Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz
Instituto Gonçalo Moniz

dQF
Departamento
de Química Fundamental

VIRTUS IMPAVIDA
UNIVERSIDADE
FEDERAL
DE PERNAMBUCO

UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS DA NATUREZA
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

MANASSÉS FRANCISCO DO NASCIMENTO FILHO

AVALIAÇÃO DOS ALGORITMOS DE APRENDIZAGEM PARA PREDIÇÃO DA ENERGIA LIVRE DE GIBBS DE INTERAÇÕES PROTEÍNA-PROTEÍNA

Recife
2025

MANASSÉS FRANCISCO DO NASCIMENTO FILHO

AVALIAÇÃO DOS ALGORITMOS DE APRENDIZAGEM PARA PREDIÇÃO DA ENERGIA LIVRE DE GIBBS DE INTERAÇÕES PROTEÍNA-PROTEÍNA

Dissertação apresentada ao Programa de Pós-Graduação em Química da Universidade Federal de Pernambuco, como requisito para obtenção do título de mestre em Química. Área de concentração: Química Teórica.

Orientador: Profº Drº Roberto Dias Lins Neto

Coorientador: Drº Elton José Ferreira Chaves

Recife

2025

.Catalogação de Publicação na Fonte. UFPE - Biblioteca Central

Nascimento Filho, Manasses Francisco do.
Avaliação dos algoritmos de aprendizagem para predição da energia livre de gibbs de interações proteína-proteína / Manasses Francisco do Nascimento Filho. - Recife, 2025.
135 f.: il.

Dissertação (Mestrado) - Universidade Federal de Pernambuco, Departamento de Química Fundamental, Programa de Pós Graduação em Química, 2025.
Orientação: Roberto Dias Lins Neto.
Coorientação: Elton José Ferreira Chaves.
Inclui referências, apêndices e anexo.

1. Aprendizado de Máquina; 2. Energia Livre de Gibbs; 3. Super Learner. I. Lins Neto, Roberto Dias. II. Chaves, Elton José Ferreira. III. Título.

UFPE-Biblioteca Central

MANASSÉS FRANCISCO DO NASCIMENTO FILHO

"AVALIAÇÃO DOS ALGORITMOS DE APRENDIZAGEM PARA PREDIÇÃO DA ENERGIA LIVRE DE GIBBS DE INTERAÇÕES PROTEÍNA-PROTEÍNA"

Dissertação apresentada ao Programa de Pós-Graduação no Departamento de Química Fundamental da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Química.

Aprovada em: 26/06/2025

BANCA EXAMINADORA

gov.br Documento assinado digitalmente
ROBERTO DIAS LINS NETO
Data: 26/06/2025 12:55:08-0300
Verifique em https://validar.iti.gov.br

Prof. Roberto Dias Lins Neto (Orientador)
Centro de Pesquisas Aggeu Magalhães (CPqAM) - FIOCRUZ

gov.br Documento assinado digitalmente
DANILO FERNANDES COELHO
Data: 26/06/2025 12:19:26-0300
Verifique em https://validar.iti.gov.br

Prof. Danilo Fernandes Coelho
Universidade Federal de Pernambuco

gov.br Documento assinado digitalmente
JOSE LICARION PINTO SEGUNDO NETO
Data: 26/06/2025 12:08:25-0300
Verifique em https://validar.iti.gov.br

Prof. José Licarion Pinto Segundo Neto
Universidade do Estado do Rio de Janeiro

AGRADECIMENTOS

Agradeço à CAPES pelo apoio financeiro a qual espero que os investimentos em educação, especialmente a básica, aumentem. Ao Dr. Roberto Lins, por abrir portas ao grupo Protein Engineering and Structural Genomics, pelo qual tenho grande admiração. Aos professores e amigos incríveis que, direta ou indiretamente, me ajudaram nessa tentativa de compreender a natureza.

À Júlia, minha companheira, por estar ao meu lado ao longo dessa jornada e trazer um pouco de calmaria. Por fim, mas não menos importante, ao Movimento Casa do Estudante (MCE) e a todas as políticas assistencialistas que ajudam a reduzir desigualdades. Foi graças a esse suporte que conclui minha graduação e cheguei até aqui, pois me deram um lar e um ambiente silencioso para estudar e me desenvolver. Hoje estou retribuindo atuando na educação básica, com amor, serei eternamente grato.

RESUMO

Neste trabalho, contribuiu-se com a avaliação sistemática dos modelos de regressão utilizados na construção de um metamodelo capaz de predizer a afinidade de ligação de complexos proteína-proteína, o qual apresentou correlação de Pearson (r) igual a 0,70. A predição da variação de energia livre de ligação (ΔG) durante a formação desses complexos constitui um desafio na bioinformática estrutural, em virtude da complexidade dessas interações e da influência de diversos fatores físico-químicos. Tal predição é essencial para a compreensão de mecanismos biomoleculares, bem como para o desenvolvimento de fármacos e o projeto de terapias baseadas em proteínas, como anticorpos e vacinas. Os métodos computacionais tradicionais, como simulações baseadas em Dinâmica Molecular e Monte Carlo, embora altamente precisos, apresentam elevado custo computacional, o que limita sua aplicabilidade na triagem de grandes quantidades de proteínas. Como alternativa, métodos baseados em redes neurais, grafos e técnicas de *Deep Learning*, fundamentados em dados de sequência ou estrutura proteica, têm sido amplamente explorados e aprimorados. Diante desse cenário, avaliou-se o desempenho de dez algoritmos de regressão distintos, majoritariamente métodos de *Machine Learning* (ML), utilizados como base na arquitetura de um metamodelo de regressão com abordagem *Super Learner* (SL), cujo objetivo é predizer valores de ΔG a partir de descritores de interface calculados por meio do software Rosetta. Os modelos foram treinados com 526 estruturas no formato .pdb e seus respectivos valores experimentais de ΔG, considerando-se apenas dados de alta resolução (≤ 3,5 Å). Como melhor desempenho, obteve-se o modelo SL_MLP (*Super Learner* acoplado ao metamodelo Multilayer Perceptron), com r = 0,70, RMSE = 1,91 kcal/mol e R^2 = 0,48, com tempo de execução para o cálculo de energia inferior a cinco minutos em um computador de uso pessoal, equipado com 8 GB de memória RAM e processador Intel(R) Core(TM) i5-7300HQ (quatro núcleos, frequência base de 2,50 GHz, 6 MB de cache L3. O desempenho dos modelos foi comparado ao de ferramentas consolidadas com a mesma proposta, como Prodigy e Area Affinity, amplamente utilizadas para estimativas de ΔG com baixo custo computacional. Nesse teste, foi observado que, mesmo uma regressão linear simples aplicada aos descritores utilizados foi capaz de superar significativamente essas ferramentas. Além disso, os avanços obtidos neste trabalho contribuíram diretamente para a

publicação do artigo "*Estimating Absolute Protein–Protein Binding Free Energies by a Super Learner Model*", na revista Journal of Chemical Information and Modeling, reconhecendo a relevância da abordagem proposta no contexto internacional da bioinformática estrutural. Os resultados obtidos reforçam o potencial dessa abordagem como ferramenta de triagem molecular, com baixo custo computacional e aplicabilidade prática. A avaliação realizada neste estudo contribui para o aprimoramento e a escolha criteriosa dos modelos que compõem o SL, visando sua aplicação em ambientes de triagem e análise de interação molecular.

Palavras-chave: Aprendizado de Máquina; *Super Learner*; Interface de Interações Proteína-Proteína; Energia Livre de Gibbs.

ABSTRACT

In this study, a systematic evaluation of regression models was carried out to construct a metamodel capable of predicting the binding affinity of protein–protein complexes, which achieved a Pearson correlation coefficient (r) of 0.70. Predicting the variation in binding free energy (ΔG) during the formation of such complexes remains a major challenge in structural bioinformatics due to the complexity of these interactions and the influence of various physicochemical factors. Such predictions are essential for understanding biomolecular mechanisms as well as for drug development and the design of protein-based therapies, including antibodies and vaccines. Traditional computational methods, such as Molecular Dynamics and Monte Carlo simulations, although highly accurate, are computationally expensive, limiting their applicability in large-scale protein screening. As an alternative, methods based on neural networks, graphs, and Deep Learning techniques—using protein sequence or structure data—have been widely explored and refined. In this context, the performance of ten different regression algorithms—primarily Machine Learning (ML) approaches—was evaluated as the foundation for the architecture of a regression metamodel using the Super Learner (SL) framework. The objective was to predict ΔG values based on interface descriptors computed via the Rosetta software. The models were trained using 526 structures in .pdb format along with their respective experimental ΔG values, considering only high-resolution data (≤ 3.5 Å). The best-performing model was the SL_MLP (Super Learner coupled with a Multilayer Perceptron metamodel), which achieved r = 0.70, RMSE = 1.91 kcal/mol, and R^2 = 0.48, with a computation time of less than five minutes on a personal computer equipped with 8 GB of RAM and an Intel(R) Core(TM) i5-7300HQ processor (four cores, base frequency of 2.50 GHz, 6 MB L3 cache). The performance of the models was compared with that of well-established tools with similar goals, such as Prodigy and Area Affinity, which are widely used for low-cost ΔG estimation. In this comparison, even a simple linear regression applied to the same descriptors significantly outperformed these tools. Furthermore, the advances achieved in this study directly contributed to the publication of the article "Estimating Absolute Protein–Protein Binding Free Energies by a Super Learner Model" in the Journal of Chemical Information and Modeling, highlighting the relevance of the proposed approach within the international structural bioinformatics community. The

results obtained reinforce the potential of this strategy as a molecular screening tool, offering low computational cost and practical applicability. The evaluation conducted in this work supports the improvement and careful selection of the models comprising the SL, aiming at its application in molecular interaction screening and analysis environments.

Keywords: Machine Learning; Super Learner; Protein-Protein Interface Interactions; Gibbs Free Energy.

LISTA DE ABREVIATURAS E SIGLAS

PPIs – Interações proteína-proteína

r – Coeficiente de correlação de Pearson

RMSE – Raiz do erro quadrático médio

R^2 – Coeficiente de determinação

ΔG – Energia livre de Gibbs

ML – *Machine Learning* (aprendizado de máquina)

SL – *Super Learner*

MLP – *Multilayer Perceptron*

AdaBoost – Método de regressão *AdaBoost Regressor*

Bagging – Método de regressão *Bagging Regressor*

Decision Tree – Método de regressão *Decision Tree Regressor*

ElasticNet – Método de regressão linear com regularização L1 e L2

ET – Método de regressão *Extra Trees Regressor*

KNN – Método de regressão *K-Nearest Neighbors Regressor*

RF – Método de regressão *Random Forest Regressor*

SVR – Método de regressão *Support Vector Regressor*

SVM – *Support Vector Machines* (máquinas de vetores de suporte)

XGBoost – Método de regressão *Extreme Gradient Boosting*

ANN – *Artificial Neural Networks* (redes neurais artificiais)

Iforest – Método de detecção de outliers *Isolation Forest*

IQR – Intervalo interquartil

PCA – *Principal Component Analysis* (análise de componentes principais)

UMAP – *Uniform Manifold Approximation and Projection*

SHAP – *Shapley Additive Explanations*

PDBbind – Banco de dados experimental de interações proteína-ligante

SKEMPI – Banco de dados experimental com mutações pontuais em interações proteína-proteína

SL MLP – *Super Learner* acoplado ao metamodelo MLP

SL_MLP_UMAP_4 – *Super Learner* com metamodelo MLP utilizando UMAP com quatro variáveis

SL_MLP_PCA – *Super Learner* com metamodelo MLP utilizando PCA

SL_MLP_SHAP_42 – *Super Learner* com metamodelo MLP utilizando as 42 variáveis mais importantes segundo o SHAP

SUMÁRIO

1 INTRODUÇÃO

As interações proteína-proteína (PPIs) estão no cerne dos principais processos celulares, como replicação do DNA, transcrição gênica, transdução de sinais, organização estrutural e respostas imunes (Akbarzadeh S. et al., 2024). Estima-se que existam entre 130.000 e 650.000 tipos distintos de PPIs no organismo humano (Shin et al., 2020), sendo que mais de 80% das proteínas estão envolvidas em processos celulares interdependentes, interagindo repetidamente entre si (Rao et al., 2014). Essas interações formam redes altamente complexas, conhecidas como interactoma, cuja função vai além da simples conexão entre proteínas: elas coordenam vias metabólicas, ajustam a cinética enzimática, regulam concentrações moleculares e promovem respostas rápidas a estímulos externos (De Las Rivas & Fontanillo, 2012).

A compreensão dessas interações é essencial não apenas para elucidar funções protéicas e mecanismos patológicos, mas também para o desenvolvimento de novas abordagens terapêuticas. Embora historicamente consideradas desafiadoras como alvos farmacológicos, avanços recentes indicam que a modulação de PPIs constitui uma estratégia promissora no tratamento de doenças complexas e refratárias (Lu, H. et al., 2020; Sedov & Zuev, 2023). As PPIs dão origem aos complexos proteína-proteína, estruturas mais estáveis formadas pela associação de duas ou mais proteínas que atuam como uma unidade funcional, sendo estes os dados de entrada de alta resolução utilizados no presente estudo. Esses complexos podem ser temporários ou permanentes, a depender da natureza da interação.

Com o crescimento exponencial de dados estruturais depositados em bases públicas, como o Protein Data Bank (PDB), e a crescente dificuldade de caracterizar experimentalmente todas as interações possíveis, abordagens computacionais tornaram-se indispensáveis na análise do interactoma. O estudo computacional dessas redes é desafiador, especialmente por três fatores: uma proteína pode desempenhar múltiplas funções, interações ocasionais podem ocorrer entre proteínas com funções distintas, e muitas interações não são estáveis ao longo do tempo (RAO et al., 2014).

Entre os parâmetros físico-químicos relevantes para caracterização das interações está a variação da energia livre de Gibbs (ΔG), frequentemente usada como indicador da afinidade entre biomoléculas (Geng et al., 2019). Para uma reação de ligação simples:

$$A + B \rightleftharpoons AB$$

A constante de dissociação K_d é dada por:

$$K_d = \frac{[A][B]}{[AB]}$$

Assim, ΔG pode ser obtido por:

$$\Delta G = RTLnK_d \qquad \text{eq[1]}$$

Onde R é a constante dos gases, T é a temperatura absoluta, e K_d é a constante de dissociação do meio. Valores negativos de ΔG indicam que haverá uma maior quantidade de formação de produto do que reagente. Variações nessa energia têm sido associadas a implicações fisiopatológicas relevantes, como alterações metabólicas em células tumorais (Golas et al., 2019) e distúrbios neurobiológicos (Keegan et al., 2021).

A obtenção experimental de ΔG pode ser realizada por técnicas como Calorimetria de Titulação Isotérmica (ITC) e Calorimetria de Varredura Diferencial (DSC), que exigem condições experimentais rigorosas, alto custo e baixa escalabilidade (GUO; YAMAGUCHI, 2022). Já técnicas como difração de raios X, Cryo-EM e ensaios do tipo ELISA são amplamente utilizadas para a caracterização estrutural de proteínas e seus complexos, fornecendo modelos tridimensionais que podem ser associados a dados experimentais de afinidade, como K_d, obtidos por outras abordagens. A integração dessas estruturas com métodos de modelagem molecular ou aprendizado de máquina permite inferências computacionais sobre ΔG, mesmo quando os valores de afinidade são fornecidos como metadados em bancos de dados públicos, como o PDB.

Algoritmos de aprendizado de máquina têm sido amplamente empregados para prever a afinidade entre biomoléculas com base em descritores derivados de sequências ou estruturas tridimensionais (CHEN et al., 2023). A composição e a

ordem dos aminoácidos em uma sequência proteica contém informações essenciais para seu dobramento e, consequentemente, para a formação de interfaces específicas e funcionais com outras proteínas (SOLEYMANI et al., 2022). Estudos também apontam que mutações coordenadas ao longo da evolução promovem complementaridade eletrostática e estrutural nas interfaces (SUN et al., 2017). Tais propriedades vêm sendo aproveitadas por algoritmos de aprendizado profundo, capazes de extrair automaticamente características relevantes de grandes volumes de dados não estruturados.

Contudo, ainda há desafios relacionados à generalização e à validação externa desses modelos, uma vez que muitos estudos baseiam-se exclusivamente em validação cruzada, sem testagem independente em conjuntos externos (DURHAM et al., 2023). Apesar disso, tais modelos já demonstram utilidade na priorização de alvos terapêuticos e no desenho racional de fármacos.

Com base nessas premissas, o presente trabalho auxiliou no desenvolvimento de um modelo computacional baseado em *Super Learner* (SL), conforme proposto por Van der Laan et al. (2007), avaliando-se 10 diferentes modelos base utilizados nesse tipo de arquitetura de forma otimizada. Cujo objetivo foi prever a variação da energia livre de Gibbs (ΔG) nas interfaces de complexos proteína-proteína, a partir de descritores estruturais extraídos com o software Rosetta. O modelo foi treinado com dados obtidos de bancos estruturais amplamente validados, como PDBbind (WANG et al., 2004), PRODIGY (XUE et al., 2016), SKEMPI2 (JANKAUSKAITĖ et al., 2019) e o Protein-Protein Docking Benchmark v5.5 (VREVEN et al., 2015).

Os resultados obtidos indicaram que o modelo SL alcançou um coeficiente de correlação de Pearson r = 0,70, e um erro quadrático médio (RMSE) de 1,91, superando abordagens anteriores com a mesma finalidade e de baixo custo computacional. Esses achados reforçam o potencial da integração entre dados estruturais e algoritmos de aprendizado de máquina na triagem de interações moleculares, contribuindo para a redução de falsos positivos e a otimização de recursos no desenvolvimento de fármacos.

2 FUNDAMENTAÇÃO TEÓRICA

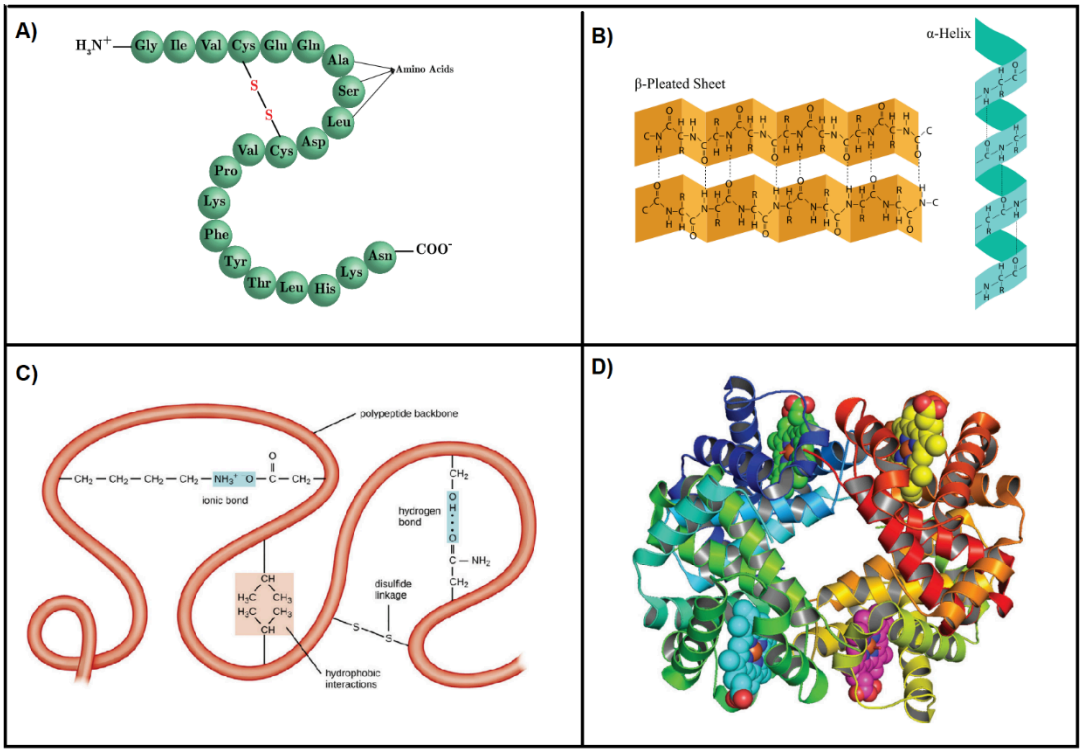
2.1 PROTEÍNAS – ESTRUTURA E FUNÇÃO

As proteínas são biomoléculas que correspondem a mais de 50% do peso seco das células (GARRETT; GRISHAM, 2016) e desempenham funções vitais nos sistemas vivos, como transporte, defesa, armazenamento, atividade enzimática e participação na composição das membranas celulares (CLARK; CHOI; DOUGLAS, 2018). Sua estrutura é determinada por uma sequência específica de aminoácidos codificada geneticamente, geralmente por um único gene. No entanto, um mesmo gene pode originar múltiplas variantes estruturais (CLARK; CHOI; DOUGLAS, 2018), capazes de interagir com DNA, RNA, íons, hormônios e outras proteínas para desempenhar funções específicas (ZHANG et al., 2020a).

Essas interações biomoleculares são complexas e podem ocorrer entre dois ou mais parceiros protéicos, envolvendo também resíduos presentes no meio. Compreender a natureza dessas interações é essencial para o desenvolvimento de fármacos e terapias.

A estrutura das proteínas pode ser descrita em quatro níveis organizacionais: primário, secundário, terciário e quaternário, que refletem o grau de complexidade na organização espacial da molécula, como pode ser visualizado.

Figura 1 – Ilustração das diferentes estruturas proteicas: (A) estrutura primária, mostrando a sequência linear de aminoácidos; (B) estrutura secundária, com alfa-hélices e folhas beta; (C) estrutura terciária, representando o dobramento tridimensional da proteína; (D) estrutura quaternária, demonstrando a associação de múltiplas subunidades na molécula de hemoglobina.



Fonte: Adaptado de (CLARK; CHOI; DOUGLAS, 2018; SOLEYMANI et al., 2022)

A estrutura primária é uma sequência linear de aminoácidos ligados por ligações peptídicas. Cada aminoácido possui um grupo carboxila e um grupo amino que, a pH fisiológico, estão carregados, influenciando a formação de interações eletrostáticas e de hidrogênio, essenciais para a estabilidade estrutural e função das proteínas (BUXBAUM, 2015). Alterações na sequência de aminoácidos podem afetar diretamente a conformação e, consequentemente, a função da proteína.

Na estrutura secundária que inclui padrões de dobramento localizados como hélices α e folhas β, estabilizados por ligações de hidrogênio entre os átomos da espinha dorsal da proteína (GARRETT; GRISHAM, 2016). As hélices α são formadas por ligações de hidrogênio intramoleculares, enquanto as folhas β resultam em interações entre segmentos adjacentes da cadeia polipeptídica.

Já a estrutura terciária refere-se ao dobramento tridimensional completo de uma única cadeia polipeptídica, resultado de interações entre as cadeias laterais de aminoácidos, como interações hidrofóbicas, ligações de hidrogênio e interações eletrostáticas (GARRETT; GRISHAM, 2016). A estrutura terciária é determinante para a função biológica das proteínas, permitindo a formação de sítios ativos e regiões de ligação a ligantes.

A estrutura quaternária consiste na organização de múltiplas subunidades polipeptídicas em um complexo funcional. A interação entre essas subunidades é essencial para a atividade de proteínas oligoméricas (GARRETT; GRISHAM, 2016).

As proteínas podem ser classificadas quanto à forma e solubilidade em três categorias principais: fibrosas, globulares e de membrana. As proteínas fibrosas são insolúveis em água e em soluções salinas diluídas. As globulares apresentam formato quase esférico e são solúveis em meios aquosos. Já as proteínas de membrana possuem cadeias laterais hidrofóbicas voltadas para o exterior, que interagem com a fase não polar das membranas, tornando-as insolúveis em água, mas solubilizáveis em detergentes. No entanto, o uso de detergentes pode comprometer a integridade da proteína, uma vez que pode induzir a desnaturação.

Além das particularidades estruturais e físico-químicas, a atividade das proteínas depende fortemente de suas interações com outras moléculas. Essas interações intermoleculares desempenham papel central nos processos biológicos e são essenciais para a manutenção da homeostase celular, como será abordado a seguir.

2.2 INTERAÇÕES INTERMOLECULARES, ASPECTOS GERAIS EM SISTEMAS BIOLÓGICOS E A HOMEOSTASIA CELULAR.

O reconhecimento molecular refere-se ao processo pelo qual macromoléculas biológicas, como proteínas, interagem com outras moléculas por meio de forças não covalentes, formando complexos específicos e funcionais. Esse reconhecimento é regido principalmente por dois fatores: a especificidade, que assegura a seletividade do parceiro de ligação, e a afinidade, que determina a força da interação, mesmo em baixas concentrações do ligante (DEMCHENKO, 2001; JANIN, 1995).

A caracterização e quantificação dessas interações, especialmente em termos energéticos, são essenciais para compreender os fatores que influenciam a formação e a estabilidade dos complexos moleculares — como a intensidade das forças envolvidas (atração ou repulsão entre cargas), o grau de organização da água ao redor das moléculas e a flexibilidade estrutural das proteínas. Esses aspectos explicam por que determinadas interações ocorrem de forma mais eficiente e são

fundamentais para prever comportamentos moleculares e viabilizar aplicações como a engenharia de proteínas e o desenho racional de fármacos (PEROZZO; FOLKERS; SCAPOZZA, 2004).

As interações entre proteínas ocorrem predominantemente por meio de ligações não covalentes, como ligações de hidrogênio e interações eletrostáticas, mas também podem envolver ligações covalentes, como as ligações dissulfeto entre resíduos de cisteína (DE LAS RIVAS; FONTANILLO, 2010; SOLEYMANI et al., 2022). Estas últimas podem conferir maior estabilidade estrutural e atuar na regulação de processos de sinalização celular (BAGCHI, 2018; GARRETT; GRISHAM, 2016).

Tais interações são fundamentais para a manutenção da homeostase celular, pois regulam processos como o enovelamento de proteínas, o tráfego intracelular, a resposta a estímulos externos e a sinalização molecular (ZHANG et al., 2020b). Sua estabilidade e dinâmica são moduladas por fatores intrínsecos, como carga e polaridade das cadeias laterais dos aminoácidos, e extrínsecos, como o pH e a presença de solventes. Por exemplo, resíduos com carga positiva, como lisina e arginina, podem interagir com aminoácidos carregados negativamente, como aspartato e glutamato, promovendo estabilidade estrutural e funcional.

A compreensão dessas interações é fundamental não apenas para a elucidação de fenômenos moleculares básicos, mas também para aplicações práticas, como o desenvolvimento de novas terapias, a investigação de mecanismos patológicos e a modelagem computacional de sistemas biológicos complexos. Diante disso, torna-se necessário aprofundar os aspectos físico-químicos que regem o reconhecimento molecular, os quais serão discutidos na próxima seção.

