Instytut Fizyki, Uniwersytet Mikołaja Kopernika
za pośrednictwem:
Rady Doskonałości Naukowej
pl. Defilad 1
00-901 Warszawa
(Pałac Kultury i Nauki, p. XXIV, pok. 2401)

Jakub Rydzewski
Instytut Fizyki
Wydział Fizyki, Astronomii i Informatyki Stosowanej
Uniwersytet Mikołaja Kopernika w Toruniu

Wniosek

z dnia 22/05/2024 r.

o przeprowadzenie postępowania w sprawie nadania stopnia doktora habilitowanego w dziedzinie nauk ścisłych i przyrodniczych w dyscyplinie[1]: nauk fizycznych

Określenie osiągnięcia naukowego będącego podstawą ubiegania się o nadanie stopnia doktora habilitowanego: cykl powiązanych tematycznie artykułów naukowych o zbiorczym tytule: “Uczenie zmiennych zbiorowych z symulacji atomistycznych”.

Wnioskuję – na podstawie art. 221 ust. 10 ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (Dz. U. z 2021 r. poz. 478 zm.) – aby komisja habilitacyjna podejmowała uchwałę w sprawie nadania stopnia doktora habilitowanego w głosowaniu ~~**tajnym**~~**/jawnym***[2]

Zostałem poinformowany, że:

Administratorem w odniesieniu do danych osobowych pozyskanych w ramach postępowania w sprawie nadania stopnia doktora habilitowanego jest Przewodniczący Rady Doskonałości Naukowej z siedzibą w Warszawie (pl. Defilad 1, XXIV piętro, 00-901 Warszawa).

Kontakt za pośrednictwem e-mail: kancelaria@rdn.gov.pl , tel. 22 656 60 98 lub w siedzibie organu.

Dane osobowe będą przetwarzane w oparciu o przesłankę wskazaną w art. 6 ust. 1 lit. c) Rozporządzenia UE 2016/679 z dnia z dnia 27 kwietnia 2016 r. w związku z art. 220 - 221 oraz art. 232 – 240 ustawy z dnia 20 lipca 2018 roku - Prawo o szkolnictwie wyższym i nauce, w celu przeprowadzenie postępowania o nadanie stopnia doktora habilitowanego oraz realizacji praw i obowiązków oraz środków odwoławczych przewidzianych w tym postępowaniu.

Szczegółowa informacja na temat przetwarzania danych osobowych w postępowaniu dostępna jest na stronie www.rdn.gov.pl/klauzula-informacyjna-rodo.html

.....
(podpis wnioskodawcy)

[1] Klasyfikacja dziedzin i dyscyplin wg. rozporządzenia Ministra Nauki i Szkolnictwa Wyższego z dnia 20 września 2018 r. w sprawie dziedzin nauki i dyscyplin naukowych oraz dyscyplin w zakresie sztuki (Dz. U. z 2018 r. poz. 1818).

[2] * Niepotrzebne skreślić.

Załączniki:

1. Potwierdzona przez jednostkę kopia dyplomu doktorskiego
2. Autoreferat w języku polskim i angielskim
3. Wykaz opublikowanych prac naukowych oraz informacja o osiągnięciach dydaktycznych, współpracy naukowej i popularyzacji nauki w języku polskim i angielskim
4. Kopie prac składających się na wyróżnione osiągnięcie naukowe
5. Dane kontaktowe

Autoreferat

(Stan na dzień 22.05.2024 r.)

I. Imię i nazwisko

Jakub Rydzewski

II. Posiadane dyplomy, stopnie naukowe lub artystyczne

- **2014–2018**: Stopień doktora nauk fizycznych (biofizyka), Wydział Fizyki, Astronomii i Informatyki Stosowanej, Uniwersytet Mikołaja Kopernika, Toruń, Polska. Tytuł rozprawy: *Rare-Event Sampling of Ligand Transport in Proteins* (wyróżnienie). Promotor: prof. dr hab. Wiesław Nowak.
- **2013–2014**: Tytuł zawodowy magistra informatyki stosowanej, Wydział Fizyki, Astronomii i Informatyki Stosowanej, Uniwersytet Mikołaja Kopernika, Toruń, Polska.
- **2011–2014**: Tytuł licencjata fizyki teoretycznej, Wydział Fizyki, Astronomii i Informatyki Stosowanej, Uniwersytet Mikołaja Kopernika, Toruń, Polska.
- **2009–2013**: Tytuł zawodowy inżyniera informatyki stosowanej, Wydział Fizyki, Astronomii i Informatyki Stosowanej, Uniwersytet Mikołaja Kopernika, Toruń, Polska.

III. Informacja o dotychczasowym zatrudnieniu w jednostkach naukowych lub artystycznych

- **X 2019–teraz**: Instytut Fizyki, Uniwersytet Mikołaja Kopernika, Toruń, Polska (adiunkt).
- **X 2018–X 2019**: Instytut Fizyki, Uniwersytet Mikołaja Kopernika, Toruń, Polska (asystent).
- **XI 2023–II 2024**: National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japonia, grupa prof. Tetsuyi Morishity.
- **X 2016–IV 2017**: Department of Theoretical and Computational Biophysics, Max Planck Institute for Biophysical Chemistry (od 2022 Max Planck Institute for Multidisciplinary Sciences), Getynga, Niemcy, grupa prof. Helmuta Grubmüllera.
- **VII 2016–X 2016**: Department of Chemistry and Applied Biosciences, Swiss Federal Institute of Technology in Zürich c/o Institute of Computational Science, Università della Svizzera italiana, Lugano, Szwajcaria, grupa prof. Michele Parrinello.

IV. Omówienie osiągnięcia, o którym mowa w art. 219 ust. 1 pkt. 2 Ustawy

Jakub Rydzewski (określany dalej jako "Autor"), wskazuje jako osiągnięcie naukowe (określane dalej jako "Osiągnięcie") cykl publikacji [H1, H2, H3, H4, H5, H6, H7], zgodnych z *Ustawą o Stopniach i Tytułach Naukowych*, dotyczących rozwoju metod statystycznego uczenia powolnych zmiennych zbiorowych z symulacji atomistycznych. Tytuł Osiągnięcia jest następujący:

Uczenie zmiennych zbiorowych z symulacji atomistycznych

Osiągnięcie jest cennym interdyscyplinarnym wkładem do współczesnej fizyki statystycznej. Poniższa lista podsumowuje najważniejsze części Osiągnięcia według Autora:

- Opracowanie spójnych podstaw teorii i metod statystycznego uczenia zmiennych zbiorowych na podstawie symulacji standardowych i wzmocnionego próbkowania, które można zastosować do zrozumienia dowolnego procesu molekularnego w eksperymentalnych skalach czasowych w sposób oparty na danych.
- W przeciwieństwie do wielu wcześniej zaproponowanych metod, podstawy te uwzględniają najważniejsze cechy fizyczne procesów dynamicznych. Obejmują one rozkład prawdopodobieństwa (równowagowy lub nierównowagowy) próbkowany przez badany układ, pojęcie odległości między konfiguracjami oraz powolną kinetykę, która jest kluczem do zrozumienia procesów na dłuższych skalach czasowych.
- Teoria leżąca u podstaw proponowanych metod łączy ogólne idee i narzędzia mechaniki statystycznej i uczenia maszynowego. Zapewnia to zastosowanym tu metodom uczenia maszynowego zdolność, której brakuje w większości nienadzorowanych metod redukcji wymiarowości: interpretowalność w kontekście fizyki.
- Proponowane podejście jest ogólne i umożliwia społeczności łatwe jego rozszerzanie, co podkreśla jego potencjał do szerokiego zastosowania i dalszego udoskonalania.

W następujących publikacjach zawartych w Osiągnięciu, z wyjątkiem ostatniej, Autor występuje jako pierwszy i korespondujący autor (oznaczony gwiazdką).

H1. ***Rydzewski, J.**, Chen, M. & Valsson, O. Manifold Learning in Atomistic Simulations: A Conceptual Review. *Mach. Learn.: Sci. Technol.* **4,** 031001 (2023).

H2. ***Rydzewski, J.** & Valsson, O. Multiscale Reweighted Stochastic Embedding: Deep Learning of Collective Variables for Enhanced Sampling. *J. Phys. Chem. A* **125,** 6286–6302 (2021).

H3. ***Rydzewski, J.**, Chen, M., Ghosh, T. K. & Valsson, O. Reweighted Manifold Learning of Collective Variables from Enhanced Sampling Simulations. *J. Chem. Theory Comput.* **18,** 7179–7192 (2022).

H4. ***Rydzewski, J.** Selecting High-Dimensional Representations of Physical Systems by Reweighted Diffusion Maps. *J. Phys. Chem. Lett.* **14,** 2778–2783 (2023).

H5. ***Rydzewski, J.** Spectral Map: Embedding Slow Kinetics in Collective Variables. *J. Phys. Chem. Lett.* **14,** 5216–5220 (2023).

H6. ***Rydzewski, J.** & Gökdemir, T. Learning Markovian Dynamics with Spectral Maps. *J. Chem. Phys.* **160** (2024).

H7. PLUMED Consortium, Promoting Transparency and Reproducibility in Enhanced Molecular Simulations. *Nat. Methods* **16,** 670–673 (2019).

Niniejszy tekst, unikający zagłębiania się w szczegóły, powinien służyć jako podsumowanie osiągnięc Autora oraz streszczenie dla recenzentów, a bardziej szczegółowe informacje można znaleźć w publikacjach związanych z Osiągnięciem.

1 Wprowadzenie

Symulacje atomistyczne, takie jak dynamika molekularna lub Monte Carlo, są powszechnie wykorzystywane w fizyce i chemii do badania złożonych układów dynamicznych [1, 2]. Pozwalają one na zdobycie szczegółowych informacji o procesach na poziomie mikroskopowym z rozdzielczością przestrzenno-czasową większą niż mogą zapewnić eksperymenty. Do takich procesów należą, na przykład, kataliza [3], oddziaływania ligandów z białkami [4–7] i DNA [8], przejścia szkliste w materiałach amorficznych [9], krystalizacja [10], czy trawienie grafitu [11]. Jednak badane w symulacjach układy często składają się z setek tysięcy atomów, co utrudnia analizę ich złożonej przestrzeni konfiguracyjnej. Aby uzyskać uproszczoną, zrozumiałą reprezentację badanego układu, często konieczne jest uśrednienie zaszumionych zmiennych, które nie są istotne z punktu widzenia skal czasowych na poziomie eksperymentalnym. Skutkuje to niskowymiarową reprezentacją, która powinna oddawać najważniejsze cechy układu. Parafrazując R. Coifmana [12], reprezentacja ta musi być fizycznie interpretowalna:

> Istnieje wrodzona prawda w niskowymiarowej reprezentacji danych. Chcielibyśmy mieć charakterystykę takiej ukrytej zmiennej fizycznej, która z natury opisuje zmiany stanów.

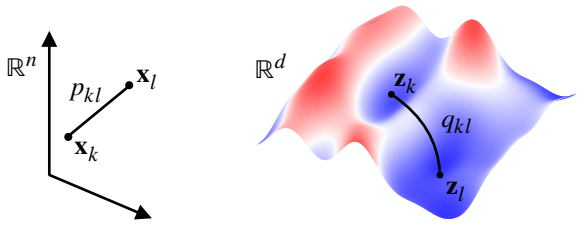
Zaproponowano wiele metod mających na celu rozwiązanie problemu wysokiej wymiarowości, takich jak teoria przejść fazowych Ginzburg'a–Landau'a [13], formalizm Mori'ego–Zwanzig'a dla transportu i ruchu kolektywnego [14–16] oraz teoria Koopman'a [17, 18]. Niedawno opracowane podejścia obejmują m.in., uczenie rozmaitości (ang. *manifold learning*) [H3, 19], klasę nieliniowych nienadzorowanych metod uczenia maszynowego i statystycznego trenowanych bezpośrednio na danych, których rozwój został zainicjowany przez innowacyjne prace Tenenbauma i in. [20] oraz Roweisa i Saula [21] opublikowanych w tym samym wydaniu Science [**290** (2000)].

Uproszczony opis dynamiki układów wieloatomowych można uzyskać wykorzystując fizykę statystyczną, gdzie współrzędne mikroskopowe składają się na przestrzeń konfiguracyjną układu. Ta reprezentacja charakteryzuje się z dużą liczbą stopni swobody. Załóżmy, że układ jest reprezentowany przez wektor *zmiennych konfiguracyjnych* o wymiarze n, $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. Zmienne konfiguracyjne są funkcjami współrzędnych mikroskopowych, przy założeniu, że przestrzeń obejmowana przez te zmienne jest wielowymiarowa. W uczeniu maszynowym zmienne konfiguracyjne są często określane jako "cechy" lub "deskryptory". Zbiór danych X składający się z K wysokowymiarowych próbek zmiennych konfiguracyjnych zarejestrowanych w kolejnych momentach dynamiki można wyrazić jako macierz o rozmiarze $n \times K$, która zwana jest *trajektorią* układu.

Bez utraty ogólności ograniczymy naszą dyskusję do zespołu kanonicznego (NVT), w którym zmienne konfiguracyjne ewoluują zgodnie z funkcją energii potencjalnej $U(\mathbf{x})$. Gdy współrzędne mikroskopowe są użyte do reprezentacji układu, jego równowagowy rozkład prawdopodobieństwa jest dany przez stacjonarny rozkład Boltzmanna [1], $p(\mathbf{x}) = \mathrm{e}^{-\beta U(\mathbf{x})}/\mathcal{Z}$, gdzie $\beta = (k_\mathrm{B}T)^{-1}$ jest temperaturą odwrotną, $k_\mathrm{B}T$, odpowiadającą temperaturze T ze stałą Boltzmanna oznaczoną przez k_B, a $\mathcal{Z} = \int \mathrm{d}\mathbf{x}\ \mathrm{e}^{-\beta U(\mathbf{x})}$ jest kanoniczną funkcją podziału. W przeciwnym razie zbiór X jest próbkowany zgodnie z nieznanym równowagowym rozkładem prawdopodobieństwa.

W fizyce statystycznej możemy uprościć opis wieloatomowego układu poprzez uśrednienie jego pewnych wysokowymiarowych właściwości. Skutkuje to makroskopowym opisem, który wykorzystuje mniej stopni swobody do scharakteryzowania zespołów mikroskopowych konfiguracji lub stanów. W symulacjach atomistycznych te uproszczone zmienne są często nazywane zmiennymi zbiorowymi (ang. *collective variables*), parametrami porządku (ang. *order parameters*) lub współrzędnymi reakcji (ang. *reaction coordinates*). Identyfikacja zmiennych zbiorowych jest trudna w przypadku złożonych układów i często wymaga uciekania się do intuicji fizycznej lub chemicznej oraz podejść typu "prób i błędów". Jednakże, poleganie wyłącznie na intuicji w celu znalezienia zmiennych zbiorowych może być niesystematyczne i utrudniać zrozumienie podstawowego procesu fizycznego, często przyczyniając się do błędnego oszacowania kinetyki układu. Takie niewłaściwe zmienne zbiorowe charakteryzują się:

1. Nakładaniem stanów metastabilnych (ang. *metastable states*), co skutkuje niedoszacowaniem



Rysunek 1: Odwzorowanie docelowe. Schematyczna ilustracja mapowania $\xi(\mathbf{x}) = \mathbf{z}$ pomiędzy wysokowymiarową przestrzenią konfiguracyjną $\mathbf{x}$ a niskowymiarową przestrzenią zmiennych zbiorowych $\mathbf{z}$ ($n \gg d$). Relacja p_{kl} między próbkami konfiguracyjnymi $\mathbf{x}_k$ i $\mathbf{x}_l$ jest zachowana w relacji q_{kl} między próbkami zmiennych zbiorowych $\mathbf{z}_k$ i $\mathbf{z}_l$. Rysunek pochodzi z Ref. [H1].

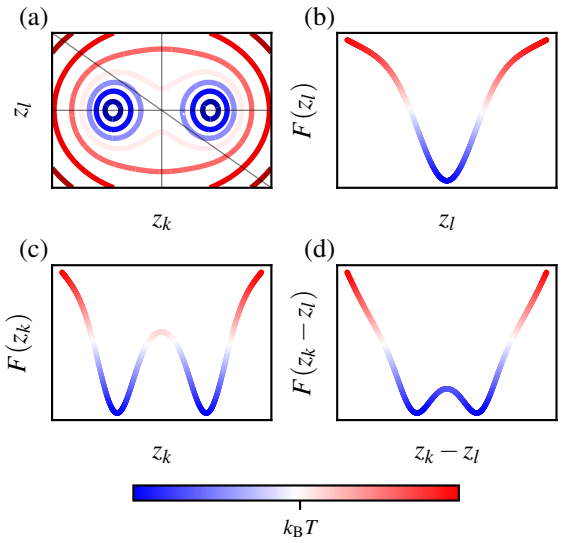
barier energii swobodnej (ang. *free energy*), niedokładnym określeniem zespołów stanów przejściowych (ang. *transition state ensemble*) i nieefektywnością wzmocnionych technik próbkowania (ang. *enhanced sampling*) ze względu na istnienie ukrytych wąskich gardeł kinetycznych [22].

2. Niezdolnością do opisu zachowania procesu na dłuższych skalach czasowych (np. mieszanie powolnych i szybkich zmiennych), a tym samym znaczne efekty niemarkowowskie, które należy następnie dodatkowo uwzględnić za pomocą uogólnionego równania Langevin'a z jądrem pamięci, jak w formalizmie Mori'ego–Zwanzig'a [14, 16].

Zmienne zbiorowe są wyrażane jako funkcje zmiennych konfiguracyjnych, co oznacza, że znalezienie ich wymaga uzyskania zestawu funkcji mapujących próbki o dużej wymiarowości do przestrzeni niskowymiarowej. Zestaw takich funkcji nazywamy *odwzorowaniem docelowym* (Rys. 1):

$$\mathbf{x} \mapsto \xi(\mathbf{x}) \equiv \left\{\xi_k(\mathbf{x})\right\}_{k=1}^{d} \tag{1}$$

gdzie $d \ll n$. Odwzorowanie docelowe $\xi(\mathbf{x})$ może być liniowe, nieliniowe, a nawet może być funkcją tożsamości (tj. redukuje to problem do selekcji). Każda metoda uczenia statystycznego zapewnia unikalną postać funkcjonalną odwzorowania docelowego używanego do zmniejszenia wymiarowości reprezentacji systemu. Od tej pory będziemy odnosić się do niskowymiarowej reprezentacji systemu jako do *przestrzeni zredukowanej* lub po prostu zmiennych zbiorowych. Aby zdefiniować rozkład prawdopodobieństwa dla zmiennych zbiorowych wyrażonych przez odwzorowanie docelowe (Rów. 1), rozważamy tylko część przestrzeni konfiguracyjnej. Rozkład zmiennych zbiorowych w stanie równowagi



Rysunek 2: Metastabilność i zmienne zbiorowe. (a) Modelowy krajobraz energii swobodnej z dwoma stanami metastabilnymi oddzielonymi barierą wyższą niż energia termiczna. Zdolność jednowymiarowych CV do rozróżniania stanów: (b) projekcja wzdłuż zmiennej z_l pokazuje pojedynczy stan, (c) wzdłuż z_k pokazuje prawidłowy krajobraz, (d) wzdłuż $z_k - z_l$ pokazuje dwa stany oddzielone niedoszacowaną barierą energetyczną ($< k_\mathrm{B}T$). Rysunek pochodzi z Ref. [H1].

uzyskuje się poprzez uśrednienie nieużywanych zmiennych. Daje nam to rozkład krańcowy:

$$p(\mathbf{z}) = \int \mathrm{d}\mathbf{x}\, \delta(\mathbf{z} - \xi(\mathbf{x}))p(\mathbf{x}). \tag{2}$$

który zazwyczaj zawiera kilka rozłącznych stanów o wysokim prawdopodobieństwie oddzielonych regionami o niskim prawdopodobieństwie, co prowadzi do metastabilności (Rys. 2). Zamiast funkcji energii potencjalnej $U(\mathbf{x})$ charakterystycznej dla wielowymiarowej reprezentacji, dynamika układu w przestrzeni zredukowanej jest zgodna z *energią swobodną* (ang. *free energy*). Zapisujemy ją jako ujemny logarytm marginalnego rozkładu zmiennych zbiorowych pomnożony przez energię termiczną:

$$F(\mathbf{z}) = -\frac{1}{\beta} \log p(\mathbf{z}) \tag{3}$$

gdzie energia swobodna jest zdefiniowana z dokładnością do stałej. Rozkład równowagowy zmiennych zbiorowych może być równoważnie zapisany jako $p(\mathbf{z}) = \mathrm{e}^{-\beta F(\mathbf{z})}/\mathcal{Z}_V$, gdzie funkcja podziału w przestrzeni zredukowanej jest dana jako $\mathcal{Z}_V = \int \mathrm{d}\mathbf{z}\, \mathrm{e}^{-\beta F(\mathbf{z})}$.

Wyczerpujące próbkowanie energii swobodnej jest wyzwaniem nawet dla prostszych układów. W skalach czasowych dostępnych dla standardowych symulacji atomistycznych (około milisekund), prze-

jścia przez wysokie bariery energii swobodnej są rzadkimi zdarzeniami. W rezultacie układ pozostaje kinetycznie uwięziony w stanie metastabilnym, ponieważ jego dynamika jest ograniczona do próbkowania szybszych fluktuacji. W tym celu w ostatnich latach opracowano kilka algorytmów wzmocnionego próbkowania, w tym próbkowanie przez temperowanie równoległe (ang. *parallel tempering*) [23–25], wariacyjne [26, 27], obciążające (ang. *biasing*) [28–33] lub ich kombinacje [34]. Aby uzyskać kompleksowy przegląd i klasyfikację tych metod, odsyłamy do artykułu Henin et al. [35].

Jako reprezentatywny przykład technik wzmocnionego próbkowania, rozważamy metody wykorzystujące zewnętrzny (tj. niefizyczny) potencjał obciążający (ang. *biasing potential*) w celu sztucznego wzmocnienia fluktuacji zmiennych zbiorowych. Pierwsze podejście tego rodzaju, zwane próbkowaniem parasolowym (ang. *umbrella sampling*) [28], zostało wprowadzone w 1977 roku przez Torrie i Valleau.[1] Po wprowadzeniu do układu potencjału obciążającego, rozkład zmiennych zbiorowych może znacznie odbiegać od stanu równowagi. Skutkuje to próbkowaniem zgodnie z obciążonym rozkładem, który z założenia jest łatwiejszy do próbkowania. Inną popularną techniką z tej kategorii jest metadynamika [30, 31], w której potencjał obciążający jest konstruowany poprzez deponowanie energii w postaci rozkładów Gaussa w przestrzeni zredukowanej.

Efektem ubocznym zwiększenia fluktuacji zmiennych zbiorowych jest znaczne odchylenie od stanu równowagowego. Aby wyodrębnić właściwości równowagowe (np. energię swobodną, kinetykę) z obciążonych symulacji, każdej próbce przypisuje się wagę statystyczną:

$$w(\mathbf{z}) \propto \frac{p(\mathbf{z})}{q(\mathbf{z})}, \tag{4}$$

gdzie $q(\mathbf{z})$ jest rozkładem prawdopodobieństwa obciążonym w przestrzeni zredukowanej. Standardowe ważenie próbek polega na wykorzystaniu wag do znalezienia stacjonarnego rozkładu równowagi z rozkładu obciążonego, który można obliczyć za pomocą histogramowania lub estymacji gęstości jądra.

2 Uczenie statystyczne zmiennych zbiorowych

Uczenie statystyczne jest obszerną dziedziną, która obejmuje metody redukcji danych wielowymiarowych do niskowymiarowych rozmaitości (ang. *manifold*) przy użyciu nieliniowych odwzorowań. Takie metody, zwane uczeniem rozmaitości, powstały na bazie metod liniowych, takich jak analiza składowych głównych (ang. *principal component analysis*) lub rozkład według wartości osobliwych (ang. *singular value decomposition*), które są powszechnie stosowane w analizie danych. W ostatnim czasie uczenie rozmaitości zyskało na znaczeniu w symulacjach atomistycznych, w szczególności w

[1] Motywacją do nazwania metody *umbrella sampling* było podkreślenie jej wszechstronności do badania szerokiego zakresu procesów fizycznych.

wyodrębnianiu właściwości fizycznych z dynamiki układów złożonych. W tym kontekście rozmaitość odnosi się do zredukowanej przestrzeni dokładnie zdefiniowanej przez kilka zmiennych zbiorowych.

Zgodnie z hipotezą rozmaitości (ang. *manifold hypothesis*), uczenie rozmaitości zakłada, że dynamika w wielowymiarowej przestrzeni może być dokładnie reprezentowana przez niskowymiarową i gładką podprzestrzeń znaną jako rozmaitość. Ten niskowymiarowy opis przypisuje się sprzężeniu między różnymi stopniami swobody, co skutkuje ograniczoną liczbą powolnie ewoluujących zmiennych, które rządzą dynamiką. Szybkie stopnie swobody są kontrolowane przez dynamikę tych powolniejszych zmiennych, co prowadzi do adiabatycznej separacji skal czasowych. Takie podejście umożliwia modelowanie złożonych układów jako procesów dyfuzyjnych, ze stochastycznymi równaniami różniczkowymi stosowanymi do opisu powolnych zmiennych. Jednocześnie szybkie stopnie swobody są traktowane jako szum termiczny.

Metody uczenia rozmaitości wykorzystują pojęcie podobieństwa między próbkami wielowymiarowymi, zwykle za pomocą metryki odległości [20, 21, 36–42]. Odległości te są następnie włączane do globalnej parametryzacji danych przy użyciu dyskretnego łańcucha Markowa, w którym podobieństwa zależą odwrotnie proporcjonalnie od odległości między próbkami. Na przykład, wspólnym punktem wyjścia jest konstrukcja łańcucha Markowa opartego na anizotropowym jądrze dyfuzji [43]:

$$L(\mathbf{x}_k, \mathbf{x}_l) = \frac{g(\mathbf{x}_k, \mathbf{x}_l)}{[\varrho(\mathbf{x}_k)]^\alpha [\varrho(\mathbf{x}_l)]^\alpha}, \tag{5}$$

gdzie $g(\mathbf{x}_k, \mathbf{x}_l) = \exp(-\|\mathbf{x}_k - \mathbf{x}_l\|^2/\varepsilon)$ jest jądrem gaussowskim ze stałą skalarną ε, $\varrho(\mathbf{x}_k) = \sum_n g(\mathbf{x}_k, \mathbf{x}_n)$ jest punktowym oszacowaniem gęstości jądra w $\mathbf{x}_k$, a α jest stałą dyfuzji anizotropowej. Następnie jądro dyfuzji anizotropowej jest znormalizowane w celu przedstawienia macierzy prawdopodobieństw Markowa:

$$p_{kl} \sim M(\mathbf{x}_k, \mathbf{x}_l) = \frac{L(\mathbf{x}_k, \mathbf{x}_l)}{\sum_n L(\mathbf{x}_k, \mathbf{x}_n)}, \tag{6}$$

która zawiera informacje o przejściu z $\mathbf{x}_k$ do $\mathbf{x}_l$. W tym ujęciu M oznacza łańcuch Markowa z prawdopodobieństwem przejścia z $\mathbf{x}_k$ do $\mathbf{x}_l$.

Stała dyfuzji anizotropowej jest związana z uwzględnieniem gęstości danych [44]. W granicy $\varepsilon \to 0$ i nieskończonej liczby próbek $K \to \infty$, możemy rozważyć następujące sytuacje:

1. $\alpha = 0$: odzyskujemy dynamikę zgodnie z potencjałem $2U(\mathbf{x})$ i rozkładem $\propto [p(\mathbf{x})]^2$ dla klasycznego znormalizowanego laplasjanu grafu [36, 38, 45].

2. $\alpha = 1$: otrzymujemy laplasjan grafu z danymi równomiernie rozłożonymi na rozmaitości. Ta normalizacja uwzględnia tylko geometrię danych, podczas gdy gęstość nie odgrywa żadnej roli.

3. $\alpha = \frac{1}{2}$: otrzymujemy dynamikę zgodnie z potencjałem $U(\mathbf{x})$ i rozkładem $\propto p(\mathbf{x})$, gdzie funkcje

własne zbudowanego łańcucha Markowa odzwierciedlają asymptotykę danych w długim czasie (tj. odpowiadają zmiennym powolnym).

Trzecia opcja jest dla nas najważniejsza, ponieważ zajmujemy się układami dynamicznymi, które ewoluują zgodnie z potencjałem $U(\mathbf{x})$ i interesują nas najwolniejsze mody takich układów. Dla anizotropowej stałej dyfuzji $\alpha = \frac{1}{2}$, asymptotycznie odzyskujemy długoczasową dynamikę układu, którego współrzędne mikroskopowe są próbkowane z rozkładu Boltzmann'a. Związane z tym tłumione równanie Langevin'a to [44]:

$$\dot{\mathbf{x}} = -\beta \nabla U(\mathbf{x}) + \sqrt{2}\eta(t), \tag{7}$$

gdzie η jest n-wymiarowym ruchem Browna. Ewolucja w czasie $\mathbf{x}$ prowadzi do wstecznego równania Fokker'a-Plank'a i związanego z nim infinitezymalnego generatora $\mathcal{L}$ tego procesu dyfuzji:

$$\mathcal{L} = \mathrm{e}^{\beta U(\mathbf{x})} \nabla \mathrm{e}^{-\beta U(\mathbf{x})} \nabla \tag{8}$$

którego wartości własne i wektory własne określają informacje kinetyczne procesu dyfuzji i mogą być wykorzystane do parametryzacji przestrzeni zredukowanej. Ponieważ generator $\mathcal{L}$ ma zwykle kilka dominujących wartości własnych dla układów metastabilnych, znalezienie podejścia do wyodrębnienia generatora z dyskretnych zbiorów danych może być wykorzystane do skonstruowania *powolnych* zmiennych zbiorowych.

Aby przedstawić każdą metodę opracowaną przez Autora, zaczynamy od sklasyfikowania odwzorowania docelowego z wielowymiarowej przestrzeni konfiguracyjnej do przestrzeni zredukowanej. W tym kontekście znajdowanie zmiennych zbiorowych jest równoważne znajdowaniu optymalnej parametryzacji odwzorowania docelowego, łącząc w ten sposób koncepcje fizyki statystycznej i uczenia nienadzorowanego.

Odwzorowanie docelowe wykonuje redukcję wymiarowości w taki sposób, że wymiarowość zredukowanej przestrzeni jest znacznie niższa niż wymiarowość przestrzeni wielowymiarowej, tj. $d \ll n$. W kontekście symulacji atomistycznych proces ten można zamknąć w następujących krokach:

1. Generowanie wysokowymiarowych próbek ze standardowych lub wzmocnionych symulacji atomistycznych.
2. Konstrukcja łańcucha Markowa na danych z prawdopodobieństwami przejścia między próbkami.
3. Parametryzacja zmiennych zbiorowych z wykorzystaniem odwzorowania docelowego osadzającego wysokowymiarowe próbki w przestrzeni zredukowanej.

To podejście do uczenia się można podzielić na dwie kategorie, w zależności od tego, w jaki sposób zredukowana przestrzeń jest konstruowana przez odwzorowanie docelowe:

I. *Optymalizacja dywergencji*, w której minimalizowana jest dywergencja (tj. odległość statystyczna między parą rozkładów prawdopodobieństw) między macierzą przejścia Markowa M zbudowaną z próbek wielowymiarowych a macierzą przejścia Markowa $Q(\mathbf{z}_k, \mathbf{z}_l)$, zbudowaną z próbek niskowymiarowych. Jako takie, metody te mogą dopasować macierze przejścia Markowa w obu przestrzeniach, umożliwiając zachowanie informacji fizycznych bez konieczności użycia dekompozycji spektralnej. Odwzorowanie docelowe wyrażone jako parametryzowalne mapowanie to:

$$\xi_w(\mathbf{x}) = \left\{\xi_k(\mathbf{x}; w)\right\}_{k=1}^{d}, \tag{9}$$

gdzie w to parametry, które są zmieniane tak, aby zminimalizować rozbieżność między M i Q. W takich metodach M jest stałe, podczas gdy Q jest szacowane przez sparametryzowane odwzorowanie docelowe. W zależności od zastosowanej metody uczenia rozmaitości, minimalizacja może być wykonywana w różny sposób, tj. metody zejścia gradientowego [41], lub stochastyczne zejścia gradientowego, jeśli odwzorowanie docelowe jest reprezentowane przez sieć neuronową [H2, 46, 47].

II. *Dekompozycja na wartości i wektory własne* (tj. dekompozycja spektralna) macierzy przejścia Markowa:

$$M\varphi_k = \lambda_k \varphi_k, \tag{10}$$

gdzie $\{\psi_k\}$ i $\{\lambda_k\}$ są odpowiednio funkcjami własnymi i wartościami własnymi macierzy M. Rozwiązanie równania Rów. 10 definiuje przestrzeń zredukowaną [40], którą można przedstawić w następujący sposób:

$$\xi(\mathbf{x}) = \left\{\lambda_k \varphi_k(\mathbf{x})\right\}_{k=1}^{d}, \tag{11}$$

gdzie k-tą współrzędną jest $\lambda_k \varphi_k(\mathbf{x})$. Wartości własne są posortowane w porządku nierosnącym i zawierają tylko d dominujących wartości własnych, ponieważ każda z nich odpowiada istotności odpowiednich współrzędnych określonych funkcjami własnymi.

Wartości własne maleją wykładniczo i mogą być powiązane z efektywnymi skalami czasowymi badanego procesu fizycznego. W związku z tym dominujące wartości własne odpowiadają również najwolniejszym procesom.

Ta ujednolicona struktura, opublikowana w artykule przeglądowym Ref. [H1], może być wykorzystana do sklasyfikowania każdej techniki opracowanej przez Autora obejmującej Osiągnięcie. Dodatkowo została ona wykorzystana do klasyfikacji metod konstruowania zmiennych zbiorowych, które nie zostały zaprojektowane przez Autora, a tym samym nie zostały wymienione w Osiągnięciu. Praca przeglądowa [H1] została napisana z perspektywy symulacji atomistycznych, co wymagało przeformułowania rozważanych metod i rozszerzenia ich o uczenie się ze standardowych i wzmocnionych symulacji atomistycznych.

3 Wieloskalowe ważone mapowanie stochastyczne

Wieloskalowe ważone mapowanie stochastyczne (ang. *multiscale reweighted stochastic embedding*) [H2] (MRSE) to niedawno opracowana przez Autora technika konstruowania niskowymiarowych zmiennych zbiorowych z symulacji. Uczenie rozmaitości nie było dotąd rozszerzone na konstruowanie zmiennych zbiorowych z symulacji z wzmocnionym próbkowaniem; podejście to zostało dopiero wprowadzone przez Autora w Ref. [H2] i rozszerzone w Ref. [H3].

MRSE nie wykonuje dekompozycji spektralnej macierzy przejścia Markowa w celu znalezienia zmiennych zbiorowych, zamiast tego koncentrując się na zachowaniu prawdopodobieństw przejścia z tej macierzy poprzez wymuszenie dopasowania między nimi zarówno w przestrzeni wysokowymiarowej M, jak i zredukowanej Q. W MRSE odwzorowanie docelowe ξ jest parametryzowane przez uniwersalne aproksymatory, znane jako sieci neuronowe, w celu przeprowadzenia nieliniowej redukcji wymiarowości. Odwzorowanie docelowe jest dane jako:

$$\mathbf{z} : \mathbf{x} \mapsto \xi_w(\mathbf{x}), \tag{12}$$

gdzie w są parametrami odwzorowania dostosowanymi w taki sposób, że zredukowana przestrzeń jest optymalna w odniesieniu do wybranej miary statystycznej. W niektórych prostych przypadkach mapowanie w Rów. 12 można również przedstawić za pomocą kombinacji liniowej. Jednak głębokie uczenie maszynowe odniosło sukces w szerokim zakresie problemów, a stosowanie bardziej skomplikowanych przybliżeń do mapowania między przestrzeniami wielowymiarowymi i niskowymiarowymi jest dość powszechne w przypadku złożonych zbiorów danych [46, 48].

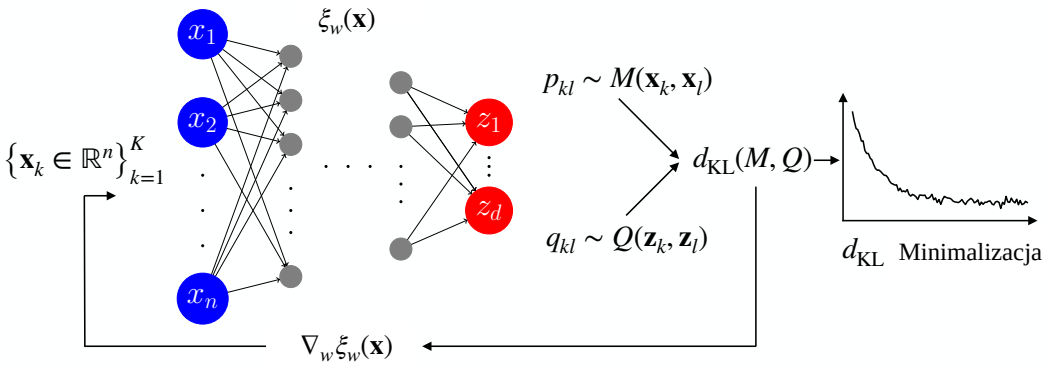
Rozważmy konstrukcję macierzy Markowa z próbek, by zakodować informacje o geometrii w przestrzeni konfiguracyjnej w prawdopodobieństwach przejścia m_{kl}:

$$m_{kl} \sim M(\mathbf{x}_k, \mathbf{x}_l) \propto L(\mathbf{x}_k, \mathbf{x}_l) \overset{\alpha=0}{=} G(\mathbf{x}_k, \mathbf{x}_l), \tag{13}$$

gdzie L to anizotropowe jądro dyfuzji (Rów. 23). W MRSE jądro L z anizotropową stałą dyfuzji $\alpha = 0$ (gęstość danych nie jest brana pod uwagę) jest po prostu jądrem Gaussa G. Ogólnie rzecz biorąc, macierz przejścia Markowa używana przez MRSE jest skonstruowana jako mieszanina rozkładów Gaussa z różnymi parametrami skali [H2]. Jednak dla uproszczenia prezentacji opisaliśmy M za pomocą pojedynczego rozkładu Gaussa.

