Universidade Federal de Pernambuco

Centro de Ciências Exatas e da Natureza

Departamento de Química Fundamental

Programa de Pós-Graduação em Química

Júlio Cesar de Melo Simões

Caracterização termodinâmica da resposta imune humoral contra a proteína do envelope do vírus da Dengue

Recife

2025

Júlio Cesar de Melo Simões

Caracterização termodinâmica da resposta imune humoral contra a proteína do envelope do vírus da Dengue

Documento de Dissertação Acadêmica apresentado ao Programa de Pós-Graduação em Química do Departamento de Química Fundamental da Universidade Federal de Pernambuco (Campus Recife) como como requisito para Exame de Defesa de Mestrado Acadêmico.

Área de Concentração: Química Teórica aplicada a sistemas biológicos

Orientador: Prof. Dr. Roberto Dias Lins Neto

Co-orientador: Prof. Dr. Danilo Fernandes Côelho

Recife

2025

.Catalogação de Publicação na Fonte. UFPE - Biblioteca Central

Simões, Júlio Cesar de Melo.
Caracterização termodinâmica da resposta imune humoral contra a proteína do envelope do vírus da Dengue / Julio Cesar de Melo Simoes. - Recife, 2025.
165f.: il.

Dissertação (Mestrado) - Universidade Federal de Pernambuco, Centro de Ciências Exatas e da Natureza, Programa de Pós-Graduação em Química, 2025.
Orientação: Roberto Dias Lins Neto.
Coorientação: Danilo Fernandes Côelho.
Inclui referências e apêndices.

1. Energia Livre de Gibbs; 2. Evasão Imune Humoral; 3. Vírus da Dengue. I. Lins Neto, Roberto Dias. II. Côelho, Danilo Fernandes. III. Título.

UFPE-Biblioteca Central

“A visão errada da ciência é traída pela sua necessidade de estar correta; pois não é a possessão de conhecimento ou de verdades absolutas que fazem o cientista, mas sua persistência descomedida na busca crítica pela verdade”.

- Karl Popper, *A Lógica da Pesquisa Científica.*

AGRADECIMENTOS

Agradeço ao dQF e à FIOCRUZ pelo suporte técnico e infraestrutura, bem como às agências de fomento pelas concessões de financiamentos para todo o grupo de pesquisa, em especial à FACEPE, pela concessão da bolsa de Mestrado Acadêmico, sem a qual o presente trabalho não seria possível.

Aos meus orientadores, que me aconselharam diligente e continuamente em matérias acadêmicas e extra-acadêmicas e se tornaram grandes amigos para além dos muros do Campus.

A todos os colegas do grupo *Protein Engeneering and Structural Genomics* (do inglês, Engenharia de Proteínas e Genômica Estrutural), pelo ambiente estimulador e acolhedor, pelas discussões e ideias diárias e, na ausência de soluções, por compartilhar dos problemas e fornecer apoio constante e necessário ao longo desta jornada.

A todos os membros da banca deste exame, por doarem seus valiosos tempo e cérebros para a avaliação interdisciplinar deste trabalho.

A todos os professores e professoras do Programa de Pós-graduação em Química (PPGQUI), que serviram como inspiração acadêmica por sua dedicação e ética de pesquisa e ensino inabaláveis, mesmo frente a cenários políticos obscuros para o progresso científico, além de todos os aprendizados acadêmicos formais em sala de aula. Aos professores e professoras da graduação, em especial à Profª Drª Gisélia Alves, que, além de ter sido guia sólido de metodologia científica e filosofia da ciência, me apoiou na mudança de carreira que motivou os presentes estudos. Aos funcionários do dQF, da FIOCRUZ e das bibliotecas do Campus, em especial aos membros das secretarias destes centros, pela grande ajuda com os processos burocráticos ao longo dos últimos meses.

Ao *University College London* (UCL), em nome dos pesquisadores Profª Drª Franca Fraternali e Dr. Carlos Cruz, pela oportunidade, ensinamentos e pesquisas realizadas em solo internacional, e a todos os colegas feitos neste curto e enriquecedor período.

À minha família e amigos, em especial aos que dividiram o mesmo teto, Víctor e Katya, pelo apoio e paciência inestimáveis nesta e em outras empreitadas.

A todos que, direta ou indiretamente, apoiam e participam da luta pelo progresso científico justo, humanístico, racional e ético.

RESUMO

Evasão imune é o fenômeno pelo qual um patógeno passa a escapar das tentativas do sistema imune do hospedeiro de combatê-lo, geralmente por acúmulo de mutações genéticas. A evasão imune humoral pode ser modelada quimicamente como um sistema de interações proteína a proteína (anticorpo-antígeno), no qual alterações no antígeno nativo podem levar ao desligamento do complexo. Variações de duas ordens de grandeza na constante de dissociação do complexo mutante em relação ao nativo, correspondentes a aumentos de energia livre de Gibbs (ΔΔG) de 2,84 kcal/mol, são suficientes para levar à evasão. A Dengue é a mais importante doença viral transmitida por mosquitos, causada por um flavivírus de RNA (DENV) com quatro sorotipos distintos. No paradigma fisiopatológico vigente, não há evasão imune por sorotipo (um paciente pode contrair cada sorotipo de DENV apenas uma vez). Porém, a situação epidemiológica de 2024 (aumento desproporcional de incidência ao curto intervalo) e evidências da possibilidade de evasão imune na Dengue, podem apontar o contrário, motivo deste trabalho. A hipótese principal foi de que o acúmulo de mutações na proteína mais antigênica do sorotipo 2 de DENV, a proteína do envelope (E), pode provocar evasão imune humoral. Tendo em vista o alto volume de dados e a necessidade de cálculos rápidos, comum em problemas biológicos, tal hipótese foi testada por três metodologias computacionais distintas, duas com cálculo por métodos físicos (Rosetta Flex ddG e FoldX), e uma por algoritmo de aprendizagem de máquinas (PBEE). Comparamos o ΔΔG de interação da proteína E com anticorpos de mamíferos àquela entre os mesmos anticorpos e proteínas E com estrutura resolvida no PDB (cepas de 2010-2014) e mais recentes (2023-2024), obtidas em parceria com a Rede Genômica da FIOCRUZ. Resultados dos métodos físicos corroboram a hipótese principal; já os do PBEE corroboram a hipótese nula, sendo os únicos compatíveis com a determinação experimental de ΔΔG de complexos da proteína E, disponível para quatro das dez estruturas testadas em nosso conjunto de dados. Tais resultados não sugerem ocorrência de evasão imune nas cepas testadas até o momento, e infere-se que as estratégias vacinais desenvolvidas e testadas experimentalmente devem manter sua eficácia por ora. O trabalho traz novidade metodológica, dado o desenvolvimento de um novo protocolo de avaliação computacional rápida de evasão imune humoral em vírus a partir de dados de sequenciamento genômico. Espera-se que este contribua para melhor compreensão da fisiopatologia e dinâmica viral, permitindo o desenvolvimento de ferramentas de vigilância epidemiológicas mais robustas, acuradas e preditivas, incluindo a previsão de eficácia vacinal frente a futuros surtos.

Palavras-chaves: Vírus da Dengue, evasão imune humoral, energia livre de Gibbs.

ASBTRACT

Immune evasion is the phenomenon by which a pathogen escapes the host's immune response, typically through the accumulation of genetic mutations. Humoral immune evasion can be chemically modeled as a protein–protein interaction system (antibody–antigen), where mutations in the native antigen may disrupt the complex. A two-order magnitude increase in the dissociation constant of the mutant complex compared to the native—corresponding to a rise in Gibbs free energy (ΔΔG) of 2.84 kcal/mol—is sufficient to induce evasion. Dengue is the most significant mosquito-borne viral disease, caused by an RNA flavivirus (DENV) with four distinct serotypes. The prevailing pathophysiological paradigm asserts that serotype-specific immune evasion does not occur (i.e., a patient can only contract each DENV serotype once). However, the disproportionate increase in incidence over a short interval in 2024, along with emerging evidence of immune evasion in Dengue, may challenge this view—motivating this investigation. Our central hypothesis was that the accumulation of mutations in the envelope (E) protein, the most antigenic protein of DENV serotype 2, may lead to humoral immune evasion. Given the need for rapid and high-throughput calculations, common in biological problems, this hypothesis was tested using three distinct computational approaches: two physics-based methods (Rosetta Flex ddG and FoldX) and one machine learning algorithm (PBEE). We compared the ΔΔG of interaction between the E protein and mammalian antibodies with that of the same antibodies and E proteins with resolved structures from the PDB (2010–2014 strains) versus newer ones (2023–2024), obtained in collaboration with FIOCRUZ's Genomic Network. Results from the physics-based methods support the central hypothesis; results from PBEE support the null hypothesis, and only the latter aligned with experimentally determined ΔΔG values of E–antibody complexes. These findings do not suggest immune evasion in the tested strains—a positive outcome, indicating that current experimentally tested vaccine strategies are likely to remain effective for now. This work also presents methodological innovation, with the development of a new protocol for rapid computational assessment of viral humoral immune evasion based on genomic sequencing data. It is expected to contribute to a better understanding of viral pathophysiology and dynamics, enabling the development of more robust, accurate, and predictive epidemiological surveillance tools, including forecasts of vaccine efficacy against future outbreaks.

Keywords: Dengue virus, humoral immune evasion, Gibbs free energy.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

Ab	Anticorpo (do inglês, *antibody*)
ADE	Efeito *antibody-dependent enhancement* (do inglês, sem equivalente em língua portuguesa, potencialização dependente de anticorpos, em tradução livre)
AF, AF1, AF2, AF3	*AlphaFold*; edições 1, 2 e 3
CHIKV	Vírus da Chikungunya (do inglês, *Chikungunya Virus*)
DENV	Vírus da Dengue (do inglês, *Dengue Virus*)
DENV1, 2, 3, 4	Sorotipos 1, 2, 3 e 4 de DENV
DNA	Ácido desoxirribonucleico
E	(Proteína do) Envelope de DENV
EM, cryo-EM	Microscopia Eletrônica, Crio Microscopia Eletrônica
Emut	(Proteína do) Envelope de DENV mutada computacionalmente
Ewt	(Proteína do) Envelope de DENV nativa, obtida do PDB
E:Ab	Complexo entre anticorpo e proteína do envelope de DENV
ED3, EDIII	Domínio 3 da Proteína do Envelope
EDO	Equação Diferencial Ordinária
ELISA	*Enzyme-Linked Immunossorbent Assay* (do inglês, ensaio imunossorvente de ligação enzimática)
Fab	Fragmento de ligação ao antígeno (do anticorpo)
Fc	Fragmento cristalizável (do anticorpo)
Fv	Fragmento variável (do anticorpo)
FIOCRUZ	Fundação Oswaldo Cruz
H0	Hipótese nula
MC	(Método de) Monte Carlo

MD	Dinâmica Molecular
ML	*Machine Learning* (Aprendizagem de Máquina)
MM	Mecânica Molecular
MUT	Mutante, Mutação
OMS	Organização Mundial de Saúde
PCR, RT-qPCR	Reação em cadeia da polimerase (PCR), PCR em tempo real (RT-qPCR)
PDB	*Protein Data Bank* (Banco de Dados de Proteínas)
pLDDT	*Predicted Local Distance Difference Test* (do inglês, teste predito das diferenças locais)
pMPNN	*Protein Message Passing Neural Network* (do inglês, pacote computacional sem equivalente em língua portuguesa)
PPI	Interface Proteína-Proteína
PRNT	*Plaque Reduction Neutralization Test* (do inglês, ensaio de neutralização por redução em placas)
QM	Mecânica Quântica
RFd	*Rosetta Fold Diffusion* (do inglês, pacote computacional sem equivalente em língua portuguesa)
RMSD	*Root Mean Square Deviation* (do inglês, desvio quadrático médio)
RNA	Ácido ribonucleico
SASA	Área de Superfície Acessível ao Solvente
SE	Semana Epidemiológica
SSM	*Site Saturation Mutagenesis* (do inglês, mutagênese por saturação local)
vdW	(Interações de) van der Waals
WT	*Wild-type* (do inglês, selvagem ou nativo)
ZIKV	Vírus da Zika

SUMÁRIO

1. INTRODUÇÃO

1.1 DENGUE

A infecção pelo vírus da Dengue (DENV) é atualmente considerada globalmente como a mais importante doença viral causada por mosquitos. A doença é causada por um vírus de RNA da família *Flaviviridae* com quatro sorotipos distintos (DENV1, DENV2, DENV3, DENV4). Um mesmo paciente pode contrair a doença até quatro vezes ao longo da sua vida, uma vez para cada sorotipo. A transmissão é vetorial e ocorre principalmente por duas espécies de mosquitos: *Aedes aegypti*, que se adaptou ao ambiente urbano altamente populado em países tropicais e subtropicais, e *Aedes albopictus*, uma ameaça a países de clima temperado pela sua adaptação a ambientes mais frios. (B.A. Seixas; Giovanni Luz; Pinto Junior, 2024) Uma história de sintomas compatíveis com dengue pode ser traçada desde a Dinastia Chin na China (265-420 DC). O vírus e seus vetores se tornaram amplamente distribuídos por regiões tropicais e subtropicais do mundo, particularmente na última metade do século XX. A expansão geográfica significativa ocorre em conjunto com um rápido aumento na taxa de incidência e hiperendemicidade, levando a formas mais severas de dengue. (Khan et al., 2023) A transmissão do vírus da dengue (DENV) está agora presente em todas as regiões da OMS do mundo todo, e é endêmica para mais de 125 países. O impacto real da dengue globalmente é difícil de ser medido devido à vigilância epidemiológica inadequada, aos diagnósticos incorretos e à subnotificação. Dados epidemiológicos atualmente disponíveis subestimam largamente o impacto social e econômico da doença. Estimativas da incidência global de dengue variam de 50 milhões a 200 milhões de casos novos por ano. Entretanto, estimativas recentes utilizando métodos cartográficos sugerem que este número está mais próximo dos 400 milhões de casos novos por ano, esperando-se um aumento ainda mais significativo no número de casos com as mudanças climáticas, globalização, viagens e importações, acordos socioeconômicos globais e evolução viral. (Murray; Quam; Wilder-Smith, 2013)

Apesar da menor incidência nos últimos anos, houve novo aumento em 2024, em que foram notificados cerca de 6,6 milhões de casos prováveis no Brasil, com 6041 óbitos dentre estes (taxa de letalidade geral de 0,09%), a maioria dos quais entre as semanas epidemiológicas (SE) 8 e 19 de 2024. A letalidade dos casos graves confirmados em 2024 foi de 5,77%. A prevalência do vírus da dengue (DENV) em 2024 é significativamente maior quando comparada com as mesmas semanas epidemiológicas do ano anterior. Após período relativamente ameno para as arboviroses entre 2020 e 2023, este novo aumento de número de

casos era esperado, dentre outros fatores, pelo próprio comportamento cíclico da doença e pelo maior controle da pandemia de síndrome respiratória aguda por coronavírus (SARS-CoV-2), permitindo o retorno às atividades regulares de vigilância epidemiológica dos demais agravos. Todos os quatro sorotipos de DENV foram isolados no Brasil, embora DENV1 e DENV2 ainda sejam os mais ubíquos em nosso território. (BRASIL, 2024; Brasil, 2025)

Estruturalmente, o DENV pode ser caracterizado como um espécime com forma esférica e diâmetro médio de 50 nm. O genoma do DENV é composto por um RNA de fita simples (ssRNA) e uma sequência aberta (em inglês, *ORF*, *Open Reading Frame*) que codifica as proteínas virais. Existem sete proteínas estruturais (cápsidio – C; pré-membrana/membrana – prM/M; envelope – E) e sete proteínas não estruturais (NS1, NS2A, NS2B, NS3, NS4B e NS5). (Kuno; Chang, 2007) As proteínas E e M formam um escudo icosaédrico que protege o nucleocapsídeo (formado pela proteína C e pelo ssRNA). As proteínas não estruturais (NS) diferem das anteriores porque não constituem a anatomia da partícula viral. No entanto, as proteínas NS desempenham um papel crucial na síntese do RNA que compõe os genomas dos novos vírions. As proteínas E e NS1 são frequentemente consideradas os principais candidatos para imunização devido à sua alta antigenicidade. (Muller; Young, 2013; Young et al., 2000)

A doença da dengue é causada pela infecção pelo vírus da dengue, e as manifestações clínicas incluem **febre da dengue** (DF), **febre hemorrágica da dengue** (DHF) ou **síndrome do choque da dengue** (DSS). Historicamente, a DF é conhecida por causar sintomas semelhantes aos do resfriado comum em humanos, como cefaleia, artralgia (dor nas articulações) e mialgia (dor muscular). Esses sintomas também são observados em outras arboviroses tropicais; portanto, a apresentação clínica isolada pode levar a muitos erros de diagnóstico, especialmente em relação ao vírus Zika (ZIKV) e ao vírus Chikungunya (CHIKV), as outras arboviroses mais comuns em circulação no Brasil. Complicações neurológicas graves (nomeadamente a Síndrome Congênita do Zika e a Síndrome de Guillain-Barré) estão principalmente associadas à infecção por ZIKV na América e na Oceania. (De Araújo et al., 2016, 2018; De Barros Miranda-Filho et al., 2016; Del Campo et al., 2017; Souza et al., 2018) Já complicações hemorrágicas e choque são mais comumente associadas ao vírus da Dengue nas supracitadas formas clínicas DHF ou DSS. Diferentes fatores e mecanismos são considerados envolvidos na apresentação da DHF e da DSS, incluindo a ocorrência de fenômeno de *antibody-dependent enhancement* (ADE; do inglês, efeito de potencialização dependente de anticorpos), desregulação imunológica, virulência viral, suscetibilidade genética do hospedeiro e anticorpos preexistentes contra a dengue. (Khan et al., 2023)

Geralmente, a infecção aguda por DENV é benigna e primariamente febril, conforme caracterizado na forma clínica DF. Entretanto, infecções sucessivas com sorotipos alternativos podem agravar a condição, levando às formas mais graves e potencialmente fatais, isto é, DHF ou DSS. Os anticorpos gerados por uma vacina não têm 100% de eficácia contra todos os sorotipos, e aqueles produzidos em infecções primárias frequentemente apresentam reatividade cruzada, porém são fracamente neutralizantes contra outro sorotipo. Durante infecções subsequentes, os anticorpos não-neutralizantes podem aumentar as chances de ocorrência do fenômeno de ADE. Este fenômeno é caracterizado pela possibilidade de anticorpos não neutralizantes (gerados em infecção prévia por outro sorotipo) facilitarem a infectividade e patogenicidade, favorecendo a ocorrência de formas mais graves. (Murray; Quam; Wilder-Smith, 2013) Apesar disso, muitos anticorpos neutralizantes têm sido identificados contra DENV, os quais são considerados úteis na redução da gravidade da doença. Para finalidades terapêuticas, um anticorpo deve estar livre de ADE, pelo aumento de risco de doença grave em reinfecções, conforme explicado. No que diz respeito à resposta imune contra DENV, grande ênfase é dada à proteína do envelope. Potenciais epítopos direcionados para a geração de anticorpos – sejam específicos por sorotipo ou com reação cruzada entre sorotipos - têm sido descritos de forma crítica na proteína E. (Sarker; Dhama; Gupta, 2023)

No geral, o fato de a apresentação clínica isolada não ser suficiente para um diagnóstico diferencial confiável é um ponto limitante na prática clínica. O diagnóstico diferencial depende de testes laboratoriais e apresenta ainda mais dificuldades quando o paciente está assintomático. Atualmente, os laboratórios de referência diagnóstica utilizam testes moleculares, baseados no isolamento e na detecção de RNA viral, por reação em cadeia da polimerase em tempo real (RT-qPCR) ou técnicas alternativas. A execução de testes moleculares exige reagentes caros e pessoal altamente qualificado. Até agora, os testes padrão-ouro para o diagnóstico diferencial de flavivírus são os testes de neutralização viral por redução de placas (PRNT), que além de exigirem elevado nível de biossegurança por lidarem com vírus vivos, são muito laboriosos, onerosos e pouco disponíveis. (Kuno; Chang, 2007) Apenas 36,2% dos casos confirmados de DENV no Brasil em 2024 o fizeram por meio de testes laboratoriais (Brasil, 2025), incluindo os testes moleculares de pesquisa viral supracitados e os testes de pesquisa de anticorpo.

Devido à ausência de medicamentos antivirais e vacinas eficazes, várias estratégias terapêuticas e de controle têm sido propostas. Embora não existam tratamentos específicos além do manejo de suporte geral ao paciente, as vacinas são uma área de grande pesquisa, com duas vacinas, Dengvaxia® (CYD-TDV) e Denvax® (TAK003), recentemente licenciadas para uso

clínico. A CYD-TDV é altamente eficaz em crianças com 9 anos ou mais que já tiveram infecção prévia por DENV (taxa de eficácia de 71% nesta população), (Diaz-Quijano et al., 2024) devido ao alto risco de doença grave em crianças soronegativas de 2 a 5 anos. Enquanto isso, a TAK003 mostrou eficácia de 97,7% e 73,7% contra DENV2 e DENV1, respectivamente, em ensaios clínicos de fase 3 na América Latina e na Ásia, em crianças saudáveis de 4 a 16 anos com dengue virologicamente confirmada. Outras vacinas, incluindo TV003 e TV005, continuam sendo desenvolvidas em todo o mundo, com a esperança de entrar em ensaios clínicos num futuro próximo. (Kallás et al., 2024; Kariyawasam et al., 2023) Outro grupo de estratégias de profilaxia consiste em controle vetorial. É possível tentar ferramentas de controle vetorial com uso de inseticidas e extratos de plantas, uso de bactérias simbióticas, modificação da população natural de mosquitos (uso de mosquitos geneticamente modificados), entre outras. (Araújo et al., 2015) Cerca de metade da população global está em risco de dengue, causando aproximadamente 100 milhões de infecções sintomáticas por ano. Com o aumento das temperaturas globais, o habitat do vetor provavelmente se expandirá. (Messina et al., 2019) Como o vírus da dengue é geneticamente diverso tanto dentro quanto entre seus quatro sorotipos, não é certo que as vacinas ou mosquitos geneticamente modificados terão eficácia ubíqua. Assim, para monitorar e ajudar a refinar essas abordagens, e fornecer dados epidemiológicos sobre linhagens emergentes, a vigilância genômica rotineira do vírus da dengue é urgentemente necessária. (Hill et al., 2023a)

1.2 DEFINIÇÕES BIOLÓGICAS PRELIMINARES

Antes de prosseguir, é importante retomar as definições de cepas, variantes, genótipos e sorotipos em virologia. Embora sejam termos de uso bastante corriqueiro na literatura da área, poucos autores propuseram definições robustas destes termos, o que pode levar a certo grau de confusão nestas definições. O próprio Comitê Internacional de Taxonomia Viral não tem diretrizes para a nomenclatura e definição de variantes virais (exceto para alguns vírus e casos específicos), justificando, em parte, a falta de homogeneidade nas definições, demarcações e nomenclaturas de variantes virais. A taxonomia viral além do nível de espécie (incluindo cepas, variantes, genótipos etc.) é bastante problemática pela falta de definição na literatura destes termos. Kuhn e colaboradores [2013] apontaram este problema taxonômico e trouxeram definições mais robustas para cepas e variantes, que adotaremos ao longo deste trabalho. Para estes autores, o termo **variante** viral se refere a um espécime isolado (ou conjunto de espécimes isolados) cuja sequência genômica de consenso é distinta da sequência de referência para aquela

espécie. Frequentemente os termos variantes e mutantes são utilizados como sinônimos, definição que também adotaremos ao longo do presente trabalho. Já o termo **cepa** se refere à variante viral reconhecível por suas características fenotípicas distintas. (Kuhn et al., 2013) **Sorotipo** viral é um termo que classifica vírus de mesma espécie de acordo com o padrão das respostas imunes humorais do hospedeiro elicitadas pelo vírus. (Simon-Loriere; Schwartz, 2022) Para cada espécie, há definições distintas de sorotipo, geralmente com base na proteína mais antigênica viral. No caso de DENV, a sorotipagem era classicamente feita pela caracterização da proteína do envelope e/ou das proteínas de membrana e NS1. (Shu et al., 2004) Atualmente, a sorotipagem é tipicamente realizada com estratégias de análise de RNA viral (mais comumente por transcriptase reversa de RNA em tempo real baseada em genoma). (Prommool et al., 2021) Dentro de cada sorotipo é possível distinguir genótipos específicos. Um **genótipo** é genericamente definido como a constituição genética de um organismo, e, no caso de vírus, o termo se aplica a formas nas quais a sequência genômica se estabilizou após um intervalo temporal prolongado e que tem capacidade replicativa. (Kramvis; Kew; François, 2005) Em geral, com a descrição de novas variantes, estas são classificadas nos genótipos já conhecidos, ou uma nova classificação genotípica é proposta, caso a variante seja suficientemente distinta das anteriores.

A evolução e a diversidade genética dos vírus podem impactar as dinâmicas de surtos e os esforços de controle, como demonstrado pela resposta à pandemia de SARS-CoV-2, especialmente com o surgimento de variantes. Monitorar os processos evolutivos pode fornecer respostas que estão ocultas à epidemiologia tradicional, tanto em pequena escala, investigando os detalhes das cadeias de transmissão, quanto desvendando os caminhos de disseminação em escala global. Dados genômicos também têm o potencial de melhorar modelos de previsão de doenças e têm sido críticos para o desenvolvimento de vacinas, terapêuticas e ensaios diagnósticos moleculares. No geral, o sequenciamento viral pode ser usado para projetar, aplicar e avaliar melhor as estratégias de mitigação de transmissão e doença. (Tosta et al., 2023a) Os vírus estão continuamente mudando. Variações genéticas ocorrem ao longo do tempo e podem levar ao surgimento de novas variantes que possam expressar características ou fenótipos alterados, alguns dos quais podem ser de interesse para a saúde pública. No entanto, a detecção e a compreensão geral dessas variantes podem ser dificultadas pela disponibilidade limitada de dados genômicos e epidemiológicos. (Hill et al., 2023a) Alguns vírus podem apresentar processo de evolução viral mais lento e conservado, enquanto outros podem ter taxa de mutações extremamente elevada e rápida, levando ao escape imune célere de novas cepas,

também conforme exemplificado pela pandemia de SARS-CoV-2. (Tosta et al., 2023a) Para o vírus da Dengue, classicamente foi proposto o modelo "um sorotipo-uma infecção", segundo o qual um espécime humano só poderá adquirir uma infecção por sorotipo, totalizando quatro possíveis infecções por Dengue, no máximo, ao longo da vida. Esta proposição inicialmente foi introduzida como axiomática no clássico trabalho de Halsted. (Halstead, 1893) Apesar de sua introdução axiomática, esta hipótese foi suportada por diversas evidências experimentais posteriores em diversas vertentes metodológicas (epidemiológica, molecular, entre outras). Atualmente, já há dados suficientes para apoiar a hipótese de uma imunidade por sorotipo sustentada por muitas décadas após a primeira exposição. (Imrie et al., 2007)

Este modelo tornou-se, assim, um paradigma biológico, mas evidências recentes sugerem que mutações pontuais nos vírus da Dengue podem provocar evasão imune. Apesar disso, o trabalho em questão realizou uma mutação baseada em intuição química, não correspondendo a uma mutação que ocorreu, de fato, em cepas circulantes. (Sukupolvi-Petty et al., 2010) Entretanto, a evidência é importante para sugerir que mutações podem chegar a novas cepas de Dengue que escapem à imunidade por sorotipo. Outros trabalhos também chamam atenção para a possibilidade de que este paradigma não seja sustentado de maneira absoluta nos próximos anos, com a evolução e celeridade da determinação genômica de novas variantes e com a melhor compreensão dos mecanismos de evasão imunológica viral. (Lee et al., 2022; Mushtaq et al., 2023)

O avanço tecnológico no sequenciamento genômico também permite uma compreensão mais profunda da relação entre estrutura, função e fisiopatologia, além de permitir categorizar os sorotipos em genótipos distintos. Para o sorotipo DENV2, foco do presente trabalho e sorotipo mais prevalente no Brasil, foram descritos 5 genótipos específicos até o momento, a saber: I (Americano), II (Cosmopolita), III (Asiático-americano), IV (Asiático II) e V (Asiático I). Este sorotipo teve relevância ainda superior recentemente, pela introdução do genótipo II em 2021 no território brasileiro (provavelmente oriundo do Peru) e sua rápida disseminação desde então, provocando aumento considerável na taxa de incidência de DENV2 desde então. (Gräf et al., 2023) O rápido aumento desta taxa chama a atenção por ter ocorrido em tempo muito inferior ao necessário para que haja mudança geracional: em 2024 houve aumento de 400% do número de casos no Brasil. (Brasil, 2025) A partir desta evidência ecológica, podemos propor a hipótese de que o paradigma clássico de "uma infecção – um sorotipo" talvez deva, em breve, ser atualizado para "uma infecção – um genótipo". Isto porque houve aumento significativo de

número de casos de DENV2 com a introdução de um novo genótipo, ainda que não tenha ocorrido mudança geracional que acompanhe e justifique.

1.3 EVASÃO IMUNE TERMODINAMICAMENTE DEFINIDA

A frequência de ocorrência de mutações no material genético viral é um critério importante para considerar na compreensão da evolução viral. Há vários mecanismos distintos de evasão imune, mas grande ênfase é dada na ocorrência de mutações, dada a sua importância e presença como pré-requisito na maioria destes mecanismos. (Chakraborty et al., 2022) Tentar considerar toda a complexidade dos possíveis mecanismos de evasão imune biologicamente para modelagem matemática e química seria tarefa árdua e infrutífera. A modelagem matemática de fenômenos biológicos é sempre desafiadora pela necessidade de simplificação dos paradigmas deste domínio cognitivo naquele, o que pode gerar resistência da comunidade científica pela possibilidade de redução do seu paradigma. O objetivo destes recursos, porém, não é o reducionismo, mas a simplificação de um fenômeno biológico complexo para permitir seu estudo sistemático e quantitativamente definido. Por esta razão, no presente trabalho, consideramos particularmente as interações antígeno-anticorpo, centrais nos processos de imunidade adaptativa humoral, como paradigma para a construção de um modelo de evasão imune termodinamicamente definido. Como o complexo antígeno-anticorpo é, quimicamente, um complexo não covalente de interações proteína-proteína (PPI), esta abordagem nos pareceu bastante adequada para o presente estudo. Além da vantagem de este modelo ser passível de descrição físico-química, há também a vantagem da sua importância nas estratégias vacinais vigentes para a doença, que dependem largamente da eficácia do sistema imune humoral do hospedeiro. Assim, considerando a complexação como um equilíbrio químico entre os estados ligado (complexo formado) e desligado (complexo não formado), é possível escrever:

$$Ag + Ab \rightleftharpoons Ag{:}Ab$$

Na representação acima o símbolo dos dois pontos indica a formação do complexo, enquanto Ag e Ab representam o antígeno e o anticorpo, respectivamente. Sendo mais específico para o vírus de DENV2, escopo do presente trabalho, tomamos a proteína do envelope (E) para estudo. Isto porque a proteína E apresenta potencial antigênico elevado, constituindo a proteína viral contra a qual ocorre maior número de anticorpos neutralizantes. (Schieffelin et al., 2010) Assim, a representação acima pode ser especificada para o presente problema como:

$$E + Ab \rightleftharpoons E{:}Ab$$

A reação nos sentidos direto e inverso tem constantes cinéticas k_{on} e k_{off}, respectivamente. A constante de equilíbrio deste sistema, que ocorre em meio aquoso biológico (solução não ideal), tomando a atividade dos componentes como proporcional à suas respectivas concentrações de equilíbrio (entre colchetes), portanto, tem a forma $K_{eq} = (k_{on}/k_{off}) = [E{:}Ab]\,[E]^{-1}[Ab]^{-1}$.

A constante de associação dos anticorpos, bem como seu recíproco (constante de dissociação de anticorpos, de uso mais rotineiro nos laboratórios de biologia), constituem os parâmetros de maior importância no par antígeno-anticorpo, além de serem extremamente importantes na caracterização geral da afinidade dos anticorpos e da performance de ensaios imunológicos. (Fischer; Kaufmann; Weller, 2024) A constante de dissociação, K_d, em particular, foi a forma mais popularizada deste parâmetro e seus valores para os complexos antígeno-anticorpo geralmente se encontram nas ordens de grandeza micromolar a nanomolar. (Karlsson; Michaelsson; Mattsson, 1991; Landry; Fei; Zhu, 2012) Valores mais baixos estão relacionados com maior afinidade e, em geral, são necessários para anticorpos neutralizantes. Valores de K_d mais altos estão relacionados com menor afinidade e são incomuns em anticorpos neutralizantes, embora apareçam em anticorpos não-neutralizantes com frequência variável. Isto provavelmente ocorre pela necessidade de a interação com tais anticorpos ser estável e altamente específica, além de competitiva. Não encontramos na literatura testes específicos de tais afirmações nos anticorpos neutralizantes de Dengue, mas estas afirmações têm validade definida para outros sistemas virais, a exemplo dos vírus SARS-CoV-2 e Influenza, responsáveis por síndromes respiratórias de elevada importância epidemiológica. (Horns; Dekker; Quake, 2020; Schmidt et al., 2021) É razoável supor, portanto, que o racional biológico subjacente supracitado possa ser estendido para outros sistemas virais. Além disso, embora não existam testes específicos para os anticorpos de Dengue, quatro das dez estruturas encontradas de complexos *Ewt:Ab* no presente trabalho têm seu K_d experimentalmente determinado. Todas estas estruturas representaram anticorpos neutralizantes e todas tiveram K_d da ordem nanomolar de grandeza. (Cockburn et al., 2012; Renner et al., 2018) Assim, é razoável esperar que mudanças em duas ordens de grandeza, representando redução de ordem de grandeza de 10E-9 para 10E-7 no K_d, seriam suficientes para evitar a formação do complexo, permitindo inferir evasão imune para este sistema. Considerando que tais mudanças decorrem de mutações na proteína E, tem-se uma reação de complexação com anticorpo para a proteína E na sua forma nativa (chamada de *wt*, do inglês, *wild-type*) e outra para a proteína E na sua forma mutante (chamada de *mut*). Os equilíbrios podem, assim, ser descritos como:

$$E_{wt} + Ab \rightleftharpoons E_{wt}{:}\,Ab$$

$$E_{mut} + Ab \rightleftharpoons E_{mut}{:}\,Ab$$

Para que a evasão imune seja termodinamicamente definida, portanto, é necessário que o K_d do segundo equilíbrio (que será representado como $K_{d,2}$) seja duas ordens de grandeza maior que o K_d do primeiro (que será representado como $K_{d,1}$). Computacionalmente, em geral, o K_d não é determinado diretamente, mas pode ser calculado pelo valor da variação de energia livre de Gibbs (ΔΔG) do sistema. Como o K_d é o recíproco do K_{eq}, dos conhecimentos de termodinâmica básica e das informações supracitadas, é possível escrever:

$$\Delta G^0 = -R\,T \ln K_{eq}$$

$$\Delta G^{0,lig} = -R\,T \ln K_{eq} = -R\,T \ln \frac{1}{K_d} = R\,T \ln K_d$$

$$\frac{K_{d,2}}{K_{d,1}} = 100$$

$$\Delta\Delta G_{1,2}^{0,lig} = \Delta G_1^{0,lig} - \Delta G_2^{0,lig} = R\,T \ln K_{d,1} - R\,T \ln K_{d,2}$$

$$\Delta\Delta G_{1,2}^{0,lig} = R\,T \ln \frac{K_{d,1}}{K_{d,2}} = R\,T \ln 0{,}01$$

$$\Delta\Delta G_{1,2}^{0,lig} = 1{,}987207 \frac{cal}{mol.K} x\ 310{,}15\ K\ x \ln 0{,}01 \cong 2{,}8383\ kcal/mol$$

Portanto, diferenças de duas ordens de grandeza no K_d podem ser estimadas a partir de diferenças de aproximadamente **2,84 kcal/mol** na energia livre de Gibbs, conforme demonstrado acima. Pelas razões demonstradas, passaremos a adotar este valor de referência ao longo de todo o trabalho para definir termodinamicamente evasão imune humoral. O racional químico acima definido, bem como o valor de corte de ΔΔG de 2,84 kcal/mol, estão de acordo com trabalho prévio do nosso grupo de pesquisa, que adotou metodologia análoga para investigação da ocorrência de evasão imune humoral nas cepas de SARS-CoV-2. (Ferraz et al., 2021)

O presente trabalho foi motivado pela importância epidemiológica da doença e pela necessidade de investigação da manutenção, frente às mutações acumuladas para as cepas de DENV2 circulantes em 2023-2024, da validade do paradigma atual de “uma infecção-um sorotipo”. Nossa hipótese principal foi de que o acúmulo de mutações na proteína do envelope

do vírus da Dengue sorotipo 2 pode provocar evasão imune humoral. Neste contexto, propomos a avaliação da resposta de anticorpos contra a proteína E de DENV2 em termos de suas propriedades termodinâmicas de interação epítopo-anticorpo. O trabalho teve início com o exame das cepas já circulantes nos anos de 2010-2014, cujas estruturas foram determinadas (em complexo com seus respectivos anticorpos) e publicadas no repositório RCSB PDB. Em seguida, em parceria com a Rede Genômica da FIOCRUZ, foram obtidas as sequências de cepas de circulação global mais recente (nos anos 2023-2024). A partir destes dados sequenciais, realizamos análise *in sílico* para determinar a afinidade das proteínas E frente aos seus respectivos anticorpos. Por fim, com os dados gerados no presente trabalho, foi avaliado se a energia livre de ligação de anticorpos de mamíferos contra a proteína do envelope de DENV2 é, de fato, a mesma ao longo do tempo, ou se ocorrem mudanças suficientes para inferir evasão imune (definidas a partir de incrementos de 2,84 kcal/mol no valor de ΔΔG, conforme previamente definido). Esses dados podem ser importantes para sustentar ou refutar a possibilidade de que humanos possam ser infectados mais de uma vez por cada sorotipo do DENV, considerando que outros vírus frequentemente evoluem de modo a levar ao escape da resposta imune e de novas tecnologias de vacinas. Tal conhecimento é de fundamental importância no estudo e implementação de estratégias de prevenção de doença em nível epidemiológico, com especial destaque para a avaliação da resposta vacinal, que depende diretamente da resposta imune humoral contra o vírus.

1.4 ARGUMENTAÇÃO E OBJETIVOS

1.4.1 Argumentação do Estudo

- **Hipótese do Estudo:** O acúmulo de mutações na proteína do envelope do vírus da Dengue sorotipo 2 provoca evasão imune humoral.
- **Hipótese Nula (H0):** O acúmulo de mutações na proteína do envelope do vírus da Dengue sorotipo 2 não provoca evasão imune humoral.
- **Argumento de refutação da hipótese nula:** Os argumentos de refutação da hipótese nula serão construídos a partir de metodologias computacionais distintas de determinação de energia de ligação (ΔΔG) da proteína E frente a anticorpos, comparando os parâmetros de interação obtidos de cepas antigas *versus* recentes do vírus da Dengue sorotipo 2.

1.4.2 Objetivo principal:

Comparar as propriedades termodinâmicas de interação entre a proteína do envelope de DENV e anticorpos de mamíferos contra sequências de aminoácidos de novas cepas de DENV.

1.4.2 Objetivos específicos:

- Procurar e catalogar complexos de proteína E com anticorpos no banco de dados do PDB
- Calcular as propriedades termodinâmicas de interação dos complexos obtidos no PDB.
- Realizar alinhamentos de sequências de aminoácidos das novas cepas de DENV (circulantes em 2023-2024) obtidas do grupo genômico da FIOCRUZ para, a partir destes alinhamentos, modelar as estruturas das novas variantes da proteína do envelope.
- Caracterizar as propriedades termodinâmicas de interação destas frente a anticorpos, e comparar com as propriedades calculadas para as cepas mais antigas, obtidas diretamente do PDB.
- Investigar e comparar as interfaces proteína-proteína nos complexos modelados em relação aos obtidos diretamente do PDB, a fim de justificar quimicamente os resultados quantitativos termodinâmicos obtidos.

2. FUNDAMENTAÇÃO TEÓRICA

2.1 ESTRUTURA TRIDIMENSIONAL DE PROTEÍNAS

2.1.1 Níveis De Organização Estrutural Proteica

Um dos paradigmas clássicos da biologia molecular moderna coloca que a função biológica de proteínas depende de sua estrutura. Inicialmente, este paradigma andava intrinsecamente relacionado ao dogma central da biologia molecular e acreditava-se numa visão enraizada na ideia "um gene - uma proteína - uma função". Atualmente, nossa compreensão evoluiu e é mais adequado utilizar um modelo de correlação definido em termos mais genéricos por um "*continuum* estrutura proteica - função biológica". Neste modelo, uma proteína existe como um *ensemble* conformacional dinâmico contendo múltiplas proteoformas (conformacional/básica, induzida/modificada, funcionante), caracterizado por um amplo espectro de características estruturais e possuindo várias funcionalidades em potencial. Este modelo está de acordo com os achados da proteômica funcional, em que o sequenciamento de proteomas funcionais de organismos é frequentemente maior do que o sequenciamento genômicos desses mesmos organismos, o que vai de encontro à clássica ideia de unidade. (Uversky, 2019)

Tradicionalmente, o estudo da bioquímica básica classifica a estrutura tridimensional de proteínas em quatro níveis de organização. A estrutura primária corresponde à sequência linear de resíduos de aminoácidos ao longo da cadeia polipeptídica em direção N para C-terminal. As proteínas são definidas quimicamente como polímeros de aminoácidos, formando cadeias ditas polipeptídicas. As definições numéricas para diferenciar proteínas de peptídeos são variáveis na literatura, mas de modo geral, quando o número de resíduos de aminoácidos é muito pequeno a estrutura correspondente é dita peptídeo e quando o número de resíduos é maior (e, consequentemente, a complexidade estrutural também) as estruturas são ditas proteínas. Proteínas podem ser compostas por uma ou mais cadeias peptídicas. Os aminoácidos se unem através de ligações peptídicas, um caso particular de reação de síntese orgânica que decorre da interação entre a hidroxila carboxílica de um aminoácido e o hidrogênio do grupo amina do aminoácido consecutivo. A reação libera água como produto de condensação e forma uma ligação amídica entre aminoácidos consecutivos. (Dekimpe; Masschelein, 2021) Após a formação das ligações, os aminoácidos passam a ser chamados de resíduos, já que houve modificação dos grupos funcionais na formação da ligação amídica. Numa sequência

polipeptídica, os resíduos de aminoácidos mantêm apenas o grupo carboxílico do último resíduo da sequência e o grupo amina do primeiro resíduo da sequência (já que os demais fazem parte da ligação peptídica), configurando as extremidades ou regiões C-terminal (por convenção, atribuída ao final da sequência polipeptídica) e N-terminal (também por convenção, no início da sequência polipeptídica). Portanto, o arranjo do *backbone* (do inglês, esqueleto ou cadeia principal) de uma cadeia polipeptídica é conservado entre as proteínas. Em contrapartida, as cadeias laterais dos resíduos (projetadas para fora do plano do *backbone*) se diferenciam pelas suas propriedades químicas, donde o sistema de classificação funcional químico. Além disso, utilizamos códigos de 3 letras e de 1 letra (o último sendo preferível para fins computacionais) para representar cada aminoácido. A classificação funcional das cadeias laterais, bem como sua estrutura e os códigos de 3 e 1 letra estão representados na Figura 1 e a consulta desta figura é recomendada ao longo da leitura do presente trabalho. (Verli, 2014)

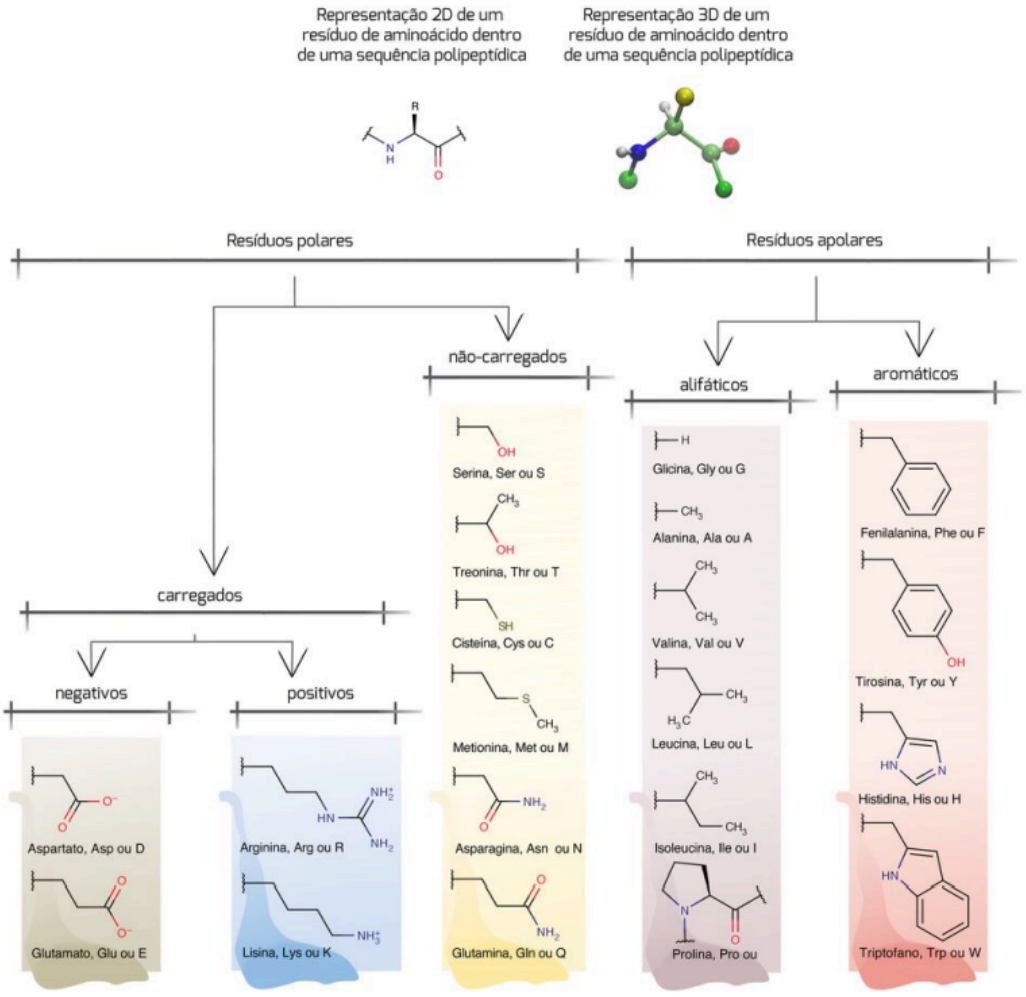


Figura 1. Estrutura dos aminoácidos codificados no genoma, organizados segundo as propriedades de suas cadeias laterais. Acima, o esqueleto peptídico é representado como encontrado dentro de uma proteína, em

representações bi e tridimensional. Nesta última, o grupo R (cadeia lateral) está apresentado como uma esfera amarela, enquanto a continuação da cadeia polipeptídica como esferas verde-escuras. As cadeias laterais estão apresentadas em sua ionização mais comum, plasmática (pH fisiológico, isto é, 7,35 - 7,45). Fonte: (Verli, 2014).

Já a estrutura secundária é decorrente da flexibilidade conformacional ao longo do *backbone* polipeptídico e é ditada pelos ângulos torcionais Φ e Ψ, informação que pode ser graficamente representada em gráficos conhecidos como *plots* de Ramachandran (Fig. 2). (Hollingsworth; Karplus, 2010) Valores repetidos destes ângulos levam a estruturas secundárias regulares conhecidas como α-hélices e β-fita. A α-hélice é o elemento mais comum da estrutura secundária encontrado em proteínas e é caracterizada por dimensões como *pitch* (isto é, distância vertical de voltas consecutivas na hélice) de 5,4 Å, distância translacional de 1,5 Å, e aproximadamente 3,6 resíduos por volta. A β-fita representa uma estrutura estendida com distância de *pitch* de cerca de 7 Å, distância translacional de 3,5 Å e menos resíduos por volta. As α-hélices são estabilizadas por ligações de hidrogênio orientadas paralelamente ao eixo da hélice e formadas entre grupos CO e NH de resíduos separados por 4 resíduos consecutivos. As β-fitas podem formar ligações de hidrogênio com outras β-fitas e formar β-folhas. Os valores de dimensões dados acima são representativos e variam em cada proteína específica. (Verli, 2014; Whitford, 2005)

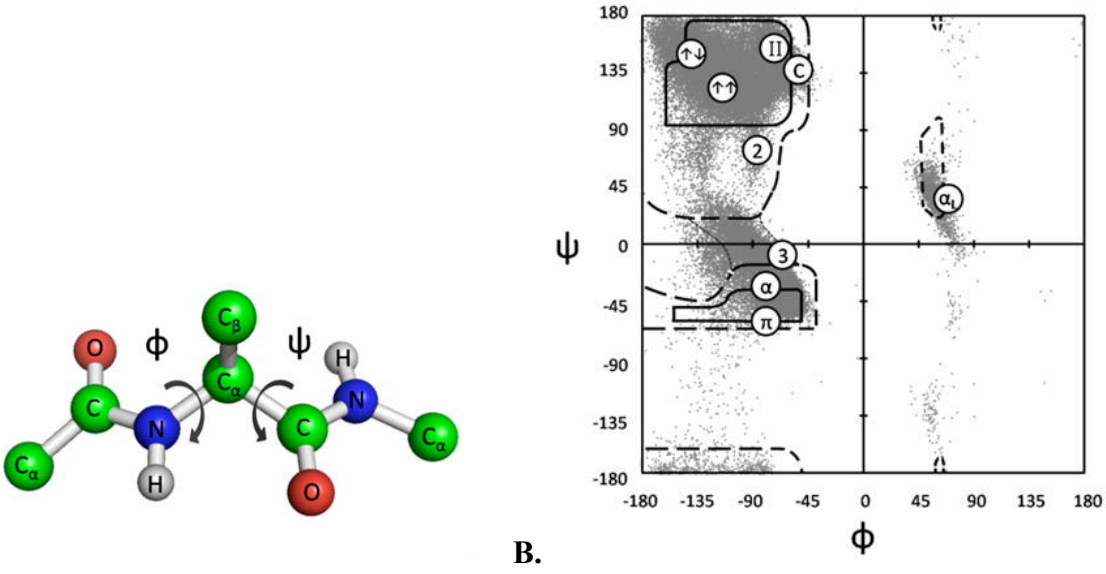


Figura 2. O plot de Ramachandran clássico, ou ϕ, ψ-plot. A) Modelo bola e palito de um dipeptídeo contendo um resíduo de alanina central, indicando que as rotações definidas pelos ângulos torsionais. O ângulo ϕ é definido como o ângulo torsional entre Ci-1-Ni-Cαi-Ci e ψ é definido por Ni-Cαi-Ci- Ni+1. B) O plot de Ramachandran canônico de Ramachandran & Sasisekharan com linhas externas definindo as estruturas classicamente permitidas

(linhas tracejadas), *core* permitido (linhas sólidas), e permitidas no limite extremo (linhas pontilhadas) para um dipeptídeo de alanina. Fonte: Adaptado de: (Hollingsworth; Karplus, 2010).

A estrutura terciária é formada pela organização da estrutura secundária em topologias mais complexas (ou *folds*, do inglês, novelos) resultantes das interações entre resíduos que frequentemente estariam distantes entre si na estrutura primária. Existem vários *folds* identificados na literatura, a exemplo do *bundle* (do inglês, pacote) de 4 hélices, o β-barril, a β-hélice, e o β-propelente. A estrutura terciária é mantida pela magnitude de interações favoráveis que supera a magnitude de interações desfavoráveis. Estas interações podem ser de natureza covalente (como nas pontes dissulfeto formadas entre os grupos tiol de resíduos de cisteína) ou, mais frequentemente, não covalentes (como interações iônicas, hidrofóbicas, vdW e ligações de hidrogênio). Estas interações serão descritas em maior detalhe na secção de Mecânica Molecular (cf. secção 2.2.1) do presente trabalho, e são chave para a compreensão conformacional de proteínas. A estrutura quaternária, por sua vez, é uma propriedade apenas de proteínas com mais de uma cadeia polipeptídica e suas cadeias interagem entre si também por meio das interações supracitadas. (Whitford, 2005) Usamos a proteína sintética Top7 (PDBid 1QYS) como exemplo para representar as estruturas secundária e terciária mais comuns (Fig. 3).

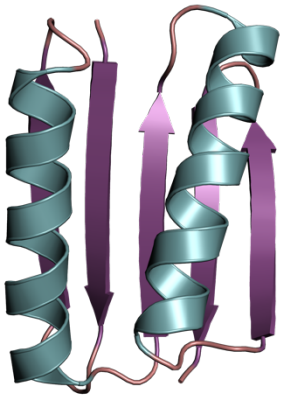


Figura 3. Representação NewCartoon da estrutura terciária da proteína Top7 (PDBid 1QYS). A proteína Top7 foi a primeira proteína sintética engenhada computacionalmente com elevadíssima estabilidade estrutural térmica. Na figura, em ciano estão destacadas as estruturas tipo α-hélice. Em magenta está representada a formação de uma estrutura de β-pregas antiparalelas (as duas primeiras distantes, na estrutura primária, das três últimas, mas próximas na estrutura terciária) coalescendo para configurar uma β-folha com 5 pregas antiparalelas. As figuras são unidas por alças e *loops* (representados em salmão). Proteína desenvolvida por Kuhlman e colaboradores (Kuhlman et al., 2003). Figura de autoria própria no software de visualização molecular PyMol.

