

Molecular Dynamics Simulations: Deep Learning Approaches

PhD. Conrado Pedebos
PhD. Pablo Ricardo Arantes

Porto Alegre, July 17th 2025

Proteína - Ligante

Summary of Course Flow:

Day 1 – Introduction to Molecular Dynamics

Day 2 – Protein-Ligand Systems

Day 3 – Coarse-Grained Molecular Dynamics and Membrane Proteins

Day 4 – Alternatives to Classical Molecular Dynamics for Conformational Sampling

Day 5 – Enhanced Sampling Methods

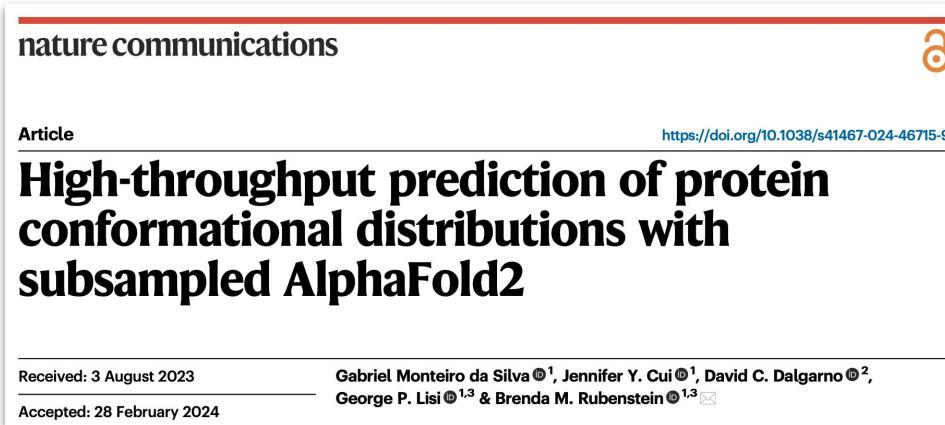
Slides + Notebooks



Deep Learning Approaches

Visão Geral dos Métodos: AlphaFold2 Subamostrado e BioEme

- **AlphaFold2 Subamostrado: Predição de Distribuições Conformativas em Alto Rendimento**
 - Uma abordagem inovadora que modifica o AlphaFold2 (AF2) para prever as populações relativas de diferentes conformações proteicas.



The screenshot shows a research article from the journal *nature communications*. The article is an Article and can be identified by the DOI <https://doi.org/10.1038/s41467-024-46715-9>. The title of the article is **High-throughput prediction of protein conformational distributions with subsampled AlphaFold2**. The authors listed are Gabriel Monteiro da Silva  ¹, Jennifer Y. Cui  ¹, David C. Dalgarno  ², George P. Lisi  ^{1,3} & Brenda M. Rubenstein  ^{1,3}  . The article was received on 3 August 2023 and accepted on 28 February 2024.

<https://doi.org/10.1038/s41467-024-46715-9>

Deep Learning Approaches

Visão Geral dos Métodos: AlphaFold2 Subamostrado e BioEmu

- **BioEmu (Biomolecular Emulator): Emulação Escalável de Dinâmica Proteica**
 - Um sistema de deep learning generativo que emula dinâmicas de proteínas, gerando milhares de conformações de equilíbrio por hora em uma única GPU.

Scalable emulation of protein equilibrium ensembles with generative deep learning

 Sarah Lewis,  Tim Hempel,  José Jiménez-Luna,  Michael Gastegger,  Yu Xie, Andrew Y. K. Foong, Victor García Satorras,  Osama Abdin, Bastiaan S. Veeling,  Iryna Zaporozhets,  Yaoyi Chen,  Soojung Yang,  Arne Schneuing, Jigyasa Nigam, Federico Barbero, Vincent Stimper, Andrew Campbell, Jason Yim, Marten Lienen, Yu Shi, Shuxin Zheng, Hannes Schulz, Usman Munir, Ryota Tomioka, Cecilia Clementi, Frank Noé

doi: <https://doi.org/10.1101/2024.12.05.626885>



Deep Learning Approaches

AlphaFold2: Revolução e Limitação Inicial

- **O que é AlphaFold2 (AF2)?**
 - Sistema de *deep learning* desenvolvido pela DeepMind para prever a estrutura 3D de proteínas a partir de suas sequências de aminoácidos.
 - Treinado com vastos dados experimentais e informações co-evolucionárias de grandes bancos de dados metagenômicos.
 - Revolucionou a biologia pela sua precisão e velocidade excepcionais na predição de estruturas, abrindo novas possibilidades para descoberta de fármacos e pesquisa básica.
- **Limitação Inicial:**
 - O algoritmo padrão do AF2 foi projetado para prever as conformações de estado fundamental (*ground state*) de proteínas.
 - É limitado em sua capacidade de prever diferentes estados conformacionais ou múltiplas conformações alternativas.

Deep Learning Approaches

Article

Highly accurate protein structure prediction with AlphaFold

<https://doi.org/10.1038/s41586-021-03819-2>

Received: 11 May 2021

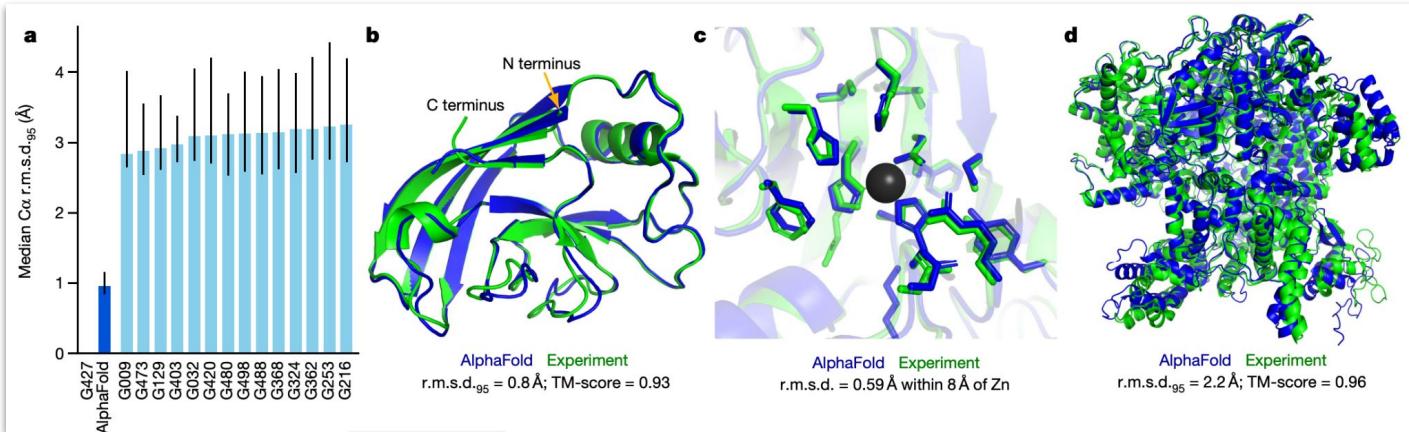
Accepted: 12 July 2021

Published online: 15 July 2021

Open access

 Check for updates

John Jumper^{1,4}, Richard Evans^{1,4}, Alexander Pritzel^{1,4}, Tim Green^{1,4}, Michael Figurnov^{1,4}, Olaf Ronneberger^{1,4}, Kathryn Tunyasuvunakool^{1,4}, Russ Bates^{1,4}, Augustin Žídek^{1,4}, Anna Potapenko^{1,4}, Alex Bridgland^{1,4}, Clemens Meyer^{1,4}, Simon A. A. Kohl^{1,4}, Andrew J. Ballard^{1,4}, Andrew Cowie^{1,4}, Bernardino Romera-Paredes^{1,4}, Stanislav Nikолов^{1,4}, Rishabh Jain^{1,4}, Jonas Adler¹, Trevor Back¹, Stig Petersen¹, David Reiman¹, Ellen Clancy¹, Michal Zielinski¹, Martin Steinegger^{2,3}, Michalina Pacholska¹, Tamas Berghammer¹, Sebastian Bodenstein¹, David Silver¹, Oriol Vinyals¹, Andrew W. Senior¹, Koray Kavukcuoglu¹, Pushmeet Kohli¹ & Demis Hassabis^{1,4}