Jednowymiarowy rozkład t [41, 46] jest używany do reprezentowania prawdopodobieństw przejścia w zredukowanej przestrzeni:

$$q_{kl} \sim Q(\mathbf{z}_k, \mathbf{z}_l) \propto \left(1 + \|\xi_{\boldsymbol{w}}(\mathbf{x}_k) - \xi_{\boldsymbol{w}}(\mathbf{x}_l)\|^2\right)^{-1}. \tag{14}$$



Rysunek 3: Schematyczne przedstawienie parametrów uczenia w dla parametrycznego odwzorowania docelowego reprezentowanego przez sieć neuronową. Procedura wstecznej propagacji szacuje błędy parametrycznego odwzorowania docelowego $\nabla_w \xi_w(\mathbf{x})$, które są wykorzystywane do korekty parametrów tak, aby dywergencja Kullbacka–Leiblera (lub jakakolwiek inna dywergencja) obliczona na podstawie macierzy przejścia Markowa $M(\mathbf{x}_k, \mathbf{x}_l)$ i $Q(\mathbf{z}_k, \mathbf{z}_l)$ zmalała do zera. Minimalna wartość dywergencji Kullbacka–Leiblera wskazuje, że relacje między próbkami w przestrzeni konfiguracyjnej i zredukowanej są zachowane. Rysunek pochodzi z Ref. [H1].

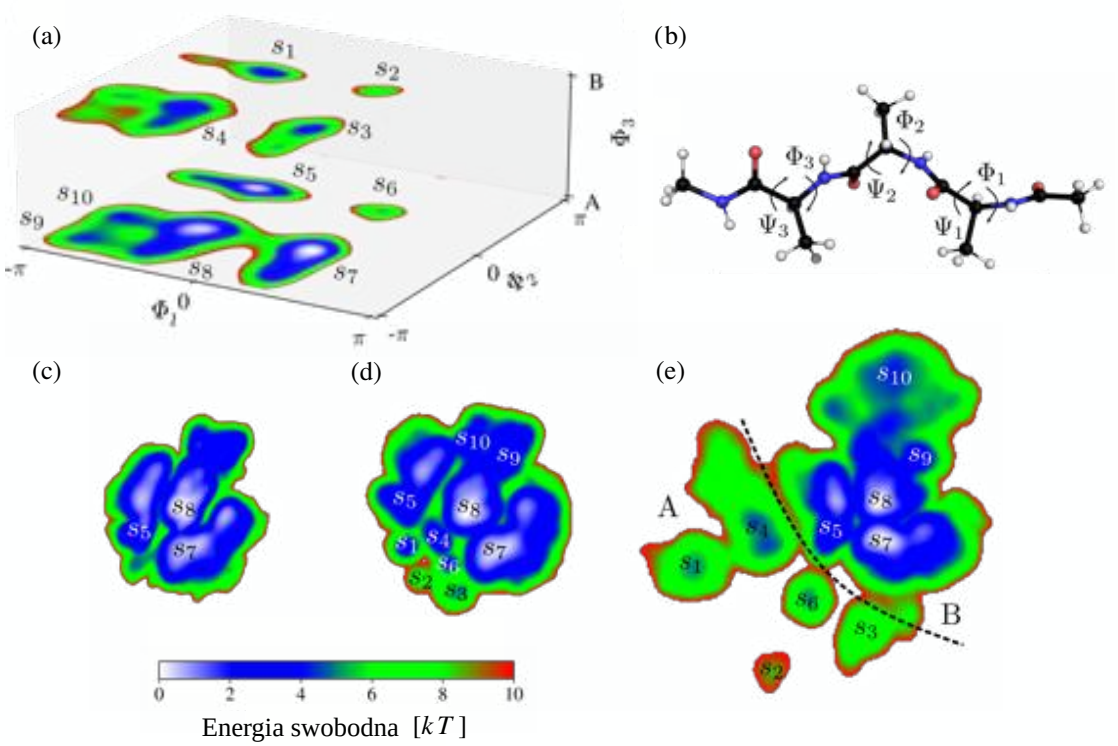
Wybór rozkładu t dla Q w MRSE jest motywowany problemem zatłoczenia [41], gdzie próbki w zredukowanej przestrzeni nie mogą być rozdzielone, gdy stosowane są rozkładu z krótkimi ogonami (np. rozkład Gaussa).

Następnie, należy porównać macierze przejścia Markowa obliczone z próbek wysokowymiarowych i niskowymiarowych. Najczęstszym wyborem dla takiej metryki jest zastosowanie odległości statystycznej, w szczególności rozbieżności Kullbacka–Leiblera:

$$D_{\mathrm{KL}}(M, Q; w) = \sum_k \sum_l m_{kl} \log\left(\frac{m_{kl}}{q_{kl}}\right), \tag{15}$$

gdzie w przeciwieństwie do standardowego sformułowania dywergencji, które porównuje dwa rozkłady prawdopodobieństwa, Rów. 15 jest obliczane dla każdej pary wierszy z M i Q, a następnie sumowane. Optymalizacja rozbieżności Kullbacka–Leiblera jest wykonywana w celu trenowania docelowego mapowania reprezentowanego przez sieć neuronową. Ponieważ docelowe odwzorowanie jest parametryczne, gradienty D_{KL} w odniesieniu do parametrów w można oszacować za pomocą wstecznej propagacji. Uproszczony schemat MRSE przedstawiono w Rys. 3.

Należy zauważyć, że Rów. 13 może być używane, gdy MRSE konstruuje zmienne zbiorowe ze standardowych symulacji, gdzie potencjał obciążający nie jest używany. Zastosowanie tej macierzy przejścia



Rysunek 4: Energia swobodna obliczona przez MRSE. (a) Stany metastabilne, które można zaobserwować wzmacniając fluktuacje kątów dwuściennych Φ_{1-3} dla tetrapeptydu alaniny. (b) Wszystkie kąty dwuścienne tetrapeptydu alaniny są używane do opisania jego przestrzeni konfiguracyjnej. Energii swobodna rozpięta przez zmienne zbiorowe dla danych (c) nieobciążonych, (d) obciążonych bez ważenia i (e) obciążonych z ważeniem. Rysunek pochodzi z Ref. [H2].

Markowa do modelowania symulacji wzmocnionego próbkowania spowodowałoby błędne uwzględnienie geometrii i gęstości danych w przestrzeni zajmowanej przez zmienne zbiorowe. Problem ten został rozwiązany dopiero w pracach Autora [H2, H3]. Aby umożliwić uczenie się na podstawie danych z symulacji wzmocnionego próbkowania, wprowadzamy następujący ansatz jako współczynnik ważenia, który skaluje macierz przejścia Markowa M:

$$r(\mathbf{x}_k, \mathbf{x}_l) = \sqrt{w(\mathbf{x}_k) w(\mathbf{x}_l)}, \tag{16}$$

będącego średnią geometryczną między dwiema wagami statystycznymi. Można to uzasadnić faktem, że potencjał odchylenia jest addytywny, a zatem średnia geometryczna jest odpowiednia do zachowania tej relacji. W związku z tym macierz przejścia Markowa jest ponownie ważona przed normalizacją,

która uwzględnia efekt odchylenia. Rów. 16, wprowadzone w Ref. [H2] bez żadnego wyprowadzenia, zostanie formalnie uzasadnione w Ref. [H3].

Jako przykład użycia MRSE, pokazujemy zmienne zbiorowe skonstruowane z symulacji tetrapeptydu alaniny i demonstrujemy jak ważenie macierzy Markowa wpływa na te zmienne (Rys. 4). Najlepszą separację między stanami metastabilnymi układu można zaobserwować w ostatnim przypadku, gdzie użyto ważenia macierzy przejścia Markowa do uwzględnienia wag z symulacji wzmocnionego próbkowania.

W Ref. [H2], Autor zaproponował i opracował metodę uczenia zredukowanej przestrzeni układu i reprezentowania jej jako zmiennych zbiorowych ze standardowych symulacji atomistycznych i rozszerzonego próbkowania. MRSE jest skonstruowana w taki sposób, że macierze przejścia skonstruowane z próbek o wysokim i niskim wymiarze są iteracyjnie ulepszane przez odwzorowanie docelowe reprezentowane przez sieć neuronową. W MRSE Autor wprowadził koncepcję ponownego ważenia w celu uwzględnienia wag statystycznych z symulacji rozszerzonego próbkowania. W pracy tej po raz pierwszy wykazano, że wykorzystanie standardowego uczenia do konstrukcji zmiennych zbiorowych z symulacji wzmocnionego próbkowania daje niewłaściwe rezultaty, ponieważ wynikowa zredukowana przestrzeń jest obciążona pod względem geometrii, gęstości i istotności próbki. Wszystkie dane, włączając dane wejściowe, wyjściowe i implementacja modułu lowlearner dla kodu symulacyjnego PLUMED za pomocą którego uzyskano wyniki, są otwarcie dostępne na `https://doi.org/10.5281/zenodo.4756093`.

4 Ważona mapa dyfuzyjna

Mapa dyfuzji (ang. *diffusion map*) zaproponowana przez Coifmana et al. [40] została zainspirowana głównie mapą własną Laplasjanu (ang. *Laplacian eigenmap*) [36, 38], która ma teoretyczne gwarancje zbieżności przy założeniu, że dane są *jednorodnie* próbkowane. Mapa dyfuzji rozszerza koncepcję mapy własnej Laplasjanu. Algorytm ten może skonstruować rodzinę mapowań do przestrzeni zredukowanej nawet gdy dane są próbkowane *niejednorodnie*.

W związku z tym mapa dyfuzji jest bardziej odpowiednia do badania złożonych układów próbkowanych przez symulacje atomistyczne, w których rozkład równowagowy jest określony przez rozkład Boltzmanna. W porównaniu z innymi metodami uczenia rozmaitości, mapa dyfuzji ma istotne podstawy teoretyczne [40, 44, 49], które pokazują, że zmienne zbiorowe obejmujące niskowymiarową rozmaitość mogą być konstruowane tak, aby odpowiadały najwolniejszym procesom relaksacji badanego układu [43].

Standardowa mapa dyfuzji może konstruować zmienne zbiorowe tylko ze standardowych symulacji atomistycznych (bez wzmocnionego próbkowania). Nietrywialne (w porównaniu ze standardowym

ważeniem próbek) podejście do uwzględniania statystycznych wag próbek jest niezbędne do uczenia się z symulacji ze wzmocnionym próbkowaniem. Skupiamy się tutaj na niedawno zaproponowanej ważonej mapie dyfuzji [H3, H4]. Wykorzystuje ona ważoną macierz przejść Markowa:

$$M(\mathbf{x}_k, \mathbf{x}_l) \propto r_{kl} \frac{g(\mathbf{x}_k, \mathbf{x}_l)}{[\varrho(\mathbf{x}_k)]^\alpha [\varrho(\mathbf{x}_l)]^\alpha}, \tag{17}$$

z dodatkowymi współczynnikami $r_{kl} \equiv r(\mathbf{x}_k, \mathbf{x}_l)$ do ważenia anizotropowego jądra dyfuzji. Również estymatory gęstości punktowej ϱ są ważone, $\varrho(\mathbf{x}_k) = \sum_l w(\mathbf{x}_l) g(\mathbf{x}_k, \mathbf{x}_l)$. Ogólnie rzecz biorąc, współczynnik ważenia przyjmuje prostą postać [H3, H4]:

$$r(\mathbf{x}_k, \mathbf{x}_l) = w(\mathbf{x}_k) w(\mathbf{x}_l). \tag{18}$$

Można wykazać, że współczynnik ważenia zaproponowany jako ansatz dla macierzy przejścia w MRSE (Rów. 16) można wyprowadzić, przyjmując przybliżenie $\varrho(\mathbf{x}_k) \approx w(\mathbf{x}_l) \sum_l L(\mathbf{x}_k, \mathbf{x}_l)$, które przypomina standardową formułę ważenia z Rów. 4, w granicy quasi-jednorodnego próbkowania. Szczegółowe wyprowadzenie i różne warianty współczynnika ważenia (w zależności od stałej dyfuzji anizotropowej) można znaleźć w Ref. [H3].

Następnie macierz prawdopodobieństwa przejścia M może zostać wykorzystana do rozwiązania problemu wartości i wektorów własnych:

$$M\varphi_k = \lambda_k \varphi_k \tag{19}$$

dla $k = 1, \ldots, K$, gdzie widmo jest określone przez wartości własne $\{\lambda_l\}$. Odpowiadające im prawe wektory własne $\{\varphi_l\}$ można wykorzystać do osadzenia układu w reprezentacji zredukowanej.

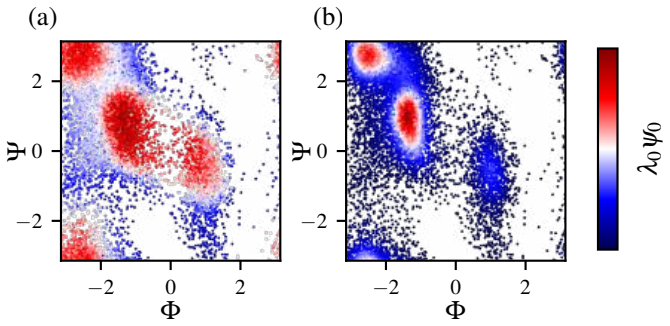
W oparciu o ten rozkład spektralny docelowe odwzorowanie $\xi(\mathbf{x})$ (Rów. 1) można zdefiniować jako *współrzędne dyfuzji* [40, 44]:

$$\mathbf{x} \mapsto \xi(\mathbf{x}) = \big(\lambda_1 \varphi_1(\mathbf{x}), \ldots, \lambda_d \varphi_d(\mathbf{x})\big), \tag{20}$$

gdzie wartości i wektory definiują współrzędne zredukowane. W Rów. 20 każda współrzędna dyfuzji jest zdefiniowana jako $z_k = \lambda_k \varphi_k$, gdzie widmo jest posortowane według nierosnącej wartości:

$$\lambda_0 = 1 > \lambda_1 \geq \cdots \geq \lambda_d \geq \cdots \geq \lambda_K, \tag{21}$$

gdzie d to indeks, przy którym obcinamy współrzędne dyfuzji w Rów. 20 i wymiarowość zredukowanej reprezentacji. W ten sposób dominujące skale czasowe występujące w dynamice układu wielowymiarowego mogą być opisane tylko przez kilka wektorów własnych odpowiadających największym wartościom własnym. Wartość własna $\lambda_0 = 1$ i pierwsza współrzędna dyfuzji $\lambda_0 \varphi_0$ odpowiada rozkładowi równowagowemu danemu przez rozkład Boltzmanna.



Rysunek 5: Ważona mapa dyfuzji. Różnica między gęstościami równowagowymi ($\lambda_0\varphi_0$) dipeptydu alaniny uzyskanymi z obciążonych danych symulacyjnych obliczonych za pomocą (a) standardowej mapy dyfuzji i (b) ważonej mapy dyfuzji. Podczas gdy ważony mapa dyfuzji poprawnie reprezentuje gęstości w stanach metastabilnych, standardowa mapa dyfuzji nie jest w stanie uzyskać poprawnego rozwiązania. W (a) przejścia między stanami są znacznie szybsze ze względu na zastosowanie potencjału obciążającego. Rysunek pochodzi z Ref. [H1].

Zastosowanie ważonej mapy dyfuzji do dipeptydu alaniny zilustrowano na rysunku Rys. 5. Można zauważyć, że gdy ważenie dyfuzji nie jest używane do konstruowania rozkładu równowagowemu z symulacji wzmocnionego próbkowania, prowadzi to do nieprawidłowych wyników. W przeciwieństwie do tego, gdy stosowane jest ważenie dyfuzyjne, rozkład równowagowy odpowiada dynamice nieobciążonej potencjałem zewnętrznym.

Podsumowując, w artykule Ref. [H3] Autor formalnie wprowadził i rozwinął teorię i implementację uczenia z wykorzystaniem ważonych rozmaitości. W przeciwieństwie do Ref. [H2], algorytm ważenia do uczenia zmiennych zbiorowych przy użyciu symulacji wzmocnionego próbkowania został wyprowadzony przy użyciu operatorów Markowa. Autor udowodnił, że w zależności od aproksymacji współczynnika ważenia można uzyskać wyrażenia zaproponowane w Ref. [H2, 47]. Pozwoliło to pokazać, które ważne właściwości (takie jak gęstość, geometria i istotność próbki) zostały zakodowane w macierzach przejścia Markowa używanych przez te metody. Ważenie dyfuzyjne może być stosowane z kilkoma metodami, w tym mapą dyfuzji.

5 Wybór przestrzeni konfiguracyjnej

Chociaż uczenie maszynowe staje się szeroko stosowane do redukcji wymiarowości w symulacjach atomistycznych i ogólnej analizie wielowymiarowych systemów w chemii fizycznej, jakość zmiennych wynikających z takich metod zależy w dużej mierze od danych wejściowych w przestrzeni zmiennych konfigu-

racyjnych. Wybór takich wysokowymiarowych reprezentacji wykorzystywanych następnie do redukcji wymiarowości jest często pomijany.

Zgodnie z hipotezą rozmaitości, zachowanie separacji skal czasowych między powolnymi i szybkimi zmiennymi jest podstawą interpretowalnej konstrukcji reprezentacji. Jak wyjaśniono wcześniej, powolne zmienne są nieodłącznie związane z kinetyką rzadkich przejść między stanami metastabilnymi. Szybkie zmienne są adiabatycznie ograniczone do dynamiki powolnych zmiennych i odpowiadają głównie krótkotrwałej dynamice w stanach metastabilnych. Dlatego możemy uznać różne reprezentacje tego samego systemu za równoważne, jeśli charakteryzuje je ta sama separacja skal czasowych.

Aby oszacować informacje zakodowane w przestrzeni wielowymiarowej, zbieramy N próbek n zmiennych konfiguracyjnych z symulacji, w celu skonstruowania macierzy przejścia Markowa. Zbiór danych składający się z tych próbek jest dany przez:

$$X = \{\mathbf{x}_k \in \mathbb{R}^n, w(\mathbf{x}_k)\}_{k=1}^N, \tag{22}$$

gdzie próbki są powiększone o wagi statystyczne w, jeśli próbkujemy obciążony rozkład z symulacji wzmocnionego próbkowania.

Podobnie jak w przypadku innych metod, ważone jądro anizotropowe ze stałą dyfuzji anizotropowej $\alpha = 1/2$ jest wprowadzane w celu wykorzystania informacji o gęstości i istotności przestrzeni konfiguracyjnej:

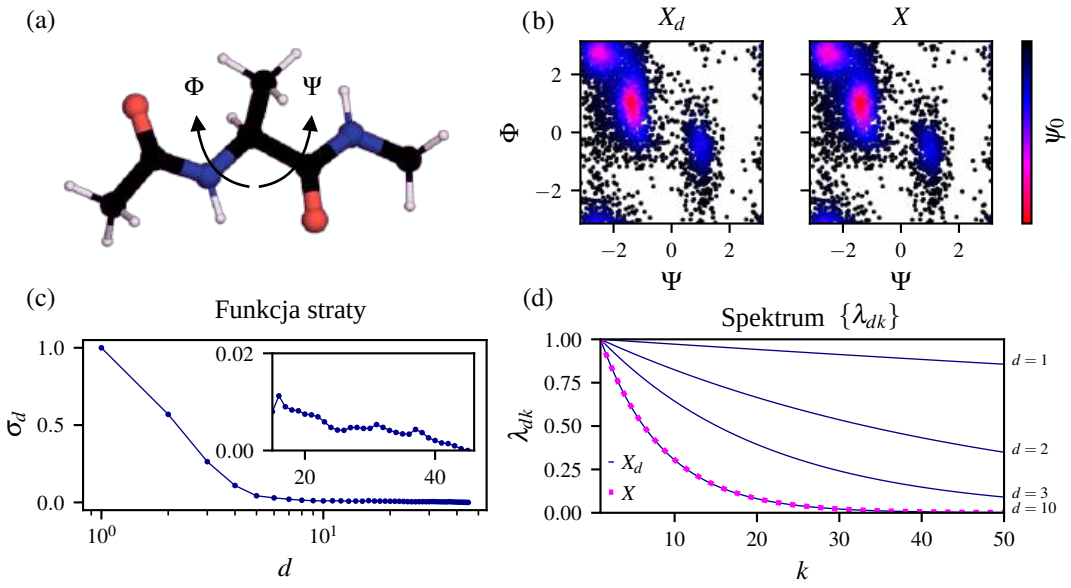
$$L(\mathbf{x}_k, \mathbf{x}_l) = r(\mathbf{x}_k, \mathbf{x}_l) \frac{g(\mathbf{x}_k, \mathbf{x}_l)}{\sqrt{\varrho(\mathbf{x}_k)\varrho(\mathbf{x}_l)}}, \tag{23}$$

gdzie $\varrho(\mathbf{x}) = \sum_k g(\mathbf{x}, \mathbf{x}_k)$ jest do stałej multiplikatywnej szacunkiem gęstości jądra w punkcie $\mathbf{x}$. Jądro anizotropowe jest ważone przy użyciu $r(\mathbf{x}_k, \mathbf{x}_l)$, który jest wprowadzany w celu skorygowania efektu próbkowania z obciążonego rozkładu prawdopodobieństwa (tj. ważenie dyfuzji [H3]). Współczynnik ważenia jest podany jako [H2, H3]:

$$r(\mathbf{x}_k, \mathbf{x}_l) = \sqrt{w(\mathbf{x}_k) w(\mathbf{x}_l)}, \tag{24}$$

gdzie $w(\mathbf{x}_k)$ i $w(\mathbf{x}_l)$ to wagi statystyczne odpowiadające odpowiednio k-tej i l-tej próbce. Dla standardowych symulacji, Rów. 24 redukuje się do anizotropowego jądra dyfuzji używanego w mapie dyfuzji [40, 50].

Aby określić, czy częściowy wybór zmiennych konfiguracyjnych zawiera podobne informacje kinetyczne jak kompletny zestaw, używamy macierzy przejścia Markowa M. Najpierw wykonujemy rozkład spektralny macierzy z pełnego zestawu $M\varphi = \lambda\varphi$ i obliczamy jej wartości $\{\lambda_k\}$ oraz funkcje własne $\{\varphi_k\}$. Wartości własne są posortowane według malejących wartości. Odpowiadające im funkcje własne zawierają informacje o układzie, ponieważ wartości własne są związane z wewnętrznymi skalami czasowymi układu.



Rysunek 6: Wybór wysokowymiarowej reprezentacji. (a) Obciążone dane symulacyjne są wygenerowane poprzez zwiększenie fluktuacji kątów dwuściennych Φ i Ψ dipeptydu alaniny. Początkowa wysokowymiarowa reprezentacja składa się z $n = 45$ odległości między ciężkimi atomami. Ważona mapa dyfuzji jest w stanie wybrać tylko $d = 10$ zmiennych, które zachowują informacje kinetyczne, co widać w (b) zachowanej gęstości równowagi (φ_0), (c) zaniku funkcji straty spektralnej σ_d do 0 oraz (d) równoważnych wartościach własnych. Rysunek pochodzi z Ref. [H4].

Następnie przeprowadzany jest rozkład spektralny dla kombinacji zmiennych konfiguracyjnych, które definiują zbiór danych X_d (d to liczba zmiennych konfiguracyjnych w częściowej reprezentacji) i porównują go z wartościami własnymi kompletnej reprezentacji wielowymiarowej. Aby opisać, ile informacji kinetycznych jest zachowanych w częściowych reprezentacjach, definiujemy funkcję straty spektralnej:

$$\sigma_d = A\left[\sum_k (\lambda_{dk} - \lambda_k)^2\right]^{1/2}, \tag{25}$$

gdzie A jest stałą normalizacji, a λ_k i λ_{dk} są wartościami własnymi, odpowiednio, pełnej reprezentacji wysokowymiarowej i częściowej reprezentacji zmiennych konfiguracyjnych d. Kombinacja zmiennych konfiguracyjnych zachowuje informacje kinetyczne zakodowane w pełnej reprezentacji, jeśli strata widmowa jest pomijalna.

Biorąc pod uwagę zbiór danych X o n zmiennych konfiguracyjnych $\mathbf{x} = (x_1, \ldots, x_n)$ i jego rozkład

spektralny $\{\lambda_k, \varphi_k\}$ powiązanej macierzy przejścia Markowa M, szukamy częściowego, wysokowymiarowego zbioru danych X_d zmiennych konfiguracyjnych d, który po spektralnej dekompozycji jego macierzy przejścia Markowa na $\{\lambda_{dk}, \varphi_{dk}\}$ zawiera podobne informacje kinetyczne jak macierz przejścia Markowa obliczona na podstawie X. Aby uniknąć wyczerpującego i wymagającego obliczeniowo przeszukiwania wszystkich kombinacji zmiennych konfiguracyjnych, używamy algorytmu, który zapewnia suboptymalny wynik [51].

W Rys. 6 pokazujemy przykład, jak wybrać wysokowymiarową reprezentację dipeptydu alaniny. Można zauważyć, że częściowy wybór odległości pomiędzy atomami układu niesie taką samą informację kinetyczną jak pełna reprezentacja. Częściowy wybór odpowiada funkcji straty spektralnej w okolicach 0 oraz temu samemu rozkładowi równowagowemu i wartościom własnym, jak te obliczone na podstawie pełnej reprezentacji.

W pracy Ref. [H4] Autor zaproponował metodę wyboru początkowej reprezentacji wysokowymiarowej do dalszego uczenia zmiennych zbiorowych. Problem konstruowania takiej reprezentacji jest często pomijany podczas uczenia lub wykonywany w sposób, który nie zachowuje ważnych informacji kinetycznych o badanym układzie. Zaproponowana metoda opiera się na równoważności kinetycznej – zachowując skale czasowe obserwowane w pełnej reprezentacji systemu. Ponadto, ponieważ proponowany algorytm opiera się na rozkładzie spektralnym ważonej macierzy przejścia Markowa, może być stosowany zarówno w standardowych symulacjach atomistycznych, jak i symulacjach ze wzmocnionym próbkowaniem.

6 Mapa spektralna

Mapa spektralna jest najnowszym osiągnięciem Autora, wprowadzonym w [H5] i rozszerzonym w [H6]. Technika ta została zaprojektowana tak, aby w szczególności skupić się na konstruowaniu *powolnych* zmiennych zbiorowych, które powstają w wyniku rozdzielenia skal czasowych w badanym układzie. Mapa spektralna łączy w sobie charakterystyki obu klas uczenia odwzorowania docelowego, które zostały wprowadzone w rozdziale 2), tj. wykorzystuje zarówno parametryzowalne mapowania, jak i rozkład spektralny macierzy przejścia Markowa.

Aby oszacować separację między efektywnymi skalami czasowymi charakterystycznymi dla złożonego układu, modelujemy jego zredukowaną dynamikę jako dyskretny łańcuch Markowa przy użyciu funkcji jądra. W celu zmierzenia podobieństwa między próbkami zmiennych zbiorowych $\mathbf{z}_k$ i $\mathbf{z}_l$, używamy jądra Gaussa:

$$g(\mathbf{z}_k, \mathbf{z}_l) = \exp\left(-\frac{1}{\varepsilon_{kl}}\|\mathbf{z}_k - \mathbf{z}_l\|^2\right) \tag{26}$$

gdzie $\|\mathbf{z}_k - \mathbf{z}_l\|$ oznacza odległości euklidesowe między każdą parą próbek $k, l = 1, \ldots, N$, a N to liczba

próbek. Jądro gaussowskie wykazuje pojęcie lokalności poprzez zdefiniowanie sąsiedztwa wokół każdej próbki o promieniu ε_{kl} [40]. Należy zwrócić uwagę, że w przeciwieństwie do mapy dyfuzji, macierz przejścia Markowa w mapie spektralnej konstruowana jest w przestrzeni zredukowanej.

Stany metastabilne często mają charakterystykę multimodalną, co skutkuje przestrzennie niejednorodnymi krajobrazami energii swobodnej. Dlatego, aby poprawić zdolność mapy spektralnej do dostosowania się do stanów metastabilnych, szacujemy zależne od próbki współczynniki skali poprzez adaptacyjne równoważenie skali lokalnej i globalnej:

$$\varepsilon_{kl}(r) = \|\mathbf{z}_k - \eta_r(\mathbf{z}_k)\| \cdot \|\mathbf{z}_l - \eta_r(\mathbf{z}_l)\|, \tag{27}$$

gdzie każdy wyraz definiuje kulę wyśrodkowaną w $\mathbf{z}$ o promieniu $\eta_r(\mathbf{z}) > 0$. Dla wygody definiujemy promień jako ułamek rozmiaru sąsiedztwa $r \in [0, 1]$, co pozwala nam zdecydować, która skala jest bardziej istotna. W szczególności, jądro Gaussa opisuje lokalne sąsiedztwo wokół każdej próbki dla wartości r bliskich 0 (tj. najbliższych sąsiadów), które odpowiadają głębokim i wąskim stanom. Dla wartości r około 1 (najdalsi sąsiedzi), bierze pod uwagę bardziej globalne informacje, odpowiadające płytkim i szerokim stanom. Pośrednie wartości r utrzymują równowagę między skalami przestrzennymi.

Ponieważ krańcowy rozkład równowagowy w przestrzeni zredukowanej jest często daleka od jednorodności dla układów dynamicznych, wprowadzamy jądro zachowującego gęstość dla danych próbkowanych z dowolnego bazowego rozkładu prawdopodobieństwa. W tym celu wykorzystujemy anizotropowe jądro dyfuzji [43]:

$$L(\mathbf{z}_k, \mathbf{z}_l) = \frac{g(\mathbf{z}_k, \mathbf{z}_l)}{\sqrt{\varrho(\mathbf{z}_k)\varrho(\mathbf{z}_l)}}, \tag{28}$$

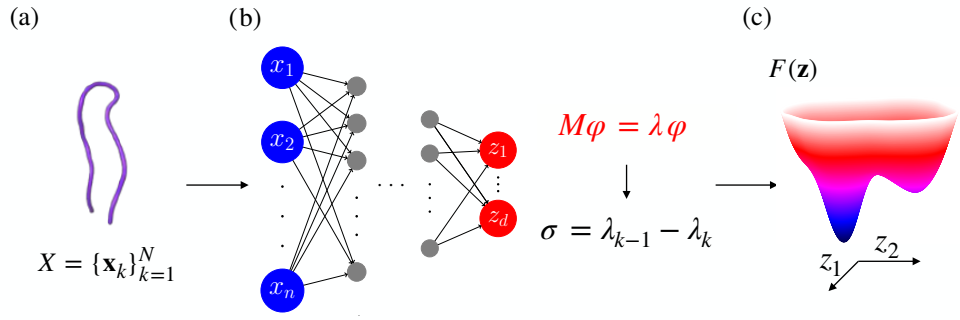
gdzie $\varrho(\mathbf{z}_k) = \sum_l g(\mathbf{z}_k, \mathbf{z}_l)$ jest oszacowaniem gęstości jądra w punkcie $\mathbf{z}_k$. Następnie budujemy macierz przejścia Markowa poprzez normalizację wierszy L:

$$m_{kl} \sim M(\mathbf{z}_k, \mathbf{z}_l) = \frac{L(\mathbf{z}_k, \mathbf{z}_l)}{\sum_n L(\mathbf{z}_k, \mathbf{z}_n)} \tag{29}$$

która modeluje dyskretny łańcuch Markowa $m_{kl} = \Pr\{\mathbf{z}_{\tau+1} = \mathbf{z}_l \mid \mathbf{z}_\tau = \mathbf{z}_k\}$ wyrażający prawdopodobieństwo przejścia między próbkami $\mathbf{z}_k$ i $\mathbf{z}_l$ w pomocniczym (niefizycznym) czasie τ. Ten łańcuch Markowa aproksymuje długoczasową asymptotykę systemu opisując dynamikę za pomocą anizotropowej dyfuzji Fokkera–Plancka [43].

By oszacować dominujące skale czasowe zakodowane w badanym układzie, przeprowadzamy rozkład spektralny macierzy przejścia Markowa:

$$M\varphi_k = \lambda_k \varphi_k, \tag{30}$$



Rysunek 7: Zarys algorytmu mapy spektralnej. (a) Zbiór danych X w wysokowymiarowej reprezentacji $\mathbf{x} = (x_1, \ldots, x_n)$ używany do opisu układu jest traktowany jako dane wejściowe do odwzorowania docelowego. (b) Odwzorowanie docelowe $\mathbf{z} = \xi_w(\mathbf{x})$ jest modelowane jako sieć neuronowa, która osadza system z jego wysokowymiarowej reprezentacji na niskowymiarowej mapie rozpiętej przez powolne zmienne zbiorowe $\mathbf{z} = (z_1, \ldots, z_d)$. Przeprowadzany jest spektralny rozkład macierzy przejścia Markowa zbudowanej z próbek ($M\varphi = \lambda\varphi$). Przerwa spektralna σ jest maksymalizowana w oparciu o różnicę między sąsiednimi wartościami własnymi $\{\lambda_k\}$ w celu oddzielenia powolnych i szybkich skal czasowych. (c) Wyszkolona sieć neuronowa może być wykorzystana do osadzenia wszystkich dostępnych próbek o wysokiej wymiarowości i obliczenia odpowiedniego krajobrazu energii swobodnej $F(\mathbf{z})$. Rysunek pochodzi z Ref. [H5].

gdzie φ_k i λ_k są odpowiednio k-tymi prawymi funkcjami i wartościami własnymi M. Wartości własne M są następujące (posortowane w porządku niemalejącym):

$$\lambda_0 = 1 > \lambda_1 \cdots \geq \lambda_N, \tag{31}$$

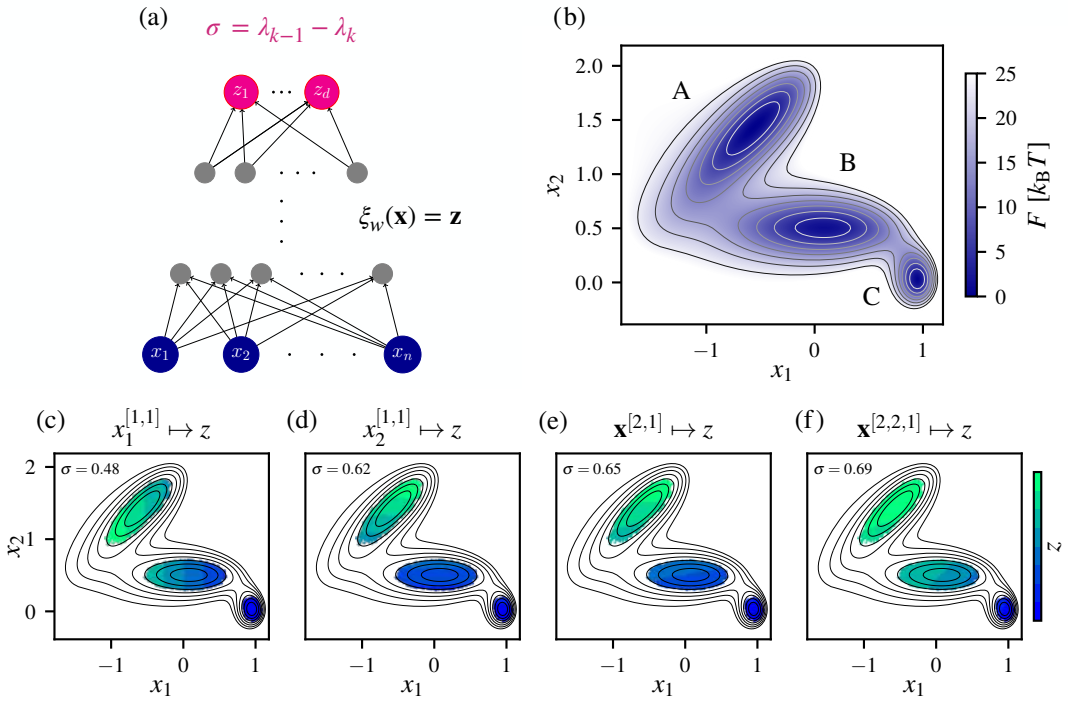
gdzie wartość własna λ_0 odpowiada rozkładowi równowagowemu łańcucha Markowa (Rów. 29) danego przez funkcję własną φ_0. Dominujące wartości własne związane z najwolniejszymi skalami czasowymi relaksacji w układzie można znaleźć, kojarząc każdą wartość własną z efektywną skalą czasową [53]:

$$t_k = -\frac{1}{\log \lambda_k}. \tag{32}$$

Największa przerwa między sąsiednimi wartościami własnymi nazywana jest przerwą spektralną i określa stopień separacji skal czasowych między wolnymi i szybkimi procesami:

$$\sigma = \lambda_{k-1} - \lambda_k, \tag{33}$$

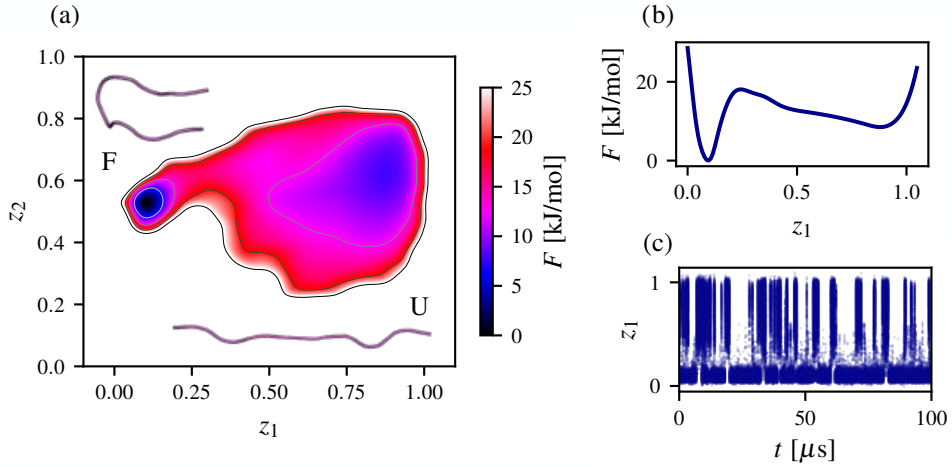
gdzie $k > 0$ oznacza liczbę stanów metastabilnych w przestrzeni zredukowanej [54].