2.1.2 Enovelamento Proteico

Compreender a estrutura tridimensional de proteínas é, portanto, central para o entendimento de sua função biológica e para a engenharia de proteínas, que pode ser definida como a manipulação da sequência primária de proteínas em nível molecular, objetivando alterar sua função ou otimizar suas propriedades. (Jozala et al., 2016) Evolutivamente, é sabido que a estrutura terciária de proteínas é mais conservada que a estrutura primária propriamente dita, sendo possível encontrar proteínas com alta similaridade estrutural terciária (conformacional), mas com estruturas primárias (isto é, sequências de resíduos de aminoácidos) distintas. (Verli, 2014) Por outro lado, a estrutura terciária depende da sequência de resíduos na proteína e pequenas modificações podem produzir efeito algum ou desfazer completamente a estrutura terciária da proteína. Este fato é de tamanha relevância para a engenharia de proteínas e sua conclusão, conhecida como hipótese de Anfinsen do enovelamento proteico (ou hipótese termodinâmica), mudou a história da engenharia de proteínas e da bioquímica computacional em geral. (Anfisen, 1973) A partir desta hipótese, segundo a qual a proteína enovelada corresponde ao estado de mínimo de energia global do sistema, foram desenvolvidos modelos termodinâmicos para explicar o enovelamento proteico, o mais famoso dos quais ficou conhecido como funil de enovelamento (ou teoria da superfície de energia), desenvolvido em 1995. Este modelo representa a energia no eixo vertical em função das coordenadas de conformação proteica nos demais eixos cartesianos, de modo que cada ponto corresponde a um determinado confôrmero ou proteoforma e sua energia associada (Fig.4). (Bryngelson et al., 1995) Algumas proteínas podem se enovelar seguindo caminho aproximadamente único (ou múltiplos caminhos com perfil energético semelhante; Figs. 4A, 4D e 4E), seja direto (Fig. 4E), gradual (ou suave; Fig. 4A), com uma (ou mais) etapas metaestáveis (Fig. 4D). Alternativamente, algumas proteínas podem ter dois principais caminhos distintos: um que leva ao estado enovelado sem passar por estados metaestáveis e outro que passa por um ou mais desses estados (Fig. 4C). Porém, o modelo mais apropriado (e menos ideal) para descrição do enovelamento é o apresentado na Fig. 4B, em que existe um panorama energético complexo, com metaestados variáveis e mínimos locais, além de estados de aumento de energia em caminhos que conduzem, no final, ao mínimo global. A proteína poderá seguir múltiplos caminhos até o estado enovelado, mas há um viés energético em direção a este, já que ele representa o mínimo global da hiperssuperfície. A proteína só será funcional, em geral, no mínimo global (isto é, no estado enovelado), mas pode haver exceções a esta regra, particularmente em proteínas que agem como motores moleculares. Como a energia livre é

função de estado, o ΔΔG associado ao enovelamento dependerá apenas dos estados inicial (desenovelado, de maior energia) e do estado final. O enovelamento só será um processo espontâneo se esta diferença for negativa, indicando que o caminho ocorreu na direção do estado menos energético (em relação ao inicial). Estados metaestáveis podem corresponder a múltiplas proteoformas com variados graus de função ou, mais comumente, corresponder a estados erroneamente enovelados (do inglês, *misfolding states*). Estados de maior energia podem fazer parte do caminho de enovelamento e são conhecidos como glóbulos fundidos (no inglês, *molten globules*). (Whitford, 2005) O glóbulo fundido corresponde ao estado de colapso inicial de uma proteína dobrada, comumente apresentando raio de giro semelhante (5 a 10% maior) ao da proteína enovelada, mas com desordem considerável das cadeias laterais e estabilidade termodinâmica insignificante (comumente representando apenas um estado intermediário para o enovelamento completo). (Voet; Voet, 2013)

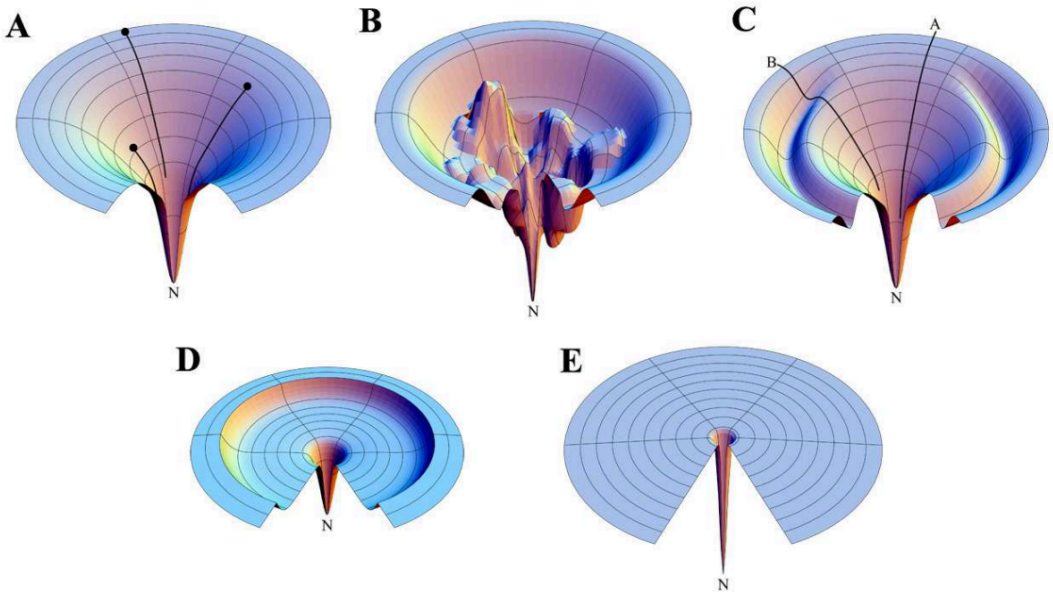


Figura 4. Representação dos possíveis funis de enovelamento para uma determinada proteína. O modelo dos funis de enovelamento ficou conhecido como teoria da paisagem da superfície de energia. Fonte: Adaptado de (Dill; Chan, 1997).

Caso a proteína procurasse por todos os caminhos possíveis de maneira exaustiva e sem tendência ao mínimo de energia local, o processo de enovelamento proteico seria extremamente demorado (e, consequentemente, sua modelagem computacional também o seria), fato conhecido como paradoxo de Levinthal. Este paradoxo tem sua fundamentação matemática na aplicação de teoremas probabilísticos básicos aos blocos de construção das proteínas: cada "posição" pode ter 20 resíduos distintos ocupando-as, resultando em 20 elevado a N

possibilidades de sequências para uma proteína de N resíduos (explosão combinatória rápida). Pode-se tratar de maneira semelhante o problema do caminho do enovelamento de energia dada uma proteína com sequência conhecida, por analogia ao problema anterior. Para uma proteína de apenas 100 resíduos, o tempo necessário para enovelamento seria da ordem de magnitude de 10E87 (em segundos), tempo superior à idade aparente do universo (da ordem de grandeza de 10E17, em segundos). (Voet; Voet, 2013) Porém, na prática, as biomoléculas não podem procurar aleatória e exaustivamente todos os estados possíveis da hiperssuperfície, já que o enovelamento proteico é um processo bastante rápido. É o viés (ou tendência) aos mínimos de energia que representa uma das possíveis soluções deste paradoxo: o dobramento por uma rota (ou conjunto de rotas) é acompanhado por um aumento abrupto na sua estabilidade conformacional (representada pela redução da energia livre). Há ainda outros modelos para resolver o paradoxo de Levinthal utilizando conhecimentos físicos (como a teoria da paisagem apresentada) ou não (como modelos de ML), que fogem ao escopo da discussão deste trabalho. (Ivankov; Finkelstein, 2020)

A identificação e manipulação das sequências de resíduos na estrutura primária é ferramenta chave para o desenvolvimento das técnicas de engenharia de proteínas, que podem ser geralmente classificadas em dois tipos de abordagem: mutagênese direcionada ou mutagênese aleatória. (Youvan, 1995) A última ocorre por meio de evolução direcionada, gerando sequências randomicamente (abordagem empírica). A primeira envolve o desenho racional de proteínas, empregando princípios de termodinâmica estatística e cálculos computacionais para guiar a sequência de uma proteína. Esta tem como desvantagem a necessidade de conhecimento da estrutura da proteína. (Ulmer, 1983) Essa desvantagem, porém, vem se mitigando gradualmente com a evolução de ferramentas computacionais de predição estrutural de proteínas com alta fidelidade, principalmente através de machine-learning, realidade em programas como AlphaFold (Abramson et al., 2024; Jumper et al., 2021), RoseTTAFold (Baek et al., 2021) e RFDiffusion (Watson et al., 2023). Discutiremos mais detalhes da engenharia de proteínas em secções subsequentes (cf. secção 2.3).

2.2 MODELAGEM MOLECULAR DE PROTEÍNAS

2.2.1 Mecânica molecular

O grande interesse em moléculas biológicas, motivado pela necessidade de solução de problemas desta natureza e pelo avanço em técnicas de resolução estrutural como difratômetros

de raios X, levou ao problema do custo computacional proibitivo para simulação de tais sistemas (frequentemente muito grandes) à época. Métodos quânticos *ab initio* lidam com os elétrons em um sistema, de modo que mesmo ignorando alguns destes (como em esquemas semiempíricos), um número muito grande de partículas ainda deve ser considerado. O tratamento da equação de Schroedinger em sistemas com grande número de partículas representava custo computacional proibitivo à época. Tais limitações prejudicam consideravelmente a aplicação de métodos quânticos a sistemas biológicos complexos, e mesmo com o avanço tecnológico recente muitas destas limitações permanecem. Surgiu, neste contexto, a necessidade de trabalhar com modelos mais simplificados que pudessem ser computados pelos sistemas de processamento da época em tempo hábil e com descrição suficientemente razoável de propriedades físico-químicas de sistemas biológicos (ou sistemas de grande número de partículas, qualquer que seja sua natureza). A partir disso, a mecânica molecular se desenvolveu com a criação de modelos de potenciais empíricos conservativos (que, em sua maioria, não trabalham explicitamente com a equação de Schroedinger), ainda que sua aplicação seja limitada a problemas que não dependem da descrição eletrônica dos sistemas. (Schlick, 2010) Na presente secção (bem como na secção de abordagens MC) serão discutidos alguns aspectos referentes à mecânica molecular (base das simulações em MD) utilizando como base os livros texto de referência na área "*Computer Simulation of Liquids*" e "*Molecular Modelling: Principles and Applications*". (Allen; Tildesley, 1989; Leach, 2001) Secções não referenciadas correspondem a materiais teóricos baseados nestas obras. Ainda assim, citações indiretas do material foram mantidas e estão devidamente sinalizadas.

A mecânica molecular é, na maioria das vezes, baseada em modelos simples de interações dentro de um sistema, cujas contribuições de processos como o alongamento de ligações (e de seus ângulos de abertura e fechamento), rotações em torno de eixos e deformações dos ângulos de diedros (conhecidas por seu termo em inglês *out of plane bending*, torções para fora do plano). Os trabalhos de mecânica molecular dependem da validade de diversas assumpções. A mais importante delas é a **aproximação de Born-Oppenheimer**, sem a qual não seria possível tratar a função de energia a partir de coordenadas nucleares. (Morgon; Coutinho, 2007) Outras aproximações envolvendo funções simples (a exemplo do oscilador harmônico clássico) ou de maior complexidade (mas, idealmente, de custo computacional inferior ao tratamento por métodos QM) são utilizadas para descrever contribuições para o campo de força com performance razoável. Transferabilidade é um atributo chave para um campo de força, por permitir que um conjunto de parâmetros desenvolvidos e testados para um

conjunto relativamente pequeno de casos seja aplicável para um conjunto de problemas mais amplo, trazendo robustez ao campo de força. (Leach, 2001)

Para um sistema de *N* partículas com posição *r*, a energia potencial *V* poderá ser expressa como uma função destes termos. As várias contribuições energéticas para esta podem ser assim representadas:

$$V(r^N) = \sum_{ligações} \frac{k_i}{2}(l_i - l_{i.0})^2 + \sum_{ângulos} \frac{k_i}{2}(\theta_i - \theta_{i,0})^2 + \sum_{torções} \frac{V_n}{2}\,(1 + \cos(\eta\omega - \gamma)) + \sum_{i=1}^{N}\sum_{j=i+1}^{N}(4\varepsilon_{ij}\left[(\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^{6}\right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}})$$

O primeiro, segundo, terceiro e quarto termos representam, respectivamente, as interações entre termos ligados modeladas por potencial harmônico em função do comprimento de ligação l; soma de todos os ângulos também modeladas por potencial harmônico; potencial rotacional correspondente à rotação da ligação interatômica; contribuições não ligadas modeladas por um termo de Lennard-Jonnes 12-6 (LJ 12-6) acoplado a um termo coulômbico para modelagem eletrostática. Inicialmente, cabe detalhar os aspectos mais importantes de cada um destes termos. Entretanto, é importante ter em mente que as equações acima são gerais e cada campo de força terá um tratamento específico para cada termo, podendo inclusive deixar de computar algum termo ou acrescentar termos novos com finalidades específicas. Idealmente, os termos de um campo de força devem ter sido parametrizados para que haja algum grau de correlação experimental, de acordo com a finalidade do campo de força desenvolvido.

O potencial harmônico de ligação é gerado considerando um sistema de esferas rígidas de dois átomos ligados por uma mola de constante de Hooke *k*. O segundo termo é modelado analogamente, mas trata do ângulo entre três esferas rígidas, ao invés da distância entre duas destas, e terá constante análoga. A modelagem em sistema de mola é compatível com potenciais mais realistas (como o potencial de Morse) próximo a regiões de mínimo de energia, embora não seja o modelo mais adequado para simulações muito distantes do mínimo, em que a correlação com resultados experimentais é reduzida (Fig. 5).

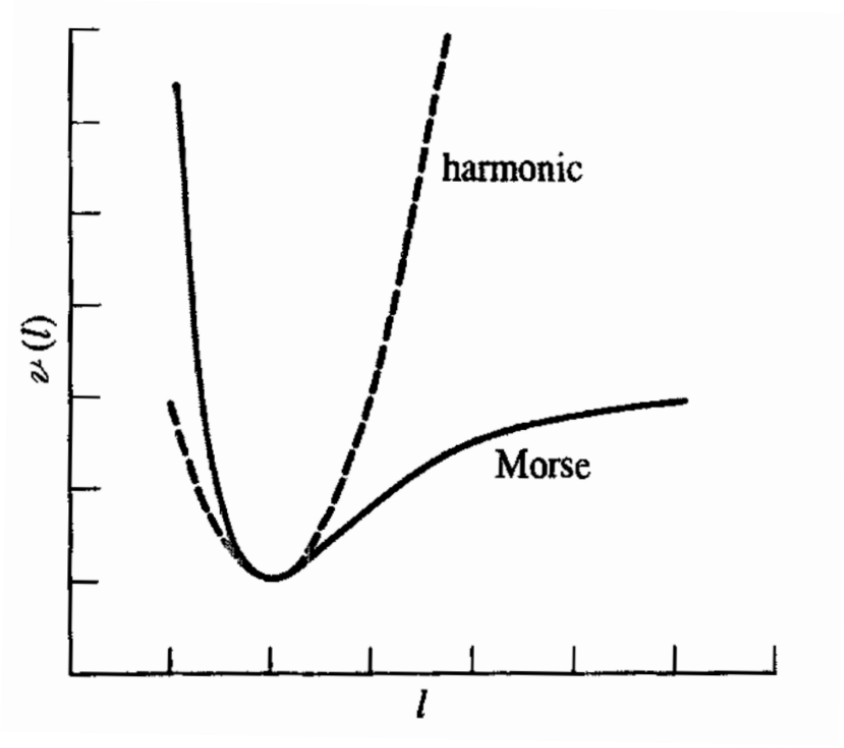


Figura 5. Comparação do potencial harmônico simples (lei de Hooke) com o potencial de Morse. Potencial harmônico marcado em linha tracejada, com legenda oriunda do inglês (*harmonic*, harmônico). Potencial de Morse marcado em linha contínua. Eixo das abscissas representando distância de ligação *l*; eixo das ordenadas representando modelagem de energia potencial *v* em função de *l*. Chama-se atenção para a superposição dos modelos nas regiões de mínimo de energia. Adaptado de (Leach, 2001).

O potencial torcional (terceiro termo acima representado) corresponde ao potencial gerado por rotações de quatro átomos consecutivos obtidas a partir de variações dos seus ângulos diedrais. Tais variações são de especial importância na modelagem de proteínas, já que, como exposto previamente (cf. secção 2.1), os ângulos de diedro são largamente responsáveis pela estrutura tridimensional proteica, bem como por sua flexibilidade e funcionalidade biológica. A modelagem tradicionalmente é feita utilizando o cosseno das variações diedrais, mas implementações mais simples podem ser atribuídas e a função cosseno pode ser simplificada por expansão de Taylor com truncamento no termo adequado, a depender da modelagem funcional do campo de força em questão.

O quarto termo representa interações entre átomos não ligados no sistema, isto é, átomos separados por uma distância superior a 4 átomos na cadeia, ou em cadeias distintas. Esta definição é aproximada e decorre do fato de outras interações (como o primeiro e o segundo termo, por exemplo) terem contribuição muito superior para átomos separados por distância inferior a 4 átomos na cadeia. A modelagem deste termo envolve uma porção coulômbica e

uma porção de interações modeladas como potencial de Lennard Jones 12-6 (LJ 12-6). A porção coulômbica pode ser tratada com maior precisão (porém maior custo computacional) por expansão do modelo eletrônico adotado para computar interações dipolares, quadripolares, entre outras. Analogamente, a porção LJ 12-6 pode ser modelada com outros exponenciais a depender do campo de força adotado (como o LJ 9-6), mas o modelo LJ 12-6 tem menor custo computacional já que o cálculo do termo de exponente 12 é efetuado simplesmente pelo quadrado do termo de exponente 6.

É possível também definir modelos de polarizabilidade explicitamente, mais comumente através do uso de três abordagens: modelo de pontos multipolares induzido, modelo de cargas flutuantes e modelo de oscilador de Drude (do inglês, respectivamente, *induced point multipole model*, *fluctuating charge model*, *Drude oscillator model*). (Allen; Tildesley, 1989) Além disso, é possível utilizar potenciais mais precisos (como o potencial de Buckingham) para modelagem das interações vdW, mas isto geralmente representa custo computacional muito maior, já que a maioria destes modelos utiliza exponenciais e requer tratamento específico com expansão de Taylor de cada potencial (embora usualmente truncamentos do termo de segunda ordem sejam suficientemente adequados para diversas situações). (Leach, 2001) Como há muitos campos de força diferentes, com sua respectiva parametrização e funcionalidade específicos, Allen & Tildesley propõem uma classificação geral de campos de força. Campos de classe I tem a forma funcional geral da equação acima exposta. Campos de classe II normalmente acrescentam termos cúbicos ou anarmônicos para os potenciais de estiramento e definem explicitamente elementos não diagnonalizados na matriz de força constante. Campos de classe III vão além destas prescrições básicas e incluem representações eletrostáticas mais acuradas entre moléculas e algum modelo de polarizabilidade. É difícil definir recomendações gerais para a escolha do campo de força, mas uma estratégia sensível para a escolha pode envolver o teste de campos de força distintos frente aos mesmos problemas para compreender a sensibilidade dos resultados à escolha. (Allen; Tildesley, 1989)

2.2.2 Geração de configurações

A gênese de conformações para uma determinada molécula é classicamente dividida em dois tipos de estratégia: as que usam modelos baseados em MD e as que usam modelos baseados em MC. As configurações geradas por MD partem de uma configuração inicial (geralmente obtida por dados experimentais de resolução estrutural seguidos de minimização de energia) e são calculadas ao longo do tempo pela integração das equações de movimento newtonianas.

Para isto, convém utilizar os modelos de mecânica molecular supracitados, logrando evolução temporal das estruturas em questão. As moléculas em configuração inicial têm suas forças calculadas pelo campo de força em questão (a partir das interações intermoleculares esperadas para o sistema) e a posição das partículas ao longo do passo temporal escolhido é calculada, obedecendo às condições impostas pelo ensemble e pelo campo de força escolhidos. São atribuídas velocidades (v) iniciais às partículas do sistema, em geral calculadas a partir da temperatura do sistema, conforme distribuição de Maxwell-Boltzmann:

$$f(v)dv = 4\pi \left(\frac{m}{2\pi kT}\right)^{3/2} v^2 \exp\left(-\frac{mv^2}{2kT}\right) dv$$

Em que $v = v_x^2 + v_y^2 + v_z^2$, e m, k e T representam a massa da partícula, a constante de Boltzmann e a temperatura do sistema. (McQuarrie, 2003) Pode-se, ainda, atribuir velocidade inicial nula a todos os átomos e deixar que o sistema evolua a partir do repouso; qualquer que seja o método de atribuição de velocidades iniciais adotado, é importante que a soma das velocidades sobre todas as partículas do sistema seja um vetor nulo, evitando o deslocamento da caixa de simulação como um todo. (Morgon; Coutinho, 2007)

A seguir, as equações newtonianas do movimento são integradas (EDO de 2º grau) conforme o algoritmo de integração escolhido, sendo os mais comuns *Leap-Frog*, Verlet clássico e Verlet velocidade. A partir destes, é possível obter as configurações a cada instante temporal, considerando o passo temporal adotado e, em geral, tomando o início da simulação como instante de tempo nulo (referencial temporal). (Martinez; Boriz; Skaf, 2007) Não é objetivo do presente trabalho discutir em detalhes tais algoritmos de integração, uma vez que as simulações de dinâmica molecular fogem ao escopo do presente estudo. As configurações geradas são armazenadas em arquivos de trajetória, que podem ser analisados com softwares de visualização molecular e ferramentas computacionais de análise estatística (como clusterização, análise de componentes principais, entre outras). (McGibbon et al., 2015) Métodos de MD são de fundamental importância em química teórica, em especial aplicada ao estudo de biomoléculas, conforme explicado em secção anterior (cf. secção 2.2.2). Entretanto, embora de grande importância geral, tais métodos não foram adotados no presente trabalho e, portanto, sua discussão detalhada foge ao escopo deste material.

Métodos MC podem gerar configurações de maneira estocástica, sendo particularmente úteis para estudo de fenômenos que independem do tempo (ou cuja influência temporal não é objeto de estudo da pesquisa em questão). Por se tratar de modelos estocásticos, a gênese de

configurações seria, *a priori*, aleatória, porém a exploração do espaço de fases é eficiente pela adoção dos critérios de Metropolis, que permitem criar vieses exploratórios em direção às configurações correspondentes aos mínimos da hiperssuperfície. O método MC será explicado em mais detalhes no tópico 2.4 do presente trabalho (inclusive a prescrição da gênese de novas configurações), dada a sua grande importância no estudo de biomoléculas em geral e, especialmente, pela sua implementação no pacote do software Rosetta, utilizado como base metodológica para esta investigação.

Atualmente, com a introdução de metodologias baseadas em aprendizagem de máquina (com especial destaque para algoritmos de redes neurais profundas), pode haver gênese de gerações baseadas nestes algoritmos. Entretanto, rotineiramente, estas só são utilizadas como input inicial para programas de MC ou MM. Comumente, configurações são geradas em programas como AF, servindo como input para simulações de MC ou MM ou do próprio programa Rosetta citado no presente trabalho. O caso do programa AF é de especial interesse por sua notória importância na pesquisa em química estrutural proteica, rendendo aos desenvolvedores o prêmio Nobel da química em 2024, compartilhado com David Baker, criador do software Rosetta. No caso do AF, a partir da estrutura primária proteica (isto é, a sequência de resíduos de aminoácidos), que é dada como *input* ao programa, são geradas configurações de possíveis enovelamentos proteicos, problema de notória dificuldade na ciência contemporânea. O enovelamento é calculado a partir de parâmetros de aprendizagem de máquina, que utilizaram como conjunto de teste/treino os dados estruturais de praticamente todas as proteínas no banco de dados do RCSB. Após a previsão dos enovelamentos, a estrutura final é relaxada em campo de força AMBER (adaptado especificamente para este fim), gerando um arquivo de PDB da proteína enovelada como *output*. Este arquivo não corresponde a uma trajetória, mas é frequentemente utilizado como configuração inicial de *input* para outros programas e simulações, geralmente após rodadas de otimização estrutural mais robustas. (Abramson et al., 2024; Jumper et al., 2021)

2.3 ENGENHARIA COMPUTACIONAL DE PROTEÍNAS

Embora os problemas do enovelamento proteico não tinham sido completamente solucionados *ab initio*, foram feitos progressos significativos na solução do problema inverso: a gênese de sequências polipeptídicas para estruturas tridimensionais específicas, isto é, a engenharia proteica. Geralmente, o processo tem início com a seleção de uma estrutura alvo e procura-se uma sequência de aminoácidos que forme essa estrutura. O peptídeo ou proteína

engenhado é então sintetizado e sua estrutura é elucidada. Uma engenharia bem-sucedida necessita não apenas que o dobramento desejado seja estável, mas que todos os demais dobramentos sejam significativamente menos estáveis (por cerca de 15 a 40 kJ/mol). (Voet; Voet, 2013) O processo é consideravelmente mais desafiador quando a estrutura alvo não é conhecida, isto é, quando o design da proteína é realizado com topologia nova (design *de novo*).

A engenharia de proteínas é um campo relativamente novo nas ciências, dada a elevada complexidade e custo do seu estudo. Apesar disso, em 2023 já era calculado que aproximadamente metade das terapias médicas em humanos utilizavam-nas (incluindo anticorpos monoclonais e policlonais, enzimas, fatores de coagulação, hormônios proteicos e citocinas inflamatórias, entre outras). A engenharia de proteínas é utilizada para melhorar propriedades desejadas de compostos existentes (como estabilidade, farmacocinética, potência, especificidade do alvo, biocompatibilidade) ou para produzir respostas a desafios terapêuticos utilizando design *de novo*. Há estratégias experimentais e computacionais para o design de proteínas. No presente trabalho focaremos nas últimas, por terem sido as utilizadas aqui e pela sua crescente importância com avanços em metodologias baseadas em ML. Apesar da existência de diversas estratégias de engenharia proteica com as finalidades supracitadas, apenas as estratégias computacionais são sistematicamente empregadas para todas as finalidades supracitadas, sem exceções. (Ebrahimi; Samanta, 2023)

2.3.1 Desenho computacional de proteínas

Costumeiramente, as estratégias de design de proteínas são classificadas em dois grupos: design racional e evolução direcionada. A metodologia *sequence saturation mutagenesis* (SeSaM) representa o paradigma das estratégias de evolução direcionada. SeSaM é um método de mutagênese quimio-enzimática aleatória aplicada para a evolução direcionada de proteínas. (Wong; Zhurina; Schwaneberg, 2006) De modo geral, as metodologias de evolução direcionada consistem em processos cíclicos que alternam entre a etapas de diversificação genética e etapas de triagem e seleção de variantes gênicas funcionais, a partir das quais as proteínas serão sintetizadas. (Packer; Liu, 2015) Este grupo de metodologias não é o foco do presente trabalho, mas as definições supracitadas serão úteis para comparação com as estratégias de design racional aqui empregadas.

As metodologias de design racional de proteínas classicamente realizam mutações em sítios específicos e avaliavam as contribuições termodinâmicas destas mutações para

parâmetros desejados de enovelamento e função proteica. As mutações podem ser completamente aleatórias ou direcionadas para resíduos específicos, baseando-se na intuição química dos pesquisadores e do tipo de problema em questão. No exemplo do presente trabalho, utilizamos uma estratégia racional de mutação de resíduos chave baseados em dados genômicos para modelar proteínas existentes (mas sem resolução estrutural) e investigar computacionalmente propriedades termodinâmicas de interação, como ficará claro na secção de procedimentos metodológicos. Para isto, utilizamos uma abordagem clássica de design racional com mutagênese de resíduos específicos. As metodologias apresentadas nesta secção tem ampla flexibilidade para a solução de problemas biológicos, mas há diversos casos na literatura de solução de problemas de outras naturezas (ambientais, industriais, entre outras).

O "ciclo de design proteico" é central nas abordagens racionais, e costumeiramente envolve metodologias computacionais e experimentais. O ponto de partida é o desenvolvimento de um modelo molecular baseado em regras de estrutura e função proteica, combinados com algoritmos para suas aplicações. Após o design racional computacional, ocorre a construção (síntese) experimental da proteína engenhada e a análise das suas propriedades. Se o resultado experimental for falho, ou parcialmente exitoso, uma nova iteração no ciclo de design se iniciará: complexidade é adicionada ao problema e às metodologias, regras e parâmetros são revistos e algoritmos de aplicação são modificados. (Hellinga, 1997) Casos de design proteico de sucesso são cada vez mais frequentes na literatura, motivados em grande parte pelos trabalhos do químico David Baker, vencedor do prêmio Nobel de Química de 2024. O primeiro caso de sucesso em engenharia *de novo* (proteína Top7, Fig. 3) já foi exposto no tópico 2.1 do presente trabalho, mas muito progresso foi feito desde então. Uma das principais contribuições deste cientista para a área foi o desenvolvimento do software Rosetta, que utiliza metodologias baseadas em MC para design proteico.

Nas próximas secções examinaremos alguns aspectos relevantes do método MC (em particular do algoritmo de Metropolis) e, a seguir, da implementação específica no Rosetta. Atualmente, as estratégias de design proteico são tão diversas e vastas que nenhum trabalho seria capaz, sozinho, de tratar detalhadamente de todas elas. Por este motivo, optou-se por focar nas metodologias de maior relevância para o presente trabalho e detalhar outras metodologias conforme forem expostas nas demais secções de desenvolvimento metodológico.

2.4 MÉTODO DE MONTE CARLO

As simulações de Monte Carlo (MC) e Metropolis (um algoritmo de amostragem direcionada para simulações de Monte Carlo) têm desempenhado um papel fundamental na investigação de sistemas complexos em diversas disciplinas científicas. O método MC, desenvolvido por von Neumann, Ulam e Metropolis (1949) para estudar a difusão de nêutrons em material fissionável, se baseia na geração de trajetórias aleatórias para análise de fenômenos físicos e químicos, permitindo a obtenção de informações estatisticamente significativas sobre o comportamento desses sistemas. (Allen; Tildesley, 1989; Metropolis; Ulam, 1949) A introdução do algoritmo de Metropolis, desenvolvido por Metropolis et al. (1953), representa um avanço fundamental no contexto das simulações de Monte Carlo. Ao contrário dos métodos tradicionais de Monte Carlo, que amostram aleatoriamente configurações sem considerar mudanças de energia, o algoritmo de Metropolis introduz um critério de aceitação probabilístico com base na diferença de energia entre as configurações propostas e atuais. (Metropolis et al., 1953) Nesta secção do presente trabalho, serão discutidos alguns aspectos referentes ao método MC. Para isto, utilizou-se principalmente, além do trabalho original de Metropolis e colaboradores (1953), os livros texto de referência na área “Computer Simulation of Liquids” e “Molecular Modelling: Principles and Applications”. (Allen; Tildesley, 1989; Leach, 2001) Secções não referenciadas correspondem a uma destas obras. Ainda assim, citações indiretas do material foram mantidas e estão devidamente sinalizadas.

O uso do algoritmo de Metropolis permite um aumento considerável na eficiência computacional em relação ao método MC puro (isto é, ao método de busca exaustiva), uma vez que introduz o critério de aceitação supracitado e a amostragem deixa de ser inteiramente aleatória *a priori* pura, passando a se tornar uma amostragem aleatória *a priori* mista (ou até mesmo *a posteriori*, em casos específicos em que a hipersuperfície deve ser conhecida de antemão, como no método *Jumping Between Wells*, JBW). O critério envolve a aceitação de conformações com energia menor e a escolha da aceitação de conformações de energia maior baseada nos pesos de Boltzmann para aquela conformação nova gerada. Essa inovação melhora significativamente a eficiência e a precisão das simulações de MC, especialmente em sistemas com paisagens energéticas complexas ou que sofrem transições de fase. Sem o algoritmo de Metropolis, as simulações de Monte Carlo frequentemente enfrentam dificuldades para explorar o espaço de configurações de forma eficaz, resultando em amostragem ineficiente e resultados enviesados, especialmente em sistemas com superfícies energéticas acidentadas ou grandes barreiras energéticas (por exemplo, estados metaestáveis). Portanto, a incorporação do

algoritmo de Metropolis representa um avanço crucial nas simulações de Monte Carlo, permitindo o estudo de uma gama mais ampla de sistemas com maior precisão e eficiência computacional. (Leach, 2001) Adicionalmente, encontramos uma síntese do algoritmo de Metropolis no livro "Métodos de Química Teórica e Modelagem Molecular", cujos organizadores produziram o software DICE® de simulação de MC, que pode ser assim descrita:

1. Especificar uma condição inicial x_i para o sitema;
2. Gerar uma nova configuração aleatória x_j;
3. Calcular a mudança de energia $\Delta\mathcal{H}_{i,j}$ em que *i* e *j* são estados consecutivos ($i + 1 = j$) e $\mathcal{H}$ representa o hamiltoniano do sistema;
4. Se $\Delta\mathcal{H}_{i,j} < 0$, aceitar o movimento e retornar ao passo 2;
5. Se $\Delta\mathcal{H}_{i,j} > 0$, gerar um número aleatório $\zeta \in [0,1]$;
6. Se $p < \exp(-\beta\,\mathcal{H}_m)$, com $\beta = {}^{1}/_{K_bT}$(em que K_bT representa o produto entre a constante de Boltzmann e a temperatura), aceitar a nova configuração e retornar ao passo 2;
7. Caso contrário, contar a configuração anterior como nova configuração e retornar ao passo 2. (Morgon; Coutinho, 2007)

No contexto da física estatística, a hipótese ergódica, conforme estabelecida por Birkhoff em 1931, desempenha um papel crucial na interpretação dos resultados obtidos por meio das simulações de Monte Carlo. A hipótese ergódica estabelece relações entre as médias temporal e ensemble, garantindo a convergência dos resultados em simulações de longo prazo para valores estatisticamente significativos. (Allen; Tildesley, 1989) A aplicação do teorema ergódico em sistemas mecânicos permite concluir resultados acerca do comportamento do sistema em períodos suficientemente longos, sem a demanda do cálculo determinístico das integrações das equações de movimento para cada partícula (o que é geralmente impossível), desde que a condição de transitividade métrica do sistema seja satisfeita. Esta condição pode ser assim definida: um sistema de um parâmetro *P* em um espaço de medida *M* apresentará transitividade métrica desde que qualquer conjunto mensurável invariante sob *P* para todo valor temporal *t* apresente medida zero, ou que seu complemento apresente medida zero. Isto significa que o fluxo não é decomponível ou irredutível, no sentido de que não poderá ser

decomposto em uma junção ou disjunção dos seus subfluxos. Além disso, não há funções mensuráveis invariantes sob o fluxo (ou transformação *P*). (Moore, 2015) Quando comparando sistemas de Dinâmica Molecular (MD) com sistemas de MC, desde que ambos tenham amostragem suficientemente exaustiva (isto é, longa para a MD e com grande número de ensembles configuracionais para a MC), os resultados de parâmetros do sistema calculados deterministicamente ao longo do tempo na MD devem ser equiparáveis aos resultados obtidos estatisticamente dos ensembles na MC, desde que a exploração do espaço amostral (ou temporal, para MD) seja suficientemente exaustiva. (Leach, 2001) Idealmente, tal condição só seria atingida com a exploração completa da hiperssuperfície, mas nenhum modelo computacional atualmente é capaz de lograr tal exploração em sistemas complexos.

O algoritmo de Metrópolis se destaca particularmente em relação às metodologias de MC tradicionais por impor uma amostragem melhorada, conhecida como *importance sampling* (do inglês, amostragem direcionada). Tal amostragem foi inicialmente desenvolvida para o ensemble NVT no trabalho original de Metropolis e colaboradores (1953), em resposta à dificuldade de encontrar um método adequado para gerar uma sequência de estados aleatórios tal que ao final da simulação cada estado tenha ocorrido com a probabilidade adequada. É possível fazê-lo sem precisar calcular o fator de normalização para $\rho(x)$ (em que ρ denota a função de densidade de probabilidade do evento *x*), isto é, a função de partição do ensemble. Para isto, utiliza-se uma cadeia de Markov dos estados da simulação construída de forma que ela tenha uma distribuição limite da função de partição do ensemble. Tomaremos como exemplo canônico no desenvolvimento das equações a seguir o ensemble NVT, por ter sido o utilizado no trabalho original de Metropolis, conforme mencionado acima. Assim, tomando ρ_m e ρ_n como as probabilidades do sistema estar nos estados *m* e *n* consecutivos (para um espaço de fases de *{1, 2, 3, ..., m, n, ...}*), constrói-se uma matriz de transição denominada π_{mn} que denota a probabilidade de o sistema transitar do estado *m* para o estado *n*, tal que:

$$\rho_n = \pi_{mn}\rho_m$$

$$\rho_m = \pi_{nm}\rho_n$$

A partir dos conhecimentos básicos de mecânica estatística, decorre que, para o ensemble canônico (NVT), as probabilidades acima podem ser escritas como:

$$\rho_m = \frac{\exp\left(-\beta\,\mathcal{H}_m\right)}{\sum_{\mathcal{H}_m}\exp\left(-\beta\,\mathcal{H}_m\right)} = \frac{\exp\left(-\beta\,\mathcal{H}_m\right)}{Z}$$

$$\rho_n = \frac{\exp\left(-\beta\,\mathcal{H}_n\right)}{\sum_{\mathcal{H}_n}\exp\left(-\beta\,\mathcal{H}_n\right)} = \frac{\exp\left(-\beta\,\mathcal{H}_m\right)}{Z}$$

Em que Z representa a função de partição do ensemble canônico. Assumindo a condição de reversibilidade microscópica de Boltzmann, isto é, da equiprobabilidade dos microestados, teremos que os valores de ρ_m e de ρ_n devem ser iguais, donde resulta que:

$$\pi_{mn}\rho_m = \pi_{nm}\rho_n$$

$$\frac{\rho_n}{\pi_{mn}} = \frac{\rho_m}{\pi_{nm}}$$

$$\frac{\pi_{mn}}{\pi_{nm}} = \frac{\rho_n}{\rho_m} = \frac{\exp(-\beta\,\mathcal{H}_n)\,/\,Z}{\exp(-\beta\,\mathcal{H}_m)\,/\,Z} = \exp\left(-\beta\,\Delta\mathcal{H}_{m,n}\right)$$

Na equação acima o termo $\Delta\mathcal{H}_{m,n}$ representa a diferença entre os hamiltonianos do sistema nas condições m e n, isto é, a diferença entre a energia final e inicial da transição considerada entre os microestados. Esta notação evidencia a dependência da transição dos estados em relação à sua energia, o que justifica algumas das etapas do método de Metropolis anteriormente mencionadas, cabendo relembrá-las. Uma nova configuração n será aceita se representar redução na energia do sistema. Caso contrário, um número aleatório ($\zeta \in [0,1]$)é gerado e comparado com os valores de $\exp\left(-\beta\,\Delta\mathcal{H}_{m,n}\right)$, expressão oriunda da dedução acima, sendo aceito se for menor que este. A escolha de uma matriz de transição apropriada e de um algoritmo adequado de gênese de números aleatórios por sistemas computacionais (ou, mais precisamente, pseudoaleatórios) é de grande importância para a implementação adequada do algoritmo, mas esta discussão foge ao escopo do presente trabalho.

As equações acima foram representadas em respeito ao ensemble canônico, conforme originalmente apresentado no trabalho de Metropolis e colaboradores (1953), porém, em sistemas biológicos é mais frequente a utilização do ensemble de pressões constantes (NpT), em que a probabilidade de um estado m pode ser assim descrita:

$$\rho_m = \frac{\exp\left(-\beta\,\mathcal{H}_m\right)}{\sum_{\mathcal{H}_m}\exp(-\beta\,\mathcal{H}_m)\exp\left(-\beta\,pV\right)} = \frac{\exp\left(-\beta\,\mathcal{H}_m\right)}{Y}$$

É evidente que, apesar da função de partição ser diferente da utilizada no artigo original, o desenvolvimento matemático anteriormente definido é válido também nesse ensemble, guardando a dependência da transição de estados com o fator exp $(-\beta\ \Delta\mathcal{H}_{m,n})$.

Novas configurações do sistema podem ser geradas utilizando movimentos aleatórios de translação e rotação combinados, ou outros movimentos específicos de acordo com o ensemble (no *NpT*, por exemplo, os movimentos são de expansão ou contração da caixa de simulação). Na prática, a nova configuração pode ser obtida transladando o centro de massa da *a*-ésima molécula ao longo dos eixos cartesianos fixos (*x, y, z*) no espaço conforme a seguinte prescrição:

$$x_{a,v'} = x_{a,v} + (2\xi_1 - 1)\Delta r_{máx}$$

$$y_{a,v'} = y_{a,v} + (2\xi_2 - 1)\Delta r_{máx}$$

$$z_{a,v'} = z_{a,v} + (2\xi_3 - 1)\Delta r_{máx}$$

Em que os ξ_κ representam os números aleatórios sorteados no intervalo [0,1], e $\Delta r_{máx}$ representa o máximo deslocamento permitido na caixa. (Morgon; Coutinho, 2007)

Adicionalmente, é valido mencionar a importância de algoritmos de amostragem otimizada em MC. O algoritmo *Jump-Walking* (JW) emergiu como uma ferramenta eficaz na realização de simulações de Monte Carlo de sistemas complexos. Seu método envolve a computação de duas corridas em paralelo em temperaturas distintas (uma delas a temperatura planejada para aquele sistema; outra delas em um valor de temperatura elevado). A temperatura mais elevada idealmente deve ser suficiente para garantir acesso probabilístico a estados de maior energia. A dependência da temperatura foi evidenciada nas equações anteriores, representada pelo fator β. O algoritmo computa ambas simultaneamente e troca as coordenadas de posição entre os dois sistemas esporadicamente, permitindo que o sistema de menor temperatura consiga explorar regiões que não seriam exploradas naquele valor térmico, pela distribuição de Boltzmann para aquele valor (isto é, garante acesso a uma distribuição de Boltzmann para temperaturas elevadas). Este algoritmo permite uma exploração eficiente do espaço configuracional de estados, resultando na geração de amostras representativas com menor esforço computacional, já que a variação de temperatura não demanda maior número de passos iterativos para cada simulação. Por outro lado, o algoritmo foi desenhado de modo que duas computações sejam realizadas em paralelo, o que de fato dobra o número de passos iterativos (já que há o dobro de simulações) e requer maior quantidade de espaço em memória

do sistema para o cômputo de ambas as cadeias de Markov em paralelo. Alternativamente a esta ideia, surgiu o modelo de JBW, mencionado previamente, que troca as coordenadas do sistema preso em situação metaestável (isto é, poço de energia potencial) com as coordenadas de outro poço, permitindo exploração de vários poços na hiperssuperfície. Entretanto, ainda que resolva o problema do custo computacional, esta implementação tem como desvantagem o fato de requerer o conhecimento prévio (*a priori*) da hiperssuperfície para realizar as trocas e os cálculos da cadeia de Markov (*a posteriori*) (Leach, 2001).

2.5 PRINCÍPIOS FUNDAMENTAIS DO ROSETTA

O software Rosetta é, na verdade, um pacote ou conjunto de aplicações distintas destinadas ao estudo e engenharia computacional de proteínas, desenvolvido pelo já mencionado grupo de pesquisa de David Baker. As aplicações têm parametrização de campo de força própria (a seguir) e as simulações são realizadas em MC, conforme os princípios já explicados do algoritmo de Metropolis. A documentação oficial do programa é bastante extensa e algumas de suas aplicações sequer foram propriamente documentadas até o momento. Assim, nesta secção do trabalho cabe examinar os princípios fundamentais comuns a todas as aplicações do Rosetta, com especial atenção ao campo de força utilizado. Não é objetivo desta etapa realizar uma revisão exaustiva de particularidades de cada aplicação, mas trazer os princípios fundamentais da literatura do grupo de Baker e apresentá-los diretamente para que possam servir para a compreensão dos protocolos adotados neste trabalho e seus resultados. Além disso, é válido comentar que o uso pleno do Rosetta envolve uma linguagem de programação própria baseada em XML, conhecida como RosettaScripts. (Fleishman et al., 2011) O programa pode ser utilizado sem este recurso, mas com consideráveis limitações, especialmente no tangente à reprodutibilidade do trabalho. Por esse motivo, adotamos os protocolos utilizando esta linguagem e estes podem ser consultados nos apêndices. Notamos também que há casos especiais que não foram tratados nesta fundamentação, além de tabelas de parametrização que também fogem ao nosso objetivo. Estes podem ser devidamente consultados nos materiais suplementares dos trabalhos aqui mencionados, mas espera-se que a existência destes casos especiais seja intuitiva para pessoas com formação química pelo próprio exame da Fig. 1, já que há aminoácidos consideravelmente distintos dos demais em sua natureza e estrutura eletrônica, a exemplo da prolina e da tirosina.

2.5.1 Campo de força

A função de energia do Rosetta é uma combinação linear de termos que modelam forças de interações entre átomos, efeitos de solvatação e ângulos torcionais. Desde o lançamento do programa, novas funções de energia são desenvolvidas e lançadas, seja com finalidades específicas (isto é, parametrizadas para uma determinada aplicação específica), seja para melhorar a correlação experimental ou o comportamento da função geral anteriormente utilizada (a ser empregada em diversas aplicações do Rosetta, não apenas em algumas específicas). A função original Score12 (padrão para computações em representação *full atom*) é composta por um termo de Lennard-Jones, um termo de solvatação implícita, um termo de ligação de hidrogênio dependente da orientação, potenciais torcionais de *backbone* e cadeia lateral (derivados do PDB), um termo de interações eletrostáticas de curto alcance e energias de referência de cada um dos 20 aminoácidos que modelam o estado não enovelado. (Leaver-Fay et al., 2013) Formalmente, dada uma configuração C, a energia total E do sistema será dada por:

$$E(C|w,\Theta) = \sum_{j}^{|T|} w_j T_j(C|\Theta_j)$$

Em que cada termo T_j tem parâmetros Θ_j e peso w_j. Uma ferramenta específica chamada de *feature analysis* (do inglês, análise de recursos) é utilizada para o refinamento dos parâmetros Θ_j e um programa específico chamado *optE* é utilizado para ajuste dos pesos w_j. (Alford et al., 2017; Leaver-Fay et al., 2013) Nas próximas secções (2.5.1.1 a 2.5.1.4), examinaremos o tratamento específico de cada termo, utilizado como referência principal o trabalho de Alford e colaboradores [2017] e a documentação original do Rosetta. Assim, trechos não referenciados dizem respeito a este trabalho.

2.5.1.1 Interações de van der Waals

Interações vdW são forças repulsivas e atrativas de curto alcance que variam em função da distância do par atômico. As forças atrativas resultam dos movimentos eletrônicos vizinhos correlatos, e as repulsivas do princípio de exclusão de Pauli. O Rosetta modela estas interações utilizando o potencial LJ 12-6 entre os átomos *i* e *j* em resíduos distintos na seguinte forma:

$$E_{vdW}(i,j) = \varepsilon_{i,j}\left[\left(\frac{\sigma_{i,j}}{d_{i,j}}\right)^{12} - 2\left(\frac{\sigma_{i,j}}{d_{i,j}}\right)^{6}\right]$$

Em que $\sigma_{i,j}$ representa a soma de seus raios atômicos, $d_{i,j}$ representa a distância do par atômico e $\varepsilon_{i,j}$ representa a média geométrica da profundidade dos poços potenciais. Os raios atômicos e as profundidades dos poços foram derivados de dados otimizados (de pequenas moléculas em fase líquida) no contexto do modelo energético adotado. Os termos atrativo (`fa_atr`) e repulsivo (`fa_rep`) são tratados separadamente pelo Rosetta no mínimo de energia do potencial LJ ($\sigma_{i,j} = d_{i,j}$). Esta separação permite mudança de pesos específicos destes componentes sem modificar a distância de mínimo de energia, nem introduzir complexidade matemática pelo tratamento de derivações na descontinuidade. Estas mudanças são úteis em protocolos específicos, em particular naqueles de amostragem conformacional, em que o aumento do peso do componente repulsivo permite atravessar regiões "acidentadas" da hiperssuperfície e prevenir desnaturação proteica durante a amostragem

2.5.1.2 Interações eletrostáticas

Interações eletrostáticas não ligadas são computadas pelo Rosetta a partir de um potencial coulômbico com cargas parciais obtidas do campo de força CHARMM e ajustadas por um esquema de otimização de grupos. Esta função (abaixo) é expressa em termos das distâncias entre os átomos *i* e *j* ($d_{i,j}$), da constante dielétrica ε, de cargas atômicas parciais para cada átomo (q_i e q_j) e da constante de Coulomb ($C_o = 322\ \text{Å}\frac{kcal}{mol}e^{-2}$, com *e* representando a carga elementar):

$$E_{coulomb}(i,j) = \frac{C_o q_i q_j}{\varepsilon}\frac{1}{d_{i,j}}$$

A constante dielétrica, entretanto, é distinta entre o *core* proteico e a superfície exposta ao solvente. Assim, o termo ε constante foi substituído por uma função sigmoide $\varepsilon(d_{i,j})$ que varia de 6 (valor no *core*) a 80 (valor na superfície acessível ao solvente) quando a distância do par atômico está entre 0 e 4 Å:

$$\varepsilon(d_{i,j}) = g\left(\frac{d_{i,j}}{4}\right)\varepsilon_{core} + \left(1 - g\left(\frac{d_{i,j}}{4}\right)\right)\varepsilon_{solvente}$$

$$g(x) = \left(1 + x + \frac{x^2}{2}\right)\exp(-x)$$

Várias aproximações heurísticas foram realizadas pelos autores para adaptar este cálculo para biomoléculas, dentre as quais destaca-se a substituição do gradiente ascendente

com a constante $E_{elec}(d_{min})$ quando $d_{i,j} < 1{,}45$ Å (evitando forças repulsivas muito fortes em curtas distâncias, o que já é computado no termo repulsivo de vdW da secção anterior) e no truncamento do potencial em $d_{max} = 5{,}5$ Å (baseado na assumpção do dielétrico dependente de distância, resultando em redução do custo computacional).

2.5.1.3 Modelo de solvatação

A modelagem explícita de solvatação para todas as interações é um procedimento computacionalmente muito caro para sua realização de maneira sistemática e eficiente em biomoléculas (em que a larga escala costuma ser fator limitante severo). Por isso, o Rosetta representa o solvente como água *bulk* (isto é, água não interfacial) baseando-se no modelo de exclusão gaussiana de Lazaridis-Karplus, adotando dois componentes (`fa_sol` e `lk_ball_wtd`). O primeiro é um componente de energia de solvatação isotrópica, que assume que a água *bulk* está uniformemente distribuída ao redor dos átomos. Já o segundo é um termo de energia de solvatação anisotrópica, que computa águas específicas próximas a átomos polares, compondo a camada de solvatação. A Figura 6 abaixo, adaptada da publicação original, representa os componentes isotrópico e anisotrópico deste modelo.

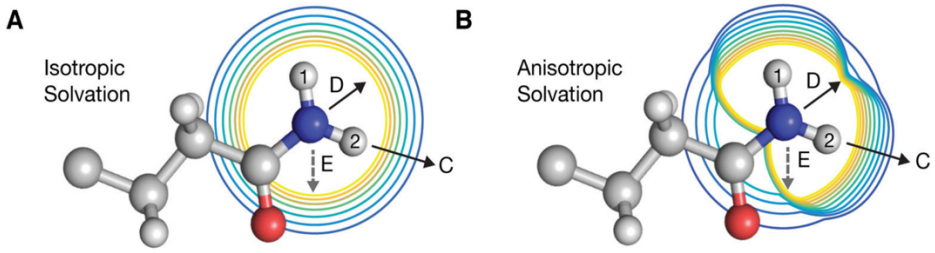


Figura 6. Modelo de solvatação do Rosetta, adaptado do modelo de Lazaridis-Karplus. As figuras demonstram a diferença entre os modelos de solvatação isotrópica e anisotrópica dos grupos NH2 ao redor do CH3 da asparagina. Os contornos variam de regiões de baixa energia (em azul) a regiões de alta energia (em amarelo). As setas representam vetores do par de potenciais. A) Modelo de solvatação isotrópica (do inglês, *isotropic solvation*). B) Modelo de solvatação anisotrópica (do inglês, *anisotropic solvation*). Fonte: Adaptado de (Alford et al., 2017).