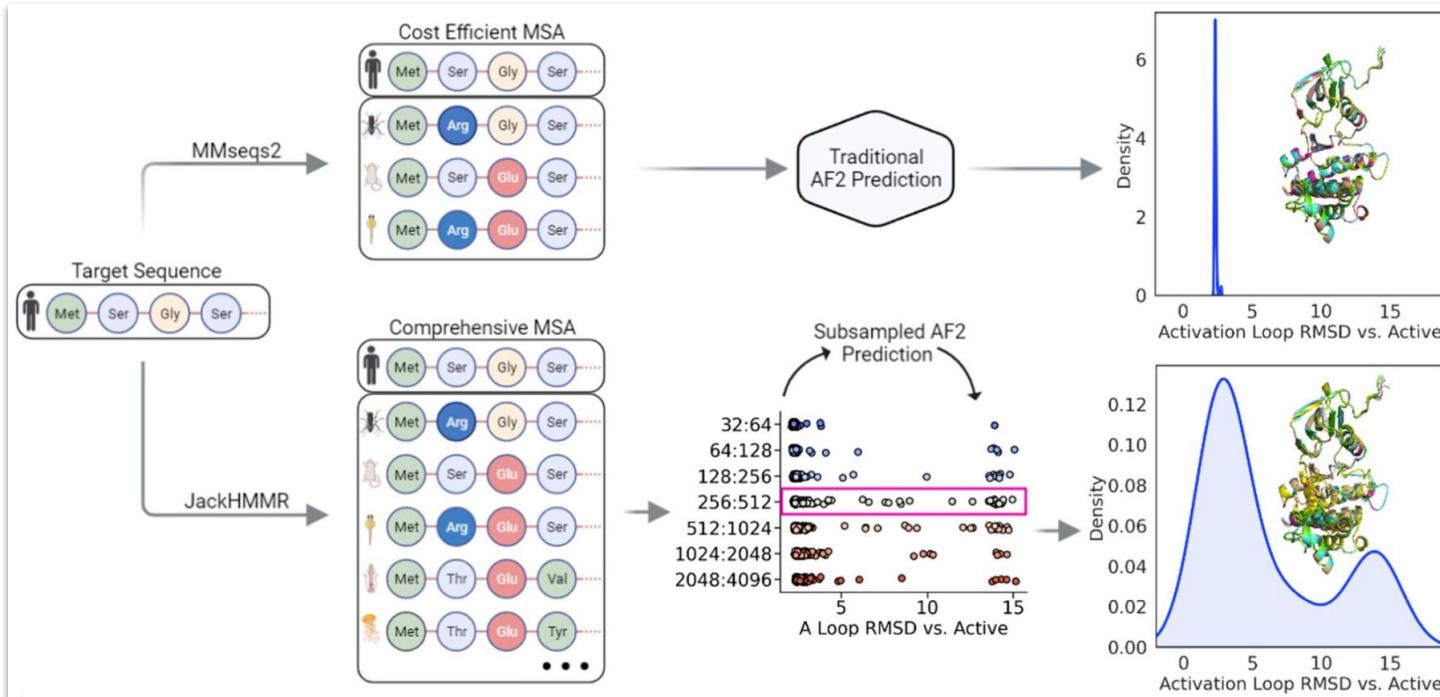
Deep Learning Approaches

Subamostragem de MSAs com AF2

- **A Necessidade de Ensembles:**
 - Fenômenos biológicos como "*fold-switching*" e transição ordem-desordem são ubíquos e ligados à função macromolecular, exigindo a compreensão de múltiplas conformações.
 - A capacidade de prever múltiplas conformações pode revolucionar a descoberta de fármacos, revelando mais alvos.
- **A Abordagem de Subamostragem de MSA:**
 - A chave é a subamostragem dos alinhamentos de sequência múltipla (MSAs) de entrada e o aumento do número de predições.
 - Isso modula os sinais co-evolucionários decodificados pelo AF2.
 - O AF2 seleciona um número de sequências (***max_seq***) aleatoriamente do MSA mestre e agrupa as restantes, usando centros de cluster e amostras adicionais (***extra_seq***) para inferência.

Deep Learning Approaches

Subamostragem de MSAs com AF2



Deep Learning Approaches

Previsão de Populações Relativas de Estados com AF2

Permitir que o AF2 preveja diretamente as populações relativas de diferentes conformações de proteínas.

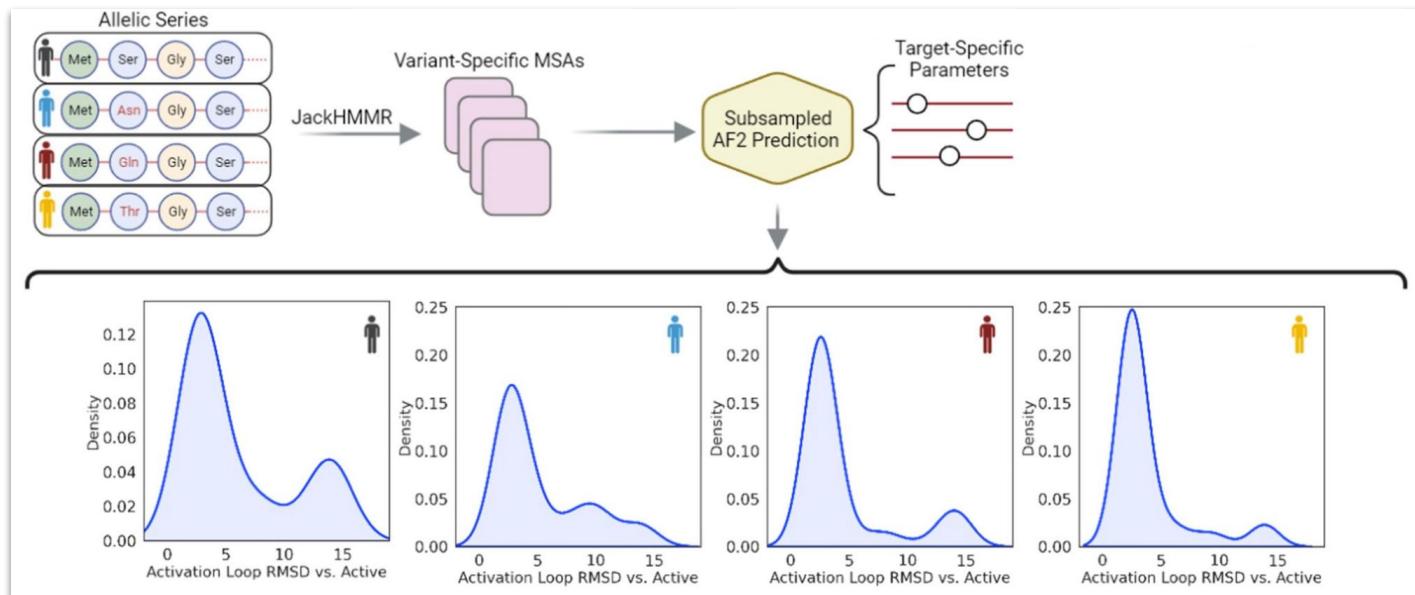
A hipótese é que o AF2 pode quantificar sinais dinâmicos codificados em sequência, tornando possível prever não apenas conformações alternativas, mas também mudanças nas populações relativas de seus estados.

