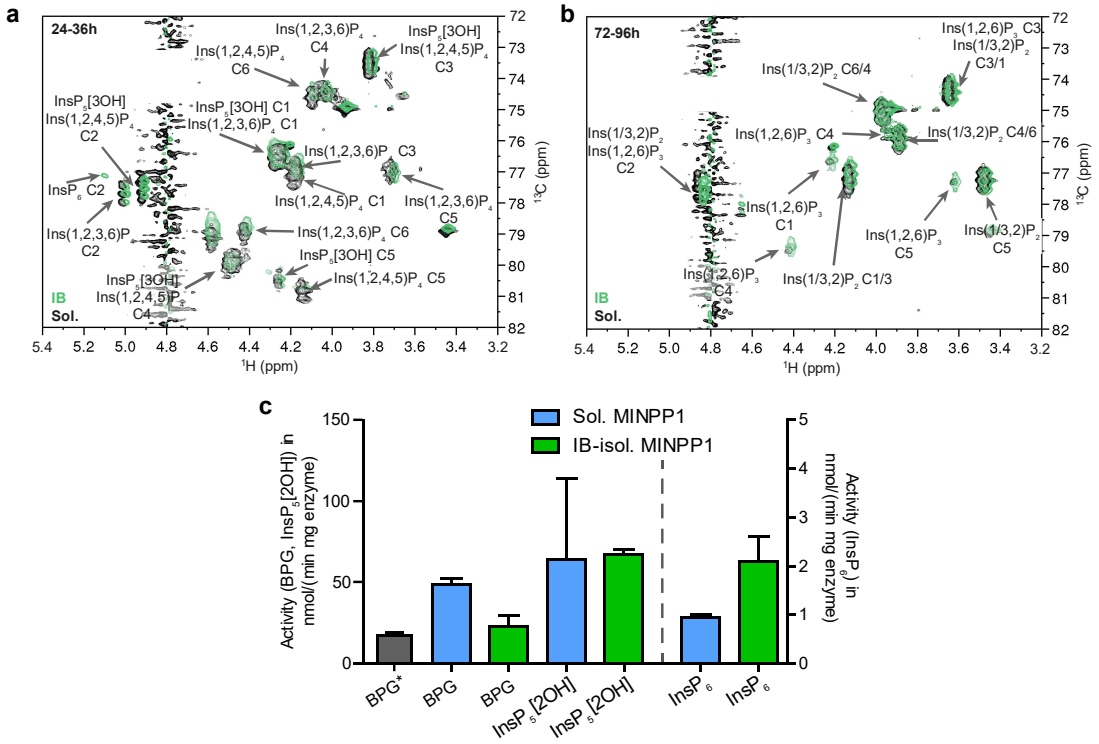


Figure S8: MINPP1 isolated from inclusion bodies (IB) exhibits similar properties to non-refolded MINPP1 obtained from the soluble fraction of *E. coli* lysate (Sol). (a) and (b): Sol. MINPP1 produces the same intermediates from [$^{13}C_6$]$InsP_6$ compared to IB MINPP1 (a: 24-36h, b: 72-96h). (c): Reaction rates of Sol. MINPP1 and IB MINPP1 against different substrates determined by Malachite green assay are similar. 2,3-bisphosphoglycerate (BPG), $InsP_5$[2OH] (left y-axis) and $InsP_6$ (right y-axis) were incubated with Sol. MINPP1 or IB MINPP1 as described in the Experimental section. Shown in grey is V_{max} determined by Cho *et al.*.[2]

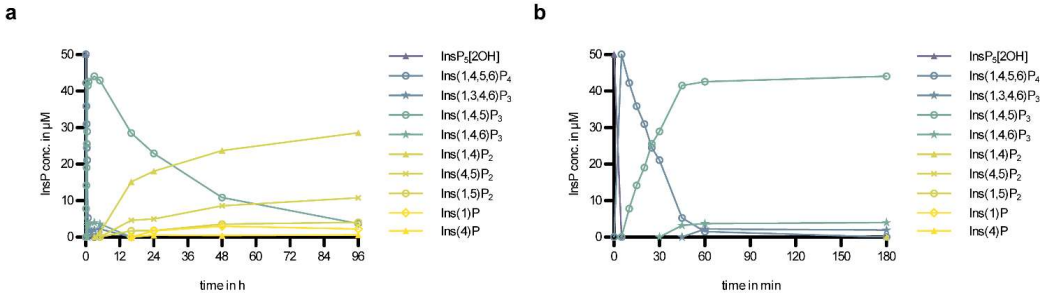


Figure S9 Progress curves of MINPP1 reaction with 50 µM $InsP_5$[2OH] (a and the first 180 min in b).

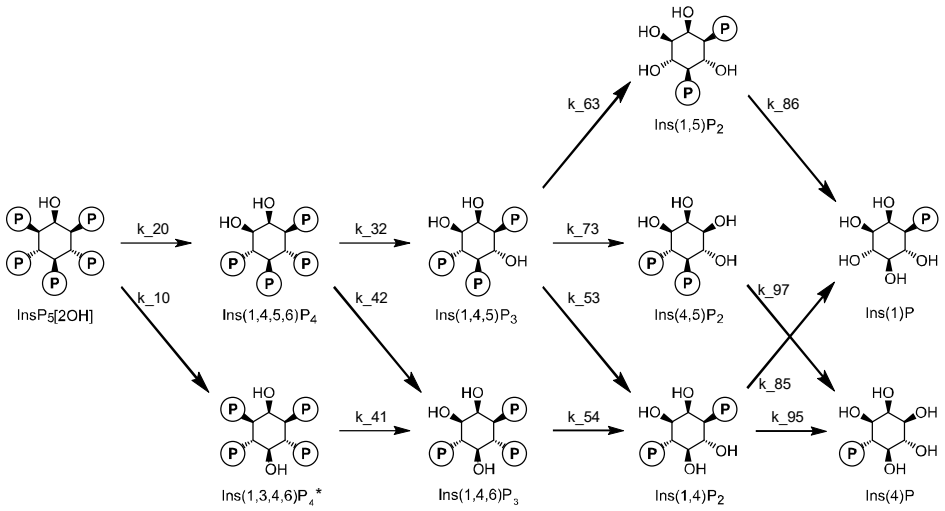


Figure S10: Complete MINPP1-mediated dephosphorylation pathway observed for $InsP_5$[2OH]

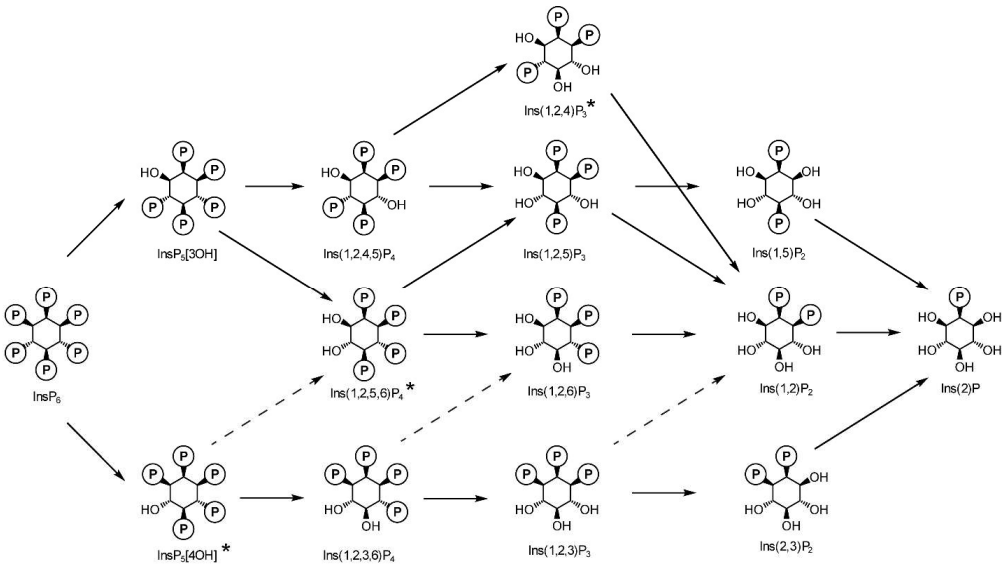


Figure S11: Complete MINPP1-mediated dephosphorylation pathway observed for $InsP_6$. The dashed arrows indicate theoretically possible paths which we assume are not relevant to the overall outcome. More investigation is needed to confirm this.

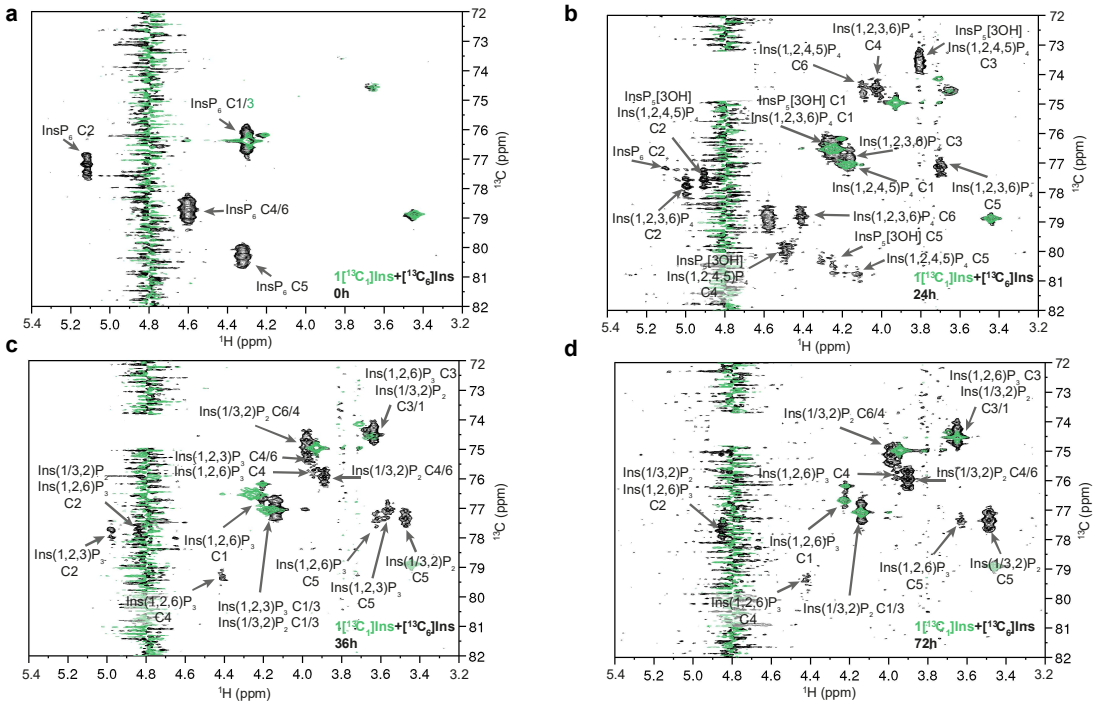


Figure S12: MINPP1 dephosphorylation of 1[$^{13}C_1$]InsP_6. The spectra show reactions in which MINPP1 was incubated with either 175 µM 1[$^{13}C_1$]InsP_6 (green) or a 1:1 mixture of [$^{13}C_6$]Ins:1[$^{13}C_1$]Ins (black). (**a**): Control sample without enzyme. InsP_6 is clearly shown to be labeled at the 1-position. The other visible signals belong to buffer components. (**b**): Reaction mixture after 24 h of incubation. The 1-position is not dephosphorylated at this stage. Thus, the shown enantiomers are enantiopure. The 3-position of Ins(1,2,3,6)P_4 is slightly shifted upfield with regards to the ^{13}C-dimension, compared to the labeled 1-position of Ins(1,2,4,5)P_4. Also, the signals at ~75 and ~79 ppm (^{13}C dimension) are buffer components from the MINPP1 stock solution. (**c**): Reaction mixture after 36 h incubation. The 1-position of Ins(1,2,6)P_3 and 3/1-position of Ins(1/3,2)P_2 overlap with the buffer component at ~75 ppm which seems to increase in intensity. The labeling of the 1-position appearing in both the region for phosphorylated and the region for dephosphorylated positions indicate that a mixture of Ins(2,3)P_2 and Ins(1,2)P_2 is formed. (**d**): Reaction mixture after 72 h of incubation. The dephosphorylated 1-position of Ins(2,3)P_2 is now evident while the 1-position of Ins(1,2)P_2 is still phosphorylated, indicating that a mix of both enantiomers has been formed, despite the enantio-specific nature of the previous dephosphorylation steps in (b). A rough integration of all labeled 1-position signals (green spectrum) resulted in a near 1:1 ratio between the dephosphorylated 1-position of Ins(2,3)P_2 and the combined phosphorylated 1-position of Ins(1,2)P_2 and Ins(1,2,6)P_3. This high ratio suggests that MINPP1 likely converts Ins(1,2,3)P_2 exclusively into Ins(2,3)P_2, while all intermediates downstream of InsP_5[3OH] must result in the other enantiomer Ins(1,2)P_2.

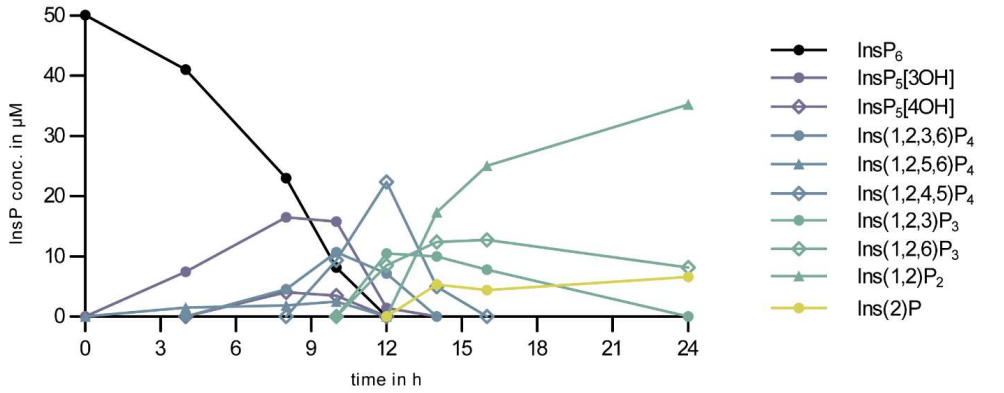


Figure S13: Due to substrate inhibition the MINPP1 progress curves for $InsP_6$ show different kinetics with respect to the dephosphorylation intermediates at lower (50 µM) initial concentrations of $InsP_6$, compare also with Figure 5c (175 µM initial concentration). Representative of 3 replicates.

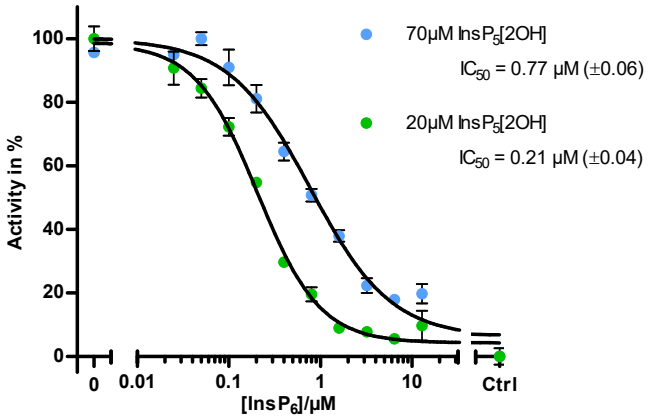


Figure S14: $InsP_6$ inhibits the MINPP1-mediated dephosphorylation of different substrate concentrations of $InsP_5$[2OH] with changing IC_{50}-values in agreement with the Cheng-Prusoff equation. Either 70 or 20 µM $InsP_5$[2OH] were incubated with 0.5 µM MINPP1 and different amounts of $InsP_6$ (two-fold dilution series ranging from 12.8 – 0.025 µM final $InsP_6$ concentration) and phosphate release was determined using a Malachite green-assay kit after 24 min reaction time (20 µM) or 1 h (70 µM). IC_{50}-values are reported with standard error of $\log_{10}IC_{50}$ in brackets. Starting from the determined IC_{50} = 1.97 (±0.02) at 175 µM substrate (see Fig 6c), the expected IC_{50}-values according to the Cheng-Prusoff equation assuming competitive inhibition are 0.77-0.82 µM for 70 µM substrate (found: 0.77 µM (±0.06)) and 0.22-0.23 µM for 20 µM substrate (found: 0.21(±0.04)). With starting concentrations far above the Michaelis-Menten constant for $InsP_5$[2OH] (40 nM), un- and non-competitive inhibition would show a substrate-concentration independent IC_{50}.[3]

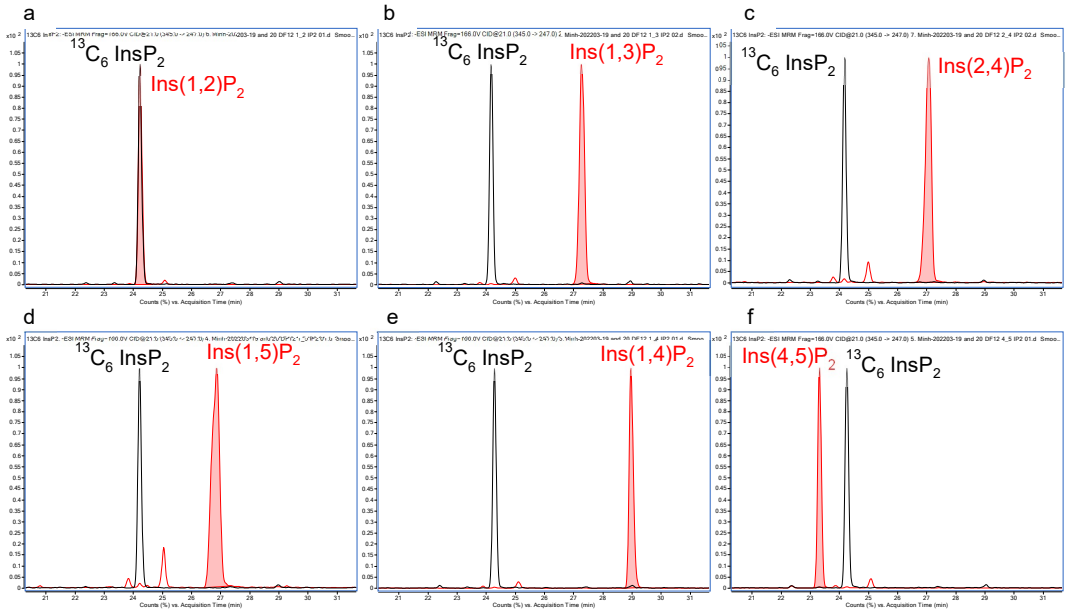


Figure S15: Confirmation of the identity of Ins(1/3,2)P_2 *via* CE-MS. The metabolic extract of [$^{13}C_6$]Ins metabolically-labeled HEK293 WT cells were spiked with commercial standards of different InsP_2 isomers and analyzed *via* CE-MS. Depicted are the extracted ion chromatograms corresponding to the masses of the intracellularly synthesized [$^{13}C_6$]InsP_2 (black) and the non-labeled InsP_2 standards. Only Ins(1,2)P_2 coelutes with the [$^{13}C_6$]InsP_2 signal in question (a) while all other tested InsP_2 standards (b: Ins(1,3)P_2, c: Ins(2,4)P_2, d: Ins(1,5)P_2, e: Ins(1,4)P_2, f: Ins(4,5)P_2) do not.

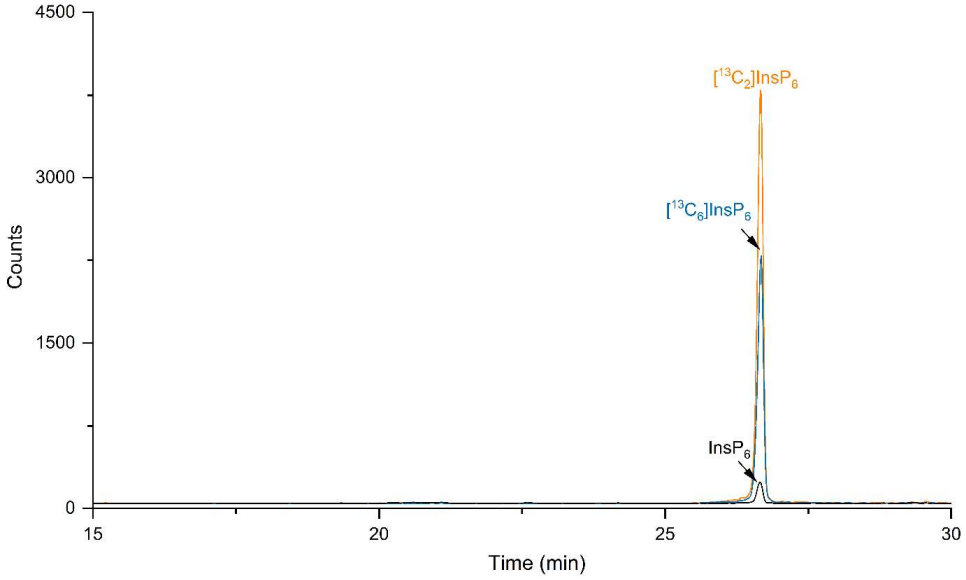


Figure S16: Example EICs (extracted ion chromatograms) of $[13C_{6/2/0}]InsP_6$ in a HEK293 WT cells which were metabolically labeled with $[^{13}C_6]$*myo*-inositol to equilibrium and then with 4,5$[^{13}C_2]$ *myo*-inositol for 48h. For the metabolic flux analysis (Figures 7b, 7c) the integrals of the respective isotopomer peaks (blue/ orange) were used for relative quantification. The InsP pools contained a constant ~3% of non-labeled InsPs (black) due to glucose-6-phosphate-dependent neogenesis of *myo*-inositol.

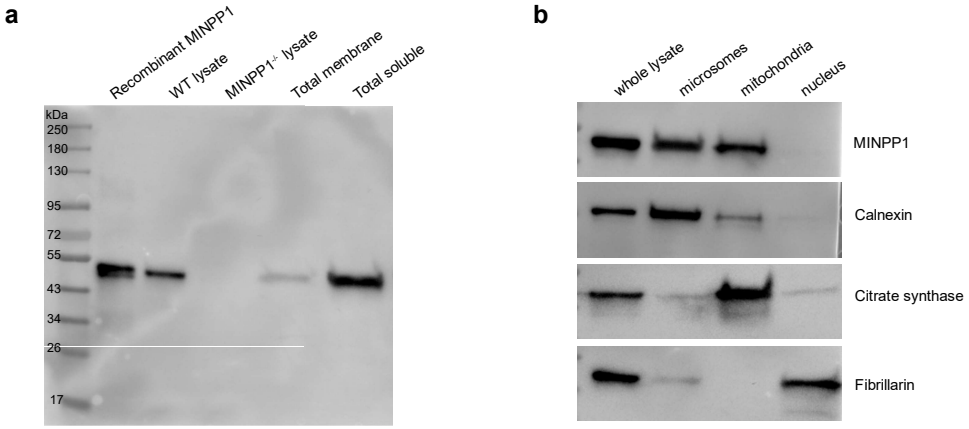


Figure S17: MINPP1 Western blots of HEK293 WT and *MINPP1*$^{-/-}$ cells and subcellular fractions. (a): MINPP1 is present in HEK293 WT cell lysates but expectedly not in MINPP1$^{-/-}$ HEK lysates and is predominantly found in the soluble fraction. (b): MINPP1 is not found in the nucleus but in ER (microsomes) and mitochondria. Calnexin was used as an ER marker, citrate synthase as a mitochondrial marker and fibrillarin as a nuclear marker.

Experimental section

Safety statement

No unexpected or unusually high safety hazards were encountered.

General Information

Chemicals were obtained from Sigma Aldrich, VWR, Roth, TCI, Thermo Scientific or Roche and used without further purification unless stated otherwise.

InsP standards were purchased as sodium, potassium, ammonium or cyclohexylammonium salts from SiChem (Ins(3,4,5,6)P_4, Ins(1,4,5,6)P_4, Ins(1,4,5)P_3, Ins(1,3,4)P_3, InsP_5[3OH], InsP_5[1OH]), Cayman chemical (Ins(2,3,5)P_3, Ins(1,2)P_2), Echelon Bioscience (Ins(1,4)P_2, Ins(1,2,6)P_3, GroPI), Biomol (Ins(1,5)P_2, Ins(1,4,6)P_2) or Sigma-Aldrich (Ins(1)P, Ins(2)P) or synthesized in-lab ([$^{13}C_6$]InsP_6, [$^{13}C_6$]InsP_5[2OH], [$^{13}C_6$]1PP-InsP_5, [$^{13}C_6$]5PP-InsP_5, [$^{13}C_6$]1,5(PP)$_2$InsP_4) as described previously.[4] Non-labeled InsPs were dissolved in a saturated $KClO_4$ solution in D_2O (pH* 6.0) to mimick the conditions of the metabolic extracts. Non-labeled standards were dissolved in the smallest volume possible for NMR measurements (min. 500 µL). All samples were adjusted to pH* 6.0 if necessary using DCl and NaOD solutions in D_2O (all deuterated solutions obtained from Eurisotop).

For NMR-based quantification purposes standards (TMPBr (Sigma, 288268) or phosphonoacetic acid (TraceCert ^{31}P-NMR standard, Supelco, 79251), respectively) were dissolved/ diluted in dry D_2O (Eurisotop D215T) and aliquots are frozen until use.

NMR data acquisition and processing

For NMR measurements and NMR data analysis TopSpin 3.5 was used. Measurements were conducted on a Bruker AV-III spectrometer (Bruker Biospin, Rheinstetten, Germany) operating at 600 MHz for ^{1}H and 151 MHz for ^{13}C nuclei equipped with a cryo-QCI probe. The pulse sequence for BIRD-{^{1}H,^{13}C}HMQC is based on the hmqcbiph pulse program from Bruker. Measurement parameters are adapted depending on sample composition. Typically, metabolic extracts were recorded with TD(^{13}C) = 1024, 140 scans, spectral width (^{13}C) limited to 40 – 100 ppm. Typically, samples from *in vitro* experiments were recorded with TD (^{13}C) = 512, 64 scans, spectral width (^{13}C) limited to 50 – 90 ppm. All samples were recorded at 310 K.

BIRD-{^{1}H,^{13}C}HMQC-NMR spectra were processed without digital water suppression with manual phasing and automatic baseline correction.

Quantification of NMR data were conducted as follows: For metabolic extracts InsPs were quantified against a known concentration of tetramethylphosphonium bromide (TMPBr). A standard curve for InsP_6 and InsP_5[2OH] against TMPBr was recorded earlier [5]. For other InsP species the standard curve for InsP_6 was used as an approximation as there are no fully ^{13}C-labeled standards available. For the samples from the *in vitro* dephosphorylation of Ins$P_{6/5}$ by MINPP1 the InsP signals were quantified relatively to each other and normalized to a total InsP concentration matching the initial substrate concentration. As the signals of the 2-positions are the sharpest and best resolved (due to the reduced coupling to the neighbouring CH groups), the 2-position signals were used for quantification. In the cases where the 2-position signals of two InsPs species are not baseline-separated, the signals were integrated together and split by the ratio of the 5-position signal integrals.

CE-MS measurement

CE-ESI-MS has been found to be an efficient platform for the analysis of inositol polyphosphate.[6] A CE-ESI-QQQ setup is used for this study, which consists of an Agilent 7100 CE, a triple quadrupole tandem mass spectrometry Agilent 6495c, connected to an Agilent Jet Stream (AJS) electrospray ionization (ESI) source. A commercial CE-MS sheath liquid coaxial interface was used, with an isocratic LC pump constantly delivering the sheath-liquid (*via* a splitter set with a ratio of 1:100). All experiments were performed on a bare fused silica capillary with a length of 100 cm (50 µm internal diameter and 365 µm outer diameter). 35 mM ammonium acetate titrated by ammonia solution to pH 9.7 was employed as

background eletrolyte (BGE). Samples were injected by applying 100 mbar pressure for 15 s, corresponding to 1.5% of the total capillary volume (30 nL).