2.3 ASPECTOS FÍSICO-QUÍMICOS DO RECONHECIMENTO MOLECULAR

Os conceitos debatidos nesta seção são amplamente estabelecidos na literatura e foram baseados nas referências de DU et al. (2016) e GARRETT; GRISHAM (2016).

A cinética de ligação proteína–ligante descreve o processo pelo qual uma proteína (P) e um ligante (L) se associam para formar o complexo PL. Esse processo é regido pelas constantes de taxa de associação (k_{on} e dissociação (k_{off}), de modo que, em equilíbrio, se tem:

$$k_{on}[P][L] = k_{off}[PL] \qquad \text{Eq. 2}$$

Define-se a constante de equilíbrio de ligação K_b como:

$$K_b = \frac{k_{on}}{k_{off}} = \frac{[PL]}{[P][L]} = \frac{1}{K_d} \qquad \text{Eq.3}$$

Onde K_d é a constante de dissociação; um valor elevado de K_b indica alta afinidade de ligação. O sistema proteína-ligante-solvente constitui um conjunto termodinâmico complexo, em que as trocas de calor e as interações não covalentes (ligação de hidrogênio, forças de Van der Waals etc.) definem a energia livre de ligação. A Energia Livre de Ligação Padrão (ΔGº) relaciona-se a K_b por meio da equação:

$$\Delta G^{\circ} = -RTlnK_b \qquad \text{Eq. 4}$$

Importa ressaltar que esta expressão termodinâmica é rigorosamente válida quando a cinética de associação se dá como reação de primeira ordem e a de dissociação como reação de segunda ordem, hipótese geralmente admitida para sistemas protótipos em equilíbrio. Em qualquer estado não-equilíbrio, o valor instantâneo de ΔG é dado por:

$$\Delta G = \Delta G^{\circ} + RTlnQ \qquad \text{Eq. 5}$$

Onde Q é o quociente de reação. No equilíbrio, Q = K_b e ΔG = 0. Além disso, ΔG pode ser decomposta em suas contribuições entálpicas (ΔH) e entrópicas (ΔS):

$$\Delta G = \Delta H - T\Delta S \qquad \text{Eq.6}$$

Aqui, ΔH reflete a energia interna do sistema sendo o balanço energético de ligações e interações não covalentes (podendo ser exotérmico ou endotérmico) e ΔS mede a variação de desordem, incluindo mudanças na solvação (ΔS_{solv}), ajustes

conformacionais (ΔS_{conf}) e perda de graus de liberdade translacionais e rotacionais ($\Delta S_{t/r}$). Frequentemente observa-se compensação entrópia–entalpia, em que ganhos em uma grandeza são parcial ou totalmente balanceados por perdas na outra, resultando em ΔG similar.

Um fenômeno relevante é a compensação entalpia-entropia, no qual mudanças favoráveis em um desses parâmetros são compensadas por variações desfavoráveis no outro, resultando em valores semelhantes de energia livre. Essa compensação é amplamente discutida na literatura, embora não seja abordada com profundidade neste trabalho.

Cabe destacar que a estimativa experimental de constantes de equilíbrio e da energia livre de ligação está sujeita a erros sistemáticos, principalmente devido à suposição de soluções ideais e à simplificação das interações moleculares envolvidas. Em sistemas reais, efeitos como força iônica do meio, pH, presença de cofatores ou competidores, e limitações instrumentais podem introduzir desvios significativos. Além disso, as técnicas experimentais utilizadas (como ITC, SPR ou métodos espectroscópicos) apresentam incertezas intrínsecas, tanto na medição de concentrações quanto na determinação de parâmetros cinéticos. Tais fontes de erro impactam diretamente na acurácia dos valores de Kb, ΔG, ΔH e ΔS, tornando essencial a interpretação cuidadosa dos resultados.

Por fim, ao compreender os fundamentos físico-químicos que regem a formação de complexos biomoleculares, torna-se essencial analisá-los a partir de dados quantitativos. A próxima seção descreve a origem e o tratamento do conjunto de dados (*dataset*) utilizado na construção e avaliação do modelo computacional proposto.

2.4 ORIGEM DO DATASET

Os dados utilizados neste trabalho foram extraídos de quatro bancos de dados distintos, com destaque para o PDBbind, que contém a maior quantidade de complexos proteína-proteína por integrar estruturas provenientes de outras bases. Para evitar redundâncias, complexos duplicados foram cuidadosamente removidos.

O PDBbind (WANG et al., 2004) é um dos principais bancos utilizados, contendo complexos proteína-ligante com dados de afinidade experimental precisos, como constantes de dissociação (K_d) e inibição (K_i), excluindo casos com valores extremos ou estimados. Apenas estruturas cristalinas com resolução inferior a 2,5 Å são incluídas, e complexos covalentes são removidos para manter a integridade estrutural dos dados.

O banco PRODIGY (PROtein binDIng enerGY prediction) (XUE et al., 2016), por sua vez, é derivado de um conjunto específico de complexos proteína-proteína, com 81 estruturas de alta confiabilidade, todas com afinidades de ligação (ΔG) experimentais e sem fragmentos ou lacunas no sítio de ligação. Este banco foi refinado para garantir que apenas técnicas experimentais confiáveis fossem consideradas, sendo que a maioria dos seus dados são provenientes do banco PDBbind.

Já o SKEMPI (JANKAUSKAITĖ et al., 2019) é um banco de dados focado em interações proteína-proteína que documenta diversas mutações e suas respectivas afinidades de ligação, utilizando métodos como espectroscopia, ELISA e calorimetria de titulação isotérmica. Este banco organiza as mutações entre diferentes regiões estruturais, como o interior e a superfície da proteína, permitindo um estudo detalhado das alterações de afinidade causadas por mutações. A versão SKEMPI2 é um aprimoramento da versão inicial, com a inclusão de novas estruturas.

Por fim, o Benchmark 5.5 (VREVEN et al., 2015), da ZLab, inclui complexos classificados em níveis de dificuldade para modelos de *docking* baseados em proteínas. Este banco abrange estruturas rígidas e flexíveis de complexos, categorizados de acordo com a alteração estrutural entre as formas ligada e não ligada das proteínas, sendo amplamente utilizado para validar algoritmos de *docking*.

Esses bancos forneceram os dados experimentais utilizados no treinamento e teste dos modelos de aprendizado de máquina. A seguir, serão descritas as métricas adotadas para avaliar o desempenho preditivo desses modelos. Vale ressaltar que uma limitação relevante deste trabalho refere-se ao tamanho do conjunto de dados, sobretudo no caso do banco SKEMPI, que apresenta um número reduzido de

complexos experimentais em comparação a outras bases. Além disso, não foi realizada distinção entre complexos proteína–proteína originados de mutações e os nativos, tampouco foi efetuada a categorização dos pares como interações antígeno–anticorpo ou de outra natureza. Tais diferenciações estruturais e funcionais podem impactar significativamente os resultados preditivos, e representam oportunidades de investigação para estudos futuros.

2.5 MÉTRICAS DE AVALIAÇÃO

Neste estudo, todas as métricas foram calculadas após a aplicação da técnica de validação cruzada k-fold com k = 10, procedimento comumente utilizado para equilibrar viés e variância na avaliação de modelos preditivos (KOHAVI, 2001). A validação foi realizada exclusivamente sobre os dados de treinamento. As métricas de avaliação adotadas para mensurar o desempenho do modelo foram o erro quadrático médio (RMSE), o coeficiente de determinação (R^2) e o coeficiente de correlação de Pearson (r). Essas métricas foram escolhidas por permitirem uma análise complementar da acurácia, do ajuste e da associação linear entre os valores preditos e os valores observados.

2.5.1 RMSE (RAIZ DO ERRO QUADRÁTICO MÉDIO)

O RMSE avalia a precisão entre os valores previstos e os valores reais. Ele é adequado para representar o desempenho do modelo quando se espera que a distribuição do erro seja uma Gaussiana. O RMSE é sensível a *outliers* (valores atípicos que fogem drasticamente do padrão dos dados) e penaliza a variância ao dar mais peso a erros com valores absolutos maiores do que a erros com valores absolutos menores (CHAI; DRAXLER, 2014).

O RMSE pode ser expresso como:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \qquad \text{Eq.7}$$

Onde “n” representa o número de amostras, “Y_i” os valores reais, e “$\hat{Y}_i$”os valores preditos. Valores menores de RMSE indicam previsões mais precisas. Essa métrica não assume linearidade entre os dados

2.5.2 R² (COEFICIENTE DE DETERMINAÇÃO)

O R² é comumente utilizado para avaliar a qualidade de modelos de regressão, incluindo aqueles baseados em técnicas de ML. Ele mede a proporção da variação na variável dependente explicada pelas variáveis independentes, mas apresenta várias limitações importantes. Conforme argumentado por Gary (KING, 1986), R², é frequentemente mal interpretado como uma medida da influência de “X” sobre “Y”, quando, na verdade, indica a dispersão dos pontos ao redor da linha de regressão. Portanto, é importante avaliar o gráfico dos resíduos juntamente com o R². Além disso, o R² não deve ser utilizado para inferir relações causais, detectar problemas de multicolinearidade ou avaliar variáveis omitidas. O R² é sensível à variação da amostra, conforme destacado por Achen (ACHEN, 1977) e Kennedy (KENNEDY, 2008), o que significa que valores altos ou baixos podem ser influenciados pelas variâncias das variáveis independentes, tornando-o inadequado para comparações entre diferentes amostras ou modelos. Embora útil como uma medida inicial de ajuste, o R² não deve ser usado isoladamente como um indicador da qualidade de um modelo de regressão em técnicas de ML. Em vez disso, é essencial complementar o R² com outras métricas e focar nos coeficientes não padronizados e seus erros estimados para uma avaliação mais robusta do modelo.

O R² é sensível a *outliers*, com um único *outlier* capaz de impactar significativamente a métrica. Um alto R² não implica causalidade; mesmo que um modelo explique bem os dados, isso não significa que uma variável causa a outra. Portanto, o R² deve ser interpretado com cautela e utilizado em conjunto com outros métodos de avaliação, essa métrica pode ser expressa como:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - Y)^{\cdots 2}} \qquad \text{Eq. 8}$$

Onde "Y_i" representa os valores observados, "$\hat{Y}_i$" os valores preditos, "$\bar{Y}$" a média dos valores observados, "$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$" a soma dos resíduos ao quadrado, e "$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$" o total da soma dos quadrados. Em alguns casos, é possível a obtenção de R² negativo, isso ocorre quando o erro do modelo (resíduo) é maior do que o erro de uma predição ingênua, como prever a média (JAMES et al., 2021).

2.5.3 CORRELAÇÃO DE PEARSON (r)

A correlação é uma medida da direção e da força da relação linear entre duas variáveis quantitativas, representada pelo coeficiente r (MOORE, 2014). A correlação de Pearson é uma medida da associação linear entre variáveis quantitativas, detectando apenas relações lineares. O coeficiente de correlação de Pearson é calculado usando a fórmula:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \qquad \text{Eq. 9}$$

Onde "x_i" e "y_i" são os pontos das amostras individuais, e "$\bar{x}$" e "$\bar{y}$" são as médias de "x" e "y" respectivamente.

Ao contrário da regressão, o coeficiente de correlação não diferencia entre variáveis independentes e dependentes. Assim, a correlação entre "x" e "y" é a mesma que entre y e x. Além disso, uma correlação de 0,8 não é duas vezes mais forte que uma correlação de 0,4, uma vez que o coeficiente de correlação de Pearson não representa uma proporção da força da relação (PETER Y, 2002).

A correlação de Pearson é interpretada por valores que variam de -1 a 1, indicando a força e a direção da relação linear entre as variáveis. Ela assume que a relação entre as variáveis é linear (PARANHOS et al., 2014).

Essas métricas foram fundamentais para avaliar a performance preditiva dos modelos propostos a partir dos descritores moleculares gerados. A seguir, apresenta-se a modelagem molecular empregada para extrair tais descritores, com ênfase no uso do software Rosetta

2.6 MODELAGEM MOLECULAR DE PROTEÍNAS COM O ROSETTA

O Rosetta é um conjunto robusto de algoritmos para modelagem e design macromolecular, amplamente utilizado em pesquisas acadêmicas e aplicações industriais. Desde sua criação, o pacote evoluiu significativamente, indo além da predição de estruturas de proteínas para abarcar funcionalidades como modelagem de complexos proteína-proteína, *docking* molecular, engenharia de enzimas e design racional de proteínas (Leman et al., 2020). Sua arquitetura modular possibilita a integração flexível de métodos de amostragem e funções de pontuação, o que favorece o avanço de aplicações em biologia estrutural computacional.

Entre os recursos disponíveis, destacam-se ferramentas para refinamento de estruturas experimentais, avaliação de efeitos de mutações e modelagem de interações intermoleculares. Para facilitar o uso dessas ferramentas, o Rosetta disponibiliza interfaces como o “*RosettaScripts”*, um framework baseado em XML que permite a personalização de protocolos, e o PyRosetta, que oferece integração com outras plataformas computacionais.

2.6.1 ESTRUTURA DO ROSETTASCRIPTS

O “*RosettaScripts”* organiza a execução dos protocolos computacionais em duas fases principais:

- Fase de declaração: define os componentes que serão utilizados, como “*Movers”*, “*Filters”*, “*ScoreFunctions”*, “*TaskOperations”* e “*SimpleMetrics”*.
- Fase de ordenação: organiza esses componentes em sequência para executar o fluxo de simulação. Esse fluxo é gerenciado pelo objeto

"*JobDistributor"*, que controla a leitura da estrutura de entrada e reinicia o protocolo se algum filtro for reprovado (FLEISHMAN et al., 2011).

A seguir, detalham-se os principais elementos utilizados neste trabalho.

2.6.2 MOVERS: MODIFICAÇÕES NA ESTRUTURA

Os "*Movers"* são módulos responsáveis por aplicar alterações diretas na estrutura molecular, como minimização de energia, acoplamento (*docking*), redesign de sequências e remodelagem estrutural.

Neste trabalho, utilizou-se o "*InterfaceAnalyzerMover"*, ferramenta que analisa interfaces de interação entre cadeias polipeptídicas. Ele calcula propriedades estruturais e energéticas como:

- Energia de ligação (ΔG): estimativa da energia envolvida na formação do complexo.
- Área de superfície enterrada ("*Buried Surface Area"*, BSA): mede a área que deixa de estar exposta ao solvente após a formação do complexo.
- Empacotamento da interface ("*Packing statistics"*): avalia o preenchimento espacial das cadeias.
- Complementaridade de forma ("*Shape Complementarity"*): verifica o encaixe tridimensional entre as superfícies moleculares. Uma boa complementaridade indica que as superfícies interagem de maneira eficiente, como peças de um quebra-cabeça.

Essas métricas são fundamentais para avaliar a estabilidade e a qualidade da interação proteína-proteína.

2.6.3 FILTERS: CRITÉRIOS DE SELEÇÃO DE CONFORMAÇÕES

Os “*Filters*” têm como função interromper ou reiniciar a execução do protocolo caso a estrutura analisada não satisfaça determinados critérios. Eles garantem que apenas conformações com propriedades desejáveis sejam consideradas nos resultados finais.

Dois filtros foram utilizados:

- *“InterfaceHoles”*: detecta buracos (vazios) na interface de ligação por meio do algoritmo “*PackStat*”, que é baseado no método geométrico “*DAlphaBall*”. Esse algoritmo verifica o quão bem preenchida está a interface entre as proteínas, identificando cavidades internas que indicam má interação e potencial instabilidade do complexo. Interfaces com muitos "*holes*" são indesejáveis, pois sugerem baixa qualidade estrutural.
- *“ContactMolecularSurface”*: calcula a área efetiva de contato entre as superfícies das duas cadeias. A interação entre proteínas depende da proximidade e complementaridade das superfícies envolvidas, e este filtro fornece uma estimativa quantitativa dessa interação. A métrica pondera os pontos de contato pela distância entre os átomos.

2.6.4 SIMPLEMETRICS: COLETA DE MÉTRICAS INFORMATIVAS

O *framework “SimpleMetrics”* permite coletar dados quantitativos ao longo do protocolo sem interferir na execução. Diferente dos “*Filters”*, os “*SimpleMetrics*” não interrompem o processo, eles apenas registram propriedades estruturais e energéticas úteis para análises posteriores. As métricas podem ser exportadas em arquivos “JSON” e analisadas com ferramentas estatísticas como “pandas”, “numpy” e “seaborn” (ADOLF-BRYFOGLE et al., 2021).

No presente trabalho, foi utilizada a “*InteractionEnergyMetric*”, que mede a energia de interação entre dois grupos de resíduos (por exemplo, duas cadeias proteicas) com base na função de energia do Rosetta. Essa métrica permite quantificar a contribuição energética da interface na formação do complexo.

2.6.5 SCOREFUNCTIONS: AVALIAÇÃO ENERGÉTICA

As “*ScoreFunctions*” são centrais para a avaliação das estruturas no Rosetta. Elas combinam diferentes termos energéticos, como forças de van der Waals, interações eletrostáticas, ligações de hidrogênio e potenciais estatísticos, para gerar um valor escalar que representa a qualidade de uma conformação.

Neste estudo, utilizou-se a “*BETA_NOV16*”, uma versão mais recente e precisa da função “*REF15”*, com melhor desempenho na modelagem de interações proteína-proteína (SHRINGARI et al., 2020). Essa função inclui melhorias como:

- Ajustes nos pesos dos termos energéticos;
- Modelagem mais precisa das interações com o solvente;
- Preferências conformacionais mais realistas.

Além disso, ativou-se o parâmetro “*auto_setup_metals*”, que permite a modelagem automática de íons metálicos presentes na estrutura, melhorando a precisão na análise de complexos com cofatores metálicos.

2.6.6 GERAÇÃO DE DESCRITORES

Com base nos componentes previamente descritos, Movers, Filters e SimpleMetrics, foram extraídos 49 descritores moleculares para cada uma das 526 estruturas avaliadas. Esses descritores refletem características energéticas, geométricas e topológicas da interface entre as cadeias polipeptídicas, incluindo medidas como energia de interação, área de superfície de contato, complementaridade de forma e qualidade do empacotamento da interface.

A seguir, a seção 2.7 apresenta os algoritmos utilizados na construção dos modelos preditivos concorrentes.

2.7 ALGORITMOS DE PREDIÇÃO DE ΔG DOS CONCORRENTES

Foram definidos critérios objetivos para selecionar os concorrentes relevantes ao algoritmo proposto nesta dissertação. Os critérios estabelecidos foram:

- O sistema deve ser um software ou aplicação cujo dado de entrada seja um arquivo no formato .pdb, não apenas um método teórico ou algoritmo publicado;
- Deve ser aplicável a interações do tipo proteína-proteína ou proteína-ligante;
- Os dados de treinamento e teste devem ser publicamente disponíveis, independentemente de haver fins lucrativos;
- Os resultados de predição devem ser produzidos em tempo reduzido, com um limite máximo de 5 minutos por estrutura, considerando execução em um computador com especificações medianas (8 GB de memória RAM e processador Intel(R) Core(TM) i5-7300HQ (quatro núcleos, frequência base de 2,50 GHz, 6 MB de cache L3).

Entre os principais concorrentes avaliados, destaca-se o AREA-AFFINITY, um webserver disponível em https://affinity.cuhk.edu.cn/index.html, que se propõe a prever a afinidade de ligação em complexos proteína-proteína e anticorpo-antígeno. Esse sistema utiliza modelos lineares e não lineares baseados em descritores como área de superfície e características de interface, aplicando 60 modelos específicos para proteína-proteína e 37 para anticorpo-antígeno.

Os dados de entrada exigem a estrutura tridimensional do complexo em formato .pdb, além da especificação das cadeias envolvidas na interação. O sistema apresenta restrições de uso, como:

- Proteínas devem conter mais de 25 resíduos;
- Não há suporte para aminoácidos não padrão;
- No caso de complexos anticorpo-antígeno, o anticorpo precisa conter pelo menos 400 resíduos, e o antígeno deve ter entre 25 resíduos e o total de resíduos do anticorpo.

Os resultados incluem a afinidade prevista em log(K) e a energia de ligação estimada em kcal/mol. O sistema é recomendado tanto para predições pontuais

quanto para ajustes baseados em dados experimentais, permitindo comparações entre os modelos disponíveis. Exemplos reportados mostram correlações elevadas entre os valores preditos e experimentais, com coeficiente de Pearson (r) de até 0,92, valor que será verificado nesta dissertação.

Contudo, é necessário considerar as limitações de generalização desses modelos, que serão discutidas em seções posteriores desta dissertação.

Outro concorrente relevante é o PRODIGY, desenvolvido por Vangone e Bonvin (XUE et al., 2016), com base na observação de que o número de contatos na interface proteica se correlaciona com ΔG (KASTRITIS et al., 2011). Esse sistema é amplamente utilizado para prever a afinidade de ligação entre proteínas, e também oferece dois outros métodos voltados à predição de interações proteína-ligante e à classificação de interações cristalográficas versus biológicas. No entanto, esses dois últimos métodos estão associados a processos de *docking* que demandam até duas semanas para entrega do resultado, não atendendo aos critérios temporais definidos neste trabalho.

A abordagem do PRODIGY se baseia na extração de descritores físico-químicos e estruturais diretamente da interface proteína-proteína, incluindo a composição de resíduos, parâmetros termodinâmicos e a contribuição de contatos interatômicos específicos. É utilizado um cutoff de 5,5 Å para considerar interações relevantes.

Seu modelo de regressão foi treinado com dados experimentais oriundos de bancos como o PDBbind, abrangendo ampla variedade de complexos proteína-proteína com ΔG determinado experimentalmente. O método apresenta precisão considerável, com coeficiente de correlação (r) de até 0,73, demonstrando uma correlação satisfatória entre predições e valores observados.

Apesar de eficiente, o PRODIGY apresenta limitações relacionadas à qualidade estrutural dos modelos de entrada e ao equilíbrio entre descritores físicos e estatísticos, o que pode impactar sua aplicação em cenários que envolvam, por exemplo, mutações específicas.

Cabe ressaltar que a literatura aponta que modelos de predição de afinidade em complexos proteína-proteína geralmente não apresentam o mesmo desempenho

quando aplicados a complexos anticorpo-antígeno (GUEST et al., 2021). Diferentemente de outras abordagens descritas na literatura (XUE et al., 2016; YANG et al., 2023), o algoritmo desenvolvido no contexto desta dissertação apresenta maior flexibilidade, uma vez que não exige a classificação prévia do tipo de interação molecular. Dessa forma, pode ser aplicado diretamente a interações proteína-proteína de diferentes naturezas, incluindo dímeros, multímeros e, potencialmente, a complexos do tipo anticorpo-antígeno.

2.8 ASPECTOS GERAIS DE ML

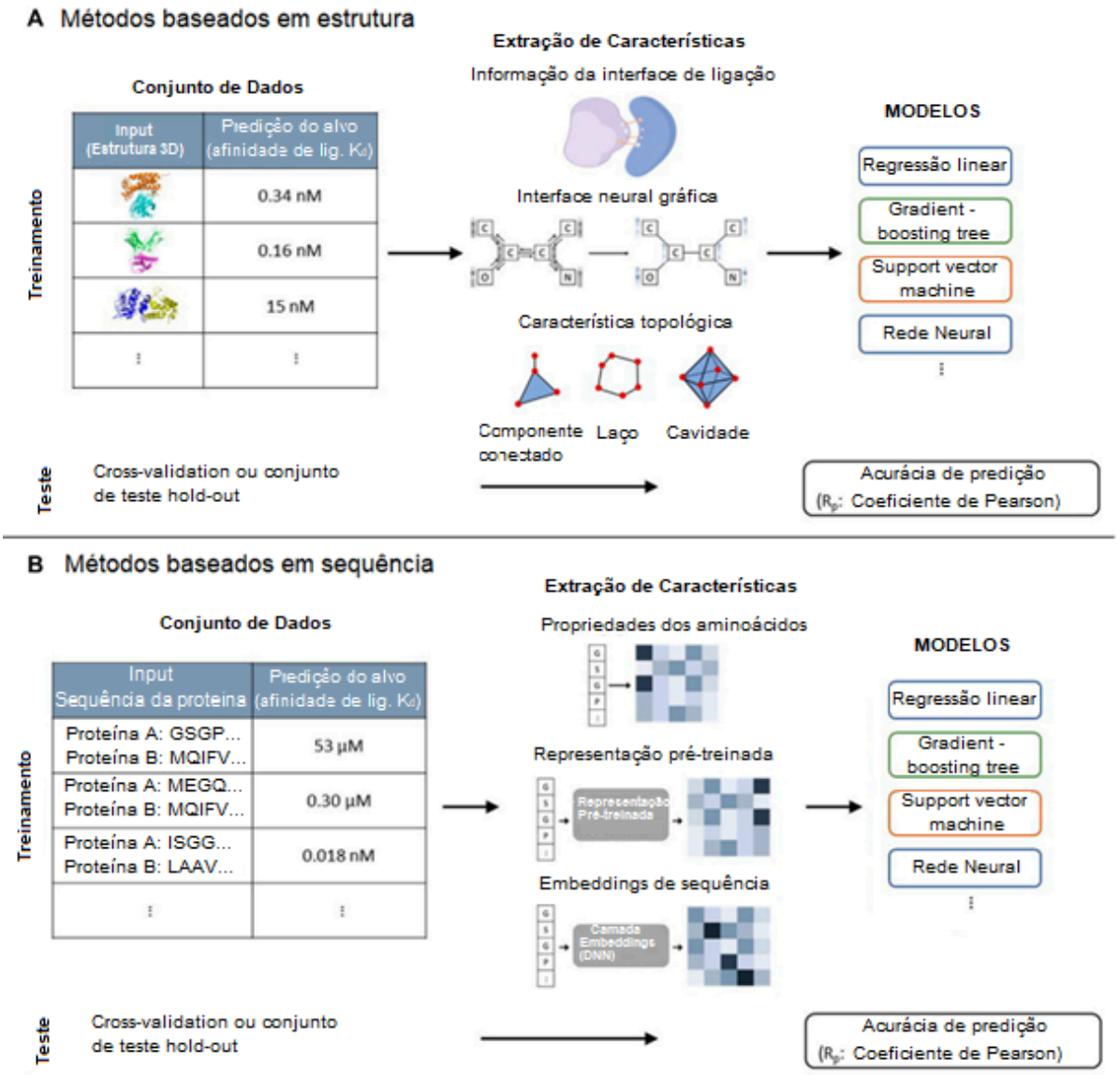
A predição de afinidades de ligação e de ΔG em sistemas biológicos pode ser realizada por meio de diferentes abordagens de machine learning (ML).

