w ilościach wystarczających dla podtrzymania procesów życiowych wówczas, gdy jesteśmy w dobrej kondycji zdrowotnej. Dodatkowa zaś suplementacja powinna być ściśle kontrolowana i raczej wynikać z zaleceń lekarza, aniżeli być wynikiem mody na przyjmowanie preparatów dodających nam zdrowia i urody.

Ciekawe: Jedynym znanym organizmem żywym obywającym się bez żelaza są bakterie z rodzaju *Lactobacillus*. Jedna komórka *Lactobacillus plantarum* zawiera zaledwie 3–4 atomy żelaza, podczas gdy w komórce Escherichia coli jest ich $2x10^5$. W organizmie człowieka żelazo występuje w ilości 3–4 g.

Dr nauk med. Marek Jurgowiak — Katedra i Zakład Biochemii Klinicznej Collegium Medicum im. L. Rydygiera w Bydgoszczy Uniwersytetu Mikołaja Kopernika w Toruniu.

Mgr Beata Ignasiak — doktorantka w Collegium Medicum im. L. Rydygiera w Bydgoszczy, Uniwersytetu Mikołaja Kopernika w Toruniu.

Jakub Rydzewski

Metody wzmocnionego próbkowania

Symulacje atomistyczne (dynamiki molekularnej lub Monte Carlo) powszechnie stosowane w fizyce, chemii i biologii zapewniają wgląd w skalę czasową i przestrzenną, które są niedostępne dla eksperymentów. Wykorzystuje się je do badania własności równowagowych procesów, takich jak zwijanie białek, wiązanie leków oraz przejścia fazowe.

Fot. Andrzej Romański

Pomimo znacznego postępu w rozwoju klastrów obliczeniowych oraz oprogramowania, klasyczne symulacje atomistyczne nie są w stanie próbkować zjawisk o dużej skali czasowej. Problem ten jest związany z wysokimi barierami energetycznymi oddzielającymi minima energetyczne, odpowiadające za realizowane funkcje badanych układów fizycznych. Jeśli bariery te są znacznie wyższe niż energia cieplna układu (kT), układ zostaje kinetycznie uwięziony w metastabilnych minimach energii. Innymi słowy, przekroczenie bariery energetycznej staje się zdarzeniem rzadkim (Valsson, 2016). Z teoretycznego punktu widzenia, aby móc próbkować zdarzenie rzadkie, czas symulacji powinien dążyć do nieskończoności. Jak można przypuszczać, jest to problem nieprzekraczalny dla obecnych inftrastruktur obliczeniowych. By rozwiązać problem ergodyczności, konieczny jest znaczny postęp metodologiczny.

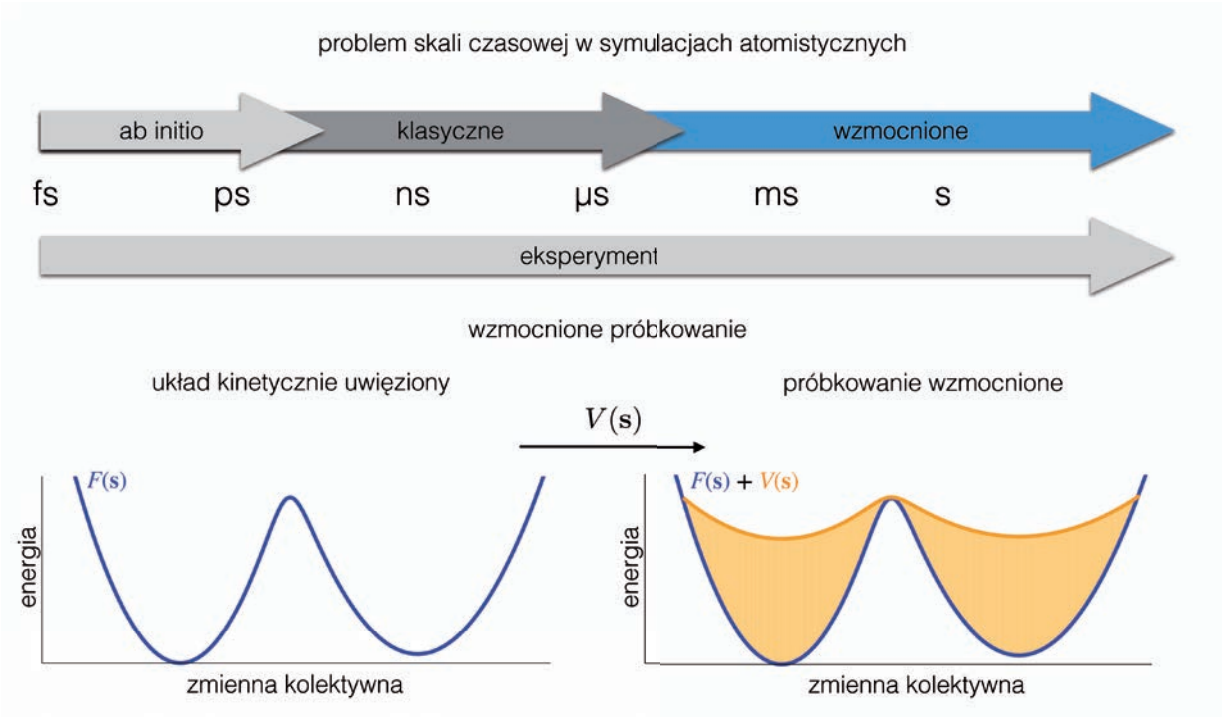
Możliwość symulowania zdarzeń rzadkich zapewniają metody tzw. wzmocnionego próbkowania (ang. *enhanced sampling*). Okazały się one kluczowe w dostępie do dłuższych skal czasowych. Jedna z klas takich metod opiera się na zidentyfikowaniu kilku gruboziarnistych zmiennych kolektywnych, dzięki którym można prześledzić badaną reakcję, tj. odpowiednio rozróżniających metastabilne minima i odpowiadających powolnym stopniom swobody. Próbkowanie zdarzeń rzadkich podczas symulacji atomistycznej opiera się na wzmocnieniu wahań zmiennych kolektywnych za pomocą zewnętrznego potencjału niefizycznego (tzn. potencjału innego,