The sheath liquid is a mixture of water-isopropanol (1/1, v/v) and with a constant flow of 10 µL/min. The MS source parameters settings were as follows: nebulizer pressure was set to 8 psi, gas temperature was 150 °C with a flow of 11 L/min, sheath gas temperature was 175 °C and with a flow of 8 L/min, capillary voltage was -2000 V with nozzle voltage 2000 V. Negative high-pressure RF and low-pressure RF (Ion Funnel parameters) were 70 V and 40 V, respectively. Mass spectrometer parameters for MRM transitions are shown below.

Compound Name	Precursor Ion	Product Ion	dwell	Frag (V)	CE (V)	Cell Acc (V)	Polarity
$[^{13}C_6]InsP_6$	331.9	486.9	60	166	13	4	Negative
$[^{13}C_2]InsP_6$	329.9	482.9	60	166	13	4	Negative
$[^{12}C_6]InsP_6$	328.9	480.9	60	166	13	4	Negative
$[^{13}C_6]InsP_5$	292	504.9	60	166	9	3	Negative
$[^{13}C_2]InsP_5$	290	500.9	60	166	9	3	Negative
$[^{12}C_6]InsP_5$	289	498.9	60	166	9	3	Negative
$[^{13}C_6]InsP_4$	252	424.9	60	166	5	1	Negative
$[^{13}C_2]InsP_4$	250	420.9	60	166	5	1	Negative
$[^{12}C_6]InsP_4$	249	418.9	60	166	5	1	Negative
$[^{13}C_6]InsP_2$	345	247	60	166	21	4	Negative
$[^{13}C_2]InsP_2$	341	243	60	166	21	4	Negative
$[^{12}C_6]InsP_2$	339	241	60	166	21	4	Negative

Data handling

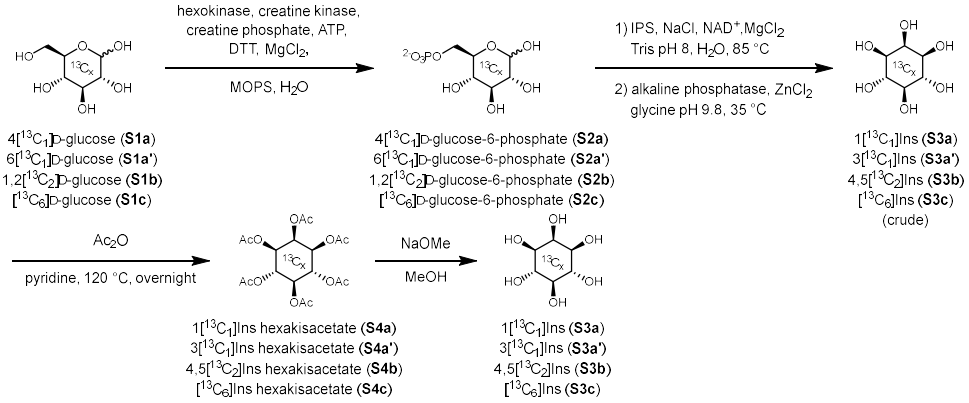
For plotting and other analyses Microsoft Excel, OriginPro 2016 and GraphPad Prism 5 were used. Bagplots were created in R (version 4.1.2) with the aplpack package (version 1.3.5).

For details of the kinetic modelling of MINPP1 see separate SI file “Supporting Information: Numerical Analysis”.

Synthesis of ^{13}C-labeled Ins and InsPs

The synthesis of ^{13}C-labeled *myo*-inositol and its derivatization to InsPs were carried out based on published procedures for $[^{13}C_6]$Ins with slight improvements of the protocol as described below.[5]

Chemoenzymatic synthesis of *myo*-inositol isotopomers



Ins isotopomers were synthesized chemoenzymatically from the respective D-glucose isotopomer: 1[$^{13}C_1$]Ins (**S3a**) was synthesized from 4[$^{13}C_1$]D-glucose (**S1a**), 3[$^{13}C_1$]Ins (**S3a'**) was synthesized from 6[$^{13}C_1$]D-glucose (**S1a'**), 4,5[$^{13}C_2$]Ins (**S3b**) was synthesized by starting from 1,2[$^{13}C_2$]D-glucose (**S1b**), and [$^{13}C_6$]Ins (**S3c**) from [$^{13}C_6$]D-glucose (**S1c**). ^{13}C-labeled material was obtained from Eurisotop/ Cambridge Isotope Labs. Generally, we observed improved yields with higher synthesis scale with 1 to 3 g glucose as starting material yielding up to 55% Ins. However, the asymmetric isotopomer **S3a** was synthesized only on a 500 mg scale.

Briefly, **S1a/a'/b/c** is first converted enzymatically to the respective D-glucose-6-phosphate (**S2a/a'/b/c**) with hexokinase and crudely purified *via* an anion exchange hand column. The subsequent lyophilization step of the eluate in the original procedure can be replaced by concentrating using a rotavap without reduction of yield while saving time. The resulting product/salt mixture is then converted to inositol-3-monophosphate (Ins(3)P) through the action of inositol monophosphate synthase (IPS), which is monitored *via* NMR. We recommend preparing recombinantly expressed IPS as closely to the protocol in [5] as possible to ensure sufficient activity of the IPS (esp. induction at high OD_{600} and purification *via* heat-treatment); prolonged reaction times causes the NAD^+ cofactor to degrade, inhibiting IPS activity even after resupplementing more IPS and NAD^+. Subsequently, Ins(3)P is dephosphorylated to Ins (**S3a/a'/b/c**) by alkaline phosphatase. The reaction progress is also monitored *via* NMR. The ion exchange treatment in the original procedure can be skipped upon complete conversion and the aqueous solution can be reduced on a rotavap instead, yielding a crude brown solid. The Ins is then purified through chemical derivatization by acetylation to *myo*-inositol hexakisacetate (**S4a/a'/b/c**), purification *via* extraction and column chromatography on silica gel (~500 mL silica gel for a 3 g synthesis scale), followed by deacetylation and precipitation in acetonitrile (the precipitation is repeated twice if necessary) to afford the desired *myo*-inositol isotopomer **S3a/a'/b/c** in pure form following the published protocol.

1[$^{13}C_1$]Ins (S3a): yield: 122 mg (starting from 500 mg **S1a**, 24%)

^{1}H NMR (600 MHz, D_2O) δ[ppm]: 3.99 (s, 1H, 2-position), 3.56 (ps-q, J = 9.8 Hz, 2.5H, 4/6-position and 1-position), 3.46 (d, J = 9.9 Hz, 1H, 3-position), 3.34 (d, J = 11.4 Hz, 0.5H, 1-position), 3.21 (t, J = 9.5 Hz, 1H, 5-position). Please note that the 1-position is coupling with ^{13}C with a coupling constant of $^1J_{CH}$ = 143.4 Hz.

^{13}C NMR (151 MHz, D_2O) δ[ppm]: 77.09 (d, *J* = 6.7 Hz, 5-position), 75.17 (d, *J* = 33.4 Hz, 6-position), 75.14 (s, 4-position), 74.91 (d, *J* = 32.4 Hz, 2-position), 73.88 (large s, satellite d, J = 39.0 Hz, 1- and 3-position).

HRMS *m/z*: $[M - H]^-$ calcd. for $^{13}C_1{}^{12}C_5H_{11}O_6$ 180.0595; found 180.0593.

3[^{13}C$_1$]Ins (S3a'): yield: 563 mg (starting from 1000 mg **S1a**, 56%)

^{1}H NMR (600 MHz, D_2O) δ[ppm]: 4.09 (dt, J = 5.3, 2.9 Hz, 1H, 2-position), 3.66 (m, 2.5H, 4/6-position and 3-position), 3.56 (dd, J = 10.1, 3.0 Hz, 1H, 1-position), 3.44 (dd, J = 9.9, 2.9 Hz, 0.5H, 3-position), 3.31 (t, J = 8.9 Hz, 1H, 5-position). Please note that the 3-position is coupling with ^{13}C with a coupling constant of $^1J_{CH} \approx 140$ Hz.

^{13}C NMR (151 MHz, D_2O) δ[ppm]: 77.12 (d, *J* = 6.8 Hz, 5-position), 75.20 (d, *J* = 33.9 Hz, 4-position), 75.17 (s, 6-position), 74.94 (d, *J* = 32.8 Hz, 2-position), 73.9 (large s, satellite d, J = 39.1 Hz, 3- and 1-position).

HRMS *m/z*: $[M - H]^-$ calcd. for $^{13}C_1{}^{12}C_5H_{11}O_6$ 180.0595; found 180.0593.

4,5[^{13}C$_2$]Ins (S3b): yield: 450 mg (starting from 1 g **S1b**, 45 %)

^{1}H NMR (600 MHz, D_2O) δ[ppm]: 4.19 (t, *J* = 3 Hz, 1H, 2-position), 3.75 (tdd, J = 144.3, 9.9, 4.2 Hz, 1, 4-position) 3.75 (td, *J* = 9.7, 4.7 Hz, 1H, 6-postition), 3.66 (d, *J* = 9.8 Hz, 2H, 1/3-position), 3.40 (tdd, *J* = 140.7, 9.3, 4.1 Hz, 1H, 5-position).

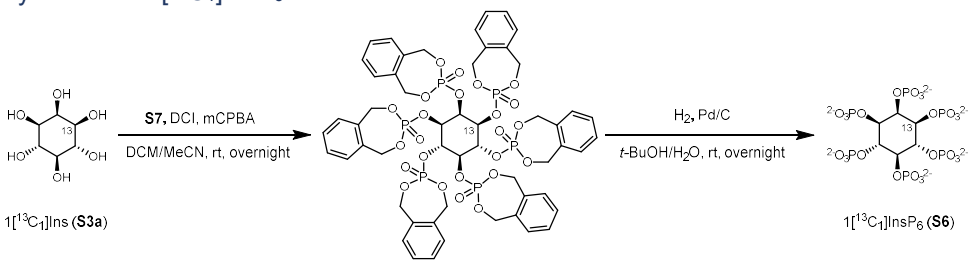
^{13}C NMR (151 MHz, D_2O) δ[ppm]: 77.23 (d, *J* = 38.8 Hz, 5-position), 75.28 (d, *J* = 38.9 Hz, 4+6position), 75.05 (2-position), 74.02 (d, *J* = 6.8 Hz, 1-position), 74.00 (dd, *J* = 39.5, 7.1 Hz, 3-position).

HRMS *m/z*: $[M - H]^-$ calcd. for $^{13}C_2{}^{12}C_4H_{11}O_6$ 181.0628; found 181.0627.

[^{13}C$_6$]Ins (S3c): yield: up to 1.55 g (starting from 3 g **S1c**, 52%)

Analytical data for [$^{13}C_6$]Ins were published previously. [5]

Synthesis of 1[^{13}C$_1$]InsP$_6$



Synthesis of 1[$^{13}C_1$]InsP$_6$ (**S6**) was carried out following published procedures [5] with slight modifications:

1[$^{13}C_1$]Ins (30 mg, 0.17 mmol) is resuspended together with commercial *o-xylylene N,N-diethylphosphoramidite (**S7**)* (Sigma Aldrich, 360 mg, 1.5 mmol) and a stirring bar in anhydrous acetonitrile under nitrogen atmosphere. To reduce water content further, the suspension is reduced and evaporated under high vacuum for an hour. The dried mixture is then resuspended in 5.5 mL of 1:1 anhydrous dichloromethane:acetonitrile and sonicated briefly. The mixture is cooled to 0 °C using an acetone bath to which dry ice was added in a controlled manner. 4,5-DCI (254 mg, 2.15 mmol) was added and the gas phase was exchanged three times against nitrogen. The reaction is allowed to warm to room temperature and stirring is continued overnight under nitrogen/argon atmosphere. The subsequent workup is identical as described preciously [5] yielding **S5** in 68% yield (144 mg, 0.116 mmol) with slight impurities.

S5:

^{1}H NMR (600 MHz, $CDCl_3$) δ[ppm]: 7.40 – 7.31 (m, 20H), 7.27 – 7.25 (m, 4H, overlaps with solvent signal signal), 5.75 (dd, *J* = 13.8, 9.3 Hz, 2H), 5.65 (dt, *J* = 13.0, 8.0 Hz, 3H), 5.59 – 5.51 (m, 6H), 5.39 (dd, *J* = 13.8, 12.3 Hz, 2H), 5.30 – 4.93 (m, 18H).

^{13}C NMR (151 MHz, $CDCl_3$) δ[ppm:] 138.54, 138.51, 138.35, 138.16, 137.29, 132.34, 132.21, 132.08, 132.07, 132.04, 132.00, 131.83, 131.75, 80.01, 79.80, 79.59, 76.59, 76.56, 76.54, 72.46, 72.40, 72.29, 72.24, 72.19, 72.14, 72.11, 72.06.

^{31}P NMR (243 MHz, $CDCl_3$) δ[ppm]: -2.81 (d, *J* = 2.9 Hz, 1P), -3.37 (d, *J* = 3.0 Hz, 2P), -4.37 (s, 1P), -4.53 (s, 1P).

HRMS *m/z*: $[M + H]^+$ calcd. for $^{13}C_1{}^{12}C_{53}H_{55}O_{24}P_6$ 1274.1537; found 1274.1526.

144 mg (**S5**, 0.116 mmol, 1 eq.) was dissolved in 28 mL *t*-BuOH and Milli-Q® water 6:1, and 250 mg of palladium black (10% Pd/C) was added. The suspension was stirred overnight under hydrogen atmosphere. Upon depletion of starting material (according to LC-MS analysis) 2 ml Milli-Q® water was added to adjust the solvent to a ratio of 4:1 *t*-BuOH:Milli-Q® and stirring under hydrogen atmosphere was continued overnight. The catalyst was removed by centrifuging the suspension in 50 ml centrifugal tubes at 3000 g for 15 min and the supernatant was passed through a PTFE syringe filter (0.45 µm). The catalyst pellet is washed once with 5 ml Milli-Q® water, centrifuged and the supernatant is again filtered and the filtrates are united and tBuOH is removed on the rotavap before the aqueous solution is lyophilized. The resulting white solid is redissolved in 200 ml water and magnesium chloride solution is added to a final concentration of 26 mM (49 eq.). The solution is adjusted with sodium hydroxide solution to a final pH of 9.0 – 9.2 which initiates precipitation of **S6** as a Mg^{2+}-complex. The mixture was incubated at 4 °C overnight. The precipitate is pelleted by centrifugation (3000 g, 15 min) in a 50 ml tube and washed twice with 20 ml of 8 mM $MgCl_2$ solution at pH 9.0. The resulting pellet is resuspended in 10 ml of water resulting in a milky solution without any clumps. Meanwhile 15 ml bed volume of Amberlite® IRC-748 (chelating) ion exchange resin (Alfa Aesar, L19570), which was washed in advance extensively with deionized water and methanol and stored in methanol until use) is loaded into a 20 ml peptide reactor column (or another small column) and equilibrated by passing through 100 ml of water. 8 ml bed volume of this Amberlite are added to the $InsP_6$ suspension and incubated at rt on a shaking platform for 30 min until the supernatant turns clear. The content of the tube was transferred onto the remaining Amberlite column and the eluate (gravity-flow) was collected. Additional 20 ml of water pushed through the Amberlite column and the eluates are combined and lyophilized. The resulting clean material was redissolved in D_2O for analysis, filtered through a 0.2 µm PTFE syringe filter and pH was adjusted by addition of DCl solution to 7.0 and dilution to a defined volume. The concentration of 1[$^{13}C_1$]$InsP_6$ was determined against a quantitative NMR-standard (phosphonoacetic acid). In total 0.105 mmol (91%) of clean 1[$^{13}C_6$]$InsP_6$ were obtained.

1[$^{13}C_6$]$InsP_6$ (S6):

^{1}H NMR (600 MHz, Deuterium Oxide) δ 4.90 (dq, *J* = 7.7, 2.5 Hz, 1H, 2-position), 4.39 (qt, *J* = 9.5, 2.9 Hz, 2H, 4/6-position), 4.11 (q, *J* = 9.6 Hz, 1H, 3-position), 4.09 (t, *J* = 9.5 Hz, 1H, 5-position), 4.09 (dt, *J* = 144.0, 9.4 Hz, 1H, 1-position).

^{13}C NMR (151 MHz, D_2O) δ[ppm]: 80.16, 78.69, 76.91, 76.24 (1-position).

^{31}P NMR (243 MHz, D_2O) δ 1.94, 1.08, 0.73.

HRMS *m/z*: $[M - 2H]^{2-}$ calcd. for $^{13}C_1{}^{12}C_5H_{16}O_{24}P_6$ 329.4251; found 329.4242.

Cloning, expression and purification of recombinant human MINPP1

A gene sequence encoding for human MINPP1 (29-487, Uniprot Q9UNW1-1) lacking the N-terminal signal peptide was designed and ordered using Thermo Fisher's GeneArt service. The sequence was codon-optimized for expression in *E. coli* and contains a NdeI (at initial ATG) and XhoI (after the stop codon) restriction site. The MINPP1 gene was cloned into the vector pET-15b using the NdeI and XhoI restriction sites. The resulting plasmid (pET-15b-MINPP1) encodes an N-terminal His-tag with a thrombin cleavage site followed by MINPP1. For plasmid preparation the *E. coli* Top10 strain was used.

The complete nucleotide sequence of the ORF of pET-15b-MINPP1 is as follows:

CCATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCATAT
GCGTTGTAGCCTGCTGGAACCGCGTGATCCGGTTGCAAGCAGCCTGAGTCCGTATTTTGGTACAA
AAACCCGTTATGAAGATGTGAATCCGGTTCTGCTGAGCGGTCCGGAAGCACCGTGGCGTGATCCT
GAACTGCTGGAAGGCACCTGTACACCGGTTCAGCTGGTTGCACTGATTCGTCATGGCACCCGTTA
TCCGACCGTTAAACAAATTCGTAAACTGCGTCAGCTGCATGGTCTGCTGCAGGCACGTGGTAGCC
GTGATGGTGGTGCCAGCAGCACCGGTAGTCGTGATCTGGGTGCAGCACTGGCAGATTGGCCTCT
GTGGTATGCAGATTGGATGGATGGTCAGCTGGTAGAAAAAGGTCGTCAGGATATGCGTCAACTG
GCACTGCGTCTGGCAAGCCTGTTTCCGGCACTGTTTAGCCGTGAAAATTATGGTCGTCTGCGTCT
GATTACCAGCAGCAAACATCGTTGTATGGATAGCAGCGCAGCATTTCTGCAAGGTCTGTGGCAGC
ATTATCATCCGGGTCTGCCTCCGCCTGATGTTGCAGATATGGAATTTGGTCCGCCTACCGTTAATG
ATAAACTGATGCGTTTTTTTGACCATTGCGAGAAGTTTCTGACCGAGGTTGAAAAAAATGCAACCG
CACTGTATCATGTGGAAGCATTTAAAACAGGTCCGGAAATGCAGAACATCCTGAAAAAAGTTGCA
GCAACCCTGCAGGTTCCGGTTAATGATCTGAATGCCGATCTGATTCAGGTTGCCTTTTTTACCTGT
TCATTTGACCTGGCCATTAAAGGTGTTAAAAGCCCGTGGTGTGATGTGTTTGATATTGATGATGCA
AAGGTGCTGGAATATCTGAACGATCTGAAACAGTATTGGAAACGCGGTTATGGCTATACCATTAA
TAGCCGTAGCAGCTGTACCCTGTTTCAGGATATTTTTCAGCATCTGGATAAAGCCGTTGAACAGAA
ACAGCGTAGCCAGCCGATTAGCAGTCCGGTTATTCTGCAGTTTGGTCATGCGGAAACCCTGCTGC
CGCTGCTGAGCCTGATGGGTTATTTCAAAGATAAAGAACCGCTGACCGCCTACAACTATAAAAAG
CAGATGCATCGTAAATTTCGCAGCGGTCTGATTGTTCCGTATGCAAGCAATCTGATTTTTGTGCTG
TATCATTGCGAAAATGCGAAAACCCCGAAAGAACAGTTTCGTGTTCAGATGCTGCTGAATGAAAA
AGTTCTGCCGCTGGCATATAGCCAAGAAACCGTTAGCTTTTATGAGGACCTGAAAAACCACTACA
AAGATATCCTGCAGAGCTGTCAGACCAGCGAAGAATGTGAACTGGCACGTGCAAATAGCACCAG
TGATGAACTGTAACTCGAGGATCC

Complete ORF of pET-15b-MINPP1. Restriction sites are highlighted (NcoI in yellow, NdeI in green, XhoI in light blue). The sequence encoding MINPP1 is shown in bold and the font colour for chosen component of the protein are changed (His-tag in blue, thrombin cleavage site in orange, catalytic histidine in red).

For protein expression *E. coli* BL21 (DE3) was used which was transformed with the MINPP1-encoding plasmid using the heat-shock method. A 5 ml-overnight culture of the transformed bacterial strain in terrific broth (TB, Formedium) and Ampicillin (100 µg/mL, Roth) at 37 °C was inoculated into 500 ml of TB and Ampicillin. The culture was cooled to 18 °C when OD_{600nm} = 0.5 was reached (~160 min after inoculation). Protein expression was induced at OD_{600nm} = 0.6 (~170 min after inoculation) with 0.6 mM Isopropyl β-D-1-thiogalactopyranoside (IPTG, Thermo Scientific). The culture was incubated at 18 °C for 18-20 h. The bacterial suspension was centrifuged (3000 g, 15 min, 4 °C) upon which a bacterial pellet of ~1.5 g wet

weight was obtained. The pellet was resuspended in 50 ml ice-cold lysis buffer (150 mM NaCl, 10 mM Tris*HCl (Roth), pH 8.0, 1 mM DTT (Roth or VWR), 1X cOmplete™ protease inhibitor cocktail (Roche)) and a spatula tip of lysozyme (Roth) and DNAse I (Roche) were added. The bacterial cells were lysed using a homogenizer (LM10 Microfluidizer, Microfluidics, 15000 psi, 5 passages). The resulting suspension was centrifuged (20 000 g, 20 min, 4 °C). The supernatant was used for purification of soluble MINPP1 while the resulting pellet was used for MINPP1 isolation from inclusion bodies (see below). Inclusion body-purification of MINPP1 yielded higher amounts.

The purification of soluble MINPP1 was adapted from Craxton *et al.* [7]: The supernatant was filtered (VWR vacuum filter, PES 0.45 µm) and the flowthrough was applied to a 5 ml Ni-NTA column (GE, HiTrap IMAC FastFlow) on a FPLC system (NGC Quest 10 Chromatography System, Bio-Rad) equilibrated to buffer A (150 mM NaCl, 10 mM Tris*HCl, pH 8.0, 1 mM DTT). The column was subsequently washed with 5 column volumes (CV) buffer A, 5 CV buffer A:B 10:7, 5 CV buffer B (1 M NaCl, 10 mM Tris*HCl, pH 8.0, 1 mM DTT), 5 CV buffer A with 2% buffer C (buffer C is identical to buffer A containing additional 500 mM imidazole (AppliChem), pH 8.0). For elution a gradient of 2% buffer C in A to 75% buffer C in A over 20 CV was applied. Fractions containing MINPP1 were united, concentrated using centrifugal filters (15 ml 10 kDa MWCO, Amicon Ultra) and dialyzed against 1 L of dialysis buffer 1 (150 mM NaCl, 10 mM Tris*HCl, pH 8.0, 1 mM DTT, 10 Vol-% glycerol (Roth), 0.25% CHAPS (Roth)) twice for 1.5 h. Protein concentration was determined using a BCA assay kit (Pierce™ BCA Protein Assay Kit). Protein solution was aliquoted and stored at -80 °C. However, soluble MINPP1 was obtained in only low amounts this way (2 mg from 1 L culture) which is a known problem with heterologous expression of MINPP1.[8]

For purification of MINPP1 from inclusion bodies: After removing the lysate, the pellet was washed by thoroughly resuspending in 35 ml ice-cold deionized water, centrifugation (20 000 g, 20 min, 4 °C) and after discarding the supernatant the resulting pellet was washed in the same manner two more times after which a pellet of 1.4 g wet weight was obtained. Per 0.7 g pellet mass, the pellet was resuspended in 30 ml resolubilization buffer (0.2 w/v-% *N*-lauroylsarcosin sodium salt (Sarkosyl, Fisher Scientific), 10 mM Tris*HCl, pH 8.0, 1 mM DTT). The suspension was incubated overnight at 4 °C in a 50 ml-tube under light agitation on a reciprocal shaker. The tube was centrifuged (3000 g, 30 min, 4 °C). 20 ml of recovered supernatant containing MINPP1 was dialyzed first against 1 L dialysis buffer 2 (150 mM NaCl, 10 mM Tris*HCl, pH 8.0, 1 mM DTT, 10 Vol-% glycerol, 0.1 % Triton-X100 (Roth)) for 3 h at 4 °C and then again against fresh dialysis buffer overnight at 4 °C. Protein concentration was determined using a BCA assay kit (Pierce™ BCA Protein Assay Kit). The dialyzed protein solution was adjusted to a final glycerol content of 30 Vol-%, aliquoted, flash-frozen in liquid nitrogen and stored at -80 °C. Using this inclusion body purification procedure 73 mg MINPP1 could be obtained from half of a 500 mL culture.

The activity of MINPP1 preparations were validated against its substrates 2,3-bisphophoglycerate (BPG), $InsP_5$[2OH] and $InsP_6$ using a Malachite green assay. The activity of inclusion body-purified MINPP1 against BPG was determined to be 22 nmol min^{-1} mg^{-1} enzyme, which is comparable to the value 16 nmol min^{-1} mg^{-1} enzyme reported in the literature.[2] The activity of soluble MINPP1 and inclusion-body purified MINPP1 did not differ drastically (see also Figure S8).

No decrease in activity was observed after over a year of storage (without freeze-thaw cycles).

Alternative solubilization buffers

Different solubilization buffers were also tested for the inclusion body purification of MINPP1 on a smaller scale. Several mild solubilization buffers [9] were unable to sufficiently resolubilize MINPP1 (40 mM TrisHCl, pH 8, with either 5 Vol-% DMSO or 5 Vol-% *n*-propanol (VWR); 90 mM TrisHCl, pH 8.6, 2 M urea). Among the resolubilization buffers only the following managed to solubilize MINPP1: 40 mM TrisHCl, pH 8 with a) 0.2% sarkosyl, b) 8 M urea or c) 6 M guanidinium chloride (see Figure S7).

The resolubilized MINPP1 solution from a) and b) were dialyzed against dialysis buffer with and without Triton-X. In general, it was observed that protein concentration with urea was higher than with Sarkosyl

(~1 mg/mL vs. ~3 mg/mL) and with Triton-X-containing dialysis buffer the protein yield was also slightly higher by ~0.2mg/mL.

For all *in vitro* experiments MINPP1 preparations were used based on resolubilization with sarkosyl and Triton-X-containing dialysis buffer (see above).

NMR-based enzymatic assays

For the *in vitro* dephosphorylation of $InsP_6$ and $InsP_5$[2OH] by **MINPP1** the following conditions were used unless stated otherwise:

The reaction buffer contained 100 mM NaCl, 100 mM Na_2SO_4, 25 mM HEPES, pH* = 7.4, 1 mM DTT, 1 mM EDTA (Sigma), 0.2 mg/mL BSA (Roth), 2 mM CHAPS, 175 µM (or 50 µM) of inositol phosphate substrate, 0.5 µM enzyme. The reactions were carried out in D_2O. For each sample (500 µL final volume), the reaction mixture was prepared without InsP substrate in a 1.5 ml microcentrifuge tube, prewarmed to 37 °C for 5 min before the reaction was started by adding the substrate. The reactions were quenched by boiling at 95 °C for 5 min. NMR spectra were recorded without further workup. For the substrate inhibition experiments, $InsP_5$[2OH] was mixed with aliquots of a dilution series of $InsP_6$ prior to addition to the reaction mixture.

