

Data: Toruń, 12/12, 2016

L. Dz.

Prof. dr hab. Wiesław Nowak
Zespół Teoretycznej Biofizyki Molekularnej
Instytut Fizyki
Uniwersytet M.Kopernika w Toruniu
87-100 Toruń, ul. Grudziądzka 5

RECENZJA rozprawy doktorskiej

Mgr Macieja Dziubińskiego

pt.

“Development of selected mesoscopic physical models with the aid of machine learning methods and their applications in studies of molecular systems”

Metody modelowania komputerowego w sposób trwały wkroczyły do badań fizycznych, w tym do biofizyki. Obszary zastosowań obejmują zarówno symulacje nanoskopowych układów kwantowych, jak i obiekty stosunkowo duże, np. biopolimery. Poziom abstrakcji obiektów zmienia się: od elektronów, przez atomy, do całych fragmentów molekularnych, aminokwasów, dużych biopolimerów czy nawet komórek. Metody modelowania mają różny stopień ogólności, ale zwykle lokują się w dość szerokich przedziałach rozmiarów i skal czasowych. W recenzowanej rozprawie doktorskiej podjęto ambitną próbę powiększenia arsenału metod symulacji komputerowych w zakresie skal mezoskopowych rozmiarów obiektów (obejmujących proste molekuly i b. duże cząsteczki) oraz skal ponad nanosekundowych w domenie czasowej. Autor rozprawy rozwija nowe metody analizy wyników symulacji dynamiki molekularnej (MD). Badania takie są niezwykle popularne i w świecie pojawiają się coraz liczniejsze depozytoria z wynikami takich czasochłonnych obliczeń. Niestety, metody analizy są wciąż stosunkowo prymitywne i archaiczne. Pomysł by wykorzystać uczenie maszynowe do wydobywania nowych informacji z tych danych jest ze wszech miar słuszny. Technika ta, klasyfikowana obecnie w obszarze metod sztucznej inteligencji, jest b. słabo wykorzystywana przez grupy badawcze działające w dziedzinie MD, a ma ona ogromny potencjał na skuteczne wspieranie analiz rezultatów symulacji komputerowych. Autor podjął próbę rozwiązania dwóch istotnych problemów: (1) identyfikacji stałych struktur w białkach zwanych domenami dynamicznymi (2) określenia zespołów atomów współdziałających w trakcie przemian konformacyjnych w cząsteczkach, tzw. trybów (czy przekładni) molekularnych. Resumując wprowadzenie do recenzji stwierdzam, że w **rozprawie doktorskiej podjęto ważny, ciekawy i aktualny problem badawczy**. Temat rozprawy, może nieco za ogólny, dobrze opisuje zawartość dzieła.

Praca doktorska (nauki fizyczne, dyscyplina biofizyka) została napisana w Zakładzie Biofizyki Wydziału Fizyki Uniwersytetu Warszawskiego, pod kierunkiem prof. dr hab. Bogdana Lesynga. Funkcje promotora pomocniczego sprawował dr Paweł Daniluk z tego samego Zakładu. Praca napisana została w ramach nowej ustawy, zatem ma formę „spójnego tematycznie zbioru artykułów opublikowanych lub przyjętych do druku w czasopismach naukowych”. Są to dwa, artykuły współautorskie opublikowane w solidnych, o znacznej renomie, międzynarodowych czasopismach specjalistycznych (1) M. Dziubiński, P. Daniluk and B. Lesyng, „ResiCon: a method for the identification of dynamic domains, hinges and interfacial regions in proteins”, *Bioinformatics* 32 (2016) 25-34. (2) M. Dziubiński and B. Lesyng, „Toward the identification of molecular cogs”, *J. Comp. Chem.* 37 (2015) 848-860. Załączone do dokumentacji oświadczenia współautorów jednoznacznie wskazują, iż dominujący wkład w powstanie obu publikacji miał pan mgr. Dziubiński, jest też w nich jasno określony zakres prac wykonanych samodzielnie przez doktoranta.