Métodos baseados em estrutura utilizam a estrutura tridimensional das proteínas, explorando descritores como a superfície de ligação, redes neurais gráficas que capturam topologia e propriedades geométricas do sítio ativo. Esses atributos são alimentados em algoritmos de ML que, após validação cruzada e teste *hold-out*, estimam a energia livre de ligação.

Métodos baseados em sequência extraem características diretamente das cadeias de aminoácidos: propriedades físico-químicas dos resíduos, representações pré-treinadas e camadas de incorporação treináveis. Da mesma forma, validação cruzada e conjuntos *hold-out* garantem a robustez das previsões.

A eficácia de ambas as abordagens depende da qualidade e representatividade dos dados de treinamento e da seleção criteriosa de descritores. A inclusão de variáveis irrelevantes ou a escassez de exemplos podem prejudicar a generalização dos modelos. Além disso, é fundamental controlar o *overfitting,* quando o modelo aprende ruído dos dados de treinamento e perde desempenho em amostras independentes (GÉRON, 2022).

Figura 2 – Comparação esquemática dos métodos baseados em estrutura (A) e em sequência (B) para predição de afinidades de ligação e ΔG.



Fonte: adaptado (GUO; YAMAGUCHI, 2022)

Dessa forma, compreender essas abordagens gerais permite contextualizar os algoritmos de ML adotados neste trabalho, detalhados na próxima seção (2.9), com foco em suas características, fundamentos e desempenho na predição de ΔG.

2.9 ALGORITMOS DE ML UTILIZADOS

Nesta seção, apresentamos os algoritmos de machine learning (ML) empregados neste estudo, todos implementados com a biblioteca scikit-learn (versão 1.7), utilizando a linguagem Python 3.13. Os experimentos foram executados em um sistema operacional Windows, garantindo reprodutibilidade ao

especificar o ambiente de desenvolvimento utilizado.

Os modelos foram treinados individualmente utilizando busca em grade (“*grid search”*) para a seleção dos principais hiperparâmetros, com exceção do Support Vector Regressor. Durante esse processo, além dos valores padrão, testaram-se duas configurações adicionais (um valor acima e outro abaixo do padrão) com o objetivo de avaliar impactos no desempenho e mitigar o *overfitting*. Para garantir a reprodutibilidade dos resultados, o parâmetro “*random_state* = 42” foi definido.

A validação cruzada com k-fold = 10 foi aplicada para estimar o desempenho dos modelos, com base na minimização do erro e na maximização do coeficiente de determinação (R^2), orientando a escolha dos hiperparâmetros. Após essa etapa, os modelos foram ajustados ao conjunto total de treinamento, e os arquivos .pkl foram gerados para posterior uso na construção do modelo de *Super Learner* (SL).

A avaliação final dos modelos foi realizada com o conjunto de teste, composto por amostras que não participaram do treinamento, assegurando uma comparação consistente e imparcial dos resultados.

2.9.1 ADABOOST REGRESSOR

O AdaBoost Regressor (*Adaptive Boosting Regressor*) (DRUCKER, 1997; FREUND; SCHAPIRE, 1997) é um algoritmo de aprendizado de máquina baseado em *ensemble*, que combina múltiplos modelos base para melhorar a capacidade preditiva em problemas de regressão. Ele ajusta iterativamente o peso dos exemplos do conjunto de treinamento, enfatizando aqueles que apresentam maiores erros em previsões anteriores, por padrão, sejam os pesos dados por $w_1,w_2,...,w_n$, inicialmente todos os pesos terão valor $w_i = \frac{1}{n}$. Assim, os modelos subsequentes concentram-se em corrigir os casos mais difíceis, culminando em uma predição final baseada na média ponderada das previsões individuais.

Nesta pesquisa, priorizamos a generalização do modelo, buscando equilíbrio entre viés e variância. Foram testados ajustes nos principais hiperparâmetros (“*n_estimators*” e “*learning_rate*”), mas os resultados indicaram que a configuração

padrão já oferecia desempenho satisfatório, não justificando modificações adicionais.

O AdaBoost Regressor (AdaBoost) apresenta vantagens como simplicidade, flexibilidade para combinar diferentes modelos base e capacidade de focar em exemplos difíceis, o que pode melhorar a acurácia em dados complexos. Quando bem ajustado, também tende a reduzir o risco de *overfitting*. No entanto, é sensível a *outliers*, que podem distorcer os pesos atribuídos e comprometer o desempenho. Além disso, o processo iterativo pode aumentar a complexidade computacional.

Portanto, o AdaBoost é uma escolha eficiente para problemas que requerem modelos robustos e interpretáveis. Seu uso, contudo, exige atenção especial à presença de ruídos nos dados e à seleção adequada dos hiperparâmetros para evitar sobreajustes.

2.9.2 BAGGING REGRESSOR

O Bagging Regressor (*Bootstrap Aggregating Regressor*) é uma técnica de aprendizado de máquina do tipo ensemble que visa reduzir a variância de modelos base, como a árvore de decisão, utilizada por padrão. Para isso, o algoritmo treina múltiplos modelos de forma independente a partir de subconjuntos aleatórios do conjunto de dados original, utilizando amostragem com reposição (bootstrap). A predição final é obtida por meio da média das previsões individuais (BREIMAN, 1996). Quando não há reposição, a técnica é denominada *Pasting* (BREIMAN, 1999). Já quando a amostragem é feita a partir das características, aplica-se o método *Random Subspaces*. Caso envolva tanto as observações quanto as características, utiliza-se o *Random Patches*.

Durante o ajuste, priorizou-se a generalização do modelo sem aumento desnecessário de complexidade. O número de estimadores (“*n_estimators*”) foi aumentado de 10 para 100, proporcionando diversidade suficiente entre os modelos e reduzindo a variância sem comprometer a eficiência computacional. O hiperparâmetro “*max_features*” foi ajustado para 0.5 (em vez do padrão 1.0), de

forma a limitar o número de características utilizadas em cada árvore, o que diminui a correlação entre os modelos e contribui para evitar o overfitting. O parâmetro "*bootstrap*" foi mantido como "*True*", pois a amostragem com reposição é essencial na estratégia de Bagging.

Os demais hiperparâmetros foram mantidos em seus valores padrão por não apresentarem impacto relevante na relação viés-variância ou já estarem adequados para o contexto de regressão. Por exemplo, "*bootstrap_features*" não foi alterado, pois o foco da variação estava nas amostras, e não nas características. O "*oob_score*" permaneceu desativado, uma vez que a avaliação do modelo foi realizada via validação cruzada, tornando essa métrica redundante.

O Bagging Regressor (Bagging) oferece vantagens importantes, como maior robustez e redução da variância, ao combinar previsões de múltiplas árvores de decisão, o que melhora a generalização e reduz o risco de *overfitting*. Em contrapartida, o modelo apresenta maior custo computacional e menor interpretabilidade, dada a complexidade do *ensemble*. É especialmente eficaz em cenários nos quais o modelo base é propenso ao *overfitting*, mas não é adequado para análises que exigem interpretação direta de coeficientes ou parâmetros únicos.

2.9.3 DECISION TREES REGRESSOR

O Decision Tree Regressor (Decision Tree) é um algoritmo de aprendizado supervisionado utilizado para prever valores contínuos por meio de uma estrutura hierárquica de decisões. O modelo divide recursivamente o conjunto de dados em subconjuntos menores, escolhendo pontos de divisão que minimizam a variância dentro dos grupos resultantes. Esse processo se repete até que um critério de parada seja alcançado, como a profundidade máxima da árvore ou o número mínimo de amostras por nó.

Entre as principais vantagens do Decision Tree estão a facilidade de interpretação e a capacidade de capturar relações não lineares entre as variáveis, sem a necessidade de pré-processamento intensivo. No entanto, o modelo pode sofrer de *overfitting*, especialmente quando gera árvores muito profundas,

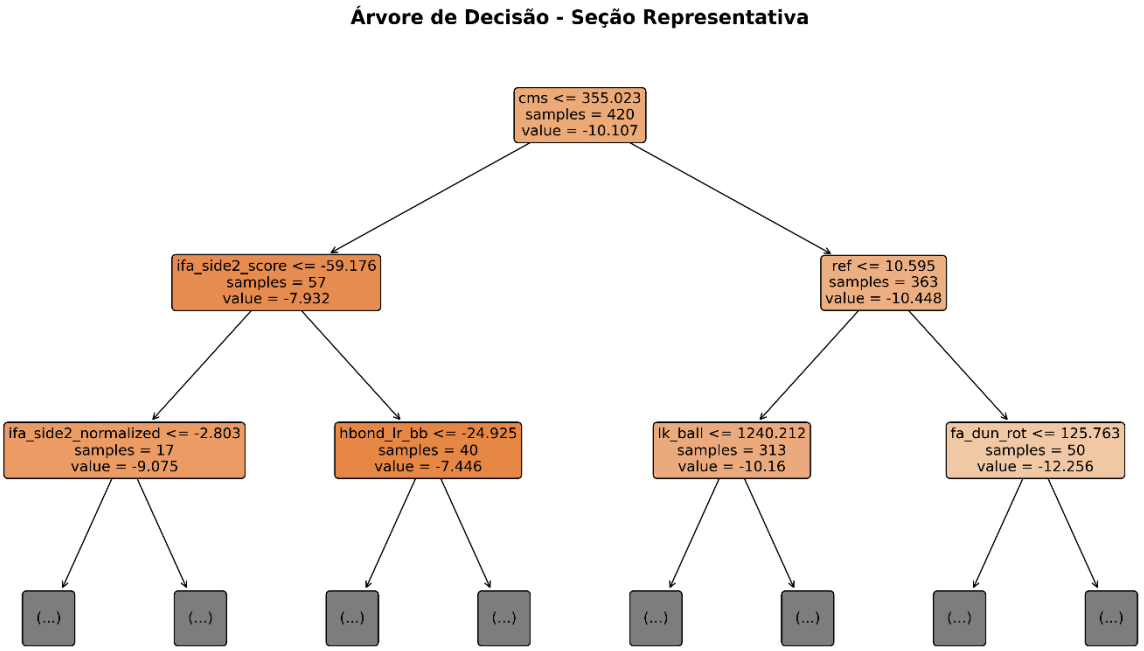
tornando-se altamente sensível a pequenas variações nos dados. Para mitigar esse problema, é comum limitar a profundidade da árvore (“*max_depth”*) e definir um número mínimo de amostras por folha (“*min_samples_leaf*”).

Neste estudo, foi utilizada a métrica R^2 como critério de avaliação durante a otimização dos hiperparâmetros. Os valores obtidos foram: “*max_depth* = 10”, “*max_features* = 'sqrt'”, “*min_samples_leaf* = 8” e “*min_samples_split* = 5”. Esses ajustes foram aplicados com o objetivo de reduzir a complexidade do modelo e aumentar sua capacidade de generalização, minimizando o risco de *overfitting* sem comprometer seu desempenho preditivo.

Alguns hiperparâmetros foram mantidos em seus valores padrão por já se mostrarem adequados ao problema. É o caso de *criterion* (“squared_error”), responsável por minimizar o erro quadrático médio, e *splitter* (“*best*”), que seleciona a melhor divisão em cada nó. Parâmetros como “*min_impurity_decrease”* e “*ccp_alpha”* também não foram ajustados, uma vez que as restrições impostas por “*max_depth”* e “*min_samples_leaf”* já contribuíram para o controle da complexidade do modelo.

Apesar das limitações aplicadas, a árvore gerada ainda apresentou considerável complexidade estrutural, com um número elevado de nós e profundidade significativa. Devido à extensão do modelo completo, optou-se por apresentar, na Figura 3, apenas os três primeiros níveis da árvore de decisão, com o objetivo de ilustrar a lógica inicial das divisões realizadas durante o processo de treinamento.

Figura 3 – Seção inicial da Árvore de Decisão gerada pelo modelo, exibindo os três primeiros níveis de divisão. A figura tem como objetivo ilustrar o funcionamento do algoritmo, apresentando os primeiros critérios de separação aplicados aos dados durante o treinamento



Fonte: O Autor (2024)

2.9.4 ELASTIC NET

O ElasticNet é um método de regressão linear presente na biblioteca Scikit Learn que regularizada que combina as penalizações L1 e L2, sendo amplamente utilizado em problemas de alta dimensionalidade, forte correlação entre variáveis, multicolinearidade ou com mais preditores do que amostras disponíveis. Ele integra os princípios do Lasso (Least Absolute Shrinkage and Selection Operator) e do Ridge, equilibrando a seleção de variáveis e a estabilização dos coeficientes, resultando em um modelo robusto e flexível (JAMES et al., 2021a).

A penalização L1, baseada na norma L1 (soma dos valores absolutos dos coeficientes), promove sparsidade ao forçar alguns coeficientes β a se tornarem exatamente zero, eliminando variáveis menos relevantes. Essa característica é a base do Lasso, ideal para criar modelos interpretáveis e com menor número de variáveis.

A função objetivo do Lasso, que tende a minimização dos pesos ω, pode ser expressa como:

$$min_{\omega}\left(\frac{1}{2n_{amostras}}\|X\omega - y\|_2^2 + \beta\|\omega\|_1\right) \qquad \text{Eq. 10}$$

Onde $\frac{1}{2n_{amostras}}\|X\omega - y\|_2^2$ representa a penalidade do erro quadrático médio que mede a distância entre a previsão e os dados reais, e $\beta\|\omega\|_1$ é o termo de regularização L1 que força os coeficientes a serem exatamente zero, promovendo sparsidade.

Por outro lado, a penalização L2, baseada na norma L2 (soma dos quadrados dos coeficientes), penaliza a magnitude dos coeficientes sem zerá-los completamente. Essa é a essência do Ridge, que estabiliza o modelo em situações de multicolinearidade ao distribuir a penalização entre os coeficientes.

Sua função objetivo é expressa por:

$$min_{\omega}(\|X\omega - y\|_2^2 + \alpha\|\omega\|_2^2) \qquad \text{Eq. 11}$$

O ElasticNet combina essas duas abordagens, permitindo que a penalização L_1 identifique variáveis importantes, enquanto a penalização L_2 estabiliza os coeficientes (ZOU; HASTIE, 2005). A função objetivo do ElasticNet é dada por:

$$min_{\omega}(\frac{1}{2n_{amostras}}\|X\omega - y\|_2^2 + \alpha\rho\|\omega\|_1 + \frac{\alpha(1-\rho)}{2}\|\omega\|_2^2) \qquad \text{Eq. 12}$$

Nessa equação, α controla a intensidade da regularização e ρ (razão entre as penalizações L1 e L2) define o balanço entre as duas penalizações. Quando ρ=1, o ElasticNet se reduz ao Lasso, e quando ρ=0, torna-se equivalente ao Ridge.

O ElasticNet é especialmente útil quando há grupos de variáveis correlacionadas, pois tende a selecionar essas variáveis em conjunto, mantendo informações relevantes. A combinação das penalizações oferece um equilíbrio entre simplicidade e estabilidade, frequentemente resultando em melhor desempenho preditivo do que o Lasso isoladamente, como demonstrado por estudos em dados simulados e reais (ZOU; HASTIE, 2005).

Durante a otimização do modelo, a melhor configuração encontrada para o parâmetro "*max_iter*" foi 100000, garantindo convergência sem interrupções prematuras. Foram testados ajustes para os hiperparâmetros "*alpha*" (penalização geral) e "*l1_ratio*" (proporção entre L1 e L2), mas as alterações não resultaram em melhora de desempenho. Em alguns casos, inclusive, causaram degradação,

indicando que os valores padrão já estavam adequados ao conjunto de dados analisados.

2.9.5 EXTRA TREES REGRESSOR (ET)

O Extra Trees Regressor (*Extremely Randomized Trees Regressor*) é um algoritmo de regressão baseado em florestas de árvores de decisão, que introduz elevada aleatoriedade no processo de construção das árvores para reduzir o viés e melhorar a generalização.

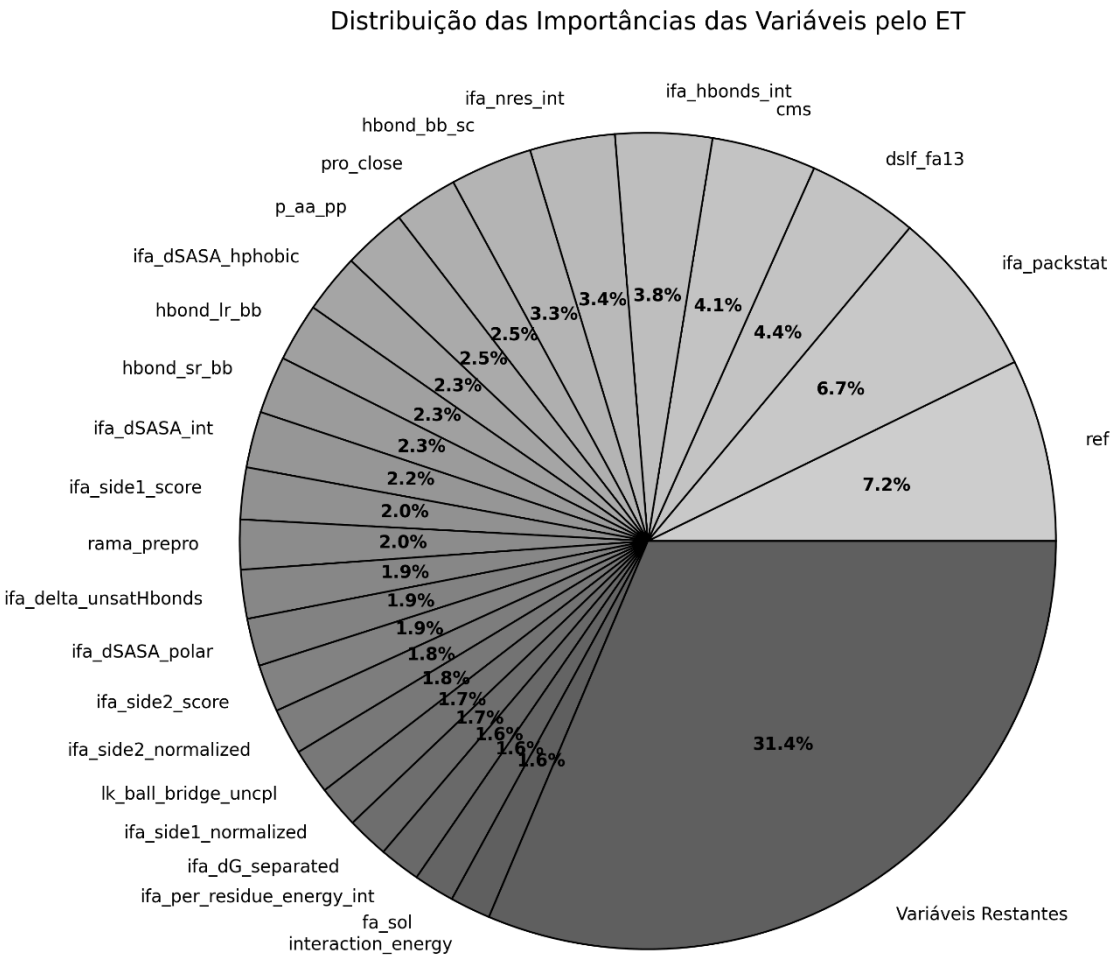
Diferentemente das florestas aleatórias convencionais, no ET os pontos de divisão são selecionados aleatoriamente a partir de um subconjunto de atributos, sem otimização por métricas como ganho de informação ou redução de variância (GEURTS; ERNST; WEHENKEL, 2006). Neste contexto, atributos referem-se às variáveis independentes ou características do conjunto de dados utilizadas na previsão da variável alvo.

O modelo constrói múltiplas árvores de decisão com amostragem aleatória do conjunto de treinamento e combina as previsões por média. Essa abordagem contribui para a redução do risco de *overfitting*, em comparação a modelos isolados, mantendo elevada capacidade preditiva.

Entre as principais vantagens do ET, destacam-se sua eficiência computacional, devido à aleatoriedade na escolha dos pontos de divisão, e menor sensibilidade a ruídos nos dados.

Além disso, o modelo permite a avaliação da importância das variáveis com base na redução da variância explicada, o que contribui para a interpretação do modelo. A Figura 4 apresenta as 25 variáveis mais relevantes segundo o critério de importância atribuído pelo modelo ET, que juntas correspondem a aproximadamente 70% da variância explicada. As demais 24 variáveis, com menor contribuição individual, foram agrupadas sob a categoria "Variáveis Restantes" para fins de simplificação visual.

Figura 4 - Importância relativa das 25 variáveis mais relevantes segundo o modelo Extra Trees Regressor, totalizando aproximadamente 70% da importância cumulativa. As demais 24 variáveis foram agrupadas sob o rótulo “Variáveis Restantes” para facilitar a visualização.



Fonte: O Autor (2024)

Apesar de suas vantagens, o ET pode apresentar maior variância em conjuntos de dados pequenos, resultando em desempenho instável. Somado a isto, a aleatoriedade nos pontos de divisão pode comprometer a adequação local, ao não capturar os melhores pontos de separação.

Entretanto, a combinação de múltiplas árvores tende a mitigar essas limitações, estabilizando as previsões.

No processo de otimização dos hiperparâmetros, verificou-se que apenas "*n_estimators*" = 1000 teve impacto relevante nos resultados. O aumento do número de estimadores, em comparação ao valor padrão de 100, não levou ao *overfitting*, mas contribuiu para uma melhor estabilidade do modelo, alinhando-se à ideia de que o ganho de desempenho tende a convergir para um limite assintótico (BREIMAN, 2001).

Hiperparâmetros como “*max_depth*”, “*min_samples_leaf*” e “*min_samples_split*” foram inicialmente ajustados, mas não demonstraram melhorias significativas. Da mesma forma, “*max_features*” foi mantido no padrão (1.0), pois não influenciou de forma relevante na previsão. Outros hiperparâmetros, como “*bootstrap*”, “*oob_score*” e “*ccp_alpha*”, também permaneceram inalterados para evitar complexidade desnecessária. Dessa forma, a escolha final dos hiperparâmetros priorizou um modelo mais eficiente e interpretável, concentrando-se no ajuste que efetivamente trouxe benefícios à performance preditiva.

2.9.6 K-NEAREST NEIGHBORS REGRESSOR (KNN)

O algoritmo *k*-Nearest Neighbors (KNN) é um método de aprendizado supervisionado amplamente utilizado, originalmente proposto por Fix e Hodges (FIX; HODGES, 1951). Ele baseia-se na proximidade entre amostras para realizar predições, classificando elementos ou estimando valores com base nos *k* vizinhos mais próximos, definidos por métricas de distância como Euclidiana, Manhattan ou Minkowski (COVER; HART, 1967). Sua simplicidade e interpretabilidade o tornam uma escolha atraente para diversas aplicações, incluindo regressão, em que valores contínuos são previstos a partir da média (ou ponderação) dos vizinhos (ALTMAN, 1992).

Apesar de suas vantagens, o KNN apresenta limitações importantes, especialmente em contextos com alta dimensionalidade. Nessas situações, o fenômeno conhecido como “maldição da dimensionalidade” pode comprometer a eficácia do modelo, tornando menos significativa a distinção entre os vizinhos (BEYER et al., 1999). Ademais, o algoritmo é sensível à presença de atributos irrelevantes, uma vez que todas as variáveis são consideradas com o mesmo peso, o que pode levar ao *overfitting*.

Por outro lado, uma das principais qualidades do KNN é sua flexibilidade, pois não impõe suposições rígidas sobre a distribuição dos dados. Além disso, sua alta interpretabilidade permite identificar diretamente quais amostras vizinhas influenciaram determinada predição.

Visando aprimorar a capacidade preditiva do modelo e reduzir o *overfitting*, foi realizada a otimização dos principais hiperparâmetros do "*KNeighborsRegressor*". Foram ajustados os parâmetros "*n_neighbors*", "*weights*" e "*p*", por influenciarem diretamente o comportamento do modelo.

O hiperparâmetro "*n_neighbors*", cujo valor padrão é 5, apresentou melhor desempenho com valor igual a 10. Esse aumento reduz a variância do modelo, tornando-o menos sensível a ruídos pontuais, já que a predição passa a considerar um grupo mais amplo de vizinhos, o que suaviza o impacto de *outliers* e amostras atípicas.

O parâmetro "*weights*", que por padrão assume o valor "*uniform*", apresentou melhor desempenho com a configuração "*distance*". Nesse caso, os vizinhos mais próximos têm maior influência na predição, o que aprimora a adaptação do modelo a padrões locais e o torna mais robusto em regiões com densidade variável de dados (JAMES et al., 2021b).

O hiperparâmetro "*p*", que define a métrica de distância, apresentou melhores resultados com "*p*=1" (distância de Manhattan), em comparação com "*p*=2" (distância Euclidiana). Essa alteração permitiu uma melhor representação das relações não lineares presentes nos dados.

Os demais hiperparâmetros ("*algorithm", "leaf_size", "metric", "metric_params" e "n_jobs"*) não foram ajustados, por apresentarem impacto reduzido na qualidade das previsões ou já estarem implicitamente relacionados aos parâmetros otimizados. Assim, a configuração final adotada resultou em um modelo mais generalizável, com menor tendência ao sobreajuste e maior precisão preditiva.

2.9.7 REGRESSÃO LINEAR

A Regressão Linear é um método estatístico clássico tendo seu conceito inicial desenvolvido por Sir Francis Galton no final do século XIX. O método dos mínimos quadrados, que é a base matemática da Regressão Linear, foi formalizado anteriormente por Adrien-Marie Legendre em 1805 e aperfeiçoado por Carl Friedrich Gauss em 1809, que também demonstrou sua aplicação em problemas de ajuste de

órbitas astronômicas (STIGLER, 1981). O objetivo da Regressão Linear é modelar a relação entre uma variável dependente y e uma ou mais variáveis independentes x, ajustando um modelo que minimiza a soma dos quadrados dos resíduos. Sua equação geral é dada por:

$$y = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \varepsilon \qquad \text{Eq. 13}$$

Onde "y" representa a variável dependente, "x_i" são as variáveis independentes, "β_0" corresponde ao intercepto, "β_i" são os coeficientes das variáveis independentes, "p" é o número total de preditores, e "ε" representa o erro residual.

A regressão linear é amplamente utilizada devido à sua simplicidade, baixo custo computacional e alta interpretabilidade. Seus coeficientes fornecem uma estimativa direta da influência de cada variável sobre a resposta. Entretanto, seu desempenho pode ser limitado em cenários complexos, pois assume linearidade, homocedasticidade dos erros e baixa multicolinearidade entre os preditores (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Na otimização do modelo, foram considerados os hiperparâmetros "*fit_intercept*" e "*positive*". O primeiro define se o modelo ajustará o intercepto. O ajuste com "*fit_intercept* = *False*" apresentou melhor desempenho, indicando que os dados já estavam centrados ou que a inclusão do intercepto não contribuía significativamente para a modelagem.