Malachite green-based enzymatic assays

For comparing the enzymatic activities of MINPP1 preparations the dephosphorylation of BPG, $InsP_5$[2OH] and $InsP_6$ were compared. The reaction buffer contained 100 mM NaCl, 100 mM Na_2SO_4, 25 mM HEPES, pH* = 7.4, 1 mM DTT, 1 mM EDTA, 0.2 mg/mL BSA, 2 mM CHAPS, 0.5 µM enzyme. 5 mM 2,3-BPG, or 50 µM InsP, respectively, were used and the reaction was carried out in Milli-Q® water (50 µL total volume per sample) at 37 °C in 0.2 mL tubes. The reaction mixtures lacking MINPP1 were preincubated at 37 °C for 5 min and reactions were started by addition of MINPP1. After 15-25 min (BPG and $InsP_5$[2OH]) and 19 h ($InsP_6$), 20 µL of the reaction mixture were transferred into a clear, flat-bottom 96-well plate, each well already containing 20 µL of Malachite green assay solution (Sigma) and 40 µL Milli-Q® water. After incubation for 30 min at room temperature, absorption was measured at 620 nm on a TECANInfinite 200 Pro M-Plex Plate Reader. For calculating the amount of released phosphate, a dilution series of inorganic phosphate standard was measured in parallel with the same buffer background and no-enzyme controls (MINPP1 preparations did not show any phosphate background). All samples were prepared in triplicate.

For determining IC_{50} values for the inhibition of MINPP1-mediated dephosphorylation of $InsP_5$[2OH] by $InsP_6$. Same buffer used as described above but 0.1 µM enzyme was used. $InsP_5$[2OH] concentrations were either 70 µM or 20 µM and $InsP_6$ concentrations ranged from 12.8 µM to 0.025 µM in a two-fold serial dilution and 0 µM $InsP_6$. Phosphate release was measured as stated above but 40 µL reaction mixture were used. The samples were quenched after 1 h (70 µM substrate) or 25 min (20 µM substrate). For calculating the amount of released phosphate, a dilution series of inorganic phosphate standard was measured in parallel with the same buffer background and no-enzyme control). All samples were prepared in triplicate.

Mammalian cell culture and metabolic labeling

HT29 WT cells were a kind gift from the lab of Jan Carette[10] (William Kaiser laboratory, RRID: CVCL_0320). HEK293 cell lines (WT and *MINPP1*$^{-/-}$) were a kind gift of the labs of Adolfo Saiardi and Vincent Cantagrel and the generation of the *MINPP1*$^{-/-}$ cells was described previously.[11] The absence of MINPP1 in *MINPP1*$^{-/-}$ cells was verified by Western blots (see Figure S17). HCT116 were obtained from ATCC. H1975 cells were a kind gift of the Klingmüller lab (originally ATCC, CRL-5908).

Unless stated otherwise all cell lines are cultivated in DMEM (Gibco DMEM high glucose, no glutamine, product no. 11960044) supplemented with streptomycin/penicillin (100 U/mL final concentration, Gibco), L-glutamine (1X, Gibco GlutaMAX™) and 10% FBS (Pan Biotech), at 37 °C in an atmosphere with 5% CO_2, and 95% humidity.

The metabolic labeling was conducted as described in a previous publication.[5] Briefly, cells are seeded at a density of $3 \cdot 10^5$ on a 15 cm culture dish in custom DMEM containing no regular inositol nor FBS but 100 µM [$^{13}C_6$]*myo*-inositol (or the respective isotopomer) and 10% dialyzed FBS (Gibco, product no. 26400044) instead (from here on referred to as "labeling medium"). Upon reaching ~85 % confluency, the cells are split into five 15 cm culture dishes in labeling medium. Upon reaching confluency cells were harvested by trypsination, collected in 50 ml tubes and washed twice with 50 ml ice-cold PBS or 0.9% NaCl solution. Packed cell volumes were determined for quantification. The collected cell pellets were either processed immediately after harvest or flash-frozen and stored at -80 °C. Metabolites were extracted by $HClO_4$-extraction. Lyophilized metabolite extracts were redissolved in D_2O, re-lyophilized, and finally measured in D_2O (dry D_2O from ampulla, Eurisotop D215T). For quantification 100 µM TMPBr was added to each sample. Standard curves for $InsP_6$ and $InsP_5$[2OH] concentrations against TMPBr are reported previously.[5] The concentrations of other ^{13}C-labeled InsP species for which no labeled standards were available were estimated using the standard curve for $InsP_6$. Cellular InsP concentrations were backcalculated from PCV.

TiO_2 enrichment of InsPs for NMR samples was adapted from published procedures.[12] Briefly, 500 µL of InsP containing sample is mixed 1:1 with ice-cold 1 M aq. perchloric acid and incubated for 30 min on ice (frozen samples are thawn directly in the perchloric acid and then incubated on ice). The sample is then centrifuged (10 min, 18 000 g, 4 °C) and the supernatant transferred into a separate 1.5 mL tube containing 5 mg of TiO_2 beads (Titanosphere 5 µm, GL Sciences), which were already washed with 500 µL Milli-Q® water and 500 µL 1 M perchloric acid ($HClO_4$, Supelco). The extract and TiO_2 beads were mixed on a rotary shaker on low speed for 5 min at 4 °C. The beads were briefly washed twice with 500 µL ice-cold 1 M perchloric acid (note: For centrifugation a table centrifuge (IKA miniG, 6000 rpm) was used at 1 min, and for transferring the supernatant without disturbing the TiO_2 beads a 2 µL Eppendorf tip attached to the tip of a 1 mL tip was used). Supernatans were united to check for unbound InsP species, neutralized roughly with 750 µL 2 M potassium hydroxide, centrifuged and the supernatant lyophilized. To eluate InsPs from the TiO_2 beads, the beads were incubated with 250 µL of 10% ammonia solution for 5 min at rt on a rotary shaker. After centrifugation the supernatant was collected in a separate tube. The elution step is repeated once more and the eluates are combined. The combined eluates are filtered through a 0.2 µM syringe filter (Sartorius Minisart RC4) which was subsequently rinsed with 150 µL of Milli-Q® water. The filtrate was collected in a new 1.5 mL tube and lyophilized. To reduce the water content for NMR analysis the lyophilized eluates were redissolved in 500 µL D_2O and lyophilized again. For NMR measurement the eluates are redissolved in 500 µL D_2O, pH* was adjusted to 6.0.

For the **metabolic flux analysis** *via* CE-MS HEK293 cells were first metabolically labeled with [$^{13}C_6$]Ins as described above over two passages. One week prior to harvest, $4 \cdot 10^5$ of the [$^{13}C_6$]Ins-labeled cells were seeded into one 15 cm dish per time point in [$^{13}C_6$]Ins-labeling medium. For each time point (72, 48, 24, 18, 12.5, 8, 4, 2 and 1 h before harvest) one plate had its medium removed and washed once with 0.9% NaCl solution. The cells were then continued to incubate in 4,5[$^{13}C_2$]Ins-containing labeling medium. For harvesting, the cells of one plate were washed with 25 mL 0.9% NaCl solution, trypsinized (3 mL), then resuspended in 7 mL 0.9% NaCl solution, then pelleted, washed once with 15 mL 0.9% NaCl per pellet and kept on ice until flash-freezing and storage at -80 °C until further processing. For preparing CE-MS each cell pellet was processed as follows: Cells were lysed by resuspending in 1 mL ice-cold 1 M $HClO_4$ (4 °C, 10 min) and centrifuged (18 000 g, 5 min, 4 °C). The supernatant was added to 4 mg of TiO_2 beads (prepared as described above) and the TiO_2 enrichment protocol was followed as described above until the first lyophilization step. Lyophilized samples were stored at -20 °C until CE-MS measurement. CE-MS measurements were carried out as described above. InsP-isotopomers were quantified relatively to each other.

Subcellular organelle isolation

For isolating cellular organelles, wild-type HEK293T were grown in complete DMEM up to 80% confluency. The cells were then harvested by scraping (cell scraper VWR, 734-2604) and were washed thrice with PBS by centrifugation at 250 g for 5 mins at 4 °C. The cell pellet was then processed for organelle isolation

using the respective protocols as mentioned below. All the buffers in the following protocols contain protease inhibitors (5 mM Benzamidine and 20 µg/mL pepstatin). Once isolated, the protein concentration was measured using a BCA assay kit following brief sonication. The organelle preparations were then stored at -80 °C until further use. Results are shown in Figure S17.

ER (microsome) Isolation

Intact microsomes were isolated according to published procedures with slight modifications.[13] Briefly, harvested cells we suspended in 2 mL SH buffer (0.25 M sucrose, 5 mM HEPES, pH 7.4) and homogenized in a dounce homogenizer with a clearance of 0.15 – 0.2 mm (tight fitting) with ten strokes. The lysate was centrifuged at 6000 g for 5 mins and the pellet was discarded. The supernatant was centrifuged at 15 000 g for 5 mins. The supernatant was transferred to a new tube while the pellet was resuspended in fresh buffer and centrifuged again at 15 000 g for 5 mins. The two supernatants were pooled and centrifuged at 105 000 g for 40 mins. The pellet was resuspended in 100 µL of SH buffer and stored at -80 °C for further use.

Mitochondria Isolation

Mitochondria were isolated following published procedures with some modifications.[14] Briefly, the harvested cell pellet was weighed and 1 mL cold T-K-Mg buffer (10 mM Tris-HCl pH 7.4, 10 mM KCl, 0.5 mM $MgCl_2$) per 0.15 g of cell pellet was used for resuspending the cells. The suspension was incubated on ice for 10 mins and passed through a 5 µm syringe filter to lyse the cells. Sucrose stock solution (1 M sucrose, 10 mM Tris-HCl pH 7.4) was added immediately to achieve a final concentration of 0.25 M sucrose (3:1 lysed suspension : 1 M sucrose). The lysate was then centrifuged at 1200 g for 3 mins to pellet unbroken cells, nuclei, and other cellular debris. The pellet was discarded and this step was repeated until no pellet was visible. The supernatant was then centrifuged at 15 000 g for 5 mins to pellet the mitochondria. The supernatant was discarded and the pellet was resuspended in STE buffer (0.32 M sucrose, 1 mM EDTA, 10 mM TrisHCl pH 7.4). The pellet was washed with STE buffer twice at 15 000 g for increased mitochondrial purity. The pellet containing mitochondria was resuspended in minimal volume of STE buffer and stored at -80 °C for further use.

Nuclei Isolation

The protocol for isolating nuclei was adapted from Hymer *et al.*[15] with slight modifications. Briefly, cells were grown and harvested by scraping and washing with PBS with centrifugation at 250 g for 5 mins. The cell pellet was resuspended in ice-cold nuclear extraction buffer (320 mM Sucrose, 5 mM $MgCl_2$, 10 mM HEPES,pH 7.4 1% Triton X-100) and incubated on ice for 10 min with mild intermittent mixing. The suspension was centrifuged at 2000 g for 5 mins at 4 °C and the supernatant was discarded. The pellet was washed with nuclear wash buffer (320 mM sucrose, 5 mM $MgCl_2$, 10 mM HEPES, pH 7.4) by centrifugation at 2000 g for 5 mins at 4 °C. The nuclei were then stored at -80 °C for further use.

Western blots

For Western blots the following primary antibodies were used: MINPP1: MIPP(A-8) sc-514214 (Santa Cruz), Calnexin: Calnexin-HRP conjugate C5C9 #40090 (Cell signalling), Fibrillarin: Fibrillarin(B-1)-HRP conjugate sc-166001 (Santa Cruz), Citrate synthase: D7V8B #14309 (Cell signaling). Following secondary antibodies were used: Anti-rabbit IgG, HRP-linked 7074S (Cell Signaling), Anti-mouse IgG, HRP-linked 7076S (Cell Signaling).

For Western blot analysis 20 µg of the respective samples (except recombinant MINPP1, 0.02 µg) were subjected to SDS-PAGE (4-20% Mini-PROTEAN® TGX Precast gels, 10 well, BioRad, 60 - 90 min, 100V, in SDS running buffer (25 mM Tris, 192 mM glycine, 3 mM sodium dodecylsulfate). A Trans-Blot® SD system (BioRad) was used for transfer onto a nitrocellulose membrane. Standard immunoblotting techniques were applied afterwards.

References

(1) Puschmann, R.; Harmel, R. K.; Fiedler, D. Analysis of Metabolically Labeled Inositol Phosphate Messengers by NMR. In *Methods in Enzymology Vol ume 641*; 2020; pp 35–52. https://doi.org/10.1016/bs.mie.2020.04.035.

(2) Cho, J.; King, J. S.; Qian, X.; Harwood, A. J.; Shears, S. B. Dephosphorylation of 2,3-Bisphosphoglycerate by MIPP Expands the Regulatory Capacity of the Rapoport-Luebering Glycolytic Shunt. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105* (16), 5998–6003. https://doi.org/10.1073/pnas.0710980105.

(3) Cheng, Y.-C.; Prusoff, W. H. Relationship between the Inhibition Constant (KI) and the Concentration of Inhibitor Which Causes 50 per Cent Inhibition (I50) of an Enzymatic Reaction. *Biochem. Pharmacol.* **1973**, *22* (23), 3099–3108. https://doi.org/10.1016/0006-2952(73)90196-2.

(4) Puschmann, R.; Harmel, R. K.; Fiedler, D. Scalable Chemoenzymatic Synthesis of Inositol Pyrophosphates. *Biochemistry* **2019**, *58* (38), 3927–3932. https://doi.org/10.1021/acs.biochem.9b00587.

(5) Harmel, R. K.; Puschmann, R.; Nguyen Trung, M.; Saiardi, A.; Schmieder, P.; Fiedler, D. Harnessing 13 C-Labeled Myo -Inositol to Interrogate Inositol Phosphate Messengers by NMR. *Chem. Sci.* **2019**, *10*, 5267–5274. https://doi.org/10.1039/C9SC00151D.

(6) Qiu, D.; Eisenbeis, V. B.; Saiardi, A.; Jessen, H. J. Absolute Quantitation of Inositol Pyrophosphates by Capillary Electrophoresis Electrospray Ionization Mass Spectrometry. *J. Vis. Exp.* **2021**, *2021* (174), 1–13. https://doi.org/10.3791/62847.

(7) Craxton, A.; Caffrey, J. J.; Burkhart, W.; Safrany, T. S.; Shears, B. S. Molecular Cloning and Expression of a Rat Hepatic Multiple Inositol Polyphosphate Phosphatase. *Biochem. J.* **1997**, *328* (1), 75–81. https://doi.org/10.1042/bj3280075.

(8) Cho, J.; Choi, K.; Darden, T.; Reynolds, P. R.; Petitte, J. N.; Shears, S. B. Avian Multiple Inositol Polyphosphate Phosphatase Is an Active Phytase That Can Be Engineered to Help Ameliorate the Planet's "Phosphate Crisis." *J. Biotechnol.* **2006**, *126* (2), 248–259. https://doi.org/10.1016/j.jbiotec.2006.04.028.

(9) Singh, A.; Upadhyay, V.; Upadhyay, A. K.; Singh, S. M.; Panda, A. K. Protein Recovery from Inclusion Bodies of Escherichia Coli Using Mild Solubilization Process. *Microb. Cell Fact.* **2015**, *14* (1), 1–10. https://doi.org/10.1186/s12934-015-0222-8.

(10) Dovey, C. M.; Diep, J.; Clarke, B. P.; Hale, A. T.; McNamara, D. E.; Guo, H.; Brown, N. W.; Cao, J. Y.; Grace, C. R.; Gough, P. J.; Bertin, J.; Dixon, S. J.; Fiedler, D.; Mocarski, E. S.; Kaiser, W. J.; Moldoveanu, T.; York, J. D.; Carette, J. E. MLKL Requires the Inositol Phosphate Code to Execute Necroptosis. *Mol. Cell* **2018**, *70* (5), 936-948.e7. https://doi.org/10.1016/j.molcel.2018.05.010.

(11) Ucuncu, E.; Rajamani, K.; Wilson, M. S. C.; Medina-Cano, D.; Altin, N.; David, P.; Barcia, G.; Lefort, N.; Banal, C.; Vasilache-Dangles, M. T.; Pitelet, G.; Lorino, E.; Rabasse, N.; Bieth, E.; Zaki, M. S.; Topcu, M.; Sonmez, F. M.; Musaev, D.; Stanley, V.; Bole-Feysot, C.; Nitschké, P.; Munnich, A.; Bahi-Buisson, N.; Fossoud, C.; Giuliano, F.; Colleaux, L.; Burglen, L.; Gleeson, J. G.; Boddaert, N.; Saiardi, A.; Cantagrel, V. MINPP1 Prevents Intracellular Accumulation of the Chelator Inositol Hexakisphosphate and Is Mutated in Pontocerebellar Hypoplasia. *Nat. Commun.* **2020**, *11* (1). https://doi.org/10.1038/s41467-020-19919-y.

(12) Wilson, M. S. C.; Bulley, S. J.; Pisani, F.; Irvine, R. F.; Saiardi, A. A Novel Method for the Purification of Inositol Phosphates from Biological Samples Reveals That No Phytate Is Present in Human Plasma or Urine. *Open Biol.* **2015**, *5* (3), 150014. https://doi.org/10.1098/rsob.150014.

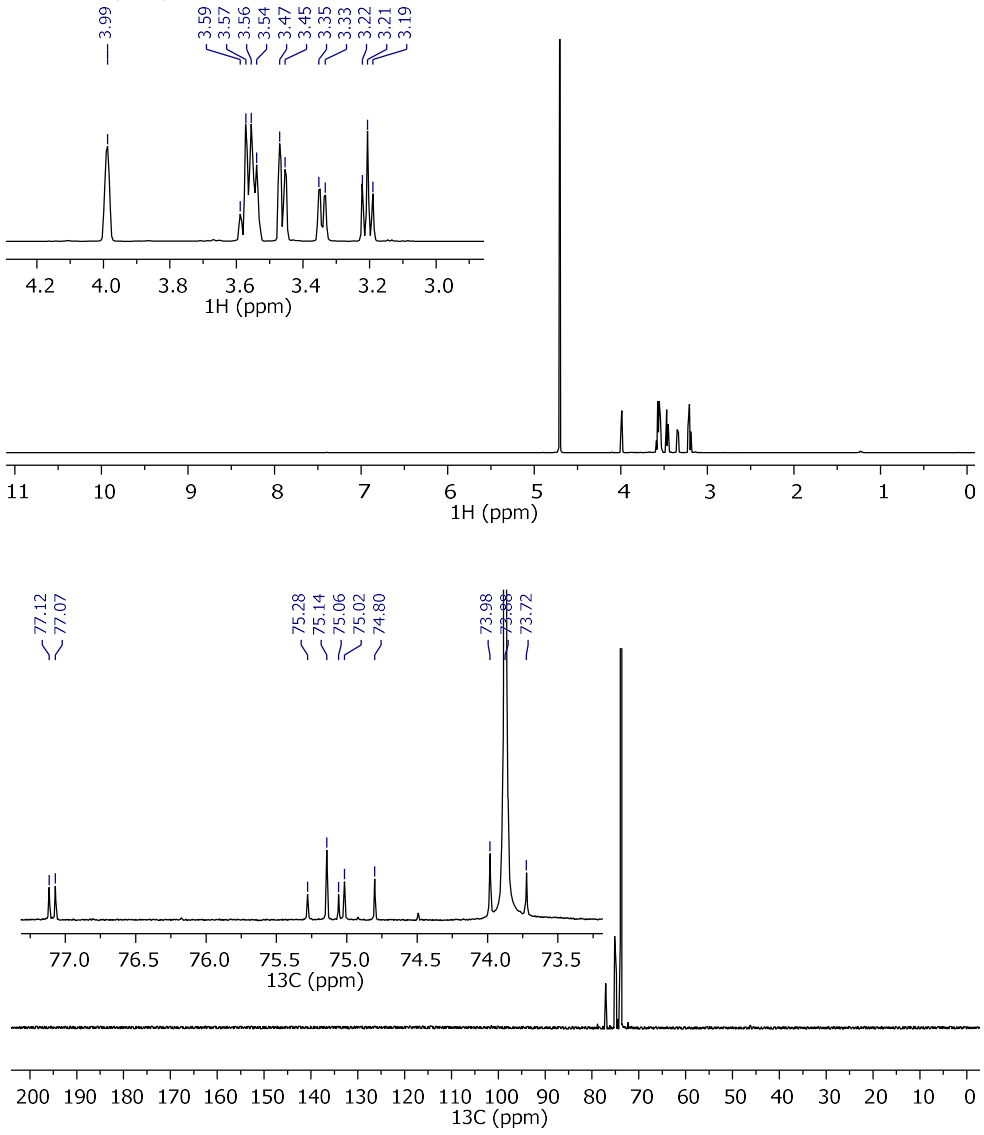
(13) Sukhodub, A. L.; Burchell, A. Preparation of Intact Microsomes from Cultured Mammalian H4IIE Cells. *J. Pharmacol. Toxicol. Methods* **2005**, *52* (3), 330–334. https://doi.org/10.1016/j.vascn.2005.04.016.

(14) Choi, A.; Barrientos, A. Sucrose Gradient Sedimentation Analysis of Mitochondrial Ribosomes. In *Methods in Molecular Biology*; 2021; pp 211–226. https://doi.org/10.1007/978-1-0716-0834-0_16.

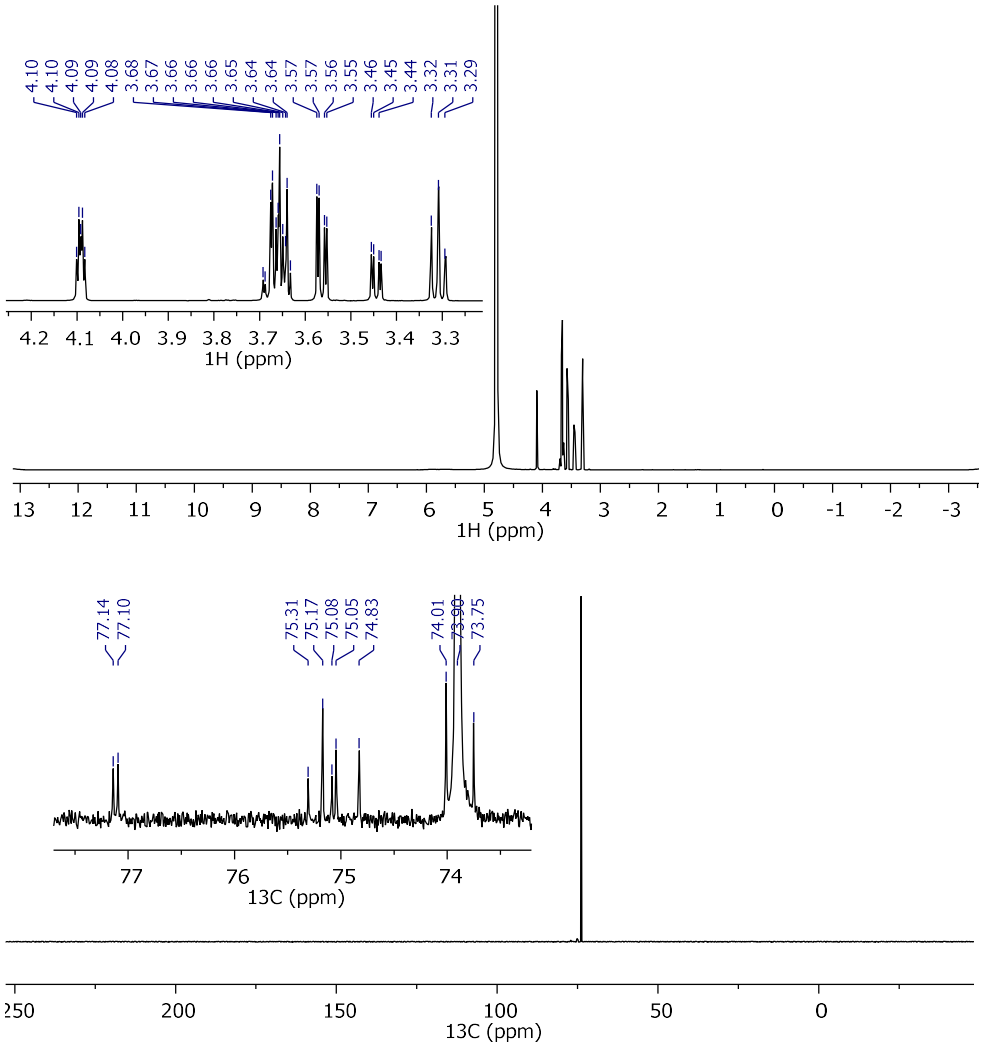
(15) HYMER, W. C.; KUFF, E. L. ISOLATION OF NUCLEI FROM MAMMALIAN TISSUES THROUGH THE USE OF TRITON X-100. *J. Histochem. Cytochem.* **1964**, *12* (5), 359–363. https://doi.org/10.1177/12.5.359.