Praca napisana jest w bardzo poprawnym języku angielskim, stylem zwięzłym i informatywnym. Ma 105 stron obejmujących: *Wstęp* (13 str.), *Rozdział 2* cytujący publikację (#1) (18 str.), *Rozdział 3* – przytoczenie publikacji (#2) (20 str.), *Wnioski* (4 str.), *Dodatki do prac #1 i #2* (34 strony) i *Bibliografię* (83 pozycje). Do pracy dołączono streszczenie rozprawy w języku polskim (4 strony). W wykazie zawartym w spisie treści brakuje odniesienia do strony z bibliografią. Podział treści rozprawy jest prawidłowy, rysunki i tabele dobrej jakości, podpisy na ogół kompletne i odpowiednie. Cele rozprawy są opisane w końcowej części *Wstępu* oraz najogólniej w streszczeniu w języku polskim, gdzie czytamy: „Celem rozprawy było opracowanie optymalnej metodologii i narzędzi pozwalających stworzyć uproszczone opisy złożonych, dynamicznych układów molekularnych”. Po zapoznaniu się z treścią obu publikacji mogę stwierdzić, że **cel ten został osiągnięty**. Rozprawa doktorska zawiera szereg oryginalnych i wartościowych wyników składających się na realizację tych celów.

Przedstawię kolejno ocenę rozdziałów i uwagi merytoryczne na ich temat. Uwagi formalne podam w dalszej części recenzji.

Wstęp zawiera ogólną prezentację tematyki: zwięzłe opisy uczenia maszynowego, uczenia pod nadzorem oraz uczenia bez nadzoru. Krótkie paragrafy poświęcone są metodzie analizy głównych składowych (PCA) oraz przeglądowi ważniejszych metod klasteryzacji danych. Autor słusznie sygnalizuje ograniczenia metody PCA i zwraca uwagę na algorytm t-SNE (nb.

stosowany w naszym zespole). Podane jest ważne w rozprawie pojęcie macierzy podobieństwa W i pojęcie potencjału średniej siły (PMF). Dążenie do zwięzłości opisu poprzez lakoniczność prowadzi, niestety, niekiedy do niekompletności. Taka sytuacja przytrafiła się Autorowi *Wstępu* w podrozdziale 1.2 poświęconemu metodzie MD: Po rzetelnym opisie sił występujących w metodzie, pewnej dyskusji potencjału, znajdujemy w tej części od razu rozważanie na temat zalet i wad MD oraz obliczeń energii swobodnej. Czytelnik mniej zorientowany nie dowie się sedna metody, tj. nie pozna odpowiedniego równania ruchu ani nie dowie się, iż symulacje MD polegają na rozwiązywaniu numerycznym układów tysięcy czy setek tysięcy takich równań (poza wzmianką o szczególnej postaci MD tj. dynamice Langevina w sekcji 1.2.1, str. 11). Podobnie lakoniczny opis uczenia maszynowego bez nadzoru zastosowanego do analizy danych molekularnych (zawarty w paragrafie 1.2.2) nie pozwala Czytelnikowi zorientować się czy w omawianych algorytmach parametr t (zwykle związany z czasem, przynajmniej w trajektoriach MD) jest ważny czy też będzie ignorowany. Omawiane pojęcia i zagadnienia są klasyczne i szkoda, że Autor nie popróbował podawać za każdym razem odnośników do monografii czy książek szerzej opisujących prezentowane treści. Pomimo tej lekkiej krytyki *Wstęp* oceniam dobrze, właściwie spełnia on swoją podstawową rolę i szybko wprowadza Czytelnika w podstawowe zagadnienia prezentowane w głównych Rozdziałach 2 i 3. Dowodzi też erudycji i dobrej orientacji w dziedzinie modelowania molekularnego doktoranta.