O modelo isotrópico de Lazaridis-Karplus é baseado na função $f_{dessolv}$ que descreve a energia requerida para dessolvatar um átomo i quando aproximado por um átomo vizinho j. No Rosetta, como a energia de referência adotada é própria, o termo ΔG_{ref} foi excluído. A energia da interação do par atômico, assim, varia com a distância de separação do par, com as energias livres de transferência de vapor para água ΔG_{livre} (determinadas experimentalmente), com os

raios atômicos somados $\sigma_{i,j}$, com o comprimento de correlação λ e com o volume atômico do átomo dessolvatado V_j:

$$f_{dessolv} = -V_j \frac{\Delta G_{i,livre}}{3\pi\lambda_i\sigma_i^2} \exp\left(-\left(\frac{d - \sigma_{i,j}}{\lambda_i}\right)^2\right)$$

Em distâncias curtas, como mencionado previamente, a superposição é evitada pelo termo `fa_rep`, porém, como esse termo pode ser reduzido ou eliminado em protocolos específicos, foi criado um artifício. Este artifício consiste no incremento gradual da função a uma constante em distâncias próximas quando ocorre superposição das esferas de vdW ($d_{i,j} = \sigma_{i,j}$). Outro artifício interessante foi o truncamento da função em 6,0 Å para acelerar o cálculo, o que é justificável pois em distâncias maiores a função se aproxima assintoticamente de zero. Além disso, a transição entre as constantes de curta e longa distância foi realizada com uma suavização da função acima, obtida pelo uso de polinômios cúbicos dependentes de distância ($f_{poli}^{solv,baixo}$ e $f_{poli}^{solv,alto}$) definidos entre os pontos constantes $c_0 = 0{,}3$ Å e $c_1 = 0{,}2$ Å, tal que:

$$g_{dessolv} = \begin{cases} f_{dessolv}(i,j,\sigma_{i,j}), & d_{i,j} \leq \sigma_{i,j} - c_0 \\ f_{poli}^{solv,baixo}(i,j,d_{i,j}), & \sigma_{i,j} - c_0 < d_{i,j} \leq \sigma_{i,j} + c_1 \\ f_{dessolv}(i,j,d_{i,j}), & \sigma_{i,j} + c_1 < d_{i,j} \leq 4{,}5\ \text{Å} \\ f_{poli}^{solv,alto}(i,j,d_{i,j}), & 4{,}5\ \text{Å} < d_{i,j} \leq 6{,}0\ \text{Å} \\ 0, & 6{,}0\ \text{Å} < d_{i,j} \end{cases}$$

O termo total de energia de solvatação isotrópica (`fa_sol`) é computado como uma soma incluindo o átomo *j* dessolvatando o átomo *i* e vice-versa, utilizando pesos específicos, de forma que:

$$E_{fa\_sol} = \sum_{i,j} w_{i,j}^{conn}(g_{dessolv}(i,j) + g_{dessolv}(j,i))$$

Há ainda um termo de energia de solvatação isotrópica intra-resíduo (`fa_intra_sol`) com a mesma forma funcional do termo `fa_sol` apresentada acima.

Em relação à solvatação anisotrópica, o Rosetta adota um modelo que aumenta o pênalti energético de dessolvatação para átomos polares volumosos (potencialmente oclusivos) próximos de locais nos quais águas poderiam formar ligações de hidrogênio. Para átomos polares, subtrai-se parte da energia isotrópica da equação acima e adiciona-se um termo

energético anisotrópico (`lk_ball_wtd`) para contabilizar a posição do átomo dessolvatado em relação às posições hipotéticas das águas.

Para o cálculo do termo `lk_ball_wtd`, primeiro é estabelecido o conjunto de locais ideais de águas ao redor do átomo *i*: $\mathcal{W}_i = \{v_{i1}, v_{i2}, \ldots\}$. Este conjunto contém 1 a 3 locais (ou sítios) de águas, dependendo do tipo do átomo *i*. Cada sítio dista 2,65 Å do átomo *i* e tem uma geometria ótima para ligações de hidrogênio, considerando a potencial superposição de um átomo dessolvatado *j* em relação a cada água. A sobreposição é consideravelmente desprezível até que o raio da esfera de vdW do átomo dessolvatante *j* (σ_j) toque a esfera de vdW da água no sítio *k* σ_w. O termo é a seguir suavemente incrementado sobre uma zona de superposição parcial de aproximadamente 0,5 Å. Então, para cada sítio de água *k* com coordenadas $v_{j,k}$, computa-se uma medida de oclusão d_k^2 para capturar o espaço (*gap*) entre a água hipotética e o átomo dessolvatante *j* conforme a equação abaixo (em que $\Omega = 3{,}7$ Å):

$$d_k^2 = ||r_j - v_{i,k}||^2 - \left(\sigma_w + \sigma_j\right)^2 + \Omega$$

A seguir, é encontrado o mínimo suavizado da função acima sobre todos os sítios de água W_i computando a média logarítmica:

$$d_{min}^2(i,j) = -\ln\left(\sum_{k \in W_i} \exp\left(-d_k^2\right)\right)$$

Então, os valores são utilizados para computar a função f_{lkfrac}, que por sua vez é utilizada para calcular a energia de solvatação anisotrópica E_{lk_ball} (que segue modelo semelhante ao da energia de solvatação isotrópica, mas com peso w próprio, tipicamente de 0,7). Também é calculado um pênalti energético isotrópico ($E_{lk\_ball\_iso}$) e a função total ($h_{dessolv}$) resulta da soma da energia de solvatação anisotrópica com pênalti isotrópico adicionado. As equações para este desenvolvimento se encontram reproduzidas abaixo:

$$f_{lkfrac}(i,j) = \begin{cases} 1, & d_{min}^2 < 0 \\ \left(1 - \left(\dfrac{d_{min}^2(i,j)}{\Omega}\right)\right)^2, & 0 \le d_{min}^2(i,j) < \Omega \\ 0, & \Omega \le d_{min}^2(i,j) \end{cases}$$

$$E_{lk_{ball}}(i,j) = w_{aniso,i} g_{dessolv}(i,j) f_{lkfrac}(i,j)$$

$$E_{lk_{ball_{iso}}}(i,j) = -w_{iso,i} g_{dessolv}(i,j)$$

$$h_{dessolv}(i,j) = E_{lk\_ball\_iso}(i,j) + E_{lk_{ball}}(i,j)$$

Analogamente ao procedimento adotado para `fa_sol`, a energia de dessolvatação de um átomo *i* por um átomo *j* e, vice versa, do átomo *j* pelo átomo *i* são somadas para resultar na energia total `lk_ball_wtd`, mas apenas levando em consideração a dessolvatação de ligações de hidrogênio polares de átomos pesados (O,N), definidas pelo conjunto *P*, tal que:

$$E_{lk\_ball\_wtd} = \sum_{i \in P} w_{i,j}^{conn}\, h_{dessolv}(i,j) + \sum_{j \in P} w_{i,j}^{conn}\, h_{dessolv}(j,i)$$

2.5.1.4 Ligação de Hidrogênio

As ligações de hidrogênio no Rosetta são modeladas como parcialmente descritas por interações eletrostáticas e parcialmente descritas por ligações covalentes, com o átomo pesado nucleofílico doando densidade eletrônica para o hidrogênio polar, salientando o caráter híbrido covalente-eletrostático deste modelo. (O'Meara et al., 2015) O modelo foi construído teoricamente a partir de dados experimentais de estudos cristalográficos de alta resolução e leva em consideração as preferências de orientação das ligações de hidrogênio encontradas nestes estudos cristalográficos. A energia resultante das ligações de hidrogênio é calculada para todos os pares de hidrogênios doadores (H) e aceptores (A) como uma função de quatro graus de liberdade:

(1) a distância *d* entre H e A;

(2) o ângulo θ formado entre H, A e o átomo pesado doador D;

(3) o ângulo θ formado entre H, A e a "base" B do átomo aceptor parental;

(4) o ângulo torcional ϕ entre A, H, e dois átomos parentais subsequentes (B e B2).

Para facilitar a visualização destes graus de liberdade, reproduzimos abaixo a figura do trabalho original de referência para estas secções:

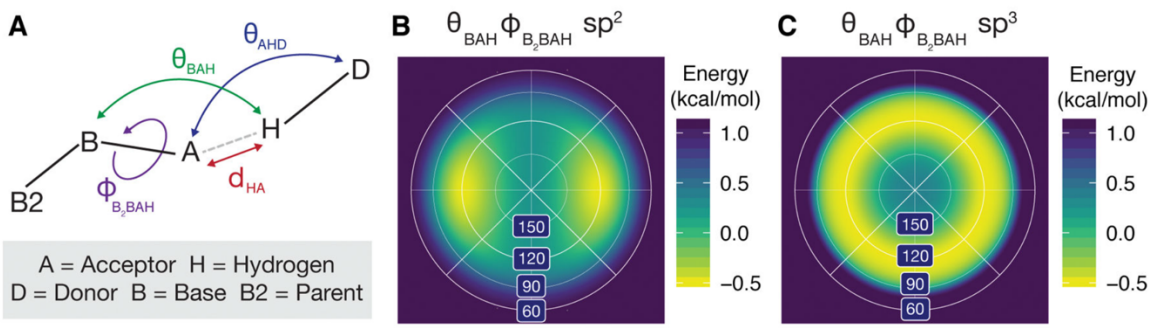


Figura 7. Modelo de ligação de hidrogênio dependente da orientação. A) Graus de liberdade avaliados pelo termo da ligação de hidrogênio; B) Representação das projeções Labert-azimutais da hiperssuperfície da energia de ligação de hidrogênio para um aceptor com hibridação sp2 e C) sp3. Fonte: (Alford et al., 2017)

Para evitar contagem redundante, as ligações de hidrogênio entre *backbone* e cadeia lateral são excluídas se o grupo do *backbone* já está envolvido em uma ligação de hidrogênio. A contribuição energética referente ao quarto grau de liberdade apresentado acima, isto é, do ângulo torcional ϕ supracitado, depende da hibridação orbital do aceptor, ρ. A função total é então suavizada com *f(x)* para evitar derivadas em descontinuidades e garantir que casos de ligações de hidrogênio próximos das bordas da caixa sejam considerados:

$$E_{hbond} = \sum_{H,A} w_H w_A f\left(E_{hbond}^{HA}(d_{HA}) + E_{hbond}^{AHD}(\theta_{AHD}) + E_{hbond}^{BAH}(\theta_{BAH}) + E_{hbond}^{B2BAH}(\rho, \phi_{B2BAH}, \theta_{BAH})\right)$$

$$f(x) = \begin{cases} x, & x < -0{,}1 \\ -0{,}025 + {}^{x}/_{2} + 2{,}5x^2, & -0{,}1 \le x < 0{,}1 \\ 0, & 0{,}1 \le x \end{cases}$$

2.5.1.5 Ligações Dissulfeto

Tipicamente, o Rosetta se baseia em um sistema arbóreo para manter as distâncias e ângulos de ligação fixos, de modo que a amostragem do espaço conformacional mude apenas as torções, dispensando termos que avaliam a energia das distâncias e ângulos de ligação para tanto. Porém, com ligações dissulfeto e com prolinas (que se destacam por serem o único resíduo iminoácido, não aminoácido), a representação arbórea falha, já que esta representação só é adequada para cadeias acíclicas, requerendo o tratamento explícito destes parâmetros. A parametrização foi realizada com base em um modelo de energia orientação-dependente chamado dslf_fa13 a partir de dados cristalográficos e de estimativas de densidade de kernel (KDE). Pela introdução dos componentes de distância e ângulo de ligação supracitados, o modelo utilizado para cálculo das energias das ligações dissulfeto requer mais graus de liberdade que o modelo adotado para as demais ligações. A energia total das ligações dissulfeto é computada como uma função de seis graus de liberdade (Fig. 8), mapeados para quatro componentes, a saber:

(1) A geometria da distância *d* entre enxofres (*S-S*);

(2) O ângulo θ formado entre os carbonos beta com relação à ligação S-S (*CSS*);

(3) O ângulo diedral ϕ formado entre os carbonos alfa e beta com relação à ligação S-S (*CaCbSS*);

(4) O ângulo diedral ϕ formado entre os carbonos betas consecutivos e à ligação S-S (*Cb1Cb2SS*).

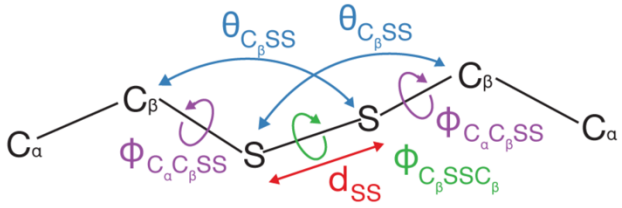


Figura 8. Representação dos graus de liberdade utilizados para a modelagem das ligações dissulfeto.

A equação que representa a energia total (obtida pela soma dos seus componentes acima explicados) está esquematizada a seguir:

$$E_{dslf\_fa13} = \sum_{S1,S2} E^{SS}_{dslf}(d_{SS}) + E^{CSS}_{dslf}(\theta_{Cb1SS}) + E^{CSS}_{dslf}(\theta_{Cb2SS}) + E^{CaCbSS}_{dslf}(\phi_{Ca1Cb1SS}) + E^{CaCbSS}_{dslf}(\phi_{Ca2Cb2SS}) + E^{CbSSCb}_{dslf}(\phi_{CbSSCb})$$

2.5.2 Amostragem conformacional das torções

A amostragem computacional do espaço conformacional classicamente pode adotar várias metodologias, incluindo busca aleatória, geometria de distâncias, construções de grades, modelos de simulação, entre outros. (Leach, 2001) A avaliação das conformações de *backbone* e cadeia lateral no espaço torcional é otimizada no Rosetta para melhorar a eficiência computacional da busca. Campos de força tradicionais descrevem as energias torcionais em termos trigonométricos que por vezes podem ter desempenho pobre ao reproduzir as distribuições diedrais em regiões desestruturadas. Visando a otimização do tempo de cálculo, o Rosetta utiliza vários termos baseados em conhecimentos prévios para os ângulos torcionais que servem como aproximações de efeitos quânticos e modelam de maneira acurada as conformações preferidas das proteínas.

2.5.2.1 Amostragem conformacional do *backbone*

Para avaliação dos ângulos diedrais do *backbone* (ϕ e ψ), o Rosetta utiliza um termo energético chamado `rama_prepro`, baseado nos mapas de Ramachandran (Fig. 2) para cada aminoácido. Os resíduos foram utilizados em KDE (estimativas de densidade de Kernel) dos mapas de Ramachandran com tamanho de passo do grid de 10° para cada ângulo diedral.

Resíduos precedendo prolina são tratados separadamente, já que esta impõe interações e restrições estéricas com relação ao seu carbono gama. A energia é calculada convertendo as probabilidades em energias nos pontos do grid por inversão de Botzmann, seguido de etapa de interpolação bicúbica das energias (não das probabilidades). A inversão de Boltzmann já é um método utilizado para resgatar informação estrutural de cadeias laterais de campos de força *coarse-grain*. (Leach, 2001) A energia calculada em função das probabilidades P pode, portanto, ser assim representada:

$$E_{rama\_pre\_pro} = \sum_i \begin{cases} -\ln\left[P_{reg}(\phi_i, \psi_i | aa_i)\right], \text{se } i+1 \text{ não é prolina} \\ -\ln\left[P_{prepro}(\phi_i, \psi_i | aa_i)\right], \text{se } i+1 \text{ é prolina} \end{cases}$$

Além disso, o Rosetta também tem um termo específico para o design do *backbone*, computado como a probabilidade de adicionar uma cadeia lateral específica dado uma conformação de *backbone* (a partir de seus ângulos diedrais), chamado de `p_aa_p`. Este termo representa a tendência de encontrar este aminoácido em relação aos 19 outros aminoácidos canônicos (Fig. 1) dados determinados valores de ângulos diedrais. O resultado é expresso em termos de regras de probabilidade bayesianas, conforme a equação a seguir, e a contribuição energética deste termo é calculada de maneira análoga à adotada para o cálculo de `rama_prepro`:

$$P(aa|\phi,\psi) = \frac{P(\phi,\psi|aa)P(aa)}{\sum_{aa'} P(\phi,\psi|aa')}$$

$$E_{p\_aa\_pp} = \sum_r -\ln\left[\frac{P(aa_r|\phi_r,\psi_r)}{P(aa_r)}\right]$$

2.5.2.2 Amostragem conformacional das cadeias laterais

As cadeias laterais proteicas ocupam, em grande parte, conformações discretas (rotâmeros), separadas por barreiras energéticas largas. As coordenadas dos rotâmeros oriundas de estudos experimentais são armazenadas em bibliotecas de rotâmeros. Cada conjunto topológico de uma biblioteca pode ser armazenado com sua respectiva probabilidade de ocorrência (obtida por análise bayesiana simples), bem como outras métricas estatísticas de interesse para a construção da biblioteca. O Rosetta utiliza as probabilidades de uma biblioteca rotamérica construída em 2010, contendo frequências, médias e desvios padrão de valores individuais de ângulos χ para cada possível ângulo (k) para cada tipo de aminoácido. A probabilidade tem três componentes:

(1) probabilidade de ocorrência de um rotâmero específico, dados os ângulos diedrais do *backbone*;
(2) probabilidade de ocorrência de ângulos χ específicos, dado um rotâmero específico;
(3) distribuição de ângulo χ terminal, que pode ser semelhante à gaussiana, ou contínua quando o ângulo χ terminal tem hibridação sp2.

A probabilidade é calculada conforme a equação a seguir, na qual T representa o número de ângulos χ rotaméricos + 1.

$$P(\chi|\phi,\psi,aa) = P(rot|\phi,\psi,aa)\left(\prod_{k<T} P(\chi_k|\phi,\psi,rot,aa)\right)P(\chi_T|\phi,\psi,rot,aa)$$

A biblioteca adotada pelo Rosetta distingue torções rotaméricas de não rotaméricas. Torções rotaméricas ocorrem quando o terceiro dos quatro átomos que definem a torção tem hibridação sp3 (preferindo valores em torno de 60, 180 e -60° com barreiras energéticas elevadas). Se o último ângulo torcional χ é rotamérico, sua probabilidade é fixa em 1. Já torções não rotaméricas são definidas pela hibridação sp2 no terceiro átomo. Nesse caso, a distribuição de probabilidades é contínua. Há oito aminoácidos categorizados como semi-rotaméricos (por conter diedros rotaméricos e não rotaméricos na cadeia lateral): Asp, Asn, Gln, Glu, His, Phe, Tyr e Trp.

A probabilidade de cada rotâmero ocorrer é derivada dos mapas de Ramachandran, conforme previamente explicado. Os valores de U e O são computados em função dos ângulos diedrais do *backbone*. A probabilidade dos valores de χ é calculada com uma forma reminiscente a gaussiana, mas sem o coeficiente de normalização. Já o logaritmo natural desta probabilidade tem forma semelhante à da lei de Hooke, utilizada para modelagem do oscilador harmônico clássico. As equações de cálculo da probabilidade dos valores de χ e a forma funcional completa de `fa_dun` são reproduzidas a seguir:

$$P(\chi_k|\phi_k,\psi_k,rot) = \exp\left(-\frac{1}{2}\left(\frac{\chi_k - \mu_{\chi k}(\phi,\psi|rot,aa)}{\sigma_{\chi k}(\phi,\psi|rot,aa)}\right)^2\right)$$

$$E_{fa\_dun} = \sum_{r} -\ln\left(P(rot_r|\phi_r,\psi_r,aa_r)\right) + \sum_{k<Tr}\frac{1}{2}\left(\frac{\chi_{k,r}-\mu_{\chi k}(\phi_r,\psi_r|rot_r,aa_r)}{\sigma_{\chi k}(\phi_r,\psi_r|rot_r,aa_r)}\right)^2 + \left(-\ln\left(P\left(\chi_{Tr,r}|\phi_r,\psi_r,rot_r,aa_r\right)\right)\right)$$

2.5.4 Interpretação das Unidades das Funções de Energia do Rosetta

As unidades de energia do Rosetta, inicialmente, eram expressas em uma unidade genérica de energia, denominada *Rosetta Energy Unit* (REU, do inglês, unidades de energia do Rosetta). Esta escolha foi feita porque alguns dos termos do Rosetta não eram calibrados com dados experimentais, mas gerou grande dificuldade de interpretação desta unidade e grande discussão na comunidade acadêmica acerca de seu uso. No trabalho utilizado como referência para esta secção, é apresentada a função de energia REF15, que foi parametrizada em estruturas de referência de alta resolução e com parâmetros mensurados em kcal/mol. Além disso, os autores relatam ter encontrado forte correlação experimental com os valores preditos pelo Rosetta (ΔΔG de mutação com coeficiente de Pearson R=0,994). Por esta razão, os autores propõem como prática padronizada a expressão das energias em kcal/mol (similarmente a realidade de campos de força como OPLS, CHARMM e AMBER). (Alford et al., 2017) Apesar disso, a experiência do nosso grupo de pesquisa não corrobora esta proposição, o que ficará mais evidente na discussão dos resultados aqui apresentados. Por este motivo, até o momento seguimos adotando neste trabalho e nos demais trabalhos do grupo as unidades REU para os cálculos realizados com o Rosetta, mesmo nos protocolos com a função REF15.

2.6 MÉTODOS COMPUTACIONAIS DE CÁLCULO DE ENERGIA LIVRE DE GIBBS

O cálculo da energia livre de Gibbs absoluta utilizando simulações de MC ou MD é extremamente desafiador, pois o termo do hamiltoniano na energia livre (inclusive na de Helmholtz) têm contribuições importantes para sua integral. Esta limitação decorre do fato de ambas as metodologias utilizarem amostragem preferencial nas regiões de mínimos de energia da paisagem energética, nunca amostrando adequadamente as regiões de alta energia que são importantes e necessárias para o cálculo da energia livre. Métodos de inserção de partículas e ensemble grã-canônico podem apresentar estratégias para o cálculo da energia livre, mas não são aplicáveis para muitos dos sistemas de interesse, que contêm moléculas complexas em altas densidades. (Leach, 2001) Apesar destas limitações, metodologias foram desenvolvidas para

cálculo das diferenças de energia livre em diferentes cenários. Começaremos esta secção examinando os métodos termodinâmicos alquímicos mais comuns (isto é, integração e perturbação termodinâmica), seguidos de métodos de particionamento de energia e dos modernos métodos baseados em aprendizagem de máquina.

2.6.1 Métodos alquímicos

Os métodos alquímicos mais comuns são a perturbação termodinâmica (ou método de Zwanzig) e a integração termodinâmica aplicados aos ciclos termodinâmicos. Tais métodos surgem da percepção de que não precisamos, em simulações computacionais, ficar restritos a seguir caminhos de integrais com sentido termodinâmico e físicos reais. Assim, enquanto experimentalmente é necessário recorrer a transições de fase (geralmente para estado gasoso ideal) para o cálculo da energia livre, computacionalmente as transições construídas podem ser alquímicas (ou até mesmo do próprio hamiltoniano, como no método *slow-growth*, que não será abordado), desde que as condições de histerese sejam razoavelmente satisfeitas, já que a variação de energia livre é função de estado. (Allen; Tildesley, 1989)

O método de perturbação termodinâmica (ou método de Zwanzig) considera dois estados X e Y contendo N partículas e estados transitórios entre estes que funcionam como caminho de integração. Os estados inicial e final serão diferentes a depender do fenômeno investigado pelo químico teórico (X e Y podem representar, por exemplo, a mutação de um aminoácido específico em uma proteína, ou representar a mesma proteína em solução aquosa *versus* em etanol, ou representar a mudança de um único átomo de uma estrutura). O número de estados intermediários é variável, e geralmente é descrito em relação aos estados inicial e final através de um parâmetro de acoplamento λ. Conforme este parâmetro varia de 0 a 1, o hamiltoniano do sistema varia de $\mathcal{H}_X$ a $\mathcal{H}_Y$. Em teoria, se os estados X e Y não possuíssem diferença energética muito superior aos valores de 1/β, não haveria necessidade de utilizar estados intermediários. Porém, como tipicamente a diferença de energia entre os estados inicial e final é muito superior a 1/β, faz-se necessário o uso de estados intermediários como caminhos de integral para o cálculo da diferença de energia livre. O número de estados deve ser suficiente para garantir uma sobreposição adequada de hamiltonianos entre os estados (respeitando a diferença de 1/β, idealmente). É possível escolher tantos estados intermediários quanto se deseje para garantir esta sobreposição, pois os termos intermediários se cancelam. (Leach, 2001) Entretanto, como cada estado representará uma simulação, o custo computacional deve

ser atentado e o mínimo de estados intermediários para garantir sobreposição adequada deve ser escolhido, idealmente.

Cada termo do campo de força para um estado intermediário pode ser escrito em termos das combinações lineares dos valores de X e Y. Assim, a descrição do campo de força para cada um de seus componentes terá os valores destas combinações lineares nos lugares dos parâmetros. A forma da equação irá depender do campo de força adotado, mas de modo geral cada parâmetro (por exemplo, distância de ligação e constante de mola para uma modelagem de oscilador harmônico clássico para a energia de ligação) será escrito como combinação linear representando o estado de transição como intermediário entre X a Y através do parâmetro de acoplamento. Para cada valor de λ uma simulação deve ser realizada (utilizando MC ou MD conforme apropriado) com os termos adequados do campo de força escolhido. Após a equilibração, tem início uma fase de produção durante a qual a diferença de energia livre entre estados consecutivos $\Delta G(\lambda_i \rightarrow \lambda_{i+1})$ é acumulada como $-k_B T \ln \langle \exp(-\beta \Delta \mathcal{H}_i) \rangle$, em que $\Delta \mathcal{H}_i = \mathcal{H}_{i+1} - \mathcal{H}_i$. A energia livre total para λ = 0 e λ = 1 é a soma das diferenças nas energias livres para os vários estados intermediários. (Leach, 2001) O procedimento acima é conhecido como *forward sampling* (do inglês, amostragem direta) pois calcula a energia livre no sentido $\lambda_i \rightarrow \lambda_{i+1}$. Porém, é possível executar o procedimento inverso, conhecido como *backward sampling* (do inglês, amostragem inversa) no sentido $\lambda_i \rightarrow \lambda_{i-1}$, chegando aos mesmos resultados, já que a variação de energia livre é função de estado.

A integração termodinâmica tem princípio semelhante ao da perturbação termodinâmica, em relação ao procedimento prático das simulações (isto é, também aqui há estados inicial X e final Y, e o parâmetro de acoplamento λ). Entretanto, a diferença de energia livre entre os estados final e inicial é calculada como uma integral contínua do caminho entre os estados final e inicial em relação ao parâmetro de acoplamento. Na prática, isto é feito tomando uma série de simulações correspondentes a valores discretos de λ variando de 0 a 1. Para cada valor de λ, a média de $\partial \mathcal{H} / \partial \lambda$ é determinada. A cada etapa, a derivada parcial do hamiltoniano pode ser calculada analiticamente utilizando programas específicos ou, mais comumente, aceitando a aproximação $\partial \mathcal{H} / \partial \lambda \approx \Delta \mathcal{H} / \Delta \lambda$ como válida. A diferença de energia livre total será igual à área sob a curva da função das médias de $\partial \mathcal{H} / \partial \lambda$.

Em ambas as metodologias, pode ser empregado um ciclo termodinâmico para cálculo das diferenças de energia livre entre estados. Por exemplo, supondo que se deseja calcular a

diferença de energia livre de ligação de um ligante *L1* ao receptor *R*. Na impossibilidade de realizar diretamente a simulação da ligação (pois este tipo de simulação envolve tantas mudanças na geometria do receptor, ligante e solvente que o custo computacional costuma ser impeditivo para a adequada exploração do espaço de fases), pode-se conduzir as simulações do ligante na presença do receptor, mas não complexado a este, e deste mesmo ligante já complexado ao receptor separadamente. Nestas simulações, porém, é necessário utilizar um artifício teórico alquímico. Utiliza-se um ligante teórico *L2* (que pode ser o mesmo receptor *L1* com mudança de apenas um átomo, por exemplo) e a simulação passa a ser conduzida em termos da mudança alquímica introduzida. Uma representação visual deste ciclo termodinâmico pode ser apreciada na Fig. 9 abaixo, retirada do artigo que propôs a implementação computacional desta metodologia, por vanGustereng e Berendsen [1987]:

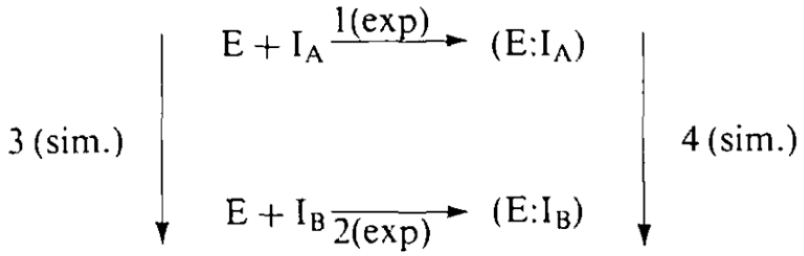


Figura 9. Ciclo termodinâmico realizado por vanGustereng e Berendsen. Neste exemplo, os autores realizaram um ensaio de competição inibitória enzimática (ligantes inibidores Ia e Ib e enzima E). O símbolo de dois pontos denota a formação do complexo. As etapas 1 e 2 foram realizadas experimentalmente (exp). As etapas 3 e 4 foram realizadas computacionalmente (sim.), validando o método proposto acima. Fonte: Adaptado de (Van Gunsteren; Berendsen, 1987)

O exemplo acima ilustra a aplicação de um ciclo termodinâmico em ensaio de competição inibitória enzimática. As etapas 1 e 2 são praticamente impossíveis de se realizar computacionalmente, considerando a dificuldade de explorar adequadamente o espaço de fases para a mudanças conformacionais do inibidor, da enzima e do solvente. Assim, o recurso de utilizar um outro inibidor para conduzir as simulações dos processos 3 e 4 é um artifício sagaz, permitido pelo fato de a diferença de energia livre ser função de estado, o que garante que a variação total no ciclo seja nula. Por este motivo, $\Delta G2 - \Delta G1 = \Delta G4 - \Delta G3$. Portanto, os autores realizaram as simulações computacionais dos processos não químicos 3 e 4, o que pode ser realizado mais facilmente se os inibidores não forem muito diferentes entre si. (Van Gunsteren; Berendsen, 1987)

Esta metodologia permite o cálculo das diferenças de energia livre relativas, mais adequadamente representadas como $\Delta\Delta G = \Delta G2 - \Delta G1 = \Delta G4 - \Delta G3$. Para o cálculo preciso dos valores de energia livre, há muitas metodologias disponíveis. As duas primeiras apresentadas, aplicadas a um processo alquímico em um ciclo termodinâmico são um exemplo disto. O problema com estas metodologias (além da escolha adequada do campo de força, do tipo de simulação, do hamiltoniano do sistema, do parâmetro de acoplamento, entre outros) é seu custo computacional elevado, já que simulações são realizadas para cada estado intermediário, além das simulações dos estados inicial e final. Tomando como exemplo o presente estudo de mestrado, em que mais de 800 mutações são realizadas para cada proteína em estudo, seria necessário (utilizando apenas 8 estados intermediários, totalizando 10 simulações por mutação) realizar mais de 8000 simulações por proteína. Por esse motivo, cabe discutir outros métodos de cálculo de energia livre, que permitam que os resultados sejam calculados com a celeridade requerida para o grande número de dados referentes ao problema em questão. Apesar disso, as discussões desta secção são de extrema importância, tanto do ponto de vista histórico, quanto do ponto de vista metodológico, já que as abordagens que serão apresentadas se baseiam largamente nos ciclos termodinâmicos aqui apresentados.

2.6.2 Métodos de Particionamento de Energia

Dos conhecimentos básicos de mecânica estatística, é sabido que uma função de partição molecular pode ser decomposta em funções de partição para cada grau de liberdade. Também as energias de uma molécula podem ser decompostas em seus graus de liberdades componentes (por exemplo, somando-se as contribuições translacional, rotacional, vibracional e eletrônica para calcular a energia total de uma molécula num determinado sistema). (McQuarrie; Simon, 1999) Em sistemas mais complexos, procedimento análogo pode ser tomado para calcular contribuições energéticas distintas e somá-las para determinar o hamiltoniano do sistema, conforme explicado na secção de mecânica molecular do presente trabalho. Este procedimento é particularmente útil para proteínas, em que o hamiltoniano completo *ab initio* nunca foi completamente elucidado (e calculá-lo com métodos QM implicaria em custo computacional impeditivo). A dificuldade deste procedimento consiste, principalmente, na escolha adequada do modelo para cálculo de cada componente. O problema da escolha adequada do hamiltoniano para o sistema deve levar em consideração um equilíbrio delicado entre a necessidade de rigor matemático na descrição do sistema e seus componentes *versus* o custo computacional para implementar este modelo. O autor Kerson Huang em suas aulas sobre a mecânica estatística do enovelamento proteico (reunidas no livro "*Huang's*

Lectures on Statistical Physics And Protein Folding" [2005]) elenca alguns dos fatores que demandam atenção na escolha do hamiltoniano adequado para a descrição de sistemas de proteínas, a saber:

(1) As propriedades atribuídas a uma única molécula de proteína, incluindo sua energia livre, dependem fortemente do seu ambiente;
(2) O meio em que a molécula está inserida que induz o enovelamento, mantendo a estrutura nativa em um estado de equilíbrio dinâmico;
(3) Além das interações entre os resíduos, a energia livre deve refletir as interações hidrofóbicas com o meio;
(4) É esperado que haja vários mínimos locais, sendo possível que a estrutura transite rapidamente entre mínimos vizinhos, desde que as barreiras energéticas não sejam altas o suficiente para impedir este processo;
(5) A energia livre não deve ser encarada como um valor absoluto, mas como uma restrição de mínimo (ou de grupos de mínimos) determinados pela cinética do enovelamento. (Huang, 2005)

Muitos dos conceitos acima foram abordados em relação ao funil de enovelamento e termodinâmica deste processo na secção 2.1, mas optou-se por sintetizar os princípios acima, elegantemente descritos pelo Prof. Huang na obra supracitada, para prosseguir com a discussão desta secção. O fundamento matemático que permite que a energia livre total seja particionada pode ser compreendido analisando a fórmula de cálculo de energia livre pelo método de integração termodinâmica abaixo:

$$\Delta A = \int_{\lambda=0}^{\lambda=1} \langle \frac{\partial \mathcal{H}}{\partial \lambda} \rangle \, d\lambda$$

O Hamiltoniano do sistema pode ser descrito como a soma das contribuições dos hamiltonianos das energias de comprimento e ângulo de ligação, entre outros componentes discutidos na secção de mecânica molecular (representados como (...) na equação abaixo):

$$\langle \frac{\partial \mathcal{H}}{\partial \lambda} \rangle = \langle \frac{\partial \mathcal{H}_{ligações}}{\partial \lambda} + \frac{\partial \mathcal{H}_{ângulos}}{\partial \lambda} + \cdots \rangle$$

$$\Delta A = \int_{\lambda=0}^{\lambda=1} \langle \frac{\partial \mathcal{H}_{ligações}}{\partial \lambda} \rangle \, d\lambda + \int_{\lambda=0}^{\lambda=1} \langle \frac{\partial \mathcal{H}_{ângulos}}{\partial \lambda} \rangle \, d\lambda + \cdots$$

Uma prática comum neste processo é adotar, além dos parâmetros de interação acima, as contribuições de diedros, interações vdW e eletrostáticas. (Leach, 2001) A parametrização específica utilizada no Rosetta foi discutida na sua respectiva secção e pode ser consultada. A partir de um hamiltoniano apropriado, é possível calcular a energia livre do sistema como uma soma das suas contribuições individuais. Um problema com este método é o fato de estas contribuições individuais não serem funções de estado por si só (apenas sua soma refletiria a função de estado da energia livre). Esse fato limita o uso destas técnicas pela descrição e implementação dos modelos adequados no sistema, um desafio comum ao cálculo de energia livre por diversos métodos (isto é, erros decorrentes da falta de acurácia no hamiltoniano escolhido). Apesar disso, o particionamento em componentes individuais (embora não tenham significado físico inteiramente conhecido) permite avaliar a importância de cada contribuição no cálculo do hamiltoniano para um processo. Além disso, embora a discussão seja válida do ponto de vista teórico, caso os autores comprovem que há histerese no hamiltoniano total do sistema (prática relativamente comum em trabalhos de parametrização de validação de funções de energia, especialmente em MD), a condição necessária para satisfazê-lo como função de estado é razoavelmente satisfeita, mesmo com o uso de aproximações heurísticas.

Como discutido anteriormente, na parametrização do Rosetta, cada componente da função de energia total tem um peso próprio, calculado geralmente a partir de dados experimentais, para garantir que haja maior correlação entre a função de energia e os dados experimentais. A energia do Rosetta é então calculada com a diferença dos hamiltonianos inicial e final em um ciclo termodinâmico, resultando em valores de $\Delta\Delta G$ expressos em REU. (Alford et al., 2017) Porém, embora os autores tenham encontrado alto grau de correlação experimental com este modelo (cf. Secção 2.5.4), não conseguimos reproduzir os resultados no nosso grupo, levando ao desenvolvimento de um método baseado em aprendizagem de máquina destinado a este fim. Uma possível razão para esta diferença consiste nas condições experimentais específicas, que são distintas em experimentos conduzidos em laboratórios diferentes. As consequências destas distinções, porém, deveriam ser mínimas ou ter pouco impacto no resultado final, caso o modelo proposto seja suficientemente robusto. Outro problema é que a descrição do hamiltoniano utilizado tem cortes e atalhos necessários para acelerar o cálculo (caso contrário, o programa seria aplicável apenas a moléculas pequenas, não a proteínas), motivo pelo qual os próprios autores utilizam o termo *score de energia* no lugar do termo *hamiltoniano* do sistema. Um outro problema, também, é que os métodos de particionamento de energia, em teoria, só poderiam ser diretamente aplicados ao cálculo de energia livre por

integração termodinâmica. Utilizá-lo diretamente envolve considerar que as diferenciais parciais (contínuas) dos componentes do sistema sejam aproximáveis aos valores discretos das variações entre estados finais e iniciais, o que nem sempre pode ser garantido. Por fim, um outro problema (e este é comum a todos os métodos de cálculo de energia livre), é que a amostragem do espaço de fases deve ser suficiente para o cálculo, e esta amostragem nunca será suficientemente exaustiva utilizando algoritmos direcionados para determinadas regiões da paisagem energética do sistema (no caso de modelos MC com implementação Metropolis ou de modelos MD, a amostragem é direcionada para os mínimos de energia). É interessante notar que os problemas discutidos acima, embora aplicados a softwares modernos desenvolvidos recentemente, já foram elegantemente elencados e apontados por van Gustereng e colaboradores em seu clássico trabalho intitulado "*Computation of free energy*" [2002]. As conclusões deste trabalho são reproduzidas a seguir por sua grande validade e atemporalidade, e como contraponto aos novos paradigmas de cálculos de energia livre utilizando métodos baseados em aprendizagem de máquina (que serão discutidos a seguir, cf. secção 2.6.3): Para a estimativa confiável das diferenças de energia livre entre dois estados de um sistema, três condições cruciais devem ser satisfeitas:

(1) A simulação deve ser realizada com campo de força suficientemente acurado;
(2) O sistema deve estar adequadamente equilibrado em todos os instantes;
(3) Um conjunto de configurações representativa do equilíbrio completo do *ensemble* deve ser amostrado a cada ponto.

Adicionalmente, os autores relatam que a seleção de fórmulas adequadas para cada aplicação em conjunto com a incorporação de funções judiciosas no hamiltoniano para acelerar a equilibração e a amostragem podem aumentar largamente a acurácia e eficiência do cálculo de energia livre. (Van Gunsteren; Daura; Mark, 2002)

Uma outra estratégia metodológica interessante para a predição das energias livres de interação entre complexos proteicos está representada pelo pacote computacional FoldX. Este pacote implementa uma metodologia também baseada nos ciclos termodinâmicos supracitados para o cálculo. O campo de força do FoldX foi descrito em 2003 e está parametrizado conforme a equação abaixo:

$$\Delta G = a\,\Delta G_{vdW} + b\,\Delta G_{solvH} + c\,\Delta G_{solvP} + d\,\Delta G_{wb} + e\,\Delta G_{hbond} + f\,\Delta G_{el} + g\,\Delta G_{kon} + h\,T\Delta S_{mc} + k\,T\Delta S_{sc} + l\,\Delta G_{clash}$$

Nesta expressão, os termos (*a, b, .., l*) são os pesos relativos de cada termo energético diferente utilizado para cálculo de energia livre. As interações com os solventes são computadas particionando-se em contribuições de grupos hidrofóbicos (ΔG_{solvH}) e polares (ΔG_{solvP}). Os parâmetros de solvatação são derivados de experimentos de transição de solvente (de água para solvente orgânico) em aminoácidos. Moléculas de água com interações persistentes com grupos da proteína (isto é, aquelas que realizam mais de duas ligações de hidrogênio com a proteína) são calculadas explicitamente no termo ΔG_{wb}. O termo de interações vdW é computado semelhantemente ao processo de dessolvatação, mas leva em conta energias de transferência de fase (líquida para vapor) experimentais. ΔG_{hbond} é inferido a partir de dois ciclos de mutação (efetuados por engenharia de proteínas). A contribuição eletrostática (ΔG_{el}) é modelada coulombicamente. Para complexos proteicos, um termo eletrostático adicional (ΔG_{kon}) é calculado, baseado na equação empírica descrita por Selzer e colaboradores [2000]. (Selzer; Albeck; Schreiber, 2000) Os termos entrópicos representam uma diferença considerável entre o FoldX e outros campos de força. Neste programa a penalidade entrópica para fixar o *backbone* em uma configuração específica (ΔS_{mc}) é derivada de uma análise estatística das distribuições dos ângulos diedrais φ e ψ de um dado aminoácido observado em estruturas cristalográficas não redundantes de alta resolução. O custo entrópico é obtido pelo escalonamento de uma série de parâmetros entrópicos calculados por Abagaynan e colaboradores [1994] para o enterramento da cadeia lateral. (Abagyan; Totrov, 1994) O último termo (ΔG_{clash}) advem da medida da superposição estérica entre átomos na estrutura. Para os cálculos de mutações pontuais no FoldX, é recomendado incorporar este termo energético repulsivo utilizando uma penalização suavizada para as superposições vdW. Para o design proteico, porém, os autores recomendam o uso de penalizações completas (não suavizadas). (Schymkowitz et al., 2005) O cálculo do ΔΔG neste programa, conforme exposto, é obtido por particionamento da função de energia do campo de força e implementado como um ciclo termodinâmico que compara as diferenças de energia livres nativa e da mutação:

$$\Delta\Delta G = \Delta G_{WT} - \Delta G_{mut}$$

Como limitação do modelo, a energia total obtida por esta metodologia não é capaz de prever resultados experimentais, mas os autores da publicação mais recente do programa sugerem adotar os resultados em kcal/mol, já que os termos do particionamento são derivados, direta ou indiretamente, conforme exposto acima, de resultados experimentais. Outras limitações conhecidas consistem na necessidade de utilizar estruturas com resolução em escala

quase-atômica para obter resultados confiáveis e na menor correlação experimental de resultados de complexos proteicos, já que o programa não computa termos específicos para a PPI explicitamente. Apesar destas limitações, a correlação experimental para predizer a ocorrência de mutações desestabilizantes em uma estrutura é melhor que aquela para predizer mutações estabilizantes. (Buß; Rudat; Ochsenreither, 2018) Isto porque o programa permite realizar mutações pontuais diretamente, o que não foi necessário no presente estudo, já que as mutações já haviam sido modeladas no protocolo Flex ddG do Rosetta (cf. secção 3.3) e os PDBs resultantes do Flex ddG foram utilizados como *input* para o FoldX.

2.6.3 Métodos Baseados em Aprendizagem de Máquina

A química computacional passou por avanços muito expressivos na década de 1990 (e continua passando desde então com revoluções cognitivas, sociais e tecnológicas que permitem cálculos progressivamente mais rápidos e de maior complexidade), fundamentada principalmente em regras teóricas e empíricas. As aplicações baseadas em ML, por sua vez, são baseadas em largos conjuntos de dados carregando a informação essencial para estimar ou calcular determinada propriedade. Estas aplicações podem acelerar bastante o desenvolvimento de trabalhos originais, seja por reduzir o custo computacional (de uma simulação, por exemplo) ou experimental, ou, ainda, por prover uma nova rota para resolução de problemas complexos de maneira racional. (Shi et al., 2023) Esta mudança de paradigma ocorreu não apenas nos softwares utilizados, mas também no próprio pensamento químico, que tende progressivamente mais à análise de dados já existentes e ao trato específico desses largos conjuntos de dados, em oposição à definição e à aplicação das regras teóricas e empíricas. A mudança de paradigmas provavelmente constitui uma revolução científica (embora só poderemos sabê-lo com certeza após alguns anos em análise retrospectiva). Na prática, a implementação computacional do ML foi facilitada pela existência de pacotes abertos (no inglês, *open source packages*), como scikit-learn, PyTorch e TensorFlow. Porém, a dificuldade de escolher uma metodologia de ML apropriada para determinada aplicação persiste. Não é objetivo deste trabalho discutir cada metodologia de ML, mas é interessante discutir a metodologia implementada aqui para estimar as energias livres de ligação.

Como mencionado na secção anterior, ao notar que as energias livres calculadas pelo Rosetta guardavam pouca correlação com dados experimentais, teve início do desenvolvimento de uma metodologia baseada em ML para refinar a função de energia do Rosetta. O primeiro passo nessa direção foi feito através da aplicação de uma rede neural artificial sobre as funções

de energia do Rosetta para determinar os valores de ΔΔG de ligação entre proteínas. Redes neurais artificiais são estabelecidas pelo fluxo de informação em camadas, com transformações ocorrendo a cada estágio, culminando em um *output* que é comparado com o resultado desejado. Havendo discrepâncias entre os resultados esperados e o *output* da rede, esta reajusta os pesos aplicados em cada camada para melhorar o desempenho do programa, em processo conhecido como *backpropagation* (do inglês, propagação retrógrada), essencial para a etapa de treinamento da rede. As redes neurais artificias aprendem com o tempo e com o treinamento. (D. K. et al., 2023) É importante que uma rede neural artificial apresente robustez, isto é, que os resultados obtidos com o conjunto de dados de treinamento sejam reprodutíveis a conjuntos de dados reais distintos, garantindo que a rede não apenas reproduziu (ou memorizou o cálculo) o que deu certo no conjunto de treino, mas que "aprendeu" a calcular determinada propriedade em conjunto com dados muito distintos. Num primeiro momento, o grupo utilizou a função de energia do Rosetta (e seus componentes do particionamento) para calcular o valor do ΔΔG de ligação entre pares proteicos, comparando o valor final (output do programa) com os valores obtidos experimentalmente. Após desenvolvimento da rede e treinamento, o modelo apresentou correlação experimental significativa, variando de 1,68 a 2,45 kcal/mol (desvio quadrático médio). (Ferraz et al., 2023)

Após a publicação dos resultados acima, o grupo implementou uma nova metodologia de ML chamada *superlearner* para estimar o ΔΔG de ligação entre proteínas em pacote computacional chamado PBEE (*Protein Binding Energy Estimator*, do inglês, programa que estima energias livres de ligação entre proteínas). Os algoritmos de *superlearner* utilizam um conjunto de outros algoritmos já conhecidos de ML para predição da variável desejada. Há pouca informação na literatura sobre a metodologia do *superlearner* (de modo geral), mas há diversos exemplos publicados de suas aplicações. (Ehwerhemuepha et al., 2021; Lee et al., 2023) O algoritmo *superlearner* utiliza, em termos gerais, outros algoritmos de ML já conhecidos e implementa um modelo de validação cruzada para estimar o desempenho destes modelos em diferentes condições, criando uma média otimizada e ponderada para estes modelos (isto é, um ensamble) a partir dos dados de desempenho de cada um. A versão mais nova do programa utiliza o software Rosetta para calcular parâmetros das PPI (implementados no módulo *Interface Analyzer* do Rosetta, a partir da função de energia e da parametrização já descritas) e utiliza estes dados como *input* para o algoritmo *superlearner*, gerando o valor do ΔΔG de ligação como resultado. O algoritmo foi treinado utilizando dados experimentais cristalográficos obtidos do PBD (e comparados com seus valores experimentais de ΔΔG). Os

resultados obtidos com este modelo apresentaram correlação experimental ainda mais significativa, e o programa está disponível para utilização no repositório do *GitHub* <<https://github.com/chavesejf/pbee/>>, e mais detalhes sobre seu uso e implementação podem ser obtidos na publicação original. (Chaves et al., 2025)

3. METODOLOGIA

Os métodos apresentados no presente trabalho foram adaptados dos protocolos estabelecidos pelo grupo de pesquisa durante a pandemia de SARS-CoV-2. Os protocolos originais estão detalhados e documentados no artigo "*The ongoing evolution of variants of concern and interest of SARS-CoV-2 in Brazil revealed by convergent indels in the amino (N)-terminal domain of the spike protein*" publicado pelo grupo em colaboração com outros membros da FIOCRUZ no Brasil. O protocolo foi adaptado e expandido desde esta publicação visando resolver nosso problema específico, mas a referência supracitada foi destacada pelo seu papel ímpar como fio condutor dos protocolos aqui utilizados. (Resende et al., 2021)

O fluxograma geral dos procedimentos metodológicos adotados no presente trabalho se encontra representado abaixo (Fig. 10). Inicialmente, partimos de sequências de RNA caracterizadas experimentalmente pela rede genômica da FIOCRUZ e as traduzimos utilizando *scripts* em Python, obtendo a sequência de resíduos de aminoácidos (estrutura primária) das variantes de DENV2 circulantes no período de 2023-2024. Como tais variantes não estão com estrutura cristalográfica resolvida no PDB, partimos de estruturas encontradas em proteínas E mais antigas para utilizar como proteínas de referência. Inicialmente, não definimos valores de referência para os anos a partir dos quais a variante seria definida como antiga. Porém, como só foram encontradas estruturas de referência no PDB (preenchendo todos os pré-requisitos deste trabalho) circulantes nos anos 2010-2014, estes anos foram adotados para definir as cepas de circulação mais antiga em comparação com as cepas mais recentes, que circularam em 2023-2024. A partir deste momento no texto, para otimizar a compreensão e fluidez da leitura, referências a cepas antigas ou recentes corresponderão aos intervalos de circulação explicitados. Isto também implica que as antigas tiveram suas estruturas obtidas diretamente do PDB, enquanto as mais recentes tiveram sua estrutura modelada computacionalmente a partir dos dados genômicos supracitados. Assim, cepas antigas e recentes poderão também ser denominadas estruturas nativas (ou estruturas de referência) e estruturas mutadas (ou estruturas mutantes), respectivamente, ao longo do texto.

As estruturas de referência do PDB foram obtidas em complexo com anticorpos. Utilizamos os dados da estrutura primária das novas cepas para introduzir mutações nas proteínas de referência e, assim, obter estruturas das proteínas E das cepas novas, em complexo com os anticorpos antigos. Este esquema visa simular computacionalmente as condições que serão realizadas pelo sistema imune: a partir do reconhecimento de proteínas E, anticorpos

elicitados pela memória de infecções anteriores pelo mesmo sorotipo seriam produzidos. Assim, podemos comparar a diferença nas propriedades termodinâmicas de interação entre as proteínas E de cepas mais antigas (obtidas do PDB) e seus anticorpos *versus* a interação entre as proteínas E de cepas mais novas (modeladas computacionalmente) e estes mesmos anticorpos. Diferenças significativas nestas propriedades, em particular transições de valores de ΔΔG negativos para positivos, poderiam indicar que as proteínas E de cepas novas não interagem mais da mesma forma com os anticorpos gerados pela infecção prévia do mesmo sorotipo. Tal resultado corroboraria nossa hipótese de que a resposta imune elicitada por cada sorotipo não confere imunidade específica permanente para aquele sorotipo. Os resultados destes cálculos foram analisados estatisticamente utilizando ferramentas de ciência de dados em Python e R.

Após a presente explanação inicial do panorama metodológico geral do presente trabalho e da nomenclatura que será adotada, cabe o detalhamento dos aspectos metodológicos em cada etapa. Para isto, será adotada estrutura de perguntas condutoras a cada tópico, tanto da apresentação da metodologia, quanto dos resultados. Esta abordagem é compatível com o método científico hipotético-dedutivo clássico de Popper, que se encontra alinhado com o presente programa de investigação científica. A estratégia para buscar respostas cada pergunta será detalhada na presente secção (Metodologia), enquanto as respostas propriamente ditas ficarão a cargo das respectivas seções de apresentação e discussão dos resultados.