NMR spectra

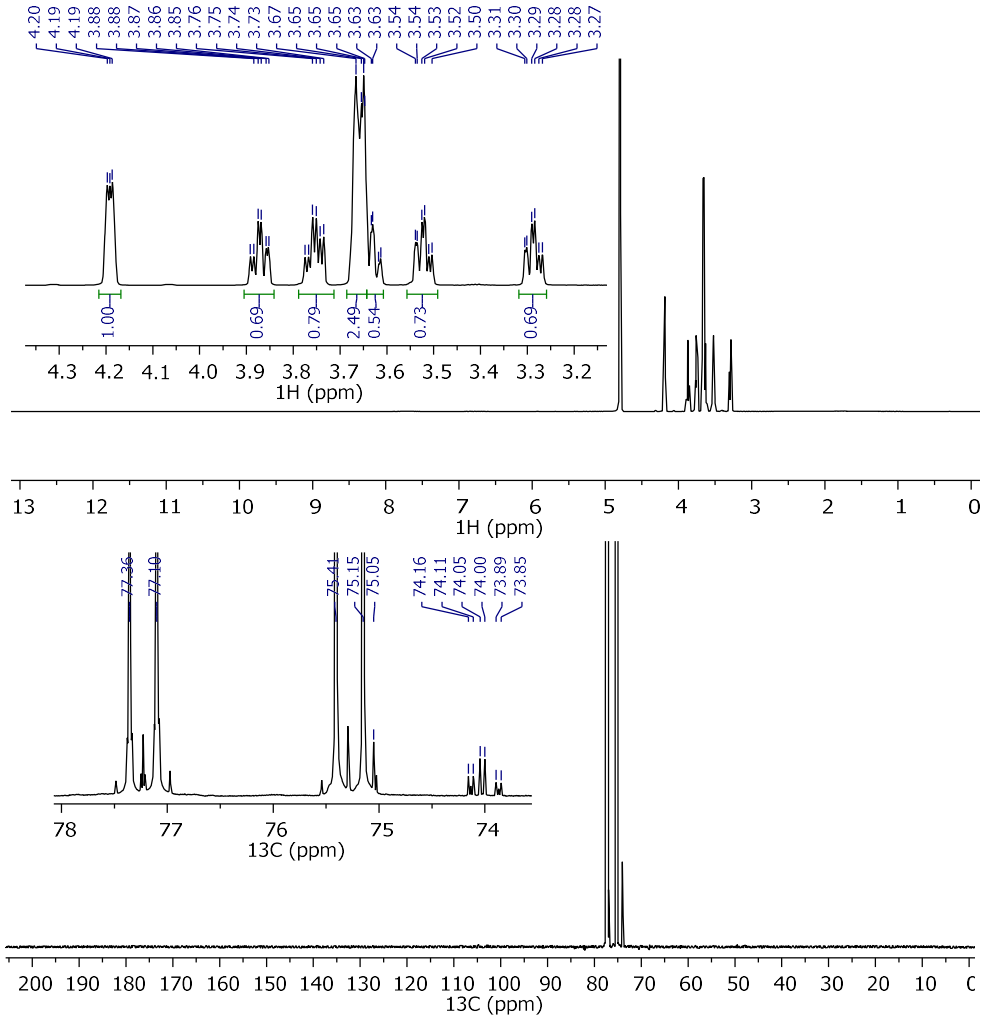
1[$^{13}C_1$]Ins (**S3a**):



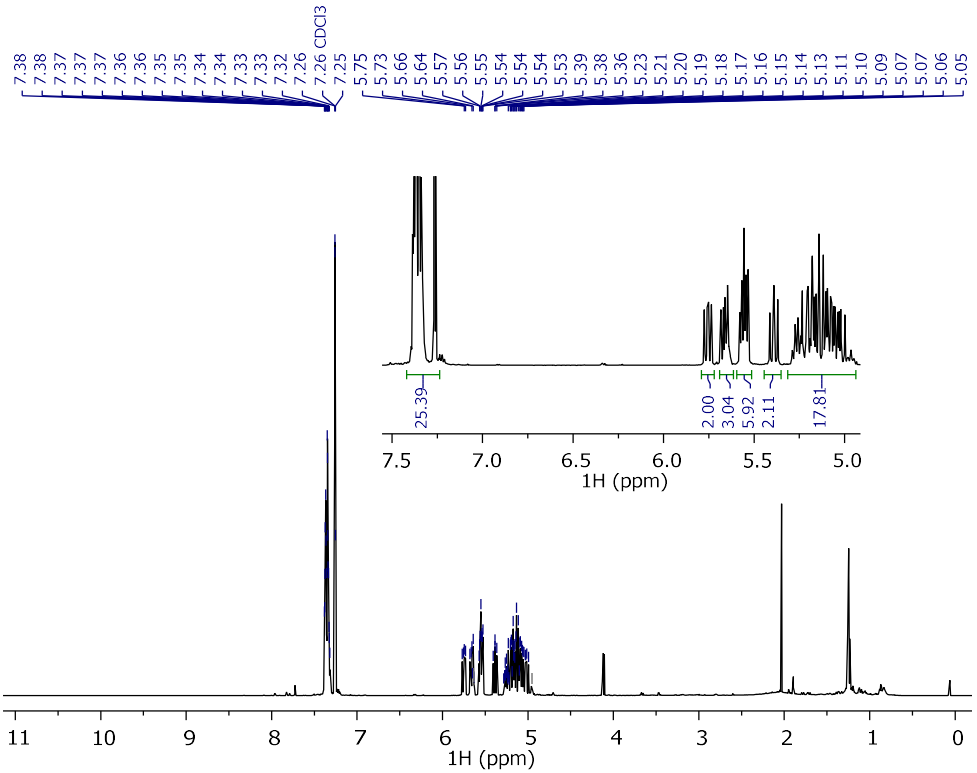
3[$^{13}C_1$]Ins (**S3a'**):

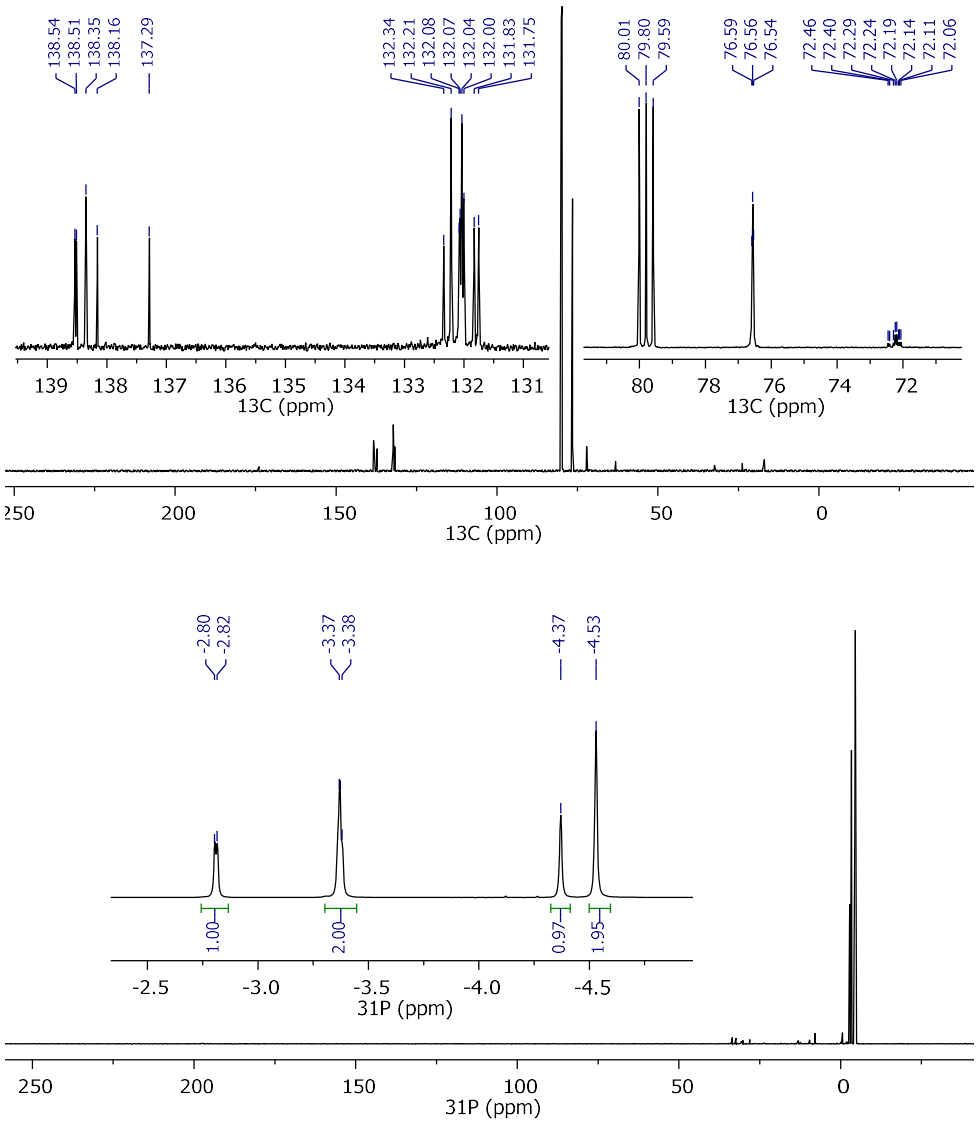


4,5[$^{13}C_2$]Ins (**S3b**):

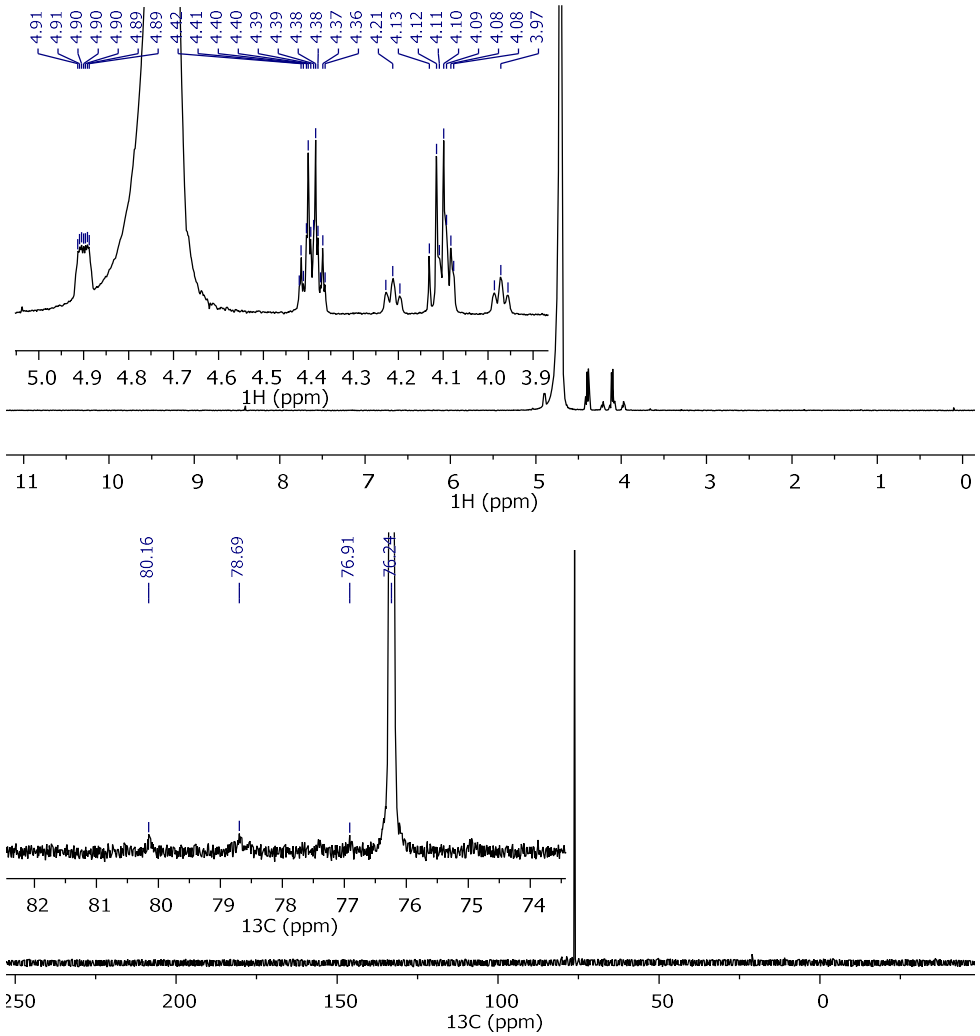


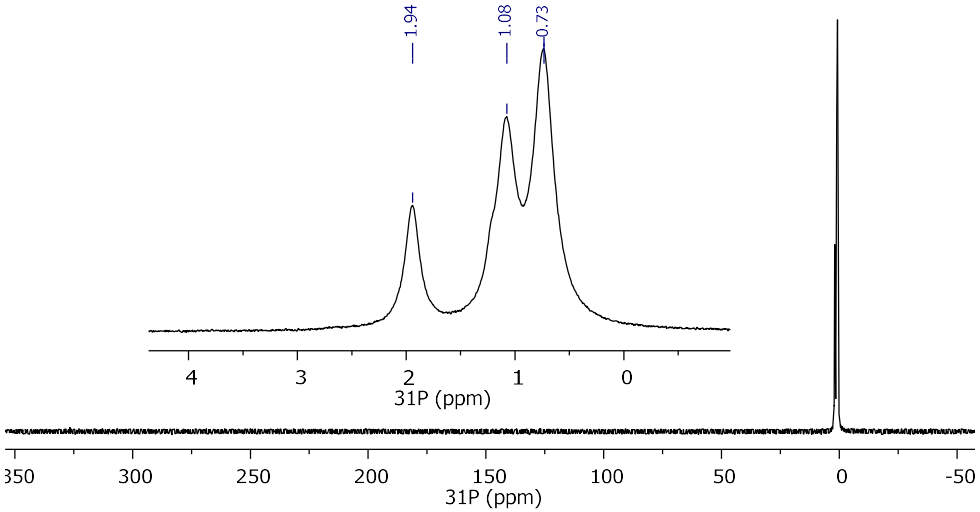
S5:


7.38
7.38
7.37
7.37
7.37
7.36
7.36
7.35
7.35
7.34
7.34
7.33
7.33
7.32
7.26
7.26 CDCl3
7.25
5.75
5.73
5.66
5.64
5.57
5.56
5.55
5.54
5.54
5.54
5.53
5.39
5.38
5.36
5.23
5.21
5.20
5.19
5.18
5.17
5.16
5.15
5.14
5.13
5.11
5.10
5.09
5.07
5.07
5.06
5.05
25.39
2.00
3.04
5.92
2.11
17.81
7.5
7.0
6.5
6.0
5.5
5.0
1H (ppm)
11
10
9
8
7
6
5
4
3
2
1
0
1H (ppm)


138.54
138.51
138.35
138.16
137.29
132.34
132.21
132.08
132.07
132.04
132.00
131.83
131.75
80.01
79.80
79.59
76.59
76.56
76.54
72.46
72.40
72.29
72.24
72.19
72.14
72.11
72.06
139 138 137 136 135 134 133 132 131
13C (ppm)
80 78 76 74 72
13C (ppm)
250 200 150 100 50 0
13C (ppm)
-2.80
-2.82
-3.37
-3.38
-4.37
-4.53
1.00
2.00
0.97
1.95
-2.5 -3.0 -3.5 -4.0 -4.5
31P (ppm)
250 200 150 100 50 0
31P (ppm)

1[$^{13}C_6$]InsP$_6$ (**S6**):




1.94
1.08
0.73
4
3
2
1
0
31P (ppm)
350
300
250
200
150
100
50
0
-50
31P (ppm)

Supporting Information: Numerical Analysis

Stable isotopomers of *myo*-inositol to uncover the complex MINPP1-dependent inositol phosphate network

Minh Nguyen Trung[1,2], Stefanie Kieninger[3], Zeinab Fandi[1], Danye Qiu[4], Guizhen Liu[4], Neelay K. Mehendale[1], Adolfo Saiardi[5], Henning Jessen[4], Bettina G. Keller[3], Dorothea Fiedler[1,2,*]

1 Leibniz-Forschungsinstitut für Molekulare Pharmakologie, Robert-Rössle-Straße 10, 13125 Berlin, Germany
2 Institut für Chemie, Humboldt-Universität zu Berlin, Brook-Taylor-Straße 2, 12489 Berlin, Germany
3 Institut für Chemie, Freie Universität Berlin, Arnimallee 22, 14195 Berlin
4 Albert-Ludwigs-Universität Freiburg
5 MRC Laboratory for Molecular Cell Biology, University College London, WC1E 6BT London, UK
* corresponding author: fiedler@fmp-berlin.de

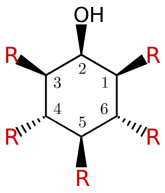
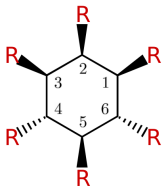
Contents

List of Abbreviations

$InsP_5$[2OH]	inositol-1,3,4,5,6-pentakisphosphate
$InsP_6$	inositol hexakisphosphate
MINPP1	Multiple Inositol Polyphosphate Phosphatase 1
InsPx	inositol polyphosphate (in general)
NMR	nuclear magnetic resonance (spectroscopy)

1 Introduction

This Supplementary Information (SI) contains all information on the numerical evaluation of the reaction rates for the $InsP_5$[2OH]- and $InsP_6$-dephosphorylation from the experimental data. For additional information on the experimental part and figures S1 - S12, please consult the other SI file. Here, we explain the theoretical background of the applied model and the procedure and assumptions that let to the final results presented in main part Fig. 6a and 6b. In this SI, we introduce a six-digit binary representation of the structure names as shown in S18. The numbers in the binary code represent the groups attached to the Cyclohexane scaffold, where the number "0" encodes the hydroxyle group -OH and the number "1" encodes the phosphoryl group $-OPO_3^{2-}$. The position of the number in the binary code (read from left to right) corresponds to the position of the corresponding group in the Cyclohexane scaffold (see S18). For example, the binary representation of Ins(1,2,5,6)P_4 reads 110011 and the binary representation of Ins(4,6)P_2 is 000101.

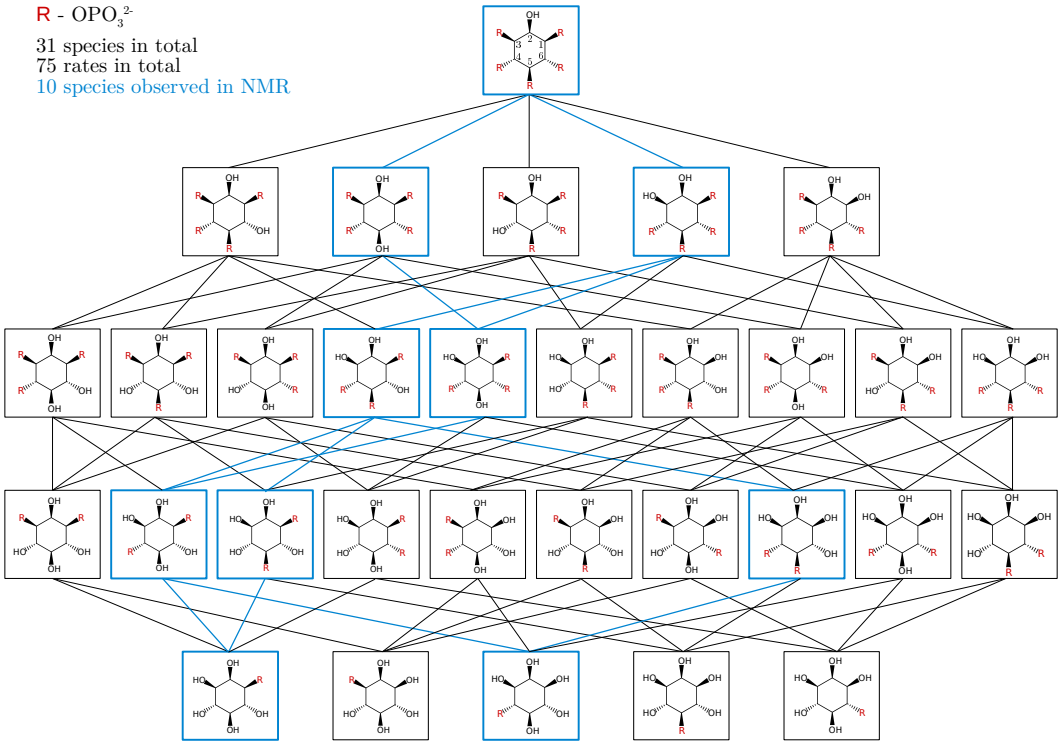
name representation:	$InsP_5$[2OH]	$InsP_6$
binary representation:	101111	111111
structure: R - OPO_3^{2-}		

S 18:
$InsP_5$[2OH] and $InsP_6$ and their corresponding binary representations and structures.

2 Theory

2.1 Full $InsP_5$[2OH] dephosphorylation network

S19 depicts the full reaction pathway of the MINPP1-mediated dephosphorylation of $InsP_5$[2OH]. The network contains all possible intermediates and products including their connection pattern. Each line in S19 represents a rate that describes the reaction from the higher phosphorylated InsPx to the lower phosphorylated InsPx. Since MINPP1 is a phosphatase, the respective reverse reactions are neglected and the network is to be read from top to bottom. The full network contains a total of 75 rates and 31 species, 10 of

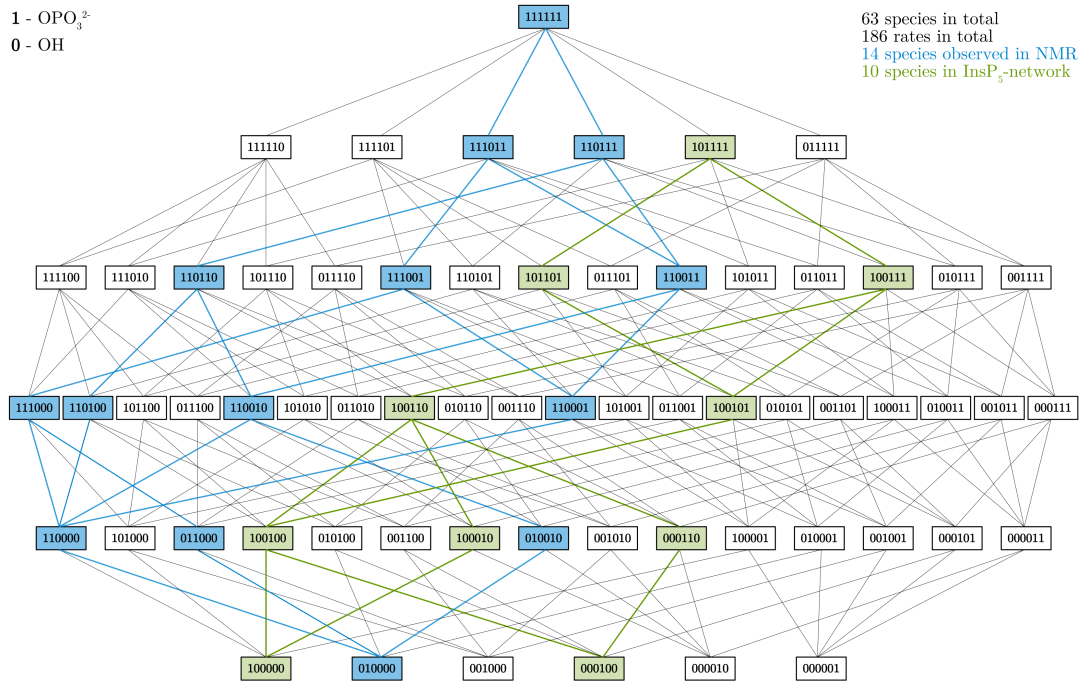


S 19:
Full reaction network with all theoretically possible intermediates and reaction rates of the MINPP1-mediated dephosphorylation of $InsP_5$[2OH]. The structures enclosed in a blue box have been identified in the NMR-experiments with the blue lines depicting the corresponding connection pattern.

which have been identified in the NMR-experiments (blue boxes). We assume that the dephosphorylation network of $InsP_5$[2OH] is dominated by these 10 observed InsPx (see S8) and the corresponding 13 rates (blue lines).

2.2 Full $InsP_6$ dephosphorylation network

S20 depicts the full reaction network of the MINPP1-mediated dephosphorylation of $InsP_6$ in the binary representation. The network contains all possible intermediates and products including their connection pattern. Each line represents a rate that describes the reaction from the higher phosphorylated InsPx to the lower phosphorylated InsPx. Since MINPP1 is a phosphatase, the respective reverse reactions are neglected and the network is to be read from top to bottom. The full network contains a total of 186 rates and 63 species, 14 of which have been identified in the NMR-experiments (blue highlighted boxes) with symmetrically and asymmetrically ^{13}C-labeled $InsP_6$. We assume that the



S 20:
Full reaction network with all theoretically possible intermediates and reaction rates of the MINPP1-mediated dephosphorylation of $InsP_6$. The structures highlighted in blue have been identified in the NMR-experiments with the blue lines depicting the corresponding connection pattern. The $InsP_5$ dephosphorylation network (S19) is highlighted in green.

$InsP_6$ dephosphorylation pathway is dominated by the observed 14 InsPx (see S9) and the corresponding 21 rates (blue lines). Additionally, S20 compares the $InsP_6$ dephosphorylation network (highlighted in blue) to the $InsP_5$[2OH] dephosphorylation network (highlighted in green). We can clearly see that the two networks do not overlap and therefore do not share a single structure or rate.

2.3 Master equation formalism

All processes in the $InsP_5$[2OH] dephosphorylation network (SI Fig 19) as well as in the $InsP_6$ dephosphorylation network (S20) are irreversible chemical reactions of the type

$$A_j \xrightarrow{k_{ij}} A_i \quad \text{with} \quad i, j = 0, 1, \ldots, N-1, \; i \neq j \,, \tag{2.1}$$

where species A_j reacts to species A_i with the rate constant k_{ij}. N is the total number of species within the network. Please note that we start counting from zero to present the theory in line with the implementation of our analysis in Python3. The rate constants k_{ij} are the matrix elements of the rate matrix $\mathbf{K} \in \mathbb{R}^{N \times N}$. To ensure mass conservation the

diagonal elements of $\mathbf{K}$ are defined as the negative of the sum of all other elements in the same column (constraints to master equation)

$$k_{ii} = -\sum_{j=1}^{N} k_{ji} \quad \text{for} \quad i = 0, 1, \ldots, N-1 \text{ and } i \neq j \tag{2.2}$$

such that the sum over each column evaluates to zero. In other words, $\mathbf{K}$ is column-normalized to zero. The vector $\boldsymbol{\phi}(t) \in \mathbb{R}^N$ collects the density (or concentration) $\phi_{A_i}(t)$ at time t of all species. The master equation that corresponds to scheme 2.1 reads

$$\dot{\boldsymbol{\phi}}(t) = \mathbf{K}\,\boldsymbol{\phi}(t)\,, \tag{2.3}$$

where $\dot{\boldsymbol{\phi}}(t)$ is the first derivative of $\boldsymbol{\phi}(t)$ with respect to time. Eq. 2.3 yields N coupled linear homogeneous first order differential equations which describe the kinetics of the entire network. Given the time series $\boldsymbol{\phi}(t)$ (e.g. from experimental data), the master equation can be used to numerically determine the corresponding rates k_{ij} and incorporate additional constraints.[1,2]

2.4 Propagator formalism

In the previous subsection, we describe the kinetics with the corresponding master equation, that is via the change of the density (or concentration) with respect to time. Next, we want to introduce the propagator formalism with which the time evolution of the density (or concentration) for N species can be described directly without the use of time derivatives.
The solution of eq. 2.3 is given as

$$\phi(\tau) = \exp(\mathbf{K}\tau)\phi_0\,, \tag{2.4}$$

where $\phi_0 = \phi(\tau = 0)$ denotes the initial condition at time $\tau = 0$. Eq. 2.4 contains the operator

$$\mathbb{P}(\tau) = \exp(\mathbf{K}\tau) \quad \text{with: } \mathbb{P}(\tau) \in \mathbb{R}^{N\times N}\,, \tag{2.5}$$

which is called propagator and acts on the initial density $\boldsymbol{\phi}(0) = \boldsymbol{\phi}_0 \in \mathbb{R}^N$ to yield the density $\boldsymbol{\phi}_\tau$ after time τ. A given propagator $\mathbb{P}(\tau)$ can only propagate the density in increments of the lag time τ to yield the time series $\boldsymbol{\phi}_0, \boldsymbol{\phi}_\tau, \boldsymbol{\phi}_{2\tau}, \ldots, \boldsymbol{\phi}_{n\tau}$ with $n \in \mathbb{N}$

according to

$$\begin{aligned}
\boldsymbol{\phi}_{\tau} &= \mathbb{P}(\tau)\,\boldsymbol{\phi}_0 \\
\boldsymbol{\phi}_{2\tau} &= \mathbb{P}(\tau)\,\boldsymbol{\phi}_{\tau} \\
\boldsymbol{\phi}_{3\tau} &= \mathbb{P}(\tau)\,\boldsymbol{\phi}_{2\tau} \\
&\vdots \\
\boldsymbol{\phi}_{n\tau} &= \mathbb{P}(\tau)\,\boldsymbol{\phi}_{(n-1)\tau}
\end{aligned} \tag{2.6}$$

By recursively inserting each equation into the other we get

$$\begin{aligned}
\boldsymbol{\phi}_{n\tau} &= \overbrace{\mathbb{P}(\tau)\mathbb{P}(\tau)\cdots\mathbb{P}(\tau)}^{n \text{ times}}\,\boldsymbol{\phi}_0 \\
\boldsymbol{\phi}_{n\tau} &= \mathbb{P}^n(\tau)\,\boldsymbol{\phi}_0\,.
\end{aligned} \tag{2.7}$$

We want to emphasize that, similar to the master equation formalism, conservation of mass is automatically incorporated into the propagator formalism via the rate matrix $\mathbf{K}$ (see eq. 2.2) such that $\mathbb{P}$ is column-normalized to one. In summary, given a propagator $\mathbb{P}(\tau)$ we can compute the density $\boldsymbol{\phi}_{n\tau}$ at time $n\tau$ either by computing all intermediate steps as described in eqs. 2.6 or evaluate $\boldsymbol{\phi}_{n\tau}$ directly via eq. 2.7. In other words, given all rates k_{ij} we can use the propagator formalism to predict the progress curves of all N species in the network.[2,3]

2.5 Minimization method to numerically determine rates

Let's assume we experimentally obtained the concentration of all N species within a network at different discrete times. In other words, we know the density (concentration) vectors $\boldsymbol{\phi}_0^{\text{exp}}, \boldsymbol{\phi}_{\tau}^{\text{exp}}, \boldsymbol{\phi}_{2\tau}^{\text{exp}}, \ldots, \boldsymbol{\phi}_{n\tau}^{\text{exp}} \in \mathbb{R}^N$ at times $0, \tau, 2\tau, \ldots, n\tau$. From this time series, we can numerically determine the time-derivatives as a finite difference

$$\dot{\boldsymbol{\phi}}_{m\tau}^{\text{exp}} = \frac{\boldsymbol{\phi}_{(m+1)\tau}^{\text{exp}} - \boldsymbol{\phi}_{m\tau}^{\text{exp}}}{\tau} \qquad \text{with} \qquad m = 0, 1, \ldots, n-1\,. \tag{2.8}$$

Additionally, we can define a set of n master equations

$$\dot{\boldsymbol{\phi}}_{m\tau} = \mathbf{K}\,\boldsymbol{\phi}_{m\tau}^{\text{exp}}\,, \tag{2.9}$$

where the elements of $\mathbf{K}$ are unknown. With eq. 2.9 we can predict $\dot{\boldsymbol{\phi}}_{m\tau}$ for a specific choice of $\mathbf{K}$. By selecting one value of m we obtain one master equation for this specific

m as

$$\mathbf{y} = \mathbf{K}\mathbf{x}^{\text{exp}}\,, \tag{2.10}$$

where we abbreviate $\boldsymbol{\phi}^{\text{exp}}_{m\tau} = \mathbf{x}^{\text{exp}} = (x^{\text{exp}}_0, x^{\text{exp}}_1, \ldots, x^{\text{exp}}_{N-1})^T$ and $\dot{\boldsymbol{\phi}}_{m\tau} = \mathbf{y} = (y_0, y_1, \ldots, y_{N-1})^T$. Let $\dot{\boldsymbol{\phi}}^{\text{exp}}_{m\tau} = y^{\text{exp}} = (y^{\text{exp}}_0, y^{\text{exp}}_1, \ldots, y^{\text{exp}}_{N-1})^T$ be the density vector at time $m\tau$ which was calculated numerically from the experimental data via eq. 2.8. We use the mean squared error $\Delta(m\tau)$

$$\Delta(m\tau) = \sum_{i=0}^{N-1} (y_i - y^{\text{exp}}_i)^2\,, \tag{2.11}$$

to measure the error of the prediction of $\mathbf{y}$ described in eq. 2.10 and the experimentally obtained $\mathbf{y}^{\text{exp}}$ (eq. 2.8).
$\mathbf{K}$ is column-normalized to zero such that we can substitute the diagonal matrix elements k_{ii} by eq. 2.2. The error in eq. 2.11 then only depends on the off-diagonal elements of $\mathbf{K}$. We can now use a least-square method to minimize $\Delta(m\tau)$ with respect to these off-diagonal elements to get a rate matrix $\mathbf{K}$ that produces $\mathbf{y}$ as close as possible to the experimentally observed $\mathbf{y}^{\text{exp}}$.