Rysunek 8: Trening odwzorowania docelowego $\xi_w(\mathbf{x}) = \mathbf{z}$ dla trójstanowego potencjału Müllera–Browna. (a) Schematyczny zarys sieci neuronowej. (b) Krajobraz energii swobodnej cząstki poruszającej się w dwuwymiarowej przestrzeni $\mathbf{x} = (x_1, x_2)$ z barierami między stanami około 20 $k_\mathrm{B}T$ i kinetycznym wąskim gardłem między stanami B i C. (c-f) Przykłady odwzorowań docelowych z różnymi architekturami sieci, np., $\mathbf{x}^{[a,\ldots,b]} \mapsto z$ oznacza mapowanie zmiennej $\mathbf{x}$ przez sieć składającą się z warstw $[a, \ldots, b]$ do zmiennej zbiorowej z. Rysunek pochodzi z Ref. [H6].

Teoria spektralnej charakterystyki stanów metastabilnych (prace Gaveau i Schulmana [54, 55]) wyjaśnia, że przerwa spektralna i stopień degeneracji w widmie wartości własnych są związane z separacją skal czasowych i dynamiką Markowa. Jeśli wartość własna macierzy przejścia Markowa jest prawie zdegenerowana $k + 1$ razy, oznacza to, że rozkład równowagowy rozpada się na k stanów metastabilnych z rzadkimi przejściami między nimi. Odwrotna relacja jest również prawdziwa: jeśli gęstość równowagowa rozpada się na stany metastabilne oddzielone barierą energii swobodnej znacznie większą niż energia termiczna $k_\mathrm{B}T$, występuje degeneracja wartości własnej.

Aby osiągnąć zredukowaną dynamikę, która jest efektywnie markowowska, kluczowe jest posiadanie przerwy spektralnej między sąsiednimi wartościami własnymi, wraz z bliską degeneracją dominującej

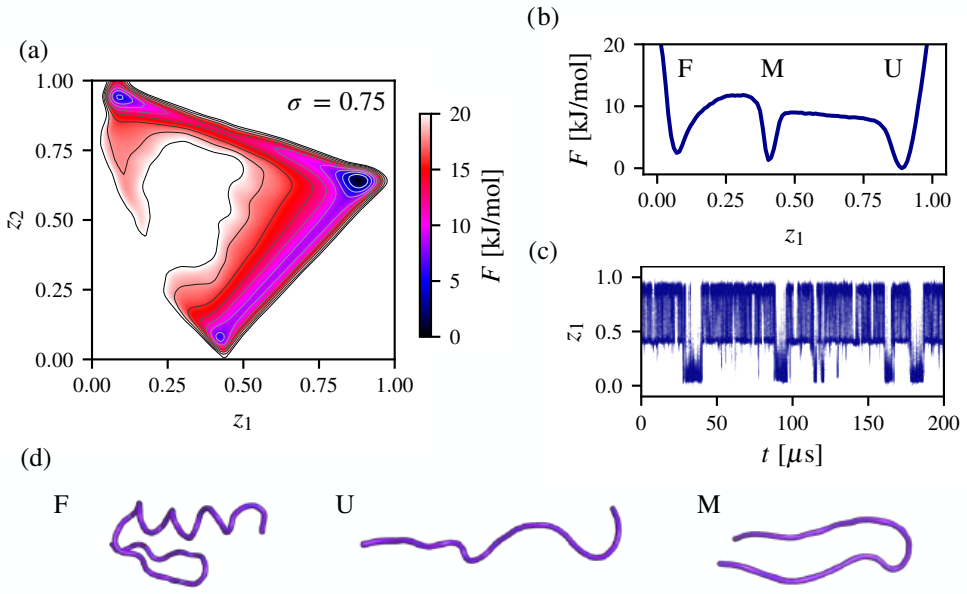


Rysunek 9: Uczenie powolnych zmiennych zbiorowych dla odwracalnego zwijania białka CLN025 próbkowanego przez 100 μ symulację dynamiki molekularnej. Zbiór danych pochodzi z Ref. [52]. (a) Mapa spektralna i odpowiadający jej krajobraz energii swobodnej F CLN025 pokazujący dwa odrębne baseny energii swobodnej: główny i dobrze zdefiniowany stan pofałdowany (F) i bardziej luźno ustrukturyzowany stan metastabilny (U), oddzielone barierą energii swobodnej około 20 kJ/mol. Przedstawiono reprezentatywne konformacje ze stanów metastabilnych. Współczynnik kształtu osi jest zachowany. (b) Profil energii swobodnej pokazany jako funkcja z_1 CV. (c) Szereg czasowy dla z_1 CV pokazujący przejścia między stanami. Rysunek pochodzi z Ref. [H6].

wartości własnej. Warunek ten jest niezbędny, aby maksymalna przerwa spektralna w k prowadziła do rozdzielenia na k stanów metastabilnych, co może pomóc zmniejszyć efekty pamięciowe w dynamice. W związku z tym rozważamy przerwę spektralną jako funkcję oceny w mapie spektralnej.

Ponieważ macierz przejścia Markowa jest szacowana w przestrzeni zredukowanej, rozważamy efektywną dynamikę, a nie dynamikę mikroskopowych współrzędnych układu. W ten sposób łańcuch Markowa zdefiniowany w przestrzeni zredukowanej jest domyślnie modelowany jako podążający za tłumioną dynamiką Langevina z gęstością równowagową krańcową $p(\mathbf{z})$. Ogólnie rzecz biorąc, ta efektywna dynamika jest albo niemarkowowska, albo ma macierz dyfuzji zależną od $\mathbf{z}$, co oznacza, że nie jest napędzana wyłącznie przez krajobraz energii swobodnej. Jednak wybierając zmienne zbiorowe, które wynikają z separacji skal czasowych w układzie jako przestrzeń zredukowaną, możemy reprezentować dynamikę współrzędnych mikroskopowych poprzez efektywną dynamikę powolnych zmiennych zbiorowych, która jest w przybliżeniu markowowska [56].

Schematyczna ilustracja dla mapy spektralnej jest pokazana w Rys. 7. Po pierwsze, zbiór danych



Rysunek 10: Mapa spektralna i krajobraz energii swobodnej zwijania białka BBA obliczone na podstawie zbioru danych z symulacji dynamiki molekularnej 200-μs w temperaturze 325 K. Wysokowymiarowa reprezentacja dana przez $n = 378$ odległości euklidesowych pomiędzy atomami Cα BBA. (a) Krajobraz energii swobodnej pokazujący stany metastabilne obejmowane przez zmienne zbiorowe obliczone za pomocą mapy spektralnej dla $k = 3$, gdzie przerwa spektralna osiąga $\sigma = 0.75$. (b) Profil energii swobodnej wzdłuż zmiennej z_1. (c) Trajektoria z_1 pokazująca zmiany między stanami metastabilnymi podczas symulacji dynamiki molekularnej. (c) Reprezentatywne konformacje białka BBA odpowiadające stanowi zwiniętemu (F), stanowi rozwiniętemu (U) i stanowi błędnie zwiniętemu (M). Rysunek pochodzi z Ref. [H5].

X w wysokowymiarowej reprezentacji $\mathbf{x} = (x_1, \ldots, x_n)$ używany do opisu systemu jest pobierany jako dane wejściowe do odwzorowania docelowego. Następnie odwzorowanie docelowe $\mathbf{z} = \xi_w(\mathbf{x})$ jest modelowane jako sieć neuronowa, która osadza system w jego wysokowymiarowej reprezentacji na niskowymiarowej mapie rozpiętej przez powolne zmienne zbiorowe $\mathbf{z} = (z_1, \ldots, z_d)$. Następnie przeprowadzany jest rozkład spektralny macierzy przejścia Markowa ($M\varphi = \lambda\varphi$). Przerwa spektralna σ jest maksymalizowana w oparciu o różnicę między sąsiednimi wartościami własnymi $\{\lambda_k\}$ w celu oddzielenia powolnych i szybkich skal czasowych. Wreszcie, wytrenowana sieć neuronowa może być wykorzystana do obliczenia energii swobodnej $F(\mathbf{z})$. Przykładowe zastosowanie mapy spektralnej do oszacowania powolnych zmiennych zbiorowych i powiązanego krajobrazu energii swobodnej pokazano na Rys. 8 (potencjał Mullera–Browna), 9 (chignolina) i 10 (białko BBA).

W artykułach Ref. [H5, H6] autor zaproponował metodę uczenia powolnych zmiennych zbiorowych, które opisują najważniejsze stopnie swobody lub mody układu, ponieważ odpowiadają przejściom w układzie występującym na dłuższych i eksperymentalnych skalach czasowych. Mapa spektralna opiera się na obliczaniu jąder anizotropowych i wykonywaniu rozkładów spektralnych przy jednoczesnym zapewnieniu, że luka spektralna, która mierzy separację w skali czasowej między wolną i szybką kinetyką w układzie, jest zmaksymalizowana. Mapa spektralna może pracować z wysokowymiarowymi reprezentacjami setek zmiennych bez wstępnego przetwarzania, zwykle wymaganego w przypadku metod uczenia nienadzorowanego. Dane i kod wymagane do odtworzenia wyników tych prac są dostępne pod adresem https://zenodo.org/records/10678142.

7 PLUMED: Promowanie przejrzystości i reprodukowalności

PLUMED to otwarcie dostępna biblioteka typu open source, która zapewnia różne metody, takie jak wzmocnione algorytmy próbkowania, metody szacowania energii swobodnej i narzędzia do analizy ogromnych ilości danych generowanych przez symulacje dynamiki molekularnej. Techniki te można łączyć z dużym zestawem innych narzędzi wykorzystywanych w fizyce, chemii, materiałoznawstwie i biologii. PLUMED może być zintegrowany z większością silników symulacyjnych.

W 2019 r. utworzono konsorcjum PLUMED, którego celem jest wspieranie przejrzystości i odtwarzalności wyników w symulacjach atomistycznych [H7]. Konsorcjum to obraca się wokół biblioteki PLUMED. Konsorcjum PLUMED to otwarta społeczność składająca się z obecnych programistów, współpracowników i wszystkich tych badaczy, których praca opiera się na symulacjach atomistycznych i napędza rozwój i rozpowszechnianie ich wyników. Głównym celem konsorcjum jest ustanowienie skuteczniejszych protokołów udostępniania informacji w symulacjach dynamiki molekularnej, promowanie odtwarzalności naukowej i utrzymywanie najwyższych standardów badawczych.

Konsorcjum PLUMED wprowadziło bazę danych PLUMED-NEST (https://www.plumed-nest.org/), które jest publicznym repozytorium użytkowników PLUMED. Zapewnia ona wszystkie dane potrzebne do odtworzenia wyników symulacji (lub analizy) dynamiki molekularnej. Ponadto, PLUMED-NEST monitoruje zgodność dostarczonych plików wejściowych do PLUMED z aktualnymi i rozwojowymi wersjami kodu. Odkąd PLUMED-NEST został otwarty, Autor zdeponował wiele danych, implementacji i wyników swoich prac.[2]

[2]Szukaj “rydzewski” na https://www.plumed-nest.org/browse.html

8 Krótkie podsumowanie

Symulacje atomistyczne, takie jak dynamika molekularna lub Monte Carlo, stały się ogólnymi metodami badania układów dynamicznych w fizyce, chemii i biologii. Takie symulacje oferują szczegółowy wgląd w procesy na poziomie mikroskopowym z większą dokładnością przestrzenno-czasową niż eksperymenty. Jednak analiza układów składających się z tysięcy atomów może stanowić wyzwanie. Aby uzyskać uproszczoną, bardziej zrozumiałą reprezentację, często konieczne jest skonstruowanie niskowymiarowej reprezentacji, która uchwyci istotne cechy fizyczne. Dostarczenie teorii i technik, które mogą wykonać to zadanie bez nadzoru, pozwala złagodzić problem polegania na doświadczeniu oraz próbach i błędach.

Osiągnięcie jest cennym interdyscyplinarnym dodatkiem do współczesnej fizyki statystycznej. Poniższa lista podsumowuje najważniejsze części Osiągnięcia według Autora:

- Opracowanie spójnych podstaw teorii i metod statystycznego uczenia zmiennych zbiorowych na podstawie symulacji standardowych i wzmocnionego próbkowania, które można zastosować do zrozumienia dowolnego procesu molekularnego w eksperymentalnych skalach czasowych w sposób oparty na danych.
- W przeciwieństwie do wielu wcześniej zaproponowanych metod, podstawy te uwzględniają najważniejsze cechy fizyczne procesów dynamicznych. Obejmują one rozkład prawdopodobieństwa (równowagowy lub nierównowagowy) próbkowany przez badany układ, pojęcie odległości między konfiguracjami oraz powolną kinetykę, która jest kluczem do zrozumienia procesów na dłuższych skalach czasowych.
- Teoria leżąca u podstaw proponowanych metod łączy ogólne idee i narzędzia mechaniki statystycznej i uczenia maszynowego. Zapewnia to zastosowanym tu metodom uczenia maszynowego zdolność, której brakuje w większości nienadzorowanych metod redukcji wymiarowości: interpretowalność w kontekście fizyki.
- Proponowane podejście jest ogólne i umożliwia społeczności łatwe jego rozszerzanie, co podkreśla jego potencjał do szerokiego zastosowania i dalszego udoskonalania.

Choć metody opracowane na styku fizyki statystycznej i uczenia maszynowego dopiero zaczynają się pojawiać i być wykorzystywane, z pewnością można zauważyć znaczący postęp. Chociaż problem konstruowania powolnych zmiennych zbiorowych z symulacji atomistycznych jest daleki od ostatecznego rozwiązania, Autor jest przekonany, że opracowane i wdrożone przez niego spójne podejście, opisane w tym Osiągnięciu, będzie szeroko stosowane przez społeczność i jest ważnym krokiem w lepszym zrozumieniu złożonych układów i szerokiego zakresu procesów fizycznych, takich jak zmiany konformacyjne białek podczas zwijania lub wiązania z lekami, kataliza, przejścia fazowe w szkle lub krystalizacja.

Literatura

H1. ***Rydzewski, J.**, Chen, M. & Valsson, O. Manifold Learning in Atomistic Simulations: A Conceptual Review. *Mach. Learn.: Sci. Technol.* **4,** 031001 (2023).

H2. ***Rydzewski, J.** & Valsson, O. Multiscale Reweighted Stochastic Embedding: Deep Learning of Collective Variables for Enhanced Sampling. *J. Phys. Chem. A* **125,** 6286–6302 (2021).

H3. ***Rydzewski, J.**, Chen, M., Ghosh, T. K. & Valsson, O. Reweighted Manifold Learning of Collective Variables from Enhanced Sampling Simulations. *J. Chem. Theory Comput.* **18,** 7179–7192 (2022).

H4. ***Rydzewski, J.** Selecting High-Dimensional Representations of Physical Systems by Reweighted Diffusion Maps. *J. Phys. Chem. Lett.* **14,** 2778–2783 (2023).

H5. ***Rydzewski, J.** Spectral Map: Embedding Slow Kinetics in Collective Variables. *J. Phys. Chem. Lett.* **14,** 5216–5220 (2023).

H6. ***Rydzewski, J.** & Gökdemir, T. Learning Markovian Dynamics with Spectral Maps. *J. Chem. Phys.* **160** (2024).

H7. PLUMED Consortium, Promoting Transparency and Reproducibility in Enhanced Molecular Simulations. *Nat. Methods* **16,** 670–673 (2019).

1. Chandler, D. *Introduction to Modern Statistical Mechanics* (Oxford University Press, Oxford, UK, 1987).
2. Battimelli, G., Battimelli, G., Ciccotti, G., Greco, P. & Scalone. *Computer Meets Theoretical Physics* (Springer, 2020).
3. Piccini, G. *et al.* Ab Initio Molecular Dynamics with Enhanced Sampling in Heterogeneous Catalysis. *Catal. Sci. Technol.* **12,** 12–37 (2022).
4. Baron, R. & McCammon, J. A. Molecular Recognition and Ligand Association. *Annu. Rev. Phys. Chem.* **64,** 151–175 (2013).
5. Bruce, N. J., Ganotra, G. K., Kokh, D. B., Sadiq, S. K. & Wade, R. C. New Approaches for Computing Ligand–Receptor Binding Kinetics. *Curr. Opin. Struct. Biol.* **49,** 1–10 (2018).
6. Bernetti, M., Masetti, M., Rocchia, W. & Cavalli, A. Kinetics of Drug Binding and Residence Time. *Annu. Rev. Phys. Chem.* **70,** 143–171 (2019).

7. Wolf, S. Predicting Protein–Ligand Binding and Unbinding Kinetics with Biased MD Simulations and Coarse-Graining of Dynamics: Current State and Challenges. *J. Chem. Inf. Model.* (2023).

8. O'Hagan, M. P., Haldar, S., Morales, J. C., Mulholland, A. J. & Galan, M. C. Enhanced Sampling Molecular Dynamics Simulations Correctly Predict the Diverse Activities of a Series of Stiff-Stilbene G-Quadruplex DNA Ligands. *Chem. Sci.* **12,** 1415–1426 (2021).

9. Van Speybroeck, V., Vandenhaute, S., Hoffman, A. E. J. & Rogge, S. M. J. Towards Modeling Spatiotemporal Processes in Metal–Organic Frameworks. *Trends Chem.* **3,** 605–619 (2021).

10. Neha, Tiwari, V., Mondal, S., Kumari, N. & Karmakar, T. Collective Variables for Crystallization Simulations–from Early Developments to Recent Advances. *ACS Omega* **8,** 127–146 (2023).

11. Aussems, D. U. B., Bal, K. M., Morgan, T. W., Van De Sanden, M. C. M. & Neyts, E. C. Atomistic Simulations of Graphite Etching at Realistic Time Scales. *Chem. Sci.* **8,** 7160–7168 (2017).

12. Coifman, R. *Harmonic Analytic Geometry in High Dimensions—Empirican Models* International Conference of Mathematicians. Lecture. 2018.

13. Hohenberg, P. C. & Krekhov, A. P. An Introduction to the Ginzburg–Landau Theory of Phase Transitions and Nonequilibrium Patterns. *Phys. Rep.* **572,** 1–42 (2015).

14. Zwanzig, R. Memory Effects in Irreversible Thermodynamics. *Phys. Rev.* **124,** 983 (1961).

15. Luttinger, J. Theory of Thermal Transport Coefficients. *Physical Review* **135,** A1505 (1964).

16. Mori, H. Transport, Collective Motion, and Brownian Motion. *Prog. Theor. Phys.* **33,** 423–455 (1965).

17. Wu, H. & Noé, F. Variational Approach for Learning Markov Processes from Time Series Data. *J. Nonlinear Sci.* **30,** 23–66. ISSN: 1432-1467 (2020).

18. Brunton, S. L., Budišić, M., Kaiser, E. & Kutz, J. N. Modern Koopman theory for dynamical systems. *arXiv preprint arXiv:2102.12086* **64,** 229–340 (2021).

19. Izenman, A. J. Introduction to Manifold Learning. *Wiley Interdiscip. Rev. Comput. Stat.* **4,** 439–446 (2012).

20. Tenenbaum, J. B., De Silva, V. & Langford, J. C. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* **290,** 2319–2323 (2000).

21. Roweis, S. T. & Saul, L. K. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* **290,** 2323–2326 (2000).

22. Valsson, O., Tiwary, P. & Parrinello, M. Enhancing Important Fluctuations: Rare Events and Metadynamics from a Conceptual Viewpoint. *Annu. Rev. Phys. Chem.* **67,** 159–184 (2016).

23. Swendsen, R. H. & Wang, J.-S. Replica Monte Carlo Simulation of Spin-Glasses. *Phys. Rev. Lett.* **57,** 2607 (1986).

24. Earl, D. J. & Deem, M. W. Parallel Tempering: Theory, Applications, and New Perspectives. *Phys. Chem. Chem. Phys.* **7,** 3910–3916 (2005).

25. Chen, M., Cuendet, M. A. & Tuckerman, M. E. Heating and Flooding: A Unified Approach for Rapid Generation of Free Energy Surfaces. *J. Chem. Phys.* **137,** 024102 (2012).

26. Valsson, O. & Parrinello, M. Variational Approach to Enhanced Sampling and Free Energy Calculations. *Phys. Rev. Lett.* **113,** 090601 (2014).

27. Reinhardt, M. & Grubmüller, H. Determining Free-Energy Differences Through Variationally Derived Intermediates. *J. Chem. Theory Comput.* **16,** 3504–3512 (2020).

28. Torrie, G. M. & Valleau, J. P. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comp. Phys.* **23,** 187–199 (1977).

29. Mezei, M. Adaptive Umbrella Sampling: Self-Consistent Determination of the Non-Boltzmann Bias. *J. Comput. Phys.* **68,** 237–248 (1987).

30. Laio, A. & Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. U.S.A.* **99,** 12562–12566 (2002).

31. Barducci, A., Bussi, G. & Parrinello, M. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys. Rev. Lett.* **100,** 020603 (2008).

32. Maragakis, P., van der Vaart, A. & Karplus, M. Gaussian-Mixture Umbrella Sampling. *J. Phys. Chem. B* **113,** 4664–4673. ISSN: 1520-5207 (2009).

33. Morishita, T., Itoh, S. G., Okumura, H. & Mikami, M. Free-Energy Calculation via Mean-Force Dynamics using a Logarithmic Energy Landscape. *Phys. Rev. E* **85,** 066702 (2012).

34. Invernizzi, M., Piaggi, P. M. & Parrinello, M. Unified Approach to Enhanced Sampling. *Phys. Rev. X* **10,** 041034 (2020).

35. Hénin, J., Lelièvre, T., Shirts, M. R., Valsson, O. & Delemotte, L. Enhanced Sampling Methods for Molecular Dynamics Simulations. *arXiv preprint arXiv:2202.04164* **4,** 1583 (2022).

36. Belkin, M. & Niyogi, P. *Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering* in *Adv. Neural Inf. Process. Syst.* **14** (MIT Press, 2001), 585–591.

37. Hinton, G. E. & Roweis, S. *Stochastic Neighbor Embedding* in *Adv. Neural Inf. Process. Syst.* (eds Becker, S., Thrun, S. & Obermayer, K.) **15** (MIT Press, 2002), 833–864.

38. Belkin, M. & Niyogi, P. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Comput.* **15,** 1373–1396 (2003).

39. Hashemian, B., Millán, D. & Arroyo, M. Modeling and Enhanced Sampling of Molecular Systems with Smooth and Nonlinear Data-Driven Collective Variables. *J. Chem. Phys.* **139,** 12B601_1 (2013).

40. Coifman, R. R. *et al.* Geometric Diffusions as a Tool for Harmonic Analysis and Structure Definition of Data: Diffusion Maps. *Proc. Natl. Acad. Sci. U.S.A.* **102,** 7426–7431 (2005).

41. Van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9,** 2579–2605 (2008).

42. Aflalo, Y. & Kimmel, R. Spectral Multidimensional Scaling. *Proc. Natl. Acad. Sci. U.S.A.* **110,** 18052–18057 (2013).

43. Nadler, B., Lafon, S., Coifman, R. R. & Kevrekidis, I. G. Diffusion Maps, Spectral Clustering and Reaction Coordinates of Dynamical Systems. *Appl. Comput. Harmon. Anal.* **21,** 113–127 (2006).

44. Coifman, R. R. & Lafon, S. Diffusion Maps. *Appl. Comput. Harmon. Anal.* **21,** 5–30 (2006).

45. Jones, P. W., Maggioni, M. & Schul, R. Manifold Parametrizations by Eigenfunctions of the Laplacian and Heat Kernels. *Proc. Natl. Acad. Sci. U.S.A.* **105,** 1803–1808 (2008).

46. Van der Maaten, L. Learning a Parametric Embedding by Preserving Local Structure. *J. Mach. Learn. Res.* **5,** 384–391 (2009).

47. Zhang, J. & Chen, M. Unfolding Hidden Barriers by Active Enhanced Sampling. *Phys. Rev. Lett.* **121,** 010601 (2018).

48. Hinton, G. E. & Salakhutdinow, R. R. Reducing the Dimensionality of Data with Neural Networks. *Science* **313,** 504–507 (2006).

49. Coifman, R. R., Kevrekidis, I. G., Lafon, S., Maggioni, M. & Nadler, B. Diffusion Maps, Reduction Coordinates, and Low Dimensional Representation of Stochastic Systems. *Multiscale Model. Simul.* **7,** 842–864 (2008).

50. Singer, A., Erban, R., Kevrekidis, I. G. & Coifman, R. R. Detecting Intrinsic Slow Variables in Stochastic Dynamical Systems by Anisotropic Diffusion Maps. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 16090–16095 (2009).

51. Pudil, P., Novovičová, J. & Kittler, J. Floating Search Methods in Feature Selection. *Pattern Recognit. Lett.* **15,** 1119–1125 (1994).

52. Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. How Fast-Folding Proteins Fold. *Science* **334,** 517–520 (2011).

53. Bovier, A., Eckhoff, M., Gayrard, V. & Klein, M. Metastability and Low Lying Spectra in Reversible Markov Chains. *Commun. Math. Phys.* **228,** 219–255 (2002).

54. Gaveau, B. & Schulman, L. S. Theory of Nonequilibrium First-Order Phase Transitions for Stochastic Dynamics. *J. Math. Phys.* **39,** 1517–1533 (1998).

55. Gaveau, B. & Schulman, L. S. Master Equation based Formulation of Nonequilibrium Statistical Mechanics. *J. Math. Phys.* **37,** 3897–3932 (1996).

56. Tiwary, P. & Berne, B. J. Spectral Gap Optimization of Order Parameters for Sampling Complex Molecular Systems. *Proc. Natl. Acad. Sci. U.S.A.* **113,** 2839 (2016).

Inne osiągnięcia naukowe

Poza publikacjami naukowymi uwzględnionymi w Osiągnięciu, od czasu uzyskania stopnia doktora, Autor zajmował się również problematyką oddziaływań ligand-białko, tj., rozwojem technik wzmocnionego próbkowania w celu obserwacji ścieżek dysocjacji ligandów z białek [A1, A2, A3], wzmocnienie fotodynamiki kanonicznego bakteriofytochromu przy użyciu zaawansowanych metod dynamiki molekularnej [A4], agregacja α-synuklein [A5] oraz badanie oporności na pyretroidy w kontroli komarów [A6].

A1. ***Rydzewski, J.** maze: Heterogeneous Ligand Unbinding along Transient Protein Tunnels. *Comp. Phys. Commun.* **247,** 106865 (2020).

A2. ***Rydzewski, J.** & Valsson, O. Finding Multiple Reaction Pathways of Ligand Unbinding. *J. Chem. Phys.* **150** (2019).

A3. ***Rydzewski, J.**, Jakubowski, R., Nowak, W. & Grubmuller, H. Kinetics of Huperzine A Dissociation from Acetylcholinesterase via Multiple Unbinding Pathways. *J. Chem. Theory Comput.* **14,** 2843–2851 (2018).

A4. ***Rydzewski, J.**, Walczewska-Szewc, K., Czach, S., Nowak, W. & Kuczera, K. Enhancing the Inhomogeneous Photodynamics of Canonical Bacteriophytochrome. *J. Phys. Chem. B* **126,** 2647–2657 (2022).

A5. Walczewska-Szewc, K., **Rydzewski, J.** & Lewkowicz, A. Inhibition-Mediated Changes in Prolyl Oligopeptidase Dynamics Possibly Related to α-Synuclein Aggregation. *Phys. Chem. Chem. Phys.* **24,** 4366–4373 (2022).

A6. Niklas, B., **Rydzewski, J.**, Lapied, B. & Nowak, W. Toward Overcoming Pyrethroid Resistance in Mosquito Control: The Role of Sodium Channel Blocker Insecticides. *Int. J. Mol. Sci.* **24,** 10334 (2023).

V. Informacja o wykazywaniu się istotną aktywnością naukową realizowaną w więcej niż jednej uczelni, instytucji naukowej lub instytucji kultury, w szczególności zagranicznej

Badania naukowe Autora w europejskich instytucjach zagranicznych były prowadzone w:

1. National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japonia, grupa prof. Tetsuyi Morishity, XI 2023–II 2024.

2. Department of Theoretical and Computational Biophysics, Max Planck Institute for Biophysical Chemistry (od 2022 Max Planck Institute for Multidisciplinary Sciences), Getynga, Niemcy,

grupa prof. Helmuta Grubmüllera, X 2016–IV 2017.

3. Department of Chemistry and Applied Biosciences, Swiss Federal Institute of Technology in Zürich c/o Institute of Computational Science, Università della Svizzera italiana, Lugano, Szwajcaria, grupa of prof. Michele Parrinello, VII 2016–X 2016.

Trwające projekty prowadzone na więcej niż jednym uniwersytecie:

1. Prof. Omar Valsson, Uniwersytet North Texas, Denton, US, 2016–. Prace związane z konstruowaniem zmiennych zbiorowych dla symulacji wzmocnionego próbkowania z wykorzystaniem optymalizacji i wyjaśnialnego uczenia maszynowego. Kontynuacja prac przedstawionych w Refs. [H1, H2, H3, A2].

2. Prof. Ming Chen, Uniwersytet Purdue, West Lafayette, USA, 2018–. Prace związane z konstruowaniem nieobciążonych macierzy przejścia Markowa na podstawie symulacji z rozszerzonym próbkowaniem. Kontynuacja prac przedstawionych w Refs. [H1, H3].

3. Prof. Alexander M. Berezhkovski, National Institutes of Health, Bethesda, Maryland, US, 2023–. Praca związana z dynamiką Markowa wzdłuż współrzędnych reakcji.

4. Prof. Michele Parrinello, Technology, Genua, Włochy, 2016–. Prace związane z opracowaniem ulepszonych metod próbkowania do symulacji wiązania ligandów z białek i ustanowieniem protokołów dla przejrzystych i powtarzalnych symulacji atomistycznych. Kontynuacja prac przedstawionych w Ref. [H7].

5. Prof. Helmut Grubmüller, Max Planck Insitutue for Multidisciplinary Sciences, Getynga, Niemcy, 2016–. Prace związane z oszacowaniem kinetyki i termodynamiki wiązania inhibitorów z enzymów. Kontynuacja prac przedstawionych w Ref. [A3].

6. Prof. Tetsuya Morishita, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japonia, 2022–. Prace związane z obliczaniem makroskopowych zmiennych opisujących tworzenie się szkła.

7. Prof. Yasuteru Shigeta i prof. Ryuhei Harada, Uniwersytet w Tsukubie, Japonia, 2019–. Prace związane z wdrożeniem metod rozszerzonego próbkowania bez obciążenia we wtyczce PLUMED [H7].

8. Prof. Bruno Lapied, Uniwersytet w Angers, Francja, 2023–. Prace związane z obliczeniowym badaniem oporności na pyretroidy w kontroli komarów. Kontynuacja prac przedstawionych w Ref. [A6].

VI. Informacja o osiągnięciach dydaktycznych, organizacyjnych oraz popularyzujących naukę

Dydaktyka:

Kursy oznaczone przez "(koordynator)" zostały zaproponowane, opracowane i koordynowane przez Autora. Począwszy od 2022 r. Autor pracuje w pełnym wymiarze godzin na stanowisku badawczym w ramach grantu Sonata NCN i w związku z tym nie prowadzi żadnych kursów.

- Metody numeryczne: 2014–2015, 2018–2020
- Radiomika (koordynator): 2019–2022
- Programowanie w języku Python (koordynator): 2018–2022
- Podstawy fizyki obliczeniowej (koordynator): 2018–2022
- Matematyka dyskretna: 2018-2022
- Algorytmy stochastyczne (koordynator): 2016–2017
- Programowanie proceduralne: 2015–2022
- Elementy bioinformatyki: 2015–2016
- Języki programowania: 2014–2017
- Problemy etyki, otwartej nauki i otwartych innowacji: 2022

Organizacja:

- Organizacja konferencji Bioinformatyka w Toruniu organizowanej wspólnie przez Polskie Towarzystwo Bioinformatyczne (PTBI) i Uniwersytet Mikołaja Kopernika w Toruniu, 2014–.
- Recenzent konkursu Polskiego Towarzystwa Bioinformatycznego na najlepszą pracę licencjacką z bioinformatyki i biologii obliczeniowej, 2022.
- Recenzje dla Proceedings of Annual Conference on Machine Learning, Optimization and Data Science (LOD), 2018-2020.

Studenci:

- Promotor pomocniczy doktorantów: Tugce Gokdemir (nauki fizyczne; 2023–; finansowany z grantu Sonata uzyskanego przez Autora), Sylwia Czach (nauki fizyczne; 2020–).
- Promotor magistrantów: Wojciech Amtmański (fizyka medyczna; 2020-2021), Karolina Kolonko (informatyka stosowana; 2021-2022), Patryk Tajs (informatyka stosowana; 2023-2024).

- Promotor licencjatów: Aleksander Oskroba (fizyka; 2019–2021).
- Promotor inżynierantów: Aleksandra Warmbier (informatyka stosowana; 2019–2022), Kajetan Krzewina (informatyka stosowana; 2019–), Bartosz Jagodziński (informatyka stosowana; 2020–2022), Patryk Tajs (informatyka stosowana; 2021–2023), Jacek Wierzejewski (informatyka stosowana; 2021–2023).

Popularyzacja:

- Wykład inauguracyjny dla studentów I roku Wydziału Fizyki, Astronomii i Informatyki Stosowanej Uniwersytetu Mikołaja Kopernika w Toruniu, październik 2018 r.
- Metody wzmocnionego próbkowania opublikowane w Głosie Uczelni 7-10, 2019, dystrybuowane przez Uniwersytet Mikołaja Kopernika w Toruniu.
- Notatka popularyzatorska dotycząca promocji grantów oferowanych przez Narodowe Centrum Nauki, 10-lecie Narodowego Centrum Nauki, 2021 r.
- Letni obóz dla studentów zagranicznych i doktorantów w ramach programu Narodowej Agencji Wymiany Akademickiej (NAWA) SPINAKER - Intensywne Międzynarodowe Programy Edukacyjne, 12-14 lipca 2022 r. w Toruniu, Polska.
- Promowanie otwartej nauki jako główny wykonawca projektu Narodowej Agencji Wymiany Akademickiej "Open NCU–Open Source, Open Science " poprzez finansowanie opłat za publikacje w otwartym dostępie dla młodych naukowców na Uniwersytecie Mikołaja Kopernika w Toruniu.
- Prowadzenie zajęć na temat otwartej nauki w ramach programu "Developing and Implementing hands-on training on Open Science and Open Innovation for Early Career Researchers (DIOSI)" finansowanego przez Unię Europejską w ramach programu Horyzont 2020, Komisja Europejska 101006318, 2020-2023.

