

UNIWERSYTET WARSZAWSKI
Międzydziedzinowa Szkoła Doktorska



ROZPRAWA DOKTORSKA

Metody wieloskalowego modelowania oddziaływań białko-białko i białko-peptyd

Mateusz Zalewski

*Praca wykonana pod kierunkiem:
prof. dr. hab. Sebastiana Kmiecika
Uniwersytet Warszawski,
Centrum Nauk Biologiczno-Chemicznych,
Wydział Chemii,
Laboratorium Biologii Obliczeniowej*

***prof. dr. hab. Dimy Kozakova**
Stony Brook University,
Department of Applied Mathematics and Statistics*

*“Don't go where the path may lead.
Go instead where there is no path
and leave a trail.”*

— Ralph Waldo Emerson

PODZIĘKOWANIA

Ta rozprawa jest zwieńczeniem kilku intensywnych lat pracy i jednocześnie kawałkiem mojego życia, którego nie da się oddzielić od ludzi, którzy byli obok i dawali mi siłę, kiedy najbardziej jej potrzebowałem.

Dziękuję Prof. dr hab. Sebastianowi Kmiecikowi.

Dziękuję Ci za opiekę naukową, zaufanie i cierpliwość oraz za to, że potrafiłeś jednocześnie stawiać wysokie wymagania i dawać przestrzeń do samodzielnego myślenia. To były lata pracy, w trakcie których Twoje uwagi i rozmowy niejednokrotnie mnie ukierunkowały i doprowadziły do tego, kim jestem dzisiaj. Bez Ciebie nie byłbym w tym punkcie ani jako naukowiec, ani jako człowiek.

Dziękuję mojej żonie.

Największe podziękowania kieruję do Ciebie za obecność, spokój i siłę, którą wносиłaś każdego dnia. Dziękuję za cierpliwość, wyrozumiałość i wsparcie w całym tym procesie także wtedy, gdy tempo było wysokie, a ja byłem pochłonięty pracą. Dziękuję, że byłaś obok, kiedy trzeba było po prostu iść dalej. To, że ta praca powstała, jest w dużej mierze także Twoją zasługą.

Dziękuję moim Rodzicom.

Dziękuję Wam za wsparcie, jakie otrzymywałem od zawsze, za fundament, który daliście mi na starcie. To dzięki Wam miałem z czego budować, gdy przyszło udźwignąć coś tak dużego.

Skróty:

A4D - Aggrescan4D
ADCP - metoda dokowania ADCP (*AutoDock CrankPep*)
AF - AlphaFold (rodzina modeli)
AF2 - AlphaFold2
AF3 - AlphaFold3
Aggrescan - predyktor agregacji sekwencyjnej Aggrescan
AI - sztuczna inteligencja (*Artificial Intelligence*)
AMBER - pole siłowe AMBER (*Assisted Model Building with Energy Refinement*)
API - interfejs programistyczny aplikacji (*Application Programming Interface*)
bGH - bydlęcy hormon wzrostu (*bovine Growth Hormone*)
CABS - model gruboziarnisty CABS
CABS-dock - metoda dokowania CABS-dock
CABS-flex - metoda symulacji elastyczności CABS-flex
CamSol - predyktor rozpuszczalności CamSol
CAPRI - *Critical Assessment of PRedicted Interactions*
CASP14 - 14. edycja konkursu CASP (*Critical Assessment of protein Structure Prediction*)
cg2all - rekonstrukcja atomowa cg2all (*coarse-grained to all-atom*)
CHARMM - pole siłowe CHARMM (*Chemistry at HARvard Macromolecular Mechanics*)
cryo-EM - krioelektronowa mikroskopia pojedynczych cząsteczek
CTD - domena C-końcowa (*C-Terminal Domain*)
ESMFold - model językowy do predykcji struktur ESMFold
FoldX - narzędzie FoldX do oceny stabilności
GPU - procesor graficzny (*Graphics Processing Unit*)
GROMOS - pole siłowe GROMOS (*GRoningen MOlecular Simulation*)
hGH - ludzki hormon wzrostu (*human Growth Hormone*)
HSA - albumina surowicy ludzkiej (*Human Serum Albumin*)
IDP - białko nieuporządkowane wewnętrznie (*Intrinsically Disordered Protein*)
iRMSD - RMSD interfejsu (*interface RMSD*)
LoRMSD - RMSD ligandu (*ligand-only RMSD*)
LRMSD - RMSD ligandu (*ligand RMSD*)
MARTINI - gruboziarniste pole siłowe MARTINI
MD - dynamika molekularna (*Molecular Dynamics*)
MSA - wielosekwencyjne ułiniowanie (*Multiple Sequence Alignment*)
NMR - jądrowy rezonans magnetyczny (*Nuclear Magnetic Resonance*)
NPT - zespół izotermiczno-izobaryczny (stałe: N, P, T)
NVT - zespół kanoniczny (stałe: N, V, T)
OPLS - pole siłowe OPLS (*Optimized Potentials for Liquid Simulations*)
PDB - *Protein Data Bank*
pLDDT - przewidywany lokalny wskaźnik dokładności (*predicted Local Distance Difference Test*)
PPI - oddziaływania białko-białko (*Protein-Protein Interactions*)
REMC - *Replica Exchange Monte Carlo*

REST - architektura REST (*Representational State Transfer*)
RMSD - średniokwadratowe odchylenie odległości (*Root-Mean-Square Deviation*)
RMSF - średniokwadratowa fluktuacja (*Root-Mean-Square Fluctuation*)
RSA - względna dostępność rozpuszczalnika (*Relative Solvent Accessibility*)
SolubiS - SolubiS (predyktor rozpuszczalności)
STAP - region podatny na agregację strukturalną (*Structural Aggregation-Prone*)
ZDOCK - zbiór testowy/metoda ZDOCK

Uwaga: STAP - liczba mnoga: STAPy.

Spis treści

CZĘŚĆ I: PODSTAWY TEORETYCZNE I METODOLOGICZNE

1. Wprowadzenie: cel i zakres pracy
 - 1.1. Oddziaływania białko-białko i białko-peptyd: znaczenie biologiczne i implikacje terapeutyczne
 - 1.2. Wyzwania modelowania struktur kompleksów białko-białko i białko-peptyd
 - 1.3. Struktura pracy i wykaz publikacji
2. Podstawy teoretyczne i metody obliczeniowe
 - 2.1. Metody eksperymentalne i obliczeniowe w określaniu struktury białek
 - 2.2. Dokowanie molekularne
 - 2.2.1. Klasyczne podejścia do dokowania
 - 2.2.2. Metoda CABS-dock
 - 2.2.3. Metody przewidywania struktur oparte na AI
 - 2.3. Dynamika białek
 - 2.3.1. Dynamika molekularna (MD)
 - 2.3.2. Symulacja elastyczności - CABS-flex
 - 2.4. Predykcja agregacji z uwzględnieniem struktury i elastyczności konformacyjnej

CZĘŚĆ II: CZĘŚĆ BADAWCZA ROZPRAWY

3. Rozwój nowych metod dokowania białek i peptydów
 - 3.1. [A1] Dokowanie białko-białko z dużą elastycznością łańcucha głównego z wykorzystaniem gruboziarnistych symulacji Monte Carlo
 - 3.1.1. Wprowadzenie do [A1]
 - 3.1.2. Artykuł [A1]
 - 3.2. [A2] Elastyczne dokowanie cyklicznych peptydów do białek z użyciem metody CABS-dock
 - 3.2.1. Wprowadzenie do [A2]
 - 3.2.2. Artykuł [A2]
 - 3.3. [A3] Metoda oceny modeli kompleksów białko-peptyd pochodzących z dokowania gruboziarnistego oparta na dynamice molekularnej
 - 3.3.1. Wprowadzenie do [A3]
 - 3.3.2. Artykuł [A3]
 - 3.4. [A4] Dokowanie białko-peptyd z wykorzystaniem modelu językowego ESMFold
 - 3.4.1. Wprowadzenie do [A4]
 - 3.4.2. Artykuł [A4]
4. Modelowanie elastyczności struktur białkowych: CABS-flex
 - 4.1. [A5] Badanie funkcji białek na podstawie ich elastyczności strukturalnej przy użyciu modelowania CABS-flex
 - 4.1.1. Wprowadzenie do [A5]
 - 4.1.2. Artykuł [A5]
 - 4.2. [A6] Rozwój metody CABS-flex 3.0: narzędzia online do symulacji elastyczności strukturalnej białek i modelowania peptydów

- 4.2.1. Wprowadzenie do [A6]
 - 4.2.2. Artykuł [A6]
- 5. Agregacja białek zależna od pH: Aggrescan4D
 - 5.1. [A7] Aggrescan4D: analiza zależnej od pH agregacji białek z uwzględnieniem informacji strukturalnej
 - 5.1.1. Wprowadzenie do [A7]
 - 5.1.2. Artykuł [A7]
 - 5.2. [A8] Aggrescan4D: kompleksowe narzędzie do analizy i projektowania podatności białek na agregację zależną od pH
 - 5.2.1. Wprowadzenie do [A8]
 - 5.2.2. Artykuł [A8]
- 6. Dynamika molekularna (MD) w biologii strukturalnej
 - 6.1. [A9] Fosforylacja białek P-stalk definiuje stan rybosomu umożliwiającą interakcję z pomocniczymi czynnikami białkowymi
 - 6.1.1. Wprowadzenie do [A9]
 - 6.1.2. Artykuł [A9]

CZĘŚĆ III: PODSUMOWANIE I WNIOSKI

- 7. Podsumowanie wyników i wnioski końcowe

Bibliografia

CZĘŚĆ I: PODSTAWY TEORETYCZNE I METODOLOGICZNE

1. Wprowadzenie: cel i zakres pracy

Biologia molekularna stoi dziś u progu rewolucji, której tempo i skala przypominają przełom komputerowy końca XX wieku. Zaledwie dekadę temu struktury białek były dobrem deficytowym, a ich modelowanie opierało się na mozolnym eksperymentowaniu i uproszczonych założeniach. Dziś, dzięki połączeniu sztucznej inteligencji, obliczeń wieloskalowych i automatyzacji, jesteśmy w stanie projektować, modyfikować i analizować molekularne procesy z precyzją, jakiej biologia nigdy wcześniej nie знаła.

W centrum tej zmiany są oddziaływania białko-białko i białko-peptyd - molekularne decyzje, które determinują wszystko: czy komórka przetrwa stres, czy wirus zostanie rozpoznany, czy lek dotrze do celu. To one odpowiadają za precyzję układu odpornościowego, działanie enzymów, kontrolę translacji i sygnalizację międzykomórkową. Zdolność do ich modelowania i modyfikowania to dziś klucz do rozwoju przełomowych terapii: peptydów cyklicznych o wysokiej selektywności, zoptymalizowanych przeciwciał, białek o zwiększonej rozpuszczalności czy nowych inhibitorów interakcji. Wkraczamy w erę, w której białka nie tylko badamy, ale potrafimy je konstruować niczym funkcjonalne maszyny - precyzyjne, efektywne i przewidywalne.

Celem niniejszej rozprawy doktorskiej było opracowanie i wdrożenie zaawansowanych metod modelowania oddziaływań molekularnych, które realistycznie odzwierciedlają złożoność biologicznych układów - z uwzględnieniem elastyczności strukturalnej, ograniczeń topologicznych, wpływu środowiska oraz modyfikacji potranslacyjnych.

Praca obejmuje rozwój narzędzi umożliwiających elastyczne dokowanie peptydów, analizę dynamiki białek, predykcję podatności na agregację, modelowanie wpływu środowiska biologicznego, integrację danych strukturalnych z modelami językowymi białek oraz symulacje mechanizmów regulowanych modyfikacjami potranslacyjnymi.

Wszystkie te obszary łączy wspólna ambicja: stworzyć narzędzia, które nie tylko opisują rzeczywistość molekularną, ale umożliwiają jej aktywne modelowanie, projektowanie i interpretację - z myślą o zastosowaniach w biologii strukturalnej, inżynierii białek i projektowaniu leków nowej generacji.

Praca ta powstała z przekonania, że współczesna biologia molekularna potrzebuje nie tylko lepszych predykcji - potrzebuje narzędzi inżynierskich. I właśnie taki zestaw narzędzi - oparty na wiedzy, kodzie i koncepcjach - starałem się stworzyć.

1.1. Oddziaływania białko-białko i białko-peptyd: znaczenie biologiczne i implikacje terapeutyczne

W tej części przedstawię zwięzły przegląd zagadnienia oraz uzasadnienie podjętych przeze mnie badań. W epoce, w której białka nie tylko opisujemy, lecz także projektujemy, kluczowe staje się precyzyjne zrozumienie interakcji białko-białko i białko-peptyd. To właśnie one przenoszą sygnały, uruchamiają enzymy, regulują transkrypcję i translację, budują struktury komórkowe i inicjują odpowiedź immunologiczną. Nie są efektem przypadkowych zdarzeń, lecz dynamicznymi, ściśle kontrolowanymi procesami decydującymi o przetrwaniu, proliferacji, adaptacji lub śmierci komórki - i stanowią praktyczny fundament nowoczesnych terapii.

Peptydy potrafią w tych układach działać jako ligandy, blokery, aktywatory i transportery. Na pierwszy plan wysuwają się szczególnie peptydy cykliczne: dzięki zamkniętej topologii są bardziej odporne na degradację, stabilniejsze konformacyjnie i mniej ruchliwe, co przekłada się na wyjątkowo trwałe i selektywne wiązanie celów.¹ Dlatego uznaje się je za obiecujących kandydatów na leki (temat rozwijam w dalszej części rozdziału).

Na skuteczność oddziaływań wpływa nie tylko sama sekwencja czy struktura, lecz także otoczenie: pH, siła jonowa i modyfikacje potranslacyjne. Niewielkie zmiany pH potrafią odwrócić ładunki powierzchniowe, zmienić stabilność i zdolność do tworzenia kompleksów; agregacja wynika ze splotu cech cząsteczek i czynników zewnętrznych. Dlatego modele muszą łączyć dane strukturalne i fizykochemiczne z dynamicznym kontekstem środowiskowym - inaczej nie oddadzą prawdziwej natury układu.

Same interakcje białko-białko obejmują rozległe, często kilkudziesięciorezultowe interfejsy i sieci oddziaływań hydrofobowych, elektrostatycznych oraz wodorowych. Wiele kontaktów bywa przejściowych i słabo stabilnych - zależnych od stężenia, środowiska i kofaktorów - i przyjmuje charakter allosteryczny, przenosząc sygnał przez całe szlaki. To podłoże tak różnych zjawisk jak aktywacja receptorów błonowych czy powstawanie kompleksów immunologicznych i enzymatycznych. Tu przewagę pokazują peptydy, zwłaszcza cykliczne: zwarta budowa utrudnia denaturację, sprzyja biodostępności i pozwala docierać do trudnych miejsc wiązania. Ich niewielki rozmiar i elastyczność ułatwiają „wejście” w kieszenie hydrofobowe, płaskie interfejsy i wnęki allosteryczne, dzięki czemu skutecznie hamują oddziaływania białko-białko (PPI) - tam, gdzie klasyczne małe cząsteczki lub przeciwciała często zawodzą.

Ta wiedza przekłada się bezpośrednio na terapię. Cele PPI są już wykorzystywane w onkologii (np. inhibitory MDM2-p53), w chorobach neurodegeneracyjnych (stabilizacja interakcji synaptycznych) i autoimmunologicznych (regulacja aktywacji limfocytów T); rozwiązania te są w badaniach klinicznych i dopuszczane do leczenia. Stąd potrzeba narzędzi, które przewidują nie tylko „czy” dojdzie do kontaktu, ale także „jak”: w jakiej geometrii, z jaką siłą i z jakim skutkiem funkcjonalnym.

1.2. Wyzwania modelowania struktur kompleksów białko-białko i białko-peptyd

Oddziaływania międzycząsteczkowe w układach biologicznych są z natury złożone - zarówno pod względem dynamiki, jak i form, jakie przyjmują. Wśród nich szczególną rolę odgrywają interakcje białko-białko (PPI) oraz białko-peptyd, których wiarygodne modelowanie musi łączyć opis strukturalny i konformacyjny z kontekstem środowiskowym. Mimo dużego postępu technologicznego, wciąż istnieją istotne ograniczenia metodologiczne i obliczeniowe.

Po pierwsze - zmienność konformacyjna peptydów i skutki cyklizacji. Wiele peptydów w roztworze współistnieje w licznych konformacjach, a ułożenie po związaniu często znacząco odbiega od formy niezwiązanej, co ogranicza przydatność metod opartych na sztywnych strukturach i wymusza elastyczne dokowanie.² Dodatkowym wyzwaniem jest cyklizacja: zwiększa stabilność i odporność na degradację, ale nakłada topologiczne ograniczenia trudne do odwzorowania w klasycznych algorytmach.

Po drugie - identyfikacja miejsc wiązania i trajektorii asocjacji. W przeciwieństwie do enzymów czy receptorów o wyraźnych kieszeniach, powierzchnie wiązania peptydów bywają płaskie, rozproszone lub allosteryczne, a ich właściwości zmieniają się w trakcie samego procesu wiązania. Predykcja musi więc uchwycić nie tylko pozycję peptydu, lecz także indukowaną dopasowalność i stany pośrednie prowadzące do stabilnego kompleksu.

Po trzecie - rozległość i dynamika interfejsów PPI. PPI obejmują duże, wielopunktowe powierzchnie kontaktu (oddziaływania hydrofobowe, elektrostatyczne, wodorowe, π -stackingowe), są często przejściowe i słabo stabilne - zależne od warunków komórkowych, kofaktorów i stężeń. Jedna struktura nie oddaje całego spektrum stanów konformacyjnych i funkcjonalnych, dlatego trzeba modelować nie tylko kontakt, lecz także jego plastyczność i allosteryczną propagację sygnału.³

Równolegle kluczowy jest kontekst środowiskowy. Parametry takie jak pH, siła jonowa czy modyfikacje potranslacyjne (np. fosforylacja) wpływają na ładunek powierzchniowy, lokalną stabilność i skłonność do tworzenia kompleksów. Pominięcie tych czynników prowadzi do uproszczeń i obniżonej trafności predykcji, co jest szczególnie problematyczne w projektowaniu terapeutycznym.

Na tym tle coraz częściej stosowane są metody oparte na sztucznej inteligencji, które znacząco poszerzyły możliwości predykcji struktur białek i ich oddziaływań. Przełomowe osiągnięcia, takie jak AlphaFold2, pokazały, że możliwe jest przewidywanie struktur pojedynczych łańcuchów białkowych z dokładnością zbliżoną do danych eksperymentalnych⁴. Jednak modelowanie układów złożonych - w tym białek wielodomenowych oraz kompleksów PPI - pozostaje znacznie trudniejsze, zwłaszcza w zakresie poprawnego odwzorowania wzajemnego ułożenia domen i geometrii interfejsów.⁵ Co istotne, predykcje generowane przez algorytmy AI mają charakter statyczny i w ograniczonym stopniu uwzględniają dynamikę układu, co

ogranicza ich zastosowanie do opisu procesów wiązania zachodzących w rzeczywistych warunkach biologicznych.

Pomimo znaczących sukcesów metod opartych na sztucznej inteligencji w predykcji struktur białek, ich zastosowanie do opisu oddziaływań białko-białko i białko-peptyd napotyka na istotne ograniczenia natury koncepcyjnej i metodologicznej. Ograniczenia te wynikają przede wszystkim ze statycznego charakteru predykcji oraz niepełnego odwzorowania procesów dynamicznych, które są kluczowe dla funkcji biologicznej białek.⁶⁻⁸

Po pierwsze, modele AI zazwyczaj zwracają pojedynczą, „najbardziej prawdopodobną” konformację, podczas gdy białka i peptydy funkcjonują jako zespoły wielu stanów konformacyjnych, reagując na zmiany środowiska, obecność ligandów i partnerów wiążących^{4,9}. Brak jawnego opisu stanów pośrednich utrudnia zrozumienie mechanizmów wiązania, regulacji allosterycznej oraz selektywności oddziaływań⁹.

Po drugie, obecne metody predykcyjne nie rekonstruują ścieżek zwijania białek ani procesów agregacji, mimo że zjawiska te odgrywają kluczową rolę w projektowaniu białek oraz w patogenezie wielu chorób związanych z nieprawidłowym fałdowaniem^{9,10}. Ograniczenie to jest szczególnie istotne w kontekście analizy stabilności i podatności struktur na zmiany środowiskowe.

Po trzecie, skuteczność modeli AI silnie zależy od dostępności bogatych informacji ewolucyjnych w postaci wielokrotnych dopasowań sekwencji (MSA)⁴. W przypadku białek *de novo*, regionów niskiej konserwatywności lub tzw. „ciemnych” obszarów proteomu, wiarygodność predykcji ulega istotnemu obniżeniu, co ogranicza uogólnialność tych metod¹¹.

Wreszcie, istotnym wyzwaniem pozostaje interpretowalność predykcji oraz ocena ich wiarygodności. Jak podkreślają Peng i wsp.⁵, nawet wysoka zgodność strukturalna z danymi referencyjnymi nie gwarantuje poprawnego odwzorowania mechanizmu działania, a brak przejrzystości decyzyjnej modeli utrudnia identyfikację regionów krytycznych funkcjonalnie^{12,13}.

Łącznie, ograniczenia te wskazują, że skuteczna predykcja interakcji białko-białko i białko-peptyd wymaga podejść wykraczających poza statyczne modele strukturalne. Najbardziej obiecujące wydają się strategie hybrydowe, łączące metody AI z symulacjami dynamiki molekularnej, modelami gruboziarnistymi oraz danymi eksperymentalnymi, co pozwala lepiej uchwycić zmienność konformacyjną, wpływ środowiska i dynamiczny charakter oddziaływań¹⁰.

1.3. Struktura pracy i wykaz publikacji

Niniejszą rozprawę oparłem na dziewięciu artykułach naukowych [A1-A9], których jestem współautorem. Publikacje te ukazały się w recenzowanych międzynarodowych czasopismach, w tym w czołowych periodykach z obszaru biologii strukturalnej, bioinformatyki i chemii obliczeniowej. Łącznie dotyczą one kluczowych wyzwań związanych z modelowaniem oddziaływań białek i peptydów, proponując jednocześnie nowe rozwiązania metodologiczne.

W rozdziale 2 przedstawiłem teoretyczne podstawy najważniejszych metod eksperymentalnych i obliczeniowych stosowanych w analizie struktury białek. Omówiłem klasyczne podejścia do dokowania, modele gruboziarniste, dynamikę molekularną oraz najnowsze metody oparte na sztucznej inteligencji, takie jak AlphaFold i ESMFold. Ten rozdział stanowi fundament dla dalszej części pracy.

W rozdziałach 3-6 zaprezentowałem pełne teksty moich artykułów, które pogrupowałem tematycznie:

- rozdział 3 - rozwój i zastosowanie metod dokowania białek i peptydów,
- rozdział 4 - modelowanie elastyczności strukturalnej przy użyciu CABS-flex,
- rozdział 5 - analiza i projektowanie podatności białek na agregację (Aggrescan4D),
- rozdział 6 - dynamika molekularna i jej zastosowania w kontekście biologii strukturalnej.

Każdy z tych rozdziałów poprzedziłem krótkim wprowadzeniem, w którym umieściłem publikację w szerszym kontekście badań i podkreśliłem jej unikalny wkład.

Poniżej przedstawiam wykaz publikacji składających się na rozprawę:

[A1] Kurcinski M, Kmiecik S, **Zalewski M**, Kolinski A. Protein-Protein Docking with Large-Scale Backbone Flexibility Using Coarse-Grained Monte-Carlo Simulations. *Int J Mol Sci*. 2021;22(14):7341.

[A2] **Zalewski M**, Badaczewska-Dawid AE, Kmiecik S. Flexible docking of cyclic peptides to proteins using CABS-dock. *J Chem Theory Comput*. 2025;21(17):8249-8254.

[A3] **Zalewski M**, Kmiecik S, Koliński M. Molecular Dynamics Scoring of Protein-Peptide Models Derived from Coarse-Grained Docking. *Molecules*. 2021;26(11):3293.

[A4] **Zalewski M**, Wallner B, Kmiecik S. Protein-peptide docking with ESMFold language model. *J Chem Theory Comput*. 2025;21(6):2817-2821.

[A5] Nithin C, Fornari RP, Pilla SP, Wroblewski K, **Zalewski M**, Madaj R, et al. Exploring protein functions from structural flexibility using CABS-flex modeling. *Protein Sci*. 2024;33(9):e5090.

[A6] Wróblewski K, **Zalewski M**, Kuriata A, Kmiecik S. CABS-flex 3.0: an online tool for simulating protein structural flexibility and peptide modeling. Nucleic Acids Res. 2024;gkaf412.

(K.W. i **M.Z.** - równorzędny wkład w pracę)

[A7] Bárcenas O, Kuriata A, **Zalewski M**, Iglesias V, Pintado-Grima C, Firlik G, et al. Aggrescan4D: structure-informed analysis of pH-dependent protein aggregation. Nucleic Acids Res. 2024;52(W1):W170-W175.

[A8] **Zalewski M**, Iglesias V, Bárcenas O, Ventura S, Kmiecik S. Aggrescan4D: a comprehensive tool for pH-dependent analysis and engineering of protein aggregation propensity. Protein Sci. 2024;33(10):e5180.

[A9] Filipek K, Blanchet S, Molestak E, Zaciura M, Wu CCC, **Zalewski M**, et al. Phosphorylation of P-stalk proteins defines the ribosomal state for interaction with auxiliary protein factors. EMBO Rep. 2025;25(12):5478-5506.

Poniżej przedstawiono syntetyczne zestawienie mojego wkładu w publikacje składające się na rozprawę. Szczegółowe informacje dotyczące zakresu prac znajdują się w pierwszych akapitach wprowadzeń do poszczególnych artykułów.

- [A1] - wkład techniczny i analityczny; implementacja oraz rozwój wybranych elementów oprogramowania do dokowania białko-białko, udział w analizie wyników oraz współudział w przygotowaniu manuskryptu.
- [A2] - **autor wiodący**; opracowanie koncepcji metody dokowania cyklicznych peptydów, implementacja algorytmu, przygotowanie i przeprowadzenie obliczeń oraz analiza i wizualizacja wyników, a także przygotowanie i redakcja manuskryptu.
- [A3] - **autor wiodący**; zaprojektowanie i przetestowanie podejścia opartego na krótkich symulacjach dynamiki molekularnej jako etapu selekcji modeli dokowania, przygotowanie i realizacja obliczeń oraz analiza i wizualizacja wyników, a także współudział w procesie redakcji manuskryptu.
- [A4] - **autor wiodący**; zaproponowanie i opracowanie koncepcji wykorzystania modelu językowego ESMFold w zadaniach dokowania peptydów, implementacja i testowanie procedury oraz przeprowadzenie analiz porównawczych, przygotowanie wizualizacji i interpretacja wyników, opracowanie pierwszej wersji manuskryptu oraz udział w jego dalszej redakcji.
- [A5] - wkład współautorski w pracy przeglądowej; współudział w przygotowaniu manuskryptu, w szczególności opracowanie kluczowego fragmentu dotyczącego integracji modelowania elastyczności strukturalnej z analizą skłonności białek do agregacji oraz przygotowanie wizualizacji.
- [A6] - **równorzędny wkład wiodący** (*equal contribution*); opracowanie i implementacja złożonego interfejsu webowego narzędzia CABS-flex 3.0 oraz udział w jego integracji, testach i udostępnieniu online, przygotowanie wizualizacji, a także udział w przygotowaniu manuskryptu.

- [A7] - wkład implementacyjny; zaprojektowanie i wdrożenie interfejsu webowego serwera Aggrescan4D oraz testowanie jego funkcjonalności i poprawności działania.
- [A8] - **autor wiodący**; koordynacja przygotowania publikacji oraz opracowanie kluczowych elementów narzędzia Aggrescan4D, obejmujących architekturę systemu, walidację metody oraz integrację z zewnętrznymi bazami danych i interfejsami API, przygotowanie wizualizacji, opracowanie zasadniczej części pierwszej wersji manuskryptu oraz udział w jego dalszej redakcji.
- [A9] - wkład obliczeniowo-analityczny (praca konsorcyjna); przygotowanie układów obliczeniowych, przeprowadzenie symulacji dynamiki molekularnej, analiza trajektorii, przygotowanie wizualizacji wyników oraz udział w analizie i interpretacji danych, a także współudział w tworzeniu pierwszej wersji manuskryptu oraz w jego dalszej redakcji.

2. Podstawy teoretyczne i metody obliczeniowe

W tej części rozprawy przedstawiam kluczowe koncepcje, techniki oraz założenia teoretyczne, które stanowią podstawę przeprowadzonych analiz. Skupiłem się zarówno na klasycznych metodach modelowania molekularnego, jak i na nowoczesnych podejściach wykorzystujących sztuczną inteligencję, symulacje dynamiki oraz bioinformatykę strukturalną.

Zależało mi nie tylko na przedstawieniu używanego aparatu metodologicznego, ale również na zarysowaniu jego ograniczeń - oraz na pokazaniu, w jaki sposób zidentyfikowane braki stały się punktem wyjścia do rozwoju i wdrożenia nowych podejść, które omówiłem w części eksperymentalnej.

2.1. Metody eksperymentalne i obliczeniowe w określaniu struktury białek

Struktura przestrzenna białka warunkuje jego funkcję, dlatego poznanie jej stało się jednym z fundamentów biologii molekularnej^{4,14,15}. Przez dekady to właśnie metody eksperymentalne stanowiły główne źródło wiedzy o trójwymiarowej architekturze białek. Wśród nich najważniejszą rolę odegrały: krystalografia rentgenowska, spektroskopia magnetycznego rezonansu jądrowego (NMR) oraz krioelektronowa mikroskopia pojedynczych cząsteczek (cryo-EM).

Każda z tych technik wnosi istotny wkład w badania strukturalne, jednocześnie niosąc ze sobą określone ograniczenia. Krystalografia rentgenowska zapewnia wysoką rozdzielczość i dużą precyzję modeli, ale wymaga uzyskania dobrze uformowanych kryształów - co często okazuje

się niemożliwe w przypadku białek błonowych, dużych kompleksów czy cząsteczek o wysokiej dynamice.¹⁶ NMR pozwala na badanie białek w warunkach zbliżonych do fizjologicznych, co umożliwia również wnioskowanie o ich lokalnej ruchliwości, ale jej zastosowanie ogranicza się do stosunkowo niewielkich układów.¹⁷ Z kolei cryo-EM świetnie sprawdza się w analizie dużych, wielopodjednostkowych kompleksów, choć jeszcze niedawno ustępowała rozdzielczością innym metodom.¹⁸

Postęp technologiczny sprawił, że eksperymentalne techniki strukturalne dostarczyły ogromnej liczby danych, które dziś zasilają bazy wiedzy takie jak PDB (Protein Data Bank)¹⁹. Te dane stanowią nie tylko fundament biologii strukturalnej, ale także niezbędny punkt odniesienia dla rozwoju narzędzi obliczeniowych, które zyskały szczególne znaczenie w erze bioinformatyki.

Metody obliczeniowe - zwłaszcza modele oparte na homologii, dopasowaniu wzorców (*threading*) oraz techniki *de novo* - przez lata rozwijano jako uzupełnienie lub alternatywę dla eksperymentów. Choć ich skuteczność była ograniczona i silnie zależała od jakości istniejących danych, już wtedy umożliwiały uzyskanie wstępnych modeli strukturalnych, przydatnych w dalszych analizach funkcjonalnych, ewolucyjnych czy chemicznych.

Wraz z pojawieniem się nowoczesnych rozwiązań opartych na sztucznej inteligencji - takich jak AlphaFold²⁴ - przewidywanie struktury białek wkroczyło na zupełnie nowy poziom. Niemniej jednak, nawet najbardziej zaawansowane algorytmy wymagają punktów odniesienia oraz weryfikacji. W tym sensie dane eksperymentalne pozostają nie tylko źródłem wiedzy, ale również standardem, względem którego ocenia się jakość modeli generowanych *in silico*.

2.2. Dokowanie molekularne

Dokowanie molekularne to jedna z kluczowych technik w obliczeniowej biologii strukturalnej, umożliwiająca przewidywanie, w jaki sposób dwie cząsteczki - najczęściej ligand (np. peptyd, mała cząsteczka) i receptor (np. białko) - mogą tworzyć stabilny kompleks. Celem tego procesu jest określenie najbardziej prawdopodobnego położenia i orientacji liganda względem powierzchni wiążącej receptora, z uwzględnieniem potencjalnych interakcji stabilizujących układ. Dokowanie znajduje zastosowanie m.in. w projektowaniu leków, analizie mechanizmów molekularnych oraz w badaniach nad selektywnością i specyficznością oddziaływań biologicznych.

Metody dokowania można klasyfikować na kilka sposobów, w zależności od zastosowanego formalizmu obliczeniowego oraz rodzaju oddziałujących cząsteczek. Z punktu widzenia aplikacyjnego wyróżnia się przede wszystkim dokowanie: (i) małocząsteczkowe (*small-molecule docking*), (ii) peptyd-białko oraz (iii) białko-białko (PPI), które różnią się skalą układu, stopniem elastyczności oraz charakterem interfejsów wiążących. Niniejsza sekcja koncentruje się głównie na metodach stosowanych w układach białko-peptyd oraz, w szerszym

kontekście, białko-białko, które stanowią szczególne wyzwanie obliczeniowe ze względu na dużą zmienność konformacyjną i rozległość powierzchni kontaktu.

2.2.1. Klasyczne podejścia do dokowania

Z perspektywy metodologicznej podejścia do dokowania można podzielić na klasyczne metody oparte na modelach fizykochemicznych oraz metody wykorzystujące sztuczną inteligencję, które zostaną omówione w sekcji 2.2.3. W ramach metod klasycznych wyróżnia się dwa podstawowe podejścia różniące się poziomem szczegółowości: dokowanie w pełnoatomowej rozdzielczości (*all-atom*) oraz dokowanie w rozdzielczości uproszczonej, gruboziarnistej (*coarse-grained*).

Metody pełnoatomowe wykorzystują modele w których uwzględniane są wszystkie atomy, ich geometria oraz interakcje międzyatomowe. Podejścia tego typu wykorzystują uproszczone modele fizykochemiczne oparte na klasycznej mechanice molekularnej i funkcjach energii oddziaływań (tzw. pola siłowe). Choć nie symulują one dynamiki czasowej układu, jak w przypadku klasycznej MD, to pozwalają na ocenę energetycznej preferencji różnych pozycji liganda względem miejsca wiązania. Narzędzia reprezentujące to podejście to m.in. AutoDock²⁰, RosettaDock²¹ oraz Glide²². Z kolei metody gruboziarniste upraszczają reprezentację cząsteczek, redukując liczbę stopni swobody, np. poprzez modelowanie grup atomów jako pojedynczych pseudo-atomów. Takie podejście pozwala na znaczne przyspieszenie obliczeń oraz efektywniejsze próbkowanie przestrzeni konformacyjnej, co ma szczególne znaczenie w przypadku układów dynamicznych i o dużej elastyczności. Choć *coarse-graining* wiąże się z utratą części szczegółów strukturalnych, często okazuje się wystarczający w kontekście wstępnego przewidywania interfejsów wiążących czy generowania hipotez strukturalnych. Przykładowe programy tego typu to HADDOCK²³ i ClusPro²⁴.

W ramach klasycznego dokowania wyróżnia się także tzw. podejścia półelastyczne, w których jedna z cząsteczek (najczęściej receptor) traktowana jest jako sztywna, podczas gdy druga może przyjmować ograniczone ruchy konformacyjne. Tego typu kompromis pozwala zachować względnie niskie koszty obliczeniowe, przy jednoczesnym uwzględnieniu podstawowych efektów elastyczności.

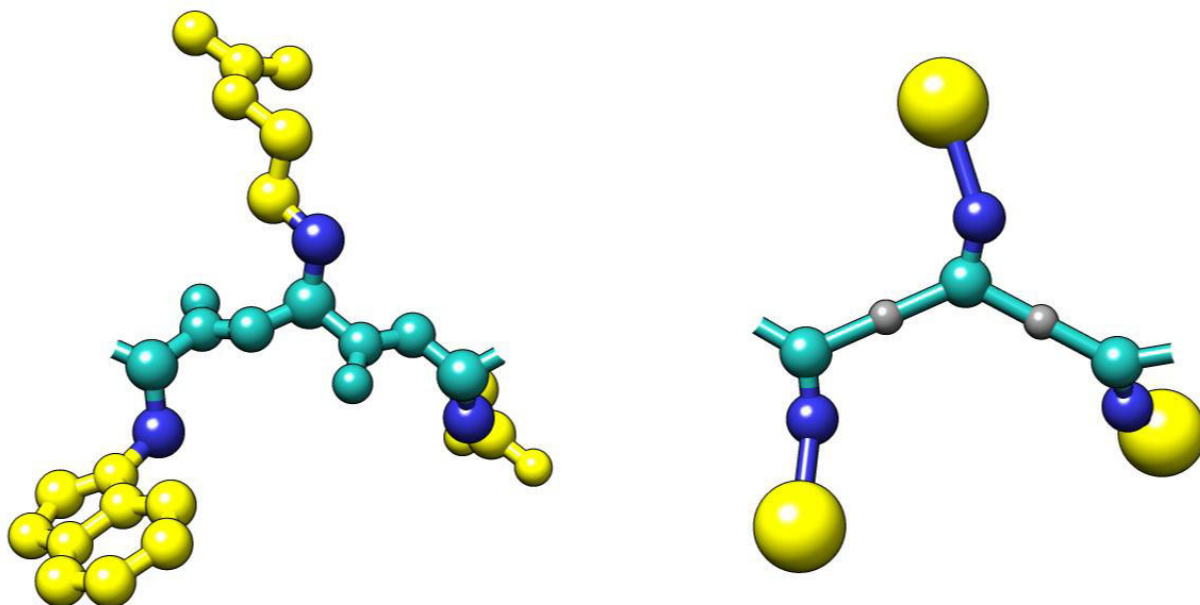
Niezależnie od poziomu szczegółowości, wszystkie klasyczne metody dokowania opierają się na dwóch kluczowych komponentach: generowaniu pozycji liganda (*sampling*) oraz ocenie ich jakości (*scoring*). *Sampling* może bazować na przeszukiwaniu przestrzeni konformacyjnej i translacyjnej przy użyciu heurystyk, dynamiki molekularnej, symulowanego wyżarzania lub algorytmów ewolucyjnych. Z kolei *scoring* polega na przypisaniu każdej pozycji wartości energetycznej lub heurystycznej funkcji dopasowania, co pozwala wyłonić najbardziej prawdopodobne konformacje.

Warto podkreślić, że mimo dużego rozwoju metod klasycznych, wciąż napotykają one na ograniczenia w kontekście elastyczności dużych cząsteczek, kontekstu środowiskowego oraz modelowania stanów pośrednich.

2.2.2. Metoda CABS-dock

CABS-dock²⁵⁻²⁷ jest metodą dokowania globalnego, która umożliwia przewidywanie struktur kompleksów białko-peptyd bez wcześniejszej wiedzy o miejscu wiązania. Jej siłą jest elastyczne podejście do reprezentacji cząsteczek, pozwalające na pełną ruchomość peptydu oraz ograniczoną, ale realistyczną elastyczność białka receptora. Podstawą tej metody jest tzw. model CABS²⁸ - gruboziarnista reprezentacja cząsteczki białkowej, w której każda reszta aminokwasowa opisana jest za pomocą czterech pseudo-atomów: atomu C α , C β , środka grupy bocznej oraz środka wiązania peptydowego (Rysunek 1). Taki poziom uproszczenia - zachowujący kluczowe cechy geometrii - pozwala znacząco obniżyć koszty obliczeniowe bez utraty jakości w analizie ogólnej topologii i dynamiki układu. Unikalną cechą metody CABS-dock jest możliwość jednoczesnego, globalnego przeszukiwania powierzchni białka oraz pełnej elastyczności peptydu, bez potrzeby wcześniejszego wskazania miejsca dokowania. Metoda ta pozostaje szeroko stosowana i wysoko ceniona w społeczności naukowej, co potwierdza jej popularność i liczne cytowania w literaturze przedmiotu.

Model CABS opiera się na potencjałach statystycznych wyprowadzonych z analizy struktur białek występujących w bazie danych struktur PDB. Uwzględniają one m.in. objętości wykluczone, kontekstowe oddziaływania łańcuchów bocznych, geometrię łańcucha głównego oraz uproszczony model wiązań wodorowych. Interakcje grup bocznych są zależne od otoczenia i wynikają z analizy częstości kontaktów w strukturach białkowych, a efekty środowiska (takie jak solwatacja) ujęte są w sposób pośredni poprzez dopasowanie parametrów kontaktów. Choć model nie nadaje się do symulacji subtelnych efektów pełnoatomowych, doskonale sprawdza się w odwzorowywaniu dużych przejść konformacyjnych, struktur globularnych oraz układów z regionami nieuporządkowanymi. Dzięki swojej konstrukcji model CABS umożliwia realistyczne odwzorowanie zarówno ruchliwości domen, jak i lokalnych zmian konformacyjnych. Rysunek 1 ilustruje porównanie reprezentacji reszty aminokwasowej w modelu CABS i w reprezentacji atomowej.



Rysunek 1. Reprezentacja pojedynczej reszty aminokwasowej w modelu atomowym (lewa strona) i modelu CABS (prawa strona). W reprezentacji CABS każda reszta aminokwasowa opisana jest za pomocą czterech pseudo-atomów: $C\alpha$ (turkusowy), $C\beta$ (granatowy), pseudo-atomu łańcucha bocznego (żółty) oraz pseudo-atomu reprezentującego środek wiązania peptydowego (szary). Źródło: [A1].

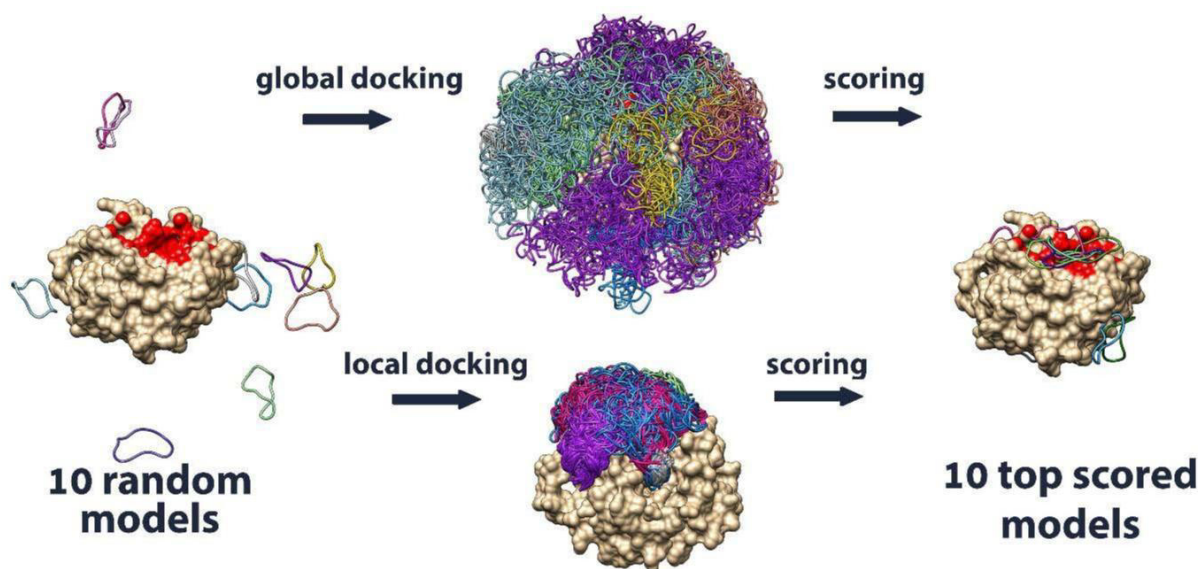
Eksploracja przestrzeni konformacyjnej w modelu CABS odbywa się przy użyciu metody *Replica Exchange Monte Carlo* (REMC), stanowiącej rozszerzenie klasycznego schematu *Monte Carlo* o mechanizm wymiany replik prowadzonych w różnych temperaturach. Polega to na losowym stosowaniu serii lokalnych przekształceń - zarówno pojedynczych zmian pozycji aminokwasów, jak i fragmentów łańcucha. Długie sekwencje takich kroków odwzorowują duże zmiany konformacyjne, pozwalając modelowi na realistyczne odtworzenie giętkości struktur.

Standardowy przebieg dokowania CABS-dock obejmuje cztery główne etapy:

1. Generowanie konformacji wejściowych peptydu.
2. Globalne dokowanie z pełną swobodą peptydu, z ograniczoną elastycznością receptora, przeprowadzane z wykorzystaniem REMC.
3. Analiza skupień (klasteryzacja) otrzymanych rozwiązań na podstawie podobieństwa strukturalnego i wybór reprezentantów najliczniejszych klastrów.
4. Odtworzenie struktury pełnoatomowej dla najlepiej ocenionych modeli.

CABS-dock jest dostępny zarówno jako aplikacja *standalone*²⁷, jak i jako serwer WWW²⁵. Dzięki temu może być wykorzystywany w szerokim zakresie zastosowań, od badań podstawowych po wstępne etapy projektowania leków.

CABS-dock umożliwia również stosowanie więzów odległości (*distance restraints*) - dodatkowych ograniczeń geometrycznych nakładanych na wybrane pary atomów lub reszt aminokwasowych. Mechanizm ten został pierwotnie wprowadzony w kontekście dokowania peptydów liniowych jako sposób na uwzględnianie wiedzy zewnętrznej podczas symulacji, m.in. w postaci danych eksperymentalnych lub hipotez funkcjonalnych²⁹. Więzy definiują docelowe zakresy odległości, a ich przestrzeganie jest wymuszane w trakcie symulacji za pomocą miękkich potencjałów energetycznych; w praktyce pozwala to uwzględniać wiedzę zewnętrzną (np. wskazówki z NMR, dane literaturowe czy hipotezy funkcjonalne). Ten sam mechanizm stanowi podstawę trybu dokowania lokalnego, który zaimplementowałem i opisałem w pracy [A2] - w tym wariancie więzy odległości ograniczają przestrzeń poszukiwań, wymuszając próbkowanie konformacji peptydu w zdefiniowanym obszarze powierzchni receptora oraz umożliwiając precyzyjne ukierunkowanie obliczeń w oparciu o dostępne informacje. Rysunek 2 ilustruje różnicę między trybem globalnym a lokalnym, ukazując rozkład położenia peptydu względem receptora.



Rysunek 2. Wizualizacja dokowania peptydu w trybie globalnym i lokalnym (z wykorzystaniem więzów odległości). Źródło: [A2].

CABS-dock stanowi unikalne połączenie zalet podejść *coarse-grained* z realistycznym odwzorowaniem dynamiki i konformacyjnej zmienności. Choć nie zastępuje on metod wykorzystujących rozdzielczość pełnoatomową, jego wszechstronność, szybkość i możliwość dokowania w pełni swobodnego peptydu bez znajomości miejsca wiązania sprawiają, że jest to narzędzie wyjątkowo użyteczne w badaniach strukturalnych białko-peptyd.

2.2.3. Metody przewidywania struktur oparte na AI

W ostatnich latach sztuczna inteligencja zrewolucjonizowała przewidywanie struktury białek⁶, a symbolem tej zmiany stała się seria modeli AlphaFold (AF)⁴. Początkowo zaprezentowany

w 2018 roku AlphaFold1 zdobył uznanie dzięki znacznemu postępowi względem wcześniejszych metod³⁰, jednak to dopiero AlphaFold2, zaprezentowany w 2020 roku przez DeepMind⁴, wyznaczył nowy standard w dziedzinie biologii strukturalnej.⁴

AlphaFold2 po raz pierwszy umożliwił przewidywanie trójwymiarowej struktury białek z dokładnością porównywalną do eksperymentów krystalograficznych - w niektórych przypadkach nawet przewyższającą jakość danych z NMR. Jego skuteczność potwierdziły wyniki konkursu CASP14³¹, gdzie model ten zdominował konkurencję, osiągając średnią RMSD poniżej 1 Å dla dużej części testowanych struktur. Tak spektakularne wyniki skłoniły wielu ekspertów do stwierdzenia, że jeden z głównych problemów biologii strukturalnej - tzw. problem przewidywania struktury białek - został *de facto* rozwiązany.

Architektura AF2 opiera się na innowacyjnym połączeniu transformera i sieci grafowych. W przeciwieństwie do klasycznych podejść, które opierały się na sztywnych modelach fizykochemicznych, AlphaFold przetwarza wielosekwencyjne ułiniowania (MSA) oraz pary reszt przy użyciu tzw. Evoformerów - bloków uwzględniających zarówno informacje ewolucyjne, jak i kontekst strukturalny. Następnie predykowane są odległości pomiędzy parami atomów (tzw. *distance maps*) i przekształcane w trójwymiarową geometrię białka. W trakcie treningu model korzystał z tysięcy eksperymentalnie rozwiązanych struktur, co pozwoliło mu uchwycić globalne zależności przestrzenne, trudne do wychwycenia klasycznymi metodami.

Na bazie tego sukcesu powstały kolejne rozszerzenia modelu AlphaFold. Model AlphaFold-Multimer³² został dostosowany do przewidywania kompleksów białkowych, a AlphaFold3¹⁰ (AF3) - zaprezentowany w 2024 roku - zintegrował przewidywanie wiązań z ligandami, kwasami nukleinowymi i modyfikacjami chemicznymi, stając się wielofunkcyjnym narzędziem do przewidywania kompleksów biologicznych. W tym podejściu całkowicie zrezygnowano z wykorzystania wielokrotnych dopasowań sekwencji (MSA) jako danych wejściowych, zastępując je wektorowymi reprezentacjami sekwencyjnymi i strukturalnymi (tzw. *embeddings*), kodującymi informacje ewolucyjne i kontekstowe w postaci ciągłych cech numerycznych. AF3 nie tylko przewiduje geometrię układu, ale także siły wiązań, uporządkowanie łańcuchów i rozkład niepewności.

Równolegle rozwijane były inne modele oparte na językowych architekturach deep learningowych, takie jak ESMFold - opracowany przez Meta AI³³. ESMFold całkowicie pomija etap MSA, zastępując go przetworzoną reprezentacją wektorową z dużego modelu językowego ESM2, wytrenowanego na dziesiątkach milionów sekwencji białkowych. To podejście pozwala uzyskiwać struktury w czasie liczącym w sekundach - znacznie szybciej niż AF2 - przy zachowaniu akceptowalnej dokładności, zwłaszcza dla globularnych białek.

Oba podejścia - AlphaFold i ESMFold - są wykorzystywane nie tylko do predykcji struktur pojedynczych białek, ale również jako narzędzia pomocnicze w takich zastosowaniach jak dokowanie, analiza interfejsów, projektowanie białek, identyfikacja miejsc wiązania, analiza stabilności czy symulacje dynamiki molekularnej. W szczególności AlphaFold2 stał się

podstawą nowych algorytmów dokowania - używa się go np. do generowania wysokiej jakości modeli receptorów w przypadku braku danych eksperymentalnych, które następnie są poddawane klasycznemu dokowaniu lub elastycznym symulacjom.

W kontekście dynamiki kompleksów, warto zauważyć, że choć AI dostarcza pojedyncze, statyczne struktury, to są one często wystarczającym punktem wyjścia do dalszych analiz konformacyjnych, np. z użyciem metody CABS-flex lub pełnoatomowej dynamiki molekularnej. Niemniej, brak możliwości bezpośredniego modelowania konformacyjnych przejść, stanów pośrednich czy dynamiki allosterycznej nadal stanowi ograniczenie obecnych narzędzi AI.³⁴⁻³⁸

Warto także zaznaczyć, że metody takie jak ESMFold wykazują większą odporność na problemy związane z brakiem homologów - czyli tzw. „ciemnych regionów proteomu” - gdzie AF2 często zawodzi z powodu słabego sygnału ewolucyjnego w MSA.^{33,39-41} Dzięki zastosowaniu dużych modeli językowych, ESMFold może generować rozsądne przewidywania nawet w przypadkach, gdzie AlphaFold pozostaje bezsilny.

Rozwój tych narzędzi nabiera tempa - kolejne wersje są stale udoskonalane, a ich integracja z innymi obszarami biologii strukturalnej i chemii obliczeniowej staje się coraz głębsza. Należy oczekiwać, że przyszłe generacje metod opartych na sztucznej inteligencji będą w stanie nie tylko przewidywać statyczne struktury białek, ale również opisywać ich pełne, dynamiczne krajobrazy konformacyjne oraz wiarygodnie szacować niepewność predykcji.

2.3. Dynamika białek

Choć struktura białka dostarcza kluczowej informacji o jego potencjalnej funkcji, to właśnie jego dynamika stanowi o rzeczywistym mechanizmie działania w warunkach biologicznych.⁴²⁻⁴⁴ Większość procesów molekularnych - takich jak wiązanie liganda, rozpoznanie partnera interakcji, kataliza enzymatyczna czy przekazywanie sygnału - zachodzi nie w statycznej geometrii, lecz w ciągłym ruchu i adaptacji strukturalnej.⁴⁵⁻⁴⁹ Białka zmieniają kształt, przemieszczają domeny względem siebie, a nawet przechodzą reorganizacje konformacyjne warunkujące aktywność lub jej brak.^{38,43,50-52} Jednocześnie bezpośrednie, eksperymentalne śledzenie tych procesów w pełnym zakresie czasowym i konformacyjnym pozostaje dużym wyzwaniem, co istotnie ogranicza możliwość kompleksowego opisu dynamiki białek wyłącznie na podstawie metod doświadczalnych.

Dlatego analiza funkcji białek wyłącznie na podstawie pojedynczych, statycznych modeli - choć przydatna - jest dalece niewystarczająca. Kluczowe okazuje się uwzględnienie dynamiki: zakresu dostępnych konformacji, częstości przejść pomiędzy nimi oraz ich energetycznego i funkcjonalnego znaczenia. Tego typu wgląd umożliwiają metody symulacyjne, które -

niezależnie od poziomu szczegółowości - pozwalają śledzić ruchy białek w czasie i analizować ich zachowanie w środowisku przypominającym warunki *in vivo*.

W kolejnych sekcjach przedstawione zostaną dwa podejścia do modelowania dynamiki białek: klasyczne, oparte na mechanice molekularnej symulacje dynamiki molekularnej (MD), oraz uproszczona metoda CABS-flex⁵³⁻⁵⁵, bazująca na modelu CABS oraz schemacie REMC. Obie strategie oferują odmienne perspektywy - od precyzyjnego śledzenia trajektorii atomów w czasie rzeczywistym, po szybkie i efektywne oszacowanie zakresu elastyczności struktur.

2.3.1. Dynamika molekularna (MD)

Klasyczna dynamika molekularna (MD) jest jedną z najczęściej stosowanych metod badania zachowania biomolekuł w czasie. Jej podstawą jest rozwiązywanie równań ruchu Newtona dla układów atomów, przy założeniu, że siły działające pomiędzy nimi wynikają z uprzednio zdefiniowanego pola siłowego (*force field*). W ramach takiego podejścia cząsteczki traktowane są jako zbiór kul połączonych sprężynami, a ich wzajemne oddziaływania opisuje się za pomocą funkcji potencjału - uwzględniających wiązania kowalencyjne, kąty, kąty dwuścienne oraz oddziaływania niewiążące, takie jak van der Waalsa czy elektrostatyczne.

Najczęściej wykorzystywane pola siłowe to AMBER⁵⁶, CHARMM⁵⁷, GROMOS⁵⁸ i OPLS⁵⁹ - różniące się sposobem parametryzacji i przeznaczeniem. Obliczenia MD wykonywane są zwykle w ustalonych warunkach termodynamicznych - izotermiczno-izobarycznych (NPT) lub izotermiczno-izochorycznych (NVT) - z wykorzystaniem odpowiednich algorytmów termostatów i barostatów (np. Nosé-Hoover, Berendsen, Parrinello-Rahman). Cały proces realizowany jest w ramach dyskretnego schematu czasowego, z typowym krokiem rzędu 1-2 fs, co pozwala uzyskać rozdzielczość atomową trajektorii. Jednak ze względu na koszt obliczeniowy, standardowe symulacje obejmują zazwyczaj zakres od nanosekund do kilku mikrosekund.

Dynamika molekularna pozwala nie tylko na analizę stabilności struktur, ale także na badanie szlaków fałdowania, wpływu mutacji, wiązania ligandów, przejść konformacyjnych, czy mechanizmów działania białek.⁶⁰⁻⁶⁴ Szczególnie istotne jest to w kontekście oddziaływań białko-białko i białko-peptyd, gdzie ruchliwość domen, regionów nieuporządkowanych czy elementów allosterycznych może mieć kluczowe znaczenie dla funkcji biologicznej.⁶⁵ W połączeniu z innymi technikami, takimi jak dokowanie molekularne, MD umożliwia ponadto analizę stabilności kompleksów i estymację ich energii wiązania.⁶⁶⁻⁷⁰

W ostatnich latach rozwój mocy obliczeniowej - zwłaszcza dzięki zastosowaniu procesorów GPU - oraz upowszechnienie przyspieszonych metod próbkowania, takich jak metadynamika^{57, 58}, rekonstrukcja trajektorii z danych eksperymentalnych czy symulacje steered MD, znacząco poszerzyły możliwości dynamiki molekularnej. Mimo to technika ta wciąż pozostaje wymagająca pod względem zasobów obliczeniowych i wiedzy eksperckiej,

niezbędnej do odpowiedniego przygotowania układów, przetwarzania danych oraz interpretacji uzyskanych wyników.

2.3.2. Symulacja elastyczności - CABS-flex

CABS-flex⁵³⁻⁵⁵ to obliczeniowo efektywna, a jednocześnie skuteczna metoda służąca do badania elastyczności białek, zbudowana na bazie opisanego wcześniej modelu CABS (rozdział 2.2.2). Jest szeroko wykorzystywana w analizach strukturalnych, stanowiąc popularną alternatywę dla klasycznych symulacji MD, szczególnie tam, gdzie kluczowe jest szybkie określenie dynamiki bez konieczności uruchamiania kosztownych obliczeniowo symulacji w reprezentacji pełnoatomowej.

Metoda CABS-flex stanowi udany kompromis między dokładnością a efektywnością obliczeniową, umożliwiając odtworzenie eksperymentalnych profili ruchliwości przy czasie obliczeń krótszym o kilka rzędów wielkości w porównaniu z symulacjami MD. Jej zastosowania obejmują zarówno analizę lokalnej ruchliwości, jak i badania dużych przejść konformacyjnych, dynamiki domen oraz identyfikację funkcjonalnie istotnych, elastycznych regionów białek.

Z biegiem lat powstały różne wersje metody: serwer WWW CABS-flex⁷¹, a następnie CABS-flex 2.0⁵³ oraz aplikacja *standalone*⁷².

2.4. Predykcja agregacji z uwzględnieniem struktury i elastyczności konformacyjnej

Jedną z najczęściej stosowanych metod przewidywania tendencji do agregacji białek są narzędzia analizujące sekwencję aminokwasową, takie jak TANGO⁷³, Zyggregator⁷⁴ czy Aggrescan⁷⁵. Bazują one na analizie ciągu aminokwasowego i pozwalają szybko wskazać tzw. „hotspoty” - krótkie fragmenty sekwencji o wysokim potencjale agregacyjnym. Ich zaletą jest szybkość i niskie wymagania obliczeniowe, jednak mają one ograniczoną skuteczność w przypadku białek globularnych, w których wiele potencjalnych regionów agregacyjnych jest osłoniętych w rdzeniu struktury.

Odpowiedzią na to ograniczenie było wprowadzenie predyktorów strukturalnych, które uwzględniają przestrzenne rozmieszczenie reszt aminokwasowych. Do najważniejszych narzędzi tej klasy należą SolubiS⁷⁶, CamSol⁷⁷ oraz Aggrescan3D (A3D)⁷⁸ - strukturalne rozszerzenie algorytmu Aggrescan. A3D łączy informacje o ekspozycji reszt i ich przestrzennej bliskości, co pozwala wykrywać tzw. STAPy (*Structural Aggregation Prone regions*) - regiony agregacyjne niedostrzegalne dla klasycznych metod sekwencyjnych. Kluczowym elementem predykcji jest tu parametr A3D-score, obliczany jako suma lokalnych wskaźników agregacji

dla wszystkich par kontaktujących się reszt, ważona ich względną dostępnością powierzchniową (RSA - *Relative Solvent Accessibility*). Wynik ten może być podany jako suma globalna dla całego białka lub jako rozkład lokalny - pozwalający na identyfikację fragmentów struktury o najwyższej tendencji do agregacji.

Wersja Aggrescan3D 2.0⁷⁸ wprowadziła szereg istotnych udoskonaleń: rozszerzenie zakresu obsługiwanych struktur do kompleksów multimerycznych, integrację z narzędziem FoldX⁷⁹ oraz tryb dynamiczny, w którym predykcja opiera się nie na jednej, statycznej strukturze, lecz na reprezentatywnym zbiorze konformacji wygenerowanych za pomocą metody CABS-flex. W ten sposób można uwzględnić wpływ elastyczności strukturalnej na odsłonięcie regionów agregacyjnych - co okazuje się kluczowe np. w analizie pętli czy domen luźno związanych.

CZEŚĆ II: CZĘŚĆ BADAWCZA ROZPRAWY

3. Rozwój nowych metod dokowania białek i peptydów

3.1. [A1] Dokowanie białko-białko z dużą elastycznością łańcucha głównego z wykorzystaniem gruboziarnistych symulacji Monte Carlo

3.1.1. Wprowadzenie do [A1]

W publikacji [A1] rozszerzono metodę CABS-dock, pierwotnie opracowaną do dokowania peptydów do białek^{25,26}, w celu umożliwienia modelowania interakcji białko-białko, obejmujących większe oraz bardziej złożone układy⁸⁰⁻⁸². Mój wkład w realizację tego projektu obejmował implementację oraz rozwój wybranych elementów oprogramowania wykorzystywanego w obliczeniach, a także udział w analizie uzyskanych wyników oraz współudział w przygotowaniu manuskryptu do publikacji.

Badania te miały na celu przezwyciężenie kluczowego ograniczenia tradycyjnych metod dokowania molekularnego białko-białko, jakim jest pominięcie lub jedynie minimalne uwzględnienie elastyczności struktur białkowych^{83,84}. Metody te dobrze sprawdzają się w prostych przypadkach, gdy zmiany konformacyjne są niewielkie, jednak w bardziej wymagających sytuacjach - gdy tworzenie kompleksu wymaga dużych przekształceń łańcucha głównego lub reorganizacji domen - ich skuteczność znacząco spada. Opracowany przez nas protokół, będący rozwinięciem metody CABS-dock w kierunku obsługi dwóch pełnych partnerów białkowych, został zaprojektowany tak, aby poszerzyć zakres możliwych do modelowania układów i umożliwić uwzględnienie istotnych zmian konformacyjnych zachodzących podczas procesu wiązania - w sytuacjach, w których klasyczne narzędzia okazują się niewystarczające.

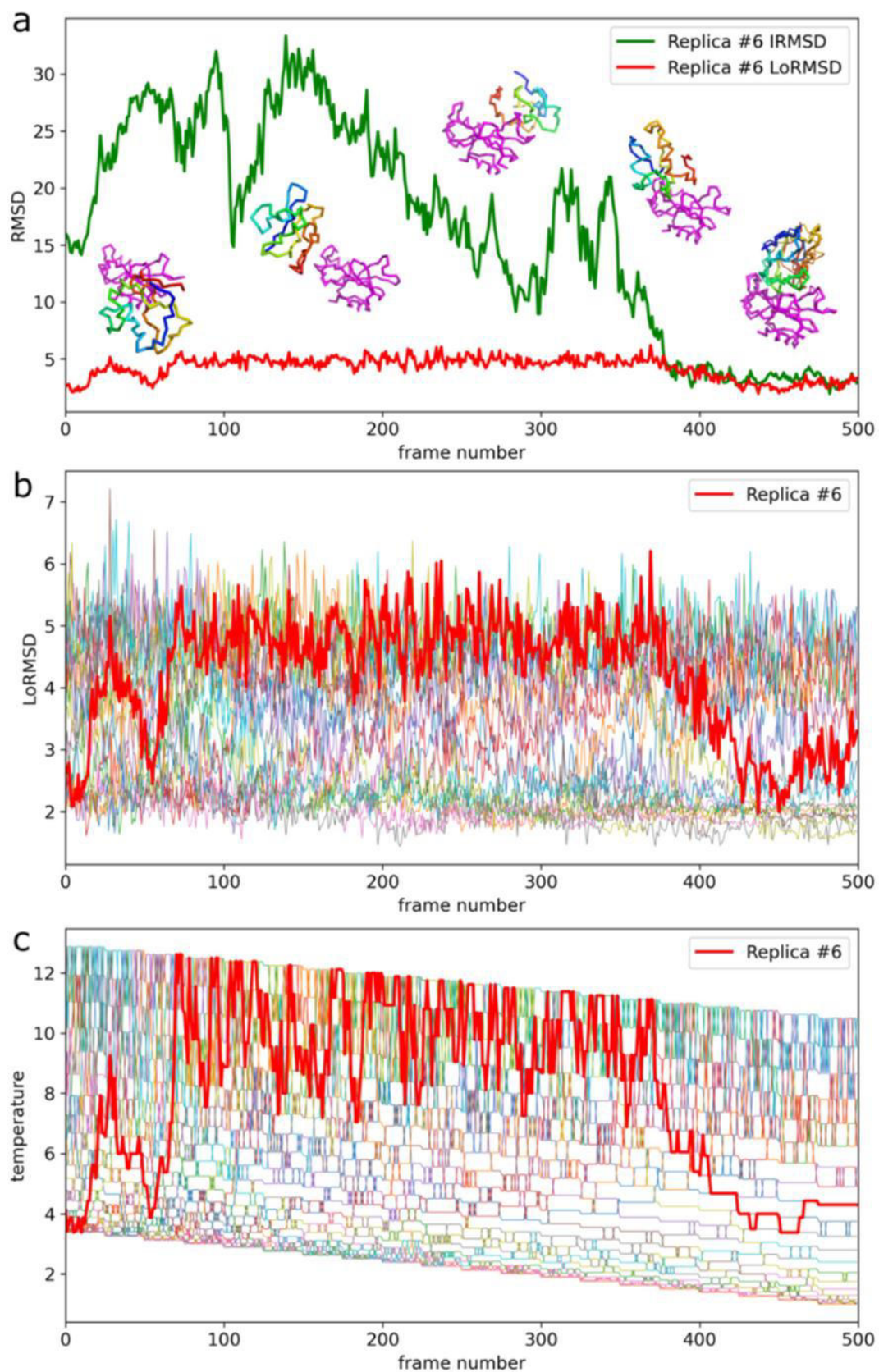
Zastosowaliśmy asymetryczną strategię dokowania, w której pełną swobodę konformacyjną przypisano wyłącznie partnerowi ruchomemu, przy jednoczesnym ograniczeniu elastyczności receptora. Takie uproszczenie znacząco zmniejszyło przestrzeń konformacyjną konieczną do przeszukania. Wybór ten był zainspirowany powszechnie obserwowanymi zjawiskami biologicznymi, w których partner o większej elastyczności lub uproszczonej architekturze domenowej dostosowuje swoją strukturę do stabilniejszej powierzchni receptora. Takie scenariusze często występują w kompleksach regulatorowych, sygnałowych i enzymatycznych.

Do realizacji tego podejścia wykorzystaliśmy model gruboziarnisty CABS, który omówiłem w rozdziale 2.2.2. Model ten pozwalał na zachowanie korzystnego kompromisu pomiędzy dokładnością strukturalną a wydajnością obliczeniową, umożliwiając realistyczne próbkowanie dużych przekształceń łańcucha głównego przy znacznie niższych kosztach

obliczeniowych niż metody pełnoatomowe. Umożliwiał on nie tylko swobodne próbkowanie pozycji i orientacji partnera ruchomego względem receptora, ale również jego znaczną elastyczność konformacyjną w ramach pojedynczej symulacji.

Integralnym elementem protokołu było zastosowanie schematu *Replica Exchange Monte Carlo* (REMC). Wykorzystanie wielu replik działających równolegle w zróżnicowanych warunkach temperaturowych pozwoliło na skuteczniejsze przeszukiwanie przestrzeni konformacyjnej i uniknięcie uwięzienia symulacji w lokalnych minimach energetycznych. Dzięki temu możliwe było efektywne próbkowanie szerokiego zakresu stanów konformacyjnych kompleksu, w tym takich, które są trudne do uchwycenia przy użyciu klasycznych metod dokowania ograniczających się do deterministycznego przeszukiwania orientacji i lokalnych, niewielkich zmian strukturalnych.

Przykładowy przebieg jednej z takich symulacji przedstawiony został na Rysunku 3. Widać na nim, że metoda pozwala stopniowo przejść od przypadkowych pozycji początkowych do uformowania wiarygodnego kompleksu, a zastosowanie wymiany replik skutecznie zwiększa różnorodność próbkowanych konformacji. Przykład barnase/barstar dobrze pokazuje, że protokół jest w stanie uchwycić duże zmiany konformacyjne i doprowadzić do rozwiązania o dokładności porównywalnej z eksperymentem.



Rysunek 3. Przykład działania metody opracowanej w [A1] na układzie barnase/barstar. Na wykresach przedstawiono przebieg dokowania globalnego prowadzonego w schemacie *Replica Exchange Monte Carlo* (REMC). Panel (a) pokazuje zmiany wartości RMSD interfejsu

(*interface root-mean-square deviation*, iRMSD) i RMSD liganda (*ligand root-mean-square deviation*, LoRMSD) w wybranej replice wraz z kilkoma etapami symulacji - od losowych pozycji początkowych po końcowy model (cienkie linie) nałożony na strukturę krystaliczną (grube linie). Panel (b) przedstawia LoRMSD we wszystkich replikach, co ilustruje różnorodność próbkowanych stanów, a panel (c) historię wymian między replikami, umożliwiającą wyjście z lokalnych minimów energetycznych. Ostateczny model osiągnął iRMSD = 1,9 Å względem struktury eksperymentalnej, co pokazuje zdolność protokołu do odtworzenia poprawnej orientacji kompleksu przy jednoczesnym uwzględnieniu dużych przekształceń łańcucha głównego. Źródło: [A1].

Do walidacji skuteczności opracowanego podejścia wykorzystaliśmy zestaw 62 kompleksów białkowych pochodzących ze zbioru testowego ZDOCK⁸⁵, stanowiącego powszechny punkt odniesienia przy ocenie metod dokowania. Zależało nam na sprawdzeniu skuteczności protokołu w szerokim spektrum przypadków, dlatego wybraliśmy kompleksy o zróżnicowanym poziomie trudności - od przypadków niskoelastycznych po scenariusze obejmujące wyraźne przekształcenia strukturalne. Wszystkie przewidywania przeprowadziliśmy na ślepo (ang. *blind docking*), bez wykorzystania informacji o miejscach wiązania czy znanych kontaktach, co pozwoliło na obiektywną ocenę efektywności próbkowania.

Dodatkową wartością naszej pracy było zastosowanie alternatywnego podejścia do oceny jakości modeli, opartego na porównaniu wzorców kontaktów pomiędzy resztami aminokwasowymi zamiast klasycznego porównania opartego na RMSD stosowanego w wyjściowej metodzie CABS-dock. Zamiast skupiać się na globalnym podobieństwie orientacji kompleksów, ocenialiśmy bezpośrednio podobieństwo interfejsu oddziaływania, analizując mapy kontaktów pomiędzy resztami tworzącymi powierzchnię wiązania. Takie podejście umożliwiało bardziej funkcjonalną ocenę poprawności modeli, szczególnie w przypadkach, gdzie zmiany orientacji całego kompleksu mogłyby maskować poprawne odwzorowanie kluczowych interakcji pomiędzy partnerami. Zastosowanie metody opartej na mapach kontaktów stanowiło istotne usprawnienie procesu selekcji modeli, pozwalając na trafniejszą identyfikację poprawnych sposobów wiązania, zwłaszcza w układach o wysokim stopniu elastyczności.

Wyniki naszych analiz pokazały, że zaproponowany schemat skutecznie radził sobie w przypadkach o niskiej i średniej elastyczności, generując modele spełniające kryteria jakości według protokołu CAPRI⁸⁶ (iRMSD poniżej 4 Å). Najważniejsze jest to, że taką dokładność osiągnęliśmy w ramach globalnego dokowania przy jednoczesnym uwzględnieniu dużej swobody konformacyjnej ruchomego partnera - czego inne zaawansowane podejścia zazwyczaj nie realizują w fazie globalnego wyszukiwania, ograniczając się na tym etapie do struktur sztywnych i wprowadzając elastyczność dopiero później, na etapie lokalnej rafinacji. Nasze rozwiązanie łączy obie te fazy w jednej symulacji, co pozwala uchwycić stany kompleksu niedostępne w tradycyjnych algorytmach.

W bardziej wymagających przypadkach, obejmujących rozległe zmiany strukturalne, zaobserwowaliśmy ograniczenia skuteczności, prawdopodobnie wynikające z zastosowania jednolitego schematu więzów odległości stabilizujących strukturę receptora. Wskazuje to na potrzebę dalszego rozwoju metody - zwłaszcza wprowadzenia lokalnie dostosowanych więzów oraz udoskonalenia procedur selekcji modeli. Uważam jednak, że już na obecnym etapie metoda ta wnosi istotną wartość: umożliwia realistyczne modelowanie wiązań białko-białko w przypadkach niedostępnych dla klasycznych narzędzi i otwiera drogę do badania trudnych układów biologicznych, generowania nowych hipotez strukturalnych oraz wspierania eksperymentów tam, gdzie dane strukturalne są ograniczone.

3.1.2. Artykuł [A1]



Article

Protein–Protein Docking with Large-Scale Backbone Flexibility Using Coarse-Grained Monte-Carlo Simulations

Mateusz Kurcinski *, Sebastian Kmiecik * , Mateusz Zalewski and Andrzej Kolinski

Biological and Chemical Research Centre, Faculty of Chemistry, University of Warsaw, 02-089 Warsaw, Poland; mateusz.zalewski.fuw@gmail.com (M.Z.); kolinski@chem.uw.edu.pl (A.K.)

* Correspondence: mkurc@chem.uw.edu.pl (M.K.); sekmi@chem.uw.edu.pl (S.K.)



Citation: Kurcinski, M.; Kmiecik, S.; Zalewski, M.; Kolinski, A. Protein–Protein Docking with Large-Scale Backbone Flexibility Using Coarse-Grained Monte-Carlo Simulations. *Int. J. Mol. Sci.* **2021**, *22*, 7341. <https://doi.org/10.3390/ijms22147341>

Academic Editors: Istvan Simon and Csaba Magyar

Received: 19 June 2021

Accepted: 4 July 2021

Published: 8 July 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Correction Statement: This article has been republished with a minor change. The change does not affect the scientific content of the article and further details are available within the back-matter of the website version of this article.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: Most of the protein–protein docking methods treat proteins as almost rigid objects. Only the side-chains flexibility is usually taken into account. The few approaches enabling docking with a flexible backbone typically work in two steps, in which the search for protein–protein orientations and structure flexibility are simulated separately. In this work, we propose a new straightforward approach for docking sampling. It consists of a single simulation step during which a protein undergoes large-scale backbone rearrangements, rotations, and translations. Simultaneously, the other protein exhibits small backbone fluctuations. Such extensive sampling was possible using the CABS coarse-grained protein model and Replica Exchange Monte Carlo dynamics at a reasonable computational cost. In our proof-of-concept simulations of 62 protein–protein complexes, we obtained acceptable quality models for a significant number of cases.

Keywords: protein–protein interactions; protein–protein binding; protein–protein complex; coarse-grained modeling; multiscale modeling

1. Introduction

Protein–protein interactions are fundamental in many biological processes. Their structural characterization is one of the biggest challenges of computational biology. A variety of docking methods are currently available for structure prediction of protein–protein complexes [1,2]. They can be divided into free (global) and template-based docking. Free (global) docking methods are designed to generate many distinct binding configurations. Template-based methods restrict docking to a binding mode found in a structural template. As demonstrated in the blind docking challenge, Critical Assessment of Prediction of Interactions (CAPRI), template-based methods generate more accurate results but only if a good quality template exists [1–5]. In some cases lacking useful templates, free global docking can yield acceptable results. According to recent estimates, the best free docking methods find adequate models among the top 10 predictions for around 40% of the targets [1]. The CAPRI analysis also indicates that protein backbone flexibility is a big challenge; protein complexes that undergo substantial conformational changes upon docking get no successful predictions from any method [3–5].

Presently, most of the free docking methods treat the backbone of input protein structures as rigid. This approximation reduces the protein–protein docking problem to a 6D (three rotational and three translational degrees of freedom) search space. Rigid-body search for the binding site most often rely on the Fast Fourier Transform [6–8]. Other successful approaches include 3D Zernike descriptor-based docking [9,10] or geometric hashing [11]. These rigid-body methods are often used as a first docking step, followed by scoring [12–16], using experimental data [17] and/or structural refinement to capture backbone flexibility [5,18]. Molecular Dynamics is perhaps the most common refinement strategy, either in classic or enhanced sampling versions [17,19–22]. Other tools use rotamer libraries to address side-chain flexibility [23] and Elastic Network Models (ENMs) for modeling backbone rearrangements [24–28]. Accounting for backbone flexibility in

the search for the binding site significantly increases the docking complexity and makes it practically intractable using conventional all-atom modeling approaches. This enormous computational complexity of flexible docking can be reduced using coarse-grained protein models [29–32]. The best-performing methods that can now include backbone flexibility during the docking calculations use coarse-grained models and/or ENM-driven simulations. These include RosettaDock combining coarse-grained generation of backbone ensembles and all-atom refinement [33–35]; ATTRACT combining coarse-grained docking with ENM and all-atom refinement [36,37]; and SwarmDock using all-atom ENM [25,38]. All these approaches show some advantages in modeling protein flexibility compared to rigid-body docking followed by structure refinements. However, effective modeling flexibility in protein–protein docking remains an unsolved problem, as demonstrated in the recent CAPRI round [25,35,37,39].

In this work, we use a well-established CABS coarse-grained protein model [29] for protein–protein docking. During the CABS docking simulation, one of the docking partners undergoes a long random process of rotations, translations, and extensive backbone conformational rearrangements that significantly modify its fold. Simultaneously, the backbone of the second protein undergoes small fluctuations.

2. Results

The most accurate models (out of the sets of 10,000 generated models and 10 top-scored) are characterized in Table 1. The table presents different metrics of similarity to the experimental structures for the set of 62 protein–peptide complexes (divided into three categories: low, medium and high flexibility cases). To assess the sampling performance, below we will use the iRMSD values for the best models out of all models. According to the iRMSD values the CABS-based docking algorithm produced a significant number of near-native protein–protein arrangements of acceptable quality ($iRMSD < 4 \text{ \AA}$, according to CAPRI criteria) for most protein–protein cases in the categories of low and medium flexibility cases. However, in the high-flexibility category, the best iRMSD values were noticeably higher (in the range of 4–12 Angstroms). This resulted from the adopted distance restraint scheme (see the Methods section), which was uniform for whole proteins and introduced a penalty for deviations of more than $1 \text{ \AA}$ from the input structures (unbound experimental structures). This penalty was very small for the protein ligands. Thus, the distance-restraints scheme allowed for the large-scale conformational changes, however, they might have prevented binding-induced conformational changes in the high-flexibility category. Therefore, there is the need to modify such a scheme for the most challenging targets.

Table 1. Summary of the docking simulations. The table characterizes X-ray data used in the docking, average ligand flexibility, and docking results. The table reports the best accuracy models out from all (10,000) and 10 top-scored models. The metrics definitions are provided in the Methods section. The table divides the presented cases on the three categories: low-flexibility, medium-flexibility and highly flexible cases.

X-ray Data (Number of Residues)				Ligand Flexibility	Results—Best from All Models			Results—Best from 10 Top-Scored Models		
Receptor *	Ligand *	Complex	RMSD **	Average LoRMSD	iRMSD	LRMSD	fNAT	iRMSD	LRMSD	fNAT
Low-flexibility cases										
5CHA (238)	2OVO (53)	1CHO	0.62	4.84	2.65	6.95	0.48	2.96	10.93	0.18
2PKA (232)	6PTI (56)	2KAI	0.91	4.64	3.32	11.34	0.19	4.75	15.76	0.12
1CHG (245)	1HPT (56)	1CGI	1.53	5.24	2.76	4.13	0.37	6.18	14.15	0.09
2PTN (223)	6PTI (58)	2PTC	0.31	5.23	2.97	11.86	0.29	4.39	15.93	0.15

Table 1. Cont.

X-ray Data (Number of Residues)			Ligand Flexibility		Results—Best from All Models			Results—Best from 10 Top-Scored Models		
Receptor *	Ligand *	Complex	RMSD **	Average LoRMSD	iRMSD	LRMSD	fNAT	iRMSD	LRMSD	fNAT
1SUP (275)	2CI2 (64)	2SNI	0.37	3.89	1.09	3.86	0.69	2.81	9.09	0.46
2ACE (532)	1FSC (61)	1FSS	0.76	4.48	3.41	7.20	0.25	15.03	32.56	0.03
1MAA (533)	1FSC (61)	1MAH	0.60	4.58	2.49	3.89	0.45	11.25	24.43	0.06
1A2P (108)	1A19 (89)	1BRS	0.47	3.33	1.94	4.19	0.64	4.01	8.74	0.14
1CCP (294)	1YCC (103)	2PCC	0.39	4.18	3.13	10.19	0.25	11.89	26.68	0.08
1SUP (275)	3SSI (107)	2SIC	0.39	4.01	4.03	18.96	0.23	4.77	19.40	0.12
1VFA (223)	1LZA (129)	1VFB	0.59	3.72	4.61	15.07	0.11	17.45	37.15	0.00
1MLB (432)	1LZA (129)	1MLC	0.85	3.74	2.82	10.47	0.36	8.04	33.09	0.04
Medium-flexibility cases										
1CHG (226)	1HPT (56)	1CGI	2.02	5.80	2.46	3.21	0.44	5.86	10.72	0.12
5C2B (241)	4ZAI (80)	5CBA	1.49	4.51	2.48	7.64	0.42	9.34	16.01	0.10
5P2 (166)	1LXD (87)	1LFD	1.79	4.12	2.87	6.76	0.27	12.47	24.24	0.00
1R6C (142)	2W9R (97)	1R6Q	1.67	9.27	7.95	11.97	0.14	13.71	35.71	0.00
1JXQ (242)	2OPY (106)	1NW9	1.97	4.09	7.05	8.69	0.23	9.33	17.55	0.00
1IAS (330)	1D6O (107)	1B6C	1.96	4.65	4.72	10.74	0.14	12.24	23.99	0.00
5E56 (116)	5E03 (113)	5E5M	1.56	4.16	3.83	9.09	0.23	10.96	20.00	0.00
2HRA (180)	2HQT (115)	2HRK	2.03	7.27	3.55	10.17	0.26	10.81	32.54	0.00
4BLM (256)	4M3J (116)	4M3K	1.77	4.41	4.96	7.41	0.10	13.75	27.11	0.03
1E78 (578)	5VNV (120)	5VNW	1.49	3.81	5.93	22.23	0.10	23.89	70.83	0.00
3BX8 (167)	3OSK (121)	3BX7	1.63	4.63	4.94	17.46	0.28	6.22	20.32	0.12
6ETL (124)	4POY (121)	4POU	1.83	4.01	2.91	10.16	0.50	6.55	19.65	0.25
4FUD (246)	5HDO (126)	5HGG	0.84	4.22	3.59	12.52	0.19	13.00	29.3	0.00
3TGR (346)	3R0M (127)	3RJQ	0.79	4.00	5.32	16.98	0.13	12.77	33.94	0.00
6EY5 (585)	5FWO (129)	6EY6	1.90	3.86	3.83	6.03	0.14	12.89	27.61	0.00
1SZ7 (159)	2BJN (141)	2CFH	1.55	5.13	1.98	4.01	0.71	2.82	5.50	0.63
3V6F (437)	3KXS (142)	3V6Z	1.83	7.11	6.12	16.68	0.15	6.66	20.06	0.06

Table 1. Cont.

X-ray Data (Number of Residues)			Ligand Flexibility		Results—Best from All Models			Results—Best from 10 Top-Scored Models		
Receptor *	Ligand *	Complex	RMSD **	Average LoRMSD	iRMSD	LRMSD	fNAT	iRMSD	LRMSD	fNAT
3CPI (437)	1G16 (156)	3CPH	2.12	4.34	4.87	15.64	0.09	15.02	27.88	0.00
1QJB (460)	1KUY (166)	1IB1	2.09	4.22	6.56	14.83	0.13	16.10	46.26	0.00
1IAM (185)	1MQ9 (173)	1MQ8	1.76	4.22	4.93	14.99	0.21	26.17	70.50	0.00
3HI5 (430)	1MJN (179)	3HI6	1.65	3.77	5.79	23.30	0.21	19.38	49.77	0.00
2G75 (429)	2GHV (183)	2DD8	2.19	5.37	5.73	13.78	0.09	17.20	34.33	0.00
1A12 (401)	1QG4 (202)	1I2M	2.12	4.19	2.84	6.43	0.51	3.58	6.97	0.47
1NOV (825)	1XK9 (204)	1ZM4	2.11	3.54	8.82	28.17	0.04	11.05	48.14	0.00
4EBQ (429)	4E9O (230)	4ETQ	0.47	3.72	7.12	14.74	0.20	8.68	19.61	0.07
1S3X (380)	1XQR (259)	1XQS	1.77	5.44	5.63	26.14	0.11	15.88	30.51	0.00
3HEC (329)	3FYK (282)	2OZA	1.89	4.29	4.35	9.32	0.33	11.24	18.8	0.03
6A0X (437)	2FK0 (322)	6A0Z	1.28	5.75	5.75	25.59	0.16	11.43	31.39	0.00
Highly flexible cases										
1CL0 (316)	2TIR (108)	1F6M	4.9	3.83	7.02	11.34	0.10	11.92	18.06	0.00
1 × 9Y (346)	1NYC (110)	1PXV	2.63	4.86	5.74	14.10	0.07	7.46	16.31	0.02
1JZO (431)	1JPE (116)	1JZD	2.71	4.65	4.98	8.13	0.28	13.38	34.05	0.00
5D7S (423)	2GMF (121)	5C7X	2.26	4.17	4.12	13.61	0.34	4.69	16.74	0.20
1FCH (302)	1C44 (123)	2C0L	2.62	5.51	5.02	5.54	0.21	10.24	24.74	0.00
1YWH (268)	2I9A (123)	2I9B	3.79	7.14	5.79	17.59	0.14	6.92	33.23	0.05
3L88 (550)	1CKL (126)	3L89	2.51	9.86	4.83	10.90	0.17	17.84	31.87	0.00
1ZM8 (239)	1J57 (143)	2O3B	3.13	6.20	4.76	16.43	0.18	15.34	31.95	0.00
1G0Y (310)	1ILR (145)	1IRA	8.38	4.07	12.97	22.24	0.08	15.86	25.46	0.05
1QUP (219)	2JCW (153)	1JK9	2.51	9.40	8.07	13.85	0.10	17.41	30.74	0.00
1SYQ (259)	3MYI (163)	1RKE	4.25	4.15	5.26	6.43	0.38	16.11	34.67	0.00
2II0 (463)	1CTQ (166)	1BKD	2.86	4.51	4.80	7.33	0.14	19.96	39.32	0.00
1ERN (416)	1BUY (166)	1EER	2.44	5.22	12.97	13.18	0.02	17.12	30.73	0.00
3AVE (419)	1FNL (173)	1E4K	2.60	5.32	3.44	10.07	0.43	7.59	24.33	0.13

Table 1. Cont.

X-ray Data (Number of Residues)			Ligand Flexibility		Results—Best from All Models			Results—Best from 10 Top-Scored Models		
Receptor *	Ligand *	Complex	RMSD **	Average LoRMSD	iRMSD	LRMSD	fNAT	iRMSD	LRMSD	fNAT
1R8M (195)	1HUR (180)	1R8S	3.73	5.50	6.67	13.41	0.09	15.15	25.10	0.00
1QFK (348)	1TFH (182)	1FAK	6.18	5.64	8.97	15.57	0.16	15.59	34.46	0.00
1F59 (440)	1QG4 (202)	1IBR	2.54	5.01	6.65	14.36	0.14	16.41	33.07	0.00
4DVB (427)	4DVA (246)	4DW2	2.27	3.85	6.61	21.91	0.14	9.94	29.27	0.00
1NG1 (294)	2IYL (271)	2J7P	2.67	4.51	8.87	18.77	0.11	18.46	48.05	0.00
1UX5 (411)	2FXU (360)	1Y64	4.69	4.15	6.42	13.50	0.27	15.50	36.42	0.00
1D0N (729)	1IJJ (371)	1H1V	6.62	3.44	7.92	31.14	0.36	29.12	65.07	0.03

* 4-letter PDB code for the crystal structures used in this study. ** The RMSD (in Å) of the interface C α atoms for input receptor and ligand after superposition onto the co-crystallized complex system.

The results analysis below focuses on the sampling performance for the selected low-flexibility barnase/barstar case. Figure 1 characterizes iRMSD versus CABS model energy values for the barnase/barstar (1BRS) and another low-flexibility case with clearly the lowest iRMSD value 1.09 Angstroms (2SNI).

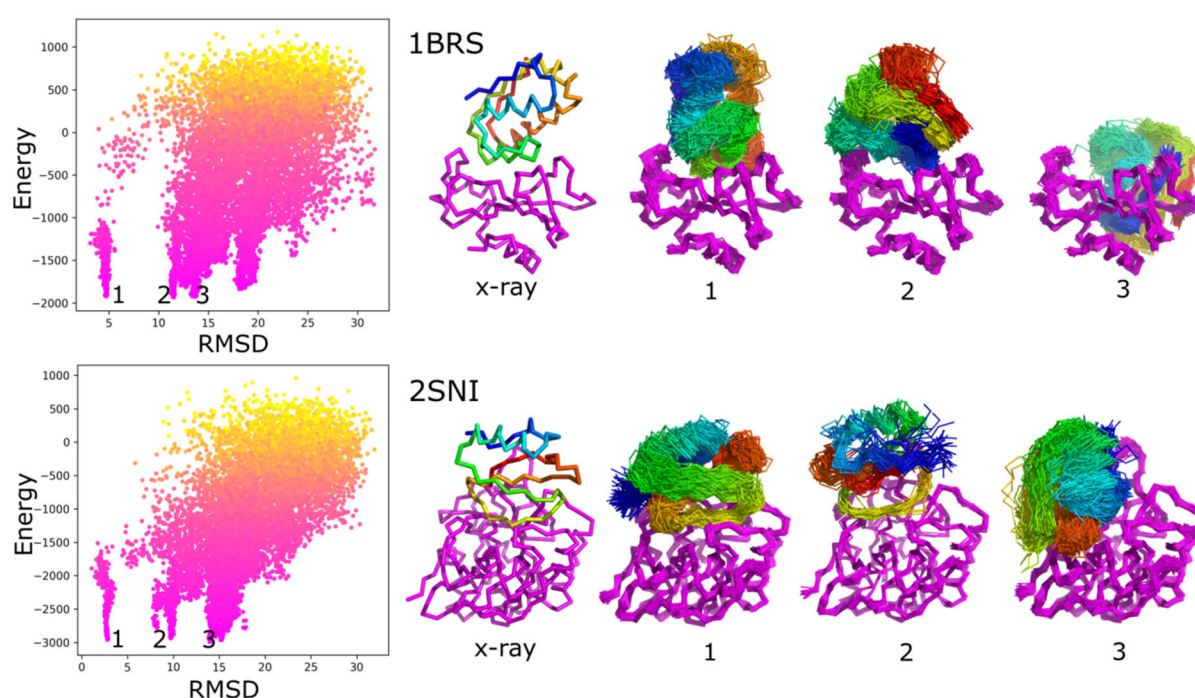


Figure 1. Characterization of docking results using RMSD to the X-ray structure and system energy. The left panels show the interface-RMSD versus CABS energy values. Point color represents the temperature—from yellow (**high**) to pink (**low**). The molecular visualizations show X-ray structures and ensembles of predicted models corresponding to selected energy minima (numbered in the picture from 1 to 3). As presented in the picture, the minima numbered as 1st corresponds to near-native protein–protein arrangements, others to non-native ensembles, as presented in the picture. The presented ensembles are the sets of similar models found in the structural clustering of contact maps (see Methods). The figure shows two modeling cases: 1BRS and 2SNI.

Figure 2 shows example ensembles of barnase/barstar models and the most accurate model (iRMSD 1.9 Å). A single system replica could explore an ample conformational space that involved significantly different binding configurations and protein-ligand conformations, as demonstrated in Figure 2c and Movie S1. Figure 3a further characterizes this single replica's using iRMSD and LoRMSD (RMSD for ligand only) values. As presented in the figure, the ligand structure fluctuated around 5 Å (the same fluctuations in the context of all replicas are shown in Figure 3b). The ligand became significantly more closer to the X-ray structure after binding to the native binding site as reflected by iRMSD values. Namely, after correct binding, LoRMSD values got noticeably lower to around 2 Å (see Figure 3a). In the following sections, we do not discuss this aspect of our method; however, it is worth mentioning that the proposed method enabled a detailed analysis of plausible docking trajectories. The described docking procedure uses REMC protocol enhanced by simulated annealing of all 20 replicas. Figure 3c shows their evolution through different temperatures. Figure 4 provide more detailed pictures of structural flexibility for protein “receptor” and “ligand”. Protein–protein contacts defining the complex assembly are characterized in Figure 5. In the presented example, the most persistent protein–protein contacts occurred in about 15% of snapshots. Therefore, they were significantly less stable compared to intramolecular protein contacts (Figure 4).

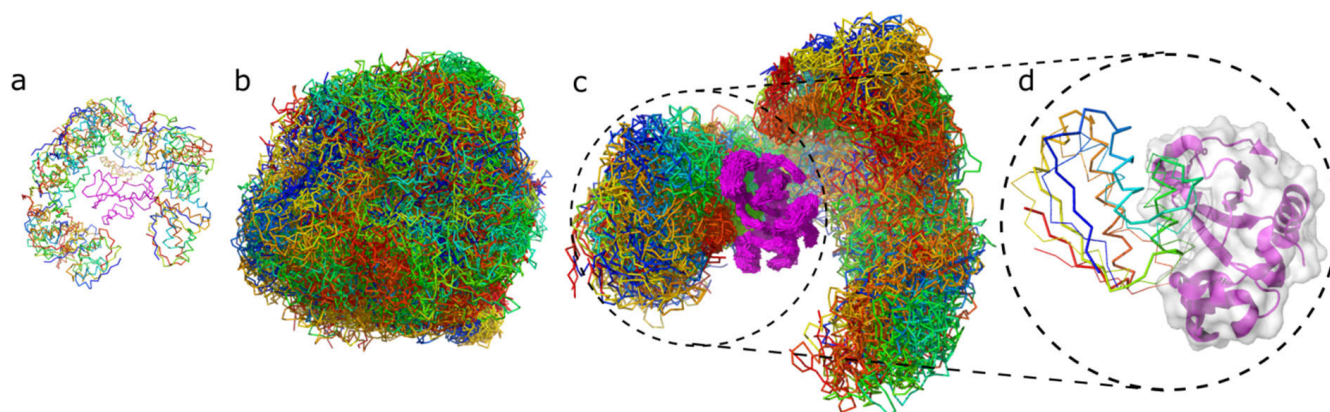


Figure 2. Protein–protein docking stages illustrated by barstar/barnase docking case. The figure shows the barnase receptor in magenta and the barstar ligand in rainbow colors. The respective panels show: (a) 20 starting structures for each replica of the system; (b) 10,000 models combined from 20 replicas (500 models per replica) in which the highly flexible ligand is covering the entire surface of the flexible receptor; (c) 500 models from one replica only, (d) the best model obtained for barnase/barstar system (the X-ray structure of the ligand is shown in thick ribbon, the modeled in thin ribbon).

An essential and unique feature of the presented docking simulations is the level of backbone flexibility during docking. In the example above, the ligand backbone fluctuations (LoRMSD) were in the range of 2–7 Å (Figure 3b), with the average LoRMSD value of 3.3 Å from the entire docking simulations. In other cases, the ligand fluctuations were at a similar level or higher (see LoRMSD values in Table 1).

Finally, using structural clustering of contact maps (see Methods), we attempted to select the set of 10 top-scored models for each case. Table 1 reports the most accurate models out of the 10 top-scored.

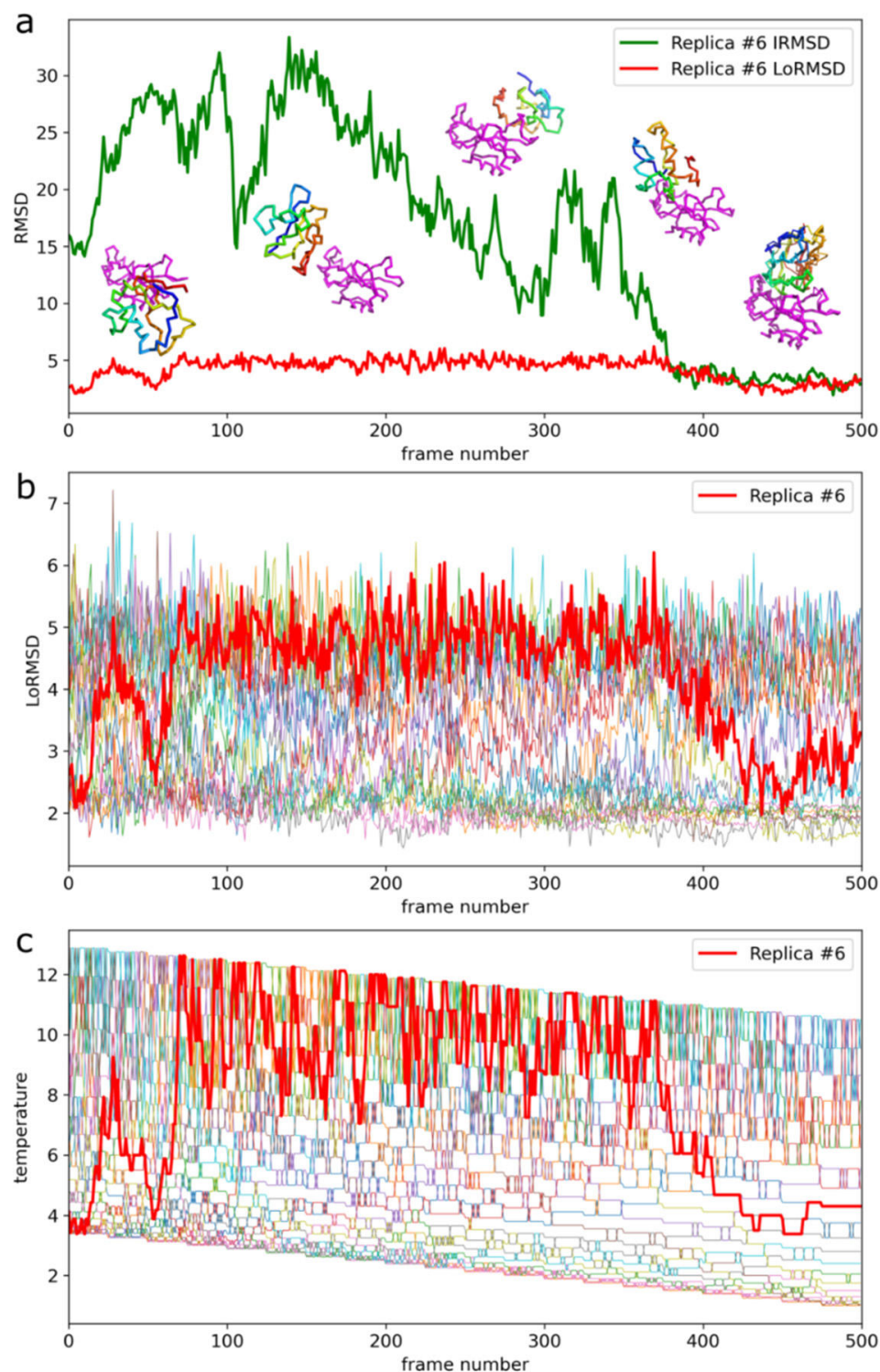


Figure 3. Docking trajectory for the selected replica of barnase/barstar system. The presented replica reached the most accurate barnase/barstar complex structure. (a) iRMSD (interface RMSD) and LoRMSD (ligand only RMSD) values. Example simulation snapshots illustrate the plot. The ligand is presented in rainbow colors, the receptor in magenta. The lowest iRMSD model (1.9 Å from X-ray structure) is presented on the right lower corner superimposed on the X-ray structure (the X-ray structure is shown in thick lines, the predicted model in thin lines). (b) Ligand only RMSD (LoRMSD) values for all replicas. The thick red line presents selected replica. (c) Exchange of system replicas between different temperatures driven by Replica Exchange Monte Carlo (REMC) system. The thick red line presents selected replica. The replica trajectory is also presented in the Video S1.