With eqs. 2.10 and 2.11 we only made use of the experimental data $\boldsymbol{\phi}^{\text{exp}}_{(m+1)\tau}$ and $\boldsymbol{\phi}^{\text{exp}}_{m\tau}$ at two distinct times m and $m+1$ to determine $\mathbf{K}$. Next, we extend our approach to all values of m such that we can include the entire experimental time series as described in eq. 2.9. In this context, we define the overall error Δ as a sum over all individual errors $\Delta(m\tau)$ (eq. 2.11)

$$\Delta = \Delta(\tau) + \Delta(2\tau) + \cdots + \Delta(n\tau)\,. \tag{2.12}$$

and minimize eq. 2.12 with respect to the off-diagonal elements k_{ij} in order to get a good estimate for the rate matrix $\mathbf{K}$. Please note, that the described least-square method is particularly effective if the unknown $\mathbf{K}$ is sparse, meaning if it contains a lot of zeros. Furthermore, the method allows for additional constraints (additional to column-normalization, e.g. fixing certain reaction rates to a predefined value) and upper and lower limits for the value of the unknown parameters (e.g. for reaction rates we have $k_{ij} \in [0, 1]$). As indicated in S20, the $InsP_5$[2OH] dephosphorylation is dominated by 10 different InsPx forming a network that includes 13 different rates. Consequently, the corresponding rate matrix $\mathbf{K}$ is 10×10-dimensional and sparse, which makes the minimization process described above a very well suited tool to determine the reaction rates of the kinetic network. The same argument holds for the $InsP_6$ dephosphorylation network which consists of 12 species and 17 rates, yielding a sparse 12×12-dimensional rate matrix. Finally, we want to emphasize that a good initial guess for all elements of $\mathbf{K}$ is crucial for the convergence behaviour of a minimization routine as described above.

2.6 Consecutive first-order kinetics

We consider the simplest example of consecutive first order kinetics which is given as

$$A_1 \xrightarrow{k_{21}} A_2 \xrightarrow{k_{32}} A_3\,. \tag{2.13}$$

where one irreversible reaction from species A_1 to A_2 with the reaction rate k_{21} is followed by a second irreversible reaction from A_2 to A_3 with the reaction rate K_{32}. The naming convention of the reaction scheme follows eq. 2.1. The corresponding master equation is defined in eq. 2.3 with $\boldsymbol{\phi}(t) = (\phi_{A_1}(t), \phi_{A_2}(t), \phi_{A_3}(t))^T \in \mathbb{R}^3$ and $\mathbf{K} \in \mathbb{R}^{3\times 3}$. The diagonal elements of $\mathbf{K}$ are given as $k_{11} = -k_{21}$ and $k_{22} = -k_{32}$ (see eq. 2.2). The master equation yields a system of three coupled linear differential equations

$$\dot{\phi}_{A_1}(t) = \frac{\mathrm{d}\phi_{A_1}(t)}{\mathrm{d}t} = -k_{21}\phi_{A_1}(t) \tag{2.14}$$

$$\dot{\phi}_{A_2}(t) = \frac{\mathrm{d}\phi_{A_2}(t)}{\mathrm{d}t} = k_{21}\phi_{A_1}(t) - k_{32}\phi_{A_2}(t) \tag{2.15}$$

$$\dot{\phi}_{A_3}(t) = \frac{\mathrm{d}\phi_{A_3}(t)}{\mathrm{d}t} = k_{32}\phi_{A_2}(t)\,, \tag{2.16}$$

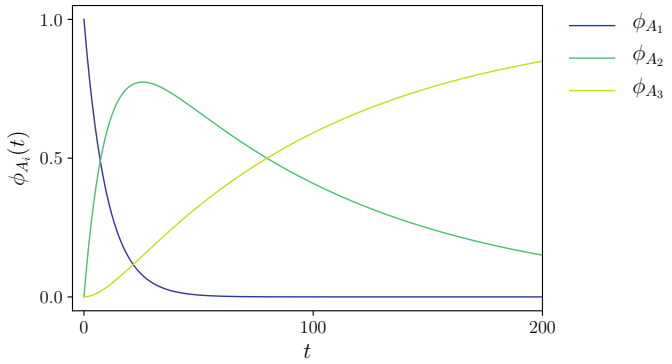
with the analytic solution

$$\phi_{A_1}(t) = \phi^0_{A_1}\ \exp(-k_{21}t) \tag{2.17}$$

$$\phi_{A_2}(t) = k_{21}\phi^0_{A_1}\frac{\exp(-k_{21}t) - \exp(-k_{32}t)}{k_{32} - k_{21}} \tag{2.18}$$

$$\phi_{A_3}(t) = \phi^0_{A_1}\left(1 + \frac{k_{21}\exp(-k_{32}t) - k_{32}\exp(-k_{21}t)}{k_{32} - k_{21}}\right)\,, \tag{2.19}$$

for the case $k_{21} \neq k_{32}$ and the initial condition $\phi_{A_1}(0) = \phi^0_{A_1}$.[4] S21 shows an example for the progress curves defined in eqs. 2.17-2.19 for randomly selected rates.



S 21:
The progress curves for two consecutive first order reactions as defined in eq. 2.13 for a set of example rates.

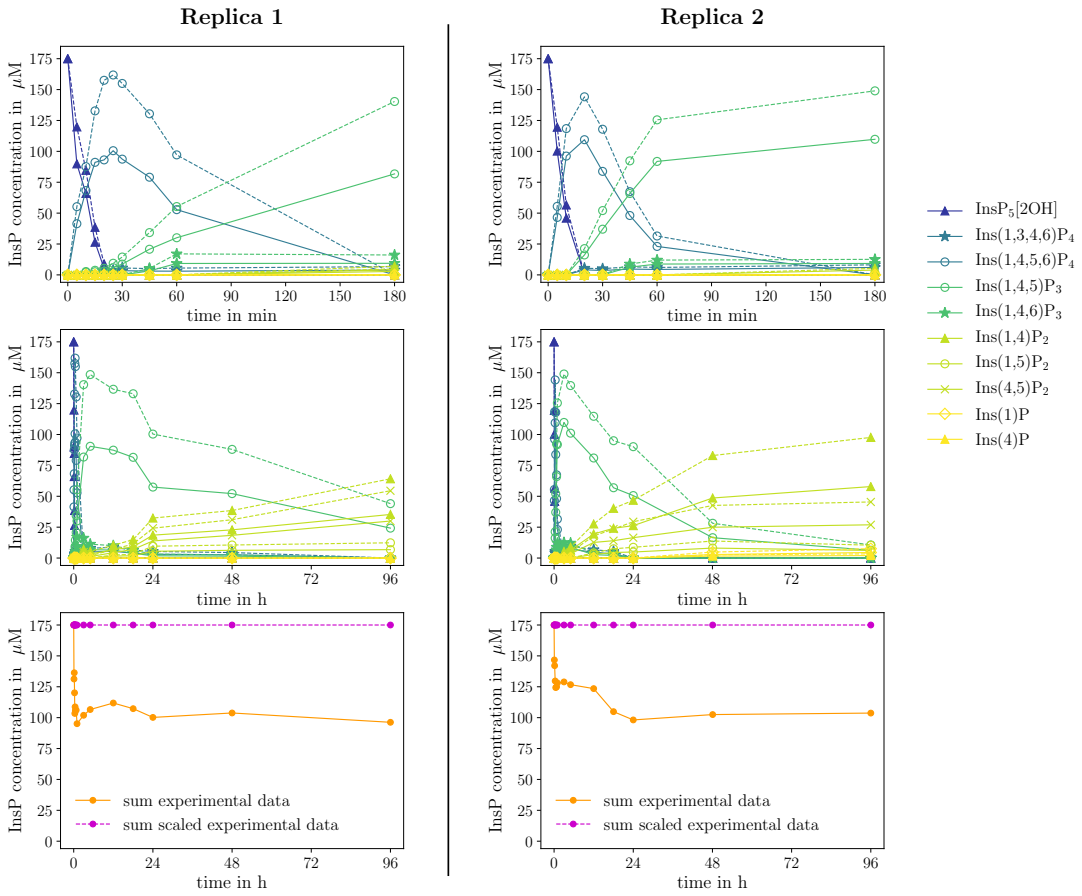
3 Analysis

3.1 Analysis protocol

To extract the reaction rates from the scaled experimental data shown in S22 and S22 we use the least-square minimization routine described in SI section 2.5 following the protocol presented below for both replicas of the $InsP_5[2OH]$ dephosphorylation respectively as well as for the $InsP_6$ dephosphorylation.

1. scale experimental data such that conservation of mass is fulfilled
2. fit scaled experimental data with analytic fit functions
3. create network assumption
4. use fit functions to generate time-equidistant series $\boldsymbol{\phi}_0^{\text{exp}}, \boldsymbol{\phi}_\tau^{\text{exp}}, \ldots, \boldsymbol{\phi}_{n\tau}^{\text{exp}}$ with resolution $\tau = 1$ min
5. use analytical time-derivatives to compute time series $\dot{\boldsymbol{\phi}}_0^{\text{exp}}, \dot{\boldsymbol{\phi}}_\tau^{\text{exp}}, \ldots, \dot{\boldsymbol{\phi}}_{n\tau}^{\text{exp}}$
6. use corresponding network to set-up rate matrix $\mathbf{K}$ and identify all elements that are not equal zero
7. set-up corresponding master equation and extract set of coupled differential equations
8. determine boundary conditions (bounds) and constraints
9. generate initial guess
10. write numerical program using `scipy.optimize.minimize`
11. compute all rates
12. use the rates to predict corresponding progress curves (eq. 2.7) and compare to scaled experimental data

3.2 Experimental data $InsP_5[2OH]$ dephosphorylation



S 22:
Measured progress curves (solid lines) and scaled progress curves (dashed lines) of MINPP1 reaction with 175 μM [$^{13}C_6$]$InsP_5$[2OH] as a concentration time series for two replicas with identical experimental setup. The top row magnifies the first 180 min, the middle row shows the full 96 hours and the bottom row represents the sum over all progress curves, for each of the experiments respectively. The dashed lines in the left column top and middle represent the same data set as main part Fig. 5a.

S22 shows the progress curves of two replicas of the MINPP1 reaction with 175 μM [$^{13}C_6$]$InsP_5$[2OH] (columns) with identical experimental setup, where we conducted NMR-measurements at 10 different points in time for replica 1 and at 8 different points in time for replica 2. The plot at the top magnifies the first 180 min of the experiment and the plot in the middle shows the full 96 hours time interval of the measurements, respectively. The solid lines represent the experimentally measured concentration time series $\phi_i^{\text{exp}}(t)$ with $i = 0, \ldots, N-1$ of the $N = 10$ species that could be identified in the NMR-experiments. The orange line in the bottom plot represents the corresponding sum $S^{\text{exp}}(t)$ over the

concentrations of all 10 species at each point in time

$$S^{\text{exp}}(t) = \sum_{i=0}^{N-1} \phi_i^{\text{exp}}(t)\,, \tag{3.1}$$

We can clearly see that $S^{\text{exp}}(t) \neq 175\ \mu\text{M}$ for all t, meaning that we "loose" mass during the course of the experiment and conservation of mass is not fulfilled by the original experimental data. Since conservation of mass is crucial for the kinetic model we use to extract rates from the experimental data, we correct for the loss of mass by scaling the experimental data according to

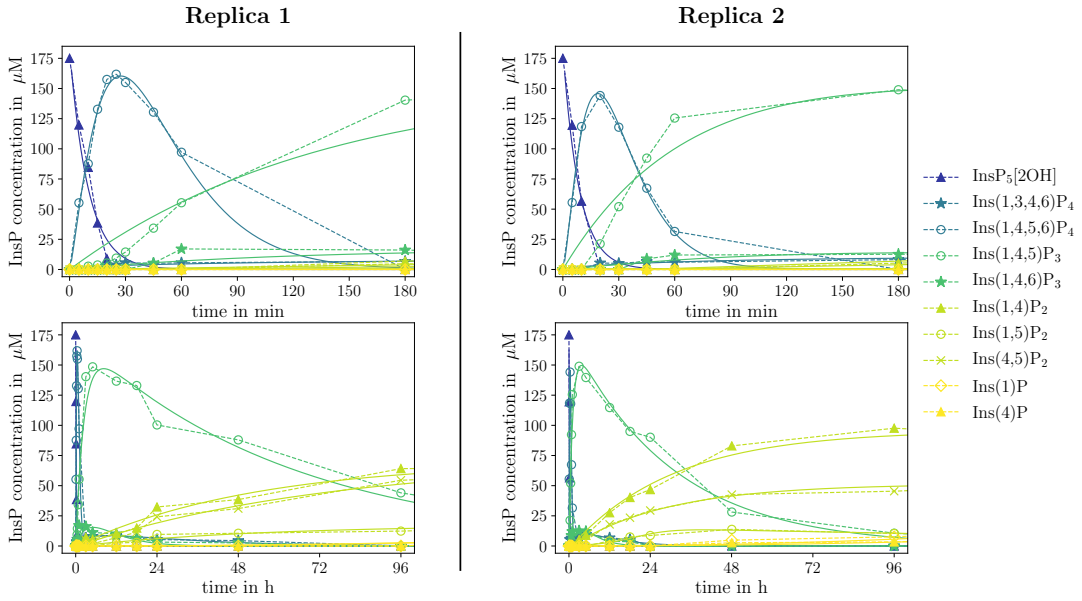
$$\phi_i^{\text{scaled}}(t) = \frac{\phi_i^{\text{original}}(t)}{S^{\text{exp}}(t)} \cdot 175\ \mu\text{M} \tag{3.2}$$

such that

$$S^{\text{scaled}}(t) = \sum_{i=0}^{N-1} \phi_i^{\text{scaled}}(t) = 175\ \mu\text{M} \qquad \forall t\,. \tag{3.3}$$

The scaled progress curves $\phi_i^{\text{scaled}}(t)$ are shown as dashed lines in S22 and in main part Fig. 5a. The results of of both replicas exhibit similar behaviour but the progress curves of replica 2 indicate slightly faster kinetics. In the main part we chose replica 1 as representative for both replicas. To extract kinetic of the $InsP_5$ dephosphorylation, we perform the numerical analysis on both replicas separately and compare the resulting rates in SI section 4.1. Please note that we solely use the scaled progress curves for the numerical analysis.

To prepare the scaled experimental data for the numerical analysis, we fitted the progress curves of each species with an analytic fit function. The fit functions provide access to more and time-equidistant data points and analytical derivatives for each progress curve (no numerical derivatives necessary). S23 compares the fit function to the scaled experimental data for both replicas and SI table 1 summarizes the fit functions and the corresponding fit parameters. Please note, that we used the kinetic function defined in eq. 2.18 to fit the progress curve of Ins(1,4,5)P_3 (dark green circles) meaning that the fit parameters k_1 and k_2 can already be interpreted as reaction rates.



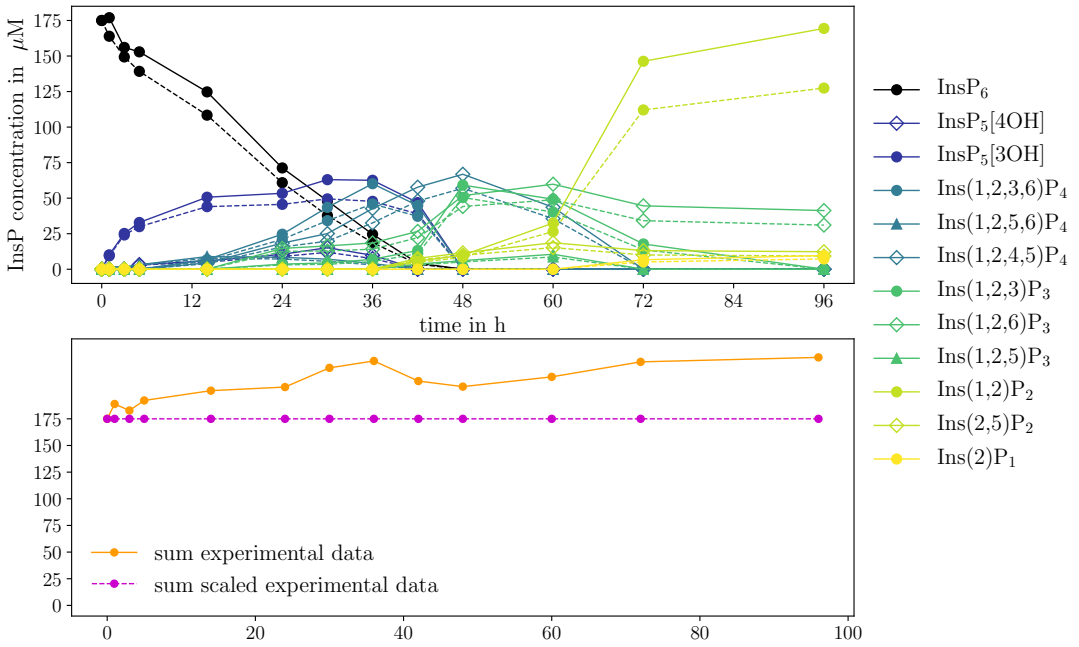
S 23:
Scaled progress curves (dashed lines) of MINPP1 reaction with 175 μM [$^{13}C_6$]InsP$_5$[2OH] as a concentration time series and corresponding fit functions (solid lines). The top row magnifies the first 180 min and the bottom row shows the full 96 hours.

SI Table 1:
Fit functions and parameters used to fit the scaled experimental data of [$^{13}C_6$]InsP$_5$[2OH] dephosphorylation for two replicas.

species	Replica 1	Replica 2
InsP$_5$[2OH]	$f(t) = a\exp(-kt)$ $a = 182.834$ $k = 0.100$	$f(t) = a\exp(-kt)$ $a = 182.020$ $k = 0.114$
Ins(1,3,4,6)P$_4$	$f(t) = a \cdot t^b \cdot \exp(-kt)$ $a = 0.734$ $b = 0.459$ $k = 0.0008$	$f(t) = a \cdot t^b \cdot \exp(-kt)$ $a = 0.575$ $b = 0.612$ $k = 0.002$
Ins(1,4,5,6)P$_4$	$f(t) = a \cdot t^b \cdot \exp(-kt)$ $a = 7.508$ $b = 1.322$ $k = 0.048$	$f(t) = a \cdot t^b \cdot \exp(-kt)$ $a = 7.722$ $b = 1.520$ $k = 0.080$
Ins(1,4,5)P$_3$	$f(t) = \frac{c_0\, k_1}{k_2 - k_1}(\exp(-k_1 t) - \exp(-k_2 t))$ $k_1 = 0.006$ $k_2 = 0.000262$ $c_0 = 167.378$	$f(t) = \frac{c_0\, k_1}{k_2 - k_1}(\exp(-k_1 t) - \exp(-k_2 t))$ $k_1 = 0.0153$ $k_2 = 0.000553$ $c_0 = 169.494$
Ins(1,4,6)P$_3$	$f(t) = a \cdot t^b \cdot \exp(-kt)$ $a = 0.181$ $b = 0.937$ $k = 0.003$	$f(t) = a \cdot t^b \cdot \exp(-kt)$ $a = 0.111$ $b = 1.115$ $k = 0.005$
Ins(1,4)P$_2$	$f(t) = S - (S - a) \cdot \exp(-bt)$ $a = -0.052$ $b = 0.0003$ $S = 70.446$	$f(t) = S - (S - a) \cdot \exp(-bt)$ $a = -2.263$ $b = 0.0005$ $S = 95.617$

species	new experiment	old experiment
$Ins(1,5)P_2$	$f(t) = S - (S - a) \cdot \exp(-bt)$ $a = -0.289$ $b = 0.0003$ $S = 17.523$	$f(t) = \frac{d}{a + b\exp(-ct)} \exp(-gt)$ $a = 2.029$ $b = 78.948$ $c = 0.003$ $d = 33.623$ $g = 0.00007$
$Ins(4,5)P_2$	$f(t) = S - (S - a) \cdot \exp(-bt)$ $a = -0.582$ $b = 0.0002$ $S = 71.352$	$f(t) = S - (S - a) \cdot \exp(-bt)$ $a = -1.303$ $b = 0.0005$ $S = 51.389$
$Ins(1)P_1$	$f(t) = a\,t^2$ $a = 7.89 \cdot 10^{-8}$	$f(t) = a\,t^2$ $a = 1.7 \cdot 10^{-7}$
$Ins(4)P_1$	$f(t) = a\,t^2$ $a = 7.06 \cdot 10^{-8}$	$f(t) = a\,t^2$ $a = 9.64 \cdot 10^{-8}$

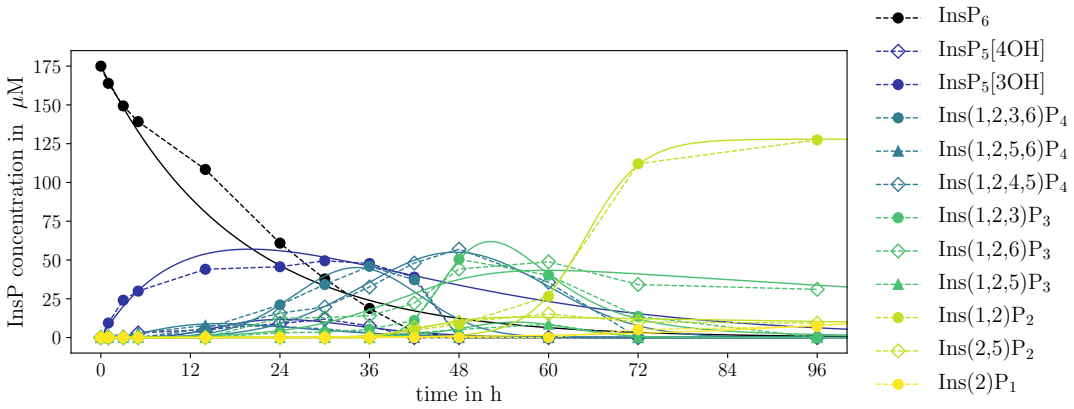
3.3 Experimental data $InsP_6$ dephosphorylation



S 24:
Measured progress curves (solid lines) and scaled progress curves (dashed lines) of MINPP1 reaction with 175 μM [$^{13}C_6$]$InsP_6$ as a concentration time series (top) and the sum over all progress curves at each point in time (bottom). The dashed lines in the top plot represent the same data set as main part Fig. 5c.

S9 and S20 depict our assumption of the complete MINPP1-mediated $InsP_6$ dephosphorylation pathway and main part Fig. 5d shows the corresponding simplified version. The pathway contains the enantiomers Ins(1,2,4)P_3 and Ins(1,2,6)P_3 and the enantiomers Ins(1,2)P_2 and Ins(2,3)P_2 which can only be distinguished in asymmertrically ^{13}C-labeled NMR experiments. However, we base our numerical analysis on the progress curves shown in S24,top which resulted from the MINPP1 reaction with symmetrically labeled [$^{13}C_6$]$InsP_6$. Consequently, both pairs of enantiomers are represented by one progress curve each, which we labeled with one representative for each pair of enantiomers. This reduces the network in S20 from 14 to 12 different species. The solid lines in S24,top, represent the experimentally measured concentration time series $\phi_i^{\text{exp}}(t)$ with $i = 0, \ldots, N-1$ and $N = 12$. The orange line in S24, bottom represents the sum $S^{\text{exp}}(t)$ over the concentrations of all 12 species at each point in time. Since the original experimental data does not obey conservation of mass over the entire time axis, we scale the data according to eq. 3.2. The scaled progress curves are shown as dashed lines and are equivalent to the solid lines in main part Fig. 5c. We want to emphasize that we solely use the scaled progress curves for all further analysis. To prepare the scaled experimental data

for the numerical analysis, we fitted the progress curves of each species with an analytic fit function. S25 compares the fit function to the scaled experimental data and SI table 2 summarizes the fit functions and the corresponding fit parameters. We used eq. 2.17 as fit function to fit the $InsP_6$ progress curve which means that we can interpret the fit parameter k as the reaction rate that quantifies the depletion of $InsP_6$ over time. Moreover, we used eq. 2.18 to fit the $InsP_5$[3OH] progress curve and thus we can interpret the fit parameters k_1 and k_2 as reaction rates that dictate the growth and the depletion of $InsP_5$[3OH] concentration in time.



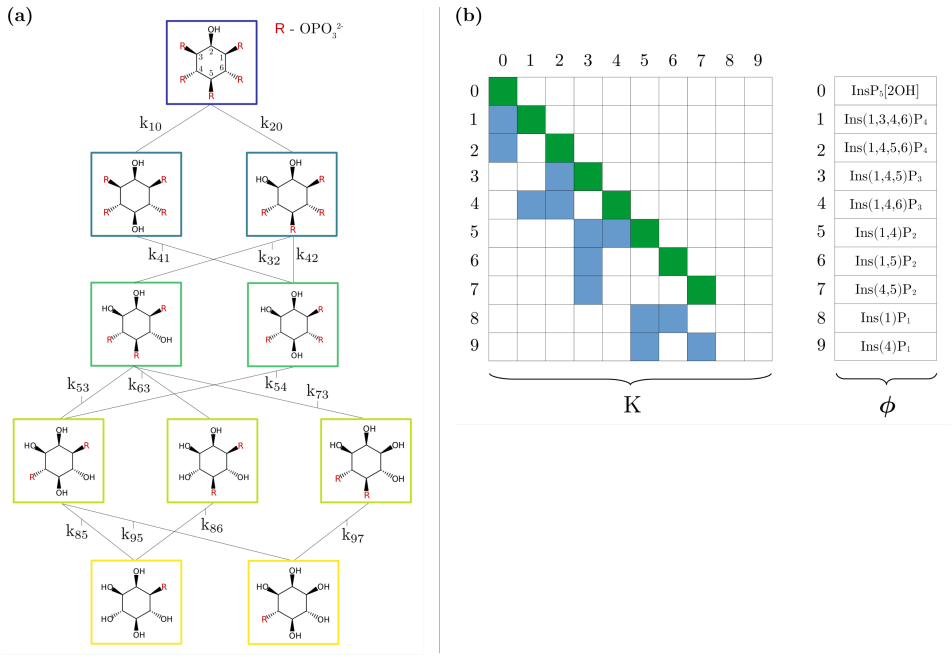
S 25:
Scaled progress curves (dashed lines) of MINPP1 reaction with 175 μM [$^{13}C_6$]$InsP_6$as a concentration time series and corresponding fit functions (solid lines).