W Rozdziale 2 rozwijana jest metoda ResiCon służąca do identyfikacji dynamicznych domen białkowych w oparciu o analizy zbiorów konfiguracji 3D białek. Pomysł polega na przyjęciu założenia, że „siła kontaktów” pomiędzy residuami aminokwasowymi zależy wyłącznie od zmian geometrycznych występujących w zadanym zbiorze struktur. ResiCon służy do identyfikacji domen dynamicznych, ale może być wykorzystana do identyfikacji połączeń między domenami czy rejonów ich kontaktów. Autorzy metody proponują pewną definicję zmienności geometrycznej (dość naturalną) i konstrukcję macierzy kontaktów opartą o funkcję sigmoidalną L (*logistic f.*). To jest dobry pomysł bowiem parametry tej funkcji pozwalają na elastyczne dopasowanie definicji kontaktów do specyfiki analizowanego zbioru konfiguracji. W ramach metody konstruuje się szkielet (graf), w którym połączenia między resztami aminokwasowymi reprezentowane są liniami o grubości (*stiffness*) proporcjonalnej do ważności (*strength*) analizowanego kontaktu. Algorytm klasteryzacji służy następnie do przecinania najsłabszych połączeń i identyfikacji zespołów „najmocniejszych” – te zestawy aminokwasów tworzą wspomniane domeny dynamiczne. W rozdziale elegancko i szczegółowo opisano metodę. Podano logiczne argumenty na rzecz stosowania klasteryzacji opartej na widmie (wartościach

własnych) macierzy podobieństwa. Przedyskutowano dogłębnie problem wyboru (założenia) liczby klastrów oraz zaproponowano nieco heurystyczny sposób wyboru optymalnej liczby klastrów. Podano też prosty i naturalny sposób definiowania obszarów łącznikowych czy zawiasów (*hinges*). Wykonano analizy 30 struktur NMR białka proteaza HIV (macierz podobieństwa 150x150) badanych wcześniej w innych publikacjach. Cenna częścią rozdziału jest staranna analiza jakości otrzymanych rozwiązań (podziału na domeny dynamiczne). Autorzy publikacji zaproponowali stosowanie indeksu VI (*Variation of Information*) i porównali swoje wyniki z niedawno opublikowanymi metodami znajdowania domen dynamicznych GeoStaS (2012 – zespół prof. J. Trylskiej) oraz PiSQRD (2009 R. Potestio i inn.). Wyniki są kompatybilne z wcześniejszymi metodami, za to nowa metoda opisana w doktoracie jest znacznie mniej czuła na szczegóły parametryzacji stosowane do klasteryzacji i wyznaczania domen. Co ciekawe, nowa metoda ResiCon pozwoliła na zauważenie pewnych cech strukturalnych budowy proteazy nie znanych wcześniej. Jest to silna wskazówka na rzecz jedynie antycypowanej na początku badań tezy, iż uczenie maszynowe może przyczynić się do nowych odkryć, jeśli zostanie umiejętnie wykorzystane do analizy danych z dynamiki strukturalnej białek.

Rozdział 2 zrobił na mnie b. dobre wrażenie, Zawiera on wiele nowych pomysłów, dyskusji istotnych problemów metodologicznych, dużo danych obliczeniowych i graficznych (obszerne SI). Problem atakowany nie jest trywialny, sam bym chętnie posiadał do dyspozycji łatwy i przyjemny w użyciu algorytm (metodę) wyszukującą niejako rutynowo i automatycznie domeny dynamiczne we wszystkich białkach badanych w moim laboratorium. Obawiam się jednak, że przejście do rutynowej pracy badawczej, najlepiej w schemacie „*black box*”, z metodą ResiCon na tym etapie rozwoju nie jest jeszcze możliwe. Byłbym ciekaw, jak ten problem widzi kandydat do stopnia doktora. W każdym razie na pewno pozytywnie należy ocenić udostępnienie kodu programu wszystkim zainteresowanym. Pewien drobny niedosyt budzi dość pobieżna dyskusja podejmowanych wcześniej prób poszukiwania domen dynamicznych. Co prawda autor odnosi się do kilku prac na ten temat, ale w spisie nie zauważyłem dość dobrze cytowanych prac M. Stepanowej z Edmonton w Kanadzie (np. „Dynamics of essential collective motions in proteins: Theory”, Maria Stepanova, Phys. Rev. E 76, 051918 – Published 26 November 2007; Molecular Simulation, Vol. 37, 2011: Nanotech 09-10.”Identification of dynamic structural domains in proteins, analysis of local bond flexibility and application for interpretation of NMR experiments”, Maria Stepanova Pages 729-732 Published online: 25 Sep 2009.)