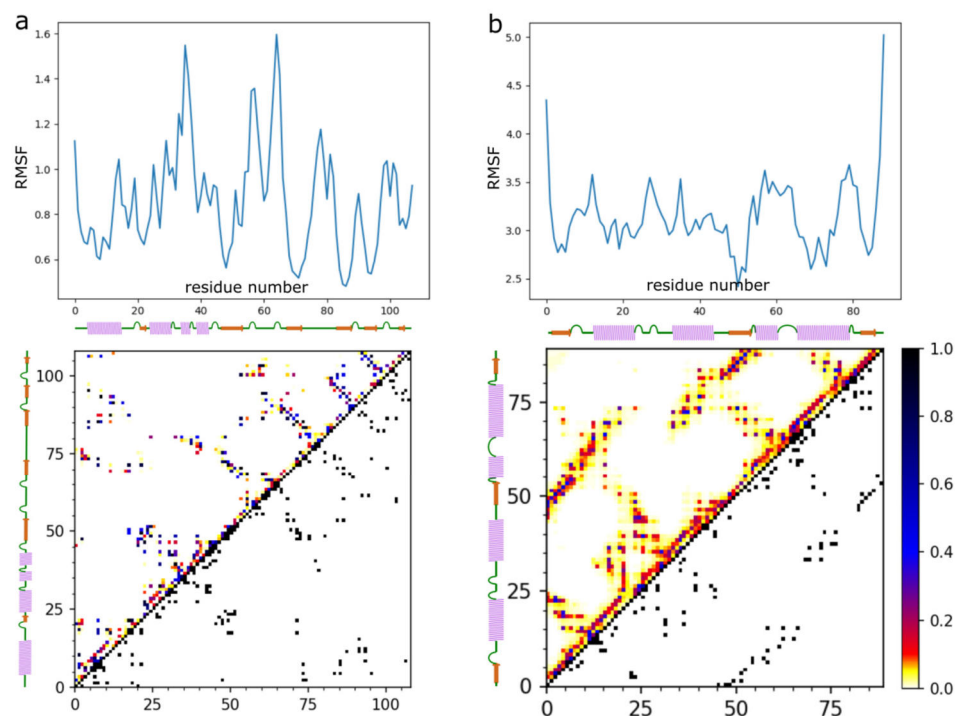


Figure 4. Characterization of barnase/barstar flexibility in the docking simulation. The figure shows RMSF plots (upper panels) and contact maps (lower panels) for (a) the barnase receptor and (b) the barstar ligand. The RMSF profile (see Methods) and contact maps showing the frequency of contacts are derived from the entire simulation (derived from 10,000 models).

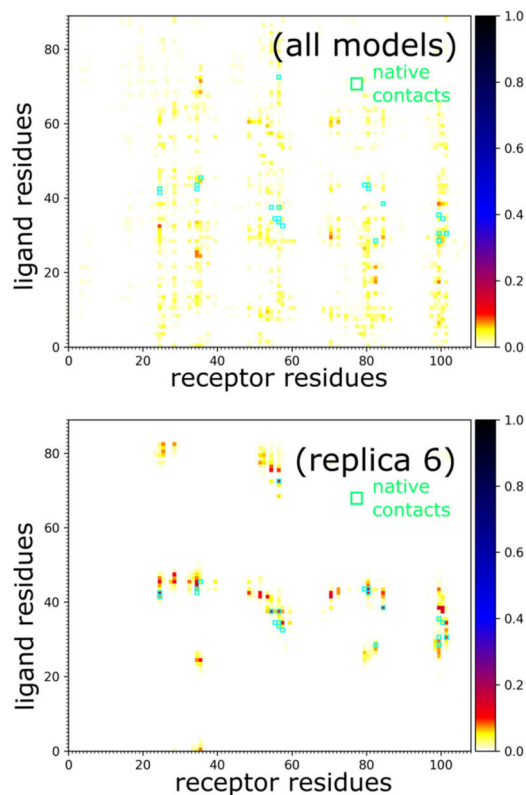


Figure 5. Characterization of barnase/barstar contacts. Panels show barnase/barstar models and contact maps for entire simulation (all models, 10,000 models) and single selected replica (replica 6, 500 models) that reached a near-native arrangement. In the maps, green circles mark the native contacts.

3. Discussion

This work demonstrates a significant improvement in the sampling of large-scale conformational transitions during global protein–protein docking compared to other state-of-the-art approaches. We show that modeling the large conformational changes is possible at a relatively low computational cost. The presented simulations took between 10 and 80 h (depending on the system size) using a single standard CPU. The proposed modeling protocol can be used as the docking engine in template-based and integrative docking protocols using experimental structural data and additional information from various sources [2,40]. We focused on the free docking of protein ligands with a highly flexible backbone in the present test simulations. Using unbound structures as the input, we produced acceptable accuracy models (iRMSD around 4 Å or lower) in low-flexibility and medium-flexibility cases. However, the selection procedure of the most accurate models needs further improvement. Namely, selecting the best-ranked models led to acceptable models in about half of the tested cases.

Presently, the most common approach to account for conformational changes in protein docking is using ENM [24–28,36–38]. The applicability of ENM to modeling protein flexibility is limited to specific systems and depends on how collective the protein motions are. Our method presents a conceptually different approach that seems to be more realistic (see review discussing coarse-grained CABS dynamics in the context of ENM approaches [24]). We demonstrated that it is possible to simulate effectively free docking of highly flexible protein ligands to quite elastic protein receptor structures. Such a significant degree of flexibility was achieved using a highly efficient simulation engine based on the coarse-grained representation of protein structures, Monte Carlo dynamics, and knowledge-based force field. CABS coarse-graining, enhanced by the discretized protein model and interaction patterns, significantly reduces the search space. Monte Carlo dynamics, enhanced by Replica Exchange annealing, leads to huge speed-up of the search procedures. Additionally, a significant (although acceptable for many problems) flattening of energy surfaces by statistical potentials of CABS model simplifies simulations. As a result the flexible docking using CABS-dock is orders of magnitude faster than equivalent simulations based on classical modeling methods. Obviously, the new method also has several limitations that have to be considered when designing new computational experiments. First, since the “ligand” protein is treated as a very elastic object (what is necessary to guarantee efficient search of the binding sites and poses) the cost of computations rapidly grows with the protein size. Thus, completely free global docking of protein ligands larger than 150 residues (see Table 1) may be impractical. Second, the coarse-graining of the sampling space and simplifying interaction patterns (so important for the huge acceleration of the simulations) makes the docking energetics less sensitive. For these reasons, the clustering procedures, refinement of the resulting structures, and final model selection become challenging and need further development. Additionally, speeding-up the entire protocol can be useful. We estimate that the simulations could be easily speeded-up at least 10 times or more through algorithm parallelization. The speed-up would enable making the protocol available as the publicly accessible and automated web service.

4. Methods

4.1. Docking Simulation Protocol

In this work, we present the protein–protein docking simulation protocol that relies on the CABS coarse-grained model. The CABS design and applications have been recently described in the reviews on protein coarse-grained [29] and protein flexibility [24,41] modeling. Here we outline only its main features. The CABS model uses a coarse-grained representation of protein chains (see Figure 6), Replica Exchange Monte Carlo (REMC) dynamics, and knowledge-based statistical potentials. Representation of protein chains is based on C-alpha traces, restricted to an underlying high-resolution lattice. The lattice spacing allows slight fluctuations of the C-alpha–C-alpha distances and many pseudo-bonds orientations. Virtual pseudo-atoms are placed in the centers of these C-alpha–C-

alpha bonds and are used to locate the main-chain hydrogen bonds. Additionally, the positions of the two pseudo-atoms representing side chains are defined by the geometry of C-alpha traces and amino-acid identities. Such fixed positions of side chains (taken from the statistics of protein databases) reduce the model's resolution. However, this limitation is less serious than it may appear since even small movements of the main chain (allowed due to the soft nature of the assumed geometrical restrictions) leads to large moves of the side chains. This way, the packing of side chains can be quite accurate. The interaction scheme of CABS consists of statistical potentials mimicking effects of main chain rotational preferences, main-chain hydrogen bonds, and side-chain contacts. All statistical potentials, derived from structural regularities observed in PDB structures, have relatively broad minima compensating the low-resolution effects and allowing a fast search for global energy minima. The solvent is treated implicitly, and its averaged effects are encoded within the above-mentioned contact potentials. Energy computation for protein chain models is very fast since many interactions could be pre-computed (and coded in large tables) due to the discretized patterns of main chains geometry. The Monte Carlo sampling of CABS uses a set of local movers. The resulting model dynamics is quite realistic for large-scale distances, allowing coarse-grained modeling of protein structures, dynamics, and protein-protein interactions.

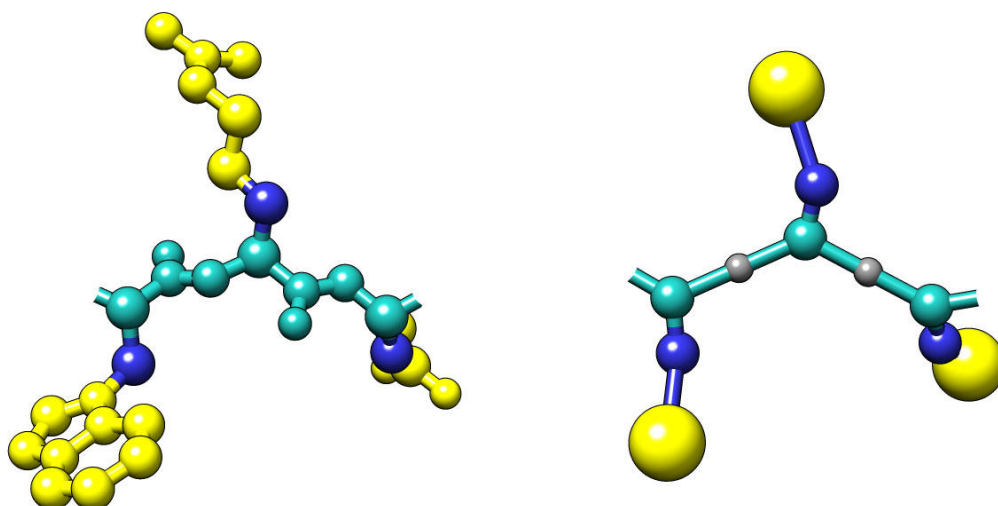


Figure 6. Comparison of the all-atom (**left**) and the CABS coarse-grained model representation (**right**) for an example tripeptide. In the CABS model, protein residues are represented using C-alpha, C-beta, united side-chain atom, and the peptide bond center [29].

The modeling protocol consists of the following steps:

1. **Preparing input structures of a protein-ligand and a protein-receptor.** The protocol requires the input of two protein structures (single- or multi-chain) in the PDB format. One of them has to be indicated as a ligand and the second as a receptor. The ligand undergoes large conformational fluctuations, translations, and rotations around the receptor within the proposed protocol. The “ligand” should be a smaller protein because the computational cost of searching its conformational space rapidly grows with the chain length. That is because the motion of the entire structure (including fold relaxation, rotation, and translation of the entire molecule) is simulated by a random sequence of local moves. The accuracy of such sampling is acceptable for not too-large proteins. On the other hand, treating the “ligand” as a fully flexible object allows approximate studies of entire docking trajectories. In some cases, it would be perhaps worth treating a larger protein (but not too large) as a flexible “ligand”, although this was out of range of the present studies.
2. **Generating starting structures.** Starting conformations are built using C-alpha coordinates only (in the CABS model C-alpha traces define the position of other united

pseudo-atoms, see details [29]). The algorithm places the protein-ligand center at 20 random positions around the protein receptor at the approximate distance of 20 Å from the protein receptor's surface. Next, these protein-ligand systems are used as starting conformations for the 20 replicas in the REMC CABS sampling scheme (each replica starts from a different ligand-receptor arrangement).

3. **Docking simulations using CABS coarse-grained model and REMC dynamics.** During simulations, the protein receptor structure is kept close to the starting structure using distance restraints. Distance restraints are generated using the input coordinates of the C-alpha atoms. Two residues are automatically restrained if two conditions are met. First, their separation along the sequence has to be at least five residues. Second, the distance between their C-alpha atoms must be within the range of 5–15 Å. During simulations, the receptor restraints imply small-scale fluctuations of the protein receptor backbone in the range of 1 Å and, accordingly, more significant fluctuations of the side-chain atoms. A similar restraints scheme is applied to the protein-ligand but with tenfold weaker weights. During simulations, the ligand moves freely within the vicinity of the receptor and internal restraint allows for large-scale fluctuations of its structure. Usually, the ligand fluctuations are within the range of 2 and 12 Å to the input structure although folding-unfolding events are possible at highest temperatures. The docking simulation is conducted using CABS REMC pseudo-dynamics with simulated annealing. In this work, 20 replicas and 20 annealing steps have been used. All the REMC scheme parameters have been adjusted to allow for large-scale conformational transitions, rotations, and translations of the protein-ligand in a reasonable computational time. The modeling protocol collects trajectories from all 20 replicas. The protocol saves only a small fraction (2%) of the generated models for further analysis i.e., 500 models from each replica, thus 10,000 models in total.
4. **Reconstructing to CABS coarse-grained representation.** The set of 10,000 models in C-alpha traces are reconstructed to complete CABS model representation using CABS algorithm [29]. In CABS, positions of C-beta and Side-Chain united atoms are defined by the positions of the three consecutive C-alpha atoms and the amino acid identity (the most probable positions from the PDB database are used).
5. **Clustering of contact maps.** First, for all of the 10,000 models the contact maps between the receptor and the ligand proteins are calculated. Two residues are considered to form a contact if their Side Chain pseudoatoms are at most 6 Å apart (for Alanines the C-beta atoms are considered as the Side Chain; for Glycines—it's the C-alpha atoms). Next, the algorithm sorts the models according to the number of the receptor-ligand contacts, and the set of top 1000 is kept for further processing. This way the transient and weakly bound complexes are removed from the solutions pool. In the next step, the 1000 contact maps are clustered together to identify the most frequently occurring contact patterns. The complete link hierarchical clustering was used with the Jaccard index as the distance metric between contact maps. Finally, the identified clusters are ranked according to their density, defined as the number of the cluster members divided by the average metric between them.
6. **Reconstructing to all-atom representation.** Representative models from the ten most dense clusters are reconstructed to all-atom representation using Modeller-based rebuilding procedure [42] (or can be reconstructed using other rebuilding strategies, see review [43]).

In recent years, the CABS model has been used for modeling the flexibility of globular proteins [44–47] and various processes leading to large-scale conformational transitions. These included: ab initio simulations of protein folding mechanisms [48,49], folding and binding mechanisms [49,50], and free protein–peptide docking within the CABS-dock tool [51–57]. The CABS-dock is a well-established peptide docking tool that has been made available as a web server [51,52] and, most recently, as a standalone application [54]. Its distinctive feature among other tools is the possibility of fast simulation of the large

backbone rearrangements of both peptide and protein receptors during binding (see the review on protein–peptide docking tools [58]). In addition, the CABS-dock has been used in multiple applications (recently reviewed [56]), including docking to receptors with disordered fragments [41,59], GPCRs [60], and modeling proteolysis mechanisms [61].

The presented protocol for protein–protein docking utilizes the CABS-dock standalone package [54] developed primarily for protein–peptide docking. In order to tackle the protein–protein docking problem, key changes have been made to the docking algorithm that aimed mainly at the improvement of the conformational sampling. First of all, the temperature distribution between replicas in the REMC scheme was adjusted. Instead of constant temperature increment between consecutive replicas, as in the original CABS-dock, here we've implemented progressive geometric raise of the temperature increment. Furthermore, the number of simulation replicas was increased to twenty versus ten in the original CABS-dock. Besides the sampling improvement, a new clustering protocol was introduced. The original CABS-dock used RMSD-based clustering, which worked well for peptides. For the protein–protein complexes, however, purely geometrical similarity condition such as the RMSD is too severe. Namely, for two binding poses, where the mobile protein was docked in the exact same pocket but is slightly tilted in one of them, the RMSD difference would be considerable. Despite representing similar binding poses, the two structures would end up in different clusters. To overcome this, the current protocol uses clustering based on the similarity between receptor–ligand contact maps.

4.2. Results Analysis and Quality Metrics

The docking simulation analysis was performed using Python and NumPy (Python library). Structural differences between experimentally determined structures and generated models were evaluated using Root Mean Square Deviations (RMSDs). Interface RMSD (iRMSD) is an RMSD calculated for interface residues of the receptor and the ligand separated by no more than 6 Å. Ligand RMSD (LRMSD) is an RMSD computed for the ligands after the superimposition of the receptors. Ligand only RMSD (LoRMSD) is an RMSD computed for the ligand structure only. Root Mean Square Fluctuation (RMSF) is a measure of the amino acid's flexibility. It is calculated for every residue as the square root of this residue's variance around the reference residue position. The fraction of native contacts (fNAT) was calculated as a number of experimental structure contacts found in the generated structure divided by the total number of contacts found in the experimental structure. Rather restrictive contact criterion, distance up to 6 Å between side-chain centers, was used. All figures presented in this work were generated using PyMOL, UCSF Chimera, and Matplotlib (Python library).

4.3. Dataset

In this docking study, we used protein–protein cases from the ZDOCK benchmark set [62] (cases in which a smaller size protein—a protein–ligand—contained more than one protein chain, or chain gaps, were discarded from our set). The set comprises the three flexibility-based subsets: low-flexible (almost rigid), medium-flexible, and highly flexible with available unbound X-ray structures of both the protein–receptor and the protein–ligand. The unbound structures were used as the docking input. As the reference for calculating various similarity measures, we used the X-ray structures of the protein–ligand complexes. Table 1 lists all the PDB IDs of X-ray structures used in the study.

5. Conclusions

In summary, the described docking procedure accounts for large-scale protein structure fluctuations during unrestrained protein–protein docking search for the binding site. The exploration of such vast conformational space has not been demonstrated before to the best of our knowledge. The approach shows unprecedented sampling possibilities; however, the accuracy of the obtained complexes is still lower than observed for state-of-the-art docking tools. Definitely, the balancing of the structural restraints scheme needs

further developments and tests. Therefore, this work is the first step towards a mature protein–protein docking tool. The next development steps would involve modifications of the distance restraints scheme, which allow for different degrees of flexibility for appropriate protein fragments (now the presented algorithm treats the entire protein–ligand as very flexible) and force-field improvements. The proposed approach is also very promising in the refinement applications when searching for the binding site is not needed, and only the protein–protein interface needs to be optimized.

Supplementary Materials: The following are available online at <https://www.mdpi.com/article/10.3390/ijms22147341/s1>, Video S1. The trajectory of a single replica from the protein–protein docking simulation of barnase/barstar system. The movie shows the barnase receptor in surface representation and the barstar ligand in ribbon. The presented replica reached the model with interface RMSD value 1.9 Å from the complex X-ray structure, shown as transparent ribbon.

Author Contributions: Conceptualization, M.K., S.K. and A.K.; Methodology, M.K. and A.K.; Software, M.K. and M.Z.; Supervision, S.K.; Visualization, M.K. and S.K.; Writing—original draft, S.K.; Writing—review & editing, M.K., S.K., M.Z. and A.K. All authors have read and agreed to the published version of the manuscript.

Funding: SK: MZ and AK acknowledge funding by the National Science Centre, Poland [MAE-STRO2014/14/A/ST6/00088]. MK acknowledges funding by the National Science Centre, Poland [2019/33/B/NZ2/02100]. SK also acknowledges funding in part by the National Science Centre, Poland [UMO-2020/39/B/NZ2/01301].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Porter, K.A.; Desta, I.; Kozakov, D.; Vajda, S. What method to use for protein–protein docking? *Curr. Opin. Struct. Biol.* **2019**, *55*, 1–7. [\[CrossRef\]](#)
- Rosell, M.; Fernández-Recio, J. Docking approaches for modeling multi-molecular assemblies. *Curr. Opin. Struct. Biol.* **2020**, *64*, 59–65. [\[CrossRef\]](#)
- Lensink, M.F.; Velankar, S.; Kryshtafovych, A.; Huang, S.; Schneidman-Duhovny, D.; Sali, A.; Segura, J.; Fernandez-Fuentes, N.; Viswanath, S.; Elber, R.; et al. Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: A CASP-CAPRI experiment. *Proteins Struct. Funct. Bioinf.* **2016**, *84*, 323–348. [\[CrossRef\]](#) [\[PubMed\]](#)
- Lensink, M.F.; Velankar, S.; Wodak, S.J. Modeling protein–protein and protein–peptide complexes: CAPRI 6th edition. *Proteins Struct. Funct. Bioinf.* **2017**, *85*, 359–377. [\[CrossRef\]](#) [\[PubMed\]](#)
- Lensink, M.F.; Nadzirin, N.; Velankar, S.; Wodak, S.J. Modeling protein–protein, protein–peptide, and protein–oligosaccharide complexes: CAPRI 7th edition. *Proteins Struct. Funct. Bioinf.* **2020**, *88*, 916–938. [\[CrossRef\]](#) [\[PubMed\]](#)
- Pierce, B.G.; Wiehe, K.; Hwang, H.; Kim, B.-H.; Vreven, T.; Weng, Z. ZDOCK server: Interactive docking prediction of protein–protein complexes and symmetric multimers. *Bioinformatics* **2014**, *30*, 1771–1773. [\[CrossRef\]](#) [\[PubMed\]](#)
- Kozakov, D.; Hall, D.R.; Xia, B.; Porter, K.A.; Padhorny, D.; Yueh, C.; Beglov, D.; Vajda, S. The ClusPro web server for protein–protein docking. *Nat. Protoc.* **2017**, *12*, 255–278. [\[CrossRef\]](#) [\[PubMed\]](#)
- Yan, Y.; Tao, H.; He, J.; Huang, S.-Y. The HDock server for integrated protein–protein docking. *Nat. Protoc.* **2020**, *15*, 1829–1852. [\[CrossRef\]](#) [\[PubMed\]](#)
- Venkatraman, V.; Yang, Y.D.; Sael, L.; Kihara, D. Protein–protein docking using region-based 3D Zernike descriptors. *BMC Bioinf.* **2009**, *10*, 407. [\[CrossRef\]](#) [\[PubMed\]](#)
- Christoffer, C.; Terashi, G.; Shin, W.; Aderinwale, T.; Maddhuri Venkata Subramaniya, S.R.; Peterson, L.; Verburt, J.; Kihara, D. Performance and enhancement of the LZerD protein assembly pipeline in CAPRI 38–46. *Proteins Struct. Funct. Bioinf.* **2020**, *88*, 948–961. [\[CrossRef\]](#)
- Estrin, M.; Wolfson, H.J. SnapDock—template-based docking by Geometric Hashing. *Bioinformatics* **2017**, *33*, i30–i36. [\[CrossRef\]](#)
- Gromiha, M.M.; Yugandhar, K.; Jemimah, S. Protein–protein interactions: Scoring schemes and binding affinity. *Curr. Opin. Struct. Biol.* **2017**, *44*, 31–38. [\[CrossRef\]](#)
- Feng, T.; Chen, F.; Kang, Y.; Sun, H.; Liu, H.; Li, D.; Zhu, F.; Hou, T. HawkRank: A new scoring function for protein–protein docking based on weighted energy terms. *J. Cheminform.* **2017**, *9*, 66. [\[CrossRef\]](#) [\[PubMed\]](#)

14. Geng, C.; Jung, Y.; Renaud, N.; Honavar, V.; Bonvin, A.M.J.J.; Xue, L.C. iScore: A novel graph kernel-based function for scoring protein–protein docking models. *Bioinformatics* **2020**, *36*, 112–121. [[CrossRef](#)] [[PubMed](#)]
15. Yan, Y.; Huang, S.-Y. Pushing the accuracy limit of shape complementarity for protein–protein docking. *BMC Bioinf.* **2019**, *20*, 696. [[CrossRef](#)]
16. Siebenmorgen, T.; Zacharias, M. Evaluation of Predicted Protein–Protein Complexes by Binding Free Energy Simulations. *J. Chem. Theory Comput.* **2019**, *15*, 2071–2086. [[CrossRef](#)]
17. van Zundert, G.C.P.; Rodrigues, J.P.G.L.M.; Trellet, M.; Schmitz, C.; Kastiris, P.L.; Karaca, E.; Melquiond, A.S.J.; van Dijk, M.; de Vries, S.J.; Bonvin, A.M.J.J. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J. Mol. Biol.* **2016**, *428*, 720–725. [[CrossRef](#)] [[PubMed](#)]
18. Lensink, M.F.; Brysbaert, G.; Nadzirin, N.; Velankar, S.; Chaleil, R.A.G.; Gerguri, T.; Bates, P.A.; Laine, E.; Carbone, A.; Grudinin, S.; et al. Blind prediction of homo- and hetero-protein complexes: The CASP13-CAPRI experiment. *Proteins Struct. Funct. Bioinf.* **2019**, *87*, 1200–1221. [[CrossRef](#)] [[PubMed](#)]
19. Harmalkar, A.; Gray, J.J. Advances to tackle backbone flexibility in protein docking. *Curr. Opin. Struct. Biol.* **2021**, *67*, 178–186. [[CrossRef](#)] [[PubMed](#)]
20. Siebenmorgen, T.; Engelhard, M.; Zacharias, M. Prediction of protein–protein complexes using replica exchange with repulsive scaling. *J. Comput. Chem.* **2020**, *41*, 1436–1447. [[CrossRef](#)]
21. Park, T.; Woo, H.; Baek, M.; Yang, J.; Seok, C. Structure prediction of biological assemblies using GALAXY in CAPRI rounds 38–45. *Proteins Struct. Funct. Bioinf.* **2020**, *88*, 1009–1017. [[CrossRef](#)]
22. Zalewski, M.; Kmiecik, S.; Koliński, M. Molecular Dynamics Scoring of Protein–Peptide Models Derived from Coarse-Grained Docking. *Molecules* **2021**, *26*, 3293. [[CrossRef](#)]
23. Peterson, L.X.; Kang, X.; Kihara, D. Assessment of protein side-chain conformation prediction methods in different residue environments. *Proteins Struct. Funct. Bioinf.* **2014**, *82*, 1971–1984. [[CrossRef](#)] [[PubMed](#)]
24. Kmiecik, S.; Kouza, M.; Badaczewska-Dawid, A.; Kloczkowski, A.; Kolinski, A. Modeling of Protein Structural Flexibility and Large-Scale Dynamics: Coarse-Grained Simulations and Elastic Network Models. *Int. J. Mol. Sci.* **2018**, *19*, 3496. [[CrossRef](#)] [[PubMed](#)]
25. Torchala, M.; Gerguri, T.; Chaleil, R.A.G.; Gordon, P.; Russell, F.; Keshani, M.; Bates, P.A. Enhanced sampling of protein conformational states for dynamic cross-docking within the protein–protein docking server SwarmDock. *Proteins Struct. Funct. Bioinf.* **2020**, *88*, 962–972. [[CrossRef](#)]
26. Jiménez-García, B.; Roel-Touris, J.; Romero-Durana, M.; Vidal, M.; Jiménez-González, D.; Fernández-Recio, J. LightDock: A new multi-scale approach to protein–protein docking. *Bioinformatics* **2018**, *34*, 49–55. [[CrossRef](#)]
27. Kurkcuoglu, Z.; Bonvin, A.M.J.J. Pre- and post-docking sampling of conformational changes using ClustENM and HADDOCK for protein–protein and protein–DNA systems. *Proteins Struct. Funct. Bioinf.* **2020**, *88*, 292–306. [[CrossRef](#)] [[PubMed](#)]
28. Schindler, C.E.M.; de Vries, S.J.; Zacharias, M. iATTRACT: Simultaneous global and local interface optimization for protein–protein docking refinement. *Proteins Struct. Funct. Bioinf.* **2015**, *83*, 248–258. [[CrossRef](#)] [[PubMed](#)]
29. Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A.E.; Kolinski, A. Coarse-Grained Protein Models and Their Applications. *Chem. Rev.* **2016**, *116*, 7898–7936. [[CrossRef](#)] [[PubMed](#)]
30. Baaden, M.; Marrink, S.J. Coarse-grain modelling of protein–protein interactions. *Curr. Opin. Struct. Biol.* **2013**, *23*, 878–886. [[CrossRef](#)]
31. Roel-Touris, J.; Bonvin, A.M.J.J. Coarse-grained (hybrid) integrative modeling of biomolecular interactions. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 1182–1190. [[CrossRef](#)]
32. Krupa, P.; Karczyńska, A.S.; Mozolewska, M.A.; Liwo, A.; Czaplewski, C. UNRES-Dock—protein–protein and peptide–protein docking by coarse-grained replica-exchange MD simulations. *Bioinformatics* **2020**. [[CrossRef](#)] [[PubMed](#)]
33. Kuroda, D.; Gray, J.J. Pushing the Backbone in Protein–Protein Docking. *Structure* **2016**, *24*, 1821–1829. [[CrossRef](#)] [[PubMed](#)]
34. Marze, N.A.; Roy Burman, S.S.; Sheffler, W.; Gray, J.J. Efficient flexible backbone protein–protein docking for challenging targets. *Bioinformatics* **2018**, *34*, 3461–3469. [[CrossRef](#)]
35. Roy Burman, S.S.; Nance, M.L.; Jeliazkov, J.R.; Labonte, J.W.; Lubin, J.H.; Biswas, N.; Gray, J.J. Novel sampling strategies and a coarse-grained score function for docking homomers, flexible heteromers, and oligosaccharides using Rosetta in CAPRI rounds 37–45. *Proteins Struct. Funct. Bioinf.* **2020**, *88*, 973–985. [[CrossRef](#)] [[PubMed](#)]
36. Zacharias, M. ATTRACT: Protein–protein docking in CAPRI using a reduced protein model. *Proteins Struct. Funct. Bioinf.* **2005**, *60*, 252–256. [[CrossRef](#)] [[PubMed](#)]
37. Glashagen, G.; Vries, S.; Uciechowska-Kaczmarzyk, U.; Samsonov, S.A.; Murail, S.; Tuffery, P.; Zacharias, M. Coarse-grained and atomic resolution biomolecular docking with the ATTRACT approach. *Proteins Struct. Funct. Bioinf.* **2020**, *88*, 1018–1028. [[CrossRef](#)] [[PubMed](#)]
38. Moal, I.H.; Bates, P.A. SwarmDock and the Use of Normal Modes in Protein–Protein Docking. *Int. J. Mol. Sci.* **2010**, *11*, 3623–3648. [[CrossRef](#)]
39. Yan, Y.; He, J.; Feng, Y.; Lin, P.; Tao, H.; Huang, S. Challenges and opportunities of automated protein–protein docking: HDock server vs. human predictions in CAPRI Rounds 38–46. *Proteins Struct. Funct. Bioinf.* **2020**, *88*, 1055–1069. [[CrossRef](#)] [[PubMed](#)]
40. Roel-Touris, J.; Don, C.G.V.; Honorato, R.; Rodrigues, J.P.G.L.M.; Bonvin, A.M.J.J. Less Is More: Coarse-Grained Integrative Modeling of Large Biomolecular Assemblies with HADDOCK. *J. Chem. Theory Comput.* **2019**, *15*, 6358–6367. [[CrossRef](#)]

41. Ciemny, M.; Badaczewska-Dawid, A.; Pikuzinska, M.; Kolinski, A.; Kmiecik, S. Modeling of Disordered Protein Structures Using Monte Carlo Simulations and Knowledge-Based Statistical Force Fields. *Int. J. Mol. Sci.* **2019**, *20*, 606. [[CrossRef](#)] [[PubMed](#)]
42. Badaczewska-Dawid, A.E.; Khramushin, A.; Kolinski, A.; Schueler-Furman, O.; Kmiecik, S. Protocols for All-Atom Reconstruction and High-Resolution Refinement of Protein–Peptide Complex Structures. *Methods Mol. Biol.* **2020**, *2165*, 273–287. [[CrossRef](#)] [[PubMed](#)]
43. Badaczewska-Dawid, A.E.; Kolinski, A.; Kmiecik, S. Computational reconstruction of atomistic protein structures from coarse-grained models. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 162–176. [[CrossRef](#)] [[PubMed](#)]
44. Jamroz, M.; Orozco, M.; Kolinski, A.; Kmiecik, S. Consistent View of Protein Fluctuations from All-Atom Molecular Dynamics and Coarse-Grained Dynamics with Knowledge-Based Force-Field. *J. Chem. Theory Comput.* **2013**, *9*, 119–125. [[CrossRef](#)] [[PubMed](#)]
45. Kuriata, A.; Gierut, A.M.; Oleniecki, T.; Ciemny, M.P.; Kolinski, A.; Kurcinski, M.; Kmiecik, S. CABS-flex 2.0: A web server for fast simulations of flexibility of protein structures. *Nucleic Acids Res.* **2018**, *46*, W338–W343. [[CrossRef](#)] [[PubMed](#)]
46. Kurcinski, M.; Oleniecki, T.; Ciemny, M.P.; Kuriata, A.; Kolinski, A.; Kmiecik, S. CABS-flex standalone: A simulation environment for fast modeling of protein flexibility. *Bioinformatics* **2019**, *35*, 694–695. [[CrossRef](#)] [[PubMed](#)]
47. Kmiecik, S.; Gront, D.; Kouza, M.; Kolinski, A. From Coarse-Grained to Atomic-Level Characterization of Protein Dynamics: Transition State for the Folding of B Domain of Protein A. *J. Phys. Chem. B* **2012**, *116*, 7026–7032. [[CrossRef](#)]
48. Kmiecik, S.; Kolinski, A. Characterization of protein-folding pathways by reduced-space modeling. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 12330–12335. [[CrossRef](#)]
49. Kmiecik, S.; Kolinski, A. Simulation of Chaperonin Effect on Protein Folding: A Shift from Nucleation–Condensation to Framework Mechanism. *J. Am. Chem. Soc.* **2011**, *133*, 10283–10289. [[CrossRef](#)]
50. Kurcinski, M.; Kolinski, A.; Kmiecik, S. Mechanism of Folding and Binding of an Intrinsically Disordered Protein As Revealed by ab Initio Simulations. *J. Chem. Theory Comput.* **2014**, *10*, 2224–2231. [[CrossRef](#)]
51. Kurcinski, M.; Jamroz, M.; Blaszczyk, M.; Kolinski, A.; Kmiecik, S. CABS-dock web server for the flexible docking of peptides to proteins without prior knowledge of the binding site. *Nucleic Acids Res.* **2015**, *43*, W419–W424. [[CrossRef](#)]
52. Blaszczyk, M.; Kurcinski, M.; Kouza, M.; Wieteska, L.; Debinski, A.; Kolinski, A.; Kmiecik, S. Modeling of protein–peptide interactions using the CABS-dock web server for binding site search and flexible docking. *Methods* **2016**, *93*, 72–83. [[CrossRef](#)]
53. Ciemny, M.P.; Kurcinski, M.; Kozak, K.; Kolinski, A.; Kmiecik, S. Highly flexible protein-peptide docking using cabs-dock. *Methods Mol. Biol.* **2017**, *1561*, 69–94. [[CrossRef](#)]
54. Kurcinski, M.; Ciemny, M.P.; Oleniecki, T.; Kuriata, A.; Badaczewska-Dawid, A.E.; Kolinski, A.; Kmiecik, S. CABS-dock standalone: A toolbox for flexible protein–peptide docking. *Bioinformatics* **2019**, *35*, 4170–4172. [[CrossRef](#)]
55. Blaszczyk, M.; Ciemny, M.P.; Kolinski, A.; Kurcinski, M.; Kmiecik, S. Protein-peptide docking using CABS-dock and contact information. *Brief. Bioinf.* **2019**, *20*, 2299–2305. [[CrossRef](#)]
56. Kurcinski, M.; Badaczewska-Dawid, A.; Kolinski, M.; Kolinski, A.; Kmiecik, S. Flexible docking of peptides to proteins using CABS-dock. *Protein Sci.* **2020**, *29*, 211–222. [[CrossRef](#)] [[PubMed](#)]
57. Ciemny, M.P.; Kurcinski, M.; Blaszczyk, M.; Kolinski, A.; Kmiecik, S. Modeling EphB4-EphrinB2 protein–protein interaction using flexible docking of a short linear motif. *Biomed. Eng. Online* **2017**, *16*, 71. [[CrossRef](#)]
58. Ciemny, M.; Kurcinski, M.; Kamel, K.; Kolinski, A.; Alam, N.; Schueler-Furman, O.; Kmiecik, S. Protein–peptide docking: Opportunities and challenges. *Drug Discov. Today* **2018**, *23*, 1530–1537. [[CrossRef](#)] [[PubMed](#)]
59. Ciemny, M.P.; Debinski, A.; Paczkowska, M.; Kolinski, A.; Kurcinski, M.; Kmiecik, S. Protein-peptide molecular docking with large-scale conformational changes: The p53-MDM2 interaction. *Sci. Rep.* **2016**, *6*. [[CrossRef](#)] [[PubMed](#)]
60. Badaczewska-Dawid, A.E.; Kmiecik, S.; Koliński, M. Docking of peptides to GPCRs using a combination of CABS-dock with FlexPepDock refinement. *Brief. Bioinf.* **2020**. [[CrossRef](#)]
61. Koliński, M.; Kmiecik, S.; Dec, R.; Piejko, M.; Mak, P.; Dzwolak, W. Docking interactions determine early cleavage events in insulin proteolysis by pepsin: Experiment and simulation. *Int. J. Biol. Macromol.* **2020**, *149*, 1151–1160. [[CrossRef](#)] [[PubMed](#)]
62. Vreven, T.; Moal, I.H.; Vangone, A.; Pierce, B.G.; Kastiris, P.L.; Torchala, M.; Chaleil, R.; Jiménez-García, B.; Bates, P.A.; Fernandez-Recio, J. Updates to the Integrated Protein-Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *J. Mol. Biol.* **2015**. [[CrossRef](#)] [[PubMed](#)]

3.2. [A2] Elastyczne dokowanie cyklicznych peptydów do białek z użyciem metody CABS-dock

3.2.1. Wprowadzenie do [A2]

W pracy [A2] przedstawiliśmy nowy protokół dokowania cyklicznych peptydów do białek, umożliwiający szerokie próbkowanie przestrzeni konformacyjnej przy jednoczesnym zachowaniu cyklicznej topologii peptydu. W projekcie tym odegrałem wiodącą rolę, odpowiadając za implementację metody, przeprowadzenie kluczowych obliczeń oraz analizę wyników, a także aktywny udział w przygotowaniu i redakcji manuskryptu. Celem pracy było rozszerzenie metody CABS-dock, pierwotnie opracowanej do dokowania peptydów liniowych, w taki sposób, aby umożliwiała poprawne generowanie modeli kompleksów z peptydami cyklicznymi, zarówno w trybie globalnym (*blind docking*, bez wiedzy o miejscu wiązania), jak i w trybie lokalnym, z wykorzystaniem orientacyjnej informacji o miejscu wiązania. Projektując zaproponowany protokół, dążyłem do połączenia szerokiego próbkowania konformacyjnego z rozsądnym kosztem obliczeniowym, tak aby metoda pozostawała użyteczna zarówno w warunkach ograniczonej wiedzy strukturalnej, jak i w przypadkach, gdy przybliżone miejsce wiązania jest znane.

Za wyborem cyklicznych peptydów jako głównego obiektu badań przemawiała ich rosnąca rola w projektowaniu bioaktywnych cząsteczek, szczególnie jako inhibitorów oddziaływań białko-białko^{87,88}. Peptydy o zamkniętej topologii łączą w sobie wysoką stabilność strukturalną, ograniczoną ruchomość konformacyjną oraz zwiększoną odporność na degradację proteolityczną⁸⁹⁻⁹¹. Z drugiej strony, ich modelowanie strukturalne stawia istotne wyzwania^{92,93} - standardowe metody dokowania peptydów często pomijają aspekt cykliczności, a wiele dostępnych narzędzi, szczególnie opartych na uczeniu maszynowym, ogranicza się do pojedynczych predykcji najlepszego modelu bez możliwości dalszego próbkowania.

W praktyce oznaczało to zaprojektowanie protokołu dokowania zdolnego do pracy zarówno w trybie globalnym („na ślepo”), jak i w trybie lokalnym, wykorzystującym informację o znanym lub przewidywanym miejscu wiązania. W trybie globalnym dokowania nie wykorzystywałem żadnych informacji o lokalizacji interfejsu, pozwalając na swobodne próbkowanie całej powierzchni białka. W trybie lokalnym ograniczałem przestrzeń próbkowania do określonego obszaru wokół miejsca wiązania. Oba tryby testowałem na zbiorze biologicznie istotnych kompleksów białko-peptyd.

CABS-dock umożliwia efektywne dokowanie peptydów liniowych do białek, przy znacząco niższym koszcie obliczeniowym w porównaniu z metodami pełnoatomowymi, co opisałem w rozdziale 2.2.2. Opracowany przeze mnie protokół opierał się na wykorzystaniu więzów odległości, które wymuszały zachowanie cyklicznej topologii peptydu podczas całej symulacji przy jednoczesnym zachowaniu jego ruchomości. Zaprojektowałem również schemat

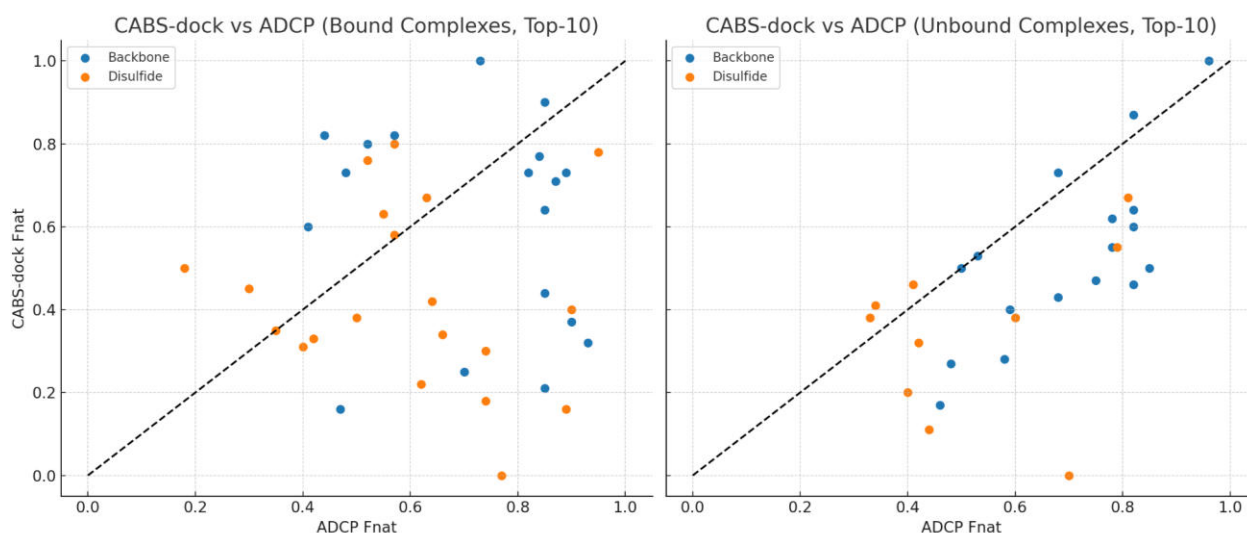
zastosowania tych więzów w taki sposób, aby umożliwić efektywne ograniczenie przestrzeni konformacyjnej wyłącznie do wskazanego miejsca wiązania.

Do oceny skuteczności opracowanego podejścia przygotowałem zbiór testowy składający się z 38 kompleksów białko-peptyd pochodzących ze zbioru ADCP⁹⁴, obejmujących zarówno peptydy cyklizowane przez zamknięcie łańcucha głównego, jak i przez mostki disiarczkowe. Dla każdego kompleksu przeprowadziłem symulacje w trybie globalnym oraz lokalnym, z wykorzystaniem zarówno struktury związanej, jak i wolnej białka receptora. Każda symulacja obejmowała generowanie 10 000 modeli, wybór 10 najlepszych (tzw. top 10), ich odbudowę do reprezentacji pełnoatomowej przy użyciu PD2⁹⁵, a następnie refinację z wykorzystaniem narzędzia Rosetta FlexPepDock⁹⁶. Do oceny jakości modeli stosowałem Fnat dla modeli po rekonstrukcji do reprezentacji pełnoatomowej (top 10) oraz LRMSD w przypadku modeli pozostających w reprezentacji C α (top 10 000). Taki schemat oceny pozwolił mi na bezpośrednie porównanie jakości modeli pełnoatomowych i gruboziarnistych, zależnie od dostępnego poziomu rekonstrukcji struktury.

Analiza wyników potwierdziła wysoką skuteczność opracowanego podejścia - w przypadkach *holo* akceptowalne kryterium Fnat ≥ 0.3 uzyskało (dla top 10) odpowiednio 83% przypadków peptydów cyklizowanych przez łańcuch główny i 80% peptydów cyklizowanych mostkiem disiarczkowym, a skuteczność była porównywalna do klasycznych metod dokowania. W przypadku struktur *apo*, gdzie modelowanie stanowiło większe wyzwanie, wartości te wynosiły odpowiednio 82% i 70%. Jednocześnie, analiza całych zespołów modeli (Top 10k) z wykorzystaniem LRMSD pokazała, że poprawne konformacje były często obecne nawet w przypadkach, gdzie klasyczny ranking energetyczny nie wskazywał ich wśród najwyżej ocenianych modeli. Należy jednak również zaznaczyć, że średnie wartości LRMSD dla struktur były znacznie niższe w trybie lokalnym niż globalnym, co sugeruje przydatność protokołu do generowania modeli wysokiej jakości przy orientacyjnym określeniu miejsca wiązania. Głównym ograniczeniem pozostawał jednak etap rankingu modeli, a nie generowanie poprawnych konformacji, co wyraźnie widoczne było w porównaniu jakości modeli Top 10 z jakością pełnego zespołu modeli.

Skuteczność opracowanego podejścia porównałem z wiodącymi metodami dokowania peptydów cyklicznych, którymi były: AutoDock CrankPep (ADCP) - w kategorii metod klasycznych, oraz HighFold⁹⁷ - wśród metod opartych na sztucznej inteligencji. Porównanie mojego protokołu z metodą ADCP przedstawiony został na Rysunku 4, gdzie dla każdego kompleksu pokazana została jakość najlepiej ocenionych modeli (Top 10, Fnat). Analiza tych danych potwierdziła, że w przypadku struktur *holo* obie metody osiągają podobne wyniki, natomiast dla struktur *apo* jakość generowanych modeli spada, przy czym spadek ten jest nieco bardziej wyraźny dla metody CABS-dock. W części układów CABS-dock uzyskał jednak lepsze wyniki niż ADCP, co wskazuje, że metody te mogłyby być traktowane jako podejścia komplementarne. Porównanie z metodą HighFold - nowoczesnym modelem głębokiego uczenia przeznaczonym do cyklicznych peptydów - potwierdziło natomiast wyraźną przewagę metod AI pod względem trafności najlepszego modelu (Top 1). HighFold uzyskał 100 % skuteczności na zestawie *apo*, podczas gdy mój protokół przekroczył granicę akceptowalności

tylko w ok. 35 % przypadków, co uwidoczniło ograniczenia klasycznego próbkowania w rankingu pojedynczych modeli. Warto jednak zaznaczyć, że moje podejście wyróżnia się możliwością generowania bogatych zespołów modeli, co czyni je wartościowym uzupełnieniem metod AI - zwłaszcza w badaniach eksploracyjnych lub w sytuacjach niepewności co do miejsca wiązania, gdzie analiza wielu możliwych stanów konformacyjnych ma szczególne znaczenie biologiczne.



Rysunek 4. Porównanie jakości modeli uzyskanych metodą CABS-dock i ADCP dla dokowania cyklicznych peptydów do białek (na podstawie [A2]). Każdy punkt odpowiada jednemu kompleksowi białko-peptyd: panel lewy przedstawia wyniki dla struktur związanych (*holo*), panel prawy dla struktur niezwiązanych (*apo*). Oś pozioma przedstawia wartość Fnat (Top 10) uzyskaną metodą ADCP, a oś pionowa - metodą CABS-dock. Punkty bliskie linii przerywanej wskazują podobną skuteczność obu metod, punkty powyżej tej linii - przewagę podejścia CABS-dock, a punkty poniżej - przewagę ADCP. Kolory oznaczają typ cyklizacji peptydu (zamknięcie łańcucha głównego - kolor niebieski; mostek disiarczkowy - kolor pomarańczowy). Źródło: [A2].

Opracowany przeze mnie protokół umożliwia dokowanie globalne i lokalne oraz generowanie zróżnicowanych stanów konformacyjnych przy zachowaniu realistycznych ograniczeń strukturalnych. Dzięki temu stanowi narzędzie przydatne zarówno w zadaniach predykcyjnych, jak i w badaniach eksploracyjnych wymagających szerokiego próbkowania przestrzeni możliwych modeli. Uzyskane wyniki pokazują, że metoda dostarcza zbiory wysokiej jakości, różnorodnych konformacji, które tworzą bogate zespoły struktur (ang. *conformational ensembles*). Takie zbiory mogą stanowić podstawę analiz konsensusowych, zwiększających wiarygodność i interpretowalność wyników, a sam protokół uzupełnia metody oparte na sztucznej inteligencji, koncentrujące się zwykle na pojedynczych predykcjach. Procedura wnosi istotny wkład w modelowanie kompleksów białko-peptyd, wspiera formułowanie nowych hipotez strukturalnych i może być potencjalnie użyta do projektowania cyklicznych peptydów o potencjale terapeutycznym.

3.2.2. Artykuł [A2]

Flexible Docking of Cyclic Peptides to Proteins Using CABS-dock

Mateusz Zalewski, Aleksandra Badaczewska-Dawid, and Sebastian Kmiecik*

Cite This: *J. Chem. Theory Comput.* 2025, 21, 8249–8254

Read Online

ACCESS |



Metrics & More

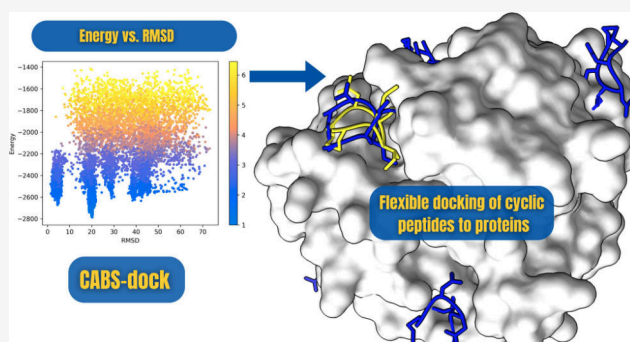


Article Recommendations



Supporting Information

ABSTRACT: Cyclic peptides are promising therapeutics, but their flexible docking remains challenging. We present a protocol based on the well-established CABS-dock method, enhanced with cyclic restraints and Rosetta refinement. The approach was evaluated on 38 benchmark complexes previously used in other docking method studies. While selecting the truly best model remains difficult, near-native solutions are frequently sampled. CABS-dock offers global, unbiased docking without prior binding site knowledge, making it valuable for pose generation, structural ensemble modeling, and integration into AI-driven peptide–protein docking workflows.



INTRODUCTION

Peptides are highly promising drug candidates due to their strong target specificity, low toxicity, and compatibility with rational design strategies.¹ However, linear peptides often suffer from major pharmacological limitations, including conformational flexibility, proteolytic instability, and poor oral bioavailability and membrane permeability.^{2,3} Cyclization—through backbone or disulfide linkages—enhances peptide rigidity, stability, and target affinity,⁴ making cyclic peptides promising tools for modulating protein–protein interactions.

Classical peptide–protein docking methods, ranging from coarse-grained^{5–7} to atomistic and hybrid tools,^{8–10} rely on explicit conformational sampling combined with scoring functions to evaluate binding poses. Many of these approaches were later adapted to address the specific challenges of cyclic peptides, such as conformational rigidity and backbone closure constraints, as exemplified by AutoDock CrankPep (ADCP)^{11,12} and other methods.^{13,14} Recent advances in deep learning—spurred by the success of AlphaFold—have led to a paradigm shift in the modeling of both linear^{15–17} and cyclic peptides,^{18,19} significantly improving prediction performance across peptide classes. For cyclic peptides, dedicated deep learning methods include HighFold¹⁸ and AfCycDesign,¹⁹ both based on the AlphaFold architecture. Among these, HighFold¹⁸ has consistently achieved higher Top-1 success rates and interface-level accuracy compared to AfCycDesign¹⁹ and classical methods such as ADCP,^{11,12} establishing a new benchmark in this field.

Despite this progress, AI-based methods like HighFold are typically deterministic, producing a single binding pose per input. This can be limiting in cases where binding is ambiguous or flexible, or where pose diversity and interpretability are

crucial. In contrast, methods, though generally inferior in scoring accuracy, can offer extensive conformational sampling that captures a broader landscape of potential binding modes.

In this work, we present a flexible docking protocol for cyclic peptide–protein complexes that supports both local and global docking modes. The global setup, which does not rely on prior knowledge of the binding site, represents—to our knowledge—the first such application for cyclic peptides. The protocol builds upon the well-established CABS-dock framework^{7,20–22} for flexible protein–peptide docking based on coarse-grained modeling and Monte Carlo sampling. To accommodate cyclic topologies, we extended CABS-dock with (1) distance restraints to preserve ring closure during sampling, (2) PD2-based reconstruction of C α traces to all-atom resolution,²³ and (3) Rosetta FlexPepDock-based refinement for improved backbone and side-chain accuracy.^{24,25} We evaluated both docking modes on a benchmark of 38 experimentally determined cyclic peptide–protein complexes from the ADCP data set,¹¹ 17 of which overlap with the HighFold study.¹⁸

MATERIALS AND METHODS

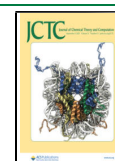
CABS-dock^{7,20–22} is a multiscale molecular docking method that employs a coarse-grained simulation engine based on the CABS model, illustrated in Supplementary Figure 1, and uses Replica Exchange Monte Carlo sampling to explore peptide–

Received: June 17, 2025

Revised: August 19, 2025

Accepted: August 19, 2025

Published: August 24, 2025



protein interactions. Within the method, the resulting models are reconstructed to all-atom resolution. Originally developed as a web server for global, flexible docking,^{7,20} it has since been adapted to a range of modeling tasks, as summarized in a review,²² including peptide docking with structural information,²⁶ protein–protein docking,^{27,28} GPCR–peptide interactions,²⁹ docking involving large receptor conformational changes,³⁰ and peptide cleavage site prediction.³¹ A standalone version²¹ allows for full control over simulation parameters and custom workflows.

To support cyclic peptides, we introduced distance restraints to maintain the closed-ring topology of backbone-cyclized and disulfide-cyclized peptides. These restraints were defined using the built-in constraint system of the standalone version (See [Supporting Methods](#) for detailed implementation and example syntax).

The simulation involves multiple steps, including conformational sampling, clustering, and all-atom reconstruction (see [Supporting Figure 2](#)). Each docking simulation generated 10,000 *Cα*-trace models, sampled without prior knowledge of the binding site. The models were scored using CABS internal energy terms, followed by structural clustering to identify the top 10 candidates.

In this work, we applied two docking modes: global and local. In global docking, the peptide was allowed to interact with the entire receptor surface. In local docking, the peptide was spatially restrained to the binding site region but remained fully flexible during sampling. This setup follows a protocol previously applied to linear peptides²⁶ (see [Supporting Methods](#) for implementation details). [Figure 1](#) presents an example case from the benchmark, highlighting the spatial extent of peptide sampling in global versus local docking setups.

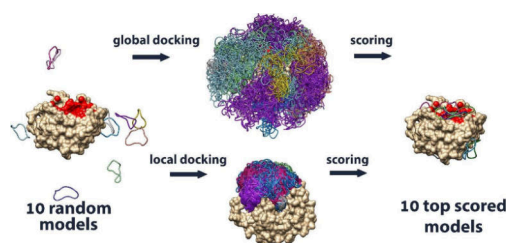


Figure 1. Visualization of peptide sampling in global and local docking modes, shown for a representative benchmark complex.

Benchmark Set and Evaluation. We evaluated the method on 38 peptide–protein complexes from the ADCP benchmark set,¹¹ comprising 18 backbone-cyclized and 20 disulfide-cyclized peptides. Each complex was redocked to both the bound (*holo*) and unbound (*apo*) receptor conformations. Success was measured using the CAPRI-defined Fn timeric,³² with additional ensemble analysis based on LRMSD (details in [Supporting Information](#)).

RESULTS AND DISCUSSION

In this work, we evaluated flexible docking of cyclic peptides to both *holo* and *apo* receptor structures using CABS-dock, in global and local modes. To our knowledge, this represents the first application of global docking to cyclic peptides without prior knowledge of the binding site. The benchmark included 38 complexes, divided into backbone-cyclized (18) and disulfide-cyclized (20) cases. *Apo* forms were available for 17

backbone-cyclized and 10 disulfide-cyclized complexes.¹¹ For each target, docking simulations were run in both modes, and the top 10 models were reconstructed from *Cα* traces to all-atom resolution and refined with Rosetta FlexPepDock. This refinement step was previously applied to linear peptides in multistage protocols for GPCR–peptide docking²⁹ (see [Supporting Information](#) for details). In our benchmark, the refined models reached acceptable quality ($\text{Fn timeric} \geq 0.3$) in 83% of backbone-cyclized and 80% of disulfide-cyclized *holo* complexes, with 67% and 35% reaching medium or high accuracy ($\text{Fn timeric} \geq 0.5$). For *apo* structures, the corresponding rates were 82% and 70%, with 56% and 50% reaching $\text{Fn timeric} \geq 0.5$. These values reflect the best-performing models selected from the top 10 scoring models in each simulation. Full results are provided in [Supporting Tables 1 and 2](#).

To complement the analysis of top-scored models, we evaluated the quality of the full CABS-dock output ensemble (top 10,000 models per run) using LRMSD, which is more suitable for C- α traces than Fn timeric, as it does not rely on atom-level contact definitions and provides a reliable measure of backbone-level structural accuracy. Average LRMSD values on bound complexes were 3.89 Å (cyclized by backbone) and 5.09 Å (cyclized by disulfide bond) for global docking, and 2.65 Å (backbone) and 4.23 Å (disulfide) for local docking. On unbound receptors, the respective averages were 4.76 Å (backbone) and 6.46 Å (disulfide) for global docking, and 2.92 Å (backbone) and 3.82 Å (disulfide) for local docking. These values confirm that even when top-scoring models miss the native state, the full ensemble often includes high-quality candidates (see [Figure 2A](#)).

Using the 5.5 Å LRMSD threshold established by Raveh et al.³³ as a reliable basin of attraction for near-native refinement, we found that for bound cases, acceptable structures (<5.5 Å) were present in 27 out of 38 cases (71%) with global docking, and in 34 out of 38 cases (89%) with local docking. For unbound cases, these numbers were 17 out of 27 (63%) for global and 26 out of 27 (96%) for local docking.

In summary, while near-native models are frequently generated, they are not always selected as top-ranked due to limitations in the scoring procedure. In CABS-dock, scoring is based on structural clustering combined with energy evaluation in the coarse-grained CABS model. Similar scoring-related issues—where accurate models were generated but not top-ranked—have also been observed previously for docking of linear peptides.^{7,20,26} [Figure 2A](#) shows that more accurate models often exist among the 10,000 generated structures than among the top 10 selected ones, indicating that the main limitation lies in scoring rather than sampling. Panels B and C illustrate a representative case where a near-native model appears in the top 10 but is not ranked first. Instead, the top-scoring structure is a misdocked conformation, likely favored due to its larger population and a deeper minimum in the coarse-grained energy landscape.

Comparison with ADCP. To assess the performance of CABS-dock relative to established docking tools, we compared it directly with AutoDock CrankPep (ADCP)—the classical state-of-the-art method for flexible docking of cyclic peptides prior to the emergence of AI-based predictors. The comparison was performed on the full benchmark set of bound (*holo*) and unbound (*apo*) receptor structures, grouped by peptide cyclization type: backbone or disulfide. For both methods, results are reported for the best model among the top 10

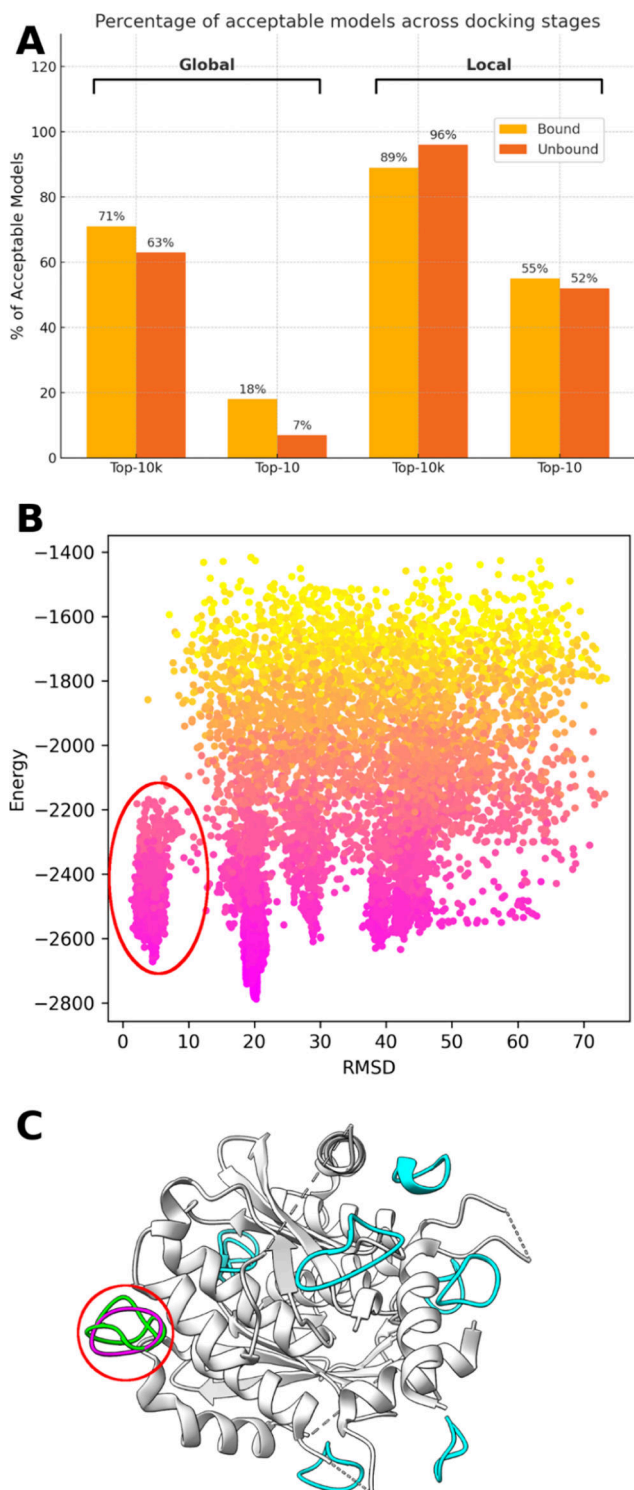


Figure 2. Sampling and scoring behavior in CABS-dock. (A) Fraction of complexes with at least one acceptable model (LRMSD < 5.5 Å) in Top-10 vs Top-10k, for global and local docking and bound/unbound receptors. (B) Energy vs LRMSD for complex 3WNE (global docking, bound receptor). Near-native models are circled in red. (C) Molecular visualization of the same case as in B. Top-10 predicted models are shown, with near-native structures in green and others in cyan. Experimental structure shown in magenta; red circles indicate near-native models.

scored predictions, following standard procedures from the original ADCP study¹¹

For bound complexes, CABS-dock showed competitive performance. In the backbone-cyclized subset ($n = 18$), it achieved an average Fnat of 0.61, with 83% acceptable (Fnat ≥ 0.3) and 67% medium-quality (Fnat ≥ 0.5) predictions. ADCP performed slightly better overall (avg Fnat 0.72; 100% acceptable; 78% medium), but CABS-dock outperformed ADCP in 7 out of 18 cases, indicating comparable strength in certain structural contexts. In the disulfide-cyclized subset ($n = 20$), CABS-dock yielded an average Fnat of 0.43 (16 acceptable, 7 medium), versus 0.60 for ADCP (19 acceptable, 15 medium). Again, CABS-dock surpassed ADCP in 7 targets, reinforcing its utility across structurally diverse or flexible systems.

For unbound complexes, both methods showed reduced performance due to increased receptor flexibility. The drop was more pronounced for CABS-dock, particularly in scoring. In the backbone-cyclized unbound set ($n = 17$), CABS-dock reached an average Fnat of 0.53 (14 acceptable, 10 medium), while ADCP maintained higher accuracy (avg Fnat 0.70; 17 acceptable; 15 medium). CABS-dock performed better on 4 targets. In the disulfide-cyclized unbound group ($n = 10$), CABS-dock averaged 0.35 (7 acceptable, 2 medium), compared to ADCP's 0.49 (10 acceptable, 4 medium), with CABS-dock outperforming ADCP in 3 cases.

While average performance favors ADCP, CABS-dock consistently generates near-native models, especially in challenging unbound scenarios, thanks to its flexible Monte Carlo sampling. However, limitations in scoring and model selection often prevent these models from being top-ranked. This is further supported by Figure 2A, where a greater number of unbound cases contain acceptable models among the Top-10 predictions compared to the bound set. Importantly, average performance does not capture the full picture of target-specific variability. As shown in Figure 3, per-target Fnat values vary widely across both methods, with each outperforming the other in a subset of cases. This suggests that their complementary strengths could be leveraged through selective or combined use to improve overall reliability.

Comparison with HighFold. To evaluate the performance of CABS-dock against modern AI-based methods, we compared it to HighFold—a recent deep learning framework for docking cyclic peptides. HighFold has shown superior predictive accuracy over previous methods, including AlphaFold-based AfCycDesign method.¹⁹ We used the same benchmark set of 17 unbound backbone-cyclized protein-peptide complexes and evaluated only the top-ranked (Top-1) models from each method, as reported in the HighFold study.¹⁸ Acceptable models were defined using an Fnat threshold of 0.2, following the HighFold benchmark. This differs from the 0.3 threshold used in earlier comparisons to ADCP, which followed its original evaluation protocol.¹¹

HighFold achieved acceptable predictions (Fnat ≥ 0.2) in all 17 cases (100%), with an average Fnat of 0.82, indicating consistently high accuracy. CABS-dock, by comparison, reached this threshold in only 7 cases (35.3%), with an average Fnat of 0.24. This gap reflects core methodological differences: HighFold directly predicts refined backbone structures using a transformer model trained on cyclic peptide complexes, while CABS-dock relies on exhaustive sampling with a coarse-grained model, followed by reconstruction and refinement. Although CABS-dock generates 10,000 models per

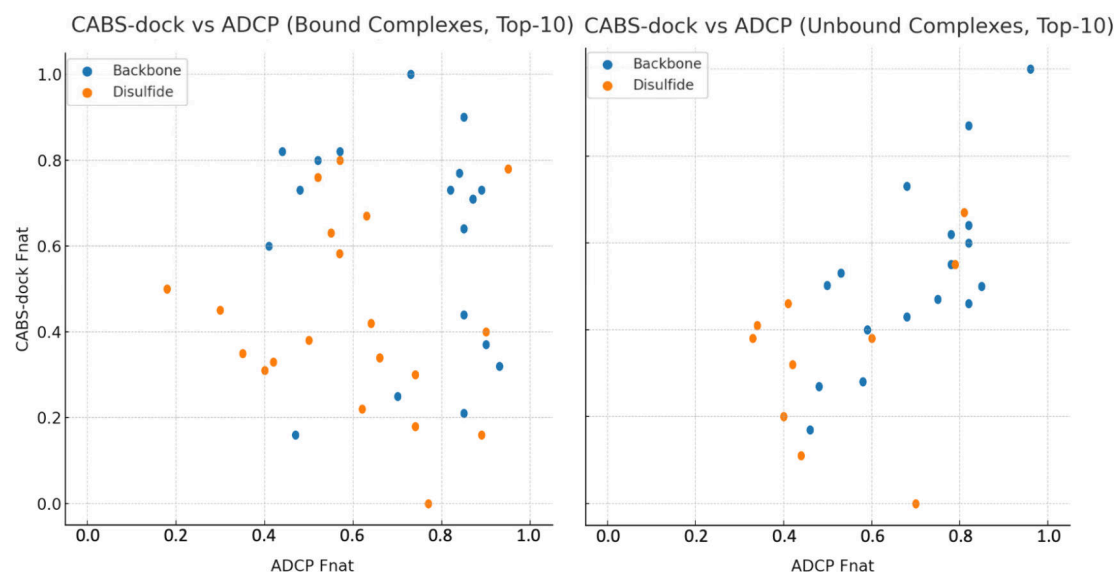


Figure 3. Per-target comparison of CABS-dock and ADCP (Top-10 Fnat) on bound (left) and unbound (right) complexes. Each point represents one target; colors indicate peptide cyclization type.

run, its scoring function often fails to identify the most accurate structures.

Nevertheless, in several cases, CABS-dock's broad sampling successfully recovered viable binding modes. A notable example is the 3ZGC complex: both methods produced acceptable Top-1 models, though none reached the medium-quality threshold ($\text{Fnat} \geq 0.5$). This was the worst-performing case in the HighFold benchmark, due to its unusual binding mode—an extended peptide conformation at the receptor rim—which is difficult for models trained primarily on pocket-like interactions.¹⁸ HighFold, AfCycDesign, and CABS-dock all achieved Fnat values around 0.44–0.47, while ADCP scored lower (0.15). Importantly, CABS-dock's Top-10 ensemble included better-quality poses, reaching $\text{Fnat} = 0.53$ (unbound) and 0.85 (bound). These results show that while CABS-dock lags behind in Top-1 accuracy, it can still generate near-native solutions in challenging cases. [Supplementary Figure 3](#) illustrates this contrast across all targets, with 3ZGC emerging as a unique outlier. Although HighFold excels in predictive precision, CABS-dock may contribute valuable structural diversity in hybrid or ensemble-based workflows.

CONCLUSIONS

In this study, we present a proof-of-concept application of the CABS-dock method for docking flexible cyclic peptides to protein targets. The protocol employs coarse-grained Monte Carlo sampling with distance-based restraints to mimic cyclic topology, followed by all-atom reconstruction. This work builds directly upon our previous studies, in which we used the standalone CABS-flex engine to predict the structures of free linear and cyclic peptides,^{34,35} as well as for docking linear peptides to globular proteins and GPCRs.²⁹ Here, we adapt the same core methodology to the docking context, demonstrating that flexible sampling can be applied to model bound-state peptide conformations under topological constraints.

Although the overall accuracy of predicted complexes does not match that of AlphaFold-based methods such as AfCycDesign or HighFold, CABS-dock performs comparably to classical methods such as ADCP, while offering a key

advantage: the ability to explore diverse peptide conformations under cyclic constraints. This makes it particularly useful in exploratory modeling where binding modes are unknown, or when structural ensembles are needed for reranking, refinement, or experimental data interpretation. Unlike many AI-based structure prediction tools, which are deterministic and often opaque, CABS-dock offers full control over flexibility, restraints, and sampling settings. This configurability supports its use in integrative or hypothesis-driven modeling, especially when peptide flexibility is biologically relevant.

Experimental evidence supports the idea that conformational plasticity plays a key role in cyclic peptide function, permeability, and binding specificity. At the same time, predictors like AlphaFold often overstabilize ordered structures and miss alternative or transient states observed by NMR or MD.³⁶ By explicitly sampling such conformational alternatives, CABS-dock has the potential to address this gap. While our evaluation focused on local docking near known interfaces, the method also supports global docking without prior knowledge of the binding site—making it applicable to challenging or noncanonical targets. Although its Top-1 ranking accuracy remains lower than that of state-of-the-art AI models, near-native solutions often appear in the generated ensembles, underscoring its value in hybrid or ensemble-based pipelines.

Looking ahead, further development could include guiding sampling with experimental restraints (e.g., from NMR or cross-linking) and improving model selection using physics-based or machine learning scoring. As such, CABS-dock should be seen not only as a docking tool, but also as a generator of diverse, physically plausible conformational states. This flexibility, combined with support for both local and global docking under cyclic constraints, gives it unique utility in exploratory modeling, data integration, and studies of systems where structural heterogeneity is functionally important.

ASSOCIATED CONTENT

Data Availability Statement

Detailed instructions for reproducing the docking protocol are provided in the Supporting Information of the paper. The standalone version of CABS-dock used in this study is available

at <https://bitbucket.org/lcbio/cabsdock>. All output structures corresponding to the Top-10 predicted models for each target are available in the associated GitHub repository: <https://github.com/ZalewskiMa/CABSdock-cyclic>.

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jctc.5c00995>.

Detailed descriptions of the docking workflow, input preparation, and reconstruction/refinement protocols used in this study; comprehensive tables reporting docking performance across all benchmark cases; and schematic figures illustrating the CABS model and simulation steps (PDF)

AUTHOR INFORMATION

Corresponding Author

Sebastian Kmiecik – Biological and Chemical Research Center, Faculty of Chemistry, University of Warsaw, 02-093 Warsaw, Poland; orcid.org/0000-0001-7623-0935; Email: sekmi@chem.uw.edu.pl

Authors

Mateusz Zalewski – Biological and Chemical Research Center, Faculty of Chemistry, University of Warsaw, 02-093 Warsaw, Poland

Aleksandra Badaczewska-Dawid – Genome Informatics Facility, Office of Biotechnology, Iowa State University, Ames, Iowa 50011, United States

Complete contact information is available at: <https://pubs.acs.org/doi/10.1021/acs.jctc.5c00995>

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Mateusz Zalewski and Sebastian Kmiecik have been supported by the National Science Centre, Poland (OPUS 2020/39/B/NZ2/01301).

REFERENCES

- (1) Muttenthaler, M.; King, G. F.; Adams, D. J.; Alewood, P. F. Trends in peptide drug discovery. *Nat. Rev. Drug Discov* **2021**, *20*, 309–325.
- (2) Craik, D. J.; Du, J. Cyclotides as drug design scaffolds. *Curr. Opin. Chem. Biol.* **2017**, *38*, 8–16.
- (3) Zorzi, A.; Deyle, K.; Heinis, C. Cyclic peptide therapeutics: past, present and future. *Curr. Opin. Chem. Biol.* **2017**, *38*, 24–29.
- (4) Abdalla, M. A.; McGaw, L. J. Natural Cyclic Peptides as an Attractive Modality for Therapeutics: A Mini Review. *Molecules* **2018**, *23*, 2080.
- (5) Lamiable, A.; Thévenet, P.; Rey, J.; Vavrusa, M.; Derreumaux, P.; Tufféry, P. PEP-FOLD3: faster de novo structure prediction for linear peptides in solution and in complex. *Nucleic Acids Res.* **2016**, *44*, W449–454.
- (6) Krupa, P.; Karczyńska, A. S.; Mozolewska, M. A.; Liwo, A.; Czaplewski, C. UNRES-Dock—protein–protein and peptide–protein docking by coarse-grained replica-exchange MD simulations. *Bioinformatics* **2021**, *37*, 1613–1615.
- (7) Kurcinski, M.; Jamroz, M.; Blaszczyk, M.; Kolinski, A.; Kmiecik, S. CABS-dock web server for the flexible docking of peptides to proteins without prior knowledge of the binding site. *Nucleic Acids Res.* **2015**, *43*, W419–424.
- (8) Xu, X.; Zou, X. Predicting Protein–Peptide Complex Structures by Accounting for Peptide Flexibility and the Physicochemical Environment. *J. Chem. Inf. Model* **2022**, *62*, 27–39.
- (9) Alam, N.; Goldstein, O.; Xia, B.; Porter, K. A.; Kozakov, D.; Schueler-Furman, O. High-resolution global peptide–protein docking using fragments-based PIPER-FlexPepDock. *PLOS Computational Biology* **2017**, *13*, No. e1005905.
- (10) Ciemny, M.; Kurcinski, M.; Kamel, K.; Kolinski, A.; Alam, N.; Schueler-Furman, O.; Kmiecik, S. Protein–peptide docking: opportunities and challenges. *Drug Discov. Today* **2018**, *23*, 1530–1537.
- (11) Zhang, Y.; Sanner, M. F. Docking flexible cyclic peptides with AutoDock CrankPep. *J. Chem. Theory Comput.* **2019**, *15*, 5161–5168.
- (12) Zhang, Y.; Sanner, M. F. AutoDock CrankPep: combining folding and docking to predict protein–peptide complexes. *Bioinformatics* **2019**, *35*, 5121–5127.
- (13) Tao, H.; Zhao, X.; Zhang, K.; Lin, P.; Huang, S.-Y. Docking cyclic peptides formed by a disulfide bond through a hierarchical strategy. *Bioinformatics* **2022**, *38*, 4109–4116.
- (14) Charitou, V.; van Keulen, S. C.; Bonvin, A. M. J. J. Cyclization and Docking Protocol for Cyclic Peptide–Protein Modeling Using HADDOCK2.4. *J. Chem. Theory Comput.* **2022**, *18*, 4027–4040.
- (15) Johansson-Åkhe, I.; Wallner, B. Improving peptide–protein docking with AlphaFold-Multimer using forced sampling. *Front. Bioinform.* **2022**, *2*, No. 959160.
- (16) Zalewski, M.; Wallner, B.; Kmiecik, S. Protein–Peptide Docking with ESMFold Language Model. *J. Chem. Theory Comput.* **2025**, *21*, 2817–2821.
- (17) Zhang, Z.; Verburgt, J.; Kagaya, Y.; Christoffer, C.; Kihara, D. Learning with Privileged Knowledge Distillation for Improved Peptide–Protein Docking. *ACS Omega* **2025**, *10*, 26684–26693.
- (18) Zhang, C.; Zhang, C.; Shang, T.; Zhu, N.; Wu, X.; Duan, H. HighFold: accurately predicting structures of cyclic peptides and complexes with head-to-tail and disulfide bridge constraints. *Briefings in Bioinformatics* **2024**, *25*, No. bbae215.
- (19) Rettie, S. A.; Campbell, K. V.; Bera, A. K.; Kang, A.; Kozlov, S.; Bueso, Y. F.; De La Cruz, J.; Ahlrichs, M.; Cheng, S.; Gerben, S. R.; et al. Cyclic peptide structure prediction and design using AlphaFold2. *Nat. Commun.* **2025**, *16*, 4730.
- (20) Blaszczyk, M.; Kurcinski, M.; Kouza, M.; Wieteska, L.; Debinski, A.; Kolinski, A.; Kmiecik, S. Modeling of protein–peptide interactions using the CABS-dock web server for binding site search and flexible docking. *Methods* **2016**, *93*, 72–83.
- (21) Kurcinski, M.; Pawel Ciemny, M.; Oleniecki, T.; Kuriata, A.; Badaczewska-Dawid, A. E.; Kolinski, A.; Kmiecik, S. CABS-dock standalone: a toolbox for flexible protein–peptide docking. *Bioinformatics* **2019**, *35*, 4170–4172.
- (22) Kurcinski, M.; Badaczewska-Dawid, A.; Kolinski, M.; Kolinski, A.; Kmiecik, S. Flexible docking of peptides to proteins using CABS-dock. *Protein Sci.* **2020**, *29*, 211–222.
- (23) Moore, B. L.; Kelley, L. A.; Barber, J.; Murray, J. W.; MacDonald, J. T. High-quality protein backbone reconstruction from alpha carbons using Gaussian mixture models. *J. Comput. Chem.* **2013**, *34*, 1881–1889.
- (24) London, N.; Raveh, B.; Cohen, E.; Fathi, G.; Schueler-Furman, O. Rosetta FlexPepDock web server—high resolution modeling of peptide–protein interactions. *Nucleic Acids Res.* **2011**, *39*, W249–253.
- (25) Ciemny, M.; Badaczewska-Dawid, A.; Pikuzinska, M.; Kolinski, A.; Kmiecik, S. Modeling of Disordered Protein Structures Using Monte Carlo Simulations and Knowledge-Based Statistical Force Fields. *IJMS* **2019**, *20*, 606.
- (26) Blaszczyk, M.; Ciemny, M. P.; Kolinski, A.; Kurcinski, M.; Kmiecik, S. Protein–peptide docking using CABS-dock and contact information. *Brief. Bioinform.* **2019**, *20*, 2299–2305.
- (27) Ciemny, M. P.; Kurcinski, M.; Blaszczyk, M.; Kolinski, A.; Kmiecik, S. Modeling EphB4–EphrinB2 protein–protein interaction using flexible docking of a short linear motif. *BioMedical Engineering OnLine* **2017**, *16*, 71.

- (28) Kurcinski, M.; Kmiecik, S.; Zalewski, M.; Kolinski, A. Protein-Protein Docking with Large-Scale Backbone Flexibility Using Coarse-Grained Monte-Carlo Simulations. *Int. J. Mol. Sci.* **2021**, *22*, 7341.
- (29) Badaczewska-Dawid, A. E.; Kmiecik, S.; Koliński, M. Docking of peptides to GPCRs using a combination of CABS-dock with FlexPepDock refinement. *Briefings in Bioinformatics* **2021**, *22*, No. bbaa109.
- (30) Ciemny, M. P.; Debinski, A.; Paczkowska, M.; Kolinski, A.; Kurcinski, M.; Kmiecik, S. Protein-peptide molecular docking with large-scale conformational changes: the p53-MDM2 interaction. *Sci. Rep* **2016**, *6*, No. 37532.
- (31) Koliński, M.; Kmiecik, S.; Dec, R.; Piejko, M.; Mak, P.; Dzwolak, W. Docking interactions determine early cleavage events in insulin proteolysis by pepsin: Experiment and simulation. *Int. J. Biol. Macromol.* **2020**, *149*, 1151–1160.
- (32) Lensink, M. F.; Velankar, S.; Wodak, S. J. Modeling protein-protein and protein-peptide complexes: CAPRI 6th edition. *Proteins* **2017**, *85*, 359–377.
- (33) Raveh, B.; London, N.; Schueler-Furman, O. Sub-angstrom modeling of complexes between flexible peptides and globular proteins. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 2029–2040.
- (34) Badaczewska-Dawid, A.; Wróblewski, K.; Kurcinski, M.; Kmiecik, S. Structure prediction of linear and cyclic peptides using CABS-flex. *Briefings in Bioinformatics* **2024**, *25*, No. bbae003.
- (35) Wróblewski, K.; Zalewski, M.; Kuriata, A.; Kmiecik, S. CABS-flex 3.0: an online tool for simulating protein structural flexibility and peptide modeling. *Nucleic Acids Res.* **2025**, *53*, W95–W101.
- (36) Fowler, N. J.; Williamson, M. P. The accuracy of protein structures in solution determined by AlphaFold and NMR. *Structure* **2022**, *30*, 925–933.e2.

3.3. [A3] Metoda oceny modeli kompleksów białko-peptyd pochodzących z dokowania gruboziarnistego oparta na dynamice molekularnej

3.3.1. Wprowadzenie do [A3]

Punktem wyjścia dla tego projektu było jedno z kluczowych ograniczeń metody CABS-dock, którym - jak omówiłem szerzej w sekcjach 3.1.1 oraz 3.2.1 niniejszej rozprawy na przykładzie prac [A1] i [A2] - pozostaje problem selekcji najlepszych modeli spośród dużych zbiorów zawierających poprawne konformacje. Już na etapie wcześniejszych badań wykazałem, że metody gruboziarniste, mimo skuteczności w generowaniu modeli zawierających rozwiązania bliskonatywne, charakteryzują się ograniczoną zdolnością do trafnej selekcji najlepszych konformacji przy użyciu standardowych procedur oceniania. W pracy [A3] skoncentrowałem się na przetestowaniu podejścia polegającego na wykorzystaniu krótkich symulacji dynamiki molekularnej (MD) jako dodatkowego etapu selekcji modeli. W realizacji tego projektu pełniłem wiodącą rolę jako pierwszy autor. Mój wkład obejmował przede wszystkim przygotowanie i przeprowadzenie obliczeń, analizę wyników oraz przygotowanie wizualizacji, a także współudział w opracowaniu szczegółowych założeń metodycznych oraz w procesie redakcji manuskryptu.

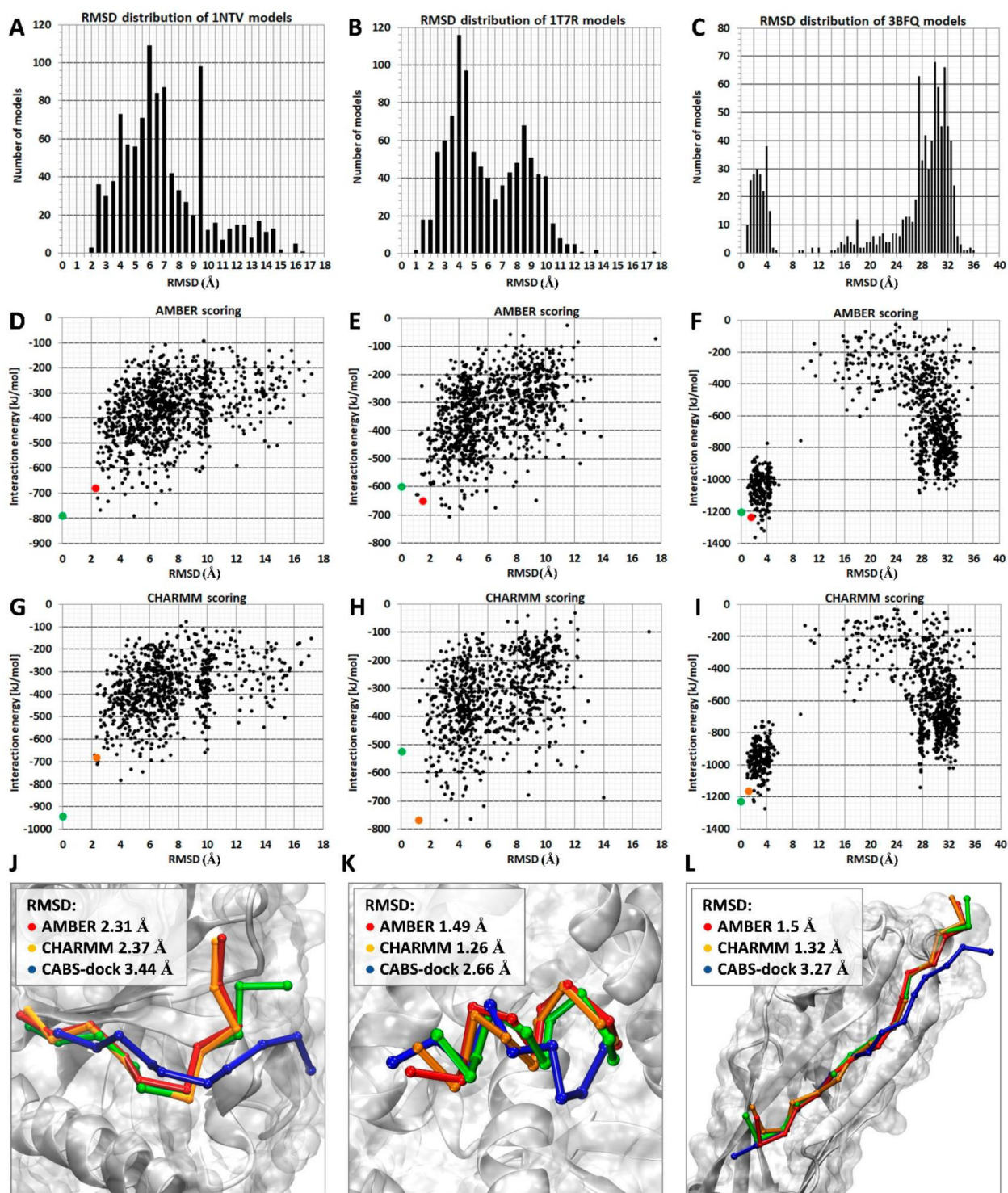
Celem pracy [A3] było opracowanie uproszczonego, praktycznego schematu wykorzystania dynamiki molekularnej do oceny interakcji białko-peptyd, przy zachowaniu rozsądnego kosztu obliczeniowego. Głównym założeniem było zastosowanie krótkich symulacji MD jako narzędzia poprawiającego selekcję modeli bez potrzeby wielogodzinnych symulacji produkcyjnych typowych dla klasycznych protokołów MD. Podstawy teoretyczne oraz założenia metodyczne symulacji dynamiki molekularnej zostały przeze mnie omówione w rozdziale 2.3.1.

Do realizacji tego celu wykorzystałem zbiór 66 kompleksów białko-peptyd pochodzących ze standardowego zbioru testowego CABS-dock²⁹, obejmującego szeroki zakres scenariuszy wiązania, w tym zarówno proste układy, jak i bardziej złożone przypadki o dużej różnorodności konformacyjnej. Wszystkie kompleksy pochodziły z globalnego dokowania peptydów do białek.

Procedura obejmowała rekonstrukcję pełnoatomową modeli CABS-dock przy użyciu narzędzia Modeller⁹⁸, po czym każdy model poddawany był krótkim symulacjom MD w środowisku wodnym, z zastosowaniem restrykcji położenia atomów C α zarówno peptydu, jak i białka receptora, co pozwalało ograniczyć globalne zmiany konformacyjne układu i skoncentrować analizę na stabilności oraz jakości oddziaływań w obrębie interfejsu białko-peptyd. Symulacje przeprowadziłem z użyciem trzech różnych pól siłowych: dwóch pełnoatomowych (CHARMM oraz AMBER) oraz gruboziarnistego MARTINI. Następnie dla każdego modelu obliczałem średnią energię oddziaływania białko-peptyd wyznaczoną na

podstawie ostatnich klatek trajektorii, a także końcowe wartości RMSD peptydu względem eksperymentalnej struktury krystalicznej.

Wyniki przeprowadzonych analiz pokazały, że ocena modeli oparta na krótkich symulacjach MD może pozwolić na lepsze rozróżnienie modeli wysokiej jakości niż klasyczna funkcja oceny CABS-dock oparta na analizie skupień (klastrowaniu strukturalnym). Największe korzyści zaobserwować można było przy ocenie zestawów 100 najlepszych modeli, gdzie procedura oceny oparta na MD z użyciem pól siłowych CHARMM oraz AMBER identyfikowała znacznie większą liczbę modeli o niższym RMSD względem struktury natywnej w przedziale 1-4 Å. W zestawach 10 najlepszych modeli zjawisko to było mniej wyraźne, ale również widoczne dla modeli wysokiej jakości ($\text{RMSD} < 2 \text{ Å}$). Szczegóły analizy skuteczności procedury oceny opisane zostały w rozdziale *Results and Discussion* publikacji [A3]. Wybrane przykłady działania tego podejścia zostały zilustrowane na Rysunku 5, który przedstawia wyniki rankingów modeli dla trzech kompleksów białko-peptyd. W tych przypadkach dodatkowy etap krótkich symulacji MD pozwolił na lepsze wyróżnienie modeli o niskich wartościach RMSD niż klasyczna procedura oceny CABS-dock, co potwierdza praktyczną przydatność takiego podejścia jako filtra wspomagającego selekcję modeli.



Rysunek 5. Przykłady działania procedury oceny opartej na krótkich symulacjach MD. Na panelach (A-C) pokazano histogramy wartości RMSD modeli wygenerowanych w zbiorze testowym CABS-dock dla trzech wybranych kompleksów (PDB ID: 1ntv, 1t7r i 3bfq). Panele (D-I) przedstawiają zależność energii oddziaływania od RMSD obliczonych dla modeli w trakcie etapu oceny z użyciem pól siłowych AMBER i CHARMM - czerwone i pomarańczowe punkty odpowiadają modelom najlepiej ocenionym w zestawach Top 10. Panele (J-L) pokazują porównanie najlepszego modelu wybranego na podstawie funkcji oceny CABS-dock (niebieski) i modeli wybranych po zastosowaniu oceny opartej na symulacjach MD (czerwony

- AMBER, pomarańczowy - CHARMM) z eksperymentalną strukturą peptydu (zielony).
Źródło: [A3].

Przeprowadzona została również dodatkowa analiza, której celem było zbadanie wpływu jakości rekonstrukcji struktury białka receptora na skuteczność oceny modeli opartej na symulacjach MD. Dla każdego kompleksu białko-peptyd wykonano dodatkowe symulacje, w których jedynie zrekonstruowaną strukturę receptora zastąpiono jego eksperymentalną strukturą krystaliczną. Wyniki jednoznacznie wykazały, że dokładność modelu receptora ma kluczowe znaczenie dla oceny energetycznej - zastosowanie struktury eksperymentalnej prowadziło do znacznego wzrostu udziału modeli o niskim RMSD wśród najlepiej ocenianych energetycznie.

Interesującym elementem badań była również analiza porównawcza procedur oceny opartych na różnych polach siłowych. W przypadkach modeli wysokiej rozdzielczości największą skuteczność wykazały pola AMBER oraz CHARMM, przy czym CHARMM wykazywał nieco lepsze wyniki dla modeli w przedziale RMSD 2-4 Å. Ocena oparta na MARTINI okazała się mniej skuteczna w selekcji najlepszych modeli w zestawach Top 10, natomiast była porównywalna do metod pełnoatomowych w szerszych zestawach Top 100. Szczegółowe dane liczbowe oraz wykresy znajdują się w materiałach uzupełniających do publikacji [A3].

Kluczową wartością opracowanego podejścia była możliwość wykorzystania stosunkowo niedrogich obliczeniowo symulacji MD jako praktycznego narzędzia wspomagającego selekcję modeli w zadaniach dokowania peptydów. Choć metodologia ta nie rozwiązuje problemu jakości rekonstrukcji, ani nie zastępuje kompleksowej oceny fizykochemicznej, pokazuje, że proste symulacje MD mogą znacząco poprawić selekcję modeli bez znacznego wzrostu kosztu obliczeniowego. Dalsze usprawnienia można byłoby prawdopodobnie osiągnąć poprzez poprawę procedur rekonstrukcyjnych oraz zastosowanie bardziej zaawansowanych schematów optymalizacji geometrii na etapie odbudowy modeli.

Uważam, że praca [A3] dobrze ilustruje potencjał łączenia szybkich metod dokowania z lekkimi symulacjami MD w ramach hybrydowych, wieloetapowych protokołów. Takie podejście nie konkuruje z metodami opartymi na sztucznej inteligencji, ale może je uzupełniać - jako dodatkowy etap walidacji i filtrowania dużych zbiorów modeli. W mojej ocenie w erze AI tego typu procedury pozostają wartościowe wszędzie tam, gdzie konieczna jest ocena wielu konkurencyjnych hipotez strukturalnych lub uzyskane modele budzą wątpliwości i wymagają szybkiego odrzucenia najbardziej błędnych konformacji. Metoda opracowana w pracy [A3] nie została jednak wykorzystana w moich kolejnych publikacjach głównie z powodów praktycznych - wymagałaby ona czasochłonnej rekonstrukcji pełnoatomowej i dodatkowych symulacji, co znacznie zwiększałoby koszt obliczeniowy przy pracy na dużych zbiorach modeli.

3.3.2. Artykuł [A3]

Article

Molecular Dynamics Scoring of Protein–Peptide Models Derived from Coarse-Grained Docking

Mateusz Zalewski ¹ , Sebastian Kmiecik ¹  and Michał Koliński ^{2,*} 

¹ Biological and Chemical Research Center, Faculty of Chemistry, University of Warsaw, 1 Pasteura St., 02-093 Warsaw, Poland; mateusz.zalewski.fuw@gmail.com (M.Z.); sekmi@chem.uw.edu.pl (S.K.)

² Bioinformatics Laboratory, Mossakowski Medical Research Centre, Polish Academy of Sciences, 5 Pawinskiego St., 02-106 Warsaw, Poland

* Correspondence: mkolinski@imdik.pan.pl

Abstract: One of the major challenges in the computational prediction of protein–peptide complexes is the scoring of predicted models. Usually, it is very difficult to find the most accurate solutions out of the vast number of sometimes very different and potentially plausible predictions. In this work, we tested the protocol for Molecular Dynamics (MD)-based scoring of protein–peptide complex models obtained from coarse-grained (CG) docking simulations. In the first step of the scoring procedure, all models generated by CABS-dock were reconstructed starting from their original C-alpha trace representations to all-atom (AA) structures. The second step included geometry optimization of the reconstructed complexes followed by model scoring based on receptor–ligand interaction energy estimated from short MD simulations in explicit water. We used two well-known AA MD force fields, CHARMM and AMBER, and a CG MARTINI force field. Scoring results for 66 different protein–peptide complexes show that the proposed MD-based scoring approach can be used to identify protein–peptide models of high accuracy. The results also indicate that the scoring accuracy may be significantly affected by the quality of the reconstructed protein receptor structures.

