Análise Comparativa de Arquiteturas de Aprendizado Profundo para Predição da Energia de Ligação em Interações Proteína-Proteína

1st Ana Luísa Araújo Bastos
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
ana.araujo@dcc.ufmg.br

2nd Adenilson Arcanjo
Instituto Federal de Educação, Ciência e Tecnologia do Ceará, Campus Sobral
Sobral, Brasil
Programa Interunidades de Pós-Graduação em Bioinformática
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
adenilson.junior@ifce.edu.br

3rd Raquel Cardoso de Melo-Minardi
Departamento de Ciência da Computação
Programa Interunidades de Pós-Graduação em Bioinformática
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
raquelcm@dcc.ufmg.br

***Abstract*—Protein-protein interactions play a central role in biological regulation, molecular recognition, and drug discovery. However, experimental measurements of binding affinity remain costly and difficult to scale. In this work, we investigate different representations for predicting the binding free energy (ΔG) of protein-protein complexes, comparing sequence-based, structure-based, and hybrid deep learning approaches under the same experimental protocol. We used a high-confidence subset of the PPB-Affinity dataset, retaining complexes with affinities measured by SPR or ITC. Mutations were explicitly applied to the corresponding sequences, and mutant complexes were modeled with AlphaFold/ColabFold before structural graph construction. To reduce homologous data leakage, we adopted a cluster-based split using MMseqs2. Sequence representations were obtained with ESM-2 and evaluated through a global embedding baseline combined with a Multi-Layer Perceptron (MLP). Structural models were represented as residue-level graphs inspired by ProAffinity-GNN, incorporating Cα-Cα distances, ESM residue embeddings, and interface descriptors such as COCαDA, CSM, and aCSM-ALL. The ESM-2 + MLP baseline achieved the best linear correlation (Pearson = 0.705), while the hybrid GNN combining COCαDA descriptors and residue-level ESM embeddings achieved the best ranking performance (Spearman = 0.719). These results indicate that sequence representations are highly competitive for absolute ΔG prediction, while structural information can provide complementary value, especially for ranking protein-protein complexes. The findings also highlight the challenge of capturing subtle local effects, particularly in mutant variants, motivating future work on improved structural integration and $\Delta\Delta G$ prediction.**

***Resumo*—Interações proteína-proteína desempenham papel central na regulação biológica, no reconhecimento molecular e na descoberta de fármacos. Entretanto, a medição experimental da afinidade de ligação ainda apresenta alto custo e baixa escalabilidade. Neste trabalho, investigamos diferentes representações para a predição da energia livre de ligação (ΔG) em complexos proteína-proteína, comparando abordagens baseadas em sequência, estrutura e combinações híbridas sob o mesmo protocolo experimental. Utilizamos um subconjunto de alta confiabilidade do PPB-Affinity, mantendo complexos com afinidades determinadas por SPR ou ITC. As mutações foram aplicadas explicitamente às sequências correspondentes, e os complexos mutados foram modelados com AlphaFold/ColabFold antes da construção dos grafos estruturais. Para reduzir vazamento de informação por homologia, adotamos uma divisão baseada em clusters gerados pelo MMseqs2. As representações de sequência foram obtidas com o ESM-2 e avaliadas em um baseline com embeddings globais e uma rede Multi-Layer Perceptron (MLP). Os modelos estruturais foram representados como grafos em nível de resíduo, inspirados no ProAffinity-GNN, incorporando distâncias Cα-Cα, embeddings ESM por resíduo e descritores de interface como COCαDA, CSM e aCSM-ALL. O baseline ESM-2 + MLP apresentou a melhor correlação linear (Pearson = 0.705), enquanto o modelo híbrido GNN + COCαDA + ESM por resíduo obteve o melhor desempenho de ranqueamento (Spearman = 0.719). Esses resultados indicam que representações de sequência permanecem altamente competitivas para a predição absoluta de ΔG, enquanto informações estruturais podem atuar de forma complementar, especialmente na ordenação relativa dos complexos. Os achados também evidenciam o desafio de capturar efeitos locais sutis, particularmente em variantes mutantes, motivando trabalhos futuros em integração estrutural e predição de $\Delta\Delta G$.**

***Index Terms*—Protein–protein interactions, binding affinity prediction, protein language models, machine learning, bioinformatics.**

I. Introdução

Interações proteína–proteína (PPIs, do inglês *protein–protein interactions*) constituem a base de inúmeros processos biológicos, governando mecanismos essenciais,

como a transmissão de sinais, a regulação gênica e a resposta imunológica [1]. A estabilidade desses complexos está associada à afinidade de interação entre as proteínas, usualmente descrita por constantes de equilíbrio, como a constante de dissociação (K_d) e a constante de associação (K_a). Para uma interação entre duas proteínas livres, P_1 e P_2, que formam o complexo P_1P_2, essas constantes são expressas como $K_d = [P_1][P_2]/[P_1P_2]$ e $K_a = [P_1P_2]/([P_1][P_2])$, sendo inversamente relacionadas ($K_a = 1/K_d$). A partir dessas constantes, é possível calcular a variação de energia livre de ligação (ΔG), que fornece uma escala termodinâmica contínua para comparar complexos.