SI Table 2:
Fit functions and parameters used to fit the scaled experimental of [$^{13}C_6$]InP$_6$ dephosphorylation.

species	fit data
InsP$_6$	$f(t) = a\exp(-kt)$ $a = 175.00$ $k = 0.000932$
InsP$_5$[4OH]	$f(t) = at^b\exp(-kt)$ $a = 6.2 \cdot 10^{-18}$ $b = 6.697$ $k = 0.00460$
InsP$_5$[3OH]	$f(t) = \frac{k_1 c_0}{k_2 - k_1}(\exp(-k_1 t) - \exp(-k_2 t))$ $k_1 = 0.00074$ $k_2 = 0.00093$ $c_0 = 175$
Ins(1,2,3,6)P$_4$	$f(t) = a \cdot \exp\left(-\frac{(t-\mu)^2}{\sigma^2}\right)$ $a = 45.149$ $\mu = 2057.099$ $\sigma = 709.875$
Ins(1,2,5,6)P$_4$	$f(t) = at^b\exp(-kt)$ $a = 0.0000021$ $b = 2.603$ $k = 0.00272$
Ins(1,2,4,5)P$_4$	$f(t) = a \cdot \exp\left(-\frac{(t-\mu)^2}{\sigma^2}\right)$ $a = 55.045$ $\mu = 2849.46$ $\sigma = 1045.39$

species	fit data
$\mathrm{Ins(1,2,3)P_3}$	$f(t) = a\,\exp\left[-b\Big(1-\exp(-c(t-d))\Big)^2\right]$ $a = 61.89$ $b = 4.578$ $c = 0.00076$ $d = 3133.40$
$\mathrm{Ins(1,2,6)P_3}$	$f(t) = \frac{d}{a+b\,\exp(-c\,t)}\,\exp(-e\,t)$ $a = 0.0973$ $b = 45.655$ $c = 0.00249$ $d = 7.480$ $e = 0.00014$
$\mathrm{Ins(1,2,5)P_3}$	$f(t) = a\cdot\exp\left(-\frac{(t-\mu)^2}{\sigma^2}\right)$ $a = 9.909$ $\mu = 3338.044$ $\sigma = 626.472$
$\mathrm{Ins(1,2)P_2}$	$f(t) = \frac{d}{a+b\,\exp(-c\,t)}$ $a = 0.000046$ $b = 1394.199$ $c = 0.00442$ $d = 0.00590$
$\mathrm{Ins(2,5)P_2}$	$f(t) = \frac{d}{a+b\,\exp(-c\,t)}\,\exp(-e\,t)$ $a = 11.970$ $b = 0.000024$ $c = -0.00497$ $d = 0.000468$ $e = -0.00486$
$\mathrm{Ins(2)P_1}$	$f(t) = \frac{d}{a+b\,\exp(-c\,t)}$ $a = -8.741$ $b = -3609.11$ $c = 0.000978$ $d = -168.34$

3.4 Analysis setup InsP$_5$[2OH] dephosphorylation



S 26:
(a) Assumed network for MINPP1 reaction with 175 μM InsP$_5$[2OH] including all reactions rates. A copy of this network is shown in S8 and a simplified version is depicted in main part Fig. 5b.
(b) Schematic representation of the corresponding rate matrix and density (concentration) vector. Matrix: The white squares mark all matrix elements that are equal to zero, the blue squares all elements that are not zero and the green squares represent the diagonal elements defined via eq. 2.2. Vector: The representation indicates which vector element is associated with which InsP.

Network:
Based on the NMR-data (see main part Fig. 3 and S22), we assume that the InsP$_5$[2OH] dephosphorylation network is dominated by 10 different InsPx that form the network depicted in S26, a. All possible reactions from a higher phosphorylated InsPx to a lower phosphorylated InsPx are indicated with a line and are associated with a reaction rate $k_{ij} \neq 0$.

Density (concentration) vector and corresponding time derivative:
We use the fit functions (SI table 1) to create time-equidistant data points $\boldsymbol{\phi}_0^{\text{exp}}, \boldsymbol{\phi}_\tau^{\text{exp}}, \ldots, \boldsymbol{\phi}_{n\tau}^{\text{exp}}$ and $\dot{\boldsymbol{\phi}}_0^{\text{exp}}, \dot{\boldsymbol{\phi}}_\tau^{\text{exp}}, \ldots, \dot{\boldsymbol{\phi}}_{n\tau}^{\text{exp}}$ (eq. 2.9) with a resolution of $\tau = 1$ min for each replica.

Rate matrix:
To build the rate matrix $\mathbf{K}$, we number all species in the network from zero to nine in a left-to-right and top-to-bottom fashion and assign the corresponding rates according to eq. 2.1. These rates are represented in S26, b as blue squares. The diagonal elements

(green squares) are defined via eq. 2.2 and given as

$$\begin{aligned} k_{00} &= -(k_{10} + k_{20}) \\ k_{11} &= -k_{41} \\ k_{22} &= -(k_{32} + k_{42}) \\ k_{33} &= -(k_{53} + k_{63} + k_{73}) \\ k_{44} &= -k_{54} \\ k_{55} &= -(k_{85} + k_{95}) \\ k_{66} &= -k_{86} \\ k_{77} &= -k_{97}\,. \end{aligned} \tag{3.4}$$

All other matrix elements are equal to zero (white squares in S26, b).

Set of differential equations:
With the rate matrix **K** defined, we can now formulate the corresponding master equation (eq. 2.3) which yields the following set of 10 coupled first-order differential equations

$$\begin{array}{rcccccccccc} \dot{\phi}_0 &=& k_{00}\phi_0 & & & & & & & & \\ \dot{\phi}_1 &=& k_{10}\phi_0 & +k_{11}\phi_1 & & & & & & & \\ \dot{\phi}_2 &=& k_{20}\phi_0 & & +k_{22}\phi_2 & & & & & & \\ \dot{\phi}_3 &=& & & +k_{32}\phi_2 & +k_{33}\phi_3 & & & & & \\ \dot{\phi}_4 &=& & +k_{41}\phi_1 & +k_{42}\phi_2 & & +k_{44}\phi_4 & & & & \\ \dot{\phi}_5 &=& & & & +k_{53}\phi_3 & +k_{54}\phi_4 & +k_{55}\phi_5 & & & \\ \dot{\phi}_6 &=& & & & +k_{63}\phi_3 & & & +k_{66}\phi_6 & & \\ \dot{\phi}_7 &=& & & & +k_{73}\phi_3 & & & & +k_{77}\phi_7 & \\ \dot{\phi}_8 &=& & & & & & +k_{85}\phi_5 & +k_{86}\phi_6 & & \\ \dot{\phi}_9 &=& & & & & & +k_{95}\phi_5 & & +k_{97}\phi_7 & \end{array} \tag{3.5}$$

Constraints and bounds:
Here, we report the applied constraints that yielded the best results for the reaction rates reported in main part Fig. 6a and 6b. In total, we constrained 4 rates (k_{10}, k_{41}, k_{73} and k_{42}), which leaves us with 9 reaction rates that have to be optimized during the minimization routine.

k_{10} and k_{41}:
Since the progress curve of Ins(1,3,4,6)P_4 evolves at very low concentrations (less than 9 μM for the entire time series), we decided to exclude this species from the analysis and

set the corresponding rates to zero, $k_{10}, k_{41} = 0$, for both replicas.

k_{73}:
As mentioned in section 3.2 , we use eq. 2.18 as fit function for the progress curve of Ins(1,3,4)P_3 (SI table 1). Since this function emerges from a kinetic model, we can interpret the corresponding fit parameters k_1 and k_2 as kinetic rates with k_1 describing the increase and k_2 the decrease of concentration. The increase in Ins(1,3,4)P_3 concentration is determined by k_{32} and the decrease is determined by $k_{53} + k_{63} + k_{73}$. We use the fit parameter k_2 ($2.62 \cdot 10^{-4}$ min^{-1} for replica 1 and $5.53 \cdot 10^{-4}$ min^{-1} for replica 2) to constrain the rate k_{73} as $k_{73} = k_2 - k_{53} - k_{63}$ and leave k_{32} unconstraint for each replica respectively.

k_{42}:
During our analysis, we found that the increase in Ins(1,4,6)InsP_3 concentration was generally overestimated by the minimization routine and thus we decided to constrain the rate k_{42} towards Ins(1,4,6)InsP_3 by hand. In an iterative procedure, we found that the constraint $k_{42} = 0.001$ yields the most promising results for both replicas.

bounds:
Since reaction rates are a real number between zero and one, we bound all rates to the interval $k_{ij} \in [10^{-6}, 1]$.

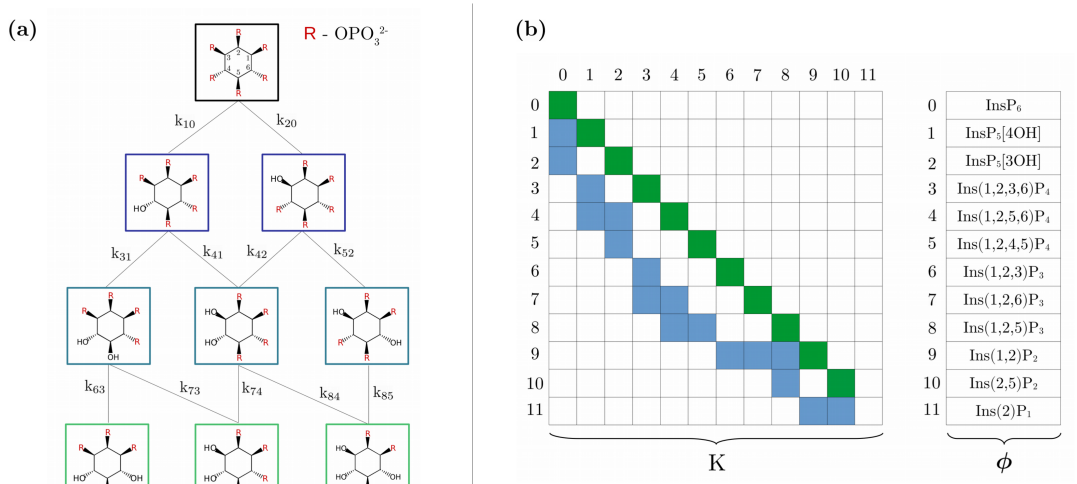
Initial guess:
We built the rate matrix **K** by formulating an initial guess for each rate, used eq. 2.7 to predict the corresponding progress curves and compared these prediction to the scaled experimental data (S22, dashed lines). In an iterative procedure, we corrected the rates by hand until the set of rates produced progress curves that roughly matched the scaled experimental progress curves. The set of rates is summarized in SI table 3 and serves as initial guess for our minimzation routine.

Technical details:
We used Python3 and `scipy.optimize.minimize`[5] to implement the minimization routine described in SI section 2.5, where we passed eq. 2.12 as objective function to be minimized, the initial guess and bounds as described above and left all other parameters at their default settings. Since we constraints 4 of the 13 rates, the implemented minimization routine optimizes the remaining 9 rates such that the resulting rate matrix **K** yields progress curves that are in excellent agreement with the scaled experiment data.

SI Table 3:
Initial guess for all rates of the $InsP_5$[2OH] dephosphorylation network in min^{-1} for both replicas.
*) The marked rates are subject to constraints.

rate	**replica 1**	**replica 2**
k_{10}	0.00^*	0.00^*
k_{20}	$9.76 \cdot 10^{-2}$	$9.76 \cdot 10^{-2}$
k_{41}	0.00^*	0.00^*
k_{32}	$6.84 \cdot 10^{-3}$	$6.84 \cdot 10^{-3}$
k_{42}	$1.00 \cdot 10^{-3*}$	$1.00 \cdot 10^{-3*}$
k_{53}	$1.09 \cdot 10^{-4}$	$1.09 \cdot 10^{-4}$
k_{63}	$4.16 \cdot 10^{-5}$	$4.16 \cdot 10^{-5}$
k_{73}	$(2.62 \cdot 10^{-4} - k_{53} - k_{63})^*$	$(5.53 \cdot 10^{-4} - k_{53} - k_{63})^*$
k_{54}	$6.11 \cdot 10^{-4}$	$6.11 \cdot 10^{-4}$
k_{85}	$1.00 \cdot 10^{-5}$	$1.00 \cdot 10^{-5}$
k_{95}	$1.08 \cdot 10^{-5}$	$1.08 \cdot 10^{-5}$
k_{86}	$5.77 \cdot 10^{-5}$	$5.77 \cdot 10^{-5}$
k_{97}	$2.40 \cdot 10^{-6}$	$1.00 \cdot 10^{-6}$

3.5 Analysis setup $InsP_6$ dephosphorylation



S 27:
(a) Assumed network for MINPP1 reaction with 175 μM $InsP_6$ including all reactions rates.
(b) Schematic representation of the corresponding rate matrix and density (concentration) vector. Matrix: The white squares mark all matrix elements that are equal to zero, the blue squares all elements that are not zero and the green squares represent the diagonal elements defined via eq. 2.2. Vector: The representation indicates which vector element is associated with which InsP.

Network:

S27a depicts the network assumption on which we base our numerical analysis. Ins(1,2,6)P_3 and Ins(1,2)P_2 are chosen as representatives for their respective pair of enantiomiers (see also section 3.3). All possible reactions from a higher phosphorylated InsPx to a lower phosphorylated InsPx are indicated with a line and are associated with a reaction rate $k_{ij} \neq 0$. The network consits of 12 different InsPx and 17 reaction rates.

Density (concentration) vector and corresponding time derivative:

We use the fit functions (SI table 2) to create time-equidistant data points $\boldsymbol{\phi}_0^{\text{exp}}, \boldsymbol{\phi}_\tau^{\text{exp}}, \ldots, \boldsymbol{\phi}_{n\tau}^{\text{exp}}$ and $\dot{\boldsymbol{\phi}}_0^{\text{exp}}, \dot{\boldsymbol{\phi}}_\tau^{\text{exp}}, \ldots, \dot{\boldsymbol{\phi}}_{n\tau}^{\text{exp}}$ (eq. 2.9) with a resolution of $\tau = 1$ min.

Rate matrix:
To build the rate matrix **K**, we number all InsPx included in the network from zero to eleven in a left-to-right and top-to-bottom fashion and assign the corresponding rates according to eq. 2.1. S27,b represents these rates as blue squares. The diagonal elements (green squares) are defined via eq. 2.2 and given as

$$\begin{aligned} k_{00} &= -(k_{10} + k_{20}) \\ k_{11} &= -(k_{31} + k_{41}) \\ k_{22} &= -(k_{42} + k_{52}) \\ k_{33} &= -(k_{63} + k_{73}) \\ k_{44} &= -(k_{74} + k_{84}) \\ k_{55} &= -k_{85} \\ k_{66} &= -k_{96} \\ k_{77} &= -k_{97} \\ k_{88} &= -(k_{98} + k_{108}) \\ k_{99} &= -k_{119} \\ k_{1010} &= -k_{1110}\,. \end{aligned} \tag{3.6}$$

All other matrix elements are equal to zero (white squares).

Set of differential equations:
With the rate matrix **K** we can formulate the corresponding master equation (eq. 2.3) which yields the following set of coupled first-order differential equations

$$\begin{array}{rcccccccccccc} y_0 &=& +k_{00}x_0 & & & & & & & & & & \\ y_1 &=& +k_{10}x_0 & +k_{11}x_1 & & & & & & & & & \\ y_2 &=& +k_{20}x_0 & & +k_{22}x_2 & & & & & & & & \\ y_3 &=& & +k_{31}x_1 & & +k_{33}x_3 & & & & & & & \\ y_4 &=& & +k_{41}x_1 & +k_{42}x_2 & & +k_{44}x_4 & & & & & & \\ y_5 &=& & & +k_{52}x_2 & & & +k_{55}x_5 & & & & & \\ y_6 &=& & & & +k_{63}x_3 & & & +k_{66}x_6 & & & & \\ y_7 &=& & & & +k_{73}x_3 & +k_{74}x_4 & & & +k_{77}x_7 & & & \\ y_8 &=& & & & & +k_{84}x_4 & +k_{85}x_5 & & & +k_{88}x_8 & & \\ y_9 &=& & & & & & & +k_{96}x_6 & +k_{97}x_7 & +k_{98}x_8 & +k_{99}x_9 & \\ y_{10} &=& & & & & & & & +k_{108}x_8 & & +k_{1010}x_{10} & \\ y_{11} &=& & & & & & & & & +k_{119}x_9 & +k_{1110}x_{10} & \end{array} \tag{3.7}$$

Constraints and bounds:
Here, we report the applied constraints that yielded the best results for the reaction rates reported in SI table 5. In total, we constrain 3 reaction rates (k_{10}, k_{20} and k_{52}) which left us with 14 reaction rates that have to be optimized during the minimization routine.

k_{20} and k_{52}:
As mentioned in section 3.3, we used eq. 2.18 as fit function for the progress curve of $InsP_5$[3OH] (SI table 2). Consequently, we can interpret the fit parameter k_1 as reaction rate that describes build-up of concentration and k_2 as reaction rate that describes the decrease of concentration. According to the network in S27a, the increase of $InsP_5$[3OH] concentration is solely determined by k_{20} and thus we constrain $k_{20} = k_1 = 7.4 \cdot 10^{-4}$ min^{-1}. The decrease of $InsP_5$[3OH] concentration is determined by $k_{42} + k_{52}$ and we set the constraint $k_{52} = k_2 - k_{42} = 9.3 \cdot 10^{-4} - k_{42}$.

k_{10}:
We used eq. 2.17 as fit function for the progress curve of $InsP_6$ (SI table 2) and thus the fit parameter k represents the reaction rate that dictates the decrease of $InsP_6$ concentration over time. According to the network in S27,a, this decrease is described by $k_{10} + k_{20}$ and we constrain $k_{10} = k - k_{20} = 1.9 \cdot 10^{-4}$ min^{-1}.

bounds: We set the bounds $k_{108} \in [10^{-3}, 1]$, $k_{119} \in [10^{-5}, 1]$ and $k_{1110} \in [10^{-4}, 1]$ to brute-force increase the influence of these rates on the network and prevent the minimization routine from setting all of them to the lowest possible value 10^{-6}. All other rates were bound to the interval $k_{ij} \in [10^{-6}, 1]$.

Initial guess:
To generate a good initial guess for the unconstrained rates, we started with a reduced network that included $InsP_6$, $InsP_5$[4OH], $InsP_5$[3OH], Ins(1,2,3,6)P_4, Ins(1,2,5,6)P_4 and Ins(1,2,4,5)P_4 and performed a minimization run. Next, we increased the network by including Ins(1,2,3)P_3, Ins(1,2,6)P_3 and Ins(1,2,5)P_3 and repeated the minimization, where we used the results from the previous run for k_{10}, k_{31} and, k_{41} and the default initial guess for the remaining rates. Finally, we repeated this step with the full network which yielded a good initial guess for all 13 unconstrained rates as shown in SI table 4.

Technical details:
For all technical details, the reader is referred to section 3.4.

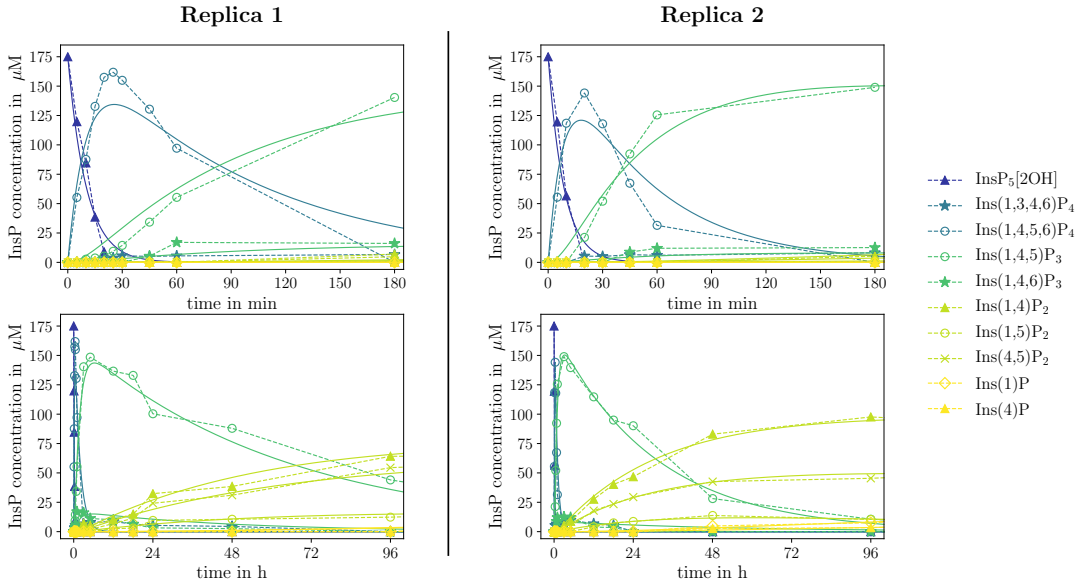
SI Table 4:
Initial guess (in min^{-1}) and bounds for all rates in the $[^{13}C_6]InsP_6$ dephosphorylation network.
*) The marked rates are subject to constraints.

rate	initial guess	bounds
k_{10}	$1.9 \cdot 10^{-4*}$	-
k_{20}	$7.4 \cdot 10^{-4*}$	-
k_{31}	$2.9 \cdot 10^{-3}$	$[10^{-6}, 1]$
k_{41}	$1.0 \cdot 10^{-5}$	$[10^{-6}, 1]$
k_{42}	$3.0 \cdot 10^{-4}$	$[10^{-6}, 1]$
k_{52}	$9.3 \cdot 10^{-4*} - k_{42}$	-
k_{63}	$5.0 \cdot 10^{-4}$	$[10^{-6}, 1]$
k_{73}	$3.7 \cdot 10^{-4}$	$[10^{-6}, 1]$
k_{74}	$1.4 \cdot 10^{-3}$	$[10^{-6}, 1]$
k_{84}	$1.5 \cdot 10^{-4}$	$[10^{-6}, 1]$
k_{85}	$4.2 \cdot 10^{-4}$	$[10^{-6}, 1]$
k_{96}	$2.0 \cdot 10^{-4}$	$[10^{-6}, 1]$
k_{97}	$2.0 \cdot 10^{-4}$	$[10^{-6}, 1]$
k_{98}	$3.9 \cdot 10^{-3}$	$[10^{-6}, 1]$
k_{108}	$1.0 \cdot 10^{-5}$	$[10^{-3}, 1]$
k_{119}	$1.0 \cdot 10^{-3}$	$[10^{-5}, 1]$
k_{1110}	$1.0 \cdot 10^{-3}$	$[10^{-4}, 1]$

4 Results

4.1 Results $InsP_5$[2OH] dephosphorylation

The numerically determined set of rates for both replicas are presented in S29. We can see that replica 2 exhibits slightly faster kinetics than replica although both share an identical experimental setup. Excluding the rates k_{95} and k_{97}, both sets of rates are in good agreement (S29, b). The fastest process is described by $k_{20} \approx 10^{-1}$ min^{-1} which governs the reaction $InsP_5$[2OH] $\rightarrow$ Ins(1,4,5,6)P_4. This result is in good agreement with MINPP1's annotation as a phosphatase that predominantly removes the phosphoryl group at the 3-position.[6] The reaction rates of the subsequent dephosphorylation steps are separated by at least one order of magnitude, where we get $k_{ij} \approx 10^{-2}$ min^{-1} for reactions of the type $InsP_4 \rightarrow InsP_3$, $k_{ij} \approx 10^{-4}$ min^{-1} for reactions of the type $InsP_3 \rightarrow InsP_2$ and $k_{ij} \approx 10^{-5}$ min^{-1} for reactions of the type $InsP_2 \rightarrow InsP_1$.
In S28, we compare the scaled experimental data (dotted lines) to the progress curves (solid lines) predicted from the numerically determined set of rates (eq. 2.7) for each replica. The predicted progress curves match the experimental data both qualitatively and quantitatively which strongly supports the assumption that the reaction rates in the $InsP_5$[2OH] dephosphorylation network are time independent. Furthermore, the results

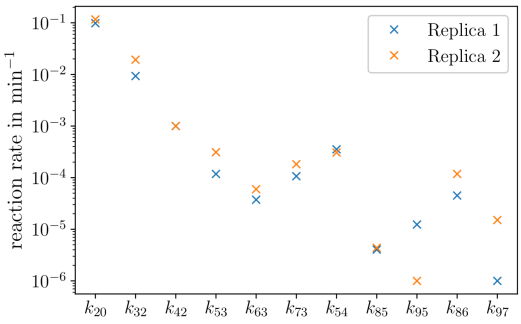


S 28:
Predicted progress curves (solid lines) obtained via minimization routine and scaled experimental data (dashed lines) for two replicas (columns) of MINPP1 reaction with 175 μM [$^{13}C_6$]InsP$_5$[2OH], where the top row magnifies the first 180 min and the bottom row the entire time axis of the experiment. The results for replica 1 are a copy of the results shown in main part Fig. 6a.

(a)

rate	Replica 1	Replica 2
k_{10}	0.00*	0.00*
k_{20}	$9.8 \cdot 10^{-2}$	$1.1 \cdot 10^{-1}$
k_{41}	0.00*	0.00*
k_{32}	$9.3 \cdot 10^{-3}$	$1.9 \cdot 10^{-2}$
k_{42}	$1.0 \cdot 10^{-3}$*	$1.0 \cdot 10^{-3}$*
k_{53}	$1.1 \cdot 10^{-4}$	$3.1 \cdot 10^{-4}$
k_{63}	$3.7 \cdot 10^{-5}$	$5.9 \cdot 10^{-5}$
k_{73}	$1.0 \cdot 10^{-4}$*	$1.8 \cdot 10^{-4}$*
k_{54}	$3.5 \cdot 10^{-4}$	$3.0 \cdot 10^{-4}$
k_{85}	$4.0 \cdot 10^{-6}$	$4.3 \cdot 10^{-6}$
k_{95}	$1.2 \cdot 10^{-5}$	$1.0 \cdot 10^{-6}$
k_{86}	$4.5 \cdot 10^{-5}$	$1.1 \cdot 10^{-4}$
k_{97}	$1.0 \cdot 10^{-6}$	$1.5 \cdot 10^{-5}$

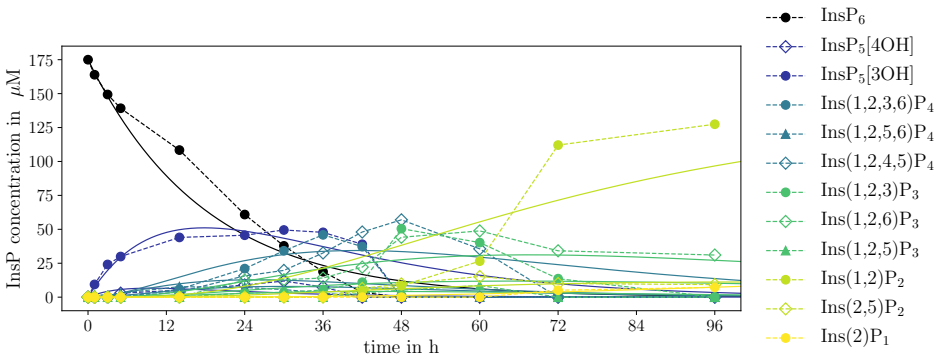
(b)



S 29:
(a) Computed rates in min^{-1} for both replicas of the [$^{13}C_6$]InsP$_5$[2OH] dephosphorylation. The column for replica 1 is a copy of the results presented in main part Fig. 6b. *) The marked rates are subject to constraints.
(b) Visual comparison of the rates computed from the scaled experimental data of replica 1 and replica 2, respectively. The rates $k_{10}, k_{41} = 0$ are not included in the representation.

confirm that the network depicted in S26 accurately describes the MINPP1-mediated dephosphorylation of InsP$_5$[2OH].