Keywords: model scoring; protein–peptide docking; molecular dynamics; CABS-dock; coarse-grained docking



Citation: Zalewski, M.; Kmiecik, S.; Koliński, M. Molecular Dynamics Scoring of Protein–Peptide Models Derived from Coarse-Grained Docking. *Molecules* **2021**, *26*, 3293. <https://doi.org/10.3390/molecules26113293>

Academic Editor: Maria Cristina De Rosa

Received: 30 April 2021

Accepted: 28 May 2021

Published: 30 May 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Protein–peptides interactions play essential roles in many cellular processes, and in recent years peptides have become attractive scaffolds for the design of new therapeutics [1]. Therefore, understanding the protein–peptide interactions and peptide-binding mechanisms is an important subject of theoretical and experimental research.

Computational prediction of protein–peptide complexes using molecular docking is a very challenging task [2]. First of all, peptides are very flexible molecules. Even relatively short peptides can adopt many distinct conformations and even marginal changes of their amino acid sequences may significantly alter their structure and dynamics. Additionally, even small changes in protein receptor three-dimensional structures can have qualitative effects on the binding mechanisms, binding energy, and structures of resulting complexes. For these reasons, molecular docking based on classical MD protocols is computationally extremely costly and, in most cases, not practical. Numerous approaches based on CG modeling strategies have been employed to deal with conformational sampling difficulties. For instance, the Rosetta FlexPepDock method [3] has proven quite effective in docking flexible peptides, providing the binding site can be approximately pre-assumed. On the contrary, another CG method, CABS-dock [4–6], enables simulations of global docking with the efficient exploration of docking poses over the entire receptor protein surface.

CABS-dock treats peptide ligands as fully flexible chains. The method allows for modeling structural changes of the receptor protein that are induced by peptide binding [6–8]. Appropriate coarse-graining of the protein/peptide structure and a very efficient Monte

Carlo sampling of the system dynamics enable an exhaustive search of the conformational space and unrestrained docking of even quite long peptides (or small proteins). Several peptides can be docked simultaneously [9]. Of course, coarse-graining of conformational space and inevitable smoothening of energy surface (CABS uses knowledge-based statistical potentials—see the Materials and Methods section and References for more details) leads to imprecise energy scaling of the docked structures. Standard CABS-dock protocols generate a large number of protein–peptide structures, which are subject to CABS-energy ranking and structural clustering of the results. Usually, it is possible to identify a set of 1000 docking results containing a good resolution of the structure/s (close to the native structure) of the protein–peptide complex [4–6], although a selection of the single best structure is essentially impossible.

The problem of scoring protein–peptide poses is one of the major challenges in the field of molecular docking [2]. Scoring functions for protein–ligand docking can be divided into four categories [10]: physics-based using classical force field; empirical, with the binding affinity estimated by counting various protein–ligand interactions; knowledge-based, derived from statistics of close contacts observed in a representative set of protein–ligand structures; and machine-learning, where binding affinity is predicted with machine-learning algorithms (random forest, deep-learning, etc.). Since every class of scoring functions has its limitations, the most efficient are hybrid scoring methods, based on mixed functions [2,10]. The majority of protein–peptide docking tools—such as Rosetta FlexPepDock [3,11], HADDOCK [12], PyDockWEB [13], pepATTRACT [14], CABS-dock [4], and BiPPred [15]—use energy-based scoring that employs physics or knowledge-based terms, or their combination [2]. Some of those methods improve structure selection by predictions based on sequence similarities [15], template structure comparison [16], mutagenesis data [17], coevolution information [18], or structural clustering [4,6].

In this paper, we describe and test the MD-based strategy for scoring protein–peptide complex models provided by the CABS-dock method. The scoring is based on protein–peptide interaction energy values estimated from a series of short MD simulations for the sets of 1000 predicted models. Extensive MD simulations for this large number of evaluated models would be extremely computationally expensive, and at least for this reason not practical. Therefore, we propose a simpler, and as we show quite effective, protocol for ranking the large number of CG CABS-dock models using AA MD. First, we use the Modeller method [19] for reconstructions of atomistic structures of protein–peptide complexes. Next, these reconstructed structures are submerged in a water environment and subjected to short MD simulations with restrained C-alpha positions. These short simulations lead to a better arrangement of the amino acid side chains. Two AA force fields, CHARMM and AMBER, and the MARTINI CG force field are used. This way, the original CABS-dock models can be re-scored using calculated values of protein–peptide interaction energy. Finally, we analyze the quality of obtained best models in the sets of 10 and 100 top-scored models, respectively, and compare them with the original CABS-dock scoring results based on structural clustering.

Additionally, we investigate the impact of the quality of reconstruction of protein receptor structure (mainly proper packing of protein side-chains) on the calculated protein–peptide interaction energy values and the resulting scoring accuracy. For this purpose, in all analyzed molecular models of protein–peptide complexes (reconstructed to AA representation), the structures of receptor proteins were substituted by their original AA crystal structures and the scoring procedure was repeated for all systems.

2. Results and Discussion

2.1. Comparison of MD-Based Scoring with CABS-Dock Scoring

An appropriate set of protein–peptide structures is needed to evaluate the performance of the scoring method. In particular, the benchmark set should contain a large population of alternative structural models of different protein–peptide complexes, with peptides that exhaustively sample the conformational space within the receptor binding site and its

vicinity. For this purpose, we have chosen docking results for 66 unique protein–peptide complexes from the CABS-dock simulations [20]. The selected benchmark set includes high-resolution models for which calculated RMSD values for the C-alpha atoms of bound peptide ligands were below 4 Å when compared to the crystal structures. The different distribution profiles of RMSD values (see the first column in Supplementary Materials, Figure S1) indicate distinct ligand sampling patterns for the analyzed complexes, which should be taken into account when interpreting the scoring results.

In this work AA and CG MD simulations are used to score a large number of molecular models of protein–peptide complexes. All models were scored according to their energy of protein–peptide interactions estimated based on short MD simulations in explicit solvent. During the MD simulation, positional restraints were imposed on C-alpha atoms. This way, the geometries of amino acid side chains were optimized and the initial positions and conformations of the protein–peptide backbones were maintained during the MD simulations. The measured changes in RMSD values for C-alpha atoms of peptide ligands were small, not exceeding 1 Å for structures resulting from the AA MD and 1.8 Å for the CG MD simulations. To account for even small changes in peptide ligand conformation the RMSD values for the bound peptides were recalculated after each MD simulation. The observed relations between the measured RMSD values and interaction energy values for the group of 66 protein–peptide complexes are shown in Supplementary Materials, Figure S1.

The same MD-based scoring procedure was used to calculate the protein–peptide interaction energy for the crystal structures taken from the PDB database and used as reference structures for all scored complexes. Obtained energy values were marked on the plots with red dots (see Supplementary Information, Figure S1). Interestingly, only in 7 out of 66 analyzed crystal structures, the estimated protein–peptide interaction energy for crystal structures was lower than the interaction energy calculated for all the predicted models (PDB ID: 1iak, 1n7f, 1nln, 1ntv, 1ou8, 2hpl, and 3bwa). Additionally, for complexes having a significant population of models with bound peptide ligands located inside the receptor-binding site or in its vicinity (e.g., PDB ID: 1d4t, 1w93, 2dze, 3bfq), the interaction energies calculated for the crystal structures were lower than the interaction energies calculated for most of the predicted models.

The scoring results obtained using three different force fields (AMBER, CHARMM, and MARTINI) are compared to CABS-dock benchmark data in Figure 1. Upper panels (Figure 1A–C) show the analysis for 10 top-scored models, and lower panels (Figure 1D–F) show the analysis for 100 top-scored models. At the top of each panel, a plot shows RMSD values for the most accurate model out of 10 or 100 top-scored models for each of 66 protein–peptides complexes. The vertical y-axis corresponds to CABS-dock-derived RMSD values, whereas the horizontal x-axis corresponds to RMSD values calculated using the MD-based scoring procedure. Dots marked with the color green, which are positioned above the black diagonal line, indicate models characterized by lower RMSD values obtained using MD-based scoring. Similarly, red dots, positioned below the black diagonal line, indicate models with the lower RMSD values derived using the CABS-dock scoring approach. At the bottom of each panel, a bar plot shows the number of identified models with lower RMSD values using either the MD-based scoring procedure or the CABS-dock method calculated for 0.5 Å intervals.

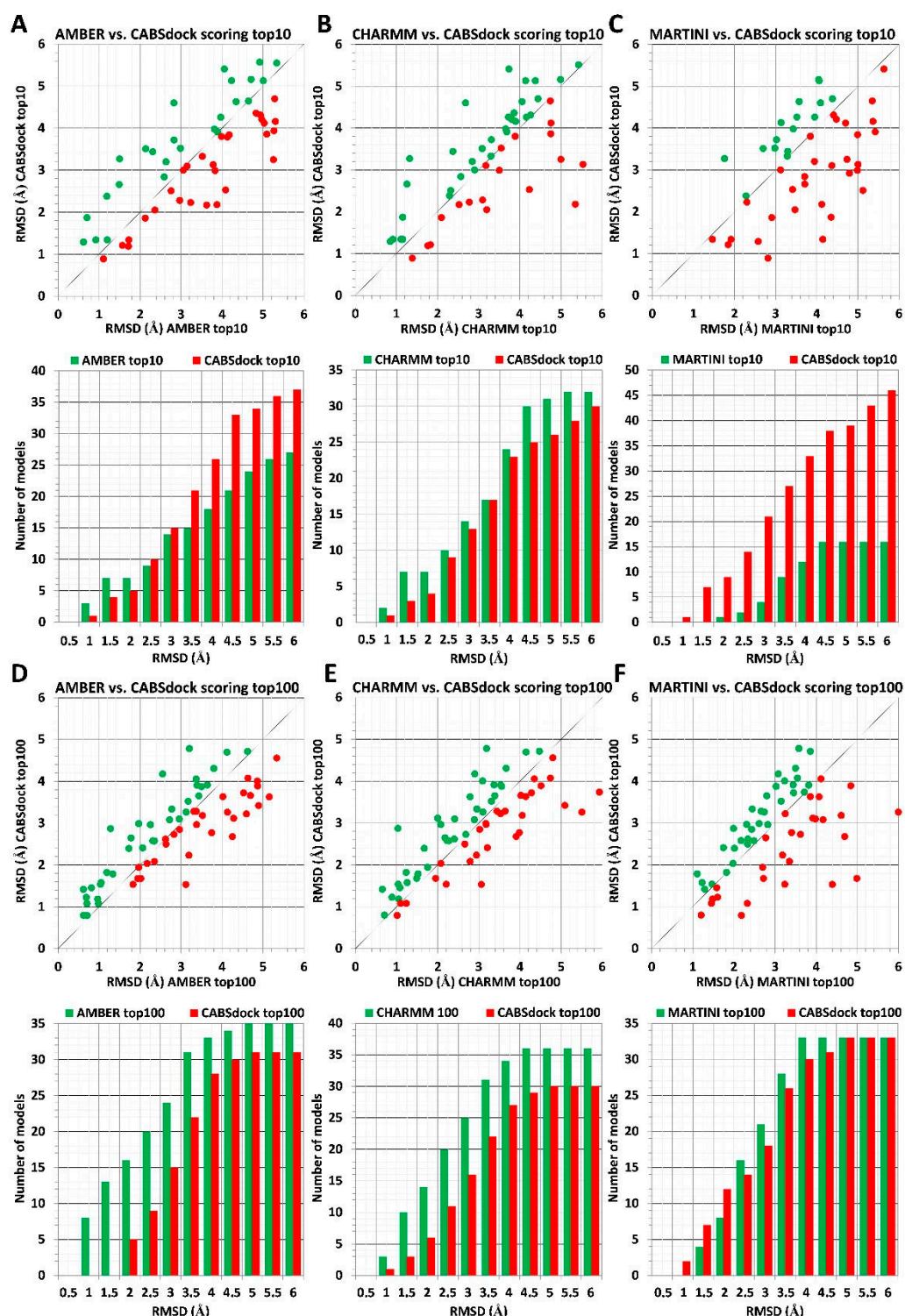


Figure 1. Comparison of MD-based versus CABS-dock scoring. The scoring data for 66 different protein–peptide complexes is shown. MD-based scoring was performed using three different force fields: AMBER, CHARMM, and MARTINI. Panels (A–C) show results obtained for sets of top 10 models, and panels (D–F) show results for sets of top 100 models. The plot at the top of each panel (A–F) shows RMSD values for best rank models according to the MD-based scoring method (color green) vs. RMSD values obtained for CABS-dock models (color red). At the bottom of each panel, the bar plot shows the number of models with the lower RMSD values identified using either the MD-based scoring method (color green) or using the CABS-dock scoring scheme (color red) for 0.5 Å intervals. Plots made based on data from Table S1.

Scoring results for the sets of 10 top-scored models indicate that the proposed MD-based scoring procedure using AA force fields of AMBER and CHARMM delivers more accurate scoring results than the default CABS-dock scoring. This is true for high-resolution complex structures with RMSD values below 2 Å (Figure 1A,B). Additionally, for identified models with resolution above 2 Å, scoring based on CHARMM parametrization performs slightly better than CABS-dock. In contrast, CABS-dock performs better for models above 2 Å when compared to MD-based scoring employing AMBER parametrization (Figure 1A,B).

When comparing the sets of 100 top-scored models, one can see the AA MD-based scoring approach's superiority over the CABS-dock scoring. In the considered range of RMSD values from 0 to 6 Å, the AA MD-based procedure identified a significantly larger number of models with lower RMSD values when compared to CABS-dock results (Figure 1D,E).

In contrast to scoring results obtained using AA MD simulations, the CABS-dock scoring method shows much better performance when compared to scoring employing CG MARTINI force field for sets of top 10 models, whereas both scoring methods showed similar accuracy for ranking sets of top 100 models (Figure 1C,F).

An example of ranking results for the three protein–peptide complexes for which the AA MD-based scoring procedure provided more accurate results than the default CABS-dock scoring method is shown in Figure 2. Distribution plots of RMSD values calculated for C-alpha atoms of bound peptides are included in panels A, B, and C. As demonstrated, the sets of models for the first two systems (PDB IDs: 1ntv, 1t7r) have a similar distribution of RMSD values in the range from 2 to 12 Å, which means that the docked ligand extensively samples the receptor-binding site and its vicinity (Figure 2A,B). The set of models for the third complex (PDB ID: 3bfq) shows a bimodal RMSD distribution pattern with two noticeable peaks (Figure 2C). For the majority of the docked peptides, the calculated RMSD values fall within two separate bins, the first ranging from 1 to 4 Å and the second ranging from 28 to 32 Å. Those two clusters suggest that the peptide ligand can accept two distinct binding modes: one that closely resembles the reference structure; and another one located on the protein surface at a considerable distance from the receptor protein binding site.

Two plots showing RMSD values and corresponding interaction energy values calculated using AMBER and CHARMM parametrization for each scored set of models are presented in Figure 2D–I. The correlation between low values of interaction energy with low RMSD values calculated for bound peptides allows identification of the high accuracy models in the sets of top 10 models (red and orange dots).

The three peptide ligands in analyzed systems have similar peptide chain lengths (1ntv: 10 aa, 1t7r 10 aa, and 3bfq 15 aa), but when forming a complex, they adopt different secondary structures, including a coil-like conformation, a short α -helix, and an extended conformation forming a long β -sheet with parts of protein receptor. In Figure 2, in the bottom panels J, K, and L, top-scored models are compared with experimental structures. In the three presented systems, the proposed MD-based scoring procedure identified more accurate models than the CABS-dock scoring. The differences for the lowest RMSD values calculated for the set of top 10 models were above 1 Å in the case of 1ntv and 1t7r systems and about 2 Å in the 3bfq system.

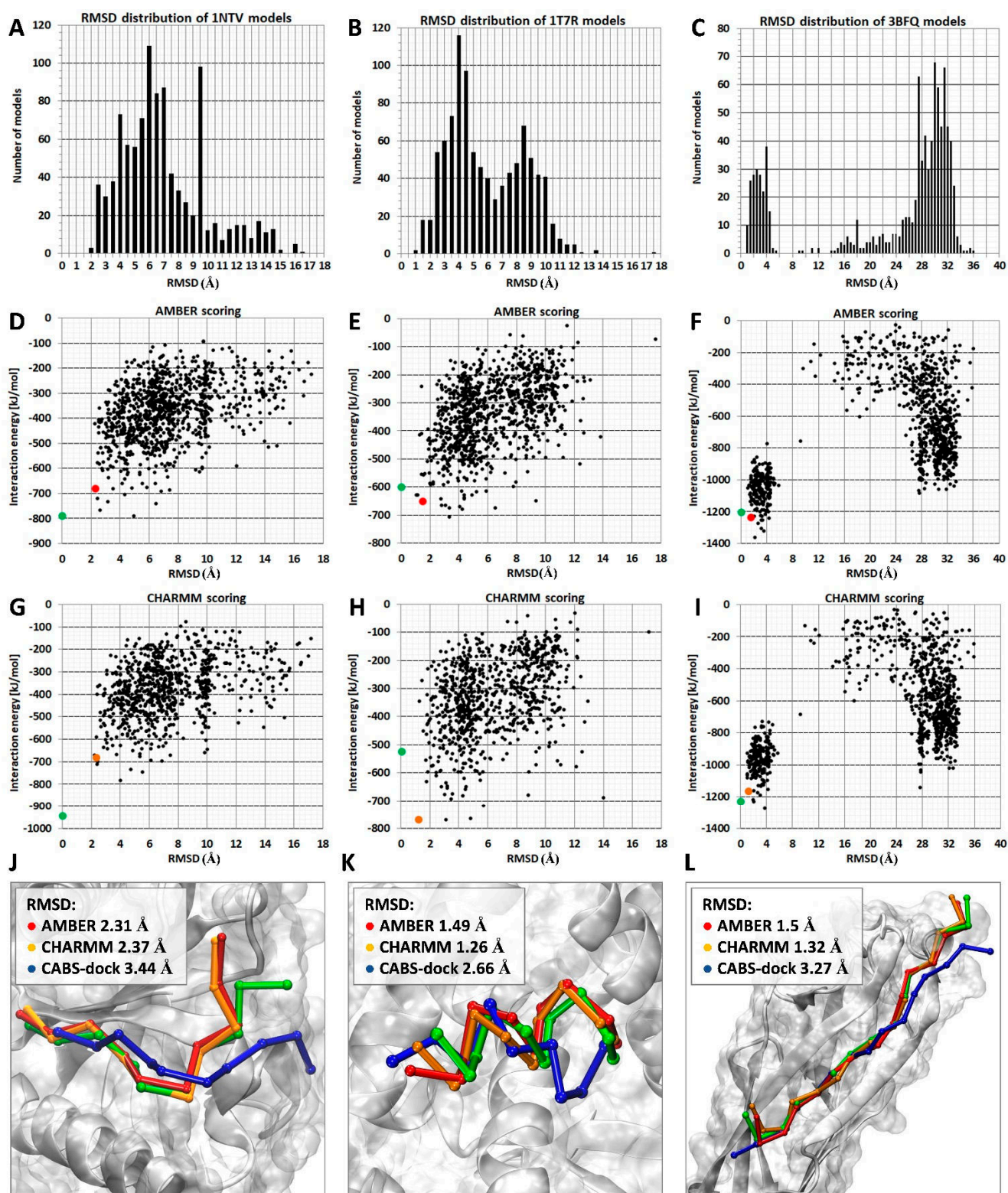


Figure 2. Examples of the MD-based scoring results. Ranking results for three complexes, 1ntv, 1t7r, and 3bfq, are shown. In panels (A–C) histograms of RMSD values calculated for model set derived from CABS-dock benchmark [20] are shown. Plots presenting RMSD values and corresponding interaction energy values calculated using AMBER and CHARMM force fields are shown in panels (D–I). Panels (J–L) show visualization of the best model found in the set of top 10 models for each of three complexes: CABS-dock model—color blue, the model scored using AMBER—color red, the model scored using CHARMM—color orange. For comparison, the crystal structure of peptide ligand is also displayed—color green.

2.2. Substitution of a Protein Receptor with Its Crystal Structure and Its Impact on Scoring Results

The CABS method uses discretized coordinates and a simplified representation of the protein chain. It allows very fast and efficient sampling of conformational space, which gives an advantage over other existing methods dedicated to protein–peptide docking. Docked ligands are fully flexible, and some structural changes are also allowed for receptor proteins that facilitate complex formation. However, this approach requires the reconstruction of full atomic details from C-alpha trace representation at the last stage of the docking procedure. This step is conducted using the Modeller method [19]. Unfortunately, the reconstruction process does not always assure the high quality of the local structure. This is due to inaccuracies in the C-alpha trace positions in the CG models and sometimes is not an optimal reconstruction of the backbone and side-chain positions (see our review [21]).

The molecular models of protein–peptide complexes used in this study were taken from the CABS-dock benchmark conducted for peptide-bound systems. It means that the initial structure of the receptor protein used for docking was taken directly from the protein–peptide crystallographic structure and presented bound conformation of the receptor protein. However, all resulting complex models contained a protein receptor whose structure was slightly different from its initial bound conformation due to the above-mentioned features of the CABS model, namely receptor flexibility during simulation followed by its reconstruction to AA representation. To analyze the effect of the changes in receptor structure (with respect to the input crystal structure) on the scoring results, we repeated the scoring procedure using AA MD with one modification—the receptor protein in scored complexes was replaced by the original structure taken from the crystallographic structure of the receptor–protein complex.

The ranking results indicate that the usage of receptor crystal structures significantly improves the scoring accuracy when compared to the CABS-dock driven data (see Figure 3). For scoring of sets of top 10 models, the MD-based procedure was able to find on average twice as many models presenting lower RMSD values in the RMSD range from 1 to 3 Å and in the range from 1 to 3.5 Å using AMBER and CHARMM parameters, respectively (Figure 3A,B). The data indicate that atomistic details of the receptor protein play a crucial role in estimating receptor–ligand interactions energy values, and even minor structural details may significantly impact the measurements. Ranking results for identified models with RMSD values above 4 Å show a similar performance to the AA MD-based and CABS-dock scoring approaches. Both methods were able to identify a similar number of best-scored models in analyzed RMSD intervals.

The AA MD-based scoring procedure showed a much better performance for scoring of sets of top 100 models. In the RMSD range from 1 to 4 Å, the significant majority of models presenting lower RMSD values were identified using the proposed scoring procedure. Besides, for less accurate complex structures presenting RMSD values above 4 Å, for about 2/3 of all analyzed cases, MD-based scoring was also able to deliver models with lower RMSD values.

The scoring results obtained using experimental receptor structures in the modeled complexes strongly suggest the need to improve all atom models' reconstruction techniques and/or optimization strategies. In our experience, the improvement can be conducted primarily by enhancing the accuracy of backbone reconstruction since it significantly affects the subsequent side-chain reconstruction and energy-based scoring, as demonstrated in protein structure prediction [21,22]

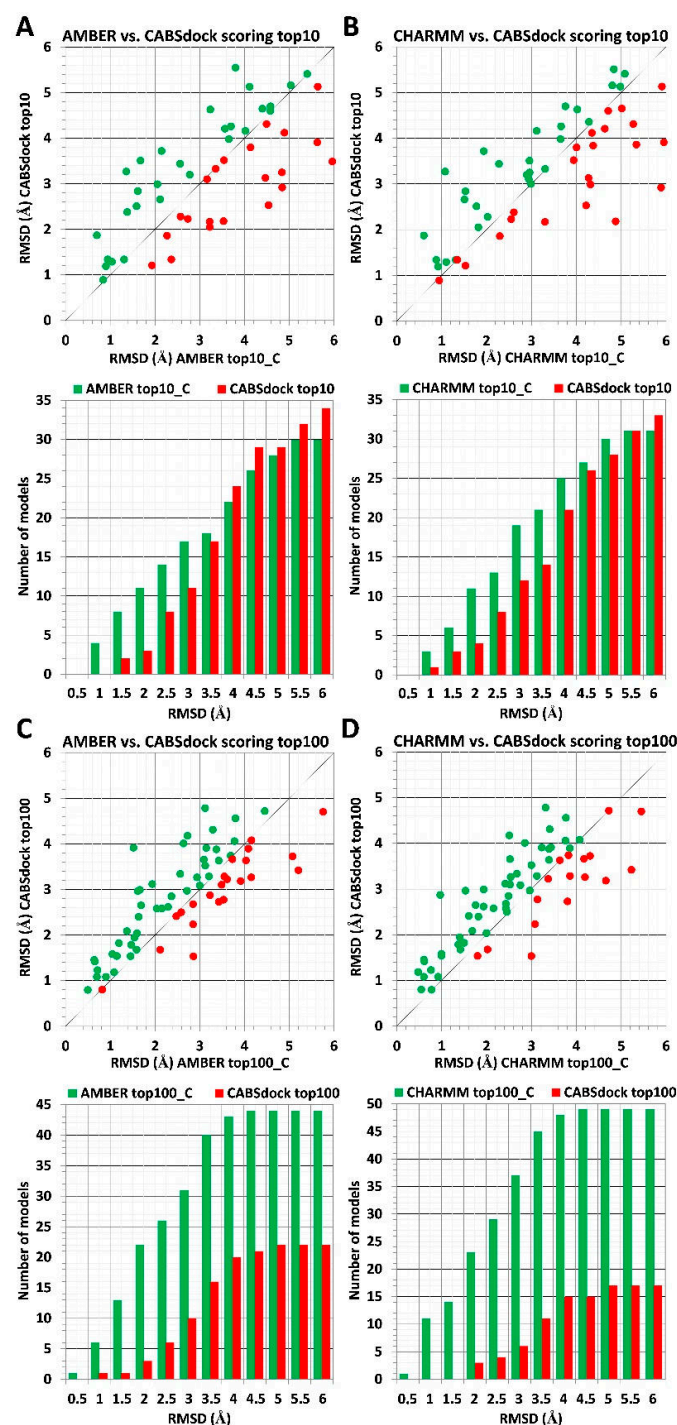


Figure 3. Comparison of MD-based versus CABS-dock scoring after substitution of a protein receptor with its crystal structure. The scoring data for 66 different protein–peptide complexes is shown. MD-based scoring was performed using two different force fields: AMBER and CHARMM. Interaction energy for each complex model was calculated using receptor–protein conformation derived from its crystal structure used in the CABS-dock benchmark (see Materials and Methods section for details). Panels (A,B) show results obtained for sets of top 10 models, and panels (C,D) show results for sets of top 100 models. The plot at the top of each panel (A–D) shows RMSD values for best rank models according to the MD scoring method (color green) vs. RMSD values obtained for CABS-dock models (color red). At the bottom of each panel, the bar plot shows numbers of models with lower RMSD values identified using either the MD scoring method (color green) or using the CABS-dock scoring scheme (color red) for 0.5 Å intervals.

3. Materials and Methods

3.1. The CABS-Dock Data Set

CABS-dock is an efficient method for global docking of peptides to known (or approximately known) receptor–protein structures. The CABS-dock is available as the web-server [4,5] or the standalone application [23]. In recent years, CABS-dock has been applied or extended to different modeling tasks. These include peptide docking with large structural changes of the receptor structure and disordered structures [7,8], docking with protein–peptide contact information [20], docking of peptides to GPCR structures [24] (see also recent review on CABS-dock development and applications [6]), and identification of peptide cleavage sites for protease–substrate systems [9].

For the docking simulations, CABS-dock uses the CABS CG model [25] based on an efficient Monte Carlo dynamics sampling scheme and knowledge-based statistical potentials describing interactions of amino acid united atoms (2–4 united atoms representing single residues, see review [26]). The water environment is treated in an implicit fashion where the solvent effects are encoded in burial and contact potentials describing interactions of united atoms. Such knowledge-based potentials work surprisingly well, although such a model of interactions is certainly imprecise and biased by “average” docking conditions. Consequently, while a subset of preferable docking results can be identified with an acceptable credibility, the best docking results are difficult to identify. Therefore, a better ranking of the docking results is highly desirable.

By default, CABS-dock generates 10,000 predictions in C-alpha trace representation of protein–peptide structures. The CABS-dock scoring procedure leads to the sets of 10 or 100 top-scored models and consists of two steps. First, 1000 top-scored models are selected based on CABS CG energy values (100 models having the lowest interaction energy are selected from every 10 replicas of a simulated system). Second, a set of 1000 top-scored models is subjected to structural clustering using a k-medoids algorithm, with the number *k* of clusters equal to 10 or 100, thus leading to 10 or 100 top-scored models.

In this work, we used the sets of models obtained from the CABS-dock docking protocol described by Blaszczyk et al. [20]. The protocol was validated on the PeptiDB benchmark set and the accuracy of individual docking cases was described in the Supplementary Information of the work [20]. Our working hypothesis is that MD-based scoring can distinguish high-accuracy models from medium-resolution models. To test this hypothesis, we selected the set of 66 different protein–peptide complexes presenting substantial content (at least a few or more) of high-accuracy models in the set of 1000 top-scored low energy models from the bound dataset. For each protein–peptide complex, the set of 1000 top-scored structures in C-alpha trace format was used as the input for the scoring procedure. For the list of all protein–peptide complexes (including PDB IDs) used in this study see Table S1.

3.2. Molecular Dynamics Scoring of Docking Results

The major stages of the scoring procedure are shown in Figure 4. The scoring has been conducted for 66 different protein–peptide complexes using the sets of 1000 molecular models predicted for each complex. First, complex structures in each set of models were reconstructed from C-alpha traces to their atomistic structures. The reconstruction was conducted using a Modeller-based protocol [19,27]. Next, for each analyzed complex, protein–peptide interaction energy was estimated for all models during short MD simulation using three different force-fields, two of the atomic resolution, CHARMM [28] and AMBER [29], and the CG MARTINI force-field [30]. Initially, each reconstructed AA structure of a complex was inserted in the cubic simulation box and solvated with water molecules. The simulation box dimensions were assigned for each system, considering the protein–peptide complex's size to prevent interaction of simulated molecules with its periodic images. The specific number of ions was added to neutralize the system charge necessary for the particle mesh Ewald (PME) procedure [31] used to calculate electrostatic interactions.

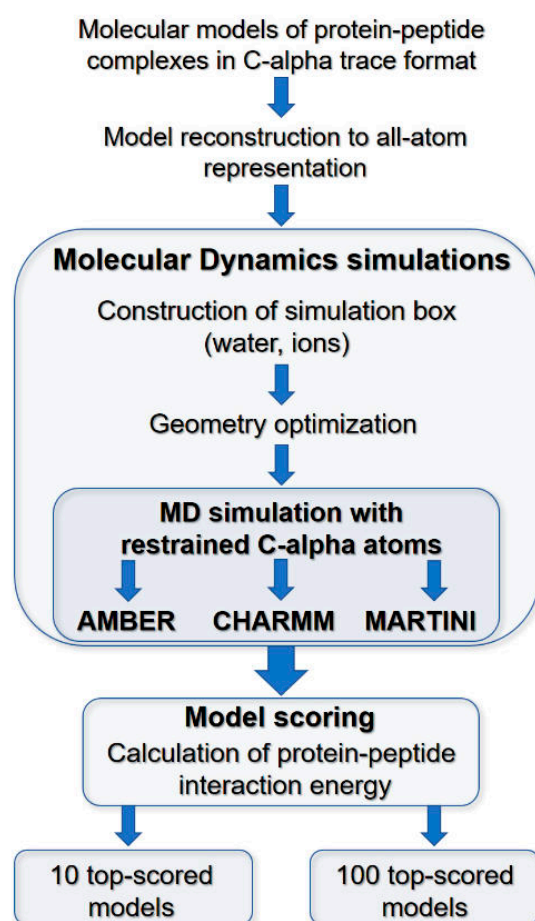


Figure 4. Main stages of the model scoring procedure. Scoring procedure includes: (i) reconstruction of C-trace models to AA representation; (ii) MD simulations of protein–peptide complexes in water environment using CHARMM, AMBER, and MARTINI parametrization; (iii) model scoring: calculation of protein–peptide interaction energy and RMSD values for bound peptide ligands.

CHARMM and AMBER-based scoring of AA models was conducted using the following procedure: First, the geometry of the system was optimized during 2000 steps of steepest descent algorithm. During the optimization, C-alpha atoms were kept at their initial positions to prevent distortions of the analyzed structure. In the next step, 100 ps MD simulations were conducted with position restraints using force constant 1000 kJ/(mol·nm²) imposed on C-alpha atoms of both the receptor protein and the peptide ligand. The simulation time step was 2 fs. Trajectory frames were recorded every 2 ps. The periodic boundary conditions were used. The temperature was set to 300 K and pressure was maintained at 1013 hPa. Long-range electrostatic interactions were approximated by the PME method [31] with a cutoff of 1.2 nm and hydrogen bonds were restrained using the LINCS algorithm [32]. In the last step, interaction energies were estimated for each of the last 25 frames of the recorded trajectory (which corresponds to the simulation period of 50 to 100 ps) as the sum of the energy of electrostatic and van der Waals interactions between protein and bound peptide. Average values of interaction energy were used to rank the analyzed model. The structure of the simulated complex derived from the last frame of the simulation trajectory was superimposed on its crystal structure (based on coordinates of receptor protein atoms only) and RMSD value was calculated for C-alpha atoms of the bound peptide.

A similar procedure for calculating the protein–peptide interaction energy described above was used for systems simulated in CG representation using the MARTINI model. Systems in atomistic coordinates were translated into CG representation and corresponding topologies were created using the martinize.py program. After optimization of the system geometry (1000 steps of the steepest descent algorithm), MD simulation was conducted with

position restraints imposed on BB united atoms using force constant 1000 kJ/(mol*nm²). Due to the CG representation of the MARTINI model, a longer simulation step of 20 fs was used and the simulation for each analyzed model lasted 10 ns. A set of simulation parameters dedicated to the MARTINI force field was applied [33]. Trajectory frames were recorded every 20 ps. Simulations were conducted at temperature 310 K. Receptor–peptide interaction energy was calculated based on 250 trajectory frames recorded during the simulation time from 5 to 10 ns. The structure of the protein–peptide complex resulting from MD simulation was compared to the crystal structure, and RMSD value was calculated for BB united atoms of the bound peptide.

Calculated interaction energy and RMSD values for each model were used to rank sets of molecular models for 66 protein–peptide complexes. Results were presented for the best model in sets of top 100 models (model presenting the lowest RMSD value for peptide molecule out of 100 models scored with the lowest interaction energy value) and for the best model in the set of top 10 models (model presenting the lowest RMSD value for peptide molecule out of 10 models scored with the lowest interaction energy value). Finally, ranking results were compared to CABS-dock scoring (the lowest RMSD values for the sets of top 100 models and sets of top 10 models) based on structural clustering of the set of 1000 generated models for each protein–peptide complex (the CABS-dock data set). All simulations and data analysis were performed using the Gromacs program suit version 5.1.4 [34]. The VMD software version 1.9.4 was used for visualization of the obtained complex structures.

To investigate how the reconstruction quality of the receptor structure may impact the scoring results, we repeated the whole scoring procedure using AA force-fields for all 66 systems with one significant modification. In each scored complex model (resulting from CABS-dock docking simulation), the protein receptor structure was replaced by its crystal structure (derived from PDB database and also initially used as input for CABS-dock docking procedure). Scoring results were compared to previously collected data.

4. Conclusions

Efficient molecular docking of peptide ligands to protein receptors is a highly expected and necessary tool for a deeper understanding of many important molecular biology problems and, in particular, for the efficient design of new peptide-based drugs and therapies. Classical molecular docking tools have proven to be very efficient in studies of the interaction of small ligands with proteins and protein complexes, provided that the receptor structure is not significantly changed upon the docking and the binding site is approximately known. Peptides are specific ligands because they can adapt to a huge number of conformations and binding poses. Due to the large conformational space that needs to be explored, free docking of peptide ligands based on classical MD tools is now not possible. Only a significant simplification of the problem, based on CG modeling approaches, opens a possibility for attempting fully unrestrained docking of peptides. CABS-dock is such a tool that has proven to be quite effective in unrestrained and fully flexible docking of even quite large peptides (or small proteins) to protein receptors. CABS-dock docking generates huge numbers of possible structures for a relatively low computational cost, although selecting the best results is not easy at the CG modeling stage (due to the low resolution of the produced models and due to the smoothened energy surface). Therefore, a kind of multiscale ranking protocol is needed that merges CG with more precise AA modeling.

In summary, we proposed and tested a multiscale modeling protocol that allows more efficient use of the CABS-dock results for a very moderate cost of short MD simulations. The proposed re-ranking of the CABS-dock results works well for high and moderate resolution models; however, it is less efficient for the generally less accurate solutions. Analysis of the re-ranking results also indicated the necessity of very accurate AA reconstruction of the CG models. Additionally, it should be pointed out that the presented modeling protocol can be easily extended by unrestrained AA MD simulations aiming at long-timescale refinement

of protein–peptide complexes. This is, however, much more expensive computationally and beyond the scope of the present work.

Supplementary Materials: The following are available online, Figure S1: Graphical representation of scoring results for 66 protein–peptide complexes, Table S1: Model scoring results obtained for 66 protein–peptide complexes.

Author Contributions: Conceptualization, M.K.; methodology, M.K. and M.Z.; formal analysis, M.Z.; investigation, M.Z. and M.K.; writing—original draft preparation, M.K. and S.K.; writing—review and editing, M.K. and S.K.; visualization, M.K. and M.Z.; supervision, M.K. and S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in supplementary material.

Conflicts of Interest: The authors declare no conflict of interest.

Sample Availability: Samples of the compounds are not available from the authors.

References

- Henninot, A.; Collins, J.C.; Nuss, J.M. The Current State of Peptide Drug Discovery: Back to the Future? *J. Med. Chem.* **2018**, *61*, 1382–1414. [[CrossRef](#)] [[PubMed](#)]
- Ciemny, M.; Kurcinski, M.; Kamel, K.; Kolinski, A.; Alam, N.; Schueler-Furman, O.; Kmiecik, S. Protein–peptide docking: Opportunities and challenges. *Drug Discov. Today* **2018**, *23*, 1530–1537. [[CrossRef](#)] [[PubMed](#)]
- Raveh, B.; London, N.; Schueler-Furman, O. Sub-angstrom modeling of complexes between flexible peptides and globular proteins. *Proteins Struct. Funct. Bioinform.* **2010**, *78*, 2029–2040. [[CrossRef](#)]
- Kurcinski, M.; Jamroz, M.; Blaszczyk, M.; Kolinski, A.; Kmiecik, S. CABS-dock web server for the flexible docking of peptides to proteins without prior knowledge of the binding site. *Nucleic Acids Res.* **2015**, *43*, W419–W424. [[CrossRef](#)] [[PubMed](#)]
- Blaszczyk, M.; Kurcinski, M.; Kouza, M.; Wieteska, L.; Debinski, A.; Kolinski, A.; Kmiecik, S. Modeling of protein–peptide interactions using the CABS-dock web server for binding site search and flexible docking. *Methods* **2016**, *93*, 72–83. [[CrossRef](#)]
- Kurcinski, M.; Badaczewska-Dawid, A.; Kolinski, M.; Kolinski, A.; Kmiecik, S. Flexible docking of peptides to proteins using CABS-dock. *Protein Sci.* **2020**, *29*, 211–222. [[CrossRef](#)] [[PubMed](#)]
- Ciemny, M.P.; Debinski, A.; Paczkowska, M.; Kolinski, A.; Kurcinski, M.; Kmiecik, S. Protein–peptide molecular docking with large-scale conformational changes: The p53-MDM2 interaction. *Sci. Rep.* **2016**, *6*, 37532. [[CrossRef](#)]
- Ciemny, M.P.; Badaczewska-Dawid, A.E.; Pikuzinska, M.; Kolinski, A.; Kmiecik, S. Modeling of Disordered Protein Structures Using Monte Carlo Simulations and Knowledge-Based Statistical Force Fields. *Int. J. Mol. Sci.* **2019**, *20*, 606. [[CrossRef](#)]
- Kolinski, M.; Kmiecik, S.; Dec, R.; Piejko, M.; Mak, P.; Dzwolak, W. Docking interactions determine early cleavage events in insulin proteolysis by pepsin: Experiment and simulation. *Int. J. Biol. Macromol.* **2020**, *149*, 1151–1160. [[CrossRef](#)]
- Li, J.; Fu, A.; Zhang, L. An Overview of Scoring Functions Used for Protein–Ligand Interactions in Molecular Docking. *Interdiscip. Sci. Comput. Life Sci.* **2019**, *11*, 320–328. [[CrossRef](#)]
- London, N.; Raveh, B.; Cohen, E.; Fathi, G.; Schueler-Furman, O. Rosetta FlexPepDock web server—High resolution modeling of peptide–Protein interactions. *Nucleic Acids Res.* **2011**, *39*, W249–W253. [[CrossRef](#)]
- Van Zundert, G.; Rodrigues, J.; Trellet, M.; Schmitz, C.; Kastiris, P.; Karaca, E.; Melquiond, A.; van Dijk, M.; de Vries, S.; Bonvin, A. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J. Mol. Biol.* **2016**, *428*, 720–725. [[CrossRef](#)] [[PubMed](#)]
- Jiménez-García, B.; Pons, C.; Fernández-Recio, J. pyDockWEB: A web server for rigid-body protein–protein docking using electrostatics and desolvation scoring. *Bioinformatics* **2013**, *29*, 1698–1699. [[CrossRef](#)] [[PubMed](#)]
- Schindler, C.E.; de Vries, S.J.; Zacharias, M. Fully Blind Peptide–Protein Docking with pepATTRACT. *Structure* **2015**, *23*, 1507–1515. [[CrossRef](#)] [[PubMed](#)]
- Schneider, M.; Rosam, M.; Glaser, M.; Patronov, A.; Shah, H.; Back, K.C.; Daake, M.A.; Buchner, J.; Antes, I. BiPPred: Combined sequence-and structure-based prediction of peptide binding to the Hsp70 chaperone BiP. *Proteins Struct. Funct. Bioinform.* **2016**, *84*, 1390–1407. [[CrossRef](#)]
- Alam, N.; Schueler-Furman, O. Modeling Peptide–Protein Structure and Binding Using Monte Carlo Sampling Approaches: Rosetta FlexPepDock and FlexPepBind. In *Methods in Molecular Biology*; Springer: Berlin/Heidelberg, Germany, 2017; Volume 1561, pp. 139–169.
- Lensink, M.F.; Velankar, S.; Wodak, S.J. Modeling protein–protein and protein–peptide complexes: CAPRI 6th edition. *Proteins Struct. Funct. Bioinform.* **2016**, *85*, 359–377. [[CrossRef](#)]

18. Yu, J.; Andreani, J.; Ochsenbein, F.; Guerois, R. Lessons from (co-)evolution in the docking of proteins and peptides for CAPRI Rounds 28–35. *Proteins Struct. Funct. Bioinform.* **2017**, *85*, 378–390. [[CrossRef](#)]
19. Webb, B.; Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Bioinform.* **2016**, *54*, 5–6. [[CrossRef](#)]
20. Blaszczyk, M.; Ciemny, M.P.; Kolinski, A.; Kurcinski, M.; Kmiecik, S. Protein-peptide docking using CABS-dock and contact information. *Brief. Bioinform.* **2018**, *20*, 2299–2305. [[CrossRef](#)] [[PubMed](#)]
21. Badaczewska-Dawid, A.E.; Kolinski, A.; Kmiecik, S. Computational reconstruction of atomistic protein structures from coarse-grained models. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 162–176. [[CrossRef](#)] [[PubMed](#)]
22. Kmiecik, S.; Gront, D.; Kolinski, A. Towards the high-resolution protein structure prediction. Fast refinement of reduced models with all-atom force field. *BMC Struct. Biol.* **2007**, *7*, 1–11. [[CrossRef](#)]
23. Kurcinski, M.; Ciemny, M.P.; Oleniecki, T.; Kuriata, A.; Badaczewska-Dawid, A.E.; Kolinski, A.; Kmiecik, S. CABS-dock standalone: A toolbox for flexible protein-peptide docking. *Bioinformatics* **2019**, *35*, 4170–4172. [[CrossRef](#)]
24. Badaczewska-Dawid, A.E.; Kmiecik, S.; Koliński, M. Docking of peptides to GPCRs using a combination of CABS-dock with FlexPepDock refinement. *Brief. Bioinform.* **2021**, *22*. [[CrossRef](#)]
25. Kolinski, A. Protein modeling and structure prediction with a reduced representation. *Acta Biochim. Pol.* **2004**, *51*, 349–371. [[CrossRef](#)]
26. Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A.E.; Kolinski, A. Coarse-Grained Protein Models and Their Applications. *Chem. Rev.* **2016**, *116*, 7898–7936. [[CrossRef](#)]
27. Badaczewska-Dawid, A.E.; Khramushin, A.; Kolinski, A.; Schueler-Furman, O.; Kmiecik, S. Protocols for All-Atom Reconstruction and High-Resolution Refinement of Protein–Peptide Complex Structures. In *Methods in Molecular Biology*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 273–287.
28. Huang, J.; MacKerell, A.D., Jr. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *J. Comput. Chem.* **2013**, *34*, 2135–2145. [[CrossRef](#)]
29. Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J.L.; Dror, R.O.; Shaw, D.E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins Struct. Funct. Bioinform.* **2010**, *78*, 1950–1958. [[CrossRef](#)]
30. Marrink, S.J.; Tieleman, D.P. Perspective on the Martini model. *Chem. Soc. Rev.* **2013**, *42*, 6801–6822. [[CrossRef](#)] [[PubMed](#)]
31. Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092. [[CrossRef](#)]
32. Hess, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4*, 116–122. [[CrossRef](#)] [[PubMed](#)]
33. Monticelli, L.; Kandasamy, S.K.; Periole, X.; Larson, R.G.; Tieleman, D.P.; Marrink, S. The MARTINI Coarse-Grained Force Field: Extension to Proteins. *J. Chem. Theory Comput.* **2008**, *4*, 819–834. [[CrossRef](#)] [[PubMed](#)]
34. Abraham, M.J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J.C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1–2*, 19–25. [[CrossRef](#)]

3.4. [A4] Dokowanie białko-peptyd z wykorzystaniem modelu językowego ESMFold

3.4.1. Wprowadzenie do [A4]

W publikacji [A4] skoncentrowałem się na analizie możliwości zastosowania językowego modelu białkowego ESMFold³³ do dokowania peptydów do białek. Głównym celem pracy było zweryfikowanie, czy ESMFold - model opracowany pierwotnie do predykcji struktur pojedynczych łańcuchów białkowych - może być skutecznie stosowany w zadaniach dokowania peptydów oraz jak jego skuteczność wypada w porównaniu z wiodącymi obecnie metodami opartymi na narzędziu AlphaFold. W realizacji tego projektu odegrałem kluczową rolę - zaproponowałem koncepcję wykorzystania ESMFold do dokowania peptydów, zaimplementowałem i przetestowałem różne warianty opracowanej procedury, przeprowadziłem obliczenia i analizy porównawcze oraz przygotowałem wizualizacje i dokonałem interpretacji wyników. Opracowałem również pierwszą wersję manuskryptu i uczestniczyłem w jego dalszej redakcji.

Punktem wyjścia dla tej pracy była obserwacja gwałtownego rozwoju metod AI do przewidywania struktur białkowych^{4,7,99–103}. Jak omówiłem szerzej w rozdziale 2.2.3, AlphaFold-Multimer oraz AlphaFold3 osiągnęły znakomite wyniki w zadaniach modelowania kompleksów, jednak ich efektywność wiązała się ze znacznymi kosztami obliczeniowymi oraz ograniczeniami związanymi z wymaganiami wstępnego przetwarzania danych (np. koniecznością obliczania MSA). ESMFold, jako model wytrenowany wyłącznie na sekwencjach białkowych i niewymagający baz danych struktur ani wielokrotnych ułiniowień sekwencji, wydawał się potencjalnie atrakcyjną alternatywą, szczególnie w kontekście szybkiego dokowania peptydów o ograniczonej dostępności danych strukturalnych.

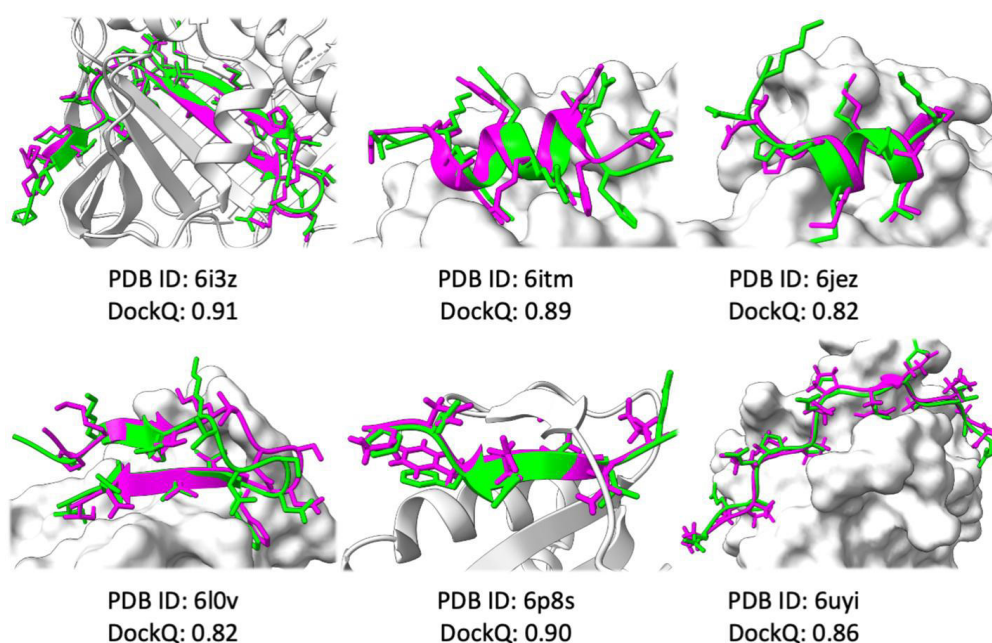
Do przeprowadzenia badania wykorzystałem dwa zestawy testowe, zdefiniowane wcześniej w literaturze. Pierwszy zestaw (*Dataset 1*), pochodzący ze zbioru z pracy Wallner i wsp.¹⁰⁴, obejmował 112 struktur kompleksów białko-peptyd dobranych według klasyfikacji ECOD. Drugi zbiór (*Dataset 2*), zaczerpnięty z pracy Manshour i wsp.¹⁰⁵, zawierał 60 struktur niedostępnych w zbiorze uczącym AlphaFold3. Program ESMFold został zaprojektowany do przewidywania struktur pojedynczych łańcuchów białkowych, dlatego nie umożliwia bezpośredniego modelowania kompleksów multimerycznych. W celu przewidzenia struktury kompleksu białko-peptyd zastosowałem podejście jednołańcuchowe, wcześniej wykorzystane m.in. w pracach z użyciem AlphaFold¹⁰⁶. Polega ono na połączeniu sekwencji receptora i peptydu za pomocą elastycznego glicynowego linkera, który następnie został usunięty z finalnego modelu.

Moim celem było nie tylko przetestowanie narzędzia ESMFold pod kątem przewidywania kompleksów białko-peptyd, ale także przeanalizowanie wpływu wybranych modyfikacji protokołu na skuteczność modelowania. W pracy przetestowałem dokowanie z użyciem

domyślnych ustawień narzędzia ESMFold, ale również dokowanie ze strategią maskowania losowych reszt w sekwencji wejściowej, a także warianty adaptacyjne ze zmieniającą się liczbą cykli oraz różnymi konfiguracjami i długością linkera.

Analiza wyników wykazała, że domyślna konfiguracja ESMFold prowadzi do stosunkowo niskiej skuteczności - jedynie 14% przypadków w Dataset 1 spełniło minimalne kryterium $\text{DockQ} \geq 0.23$. Istotnym problemem była tendencja modelu do generowania konformacji, w których peptyd nie wchodził w kontakt z receptorem. Wprowadzenie losowego maskowania pozwoliło znacząco poprawić wyniki: skuteczność wzrosła do 21% przypadków spełniających kryterium $\text{DockQ} \geq 0.23$ w Dataset 1 i do 28% w Dataset 2. Najlepsze wyniki udało mi się uzyskać korzystając ze strategii losowego maskowania, która pozwoliła na wygenerowanie ośmiu różnorodnych konformacji, ustawiając długość linkerów na 30 reszt aminokwasowych oraz wykorzystując opracowaną przeze mnie metodę oceny pozwalającą na wybór najlepszego z wygenerowanych przeze mnie modeli.

Wyniki uzyskane dla pierwszego zbioru struktur porównałem bezpośrednio z wynikami otrzymanymi dla narzędzia AlphaFold-Multimer przy zastosowaniu wymuszonego próbkowania¹⁰⁴. ESMFold uzyskał zauważalnie niższą skuteczność w zakresie modeli wysokiej jakości ($\text{DockQ} \geq 0.8$), natomiast pojedyncze przypadki wykazały wyraźną przewagę ESMFold nad AlphaFold-Multimer, szczególnie w układach trudnych do modelowania dla narzędzi AI. Korzystając z drugiego zbioru struktur porównałem skuteczność zaproponowanego przeze mnie protokołu z narzędziami AlphaFold3, AlphaFold-Multimer oraz ColabFold - ESMFold osiągał niższy odsetek modeli wysokiej jakości, jednak odsetek modeli akceptowalnych ($\text{DockQ} \geq 0.23$) był podobny do otrzymanego dla pierwszego zbioru struktur. Przykłady poprawnie przewidzianych kompleksów, w których konformacja peptydu została dobrze odtworzona przy użyciu protokołu opartego na ESMFold, przedstawiłem na Rysunku 6.



Rysunek 6. Przykłady wysokiej jakości wyników dokowania białko-peptyd uzyskanych z użyciem metody ESMFold (na podstawie [A4]). Powierzchnia przedstawia receptor białkowy; peptyd w strukturze eksperymentalnej zaznaczono kolorem magenta, a przewidziany przez ESMFold - kolorem zielonym. Widoczna zgodność obu konformacji pokazuje, że mimo ogólnych ograniczeń skuteczności ESMFold potrafi generować modele o jakości porównywalnej z danymi eksperymentalnymi. Źródło: [A4].

Z punktu widzenia praktycznego zastosowania, kluczową zaletą ESMFold była jego szybkość. Symulacje pojedynczego kompleksu trwały średnio od 20 do 60 sekund. ESMFold nie wymagał generowania MSA ani przeszukiwania baz danych, co czyniło go potencjalnie przydatnym w zadaniach przesiewowych lub do szybkiej oceny dużych zbiorów sekwencji.

W pracy [A4] zwróciłem uwagę również na ograniczenia testowanej metody: częste tworzenie modeli, w których peptyd nie kontaktował się z receptorem, niską skuteczność w identyfikacji modeli wysokiej rozdzielczości oraz silną zależność wyników od przyjętego protokołu maskowania i konfiguracji wejściowych. Mimo że ESMFold nie dorównuje dokładnością metodom AlphaFold3 czy AlphaFold-Multimer, okazał się wartościowym uzupełnieniem narzędzi do dokowania - szczególnie w zastosowaniach wymagających dużej szybkości i możliwości generowania modeli bez sekwencji homologicznych..

Uważam, że wyniki zaprezentowane w pracy [A4] pokazują, że nawet model językowy o niskich wymaganiach obliczeniowych, jakim jest ESMFold, może wnieść nową jakość do modelowania kompleksów białko-peptyd. Choć nie zastąpi on najbardziej zaawansowanych rozwiązań z rodziny AlphaFold, to jego zdolność do szybkiego generowania sensownych modeli otwiera możliwość zupełnie nowego sposobu pracy - szybkiego przeglądu dużych zbiorów sekwencji, formułowania wstępnych hipotez strukturalnych i poszukiwania nietypowych sposobów wiązania, które mogą umknąć cięższym, kosztownym obliczeniowo metodom. Co istotne, praca ta odbiła się szerokim echem w środowisku naukowym i była aktywnie dyskutowana w mediach społecznościowych - moim zdaniem dlatego, że jako jedna z pierwszych wyraźnie pokazała potencjał modeli językowych, do tej pory kojarzonych głównie z predykcją struktur pojedynczych białek, w zadaniach znacznie bardziej złożonych. Wskazuje to na otwarcie nowego kierunku badań na styku sztucznej inteligencji i modelowania molekularnego, który w przyszłości może istotnie zmienić sposób, w jaki podchodzimy do problemu dokowania.

3.4.2. Artykuł [A4]

Protein–Peptide Docking with ESMFold Language Model

Mateusz Zalewski, Björn Wallner, and Sebastian Kmieciak*

Cite This: *J. Chem. Theory Comput.* 2025, 21, 2817–2821

Read Online

ACCESS |



Metrics & More

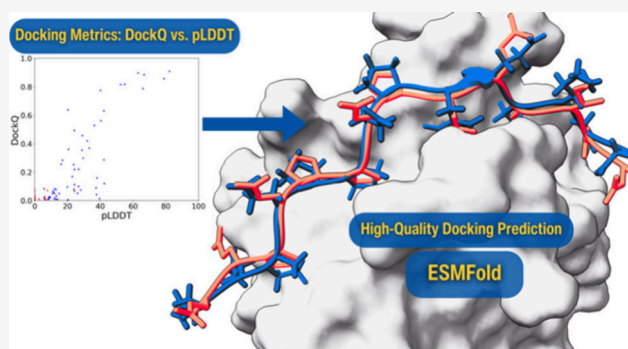


Article Recommendations



Supporting Information

ABSTRACT: Designing peptide therapeutics requires precise peptide docking, which remains a challenge. We assessed the ESMFold language model, originally designed for protein structure prediction, for its effectiveness in protein–peptide docking. Various docking strategies, including polyglycine linkers and sampling-enhancing modifications, were explored. The number of acceptable-quality models among top-ranking results is comparable to traditional methods and generally lower than AlphaFold-Multimer or AlphaFold 3, though ESMFold surpasses it in some cases. The combination of result quality and computational efficiency underscores ESMFold's potential value as a component in a consensus approach for high-throughput peptide design.



INTRODUCTION

Protein–peptide interactions are essential to many biological processes, or can play a key role when designed as peptide therapeutics. A comprehensive understanding of these interactions is crucial for unraveling disease mechanisms and guiding the rational design of therapeutics. Peptides, with their high specificity, selectivity, and favorable safety profiles, stand out as unique therapeutic candidates compared to small molecules and antibodies due to their natural origin and minimal toxicity.¹

While traditional experimental studies of protein–peptide complexes using X-ray crystallography and NMR spectroscopy are insightful, they are also expensive, time-consuming, and face scalability issues.² These challenges have spurred efforts toward computational methods in rational drug design, which leverage computer modeling to understand and predict protein structures and interactions.³ Despite the complexities presented by the flexible and disordered nature of peptides, these computational techniques are increasingly important for drug development.

In recent years, structural bioinformatics has been transformed by AlphaFold2 (AF2), an AI tool developed by DeepMind that achieved near-experimental accuracy in protein structure prediction.⁴ AF2's applications soon expanded to include protein–peptide interaction modeling. Initially, this involved using a monomeric version of AF2 with a polyglycine linker to connect peptides and proteins, effectively modeling peptide–protein complexes.⁵ Later efforts shifted to AlphaFold-Multimer (AFM), an AF2 extension trained on protein complexes.⁶ AFM docking, including forced sampling,⁷ achieved 59% accuracy in protein–peptide docking, outperforming traditional methods like ZDOCK,⁸ CABS-dock,⁹

PIPER-FlexPepDock,¹⁰ and InterPep2.¹¹ When benchmarked against other methods, AFM showed 53% accuracy, which increased to 60% when combined with AutoDockCrankPep,¹² highlighting the potential of hybrid approaches for improving docking predictions and enhancing the study of protein–peptide interactions.

Following the strides made by AlphaFold, ESMFold emerged as another robust tool for protein structure prediction, utilizing a distinct approach.¹³ Unlike AlphaFold, which relies heavily on extensive databases of protein structures and sequence alignments, ESMFold employs embeddings from protein language models (pLMs) derived from vast sequences. This method allows ESMFold to excel particularly in scenarios where limited structural data exists, as it captures more generalized sequence features and patterns through its advanced language modeling techniques. This shift from reliance on direct structural analogs to leveraging learned sequence contexts enables ESMFold to offer unique advantages in predicting novel or less-characterized protein structures.

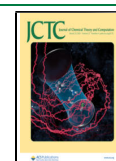
To our knowledge, ESMFold has not yet been tested specifically for protein–peptide docking. In this work, we investigate the potential of the ESMFold language model for this application and compare its performance with AlphaFold-

Received: November 21, 2024

Revised: March 2, 2025

Accepted: March 5, 2025

Published: March 7, 2025



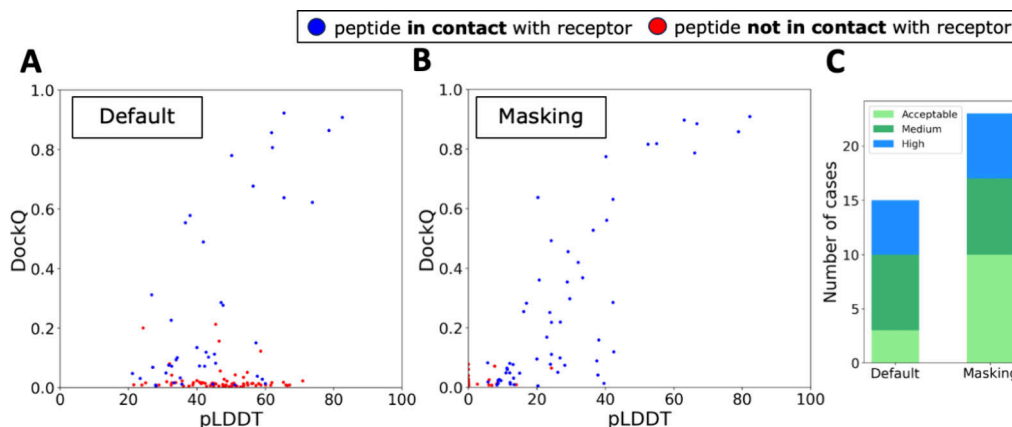


Figure 1. ESMFold docking results with default settings and enhanced sampling variants. (A, B) DockQ vs pLDDT scatter plots for different ESMFold simulation variants: (A) default settings and (B) using a masking approach. Blue dots indicate peptides in contact with the receptor (within 8 Å), while red dots indicate those not in contact (over 8 Å). pLDDT values are the average for each peptide, with (B) showing the weighted mean pLDDT for the top-ranked model out of eight generated per complex. (C) Distribution of high-quality (DockQ ≥ 0.8), medium-quality ($0.5 \leq \text{DockQ} < 0.8$), and acceptable ($0.23 \leq \text{DockQ} < 0.5$) docking models for different ESMFold simulation approaches (default, random masking) across 111 complexes, with colors indicating quality.

Multimer with an improved sampling protocol⁷ and AlphaFold 3.¹⁴

MATERIALS AND METHODS

Dataset. In this study, to evaluate ESMFold's performance in predicting protein–peptide interactions and to compare it with AlphaFold-Multimer (AFM) augmented by an extended sampling approach,⁷ we used the dataset created and described by Wallner et al.⁷ This dataset was assembled by selecting one representative structure per ECOD (Evolutionary Classification of Protein Domains)¹⁵ family from the test set in Lei et al.¹⁶ Referred to here as Dataset 1, it comprises 112 experimental protein–peptide complex structures obtained from the PDB¹⁷ (see [Supplementary Table 1](#)). The 6UEB structure was excluded due to memory constraints in the Colab environment.

Additionally, to compare our results with AlphaFold 3 (AF3) and other AlphaFold-based protocols, we used the dataset introduced by Manshour et al.¹⁴ Referred to here as Dataset 2, this benchmark consists of 60 protein–peptide structures deposited in the PDB after the AF3 training set cutoff date. Details are provided in the [Supporting Information](#) in section “Dataset 2”.

ESMFold. ESMFold is a deep-learning method that uses the ESM-2 protein language model to predict protein structures directly from amino acid sequences. Unlike AF2, ESMFold does not rely on external databases, template searches, or Multiple Sequence Alignments (MSA). It provides accuracy and resolution comparable to AlphaFold while being up to 60 times faster, depending on the sequence length.¹³ In this study, we used the version of ESMFold implemented in ColabFold notebook.¹⁸ Since ESMFold is designed for single-chain predictions, we performed protein–peptide docking by introducing a flexible linker (a polyglycine chain) between amino acid chains, similar to early AF2 applications.⁵ This linker was removed from the model after prediction.

Evaluation Metrics. To evaluate docking success, we used the DockQ score,¹⁹ which measures protein–peptide docking quality. The DockQ score, ranging from 0 to 1, combines CAPRI criteria:²⁰ LRMSD, IRMSD, and FNAT. Scores

between 0.23 and 0.5 are considered acceptable, 0.5 to 0.8 medium quality, and 0.8 or above high quality.

We assessed prediction confidence and reliability using the predicted Local Distance Difference Test (pLDDT) value from ESMFold. While pLDDT accuracy can vary based on the protein's characteristics, ESMFold predictions often reach experimental-like accuracy for high-confidence predictions.¹³ There is also a high correlation between pLDDT scores from ESMFold and AlphaFold.¹³ A prediction is considered high-confidence if the mean pLDDT value exceeds 70.¹³

RESULTS AND DISCUSSION

This study aimed to assess the viability of using ESMFold for molecular docking. It involves inserting a polyglycine linker between the receptor and peptide sequences to facilitate multichain structure predictions. In this work, we tested various modeling variants and reported the quality of the top-ranked models. The output from ESMFold was either a single model, or, if multiple models were available, we applied a scoring scheme based on the pLDDT confidence scores.

ESMFold Default Protocol. The initial run, performed using a 30-amino-acid polyglycine linker and default settings, produced a limited number of high-quality structures. Out of 111 cases, 15 structures met or exceeded the acceptable threshold with DockQ scores of ≥ 0.23 , among which only 5 were classified as high quality, as shown in [Figure 1A](#). This figure presents a scatter plot that illustrates the relationship between DockQ scores and pLDDT values.

Many instances placed the peptide more than 8 Å away from the receptor, indicating incorrect docking. These cases were relatively straightforward to identify and exclude. After discarding these, we retained 42 models. This results in an effective success rate of 36% for the viable models: 5 high quality, 7 medium quality, and 3 acceptable.

It is also important to note that predicting a protein–peptide complex of 252 residues (excluding the linker), which is the median length in our dataset, requires only 21 s on an A100 GPU when using a Colab notebook. Additionally, there is an initial setup time of approximately 3 min to install ESMFold on the Colab notebook, but this setup is required only once per session.

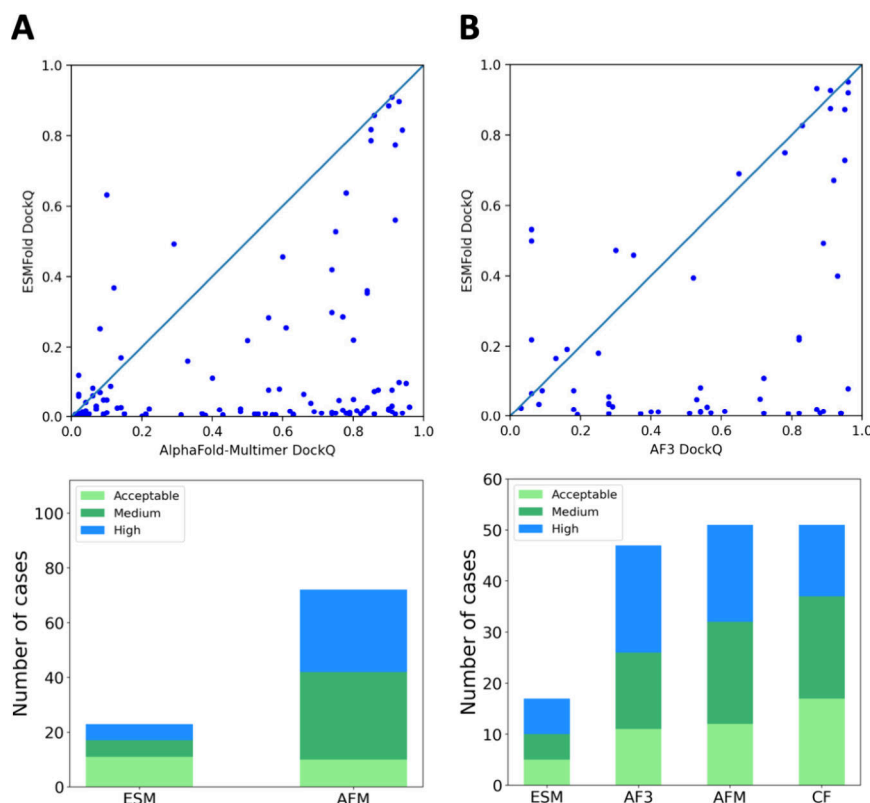


Figure 2. Comparison of ESMFold and AlphaFold-based docking tools on Dataset 1 (A) and Dataset 2 (B). The upper panels show DockQ scatter plots, comparing ESMFold with (A) AlphaFold Multimer (AFM) using enhanced sampling⁷ and (B) AlphaFold 3 (AF3).¹⁴ The lower panels present the distribution of high (DockQ ≥ 0.8), medium ($0.5 \leq \text{DockQ} < 0.8$), and acceptable ($0.23 \leq \text{DockQ} < 0.5$) quality models, comparing ESMFold with (A) AFM with enhanced sampling and (B) AFM, AF3, and ColabFold (CF). Data for AF-based tools are taken from previous studies.^{7,14}

Random Masking Approach. With default settings, a significant number of models failed to position the peptide in direct contact with the receptor (Figure 1A). To address this, we expanded our efforts to generate a larger pool of models. Despite the use of different seed values, ESMFold consistently produced nearly identical output models for each run, which highlighted the need for a strategy to enhance diversity in the generated structures. We adopted a random masking strategy²¹ that involves masking random residues within the input sequence. This approach, using a masking rate of 0.25 and generating 8 structures for each of the 111 complexes, encouraged the production of diverse structural predictions. This method yielded 27 acceptable or better structures, with 6 classified as high-quality from the best of the 8 generated per complex.

Analysis showed the greatest improvement in cases with a low but nonzero initial DockQ score (see Supplementary Figure 1). The challenge, however, lay in selecting the best structure among the 8 generated in each run. We experimented with various scoring strategies based on pLDDT values. Selecting structures with the lowest average pLDDT values across all peptide residues resulted in 20 structures being acceptable or better, including 5 high-quality ones (see Supplementary Figure 2B). A more targeted approach, focusing only on residues within 5 Å of the receptor, produced fewer but more relevant structures, with 2 being acceptable or better, and 5 high-quality (see Supplementary Figure 2C). The most effective method involved weighting the pLDDT scores by the proportion of residues in contact with the receptor,

assigning a score of zero to noncontacting residues (see Supplementary Figure 2A). This strategy yielded 23 acceptable or better structures, with 6 high-quality (see Figure 1B and Supplementary Figure 2D).

After discarding cases where the peptide was clearly misdocked (more than 8 Å away from the receptor surface), 57 cases remained. This results in an effective success rate of 40% for the viable models: 6 high-quality, 7 medium-quality, and 10 acceptable structures.

Performing random masking to generate 8 structures for a median dataset protein–peptide complex of 252 residues takes about 61 s on an A100 GPU, a significant increase from the 21 s required for a default run but still efficient for the task. It is important to remember that an additional setup time of about 3 min is required for installing ESMFold on a Colab notebook, but this installation is necessary only once per session.

Testing Adaptive Recycling and Various Linker Configurations. Apart from masking, we explored an alternative strategy to improve sampling. The default simulations utilized 3 recycles; we extended this to a maximum of 12 recycles by initially performing the standard 3 recycles and then continuing only if the peptide remained more than 8 Å from the receptor. This approach resulted in improved outcomes compared to the default settings but was less effective than masking. Detailed results are provided in Supplementary Figure 3.

Additionally, we explored various polyglycine linker configurations to enhance peptide–receptor contact, testing lengths of 30, 100, and 200 residues at different termini with

the masking approach. The results indicated that a 30-residue linker at the C-terminus was optimal, as longer or differently placed linkers did not significantly improve docking quality. Description of detailed results is available in [Supplementary Figure 4](#).

Comparison to AlphaFold-Based Tools. For comparisons with AlphaFold-based protocols described in this section, we used the best-performing ESMFold variant (random masking, 30-residue polyglycine linker, weighted pLDDT). First, on Dataset 1, we compared ESMFold to AlphaFold-Multimer with enhanced sampling, which produced 75 acceptable or better models (out of 112), while AlphaFold-Multimer v2.1.0 generated 66.⁷ ESMFold exhibited generally lower overall precision generating 23 acceptable or better structures, including 6 of high quality (see [Figure 2A](#), examples of high-quality models are presented in [Figure 3](#)). Notably,

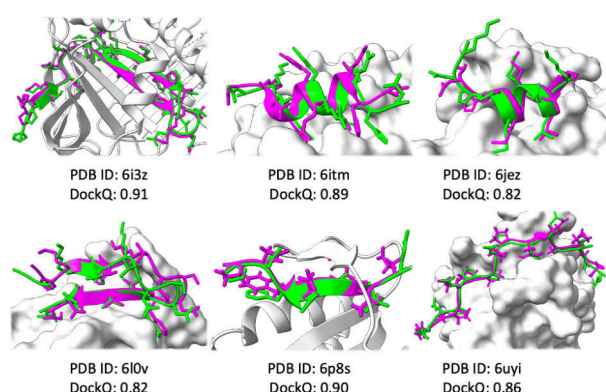


Figure 3. Example high-quality protein–peptide docking results using ESMFold. Protein receptor shown as surface: experimental peptide in magenta; predicted peptide in lime.

ESMFold successfully produced acceptable or better models in three instances where AlphaFold's predictions were of unacceptable quality (DockQ scores: ESMFold 0.25, 0.63, and 0.37 versus AlphaFold's 0.08, 0.10, and 0.12; see [Figure 2A](#)). Although primarily developed for single-sequence predictions and not trained on multimeric structures or short peptides, ESMFold's performance here is comparable to non-AI traditional methods^{8–11} (see performance data presented in the work⁷). This demonstrates the potential of evolutionary-scale models to predict biological properties from sequence patterns, suggesting that ESMFold could be further enhanced by including training data on multimeric complexes and short peptides.

Next, to compare our approach with the latest AlphaFold 3 (AF3), we drew upon a recent study by Manshour et al.,¹⁴ which comprehensively evaluated AlphaFold-Multimer, ColabFold, and AF3 for protein–peptide structure prediction. Using the Manshour et al. benchmark (Dataset 2), ESMFold achieved performance that was comparable to, or even better than, Dataset 1, reaching 28.3% acceptable or better models ([Figure 2B](#)) compared to 20.7% in Dataset 1 ([Figure 2A](#)). Notably, Manshour et al. found that AlphaFold-based tools yield similar rates of acceptable models overall, although AF3 typically generates a noticeably larger number of high-resolution structures than any other method. Consistent with our findings on Dataset 1, ESMFold also outperformed AF3 in several cases within Dataset 2.

CONCLUSIONS

In this study, we tested various parameters and settings of the ESMFold tool for protein–peptide docking applications. The optimal approach – combining random masking, a 30-residue polyglycine linker, and weighted pLDDT scoring—successfully yielded a significant fraction of acceptable-quality results (about 20.7% for Dataset 1 and 28.3% for Dataset 2, respectively). Although ESMFold did not match the overall accuracy of AlphaFold-based tools, its ability to produce acceptable models—and occasionally outperform AlphaFold-based tools—makes it a valuable component in a consensus-based approach. Furthermore, its speed—about 1 min per run—renders it particularly useful in high-throughput settings. Nevertheless, many of the generated models did not dock properly, underscoring the need for further improvements in ESMFold's docking accuracy.

In summary, ESMFold's ability to leverage sequence embeddings and identify crucial binding motifs shows significant promise for protein–peptide docking. Its rapid processing capability is particularly advantageous for applications requiring quick results, such as in peptide-based therapeutic development. Further development could enhance its utility, making it a robust complement to existing methods.

ASSOCIATED CONTENT

Data Availability Statement

The input FASTA sequences used for structure prediction are publicly accessible from the Protein Data Bank (PDB) database. All PDB codes corresponding to the dataset are provided in [Supporting Information \(SI\) Table 1](#). The predictions in this study were performed using the ESMFold model, which is available through ColabFold and can be accessed publicly on <https://github.com/sokrypton/ColabFold>. Additionally, a modified version of the ESMFold script that includes recycling for enhanced sampling (as described in this paper) as well as all the output structures generated during this study are available on our GitHub repository: <https://github.com/ZalewskiMa/ESMFold-docking>.

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jctc.4c01585>.

Additional figures and tables providing detailed analysis of protein–peptide docking with ESMFold. Supplementary Figures 1–5: DockQ distributions, scoring method comparisons, adaptive recycling results, and linker configuration effects. Supplementary Table 1: dataset of protein–peptide complexes used in the study ([PDF](#))

AUTHOR INFORMATION

Corresponding Author

Sebastian Kmiecik – Biological and Chemical Research Center, Faculty of Chemistry, University of Warsaw, 02-093 Warsaw, Poland; orcid.org/0000-0001-7623-0935; Email: sekmi@chem.uw.edu.pl

Authors

Mateusz Zalewski – Biological and Chemical Research Center, Faculty of Chemistry, University of Warsaw, 02-093 Warsaw, Poland

Björn Wallner – Department of Physics, Chemistry and Biology, Linköping University, Linköping S8 183, Sweden

Complete contact information is available at:
<https://pubs.acs.org/10.1021/acs.jctc.4c01585>

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We sincerely thank Negin Manshour for providing the data used in Figure 2. Mateusz Zalewski has been supported by the National Science Centre, Poland (Sheng 2021/40/Q/NZ2/00078). Sebastian Kmiecik has been supported by the National Science Centre, Poland (OPUS 2020/39/B/NZ2/01301).

REFERENCES

- (1) Wang, L.; Wang, N.; Zhang, W.; Cheng, X.; Yan, Z.; Shao, G.; Wang, X.; Wang, R.; Fu, C. Therapeutic peptides: current applications and future directions. *Sig Transduct Target Ther* **2022**, *7*, 1–27.
- (2) Lee, A.C.-L.; Harris, J. L.; Khanna, K. K.; Hong, J.-H. A Comprehensive Review on Current Advances in Peptide Drug Development and Design. *International Journal of Molecular Sciences* **2019**, *20*, 2383.
- (3) Ciemny, M.; Kurcinski, M.; Kamel, K.; Kolinski, A.; Alam, N.; Schueler-Furman, O.; Kmiecik, S. Protein-peptide docking: opportunities and challenges. *Drug Discov Today* **2018**, *23*, 1530–1537.
- (4) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Židek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589.
- (5) Tsaban, T.; Varga, J. K.; Avraham, O.; Ben-Aharon, Z.; Khrumushin, A.; Schueler-Furman, O. Harnessing protein folding neural networks for peptide–protein docking. *Nat. Commun.* **2022**, *13*, 176.
- (6) Evans, R.; O'Neill, M.; Pritzel, A.; Antropova, N.; Senior, A.; Green, T.; Židek, A.; Bates, R.; Blackwell, S.; Yim, J., et al. Protein complex prediction with AlphaFold-Multimer. *bioRxiv Preprint*, 2022. <https://www.biorxiv.org/content/10.1101/2021.10.04.463034v2>.
- (7) Johansson-Åkhe, I.; Wallner, B. Improving peptide-protein docking with AlphaFold-Multimer using forced sampling. *Front Bioinform* **2022**, *2*, 959160.
- (8) Pierce, B. G.; Wiehe, K.; Hwang, H.; Kim, B.-H.; Vreven, T.; Weng, Z. ZDOCK server: interactive docking prediction of protein–protein complexes and symmetric multimers. *Bioinformatics* **2014**, *30*, 1771–1773.
- (9) Kurcinski, M.; Badaczewska-Dawid, A.; Kolinski, M.; Kolinski, A.; Kmiecik, S. Flexible docking of peptides to proteins using CABS-dock. *Protein Sci.* **2020**, *29*, 211–222.
- (10) Alam, N.; Goldstein, O.; Xia, B.; Porter, K. A.; Kozakov, D.; Schueler-Furman, O. High-resolution global peptide-protein docking using fragments-based PIPER-FlexPepDock. *PLOS Computational Biology* **2017**, *13*, No. e1005905.
- (11) Johansson-Åkhe, I.; Mirabello, C.; Wallner, B. InterPep2: global peptide–protein docking using interaction surface templates. *Bioinformatics* **2020**, *36*, 2458–2465.
- (12) Shanker, S.; Sanner, M. F. Predicting Protein–Peptide Interactions: Benchmarking Deep Learning Techniques and a Comparison with Focused Docking. *J. Chem. Inf. Model.* **2023**, *63*, 3158–3170.
- (13) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **2023**, *379*, 1123–1130.
- (14) Manshour, N.; Ren, J. Z.; Esmaili, F.; Bergstrom, E.; Xu, D. Comprehensive Evaluation of AlphaFold-Multimer AlphaFold3 and ColabFold, and Scoring Functions in Predicting Protein–Peptide Complex Structures. *bioRxiv Preprint*, 2024. DOI: 10.1101/2024.11.11.622992.
- (15) Schaeffer, R. D.; Liao, Y.; Cheng, H.; Grishin, N. V. ECOD: new developments in the evolutionary classification of domains. *Nucleic Acids Res.* **2017**, *45*, D296–D302.
- (16) Lei, Y.; Li, S.; Liu, Z.; Wan, F.; Tian, T.; Li, S.; Zhao, D.; Zeng, J. A deep-learning framework for multi-level peptide–protein interaction prediction. *Nat. Commun.* **2021**, *12*, 5465.
- (17) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
- (18) Mirdita, M.; Schütze, K.; Moriwaki, Y.; Heo, L.; Ovchinnikov, S.; Steinegger, M. ColabFold: making protein folding accessible to all. *Nat. Methods* **2022**, *19*, 679–682.
- (19) Basu, S.; Wallner, B. DockQ: A Quality Measure for Protein–Protein Docking Models. *PLoS One* **2016**, *11*, No. e0161879.
- (20) Lensink, M. F.; Wodak, S. J. Docking, scoring, and affinity prediction in CAPRI. *Proteins: Structure, Function, and Bioinformatics* **2013**, *81*, 2082–2095.
- (21) Jeliakov, J. R., del Alamo, D., Karpiak, J. D. ESMFold Hallucinates Native-Like Protein Sequences. *bioRxiv Preprint*, 2023. <https://www.biorxiv.org/content/10.1101/2023.05.23.541774v1>.

4. Modelowanie elastyczności struktur białkowych: CABS-flex

4.1. [A5] Badanie funkcji białek na podstawie ich elastyczności strukturalnej przy użyciu modelowania CABS-flex

4.1.1. Wprowadzenie do [A5]

W pracy [A5] przedstawiliśmy artykuł przeglądowy poświęcony metodzie CABS-flex oraz jej zastosowaniom w modelowaniu elastyczności białek w kontekście ich funkcji biologicznych. Publikacja ukazała się w czasopiśmie *Protein Science* w ramach specjalnej serii *Tools for Protein Science*, co stanowiło wyraz uznania dla szerokiego wykorzystania metody CABS-flex w badaniach biologii strukturalnej. Artykuł miał na celu nie tylko syntetyczne omówienie podstaw metody, lecz także przegląd jej różnorodnych zastosowań. Mój wkład w realizację tego projektu obejmował współudział w przygotowaniu manuskryptu, ze szczególnym uwzględnieniem fragmentu dotyczącego integracji modelowania elastyczności z analizą skłonności do agregacji białek (Aggrescan4D), oraz udział w przygotowaniu wizualizacji.

Punkt wyjścia stanowiła obserwacja, że ruchliwość struktury białek - mimo trudności eksperymentalnej charakteryzacji - ma często kluczowe znaczenie funkcjonalne^{42,43,107–109}. Symulacje dynamiki molekularnej (MD) pozwalają na uchwycenie tych zjawisk, jednak wysoki koszt obliczeniowy ogranicza ich zastosowanie do stosunkowo małych układów i krótkich czasów symulacji^{110,111}, co bardziej szczegółowo omówiłem w rozdziale 2.3.1. W związku z tym niezbędne są metody mniej wymagające obliczeniowo, zdolne do odwzorowania realistycznych zmian konformacyjnych w większej skali. Takim narzędziem jest właśnie CABS-flex - metoda umożliwiająca szybkie i elastyczne modelowanie ruchów struktur białkowych.

CABS-flex opiera się na połączeniu odpowiednio zaprojektowanych symulacji wykorzystujących gruboziarnisty model CABS z rekonstrukcją uzyskanych struktur do reprezentacji pełnoatomowej. Metoda ta została udostępniona zarówno w formie serwera internetowego (CABS-flex⁷¹, a następnie CABS-flex 2.0⁵³), jak i jako samodzielna aplikacja *standalone*⁷². Obie wersje umożliwiają przeprowadzenie symulacji na bazie struktury wejściowej, analizę trajektorii oraz rekonstrukcję atomową wygenerowanych modeli. Metoda CABS-flex została przeze mnie omówiona w rozdziale 2.3.2, a model CABS w rozdziale 2.2.2.

Przegląd w [A5] porządkuje główne obszary zastosowań metody CABS-flex, wskazując ją jako efektywne narzędzie pośrednie pomiędzy kosztownymi obliczeniowo symulacjami pełnoatomowej dynamiki molekularnej a metodami dostarczającymi predykcji statycznych, w tym podejściami opartymi na sztucznej inteligencji. Pokazaliśmy, że profile fluktuacji generowane w CABS-flex dobrze odwzorowują dane eksperymentalne oparte na

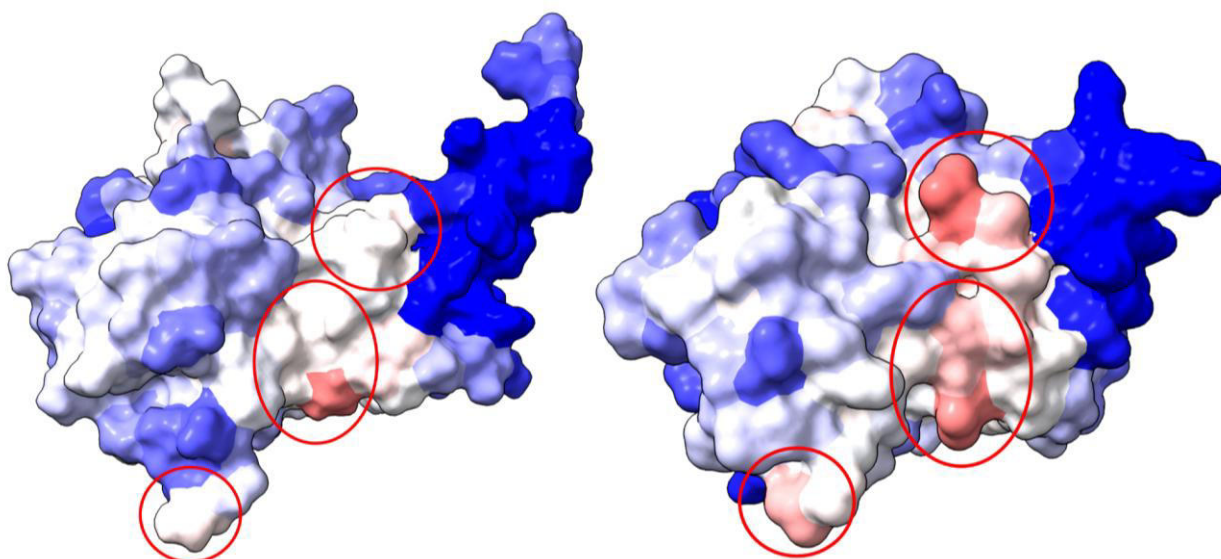
przesunięciach chemicznych NMR (RCI), szczególnie w identyfikacji regionów nadmiernie ruchliwych. Zwróciliśmy również uwagę na praktyczną wartość modeli AlphaFold jako punktu wyjścia do symulacji elastyczności - mimo ich statycznego charakteru, połączenie z CABS-flex umożliwia uzyskanie informacji o zmienności konformacyjnej w sytuacjach braku danych eksperymentalnych.

W kontekście układów wielocząsteczkowych omówiliśmy zastosowania metody do analizy dynamiki kompleksów białkowych, w tym białka kolca SARS-CoV-2. Trajektorie CABS-flex pozwalały tam ocenić wpływ mutacji w domenie RBD na ruchliwość regionów kontaktowych oraz potencjalne konsekwencje dla oddziaływań z receptorem ACE2. Przegląd uzupełniliśmy przykładami analizy komunikacji allosterycznej, m.in. w białkach Hsp90 i Hsp70, gdzie połączenie modelowania strukturalnego z analizą sieciową umożliwiało identyfikację regionów funkcjonalnych i szlaków przekazywania sygnału.

Istotnym elementem przeglądu są także zastosowania aplikacji CABS-flex *standalone*⁷² do modelowania peptydów⁵⁵. Rozszerzenie narzędzia o generowanie konformacji peptydów liniowych i cyklicznych umożliwia szybkie badanie krótkich, nieustrukturyzowanych fragmentów, które nie przyjmują jednej dominującej konformacji. W takich przypadkach metody oparte na głębokim uczeniu, nastawione na predykcję pojedynczej struktury, dostarczają ograniczonej informacji o ich rzeczywistej zmienności konformacyjnej.

Rozszerzenie zastosowań CABS-flex na modelowanie peptydów ujawniło również ograniczenia standardowych procedur rekonstrukcji pełnoatomowej. Wskazaliśmy, że podejścia oparte na Modellerze bywają wrażliwe na deformacje geometrii łańcucha głównego, co utrudnia dalsze analizy energetyczne. Jako alternatywę omówiliśmy metodę cg2all¹¹², opartą na głębokim uczeniu, która lepiej zachowuje geometrię struktur, redukuje kolizje i poprawia identyfikację elementów drugorzędowych. Uzupełnieniem tej analizy była dyskusja relacji między profilami elastyczności a wskaźnikiem pLDDT stosowanym w modelach AlphaFold, z podkreśleniem, że wysokie pLDDT nie wyklucza istnienia alternatywnych konformacji.

Ważny w kontekście niniejszej rozprawy pozostaje jednak obszar zastosowań dotyczący predykcji regionów podatnych na agregację (ang. *aggregation-prone regions*). CABS-flex został zintegrowany z metodą Aggrescan3D w trybie „dynamicznym” (ang. *dynamic mode*), w którym symulacje elastyczności pozwalają na bardziej realistyczną ocenę ryzyka agregacji. Pokazałem, że takie podejście może ujawnić regiony agregacyjne niewidoczne w analizach struktur statycznych, jak w przypadku domeny wiążącej DNA białka p53. Przykład tego podejścia pokazano na Rysunku 7, gdzie porównano wyniki oceny statycznej i dynamicznej. Samą metodę Aggrescan szerzej opisałem w rozdziale 2.4, a jej zastosowania omówiłem szczegółowo w rozdziale 5.



Rysunek 7. Integracja modelowania elastyczności z oceną skłonności do agregacji białek. Lewy panel przedstawia wynik klasycznej analizy Aggrescan3D opartej na pojedynczej strukturze białka p53, natomiast prawy panel pokazuje wynik analizy w trybie „dynamicznym”, w którym wykorzystano trajektorie CABS-flex, aby uwzględnić elastyczność konformacyjną białka. Ujęcie dynamiczne pozwala ujawnić dodatkowe regiony podatne na agregację (zaznaczone na czerwono), które są niewidoczne w analizie opartej wyłącznie na pojedynczym modelu. Źródło: [A5].

Uważam, że przedstawiony w pracy [A5] przegląd wyraźnie pokazuje, jak szerokie możliwości oferuje metoda CABS-flex w badaniach nad elastycznością i funkcją białek. Zaprezentowane przykłady potwierdzają jej wszechstronność - od integracji z danymi eksperymentalnymi i wynikami AlphaFold, przez analizę dynamiki kompleksów, mutacji i allosteryczności, po zastosowania specjalistyczne, takie jak ocena ryzyka agregacji czy predykcja modyfikacji potranslacyjnych. W mojej ocenie CABS-flex jest obecnie jednym z najbardziej uniwersalnych i efektywnych narzędzi do badania ruchliwości białek, stanowiąc cenne uzupełnienie klasycznych symulacji dynamiki molekularnej i metod opartych na sztucznej inteligencji. Pozwala w krótkim czasie uzyskać realistyczne modele ruchów konformacyjnych w skali niedostępnej dla standardowych symulacji MD. Wnioski płynące z tego przeglądu stały się również punktem wyjścia do opracowania kolejnej generacji narzędzia - CABS-flex 3.0 - której szczegóły omawiam w dalszej części niniejszej rozprawy.

4.1.2. Artykuł [A5]

Exploring protein functions from structural flexibility using CABS-flex modeling

Chandran Nithin¹  | Rocco Peter Fornari¹  | Smita P. Pilla¹  |
Karol Wroblewski¹  | Mateusz Zalewski¹  | Rafał Madaj²  |
Andrzej Kolinski¹  | Joanna M. Macnar¹  | Sebastian Kmiecik¹ 

¹Biological and Chemical Research Centre, Faculty of Chemistry, University of Warsaw, Warsaw, Poland

²Institute of Evolutionary Biology, Biological and Chemical Research Centre, Faculty of Biology, University of Warsaw, Warsaw, Poland

Correspondence

Sebastian Kmiecik, Biological and Chemical Research Center, Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland.
Email: sekmi@chem.uw.edu.pl

Present address

Joanna M. Macnar, Ryvu Therapeutics, Cracow, Poland.

Funding information

National Science Centre, Poland, Grant/Award Numbers: OPUS 2020/39/B/NZ2/01301, 2021/40/Q/NZ2/00078; PLGrid, Grant/Award Number: PLG/2024/016931

Review Editor: Nir Ben-Tal

Abstract

Understanding protein function often necessitates characterizing the flexibility of protein structures. However, simulating protein flexibility poses significant challenges due to the complex dynamics of protein systems, requiring extensive computational resources and accurate modeling techniques. In response to these challenges, the CABS-flex method has been developed as an efficient modeling tool that combines coarse-grained simulations with all-atom detail. Available both as a web server and a standalone package, CABS-flex is dedicated to a wide range of users. The web server version offers an accessible interface for straightforward tasks, while the standalone command-line program is designed for advanced users, providing additional features, analytical tools, and support for handling large systems. This paper examines the application of CABS-flex across various structure–function studies, facilitating investigations into the interplay among protein structure, dynamics, and function in diverse research fields. We present an overview of the current status of the CABS-flex methodology, highlighting its recent advancements, practical applications, and forthcoming challenges.

KEYWORDS

coarse-grained model, integrative modeling, molecular modeling, Monte Carlo simulation, multiscale simulation, protein aggregation, protein flexibility

1 | INTRODUCTION

The flexibility of a protein's structure can be pivotal to its biological function. Unfortunately, experimental investigation of protein dynamics is often challenging or unfeasible. Consequently, computer simulations play an important role here; however, the computational cost restricts the application of classical simulations like

all-atom molecular dynamics (MD) to small systems and short timescales (Kmiecik et al., 2016). In practice, for most biologically relevant proteins, classical simulations of structural flexibility demand significant computational resources, such as high-performance computer clusters.

In the age of AlphaFold, the field of structural biology has witnessed a profound transformation in our ability to predict protein structures with remarkable accuracy

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2024 The Author(s). *Protein Science* published by Wiley Periodicals LLC on behalf of The Protein Society.

(Varadi et al., 2022). However, the importance of dynamic and flexibility predictions cannot be overstated. Thanks to AlphaFold, a myriad of previously elusive systems (Varadi et al., 2022), including disordered regions (Ruff and Pappu, 2021), are now accessible for comprehensive dynamic simulations. The confidence scores provided by AlphaFold not only inform us about the accuracy of structural predictions but also offer invaluable insights into protein flexibility (Schwarz et al., 2021). Numerous pioneering studies have emerged, aiming to harness AlphaFold's predictive power for flexibility assessments (Guo et al., 2022; Ma et al., 2023). However, these endeavors are not without challenges, such as reconciling the static nature of structural predictions with the inherently dynamic nature of proteins (Saldaña et al., 2022) or better prediction of protein regions that are disordered under ambient conditions (Ma et al., 2023).

A decade ago we proposed CABS-flex, a method for fast simulations of protein flexibility (Jamroz et al., 2013a). This method has demonstrated a speed advantage of three to four orders of magnitude compared to all-atom MD (Jamroz et al., 2013a) and continues to be competitive. The CABS-flex method is built upon the well-established C-alpha, beta, and side chain (CABS) coarse-grained protein model, the applications of which have been comprehensively reviewed (Kmieciak et al., 2016). The dynamics observed in CABS Monte Carlo simulations align well with the dynamics observed in MD simulations of folded globular proteins over nanosecond timescales (Jamroz et al., 2013a, 2013b), fluctuations detected in NMR ensembles (Jamroz et al., 2014), and experimental data on folding and binding dynamics (Ciemny et al., 2016; Kmieciak and Kolinski, 2007, 2008; Kmieciak et al., 2012, 2016; Kurcinski et al., 2014, 2020). CABS-flex was made available as a publicly accessible web server (Jamroz et al., 2013b) and later updated to version 2.0 (Kuriata et al., 2018) available at <http://biocomp.chem.uw.edu.pl/CABSflex2>. Subsequently, the CABS-flex method was made available as a standalone application (Kurcinski et al., 2019)—a Python package engineered to provide command-line access to CABS computations, granting users full control over the simulation process. CABS-flex also serves as a component responsible for modeling protein flexibility in the Aggrescan3D method (Zambrano et al., 2015) for structure-based predictions of protein aggregation preferences. The CABS-flex methodology was also used in the protocol for flexible docking of protein-peptide complexes, available as CABS-dock (Kurcinski et al., 2020), which enables modeling of large-scale conformational changes during the docking simulation process (Ciemny et al., 2016; Puławski et al., 2023).

In this paper, we provide a concise overview of the CABS-flex method (Section 2) and showcase its practical

applications, supported by recent case studies (Section 3). Our demonstrations confirm the efficacy of the CABS-flex method in real-world applications, including analyzing the structural dynamics of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) spike proteins, modeling protein flexibility, exploring allosteric sites, and predicting regions prone to aggregation or nitrosylation. We also illustrate how CABS-flex can be used to combine flexibility analysis with predicted local distance difference test (pLDDT) predictive confidence scores. Finally, we discuss the important aspect of all-atom reconstruction from coarse-grained models and conclude with remarks and challenges for future research.

2 | CABS-FLEX METHOD

The CABS-flex methodology relies on a high-resolution coarse-grained model to represent protein chains, where each protein residue is characterized by up to four atoms. This representation includes the C α and C β atoms, as well as two virtual pseudo-atoms: one representing the center of mass of the side chain, and the other positioned at the center of the C α –C α virtual bond (Kmieciak et al., 2016; Kolinski, 2004). Notably, CABS-flex leverages knowledge-based statistical potentials, encompassing sequence-dependent short-range conformational preferences, context-dependent pairwise interactions of side chains, and a comprehensive model for hydrogen bonds within the protein's main chain. Importantly, our tool incorporates solvent effects implicitly into its calculations, enhancing the accuracy of its simulations. Through the implementation of a Monte Carlo-based scheme that controls a random series of small local movements, CABS-flex captures the intricate dynamics of proteins in long-term evolution. CABS-flex simulations demonstrate computational efficiency, with simulation times dependent on the size of the system. For example, 20 amino acid peptides require just over a minute, 500 amino acid proteins take less than an hour, and 1000 amino acid proteins about 2 h on an Intel Xeon 2.50 GHz processor. Coarse-grained structures produced during CABS-flex simulations can be rapidly translated onto atomistic models of reasonable fidelity (see discussion in Section 3.7). This unique combination of techniques makes CABS-flex an invaluable tool for protein research.