A ligação entre proteínas é sensível a alterações na sequência de aminoácidos. Mutações pontuais podem afetar o enovelamento da proteína, modificar a interface de interação e, consequentemente, influenciar a estabilidade do complexo, sendo frequentemente associadas à origem de diversas doenças [2]. Nesse contexto, a predição computacional de ΔG é relevante tanto para investigar os efeitos de mutações e mecanismos patológicos quanto para caracterizar a estabilidade de complexos proteína–proteína em diferentes contextos biológicos e biotecnológicos.

Apesar de sua importância, medições experimentais de afinidade, como *Surface Plasmon Resonance* (SPR) e *Isothermal Titration Calorimetry* (ITC), possuem alto custo e baixa escalabilidade [3]. Nesse contexto, abordagens computacionais podem auxiliar na priorização de interações e na análise de conjuntos extensos de proteínas.

Do ponto de vista computacional, prever a energia livre de ligação ainda é um desafio, pois envolve traduzir a complexidade biofísica da interface de interação em uma representação adequada para modelos de aprendizado de máquina. Tradicionalmente, são utilizados dois tipos de representação: métodos baseados nas estruturas tridimensionais das proteínas e métodos baseados exclusivamente na sequência.Este trabalho realiza uma comparação sistemática entre representações baseadas em sequência, estrutura e abordagens híbridas sob um protocolo experimental rigoroso que reduz vazamento por homologia, permitindo uma avaliação mais realista da capacidade de generalização dos modelos.

Durante o POC I, investigamos o uso dos Modelos de Linguagem de Proteínas (pLMs, do inglês *protein language models*), especialmente o ESM-2 [4]. Esses modelos tratam sequências de aminoácidos como uma linguagem biológica e aprendem representações vetoriais ricas (*embeddings*) que capturam padrões evolutivos e propriedades físico-químicas, sem necessidade de estruturas tridimensionais. Desse modo, combinamos os *embeddings* do ESM-2 a uma rede *Multi-Layer Perceptron* (MLP), estabelecendo um *baseline* baseado em sequência para o problema.

Neste trabalho, além das abordagens baseadas em sequência, também investigamos o uso de informações estruturais. Métodos baseados em estrutura permitem representar diretamente a geometria da interface e os contatos entre resíduos, aspectos biologicamente relevantes para a formação do complexo. Assim, avaliamos diferentes representações para a predição de ΔG em interações proteína–proteína, comparando abordagens baseadas em sequência, grafos estruturais e descritores de interface sob o mesmo conjunto de dados e o mesmo protocolo experimental.

Os resultados indicam que as representações baseadas em sequência permanecem altamente competitivas, enquanto descritores estruturais isolados apresentam limitações quando utilizados diretamente. No entanto, a combinação entre embeddings por resíduo e descritores estruturais de interface mostrou potencial complementar, especialmente para o ranqueamento das amostras.

O restante deste artigo está organizado da seguinte forma: na Seção 2, apresentamos o referencial teórico e trabalhos relacionados. Na Seção 3, descrevemos o conjunto de dados e o *pipeline* proposto. Na Seção 4, discutimos os resultados obtidos e, por fim, na Seção 5, apresentamos as conclusões e direções futuras.

II. Trabalhos Relacionados

O problema da predição de afinidade de ligação e do impacto de mutações em complexos proteicos tem sido abordado sob diferentes perspectivas, que vão desde métodos baseados em física até métodos de aprendizado profundo. Nesta seção, discutimos a evolução dos conjuntos de dados e duas vertentes principais de modelagem: abordagens baseadas em estrutura e abordagens baseadas em sequência.