W Rozdziale 3 podano próbę rozwiązania jeszcze trudniejszego problemu: znalezienia odpowiedzi na pytanie co napędza przemiany konformacyjne w cząsteczkach. Przemiany takie są faktem. Zgodnie z powszechnym odczuciem w molekułach, zwłaszcza większych, poważne grupy atomów przemieszczają się względem siebie by zrealizować przemianę z konformacji A do B. Autorzy stawiają hipotezę, że muszą wobec tego być w systemie zestawy atomów „popychających” w kierunku tego przejścia a także mogą być atomy „hamujące”. Osobiście uważam, że jest to bardzo odważna i mocno „antropomorficzna” hipoteza, jednak czytając argumentację przedstawioną w publikacji, przypuszczam, że w węższym kontekście specyficznej analizy klasycznych pól siłowych tego typu idee mogą być pożyteczne.

Punktem wyjścia są rozważania na temat hiperpowierzchni energii swobodnej układu molekularnego. W przestrzeni konformacyjnej możliwe są przejścia i dają się one przedstawić w przestrzeniach niskowymiarowych, o ile określi się odpowiednie zmienne kolektywne, dające tzw. współrzędne reakcji. Autorzy w sprytny sposób rozkładają wkłady do energii swobodnej na wkłady dwuciałowe (dwuatomowe), wprowadzając przy tym człon kooperatywny (*cooperation term*). Definiują też nową macierz kooperatywności (*transient cooperation matrix*). W tej macierzy szukają klastrów (grup atomów) i te klastry nazywają trybami molekularnymi – jeśli kooperatywność między danymi atomami z grupy obniża energię swobodną są to tryby napędzające (*forward cogs*), zaś jeśli podwyższa – hamujące czy może wsteczne (*reverse cogs*). Ważną formalnie częścią pracy jest próba klasteryzacji danych reprezentowanych w macierzy TCM. Nie jest to trywialne zadanie, Autor wykazał się tutaj dużą pomysłowością, znajomością cech metod klasteryzacji i sporą dozą dociekliwości i potrzebnego krytycyzmu. W rozdziale stosowane są znane algorytmy AP (*Affinity Propagation*) i specyficzny algorytm genetyczny zorientowany na maksymalizację nowo zdefiniowanej funkcji celu SCORE, opartej na molekularnym obrazie postępu reakcji i wkładach hipotetycznych trybów do tej reakcji. Wprowadzono własną graficzną reprezentację wkładów atomów do energii swobodnej związanej z badaną przemianą konformacyjną. Pokazano na przykładzie, że autorski algorytm genetyczny znajduje optymalne rozwiązania. Rozwiniętą metodę analizy zastosowano do prostej cząsteczki chemicznej: kationu 2,2 diiodo-1-amino-etanu, oddziaływania przy rotacji wybranej grupy funkcyjnej były obliczane polem siłowym GAFF z ładunkami AM1-BCC. Niestety, nie zauważyłem w tym rozdziale uzasadnienia wyboru akurat tej molekuly do głębszych badań. Nie jest też dla mnie jasne, czy doktorant sam pisał kod do metody obliczeniowej cMD, czy też może skorzystał z gotowego rozwiązania? (str. 49). W tabelach 3. i 3.4 przedstawiono w pomysłowy sposób główne wyniki analiz. Można było pokusić się o ocenę, które atomy sprzyjają przejściu konformacyjnemu, a które nie oraz rozbić oddziaływania (wkłady

do zmian energii) na przyczynki. Autorzy poddali również analizie porównawczej wyniki obliczeń dla tego samego przejścia konformacyjnego w 4 innych bardzo zbliżonych pochodnych.

Ocena ogólna tego rozdziału budzi we mnie nieco wątpliwości. Z jednej strony cenny jest oryginalny pomysł analizy wkładów par atomów do zmian energii swobodnej. Widać w pracy dużą biegłość formalną i rozumienie problemów modelowania i klasteryzacji. Z drugiej strony wyniki są jednak trudne do uzyskania, interpretacji i weryfikacji. Analiza takiego prostego przypadku, na moje wyczucie, wnosi niewiele do zrozumienia przyrody, może poucza nas więcej o naturze konkretnych pól siłowych i modeli molekularnych. Byłoby bardzo ciekawym rozwinąć podobny formalizm, ale dla naprawdę dużych systemów, chociażby peptydów. Tam jednak, jak sami autorzy zauważają, mamy poważną trudność z wyborem zmiennych kolektywnych, a tym samym zdefiniowaniem rozsądnych współrzędnych reakcji. W oddali mający nierozwiązany problem efektywnej analizy procesu zwijania białek, też byłoby dobrze poznać "tryby" tam działające, jednak złożoność tego zagadnienia może być po prostu za wielka na zdolności rozdzielcze algorytmów klasteryzujących opartych na podanej w rozprawie funkcji SCORE.