2.1 | CABS-flex web server

The CABS-flex web server prioritizes accessibility and user-friendliness for nonspecialists in programming (Jamroz et al., 2013b; Kuriata et al., 2018). Simply providing a file containing a full-atom model of a protein in PDB format is

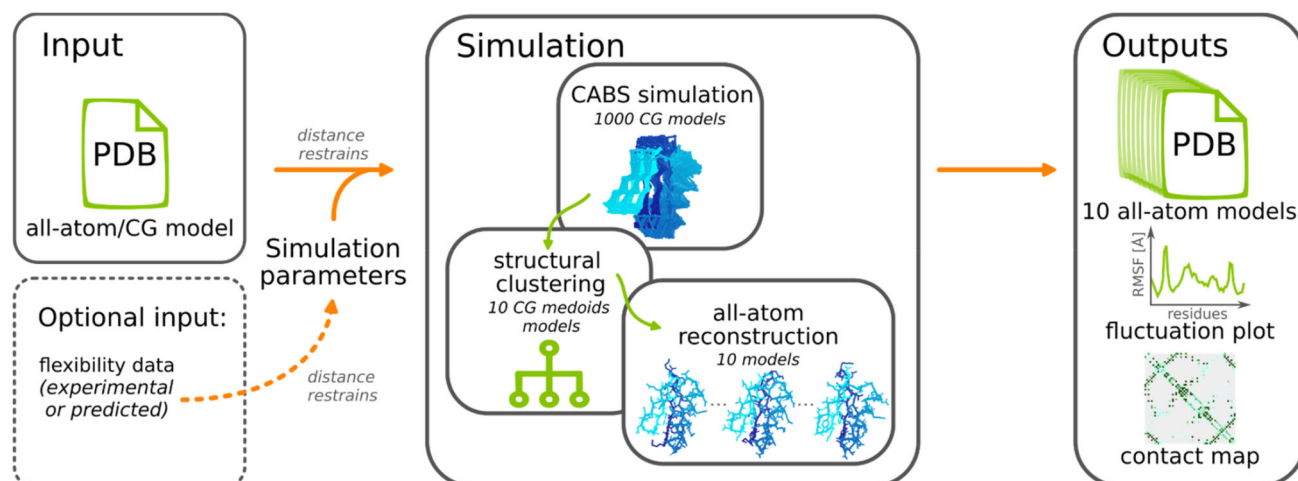
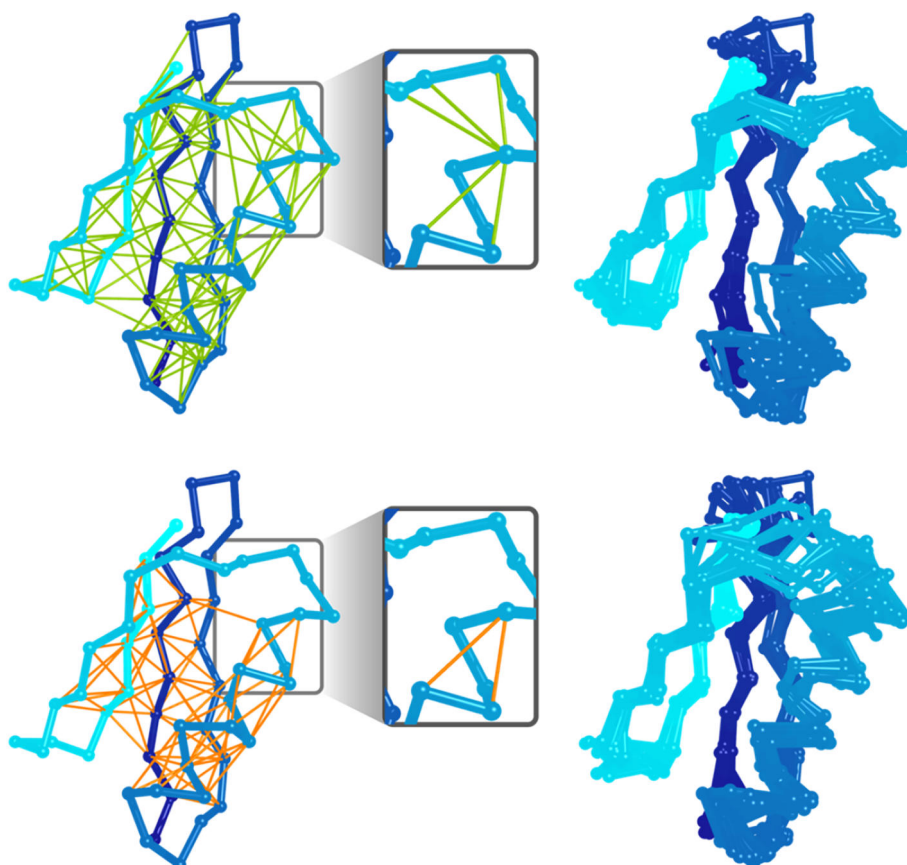


FIGURE 1 CABS-flex pipeline in default mode. The input protein structure serves as the starting point for the simulation and defines distance restraints based on either a default or a selected scheme. Optional inputs may include user-defined distance restraints. The simulation employs a coarse-grained CABS model. The protein dynamics trajectory is analyzed and clustered into 10 representative models, each capturing the essential variability of its cluster. These representative models are then reconstructed into all-atom models. The outputs of this process include PDB structures of the representative models, fluctuation profiles (illustrating RMSF vs. amino acid residue number), and contact maps derived from the trajectory. RMSF, root mean square fluctuation.

FIGURE 2 Distance restraints imposed on the protein (PDB ID: 2gb1) in the CABS-flex default modes SS1 (upper panel) and SS2 (lower panel). The left panels show the complete network of restraints shown as green and orange sticks, respectively for SS1 and SS2. In the middle, the restraints for a single residue are shown. The right panels show 10 protein models representing clusters of structures collected during one CABS-flex run with given default parameters. One can observe here a difference in the movement of the part depicted in the middle panel resulting from a different number of restraints.



enough to simulate the dynamics of the protein and to obtain an analysis of the fluctuations and their visualization. The server's capabilities have been significantly enhanced following a major upgrade to the 2.0 version available at

<http://biocomp.chem.uw.edu.pl/CABSflex2>. In response to user feedback, we have expanded the restrictions on protein structure input from 400 to 2000 amino acids and, from single-chain proteins to those comprising up to 10 chains.

This progress was facilitated by increased computing resources. Additionally, we enhanced our result analysis by offering protein contact maps for the simulation trajectory that display the frequency of residue–residue contacts during the simulation process and for 10 selected protein models (Figure 1).

Furthermore, a customizable panel has been implemented, offering a range of simulation parameters and options to provide better control over the simulation process. This includes user-defined alterations to distance restraints. Through this approach, it is possible to incorporate experimental data into the simulation, immobilize specific segments of the model, or eliminate selected restraints. Within the web server, users have the option to choose from various restraint modes. Each “Mode” directs the algorithm in generating restraints based on the assignment of secondary structures to the residues. Specifically, if set to “SS1,” the algorithm generates restraints when at least one of the residues is designated with a regular secondary structure, such as a helix or sheet. The “SS2” mode requires both residues to be assigned a regular secondary structure for restraints to be generated (Figure 2 presents a comparison of SS1 and SS2 restrain schemes for an exemplary small protein). Selecting “All” instructs the algorithm to produce restraints for every residue, whereas choosing “None” results in no restraint generation. By default, the system is set to “SS2.” Beyond the “Mode” setting, a collection of restraints is automatically produced based on additional parameters. These parameters determine the minimum distance along the protein chain required for two residues to be connected by a restraint, as well as the length and strength of these restraints (Kuriata et al., 2018).

2.2 | CABS-flex standalone application

Compared to the web server version, the CABS-flex standalone tool (Kurcinski et al., 2019) provides additional capabilities for modeling larger systems and customizing protocols. These enhancements enable full command-line access to various settings, such as the CABS coarse-grained model parameters, geometrical restraints utilized in CABS simulations, and result analysis. There are several options for analyzing the results, including customized output parameters, optional all-atom reconstruction (Badaczewska-Dawid et al., 2020a) with MODELLER (Webb and Sali, 2016) (including hydrogen atoms), personalized clustering and filtering, as well as contact map calculation and visualization, enabling users to analyze residue-to-residue contact frequencies with user-defined cutoffs.

As an additional feature, users can apply custom distance restraints to simulated models based on

experimental data. It is also possible to manually adjust the flexibility of specific regions within a protein. Users can adjust flexibility factors for targeted residues at the local level, or they can opt for global modifications to the restraint generation system to modify the protein's overall flexibility across its structure.

With the standalone version of CABS-flex being written in Python, it can be seamlessly integrated into bioinformatics workflows and connected to additional libraries. This allows users to incorporate their data and analyses into more complex systems. The software is compatible with Linux, macOS, and Windows, and requires Python 2.7 and GFortran as the only essential additional software. While Python 2.7 is currently a limitation due to its obsolescence, an in-house version utilizing Python 3.12 is under development, indicating future updates to the standalone version. Installing MODELLER is recommended for full functionality. Importantly, CABS-flex can be operated on a standard up-to-date home computer without any special hardware requirements, making it both practical and accessible. Moreover, CABS-flex is free and open-source, ensuring wide accessibility. For more information, including an overview of all CABS-flex options, examples of use, and installation instructions, visit <https://bitbucket.org/lcbio/cabsflex>.

2.3 | Selecting optimal inputs for CABS-flex protein simulations

The primary and only required input for CABS-flex is the protein structure. When predicting the flexibility pattern of a protein with multiple conformations available, the following approach is recommended: First, employ an ensemble approach by analyzing multiple conformations, which is crucial for proteins with substantial structural diversity to ensure a comprehensive representation of their flexibility. Second, prioritize structures obtained under physiological conditions comparable to those of the biological study, enhancing the simulation's relevance. Third, utilize databases such as CoDNAS (Monzon et al., 2016) and PDB (Burley et al., 2023) to select a variety of conformations that demonstrate different functional states, enriching the foundation for predictive modeling.

The variability in flexibility patterns that arise from using different starting structures underscores the dynamic nature of proteins. Each conformation can distinctively affect simulation outcomes, identifying unique flexibility hotspots influenced by both structural and environmental factors.

Regarding simulation settings, the default parameters were set to mirror fluctuation patterns similar to those seen in 10-ns MD simulations using various force fields,

as referenced in earlier studies (Jamroz et al., 2013a). These settings are particularly calibrated for proteins up to approximately 200 amino acids, where the fluctuation profiles tend to be qualitatively consistent, with minimal differences due to the stochastic nature of the simulation algorithm. For proteins with high conformational diversity or larger sizes, it is advisable to increase the number of simulation cycles, potentially by tenfold, to capture more accurate dynamics (the advanced simulation options are discussed in the CABS-flex 2.0 documentation) (Kuriata et al., 2018).

3 | APPLICATIONS—CASE STUDIES

This section presents case studies showing the capabilities of CABS-flex in modeling protein systems. These examples are intended to inspire readers to explore CABS-flex applications in structure–function–flexibility relationship studies.

3.1 | Modeling the flexibility of globular protein structures

CABS-flex was initially employed to simulate the flexibility of folded globular proteins, demonstrating results in line with MD simulations across nanosecond timescales (Jamroz et al., 2013a, 2013b) and consistent with fluctuations observed in NMR ensembles (Jamroz et al., 2014). This section presents two case studies that compare the flexibility profiles predicted by CABS-flex simulations to those measured by the random coil index (RCI), which is derived from backbone chemical shifts (Fowler et al., 2020).

The RCI serves as a reliable measure of local protein rigidity (Berjanskii and Wishart, 2005, 2008), and is frequently referenced for assessing the structural accuracy of both NMR-derived and AlphaFold 2 (AF2)-predicted models (Fowler et al., 2021; Fowler and Williamson, 2022). Notably, it has been observed (Fowler and Williamson, 2022) that AF2-generated structures generally surpass NMR-derived structures in accuracy. However, the application of RCI is contingent upon the availability of backbone chemical shift assignments. In the absence of such data, AF2-predicted structures offer a provisional model for the protein's tertiary structure, whose flexibility can then be swiftly approximated using CABS-flex.

For our analysis, we have chosen two examples from Fowler and Williamson (2022) to explore the correlation between flexibility profiles from CABS-flex simulations and those derived from RCI, which stands as a credible

experimental benchmark. The first case, PDB ID 2ct5, illustrates an instance where the AF2-predicted structure is more accurate than the highest-scoring NMR structure. Conversely, PDB ID 2y4q represents a case where the NMR structure is more precise. The CABS-flex simulations were conducted using standard parameters, and the root mean square fluctuation (RMSF) was normalized to a unit scale to facilitate a direct comparison with the RCI-measured flexibility. These findings are illustrated in Figure 3. For PDB ID 2ct5, the flexibility profiles generated by CABS-flex for both the NMR and AF2 models accurately pinpoint the locations of flexible regions. Specifically, the segment between residues 18 and 28 and the terminal regions are identified as overly flexible in the NMR model predictions, with a notably better quantitative alignment observed with the AF2 structure's predictions. In the case of PDB ID 2y4q, CABS-flex accurately identifies the flexible regions as well. The AF2 model lacks a short beta-sheet found in the NMR model (Fowler and Williamson, 2022), which contributes to the AF2 model's diminished accuracy for this protein. Intriguingly, CABS-flex predicts the region encompassing residues 760–770 to exhibit excessive flexibility in both models. A further distinct difference is observed in the terminal region, where the AF2 model features a longer alpha-helix compared to the NMR model.

In summary, our evaluation indicates that CABS-flex serves as an effective tool for estimating backbone flexibility at a modest computational expense, offering accuracy on par with the RCI method for proteins lacking backbone chemical shift data. For such instances, employing AF2-derived structures—or x-ray structures when accessible—as the basis for analysis is expected to provide the most accurate estimates of flexibility. This approach underscores the utility of CABS-flex in complementing other flexibility measurement methods, particularly in scenarios where direct experimental data may be limited or unavailable.

3.2 | Structure–function analysis of protein complexes: Interactions and mutational dynamics

In this section, we have compiled several examples demonstrating applications of the CABS-flex method to explore the critical interactions and structural dynamics of the SARS-CoV-2 spike (S) protein. The S glycoprotein is a multidomain protein that binds to the angiotensin-converting enzyme 2 (ACE2) (Huang et al., 2020). Comprising 1273 residues, the S protein consists of an amino (N)-terminal S1 subunit and a carboxyl (C)-terminal S2 subunit. The S1 domain, which is responsible for interactions with the host receptor, includes

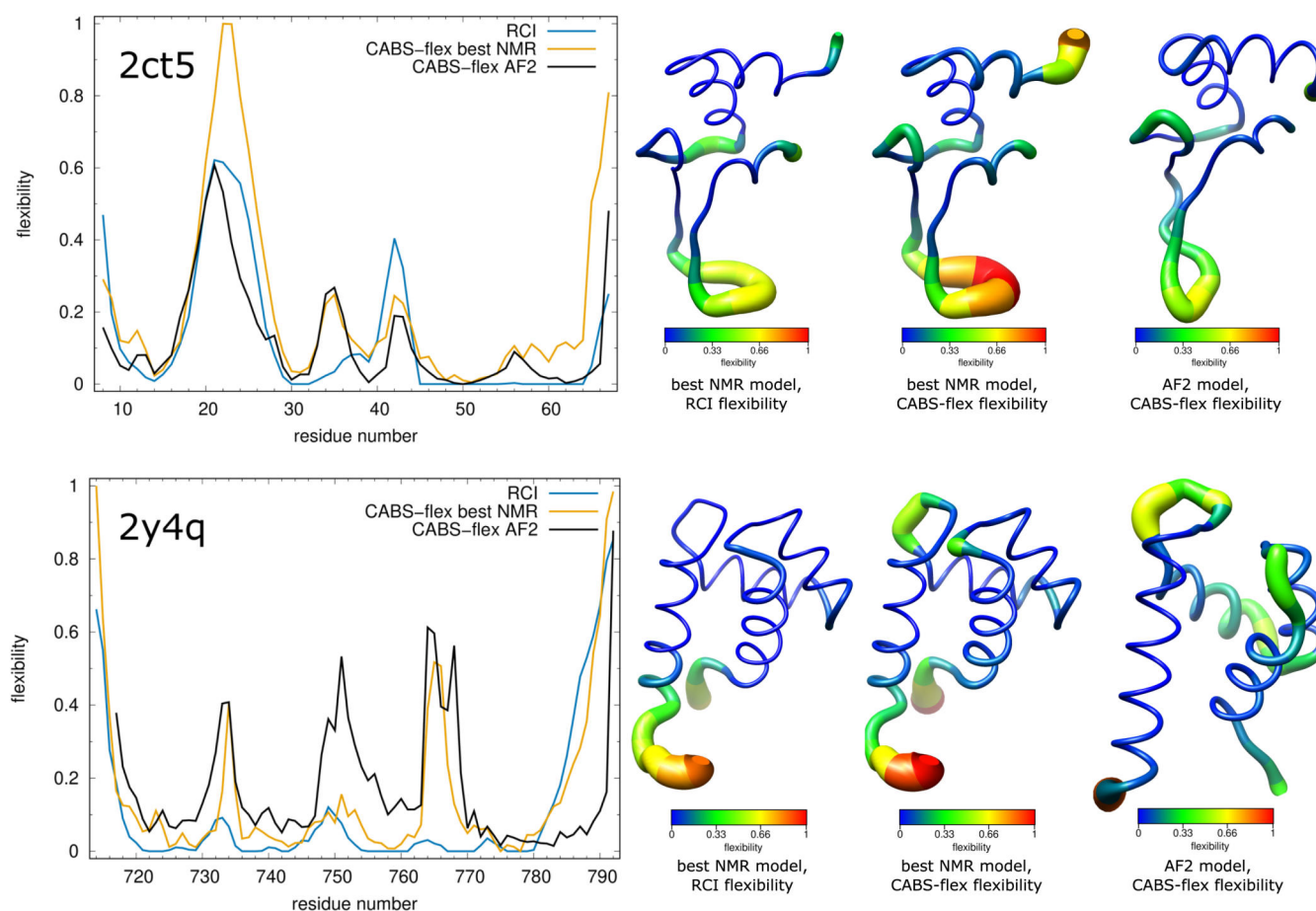


FIGURE 3 Comparison of flexibility profiles computed from the RCI and obtained from CABS-flex simulations. Top: PDB ID **2ct5** NMR model 3 and AF2 model (UniProt: O96006). Bottom: PDB ID **2y4q** NMR model 3 and AF2 model (UniProt: Q13563). In the right panels, the color and the thickness of the tube representation are proportional to the flexibility of each residue. RCI, random coil index.

an N-terminal domain (NTD), a receptor binding domain (RBD), and a C-terminal domain (CTD) with two structurally conserved subdomains: CTD1 and CTD2. The S protein forms a homotrimer and the virus infection mechanism involves conformational transitions. Specifically, the RBDs alternate between a receptor-accessible “up” state (Figure 4a) and a receptor-inaccessible “down” state (Figure 4b) (Wrapp et al., 2020). Figure 4c highlights the residues involved in the binding of the RBD with ACE2 (using contact map MAPIYA web server [Badaczewska-Dawid et al., 2022a] with a distance cutoff of 8 Å). The mutations in the RBD domain significantly influence the virus's evolution.

CABS-flex can be particularly valuable for investigating the dynamic aspects of protein interactions that are not evident from static structural data. Although RMSF derived from multiple crystallized structures offers important insights into observed protein flexibility, CABS-flex provides additional value by modeling dynamics and interactions that may not be fully resolved in crystal structures. For example, in the study of the SARS-CoV-2 spike protein (Sanyal et al., 2022), CABS-

flex highlighted significant flexibility in the distal ends of the ACE2 binding motif—a region often unresolved in crystallographic data. This modeling was crucial for understanding the enhanced binding affinity of the Omicron variant's spike protein to the ACE2 receptor (Ortega et al., 2022), which is speculated to contribute to its higher transmissibility. The insights gained from CABS-flex about the dynamic behavior of these critical regions complement static RMSF analyses, offering a more complete understanding of how mutations may influence virus–host interactions.

Beyond exploring the dynamics of entire regions of protein–protein interactions, CABS-flex has also been extensively used to study the effects of mutations on protein dynamics. It is important to clarify that CABS-flex does not inherently modify amino acid sequences or generate coordinates for newly substituted amino acids. Instead, users must prepare mutated structures prior to simulation, often employing tools such as Modeller (Webb and Sali, 2016) or FoldX (Büß et al., 2018). This crucial preparatory step ensures that the input structures

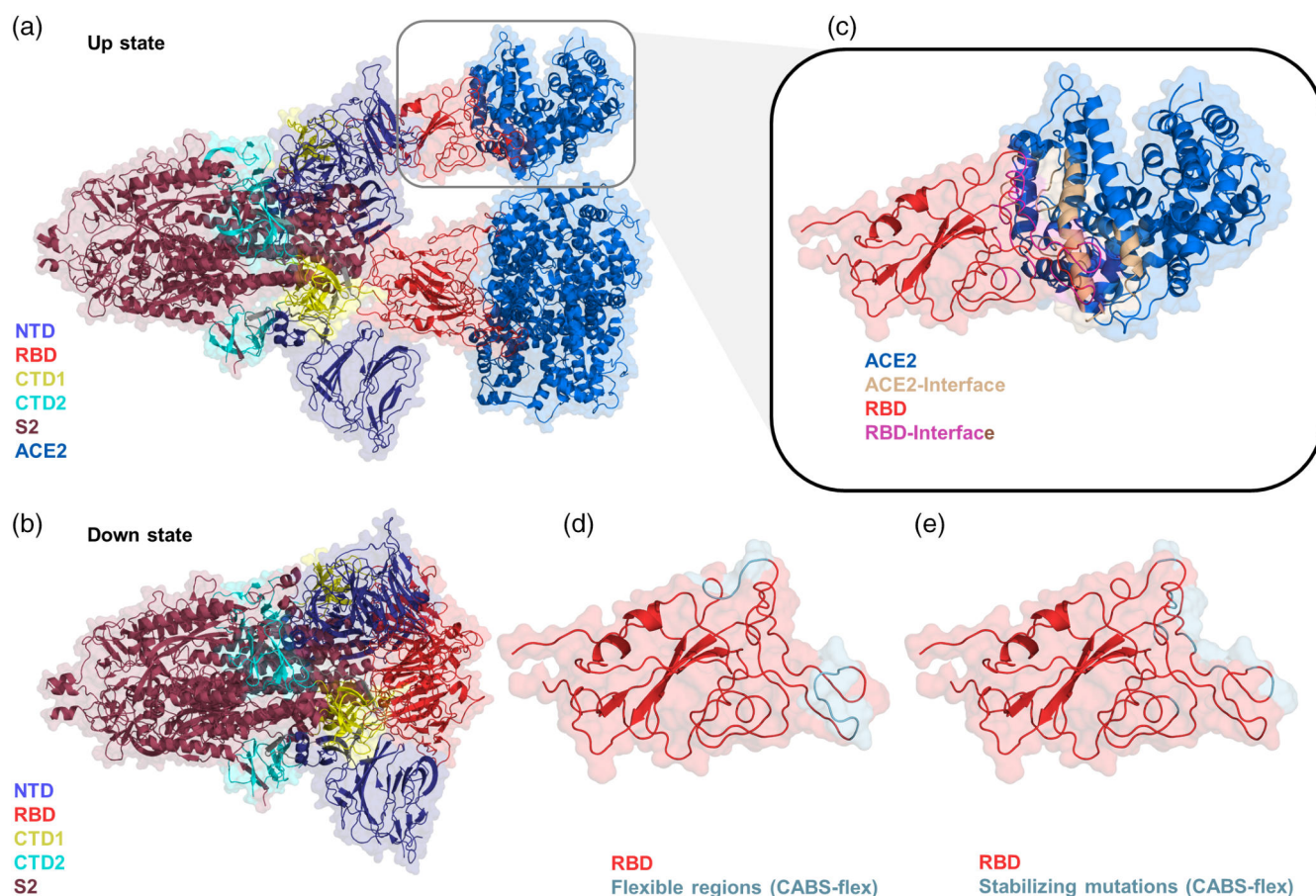


FIGURE 4 The SARS-CoV-2 spike (S) protein in (a) “up” conformation (PDB ID: 7a98) bound to angiotensin-converting enzyme 2 (ACE2) and in (b) “down” conformation (PDB ID: 7a98) which is without binding to the receptor. The N-terminal S1 subunit consists of NTD (residues 14–306) in blue, RBD (residues 331–528) in tv_red, CTD1 (residues 529–591) in yellow, and CTD2 (residues 592–686) in cyan. The remaining residues form the S2 subunit and are shown in raspberry. (c) The interaction between S and ACE2 happens through the residues shown in light magenta on the RBD and wheat on the ACE2 receptor. The residues identified from CABS-flex trajectories as (d) highly flexible and mutations could have significant effects on the evolution of the virus (Sanyal et al., 2022) and (e) have a significant stabilization effect on S-ACE2 complex (Mishra et al., 2022) are highlighted in sky blue. SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

accurately reflect the physical characteristics of the mutated proteins, allowing for reliable simulations of their dynamics and interactions. This detail underscores the practical utility of CABS-flex in mutation analysis, demonstrating its capability to predict the functional consequences of genetic variations.

In combination with other analyses, CABS-flex has been employed to study the effects of mutations in the RBD of the spike protein. Sanyal et al. (2022) used CABS-flex trajectories to explore and identify the regions in the RBD that can have significant impacts on the stabilization of the RBD domain in its interactions with ACE2 reception. This study identified the residue stretches 441–445 and 477–484 (Figure 4d) as highly flexible and mutations in this region will have a significant effect on the evolution of SARS-CoV-2. Notably, the residues 477, 478, and 484 were mutated in the Omicron variant (Karim and Karim, 2021). In another study, Mishra et al. (2022) used CABS-flex to

study the impact of mutations in S proteins on the stabilization or destabilization of the S-ACE2 complex. This study pinpointed 36 mutations involving 12 residues that could enhance the stability of the S-ACE2 complex (Figure 4e). The V503D mutation exhibited the highest stabilization of the S-ACE2 complex and is known to resist neutralization (Tortorici et al., 2021). Moreover, mutations K417Y, E484A, N501Y, D614G, and P681H identified in this study align with those observed in the Omicron variant and its sub-lineages (Cao et al., 2022; He et al., 2021; Zhang et al., 2022). These two studies used CABS-flex trajectories to identify the residues that could play a significant role in virus evolutions. Multiple residues and substitutions identified by these studies were also observed in the SARS-CoV-2 variants.

The CABS-flex was also used to investigate the impact of mutations on the ability of the S protein to bind to the ACE2 receptor. For example, the D614G substitution

introduces higher infectivity (Korber et al., 2020) and to understand the possible reasons for this, Verkhivker et al. (2022) used trajectories from 1000 independent CABS-flex simulations. The authors performed a principal component analysis (PCA) of the CABS-flex trajectories and used elastic network model analysis to determine the dynamics profiles from these trajectories. The D614G mutation introduces significant conformational adaptability to the “up” states, without compromising the folding stability or integrity of the S proteins (Cai et al., 2021). Based on the dynamic profiles from CABS-flex trajectories, the study concluded that the D614G mutation allows the S protein to exist in a dynamic equilibrium between closed and open states, with the RBD favoring the “up” conformation before ACE2 binding. In another study (Verkhivker, 2022a) exploring the effects of amino acid substitutions in SARS-CoV-2 variants, trajectories from 100 independent CABS-flex simulations were used to identify the diversity of “up” conformations. The study concluded that both A570D and D614G substitutions act as potential regulatory switches and play a role in modulating the dynamics of the S protein and its interaction with host receptors. Moreover, the study postulates that the two residues play an important role in the transitions between the “down” and “up” conformations of the RBD (Verkhivker, 2022a). Together, these two studies exemplify the use of CABS-flex trajectories to understand the implications of amino acid substitution introducing changes in the adaptability and evolution of the virus.

3.3 | Exploring allosteric mechanisms through integrative computational modeling

CABS-flex simulations have been utilized as a tool for investigating dynamics and elucidating allosteric regulation mechanisms. Although CABS-flex does not directly model ligand interactions, it effectively pinpoints regions of significant conformational flexibility within proteins that suggest potential allosteric sites. For instance, Verkhivker (2022b) applied CABS-flex in conjunction with atomistic simulations to explore the dynamic conformational landscapes of the Hsp90 chaperone complexes. This integrated approach, which demonstrated the compatibility of CABS-flex with MD simulations, helped identify critical dynamic areas essential for understanding allosteric regulation and client protein integration in Hsp90.

Similarly, Amusengeri et al. (2019) employed CABS-flex to analyze the intrinsic dynamics of heat shock proteins (Hsp70s) from *Plasmodium falciparum*, highlighting regions of conformational flexibility indicative

of potential allosteric sites. Their study effectively bridges the resolution gap between coarse-grained simulations and all-atom MD simulations, confirming the consistency of allosteric sites identified by CABS-flex with those observed in more detailed simulations. Furthermore, Amusengeri et al. (2019) verified the interactions of ligands with the studied proteins through additional all-atom MD simulations. These simulations, focusing on both apo (ligand-free) and holo (ligand-bound) states, provided a detailed view of how ligands influence protein conformation and function.

Another significant contribution comes from a study focusing on the Hsp90 chaperone interactions with cochaperones and client proteins (Verkhivker, 2022c), which provided detailed atomistic insights from integrative modeling (using CABS-flex simulations followed by MODELLER-based all-atom reconstruction) and network analysis of conformational landscapes. This study reveals that dynamic-based allostery, driven by the interplay of conformational states and interactions within the chaperone complex, is crucial for the efficient recruitment and integration of client proteins.

CABS-flex was also adeptly employed to model the S protein's ability to switch between functional states, highlighting its utility in mapping dynamic regulatory switches within the protein structure (Verkhivker and Di Paola, 2021a, 2021b). Simulation trajectories from CABS-flex were validated by comparing them with those from all-atom MD simulations and experimental data, ensuring reliability in modeling the spike protein's conformational dynamics. These were further analyzed alongside atomistic reconstructions and PCA to identify hinge regions and characterize collective motions within the SARS-CoV-2 spike protein complexes, providing insights into allosteric communication pathways.

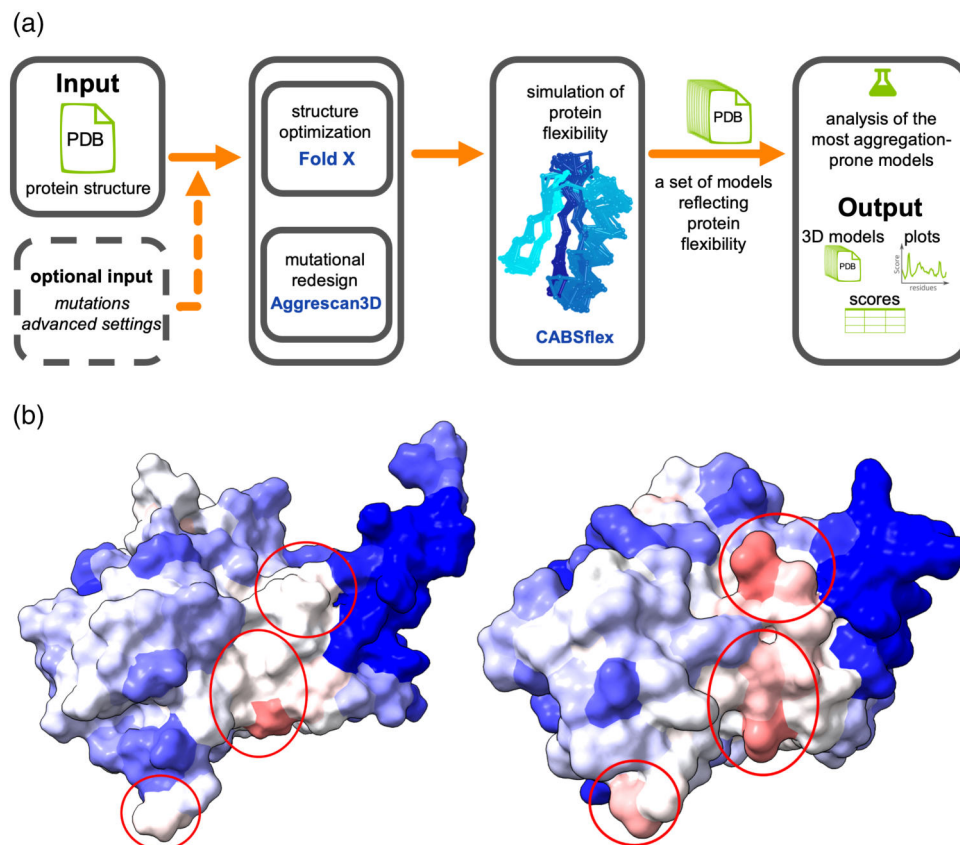
Ultimately, this collection of studies highlights the utility of CABS-flex in identifying potential allosteric sites and demonstrates that integrative simulation strategies can provide crucial insights into the complex mechanisms of allostery in protein dynamics, essential for advancing targeted drug development.

3.4 | Structure-based prediction of aggregation-prone regions (APRs)

CABS-flex, as a tool for simulating near-native flexibility, can be easily used to introduce the aspect of protein flexibility to structure-based studies of static conformations of globular proteins. This capability was used in the Aggrescan3D (A3D) (Kuriata et al., 2019a) method for structure-based prediction of protein aggregation properties. By default, A3D runs in “static mode” where it evaluates aggregation tendencies of a static protein structure. However, A3D

FIGURE 5 Aggrescan3D

“dynamic mode.” (a) A3D pipeline: the input structure’s flexibility is simulated using CABS-flex to generate a set of predicted models that reflect the varying conformations of the structure; these models are then analyzed and scored based on their aggregation propensity, and the model with the highest aggregation propensity is presented as the final prediction. (b) Human p53 DNA-binding domain (PDB ID: [2fej](#)) aggregation propensity predictions using A3D “static mode” (left side) and “dynamic mode” (right side). Residues are colored in A3D scale (blue—solubility; red—aggregation). Aggregation prone regions are encircled. A3D, Aggrescan3D.



calculations can be also performed using “dynamic mode” (Figure 5a) which harnesses CABS-flex to simulate protein fluctuations. In this mode, an initial static structure serves as a starting point for a CABS-flex simulation, and aggregation assessments are performed on CABS-flex output models, which reflect protein flexibility. The final structure is determined as the model with the highest aggregation score, offering a more comprehensive and dynamic view of protein aggregation tendencies. This dynamic approach is particularly valuable as in solution protein aggregation may arise not only from global unfolding but also as a result of local structural changes and fluctuations from the protein’s native state (Chiti and Dobson, 2009). To emphasize the utility of this approach we can consider the illustrative example of the human p53 DNA-binding domain (PDB ID: [2fej](#)). Calculations performed using A3D “static mode” (Figure 5b, left panel) do not reveal all potential APRs. However, after harnessing A3D “dynamic mode” (Figure 5b, right panel), CABS-flex unveils APRs that mirror experimental observations (Navarro and Ventura, 2019). In addition, A3D provides robust tools for introducing and analyzing mutations within protein sequences. Users can utilize the “Mutate residues” feature to manually input specific sequence changes or employ the automatic mutations functionality. This automatic tool identifies APRs and suggests strategic point mutations to charged residues, aiming to enhance protein solubility. Such modifications are crucial

for exploring how changes in the flexibility of flanking regions can influence protein stability and aggregation propensity.

A3D method is available both as a web server (Kuriata et al., 2019a) and as a standalone tool (Kuriata et al., 2019b), offering a wide range of applications including automatic redesign toward soluble variants. Recently, A3D has been applied to predict regions responsible for aggregation across entire proteomes, a feat made possible through the use of structural predictions generated by AlphaFold2 (AF2) (Varadi et al., 2022). A3D databases have been created to seamlessly integrate A3D’s analytical infrastructure with structural predictions for human (Badaczewska-Dawid et al., 2022b), yeast (Garcia-Pardo et al., 2023), and other proteomes (Badaczewska-Dawid et al., 2024a). Finally, the A3D method has recently undergone significant enhancements, notably including the ability to consider environmental pH. These advancements are now accessible through the A4D web service (Barcenas et al. 2024).

3.5 | Structure-based prediction of S-nitrosylation sites

In Section 3.4, we discussed the utilization of CABS-flex for revealing APRs on protein surfaces. A similar approach, integrating CABS-flex into structure-based

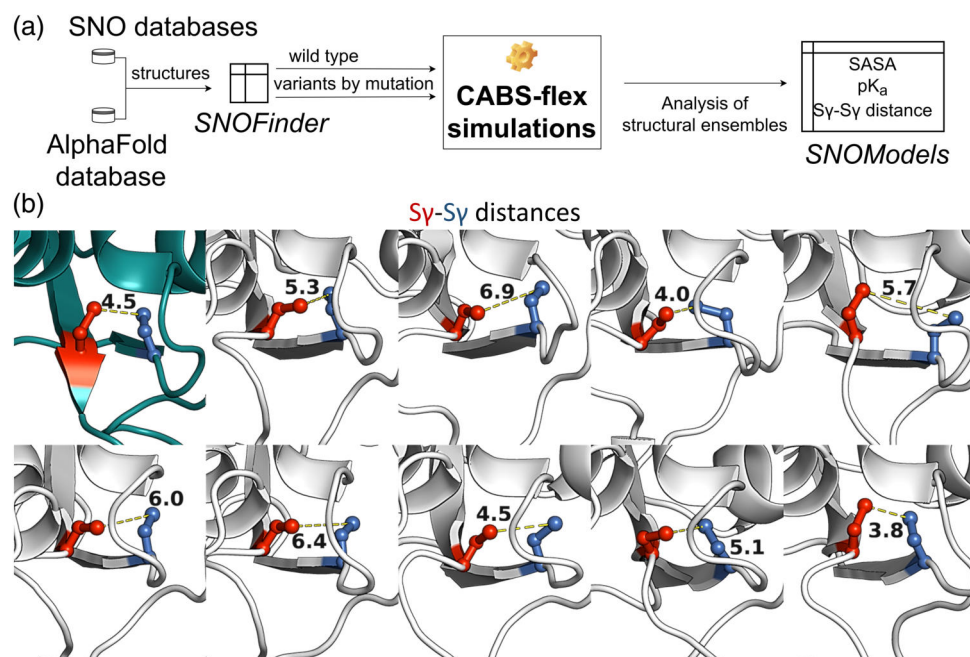


FIGURE 6 CABS-flex ensembles provided valuable insights into the structural dynamics of S-nitrosylated proteins. (a) The workflow utilizes SNOfinder to identify proteins susceptible to S-nitrosylation, generates a structural ensemble of 20 models per protein with CABS-flex to highlight their flexibility, and analyzes these models with SNOmodel to calculate key structural and biochemical parameters, thereby enhancing understanding of S-nitrosylation mechanisms. (b) The cartoon representations show the models from the CABS-flex ensemble (white) of the mitochondrial chaperone TRAP1 (deep teal). The stick and ball representations highlight the conformational dynamics of SNO cysteine (SNO, red), proximal cysteine (proximal, marine) of the ensembles, and the Sy-Sy distances between them.

analyses, was also employed by Papaleo et al. (2023) for identifying S-nitrosylation sites. S-nitrosylation—a reversible modification of cysteine residues to form S-nitrosothiols—is a key posttranslational modification (PTM) affecting cellular functions and is implicated in various diseases, making it a target for therapeutic interventions. The method involves identifying cysteines close in three-dimensional space, not just sequence, highlighting the importance of protein flexibility for exposing potential S-nitrosylation sites and influencing protein behavior and interactions within cellular pathways. For instance, in the mitochondrial chaperone TRAP1, SNO at Cys501 induces conformational changes at the distal site in the protein structure which results in altering the open and closing motions necessary for its chaperone function. This modification reduces its ATPase activity, and with its proximity to Cys527 suggests potential disulfide bridge formation (Faenza et al., 2020; Rizza et al., 2016).

Figure 6a presents a workflow devised by Papaleo et al. (2023), showcasing a two-step bioinformatic approach for studying S-nitrosylation in proteins. Initially, the workflow employs SNOfinder, a tool specifically designed for the identification of proteins susceptible to S-nitrosylation by leveraging databases of known PTMs and structural data. This identification process is critical for pinpointing proteins where S-nitrosylation plays a potential role in regulating

protein function and signaling pathways. Following the identification, CABS-flex is utilized to generate a structural ensemble comprising 20 models for each identified protein. These models collectively represent the protein's structural flexibility and potential conformational states, taking into account the dynamic nature of protein structures in biological systems. The resulting ensemble is then analyzed using the SNOmodel pipeline, which is tailored to calculate specific structural and biochemical parameters, such as dihedral angles and solvent accessibility of cysteine residues involved in S-nitrosylation. This analysis is aimed at understanding the structural context of S-nitrosylation sites, including their proximity to other functional groups within the protein and potential interactions with cofactors or metal ions, thereby shedding light on the mechanistic underpinnings of S-nitrosylation in protein function and regulation.

For the mitochondrial chaperone TRAP1, the CABS-flex simulations focused on the cysteines C501 and C527. The Sy-Sy distance between the two sulfur groups of these cysteines is highlighted in Figure 6b. It showed that SNO at C501 could lead to a disulfide bridge with C527, a finding supported by experimental validation. This insight was important for understanding the redox regulation of TRAP1 and its role in diseases. The CABS-flex approach, a computationally inexpensive method, effectively

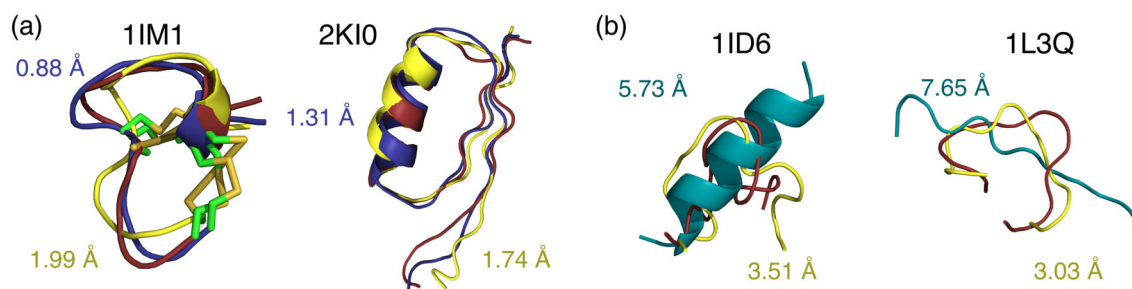


FIGURE 7 (a) Example CABS-flex predictions (best prediction out of 10,000 simulated structures—blue; best prediction out of top 10 structures ranked by CABS-flex—yellow) superimposed on the experimental PDB structures (red). (b) Comparison between CABS-flex predictions (yellow) and AlphaFold2 predictions (cyan) superimposed on the experimental PDB structures (red).

identified candidates with structural constraints that limit SNO or disulfide bridge formation. This study analyzed multiple SNO proteins, including the tumor protein TP53, focusing on its DNA-binding domain. For example, CABS-flex simulations were used to determine how mutations near the SNO and adjacent cysteine sites in TP53 influence its stability and DNA-binding capabilities. Key findings include the stabilizing N235H variant promoting S-nitrosylation and disulfide bridge formation, enhancing TP53's structure. On the other hand, the destabilizing variants could impair the structural mechanism.

3.6 | Structure prediction of linear and cyclic peptides

Recently, we introduced a new CABS-flex-based protocol that enables de novo structure prediction of linear and cyclic peptides (Badaczewska-Dawid et al., 2024b). The protocol has been introduced as an extension of the CABS-flex standalone application; it uses the CABS-flex methodology to perform an efficient exploration of a peptide's conformational space (Badaczewska-Dawid et al., 2024b). The protocol was extensively tested against other methods (AF2 [Jumper et al., 2021] implemented in the ColabFold [Mirdita et al., 2022], ESMFold [Lin et al., 2023], APPTTEST [Timmons and Hewage, 2021], PEP-FOLD [Shen et al., 2014]) on a benchmark consisting of 159 linear and cyclic peptides ranging from short peptides of 9 residues to mini-proteins consisting of 49 residues. The two examples of CABS-flex predictions are highlighted in Figure 7. In most benchmarked cases, AlphaFold-based protocol outperformed CABS-flex, especially for longer peptides, due to its ability to capture complex long-range interactions. However, the diverse structural behaviors of short peptides, which range from disordered regions to well-defined secondary structures, pose a challenge for deep learning methods like AF2, as experimental structures sometimes inadequately

represent this diversity. We have demonstrated that in these instances, coarse-grained methods such as CABS-flex remain competitive by effectively navigating this structural diversity through simplifications and reduced granularity. Figure 7b illustrates two examples in which CABS-flex was able to predict conformations that turned out to be closer to the experimental ones than those proposed by AF2.

In a standard CABS-flex simulation, the only required input is the initial structure provided in the form of a PDB code or PDB file. In the CABS-flex peptide structure prediction protocol, the only explicitly required input is an amino acid sequence. The starting structure is generated as a random coil based on that input sequence. There are other optional inputs, such as secondary structure in three-letter notation (H, helix; E, extended; C, coil). If it is not supplied, the secondary structure would be predicted by Psipred (Jones, 1999), though we recommend users to predict it using state-of-the-art deep-learning methods for enhanced accuracy (i.e., NetSurfP-3.0 [Høie et al., 2022] or SPOT-1D-LM [Singh et al., 2022]). The versatility of the presented CABS-flex modeling pipeline is a notable advantage, as it can be readily expanded by incorporating additional data or testing alternative configurations. Additional simulation data can be easily introduced as distance constraints between C-alpha atoms or side group centers of mass with appropriate confidence levels that may force a precisely known distance or provide a weak bias (Badaczewska-Dawid et al., 2024b). Potential pipeline extensions include the use of techniques like MD to investigate the peptide dynamics or explore the design of peptide-based drugs with modifications extending beyond natural amino acids, providing a broader scope for research and applications.

3.7 | Merging with high-accuracy all-atom reconstruction

CABS-flex employs the CABS coarse-grained model to create C-alpha protein models (refer to Section 2), which

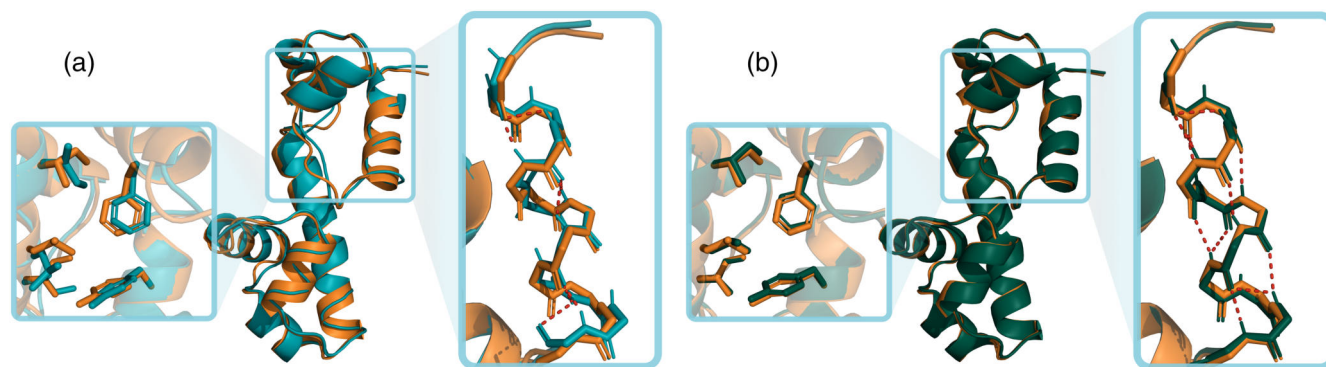


FIGURE 8 CABS-flex output structures rebuilt by two methods: (a) Modeller in cyan and (b) cg2all⁷² in dark green, superimposed on the experimental PDB structure in orange (PDB ID: 3oou). We highlight the accuracy of side-chain reconstruction (left panels) and helical hydrogen bond networks (right panels) in the rebuilt structures.

are subsequently reconstructed into full-atom structures. Despite the advances in reconstruction methods, they are not without flaws, and as a result, suboptimal chain geometries can emerge in the final output (Badaczewska-Dawid et al., 2020b). Moreover, the process of reconstructing from coarse-grained models introduces additional challenges. Namely, the initial C-alpha trace may include unphysical distortions and different backbone reconstruction methods exhibit varying degrees of resilience to that challenge (Badaczewska-Dawid et al., 2020b). This can lead to missing portions of the rebuilt backbone or amplified distortions, which can substantially affect the quality of subsequent side chain reconstruction, all-atom energy minimization, and scoring (Badaczewska-Dawid et al., 2020b).

As of now, CABS-based tools are integrated with Modeller-based protocol (Webb and Sali, 2016), utilizing the coarse-grained models as templates and the automodel class from Modeller (Webb and Sali, 2016), to generate a comprehensive set of restraints for reconstructing all heavy atoms (Badaczewska-Dawid et al., 2020a). Unfortunately, the Modeller protocol often faces the reconstruction difficulties mentioned above. This prompted our exploration of a new reconstruction method. Recently, we obtained very promising results using the deep-learning-based method cg2all (Heo and Feig, 2024). The method, inspired by the success of AF2, employs a graph neural network and utilizes rigid-body blocks for 3D structure generation. It incorporates physical constraints and structural features, learned from known protein conformations during model training. Preliminary results indicate an improvement in C-alpha RMSDs, fewer clashes, and better MolProbity (Chen et al., 2010) scores of the rebuilt models by the cg2all method. This can be seen in Figure 8, where the quality of both the main chain and side chains rebuilt by Modeller (Figure 8a) is worse than in the structure rebuilt by

cg2all (Figure 8b). The reconstructed structures demonstrate improved local quality and stability, assessed by using the reconstructed structures as a starting point for MD simulations. Furthermore, we observe a significant improvement (Figure 8b, right panel) in the previously observed issue of DSSP (Kabsch and Sander, 1983) being unable to identify secondary structure elements in the Modeller-reconstructed structures, due to local distortions (Figure 8a, right panel).

Therefore, one of the shortcomings of using coarse-grained models, including the CABS-flex method, is the difficulty in reconstructing physically meaningful full-atom structures. We believe that this problem can be addressed by employing new deep learning networks, which would allow for the utilization of extensive knowledge of known structures in the task of reconstruction based on the coarse-grained data.

Another common challenge in reconstructing protein structures is the incorporation of nonstandard amino acids or PTMs. While CABS-flex is primarily designed to work with natural amino acids and does not directly support the integration of nonstandard amino acids or PTMs, users can approximate the effects of such modifications. For instance, users can substitute phosphoserine with aspartic acid or acetylsine with lysine to mimic the effects of these modifications. For a more precise analysis, researchers may start with a CABS-flex generated ensemble of structures and subsequently refine these models using more detailed MD simulations. Tools like CHARMM-GUI, which has recently incorporated the PDB Manipulator for easier handling of nonstandard amino acids and PTMs (Park et al., 2023), offer an advanced platform for such detailed studies. Additionally, the PyMOL plugin pyTM can be used to visualize and introduce common PTMs into models (Warnecke et al., 2014). This layered approach leverages the rapid simulation capabilities of CABS-flex alongside the detailed analysis possible with MD,

providing a framework for studying the structural and functional impacts of PTMs.

3.8 | AlphaFold's pLDDT: Bridging predictive confidence and protein flexibility

The pLDDT score, developed by AlphaFold, serves as a confidence metric for the predicted positions of amino acids within a protein structure. Regions with high pLDDT scores generally indicate high structural confidence and presumed rigidity, whereas lower scores often suggest potential flexibility where structural predictions are uncertain. Importantly, pLDDT scores can indicate flexibility even in regions with high scores if the predictive model, like AlphaFold, has incorporated multiple conformations in its training dataset.

This underscores the complexity of interpreting pLDDT scores, necessitating a nuanced approach to assessing protein flexibility. AF2 adaptations for modeling multiple protein states enhance the structural heterogeneity of predicted ensembles (Sala et al., 2023). These adaptations allow for the simulation of various conformational states beyond a

single rigid configuration, which is essential for proteins undergoing significant structural changes. The pLDDT metric reliably assesses the structural confidence of these predicted models, assigning higher values to more accurately predicted conformations.

Here, we analyze two example proteins from the ATLAS database (Vander Meersche et al., 2024), 6o2v_A and 1d3y_B, which were subjected to three 100 ns simulations using GROMACS v2019.4 with the CHARMM36m force field. By scaling the RMSF values yielded by CABS-flex alongside those from MD simulations, we maintained comparability and shape consistency of fluctuation values. Spearman's correlation was employed as a metric to measure the similarity between these two sets of fluctuation data.

For 6o2v_A, the correlation between CABS-flex and ATLAS's MD simulations was notably high at 0.86, indicating strong predictive accuracy by CABS-flex (Figure 9a). Interestingly, the shape of the pLDDT curve inversely mirrored the RMSF curves, leading to further analysis where RMSF and pLDDT values from MD were scaled together. This analysis resulted in a Spearman correlation of -0.80 , suggesting a strong negative interdependence, with most fluctuations aligning with protein loops that are prone to movement.

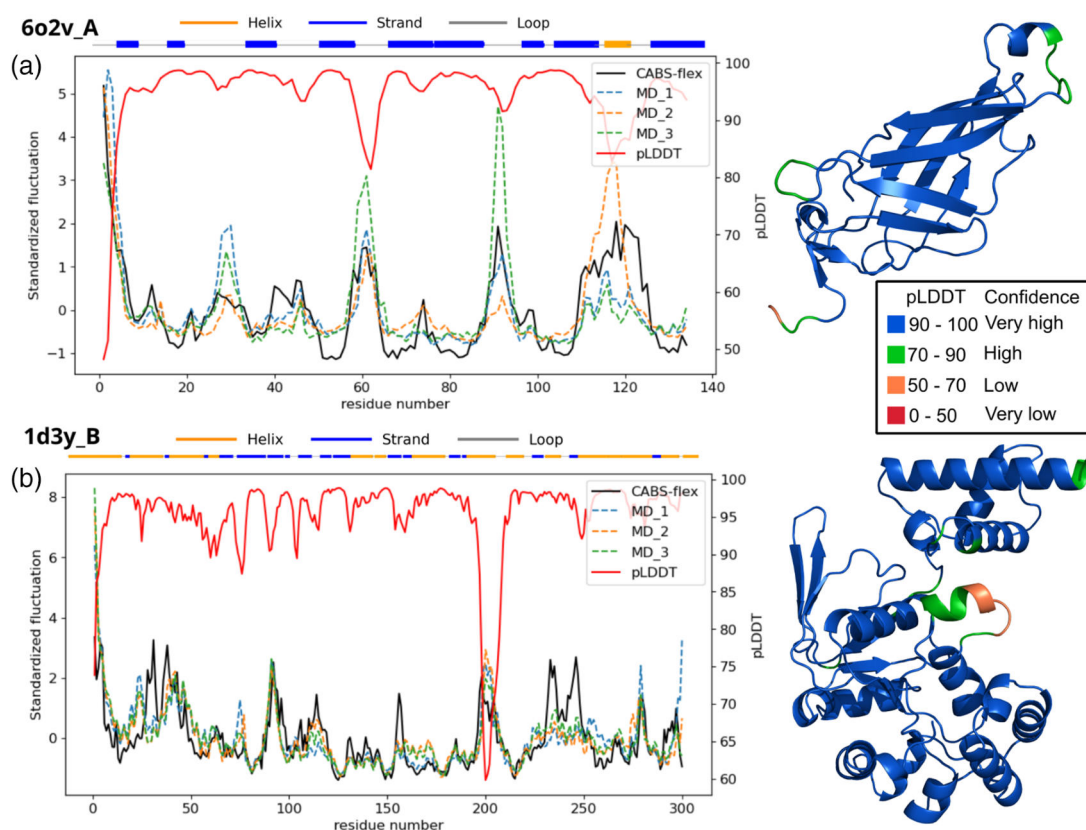


FIGURE 9 Examples of protein structure fluctuations from CABS-flex and molecular dynamics (MD) runs, and pLDDT confidence scores for proteins: (a) 6o2v_A and (b) 1d3y_B. MD runs were retrieved from the ATLAS database (Vander Meersche et al., 2024). On the right, the figure shows pLDDT confidence scores visualized on the protein structures. pLDDT, predicted local distance difference test.

Conversely, 1d3y_B, which is rich in secondary structure elements and has fewer loops, showed significant fluctuations loosely correlated with pLDDT scores (Figure 9b). Particularly in the region 195–205, a significant drop in pLDDT scores was observed. This region was not originally present in the PDB data and was computationally modeled by AF2 for inclusion in the ATLAS database (Vander Meersche et al., 2024), explaining the lower pLDDT scores. Such scores often suggest structural uncertainty, which, in this case, is a direct result of the absence of empirical data for this segment. Despite the low pLDDT, this region exhibits a level of flexibility comparable to other regions with much higher pLDDT scores, emphasizing the importance of source data in interpreting both pLDDT scores and flexibility.

These findings underscore the complex and context-dependent nature of using pLDDT scores to infer flexibility in protein structures. They highlight the importance of integrating these scores with dynamic simulation data from CABS-flex to obtain a comprehensive understanding of protein behavior, facilitating more informed predictions about protein functionality under diverse physiological conditions.

4 | CONCLUSIONS

CABS-flex has established itself as an efficient and cost-effective tool for simulating the backbone flexibility of proteins across diverse biological contexts. It has been particularly useful for applications in the analysis of globular proteins, mechanisms of viral adaptation, allostery, prediction of aggregation-prone and S-nitrosylation sites, as well as in structure prediction for linear and cyclic peptides, thereby enhancing our understanding of protein dynamics.

In this work, we have shown how CABS-flex effectively integrates into research workflows, complementing traditional experimental techniques such as x-ray crystallography, NMR spectroscopy, and cryo-electron microscopy (cryo-EM). This integration proves especially valuable in scenarios where experimental data are sparse, or where traditional methods like crystallization fail due to the dynamic nature of the target proteins. Additionally, CABS-flex has demonstrated its utility in studying the effects of mutations, further enhancing its versatility in addressing various biological questions. In the various workflows presented, CABS-flex is primarily used to simulate the flexibility of globular proteins at a low computational cost, offering a practical and less resource-intensive alternative to the more computationally demanding all-atom methods. This capability allows for the exploration of a wide range of protein conformations

and provides insights into dynamics that are often inaccessible through conventional experimental approaches alone, making CABS-flex a useful tool in computational biology for studying protein dynamics with reduced computational overhead.

Despite its advantages, the coarse-grained nature of CABS-flex, which relies on a knowledge-based force field, might not capture all detailed molecular interactions. This limitation is particularly evident in scenarios involving transient interactions, highly dynamic regions, or intricate regulatory mechanisms within proteins. Such interactions often require a detailed atomic-level understanding that coarse-grained models may oversimplify. This gap highlights the importance of supplementing CABS-flex with higher-resolution methods, such as all-atom MD simulations, to validate and refine computational predictions. Integrating these methods ensures a more comprehensive understanding of protein dynamics, particularly in understanding the subtle nuances of allosteric regulations and the complexity of protein–protein interaction interfaces.

From our experience, the adept application of distance restraints in CABS-flex simulations is crucial yet frequently overlooked by users. Often, a default set of distance restraints and simulation parameters is employed, optimized for the short-term dynamics of folded globular proteins as derived from all-atom MD simulations. Adjustments to these restraints are necessary for different protein systems and timescales to accommodate diverse functional states and conformations. Modifying these restraints presents challenges, requiring expertise and sometimes extensive customization. An interesting and important challenge is the design of distance restraints based on low-resolution experimental data, which can include techniques such as small-angle x-ray scattering (SAXS), cryo-EM, or cross-linking mass spectrometry (XL-MS). These methods provide valuable structural information that can guide the placement and adjustment of distance restraints. Recent significant advances in AI-based structural prediction have further expanded the possibilities for integrative modeling pipelines (Rout and Sali, 2019), which combine modeling methods with experimental data. For instance, AI-predicted models can now be integrated with data from MS (Allison et al., 2022) or cryo-electron tomography (Mosalaganti et al., 2022), setting the stage for simulating large-scale dynamics. Simplifying the process of adjusting restraints and their weights remains a key area for future development to enhance user accessibility and flexibility in simulation setups.

Additionally, exploring the integration of CABS-flex with advanced AI-based structural predictions, such as those from AF2, could be promising. The conceptual use

of AlphaFold's pLDDT scores to inform distance restraints in CABS-flex simulations might lead to more nuanced depictions of protein flexibility, particularly in regions where the predictive model's confidence varies. This approach could potentially enhance CABS-flex's capability to model protein dynamics more accurately, especially in areas that are critical for understanding functional mechanisms. Such integration could improve the utility of simulations, providing insights into protein behavior that are complementary to traditional experimental approaches.

As we advance, refining the CABS-flex algorithm and simplifying its user interface will be crucial for facilitating more precise simulations of large biological systems. Continued development will ensure that CABS-flex remains an useful tool in computational biology, aiding researchers in effectively exploring the dynamic nature of proteins and their complexes.

AUTHOR CONTRIBUTIONS

Chandran Nithin: Writing – original draft; writing – review and editing; visualization. **Rocco Peter Fornari:** Conceptualization; investigation; formal analysis; visualization; writing – original draft; writing – review and editing. **Smita P. Pilla:** Conceptualization; investigation; formal analysis; writing – original draft; writing – review and editing; visualization. **Karol Wroblewski:** Writing – original draft; investigation; formal analysis; visualization. **Mateusz Zalewski:** Investigation; visualization; writing – original draft. **Rafał Madaj:** Writing – original draft; investigation; visualization. **Andrzej Kolinski:** Writing – review and editing. **Joanna M. Macnar:** Writing – original draft; writing – review and editing; visualization. **Sebastian Kmiecik:** Conceptualization; writing – review and editing; writing – original draft; funding acquisition.

ACKNOWLEDGMENTS

S.K., R.P.F., S.P.P., C.N., and K.W. acknowledge funding from the National Science Centre, Poland [OPUS 2020/39/B/NZ2/01301]; S.K., J.M.M., and M.Z. acknowledge funding from the National Science Centre, Poland [2021/40/Q/NZ2/00078]; C.N. and S.P.P. acknowledge Polish high-performance computing infrastructure PLGrid (HPC Centers: ACK Cyfronet AGH) for providing computer facilities and support [computational grant no. PLG/2024/016931].

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

ORCID

Chandran Nithin  <https://orcid.org/0000-0001-8212-6093>

Rocco Peter Fornari  <https://orcid.org/0000-0002-3792-8262>

Smita P. Pilla  <https://orcid.org/0000-0002-4520-3210>

Karol Wroblewski  <https://orcid.org/0000-0001-5182-1376>

Mateusz Zalewski  <https://orcid.org/0000-0002-3436-8829>

Rafał Madaj  <https://orcid.org/0000-0002-6330-6344>

Andrzej Kolinski  <https://orcid.org/0000-0002-8830-2315>

Joanna M. Macnar  <https://orcid.org/0000-0003-1760-2668>

Sebastian Kmiecik  <https://orcid.org/0000-0001-7623-0935>

REFERENCES

- Allison TM, Degiacomi MT, Marklund EG, Jovine L, Elofsson A, Benesch JLP, et al. Complementing machine learning-based structure predictions with native mass spectrometry. *Protein Sci.* 2022;31:e4333.
- Amusengeri A, Astl L, Lobb K, Verkhivker GM, Tasthan Bishop Ö. Establishing computational approaches towards identifying malarial allosteric modulators: a case study of *Plasmodium falciparum* Hsp70s. *Int J Mol Sci.* 2019;20:5574.
- Badaczewska-Dawid A, Wróblewski K, Kurcinski M, Kmiecik S. Structure prediction of linear and cyclic peptides using CABS-flex. *Brief Bioinform.* 2024b;25:bbae003.
- Badaczewska-Dawid AE, Garcia-Pardo J, Kuriata A, Pujols J, Ventura S, Kmiecik S. A3D database: structure-based predictions of protein aggregation for the human proteome. *Bioinformatics.* 2022b;38:3121–3.
- Badaczewska-Dawid AE, Khramushin A, Kolinski A, Schueler-Furman O, Kmiecik S. Protocols for all-atom reconstruction and high-resolution refinement of protein–peptide complex structures. In: Kihara D, editor. *Protein structure prediction. Methods in molecular biology.* New York, NY: Springer US; 2020a. p. 273–87. https://doi.org/10.1007/978-1-0716-0708-4_16
- Badaczewska-Dawid AE, Kolinski A, Kmiecik S. Computational reconstruction of atomistic protein structures from coarse-grained models. *Comput Struct Biotechnol J.* 2020b;18:162–76.
- Badaczewska-Dawid AE, Kuriata A, Pintado-Grima C, Garcia-Pardo J, Burdukiewicz M, Iglesias V, et al. A3D model organism database (A3D-MODB): a database for proteome aggregation predictions in model organisms. *Nucleic Acids Res.* 2024a;52:D360–7.
- Badaczewska-Dawid AE, Nithin C, Wroblewski K, Kurcinski M, Kmiecik S. MAPIYA contact map server for identification and visualization of molecular interactions in proteins and biological complexes. *Nucleic Acids Res.* 2022a;50:W474–82.
- Bárceñas, O., Kuriata, A., Zalewski, M., Iglesias, V., Pintado-Grima, C., Firlik, G., Burdukiewicz, M., Kmiecik, S., & Ventura, S. Aggrescan4D: structure-informed analysis of pH-dependent protein aggregation. *Nucleic Acids Research.* 2024;52:W170–5. <https://doi.org/10.1093/nar/gkae382>
- Berjanskii MV, Wishart DS. A simple method to predict protein flexibility using secondary chemical shifts. *J Am Chem Soc.* 2005;127:14970–1.

- Berjanskii MV, Wishart DS. Application of the random coil index to studying protein flexibility. *J Biomol NMR*. 2008;40:31–48.
- Burley SK, Bhikadiya C, Bi C, Bittrich S, Chao H, Chen L, et al. RCSB Protein Data Bank (RCSB.Org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Res*. 2023;51:D488–508.
- Buß O, Rudat J, Ochsenreither K. FoldX as protein engineering tool: better than random based approaches? *Comput Struct Biotechnol J*. 2018;16:25–33.
- Cai Y, Zhang J, Xiao T, Lavine CL, Rawson S, Peng H, et al. Structural basis for enhanced infectivity and immune evasion of SARS-CoV-2 variants. *Science*. 2021;373:642–8.
- Cao Y, Wang J, Jian F, Xiao T, Song W, Yisimayi A, et al. Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. *Nature*. 2022;602:657–63.
- Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr*. 2010;66:12–21.
- Chiti F, Dobson CM. Amyloid formation by globular proteins under native conditions. *Nat Chem Biol*. 2009;5:15–22.
- Ciemny MP, Debinski A, Paczkowska M, Kolinski A, Kurcinski M, Kmiecik S. Protein-peptide molecular docking with large-scale conformational changes: the p53-MDM2 interaction. *Sci Rep*. 2016;6:37532.
- Faienza F, Lambrugh M, Rizza S, Pecorari C, Giglio P, Salamanca Vilorio J, et al. S-nitrosylation affects TRAP1 structure and ATPase activity and modulates cell response to apoptotic stimuli. *Biochem Pharmacol*. 2020;176:113869.
- Fowler NJ, Sljoka A, Williamson MP. A method for validating the accuracy of NMR protein structures. *Nat Commun*. 2020;11:6321.
- Fowler NJ, Sljoka A, Williamson MP. The accuracy of NMR protein structures in the Protein Data Bank. *Structure*. 2021;29:1430–1439.e2.
- Fowler NJ, Williamson MP. The accuracy of protein structures in solution determined by AlphaFold and NMR. *Structure*. 2022;30:925–933.e2.
- Garcia-Pardo J, Badaczewska-Dawid AE, Pintado-Grima C, Iglesias V, Kuriata A, Kmiecik S, et al. A3DyDB: exploring structural aggregation propensities in the yeast proteome. *Microb Cell Factories*. 2023;22:186.
- Guo H-B, Perminov A, Bekele S, Kedziora G, Farajollahi S, Varaljay V, et al. AlphaFold2 models indicate that protein sequence determines both structure and dynamics. *Sci Rep*. 2022;12:10696.
- He X, Hong W, Pan X, Lu G, Wei X. SARS-CoV-2 omicron variant: characteristics and prevention. *MedComm*. 2021;2:838–45.
- Heo L, Feig M. One bead per residue can describe all-atom protein structures. *Structure*. 2024;32:97–111.e6.
- Hoie MH, Kiehl EN, Petersen B, Nielsen M, Winther O, Nielsen H, et al. NetSurfP-3.0: accurate and fast prediction of protein structural features by protein language models and deep learning. *Nucleic Acids Res*. 2022;50:W510–5.
- Huang Y, Yang C, Xu X, Xu W, Liu S. Structural and functional properties of SARS-CoV-2 spike protein: potential antiviral drug development for COVID-19. *Acta Pharmacol Sin*. 2020;41:1141–9.
- Jamroz M, Kolinski A, Kmiecik S. CABS-flex: server for fast simulation of protein structure fluctuations. *Nucleic Acids Res*. 2013b;41:W427–31.
- Jamroz M, Kolinski A, Kmiecik S. CABS-flex predictions of protein flexibility compared with NMR ensembles. *Bioinformatics*. 2014;30:2150–4.
- Jamroz M, Orozco M, Kolinski A, Kmiecik S. Consistent view of protein fluctuations from all-atom molecular dynamics and coarse-grained dynamics with knowledge-based force-field. *J Chem Theory Comput*. 2013a;9:119–25.
- Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*. 1999;292:195–202.
- Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596:583–9.
- Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 1983;22:2577–637.
- Karim SSA, Karim QA. Omicron SARS-CoV-2 variant: a new chapter in the COVID-19 pandemic. *Lancet*. 2021;398:2126–8.
- Kmiecik S, Gront D, Kolinski M, Wieteska L, Dawid AE, Kolinski A. Coarse-grained protein models and their applications. *Chem Rev*. 2016;116:7898–936.
- Kmiecik S, Gront D, Kouza M, Kolinski A. From coarse-grained to atomic-level characterization of protein dynamics: transition state for the folding of B domain of protein A. *J Phys Chem B*. 2012;116:7026–32.
- Kmiecik S, Kolinski A. Characterization of protein-folding pathways by reduced-space modeling. *Proc Natl Acad Sci*. 2007;104:12330–5.
- Kmiecik S, Kolinski A. Folding pathway of the B1 domain of protein G explored by multiscale modeling. *Biophys J*. 2008;94:726–36.
- Kolinski A. Protein modeling and structure prediction with a reduced representation. *Acta Biochim Pol*. 2004;51:349–71.
- Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell*. 2020;182:812–827.e19.
- Kurcinski M, Badaczewska-Dawid A, Kolinski M, Kolinski A, Kmiecik S. Flexible docking of peptides to proteins using CABS-dock. *Protein Sci*. 2020;29:211–22.
- Kurcinski M, Kolinski A, Kmiecik S. Mechanism of folding and binding of an intrinsically disordered protein as revealed by ab initio simulations. *J Chem Theory Comput*. 2014;10:2224–31.
- Kurcinski M, Oleniecki T, Ciemny MP, Kuriata A, Kolinski A, Kmiecik S. CABS-flex standalone: a simulation environment for fast modeling of protein flexibility. *Bioinformatics*. 2019;35:694–5.
- Kuriata A, Gierut AM, Oleniecki T, Ciemny MP, Kolinski A, Kurcinski M, et al. CABS-flex 2.0: a web server for fast simulations of flexibility of protein structures. *Nucleic Acids Res*. 2018;46:W338–43.
- Kuriata A, Iglesias V, Kurcinski M, Ventura S, Kmiecik S. Aggrescan3D standalone package for structure-based prediction of protein aggregation properties. *Bioinformatics*. 2019b;35:3834–5.
- Kuriata A, Iglesias V, Pujols J, Kurcinski M, Kmiecik S, Ventura S. Aggrescan3D (A3D) 2.0: prediction and engineering of protein solubility. *Nucleic Acids Res*. 2019a;47:W300–7.
- Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*. 2023;379:1123–30.
- Ma P, Li D-W, Brüschweiler R. Predicting protein flexibility with AlphaFold. *Proteins*. 2023;91:847–55.
- Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. *Nat Methods*. 2022;19:679–82.

- Mishra PM, Anjum F, Uversky VN, Nandi CK. SARS-CoV-2 spike mutations modify the interaction between virus spike and human ACE2 receptors. *Biochem Biophys Res Commun.* 2022; 620:8–14.
- Monzon AM, Rohr CO, Fornasari MS, Parisi G. CoDNas 2.0: a comprehensive database of protein conformational diversity in the native state. *Database.* 2016;2016:baw038.
- Mosalaganti S, Obarska-Kosinska A, Siggel M, Taniguchi R, Turoňová B, Zimmerli CE, et al. AI-based structure prediction empowers integrative structural analysis of human nuclear pores. *Science.* 2022;376:eabm9506.
- Navarro S, Ventura S. Computational re-design of protein structures to improve solubility. *Expert Opin Drug Discov.* 2019;14:1077–88.
- Ortega JT, Jastrzebska B, Rangel HR. Omicron SARS-CoV-2 variant spike protein shows an increased affinity to the human ACE2 receptor: an in silico analysis. *Pathogens.* 2022;11:45.
- Papaleo E, Tiberti M, Arnaudi M, Pecorari C, Faienza F, Cantwell L, et al. TRAP1 S-nitrosylation as a model of population-shift mechanism to study the effects of nitric oxide on redox-sensitive oncoproteins. *Cell Death Dis.* 2023;14:284.
- Park S-J, Kern N, Brown T, Lee J, Im W. CHARMM-GUI PDB manipulator: various PDB structural modifications for biomolecular modeling and simulation. *J Mol Biol.* 2023;435:167995.
- Puławski W, Koliński A, Koliński M. Integrative modeling of diverse protein-peptide systems using CABS-dock. *PLoS Comput Biol.* 2023;19:e1011275.
- Rizza S, Montagna C, Cardaci S, Maiani E, Di Giacomo G, Sanchez-Quiles V, et al. S-nitrosylation of the mitochondrial chaperone TRAP1 sensitizes hepatocellular carcinoma cells to inhibitors of succinate dehydrogenase. *Cancer Res.* 2016;76:4170–82.
- Rout MP, Sali A. Principles for integrative structural biology studies. *Cell.* 2019;177:1384–403.
- Ruff KM, Pappu RV. AlphaFold and implications for intrinsically disordered proteins. *J Mol Biol.* 2021;433:167208.
- Sala D, Engelberger F, Mchaourab HS, Meiler J. Modeling conformational states of proteins with AlphaFold. *Curr Opin Struct Biol.* 2023;81:102645.
- Saldaño T, Escobedo N, Marchetti J, Zea DJ, Mac Donagh J, Velez Rueda AJ, et al. Impact of protein conformational diversity on AlphaFold predictions. *Bioinformatics.* 2022;38:2742–8.
- Sanyal D, Banerjee S, Bej A, Chowdhury VR, Uversky VN, Chowdhury S, et al. An integrated understanding of the evolutionary and structural features of the SARS-CoV-2 spike receptor binding domain (RBD). *Int J Biol Macromol.* 2022;217:492–505.
- Schwarz D, Georges G, Kelm S, Shi J, Vangone A, Deane CM. Co-evolutionary distance predictions contain flexibility information. *Bioinformatics.* 2021;38:65–72.
- Shen Y, Maupetit J, Derreumaux P, Tufféry P. Improved PEP-FOLD approach for peptide and miniprotein structure prediction. *J Chem Theory Comput.* 2014;10:4745–58.
- Singh J, Paliwal K, Litfin T, Singh J, Zhou Y. Reaching alignment-profile-based accuracy in predicting protein secondary and tertiary structural properties without alignment. *Sci Rep.* 2022;12:7607.
- Timmons PB, Hewage CM. APPTTEST is a novel protocol for the automatic prediction of peptide tertiary structures. *Brief Bioinform.* 2021;22:bbab308.
- Tortorici MA, Czudnochowski N, Starr TN, Marzi R, Walls AC, Zatta F, et al. Broad sarbecovirus neutralization by a human monoclonal antibody. *Nature.* 2021;597:103–8.
- Vander Meersche Y, Cretin G, Gheeraert A, Gelly J-C, Galochkina T. ATLAS: protein flexibility description from atomistic molecular dynamics simulations. *Nucleic Acids Res.* 2024;52:D384–92.
- Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* 2022;50:D439–44.
- Verkhivker GM. Conformational flexibility and local frustration in the functional states of the SARS-CoV-2 spike B.1.1.7 and B.1.351 variants: mutation-induced allosteric modulation mechanism of functional dynamics and protein stability. *Int J Mol Sci.* 2022a;23:1646.
- Verkhivker GM. Conformational dynamics and mechanisms of client protein integration into the Hsp90 chaperone controlled by allosteric interactions of regulatory switches: perturbation-based network approach for mutational profiling of the Hsp90 binding and allostery. *J Phys Chem B.* 2022b;126:5421–42.
- Verkhivker GM. Exploring mechanisms of allosteric regulation and communication switching in the multiprotein regulatory complexes of the Hsp90 chaperone with cochaperones and client proteins: atomistic insights from integrative biophysical modeling and network analysis of conformational landscapes. *J Mol Biol.* 2022c;434:167506.
- Verkhivker GM, Agajanian S, Kassab R, Krishnan K. Landscape-based protein stability analysis and network modeling of multiple conformational states of the SARS-CoV-2 spike D614G mutant: conformational plasticity and frustration-induced allostery as energetic drivers of highly transmissible spike variants. *J Chem Inf Model.* 2022;62:1956–78.
- Verkhivker GM, Di Paola L. Dynamic network modeling of allosteric interactions and communication pathways in the SARS-CoV-2 spike trimer mutants: differential modulation of conformational landscapes and signal transmission via cascades of regulatory switches. *J Phys Chem B.* 2021a;125:850–73.
- Verkhivker GM, Di Paola L. Integrated biophysical modeling of the SARS-CoV-2 spike protein binding and allosteric interactions with antibodies. *J Phys Chem B.* 2021b;125:4596–619.
- Warnecke A, Sandalova T, Achour A, Harris RA. PyTMs: a useful PyMOL plugin for modeling common post-translational modifications. *BMC Bioinformatics.* 2014;15:370.
- Webb B, Sali A. Comparative protein structure modeling using MODELLER. *Curr Protoc Bioinformatics.* 2016;54:5.6.1–5.6.37.
- Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh C-L, Abiona O, et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science.* 2020;367:1260–3.
- Zambrano R, Jamroz M, Szczasiuk A, Pujols J, Kmiecik S, Ventura S. AGGREGSCAN3D (A3D): server for prediction of aggregation properties of protein structures. *Nucleic Acids Res.* 2015;43:W306–13.
- Zhang N-N, Zhang R-R, Zhang Y-F, Ji K, Xiong X-C, Qin Q-S, et al. Rapid development of an updated mRNA vaccine against the SARS-CoV-2 omicron variant. *Cell Res.* 2022;32:401–3.

How to cite this article: Nithin C, Fornari RP, Pilla SP, Wroblewski K, Zalewski M, Madaj R, et al. Exploring protein functions from structural flexibility using CABS-flex modeling. *Protein Science.* 2024;33(9):e5090. <https://doi.org/10.1002/pro.5090>

4.2. [A6] Rozwój metody CABS-flex 3.0: narzędzia online do symulacji elastyczności strukturalnej białek i modelowania peptydów

4.2.1. Wprowadzenie do [A6]

W pracy [A6] przedstawiliśmy nową wersję narzędzia - CABS-flex 3.0 - stanowiącą bezpośrednie rozwinięcie wcześniejszych badań nad metodą CABS-flex. Projekt miał charakter złożony oraz długoterminowy. Mój wkład w publikację był jednym z wiodących i porównywalny z wkładem Karola Wróblewskiego, co zostało formalnie podkreślone adnotacją o równorzędnym wkładzie (tzw. *equal contribution*). Skupiliśmy się jednak na odmiennych zakresach zadań - K. Wróblewski odpowiadał głównie za rozwój algorytmiczny i implementacyjny metody, natomiast ja za opracowanie oraz dopracowanie złożonego interfejsu webowego.

CABS-flex 2.0, dostępny wcześniej jako ogólnodostępny serwer internetowy, był szeroko wykorzystywany w środowisku naukowym i znalazł zastosowanie w licznych badaniach: od analizy elastyczności białek globularnych, poprzez ocenę wpływu mutacji i identyfikację miejsc allosterycznych, aż po integrację z predykcjami AlphaFold i wykorzystanie w protokołach dynamicznych do analizy agregacji i S-nitrozylacji.

Znaczące zainteresowanie społeczności naukowej, potwierdzone licznymi cytowaniami, szerokie zastosowanie narzędzia oraz opinie użytkowników i ograniczenia wersji 2.0 stały się naturalną motywacją do opracowania jego kolejnej generacji. Wersja 3.0 wprowadziła szereg udoskonaleń, takich jak nowe tryby elastyczności z uwzględnieniem pLDDT, możliwość predykcji struktury i elastyczności peptydów (zarówno liniowych, jak i cyklicznych), nowy algorytm rekonstrukcji pełnoatomowej cg2all¹¹², a także całkowicie przebudowany, interaktywny interfejs webowy. Kluczowym celem było zwiększenie dokładności, elastyczności i użyteczności metody - przy zachowaniu jej podstawowej zalety, jaką jest wysoka efektywność obliczeniowa wynikająca z wykorzystania modelu gruboziarnistego CABS, który opisałem w rozdziale 2.2.2.

W wyniku prac nad serwerem CABS-flex 3.0 zaproponowaliśmy cztery zróżnicowane tryby definiowania więzów odległości, odpowiadające różnym zakresom ruchliwości i sztywności struktur białkowych:

- Flexible - tryb domyślny, w którym więzy odległości definiowane są wyłącznie pomiędzy resztami należącymi do stabilnych elementów struktury drugorzędowej (helisy i β -harmonijek). Pozwala to na zachowanie lokalnej stabilności tych regionów przy jednoczesnym umożliwieniu swobodnych zmian konformacyjnych w obrębie pętli i regionów nieuporządkowanych^{53,55,71,113}.

- Rigid - tryb o jednolitym poziomie sztywności, w którym więzy odległości nakładane są na całą strukturę białka, a nie tylko na elementy struktury drugorzędowej, co utrzymuje ją w możliwie niezmięnionej konformacji. Tryb ten został opracowany i skalibrowany w oparciu o dane z bazy ATLAS MD¹¹⁴, zawierającej wysokorozdzielcze monomery białkowe o ograniczonych fluktuacjach konformacyjnych.⁵⁴
- Rigid-pLDDT - tryb hybrydowy, w którym siła ograniczeń modulowana jest na podstawie wartości pLDDT pochodzących z predykcji AlphaFold. Podobnie jak tryb Rigid, został on opracowany i zweryfikowany w oparciu o dane z bazy ATLAS MD¹¹⁴, co pozwoliło na dopasowanie parametrów ograniczeń do rzeczywistych poziomów ruchliwości obserwowanych w monomerach białkowych⁵⁴.
- Unleashed - tryb całkowicie pozbawiony ograniczeń, pozwalający na maksymalnie swobodne próbkowanie przestrzeni konformacyjnej. W tym trybie to pole siłowe CABS w pełni decyduje o generowanej dynamice. Jak pokazano w pracy¹¹⁵, nawet przy całkowitym braku więzów fluktuacje uzyskiwane w modelu CABS dobrze korelują z wynikami pełnoatomowych symulacji MD.

W artykule [A6] wyjaśniliśmy szczegółowo zalecany sposób wyboru trybu - w zależności od typu białka i dostępnych danych, a także możliwość równoległego testowania różnych trybów w celu identyfikacji najbardziej adekwatnego modelu

CABS-flex 3.0 oferuje dedykowany tryb do modelowania peptydów, opracowany w oparciu o pracę Badaczewskiej i wsp.⁵⁵, który umożliwia generowanie struktur peptydów liniowych i cyklicznych. Protokół uwzględnia obecność mostków disiarczkowych oraz możliwość zamknięcia łańcucha głównego. W przypadku braku informacji o strukturze drugorzędowej serwer automatycznie wykonuje jej predykcję z wykorzystaniem narzędzia NetSurfP-3.0¹¹⁶.

Zaimplementowany protokół został oceniony na szerokim zbiorze testowym obejmującym peptydy o zróżnicowanej długości, cykliczności i charakterze strukturalnym.⁵⁵ Wykazaliśmy, że CABS-flex osiąga wyniki porównywalne z innymi narzędziami do predykcji struktury peptydów - szczególnie w przypadku krótkich peptydów liniowych - przy istotnie niższym koszcie obliczeniowym niż metody takie jak AlphaFold¹⁰⁴, PEP-FOLD¹¹⁷, APPTTEST¹¹⁸ czy ESMFold³³. Dodatkowym atutem CABS-flex, odróżniającym je od większości wymienionych metod, jest możliwość analizy elastyczności peptydów, a nie wyłącznie predykcji pojedynczej struktury statycznej.

Ważnym aspektem rozwoju wersji CABS-flex 3.0 była również rekonstrukcja pełnoatomowa struktur uzyskiwanych z symulacji gruboziarnistych. W dotychczasowych wersjach narzędzia CABS-flex proces ten realizowany był z wykorzystaniem programu Modeller, który w wielu przypadkach prowadził do lokalnych zniekształceń geometrii łańcucha głównego oraz suboptymalnych konformacji łańcuchów bocznych. W wersji 3.0 podejście to zostało zastąpione metodą cg2all¹¹², opartą na architekturze głębokiego uczenia typu SE(3)

transformer, wykorzystującej sztywną reprezentację bloków strukturalnych inspirowaną rozwiązaniami zastosowanymi w AlphaFold2.

Testy rekonstrukcji struktur atomowych po modelowaniu w CABS wykazały, że metoda cg2all istotnie poprawia jakość lokalnej geometrii w porównaniu do wcześniejszych procedur, redukując kolizje atomowe oraz lepiej zachowując poprawne relacje przestrzenne zarówno w szkieletcie peptydowym, jak i w konformacjach łańcuchów bocznych. Zastosowanie cg2all przekłada się tym samym na zwiększoną precyzję rekonstrukcji pełnoatomowej, wyższą przydatność wygenerowanych modeli do dalszych analiz strukturalnych i energetycznych, a także znaczną poprawę jakości wizualizacji wyników symulacji.

Interfejs serwera CABS-flex 3.0 zaprojektowałem i wykonałem tak, aby umożliwić pełną obsługę symulacji bez potrzeby znajomości linii komend. Dostępne są dwie zakładki wejściowe (*Flexibility modelling* oraz *Peptide modelling*), szereg opcji zaawansowanych (modyfikacja parametrów symulacji, ograniczenia niestandardowe, edytor elastyczności na poziomie reszt), a także cztery interaktywne zakładki wynikowe:

- Models - wizualizacja 10 wygenerowanych modeli w przeglądarce mol* z opcją kolorowania wg RMSF, pLDDT, B-factor itd.
- Contact maps - interaktywne mapy kontaktów generowane z użyciem Plotly, zsynchronizowane z wizualizacją 3D.
- Fluctuation plot - wykresy RMSF i pLDDT z adnotacją struktury drugorzędowej, umożliwiające lokalizację regionów stabilnych i zmiennych.
- Project info - zbiorcze informacje o symulacji i wprowadzonych danych.

Przykładowy wygląd interfejsu serwera CABS-flex 3.0 przedstawiono na Rysunku 8. Serwer umożliwia nie tylko przeprowadzenie symulacji, ale również interaktywną analizę i wizualizację wyników.

Project information

Models

Contact maps

Fluctuation plot

Structure

Models

All models

Flexibility coloring:

Structural coloring:

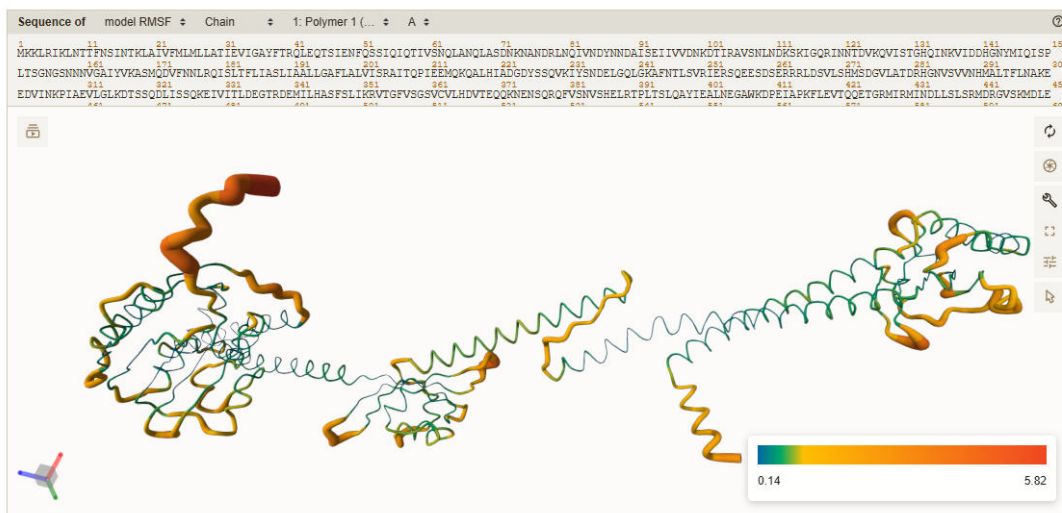
RMSF

Download data [tar.gz]

Show initial structure

Disulfide Bridges

Reset View



Project information

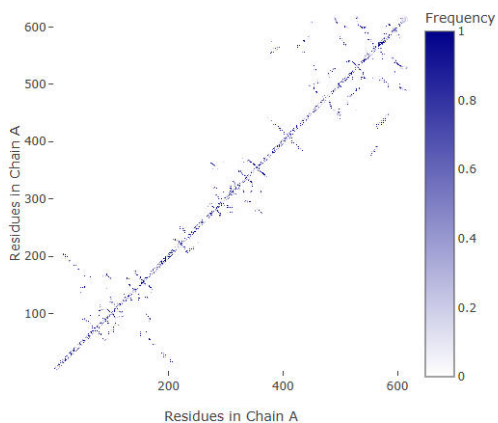
Models

Contact maps

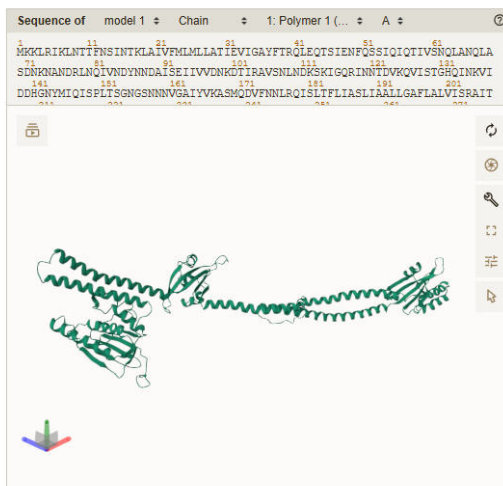
Fluctuation plot

A vs A

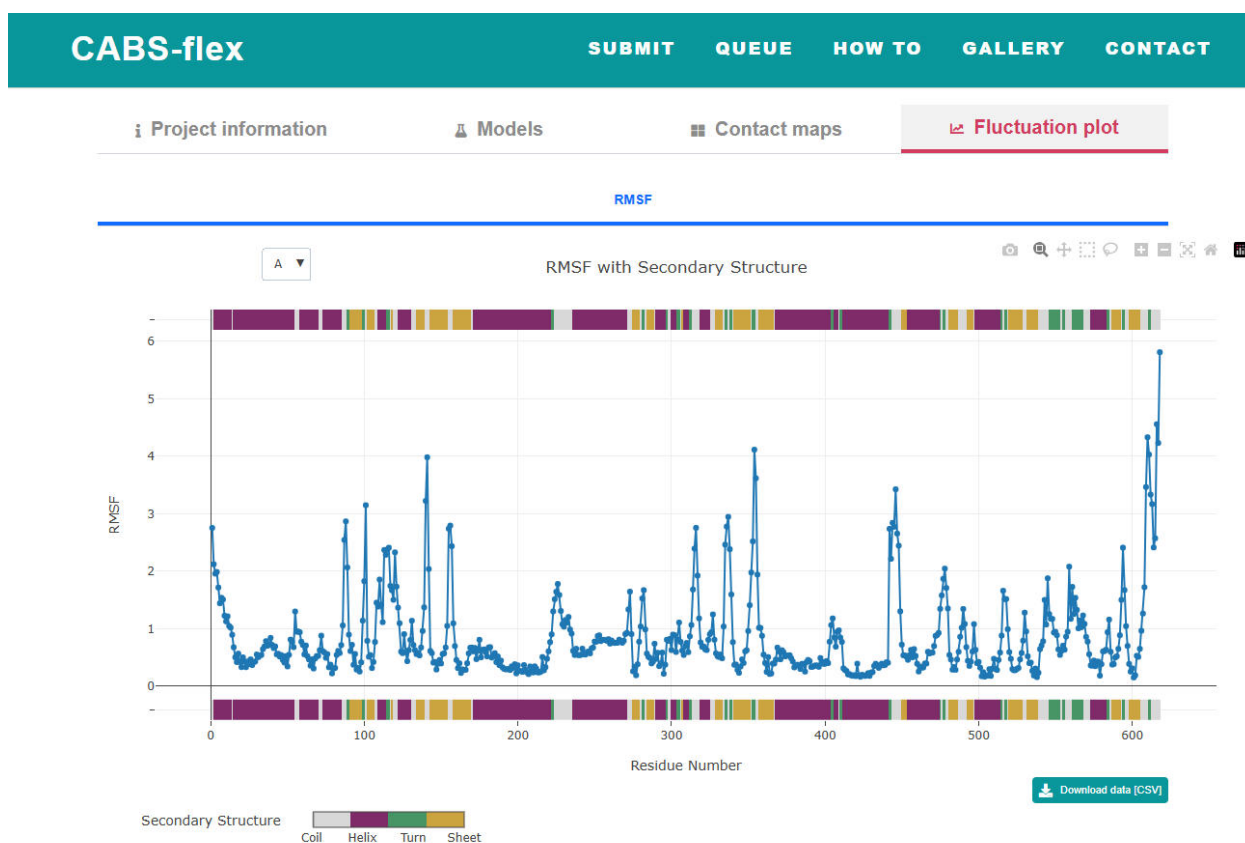
Contact Map: A



Download data [tar.gz]



Select Disulfide Bridges



Rysunek 8. Przykłady możliwości wizualizacji w ramach interfejsu CABS-flex 3.0: wizualizacja 3D wygenerowanych modeli (zakładka Models), interaktywna mapa kontaktów (zakładka Contact maps) oraz wykres fluktuacji RMSF wraz z adnotacjami dotyczącymi struktury drugorzędowej (zakładka Fluctuation plot). Tak zaprojektowany interfejs pozwala w prosty i intuicyjny sposób analizować wyniki bez konieczności korzystania z dodatkowego oprogramowania. Źródło: [A6].

Architektura serwera bazuje na frameworku Flask oraz silniku szablonów Jinja2, z kolejką obliczeniową obsługiwaną przez MySQL. Do renderowania struktur wykorzystałem bibliotekę mol*¹¹⁹.

Wprowadzone w CABS-flex 3.0 udoskonalenia stanowiły odpowiedź na konkretne potrzeby badaczy analizujących powiązania pomiędzy strukturą, elastycznością i funkcją białek oraz peptydów. Nowa wersja narzędzia zwiększyła dostępność symulacji elastyczności również dla mniej doświadczonych użytkowników, oferując jednocześnie tryby zaawansowane umożliwiające precyzyjne dostrojenie symulacji do danych eksperymentalnych.

Dzięki dużej szybkości, możliwości pracy na strukturach otrzymanych za pomocą narzędzia AlphaFold, opcji analizy interakcji i dostępności bez logowania, CABS-flex 3.0 znajduje zastosowanie zarówno w badaniach podstawowych (np. zmienność konformacyjna, wiązanie

peptydów), jak i w badaniach aplikacyjnych (np. projektowanie peptydów bioaktywnych, ocena wpływu mutacji).

W mojej ocenie znaczenie wersji 3.0 wykracza jednak poza samo rozszerzenie funkcjonalności i stanowi istotny wkład w rozwój metod szybkiej analizy elastyczności struktur białkowych. Obserwowany w ostatnich latach gwałtowny wzrost liczby modeli białek dostępnych dzięki metodzie AlphaFold przełożył się na ogromne zainteresowanie serwerem CABS-flex - najpierw wersją 2.0, a obecnie 3.0 - który pozwala szybko badać i wizualizować dynamikę tych struktur na zupełnie nowym poziomie. Dzięki temu narzędzie to znajduje zastosowanie w badaniach nad mechanizmami działania białek, projektowaniu leków i analizie wiarygodności modeli generowanych metodami AI. Uważam, że CABS-flex 3.0 jest dziś jednym z najbardziej uniwersalnych narzędzi do badania elastyczności białek i peptydów, łącząc efektywność obliczeniową z funkcjonalnością, która realnie wspiera badania od etapu wstępnych hipotez aż po szczegółowe analizy strukturalne.

4.2.2. Artykuł [A6]

CABS-flex 3.0: an online tool for simulating protein structural flexibility and peptide modeling

Karol Wróblewski[†], Mateusz Zalewski[†], Aleksander Kuriata, Sebastian Kmiecik^{✉*}

University of Warsaw, Biological and Chemical Research Centre, Faculty of Chemistry, 02-089 Warsaw, Poland

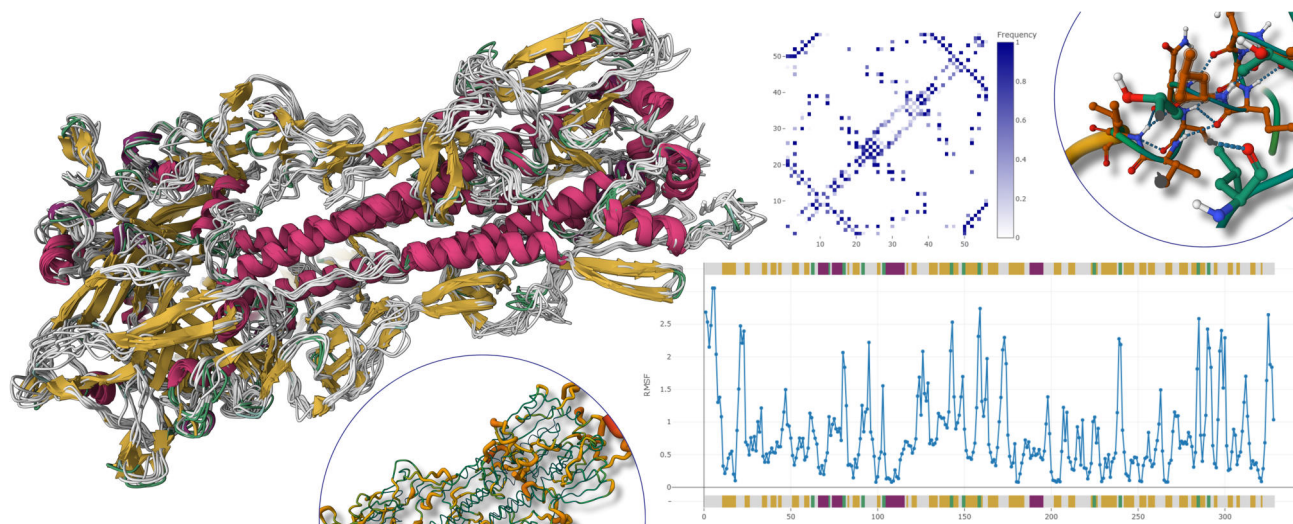
*To whom correspondence should be addressed. Tel: +48 22 552 6607; Fax: +48 22 552 6607; Email: sekmi@chem.uw.edu.pl

[†]The first two authors should be regarded as Joint First Authors.