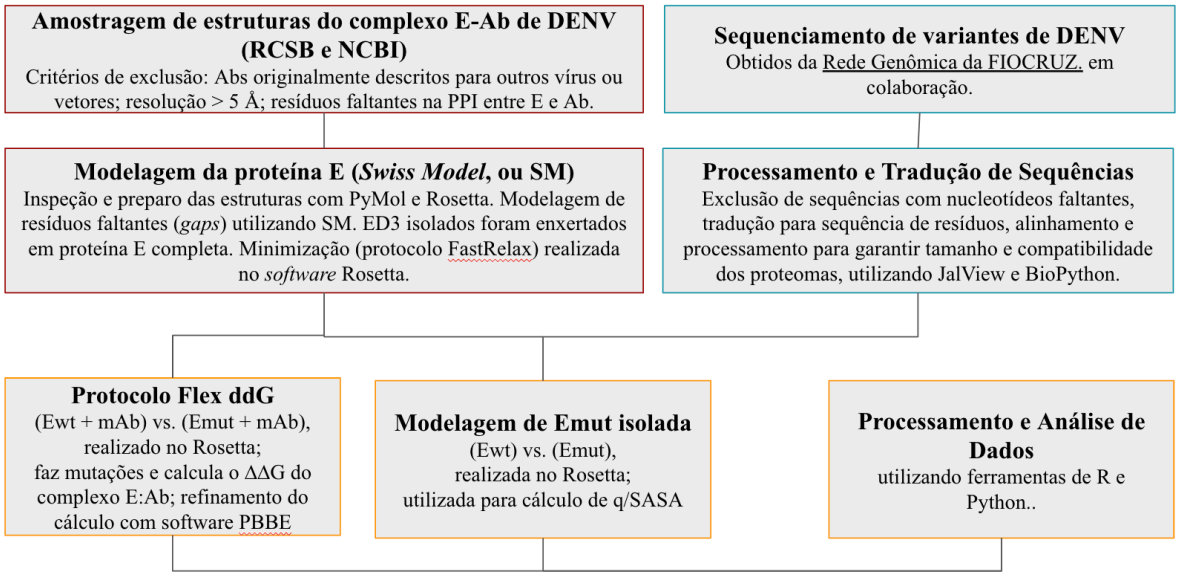


Figura 10. Esquema (*pipeline*) geral dos procedimentos metodológicos adotados na presente investigação. Em contorno vermelho, representamos os dados cristalográficos obtidos dos bancos de dados RCSB e NCBI,

representando estruturas de cepas mais antigas em complexos com seus respectivos anticorpos. Em contorno turquesa, representamos as etapas de processamento de dados das cepas mais novas (obtenção das estruturas primárias). Em amarelo, as etapas finais, em que os dados foram reunidos (introdução das mutações de cepas novas nas antigas) e comparados para teste de hipóteses.

3.1 HÁ ESTRUTURAS DE COMPLEXOS E:AB NO PDB?

Para responder a esta pergunta condutora inicial, buscamos estruturas cristalográficas do complexo formado entre a proteína E e seus anticorpos de ligação específicos. Incluímos apenas estruturas contendo a proteína E (ou apenas partes dela, por exemplo, domínios de ligação específicos) em complexo com um anticorpo de mamífero ligado a ela. Foram critérios de exclusão do presente estudo: estruturas que continham complexos com anticorpos que foram originalmente descritos para outros vírus ou outras espécies de vetores/recipientes; estruturas com resolução ruim (ou seja, maior que 5,00 A); estruturas com resíduos ausentes na superfície de interação (entre a proteína E e o respectivo anticorpo). Os termos de busca no PDB foram resumidos da seguinte forma:

```
<< QUERY: Full Text = "DENV envelope" AND ( ( Scientific Name of the Source Organism = "dengue virus type 2" OR Scientific Name of the Source Organism = "dengue virus type 3" OR Scientific Name of the Source Organism = "dengue virus type 1" OR Scientific Name of the Source Organism = "dengue virus type 4" ) AND Polymer Entity Type = "Protein" ) >>.
```

Os critérios de exclusão foram testados visando reduzir possíveis fontes de vieses metodológicos, mantendo o rigor da pesquisa conduzida, mas sem reduzir excessivamente a quantidade de estruturas geradas, o que resultaria em análise estatística pobre por insuficiência de dados. Tal equilíbrio é difícil de ser atingido, considerando a escassez de estruturas resolvidas (especialmente em complexo com seus respectivos anticorpos) com boa resolução nas bases de dados, notadamente para doenças tropicais negligenciadas. Assim, optamos por manter estruturas que contivessem apenas um dos domínios da proteína do envelope (além das estruturas que continham toda a proteína E) em complexo com anticorpo. Para reduzir vieses metodológicos (do uso da proteína E inteira para algumas estruturas referências e apenas do domínio imunogênico/funcional em outras), as estruturas contendo apenas o domínio da proteína E foram enxertadas em PDB de referência. Desta forma, foi gerada computacionalmente proteína E completa em complexo com anticorpo, conforme detalharemos

na próxima secção. Através desse procedimento, foi possível manter um número maior de estruturas cristalográficas, permitindo adequada análise estatística em etapas suficientes, sem perder significativamente o rigor metodológico ou introduzir vieses. O critério de resolução também foi necessário para manter estruturas com qualidade suficiente para modelagem de estrutura terciária. Já o critério de resíduos ausentes teve finalidade específica: é relativamente comum que estruturas cristalográficas não disponham de todos os resíduos resolvidos (por questões inerentes ao método, geralmente relacionadas ao problema de fases de Bragg). Tais resíduos faltantes costumam ser modelados computacionalmente, mitigando o problema de sua falta nas estruturas trabalhadas; entretanto, tal processo pode gerar diferenças de ΔΔG. Portanto, como as propriedades termodinâmicas calculadas refletem as interações entre pares (proteína E:Ab), estruturas cristalográficas contendo resíduos faltantes na superfície de interação poderiam ser fonte de vieses, e foram excluídas. Já estruturas com resíduos faltantes fora da superfície de interação foram devidamente aceitas e modeladas, conforme será descrito a seguir (cf. secção 3.3).

Como mencionado previamente, não foram encontradas estruturas de circulação recente no PDB. Como a avaliação do ΔΔG destas é essencial para responder a hipótese principal do presente trabalho, considerando-se que tomamos o ΔΔG para definir quantitativamente evasão imune, as estruturas foram geradas de outra forma. Partimos de um levantamento genômico de amostras de DENV2 que fora sequenciado pela Rede de Vigilância Genômica da Fiocruz. Obtivemos as sequências de RNA específicas para a proteína E das cepas de DENV em circulação global atual (nos anos de 2023-2024, em que este estudo foi conduzido). As sequências foram então traduzidas em sequências de aminoácidos usando pacotes de *biopython* e depois inspecionadas usando JalView®. As sequências que traduziram resíduos para além da sequência da proteína E foram ajustadas (ou seja, os resíduos "excessivos" foram descartados) e alinhadas no JalView®. As sequências redundantes foram descartadas para evitar computações repetidas. Por fim, obtivemos 102 sequências não redundantes para DENV2. O descarte de sequências foi também realizado em script em Python, utilizando as menores referências incluídas (isto é, sem resíduos não resolvidos) na análise. Aquela que tivesse início mais tardio (em relação ao número de resíduos no alinhamento de sequências) foi selecionada como referência para o resíduo inicial da proteína E. Já aquela que tivesse término mais precoce de sequência foi selecionada como referência para o resíduo final da proteína E. O corte foi feito tomando estas referências, eliminando resíduos a jusante da referência de início e a montante da referência de término. Este procedimento foi necessário pois, em etapas

posteriores, como comparam-se os valores de ΔΔG calculados pelo mesmo método para proteínas nativas e para mutantes, proteínas de envelope com tamanhos diferentes poderiam introduzir vieses de cálculo, levando a diferenças de ΔΔG que poderiam ser atribuíveis às diferenças de tamanho de sequência, não das mutações introduzidas. Inicialmente, foram testadas outras estruturas e scripts separados para esta etapa. Após otimização de protocolo, os diversos scripts em suas versões finais foram reunidos em um único, visando facilitar a reprodutibilidade dos procedimentos aqui apresentados (ver Apêndices, Script 1).

3.2 QUAL É O MODO DE LIGAÇÃO DAS PPI NOS COMPLEXOS DE REFERÊNCIA?

Para responder a esta pergunta condutora, foram utilizados softwares de visualização molecular VMD (Humphrey; Dalke; Schulten, 1996) e PyMol. Além disso, foram utilizados os pacotes computacionais *PLIP* (Avaliador do Perfil de Interações Proteína-Ligante , do inglês, *Protein-Ligand Interaction Profiler*) e *APBS Electrostatics* (Resolvedor Eletrostático de Poisson-Boltzmann Adaptativo, do inglês, *Adaptative Poisson-Boltzmann Solver Electrostatics*) implementados no PyMol para estudar as interações específicas dos resíduos nas PPI e para avaliar o comportamento eletrostático da superfície dos componentes da PPI, respectivamente. O pacote *PLIP* utiliza um conjunto de regras para classificar as interações que compõem a PPI em sete grupos possíveis, a saber ligações de hidrogênio, contatos hidrofóbicos, interações pi-pi (conhecidas, em inglês, como *pi-stacking*), interações pi-cátion, pontes salinas, pontes aquosas e ligações halogênicas. (Salentin et al., 2015) Através deste, é possível caracterizar os tipos de interações químicas que contribuem para a formação dos complexos. Já o pacote *APBS Electrostatics* resolve equações de eletrostática continua para sistemas biomoleculares grandes rapidamente, permitindo a avaliação de modelos acurados de distribuição de cargas na superfície de biomoléculas. (Jurrus et al., 2018)

3.3 HÁ DIFERENÇAS DE ΔΔG DOS COMPLEXOS E:AB DE DENV2 ENTRE ESTRUTURAS NATIVAS E MUTADAS?

Para responder a esta pergunta, foram selecionadas três estratégias de cálculo de ΔΔG dos complexos a partir de informações estruturais. Todas as estruturas precisaram, inicialmente, ser preparadas para evitar cálculos inadequados e erros não sistemáticos. A estrutura cristalográfica resolvida da proteína E ligada aos anticorpos específicos foi recuperada do PDB conforme declarado anteriormente. As estruturas “brutas” foram então inspecionadas e limpas

(ou seja, foram removidos água e outros heteroátomos das estruturas cristalográficas) usando VMD® e PyMol®.(Humphrey; Dalke; Schulten, 1996) Resíduos ausentes de cada estrutura foram modelados usando o modo de modelo de usuário do servidor web Swiss-Model (https://swissmodel.expasy.org/) (Waterhouse et al., 2018) e as proteínas modeladas foram utilizadas como estrutura inicial para DENV WT. Em outras palavras, cada estrutura selvagem continha um complexo limpo (e remodelado, quando necessário) da proteína E de DENV com o respectivo anticorpo.

Então, essas estruturas foram usadas como referência para a modelagem estrutural das variantes da proteína E usando a estratégia de design Flex ddG do software Rosetta (Barlow et al., 2018; Leaver-Fay et al., 2011), adaptada para nossos requisitos específicos. Nessa estratégia, usamos o complexo Ewt:Ab como referência, que é primeiramente minimizada usando o protocolo FastRelax (Conway et al., 2014). Então, o software Rosetta realiza mutações de resíduos para cada uma das variantes DENV cujo genoma foi sequenciado. Para DENV2, 10 estruturas cristalográficas de referência foram encontradas no PDB. Para cada tipo selvagem, o protocolo foi realizado usando as 102 sequências correspondentes às mutações, resultando em 102 x 10 = 1020 espécimes mutantes. O esquema do protocolo Flex ddG proposto em seu artigo de referência se encontra reproduzido abaixo:

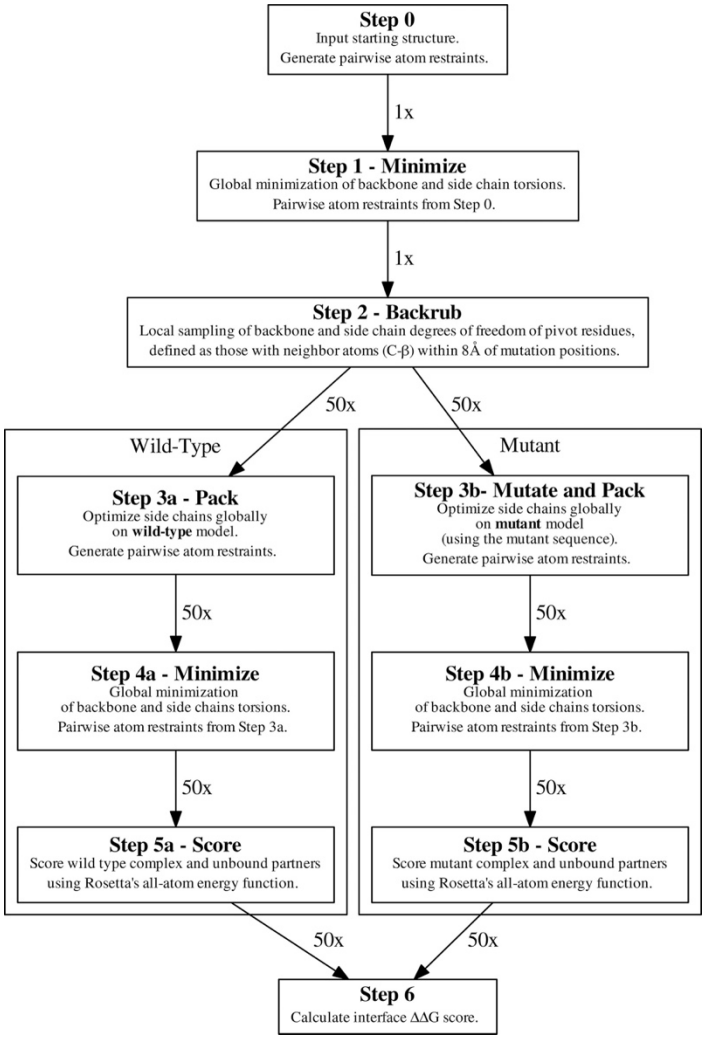


Figura 11. Representação esquemática do protocolo Flex ddG, conforme apresentado por seus autores. A primeira etapa consiste em minimização e backrub das estruturas de input. A seguir, as mutações são realizadas conforme arquivos de resfiles fornecidos pelo usuário. As mutações se acompanham de packing (otimização de cadeia lateral). A seguir, os grupos mutados e não mutados (WT) passam por etapas sucessivas de minimização e cálculo de função de energia. Ao final, o ΔΔG de interface é calculado por método de particionamento de energia, conforme implementado na função de energia do Rosetta. Fonte: extraído de (Barlow et al., 2018).

Em cada variante, o programa procura resíduos diferentes (quando comparado ao nosso modelo de tipo selvagem) e muta os resíduos específicos para seu correspondente no proteoma sequenciado. O protocolo Flex ddG também faz rodadas sucessivas de minimização de ambas as estruturas (ou seja, tipo selvagem e mutante), implementadas com restrições harmônicas às distâncias de pares atômicos da estrutura cristalográfica original, e realizadas até a convergência (definida como diferenças de *score* de energia entre as rodadas inferior a 1 REU). No final, o ΔΔG é calculado para o tipo selvagem e para as mutações, permitindo a comparação entre eles. O cálculo é feito aplicando a função de energia do Rosetta (aproximada como Hamiltoniano do sistema) em ciclo termodinâmico, conforme explicado na fundamentação teórica (cf. secção

2.6.2). Para cálculo de múltiplas mutações, a documentação oficial do programa aponta coeficiente de Pearson R = 0,62 e erro médio absoluto (MAE) de 1,62. (Barlow et al., 2018) As estruturas modeladas das variantes da proteína E foram sobrepostas nas coordenadas dos PDBs obtidos para visualizar as diferenças entre as interfaces de interação proteína E:Ab. A renderização das imagens foi realizada utilizando os softwares VMD® e PyMol®. (Humphrey; Dalke; Schulten, 1996) Os resultados do protocolo Flex ddG foram analisados utilizando o script em R no RStudio® (cf. Apêndices, script 6). A implementação deste protocolo tem como vantagem adicional o fato de o programa gerar como *output*, além do valor de ΔΔG, os PDBs das estruturas mutantes. Para realizar as mutações, geramos *resfiles* distintos para as 102 cepas recentes de DENV sequenciadas pela Rede Genômica da FIOCRUZ. Um *resfile* é um arquivo de texto com formatação específica para ser utilizado como *input* do Rosetta para especificar as mutações em cada posição. Um *resfile* pode ser interpretado como um "guia" para o programa realizar as mutações sobre a proteína de *input*, especificando quais resíduos serão modificados em cada posição ao longo da sequência. Conforme explicado na secção 3.1, a partir dos dados de sequenciamento genômico, obtivemos estruturas primárias para as proteínas E de novas cepas circulantes de DENV. Estes dados foram utilizados para implementar as mutações nesta etapa. Inicialmente, geramos os *resfiles* manualmente para cada mutação. Porém, com o maior número de mutações, a gênese de *resfiles* a partir da estrutura primária das proteínas E novas foi implementada como parte do Script 1, conforme descrito em 3.1. Assim, como output final do protocolo Flex ddG, foram obtidos os PDBs das proteínas E de novas cepas (já em complexo com os anticorpos antigos, isto é, complexos Emut:Ab) e os valores de ΔΔG de interface (entre a proteína E e o anticorpo) quando comparadas com as proteínas E de referência (isto é, das cepas antigas).

Após a implementação deste protocolo, os PDBs gerados dos complexos proteína E mutadas (representando as cepas novas das proteínas E) e seus respectivos anticorpos foram utilizados para calcular o ΔΔG por outros métodos (a saber, FoldX e PBEE), uma vez que o método de particionamento de energia do Rosetta tem baixíssima correlação experimental. Inicialmente, realizamos teste com outro programa que também utiliza particionamento de energia, o FoldX (cf. secção 2.6.2). Esta aplicação aceita PDB de complexos proteicos como *input* e consiste em vários programas com funções distintas. Para o nosso caso, foi necessário utilizar os PDBs de complexos Emut:Ab gerados no Rosetta como *input*. Foi utilizado o programa *AnalyseComplex* (do inglês, Análise de Complexo) da aplicação mencionada. Para obter valores de referência, também fornecemos as estruturas de referência (complexos

Ewt:Ab) devidamente modeladas e minimizadas como *input* e os resultados da energia livre de ligação também foram expressos relativamente (ΔΔG da mutação em relação ao nativo). A documentação oficial do programa traz os coeficientes de Pearson e desvio padrão encontrados em outros trabalhos na literatura, que variam de 0,29 a 0,73 e 1,0 a 1,78 kcal/mol, respectivamente. (Broom et al., 2017; Dehouck et al., 2009; Yue et al., 2023) Para manter o rigor metodológico, na secção resultados e discussão consideramos os valores de menor confiança para a discussão (isto é, R = 0,29 e desvio padrão de 1,78 kcal/mol). Não foram encontrados registros de erro médio absoluto para este programa na literatura até o momento.

Além do FoldX, como mencionado, o *output* do Rosetta serviu como *input* para outro programa de cálculo de ΔΔG, o PBEE. Idealmente, o cálculo de ΔΔG poderia ser implementado utilizando simulações de MD para cada uma das proteínas (nativa e mutada) por métodos alquímicos, porém tais métodos tem custo computacional impeditivo para a grande quantidade de proteínas analisadas neste trabalho (1030 estruturas mutadas + 10 estruturas de referência = 1040 cálculos alquímicos), conforme discutido previamente (cf. secção 2.6.2). Por isso, optamos por método novo, baseado em aprendizagem de máquina, que apresenta boa correlação experimental, mas custo computacional baixo, desde que o usuário forneça os PDBs já mutados como *input*. Este método foi desenvolvido pelo grupo, utilizando aprendizagem de máquina sobre as partes da função de energia calculadas do Rosetta, no pacote computacional PBEE (*Protein Binding Energy Estimator*). Este pacote foi submetido para publicação, mas sua documentação já se encontra disponível em <<https://github.com/chavesejv/PBEE>>. Nele, o PDB fornecido como *input* é analisado, procurando por *gaps* no *backbone* proteico e íons próximos a átomos da proteína. Na presença dos primeiros, o processo é terminado, na dos segundos, estes são removidos. Os PDBs que utilizamos, entretanto, foram limpos previamente e tiveram seus resíduos faltantes remodelados, conforme descrito anteriormente. A seguir, o PBEE aciona o Rosetta para otimização geométrica e análise de interface, conforme implementado nos pacotes de *Geometry Optimization* e *Interface Analysis* do Rosetta. A partir dos dados obtidos da função de energia do Rosetta e da análise de interface, uma arquitetura de *superlearner* implementada em Python3 foi desenvolvida para estimar valores de ΔΔG com excelente correlação experimental (cf. secção 2.6.3). Utilizamos este programa para calcular os valores de ΔΔG dos mutantes e compará-los com os valores de referência. O resultado gerado pelo protocolo anterior, do Flex ddG, já é relativo, isto é, já representa a diferença nos valores de ΔΔG de interface introduzidas pelas mutações realizadas (ou seja, a diferença entre os valores de ΔΔG dos complexos Ewt:Ab e Emut:Ab). Já no PBEE, os resultados são do ΔΔG de

interface da proteína E com o anticorpo. Por isso, utilizamos como *input* para o PBEE os PBDs de referência (assim como o procedimento adotado para o Flex ddG) e os PDBs das mutações geradas como *output* do Flex ddG. Na análise estatística do PBEE, realizamos uma etapa adicional de cálculo do ΔΔG relativo (com base nas proteínas de referência), por subtração simples dos valores obtidos para as mutantes em relação às suas respectivas estruturas WT. Os resultados foram analisados e convertidos em gráficos utilizando RStudio (Script 6). A documentação oficial do programa traz erro quadrático médio (RMSE) de 1,98 e coeficiente de Pearson (R) de 0,68. Interessantemente, embora a documentação oficial do Flex ddG do Rosetta relate que este tem coeficiente de Pearson de 0,62, ao testar sua performance no mesmo conjunto de dados que o seu programa, os autores do programa PBEE encontraram R = 0,08. Desta forma, para o conjunto de dados utilizado na documentação oficial do PBEE, este apresentou correlação experimental (R = 0,68), mas o Rosetta não (R = 0,08).

3.4 DIFERENÇAS DE ΔΔG DOS COMPLEXOS E:AB DE DENV2 PODEM SER EXPLICADAS QUIMICAMENTE?

Para responder a esta pergunta condutora, a inspeção cuidadosa dos complexos nos softwares de visualização molecular VMD e PyMol foi realizada, conforme metodologia já descrita na secção 3.2. Para avaliar esta pergunta sistematicamente, porém, escolhemos especificamente os candidatos que apresentaram maior valor de aumento de ΔΔG em cada método (caso haja, no método em questão, algum candidato com aumento de ΔΔG acima do valor de referência de 2,81 kcal/mol, conforme definido na secção 1.3). Os candidatos selecionados foram cuidadosamente inspecionados comparando as PPI dos complexos das proteínas nativas com as PPI dos complexos das proteínas mutadas. Os resultados quantitativos obtidos na secção anterior são, então, qualitativamente avaliados quanto à sua plausibilidade química. Alguns dos complexos E:Ab nativos têm seus valores experimentais de K_d documentados nos artigos de origem da estrutura. Nestes casos, é possível calcular o valor do ΔΔG experimental, utilizando as equações da secção 1.3, e, assim, comparar o desempenho de métodos distintos frente ao conjunto de dados do presente trabalho. Esta etapa não responde diretamente à pergunta condutora, mas foi incluída posteriormente para explicar divergências de resultados entre métodos. Assim, sua inclusão na metodologia foi importante para complementar perguntas que surgiram como consequência dos resultados obtidos para as perguntas condutoras 3.3 e 3.4, por motivos que serão pormenorizados na secção correspondente.

4. RESULTADOS E DISCUSSÃO

Continuaremos ao longo desta secção com o mesmo procedimento adotado na secção de metodologia: cada subseção será intitulada com a pergunta condutora que a motivou. Logo a seguir, no texto em itálico, estará destacada a resposta sucinta da pergunta conduta. Após a resposta sucinta será fornecido detalhamento da resposta, em texto normal, geralmente em ordem cronológica das descobertas de cada etapa.

4.1 HÁ ESTRUTURAS DE COMPLEXOS E:AB NO PDB?

Encontramos 10 estruturas de referência no PDB para complexos E:Ab de cepas de DENV2 circulantes entre 2010-2014 e nenhuma estrutura correspondente de cepas circulantes entre 2023-2024. A partir de dados de sequenciamento da Rede Genômica da FIOCRUZ, foi possível gerar computacionalmente 102 estruturas mutantes distintas para cada uma das estruturas de referência, totalizando 1020 estruturas correspondentes a variantes circulantes entre 2023 e 2024.

Os resultados do catálogo de estruturas obtidas do RCSB e NCBI são descritos na Tabela 1. A busca levou a 164 estruturas no total, 25 das quais passaram pelos nossos critérios de seleção (cf. Métodos, secção 3.2) e são descritas na Tabela 1 abaixo. Inicialmente, ainda não havíamos definido DENV2 como sorotipo alvo para este estudo, portanto a catalogação foi feita para todos os sorotipos de DENV. Após definido DENV2 como sorotipo alvo para o presente estudo, pela sua importância epidemiológica no Brasil e pela existência de quantidade suficiente de estruturas para seu estudo sistemático, mantivemos o catálogo das demais para referência e comparação.

Sorotipo de DENV (cepa)	**PDB id (resolução em Å)**	**Constante de Dissociação (Kd)**	**Observações**
1 (Western Pacific-74)	4FFZ (3.80)	416±24 nM	Fragmento Fab de DENV1-E111 ligado ao ED3 de DENV1; pontos de contato: S338, G344, e A345. Ab de camundongo

			Neutralizante (DENV1 apenas)
1 (16007)	4FFY (2.50)	18±0.08 nM	Fragmento Fv de cadeia única de DENV1-E111 ligado ao ED3 de DENV1. Ab de camundongo Neutralizante (DENV1 apenas)
1 (Desc.)	4L5F (2.45)	Desc.	A ser publicado.
1 (Hawai H241)	4AL8 (1.66)	0.45 nM	ED3 de DENV1 em complex com Fab 2H12; Kd determinado por ELISA de captura de antígeno; Ab de camundongo Reatividade cruzada (exceto DENV2)
1 (Guiana/FGA89/1989)	3UZQ (1.60)	0.082±0.005 nM	Complexo ED3 de DENV1 com Fv de Ab monoclonal 4E11. Ab de camundongo Reatividade cruzada
2 (NGC)	6FLA (2.90)	0.92 nM	Fab 3H5 ligado ao ED3 de DENV2; Fab de 3H5 se liga rapidamente ao ED3 (ka ~ 9.0x10ˆ5 M-1s-1) e se dissocia lentamente (kd ~ 8.3x10ˆ-4 s-1). Ab de camundongo Reatividade cruzada
2 (NGC)	6FLB (2.20)	0.92 nM	Fab 3H5 ligado ao ED3 de DENV2. Ab de camundongo Reatividade cruzada

2 (NGC)	6FLC (2.00)	1.6 x 10ˆ-7 M (~160 nM)	Fab 2C8 Fab ligado ao ED3 de DENV2; 2C8 se associa lentamente (ka ~ 6.2x10ˆ5 M-1s-1) e se dissocia rapidamente (kd ~ 0.1 s-1) Ab de camundongo. Neutralizante (DENV2 apenas)
2 (Jamaica/1409 /1983)	3UZV (2.10)	0.43 ± 0.04 nM	ED3 em complex com Fv do Ab 4E11 Ab de camundongo Reatividade cruzada
2 (FGA-02)	4UT9 (3.20)	Desc.	E de DENV2 em complex com Fv do Ab EDE1 C10. Ab humano. Pan-neutralizante.
2 (FGA-02)	4UT6 (3.20)	Desc.	E de DENV2 em complexo com Fab de Ab EDE2 B7. Ab humano. Pan-neutralizante.
2 (FGA-02)	4UTA (3.00)	Desc.	E de DENV2 em complex com Fab do Ab EDE1 C8. Ab humano. Pan-neutralizante.
2 (FGA-02)	4UTB (3.85)	Desc.	E de DENV2 em complex com Fab do Ab EDE2 A11. Ab humano. Pan-neutralizante.

2 (Tailandia/16681/84)	2R69 (3.80)	Desc.	E de DENV2 em complex com Ab monoclonal 1A1D-2. Ab murino. Neutralizante (especialmente para DENV1, DENV2, DENV3)
2 (Tailandia/16681/84)	2R29 (3.00	Desc.	E de DENV2 em complex com Ab monoclonal 1A1D-2. Ab murino. Neutralizante (especialmente para DENV1, DENV2, DENV3)
3 (H87)	4ALA (1.84)	0.47 nM	ED3 de DENV3 em complex com Fab 2H12; Kd determinado por ELIAS de captura de antígeno. Ab de camundongo. Reatividade cruzada (exceto DENV2)
3 (Tailandia/PaH881/ 1988)	3UZE (2.04)	8.0±1.0 nM	ED3 de DENV3 em complexo com Fv do Ab monoclonal 4E11; Kd determinado por ELISA de competição. Ab de camundongo. Pan-neutralizante.
3 (EHIE46200Y19)	8JN1 (3.50)	Desc.	E de DENV3 em complexo com Ab DENV-115 IgG; Ab humano; Neutralizante (DENV3); A ser publicado.

3 (863DK)	8JN2 (4.10)	Desc.	E de DENV3 em complexo com Fab do Ab DENV-115; Ab humano; Neutralizante (DENV3); A ser publicado.
3 (853DK)	8JN4 (3.50)	Desc.	E de DENV3 em complexo com Fab de Ab DENV-290; Ab humano; Neutralizante (DENV3); A ser publicado.
3 (853DK)	8JN5 (3.60)	Desc.	E de DENV3 em complexo com Fab de Ab DENV-290; Ab humano; Neutralizante (DENV3); A ser publicado.
4 (desc.)	4BZ1 (2.15)	Desc.	ED3 de DENV4 em complexo com Fab 3e31; Ab de camundongo; Pan-neutralizante. A ser publicado.
4 (desc.)	4BZ2 (2.03)	Desc.	ED3 de DENV4 em complexo com Fab 3e31; Ab de camundongo; Pan-neutralizante. A ser publicado.

4 (Burma/63632/1976)	3UYP (2.00)	4,100±700 nM	ED3 de DENV4 em complexo com Fab do Ab monoclonal 4E11; Kd avaliado por ELISA de competição; Ab de camundongo; Pan-neutralizante. A ser publicado.
4 (H241)	4AM0 (3.02)	0.42 nM	ED3 de DENV4 em complexo com Fab do Ab 2H12; Kd determinado por ELISA de captura; Ab de camundongo; Pan-neutralizante.

Tabela 1. Catálogo das estruturas de complexos Ewt:Ab obtidos do PDB. Desc. = desconhecido.

Das 25 estruturas acima, selecionamos as 10 estruturas de DENV2 para servir de referência para as mutações (guiadas pelos dados genômicos mencionados na secção 3.1). Destas 10 estruturas, 4 representavam a proteína do envelope completa (PDBid: 4UTA, 4UTb, 4UT6 e 4UT9), e as 6 restantes representavam apenas o domínio 3 (ED3), necessitando de enxertia na glicoproteína E de referência (adotamos aleatoriamente a estrutura 4UTB para este fim). Examinaremos o perfil de interações destas sequências na próxima secção do trabalho.

Por hora, respondemos à pergunta condutora atual em relação a cepas mais antigas (isto é, circulantes entre 2010-2014) de DENV2, já que 10 estruturas de complexos foram encontradas para estas. Cabe ainda responder formalmente à pergunta condutora para as cepas mais recentes de DENV2 (afinal, o estudo se propõe a comparar o perfil de ΔΔG de cepas circulantes em 2023-2024 com aquelas que circulavam previamente). Como mencionado, entretanto, não encontramos no PDB estruturas das variantes circulantes em 2023-2024, referentes ao fenômeno epidemiológico que serviu como provocação inicial para a hipótese principal deste trabalho. Para obter estruturas destas cepas, partimos de dados de sequenciamento genômico da proteína E de variantes de DENV2 que circularam no período de 2023-2024. Tais sequências foram obtidas da rede genômica da FIOCRUZ, com os correspondentes nucleotídeos dos RNA-vírus respectivos. Obtivemos 856 sequências, inicialmente. As sequências foram traduzidas utilizando *script* em Python para arquivos em formato *fasta* com as sequências de resíduos de aminoácidos correspondentes. Após eliminar

sequências repetidas (pois cepas diferentes com a mesma sequência de nucleotídeos resultariam em cálculos redundantes de ΔΔG) e aquelas com resíduos faltantes, obtivemos 102 sequências. As 102 sequências (das quais, 47 correspondem a estruturas da Ásia, 33 da América do Sul e 17 da América Central) após tradução e processamento foram alinhadas e representadas nos Apêndices (cf. Secção 8.2).

A partir da inspeção das sequências, notamos mutações ocorrendo em todas as partes da proteína do envelope, mas, interessantemente, notamos que muitas destas se reuniram próximos do resíduo 310, no domínio III (EDIII). A literatura aponta que o domínio mais antigênico da proteína E é o domínio II. Entretanto, este é caracterizado pela ligação a anticorpos não neutralizantes (embora não neutralizantes, estes têm importância no fenômeno de ADE). Quando tomando apenas anticorpos neutralizantes (de fundamental importância na avaliação de evasão imune, já que levam à imunidade protetiva propriamente dita), é o domínio EDIII que tem maior quantidade de anticorpos que se liguem. (Sarker; Dhama; Gupta, 2023) Em outras palavras, embora não sejam o domínio mais antigênico em geral (considerando anticorpos neutralizantes e não neutralizantes), EDIII é o sítio mais antigênico quando considerados apenas anticorpos neutralizantes, donde sua importância. É provável que isto decorra da importância destes domínios na interação da proteína do envelope com proteínas receptoras do hospedeiro, as quais formam um complexo conhecido como receptor de entrada de flavivírus (FVCR, do inglês, Flavivirus Cell Receptor). O EDII participa principalmente da ligação da proteína do envelope a um destes receptores em células dendríticas, chamado de molécula de adesão 3 intercelular específica de células dendríticas (DC-SIGN; do inglês, *Dendritic Cell-Specific Intercellular Adhesion Molecule-3-Grabbing Non-Integrin*). Esta interação é um processo patogênico crucial que facilita a entrada do vírus em células do hospedeiro. Já o EDIII interage com outros receptores que participam do complexo FVCR, promovendo a entrada viral nas células do hospedeiro. Anticorpos neutralizantes contra este receptor impedem a entrada do vírus no hospedeiro, além de ativar o sistema imune do último para destruir o vírus. (Penteado et al., 2024) Desta forma, é provável que o maior número de mutações encontradas na região do EDIII se deva à pressão evolutiva seletiva por anticorpos neutralizantes, dada a importância da sua ligação aos receptores supracitados.

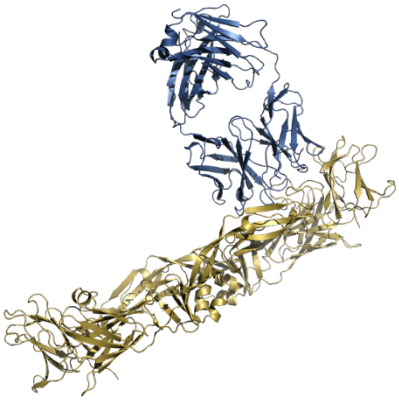
Desta forma, nesta etapa possuíamos 102 sequências de aminoácidos referentes a cepas circulantes em 2023 e 2024, mas nenhuma estrutura destas. Ora, como possuíamos 10 estruturas cristalográficas das cepas circulantes em 2010-2014 (ditas cepas de referência), implementamos mutações nas cepas de referência, guiadas pelas sequências de aminoácidos obtidas para as

cepas mais recentes, sobre cada uma das estruturas de referência, totalizando 102 x 10 = 1020 estruturas de variantes de circulação mais recente de DENV2 (entre 2023 e 2024). Assim, ao final desta etapa, possuíamos 10 estruturas de referência diretamente do PDB (correspondentes às cepas circulantes entre 2010 e 2014) e 1020 estruturas de variantes de circulação mais recente de DENV2 (entre 2023 e 2024), modeladas computacionalmente a partir das estruturas de referência supracitadas. Para cada estrutura de referência, geramos 102 estruturas de mutantes para comparação em etapas futuras.

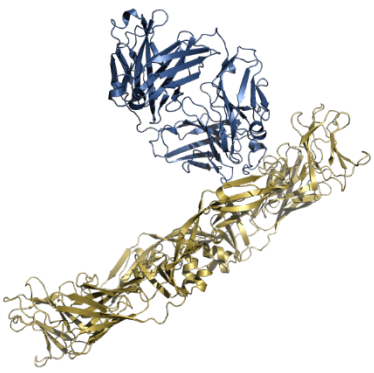
4.2 QUAL É O MODO DE LIGAÇÃO DAS PPI NOS COMPLEXOS DE REFERÊNCIA?

Os 10 complexos de referência apresentam modos de interação distintas entre si, mas todas as PPI dependem de interações hidrofóbicas, eletrostáticas e ligações de hidrogênio. Foi possível agrupá-los geometricamente em complexos com modo de interação perpendicular ou paralela.

A segunda pergunta condutora do trabalho surgiu naturalmente após a obtenção das estruturas. Não seria viável ou eficiente inspecionar cada uma das 1020 estruturas mutantes encontradas em buscas de padrões, tampouco seria esperado que alguma informação valiosa surgisse desta tarefa homérica. Entretanto, há apenas 10 estruturas nativas de complexos E:Ab para DENV2, obtidas de estudos cristalográficos. A inspeção destas é tarefa muito menos demorada e potencialmente muito mais frutífera, já que, ao estudar os complexos nativos das proteínas E com seus anticorpos, talvez fosse possível divisar algum tipo de padrão ou comportamento químico comum que regesse as interações. Reproduzimos abaixo as 10 estruturas de referência a seguir:



A)



B)

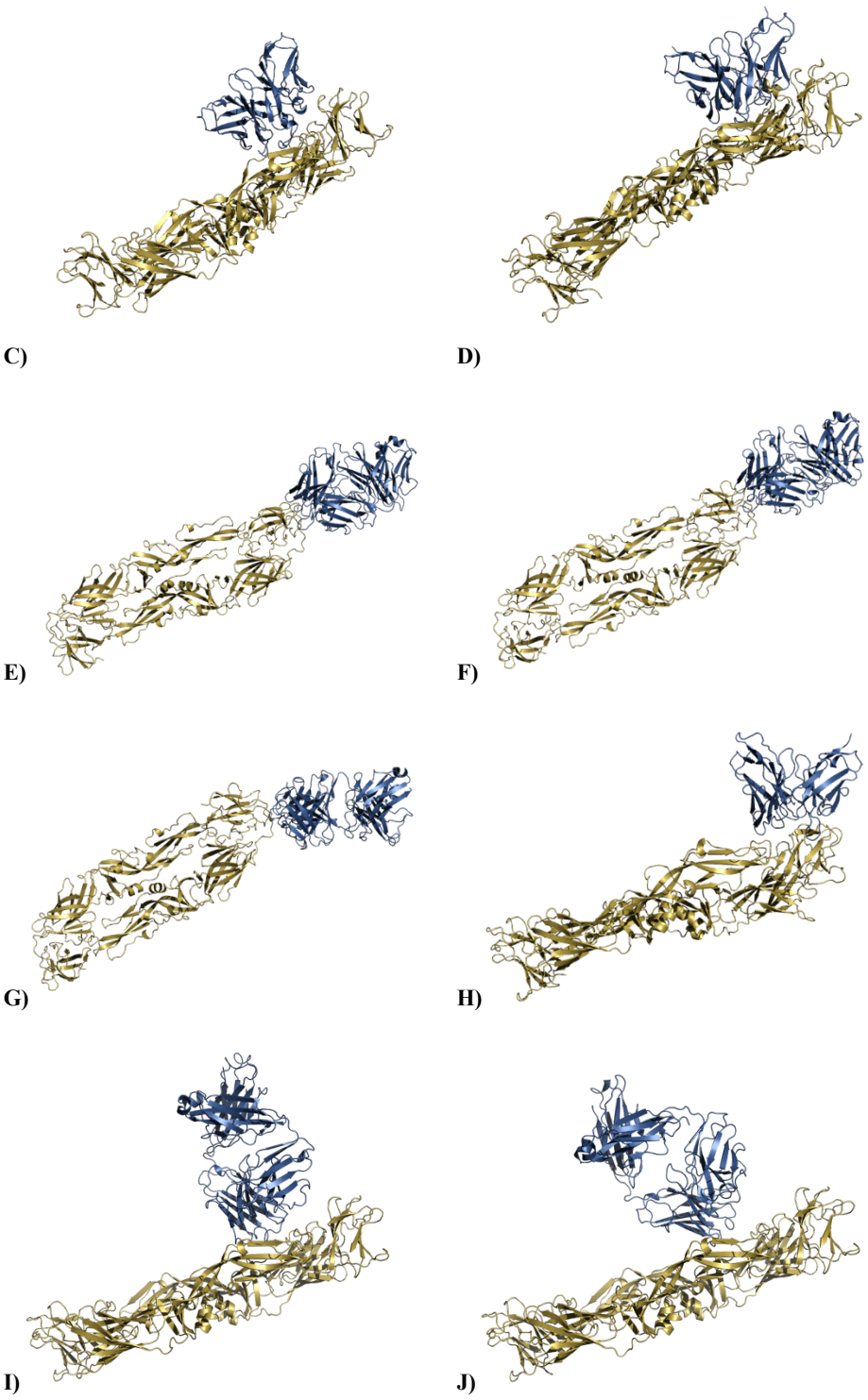


Figura 12. Estruturas de referência dos 10 complexos de proteína E com seus respectivos anticorpos, representação *NewCartoon*. Figuras acima de autoria própria, confeccionadas no software de visualização molecular PyMol, com proteínas do envelope representadas em dourado e anticorpos representados em azul marinho. Os complexos com código 4UTA, 4UTB, 4UT6 e 4UT9 (Figs. 12 A-D) apresentam anticorpos humanos. Os demais complexos (Figs. 12 E-J) apresentam anticorpos murinos. A) PDBid: 4UTA, artigo de referência da estrutura: doi:10.1038/nature14130; B) PDBid: 4UTB, artigo de referência da estrutura: doi:10.1038/nature14130; C) PDBid: 4UT6, artigo de referência da estrutura: doi:10.1038/nature14130; D) PDBid: 4UT9, artigo de referência da estrutura:

doi:10.1038/nature14130; E) PDBid: 6FLA, artigo de referência da estrutura: doi:10.1038/s41590-018-0227-7; F) PDBid: 6FLB, artigo de referência da estrutura: doi:10.1038/s41590-018-0227-7; G) PDBid: 6FLC, artigo de referência da estrutura: doi:10.1038/s41590-018-0227-7; H) PDBid: 3UZV, artigo de referência da estrutura: doi:10.1016/j.str.2012.01.001. I) PDBid: 2R29, artigo de referência: doi:10.1038/nsmb.1382; J) PDBid: 2R69, artigo de referência: doi:10.1038/nsmb.1382.

Na Fig. 12 é notável que o modo de ligação dos anticorpos é distinto, o que é esperado considerando que cada estudo apresentou estrutura cristalográfica própria com anticorpos diferentes. Nas Fig. 11A-11B, que representam estruturas resolvidas com anticorpos humanos, os anticorpos interagem com a proteína E dimérica com relativa ortogonalidade (considerando os eixos de maior dimensão de cada proteína). Já nas demais estruturas, resolvidas com anticorpos de outros mamíferos, há maior variedade no modo de ligação dos anticorpos, podendo ter padrão predominantemente ortogonal (Fig. 12H-11J) ou mais paralelo (isto é, segmentos dos maiores eixos aproximadamente colineares; Fig. 12E-11G). A afirmação de que o modo de ligação ortogonal para anticorpos humanos e variável para anticorpos murinos é restrita a este conjunto de dados e não pode ser generalizada. Isto porque a espécie de origem não é suficiente para explicar as diferenças estruturais supracitadas. Além disso, como há número pequeno de estruturas para cada espécie, considerando a diversidade biológica e genética ampla entre os hospedeiros, que foge do escopo deste trabalho, é provável que os dados reportados nesta secção sejam decorrentes apenas da disponibilidade de estruturas encontradas. Assim, as informações aqui, embora sejam valiosas do ponto de vista exploratório e qualitativo, não permitem traçar correlações específicas entre espécies, tampouco haveria racional biológico para suportá-las. Apesar disso, além da utilidade exploratória, a observação desta diferença permitiu sistematizar (ou categorizar) de alguma forma esta etapa da investigação, na falta de outras diferenças perceptíveis (e categorizáveis em subgrupos) à inspeção inicial. Apesar de haver considerável diferença entre os anticorpos em cada estrutura e em seus modos de ligação, modelamos a estrutura representativa de dois deles (reproduzidas a seguir, Fig. 13), um com modo de ligação de padrão predominantemente ortogonal e outro colinear.

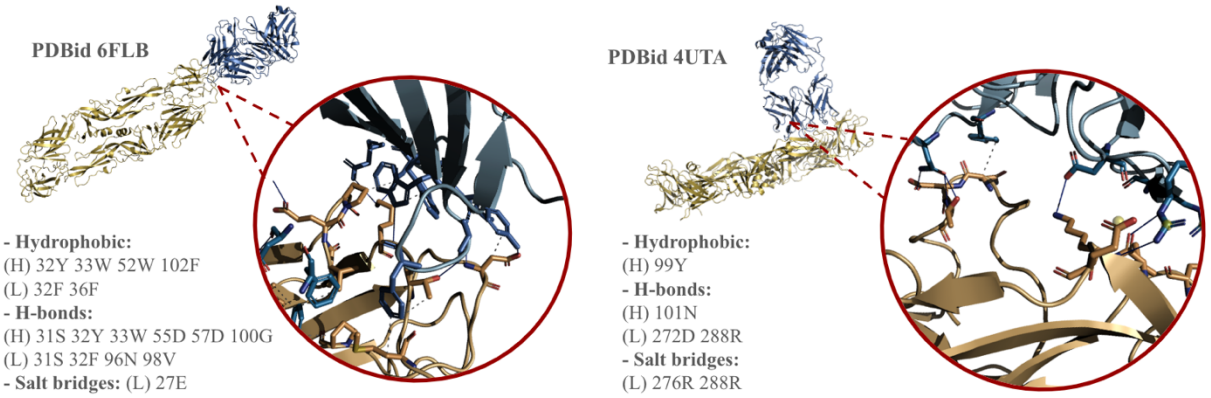


Figura 13. Interações (PPIs) entre a proteína E seus respectivos anticorpos. A esquerda, representamos um anticorpo de ligação predominantemente colinear em relação à proteína E (maiores eixos) e, a direita, um anticorpo humano de ligação predominantemente ortogonal em relação à proteína E (maiores eixos). As figuras foram produzidas utilizando o software de visualização molecular PyMol, com destaque para interações relevantes (na circunferência em vermelho). As interações foram previstas utilizando o plugin PLIP, agrupadas em hidrofóbicas (em inglês, no programa, *Hydrophobic*), ligações de hidrogênio (em inglês, no programa, *H-bonds*) e interações eletrostáticas (em inglês, no programa, *salt bridges*, pontes salinas). Utilizamos código de 1 letra dos aminoácidos (cf. Fig. 1) e, entre parênteses, as letras (H) e (L) representam as cadeias pesada e leve do anticorpo, respectivamente. Destacamos a importância das interações eletrostáticas: à direita, entre duas das argininas (cadeia lateral positivamente carregada; cf. Fig. 1) da cadeia leve do anticorpo e um ácido glutâmico (cadeia lateral negativamente carregada) da proteína E; à esquerda, entre o ácido glutâmico negativamente carregada) da cadeia leve do anticorpo e um resíduo de histidina (cadeia lateral positivamente carregada) da proteína E.

Em ambos os casos todos os principais tipos de interações químicas contempladas pelo programa (interações eletrostáticas, ligações de hidrogênio e interações hidrofóbicas) têm contribuição na interação, conforme esperado em PPI. Este é um dos motivos pelos quais o estudo de PPI é bastante complexo. Esta complexidade pode ser ainda mais considerável em interações do sistema imune, as quais foram refinadas ao longo de muitos anos pelo processo evolutivo das espécies para resultar em utilidade biológica para os organismos (e refinadas especificamente em cada organismo para que o anticorpo reconheça os epítopos do patógeno de interesse biológico do hospedeiro). Apesar desta complexidade, a diferenciação didática das PPI em seus componentes (o que é feito computacionalmente pelo particionamento da função de energia de interações, conforme discutido na secção 2.6.2) é bastante útil para estudo dos padrões de interação entre sistema imune e seu hospedeiro.

Dentro das interações acima definidas, inferimos que a interação eletrostática possa ter contribuição importante candidatos que representam complexos com anticorpos humanos. Tal inferência surgiu da inspeção dos mapas de distribuição de cargas nas superfícies das estruturas (Fig. 14), conforme será detalhado a seguir. Todos os principais tipos de interações químicas são importantes nas PPI acima, conforme visto na Fig. 13. Porém, há algumas peculiaridades na interação eletrostática observadas a partir da análise visual qualitativa do potencial eletrostático de superfície dos candidatos. De antemão, é importante deixar claro o caráter exploratório e qualitativo desta etapa da análise. Assim, as inferências feitas nesta secção a partir da inspeção dos candidatos não devem ser tomadas como absolutas, tampouco conferem alguma utilidade se avaliadas isoladamente, mas podem ser úteis na discussão dos resultados obtidos nas etapas vindouras e podem, também, fornecer suporte qualitativo para as evidências quantitativas que serão apresentadas nas demais secções. Para facilitar a visualização desta

estrutura, separamos o complexo E:Ab e representamos cada proteína separadamente, destacando a superfície da PPI. Nela (Fig. 14), fica evidente que há uma região de cargas negativas (em vermelho) na proteína E destes candidatos complementada por uma região similar de cargas positivas (em azul) nos anticorpos humanos. Estas regiões de complementariedade de carga encontradas (Fig. 14) correspondem ao local dos *hotspots* propostos no PLIP para as interações eletrostáticas importantes para a PPI (Fig. 13), destacadas com asteriscos e cruzes simples para as áreas de carga positiva e negativa, respectivamente. Além destas regiões destacadas, também é notável que há uma região proeminente convexa de cargas negativas no anticorpo complementada por uma região aprofundada côncava de cargas positivas na proteína E não previstas pelo PLIP, mas que apresentam complementariedade de carga e de geometria, parecendo também ter importância na PPI em questão. Esta região não prevista pelo PLIP foi destacada com cruzes duplas.

Por este motivo, um outro braço do presente estudo visou analisar a influência dos potenciais eletrostáticos de superfície nas mutações geradas pelo Rosetta. Estes resultados foram obtidos pelo colega de grupo Whendel Muniz como parte de seu trabalho de mestrado e, por esta razão, foram omitidos do presente trabalho. Entretanto, é válido mencionar a obtenção destes resultados pois eles justificam, em parte, os achados do presente trabalho, conforme discutiremos adiante na secção 4.3. Além disso, seus resultados partiram da análise da proteína E isolada (isto é, não frente aos anticorpos). Assim, para evitar computações redundantes, as proteínas de referência e suas mutantes geradas neste trabalho foram separadas de seus anticorpos e minimizadas novamente para uso neste outro braço do estudo, motivo pelo qual a caixa inferior central no fluxograma da Fig. 10 foi mantida.

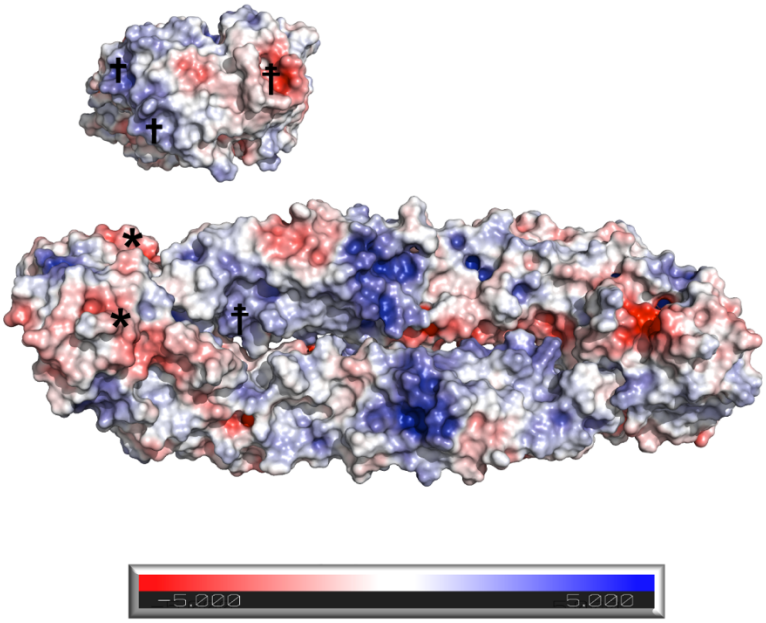


Figura 14. Estruturas de referência da proteína E (abaixo, vista de cima) com seu respectivo anticorpo humano (acima, visto de baixo), com destaque para a região da PPI de interesse eletrostático. Figuras acima de autoria própria, confeccionadas no software de visualização molecular PyMol. Os cálculos do potencial eletrostático da superfície foram realizados no plugin APBS-electrostatics, que utiliza modelo de Poisson-Boltzmann. A escala gráfica acima está representada em unidades $\boldsymbol{K_bT/e_c}$, em que $\boldsymbol{e_c}$ representa a carga elementar. PDBid: 4UTA, artigo de referência da estrutura: doi:10.1038/nature14130. Abaixo, representação da face superior da proteína E, com as regiões negativas (em vermelho) de complementariedade de carga marcadas com asteriscos. Acima, representação da face inferior do Ab, com as regiões positivas (em azul) de complementariedade de carga marcadas com cruzes. Além destas, há uma região de complementariedade de cargas e de geometria que não foi prevista pelo PLIP, destacada com cruzes duplas.