O hiperparâmetro "*positive*", que restringe os coeficientes a valores não negativos, foi mantido como "*False*", permitindo coeficientes negativos. Essa configuração foi importante para modelar adequadamente as interações proteína-proteína, cujos descritores podem representar efeitos atrativos ou repulsivos.

Outros hiperparâmetros, como "*copy_X*", não influenciam diretamente a qualidade da regressão e, portanto, não foram otimizados.

2.9.8 RANDOM FOREST REGRESSOR (RF)

O Random Forest (RF), proposto por Breiman (2001), é um modelo de aprendizado de máquina baseado em árvores de decisão e no conceito de *ensemble learning*. O método consiste na construção de múltiplas árvores de decisão, cada uma treinada com subconjuntos de dados gerados por *bootstrap*. As previsões individuais são combinadas por média, reduzindo a variância do modelo. Além disso, cada divisão em uma árvore considera apenas um subconjunto aleatório de variáveis, promovendo diversidade entre as árvores e favorecendo a generalização. A impureza nos nós é minimizada com base no erro quadrático médio (MSE), tornando o algoritmo adequado para problemas de regressão.

Apesar de sua robustez contra o overfitting e da capacidade de capturar relações não lineares complexas, o RF pode apresentar custo computacional elevado em conjuntos de dados muito grandes (BREIMAN, 2001). Para otimizar seu desempenho, foram ajustados os principais hiperparâmetros que afetam diretamente a complexidade e a capacidade de generalização do modelo. Foram investigados: "*n_estimators", "max_depth", "min_samples_split", "min_samples_leaf" e "max_features"*. Por outro lado, hiperparâmetros como "*bootstrap*", "*oob_score*" e "*ccp_alpha*" não foram considerados, pois o modelo apresentou bom desempenho com as configurações padrão de amostragem, pois a validação cruzada substituiu a estimativa *out-of-bag* (OOB).

Os melhores resultados foram obtidos com "*n_estimators* = 1000" e "*max_features = 'log2*'", superando as configurações padrão de "*n_estimators* = 100" e "*max_features* = 1.0". O aumento no número de árvores reduziu a variância do modelo, tornando as previsões mais estáveis, sem comprometer significativamente o custo computacional. A adoção de "*max_features* = 'log2'" promoveu a seleção de subconjuntos menores de variáveis em cada divisão, reduzindo o risco de que variáveis com alta importância dominassem as decisões em múltiplas árvores, o que contribuiu para maior diversidade e melhor generalização do modelo.

Os demais hiperparâmetros foram mantidos em seus valores padrão. A profundidade ilimitada ("*max_depth = None*") não gerou *overfitting* significativo, graças ao número elevado de árvores e à seleção aleatória de variáveis. Os valores de "*min_samples_split* = 2" e "*min_samples_leaf* = 1" também foram preservados, pois restrições adicionais não trouxeram ganhos na validação cruzada, indicando

que o modelo já estava capturando adequadamente as relações presentes nos dados.

Uma comparação relevante ao RF é o modelo Extra Trees (ET), que também se baseia em múltiplas árvores de decisão, mas se diferencia pelo critério de divisão dos nós. Enquanto o RF seleciona os pontos de corte com base na maior redução da impureza, o ET escolhe os pontos de forma aleatória dentro dos limites das variáveis (GEURTS; ERNST; WEHENKEL, 2006). Essa abordagem pode reduzir ainda mais a variância e melhorar a generalização em certos contextos, porém, com possível perda de precisão individual em cada árvore, especialmente em conjuntos de dados com padrões mais definidos.

2.9.9 SUPPORTING VECTOR REGRESSOR (SVR)

O Support Vector Regressor (SVR) é um método de aprendizado supervisionado derivado do algoritmo de Support Vector Machines (SVM), desenvolvido por Vladimir Vapnik e Alexey Chervonenkis na década de 1990. Enquanto o SVM visa encontrar um hiperplano ótimo para separação de classes, o SVR adapta esse princípio à regressão, construindo uma função que busca manter os resíduos das predições dentro de uma margem de tolerância chamada *epsilon-tube*. Erros que permanecem dentro dessa margem não são penalizados, enquanto desvios superiores a ε geram penalidades proporcionais à magnitude do erro (PLATT, 2000).

Em problemas com múltiplos descritores, o SVR interpreta cada variável como uma dimensão no espaço de entrada, e estima um hiperplano de regressão que melhor representa a relação entre as variáveis preditoras e a variável alvo. Para modelar relações não lineares, o SVR pode empregar funções de kernel, que mapeiam os dados de entrada para espaços de maior dimensionalidade, onde padrões complexos tornam-se linearmente separáveis. Essa característica permite ao SVR capturar dependências não lineares com elevado grau de precisão.

A implementação do SVR na biblioteca scikit-learn é baseada na libsvm, que oferece uma otimização eficiente para o uso de kernels como o “*radial basis*

function” (RBF), linear e polinomial. Embora o treinamento com parâmetros padrão seja relativamente rápido, a otimização dos hiperparâmetros do SVR, especialmente do kernel, do parâmetro de penalização C e da largura ε, pode ser altamente custosa em termos computacionais. Isso ocorre devido à complexidade quadrática do algoritmo em relação ao número de amostras, o que dificulta sua aplicação em bases de dados maiores ou em tarefas que demandem validação extensiva.

Nos testes realizados, a busca pelos melhores hiperparâmetros por meio de “*grid_search”* foi interrompida após mais de uma semana sem conclusão, evidenciando a necessidade de estratégias de otimização mais eficientes, como “*random_search”*, “*Bayesian_optimization”* ou a aplicação de métodos de redução de dimensionalidade antes do ajuste fino do modelo.

2.9.10 EXTREME GRADIENT BOOSTING (XGBoost)

O Extreme Gradient Boosting (XGBoost) é um algoritmo avançado de aprendizado de máquina baseado na técnica de *boosting* de gradiente. Desenvolvido por Tianqi Chen em 2016, o XGBoost foi projetado com foco em desempenho, flexibilidade e escalabilidade, superando as limitações computacionais e preditivas de métodos anteriores como o AdaBoost e o Gradient Boosting Machine (GBM) (CHEN; GUESTRIN, 2016).

O XGBoost constrói modelos de forma sequencial, onde cada nova árvore de decisão é treinada para corrigir os erros residuais cometidos pelas árvores anteriores. Para mitigar o risco de *overfitting*, o algoritmo incorpora uma função objetivo regularizada, que penaliza a complexidade das árvores (por exemplo, número de folhas) e os pesos atribuídos a cada folha. Essa regularização estruturada contribui para uma melhor generalização, especialmente em conjuntos de dados com alta dimensionalidade.

Sua eficiência computacional decorre de várias otimizações implementadas na biblioteca: paralelização do processo de construção das árvores em múltiplos núcleos (CPUs), compressão de memória para armazenamento eficiente dos gradientes e pesos, e uso de histogramas discretos que agrupam os valores das variáveis, permitindo uma busca mais rápida pelos melhores pontos de divisão. Tais

estratégias tornam o XGBoost especialmente eficaz em ambientes com grandes volumes de dados, razão pela qual o algoritmo é amplamente utilizado tanto em aplicações industriais quanto em competições de *machine learning* (KE et al., 2017).

Durante os testes realizados neste trabalho, a configuração padrão do XGBoost mostrou-se adequada em termos de desempenho e eficiência computacional. A principal modificação que resultou em melhoria significativa foi a definição do objetivo da função de perda como "*objective* = '*reg:squarederror*'", apropriada para tarefas de regressão. Demais hiperparâmetros, como a taxa de aprendizado ("*eta"*), regularização ("*gamma"*), amostragem de atributos ("*colsample_bytree"*), profundidade máxima das árvores ("*max_depth*"), peso mínimo da folha ("*min_child_weight*") e número de estimadores ("*n_estimators"*), apresentaram desempenho estável em suas configurações padrão, sem ganhos expressivos após ajuste. A remoção dos termos de regularização L2 ("*reg_lambda* = 0") e L1 ("*reg_alpha* = 0") também não comprometeu a capacidade de generalização do modelo, indicando que os dados utilizados não exigiam penalização adicional para evitar sobreajuste.

A busca por configurações mais eficientes foi conduzida utilizando o método "*RandomizedSearchCV"* com validação cruzada *k*-fold (k = 10), o que permitiu explorar o espaço de hiperparâmetros com custo computacional reduzido em relação ao *grid search*. Dessa forma, a robustez da configuração padrão do XGBoost, aliada à seleção apropriada da função de perda, revelou-se suficiente para atingir bons resultados preditivos no conjunto de dados avaliado.

Por fim, devido ao seu desempenho competitivo e à estabilidade frente a diferentes configurações, o XGBoost foi incluído como um dos modelos base no *ensemble* preditivo desenvolvido neste trabalho. O próximo capítulo apresenta a arquitetura do metamodelo denominado *Super Learner*, que combina os pontos fortes dos 10 algoritmos de regressão descritos, para melhorar a acurácia na predição da energia livre de Gibbs da interface da interação proteína-proteína.

2.10 SUPER LEARNER (SL)

O Super Learner (SL) é um método de *ensemble* que combina diversos modelos de aprendizado de máquina, buscando gerar previsões mais precisas do que qualquer modelo individual. O teorema fundamental do Super Learner garante que sua performance será sempre pelo menos tão boa quanto a do melhor modelo individual da biblioteca. Além disso, para conjuntos de dados suficientemente grandes, o SL se aproxima da melhor combinação possível de modelos, o que justifica seu nome (VAN DER LAAN; POLLEY; HUBBARD, 2007).

A construção do SL inicia-se com a seleção de modelos de regressão. Neste estudo, empregamos um conjunto de 10 algoritmos de aprendizado de máquina previamente discutidos: AdaBoost Regressor, Bagging Regressor, Decision Tree Regressor, Elastic Net, Extra Trees Regressor, K-Nearest Neighbors, Regressão Linear, Random Forest Regressor, Support Vector Regressor e XGBoost. Cada um desses modelos contribui com diferentes abordagens de modelagem, permitindo que sua combinação leve a uma predição mais robusta e precisa da energia livre de ligação (ΔG).

O objetivo principal do SL é encontrar a melhor estimativa para a função $\psi_0(W)$ que prevê um resultado Y com base nas variáveis W. Essa função pode ser encontrada minimizando o erro de previsão (por exemplo, o erro quadrático médio) sendo a previsão final uma média ponderada das previsões individuais:

$$\hat{\psi}_{SL}(W) = \sum_{k=1}^{K} \alpha_k \hat{\psi}_k(W) \qquad \text{Eq. 14}$$

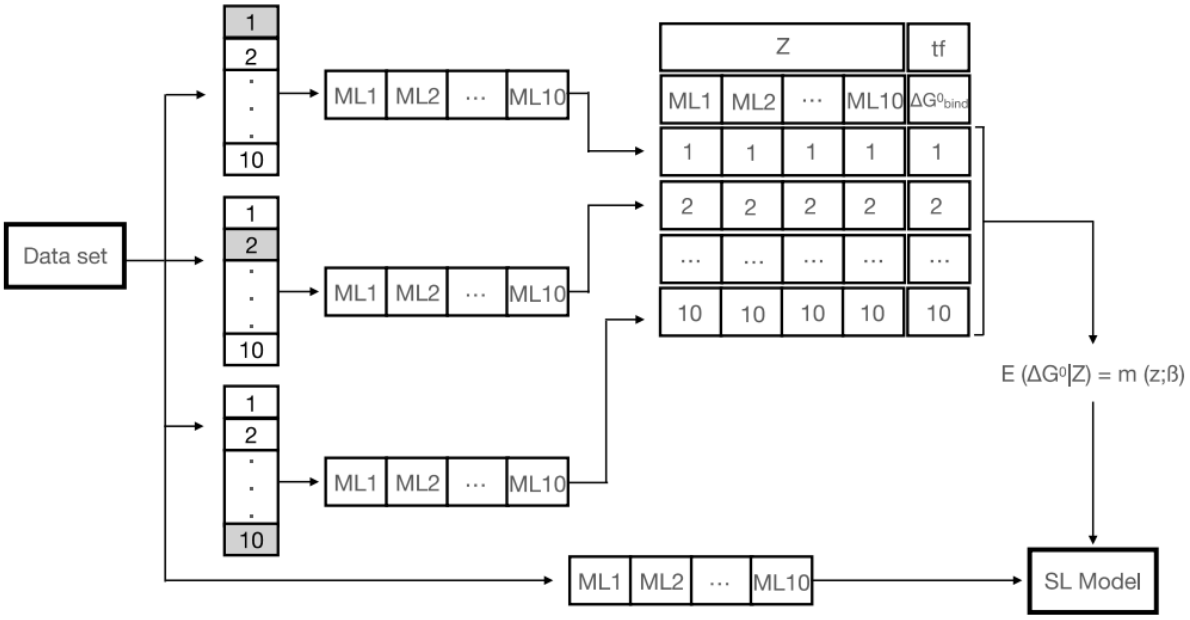
Os pesos α_k atribuídos a cada modelo base no Super Learner (SL) são não negativos e somam 1. Esse mecanismo permite que o SL reduza a influência de modelos menos eficazes, melhorando a estabilidade da previsão final. Em alguns casos, o SL pode atribuir todo o peso a um único modelo, caracterizando o chamado Super Learner discreto. No entanto, na maioria das situações, a combinação ponderada de múltiplos modelos tende a apresentar melhor desempenho.

O funcionamento do SL envolve duas etapas principais: (i) a aplicação de validação cruzada k-fold para gerar previsões fora da dobra (*out-of-fold predictions*) de cada modelo candidato; e (ii) o uso dessas previsões para treinar um metamodelo $m(z;\beta)$, que aprende a combinar os modelos base. O vetor z representa as covariáveis (descritores), e o metamodelo estima diretamente o valor esperado

da variável alvo $E(\Delta G^{o}|Z)$, sendo essa a energia livre de ligação padrão no presente estudo.

Na Figura 5, apresenta-se uma ilustração esquemática do processo de construção do SL, destacando as fases de treinamento, validação e combinação dos modelos base.

Figura 5 – Esquema do funcionamento do Super Learner. Cada algoritmo de ML é treinado com o conjunto de dados usando validação cruzada k-fold. As previsões fora da dobra são reunidas e usadas por um metamodelo que aprende a melhor forma de combinar os modelos base, resultando na predição final..



Fonte: adaptado(VAN DER LAAN; POLLEY e HUBBARD, 2007).

Com base no trabalho de Ferraz et al. (2023), que demonstraram o potencial dos descritores gerados pelo Rosetta na construção de redes neurais artificiais (ANN), foram calculados os mesmos descritores para um conjunto de dados expandido contendo 526 estruturas de alta resolução (resolução inferior a 3,5 Å), com valores experimentais de ΔG para dímeros e multímeros.

O Super Learner foi então treinado com esse conjunto, sendo a otimização direcionada à maximização do coeficiente de determinação (R^2).

3 OBJETIVOS

3.1 OBJETIVO GERAL

Contribuir para a avaliação e aprimoramento de um algoritmo de ML voltado à predição da afinidade de ligação em complexos proteína-proteína, com foco em alta precisão e baixo custo computacional.

3.2 OBJETIVOS ESPECÍFICOS

- Quantificar e comparar o potencial preditivo dos modelos de regressão que compõem o metamodelo em relação a ferramentas concorrentes, utilizando um mesmo conjunto de dados.
- Investigar o impacto de *outliers* na qualidade das predições.
- Avaliar o efeito da redução de dimensionalidade no desempenho preditivo do metamodelo.
- Identificar as variáveis de maior relevância na predição, visando ampliar a interpretabilidade e a extração de *insights* biológicos.

4 METODOLOGIA

4.1 DATASET

Os dados utilizados neste estudo foram extraídos de quatro bancos de dados distintos, conforme apresentado na Tabela 1. As informações de afinidade experimental foram obtidas diretamente das bases originais. Quando necessário, a energia livre de ligação (∆G) foi calculada a partir da constante de dissociação (K_d), utilizando a relação descrita na Equação 4, considerando $K_b = \frac{1}{K_d}$ e a temperatura de 298,15 K.

Tabela 1 - Tamanho do conjunto de dados não redundantes de complexos proteína-proteína utilizados para treinamento e validação dos modelos de aprendizado de máquina.

Banco de Dados	**Dados de Treinamento**	**Dados de Teste**
PDBbind[a]	366	92
PRODIGY[b]	2	-
SKEMPI2[c]	16	5
Benchmark 5.5[d]	36	9
Total	420	106

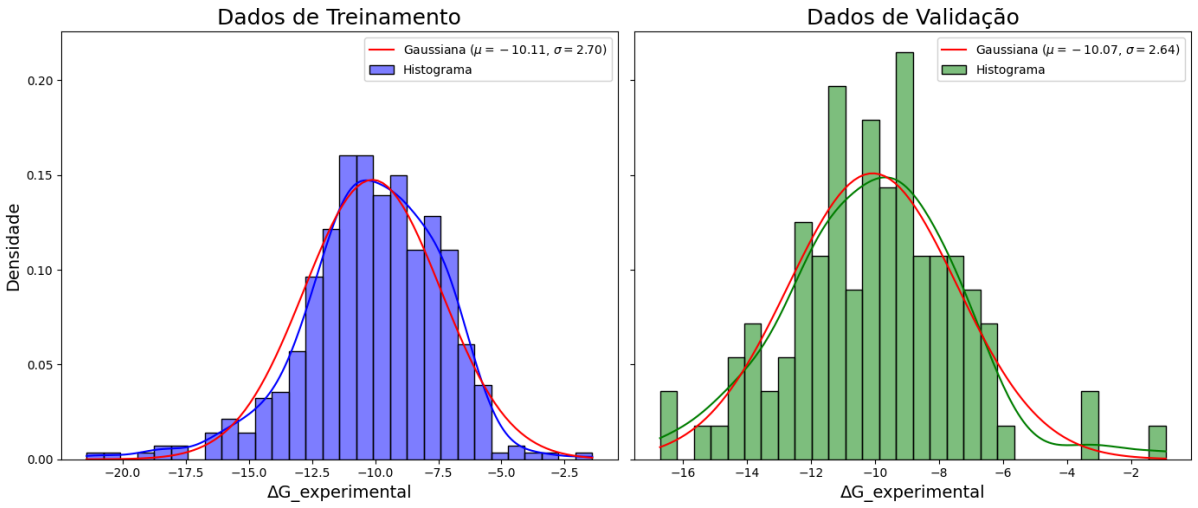
Fonte: O autor (2024)

Todas as estruturas utilizadas neste trabalho, bem como suas respectivas resoluções, estão listadas no Apêndice A. Foram descartadas as proteínas cuja resolução era inferior a 3,5 Å.

Os dados foram divididos aleatoriamente em dois subconjuntos: 80% para treinamento e 20% para validação. Essa proporção foi mantida também para cada banco de dados individualmente, exceto no caso do PRODIGY, cujos dados estão amplamente representados nas demais bases. Essa estratégia contribuiu para

manter um perfil de distribuição dos valores de ΔG experimental semelhante ao de uma distribuição normal, como ilustrado na Figura 6.

Figura 6 –Distribuição dos dados de treinamento e teste evidenciando comportamento semelhante ao de uma distribuição gaussiana (linha vermelha).



Fonte: O autor (2024)

Essa semelhança estatística entre os conjuntos de treinamento e teste fortalece a hipótese de que ambos compartilham propriedades estatísticas comparáveis, o que é desejável para a generalização dos modelos de aprendizado de máquina.

4.2 PREPARAÇÃO DAS ESTRUTURAS

O script completo utilizado para calcular os descritores está disponível no Anexo A, junto à Tabela AA1, que lista todos os descritores utilizados. A seguir, são brevemente descritas as principais etapas de seu funcionamento.

4.2.1 SCRIPT DE PRÉ - PROCESSAMENTO

O processo de pré-processamento das estruturas seguiu os seguintes passos:

i. Foi realizada uma busca por lacunas na estrutura do esqueleto proteico;
ii. Adicionaram-se hidrogênios ausentes na estrutura;
iii. Identificaram-se átomos iônicos estruturais, como Mg^{2+}, Ca^{2+}, Na^{+}, Cl^{-}, $Fe^{2+/3+}$, K^{+} e Zn^{2+}, que estivessem a uma distância euclidiana inferior a 2,5 Å de qualquer átomo da proteína;
iv. Foram eliminados outros tipos de heteroátomos, como íons não estruturais, moléculas de água, pequenos ligantes, cofatores e componentes de glicosilação, incluindo açúcares. Essa remoção visou evitar ruídos nos descritores gerados, uma vez que a presença desses elementos poderia comprometer a representação física da interação

Adicionalmente, durante a análise preliminar dos dados, observou-se que algumas estruturas apresentavam valores de ΔG superiores a três ordens de grandeza em relação ao intervalo esperado. Por esse motivo, essas estruturas foram excluídas do conjunto final por se considerar que tais desvios representavam potenciais erros de cálculo. Essa curadoria visou garantir que os descritores calculados refletissem corretamente o comportamento físico-químico das interações e não fossem influenciados por valores atípicos de energia.

4.2.2 SCRIPT DE PÓS – PROCESSAMENTO

4.2.2.1 OTIMIZAÇÃO GEOMÉTRICA

As estruturas resultantes do pré-processamento foram otimizadas utilizando o software Rosetta, conforme o protocolo descrito a seguir:

1 - Realizou-se uma primeira otimização das cadeias laterais, mantendo o esqueleto principal fixo;
2 - Posteriormente, todos os átomos foram otimizados com o algoritmo de minimização padrão do Rosetta, configurado com 50.000 etapas e critério de convergência de 10^{-4} unidades de energia do Rosetta;
3 - Durante a otimização da colocação de hidrogênios, as orientações dos resíduos de histidina, asparagina e glutamina foram ajustadas,

além de serem amostrados os ângulos rotâmeros χ_1 e χ_2 de todos os resíduos.

4 - Os potenciais da função *score* "*BETA_NOV16*" foram empregados para avaliar a geometria durante a otimização.

4.2.2.2 DESCRITORES DE INTERAÇÃO PROTEÍNA-PROTEÍNA

As propriedades das interfaces proteína–proteína foram calculadas por meio das aplicações do Rosetta: "*InteractionEnergyMetric*", "*ContactMolecularSurface*", "*InterfaceHoles*" e "*InterfaceAnalyzerMover*".

O parâmetro "*auto_setup_metals*" foi ativado nos casos em que foram detectados íons nas estruturas. Todas as propriedades foram calculadas utilizando os potenciais "*BETA_NOV16"*.

O conjunto final de dados incluiu 49 descritores (listados na Tabela AA1), obtidos para cada uma das 526 estruturas, juntamente com seus respectivos valores experimentais de ΔG. .

4.3 SUPER LEARNER

Cada modelo base foi treinado previamente e utilizado na construção do SL. A matriz de covariância gerada possui exemplos como linhas e previsões dos modelos como colunas, fornecendo uma representação abrangente das respostas dos modelos. A partir dessa matriz de previsões, foi construído um metamodelo $m(z;\beta)$ com o objetivo de estimar $E(\Delta G_0|Z)$, que representa a função-alvo ΔG_0, no bloco de validação correspondente ao bloco de treinamento do modelo candidato.

Esse metamodelo atua como um segundo nível de aprendizado, combinando as previsões dos modelos base por meio de uma curva de regressão. Assim, o SL aprende a ponderar os modelos individuais de forma otimizada, buscando melhorar a precisão e a estabilidade da predição.

A implementação do SL foi realizada conforme os passos descritos abaixo:

I. Preparação dos Dados:

- Os dados foram carregados a partir do arquivo de treinamento ".csv", removendo colunas irrelevantes ("*pdb*", "*database*", "*partner1*", "*partner2*"). A variável alvo foi definida como "*dG_exp*".

II. Carregamento dos Modelos Base:
    - Os modelos foram previamente treinados e salvos em arquivos ".pkl", garantindo que as predições fossem feitas sem a necessidade de reprocessamento.

III. Treinamento dos Metamodelos:

Foram testados diferentes metamodelos, incluindo:

- Regressão Linear
- Elastic Net
- XGBoost
- Extra Trees
- Ridge Regression
- MLP (Perceptron Multicamadas)

Cada metamodelo foi ajustado sobre a matriz de previsões dos modelos base, gerando a predição final do SL.

IV. Salvamento dos Modelos:

Após o treinamento, os metamodelos foram armazenados em arquivos .pkl para uso em predições futuras.

Dessa forma, o SL não apenas combina as previsões dos modelos individuais, mas também aprende a ponderá-las de maneira otimizada, promovendo maior robustez e precisão na predição dos valores de ΔG.

O código utilizado tanto para o treinamento dos modelos individuais quanto para a implementação do SL encontra-se disponível no Apêndice B.

4.4 IDENTIFICAÇÃO E REMOÇÃO DE *OUTILERS*

Para avaliar a presença de *outliers* no conjunto de dados de treinamento, foram analisadas as distribuições dos 49 descritores em relação à variável alvo, ΔG experimental. A normalidade dos descritores foi verificada por meio do teste de Shapiro-Wilk, adequado para amostras com menos que 5000 observações.

Para reduzir o erro “Tipo I” proveniente da realização de múltiplos testes, foi aplicada a Correção de Bonferroni ao nível de significância (α), ajustando-o para $\alpha' = \frac{0,05}{N}$, onde N representa o número total de descritores testados.

Os resultados indicaram que apenas dois descritores, “*ifa_hbond_E_fraction*” e “*ifa_packstat*”, seguem uma distribuição normal. Para esses, foram identificados três outliers considerando um intervalo de confiança de 99,7%.

Para os demais descritores, que não apresentaram aderência à normalidade, a identificação de *outliers* foi realizada utilizando dois métodos complementares:

- Intervalo Interquartil (IQR): valores situados além de 1,5 vezes o intervalo interquartil (IQR) abaixo do primeiro quartil (Q1) ou acima do terceiro quartil (Q3) foram classificados como outliers. Esse critério é amplamente utilizado por sua robustez na detecção de valores extremos em distribuições não normais. A aplicação desse método resultou na identificação de 152 *outliers*.
- Isolation Forest (IForest)**:** algoritmo baseado na facilidade de isolamento de valores discrepantes em árvores de decisão. Para garantir estabilidade na detecção, os hiperparâmetros foram definidos como "*n_estimators*" = 1000, assegurando um número adequado de árvores, e "contamination" = *'auto'*, permitindo que o modelo ajustasse automaticamente a fração de *outliers* de acordo com a distribuição dos dados. O valor "*random_state* = 42" foi utilizado para garantir a reprodutibilidade dos resultados. Com essa configuração, 35 *outliers* foram identificados.