Abstract

Simulating protein structure flexibility using classical methods is computationally demanding, especially for large proteins. To address this challenge, we have been developing the CABS-flex method, which enables fast simulations of protein structural flexibility by combining a coarse-grained simulation approach with all-atom detail. Previously available as the CABS-flex 2.0 web server, the method has now undergone a major upgrade with the release of CABS-flex 3.0. Key improvements include the introduction of intuitive flexibility modes that simplify the control of distance restraints and allow users to reflect known or expected dynamic regions; improved all-atom reconstruction for higher-quality model generation; a new feature for *de novo* peptide structure prediction, supporting both linear and cyclic peptides along with their conformational flexibility; and new tools for result analysis and visualization, facilitating deeper insights into structural flexibility. Additionally, AlphaFold pLDDT-derived restraints can be used as optional input for guiding simulations. The method accepts input as either a PDB/mmCIF structure or a sequence (for peptide modeling). Advanced options allow users to incorporate experimental or computational restraints. The CABS-flex 3.0 web server is available at <https://lcbio.pl/cabsflex3>. This website is free and open to all users, with no login requirement.

Graphical abstract



Introduction

Protein structural flexibility is essential for their biological functions, yet studying protein dynamics experimentally is often challenging or even impossible. Consequently, computational simulations have become indispensable. However, classical approaches, such as all-atom molecular dynamics (MD), are constrained by their high computational cost, making the simulation of structural flexibility for most biologically relevant

proteins particularly demanding. Performing such simulations typically requires specialized hardware, such as supercomputers, or substantial computational resources.

To address these challenges, we introduced CABS-flex approximately a decade ago as a method designed for rapid simulations of protein flexibility. CABS-flex offers a computational speed advantage of three to four orders of magnitude over all-atom MD while maintaining competitive

Received: March 13, 2025. Revised: April 22, 2025. Editorial Decision: April 29, 2025. Accepted: May 2, 2025

© The Author(s) 2025. Published by Oxford University Press on behalf of Nucleic Acids Research.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License

(<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

accuracy in predicting structural flexibility [1]. Built upon the well-established CABS coarse-grained model, this approach effectively captures key dynamic properties observed in experimental and MD-derived datasets [1–3]. Its integration with all-atom reconstruction enables a seamless transition from coarse-grained flexibility simulations to detailed atomic-level representations. Over the years, we have continuously refined the method, introducing two iterations of the CABS-flex web server [4, 5], a standalone application with enhanced control and scalability [6], and, more recently, extensions that support peptide modeling and refined structural analysis [7]. Today, CABS-flex has gained widespread adoption, particularly through its popular web server 2.0, making flexibility simulations more accessible to the scientific community. Furthermore, it has become an integral component of various computational protocols, enabling researchers to explore the interplay between protein structure, flexibility, and function across diverse fields, as summarized in a recent review [8].

Previous versions of CABS-flex, although widely adopted, had several limitations. These included limited options for customizing flexibility restraints, no dedicated support for peptide modeling, and an all-atom reconstruction method that could lead to suboptimal side-chain conformations and local backbone geometry. Moreover, earlier versions did not incorporate structural confidence metrics such as AlphaFold's pLDDT scores. CABS-flex 3.0 addresses these issues by introducing new flexibility modes with adjustable restraint schemes, a protocol for modeling both linear and cyclic peptides, and an advanced deep learning-based all-atom reconstruction method (cg2all) that substantially improves structural detail [9]. The integration of pLDDT-derived restraints further enhances the biological relevance of flexibility predictions. Together, these improvements significantly expand the scope, accuracy, and usability of CABS-flex for modeling the dynamic behavior of proteins and peptides. These enhancements build on our recent research and are introduced in the “Materials and methods” section.

Materials and methods

CABS-flex combines a fast coarse-grained modeling engine with all-atom modeling, providing an efficient approach for simulating protein flexibility. Our recent review [8] provides a comprehensive overview of the CABS-flex methodology, its applications in structure-function studies, and its role in investigating the relationship between protein structure, dynamics, and function. The current version (3.0) of the CABS-flex web server builds upon the capabilities of its predecessor, CABS-flex 2.0 [5], as well as the CABS-flex standalone application [6], introducing significant improvements in structural flexibility modeling and a new peptide modeling mode. These advancements in CABS-flex 3.0 are based on recent studies, which we briefly reference below as the foundation for its key developments.

In CABS-flex 3.0, we implemented a new approach to distance restraints, as described in a recent study [3]. This work introduces a classification system for different flexibility categories, assigning distance restraints based on secondary structure and AlphaFold-derived pLDDT scores. This methodology enhances the predictive accuracy of CABS-flex while maintaining computational efficiency. We validated this approach by comparing fluctuation profiles obtained from CABS simulations with data from the ATLAS database [10], which

contains MD simulations of nearly 1400 proteins. By leveraging pLDDT scores, which indicate the confidence of AlphaFold's structure predictions, CABS-flex achieves improved accuracy in structural flexibility modeling while preserving its speed advantage over traditional all-atom MD simulations [3]. Nevertheless, when pLDDT scores are unavailable or unreliable, users can choose from the established Flexible and Rigid modes, which do not depend on AlphaFold predictions. These modes have been extensively benchmarked in previous studies [1–3, 8], demonstrating reliable performance across a wide range of protein systems.

Another major addition in CABS-flex 3.0 is the peptide modeling protocol, which was developed using the CABS-flex standalone application [6] and is now incorporated into the web server. This protocol has been benchmarked against state-of-the-art tools such as AlphaFold, PEP-FOLD, APPTTEST, and ESMFold [7]. While AlphaFold generally provided the highest accuracy, CABS-flex proved competitive, particularly for short linear peptides. The proposed approach also enables the modeling of cyclic peptides and allows the incorporation of disulfide bridge information. Apart from structural prediction, the protocol also provides insights into peptide flexibility.

A further significant improvement in CABS-flex 3.0 is the integration of a deep-learning-based cg2all method for all-atom reconstruction [9]. In our practice, cg2all provides a substantial enhancement over the previous reconstruction approach based on Modeller [11], particularly in terms of local structural quality, and we consider it the best currently available method. This improvement likely results from cg2all's ability to learn complex side-chain and backbone reconstruction patterns beyond what is possible with traditional optimization methods. The cg2all approach employs an SE(3) transformer architecture and a rigid-body block representation inspired by AlphaFold2. Its integration into CABS-flex 3.0 significantly enhances the accuracy and reliability of reconstructed structures, improving both visualization and downstream analyses of protein flexibility.

Web server input options

The CABS-flex 3.0 web server is designed to operate in two distinct modes: flexibility modeling and peptide modeling, see pipeline in Fig. 1. For ease of use, each mode is accessible through its own dedicated tab on the main page, where users can enter their input data.

Flexibility modeling

Basic input options

The only required input for flexibility modeling is the protein structure, provided either as a PDB code or an uploaded file in PDB/mmCIF format. Due to computational limitations, input structures must contain no more than 2000 residues and may include up to 10 chains. Users may optionally select specific chains using the “Chain(s)” field. To help manage submitted jobs, additional optional fields include a “Project name” (displayed in the queue list unless the “Do not show my job on the results page” option is selected) and an “Email address” for receiving job completion notifications.

Flexibility modes

CABS-flex 3.0 introduces four flexibility modes—Flexible, Rigid, Rigid-pLDDT, and Unleashed—that define how dis-

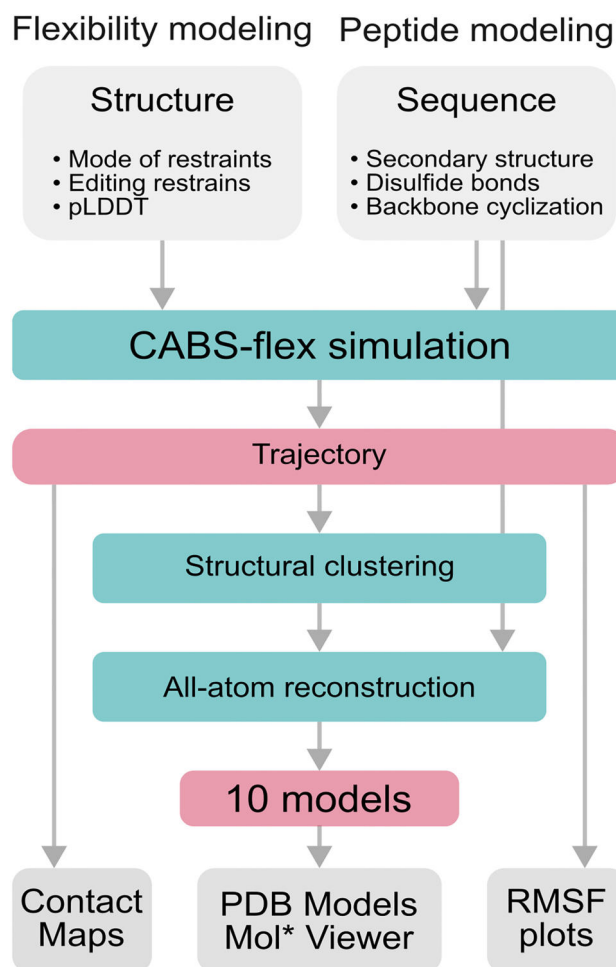


Figure 1. CABS-flex 3.0 pipeline. The image shows some of the input options. The primary output, which is then analyzed using the web interface, includes the simulation trajectory and 10 models.

tance restraints are applied across the structure and allow users to control the extent of conformational sampling. A representative example of their effect on backbone fluctuations is shown in Fig. 2. Each mode may be appropriate for different modeling contexts, with names that broadly reflect the expected levels of structural flexibility. All modes have been benchmarked against MD and experimental data, as described below:

- **Flexible mode** applies distance restraints only to residues forming secondary structure elements, allowing greater flexibility in loops and unstructured regions. Previously referred to as “SS2” in earlier versions of CABS-flex 2.0 [5] and standalone application [6], it provides a balanced representation of protein dynamics. This mode has been validated against crystallographic B-factors and atomistic MD simulations using various force fields [1], NMR ensembles [2], as well as in numerous structure-flexibility-function studies supported by experimental evidence, including Cryo-EM-derived conformational variability and functional analyses, related to aggregation propensity and S-nitrosylation sensitivity [8]. Importantly, the ability of the CABS model to predict loop structures *ab initio* has been demonstrated across vari-

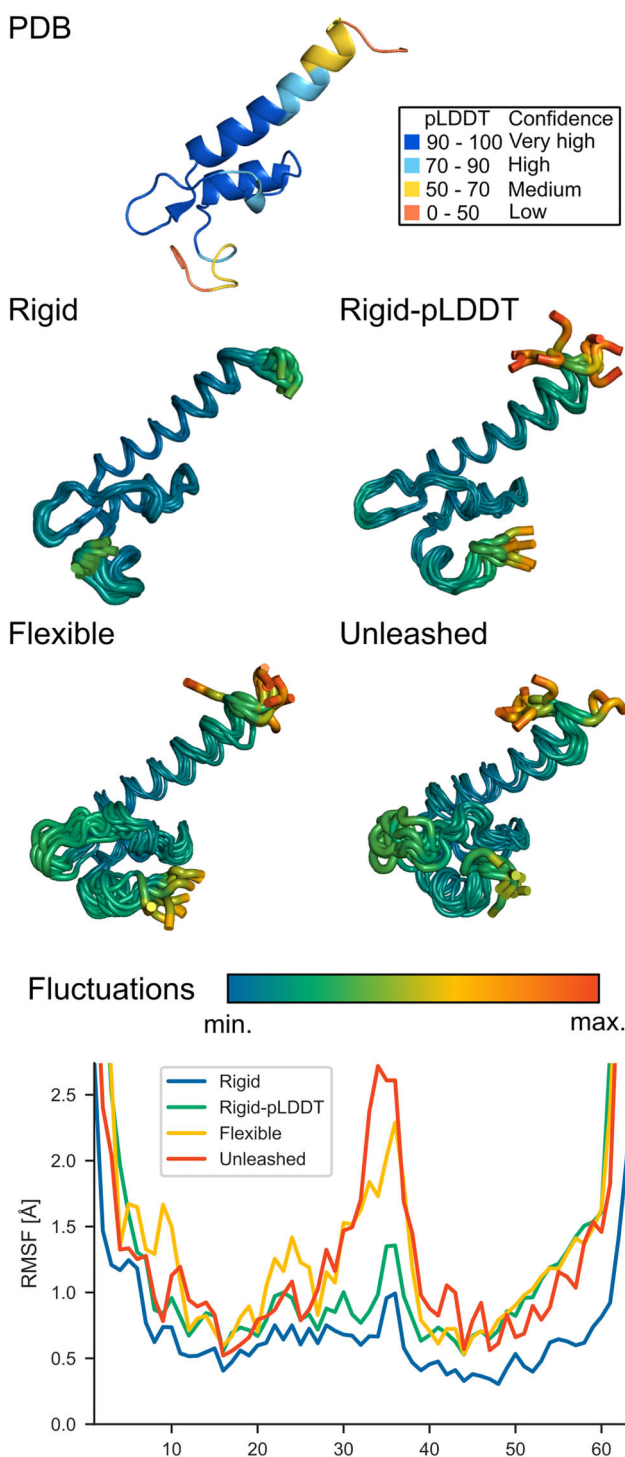


Figure 2. Example of the effect of distance restraint mode versus backbone fluctuations in the simulation. The figure presents, from the top left: the starting protein structure (PDB ID: 2f60 chain K) colored according to pLDDT, followed by the output structures obtained using restraints modes: Rigid, Rigid-pLDDT, Flexible, and Unleashed. Further details on the presented restraints were recently discussed in [3].

ous protein systems [12, 13], including in the context of peptide interactions [14], highlighting its suitability for modeling local structural variability.

- **Rigid mode** imposes uniform restraints on all residues, minimizing fluctuations throughout the structure. Equivalent to the ‘All’ mode described previously [5, 6], it effectively preserves native-like conformational constraints. Benchmarking with CHARMM36m-based ATLAS MD simulations showed high correlation with MD-derived fluctuation profiles, particularly in structured, globular proteins [3]. It is worth noting that the selection criteria of the ATLAS dataset favor well-folded, high-resolution monomeric proteins, which may bias the dataset toward more structurally stable and less flexible systems. However, some degree of intrinsic flexibility is still likely present in the selected structures.
- **Rigid-pLDDT mode** improves flexibility predictions by integrating structural confidence into the simulation process. In this mode, restraint strength is modulated according to per-residue pLDDT scores and secondary structure classification, as introduced in our recent work [3]. It is applicable when pLDDT scores are available, as in AlphaFold-predicted structures. Users may supply pLDDT values either through the B-factor field in the PDB file or via an external .json or .tsv file. Based on general observations, the overall level of flexibility generated by this mode tends to be closer to **Rigid** than **Flexible**, which is reflected in the naming convention, and the mode has been validated using ATLAS MD simulation data [3].
- **Unleashed mode** applies no restraints, allowing for fully unrestricted conformational sampling. The resulting ensemble is governed solely by the intrinsic properties of the CABS coarse-grained force field. This mode is designed as an advanced option, recommended primarily for exploratory simulations where it is important to observe the behavior of a system in the complete absence of restraints. While this mode typically shows lower agreement with MD-derived fluctuation profiles and may produce exaggerated motions, it is valuable for modeling folding/unfolding processes or large-scale transitions in disordered or flexible systems. In such cases, meaningful results may require careful tuning of additional simulation parameters, such as temperature and number of cycles. The applicability of CABS-based simulations to disordered and unfolded protein systems has been reviewed in our previous work [15], which discusses several case studies involving disordered binding partners and unstructured regions.

Mode selection guidance

By default, Flexible mode is enabled in CABS-flex 3.0 due to its versatility and proven accuracy. For stable, well-folded proteins, the Rigid mode may be most appropriate. When pLDDT scores are available, we recommend using Rigid-pLDDT, which incorporates structural confidence into simulations and improves predictive accuracy. Conversely, for systems expected to undergo large-scale motions or unfolding, the Unleashed mode can be used, with caution and additional parameter tuning. When in doubt, we recommend running simulations with multiple flexibility modes, particularly Rigid, Flexible, and Rigid-pLDDT, and comparing the resulting fluctuation profiles with available structural or functional knowl-

edge about the protein. This comparative approach can help identify the most appropriate mode for a given system.

Advanced flexibility control

CABS-flex 3.0 also offers a residue-level flexibility editor that allows users to manually assign flexibility categories prior to simulation. This option, available through a checkbox before job submission, uses the same underlying scheme as the Rigid-pLDDT mode [3] and enables expert-level customization when specific flexibility patterns are known.

Advanced input options

Advanced options allow users to modify default settings based on their needs and available information about the modeled system. Below we outline the main categories of advanced input options available in CABS-flex 3.0. Detailed explanations, usage tips, and practical recommendations are available in the ‘How To’ page (<https://lcbio.pl/cabsflex3/howto>) on the CABS-flex website. The influence of key parameters, such as restraint schemes and temperature, on simulation outcomes has also been systematically analyzed in our recent study [3].

- **Flexibility Restraint Settings**—a set of restraints is automatically generated based on the selected Flexibility Mode. However, additional parameters in this section allow users to modify the rule set for generating restraints, enabling fine-tuned control over how flexibility is applied in the simulation.
- **Custom Restraints**—users can add restraints beyond the automatically generated ones, incorporating experimental data. CABS-flex 3.0 supports restraints for both C α and side chains, which can be entered manually or uploaded as a text file.
- **Simulation Settings**—this section provides control over general simulation parameters, such as trajectory length and temperature. Temperature is one of the key factors affecting protein flexibility.

Peptide modeling

The only required input is the peptide sequence. If secondary structure information is not provided, it will be predicted using the NetSurfP-3.0 method [16]. Alternatively, users can provide the secondary structure in HEC format (H for helix, E for beta-strand, C for coil) alongside the sequence, separated by a colon (e.g. ALALA:CHHHH). Additional options include ‘Add disulfide bonds,’ which allows users to select cysteines from the sequence, and ‘Model cyclic backbone.’ Both options act at the coarse-grained simulation level and during all-atom reconstruction with Modeller, ensuring the correct closure of disulfide and/or cyclic backbone bonds in the final structure.

Similar to the flexibility mode, users can specify a ‘Project name,’ choose ‘Do not show my job on the results page,’ provide an ‘Email address,’ and set a ‘Seed for random number generation.’ As in the flexibility modeling mode, CABS-flex generates 10 final models, representing alternative predictions, which can also illustrate the flexibility of the studied peptides.

This modeling protocol was previously benchmarked and evaluated in detail in a recent study [7], which included comparisons with other state-of-the-art tools such as AlphaFold and PEP-FOLD. The benchmark included linear and cyclic peptides, with varying lengths and different types of cyclization. While CABS-flex 3.0 performs best for short, linear, and

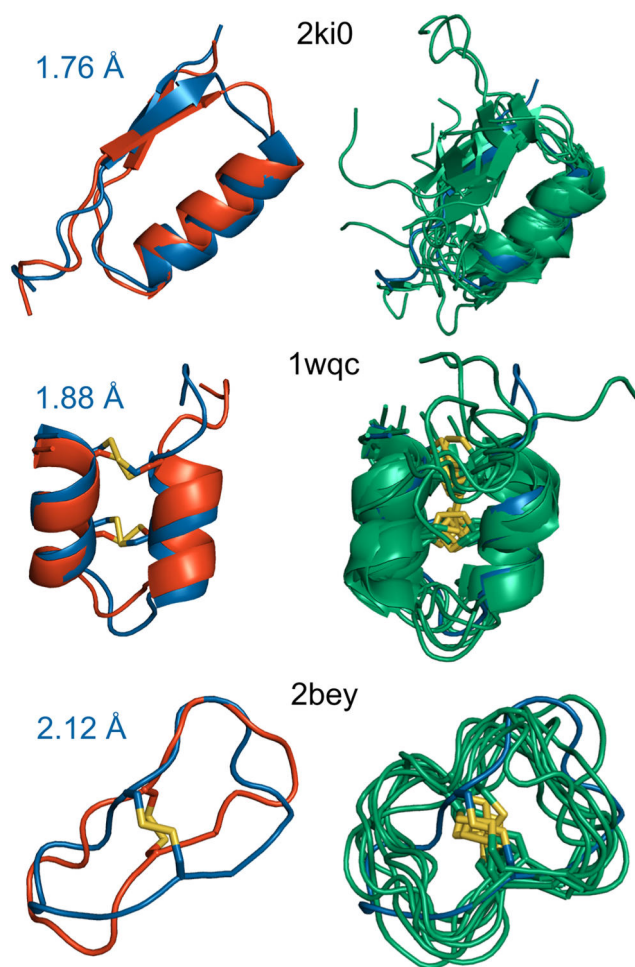


Figure 3. Example structure predictions of, from the top: linear, cyclic disulfide, and cyclic backbone peptides. On the left, best predicted structure versus experimental structure (lowest RMSD) and on the right, 10 final models superimposed on each other. The method's performance has been described in [7].

flexible peptides, it also demonstrates reliable performance for other peptide types, such as cyclic peptides and longer peptides with more complex structures. As of now, the protocol supports only canonical residues and does not include post-translational modifications. These limitations, along with possible workarounds, are discussed in detail in Section 3.7 (Merging with High-accuracy All-atom Reconstruction) of our recent review [8]. Example predictions from this protocol are shown in Fig. 3.

Web server output data

Upon job completion, users gain access to four analysis tabs: Project information, Models, Contact Maps, and Fluctuation Plot, which integrate advanced visualization features (see Fig. 4).

- **Project information tab** contains details about the job, basic input information, and the options introduced by a user.
- **Models tab** enables users to visualize 3D structural outputs using the mol* viewer [17]. Users can toggle between superimposed view of all 10 models and detailed

individual views. This tab offers extensive customization, allowing structures to be colored according to various experimental and simulation-derived metrics such as B-factor, pLDDT, and RMSF, as well as by sequence, chain, or secondary structure attributes. Additionally, users can adjust visual settings, like background, shadows, and outlines. All structural data, including trajectory and output models, are available for download in a single compressed package (.tar.gz).

- **Contact maps tab** integrates interactive Plotly generated contact matrices with a synchronized mol* viewer [17] for an in-depth exploration of residue interactions across the simulation trajectory. It generates contact maps for each combination of chains, including intra-chain interactions, with contact frequencies normalized between 0 and 1 and visually distinguished by color. Hover tooltips provide detailed residue pair information and interaction frequencies, while selecting any contact point dynamically highlights the corresponding residues in the 3D structure and zooms in for closer inspection. This dual-view approach merges sequence-based interaction patterns with structural insights, aiding in the identification of critical inter-residue contacts. Additionally, for analyses related to the physicochemical nature of these contacts, the Mapiya server can be used [18].
- **Fluctuation plot tab** plots RMSF (and pLDDT, if applicable) values per residue as interactive 2D graphs, annotated with secondary structure elements. This detailed visualization enables users to zoom, drag, and hover over data points to access specific information about residue fluctuations. The overlay of secondary structure elements allows for a direct correlation between observed flexibility and distinct structural motifs, making it easier to pinpoint areas of inherent stability as well as regions exhibiting significant conformational variability.

Server architecture and documentation

The CABS-flex 3.0 server features an intuitive Flask-based HTML interface dynamically rendered via Jinja2 templating system. User submissions are validated and stored in a MySQL database, with jobs executed immediately or queued based on resource availability. Users can track computational progress with real-time status updates (pending, queued, running, and completed). Molecular structures are rendered using the Mol* library [17] (HTML5/JavaScript), while interactive contact maps and simulation data are plotted with Plotly (<https://plotly.com/>). Protein data are fetched directly from the PDB via RESTful APIs. The server is hosted on Apache2 and manages the user queue via MySQL. At its core, CABS-flex 3.0 employs the CABS-flex standalone package [6]. The service is freely accessible without login, with detailed documentation provided on the “How to” page: <https://lcbio.pl/cabsflex3/howto>.

Summary

CABS-flex 3.0 enables efficient simulations of protein and peptide dynamics using a coarse-grained model combined with all-atom reconstruction. It introduces flexibility modes (Flexible, Rigid, Rigid-pLDDT, and Unleashed), allowing users to control the extent of conformational freedom in simulations. The method integrates pLDDT-derived restraints, improving

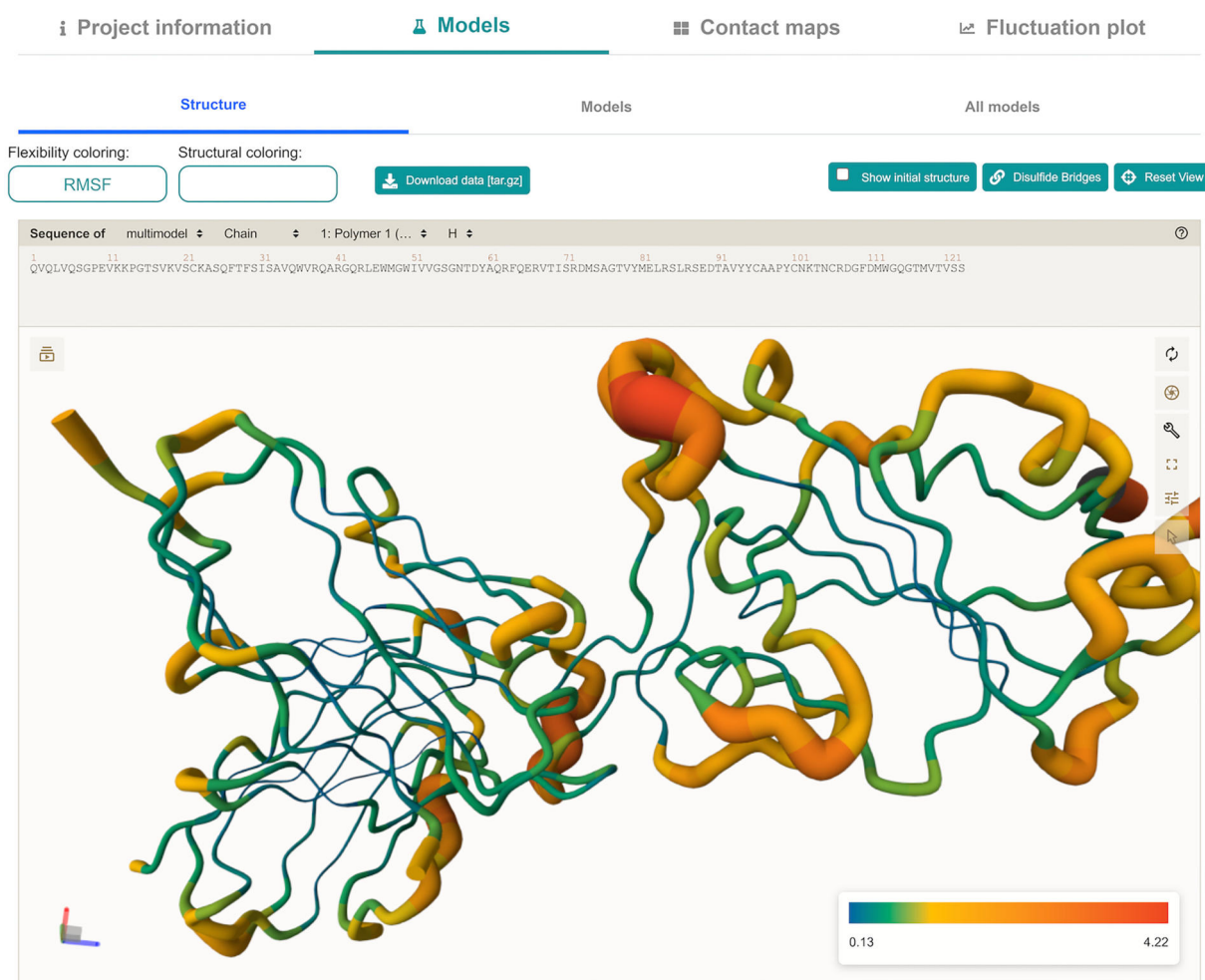


Figure 4. The CABS-flex 3.0 web interface presenting example modeling results in Flexibility Modeling mode. The interface displays the “Models” tab, showing a protein model colored according to RMSF.

accuracy by incorporating structural confidence levels. It also supports peptide modeling, enabling structure prediction for both linear and cyclic peptides. Users can input a protein structure or peptide sequence, while output includes simulation trajectories and model ensembles, which can be analyzed via the web interface. Default settings are optimized, but users can adjust flexibility levels based on experimental or computational data. The intuitive web interface, combined with optimized default settings and flexible control options, makes CABS-flex 3.0 a versatile tool for studying protein and peptide dynamics.

Acknowledgements

We would like to express our sincere gratitude to Grzegorz Firlik for his invaluable assistance in setting up the web server infrastructure and for enabling computations on the Funk supercomputer at the Biological and Chemical Research Centre, University of Warsaw. We also extend our thanks to Rafał Madaj for his support during the early stages of web service development..

Author contributions: K.W. (Conceptualization [equal], investigation [equal], methodology [equal], project administration [equal], software [equal], validation [equal], visualization [equal], writing – original draft [equal], writing – review & editing [equal]), M.Z. (Conceptualization [equal], in-

vestigation [equal], methodology [equal], software [equal], validation [equal], visualization [equal], writing – original draft [equal]), A.K. (Conceptualization [equal], investigation [equal], methodology [equal], software [equal], validation [equal]), S.K. (Conceptualization [equal], funding acquisition [equal], resources [equal], supervision [equal], validation [equal], writing – original draft [equal], writing – review & editing [equal]).

Conflict of interest

None declared.

Funding

K.W., M.Z., and S.K. acknowledge funding from the National Science Centre, Poland [2020/39/B/NZ2/01301] OPUS.

Data availability

This website is free and open to all users and there is no login requirement. The web server is available at <https://lcbio.pl/cabsflex3>. An extended CABS-flex 3.0 description and online tutorials are available at the CABS-flex 3.0 “How to” page: <https://lcbio.pl/cabsflex3/howto>.

References

1. Jamroz M, Orozco M, Kolinski A *et al.* Consistent view of protein fluctuations from all-atom molecular dynamics and coarse-grained dynamics with knowledge-based force-field. *J Chem Theory Comput* 2013;9:119–25. <https://doi.org/10.1021/ct300854w>
2. Jamroz M, Kolinski A, Kmiecik S. CABS-flex predictions of protein flexibility compared with NMR ensembles. *Bioinformatics* 2014;30:2150–4. <https://doi.org/10.1093/bioinformatics/btu184>
3. Wróblewski K, Kmiecik S. Integrating AlphaFold pLDDT Scores into CABS-flex for enhanced protein flexibility simulations. *Comput Struct Biotechnol J* 2024;23:4350–6. <https://doi.org/10.1016/j.csbj.2024.11.047>
4. Jamroz M, Kolinski A, Kmiecik S. CABS-flex: server for fast simulation of protein structure fluctuations. *Nucleic Acids Res* 2013;41:W427–31. <https://doi.org/10.1093/nar/gkt332>
5. Kuriata A, Gierut AM, Oleniecki T *et al.* CABS-flex 2.0: a web server for fast simulations of flexibility of protein structures. *Nucleic Acids Res* 2018;46:W338–43. <https://doi.org/10.1093/nar/gky356>
6. Kurcinski M, Oleniecki T, Ciemny MP *et al.* CABS-flex standalone: a simulation environment for fast modeling of protein flexibility. *Bioinformatics* 2019;35:694–5. <https://doi.org/10.1093/bioinformatics/bty685>
7. Badaczewska-Dawid A, Wróblewski K, Kurcinski M *et al.* Structure prediction of linear and cyclic peptides using CABS-flex. *Brief Bioinform* 2024;25:bbae003. <https://doi.org/10.1093/bib/bbae003>
8. Nithin C, Fornari RP, Pilla SP *et al.* Exploring protein functions from structural flexibility using CABS-flex modeling. *Protein Sci* 2024;33:e5090. <https://doi.org/10.1002/pro.5090>
9. Heo L, Feig M. One bead per residue can describe all-atom protein structures. *Structure* 2024;32:97–111.e6. <https://doi.org/10.1016/j.str.2023.10.013>
10. Vander Meersche Y, Cretin G, Gheeraert A *et al.* ATLAS: protein flexibility description from atomistic molecular dynamics simulations. *Nucleic Acids Res* 2024;52:D384–92. <https://doi.org/10.1093/nar/gkad1084>
11. Webb B, Sali A. Comparative protein structure modeling using MODELLER. *Curr Protoc Bioinform* 2016;54:5.6.1–5.6.37. <https://doi.org/10.1002/cpbi.3>
12. Jamroz M, Kolinski A. Modeling of loops in proteins: a multi-method approach. *BMC Struct Biol* 2010;10:5. <https://doi.org/10.1186/1472-6807-10-5>
13. Kmiecik S, Jamroz M, Kolinski M. Structure prediction of the second extracellular loop in G-protein-coupled receptors. *Biophys J* 2014;106:2408–16. <https://doi.org/10.1016/j.bpj.2014.04.022>
14. Blaszczyk M, Kurcinski M, Kouza M *et al.* Modeling of protein-peptide interactions using the CABS-dock web server for binding site search and flexible docking. *Methods* 2016;93:72–83. <https://doi.org/10.1016/j.ymeth.2015.07.004>
15. Ciemny M, Badaczewska-Dawid A, Pikuzinska M *et al.* Modeling of disordered protein structures using Monte Carlo simulations and knowledge-based statistical force fields. *Int J Mol Sci* 2019;20:606. <https://doi.org/10.3390/ijms20030606>
16. Hoie MH, Kiehl EN, Petersen B *et al.* NetSurfP-3.0: accurate and fast prediction of protein structural features by protein language models and deep learning. *Nucleic Acids Res* 2022;50:W510–5. <https://doi.org/10.1093/nar/gkac439>
17. Sehnal D, Bittrich S, Deshpande M *et al.* Mol* viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Res* 2021;49:W431–7. <https://doi.org/10.1093/nar/gkab314>
18. Badaczewska-Dawid AE, Nithin C, Wroblewski K *et al.* MAPIYA contact map server for identification and visualization of molecular interactions in proteins and biological complexes. *Nucleic Acids Res* 2022;50:W474–82. <https://doi.org/10.1093/nar/gkac307>

Received: March 13, 2025. Revised: April 22, 2025. Editorial Decision: April 29, 2025. Accepted: May 2, 2025

© The Author(s) 2025. Published by Oxford University Press on behalf of Nucleic Acids Research.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

5. Agregacja białek zależna od pH: Aggrescan4D

5.1. [A7] Aggrescan4D: analiza zależnej od pH agregacji białek z uwzględnieniem informacji strukturalnej

5.1.1. Wprowadzenie do [A7]

Publikacja [A7] dotyczy nowej wersji narzędzia Aggrescan3D - serwera Aggrescan4D (A4D), rozwijanego wspólnie przez nasz zespół oraz grupę Salvadora Ventury z Autonomicznego Uniwersytetu w Barcelonie. Jest to kolejna generacja dobrze znanego i szeroko stosowanego narzędzia Aggrescan3D¹²⁰, zaprojektowana do analizy skłonności białek do agregacji w sposób bardziej realistyczny biologicznie, z uwzględnieniem zarówno wpływu pH, jak i dynamiki konformacyjnej struktur. A4D stanowi rozwinięcie wcześniejszych podejść^{78,120} i rozszerza ich możliwości, czyniąc je bardziej funkcjonalnymi w badaniach nad układami białkowymi. Mój wkład w realizację tego projektu obejmował przede wszystkim zaprojektowanie i implementację nowego interfejsu webowego, umożliwiającego intuicyjną i wygodną pracę z rozszerzonymi funkcjonalnościami serwera, a także testowanie jego działania oraz weryfikację poprawności wdrożonych funkcji.

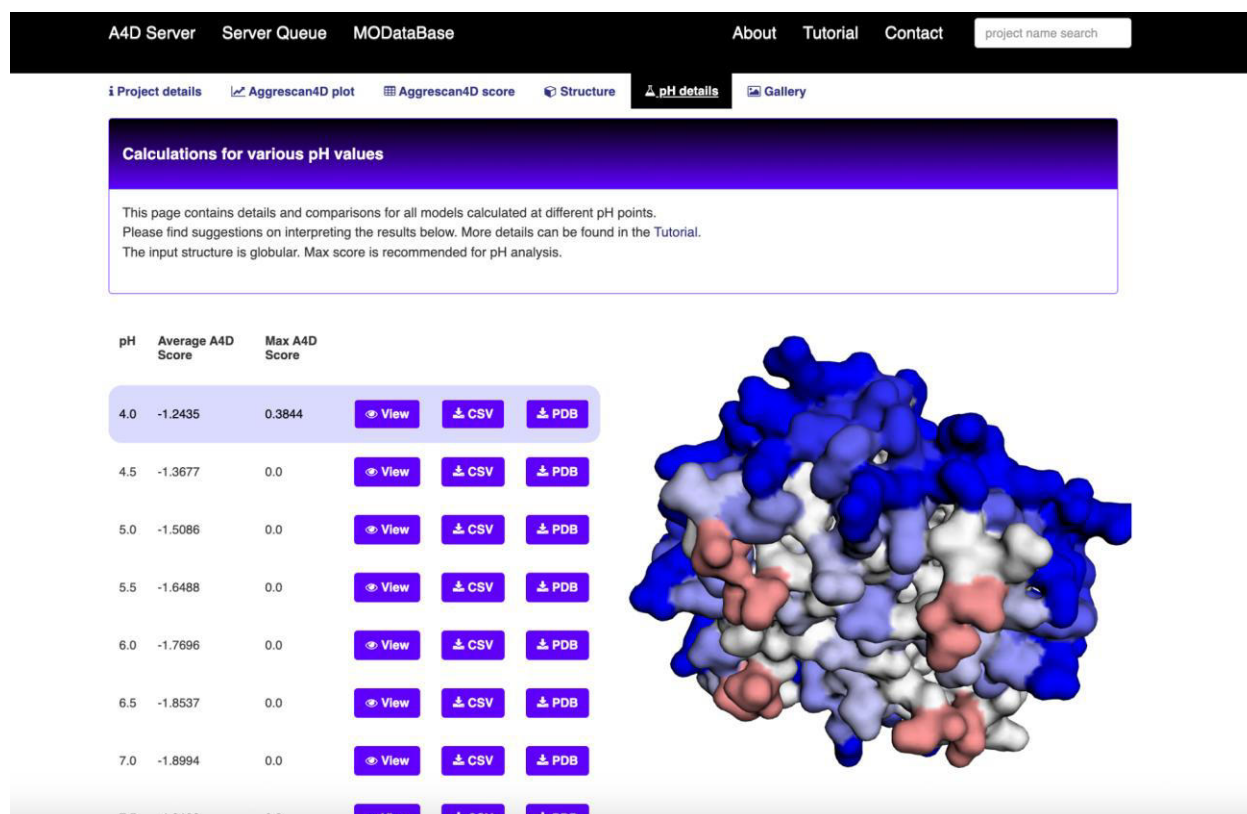
Istotnym elementem A4D jest metoda CABS-flex, szczegółowo omówiona w publikacjach [A5] i [A6] (teoretyczne podstawy metody oraz modelu gruboziarnistego CABS przedstawiłem odpowiednio w rozdziałach 2.3.2 i 2.2.2). A4D zawiera zoptymalizowany schemat trybu dynamicznego, w którym dane wejściowe mogą pochodzić bezpośrednio z predykcji AlphaFold lub innych źródeł strukturalnych. Po wygenerowaniu zespołu konformacji z użyciem metody CABS-flex, dla każdej struktury wyznaczany jest lokalny rozkład wartości agregacji, a następnie - poprzez uśrednienie lub analizę rozrzutu - tworzone są mapy agregacyjności uwzględniające fluktuacje strukturalne. Takie podejście pozwala na identyfikację tzw. konformacyjnych przełączeń agregacyjnych, czyli sytuacji, w których dana reszta zyskuje lub traci potencjał agregacyjny w wyniku zmiany swojej orientacji przestrzennej lub otoczenia.

A4D zawiera również zintegrowany moduł mutagenezy, umożliwiający ocenę wpływu podstawień punktowych na profil agregacyjny struktury, co znajduje zastosowanie m.in. w projektowaniu bardziej stabilnych wariantów białek lub optymalizacji sekwencji białek terapeutycznych.

Głównym celem prac nad A4D było połączenie przewidywań agregacji wynikających z właściwości sekwencji i struktury przestrzennej z dodatkowym kontekstem biochemicznym - pH otoczenia. Agregacja białek ma charakter wieloczynnikowy i zależy nie tylko od wewnętrznych determinant sekwencji, lecz także od środowiska, w którym białko funkcjonuje.

To środowisko (temperatura, siła jonowa, pH) w istotny sposób wpływa na solubilność i tendencję do agregacji.^{121–124}

Dotychczasowe narzędzia predykcyjne pomijały te czynniki, opierając się wyłącznie na informacji sekwencyjnej, czasem strukturalnej, ale wciąż statycznej. W A4D zastosowaliśmy podejście bardziej realistyczne biologicznie, integrując: konformacyjną zmienność białek, predykcję pKa reszt aminokwasowych na bazie uczenia głębokiego oraz nowe metryki uwzględniające wpływ pH na tendencję do agregacji. Przykład modułu służącego do analizy wpływu pH na serwerze A4D przedstawiono na Rysunku 9.



Rysunek 9. Interfejs serwera Aggrescan4D (A4D) pokazujący moduł analizy wpływu pH na skłonność białka do agregacji. Tabela po lewej stronie prezentuje wartości średnie (Average A4D Score) i maksymalne (Max A4D Score) dla różnych wartości pH, wraz z możliwością pobrania danych i ich podglądu. Po prawej stronie znajduje się interaktywna wizualizacja trójwymiarowej powierzchni białka, w której kolory odpowiadają regionom o zróżnicowanej skłonności do agregacji. Źródło: [A7].

Wersja A4D wprowadza dwie metryki agregacyjności: A4D-max (maksymalna wartość punktowa w strukturze) i A4D-avg (średnia wartość dla wszystkich reszt). W zależności od tego, czy analizujemy białko globularne, czy IDP (ang. *intrinsically disordered protein*), preferowana metryka będzie inna. Dla białek o zwartej strukturze, lepiej sprawdza się A4D-max, natomiast w przypadku IDP - A4D-avg.

Do oceny tych miar przeprowadziliśmy analizę korelacji między wartościami A4D a 18 eksperymentalnymi zestawami danych dotyczącymi wpływu pH na agregację lub rozpuszczalność 14 różnych białek. Wyniki pokazały wyraźne różnice w skuteczności zależne od typu białka, co doprowadziło nas do opracowania strategii konsensusowej (A4D-consensus), uwzględniającej zwartość struktury przy wyborze odpowiedniej miary.

W A4D zaimplementowaliśmy także nowy protokół modyfikacji konserwatywnych, oparty na macierzy PAM250. W odróżnieniu od podejścia zastosowanego wcześniej w A3D 2.0, które sugerowało mutacje do silnie naładowanych aminokwasów (co mogło negatywnie wpływać na stabilność), nowy schemat opiera się na zmianach akceptowanych ewolucyjnie i uwzględnia nie tylko skuteczność w redukcji agregacji ($\Delta A4D < 0$), ale również potencjalny wpływ na strukturę i funkcję białka. Przykładowo, dla bGH (*bovine growth hormone*) wskazane przez protokół mutacje odpowiadały naturalnym różnicom pomiędzy bardziej rozpuszczalną hGH (*human growth hormone*) i mniej rozpuszczalną bGH.

W serwerze A4D udostępniliśmy szereg nowych funkcjonalności zwiększających jego dostępność i przyjazność dla użytkownika:

- integrację z bazą AlphaFold¹²⁵ oraz A3D-MODB¹²⁶, co pozwoliło na pobieranie gotowych struktur z bazy lub wykorzystanie uprzednio obliczonych wyników dla organizmów modelowych,
- automatyczne przycinanie struktur AlphaFold za pomocą narzędzia AlphaCutter¹²⁷, aby usunąć sztucznie wygenerowane regiony nieuporządkowane,
- nowy interfejs graficzny,
- przykładowe wyniki przewidywań,
- rozszerzoną dokumentację.

Zastosowanie struktur AlphaFold mogłoby wprowadzać błędy wynikające z obecności nieuporządkowanych regionów o przypadkowym ułożeniu przestrzennym, dlatego pre-processing (AlphaCutter) stał się kluczowym elementem metody A4D w celu eliminacji artefaktów predykcyjnych.

A4D jest pierwszym strukturalnym predyktorem agregacji białek uwzględniającym wpływ pH. Dotychczas podobne podejścia obejmowały jedynie białka nieustrukturyzowane (np. SolupHred¹²⁸) lub opierały się na analizie sekwencyjnej (np. CamSol 3.0^{129(p3)}), bez pełnej integracji informacji przestrzennej. Zbiory testowe A4D wykazały wysoką zgodność z danymi eksperymentalnymi - szczególnie w podejściu konsensusowym (100% zgodności).

Ograniczeniem pozostaje ograniczona liczba dostępnych danych eksperymentalnych oraz ich różnorodność. Niemniej jednak, A4D stanowi nową jakość w analizie i projektowaniu białek

pod kątem ich rozpuszczalności, stabilności i podatności na agregację w określonych warunkach środowiskowych.

Połączenie informacji ewolucyjnych, strukturalnych i fizykochemicznych w jednym, łatwo dostępnym narzędziu, które może wspierać zarówno badania podstawowe, jak i zastosowania biotechnologiczne było szczególnie ważnym zadaniem. Dzięki wszechstronności i otwartości (brak logowania, REST API, dostępność danych), A4D może być z powodzeniem stosowane w zadaniach optymalizacji konstrukcji białkowych, oceny stabilności form terapeutycznych oraz nauczania zagadnień związanych z agregacją białek.

Uważam, że narzędzie Aggrescan4D przedstawione w pracy [A7] stanowi ważny krok naprzód w analizie zjawiska agregacji białek. Integracja informacji strukturalnych, dynamicznych i środowiskowych w jednym protokole otwiera nowe możliwości - pozwala nie tylko dokładniej przewidywać skłonność do agregacji, ale także lepiej zrozumieć mechanizmy leżące u jej podstaw. Moim zdaniem A4D ma potencjał, by stać się narzędziem pierwszego wyboru zarówno w badaniach podstawowych, jak i aplikacyjnych - od projektowania stabilniejszych białek terapeutycznych po analizy biofizyczne dużych zbiorów danych strukturalnych. Warto podkreślić, że dzięki swojej dostępności i łatwości użycia serwer już teraz cieszy się dużym zainteresowaniem i jest aktywnie wykorzystywany w środowisku naukowym, a połączenie doświadczenia dwóch zespołów (naszego i grupy Salvadora Ventury) sprawiło, że powstało narzędzie łączące rzetelną metodologię z praktycznym podejściem do analizy problemu, który w ostatnich latach nabiera coraz większego znaczenia.

5.1.2. Artykuł [A7]

Aggrescan4D: structure-informed analysis of pH-dependent protein aggregation

Oriol Bárcenas¹, Aleksander Kuriata², Mateusz Zalewski², Valentín Iglesias^{1,3}, Carlos Pintado-Grima¹, Grzegorz Firlik², Michał Burdukiewicz^{1,3}, Sebastian Kmiecik^{2,*} and Salvador Ventura^{1,*}

¹Institut de Biotecnologia i de Biomedicina and Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain

²Biological and Chemical Research Center, Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland

³Clinical Research Centre, Medical University of Białystok, Kilińskiego 1, 15-369 Białystok, Poland

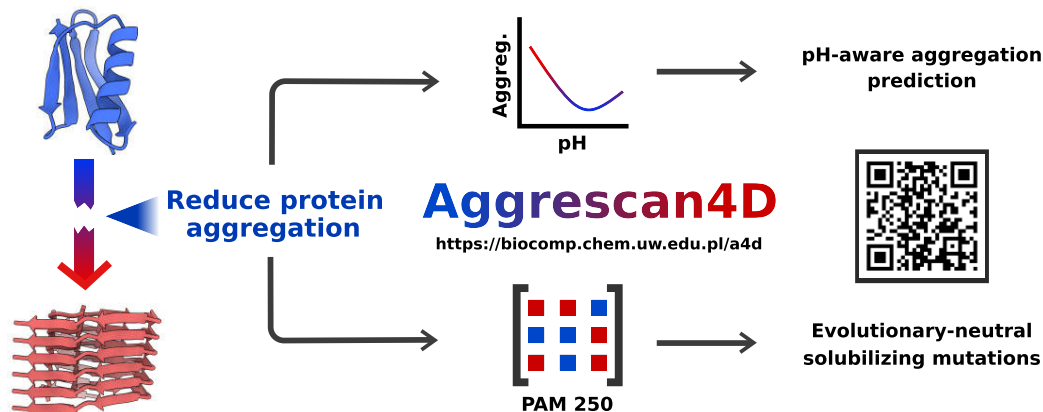
*To whom correspondence should be addressed. Tel: +34 93 586 8956; Fax: +34 93 581 2011; Email: salvador.ventura@uab.es

Correspondence may also be addressed to Sebastian Kmiecik. Tel: +48 22 8220211 (Ext. 310); Fax: +48 22 8220211 (Ext. 320); Email: sekmi@chem.uw.edu.pl

Abstract

Protein aggregation is behind the genesis of incurable diseases and imposes constraints on drug discovery and the industrial production and formulation of proteins. Over the years, we have been advancing the Aggrescan3D (A3D) method, aiming to deepen our comprehension of protein aggregation and assist the engineering of protein solubility. Since its inception, A3D has become one of the most popular structure-based aggregation predictors because of its performance, modular functionalities, RESTful service for extensive screenings, and intuitive user interface. Building on this foundation, we introduce Aggrescan4D (A4D), significantly extending A3D's functionality. A4D is aimed at predicting the pH-dependent aggregation of protein structures, and features an evolutionary-informed automatic mutation protocol to engineer protein solubility without compromising structure and stability. It also integrates precalculated results for the nearly 500,000 jobs in the A3D Model Organisms Database and structure retrieval from the AlphaFold database. Globally, A4D constitutes a comprehensive tool for understanding, predicting, and designing solutions for specific protein aggregation challenges. The A4D web server and extensive documentation are available at <https://biocomp.chem.uw.edu.pl/a4d/>. This website is free and open to all users without a login requirement.

Graphical abstract



Introduction

Protein aggregation is a multifactorial process, directed by the intrinsic properties of proteins (1) and heavily influenced by environmental factors (2). This phenomenon accounts for the onset of highly debilitating proteinopathies, posing a significant challenge to human health. Additionally, it hinders the development and implementation of protein-based biotechnological and biomedical applications (3).

Classical computational methods that predict protein aggregation rely on sequential information to assess the propensity of a protein to aggregate. These algorithms belong to four main categories: (i) heuristic detection of molecular determinants of aggregation (4–6), (ii) assessment of the conformational compatibility with amyloid-like architectures (7–9), (iii) consensus methods employing weighted predictions from other algorithms (10,11) and (iv) machine-learning

Received: March 11, 2024. Revised: April 11, 2024. Editorial Decision: April 27, 2024. Accepted: April 29, 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of Nucleic Acids Research.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License

(<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

methods (12–15). While some of these algorithms provide a binary decision on whether a given sequence will aggregate, in most cases, they identify and weight the contribution of aggregation-prone regions (APRs), as is the case for Aggrescan (6). While highlighting APRs provides information on the regions driving aggregation, this strategy assumes that the proximity between residues is directly proportional to their sequential distance, which might apply for disordered or denatured protein states but is an imprecise approximation for the native state of globular proteins.

Being structure agnostic, these methods treat the aggregation contribution of amino acids residing in collapsed or inaccessible regions (such as the hydrophobic core) equally to exposed residues. To overcome these limitations, we developed a three-dimensional structure aggregation predictor, Aggrescan3D (A3D), which projects classical Aggrescan scoring to proteins' native structure, introducing the concept of STructural APRs (STAPs (16)), which correspond to the main drivers of protein aggregation in natively folded structures (17). To further enhance the biological relevance of its predictions, A3D incorporates CABS-flex (18,19). CABS-flex is a cost-efficient conformational sampling method designed to capture the inherent flexibility of proteins in solution. It accounts for the influence of multiple conformations on aggregation, moving beyond the limitations of a single, static structure description.

Building on these improvements, here we present Aggrescan4D (A4D), which adds a new dimension to protein aggregation prediction: environmental pH. This new variable contextualizes aggregation as a phenomenon modulated by environmental conditions. This paves the way to studying the relationship between pH and aggregation in biological contexts (20,21), while providing a tool to modulate protein aggregation in biotechnological setups. We have also added a new, evolutionary-based automatic mutation protocol to reduce protein aggregation while minimally impacting protein thermodynamic stability. Additionally, other user-centric improvements have been incorporated to streamline A4D use. Collectively, this update brings experimental results and computational predictions closer by providing a better description of the protein's biochemical context.

New features and updates

pH-dependent aggregation prediction

Proteins are, in general terms, evolutionarily selected to be soluble in their natural contexts (22), but the protein microenvironment (temperature, ions, or pH) can profoundly impact this capacity. This is highly relevant for engineered proteins (such as antibodies or enzymes), which have not undergone natural selection to ensure solubility under their intended conditions of use. Despite their importance, aggregation predictors have historically disregarded environmental conditions and focused solely on the proteins' physicochemical characteristics to model protein aggregation. In 2020, we inferred that environmental changes in pH impacted amino acids' charge and lipophilicity, and by applying a pH-dependent scale for proteinogenic amino acids (23), we fitted an empirical equation capable of capturing pH-dependent aggregation in intrinsically disordered proteins (IDPs) (24). This led us to develop SolupHred, the first disordered protein aggregation predic-

tor that systematically incorporated the pH variable into its logic (25). Later in 2023, CamSol 3.0 adopted the same rationale to introduce pH-dependent solubility predictions for IDPs and globular proteins (26). Thus, we sought to adapt SolupHred's methodology to structured proteins by calculating structural pKa values with pKa-ANI, an algorithm based on deep representation learning (27), to predict the influence of pH on the aggregation of proteins in their folded configurations.

To model the influence of pH on aggregation propensity, we used two metrics: the maximum aggregation propensity ($A4D_{\max}$) and the average aggregation propensity ($A4D_{\text{avg}}$). These two measures delineate two different aggregation regimes: $A4D_{\max}$ represents the highest propensity observed across all residues and is better suited for proteins displaying defined STAPs. $A4D_{\text{avg}}$ represents the mean propensity across all residues and is more appropriate for polypeptides with an aggregation tendency that is more evenly distributed throughout the protein structure. Therefore, these two measures address the innate difference in which aggregation propensity is encoded in globular proteins and IDPs. $A4D_{\max}$ is the most appropriate metric to describe pH-dependent aggregation for globular proteins, while $A4D_{\text{avg}}$ is better suited for IDPs. To aid in selecting the appropriate metric, we provide an objective criterion capturing the compactness of the protein structure for discriminating between ordered and disordered protein structures (Supplementary Information SI 1).

We assessed the performance of $A4D_{\max}$ and $A4D_{\text{avg}}$ by analyzing data collected from 18 different experiments studying the impact of pH changes on the aggregation or solubility of 14 different proteins. Eleven of these experiments were conducted with globular proteins and 7 on IDPs (summarized in SI 2 and Supplementary Table S1; STs S2-S19 contain collected experimental data). These experiments used different metrics to measure the aggregation/solubility of the protein in the experimental setup, hereby referred to as experimental reporter. This umbrella term allows us to gather together aggregation and solubility measurements, such as Thioflavin T fluorescence (28) or solubility in PEG (29), respectively. We used linear regression to evaluate the relationship between the experimental reporters of aggregation/solubility and both A4D metrics.

The final benchmark covered $A4D_{\max}$ and $A4D_{\text{avg}}$ performance. Although protein solubility cannot always be assumed to be the inverse of aggregation propensity, for the sake of completeness, we decided to include as well CamSol 3.0, a web server for the sequence-based prediction of pH-dependent protein solubility (26). A4D scores increase with a higher aggregation tendency. Conversely, as CamSol 3.0 predicts solubility, its scores should be lower for higher aggregation propensities.

To assess the algorithm's ability to discriminate the influence of pH on protein solubility/aggregation, we analyzed the concordance of the linear regression slope between the experimental reporters and the benchmarked metrics (summarized in SI3 and STs S20-S23).

The overall concordance of $A4D_{\max}$, $A4D_{\text{avg}}$ and CamSol 3.0 were 78%, 89% and 78%, respectively. However, the degree of concordance varied depending on the proteins' conformation. Notably, for IDPs, both $A4D_{\text{avg}}$ and CamSol 3.0 demonstrated perfect concordance (100%); while, as expected, $A4D_{\max}$ performed significantly worse, reaching only

43%. The situation is the opposite in the case of globular proteins. Here, A4D_{max} reached a concordance of 100%, whereas A4D_{avg} and CamSol 3.0 showed worse performance, 82% and 64%, respectively.

Thus, we have computed a novel structure-informed consensus measure, A4D_{consensus}, which considers the protein structure's compactness and based on that, selects the most appropriate metric for each query (A4D_{max} or A4D_{avg}), reaching a 100% overall concordance in the evaluated experiments (Figure 1A). Even though A4D_{max} is recommended for globular proteins, it is worth noting that A4D_{avg} remains a valid metric across all scenarios, reporting an overall concordance of 89%. Details of the benchmark are provided in SI 3 (STs S24-S41 and Supplementary Figures S1- S18). A case study is presented in SI 4 and SF S19.

To cover all tools aimed at predicting pH-dependent protein aggregation, we integrated a benchmark of SolupHred with the disordered proteins used in the general benchmark. SolupHred, like A4D_{avg} and CamSol 3.0, demonstrated 100% concordance on this dataset, something expected as it was specifically designed for IDPs.

Evolutionary-aware protein solubility improvement

The automated mutation mode to enhance solubility was introduced in A3D 2.0 (30). This method optimizes the input structure using FoldX (31) and ranks the most aggregation-prone residues according to their A3D score. Subsequently, the less soluble residues are mutated to charged amino acids (lysine, arginine, aspartic acid, or glutamic acid). While these mutations might turn optimal for reducing protein aggregation, they often induce significant changes in side chain physicochemical properties, which can negatively impact folding and stability, potentially disrupting protein function.

A4D provides an alternative to circumvent this limitation by applying a solubilizing strategy informed on the evolutionary conservation of point substitutions. Point accepted mutation (PAM) matrices store the sum of accepted substitutions in closely related sequences (identity ≥ 0.85), ensuring changes are a consequence of a single point mutation ($A \rightarrow B$) and not an accumulation of multiple mutations ($A \rightarrow C \rightarrow B$) (32,33). This translates to evolutionary favored changes that should have minimal impact on protein structure and stability. Despite this generalization, the server allows restraining the algorithm from mutating residues known by users as critical for maintaining protein function.

This protocol integrates evolutionary and aggregation information to suggest conservative mutations that reduce the aggregation propensity of the protein. First, conserved and neutral mutations are chosen based on the PAM250 substitution matrix (log-transformed PAM value ≥ 0). Out of these, only solubilizing mutations ($\Delta A4D$ score < 0) are considered. Finally, up to three mutations are suggested for each residue, prioritizing solubilization. We conducted a case study on bovine growth hormone (bGH) predicted structure (AF-A4GX95, residues 27–217). Despite sharing 90% sequential identity, whereas the human protein (hGH) is very soluble, bGH exhibits extreme aggregation across different experimental conditions (34). Notably, the analysis revealed different instances where the automated protocol's recommended mutations aligned with the residue identities found in hGH (V76T, Figure 1B). A comparison of this protocol and the protocol added in A3D 2.0 is provided in SI 5 and SF S20.

Other improvements

Beyond core functionalities, this update also contains numerous quality-of-life improvements. First, we expanded file compatibility to accept CIF/mmCIF file formats. Moreover, we have added seamless cross-database linking in the form of automatic AlphaFold structure fetching from UniProt ID (35) and A3D-Model Organisms DataBase (A3D-MODB) (36) entry retrieval. A3D-MODB includes precomputed predictions for over 164k proteins across 12 model organisms. Avoiding unnecessary calculations speeds up the tool and reduces its environmental impact, especially for recurrent jobs or training scenarios (37).

An essential consideration when using protein structures as inputs is the presence of long, disordered regions in proteins from the AlphaFold database, which are absent in most experimentally resolved protein structures. These disordered regions are occupying (pseudo)-random positions, generally close to the globular core (if it exists) (38). The static nature of these models leads to the generation of artificial surface regions stemming from the arbitrary and fixed position of disordered regions, which might lead to artifactual results, such as protecting or generating STAPs (36). To mitigate this, AlphaCutter is implemented as a pre-processing option, generating protein structures containing only globular regions (39).

In addition to these technical improvements, we have undertaken a thorough redesign of our user interface to improve the tool's usability and user experience. Moreover, A4D now includes a selection of example jobs showcasing the offered functionalities, complementing the detailed tutorial and FAQ sections already present in the server.

Limitations

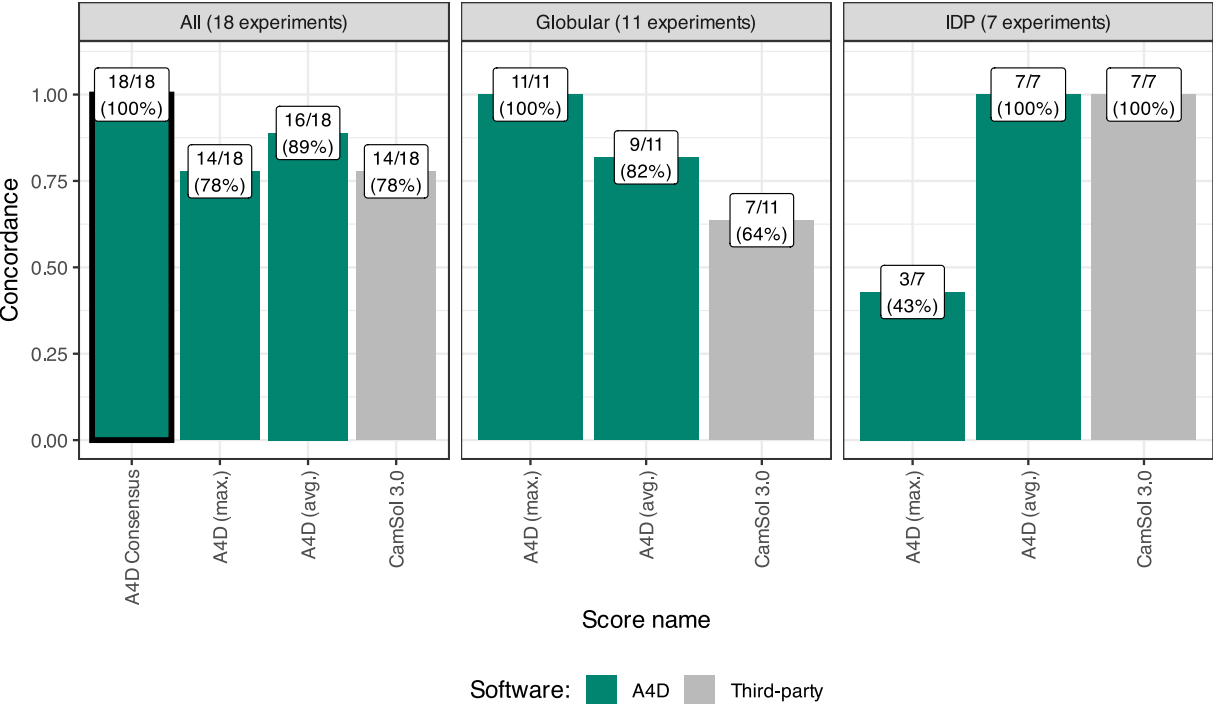
We acknowledge that limitations associated with A4D pH-dependent aggregation predictions may impact the overall performance of the tool and subsequent benchmark.

First and foremost, the scarcity of well-annotated available data has been one of the most limiting factors for this work. Although we aimed to amend this limitation by enriching existing datasets with our unpublished data, we still cover only 14 different proteins, representing a considerably small fraction of the protein landscape. A second limiting factor was the high diversity of experimental reporters of protein aggregation provided by the authors. Given this problem, we interchangeably use measures of aggregation and solubility. Third, the heterogeneity of data sources has also contributed to the noise in the performance assessment (40).

Discussion and conclusions

Before the advent of bioinformatic tools, understanding and modifying the aggregation tendency of proteins was a very challenging and expensive task. For instance, to produce soluble, protein-based pharmaceuticals, techniques such as phage display (41) were often used. Services like A4D help users understand the causes behind aggregation in their proteins and assist in devising strategies to prevent it. While this behavior is primarily encoded in a protein's sequence and structure, other factors such as protein or crowding agent concentration, temperature, or pH can heavily influence the aggregation of polypeptides. Recognizing the importance of pH in

A



B

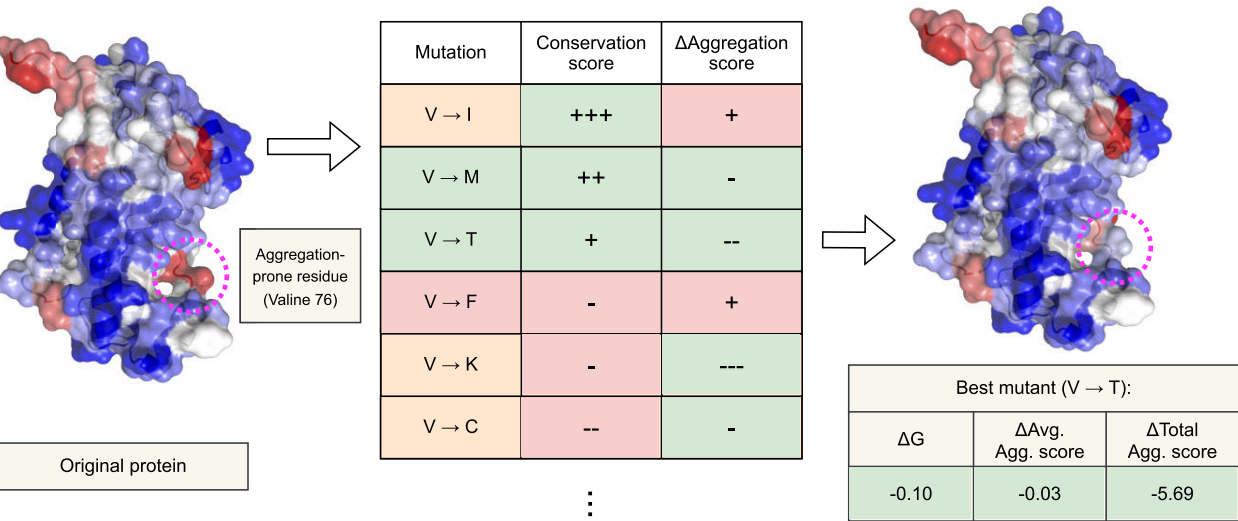


Figure 1. (A) Benchmark results for A4D (green) and CamSol 3.0 (gray) pH predictions. Results are presented jointly for all proteins and separately for globular proteins and IDPs. **(B)** Schematic representation of the evolutionary-aware solubility improvement protocol applied to bovine growth hormone. Mutations are suggested for the protein's most aggregation-prone residues, considering their relative conservation and aggregation score. If the new residue is predicted to be conserved or neutral and improve the solubility of the protein, it is considered for mutation (green). On the contrary, if the residue is not conserved or does not reduce the aggregation potential (orange) or both (red), it is not considered.

modulating protein aggregation, we updated our existing algorithm, A3D, to model this effect. This represents a categorical shift, moving beyond protein sequence (Aggrescan) and structure (Aggrescan3D) analyses, to provide a more comprehensive vision of protein aggregation. To our knowledge, SolupHred (2020) (25), and CamSol 3.0 (2023) (26) are the only existing resources that systematically model the effect of pH on solubility/aggregation of protein structures. This lim-

ited landscape reflects the difficulty of modeling the contribution of solvent conditions to protein aggregation, and the scarcity of available experimental data. We also introduce a novel protocol that suggests mild, evolutionary-neutral mutations to solubilize and, eventually, stabilize proteins without requiring a deep understanding of the protein at hand. A4D provides a platform for researchers to optimize their experimental conditions and generate more consistent data,

which can feed back the present algorithm and trigger the development of novel computational tools and standardized benchmarks.

A3D has been successfully applied in the redesign of less aggregation-prone variants of biotherapeutics, engineering novel self-assembled nanomaterials, understanding of pathological and physiological protein aggregation, and as a tool for teaching protein aggregation in university courses (42). While mutation protocols are the most straightforward and potent to prevent protein aggregation, they may face limitations such as the impossibility of mutating aggregation-prone but functionally important residues, like CDRs in antibodies or the existence of intellectual protection issues, like in the case of biosimilars. Thus, modifying the aggregation behavior of proteins without altering the protein sequence itself is key in many biotechnological and medical applications. Given the new functionalities implemented in this update, we anticipate that A4D will become a widely used tool within the protein science community.

Data availability

All data on the predicted aggregation propensity supporting the findings of this study are included within this paper and its [Supplementary Information](#) files.

Supplementary data

[Supplementary Data](#) are available at NAR Online.

Funding

O.B. was supported by the Spanish Ministry of Science and Innovation via a doctoral grant [FPU22/03656]; V.I. was supported by Spanish Ministry of Universities; European Union-Next Generation EU [ruling 02/07/2021, Universitat Autònoma de Barcelona]; C.P.-G. was supported by the Secretariat of Universities and Research of the Catalan Government and the European Social Fund [2023 FI_3 00018]; M.B. was supported by the Maria Zambrano grant funded by the European Union-NextGenerationEU; S.V. was supported by Spanish Ministry of Science and Innovation [PID2022-137963OB-I00]; ICREA, ICREA-Academia 2020; A.K., M.Z. and S.K. acknowledge funding from the National Science Centre, Poland [Sheng 2021/40/Q/NZ2/00078]. Funding for open access charge: National Science Centre, Poland [Sheng 2021/40/Q/NZ2/00078 and PID2022-137963OB-I00].

Conflict of interest statement

None declared.

References

- López de la Paz, M. and Serrano, L. (2004) Sequence determinants of amyloid fibril formation. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 87–92.
- Owen, M.C., Gnutt, D., Gao, M., Wärmländer, S.K.T.S., Jarvet, J., Gräslund, A., Winter, R., Ebbinghaus, S. and Strodel, B. (2019) Effects of in vivo conditions on amyloid aggregation. *Chem. Soc. Rev.*, **48**, 3946–3996.
- Chiti, F. and Dobson, C.M. (2017) Protein misfolding, amyloid formation, and human disease: a summary of progress over the last decade. *Annu. Rev. Biochem.*, **86**, 27–68.
- Fernandez-Escamilla, A.-M., Rousseau, F., Schymkowitz, J. and Serrano, L. (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.*, **22**, 1302–1306.
- Tartaglia, G.G. and Vendruscolo, M. (2008) The Zyggregator method for predicting protein aggregation propensities. *Chem. Soc. Rev.*, **37**, 1395–1401.
- Conchillo-Solé, O., de Groot, N.S., Avilés, F.X., Vendrell, J., Daura, X. and Ventura, S. (2007) AGGRESCAN: a server for the prediction and evaluation of ‘hot spots’ of aggregation in polypeptides. *BMC Bioinformatics*, **8**, 65.
- Garbuzynskiy, S.O., Lobanov, M.Y. and Galzitskaya, O.V. (2010) FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics (England)*, **26**, 326–332.
- Ahmed, A.B., Znassi, N., Château, M.-T. and Kajava, A.V. (2015) A structure-based approach to predict predisposition to amyloidosis. *Alzheimers Dementia*, **11**, 681–690.
- Louros, N., Orlando, G., De Vleeschouwer, M., Rousseau, F. and Schymkowitz, J. (2020) Structure-based machine-guided mapping of amyloid sequence space reveals uncharted sequence clusters with higher solubilities. *Nat. Commun.*, **11**, 3314.
- Tsolis, A.C., Papandreou, N.C., Iconomidou, V.A. and Hamodrakas, S.J. (2013) A consensus method for the prediction of ‘aggregation-prone’ peptides in globular proteins. *PLoS One*, **8**, e54175.
- Emily, M., Talvas, A. and Delamarche, C. (2013) MetAmyl: a METa-predictor for AMYloid proteins. *PLoS One*, **8**, e79722.
- Gasior, P. and Kotulska, M. (2014) FISH Amyloid - a new method for finding amyloidogenic segments in proteins based on site specific co-occurrence of aminoacids. *BMC Bioinformatics*, **15**, 54.
- Familia, C., Dennison, S.R., Quintas, A. and Phoenix, D.A. (2015) Prediction of peptide and protein propensity for amyloid formation. *PLoS One*, **10**, e0134679.
- Burdukiewicz, M., Sobczyk, P., Rödiger, S., Duda-Madej, A., Mackiewicz, P. and Kotulska, M. (2017) Amyloidogenic motifs revealed by n-gram analysis. *Sci. Rep.*, **7**, 12961.
- Charoenkwan, P., Ahmed, S., Nantasenamat, C., Quinn, J. M.W., Moni, M.A., Lio, P. and Shoombutong, W. (2022) AMYPred-FRL is a novel approach for accurate prediction of amyloid proteins by using feature representation learning. *Sci. Rep.*, **12**, 7697.
- Santos, J., Pujols, J., Pallarès, I., Iglesias, V. and Ventura, S. (2020) Computational prediction of protein aggregation: advances in proteomics, conformation-specific algorithms and biotechnological applications. *Comput. Struct. Biotechnol. J.*, **18**, 1403–1413.
- Zambrano, R., Jamroz, M., Szczasiuk, A., Pujols, J., Kmiecik, S. and Ventura, S. (2015) AGGRESCAN3D (A3D): server for prediction of aggregation properties of protein structures. *Nucleic Acids Res.*, **43**, W306–W313.
- Jamroz, M., Orozco, M., Kolinski, A. and Kmiecik, S. (2013) Consistent view of protein fluctuations from all-atom molecular dynamics and coarse-grained dynamics with knowledge-based force-field. *J. Chem. Theory Comput.*, **9**, 119–125.
- Kuriata, A., Gierut, A.M., Oleniecki, T., Ciemny, M.P., Kolinski, A., Kurcinski, M. and Kmiecik, S. (2018) CABS-flex 2.0: a web server for fast simulations of flexibility of protein structures. *Nucleic Acids Res.*, **46**, W338–W343.
- Horváth, D., Dürvanger, Z., K. Menyhárd, D., Sulyok-Eiler, M., Bencs, F., Gyulai, G., Horváth, P., Taricska, N. and Perczel, A. (2023) Polymorphic amyloid nanostructures of hormone peptides involved in glucose homeostasis display reversible amyloid formation. *Nat. Commun.*, **14**, 4621.
- Raposo, G., Tenza, D., Murphy, D.M., Berson, J.F. and Marks, M.S. (2001) Distinct protein sorting and localization to premelanosomes, melanosomes, and lysosomes in pigmented melanocytic cells. *J. Cell Biol.*, **152**, 809–824.
- Monsellier, E. and Chiti, F. (2007) Prevention of amyloid-like aggregation as a driving force of protein evolution. *EMBO Rep.*, **8**, 737–742.

23. Zamora,W.J., Campanera,J.M. and Luque,F.J. (2019) Development of a structure-based, ph-dependent lipophilicity scale of amino acids from continuum solvation calculations. *J. Phys. Chem. Lett.*, **10**, 883–889.
24. Santos,J., Iglesias,V., Santos-Suárez,J., Mangiagalli,M., Brocca,S., Pallarès,I. and Ventura,S. (2020) pH-dependent aggregation in intrinsically disordered proteins is determined by charge and lipophilicity. *Cells*, **9**, 145.
25. Pintado,C., Santos,J., Iglesias,V. and Ventura,S. (2021) SolupHred: a server to predict the ph-dependent aggregation of intrinsically disordered proteins. *Bioinformatics*, **37**, 1602–1603.
26. Oeller,M., Kang,R., Bell,R., Ausserwöger,H., Sormanni,P. and Vendruscolo,M. (2023) Sequence-based prediction of pH-dependent protein solubility using camisol. *Brief. Bioinform.*, **24**, bbad004.
27. Gokcan,H. and Isayev,O. (2022) Prediction of protein pKa with representation learning. *Chem. Sci.*, **13**, 2462–2474.
28. Gade Malmos,K., Blancas-Mejia,L.M., Weber,B., Buchner,J., Ramirez-Alvarado,M., Naiki,H. and Otzen,D. (2017) ThT 101: a primer on the use of thioflavin T to investigate amyloid formation. *Amyloid*, **24**, 1–16.
29. Oeller,M., Sormanni,P. and Vendruscolo,M. (2021) An open-source automated PEG precipitation assay to measure the relative solubility of proteins with low material requirement. *Sci. Rep.*, **11**, 21932.
30. Kuriata,A., Iglesias,V., Pujols,J., Kurcinski,M., Kmiecik,S. and Ventura,S. (2019) Aggrescan3D (A3D) 2.0: prediction and engineering of protein solubility. *Nucleic Acids Res.*, **47**, W300–W307.
31. Schymkowitz,J., Borg,J., Stricher,F., Nys,R., Rousseau,F. and Serrano,L. (2005) The FoldX web server: an online force field. *Nucleic Acids Res.*, **33**, W382–W388.
32. Dayhoff,M.O. (1972) A model of evolutionary change in proteins. *Atlas Protein Sequence Struct.*, **5**, 89–99.
33. Liò,P. and Goldman,N. (1998) Models of molecular evolution and phylogeny. *Genome Res.*, **8**, 1233–1244.
34. Dellacha,J.M., Santomé,J.A. and Paladini,A.C. (1968) Physicochemical and structural studies of bovine growth hormone. *Ann. NY Acad. Sci. U.S.A.*, **148**, 313–327.
35. Varadi,M., Anyango,S., Deshpande,M., Nair,S., Natassia,C., Yordanova,G., Yuan,D., Stroe,O., Wood,G., Laydon,A., *et al.* (2022) AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.*, **50**, D439–D444.
36. Badaczewska-Dawid,A.E., Kuriata,A., Pintado-Grima,C., Garcia-Pardo,J., Burdukiewicz,M., Iglesias,V., Kmiecik,S. and Ventura,S. (2024) A3D model organism database (A3D-MODB): a database for proteome aggregation predictions in model organisms. *Nucleic Acids Res.*, **52**, D360–D367.
37. Lucivero,F. (2020) Big data, big waste? a reflection on the environmental sustainability of big data initiatives. *Sci. Eng. Ethics*, **26**, 1009–1030.
38. Ruff,K.M. and Pappu,R.V. (2021) AlphaFold and implications for intrinsically disordered proteins. *J. Mol. Biol.*, **433**, 167208.
39. Tam,C. and Iwasaki,W. (2023) AlphaCutter: efficient removal of non-globular regions from predicted protein structures. *Proteomics*, **23**, e2300176.
40. Landrum,G.A. and Riniker,S. (2024) Combining IC50 or Ki values from different sources is a source of significant noise. *J. Chem. Inf. Model.*, **64**, 1560–1567.
41. Sidhu,S.S. (2000) Phage display in pharmaceutical biotechnology. *Curr. Opin. Biotechnol.*, **11**, 610–616.
42. Pintado-Grima,C., Bárcenas,O., Bartolomé-Nafria,A., Fornt-Suñé,M., Iglesias,V., Garcia-Pardo,J. and Ventura,S. (2023) A review of fifteen years developing computational tools to study protein aggregation. *Biophysica*, **3**, 1–20.

5.2. [A8] Aggrescan4D: kompleksowe narzędzie do analizy i projektowania podatności białek na agregację zależną od pH

5.2.1. Wprowadzenie do [A8]

Artykuł [A8], opublikowany w *Protein Science* w sekcji *Tools for Protein Science*, stanowi syntetyczne i zintegrowane omówienie narzędzi Aggrescan3D (A3D) i Aggrescan4D (A4D). Publikacja prezentuje wspólne założenia metodologiczne, architekturę, zastosowane metryki oraz przykłady wykorzystania tych narzędzi w analizie stabilności i skłonności białek do agregacji, podkreślając ich komplementarność oraz znaczenie dla badań nad właściwościami strukturalnymi białek. Artykuł [A8] nawiązuje bezpośrednio do wcześniejszej publikacji [A7] w *Nucleic Acids Research*, w której po raz pierwszy zaprezentowano serwer Aggrescan4D, jego koncepcję oraz wstępną implementację. W publikacji [A8] opracowałem omówienie pełnej funkcjonalności A4D, jego architektury, praktycznych zastosowań oraz wyników walidacji z wykorzystaniem biologicznie istotnych układów, a także szczegółowy opis zastosowanych metryk, algorytmów mutagenezy oraz mechanizmów integracji z bazami zewnętrznymi i interfejsami API.

Naszym celem było nie tylko dostarczenie narzędzia do oceny agregacji białek, ale również stworzenie zintegrowanego środowiska wspierającego projektowanie wariantów białek o lepszej rozpuszczalności. Aggrescan4D, łącząc dużą funkcjonalność z prostotą obsługi, może być wykorzystywany zarówno w badaniach akademickich, jak i przemysłowych.

A4D implementuje trzy główne tryby działania:

- statyczny, oparty na pojedynczej strukturze wejściowej (np. PDB lub model AlphaFold),
- dynamiczny, wykorzystujący CABS-flex do wygenerowania zespołów stanów konformacyjnych,
- pH-zależny, w którym użyto specjalnie opracowanej skali lipofilowości oraz predykcji pKa do oceny wpływu pH na lokalną agregacyjność.

Dla każdej reszty obliczana jest lokalna wartość A4D score, będąca funkcją jej własnych właściwości (agregacja, ekspozycja) oraz wpływu otoczenia przestrzennego w promieniu 10 Å. Ta metoda oceny została dokładnie zdefiniowana w pracy [A8] i umożliwia wyznaczanie agregacji zarówno na poziomie lokalnym (np. pojedynczej reszty lub segmentu), jak i globalnym - poprzez uśrednienie lub wybór wartości maksymalnej.

Na potrzeby trybu pH-zależnego, opracowaliśmy interpolowaną skalę lipofilowości opartą na danych SolupHred¹³⁰ oraz zastosowaliśmy predykcję pKa reszt kwasowych i zasadowych przy użyciu modelu pKa-ANI, opartego na głębokich sieciach neuronowych.

A4D umożliwia również identyfikację segmentów agregacji, czyli ciągłych fragmentów białka zawierających przynajmniej pięć kolejnych reszt o dodatnim A4D score, co może ułatwiać interpretację biologiczną wyników oraz projektowanie celowanych mutacji.

Jedną z najbardziej zaawansowanych funkcji A4D jest moduł automatycznego projektowania mutacji, dostępny w dwóch trybach:

- konserwatywnym ewolucyjnie, opartym na macierzy PAM250,
- zasadowo/kwasowym, sugerującym mutacje naładowanych reszt.

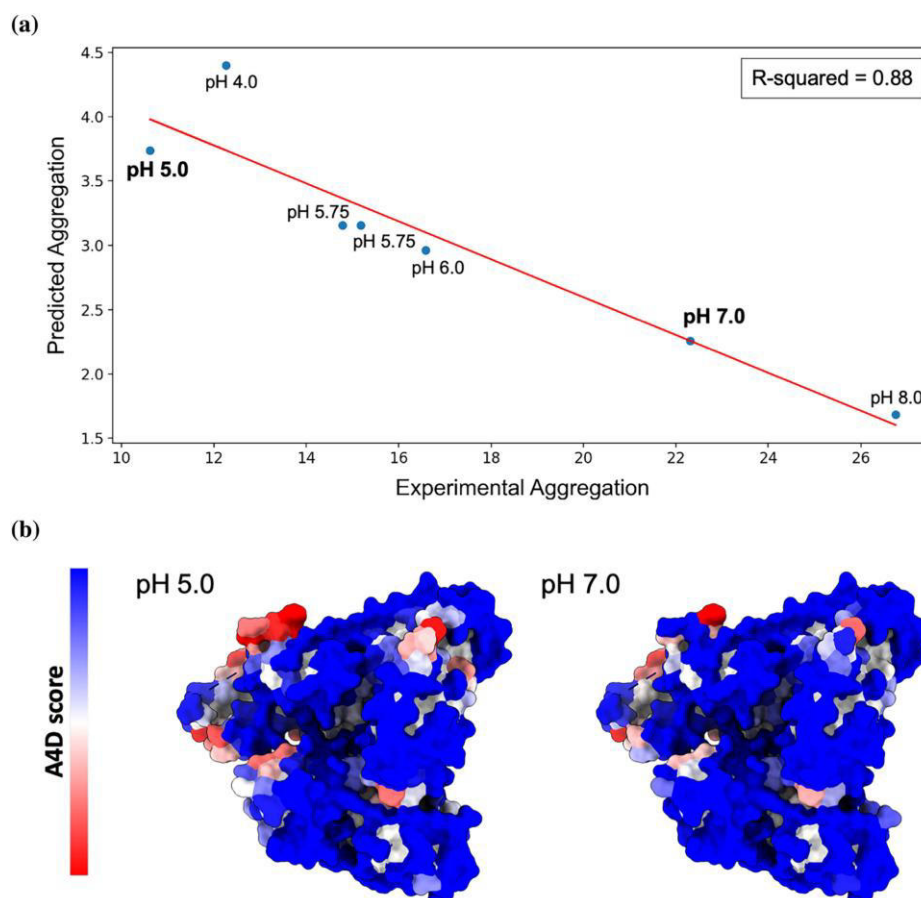
Mutacje są klasyfikowane według skuteczności w obniżaniu lokalnego A4D score ($\Delta A4D$), przy czym narzędzie pozwala również na ocenę wpływu mutacji na stabilność struktury za pomocą FoldX⁷⁹. Użytkownik może filtrować mutacje m.in. według ich pozycji, lokalizacji w domenie, typu chemicznego i efektu predykcji. Istnieje także możliwość zaprojektowania mutacji ręcznie, a następnie ich oceny - co czyni A4D elastycznym narzędziem do celowanej inżynierii białek.

W publikacji [A8] zaprezentowaliśmy przykłady zastosowania tej funkcji m.in. do zaprojektowania wariantów białka bGH, w których zasugerowane mutacje pokrywały się z naturalnymi różnicami między bardziej rozpuszczalną ludzką hGH i mniej rozpuszczalną bGH.

W artykule [A8] przedstawiliśmy również trzy pogłębione analizy przypadków:

- *Human Serum Albumin* (HSA) - Wartości A4D-max wzrastały przy obniżeniu pH, zgodnie z danymi eksperymentalnymi, osiągając $R^2 = 0,88$. Ponadto analiza przestrzenna wykazała, że główny region odpowiedzialny za agregację znajduje się na granicy domen i zwiększa swoją ekspozycję przy niższym pH. Wyniki tej analizy zostały zilustrowane na Rysunku 10, który pokazuje zarówno zależność wartości A4D-max od pH, jak i lokalizację przestrzenną regionów agregacyjnych ujawnionych w tych warunkach.
- Fragment *Fab* przeciwciała *bevacizumab* - Tryb dynamiczny ujawnił agregacyjne segmenty, które nie były widoczne w analizie pojedynczej struktury. Pokazuje to znaczenie konformacyjnej elastyczności dla trafnej predykcji.
- Zestaw 18 wariantów mAb (MEDI1912) - A4D uzyskało najlepsze wyniki predykcyjne w porównaniu z SolubiS i CamSol (trafność 82%, precyzja 80%). Co istotne, A4D poprawnie sklasyfikowało przypadki, których inne narzędzia nie rozpoznały.

Przeanalizowaliśmy także wartość diagnostyczną różnych metryk scoringowych oraz strategii interpretacji danych.



Rysunek 10. Analiza wpływu pH na skłonność do agregacji białka HSA uzyskana przy użyciu serwera Aggrescan4D (na podstawie [A8]). Górny panel (a) przedstawia zmianę wartości A4D-max w funkcji pH, a dolny panel (b) - lokalizację przestrzenną regionów agregacyjnych, które stają się bardziej eksponowane przy niższym pH. Wynik ten pokazuje, jak tryb pH-zależny A4D pozwala na uwzględnienie wpływu środowiska na agregacyjność białek. Źródło: [A8].

Dzięki tym przykładom wykazaliśmy, że A4D może być z powodzeniem stosowane nie tylko do predykcji, ale i do optymalizacji wariantów białek terapeutycznych.

A4D jest dostępny jako publiczny serwis webowy, z intuicyjnym GUI; REST API, umożliwiające automatyczne analizy dużych zbiorów danych (np. całych proteomów); oraz wersja *standalone*, kompatybilna z Linux, Windows i MacOS.

Dzięki REST API możliwe jest zautomatyzowanie przetwarzania struktury, mutagenezy, analizy domen i wizualizacji. W [A8] opisałem formaty wejściowe/wyjściowe oraz przykładowe zapytania HTTP, co czyni to rozwiązanie wyjątkowo elastycznym - zarówno dla użytkowników akademickich, jak i przemysłowych. Funkcja automatycznego pobierania

modeli AlphaFold i przycinania regionów nieustrukturyzowanych (AlphaCutter) minimalizuje błędy predykcyjne i skraca czas analizy.

Uważam, że praca [A8] stanowi istotny krok w stronę ujednolicenia i podniesienia standardu analiz agregacyjności białek. Dzięki połączeniu wielu źródeł informacji - od sekwencji i struktury, przez zmienność konformacyjną, aż po wpływ pH - udało się stworzyć narzędzie, które jest nie tylko predyktorem, ale także środowiskiem wspierającym projektowanie i optymalizację wariantów białek. Moim zdaniem właśnie ta integracja - dostępna zarówno w intuicyjnym interfejsie graficznym, jak i poprzez REST API - sprawia, że A4D jest dziś jednym z najbardziej wszechstronnych narzędzi w tej dziedzinie. Widać też, że przedstawione w artykule przykłady praktycznych zastosowań mają potencjał, by w najbliższych latach realnie wpłynąć na sposób projektowania białek terapeutycznych oraz analizę stabilności konstrukcji badanych zarówno w środowisku akademickim, jak i w przemyśle biotechnologicznym.

5.2.2. Artykuł [A8]

Aggrescan4D: A comprehensive tool for pH-dependent analysis and engineering of protein aggregation propensity

Mateusz Zalewski¹  | Valentin Iglesias^{2,3} | Oriol Bárcenas^{2,4} |
Salvador Ventura^{2,5}  | Sebastian Kmiecik¹ 

¹Faculty of Chemistry, Biological and Chemical Research Center, University of Warsaw, Warsaw, Poland

²Departament de Bioquímica i Biologia Molecular, Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Barcelona, Spain

³Clinical Research Centre, Medical University of Białystok, Białystok, Poland

⁴Institute of Advanced Chemistry of Catalonia (IQAC), CSIC, Barcelona, Spain

⁵Hospital Universitari Parc Taulí, Institut d'Investigació i Innovació Parc Taulí (I3PT-CERCA), Universitat Autònoma de Barcelona, Sabadell, Spain

Correspondence

Sebastian Kmiecik, Biological and Chemical Research Center, Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland.
Email: sekmi@chem.uw.edu.pl

Salvador Ventura, Institut de Biotecnologia i de Biomedicina and Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, Barcelona, Spain.
Email: salvador.ventura@uab.cat

Funding information

Polish National Agency for Academic Exchange under the ULAM NAWA Programme, Grant/Award Number: BPN/ULM/2023/1/00189/U/00001; Spanish Ministry of Science and Innovation, Grant/Award Number: PID2022-137963OB-I00; Generalitat de Catalunya, Grant/Award Number: 2021-SGR-00635 AGA; ICREA, ICREA-Academia 2020; National Science Center, Grant/Award Number: Sheng2021/40/Q/NZ2/00078

Review Editor: Nir Ben-Tal.

Abstract

Aggrescan4D (A4D) is an advanced computational tool designed for predicting protein aggregation, leveraging structural information and the influence of pH. Building upon its predecessor, Aggrescan3D (A3D), A4D has undergone numerous enhancements aimed at assisting the improvement of protein solubility. This manuscript reviews A4D's updated functionalities and explains the fundamental principles behind its pH-dependent calculations. Additionally, it presents an antibody case study to evaluate its performance in comparison with other structure-based predictors. Notably, A4D integrates advanced protein engineering protocols with pH-dependent calculations, enhancing its utility in advising solubility-enhancing mutations. A4D considers the impact of structural flexibility on aggregation propensities, and includes a large set of precalculated predictions. These capabilities should help to open new avenues for both understanding and managing protein aggregation. A4D is accessible through a dedicated web server at <https://biocomp.chem.uw.edu.pl/a4d/>.

KEYWORDS

computational biology, monoclonal antibodies, pH-dependent aggregation, protein aggregation, protein engineering, protein solubility, protein stability, structural bioinformatics

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Author(s). *Protein Science* published by Wiley Periodicals LLC on behalf of The Protein Society.

1 | INTRODUCTION

Protein aggregation is a significant biochemical process with implications for both health and technology. It occurs when soluble proteins form insoluble aggregates, often leading to diseases such as Parkinson's, Alzheimer's, ALS, type II diabetes, and various cancers (Chiti & Dobson, 2017). From an evolutionary perspective, protein aggregation can limit the diversity of protein sequences that can be stably expressed in cells. This is because the aggregation process affects several crucial aspects, including the proper folding of proteins into their native states and the concentration at which they can exist without forming aggregates (Carija et al., 2019; de Groot & Ventura, 2010; Santos, Iglesias, & Ventura, 2020). In industry, although protein aggregation poses challenges in biotechnology, particularly in the production and purification of biotherapeutics, it also presents unique opportunities (Shukla & Thömmes, 2010). Amyloid fibers, utilized as scaffolds (Díaz-Caballero et al., 2018; Knowles & Mezzenga, 2016) and enzymatic platforms (Díaz-Caballero et al., 2018; Knowles & Mezzenga, 2016), exemplify how biological challenges can be transformed into technological advantages. Understanding these aggregation processes is essential for both mitigating their harmful effects and harnessing their potential benefits.

The growing need to understand protein aggregation across various contexts has spurred the development of sequence-based bioinformatics tools designed to pinpoint regions in proteins most susceptible to aggregation (Santos, Pujols, Pallarès, et al., 2020). Tools such as TANGO (Fernandez-Escamilla et al., 2004), Zygggregator (Tartaglia & Vendruscolo, 2008), and Aggrescan (Conchillo-Solé et al., 2007) are highly valued for their speed and cost-effectiveness, facilitating the rapid redesign of protein variants. TANGO (Fernandez-Escamilla et al., 2004) utilizes a statistical mechanics approach to predict protein aggregation by evaluating beta-sheet formation tendencies, focusing on peptide backbone energetics and side chain interactions. Zygggregator (Tartaglia & Vendruscolo, 2008) determines aggregation propensities using physicochemical properties and the context within protein sequences. Meanwhile, Aggrescan (Conchillo-Solé et al., 2007) identifies critical "hot spots" for aggregation using an aggregation propensity scale derived from *in vivo* data, making it particularly effective for detecting short sequence stretches that significantly influence protein aggregation. These tools are particularly useful for proteins in extended conformations or intrinsically disordered proteins (IDPs), where exposed aggregation prone regions (APRs) are directly related to the sequence (Belli et al., 2011). However,

sequence-based tools often misidentify APRs in folded proteins, where hydrophobic regions involved in interactions or buried within the protein are less likely to contribute to aggregation.

This limitation led to the development of structure-based predictors that utilize spatial information from experimental structures to identify aggregation-prone regions where residues are closely packed in three dimensions (Chennamsetty et al., 2009). Key advancements in this area include SolubiS (De Baets et al., 2015), CamSol (Sormanni et al., 2015), and Aggrescan3D (A3D) (Zambrano et al., 2015). These tools have improved the accuracy of predicting protein aggregation by effectively using structural data to pinpoint potential hotspots. SolubiS (De Baets et al., 2015) builds on TANGO's (Fernandez-Escamilla et al., 2004) methodology to improve protein solubility by mutating aggregation-prone regions to more stable residues, utilizing the FoldX (Schymkowitz et al., 2005) force field to assess the impact of these changes. CamSol (Sormanni et al., 2015), on the other hand, predicts protein solubility by taking into account factors such as hydrophobicity, charge, and structural context, and it provides users with protein variants engineered for increased solubility. A3D (Zambrano et al., 2015) builds on the original Aggrescan (Conchillo-Solé et al., 2007) by adding three-dimensional structural information to its sequence-based aggregation propensity analysis. This allows A3D to identify aggregation-prone regions based on both the sequence and the spatial proximity of residues, giving a more accurate depiction of potential aggregation hotspots in the three-dimensional structure of proteins.

Building on the strengths of A3D (Zambrano et al., 2015), our new in-house algorithm, referred to as Aggrescan4D (A4D) (Bárcenas et al., 2024), represents the next step in the evolution of aggregation prediction tools. A4D not only incorporates all the features of A3D but also introduces environmental pH as a variable in its calculations, offering a more nuanced view of protein aggregation under different physiological conditions. Additionally, A4D incorporates further enhancements for automated protein design and user-friendly operation.

In this overview, we explore the capabilities of A4D (Bárcenas et al., 2024), a state-of-the-art tool for predicting protein aggregation propensity, and compare its strengths and weaknesses with other aggregation prediction algorithms. Section 2 highlights the transformative advancements from A3D (Zambrano et al., 2015) to A4D (Bárcenas et al., 2024), emphasizing its key features, methodologies, and applications. Section 3 offers a concise overview of A4D's options, including its most important functionalities such as: pH calculations (Section 3.1), dynamic mode simulations (Section 3.4), and automatic

mutations (Section 3.7). Each section delves into A4D's capabilities, underscoring its significance in protein aggregation research and potential impact on biotechnology and biomedical science. Additionally, in Section 4, we describe an antibody case study evaluating the predictive accuracy of A4D, alongside other tools, providing valuable insights into its predictive accuracy. Finally, we also discuss potential future enhancements that could improve our ability to predict and manage protein aggregation.