- **Vantagem sobre MD Tradicional:**

- Essa abordagem pode tornar as etapas de simulação de MD desnecessárias para a descoberta de alto rendimento de grandes conformações alternativas e suas populações relativas.

Deep Learning Approaches

- Otimização dos parâmetros de subamostragem (*max_seq*, *extra_seq*) para gerar a maior diversidade de conformações.
- Executar múltiplas predições com *seeds* independentes e "dropouts" durante a inferência para amostrar a incerteza dos modelos.

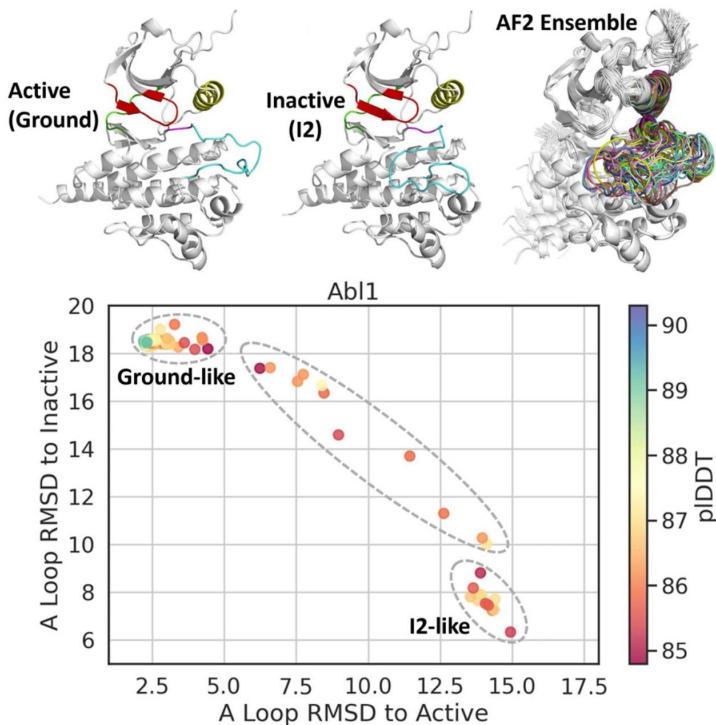


Deep Learning Approaches

Estudos de Caso do AF2: Abl1 Kinase e GMCSF

- **Abl1 Kinase:**

- Previsão qualitativa das populações de estado: o AF2 subamostrado previu corretamente a diferença nas distribuições conformacionais entre os núcleos das quinases Abl1 e Src, e entre Abl1, Anc-AS e Src.
- Efeitos de Mutação: Previu a mudança na população de estado relativo e sua direção em mais de 80% dos casos testados para uma série alélica de Abl1.
- Conformações Intermediárias: O ensemble de conformações do laço de ativação previsto pelo AF2 cobriu bem a transição do estado fundamental para o estado I2, sugerindo a capacidade de amostrar estados intermediários e descobrir vias.



Deep Learning Approaches

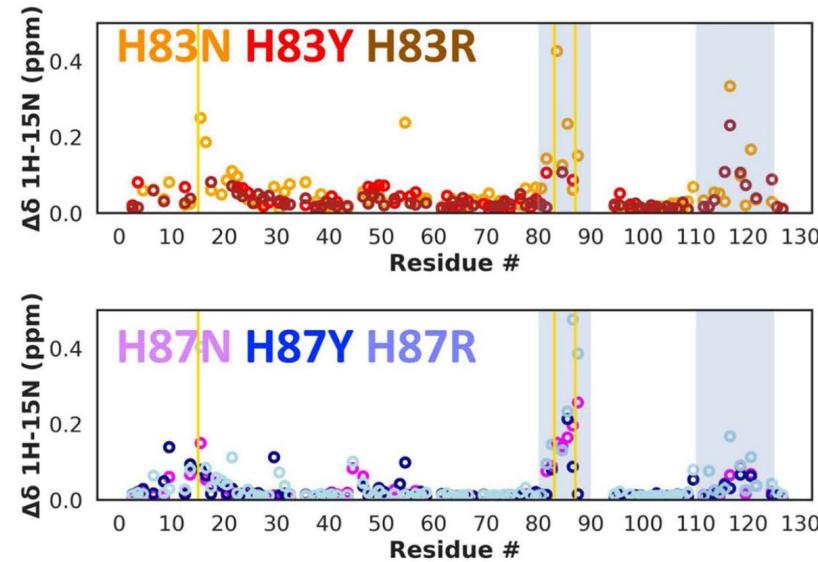
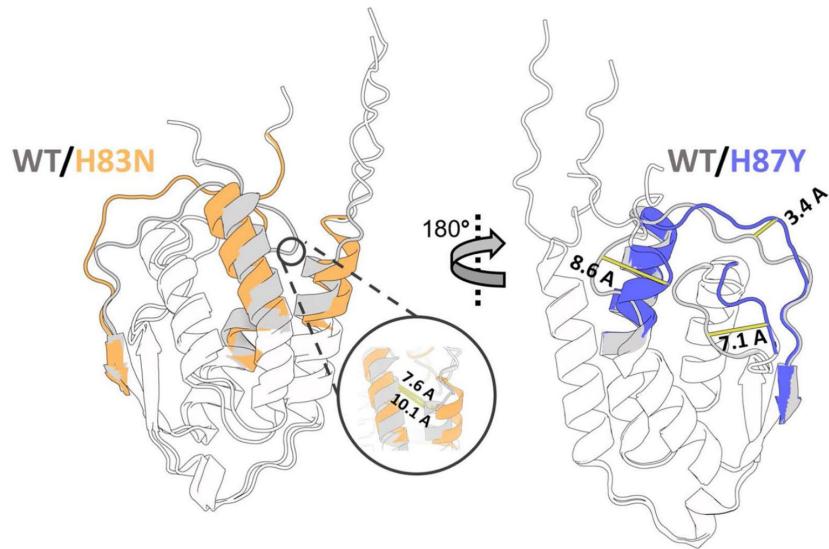
Estudos de Caso do AF2: Abl1 Kinase e GMCSF

- **GMCSF (Fator Estimulador de Colônias de Granulócitos-Macrófagos):**
 - Apesar da escassez de dados de sequência (MSA de apenas 112 sequências vs. 600.000+ para Abl1), o AF2 subamostrado previu ensembles conformacionais que se correlacionaram com experimentos de RMN, mostrando robustez com dados limitados.
 - Identificou resíduos mais sensíveis à mutação (H15 e H83) e previu o impacto significativo de certas mutações na dinâmica da cadeia principal.