A. Conjuntos de Dados

A eficácia de modelos supervisionados para afinidade em PPIs é historicamente limitada pela escassez de dados de alta qualidade. O banco de dados SKEMPI v2.0 [5] tem servido como o padrão-ouro para a predição de mudanças na energia livre de ligação ($\Delta\Delta G$) após mutações, contendo dados termodinâmicos para milhares de variantes. No entanto, o número restrito de estruturas limita a capacidade de generalização de modelos profundos. Recentemente, o *dataset* PPB-Affinity [6] foi proposto com o objetivo de mitigar esse gargalo. Ele integra e padroniza dados de múltiplas fontes (PDBbind, SAbDab, ATLAS), resultando em mais de 12.000 complexos com K_D experimental. Além do volume, um diferencial importante do PPB-Affinity é a anotação rigorosa de cadeias de receptores e ligantes, fundamental para o treinamento de modelos [6].

B. Abordagens Baseadas em Estrutura

Métodos estruturais utilizam explicitamente a geometria tridimensional do complexo proteico. O ProAffinity-GNN [7] é um exemplo recente dessa classe, modelando o complexo como grafos intra e intermoleculares distintos e utilizando redes de atenção em grafos (GATs) para capturar dependências espaciais na interface. Além das Redes Neurais Baseadas em Grafos (GNNs), alguns trabalhos exploram as chamadas assinaturas estruturais para extrair informações da estrutura. Nessa linha, o mCSM-PPI2 [8] baseia-se nos métodos CSM [9,10] e combina padrões de distância, propriedades físico-químicas, informação evolutiva, contatos não covalentes e termos energéticos para representar o ambiente molecular de resíduos mutados e assim predizer o $\Delta\Delta G$ após mutações.

C. Abordagens Baseadas em Sequência

A popularização dos pLMs, como o ESM-2, permitiu a extração de características ricas diretamente da sequência, contornando a necessidade de modelagem estrutural explícita. Trabalhos como o PLM-interact [11] e o PPLM [12], demonstram que concatenar as sequências do par interagente na entrada do Transformer é mais eficiente para tarefas de PPI do que processá-las de forma independente. O PLM-interact alcançou melhorias de até 28% em AUPR sobre métodos anteriores ao permitir que o mecanismo de atenção do ESM-2 opere através da fronteira das duas proteínas, capturando implicitamente a compatibilidade da interface mutante [11]. O PPLM refina essa ideia introduzindo codificação posicional híbrida para distinguir interações intra e intermoleculares, atingindo correlações de Pearson em torno de 0,76 na tarefa de predição de afinidade [12].

Outra direção busca incorporar informações estruturais ou relacionais aos modelos baseados em sequência. O TT3D [13], por exemplo, enriquece a entrada do modelo de linguagem com sequências "3Di", um alfabeto estrutural gerado pelo Foldseek [14], buscando incorporar informação estrutural local de forma eficiente.

Nossa proposta consiste em comparar diretamente modelos baseados em sequência, estrutura e combinações híbridas sob um mesmo protocolo experimental. Para isso, avaliamos representações baseadas em pLMs, GNNs e assinaturas estruturais, buscando identificar os potenciais e limitações de cada abordagem na predição de ΔG em interações proteína-proteína.

Embora diversos trabalhos tenham explorado modelos baseados em sequência ou estrutura, poucos realizam uma comparação controlada dessas diferentes representações utilizando exatamente o mesmo conjunto de dados e protocolo experimental. Este trabalho busca preencher essa lacuna.

III. Metodologia

A metodologia deste trabalho foi construída a partir do *pipeline* desenvolvido no POC I, expandindo-o para incorporar informações estruturais dos complexos proteína-proteína. Todos os modelos foram avaliados em termos das correlações de Pearson(r) e Spearman(ρ) sobre o mesmo conjunto de dados e sob o mesmo protocolo de divisão, de modo a permitir uma comparação controlada entre representações baseadas em sequência, estrutura e combinações híbridas.

A. Construção do Conjunto de Dados

Utilizamos como base o *dataset* PPB-Affinity [6], que reúne complexos proteína-proteína com medidas experimentais de afinidade. Para assegurar a qualidade, aplicou-se um filtro que manteve apenas amostras cujas afinidades foram determinadas por métodos de alta precisão (SPR e ITC). De um total inicial de 12.062 pares, restaram 4.209 complexos de alta confiança. Os valores de constante de dissociação (K_d) foram convertidos para Energia Livre de Gibbs (ΔG), padronizando a escala de regressão, conforme:

$$[\ \Delta G = RT \ln(K_d)]$$

Para amostras sem temperatura experimental anotada, assumiu-se o padrão de $298\,K$.

Além disso, como grande parte das amostras contém padrões de mutação associados ao complexo, foi aplicado um procedimento de substituição dos resíduos mutados por meio de um *script* em Python que processa as strings de mutação (ex: `A_Y29F`). O algoritmo identifica a cadeia alvo, localiza o resíduo na posição indicada (ou próximo) e realiza a substituição na sequência de caracteres.