Widząc u siebie te wątpliwości w próbie podsumowania jakości Rozdziału 3 uznałem jednak dość szybko, że ocena tej części rozprawy powinna być też wysoka: po prostu publikacja ta prowokuje do zadawania pytań, daje nowe spojrzenie na stuletni problem zmiany konformacyjnej i wprowadza nowe techniki do badania problemu. Jest to niewątpliwie wysokiej jakości wkład do nauki, nawet jeśli badana cząsteczka nie jest lekiem na wszelkie zło. Uzyskanie tych rezultatów na pewno wymagało dużej biegłości warsztatowej, rozumienia metod pracy naukowej, solidnego programowania i dobrej analizy wyników. Praca przebiła się do J. Comp. Chem. co też świadczy o randze rezultatów i spełnieniu standardów międzynarodowych jakości badań.

Uwagi szczegółowe:

1. W Rozdziale 3 Autor posługuje się stale terminem energia swobodna, ale już na stronie 42 deklaruje, iż w tak małym systemie możemy zaniedbać wkłady entropowe, co zresztą czyni. Czy nie lepiej byłoby zatem dalej pisać po prostu o energii wewnętrznej?
2. Rys. 3.5. Nie jest dla mnie jasne skąd wziął się „prawdziwy” PMF? Na wykresie widzimy minimum potencjału w ok. -170 deg. zaś na tym samym rysunku pokazano strukturę dla kąta -150 deg. Dlaczego? Nie podano jakim programem graficznym wykonano rysunki molekuł.

3. Str. 56, 9d, literówka „of fewer”, a powinno być „or fewer”
4. Str. 58, „conformational interactions” – wyjątkowo dla mnie niejasny typ oddziaływań, są takie?
5. W Podsumowaniu, na str. 62 czytamy „groups of atoms, cooperating to propel the system forwards along a given collective variable”. Czy atomy “popychają” system, czy raczej są same “popychane” siłami wynikającymi z gradientu potencjału jaki “czują”?

Uwagi redakcyjne

Praca zredagowana jest starannie, nie znam na tyle języka angielskiego (non-native..) by krytykować kompetentnie styl czy gramatykę, poważnych uchybień wg mnie ma, drobne uwagi są następujące:

1. Alfabetyczny spis pozycji bibliograficznych połączony z enumeratywnym sposobem odwoływania się do literatury jest BARDZO niewygodny dla czytelnika.
2. Spis literatury nie ma jednolitego stylu.
3. Tabele w Dodatkach są źle rozmieszczone. Orientacja stron „do środka” jest b. niepraktyczna i nie spotykana w książkach ani w publikacjach.

KONKLUZJA

Przedłożona mi od recenzji rozprawa zawiera nowe, wartościowe pomysły oraz szereg nowych i cennych wyników. Sposób prezentacji zagadnienia i wyników jest dobry, a momentami bardzo dobry. Wyniki zostały opublikowane w prestiżowych czasopismach międzynarodowych. Wiele też pobudza do dalszych badań. Nakład pracy był znaczny, wymagał znakomitych umiejętności komputerowych, dużej wiedzy fizycznej i pewnej wiedzy biologiczno-chemicznej. Tematyka łączy się z dużym nurtem badań biofizycznych modelowaniem komputerowym biomolekuł.

Resumując, stwierdzam, że praca stanowi oryginalne rozwiązanie problemu naukowego, wykazuje znakomitą wiedzę teoretyczną i praktyczną kandydata w dziedzinie modelowania komputerowego i wskazuje dobitnie na umiejętność samodzielnego prowadzenia pracy naukowej. Spełnia wszelkie kryteria ustawowe („nowy tryb”) i zwyczajowe stawiane rozprawom doktorskim w dziedzinie nauk fizycznych, dyscyplinie biofizyka, wnoszę zatem o dopuszczenie mgr. Macieja Dziubińskiego do dalszych etapów przewodu doktorskiego.

W. Nowak