Summary of Professional Achievements

(As of May 22, 2024)

I. Name and Surname

Jakub Rydzewski

II. Diplomas, Degrees Conferred in Specific Areas of Science

- **2014–2018**: PhD in Physical Sciences (Biophysics), Faculty of Physics, Astronomy and Informatics, Nicolaus Copernicus University, Torun, Poland, Thesis: *Rare-Event Sampling of Ligand Transport in Proteins* (with honors), Supervised by Prof. W. Nowak
- **2013–2014**: MSc in Informatics, Faculty of Physics, Astronomy and Informatics, Nicolaus Copernicus University, Torun, Poland
- **2011–2014**: BSc in Theoretical Physics, Faculty of Physics, Astronomy and Informatics, Nicolaus Copernicus University, Torun, Poland
- **2009–2013**: BSc Eng. in Informatics, Faculty of Physics, Astronomy and Informatics, Nicolaus Copernicus University, Torun, Poland

III. Information on Employment in Research Institutes or Faculties

- **Since X 2019**: Institute of Physics, Nicolaus Copernicus University, Torun, Poland (assistant professor)
- **X 2018–X 2019**: Institute of Physics, Nicolaus Copernicus University, Torun, Poland (assistant)
- **XI 2023–II 2024**: National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan, Prof. Tetsuya Morishita's group
- **X 2016–IV 2017**: Department of Theoretical and Computational Biophysics, Max Planck Institute for Biophysical Chemistry (since 2022 Max Planck Institute for Multidisciplinary Sciences), Gottingen, Germany, Prof. Helmut Grubmüller's group
- **VII 2016–X 2016**: Department of Chemistry and Applied Biosciences, Swiss Federal Institute of Technology (ETH) in Zürich c/o Institute of Computational Science, Università della Svizzera italiana, Lugano, Switzerland, Prof. Michele Parrinello's group

IV. Description of the Achievement, set out in art. 219 para 1 point 2 of the Act

Jakub Rydzewski (referred to as "the Author"), presents a main scientific achievement (referred as to "the Achievement") in accordance to the *Act on Scientific Degrees and Titles* as a series of publications [H1, H2, H3, H4, H5, H6, H7] on developing techniques for statistical learning of slow collective variables from atomistic simulations. The title of the Achievement is:

Learning Collective Variables from Atomistic Simulations

In the following publications included in the Achievement, excluding the last, the Author acts as the first and corresponding author (indicated by an asterisk).

H1. ***Rydzewski, J.**, Chen, M. & Valsson, O. Manifold Learning in Atomistic Simulations: A Conceptual Review. *Mach. Learn.: Sci. Technol.* **4,** 031001 (2023).

H2. ***Rydzewski, J.** & Valsson, O. Multiscale Reweighted Stochastic Embedding: Deep Learning of Collective Variables for Enhanced Sampling. *J. Phys. Chem. A* **125,** 6286–6302 (2021).

H3. ***Rydzewski, J.**, Chen, M., Ghosh, T. K. & Valsson, O. Reweighted Manifold Learning of Collective Variables from Enhanced Sampling Simulations. *J. Chem. Theory Comput.* **18,** 7179–7192 (2022).

H4. ***Rydzewski, J.** Selecting High-Dimensional Representations of Physical Systems by Reweighted Diffusion Maps. *J. Phys. Chem. Lett.* **14,** 2778–2783 (2023).

H5. ***Rydzewski, J.** Spectral Map: Embedding Slow Kinetics in Collective Variables. *J. Phys. Chem. Lett.* **14,** 5216–5220 (2023).

H6. ***Rydzewski, J.** & Gökdemir, T. Learning Markovian Dynamics with Spectral Maps. *J. Chem. Phys.* **160** (2024).

H7. PLUMED Consortium, Promoting Transparency and Reproducibility in Enhanced Molecular Simulations. *Nat. Methods* **16,** 670–673 (2019).

The Achievement is a valuable interdisciplinary addition to modern statistical physics. The following list summarizes the most important parts of the Achievement according to the Author:

- Developing a self-consistent framework of theory and methods for the statistical learning of collective variables from standard and enhanced sampling simulations, which can be applied to understand any molecular process on the experimental timescales in a data-driven manner.

- The framework takes into account the most important physical characteristics of dynamical processes, unlike many previously proposed methods. These include the probability distribution (equilibrium or nonequilibrium) sampled by the studied system, a notion of distance between the configurations, and slow kinetics, which is key to understanding processes on longer timescales.
- The theory underlying the proposed framework combines general ideas and concepts from statistical mechanics and machine learning. This provides the machine learning methods used here with an ability that is lacking in most unsupervised dimensionality reduction methods: interpretability in the context of physics.
- The proposed framework is general and enables an easy extension by the community, which underlines its potential for widespread use and further improvement.

This text serves as a summary and a guideline for the reviewers without delving into the specifics; more detailed information can be found in the publications forming the Achievement.

1 Background

Atomistic simulations, such as molecular dynamics or Monte Carlo, are commonly used in physics, chemistry, and biology to explore intricate dynamical systems [1, 2]. These simulations provide detailed information about processes at the microscopic level with greater spatiotemporal accuracy than experiments. To such processes we can include, for instance, catalysis [3], ligand interactions with proteins [4–7] and DNA [8], glass transitions in amorphous materials [9], crystallization [10], and graphite etching [11]. However, analyzing systems consisting of thousands of atoms can be challenging. To obtain a simplified, more understandable representation, it is often necessary to average over noisy variables to arrive at a low-dimensional representation that captures essential characteristics. As R. Coifman [12] stated:

> There is inherent truth in the data, and we seek to characterize some latent physical variable that intrinsically describes the changes of states.

Many methods have been introduced to alleviate the apparent problem of high dimensionality, such as the Ginzburg–Landau theory of phase transitions [13], the Mori–Zwanzig formalism for transport and collective motion [14–16], and Koopman's theory [17, 18]. More recently developed approaches include manifold learning [H3, 19], a class of nonlinear unsupervised statistical learning methods trained directly on data, whose development was instigated by the innovative works of Tenenbaum et al. [20] and Roweis and Saul [21]; both published in the same issue of Science [**290** (2000)].

A simplified description of the dynamics of complex many-body systems can be achieved with statistical

mechanics, where the microscopic coordinates comprise the configuration space of the system (i.e., atom positions). Such a representation involves a high number of degrees of freedom. Let us suppose that the system is represented by an n-dimensional vector of *configuration variables*, $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. Generally, the configuration variables are functions of the microscopic coordinates, assuming that the space spanned by these variables is high-dimensional. In machine learning, the configurational variables are often referred to as *features* or *descriptors*. Then, a dataset X of K high-dimensional samples of the configuration variables recorded at consecutive times during the dynamics can be expressed as a matrix of size $n \times K$ called a *trajectory*.

Without loss of generality and for presentation purposes, we limit our discussion to the canonical ensemble (NVT), in which the configuration variables evolve according to a high-dimensional potential energy function $U(\mathbf{x})$. When the microscopic coordinates represent the system, its equilibrium density is given by the stationary Boltzmann distribution [1], $p(\mathbf{x}) = \mathrm{e}^{-\beta U(\mathbf{x})}/\mathcal{Z}$, where $\beta = (k_\mathrm{B}T)^{-1}$ is the inverse of the thermal energy $k_\mathrm{B}T$ corresponding to the temperature T with the Boltzmann constant denoted by k_B, and $\mathcal{Z} = \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta U(\mathbf{x})}$ is the canonical partition function. Otherwise, the set X is sampled according to an unknown high-dimensional equilibrium density.

In statistical physics, we can simplify the description of a system by averaging over its certain high-dimensional properties. This results in a macroscopic description that uses fewer degrees of freedom to characterize ensembles of microscopic configurations or states. In atomistic simulations, these simplified variables are often called collective variables (CVs), order parameters, or reaction coordinates. Identifying CVs is challenging for complex systems and often involves resorting to physical or chemical intuition and trial-and-error approaches [22]. Relying solely on intuition or trial and error to identify CVs can be unsystematic and hinder our understanding of the underlying physical process often contributing to erroneously estimated kinetics. This can lead to:

1. Overlapping metastable states, which results in the underestimation of free-energy barriers, inaccurate determination of transition state ensembles, and inefficiency of enhanced sampling techniques due to the existence of hidden bottlenecks [23].

2. Inability to extract the behavior of the process on longer timescales (e.g., mixing slow and fast variables), and thus considerable non-Markovian effects that should then be additionally accounted for using a generalized Langevin equation with a memory kernel as in the Mori–Zwanzig formalism [14, 16].

CVs are typically expressed as functions of the configuration variables, meaning that finding CVs involves obtaining a set of functions that embed high-dimensional samples into a low-dimensional CV

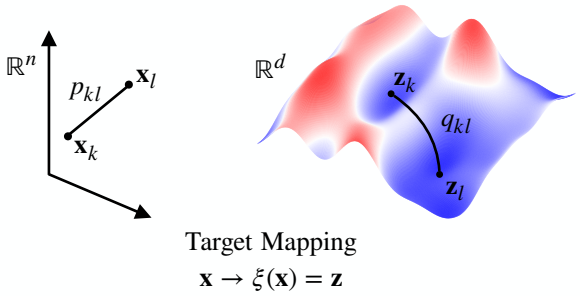


Figure 1: Target mapping. Schematic illustration of mapping $\xi(\mathbf{x}) = \mathbf{z}$ between the high-dimensional configuration space $\mathbf{x}$ and the low-dimensional CV space $\mathbf{z}$ ($n \gg d$). The relation p_{kl} between configuration samples $\mathbf{x}_k$ and $\mathbf{x}_l$ is conserved in the relation q_{kl} between CV samples $\mathbf{z}_k$ and $\mathbf{z}_l$. Figure from Ref. [H1].

space. We call this set of functions the *target mapping* (Fig. 1):

$$\mathbf{x} \mapsto \xi(\mathbf{x}) \equiv \left\{\xi_k(\mathbf{x})\right\}_{k=1}^{d} \tag{1}$$

where $d \ll n$. The target mapping $\xi(\mathbf{x})$ can be linear, nonlinear, or even an identity function (i.e., this reduces the problem to selection). Each statistical learning method provides a unique functional form of the target mapping used to reduce the dimensionality of the system representation. From now on, we will refer to the low-dimensional representation of the system as to *the reduced space* or simply CVs. To define a probability density for CVs expressed by the target mapping (Eq. 1), we consider only a part of the configuration space. The equilibrium distribution of CVs is obtained by averaging over unused variables. This gives us a marginal density:

$$p(\mathbf{z}) = \int \mathrm{d}\mathbf{x}\, \delta(\mathbf{z} - \xi(\mathbf{x}))p(\mathbf{x}). \tag{2}$$

which typically contains several disconnected high-probability states separated by low-probability regions, leading to metastability (Fig. 2). Instead of the potential energy function $U(\mathbf{x})$ characteristic for a rugged high-dimensional representation, the reduced dynamics of the system in the CV space follows the underlying *free-energy landscape*. We write it as the negative logarithm of the marginal distribution of CVs multiplied by the thermal energy:

$$F(\mathbf{z}) = -\frac{1}{\beta} \log p(\mathbf{z}) \tag{3}$$

which is defined up to an immaterial constant. The equilibrium density of CVs can be equivalently written as $p(\mathbf{z}) = \mathrm{e}^{-\beta F(\mathbf{z})} / \mathcal{Z}_V$, where the partition function in the CV space is given as $\mathcal{Z}_V = \int \mathrm{d}\mathbf{z}\, \mathrm{e}^{-\beta F(\mathbf{z})}$.

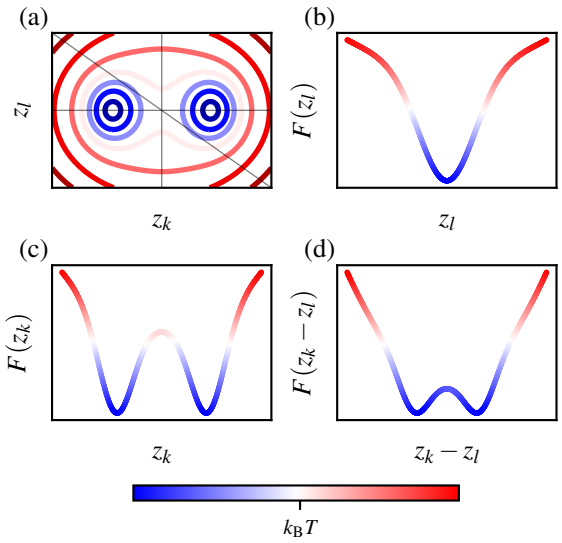


Figure 2: Metastability and CVs. (a) Model free-energy landscape with two metastable states separated by a barrier higher than the thermal energy. The ability of the one-dimensional CVs to discriminate between states: (b) projection along the z_l variable shows a single state, (c) along the z_k shows the correct landscape, (d) along $z_k - z_l$ shows two states separated by an underestimated energy barrier ($< k_\mathrm{B}T$). Figure from Ref. [H1].

Sampling free-energy landscapes exhaustively is challenging, even for simple systems. On the timescales accessible for standard atomistic simulations (around milliseconds), crossings over high free-energy barriers are rare events. As a result, the system remains kinetically trapped in a metastable state as its dynamics is restricted to sampling fast equilibration. Enhanced sampling methods can alleviate the sampling problem and overcome kinetic bottlenecks. Over recent years, several such enhanced sampling algorithms have been developed, including tempering [24–26], variational [27, 28], biasing [29–34] approaches, or combinations of these [35]. For a comprehensive review and classification of these methods, we refer to an article by Henin et al. [36].

As a representative example of enhanced sampling techniques, we consider methods employing external (i.e., nonphysical) bias potential to enhance CV fluctuations artificially. The first approach of this kind, called umbrella sampling [29], was introduced in 1977 by Torrie and Valleau[1]. When the bias potential is introduced to the system, the distribution of CVs can deviate significantly from equilibrium. This

[1]The motivation for naming the method "umbrella sampling" was to highlight the method's versatility to investigate a wide range of physical processes.

results in sampling according to a biased distribution that is, by design, easier to sample. Another popular technique from this category is metadynamics [31, 32], where the bias potential is constructed by depositing Gaussians in the CV space. The side effect of enhancing the fluctuations of CVs is a significant deviation from the equilibrium density. To extract the equilibrium properties (e.g., free-energy landscapes, kinetics) from biased simulations, each sample is given a statistical weight to account for the effect of the biasing:

$$w(\mathbf{z}) \propto \frac{p(\mathbf{z})}{q(\mathbf{z})}, \tag{4}$$

where $q(\mathbf{z})$ is the biased density in the CV space. Standard reweighting involves employing the weights to find the stationary equilibrium distribution from the biased CV distribution, which can be computed by histogramming or kernel density estimation.

2 Statistical Learning of Collective Variables

Statistical learning is a comprehensive field that includes methods for simplifying high-dimensional data to low-dimensional manifolds using nonlinear mappings. Such methods called manifold learning, build on linear methods like principal component analysis (PCA) or singular value decomposition (SVD), which are commonly applied in data analysis. Recently, manifold learning has gained attention in atomistic simulations, particularly in extracting physical properties from complex systems. In this context, a manifold refers to the reduced space accurately defined by a few CVs.

According to the manifold hypothesis, manifold learning assumes that the dynamics in a high-dimensional space can be accurately represented by a reduced and smooth subspace known as a manifold. This low-dimensional description is attributed to the coupling between various degrees of freedom, resulting in a limited number of slowly evolving variables that govern the dynamics. Fast degrees of freedom are controlled by the dynamics of these slower variables, leading to an adiabatic timescale separation. This approach enables the modeling of complex systems as diffusion processes, with stochastic differential equations employed to describe the slow variables. At the same time, the fast degrees of freedom are represented as thermal noise.

Manifold learning methods incorporate a notion of similarity between high-dimensional samples, usually through a distance metric [20, 21, 37–43]. The distances are subsequently integrated into a global parameterization of the data using a discrete Markov chain, where the similarities depend on distances between the samples. For example, a common starting point is the construction of the Markov chain based on an anisotropic diffusion kernel [44] that we use frequently throughout the text:

$$L(\mathbf{x}_k, \mathbf{x}_l) = \frac{g(\mathbf{x}_k, \mathbf{x}_l)}{[\varrho(\mathbf{x}_k)]^\alpha [\varrho(\mathbf{x}_l)]^\alpha}, \tag{5}$$

where $g(\mathbf{x}_k, \mathbf{x}_l) = \exp(-\|\mathbf{x}_k - \mathbf{x}_l\|^2/\varepsilon)$ is a Gaussian kernel with a scale constant ε, $\varrho(\mathbf{x}_k) = \sum_n g(\mathbf{x}_k, \mathbf{x}_n)$ is a pointwise kernel density estimate at $\mathbf{x}_k$, and α is an anisotropic diffusion constant. Then, the anisotropic diffusion kernel is row-normalized to represent Markov probabilities:

$$p_{kl} \sim M(\mathbf{x}_k, \mathbf{x}_l) = \frac{L(\mathbf{x}_k, \mathbf{x}_l)}{\sum_n L(\mathbf{x}_k, \mathbf{x}_n)}, \tag{6}$$

which contain information about the transition from $\mathbf{x}_k$ to $\mathbf{x}_l$. Under this view, M denotes a Markov chain with the transition probability from $\mathbf{x}_k$ to $\mathbf{x}_l$.

The anisotropic diffusion constant is related to the data density [45]. In the limit of $\varepsilon \to 0$ and the infinite number of samples $K \to \infty$, we can consider the following:

1. $\alpha = 0$: we recover dynamics according to the potential $2U(\mathbf{x})$ and the density $\propto [p(\mathbf{x})]^2$ for the classical normalized graph Laplacian [37, 39, 46].

2. $\alpha = 1$: we get the graph Laplacian with data uniformly distributed on a manifold. This normalization accounts only for the data geometry, while density does not play a role.

3. $\alpha = \frac{1}{2}$: we obtain dynamics according to the underlying potential $U(\mathbf{x})$ and the density $\propto p(\mathbf{x})$ whose eigenfunctions capture the long-time asymptotics of data (i.e., correspond to slow variables).

The third option is the most important case for us, as we deal with dynamical systems that evolve according to the potential $U(\mathbf{x})$ and are interested in the slowest modes of such systems. For the anisotropic diffusion constant $\alpha = \frac{1}{2}$, we asymptotically recover the long-time dynamics of the system whose microscopic coordinates are sampled from the Boltzmann distribution. The related overdamped Langevin equation is [45]:

$$\dot{\mathbf{x}} = -\beta \nabla U(\mathbf{x}) + \sqrt{2}\eta(t), \tag{7}$$

where η is an n-dimensional Brownian motion. The evolution in time of $\mathbf{x}$ leads to the backward Fokker-Plank equation, which describes the infinitesimal generator $\mathcal{L}$ of this diffusion process:

$$\mathcal{L} = \mathrm{e}^{\beta U(\mathbf{x})} \nabla \, \mathrm{e}^{-\beta U(\mathbf{x})} \nabla \tag{8}$$

whose eigenvalues and eigenvectors determine the kinetic information of the diffusion process and can be used to parametrize the reduced space. As the generator $\mathcal{L}$ usually has several dominant eigenvalues for metastable systems, finding an approach to extract the generator from discrete datasets can be used to construct *slow* CVs.

To systematically introduce each method developed by the Author, we start by classifying the target mapping from the high-dimensional configuration space to the reduced space (or CVs). Under this

framework, finding CVs is equivalent to finding an optimal parametrization of the target mapping, thereby merging concepts from statistical physics and unsupervised learning.

The target mapping performs dimensionality reduction such that the dimensionality of the reduced space is much lower than that of the high-dimensional space, i.e., $d \ll n$. In the context of atomistic simulations, this process can generally be enclosed in the following steps:

1. Generation of high-dimensional samples from unbiased or biased atomistic simulations.
2. Construction of a Markov chain on the data with pairwise transition probabilities between samples.
3. Parametrization of CVs using a mapping that embeds high-dimensional samples to the reduced space.

This learning approach can be divided into two categories, depending on how the reduced space is constructed by the target mapping:

I. *Divergence optimization* where a divergence (i.e., a statistical distance between a pair of probability distributions) between the Markov transition matrix M built from high-dimensional samples and a Markov transition matrix $Q(\mathbf{z}_k, \mathbf{z}_l)$, constructed from low-dimensional samples, is minimized. As such, these methods can match Markov transition matrices in both spaces, enabling the preservation of physical information by excluding spectral decomposition. The target mapping expressed as a parametrizable embedding is:

$$\xi_w(\mathbf{x}) = \left\{\xi_k(\mathbf{x}; w)\right\}_{k=1}^{d}, \tag{9}$$

where w are parameters that are varied such that the divergence between M and Q is minimized. In such methods, M is fixed while Q is estimated by the parametrized target mapping. Depending on the manifold learning method used, the minimization can be performed differently, i.e., gradient descent [42], or stochastic gradient descent if the target mapping is represented by a neural network [H2, 47, 48].

II. *Eigendecomposition* (i.e., spectral decomposition) of the Markov transition matrix:

$$M\varphi_k = \lambda_k \varphi_k, \tag{10}$$

where $\{\psi_k\}$ and $\{\lambda_k\}$ are the corresponding eigenfunctions and eigenvalues, respectively. The solution of Eq. 10 determines the reduced space [41]. For instance, the target mapping can be parametrized as follows:

$$\xi(\mathbf{x}) = \left\{\lambda_k \varphi_k(\mathbf{x})\right\}_{k=1}^{d}, \tag{11}$$

where the k-th coordinate is $\lambda_k \varphi_k(\mathbf{x})$. The eigenvalues are sorted in non-increasing order and include only d dominant eigenvalues as each corresponds to the importance of respective coordinates spanned by eigenfunctions.

The eigenvalues decrease exponentially and can be related to the effective timescales of the studied physical process, as multiple timescales frequently characterize complex systems. As such, the dominant eigenvalues also correspond to the slowest processes.

This unified framework, as introduced in Ref. [H1], can be used to classify every technique developed by the Author encompassing the Achievement. Additionally, it has been used to classify methods for constructing CVs, not designed by the Author and thus not mentioned in the Achievement. The review [H1] is written from the perspective of atomistic simulations, which required reframing the considered methods, and extending them to include learning for unbiased and *enhanced sampling* simulations.

3 Multiscale Reweighted Stochastic Embedding

Multiscale reweighted stochastic embedding (MRSE) [H2] is a recent technique developed by the Author to construct low-dimensional CVs from standard and *enhanced sampling* atomistic simulations. Note that manifold learning had not been extended to construct CVs from enhanced sampling simulations; this development has recently been introduced in Ref. [H2].

MRSE does not perform an eigendecomposition of Markov transition matrices to find CVs, instead focusing on preserving transition probabilities from a matrix by enforcing matching between them in both high- and reduced spaces. In MRSE, the target mapping ξ is parametrized by universal approximators, known as neural networks, to perform nonlinear dimensionality reduction. The target mapping is given as:

$$\mathbf{z} : \mathbf{x} \mapsto \xi_{\boldsymbol{w}}(\mathbf{x}), \tag{12}$$

where w are parameters of the target mapping adjusted such that the reduced space is optimal with respect to a selected statistical measure. Note that in some simple cases, the mapping in Eq. 12 can also be represented using a linear combination. However, deep learning has been successful in a broad range of learning problems, and using more intricate approximations for the mapping between high-dimensional and low-dimensional spaces is quite common for complex data sets [47, 49].

Let us consider the construction of the Markov transition matrix from configuration samples. For this, MRSE builds a Laplacian matrix to encode information about geometry in transition probabilities m_{kl}:

$$m_{kl} \sim M(\mathbf{x}_k, \mathbf{x}_l) \propto L(\mathbf{x}_k, \mathbf{x}_l) \overset{\alpha=0}{=} G(\mathbf{x}_k, \mathbf{x}_l), \tag{13}$$

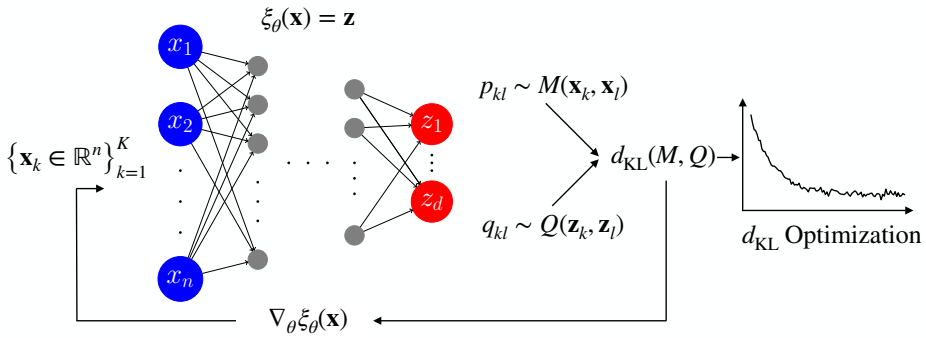


Figure 3: Multiscale reweighted stochastic embedding. Schematic depiction of learning parameters θ for a parametric target mapping represented by a neural network. The backpropagation procedure estimates errors of the parametric target mapping $\nabla_\theta \xi_\theta(\mathbf{x})$ which are used to correct the parameters so that the Kullback–Leibler divergence (or any other divergence) calculated from the Markov transition matrix $M(\mathbf{x}_k, \mathbf{x}_l)$ and $Q(\mathbf{z}_k, \mathbf{z}_l)$ decreases to zero. A minimum value of the Kullback–Leibler divergence indicates that the relations between samples in the configuration and reduced spaces are preserved. Figure from Ref. [H1].

where we skip the row normalization for brevity. In MRSE, the kernel L with the anisotropic diffusion constant $\alpha = 0$ (data density is not considered) is simply the Gaussian kernel G. Generally, the Markov transition matrix used by MRSE is constructed as a mixture of Gaussian with different scale parameters [H2]. However, for the simplicity of our presentation, we described M by a single Gaussian.

Next, a one-dimensional t-distribution [42, 47] is taken to represent transition probabilities in the reduced space:

$$q_{kl} \sim Q(\mathbf{z}_k, \mathbf{z}_l) \propto \left(1 + \|\xi_{\boldsymbol{w}}(\mathbf{x}_k) - \xi_{\boldsymbol{w}}(\mathbf{x}_l)\|^2\right)^{-1}. \tag{14}$$

The choice of the t-distribution for Q in MRSE is motivated by the apparent crowding problem [42] where samples in the reduced space cannot be separated if the distribution does not have a long tail (e.g., Gaussian).

Finally, the Markov transition matrices computed from the high-dimensional and low-dimensional samples must be compared. The most common choice for such a metric is employing a statistic distance, particularly the Kullback–Leibler divergence:

$$D_{\mathrm{KL}}(M, Q; w) = \sum_k \sum_l m_{kl} \log\left(\frac{m_{kl}}{q_{kl}}\right), \tag{15}$$

where in contrast to the standard formulation of the Kullback–Leibler divergence that compares two

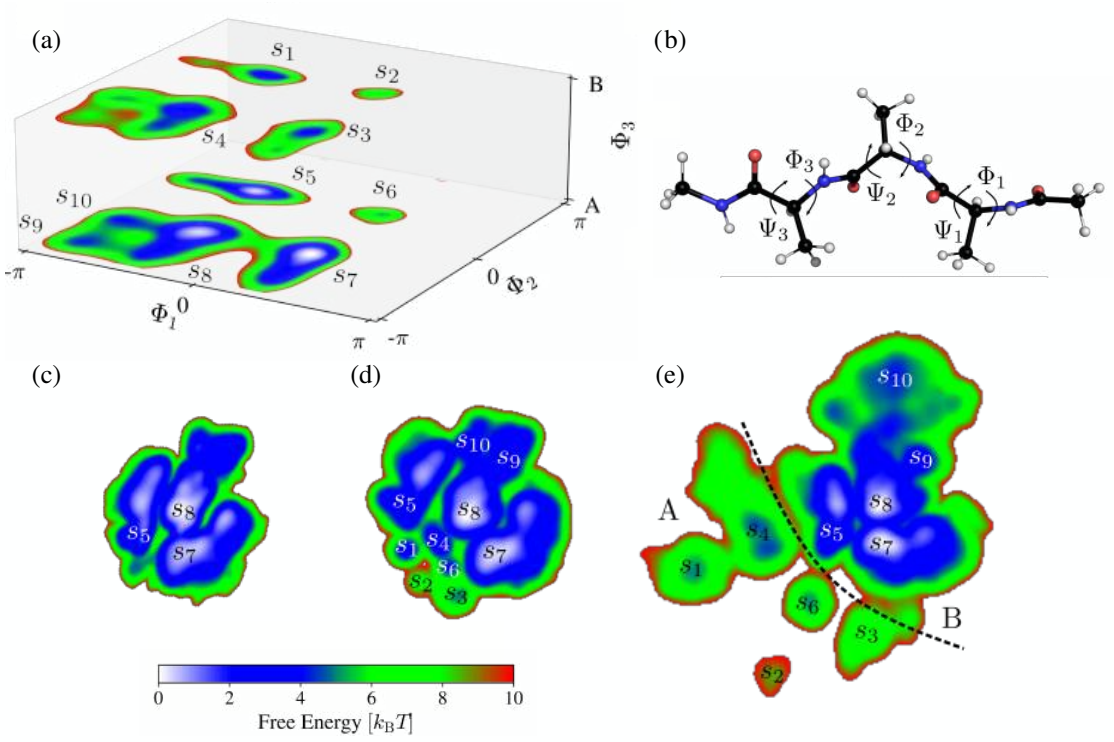


Figure 4: Free-energy landscape constructed by MRSE. (a) Metastable states that can be observed by increasing the fluctuations of Φ_1, Φ_2, and Φ_3 dihedral angles or alanine tetrapeptide. (b) All the dihedral angles of alanine tetrapeptide are used to describe its configuration space. Free-energy landscapes spanned by CVs constructed using MRSE using (c) unbiased simulation data, (d) biased data without reweighting, and (e) biased data with reweighting. Figure from Ref. [H2].

probability distributions, Eq. 15 is computed for every pair of rows from M and Q and then summed. The Kullback–Leibler divergence optimization is performed to train the target mapping represented by a neural network. As the target mapping is parametric, the gradients of D_{KL} with respect to the parameters w can be estimated using backpropagation. A simplified scheme of MRSE presented in Fig. 3.

It is important to note that Eq. 13 can be used when MRSE constructs CVs from unbiased simulations. Applying this Markov transition matrix to model biased simulations would result in reduced space spanned by CVs, which is also biased, i.e., its geometry and density. To enable learning from biased data, we introduce the following ansatz reweighting factor, which multiplies the Markov transition

matrix:

$$r(\mathbf{x}_k, \mathbf{x}_l) = \sqrt{w(\mathbf{x}_k)w(\mathbf{x}_l)}, \tag{16}$$

which is written as a geometric mean between two statistical weights. It can be justified by the fact that the bias potential is additive, and thus, a geometric mean is appropriate to preserve this relation. As such, the Markov transition matrix is reweighted before normalization, which takes into account the effect of biasing. Eq. 16, introduced in Ref. [H2] without any derivation, will be formally justified in Ref. [H3].

An application to alanine tetrapeptide and a demonstration showing how the reweighting influences the CVs are shown in Fig. 4. We can see that the best separation between the metastable states can be observed for the case of using the reweighting to unbias the Markov transition matrix constructed from biased data.

In summary, in Ref. [H2], the Author proposed and developed a method to learn the reduced space and represent it as CVs from standard atomistic and enhanced sampling simulations. MRSE is constructed such that the Laplacians constructed from high- and low-dimensional samples are iteratively improved by the target mapping represented by a neural network. In MRSE, the Author introduced the concept of using a reweighting ansatz to account for learning CVs from enhanced sampling simulations using statistical weights. This work, for the first time, showed that using standard manifold learning to estimate CVs from enhanced sampling simulations is ill-posed as the resulting reduced space is biased in geometry, density, and sample importance. All data, inputs, and the implementation of the lowlearner module for PLUMED, with which the results were obtained, are openly available from `https://doi.org/10.5281/zenodo.4756093`.

4 Reweighted Diffusion Map

Diffusion map was inspired mainly by Laplacian eigenmap [37, 39] that has theoretical convergence guarantees as the discrete operator approaches the Laplacian on the underlying manifold assuming the data are *uniformly* sampled. Diffusion map proposed by Coifman et al. [41] expands the concept of Laplacian eigenmap. This algorithm yields a family of embeddings even when the data are *non-uniformly* sampled. Compared to other manifold learning methods, diffusion map has a substantial theoretical background [41, 45, 50] that shows that slow CVs spanning the low-dimensional manifold can be constructed by diffusion map so that they correspond to the slowest relaxation processes given by a probability distribution evolving under a random walk over the data [44].

The standard diffusion map can construct CVs only from unbiased atomistic simulations (without enhanced sampling). A non-trivial approach for incorporating statistical sample weights is necessary

to learn from enhanced sampling simulations where sampling follows a biased probability distribution. Here, we focus on the recently proposed reweighted diffusion map, which implements *diffusion reweighting* to unbias Markov transitions [H3, H4]. It employs a weighted Markov transition matrix:

$$M(\mathbf{x}_k, \mathbf{x}_l) \propto r_{kl} \frac{g(\mathbf{x}_k, \mathbf{x}_l)}{[\varrho(\mathbf{x}_k)]^\alpha [\varrho(\mathbf{x}_l)]^\alpha}, \tag{17}$$

that uses a pairwise reweighting factor $r_{kl} \equiv r(\mathbf{x}_k, \mathbf{x}_l)$ to reweight the anisotropic diffusion kernel and each pointwise density estimate ϱ is also reweighted with weights [e.g., $\varrho(\mathbf{x}_k) = \sum_l w(\mathbf{x}_l) g(\mathbf{x}_k, \mathbf{x}_l)$]. Generally, the pairwise reweighting factor takes the following simple form [H3, H4]:

$$r(\mathbf{x}_k, \mathbf{x}_l) = w(\mathbf{x}_k) w(\mathbf{x}_l). \tag{18}$$

It can be shown that the reweighting factor proposed as an ansatz for unbiasing transition matrices in MRSE (Eq. 16) can be derived by taking the approximation such as $\varrho(\mathbf{x}_k) \approx w(\mathbf{x}_l) \sum_l L(\mathbf{x}_k, \mathbf{x}_l)$, which resembles standard reweighting formula from Eq. 4, and in the limit of quasi-uniform sampling. For a detailed derivation and different variants of the reweighting factor (depending on the anisotropic diffusion constant), we refer to Ref. [H3].

Next, the transition probability matrix M can be used to solve the eigenvalue problem:

$$M\varphi_k = \lambda_k \varphi_k \tag{19}$$

for $k = 1, \ldots, K$, where the spectrum is given by the eigenvalues $\{\lambda_l\}$. The corresponding right eigenvectors $\{\varphi_l\}$ can be used to embed the system in a low-dimensional representation (or CVs).

Based on this eigendecomposition, the target mapping $\xi(\mathbf{x})$ (Eq. 1) can be defined as *diffusion coordinates* [41, 45]:

$$\mathbf{x} \mapsto \xi(\mathbf{x}) = \big(\lambda_1 \varphi_1(\mathbf{x}), \ldots, \lambda_d \varphi_d(\mathbf{x})\big), \tag{20}$$

where the eigenvalues and eigenvectors are given by $\{\lambda_l\}$ and $\{\varphi_l\}$, respectively, and define reduced coordinates. In Eq. 20, each diffusion coordinate is defined as $z_k = \lambda_k \varphi_k$, where the spectrum is sorted by non-increasing value:

$$\lambda_0 = 1 > \lambda_1 \geq \cdots \geq \lambda_d \geq \cdots \geq \lambda_K, \tag{21}$$

where d is the index at which we truncate the diffusion coordinates in Eq. 20 and the dimensionality of the reduced representation. Thus, the dominant timescales found in the dynamics of the high-dimensional system can be described only by several eigenvectors corresponding to the largest eigenvalues. The eigenvalue $\lambda_0 = 1$ and the first diffusion coordinate $\lambda_0 \varphi_0$ corresponds to the Boltzmann equilibrium distribution.

An application of reweighted diffusion map to alanine dipeptide is illustrated in Fig. 5. It can be seen that when diffusion reweighting is not used to construct the equilibrium density from enhanced

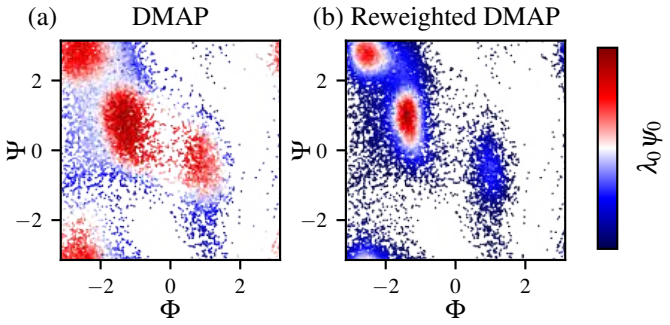


Figure 5: Reweighted diffusion map. Difference between equilibrium densities ($\lambda_0\varphi_0$) of alanine dipeptide obtained from biased simulation data calculated by (a) the standard DMAP and (b) reweighted DMAP. While reweighted DMAP correctly represents densities in metastable states, the standard DMAP is unable to provide the correct solution. In (a), transitions between states are much faster due to the use of bias potential. Figure from Ref. [H1].

sampling simulations, it leads to incorrect results. In contrast, when diffusion reweighting is used, the equilibrium density corresponds to the unbiased dynamics.

In summary, in Ref. [H3], the Author formally introduced and developed the theory and implementation of reweighted manifold learning. In contrast to Ref. [H2], the reweighting algorithm for learning CVs using manifold learning from biased enhanced sampling simulations was derived using Markov operators. The Author proved that, depending on the approximation of the manifold reweighting factor, one can obtain expressions proposed in Ref. [H2, 48]. This allowed us to show which important properties (such as density, geometry, and sample importance) were encoded in the Markov transition matrices used by those methods. The introduced reweighting was called diffusion reweighting, based on anisotropic diffusion kernels considered in diffusion map. From this point of view, the Author showed that diffusion reweighting can be used with several methods, including diffusion map.

5 Selecting High-Dimensional Representation

Although machine learning is becoming widely used for performing dimensionality reduction in atomistic simulations and general analysis of high-dimensional systems in physical chemistry, the quality of variables resulting from such methods depends heavily on the input data of many configuration variables. The selection of such high-dimensional representations used subsequently for dimensionality reduction is often overlooked.