A interseção dos métodos identificou 35 outliers que foram detectados tanto pelo IForest quanto pelo IQR. Foram removidas 37 estruturas (35 da interseção e 2 detectadas por distribuição Gaussiana) para avaliação do impacto no desempenho.

A amostra '4o27' foi o único *outlier* identificado por todos os métodos. A Tabela C1, no Apêndice C, lista os *outliers* removidos e seus respectivos métodos de identificação.

Como perspectivas futuras, sugere-se a investigação do uso do algoritmo t-SNE (*t-distributed Stochastic Neighbor Embedding*) como uma abordagem complementar para identificação de *outliers* em espaços de alta dimensionalidade. Essa técnica de redução de dimensionalidade pode auxiliar na visualização e detecção de padrões incomuns, possibilitando a identificação de amostras que se

comportam de maneira discrepante em relação ao restante do conjunto de dados em distribuições que não seguem a normalidade.

4.5 ANÁLISE DE REDUÇÃO DE DIMENSIONALIDADE

Todos os modelos de classificação e de redução de dimensionalidade foram avaliados utilizando-se o conjunto completo dos dados de teste. Muitas variáveis apresentaram forte correlação entre si (Figura 8), o que motivou a aplicação de técnicas de redução de dimensionalidade. Foram utilizados os métodos de Análise de Componentes Principais (PCA), *Uniform Manifold Approximation and Projection* (UMAP) e a adaptação do classificador de variáveis *Shapley Additive Explanations* (SHAP) como abordagem para seleção de atributos.

Outros métodos não foram considerados, pois os descritores não seguem uma distribuição normal, conforme detalhado na Seção 5.1. Ressalta-se, no entanto, que métodos não lineares como o t-SNE (*t-distributed Stochastic Neighbor Embedding*), que não exigem normalidade dos dados, poderiam ter sido empregados, contudo, sua aplicação não foi explorada neste estudo.

4.5.1 PCA (PRINCIPAL COMPONENT ANALYSIS)

O método PCA foi aplicado para reduzir a dimensionalidade dos dados por meio da decomposição da matriz de dados via Singular Value Decomposition (SVD), técnica que não exige pressupostos de normalidade das variáveis. O modelo foi ajustado exclusivamente com os dados de treinamento, e os dados de teste foram posteriormente projetados no espaço definido pelos componentes principais extraídos. Apenas os componentes que explicaram 95 % da variância total foram mantidos, o que resultou na redução para 11 variáveis.

4.5.2 UMAP (UNIFORM MANIFOLD APPROXIMATION AND PROJECTION)

O método UMAP foi empregado para preservar tanto a estrutura local quanto a global dos dados em um espaço de menor dimensão. O número ideal de dimensões a ser mantido foi determinado por meio da otimização do coeficiente de determinação R^2 , utilizando como referência o desempenho do modelo Extra Trees

(ET). O processo iterativo variou o número de dimensões de 2 até o total de variáveis (49), resultando na seleção de 39 dimensões. Os dados de teste foram, então, reduzidos para a mesma dimensionalidade para possibilitar a avaliação.

Adicionalmente, duas reduções alternativas foram realizadas: a primeira utilizando como referência o desempenho do modelo ElasticNet, que resultou em 4 variáveis; e a segunda utilizando um valor intermediário de 23 variáveis, escolhido arbitrariamente para fins comparativos.

4.5.3 SHAP (SHAPLEY ADDITIVE EXPLANATIONS)

Após a classificação e normalização das variáveis principais (Seção 4.6.1) por meio do método SHAP, as variáveis foram ordenadas em ordem decrescente de importância, selecionando-se o subconjunto necessário para explicar até 95% da importância acumulada. Como resultado, foram selecionadas 42 variáveis.

A importância relativa de cada variável foi documentada no Apêndice F. Os dados de teste foram preparados removendo-se as mesmas 7 colunas irrelevantes para análise e que foram previamente excluídas no conjunto de treinamento. Posteriormente, os modelos foram treinados com o novo conjunto reduzido e comparados por meio das métricas de avaliação.

4.6 CLASSIFICAÇÃO DAS VARIÁVEIS PRINCIPAIS PELO SHAP

O SHAP é um método derivado da teoria dos valores de Shapley, originalmente desenvolvida no âmbito da teoria dos jogos cooperativos (Lundberg; Lee, 2017). Essa teoria fornece uma formulação matemática para distribuir, de maneira justa, os ganhos ou as contribuições entre os participantes de um sistema. No contexto do aprendizado de máquina, os valores de Shapley são utilizados para atribuir a cada variável preditora uma importância proporcional ao seu impacto nas previsões do modelo. Isso garante que a influência de cada variável seja calculada de forma consistente, considerando diferentes combinações de entrada. A eficácia desse método já foi validada em estudos anteriores, como na análise de fatores associados à mortalidade por COVID-19 (Nohara et al., 2022).

O SHAP foi aplicado por meio da função “*KernelExplainer*”, disponibilizada em sua biblioteca, em conjunto com o modelo Extra Trees (ET) para classificar as variáveis, como perspectivas futuras, seria interessante associar o “*KernelExplainer*” do SHAP à outros modelos de regressão para verificar se há convergência no método de seleção, algo que não foi explorado neste trabalho, escolheu-se o ET por ser o modelo que apresentou melhor desempenho individual geral.

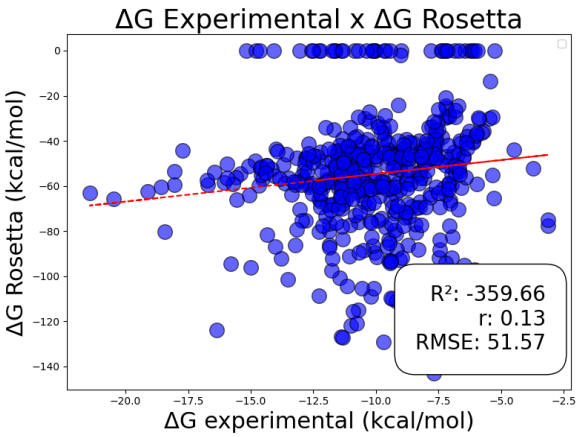
Inicialmente, o modelo ET foi treinado para prever os valores de energia livre de Gibbs experimental (ΔG_exp) a partir dos descritores estruturais. Após o ajuste do modelo, o “*KernelExplainer*” foi utilizado para estimar os valores SHAP das variáveis. Esse método, baseado em amostragem, calcula a influência de cada variável ao analisar como as previsões do modelo variam frente a diferentes combinações de entradas.

Os valores SHAP absolutos médios foram calculados para cada variável, refletindo sua influência média nas previsões. Para facilitar a interpretação, esses valores foram normalizados, de modo que a soma total das importâncias resultasse em 100%. A normalização foi realizada dividindo-se a importância individual de cada variável pela soma total e multiplicando-se por 100. Os resultados dessa análise estão disponíveis no Apêndice F.

5 RESULTADOS E DISCUSSÃO

O Rosetta foi utilizado para gerar os descritores de interface neste estudo. A correlação entre ΔG de ligação calculada pelo Rosetta e os dados experimentais está mostrada na Figura 7. Observa-se uma baixa correlação entre os valores previstos e experimentais, o que reforça as limitações do Rosetta para essa finalidade específica. Isso ocorre porque, embora o Rosetta seja amplamente utilizado para modelagem estrutural e cálculo de energias relativas, ele não foi originalmente desenvolvido com o objetivo de estimar com precisão a energia livre de ligação de complexos proteína–proteína em valores absolutos. A falta de correlação não é inesperada, uma vez que o Rosetta não foi desenvolvido especificamente para prever energias de ligação de complexos proteína-proteína, o que justifica a aplicação desse trabalho. Nesse contexto, os descritores derivados do Rosetta ainda podem capturar informações estruturais úteis da interface, mas exigem a aplicação de métodos de aprendizado de máquina para traduzir essas características em estimativas quantitativas mais precisas de ΔG. No entanto, seu uso pode ser combinado com técnicas de redes neurais (FERRAZ et al., 2023) ou aprendizado de máquina para aprimorar as predições de ΔG (SHRINGARI et al., 2020). A escolha da "*score function*" BETA_NOV16, em vez da REF15, baseia-se em um estudo que demonstrou que a BETA_NOV16 apresenta melhores descritores para a predição de ΔG para mutações que ocorrem na interface da interação proteína-proteína (SHRINGARI et al., 2020).

Figura 7 – Correlação r entre o valor de ΔG experimental e o valor calculado diretamente pelo software Rosetta.



Fonte: O autor (2024)

Embora o coeficiente de determinação R^2 seja comumente interpretado como a proporção da variância explicada pelo modelo, valores negativos podem ocorrer. Um R^2 negativo indica que o modelo teve desempenho pior do que uma predição baseada apenas na média dos valores observados, ou seja, a soma dos erros quadráticos do modelo superou a soma dos erros de um modelo nulo.

Esse resultado pode refletir a baixa capacidade preditiva do modelo frente ao conjunto de dados avaliado, especialmente quando há *overfitting*, subajuste (*underfitting*) ou pouca correlação entre os descritores e a variável-alvo. Assim, o valor negativo de R^2 sugere que o modelo não conseguiu capturar adequadamente os padrões subjacentes aos dados.

5.1 DATASET E SUAS CORRELAÇÕES

Inicialmente, com foco na premissa de que correlação não implica em causalidade, avaliou-se, em pares, a correlação entre os 49 descritores, por meio da construção da matriz de correlação de Pearson utilizando a biblioteca seaborn versão 0.13-2. Foi constatado que diversas variáveis apresentam correlações significativas, como ilustrado na Figura 8 pelo mapa de calor.

Entre as 1176 combinações possíveis de pares, 143 apresentaram correlações com $|r \geq 0,9|$, indicando uma possível relação linear. Isso corresponde a 12,2% das combinações entre pares de descritores, sendo 27 descritores diferentes listados na Tabela 2.

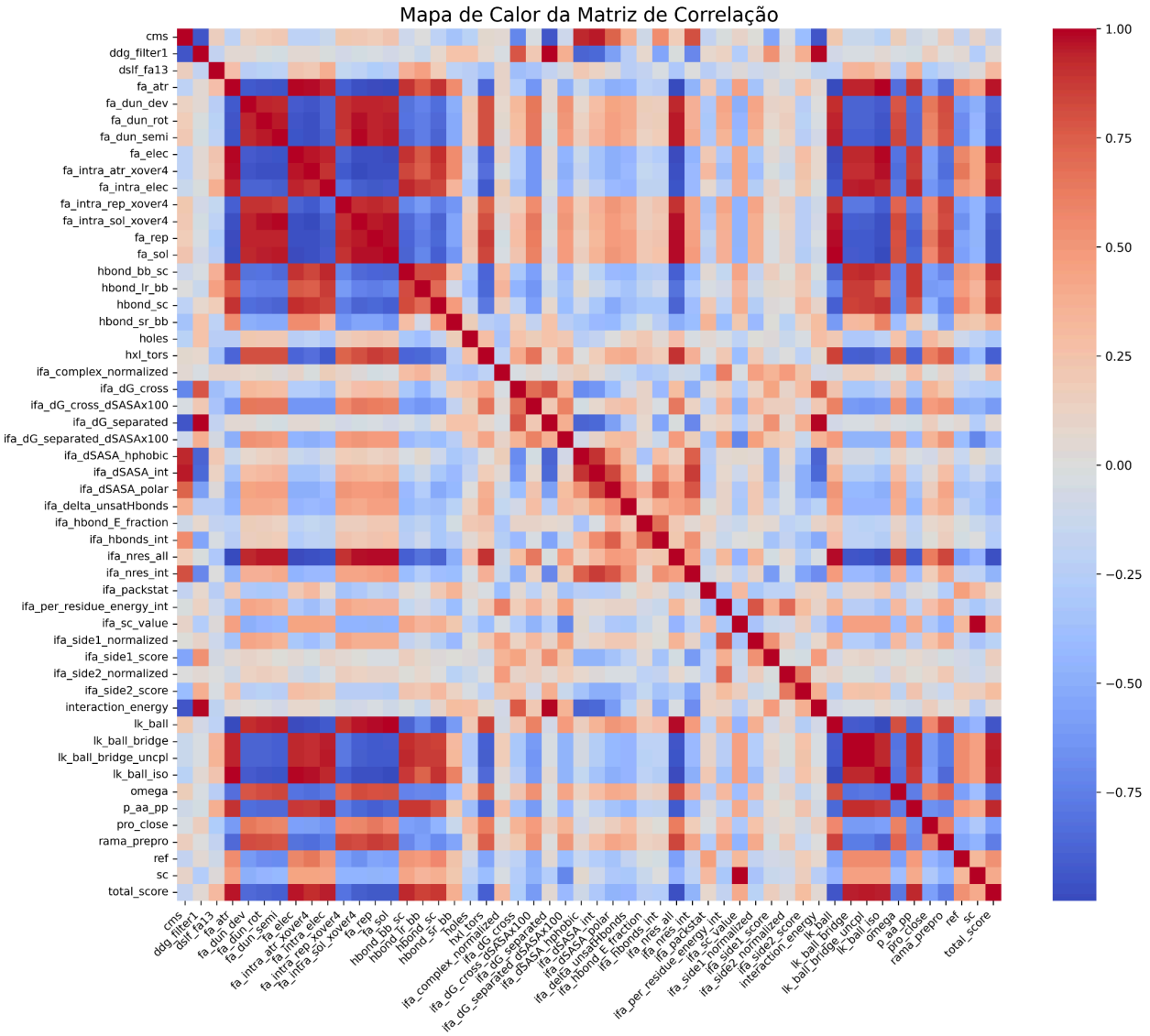
Tabela 2 – Lista com todos os 27 descritores que possuem 2 ou mais pares de correlações de Pearson (r), em módulo, iguais ou superiores a 0,9.

Descritores com $\|r \geq 0,9\|$		
Cms	fa_dun_semi	lk_ball
ddg_filter1	fa_elec	lk_ball_bridge
ifa_dSASA_hphobic	fa_intra_atr_xover4	lk_ball_bridge_uncpl

ifa_dSASA_int	fa_intra_elec	lk_ball_iso
interaction_energy	fa_intra_rep_xover4	total_score
ifa_dG_separated	fa_intra_sol_xover4	hbond_sc
fa_atr	fa_rep	hbond_bb_sc
fa_dun_dev	fa_sol	hxl_tors
fa_dun_rot	ifa_nres_all	p_aa_pp
Total = 27 Descritores diferentes		

Fonte: O autor (2024)

Figura 8 – Mapa de calor da matriz de correlação.



Fonte: O autor (2024)

Essa observação, sugeriu a importância de se avaliar o modelo após a aplicação de técnicas de redução de dimensionalidade conforme é discutido na seção 5.4.

5.2 AVALIAÇÃO DOS MODELOS

Inicialmente, avaliou-se o desempenho individual dos dez modelos de regressão que compõem o Super Learner (SL), utilizando os hiperparâmetros otimizados conforme descrito na Seção 2.9. Cada execução do experimento segue as etapas abaixo:

1. Inicialização e aleatoriedade – em cada uma das 100 repetições, um valor de semente pseudo-aleatória é gerado para garantir variações independentes nas partições de dados.
2. Validação cruzada 10-*fold* – dentro de cada repetição, o conjunto de treinamento é particionado em 10 folds estratificados, sendo o modelo ajustado em nove folds e validado no fold restante, percorrendo todos os folds para reduzir viés de partição.
3. Treinamento completo – após a validação cruzada, cada modelo é treinado novamente usando 100% do conjunto de treinamento.
4. Predição e métricas – os modelos ajustados geram predições sobre o próprio conjunto de treinamento completo e um conjunto de teste externo fixo. Em ambos os conjuntos calculam-se três métricas: coeficiente de determinação (R^2), correlação de Pearson (r) e erro quadrático médio (RMSE).
5. Armazenamento de modelos – os coeficientes/árvores do modelo ajustado na primeira repetição são salvos em arquivos “.pkl”, possibilitando seu reaproveitamento posterior no SL.
6. Estimativa do erro associado – ao final das 100 repetições, para cada modelo e métrica (em treino e teste) calcula-se:
    a. A média dos 100 valores;

b. O erro definido como metade da amplitude do intervalo empírico de 95 % (percentis 2,5 % e 97,5 %), obtido diretamente da distribuição observada das repetições.

A justificativa para o cálculo do erro seguindo este procedimento ocorre pois, dos 49 descritores utilizados, apenas dois apresentaram distribuição compatível com a normalidade (teste de Shapiro–Wilk, $\alpha = 0,05$). Portanto, estimar a incerteza via desvio-padrão (que pressupõe a normalidade) seria inadequado. O intervalo empírico de 95 % foi calculado sem hipóteses paramétricas, simplesmente ordenando-se os 100 valores obtidos e selecionando-se a faixa central que contém 95 % deles. Obtendo-se a variabilidade real causada por diferentes partições de dados e pelo processo de treinamento, fornecendo um erro (“±”) robusto mesmo diante de distribuições assimétricas ou com caudas longas, obtendo-se, dessa forma, a estimativa de erro associada às medições neste trabalho. Todos os resultados com seus respectivos erros de treinamento e teste dessa análise para os modelos individuais ou metamodelos componentes do SL podem ser consultados no Apêndice G, onde observou-se que os dados de treinamento foram melhores generalizados pelos modelos baseados em árvores de decisão, como o XGBoost, RF, ET e também pelo KNN, método baseado nos k vizinhos mais próximos.

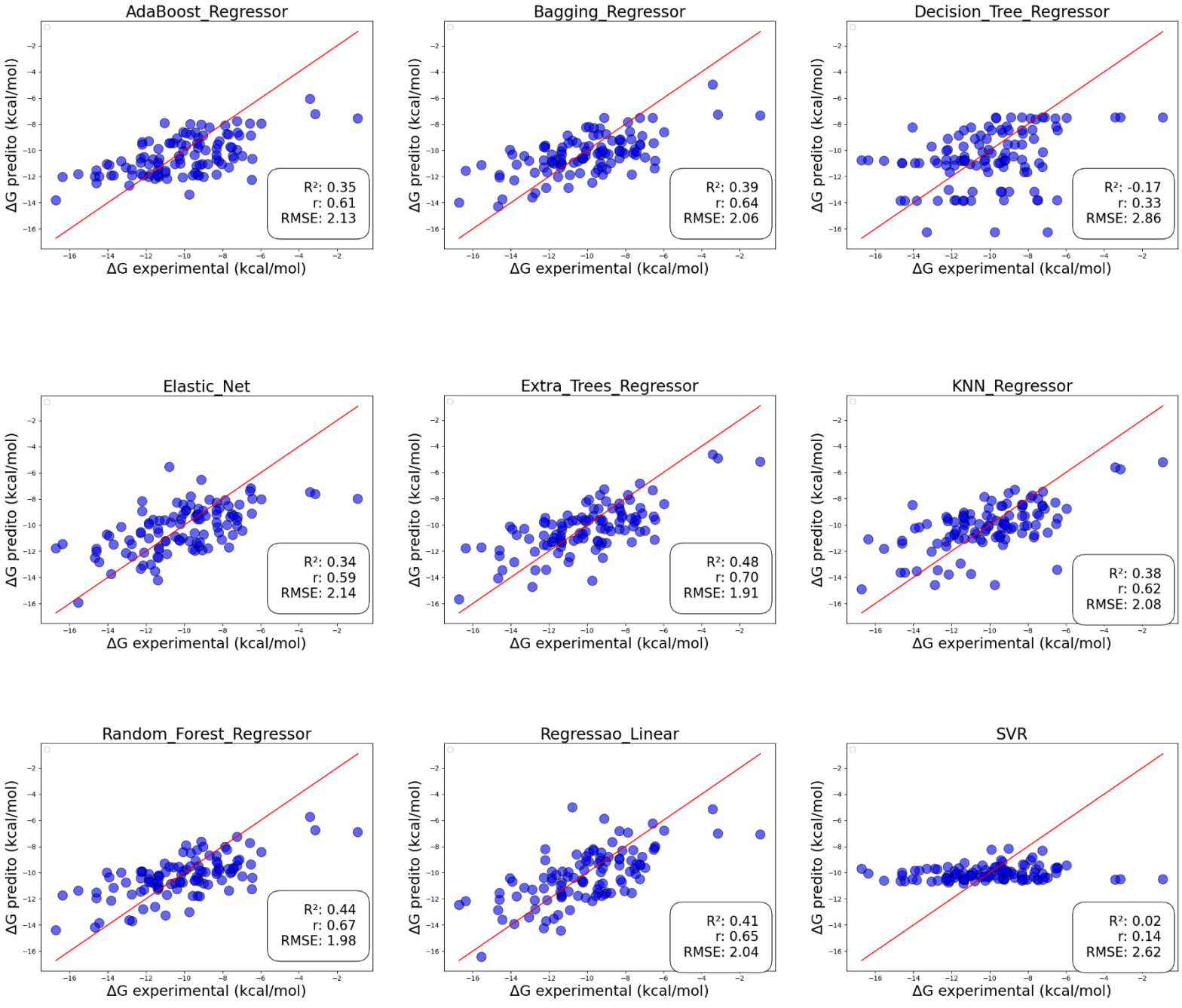
O SL seguiu um processo semelhante, combinando as previsões dos modelos base previamente treinados por meio de um conjunto de meta-modelos, incluindo Regressão Linear, ElasticNet, XGBoost, Extra Trees, Ridge e um Perceptron Multicamadas (MLP). Cada um desses meta-modelos aprendeu a ajustar pesos para minimizar o erro global.

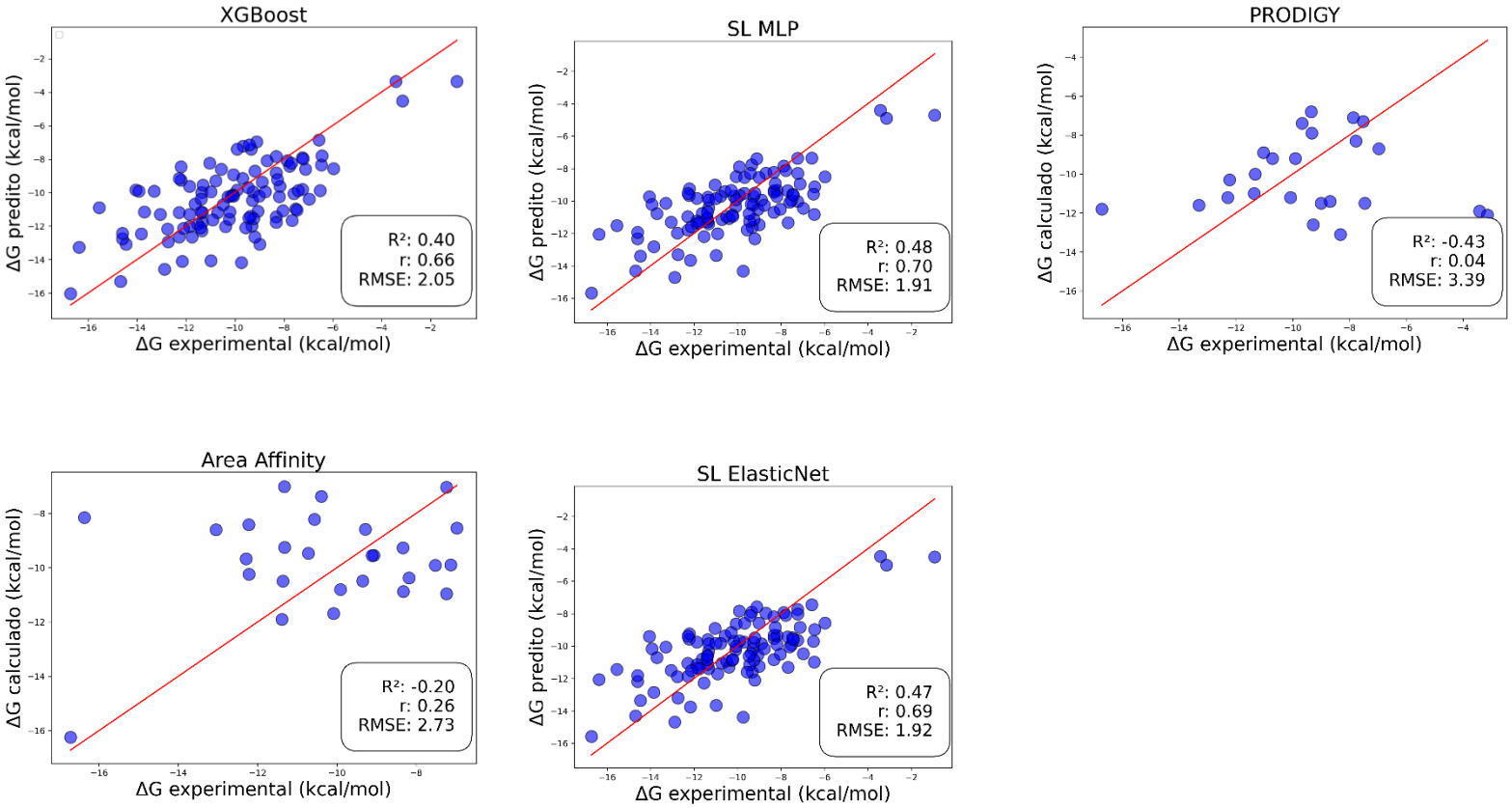
Para validar a eficácia do SL em comparação com abordagens concorrentes, foram utilizados os métodos "Area Affinity" e "PRODIGY". Para isso, selecionaram-se aleatoriamente 25 amostras do conjunto de teste, garantindo que nenhuma tivesse sido utilizada no treinamento desses métodos. No caso do PRODIGY, seguiu-se um critério mais rigoroso, separando interações proteína-proteína de interações proteína-ligante. Assim, o PRODIGY foi aplicado exclusivamente a interações proteína-proteína, considerando apenas estruturas onde ambas as cadeias eram representadas por letras simples (ex.: A, B, C). Já as interações proteína-ligante foram identificadas pela presença de códigos múltiplos (ex.: ABC, BD) ou caracteres não convencionais para proteínas (ex.: T, H, X).

Após a análise dos resultados, observou-se que os meta-modelos MLP e ElasticNet apresentaram os melhores desempenhos dentro do SL. Dessa forma, as análises posteriores focaram nessas duas versões do SL.

Os resultados comparativos dos modelos individuais, do SL e dos métodos concorrentes são apresentados na Figura 9.

Figura 9 - Gráficos de ΔG predito vs. ΔG experimental para cada modelo individual, o SL MLP, o SL ElasticNet e os métodos concorrentes, incluindo métricas de regressão e a linha de tendência (vermelha) ajustada aos dados não utilizados no treinamento.