4.3 HÁ DIFERENÇAS DE ΔΔG DOS COMPLEXOS E:AB DE DENV2 ENTRE ESTRUTURAS NATIVAS E MUTADAS?

A variação máxima de ΔΔG obtida por métodos de particionamento de energia foi de 13,7 kcal/mol (Rosetta) e 5,1 kcal/mol (FoldX). A variação máxima de ΔΔG obtida no PBEE, que usa método de aprendizagem de máquina, foi de 1,9 kcal/mol.

Foi realizada a preparação e modelagem de amostras com os PDBs apresentados na Tabela 1. A preparação de amostras incluiu a limpeza de PDBs, renomeando seus resíduos (para começar a contagem no 1º resíduo) e cadeias (para garantir que a cadeia A represente a proteína E alvo e a cadeia B represente o anticorpo ligante), assim como a modelagem dos resíduos ausentes (gaps na estrutura). Os PDBs preparados estão representados na Fig. 12 E-J.

Após a preparação da amostra, realizamos os protocolos Fast Relax e Flex ddG das amostras para candidatos DENV2 visando, respectivamente, à minimização e à gênese das mutações (com cálculo simultâneo das diferenças de ΔΔG delas oriundas). Os resultados dos cálculos de ΔΔG obtidos no protocolo Flex ddG do Rosetta são mostrados na Figura 15. Nossos dados iniciais sugeriram que algumas das mutações levam a um aumento de ΔΔG de até 13,7 kcal/mol, indicando que tais mutações podem levar à evasão imunológica, ao contrário do paradigma de imunidade pós-infecção garantida para cada sorotipo.

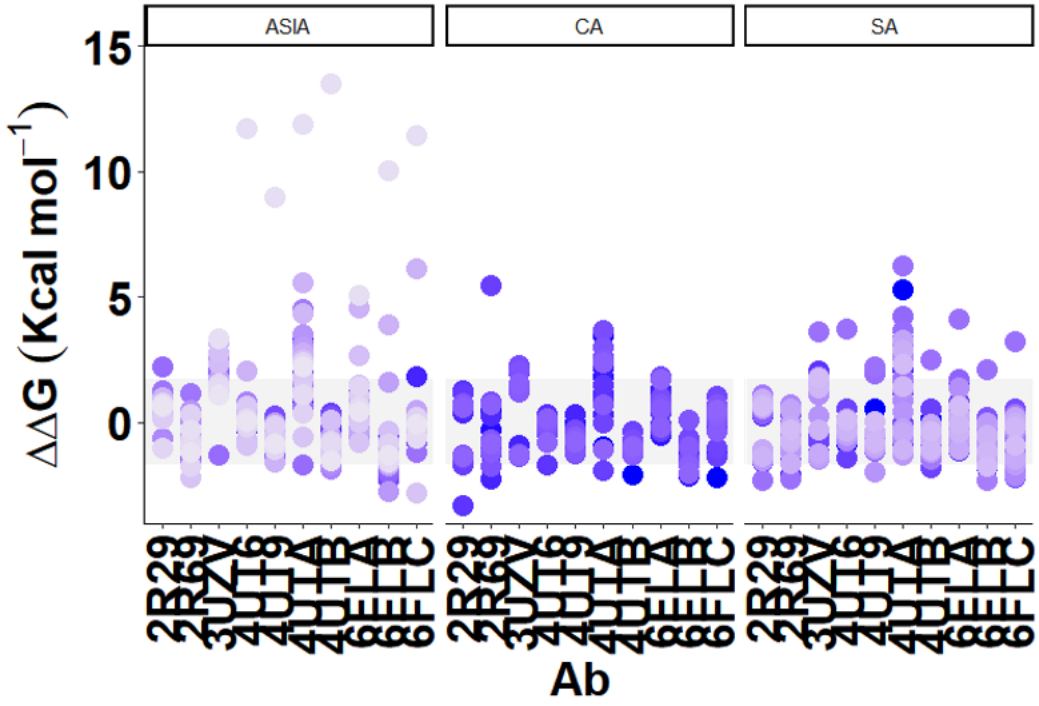


Figura 15. Valores de ΔΔG das variantes de DENV2 calculadas com o protocolo FlexddG no Rosetta. Os valores foram representados no gráfico em Kcal/mol, conforme orientado pelos autores do programa, mas, conforme discutido em secções anteriores, adotamos os valores em REU ao longo do texto. Destacamos em cinza a faixa correspondente ao erro médio padrão, de cerca de 1,62 kcal/mol, conforme referenciado no trabalho original deste protocolo. (Barlow et al., 2018) Cada ponto corresponde a uma variante diferente (complexo Emut:Ab) implementada no PDB de referência daquela coluna (no eixo das abcissas). A linha de ΔΔG nulo corresponde aos valores próprios das estruturas de PDB de referência (complexo Ewt:Ab). Assim, pontos mais acima da região acinzentada sugerem elevação do ΔΔG de interação do complexo, levando, consequentemente à evasão imune. SA, ASIA e CA (acima) correspondem às variantes da América do Sul, Ásia e América Central, respectivamente.

Para exploração adicional desta hipótese, as sequências de resíduos de aminoácidos dos mutantes foram avaliadas, evidenciando que algumas mutações aconteceram dentro ou perto dos principais epítopos (incluindo os resíduos importantes para o componente eletrostático da PPI evidenciados na Fig. 13 e na Fig. 14). Isso sugere que mutações ocorrendo nesses principais

aminoácidos (ou perto deles) podem levar à evasão imunológica, pois podem diminuir a afinidade do anticorpo. Um dos artigos mencionados na introdução (cf. secção 1) também propôs essa hipótese, e os autores de fato a provaram experimentalmente realizando pseudomutações em proteínas do envelope DENV, demonstrando que a evasão imunológica é possível neste cenário hipotético de mutação. O trabalho deles, no entanto, não realizou uma mutação baseada em evidências genômicas, mas uma mutação aleatória de ponto único (cuja ocorrência até o momento nunca foi demonstrada no DENV do tipo selvagem). (Soila et al., 2010) Com nossos dados, as mutações que foram introduzidas a partir dos dados de vigilância genômica de cepas de DENV realmente circulantes e nossos dados computacionais iniciais sugerem que a evasão imunológica é possível de ocorrer dentro dessas novas cepas circulantes. Dados de ambos os estudos corroboram a possibilidade de existência de evasão imune, embora nenhum deles tenha sido capaz de prová-la definitivamente (no trabalho supracitado, pelo fato de a mutação introduzida ter sido artificial e ainda não encontrada em cepas circulantes de DENV2 até o momento; no nosso caso, pelo fato de o escape não ter sido encontrado em métodos computacionais mais precisos).

Também testamos a hipótese principal do trabalho utilizando outro método de cálculo de ΔΔG implementado no programa FoldX. Como discutido anteriormente, este programa também utiliza método de particionamento de energia aplicado a ciclos termodinâmicos para realizar o cálculo do ΔΔG de complexos (cf. Secção 2.6.2), conforme implementado no pacote *Analyze Complex* do programa. Os resultados encontrados no FoldX se encontram representados na Fig. 16 (abaixo). Algumas das variantes analisadas por esta metodologia também corroboram a hipótese de evasão imune, embora nenhuma tenha chegado à magnitude de 13,7 kcal/mol obtida no Flex ddG do Rosetta. No FoldX, encontramos tendência ao escape imune significativa em variantes das 3 regiões apresentadas (Ásia, América Central e América do Sul). Também é notável que a tendência ao escape imune foi maior nos casos em que as mutações ocorreram nos resíduos chave anteriormente discutidos ou, pelo menos, perto destes. Estes resíduos se localizam na região do EDIII da proteína do envelope. A maioria dos anticorpos contra a proteína E se liga à região do domínio II da proteína do envelope, não na região do domínio III, porém, estes anticorpos apresentam reatividade cruzada considerável e não são neutralizantes. Embora a região do EDIII tenha reatividade com menos anticorpos do que a região do domínio II, os anticorpos contra EDIII são extremamente neutralizantes para sorotipos específicos. (Sarker; Dhama; Gupta, 2023) Assim, embora inicialmente a região do

EDIII não pareça ter nada em especial que justifique a tendência encontrada, o fato de esta região ser o principal alvo de anticorpos neutralizantes pode conferir suporte biológico a esta.

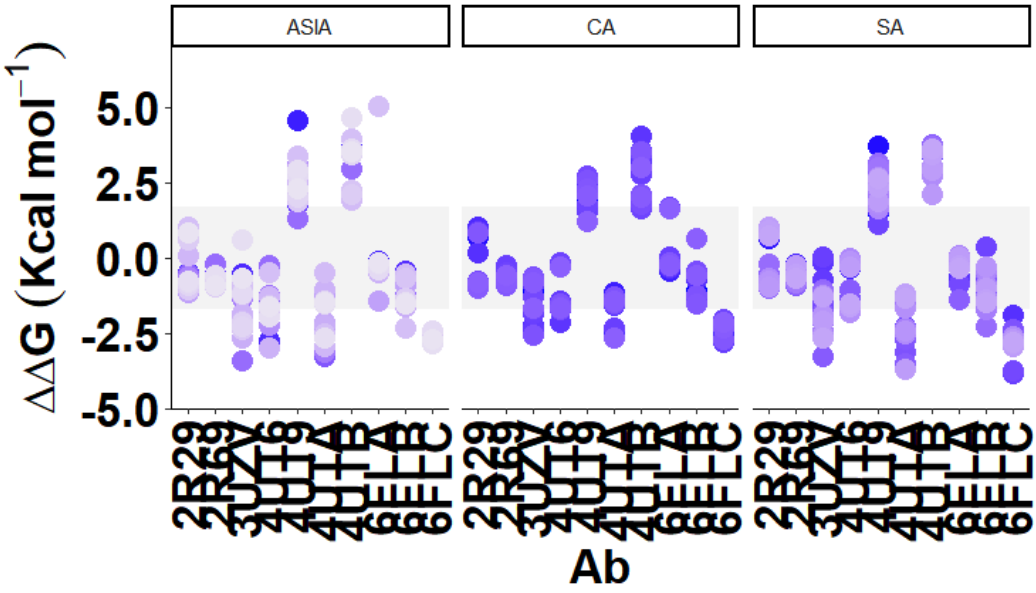


Figura 16. Valores de ΔΔG das variantes de DENV2 calculadas com o programa FoldX. Os valores foram representados no gráfico em Kcal/mol. Destacamos em cinza a faixa correspondente ao desvio padrão do método, de cerca de 1,7 kcal/mol, conforme referenciado no trabalho original deste protocolo. (Buß; Rudat; Ochsenreither, 2018) Cada ponto corresponde a uma variante diferente (complexo Emut:Ab) implementada no PDB de referência daquela coluna (no eixo das abcissas). A linha de ΔΔG nulo corresponde aos valores próprios das estruturas de PDB de referência (complexo Ewt:Ab). Assim, pontos mais acima da região acinzentada sugerem elevação do ΔΔG de interação do complexo, levando, consequentemente à evasão imune.

Além das metodologias de particionamento de energia acima, os cálculos também foram realizados por metodologia distinta. Nosso grupo desenvolveu um método baseado em aprendizado de máquina (ML) para prever o ΔΔG de interações proteína-proteína, implementado no software PBEE (Protein Binding Energy Estimator), que mostrou correlação experimental satisfatória, com valores de R (correlação de Pearson) de 0,68 (cf. Secção 2.6.3). (Chaves et al., 2025) Conforme discutido na secção 3.3, embora o Rosetta também defina correlação experimental (no seu artigo original) quando comparado com o PBEE no artigo original deste, apresentou coeficiente de Pearson de 0,08, não sendo possível traçar correlação experimental com o Rosetta nestas condições. (Barlow et al., 2018; Chaves et al., 2025) Como resultado, incorporamos o método ML do *superlearner* acima (PBEE) em nossos cálculos e comparamos os resultados obtidos no PBEE com os atuais obtidos do Flex ddG e FoldX. Todavia, os resultados obtidos utilizando este método (Fig. 17) não foram suficientes para apontar mutações que levam à evasão imunológica até o momento. A maior elevação de ΔΔG

avaliada por este método foi de 1,9 kcal/mol, estando, assim, abaixo da definição termodinâmica de evasão imune adotada (de 2,84 kcal/mol, cf. secção 1.3) no presente trabalho.

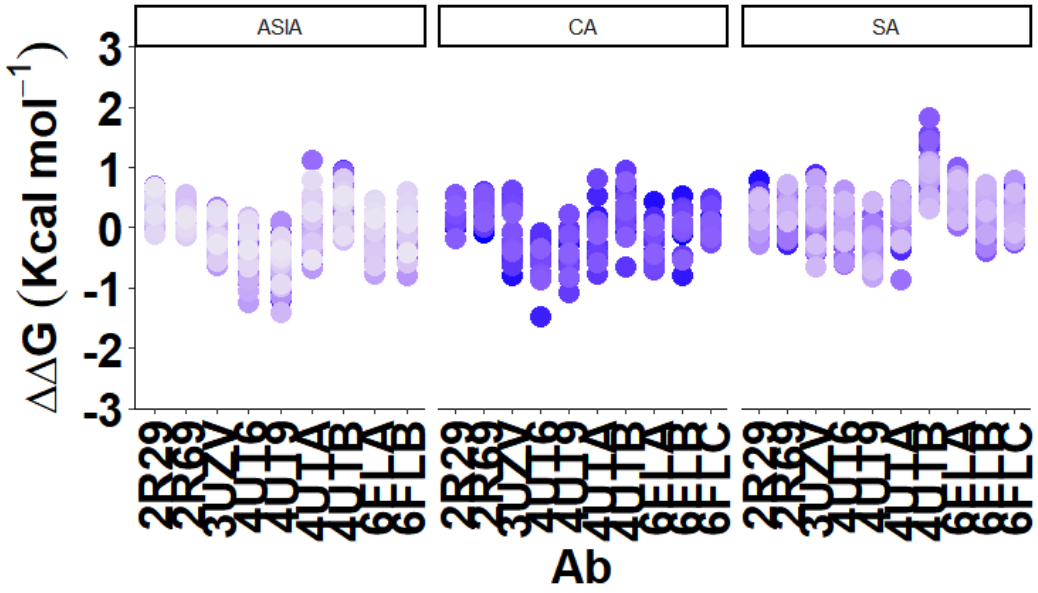


Figura 17. ΔΔG de variantes de DENV2, calculadas utilizando o software PBEE. Os valores na linha de 0 representam o ΔΔG nativo calculado para cada um dos candidatos de referência. Já os pontos plotados acima e abaixo da linha zero representam o valor de ΔΔG ganhado ou perdido com as mutações executadas (isto é, cada ponto representa a diferença de ΔΔG dos complexos E:Ab entre Ewt:Ab e Emut:Ab). ASIA representa dados coletados de variantes asiática; CA de variantes da América Central e AS de variantes da América do Sul (a maioria dos quais é oriundo do Brasil).

Para investigar estes achados de maneira mais aprofundada, classificamos os resultados destes cálculos por genótipo nas cepas brasileiras, que foram as que levaram a diferenças mais significativas nos valores de ΔΔG (Fig. 18). Nossos dados evidenciam que a tendência à evasão imune é maior em variantes correspondentes ao genótipo 3 (asiático-americano), presente no Brasil desde 1990. (Gräf et al., 2023)

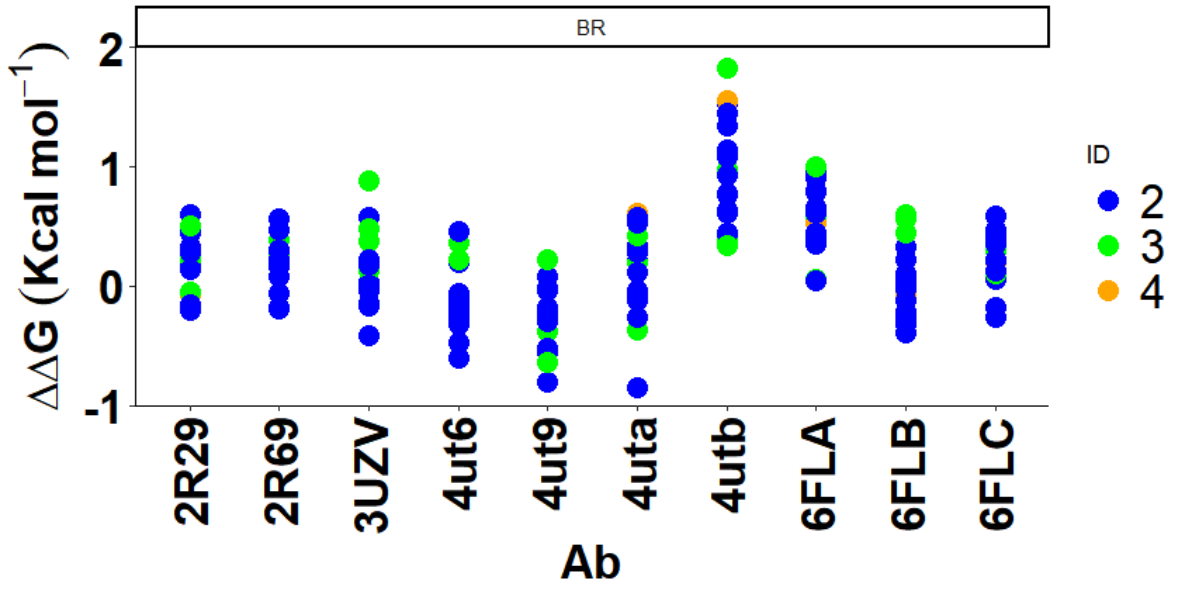


Figura 18. ΔΔG de ligação das variantes de DENV2, coloridas por genótipo, calculados utilizando o software PBEE.

4.4 DIFERENÇAS DE ΔΔG DOS COMPLEXOS E:AB DE DENV2 PODEM SER EXPLICADAS QUIMICAMENTE?

Os resultados encontrados pelas metodologias de particionamento de energia, embora corroborem a hipótese principal, não podem ser justificados quimicamente por mudanças locais nas PPIs. Diante da discordância dos resultados entre métodos, reunimos evidências e argumentos químicos, lógico-argumentativos (Popperianos e Kuhnianos) e experimentais (a partir do próprio conjunto de estruturas de referência) para avaliar o melhor método para o problema atual: todos os argumentos favorecem o uso do PBEE em relação aos demais métodos neste cenário. Assim, a hipótese nula segue confirmada, a hipótese principal, falseada e o paradigma fisiopatológico atual, embora questionado, mantido.

Ao verificar a grande diferença de resultados entre os métodos encontrados, com algum grau de concordância entre os métodos de particionamento de energia, mas com discordância total dos resultados encontrados por aprendizagem de máquina, realizou-se exame sistemático das possíveis causas para as diferenças encontradas. Inicialmente, avaliamos o sentido químico dos resultados encontrados nas metodologias de particionamento de energia. Não tomamos procedimento semelhante para o método baseado em aprendizagem de máquina por dois motivos, o primeiro, metodológico e o segundo, lógico: por não ser baseado inteiramente em métodos físicos, não é adequado discutir o racional químico dos resultados encontrados como consequência direta de um ou outro componente do hamiltoniano do sistema; por este método

ter encontrado resultados que corroborem a hipótese nula, enquanto que os outros métodos corroboraram a hipótese principal, a refutação destes resultados por si só deve implicar na valoração daqueles. A seguir, examinamos do ponto de vista lógico, sob perspectiva argumentativa dos filósofos da ciência Karl Popper e Thomas Khun, os resultados encontrados. Por fim, comparamos os resultados dos três métodos com os resultados experimentais encontrados na literatura para alguns dos complexos nativos. Ao fim destes três exames de naturezas distintas (química, lógica e experimental), poderemos chegar às conclusões definitivas desta secção.

4.4.1 Os resultados das metodologias obtidas por particionamento de energia podem ser quimicamente explicados?

Os resultados encontrados pelas metodologias de particionamento de energia, embora corroborem a hipótese principal, não podem ser justificados quimicamente por mudanças locais nas PPIs.

Para exploração adicional das diferenças entre os dois métodos, optou-se por selecionar o candidato de maior diferença (e, consequentemente, aquele com maior tendência a escape imune) obtido por cada um dos métodos e discutir as mutações apresentadas por estes individualmente. Esta análise inicialmente parece pouco relevante, mas é através dela que serão buscadas evidências e racional químico que deem suporte (ou não) aos resultados encontrados por estes métodos, permitindo interpretá-los. Este tipo de busca racional é mais difícil de ser realizada em algoritmos de ML com redes neurais com muitas camadas implícitas, pela própria natureza destas camadas "escondidas" (do inglês, *hidden layers*). (Shi et al., 2023) Porém, para os protocolos baseados em particionamento de energia, é possível tentar traçar tendências de explicações razoáveis para os resultados baseando-se no conhecimento químico dos aminoácidos e nos fundamentos teóricos destes métodos de maneira mais intuitiva. A escolha do candidato com maior tendência a escape imune também é justificável: analisar todas as 1020 estruturas detalhadamente seria infrutífero (cf. secção 4.2); além disso, ao tentar divisar padrões que justifiquem a magnitude dos resultados obtidos, parece mais apropriado tomar por referência elementos que a apresentem em maior evidência (isto é, aqueles para os quais a magnitude foi superior aos demais).

Para o protocolo Flex ddG do Rosetta, a maior diferença encontrada correspondeu às mutações implementadas da única variante do Paquistão (identificador *Pakistan_NUST-2_2023_EPI_ISL_19066415_2023-10-15*) na estrutura de referência 4UTB. Este mesmo

candidato (que apresentou variações de 13,7 kcal/mol no protocolo Flex ddG do Rosetta), apresentou variações de ΔΔG de 4,7 kcal/mol (no Fold X) e 0,26 kcal/mol (no PBEE), ilustrando novamente a discordância entre os métodos, já observada nas Figs. 14-16: ambos os métodos baseados em particionamento de energia corroboram a hipótese principal (embora o Fold X apresente geralmente valores de ΔΔG menos magnificados que os obtidos no Rosetta), enquanto que o método baseado em aprendizagem de máquina corrobora a hipótese nula. Comparando com a sequência original da estrutura de referência (PDBid 4UTB), ocorreram as seguintes mutações (com notação de amino ácidos de uma letra, apresentando o resíduo nativo (WT) a esquerda e o mutante (MUT) a direita do número da posição do resíduo) na Tabela 3 abaixo:

E71A	I164V	I322V
D98G	M260L	V324I
N103D	*T276I*	*T340M*
I141V	*D283N*	V380I
H149N	I308V	N390S

Tabela 2. Lista de mutações implementadas para variante *Pakistan_NUST-2_2023_EPI_ISL_19066415_2023-10-15* no PDB de referência 4UTB. Em itálico, destacam-se mutações em que houve mudança de classificação química do aminoácido correspondente no número do resíduo.

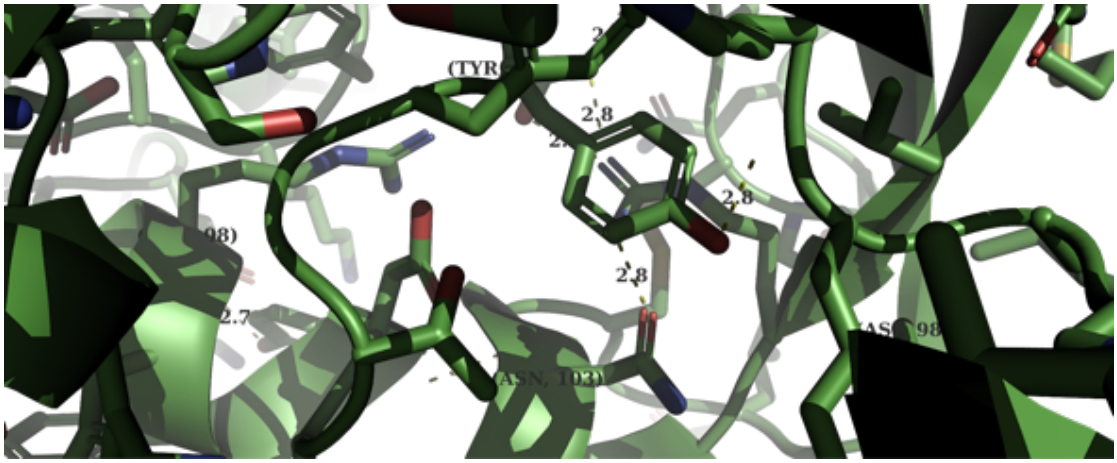
Algumas das mutações acima não representam mudança significativa esperada na PPI, tanto por não configurarem mudança de classe do resíduo de aminoácido, quanto por não estarem em regiões de epítopos previstos para a interação, a exemplo das mutações I141V e M260L. Porém, interessantemente, notamos mutações em epítopos previstos com trocas de classe do resíduo, como ocorre em E71A, D98G, N103D e T340M. Na primeira e na segunda, houve troca de resíduo com cadeia lateral negativa (E e D, respectivamente) por resíduo com cadeia lateral apolar (A e G, respectivamente). Na terceira, houve troca de resíduo polar (N) por resíduo de cadeia lateral de carga negativa (D). Na quarta, houve troca de resíduo polar (T) por apolar (M).

Estas mutações podem explicar os resultados encontrados, já que, em conjunto, são suficientes para elevar o ΔΔG da interação do complexo nativo Ewt:Ab a valores positivos no complexo mutante Emut:Ab. Entretanto, ao examinar em detalhes estas mutações e suas influências prováveis nas interações E:Ab utilizando software de visualização molecular (PyMol), nota-se que o cenário é mais complexo. Algumas das mutações apresentadas se encontram distantes na PPI (a exemplo do resíduo 340, que dista aproximadamente 30 Å do anticorpo), sendo improvável que mutações nestes exerçam grande influência na energia livre

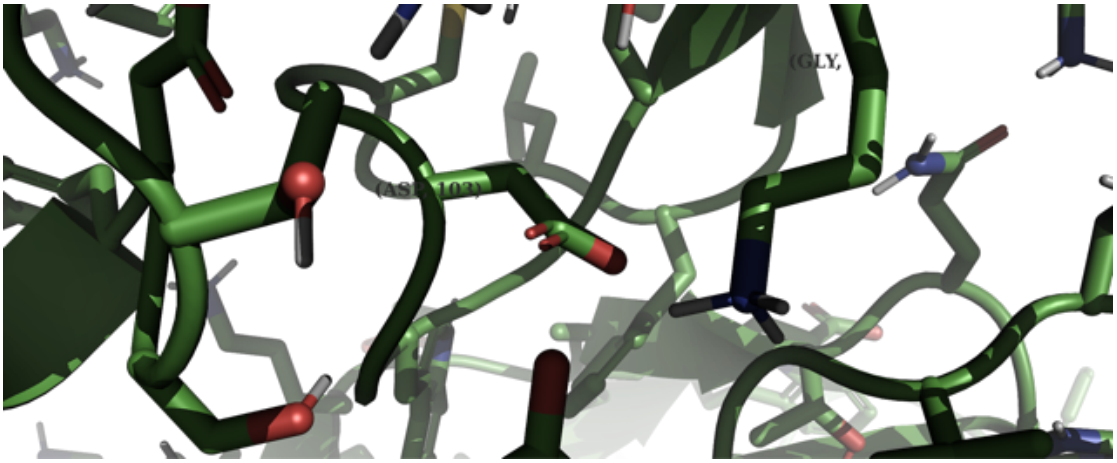
de afinidade e, consequentemente, na avidez, já que esta tem correlação com a constante de dissociação. (Tsuji et al., 2014) Não encontramos na literatura valor de referência específico para definir quando as PPI ocorrem ou não (em geral, cada campo de força tem suas próprias definições específicas de corte para cada parâmetro e equação modelada, mas não há consenso na literatura). Entretanto, as forças moleculares descritas na secção 2.2.1, em sua maioria, decrescem com a distância. (Allen; Tildesley, 1989) Embora não haja raio de corte específico de distância para definir as interações na literatura, podemos tomar emprestada a definição de contatos do *CASP* (avaliação crítica de predição estrutural proteica, do inglês, *Critical Assessment of protein Structure Prediction*), que usa distância de corte de 8 Å, para afirmar que os 30 Å supracitados estão muito além deste valor. Embora a definição de contatos seja diferente da definição de interações (a primeira é utilizada para avaliar o número de contatos dentro de uma proteína, correlacionando-se com o seu grau de compactação no enovelamento), tomamo-la por empréstimo pela falta de definição de distância de corte para definir uma PPI e pelo racional químico semelhante subjacente a ambas as medidas: em suma, tratam de uma definição de raio geométrico de corte a partir do qual há interações significativas entre aminoácidos. Já o resíduo 98 está parcialmente enterrado na proteína e, por isso, parece exercer maior influência na interação de cada subunidade da proteína E com ela mesma do que com o Ab propriamente dito. Assim, notamos que apenas duas das mutações elencadas na Tabela 3 parecem ter papel fundamental na grande diferença de ΔΔG observado nesta variante, os resíduos 71 e 103 (mutações E71A e N103D). Estas mutações foram representadas nas Figuras 19A-D. O primeiro resíduo (E71, na estrutura nativa; Fig. 19C) é estabilizado por interações eletrostáticas com oo resíduo R73 da própria proteína E. Já o segundo (N103, na estrutura nativa; Fig. 19A), é estabilizado por contatos polares com o resíduo Y110 no Ab. Assim, a mutação destes para resíduos com outra classificação implica numa perda importante destas interações (Figs. 19B e D), justificando em parte o aumento significativo de ΔΔG observado nesta variante. Porém, não é esperando que apenas estas duas mutações sejam suficientes para provocar elevações de ΔΔG tão expressivas que ultrapassem 10 kcal/mol em relação à estrutura nativa. Assim, pelos argumentos acima apresentamos, concluímos que modificações locais nas PPI não justificam a magnitude da elevação do ΔΔG encontradas pelo protocolo Rosetta Flex ddG.

Avaliamos também a similaridade estrutural global entre a proteína E nativa e sua respectiva variante. Tal similaridade foi avaliada calculando o RMSD (desvio quadrático médio, do inglês, *Root Mean Square Deviation*) entre as estruturas terciárias das duas proteínas

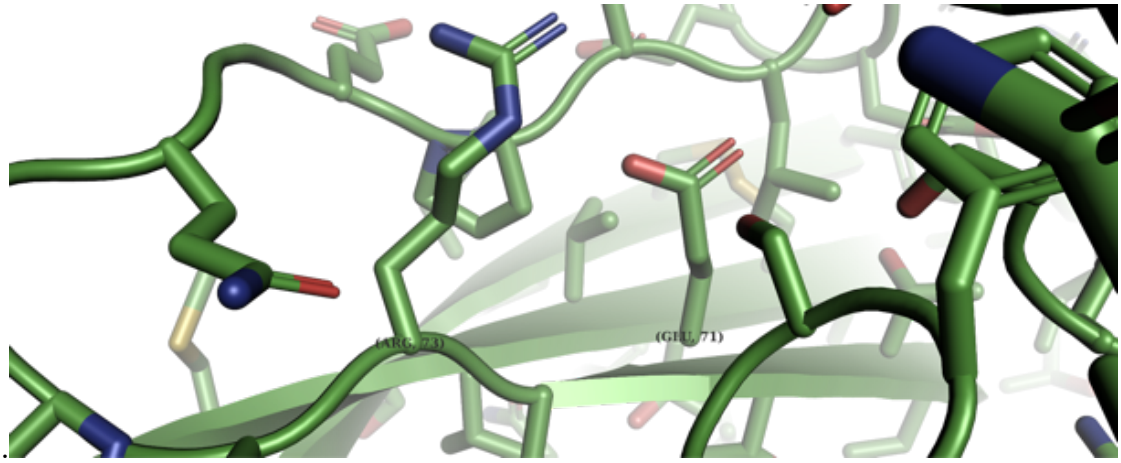
do envelope. Valores baixos de RMSD (próximos de 1 Å) correspondem a alta similaridade estrutural, e valores elevados correspondem a baixa similaridade estrutural. O RMSD encontrado para esta variante, em relação à estrutura nativa de referência, foi de 3,87 Å. Inferimos, portanto, que embora tenham ocorrido poucas mutações na proteína, a estrutura global do complexo teve variações consideráveis induzidas por estas mutações. Assim, apesar de as modificações locais na PPI não justificarem a magnitude da elevação de ΔΔG encontrada, é possível que as modificações estruturais globais (cuja contribuição específica no escore de energia do Rosetta não pode ser mensurada diretamente, já que todos os componentes da função de energia dependem das coordenadas atômicas do sistema) tenham justificado a elevação observada neste programa. Esta afirmação, porém, carece de comprovação quantitativa, pelo motivo supracitado, e maior confiança seria atribuída ao método se os resultados encontrados fossem justificados em conjunto pelas modificações estruturais locais (na PPI) e pelas globais, não apenas pela última, que não é mensurável. Assim, conclui-se que mudanças locais nas PPI não justificam a magnitude do resultado obtido pelo método Flex ddG, mas mudanças globais conformacionais poderiam explicá-lo, embora essa proposição seja inferencial e não possa ser mensurada.

A. 

B.



C.



D.

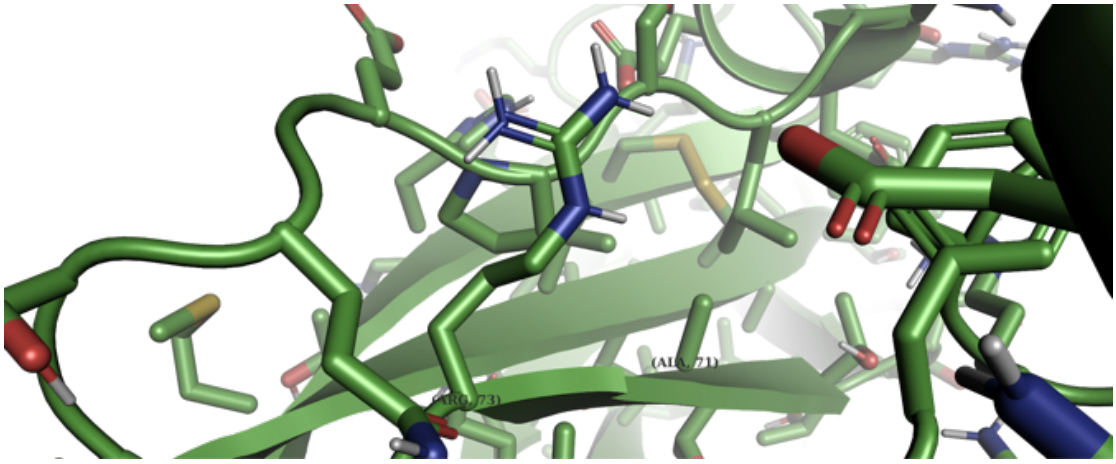


Figura 19. Representação esquemática de modificações na estrutura de referência 4UTB implementadas a partir da variante com maior variação de ΔΔG estimada pelo protocolo Flex ddG. Variante de identificador *Pakistan_NUST-2_2023_EPI_ISL_19066415_2023-10-15*, com valores de ΔΔG calculados de 13,7 kcal/mol, 4,7 kcal/mol e 0,26 kcal/mol obtidos nos métodos Rosetta Flex ddG, Fold X e PBEE, respectivamente. As figuras A e B representam a mutação N103D, com o resíduo nativo (N103) na figura A e o resíduo mutado (103D) na figura B. Além da troca de classe do aminoácido, houve perda de contato polar com o resíduo Y100, que passou a estar mais enterrado. Nas figuras C e D está representada a mutação E71A, com o resíduo nativo (E71) na figura C e o

resíduo mutado (71A) na figura D. Além da troca de classe do aminoácido, houve perda de ponte salina intracadeia com R73 após esta mutação.

Já no caso do cálculo utilizando o programa FoldX, a maior diferença de ΔΔG foi obtida para a variante *Indonesia_KS-NIHRD-WD106_2023_EPI_ISL_18462816_2023-09-23*, também da Ásia, em relação à estrutura de referência 6FLA. Este mesmo candidato (que apresentou variações de 5,1 kcal/mol no protocolo Fold X), apresentou variações de ΔΔG de 0,4 kcal/mol (no Flex ddG) e 0,19 kcal/mol (no PBEE), ilustrando discordância completa deste resultado com os demais métodos. Representamos as mutações esquematicamente na Tabela 4, de maneira análoga à Tabela 3.

E383D	*K394Q*
N390S	*G395S*
F392L	*S396P*
K393T	*S397A*

Tabela 3. Lista de mutações implementadas para variante *Indonesia_KS-NIHRD-WD106_2023_EPI_ISL_18462816_2023-09-23* no PDB de referência 6FLA. Em itálico, destacam-se mutações em que houve mudança de classe do resíduo.

Neste caso, cinco mutações (K393T, K394Q, G395S, S396P, S397A) representaram mudança de classe do resíduo. Entretanto, todas as mutações representam resíduos da extremidade C-terminal da proteína (Fig. 20). Tais resíduos muito provavelmente tem alta flexibilidade estrutural, que pode ser inferida pelos seus fatores B superiores a 50 Å², o que argumenta contra sua participação nas interações com o anticorpo. Em estruturas cristalográficas, as amplitudes de oscilação dos átomos ao redor de sua posição de equilíbrio são monitorizadas pelos fatores B, que se relacionam com a amplitude quadrática média de oscilação (u) dos átomos ao redor de sua posição de equilíbrio pela equação $B = 8\pi^2 u^2$. (Carugo, 2018) Valores muito altos podem indicar, portanto, maior flexibilidade estrutural da região (que poderá ser comprovada experimentalmente verificando mapas de densidade eletrônica ou computacionalmente realizando simulações de MD da proteína no solvente). Valores elevados de fator B são esperados em regiões terminais proteicas, que costumeiramente apresentam elevada flexibilidade estrutural associada. Outro fator problemático para estas mutações é o fato de elas, assim como no caso anterior, do Rosetta, estarem distantes da PPI. O resíduo mais próximo da PPI (de número 393) dista 22 Å do resíduo mais próximo do anticorpo. Novamente, reitera-se que não há valor de corte definido na literatura para definir a importância

de resíduos nas PPI, mas, como discutido no caso anterior, os 22 Å citados estão muito acima do corte de definição de contatos, de 8 Å. Portanto, mudanças locais nas PPI não justificam a magnitude dos resultados encontrados no Fold X.

Analisando as mudanças conformacionais globais para este caso, notamos que o RMSD é de 1,247 Å para a estrutura mutada em relação à estrutura de referência. Como os valores de RMSD se encontram próximos de 1 Å, é possível afirmar que as estruturas de referência e mutada tem elevada similaridade estrutural. Portanto, mudanças conformacionais globais também não justificam a magnitude dos resultados encontrados por este método.

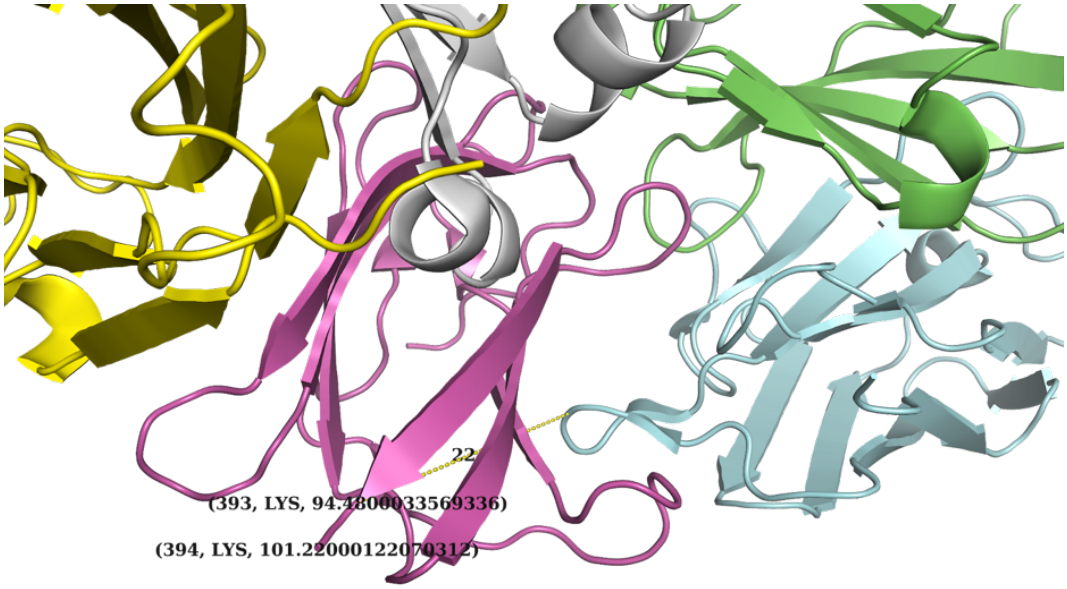


Figura 20. Representação esquemática de modificações na estrutura de referência 6FLA implementadas a partir da variante com maior variação de ΔΔG estimada pelo protocolo Fold X. Variante de identificador *Indonesia_KS-NIHRD-WD106_2023_EPI_ISL_18462816_2023-09-23*, com valores de ΔΔG calculados de 0,4 kcal/mol, 5,1 kcal/mol e 0,19 kcal/mol obtidos nos métodos Rosetta Flex ddG, Fold X e PBEE, respectivamente. As cadeias do anticorpo estão representadas a direita em ciano e verde. As cadeias da proteína do envelope estão representadas a esquerda em amarelo e magenta. Os dois objetos inferiores na figura incluem o nome dos dois resíduos C-terminais mais próximos da PPI, com seus fatores B correspondentes. O objeto superior marca a distância de 22 Å entre tais resíduos e o anticorpo. Esta é a menor distância encontrada entre os resíduos nos quais houve mutação e a PPI, sugerindo que as mutações encontradas não influenciem na dinâmica da PPI.

Portanto, os resultados encontrados pelos métodos Fold X e Flex ddG não podem ser justificados pelas modificações locais nas PPI. No caso do protocolo Flex ddG, é possível que mudanças conformacionais globais induzidas pelas mutações justifiquem os valores de ΔΔG encontrados, mas não é possível mensurar sua influência específica. Os argumentos químicos aqui apresentados, portanto, vão de encontro aos resultados obtidos utilizando métodos

baseados em particionamento de energia. Como estes corroboraram a hipótese principal do trabalho, os argumentos químicos apresentados nesta secção corroboram a hipótese nula. Assim, ainda não há resposta acerca da escolha do método que responda com maior confiança às perguntas condutoras do presente estudo, mas os argumentos químicos não foram favoráveis para os métodos Flex ddG do Rosetta e Fold X. Na próxima secção, examinam-se racionais lógico-argumentativos para esta escolha e, a partir destes, aprofunda-se a arguição química desta secção, correlacionando os dados aqui apresentados com possíveis justificativas metodológicas (com destaque para a forma funcional dos campos de força). Na secção seguinte, serão reunidos argumentos experimentais e tecidas conclusões acerca da pergunta conduta de que trata a secção 4.4.

4.4.2 A comparação dos resultados dos três métodos pode ser logicamente explicada?

Do ponto de vista Popperiano e Kuhniano, argumentos em favor da hipótese nula são mais fortes que os argumentos contra esta.

Não encontrando justificativa disciplinar (neste caso, química) que satisfaça o cientista de dada disciplina frente aos seus resultados, o raciocínio hipotético dedutivo baseado em modelo Popperiano (Popper, 2013) deve conduzir a investigação para as seguintes proposições: (1) a justificativa química para os resultados pode ser mais obscura, profunda ou complexa do que se havia pensado; (2) os paradigmas químicos utilizados para o racional (no nosso caso, os princípios básicos da físico-química de proteínas) podem estar errados; (3) os resultados encontrados podem estar errados. A primeira proposição, em estando correta, foge à percepção dos autores e, portanto, é logicamente impossível de ser discutida. A segunda proposição é extremamente improvável, já que o conjunto de evidências para os princípios básicos aqui discutido é extremamente vasto, consagrado e suportado por toda a literatura até então (e colocá-la em xeque também seria logicamente implausível, já que a defesa e argumentação de todo o trabalho cairia por terra). Já a terceira proposição é mais plausível e razoável. Ao assumi-la como verdadeira, há que se investigar sistematicamente a possível ocorrência de erros nos métodos propriamente ditos (primeiro caso), ou nos procedimentos tomados para sua execução (segundo caso). Este segundo caso parece improvável, pois a metodologia do presente trabalho foi revista em todas as suas etapas pelos professores orientadores e pelo próprio autor e está compatível com os procedimentos metodológicos encontrados na literatura e com as documentações originais dos programas utilizados. Resta a análise do primeiro caso. A única

resposta logicamente razoável é, portanto, o questionamento da validade dos métodos aplicados para responder ao problema proposto.

Nenhum dos métodos de particionamento de energia adotados parece responder satisfatoriamente ao problema com racional químico que o embase e justifique. Na literatura recente, alguns autores apontam para limitações nestes métodos, incluindo a amostragem insatisfatória das conformações na região das mutações e a eficiência conformacional baixa, além do problema da tendência destas abordagens em favorecer energeticamente a presença de amino ácidos hidrofóbicos, especialmente em áreas de PPI. (Schymkowitz et al., 2005) Esta última provavelmente decorre dos pesos empíricos dados a cada um dos componentes da função de energia avaliada pelo programa. Assim, mutações que troquem aminoácidos apolares por outras classes podem ter pênalti energético elevado no escore de energia adotado pelo programa, levando a valores de ΔΔG mais elevados do que seria esperado em dados experimentais. Estes pesos, entretanto, fazem sentido para compensar a falta de exploração amostral conformacional levantada anteriormente, configurando uma espécie de *tradeoff* energético: frente ao grande custo computacional da exploração adequada do espaço amostral, é recurso inteligente favorecer a manutenção de resíduos de conferem maior estabilidade nas estruturas nativas (como é o caso de resíduos apolares em PPI ou enterrados). Esse *tradeoff* energético, no entanto, pode ser sustentável para a implementação de mutações pontuais em complexos de menor flexibilidade conformacional, mas não foi satisfatório para os cálculos deste trabalho, já que envolvem as complexas interações antígeno-anticorpo e a implementação de muitas mutações simultaneamente.

Além da já mencionada motivação lógica (Popperiana) para continuar tentando encontrar evidências contra hipótese nula por outros métodos, há outra motivação cognitivo-social (e Kuhniana) para fazê-lo (Kuhn, 2018): a hipótese nula está alinhada com o já discutido paradigma de um sorotipo-uma infecção; tal proposição é de tamanha força na literatura atual (além das implicações sociais e epidemiológicas em falseá-la, já discutidas na introdução) que, para derrubá-la, será necessário coletar um conjunto de evidências por métodos mais rigorosos e robustos. Assim, a colocamos à prova utilizando métodos de previsão de ΔΔG com melhor correlação experimental, pois as metodologias anteriores, embora robustas, dependem de uma função de pontuação teórica que trata o cálculo de ΔΔG como um problema de partição, com termos distintos representando cada propriedade química de importância e o ΔΔG sendo calculado como a soma dos termos individuais. Apesar dos métodos anteriores terem sua validade amplamente testada e constituírem boa avaliação por comparação, que é o que

realizamos confrontando os valores de ΔΔG dos complexos Ewt:Ab contra Emut:Ab, eles não mantêm correlação experimental suficientemente forte para derrubar um paradigma clássico com décadas de evidências teóricas e experimentais que o suportam. Assim, mesmo que sua validade seja vasta, para este tipo de pergunta específica, é necessário recorrer a ferramentas que tenham correlação experimental melhor. Voltaremos a examinar a correlação experimental em mais detalhes na secção 4.4.3.

Notamos, interessantemente, que os resultados de maior diferença de ΔΔG (isto é, aqueles em que foi sugerido escape imune), pelos cálculos dos protocolos Flex ddG e FoldX foram aqueles contendo anticorpos humanos, quando representando mutações introduzidas por variantes genéticas muito distintas (geográfica, filogenética e quimicamente, isto é, na sequência de resíduos) da cepa original. Como discutido na secção 4.2, estes foram os anticorpos de modo de ligação predominantemente perpendicular, nos quais as interações eletrostáticas de superfície são muito significativas para a interação. Nas mutações de maior tendência à evasão imune, houve substituição de resíduos com troca de carga. Como a função de energia do Rosetta tem pesos significativos para interações eletrostáticas (o que é esperado, já que trocas de carga de um único resíduo importante para a interação, na experiência do nosso grupo, podem levar a diferenças de até 3 kJ/mol no cálculo do ΔΔG), é provável que o escape imunológico encontrado no protocolo do Flex ddG, mas não corroborado no PBEE (em que os pesos da análise de interface do Rosetta são ajustados por modelo de ML), se deva a estas diferenças no potencial eletrostático de superfície. Este dado é corroborado por achados de um outro braço deste mesmo estudo realizado por outro membro do grupo (também já mencionado na secção 4.2), em que as diferenças no potencial eletrostático de superfície da proteína E viral foram comparadas ao longo do tempo. Apesar de termos encontrado esta correlação que pode ajudar a explicar os resultados acima, ao tentar esclarecer se a correlação entre as diferenças de ΔΔG e a variação no potencial eletrostático de superfície normalizado pela área de superfície exposta ao solvente (parâmetro chamado de q/SASA) é ubíqua no nosso conjunto de dados, encontramos coeficiente de Pearson de apenas 0,32. Isto indica que, embora esta interação seja importante para alguns grupos de anticorpos específicos (em especial nos resultados encontrados com anticorpos humanos), uma correlação geral e direta entre estes parâmetros não é adequada, ainda que sua contribuição no escore de energia seja inegável.

Estes achados sugerem que é muito provável que as respostas imunes humorais humanas à DENV sejam mais adequadas para cada genótipo específico. Estes argumentos alimentam a discussão sobre se deveríamos tratar o modelo patogênico de “um sorotipo – uma

infecção" como absoluto. A evasão imune por sorotipo ainda não foi comprovada, como discutido, e os argumentos levantados até o momento corroboram a hipótese nula, que esta alinhada com o paradigma clássico de Halsted. Porém, com a evolução e disseminação de técnicas genômicas, é possível que em breve comprovemos que, como ocorre em outras doenças virais, seja o genótipo, e não necessariamente o sorotipo, que elicite respostas imunes humorais de longa duração. A percepção desta possibilidade é importante em termos de políticas públicas de saúde em vigilância epidemiológica, em especial no desenvolvimento de novas estratégias vacinais, terapêuticas e diagnósticas para DENV.

Assim, muito embora ambos os métodos de Flex ddG e FoldX falem contra a hipótese nula do trabalho (e, portanto, poderiam ser considerados exitosos do ponto de vista da defesa da hipótese principal), o racional lógico-argumentativo e o racional químico da secção anterior (e aprofundado nesta), tomando como exemplo seus casos mais extremos, não suportam a magnitude destes achados. Por utilizarem campos de força distintos, não é esperado que haja concordância absoluta nos resultados (isto é, nas variantes específicas e nos seus respectivos valores de ΔΔG), o que reforça ainda mais a importância da tendência semelhante encontrada nestes (ambos encontraram maiores diferenças nas variantes asiáticas). Esta tendência semelhante nos dois programas também foi encontrada em trabalho anterior do grupo cuja metodologia foi semelhante à do presente trabalho, argumento a favor da robustez destes métodos. (Ferraz et al., 2021) Porém, mesmo que estes programas tenham resultados validados na literatura, e que suas tendências tenham sido semelhantes, é possível que a magnitude das diferenças de interação destes tenha sido engrandecida, presumivelmente pela própria forma funcional dos seus campos de força, conforme sugerido pelo racional arguido. Por fim, o exame argumentativo sistemático dos pontos de vista químico e lógico corroboram a hipótese nula, não a hipótese principal, por irem de encontro direto aos resultados que suportariam esta. Na próxima secção, serão examinados argumentos experimentais e, considerando o conjunto destes argumentos (arrazoado), os resultados apresentados serão finalmente julgados.

4.4.3 Frente a evidências experimentais, qual método estima o ΔΔG com melhor acurácia?

Em análise comparativa, o método PBEE apresentou melhor exatidão e precisão nas estimativas do ΔΔG do que os demais métodos para o conjunto de dados experimentais disponível.