Por fim, utilizamos as sequências com aplicação das mutações para modelar a estrutura dos complexos com o AlphaFold2/ColabFold [15]. Para isso, as sequências do receptor e do ligante foram fornecidas como entrada pareada, permitindo a predição do complexo proteína-proteína já na forma mutada.

Após a remoção de entradas com falhas, o conjunto final consistiu em 4.115 pares.

B.

Estratégia de Partição (*Cluster-Split)* A divisão aleatória de dados biológicos frequentemente resulta em *data leakage* devido à homologia evolutiva, ou seja, quando proteínas são muito parecidas por terem origem evolutiva comum. Assim, proteínas semelhantes podem aparecer tanto no treino quanto no teste, fazendo com que o desempenho do modelo pareça melhor do que realmente é.. Para mitigar isso, implementamos uma estratégia de particionamento baseada em pares de *clusters*. Todas as sequências únicas presentes no *dataset* foram agrupadas utilizando MMseqs2 [16], com parâmetros de 50% de identidade de sequência e 80% de cobertura. Cada par de interação (P_A, P_B) foi então mapeado para uma tupla de clusters (C_A, C_B). Em outras palavras, pares (P_A, P_B) e (Q_A, Q_B), tais que P_A e Q_A são similares, e, P_B e Q_B são similares, de acordo com as definições acima, então os pares (P_A, P_B) e (Q_A, Q_B) são alocados no mesmo subconjunto (treino, validação or teste). A divisão em Treino (80%), Validação (10%) e Teste (10%) foi realizada de forma aleatória sobre as tuplas (C_A, C_B)

C. Representações de Sequência

As representações de sequência foram geradas com o modelo ESM-2 de 8M de parâmetros e *embeddings* de tamanho 320 [4], utilizando as bibliotecas Hugging Face Transformers e PyTorch. Para o baseline global, os *embeddings* por resíduo foram agregados por *mean pooling*, produzindo um vetor fixo para cada cadeia. Os vetores do receptor e do ligante foram concatenados e utilizados como entrada para uma rede *Multi-Layer Perceptron* (MLP), responsável pela predição de ΔG.

Também avaliamos o uso de *embeddings* ESM por resíduo nos modelos estruturais, onde cada resíduo do grafo recebe sua representação local.

D. Representações Estruturais e Arquitetura das GNNs

Nos modelos estruturais, cada complexo proteína-proteína foi representado por grafos em nível de resíduo. Essa estratégia foi inspirada no ProAffinity-GNN [7], que modela

separadamente as relações intramoleculares e intermoleculares do complexo. Assim, para cada amostra, foram construídos três grafos: um grafo intermolecular, representando os contatos entre resíduos do receptor e do ligante, e dois grafos intramoleculares, representando as relações internas de cada uma das cadeias.

Em todos os grafos, os nós correspondem aos resíduos das proteínas. As arestas foram definidas a partir de critérios geométricos, principalmente por distâncias entre átomos Cα, que correspondem ao carbono central da estrutura principal de cada aminoácido. No grafo intermolecular, as arestas conectam resíduos pertencentes a proteínas diferentes, representando a interface de interação. Nos grafos intramoleculares, as arestas conectam resíduos da mesma cadeia, capturando o contexto estrutural interno do receptor e do ligante.

Cada um dos três grafos é processado por módulos de GNN de forma independente. Em seguida, as representações aprendidas para o grafo intermolecular, para o grafo do receptor e para o grafo do ligante são agregadas e concatenadas, formando uma representação única do complexo. Essa representação final é enviada a uma rede MLP, responsável pela regressão do valor de ΔG.

Além da distância Cα-Cα, foram avaliados descritores estruturais adicionais. O COCαDA foi incorporado como atributo de aresta, representando contatos atômicos e tipos de interação entre resíduos. As assinaturas CSM e aCSM-ALL foram utilizadas como atributos de nó, descrevendo o ambiente estrutural local dos resíduos por meio de padrões de distância e propriedades físico-químicas.

E. Modelos Avaliados

Foram avaliadas as seguintes configurações:

- **1 - ESM-2 + MLP**: baseline com *embeddings* globais de sequência;
- **2 - GNN simples**: arquitetura baseada em três grafos, utilizando *one-hot* dos 20 aminoácidos como atributo de nó e distância Cα-Cα como atributo de aresta;
- **3 - GNN + ESM por resíduo**: arquitetura estrutural com *embeddings* ESM associados aos nós;
- **4 - GNN + COCαDA**: arquitetura estrutural com contatos calculados pelo COCαDA incorporados às arestas;
- **5 - GNN + CSM**: arquitetura estrutural com assinaturas CSM utilizadas como atributos dos nós;
- **6 - GNN + aCSM-ALL**: arquitetura estrutural com assinaturas aCSM-ALL utilizadas como atributos dos nós;
- **7 - GNN + COCαDA + ESM por resíduo**: modelo híbrido combinando *embeddings* ESM nos nós e COCαDA nas arestas.