2 | FROM Aggrescan3D TO Aggrescan4D: TRANSFORMATIVE ADVANCES IN PROTEIN AGGREGATION PREDICTION TECHNOLOGY

A3D (Zambrano et al., 2015) integrates the Aggrescan *in vivo*-derived scale (Conchillo-Solé et al., 2007) into the 3D space of protein structures by leveraging distances between amino acid residues and their overall surface exposure. This method overcomes the main limitations of sequence-based predictive approaches by disregarding solvent-protected residues that have negligible impact on protein aggregation, identifying spatially clustered residues that form Structural Aggregation Prone regions (STAPs) despite being sequentially distant, and evaluating the decay of residues' impact on STAP potency due to distance between alpha-carbon atoms (Conchillo-Solé et al., 2007; Sánchez de Groot et al., 2005). A3D's distinctive feature CABS-flex (Jamroz et al., 2014; Kurcinski et al., 2019; Kuriata et al., 2018) provides a dynamic model of protein flexibility which allows for evaluating the aggregation propensity of proteins by identifying potential STAPs that may be hidden in the static input structure (Nithin et al., 2024; Zambrano et al., 2015).

The first web server for A3D (Zambrano et al., 2015) was launched in 2015, marking a significant milestone in making this innovative tool accessible to the wider scientific community. Since its launch, A3D has been widely accepted in the scientific community, leading to numerous significant publications. Researchers have utilized A3D for the *in silico* redesign of antibodies and other proteins, enhancing their structural and functional properties for various biotechnological applications (Ejaz et al., 2023; Gil-Garcia et al., 2018; Xia et al., 2016). Additionally, the platform has been instrumental in understanding the aggregation mechanisms of both pathogenic and non-pathogenic proteins as well as biotherapeutics, providing insights into their behaviors and impacts (Berdyński et al., 2022; Bhandare & Ramaswamy, 2018; Gil-Garcia et al., 2020; Katina et al., 2017; Ko et al., 2023;

Leppert et al., 2023; Petrlova et al., 2022; Zakeri et al., 2023; Žerovnik, 2017). A3D has also supported studies on the interactions between antibacterial proteins and lipidic bilayers, as well as disaggregase-client recognition dynamics, aiding in the development of multi-epitope vaccines for tropical diseases and enhancing our understanding of protein evolution and its impact on aggregation (Badaczewska-Dawid et al., 2023; Carija et al., 2019; Flores-León et al., 2021; Garcia-Pardo et al., 2023; Ishwarlall et al., 2022; Pulido, Arranz-Trullén, et al., 2016; Pulido, Llamas, et al., 2016; Ruiz-Solaní et al., 2023).

In 2019, the A3D 2.0 web server (Kuriata, Iglesias, Pujols, et al., 2019) was launched with major improvements, including the new CABS-flex package and the FoldX (Schymkowitz et al., 2005) force field. This version's enhanced capabilities enabled it to handle larger proteins and complexes, facilitating rational protein redesign by balancing aggregation propensity with thermodynamic stability (Castillo et al., 2010; Espargaró et al., 2008; Gil-Garcia et al., 2018; Kurcinski et al., 2019; Kuriata, Iglesias, Kurcinski, et al., 2019; Kuriata, Iglesias, Pujols, et al., 2019; Schymkowitz et al., 2005). New automated mutation tools were also introduced to streamline protein engineering processes, making them more accessible to users with varying levels of expertise in protein chemistry.

In response to growing demand for local and customizable versions, the A3D Standalone (Kuriata, Iglesias, Kurcinski, et al., 2019) version was developed, offering the full functionalities of A3D 2.0 (Kuriata, Iglesias, Pujols, et al., 2019) but with the capability for users to operate locally and customize features to suit their specific needs. This standalone version, especially useful for integrating into larger bioinformatics platforms, was well-received by researchers preferring to work within their own computational environments.

In 2024, A3D (Zambrano et al., 2015) underwent a significant update leading to its rebranding as A4D web server (Bárcenas et al., 2024), driven by the inclusion of pH calculations as the "fourth dimension." This transformative enhancement, along with the integration of environmental factors like pH, allows for more realistic simulations of physiological conditions and provides a comprehensive understanding of how these conditions influence protein stability and aggregation (Arseni et al., 2022; Arseni et al., 2023; Garcia-Pardo & Ventura, 2024; Oeller et al., 2023; Pintado et al., 2021; Santos, Iglesias, Santos-Suárez, et al., 2020; Strohäker et al., 2019). A4D also capitalized on the AlphaFold database (Jumper et al., 2021), enhancing the accuracy of structural predictions by addressing the representation of intrinsically disordered regions (IDRs) and allowing

seamless integration of structural data from various databases (Badaczewska-Dawid et al., 2022; Badaczewska-Dawid et al., 2023; Garcia-Pardo et al., 2023; Ruff & Pappu, 2021; Varadi et al., 2024). Additionally, new automated mutation modes and expanded file compatibility have further enhanced A4D's utility in protein solubility and aggregation studies, marking a pivotal development in the field of protein aggregation prediction.

3 | A4D PIPELINE AND OPTIONS

A4D web server (Bárceñas et al., 2024) (available at <https://biocomp.chem.uw.edu.pl/a4d/>) offers users the choice between utilizing the original A3D scoring method (Zambrano et al., 2015), featuring Static mode and Dynamic mode, or adopting the A4D algorithm, which integrates pH calculations (see Figure 1).

A4D (Bárceñas et al., 2024) defaults to conducting a pH-dependent analysis of aggregation properties and structural stability assessment. It also facilitates input optimization using the FoldX (Schymkowitz et al., 2005) tool. The aggregation propensity calculations are centered around each amino acid α -Carbon atom within a default spherical radius of 10 Å. The A4D score quantifies aggregation propensity for each amino acid i using the following formula:

$$A4D \text{ score} = \text{Agg}_i(\alpha e^{\beta \text{RSA}_i}) + \sum_e [\text{Agg}_e(\alpha e^{\beta \text{RSA}_e}) (\gamma e^{-\delta \text{dist}})] \quad (1)$$

In this equation:

- Agg_i represents the aggregation propensity of the central residue i , and RSA_i denotes its relative surface area exposed to solvent. For static and dynamic mode pathways (Figure 1), the Agg_i score is based on experimental data from *in vivo* studies (Conchillo-Solé et al., 2007; Sánchez de Groot et al., 2005), whereas, in pH calculation mode, Agg_i is determined using a pH-dependent lipophilicity scale (details in Section 3.1).
- The summation includes each neighboring residue e within a 10 Å radius around i , where Agg_e and RSA_e are their respective aggregation propensities and solvent exposure.
- dist refers to the distance from the central residue i to each neighboring residue e .
- α , β , γ , and δ are numerical parameters used to adjust the influence of solvent exposure and distance on the aggregation score, reflecting residue interactions as outlined in work (Zambrano et al., 2015).

For most applications, the A4D score above 0 indicates potential aggregation, but scores higher than 0.5 are more likely to indicate residues that are critical to aggregation processes. A sphere radius of 10 Å is typically used to identify residues forming aggregation-prone regions, though a smaller radius of 5 Å may be employed for detailed analysis of specific residues within these patches (see Section 3.6).

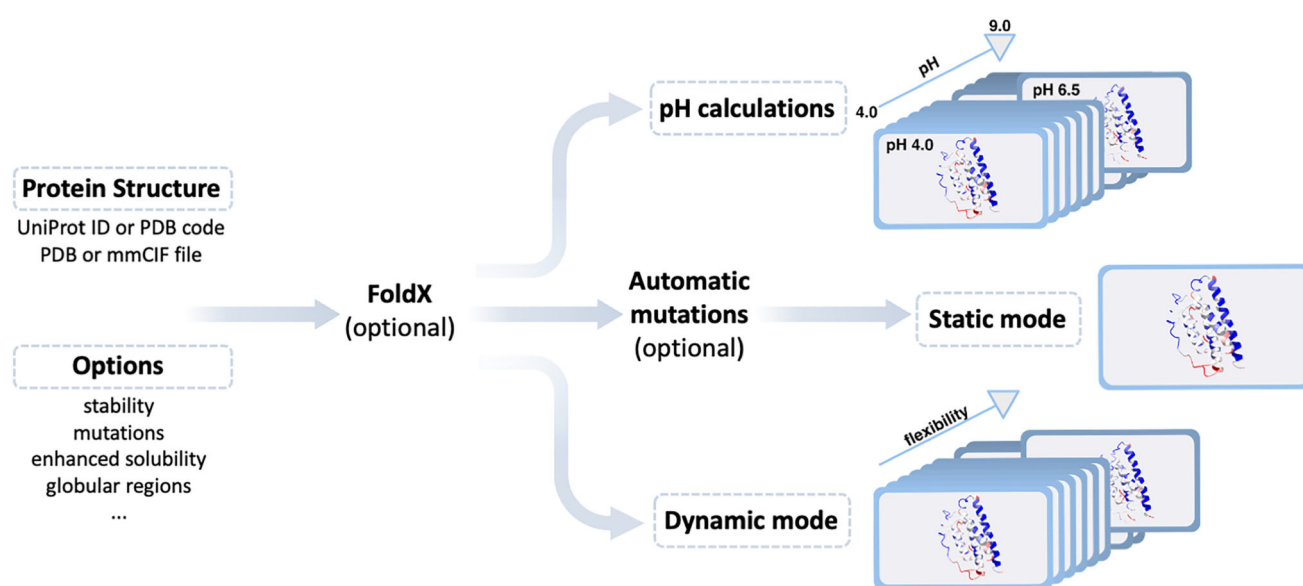


FIGURE 1 A4D pipeline.

TABLE 1 A4D advanced options list.

Name	Description	Value	Note	Section
pH calculations	The simulation will be run across a range of pH values (4–9), results for each pH value will be presented.	Yes/No	Cannot be used alongside Dynamic mode and Improve/Enhance protein stability options.	3.1
Stability calculations	FoldX (Schymkowitz et al., 2005) is used to optimize input structure.	Yes/No	Required for mutation calculations.	3.2
Analysis of the globular regions	Use the AlphaCutter tool (Tam & Iwasaki, 2023) to remove the disordered regions of the protein.	Yes/No	Recommended for structures obtained from the AlphaFold database.	3.3
Dynamic mode	If Yes, A4D will take into account the flexibility of an input structure. If No, A4D will perform analysis on a static input structure.	Yes/No	Cannot be used alongside pH calculations and Improve/Enhance protein stability options.	3.4
Mutate residues	Mutate your input protein, mutation options will be provided in the next step.	Yes/No	–	3.5
Distance of aggregation analysis:	Select the radius (R) of the sphere used in the analysis.	5 Å/10 Å	–	3.6
Improve protein solubility with evolutionary conserved mutations	Automatically select aggregation prone residues and mutates them into more soluble amino acids using a conservative evolutionary approach.	Yes/No	Cannot be used alongside pH calculations, Dynamic mode and Enhance protein solubility options.	3.7
Enhance protein solubility with charged mutations	Automatically selects aggregation prone residues and mutates them into charged amino acids.	Yes/No	Cannot be used alongside pH calculations, Dynamic mode and Improve protein solubility options.	3.7

Note: Each option is described in the corresponding subsection of this manuscript, with the subsection number listed in the last column. Default options are highlighted in bold.

The following section provides a comprehensive explanation of A4D's (Bárcenas et al., 2024) modes and features, ensuring users have a clear understanding of its capabilities for predicting protein aggregation propensities. All available options and their default settings are presented in Table 1.

On the A4D web server, computation times vary based on protein size and the selected options. For most settings, smaller proteins (around 100 amino acids) typically require only a few minutes, while larger proteins (around 300 amino acids) usually take between 10 and 15 min. However, dynamic mode is particularly time-consuming, taking about 20 min for smaller proteins and over an hour for larger ones. Note that these times may vary if the job is queued on the A4D web server.

3.1 | pH calculations

The possibility of including pH-dependent calculations was introduced in 2024 (Bárcenas et al., 2024), with the rebranding of A3D to A4D, marking the integration of environmental pH as the “4th dimension.” The integration of pH calculations into A4D marks a significant leap forward in protein research and

engineering. It empowers researchers to explore the impact of pH variations on protein aggregation in biotechnological contexts, aids in optimizing protein solubility across diverse environmental conditions crucial for industrial applications, and enhances our understanding of disease mechanisms, particularly in conditions where pH dysregulation contributes to protein aggregation-related pathologies.

The propensity for aggregation of pH-sensitive amino acids is influenced by changes in their charge and hydrophobicity, which vary with pH. This results in a pH-dependent alteration in their contribution to aggregation (Santos, Iglesias, Santos-Suárez, et al., 2020). SolupHred has used these factors successfully to predict how pH influences the aggregation of intrinsically disordered proteins (IDPs) (Pintado et al., 2021). The residues neighboring the pH-sensitive residues can shift the pKa values. A4D combines the pH-dependent lipophilicity scale used in SolupHred with 3D-informed pKa calculations for pH-sensitive residues available through the pKa-ANI method (Gokcan & Isayev, 2022) (Figure 2). To calculate the aggregation propensity index (Agg_i) for each amino acid at different pH values, we use these combined data and substitute them into the Equation (1). It's important to note that while there is a strong correlation between the

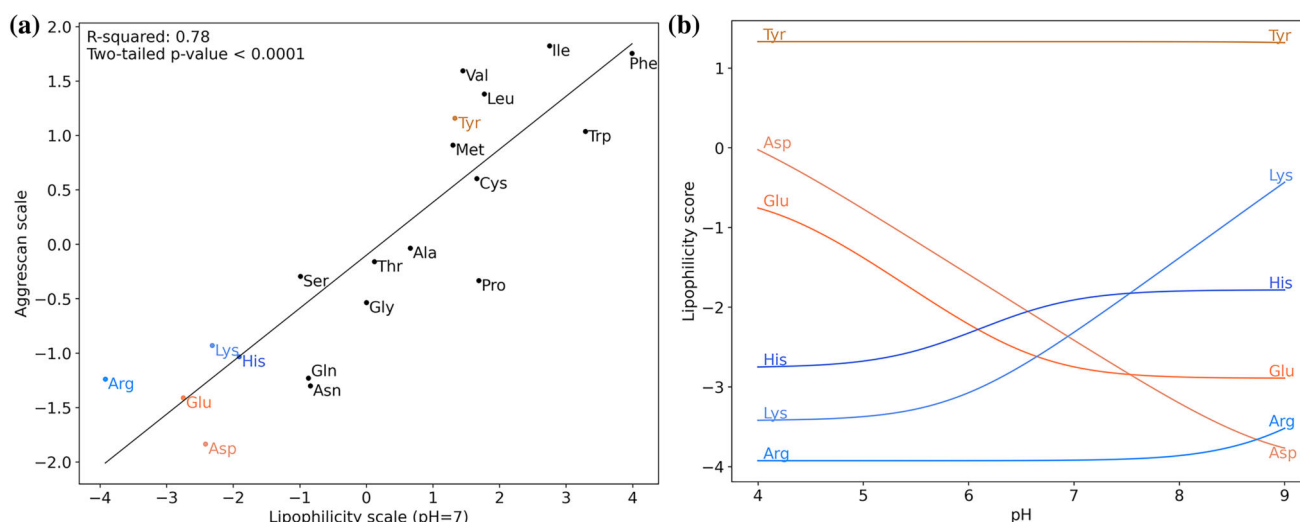


FIGURE 2 Incorporation of pH modulated changes in aggregation propensity in A4D. (a) Linear correlation between the pH-dependent lipophilicity scale used for pH calculations at neutral pH and the *in vivo* derived Aggrescan scale used for non-pH calculations. (b) pH-driven changes in the lipophilicity score. Due to limitations on satisfactory 3D pKa estimation for Arginine, its score is considered constant in the A4D assayed pH-range.

primary aggregation scale of A4D and the pH-dependent lipophilicity scale, their dynamic ranges are different. As a result, A4D reports different numerical values for the aggregation score at neutral pH compared to the “classical” A4D score (i.e., when pH calculations are not considered) (Figure 2a). Regarding pH calculations, A4D considers the pH score variation of aspartic and glutamic acids, lysine, histidine and tyrosine (see Figure 2b), while the score of other amino acids remains unchanged.

If the ‘pH calculations’ option is enabled, A4D facilitates aggregation predictions across a range of 11 pH values from 4 to 9 with increments of 0.5. Subsequently, users gain access to detailed pH-specific predictions on the result page, including structural visualization along with average and max A4D scores for each prediction. Through performance analysis, we determined that for globular proteins, the maximum value, and for IDPs, the average value, better aligns with experimental observations (Bárceñas et al., 2024). Furthermore, A4D adeptly recognizes input structures and recommends the most suitable metric for use. It is important to note that this functionality cannot be combined in a single prediction run with Dynamic mode and automatic mutations, as illustrated in Figure 1.

In our study covering A4D web server (Bárceñas et al., 2024), we examined the accuracy of A4D’s pH-dependent predictions across 18 diverse proteins, including 11 globular proteins and 7 IDPs. We measured the performance by comparing the slopes of the linear regressions between the experimental data and A4D metrics. Specifically, we used the A4D max score for globular proteins and the A4D average score for IDPs, achieving 100%

agreement between the predictions and experimental results for both categories. In comparison, CamSol had an overall agreement of 78% (64% for globular proteins and 100% for disordered proteins).

The usage of A4D in predicting pH-dependent aggregation behaviors can be illustrated through the example of Human Serum Albumin (HSA), a critical blood protein involved in substance transportation. Although HSA tends to aggregate into amyloid fibrils under destabilizing conditions, such as temperature or pH changes, the precise impact of pH on its structure and the consequent exposure of the hydrophobic core—which can increase aggregation—remains complex and unpredictable. Therefore, HSA serves as a useful but challenging model for studying aggregation-related phenomena (Malik et al., 2022). Recent studies have illuminated the aggregation behavior of Human Serum Albumin (HSA) under pH alterations, particularly in the shift from pH 7 to 5. Notably, the predictions generated by A4D regarding the aggregation tendencies of HSA (PDB ID: 6YG9) are supported by experimental findings. Utilizing the pH calculations feature of A4D, it was found that at pH 5, HSA shows a heightened propensity for aggregation compared to pH 7. These findings are reflected in the scatter plot shown in Figure 3a, which illustrates a strong concordance between predicted aggregation propensity scores and experimental data, evidenced by an *R*-squared value of 0.88, indicating a robust correlation.

In Figure 3, we present an example where the correlation between experimental data and the predicted A4D score is high across the entire pH range studied. After analyzing the pH-dependent aggregation profile of

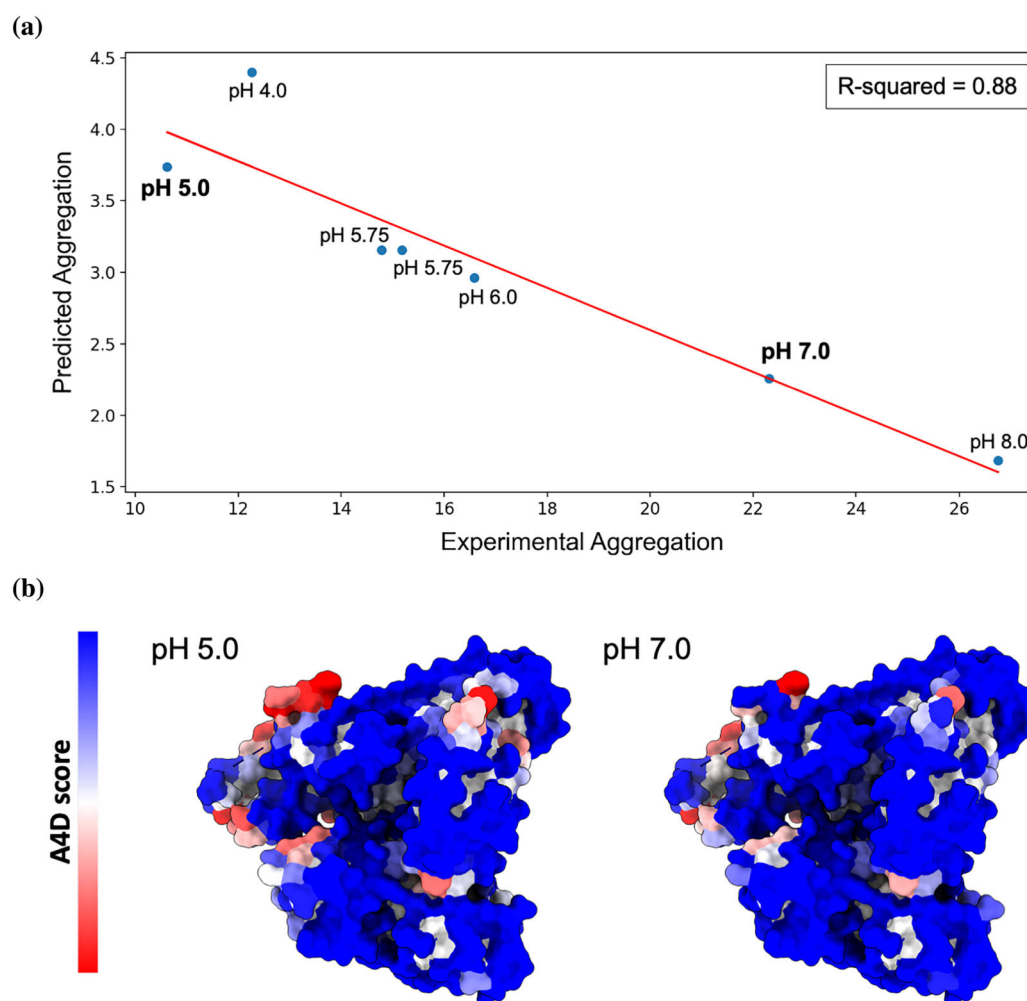


FIGURE 3 Human Serum Albumin (PDB ID: 6YG9) aggregation propensity predictions using A4D pH calculations. (a) Scatter plot illustrating the relationship between experimental data (X-axis) (Oeller et al., 2023) and predicted aggregation propensity scores (Y-axis). Each data point represents a distinct pH, with the experimental values plotted against their corresponding predicted aggregation propensity scores, while the red line represents the linear regression fit. The experimental aggregation value represents the measured aggregation propensity of proteins under specific conditions, as reported in the original A4D work (Bárcenas et al., 2024). For BSA, it is the percentage of aggregated proteins, and for lysozyme and RNase Sa, it is ThT absorbance at 595 nm. All values are normalized. The linear regression model was fitted to the experimental data to estimate the relationship between the experimental values and the predicted aggregation propensity scores. The close alignment between the points and the fitted line indicates a high level of concordance between the experimental data and the predicted aggregation propensity scores. (b) Structural representations of HSA at pH 5.0 (left) and 7.0 (right). Residues are colored by A4D scale: Blue for solubility, red for aggregation, white for neutral regions. Higher predicted propensity for aggregation at pH 5.0, compared to pH 7.0, aligns with experimental findings.

multiple proteins (Bárcenas et al., 2024), various scenarios are possible regarding this correlation. In Figure 4, we have compiled predicted and experimental aggregation values from three proteins: Bovine Serum Albumin (BSA) (Li et al., 2016), Lysozyme (Bárcenas et al., 2024) (supplementary materials), and Ribonuclease Sa (RNase Sa) (Schmittschmitt & Scholtz, 2003). For BSA, the predicted A4D values accurately model the observed experimental aggregation. When analyzing the lysozyme data, we observe that A4D accurately models most of the pH range (pH 3.5–9.5) but fails for moderately basic pH

values (10 and above). This discrepancy is attributed to the partial unfolding of lysozyme at such pH values (Petersen et al., 2004), which is why the A4D server limits pH prediction range to values between 4.0 and 9.0. Finally, in the case of RNase Sa, the predicted and experimental values agree, showing a decrease in aggregation with an increase in pH. However, the rate of change differs. This is a known limitation in the algorithm for some protein case studies and may result from the limitations in A4D's aggregation modeling capabilities or differences in the sensitivity of the experimental value reported.

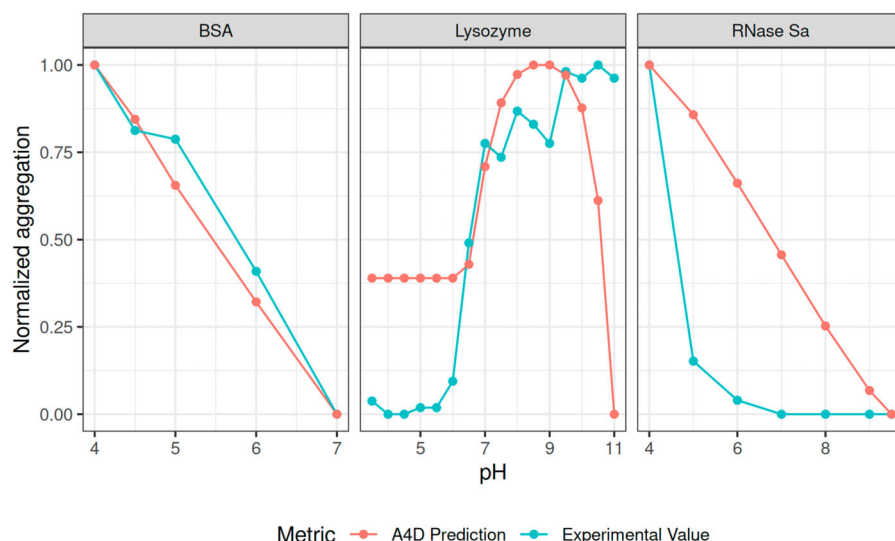


FIGURE 4 Examples of different pH-dependent normalized aggregation profiles, showcasing different behaviors in protein aggregation. Bovine Serum Albumin (BSA), Lysozyme, Ribonuclease Sa (RNase Sa); PDB IDs: 4F5S, 1LYZ chain A, 1C54 chain A, respectively. The predicted values align very well with the experimental data (BSA), fail in extreme pHs (basic pH, lysozyme), or show a different range of variability between experimental and predicted data (RNase SA).

3.2 | Stability calculations

When activated, this feature harnesses FoldX (Schymkowitz et al., 2005) to optimize input structure, consequently enhancing prediction accuracy. The setting should be toggled to “No” only for input structures that have been already optimized, in order to expedite calculations. To perform any mutation, or dynamic mode analysis, this feature has to be enabled. By default, the setting is preset to “Yes.”

3.3 | Analysis of the globular regions

Structures obtained from the AlphaFold database (Jumper et al., 2021) often present elongated, disordered regions that are not typically observed in experimentally resolved structures. These regions tend to congregate around the globular core, introducing potential surface features that could mislead the interpretation of STAPs (Ruff & Pappu, 2021) and undermine the accuracy of A4D predictions (Badaczewska-Dawid et al., 2022; Badaczewska-Dawid et al., 2023).

To address this concern, A4D introduces the “Analysis of the globular regions” feature, which employs AlphaCutter (Tam & Iwasaki, 2023) to selectively remove these disordered segments. AlphaCutter employs DSSP to pinpoint long loops and helices, followed by the calculation of a contact map. Additionally, it applies an AlphaFold pLDDT threshold to refine the domains further. By focusing on globular regions, this approach refines the structural assessment, thereby reducing the impact of artificially generated surface areas. By default, this option is deactivated.

3.4 | Dynamic mode

In some cases, analyzing static structures alone is insufficient; we must also consider their dynamics to accurately predict aggregation capabilities. Protein aggregation onset may not necessarily require global unfolding and transitioning across major energy barriers but can originate from locally unfolded states accessible under physiological conditions, such as thermal fluctuations of native states (Chiti & Dobson, 2009). A4D Dynamic mode is designed to capture and analyze this kind of structural changes, and thereby to facilitate a more comprehensive and dynamic perspective of protein aggregation tendencies. A4D Dynamic mode harnesses CABS-flex tool (Kurcinski et al., 2019; Nithin et al., 2024) to simulate protein fluctuations, allowing for the incorporation of information about protein flexibility into their aggregative properties. In this mode, an input structure, after FoldX (Schymkowitz et al., 2005) optimization, serves as a starting point for CABS-flex simulation. During this stage, a set of all-atom structures representing the predominant fluctuations observed throughout the CABS-flex conformational space sampling is produced. For each of these models, the A4D score is calculated using the same method as in the static mode, and models are ranked from highest to lowest aggregation score, including the original structure. Following the completion of calculations, a comprehensive summary of the Dynamic mode computations, including the models generated via CABS-flex and their corresponding A4D scores, is accessible in the “Dynamic mode details” tab. This summary also includes an RMSF plot, where higher RMSF values generally indicate more flexible regions, potentially exposing hydrophobic patches or aggregation-prone sequences that are otherwise buried. It's important to

note that the Dynamic mode does not support pH calculations or automatic mutations (see Figure 1). By default, Dynamic mode is set to “No.”

Below, we present an example of the bevacizumab Fab fragment (PDB ID: 1BJ1), a monoclonal antibody crucial for inhibiting vascular endothelial growth factor A (VEGF-A) in the treatment of various cancers and eye diseases. The Fab domain of bevacizumab is known for its high aggregation propensity, necessitating reduced concentrations in its therapeutic applications. While A4D static analysis of the Fab domain indicates a certain level of aggregation tendency, the incorporation of Dynamic mode unravels undisclosed aggregation-prone regions not apparent in static structural evaluation (see Figure 5). This suggests that structural fluctuations contribute to an expanded aggregation-prone area, emphasizing the significance of considering protein dynamics in aggregation predictions. The Fab fragment example, along with analogous observation for different antibodies has been comprehensively described in ref (Kuriata, Iglesias, Pujols, et al., 2019), in which the A3D 2.0 tool was used in the analysis of protein flexibility on aggregation properties.

3.5 | Mutations

A4D offers the capability to introduce mutations both before and after conducting calculations, enhancing users' control over their analyses. By default, the Mutate residues option is initially set to “No.” However, users can activate this feature to access a dedicated window where they can perform individual or multiple mutations. This process involves selecting the desired amino acid from a provided list and specifying the target amino acid for mutation. Importantly, unlike automated mutation methods, the Mutate residues option seamlessly integrates with pH calculations and Dynamic mode. Notably, enabling the Mutate residues option requires the activation of Stability calculations, leveraging the FoldX (Schymkowitz et al., 2005) tool for mutation provision.

Mutations can be also performed, through the “Aggrescan4D score” tab, accessible on the results page after completing calculations. This tab features an interactive table displaying the A4D values for each residue, along with an intuitive interface enabling users to easily select desired mutations and resubmit calculation for mutated protein.

3.6 | Distance of aggregation analysis

Aggregation propensity is determined within spherical regions centered on each residue's C α carbon. The

“Distance of aggregation analysis” option offers a choice between two distinct distance settings: 5 or 10 Å.

The selection between the 5 and 10 Å settings depends on the desired level of precision and detail in the analysis. The default 10 Å radius provides a comprehensive overview of aggregation tendencies throughout the protein structure, while opting for the 5 Å radius allows for a more refined examination, enabling researchers to delve into individual residue impacts with greater precision.

3.7 | Automated mutations to improve protein solubility

A4D offers an Automated mutation feature aimed at identifying more soluble variants of functional therapeutic proteins. This feature serves as an efficient and user-friendly method, accessible via “Enhance protein solubility with charged mutations” and “Improve protein solubility with evolutionary conserved mutations” input options. Its purpose is to offer a computational alternative to time-consuming and costly experimental approaches such as phage display.

Under the setting “Enhance protein solubility with charged mutations,” A4D identifies aggregation-prone regions within the protein structure and virtually introduces mutations by incorporating charged amino acids at corresponding residues to impede protein self-association. Upon analysis completion, A4D generates a ranked list of suggested point mutations, accompanied by an evaluation of their solubilizing effects and energetic implications. This empowers researchers to manually select mutations that enhance solubility while considering potential destabilization. Furthermore, users can exclude specific residues from the automated mutation process to prevent mutations of critical functional fragments within the protein sequence.

The setting “Improve protein solubility with evolutionarily conserved mutations” focuses on minimizing changes in physicochemical properties that could disrupt protein folding and stability. This option ensures that suggested mutations result from a single point mutation (A $\rightarrow$ B) rather than a series of successive changes (A $\rightarrow$ C $\rightarrow$ B), thus minimizing impacts on protein structure and stability. To achieve this, A4D employs Point Accepted Mutation (PAM) matrices. These matrices store the sum of accepted substitutions in closely related sequences, with an identity threshold of 0.85 or higher. By utilizing PAM matrices, A4D ensures that the suggested mutations preserve evolutionary conservation while enhancing protein solubility. This approach optimizes protein stability and maintains biological functionality.

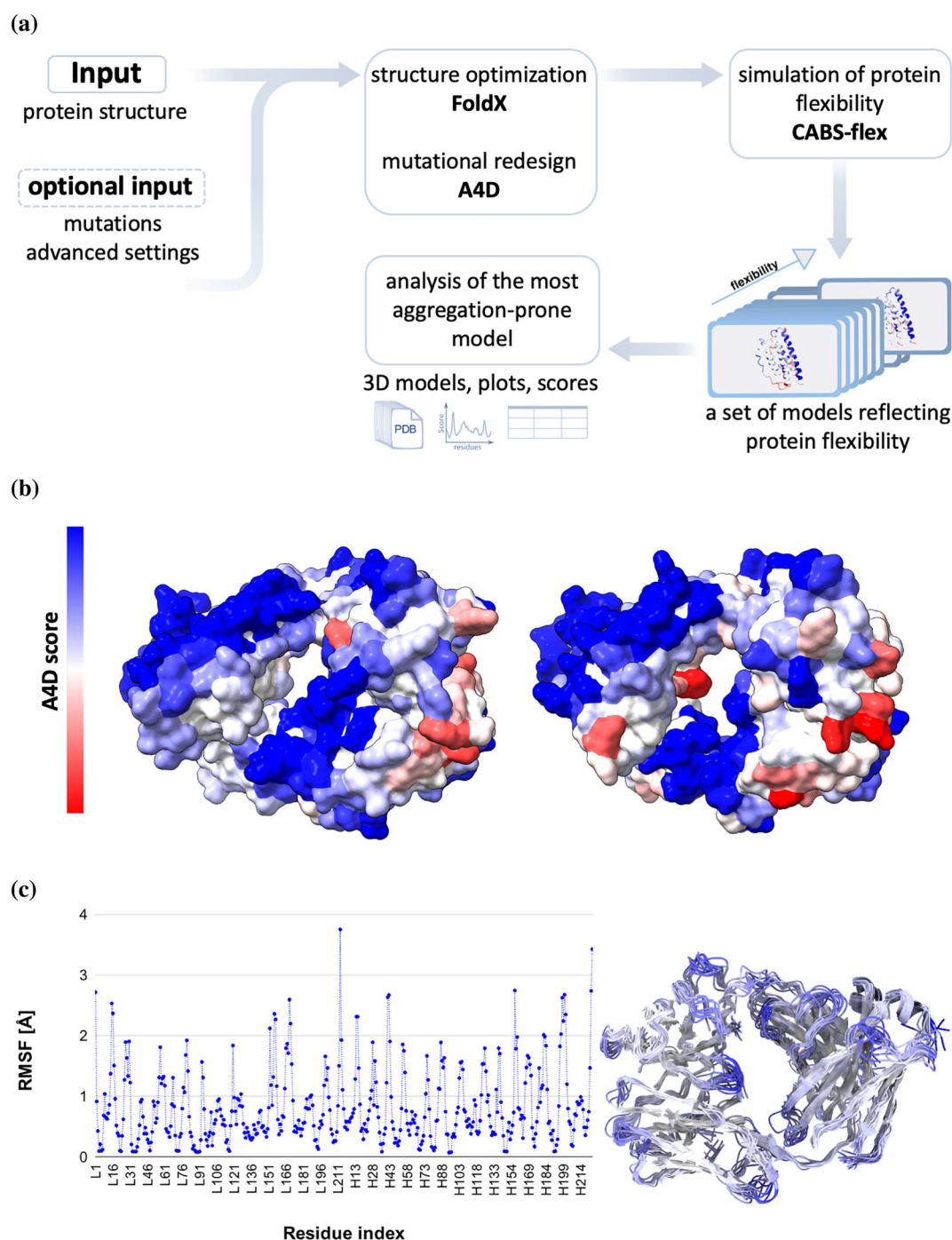


FIGURE 5 Aggrescan4D Dynamic mode. (a) Dynamic mode pipeline. (b) Bevacizumab Fab fragment (PDB ID: 1BJ1) aggregation propensity predictions using A4D Static mode (left side) and Dynamic mode (right side). Residues are colored by A4D scale: Blue for solubility, red for aggregation, white for neutral regions. Dynamic mode reveals aggregation-prone regions that are not evident in Static mode. (c) RMSF (Root Mean Square Fluctuation) plot from the CABS-flex simulation (left) and superposition of models produced by CABS-flex (right).

These functionalities are summarized in Figure 6. By default, both mutation options are disabled. It's noteworthy that automated mutations are not compatible with pH calculations and Dynamic mode (see Figure 1).

3.8 | Privacy

To prioritize privacy, users can select the “Do not show my job on the results page” checkbox, found on the main page's options panel. This feature allows users to control

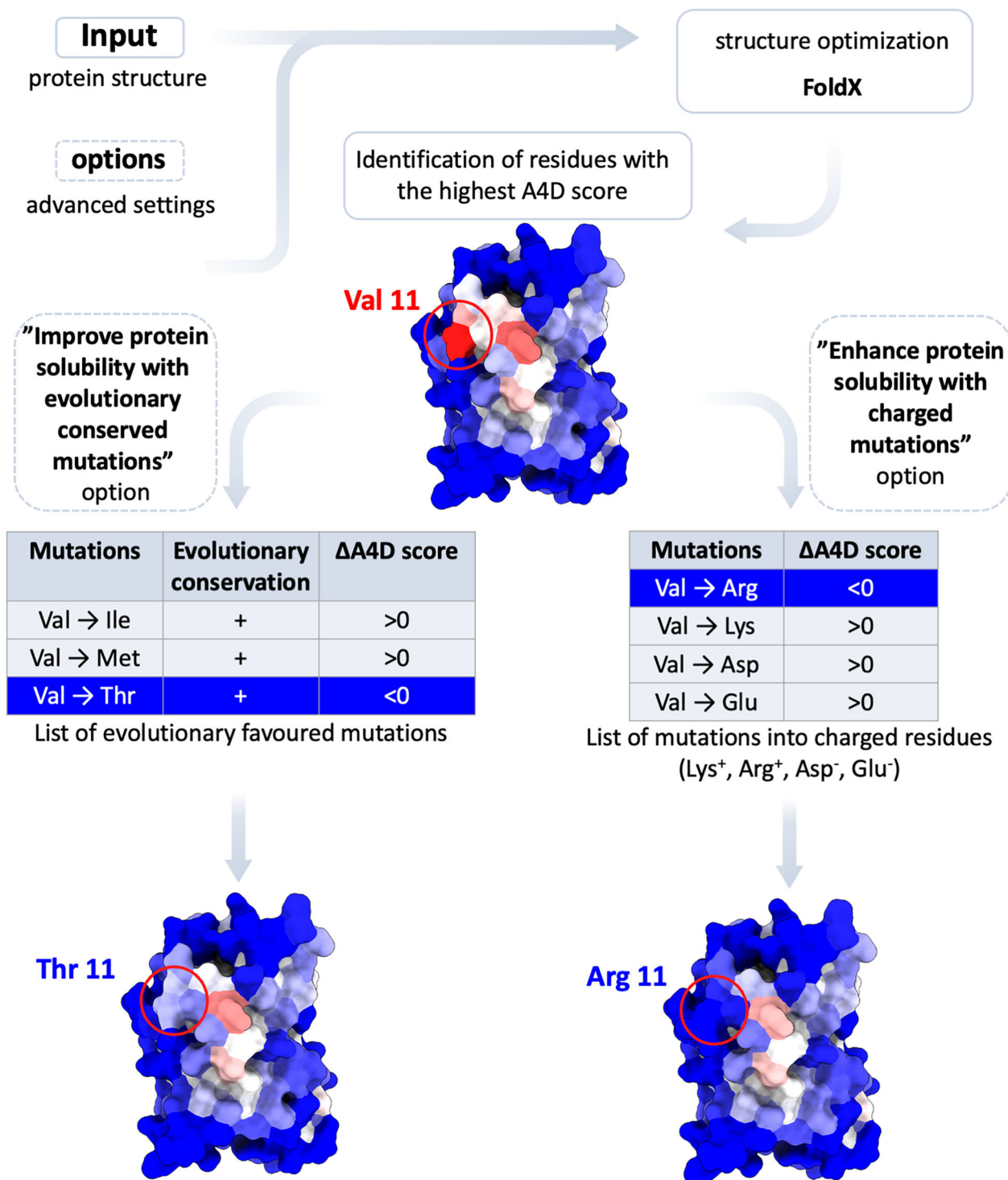


FIGURE 6 Aggrescan4D automated mutations pipeline. In the automated mutations mode, A4D identifies residues with the highest A4D score, subsequently mutating them into positively charged residues (Lysine and Arginine) or negatively charged residues (Aspartic and Glutamic acids). It then calculates A4D scores and stability increments for each variant, providing a ranked list of mutations. Mutations that do not negatively impact stability are prioritized, appearing at the top of the list, while those that affect stability are highlighted in orange (energetic effect >0.5 kcal/mol) and red (energetic effect >1.0 kcal/mol). In the "Improving protein solubility with evolutionary conserved mutations" mode, A4D performs mutations which impact less on protein stability based on protein residue based on PAM250 matrix residue distances.

the visibility of their analyses, which is especially beneficial for handling sensitive data. Furthermore, the standalone version of A3D (outlined in Methods Section 3.11)

allows users to conduct analyses locally on their own machines, granting complete control over the research process. This eliminates the need to upload sensitive data

to a remote server, ensuring maximum confidentiality for researchers.

3.9 | A4D RESTful service

The A4D web server extends its capabilities through a programmatically accessible interface, providing researchers with the capability of integrating aggregation predictions into their computational workflows. RESTful URLs, featured by A4D web server, allow users to initiate new jobs either by uploading a custom PDB/CIF file or with specific PDB/UniProt IDs. RESTful service comprehensively supports the entire A4D web server functionality, enabling users defining project names, receiving email notifications and configuring input parameters. Upon job submission, using cURLs, users can monitor progress, access detailed information and download complete A4D output. The RESTful API can be harnessed through Python or other tools capable of POST requests.

For a comprehensive overview of A4D RESTful service options and usage examples, please refer to ref (Pujols et al., 2021). Additionally, detailed RESTful guides and examples can be found on: <http://biocomp.chem.uw.edu.pl/a4d/tutorial>.

3.10 | Database for model organisms

The introduction of AlphaFold and the accompanying AlphaFold Protein Structure Database (Varadi et al., 2024), which includes over 200 million protein structures, has made a significant impact on structural biology. This development provides access to accurate protein structures, especially for proteins that had not been structurally characterized before, enhancing our understanding and opening new avenues in scientific research. Leveraging the availability of AlphaFold's structures, we initially proposed A3D predictions for the human proteome (Badaczewska-Dawid et al., 2022) followed by predictions for yeast (Garcia-Pardo et al., 2023). Subsequently, we developed the A3D Model Organism Database (A3D-MODB) (Badaczewska-Dawid et al., 2023). This resource contains a vast collection of over 160,000 proteins from the proteomes of 12 commonly studied organisms: *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Danio rerio*, *Escherichia coli*, *Mycoplasma genitalium*, *Mus musculus*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Homo sapiens*, *Rattus norvegicus*, *Drosophila melanogaster*, and SARS-CoV-2. The primary goal of A3D-MODB is to facilitate extensive cross-species analysis related to protein aggregation. It provides

valuable predictions regarding solubility and stability, as well as confidence scores for each residue modeled by AlphaFold (pLDDT). Additionally, it offers topology predictions for membrane proteins, calculated using TOPCONS (Tsirigos et al., 2015).

Each entry in the A3D-MODB database (Badaczewska-Dawid et al., 2023) is linked to the UniProt Knowledgebase (UniProtKB) (UniProt Consortium, 2023) and species-specific databases, providing comprehensive coverage and making it an indispensable resource for researchers in various scientific fields. The A3D-MODB database is freely accessible, requiring no registration, and can be reached at <http://biocomp.chem.uw.edu.pl/A3D2/MODB>. Complementing these capabilities, the A4D web server allows users to input UniProt IDs for automatic retrieval of AlphaFold structures. It also integrates seamlessly with A3D-MODB, providing easy access to nearly 500,000 precalculated jobs. This integration simplifies research processes and enhances access to valuable structural biology data.

3.11 | A3D standalone package

A3D is also available as an independent, self-contained software application (Kuriata, Iglesias, Kurcinski, et al., 2019) designed for local use, ensuring enhanced privacy. This standalone version (Kuriata, Iglesias, Kurcinski, et al., 2019) integrates seamlessly with researchers' workflows, providing complete control over every stage of the process (the standalone version tutorial has been presented in the monograph (Kuriata et al., 2022)). Additionally, its extensional integration with the CABS-flex tool allows for interactive analysis and visualization of models generated by CABS-flex, as well as comprehensive control over CABS-flex simulations and their parameters.

Despite the robust features of A3D standalone (Kuriata, Iglesias, Kurcinski, et al., 2019), it can be easily installed via Python package managers such as pip or Conda, or as a Docker image. This installation flexibility ensures compatibility with Windows, MacOS, and Linux operating systems. The standalone version includes an integrated local server and offers a user-friendly graphical interface that mirrors the A3D web-server (Zambrano et al., 2015) interface.

It is important to note that the current standalone version of A3D operates on Python 2.7. While this is not the latest version of Python, the next iteration of the standalone A4D, is currently under development. A4D standalone will be fully compatible with modern Python releases, ensuring smooth integration and enhanced functionality. Installation guides, documentation, and

usage examples are available in the A3D repository at <https://bitbucket.org/lcbio/aggrescan3d>.

4 | CASE STUDY: EVALUATION OF AGGREGATION PREDICTION METHODS WITH MEDI1912 MONOCLONAL ANTIBODY VARIANTS

Along with A4D (Bárcenas et al., 2024), SolubiS (De Baets et al., 2015) and CamSol (Sormanni et al., 2015) are widely employed tools to understand and rationally design structural aggregation propensity. The three tools offer aggregation propensity reporters and the possibility to inform users with mutational variants predicted to have less propensity to self-association.

Recently, David Brockwell and collaborators developed an experimental pipeline for the *in vivo* development of more soluble proteoforms focusing on highly aggregating MEDI1912 monoclonal antibody which targets human nerve growth factor with picomolar affinity (Ebo et al., 2020) (Figure 7). The authors developed

18 different variants and recorded the full-length IgG aggregation propensity through retention time on high-performance size exclusion chromatography (HP-SEC), a reporter of aggregation-driven monomer loss. Although the described methodology allowed the development of novel variants, the lack of control over the mutation sites caused these less aggregating variants to impact protein functionality (Ebo et al., 2020). This dataset reporting experimental evidence for protein aggregation variants allowed us to evaluate the performance of the A4D, SolubiS, and CamSol structural aggregation prediction methods on protein mutations (antibody structure was a courtesy of the Brockwell lab at the University of Leeds). We categorized the mutants based on their aggregation propensity considered positives (9 variants) or negatives (8 variants) those that increased or decreased the retention time with respect to the midpoint. As reported in Table 2, A4D displays the highest sensitivity, accuracy, and specificity in predicting the effects of mutational variants on experimental aggregation, compared with structural-based state-of-the-art prediction algorithms.

The three tools also differ in the additional possibilities they provide. A4D like SolubiS allow users different

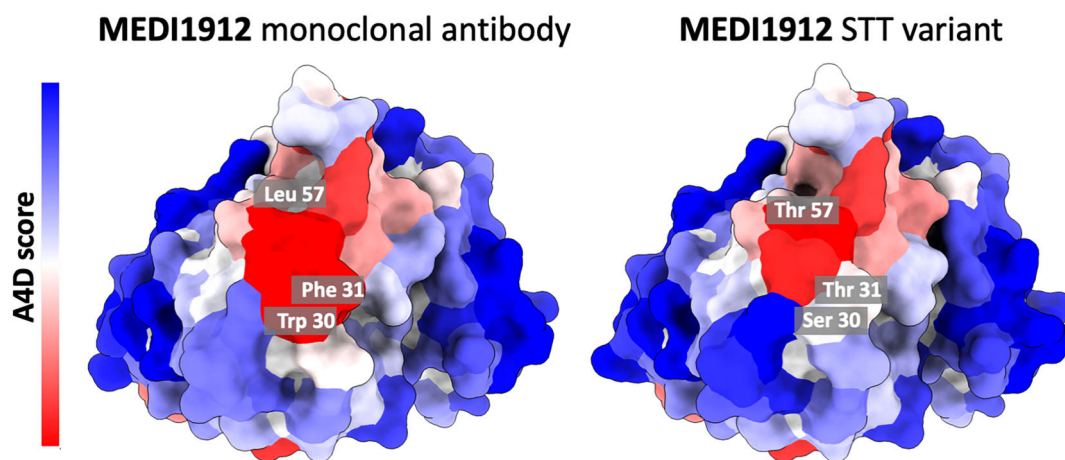


FIGURE 7 Monoclonal antibody MEDI1912 and its more soluble engineered variant STT. Residues are colored by A4D scale: Blue for solubility, red for aggregation, white for neutral regions. MEDI1912 STT differs from wild type protein with three mutations in the complementary complementarity determining region; Trp30Ser, Phe31Thr and Leu57Thr (Dobson et al., 2016).

TABLE 2 Performance of A4D, SolubiS and CamSol for a dataset of 19 mutants of the MEDI1912 monoclonal antibody as reported in ref (Ebo et al., 2020).

	Sensitivity	Specificity	Precision	Accuracy	MCC ^a	FDR ^b
A4D	0.89	0.75	0.80	0.82	0.65	0.20
CamSol	0.78	0.25	0.54	0.53	0.03	0.46
SolubiS	0.89	0.38	0.62	0.65	0.31	0.39

Note: The best performance according to each particular parameter is shown in bold. Mutational impact for A4D was assessed by the maximum score (the best reporter for experimental properties of folded proteins), for SolubiS the Tango score and for CamSol the structurally corrected score.

^aMatthews correlation coefficient.

^bFalse discovery rate.

methods of inputting structures through user-uploaded PDB/CIF files or PDB/UniProt ID without requiring registration. A4D nonetheless also allows for large-scale analysis in an automatic way through its RESTful API. Additionally, A4D allows carrying out a diversity of *in situ* analysis of the results through a potent yet user-friendly user-interface for instance allowing tagging residues in the A4D colored protein structure or calculating impact of mutations from the results page.

5 | CONCLUSIONS

Aggrescan4D (A4D) (Bárceñas et al., 2024) represents a refined evolution of Aggrescan3D (A3D) (Zambrano et al., 2015), as it considers the effect of the pH, an important environmental factor heavily impacting protein aggregation. This is important, since protein aggregation underlies the onset of human diseases, challenges drug development and protein formulation.

A4D (Bárceñas et al., 2024) enhances the functionalities of A3D, maintaining the user-friendly interface that facilitated its wide acceptance within the protein science community. It introduces innovative features like pH-dependent aggregation prediction and an evolutionary-informed automatic mutation protocol. These enhancements aim to improve the design of protein solubility and stability without compromising structural integrity, complemented with A4D's integration with comprehensive databases like the A3D-MODB (Badaczewska-Dawid et al., 2023) and AlphaFold (Jumper et al., 2021).

Beyond its core functions, A3D's application has been extended through its integration into various bioinformatics tools, broadening its impact across different scientific areas. Its incorporation into the Protein Solubilizing Server (PSS) (Tao et al., 2022) aids in addressing the solubilization challenges of structurally complex membrane proteins, utilizing the innovative QTY code-based engineering approach to enhance their water solubility. This integration not only facilitates the design of more effective biopharmaceuticals but also supports basic research into membrane protein structures and functions. Similarly, DE-STRESS (Stam & Wood, 2021) utilizes A3D to rigorously evaluate structural models of proteins, providing researchers with critical insights that guide the engineering of proteins with optimized stability and function. This tool enables a systematic assessment of protein designs, enhancing the predictability and success rate of protein engineering efforts. We expect that A4D's versatility would be soon exploited by the bioinformatics ecosystem beyond aggregation prediction in the broader context of protein design and analysis.

Looking to the future, there are several potential enhancements to the A4D platform that could substantially improve its utility in the design of therapeutic proteins. One key advancement would involve the integration of machine learning algorithms to perform mutational analysis. These algorithms can analyze extensive datasets of known protein mutations, extracting patterns to predict their effects on protein behavior (Pang et al., 2023). Furthermore, the incorporation of empirical aggregation data into these models could refine the predictive capabilities, leveraging real-world experimental results to increase accuracy (Housmans et al., 2023). Updating the A4D database to include recent discoveries on aggregation-prone regions and resistant mutations would also be advantageous. Additionally, expanding the structural dynamics suite to include longer-term dynamics and more extensive conformational changes could enhance simulations and approach them to biologically/industrially relevant scenarios.

These prospective updates would enhance A4D's functionality and expand its application in biotechnology and biomedical research. As A4D continues to evolve, it can play an important role in uncovering new therapeutic strategies and deepening our understanding of protein aggregation-related disorders.

AUTHOR CONTRIBUTIONS

Mateusz Zalewski: Writing – original draft; investigation; visualization; conceptualization; writing – review and editing. **Valentin Iglesias:** Writing – original draft; formal analysis; investigation; visualization; conceptualization; writing – review and editing. **Oriol Bárceñas:** Writing – review and editing; writing – original draft; visualization. **Salvador Ventura:** Conceptualization; writing – review and editing. **Sebastian Kmiecik:** Conceptualization; writing – review and editing; writing – original draft; funding acquisition.

ACKNOWLEDGMENTS

The authors would like to thank Prof. David Brockwell for providing the structural model for MEDI1912 antibody. Mateusz Zalewski has been supported by the National Science Centre, Poland (Sheng 2021/40/Q/NZ2/00078). Valentin Iglesias has been supported by the Polish National Agency for Academic Exchange under the ULAM NAWA Programme (Grant agreement no. BPN/ULM/2023/1/00189/U/00001). Oriol Bárceñas has been supported by the Spanish Ministry of Science and Innovation via a doctoral grant (FPU22/03656). Salvador Ventura has been supported by the Spanish Ministry of Science and Innovation (PID2022-137963OB-I00), Generalitat de Catalunya (2021-SGR-00635 AGAUR) (CERCA Programme) and by ICREA, ICREA-Academia

2020. Sebastian Kmiecik has been supported by the National Science Centre, Poland (Sheng 2021/40/Q/NZ2/00078).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

ORCID

Mateusz Zalewski  <https://orcid.org/0000-0002-3436-8829>

Salvador Ventura  <https://orcid.org/0000-0002-9652-6351>

Sebastian Kmiecik  <https://orcid.org/0000-0001-7623-0935>

REFERENCES

- Arseni D, Chen R, Murzin AG, Peak-Chew SY, Garringer HJ, Newell KL, et al. TDP-43 forms amyloid filaments with a distinct fold in type A FTLTDP. *Nature*. 2023;620:898–903.
- Arseni D, Hasegawa M, Murzin AG, Kametani F, Arai M, Yoshida M, et al. Structure of pathological TDP-43 filaments from ALS with FTLTDP. *Nature*. 2022;601:139–43.
- Badaczewska-Dawid AE, Garcia-Pardo J, Kuriata A, Pujols J, Ventura S, Kmiecik S. A3D database: structure-based predictions of protein aggregation for the human proteome. *Bioinformatics*. 2022;38:3121–3.
- Badaczewska-Dawid AE, Kuriata A, Pintado-Grima C, Garcia-Pardo J, Burdukiewicz M, Iglesias V, et al. A3D model organism database (A3D-MODB): a database for proteome aggregation predictions in model organisms. *Nucleic Acids Res*. 2023; 52:D360–D367.
- Bárceñas O, Kuriata A, Zalewski M, Iglesias V, Pintado-Grima C, Firlik G, et al. Aggrescan4D: structure-informed analysis of pH-dependent protein aggregation. *Nucleic Acids Res*. 2024;1762:427–443.
- Belli M, Ramazzotti M, Chiti F. Prediction of amyloid aggregation in vivo. *EMBO Rep*. 2011;12:657–63.
- Berdyński M, Misztal P, Safranow K, Andersen PM, Morita M, Filippek S, et al. SOD1 mutations associated with amyotrophic lateral sclerosis analysis of variant severity. *Sci Rep*. 2022; 12:103.
- Bhandare VV, Ramaswamy A. The proteinopathy of D169G and K263E mutants at the RNA recognition motif (RRM) domain of tar DNA-binding protein (tdp43) causing neurological disorders: a computational study. *J Biomol Struct Dyn*. 2018;36:1075–93.
- Carija A, Pinheiro F, Iglesias V, Ventura S. Computational assessment of bacterial protein structures indicates a selection against aggregation. *Cells*. 2019;8:856.
- Castillo V, Espargaró A, Gordo V, Vendrell J, Ventura S. Deciphering the role of the thermodynamic and kinetic stabilities of SH3 domains on their aggregation inside bacteria. *Proteomics*. 2010; 10:4172–85.
- Chennamsetty N, Voynov V, Kayser V, Helk B, Trout BL. Design of therapeutic proteins with enhanced stability. *Proc Natl Acad Sci U S A*. 2009;106:11937–42.
- Chiti F, Dobson CM. Amyloid formation by globular proteins under native conditions. *Nat Chem Biol*. 2009;5:15–22.
- Chiti F, Dobson CM. Protein misfolding, amyloid formation, and human disease: a summary of Progress over the last decade. *Annu Rev Biochem*. 2017;86:27–68.
- Conchillo-Solé O, de Groot NS, Avilés FX, Vendrell J, Daura X, Ventura S. AGGRESKAN: a server for the prediction and evaluation of “hot spots” of aggregation in polypeptides. *BMC Bioinformatics*. 2007;8:65.
- De Baets G, Van Durme J, van der Kant R, Schymkowitz J, Rousseau F. Solubis: optimize your protein. *Bioinformatics*. 2015;31:2580–2.
- de Groot NS, Ventura S. Protein aggregation profile of the bacterial cytosol. *PLoS One*. 2010;5:e9383.
- Díaz-Caballero M, Navarro S, Fuentes I, Teixidor F, Ventura S. Minimalist prion-inspired polar self-assembling peptides. *ACS Nano*. 2018;12:5394–407.
- Dobson CL, Devine PWA, Phillips JJ, Higazi DR, Lloyd C, Popovic B, et al. Engineering the surface properties of a human monoclonal antibody prevents self-association and rapid clearance in vivo. *Sci Rep*. 2016;6:38644.
- Ebo JS, Saunders JC, Devine PWA, Gordon AM, Warwick AS, Schiffrin B, et al. An in vivo platform to select and evolve aggregation-resistant proteins. *Nat Commun*. 2020;11:1816.
- Ejaz S, Paracha RZ, Ejaz S, Jamal Z. Antibody designing against IIIabc junction (IIIabc) of HCV IRES through affinity maturation; RNA-antibody docking and interaction analysis. *PLoS One*. 2023;18:e0291213.
- Espargaró A, Castillo V, de Groot NS, Ventura S. The in vivo and in vitro aggregation properties of globular proteins correlate with their conformational stability: the SH3 case. *J Mol Biol*. 2008;378:1116–31.
- Fernandez-Escamilla A-M, Rousseau F, Schymkowitz J, Serrano L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotechnol*. 2004; 22:1302–6.
- Flores-León M, Lázaro DF, Shvachiy L, Krisko A, Outeiro TF. In silico analysis of the aggregation propensity of the SARS-CoV-2 proteome: insight into possible cellular pathologies. *Biochim Biophys Acta Proteins Proteom*. 2021;1869:140693.
- Garcia-Pardo J, Badaczewska-Dawid AE, Pintado-Grima C, Iglesias V, Kuriata A, Kmiecik S, et al. A3DyDB: exploring structural aggregation propensities in the yeast proteome. *Microb Cell Fact*. 2023;22:186.
- Garcia-Pardo J, Ventura S. Cryo-EM structures of functional and pathological amyloid ribonucleoprotein assemblies. *Trends Biochem Sci*. 2024;49:119–33.
- Gil-Garcia M, Baño-Polo M, Varejão N, Jamroz M, Kuriata A, Díaz-Caballero M, et al. Combining structural aggregation propensity and stability predictions to redesign protein solubility. *Mol Pharm*. 2018;15:3846–59.
- Gil-Garcia M, Navarro S, Ventura S. Coiled-coil inspired functional inclusion bodies. *Microb Cell Fact*. 2020;19:117.
- Gokcan H, Isayev O. Prediction of protein pKa with representation learning. *Chem Sci*. 2022;13:2462–74.
- Housmans JAJ, Wu G, Schymkowitz J, Rousseau F. A guide to studying protein aggregation. *FEBS J*. 2023;290:554–83.
- Ishwarlall TZ, Adeleke VT, Maharaj L, Okpeku M, Adeniyi AA, Adeleke MA. Identification of potential candidate vaccines against *Mycobacterium ulcerans* based on the major facilitator superfamily transporter protein. *Front Immunol*. 2022;13:1023558.

- Jamroz M, Kolinski A, Kmiecik S. CABS-flex predictions of protein flexibility compared with NMR ensembles. *Bioinformatics*. 2014;30:2150–4.
- Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596:583–9.
- Katina NS, Balobanov VA, Ilyina NB, Vasiliev VD, Marchenkov VV, Glukhov AS, et al. Sw ApoMb amyloid aggregation under nondenaturing conditions: the role of native structure stability. *Biophys J*. 2017;113:991–1001.
- Knowles TPJ, Mezzenga R. Amyloid fibrils as building blocks for natural and artificial functional materials. *Adv Mater*. 2016;28:6546–61.
- Ko SK, Björkengren G, Berner C, Winter G, Harris P, Peters GHJ. Combining molecular dynamics simulations and biophysical characterization to investigate protein-specific excipient effects on Reteplase during freeze drying. *Pharmaceutics*. 2023;15:1854.
- Kurcinski M, Oleniecki T, Ciemny MP, Kuriata A, Kolinski A, Kmiecik S. CABS-flex standalone: a simulation environment for fast modeling of protein flexibility. *Bioinformatics*. 2019;35:694–5.
- Kuriata A, Badaczewska-Dawid AE, Pujols J, Ventura S, Kmiecik S. Protocols for rational design of protein solubility and aggregation properties using Aggrescan3D standalone. In: Li MS, Kloczkowski A, Cieplak M, Kouza M, editors. *Computer simulations of aggregation of proteins and peptides*. New York, NY: Springer US; 2022. p. 17–40. https://doi.org/10.1007/978-1-0716-1546-1_2
- Kuriata A, Gierut AM, Oleniecki T, Ciemny MP, Kolinski A, Kurcinski M, et al. CABS-flex 2.0: a web server for fast simulations of flexibility of protein structures. *Nucleic Acids Res*. 2018;46:W338–43.
- Kuriata A, Iglesias V, Kurcinski M, Ventura S, Kmiecik S. Aggrescan3D standalone package for structure-based prediction of protein aggregation properties. *Bioinformatics*. 2019;35:3834–5.
- Kuriata A, Iglesias V, Pujols J, Kurcinski M, Kmiecik S, Ventura S. Aggrescan3D (A3D) 2.0: prediction and engineering of protein solubility. *Nucleic Acids Res*. 2019;47:W300–7.
- Leppert A, Chen G, Lama D, Sahin C, Railaite V, Shilkova O, et al. Liquid-liquid phase separation primes spider silk proteins for fiber formation via a conditional sticker domain. *Nano Lett*. 2023;23:5836–41.
- Li R, Wu Z, Wangb Y, Ding L, Wang Y. Role of pH-induced structural change in protein aggregation in foam fractionation of bovine serum albumin. *Biotechnol Rep (Amst)*. 2016;9:46–52.
- Malik S, Zaidi N, Siddiqi MK, Majid N, Masroor A, Salam S, et al. Mechanistic insight into inhibition of amyloid fibrillation of human serum albumin by vildagliptin. *Colloids Surf B Biointerfaces*. 2022;216:112563.
- Nithin C, Fornari RP, Pilla SP, Wroblewski K, Zalewski M, Madaj R, et al. Exploring protein functions from structural flexibility using CABS-flex modeling. *Protein Sci*. 2024;33:e5090.
- Oeller M, Kang R, Bell R, Ausserwöger H, Sormanni P, Vendruscolo M. Sequence-based prediction of pH-dependent protein solubility using CamSol. *Brief Bioinform*. 2023;24:bbad004.
- Pang KT, Yang YS, Zhang W, Ho YS, Sormanni P, Michaels TCT, et al. Understanding and controlling the molecular mechanisms of protein aggregation in mAb therapeutics. *Bio-technol Adv*. 2023;67:108192.
- Petersen SB, Jonson V, Fojan P, Wimmer R, Pedersen S. Sorbitol prevents the self-aggregation of unfolded lysozyme leading to and up to 13 degrees C stabilisation of the folded form. *J Biotechnol*. 2004;114:269–78.
- Petrlova J, Samsudin F, Bond PJ, Schmidtchen A. SARS-CoV-2 spike protein aggregation is triggered by bacterial lipopolysaccharide. *FEBS Lett*. 2022;596:2566–75.
- Pintado C, Santos J, Iglesias V, Ventura S. SolupHred: a server to predict the pH-dependent aggregation of intrinsically disordered proteins. *Bioinformatics*. 2021;37:1602–3.
- Pujols J, Iglesias V, Santos J, Kuriata A, Kmiecik S, Ventura S. A3D 2.0 update for the prediction and optimization of protein solubility. 2021.04.13.439600. 2021 <https://doi.org/10.1101/2021.04.13.439600v1>
- Pulido D, Arranz-Trullén J, Prats-Ejarque G, Velázquez D, Torrent M, Moussaoui M, et al. Insights into the antimicrobial mechanism of action of human RNase6: structural determinants for bacterial cell agglutination and membrane permeation. *Int J Mol Sci*. 2016;17:552.
- Pulido P, Llamas E, Llorente B, Ventura S, Wright LP, Rodríguez-Concepción M. Specific Hsp100 chaperones determine the fate of the first enzyme of the Plastidial isoprenoid pathway for either refolding or degradation by the stromal Clp protease in Arabidopsis. *PLoS Genet*. 2016;12:e1005824.
- Ruff KM, Pappu RV. AlphaFold and implications for intrinsically disordered proteins. *J Mol Biol*. 2021;433:167208.
- Ruiz-Solaní N, Salguero-Linares J, Armengot L, Santos J, Pallarès I, van Midden KP, et al. Arabidopsis metacaspase MC1 localizes in stress granules, clears protein aggregates, and delays senescence. *Plant Cell*. 2023;35:3325–44.
- Sánchez de Groot N, Pallarès I, Avilés FX, Vendrell J, Ventura S. Prediction of “hot spots” of aggregation in disease-linked polypeptides. *BMC Struct Biol*. 2005;5:18.
- Santos J, Iglesias V, Santos-Suárez J, Mangiagalli M, Brocca S, Pallarès I, et al. pH-dependent aggregation in intrinsically disordered proteins is determined by charge and lipophilicity. *Cells*. 2020;9:145.
- Santos J, Iglesias V, Ventura S. Computational prediction and redesign of aberrant protein oligomerization. *Prog Mol Biol Transl Sci*. 2020;169:43–83.
- Santos J, Pujols J, Pallarès I, Iglesias V, Ventura S. Computational prediction of protein aggregation: advances in proteomics, conformation-specific algorithms and biotechnological applications. *Comput Struct Biotechnol J*. 2020;18:1403–13.
- Schmittschmitt JP, Scholtz JM. The role of protein stability, solubility, and net charge in amyloid fibril formation. *Protein Sci*. 2003;12:2374–8.
- Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. *Nucleic Acids Res*. 2005;33:W382–8.
- Shukla AA, Thömmes J. Recent advances in large-scale production of monoclonal antibodies and related proteins. *Trends Biotechnol*. 2010;28:253–61.
- Sormanni P, Aprile FA, Vendruscolo M. The CamSol method of rational design of protein mutants with enhanced solubility. *J Mol Biol*. 2015;427:478–90.

- Stam MJ, Wood CW. DE-STRESS: a user-friendly web application for the evaluation of protein designs. *Protein Eng Des Select*. 2021;34:gzab029.
- Strohäker T, Jung BC, Liou S-H, Fernandez CO, Riedel D, Becker S, et al. Structural heterogeneity of α -synuclein fibrils amplified from patient brain extracts. *Nat Commun*. 2019;10:5535.
- Tam C, Iwasaki W. AlphaCutter: efficient removal of non-globular regions from predicted protein structures. *Proteomics*. 2023;23:e2300176.
- Tao F, Tang H, Zhang S, Li M, Xu P. Enabling QTY server for designing water-soluble α -helical transmembrane proteins. *MBio*. 2022;13:e0360421.
- Tartaglia GG, Vendruscolo M. The Zyggregator method for predicting protein aggregation propensities. *Chem Soc Rev*. 2008;37:1395–401.
- Tsirigos KD, Peters C, Shu N, Käll L, Elofsson A. The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res*. 2015;43:W401–7.
- UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res*. 2023;51:D523–31.
- Varadi M, Bertoni D, Magana P, Paramval U, Pidruchna I, Radhakrishnan M, et al. AlphaFold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res*. 2024;52:D368–75.
- Xia X, Kumru OS, Blaber SI, Middaugh CR, Li L, Ornitz DM, et al. Engineering a cysteine-free form of human fibroblast growth Factor-1 for “second generation” therapeutic application. *J Pharm Sci*. 2016;105:1444–53.
- Zakeri F, Latifi-Navid H, Soheili Z-S, Sadeghi M, Arab SS, Samiei S, et al. Design, construction and in vivo functional assessment of a hinge truncated sFLT01. *Gene Ther*. 2023;30:347–61.
- Zambrano R, Jamroz M, Szczasiuk A, Pujols J, Kmiecik S, Ventura S. AGGRESKAN3D (A3D): server for prediction of aggregation properties of protein structures. *Nucleic Acids Res*. 2015;43:W306–13.
- Žerovnik E. Putative alternative functions of human stefin B (cystatin B): binding to amyloid-beta, membranes, and copper. *J Mol Recognit*. 2017;30:e2562.

How to cite this article: Zalewski M, Iglesias V, Bárcenas O, Ventura S, Kmiecik S. Aggrescan4D: A comprehensive tool for pH-dependent analysis and engineering of protein aggregation propensity. *Protein Science*. 2024;33(10):e5180. <https://doi.org/10.1002/pro.5180>

6. Dynamika molekularna (MD) w biologii strukturalnej

6.1. [A9] Fosforylacja białek P-stalk definiuje stan rybosomu umożliwiającą interakcję z pomocniczymi czynnikami białkowymi

6.1.1. Wprowadzenie do [A9]

W publikacji [A9] skupiliśmy się na biologicznej roli fosforylacji białek szypuły rybosomalnej (P-stalk) -kluczowego kompleksu białkowego zlokalizowanego na podjednostce 60S rybosomu eukariotycznego. Białka te, zbudowane z podjednostki uL10 oraz heterodimerów P1-P2, pełnią funkcje aktywacyjne względem translacyjnych GTPaz (trGTPaz), a także pośredniczą w sygnalizacji stresowej za pośrednictwem kinazy Gcn2, co czyni je elementem łączącym maszynę translacyjną z odpowiedzią komórkową typu ISR (*Integrated Stress Response*).^{131,132}

Badania przedstawione w pracy [A9] były wynikiem szerokiej współpracy kilku wiodących laboratoriów zajmujących się badaniem rybosomów. Nasz zespół odpowiadał za część obliczeniową, opartą na symulacjach dynamiki molekularnej (MD), które umożliwiły szczegółowe scharakteryzowanie zmian konformacyjnych kompleksu P1-P2 w zależności od stanu fosforylacji. Mój wkład w realizację tego projektu obejmował przygotowanie układów obliczeniowych, przeprowadzenie symulacji, analizę trajektorii oraz przygotowanie wizualizacji wyników, a także udział w interpretacji danych i redakcji manuskryptu.

Wcześniejsze badania sugerowały, że białka P-stalk podlegają fosforylacji przez kinazę CK2, jednak nie było jasne, jaki jest ich rzeczywisty status fosforylacji *in vivo*, ani jaka jest fizjologiczna rola tej modyfikacji. W naszej pracy wykazaliśmy, że białka P-stalk w warunkach fizjologicznych występują niemal wyłącznie w postaci ufosforylowanej - zarówno w komórkach drożdży, jak i ssaków, w tym w rybosomach z linii komórkowej MEF i z wątroby myszy. Potwierdziliśmy to metodą Phos-tag/SDS-PAGE, a także niezależną metaanalizą danych proteomicznych z bazy PeptideAtlas, co wskazuje na ewolucyjnie konserwowany i stabilny charakter tej modyfikacji.

W kolejnych etapach pracy zbadaliśmy funkcjonalne skutki tej fosforylacji. Zaskakująco, defosforylowane kompleksy P-stalk wiązały trGTPazy z większym powinowactwem oraz prowadziły do silniejszej stymulacji hydrolizy GTP, ale jednocześnie obniżały dokładność dekodowania. Tym samym, stała fosforylacja P-stalk wydaje się optymalizować działanie rybosomu, zapewniając równowagę między szybkością a wiernością translacji. Co istotne, w warunkach stresu wywołanego tunikamycyną zaobserwowaliśmy, że fosforylacja P-stalk moduluje aktywację kinazy Gcn2. W szczepie pozbawionym możliwości fosforylacji CTD białek P-stalk (mutant SDPM), aktywacja Gcn2 była opóźniona i wykazywała inną dynamikę

niż w szczepie dzikim. Ustaliliśmy również, że fosforylacja wspomaga akumulację Gen2 na frakcji 60S, co może stanowić jeden z mechanizmów umożliwiających szybką adaptację translacyjną do sygnałów stresowych.

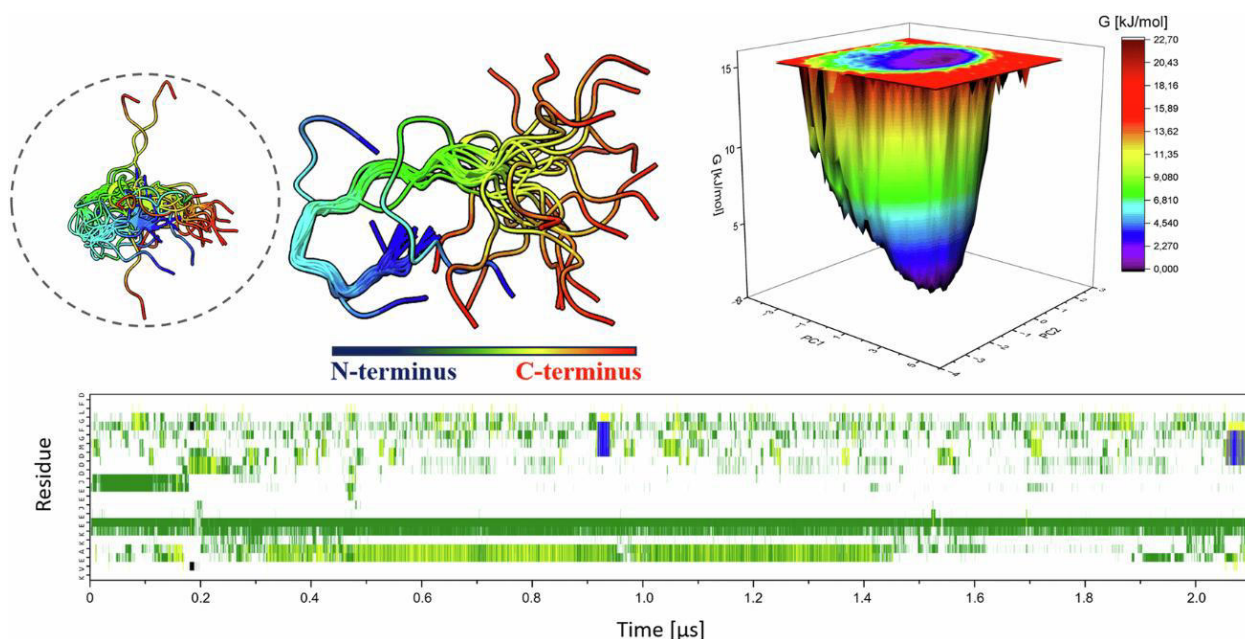
Aby pogłębić zrozumienie mechanistycznego wpływu fosforylacji na strukturalne właściwości białek szypuły, przeprowadziłem serię symulacji dynamiki molekularnej (MD) fragmentów C-końcowych (CTD) białek P1, P2 oraz uL10. Punktem wyjścia do tych analiz były sekwencje 22 C-końcowych aminokwasów, obejmujące regiony zawierające seryny podlegające fosforylacji, zarówno dla białek ludzkich (dwumiejscowa fosforylacja), drożdżowych (jedno miejsce fosforylacji), jak i archealnych (z zamianą seryn na glutaminian w pozycji homologicznej).

Symulacje przeprowadziłem osobno dla wersji ufosforylowanych oraz defosforylowanych każdego peptydu. W przypadku białek eukariotycznych szczególną uwagę poświęciłem ludzkim CTD, gdzie analizowałem wpływ obecności dwóch reszt fosfoseryny oraz ich otoczenia elektrostatycznego. Wykazałem, że fosforylacja prowadzi do ustalenia stabilnej konformacji peptydu, którą określiłem mianem „U-bent” - ze względu na charakterystyczny kształt przypominający literę „U”. Struktura ta wynika z sieci oddziaływań elektrostatycznych między fosfoserynami (pozycje 9 i 12) a dodatnimi resztami lizyny (K1 i K5). Ta konformacja jest utrzymywana przez stabilne mostki solne i skutkuje ograniczoną swobodą konformacyjną w części N-końcowej peptydu. Analiza przebiegu zmian struktury drugorzędowej w czasie symulacji wykazała długotrwałe utrzymywanie się motywów typu bend i turn charakterystycznych dla tej formy, co wskazuje na jej wysoką stabilność.

Dodatkowo przeprowadziłem symulacje CTD rozpoczynając od formy ustrukturyzowanej (α -helisy) - dla oceny wpływu fosforylacji na zdolność do utrzymania struktury drugorzędowej. Stwierdziłem, że fosforylacja w przypadku eukariotycznych CTD przyspiesza przejście do struktury nieuporządkowanej, podczas gdy w przypadku peptydów archealnych konformacja helikalna była utrzymywana znacznie dłużej. Jest to zgodne z wcześniejszymi obserwacjami eksperymentalnymi, sugerującymi zdolność CTD archeonów do formowania α -helis podczas interakcji z partnerami białkowymi.

Aby lepiej wyjaśnić wpływ otoczenia elektrostatycznego na kształtowanie się konformacji, przygotowałem także symulacje mutacyjne z podstawieniem lizyn na alaniny. W przypadku mutacji tych reszt w ludzkich CTD zaobserwowałem zanik stabilnych mostków solnych i brak charakterystycznego zagięcia struktury, co potwierdziło, że obecność dodatnio naładowanych reszt obok fosfoseryn jest kluczowa dla formowania struktury typu U-bent.

Wyniki tych analiz zostały zilustrowane na Rysunku 11, który przedstawia powstawanie charakterystycznej konformacji typu U-bent w przebiegu symulacji dynamiki molekularnej oraz jej stabilizację przez sieć mostków solnych pomiędzy fosfoserynami i resztami lizyny.



Rysunek 11. Konformacja typu U-bent w C-końcowych fragmentach białek P-stalk (na podstawie [A9]). Górny panel: przebieg trajektorii symulacji MD pokazujący formowanie się charakterystycznego zagięcia peptydu w obecności dwóch fosfoseryn. Dolny panel: wizualizacja konformacji stabilizowanej siecią mostków solnych pomiędzy fosfoserynami (S9, S12) a resztami lizyny (K1, K5). Źródło: [A9].

W oparciu o wyniki uzyskane dla fragmentów C-końcowych kontynuowaliśmy te badania na wyższym poziomie złożoności, rozszerzając analizy MD na cały dimer P1-P2. Kolejne prace (obecnie na etapie przygotowania manuskryptu) koncentrują się na tym, jak fosforylacja wpływa na dynamikę i interakcje całego kompleksu. Dwumikrosekundowe symulacje all-atom ujawniły, że obecność fosforylacji nie tylko kształtuje lokalną strukturę CTD, ale także przeorganizowuje krajobraz dynamiczny całego dimeru, w tym orientację motywów GFGLFD i zachowanie regionu N-końcowego P2. Wyniki te pozwalają powiązać odkryty w [A9] mechanizm konformacyjny typu U-bent z globalnymi zmianami w dynamice kompleksu P1-P2, co ma kluczowe znaczenie dla zrozumienia regulacji funkcji białek szypuły przez fosforylację.

Uważam, że jednym z najważniejszych osiągnięć pracy [A9] było pokazanie - dzięki symulacjom dynamiki molekularnej - w jaki sposób fosforylacja zmienia zachowanie nieuporządkowanych fragmentów CTD białek P-stalk. Analizy MD pozwoliły uchwycić powstawanie charakterystycznej i wcześniej nieopisanej konformacji typu U-bent (żartobliwie w naszym zespole nazywanej „Zalewski's bend”), utrzymywanej przez sieć mostków solnych pomiędzy fosfoserynami i resztami lizyny. To właśnie symulacje ujawniły mechanizm, którego nie dało się wywnioskować na podstawie samych danych eksperymentalnych - fosforylacja nie tylko zmienia ładunek i dynamikę tych fragmentów, lecz aktywnie stabilizuje określony kształt przestrzenny, wpływając na sposób interakcji CTD z partnerami rybosomalnymi i regulatorowymi. Bez tego komponentu obliczeniowego obraz roli fosforylacji byłby niepełny i znacznie bardziej spekulatywny.

6.1.2. Artykuł [A9]

Phosphorylation of P-stalk proteins defines the ribosomal state for interaction with auxiliary protein factors

Kamil Filipek^{ID 1,13}, Sandra Blanchet^{ID 2,11,13}, Eliza Molestak¹, Monika Zaciura¹, Colin Chih-Chien Wu^{3,4}, Patrycja Horbowicz-Drożdżał¹, Przemysław Grela^{ID 1}, Mateusz Zalewski^{ID 5}, Sebastian Kmiecik⁵, Alan González-Ibarra¹, Dawid Krokowski¹, Przemysław Latoch^{ID 6}, Agata L Starosta^{ID 6}, Mateusz Mołoń^{ID 7}, Yutian Shao^{ID 1}, Lidia Borkiewicz^{ID 1,12}, Barbara Michalec-Wawiórka¹, Leszek Wawiórka¹, Konrad Kubiński^{ID 8}, Katarzyna Socąła^{ID 9}, Piotr Wlaź^{ID 9}, Kyle W Cunningham^{ID 10}, Rachel Green^{ID 3}, Marina V Rodnina^{ID 2} & Marek Tchórzewski^{ID 1}

Abstract

Ribosomal action is facilitated by the orchestrated work of *trans*-acting factors and ribosomal elements, which are subject to regulatory events, often involving phosphorylation. One such element is the ribosomal P-stalk, which plays a dual function: it activates translational GTPases, which support basic ribosomal functions, and interacts with the Gcn2 kinase, linking the ribosomes to the ISR pathway. We show that P-stalk proteins, which form a pentamer, exist in the cell exclusively in a phosphorylated state at five C-terminal domains (CTDs), ensuring optimal translation (speed and accuracy) and may play a role in the timely regulation of the Gcn2-dependent stress response. Phosphorylation of the CTD induces a structural transition from a collapsed to a coil-like structure, and the CTD gains conformational freedom, allowing specific but transient binding to various protein partners, optimizing the ribosome action. The report reveals a unique feature of the P-stalk proteins, indicating that, unlike most ribosomal proteins, which are regulated by phosphorylation in an on/off manner, the P-stalk proteins exist in a constantly phosphorylated state, which optimizes their interaction with auxiliary factors.

Keywords Ribosomal Proteins; Ribosomal Stalk; Ribosome; Phosphorylation; Gcn2 Kinase

Subject Categories Post-translational Modifications & Proteolysis; Translation & Protein Quality

<https://doi.org/10.1038/s44319-024-00297-1>

Received 22 April 2024; Revised 3 October 2024;

Accepted 14 October 2024

Published online: 28 October 2024

Introduction

Protein synthesis is a highly energy-consuming process where the ribosome acting as a Brownian machine harnesses thermal-structural fluctuations into directed motion on mRNA (Frank and Gonzalez, 2010; Rodnina et al, 1997; Wilson and Noller, 1998). The unidirectional trajectory for the translational apparatus is conferred by GTP hydrolysis catalyzed by auxiliary translational GTPases (trGTPases). The eukaryotic ribosomal P-stalk (and its analog L7/12 in bacteria) plays a pivotal role in the recruitment and activation of trGTPase to the ribosome, which leads to subsequent structural rearrangements, driving mRNA decoding and the ribosome movement on the mRNA (Liljas and Sanyal, 2018; Mohr et al, 2002; Savelsbergh et al, 2000). Recent findings show that, the eukaryotic P-stalk is also responsible for the activation of the stress-related kinase Gcn2, which couples the translational machinery with the Integrated Stress Response (ISR) pathway (Gupta and Hinnebusch, 2023; Harding et al, 2019; Inglis et al, 2019; Masson, 2019). Thus, the ribosome is not only responsible for protein biosynthesis per se, but has also been recognized as a sentinel for stress sensing able to trigger ISR (Kim and Zaher, 2022). In this respect, the ribosomal P-stalk appears as a key player in both

¹Department of Molecular Biology, Institute of Biological Sciences, Maria Curie-Skłodowska University, Lublin, Poland. ²Department of Physical Biochemistry, Max Planck Institute for Multidisciplinary Sciences, Göttingen, Germany. ³Department of Molecular Biology and Genetics, Howard Hughes Medical Institute, Johns Hopkins University School of Medicine, Baltimore, MD, USA. ⁴Section of Translational Control of Gene Expression, RNA Biology Laboratory, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Frederick, MD, USA. ⁵Biological and Chemical Research Center, Faculty of Chemistry, University of Warsaw, Warsaw, Poland. ⁶Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw, Poland. ⁷Institute of Biology, University of Rzeszów, Rzeszów, Poland. ⁸Department of Molecular Biology, Institute of Biological Sciences, John Paul II Catholic University of Lublin, Lublin, Poland. ⁹Department of Animal Physiology and Pharmacology, Institute of Biological Sciences, Maria Curie-Skłodowska University, Lublin, Poland. ¹⁰Department of Biology, Johns Hopkins University, Baltimore, MD, USA. ¹¹Present address: Institute for Integrative Biology of the Cell, I2BC, CEA, CNRS, Université Paris-Saclay, Gif-sur-Yvette, France. ¹²Present address: Department of Biochemistry and Molecular Biology, Medical University of Lublin, Aleje Racławickie 1, 20-059 Lublin, Poland. ¹³These authors contributed equally: Kamil Filipek, Sandra Blanchet. ✉E-mail: kwc@jhu.edu; ragreen@jhmi.edu; rodnina@mpinat.mpg.de; marek.tchorzewski@mail.umcs.pl

ribosomal activities, displaying dual functionality and being involved in the interplay with trGTPases and Gcn2 kinase. The overall architecture of the stalk is similar in all domains of life: the stalk is composed exclusively of ribosomal proteins, the uL11 and the heterocomplex composed of uL10 in a complex with lateral bL12 in bacteria and P1-P2 dimers in eukaryotes. The uL11 and uL10 proteins belong to the conserved ribosomal proteins found in all domains of life and form the base of the stalk by interacting with domain IV of rRNA (Liljas and Sanyal, 2018). In Archaea/Eukarya clades, the P-stalk forms heptamers uL10(aP1)₆ or pentamers uL10(P1-P2)₅, respectively, whereas the bL12-stalk in bacteria has a bimodal distribution between two or three lateral bL12 dimers, uL10(bL12)₄₍₆₎, except for 8 bL12 copies in Cyanobacteria (Davydov et al, 2013). Despite the overall structural similarity, the stalk complexes are not functionally exchangeable between organisms from the three life domains, being compatible only with homologous trGTPases (Mochizuki et al, 2012). Such specificity is conferred by the Eukarya-specific domain II of the uL10 protein (Mochizuki et al, 2012) along with uL11 (Uchiumi et al, 2002), and by the lateral stalk elements bL12 and P1-P2/aP1 for bacterial and archaeal/eukaryotic ribosomes, respectively (Mochizuki et al, 2012). The uL11 forms a separate stalk entity and represents an element shaping the architecture of the stalk (Wawiora et al, 2016). Functionally, uL11, together with uL10, serves as an anchoring point for trGTPases. As shown in bacterial ribosomes, uL11 can form an Arc-like connection with the GTPase domain of trGTPases (Datta et al, 2005), while in the eukaryotic ribosome, uL11 interacts with domain II of uL10, which can stabilize the translation factors on the ribosome (Shao et al, 2022; Yang et al, 2022). On the other hand, the uL10 anchor the lateral stalk proteins bL12 or P1-P2/aP1, which form protruding structural element. Regardless of their origin, the bL12 or P1-P2/aP1 proteins display characteristic topology with the N-terminal domain (NTD) and the C-terminal domain (CTD) connected by a flexible linker region (Bernado et al, 2010; Diaconu et al, 2005; Grela et al, 2007; Lee et al, 2012). Both NTDs and CTDs are evolutionarily conserved within each domain of life, but not across, showing that the lateral proteins should be regarded as analogous (Grela et al, 2008a). The NTDs of lateral stalk proteins form a compact dimeric structure, which is responsible for the interaction with the helical regions of the uL10 scaffold (Diaconu et al, 2005; Grela et al, 2012; Lee et al, 2012). The striking structural difference between bacterial and eukaryotic lateral proteins refers to their CTDs. Even though CTDs in bacteria and eukaryotes are critical for the recruitment of trGTPases, they are neither evolutionarily nor structurally related (Grela et al, 2008a). The bacterial CTD is a globular domain (Diaconu et al, 2005; Leijonmarck and Liljas, 1987) involved in binding of trGTPases and facilitating GTP hydrolysis and stabilization of trGTPase binding to the ribosome (Diaconu et al, 2005); whereas the archaeal/eukaryotic CTD lacks a defined structural organization (Lee et al, 2012), displaying properties of an intrinsically disordered protein (IDP) (Mishra and Hosur, 2015). The eukaryotic CTD function has not been resolved as yet; however, the archaeal CTD is directly involved in binding of various trGTPases, including eIF5B (Murakami et al, 2018), eEF1A (Nomura et al, 2012), eEF2 (Tanzawa et al, 2018), and the ABCE1 recycling factor (Imai et al, 2018). The most unique feature of the P-stalk protein CTDs is the highly conserved amino acid motif EESEESDDDMGFGFLFD (Tchorzewski, 2002), which is almost

identical for uL10, P1, and P2. The sequence is attributed to IDPs, as supported by structural analyses (Grela et al, 2007; Lee et al, 2012); however, as shown for the archaeal CTD, it can also adopt a helical conformation upon trGTPase binding (Murakami et al, 2018; Nomura et al, 2012; Tanzawa et al, 2018).

All eukaryotic P-stalk proteins, uL10, P1, and P2, are specific substrates for the CK2 protein kinase, which can modify two serine residues in the SEESDDD motif in the CTD (Hasler et al, 1991; Zambrano et al, 1997; Zinker and Warner, 1976). While the first serine in the motif is found mostly in mammals, the second serine residue in the motif is highly conserved across all eukaryotes, with only a few notable exceptions, such as pathogenic protists *Trypanosoma* sp. and *Plasmodium* sp., where serine is substituted with glutamic acid. In Archaea, such as extremophilic *Haloarcula morismortui*, the serine residue is substituted with glutamic residue as well (Tchorzewski, 2002) (Fig. 1). Importantly, although the phosphorylation of P-stalk proteins was discovered more than five decades ago, the phospho-status of the stalk proteins and the physiological role of this modification has not been revealed. The only observation was that phosphorylation of the P1/P2 proteins is not required for ribosome binding, but P2 phosphorylation can promote eEF2 binding (Vard et al, 1997). Protein phosphorylation represents an ubiquitous molecular strategy used by the cell to regulate metabolic activity, affecting the conformation of the proteins and, at the same time, influencing numerous aspects of protein functions (Nishi et al, 2011). Reversible phosphorylation and dephosphorylation can drive structural transitions in the IDPs (Iakouchcheva et al, 2004); for example, the CK2 kinase plays a role in regulating the structural transitions of various IDPs and affects the balance between their aggregation and disaggregation as well as their binding to other proteins, DNA, and RNA (Zhang et al, 2020). It is well established that the phosphorylation of proteins involved in all major steps of translation, including translation factors, regulates the performance of the translational machinery and, consequently, the cellular proteome (Simsek and Barna, 2017). The most prominent example is the eIF2α phosphorylation by the Gcn2 kinase upon ISR pathway induction, which blocks the translation of house-keeping mRNAs and stimulates the expression of mRNAs involved in stress adaptation (Masson, 2019). Another example is the phosphorylation of the 4E-BP protein family. In its non-phosphorylated state, 4E-BP binds to eIF4E, and this binding prevents efficient initiation of translation. Conversely, phosphorylation of 4E-BP by mTORC1 kinase disrupts its interaction with eIF4E, thereby stimulating the initiation of translation (Bah et al, 2015). In addition, several ribosomal proteins undergo phosphorylation, including S6, which was discovered as the first ribosomal protein to be phosphorylated (Gressner and Wool, 1974) and its phosphorylation promotes initiation of translation (Bohlen et al, 2021). In turn, phosphorylation of uL13 governs its association with the ribosome (Mazumder et al, 2003).

In this study, we address the long-standing unresolved question of phosphorylation in eukaryotic ribosomal P-stalk proteins to understand the biological role of this post-translational modification. We show that P-stalk proteins in cells predominantly exist in a phosphorylated state. This post-translational modification lowers the interactions with trGTPases, leading to an optimal translation rate, particularly during the decoding step. Furthermore, phosphorylated P-stalk proteins may play an important role in the Gcn2-dependent stress response by facilitating ribosome

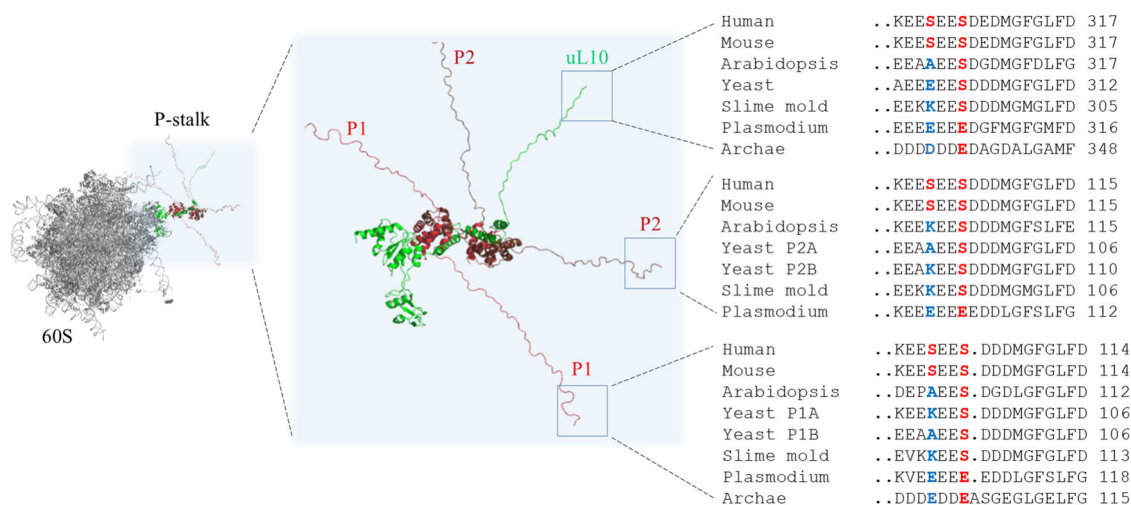


Figure 1. Structural representation of the ribosomal P-stalk with the C-terminal amino acid alignment.

The left panel shows the structural model of the 60S ribosomal subunit from *H. sapiens* (PDB: 4V6X) with the ribosomal P-stalk proteins complex (boxed). The P-stalk proteins are indicated as follows: uL10—green, P1—red, P2—brown, shown as a cartoon representation. Middle panel, zoom-in on the P-stalk pentameric structure; the structure represents compilation of the data for the 60S subunit (PDB: 4V6X for uL10) and P1/P2 dimer (PDB: 4BEH) with the C-terminal motif involved in ligand binding indicated in boxes. The right panel shows amino acid alignments of the C-terminal motifs of the uL10 protein (top), P2 (middle) and P1 (bottom); human - *Homo sapiens* (uL10—accession number #P05388, P1—#P05386, P2—#P05387), mouse - *Mus musculus* (uL10—#P14869, P1—#P47955, P2—#P99027), Arabidopsis, *Arabidopsis thaliana* (uL10—#O04204, P1—#Q8LCW9, P2—#P51407), Yeast, *Saccharomyces cerevisiae* (uL10—#P05317, P1A—#P05318, P1B—#P10622, P2A—#P05319, P2B—#P02400), Slime mold, *Dictyostelium discoideum* (uL10—#P22685, P1—#P22684, P2—#P22683), Plasmodium, *Plasmodium falciparum* (uL10—#Q94660, P1—#Q8IIX0, P2—#O00806), Archaea - *Haloarcula marismortui* (uL10—#P15825, P1—#P15772). The sequences of the P-stalk proteins were retrieved from the UniProt database. The numbers on the right side of the alignment represent total numbers of amino acid residues in the proteins. The amino acid residues marked in red show the position of serine residues that are phosphorylated by the CK2 kinase; in other species one serine can be substituted with alanine, glutamic acid, or lysine, shown in blue.

interaction with Gcn2, which allows timely adaptation to stress conditions. Molecular dynamics simulations demonstrate that the phosphorylation of the P-stalk CTD induces structural transitions, shifting the CTD conformation from a collapsed/globule structure to disordered coils. Consequently, we postulate, that by altering the net charge through CTD phosphorylation within the P-stalk proteins, the CTD gains conformational freedom allowing specific but transient binding to various protein partners.