Deep Learning Approaches

Estudos de Caso do AF2: Abl1 Kinase e GMCSF

- GMCSF (Fator Estimulador de Colônias de Granulócitos-Macrófagos):



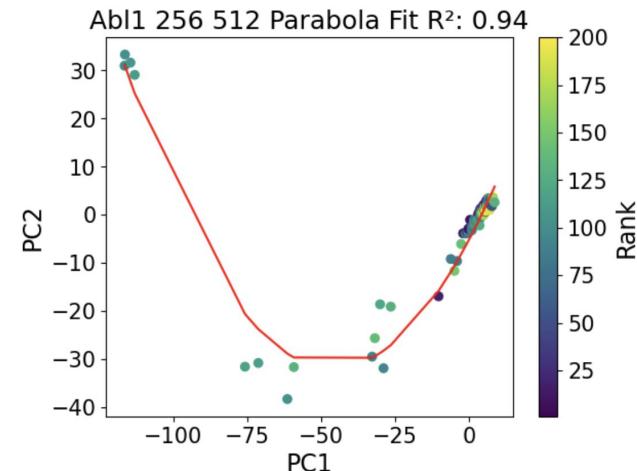
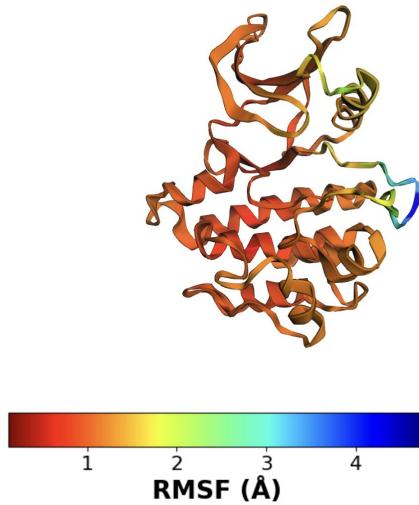
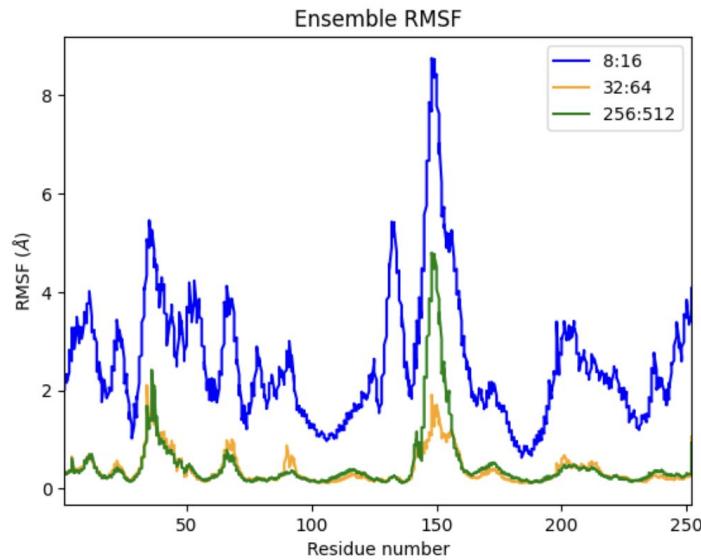
Deep Learning Approaches

Limitações e Desafios do AF2 na Geração de Ensembles

- **Limitação de Resolução:**
 - Conformações com ocupação muito baixa podem ser perdidas pelo algoritmo.
 - Sugere-se que há um limite mínimo entre 2% e 6% na população relativa para detectar estados de alta energia usando o AF2 subamostrado.
- **Acurácia Quantitativa:**
 - A acurácia se aplica principalmente no sentido qualitativo; pode subestimar ou prever efeitos incorretos para certas mutações.
- **Outras Desvantagens:**
 - Pode falhar na predição de todos os elementos estruturais esperados ou na amplitude correta das mudanças.
 - Ocasional predição de conformações parcialmente desenoveladas sem análogos experimentais, especialmente com MSAs mais rasos.
 - Desempenho melhor em mudanças conformativas grandes e lentas; limitado para mudanças mais rápidas ou menos significativas.

Deep Learning Approaches

Exemplo prático de AF2 na Geração de Ensembles (Abl1 Kinase)



Deep Learning Approaches

- **FastConformation: Uma Ferramenta para Exploração de Ensembles Conformatoriais com AlphaFold2 Subamostrado**

New Results

 [Follow this preprint](#)

FastConformation: A Standalone ML-Based Toolkit for Modeling and Analyzing Protein Conformational Ensembles at Scale

Flavia Maria Galeazzi,  Gabriel Monteiro da Silva, Pablo Arantes, Iz Varghese, Ananya Shukla,
 Brenda M. Rubenstein

doi: <https://doi.org/10.1101/2025.05.09.653048>

<https://doi.org/10.1101/2025.05.09.653048>

Deep Learning Approaches

- **FastConformation: Uma Ferramenta para Exploração de Ensembles Conformatoriais com AlphaFold2 Subamostrado**
 - Toolkit de Machine Learning autônomo e com Interface Gráfica de Usuário (GUI) amigável, projetado para modelar e analisar ensembles conformacionais de proteínas em escala.
 - Ele integra geração de MSA, predição de estrutura via AlphaFold2 (AF2) e análise interativa, tudo em um só lugar.
 - Previsões e análises em poucas horas em máquinas locais, drasticamente mais rápido que as simulações de Dinâmica Molecular (MD) tradicionais.
 - Insights Rápidos: Permite a predição qualitativa dos efeitos de mutações ou evolução na paisagem conformacional, mesmo com dados de sequência escassos.
 - Acessibilidade: Desenvolvido para não-programadores através de sua GUI, tornando métodos avançados de predição acessíveis.

Deep Learning Approaches

- **FastConformation: Uma Ferramenta para Exploração de Ensembles Conformativos com AlphaFold2 Subamostrado**



Deep Learning Approaches

- **FastConformation: Uma Ferramenta para Exploração de Ensembles Conformatoriais com AlphaFold2 Subamostrado**



https://youtu.be/a7_y4lxic8w?si=ywTe3UzGpaEW6gpg

Deep Learning Approaches

- **BioEmu (Biomolecular Emulator): Emulação Escalável de Dinâmica Proteica**
 - Um sistema de deep learning generativo que emula dinâmicas de proteínas, gerando milhares de conformações de equilíbrio por hora em uma única GPU.

Scalable emulation of protein equilibrium ensembles with generative deep learning

 Sarah Lewis,  Tim Hempel,  José Jiménez-Luna,  Michael Gastegger,  Yu Xie, Andrew Y. K. Foong, Victor García Satorras,  Osama Abdin, Bastiaan S. Veeling,  Iryna Zaporozhets,  Yaoyi Chen,  Soojung Yang,  Arne Schneuing,  Jigyasa Nigam, Federico Barbero,  Vincent Stimper, Andrew Campbell,  Jason Yim, Marten Lienen, Yu Shi,  Shuxin Zheng, Hannes Schulz,  Usman Munir,  Ryota Tomioka,  Cecilia Clementi,  Frank Noé

doi: <https://doi.org/10.1101/2024.12.05.626885>



<https://doi.org/10.1101/2024.12.05.626885>

Deep Learning Approaches

- **BioEmu (Biomolecular Emulator): Emulação Escalável de Dinâmica Proteica**
 - Um sistema de deep learning generativo que emula dinâmicas de proteínas, gerando milhares de conformações de equilíbrio por hora em uma única GPU.

Science RESEARCH ARTICLES

Cite as: S. Lewis *et al.*, *Science* 10.1126/science.adv9817 (2025).

Scalable emulation of protein equilibrium ensembles with generative deep learning

Sarah Lewis^{1†}, Tim Hempel^{1†}, José Jiménez-Luna^{1†}, Michael Gastegger^{1†}, Yu Xie^{1†}, Andrew Y. K. Foong^{1†}, Victor García Satorras^{1†}, Osama Abdin^{1†}, Bastiaan S. Veeling^{1†}, Iryna Zaporozhets^{1,2}, Yaoyi Chen^{1,2}, Soojung Yang¹, Adam E. Foster¹, Arne Schneuing¹, Jigyasa Nigam¹, Federico Barbero¹, Vincent Stimper¹, Andrew Campbell¹, Jason Yim¹, Marten Lienen¹, Yu Shi¹, Shuxin Zheng¹, Hannes Schulz¹, Usman Munir¹, Roberto Sordillo¹, Ryota Tomioka¹, Cecilia Clementi^{1,2,3}, Frank Noe^{1,2,3*}

¹AI for Science, Microsoft Research. ²Freie Universität Berlin, Berlin, Germany. ³Department of Chemistry, Rice University, Houston, TX, USA.
†These authors contributed equally to this work.
*Corresponding author. Email: franknoe@microsoft.com