Essa organização permite comparar, sob a mesma arquitetura geral, o efeito de diferentes fontes de informação: geometria simples, descritores estruturais, representações de sequência e combinações híbridas.

IV. Resultados e Discussão

Esta seção apresenta os resultados obtidos para as diferentes representações avaliadas na predição de ΔG em interações proteína–proteína. Todos os modelos foram avaliados no mesmo conjunto de teste, utilizando as correlações de Pearson e Spearman como métricas principais, com os resultados completos apresentados na Tabela I. A correlação de Pearson mede a associação linear entre valores preditos e experimentais, enquanto a correlação de Spearman avalia a preservação do ranqueamento relativo das amostras.

Antes de analisar os resultados, é importante destacar que a estratégia de *cluster split* utilizada neste trabalho torna o problema significativamente mais desafiador do que uma divisão aleatória tradicional. Ao agrupar proteínas homólogas e garantir que proteínas do mesmo cluster não apareçam simultaneamente nos conjuntos de treinamento e teste, reduz-se a similaridade entre os dados utilizados para treinamento e avaliação. Dessa forma, os resultados obtidos fornecem uma estimativa mais realista da capacidade de generalização dos modelos para proteínas ainda não observadas, evitando superestimações de desempenho decorrentes de vazamento de informação por homologia.

TABLE I
Desempenho dos modelos avaliados no conjunto de teste.

Modelo	Pearson	Spearman
ESM-2 + MLP	0.705	0.653
GNN simples	0.616	0.535
GNN + ESM	0.682	0.656
GNN + COCaDA	0.335	0.323
GNN + CSM	0.274	0.250
GNN + aCSM-ALL	0.328	0.303
GNN + COCaDA + ESM	0.638	0.719

O baseline baseado em sequência, composto por embeddings globais do ESM-2 e uma MLP, apresentou (r) de 0.705 e (ρ) de 0.653. Esse resultado confirma que representações globais de sequência são altamente informativas para a predição de ΔG, mesmo sem uso explícito de estrutura tridimensional, o que condiz com o Dogma de Anfinsen [17], que diz que toda a informação necessária para enovelar uma proteína está contida em sua sequência de aminoácidos. Além disso, o ESM-2 foi treinado em um conjunto massivo de dados de cerca de 250 milhões de proteínas [4], sendo altamente capaz de capturar padrões evolutivos e biofísicos relevantes.

Já nos modelos de estrutura, a GNN simples, baseada apenas na geometria Cα-Cα e no one-hot, apresentou (r) de 0.616 e (ρ) de 0.535. Embora inferior ao modelo de sequência, esse resultado indica que a representação estrutural por grafos contém sinal relevante para a tarefa. Ainda assim, a geometria sozinha parece não ser suficiente para superar as representações aprendidas pelo ESM-2, pois, ao contrário dos seus *embeddings*, não incorpora uma representação contextual aprendida a partir de padrões evolutivos, considerando apenas a estrutura do complexo específico.

Os modelos que utilizaram descritores estruturais isolados apresentaram desempenho ainda mais baixo. A GNN + COCαDA obteve (r) de 0.335 e (ρ) de 0.323, enquanto a GNN

+ CSM e a GNN + aCSM-ALL obtiveram, respectivamente, correlação de Spearman de 0.250 e 0.303. Isso não significa que não sejam relevantes para a análise de interações, mas provavelmente as representações de alta dimensionalidade dos CSMs ou esparsas do COCαDA introduziram ruído.

Já a incorporação de *embeddings* ESM por resíduo à GNN inicial elevou o desempenho para (r)=0.682 e (ρ)=0.656. Isso sugere que uma abordagem híbrida entre informação local de sequência e estrutura melhora a representação do complexo, ficando bem mais próxima do baseline de sequência.

O melhor desempenho em ranqueamento foi obtido pelo modelo híbrido(GNN + COCαDA + ESM por resíduo), que alcançou (r)=0.638 e (ρ)=0.719. Esse resultado é relevante porque, embora a correlação de Pearson seja inferior ao baseline ESM-2 + MLP, a de Spearman é a maior entre todos os modelos avaliados. Esse comportamento sugere que a incorporação de informações estruturais melhora a ordenação relativa dos complexos, mas ainda não produz calibração suficiente para prever corretamente os valores absolutos de ΔG.