Results

P-stalk proteins are fully phosphorylated in the steady-state conditions in eukaryotic cells

While it is established that P-stalk proteins are modified by CK2, neither the actual proportion of phospho- and dephospho-forms of the P-stalk proteins nor the physiological importance of such modification are known. We determined the actual phosphorylation status of P-stalk proteins in vivo in yeast *S. cerevisiae* and in a mammalian cell line (MEF cells) using the Phos-tag/SDS-PAGE approach. In this approach, the migration of phosphorylated proteins is changed by the Phos-tag reagent, which facilitates the separation and relative quantification of phosphorylated and dephosphorylated protein populations, followed by Western Blotting analysis with specific antibodies against individual P-stalk proteins. To provide a snapshot of the P-stalk protein phospho-state, we first analyzed bulk cell extracts (Fig. 2A). The protocol developed for rapid cell lysis allowed us to avoid exposure

of the protein to phosphatase/kinase activity. As a control, the protein fractions were treated with either the AP-phosphatase (FastAP, which is active in a broad range of buffers) or the CK2 kinase. The P-stalk proteins from the untreated cell extract migrated more slowly than the AP-treated fractions in both the yeast (left panel) and mammalian (right panel) cells, while the CK2 treatment had no influence on the P-stalk protein mobility. This suggests that the P-stalk proteins in the lysate are phosphorylated. To validate this conclusion, we used a yeast strain (SDPM) where all serine residues within the CTDs were mutated to alanine residues, which abolished the phosphorylation of the P-stalk proteins. The P-stalk proteins from the SDPM strain migrated as a single band at the same level as those from the AP-treated cell extracts, confirming that the lower bands corresponded to the unphosphorylated P-stalk proteins (Fig. 2A). In the next step, we analyzed purified ribosomes from the yeast and mammalian cells. The Phos-tag/SDS-PAGE analysis showed that purified yeast ribosomes constituted a mixed population with phosphorylated and dephosphorylated forms of the P-stalk proteins (Fig. 2B, left panel). This may have represented the heterogeneity of the P-stalk proteins in the cells used for ribosome isolation or the result of dephosphorylation during ribosome purification. In contrast, the ribosomal fraction isolated from the MEF cells displayed a single band of the P-stalk proteins corresponding to the phospho-state (Fig. 2B, right panel).

The experimental data for the human stalk were further supported by the meta-analysis of available high throughput mass spectrometry data from the PeptideAtlas database (Desiere et al, 2006). The human 2024-01 build datasets were used, reference

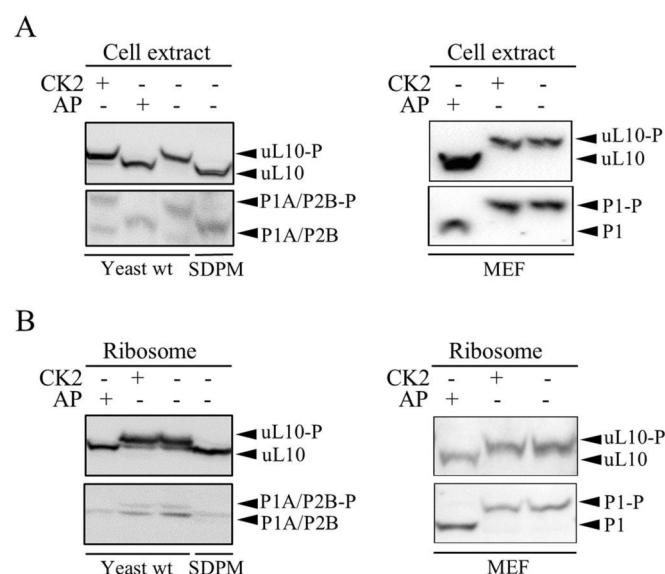


Figure 2. Quantification of P-stalk phosphorylation in yeast and MEF cell line.

Immunoblot analyses of the total cell extracts (A) or total ribosome fractions (B). The electrophoretic mobility of phosphorylated proteins is changed upon treatments by phos-tag. In addition to the yeast WT strain, the SDPM mutant was used that lacks phosphorylation sites. (A) left panel: line 1, yeast cell extract treated with CK2 kinase (CK2+); lane 2, yeast cell extract treated with phosphatase (AP+); lane 3, untreated cell extract from the WT strain; lane 4, extract from SDPM strain. Right panel: lanes 1, 2, and 3, MEF cell extracts treated with: AP, CK2 and untreated, respectively. (B) Total ribosomal fraction from yeast cells. Lanes 1, 2, 3: ribosomes treated with AP, CK2 and untreated, respectively; lane 4, ribosomes from SDPM strain. Right panel: total ribosomes from MEF cell line. Lanes 1, 2 and 3, AP- or CK2-treated and untreated, respectively. The P-stalk proteins were detected using specific antibodies recognizing yeast uL10 and P1A/P2B proteins or human uL10 and P1 proteins. Black arrows show the position of dephospho- and phospho-forms of the P-stalk proteins unmarked and marked with (P), respectively. Source data are available online for this figure.

database: THISP_2024-01-01. Two specific peptides were identified for the human P-stalk proteins: a C-terminal unique peptide KEESSESDDDMGFGFLD shared by P1 and P2 proteins and a unique peptide VEAKEESSESDDDMGFGFLD from uL10. The unbiased MS data showed that P1/P2 and uL10 are mostly double-phosphorylated in the human proteome. Single-site phosphorylation represented about 25% of the population, whereas the dephosphorylated peptide represented only 6% of all phosphopeptide signatures (Fig. EV1, pie-chart, blue part). Similarly, the MS data of the human phosphoproteome suggested that P-stalk proteins exist predominantly (>70%) in the double-phosphorylated state (Fig. EV1, pie-chart, green part). The complementary MS data (PeptideAtlas) for yeast are limited to phospho-proteomic analysis only (YEAST 2022-06 build) (Hollenstein et al, 2021) and do not allow such quantitative estimation.

Having discovered P-stalk proteins in a phosphorylated state within the bulk cytoplasmic ribosomes, our next step was to verify their status in translationally active ribosomes in the polysome fractions. The fractionation of the yeast polysomes showed that the P-stalk proteins were present in the 60S, 80S, and polysome fractions, and the distribution overlapped with the uL23 protein

used as a marker for the 60S subunit. A small fraction of P1/P2 and uL10 proteins was also detected in the ribosome-free fraction (Fig. 3A, left). The analysis of the MEF cells showed a very similar distribution (Fig. 3A, middle panel). We further validated these results using native tissue and examined polysome profiles from mouse liver. As in the yeast and MEF samples, the P-stalk proteins were detected across all polysomal fractions; in this case, no ribosome-free population of the P-stalk protein was detected (Fig. 3A, right panel). The Phos-tag/SDS-PAGE analysis revealed that the P-stalk proteins were mostly present in a phosphorylated form in the yeast, MEF cells, and mouse liver tissue (Fig. 3B). As a final approach to characterize the phospho-state of the P-stalk proteins, the whole cell lysates from the yeast and MEF cells were subjected to AP treatment and, subsequently, the polysomes were resolved by ultracentrifugation. The AP treatment of the isolated cell extracts prior to fractionation did not affect the overall shape of polysomal profiles in the yeast (Fig. 4A, left) or mammalian (Fig. 4A, right) cells, however, the yeast fractions showed a slight decrease in the amount of heavy polysomes. The AP treatment caused a clear shift of P-stalk protein bands toward lower molecular mass, indicating their dephosphorylation by AP (Fig. 4B, lower panel).

Taken together, these data show that, in the steady-state conditions, the P-stalk proteins are predominantly phosphorylated.

Phosphorylation of P-stalk proteins optimizes the activity of translation factors without a significant effect on the rate of protein synthesis

P-stalk proteins play an important role in the recruitment of trGTPases to the ribosome and contribute to GTPase activation, which prompted us to study the effect of phosphorylation on the binding affinity for trGTPases and GTPase activity using in vitro assays. First, we assessed the interactions of the isolated P-stalk with trGTPase eIF5B as a model system using the MicroScale Thermophoresis (MST) approach. The natively assembled yeast pentameric P-stalk complex was obtained by thrombin cleavage and release of the pentameric complex using an established protocol (Grela et al, 2010) (Fig. EV2A). Two forms of P-stalk complexes were prepared, a native one (phosphorylated) and an AP-treated (dephosphorylated) complex (Fig. EV2B,D). The native-MS and Phos-tag analysis showed that the majority of the P-stalk proteins in the native complex were phosphorylated, however, in the case of uL10, mixed fractions were observed, whereas the P1/P2 proteins were mostly phosphorylated. The AP treatment led to complete dephosphorylation of the proteins (Fig. EV2D). For the MST analysis, the P-stalk complex was titrated with increasing concentrations of eIF5B. The saturation curves were obtained for phosphorylated and dephosphorylated P-stalks, which allowed calculation of the dissociation constant, with $K_d = 4.6 \pm 0.8 \mu\text{M}$ and $11.2 \pm 2 \mu\text{M}$ for dephospho- and phospho-P-stalk complexes, respectively (Fig. 5A). Surprisingly, the native phosphorylated P-stalk showed two times lower affinity for eIF5B than the dephosphorylated one, but considering the mixed fraction of uL10, the result may be underestimated.

In the next experiment, we examined the effect of the phosphorylation of P-stalk proteins on the rate of trGTPase-dependent GTP hydrolysis. Two types of ribosomes were used in the analysis: the WT with phosphorylated P-stalk proteins and

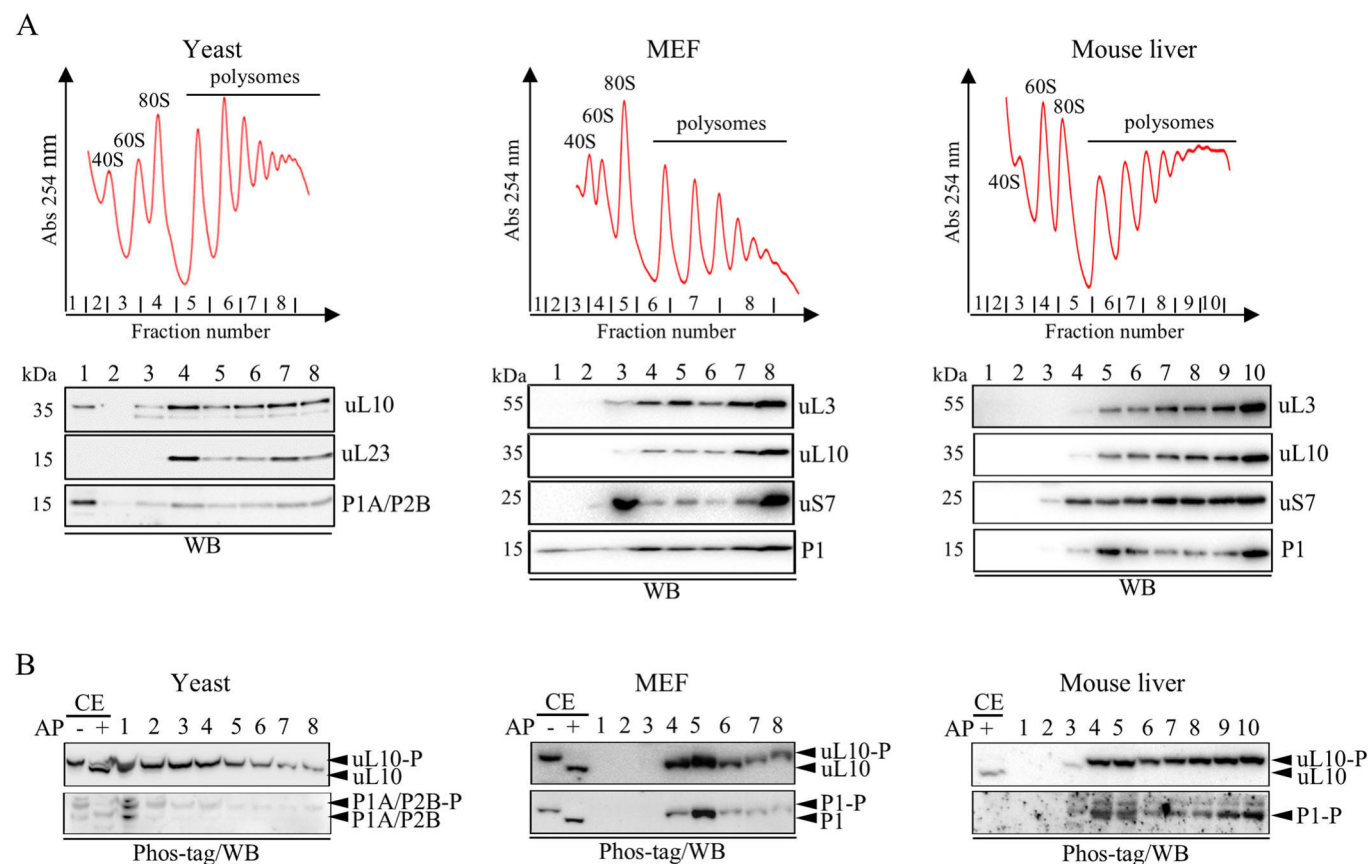


Figure 3. Phosphorylation of P-stalk proteins in 60S subunits, 80S monosomes and polysomes.

(A) Polysome profiles obtained from yeast, MEF cells and mouse liver cells (upper panel). The position of 40S, 60S, 80S and polysomes is indicated. The P-proteins and marker ribosomal proteins were visualized by immunoblotting (lower panel). (B) Phos-tag analysis of protein phosphorylation. Alkaline phosphatase treated (+AP) cell extracts (CE) were used to determine position of phosphorylated (–AP, upper band) and dephosphorylated (+AP, lower band) forms of P-proteins. Bands corresponding to both forms are marked with arrows. Source data are available online for this figure.

ribosomes from the SDPM yeast mutant strain with S to A mutations within all the P-stalk proteins. The phospho-status of the P-stalk proteins was confirmed by the Phos-tag/SDS-PAGE analysis, showing homogenous sample preparation, with two forms of ribosomes: phosphorylated isolated from the WT cells and non-phosphorylated from the SDPM mutant strain (Fig. EV2C). In addition, r-protein content was verified by MS analysis, showing that ribosomes from WT and SDPM mutant strains contained the full set of r-proteins. The purified 80S ribosomes were used to stimulate the GTP hydrolysis of various trGTPases, including eIF5B, eEF1A, and eEF2. The stimulatory effect of the ribosomes was observed for all the trGTPases tested, regardless of the P-stalk configuration. To our surprise, the ribosomes with the dephosphomimicking P-stalk displayed a higher rate of GTP hydrolysis stimulation than the phosphorylated form; in all cases, we observed an approx. 2-fold increase in the level of stimulation, compared to the WT ribosomes (Fig. 5B–D). This can be explained by the higher affinity of the dephosphorylated P-stalk for the trGTPase eIF5B, which is likely to lead to higher GTPase rates at the sub-saturating ribosome concentrations (k_{cat}/K_M conditions) used in these experiments. The higher rate of the GTP hydrolysis in the case of the SDPM mutant compared to the WT ribosomes prompted us to

investigate the effect of P-stalk phosphorylation on the accuracy of decoding, because decoding fidelity is linked to the rate of elongation, and the P-stalk is also involved in maintaining translational fidelity (Wawiora et al, 2017). The dual-luciferase reporter assay was used to evaluate the decoding error rate in vivo as misincorporation events of near-cognate codons placed at different positions in the reporter gene. The analysis showed increased misreading in the SDPM mutant, showing that the lack of P-stalk phosphorylation reduces the accuracy of decoding (Fig. EV3).

Next, we tested the functional effect of the P-stalk phosphorylation on translation. A fully reconstituted translation system was used to study each step of translation elongation in a codon-resolved manner (Ranjan et al, 2021). As above, two types of ribosomes were used: the WT with phosphorylated P-stalk proteins and the ribosomes from the SDPM yeast mutant strain (Fig. EV2C). First, we tested whether the phosphorylation of the P-stalk proteins affects the first round of translation elongation by monitoring the formation of dipeptides (Met-Phe) in real-time. The rate of Met-Phe formation was comparable in both types of the ribosomes tested (Fig. 6A). These data indicate that the presence or absence of the phosphate group on the P-stalk proteins has no significant

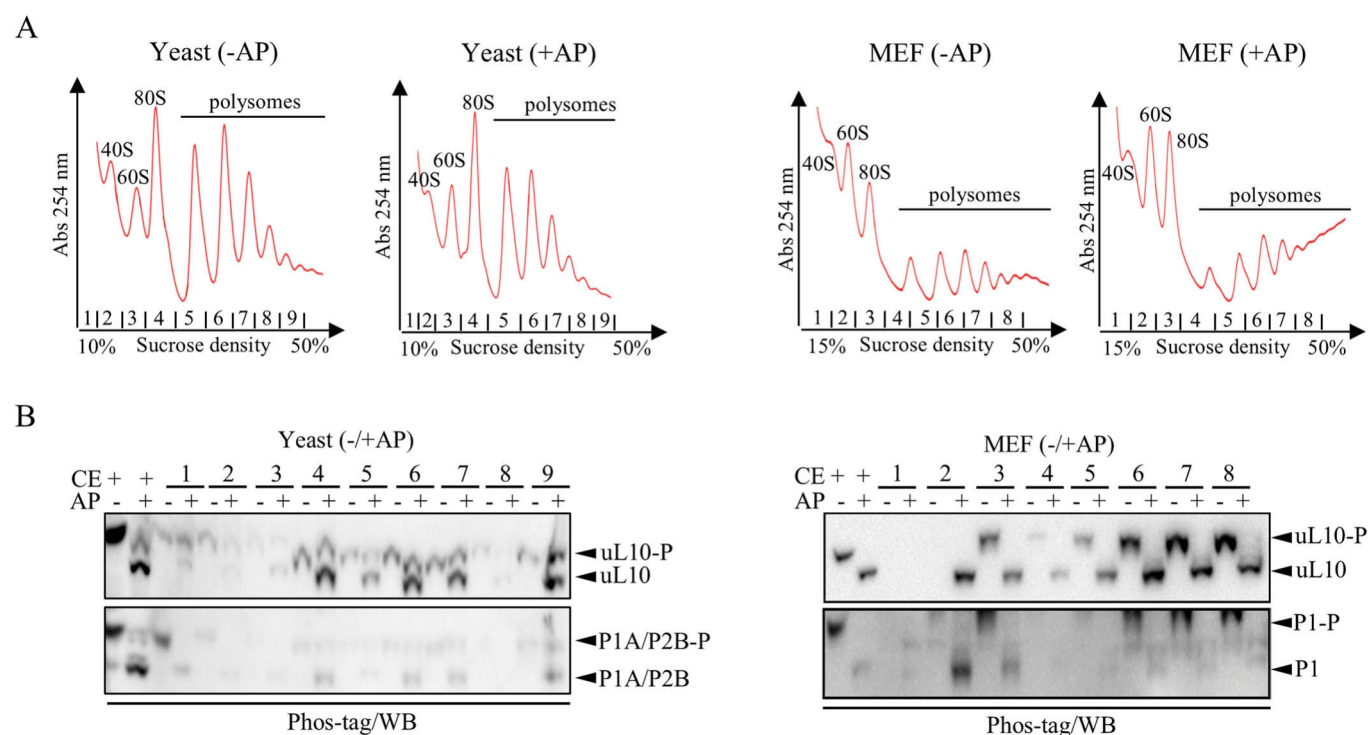


Figure 4. Analysis of the phosphorylation state of P-stalk proteins bound to ribosomes in yeast and MEF fractions.

(A) Polysome profiles obtained from yeast and MEF cells upon alkaline phosphatase (AP) treatment. The position of 40S, 60S, 80S and polysomes is indicated. (B) Phos-tag analysis of the P-proteins within each fraction. Corresponding fractions from non-treated (–AP) and treated (+AP) samples were loaded back-to-back for better comparison. The positions of P-proteins are indicated by arrows. Alkaline phosphatase-treated cell extracts (CE; +AP) were used as control to determine the electrophoretic mobility of dephosphorylated proteins. Source data are available online for this figure.

effect on the first round of decoding and peptide bond formation. To probe the influence of the P-stalk phosphorylation on the second round of translation elongation including a translocation event, the same assay was conducted using a preformed 80S complex (with WT or SDPM ribosomes) carrying a dipeptidyl-tRNA (MetPhe-tRNA^{Phe}) in the A site and deacylated tRNA_i^{Met} in the P site and a Val-tRNA^{Val} ternary complex reading the next GUA codon (Fig. 6B). The results showed that the rate of tripeptide formation (Met-Phe-Val) for both types of ribosomes was also comparable. These data indicate that the presence or absence of the phosphate group on P-stalk proteins has no significant effect on the analyzed steps of translation elongation, including the activity of eEF2 and eEF3.

Overall, these results indicate that phosphorylation decreases the affinity of the P-stalk complex for trGTPase, leading to slower GTP hydrolysis at sub-saturating ligand concentrations; this, in turn, has a positive effect on decoding accuracy. However, there was no significant effect on the rate of translation elongation, including peptide bond formation and translocation.

Phosphorylation of stalk proteins optimizes ribosome function

In the next set of experiments, we investigated the effect of the phospho-state of the P-stalk on translation *in vivo* using the WT and SDPM mutant strains. First, the classical growth rate analysis showed that the absence of P-stalk phosphorylation had a small but

statistically significant negative effect on cell growth (Fig. EV4A). We then analyzed the growth of the cells at the single-cell level, evaluating several aspects of the yeast lifespan, which showed that the SDPM mutant strain had similar parameters to the WT strain, albeit with a reduced reproductive lifespan (Fig. EV4B). The overall translational efficiency *in vivo* was not significantly affected, as the kinetics of the [³⁵S]-methionine incorporation into the newly synthesized polypeptides showed no significant differences between WT and SDPM, supporting the *in vitro* analysis performed at the single-codon resolution (Fig. 7A). Furthermore, the ribosome half-transit times ($t_{1/2}$), representing the average processivity of the ribosome, were similar for WT and SDPM (Fig. 7B). Also, the quantification of the WT and SDPM polysome profiles showed an indistinguishable polysome-to-monomosome ratio (P/M ratio) on both rich (Fig. 7C) and minimal (Appendix Fig. S1) media. Overall, the yeast fitness analyses indicate that the phosphorylation status of the P-stalk has a small but significant effect on cell growth, but not on the overall performance of the translational machinery assessed *in vivo*.

Next, we compared the translational processivity of the WT and SDPM mutant strains on the genome-wide scale by high-resolution ribosome profiling, HrRP (Wu et al, 2019), at the logarithmic growth phase. The ribosome footprint libraries were prepared using a combination of elongation inhibitors, tigecycline and cycloheximide, which allowed distinguishing two distinct ribosome functional states, reflecting the decoding and translocation steps of the elongation cycle. The pre-accommodation state (PreAcc) entails

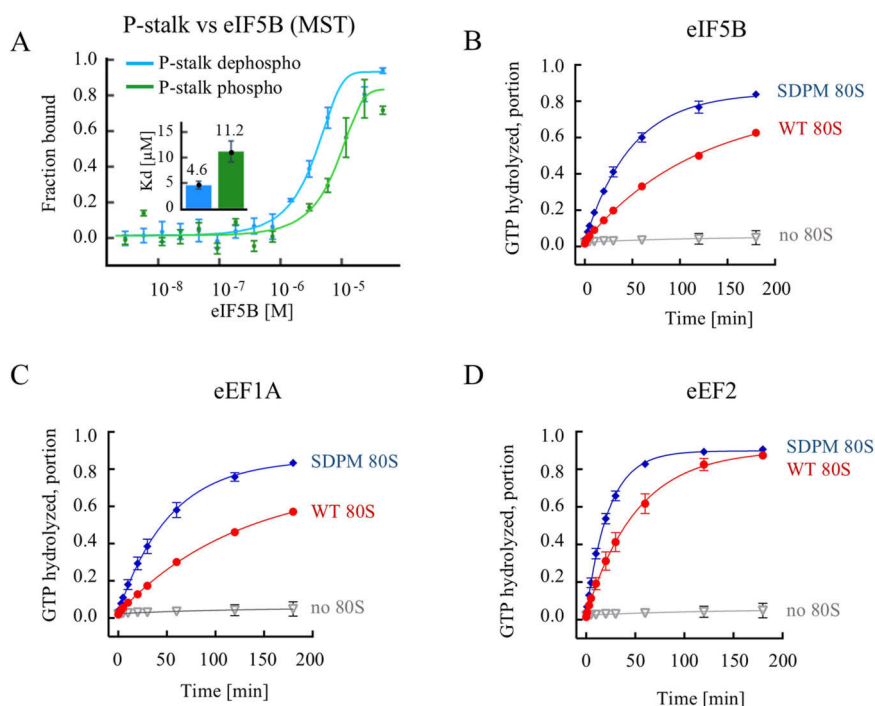


Figure 5. Effect of phosphorylation on the interaction of P-stalk proteins with trGTPases.

(A) Binding affinity of eIF5B to phospho P-stalk (green) or dephospho P-stalk (blue) estimated by MST. The P-stalk complex was titrated with the excess of eIF5B. Data presented as mean $\pm$ SEM of $n = 3$, technical replicates. Inset, equilibrium dissociation constants (K_d): $4.5 \pm 0.8 \mu\text{M}$, $11 \pm 2 \mu\text{M}$, respectively. (B–D) GTPase activity of eIF5B (B), eEF1A (C), and eEF2 (D) in the presence of 80S ribosomes from WT (red) or SDPM (blue) strains or in the absence of the ribosomes (gray). Data are presented as mean $\pm$ SEM of $n = 3$, biological replicates. The average GTPase rates were $0.008 \pm 0.001 \text{ min}^{-1}$ with WT and $0.020 \pm 0.002 \text{ min}^{-1}$ with SDPM ribosomes for eIF5B (B); $0.008 \pm 0.001 \text{ min}^{-1}$ with WT and $0.019 \pm 0.003 \text{ min}^{-1}$ with SDPM ribosomes for eEF1A (C); and $0.019 \pm 0.003 \text{ min}^{-1}$ with WT and $0.044 \pm 0.004 \text{ min}^{-1}$ with SDPM ribosomes for eEF2 (D). Source data are available online for this figure.

ribosomes with an open A site, which provides information on the decoding process, whereas the pre-translocation state (PreTrans) corresponds to ribosomes with an occupied A site before tRNA translocation is completed. The ribosome-protected footprints of the two states differ in the ribosome-protected footprint size (RPF), i.e., 21 nt RPFs corresponding to PreAcc and 28 nt RPFs corresponding to PreTrans states (Wu et al, 2019). The initial correlation analysis using scatter plots of RNAseq and RIBOseq data showed no substantial changes in the transcriptome and translational efficiency, showed no changes either (Fig. EV5C). To characterize the ribosome-specific processivity, we analyzed the distribution of the 21 nt and 28 nt RPFs describing two discrete functional states of elongating ribosomes. The analysis showed a similar bimodal distribution of the 21 and 28 nt RPFs in WT and SDPM, indicating that, in the steady-state conditions, translation elongation in the SDPM mutant did not differ from that in WT (Fig. 8A, left panel), and the relative abundance of the 21 nt and 28 nt RPFs was similar in both strains. Thus, our results indicate that the lack of P-stalk phosphorylation does not cause any significant functional perturbations for elongating ribosomes. Additionally, the codon-specific ribosome occupancies for the 21 nt RPFs were correlated between the WT and SDPM samples (Fig. 8B, left panel); the same behavior was also observed for the 28 nt RPFs (Fig. EV5D, left panels). Once

again, all these data support the idea that the phosphorylation of P-stalk proteins is dispensable for translational machinery performance in optimal conditions.

To gain further insight, we used HrRP to analyze the distribution of two RPFs, 21 and 28 nt, to determine the two functional states of elongating ribosomes following induction of ER stress by tunicamycin (Tun). Tun was used as a general stressor inducing proteotoxic stress, which is considered to occur frequently in cells upon various environmental insults. Tun evokes several types of cellular reactions in the cell, especially ER-stress, inducing the UPR-response (Walter and Ron, 2011), a phenomenon found in natural metabolic transitions in yeast, such as the diauxic shift from the fermentative to mitochondrial respiration phase of growth, also driven by Gcn2 (Patil et al, 2004; Tran et al, 2019). Several time points were used for the analysis because, as we have shown, Tun exerts an immediate effect, with HAC1 mRNA splicing already occurring after 15 min of treatment (Appendix Fig. S2). Changes in the abundance of RPFs in the presence of Tun are regarded as stress markers; in particular, the reduction of the 21 nt RPFs is induced by environmental insults, including osmotic or oxidative stresses (Wu et al, 2019). In contrast to the steady-state conditions, we found reduced levels of the 21 nt RPFs upon Tun-induced stress in the WT cells after 15 min of treatment (Fig. 8A, middle panel). In contrast, the SDPM strain was insensitive to the short Tun treatment, as the ratio between the 21 nt and 28 nt RPFs was similar to that in the untreated sample. After 60 min of the Tun

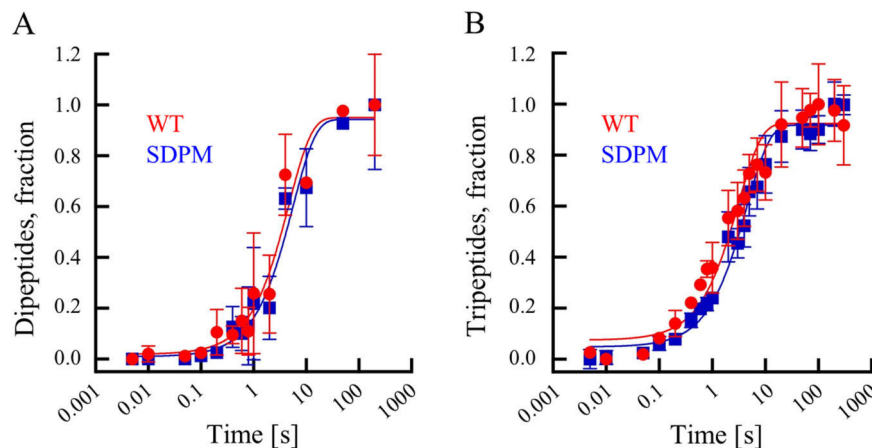


Figure 6. Translation elongation on ribosomes from the WT and SDPM strains.

(A) Dipeptide (Met-Phe) formation monitored upon rapidly mixing initiation complexes (80S IC) using ribosomes from the WT (red line) and SDPM (blue line) yeast cells, with ternary complexes eEF1A-GTP- ^{14}C Phe-tRNA^{Phe} in a quench-flow apparatus. The extent of peptide formation was analyzed by HPLC and radioactivity counting. Data are normalized to Met-Phe formation with the maximum value in the dataset set to 1 and presented as mean $\pm$ SEM of $n = 3$ biological replicates; the average rate is $0.23 \pm 0.14 \text{ s}^{-1}$ and $0.19 \pm 0.10 \text{ s}^{-1}$ for WT and SDPM, respectively. (B) Tripeptides (Met-Phe-Val) formation monitored upon rapidly mixing 80S ribosomes from WT and SDPM cells carrying MetPhe-tRNA^{Phe} (80S 2C) with ternary complexes eEF1A-GTP- ^{14}C Phe-tRNA^{Phe}. Data are presented as mean $\pm$ SEM of $n = 3$, biological replicates; the average rate is $0.34 \pm 0.13 \text{ s}^{-1}$ and $0.22 \pm 0.05 \text{ s}^{-1}$ for WT and SDPM, respectively. Source data are available online for this figure.

treatment, the 21 nt to 28 nt RPF ratio was reduced in both the WT and SDPM strains, which indicates that the ribosome response to ER-stress is delayed, but not abolished, by the lack of P-stalk phosphorylation.

Next, to evaluate whether the change in the abundance of the 21 nt RPF is related to specific ribosomal aberrations during elongation, the correlation of codon-specific occupancy for the 21 and 28 nt RPFs was examined. As already shown, the codon-specific occupancy can be a very sensitive measure of ribosomal processivity at the elongation steps, shown as a loss of correlation, with enrichment in several codons, such as proline codons, especially shown for 21 nt RPFs (Wu et al, 2019). In the case of WT and SDPM strains, under control conditions, the codon-specific occupancy showed no enrichment at specific codons (Fig. 8B, left panel). While, after the 15-min Tun treatment, a loss of correlation was recorded, and the occupancies at three codons encoding arginine (CGA, CGG) and proline (CCG) were increased in the WT strain, indicating altered decoding of these codons (Fig. 8B, middle panel). In turn, no effect was observed in the SDPM mutant strain after the 15-min Tun treatment (Fig. 8B, middle panel), but a loss of correlation was observed after 60 min (Fig. 8B, right). After the 60-min treatment, the decoding of the CGA, CGA, and CCG codons was defective in both the WT and SDPM cells. In addition, the analysis of the 28 nt RPFs also revealed an impact on the translocation at the CCG, CGG, and CCC codons (Fig. EV5D), although this effect was more pronounced after 60 min, with a delay in the response in the case of the SDPM strain, as was the case for the 21 nt RPFs.

Taken together, these data show that, although in conditions of normal growth the overall rate of translation is not significantly affected by the phosphorylation of the P-stalk proteins, upon exposure to ER-stress, the ribosome starts to pause at proline and arginine codons, and this response is delayed in SDPM mutant cells. Thus, the phosphorylation of the P-stalk protein plays a role

in stress response by providing a timely translational response to environmental cues.

Phosphorylation of P-stalk proteins affects Gcn2 activation in ER stress conditions

P-stalk proteins not only recruit trGTPases but also play a role in Gcn2 kinase activation, linking translation to the ISR pathway (Gupta and Hinnebusch, 2023; Inglis et al, 2019; Kim and Zaher, 2022; Masson, 2019). To test the effect of P-stalk phosphorylation on the activation of the ISR pathway, we monitored the phosphorylation status of eIF2 α , the target of the Gcn2 kinase, in the WT and SDPM strains upon the Tun treatment. In the WT strain, the eIF2 α phosphorylation increased up to 2 h of the Tun treatment and then decreased (Fig. 9A). In contrast, the SDPM strain showed a markedly altered temporal response, with the eIF2 α phosphorylation constantly increasing over 8 h (Fig. 9A). In yeast, eIF2 α is thought to represent the only substrate of Gcn2 kinase (Dever et al, 1992), but the kinase can be activated via different mechanisms, e.g., by direct interaction with deacylated tRNA that accumulates upon amino acid starvation (Dong et al, 2000; Wek et al, 1995) or by ribosomes carrying uncharged tRNAs in the A site (Sattlegger and Hinnebusch, 2000). Alternatively, Gcn2 can be activated as a result of ribosome stalling independent of the presence of deacylated tRNA (Gupta and Hinnebusch, 2023; Wu et al, 2020). It was suggested, that in yeast cells, Gcn2 is associated with 60S subunits and polysomes (Dong et al, 2000; Ramirez et al, 1991; Zhu and Wek, 1998) and may accumulate on the 60S subunits in stress conditions (Ramirez et al, 1991). To test whether P-stalk phosphorylation modulates this response, we tested the Gcn2 binding to the ribosomes upon the Tun treatment. A 2-h time point was taken for the analysis, where we detected the highest eIF2 α phosphorylation in the WT strain. The fractions from untreated (–Tun) and treated (+Tun) samples were loaded side by

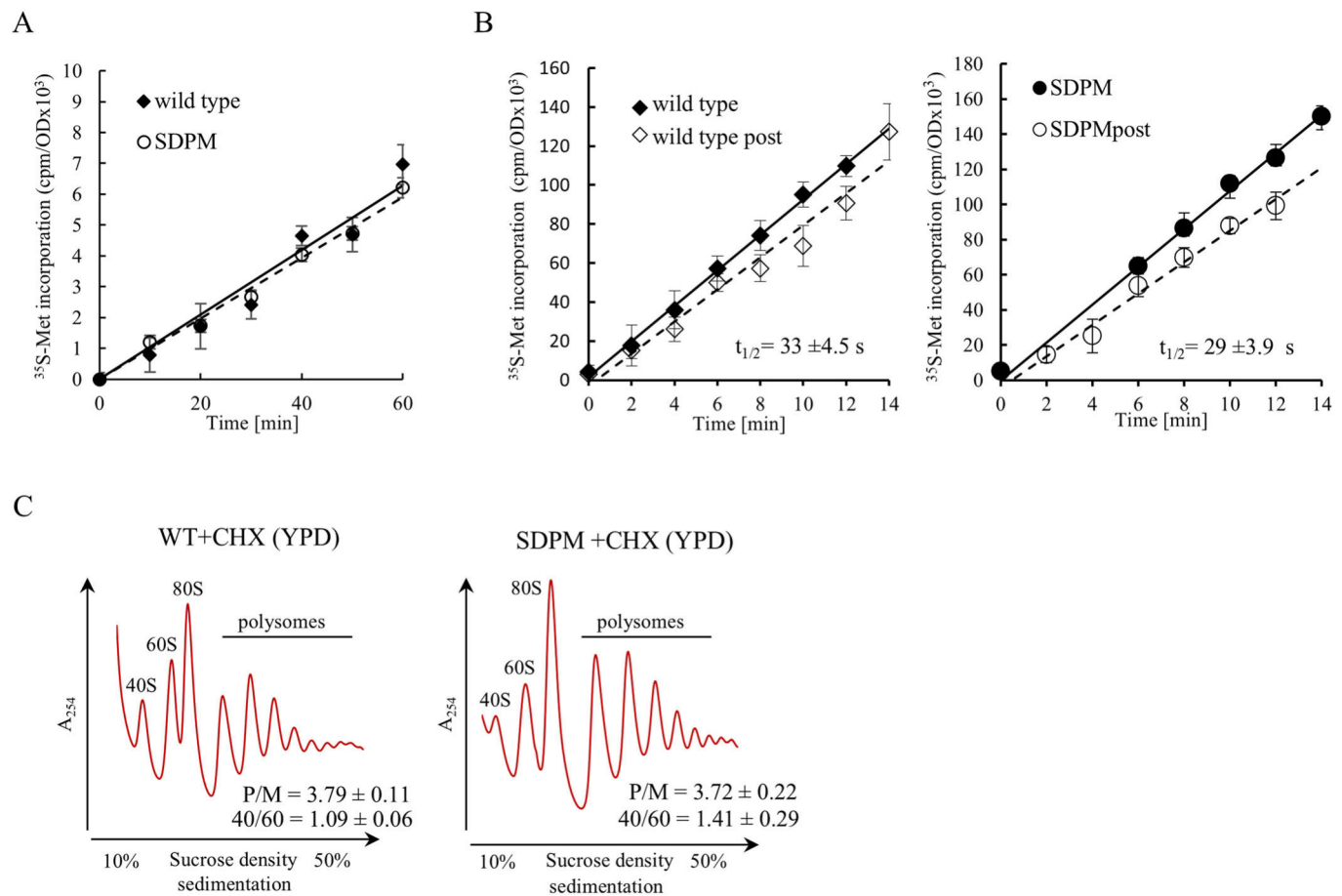


Figure 7. Translational fitness of WT and SDPM yeast strains in vivo.

(A) Kinetics of [³⁵S]methionine incorporation into newly synthesized polypeptides for wild type (◆) and SDPM (○) cells. Error bars represent $\pm$ SD, $n = 3$, technical replicates. (B) Ribosome half-transit times, $t_{1/2}$. The incorporation of [³⁵S]methionine into total proteins (nascent and completed) (filled symbols) and completed proteins (open symbols) is shown for each strain. The half-transit time was measured as the displacement between two lines by linear regression analysis and is shown in each panel as an inset. The $t_{1/2}$ values are means $\pm$ SD, $n = 3$, technical replicates. (C) Polysome profiles of WT and SDPM strains after CHX treatment. The polysome-to-monomosome (P/M) ratio was calculated by dividing the area of the first four polysomal peaks by the area of the peak for the 80S monosome; P/M value is presented as means $\pm$ SD, $n = 3$, technical replicates. The positions of individual ribosomal subunits are indicated. Source data are available online for this figure.

side for better comparison. Gcn2 was probed with anti-Gcn2 antibodies (Fig. 9B). The analysis showed that upon the Tun treatment, Gcn2 accumulated on the 60S fraction in the WT strain, whereas this effect was not prominently observed in the SDPM strain (Fig. 9B).

These experiments show that the phosphorylation of P-stalk proteins may play an important role in the timely activation of Gcn2 kinase. The timing of Gcn2 activation is modulated by the possible accumulation of Gcn2 on the 60S subunit, depending on the phosphorylation of the P-stalk, with more rapid activation and adaptation of cells correlating with the phosphorylated P-stalk.

Structural dynamics of P-stalk protein C-termini is affected by phosphorylation

The CTDs of P-stalk proteins, i.e., the parts proposed to interact with trGTPases, were shown to have unstructured properties in the solution (Lee et al, 2012; Mishra and Hosur, 2015); however, the homologous CTD of the archaeal P1-protein is able to form an α -

helix upon trGTPase binding (Murakami et al, 2018; Nomura et al, 2012; Tanzawa et al, 2018). To understand the effect of phosphorylation on the CTD structure, we performed molecular dynamics simulations (MD) using CTD sequences comprising the C-terminal 22 amino acid residues (Appendix Fig. S3), starting the simulations with a linear chain or an α -helix. We used sequences from human P-stalk proteins, which are doubly phosphorylated, yeast *Saccharomyces cerevisiae*, which have only one phosphorylation site, and an Archaeon *Haloarcula morismortui*, where glutamic acid replaces the invariant serine residues present in eukaryotic sequences (Appendix Fig. S3).

In the CTD peptides, phosphoserine residues are located at positions 9 (human specific) and 12 (conserved in all eukaryotes), framed by lysine and glutamic acid residues (Appendix Fig. S3). The MD analysis revealed that phosphorylation of the serine residues in human P1/P2 and yeast P1 proteins induces a conformational constraint leading to a bent conformation at the N-terminal part of the peptide (Fig. 10A,B, left panel) and (Appendix Figs. S4 and S7), while P1 from Archaeon protein

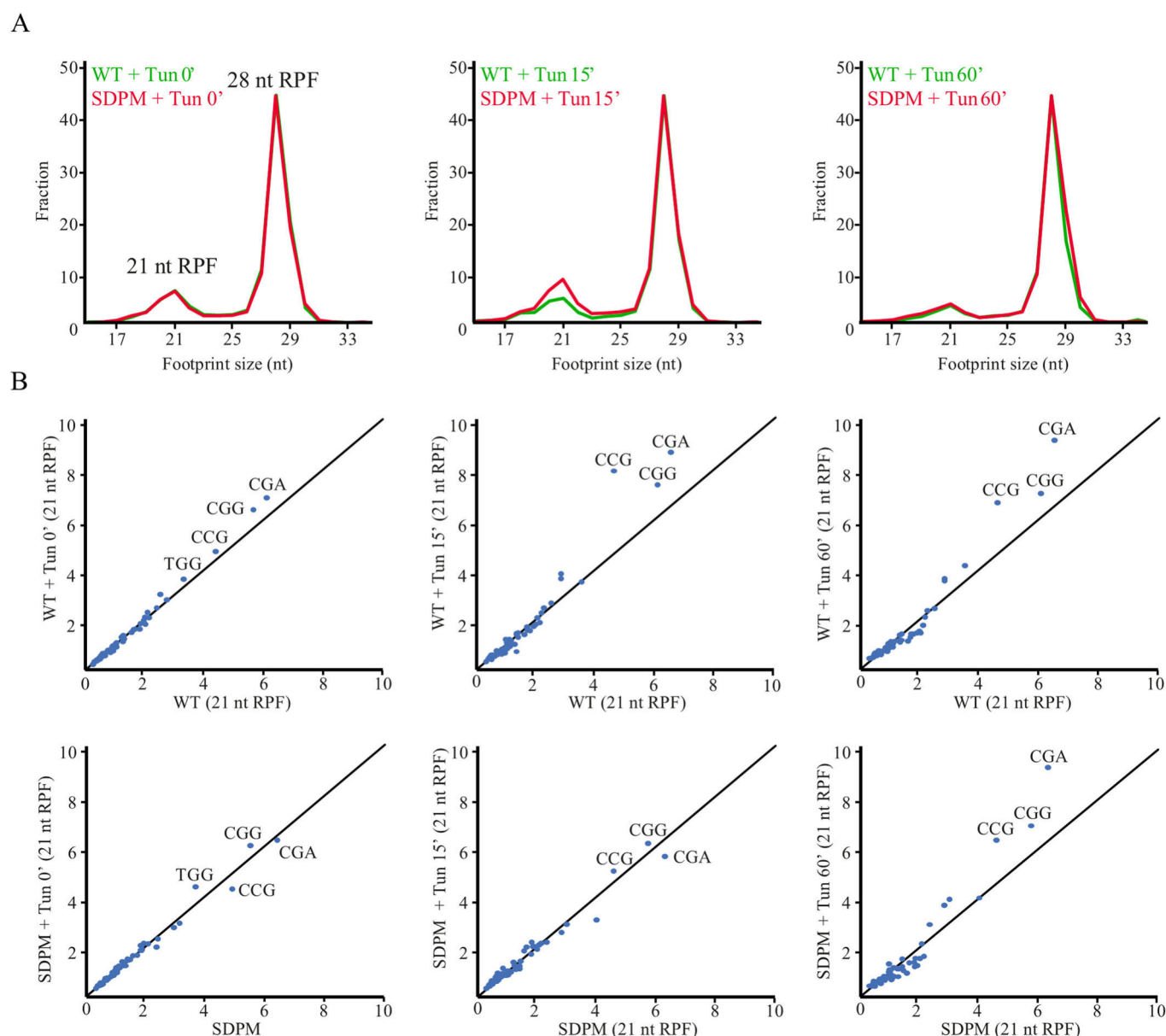


Figure 8. The relative abundance of 21 nt and 28 nt RPF and specific codon occupancies within 21 nt RPF in steady-state conditions and upon Tun treatment.

(A) The relative abundance of 21 nt and 28 nt RPF in the course of Tun treatment of WT (green) and SDPM (red). (B) The correlation of specific codon occupancies within 21 nt RPF, corresponding to part A. Specific overrepresented codons CGA, CGG for arginine and CCG for proline are indicated. Source data are available online for this figure.

showed relaxed, extended conformation (Fig. 10C). This bent conformation arises from the network of electrostatic interactions. Specifically, phospho-serine #9 and #12 form salt bridges with positively charged lysine residues, Lys5 and Lys1, respectively, which stabilizes the U-bent shape (Appendix Fig. S4, inset). The conformation is further stabilized by additional salt bridges, e.g., between Lys5 and Glu3 and between Lys1 and Asp14. The formation of stable salt bridges between phosphorylated serine and neighboring positively charged residues restricts the conformational freedom and decreased structural heterogeneity at the N-terminal part of the analyzed eukaryotic peptides. Consequently, the applied structural constraints lead to a structural transition

from the collapsed/globule-to-coil extended structure (Fig. 10A,B, left panels). Analyzing secondary structure transitions over time during the MD simulations, as shown in the bottom panel of Fig. 10A,B, especially for human proteins, shows that the U-bent conformation is evident, as bends (marked in green) and turns (marked in yellow) at the N-terminal part of the peptide, which persist almost throughout the entire simulation. This persistence indicates that the U-bent conformation is highly stable and suggests that the phosphorylated peptide adopts a consistent and reproducible structural motif. This stability and the presence of these bends imply that phosphorylation induces a robust conformational constraint, promoting a uniform structural state that minimizes

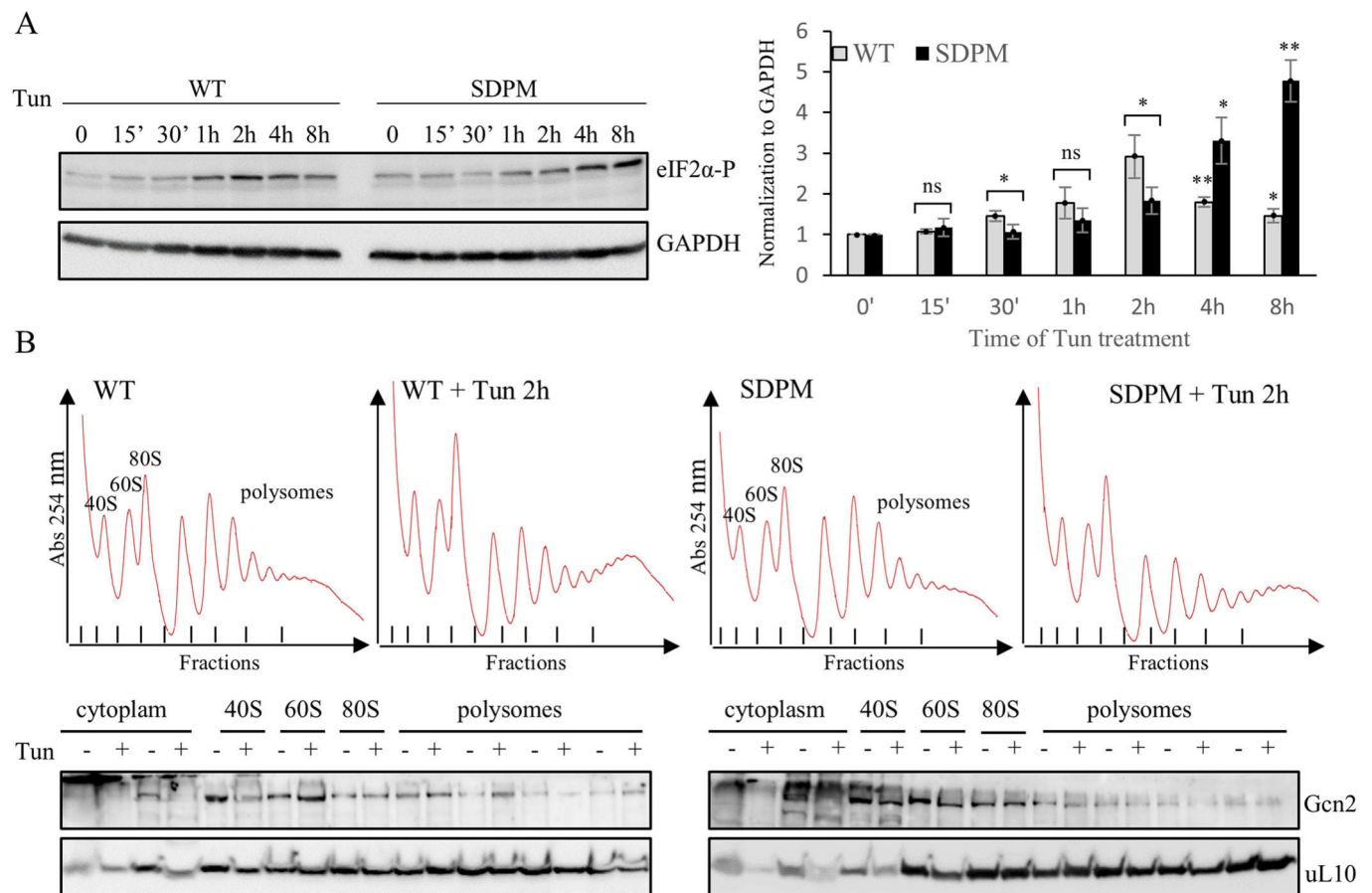


Figure 9. Ribosome-dependent eIF2α phosphorylation and Gcn2 association with the ribosomes.

(A) eIF2α phosphorylation after Tun-induced ER-stress probed with specific anti-phospho-eIF2α antibodies; GAPDH was used as loading control. Left panel, time course of phosphorylation. Right panel, quantification of eIF2α phosphorylation relative to GAPDH loading control, normalized to non-treated cells. Error bars represent SD * $p < 0.05$; ** $p < 0.001$, determined by Student's *t*-test, $n = 4$, biological replicates. (B) Gcn2 binding to the ribosomes. Top panel, polysome profiles from non-treated (–Tun) and Tun-treated (+Tun) (2.5 μg/ml of tunicamycin for 2 h) WT and SDPM yeast cells. Lower panel, Western blot analysis with antibodies against Gcn2 kinase and uL10 ribosomal protein. Fractions from control and Tun exposed cells are marked with –/+ . Source data are available online for this figure.

conformational variability. This structure corresponds to a new global minimum, leading to a significantly less rugged conformational landscape and a narrower and simpler peptide free energy landscape (FEL) (Fig. 10A,B, right panel). This stabilizing effect and structural transition is also observed in the peptides of human P2 and uL10 (Appendix Figs. S4 and 5), having similar changes in FEL minima (Appendix Fig. S6).

Despite very conserved features, the sequences of the P-stalk proteins display some differences in human and yeast (Fig. 1; Appendix Fig. S3). In yeast, the P1/P2 proteins diverge into four distinct elements, forming two hetero-dimers, P1A-P2B and P1B-P2A (Tchorzewski et al, 2000a; Tchorzewski et al, 2003). Contrary to the human counterparts, each of the P-stalk proteins contains only one phosphorylated serine residue and only one closely placed Lys residue in the P1A-P2B dimer, while P1B-P2A has only a distal Lys and uL10 has none at all (Appendix Fig. S3). As a result, the U-bent occurs only sporadically and is not as stable as the human CTD. Notably, the presence of a Lys residue in the vicinity of the phosphorylated serine makes the U-bent conformation more stable during simulations conducted for P1A and P2B than for other

peptides (Fig. 10B; Appendix Figs. S7 and 8). The peptides from the P1B and P2A proteins do not display the fixed U-shape bent but can adopt a variety of conformations (Appendix Fig. S7), resulting in a more even FEL distribution (Appendix Fig. S9). In Archaea, there are even two different P-stalk proteins, uL9 and P1, which form a homodimer attached to the uL10. P1 and uL10 share similarity with the CTDs of eukaryotic P-stalk proteins but are not phosphorylated (Fig. 1; Appendix Fig. S3). The MD simulations revealed that, due to the absence of phosphorylated serine residues, there was no discernible tendency towards forming a U-bent conformation (Fig. 10C; Appendix Figs. S10 and 11) with an even FEL distribution (Appendix Fig. S12).

The MD simulations were also conducted on all types of CTDs with peptides initially formed as α-helices. Although all the peptides ultimately transitioned to an unstructured state, those from Archaea retained their helical conformation significantly longer than the CTDs from the other species (Appendix Figs. S13, 14 and 15). This enhanced stability of the archaeal α-helices is consistent with a reported structural analysis, which indicates that archaeal peptides can form an α-helix when interacting with

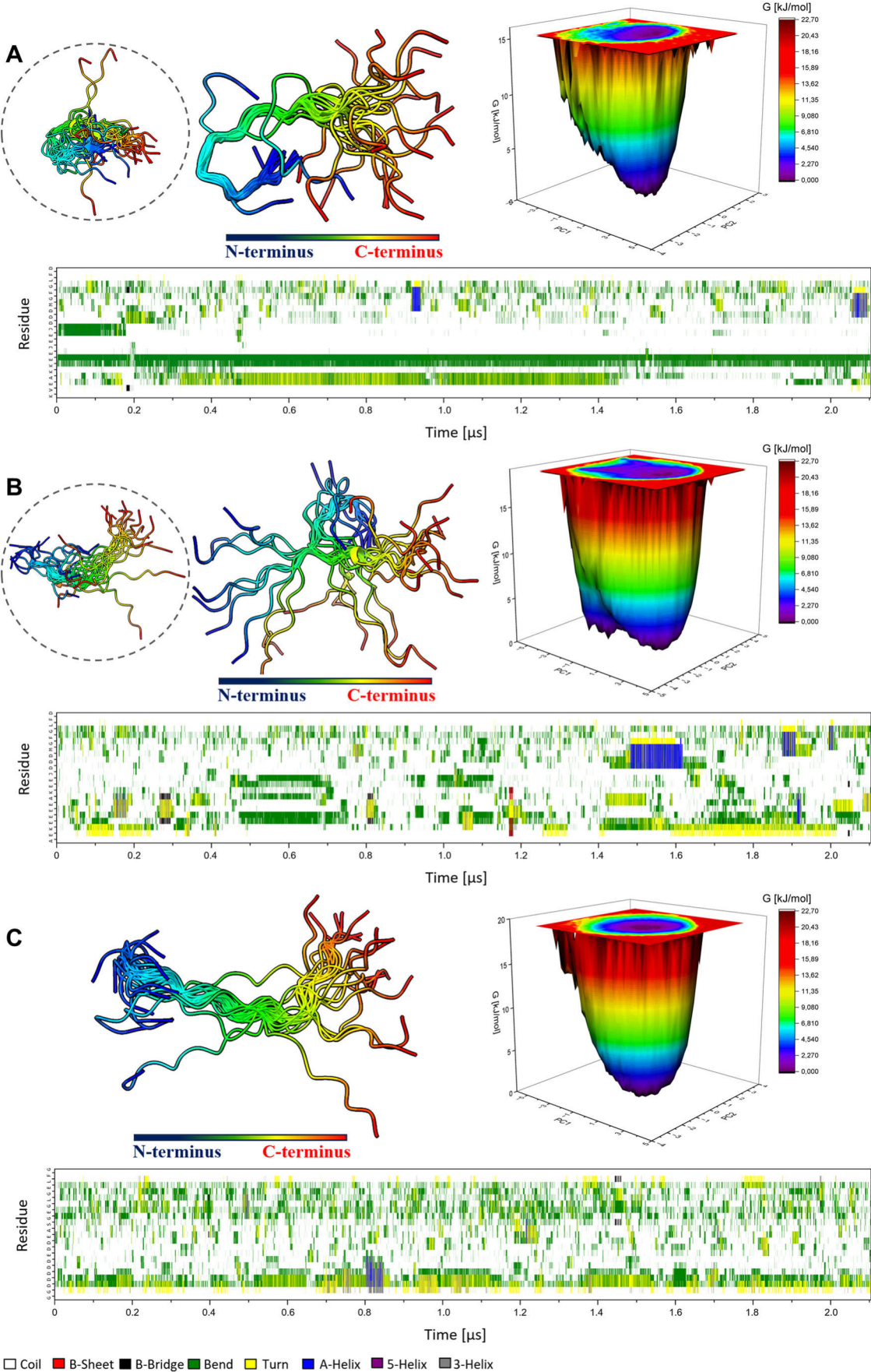


Figure 10. MD simulation of structures of the unphosphorylated and phosphorylated forms of P-stalk C-terminal peptides.

(A–C) Representative structures of the most populated clusters from the MD simulation of the human, yeast and archaeal P1 proteins, respectively. Insets in (A) and (B) display the representative structures of the most populated clusters from the MD simulation of the unphosphorylated P1-proteins, respectively. The right panels in (A), (B), and (C) present free energy landscape plots of C-termini peptides folding in unphosphorylated and phosphorylated state. The bottom panels in (A), (B), and (C) depict C-termini peptide secondary structure transitions over time during the MD simulations.

trGTPases (Murakami et al, 2018; Nomura et al, 2012; Tanzawa et al, 2018). It seems that archaeal CTDs tend to maintain α -helices, while the phosphorylation of eukaryotic CTDs destabilizes α -helix formation, resulting in a U-bent conformation with the very C-termini of the CTD adopting a coil-like conformation. To address the role of lysine residues, the MD simulations were performed with lysine-to-alanine mutant forms of the C-terminal P-stalk peptides (Appendix Fig. S16). The analysis showed, that the absence of lysine made the peptides even less likely to form the bends and turns responsible for the U-bend structure in the N-terminal part of the peptide, suggesting that lysine residues play a crucial role in stabilizing the U-bend conformation.

In conclusion, the MD simulations demonstrate that the presence of phospho-serine sidechains and Lys residues in their vicinity stabilizes the U-bent structure in the CTD and stimulates the collapsed/globule-to-coil transition, an effect particularly evident in the case of human P-stalk proteins. The simulations demonstrate the remarkable evolutionary transition of CTDs. In Archaea, which lack CTD phosphorylation, the coil-like structure is readily maintained and may also form a stable α -helix. In yeast, the CTDs exhibit transitional properties; the CTDs from the dimer (P1B-P2A) are similar to the archaeal type, while the others (P1A-P2B) resemble properties of human CTDs. In human, the double phosphorylation of the CTDs is crucial for stabilizing the U-bent structure, resulting in an extended coil conformation.

Discussion

The ribosomal P-stalk located on the 60S subunit is primarily responsible for interactions with trGTPases and facilitate GTP hydrolysis. However, the eukaryotic ribosomal P-stalk has additional functions beyond its canonical role in translation. It also provides an interacting platform for several additional auxiliary factors, such as Ribosome Inactivating Proteins, RIPs (Grela et al, 2017; Grela et al, 2019) and also interacts with the Gcn2 kinase, which plays an important role in cellular response to adverse growth conditions (Masson, 2019) and can be perceived as a key link between the Integrated Stress Response signal transduction pathways and the ribosome (Wu et al, 2020). Consequently, the molecular characteristics of the P-stalk may be tailored to interact with a wide range of diverse partners.

Thus, the question arises: how did the eukaryotic P-stalk achieve such unique, pleiotropic, and adaptive characteristics to interact with a variety of protein partners? Considering the high degree of flexibility of CTDs from the P-stalk proteins, which have IDP properties, these unique CTD architecture may allow promiscuous interactions with different partners on various occasions, following the so-called “fly-casting” mechanism (Shoemaker et al, 2000). This phenomenon was already proposed for the bacterial L7/12 stalk, which is able to sample a large volume around the ribosome, being

critical for functional interactions with trGTPases (Bernado et al, 2010; Diaconu et al, 2005). The key aspect of the IDP is its ability to interact with a broad range of partners with high specificity and modest affinity (van der Lee et al, 2014), and these interactions are often achieved or regulated by phosphorylation. P-stalk proteins belong to the group of ribosomal proteins that undergo phosphorylation (Hasler et al, 1991; Zambrano et al, 1997; Zinker and Warner, 1976), but the phosphorylation state and the role of this post-translational modification has not been revealed.

Here, we demonstrate that, in optimal growth conditions, P-stalk proteins are found exclusively in a phosphorylated state, and this feature is conserved in Eukaryotes (Fig. 1). Considering the functional implications of phosphorylation, such modifications are typically transient (Nussinov et al, 2012), and several studies have indicated that the phosphorylation levels of P-stalk proteins may vary (Maniratanachote et al, 2006; Pan et al, 2014; Pan et al, 2013; Zampieri et al, 2001). However, despite examining a wide range of stress conditions, we found that these proteins remained phosphorylated, regardless of the stress conditions tested. This observation is supported also by recent studies examining the effect of UV on cellular metabolism (Sinha et al, 2024), where P-stalk proteins were found phosphorylated.

Our earlier study suggested that phosphorylation decreases the affinity of the P-stalk proteins toward RIPs (Horbowicz-Drozdal et al, 2021). Importantly, here we show that the phosphorylation of the P-stalk also decreases its affinity toward the trGTPase, eIF5B. However, this information is in contrast to the previous report, which showed that phosphorylation of the P2 protein can promote interaction with eEF2 (Vard et al, 1997). In this case, the recombinant P2 protein was used as the analyte, which may have biased the results, as P2 protein is known to tend to form high-molecular mass oligomers without its natural P1 partner (Grela et al, 2008b; Tchorzewski et al, 2000a; Tchorzewski et al, 2000b). The observed lower affinity has a biological impact. We observed that it reduces the stimulatory effect on GTP hydrolysis by trGTPases. Considering the ribosome performance, the speed and accuracy of decoding are fundamental elements ensuring correct metabolic fitness of living cells. The GTP hydrolysis rate is crucial for optimizing both the speed and accuracy of ribosome action and represents a balance between these two fundamental parameters (Wohlgemuth et al, 2010). However, the measured rate of translation in a codon-resolved manner, showed that the overall speed of translation is not regulated by the CTD phosphorylation. However, as we have already reported, the P-stalk also contributes to the accuracy of decoding (Wawiora et al, 2017), and our finding that the SDPM mutant has a higher error rate due to the higher GTP hydrolysis rate further supports this idea. This finding is supported by reports showing that, in contrast to the bacterial translation, which seems to be optimized towards a high speed of translation at the cost of fidelity (Wohlgemuth et al, 2010), the ribosomes in Eukaryotes, are slower than in bacteria. The timing of

the reactions occurring during the decoding has a significant role in adjusting the decoding accuracy, which provides superior decoding fidelity in eukaryotes (Holm et al, 2023). Thus, it can be concluded that the phosphorylation of P-stalk proteins optimizes the decoding, probably by controlling the timing of interactions with the ternary complex.

At this point, another question arises: what is the biological outcome of the P-stalk protein phosphorylation beyond translation? We have shown that P-stalk proteins are phosphorylated regardless of the metabolic conditions. While the overall translation rate is not changed, considering phospho- or dephospho-states, the functional effect of the P-stalk phosphorylation is more pronounced upon ER-stress, as shown by the genome-wide analysis. Upon stress, we have shown that ribosomes start to pause specifically at CGA arginine codons, which is in line with a previous report showing that the decoding of the CGA codons may be challenging due to the low abundance of tRNA^{Arg} and its inosine wobble pairing interaction (Letzring et al, 2013). Also, as it was also already reported, successive CGA codons in yeast trigger ribosome stalling, likely through steric effects at both decoding and peptidyl transferase centers (Matsuo et al, 2020), which may lead to the ribosome stalling-dependent activation of Gcn2 kinase (Kim and Zaher, 2022). Indeed, the presented ribosome profiling data were supported by the evaluation of eIF2 α phosphorylation upon the Tun treatment. In the WT cells, the eIF2 α phosphorylation increased at the onset of stress but decreased over time due to adaptation. In contrast, the SDPM mutant exhibited a delayed response and a constant increase in the eIF2 α phosphorylation, indicating the lack of proper adaptation. Interestingly, the effect on eIF2 α phosphorylation shows a significant delay in relation to the onset of stress. Although the stress response begins within 15 min, the effect on eIF2 α phosphorylation is not observed until after one hour. This suggests that signaling from ribosomes to downstream effectors is buffered, possibly acting like a rheostat, as suggested by recent studies (Sinha et al, 2024). In addition, it has been reported that Gcn2 in yeast may associated with the 60S subunit, especially when cells are exposed to stress conditions (Dong et al, 2000; Ramirez et al, 1991; Zhu and Wek, 1998). Our findings indicate that Gcn2 has a potential ability to bind to the ribosomal 60S subunit, which is enhanced by the Tun-induced stress, and that phosphorylation of the P-protein further strengthens this binding, suggesting that P-stalk phosphorylation may have a beneficial impact on the interaction with Gcn2 and consequently, on the proper development of the stress response.

Another intriguing functional aspect of P-stalk proteins is their ability to interact with a variety of partners. Structural work suggests that the CTD of P-stalk proteins can adopt different structural configurations upon interaction with different protein partners, such as the trGTPases eIF5B (Murakami et al, 2018), eEF1A (Nomura et al, 2012), eEF2 (Tanzawa et al, 2018), and ABCE1 (Imai et al, 2018) and RIPs, such as ricin (Fan et al, 2016), trichosanthin (Too et al, 2009), and Shiga toxin (Kulczyk et al, 2023). Therefore, how does the CTD achieve such adaptability to different partners? An NMR study of human P-stalk proteins has demonstrated that the CTD is unstructured (Lee et al, 2012) and behaves as the IDR in the solution (Mishra and Hosur, 2015). On the other hand, the crystallographic structural studies have shown that the CTD can adopt an α -helical structure when binding to

trGTPases or adopt various extended conformations when binding to RIPs. Our MD simulations suggest that such structural adaptability can be achieved through phosphorylation of the CTD. As it has already been shown, several functional modules within the IDR can be defined as a short linear motif (SLiM); the first one is usually involved in direct interactions with partners, and the other is engaged in the regulation of the interplay with partner protein (Davey et al, 2012). The C-terminal element of human P-stalk proteins (...KVEAKKEESEESDDDDMGFGLFD) entails two characteristic elements: a highly charged cluster of amino acid residues including phosphorylated serine residues (marked in red) and a hydrophobic part (marked in blue), which represents the main interaction element for various protein partners. The latter sequence, MGFGLFD, can be considered a SLiM #1 involved in the binding of protein partners and adopting various structural organizations upon binding, ranging from linear extended structures (Fan et al, 2016; Too et al, 2009) (as reported in the case of RIPs) to an α -helical fold (Kulczyk et al, 2023; Tanzawa et al, 2018) (characteristic for trGTPases or RIPs). The SLiM #2 (in red) within the CTD is highly charged and contains phosphorylated serine residues. The structure of this peptide has not been solved experimentally, probably due to the high intrinsic flexibility of this region. Based on the MD simulation, we showed that phosphorylation changes the structure of SLiM #2 by stabilizing the U-bent conformation, which provides structural freedom for SLiM #1. The unphosphorylated CTD displays a continuum of conformational ensembles transiently sampling coil-like states and more compact collapsed globules, with a broad range of secondary structures. Thus, the presence of phosphate group, especially in higher Eukaryotes, induces the U-bent shape within the SLiM #2, which provides conformational freedom to the very C-terminal patch of amino acid residues within the SLiM #1. We postulated, that this transition may increase the propensity to interact with a broad repertoire of partners and, on the other hand, destabilize the interactions, providing short-lived transient binding. As demonstrated for other IDPs, particularly those possessing polyelectrolyte properties with numerous Glu and Asp residues, the transition from collapsed/globule to expanded unstructured coils is sharp, indicating that minor changes in the net charge, such as those resulting from phosphorylation, can trigger globule-to-coil transitions (van der Lee et al, 2014). Therefore, it can be postulated that the phosphorylated CTD of P-stalk proteins, with the coil-like conformation, gains conformational freedom, which in turn may allow robust and promiscuous but transient binding to the various protein partners. The proposed CTD structural behavior may be supported by the fact that binding of P-stalk protein CTDs to trGTPases and RIPs is taking place on the various structurally distinct concave extended grooves on the surface of partner proteins (Fan et al, 2016; Imai et al, 2018; Kulczyk et al, 2023; Murakami et al, 2018; Nomura et al, 2012; Tanzawa et al, 2018; Too et al, 2009), acting as an 'entropic trap', reducing the overall free energy of binding.

Taken together, our findings reveal that P-stalk proteins exist exclusively in a phosphorylated state on the ribosomal particles, and the presence of phosphate groups leads to collapsed/globule-to-coil conformational transitions within the intrinsically disordered region of the CTD. This structural shift yields functional outcomes by optimization of ribosome interactions with auxiliary factors, such as trGTPases; however, this post-translational modification

does not affect the overall rate of translation, apart from decoding fidelity. On the other hand, ribosomal particles with phosphorylated P-stalk proteins effectively activate the Gcn2 kinase, contributing to the correct timing of kinase activation. Therefore, we propose that the CTD phosphorylation not only optimizes the ribosome performance in translation per se but has also triggered the expansion of the ribosomal interactome beyond the trGTPases, allowing timely interaction with the Gcn2 kinase. Thus, P-stalk proteins with the IDP region enable eukaryotic ribosomes to serve as an adaptable regulatory hub linking translation with regulatory pathways.

Methods

Reagents and tools table

Reagent/Resource	Reference or Source	Identifier or Catalog Number
Experimental Models		
Yeast strain - BY4741/2 (MATa/alpha his3Δ1 leu2Δ0 met15Δ0 ura3Δ0)	EUROSCARF	Y00000/Y10000
Yeast strain - uL10TH199 (MATa; RPP0-TH199; his3Δ1; leu2Δ0; lys2Δ0; and ura3Δ0)	Grela et al, 2010	Not applicable
Yeast strain - SDPM (BY4741 with mutations S-to-A within all P-stalk proteins at the C-termini, uL10, P1A, P1B, P2A and P2B)	This study	Not applicable
Cell line - MEF (DR-Wildtype)	ATCC	CRL-2977
Cell line - NIH 3T3	ECACC	93061524
Male albino Swiss mice	Laboratory Animals Breeding, Ilkowiec, Poland	Not applicable
Recombinant DNA		
pDB688	Provided by Bedwell laboratory	Not applicable
pDB868 CGC ₂₄₅	Provided by Bedwell laboratory	Not applicable
pJD375	Provided by Dinman laboratory	Not applicable
pJD643 AGC ₂₁₈	Provided by Dinman laboratory	Not applicable
Antibodies		
Mouse anti-tubulin (DM1A)	Abcam	Ab7291
Rabbit anti-RPLP0	Abcam	Ab192866
Mouse anti-eIF2α (D-3)	Santa Cruz Biotechnology	Sc-133132
Rabbit anti-RPS5	Santa Cruz Biotechnology	Sc-100832
Rabbit anti-RPL3	Proteintech	11005-1-AP
Rabbit anti-RPLP1	Atlas Antibody	HPA003368
Rabbit anti-P1A/P2B	In house, Boguszewska et al, 2002	Not applicable
Rabbit anti-RPL25/uL23	Provided by Ed Hurt Laboratory	Not applicable
Rabbit anti-GAPDH	Proteintech	60004-1-Ig

Reagent/Resource	Reference or Source	Identifier or Catalog Number
Rabbit anti-EIF2S1 (phospho S51)	Abcam	Ab32157
Goat anti-Rabbit IgG (H + L)-HRP Conjugate	Bio-Rad	1706515
Goat anti-Mouse IgG (H + L)-HRP Conjugate	Bio-Rad	1706516
Rabbit anti-Gcn2	Provided by Alan Hinnebusch lab	Not applicable
Chemicals, Enzymes and other reagents		
YPD medium	MP Biomedicals	114001032
Dulbecco's Modified Eagle Medium- high glucose	Merck	D6429
Fetal bovine serum	Merck	F0804
Penicillin/Streptomycin	Merck	P4333
Minimal Eagle's Medium	Merck	M2279
GlutaMAX	Gibco	35050-061
Complete Mini EDTA-free Protease Inhibitor Cocktail	Roche	04693159001
Complete Mini EDTA-free Protease Inhibitor Cocktail	Roche	5056489001
Bio-Rad Protein Assay Dye Reagent Concentrate	Bio-Rad	5000006
Bovine Serum Albumin	Merck	A4503
DNase I	Jena Bioscience	EN-1735
Puromycin	Serva	33835
Cycloheximide	Merck	C7698
RNase inhibitor	Merck	R7397
Dulbecco's Buffered Saline	Gibco	14040133
FastAP Thermosensitive Alkaline Phosphatase	Thermo Fisher Scientific	EF0652
Phos-tag	Wako Chemicals	AAL-107
Clarity™ Western ECL	Bio-Rad	170-5061
Thrombin	Merck	T4648-1KU
Software		
Microcal Origin 6.0	OriginLab Corporation (https://www.originlab.com)	Not applicable
ImageLab	Bio-Rad	Not applicable
RStudio environment version 4.2.1	Posit (https://posit.co)	Not applicable
NanoTemper NT version 2.2.4	NanoTemper Technologies	Not applicable
UCSF ChimeraX tool	University of California (https://www.cgl.ucsf.edu/chimerax/)	Not applicable
Other		
Glass beads	Merck	G8772
Corundum	Merck	8334
Centrifuge MPW-150R	MPW Med Instruments	Not applicable

Reagent/Resource	Reference or Source	Identifier or Catalog Number
Optima™ MAX-TL	Beckman-Coulter	Not applicable
Rotor MLA-80	Beckman-Coulter	Not applicable
Optima™ L-100XP	Beckman-Coulter	Not applicable
Rotor SW32 Ti	Beckman-Coulter	Not applicable
Gradient master 108	Biocomp	Not applicable
Gradient fractionator	ISCO Brendel	Not applicable
PVDF membrane	Millipore	IPVH00010
Mini Trans-Blot Cell Module	Bio-Rad	1703930
ChemiDoc™ MP analyzer	Bio-Rad	Not applicable
Monolith NT.115	NanoTemper	Not applicable
Nikon Eclipse E200	Nikon	Not applicable
Monolith His-Tag Labeling Kit RED-tris-NTA 2 nd Generation Kit	NanoTemper Technologies	Not applicable

Yeast strains

Yeast haploid strains BY4741/2 (*MATa/alpha his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*) purchased from EUROSCARF (Scientific Research and Development, Germany) were used as wild-type strains. The SDPM mutant strain with serine to alanine substitutions within the C-termini of all stalk proteins: uL10, P1A, P1B, P2A, and P2B was derived from the BY4741/2 strains. This mutant was constructed in a stepwise fashion with the use of a pop-in/pop-out strategy (Boeke et al, 1987) using the BY4741 and BY4742 parent strains and URA3-based plasmids containing the mutation within synthetic DNA (Genscript Inc). After confirmation of three different substitutions in each strain, two strains were mated, sporulated, and subjected to tetrad analysis. The resulting spores were genotyped by PCR-seq to identify the quintuple SDPM mutant. The yeast uL10 mutant strain uL10TH199 (*MATa; RPP0-TH199; his3Δ1; leu2Δ0; lys2Δ0; and ura3Δ0*) used for purification of P-stalk complexes was described previously (Grela et al, 2010). The yeast cells were cultivated in YPD medium (1% yeast extract, 2% peptone, 2% dextrose) at 30 °C.

Cell lines

MEF cells were purchased from ATCC and were grown in Dulbecco’s modified Eagle medium - high glucose (DMEM-Hi) (Merck, cat#D6429) supplemented with 10% FBS (Merck, cat#F0804) and with antibiotics penicillin/streptomycin (Merck, cat#P4333). The NIH3T3 cell line was purchased from ECACC and was grown in Minimal Eagle’s Medium (Merck, cat#M2279)

supplemented with 10% FBS, penicillin/streptomycin, and 1X GlutaMAX solution (Gibco, cat#35050-061).

Yeast extracts and ribosome purification

To obtain cell extracts, the yeast cells were grown to OD₆₀₀ of 0.4–0.6 and harvested by centrifugation at 1000 × g for 3 min at room temperature (RT). Pelleted cells were suspended in RIPA lysis buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 0.1% SDS, 1% NP-40) supplemented with 1x complete Mini EDTA-free Protease Inhibitor Cocktail (Merck, Roche cat#04693159001) and disrupted by vigorous shaking with glass beads (Merck, Sigma-Aldrich #catG8772) followed by centrifugation at 15,000 × g for 20 min at RT. The protein concentration in the supernatant was determined by the Bradford procedure using Bio-Rad Protein Assay Dye Reagent Concentrate (Bio-Rad, cat#5000006) and bovine serum albumin (BSA, Merck, Sigma-Aldrich cat#A4503) used for preparation of the standard curve.

For purification of ribosomes, the yeast cells were harvested by centrifugation in the logarithmic phase of growth when OD₆₀₀ reached 0.5. Then, the cells were washed with ice-cold buffer A (50 mM Tris-HCl pH 7.5, 80 mM KCl, 12.5 mM MgCl₂, 1 mg/ml heparin, 2 mM DTT, 1 mM PMSF, 0.5 μM puromycin, 1x complete Mini EDTA-free Protease Inhibitor Cocktail), frozen in liquid nitrogen, and preserved at –70 °C. The following day, the frozen cells were ground with corundum (Merck, Sigma-Aldrich cat#8334) in buffer A supplemented with DNase I (Jena Bioscience, cat#EN-173S). The obtained cell lysate was centrifuged at 105,000 × g (MPW-150R, MPW Med Instruments) for 30 min at 4 °C. For ribosome purification, the clarified lysate was loaded on a 30% sucrose cushion in buffer containing 50 mM Tris-HCl pH 7.5, 500 mM KCl, 12.5 mM MgCl₂, 1 mg/ml heparin, 2 mM DTT, and 1 mM PMSF and centrifuged at 200,000 × g for 5 h at 4 °C (MLA80 rotor, Beckman-Coulter). The pellet of purified ribosomes was resuspended in storage buffer containing 50% glycerol, 50 mM Tris-HCl pH 7.5, 80 mM KCl, 10 mM MgCl₂, 0.5 mM EDTA, and 5 mM DTT.

Mammalian cell extracts and ribosome purification

The whole cell extracts were prepared as follows: the cells were washed three times with ice-cold TBS buffer (50 mM Tris-HCl pH 7.4, 150 mM NaCl) and lysed with RIPA buffer (50 mM Tris-HCl pH 7.4, 150 mM NaCl, 1% NP-40, 0.1% SDS) directly on plates. Then, the cell lysates were transferred to 1.5 ml tubes, clarified by centrifugation at 12,000 × g for 15 min at 4 °C, sonicated (two rounds of 10 sec-pulse with 30% amplitude), and centrifuged again with the same parameters as above. After clarification of the lysates, the protein concentration was measured using Bio-Rad Protein Assay Dye Reagent Concentrate (Bio-Rad, cat#5000006) and BSA (Merck, Sigma-Aldrich cat.#A4503) used for preparation of the standard curve.

Ribosomes from the MEF cell line were purified according to the procedure described in (Belin et al, 2010) with modifications. Briefly, the MEF cells were grown to about 75% confluency, washed three times with ice-cold TBS, harvested with a cell scraper, and spun down by centrifugation: 500 × g, 5 min at 4 °C. In the next step, the cells were resuspended in hypotonic buffer (50 mM

Tris-HCl pH 7.4, 250 mM sucrose, 250 mM KCl, 5 mM MgCl₂) containing 0.5 μM puromycin (Serva, cat#33835) and 1x complete Mini EDTA-free Protease Inhibitor Cocktail (Merck, Roche cat#5056489001) and incubated for 15 min on ice. Then, the cell homogenate was cleared in two steps. First, the nuclei were pelleted at 800 × g for 10 min at 4 °C and, secondly, the mitochondria and other membrane fractions were sedimented using 16,000 × g for 15 min at 4 °C. The obtained post-mitochondrial fraction was supplemented with KCl to the final concentration of 0.5 M, layered onto a sucrose cushion containing 0.5 M KCl, and centrifuged at 250,000 × g for 3.5 h at 4 °C (MLA-80 rotor; Beckman-Coulter) to sediment the ribosomal fraction. The ribosomal pellet was washed with ice-cold MilliQ water and resuspended in storage buffer containing 50% glycerol, 50 mM Tris-HCl pH 7.5, 80 mM KCl, 10 mM MgCl₂, 0.5 mM EDTA, and 5 mM DTT.

Polysome profiles of yeast cells

To obtain cell extracts for polysome profile analysis, the yeast cells were grown to OD₆₀₀ = 0.4–0.6 with subsequent 20-min treatment with 100 μg/ml of cycloheximide (CHX, Merck, Sigma cat#C7698) and harvested by centrifugation. The cell pellet was resuspended in lysis buffer (10 mM Tris-HCl pH 7.5, 100 mM NaCl, 30 mM MgCl₂, 100 μg/ml CHX, 1 mM PMSF, 6 mM 2-mercaptoethanol, 1 nM pepstatin A, 10 nM leupeptin, 10 ng/ml aprotinin, 200 ng/ml heparin, and 1xRNase inhibitor (Merck, Sigma cat#R7397)) and the cells were disrupted by vigorous shaking with glass beads at 4 °C. The obtained cell lysate was cleared by centrifugation at 12,000 × g for 10 min at 4 °C. The RNA content was measured at 260 nm and 15 U of RNA were run through a 10–50% sucrose gradient using a Beckman Coulter SW32 Ti rotor for 4.5 h at 26,500 RPM at 4 °C. The gradients were analyzed using an ISCO Brendel density gradient fractionator. The analyses of polysome profiles were conducted using MicrocalOrigin 6.0 software. The area of polysome fractions was calculated with the integration method to assess the total surface area and to determine the polysome-to-monomosome (P/M) and 40S/60S ratios. For immunodetection of the fractions, proteins from each fraction were precipitated with TCA, centrifuged at 12,000 × g for 30 min at 4 °C, washed twice with acetone, and analyzed by Western Blotting.

Polysome profiles of mammalian cells and tissue

To prepare cell extracts for the polysome profile analysis, the cell lines were grown to about 75% confluency and the medium was refreshed 3 h before harvesting. Before harvesting, the cells were treated with cycloheximide (CHX, 100 μg/ml) (Merck, cat#C7698) for 3 min to stall the polysomal fraction. Next, the cells were washed twice with ice-cold DPBS (Gibco, cat#14040133) supplemented with 100 μg/ml of CHX and harvested by centrifugation at 500 × g for 5 min at 4 °C. The pelleted cells were suspended in polysome lysis buffer (20 mM HEPES-KOH pH 7.4, 200 mM KCl, 5 mM MgCl₂, 0.5% NP-40, CHX 100 μg/ml) supplemented with EDTA-free Protease Inhibitor Cocktail (Roche, cat#5056489001) and RNase Inhibitor (Merck, cat#R7397) and incubated on ice for 30 min, with 15 s vortexing every 5 min. Then, the cell lysate was passed 20 times through a 23 G needle and incubated on ice for 5 min with 3 rounds of 1 min vortexing; the lysate was centrifuged at 10,000 × g for 15 min at 4 °C. The RNA concentration was

measured at 260 nm, and six OD₂₆₀ units of cell extracts were loaded onto a 15–50% sucrose gradient (20 mM HEPES-KOH pH 7.4, 200 mM KCl, 5 mM MgCl₂) prepared on a gradient mixer (BioComp). The samples were centrifuged at 26,500 rpm using a Beckman Coulter SW32 Ti rotor for 4.5 h at 4 °C. The gradients were fractionated using an ISCO Brendel Density Gradient Fractionator, and the absorbance at 254 nm was recorded. Proteins in the collected fractions were precipitated overnight using TCA at 4 °C. The following day, the TCA-treated-fractions were centrifuged at 12,000 × g for 20 min at 4 °C and then precipitated proteins were washed twice with ice-cold acetone and spun down at 12,000 × g for 10 min at 4 °C. Dried protein pellets were resuspended in SDS-PAGE loading buffer, heated at 95 °C for 5 min, and analyzed by SDS-PAGE/Western Blotting.

The phosphatase-treated samples were prepared as follows: the cell lysate containing six OD₂₆₀ units of RNA was incubated with 70 U of FastAP Thermosensitive Alkaline Phosphatase (Thermo Fisher Scientific, cat#EF0652) for 1 h at 37 °C; lysate supplemented with FastAP buffer was used as a control. Both samples were processed following the protocol described above for the standard cell extracts used for the polysome profile analysis.

Male albino Swiss mice (weighing 20–30 g) obtained from a licensed breeder (Laboratory Animals Breeding, Ilkowiec, Poland) were used for preparation of the polysomal fraction from mouse liver. The mice were adapted to the laboratory conditions for at least 1 week before the experiments. The animals were housed in groups of 8 per cage in controlled environmental conditions (temperature: 21–24 °C, relative humidity: 45–65%) with an artificial 12/12 h light/dark cycle (light on at 6:00 a.m.) and free access to food pellets and water. The housing and experimental procedures were conducted under the guidelines provided by the European Union Directive of 22 September 2010 (2010/63/EU) and Polish legislation concerning animal experimentation. All experimental procedures were approved by the Local Ethics Committee in Lublin, Poland (license no 107/2019). The freshly prepared liver was washed with 0.9% of ice-cold NaCl, cut into small pieces, weighted, and frozen in liquid nitrogen. Next, the liver was homogenized on ice in homogenization buffer (50 mM HEPES-KOH pH 7.4, 250 mM KCl, 5 mM MgCl₂, 250 mM sucrose, 100 μg/ml of CHX, 1 × RNase Inhibitor) using a glass-teflon Dounce homogenizer with 50 strokes. For 1 g of mouse liver, 3 ml of homogenization buffer was used. The tissue lysate was clarified by centrifugation at 3000 × g for 15 min at 4 °C, supplemented with both 10% Triton X-100 and 13% sodium deoxycholate at a concentration of 100 μg/ml, and mixed gently. The RNA concentration was measured at 260 nm and ten OD₂₆₀ units were loaded onto a 15–50% sucrose gradient and separated using a Beckman Coulter SW32 Ti rotor at 26,500 rpm for 4.5 h at 4 °C. The gradient fractionation and analysis of polysomal fractions were performed as described above.

Phos-tag analysis

Samples analyzed with the Phos-tag/Western Blotting method including purified ribosomes, total cell extracts, and cell extract used for polysome fractionation were prepared in buffers without EDTA. Phos-tag gels: 10% polyacrylamide supplemented with 50 μM Phos-tag (Wako Chemicals, cat#AAL-107) and 100 μM ZnCl₂ (Acros Organics) were run with the constant current of

25 mV for 3 h at 4 °C. Dephosphorylated and phosphorylated samples were used as controls. For dephosphorylation, 20 µg of ribosomes and 10 µg of total cell extracts were incubated for 1 h at 37 °C with 10 U of FastAP Thermosensitive Alkaline Phosphatase (ThermoFisher Scientific, cat#EF0652), whereas cell extracts (six OD₂₆₀ units of RNA) used for polysome fractionation were treated with 70 U of FastAP and processed analogously as the untreated samples until proteins from the polysome profile fractions were precipitated. Finally, all samples were diluted in SDS-PAGE loading buffer and heated for 5 min at 95 °C. Phosphorylated samples were prepared with CK2 kinase; to this end, 20 µg of purified ribosomes were incubated with 10 U of CK2 alpha subunit (kindly provided by Dr. K. Kubiński, KUL, Lublin, Poland) for 1 h at 37 °C, diluted in SDS-PAGE loading buffer, and boiled for 5 min at 95 °C. After electrophoresis, the gels were washed twice for 10 min with transfer buffer (25 mM Tris, 236 mM glycine, 20% methanol) supplemented with 10 mM EDTA pH 8.0, followed by washing in plain transfer buffer to remove EDTA. The transfer of proteins to PVDF membranes was carried out with a constant current of 300 mA for 3 h using wet-transfer, followed by immunodetection.

Immunodetection

Protein samples (10 µg of cell extracts and 20 µg ribosomes) were resolved by SDS-PAGE electrophoresis in polyacrylamide gels (10% for the Phos-tag analysis and GCN2 detection and 12% for the other analyses) and transferred to PVDF membranes (Millipore, cat#IPVH00010) using a Mini Trans-Blot Cell Module (Bio-Rad). The membranes were blocked with 5% of non-fat dry milk in TBST buffer (50 mM Tris-HCl pH 7.4, 150 mM NaCl, 0.1% Tween-20) for 1 h with gentle orbital shaking at RT and, after washing with TBST, incubated with indicated primary antibodies diluted in TBST supplemented with 3% BSA for 2 h at RT. Then, the membranes were washed three times in TBST for 15 min followed by incubation with secondary antibodies diluted in TBST supplemented with 5% of non-fat milk for 1 h at RT. After three washes in TBST (15 min each), chemiluminescent signals were developed using Clarity™ Western ECL Substrate (Bio-Rad, cat#170-5061) and visualized with the aid of the Bio-Rad ChemiDoc™ MP analyzer and Image Lab software (Bio-Rad). For quantitative analysis, chemiluminescent blot signals developed with horseradish peroxidase-conjugated secondary antibodies were acquired and analyzed with the system and software mentioned above and plotted using Microsoft Excel. For statistical analysis, one-sided Welch's two-sample t-tests were carried out at a fixed significance level of $\alpha=0.05$. The tests were performed in the RStudio environment (R version 4.2.1) (<http://www.rstudio.com>).

Antibodies

Antibodies directed toward tubulin (DM1A, #ab7291) and uL10 (#ab192866) were purchased from Abcam. Antibodies for eIF2 α (D-3, #sc-133132) and uS7 (#sc-100832) detection were supplied by Santa Cruz Biotechnology. Antibodies for uL3 (#11005-1-AP) were purchased from Proteintech. Antibodies recognizing P1 (#HPA003368) were acquired from Atlas Antibody, and were used as a readout for the P1-P2 dimer. Antibodies for yeast uL10 and P1A/P2B were prepared in the Department of Molecular Biology, Lublin, Poland, based on purified recombinant proteins

(Tchorzewski et al, 1999) and were used as described previously (Boguszewska et al, 2002). Goat anti-rabbit IgG (H + L)-HRP Conjugate (#1706515) and goat anti-mouse IgG (H + L)-HRP Conjugate (#1706516) were purchased from Bio-Rad. Antibodies recognizing Gcn2 kinase were kindly provided by Alan Hinnebusch (National Institutes of Health, Bethesda, MD, USA). Antibodies against uL23 were kindly provided by Ed Hurt (University of Heidelberg, Germany).

Meta-analysis of proteomic data for P-stalk protein phosphorylation

The PeptideAtlas platform was used to map the ribosomal P-stalk protein phosphorylation (Desiere et al, 2006). The Human 2022-01-build was used to search for experimentally observed peptides with the C-terminal group and two C-terminal serine residues. The following dataset was used: build name—human 2024-01, reference_database—THISP_2024-01-01 + gEVE1.1 + TEs, with distinct_peptides 3,945,444, #Experiments 3160, canonical_proteins 17,416. Once the list was derived, the peptides with the highest number of observations (spectra) were selected for each protein. The VEAKEESESEDEDMGFGLFD peptide was used for the human protein uL10, whereas the single common peptide with the sequence KEESEESDDDMGFGLFD was used for the human P1 and P2 proteins. The phosphorylation observed within the peptide was classified into four groups: first serine residue modified calculated from NTD (S304, S101, and S102, for uL10, P1, and P2, respectively), second serine residue modified (S307, S104, and S105 for uL10, P1, and P2, respectively), both serine residues modified, or no serine residue modified. The peptides were searched in the Human Phosphoproteomics 2022-04 build. In the 2022-04 build, the phosphorylation observed in the peptides was divided into three groups: first, second, and both serine residues modified.

Isolation and in vitro dephosphorylation of the ribosomal P-stalk

The purification of the P-stalk complex was performed using *S. cerevisiae* uL10 mutant uL10TH199 and the procedure described previously (Grela et al, 2010). In brief, the mutant strain was used where uL10 has a short amino acid sequence recognized by thrombin placed at position 199 with 6xTag, and upon thrombin treatment the truncated uL10 protein (amino acids 199–312) can be released with P1A-P2B and P1B-P2A dimers (Δ uL10₁₉₉₋₃₁₂-(P1A-P2B)-(P1B-P2A)). Thus, to release the pentameric P-stalk complex, 80S ribosomal particles were isolated from P0TH199 yeast cells and treated with activated thrombin from bovine plasma (Sigma, # T4648-1KU) (1 U of thrombin/10 mg of 80S) for 12 h at 4 °C. In vitro P-stalk dephosphorylation was performed during the ribosome isolation procedure. Briefly, the ribosomes purified with the sucrose cushion were treated with AP according to the protocol applied for the preparation of Phos-tag samples described above. The wild-type and AP-dephosphorylated ribosomes were ultracentrifuged (100,000 × g) and resuspended in buffer (10 mM Tris pH 7.4, 150 mM NaCl, 10 mM MgCl₂), and the thrombin-cleaved P-stalk complexes were purified to homogeneity using affinity (Ni-NTA agarose from QIAGEN) and size-exclusion chromatography (Superose 12, 12/300HR column; GE Healthcare). The purity of the

ribosomal stalk complexes was verified by 12% SDS-PAGE, 10% native PAGE, and native mass spectrometry.

Native mass spectrometry of P-stalk complexes

All complexes were analyzed using a SYNAPT G2-Si High-Definition Mass Spectrometer (Waters, Manchester, UK). For the analysis, all protein solutions were buffer-exchanged into 200 mM ammonium acetate (pH 7.5) using Micro Bio-Spin chromatography columns (Bio-Rad). Aliquots of 2 μ l were introduced into the mass spectrometer via nanoflow capillaries, and the following conditions were applied: capillary voltage 1.2 kV, sampling cone 120 V, and source offset 20 V. The source temperature was set at 25 °C. The collision voltage was adjusted for an optimal signal level. Maximum entropy (MaxEnt, Waters) deconvolution was applied to the electrospray data to recalculate the gas phase-existing masses according to previously published results (Grela et al, 2010).

Microscale thermophoresis (MST)

Purified P-stalk complexes were fluorescently labeled according to the manufacturer's instructions with a Monolith His-Tag Labeling Kit RED-tris-NTA 2nd Generation kit (NanoTemper Technologies, GmbH, Germany). The concentration of the labeled P-stalk protein complexes was adjusted to 5 nM. A dilution series of unlabeled eIF5B protein was prepared in a final volume of 20 μ l for the MST measurements, starting from the highest 1.48 μ M concentration in PBS buffer with 0.05% Tween-20 and 1 mg/ml of bovine serum albumin. Thermophoresis was performed using a Monolith NT.115 Pico instrument (NanoTemper Technologies) at an ambient temperature of 25 °C and 1/14/1 s laser off/on/off cycles, respectively. The instrument parameters were adjusted to 100% excitation power and a medium MST detection range. The data were analyzed using NanoTemper NT analysis software (version 2.2.4) at the standard MST time of 30 s (thermophoresis plus T-jump). The K_d of the interaction was determined by fitting the data using fit algorithms implemented into the MST analytical software. To calculate the fraction bound, the ΔF_{norm} value of each point was divided by the amplitude of the fitted curve, resulting in values from 0 to 1 (0 = unbound, 1 = bound).

Assays of dipeptides and tripeptides and GTP hydrolysis

The rate of the first and second peptide bond formation was determined from time courses using the 80S initiation complex (ICs) prepared with the wild-type WT (phosphorylated) or dephospho-mimicking ribosomes (SDPM) as described previously (Ranjan et al, 2021). The integrity of ribosomal subunits was verified by mass-spectrometry analysis of ribosomal proteins (r-proteins), subjected to trypsin digestion (Appendix Table S1). To form the ternary complex, 8 μ M of eIF2 was incubated with 1 mM GTP and 4 μ M [^3H]Met-tRNA $^{\text{Met}}$ in YT buffer (30 mM HEPES-KOH pH 7.5, 100 mM KOAc, 3 mM MgCl $_2$) at 26 °C for 15 min. To form 80S IC, 2 μ M 40S, 10 μ M mRNA, 10 μ M eIF-mix (mixture of initiation factors eIF1, eIF1A, eIF3, eIF5), 2 mM DTT, 0.25 mM spermidine, and 1 mM GTP were incubated for 5 min at 26 °C

before adding 3 μ M 60S subunits and 6 μ M eIF5B, followed by addition of the ternary complex with simultaneous adjustment of the MgCl $_2$ concentration to 9 mM (YT9 buffer). 80S ICs were purified by ultracentrifugation through a 1.1 M sucrose cushion, and the pellets were dissolved in YT9 buffer. 80S ICs were diluted to 3 mM MgCl $_2$ prior to the experiments. Ternary complexes eEF1A-GTP-[^{14}C]Phe-tRNA $^{\text{Phe}}$ and eEF1A-GTP-[^{14}C]Val-tRNA $^{\text{Val}}$ were prepared by incubating 1 μ M eEF1A, 1 μ M eEF1B α , 3 mM PEP, 1% PK, 1 mM DTT, and 2 mM GTP in YT buffer for 15 min at 26 °C. 0.2 μ M [^{14}C]Phe-tRNA $^{\text{Phe}}$ or [^{14}C]Val-tRNA $^{\text{Val}}$ (5 eEF1A:1 aa-tRNA) was then added and incubated for additional 5 min at 26 °C, followed by addition of 2 μ M modified eIF5A. For each time point, 0.7 μ M of the pre-formed 80S IC with [^3H]Met-tRNA $^{\text{Met}}$ in the P site was rapidly mixed with 0.2 μ M eEF1A-GTP-[^{14}C]Phe-tRNA $^{\text{Phe}}$ in the quench-flow apparatus. After the desired incubation time, the reaction was quenched with 0.5 M KOH and the dipeptides were analyzed by Reverse-Phase High Performance Liquid Chromatography (HPLC) following [^3H] and [^{14}C] counts (Wohl-gemuth et al, 2008). For the analysis of tripeptide formation, 0.7 μ M of the initiation complexes ([^3H]Met-tRNA $^{\text{Met}}$) was mixed with 1 μ M eEF1A-GTP-Phe-tRNA $^{\text{Phe}}$ and incubated for 15 min at 26 °C to form dipeptides. The dipeptides were then mixed with 0.7 μ M eEF1A-GTP-[^{14}C]Val-tRNA $^{\text{Val}}$, 2 μ M eEF2, 4 μ M eEF3 and with 40 μ M ATP, in the quench-flow apparatus and the reaction was quenched with KOH after the desired time point. The samples were further processed as for the dipeptide analysis. The mRNA used for all analyses was purchased from IBA Lifesciences GmbH, with sequence as follows: 5'-GGUCUCUCUCUCUCUCUAUGUUU GUUUCUCUCUCUCUC-3'.