<https://www.science.org/doi/10.1126/science.adv9817>

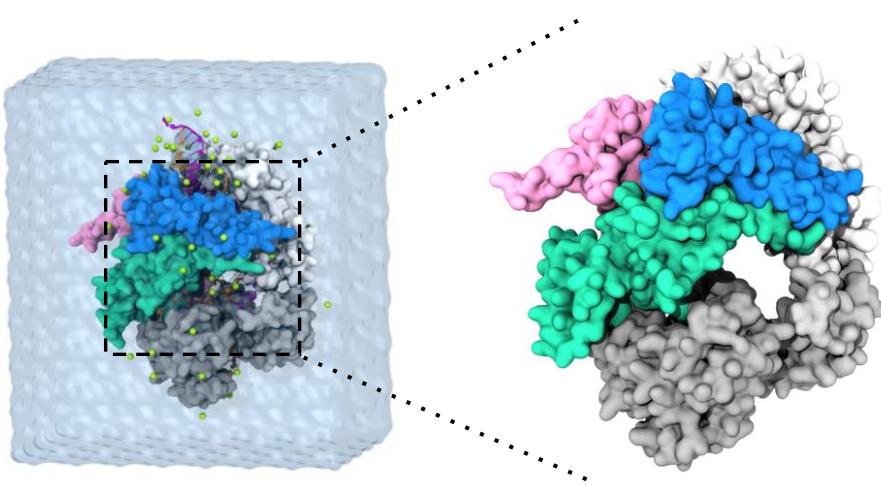
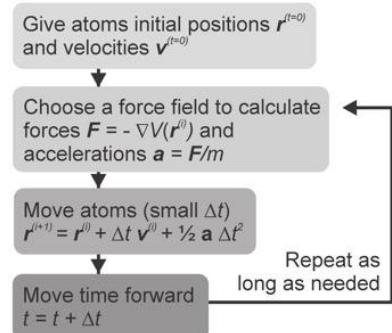
Deep Learning Approaches

BioEmu: Biomolecular Emulator para Distribuições de Equilíbrio

- **O Desafio da Função Proteica:**

- Apesar dos avanços em sequência e estrutura, prever os mecanismos dinâmicos de proteínas que implementam funções biológicas continua sendo um desafio científico.
- Técnicas experimentais e simulações de MD são limitadas por baixo rendimento (throughput).

Molecular Dynamics (MD)



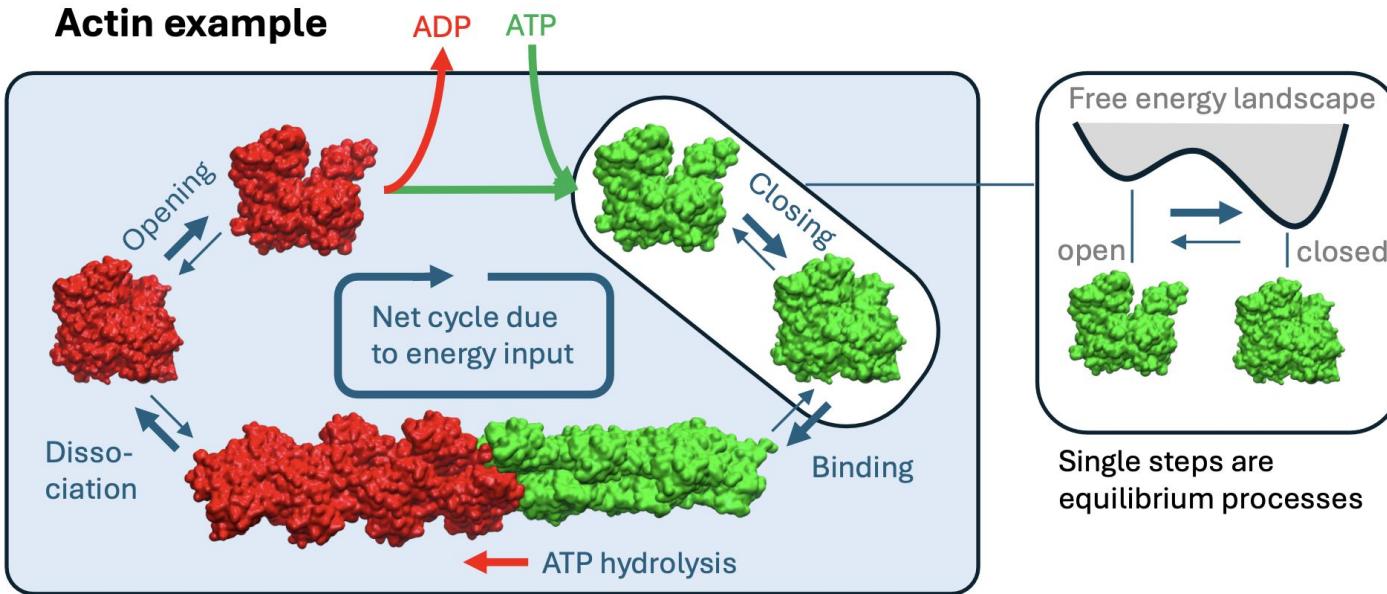
Deep Learning Approaches

BioEmu: *Biomolecular Emulator* para Distribuições de Equilíbrio

- **O que é BioEmu?**
 - Um sistema generativo de deep learning (Aprendizado Profundo) desenvolvido para emular ensembles de proteínas.
 - Capaz de gerar milhares de amostras estatisticamente independentes do ensemble de estrutura da proteína por hora em uma única GPU.
 - Objetivo: Alcançar a acurácia de uma simulação de MD convergida ou experimento cryo-EM (com análise multi-conformacional), mas em horas/dólares, em vez de dias/milhões.
- **Inovação:**
 - Representa o equilíbrio em uma gama de métricas desafiadoras e relevantes.
 - Revela insights mecânicos, como causas de desestabilização de mutantes, e pode fornecer hipóteses testáveis experimentalmente.

Deep Learning Approaches

BioEmu: *Biomolecular Emulator para Distribuições de Equilíbrio*

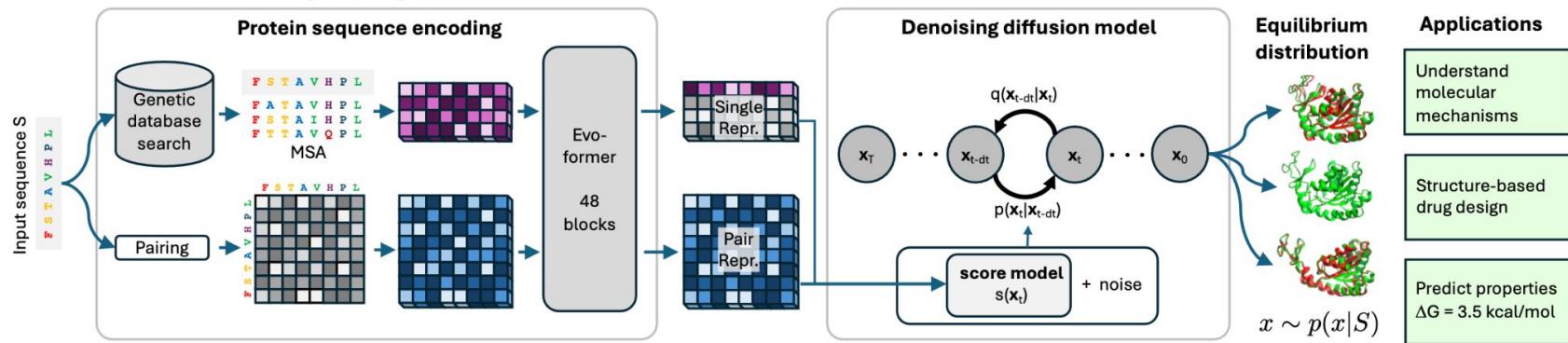


Deep Learning Approaches

Arquitetura e Treinamento Abrangente do BioEmu

- **Arquitetura do Modelo:**
 - Utiliza uma arquitetura de modelo similar ao Distributional Graphomer.
 - Baseado no evoformer do AlphaFold2, computa representações de sequência única e de par.
 - Modelo de difusão denoising: Gera estruturas de proteínas de *coarse-grained* na representação de quadro (*frame*) da cadeia principal.
 - Eficiência de Amostragem: 10.000 estruturas independentes podem ser amostradas em minutos a poucas horas em uma única GPU, dependendo do tamanho da proteína.