niż potencjał wynikający z pola siłowego, wykorzystywanego w dynamice molekularnej), który ułatwia przekraczanie barier energetycznych. Pierwszym algorytem należącym do tej klasy metod jest próbkowanie parasolowe (ang. *umbrella sampling*), który swoją nazwę zawdzięcza serii zewnętrznych potencjałów harmonicznych wzdłuż wybranej zmiennej kolektywnej (Torrie, 1977).

Wydajność metod wzmocnionego próbkowania opartych na zmiennych kolektywnych zależy w dużym stopniu od jakości wybranych zmiennych kolektywnych. Niestety, wybór ten dla złożonego układu biologicznego, takiego jak białko, nie jest trywialny. Przykładowo, problem ten pojawia się w badaniu dysocjacji kompleksu ligand-białko, ponieważ zdarzenie to może realizować się na kilka różnych sposobów. Jest to związane z faktem mówiącym, że białko może mieć wiele tunelów, przez które mogą podróżować małe cząstki organiczne, a niektóre drogi dysocjacji mogą być nieidentyfikowalne przez badanie statycznych struktur krystalograficznych białek. Błędny dobór zmiennych kolektywnych spowodowałby zafałszowanie skali czasowej, w której zachodzi dysocjacja, a jest ona proporcjonalna do czasu działania leku w organizmie. Takie rozważania teoretyczne i obliczeniowe są kluczowe dla projektowania nowych leków – symulacje mogą być w przyszłości stosowane w farmakologii. Z tego względu projektowanie skutecznych metod uzyskiwania kinetyki zdarzeń rzadkich jest niezwykle ważne.

Ostatnimi czasy można zaobserwować wzrost użycia metod „sztucznej inteligencji" w naukach przyrodniczych. Warto zaznaczyć, że tak chętnie używane przez media określenie „sztuczna inteligencja" jest coraz częściej niewłaściwie używane, także przez badaczy. Określenie jest to obecnie używane jako termin parasolowy (ang. *umbrella term*), do którego wrzucane są wszystkie metodologiczne rozwiązania związane z „uczeniem" się komputerów. Wg jednego z pionierów uczenia maszynowego, Maxa Wellinga (Welling, 2019), „The term Artificial Intelligence is overloaded, the field is overhyped, and the media could be more objective about the way they report", z czego płynie wniosek, że powinno się używać terminu „uczenie maszynowe", który jest właściwie zdefiniowany. Arthur Samuel, któremu przypisuje się ukucie terminu „uczenie maszynowe", zaproponował definicję, według której algorytmy uczenia maszynowego budują model matematyczny, bazując na próbkach danych, by przewidywać, nie będąc jednocześnie wprost zaprogramowanymi, by realizować to zadanie.