Dessa forma, os resultados indicam que a escolha do modelo mais adequado depende do objetivo da análise. Para aplicações em que é mais importante prever valores absolutos de ΔG com maior precisão, o baseline ESM-2 + MLP tende a ser a melhor opção, por apresentar o maior Pearson. Por outro lado, em cenários nos quais o interesse principal é comparar diferentes pares proteína-proteína e identificar quais apresentam interações relativamente mais fortes, a GNN + COCαDA + ESM por resíduo se destaca.

A. Comparação com Modelos da Literatura

A Tabela II apresenta uma comparação com modelos relacionados da literatura. Essa comparação deve ser interpretada com cautela, pois os protocolos, conjuntos de dados e estratégias de validação não são idênticos. Enquanto trabalhos como PPLM e PPB-Affinity utilizam validação cruzada sobre o PPB-Affinity completo, este trabalho utiliza um subconjunto curado e uma divisão por clusters, projetada para reduzir a contaminação por homologia entre treino e teste. Já o Pro-Affinity-GNN utiliza ainda outros conjuntos.

TABLE II
COMPARAÇÃO COM MODELOS RELACIONADOS DA LITERATURA.

Modelo	Tipo	Pearson	Spearman
ProAffinity-GNN [7]	Híbrido/GNN	0.669	–
ESM2-Affinity [12]	Sequência	0.669	0.649
PPLM-Affinity [12]	Seq. pareada	0.764	0.752
ESM-2 + MLP*	Sequência	0.705	0.653
GNN + ESM*	Híbrido/GNN	0.682	0.656
GNN + COCaDA + ESM*	Híbrido/GNN	0.638	0.719

* Modelos desenvolvidos neste trabalho.

Em relação aos modelos da literatura, o baseline ESM-2 + MLP apresentou desempenho competitivo, especialmente considerando que utiliza uma arquitetura simples baseada em concatenação de *embeddings* globais. O modelo híbrido GNN + COCaDA + ESM por resíduo, por sua vez, apresentou o melhor Spearman entre os modelos desenvolvidos neste trabalho, aproximando-se do PPLM-Affinity na capacidade de ranqueamento, embora ainda com menor correlação linear.

Esses resultados reforçam duas observações principais. Primeiro, modelos de sequência continuam sendo muito fortes para a predição de afinidade proteína–proteína, especialmente quando treinados em grandes bases e capazes de capturar relações evolutivas. Segundo, a informação estrutural pode ser complementar, mas sua contribuição depende fortemente da forma de incorporação ao modelo. No presente estudo, descritores estruturais isolados não melhoraram o desempenho, enquanto sua combinação com *embeddings* por resíduo resultou em ganho expressivo no ranqueamento.

B. Análise Qualitativa: Limites de Representações Globais de Sequência

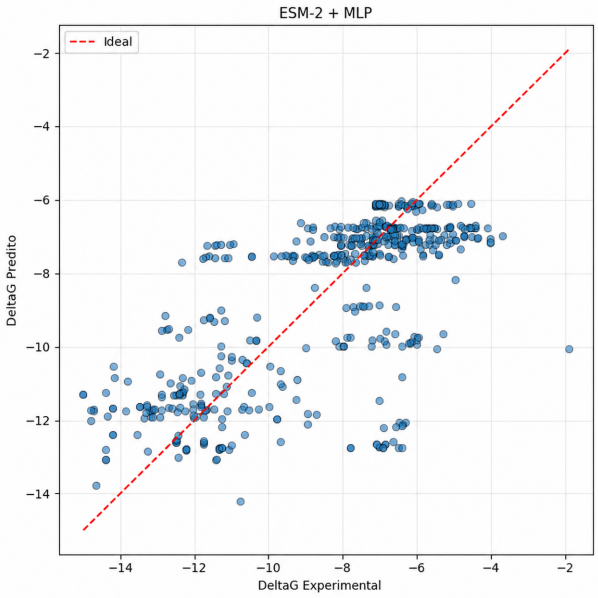

Fig. 1. Dispersão entre valores experimentais e preditos de ΔG para o modelo baseado apenas em sequência com (r)=0.705 e (ρ)=0.653. A linha tracejada representa a identidade ($y = x$).

A inspeção visual das Figuras 1 e 2 mostra que ambos os modelos apresentam agrupamentos de pontos aproximadamente alinhados horizontalmente, nos quais os valores preditos variam menos do que os valores experimentais. Esses agrupamentos provavelmente correspondem a variantes de um mesmo complexo proteico, o que sugere que os modelos capturam tendências médias associadas aos complexos ou famílias proteicas, mas ainda apresentam dificuldade em distinguir variações pontuais.