Biomolecular Emulator (BioEmu)

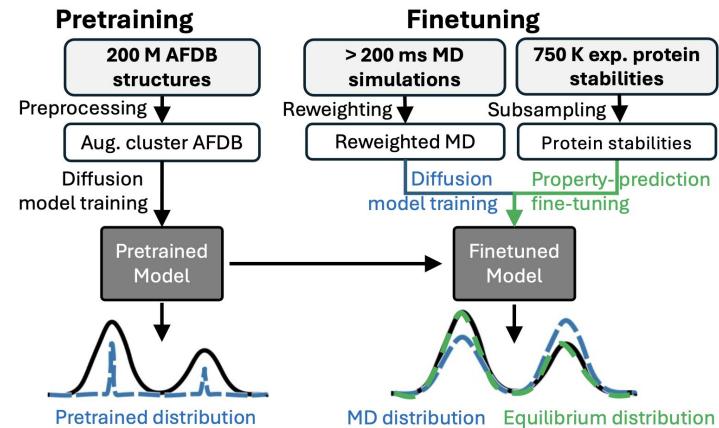


Deep Learning Approaches

Arquitetura e Treinamento Abrangente do BioEmu

- Estratégia de Treinamento (Dados Heterogêneos):**

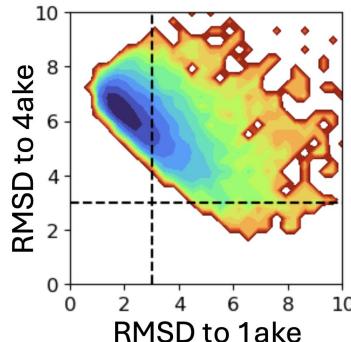
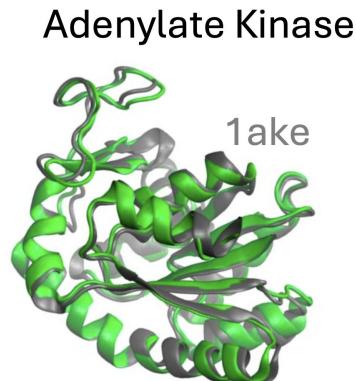
- Pré-treinamento: Em uma versão agrupada do banco de dados AlphaFold (AFDB), com uma estratégia de aumento de dados que incentiva a amostragem de diversas conformações.
- Fine-tuning: Continua o treinamento em uma mistura de dados de simulação de MD (mais de 200 milissegundos) e medições experimentais de estabilidade de proteínas (mais de 750.000 medições do dataset MEGAscale).
- Dados de MD são reponderados para o equilíbrio usando *Markov State Models* (MSMs) ou pesos de dados experimentais.



Deep Learning Approaches

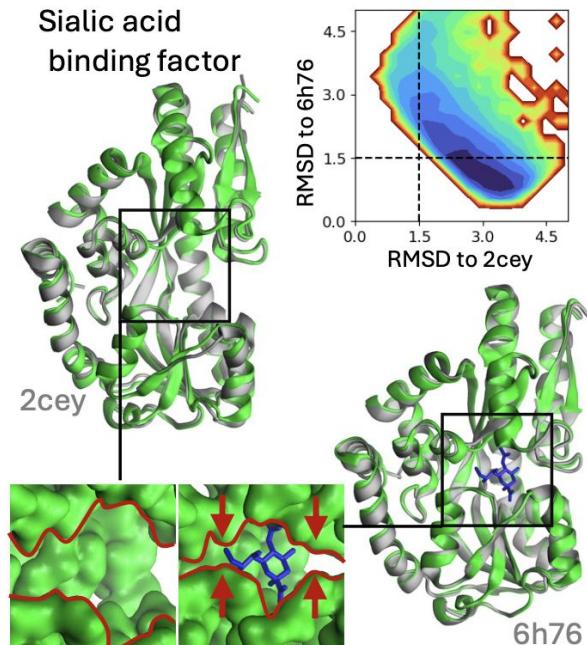
BioEmu: Amostrando Mudanças Conformativas Funcionais

- **Amostragem Qualitativa de Conformações Relevantes:**
 - BioEmu amostra muitas mudanças conformacionais funcionalmente relevantes.
 - Supera métodos baseline como AFCluster e AlphaFlow em conjuntos de teste.
 - Grandes movimentos de domínio / rearranjos
 - Transições de desenovelamento local
 - Formação de "bolsões crípticos" de ligação que não estão presentes no apo

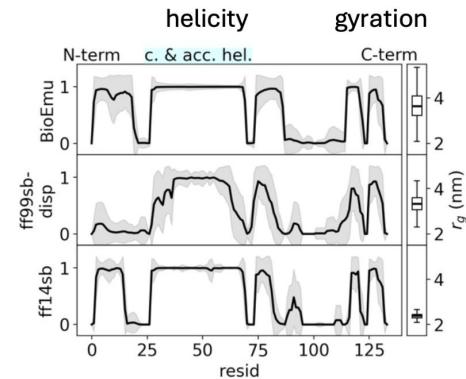
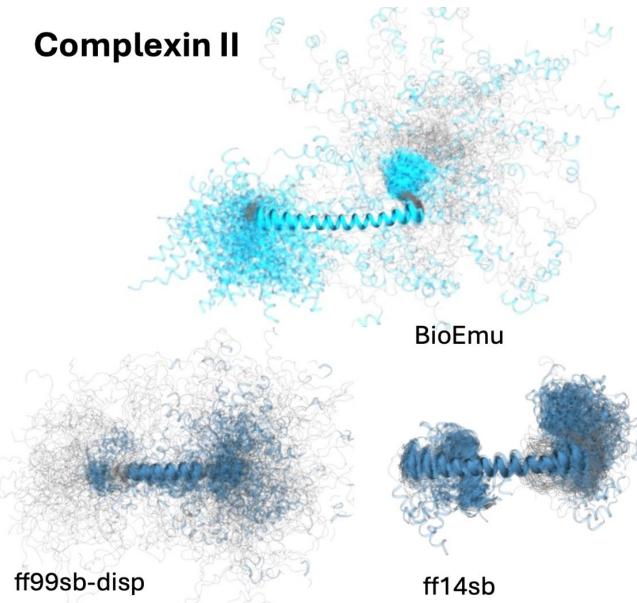


Deep Learning Approaches

BioEmu: Amostrando Mudanças Conformativas Funcionais



Complexin II



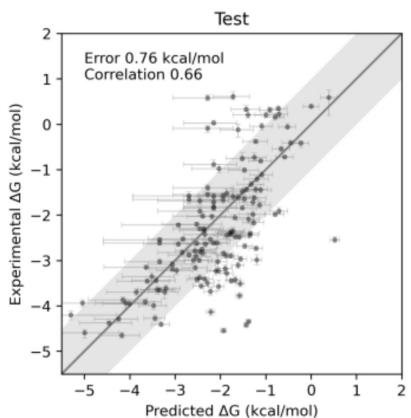
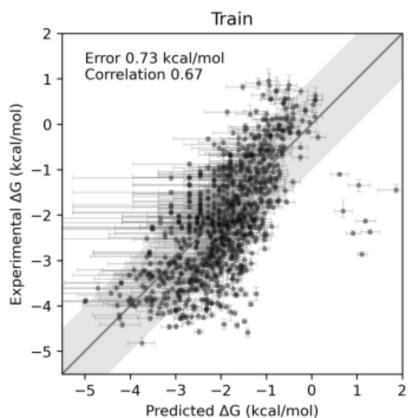
Deep Learning Approaches