Pelos motivos anteriormente expostos, os resultados obtidos com o PBEE, que corroboram a hipótese nula, nos pareceram mais confiáveis do que os resultados obtidos pelos

métodos de particionamento de energia. Nas secções anteriores apresentamos argumentos lógicos para tal afirmação, bem como argumentos químicos (ou, sendo mais específico, a falta de justificativa química dos métodos baseados em particionamento de energia, que diminui sua credibilidade para o nosso conjunto de dados). Embora os métodos tenham validação ampla na literatura, conforme já discutido nesta secção e nos procedimentos metodológicos, é importante saber se os resultados apresentados nos seus trabalhos fundadores são extrapoláveis para responder às perguntas condutoras do presente estudo. Embora todos os programas defendam correlação experimental, com coeficiente de Pearson acima de 0,5 relatado em seus trabalhos fundadores, já apresentamos anteriormente um cenário em que o programa Rosetta apresentou desempenho bastante diferente do relatado em seu trabalho original quando confrontado com o conjunto de dados para os quais o PBEE apresentou correlação experimental adequada. Cabe, portanto, retomar o cenário previamente apresentado e introduzir outros cenários semelhantes encontrados na literatura. Os próprios autores dos métodos Flex ddG e Fold X em seus trabalhos originais reportarem coeficientes de Pearson acima de 0,50 (R = 0,63 para o Flex ddG e R = 0,54 para o Fold X), sugerindo que exista correlação experimental dos resultados. (Barlow et al., 2018; Schymkowitz et al., 2005) Na literatura, porém, os valores experimentais destes programas variam amplamente. Para o Fold X, foram encontrados coeficientes de Pearson de 0,29 a 0,73, contra valores de R = 0,26 a 0,54 para o Flex ddG. (Buß; Rudat; Ochsenreither, 2018; Kumar et al., 2017) Também fora comentado previamente (cf. secção 3.4) o fato de o coeficiente de Pearson do protocolo Flex ddG ter sido muito inferior (R=0,08) quando submetido ao conjunto de dados de treino utilizado na publicação original do PBEE. (Chaves et al., 2025)

Nestas situações, é muito difícil definir qual método será mais adequado, pois todos eles apresentaram correlação experimental adequada em seus trabalhos fundadores, mas é notório que o desempenho deles varia conforme o conjunto de dados utilizado. A extrapolação de um programa investigativo testado em um cenário para cenários distintos deste é referida como robustez quotidianamente nas discussões científicas. O termo inclusive foi oficialmente adotado em algumas referências de aprendizagem de máquina, com significado semelhante, embora mais específico, conforme previamente explicado (cf. secção 2.6.3). Nos parece tentador afirmar que um programa é mais robusto que o outro e, portanto, mais confiável. Porém, este tipo de afirmação, embora possa ser verdadeira sob certa medida, pode trazer reducionismo à análise científica cautelosa. Isto é, a robustez, se interpretada isoladamente, pode levar a conclusões científicas inapropriadas. Um programa pode ser vastamente robusto para um

conjunto de problemas muito grande, mas ainda assim não ser aplicável ao seu problema em questão, especificamente. Assim, não cabe apenas utilizar o programa mais robusto e confiar cegamente que seus resultados levem à verdade (a própria ideia de verdades absolutas em ciência é extremamente problemática, pois é a busca pela verdade que caracteriza o método, não a sua obtenção ou logro). No caso do problema aqui apresentado, isto fica muito claro, pois os três programas têm robustez defendida e argumentada em literatura, mas os resultados de um deles são diametralmente discordantes dos demais. Isto é ainda mais significativo ao se recordar que nenhum dos programas apresentados teve seu conjunto de dados testado especificamente para anticorpos. (Barlow et al., 2018; Buß; Rudat; Ochsenreither, 2018; Chaves et al., 2025)

Assim, mesmo que nossas discussões prévias apontem que, para o problema em questão, é mais provável que o resultado do PBEE seja aplicável, mas não o obtido nos demais, o cientista deve colecionar exaustivamente o máximo possível de evidências que corroborem a hipótese nula. É esta busca incansável que qualifica o trabalho científico idôneo, para Karl Popper, tendo-se como finalidade tentar comprovar a hipótese nula com todos os recursos possíveis e, só ao coletar múltiplas evidências de que esta não se sustenta, defender a hipótese principal. Assim, partimos em outra frente argumentativa para definir qual programa teria melhor comportamento frente ao nosso conjunto de dados específico. Como algumas das estruturas nativas obtidas diretamente do PDB (mais especificamente, as estruturas de PDBid 3UZV, 6FLA, 6FLB, 6FLC) tinham seus valores de $\Delta\Delta G$ de ligação determinados na literatura, utilizamos estas estruturas para comparar o desempenho dos três métodos na resposta ao nosso problema. (Cockburn et al., 2012; Renner et al., 2018) Nos seus trabalhos de publicação das estruturas, os autores calcularam a constante de dissociação (ou as constantes cinéticas no sentido direto e inverso). A partir destas métricas, e das equações da seção 1.3, foi possível calcular o valor experimental do $\Delta\Delta G$ e, assim, comparar os valores absolutos obtidos por estes métodos. Como não havia etapa de mutação a ser realizada, o protocolo Flex ddG foi substituído por um cálculo do escore de energia do Rosetta da estrutura. Nos demais métodos, nenhuma mudança adicional foi necessária, pois, como visto na secção de procedimentos metodológicos, estes aceitam os PDBs diretamente como *input*, sem necessidade de fornecer informações sobre mutações (cf. secção 3.3). Em condições ideais, caso houvesse quantidade suficientemente volumosa de dados experimentais resolvidos para os complexos deste estudo, seria simples comparar a correlação experimental dos métodos distintos e definir estatisticamente qual método melhor se aplicaria ao problema em questão. Porém, nas condições reais expostas, o

procedimento adotado não tem poder estatístico para definir, sozinho, qual dos métodos utilizados melhor se aplica para este tipo de investigação, pois o número de estruturas com resultados experimentais disponíveis na literatura é diminuto para uma análise estatística precisa. Porém, fazê-lo pode representar mais uma evidência que advogue a favor de um ou outro método e a coleção desta com as evidências previamente defendidas nas secções 4.4.1 e 4.4.2, se concordantes, poderá guiar a escolha do método. Representamos os resultados experimentais e os resultados obtidos nas três metodologias distintas na Tabela 2 abaixo. Como informado previamente, o número de dados é muito pequeno para uma análise estatística robusta e para definir, com poder estatístico adequado, o método mais aplicável para o problema em questão. Entretanto, alguns parâmetros estatísticos simples podem ser úteis para propor tendências e coletar evidências que orientem a escolha.

PDBid	**Kd experimental (nM)**	**Valores de ΔΔG**			
		Experimental	**Flex ddG**	**FoldX**	**PBEE**
3UZV	0.43	-13	-26.584	-5.849	-10.081
6FLA	0.92	-13	-141.248	-12.255	-10.354
6FLB	0.92	-13	-142.202	-11.094	-9.273
6FLC	160	-9.64	-142.389	-4.180	-8.300
		MAE	100.94	3.82	2.66
		RMSE	201.89	7.631	5.32
		SD	49.96	3.41	0.80

Tabela 4. Comparação entre a energia livre de Gibbs (ΔΔG) das estruturas dos complexos E:Ab nativos obtidos do PDB. A primeira coluna mostra o PDBid de cada candidato nativo, enquanto a segunda coluna apresenta os valores experimentais de Kd (em unidades de nanomolar) reportado na literatura, isto é, obtido dos trabalhos originais de publicação das estruturas. (Cockburn et al., 2012; Renner et al., 2018) A coluna seguinte representa o valor de ΔΔG determinado experimentalmente, obtido da fórmula ΔΔG = RT ln Kd apresentada na Introducção (cf. secção 1.3). As demais colunas apresentam os valores de estimativas de ΔΔG calculadas pelos métodos computacionais do presente trabalho, isto é, Rosetta, FoldX e PBEE. Os dois primeiros utilizam métodos de embasamento físico, baseados no particionamento de energia. O último utiliza método baseado em aprendizagem de máquina. As últimas três linhas da tabela apresentam os valores de erro médio absoluto em relação aos dados experimentais (MAE, do inglês, *mean alignment error*), o erro quadrático médio (RMSE, do inglês, *root mean square error*), e o desvio padrão (do inglês, *standar deviation*) de cada método computacional.

Fica evidente na tabela acima que os valores obtidos pelo programa Rosetta discordam completamente dos dados experimentais para este conjunto de dados, isto é, não apresentam exatidão adequada. Além disso, os valores também apresentam pouca coerência interna: as estruturas 3UZV, 6FLA e 6FLB apresentam resultados experimentais de ΔΔG iguais

(considerando as aproximações para o número correto de algarismos significativos), porém os resultados do escore de energia do Rosetta variam de uma ordem de grandeza da primeira estrutura para as demais. Assim, o seu desempenho também não foi preciso para o conjunto de dados apresentado. É interessante notar que as funções de energia do Rosetta inicialmente não se propunham a ter significado físico real, como discutido previamente (cf. Secção 2.5.4). As funções de energia do Rosetta eram dadas em unidades teóricas conhecidas como unidades de energia do Rosetta (REU, do inglês, *Rosetta Energy Units*), sem correspondência física. Nas funções de energia mais recentes, pelo uso extensivo de dados experimentais no cálculo estatístico da maioria dos componentes, os autores do programa propõem que se adote a função de energia em kcal/mol diretamente (embora alertem que o significado físico não seja absoluto). (Alford et al., 2017) Porém, como discutido previamente, a experiência prévia do grupo não permite que tal extrapolação seja sustentada na nossa prática rotineira. Assim, embora a evidência publicada sugira a adoção das unidades em kcal/mol (razão pela qual a adotamos neste trabalho), a evidência heurística do grupo fala em contrário, como exemplificado na Tabela 2.

Em relação ao desempenho do *FoldX*, algumas tendências interessantes podem ser notadas: a precisão do método foi aquém do esperado, também com previsões errando em uma ordem de grandeza para estruturas que deveriam apresentar valores semelhantes (isto é, para as já citadas estruturas de 3UZV, 6FLA, 6FLB). Embora estes efeitos não sejam de magnitude tão esdrúxula quanto a observada para o Rosetta (com RMSE de 201,89 para este e 7,631 para aquele), o programa ainda tem resultados imprecisos, com variações de mesma ordem de grandeza. Porém, é possível definir que, para o conjunto de dados apresentado neste trabalho, sua exatidão foi consideravelmente melhor que a daquele, como evidenciado pelo MAE de 3,82 deste (contra 100,94 daquele). É também interessante notar que a função de energia do Rosetta magnifica muito mais os resultados (em relação aos experimentais) do que o FoldX, chegando a resultados de superiores a dez vezes os valores experimentais para três (6FLA, 6FLB, 6FLC) dos quatro candidatos. Esta tendência parece estar de acordo com a discussão previamente realizada para as Figuras 14 e 15: ambos os programas apresentaram certo grau de magnificação das estimativas de $\Delta\Delta G$, mas o Rosetta chegou a resultados ainda maiores, em módulo, que o FoldX, com máximas de 13,7 kcal/mol no Rosetta e 5,1 kcal/mol no FoldX.

O PBEE apresentou valores de MAE (2,66) inferior aos demais candidatos, o que advoga em favor de sua exatidão superior em relação a estes. Além disso, os valores de RMSE (5,32) e SD (0,80) também foram consideravelmente inferiores aos demais candidatos, o que

indica melhor precisão do método para este problema. Além disso, o problema da discordância das previsões da estrutura 3UZV em relação às demais não ocorreu no PBEE: embora os valores de ΔΔG experimental de -13 kcal/mol não tenham sido previstos, o programa encontrou valores mais próximos dos experimentais (em relação aos demais programas) e mais concordantes entre si, confirmando sua já mencionada superioridade em exatidão e precisão, respectivamente.

Dos três programas, o PBEE apresentou o melhor desempenho nas métricas calculadas para este conjunto de dados, o que sugere que, para o problema em questão, sua exatidão e precisão são superiores a outros métodos. Ainda assim, nenhum dos três métodos teve RMSE próximo de zero, indicando que nenhum deles foi acurado nas estimativas de ΔΔG. A falta de acurácia de métodos de cálculo mais rápido (quando comparados com métodos baseados em MD, cf. secção 2.6.3) é esperada. Idealmente, o cálculo de ΔΔG deveria ser realizado por métodos de dinâmica molecular, conforme explanado na introdução (cf. secção 2.6). Porém, o altíssimo volume de dados em problemas biológicos, em combinação com a necessidade de respostas rápidas (para medidas epidemiológicas efetivas em tempo hábil) em problemas de saúde humana precludem seu uso neste tipo de problema. Estes cenários de cálculos rápidos com alto volume de informações tem termo em inglês *high throughput*, sem equivalente em língua portuguesa. Nos cenários *high throughput*, o maior desafio é garantir que os cálculos sejam feitos na velocidade adequada sem sacrificar consideravelmente a sua acurácia. É impossível, com as tecnologias atuais, que os cálculos sejam feitos de maneira mais rápida que os métodos alquímicos sem perder acurácia alguma: esta troca de custo computacional e acurácia é clássico em simulações computacionais (de natureza química ou não) e surgem da própria natureza dos modelos. Até o momento, nossa discussão foi baseada amplamente nos fundamentos metodológicos de Popper e Khun, porém, cabe uma pequena digressão para a discussão breve da importância dos modelos na dinâmica científica, elegantemente estudados no período de estruturalismo metateórico de Sneed e Stegmüller, na década de 1970. Em ciência, modelos são estruturas compostas por sequência de conjuntos de entidades (chamadas domínios do modelo) e as relações entre os domínios. Um modelo deve ter uma relação de representação com uma parcela da realidade: escolhe-se o modelo mais adequado para representar um fenômeno. Em geral, escolhe-se um modelo mais ou menos pormenorizado para fornecer explicações para perguntas científicas. Não é prudente (ou custo-efetivo), escolher um modelo excessivamente pormenorizado quando um modelo mais simples já responde às perguntas propostas na investigação. Similarmente, não é adequado escolher um modelo tão simples que não responda às perguntas investigativas a contento. (Sneed, 1971; Stegmüller,

1979) Na modelagem computacional não é diferente. A relação entre custo computacional e acurácia do modelo adotado é uma consequência natural da própria dinâmica das evoluções científicas, conforme defendido pela perspectiva estruturalista metateórica mencionada.

Assim, embora o problema do cálculo computacional da energia livre de Gibbs em cenários biológicos pareça recente, trata-se apenas de um caso particular do problema da relação de representação dos modelos em ciência (que, embora só descrita na década de 1970, é exemplificada desde os primórdios do pensamento científico iluminista). Não há solução clara para o problema agora (tanto quanto não havia para os iluministas no século XVII) e é possível que jamais haja. O cientista deve usar seu arsenal de habilidades para definir o melhor modelo para cada situação e esta decisão inclui argumentos factuais experimentais (como os aqui apresentados), lógicos (como os apresentados na secção anterior), argumentos específicos do conhecimento e do léxico de seu domínio cognitivo (como os argumentos químicos apresentados anteriormente) e, por fim, um conjunto menos estruturado, mais igualmente importante, de argumentos heurísticos construídos com sua experiência individual (de carreira) e coletiva (de seu grupo, de seus orientadores, da literatura e da comunidade científica como um todo). Com o conjunto de argumentos apresentado, parece seguro tecer algumas afirmações:

1. Embora tenha ocorrido discordância entre os métodos apresentados, o método do PBEE parece o mais adequado para cálculo de $\Delta\Delta G$ em cenários *high throughput*, de ocorrência comum em problemas de natureza biológica.
2. Como consequência direta da primeira afirmação, o método aqui apresentado para definir evasão imune em vírus a partir de informações genômicas e estruturais de suas proteínas antigênicas definiu que não há evidências computacionais, até o momento, de evasão imune no vírus DENV2, tomando sua proteína mais antigênica (proteína E) para o cálculo.
3. Ainda que a hipótese nula tenha sido corroborada, refutando a hipótese principal deste trabalho, do ponto de vista metodológico é interessante apontar que os procedimentos aqui tomados podem ser realizados para outros sistemas virais. Para isso, entretanto, o cientista deverá reunir argumentos sólidos em torno da escolha do modelo que melhor representa o seu problema (problema da relação de representação).

Embora esta última argumentação pareça particular a este cenário, ela é, como exposto, fundamental em todos os programas de investigação científica e ignorá-la pode trazer

consequências severas para a investigação, pois ignorar a relação de representação é ignorar as próprias limitações do conhecimento e do método, indo de encontro ao próprio fazer científico em aspectos de cognição e práxis. É importante também, como demonstrado neste trabalho, que o cientista não se restrinja a termos genéricos como robustez e precisão para definir o modelo a ser utilizado, mas que se utilize, além destes termos e definições, de um conjunto de argumentos de várias naturezas para defender o seu modelo adotado. No entanto, como o presente trabalho já o fez para o método PBEE na avaliação de evasão imune humoral em um sistema viral, é possível sugerir que este método possa ser extrapolado para a avaliação deste mesmo fenômeno em outros vírus. Os métodos aqui apresentados podem ser facilmente reproduzidos para a investigação de outras doenças com fisiopatologia semelhante, desde que seu paradigma fisiopatológico seja redutível à caracterização termodinâmica das interações proteína-proteína, como ocorreu neste trabalho. A extrapolação de resultados (robustez), porém, deve ser sempre vista com cautela, pelas razões já explanadas: para cada cenário, é importante utilizar criticamente os métodos e ser capaz de analisá-los quanto à falsificação da hipótese principal. Aceitar resultados sem esta análise crítica pode levar a hipóteses infalsificáveis, o que jamais deve ocorrer na prática científica séria.

Assim, conclui-se não apenas que o PBEE teve melhor desempenho para o problema apresentado aqui, mas também que seu desempenho provavelmente pode ser extrapolado para avaliar a evasão imune em outros vírus. Embora a hipótese nula tenha sido favorecida por estes achados, a avaliação sistemática destes métodos distintos aqui apresentada pode servir de auxílio para a comunidade científica direcionar seus esforços para a implementação desta metodologia frente a outros vírus. Esperamos, assim, contribuir para a vigilância epidemiológica com um novo método computacional de previsão de evasão imune em tempo quase real. Além disso, os resultados apresentados aqui têm impacto socioeconômico positivo do ponto de vista vacinal. Como a evasão imune não foi comprovada para DENV2 até o momento, é esperado que a eficácia vacinal discutida na introdução seja mantida e que, portanto, as vacinas já existentes continuem constituindo estratégia de saúde pública eficaz no combate à doença.

4.4 PERSPECTIVAS FUTURAS

Apesar de os nossos dados computacionais serem insuficientes para comprovar a hipótese de “um genótipo – uma infecção”, também almejamos realizar experimentos em laboratório para avaliar a falseabilidade desta. Tais ensaios serão realizados por metodologia

de neutralização viral em placas, comparando a resposta de cepas antigas de DENV2 frente a anticorpos contra a resposta de cepas mais novas (e circulantes) de DENV2 frente aos mesmos anticorpos. Esta metodologia permitirá documentar, se existirem, as diferenças de afinidades de interação com anticorpos entre as cepas antigas e novas.

É também de grande importância avaliar se mutações específicas poderiam gerar evasão imune caso ocorressem, tendo em vista planejamento de vigilância genômica. Ao saber de antemão quais mutações poderiam gerar evasão imune, as estratégias de vigilância epidemiológica podem antever o escape imunológico e potencialmente agir de acordo antes de isto se tornar um problema em larga escala epidemiológica. Ter uma ferramenta que investiga sistematicamente quais mutações poderiam gerar evasão imune antes mesmo de elas acontecerem permite que dados genômicos sozinhos já sejam suficientes para adoção de medidas de saúde pública de controle de disseminação destas variantes mutantes em tempo hábil. Portanto, propomos a realização de SSM de resíduos de aminoácidos chave para as PPI (isto é, *hotspots*), para avaliar se possíveis mutações poderiam levar à evasão imune. Assim, planejamos converter essas mutações calculadas sobre os resíduos de aminoácidos para sequências de DNA a elas correspondentes. Esta técnica permitiria uma vigilância genômica translacional direta e preditiva.

Vigilância genômica é uma área relativamente recente na integração entre biologia molecular/bioquímica e saúde pública. Muitas das técnicas de vigilância genômica foram desenvolvidas no contexto da pandemia de SARS-CoV-2. No melhor dos cenários, tais técnicas permitiram monitorar e prever a evolução dos vírus em tempo real, mas não permitem prever mutações que levem a variantes de preocupação. (Tosta et al., 2023b) Não foram encontrados métodos similares aos propostos aqui na literatura até o momento. Assim, propõe-se a aplicação desta metodologia para vigilância genômica de DENV e recomenda-se seu uso para que nos preparemos para mutações virais no contexto de outras patologias. Isto pode ser especificamente útil para doenças infecciosas virais, dada a capacidade notável de mutação e adaptação viral rápida em períodos muito curtos. Com a chegada das ciências ômicas e de técnicas de vigilância genômica, ambas as quais foram bem desenvolvidas durante a pandemia de SARS-CoV-2, a literatura ainda não tem informações suficientes sobre a visão interdisciplinar destas. (Hill et al., 2023b) Ser capaz de transitar entre vigilância genômica e proteômica utilizando metodologias da química computacional, como propomos no presente trabalho, é, além de inovador, primordial para a expansão da nossa compreensão acerca das conexões e dependências destas áreas específicas. Além disso, problemas do mundo real

raramente respeitam as barreiras entre as áreas científicas distintas, de modo que tal compreensão pode ser bastante útil no desenvolvimento de ferramentas aplicáveis à espécie humana.

Neste contexto, objetivamos iniciar a aplicação da metodologia de vigilância genômica apresentada acima e, esperançosamente, obter um protocolo geral que seja facilmente aplicável para outros sorotipos de DENV e, posteriormente, para outras doenças virais. Para isto, o grupo está desenvolvendo um *pipeline* único em Python que automatize a maior parte do processo, garantindo cálculos rápidos e gráficos para as novas cepas e sorotipos de DENV (e, futuramente, de outros vírus). Este *pipeline* geral em etapa única está sendo desenvolvido pelo nosso grupo de pesquisa tendo como base a metodologia aqui utilizada e está sendo implementado em plataforma de acesso amplo e rápido (*Google Colab*) pelo colega de grupo Whendel Muniz como parte de seu projeto de mestrado.

Na secção 4.1 discutimos alguns dos aspectos estruturais da proteína E. Entretanto, focamos nos aspectos úteis à compreensão dos resultados das secções futuras. Para os softwares e metodologias que utilizamos, os cálculos são realizados apenas para proteínas, peptídeos ou resíduos de amino ácidos. Porém, a proteína E é glicosilada e carboidratos não são computados nestes programas, sendo excluídos na etapa de limpeza dos PBDs, o que é necessário para o uso do Rosetta. Notamos, na inspeção dos PDBs que utilizamos, que a proteína E em complexo com anticorpos tem padrões de glicosilação específicos próximos aos epítopos que podem participar das PPI entre proteína E e anticorpos. Não somos capazes de determinar, utilizando as metodologias aqui apresentadas, qual é o papel destas glicosilações na resposta imune à DENV. Também não encontramos nenhuma discussão semelhante na literatura específica de DENV, embora o fato de a proteína E ser uma glicoproteína já ser bem estabelecido. Entretanto, há diversos relatos na literatura dos efeitos da glicosilação de proteínas nas interações com anticorpos em geral, mas nenhum relato específico para DENV ou outras arboviroses. (He; Zhou; Wang, 2024; Lee; Qi; Im, 2015; Sun; Suttapitugsakul; Wu, 2021)

Uma possibilidade de investigação da hipótese da influência dos padrões de glicosilação nas PPI dos complexos E:Ab é o uso de simulações de dinâmica molecular das proteínas do complexo. As simulações devem ser realizadas na presença e na ausência destes carboidratos, permitindo compreender seu papel nas interações e seus impactos nos parâmetros termodinâmicos. Uma das limitações desta abordagem é a ausência de estruturas contendo toda a cadeia de glicídios da proteína E, já que apenas os primeiros carboidratos ligados à cadeia

peptídica são mantidos nos experimentos de cristalografia convencionais. É possível contornar esta limitação procurando as sequências dos carboidratos em estudos de espectrometria de massas e reconstruindo computacionalmente suas sequências completas na estrutura. Porém, outro problema mais importante nesta investigação é o mesmo discutido nas metodologias de cálculo de ΔΔG: o custo computacional de cálculo de energia livre de interação (entre outras propriedades termodinâmicas de interesse) utilizando MD é muito elevado para ser realizado para todas as mutações e sequências virais obtidas, especialmente visando o uso de métodos alquímicos, nos quais muitas simulações são realizadas para um único cálculo de ΔΔG. Outra possibilidade de investigação (que, como já mencionado, está objetivamente nos nossos planos) é experimental. Os ensaios de neutralização supracitados serão realizados com vírus vivos, nos quais a proteína E está em sua forma ativa, dimérica e glicosilada. Assim, os próprios ensaios de neutralização podem trazer algum *insight* sobre a glicosilação, embora sua influência específica siga sendo desconhecida. A partir deles será possível, se necessário, investigar mais a fundo computacionalmente esta influência (já que em MD conseguimos comparar a presença e ausência destes padrões de glicosilação) utilizando as metodologias supracitadas apenas em mutações selecionadas com base nos resultados dos ensaios laboratoriais, no lugar de testar todas (custo computacional proibitivo) ou algumas ao acaso (baixas chances de encontrar resultados satisfatórios). Também é importante notar que, embora os cálculos aqui realizados ignorem os padrões de glicosilação, os erros possivelmente advindos disto são sistemáticos em sua essência. Como estamos comparando WT versus MUT, estes não devem impor problemas estatísticos ou invalidar o teste de hipóteses realizado.

5. CONCLUSÕES

A partir do presente estudo, foi possível determinar a energia livre de Gibbs de interação entre proteína do envelope e anticorpos de mamíferos, comparando as diferenças destas em cepas antigas e recentes do vírus de DENV. Não foi possível estabelecer a ocorrência de evasão imune até o momento. Isto porque apenas metodologias com menor precisão e exatidão foram capazes de derrubar a hipótese nula, mantendo o paradigma de uma infecção por sorotipo. A pressão evolutiva menor para a ocorrência de mutações que levem a escape imune (pela existência de 4 sorotipos que se alternam entre as gerações de humanos susceptíveis à infecção) em relação a modelos virais de sorotipo único pode ser a justificativa para a lentidão no acúmulo de mutações que levem à evasão imune. Embora ainda não comprovada, é possível que esta venha a ocorrer em breve e chama-se a atenção para a necessidade de políticas públicas de vigilância epidemiológica, prevenção e controle que levem em consideração esta possibilidade. Outra conclusão interessante a ser levantada é em relação à eficácia das vacinas licenciadas para uso clínico (Dengvaxia e Denvax). Embora esta não tenha sido objeto do presente estudo, como ainda não há evidência de ocorrência de evasão imune, é razoável esperar que os estudos de eficácia vacinal para sigam válidos e que, portanto, estas vacinas tenham sua eficácia mantida. Desta forma, as vacinas atuais para DENV continuam constituindo estratégia de saúde pública de grande importância no combate à doença.

Como perspectivas futuras, propõe-se a utilização da metodologia inovadora aqui proposta para monitorar a evolução viral de outras cepas em tempo real, o que demandará uma plataforma acessível e disponível para a comunidade científica. Tal plataforma está sendo desenvolvida pelo grupo, como parte do projeto de Mestrado do discente Whendel Muniz. Além disso, esperamos implementar estratégias de SSM para prever mutações antes mesmo do seu acontecimento, estratégia inovadora ainda não encontrada na literatura até então.

REFERÊNCIAS

ABAGYAN, Ruben; TOTROV, Maxim. Biased Probability Monte Carlo Conformational Searches and Electrostatic Calculations for Peptides and Proteins. **Journal of Molecular Biology**, v. 235, n. 3, p. 983–1002, 1994.

ABRAMSON, Josh *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. **Nature**, v. 630, n. 8016, p. 493–500, 13 jun. 2024.

ALFORD, Rebecca F. *et al.* The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. **Journal of Chemical Theory and Computation**, v. 13, n. 6, p. 3031–3048, 13 jun. 2017.

ALLEN, Michael. P.; TILDESLEY, Dominic J. **Computer Simulation of Liquids**. 1. ed. *[S.l.: S.n.]*.

ANFISEN, Cristian. **Principles that Govern the Folding of Protein Chains**. *[S.l.: S.n.]*. Disponível em: <https://www.science.org>.

ARAÚJO, Helena *et al.* Aedes aegypti Control Strategies in Brazil: Incorporation of New Technologies to Overcome the Persistence of Dengue Epidemics. **Insects**, v. 6, n. 2, p. 576–594, 11 jun. 2015.

B.A. SEIXAS, Jorge; GIOVANNI LUZ, Kleber; PINTO JUNIOR, Vitor. Atualização Clínica sobre Diagnóstico, Tratamento e Prevenção da Dengue. **Acta Médica Portuguesa**, v. 37, n. 2, p. 126–135, 1 fev. 2024.

BAEK, Minkyung *et al.* **Accurate prediction of protein structures and interactions using a three-track neural networkScience**. *[S.l.: S.n.]*. Disponível em: <https://predictioncenter.org/casp14/>.

BARLOW, Kyle A. *et al.* Flex ddG: Rosetta Ensemble-Based Estimation of Changes in Protein-Protein Binding Affinity upon Mutation. **Journal of Physical Chemistry B**, v. 122, n. 21, p. 5389–5399, 31 maio 2018.

BRASIL. **Nº DE CASOS PROVÁVEIS DE DENGUE POR SEMANA EPIDEMIOLÓGICA, BRASIL, 2023 E 2024**. *[S.l.: S.n.]*. Disponível em: <www.gov.br/saude/pt->.

BRASIL. **Painel de Monitoramento das Arboviroses**.

BROOM, Aron *et al.* Computational tools help improve protein stability but with a solubility tradeoff. **Journal of Biological Chemistry**, v. 292, n. 35, p. 14349–14361, 1 set. 2017.

BRYNGELSON, Joseph D. *et al.* **RESEARCH ARTICLES Funnels, Pathways, and the Energy Landscape of Protein Folding: A SynthesisPROTEINS: Structure, Function, and Genetics**. *[S.l.: S.n.]*.

BUSS, Oliver; RUDAT, Jens; OCHSENREITHER, Katrin. FoldX as Protein Engineering Tool: Better Than Random Based Approaches? **Computational and Structural Biotechnology Journal**, v. 16, p. 25–33, 2018.

CARUGO, Oliviero. How large B-factors can be in protein crystal structures. **BMC Bioinformatics**, v. 19, n. 1, p. 61, 23 dez. 2018.

CHAKRABORTY, Chiranjib *et al.* **A Detailed Overview of Immune Escape, Antibody Escape, Partial Vaccine Escape of SARS-CoV-2 and Their Emerging Variants With Escape Mutations**. **Frontiers in Immunology**Frontiers Media S.A., , 9 fev. 2022.

CHAVES, Elton J. F. *et al.* Estimating Absolute Protein-Protein Binding Free Energies by a Super Learner Model. **Journal of Chemical Information and Modeling**, 2025.

COCKBURN, Joseph J. B. *et al.* Mechanism of Dengue Virus Broad Cross-Neutralization by a Monoclonal Antibody. **Structure**, v. 20, n. 2, p. 303–314, fev. 2012.

CONWAY, Patrick *et al.* Relaxation of backbone bond geometry improves protein energy landscape modeling. **Protein Science**, v. 23, n. 1, p. 47–55, 2014.

D. K., Shreyas *et al.* A Review on Neural Networks and its Applications. **Journal of Computer Technology & Applications**, 2023.

DE ARAÚJO, Thalia Velho Barreto *et al.* Association between Zika virus infection and microcephaly in Brazil, January to May, 2016: preliminary report of a case-control study. **The Lancet Infectious Diseases**, v. 16, n. 12, p. 1356–1363, 1 dez. 2016.

DE ARAÚJO, Thalia Velho Barreto *et al.* Association between microcephaly, Zika virus infection, and other risk factors in Brazil: Final report of a case-control study. **The Lancet Infectious Diseases**, p. 328–336, 1 mar. 2018.

DE BARROS MIRANDA-FILHO, Demócrito *et al.* Initial Description of the Presumed Congenital Zika Syndrome. **Public Health**, v. 106, p. 598–600, 2016.

DEHOUCK, Yves *et al.* Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. **Bioinformatics**, v. 25, n. 19, p. 2537–2543, 2009.

DEKIMPE, Sofie; MASSCHELEIN, Joleen. Beyond peptide bond formation: the versatile role of condensation domains in natural product biosynthesis. **Natural Product Reports**, v. 38, n. 10, p. 1910–1937, 2021.

DEL CAMPO, Miguel *et al.* The phenotypic spectrum of congenital Zika syndrome. **American Journal of Medical Genetics, Part A**, v. 173, n. 4, p. 841–857, 1 abr. 2017.

DIAZ-QUIJANO, Fredi Alexander *et al.* Effectiveness of mass dengue vaccination with CYD-TDV (Dengvaxia®) in the state of Paraná, Brazil: integrating case-cohort and case-control designs. **The Lancet Regional Health - Americas**, v. 35, p. 100777, jul. 2024.

DILL, Ken A.; CHAN, Hue Sun. From Levinthal to pathways to funnels. **Nature Structural & Molecular Biology**, v. 4, n. 1, p. 10–19, 1 jan. 1997.

EBRAHIMI, Sasha B.; SAMANTA, Devleena. **Engineering protein-based therapeutics through structural and chemical design**. **Nature Communications**Nature Research, , 1 dez. 2023.

EHWERHEMUEPHA, Louis *et al.* A super learner ensemble of 14 statistical learning models for predicting COVID-19 severity among patients with cardiovascular conditions. **Intelligence-Based Medicine**, v. 5, p. 100030, 2021.

FERRAZ, Matheus V. F. *et al.* Immune evasion of SARS-CoV-2 variants of concern is driven by low affinity to neutralizing antibodies. **Chemical Communications**, v. 57, n. 49, p. 6094–6097, 2021.

FERRAZ, Matheus V. F. *et al.* An artificial neural network model to predict structure-based protein–protein free energy of binding from Rosetta-calculated properties. **Physical Chemistry Chemical Physics**, v. 25, n. 10, p. 7257–7267, 2023.

FISCHER, Janina; KAUFMANN, Jan Ole; WELLER, Michael G. Simple Determination of Affinity Constants of Antibodies by Competitive Immunoassays. **Methods and Protocols**, v. 7, n. 3, 1 jun. 2024.

FLEISHMAN, Sarel J. *et al.* Rosettascripts: A scripting language interface to the Rosetta Macromolecular modeling suite. **PLoS ONE**, v. 6, n. 6, 2011.

GRÄF, Tiago *et al.* Multiple introductions and country-wide spread of DENV-2 genotype II (Cosmopolitan) in Brazil. **Virus Evolution**, v. 9, n. 2, p. vead059, 1 jan. 2023.

HALSTEAD, Scott B. **Pathogenesis of Dengue: Challenges to Molecular BiologyProc. R. Soc. London Ser. A**. *[S.l.]*: Cambridge Univ. Press, 1893. Disponível em: <https://www.science.org>.

HE, Mengyuan; ZHOU, Xiangxiang; WANG, Xin. Glycosylation: mechanisms, biological functions and clinical implications. **Signal Transduction and Targeted Therapy**, v. 9, n. 1, p. 194, 2024.

HELLINGA, H. W. Rational protein design: Combining theory and experiment. **Proceedings of the National Academy of Sciences**, v. 94, n. 19, p. 10015–10017, 16 set. 1997.

HILL, Verity *et al.* **Toward a global virus genomic surveillance network**. **Cell Host and Microbe**Cell Press, , 14 jun. 2023a.

HILL, Verity *et al.* **Toward a global virus genomic surveillance network**. **Cell Host and Microbe**Cell Press, , 14 jun. 2023b.

HOLLINGSWORTH, Scott A.; KARPLUS, P. Andrew. **A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins**. **Biomolecular Concepts**De Gruyter Mouton, , 1 out. 2010.

HORNS, Felix; DEKKER, Cornelia L.; QUAKE, Stephen R. Memory B Cell Activation, Broad Anti-influenza Antibodies, and Bystander Activation Revealed by Single-Cell Transcriptomics. **Cell Reports**, v. 30, n. 3, p. 905- 913.e6, 21 jan. 2020.

HUANG, Kerson. **Lectures On Statistical Physics And Protein Folding**. 1. ed. *[S.l.]*: World Scientific, 2005.

HUMPHREY, William; DALKE, Andrew; SCHULTEN, Klaus. **VMD: Visual Molecular Dynamics**. *[S.l.: S.n.]*.

IMRIE, Allison *et al.* Antibody to dengue 1 detected more than 60 years after infection. **Viral Immunology**, v. 20, n. 4, p. 672–675, 1 dez. 2007.

IVANKOV, Dmitry N.; FINKELSTEIN, Alexei V. **Solution of levinthal's paradox and a physical theory of protein folding times**. **Biomolecules**MDPI AG, , 1 fev. 2020.

JOZALA, Angela Faustino *et al.* **Biopharmaceuticals from microorganisms: from production to purification**. **Brazilian Journal of Microbiology**Elsevier Editora Ltda, , 1 dez. 2016.

JUMPER, John *et al.* Highly accurate protein structure prediction with AlphaFold. **Nature**, v. 596, n. 7873, p. 583–589, 26 ago. 2021.

JURRUS, Elizabeth *et al.* Improvements to the APBS biomolecular solvation software suite. **Protein Science**, v. 27, n. 1, p. 112–128, 24 jan. 2018.

KALLÁS, Esper G. *et al.* Live, Attenuated, Tetravalent Butantan–Dengue Vaccine in Children and Adults. **New England Journal of Medicine**, v. 390, n. 5, p. 397–408, fev. 2024.

KARIYAWASAM, Ruwandi *et al.* **A dengue vaccine whirlwind update**. **Therapeutic Advances in Infectious Disease**SAGE Publications Ltd, , 1 jan. 2023.

KARLSSON, Robert; MICHAELSSON, Anne; MATTSSON, Lars. Kinetic analysis of monoclonal antibody-antigen interactions with a new biosensor based analytical system. **Journal of Immunological Methods**, v. 145, n. 1–2, p. 229–240, dez. 1991.

KHAN, Muhammad Bilal *et al.* **Dengue overview: An updated systemic review**. **Journal of Infection and Public Health**Elsevier Ltd, , 1 out. 2023.

KRAMVIS, Anna; KEW, Michael; FRANÇOIS, Guido. Hepatitis B virus genotypes. **Vaccine**, v. 23, n. 19, p. 2409–2423, 2005.

KUHLMAN, Brian *et al.* **Design of a Novel Globular Protein Fold with Atomic-Level Accuracy**. *[S.l.: S.n.]*. Disponível em: <www.sciencemag.org>.

KUHN, Jens H. *et al.* Virus nomenclature below the species level: a standardized nomenclature for natural variants of viruses assigned to the family Filoviridae. **Archives of Virology**, v. 158, n. 1, p. 301–311, 23 jan. 2013.

KUHN, Thomas S. **A estrutura das revoluções científicas**. 13. ed. *[S.l.]*: São Paulo: Perspectiva (Debates), 2018.

KUMAR, Vijay *et al.* Computing disease-linked SOD1 mutations: deciphering protein stability and patient-phenotype relations. **Scientific Reports**, v. 7, n. 1, p. 4678, 5 jul. 2017.

KUNO, G.; CHANG, G. J. J. Full-length sequencing and genomic characterization of Bagaza, Kedougou, and Zika viruses. **Archives of Virology**, v. 152, n. 4, p. 687–696, abr. 2007.

LANDRY, James P.; FEI, Yiyan; ZHU, Xiangdong. Simultaneous Measurement of 10,000 Protein-Ligand Affinity Constants Using Microarray-Based Kinetic Constant Assays. **ASSAY and Drug Development Technologies**, v. 10, n. 3, p. 250–259, jun. 2012.

LEACH, Andrew. **Molecular Modelling: Principles and Applications**. 2. ed. Edinburgh: Peasron Education Limited, 2001.

LEAVER-FAY, Andrew *et al.* Rosetta3: An object-oriented software suite for the simulation and design of macromolecules. *In*: **Methods in Enzymology**. *[S.l.]*: Academic Press Inc., 2011. v. 487 p. 545–574.

LEAVER-FAY, Andrew *et al.* Scientific benchmarks for guiding macromolecular energy function improvement. *In*: **Methods in Enzymology**. *[S.l.]*: Academic Press Inc., 2013. v. 523 p. 109–143.

LEE, Hui Sun; QI, Yifei; IM, Wonpil. Effects of N-glycosylation on protein conformation and dynamics: Protein Data Bank analysis and molecular dynamics simulation study. **Scientific Reports**, v. 5, n. 1, p. 8926, 2015.

LEE, Michelle Felicia *et al.* **Innate and adaptive immune evasion by dengue virus**. **Frontiers in Cellular and Infection Microbiology**Frontiers Media S.A., , 16 set. 2022.

LEE, Seunghye *et al.* Super learner machine-learning algorithms for compressive strength prediction of high performance concrete. **Structural Concrete**, v. 24, n. 2, p. 2208–2228, 1 abr. 2023.

MARTINEZ, Leandro; BORIZ, Ivana; SKAF, Munir. Fundamentos de Simulação por Dinâmica Molecular. *In*: **Métodos de Química Teórica e Modelagem Molecular**. 1. ed. *[S.l.: S.n.]*. p. 413–452.

MCGIBBON, Robert T. *et al.* MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. **Biophysical Journal**, v. 109, n. 8, p. 1528–1532, 23 out. 2015.

MCQUARRIE, Donald. **Statistical Mechanics**. 1. ed. *[S.l.]*: University Science Books, 2003.

MCQUARRIE, Donald A.; SIMON, John D. **Molecular Thermodynamics**. 1. ed. *[S.l.]*: University Science Books, 1999.

MESSINA, Jane P. *et al.* The current and future global distribution and population at risk of dengue. **Nature Microbiology**, v. 4, n. 9, p. 1508–1515, 1 set. 2019.

METROPOLIS, Nicholas *et al.* Equation of State Calculations by Fast Computing Machines. **The Journal of Chemical Physics**, v. 21, n. 6, p. 1087–1092, 1 jun. 1953.

METROPOLIS, Nicholas; ULAM, S. The Monte Carlo Method. **Journal of The American Statistical Association**, v. 44, n. 247, p. 335–341, set. 1949.

MOORE, Calvin C. **Ergodic theorem, ergodic theory, and statistical mechanics**. **Proceedings of the National Academy of Sciences of the United States of America**National Academy of Sciences, , 17 fev. 2015.

MORGON, Nelson; COUTINHO, Kaline. **Métodos de Química Teórica e Modelagem Molecular**. 1. ed. *[S.l.: S.n.]*.

MULLER, David A.; YOUNG, Paul R. **The flavivirus NS1 protein: Molecular and structural biology, immunology, role inpathogenesis and application asadiagnostic biomarker. Antiviral Research**Elsevier B.V., , 2013.

MURRAY, Natasha Evelyn Anne; QUAM, Mikkel B.; WILDER-SMITH, Annelies. **Epidemiology of dengue: Past, present and future prospects**. **Clinical Epidemiology**, 19 ago. 2013.

MUSHTAQ, Saira *et al.* Novel mutations in structural proteins of dengue virus genomes. **Journal of Infection and Public Health**, v. 16, n. 12, p. 1971–1981, 1 dez. 2023.

O'MEARA, Matthew J. *et al.* Combined Covalent-Electrostatic Model of Hydrogen Bonding Improves Structure Prediction with Rosetta. **Journal of Chemical Theory and Computation**, v. 11, n. 2, p. 609–622, 2015.

PACKER, Michael S.; LIU, David R. **Methods for the directed evolution of proteins**. **Nature Reviews Genetics**Nature Publishing Group, , 19 jul. 2015.

PENTEADO, André Berndt *et al.* Binding Evolution of the Dengue Virus Envelope Against DC-SIGN: A Combined Approach of Phylogenetics and Molecular Dynamics Analyses Over 30 Years of Dengue Virus in Brazil. **Journal of Molecular Biology**, v. 436, n. 11, 1 jun. 2024.

POPPER, Karl Raimund, Sir. **A lógica da pesquisa científica**. 2. ed. *[S.l.]*: São Paulo: Cultrix, 2013.

PROMMOOL, Tanapan *et al.* High performance dengue virus antigen-based serotyping-NS1-ELISA (plus): A simple alternative approach to identify dengue virus serotypes in acute dengue specimens. **PLOS Neglected Tropical Diseases**, v. 15, n. 2, p. e0009065, 26 fev. 2021.

RENNER, Max *et al.* Characterization of a potent and highly unusual minimally enhancing antibody directed against dengue virus. **Nature Immunology**, v. 19, n. 11, p. 1248–1256, 15 nov. 2018.

RESENDE, Paola Cristina *et al.* The ongoing evolution of variants of concern and interest of SARS-CoV-2 in Brazil revealed by convergent indels in the amino (N)-terminal domain of the spike protein. **Virus Evolution**, v. 7, n. 2, 2021.

SALENTIN, Sebastian *et al.* PLIP: fully automated protein–ligand interaction profiler. **Nucleic Acids Research**, v. 43, n. W1, p. W443–W447, 1 jul. 2015.

SARKER, Animesh; DHAMA, Nidhi; GUPTA, Rinkoo Devi. **Dengue virus neutralizing antibody: a review of targets, cross-reactivity, and antibody-dependent enhancement**. **Frontiers in Immunology**Frontiers Media S.A., , 2023.

SCHIEFFELIN, John S. *et al.* Neutralizing and non-neutralizing monoclonal antibodies against dengue virus E protein derived from a naturally infected patient. **Virology Journal**, v. 7, n. 1, p. 28, 4 dez. 2010.

SCHLICK, Tamar. **Molecular Modeling and Simulation: An Interdisciplinary Guide**. New York, NY: Springer New York, 2010. v. 21

SCHMIDT, Fabian *et al.* High genetic barrier to SARS-CoV-2 polyclonal neutralizing antibody escape. **Nature**, v. 600, n. 7889, p. 512–516, 16 dez. 2021.

SCHYMKOWITZ, Joost *et al.* The FoldX web server: An online force field. **Nucleic Acids Research**, v. 33, n. SUPPL. 2, jul. 2005.

SELZER, Tzvia; ALBECK, Shira; SCHREIBER, Gideon. **letters Rational design of faster associating and tighter binding protein complexes**. *[S.l.: S.n.]*. Disponível em: <http://structbio.nature.com>.

SHI, Yun Fei *et al.* **Machine Learning for Chemistry: Basics and Applications**. **Engineering**Elsevier Ltd, , 1 ago. 2023.

SHU, Pei-Yun *et al.* Dengue Virus Serotyping Based on Envelope and Membrane and Nonstructural Protein NS1 Serotype-Specific Capture Immunoglobulin M Enzyme-Linked

Immunosorbent Assays. **Journal of Clinical Microbiology**, v. 42, n. 6, p. 2489–2494, jun. 2004.

SIMON-LORIERE, Etienne; SCHWARTZ, Olivier. Towards SARS-CoV-2 serotypes? **Nature Reviews Microbiology**, v. 20, n. 4, p. 187–188, 18 abr. 2022.

SNEED, J. D. **The Logical Structure of Mathematical Physics**. 1. ed. *[S.l.]*: Dordecht: Reidel, 1971.

SOILA, Sukupolvi-Petty *et al.* Structure and Function Analysis of Therapeutic Monoclonal Antibodies against Dengue Virus Type 2. **Journal of Virology**, v. 84, n. 18, p. 9227–9239, 15 set. 2010.

SOUZA, Wayner Vieira De *et al.* Microcephaly epidemic related to the Zika virus and living conditions in Recife, Northeast Brazil. **BMC Public Health**, v. 18, n. 1, 12 jan. 2018.

STEGMÜLLER, W. **The Structuralist View: Survey, Recent Developments and Answers to Some Criticisms**. 1. ed. *[S.l.]*: Amsterdam: North Holland, 1979.

SUKUPOLVI-PETTY, Soila *et al.* Structure and Function Analysis of Therapeutic Monoclonal Antibodies against Dengue Virus Type 2. **Journal of Virology**, v. 84, n. 18, p. 9227–9239, 15 set. 2010.

SUN, Fangxu; SUTTAPITUGSAKUL, Suttipong; WU, Ronghu. Unraveling the surface glycoprotein interaction network by integrating chemical crosslinking with MS-based proteomics. **Chemical Science**, v. 12, n. 6, p. 2146–2155, 2021.

TOSTA, Stephane *et al.* Global SARS-CoV-2 genomic surveillance: What we have learned (so far). **Infection, Genetics and Evolution**, v. 108, 1 mar. 2023a.

TOSTA, Stephane *et al.* Global SARS-CoV-2 genomic surveillance: What we have learned (so far). **Infection, Genetics and Evolution**, v. 108, 1 mar. 2023b.

TSUJI, Shoichiro *et al.* TACI deficiency enhances antibody avidity and clearance of an intestinal pathogen. **Journal of Clinical Investigation**, v. 124, n. 11, p. 4857–4866, 3 nov. 2014.

ULMER, Kevin M. Protein Engineering. **Science**, v. 219, n. 4585, p. 666–671, 11 fev. 1983.

UVERSKY, Vladimir N. Chapter One - Protein intrinsic disorder and structure-function continuum. *In*: UVERSKY, Vladimir N. (Org.). **Progress in Molecular Biology and Translational Science**. *[S.l.]*: Academic Press, 2019. v. 166 p. 1–17.

VAN GUNSTEREN, W. F.; BERENDSEN, H. J. C. **Thermodynamic cycle integration by computer simulation as a tool for obtaining free energy differences in molecular chemistryJournal of Computer-Aided Molecular Design**. *[S.l.: S.n.]*.

VAN GUNSTEREN, Wilfred F.; DAURA, Xavier; MARK, Alan E. Computation of free energy. **Helvetica Chimica Acta**, v. 85, n. 10, p. 3113–3129, 2002.

VERLI, Hugo. Bioinformática: da Biologia à Flexibilidade Molecular. **Sociedade Brasileira de Bioquímica e Biologia Molecular**, 2014.

VOET, Donald; VOET, Judith G. **Bioquímica**. 4. ed. *[S.l.: S.n.]*.

WATERHOUSE, Andrew *et al.* SWISS-MODEL: Homology modelling of protein structures and complexes. **Nucleic Acids Research**, v. 46, n. W1, p. W296–W303, 2 jul. 2018.

WATSON, Joseph L. *et al.* De novo design of protein structure and function with RFdiffusion. **Nature**, v. 620, n. 7976, p. 1089–1100, 31 ago. 2023.

WHITFORD, David. **Proteins: Structure and Function**. *[S.l.: S.n.]*.

WONG, Tuck; ZHURINA, Daria; SCHWANEBERG, Ulrich. The Diversity Challenge in Directed Protein Evolution. **Combinatorial Chemistry & High Throughput Screening**, v. 9, n. 4, p. 271–288, 1 maio 2006.

YOUNG, Paul R. *et al.* **An Antigen Capture Enzyme-Linked Immunosorbent Assay Reveals High Levels of the Dengue Virus Protein NS1 in the Sera of Infected PatientsJOURNAL OF CLINICAL MICROBIOLOGY**. *[S.l.: S.n.]*.

YOUVAN, Douglas C. Searching Sequence Space. **Bio/Technology**, v. 13, n. 8, p. 722–723, 1995.

YUE, Yang *et al.* MpbPPI: A multi-Task pre-Training-based equivariant approach for the prediction of the effect of amino acid mutations on protein-protein interactions. **Briefings in Bioinformatics**, v. 24, n. 5, 1 set. 2023.