According to the manifold hypothesis, preserving the timescale separation between slow and fast variables is a basis for the interpretable construction of high-dimensional representations. As previously explained, the slow variables intrinsically relate to the kinetics of rare transitions between long-lived metastable states. The fast variables, however, are adiabatically constrained to the dynamics of the slow variables and correspond mainly to short-time equilibration within metastable states. Therefore, we can consider different representations of the same system equivalent if the same timescale separation characterizes them.

To estimate the kinetic information encoded in the high-dimensional space, we collect N samples of n configuration variables from a simulation to construct the Markov transition matrix and perform its spectral decomposition. To this aim, a data set consisting of these samples is:

$$X = \{\mathbf{x}_k \in \mathbb{R}^n, w(\mathbf{x}_k)\}_{k=1}^{N}, \tag{22}$$

where the samples are augmented by statistical weights w if we sample a biased probability distribution, such as in enhanced sampling simulations.

Similar to the other methods, the reweighted anisotropic kernel with the anisotropic diffusion constant $\alpha = 1/2$ is introduced to employ information about the density and importance of the configuration space:

$$L(\mathbf{x}_k, \mathbf{x}_l) = r(\mathbf{x}_k, \mathbf{x}_l) \frac{g(\mathbf{x}_k, \mathbf{x}_l)}{\sqrt{\varrho(\mathbf{x}_k)\varrho(\mathbf{x}_l)}}, \tag{23}$$

where $\varrho(\mathbf{x}) = \sum_k g(\mathbf{x}, \mathbf{x}_k)$ is up to a multiplicative constant a kernel density estimate. We reweight the anisotropic kernel using $r(\mathbf{x}_k, \mathbf{x}_l)$, which is introduced to correct the effect of sampling from the biased probability distribution q (i.e., diffusion reweighting [H3]). The reweighting factor is given as [H2, H3]:

$$r(\mathbf{x}_k, \mathbf{x}_l) = \sqrt{w(\mathbf{x}_k) w(\mathbf{x}_l)}, \tag{24}$$

where $w(\mathbf{x}_k)$ and $w(\mathbf{x}_l)$ are the statistical weights corresponding to the k-th and l-th samples, respectively. For unbiased simulations, Eq. 24 reduces to the anisotropic diffusion kernel used in diffusion map [41, 51].

Equation 23 asymptotically corresponds to a reversible, overdamped approximation to the slow dynamics with the unbiased probability $p(\mathbf{x})$ as the stationary density [41, 51] even if the underlying dynamics proceeds according to the biased probability distribution [H3].

To determine whether a partial selection of configuration variables contains similar kinetic information as the complete set, we use the Markov transition matrix M. First, we perform an eigendecomposition of the matrix from the complete set of variables $M\varphi = \lambda\varphi$ and calculate its eigenvalues $\{\lambda_k\}$ and

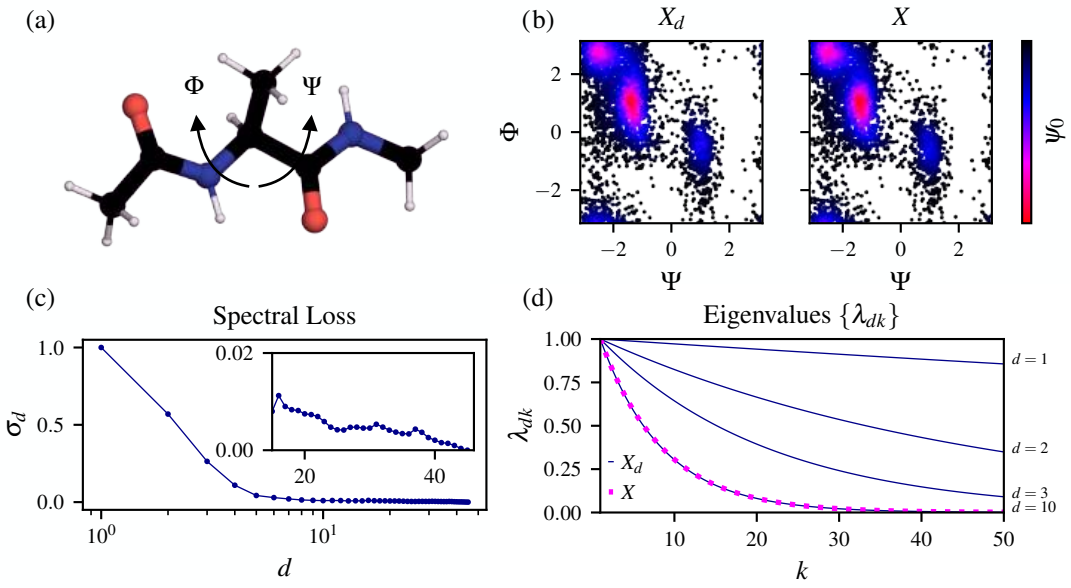


Figure 6: Selecting high-dimensional representation. (a) Biased simulation data is generated by enhancing fluctuations of the dihedral angles Φ and Ψ of alanine dipeptide. The initial high-dimensional representation consists of $n = 45$ pairwise distances between heavy atoms. Reweighted diffusion map is able to select only $d = 10$ variables which preserve kinetic information as can be seen in (b) conserved equilibrium density (φ_0), (c) decay of the spectral loss σ_d to 0, and (d) equivalent eigenvalues. Figure from Ref. [H4].

eigenfunctions $\{\varphi_k\}$. The eigenvalues are sorted by decreasing values. The corresponding eigenfunctions contain kinetic information about the system as the eigenvalues are related to the intrinsical timescales of the system.

Next, the eigendecomposition is carried out for combinations of the configuration variables, which define a data set X_d (d is the number of the configuration variables in the partial representation) and compare it to the eigenvalues of the complete high-dimensional representation. To describe how much kinetic information is conserved in the partial representations, we define a spectral loss:

$$\sigma_d = \alpha \left[\sum_k (\lambda_{dk} - \lambda_k)^2 \right]^{1/2}, \tag{25}$$

where α is a normalization constant and λ_k and λ_{dk} are the eigenvalues of the complete high-dimensional representation and the partial representation of d configuration variables, respectively. A combination of the configuration variables preserves kinetic information encoded in the complete

representation if the spectral loss is negligible.

Given the data set X of n configuration variables $\mathbf{x} = (x_1, \ldots, x_n)$ and its spectral decomposition $\{\lambda_k, \varphi_k\}$ of the related Markov transition matrix M, we search for the partial high-dimensional data set X_d of d configuration variables that upon spectral decomposition of its Markov transition matrix into $\{\lambda_{dk}, \varphi_{dk}\}$ contains similar kinetic information as the Markov transition matrix calculated from X. To avoid an exhaustive and computationally demanding search through all combinations of the configuration variables, we use an algorithm that provides a suboptimal result [52].

In Fig. 6, we show an example of how to select a high-dimensional alanine dipeptide representation. It can be seen that a partial selection of pairwise distances carries the same kinetic information as the complete representation. The partial selection corresponds to the spectral loss at around 0 and the same equilibrium distribution and eigenvalues as those calculated from the complete representation.

In summary, in Ref. [H4], the Author proposed and developed a method to select an initial high-dimensional representation for further learning CVs. The problem of constructing such a representation is often neglected when learning CVs or performed in a manner that does not preserve important kinetic information about the studied system, which often leads to invalid learned CVs. The method is based on kinetic equivalence – preserving timescales observed in the complete representation of the system. Additionally, as the proposed algorithm is based on the eigendecomposition of the reweighted Markov transition matrix, it can be used on both standard atomistic and enhanced sampling simulations.

6 Spectral Map

Spectral map is the latest development introduced by the Author in Ref. [H5] and improved in [H6]. By combining both kinds of manifold learning for molecular simulations (see Sec. 2), the technique is designed to especially focus on constructing *slow* CVs that arise as a result of the timescale separation in the investigated system.

To estimate the separation between effective timescales characteristic of the system, we model its reduced dynamics as a Markov chain using kernel functions. To measure the similarity between CV samples $\mathbf{z}_k$ and $\mathbf{z}_l$, we use the Gaussian kernel for the reduced representation:

$$g(\mathbf{z}_k, \mathbf{z}_l) = \exp\left(-\frac{1}{\varepsilon}\|\mathbf{z}_k - \mathbf{z}_l\|^2\right) \tag{26}$$

where $\|\mathbf{z}_k - \mathbf{z}_l\|$ denotes pairwise Euclidean distances between every pair of CV samples $k, l = 1, \ldots, N$ and N is the number of samples. The Gaussian kernel exhibits a notion of locality by defining a neighborhood around each sample of radius ε [41].

Metastable states often have multimodal characteristics, resulting in spatially heterogeneous free-energy landscapes. Therefore, to improve the ability of spectral map to adjust to metastable states, we estimate the sample-dependent scale factors by adaptively balancing local and global scales as:

$$\varepsilon_{kl}(r) = \|\mathbf{z}_k - \eta_r(\mathbf{z}_k)\| \cdot \|\mathbf{z}_l - \eta_r(\mathbf{z}_l)\|, \tag{27}$$

where each term defines a ball centered at $\mathbf{z}$ of radius $\eta_r(\mathbf{z}) > 0$. For convenience, we define the radius by the fraction of the neighborhood size $r \in [0, 1]$, allowing us to decide which scale is more relevant. Specifically, the Gaussian kernel describes a local neighborhood around each sample for values r close to 0 (i.e., the nearest neighbors), which correspond to deep and narrow states. For values of r around 1 (the farthest neighbors), it considers more global information, corresponding to shallow and wide states. Intermediate values of r maintain a balance between spatial scales.

As the marginal equilibrium density in the CV space is often far from uniform for dynamical systems with complex free-energy landscapes, we need a density-preserving kernel for data sampled from any underlying probability distribution. For this, we employ the anisotropic diffusion kernel [44]:

$$L(\mathbf{z}_k, \mathbf{z}_l) = \frac{g(\mathbf{z}_k, \mathbf{z}_l)}{\sqrt{\varrho(\mathbf{z}_k)\varrho(\mathbf{z}_l)}}, \tag{28}$$

where $\varrho(\mathbf{z}_k) = \sum_l g(\mathbf{z}_k, \mathbf{z}_l)$ is a kernel density estimate. Next, we build a Markov transition matrix by row-normalizing K:

$$m_{kl} \sim M(\mathbf{z}_k, \mathbf{z}_l) = \frac{L(\mathbf{z}_k, \mathbf{z}_l)}{\sum_n L(\mathbf{z}_k, \mathbf{z}_n)} \tag{29}$$

which models a discrete Markov chain in the CV space $m_{kl} = \Pr\{\mathbf{z}_{\tau+1} = \mathbf{z}_l \mid \mathbf{z}_\tau = \mathbf{z}_k\}$ expressing a probability of transition between CV samples $\mathbf{z}_k$ and $\mathbf{z}_l$ in an auxiliary (non-physical) time τ. The Markov chain approximates the long-time asymptotics of the system by describing the dynamics by the Fokker–Planck anisotropic diffusion [44].

Finally, to estimate the dominant timescales encoded in the system, we perform a spectral decomposition of the Markov transition matrix:

$$M\varphi_k = \lambda_k \varphi_k, \tag{30}$$

where φ_k and λ_k are the k-th right eigenfunctions and eigenvalues of M, respectively. The real-valued eigenvalues of M are (sorted in non-ascending order):

$$\lambda_0 = 1 > \lambda_1 \cdots \geq \lambda_N, \tag{31}$$

where the eigenvalue λ_0 corresponds to the equilibrium distribution of the Markov chain (Eq. 29) given by the eigenfunction φ_0. The dominant eigenvalues related to the slowest relaxation timescales in the system can be found by associating each eigenvalue with an effective timescale [53]:

$$t_k = -\frac{1}{\log \lambda_k}. \tag{32}$$

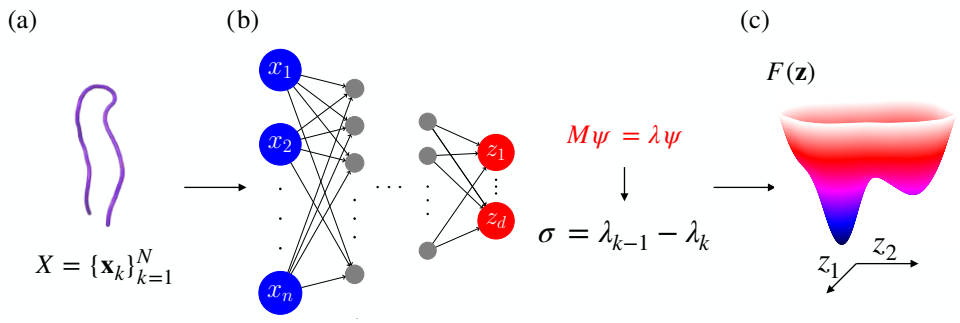


Figure 7: Outline of spectral map. (a) Dataset X in a high-dimensional representation $\mathbf{x} = (x_1, \ldots, x_n)$ used to describe the system is taken as an input for the target mapping. (b) Target mapping $\mathbf{z} = \xi_w(\mathbf{x})$ is modeled as a neural network that embeds the system in its high-dimensional representation to a low-dimensional map spanned by slow CVs $\mathbf{z} = (z_1, \ldots, z_d)$. An eigendecomposition of a Markov transition constructed from CV samples is performed ($M\varphi = \lambda\varphi$). The spectral gap σ is maximized based on the difference between neighboring eigenvalues $\{\lambda_k\}$ to separate the slow and fast timescales. (c) A trained neural network can be used to evaluate all available high-dimensional samples and calculate the corresponding free-energy landscape $F(\mathbf{z})$. Figure from Ref. [H5].

The largest gap between neighboring eigenvalues is called the spectral gap and determines the degree of the timescale separation between the slow and fast processes:

$$\sigma = \lambda_{k-1} - \lambda_k, \tag{33}$$

where $k > 0$ indicates the number of metastable states in the CV space [54].

The theory of spectral characterization of metastable states (see works by Gaveau and Schulman [54, 55] and references therein) explains that the spectral gap and degree of degeneracy in eigenvalue spectrum are related to the timescale separation and Markovian dynamics. If the eigenvalue of the Markov transition matrix is nearly degenerate $k+1$ times, it indicates that the equilibrium distribution breaks into k metastable states with infrequent transitions between them. The converse is also true: if the equilibrium density breaks into metastable states separated by a free-energy barrier much larger than the thermal energy $k_\mathrm{B}T$, there is eigenvalue degeneracy.

To achieve the reduced dynamics that is effectively Markovian, it is crucial to have a spectral gap between neighboring eigenvalues, along with the near degeneracy of the dominant eigenvalue. This condition is essential for the maximal spectral gap at k to lead to the separation into k metastable states, which can help reduce memory effects in the dynamics. Therefore, we consider the spectral gap as a scoring function in spectral map.

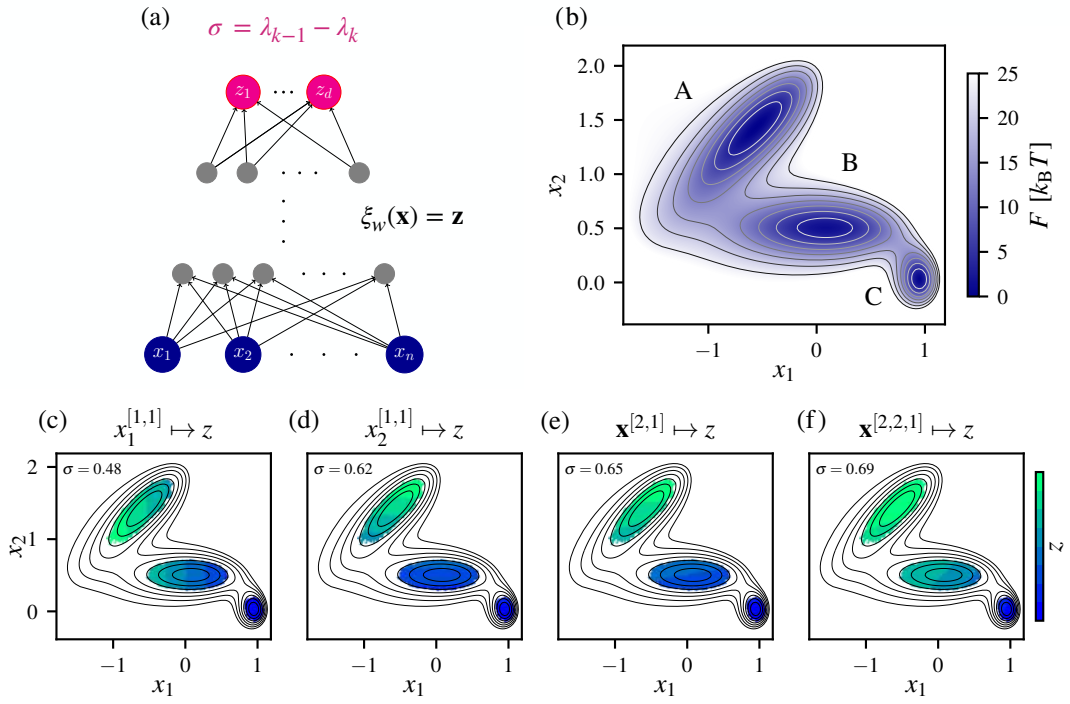


Figure 8: Training the target mapping $\xi_w(\mathbf{x}) = \mathbf{z}$ for the three-state Müller–Brown potential. (a) Schematic outline. (b) Free-energy landscape of a particle moving in the two-dimensional space $\mathbf{x} = (x_1, x_2)$ with barriers between the states of around 20 $k_\mathrm{B}T$ and a kinetic bottleneck between states B and C. (c-f) Examples of training the target mappings with different network architectures, e.g., $\mathbf{x}^{[a,\ldots,b]} \mapsto z$ denotes mapping the $\mathbf{x}$ variable through a network consisting of $[a, \ldots, b]$ layers to the z CV. Figure from Ref. [H6].

As the Markov transition matrix is estimated in the reduced space, we consider the effective dynamics rather than the dynamics of the microscopic coordinates of the system. Thus, the Markov chain defined in the CV space is implicitly modeled as following an overdamped Langevin dynamics with the marginal equilibrium density $p(\mathbf{z})$. Generally, this effective dynamics is either non-Markovian or has a $\mathbf{z}$-dependent diffusion matrix, meaning it is not driven exclusively by the free-energy landscape. However, by selecting CVs that arise from the timescale separation in the system as the reduced space, we can represent the dynamics of microscopic coordinates through the effective dynamics of slow CVs, which is approximately Markovian [56].

A schematic illustration of the workflow implemented in spectral map is shown in Fig. 7. First, a

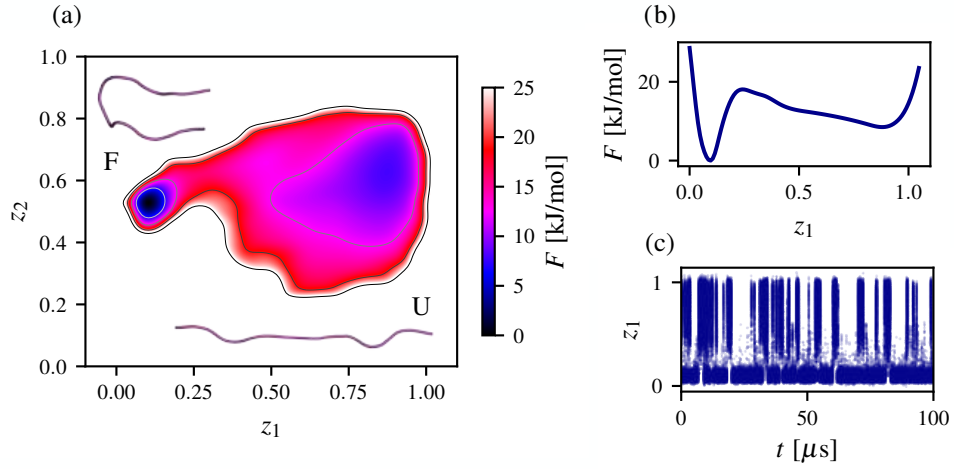


Figure 9: Learning slow CVs for the reversible folding of CLN025 in solvent sampled through a 100-μ molecular dynamics simulation. (a) Spectral map and the corresponding free-energy landscape F of CLN025 showing two distinct free-energy basins: the main and well-defined folded (F) and more loosely structured unfolded (U) metastable states, separated by a free-energy barrier of around 20 kJ/mol. Representative conformations from the metastable states are shown. The aspect ratio of axes is preserved. (b) Free-energy profile shown as a function of the z_1 CV. (c) Time series for the z_1 CV showing transitions between the states. Figure from Ref. [H6].

dataset X in a high-dimensional representation $\mathbf{x} = (x_1, \ldots, x_n)$ used to describe the system is taken as an input for the target mapping. Then, the target mapping $\mathbf{z} = \xi_w(\mathbf{x})$ is modeled as a neural network that embeds the system in its high-dimensional representation to a low-dimensional map spanned by slow CVs $\mathbf{z} = (z_1, \ldots, z_d)$. Subsequently, an eigendecomposition of a Markov transition constructed from CV samples is performed ($M\varphi = \lambda\varphi$). The spectral gap σ is maximized based on the difference between neighboring eigenvalues $\{\lambda_k\}$ to separate the slow and fast timescales. Finally, the trained neural network can be used to evaluate all available high-dimensional samples and calculate the corresponding free-energy landscape $F(\mathbf{z})$. An example application of using spectral map to estimate slow CVs and the related free-energy landscape are shown in Fig. 8 (Muller–Brown potential), 9 (chignolin), and 10 (BBA protein).

In summary, in Ref. [H5, H6], the Author proposed and developed a method for learning slow CVs. Learning slow CVs is a very difficult problem; thus, CVs often do not satisfy this requirement. However, slow CVs describe the most important degrees of freedom (or modes) of the system, as they correspond to the most interesting transitions in the system that occur on longer and experimental timescales.

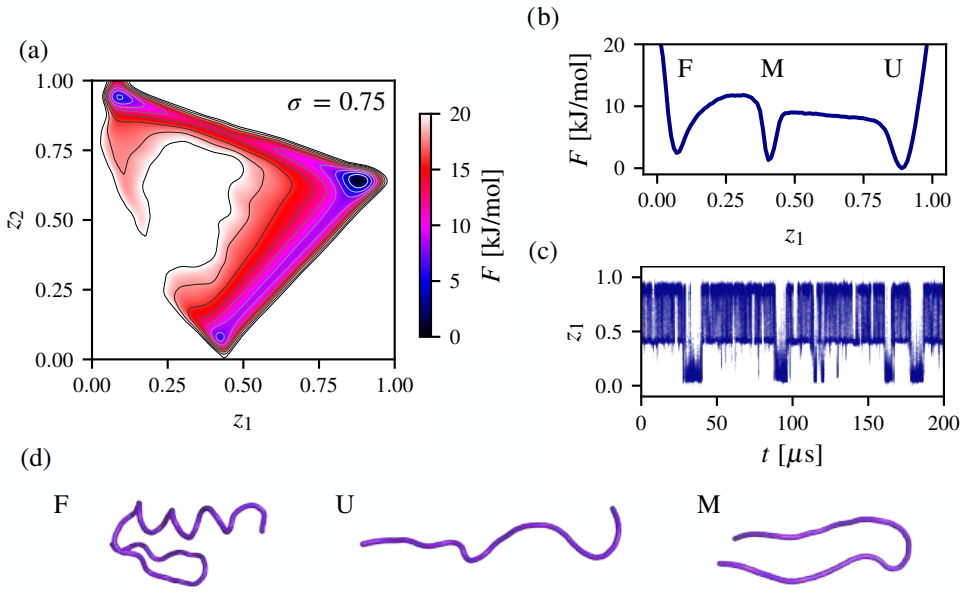


Figure 10: Spectral map and free-energy landscape of the folding of the BBA protein from a 200-μs molecular dynamics simulation. A high-dimensional representation given by $n = 378$ pairwise Euclidean distances between the Cα atoms of BBA. (a) Free-energy landscape showing metastable states spanned by CVs calculated by spectral map for $k = 3$ where the corresponding spectral gap reaches $\sigma = 0.75$. (b) Free-energy profile along the z_1 CV with z_2 integrated out. (c) Time series of z_1 showing changes between metastable states during the molecular dynamics simulation. (c) Representative conformations of the BBA protein corresponding to the folded state (F), the unfolded state (U), and the misfolded state (M). Figure from Ref. [H5].

Additionally, spectral map is very easy to implement as it is based on calculating anisotropic kernels and performing spectral decompositions while ensuring that the spectral gap, which measures the timescale separation between the slow and fast kinetics in the system, is maximized. Spectral map can work with high-dimensional representations of hundreds of variables without any preprocessing usually required for manifold learning methods. Data and code required to reproduce the results are available from `https://zenodo.org/records/10678142`.

7 PLUMED: Promoting transparency and reproducibility

PLUMED is a freely available, open-source library that provides various methods, such as enhanced sampling algorithms, free-energy methods, and tools to analyze the vast amounts of data produced

by MD simulations. More information can be found here. These techniques can be combined with a large toolbox of collective variables that describe complex processes in physics, chemistry, material science, and biology. PLUMED can be patched with the majority of simulation engines used by the community.

In 2019, the PLUMED consortium was established to encourage transparency and reproducibility in atomistic simulations [H7]. This consortium revolves around the open-source plugin PLUMED, which is used in various fields where atomistic simulations are conducted. The PLUMED consortium is an open community composed of current developers, contributors, and all those researchers whose work builds on atomistic simulations and drives the development and dissemination of their results. The core objective of the consortium is establishing more effective protocols for sharing information in molecular dynamics simulations, promoting scientific reproducibility, and upholding the highest research standards.

PLUMED consortium introduced PLUMED-NEST (https://www.plumed-nest.org/), which is the public repository of the PLUMED users. It provides all the data needed to reproduce the results of a PLUMED-enhanced molecular dynamics simulation (or analysis) in a published paper. Furthermore, PLUMED-NEST monitors the compatibility of the provided PLUMED input files with the current and development versions of the code. Since PLUMED-NEST has been opened, the Author has deposited many results, datasets, and implementations. [2]

8 Brief Conclusion

Atomistic simulations, such as molecular dynamics or Monte Carlo, have emerged as general methods to investigate dynamical systems in physics, chemistry, and biology. Such simulations offer detailed insight into processes at the microscopic level with greater spatiotemporal accuracy than experiments. However, analyzing systems consisting of thousands of atoms can be challenging. To obtain a simplified, more understandable representation, it is often necessary to construct a low-dimensional representation that captures essential physical characteristics. Providing theory and techniques that can perform this task without supervision alleviates the problem of relying on experience and trial and error.

The Achievement is a valuable interdisciplinary addition to modern statistical physics. The following list summarizes the most important parts of the Achievement according to the Author:

- Developing a self-consistent framework of theory and methods for the statistical learning of CVs from standard and enhanced sampling simulations, which can be applied to understand any molecular process on the experimental timescales in a data-driven manner.

[2] See https://www.plumed-nest.org/browse.html and search "rydzewski"

- The framework takes into account the most important physical characteristics of dynamical processes, unlike many previously proposed methods. These include the probability distribution (equilibrium or nonequilibrium) sampled by the studied system, a notion of distance between the configurations, and slow kinetics, which is key to understanding processes on longer timescales.
- The theory underlying the proposed framework combines general ideas and tools from statistical mechanics and machine learning. This provides the machine learning methods used here with an ability that is lacking in most unsupervised dimensionality reduction methods: interpretability in the context of physics.
- The proposed framework is general and enables the community to extend it easily, which underlines its potential for widespread use and further improvement.

Even though methods developed at the intersection of statistical physics and machine learning have just started to emerge and be used, significant progress can certainly be noticed. Although the problem of constructing slow CVs from atomistic simulations is far from being definitively solved, the Author is confident that his developed and implemented consistent framework, described in this Achievement, will be widely used by the community and is an important step in gaining a better understanding of complex systems and a wide range of physical processes such as protein conformational changes during folding or binding to drugs, catalysis, glass transitions, or crystallization.

References

H1. ***Rydzewski, J.**, Chen, M. & Valsson, O. Manifold Learning in Atomistic Simulations: A Conceptual Review. *Mach. Learn.: Sci. Technol.* **4,** 031001 (2023).

H2. ***Rydzewski, J.** & Valsson, O. Multiscale Reweighted Stochastic Embedding: Deep Learning of Collective Variables for Enhanced Sampling. *J. Phys. Chem. A* **125,** 6286–6302 (2021).

H3. ***Rydzewski, J.**, Chen, M., Ghosh, T. K. & Valsson, O. Reweighted Manifold Learning of Collective Variables from Enhanced Sampling Simulations. *J. Chem. Theory Comput.* **18,** 7179–7192 (2022).

H4. ***Rydzewski, J.** Selecting High-Dimensional Representations of Physical Systems by Reweighted Diffusion Maps. *J. Phys. Chem. Lett.* **14,** 2778–2783 (2023).

H5. ***Rydzewski, J.** Spectral Map: Embedding Slow Kinetics in Collective Variables. *J. Phys. Chem. Lett.* **14,** 5216–5220 (2023).

H6. ***Rydzewski, J.** & Gökdemir, T. Learning Markovian Dynamics with Spectral Maps. *J. Chem. Phys.* **160** (2024).

H7. PLUMED Consortium, Promoting Transparency and Reproducibility in Enhanced Molecular Simulations. *Nat. Methods* **16,** 670–673 (2019).

1. Chandler, D. *Introduction to Modern Statistical Mechanics* (Oxford University Press, Oxford, UK, 1987).
2. Battimelli, G., Battimelli, G., Ciccotti, G., Greco, P. & Scalone. *Computer Meets Theoretical Physics* (Springer, 2020).
3. Piccini, G. *et al.* Ab Initio Molecular Dynamics with Enhanced Sampling in Heterogeneous Catalysis. *Catal. Sci. Technol.* **12,** 12–37 (2022).
4. Baron, R. & McCammon, J. A. Molecular Recognition and Ligand Association. *Annu. Rev. Phys. Chem.* **64,** 151–175 (2013).
5. Bruce, N. J., Ganotra, G. K., Kokh, D. B., Sadiq, S. K. & Wade, R. C. New Approaches for Computing Ligand–Receptor Binding Kinetics. *Curr. Opin. Struct. Biol.* **49,** 1–10 (2018).
6. Bernetti, M., Masetti, M., Rocchia, W. & Cavalli, A. Kinetics of Drug Binding and Residence Time. *Annu. Rev. Phys. Chem.* **70,** 143–171 (2019).

7. Wolf, S. Predicting Protein–Ligand Binding and Unbinding Kinetics with Biased MD Simulations and Coarse-Graining of Dynamics: Current State and Challenges. *J. Chem. Inf. Model.* (2023).

8. O'Hagan, M. P., Haldar, S., Morales, J. C., Mulholland, A. J. & Galan, M. C. Enhanced Sampling Molecular Dynamics Simulations Correctly Predict the Diverse Activities of a Series of Stiff-Stilbene G-Quadruplex DNA Ligands. *Chem. Sci.* **12,** 1415–1426 (2021).

9. Van Speybroeck, V., Vandenhaute, S., Hoffman, A. E. J. & Rogge, S. M. J. Towards Modeling Spatiotemporal Processes in Metal–Organic Frameworks. *Trends Chem.* **3,** 605–619 (2021).

10. Neha, Tiwari, V., Mondal, S., Kumari, N. & Karmakar, T. Collective Variables for Crystallization Simulations–from Early Developments to Recent Advances. *ACS Omega* **8,** 127–146 (2023).

11. Aussems, D. U. B., Bal, K. M., Morgan, T. W., Van De Sanden, M. C. M. & Neyts, E. C. Atomistic Simulations of Graphite Etching at Realistic Time Scales. *Chem. Sci.* **8,** 7160–7168 (2017).

12. Coifman, R. *Harmonic Analytic Geometry in High Dimensions—Empirican Models* International Conference of Mathematicians. Lecture. 2018.

13. Hohenberg, P. C. & Krekhov, A. P. An Introduction to the Ginzburg–Landau Theory of Phase Transitions and Nonequilibrium Patterns. *Phys. Rep.* **572,** 1–42 (2015).

14. Zwanzig, R. Memory Effects in Irreversible Thermodynamics. *Phys. Rev.* **124,** 983 (1961).

15. Luttinger, J. Theory of Thermal Transport Coefficients. *Physical Review* **135,** A1505 (1964).

16. Mori, H. Transport, Collective Motion, and Brownian Motion. *Prog. Theor. Phys.* **33,** 423–455 (1965).

17. Wu, H. & Noé, F. Variational Approach for Learning Markov Processes from Time Series Data. *J. Nonlinear Sci.* **30,** 23–66. ISSN: 1432-1467 (2020).

18. Brunton, S. L., Budišić, M., Kaiser, E. & Kutz, J. N. Modern Koopman theory for dynamical systems. *arXiv preprint arXiv:2102.12086* **64,** 229–340 (2021).

19. Izenman, A. J. Introduction to Manifold Learning. *Wiley Interdiscip. Rev. Comput. Stat.* **4,** 439–446 (2012).

20. Tenenbaum, J. B., De Silva, V. & Langford, J. C. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* **290,** 2319–2323 (2000).

21. Roweis, S. T. & Saul, L. K. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* **290,** 2323–2326 (2000).

22. Peters, B. Reaction Coordinates and Mechanistic Hypothesis Tests. *Annu. Rev. Phys. Chem.* **67,** 669–690 (2016).

23. Valsson, O., Tiwary, P. & Parrinello, M. Enhancing Important Fluctuations: Rare Events and Metadynamics from a Conceptual Viewpoint. *Annu. Rev. Phys. Chem.* **67,** 159–184 (2016).

24. Swendsen, R. H. & Wang, J.-S. Replica Monte Carlo Simulation of Spin-Glasses. *Phys. Rev. Lett.* **57,** 2607 (1986).

25. Earl, D. J. & Deem, M. W. Parallel Tempering: Theory, Applications, and New Perspectives. *Phys. Chem. Chem. Phys.* **7,** 3910–3916 (2005).

26. Chen, M., Cuendet, M. A. & Tuckerman, M. E. Heating and Flooding: A Unified Approach for Rapid Generation of Free Energy Surfaces. *J. Chem. Phys.* **137,** 024102 (2012).

27. Valsson, O. & Parrinello, M. Variational Approach to Enhanced Sampling and Free Energy Calculations. *Phys. Rev. Lett.* **113,** 090601 (2014).

28. Reinhardt, M. & Grubmüller, H. Determining Free-Energy Differences Through Variationally Derived Intermediates. *J. Chem. Theory Comput.* **16,** 3504–3512 (2020).

29. Torrie, G. M. & Valleau, J. P. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comp. Phys.* **23,** 187–199 (1977).

30. Mezei, M. Adaptive Umbrella Sampling: Self-Consistent Determination of the Non-Boltzmann Bias. *J. Comput. Phys.* **68,** 237–248 (1987).

31. Laio, A. & Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. U.S.A.* **99,** 12562–12566 (2002).

32. Barducci, A., Bussi, G. & Parrinello, M. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys. Rev. Lett.* **100,** 020603 (2008).

33. Maragakis, P., van der Vaart, A. & Karplus, M. Gaussian-Mixture Umbrella Sampling. *J. Phys. Chem. B* **113,** 4664–4673. ISSN: 1520-5207 (2009).

34. Morishita, T., Itoh, S. G., Okumura, H. & Mikami, M. Free-Energy Calculation via Mean-Force Dynamics using a Logarithmic Energy Landscape. *Phys. Rev. E* **85,** 066702 (2012).

35. Invernizzi, M., Piaggi, P. M. & Parrinello, M. Unified Approach to Enhanced Sampling. *Phys. Rev. X* **10,** 041034 (2020).

36. Hénin, J., Lelièvre, T., Shirts, M. R., Valsson, O. & Delemotte, L. Enhanced Sampling Methods for Molecular Dynamics Simulations. *arXiv preprint arXiv:2202.04164* **4,** 1583 (2022).

37. Belkin, M. & Niyogi, P. *Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering* in *Adv. Neural Inf. Process. Syst.* **14** (MIT Press, 2001), 585–591.

38. Hinton, G. E. & Roweis, S. *Stochastic Neighbor Embedding* in *Adv. Neural Inf. Process. Syst.* (eds Becker, S., Thrun, S. & Obermayer, K.) **15** (MIT Press, 2002), 833–864.

39. Belkin, M. & Niyogi, P. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Comput.* **15,** 1373–1396 (2003).

40. Hashemian, B., Millán, D. & Arroyo, M. Modeling and Enhanced Sampling of Molecular Systems with Smooth and Nonlinear Data-Driven Collective Variables. *J. Chem. Phys.* **139,** 12B601_1 (2013).

41. Coifman, R. R. *et al.* Geometric Diffusions as a Tool for Harmonic Analysis and Structure Definition of Data: Diffusion Maps. *Proc. Natl. Acad. Sci. U.S.A.* **102,** 7426–7431 (2005).

42. Van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9,** 2579–2605 (2008).

43. Aflalo, Y. & Kimmel, R. Spectral Multidimensional Scaling. *Proc. Natl. Acad. Sci. U.S.A.* **110,** 18052–18057 (2013).

44. Nadler, B., Lafon, S., Coifman, R. R. & Kevrekidis, I. G. Diffusion Maps, Spectral Clustering and Reaction Coordinates of Dynamical Systems. *Appl. Comput. Harmon. Anal.* **21,** 113–127 (2006).