Fonte: O autor (2024)

Os métodos concorrentes, “Area Affinity” e PRODIGY, apresentaram os piores desempenhos nas métricas avaliadas com o conjunto de teste, ambos resultando em um valor negativo de R^2. Isso indica que suas previsões foram inferiores à simples média dos dados. Além disso, os coeficientes de correlação (r) foram significativamente menores do que os reportados na literatura pelos respectivos autores (0,92 e 0,73, respectivamente). Entre os modelos individuais desenvolvidos, o SVR apresentou o pior desempenho. Esse resultado era esperado, pois o SVR tende a obter melhor performance em cenários com menor volume de dados e menor dimensionalidade dos descritores.

Um ponto interessante foi o desempenho da Regressão Linear Simples, que, apesar de sua simplicidade, obteve métricas competitivas, com r = 0,65, R^2 = 0,41 e RMSE = 2,04. Esses valores sugerem que os dados possuem uma estrutura definida, permitindo que até mesmo métodos baseados em mínimos quadrados apresentem resultados razoáveis.

Entre os modelos individuais, o ExtraTrees obteve o melhor desempenho global, com r = 0,70, R^2 = 0,48 e RMSE = 1,91, desempenho comparável ao do SL MLP. Esse resultado indica que o modelo conseguiu capturar padrões relevantes nos dados, apresentando um ajuste linear relativamente bom e um erro médio reduzido.

A análise bayesiana foi empregada com o objetivo de investigar a presença de erros sistemáticos nos resíduos dos modelos preditivos. Para isso, utilizou-se o pacote PyMC, que permitiu a construção de um modelo hierárquico bayesiano de regressão simples entre valores preditos e resíduos nos dados de teste. Foram utilizados 3000 passos de amostragem MCMC com tune=1000 e 4 cadeias paralelas para robustez estatística.

A Tabela 3 apresenta os resultados da análise bayesiana aplicada aos resíduos dos modelos no conjunto de teste. Para cada modelo, são mostrados a média posterior estimada do parâmetro μ, o intervalo de credibilidade a 95% e a probabilidade posterior de $|\mu|>0$, interpretada como a evidência de viés sistemático.

Tabela 3 - Estatísticas dos resíduos dos modelos de regressão estimadas via inferência bayesiana.

Modelo	**$\|\mu\|$**	**HDI95_Baixo**	**HDI95_Alto**	**P(|mu|>0)**
Regressao_Linear	0,210	-0,183	0,595	1,00
Elastic_Net	0,197	-0,181	0,616	1,00
KNN	0,025	-0,381	0,451	1,00
Decision Tree	0,196	-0,342	0,719	1,00
SVR	-0,028	-0,551	0,444	1,00
AdaBoost	0,326	-0,083	0,722	1,00
Bagging	0,231	-0,105	0,632	1,00
RF	0,206	-0,152	0,634	1,00

ET	0,212	-0,145	0,587	1,00
XGBoost	0,270	-0,140	0,624	1,00

Fonte: O Autor (2025)

Apesar de 47 dos 49 modelos apresentarem resíduos que não seguem a distribuição normal, a inferência bayesiana foi mantida com base na robustez do método e na adequação do modelo de média com suposições condicionais. Além disso, o objetivo da análise foi estimar a tendência média dos resíduos, e não realizar testes de hipóteses clássicos que dependam fortemente da normalidade.

Dentre os modelos analisados, o AdaBoost Regressor apresentou o maior viés médio positivo (MediaPosterior = 0,326), indicando uma tendência sistemática de subestimar os valores reais. Em seguida, destacam-se o XGBoost (0,270) e o Bagging Regressor (0,231), com comportamento semelhante. Outros modelos com viés positivo de menor magnitude incluem a Regressão Linear, Random Forest Regressor e Extra Trees Regressor, todos com média posterior entre 0,206 e 0,212.

O único modelo a apresentar viés médio negativo foi o SVR, com valor estimado de -0,028, sugerindo uma leve tendência à superestimação. No entanto, a proximidade de zero e a amplitude do intervalo de credibilidade indicam que essa tendência tem baixa relevância prática.

Todos os modelos apresentaram $P(|\mu| > 0) = 1.0$, evidenciando alta confiança de que o viés médio dos resíduos é diferente de zero. Entretanto, a largura dos intervalos de credibilidade varia entre os modelos. Os modelos KNN Regressor e Decision Tree Regressor, por exemplo, apresentaram intervalos consideravelmente mais amplos, indicando maior incerteza sobre a estimativa de viés.

Os resultados reforçam que, embora todos os modelos apresentem algum grau de viés sistemático, a intensidade e a direção do viés variam significativamente entre eles, o que é relevante para a seleção do modelo mais confiável, especialmente em aplicações sensíveis à precisão e interpretação dos resultados.

Posteriormente, realizou-se o teste de permutação com a finalidade de avaliar a significância estatística das métricas de desempenho dos modelos de regressão (RMSE, R^2 e coeficiente de Pearson) no conjunto de teste. O procedimento foi

implementado por meio da permutação dos vetores de predição (y_pred) em 1 000 iterações, utilizando as funções do scikit-learn (mean_squared_error, r2_score, shuffle) e da SciPy (pearsonr) para recalcular, em cada permutação, o RMSE, o R^2 e a correlação.

Cujos resultados revelaram que, com exceção do SVR ($p > 0,05$ em todas as métricas), todos os demais algoritmos — em especial os métodos de ensemble (Bagging, Random Forest, Extra Trees e XGBoost), apresentaram p-valores inferiores a 0,001, indicando que seu desempenho observado ($R^2 \approx 0,45–0,47$; $r \approx 0,68–0,69$; RMSE significativamente menor que o acaso) não pode ser atribuído ao mero acaso. Já o Decision Tree, apesar de $p < 0,01$, exibiu R^2 negativo e baixa correlação ($r \approx 0,27$), o que demonstra que significância estatística não equivale necessariamente a bom ajuste. Os resultados completos dessa análise podem ser encontrados no Apêndice H.

5.2.1 AVALIAÇÃO DOS MODELOS POR BANCOS DE DADOS

Para verificar a robustez dos resultados, os dados de teste foram segmentados conforme suas bases de origem, resultando em 92 estruturas do PDBbind, 9 do Benchmark e 5 do SKEMPI. A análise individual de cada banco revelou que os modelos mantiveram tendências semelhantes às observadas no conjunto completo, reforçando a consistência dos padrões de predição.

No PDBbind, os resultados foram alinhados com aqueles observados no conjunto total, indicando estabilidade do modelo. Os modelos ExtraTrees (ET) e SL MLP apresentaram os melhores desempenhos, com métricas equivalentes de $r = 0,71$, $R^2 = 0,49$ e RMSE = 1,88. Esse resultado reforça a capacidade desses modelos de capturar relações não lineares entre os descritores e ΔG.

No Benchmark, o modelo ElasticNet destacou-se, alcançando $r = 0,87$, $R^2 = 0,74$ e RMSE = 1,42, superando ligeiramente o desempenho da Regressão Linear Simples que apresentou o segundo melhor resultado geral nesse banco. Esse comportamento é coerente com as propriedades do ElasticNet, que combina regularização L1 e L2, promovendo a seleção de descritores mais relevantes

enquanto mitiga o sobreajuste (ZOU; HASTIE, 2005). No entanto, a principal limitação desse modelo é sua sensibilidade à escolha dos hiperparâmetros, o que pode comprometer a generalização para outros conjuntos de dados. O modelo ET também apresentou desempenho competitivo, sendo levemente superior ao SL MLP, que obteve r = 0,71, R^2 = 0,41 e RMSE = 2,12.

No SKEMPI, a maioria dos modelos individuais apresentou desempenhos baixos, com exceção do SVR e da Regressão Linear, que obtiveram valores razoáveis. A Regressão Linear foi o modelo individual mais eficaz, superando ligeiramente o SVR com métricas de r = 0,65, R^2 = 0,24 e RMSE = 1,99. Esse resultado corrobora a hipótese de que o SVR tende a se adaptar melhor a conjuntos menores devido à sua abordagem baseada em margens de suporte, que pode ser menos sensível à alta dimensionalidade quando há poucas amostras (SMOLA; SCHÖLKOPF, 2004). Entretanto, o SKEMPI contou com apenas 28 amostras no total (treinamento e teste), o que sugere que os resultados podem não ser estatisticamente representativos. Estudos com um volume maior de dados seriam necessários para validar essas observações. Entre os metamodelos avaliados no SKEMPI, o SL baseado na Regressão Linear apresentou as melhores métricas, com r = 0,54, R^2 = 0,28 e RMSE = 1,93, enquanto o SL MLP obteve resultados ligeiramente inferiores. Esses valores indicam que, para um conjunto de dados pequeno como o SKEMPI, um modelo mais simples como a regressão linear pode ser mais eficiente do que abordagens mais complexas, que podem sofrer com a escassez de amostras para um treinamento adequado.

Os resultados detalhados dos modelos mencionados podem ser consultados na Tabela 4, sendo os melhores por banco de dados destacados em negrito, enquanto a performance de todos os modelos está disponível no Apêndice D.

Tabela 4 – Desempenho dos principais modelos por base de dados, segundo as métricas RMSE, r e R^2.

	Teste set completo			**Pdbind**			**Benchmark**			**Skempi**		
Modelos	r	R^2	RMSE	r	R^2	RMSE	r	R^2	RMSE	r	R^2	RMSE
Extra Trees	**0.70**	**0.48**	**1.91**	**0.71**	**0.49**	**1.88**	0.75	0.45	2.04	0.36	0.05	2.23
SL MLP	**0.70**	**0.48**	**1.91**	**0.71**	**0.49**	**1.88**	0.71	0.41	2.12	0.47	0.16	2.09
SL Reg. Linear	0.62	0.38	2.09	0.62	0.38	2.07	0.54	0.28	2.35	0.54	0.28	1.93
SVR	0.15	0.02	2.62	0.10	0.00	2.63	0.38	-0.03	2.81	0.65	0.19	2.06

ElasticNet	0.59	0.34	2.14	0.56	0.30	2.19	**0.87**	**0.74**	**1.42**	0.61	-0.03	2.32
Reg. Linear	0.65	0.41	2.04	0.62	0.37	2.08	0.87	0.69	1.53	**0.65**	**0.24**	**1.99**

Fonte: O autor (2024)

Além disso, investigou-se se os dados do SKEMPI apresentavam características significativamente diferentes das demais bases analisadas. Para isso, foi treinado um modelo utilizando exclusivamente os dados do Benchmark e do PDBbind e, em seguida, esse modelo foi aplicado para prever as estruturas do SKEMPI. Os resultados mostraram um desempenho insatisfatório para a maioria dos métodos, com exceção do SVR e do XGBoost. O SVR obteve métricas razoáveis, com r = 0.77, R^2 = 0.29 e RMSE = 1.92. No entanto, o baixo valor de R^2 sugere que, embora o modelo tenha captado alguma tendência linear na relação entre os descritores e a energia livre de Gibbs, uma parte substancial da variabilidade nos dados do SKEMPI permaneceu sem explicação, indicando que os padrões identificados no Benchmark e no PDBbind não se transferem completamente para esta base.

O XGBoost também apresentou um desempenho relativamente melhor do que os demais modelos, mas ainda assim limitado, com r = 0.52, R^2 = 0.10 e RMSE = 2.17. O baixo coeficiente de determinação reforça a hipótese de que os dados do SKEMPI possuem características distintas, dificultando a generalização dos modelos treinados em outras bases. Além disso, o reduzido número de amostras do SKEMPI pode ter contribuído para a instabilidade dos resultados, uma vez que modelos de aprendizado de máquina tendem a sofrer com maior variabilidade estatística quando treinados ou avaliados em conjuntos pequenos. Dessa forma, a análise sugere que a baixa performance observada pode ser resultado tanto de diferenças intrínsecas entre os dados do SKEMPI e os das demais bases quanto da limitação imposta pelo pequeno volume amostral, tornando necessária uma maior quantidade de dados para validar essas observações com mais robustez

5.3 IMPACTO DA REMOÇÃO DE OUTLIERS NOS MODELOS

Ao treinar novamente os modelos sem os *outliers* para predizer os dados de teste, observou-se que, na maioria dos casos, o desempenho dos modelos foi equivalente ou inferior ao obtido com os dados contendo os *outliers*, conforme apresentado na Tabela 5. As únicas exceções foram os modelos Bagging e Decision Trees, que apresentaram um leve aumento no desempenho, embora ainda inferiores aos melhores resultados obtidos por outros modelos sem a remoção dos *outliers*. Além disso, todos os meta-modelos do SL demonstraram desempenho inferior quando treinados sem os *outliers*.

A remoção de *outliers* pode, em alguns casos, reduzir a variabilidade dos dados e melhorar a generalização dos modelos, mas também pode levar à perda de informações relevantes, especialmente em conjuntos de dados pequenos ou moderados, onde cada amostra contém características importantes para a modelagem da relação entre variáveis (GARCÍA; LUENGO; HERRERA, 2015). No contexto deste trabalho, a piora no desempenho dos modelos sem os *outliers* sugere que essas observações podem conter padrões relevantes para a predição de ΔG e que sua remoção pode ter reduzido a capacidade dos modelos de capturar a complexidade subjacente dos dados. Dado esse impacto negativo na performance geral, optou-se por manter os *outliers* no modelo final treinado, garantindo uma melhor capacidade preditiva. Todos os resultados detalhados podem ser encontrados no Apêndice E.

Tabela 5 – Comparação do desempenho dos principais modelos treinados com e sem *outliers* para predizer os dados de teste.

	Dados s/outliers			**Dados c/ outliers**		
Modelos	**r**	**R^2**	**RMSE**	**R**	**R^2**	**RMSE**
Extra Trees	0.67	0.43	1.99	0.70	0.48	1.91
SL MLP	0.68	0.45	1.96	0.70	0.48	1.91
SL Reg. Linear	0.61	0.37	2.10	0.62	0.38	2.09
SVR	0.16	0.02	2.62	0.15	0.02	2.62
ElasticNet	0.57	0.32	2.18	0.59	0.34	2.14
Reg. Linear	0.61	0.32	2.17	0.65	0.41	2.04

Fonte: O autor (2024)

5.4 IMPACTO DA REDUÇÃO DE DIMENSIONALIDADE NA PREDIÇÃO DOS MODELOS

5.4.1 PCA

Para evitar vazamento de dados e garantir a validade da análise, o PCA foi ajustado exclusivamente nos dados de treinamento, e os dados de teste foram posteriormente projetados no espaço definido pelos componentes principais extraídos, o resultado pode ser verificado na Tabela 6. Esse procedimento assegura que ambas as amostras compartilhem a mesma base ortonormal.

Apesar de permitir uma compactação eficiente dos dados, a redução dimensional levou à perda de informações relevantes para a modelagem, resultando em um desempenho inferior ao modelo que utilizava todas as variáveis. Isso sugere que as variáveis originais continham relações importantes que não foram preservadas nos componentes principais, tornando o PCA inviável para otimização do modelo.

5.4.2 UMAP

A aplicação do UMAP resultou em perda de desempenho em comparação aos modelos sem redução de dimensionalidade, indicando que o método não preservou integralmente a estrutura relevante dos dados para a predição de ΔG, o método leva os dados para um espaço de menor dimensão, preservando as relações locais segundo a teoria dos grafos. No entanto, os resultados demonstraram uma convergência consistente para o melhor desempenho quando o Super Learner (SL) utilizou o metamodelo MLP, validando a robustez da abordagem de combinação de modelos (VAN DER LAAN; POLLEY; HUBBARD, 2007).

Na Tabela 6, são apresentados os desempenhos das versões SL_MLP com diferentes níveis de redução dimensional: 4, 23 e 39 variáveis selecionadas pelos métodos UMAP associado ao modelo Elastic, ponto arbitrário intermediário escolhido e UMAP associado ao ET, além do SL_MLP_PCA para fins de comparação. Entre os dados selecionados pelo UMAP, curiosamente, a melhor performance foi obtida com 23 dimensões, um valor intermediário não diretamente otimizado pela métrica adotada na seleção das variáveis. Esse resultado sugere que

a relação entre o número de dimensões e a qualidade do modelo não segue uma tendência linear previsível. Além disso, levanta a hipótese de que o critério de otimização usado para a escolha do número de variáveis pode não capturar totalmente as características mais relevantes para o desempenho final do SL. No entanto, em todas as configurações, os modelos treinados com dados reduzidos via UMAP apresentaram desempenho inferior ao conjunto sem redução, ficando também abaixo do SL_MLP_PCA. Isso sugere que, para este conjunto de dados, o UMAP pode não ter sido a abordagem mais eficiente para preservar as informações críticas à predição de ΔG.

Tabela 6 – Desempenho do SL com o metamodelo MLP: comparação entre PCA, UMAP e modelo sem redução de dimensionalidade.

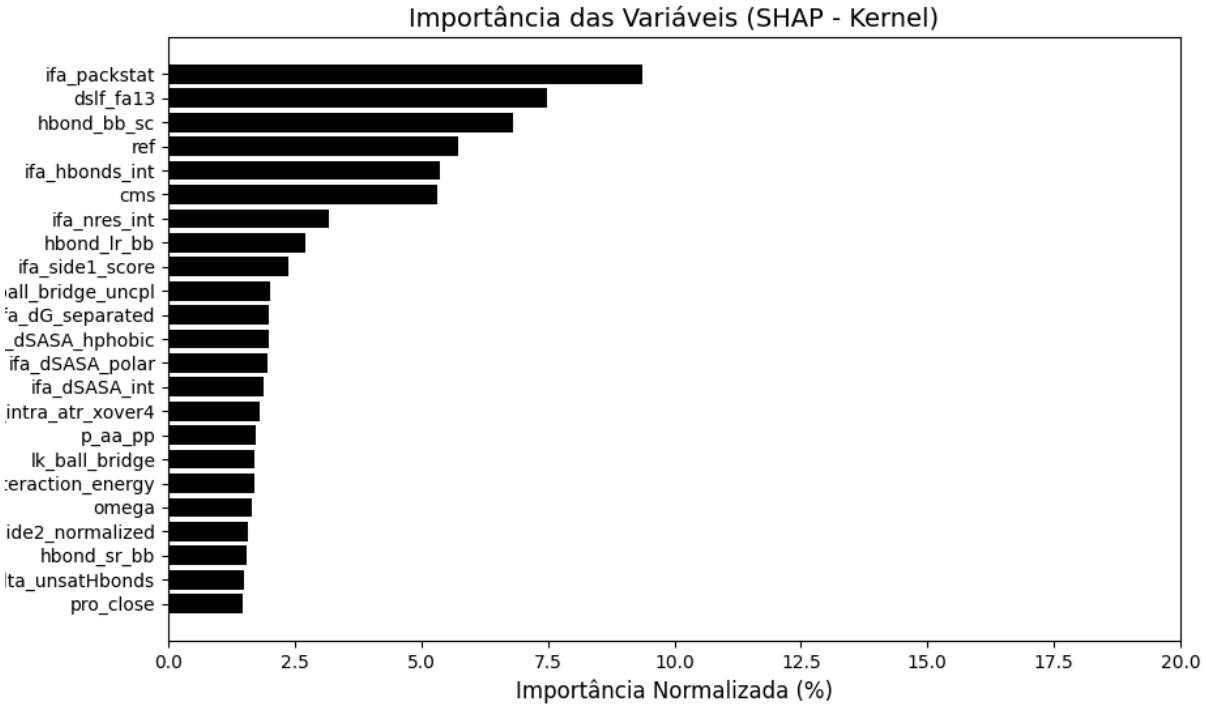
	Métricas de Avaliação		
Modelos	R	R^2	RMSE
SL_MLP_UMAP_4	0.47	0.17	2.41
SL_MLP_UMAP_23	0.57	0.30	2.21
SL_MLP_UMAP_39	0.54	0.25	2.28
SL_MLP_PCA	0.62	0.37	2.09
SL_MLP	**0.70**	**0.48**	**1.91**

Fonte: O autor (2025)

5.4.3 SHAP

Após normalização, as variáveis foram ordenadas por relevância e selecionadas até atingir 95% da importância total. Como resultado, 42 variáveis foram escolhidas, sendo as 23 mais relevantes apresentadas na Figura 10. A importância de todas as variáveis pode ser consultada no Apêndice F.

Figura 10 – Gráfico contendo as 23 principais variáveis e sua respectiva importância (%) segundo a análise SHAP aplicada ao modelo Extra Trees.



Fonte: O autor (2024)

As 42 variáveis selecionadas, foram então utilizadas para treinar o SL e posteriormente validar com os dados de treinamento, o modelo apresentou desempenho levemente inferior ao do melhor modelo obtido até o momento (SL_MLP), esse resultado é interessante pois valida o método SHAP como classificador de variáveis importantes, porém como método de redução de dimensionalidade ao avaliar o desempenho por banco de dados, observou-se uma queda de rendimento em relação ao modelo sem redução, optando-se nesse estudo por manter todas as variáveis conforme pode ser visualizado na Tabela 7.

Tabela 7 – Comparação entre o SL_MLP_SHAP_42 e SL_MLP por banco de dados.

	Teste set completo			**Pdbbind**			**Benchmark**			**Skempi**		
Modelos	**r**	**R²**	**RMSE**	**R**	**R²**	**RMSE**	**R**	**R²**	**RMSE**	**r**	**R²**	**RMSE**
SL_MLP	0.70	0.48	1.91	0.71	0.49	1.88	0.71	0.41	2.12	0.47	0.16	2.09
SL_MLP_SHAP_42	0.69	0.48	1.91	0.71	0.50	1.86	0.66	0.40	2.14	0.24	-0.04	2.33

Fonte: O autor (2025)

5.5 INVESTIGAÇÃO DA CORRELAÇÃO ENTRE OS MÉTODOS DE SELEÇÃO SHAP E UMAP

Para investigar a relação entre a seleção de variáveis pelo SHAP e a redução dimensional pelo UMAP, analisaram-se três variações do SHAP: SHAP_4, SHAP_23 e SHAP_39, correspondendo, respectivamente, à otimização via Elastic Net, a um ponto intermediário entre os valores obtidos e à otimização via Extra Trees (ET) para R^2. Todas essas seleções apresentaram desempenho superior às respectivas reduções dimensionais pelo UMAP, sugerindo a ausência de uma correlação direta entre as variáveis selecionadas e a projeção gerada pelo UMAP conforme pode ser visualizado na Tabela 8.

Tabela 8 – Comparação do desempenho do SL entre os métodos de redução de dimensionalidade UMAP e SHAP.

	Métricas de Avaliação						
Modelo	**r**	**R^2**	**RMSE**	**R**	**R^2**	**RMSE**	**Modelo**
SL_MLP_UMAP_4	0.47	0.17	2.41	0.55	0.28	2.25	SL_MLP_SHAP_4
SL_MLP_UMAP_23	0.57	0.30	2.21	0.68	0.46	1.94	SL_MLP_SHAP_23
SL_MLP_UMAP_39	0.54	0.25	2.28	0.70	0.48	1.90	SL_MLP_SHAP_39

Fonte: O autor (2025)

Um resultado interessante foi observado no SL_MLP_SHAP_23, cujo desempenho foi ligeiramente inferior ao do SL_MLP_SHAP_39. Surpreendentemente, este último apresentou um desempenho equivalente ao SL_MLP considerando o conjunto completo de variáveis, implicando uma simplificação significativa do modelo sem perda substancial de precisão. No entanto, ao avaliar os resultados por banco de dados, verificou-se que o SL_MLP_SHAP_39 teve desempenho inferior ao SL_MLP, razão pela qual optou-se por manter todas as variáveis neste estudo. Esses resultados estão apresentados na Tabela 9.

Tabela 9 - Desempenho comparativo do SL_MLP_SHAP_39 e SL_MLP por banco de dados.

	Teste set completo			**Pdbbind**			**Benchmark**			**Skempi**		
Modelos	**r**	**R^2**	**RMSE**	**r**	**R^2**	**RMSE**	**r**	**R^2**	**RMSE**	**r**	**R^2**	**RMSE**
SL_MLP	0.70	0.48	1.91	0.71	0.49	1.88	0.71	0.41	2.12	0.47	0.16	2.09
SL_MLP_SHAP_39	0.70	0.48	1.90	0.70	0.50	1.85	0.65	0.38	2.17	0.30	0.01	2.27

Fonte: O autor (2025)

6 CONCLUSÃO

Com o auxílio desse trabalho, foi desenvolvido um modelo de aprendizado de máquina baseado em *Super Learner* (SL) para a predição de ΔG de interações proteína-proteína, utilizando descritores gerados pelo software Rosetta. Embora o Rosetta não seja especificamente adequado para o cálculo de ΔG, seus descritores já demonstraram relevância em estudos anteriores. A metodologia empregada permitiu a análise de diferentes abordagens de modelagem, a avaliação de técnicas de redução de dimensionalidade e a de remoção de *outliers*.

Os resultados indicaram que, apesar da forte correlação entre diversos pares de descritores, a aplicação de métodos de redução de dimensionalidade, como PCA e UMAP, não promoveram a melhora no desempenho dos modelos. O método SHAP, acoplado ao Extra Trees (ET), permitiu reduzir o número de variáveis de 49 para 42 sem perda significativa de desempenho, porém com desempenho ligeiramente inferior quando comparado às métricas considerando-se a segmentação por banco de dados. Além disso, curiosamente, a seleção arbitrária de 23 variáveis pelo SHAP resultou em um modelo mais interpretável, embora com um leve decréscimo no rendimento preditivo em relação ao modelo SL_MLP_SHAP_42. Foi constatado também que as variáveis selecionadas pelo SHAP tiveram desempenho superior às selecionadas pelo UMAP, indicando que não há correlação entre os métodos para a natureza dos dados e os descritores utilizados.

Os modelos concorrentes, como "Area Affinity" e PRODIGY, apresentaram desempenhos insatisfatórios, com R^2 negativo no conjunto de teste, indicando predições inferiores à simples média dos dados, com desempenho muito aquém do reportado na literatura. Em contrapartida, a Regressão Linear Simples obteve resultados competitivos ($r = 0,65$, $R^2 = 0,41$, RMSE = 2,04), sugerindo uma estrutura bem definida quando utilizado o conjunto completo de dados de teste, sem segmentação por banco.

Dentre os modelos testados, o Extra Trees (ET) apresentou o melhor desempenho geral, especialmente quando os dados foram segmentados no banco PDBbind, sendo equivalente ao modelo SL_MLP. No banco Benchmark, o modelo Elastic Net obteve os melhores resultados ($r = 0,87$, RMSE = 1,42), indicando uma possível relação linear entre os dados desse banco. No banco SKEMPI, os modelos SVR e SL_MLP se destacaram, mas a baixa quantidade de dados dificultou uma

análise mais profunda, sugerindo que dados oriundos de mutações podem apresentar um comportamento distinto na predição de ΔG, sendo necessário reproduzir os estudos com uma quantidade superior de dados desse banco.

A remoção de *outliers* não resultou em melhorias significativas no desempenho dos modelos e, em alguns casos, levou a uma piora na capacidade preditiva, reforçando a importância dessas observações na modelagem do fenômeno. Com base nesses resultados, optou-se por manter os *outliers* no modelo final.