4.2 Results $InsP_6$ dephosphorylation



S 30:
Predicted progress curves (solid lines) obtained via minimization routine and scaled experimental data (dashed lines) of MINPP1 reaction with 175 μM [$^{13}C_6$]$InsP_6$.

SI Table 5:
Computed reaction rates in min^{-1} for [$^{13}C_6$]$InsP_6$ dephosphorylation network.
*) The marked rates were subject to constraints.

rate	reaction rate
k_{10}	$1.90 \cdot 10^{-4*}$
k_{20}	$7.42 \cdot 10^{-4*}$
k_{31}	$2.95 \cdot 10^{-3}$
k_{41}	$1.00 \cdot 10^{-6}$
k_{42}	$2.88 \cdot 10^{-4}$
k_{52}	$6.49 \cdot 10^{-4*}$
k_{63}	$5.10 \cdot 10^{-4}$
k_{73}	$3.67 \cdot 10^{-4}$
k_{74}	$1.56 \cdot 10^{-3}$
k_{84}	$2.13 \cdot 10^{-5}$
k_{85}	$5.99 \cdot 10^{-4}$
k_{96}	$2.02 \cdot 10^{-4}$
k_{97}	$2.22 \cdot 10^{-4}$
k_{98}	$3.88 \cdot 10^{-3}$
k_{108}	$1.00 \cdot 10^{-3}$
k_{119}	$1.00 \cdot 10^{-5}$
k_{1110}	$1.49 \cdot 10^{-4}$

S30 shows the comparison between the scaled experimental data (dashed lines) and the progress curves predicted by the numerically determined rates (solid lines). We can clearly see that the computed rates yield a poor representation of the experimental progress curves which strongly indicates that the applied time-constant rates model is insufficient

to describe the $InsP_6$ dephosphorylation. The shapes of the experimental progress curves already indicate a kinetic network with time-dependent rates, e.g. the $InsP_6$ progress curve does not represent an exponential decay as we would expect from a first-order reaction. Instead, we observe a damped decrease which could emerge from a inhibition process. As mentioned in main part (Fig. 6c) we suggest that $InsP_6$ itself could act as an inhibitor for the dephosphorylation of its own MINPP1-mediated intermediates.[7] This assumption is further supported by the fact, that the kinetics clearly accelerate as soon as $InsP_6$ is fully depleted. However, we can roughly approximate the rate at which $InsP_6$ is depleted as $k_{10} + k_{20} = 9.3 \cdot 10^{-4}$ min^{-1}. Based on our results and the discussion above, we conclude that our master equation ansatz (SI section 2.5) is not capable to capture the true kinetics for the MINPP1-mediated dephosphorylation of $InsP_6$ and thus does not provide any more insight into the main pathways that generate the enantiomers Ins(1,2)P_2 and Ins(2,3)P_2. For the sake of completeness, we report the numerically determined rates in SI table 5 but did not include these results in the main part of our work. We renounce to extend our ansatz in order to include inhibition processes but this kind of analysis is beyond the scope of this paper.

References

[1] V. (Ed) Patel, *Chemical Kinetics*, IntechOpen, London, **2012**.

[2] C. W. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*, 2nd, Springer Verlag, Berlin Heidelberg, **1983**.

[3] N. G. van Kampen, *Stochastic Processes in Physics and Chemistry*, North-Holland Publishing Company Amsterdam, New York, Oxford, **1981**.

[4] D. B. Ball, *J. Chem. Educ.* **1998**, *75*, 917–918.

[5] Manual scipy.optimize.minimize, `https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.minimize.html`, **accessed: 05.07.2022**.

[6] A. Craxton, J. Caffrey, B. W., T. Safrany, B. Shears, *Biochem J.* **1997**, *328*, 75–81.

[7] K. Nogimori, P. Hughes, M. Glennon, M. Hodgson, J. Putney, S. Shears, *J. Biol. Chem.* **1991**, *266*, 16499–16506.

4 Conclusion

This thesis presented different methods to compute dynamical quantities. Part **A** focused on underdamped Langevin integrators for MD simulations and path reweighting methods to reweight biased dynamics in order to build the unbiased MSM. We reported that the underdamped Langevin integrators BAOA and GSD are equivalent sampling algorithms and thus produce configurational and kinetic properties with identical accuracy. Furthermore, we argued that BAOA/GSD and BAOAB are equivalent configurational sampling algorithms. These results are of particular relevance for the MD community because BAOA, BAOAB and GSD belong to the most widely used Langevin integrators. BAOA (also called LFMiddle[37]) is implemented in the MD packages OpenMM[114] and AMBER[118], GSD is the standard Langevin integrator in GROMACS[119] and BAOAB is frequently used in atomistic MD simulations (available in OpenMM via the toolkit OpenMMTools[113]). In other words, all three MD packages generate trajectories with equivalent configurational sampling accuracy and we can straightforwardly transfer benchmark studies of one of the integrators to the other two. Our studies also indicated that BAOA/GSD sample the marginal distribution of momenta more accurately than BAOAB. This observation is surprising because BAOAB is a symmetric Langevin integrator[32] and thus would be expected to yield a more accurate momenta distribution than unsymmetrical Langevin integrators like BAOA/GSD.[108] Additionally, it could be possible that the accuracy of the marginal distributions might not extend to correlations between positions and momenta. Further studies are required to confirm or falsify this assumption and to provide further explanation as to why unsymmetrical Langevin integrators yield excellent uncorrelated momenta.

MD trajectories that have been generated with stochastic integrators like Langevin integrators can be reweighted by dynamical reweighting methods. This thesis presented an overview of state-of-the-art potential reweighting techniques which can be used to recover the unbiased dynamics from simulations that have been conducted at a biased potential. One of these methods is path reweighting which is based on calculating the weights that the simulated paths would have in the unbiased potential. The weight is composed of the state space reweighting factor g and the path space reweighting factor M. So far, we used the path reweighting factor M_{approx} which is based on the expression of the random number difference $\Delta\eta$ derived for overdamped Langevin dynamics even though the MD simulation was conducted with an underdamped Langevin integrator. In this thesis, we derived M_L for a simplified version of an underdamped Langevin integrator that we called ISP.[39] With M_L we performed exact reweighting for underdamped Langevin dynamics for the first time. Additionally, we were able to explain why M_{approx} yielded excellent results for underdamped Langevin dynamics and proved mathematically that the approximation $M_L \approx M_{\text{approx}}$ is $\mathcal{O}(\xi^4\Delta t^4)$ order accurate. Here, ξ represents the collision rate and Δt the simulation time step.

Since $\xi\Delta t < 1$ usually holds in MD simulations, we can readily apply M_{approx} to reweight trajectories generated by the ISP scheme. In this context, we strongly assume that $M_L \approx M_{\text{approx}}$ also holds for other underdamped Langevin integrators. From the results mentioned above, we already know that BAOA and GSD have an identical path reweighting factor which we expect to be similar if not identical to the expression for BAOAB. Furthermore, we presented two strategies that can be used as blueprints to derive the path reweighting factor for other underdamped Langevin integrators. The strategy that makes use of the random number difference $\Delta\eta$ is particularly promising in order to simply and quickly derive M which we demonstrated with the ABOBA integrator. In a future work, this strategy could be used to derive M_L for other integrators. The expressions could then be used to address the previously mentioned assumption that $M_L \approx M_{\text{approx}}$ also holds for other underdamped Langevin integrators.
In general, path reweighting is a powerful tool that can be combined with methods like enhanced sampling techniques[56–60] and Markov State Models[76–81] (MSMs) to study the dynamics of a given system.[74,75] In this context, reweighting is particularly useful when the uncertainty of transition counts estimated from a direct simulation is larger than the uncertainty of the reweighted transition counts. This is the case if a direct simulation does not have enough energy to cross large energy barriers a statistically meaningful amount of times. Combining the MD simulation with an enhanced sampling technique increases the amount of barrier crossings significantly and we can reweight the unbiased transition counts. If the path reweighting factor matches the chosen integrator and if absolute continuity in configuration as well as in path space is fulfilled, the reweighting does not introduce any additional uncertainty. In future work, we could take maximum advantage of this fact and combine transition path sampling[159–163] (TPS) methods with metadynamics and path reweighting. Let's assume the system of interest comes with a large energy barrier between two metastable states along a relevant coordinate, e.g. the dissociation of an an ion pair. With TPS we could start several independent trajectories in one of the metastable states and use metadynamics to push the system over the energy barrier into the second state. We could then use path reweighting to compute a count matrix per trajectory and sum them up to an overall count matrix from which we then compute the MSM. This combination might yield a new powerful technique to study rare events like dissociation processes.

Another field of application where path reweighting can be valuable is force field optimization. In this context, Ref. [75] already studied how the Coulomb potential and the corresponding electric constant influences the dynamics of the system using path reweighting. To date, force fields have been optimized to reproduce experimentally measured structural and thermodynamic properties. However, dynamical properties such as rates of interconversion between metastable states were not subject to the optimization process so far, although experimental knowledge about these rates exists. Ref. [164] published a method that can be used to improve force field optimization in this regard. The authors combined path reweighting with a maximum caliber approach to impose a dynamical constraint in a complex molecular transition.

The method optimizes the force field parameters with respect to the dynamical constraint while keeping the prior trajectory as unperturbed as possible. So far, all path reweighting methods are formulated for the Euler-Maruyama (EM) integrator for overdamped Langevin dynamics and work with the corresponding path reweighting factor. Since the EM integrator is not a good choice to describe the dynamics of a molecular system, it is crucial to derive M for underdamped Langevin integrators. Additionally, the confirmation of the previously made assumption that $M_L \approx M_{\text{approx}}$ would make the path reweighting factor independent of the chosen integrator and would consequently facilitate the handling of path reweighting methods significantly. This would facilitate the handling of path reweighting methods significantly and methods that already work with M_{approx} would not need to be adapted.

Part **B** investigated the MINPP1 mediated dephosphorylation pathways of $InsP_5$[2OH] and $InsP_6$ with ^{13}C-labeling experiments combined with BIRD-{^{1}H-^{13}C}HMQC-NMR measurements and extracted the corresponding reactions rates via a kinetic scheme model.[83,84] We assumed a Markovian kinetic scheme, meaning the reaction rates are time independent. This assumption yielded excellent reaction rates for the $InsP_5$[2OH] dephosphorylation pathway but only poor results in the case of $InsP_6$. We assumed that $InsP_6$ could act as an inhibitor for the dephosphorylation of the MINPP1-generated intermediates. Our experiments confirmed this assumption and we observed a clear inhibitory effect of $InsP_6$ on the MINPP1 mediated dephosphorylation of $InsP_5$[2OH]. In future work, we could incorporate this inhibitory effect into the kinetic scheme via competitive inhibition. Competitive inhibition is a model that modifies Michaelis-Menten kinetics to include the binding of one or more inhibitors to the free enzyme. With this model, we could describe each reaction rate in the dephosphorylation pathway with respect to the concentrations of the other InsP intermediates and thus include an implicit time dependence. From an experimental point of view, the next step could be to use our experimental and analytical setup to investigate the InsP metabolism in other biological contexts. For example, the role of inositol (phosphates) in pathogenic parasites such as *T. cruzi* could be investigated. It is known that the InsP metabolism is essential for the development cycle of the parasite.[165] An understanding of the respective metabolism in the host and in the parasite and how they influence each other might help in the development of new therapies for these parasitoses. Another question that could be addressed is related to MINPP1's upregulation during endoplasmic reticulum-related stress. It has yet to be explored how this affects the InsP pool and how everything might be correlated with the onset of apoptosis.[150]
Finally, we want to mention that the MINPP1 mediated dephosphorylation of $InsP_5$[2OH] and $InsP_6$ are a biochemical dynamics which could be modeled as reaction-diffusion processes. Consequently, the dephosphorylation networks could be an excellent application for methods that are based on the chemical diffusion master equation (CDME).[166,167]

A Appendix

A.1 From stochastic differential equations to Fokker-Planck equations: Integration by parts

This part of the appendix shows the integration by parts that was performed to get from eq. 2.82 to eq. 2.83. The integral on the left hand side of eq. 2.82 can be computed as

$$\begin{aligned}&\int \mathrm{d}x \, \frac{\partial f(x,t)}{\partial x}\Big(A(x,t)P(x,t|x_0,t_0)\Big)\\&= f(x,t)A(x,t)P(x,t|x_0,t_0)\Big|_{-\infty}^{+\infty} - \int \mathrm{d}x f(x,t)\frac{\partial}{\partial x}A(x,t)P(x,t|x_0,t_0)\,.\end{aligned} \tag{A.1}$$

The integral on the right hand side of eq. 2.82 evaluates to

$$\begin{aligned}&\frac{\sigma^2}{2}\int \mathrm{d}x \, \frac{\partial^2 f(x,t)}{\partial x^2}P(x,t|x_0,t_0)\\&= \frac{\sigma^2}{2}\left[\frac{\partial f(x,t)}{\partial x}P(x,t|x_0,t_0)\Big|_{-\infty}^{+\infty} - \int \mathrm{d}x \frac{\partial f(x,t)}{\partial x}\frac{\partial P(x,t|x_0,t_0)}{\partial x}\right]\\&= \frac{\sigma^2}{2}\left[\frac{\partial f(x,t)}{\partial x}P(x,t|x_0,t_0)\Big|_{-\infty}^{+\infty} - f(x,t)\frac{\partial P(x,t|x_0,t_0)}{\partial x}\Big|_{-\infty}^{+\infty} + \int \mathrm{d}x f(x,t)\frac{\partial^2 P(x,t|x_0,t_0)}{\partial x^2}\right].\end{aligned} \tag{A.2}$$

All surface terms in eqs. A.1 and A.2 are equal to zero because we defined the boundary conditions in eq. 2.79. Discarding the surface terms and adding eqs. A.1 and A.2 yields

$$-\int \mathrm{d}x f(x,t)\frac{\partial}{\partial x}A(x,t)P(x,t|x_0,t_0) + \frac{\sigma^2}{2}\int \mathrm{d}x f(x,t)\frac{\partial^2 P(x,t|x_0,t_0)}{\partial x^2}\,, \tag{A.3}$$

which represents the right hand side of eq. 2.83.

A.2 Integrators for underdamped Langevin dynamics

This part of the appendix summarizes the integrator equations for the underdamped Langevin integrators. Δt is the time step, m is the mass, q_k the position, p_k the momentum and η_k the random number at iteration step k, ξ the collision rate, $V(q_k)$ the potential energy function and ∇_q the gradient and η_k the random number.

The BAOAB integrator

$$
\begin{aligned}
p_{k+1/3} &= p_k - \frac{\Delta t}{2}\nabla_q V(q_k) && \text{(A.4a)}\\
q_{k+1/2} &= q_k + \frac{\Delta t}{2m}\, p_{k+1/3} && \text{(A.4b)}\\
p_{k+2/3} &= e^{-\xi\Delta t}\, p_{k+1/3} + \sqrt{k_B T m(1-e^{-2\xi\Delta t})}\,\eta_k && \text{(A.4c)}\\
q_{k+1} &= q_{k+1/2} + \frac{\Delta t}{2m}\, p_{k+2/3} && \text{(A.4d)}\\
p_{k+1} &= p_{k+2/3} - \frac{\Delta t}{2}\nabla_q V(q_{k+1}) && \text{(A.4e)}
\end{aligned}
$$

The BAOA integrator

$$
\begin{aligned}
p_k &= p_{k-\frac{1}{2}} - \Delta t \nabla_q V(q_k) && \text{(A.5a)}\\
q_{k+\frac{1}{2}} &= q_k + \frac{\Delta t}{2m} p_k && \text{(A.5b)}\\
p_{k+\frac{1}{2}} &= e^{-\xi\Delta t} p_k + \sqrt{k_B T m\,(1-e^{-2\xi\Delta t})}\eta_k && \text{(A.5c)}\\
q_{k+1} &= q_{k+\frac{1}{2}} + \frac{\Delta t}{2m} p_{k+\frac{1}{2}} && \text{(A.5d)}
\end{aligned}
$$

The AOBOA integrator

$$
\begin{aligned}
q_{k+1/2} &= q_k + \frac{\Delta t}{2m} p_k && \text{(A.6a)}\\
p_{k+1/3} &= e^{-\frac{\xi\Delta t}{2}} p_k + \sqrt{k_B T m(1-e^{-\xi\Delta t})}\,\eta_k^{(1)} && \text{(A.6b)}\\
p_{k+2/3} &= p_{k+1/3} - \Delta t \nabla_q V(q_{k+1/2}) && \text{(A.6c)}\\
p_{k+1} &= e^{-\frac{\xi\Delta t}{2}} p_{k+2/3} + \sqrt{k_B T m(1-e^{-\xi\Delta t})}\,\eta_k^{(2)} && \text{(A.6d)}\\
q_{k+1} &= q_{k+1/2} + \frac{\Delta t}{2m}\, p_{k+1} && \text{(A.6e)}
\end{aligned}
$$

The BOAOB integrator

$$
\begin{aligned}
p_{k+1/4} &= p_k - \frac{\Delta t}{2}\nabla_q V(q_k) && \text{(A.7a)}\\
p_{k+2/4} &= e^{-\frac{\xi\Delta t}{2}}\, p_{k+1/4} + \sqrt{k_B T m\,(1-e^{-\xi\Delta t})}\,\eta_k^{(1)} && \text{(A.7b)}\\
q_{k+1} &= q_k + \frac{\Delta t}{m} p_{k+2/4} && \text{(A.7c)}\\
p_{k+3/4} &= e^{-\frac{\xi\Delta t}{2}}\, p_{k+2/4} + \sqrt{k_B T m\,(1-e^{-\xi\Delta t})}\,\eta_k^{(2)} && \text{(A.7d)}\\
p_{k+1} &= p_{k+3/4} - \frac{\Delta t}{2}\nabla_q V(q_{k+1})\,. && \text{(A.7e)}
\end{aligned}
$$

The OBABO/Bussi-Parrinello integrator

$$
\begin{aligned}
p_{k+1/4} &= e^{-\frac{\xi\Delta t}{2}}\, p_k + \sqrt{k_B T m\,(1-e^{-\xi\Delta t})}\,\eta_k^{(1)} && \text{(A.8a)}\\
p_{k+2/4} &= p_{k+1/4} - \frac{\Delta t}{2}\nabla_q V(q_k) && \text{(A.8b)}\\
q_{k+1} &= q_k + \frac{\Delta t}{m} p_{k+2/4} && \text{(A.8c)}\\
p_{k+3/4} &= p_{k+2/4} - \frac{\Delta t}{2}\nabla_q V(q_{k+1}) && \text{(A.8d)}\\
p_{k+1} &= e^{-\frac{\xi\Delta t}{2}}\, p_{k+3/4} + \sqrt{k_B T m\,(1-e^{-\xi\Delta t})}\,\eta_k^{(2)} && \text{(A.8e)}
\end{aligned}
$$

The OABAO integrator

$$
\begin{aligned}
p_{k+1/3} &= e^{-\frac{\xi\Delta t}{2}}\, p_k + \sqrt{k_B T m(1-e^{-\xi\Delta t})}\,\eta_k^{(1)} && \text{(A.9a)}\\
q_{k+1/2} &= q_k + \frac{\Delta t}{2m}\, p_{k+1/3} && \text{(A.9b)}\\
p_{k+2/3} &= p_{k+1/3} - \Delta t\nabla_q V(q_{k+1/2}) && \text{(A.9c)}\\
q_{k+1} &= q_{k+1/2} + \frac{\Delta t}{2m}\, p_{k+2/3} && \text{(A.9d)}\\
p_{k+1} &= e^{-\frac{\xi\Delta t}{2}}\, p_{k+2/3} + \sqrt{k_B T m(1-e^{-\xi\Delta t})}\,\eta_k^{(2)} && \text{(A.9e)}
\end{aligned}
$$

References

[1] A. Bouvier, M. Wadhwa, "The age of the Solar System redefined by the oldest Pb–Pb age of a meteoritic inclusion", *Nat. Geosci.* **2010**, *3*, 637.

[2] G. Zhao, M. Sun, S. A. Wilde, S. Li, "A Paleo-Mesoproterozoic supercontinent: assembly, growth and breakup", *Earth-Sci. Rev.* **2004**, *67*, 91.

[3] H. Tavakolifar, E. Shahghasemi, S. Nazif, "Evaluation of climate change impacts on extreme rainfall events characteristics using a synoptic weather typing-based daily precipitation downscaling model", *J. Water Clim. Chang.* **2017**, *8*, 388.

[4] S. Olsson, D. Ekonomiuk, J. Sgrignani, A. Cavalli, "Molecular Dynamics of Biomolecules through Direct Analysis of Dipolar Couplings", *J. Am. Chem. Soc.* **2015**, *137*, 6270.

[5] G. R. Bowman, V. A. Voelz, V. S. Pande, "Taming the complexity of protein folding", *Curr. Opin. Struct. Biol.* **2011**, *21*, 4.

[6] J. R. Lewandowski, M. E. Halse, M. Blackledge, L. Emsley, "Direct observation of hierarchical protein dynamics", *Science* **2015**, *238*, 578.

[7] V. S. Shaw, H. Mohammadiarani, H. Vashisth, R. R. Neubig, "Differential protein dynamics of regulators of g-protein signaling: Role in specificity of small-molecule inhibitors", *J. Am. Chem. Soc.* **2018**, *140*, 3454.

[8] M. Hesse, H. Meier, B. Zeeh, *Spectroscopic Methods in Organic Chemistry*, Thieme/Houben-Weyl Series, **2008**.

[9] F. Siebert, P. Hildebrandt, *Vibrational Spectroscopy in Life Science*, WILEY-VCH Verlag GmbH & Co. KGaA Weinheim, **2008**.

[10] J. Kubelka, T. K. Chiu, D. R. Davies, W. A. Eaton, J. Hofrichter, "Sub-microsecond protein folding", *J. Mol. Biol.* **2006**, *359*, 546.

[11] R. D. Schaeffer, A. Fersht, V. Daggett, "Sub-microsecond protein folding", *Curr. Opin. Struct. Biol.* **2008**, *18*, 4.

[12] P. L. Freddolino, K. Schulten, "Common structural transitions in explicit-solvent simulations of villin headpiece folding", *Biophys. J.* **2009**, *97*, 2338.

[13] M. K. Gilson, H. X. Zhou, "Calculation of protein-ligand binding affinities", *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21.

[14] E. Lindahl, M. S. Sansom, "Membrane proteins: molecular dynamics simulations", *Curr. Opin. Struct. Biol.* **2008**, *18*, 425.

[15] F. Khalili-Araghi, J. Gumbart, P. C. Wen, M. Sotomayor, E. Tajkhorshid, K. Schulten, "Molecular dynamics simulations of membrane channels and transporters", *Curr. Opin. Struct. Biol.* **2009**, *19*, 128.

[16] T. R. Sosnick, L. Mayne, R. Hiller, S. W. Englander, "The barriers in protein folding", *Nat. Struct. Biol.* **1994**, *1*, 149.

[17] M. Karplus, J. Kuriyan, "Molecular dynamics and protein function", *J. Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6679.

[18] A. R. Leach, *Molecular Modelling: Principles and Applications*, 2nd ed., Pearson Education Limited, **2001**.

[19] B. Leimkuhler, C. Matthews, *Molecular Dynamics With Deterministic and Stochastic Numerical Methods*, Springer Cham Heidelberg New York Dordrecht London, **2015**.

[20] D. Frenkel, B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*, 2nd ed., ACADEMIC PRESS, **2002**.

[21] E. P. Barros, L. Casalino, Z. Gaieb, A. C. Dommer, Y. Wang, L. Fallon, L. Raguette, K. Belfon, C. Simmerling, R. E. Amaro, "The flexibility of ACE2 in the context of SARS-CoV-2 infection", *Biophys. J.* **2020**, *120*, 1072.

[22] T. J. Harpole, L. Delemotte, "Conformational landscapes of membrane proteins delineated by enhanced sampling molecular dynamics simulations", *Biochim. Biophys. Acta Biomembr.* **2018**, *1860*, 909.

[23] Z. Cournia, B. Allen, W. Sherman, "Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations", *J. Chem. Inf. Model.* **2017**, *57*, 2911.

[24] M. Badaoui, A. Kells, C. Molteni, C. J. Dickson, V. Hornak, E. Rosta, "Calculating Kinetic Rates and Membrane Permeability from Biased Simulations", *J. Phys. Chem. B* **2018**, *122*, 11571.

[25] J.-O. Joswig, J. Anders, H. Zhang, C. Rademacher, B. G. Keller, "The molecular basis for the pH-dependent calcium affinity of the pattern recognition receptor langerin", *J. Biol. Chem.* **2021**, *296*, 100718.

[26] D. S. Lemons, A. Gythiel, "Paul Langevin's 1908 paper "On the Theory of Brownian Motion" ["Sur la théorie du mouvement brownien," C. R. Acad. Sci. (Paris) 146, 530–533 (1908)]", *Am. J. Phys.* **1997**, *65*, 1079.

[27] P. H. Hünenberger, *"Thermostat Algorithms for Molecular Dynamics Simulations" in Advanced Computer Simulation: Approaches for Soft Matter Sciences I*, Springer Berlin Heidelberg, **2005**.

[28] P. E. Kloeden, E. Platen, *Numerical Solution of Stochastic Differential Equations*, Springer, Berlin, **1992**.

[29] B. Øksendal, *Stochastic Differential Equations: An Introduction with Applications*, 6th, Springer Verlag, Berlin, **2003**.

[30] B. Leimkuhler, C. Matthews, "Rational Construction of Stochastic Numerical Methods for Molecular Sampling", *Appl. Math. Res. eXpress* **2013**, *2013*, 34.

[31] B. Leimkuhler, C. Matthews, "Robust and efficient configurational molecular sampling via Langevin dynamics", *J. Chem. Phys.* **2013**, *138*, 174102.

[32] B. Leimkuhler, C. Matthews, "Efficient molecular dynamics using geodesic integration and solvent–solute splitting", *Proc. R. Soc. A: Math. Phys. Eng. Sci.* **2016**, *472*, 20160138.

[33] D. A. Sivak, J. D. Chodera, G. E. Crooks, "Time step rescaling recovers continuous-time dynamical properties for discrete-time Langevin integration of nonequilibrium systems", *J. Phys. Chem. B* **2014**, *118*, 6466.

[34] G. Bussi, M. Parrinello, "Accurate sampling using Langevin dynamics", *Phys. Rev. E* **2007**, *75*, 056707.

[35] N. Goga, A. J. Rzepiela, A. H. De Vries, S. J. Marrink, H. J. C. Berendsen, "Efficient algorithms for Langevin and DPD dynamics", *J. Chem. Theory Comput.* **2012**, *8*, 3637.

[36] N. Bou-Rabee, H. Owhadi, "Long-run accuracy of variational integrators in the stochastic context", *SIAM J. Numer. Anal.* **2010**, *48*, 278.

[37] Z. Zhang, X. Liu, K. Yan, M. E. Tuckerman, J. Liu, "Unified Efficient Thermostat Scheme for the Canonical Ensemble with Holonomic or Isokinetic Constraints via Molecular Dynamics", *J. Phys. Chem. A* **2019**, *123*, 6056.