APÊNDICES

I. *Scripts* Utilizados

```
from Bio import SeqIO

from Bio.Seq import Seq

from Bio.SeqRecord import SeqRecord

import re


my_file = "\path\to\your\fasta\file"


def protein_process(input_filename):

    # Lists

    translated_list = []  # List to keep the translated sequences in step 1

    removedup1_list = []  # List to keep the unique sequences in step 2 (without duplication)

    removedup2_list = []  # List to keep the sequences (already kept in removedup1_list) with the same number of amino acids


    # STEP1: DNA to Protein

    with open(input_filename, "r") as handle:

        for record in SeqIO.parse(handle, "fasta"):

            sequence = str(record.seq)

            translated_sequence = []

            # If it is not divisible by 3

            if len(sequence) % 3 != 0:

                # Warning Message (incomplete codons)

                print("INCOMPLETE CODONS VERIFICATION")

                print(f"This sequence (Id>{record.id}) has 1 incomplete codon. The code will delete this one, but it may become an error in future.")
```

```
                print()
            sequence = sequence.replace('n', '-').replace('N', '-').replace('x', '-
').replace('X', '-').replace('*', '-')
            # Iterating all codons
            codons = [sequence[i:i+3] for i in range(0, len(sequence), 3)]
            for codon in codons:
                # Find those codons with "-" and replace them with "---"
                if '-' in codon:
                    codon = '---'
                    translated_sequence.append('-')
                else:
                    translated_sequence.append(str(Seq(codon).translate()))
            # In some sequences, the last character is "-" and it's considered a
"incomplete codon" too. For all incomplete codons, they're translated for "".
            if translated_sequence and translated_sequence[-1] == '-':
                translated_sequence[-1] = ''
            translated_sequence_str = ''.join(translated_sequence)
            translated_list.append(SeqRecord(Seq(translated_sequence_str),
id=record.id, description=record.description))


            # Common length
            expected_length = (len(sequence) // 3)
            # Remove incomplete codons length.
            remove = translated_sequence.count("")
            # Count of all amino acids removing the last one from incomplete codons
            actual_length = len(translated_sequence) - remove
            if actual_length != expected_length:
                print("VERIFYING TRANSLATION")
                print(f"Wrong translation! Length does not match in {record.id}")
                print(f"Expected   length:   {expected_length},   Actual   length:
{actual_length}")
```

```
                print()
            else:
                print("VERIFYING TRANSLATION")
                print(f"Correct translation for record {record.id}")
                print()

    # Creating a fasta file with all amino acids and their IDs
    translated_filename = input_filename.split(".")[0] + "_translated.fasta"
    with open(translated_filename, "w") as output_handle:
        SeqIO.write(translated_list, output_handle, "fasta")
        print()
        print("Finished translation")

    # STEP2: Remove duplications, MRCI initialization and leftover amino acids
    with open(translated_filename, 'r') as handle:
        for record in SeqIO.parse(handle, "fasta"):
            sequence = str(record.seq)
            # The sequence is unique in list
            if sequence not in [str(rec.seq) for rec in removedup1_list]:
                nuc_find = sequence.find("MRCI")
                if nuc_find >= 0:
                    sequence = sequence[nuc_find:]
                    removedup1_list.append(SeqRecord(Seq(sequence),  id=record.id,
description=record.description))
            # Duplicated sequence (It won't be considered)
            else:
                continue

    # The number of smallest sequences to model all others.
    minlength_seq = 391
```

```
    for rec in removedup1_list:
        # Remove leftover amino acids based on smallest number.
        removedup2_seq = str(rec.seq)[:minlength_seq]
        #Find any deletions in sequences
        del_find = removedup2_seq.find("-")
        if del_find > 0:
          continue
        #Add aminoacid sequences without deletions.
        else:
          removedup2_list.append(SeqRecord(Seq(removedup2_seq),          id=rec.id,
description=rec.description))


    # Create a file with all sequences without duplications and MRCI/smallest number
settings
    removedup_filename = input_filename.split(".")[0] + "_removedup.fasta"
    with open(removedup_filename, "w") as output_handle:
        SeqIO.write(removedup2_list, output_handle, "fasta")
        print()
        print("The duplicated sequences have been deleted and all of the unique
sequences have the same number of amino acids.")
        print()


    # Gen resfiles
    #Ref sequence (Put your ref sequence)
    reference_sequence                                                              =
"MRCIGISNRDFVEGVSGGSWVDIVLEHGSCVTTMAKNKPTLDFELIKTEAKQPATLRKYCIEAKLTNTTTESRCPTQGEPT
LNEEQDKRFVCKHSMVDRGWGNGCGLFGKGGIVTCAMFTCKKNMEGKIVQPENLEYTVVITPHSGEEHAVGNDTGKHGKEVK
ITPQSSITEAELTGYGTVTMECSPRTGLDFNEMVLLQMKDKAWLVHRQWFLDLPLPWLPGADTQGSNWIQKETLVTFKNPHA
KKQDVVVLGSQEGAMHTALTGATEIQMSSGNLLFTGHLKCRLRMDKLQLKGMSYSMCTGKFKVVKEIAETQHGTIVIRVQYE
GDGSPCKIPFEIMDLEKRHVLGRLITVNPIVTEKDSPVNIEAEPPFGDSYIIIGVEPGQLKLNWFKK"
    with open(removedup_filename, "r") as handle:
        for record in SeqIO.parse(handle, "fasta"):
            sequence = str(record.seq)
```

```
            header = str(record.id)
            # Create an file for each sequence with the mutations.
            finalprotein_filename = re.sub(r'[^\w\-_\.]', '_', header) + ".txt"
            with open(finalprotein_filename, "w") as output_handle:
                output_handle.write("NATRO\n")
                output_handle.write("START\n")
                for iter, (ref_aa, seq_aa) in enumerate(zip(reference_sequence,
sequence)):
                    if ref_aa != seq_aa:
                        output_handle.write(f"{iter + 1} A PIKAA {seq_aa}\n")
                    output_handle.write(f"{iter + 1} B PIKAA {seq_aa}\n")
    print("The program has finished")


# Example of use
python3 protein_process(my_file)
```

Script 1. Processamento de sequências em FASTA obtidas da rede genômica. Inclui tradução das sequências em resíduos de aminoácidos, processamento das estruturas (para garantir que todas tenham o mesmo tamanho e comecem e terminem nas mesmas posições), e gênese de arquivos de texto do tipo *resfiles*, que serão utilizados como *input* para o programa Rosetta. Script em *Python3*.

```
-in:file:s PDB_id.pdb
-parser:protocol fastrelax.xml
-nstruct 1
-out:file:scorefile PDB_id_0001.sc
-ex1
-ex2aro
-use_input_sc
-corrections::beta_nov16
```

```
-overwrite
```

Script 2. *Flags* para o protocolo *fast relax*. Script em XML padrão do Rosetta (*Rosetta Scripts*).

```
<ROSETTASCRIPTS>
        <SCOREFXNS>
                <ScoreFunction name="sfxn" weights="beta_nov16"/>
        </SCOREFXNS>

     <FILTERS>
        <Geometry name="omega" omega="150" cart_bonded="100" confidence="0"/>
        <Rmsd name="rmsd" confidence="0" superimpose="1"/>
     </FILTERS>
        <MOVERS>
                <FastRelax     name="FastRelax"     scorefxn="sfxn"     repeats="5"
batch="false"  ramp_down_constraints="false"  cartesian="false"  bondangle="false"
bondlength="false" min_type="dfpmin_armijo_nonmonotone" />
                <InterfaceAnalyzerMover         name="ifa"         scorefxn="sfxn"
interface_sc="1"  interface="AB_C"  pack_separated="1"  pack_input="1"  packstat="1"
/>
        </MOVERS>
        <PROTOCOLS>
                 <Add filter_name="omega"/>
                <Add mover="FastRelax" />
                <Add mover="ifa" />
                <Add filter_name="rmsd"/>
        </PROTOCOLS>
        <OUTPUT />
</ROSETTASCRIPTS>
```

Script 3. Protocolo Fast relax do Rosetta. Script em XML padrão do Rosetta (*Rosetta Scripts*).

```
$ROSETTA313_BIN/rosetta_scripts.mpi.linuxgccrelease -database $ROSETTA313_DB -in:file:s 3uzv_sc_AB.pdb -restore_talaris_behavior -use_input_sc -parser:protocol ddg.xml -ex1 -ex2 -ex2aro -out:file:scorefile DENV2.sc -overwrite -script_vars in_resfile={input_file} -out:suffix _{l} -overwrite
```

Script 4. Linha de comando no terminal para executar protocolo Flex ddG do Rosetta.

```
<ROSETTASCRIPTS>
        <SCOREFXNS>
                <ScoreFunction name="fa_talaris2014" weights="talaris2014"/>
                <ScoreFunction name="fa_talaris2014_cst" weights="talaris2014">
                <Reweight scoretype="atom_pair_constraint" weight="1.0"/>
                <Set fa_max_dis="9.0"/>
                </ScoreFunction>
        </SCOREFXNS>

        <TASKOPERATIONS>
                <InitializeFromCommandline name="init"/>
                <ReadResfile name="resfile" filename="%%file%%" />
        </TASKOPERATIONS>

        <RESIDUE_SELECTORS>
                <Task name="resselector" fixed="0" packable="0" designable="1" task_operations="resfile"/>
                <Neighborhood name="bubble" selector="resselector" distance="8.0"/>
                <PrimarySequenceNeighborhood name="bubble_adjacent" selector="bubble" lower="1" upper="1"/>
```

```
                <StoredResidueSubset                    name="restore_neighbor_shell"
subset_name="neighbor_shell"/>

                <Not name="everythingelse" selector="restore_neighbor_shell"/>

        </RESIDUE_SELECTORS>


        <TASKOPERATIONS>

                <OperateOnResidueSubset                              name="repackonly"
selector="restore_neighbor_shell">

                        <RestrictToRepackingRLT/>

                </OperateOnResidueSubset>

                <OperateOnResidueSubset name="norepack" selector="everythingelse">

                        <PreventRepackingRLT/>

                </OperateOnResidueSubset>

                <UseMultiCoolAnnealer name="multicool" states="6"/>

                <ExtraChiCutoff name="extrachizero" extrachi_cutoff="0"/>

                <InitializeFromCommandline name="commandline_init"/>

                <RestrictToRepacking name="restrict_to_repacking"/>

   </TASKOPERATIONS>


   <FILTERS>

           <Ddg  name="ddg_filter"  threshold="1000"  repeats="3"  chain_num="3,4"
repack="1" repack_bound="0"/>

                # adjust chain_num to determine the interface, eg, for a protein E
with chains A (1) and B (2) and an antibody

                # with chains H (3) and L (4), the chain_num should be "1,2" or
"3,4";

                # for a protein E with only chain A (1) and an antibody with chains
H (2) and L(3), the chain_num should be "1" or "2,3"

   </FILTERS>


   <MOVERS>
```

```
        <MinMover name="minimize" scorefxn="fa_talaris2014_cst" chi="1" bb="1"
type="lbfgs_armijo_nonmonotone" tolerance="0.00001" max_iter="500"/>

            <StoreResidueSubset                    name="neighbor_shell_storer"
subset_name="neighbor_shell" residue_selector="bubble_adjacent" />

            <PackRotamersMover     name="repack"     scorefxn="fa_talaris2014"
task_operations="commandline_init,repackonly,norepack,multicool"/>

            <PackRotamersMover                              name="mut_and_pack"
task_operations="resfile,multicool,norepack"/>

            <FilterReportAsPoseExtraScoresMover                   name="dg_wt"
report_as="ddg_wt" filter_name="ddg_filter" />

            <FilterReportAsPoseExtraScoresMover                  name="dg_mut"
report_as="ddg_mut" filter_name="ddg_filter" />

            <ScoreMover    name="apply_score"    scorefxn="fa_talaris2014_cst"
verbose="0"/>

            <AddConstraintsToCurrentConformationMover            name="addcst"
use_distance_cst="1" coord_dev="0.5" min_seq_sep="0" max_distance="9" CA_only="1"
bound_width="0.0" cst_weight="0.0"/>

            <ClearConstraintsMover name="clearcst"/>

            </MOVERS>

        <PROTOCOLS>

            <Add mover_name="addcst"/>

            <Add mover_name="apply_score"/>

            <Add mover_name="neighbor_shell_storer"/>

            <Add mover_name="repack"/>

            <Add mover_name="addcst"/>

            <Add mover_name="minimize"/>

            <Add mover_name="clearcst"/>

            <Add mover_name="dg_wt"/>

            <Add mover_name="mut_and_pack"/>

            <Add mover_name="addcst"/>

            <Add mover_name="minimize"/>

            <Add mover_name="clearcst"/>

            <Add mover_name="dg_mut"/>
```

```
        </PROTOCOLS>

</ROSETTASCRIPTS>
```

Script 5. Protocolo Flex ddG do Rosetta. Script em XML padrão do Rosetta (*Rosetta Scripts*).

```
##########################

install.packages("/private/var/folders/fh/lq9zthkd3hb02n_64g0kql2w0000gn/T/Rtmp32Ob61/downloaded_packages/xml2_1.3.6.tar.gz", repos = NULL, type = "source")

install.packages("/private/var/folders/fh/lq9zthkd3hb02n_64g0kql2w0000gn/T/Rtmp32Ob61/downloaded_packages/tidyverse_2.0.0.tar.gz", repos = NULL, type = "source")

install.packages("xml2")

install.packages("tidyverse")

install.packages("ggplot2")

install.packages("ggpubr")

install.packages("data.table")

install.packages("reshape2")

install.packages("dplyr")

install.packages("devtools")

library(tidyverse)

library(ggplot2)

library(ggpubr)

library(data.table)

library(reshape2)

library(dplyr)


# Carregando os dados

ASIA <- read.csv("/Users/juliosimoes/Documents/DENV_Manhattan_Project/PBEE_results/ASIA_final.sc", sep = "", skip = 1, header = TRUE)

CA <- read.csv("/Users/juliosimoes/Documents/DENV_Manhattan_Project/PBEE_results/CA_final.sc", sep = "", skip = 1, header = TRUE)
```

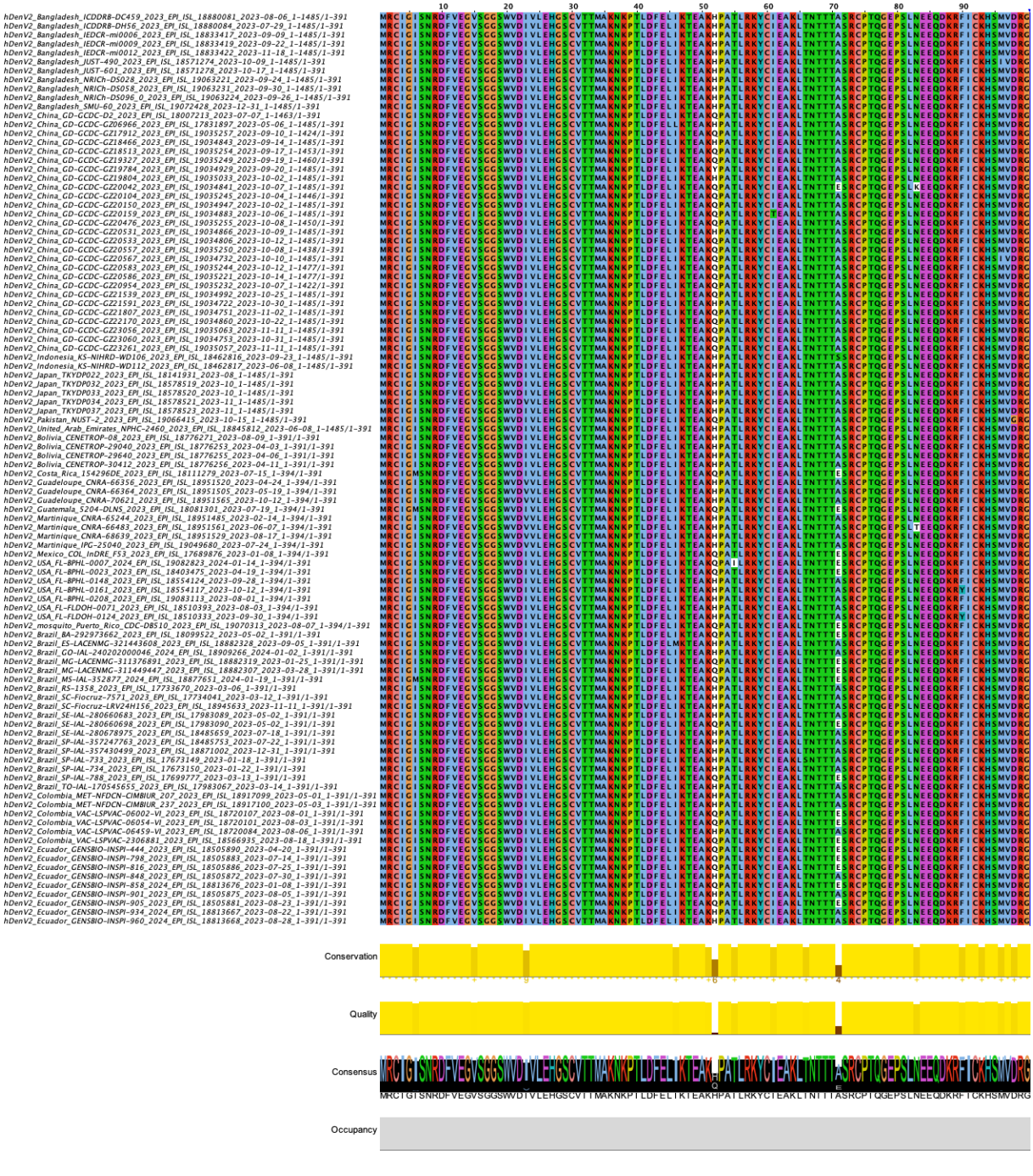
```
SA <- read.csv("/Users/juliosimoes/Documents/DENV_Manhattan_Project/PBEE_results/SA_final.sc", sep = "", skip = 1, header = TRUE)


# Criando novos data frames com os dados necessários e extraindo os quatro primeiros caracteres da descrição

ASIA <- data.frame(ddg = ASIA$ddg_mut - ASIA$ddg_wt, name = substr(ASIA$description, 1, 4), variants = ASIA$description, ID = ASIA$ID)

CA <- data.frame(ddg = CA$ddg_mut - CA$ddg_wt, name = substr(CA$description, 1, 4), variants = CA$description, ID = CA$ID)

SA <- data.frame(ddg = SA$ddg_mut - SA$ddg_wt, name = substr(SA$description, 1, 4), variants = SA$description, ID = SA$ID)


# Adicionando a coluna 'id' para identificar a origem dos dados

ASIA$id <- "ASIA"

CA$id <- "CA"

SA$id <- "SA"


# Combinando os data frames

df <- rbind(ASIA, CA, SA)


df = df %>%  mutate(name= toupper(name))

# Carregando a biblioteca ggplot2 e forcats

library(ggplot2)

library(forcats)

library(scales)


# Criando o gráfico

r = ggplot(df, aes(x = fct_inorder(factor(name)), y = as.numeric(ddg))) +

  annotate(ymin = -1.7, ymax = 1.7,

           xmin = -Inf, xmax = Inf,
```

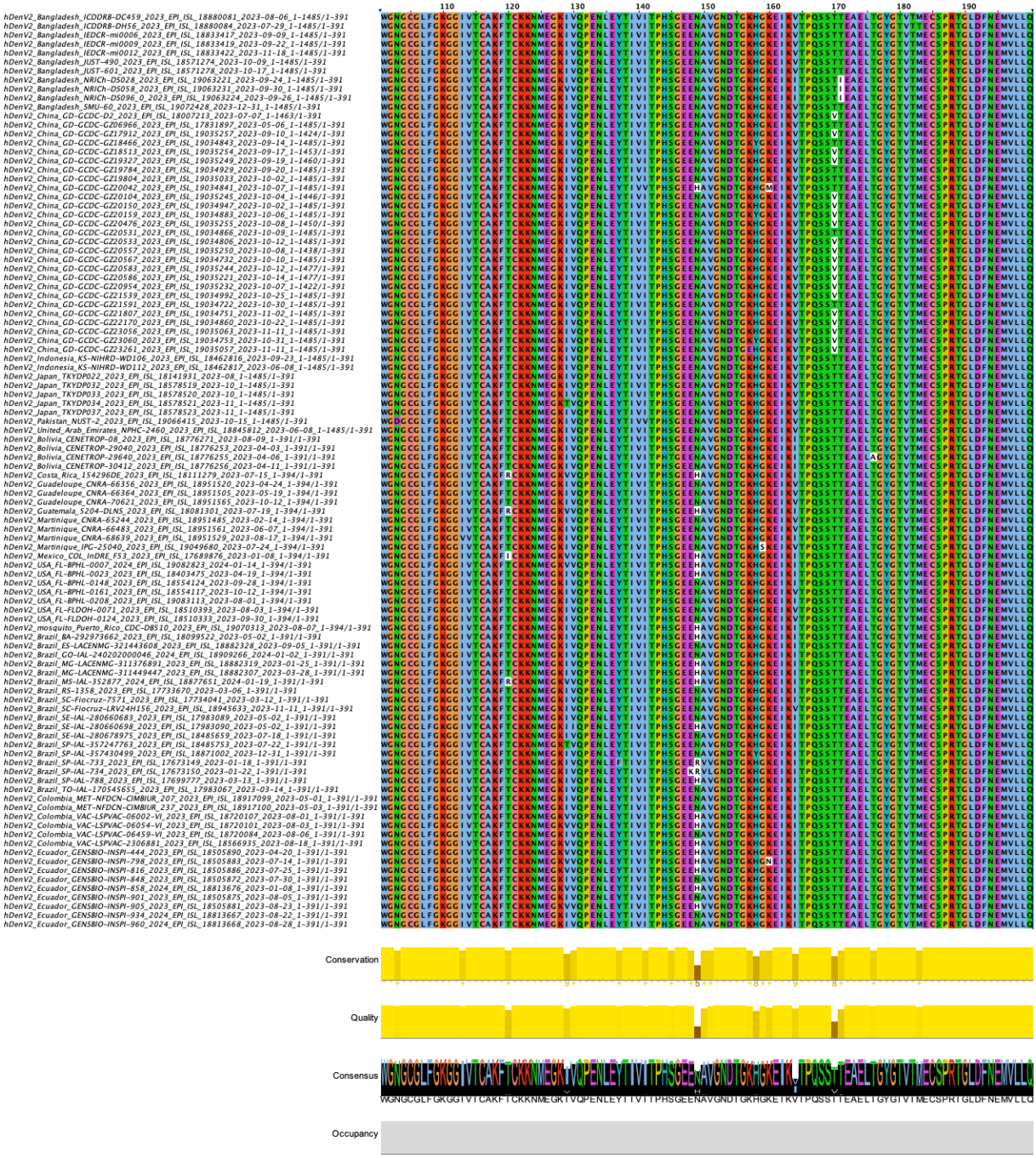
```
          geom = "rect", alpha = 0.07)+ geom_point(position = position_dodge(),
size = 4, aes(color = ID)) +  theme_classic() + scale_color_gradient2(midpoint =
50, low = "blue", mid = "gray94", high = "red", space = "Lab", name="Variant")  +
xlab('Ab') + ylab(expression(bold(Delta*Delta*G~(Kcal~mol^-1)))) + facet_wrap(~ id,
scales = "free_x")  + scale_y_continuous(expand = c(0, 0), limits = c(-4, 15))

r + theme(legend.position = "right", axis.title = element_text(size = 20, face =
'bold'), axis.text = element_text(size = 20, colour = "black", face = "bold"),
legend.text = element_text(size = 20, colour = "black"), panel.grid.major =
element_blank(),   panel.grid.minor   =   element_blank(),      axis.text.y   =
element_text(face  =  "bold",  color  =  "black",  size  =  20),    axis.text.x  =
element_text(face = "bold", color = "black", size = 20, angle = 90, hjust = 1, vjust
= 0.5), plot.title = element_text(color = "black", size = 20, face = "bold"))
```

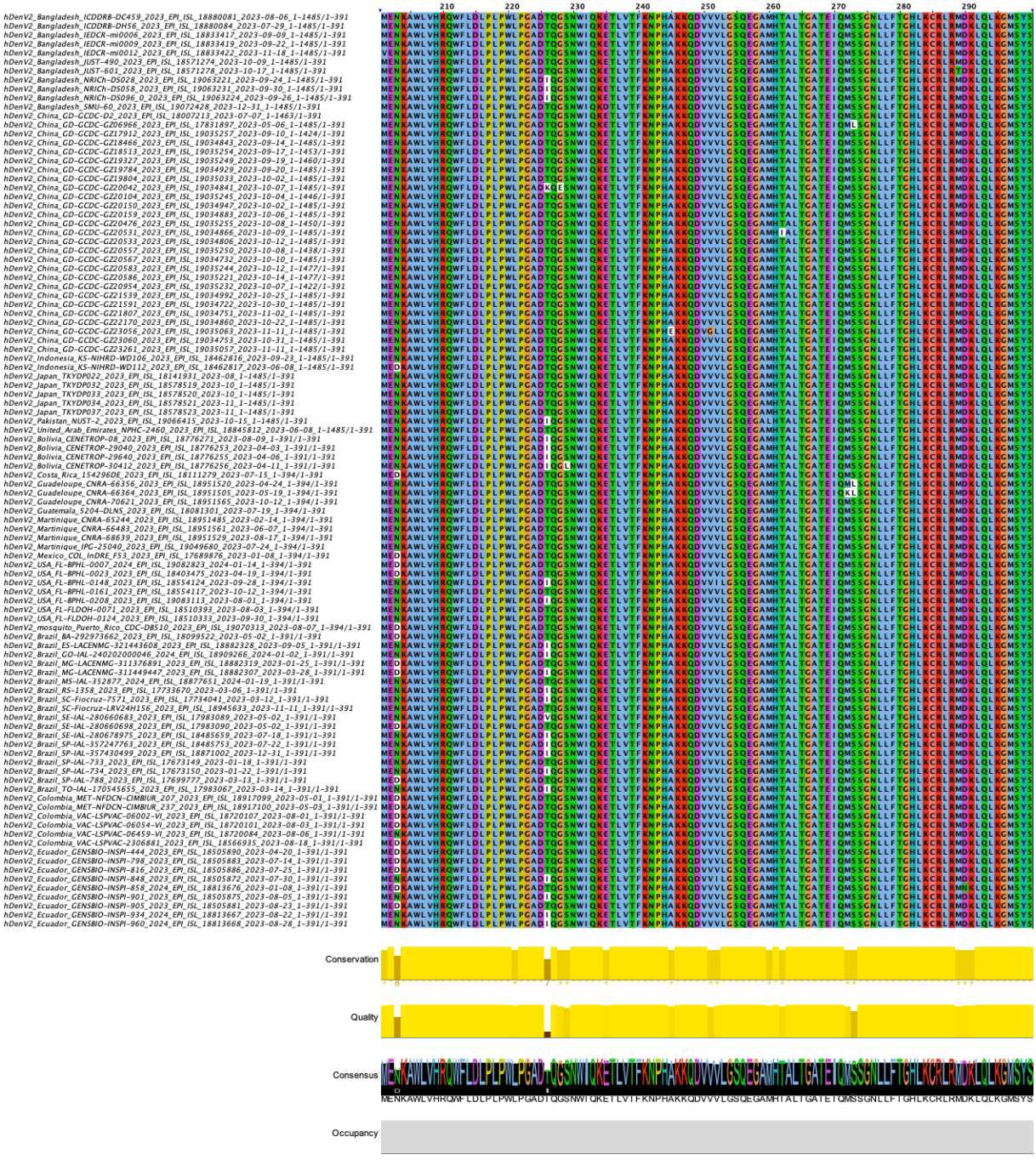
Script 6. Geração de gráficos no RStudio. Script em *R*.

II. Alinhamento de Sequências

As sequências alinhadas no programa JalView a partir dos 102 sequenciamentos genômicos que deram origem aos resfiles para as mutações do Rosetta se encontram representadas a seguir. A imagem foi dividida em quatro partes para facilitar a visualização.




110
120
130
140
150
160
170
180
190
hDenV2_Bangladesh_ICDDRB-DC459_2023_EPI_ISL_18880081_2023-08-06_1-1485/1-391
hDenV2_Bangladesh_ICDDRB-DH56_2023_EPI_ISL_18880084_2023-07-29_1-1485/1-391
hDenV2_Bangladesh_IEDCR-mi0006_2023_EPI_ISL_18833417_2023-09-09_1-1485/1-391
hDenV2_Bangladesh_IEDCR-mi0009_2023_EPI_ISL_18833419_2023-09-22_1-1485/1-391
hDenV2_Bangladesh_IEDCR-mi0012_2023_EPI_ISL_18833422_2023-11-18_1-1485/1-391
hDenV2_Bangladesh_JUST-490_2023_EPI_ISL_18571274_2023-10-09_1-1485/1-391
hDenV2_Bangladesh_JUST-601_2023_EPI_ISL_18571278_2023-10-17_1-1485/1-391
hDenV2_Bangladesh_NRICh-DS028_2023_EPI_ISL_19063221_2023-09-24_1-1485/1-391
hDenV2_Bangladesh_NRICh-DS058_2023_EPI_ISL_19063231_2023-09-30_1-1485/1-391
hDenV2_Bangladesh_NRICh-DS096_0_2023_EPI_ISL_19063224_2023-09-26_1-1485/1-391
hDenV2_Bangladesh_SMU-60_2023_EPI_ISL_19072428_2023-12-31_1-1485/1-391
hDenV2_China_GD-GCDC-02_2023_EPI_ISL_18007213_2023-07-07_1-1463/1-391
hDenV2_China_GD-GCDC-GZ06966_2023_EPI_ISL_17831897_2023-05-06_1-1485/1-391
hDenV2_China_GD-GCDC-GZ17912_2023_EPI_ISL_19035257_2023-09-10_1-1424/1-391
hDenV2_China_GD-GCDC-GZ18466_2023_EPI_ISL_19034883_2023-09-14_1-1485/1-391
hDenV2_China_GD-GCDC-GZ18513_2023_EPI_ISL_19035254_2023-09-17_1-1453/1-391
hDenV2_China_GD-GCDC-GZ19327_2023_EPI_ISL_19035249_2023-09-19_1-1460/1-391
hDenV2_China_GD-GCDC-GZ19784_2023_EPI_ISL_19034929_2023-09-20_1-1485/1-391
hDenV2_China_GD-GCDC-GZ19804_2023_EPI_ISL_19035033_2023-10-02_1-1485/1-391
hDenV2_China_GD-GCDC-GZ20042_2023_EPI_ISL_19034841_2023-10-07_1-1485/1-391
hDenV2_China_GD-GCDC-GZ20104_2023_EPI_ISL_19035245_2023-10-04_1-1446/1-391
hDenV2_China_GD-GCDC-GZ20150_2023_EPI_ISL_19034947_2023-10-02_1-1485/1-391
hDenV2_China_GD-GCDC-GZ20159_2023_EPI_ISL_19034883_2023-10-06_1-1485/1-391
hDenV2_China_GD-GCDC-GZ20476_2023_EPI_ISL_19035255_2023-10-08_1-1450/1-391
hDenV2_China_GD-GCDC-GZ20531_2023_EPI_ISL_19034866_2023-10-09_1-1485/1-391
hDenV2_China_GD-GCDC-GZ20533_2023_EPI_ISL_19034806_2023-10-12_1-1485/1-391
hDenV2_China_GD-GCDC-GZ20557_2023_EPI_ISL_19035250_2023-10-08_1-1438/1-391
hDenV2_China_GD-GCDC-GZ20567_2023_EPI_ISL_19034732_2023-10-10_1-1485/1-391
hDenV2_China_GD-GCDC-GZ20583_2023_EPI_ISL_19035244_2023-10-12_1-1477/1-391
hDenV2_China_GD-GCDC-GZ20586_2023_EPI_ISL_19035221_2023-10-14_1-1477/1-391
hDenV2_China_GD-GCDC-GZ20954_2023_EPI_ISL_19035232_2023-10-07_1-1422/1-391
hDenV2_China_GD-GCDC-GZ21539_2023_EPI_ISL_19034992_2023-10-25_1-1485/1-391
hDenV2_China_GD-GCDC-GZ21591_2023_EPI_ISL_19034722_2023-10-30_1-1485/1-391
hDenV2_China_GD-GCDC-GZ21807_2023_EPI_ISL_19034751_2023-11-02_1-1485/1-391
hDenV2_China_GD-GCDC-GZ22170_2023_EPI_ISL_19034860_2023-10-22_1-1485/1-391
hDenV2_China_GD-GCDC-GZ23056_2023_EPI_ISL_19035063_2023-11-11_1-1485/1-391
hDenV2_China_GD-GCDC-GZ23060_2023_EPI_ISL_19034753_2023-10-31_1-1485/1-391
hDenV2_China_GD-GCDC-GZ23261_2023_EPI_ISL_19035057_2023-11-11_1-1485/1-391
hDenV2_Indonesia_KS-NIHRD-WD106_2023_EPI_ISL_18462816_2023-09-23_1-1485/1-391
hDenV2_Indonesia_KS-NIHRD-WD112_2023_EPI_ISL_18462817_2023-06-08_1-1485/1-391
hDenV2_Japan_TKYDP022_2023_EPI_ISL_18141931_2023-08_1-1485/1-391
hDenV2_Japan_TKYDP032_2023_EPI_ISL_18578519_2023-10_1-1485/1-391
hDenV2_Japan_TKYDP033_2023_EPI_ISL_18578520_2023-10_1-1485/1-391
hDenV2_Japan_TKYDP034_2023_EPI_ISL_18578521_2023-11_1-1485/1-391
hDenV2_Japan_TKYDP037_2023_EPI_ISL_18578523_2023-11_1-1485/1-391
hDenV2_Pakistan_NUST-2_2023_EPI_ISL_19066415_2023-10-15_1-1485/1-391
hDenV2_United_Arab_Emirates_NPHC-2460_2023_EPI_ISL_18845812_2023-06-08_1-1485/1-391
hDenV2_Bolivia_CENETROP-08_2023_EPI_ISL_18776271_2023-08-09_1-391/1-391
hDenV2_Bolivia_CENETROP-29040_2023_EPI_ISL_18776253_2023-04-03_1-391/1-391
hDenV2_Bolivia_CENETROP-29640_2023_EPI_ISL_18776255_2023-04-06_1-391/1-391
hDenV2_Bolivia_CENETROP-30412_2023_EPI_ISL_18776256_2023-04-11_1-391/1-391
hDenV2_Costa_Rica_154296DE_2023_EPI_ISL_18111279_2023-07-15_1-394/1-391
hDenV2_Guadeloupe_CNRA-66356_2023_EPI_ISL_18951520_2023-04-24_1-394/1-391
hDenV2_Guadeloupe_CNRA-66364_2023_EPI_ISL_18951505_2023-05-19_1-394/1-391
hDenV2_Guadeloupe_CNRA-70621_2023_EPI_ISL_18951565_2023-10-12_1-394/1-391
hDenV2_Guatemala_5204-DLNS_2023_EPI_ISL_19081301_2023-07-19_1-394/1-391
hDenV2_Martinique_CNRA-65244_2023_EPI_ISL_18951485_2023-02-14_1-394/1-391
hDenV2_Martinique_CNRA-66483_2023_EPI_ISL_18951561_2023-06-07_1-394/1-391
hDenV2_Martinique_CNRA-68639_2023_EPI_ISL_18951529_2023-08-17_1-394/1-391
hDenV2_Martinique_IPC-25040_2023_EPI_ISL_19049680_2023-07-24_1-394/1-391
hDenV2_Mexico_COL_InDRE_FS3_2023_EPI_ISL_17689876_2023-01-08_1-394/1-391
hDenV2_USA_FL-BPHL-0007_2024_EPI_ISL_19082823_2024-01-14_1-394/1-391
hDenV2_USA_FL-BPHL-0023_2023_EPI_ISL_18403475_2023-04-19_1-394/1-391
hDenV2_USA_FL-BPHL-0148_2023_EPI_ISL_18554124_2023-09-28_1-394/1-391
hDenV2_USA_FL-BPHL-0161_2023_EPI_ISL_18554117_2023-10-12_1-394/1-391
hDenV2_USA_FL-BPHL-0208_2023_EPI_ISL_19083113_2023-08-01_1-394/1-391
hDenV2_USA_FL-FLDOH-0071_2023_EPI_ISL_18510393_2023-08-03_1-394/1-391
hDenV2_USA_FL-FLDOH-0124_2023_EPI_ISL_18510333_2023-09-30_1-394/1-391
hDenV2_mosquito_Puerto_Rico_CDC-DB510_2023_EPI_ISL_19070313_2023-08-07_1-394/1-391
hDenV2_Brazil_BA-292973662_2023_EPI_ISL_18099522_2023-05-02_1-391/1-391
hDenV2_Brazil_ES-LACENMG-3214443608_2023_EPI_ISL_18882328_2023-09-05_1-391/1-391
hDenV2_Brazil_GO-IAL-240202000046_2024_EPI_ISL_18909266_2024-01-02_1-391/1-391
hDenV2_Brazil_MG-LACENMG-311376891_2023_EPI_ISL_18882319_2023-01-25_1-391/1-391
hDenV2_Brazil_MG-LACENMG-311449447_2023_EPI_ISL_18882307_2023-03-28_1-391/1-391
hDenV2_Brazil_MS-IAL-352877_2024_EPI_ISL_18877651_2024-01-19_1-391/1-391
hDenV2_Brazil_RS-1358_2023_EPI_ISL_17733670_2023-03-06_1-391/1-391
hDenV2_Brazil_SC-Fiocruz-7577_2023_EPI_ISL_17734041_2023-03-12_1-391/1-391
hDenV2_Brazil_SC-Fiocruz-LRV24H156_2023_EPI_ISL_18945633_2023-11-11_1-391/1-391
hDenV2_Brazil_SE-IAL-280660683_2023_EPI_ISL_17983089_2023-05-02_1-391/1-391
hDenV2_Brazil_SE-IAL-280660698_2023_EPI_ISL_17983090_2023-05-02_1-391/1-391
hDenV2_Brazil_SE-IAL-280678975_2023_EPI_ISL_18885659_2023-07-18_1-391/1-391
hDenV2_Brazil_SP-IAL-357247763_2023_EPI_ISL_18485753_2023-07-22_1-391/1-391
hDenV2_Brazil_SP-IAL-357430499_2023_EPI_ISL_18871002_2023-12-31_1-391/1-391
hDenV2_Brazil_SP-IAL-733_2023_EPI_ISL_17673149_2023-01-18_1-391/1-391
hDenV2_Brazil_SP-IAL-734_2023_EPI_ISL_17673150_2023-01-22_1-391/1-391
hDenV2_Brazil_SP-IAL-788_2023_EPI_ISL_17699777_2023-03-13_1-391/1-391
hDenV2_Brazil_TO-IAL-170545655_2023_EPI_ISL_17983067_2023-03-14_1-391/1-391
hDenV2_Colombia_MET-NFDCN-CIMBIUR_207_2023_EPI_ISL_18917099_2023-05-01_1-391/1-391
hDenV2_Colombia_MET-NFDCN-CIMBIUR_237_2023_EPI_ISL_18917101_2023-05-03_1-391/1-391
hDenV2_Colombia_VAC-LSPVAC-06002-VI_2023_EPI_ISL_18720107_2023-08-01_1-391/1-391
hDenV2_Colombia_VAC-LSPVAC-06054-VI_2023_EPI_ISL_18720101_2023-08-03_1-391/1-391
hDenV2_Colombia_VAC-LSPVAC-06459-VI_2023_EPI_ISL_18720084_2023-08-06_1-391/1-391
hDenV2_Colombia_VAC-LSPVAC-23288-VI_2023_EPI_ISL_18566935_2023-08-18_1-391/1-391
hDenV2_Ecuador_GENSBIO-INSPI-444_2023_EPI_ISL_18505890_2023-04-20_1-391/1-391
hDenV2_Ecuador_GENSBIO-INSPI-798_2023_EPI_ISL_18505883_2023-07-14_1-391/1-391
hDenV2_Ecuador_GENSBIO-INSPI-816_2023_EPI_ISL_18505886_2023-07-25_1-391/1-391
hDenV2_Ecuador_GENSBIO-INSPI-848_2023_EPI_ISL_18505872_2023-07-30_1-391/1-391
hDenV2_Ecuador_GENSBIO-INSPI-858_2024_EPI_ISL_18813676_2023-01-08_1-391/1-391
hDenV2_Ecuador_GENSBIO-INSPI-901_2023_EPI_ISL_18505875_2023-08-05_1-391/1-391
hDenV2_Ecuador_GENSBIO-INSPI-905_2023_EPI_ISL_18505881_2023-08-23_1-391/1-391
hDenV2_Ecuador_GENSBIO-INSPI-934_2024_EPI_ISL_18813667_2023-08-22_1-391/1-391
hDenV2_Ecuador_GENSBIO-INSPI-960_2024_EPI_ISL_18813668_2023-08-28_1-391/1-391
Conservation
Quality
Consensus
Occupancy


210
220
230
240
250
260
270
280
290
Conservation
Quality
Consensus
Occupancy

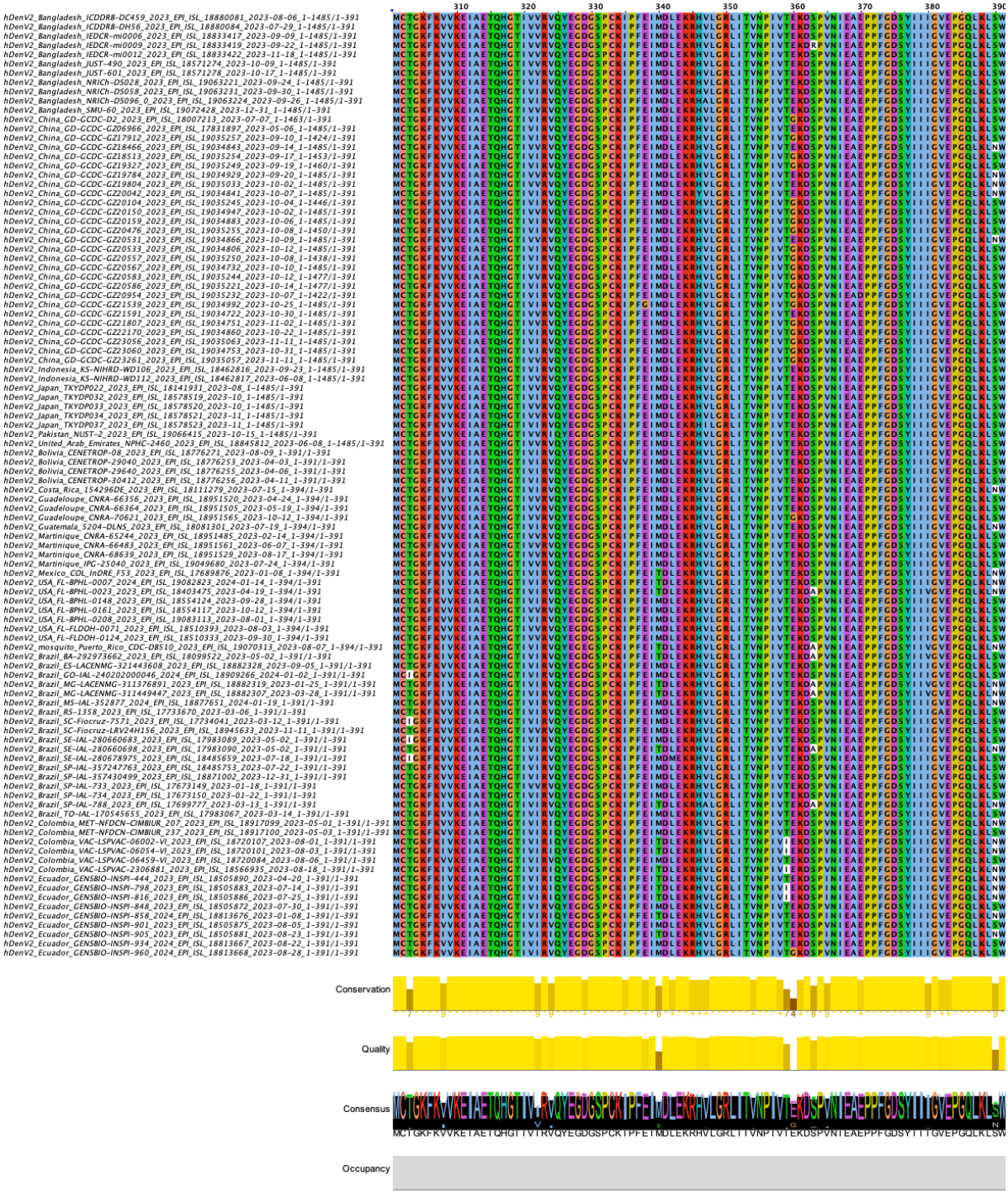


Figura 21. Alinhamento de Sequências no JalView. Representação em esquema de cores CLUSTAL das sequências de aminoácidos (estrutura primária) das 102 variantes proteicas utilizadas para modelagem das proteínas E de DENV2. A proteína E está incompleta pois as proteínas tiveram suas extremidades C e N terminal cortadas para que todas tenham o mesmo tamanho (391 resíduos), evitando erros não sistemáticos de cálculo. Das sequências obtidas da rede genômica da FIOCRUZ, apenas 102 foram utilizadas pois sequenciamentos redundantes (isto é, com mesma sequência) e aqueles contendo nucleotídeos não resolvidos dentro da zona de interesse (isto é, da região que codifica os 391 resíduos acima) foram excluídos, para aumentar a eficiência de cálculo e para evitar erros não sistemáticos, respectivamente.

III. Participação em Evento

Os resultados parciais do trabalho de Mestrado Acadêmico foram apresentados no evento "XI Escola de Modelagem Molecular em Sistemas Biológicos 2024" na forma pôster, intitulado "Caracterização Termodinâmica da Resposta Imune Humoral Contra a Proteína Envelope do Vírus da Dengue" entre 19 e 23 de agosto de 2024 no Laboratório Nacional de Computação Científica – LNCC/MCTI. Neste mesmo evento, houve participação nos minicursos abaixo:

- Dinâmica Molecular Básica.
- Conceitos e Aplicações de Aprendizado de Máquina.
- Cálculo de Energia Livre por Métodos Alquímicos Utilizando AMBER24.

IV. Período de Intercâmbio Acadêmico

Durante o mês de Outubro de 2024 foi realizado período de intercâmbio de pesquisa no *University College London* (UCL) – Reino Unido, no Laboratório de Biologia Computacional do Departamento de Biologia Molecular da UCL, sob orientação de Profª Drª Franca Fraternali (diretora deste departamento) e sob valioso auxílio de Dr. Carlos Cruz (pesquisador assistente deste departamento e egresso da PPGQUI - UFPE). Durante este período foram produzidos alguns dos resultados do Projeto de Doutorado aqui referenciado, e a Profª Drª Franca Fraternali aceitou orientação em período de doutorado sanduíche no futuro, caso haja abertura de edital respectivo do período e aprovação neste.

V. Nota de Imprensa

Pesquisa da UFPE investiga possibilidade de fuga imunológica no vírus da Dengue sorotipo 2 usando métodos computacionais

Você sabia que alguns vírus podem mudar ao longo do tempo e, com isso, "driblar" nosso sistema imunológico? Esse fenômeno é conhecido como **fuga ou evasão imunológica**. Este fenômeno é comum em vários vírus, como o vírus da gripe. Quando ocorre evasão imune, muitas vezes é necessário desenvolver novas vacinas para combater o vírus. Sempre se acreditou que isso não acontecesse com os quatro sorotipos de Dengue — o que sustentava a ideia de que uma pessoa só contrai cada tipo uma vez na vida. No entanto, o aumento inesperado de casos em 2024 e novos dados genéticos desafiaram essa visão e foi necessário investigar se o vírus da Dengue apresentaria evasão imune ou não.

Para investigar essa possibilidade, o mestrando **Júlio Cesar de Melo Simões**, sob orientação do professor **Roberto Dias Lins Neto**, no Programa de Pós-Graduação em Química (PPGQUI) da **Universidade Federal de Pernambuco (UFPE)**, desenvolveu um estudo inovador utilizando simulações computacionais. O objetivo foi testar se mutações na proteína de envelope (E) do vírus da Dengue tipo 2 – responsável por se ligar a anticorpos – poderiam levar à fuga da resposta imunológica. Este tipo de investigação é importante, pois a ocorrência de evasão imune pode diminuir a eficácia de vacinas (e, nesse caso, pode ser necessário desenvolver novas vacinas para aquele vírus específico).

A pesquisa utilizou três métodos distintos: dois métodos baseados em equações físico-químicas e um método baseado em aprendizagem de máquina (PBEE). Foram comparadas interações entre anticorpos e proteínas virais de diferentes épocas (2010–2014 versus 2023–2034), com dados genômicos (isto é, informações sobre o material genético dos vírus) fornecidos pela **Rede Genômica da FIOCRUZ**. Os resultados mostraram que os modelos físicos previram possíveis indícios de evasão, mas o modelo baseado em inteligência artificial — que mais se alinhou com resultados experimentais e apresentou maior acurácia no cálculo — **não confirmou essa hipótese**. Isso indica que, nas cepas analisadas, **as vacinas atuais ainda devem ser eficazes** e que, até o momento, só é possível contrair o vírus da dengue até quatro vezes ao longo da vida (uma vez para cada sorotipo). Além de ajudar a compreender melhor a evolução viral, o trabalho também apresenta uma **nova metodologia para prever riscos de fuga imunológica de forma rápida**, a partir de dados genéticos. Essa abordagem pode ser essencial para antecipar mudanças nos vírus e orientar estratégias de vacinação, especialmente em surtos futuros.

Este estudo é parte da **dissertação de mestrado de Júlio Cesar de Melo Simões**, sob **orientação do Professor Dr. Roberto Dias Lins Neto**, com **bolsa financiada pela FACEPE**. A pesquisa destaca a importância da ciência interdisciplinar para a saúde pública, unindo biologia, computação e vigilância epidemiológica em benefício da sociedade.

VI. Artigo de Revisão Publicado

Foi publicado o artigo de revisão "*Structure-based computational design of antibody mimetics: challenges and perspectives*" (doi:10.1002/2211-5463.13855) na revista *FEBS Open Bio* (Fator de Impacto 2,8 em 2023). O trabalho revisa o uso de miméticos de anticorpos para a engenharia de proteínas e teve papel importante nas primeiras etapas do Projeto de Doutorado, além de fundamentar princípios de engenharia de proteínas e de interações anticorpo-antígeno que foram de grande contribuição para o presente trabalho.

Os resultados do trabalho de mestrado foram escritos em formato de artigo e estão em fase de ajustes e revisão para a submissão. Os resultados do trabalho de doutorado ainda são parciais, não finais, e, assim, ainda não estão prontos para escrita e submissão.

Check for updates

FEBS openbio

FEBS PRESS science publishing by scientists

REVIEW

Structure-based computational design of antibody mimetics: challenges and perspectives

Elton J. F. Chaves[1], Danilo F. Coêlho[2], Carlos H. B. Cruz[3], Emerson G. Moreira[4], Júlio C. M. Simões[1,2], Manassés J. Nascimento-Filho[1,2] and Roberto D. Lins[1,2,4]

1 Aggeu Magalhães Institute, Oswaldo Cruz Foundation, Recife, Brazil
2 Department of Fundamental Chemistry, Federal University of Pernambuco, Recife, Brazil
3 Institute of Structural and Molecular Biology, University College London, UK
4 Fiocruz Genomics Network, Brazil

Keywords
de novo design; deep learning; machine learning; protein engineering; protein structure

Correspondence
R. D. Lins, Aggeu Magalhães Institute, Oswaldo Cruz Foundation, Recife, Brazil
E-mail: roberto.neto@fiocruz.br

Elton J. F. Chaves and Danilo F. Coêlho contributed equally to this article.

(Received 3 March 2024, revised 17 May 2024, accepted 19 June 2024)

doi:10.1002/2211-5463.13855

Edited by Claudio Soares

The design of antibody mimetics holds great promise for revolutionizing therapeutic interventions by offering alternatives to conventional antibody therapies. Structure-based computational approaches have emerged as indispensable tools in the rational design of those molecules, enabling the precise manipulation of their structural and functional properties. This review covers the main classes of designed antigen-binding motifs, as well as alternative strategies to develop tailored ones. We discuss the intricacies of different computational protein–protein interaction design strategies, showcased by selected successful cases in the literature. Subsequently, we explore the latest advancements in the computational techniques including the integration of machine and deep learning methodologies into the design framework, which has led to an augmented design pipeline. Finally, we verse onto the current challenges that stand in the way between high-throughput computer design of antibody mimetics and experimental realization, offering a forward-looking perspective into the field and the promises it holds to biotechnology.

Recent advancements in therapeutic antibody research have led to significant progress in both key technologies and theoretical innovations. This encompasses the development of antibody-drug conjugates, antibody-conjugated nucleotides, bispecific antibodies, nanobodies, and various other antibody derivatives. Furthermore, therapeutic antibodies have been effectively combined with technologies from other fields, giving rise to novel interdisciplinary applications, including cell-based therapies [1]. In fact, the biopharmaceutical industry is one of the most dynamic innovation and business ecosystems, with an estimated investment of hundreds of billions of dollars annually. In the United States alone, it accounted for 17% of dollars spent on domestic research and development (R&D) in the year of 2020, nearly doubling the investment on software development in the country [2]. Its main product,

Abbreviations
ΔΔG, binding free energy difference; AF2, AlphaFold2; AF3, AlphaFold3; AI, artificial intelligence; ANN, artificial neural network; dArmRP, designed armadillo repeat proteins; DARPIN, designed ankiryn repeat proteins; DL, deep learning; DPPM, denoising diffusion probability model; FN3, fibronecting type-III; GN, generative model; hACE2, human angiotensin-converting enzyme 2; IL-17A, interleukin-17A; kDa, kilodalton; mAb, monoclonal antibody; MC, Monte Carlo; ML, machine learning; MSA, multiple sequence alignment; MSE, mean square error; NMR, nuclear magnetic resonance; PDB, protein databank; R&D, research and development; RBD, receptor binding domain; RIF, rotamer interaction field; RMSD, root mean square deviation; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; SASA, solvent accessible surface area; VEGF-A, vascular endothelial growth factor A; VHH, variable heavy domain.

FEBS Open Bio (2024) © 2024 The Author(s). *FEBS Open Bio* published by John Wiley & Sons Ltd on behalf of Federation of European Biochemical Societies.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Table 1. Comparison of the main advantages and disadvantages between antibody mimetics and conventional antibodies.