Por fim, os resultados demonstraram que a redução de dimensionalidade não foi benéfica para a melhora da performance preditiva e que a remoção de *outliers* pode comprometer informações importantes. O modelo SL_MLP foi estabelecido como o modelo final de predição, sem redução de dimensionalidade e sem remoção de *outliers*.

Um ponto relevante foi a convergência entre as variáveis selecionadas pelo ET e pelo SHAP, indicando robustez no processo de seleção de variáveis. Das 23 melhores variáveis selecionadas pelo ET, 19 coincidiram com as selecionadas pelo SHAP, sugerindo que tais abordagens podem ser exploradas em estudos futuros para aprimorar a modelagem de ΔG em interações proteína-proteína e melhorar a interpretabilidade dos modelos.

A análise bayesiana realizada sobre os resíduos dos modelos no conjunto de teste para os modelos individuais permitiu investigar a presença de viés sistemático de forma quantitativa. A estimação da média posterior (μ) e da probabilidade $P(|\mu|>0)$ revelou que todos os modelos apresentaram algum grau de viés, com destaque para o AdaBoost Regressor, que exibiu o maior viés positivo (μ = 0,3261), seguido por XGBoost e Bagging Regressor. O único modelo a apresentar viés negativo foi o SVR (μ = -0,0278), embora com magnitude considerada irrelevante do ponto de vista prático. Esses resultados destacam a importância de avaliar não apenas o desempenho preditivo global, mas também as tendências sistemáticas que podem comprometer a acurácia em aplicações sensíveis. A aplicação da inferência bayesiana, mesmo sob violação da normalidade dos resíduos, foi justificada pela robustez do modelo e pela estabilidade dos parâmetros estimados, contribuindo significativamente para o diagnóstico e a escolha criteriosa dos modelos mais confiáveis.

Além disso, os avanços obtidos neste trabalho resultaram na publicação do artigo intitulado *"Estimating Absolute Protein–Protein Binding Free Energies by a Super Learner Model*" na revista Journal of Chemical Information and Modeling, o que evidencia a originalidade e relevância científica da abordagem proposta, especialmente no contexto internacional da bioinformática estrutural e da predição de afinidade molecular.

REFERÊNCIAS

ACHEN, C. H. Measuring Representation: Perils of the Correlation Coefficient. **American Journal of Political Science**, v. 21, n. 4, p. 805, nov. 1977.

ADOLF-BRYFOGLE, Jared et al. Growing glycans in Rosetta: Accurate de novo glycan modeling, density fitting, and rational sequon design. bioRxiv, 2021. Disponível em: https://doi.org/10.1101/2021.09.27.462000.

AGIUS, R. et al. Characterizing Changes in the Rate of Protein-Protein Dissociation upon Interface Mutation Using Hotspot Energy and Organization. **PLoS Computational Biology**, v. 9, n. 9, p. e1003216, 5 set. 2013.

ALTMAN, N. S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. **The American Statistician**, v. 46, n. 3, p. 175–185, ago. 1992.

ARKIN, M. R.; WELLS, J. A. Small-molecule inhibitors of protein–protein interactions: progressing towards the dream. **Nature Reviews Drug Discovery**, v. 3, n. 4, p. 301–317, abr. 2004.

AKBARZADEH, S., Coşkun, Ö., & Günçer, B. (2024). Studying protein–protein interactions: Latest and most popular approaches. *Journal of Structural Biology*, *216*(4), 108118. https://doi.org/10.1016/j.jsb.2024.108118

BAGCHI, A. Protein-protein Interactions: Basics, Characteristics, and Predictions. Em: **Soft Computing for Biological Systems**. Singapore: Springer Singapore, 2018. p. 111–120.

BEYER, K. et al. When Is “Nearest Neighbor” Meaningful? Em: [s.l: s.n.]. p. 217–235.

BOGAN, A. A.; THORN, K. S. Anatomy of hot spots in protein interfaces. **Journal of Molecular Biology**, v. 280, n. 1, p. 1–9, jul. 1998.

BREIMAN, L. Bagging Predictors. **Machine Learning**, v. 24, n. 2, p. 123–140, ago. 1996.

BREIMAN, L. Pasting Small Votes for Classification in Large Databases and On-Line. **Machine Learning**, v. 36, n. 1/2, p. 85–103, jul. 1999.

BREIMAN, L. Random Forests. Machine Learning. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

BUXBAUM, E. **Fundamentals of Protein Structure and Function**. Cham: Springer International Publishing, 2015.

CHAI, T.; DRAXLER, R. R. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. **Geoscientific Model Development**, v. 7, n. 3, p. 1247–1250, 30 jun. 2014.

CHAVES, E. et al. **Prediction of Absolute Protein-Protein Binding Free Energy by a Super Learner Model**. , 14 ago. 2023.

CHEN, T.; GUESTRIN, C. **XGBoost**. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. **Anais**...New York, NY, USA: ACM, 13 ago. 2016.

CHO, K.; KIM, D.; LEE, D. A feature-based approach to modeling protein–protein interaction hot spots. **Nucleic Acids Research**, v. 37, n. 8, p. 2672–2687, maio 2009.

CLACKSON, T.; WELLS, J. A. A Hot Spot of Binding Energy in a Hormone-Receptor Interface. **Science**, v. 267, n. 5196, p. 383–386, 20 jan. 1995.

CLARK, M. A.; CHOI, J.; DOUGLAS, M. **Biology 2e**. [s.l: s.n.].

COVER, T.; HART, P. Nearest neighbor pattern classification. **IEEE Transactions on Information Theory**, v. 13, n. 1, p. 21–27, jan. 1967.

DAVID, A. et al. Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. **Human Mutation**, v. 33, n. 2, p. 359–363, fev. 2012.

DE LAS RIVAS, J.; FONTANILLO, C. Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks. **PLoS Computational Biology**, v. 6, n. 6, p. e1000807, 24 jun. 2010.

DE LAS RIVAS, J., & FONTANILLO, C. (2012). Protein-protein interaction networks: unraveling the wiring of molecular machines within the cell. *Briefings in Functional Genomics*, *11*(6), 489–496. https://doi.org/10.1093/bfgp/els036

DEMCHENKO, A. P. Recognition between flexible protein molecules: induced and assisted folding. **Journal of Molecular Recognition**, v. 14, n. 1, p. 42–61, jan. 2001.

D'IMPRIMA, E.; KÜHLBRANDT, W. Current limitations to high-resolution structure determination by single-particle cryoEM. **Quarterly Reviews of Biophysics**, v. 54, p. e4, 11 mar. 2021.

DRUCKER, H. Improving Regressors Using Boosting Techniques. **Proceedings of the 14th International Conference on Machine Learning**, ago. 1997.

DU, X. et al. Insights into Protein–Ligand Interactions: Mechanisms, Models, and Methods. **International Journal of Molecular Sciences**, v. 17, n. 2, p. 144, 26 jan. 2016.

DURHAM, J., ZHANG, J., HUMPHREYS, I. R., PEI, J., & CONG, Q. (2023). Recent advances in predicting and modeling protein–protein interactions. *Trends in Biochemical Sciences*, *48*(6), 527–538. https://doi.org/10.1016/j.tibs.2023.03.003

FERRAZ, M. 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.

FIX, E.; HODGES, J. L. Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties. Technical Report. **USAF School of Aviation Medicine**, 1951.

FLEISHMAN SJ, LEAVER-FAY A, CORN JE, STRAUCH EM, KHARE SD, et al. (2011) RosettaScripts: A Scripting Language Interface to the Rosetta Macromolecular Modeling Suite. PLOS ONE 6(6): e20161. https://doi.org/10.1371/journal.pone.0020161

FREUND, Y.; SCHAPIRE, R. E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. **Journal of Computer and System Sciences**, v. 55, n. 1, p. 119–139, ago. 1997.

GARCÍA, S.; LUENGO, J.; HERRERA, F. **Data Preprocessing in Data Mining**. Cham: Springer International Publishing, 2015. v. 72

GARRETT, R. H.; GRISHAM, C. M. **biochemistry**. 6ª edition ed. [s.l: s.n.].

GENG, C. et al. Finding the $\Delta\Delta G$ spot: Are predictors of binding affinity changes upon mutations in protein–protein interactions ready for it? **WIREs Computational Molecular Science**, v. 9, n. 5, 15 set. 2019.

GÉRON, A. **Hands-On Machine Learning with scikit-learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems**. 3rd edition ed. [s.l.] O'Reilly Media, 2022.

GEURTS, P.; ERNST, D.; WEHENKEL, L. Extremely randomized trees. **Machine Learning**, v. 63, n. 1, p. 3–42, 2 abr. 2006.

GOLAS, S. M. et al. Gibbs free energy of protein-protein interactions correlates with ATP production in cancer cells. **Journal of Biological Physics**, v. 45, n. 4, p. 423–430, 16 dez. 2019.

GROSDIDIER, S.; FERNÁNDEZ-RECIO, J. Identification of hot-spot residues in protein-protein interactions by computational docking. **BMC Bioinformatics**, v. 9, n. 1, p. 447, 21 dez. 2008.

GUEST, J. D. et al. An expanded benchmark for antibody-antigen docking and affinity prediction reveals insights into antibody recognition determinants. **Structure**, v. 29, n. 6, p. 606- 621.e5, jun. 2021.

GUO, Z.; YAMAGUCHI, R. Machine learning methods for protein-protein binding affinity prediction in protein design. **Frontiers in Bioinformatics**, v. 2, 16 dez. 2022.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning**. New York, NY: Springer New York, 2009.

JAMES, G. et al. **An Introduction to Statistical Learning**. New York, NY: Springer US, 2021a.

JAMES, G. et al. **An Introduction to Statistical Learning**. New York, NY: Springer US, 2021b.

JANIN, J. Protein-protein recognition. **Progress in Biophysics and Molecular Biology**, v. 64, n. 2–3, p. 145–166, 1995.

JANKAUSKAITĖ, J. et al. SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. **Bioinformatics**, v. 35, n. 3, p. 462–469, 1 fev. 2019.

KASTRITIS, P. L. et al. A structure-based benchmark for protein–protein binding affinity. **Protein Science**, v. 20, n. 3, p. 482–491, 16 mar. 2011.

KE, G. et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. **Neural Information Processing Systems**, 2017.

KEEGAN, M. et al. Gibbs Free Energy, a Thermodynamic Measure of Protein–Protein Interactions, Correlates with Neurologic Disability. **BioMedInformatics**, v. 1, n. 3, p. 201–210, 14 dez. 2021.

KENNEDY, PETER. **A Guide to Econometrics**. sixth edition ed. [s.l: s.n.].

KESKIN, O.; MA, B.; NUSSINOV, R. Hot Regions in Protein–Protein Interactions: The Organization and Contribution of Structurally Conserved Hot Spot Residues. **Journal of Molecular Biology**, v. 345, n. 5, p. 1281–1294, fev. 2005.

KING, G. How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science. **American Journal of Political Science**, v. 30, n. 3, p. 666, ago. 1986.

KOHAVI, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. mar. 2001.

L. GARNER, A.; D. JANDA, K. Protein-Protein Interactions and Cancer: Targeting the Central Dogma. **Current Topics in Medicinal Chemistry**, v. 11, n. 3, p. 258–280, 1 fev. 2011.

LEMAN, J.K., WEITZNER, B.D., LEWIS, S.M. et al. Macromolecular modeling and design in Rosetta: recent methods and frameworks. Nat Methods 17, 665–680 (2020). https://doi.org/10.1038/s41592-020-0848-2

LUNDBERG, S.; LEE, S.-I. A Unified Approach to Interpreting Model Predictions. 22 maio 2017.

LU, H., ZHOU, Q., HE, J., JIANG, Z., PENG, C., TONG, R., & SHI, J. (2020). Recent advances in the development of protein–protein interactions modulators: mechanisms and clinical trials. *Signal Transduction and Targeted Therapy*, *5*(1), 213. https://doi.org/10.1038/s41392-020-00315-3

MOORE, D. S. , M. G. P. , & C. B. A. **Introduction to the practice of statistics**. Eighth edition ed. New York: Macmillan Higher Education Company, 2014. v. 1

MOREIRA, I. S.; FERNANDES, P. A.; RAMOS, M. J. Computational alanine scanning mutagenesis—An improved methodological approach. **Journal of Computational Chemistry**, v. 28, n. 3, p. 644–654, 28 fev. 2007a.

MOREIRA, I. S.; FERNANDES, P. A.; RAMOS, M. J. Hot spots—A review of the protein–protein interface determinant amino-acid residues. **Proteins: Structure, Function, and Bioinformatics**, v. 68, n. 4, p. 803–812, set. 2007b.

NISHI, H.; HASHIMOTO, K.; PANCHENKO, A. R. Phosphorylation in Protein-Protein Binding: Effect on Stability and Function. **Structure**, v. 19, n. 12, p. 1807–1815, dez. 2011.

NOHARA, Y. et al. Explanation of machine learning models using shapley additive explanation and application for real data in hospital. **Computer Methods and Programs in Biomedicine**, v. 214, p. 106584, fev. 2022.

PARANHOS, R. et al. Desvendando os Mistérios do Coeficiente de Correlação de Pearson: o Retorno. **Leviathan (São Paulo)**, n. 8, p. 66, 13 ago. 2014.

PEROZZO, R.; FOLKERS, G.; SCAPOZZA, L. Thermodynamics of Protein–Ligand Interactions: History, Presence, and Future Aspects. **Journal of Receptors and Signal Transduction**, v. 24, n. 1–2, p. 1–52, 23 jan. 2004.

PETER Y, C. P. M. P. **Correlation : parametric and nonparametric measures**. [s.l: s.n.].

PLATT, J. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. **Adv. Large Margin Classif.**, jun. 2000.

RAO, V. S., SRINIVAS, K., SUJINI, G. N., & KUMAR, G. N. S. (2014). Protein-Protein Interaction Detection: Methods and Analysis. *International Journal of Proteomics*, *2014*, 1–12. https://doi.org/10.1155/2014/147648

RODA, A. Biotechnological applications of bioluminescence and chemiluminescence. **Trends in Biotechnology**, v. 22, n. 6, p. 295–303, jun. 2004.

SEDOV, I. A., & ZUEV, Y. F. (2023). Recent Advances in Protein–Protein Interactions. *International Journal of Molecular Sciences*, *24*(2), 1282. https://doi.org/10.3390/ijms24021282

SHIN, W.-H., KUMAZAWA, K., IMAI, K., HIROKAWA, T., & KIHARA, D. (2020). Current Challenges and Opportunities in Designing Protein–Protein Interaction Targeted Drugs. *Advances and Applications in Bioinformatics and Chemistry*, *Volume 13*, 11–25. https://doi.org/10.2147/AABC.S235542

SHRINGARI, S. R. et al. Rosetta custom score functions accurately predict ΔΔ *G* of mutations at protein–protein interfaces using machine learning. **Chemical Communications**, v. 56, n. 50, p. 6774–6777, 2020.

SMOLA, A. J.; SCHÖLKOPF, B. A tutorial on support vector regression. **Statistics and Computing**, v. 14, n. 3, p. 199–222, ago. 2004.

SOLEYMANI, F. et al. Protein–protein interaction prediction with deep learning: A comprehensive review. **Computational and Structural Biotechnology Journal**, v. 20, p. 5316–5341, 2022.

STIGLER, S. M. Gauss and the Invention of Least Squares. **The Annals of Statistics**, v. 9, n. 3, 1 maio 1981.

THANGARATNARAJAH, C.; RHEINBERGER, J.; PAULINO, C. Cryo-EM studies of membrane proteins at 200 keV. **Current Opinion in Structural Biology**, v. 76, p. 102440, out. 2022.

VAN DER LAAN, M. J.; POLLEY, E. C.; HUBBARD, A. E. Super Learner. **Statistical Applications in Genetics and Molecular Biology**, v. 6, n. 1, 16 jan. 2007.

VREVEN, T. et al. Updates to the Integrated Protein–Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. **Journal of Molecular Biology**, v. 427, n. 19, p. 3031–3041, set. 2015.

WALTER, M. et al. Visualization of protein interactions in living plant cells using bimolecular fluorescence complementation. **The Plant Journal**, v. 40, n. 3, p. 428–438, 14 nov. 2004.

WANG, R. et al. The PDBbind Database: Collection of Binding Affinities for Protein−Ligand Complexes with Known Three-Dimensional Structures. **Journal of Medicinal Chemistry**, v. 47, n. 12, p. 2977–2980, 1 jun. 2004.

XUE, L. C. et al. PRODIGY: a web server for predicting the binding affinity of protein–protein complexes. **Bioinformatics**, v. 32, n. 23, p. 3676–3678, 1 dez. 2016.

YANG, Y. X. et al. *AREA-AFFINITY* : A Web Server for Machine Learning-Based Prediction of Protein–Protein and Antibody–Protein Antigen Binding Affinities. **Journal of Chemical Information and Modeling**, v. 63, n. 11, p. 3230–3237, 12 jun. 2023.

YIP, K. M. et al. Atomic-resolution protein structure determination by cryo-EM. **Nature**, v. 587, n. 7832, p. 157–161, 5 nov. 2020.

ZHANG, C.; LAI, L. Automatch: Target-binding protein design and enzyme design by automatic pinpointing potential active sites in available protein scaffolds. **Proteins: Structure, Function, and Bioinformatics**, v. 80, n. 4, p. 1078–1094, 7 abr. 2012.

ZHANG, F. et al. PROBselect: accurate prediction of protein-binding residues from proteins sequences via dynamic predictor selection. **Bioinformatics**, v. 36, n. Supplement_2, p. i735–i744, 30 dez. 2020a.

ZHANG, F. et al. PROBselect: accurate prediction of protein-binding residues from proteins sequences via dynamic predictor selection. **Bioinformatics**, v. 36, n. Supplement_2, p. i735–i744, 30 dez. 2020b.

ZOU, H.; HASTIE, T. Regularization and Variable Selection Via the Elastic Net. **Journal of the Royal Statistical Society Series B: Statistical Methodology**, v. 67, n. 2, p. 301–320, 1 abr. 2005.

APÊNDICE A – LISTA DE ESTRUTURAS DE COMPLEXOS PROTEICOS COM RESOLUÇÃO CRISTALOGRÁFICA

Índice	ID PDB	Resolução (Å)	Descrição
1	3TSR	2,20	Estrutura de Raios-X do Inibidor de Ribonuclease de Camundongo Complexado com Ribonuclease de Camundongo 1
2	3U43	1.72	Estrutura Cristalina do Complexo Colicina E2 DNase-Im2
3	1B27	2,10	Resposta Estrutural à Mutação em uma Interface Proteína-Proteína
4	1EMV	1,70	Estrutura Cristalina do Domínio DNase da Colicina E9 com Sua Proteína de Imunidade Cognata Im9 (1.7 Å)
5	1DFJ	2.5	Inibidor de Ribonuclease Complexado com Ribonuclease A
6	3FP6	1.49	Tripsina Aniônica em Complexo com Inibidor de Tripsina Pancreática Bovino (BPTI) Determinada com Resolução de 1.49 Å
7	2VLP	2.00	Mutante R54A do Domínio DNase E9 em Complexo com Im9
8	2VLO	1.80	Mutante K97A do Domínio DNase E9 em Complexo com Im9
9	1MXE	1.70	Estrutura do Complexo de Calmodulina com a Sequência Alvo de CamK1
10	4Q5U	1.95	Estrutura da Calmodulina Ligada ao Seu Local de Reconhecimento da Calcineurina
11	3QHY	2.06	Análise Estrutural, Termodinâmica e Cinética da Interação de Alta Afinidade do Inibidor de Beta-Lactamase Proteína-II (BLIP-II) com Beta-Lactamases Classe A
12	2VLN	1.60	Mutante N75A do Domínio DNase E9 em Complexo com Im9
13	1LW6	1.50	Estrutura Cristalina do Complexo de Subtilisina BPN' com o Inibidor de Quimotripsina 2 a 1.5 Å de Resolução
14	3LNZ	1.95	A proteína 3LNZ é a estrutura cristalina do MDM2 humano complexado com um inibidor peptídico PMI (mutante N8A), envolvido na interação p53-MDM2/MDMX
15	1TM1	1.70	Estrutura Cristalina do Complexo de Subtilisina BPN' com o Inibidor de Quimotripsina 2
16	2O3B	2.30	Estrutura Cristalina do Complexo de Nuclease A (NucA) com o Inibidor Intracelular NuiA
17	1TM7	1.59	Estrutura Cristalina do Complexo de Subtilisina BPN' com o Mutante M59Y do Inibidor de Quimotripsina 2
18	1JIW	1.74	Estrutura Cristalina do Complexo Apr-Aprin
19	4A94	1.70	Estrutura do Inibidor de Carboxipeptidase de Nerita Versicolor em Complexo com CPA4 Humana
20	6KBR	2.00	Estrutura Cristalina do Complexo do Inibidor Derivado de SPINK2 com KLK4 Humano

21	1BVN	2.50	Alfa-Amilase Pancreática Suína em Complexo com o Inibidor Proteico Tendamistat
22	4M5F	2.50	Estrutura do Complexo Tse3-Tsi3
23	1TM5	1.45	Estrutura Cristalina do Complexo de Subtilisina BPN' com o Mutante M59A do Inibidor de Quimotripsina 2
24	3SGB	1.80	Estrutura do Complexo da Protease B de Streptomyces Griseus com o Terceiro Domínio do Inibidor de Ovomucoide de Peru a 1.8 Å de Resolução
25	1TM3	1.57	Estrutura Cristalina do Complexo de Subtilisina BPN' com o Mutante M59K do Inibidor de Quimotripsina 2
26	2XTT	0.93	Tripsina Bovina em Complexo com o Inibidor de Protease de Schistocerca Gregaria Aprimorado Evolutivamente (SGPI-1-P02)
27	1TAW	1.80	Tripsina Bovina Complexada com APPI
28	2VLQ	1.60	Mutante F86A do Domínio DNase E9 em Complexo com Im9
29	1Y3C	1.69	Estrutura Cristalina do Complexo de Subtilisina BPN' com o Mutante R62A do Inibidor de Quimotripsina 2
30	1JTD	2.3	Estrutura Cristalina do Inibidor de Beta-Lactamase Proteína-II em Complexo com a Beta-Lactamase TEM-1
31	1SGN	1.80	Variante do inibidor de ovo-mucoíde da peru, domínio terceiro complexado com proteinase b de *streptomyces griseus*
32	3ZRZ	1.70	Estrutura cristalográfica dos módulos F1 da fibronectina em complexo com um fragmento de Streptococcus pyogenes Sfbi-5
33	2PSM	2.19	Estrutura cristalográfica da interleucina 15 em complexo com o receptor alfa da interleucina 15
34	4KJY	1.93	Complexo da variante de alta afinidade SIRP alfa FD6 com CD47
35	3GMW	2.10	Estrutura cristalográfica da proteína inibidora de beta-lactamase I (BLIP-I) em complexo com beta-lactamase TEM-1
36	3VV2	1.83	Estrutura cristalográfica do complexo entre S324A-subtilisina e o mutante TkPro
37	6HAR	1.50	Estrutura cristalográfica da mesotripsina em complexo com APPi-M17C/I18F/F34C
38	2SIC	1.80	Estrutura refinada do complexo entre subtilisina BPN' e o inibidor de subtilisina de Streptomyces com 1.8 Å de resolução
39	5C67	1.83	Mesotripsina humana em complexo com a variante inibidora APPi-M17G/I18F/F34V da proteína precursora de amiloide
40	1ZVY	1.63	Estrutura cristalográfica do VHH D3-L11 em complexo com a lisozima da clara de ovo
41	2SGQ	1.80	Variante do inibidor de ovomucóide de peru (GLN 18) complexada com a proteína B de Streptomyces griseus, cristalizada a pH 6,5, com mutações e expressa em Escherichia coli.