45. Coifman, R. R. & Lafon, S. Diffusion Maps. *Appl. Comput. Harmon. Anal.* **21,** 5–30 (2006).

46. Jones, P. W., Maggioni, M. & Schul, R. Manifold Parametrizations by Eigenfunctions of the Laplacian and Heat Kernels. *Proc. Natl. Acad. Sci. U.S.A.* **105,** 1803–1808 (2008).

47. Van der Maaten, L. Learning a Parametric Embedding by Preserving Local Structure. *J. Mach. Learn. Res.* **5,** 384–391 (2009).

48. Zhang, J. & Chen, M. Unfolding Hidden Barriers by Active Enhanced Sampling. *Phys. Rev. Lett.* **121,** 010601 (2018).

49. Hinton, G. E. & Salakhutdinow, R. R. Reducing the Dimensionality of Data with Neural Networks. *Science* **313,** 504–507 (2006).

50. Coifman, R. R., Kevrekidis, I. G., Lafon, S., Maggioni, M. & Nadler, B. Diffusion Maps, Reduction Coordinates, and Low Dimensional Representation of Stochastic Systems. *Multiscale Model. Simul.* **7,** 842–864 (2008).

51. Singer, A., Erban, R., Kevrekidis, I. G. & Coifman, R. R. Detecting Intrinsic Slow Variables in Stochastic Dynamical Systems by Anisotropic Diffusion Maps. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 16090–16095 (2009).

52. Pudil, P., Novovičová, J. & Kittler, J. Floating Search Methods in Feature Selection. *Pattern Recognit. Lett.* **15,** 1119–1125 (1994).

53. Bovier, A., Eckhoff, M., Gayrard, V. & Klein, M. Metastability and Low Lying Spectra in Reversible Markov Chains. *Commun. Math. Phys.* **228,** 219–255 (2002).

54. Gaveau, B. & Schulman, L. S. Theory of Nonequilibrium First-Order Phase Transitions for Stochastic Dynamics. *J. Math. Phys.* **39,** 1517–1533 (1998).

55. Gaveau, B. & Schulman, L. S. Master Equation based Formulation of Nonequilibrium Statistical Mechanics. *J. Math. Phys.* **37,** 3897–3932 (1996).

56. Tiwary, P. & Berne, B. J. Spectral Gap Optimization of Order Parameters for Sampling Complex Molecular Systems. *Proc. Natl. Acad. Sci. U.S.A.* **113,** 2839 (2016).

Other scientific achievements

Apart from the scientific publications taken into account in the Achievement, since obtaining PhD, the Author has also worked on ligand-protein interaction problems, i.e., developing enhanced sampling techniques to observe multiple ligand dissociation pathways from proteins [A1, A2, A3], enhancing photodynamics of a photoswitch canonical bacteriophytochrome using advanced molecular dynamics methods [A4], aggregation of α-synucleins [A5], and overcoming pytethroid resistance in mosquito control [A6].

A1. ***Rydzewski, J.** maze: Heterogeneous Ligand Unbinding along Transient Protein Tunnels. *Comp. Phys. Commun.* **247,** 106865 (2020).

A2. ***Rydzewski, J.** & Valsson, O. Finding Multiple Reaction Pathways of Ligand Unbinding. *J. Chem. Phys.* **150** (2019).

A3. ***Rydzewski, J.**, Jakubowski, R., Nowak, W. & Grubmuller, H. Kinetics of Huperzine A Dissociation from Acetylcholinesterase via Multiple Unbinding Pathways. *J. Chem. Theory Comput.* **14,** 2843–2851 (2018).

A4. ***Rydzewski, J.**, Walczewska-Szewc, K., Czach, S., Nowak, W. & Kuczera, K. Enhancing the Inhomogeneous Photodynamics of Canonical Bacteriophytochrome. *J. Phys. Chem. B* **126,** 2647–2657 (2022).

A5. Walczewska-Szewc, K., **Rydzewski, J.** & Lewkowicz, A. Inhibition-Mediated Changes in Prolyl Oligopeptidase Dynamics Possibly Related to α-Synuclein Aggregation. *Phys. Chem. Chem. Phys.* **24,** 4366–4373 (2022).

A6. Niklas, B., **Rydzewski, J.**, Lapied, B. & Nowak, W. Toward Overcoming Pyrethroid Resistance in Mosquito Control: The Role of Sodium Channel Blocker Insecticides. *Int. J. Mol. Sci.* **24,** 10334 (2023).

V. Presentation of Significant Scientific Activity Carried Out at More than One University, Scientific or Cultural Institution, Especially at Foreign Institutions

The Author's scientific research at European foreign institutions was conducted at:

1. National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan, Group of Prof. Tetsuya Morishita, XI 2023–II 2024.

2. Department of Theoretical and Computational Biophysics, Max Planck Institute for Biophysical Chemistry (since 2022 Max Planck Institute for Multidisciplinary Sciences), Gottingen, Germany, Group of Prof. Helmut Grubmüller, X 2016–IV 2017.

3. Department of Chemistry and Applied Biosciences, Swiss Federal Institute of Technology in Zürich c/o Institute of Computational Science, Università della Svizzera italiana, Lugano, Switzerland, Group of Prof. Michele Parrinello, VII 2016–X 2016.

Ongoing projects carried at more than one university:

1. Prof. Omar Valsson, University of North Texas, Denton, US, 2016–. Work related to constructing collective variables for enhanced sampling simulations using optimization and explainable machine learning. A continuation of works shown in Refs. [H1, H2, H3, A2].

2. Prof. Ming Chen, Purdue University, West Lafayette, US, 2018–. Work related to constructing unbiased Markov transition matrices from enhanced sampling simulations. A continuation of works shown in Refs. [H1, H3].

3. Prof. Alexander M. Berezhkovski, National Institutes of Health, Bethesda, Maryland, US, Since 2023. Work related to Markovian dynamics along reaction coordinates.

4. Prof. Michele Parrinello, Italian Institute of Technology, Genova, Italy, 2016–. Work related to developing enhanced sampling methods to simulate ligand unbinding from proteins and establishing protols for transparent and reproducible atomistic simulations. A continuation of works shown in Ref. [H7].

5. Prof. Helmut Grubmüller, Max Planck Insitutue for Multidisciplinary Sciences, Gottingen, Germany, 2016–. Work related to estimating kinetics and thermodynamics of inhibitor unbinding from enzymes. A continuation of works shown in Ref. [A3].

6. Prof. Tetsuya Morishita, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan, 2022–. Work related to calculating macroscopic variables describing glass formation.

7. Prof. Yasuteru Shigeta and Prof. Ryuhei Harada, University of Tsukuba, Tsukuba, Japan, 2019–. Work related to implementing bias-free enhanced sampling methods into the PLUMED plugin [H7].

8. Prof. Bruno Lapied, University Angers, Angers, France, 2023–. Work related to computationally overcoming pyrethroid resistance in mosquito control. A continuation of works shown in Ref. [A6].

VI. Presentation of Teaching, Organizational Achievements, and Achievements in the Popularization of Science

Teaching:

The courses indicated by "(coordinator)" were proposed, developed, and coordinated by the Author. Starting from 2022 the Author have been working full time in a research position under his Sonata NCN grant and therefore has not been teaching any courses.

- Numerical methods: 2014–2015, 2018–2020
- Radiomics (coordinator): 2019–2022
- Programming in Python (coordinator): 2018–2022
- Computational physics (coordinator): 2018–2022
- Discrete mathematics: 2018-2022
- Stochastic algorithms (coordinator): 2016–2017
- Procedural programming: 2015–2022
- Elements of bioinformatics: 2015–2016
- Programming languages: 2014–2017
- Problems of ethics, open science and open innovation: 2022

Organization:

- Organization of the conference Bioinformatics in Torun held jointly by the Polish Bioinformatics Society (PTBI) and the Nicolaus Copernicus University in Torun, 2014–2023.
- Reviewer for the Polish Bioinformatics Society competition for the best bachelor thesis in bioinformatics and computational biology, 2022.
- Reviews for the proceedings of Annual Conference on Machine Learning, Optimization and Data Science (LOD), 2018–2020.

Students:

- Assistant supervisor of PhD students: Tugce Gokdemir (Physical sciences; 2023–; funded by the Sonata grant obtained by the Author), Sylwia Czach (Physical sciences; 2020–)
- Supervisor of MSc students: Wojciech Amtmański (Medical physics; 2020–2021), Karolina Kolonko (Informatics; 2021–2022), Patryk Tajs (Informatics; 2023–2024).

- Supervisor of BSc students: Aleksander Oskroba (Physics; 2019–2021)
- Supervisor of BSc Eng students: Aleksandra Warmbier (Informatics; 2019–2022), Kajetan Krzewina (Informatics; 2019–), Bartosz Jagodziński (Informatics; 2020–2022), Patryk Tajs (Informatics; 2021–2023), Jacek Wierzejewski (Informatics; 2021–2023)

Popularization:

- Inaugural lecture for first-year students of the Faculty of Physics, Astronomy and Informatics at Nicolaus Copernicus University, Torun, October 2018.
- Enhanced Sampling Methods (in Polish) published in the Voice of the University 7-10, 2019, distributed by Nicolaus Copernicus University, Torun.
- Popularization note about promoting grants offered by the National Science Center, 10-th aniversary of the National Science Center, 2021.
- Summer camp for international students and Ph.D. students in the program of the Polish National Agency for Academic Exchange (NAWA) SPINAKER – Intensive International Education Programs, 12-14th July 2022 in Torun, Poland.
- Promoting open science as the PI of the Polish National Agency for Academic Exchange's project "Open NCU–Open Source, Open Science" by funding open access publishing fees for young researchers at Nicolaus Copernicus University, Torun.
- Teaching classes about open science in the program "Developing and Implementing hands-on training on Open Science and Open Innovation for Early Career Researchers (DIOSI)" funded by the European Union under the Horizon 2020, European Commission 101006318, 2020–2023.

Wykaz osiągnięć naukowych stanowiących znaczny wkład w rozwój określonej dyscypliny

(Stan na dzień 22/05/2024 r.)

Imię i nazwisko habilitanta: Jakub Rydzewski

Informacje naukometryczne:

- ResearcherID: N-9160-2019
- ORCID: 0000-0003-4325-4177
- Google Scholar: dEMXOpcAAAAJ
- Hirsch Index: 9 [Web of Science (WoS)] 10 [Google Scholar (GS)]
- 10-Index: 13 [GS]
- Liczba publikacji: 25

Całkowita liczba cytowań: 780 [WoS] 981 [GS]
Całkowita liczba cytowań bez autocytowań: 682 [WoS]
Całkowita liczba punktów Ministerstwa, w tym:

- Publikacje po roku 2018 (skala do 200 pkt.): 1860
- Publikacje do roku 2018 (skala do 50 pkt.): 370

Pełen wykaz artykułów habilitanta w czasopismach naukowych obejmuje artykuły wymienione w punkcie I.1, które przyczyniają się do osiągnięcia habilitacyjnego. Oprócz tych publikacji, od czasu uzyskania stopnia doktora, Autorka uzyskał również inne osiągnięcia naukowe, na które składają się publikacje [A1-A6] podane w punkcie II.1.

I. Informacja o osiągnięciach naukowych, o których mowa w art. 219 ust. 1. Pkt 2 Ustawy.

Tytuł osiągnięcia:

Uczenie zmiennych zbiorowych z symulacji atomistycznych

I.1. Cykl powiązanych tematycznie artykułów naukowych, zgodnie z art. 219 ust. 1. pkt 2b Ustawy

Numery [H1-H6] przypisane poniżej są również używane w podsumowaniu osiągnięć zawodowych. Informacje o Impact Factor zostały zaczerpnięte z bazy Web of Science (WoS) dla roku 2022. Gwiazdka przy nazwisku oznacza publikacje, w których kandydat jest autorem korespondencyjnym.

Wkład autora jest przypisywany zgodnie z klasyfikacją Contributor Roles Taxonomy (CRediT) (https://credit.niso.org/) stosowaną przez American Institute of Physics.

H1. ***Rydzewski, J.**, Chen, M. & Valsson, O. Manifold Learning in Atomistic Simulations: A Conceptual Review. *Mach. Learn. Sci. Technol.* **4**, 031001 (2023).
Impact factor: 6.8
Ilość punktów Ministerialnych: 20
Liczba cytowań: 4 [WoS] 10 [GS]

Analiza dużych zbiorów wysokowymiarowych danych wymaga redukcji wymiarowości: znalezienia znaczących struktur niskowymiarowych ukrytych w ich wysokowymiarowych obserwacjach. Taka praktyka jest potrzebna w symulacjach układów dynamicznych, w których próbkowane są nawet tysiące stopni swobody. Obfitość takich danych utrudnia uzyskanie wglądu w konkretny problem fizyczny. Naszym głównym celem w tym przeglądzie jest skupienie się na nienadzorowanych metodach uczenia maszynowego, które mogą być wykorzystane na danych symulacyjnych w celu znalezienia niskowymiarowej rozmaitości zapewniającej zbiorczą i informacyjną charakterystykę badanego procesu. Takie rozmaitości mogą być wykorzystywane do próbkowania procesów o długiej skali czasowej i szacowania energii swobodnej. Opisujemy metody, które mogą działać na zbiorach danych ze standardowych i rozszerzonych symulacji atomistycznych. W porównaniu z ostatnimi przeglądami dotyczącymi uczenia się rozmaitości dla symulacji atomistycznych, rozważamy tylko metody, które konstruują niskowymiarowe rozmaitości oparte na prawdopodobieństwach przejścia Markowa między próbkami wielowymiarowymi. Omawiamy te techniki z koncepcyjnego punktu widzenia, w tym ich podstawowe ramy teoretyczne i możliwe ograniczenia.

Wkład:

- Konceptualizacja (wiodący): Opracował teorię, metody, przykłady i koncepcję ogólnych ram uczenia nienadzorowanego w standardowych i wzmocnionych symulacjach atomistycznych; omówił koncepcję ze współautorami i uwzględnił ich sugestie.
- Zasoby (wiodący): Pozyskał środki na publikację w ramach otwartego dostępu.

- Nadzór (wiodący): Pełnił rolę autora korespondencyjnego oraz organizował spotkania i dyskusje.
- Wizualizacja (wiodący): Wybrał, przygotował i zwizualizował wszystkie wyniki do publikacji.
- Pisanie/Przygotowanie wersji roboczej pracy (wiodący): Napisał wersję roboczą.
- Pisanie/Recenzja i edycja (równy): Poprawił wersję roboczą pracy z współautorami.

H2. ***Rydzewski, J.** & Valsson, O. Multiscale Reweighted Stochastic Embedding: Deep Learning of Collective Variables for Enhanced Sampling. *J. Phys. Chem. A* **125**, 6286–6302 (2021).
Impact factor: 2.9
Ilość punktów Ministerialnych: 100
Liczba cytowań: 19 [WoS] 24 [GS]

Metody uczenia maszynowego zapewniają ogólne ramy do automatycznego znajdowania i reprezentowania podstawowych cech danych symulacyjnych. Zadanie to jest szczególnie istotne w symulacjach z wzmocnionym próbkowaniem. Tam szukamy kilku uogólnionych stopni swobody, zwanych zmiennymi zbiorowymi, aby reprezentować i kierować próbkowaniem krajobrazu energii swobodnej. Teoretycznie, te zmienne zbiorowe powinny oddzielać różne stany metastabilne i odpowiadać powolnym stopniom swobody badanego procesu fizycznego. W tym celu proponujemy nową metodę, którą nazywamy *multiscale reweighted stochastic embedding* (MRSE). Nasza praca opiera się na parametrycznej wersji stochastycznego osadzania sąsiadów. Technika ta automatycznie uczy się zmiennych zbiorowych, które mapują wielowymiarową przestrzeń cech do niskowymiarowej przestrzeni ukrytej za pośrednictwem głębokiej sieci neuronowej. Wprowadzamy kilka nowych ulepszeń do metod stochastycznego osadzania sąsiadów, które sprawiają, że MRSE szczególnie nadaje się do ulepszonych symulacji próbkowania: (1) próbkowanie losowe z uwzględnieniem wagi jako schemat wyboru punktów orientacyjnych w celu uzyskania zestawów danych szkoleniowych, które zapewniają równowagę między reprezentacją równowagi a wychwytywaniem ważnych stanów metastabilnych leżących wyżej w energii swobodnej; (2) wieloskalowa reprezentacja wielowymiarowej przestrzeni cech za pomocą gaussowskiego modelu prawdopodobieństwa mieszanki; oraz (3) procedura ponownego ważenia w celu uwzględnienia danych szkoleniowych z tendencyjnego rozkładu prawdopodobieństwa. Pokazujemy, że MRSE konstruuje niskowymiarowe zmienne zbiorowe, które mogą poprawnie charakteryzować różne stany metastabilne w trzech układach modelowych: potencjał Mullera-Browna, dipeptyd alaniny i tetrapeptyd alaniny.

Wkład:

- Konceptualizacja (równy): Wraz z współautorem opracował teorię, pomysł, metody, przykłady

i koncepcję wykorzystania parametrycznego uczenia sieci neuronowej do mapowania zmiennych zbiorowych.

- Zarządzanie danymi (wiodący): Przygotował i przeprowadził wszystkie rozszerzone symulacje próbkowania i zastosował opracowane metody uczenia się zmiennych zbiorowych z tych symulacji.
- Analiza formalna (wiodący): Przeanalizował wszystkie wyniki.
- Metodologia (równy): Wraz ze współautorem wyprowadził i zweryfikował numerycznie wszystkie równania.
- Zasoby (wiodący): Zapewnił czas obliczeniowy i pozyskał fundusze na otwarty dostęp.
- Oprogramowanie (wiodący): Zaimplementowane kody do uczenia się zmiennych zbiorowych we wtyczce PLUMED (moduł lowlearner, otwarcie dostępny) oraz dodatkowe metody numeryczne do analizy.
- Nadzór (wiodący): Działał jako autor korespondujący.
- Walidacja (równy): Wraz z współautorem zweryfikował wszystkie wyniki.
- Wizualizacja (wiodący): Wybrał, przygotował i zwizualizował wszystkie wyniki do publikacji.
- Pisanie/Przygotowanie wersji roboczej (wiodący): Napisał wersję roboczą.
- Pisanie/Recenzja i edycja (równy): Poprawiono wersję roboczą z współautorem.

H3. ***Rydzewski, J.**, Chen, M., Ghosh, T.K. & Valsson, O. Reweighted Manifold Learning of Collective Variables from Enhanced Sampling Simulations. *J. Chem. Theory Comput.* **18**, 7179–7192 (2022).
Impact factor: 5.5
Ilość punktów Ministerialnych: 140
Liczba cytowań: 8 [WoS] 10 [GS]

Wzmocnione metody próbkowania są niezbędne w chemii obliczeniowej i fizyce, gdzie symulacje atomistyczne nie mogą wyczerpująco próbkować wielowymiarową przestrzeń konfiguracyjną układów dynamicznych ze względu na problem próbkowania. Klasa takich ulepszonych metod próbkowania działa poprzez identyfikację kilku powolnych stopni swobody, zwanych zmiennymi zbiorowymi, i ulepszenie próbkowania wzdłuż tych zmiennych. Wybór zmiennych zbiorowych do analizy i napędzania próbkowania nie jest trywialny i często opiera się na intuicji chemicznej.

Pomimo rutynowego obchodzenia tej kwestii przy użyciu uczenia się rozmaitości w celu oszacowania zmiennych zbiorowych bezpośrednio ze standardowych symulacji, takie metody nie mogą zapewnić odwzorowań na niskowymiarową rozmaitość z symulacji wzmocnionego próbkowania, ponieważ geometria i gęstość wyuczonej rozmaitości są obciążone. Tutaj zajmujemy się tą kluczową kwestią i zapewniamy ogólną strukturę ponownego ważenia opartą na anizotropowych mapach dyfuzji dla uczenia rozmaitości, która uwzględnia fakt, że zbiór danych do nauki jest próbkowany z obciążonego rozkładu prawdopodobieństwa. Rozważamy metody uczenia rozmaitości oparte na konstruowaniu łańcucha Markowa opisującego prawdopodobieństwa przejścia między próbkami o wysokiej wymiarowości. Pokazujemy, że nasza metoda odwraca efekt obciążenia, dając zmienne zbiorowe, które poprawnie opisują rozkład równowagowy. Ten postęp umożliwia konstruowanie niskowymiarowych zmiennych zbiorowych przy użyciu uczenia rozmaitości bezpośrednio z danych wygenerowanych przez symulacje wzmocnionego próbkowania. Nazywamy naszą metodę *reweighted manifold learning*. Pokazujemy, że może być ona wykorzystywana w wielu technikach uczenia rozmaitości na danych pochodzących zarówno ze standardowych, jak i symulacji wzmocnionego próbkowania.

Wkład:

- Konceptualizacja (wiodący): Opracował teorię, metody, ramy, przykłady i koncepcję ważenia i uczenia rozmaitości z symulacji wzmocnionego próbkowania; omówił koncepcję ze współautorami i odniósł się do ich sugestii.
- Zarządzanie danymi (wiodący): Przygotował i przeprowadził wszystkie symulacje rozszerzonego próbkowania do uczenia zmiennych zbiorowych z tych symulacji.
- Analiza formalna (wiodący): Przeanalizował wszystkie wyniki.
- Metodologia (równy): Opracował i zweryfikował numerycznie wszystkie równania z pomocą współautorów.
- Zasoby (wiodący): Zapewnienie czasu obliczeniowego i pozyskanie funduszy na otwarty dostęp.
- Oprogramowanie (wiodący): Zaimplementowane kody do uczenia we wtyczce PLUMED (moduł lowlearner, swobodnie dostępny) i dodatkowe metody numeryczne do analizy.
- Nadzór (wiodący): Pełnił rolę autora korespondencyjnego oraz organizował spotkania i dyskusje.
- Walidacja (równy): Zatwierdził wszystkie wyniki z pomocą współautorów.

- Wizualizacja (wiodący): Wybrał, przygotował i zwizualizował wszystkie wyniki do publikacji.
- Pisanie/Przygotowanie wersji roboczej (wiodący): Napisał wersję roboczą.
- Pisanie/Recenzja i edycja (równy): Poprawił wersję roboczą z współautorami.

H4. ***Rydzewski, J.** Selecting High-Dimensional Representations of Physical Systems by Reweighted Diffusion Maps. *J. Phys. Chem. Lett.* **14**, 2778—2783 (2023).
Impact factor: 5.7
Ilość punktów Ministerialnych: 200
Liczba cytowań: 4 [WoS] 5 [GS]

Konstruowanie zredukowanych reprezentacji układów wielowymiarowych jest podstawowym problemem w chemii fizycznej. Wiele nienadzorowanych metod uczenia maszynowego może automatycznie znaleźć takie niskowymiarowe reprezentacje. Jednak często pomijanym problemem jest to, jaka wielowymiarowa reprezentacja powinna być użyta do opisu układów przed redukcją wymiarowości. Tutaj zajmujemy się tą kwestią przy użyciu niedawno opracowanej metody zwanej ważoną mapą dyfuzji [J. Chem. Theory Comput. 2022, 18, 7179-7192]. Pokazujemy, w jaki sposób wysokowymiarowe reprezentacje mogą być wybierane ilościowo poprzez badanie rozkładu spektralnego macierzy przejścia Markowa zbudowanych z danych uzyskanych ze standardowych lub wzmocnionych symulacji atomistycznych. Demonstrujemy wydajność metody na kilku przykładach o wysokiej wymiarowości.

Wkład: Konceptualizacja; Zarządzanie danymi; Analiza formalna; Metodologia; Administracja projektem; Zasoby; Oprogramowanie; Nadzór; Walidacja; Wizualizacja; Pisanie/Przygotowanie pierwotnego szkicu; Pisanie/Recenzja & Edycja zostały wykonane wyłącznie przez Autora.

H5. ***Rydzewski, J.** Spectral Map: Embedding Slow Kinetics in Collective Variables. *J. Phys. Chem. Lett.* **14**, 2778—2783 (2023).
Impact factor: 5.7
Ilość punktów Ministerialnych: 200
Liczba cytowań: 3 [WoS] 5 [GS]

Dynamika układów fizycznych, które wymagają wielowymiarowej reprezentacji, może być często uchwycona w kilku znaczących stopniach swobody zwanych zmiennymi zbiorowymi. Identyfikacja zmiennych zbiorowych jest jednak trudna i stanowi fundamentalny problem w chemii fizycznej. Problem ten jest jeszcze bardziej wyraźny, gdy zmienne zbiorowe muszą dostarczać informacji o powolnej kinetyce związanej z rzadkimi przejściami pomiędzy długożyciowymi stanami metastabilnymi. Aby rozwiązać ten problem, proponujemy nienadzorowaną metodę głębokiego

uczenia zwaną mapą spektralną. Nasza metoda konstruuje powolne zmienne zbiorowe poprzez maksymalizację przerwy spektralnej między wolnymi i szybkimi wartościami własnymi macierzy przejścia oszacowanej przez anizotropowe jądro dyfuzji. Demonstrujemy naszą metodę w kilku wysokowymiarowych procesach odwracalnego zwijania białek.

Wkład: Konceptualizacja; Zarządzanie danymi; Analiza formalna; Metodologia; Administracja projektem; Zasoby; Oprogramowanie; Nadzór; Walidacja; Wizualizacja; Pisanie/Przygotowanie pierwotnego szkicu; Pisanie/Recenzja & Edycja zostały wykonane wyłącznie przez Autora.

H6. ***Rydzewski, J.** & Gokdemir, T. Learning Markovian Dynamics with Spectral Maps. *J. Chem. Phys.* **160**, 091102 (2024).
Impact factor: 4
Ilość punktów Ministerialnych: 140
Liczba cytowań: 0 [WoS] 0 [GS]

Długoczasowe zachowanie wielu złożonych układów molekularnych można często opisać za pomocą dynamiki Markowa w powolnej podprzestrzeni rozpiętej przez kilka współrzędnych reakcji, zwanych zmiennymi zbiorowymi. Wyznaczenie zmiennych zbiorowych stanowi jednak fundamentalne wyzwanie w fizyce chemicznej. Poleganie na intuicji lub metodzie prób i błędów w celu skonstruowania zmiennych zbiorowych może prowadzić do dynamiki niemarkowowskiej z efektami pamięci, co utrudnia analizę. Aby rozwiązać ten problem, kontynuujemy rozwój niedawno wprowadzonej techniki głębokiego uczenia zwanej mapą spektralną [J. Rydzewski, J. Phys. Chem. Lett. 14, 5216-5220 (2023)]. Mapa spektralna uczy się powolnych zmiennych zbiorowych poprzez maksymalizację przerwy spektralnej macierzy przejścia Markowa opisującej dyfuzję anizotropową. Aby reprezentować heterogeniczne i wieloskalowe krajobrazy energii swobodnej za pomocą mapy spektralnej, wprowadzamy algorytm adaptacyjny do szacowania prawdopodobieństw przejścia. Poprzez analizę modelu stanów Markowa dowodzimy, że mapa spektralna uczy się powolnych zmiennych zbiorowych związanych z dominującymi skalami czasowymi relaksacji i rozróżnia długotrwałe stany metastabilne.

Wkład:

- Konceptualizacja (wiodący): Opracował teorię, pomysł, metody, przykłady i koncepcję.
- Zarządzanie danymi (równy): Opracował metody do nauki zmiennych zbiorowych z symulacji z współautorem.
- Analiza formalna (wiodący): Przeanalizował wszystkie wyniki.
- Metodologia (wiodący): Opracował i numerycznie zweryfikował wszystkie równania.

- Zasoby (wiodący): Zapewnił czas obliczeniowy.
- Oprogramowanie (wiodący): Zaimplementował kod do nauki zmiennych zbiorowych i dodatkowe metody numeryczne do analizy.
- Nadzór (wiodący): Pełnił funkcję autora korespondencyjnego.
- Walidacja (równy): Wraz z współautorem zatwierdził wszystkie wyniki.
- Wizualizacja (wiodący): Wybrał, przygotował i zwizualizował wszystkie wyniki do publikacji.
- Pisanie/Przygotowanie wersji roboczej (wiodący): Napisanie wersji roboczej.
- Pisanie/Recenzja i edycja (równy): Poprawił wersję roboczą z współautorem.

H7. **Rydzewski, J.** as part of PLUMED Consortium[1]. Promoting Transparency and Reproducibility in Enhanced Molecular Simulations. *Nat. Methods* **16**, 670-–673 (2019).
Impact factor: 48
Ilość punktów Ministerialnych: 200
Liczba cytowań: 543 [WoS] 489 [GS]

Konsorcjum PLUMED zrzesza programistów i współtwórców PLUMED, biblioteki open-source do obliczeń dynamiki molekularnej z wzmocnionym próbkowaniem, obliczeń energii swobodnej i analizy symulacji dynamiki molekularnej. Poniżej przedstawiamy nasze wysiłki na rzecz promowania przejrzystości i odtwarzalności poprzez rozpowszechnianie protokołów dla symulacji molekularnych z zaawansowanym próbkowaniem. Konsorcjum PLUMED zapewnia algorytm automatycznej weryfikacji danych wejściowych do symulacji atomistycznych, interfejs swobodnie otwartego repozytorium danych symulacyjnych oraz narzędzia umożliwiające odtwarzanie wyników z publikacji.

Wkład:

- Konceptualizacja (wspierający): Uczestniczył we wszystkich dyskusjach i spotkaniach online.
- Zarządzanie danymi (wspierający): Zaimplementował testy regresji w celu weryfikacji wyników uzyskanych przy użyciu modułu maze do ulepszonego próbkowania ścieżek dysocjacji ligandów i dostarczenie dokumentacji (https://www.plumed.org/doc-v2.8/user-doc/html/_m_a_z_e.html).
- Metodologia (wspierający): Omówił zasady odtwarzalności i przedstawił swoje sugestie.

[1]Full list of contributors: plumed-nest.org/consortium.html

- Oprogramowanie (równy): Zaimplementował moduł maze w PLUMED w celu ulepszonego próbkowania ścieżek dysocjacji ligandów - jeden z najbardziej zaawansowanych modułów PLUMED, zawierający kilka różnych algorytmów. Moduł jest dostępny bezpłatnie w każdym wydaniu PLUMED od wersji 2.7 (https://github.com/plumed/plumed2/tree/master/src/maze).
- Pisanie/Przygotowanie wersji roboczej (wspierający): Napisał o module maze.
- Pisanie/Recenzja i edycja (równy): Redagował tekst wraz z współautorami.

II. Informacja o aktywności naukowej

II.1. Wykaz artykułów opublikowanych w czasopismach naukowych, niewymienionych w punkcie I.1

Informacja o wskaźniku Impact Factor została zaczerpnięta z bazy Web of Science (WoS), a gwiazdką przy nazwisku oznaczono publikacje, w których kandydat jest autorem korespondencyjnym.

Wkład autora jest przypisywany zgodnie z klasyfikacją Contributor Roles Taxonomy (CRediT) (https://credit.niso.org/) stosowaną przez American Institute of Physics.

A1. ***Rydzewski, J.** maze: Heterogeneous Ligand Unbinding along Transient Protein Tunnels. *Comput. Phys. Commun.* **247**, 106865 (2020).
Impact factor: 6.3
Ilość punktów Ministerialnych: 140
Liczba cytowań: 9 [WoS] 7 [GS]

Niedawne osiągnięcia w metodach wzmocnionego próbkowania pokazały, że możliwe jest zrekonstruowanie ścieżek dysocjacji ligandów z rozdzielczością przestrzenną i czasową niedostępną dla eksperymentów. Idealnie byłoby, gdyby takie techniki zapewniały atomistyczną definicję możliwie wielu ścieżek reakcji, ponieważ prymitywne szacunki mogą prowadzić albo do przeszacowania barier energetycznych, albo do niemożności próbkowania ukrytych barier energetycznych. Przedstawiamy implementację nowej metody [Rydzewski and Valsson, J. Chem. Phys. 150, 221101 (2019)] przeznaczonej w całości do próbkowania ścieżek reakcji procesu dysocjacji ligand-białko. Program, nazwany maze, jest zaimplementowany jako oficjalny moduł dla PLUMED 2, biblioteki open source do wzmocnionego próbkowania w układach molekularnych, i zawiera algorytmy do znajdowania wielu heterogenicznych ścieżek reakcji dysocjacji liganda z białek podczas symulacji atomistycznych. Moduł maze wymaga jedynie struktury krystalograficznej do rozpoczęcia symulacji i nie zależy od wielu parametrów. Program opiera się na wzmocnionym próbkowaniu i niewypukłych metodach optymalizacji. Aby zaprezentować jego zastosowanie i elasty-

czność, przedstawiamy kilka przykładów ścieżek wiązania ligandów wzdłuż przejściowych tuneli białkowych zrekonstruowanych przez maze w modelowym układzie ligand-białko i omawiamy szczegóły implementacji.

Wkład: Konceptualizacja; Zarządzanie danymi; Analiza formalna; Metodologia; Administracja projektem; Zasoby; Oprogramowanie; Nadzór; Walidacja; Wizualizacja; Pisanie/Przygotowanie wersji roboczej; Pisanie/Recenzja& Edycja zostały wykonane wyłącznie przez Autora.

A2. ***Rydzewski, J.** & Valsson, O. Finding multiple reaction pathways of ligand unbinding. *J. Chem. Phys.* **150** (2019).
Impact factor: 4.4
Ilość punktów Ministerialnych: 100
Liczba cytowań: 21 [WoS] 27 [GS]

Poszukiwanie ścieżek reakcji opisujących rzadkie zdarzenia w dużych układach stanowi długotrwałe wyzwanie w chemii i fizyce. Nieprawidłowo obliczone ścieżki reakcji skutkują degeneracją konfiguracji mikroskopowych i niemożnością próbkowania ukrytych barier energetycznych. W tym celu przedstawiamy ogólną metodę wzmocnionego próbkowania w celu znalezienia wielu różnych ścieżek reakcji odłączania ligandu poprzez niewypukłą optymalizację funkcji straty opisującej interakcje ligand-białko. Metoda z powodzeniem pokonuje duże bariery energetyczne przy użyciu adaptacyjnego potencjału stronniczości i konstruuje możliwe ścieżki reakcji wzdłuż tuneli przejściowych bez wstępnego zgadywania stanów pośrednich lub końcowych, wymagając jedynie informacji krystalograficznych. Badamy metodę na mutancie lizozymu T4 L99A, który jest często używany jako system modelowy do badania wiązania ligandów z białkami, zapewniamy nieznaną wcześniej ścieżkę reakcji i pokazujemy, że dzięki wykorzystaniu potencjału polaryzacji i szerokości tunelu możliwe jest uchwycenie heterogeniczności mechanizmów odłączania pomiędzy znalezionymi przejściowymi tunelami białkowymi.

Wkład:

- Konceptualizacja (wiodąca): Opracował teorię, metody, ramy, przykłady i koncepcję próbkowania CV dla zdarzeń dysocjacji ligandów ze wzmocnionych symulacji próbkowania; omówił koncepcję ze współautorem i uwzględnił jego sugestie.
- Zarządzanie danymi (wiodący): Przygotował i przeprowadził wszystkie symulacje wzmocnionego próbkowania.
- Analiza formalna (wiodący): Przeanalizował wszystkie wyniki.
- Zasoby (wiodąca): Zapewnił czas obliczeniowy.

- Oprogramowanie (wiodący): Zaimplementował kody do próbkowania zmiennych zbiorowych opisujących dysocjację ligandów we wtyczce PLUMED i dodatkowe metody numeryczne do analizy.
- Nadzór (wiodący): Pełnił rolę autora korespondencyjnego oraz organizował spotkania i dyskusje.
- Walidacja (równy): Zatwierdził wszystkie wyniki wraz ze współautorem.
- Wizualizacja (wiodący): Wybrał, przygotował i zwizualizował wszystkie wyniki do publikacji.
- Pisanie/Przygotowanie wersji roboczej (wiodący): Napisał wersję roboczą.
- Pisanie/Recenzja i edycja (wiodący): Poprawił wersję roboczą zgodnie z sugestiami współautora.

A3. ***Rydzewski, J.**, Jakubowski, R., Nowak, W. & Grubmuller, H. Kinetics of Huperzine A Dissociation from Acetylcholinesterase via Multiple Unbinding Pathways. *J. Chem. Theory Comput.* **14**, 2843–2851 (2018).
Impact factor: 5.5
Ilość punktów Ministerialnych: 140
Liczba cytowań: 21 [WoS] 31 [GS]

Dysocjacja huperziny A (hupA) z acetylocholinoesterazy Torpedo californica (TcAChE) została zbadana za pomocą 4 μs nieobciążonych i obciążonych symulacji dynamiki molekularnej (MD). Przeprowadziliśmy nasze badanie przy użyciu próbkowania memetycznego (MS) w celu określenia ścieżek reakcji (RP), metadynamiki do obliczenia energii swobodnej i estymacji maksymalnego prawdopodobieństwa (MLE) w celu odzyskania szybkości kinetycznych z nieobciążonych symulacji MD. Nasze symulacje sugerują, że dysocjacja hupA zachodzi głównie przez dwa RP: przednie drzwi wzdłuż osi wąwozu miejsca aktywnego (pwf) i przez nowe przejściowe drzwi boczne (pws), tj. utworzone przez pętlę Ω (67-94 TcAChE). Analiza wiązania inhibitora wzdłuż RP sugeruje, że pws jest otwierane przejściowo po tym, jak hupA i pętla Ω osiągną stan przejściowy o niskiej energii swobodnej, charakteryzujący się orientacją grupy pirydonowej inhibitora skierowaną w stronę płaszczyzny pętli Ω. W przeciwieństwie do pws, pwf nie wymaga dużych zmian strukturalnych w TcAChE, aby być dostępnym. Oszacowane energie swobodne i szybkości dysocjacji są zgodne z dostępnymi danymi eksperymentalnymi. Szybkości dysocjacji wzdłuż ścieżek wiązania są podobne, co sugeruje, że dysocjacja hupA wzdłuż pws jest prawdopodobnie istotna. Wskazuje to, że perturbacje w interakcjach hupA-TcAChE mogą potencjalnie indukować przeskakiwanie ścieżek. Podsumowując, nasze wyniki charakteryzują powolne hamowanie TcAChE przez hupA,

co może zapewnić strukturalne i energetyczne podstawy do racjonalnego projektowania powolnych inhibitorów następnej generacji o zoptymalizowanych właściwościach farmakokinetycznych do leczenia choroby Alzheimera.