BioEmu: Emulando Distribuições de Equilíbrio de MD

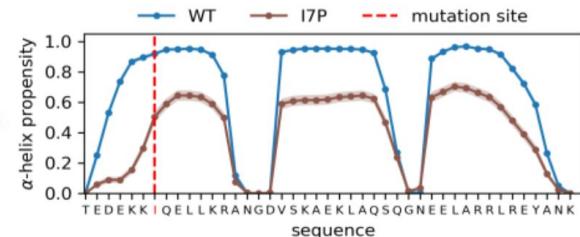
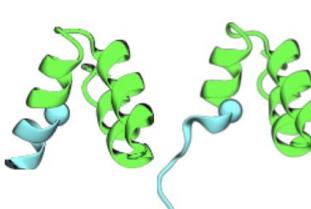
- **Superando o Problema de Amostragem da MD:**
 - Uma motivação importante para o BioEmu é contornar o problema de amostragem da MD, que exige tempos de simulação de milissegundos.
- **Resultados de Emulação:**
 - Excelente concordância das superfícies de energia livre e propensões de estrutura secundária com simulações de MD em larga escala.
 - Erro médio absoluto (MAE) das superfícies de energia livre de 2D de apenas ~0.74-0.91 kcal/mol, comparável às diferenças esperadas entre diferentes campos de força de MD.
 - Vantagem de Custo Computacional: Aceleração de quatro a cinco ordens de magnitude sobre a MD, com custos de < 1 minuto de GPU para proteínas pequenas a 1 hora de GPU para proteínas maiores.
 - Exemplos de proteínas maiores: Eficientemente emula ensembles de Complexin II (proteína intrinsecamente desordenada - IDP) e reproduz o motivo HEXXH+E estável da ACE2.

Deep Learning Approaches

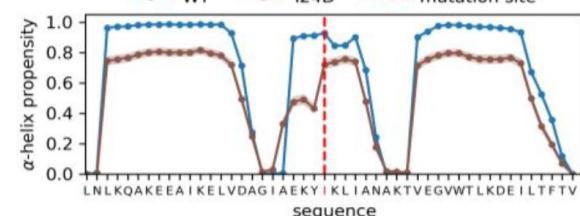
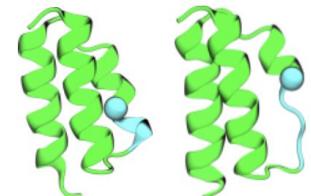
BioEmu: Emulando Distribuições de Equilíbrio de MD



Wild type HHH_rd1_0335, Mutation I7P



Wild type 2JWS, Mutation I24D



Deep Learning Approaches

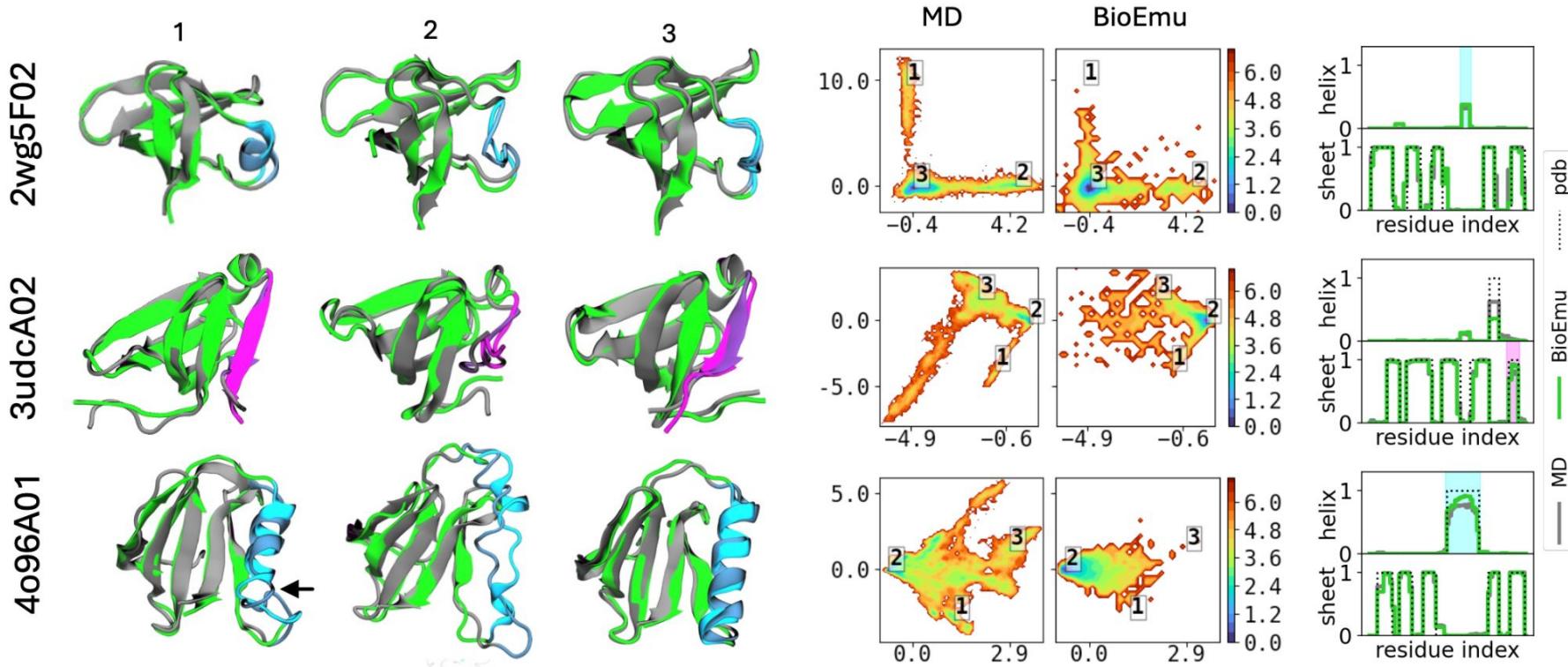
BioEmu: Previsão de Estabilidades Proteicas e Perspectivas

- **Previsão de Estabilidade Proteica:**

- BioEmu é treinado para que a proporção de amostras em estados enovelados e desenovelados corresponda à estabilidade proteica medida experimentalmente.
- Atinge um erro absoluto médio abaixo de 0.8 kcal/mol para energias livres de dobramento.
- Algoritmo inovador que incorpora eficientemente medições experimentais (como ΔG) no treinamento do modelo de difusão sem exigir estruturas de proteínas, por meio de retropropagação do erro de dobramento amostrado.