Uczenie maszynowe zostało z powodzeniem użyte w celu rozwiązania najróżniejszych problemów fizycznych (Goodfellow, 2016), w tym także tych związanych z metodami wzmocnionego próbkowania bazujących na zmiennych kolektywnych. Metody uczenia maszynowego, zastosowane w symulacjach, mają za zadanie dostarczyć niskowymia-



Ilustracja przygotowana przez autora

rowy zbiór zmiennych kolektywnych opisujących ważne stopnie swobody badanego układu. Niskowymiarowość tego zbioru związana jest z wymaganym ograniczeniem czasu obliczeniowego. Niewystarczająca definicja zestawu zmiennych kolektywnych może prowadzić do degeneracji barier energetycznych, co oznacza, że ważne stopienie swobody badanego procesu zostały pominięte w tej definicji. Obserwable kinetyczne i termodynamiczne otrzymane w procesie próbkowania błędnych zmiennych kolektywnych są niefizyczne, co wyklucza porównanie wyników symulacji z eksperymentami.

Największym obecnie wyzwaniem w użyciu metod uczenia maszynowego w metodach wzmocnionego próbkowania jest stwierdzenie, jak algorytm zapewni definicję zmiennych kolektywnych, skoro idealny zestaw zmiennych kolektywnych możliwy jest do otrzymania po nieskończonym czasie obliczeń? Ten tzw. paradoks ergodyczny doprowadził do dynamicznej definicji zmiennych kolektywnych, które są automatycznie ulepszane podczas trwania symulacji. Początkowo, zmienne kolektywne, jak i obserwable fizyczne badanego układu nie są właściwe, lecz im więcej próbek zostaje zebranych podczas symulacji, tym lepsza jest definicja zmiennych kolektywnych. Takie podejście do metod wzmocnionego próbkowania, bazujących na zmiennych kolektywnych, nazywane jest aktywnym wzmocnionym próbkowaniem (ang. *active enhanced sampling*) i wykorzystuje ono metody z fizyki obliczeniowej, uczenia maszynowego oraz teorii informacji (Zhang, 2018). Wynikowy zestaw zmiennych kolektywnych jest tworzony poprzez dopasowanie rozkładu prawdopodobieństwa niskowymiarowych zmiennych kolektywnych do rozkładu prawdopodobieństwa wysokowymiarowych zmiennych badanego układu, takich jak, np. położenia atomów układu. Interpretacja tego procesu jest prosta z punktu widzenia teorii informacji – strata informacji o układzie zakodowanej w wysokowymiarowych zmiennych poprzez zastąpienie go niskowymiarowymi zmiennymi kolektywnymi jest jak najmniejsza. Algortymy tego typu umożliwiają próbkowanie zdarzeń rzadkich w skończonym czasie obliczeniowym.

Celem badań prowadzonych w Instytucie Fizyki UMK jest opracowanie nowych rozwiązań metodologicznych koncentrujących się na udoskonalonych metodach próbkowania, w tym na wykorzystaniu pomysłów z uczenia maszynowego w celu opracowania metod wyszukiwania wolnych zmiennych kolektywnych oraz udoskonalonych metod uzyskiwania kinetyki rzadkich zdarzeń z symulacji atomistycznych. Niniejsze badania są prowadzone zgodnie z modelem „otwartej nauki" (ang. *open science*), który jest odpowiedzią na tzw. „kryzys replikacyjny", według którego większość rezultatów publikacji naukowych z dziedziny nauki przyrodniczych jest niemożliwa do odtworzenia przez niezależnych badaczy, a także samych autorów (Staddon, 2017). Model otwartej nauki jest bardzo ważnym elementem procesu badawczego, szczególnie dla nauk wykorzystujących metody obliczeniowe. Pomocne w realizacji badań tego typu jest nowopowstałe konsorcjum PLUMED, którego UMK jest aktywnym członkiem. Konsorcjum to ma na celu publikowanie protokołów badań, danych i modelów wykorzystywanych w symulacjach atomistycznych, które są niezbędne do zreprodukowania wyników badań (The PLUMED Consortium, 2019). Jest to pierwsza tego typu inicjatywa na świecie w kontekście nauk obliczeniowych.

Dr inż. Jakub Rydzewski — Wydział Fizyki, Astronomii i Informatyki Stosowanej UMK.

Podziękowania

Autor dziękuje za wsparcie finansowe Fundacji na rzecz Nauki Polskiej (START FNP).

Literatura

Goodfellow I., Bengio Y., Courville A. (2016). Deep Learning. *MIT Press*.

Staddon J. (2017). How Science Works, Fails to Work, and Pretends to Work. *Routledge*.

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

Torrie G. M., Valleau J. P. (1977). Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *Journal of Computational Physics* **23**, 187–199.

Valsson O., Tiwary P., Parrinello M. (2016). Enhancing Important Fluctuations: Rare Events and Metadynamics from a Conceptual Viewpoint. *Annual Review of Physical Chemistry* **67**, 159–184.

Welling, M. (2019). Artificial Intelligence versus Intelligence Engineering. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.364ce476

Zhang J., Chen M. (2018). Unfolding Hidden Barriers by Active Enhanced Sampling. *Physical Review Letters* **121,** 010601.