[38] R. D. Skeel, J. A. Izaguirre, "An impulse integrator for Langevin dynamics", *Mol. Phys.* **2002**, *100*, 3885.

[39] J. A. Izaguirre, C. R. Schweet, V. S. Pande, "Multiscale Dynamics of Macromolecules using Normal Mode Langevin", *Pacific Symposium on Biocomputing* **2010**, *15*, 240.

[40] N. Grønbech-Jensen, O. Farago, "A simple and effective Verlet-type algorithm for simulating Langevin dynamics", *Mol. Phys.* **2013**, *111*, 983.

[41] L. F. G. Jensen, N. Grønbech-Jensen, "Accurate configurational and kinetic statistics in discrete-time Langevin systems", *Mol. Phys.* **2019**, *117*, 2511.

[42] A. Brünger, C. L. Brooks III, M. Karplus, "Stochastic boundary conditions for molecular dynamics simulations of ST2 water", *Chem. Phys. Lett.* **1984**, *105*, 495.

[43] E. Hershkovitz, "A fourth-order numerical integrator for stochastic Langevin equations", *J. Chem. Phys.* **1998**, *108*, 9253.

[44] E. Vanden-Eijnden, G. Ciccotti, "Second-order integrators for Langevin equations with holonomic constraints", *Chem. Phys. Lett.* **2006**, *429*, 310.

[45] M. G. Paterlini, D. M. Ferguson, "Constant temperature simulations using the Langevin equation with velocity Verlet integration", *Chem. Phys.* **1998**, *236*, 243.

[46] J. Finkelstein, C. Cheng, G. Fiorin, B. Seibold, N. Grønbech-Jensen, "Bringing discrete-time Langevin splitting methods into agreement with thermodynamics", *J. Chem. Phys.* **2021**, *155*, 184104.

[47] J. Fass, D. A. Sivak, G. E. Crooks, K. A. Beauchamp, B. Leimkuhler, J. D. Chodera, "Quantifying Configuration-Sampling Error in Langevin Simulations of Complex Molecular Systems", *Entropy* **2018**, *20*, 318.

[48] Matthews, C., Dissertation: "The error in the invariant measure of numerical discretization schemes for canonical sampling of molecular dynamics", *The University of Edinburgh*, **2013**, `https://era.ed.ac.uk/handle/1842/8949`.

[49] D. Li, X. Han, Y. Chai, C. Wang, Z. Zhang, Z. Chen, J. Liu, J. Shao, "Stationary state distribution and efficiency analysis of the Langevin equation via real or virtual dynamics", *J. Chem. Phys.* **2017**, *147*, 184104.

[50] Z. Song, Z. Tan, "On Irreversible Metropolis Sampling Related to Langevin Dynamics", *SIAM J. Sci. Comput.* **2022**, *44*, A2089.

[51] E. Marinari, G. Parisi, "Simulated Tempering: A New Monte Carlo Scheme", *Europhys. Lett.* **1992**, *19*, 451.

[52] A. P. Lyubartsev, A. A. Martsinovski, S. V. Shevkunov, P. Vorontsov-Velyaminov, "New approach to Monte Carlo calculation of the free energy: Method of expanded ensembles", *J. Chem. Phys.* **1992**, *96*, 1776.

[53] A. Mitsutake, Y. Okamoto, "Replica-exchange simulated tempering method for simulations of frustrated systems", *Chem. Phys. Lett.* **2000**, *332*, 131.

[54] C. J. Geyer, E. A. Thompson, "Annealing Markov Chain Monte Carlo with Applications to Ancestral Inference", *J. Am. Stat. Assoc.* **1995**, *90*, 909.

[55] U. H. E. Hansmann, "Parallel tempering algorithm for conformational studies of biological molecules", *Chem. Phys. Lett.* **1997**, *281*, 1604.

[56] G. Torrie, J. Valleau, "Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling", *J. Comput. Phys.* **1977**, *23*, 187.

[57] J. Kästner, W. Thiel, "Bridging the gap between thermodynamic integration and umbrella sampling provides a novel analysis method: "Umbrella integration"", *J. Chem. Phys.* **2005**, *123*, 144104.

[58] T. Huber, A. Torda, W. van Gunsteren, "Local elevation: A method for improving the searching properties of molecular dynamics simulation", *J. Comput. Aided Mol. Des.* **1994**, *8*, 695.

[59] A. Laio, M. Parrinello, "Escaping free-energy minima", *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 12562.

[60] A. Barducci, G. Bussi, M. Parrinello, "Well-tempered metadynamics: A smoothly converging and tunable free-energy method", *Phys. Rev. Lett.* **2008**, *100*, 020603.

[61] J. D. Chodera, W. C. Swope, F. Noé, J.-H. Prinz, M. R. Shirts, V. S. Pande, "Dynamical reweighting: Improved estimates of dynamical properties from simulations at multiple temperatures", *J. Chem. Phys.* **2011**, *134*, 244107.

[62] J.-H. Prinz, J. D. Chodera, V. S. Pande, W. C. Swope, J. C. Smith, F. Noé, "Optimal use of data in parallel tempering simulations for the construction of discrete-state Markov models of biomolecular dynamics", *J. Chem. Phys.* **2011**, *134*, 244108.

[63] L. S. Stelzl, G. Hummer, "Kinetics from Replica Exchange Molecular Dynamics Simulations", *J. Chem. Theory Comput.* **2017**, *13*, 3927.

[64] L. Onsager, S. Machlup, "Fluctuations and Irreversible Processes", *Phys. Rev.* **1953**, *91*, 1505.

[65] D. Dürr, A. Bach, "The Onsager-Machlup function as Lagrangian for the most probable path of a diffusion process", **1978**, *60*, 153.

[66] H. Haken, "Generalized Onsager-Machlup Function and Classes of Path Integral Solutions of the Fokker-Planck Equation and the Master Equation", *Z. Physik B* **1976**, *24*, 321.

[67] T. B. Woolf, "Path corrected functionals of stochastic trajectories: towards relative free energy and reaction coordinate calculations", *Chem Phys. Lett.* **1998**, *289*, 433.

[68] D. M. Zuckerman, T. B. Woolf, "Dynamic reaction paths and rates through importance-sampled stochastic dynamics", *J. Chem. Phys.* **1999**, *111*, 9475.

[69] D. M. Zuckerman, T. B. Woolf, "Efficient dynamic importance sampling of rare events in one dimension", *Phy. Rev. E* **2000**, *63*, 016702.

[70] A. B. Adib, "Stochastic Actions for Diffusive Dynamics: Reweighting, Sampling, and Minimization", *J. Phys. Chem. B* **2008**, *112*, 5910.

[71] C. Xing, I. Andricioaei, "On the calculation of time correlation functions by potential scaling", **2006**, *124*, 034110.

[72] C. Schütte, A. Nielsen, M. Weber, "Markov state models and molecular alchemy", *Mol. Phys.* **2015**, *113*, 69.

[73] L. Donati, C. Hartmann, B. G. Keller, "Girsanov reweighting for path ensembles and Markov state models", *J. Chem. Phys.* **2017**, *146*, 244112.

[74] L. Donati, B. G. Keller, "Girsanov reweighting for metadynamics simulations", *J. Chem. Phys.* **2018**, *149*, 072335.

[75] L. Donati, M. Weber, B. G. Keller, "A review of Girsanov reweighting and of square root approximation for building molecular Markov state models", *J. Math. Phys.* **2022**, *63*, 123306.

[76] C. Schütte, W. Huisinga, P. Deuflhard, *"Transfer Operator Approach to Conformational Dynamics" in Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems*, (Ed.: B. Fiedler), Springer Berlin, **2001**.

[77] W. C. Swope, J. W. Pitera, F. Suits, M. Pitman, M. Eleftheriou, B. G. Fitch, R. S. Germain, A. Rayshubski, T. J. C. Ward, Y. Zhestkov, R. Zhou, "Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 2. Example Applications to Alanine Dipeptide and β-Hairpin Peptide", *J. Phys. Chem. B* **2004**, *108*, 6582.

[78] N.-V. Buchete, G. Hummer, "Coarse master equations for peptide folding dynamics", *J. Phys. Chem. B* **2008**, *112*, 6057.

[79] B. Keller, X. Daura, W. F. Van Gunsteren, "Comparing geometric and kinetic cluster algorithms for molecular simulation data", *J. Chem. Phys.* **2010**, *132*, 074110.

[80] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, F. Noé, "Markov models of molecular kinetics: generation and validation", *J. Chem. Phys.* **2011**, *134*, 174105.

[81] W. Wang, S. Cao, L. Zhu, X. Huang, "Constructing Markov State Models to elucidate the functional conformational changes of complex biomolecules", *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2018**, *8*, e1343.

[82] J. Quer, L. Donati, B. G. Keller, "An automatic adaptive importance sampling algorithm for molecular dynamics in reaction coordinates", *SIAM J. Sci. Comput.* **2018**, *40*, A653.

[83] N. G. van Kampen, *Stochastic Processes in Physics and Chemistry*, North-Holland Publishing Company Amsterdam, New York, Oxford, **1981**.

[84] H. Risken, *The Fokker-Planck Equation. Methods of Solution and Applications*, 2nd ed., Springer Berlin, Heidelberg, **1984**.

[85] P. J. Watson, L. Fairall, G. M. Santos, J. W. R. Schwabe, "Structure of HDAC3 bound to co-repressor and inositol tetraphosphate", *Nature* **2012**, *481*, 335.

[86] Q. Wang, E. M. Vogan, L. M. Nocka, C. E. Rosen, J. A. Zorn, S. C. Harrison, J. Kuriyan, "Autoinhibition of Bruton's tyrosine kinase (Btk) and activation by soluble inositol hexakisphosphate", *Elife* **2015**, *2015*, 1.

[87] R. D. Blind, "Structural analyses of inositol phosphate second messengers bound to signaling effector proteins", *Adv. Biol. Regul.* **2020**, *75*, 100667.

[88] M. Kanehisa, "Toward understanding the origin and evolution of cellular organisms", *Protein Sci.* **2019**, *28*, 1947.

[89] R. F. Irvine, M. J. Schell, "Back in the water: the return of the inositol phosphates", *Nat. Rev. Mol. Cell Biol.* **2001**, *2*, 327.

[90] S. Chatree, N. Thongmaen, K. Tantivejkul, C. Sitticharoon, I. Vucenik, "Role of inositols and inositol phosphates in energy metabolism", *Molecules* **2020**, *25*, 1.

[91] C. J. Barker, J. Wright, P. J. Hughes, C. J. Kirk, R. H. Michell, "Complex changes in cellular inositol phosphate complement accompany transit through the cell cycle", *Biochem. J.* **2004**, *380*, 465.

[92] M. Nguyen Trung, D. Furkert, D. Fiedler, "Versatile signaling mechanisms of inositol pyrophosphates", *Curr. Opin. Chem. Biol.* **2022**, *70*, 102177.

[93] R. A. Dick, K. K. Zadrozny, C. Xu, F. K. M. Schur, T. D. Lyddon, C. L. Ricana, J. M. Wagner, J. R. Perilla, B. K. Ganser-Pornillos, M. C. Johnson, et al., "Inositol phosphates are assembly co-factors for HIV-1", *Nature* **2018**, *560*, 509.

[94] P. J. Watson, C. J. Millard, A. Riley, N. Robertson, L. Wright, H. Godage, S. Cowley, A. Jamieson, B. V. L. Potter, J. W. R. Schwabe, "Insights into the activation mechanism of class I HDAC complexes by inositol phosphates", *Nat. Commun.* **2016**, *7*, 1.

[95] H. Lin, Y. Yan, Y. Luo, W. Y. So, X. Wei, X. Zhang, X. Yang, J. Zhang, Y. Su, X. Yang, et al., "IP6-assisted CSN-COP1 competition regulates a CRL4-ETV5 proteolytic checkpoint to safeguard glucose-induced insulin secretion", *Nat. Commun.* **2021**, *12*, 1.

[96] H. Lin, X. Zhang, L. Liu, Q. Fu, C. Zang, Y. Ding, Y. Su, Z. Xu, S. He, X. Yang, et al., "Basis for metabolite-dependent Cullin-RING ligase deneddylation by the COP9 signalosome", *Proc. Natl. Acad. Sci. U.S.A* **2020**, *117*, 4117.

[97] M. Köhn, "Turn and Face the Strange: A New View on Phosphatases", *ACS Cent. Sci.* **2020**, *6*, 467.

[98] B. Appelhof, M. Wagner, J. Hoefele, A. Heinze, T. Roser, M. Koch-Hogrebe, S. D. Roosendaal, M. Dehghani, M. Y. V. Mehrjardi, E. Torti, et al., "Pontocerebellar hypoplasia due to bi-allelic variants in MINPP1", *Eur. J. Hum. Genet.* **2021**, *29*, 411.

[99] E. Ucuncu, K. Rajamani, M. S. C. Wilson, D. Medina-Cano, N. Altin, P. David, G. Barcia, N. Lefort, C. Banal, M. T. Vasilache-Dangles, et al., "MINPP1 prevents intracellular accumulation of the chelator inositol hexakisphosphate and is mutated in Pontocerebellar Hypoplasia", *Nat. Commun.* **2020**, *11*, 6087.

[100] H. C. Andersen, "Molecular dynamics simulations at constant pressure and/or temperature", *J. Chem. Phys.* **1980**, *72*, 2384.

[101] C. W. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*, 2nd ed., Springer Verlag, Berlin Heidelberg, **1983**.

[102] M. Chaichian, A. Demichev, *Path Integrals in Physics Volume I: Stochastic Processes and Quantum Mechanics*, IOP Publishing Ltd, Bristol and Philadelphia, **2001**.

[103] P. C. Bressloff, *Stochastic Processes in Cell Biology*, Springer, New York, **2014**.

[104] W. Paul, J. Baschnagel, *Stochastic Processes: From Physics to Finance*, Springer Science and Business Media, **2013**.

[105] E. R. Dougherty, *Random processes for image and signal processing*, SPIE Optical Engineering Press, **1999**.

[106] K. Huang, *Statistical Mechanics*, 2nd ed., John Wiley and Sons, **1987**.

[107] R. Zwanzig, *Nonequilibrium Statistical Mechanics*, Oxford University Press, **2001**.

[108] H. Jia, K. Li, "A third accurate operator splitting method", *Math. Comput. Model.* **2011**, *53*, 387.

[109] G. Strang, "On the construction and comparison of difference schemes", *SIAM J. Numer. Anal.* **1968**, *5*, 506.

[110] C. C. Chow, M. A. Buice, "Path Integral Methods for Stochastic Differential Equations", *J. Math. Neurosci.* **2015**, *5*, 8.

[111] I. V. Girsanov, "On transforming a certain class of stochastic processes by absolutely continuous substitution of measures", *Theory Probab. Its Appl.* **1960**, *5*, 285.

[112] G. R. Bowman, V. S. Pande, F. Noé, *"An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation" in Advances in Experimental Medicine and Biology, vol. 797*, Springer Dordrecht, Heidelberg, New York, London, **2014**.

[113] OpenMMTools Github, `https://github.com/choderalab/openmmtools`, **accessed: 12.10.2022**.

[114] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, V. S. Pande, "OpenMM 7: Rapid development of high performance algorithms for molecular dynamics", *PLOS Comp. Biol.* **2017**, *13*, 1–17.

[115] D. A. Sivak, J. D. Chodera, G. E. Crooks, "Using Nonequilibrium Fluctuation Theorems to Understand and Correct Errors in Equilibrium and Nonequilibrium Simulations of Discrete Langevin Dynamics", *Phys. Rev. X* **2013**, *3*, 011007.

[116] Z. F. Brotzakis, P. G. Bolhuis, "Approximating free energy and committor landscapes in standard transition path sampling using virtual interface exchange", *J. Chem. Phys.* **2019**, *151*, 174111.

[117] T. N. Starr, N. Czudnochowski, Z. Liu, F. Zatta, Y.-J. Park, A. Addetia, D. Pinto, M. Beltramello, P. Hernandez, A. J. Greaney, et al., "SARS-CoV-2 RBD antibodies that maximize breadth and resistance to escape", *Nature* **2021**, *597*, 97.

[118] D. A. Case, H. M. Aktulga, K. Belfon, I. Ben-Shalom, S. R. Brozell, D. S. Cerutti, T. E. Cheatham III, V. W. D. Cruzeiro, T. A. Darden, R. E. Duke, et al., *Amber 2021*, University of California, San Francisco, **2021**.

[119] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, E. Lindahl, "GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers", *SoftwareX* **2015**, *1*, 19.

[120] S. Kieninger, L. Donati, B. G. Keller, "Dynamical reweighting methods for Markov models", *Curr. Opin. Struct. Biol.* **2020**, *61*, 124.

[121] OpenMM LangevinMiddleIntegrator, `http://docs.openmm.org/development/api-python/generated/openmm.openmm.LangevinMiddleIntegrator.html`, **2015 (accessed: 12.10.2022)**.

[122] Amber 2021 Reference Manual, section 21.6.10. `https://ambermd.org/doc12/Amber21.pdf`, **2021 (accessed: 12.10.2022)**.

[123] L. S. Bigman, Y. Levy, "Proteins: molecules defined by their trade-offs", *Curr. Opin. Struct. Biol.* **2020**, *60*, 50.

[124] C. A. F. de Oliveira, D. Hamelberg, J. A. McCammon, "Estimating kinetic rates from accelerated molecular dynamics simulations: Alanine dipeptide in explicit solvent as a case study", *J. Chem. Phys.* **2007**, *127*, 175105.

[125] A. T. Frank, I. Andricioaei, "Reaction Coordinate-Free Approach to Recovering Kinetics from Potential-Scaled Simulations: Application of Kramers' Rate Theory", *J. Phys. Chem. B* **2016**, *120*, 8600.

[126] P. Tiwary, M. Parrinello, "From Metadynamics to Dynamics", *Phys. Rev. Lett.* **2013**, *111*, 230602.

[127] D. Chandler, "Statistical mechanics of isomerization dynamics in liquids and the transition state approximation", *J. Chem. Phys.* **1978**, *68*, 2959.

[128] P. Hänggi, P. Talkner, M. Borkovec, "Reaction-rate theory: fifty years after Kramers", *Rev. Mod. Phys.* **1990**, *62*, 251.

[129] R. Casasnovas, V. Limongelli, P. Tiwary, P. Carloni, M. Parrinello, "Unbinding Kinetics of a p38 MAP Kinase Type II Inhibitor from Metadynamics Simulations", *J. Am. Chem. Soc.* **2017**, *139*, 4780.

[130] D. Pramanik, Z. Smith, A. Kells, P. Tiwary, "Can One Trust Kinetic and Thermodynamic Observables from Biased Metadynamics Simulations?: Detailed Quantitative Benchmarks on Millimolar Drug Fragment Dissociation", *J. Phys. Chem. B* **2019**, *123*, 3672.

[131] Y. Wang, O. Valsson, P. Tiwary, M. Parrinello, K. Lindorff-Larsen, "Frequency adaptive metadynamics for the calculation of rare-event kinetics", *J. Chem. Phys.* **2018**, *7*, 072309.

[132] E. Rosta, G. Hummer, "Free Energies from Dynamic Weighted Histogram Analysis Using Unbiased Markov State Model", *J. Chem. Theory Comput.* **2014**, *11*, 276.

[133] H. Wan, G. Zhou, V. A. Voelz, "A Maximum-Caliber Approach to Predicting Perturbed Folding Kinetics Due to Mutations", *J. Chem. Theory Comput.* **2016**, *12*, 5768.

[134] H. Wu, A. S. J. S. Mey, E. Rosta, F. Noé, "Statistically optimal analysis of state-discretized trajectory data from multiple thermodynamic states", *J. Chem. Phys.* **2014**, *141*, 214106.

[135] H. Wu, F. Paul, C. Wehmeyer, F. Noé, "Multiensemble Markov models of molecular thermodynamics and kinetics", *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113*, E3221.

[136] L. S. Stelzl, A. Kells, E. Rosta, G. Hummer, "Dynamic Histogram Analysis to determine free Energies and rates from biased simulations", *J. Chem. Theory Comput.* **2017**, *13*, 6328.

[137] F. Paul, C. Wehmeyer, E. T. Abualrous, H. Wu, M. D. Crabtree, J. Schöneberg, J. Clarke, C. Freund, T. R. Weikl, F. Noé, "Protein-peptide association kinetics beyond the seconds timescale from atomistic simulations", *Nat. Commun.* **2017**, *8*, 1095.

[138] D. M. Zuckerman, L. T. Chong, "Weighted Ensemble Simulation: Review of Methodology, Applications, and Software", *Annu. Rev. Biophys.* **2017**, *46*, 43.

[139] L. T. Chong, A. S. Saglam, D. M. Zuckerman, "Path-sampling strategies for simulating rare events in biomolecular systems", *Curr. Opin. Struct. Biol.* **2017**, *43*, 88.

[140] L. Donati, M. Weber, B. G. Keller, "Markov models from the Square Root Approximation of the Fokker-Planck equation: calculating the grid-dependent flux", *J. Phys.: Condens. Matter* **2021**, *33*, 115902.

[141] S. Kieninger, B. G. Keller, "GROMACS Stochastic Dynamics and BAOAB are equivalent configurational sampling algorithms", *J. Chem. Theory Comput.* **2022**, *18*, 5797.

[142] A. Nunes-Alves, D. M. Zuckerman, G. Menegon Arantes, "Escape of a Small Molecule from Inside T4 Lysozyme by Multiple Pathways", *Biophys. J.* **2018**, *114*, 1058.

[143] S. Wolf, B. Lickert, S. Bray, G. Stock, "Multisecond ligand dissociation dynamics from atomistic simulations", *Nature Commun.* **2020**, *11*, 2918.

[144] D. B. Kokh, R. C. Wade, "G Protein-Coupled Receptor-Ligand Dissociation Rates and Mechanisms from τRAMD Simulations", *J. Chem. Theory Comput.* **2021**, *17*, 6610.

[145] S. Kieninger, B. G. Keller, "Path probability ratios for Langevin dynamics-Exact and approximate", *J. Chem. Phys.* **2021**, *154*, 094102.

[146] D. Qiu, M. S. Wilson, V. B. Eisenbeis, R. K. Harmel, E. Riemer, T. M. Haas, C. Wittwer, N. Jork, C. Gu, S. B. Shears, G. Schaaf, B. Kammerer, A. Fiedler, D. Saiardi, H. J. Jessen, "Analysis of inositol phosphate metabolism by capillary electrophoresis electrospray ionization mass spectrometry", *Nat. Commun.* **2020**, *11*, 6035.

[147] A. J. Letcher, M. J. Schell, R. F. Irvine, "Do mammals make all their own inositol hexakisphosphate?", *Biochem. J.* **2008**, *416*, 263.

[148] H. Chi, X. Yang, P. D. Kingsley, R. J. O'Keefe, J. E. Puzas, R. N. Rosier, S. B. Shears, P. R. Reynolds, "Targeted Deletion of Minpp1 Provides New Insight into the Activity of Multiple Inositol Polyphosphate Phosphatase In Vivo", *Mol. Cell. Biol.* **2000**, *20*, 6496.

[149] S. P. Kilaparty, R. Agarwal, P. Singh, K. Kannan, N. Ali, "Endoplasmic reticulum stress-induced apoptosis accompanies enhanced expression of multiple inositol polyphosphate phosphatase 1 (Minpp1): a possible role for Minpp1 in cellular stress response", *Cell Stress Chaperones* **2016**, *21*, 593.

[150] S. P. Kilaparty, A. Singh, W. H. Baltosser, N. Ali, "Computational analysis reveals a successive adaptation of multiple inositol polyphosphate phosphatase 1 in higher organisms through evolution", *Evol. Bioinforma.* **2014**, *10*, 239.

[151] K. Nogimori, P. Hughes, M. Glennon, M. Hodgson, J. Putney, S. Shears, "Purification of an Inositol (1,3,4,5)-tetrakisphosphate 3-phosphatase activity from rat liver and the evaluation of its substrate specificity", *J. Biol. Chem.* **1991**, *266*, 16499.

[152] A. Craxton, J. J. Caffrey, W. Burkhart, T. S. Safrany, B. S. Shears, "Molecular cloning and expression of a rat hepatic multiple inositol polyphosphate phosphatase", *Biochem. J.* **1997**, *328*, 75.

[153] "Gene ID: 9562, Homo sapiens, MINPP1 multiple inositol-polyphosphate phosphatase", *Gene [Internet] Bethesda Natl Libr Med (US), Natl Cent Biotechnol Information* **1988**, `https://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=DetailsSearch&Term=9562#bibliography`, **accessed: 23.08.2022**.

[154] J. Yu, B. Leibiger, S. N. Yang, J. J. Caffery, S. B. Shears, I. B. Leibiger, C. J. Barker, P.-O. Berggren, "Cytosolic Multiple Inositol Polyphosphate Phosphatase in the Regulation of Cytoplasmic Free Ca^{2+} Concentration", *J. Biol. Chem.* **2003**, *278*, 46210.

[155] C. J. Barker, C. Illies, P.-O. Berggren, *"HPLC Separation of Inositol Polyphosphates" in Inositol Phosphates and Lipids*, (Ed.: C. J. Barker), Humana Press, **2010**.

[156] S. B. Shears, *"A Short Historical Perspective of Methods in Inositol Phosphate Research" in Inositol Phosphates*, (Ed.: G. Miller), Springer US, **2020**.

[157] R. K. Harmel, R. Puschmann, M. Nguyen Trung, A. Saiardi, P. Schmieder, D. Fiedler, "Harnessing ^{13}C-labeled *myo*-inositol to interrogate inositol phosphate messengers by NMR", *Chem. Sci.* **2019**, *10*, 5267.

[158] C. J. Barker, P. J. French, A. J. Moore, T. Nilsson, P. O. Berggren, C. M. Bunce, C. J. Kirk, R. H. Michell, "Inositol 1,2,3-trisphosphate and inositol 1,2- and/or 2,3-bisphosphate are normal constituents of mammalian cells", *Biochem. J.* **1995**, *306*, 557.

[159] C. Dellago, P. G. Bolhuis, F. S. Csajka, D. Chandler, "Transition path sampling and the calculation of rate constants", *J. Chem. Phys.* **1998**, *108*, 1964.

[160] C. Dellago, P. G. Bolhuis, D. Chandler, "On the calculation of reaction rate constants in the transition path ensemble", *J. Chem. Phys.* **1998**, *108*, 9263.

[161] C. Dellago, P. G. Bolhuis, D. Chandler, "Sampling ensembles of deterministic transition pathways", *Faraday Discuss.* **1998**, *110*, 421.

[162] C. Dellago, P. G. Bolhuis, P. L. Geissler, *Transition Path Sampling*, John Wiley & Sons, Inc., **2002**.

[163] C. Dellago, P. Bolhuis, *"Transition Path Sampling and Other Advanced Simulation Techniques for Rare Events" in Advanced Computer Simulation Approaches for Soft Matter Sciences III. Advances in Polymer Science, vol. 221*, (Eds.: C. Holm, K. Kremer), Springer Berlin, Heidelberg, **2000**.

[164] P. G. Bolhuis, Z. F. Brotzakis, B. G. Keller, "Force field optimization by imposing kinetic constraints with path reweighting", *arXiv*, **2022**, `https://doi.org/10.48550/arXiv.2207.04558`.

[165] B. S. Mantilla, L. D. D. Amaral, H. J. Jessen, R. Docampo, "The Inositol Pyrophosphate Biosynthetic Pathway of Trypanosoma cruzi", *ACS Chem. Biol.* **2021**, *16*, 283.

[166] M. J. del Razo, D. Frömberg, V. A. Straube, C. Schütte, F. Höfling, S. Winkelmann, "A probabilistic framework for particle–based reaction-diffusion dynamics using classical Fock space representations", *Lett. Math. Phys.* **2022**, *112*, 46.

[167] M. J. del Razo, S. Winkelmann, R. Klein, F. Höfling, "Chemical diffusion master equation: formulations of reaction–diffusion processes on the molecular level", *arXiv*, **2022**, https://doi.org/10.48550/arXiv.2210.02268.