	Main advantages	Main disadvantages
Antibody mimetics	Size control: typically, smaller and simpler in structure compared to conventional antibodies, which allows for better tissue penetration and potentially reduced immunogenicity	Achieving high affinity: compared to conventional antibodies, achieving binding affinities and target specificity is challenging, often requiring multiple rounds of design
	Engineering flexibility: mimetics can be engineered with specific properties tailored to their intended applications (e.g., enhanced stability and/or binding affinity)	Clinical validation: extensive validation and clinical track record are needed, compared to conventional antibodies, potentially raising efficacy and safety concerns
	Versatility of administration: mimetics can be designed to control the *via* of administration. They are often small enough to be orally administered, if desired, offering advantages in terms of patient convenience and compliance	Shorter half-life: mimetics may have shorter half-lives in circulation compared to conventional antibodies, requiring either more frequent dosing for therapeutic applications or fusing to other proteins to enhance their half-life
	Structural diversity: mimetics can be derived from various sources, including synthetic peptides, small proteins, or non-protein molecules, and even novel scaffolds, providing a wide range of options for development	High development cost: development of mimetics by experimental means involves great financial risk, historically being mostly undertaken by the private sector
	Production cost: they can be engineered to be produced in prokaryotes and to yield large quantities	Lack of effector function: mimetics do not carry the antibody constant fraction region
Conventional antibodies	High specificity: conventional antibodies, particularly monoclonal antibodies, exhibit high specificity for their target antigens, which minimizes off-target effects	Complex structure: conventional antibodies have a complex structure, making them expensive and challenging to produce at scale
	Natural recognition: they rely on the natural immune system's mechanisms for target recognition, ensuring biocompatibility	Immunogenicity: antibodies derived from non-human sources can provoke immune responses, potentially limiting their therapeutic use
	Versatility: antibodies can be modified and engineered for various applications, including therapeutics, diagnostics, and research tools	Limited tissue penetration: their large size can hinder tissue penetration, affecting efficacy in certain therapeutic applications
	Long half-life: IgG antibodies have a relatively long half-life in bloodstream, providing sustained therapeutic effects	Storage and stability: antibodies require specific storage conditions and can degrade over time, affecting their shelf life and efficacy
	Well-established production: large-scale production methods for conventional antibodies are well-established, facilitating manufacturing for commercial purposes	Production cost: it requires eukaryotic cell lines for production due to post-translational modifications

monoclonal antibodies (mAbs), can be designed to specifically target disease-causing molecules or cells, minimizing off-target effects. According to a report from Future Market Insights, the antibody therapy market in 2023 accounted for USD 235 billion and it is expected to reach USD 824 billion in the next decade. Most mAbs come from natural sources, offering biocompatibility advantage, and reducing the risk of adverse reactions when employed *in vivo*. They have been developed to treat a wide range of diseases, including cancer, autoimmune disorders, and infectious diseases. However, producing mAbs requires complex and highly specialized protein production technology, and its cost precludes population-wide use of this class of molecules.

The development of synthetic antibody-mimetics (proteins structurally not related to antibodies, but capable of exerting similar function) has been explored as an alternative to the limitations above. Unlike biopharmaceuticals, antibody mimetics offer simpler and scalable production *via* chemical synthesis or microbial fermentation. However, the development of an antibody mimetic typically required a significant investment in R&D. In addition to designing and optimizing novel structures (e.g., design target properties, engineer stability, solubility, and improve biocompatibility), validating their efficacy and safety *in vivo* may be challenging due to their novel and engineered nature, requiring extensive preclinical and clinical testing. Nevertheless, its versatility potential and cost of production are unmatched. A comparison of the main advantages and disadvantages of using antibody mimetics and conventional antibodies is summarized in Table 1. To date, nearly two-dozen scaffold classes of antibody-mimetics have been explored, in addition to a few tailored designs. Figure 1 illustrates the structure of the current most used scaffold classes, highlighting their binding domains. (As the focus of these review is on the computational design

FEBS Open Bio (2024) © 2024 The Author(s). *FEBS Open Bio* published by John Wiley & Sons Ltd on behalf of Federation of European Biochemical Societies.

22115463, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/2211-5463.13855 by CAPES, Wiley Online Library on [31/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

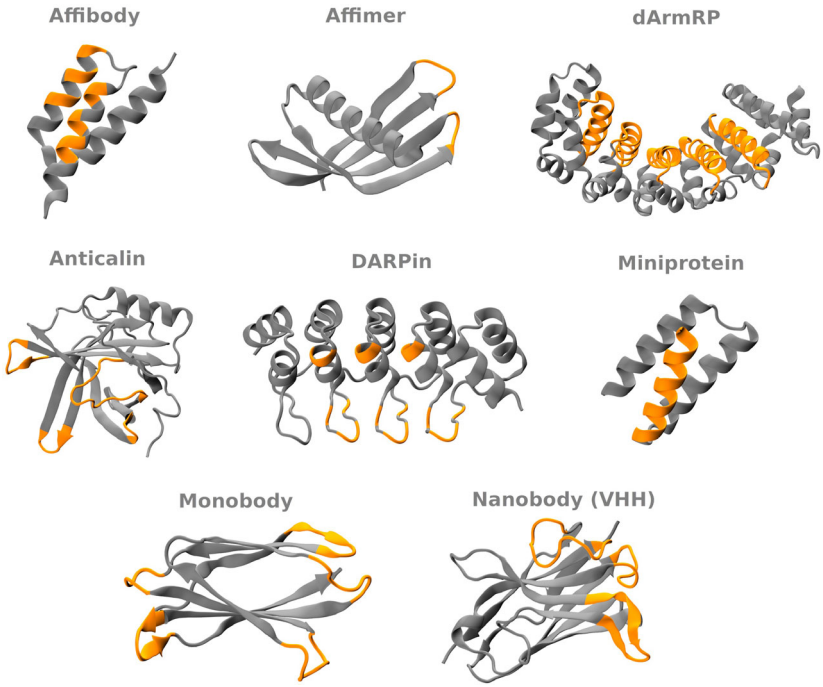


Fig. 1. Protein structure scaffolds of the main antibody mimetics. Structures are shown in cartoon model, where the framework and binding domains are represented in gray and orange, respectively. The accession codes for each structure in the PDB and the amino acids on the binding domain are the following: 8DA4 (Affibody), residues 9–11, 13, 14, 17, 18, 24, 25, 27, 28, 31, and 32; 4N6T (Affimer), residues 60–71, and 98–100; 5AEI (dArmRP), residues 65–83, 108–126, 149–168, 192–210, 233–252, 91–101, 115–121, and 141–156; 1N0S (Anticalin), residues 31–43, 62–65, 90–93, and 117–123; 2XEE (DARPin), residues 43, 45, 46, 48, 56, 57; 7S5B (Mini-protein), residues 1–16; 1TEN (Monobody), residues 813–818, 827–831, 840–846, 862–867, and 877–882; 1I3V (Nanobody), residues 26–32, 52–62, and 105–116.

approaches for antibody mimetics, we suggest the review by Yu and colleagues for a more in-depth biomedical applications for these molecules) [3].

As the field of computational protein design developed, these techniques have been employed as a way of speeding up the achievement of suitable structural properties (such as those mentioned above), thus reducing R&D-related costs, especially those associated to the early stages of development. These methods have been mainly used to leverage the potential of already validated classes of antibody mimetics by harnessing, *in silico*, all possible sequences that fit the desired function criteria. However, the recent association of AI technology to computational protein design now allows the development of novel binders that do not rely on predetermined protein templates. It unfolds an unprecedent potential for exploration of the flourishing field beyond nature's protein portfolio.

Main classes of antibody mimetics

Affibody

Based on the B-domain of staphylococcal protein A, it has a molecular weight of ca. 6 kDa. Affibodies are designed to bind to specific target molecules, such as proteins or peptides, with high affinity and specificity. Advantages over conventional antibodies include smaller size, simpler structure, and ease of engineering. Affibodies have applications in various areas including diagnostics, imaging, drug delivery, and targeted therapy [4]. Izokibep, an affibody-based biopharmaceutical, was shown to bind to and inhibit the activity of IL-17A in *in vitro* and *in vivo* assays using a murine model. It was also found to be safe and well-tolerated in phase I and II clinical studies for the treatment of psoriatic arthritis [5].

Affimer

Previously known as Adhiron, its scaffold is based on the human protease inhibitor, Stefin A [6]. Affimers have a molecular weight of ca. 11 kDa, and their structure contains four β-sheets, one α-helix, and two variable loops. These loops consist of nine amino acids each, used to design binding interfaces for a desired target. Up to date, affimers have been mainly designed by molecular or directed evolution techniques, where a diverse library of potential binding proteins is created and screened for those with the desired properties. They have been used as molecular probes for studying protein interactions, as diagnostic tools for detecting biomarkers or pathogens, and as therapeutic agents for targeting specific molecules involved in diseases such as cancer or inflammatory disorders [6,7].

dArmRP

Designed Armadillo Repeat Proteins (dArmRPs) have an armadillo domain, consisting of sequential

22115463, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/2211-5463.13855 by CAPES, Wiley Online Library on [31/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

armadillo repeats (8–12 internal repeats), each containing approximately 42 amino acids. They vary in molecular weight by 39 and 58 kDa. Each repeat consists of three α-helices, designated as H1, H2, and H3. Inserted between the N and C terminus, the helical repeats protect the hydrophobic core from exposure to the solvent. They have been computationally designed to recognize and bind to peptide ligands, overcoming the limitation of antibody specificity upon peptide flexibility. Other uses include drug delivery, and molecular imaging [8–10].

Anticalin

Derived from lipocalins, a family of naturally occurring proteins that typically bind to small hydrophobic molecules, anticalins were created through protein engineering techniques to have binding sites tailored for specific targets, such as drugs, metabolites, or other molecules of interest. Their structure consists of a cup-shaped pocket weighing about 20 kDa. Biomedical applications include the delivery of biopharmaceuticals across the blood–brain barrier [11], and theranostic applications [12].

DARPIN

Proteins composed by 33 amino acids ankyrin repeat motifs, arranged into two linked α-helices in opposite directions, and connected to the subsequent repeat through an elongated β-turn. DARPins are typically synthesized through combinatorial protein design using libraries comprising two to three ARPs repeated motifs. These motifs are sandwiched between positively and negatively charged N- and C-terminal caps, typically incorporating six random positions within the β-turn and the first α-helix of each repeat. As the number of monomers can vary, DARPINs will have a minimum molecular weight of 14 kDa. The scaffold has been designed for several applications, such as antivirals [13–15], and cancer treatment. The most progressed DARPin compound in clinical development is Abicipar pegol, an antagonist of VEGF-A. It is currently undergoing phase III trial to explore its potential to treat ophthalmic conditions including neovascular age-related macular degeneration and diabetic macular edema [16].

Miniprotein

This is a case of a tailored *de novo* design conceptualized by the group of David Baker [17,18]. These small proteins are typically formed by fewer than 50 residues. Despite their small size, they can fold into three-dimensional structures, often 3- or 4 helix bundles, with a molecular weight ranging from 5 to 10 kDa. As showcase, the Baker Labs developed designs against the RBD (receptor binding domain of spike protein) of SARS-CoV-2, with affinities ranging from 100 pM to 10 nM, and able to block virus infection *in vitro* [19].

Monobody

Based on the human fibronectin type III domain (FN3), this scaffold has an immunoglobulin-like fold, with a molecular weight ca. 10–15 kDa. Monobodies lack disulfide bonds, and thus, they are particularly suited as genetically encoded reagents to be used intracellularly [20], while the small and simple structure of monomeric monobodies confers increased tissue distribution. When designed in a bead-on-a-string-like assembly, multiple domains of FN3 can bind to different targets, overcoming the multi-specificity challenge of conventional antibodies. Furthermore, full-length fibronectin can fold into multiple conformations as part of its natural function, providing structural and sequence versatility to monobodies [21], with affinity and specificity that rival those of antibodies [22,23].

Nanobody

Nanobodies, also known as VHH or single-domain antibodies, are a class of antibody fragments derived from the variable region of heavy-chain antibodies found in camelids, such as camels, llamas, and alpacas. The proteins consist of a single monomeric antibody domain. With a molecular weight around 12–15 kDa, it makes them one-tenth the size of conventional antibodies. Despite their small size, nanobodies retain high specificity and affinity for their target antigens, making them valuable tools in various biomedical applications. While they are generally obtained by animal immunization or phage display techniques, computer design of nanobodies have recently become popular [24]. *In silico* affinity maturation has also been used to improve the thermal stability and binding affinity of natural nanobodies [25].

Computational protein design as the foundation for antibody mimetic development

The design of antibody mimetics is based on protein engineering principles. Our understanding of protein structure and function has matured significantly since the

FEBS Open Bio (2024) © 2024 The Author(s). *FEBS Open Bio* published by John Wiley & Sons Ltd on behalf of Federation of European Biochemical Societies.

22115463, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/2211-5463.13855 by CAPES, Wiley Online Library on [31/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

groundwork of Linus Pauling and Francis Crick in 1950s, allowing us to design proteins with specific properties. Protein engineering techniques can be classified as empiricism-based (e.g., directed evolution, phage display) or mechanism-based (e.g., comparative modeling, *de novo* design). Although purely experimental designs have been largely successful, they are cost and labor-consuming, and the lessons learned are usually not applicable to unrelated systems [26]. In addition, evolution explores limited protein sequence space, leading to clustered natural protein families [27]. *De novo* design allows exploration of broader sequence space, leveraging protein biophysics principles.

Computational protein design methods are based on thermodynamics principles and biological observations, and they can be used for practically all classes of proteins [27,28]. However, these methods rely on Anfinsen's Thermodynamic Hypothesis, that is, proteins fold into the lowest energy states that are accessible to their amino acid sequences [29]. In the last decade, rational design based on computational modeling and structural analysis emerged as a powerful strategy to engineer proteins with enhanced stability, activity, and/or specificity. It is based on the assumption that the geometry of a protein, together with the specific presentation of charges and molecular groups on its surface, determine its function. A given sequence of amino acids (primary structure) often leads to a specific three-dimensional structure. On the other hand, different combinations of residues with similar properties can also lead to the same final topological structure. Protein design aims to determine an amino acid sequence that will fold into a three-dimensional structure to perform a specific function. Thus, the key points are the configuration sampling method and the energy function used to predict stability and binding for searching the lowest energy model.

The advancement in rational protein design has historically been related to the progress of structure prediction methods and the increase in the number of experimentally determined target structures. Until three decades ago, most of computational techniques were limited to modifications through site-directed mutagenic assays at binding sites, surfaces, and interfaces of well-defined and untouched frameworks. This scenario has changed since 2003 after design of a novel protein from scratch called Top7. Using Monte Carlo search with molecular force fields and scoring functions, a novel protein was designed with unprecedented topology at the time [30]. To achieve the final model, several sequence design iterative cycles and backbone optimizations were performed. X-ray crystallography and NMR resolved the structure of the unnatural 93-mer α/β fold protein. Comparison with the model showed a backbone RMSD of approximately 1 Å, and the protein also exhibited remarkable thermodynamic stability [30]. That provided Top7 with a rich work portfolio that has vast implications on the most diverse areas of medicine and biotechnology as an ultra-stable scaffold [31,32]. Since then, structure-based protein design has been marked by exceptional advances and significant increase in the number of proteins designed with high levels of complexity. However, we are just now experiencing a further substantial advance in the field. The recent use of neural networks on computational protein prediction (e.g., AlphaFold2 [33] and RoseTTAFold [34]) have allowed solving large protein structures with atomic precision and remarkable rapidness, overcoming half a century of challenges. These advances have paved ways for the development of robust, yet efficient, sampling algorithms, and sophisticate design methods.

Design of antibody mimetics

While protein prediction methods have matured in the last few years, protein complex prediction has lacked behind. Success is highly dependent on the strategy to estimate parameters that affect the binding affinity. Predicting association strength and complex structure accurately often requires combined techniques and experimental data guidance [35]. This is likely based on the electrostatic diversity of protein–protein interactions. Algorithms were initially designed to fold proteins, which almost invariably are formed by a hydrophobic core and a mostly hydrophilic solvent accessible area. In contrast, protein–protein interactions are seldom characterized by hydrophobic contacts only. Designing antibody-mimetics considering target epitope molecular signature yields higher success rates than postdesign optimization. Based on that strategy, a number of target-oriented computer protein design methods have been used to the development of antibody mimetics. The main techniques are discussed below, and a schematic workflow of each technique is shown in Fig. 2.

Docking and design

When designing a new protein, sometimes the goal is just to modify positions in the amino acid sequence of an already known protein structure complex. In this case, the template is the native protein complex itself, and the design is performed on top of it. This method is commonly used to design proteins with improved binding affinities to other proteins or ligands [36]. Alternatively, key-interacting residues in the antigen

22115463, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/2211-5463.13855 by CAPES, Wiley Online Library on [31/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

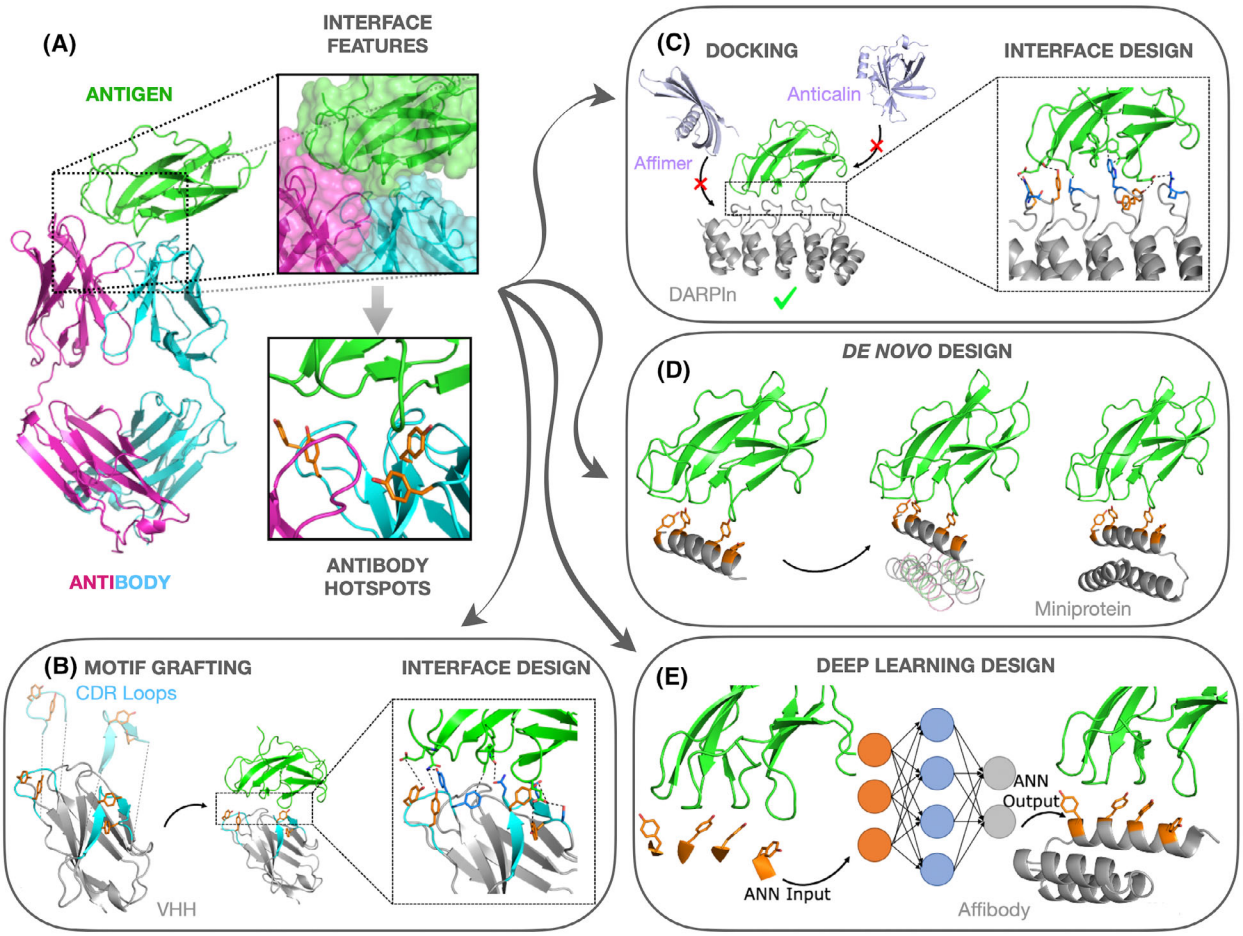


Fig. 2. Computational workflow strategies for the design of antibody mimetics. (A) Initial step consists of obtaining the structural and physicochemical characteristics of the antigen target interface (or antigen–antibody interface), and if possible, to identify key-residues for recognition. This information will be selectively passed on the design strategy used. (B) The Motif Grafting method requires structural data of antigen–antibody interface. Selected regions of the paratope that are considered important to recognition will be attempted to be grafted onto a variety of scaffolds. A scaffold that can withstand grafting will have its interface and topology sequence optimized to confer the desired affinity and protein stability, respectively. This example shows the grafting of CDR from conventional antibodies onto a VHH. (C) The Docking & Design method requires structural information from the antigen only. Scaffolds from a pre-triaged structural library will be attempted to dock onto the antigen target surface, considering their shape complementarity. Once the two proteins are docked, mimetic will undergo interface optimization to achieve high affinity, and subsequently optimization of its scaffold (where the pre-optimized interface sequence is kept fixed) to confer protein stability. It is exemplified here by attempting to dock the affimer, anticalin and DARPin mimetic classes. (D) *De Novo* Design creates a topology from scratch. It uses only structural data from the antigen only. However, a higher chance of success is achieved when antigen–antibody structural data is available. In the latter, a structural motif carrying the key-interacting residues of the antibody is used as a starting point. Proteins are designed to vehicle the chosen motif, as close as possible from its native form. The example shows the building of a miniprotein. (E) As in the *De Novo* Design, one can use either structural data from the antigen only or antigen–antibody, if available. The example depicts the generation of an affibody using ANN, that uses as starting point selected key-interacting residues from the antibody. Finally, it is worth noting that, even when key-interacting residues from the antibody are used as starting point, they can be optimized (mutated) to generate a molecule with enhanced binding affinity (all presented examples are merely illustrative and therefore do not belong to any published study).

interface are selected as starting point for the design of the new protein (antibody mimetic). The idea is to search high-resolution protein structures to be used as scaffolds with complementary shapes to antigens. Then, engineer interaction interfaces using local docking and designing steps. As in protein folding, protein association is driven by energy minimization, induced by van der Waals interactions, hydrophobic effect, electrostatic interaction, hydrogen bonding, shape, and chemical complementarity of the

FEBS Open Bio (2024) © 2024 The Author(s). *FEBS Open Bio* published by John Wiley & Sons Ltd on behalf of Federation of European Biochemical Societies.

22115463, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/2211-5463.13855 by CAPES, Wiley Online Library on [31/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

interaction partners [37]. Sequence optimization is performed every cycle, and a scoring function is used to find residues that stabilize the interaction and improve the binding propensity. Subsequently, the interaction interface sequence of the antibody mimetic remains fixed. Sequence optimization follows for remaining scaffold residues to ensure stability and solubility. A similar strategy to design interface is the so-called hotspot-centric approach, which consists of docking disembodied residues, selecting suitable scaffolds displaying residues at similar positions of the hotspots, and refining the interface [38,39]. This approach has recently been used to design a biopharmaceutical targeting the conserved fusion loop region of the Envelope proteins of flaviviruses. The protein neutralized infections by Zika and Dengue serotype 1 and 2 viruses *in vitro*, with an EC_{50} on pair of human monoclonal antibodies [39]. Another alternative is the rotamer interaction field (RIF) docking, which searches for hotspots for protein interface interaction from scratch (*de novo*). Disembodied residue conformations are docked onto the target interface to create favorable hydrogen bonds and hydrophobic interactions [40]. Scaffolds that displays residues at similar positions are subsequently used as starting point for protein design. Residues onto the scaffold are mutated by the residues outputted by RIF docking, followed by interface and scaffold optimizations.

Motif grafting design

Knowledge of the antigen–antibody interface structure is required as a starting point for the motif grafting method. Typically, antibody loops or selected regions of the interacting interface are grafted onto a protein scaffold [41]. The method consists of the following steps: define the motif to be transplanted, structurally align the motif to a putative carrier protein to determine the best region for grafting, transplant the motif, and redesign the scaffold protein around the grafted motif to ensure folding and stability of the chimeric protein. After the three-dimensional alignment, one can choose between keeping the backbone scaffold and grafting only the side chains of the antibody (side-chain grafting) or discarding the structural elements of the carrier, replacing it completely with the structural motif of the antibody (backbone-grafting). A flexible-backbone remodeling is employed to optimize the conformation of the protein skeleton after its modification [42]. Nevertheless, motif grafting by side-chain or backbone replacement faces limitations when the motif is too complex to find a structurally compatible protein scaffold. Proteins sporting a single immunoglobulin domain motif (e.g., VHH, selectins, ankyrin repeat proteins) are more commonly used with this technique.

De novo design

Template-based techniques have found several critical restrictions, imposed by strict adjustments of motifs onto scaffolds as well as the inconvenient requirement for well-defined and suitable frameworks. Comparative methods rely on known homologous structures, limiting design to pre-existing interfaces and excluding exploration of new sites [37]. To address the design of antibody mimetics to more complex epitopes, *de novo* methods represents a powerful alternative. It has been used to design new proteins around a given motif/epitope [43,44].

De novo proteins began to be designed around binding motifs using minimal knowledge of their interactions, but without prior knowledge of scaffold atom-positions. At first, these methods were considered *ab initio* or template-free techniques, as they did not use any known structure as a base. As techniques have become more sophisticated, the boundaries between different categories have become less clear [45]. There are methods today that are considered hybrids between template-free, and template-based. In *de novo* design, the target sequence is not compared with known proteins but to a blueprint of α-helix and β-sheet fragments, creating low-energy structures [27]. Folding is based on the basic principles of the diffusion-collision model developed by Karplus and Weaver [46], where once a local thermodynamically favorable interaction is formed, it is maintained, creating a bias by bringing together other contacts into spatial proximity. The process is repeated until the complete folding of the protein. Generally, pipelines are conceived through three steps: (a) The core is built first by an assembling process starting from valine-based secondary structures in the presence of the target; (b) the sequence is subsequently tailored on fixed backbone, taking into account the chemical environment (which also includes the antigen interface) and amino acid occurrence in secondary structure elements through layer-based design approaches; and (c) the sampling of possible backbone conformations is done using a library of short amino acid fragments based on the distribution of homologous structures taken from the PDB [45]. The fragment library covers accessible local structures, while side chains are assembled from a rotameric library. The final structure is assembled using Metropolis/Monte Carlo [47,48]. Binding affinity is achieved through successive interface design calculations at the

22115463, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/2211-5463.13855 by CAPES, Wiley Online Library on [31/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

interface with the aim of maximizing the number of buried hydrogen-bonds and creating hydrophobic contacts as good as commonly found in native protein–protein complexes [49]. Although binding free energy would be expected to be the most relevant feature, due to the empiricism nature of the scoring functions, other interface variables must be considered for a proper and more realistic interface description. Interfaces are evaluated by comparing the difference in binding free energies (ΔΔG), solvent accessibility surface area (SASA), shape complementarity, number of total and unsaturated hydrogen bonds, and ΔΔG/SASA of each decoy to the pool of generated structures. Although these energetic quantities do not find adherence in reality and therefore cannot be directly related to experimental measurements, they are crucial to triage the best candidate out of several thousand, or even a few million, generated designs. We highlight as showcase a work from the Baker's group that have recently harnessed the power of *de novo* methods to design a potent antibody mimetic against infection by the SARS-CoV-2 virus [19]. Miniproteins were developed to specifically block the interaction of human angiotensin-converting enzyme 2 (hACE2) to the receptor binding domain (RBD) from SARS-CoV-2 Spike protein, and thereby preventing the entrance of the virus into host cells. They showed that two selected candidates, from several thousand designs, prevented virus entry into host cells, and prevented lung disease and pathology in mice [50].

Machine and deep learning–driven design

Predicting binder-target strengths with experimental precision is crucial for picking out the best candidates. Nevertheless, accurate calculation of the absolute binding free energies for protein–protein interactions has posed a challenge in protein design methods. Historically, calculation of protein–protein free energies at experimental accuracy were restricted to enhanced sampling methods, which are highly demanding in terms of computational requirements. The advent of artificial intelligence (AI) in the most diverse areas of knowledge has inspired the development of new approaches with great potential for engineering and triage protein–protein complexes without prior structural knowledge, at exceptional performance. In the last few years, the community has resourced from machine and deep learning algorithms to reweigh the molecular feature contributions to experimental binding free energies. The approach has allowed for similar accuracy to more costly methods, at an unprecedent efficiency, making it possible to triage high-affinity binders, from thousands or even millions of candidates, in a high-throughput fashion [51–53]. In addition, AI has also been used to design antibody mimetic scaffolds. It provides advantage over template-based and *ab initio* methods. The former cannot predict new protein folds. While the latter may provide a greater range of new topology possibilities, it requires extensive sampling that results in a higher computational cost.

AI refers to systems or machines that imitate human intelligence. More specifically, machine learning (ML) and deep learning (DL) are subsets of AI that focus on building or improving predictive models that learn from data or identify informative groupings within data. ML has been used to improve existing antibodies, but most developed algorithms rely on deep sequencing or deep mutational scans for training data, and a shortcoming of these methods is that they are often specific classes of antibodies and applying them to other antibodies would require training with new data [54]. Few examples have been able to generate new antibodies and antigens without antibody-specific sequencing data [55–57].

Among AI methods, artificial neural network (ANN) and generative models (GM) have gained a lot of attention in the development of methods for structural biology, whether in the description of interactions between biomolecules or for structure modeling [58–62]. Neural networks, a set of DL techniques, are mathematical models that mimic the connectivity and behavior of neurons in the brain. Artificial neurons, which are the building blocks of ANNs, are simply mathematical functions that convert inputs to outputs in a specific way. To create an ANN, artificial neurons are organized into layers, with the output of one layer being the input of the next. On the other hand, GM is designed to generate new data instances that resemble the training data they were trained on. They learn the underlying structure of the data and then use this knowledge to generate new samples. Among the ANN and GM models available in the literature for protein modeling, two stand out: AlphaFold2 [33] and RFdiffusion [34].

AlphaFold2 (AF2) predicts protein structures by leveraging neural networks and training procedures based on evolutionary, physical, and geometric constraints [33]. The core of AF2's architecture is the Evoformer block, understanding how different parts of the protein are related to each other in 3D space and represents the protein as a graph where each amino acid is a node and the connections between them are edges. Evoformer uses two input representations: the Pair Representation and the Multiple Sequence Alignment

FEBS Open Bio (2024) © 2024 The Author(s). *FEBS Open Bio* published by John Wiley & Sons Ltd on behalf of Federation of European Biochemical Societies.

22115463, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/2211-5463.13855 by CAPES, Wiley Online Library on [31/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

(MSA). The former captures the relationships between pairs of amino acids in a matrix describing how two amino acids interact to each other, while the MSA matrix represents homologous sequences, identifying positions prone to simultaneous mutations (coevolution) and potential contact points. Evoformer updates these representations through 48 blocks, ensuring accurate 3D structure representation by applying geometric constraints and using axial attention to focus on critical information. This continuous information exchange between the MSA and pair representations enables precise structure predictions. AF2 achieved top performance in CASP14 with a median backbone accuracy of 0.96 Å RMSD and an all-atom accuracy of 1.5 Å RMSD [33]. AF2 was also adapted to predict multichain complexes (AlphaFold-Multimer) [63] and benchmarked for antibody–antigen interactions, achieving 43% top-ranked results for various protein complexes [64]. Recently, AlphaFold3 (AF3) was released [65], featuring a diffusion-based architecture capable of predicting structures containing proteins, ions, nucleic acids, and small molecules, with improved accuracy for antibody–antigen complexes. AF3 replaces Evoformer with a simpler module and predicts atom coordinates using a diffusion module. The MSA is processed in four blocks and incorporated into the pairwise representation, which is then processed through 48 blocks in the new Pairformer module. Following, the diffusion module refines the initial atom coordinates through a denoising process. AF3 surpasses classical docking tools and shows enhanced accuracy for protein complexes, including antibody-protein interactions [65].

The RFdiffusion method replaces the physical-based Rosetta methods to DL approaches, aiming to predict diverse proteins and scaffold-free binder interactions with atomic accuracy and unprecedented success [34]. RFdiffusion is an updated version of RoseTTAFold [66] that uses denoising diffusion probability models (DPPMs) to generate low-resolution backbone models, and then use the ProteinMPNN network [58] to subsequently design sequences encoding these structures. Binders are designed similarly to creating photorealistic images from textual instructions, resulting in novel proteins with higher binding potential and experimental success. The denoising process acts on a random sample of residue backbone, through an interactive DL-based design workflow, which disrupts coordinates toward true proteins, by minimizing the mean square error (MSE) to design the sequences. The authors demonstrated that RFDifusion is capable of generating binders for proteins used as target context, by selecting input residues in the target chain (defined as hotspots) to which the designed chain binds. The proof of concept was carried out with the target proteins Hemagglutinin Influenza A H1, Interleukin-7-α Receptor, Programmed Death Ligand 1, Insulin Receptor and Tropomyosin Kinase A Receptor, showing the potential of RFdiffusion for binders designing [34].

More recently, a new generative AI model was released to the public for use in structural biology. The 310 CoPILOT model was developed by 310 AI as an AI Chat for Designer Bio (https://310.ai/). This model allows you to perform tasks in structural biology through a web-based chat platform, making use of third-party tools such as search for and load proteins from the UniProt database, compare proteins using the TM-Align method [67], fold proteins using the ESM Fold method [68], and design with ProteinMPNN model [58]. Furthermore, it also allows the use of 310.ai's own algorithm for designing new proteins. For binder design, the model allows the user to fold an antibody mimetic and then dock it with the target structure. Up to date, it is a tool still under development that requires a considerable amount of work to deliver its promises.

Challenges and perspectives

The recent progress on the computer engineering of antibody mimetics has eased, but not eliminated the challenges that stand between direct design and experimental application. Successful design of antibody mimetics often requires several rounds of experimental validation. Through this process, it is possible to identify the required changes in the design or computational protocol to fine tune the structural changes needed to achieve its desired biological function [45].

In computational protein design, the quality of the final model depends on the efficiency of sampling and on the accuracy of the energy function [28]. Commonly, the configuration sampling method is a time-independent stochastic process, such as Monte Carlo (MC) [69]. However, the stochastic nature of the method leads to not sampling every regional energy minima [70]. The free energy of binding, which is directly related to the binding affinity, is the most important indicator of a protein-binding strength, and the most challenging to predict. The energy function is usually based on classical molecular mechanics forcefields associated with other empirical terms that employ simplified interaction potentials, offering computational speed but at the expense of a certain degree of accuracy [27,70]. Another well-known limitation for computational protein design is the solvent treatment.

22115463, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/2211-5463.13855 by CAPES, Wiley Online Library on [31/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Water molecules are important for the structure, stability, dynamics, and function of proteins [71] and generally should not be treated as an implicit interaction agent due to importance of desolvation penalty. The most frequent reasons for failure in protein design are insolubility and the formation of unintended oligomeric states. Proteins that bind to other proteins usually have hydrophobic residues on their surface, which may lead to unanticipated intermolecular hydrophobic interactions and aggregation. Increasing the robustness of designs will require improvements in the accuracy of the energy function not only for free energy binding, but also thermodynamic stability of monomeric proteins [27]. In addition, for such calculations to be useful in protein–protein binding discovery, where it is common to produce *in silico* millions of candidates, the predictions must be rapidly computed, preferably within at most a few hours, and they should also be accurate and reproducible.

In the last few years, AI methods have found use in all aspects of protein design, from weighting scoring functions and sampling space enhancement to the design process itself. These models can achieve high accuracy when predicting binding free energies. However, this accuracy is highly dependent on the quality of the experimental data used for training [72,73]. In addition, while predicting binding free energies is key to design antibody mimetics, other aspects of developability need to be addressed, which include selectivity, stability, aggregation prevention, solubility, biocompatibility, deimmunization, bioavailability, and clearance rate. Other challenges that cannot be predicted, so far, include some proteins that are expressed recombinantly might be the toxic for the prokaryotic cell line used or pose other complexities of the bacterium's biology [27]. Therefore, unless these issues are met, the process of protein validation and progression to preclinical and clinical studies will still be largely hindered.

On the experimental side, current attempts to reduce the time required to validate a protein's usefulness in clinical settings include the implementation of automated processes aimed at promoting scalability of testing in a compatible time frame. Successful examples have shown that fully automated laboratories (also known as self-driving labs) can achieve higher standards of accuracy and quality controls in *in vitro* assays to determine protein biocompatibility and function, when compared to their conventional counterparts [74,75]. Although automation offers a glimpse of hope toward accelerating the translatability of computationally designed antibody mimetics, this approach is hardly a tangible solution for the vast majority of research groups due to its inherent high cost, need for stringent cybersecurity and the intrinsic characteristics of a molecular biology laboratory [74].

Considering the current issues, the efficient translation from computational design of synthetic antibody mimetics to real-world applications seem to lie on the comprehensive implementation of AI methods into the design framework so that it goes beyond achieving binding affinity and stability. Toward this end, we envision that a multiple context optimization engine that combines protein structure, protein language, protein images, and biological labels may prove crucial at designing the next generation of novel antibody mimetics.

Acknowledgements

This work was supported by grants from FACEPE (APQ-0346-2.09/19); CNPq (303833/2022-0, 151860/2022-0, INCT-FCx); and the Oswaldo Cruz Foundation through its Innovation Program (VPPCB-007-FIO-18-2-134 and IAM-005-FIO-22-2-44).

Conflict of interest

The authors declare no conflict of interest.

Author contributions

EJFC, EGM, JCMS, and MJN performed an in-depth review of antibody mimetics; DFC and CHBC have reviewed the computational methodologies to design antibody mimetics. RDL has conceptualized the manuscript, oversaw the work, and wrote the final version. All coauthors have approved the submission of this manuscript.

References

1 Wang Z, Wang G, Lu H, Li H, Tang M and Tong A (2022) Development of therapeutic antibodies for the treatment of diseases. *Mol Biomed* **3**, 35.
2 America, P RaMo (2020) Biopharmaceuticals in perspective.
3 Yu X, Yang YP, Dikici E, Deo SK and Daunert S (2017) Beyond antibodies as binding partners: the role of antibody mimetics in bioanalysis. *Annu Rev Anal Chem (Palo Alto Calif)* **10**, 293–320.
4 Du W, Jiang P, Li Q, Wen H, Zheng M, Zhang J, Guo Y, Yang J, Feng W, Ye S *et al.* (2023) Novel affibody molecules specifically bind to SARS-CoV-2 spike protein and efficiently neutralize delta and omicron variants. *Microbiol Spectr* **11**, e0356222.

FEBS Open Bio (2024) © 2024 The Author(s). *FEBS Open Bio* published by John Wiley & Sons Ltd on behalf of Federation of European Biochemical Societies.

22115463, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/2211-5463.13855 by CAPES, Wiley Online Library on [31/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

5 Klint S, Feldwisch J, Gudmundsdotter L, Dillner Bergstedt K, Gunneriusson E, Höidén Guthenberg I, Wennborg A, Nyborg AC, Kamboj AP, Peloso PM *et al.* (2023) Izokibep: preclinical development and first-in-human study of a novel IL-17A neutralizing affibody molecule in patients with plaque psoriasis. *MAbs* **15**, 2209920.
6 Stadler LKJ, Hoffmann T, Tomlinson DC, Song Q, Lee T, Busby M, Nyathi Y, Gendra E, Tiede C, Flanagan K *et al.* (2011) Structure–function studies of an engineered scaffold protein derived from Stefin A. II: development and applications of the SQT variant. *Protein Eng Des Sel* **24**, 751–763.
7 Ackermann M, Morse BA, Delventhal V, Carvajal IM and Konerding MA (2012) Anti-VEGFR2 and anti-IGF-1R-adnectins inhibit Ewing's sarcoma A673-xenograft growth and normalize tumor vascular architecture. *Angiogenesis* **15**, 685–695.
8 Parmeggiani F, Pellarin R, Larsen AP, Varadamsetty G, Stumpp MT, Zerbe O, Caflisch A and Plückthun A (2008) Designed armadillo repeat proteins as general peptide-binding scaffolds: consensus design and computational optimization of the hydrophobic core. *J Mol Biol* **376**, 1282–1304.
9 Alfarano P, Varadamsetty G, Ewald C, Parmeggiani F, Pellarin R, Zerbe O, Plückthun A and Caflisch A (2012) Optimization of designed armadillo repeat proteins by molecular dynamics simulations and NMR spectroscopy. *Protein Sci* **21**, 1298–1314.
10 Madhurantakam C, Varadamsetty G, Grütter MG, Plückthun A and Mittl PRE (2012) Structure-based optimization of designed armadillo-repeat proteins. *Protein Sci* **21**, 1015–1028.
11 Nästle L, Deuschle FC, Morath V and Skerra A (2023) FerryCalin: an engineered lipocalin protein directed against the transferrin receptor with potential for brain drug delivery. *Chembiochem* **24**, e202200795.
12 Deuschle F-C, Morath V, Schiefner A, Brandt C, Ballke S, Reder S, Steiger K, Schwaiger M, Weber W and Skerra A (2020) Development of a high affinity anticalin® directed against human CD98hc for theranostic applications. *Theranostics* **10**, 2172–2187.
13 Walser M, Mayor J and Rothenberger S (2022) Designed ankyrin repeat proteins: a new class of viral entry inhibitors. *Viruses* **14**, 2242.
14 Rothenberger S, Hurdiss DL, Walser M, Malvezzi F, Mayor J, Ryter S, Moreno H, Liechti N, Bosshart A, Iss C *et al.* (2022) The trispecific DARPin ensovibep inhibits diverse SARS-CoV-2 variants. *Nat Biotechnol* **40**, 1845–1854.
15 Stojcheva N, Gladman S, Soergel M, Zitt C, Drake R, Lockett T, Marchand C, Fustier P, Stavropoulou V, Fernandez E *et al.* (2023) Ensovibep, a SARS-CoV-2 antiviral designed ankyrin repeat protein, is safe and well tolerated in healthy volunteers: results of a first-in-human, ascending single-dose phase 1 study. *Br J Clin Pharmacol* **89**, 2295–2303.
16 Smithwick E and Stewart MW (2017) Designed ankyrin repeat proteins: a look at their evolving use in medicine with a focus on the treatment of chorioretinal vascular disorders. *Antiinflamm Antiallergy Agents Med Chem* **16**, 33–45.
17 Baker EG, Bartlett GJ, Porter Goff KL and Woolfson DN (2017) Miniprotein design: past, present, and prospects. *Acc Chem Res* **50**, 2085–2092.
18 Ożga K and Berlicki Ł (2022) Design and engineering of miniproteins. *ACS Bio Med Chem Au* **2**, 316–327.
19 Cao L, Goreshnik I, Coventry B, Case James B, Miller L, Kozodoy L, Chen Rita E, Carter L, Walls Alexandra C, Park Y-J *et al.* (2020) De novo design of picomolar SARS-CoV-2 miniprotein inhibitors. *Science* **370**, 426–431.
20 Hantschel O, Biancalana M and Koide S (2020) Monobodies as enabling tools for structural and mechanistic biology. *Curr Opin Struct Biol* **60**, 167–174.
21 Chandler PG and Buckle AM (2020) Development and differentiation in monobodies based on the fibronectin type 3 domain. *Cells* **9**, 610.
22 Diem MD, Hyun L, Yi F, Hippensteel R, Kuhar E, Lowenstein C, Swift EJ, O'Neil KT and Jacobs SA (2014) Selection of high-affinity centyrin FN3 domains from a simple library diversified at a combination of strand and loop positions. *Protein Eng Des Sel* **27**, 419–429.
23 Sha F, Salzman G, Gupta A and Koide S (2017) Monobodies and other synthetic binding proteins for expanding protein science. *Protein Sci* **26**, 910–924.
24 Bai Z, Wang J, Li J, Yuan H, Wang P, Zhang M, Feng Y, Cao X, Cao X, Kang G *et al.* (2023) Design of nanobody-based bispecific constructs by in silico affinity maturation and umbrella sampling simulations. *Comput Struct Biotechnol J* **21**, 601–613.
25 Tam C, Kukimoto-Niino M, Miyata-Yabuki Y, Tsuda K, Mishima-Tsumagari C, Ihara K, Inoue M, Yonemochi M, Hanada K, Matsumoto T *et al.* (2023) Targeting Ras-binding domain of ELMO1 by computational nanobody design. *Commun Biol* **6**, 284.
26 Pantazes RJ, Grisewood MJ and Maranas CD (2011) Recent advances in computational protein design. *Curr Opin Struct Biol* **21**, 467–472.
27 Huang P-S, Boyken SE and Baker D (2016) The coming of age of de novo protein design. *Nature* **537**, 320.
28 Gainza P, Nisonoff HM and Donald BR (2016) Algorithms for protein design. *Curr Opin Struct Biol* **39**, 16–26.
29 Anfinsen CB (1973) Principles that govern the folding of protein chains. *Science* **181**, 223.
30 Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL and Baker D (2003) Design of a novel globular

22115463, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/2211-5463.13855 by CAPES, Wiley Online Library on [31/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

protein fold with atomic-level accuracy. *Science* **302**, 1364–1368.
31 Soares TA, Boschek CB, Apiyo D, Baird C and Straatsma TP (2010) Molecular basis of the structural stability of a Top7-based scaffold at extreme pH and temperature conditions. *J Mol Graph Model* **28**, 755–765.
32 Viana IFT, Soares TA, Lima LFO, Marques ETA, Krieger MA, Dhalia R and Lins RD (2013) De novo design of immunoreactive conformation-specific HIV-1 epitopes based on Top7 scaffold. *RSC Adv* **3**, 11790–11800.
33 Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589.
34 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.
35 Drake ZC, Seffernick JT and Lindert S (2022) Protein complex prediction using Rosetta, AlphaFold, and mass spectrometry covalent labeling. *Nat Commun* **13**, 7846.
36 Tinberg CE, Khare SD, Dou J, Doyle L, Nelson JW, Schena A, Jankowski W, Kalodimos CG, Johnsson K, Stoddard BL *et al.* (2013) Computational design of ligand-binding proteins with high affinity and selectivity. *Nature* **501**, 212–216.
37 Marchand A, Van Hall-Beauvais AK and Correia BE (2022) Computational design of novel protein–protein interactions – an overview on methodological approaches and applications. *Curr Opin Struct Biol* **74**, 102370.
38 Schreiber G and Fleishman SJ (2013) Computational design of protein–protein interactions. *Curr Opin Struct Biol* **23**, 903–910.
39 Viana IFT, Cruz CHB, Athayde D, Adan WCS, Xavier LSS, Archer M and Lins RD (2023) In vitro neutralisation of zika virus by an engineered protein targeting the viral envelope fusion loop. *Mol Syst Design Eng* **8**, 516–526.
40 Dou J, Vorobieva AA, Sheffler W, Doyle LA, Park H, Bick MJ, Mao B, Foight GW, Lee MY, Gagnon LA *et al.* (2018) De novo design of a fluorescence-activating beta-barrel. *Nature* **561**, 485–491.
41 Silva DA, Correia BE and Procko E (2016) Motif-driven design of protein-protein interfaces. *Methods Mol Biol* **1414**, 285–304.
42 Correia BE, Ban YE, Friend DJ, Ellingson K, Xu H, Boni E, Bradley-Hewitt T, Bruhn-Johannsen JF, Stamatatos L, Strong RK *et al.* (2011) Computational protein design using flexible backbone remodeling and resurfacing: case studies in structure-based antigen design. *J Mol Biol* **405**, 284–297.
43 Bonet J, Wehrle S, Schriever K, Yang C, Billet A, Sesterhenn F, Scheck A, Sverrisson F, Veselkova B, Vollers S *et al.* (2018) Rosetta FunFolDes – a general framework for the computational design of functional proteins. *PLoS Comput Biol* **14**, e1006623.
44 Correia BE, Bates JT, Loomis RJ, Baneyx G, Carrico C, Jardine JG, Rupert P, Correnti C, Kalyuzhniy O, Vittal V *et al.* (2014) Proof of principle for epitope-focused vaccine design. *Nature* **507**, 201–206.
45 Kuhlman B and Bradley P (2019) Advances in protein structure prediction and design. *Nat Rev Mol Cell Biol* **20**, 681–697.
46 Karplus M and Weaver DL (1994) Protein folding dynamics: the diffusion-collision model and experimental data. *Protein Sci* **3**, 650–658.
47 Rohl CA, Strauss CEM, Misura KMS and Baker D (2004) Protein structure prediction using Rosetta. *Methods Enzymol* **383**, 66–93.
48 Simons KT, Kooperberg C, Huang E and Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J Mol Biol* **268**, 209–225.
49 Cao L, Coventry B, Goreshnik I, Huang B, Sheffler W, Park JS, Jude KM, Marković I, Kadam RU, Verschueren KHG *et al.* (2022) Design of protein-binding proteins from the target structure alone. *Nature* **605**, 551–560.
50 Case JB, Chen RE, Cao L, Ying B, Winkler ES, Johnson M, Goreshnik I, Pham MN, Shrihari S, Kafai NM *et al.* (2021) Ultrapotent miniproteins targeting the SARS-CoV-2 receptor-binding domain protect against infection and disease. *Cell Host Microbe* **29**, 1151–1161.e5.
51 Chaves E, Mhrous E, Nascimento-Filho M, Cruz C, Ferraz M and Lins R (2023) Prediction of absolute protein–protein binding free energy by a super learner model. *ChemRxiv*. doi: 10.26434/chemrxiv-2023-zq1nj
52 Ferraz MVF, Neto JCS, Lins RD and Teixeira ES (2023) An artificial neural network model to predict structure-based protein–protein free energy of binding from Rosetta-calculated properties. *Phys Chem Chem Phys* **25**, 7257–7267.
53 Ferraz MVF, Viana IFT, Coêlho DF, da Cruz CHB, de Arruda Lima M, de Luna Aragão MA and Lins RD (2022) Association strength of E6 to E6AP/p53 complex correlates with HPV-mediated oncogenesis risk. *Biopolymers* **113**, e23524.
54 Notin P, Rollins N, Gal Y, Sander C and Marks D (2024) Machine learning for functional protein design. *Nat Biotechnol* **42**, 216–228.
55 Varun RS, Theodora UJB, Brian LH and Peter SK (2023) Inverse folding of protein complexes with a structure-informed language model enables unsupervised antibody evolution. *bioRxiv*. doi: 10.1101/2023.12.19.572475

FEBS Open Bio (2024) © 2024 The Author(s). *FEBS Open Bio* published by John Wiley & Sons Ltd on behalf of Federation of European Biochemical Societies.

22115463, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/2211-5463.13855 by CAPES, Wiley Online Library on [31/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

56 Amir S, Matt M, George K, Andrea KS, John MS, Edriss Y, Cailen M, Robel H, Richard S, Julian A *et al.* (2024) Unlocking *de novo* antibody design with generative artificial intelligence. *bioRxiv*. doi: 10.1101/2023.01.08.523187
57 Hie BL, Shanker VR, Xu D, Bruun TUJ, Weidenbacher PA, Tang S, Wu W, Pak JE and Kim PS (2024) Efficient evolution of human antibodies from general protein language models. *Nat Biotechnol* **42**, 275–283.
58 Dauparas J, Anishchenko I, Bennett N, Bai H, Ragotte RJ, Milles LF, Wicky BIM, Courbet A, de Haas RJ, Bethel N *et al.* (2022) Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **378**, 49–56.
59 Bennett NR, Coventry B, Goreshnik I, Huang B, Allen A, Vafeados D, Peng YP, Dauparas J, Baek M, Stewart L *et al.* (2023) Improving de novo protein binder design with deep learning. *Nat Commun* **14**, 2625.
60 Bertoline LMF, Lima AN, Krieger JE and Teixeira SK (2023) Before and after AlphaFold2: an overview of protein structure prediction. *Front Bioinform* **3**, 1120370.
61 Ferruz N, Heinzinger M, Akdel M, Goncearenco A, Naef L and Dallago C (2023) From sequence to function through structure: deep learning for protein design. *Comput Struct Biotechnol J* **21**, 238–250.
62 Krapp LF, Abriata LA, Cortes Rodriguez F and Dal Peraro M (2023) PeSTo: parameter-free geometric deep learning for accurate prediction of protein binding interfaces. *Nat Commun* **14**, 2175.
63 Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, Žídek A, Bates R, Blackwell S, Yim J *et al.* (2022) Protein complex prediction with AlphaFold-multimer. *bioRxiv*. doi: 10.1101/2021.10.04.463034
64 Yin R, Feng BY, Varshney A and Pierce BG (2022) Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants. *Protein Sci* **31**, e4379.
65 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.
66 Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD *et al.* (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876.
67 Zhang Y and Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* **33**, 2302–2309.
68 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.
69 Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH and Teller E (1953) Equation of state calculations by fast computing machines. *J Chem Phys* **21**, 1087.
70 Bender BJ, Cisneros A, Duran AM, Finn JA, Fu D, Lokits AD, Mueller BK, Sangha AK, Sauer MF, Sevy AM *et al.* (2016) Protocols for molecular modeling with Rosetta3 and RosettaScripts. *Biochemistry* **55**, 4748–4763.
71 Bellissent-Funel M-C, Hassanali A, Havenith M, Henchman R, Pohl P, Sterpone F, van der Spoel D, Xu Y and Garcia AE (2016) Water determines the structure and dynamics of proteins. *Chem Rev* **116**, 7673–7697.
72 Guo Z and Yamaguchi R (2022) Machine learning methods for protein–protein binding affinity prediction in protein design. *Front Bioinform* **2**, 1065703.
73 Wan S, Bhati AP, Zasada SJ and Coveney PV (2020) Rapid, accurate, precise and reproducible ligand–protein binding free energy prediction. *Interface Focus* **10**, 20200007.
74 Martin HG, Radivojevic T, Zucker J, Bouchard K, Sustarich J, Peisert S, Arnold D, Hillson N, Babnigg G, Marti JM *et al.* (2023) Perspectives for self-driving labs in synthetic biology. *Curr Opin Biotechnol* **79**, 102881.
75 Rapp JT, Bremer BJ and Romero PA (2024) Self-driving laboratories to autonomously navigate the protein fitness landscape. *Nat Chem Eng* **1**, 97–107.

22115463, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/2211-5463.13855 by CAPES, Wiley Online Library on [31/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License