Deep Learning Approaches

BioEmu: Previsão de Estabilidades Proteicas e Perspectivas



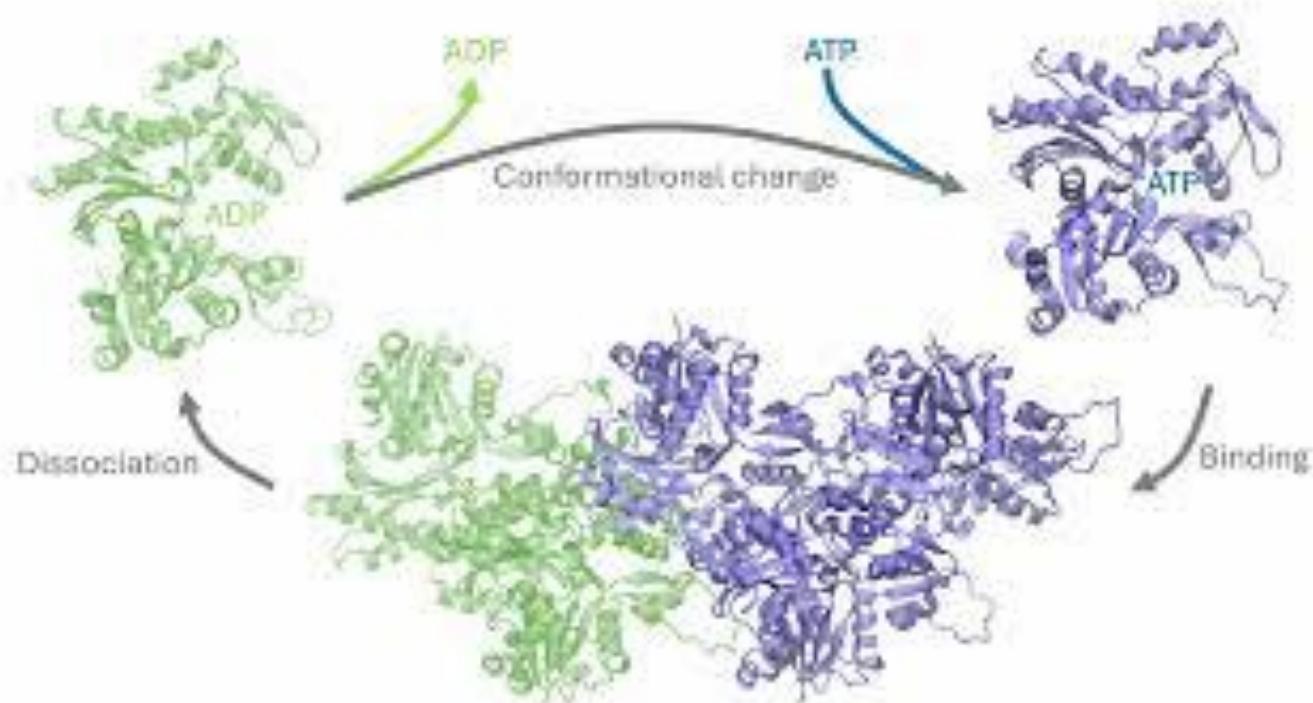
Deep Learning Approaches

BioEmu: Previsão de Estabilidades Proteicas e Perspectivas

- **Insights Mecanísticos:**
 - Ao contrário de modelos "caixa preta", BioEmu permite analisar o ensemble de estruturas gerado para revelar as causas estruturais das mudanças de estabilidade induzidas por mutações.
- **Relação Complementar com MD:**
 - BioEmu gera uma primeira estimativa da distribuição de equilíbrio; trajetórias de MD podem ser lançadas a partir do ensemble do BioEmu para refinar estruturas e calcular propriedades dinâmicas.
 - Não substitui a MD, mas muda seu papel para ferramenta de geração e validação de dados.
 - Limitações: Distribuições geradas empiricamente (sem função de energia potencial subjacente), atualmente emula apenas cadeias de proteínas únicas em condições termodinâmicas fixas (300K), não representa ambientes de membrana ou ligantes de pequenas moléculas. A falta de dados de treinamento para complexos maiores e afinidades de ligação continua sendo um obstáculo.

Deep Learning Approaches

BioEmu: Previsão de Estabilidades Proteicas e Perspectivas



Deep Learning Approaches

Avanços e Perspectivas Futuras

AlphaFold2 Subamostrado: Demonstra uma capacidade surpreendente de prever distribuições conformacionais em alto rendimento e populações de estados relativos, mesmo para proteínas com dados de sequência escassos. Capta transições importantes e estados intermediários, oferecendo uma ferramenta valiosa para a descoberta de medicamentos e a compreensão de mecanismos biológicos.

BioEmu: Representa um avanço significativo na emulação de conjuntos de equilíbrio proteicos, oferecendo uma velocidade de inferência muitas ordens de magnitude maior (quatro a cinco ordens de magnitude) do que as simulações de MD tradicionais. Atinge precisão quantitativa na previsão de energias livres de dobramento (erros inferiores a 1 kcal/mol), comparável à precisão experimental. Permite insights mecanicistas sobre a relação estrutura-estabilidade de mutantes.

Deep Learning Approaches

Limitações e Desafios

AlphaFold2 Subamostrado: Embora precise, a previsão é qualitativa, não estritamente quantitativa, e pode ter limitações de resolução para conformações de muito baixa ocupação. Pode, ocasionalmente, prever conformações parcialmente desenoveladas ou não físicas, especialmente com MSAs superficiais. Apresenta melhor desempenho na previsão de mudanças conformacionais grandes e lentas.

BioEmu: Atualmente, emula apenas cadeias proteicas únicas em condições termodinâmicas fixas (300K) e não modela explicitamente ambientes de membrana ou ligantes de pequenas moléculas. A precisão e o escopo podem ser aprimorados com a disponibilidade de mais dados de treinamento diversos e escaláveis.

Deep Learning Approaches

Impacto e Perspectivas Futuras:

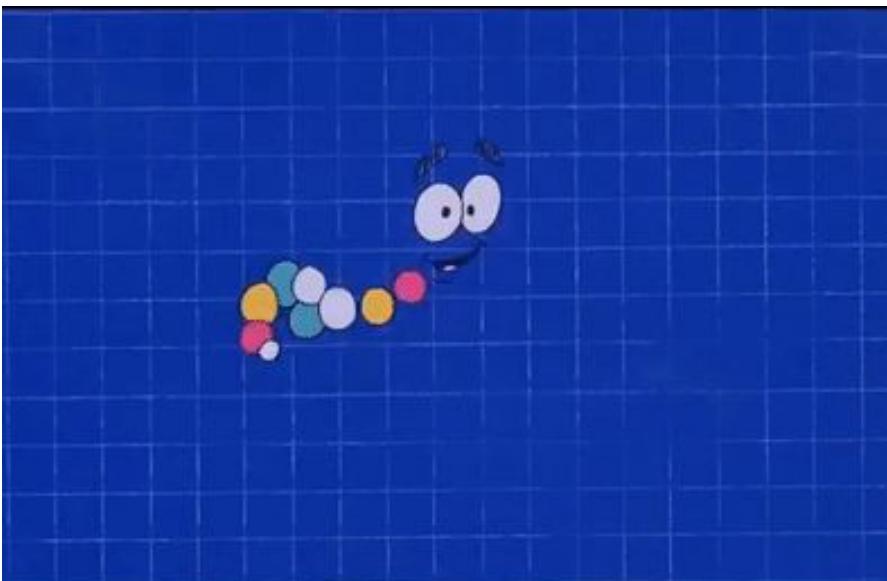
- Ambos os métodos são complementares à Dinâmica Molecular (MD), com o papel da MD potencialmente se deslocando para a geração e validação de dados.
- O AlphaFold2 Subamostrado abre novas possibilidades para o design de proteínas, a inferência de mecanismos de resistência a medicamentos , além de identificar novos estados metaestáveis.
- A capacidade do BioEmu de amortizar custos iniciais de MD e dados experimentais indica um caminho para a predição de funções biomoleculares em escala genômica.
- O avanço na integração de dados experimentais em larga escala e o aprimoramento da capacidade de modelar ambientes complexos (temperatura, pH, múltiplas moléculas) serão cruciais para o futuro desses emuladores.

pablitoarantes@gmail.com
pablo.arantes@ems.com.br
<https://pablo-arantes.github.io>

 website



 SCAN ME



 twitter



 SCAN ME

Obrigado!