Wkład:

- Konceptualizacja (wiodący): Opracował teorię, metody, ramy, przykłady i koncepcję próbkowania wielu zdarzeń dysocjacji oraz obliczania profili energii swobodnej i kinetyki na podstawie wzmocnionych symulacji próbkowania; omówił koncepcję ze współautorami i uwzględnił ich sugestie.
- Zarządzanie danymi (wiodący): Przygotował i przeprowadził wszystkie symulacje i analizy wzmocnionego próbkowania.
- Analiza formalna (wiodący): Przeanalizował wszystkie wyniki.
- Metodologia (wiodący): Opracował i zweryfikował numerycznie wszystkie równania.
- Zasoby (wiodący): Zapewnienie czasu obliczeniowego.
- Oprogramowanie (wiodący): Zaimplementował kody do obliczania wielu ścieżek dysocjacji, szybkości kinetycznych i dodatkowych metod numerycznych do analizy.
- Nadzór (równy): Pełnił rolę autora korespondencyjnego oraz organizował spotkania i dyskusje.
- Walidacja (wiodący): Zatwierdził wszystkie wyniki.
- Wizualizacja (wiodący): Wybrał, przygotował i zwizualizował wszystkie wyniki do publikacji.
- Pisanie/Przygotowanie wersji roboczej (wiodący): Napisanie wersji roboczej.
- Pisanie/Recenzja i edycja (równy): Poprawił wersję roboczą ze współautorami.

A4. ***Rydzewski, J.**, Walczewska-Szewc, K., Czach, S., Nowak, W. & Kuczera, K. Enhancing the Inhomogeneous Photodynamics of Canonical Bacteriophytochrome. *J. Phys. Chem. B* **126**, 2647—2657 (2022).
Impact factor: 3.3
Ilość punktów Ministerialnych: 140
Liczba cytowań: 2 [WoS] 5 [GS]

Zdolność fitochromów do działania jako fotoprzełączniki w roślinach i mikroorganizmach zależy od interakcji między chromoforem podobnym do biliny a białkiem gospodarza. Interkonwersja zachodzi pomiędzy spektralnie różnymi konformerami czerwonym (Pr) i dalekiej czerwieni (Pfr). Ta

zmiana konformacyjna jest wywoływana przez fotoizomeryzację pirolu pierścienia D chromoforu. W tym badaniu, jako reprezentatywny przykład układu fitochrom-bilina, rozważamy biliwerdynę IXα (BV) związaną z bakteriofytochromem (BphP) z Deinococcus radiodurans. W przypadku braku światła stosujemy metodę dynamiki molekularnej (MD) ze wzmocnionym próbkowaniem, aby pokonać barierę energetyczną fotoizomeryzacji. Stwierdzamy, że obliczone bariery energii swobodnej (FE) między podstawowymi stanami metastabilnymi są zgodne z wynikami spektroskopowymi. Pokazujemy, że zwiększona dynamika chromoforu BV w BphP przyczynia się do wyzwalania ruchów konformacyjnych w skali nanometrowej, które rozprzestrzeniają się przez dwie eksperymentalnie określone ścieżki transdukcji sygnału. Co najważniejsze, opisujemy, w jaki sposób stany metastabilne umożliwiają przejście termiczne znane jako ciemna rewersja między Pfr i Pr, poprzez nieznany wcześniej stan pośredni Pfr. Przedstawiamy heterogeniczność zależnych od temperatury stanów Pfr na poziomie atomistycznym. Praca ta toruje drogę do zrozumienia pełnego mechanizmu fotoizomeryzacji chromoforu podobnego do biliny w fitochromach.

Wkład:

- Konceptualizacja (wiodący): Opracował koncepcję zwiększenia fluktuacji BV; omówił koncepcję ze współautorami i uwzględnił ich sugestie.
- Zarządzanie danymi (wiodący): Przygotował i przeprowadził wszystkie symulacje wzmocnionego próbkowania.
- Analiza formalna (wiodący): Przeanalizował wszystkie wyniki.
- Metodologia (wiodący): Wybór wszystkich metod.
- Zasoby (wiodący): Zapewnienie czasu obliczeniowego i pozyskanie funduszy na otwarty dostęp.
- Oprogramowanie (wiodący): Zaimplementował metody numeryczne do analizy.
- Nadzór (równy): Pełnił rolę autora korespondencyjnego oraz organizował spotkania i dyskusje.
- Walidacja (równy): Zatwierdził wszystkie wyniki ze współautorami.
- Wizualizacja (równy): Wybrał, przygotował i zwizualizował wszystkie wyniki do publikacji wraz ze współautorami.
- Pisanie/Przygotowanie wersji roboczej (wiodący): Napisał wersję roboczą.
- Pisanie/Recenzja i edycja (równy): Poprawił wersję roboczą wraz ze współautorami.

A5. Walczewska-Szewc, K., **Rydzewski, J.** & Lewkowicz, A. Inhibition-Mediated Changes in Prolyl Oligopeptidase Dynamics Possibly Related to α-Synuclein Aggregation. *Phys. Chem. Chem.*

Phys. **24**, 4366—4373 (2022).
Impact factor: 3.3
Ilość punktów Ministerialnych: 100
Liczba cytowań: 2 [WoS] 4 [GS]

Tworzenie się agregatów białkowych jest jedną z głównych przyczyn nieprawidłowego funkcjonowania neuronów i późniejszego uszkodzenia mózgu w wielu chorobach neurodegeneracyjnych. W chorobie Parkinsona w akumulację agregatów zaangażowane są β-synukleiny. Pochodzenie agregacji jest nieznane, ale istnieją przekonujące dowody na to, że można ją zmniejszyć poprzez hamowanie oligopeptydazy prolilowej (PREP). Efekt ten nie może być po prostu związany z hamowaniem funkcji katalitycznej enzymu, ponieważ nie wszystkie inhibitory PREP zatrzymują agregację α-synukleiny. Znalezienie różnic w dynamice enzymu hamowanego przez różne związki pozwoliłoby nam zidentyfikować regiony białkowe zaangażowane w interakcję między PREP a α-synukleiną. Tutaj badamy wpływ trzech inhibitorów PREP, z których każdy wpływa na agregację α-synukleiny w różnym stopniu. Wykorzystujemy modelowanie dynamiki molekularnej, aby zidentyfikować mechanizmy molekularne leżące u podstaw inhibicji PREP i znaleźć różnice strukturalne między układami inhibitor-PREP. Sugerujemy, że nawet subtelne zmiany w dynamice enzymu wpływają na jego interakcje z α-synukleinami. Nasza identyfikacja tych regionów może być zatem biologicznie istotna w zapobieganiu tworzeniu się agregatów α-synuklein.

Wkład:

- Konceptualizacja (wspierający): Planował badania związane ze wzmocnionym próbkowaniem zdarzeń dysocjacji ligandów.
- Zarządzanie danymi (równy): Przygotował i przeprowadził symulacje wzmocnionego próbkowania w celu odłączenia kilku ligandów.
- Analiza formalna (równy): Analizował wyniki związane ze wzmocnionym próbkowaniem zdarzeń dysocjacji ligandów i profili energii swobodnej.
- Metodologia (równy): Numerycznie zweryfikował wyniki związane ze wzmocnionym próbkowaniem zdarzeń dysocjacji ligandów.
- Oprogramowanie (wspierające): Wdrożył dodatkowe metody numeryczne do analizy.
- Walidacja (równy): Zatwierdził wszystkie wyniki związane ze wzmocnionym próbkowaniem zdarzeń dysocjacji ligandów.
- Wizualizacja (wspierający): Przygotował i zwizualizował wszystkie wyniki związane ze wzmocnionym próbkowaniem zdarzeń dysocjacji ligandów.

- Pisanie/Przygotowanie wersji roboczej (wspierający): Napisał wersje robocze części związanych ze wzmocnionym próbkowaniem zdarzeń dysocjacji ligandów.
- Pisanie/Recenzja i edycja (równy): Poprawił wersję roboczą ze współautorami.

A6. Niklas, B., **Rydzewski, J.**, Lapied, B. & Nowak, W. Toward Overcoming Pyrethroid Resistance in Mosquito Control: The Role of Sodium Channel Blocker Insecticides. *Int. J. Mol. Sci.* **24**, 10334 (2023).
Impact factor: 5.6
Ilość punktów Ministerialnych: 140
Liczba cytowań: 0 [WoS] 0 [GS]

Choroby przenoszone przez komary prowadzą do śmierci 700 000 osób rocznie. Głównym sposobem ograniczenia transmisji jest kontrola wektorów poprzez zapobieganie ukąszeniom za pomocą środków chemicznych. Jednak najczęściej stosowane środki owadobójcze tracą skuteczność z powodu rosnącej odporności. Kanały sodowe bramkowane napięciem (VGSC), białka błonowe odpowiedzialne za fazę depolaryzacji potencjału czynnościowego, są celem szerokiej gamy neurotoksyn, w tym pyretroidów i insektycydów blokujących kanały sodowe (SCBI). Zmniejszona wrażliwość białka docelowego z powodu mutacji punktowych zagroziła kontroli malarii za pomocą pyretroidów. Chociaż SCBI - indoksakarb (pre-insektycyd bioaktywowany do DCJW u owadów) i metaflumizon - są stosowane wyłącznie w rolnictwie, pojawiają się jako obiecujący kandydaci w kontroli komarów. Dlatego dokładne zrozumienie molekularnych mechanizmów działania SCBI jest pilnie potrzebne, aby przełamać oporność i zatrzymać przenoszenie chorób. W tym badaniu, wykonując obszerną kombinację symulacji dynamiki molekularnej i wzmocnionego próbkowania, stwierdziliśmy, że fenestracja DIII-DIV jest najbardziej prawdopodobną drogą wejścia DCJW do centralnej wnęki VGSC komara. Nasze badania wykazały, że F1852 ma kluczowe znaczenie w ograniczaniu dostępu SCBI do ich miejsca wiązania. Nasze wyniki wyjaśniają rolę mutacji F1852T występującej u opornych owadów oraz zwiększoną toksyczność DCJW w porównaniu z jego większym związkiem macierzystym, indoksakarbem. Określiliśmy również reszty, które przyczyniają się zarówno do wiązania SCBI, jak i nieestrowego pyretroidu etofenproksu, a tym samym mogą być zaangażowane w odporność krzyżową w miejscu docelowym.

Wkład:

- Konceptualizacja (wspierający): Zaplanował badania związane ze wzmocnionym próbkowaniem zdarzeń dysocjacji ligandów.
- Zarządzanie danymi (równy): Przygotował i przeprowadził symulacje wzmocnionego próbkowania w celu odłączenia kilku ligandów.

- Analiza formalna (równy): Zanalizował wyniki związane ze wzmocnionym próbkowaniem zdarzeń dysocjacji ligandów i profili energii swobodnej.
- Metodologia (równy): Numerycznie zweryfikował wyniki związane ze wzmocnionym próbkowaniem zdarzeń dysocjacji ligandów.
- Oprogramowanie (wspierający): Wdrożył dodatkowe metody numeryczne do analizy.
- Walidacja (równy): Zatwierdził wszystkie wyniki związane ze wzmocnionym próbkowaniem zdarzeń dysocjacji ligandów.
- Wizualizacja (wspierający): Przygotował i zwizualizował profile energii swobodnej dysocjacji ligandów.
- Pisanie/Przygotowanie wersji roboczej (wspierający): Napisał oryginalne szkice związane ze wzmocnionym próbkowaniem zdarzeń dysocjacji ligandów.
- Pisanie/Recenzja i edycja (równy): Poprawił wersję roboczą wraz ze współautorami.

Publikacje w recenzowanych czasopismach przed uzyskaniem stopnia doktora: 12 publikacji [B1-B12], głównie koncentrujących się na opracowaniu wzmocnionych technik próbkowania w celu obserwacji ścieżek dysocjacji ligandów z białek.

B1. Carrascoza Mayén, J. F., **Rydzewski, J.**, Szostak, N., *Blazewicz, J., and *Nowak, W. Prebiotic Soup Components Trapped in Montmorillonite Nanoclay Form New Molecules: Car-Parrinello Ab Initio Simulations. *Life* **9**, 46 (2019).

B2. ***Rydzewski, J.**, and Nowak W. Photoinduced Transport in an H64Q Neuroglobin Antidote for Carbon Monoxide Poisoning. *J. Chem. Phys.* **148**, 115101 (2018).

B3. ***Rydzewski, J.**, Jakubowski, R., Nicosia G., and Nowak, W. Conformational Sampling of a Biomolecular Rugged Energy Landscape. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* **15**, 732 (2018).

B4. ***Rydzewski, J**. and Nowak, W. Rare-Event Sampling in Ligand Diffusion. *Phys. Life Rev.* **22-23**, 85 (2017).

B5. ***Rydzewski, J**. and Nowak, W. Ligand Diffusion in Proteins via Enhanced Sampling in Molecular Dynamics. *Phys. Life Rev.* **22-23**, 58 (2017).

B6. ***Rydzewski, J**. and Nowak, W. Thermodynamics of Camphor Migration in Cytochrome P450cam by Atomistic Simulations. *Sci. Rep.* **7**, 7736 (2017).

B7. ***Rydzewski, J**. and Nowak, W. Machine Learning Based Dimensionality Reduction Facilitates Ligand Diffusion Paths Assessment: A Case of Cytochrome P450cam. *J. Chem. Theory Comput.* **12**, 2110 (2016).

B8. ***Rydzewski, J**. and Nowak, W. Molecular Dynamics Simulations of Large Systems in Electronic Excited States. *Handbook of Computational Chemistry* (Vol. I), Second Edition (2016) (Editior: J. Leszczynski).

B9. ***Rydzewski, J.**, Nowak, W., Nicosia G. Inferring Pathological States in Cortical Neuron Microcircuits. *J. Theor. Biol.* **386**, 34 (2015).

B10. **Rydzewski, J.**, Jakubowski, R., and *Nowak, W. Entropic Measure to Prevent Energy Over-Minimization in Molecular Dynamics Simulations, *J. Chem. Phys.* **143**, 171103 (2015).

B11. **Rydzewski, J**. and *Nowak, W. Memetic Algorithms for Ligand Expulsion from Protein Cavities. *J. Chem. Phys.* **143**, 124101 (2015).

B12. **Rydzewski, J.**, Strzałka, W., and *Nowak, W. Nanomechanics of PCNA: A Protein-Made DNA Sliding Clamp. *Chem. Phys. Lett.* **634**, 236 (2015).

II.3. Działania popularnonaukowe i promocyjne

- Wykład inauguracyjny dla studentów I roku Wydziału Fizyki, Astronomii i Informatyki Uniwersytetu Mikołaja Kopernika w Toruniu, październik 2018.
- Metody wzmocnionego próbkowania (w języku polskim) opublikowane w Głosie Uczelni 7-10, 2019, dystrybuowane przez Uniwersytet Mikołaja Kopernika w Toruniu.
- Notatka popularyzatorska dotycząca promocji grantów oferowanych przez Narodowe Centrum Nauki, 10-lecie Narodowego Centrum Nauki, 2021 r.
- Letni obóz dla studentów zagranicznych i doktorantów w ramach programu Narodowej Agencji Wymiany Akademickiej (NAWA) SPINAKER - Intensywne Międzynarodowe Programy Edukacyjne, 12-14 lipca 2022 r. w Toruniu.
- Promowanie otwartej nauki w ramach projektu Narodowej Agencji Wymiany Akademickiej "Open NCU–Open Source, Open Science" poprzez finansowanie opłat za publikacje w otwartym dostępie dla młodych naukowców z Uniwersytetu Mikołaja Kopernika w Toruniu.
- Prowadzenie zajęć na temat otwartej nauki w ramach programu "Opracowanie i wdrożenie praktycznego szkolenia na temat otwartej nauki i otwartych innowacji dla początkujących naukowców

(DIOSI)" finansowanego przez Unię Europejską w ramach programu Horyzont 2020, Komisja Europejska 101006318, 2020-2023.

II.4. Informacja o wystąpieniach na krajowych lub międzynarodowych konferencjach naukowych

Przedstawiona poniżej lista obejmuje wyłącznie wystąpienia konferencyjne/naukowe, np. wykłady zaproszone, seminaria oraz prezentacje ustne i plakatowe Autora *po uzyskaniu stopnia doktora*.

Wystąpienia konferencyjne/naukowe:

1. Spectral Map: Embedding Slow Kinetics in Collective Variables, *Molecular Simulation Society of Japan 2023*, 2023-12-4 – 2023-12-6, Fukui, Japonia (Zaproszenie).
2. Recent Advances in Reweighted Manifold Learning for Molecular Dynamics, *Rare Event Workshop 2023*, 2023-12-3, Kanazawa, Japonia (Zaproszenie).
3. Reweighted Manifold Learning for Simulations and Enhanced Sampling, *34th IUPAP Conference on Computational Physics (CCP2023)*, 2023-08-04 – 2023-08-08, Kobe, Japonia.
4. Multiscale Reweighted Stochastic Embedding: Deep Learning of Collective Variables for Enhanced Sampling, *Recent Advances in Modelling Rare Events*, 2021-12-15 – 2021-12-18, wirtualna konferencja, Indie.
5. Multiscale Reweighted Stochastic Embedding (MRSE): Deep Learning of Generalized Variables for Statistical Physics, *ML in PL (Machine Learning in Poland)*, 2021-11-05 – 2021-11-07, Warszawa, Polska.
6. Multiscale Reweighted Stochastic Embedding: Deep Learning of Collective Variables for Enhanced Sampling, *Winter Workshop on Multiscale Modeling, Karlsruhe Institute of Technology (KIT)*, 2021-11-22 – 2021-11-23, wirtualna konferencja, Niemcy.
7. Multiscale Reweighted Stochastic Embedding: Deep Learning of Collective Variables for Enhanced Sampling, *American Physical Society March Meeting 2021*, 2021-03-15 – 2021-03-19, wirtualna konferencja, Stany Zjednoczone.
8. Multiscale Reweighted Stochastic Embedding: Deep Learning of Collective Variables for Enhanced Sampling, *Bunsen-Tagung 2021*, 2021-05-10 – 2021-05-12, wirtualna konferencja, Niemcy.
9. Sampling Ligand Unbinding Pathways, *Bioinformatics in Torun (BIT)*, 2019-06-21 – 2019-06-25, Toruń, Polska.

Plakaty:

1. Enhanced Sampling Methods for Ligand Unbinding, *Mainz Materials Simulation Days 2019 (CECAM Event)*, 2019-06-05 – 2019-06-07, Moguncja, Niemcy.

Zaproszone seminaria:

1. Manifold Learning for Atomistic Simulations, *The National Institute of Advanced Industrial Science and Technology (AIST)*, 2023-11-23, Tsukuba, Japonia.
2. Heterogeneous Ligand Unbinding from Proteins, *Insitutue of Computational Sciences*, 2019, Tsukuba, Japonia.
3. Multiscale Reweighted Stochastic Embedding: Deep Learning of Collective Variables for Enhanced Sampling, *Soft Matter Seminar*, 2019, Amsterdam, Holandia.

II.5. Udział w komitetach organizacyjnych i naukowych na konferencjach

- Organizacja konferencji Bioinformatyka w Toruniu organizowanej wspólnie przez Polskie Towarzystwo Bioinformatyczne (PTBI) i Uniwersytet Mikołaja Kopernika w Toruniu, 2014-2023.
- Recenzent konkursu Polskiego Towarzystwa Bioinformatycznego na najlepszą pracę licencjacką z bioinformatyki i biologii obliczeniowej, 2022.
- Recenzent Annual Conference on Machine Learning, Optimization and Data Science (LOD), 2018–2020.

II.6. Udział w pracach zespołów badawczych realizujących projekty finansowane w drodze konkursów krajowych lub zagranicznych

Kierownik projektu:

1. Sonata 2021/43/D/ST4/00920, Narodowe Centrum Nauki (NCN), *Statistical Learning of Slow Collective Variables from Atomistic Simulations*, 2022–.
2. Wyłaniające się pola badawcze "Initiative of Excellence — Research University," *Nanoscale Biophysics*, Co-PI, 2022– .
3. Promocja Zagraniczna PPI/PZA/2019/1/00037, Narodowe Centrum Wymiany Akademickiej, *Open Source–Open Science*, 2019—2021.
4. Etiuda 2016/20/T/ST3/00488, Narodowe Centrum Nauki (NCN), *Reconstruction of Ligand Unbiding Pathways from Proteins*, 2016–2017.
5. Preludium 2015/19/N/ST3/02171, Narodowe Centrum Nauki (NCN) *Memetic Algorithms for Ligand Expulsion in Electronic Excited State from Condensed Biological Matter*, 2016–2019.

Wykonawca:

1. NIH Grant P30GM110761, National Institute of Health in US, 2018–2022, PI: Prof. K. Kuczera (University of Kansas).
2. Opus 2016/23/B/ST4/01770, Narodowe Centrum Nauki (NCN), *Structural Determinants of Optical Control of Insulin and Neuroligin Release by Photoactive Protein Ligands*, 2017–2022, PI: Prof. W. Nowak (Uniwersytet Mikołaja Kopernika).
3. Grant N202 262038 by the Ministry of Science and Higher Education in Poland, *Nanomechanics of Modular Adhesive Proteins*, 2010–2014, PI: Prof. W. Nowak (Uniwersytet Mikołaja Kopernika).
4. LIDER/28/54/L-/10/NCBiR/2011, The National Centre for Research and Development, *Searching for a Novel Anticancer Drug based on PCNA Inhibitors*, 2013–2015, PI: Prof. W. Strzałka (Uniwersytet Jagielloński).
5. Grant POKL.04.01.01-00-081/10 by the European Union, *Enhancing Educational Potential of Nicolaus Copernicus University in the Disciplines of Mathematical and Natural Sciences*, 2015, PI: Uniwersytet Mikołaja Kopernika.
6. Horizon 2020 by European Commission 101006318, *Developing and Implementing hands-on training on Open Science and Open Innovation for Early Career Researchers (DIOSI)*, 2021–2022, PI: Universytet w Antwerpii.

II.7. Informacja o udziale w programach europejskich lub innych programach międzynarodowych

1. Prowadzenie zajęć na letnim obozie dla zagranicznych studentów i doktorantów w ramach programu Narodowej Agencji Wymiany Akademickiej SPINAKER - Intensywne Międzynarodowe Programy Edukacyjne 4-22 lipca 2022 r. współfinansowane przez European Social Fund under the Operational Programme Knowledge Education Development.
2. Prowadzenie zajęć na temat otwartej nauki w ramach programu “Developing and Implementing hands-on training on Open Science and Open Innovation for Early Career Researchers (DIOSI)” finansowany przez UE w ramach programu the Horizon 2020 (European Commission 101006318, 2020–2023).
3. Współpraca z Uniwersytetem w Taorminie we Włoszech w ramach europejskiego programu “Enhancing Educational Potential of Nicolaus Copernicus University in the Disciplines of Mathematical and Natural Sciences” (POKL.04.01.01-00-081/10, 2015)

4. Koordynacja działań pozwalających na finansowanie opłat za publikacje w otwartym dostępie dla młodych naukowców w Uniwersytetu Mikołaja Kopernika jako PI projektu Polskiej Narodowej Agencji Wymiany Akademickiej "Open NCU–Open Source, Open Science" (2019–2022).

II.8. Członkostwo w międzynarodowych lub krajowych organizacjach i towarzystwach naukowych wraz z informacją o pełnionych funkcjach

1. Konsorcjum PLUMED (2019–): otwarta społeczność dla naukowców, których praca opiera się na symulacjach atomistycznych, ustanawiająca skuteczne protokoły udostępniania danych, promująca odtwarzalność naukową i przestrzegająca najwyższych standardów badawczych (www.plumed-nest.org/consortium).
2. Polskie Towarzystwo Bioinformatyczne (PTBI) (2014–2019).

II.9. Informacja o odbytych stażach w instytucjach naukowych, w tym zagranicznych, z podaniem miejsca, terminu, czasu trwania stażu i jego charakteru

Staże naukowe:

1. National Institute of Advanced Industrial Science and Technology, Tsukuba, Japonia, grupa Prof. Tetsuyi Morishity (XI 2023–II 2024).
2. Department of Theoretical and Computational Biophysics, Max Planck Institute for Biophysical Chemistry, Getynga, Niemcy, grupa Prof. Helmuta Grubmüllera (X 2016–IV 2017).
3. Department of Chemistry and Applied Biosciences, Swiss Federal Institute of Technology in Zürich c/o Institute of Computational Science, Università della Svizzera italiana, Lugano, Szwajcaria, grupa Prof. Michele Parrinello (VII 2016–X 2016).

Wizyty naukowe (do miesiąca):

1. Department of Physics, University of Tsukuba, Tsukuba, Japonia, grupa Prof. Yasuteru Shigety, 2019 (1 miesiąc).
2. Max Planck Institute for Polymer Research, Moguncja, Niemcy, grupa Prof. Omara Valssona, 2019 (1 tydzień).
3. Quantum Chemistry Research Institute, Kioto, Japonia, 2017, grupa Prof. Hiroshi'ego Nakatsuji (1 miesiąc).
4. Institute of Fluid-Flow Machinery - Polska Akademia Nauk, Gdańsk, 2011 (1 miesiąc).

II.10. Informacja o recenzowanych pracach naukowych lub artystycznych, w szczególności publikowanych w czasopismach międzynarodowych

1. American Chemical Society (ACS): Journal of Physical Chemistry Letters, Journal of Chemical Information and Modeling, Journal of Chemical Theory and Computation, Journal of Physical Chemistry B
2. American Institute of Physics (AIP): Journal of Chemical Physics
3. Cell Press: Biophysical Journal
4. Elsevier: Computational and Structural Biotechnology Journal, Journal of Molecular Graphics and Modeling
5. PLOS: PLOS ONE, PLOS Computational Biology
6. Journal of Open-Source Software

II.11. Informacja o udziale w zespołach badawczych, realizujących projekty inne niż określone w pkt. II.6

Autor uczestniczy w kilku projektach badawczych we współpracy z grupami naukowymi:

1. Prof. Omar Valsson, University of North Texas, Denton, USA, od 2016 r. Praca związana z konstruowaniem zmiennych zbiorowych do wzmocnionego próbkowania symulacji przy użyciu optymalizacji i wyjaśnialnego uczenia maszynowego [H2, H3, H1, A2].
2. Prof. Ming Chen, Uniwersytet Purdue, West Lafayette, USA, od 2018. Praca związana z konstruowaniem macierzy przejścia Markowa na podstawie wzmocnionego próbkowania symulacji [H3, H1].
3. Prof. Alexander M. Berezhkovski, National Institutes of Health, Bethesda, Maryland, USA, od 2023 r. Praca związana z dynamiką Markowa wzdłuż współrzędnych reakcji.
4. Prof. Tetsuya Morishita, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japonia, od 2022 r. Prace związane z obliczaniem makroskopowych zmiennych opisujących tworzenie się szkła.
5. Prof. Michele Parrinello, Italian Institute of Technology, Genua, Włochy, od 2016 roku. Praca związana z opracowywaniem wzmocnionych metod próbkowania do symulacji wiązania ligandów z białek i ustanawianiem protokołów do przejrzystych i powtarzalnych symulacji atomistycznych [H7].
6. Prof. Helmut Grubmüller, Max Planck Insitutue for Multidisciplinary Sciences, Getynga, Niemcy, od 2016 roku. Praca związana z oszacowaniem kinetyki i termodynamiki odłączania inhibitora

od enzymów [A3].

7. Prof. Yasuteru Shigeta i prof. Ryuhei Harada, Uniwersytet Tsukuba, Tsukuba, Japonia, od 2019 r. Prace związane z wdrażaniem metod wzmocnionego próbkowania we wtyczce PLUMED [H7].

8. Prof. Bruno Lapied, Uniwersytet w Angers, Angers, Francja, od 2023 r. Praca związana z obliczeniowym przezwyciężaniem oporności na pyretroidy w zwalczaniu komarów [A6].

II.12. Informacja o uczestnictwie w zespołach oceniających wnioski o finansowanie badań, wnioski o przyznanie nagród naukowych, wnioski w innych konkursach mających charakter naukowy lub dydaktyczny

1. Rozpatrzenie wniosków w konkursie na najlepszą pracę licencjacką z bioinformatyki finansowanym przez Polskie Towarzystwo Bioinformatyczne (2022).

2. Recenzowanie (jako kierownik projektu) wniosków o finansowanie opłat za publikowanie w otwartym dostępie dla młodych naukowców w Uniwersytetu Mikołaja Kopernika (2019-2022). Finansowanie zapewnione przez Narodową Agencję Wymiany Akademickiej w ramach projektu "Open NCU–Open Source, Open Science".

3. Komisja rekrutacyjna ds. wyboru kandydatów do szkoły doktorskiej Uniwersytetu Mikołaja Kopernika w ramach projektów badawczych przyznanych przez Narodowe Centrum Nauki oraz Ministerstwo Nauki i Szkolnictwa Wyższego.

II.13. Inne

Nagrody i stypendia:

1. Stypendium stażowe przyznane przez Japan Society for the Promotion of Science (JSPS), 2022.

2. Stypendium za wybitne osiągnięcia dla młodych naukowców finansowane przez Ministerstwo Nauki i Szkolnictwa Wyższego, 2022.

3. Stypendium finansowane przez Fundację na rzecz Nauki Polskiej (START FNP) za wybitne osiągnięcia dla naukowców, którzy nie ukończyli 30. roku życia, 2018.

4. Stypendium doktoranckie za wybitne osiągnięcia finansowane przez Ministerstwo Nauki i Szkolnictwa Wyższego, 2017.

5. Stypendium doktoranckie za wybitne osiągnięcia finansowane przez Ministerstwo Nauki i Szkolnictwa Wyższego, 2016.

6. Stypendium Marszałka Województwa Kujawsko-Pomorskiego ("Krok w przyszłość"), 2015.

7. I miejsce w konkursie na najlepszą pracę magisterską z bioinformatyki i biofizyki obliczeniowej ufundowane przez Polskie Towarzystwo Bioinformatyczne (PTBI), 2015.
8. Wybrany najlepszym studentem Wydziału Fizyki, Astronomii i Informatyki UMK w 2015 roku.
9. Wiele nagród i stypendiów Rektora Uniwersytetu Mikołaja Kopernika.

III. Informacja o współpracy z otoczeniem społecznym i gospodarczym

III.1. Informacja o współpracy z sektorem gospodarczym

1. Badacz w projekcie Narodowego Centrum Badań i Rozwoju LIDER/28/54/L-/10/NCBiR/2011, Poszukiwanie nowego leku przeciwnowotworowego opartego na inhibitorach PCNA; PI: Prof. Wojciech Strzałka.

List of Scientific Achievements which Present a Major Contribution to the Development of a Specific Discipline

(As of May 21, 2024)

Name of the candidate: Jakub Rydzewski

Scientometric information:

- ResearcherID: N-9160-2019
- ORCID: 0000-0003-4325-4177
- Google Scholar: dEMXOpcAAAAJ
- Hirsch Index: 9 [Web of Science (WoS)] 10 [Google Scholar (GS)]
- 10-Index: 13 [GS]
- Number of publications: 25

Total number of citations: 780 [WoS] 981 [GS]
Total number of citations excluding self-citations: 682 [WoS]
Total number of points by the Ministry, including:

- Publications after 2018 (scale 0-200 points per publication): 1860
- Publications before 2018 (scale 0-50 points per publication): 370

The full list of the Author's articles in scientific journals that contribute to the main habilitation Achievement includes the articles [H1-H7] listed in section I.1. Apart from these publications, since obtaining PhD, the Author has also obtained other scientific achievements, consisting of publications [A1-A6] given in section II.1.

I. Information on scientific achievements set out in art. 219 para 1. point 2 of the Act

Title of the Achievement:

Learning Collective Variables from Atomistic Simulations

I.1. Cycle of scientific articles related thematically, pursuant to art. 219 para 1. point 2b of the Act

The numbers [H1-H7] assigned below are also used in the Summary of Professional Accomplishments. Information about the Impact Factor was taken from the Web of Science (WoS) database for the 2022 year. The asterisk next to the name denotes the publications in which the candidate acts as a corresponding author.

The Author contribution is assigned similarily to Contributor Roles Taxonomy (CRediT) Classification (https://credit.niso.org/) used by the American Institute of Physics.

H1. ***Rydzewski, J.**, Chen, M. & Valsson, O. Manifold Learning in Atomistic Simulations: A Conceptual Review. *Mach. Learn. Sci. Technol.* **4**, 031001 (2023).
Journal impact factor: 6.8
Number of Ministry points: 20
Number of citations: 4 [WoS] 10 [GS]

Analyzing large volumes of high-dimensional data requires dimensionality reduction: finding meaningful low-dimensional structures hidden in their high-dimensional observations. Such practice is needed in atomistic simulations of dynamical systems where even thousands of degrees of freedom are sampled. An abundance of such data makes it strenuous to gain insight into a specific physical problem. Our primary aim in this review is to focus on unsupervised machine learning methods that can be used on simulation data to find a low-dimensional manifold providing a collective and informative characterization of the studied process. Such manifolds can be used for sampling long-timescale processes and free-energy estimation. We describe methods that can work on data sets from standard and enhanced sampling atomistic simulations. Compared to recent reviews on manifold learning for atomistic simulations, we consider only methods that construct low-dimensional manifolds based on Markov transition probabilities between high-dimensional samples. We discuss these techniques from a conceptual point of view, including their underlying theoretical frameworks and possible limitations.

Contribution:

- Conceptualization (leading): Devised theory, methods, examples, and concept for general framework of manifold learning in standard and enhanced atomistic simulations; discussed concept with co-authors and addressed their suggestions.
- Resources (leading): Acquired funding for open access.
- Supervision (leading): Acted as corresponding author and organized meetings and discus-

sions.

- Visualization (leading): Selected, prepared, and visualized all results for publication.
- Writing/Original Draft Preparation (eqal): Wrote original draft.
- Writing/Review & Editing (equal): Revised original draft with co-authors.

H2. ***Rydzewski, J.** & Valsson, O. Multiscale Reweighted Stochastic Embedding: Deep Learning of Collective Variables for Enhanced Sampling. *J. Phys. Chem. A* **125**, 6286–6302 (2021).
Journal impact factor: 2.9
Number of Ministry points: 100
Number of citations: 19 [WoS] 24 [GS]

Machine learning methods provide a general framework for automatically finding and representing the essential characteristics of simulation data. This task is particularly crucial in enhanced sampling simulations. There we seek a few generalized degrees of freedom, referred to as collective variables (CVs), to represent and drive the sampling of the free energy landscape. In theory, these CVs should separate different metastable states and correspond to the slow degrees of freedom of the studied physical process. To this aim, we propose a new method that we call multiscale reweighted stochastic embedding (MRSE). Our work builds upon a parametric version of stochastic neighbor embedding. The technique automatically learns CVs that map a high-dimensional feature space to a low-dimensional latent space via a deep neural network. We introduce several new advancements to stochastic neighbor embedding methods that make MRSE especially suitable for enhanced sampling simulations: (1) weight-tempered random sampling as a landmark selection scheme to obtain training data sets that strike a balance between equilibrium representation and capturing important metastable states lying higher in free energy; (2) a multiscale representation of the high-dimensional feature space via a Gaussian mixture probability model; and (3) a reweighting procedure to account for training data from a biased probability distribution. We show that MRSE constructs low-dimensional CVs that can correctly characterize the different metastable states in three model systems: the Müller-Brown potential, alanine dipeptide, and alanine tetrapeptide.

Contribution:

- Conceptualization (equal): Together with co-author devised theory, idea, methods, examples, and concept for using a parametric embedding via a neural network to learn CVs.
- Data Curation (leading): Prepared and performed all enhanced sampling simulations and applied developed methods to learn CVs from those simulations.

- Formal Analysis (leading): Analyzed all results.
- Methodology (equal): Together with co-author derived and numerically validated all equations.
- Resources (leading): Provided computing time and acquired funding for open access.
- Software (leading): Implemetated codes for learning CVs into the PLUMED plugin (the lowlearner module, freely accessible) and additional numerical methods for analysis.
- Supervision (leading): Acted as corresponding author.
- Validation (equal): Together with co-author validated all results.
- Visualization (leading): Selected, prepared, and visualized all results for publication.
- Writing/Original Draft Preparation (leading): Wrote original draft.
- Writing/Review & Editing (equal): Revised original draft with co-author.