No modelo baseado apenas em sequência, uma possível explicação é o uso de *mean pooling* global, que agrega os *embeddings* de todos os resíduos da cadeia em um único vetor. Nesse processo, o efeito de uma mutação pontual pode ser diluído no vetor médio, tornando a representação menos sensível a alterações locais na interface. Já no modelo híbrido, a presença de *embeddings* por resíduo e descritores estruturais

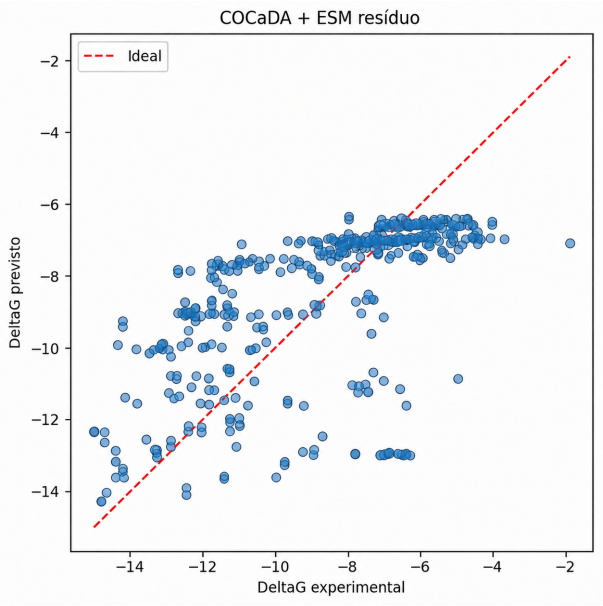

Fig. 2. Dispersão entre valores experimentais e preditos de ΔG para o modelo final combinando sequência e estrutura com (r)=0.638 e (ρ)=0.719. A linha tracejada representa a identidade $(y = x)$.

melhora o ranqueamento das amostras, mas não elimina totalmente a limitação na variação dos valores preditos.

V. Conclusão

A predição computacional da energia livre de ligação (ΔG) em interações proteína-proteína permanece como um problema relevante em bioinformática, com aplicações em engenharia de proteínas, biomedicina e descoberta de fármacos. Neste trabalho, investigamos diferentes formas de representar complexos proteína–proteína para essa tarefa, comparando abordagens baseadas em sequência, estrutura e combinações híbridas sob o mesmo protocolo experimental.

De forma geral, os resultados mostraram que, sob um protocolo rigoroso de generalização, modelos baseados exclusivamente em representações de sequência, como o ESM-2, permanecem extremamente competitivos para a predição da afinidade de interação proteína-proteína. Por outro lado, a integração entre representações de sequência e informações estruturais mostrou-se capaz de melhorar o ranqueamento relativo dos complexos, indicando que abordagens híbridas representam um caminho promissor para o desenvolvimento de modelos mais robustos e generalizáveis.

No caso do modelo baseado apenas em sequência, esses resultados reforçam que os *embeddings* globais do ESM-2 foram capazes de fornecer uma representação robusta dos complexos. Mesmo com uma arquitetura relativamente simples, o modelo capturou padrões bioquímicos e evolutivos relevantes a partir da sequência primária.

A análise das abordagens estruturais, entretanto, mostrou que a simples adição de informação tridimensional não garante ganhos de desempenho. A contribuição da estrutura mostrou-se dependente tanto da forma de representação quanto da estratégia de integração ao modelo, e a geometria Cα-Cα e os descritores estruturais isolados não foram suficientes para superar o baseline de sequência. Por outro lado, a combinação entre *embeddings* por resíduo e descritores de contato apresentou melhor capacidade de ranqueamento, sugerindo que a informação estrutural é mais efetiva quando incorporada de forma complementar à informação local de sequência, especialmente no contexto da interface de interação.

Esses achados sugerem que o principal desafio não está apenas em qual informação utilizar, mas em como incorporá-la ao modelo. Em especial, a predição de ΔG em variantes mutantes exige sensibilidade a alterações locais sutis, que podem ser diluídas por representações globais ou pouco aproveitadas por descritores estruturais esparsos. Assim, modelos híbridos parecem ser um caminho promissor, desde que consigam integrar adequadamente sequência, estrutura e contatos de interface.

Como limitações, o estudo utiliza estruturas previstas pelo AlphaFold para os complexos com mutações, e não estruturas experimentais, além de os resultados não serem diretamente comparáveis com outros trabalhos da literatura.