TrGTPases eIF5B, eEF1A, and eEF2 as well as ribosomes were purified as previously described (Blanchet and Ranjan, 2022). GTP hydrolysis reactions were performed by mixing 0.05 μ M GTPases with 10 μ M [$\gamma^{32}\text{P}$]GTP with 0.2 μ M wild-type ribosomes or ribosomes from the SDPM mutant strain mimicking the non-phosphorylated states of the ribosomes. The reactions were quenched with 50% formic acid (25% final). The amount of hydrolyzed [$\gamma^{32}\text{P}$] GTP was analyzed by thin-layer chromatography using Polygram CEL300 TLC plates (Gromadski and Rodnina, 2004). The [Pi/(Pi + GTP)] ratio was used to estimate the time dependence of the GTP hydrolysis.

Quantification of misincorporation errors using the dual-luciferase assay

The efficiency of misincorporation was measured using the dual-luciferase assay according to an established procedure (Harger and Dinman, 2003) with modifications described earlier (Wawiorka et al, 2017). Briefly, each yeast strain was transformed with a control plasmid or with appropriate vectors: for misincorporation, the pDB688 vector was used as a control; pDB868, with a near-cognate CGC $_{245}$ codon, having replaced a cognate CAC codon encoding the His amino acid at position 245 of the firefly luciferase with a CGC codon for Arg; all the genetic constructs were kindly provided by Dr David Bedwell (Salas-Marco and Bedwell, 2005). The pJD375 vector was also used as a control plasmid with cognate AGA at position 218 (coding Arg) of the firefly luciferase, and pJD643 with near-cognate AGC $_{218}$ coding Ser, provided as a kind gift by Dr Jonathan Dinman (Plant et al, 2007).

Determination of the yeast lifespan

The total yeast cell lifespan is calculated as the sum of the reproductive (i.e., time between the first and last budding) and post-reproductive (time after the last budding until death) lifespan. The total lifespan of the *S. cerevisiae* yeast was determined as previously described (Minois et al, 2005) with further modification (Molon and Zebrowski, 2017). A fresh logarithmic yeast culture ($OD_{600} = 0.6\text{--}0.8$) was collected and transferred to YPD plates with solid medium containing Phloxine B (10 $\mu\text{g/ml}$). During manipulation, the plates were kept at 28 °C for 15 h and at 4 °C during the night. The results represent measurements for at least 100 cells analyzed in at least three independent experiments. The analysis was performed by micromanipulation using the Nikon Eclipse E200 optical microscope with an attached micromanipulator. The budding lifespan of the yeast cells was assessed as follows: yeast cells were cultured on YPD medium plates after an overnight growth period using a micromanipulator. The budding lifespan was determined by microscopic examination, following a known procedure (Molon and Zebrowski, 2017). The number of buds produced by each mother cell served as an indicator of the budding lifespan. A minimum of 90 cells were meticulously examined in two independent experiments. The analysis was performed using a Nikon Eclipse E200 optical microscope (Nikon, Amsterdam, Netherlands) equipped with an attached micromanipulator. Statistical significances were assessed using ANOVA and Dunnett's post hoc test ($*p < 0.05$). Translational efficiency was measured as incorporation of ^{35}S -methionine, and the half-transit time was calculated according to previously described protocols (Wawiora et al, 2017).

Ribo-seq

For ribosome profiling, yeast cells (BY4741 and SDPM strains) were grown in synthetic complete medium supplemented with glucose (2%) and inositol (200 mM). 1 l of the medium was inoculated to OD_{600} of 0.06–0.08 from a saturated starting culture. The cells were grown to OD_{600} of 0.4–0.5 and treated with tunicamycin at a concentration of 2.5 $\mu\text{g/ml}$. The cells were harvested by fast filtration and frozen in liquid nitrogen. Isolation of ribosome-protected footprints and generation of libraries for Illumina sequencing are described in detail elsewhere (Shafieinouri et al, 2022). Analysis of RNA-seq and Ribo-seq Data was performed as follows: FastQC reports were employed to assess the quality of the raw data. Following this initial step, TrimGalore software was used to remove adapters and low-quality sequences. The Bowtie tool was applied to eliminate sequences corresponding to tRNA and rRNA based on genome reference data (RefSeq: GCF000146045.2). After the preprocessing steps, the FastQC tool was reapplied for quality control (www.bioinformatics.babraham.ac.uk/projects/fastqc/). Read length distribution plots were generated to analyze the distribution of read lengths in the Ribo-seq data. The trimmed data were then mapped to the genome using STAR, allowing for a tolerance of up to 4% mismatch in the read length. Feature Counts was employed to count uniquely mapped reads, utilizing gene annotations from the RefSeq database. Subsequently, read count normalization was carried out using the transcripts per million (TPM) method specifically for coding sequences (CDS). For the Ribo-seq data, codon analysis based on 21 nt and 28 nt ribosome

protected fragments (RPFs) and ribosome occupancy on CDS were assessed using RiboToolkit (Liu et al, 2020).

Molecular dynamics simulations

For MD simulation, initial structures were built using the UCSF ChimeraX tool (Goddard et al, 2018) as a linear chain or an α -helix. The MD simulations were performed using the GROMACS (Abraham et al, 2015) package with the CHARMM36 (Best et al, 2012) all-atom force field and the TIP3P (Jorgensen et al, 1983) water model. Each peptide was solvated in a periodic box with a 0.9 nm distance between the peptide and the box walls and neutralized using sodium ions. All the simulations were performed in the following steps: energy minimization was performed using the steepest descent algorithm until forces have converged to within the available machine precision; system heating was carried out for 500 ps, from 0 to 298 K, in NVT conditions (canonical ensemble); system equilibration was performed for 1000 ps, at 300 K and 1.0 bar, in NPT conditions; production was performed for at least 2 μs , with the time step set to 2 fs at constant pressure 1 bar and constant temperature 298 K. All the simulation analyses presented in this paper were performed using GROMACS package tools, the DSSP program, and the Python programming language. Structural visualization and graphical representation of data were rendered using the UCSF ChimeraX tool and Origin program [Origin, Version 2022b, OriginLab Corporation, Northampton, MA, USA]. Clustering was carried out using the Gromos clustering algorithm (Daura et al, 1999) with a 0.3 nm clustering cut-off. Three-dimensional Free Energy Landscapes (FELs) were generated by trajectory projection on eigenvector 1 and eigenvector 2 (PC1 and PC2) (Campos and Baptista, 2009).

Data availability

The RNA-seq and Ribo-seq data have been deposited in NCBI's SRA database (Sayers et al, 2022) and are accessible through BioProject accession number: PRJNA1071838 (<https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1071838>).

The source data of this paper are collected in the following database record: [biostudies:S-SCDT-10_1038-S44319-024-00297-1](https://www.ebi.ac.uk/biostudies/studies/S-SCDT-10_1038-S44319-024-00297-1).

Expanded view data, supplementary information, appendices are available for this paper at <https://doi.org/10.1038/s44319-024-00297-1>.

Peer review information

A peer review file is available at <https://doi.org/10.1038/s44319-024-00297-1>

References

- Bah A, Vernon RM, Siddiqui Z, Krzeminski M, Muhandiram R, Zhao C, Sonenberg N, Kay LE, Forman-Kay JD (2015) Folding of an intrinsically disordered protein by phosphorylation as a regulatory switch. *Nature* 519:106–109
- Belin S, Hacot S, Daudignon L, Therizols G, Pourpe S, Mertani HC, Rosa-Calatrava M, Diaz JJ (2010) Purification of ribosomes from human cell lines. *Curr Protoc Cell Biol* 49:3–40

- Bernado P, Modig K, Grela P, Svergun DI, Tchorzewski M, Pons M, Akke M (2010) Structure and dynamics of ribosomal protein L12: an ensemble model based on SAXS and NMR relaxation. *Biophys J* 98:2374–2382
- Best RB, Zhu X, Shim J, Lopes PE, Mittal J, Feig M, Mackerell AD Jr (2012) Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone phi, psi and side-chain chi(1) and chi(2) dihedral angles. *J Chem Theory Comput* 8:3257–3273
- Blanchet S, Ranjan N (2022) Translation phases in eukaryotes. *Methods Mol Biol* 2533:217–228
- Boeke JD, Trueheart J, Natsoulis G, Fink GR (1987) 5-Fluoroorotic acid as a selective agent in yeast molecular genetics. *Methods Enzymol* 154:164–175
- Boguszewska A, Tchorzewski M, Dukowski P, Winiarczyk S, Grankowski N (2002) Subcellular distribution of the acidic ribosomal P-proteins from *Saccharomyces cerevisiae* in various environmental conditions. *Biol Cell* 94:139–146
- Bohlen J, Roiuk M, Teleman AA (2021) Phosphorylation of ribosomal protein S6 differentially affects mRNA translation based on ORF length. *Nucleic acids Res* 49:13062–13074
- Campos SR, Baptista AM (2009) Conformational analysis in a multidimensional energy landscape: study of an arginylglutamate repeat. *J Phys Chem B* 113:15989–16001
- Datta PP, Sharma MR, Qi L, Frank J, Agrawal RK (2005) Interaction of the G' domain of elongation factor G and the C-terminal domain of ribosomal protein L7/L12 during translocation as revealed by cryo-EM. *Mol Cell* 20:723–731
- Daura X, Gademann K, Jaun B, Seebach D, van Gunsteren WF, Mark AE (1999) Peptide folding: when simulation meets experiment. *Angew Chem Int Ed* 38:236–240
- Davey NE, Van Roey K, Weatheritt RJ, Toedt G, Uyar B, Altenberg B, Budd A, Diella F, Dinkel H, Gibson TJ (2012) Attributes of short linear motifs. *Mol Biosyst* 8:268–281
- Davydov II, Wohlgemuth I, Artamonova II, Urlaub H, Tonevitsky AG, Rodnina MV (2013) Evolution of the protein stoichiometry in the L12 stalk of bacterial and organellar ribosomes. *Nat Commun* 4:1387
- Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, Chen S, Eddes J, Loevenich SN, Aebersold R (2006) The PeptideAtlas project. *Nucleic Acids Res* 34:D655–658
- Dever TE, Feng L, Wek RC, Cigan AM, Donahue TF, Hinnebusch AG (1992) Phosphorylation of initiation factor 2 alpha by protein kinase GCN2 mediates gene-specific translational control of GCN4 in yeast. *Cell* 68:585–596
- Diaconu M, Kothe U, Schlunzen F, Fischer N, Harms JM, Tonevitsky AG, Stark H, Rodnina MV, Wahl MC (2005) Structural basis for the function of the ribosomal L7/12 stalk in factor binding and GTPase activation. *Cell* 121:991–1004
- Dong J, Qiu H, Garcia-Barrio M, Anderson J, Hinnebusch AG (2000) Uncharged tRNA activates GCN2 by displacing the protein kinase moiety from a bipartite tRNA-binding domain. *Mol Cell* 6:269–279
- Fan X, Zhu Y, Wang C, Niu L, Teng M, Li X (2016) Structural insights into the interaction of the ribosomal P stalk protein P2 with a type II ribosome-inactivating protein ricin. *Sci Rep* 6:37803
- Frank J, Gonzalez Jr RL (2010) Structure and dynamics of a processive Brownian motor: the translating ribosome. *Annu Rev Biochem* 79:381–412
- Goddard TD, Huang CC, Meng EC, Pettersen EF, Couch GS, Morris JH, Ferrin TE (2018) UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci* 27:14–25
- Grela P, Bernado P, Svergun D, Kwiatowski J, Abramczyk D, Grankowski N, Tchorzewski M (2008a) Structural relationships among the ribosomal stalk proteins from the three domains of life. *J Mol Evol* 67:154–167
- Grela P, Gajda MJ, Armache JP, Beckmann R, Krokowski D, Svergun DI, Grankowski N, Tchorzewski M (2012) Solution structure of the natively assembled yeast ribosomal stalk determined by small-angle X-ray scattering. *Biochem J* 444:205–209
- Grela P, Helgstrand M, Krokowski D, Boguszewska A, Svergun D, Liljas A, Bernado P, Grankowski N, Akke M, Tchorzewski M (2007) Structural characterization of the ribosomal P1A-P2B protein dimer by small-angle X-ray scattering and NMR spectroscopy. *Biochemistry* 46:1988–1998
- Grela P, Krokowski D, Gordiyenko Y, Krowarsch D, Robinson CV, Otlewski J, Grankowski N, Tchorzewski M (2010) Biophysical properties of the eukaryotic ribosomal stalk. *Biochemistry* 49:924–933
- Grela P, Li XP, Horbowicz P, Dzwierzynska M, Tchorzewski M, Tumer NE (2017) Human ribosomal P1-P2 heterodimer represents an optimal docking site for ricin A chain with a prominent role for P1 C-terminus. *Sci Rep* 7:5608
- Grela P, Sawa-Makarska J, Gordiyenko Y, Robinson CV, Grankowski N, Tchorzewski M (2008b) Structural properties of the human acidic ribosomal P proteins forming the P1-P2 heterocomplex. *J Biochem* 143:169–177
- Grela P, Szajwaj M, Horbowicz-Drozdal P, Tchorzewski M (2019) How ricin damages the ribosome. *Toxins* 11:241
- Gressner AM, Wool IG (1974) The phosphorylation of liver ribosomal proteins in vivo. Evidence that only a single small subunit protein (S6) is phosphorylated. *J Biol Chem* 249:6917–6925
- Gromadski KB, Rodnina MV (2004) Kinetic determinants of high-fidelity tRNA discrimination on the ribosome. *Mol Cell* 13:191–200
- Gupta R, Hinnebusch AG (2023) Differential requirements for P stalk components in activating yeast protein kinase Gcn2 by stalled ribosomes during stress. *Proc Natl Acad Sci USA* 120:e2300521120
- Harding HP, Ordonez A, Allen F, Parts L, Inglis AJ, Williams RL, Ron D (2019) The ribosomal P-stalk couples amino acid starvation to GCN2 activation in mammalian cells. *Elife* 8(8):e50149
- Harger JW, Dinman JD (2003) An in vivo dual-luciferase assay system for studying translational recoding in the yeast *Saccharomyces cerevisiae*. *RNA* 9:1019–1024
- Hasler P, Brot N, Weissbach H, Parnassa AP, Elkon KB (1991) Ribosomal proteins PO, P1, and P2 are phosphorylated by casein kinase II at their conserved carboxyl termini. *J Biol Chem* 266:13815–13820
- Hollenstein DM, Gerecova G, Romanov N, Ferrari J, Veis J, Janschitz M, Beyer R, Schuller C, Ogris E, Hartl M et al (2021) A phosphatase-centric mechanism drives stress signaling response. *EMBO Rep* 22:e52476
- Holm M, Natchiar SK, Rundlet EJ, Myasnikov AG, Watson ZL, Altman RB, Wang HY, Taunton J, Blanchard SC (2023) mRNA decoding in human is kinetically and structurally distinct from bacteria. *Nature* 617(7959):200–207
- Horbowicz-Drozdal P, Kamel K, Kmiecik S, Borkiewicz L, Tumer NE, Shaw PC, Tchorzewski M, Grela P (2021) Phosphorylation of the conserved C-terminal domain of ribosomal P-proteins impairs the mode of interaction with plant toxins. *FEBS Lett* 595:2221–2236
- Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker AK (2004) The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res* 32:1037–1049
- Imai H, Abe T, Miyoshi T, Nishikawa SI, Ito K, Uchiyumi T (2018) The ribosomal stalk protein is crucial for the action of the conserved ATPase ABCE1. *Nucleic Acids Res* 46:7820–7830
- Inglis AJ, Masson GR, Shao S, Perisic O, McLaughlin SH, Hegde RS, Williams RL (2019) Activation of GCN2 by the ribosomal P-stalk. *Proc Natl Acad Sci USA* 116:4946–4954
- Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 79:926–935
- Kim KQ, Zaher HS (2022) Canary in a coal mine: collided ribosomes as sensors of cellular conditions. *Trends Biochem Sci* 47:82–97

- Kulczyk AW, Sorzano COS, Grela P, Tchorzewski M, Tumer NE, Li XP (2023) Cryo-EM structure of Shiga toxin 2 in complex with the native ribosomal P-stalk reveals residues involved in the binding interaction. *J Biol Chem* 299:102795
- Lee KM, Yu CW, Chiu TY, Sze KH, Shaw PC, Wong KB (2012) Solution structure of the dimerization domain of the eukaryotic stalk P1/P2 complex reveals the structural organization of eukaryotic stalk complex. *Nucleic Acids Res* 40:3172–3182
- Leijonmarck M, Liljas A (1987) Structure of the C-terminal domain of the ribosomal protein L7/L12 from *Escherichia coli* at 1.7 Å. *J Mol Biol* 195:555–579
- Letzring DP, Wolf AS, Brule CE, Grayhack EJ (2013) Translation of CGA codon repeats in yeast involves quality control components and ribosomal protein L1. *RNA* 19:1208–1217
- Liljas A, Sanyal S (2018) The enigmatic ribosomal stalk. *Q Rev Biophys* 51:e12
- Liu Q, Shvarts T, Sliz P, Gregory RI (2020) RiboToolkit: an integrated platform for analysis and annotation of ribosome profiling data to decode mRNA translation at codon resolution. *Nucleic Acids Res* 48:W218–W229
- Maniratanachote R, Minami K, Katoh M, Nakajima M, Yokoi T (2006) Dephosphorylation of ribosomal protein P0 in response to troglitazone-induced cytotoxicity. *Toxicol Lett* 166:189–199
- Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, Lindahl E (2015) GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 1–2:19–25
- Masson GR (2019) Towards a model of GCN2 activation. *Biochem Soc Trans* 47:1481–1488
- Matsuo Y, Tesina P, Nakajima S, Mizuno M, Endo A, Buschauer R, Cheng J, Shounai O, Ikeuchi K, Saeki Y et al (2020) RQT complex dissociates ribosomes collided on endogenous RQC substrate SDD1. *Nat Struct Mol Biol* 27:323–332
- Mazumder B, Sampath P, Seshadri V, Maitra RK, DiCorleto PE, Fox PL (2003) Regulated release of L13a from the 60S ribosomal subunit as a mechanism of transcript-specific translational control. *Cell* 115:187–198
- Minois N, Frajnt M, Wilson C, Vaupel JW (2005) Advances in measuring lifespan in the yeast *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA* 102:402–406
- Mishra P, Hosur RV (2015) Urea dependent (15)N NMR-relaxation studies on Pfp2 multimers reveal that the C-terminal behaves like an independent intrinsically disordered peptide. *Protein Pept Lett* 22:425–431
- Mochizuki M, Kitamyo M, Miyoshi T, Ito K, Uchiyumi T (2012) Analysis of chimeric ribosomal stalk complexes from eukaryotic and bacterial sources: structural features responsible for specificity of translation factors. *Genes Cells* 17:273–284
- Mohr D, Wintermeyer W, Rodnina MV (2002) GTPase activation of elongation factors Tu and G on the ribosome. *Biochemistry* 41:12520–12528
- Molon M, Zebrowski J (2017) Phylogenetic relationship and Fourier-transform infrared spectroscopy-derived lipid determinants of lifespan parameters in the *Saccharomyces cerevisiae* yeast. *FEMS Yeast Res* 17:fox031
- Murakami R, Singh CR, Morris J, Tang L, Harmon I, Takasu A, Miyoshi T, Ito K, Asano K, Uchiyumi T (2018) The interaction between the ribosomal stalk proteins and translation initiation factor 5B promotes translation initiation. *Mol Cell Biol* 38:e00067–18
- Nishi H, Hashimoto K, Panchenko AR (2011) Phosphorylation in protein-protein binding: effect on stability and function. *Structure* 19:1807–1815
- Nomura N, Honda T, Baba K, Naganuma T, Tanzawa T, Arisaka F, Noda M, Uchiyama S, Tanaka I, Yao M et al (2012) Archaeal ribosomal stalk protein interacts with translation factors in a nucleotide-independent manner via its conserved C terminus. *Proc Natl Acad Sci USA* 109:3748–3753
- Nussinov R, Tsai CJ, Xin F, Radivojac P (2012) Allosteric post-translational modification codes. *Trends Biochem Sci* 37:447–455
- Pan X, Whitten DA, Wilkerson CG, Pestka JJ (2014) Dynamic changes in ribosome-associated proteome and phosphoproteome during deoxynivalenol-induced translation inhibition and ribotoxic stress. *Toxicol Sci* 138:217–233
- Pan X, Whitten DA, Wu M, Chan C, Wilkerson CG, Pestka JJ (2013) Global protein phosphorylation dynamics during deoxynivalenol-induced ribotoxic stress response in the macrophage. *Toxicol Appl Pharm* 268:201–211
- Patil CK, Li H, Walter P (2004) Gcn4p and novel upstream activating sequences regulate targets of the unfolded protein response. *PLoS Biol* 2:E246
- Plant EP, Nguyen P, Russ JR, Pittman YR, Nguyen T, Quesinberry JT, Kinzy TG, Dinman JD (2007) Differentiating between near- and non-cognate codons in *Saccharomyces cerevisiae*. *PLoS ONE* 2:e517
- Ramirez M, Wek RC, Hinnebusch AG (1991) Ribosome association of GCN2 protein kinase, a translational activator of the GCN4 gene of *Saccharomyces cerevisiae*. *Mol Cell Biol* 11:3027–3036
- Ranjan N, Pochopien AA, Chih-Chien Wu C, Beckert B, Blanchet S, Green R, M VR, Wilson DN (2021) Yeast translation elongation factor eEF3 promotes late stages of tRNA translocation. *EMBO J* 40:e106449
- Rodnina MV, Savelsbergh A, Katunin VI, Wintermeyer W (1997) Hydrolysis of GTP by elongation factor G drives tRNA movement on the ribosome. *Nature* 385:37–41
- Salas-Marco J, Bedwell DM (2005) Discrimination between defects in elongation fidelity and termination efficiency provides mechanistic insights into translational readthrough. *J Mol Biol* 348:801–815
- Sattlegger E, Hinnebusch AG (2000) Separate domains in GCN1 for binding protein kinase GCN2 and ribosomes are required for GCN2 activation in amino acid-starved cells. *EMBO J* 19:6622–6633
- Savelsbergh A, Mohr D, Wilden B, Wintermeyer W, Rodnina MV (2000) Stimulation of the GTPase activity of translation elongation factor G by ribosomal protein L7/12. *J Biol Chem* 275:890–894
- Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Connor R, Funk K, Kelly C, Kim S et al (2022) Database resources of the national center for biotechnology information. *Nucleic Acids Res* 50:D20–D26
- Shafieinouri M, Membrano BS, Wu CC (2022) High-resolution ribosome profiling for determining ribosome functional states during translation elongation. *Methods Mol Biol* 2428:173–186
- Shao Y, Molestak E, Su W, Stankevicius M, Tchorzewski M (2022) Sordarin—an anti-fungal antibiotic with a unique modus operandi. *Br J Pharm* 179:1125–1145
- Shoemaker BA, Portman JJ, Wolynes PG (2000) Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. *Proc Natl Acad Sci USA* 97:8868–8873
- Simsek D, Barna M (2017) An emerging role for the ribosome as a nexus for post-translational modifications. *Curr Opin Cell Biol* 45:92–101
- Sinha NK, McKenney C, Yeow ZY, Li JJ, Nam KH, Yaron-Barir TM, Johnson JL, Huntsman EM, Cantley LC, Ordureau A et al (2024) The ribotoxic stress response drives UV-mediated cell death. *Cell* 187(14):3652–3670.e40
- Tanzawa T, Kato K, Girodat D, Ose T, Kumakura Y, Wieden HJ, Uchiyumi T, Tanaka I, Yao M (2018) The C-terminal helix of ribosomal P stalk recognizes a hydrophobic groove of elongation factor 2 in a novel fashion. *Nucleic Acids Res* 46:3232–3244
- Tchorzewski M (2002) The acidic ribosomal P proteins. *Int J Biochem Cell Biol* 34:911–915
- Tchorzewski M, Boguszczyńska A, Abramczyk D, Grankowski N (1999) Overexpression in *Escherichia coli*, purification, and characterization of recombinant 60S ribosomal acidic proteins from *Saccharomyces cerevisiae*. *Protein Expr Purif* 15:40–47

- Tchorzewski M, Boguszewska A, Dukowski P, Grankowski N (2000a) Oligomerization properties of the acidic ribosomal P-proteins from *Saccharomyces cerevisiae*: effect of P1A protein phosphorylation on the formation of the P1A-P2B hetero-complex. *Biochim Biophys Acta* 1499:63–73
- Tchorzewski M, Boldyreff B, Issinger O, Grankowski N (2000b) Analysis of the protein-protein interactions between the human acidic ribosomal P-proteins: evaluation by the two hybrid system. *Int J Biochem Cell Biol* 32:737–746
- Tchorzewski M, Krokowski D, Boguszewska A, Liljas A, Grankowski N (2003) Structural characterization of yeast acidic ribosomal P proteins forming the P1A-P2B heterocomplex. *Biochemistry* 42:3399–3408
- Too PH, Ma MK, Mak AN, Wong YT, Tung CK, Zhu G, Au SW, Wong KB, Shaw PC (2009) The C-terminal fragment of the ribosomal P protein complexed to trichosanthin reveals the interaction between the ribosome-inactivating protein and the ribosome. *Nucleic Acids Res* 37:602–610
- Tran DM, Ishiwata-Kimata Y, Mai TC, Kubo M, Kimata Y (2019) The unfolded protein response alongside the diauxic shift of yeast cells and its involvement in mitochondria enlargement. *Sci Rep* 9:12780
- Uchiyumi T, Honma S, Nomura T, Dabbs ER, Hachimori A (2002) Translation elongation by a hybrid ribosome in which proteins at the GTPase center of the *Escherichia coli* ribosome are replaced with rat counterparts. *J Biol Chem* 277:3857–3862
- van der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, Fuxreiter M, Gough J, Gsponer J, Jones DT et al (2014) Classification of intrinsically disordered regions and proteins. *Chem Rev* 114:6589–6631
- Vard C, Guillot D, Bargis P, Lavergne JP, Reboud JP (1997) A specific role for the phosphorylation of mammalian acidic ribosomal protein P2. *J Biol Chem* 272:20259–20262
- Walter P, Ron D (2011) The unfolded protein response: from stress pathway to homeostatic regulation *Science* 334:1081–1086
- Wawiora L, Molestak E, Szajwaj M, Michalec-Wawiora B, Boguszewska A, Borkiewicz L, Liudkovska V, Kufel J, Tchorzewski M (2016) Functional analysis of the uL11 protein impact on translational machinery *Cell Cycle* 15:1060–1072
- Wawiora L, Molestak E, Szajwaj M, Michalec-Wawiora B, Molon M, Borkiewicz L, Grela P, Boguszewska A, Tchorzewski M (2017) Multiplication of ribosomal P-stalk proteins contributes to the fidelity of translation. *Mol Cell Biol* 37:e00060–00017
- Wek SA, Zhu S, Wek RC (1995) The histidyl-tRNA synthetase-related sequence in the eIF-2 alpha protein kinase GCN2 interacts with tRNA and is required for activation in response to starvation for different amino acids. *Mol Cell Biol* 15:4497–4506
- Wilson KS, Noller HF (1998) Molecular movement inside the translational engine. *Cell* 92:337–349
- Wohlgemuth I, Brenner S, Beringer M, Rodnina MV (2008) Modulation of the rate of peptidyl transfer on the ribosome by the nature of substrates. *J Biol Chem* 283:32229–32235
- Wohlgemuth I, Pohl C, Rodnina MV (2010) Optimization of speed and accuracy of decoding in translation. *EMBO J* 29:3701–3709
- Wu CC, Peterson A, Zinshteyn B, Regot S, Green R (2020) Ribosome collisions trigger general stress responses to regulate cell fate. *Cell* 182:404–416.e414
- Wu CC, Zinshteyn B, Wehner KA, Green R (2019) High-resolution ribosome profiling defines discrete ribosome elongation states and translational regulation during cellular stress. *Mol Cell* 73:959–970.e955
- Yang L, Lee KM, Yu CW, Imai H, Choi AK, Banfield DK, Ito K, Uchiyumi T, Wong KB (2022) The flexible N-terminal motif of uL11 unique to eukaryotic ribosomes interacts with P-complex and facilitates protein translation. *Nucleic Acids Res* 50:5335–5348
- Zambrano R, Briones E, Remacha M, Ballesta JP (1997) Phosphorylation of the acidic ribosomal P proteins in *Saccharomyces cerevisiae*: a reappraisal. *Biochemistry* 36:14439–14446
- Zampieri S, Degen W, Ghiradello A, Doria A, van Venrooij WJ (2001) Dephosphorylation of autoantigenic ribosomal P proteins during Fas-L induced apoptosis: a possible trigger for the development of the autoimmune response in patients with systemic lupus erythematosus. *Ann Rheum Dis* 60:72–76
- Zhang L, Zhang D, Liu D, Li Y, Li H, Xie Y, Wang Z, Hansen BO, Olsson S (2020) Conserved eukaryotic kinase CK2 chaperone intrinsically disordered protein interactions. *Appl Environ Microbiol* 86:e02191–19
- Zhu S, Wek RC (1998) Ribosome-binding domain of eukaryotic initiation factor-2 kinase GCN2 facilitates translation control. *J Biol Chem* 273:1808–1814
- Zinker S, Warner JR (1976) The ribosomal proteins of *Saccharomyces cerevisiae*. Phosphorylated and exchangeable proteins. *J Biol Chem* 251:1799–1807

Acknowledgements

We are grateful to Emilia Łabuć for excellent technical assistance in data processing. We thank Tessa Hübner, Olaf Geintzer, and Thomas Shulz for expert technical assistance. This work was supported by grants from the National Science Center in Poland [UMO-2018/29/B/NZ1/01728] to MT; 2018/30/E/NZ1/00605 to DK; 2022/45/B/NZ3/02353 to PG, as well as by the Leibniz Prize and SFB 1565 (project number 469281184) of the Deutsche Forschungsgemeinschaft and the Max Planck Society to MVR.

Author contributions

Kamil Filipek: Investigation; Visualization; Methodology; Writing—original draft; Writing—review and editing. **Sandra Blanchet**: Investigation; Visualization; Methodology; Writing—original draft; Writing—review and editing. **Eliza Molestak**: Investigation; Methodology; Writing—review and editing. **Monika Zaciura**: Investigation; Methodology; Writing—review and editing. **Colin Chih-Chien Wu**: Data curation; Investigation. **Patrycja Horbowicz-Drożdżał**: Investigation; Visualization; Writing—review and editing. **Przemysław Grela**: Data curation; Funding acquisition; Investigation; Visualization; Methodology; Writing—original draft; Writing—review and editing. **Mateusz Zalewski**: Software; Investigation; Visualization; Methodology; Writing—original draft; Writing—review and editing. **Sebastian Kmiecik**: Data curation; Investigation; Methodology; Writing—original draft; Writing—review and editing. **Alan González-Ibarra**: Investigation. **Dawid Krokowski**: Data curation; Funding acquisition; Writing—review and editing. **Przemysław Łatoch**: Methodology; Writing—review and editing. **Agata Starosta**: Methodology; Writing—review and editing. **Mateusz Mołoń**: Investigation; Writing—review and editing. **Yutian Shao**: Investigation. **Lidia Borkiewicz**: Investigation. **Barbara Michalec-Wawiora**: Investigation; Visualization; Writing—review and editing. **Leszek Wawiora**: Conceptualization; Data curation; Writing—original draft; Writing—review and editing. **Konrad Kubiński**: Investigation; Writing—review and editing. **Katarzyna Socala**: Resources. **Piotr Wlaz**: Resources. **Kyle W Cunningham**: Conceptualization; Writing—original draft; Writing—review and editing. **Rachel Green**: Conceptualization; Data curation; Writing—original draft; Writing—review and editing. **Marina V Rodnina**: Conceptualization; Data curation; Funding acquisition; Writing—original draft; Writing—review and editing. **Marek Tchorzewski**: Conceptualization; Data curation; Supervision; Funding acquisition; Visualization; Writing—original draft; Project administration; Writing—review and editing.

Source data underlying figure panels in this paper may have individual authorship assigned. Where available, figure panel/source data authorship is listed in the following database record: [biostudies:S-SCDT-10_1038-S44319-024-00297-1](https://www.ebi.ac.uk/biostudies/studies/S-SCDT-10_1038-S44319-024-00297-1).

Disclosure and competing interests statement

The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this

licence, visit <http://creativecommons.org/licenses/by/4.0/>. Creative Commons Public Domain Dedication waiver <http://creativecommons.org/public-domain/zero/1.0/> applies to the data associated with this article, unless otherwise stated in a credit line to the data, but does not extend to the graphical or creative elements of illustrations, charts, or figures. This waiver removes legal barriers to the re-use and mining of research data. According to standard scholarly practice, it is recommended to provide appropriate citation and attribution whenever technically possible.

© The Author(s) 2024

Expanded View Figures

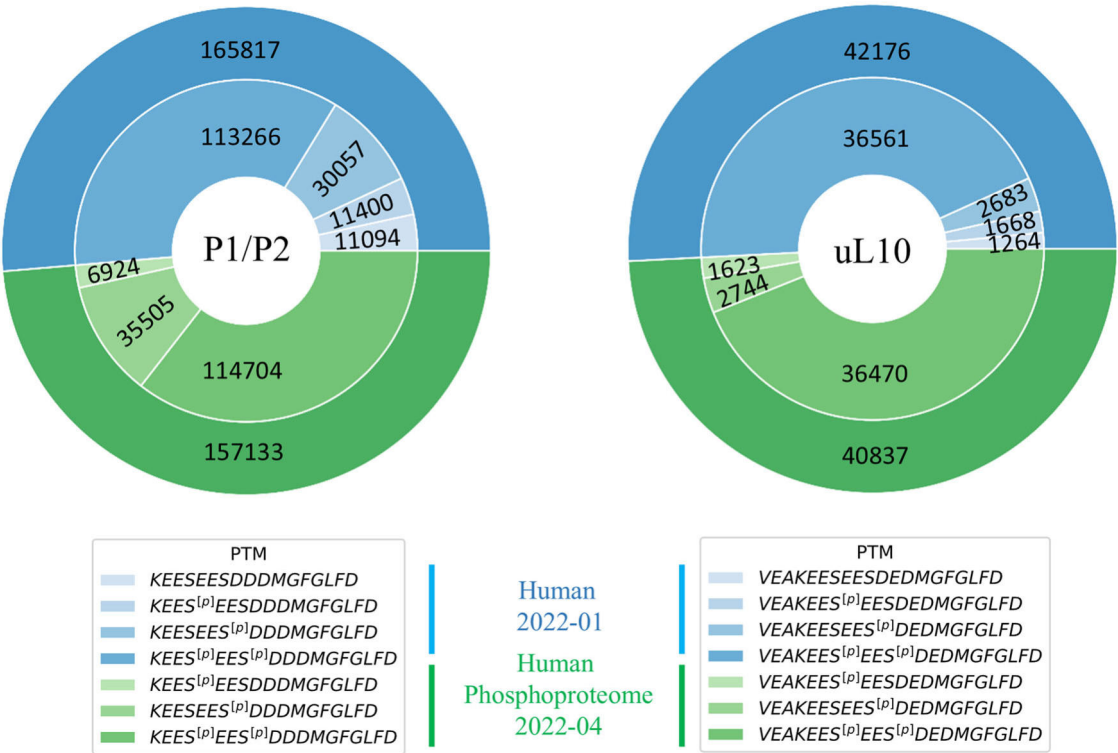


Figure EV1. Meta-analysis of CTD P-stalk protein phosphorylation in mammalian cells using mass spectrometry data implemented in the PeptideAtlas database.

The nested pie charts show the identified peptides of the P1/P2 (left chart) and uL10 (right chart) proteins. The upper, blue part of the plots represents the observation from the entire Human Proteome 2022-01 assembly. The lower, green part of the plots shows the results from the Human Phosphoproteome 2022-04 build. The Human Phosphoproteome 2022-04 contains only phosphorylated P1/P2 and uL10 proteins. The complete Human Proteome 2022-01 includes the observation for non-phosphorylated P-stalk proteins. The total number of peptide observations is shown in the outer ring of the plots. The inner ring shows the number of results with different patterns of serine residue phosphorylation. The higher the number of observations, the darker the color (blue or green). The lower panel shows the identified peptides within the two datasets, 2022-01 and 2022-04 builds, in the corresponding colors; phosphorylation site is marked with [p].

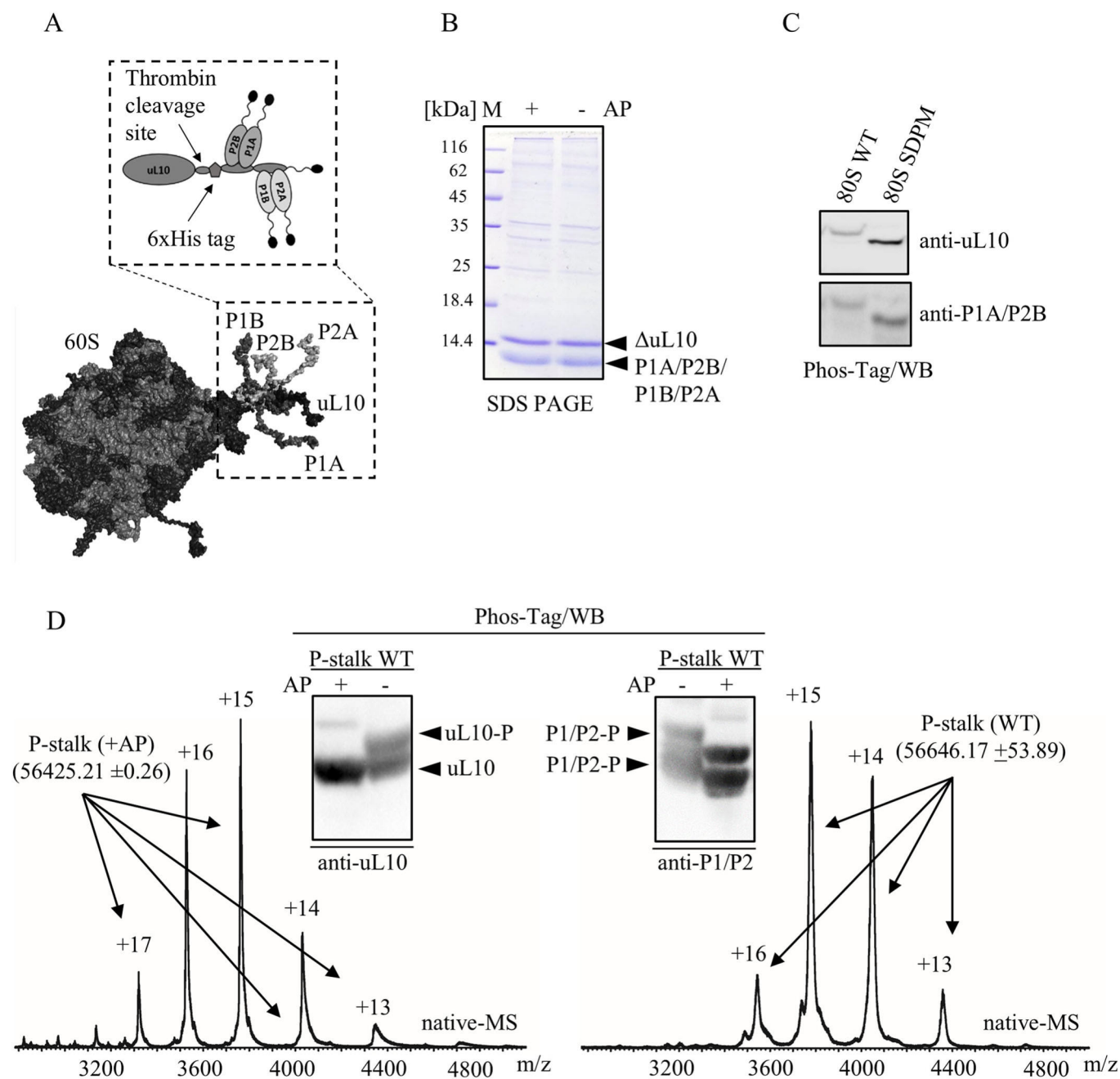


Figure EV2. Purification of phosphorylated (native) and unphosphorylated (AP-treated) P-stalk complexes and 80S ribosomal particles from WT and SDPM mutant strains.

(A) Schematic representation of the genetically engineered ribosomal particles with the P-stalk scheme. Specific thrombin cleavage site and 6xHis-tag used for P-stalk release and subsequent purification are indicated with arrows. (B) SDS-PAGE analysis of purified native (-AP) and dephosphorylated (+AP) stalk complexes. The positions of ΔuL10 and P1A/P2B/P1B/P2A are pointed with arrows. (C) Phos-tag analysis of purified 80S ribosomal particles; 80S WT - ribosomes from WT strain, 80S SDPM - ribosomes from SDPM mutant strain harboring S to A mutations within all P-protein; anti-uL10 and anti-P1A/P2B antibodies were used. (D) Native mass spectrometry analysis (native-MS) of purified intact P-stalk complexes; left and right panels, MS spectra of the complexes of WT and +AP, respectively; numbers next to the peaks indicate the charge states of the complexes; molecular masses of the P-stalk complexes were calculated by MaxEnt deconvolution software (Waters), and are 56646 and 56425 Da for WT and +AP complexes, respectively; insert - phos-tag analysis of purified stalk complexes, WT and +AP. The complexes were analyzed by SDS-PAGE/phos-tag/WB and subsequently proteins were detected with specific antibodies against uL10 and P1/P2 proteins. The positions of phospho- and dephospho-forms are indicated with arrows.

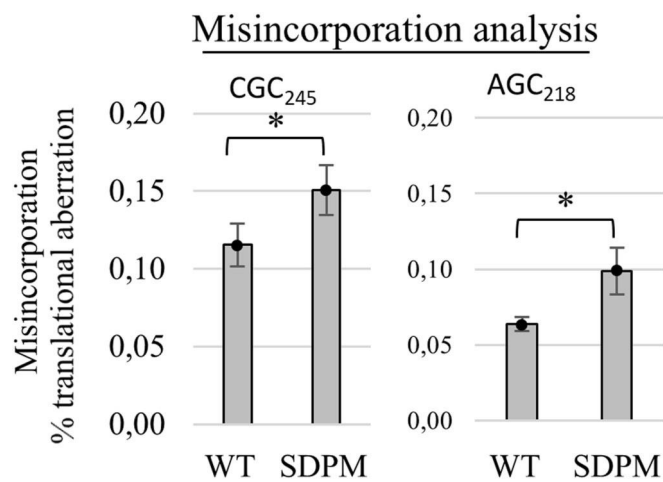


Figure EV3. Misincorporation analysis using a dual-luciferase reporter assay.

CGC₂₄₅ and AGC₂₁₈ describe near-cognate codons at positions 245 and 218 of the firefly reporter enzyme. All data are presented as the percentage of translational aberration; error bars, standard deviations ($n = 3$, technical replicates); $*p < 0.05$ by Student's t -test is indicated by asterisks.

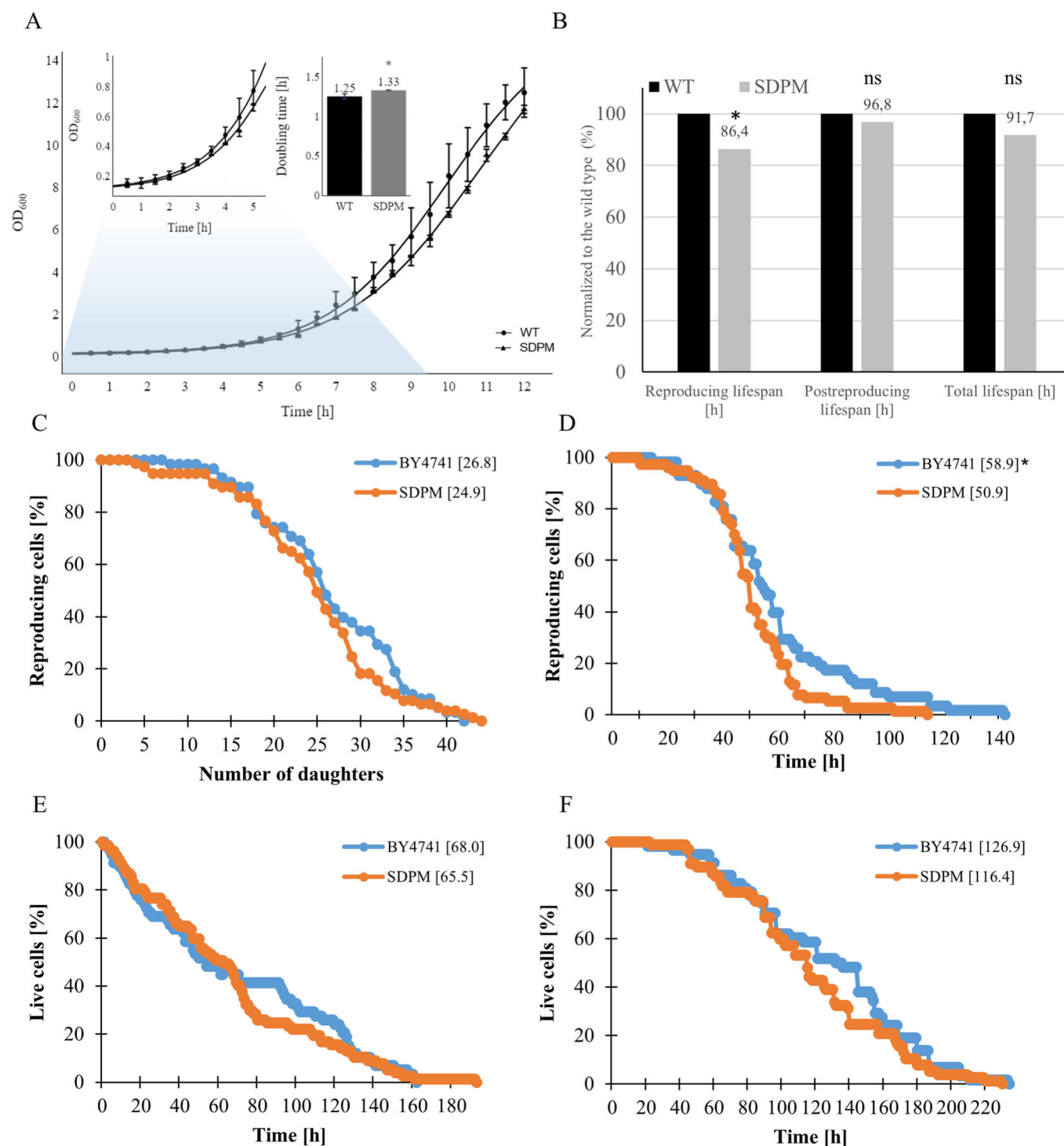


Figure EV4. Cellular fitness of the WT and SDPM yeast strains.

(A) Growth curves for WT and SDPM strains. The exponential fragment of growth curves used for doubling time calculation and average doubling time for WT and SDPM are shown in insets. The statistical analysis was done using one-tailed t-Welch test ($p < 0.05$), data are presented as mean $\pm$ SEM of $n = 3$, technical replicates. (B) Yeast lifespan analysis of WT and SDPM strains on the single cell level; the values were normalized to the WT referred as 100%, using data presented in (D-F). To assess differences between the WT and SDPM strains, one-way ANOVA and Dunnett's post hoc tests were used ($p < 0.05$), ns not significant. Comparison of the reproductive potential (A), reproductive lifespan (B), post-reproductive lifespan (C), and total lifespan (D) of the haploid reference yeast strain BY4741 (wild-type—WT) and the mutant SDPM strain. Statistical significances were assessed using ANOVA and Dunnett's post hoc test ($p < 0.05$). The mean value for a total of 90 cells from two independent experiments is shown in parentheses.

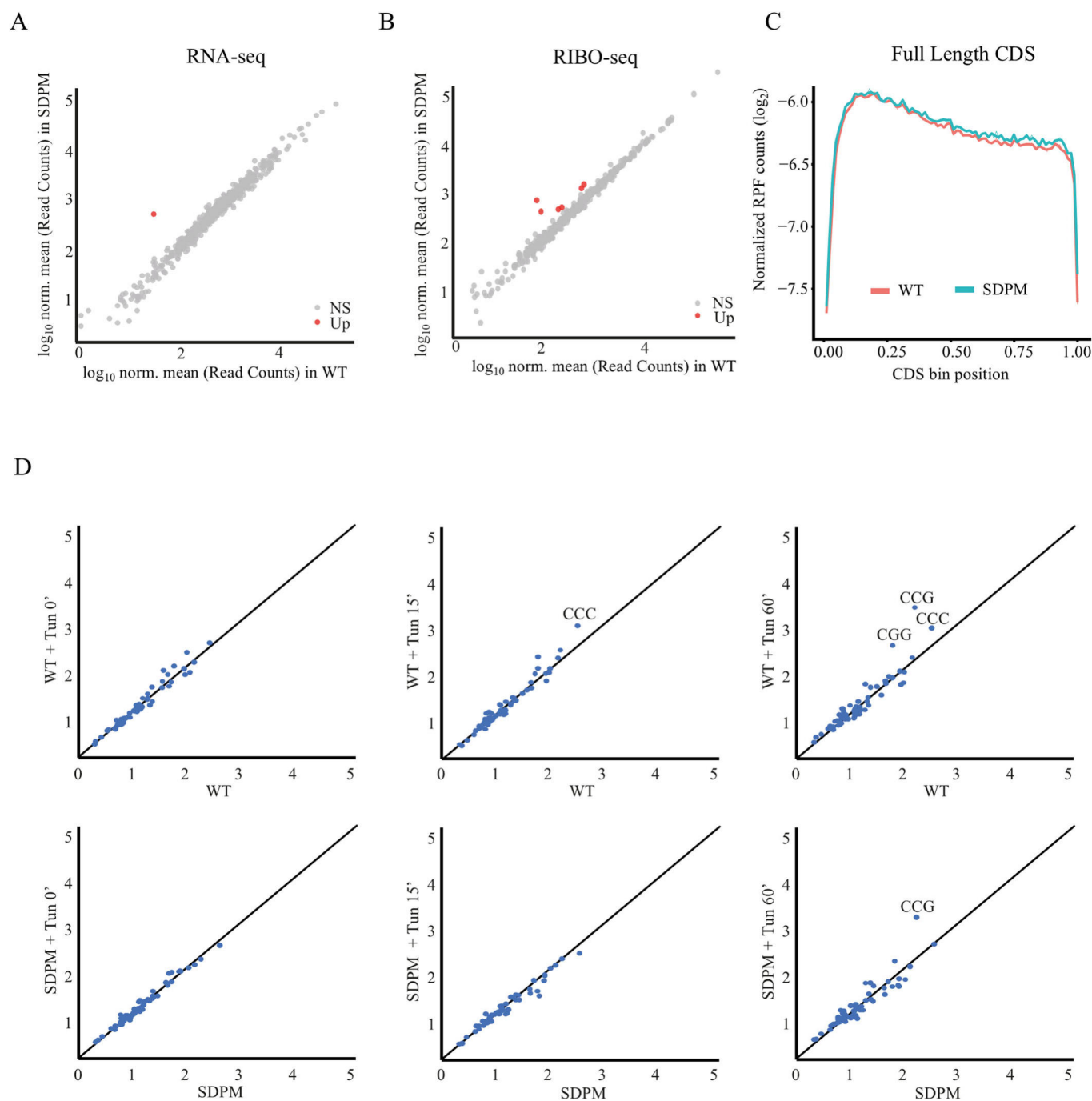


Figure EV5. The ribosomal footprints distribution obtained based on a genome-wide scale analysis of WT and SDPM yeast strains.

(A, B) Cross-correlation of gene expression analyses at the level of transcription (RNA-seq - A) and translation (RIBO-seq - B) for WT vs SDPM mutant strains; (C) average ribosome occupancy from all genes aligned from start to stop codons within the coding sequence (CDS) for WT (red line) and SDPM mutant strain (blue line); ribosome occupancy was normalized to show a mean value of 1 for each codon; the footprint occupancy was shown for the whole CDS. (D) The correlation of specific codon occupancies within 28 nt RPF. Specific overrepresented codons, CGG for arginine and CCG, CCC for proline are indicated.

CZEŚĆ III: PODSUMOWANIE I WNIOSKI

7. Podsumowanie wyników i wnioski końcowe

Na koniec chciałbym podsumować całą pracę - podkreślić najważniejsze osiągnięcia, wskazać możliwe kierunki dalszych badań oraz dodać kilka słów od siebie.

Podczas doktoratu pracowałem nad wieloma zagadnieniami - zarówno nad tworzeniem i publicznym udostępnianiem narzędzi wspierających projektowanie nowych leków, jak i nad komputerowym wsparciem badań *in vitro* z wykorzystaniem istniejących metod. Niezależnie od tematu przyświecał mi ten sam cel: tworzyć narzędzia, które nie tylko opisują rzeczywistość molekularną, ale pozwalają ją aktywnie modelować, projektować i interpretować - z myślą o zastosowaniach w biologii strukturalnej, inżynierii białek i projektowaniu leków nowej generacji.

Wspólnym mianownikiem wszystkich prac są białka oraz ich oddziaływania. Wybór tego kierunku badań doktorskich był motywowany przekonaniem, że leki białkowe stanowią ważną część przyszłości medycyny - pomimo świadomości ich ograniczeń i złożoności modelowania interakcji. To przekonanie motywowało rozwój metod prezentowanych w rozprawie. Moim głównym celem było adresowanie konkretnych problemów i ograniczeń utrudniających projektowanie leków białkowych.

Artykuły [A1-A4] dotyczą dokowania białko-białko i białko-peptyd. W artykule [A1] przedstawiona została strategia dokowania białko-białko z wykorzystaniem narzędzia CABS-dock (pierwotnie przeznaczonego do dokowania białko-peptyd), pozwalająca na pełną swobodę konformacyjną jednego z partnerów białkowych oraz ograniczoną elastyczność drugiego. W artykule [A2] zaprezentowany został protokół dokowania cyklicznych peptydów do białek, umożliwiający szerokie próbkowanie ich konformacji przy jednoczesnym zachowaniu cyklicznej struktury. Temat ten jest szczególnie interesujący ze względu na rosnące znaczenie peptydów cyklicznych, które stanowią odpowiedź na główne ograniczenia peptydów liniowych - zwykle dyskwalifikujących je jako kandydatów na leki^{90,133,134}. Peptydy o zamkniętej topologii łączą w sobie wysoką stabilność strukturalną, ograniczoną ruchliwość konformacyjną oraz zwiększoną odporność na degradację proteolityczną, co czyni je obiecującą klasą cząsteczek terapeutycznych. W artykule [A3] zaproponowano alternatywne wykorzystanie dynamiki molekularnej do oceny interakcji białko-peptyd, które przetestowano jako metodę walidacji wyników CABS-dock. Zagadnienie to ma szczególne znaczenie, ponieważ etap oceny modeli jest zazwyczaj najsłabszym ogniwem programów do dokowania peptydów: choć podczas próbkowania powstają poprawne konformacje, często zostają one odrzucone w trakcie filtrowania i wyboru najlepszego rozwiązania.

Artykuły [A1-A3] dotyczyły rozwoju klasycznych metod dokowania. W trakcie trwania mojego doktoratu cała dziedzina doświadczyła rewolucji związanej z pojawieniem się metod

predykcji struktur opartych na sztucznej inteligencji. Początkowo dotyczyły one pojedynczych łańcuchów, a z czasem również całych kompleksów, w tym kompleksów białko-peptyd. Moim wkładem w ten nurt był artykuł [A4], w którym zaproponowany został protokół wykorzystania ESMFold do dokowania peptydów do białek. Otrzymane wyniki nie wskazują, aby narzędzie to mogło zastąpić najbardziej zaawansowane rozwiązania z rodziny AlphaFold, jednak jego zdolność do szybkiego generowania sensownych modeli otwiera zupełnie nowy sposób pracy - umożliwia szybki przegląd dużych zbiorów sekwencji, formułowanie wstępnych hipotez strukturalnych oraz poszukiwanie nietypowych trybów wiązania, które mogą pozostać niewykryte w analizach z użyciem bardziej złożonych i kosztownych obliczeniowo metod. W szczególności pokazuje to, że modele językowe oparte wyłącznie na sekwencji mogą dostarczać użytecznych wskazówek strukturalnych także w kontekście interakcji białko-peptyd, co czyni je wartościowym elementem przyszłych podejść konsensusowych i hybrydowych łączących szybkość AI z precyzją metod klasycznych.

W tym miejscu czuję się w obowiązku skomentować wartość wyników przedstawionych w [A1-A3] w odniesieniu do metod opartych na sztucznej inteligencji. Czy te badania są nieaktualne? Zdecydowanie nie. Choć metody AI osiągnęły w ostatnich latach spektakularne sukcesy - również w generowaniu alternatywnych konformacji i pierwszych próbach wykorzystania ich do analizy dynamiki^{43,135} - wciąż mają one istotne ograniczenia. Modele te najczęściej koncentrują się na pojedynczych, „najbardziej prawdopodobnych” predykcjach, podczas gdy problem tworzenia bogatych i różnorodnych zbiorów struktur pozostaje domeną klasycznych metod próbkowania przestrzeni konformacyjnej, takich jak te opisane w [A1-A3]. Dlatego wśród przyszłych kierunków badań widzę nie tylko dalszy rozwój narzędzi AI (np. dostosowanie ESMFold do kompleksów z udziałem peptydów cyklicznych), ale również kontynuację prac nad metodami klasycznymi. Co więcej, szczególnie obiecujące wydaje się sprzężenie obu podejść - poprzez strategie konsensusowe lub integrację informacji *a priori* pochodzących z modeli językowych (np. ESMFold) z tradycyjnymi symulacjami próbkowania i dynamiki.

Artykuł [A5] stanowi przegląd poświęcony metodzie CABS-flex oraz jej zastosowaniom w modelowaniu elastyczności białek w kontekście ich funkcji biologicznych. Publikacja ta była punktem wyjścia do opracowania nowej wersji - CABS-flex 3.0 - zaprezentowanej w artykule [A6]. Metoda CABS-flex w ostatnich latach zyskała szerokie uznanie w środowisku naukowym, co wynikało m.in. z dynamicznego rozwoju metod sztucznej inteligencji oraz rosnącej liczby wysokiej jakości modeli białek generowanych przez algorytmy takie jak AlphaFold czy ESMFold. Wersja 2.0 była powszechnie wykorzystywana i często cytowana, co jednoznacznie potwierdza jej znaczenie w badaniach nad elastycznością struktur białkowych. CABS-flex 3.0 stanowi wyraźny krok naprzód - to narzędzie łączące wysoką efektywność obliczeniową z bogatą funkcjonalnością, które realnie wspiera proces badawczy: od formułowania pierwszych hipotez po szczegółowe analizy strukturalne. W nowej wersji wprowadzono szereg usprawnień, obejmujących m.in. zwiększoną jakość odwzorowania lokalnych struktur, nowy interfejs użytkownika umożliwiający łatwą analizę różnych aspektów dynamiki oraz bezpośrednie wykorzystanie danych eksperymentalnych (np. restrykcyjnych odległościowych lub profili fluktuacji). Dzięki tym modyfikacjom użytkownik może w

większym stopniu kontrolować przebieg symulacji i interpretację wyników, co czyni metodę bardziej elastyczną i interaktywną. W rezultacie można oczekiwać, że wersja 3.0 zyska jeszcze większą popularność i znajdzie zastosowanie w coraz szerszym zakresie badań biologicznych.

W artykule [A7] przedstawiona została nowa wersja narzędzia Aggrescan3D (A3D) - Aggrescan4D (A4D). Narzędzie A3D umożliwia przewidywanie skłonności białek do agregacji w oparciu o ich strukturę trójwymiarową oraz giętkość, natomiast A4D rozszerza tę koncepcję o możliwość analizy wpływu pH. Artykuł [A8] stanowi publikację o szerszym zakresie, podsumowującą i integrującą rozwiązania opracowane w ramach narzędzi A3D i A4D. Uważam, że narzędzie A4D stanowi ważny krok naprzód w analizie zjawiska agregacji białek. Integracja informacji strukturalnych, dynamicznych i środowiskowych w jednym protokole otwiera nowe możliwości - pozwala nie tylko dokładniej przewidywać skłonność do agregacji, ale także lepiej zrozumieć mechanizmy leżące u jej podstaw. Moim zdaniem A4D ma potencjał, by stać się narzędziem pierwszego wyboru zarówno w badaniach podstawowych, jak i aplikacyjnych - od projektowania stabilniejszych białek terapeutycznych po analizy biofizyczne dużych zbiorów danych strukturalnych. Warto podkreślić, że dzięki swojej dostępności i łatwości użycia serwer już teraz cieszy się dużym zainteresowaniem i jest aktywnie wykorzystywany w środowisku naukowym. Od momentu jego uruchomienia, w ciągu pierwszego roku działania, wykonano ponad 12 tysięcy predykcji, co potwierdza rosnące zapotrzebowanie na narzędzia umożliwiające analizę agregacji białek w sposób zależny od warunków środowiskowych.

Artykuł [A9] dotyczy roli fosforylacji białek szypuły rybosomalnej. Kluczowym wynikiem, uzyskanym dzięki długim symulacjom dynamiki molekularnej, było pokazanie, że fosforylacja przeprogramowuje elastyczne fragmenty domeny C-końcowej, stabilizując charakterystyczną konformację w kształcie litery U. Stabilizacja ta wynika z powstawania sieci oddziaływań elektrostatycznych pomiędzy resztami fosfoseryny a lizynami. Mechanizm ten nie był uchwytny w danych eksperymentalnych, a symulacje pozwoliły po raz pierwszy zrozumieć, w jaki sposób modyfikacje potranslacyjne wpływają na zdolność tych fragmentów do oddziaływania z innymi elementami rybosomu i czynnikami regulacyjnymi. Badania okazały się na tyle obiecujące, że obecnie są kontynuowane - we współpracy z grupą prof. Marka Tchórzewskiego z Uniwersytetu Marii Curie-Skłodowskiej w Lublinie rozszerzamy je z poziomu pojedynczych fragmentów na cały kompleks białkowy, aby powiązać lokalne efekty fosforylacji ze zmianami w architekturze zespołu i dostępnością miejsc wiązania dla czynników translacyjnych. Wstępne wyniki są bardzo obiecujące i pokazują, że połączenie klasycznych metod symulacyjnych z analizą strukturalną otwiera nowe możliwości interpretacji danych oraz dostarcza wglądu niedostępnego dla tradycyjnych technik eksperymentalnych. Współpraca ta pokazała również, jak wiele aspektów funkcjonowania szypuły rybosomalnej pozostaje jeszcze do odkrycia i jak duży potencjał tkwi w integracji metod obliczeniowych i eksperymentalnych.

Naturalnym pytaniem, które pojawia się po zakończeniu tego etapu badań, jest to, w jakim kierunku będzie rozwijać się dziedzina modelowania struktur i oddziaływań białkowych w najbliższych latach. Odpowiedź na to pytanie nie jest prosta, ponieważ obszar ten znajduje się

w fazie wyjątkowo dynamicznych przemian. Pojawienie się narzędzi opartych na sztucznej inteligencji, przede wszystkim modeli z rodziny AlphaFold^{4,10,32}, całkowicie zmieniło sposób, w jaki podchodzi się do przewidywania struktur i analizy interakcji międzycząsteczkowych. W trakcie trwania moich badań doktoranckich obserwowałem, jak metody, które jeszcze niedawno uchodziły za nowoczesne, zostały w dużej mierze zastąpione przez rozwiązania AI. Wymagało to ciągłej adaptacji i elastyczności, co dobrze oddaje charakter tej dziedziny. Początkowo modele uczenia maszynowego pozwalały przewidywać jedynie statyczne struktury pojedynczych białek, dziś potrafią już opisywać złożone kompleksy, wiązania ligandów, a nawet przewidywać wpływ mutacji czy modyfikacji potranslacyjnych^{136–138}. Postęp jest bardzo szybki, a kierunek dalszego rozwoju trudny do jednoznacznego przewidzenia. W mojej ocenie do najważniejszych obecnie kierunków badań należą:

- **Wiarygodne modelowanie wielu stanów konformacyjnych i pełnej dynamiki układów białkowych.**

Obecne modele AI, mimo bardzo wysokiej dokładności przewidywania struktur, nadal opisują głównie pojedyncze, statyczne konformacje. Tymczasem białka funkcjonują jako dynamiczne układy, których aktywność często zależy od przejść pomiędzy wieloma stanami energetycznymi. Kluczowym kierunkiem rozwoju jest więc nie tylko łączenie metod opartych na sztucznej inteligencji z klasycznymi i uproszczonymi modelami symulacyjnymi, ale również rozwój nowych architektur AI ukierunkowanych na modelowanie dynamiki - takich jak AlphaFlow^{139,140} czy inne podejścia integrujące informacje o czasie, ruchliwości i krajobrazie energetycznym białek^{43,110,141}. Celem tych prac jest wiarygodne odtwarzanie dróg przejść konformacyjnych, mechanizmów allosterycznych oraz pełnych krajobrazów energetycznych cząsteczek. Takie podejście pozwoli lepiej zrozumieć zależność między strukturą, fluktuacjami i funkcją białek, a w konsekwencji przyczyni się do tworzenia bardziej realistycznych modeli procesów biologicznych.

- **Uwzględnianie warunków fizjologicznych, takich jak pH, siła jonowa, błony czy obecność kofaktorów.**

Większość obecnych predykcji wykonywana jest w warunkach uproszczonych, nieodzwierciedlających rzeczywistego środowiska komórkowego. Włączenie do modeli czynników fizykochemicznych - takich jak zmienne pH, obecność jonów, błon lipidowych, redoksu czy efektu tłoku molekularnego - stanowi klucz do zbliżenia predykcji do realiów *in vivo*. Szczególnie ważne jest też uwzględnienie wpływu modyfikacji potranslacyjnych i kofaktorów metalicznych, które mogą istotnie zmieniać stabilność i funkcję białek.

- **Dokładne przewidywanie powinowactwa i energii wiązania w interakcjach białko-białko i białko-peptyd.**

Choć najnowsze modele AI potrafią z dużą precyzją przewidzieć geometrię kompleksów, wciąż nie dostarczają wiarygodnych ocen energii wiązania ani nie pozwalają skutecznie rozróżniać alternatywnych pól dokowania.^{142,143} Rozwój nowych funkcji scoringowych, łączących elementy fizykochemiczne z reprezentacjami uczenia

głębokiego, jest niezbędny, by móc wiarygodnie oceniać stabilność i specyficzność interakcji. Postęp w tym zakresie będzie kluczowy zarówno dla zrozumienia regulacji oddziaływań międzybiałkowych, jak i dla projektowania skutecznych inhibitorów peptydowych oraz leków białkowych.

- **Integracja danych eksperymentalnych z modelami AI.**

Coraz wyraźniej rysuje się potrzeba łączenia przewidywań obliczeniowych z danymi pochodzącymi z metod eksperymentalnych, takich jak krioelektronowa mikroskopia (cryo-EM), spektrometria masowa z cross-linkingiem (XL-MS), wymiana deuterowa (HDX-MS) czy rozpraszanie promieni rentgenowskich w małym kącie (SAXS).¹⁴⁴ Integracja tych źródeł informacji pozwoli tworzyć spójne, multimodalne modele strukturalne, które będą nie tylko zgodne z obserwacjami, ale również umożliwią ich interpretację na poziomie molekularnym. W przyszłości dane eksperymentalne mogą też stanowić bezpośrednie ograniczenia wykorzystywane podczas uczenia modeli AI.

- **Projektowanie funkcji białek, a nie jedynie ich struktur.**

Kolejnym krokiem po opanowaniu przewidywania struktur jest możliwość aktywnego projektowania białek o określonych właściwościach funkcjonalnych. Nowe modele generatywne już teraz pozwalają tworzyć stabilne struktury *de novo*¹⁴⁵, jednak pełne projektowanie funkcji - takich jak kataliza, specyficzność wiązania czy aktywność regulatorowa - wymaga połączenia predykcji struktury, dynamiki i energetyki z wiedzą biologiczną. W perspektywie najbliższej dekady można spodziewać się rozwoju platform umożliwiających automatyczne projektowanie kompleksów białkowych o z góry zaplanowanej funkcji biologicznej.

Dziedzina bioinformatyki strukturalnej rozwija się w zawrotnym tempie, a utrzymanie konkurencyjności wymaga nieustannego śledzenia nowych osiągnięć i gotowości do ich wdrażania w praktyce. Dlatego najważniejsze jest pozostawanie na bieżąco, otwartość na zmiany i umiejętność łączenia tradycyjnych metod z najnowszymi rozwiązaniami sztucznej inteligencji.

Wierzę, że przedstawione w tej rozprawie rozwiązania stanowią solidny fundament dla dalszego rozwoju, a ich otwartość oraz praktyczna użyteczność przełożą się na realne wsparcie dla biologii strukturalnej i projektowania leków białkowych. Najcenniejsza lekcja, jaką wynoszę z tej pracy, to przekonanie, że prawdziwy postęp rodzi się z integracji metod, wiedzy i współpracy - i właśnie tam chcę rozwijać swoje dalsze badania.

Bibliografia

1. Fosgerau K, Hoffmann T (2015) Peptide therapeutics: current status and future directions. *Drug Discovery Today* 20:122–128.
2. Ciemny M, Kurcinski M, Kamel K, Kolinski A, Alam N, Schueler-Furman O, Kmiecik S (2018) Protein-peptide docking: opportunities and challenges. *Drug Discov Today* 23:1530–1537.
3. Li H, Nithin C, Kmiecik S, Huang S-Y (2025) Computational methods for modeling protein–protein interactions in the AI era: Current status and future directions. *Drug Discovery Today* 30:104382.
4. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Židek A, Potapenko A, et al. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* 596:583–589.
5. Peng C-X, Liang F, Xia Y-H, Zhao K-L, Hou M-H, Zhang G-J (2024) Recent Advances and Challenges in Protein Structure Prediction. *J. Chem. Inf. Model.* 64:76–95.
6. Yang Z, Zeng X, Zhao Y, Chen R (2023) AlphaFold2 and its applications in the fields of biology and medicine. *Sig Transduct Target Ther* 8:1–14.
7. Bertoline LMF, Lima AN, Krieger JE, Teixeira SK (2023) Before and after AlphaFold2: An overview of protein structure prediction. *Front. Bioinform.* [Internet] 3. Available from: <https://www.frontiersin.org/journals/bioinformatics/articles/10.3389/fbinf.2023.1120370/full>
8. Li MQC, Wang S, Lin S-R, Ting LEN, Wan Z-H, Xie G, Zhang J (2025) Advantages and Limitations of AlphaFold in Structural Biology: Insights from Recent Studies. *Protein J* [Internet]. Available from: <https://doi.org/10.1007/s10930-025-10310-8>
9. Terwilliger TC, Liebschner D, Croll TI, Williams CJ, McCoy AJ, Poon BK, Afonine PV, Oeffner RD, Richardson JS, Read RJ, et al. (2024) AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. *Nat Methods* 21:110–116.
10. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, Ronneberger O, Willmore L, Ballard AJ, Bambrick J, et al. (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 630:493–500.
11. Zheng W, Wuyun Q, Li Y, Zhang C, Freddolino L, Zhang Y (2024) Improving deep learning protein monomer and complex structure prediction using DeepMSA2 with huge metagenomics data. *Nat Methods* 21:279–289.
12. Vander Meersche Y, Diharce J, Gelly J-C, Galochkina T (2025) Flexibility or uncertainty? A critical assessment of AlphaFold 2 pLDDT. *Structure* 33:2157-2163.e2.
13. Bryant P, Pozzati G, Elofsson A (2022) Improved prediction of protein-protein interactions using AlphaFold2. *Nat Commun* 13:1265.

14. Lin B, Luo X, Liu Y, Jin X (2024) A comprehensive review and comparison of existing computational methods for protein function prediction. *Brief Bioinform* 25:bbae289.
15. Morris R, Black KA, Stollar EJ (2022) Uncovering protein function: from classification to complexes. *Essays Biochem* 66:255–285.
16. Billings E, Lundquist K, Overly C, Srinivasan K, Noinaj N (2021) Structure Determination of Membrane Proteins Using X-Ray Crystallography. *Methods Mol Biol* 2302:101–136.
17. Puthenveetil R, Vinogradova O (2019) Solution NMR: A powerful tool for structural and functional studies of membrane proteins in reconstituted environments. *J Biol Chem* 294:15914–15931.
18. Lyumkis D (2019) Challenges and opportunities in cryo-EM single-particle analysis. *J Biol Chem* 294:5181–5197.
19. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Research* 28:235–242.
20. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ (2009) AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry* 30:2785–2791.
21. Lyskov S, Gray JJ (2008) The RosettaDock server for local protein–protein docking. *Nucleic Acids Res* 36:W233–W238.
22. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, et al. (2004) Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* 47:1739–1749.
23. de Vries SJ, van Dijk M, Bonvin AMJJ (2010) The HADDOCK web server for data-driven biomolecular docking. *Nat Protoc* 5:883–897.
24. Comeau SR, Gatchell DW, Vajda S, Camacho CJ (2004) ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics* 20:45–50.
25. Kurcinski M, Jamroz M, Blaszczyk M, Kolinski A, Kmiecik S (2015) CABS-dock web server for the flexible docking of peptides to proteins without prior knowledge of the binding site. *Nucleic Acids Research* 43:W419–W424.
26. Blaszczyk M, Kurcinski M, Kouza M, Wieteska L, Debinski A, Kolinski A, Kmiecik S (2016) Modeling of protein–peptide interactions using the CABS-dock web server for binding site search and flexible docking. *Methods* 93:72–83.
27. Kurcinski M, Pawel Ciemny M, Oleniecki T, Kuriata A, Badaczewska-Dawid AE, Kolinski A, Kmiecik S (2019) CABS-dock standalone: a toolbox for flexible protein–peptide docking. *Bioinformatics* 35:4170–4172.
28. Kolinski A (2004) Protein modeling and structure prediction with a reduced representation. *Acta Biochim Pol* 51:349–371.

29. Blaszczyk M, Ciemny MP, Kolinski A, Kurcinski M, Kmiecik S (2019) Protein-peptide docking using CABS-dock and contact information. *Brief Bioinform* 20:2299–2305.
30. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Židek A, Nelson AWR, Bridgland A, et al. (2020) Improved protein structure prediction using potentials from deep learning. *Nature* 577:706–710.
31. Pereira J, Simpkin AJ, Hartmann MD, Rigden DJ, Keegan RM, Lupas AN (2021) High-accuracy protein structure prediction in CASP14. *Proteins: Structure, Function, and Bioinformatics* 89:1687–1699.
32. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, Židek A, Bates R, Blackwell S, Yim J, et al. (2022) Protein complex prediction with AlphaFold-Multimer. :2021.10.04.463034. Available from: <https://www.biorxiv.org/content/10.1101/2021.10.04.463034v2>
33. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, et al. (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 379:1123–1130.
34. Nussinov R, Zhang M, Liu Y, Jang H (2022) AlphaFold, Artificial Intelligence (AI), and Allostery. *J Phys Chem B* 126:6372–6383.
35. Agarwal V, McShan AC (2024) The power and pitfalls of AlphaFold2 for structure prediction beyond rigid globular proteins. *Nat Chem Biol* 20:950–959.
36. Saldaño T, Escobedo N, Marchetti J, Zea DJ, Mac Donagh J, Velez Rueda AJ, Gonik E, García Melani A, Novomisky Nechcoff J, Salas MN, et al. (2022) Impact of protein conformational diversity on AlphaFold predictions. *Bioinformatics* 38:2742–2748.
37. Chakravarty D, Lee M, Porter LL (2025) Proteins with alternative folds reveal blind spots in AlphaFold-based protein structure prediction. *Current Opinion in Structural Biology* 90:102973.
38. Guan X, Tang Q-Y, Ren W, Chen M, Wang W, Wolynes PG, Li W (2024) Predicting protein conformational motions using energetic frustration analysis and AlphaFold2. *Proceedings of the National Academy of Sciences* 121:e2410662121.
39. Manfredi M, Vazzana G, Savojardo C, Martelli PL, Casadio R (2025) AlphaFold2 and ESMFold: A large-scale pairwise model comparison of human enzymes upon Pfam functional annotation. *Comput Struct Biotechnol J* 27:461–466.
40. Rosignoli S, Pacelli M, Manganiello F, Paiardini A (2024) An outlook on structural biology after AlphaFold: tools, limits and perspectives. *FEBS Open Bio* 15:202–222.
41. Niazi SK, Mariam Z, Paracha RZ (2024) Limitations of Protein Structure Prediction Algorithms in Therapeutic Protein Development. *BioMedInformatics* 4:98–112.
42. Nam K, Wolf-Watz M (2023) Protein dynamics: The future is bright and complicated! *Struct Dyn* 10:014301.

43. Cui X, Ge L, Chen X, Lv Z, Wang S, Zhou X, Zhang G (2025) Beyond static structures: protein dynamic conformations modeling in the post-AlphaFold era. *Brief Bioinform* 26:bbaf340.
44. Hu G, Doruker P, Li H, Demet Akten E (2021) Editorial: Understanding Protein Dynamics, Binding and Allostery for Drug Design. *Front. Mol. Biosci.* [Internet] 8. Available from: <https://www.frontiersin.org/journals/molecular-biosciences/articles/10.3389/fmolb.2021.681364/full>
45. Du X, Li Y, Xia Y-L, Ai S-M, Liang J, Sang P, Ji X-L, Liu S-Q (2016) Insights into Protein–Ligand Interactions: Mechanisms, Models, and Methods. *Int J Mol Sci* 17:144.
46. Lu W, Zhang J, Huang W, Zhang Z, Jia X, Wang Z, Shi L, Li C, Wolynes PG, Zheng S (2024) DynamicBind: predicting ligand-specific protein-ligand complex structure with a deep equivariant generative model. *Nat Commun* 15:1071.
47. Ahmed M, Maldonado AM, Durrant JD (2023) From byte to bench to bedside: molecular dynamics simulations and drug discovery. *BMC Biol* 21:299.
48. Schmidt C, Robinson CV (2014) Dynamic protein ligand interactions – insights from MS. *FEBS J* 281:1950–1964.
49. Pang X, Zhou H-X (2017) Rate Constants and Mechanisms of Protein–Ligand Binding. *Annu Rev Biophys* 46:105–130.
50. Yang L-Q, Sang P, Tao Y, Fu Y-X, Zhang K-Q, Xie Y-H, Liu S-Q (2014) Protein dynamics and motions in relation to their functions: several case studies and the underlying mechanisms. *J Biomol Struct Dyn* 32:372–393.
51. Haran G, Riven I (2023) Perspective: How Fast Dynamics Affect Slow Function in Protein Machines. *J. Phys. Chem. B* 127:4687–4693.
52. Peng C, Zhang X, Xu Z, Chen Z, Yang Y, Cai T, Zhu W (2022) D3PM: a comprehensive database for protein motions ranging from residue to domain. *BMC Bioinformatics* 23:70.
53. Kuriata A, Gierut AM, Oleniecki T, Ciemny MP, Kolinski A, Kurcinski M, Kmiecik S (2018) CABS-flex 2.0: a web server for fast simulations of flexibility of protein structures. *Nucleic Acids Res* 46:W338–W343.
54. Wróblewski K, Kmiecik S (2024) Integrating AlphaFold pLDDT Scores into CABS-flex for enhanced protein flexibility simulations. *Computational and Structural Biotechnology Journal* 23:4350–4356.
55. Badaczewska-Dawid A, Wróblewski K, Kurcinski M, Kmiecik S (2024) Structure prediction of linear and cyclic peptides using CABS-flex. *Brief Bioinform* 25:bbae003.
56. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA (2004) Development and testing of a general amber force field. *J Comput Chem* 25:1157–1174.
57. Huang J, MacKerell Jr AD (2013) CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *Journal of Computational Chemistry* 34:2135–2145.

58. van Gunsteren WF, Daura X, Mark AE GROMOS Force Field. In: Encyclopedia of Computational Chemistry. John Wiley & Sons, Ltd; 2002. Available from: <https://onlinelibrary.wiley.com/doi/abs/10.1002/0470845015.cga011>
59. Jorgensen WL, Maxwell DS, Tirado-Rives J (1996) Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* 118:11225–11236.
60. Hollingsworth SA, Dror RO (2018) Molecular Dynamics Simulation for All. *Neuron* 99:1129–1143.
61. Araki M, Matsumoto S, Bekker G-J, Isaka Y, Sagae Y, Kamiya N, Okuno Y (2021) Exploring ligand binding pathways on proteins using hypersound-accelerated molecular dynamics. *Nat Commun* 12:2793.
62. Kappel K, Miao Y, McCammon JA (2015) Accelerated Molecular Dynamics Simulations of Ligand Binding to a Muscarinic G-protein Coupled Receptor. *Q Rev Biophys* 48:479–487.
63. Martínez L, Sonoda MT, Webb P, Baxter JD, Skaf MS, Polikarpov I (2005) Molecular Dynamics Simulations Reveal Multiple Pathways of Ligand Dissociation from Thyroid Hormone Receptors. *Biophysical Journal* 89:2011–2023.
64. Nemchinova M, Schuurman-Wolters GK, Whittaker JJ, Arkhipova V, Marrink SJ, Poolman B, Guskov A (2024) Exploring the Ligand Binding and Conformational Dynamics of the Substrate-Binding Domain 1 of the ABC Transporter GlnPQ. *J. Phys. Chem. B* 128:7822–7832.
65. Delaforge E, Milles S, Huang J, Bouvier D, Jensen MR, Sattler M, Hart DJ, Blackledge M (2016) Investigating the Role of Large-Scale Domain Dynamics in Protein-Protein Interactions. *Front. Mol. Biosci.* [Internet] 3. Available from: <https://www.frontiersin.org/journals/molecular-biosciences/articles/10.3389/fmolb.2016.00054/full>
66. Paggi JM, Pandit A, Dror RO (2024) The Art and Science of Molecular Docking. *Annual Review of Biochemistry* 93:389–410.
67. Meng X-Y, Zhang H-X, Mezei M, Cui M (2011) Molecular Docking: A powerful approach for structure-based drug discovery. *Curr Comput Aided Drug Des* 7:146–157.
68. Cournia Z, Allen B, Sherman W (2017) Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *J. Chem. Inf. Model.* 57:2911–2937.
69. Pantsar T, Poso A (2018) Binding Affinity via Docking: Fact and Fiction. *Molecules* 23:1899.
70. Sulimov VB, Kutov DC, Sulimov AV (2019) Advances in Docking. *Curr Med Chem* 26:7555–7580.
71. Jamroz M, Kolinski A, Kmiecik S (2013) CABS-flex: server for fast simulation of protein structure fluctuations. *Nucleic Acids Res* 41:W427–W431.

72. Kurcinski M, Oleniecki T, Ciemny MP, Kuriata A, Kolinski A, Kmiecik S (2019) CABS-flex standalone: a simulation environment for fast modeling of protein flexibility. *Bioinformatics* 35:694–695.
73. Fernandez-Escamilla A-M, Rousseau F, Schymkowitz J, Serrano L (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotechnol* 22:1302–1306.
74. Tartaglia GG, Vendruscolo M (2008) The Zyggregator method for predicting protein aggregation propensities. *Chem Soc Rev* 37:1395–1401.
75. Conchillo-Solé O, de Groot NS, Avilés FX, Vendrell J, Daura X, Ventura S (2007) AGGRESCAN: a server for the prediction and evaluation of “hot spots” of aggregation in polypeptides. *BMC Bioinformatics* 8:65.
76. De Baets G, Van Durme J, van der Kant R, Schymkowitz J, Rousseau F (2015) Solubis: optimize your protein. *Bioinformatics* 31:2580–2582.
77. Sormanni P, Aprile FA, Vendruscolo M (2015) The CamSol method of rational design of protein mutants with enhanced solubility. *J Mol Biol* 427:478–490.
78. Kuriata A, Iglesias V, Pujols J, Kurcinski M, Kmiecik S, Ventura S (2019) Aggrescan3D (A3D) 2.0: prediction and engineering of protein solubility. *Nucleic Acids Research* 47:W300–W307.
79. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L (2005) The FoldX web server: an online force field. *Nucleic Acids Res* 33:W382–388.
80. Lensink MF, Velankar S, Wodak SJ (2017) Modeling protein-protein and protein-peptide complexes: CAPRI 6th edition. *Proteins* 85:359–377.
81. Lensink MF, Nadzirin N, Velankar S, Wodak SJ (2020) Modeling protein-protein, protein-peptide, and protein-oligosaccharide complexes: CAPRI 7th edition. *Proteins* 88:916–938.
82. Lensink MF, Velankar S, Kryshtafovych A, Huang S-Y, Schneidman-Duhovny D, Sali A, Segura J, Fernandez-Fuentes N, Viswanath S, Elber R, et al. (2016) Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: A CASP-CAPRI experiment. *Proteins* 84 Suppl 1:323–348.
83. Desta IT, Porter KA, Xia B, Kozakov D, Vajda S (2020) Performance and Its Limits in Rigid Body Protein-Protein Docking. *Structure* 28:1071–1081.e3.
84. Christoffer C, Kihara D (2022) Domain-Based Protein Docking with Extremely Large Conformational Changes. *J Mol Biol* 434:167820.
85. Vreven T, Moal IH, Vangone A, Pierce BG, Kastitis PL, Torchala M, Chaleil R, Jiménez-García B, Bates PA, Fernandez-Recio J, et al. (2015) Updates to the Integrated Protein–Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *Journal of Molecular Biology* 427:3031–3041.

86. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJE, Vajda S, Vakser I, Wodak SJ (2003) CAPRI: A Critical Assessment of PRedicted Interactions. *Proteins: Structure, Function, and Bioinformatics* 52:2–9.
87. Wang H, S. Dawber R, Zhang P, Walko M, J. Wilson A, Wang X (2021) Peptide-based inhibitors of protein–protein interactions: biophysical, structural and cellular consequences of introducing a constraint. *Chemical Science* 12:5977–5993.
88. Dougherty PG, Qian Z, Pei D (2017) Macrocycles as Protein-Protein Interaction Inhibitors. *Biochem J* 474:1109–1125.
89. Joo SH (2012) Cyclic Peptides as Therapeutic Agents and Biochemical Tools. *Biomol Ther (Seoul)* 20:19–26.
90. Merz ML, Habeshian S, Li B, David J-AGL, Nielsen AL, Ji X, Il Khwildy K, Duany Benitez MM, Phothirath P, Heinis C (2024) De novo development of small cyclic peptides that are orally bioavailable. *Nat Chem Biol* 20:624–633.
91. Ding Y, Lambden E, Peate J, Picken LJ, Rees TW, Perez-Ortiz G, Newgas SA, Spicer LAR, Hicks T, Hess J, et al. (2024) Rapid Peptide Cyclization Inspired by the Modular Logic of Nonribosomal Peptide Synthetases. *J. Am. Chem. Soc.* 146:16787–16801.
92. Tao H, Zhao X, Zhang K, Lin P, Huang S-Y (2022) Docking cyclic peptides formed by a disulfide bond through a hierarchical strategy. *Bioinformatics* 38:4109–4116.
93. Charitou V, van Keulen SC, Bonvin AMJJ (2022) Cyclization and Docking Protocol for Cyclic Peptide–Protein Modeling Using HADDOCK2.4. *J. Chem. Theory Comput.* 18:4027–4040.
94. Zhang Y, Sanner MF (2019) Docking flexible cyclic peptides with AutoDock CrankPep. *J Chem Theory Comput* 15:5161–5168.
95. Moore BL, Kelley LA, Barber J, Murray JW, MacDonald JT (2013) High–quality protein backbone reconstruction from alpha carbons using Gaussian mixture models. *Journal of Computational Chemistry* 34:1881–1889.
96. London N, Raveh B, Cohen E, Fathi G, Schueler-Furman O (2011) Rosetta FlexPepDock web server--high resolution modeling of peptide-protein interactions. *Nucleic Acids Res* 39:W249-253.
97. Zhang C, Zhang C, Shang T, Zhu N, Wu X, Duan H (2024) HighFold: accurately predicting structures of cyclic peptides and complexes with head-to-tail and disulfide bridge constraints. *Briefings in Bioinformatics* 25:bbae215.
98. Webb B, Sali A (2016) Comparative Protein Structure Modeling Using MODELLER. *Current Protocols in Bioinformatics* 54:5.6.1-5.6.37.
99. Meng Y, Zhang Z, Zhou C, Tang X, Hu X, Tian G, Yang J, Yao Y (2025) Protein structure prediction via deep learning: an in-depth review. *Front Pharmacol* 16:1498662.
100. AlQuraishi M (2021) Machine learning in protein structure prediction. *Current Opinion in Chemical Biology* 65:1–8.

101. Balasco N, Esposito L, Vitagliano L (2025) Structural Biology in the AlphaFold Era: How Far Is Artificial Intelligence from Deciphering the Protein Folding Code? *Biomolecules* 15:674.
102. Chaaban S, Ratkevičiūtė G, Lau C (2024) AI told you so: navigating protein structure prediction in the era of machine learning. *Biochem (Lond)* 46:7–12.
103. Meng Y, Zhang Z, Zhou C, Tang X, Hu X, Tian G, Yang J, Yao Y (2025) Protein structure prediction via deep learning: an in-depth review. *Front. Pharmacol.* [Internet] 16. Available from: <https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2025.1498662/full>
104. Johansson-Åkhe I, Wallner B (2022) Improving peptide-protein docking with AlphaFold-Multimer using forced sampling. *Front Bioinform* 2:959160.
105. Manshour N, Ren JZ, Esmaili F, Bergstrom E, Xu D (2024) Comprehensive Evaluation of AlphaFold-Multimer, AlphaFold3 and ColabFold, and Scoring Functions in Predicting Protein-Peptide Complex Structures. :2024.11.11.622992. Available from: <https://www.biorxiv.org/content/10.1101/2024.11.11.622992v1>
106. Tsaban T, Varga JK, Avraham O, Ben-Aharon Z, Khrumushin A, Schueler-Furman O (2022) Harnessing protein folding neural networks for peptide–protein docking. *Nat Commun* 13:176.
107. Roca-Martinez J, Lazar T, Gavalda-Garcia J, Bickel D, Pancsa R, Dixit B, Tzavella K, Ramasamy P, Sanchez-Fornaris M, Grau I, et al. (2022) Challenges in describing the conformation and dynamics of proteins with ambiguous behavior. *Front Mol Biosci* 9:959956.
108. Schmid S, Hugel T (2020) Controlling protein function by fine-tuning conformational flexibility Spies M, Faraldo-Gómez JD, Schlierf M, Craggs T, editors. *eLife* 9:e57180.
109. Klyshko E, Kim JS-H, McGough L, Valeeva V, Lee E, Ranganathan R, Rauscher S (2024) Functional protein dynamics in a crystal. *Nat Commun* 15:3244.
110. Janson G, Valdes-Garcia G, Heo L, Feig M (2023) Direct generation of protein conformational ensembles via machine learning. *Nat Commun* 14:774.
111. Salo-Ahen OMH, Alanko I, Bhadane R, Bonvin AMJJ, Honorato RV, Hossain S, Juffer AH, Kabadev A, Lahtela-Kakkonen M, Larsen AS, et al. (2021) Molecular Dynamics Simulations in Drug Discovery and Pharmaceutical Development. *Processes* 9:71.
112. Heo L, Feig M (2023) One particle per residue is sufficient to describe all-atom protein structures. :2023.05.22.541652. Available from: <https://www.biorxiv.org/content/10.1101/2023.05.22.541652v1>
113. Jamroz M, Kolinski A, Kmiecik S (2014) CABS-flex predictions of protein flexibility compared with NMR ensembles. *Bioinformatics* 30:2150–2154.
114. Vander Meersche Y, Cretin G, Gheeraert A, Gelly J-C, Galochkina T (2024) ATLAS: protein flexibility description from atomistic molecular dynamics simulations. *Nucleic Acids Research* 52:D384–D392.

115. Jamroz M, Orozco M, Kolinski A, Kmiecik S (2013) Consistent View of Protein Fluctuations from All-Atom Molecular Dynamics and Coarse-Grained Dynamics with Knowledge-Based Force-Field. *J. Chem. Theory Comput.* 9:119–125.
116. Høie MH, Kiehl EN, Petersen B, Nielsen M, Winther O, Nielsen H, Hallgren J, Marcatili P (2022) NetSurfP-3.0: accurate and fast prediction of protein structural features by protein language models and deep learning. *Nucleic Acids Res* 50:W510–W515.
117. Anon Improved PEP-FOLD Approach for Peptide and Miniprotein Structure Prediction | *Journal of Chemical Theory and Computation*. Available from: https://pubs.acs.org/doi/full/10.1021/ct500592m?casa_token=3TfJHz2x5pUAAAAA%3AbOVil5Nb5VrVnYYcP6oK9yEihnlInLlg7aTAn6SYu4sktJ-aIrFhwsSzzr98ZEHrMRcdWuNWCa8iYG1M4
118. Timmons PB, Hewage CM (2021) APPTTEST is a novel protocol for the automatic prediction of peptide tertiary structures. *Brief Bioinform* 22:bbab308.
119. Sehnal D, Bittrich S, Deshpande M, Svobodová R, Berka K, Bazgier V, Velankar S, Burley SK, Koča J, Rose AS (2021) Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Res* 49:W431–W437.
120. Pujols J, Peña-Díaz S, Ventura S AGGREGSCAN3D: Toward the Prediction of the Aggregation Propensities of Protein Structures. In: Gore M, Jagtap UB, editors. *Computational Drug Discovery and Design*. New York, NY: Springer; 2018. pp. 427–443. Available from: https://doi.org/10.1007/978-1-4939-7756-7_21
121. Grossmann L, McClements DJ (2023) Current insights into protein solubility: A review of its importance for alternative proteins. *Food Hydrocolloids* 137:108416.
122. Qing R, Hao S, Smorodina E, Jin D, Zalevsky A, Zhang S (2022) Protein Design: From the Aspect of Water Solubility and Stability. *Chem Rev* 122:14085–14179.
123. Kramer RM, Shende VR, Motl N, Pace CN, Scholtz JM (2012) Toward a Molecular Understanding of Protein Solubility: Increased Negative Surface Charge Correlates with Increased Solubility. *Biophys J* 102:1907–1915.
124. Roberts CJ (2014) Therapeutic Protein Aggregation: Mechanisms, Design, and Control. *Trends Biotechnol* 32:372–380.
125. Varadi M, Bertoni D, Magana P, Paramval U, Pidruchna I, Radhakrishnan M, Tsenkov M, Nair S, Mirdita M, Yeo J, et al. (2024) AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res* 52:D368–D375.
126. Badaczewska-Dawid AE, Kuriata A, Pintado-Grima C, Garcia-Pardo J, Burdukiewicz M, Iglesias V, Kmiecik S, Ventura S (2023) A3D Model Organism Database (A3D-MODB): a database for proteome aggregation predictions in model organisms. *Nucleic Acids Research*:gkad942.
127. Tam C, Iwasaki W (2023) AlphaCutter: Efficient removal of non-globular regions from predicted protein structures. *Proteomics* 23:e2300176.

128. Pintado C, Santos J, Iglesias V, Ventura S (2021) SolupHred: a server to predict the pH-dependent aggregation of intrinsically disordered proteins. *Bioinformatics* 37:1602–1603.
129. Oeller M, Kang R, Bell R, Ausserwöger H, Sormanni P, Vendruscolo M (2023) Sequence-based prediction of pH-dependent protein solubility using CamSol. *Briefings in Bioinformatics* 24:bbad004.
130. Pintado C, Santos J, Iglesias V, Ventura S (2021) SolupHred: a server to predict the pH-dependent aggregation of intrinsically disordered proteins. *Bioinformatics* 37:1602–1603.
131. Inglis AJ, Masson GR, Shao S, Perisic O, McLaughlin SH, Hegde RS, Williams RL (2019) Activation of GCN2 by the ribosomal P-stalk. *Proc Natl Acad Sci U S A* 116:4946–4954.
132. Harding HP, Ordonez A, Allen F, Parts L, Inglis AJ, Williams RL, Ron D (2019) The ribosomal P-stalk couples amino acid starvation to GCN2 activation in mammalian cells. *Sonenberg N, Manley JL, Sattlegger E, editors. eLife* 8:e50149.
133. Ji X, Nielsen AL, Heinis C (2024) Cyclic Peptides for Drug Development. *Angewandte Chemie International Edition* 63:e202308251.
134. Rettie SA, Campbell KV, Bera AK, Kang A, Kozlov S, Bueso YF, De La Cruz J, Ahlrichs M, Cheng S, Gerben SR, et al. (2025) Cyclic peptide structure prediction and design using AlphaFold2. *Nat Commun* 16:4730.
135. Kalakoti Y, Wallner B (2025) AFsample2 predicts multiple conformations and ensembles with AlphaFold2. *Commun Biol* 8:373.
136. Pak MA, Markhieva KA, Novikova MS, Petrov DS, Vorobyev IS, Maksimova ES, Kondrashov FA, Ivankov DN (2023) Using AlphaFold to predict the impact of single mutations on protein stability and function. *PLoS One* 18:e0282689.
137. Masters MR, Mahmoud AH, Lill MA (2025) Investigating whether deep learning models for co-folding learn the physics of protein-ligand interactions. *Nat Commun* 16:8854.
138. Krokidis MG, Koumadorakis DE, Lazaros K, Ivantsik O, Exarchos TP, Vrahatis AG, Kotsiantis S, Vlamos P (2025) AlphaFold3: An Overview of Applications and Performance Insights. *Int J Mol Sci* 26:3671.
139. Jing B, Berger B, Jaakkola T (2024) AlphaFold Meets Flow Matching for Generating Protein Ensembles. Available from: <http://arxiv.org/abs/2402.04845>
140. Li S, Li M, Wang Y, He X, Zheng N, Zhang J, Heng P-A (2024) Improving AlphaFlow for Efficient Protein Ensembles Generation. Available from: <http://arxiv.org/abs/2407.12053>
141. Dao TM, Rahman T (2025) Deep Generative Modeling of Protein Conformations: A Comprehensive Review. *BioChem* 5:32.
142. Niazi SK (2025) Critical assessment of AI-based protein structure prediction: Fundamental challenges and future directions in drug discovery. *Computational and Structural Biotechnology Reports* 2:100064.

143. Valsson Í, Warren MT, Deane CM, Magarkar A, Morris GM, Biggin PC (2025) Narrowing the gap between machine learning scoring functions and free energy perturbation using augmented data. *Commun Chem* 8:41.
144. Xie Y, Zhang C, Li S, Du X, Lu Y, Wang M, Hu Y, Chen Z, Liu S, Gao YQ (2025) Integrating diverse experimental information to assist protein complex structure prediction by GRASP. *Nat Methods*:1–13.
145. Watson JL, Juergens D, Bennett NR, Trippe BL, Yim J, Eisenach HE, Ahern W, Borst AJ, Ragotte RJ, Milles LF, et al. (2023) De novo design of protein structure and function with RFdiffusion. *Nature* 620:1089–1100.