H3. ***Rydzewski, J.**, Chen, M., Ghosh, T.K. & Valsson, O. Reweighted Manifold Learning of Collective Variables from Enhanced Sampling Simulations. *J. Chem. Theory Comput.* **18**, 7179–7192 (2022).
Journal impact factor: 5.5
Number of Ministry points: 140
Number of citations: 8 [WoS] 10 [GS]

Enhanced sampling methods are indispensable in computational chemistry and physics, where atomistic simulations cannot exhaustively sample the high-dimensional configuration space of dynamical systems due to the sampling problem. A class of such enhanced sampling methods works by identifying a few slow degrees of freedom, termed collective variables (CVs), and enhancing the sampling along these CVs. Selecting CVs to analyze and drive the sampling is not trivial and often relies on chemical intuition. Despite routinely circumventing this issue using manifold learning to estimate CVs directly from standard simulations, such methods cannot provide mappings to a low-dimensional manifold from enhanced sampling simulations, as the geometry and density of the learned manifold are biased. Here, we address this crucial issue and provide a general reweighting framework based on anisotropic diffusion maps for manifold learning that takes into account that the learning data set is sampled from a biased probability distribution. We consider manifold learning methods based on constructing a Markov chain describing transition probabilities between high-dimensional samples. We show that our framework reverts the biasing effect, yielding CVs that correctly describe the equilibrium density. This advancement enables

the construction of low-dimensional CVs using manifold learning directly from the data generated by enhanced sampling simulations. We call our framework reweighted manifold learning. We show that it can be used in many manifold learning techniques on data from both standard and enhanced sampling simulations.

Contribution:

- Conceptualization (leading): Devised theory, methods, framework, examples, and concept for reweighting and manifold learning CVs from enhanced sampling simulations; discussed concept with co-authors and addressed their suggestions.
- Data Curation (leading): Prepared and performed all enhanced sampling simulations and applied the framework to learn CVs from those simulations.
- Formal Analysis (leading): Analyzed all results.
- Methodology (equal): Derived and numerically validated all equations with the help of coauthors.
- Resources (leading): Provided computing time and acquired funding for open access.
- Software (leading): Implemetated codes for learning CVs into the PLUMED plugin (the lowlearner module, freely accessible) and additional numerical methods for analysis.
- Supervision (leading): Acted as corresponding author and organized meetings and discussions.
- Validation (equal): Validated all results with the help of coauthors.
- Visualization (leading): Selected, prepared, and visualized all results for publication.
- Writing/Original Draft Preparation (leading): Wrote original draft.
- Writing/Review & Editing (equal): Revised original draft with co-authors.

H4. ***Rydzewski, J.** Selecting High-Dimensional Representations of Physical Systems by Reweighted Diffusion Maps. *J. Phys. Chem. Lett.* **14**, 2778—2783 (2023).
Journal impact factor: 5.7
Number of Ministry points: 200
Number of citations: 4 [WoS] 5 [GS]

Constructing reduced representations of high-dimensional systems is a fundamental problem in physical chemistry. Many unsupervised machine learning methods can automatically find such low-dimensional representations. However, an often overlooked problem is what high-dimensional

representation should be used to describe systems before dimensionality reduction. Here, we address this issue using a recently developed method called the reweighted diffusion map [J. Chem. Theory Comput. 2022, 18, 7179–7192]. We show how high-dimensional representations can be quantitatively selected by exploring the spectral decomposition of Markov transition matrices built from data obtained from standard or enhanced sampling atomistic simulations. We demonstrate the performance of the method in several high-dimensional examples.

Contribution: Conceptualization; Data Curation; Formal Analysis; Methodology; Project Administration; Resources; Software; Supervision; Validation; Visualization; Writing/Original Draft Preparation; Writing/Review & Editing were done solely by the Author.

H5. ***Rydzewski, J.** Spectral Map: Embedding Slow Kinetics in Collective Variables. *J. Phys. Chem. Lett.* **14**, 2778—2783 (2023).
Journal impact factor: 5.7
Number of Ministry points: 200
Number of citations: 3 [WoS] 5 [GS]

The dynamics of physical systems that require high-dimensional representation can often be captured in a few meaningful degrees of freedom called collective variables (CVs). However, identifying CVs is challenging and constitutes a fundamental problem in physical chemistry. This problem is even more pronounced when CVs need to provide information about slow kinetics related to rare transitions between long-lived metastable states. To address this issue, we propose an unsupervised deep-learning method called spectral map. Our method constructs slow CVs by maximizing the spectral gap between slow and fast eigenvalues of a transition matrix estimated by an anisotropic diffusion kernel. We demonstrate our method in several high-dimensional reversible folding processes.

Contribution: Conceptualization; Data Curation; Formal Analysis; Methodology; Project Administration; Resources; Software; Supervision; Validation; Visualization; Writing/Original Draft Preparation; Writing/Review & Editing were done solely by the Author.

H6. ***Rydzewski, J.** & Gokdemir, T. Learning Markovian Dynamics with Spectral Maps. *J. Chem. Phys.* **160**, 091102 (2024).
Journal impact factor: 4.4
Number of Ministry points: 100
Number of citations: 0 [WoS] 0 [GS]

The long-time behavior of many complex molecular systems can often be described by Markovian dynamics in a slow subspace spanned by a few reaction coordinates referred to as collective

variables (CVs). However, determining CVs poses a fundamental challenge in chemical physics. Depending on intuition or trial and error to construct CVs can lead to non-Markovian dynamics with long memory effects, hindering analysis. To address this problem, we continue to develop a recently introduced deep-learning technique called spectral map [J. Rydzewski, J. Phys. Chem. Lett. 14, 5216–5220 (2023)]. Spectral map learns slow CVs by maximizing a spectral gap of a Markov transition matrix describing anisotropic diffusion. Here, to represent heterogeneous and multiscale free-energy landscapes with spectral map, we implement an adaptive algorithm to estimate transition probabilities. Through a Markov state model analysis, we validate that spectral map learns slow CVs related to the dominant relaxation timescales and discerns between long-lived metastable states.

Contribution:

- Conceptualization (leading): Devised theory, idea, methods, examples, and concept.
- Data Curation (equal): Applied developed methods to learn CVs from simulations with co-author.
- Formal Analysis (leading): Analyzed all results.
- Methodology (leading): Derived and numerically validated all equations.
- Resources (leading): Provided computing time.
- Software (leading): Implemetated codes for learning CVs and additional numerical methods for analysis.
- Supervision (leading): Acted as corresponding author.
- Validation (equal): Together with co-author validated all results.
- Visualization (leading): Selected, prepared, and visualized all results for publication.
- Writing/Original Draft Preparation (leading): Wrote original draft.
- Writing/Review & Editing (equal): Revised original draft with co-author.

H7. **Rydzewski, J.** as part of PLUMED Consortium[1]. Promoting Transparency and Reproducibility in Enhanced Molecular Simulations. *Nat. Methods* **16**, 670-–673 (2019).
Journal impact factor: 48
Number of Ministry points: 200
Number of citations: 543 [WoS] 489 [GS]

[1]Full list of contributors: plumed-nest.org/consortium.html

The PLUMED consortium unifies developers and contributors to PLUMED, an open-source library for enhanced-sampling, free-energy calculations and the analysis of molecular dynamics simulations. Here, we outline our efforts to promote transparency and reproducibility by disseminating protocols for enhanced-sampling molecular simulations. The PLUMED consortium provides an automatic verification algorithm to check input for atomistic simulations, an interface of freely open repository for simulation data, and tools that enable to reproduce results from publications.

Contribution:

- Conceptualization (supporting): Participated and contributed in every discussion and online meetings.
- Data Curation (supporting): Implemented regression tests for verification of results obtained using the maze module for enhanced sampling of ligand dissociation pathways and provided documentation (https://www.plumed.org/doc-v2.8/user-doc/html/_m_a_z_e.html).
- Methodology (supporting): Discussed framework for reproducibility and contributed suggestions.
- Software (equal): Implemented the maze module in PLUMED for enhanced sampling of ligand dissociation pathways – one of the most advanced PLUMED modules, containing several different algorithms. The module is freely available within every PLUMED release since 2.7 (https://github.com/plumed/plumed2/tree/master/src/maze).
- Writing/Original Draft Preparation (supporting): Wrote about the maze module.
- Writing/Review & Editing (equal): Edited text together with co-authors.

II. Information on scientific or artistic activity

II.1. List of articles published in scientific journals, not mentioned in section I.1

Information on the Impact Factor was taken from the Web of Science (WoS) database from the 2022 year. The asterisk next to the name denotes the publications in which the candidate is a corresponding author.

The Author contribution is assigned according to Contributor Roles Taxonomy (CRediT) Classification (https://credit.niso.org/) used by the American Institute of Physics.

A1. ***Rydzewski, J.*** maze: Heterogeneous Ligand Unbinding along Transient Protein Tunnels. *Comput. Phys. Commun.* **247**, 106865 (2020).

Journal impact factor: 6.3
Number of Ministry points: 140
Number of citations: 9 [WoS] 7 [GS]

Recent developments in enhanced sampling methods showed that it is possible to reconstruct ligand unbinding pathways with spatial and temporal resolution inaccessible to experiments. Ideally, such techniques should provide an atomistic definition of possibly many reaction pathways, because crude estimates may lead either to overestimating energy barriers, or inability to sample hidden energy barriers that are not captured by reaction pathway estimates. Here we provide an implementation of a new method [Rydzewski and Valsson, J. Chem. Phys. 150, 221101 (2019)] dedicated entirely to sampling the reaction pathways of the ligand–protein dissociation process. The program, called maze, is implemented as an official module for PLUMED 2, an open source library for enhanced sampling in molecular systems, and comprises algorithms to find multiple heterogeneous reaction pathways of ligand unbinding from proteins during atomistic simulations. The maze module requires only a crystallographic structure to start a simulation, and does not depend on many ad hoc parameters. The program is based on enhanced sampling and non-convex optimization methods. To present its applicability and flexibility, we provide several examples of ligand unbinding pathways along transient protein tunnels reconstructed by maze in a model ligand–protein system, and discuss the details of the implementation.

Contribution: Conceptualization; Data Curation; Formal Analysis; Methodology; Project Administration; Resources; Software; Supervision; Validation; Visualization; Writing/Original Draft Preparation; Writing/Review & Editing were done solely by the Author.

A2. ***Rydzewski, J.** & Valsson, O. Finding multiple reaction pathways of ligand unbinding. *J. Chem. Phys.* **150** (2019).
Journal impact factor: 4.4
Number of Ministry points: 100
Number of citations: 21 [WoS] 27 [GS]

Searching for reaction pathways describing rare events in large systems presents a long-standing challenge in chemistry and physics. Incorrectly computed reaction pathways result in the degeneracy of microscopic configurations and inability to sample hidden energy barriers. To this aim, we present a general enhanced sampling method to find multiple diverse reaction pathways of ligand unbinding through nonconvex optimization of a loss function describing ligand-protein interactions. The method successfully overcomes large energy barriers using an adaptive bias potential and constructs possible reaction pathways along transient tunnels without the initial guesses of intermediate or final states, requiring crystallographic information only. We examine

the method on the T4 lysozyme L99A mutant which is often used as a model system to study ligand binding to proteins, provide a previously unknown reaction pathway, and show that by using the bias potential and the tunnel widths, it is possible to capture heterogeneity of the unbinding mechanisms between the found transient protein tunnels.

Contribution:

- Conceptualization (leading): Devised theory, methods, framework, examples, and concept for sampling CVs for ligand dissociation events from enhanced sampling simulations; discussed concept with co-author and addressed their suggestions.
- Data Curation (leading): Prepared and performed all enhanced sampling simulations.
- Formal Analysis (leading): Analyzed all results.
- Resources (leading): Provided computing time.
- Software (leading): Implemetated codes for sampling CVs describing ligand dissociation into the PLUMED plugin and additional numerical methods for analysis.
- Supervision (leading): Acted as corresponding author and organized meetings and discussions.
- Validation (equal): Validated all results together with co-author.
- Visualization (leading): Selected, prepared, and visualized all results for publication.
- Writing/Original Draft Preparation (leading): Wrote original draft.
- Writing/Review & Editing (leading): Revised original draft according to co-author's suggestions.

A3. ***Rydzewski, J.**, Jakubowski, R., Nowak, W. & Grubmuller, H. Kinetics of Huperzine A Dissociation from Acetylcholinesterase via Multiple Unbinding Pathways. *J. Chem. Theory Comput.* **14**, 2843–2851 (2018).
Journal impact factor: 5.5
Number of Ministry points: 140
Number of citations: 21 [WoS] 31 [GS]

The dissociation of huperzine A (hupA) from Torpedo californica acetylcholinesterase (TcAChE) was investigated by 4 μs unbiased and biased all-atom molecular dynamics (MD) simulations in explicit solvent. We performed our study using memetic sampling (MS) for the determination of reaction pathways (RPs), metadynamics to calculate free energy, and maximum-likelihood estimation (MLE) to recover kinetic rates from unbiased MD simulations. Our simulations

suggest that the dissociation of hupA occurs mainly via two RPs: a front door along the axis of the active-site gorge (pwf) and through a new transient side door (pws), i.e., formed by the Ω-loop (residues 67–94 of TcAChE). An analysis of the inhibitor unbinding along the RPs suggests that pws is opened transiently after hupA and the Ω-loop reach a low free-energy transition state characterized by the orientation of the pyridone group of the inhibitor directed toward the Ω-loop plane. Unlike pws, pwf does not require large structural changes in TcAChE to be accessible. The estimated free energies and rates agree well with available experimental data. The dissociation rates along the unbinding pathways are similar, suggesting that the dissociation of hupA along pws is likely to be relevant. This indicates that perturbations to hupA-TcAChE interactions could potentially induce pathway hopping. In summary, our results characterize the slow-onset inhibition of TcAChE by hupA, which may provide the structural and energetic bases for the rational design of the next-generation slow-onset inhibitors with optimized pharmacokinetic properties for the treatment of Alzheimer's disease.

Contribution:

- Conceptualization (leading): Devised theory, methods, framework, examples, and concept for sampling multiple dissociation events and calculating free-energy profiles and kinetics from enhanced sampling simulations; discussed concept with co-authors and addressed their suggestions.
- Data Curation (leading): Prepared and performed all enhanced sampling simulations and analyses.
- Formal Analysis (leading): Analyzed all results.
- Methodology (leading): Derived and numerically validated all equations.
- Resources (leading): Provided computing time.
- Software (leading): Implemetated codes for calculating multiple dissociation pathways, kinetic rates and additional numerical methods for analysis.
- Supervision (equal): Acted as corresponding author and organized meetings and discussions.
- Validation (leading): Validated all results.
- Visualization (leading): Selected, prepared, and visualized all results for publication.
- Writing/Original Draft Preparation (leading): Wrote original draft.
- Writing/Review & Editing (equal): Revised original draft with co-authors.

A4. ***Rydzewski, J.**, Walczewska-Szewc, K., Czach, S., Nowak, W. & Kuczera, K. Enhancing the Inhomogeneous Photodynamics of Canonical Bacteriophytochrome. *J. Phys. Chem. B* **126**, 2647—2657 (2022).
Journal impact factor: 3.3
Number of Ministry points: 140
Number of citations: 2 [WoS] 5 [GS]

The ability of phytochromes to act as photoswitches in plants and microorganisms depends on interactions between a bilin-like chromophore and a host protein. The interconversion occurs between the spectrally distinct red (Pr) and far-red (Pfr) conformers. This conformational change is triggered by the photoisomerization of the chromophore D-ring pyrrole. In this study, as a representative example of a phytochrome-bilin system, we consider biliverdin IXα (BV) bound to bacteriophytochrome (BphP) from Deinococcus radiodurans. In the absence of light, we use an enhanced sampling molecular dynamics (MD) method to overcome the photoisomerization energy barrier. We find that the calculated free energy (FE) barriers between essential metastable states agree with spectroscopic results. We show that the enhanced dynamics of the BV chromophore in BphP contributes to triggering nanometer-scale conformational movements that propagate by two experimentally determined signal transduction pathways. Most importantly, we describe how the metastable states enable a thermal transition known as the dark reversion between Pfr and Pr, through a previously unknown intermediate state of Pfr. We present the heterogeneity of temperature-dependent Pfr states at the atomistic level. This work paves a way toward understanding the complete mechanism of the photoisomerization of a bilin-like chromophore in phytochromes.

Contribution:

- Conceptualization (leading): Devised concept for enhancing fluctuations of BV; discussed concept with co-authors and addressed their suggestions.
- Data Curation (leading): Prepared and performed all enhanced sampling simulations.
- Formal Analysis (leading): Analyzed all results.
- Methodology (leading): Selected all methods.
- Resources (leading): Provided computing time and acquired funding for open access.
- Software (leading): Implemetated numerical methods for analysis.
- Supervision (equal): Acted as corresponding author and organized meetings and discussions.
- Validation (equal): Validated all results with co-authors.

- Visualization (leading): Selected, prepared, and visualized all results for publication.
- Writing/Original Draft Preparation (leading): Wrote original draft.
- Writing/Review & Editing (equal): Revised original draft with co-authors.

A5. Walczewska-Szewc, K., **Rydzewski, J.** & Lewkowicz, A. Inhibition-Mediated Changes in Prolyl Oligopeptidase Dynamics Possibly Related to α-Synuclein Aggregation. *Phys. Chem. Chem. Phys.* **24**, 4366—4373 (2022).
Journal impact factor: 3.3
Number of Ministry points: 100
Number of citations: 2 [WoS] 4 [GS]

The formation of protein aggregates is one of the leading causes of neuronal malfunction and subsequent brain damage in many neurodegenerative diseases. In Parkinson's disease, α-synucleins are involved in the accumulation of aggregates. The origin of aggregation is unknown, but there is convincing evidence that it can be reduced by prolyl oligopeptidase (PREP) inhibition. This effect cannot simply be related to the inhibition of the enzyme's catalytic function since not all PREP inhibitors stop α-synuclein aggregation. Finding differences in the dynamics of the enzyme inhibited by different compounds would allow us to identify the protein regions involved in the interaction between PREP and α-synuclein. Here, we investigate the effects of three PREP inhibitors, each of which affects α-synuclein aggregation to a different extent. We use molecular dynamics modelling to identify the molecular mechanisms underlying PREP inhibition and find structural differences between inhibitor-PREP systems. We suggest that even subtle variations in enzyme dynamics affect its interactions with α-synucleins. Our identification of these regions may therefore be biologically relevant in preventing α-synuclein aggregate formation.

Contribution:

- Conceptualization (supporting): Planned research related to enhanced sampling of ligand dissociation events.
- Data Curation (equal): Prepared and performed enhanced sampling simulations for unbinding several ligands.
- Formal Analysis (equal): Analyzed results related to enhanced sampling of ligand dissociation events and free-energy profiles.
- Methodology (equal): Numerically validated results related to enhanced sampling of ligand dissociation events.
- Software (supporting): Implemetated additional numerical methods for analysis.

- Validation (equal): Validated all results related to enhanced sampling of ligand dissociation events.
- Visualization (supporting): Prepared, and visualized all results related to enhanced sampling of ligand dissociation events.
- Writing/Original Draft Preparation (supporting): Wrote original draft parts related to enhanced sampling of ligand dissociation events.
- Writing/Review & Editing (equal): Revised original draft with co-authors.

A6. Niklas, B., **Rydzewski, J.**, Lapied, B. & Nowak, W. Toward Overcoming Pyrethroid Resistance in Mosquito Control: The Role of Sodium Channel Blocker Insecticides. *Int. J. Mol. Sci.* **24**, 10334 (2023).
Journal impact factor: 5.6
Number of Ministry points: 140
Number of citations: 0 [WoS] 0 [GS]

Diseases spread by mosquitoes lead to the death of 700,000 people each year. The main way to reduce transmission is vector control by biting prevention with chemicals. However, the most commonly used insecticides lose efficacy due to the growing resistance. Voltage-gated sodium channels (VGSCs), membrane proteins responsible for the depolarizing phase of an action potential, are targeted by a broad range of neurotoxins, including pyrethroids and sodium channel blocker insecticides (SCBIs). Reduced sensitivity of the target protein due to the point mutations threatened malaria control with pyrethroids. Although SCBIs—indoxacarb (a pre-insecticide bioactivated to DCJW in insects) and metaflumizone—are used in agriculture only, they emerge as promising candidates in mosquito control. Therefore, a thorough understanding of molecular mechanisms of SCBIs action is urgently needed to break the resistance and stop disease transmission. In this study, by performing an extensive combination of equilibrium and enhanced sampling molecular dynamics simulations, we found the DIII-DIV fenestration to be the most probable entry route of DCJW to the central cavity of mosquito VGSC. Our study revealed that F1852 is crucial in limiting SCBI access to their binding site. Our results explain the role of the F1852T mutation found in resistant insects and the increased toxicity of DCJW compared to its bulkier parent compound, indoxacarb. We also delineated residues that contribute to both SCBIs and non-ester pyrethroid etofenprox binding and thus could be involved in the target site cross-resistance.

Contribution:

- Conceptualization (supporting): Planned research related to enhanced sampling of ligand

dissociation events.

- Data Curation (equal): Prepared and performed enhanced sampling simulations for unbinding several ligands.
- Formal Analysis (equal): Analyzed results related to enhanced sampling of ligand dissociation events and free-energy profiles.
- Methodology (equal): Numerically validated results related to enhanced sampling of ligand dissociation events.
- Software (supporting): Implemetated additional numerical methods for analysis.
- Validation (equal): Validated all results related to enhanced sampling of ligand dissociation events.
- Visualization (supporting): Prepared, and visualized free-energy profiles of ligand dissociation.
- Writing/Original Draft Preparation (supporting): Wrote original draft parts related to enhanced sampling of ligand dissociation events.
- Writing/Review & Editing (equal): Revised original draft with co-authors.

Publications in peer-reviewed journals before obtaining the doctoral degree: 12 publications [B1-B12], mainly focused on developing enhanced sampling techniques to observe ligand dissociation pathways from proteins:

B1. Carrascoza Mayén, J. F., **Rydzewski, J.**, Szostak, N., *Blazewicz, J., and *Nowak, W. Prebiotic Soup Components Trapped in Montmorillonite Nanoclay Form New Molecules: Car-Parrinello Ab Initio Simulations. *Life* **9**, 46 (2019).

B2. ***Rydzewski, J.**, and Nowak W. Photoinduced Transport in an H64Q Neuroglobin Antidote for Carbon Monoxide Poisoning. *J. Chem. Phys.* **148**, 115101 (2018).

B3. ***Rydzewski, J.**, Jakubowski, R., Nicosia G., and Nowak, W. Conformational Sampling of a Biomolecular Rugged Energy Landscape. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* **15**, 732 (2018).

B4. ***Rydzewski, J**. and Nowak, W. Rare-Event Sampling in Ligand Diffusion. *Phys. Life Rev.* **22-23**, 85 (2017).

B5. ***Rydzewski, J**. and Nowak, W. Ligand Diffusion in Proteins via Enhanced Sampling in Molecular Dynamics. *Phys. Life Rev.* **22-23**, 58 (2017).

B6. ***Rydzewski, J**. and Nowak, W. Thermodynamics of Camphor Migration in Cytochrome P450cam by Atomistic Simulations. *Sci. Rep.* **7**, 7736 (2017).

B7. ***Rydzewski, J**. and Nowak, W. Machine Learning Based Dimensionality Reduction Facilitates Ligand Diffusion Paths Assessment: A Case of Cytochrome P450cam. *J. Chem. Theory Comput.* **12**, 2110 (2016).

B8. ***Rydzewski, J**. and Nowak, W. Molecular Dynamics Simulations of Large Systems in Electronic Excited States. *Handbook of Computational Chemistry* (Vol. I), Second Edition (2016) (Editior: J. Leszczynski).

B9. ***Rydzewski, J.**, Nowak, W., Nicosia G. Inferring Pathological States in Cortical Neuron Microcircuits. *J. Theor. Biol.* **386**, 34 (2015).

B10. **Rydzewski, J.**, Jakubowski, R., and *Nowak, W. Entropic Measure to Prevent Energy Over-Minimization in Molecular Dynamics Simulations, *J. Chem. Phys.* **143**, 171103 (2015).

B11. **Rydzewski, J**. and *Nowak, W. Memetic Algorithms for Ligand Expulsion from Protein Cavities. *J. Chem. Phys.* **143**, 124101 (2015).

B12. **Rydzewski, J.**, Strzałka, W., and *Nowak, W. Nanomechanics of PCNA: A Protein-Made DNA Sliding Clamp. *Chem. Phys. Lett.* **634**, 236 (2015).

II.3. Popular and promotional actions

- Inaugural lecture for first-year students of the Faculty of Physics, Astronomy and Informatics at Nicolaus Copernicus University, Torun, October 2018.
- Enhanced Sampling Methods (in Polish) published in the Voice of the University 7-10, 2019, distributed by Nicolaus Copernicus University, Torun.
- Popularization note about promoting grants offered by the National Science Center, 10-th aniversary of the National Science Center, 2021.
- Summer camp for international students and Ph.D. students in the program of the Polish National Agency for Academic Exchange (NAWA) SPINAKER — Intensive International Education Programs, 12-14th July 2022 in Torun, Poland.
- Promoting open science as the PI of the Polish National Agency for Academic Exchange's project "Open NCU–Open Source, Open Science" by funding open access publishing fees for young researchers at Nicolaus Copernicus University, Torun.

- Teaching classes about open science in the program "Developing and Implementing hands-on training on Open Science and Open Innovation for Early Career Researchers (DIOSI)" funded by the European Union under the Horizon 2020, European Commission 101006318, 2020–2023.

II.4. Presentations given at national or international scientific conferences

The list presented below includes only conference/scientific contributions, e.g., invited lectures, seminars, and oral and poster presentations by the Author *after obtaining PhD.*

Oral presentations:

1. Spectral Map: Embedding Slow Kinetics in Collective Variables, *Molecular Simulation Society of Japan 2023*, 2023-12-4 – 2023-12-6, Fukui, Japan (Invited).
2. Recent Advances in Reweighted Manifold Learning for Molecular Dynamics, *Rare Event Workshop 2023*, 2023-12-3, Kanazawa, Japan (Invited).
3. Reweighted Manifold Learning for Simulations and Enhanced Sampling, *34th IUPAP Conference on Computational Physics (CCP2023)*, 2023-08-04 – 2023-08-08, Kobe, Japan.
4. Multiscale Reweighted Stochastic Embedding: Deep Learning of Collective Variables for Enhanced Sampling, *Recent Advances in Modelling Rare Events*, 2021-12-15 – 2021-12-18, Virtual, India.
5. Multiscale Reweighted Stochastic Embedding (MRSE): Deep Learning of Generalized Variables for Statistical Physics, *ML in PL (Machine Learning in Poland)*, 2021-11-05 – 2021-11-07, Warsaw, Poland.
6. Multiscale Reweighted Stochastic Embedding: Deep Learning of Collective Variables for Enhanced Sampling, *Winter Workshop on Multiscale Modeling, Karlsruhe Institute of Technology (KIT)*, 2021-11-22 – 2021-11-23, Germany.
7. Multiscale Reweighted Stochastic Embedding: Deep Learning of Collective Variables for Enhanced Sampling, *American Physical Society March Meeting 2021*, 2021-03-15 – 2021-03-19, Virtual, US.
8. Multiscale Reweighted Stochastic Embedding: Deep Learning of Collective Variables for Enhanced Sampling, *Bunsen-Tagung 2021*, 2021-05-10 – 2021-05-12, Virtual, Germany.
9. Sampling Ligand Unbinding Pathways, *Bioinformatics in Torun (BIT)*, 2019-06-21 – 2019-06-25, Torun, Poland.

Poster presentations:

1. Enhanced Sampling Methods for Ligand Unbinding, *Mainz Materials Simulation Days 2019 (CECAM Event)*, 2019-06-05 – 2019-06-07, Mainz, Germany.

Invited seminars:

1. Manifold Learning for Atomistic Simulations, *The National Institute of Advanced Industrial Science and Technology (AIST)*, 2023-11-23, Tsukuba, Japan.
2. Heterogeneous Ligand Unbinding from Proteins, *Institute of Computational Sciences, University of Tsukuba*, 2019, Tsukuba, Japan.
3. Multiscale Reweighted Stochastic Embedding: Deep Learning of Collective Variables for Enhanced Sampling, *Soft Matter Seminar*, 2019, Amsterdam, Netherlands.

II.5. Participation in organizational and scientific committees at conferences

- Organization of the conference Bioinformatics in Torun held jointly by the Polish Bioinformatics Society (PTBI) and the Nicolaus Copernicus University in Torun, 2014–2023.
- Reviewer for the Polish Bioinformatics Society competition for the best bachelor thesis in bioinformatics and computational biology, 2022.
- Reviews for the proceedings of Annual Conference on Machine Learning, Optimization and Data Science (LOD), 2018–2020.

II.6. Participation in the works of research teams realizing projects financed through national and international competitions

Principal investigator:

1. Sonata 2021/43/D/ST4/00920, National Science Center in Poland, *Statistical Learning of Slow Collective Variables from Atomistic Simulations*, Since 2022.
2. Emerging Research Fields under the "Initiative of Excellence — Research University," *Nanoscale Biophysics*, Co-PI, Since 2022.
3. Promocja Zagraniczna PPI/PZA/2019/1/00037, National Center of Academic Exchange, *Open Source–Open Science*, 2019—2021.
4. Etiuda 2016/20/T/ST3/00488, National Science Center in Poland, *Reconstruction of Ligand Unbiding Pathways from Proteins*, 2016–2017.
5. Preludium 2015/19/N/ST3/02171, National Science Center in Poland *Memetic Algorithms for Ligand Expulsion in Electronic Excited State from Condensed Biological Matter*, 2016–2019.

Investigator:

1. NIH Grant P30GM110761, National Institute of Health in US, 2018–2022, PI: Prof. K. Kuczera (University of Kansas).
2. Opus 2016/23/B/ST4/01770, National Science Center in Poland, *Structural Determinants of Optical Control of Insulin and Neuroligin Release by Photoactive Protein Ligands*, 2017–2022, PI: Prof. W. Nowak (Nicolaus Copernicus University).
3. Grant N202 262038 by the Ministry of Science and Higher Education in Poland, *Nanomechanics of Modular Adhesive Proteins*, 2010–2014, PI: Prof. W. Nowak (Nicolaus Copernicus University).
4. LIDER/28/54/L-/10/NCBiR/2011, The National Centre for Research and Development, *Searching for a Novel Anticancer Drug based on PCNA Inhibitors*, 2013–2015, PI: Prof. W. Strzałka (Jagiellonian University).
5. Grant POKL.04.01.01-00-081/10 by the European Union, *Enhancing Educational Potential of Nicolaus Copernicus University in the Disciplines of Mathematical and Natural Sciences*, 2015, PI: Nicolaus Copernicus University.
6. Horizon 2020 by European Commission 101006318, *Developing and Implementing hands-on training on Open Science and Open Innovation for Early Career Researchers (DIOSI)*, 2021–2022, PI: Universiteit Antwerpen.

II.7. Information on participation in European or other international programs

1. Teaching at the summer camp for international students and PhD students in the program of the Polish National Agency for Academic Exchange SPINAKER — Intensive International Education Programs 4-22th July 2022 (Torun, Poland) co-financed by the European Social Fund under the Operational Programme Knowledge Education Development.
2. Teaching classes about open science in the program "Developing and Implementing hands-on training on Open Science and Open Innovation for Early Career Researchers (DIOSI)" funded by the European Union under the Horizon 2020 (European Commission 101006318, 2020–2023).
3. Collaboration with the University of Taormina in Italy under the European program "Enhancing Educational Potential of Nicolaus Copernicus University in the Disciplines of Mathematical and Natural Sciences" (POKL.04.01.01-00-081/10, 2015)
4. Coordination of actions allowing funding open access publishing fees for young researchers at NCU as the PI of the Polish National Agency for Academic Exchange's project "Open NCU–Open Source, Open Science" (2019–2022).

II.8. Membership in international or national organizations and scientific societies

1. PLUMED consortium (Since 2019): an open community for researchers whose work builds on atomistic simulations, establishing effective protocols for sharing data, promoting scientific reproducibility, and upholding the highest research standards (www.plumed-nest.org/consortium).
2. Polish Bioinformatics Society (2014–2019).

II.9. Information on internships completed in scientific institutions

Fellowships:

1. National Institute of Advanced Industrial Science and Technology (AIST), Japan, Group of Prof. Tetsuya Morishita (XI 2023–II 2024), funded by the Japan Society for the Promotion of Science (JSPS).
2. Department of Theoretical and Computational Biophysics, Max Planck Institute for Biophysical Chemistry, Germany, Group of Prof. Helmut Grubmüller (X 2016–IV 2017), funded by the Etuida grant received from the National Science Center (NCN),
3. Department of Chemistry and Applied Biosciences, Swiss Federal Institute of Technology in Zürich c/o Institute of Computational Science, Università della Svizzera italiana, Switzerland, Group of Prof. Michele Parrinello (VII 2016–X 2016), funded by the Rector Fellowship (NCU).

Visits (up to month):

1. Department of Physics, University of Tsukuba, Japan, the group of Prof. Yasuteru Shigeta, 2019 (1 month).
2. Max Planck Institute for Polymer Research, Mainz, Germany, the group of Prof. Omar Valsson, 2019 (1 week).
3. Quantum Chemistry Research Institute, Kyoto, Japan, 2017, the group of Prof. Hiroshi Nakatsuji (1 month).
4. Institute of Fluid-Flow Machinery - Polish Academy of Sciences, Gdańsk, 2011 (1 month).

II.10. Information on scientific or artistic works reviewed, in particular, those published in international journals

1. American Chemical Society (ACS): Journal of Physical Chemistry Letters, Journal of Chemical Information and Modeling, Journal of Chemical Theory and Computation, Journal of Physical Chemistry B

2. American Institute of Physics (AIP): Journal of Chemical Physics
3. Cell Press: Biophysical Journal
4. Elsevier: Computational and Structural Biotechnology Journal, Journal of Molecular Graphics and Modeling
5. PLOS: PLOS ONE, PLOS Computational Biology
6. Journal of Open-Source Software

II.11. Information on participation in research teams realizing projects other than those defined in section II.6

The Author participates in several research projects in collaboration with scientific groups:

1. Prof. Omar Valsson, University of North Texas, Denton, US, Since 2016. Work related to constructing collective variables for enhanced sampling simulations using optimization and explainable machine learning [H2, H3, H1, A2].
2. Prof. Ming Chen, Purdue University, West Lafayette, US, Since 2018. Work related to constructing Markov transition matrices from enhanced sampling simulations [H3, H1].
3. Prof. Alexander M. Berezhkovski, National Institutes of Health, Bethesda, Maryland, US, Since 2023. Work related to Markovian dynamics along reaction coordinates.
4. Prof. Tetsuya Morishita, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan, Since 2022. Work related to calculating macroscopic variables describing glass formation.
5. Prof. Michele Parrinello, Italian Institute of Technology, Genova, Italy, Since 2016. Work related to developing enhanced sampling methods to simulate ligand unbinding from proteins and establishing protols for transparent and reproducible atomistic simulations [H7].
6. Prof. Helmut Grubmüller, Max Planck Insitutue for Multidisciplinary Sciences, Gottingen, Germany, Since 2016. Work related to estimating kinetics and thermodynamics of inhibitor unbinding from enzymes [A3].
7. Prof. Yasuteru Shigeta and Prof. Ryuhei Harada, University of Tsukuba, Tsukuba, Japan, Since 2019. Work related to implementing bias-free enhanced sampling methods into the PLUMED plugin [H7].
8. Prof. Bruno Lapied, University Angers, Angers, France, Since 2023. Work related to computationally overcoming pyrethroid resistance in mosquito control [A6].

II.12. Information on membership in the teams assessing applications for financing of research projects, applications for scientific awards, applications in other competitions of scientific or didactic character

1. Reviewing applications for the best bachelor thesis in bioinformatics funded by the Polish Bioinformatics Society (2022).
2. Reviewing (as PI of the project) applications for funding open access publishing fees for young researchers at NCU (2019–2022). Funding provided by the Polish National Agency for Academic Exchange's project "Open NCU–Open Source, Open Science."
3. Recruitment committee for selecting candidates for the NCU doctoral school for the research projects granted by the National Science Centre, Poland and the Ministry of Science and Higher Education.

II.13. Others

Awards and scholarships:

1. Scholarship funded by the Japan Society for the Promotion of Science (JSPS), 2022.
2. Scholarship for outstanding achievements by young researchers funded by the Ministry of Science and Higher Education, 2022.
3. Stipend funded by the Foundation of Polish Science (START FNP) for oustanding achievements by researchers under age 30, 2018.
4. PhD student scholarship for outstanding achievements funded by the Ministry of Science and Higher Education, 2017.
5. PhD student scholarship for outstanding achievements funded by the Ministry of Science and Higher Education, 2016.
6. The Marshal of Kuyavian-Pomeranian Voivodeship ("Krok w przyszlosc"), 2015.
7. First place in a competition for the best master thesis in bioinformatics and computational biophysics funded by the Polish Bioinformatics Society (PTBI), 2015.
8. Elected as the best student of Faculty of Physics, Astronomy and Informatics at the Nicolaus Copernicus University, 2015.
9. Many NCU Rector Awards and scholarships.

III. Information on cooperation with social and economic environment

III.1. List of technological works

1. Investigator in the project of the Polish National Centre for Research and Development LIDER/28/54/L-/10/NCBiR/2011, Searching for a Novel Anticancer Drug based on PCNA Inhibitors; PI: Prof. Wojciech Strzałka.