Como trabalhos futuros, portanto, pretende-se avaliar o mesmo *pipeline* sob o protocolo tradicional de validação cruzada em 5 *folds*, permitindo uma comparação direta com PPB-Affinity e PPLM-Affinity. Também será relevante aplicar o PPLM sob o protocolo de divisão por clusters adotado neste estudo, a fim de verificar se modelos de linguagem especializados em pares de proteínas mantêm seu desempenho em cenários de generalização mais restritivos.

É importante também comparar diretamente a abordagem proposta com arquiteturas estruturais da literatura, como o ProAffinity-GNN, sob um mesmo conjunto de dados e protocolo experimental. Além disso, os resultados com COCαDA indicam a necessidade de investigar formas alternativas de incorporação dos contatos atômicos e estratégias de agregação mais robustas.

Por fim, uma extensão deste trabalho é adaptar a arquitetura para a predição de $\Delta\Delta G$, permitindo modelar diretamente o impacto de mutações específicas na afinidade de ligação. Essa direção aproxima o método de aplicações em engenharia de proteínas e análise funcional de variantes, nas quais pequenas alterações locais podem produzir efeitos relevantes na estabilidade e na função dos complexos proteína-proteína.

References

[1] J. F. Greenblatt, B. M. Alberts, and N. J. Krogan, “Discovery and significance of protein-protein interactions in health and disease,“ *Cell*, vol. 187, no. 23, pp. 6501-6517, 2024.

[2] H. C. Jubb, A. P. Pandurangan, M. A. Turner, *et al.*, “Mutations at protein–protein interfaces: small changes over big surfaces have large impacts on human health,” *Prog. Biophys. Mol. Biol.*, vol. 128, pp. 3-13, 2017.

[3] V. Upadhyay, A. Lucas, C. Patrick, and K. M. G. Mallela, “Isothermal titration calorimetry and surface plasmon resonance methods to probe protein-protein interactions,“ *Methods*, vol. 225, pp. 52-61, 2024.

[4] Z. Lin *et al.*, “Evolutionary-scale prediction of atomic-level protein structure with a language model,“ *Science*, vol. 379, no. 6637, pp. 1123-1130, 2023.

[5] J. Jankauskaitė *et al.*, "SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation," *Bioinformatics*, vol. 35, no. 3, pp. 462-469, 2018.

[6] H. Liu *et al.*, "PPB-Affinity: Protein-Protein Binding Affinity dataset for AI-based protein drug discovery," *Sci. Data*, vol. 11, no. 1, p. 1316, 2024.

[7] Z. Zhou *et al.*, "ProAffinity-GNN: A Novel Approach to Structure-Based Protein–Protein Binding Affinity Prediction via a Curated Data Set and Graph Neural Networks," *J. Chem. Inf. Model.*, vol. 64, no. 23, pp. 8796-8808, 2024.

[8] C. H. M. Rodrigues, Y. Myung, D. E. V. Pires, D. B. Ascher, "mCSM-PPI2: predicting the effects of mutations on protein–protein interactions," *Nucleic Acids Research*, vol. 47, pp. W338–W344, 2019.

[9] D.E. Pires, R.C. de Melo-Minardi, M.A. dos Santos, *et al.*," Cutoff Scanning Matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns," *BMC Genomics*, vol. 12, 2011.

[10] D.E. Pires, R.C. de Melo-Minardi*et al.*, " aCSM: noise-free graph-based signatures to large-scale receptor-based ligand prediction, ", *Bioinformatics*, vol. 29, pp. 855-861, 2013.

[11] D. Liu *et al.*, "PLM-interact: extending protein language models to predict protein-protein interactions," *bioRxiv*, 2024.

[12] J. Liu, H. Chen, and Y. Zhang, "A paired sequence language model for protein-protein interaction modeling," *Nat Commun*, 2026.

[13] S. Sledzieski, K. Devkota, R. Singh, L. Cowen, and B. Berger, "TT3D: Leveraging precomputed protein 3D sequence models to predict protein-protein interactions," *Bioinformatics*, vol. 39, no. 11, p. btad663, 2023.

[14] M. van Kempen *et al.*, "Fast and accurate protein structure search with Foldseek," *Nature Biotechnology*, vol. 42, pp. 243–246, 2024.

[15] J. Jumper *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, pp. 583–589, 2021.

[16] M. Steinegger & J. Söding, "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets," *Nature Biotechnology*, vol. 35, pp. 1026-1028, 2017.

[17] Anfinsen, Christian B. and Haber, Edgar and Sela, Michael and White, F. H. Jr., "The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 47, pp. 1309-1314, 1961.