42	1B3S	2.39	Resposta estrutural à mutação em uma interface proteína-proteína
43	3FPU	1.76	Estrutura cristalográfica do complexo entre Evasin-1 e CCL3
44	6JB8	1.65	Estrutura cristalográfica do nanocorpo D3-L11 em complexo com a lisozima da clara de ovo
45	1TM4	1.70	Estrutura cristalográfica do complexo entre subtilisina BPN' e o inibidor de quimotripsina 2 mutante M59G
46	1B2S	1.82	Resposta estrutural à mutação em uma interface proteína-proteína
47	3M18	1.95	Estrutura cristalográfica do receptor linfocitário variável VLRa.R2.1 em complexo com a lisozima da clara de ovo
48	2P4A	1.9	Estrutura cristalográfica de um fragmento de anticorpo VHH camelídeo maturado por afinidade em complexo com RNase A
49	6NDZ	1.90	Proteína repetitiva projetada em complexo com FZ8
50	6DWH	2.00	Estrutura cristalográfica do complexo entre BBKI e tripsina bovina
51	4OYD	1.80	Estrutura cristalina de um inibidor projetado computacionalmente para uma proteína Bcl-2 viral de Epstein-Barr
52	1AVA	1.90	Complexo proteína-proteína Amy2/Basi da semente de cevada
53	1TMG	1.67	Estrutura cristalina do complexo entre subtilisina BPN' e o inibidor de quimotripsina M59F mutante, obtida por difração de raios X, com mutações e expressa em Bacillus subtilis e Escherichia coli BL21(DE3).
54	4LGP	2.40	Cadeia A da ricina ligada a um nanocorpo camelídeo (VHH1)
55	4ML7	1.80	Estrutura cristalina de Brucella abortus Plic em complexo com lisozima humana
56	1Y1K	1.56	Estrutura cristalina do complexo de Subtilisina BPN' com o inibidor de quimotripsina 2 mutante T58A
57	1PXV	1.80	Complexo Stafostatina-Stafopaina: um inibidor de ligação direta em complexo com sua protease alvo de cisteína
58	1Y3B	1.80	Estrutura cristalina do complexo de Subtilisina BPN' com o inibidor de quimotripsina 2 mutante E60S
59	5U4M	2.50	RTA-V1C7-G29R sem sal
60	1TO1	1.68	Estrutura cristalina do complexo de Subtilisina BPN' com o inibidor de quimotripsina 2 mutante Y61A
61	6GFI	2.30	Estrutura da mesotripsina humana em complexo com a variante APPI T11V/M17R/I18F/F34V
62	4BQD	2.48	KD1 do TFPI humano em complexo com um peptídeo sintético
63	5AYS	2.09	Estrutura cristalina do complexo Saugi/HSV UDG
64	1Y34	1.55	Estrutura cristalina do complexo de Subtilisina BPN' com o inibidor de quimotripsina 2 mutante E60A
65	1Y3D	1.80	Estrutura cristalina do complexo de Subtilisina BPN' com o inibidor de quimotripsina 2 mutante R67A

66	4K5B	1.85	Co-cristalização com proteínas de repetição anquirina projetadas para explicar a flexibilidade conformacional de Bcl-W
67	4GI3	1.75	Estrutura cristalina de Greglin em complexo com Subtilisina
68	1OP9	2.20	Complexo de lisozima humana com fragmento de anticorpo VHH HL6 de camelídeo
69	2GOX	2.20	Estrutura cristalina do complexo Efb-C / C3d
70	6IU7	1.90	Estrutura cristalina da Importina-Alfa1 ligada ao sinal de localização nuclear 53BP1 (wild-type)
71	1AN1	2.03	Complexo inibidor de triptase derivado de sanguessuga/tripsina
72	4NSO	2.40	Estrutura cristalina do complexo proteína efetora-imunidade
73	4C7N	2.10	Estrutura cristalina do peptídeo sintético IM10 em complexo com a região de hélice enrolada de MITF
74	3K1R	2.30	Estrutura da Harmonina NPDZ1 em complexo com o SAM-PBM de SANS
75	2VOH	1.90	Estrutura de Mouse A1 ligada ao domínio Bak BH3
76	1CLV	2.00	Amilase alfa do besouro Tenebrio molitor em complexo com o inibidor de amilase alfa da amaranto
77	4EIG	2.50	Fragmento de anticorpo camelídeo CA1698 em complexo com DHFR
78	2I25	1.8	Análise estrutural cristalina do domínio variável PBLA8 do novo receptor de antígeno do tubarão-enfermeira em complexo com lisozima
79	3N4I	1.56	Estrutura cristalina do complexo SHV-1 D104E beta-lactamase/inibidor de beta-lactamase (BLIP)
80	2VOI	2.10	Estrutura de Mouse A1 ligada ao domínio Bid BH3
81	3WDG	2.20	Complexo UDG/Ugi de Staphylococcus aureus
82	1Y4A	1.60	Estrutura cristalina do complexo de Subtilisina BPN' com o inibidor de quimotripsina 2 mutante M59R/E60S
83	1Y4D	2.00	Estrutura cristalina do complexo de Subtilisina BPN' com o inibidor de quimotripsina 2 mutante M59R/E60S
84	6NE2	1.30	Proteína de repetição projetada em complexo com FZ7
85	6DWF	1.94	Estrutura cristalina do complexo do mutante BBKI, L55R, com tripsina bovina
86	4W6Y	1.57	Estrutura co-complexada do domínio lectina da adesina fimbrial F18 FedF com nanocorpo inibidor NBFEDF9
87	6NE4	1.65	Proteína de repetição projetada especificamente em complexo com FZ7CRD
88	3FJU	1.60	Inibidor de carboxipeptidase de Ascaris suum em complexo com carboxipeptidase A1 humana
89	2OUL	2.20	Estrutura do Chagasin em complexo com uma protease cisteína esclarece o modo de ligação e a evolução de uma nova família de inibidores
90	1Y33	1.80	Estrutura cristalina do complexo de Subtilisina BPN' com o inibidor de quimotripsina 2 mutante T58P

91	1Y48	1.84	Estrutura cristalina do complexo de Subtilisina BPN' com o inibidor de quimotripsina 2 mutante R65A
92	2D10	2.50	Estrutura cristalina do domínio FERM de Radixin complexado com o peptídeo da cauda C-terminal de NHERF-1
93	6MOE	2.09	Complexo monomérico de DARPIN E2 com EPOR
94	5AYR	2.40	Estrutura cristalina do complexo Saugi/humana UDG
95	4AYI	2.31	Estrutura do complexo entre CCPs 6 e 7 do fator H do complemento humano e FHbp variante 3 de Neisseria meningitidis
96	3KJ0	1.70	MCL-1 em complexo com mutante Bim BH3 I2DY
97	3KJ1	1.95	MCL-1 em complexo com mutante Bim BH3 I2DA
98	6IUA	1.70	Estrutura cristalina da Importina-Alfa1 ligada ao sinal de localização nuclear 53BP1 (S1678D)
99	3KJ2	2.35	MCL-1 em complexo com mutante Bim BH3 F4AE
100	4Z9K	1.50	Cadeia A da Ricina ligada a um nanocorpo camelídeo (VHH2) (F5)
101	3ONW	2.38	Estrutura de um mutante de G-Alpha-I1 com afinidade aprimorada pelo motivo Goloco de RGS14
102	1R6Q	2.35	A proteína 1R6Q é a estrutura cristalina do ClpNS de *Escherichia coli* complexado com fragmentos, resolvida por difração de raios X a 2,35 Å
103	3EBA	1.85	Mutante CabHUL6 FGLW (humanizado) em complexo com lisozima humana
104	3UKZ	2.30	Importina alfa de camundongo em complexo com CBP80 CNLS de camundongo
105	4AFQ	1.51	Complexo de quimase humana com Fynomer
106	4EQA	1.60	Estrutura cristalina de PA1844 em complexo com PA1845 de Pseudomonas aeruginosa PAO1
107	3VPJ	2.50	Estrutura cristalina do efetor Tse1 do Tipo VI de Pseudomonas aeruginosa em complexo com a proteína imune Tsi1
108	1RI8	1.85	Estrutura cristalina do anticorpo de domínio único camelídeo 1D2L19 em complexo com lisozima de clara de ovo de galinha
109	5JDS	1.70	Estrutura cristalina de PD-L1 complexado com um nanocorpo a 1,7 Å de resolução
110	2VIR	3.25	A proteína 2VIR é o complexo entre a hemaglutinina do vírus da influenza A e um anticorpo neutralizante, com mutações, depositado em 1997 e liberado em 1998.
111	6E3I	1.48	BFL-1 humano em complexo com o peptídeo projetado específico para BFL-1 SRT.F4
112	1TE1	2.5	Estrutura cristalina da xilanase da família 11 em complexo com o inibidor (XIP-I)
113	1KXQ	2.50	Domínio VHH camelídeo em complexo com alfa-amilase pancreática suína

114	5IMM	1.20	Nanocorpo direcionado ao VSIG4 de camundongo no grupo espacial P212121
115	1KXV	1.60	Domínios VHH camelídeos em complexo com alfa-amilase pancreática suína
116	5YIP	1.85	Estrutura cristalina do complexo ANKG LIR/GABARAPL1
117	4JW2	1.90	Seleção de ligantes proteicos específicos para alvos pré-definidos a partir de uma biblioteca otimizada de proteínas repetidas helicoidais artificiais (AlphaRep)
118	5VKO	1.80	SPT6 TSH2-RPB1 1468-1500 PT1471, PS1493
119	1OEY	2.00	Heterodímero dos domínios PB1 de P40PHOX e P67PHOX da NADPH oxidase humana
120	5YIS	2.20	Estrutura cristalina do complexo ANKB LIR/LC3B
121	3EQY	1.63	A proteína 3EQY é a estrutura cristalina do MDMX humano complexado com um inibidor peptídico de 12 aminoácidos, resolvida por difração de raios X a 1,63 Å, com mutações e liberada em 2009.
122	3C4P	1.75	Estrutura cristalina do complexo SHV-1 beta-lactamase/inibidor de beta-lactamase (BLIP) E73M
123	4XWJ	2.10	Complexo entre a proteína histidina-portadora de fosfo (HPR) e o fator anti-sigma RSD
124	3BX1	1.85	Complexo entre o inibidor de alfa-amilase/subtilisina de cevada e a subtilisina Savinase
125	3C4O	1.70	Estrutura cristalina do complexo SHV-1 beta-lactamase/inibidor de beta-lactamase (BLIP) E73M/S130K/S146M
126	6FU9	1.20	Complexo da proteína efetora do arroz (Magnaporthe oryzae) Avr-Pikd com o domínio HMA de Pikm-1 do arroz (Oryza sativa)
127	2Z8W	2.45	Estrutura de um complexo IgNAR-AMA1
128	5UUL	1.33	BFL-1 humano em complexo com PUMA BH3
129	2Z8V	2.35	Estrutura de um complexo IgNAR-AMA1
130	4ZW2	1.86	Estrutura cristalina da subunidade beta do canal de cálcio dependente de voltagem de camundongo isoforma 1A em complexo com peptídeo do domínio de interação alfa
131	4NOO	2.30	Mecanismo molecular de autoproteção contra o sistema de secreção tipo VI em Vibrio cholerae
132	4A1S	2.10	Estrutura cristalográfica do complexo Pins:Insc
133	1XG2	1.90	Estrutura cristalina do complexo entre pectina metilesterase e sua proteína inibidora
134	5N48	1.60	Estrutura do Anticalina N9B em complexo com o domínio extra B da fibronectina oncofetal humana

135	2NU2	1.65	A proteína 2NU2 é a estrutura cristalina da SGPB complexada com variantes de OMTKY3 (Lys18I e Arg18I), resolvida por difração de raios X a 1,65 Å, com mutações e expressa em Escherichia coli.
136	4W6X	1.88	Estrutura do co-complexo do domínio lectina da adesina fimbrial F18 FedF com nanocorpo inibitório NBFedF7
137	2UUY	1.15	Estrutura de um inibidor de tripsina de carrapato em complexo com tripsina bovina
138	2OMW	1.85	A proteína 2OMW é a estrutura cristalina do complexo InlA S192N Y369S/mEC1 de *Listeria monocytogenes* e *Mus musculus*, resolvida por difração de raios X a 1,85 Å, liberada em 2007
139	3P92	1.60	Mesotripsina humana complexada com inibidor de tripsina pancreática bovina variante (BPTI-K15R/R17G)
140	1FLE	1.90	Estrutura cristalina da Elafina complexada com elastase pancreática suína
141	3ZU7	1.97	Estrutura cristalina de uma proteína repetida anquirina projetada em complexo com a quinase MAP ERK2
142	1VRK	1.90	Estrutura de 1.9 Å do complexo E84K-Calmodulina com peptídeo RS20
143	5MTM	2.41	Monocorpo MB(Lck_3) ligado ao domínio SH2 de Lck
144	3K2M	1.75	Estrutura cristalina do monocorpo HA4/Abl1 em complexo com o domínio SH2
145	3WN7	1.57	Estrutura cristalina de Keap1 em complexo com a região N-terminal do fator de transcrição Nrf2
146	3C9A	1.60	Estrutura cristalina de alta resolução de Argos ligado ao domínio EGF de Spitz
147	2VAY	1.94	Calmodulina complexada com peptídeo IQ de Cav1.1
148	2J8X	2.30	Uracil-DNA glicosilase do vírus Epstein-Barr em complexo com UGI de PBS-2
149	1SGP	1.40	A proteína 1SGP é a estrutura cristalina do complexo entre a variante Ala18 do inibidor de ovomucóide de peru e a proteinase B de *Streptomyces griseus*, resolvida por difração de raios X a 1,4 Å, com interações mediadas por moléculas de água e variação na estrutura do bolso de ligação S1
150	6B12	1.71	Estrutura de Tne2 em complexo com Tni2
151	6JJW	2.40	Estrutura cristalina do complexo Kibra e PTPN14
152	2V9T	1.70	Complexo entre o segundo domínio LRR de Slit2 e o primeiro domínio Ig de Robo1
153	1C1Y	1.90	A proteína 1C1Y é a estrutura cristalina do complexo entre RAP.GMPPNP e o domínio de ligação a RAS da quinase C-Raf1 (RAFRBD), obtida por difração de raios X, liberada em 1999 e expressa em *Escherichia coli*.
154	4MP0	2.10	Estrutura de um segundo holoenzima nuclear PP1, forma cristalina 2

155	5XCO	1.25	Estrutura cristalina do mutante G12D da K-Ras humana em complexo com GDP e um peptídeo inibitório cíclico
156	5V5G	2.10	Protease OTU do vírus da febre hemorrágica da Crimeia-Congo ligada a variante de ubiquitina CC.4
157	1TA3	1.70	Estrutura cristalina da xilanase (GH10) em complexo com inibidor (XIP)
158	2HRK	2.05	Base estrutural da formação do complexo de aminoacil-tRNA sintetase de levedura revelada por estruturas cristalinas de dois subcomplexos binários
159	2I26	2.50	Análise da estrutura cristalina do domínio ancestral da nova região de antígeno do tubarão-enfermeira em complexo com lisozima
160	4C2A	2.08	Estrutura cristalográfica do domínio A1 do fator de von Willebrand com mutações R1306Q e I1309V em complexo com GPIb alfa de alta afinidade
161	4HFK	2.10	A proteína 4HFK é a estrutura cristalina do complexo de efetor-imunidade tipo VI Tae4-Tai4 de *Enterobacter cloacae*, resolvida por difração de raios X a 2,10 Å, liberada em 2013 e expressa em *Escherichia coli*.
162	5VT9	1.85	Complexo entre a cadeia leve de miosina 1 e MyoA
163	1ZV5	2.00	Estrutura cristalográfica do domínio variável do anticorpo de cadeia pesada de camelídeo D2-L29 em complexo com a lisozima da clara de ovo de galinha
164	4K5A	1.50	Co-cristalização com proteínas de repetição de anquirina projetadas para conformação específica explica a flexibilidade conformacional de Bcl-W
165	1SV0	2.07	Estrutura cristalográfica do complexo Yan-SAM/Mae-SAM
166	5DJU	2.10	Estrutura cristalográfica do domínio LOV2 (C450A) em complexo com ZDK3
167	3UL4	1.95	Estrutura cristalográfica do complexo Coh-Olpa(Cthe_3080)-Doc918(Cthe_0918): um novo tipo de complexo coesina-dockerina tipo I de *Clostridium thermocellum* ATCC 27405
168	4Z80	1.53	Estrutura cristalográfica do domínio AMA4 DI-DII-EGF1 de *Toxoplasma gondii* em complexo com um peptídeo TGRON2L1 de 33 resíduos
169	3RBB	2.35	Proteína Nef do HIV-1 em complexo com o domínio SH3 de Hck modificado
170	1VEU	2.15	Estrutura cristalográfica do complexo P14/MP1 a 2,15 Å de resolução
171	1VET	1.90	Estrutura cristalográfica do complexo P14/MP1 a 1,9 Å de resolução
172	6BXC	2.50	Estrutura cristalográfica do fragmento N-terminal do receptor Toll-like 5 (*TLR5*) de *Danio rerio* em complexo com o receptor linfocitário variável 9 (*VLR9*) de lampreia
173	1X1U	2.30	Interação mediada por água em uma interface proteína-proteína
174	6E3J	1.48	*BFL-1* humano em complexo com o peptídeo projetado específico para BFL-1, *SRT.F10*
175	5NVM	2.00	Estrutura cristalográfica do complexo humano 4EHP-GIGYF2 sem as sequências auxiliares

176	4JE4	2.31	Estrutura cristalográfica do complexo entre o monobody NSA1 e o domínio N-SH2 de SHP2
177	1SYQ	2.42	Domínio cabeça da vinculina humana (*VH1*, resíduos 1-258) em complexo com o domínio de ligação à vinculina da talina humana (*resíduos 607-636*)
178	4ZQU	2.09	Complexo entre a toxina CdiA-CT e o fator de imunidade CdiI de *Yersinia pseudotuberculosis*
179	3P9W	2.41	Estrutura cristalográfica de um domínio VH humano autônomo modificado em complexo com VEGF
180	3K6G	1.95	Estrutura cristalográfica do complexo Rap1-TRF2
181	5DJT	1.40	Estrutura cristalográfica do domínio LOV2 (C450A) em complexo com ZDK2
182	6J4O	2.30	Base estrutural da detirossinação da tubulina pelo complexo enzimático Vasohibins-SVBP e suas implicações funcionais
183	2WH6	1.50	Estrutura cristalográfica da proteína antiapoptótica *BHRF1* em complexo com o domínio BH3 de *BIM*
184	6JHW	2.04	Estrutura do complexo anti-CRISPR *AcrIIc3* e *NmCas9* HNH
185	1GPW	2.40	Evidência estrutural do túnel de amônia através do barril (β/α)8 da imidazol glicerol fosfato sintetase bi-enzimática
186	3D5R	2.10	Estrutura cristalográfica do complexo *Efb-C (N138A)* / *C3d*
187	3FXD	2.10	Estrutura cristalográfica dos domínios de interação de *IcmR* e *IcmQ*
188	2YGG	2.23	Complexo entre *CamBR* e *CaM*
189	2J12	1.50	Cabeça da fibra do adenovírus tipo 37 em complexo com o domínio D1 do receptor CAR
190	4IU3	1.97	Complexo entre coesina, dockerina e domínio X de *Ruminococcus flavefaciens*
191	3H8K	1.80	Estrutura cristalográfica da *UBE2G2* em complexo com o domínio *G2BR* de *Gp78* a 1,8 Å de resolução
192	5F5S	2.40	Estrutura cristalográfica do complexo *PRP38-MFAP1* de *Homo sapiens*
193	2P42	1.80	Complexo entre um fragmento de anticorpo VHH de camelídeo e a *RNase A* a 1,8 Å de resolução (*Se3-Mono-2*, forma cristalina com três sítios de Se-Met: M34, M51, M83 no scaffold VHH)
194	2P43	1.65	Complexo de um fragmento de anticorpo de domínio único de camelídeo com RNase A a 1,65 Å de resolução: forma cristalina Se3-Mono-1 com três locais Se-Met (M34, M51, M83) na estrutura Vhh
195	1UUZ	1.80	Ivy: uma nova família de proteínas
196	3D5S	2.30	Estrutura cristalina do complexo Efb-C (R131A) / C3d
197	1X1W	2.10	Interação mediada por água em uma interface proteína-proteína
198	2JEL	2.50	A proteína 2JEL é a estrutura cristalina do complexo entre o anticorpo JEL42 FAB e HPR, resolvida por difração de raios X a 2,50 Å, liberada em 1998 e expressa em *Mus musculus* e *Escherichia coli*.

199	5Y4R	2.30	Estrutura de um complexo de metiltransferase
200	6FUB	1.30	Complexo da proteína efetora Avr-PikE do fungo Magnaporthe oryzae com o domínio HMA de Pikm-1 do arroz (Oryza sativa)
201	5TZP	1.35	Estrutura cristalina do complexo FPV039:Bik BH3
202	6H46	2.22	KRAS humano em complexo com DARPIN K13
203	1YCS	2.20	Complexo P53-53BP2
204	3V1C	1.13	Estrutura cristalina de uma proteína Mid1-Zinc projetada de novo
205	4LYL	1.93	Estrutura cristalina da uracil-DNA glicosilase do bacalhau (Gadus morhua) em complexo com o inibidor proteico Ugi
206	6SAK	2.00	Estrutura do complexo OTULINCAT C129A - SNX27 PDZ
207	4BD9	2.20	Estrutura do complexo entre SmcI e carboxipeptidase A4 humana
208	4DH2	1.75	Estrutura cristalina do complexo Coh-OlpC (Cthe_0452) - Doc435 (Cthe_0435): um novo tipo de complexo coesina-dockerina tipo I de Clostridium thermocellum ATCC 27405
209	5M72	1.60	Estrutura do domínio de ligação de proteínas SRP68-72 humano
210	3GQI	2.50	Estrutura cristalina de uma quinase de receptor tirosina ativada em complexo com substratos
211	3G3B	2.40	Estrutura de um receptor linfocitário variável de lampreia mutante em complexo com um antígeno proteico
212	2YQ7	1.90	Estrutura do Bcl-XL ligado ao BimLock
213	5DC4	1.48	Estrutura cristalina do complexo Monobody As25/Abl1 SH2, cristal A
214	2W8B	1.86	Estrutura cristalina do TolB processado em complexo com Pal
215	1T01	2.06	Vinculina complexada com a hélice VBS1 de Talina
216	5TDY	2.11	Estrutura do complexo coenovelado FliFC:FliGN de Thermotoga maritima
217	1GUA	2.00	Rap1A humano (resíduos 1-167), duplo mutante (E30D, K31E) complexado com GppNHp e o domínio de ligação de Ras de C-Raf1 humano (resíduos 51-131)
218	5E3E	1.70	Estrutura cristalina do complexo CdiA-CT/CdiI de Y. kristensenii 33638
219	6J9H	2.31	Estrutura cristalina do complexo SVBP-VASH1
220	1ZC3	2.00	Estrutura cristalina do domínio de ligação de Ral de Exo84 em complexo com RalA ativo
221	1ZC4	2.50	Estrutura cristalina do domínio de ligação de Ral de Exo84 em complexo com RalA ativo
222	6FHP	1.70	DaIP em complexo com um fragmento C-terminal de termolisina
223	1NF3	2.10	Estrutura de CDC42 em complexo com o domínio de ligação de GTPase da proteína de polaridade celular, Par6
224	3T04	2.10	Estrutura cristalina do complexo Monobody 7C12/Abl1 SH2
225	1X1X	2.30	Interação mediada por água em uma interface proteína-proteína
226	1PK1	1.80	Estrutura do domínio SAM heterodimérico de PH e SCM

227	6F0G	2.30	Estrutura cristalina ASF1-IP3
228	2A9K	1.73	Estrutura cristalina do complexo C3bot-NAD-RalA, revelando um novo tipo de ação de uma exoenzima bacteriana
229	2A78	1.81	Estrutura cristalina do complexo C3bot-RalA, revelando um novo tipo de ação de uma exoenzima bacteriana
230	4AOR	1.70	Tripsina catiônica em complexo com o inibidor de tripsina Spinacia oleracea III (SOTI-III)
231	2F4M	1.85	O complexo Pngase-Hr23 de camundongo revela uma remodelação completa da interface proteína-proteína em comparação com seus ortólogos de levedura
232	3RZW	2.15	Estrutura cristalina do monobody YSMB-9 ligado ao SUMO1 humano
233	2V5Q	2.30	Estrutura cristalina do domínio quinase PLK-1 selvagem em complexo com um DARPIN seletivo
234	5J4A	2.00	Toxina CdiA-CT de Burkholderia pseudomallei E479 em complexo com a proteína de imunidade Cdii cognata
235	6FUD	1.30	Complexo da proteína efetora Avr-PikA do fungo Magnaporthe oryzae com o domínio HMA de Pikm-1 do arroz (Oryza sativa)
236	6ERE	2.25	Estrutura cristalina de um par endonuclease e imunidade Colicin projetado computacionalmente, Coledes3/Imdes3
237	2WWX	1.50	Estrutura cristalina do complexo SidM/DRRa (domínio GEF/GDF) - Rab1 (domínio GTPase)
238	1PVH	2.50	Estrutura cristalina do fator inibidor de leucemia em complexo com GP130
239	3QHT	2.40	Estrutura cristalina do monobody YSMB-1 ligado ao SUMO de levedura
240	2UYZ	1.40	Complexo não covalente entre Ubc9 e SUMO1
241	4YN0	2.20	Estrutura cristalina do domínio E2 de APP em complexo com o domínio CRD de DR6
242	5JW9	2.00	Estrutura cristalina do domínio Oclludin de ELL2 e peptídeo AFF4
243	3IXS	1.70	Complexo do domínio C-terminal de Ring1b e o domínio C-terminal de RYBP
244	3BS5	2.00	Estrutura cristalina do complexo HCNK2-SAM/DHYP-SAM
245	1MCV	1.80	Análise estrutural de um inibidor de squash híbrido em complexo com elastase pancreática suína
246	1B2U	2.10	Resposta estrutural à mutação em uma interface proteína-proteína
247	5ET1	1.65	Estrutura cristalina de Myo3b-Arb1 em complexo com Espin1-AR
248	4REY	1.96	Estrutura cristalina do complexo peptídeo GRASP65-GM130 C-terminal
249	5VKL	2.20	Estrutura do complexo SPT6 TSH2-RPB1 1476-1500 PS1493
250	2XZE	1.75	Base estrutural para a interação entre AMSH e ESCRT-III CHMP3

251	2P47	2.50	Complexo de um fragmento de anticorpo de domínio único de camelídeo com RNase A a 2.5Å de resolução: forma cristalina Se5b-Tri com cinco locais Se-Met (L4M, M34, M51, F68M, M83) no andaime VHH
252	2P45	1.10	Complexo de um fragmento de anticorpo de domínio único de camelídeo com RNase A a 1.1Å de resolução: forma cristalina Se5b-Ortho-1 com cinco locais Se-Met (L4M, M34, M51, F68M, M83) no VHH
253	2P48	2.30	Complexo de um fragmento de anticorpo de domínio único de camelídeo com RNase A a 2.3Å de resolução: forma cristalina Se5b-Tetra com cinco locais Se-Met (L4M, M34, M51, F68M, M83) no andaime VHH
254	1G9I	2.20	Estrutura cristalina do complexo Beta-tripsina em ciclohexano
255	5GJK	2.05	Estrutura cristalina do complexo BAF47 e BAF155
256	1SMF	2.10	Estudos sobre um inibidor de tripsina artificial derivado do inibidor de feijão-mungo
257	5INB	1.30	Complexo holoenzima Repoman-PP1G (fosfatase de proteína 1, isoforma gama)
258	2nu4	1.75	Estruturas cristalinas do SGPB em complexo com as variantes OMTKY3 Lys18I e Arg18I, destacando a acomodação de resíduos positivamente carregados em um bolso de especificidade hidrofóbico. Proteína originária de *Streptomyces griseus* e *Meleagris gallopavo*. Método: Difração de raios X.
259	4PAS	1.62	Estrutura de bobina enrolada heterodimérica do receptor humano GABA(B)
260	3OJM	2.10	Estrutura cristalina de FGF1 complexada com o ectodomínio de FGFR2b contendo a mutação P253R da síndrome de Apert
261	3OJ2	2.20	Estrutura cristalina de FGF1 complexada com o ectodomínio de FGFR2b contendo a mutação A172F da síndrome de Pfeiffer
262	3DA7	2.25	Um permutante circular conformacionalmente tensionado de barnase
263	3L33	2.48	Mesotripsina humana complexada com o inibidor da proteína precursora amiloide (APPI)
264	4U30	2.50	Mesotripsina humana complexada com o domínio Kunitz 2 de Bikunin
265	5EB1	1.80	O complexo Yfib-Yfir
266	3TZ1	1.80	Estrutura cristalina do domínio C-terminal saturado com Ca^{2+} da troponina C da vieira Akazara em complexo com um fragmento da troponina I
267	4P3Y	2.15	Estrutura cristalina da DsbA de *Acinetobacter baumannii* em complexo com Ef-Tu
268	5V5H	1.50	Protease OTU do vírus da febre hemorrágica da Crimeia-Congo ligada a uma variante de ubiquitina CC.2
269	6AAF	2.20	Estrutura cristalina da Atg8 da levedura de fissão complexada com a hélice AIM de Hfl1

270	5F5O	2.20	Estrutura cristalina do domínio central da nucleoproteína do vírus de Marburg ligada ao peptídeo regulador Vp35
271	5O2T	2.19	KRAS humano em complexo com DARPIN K27
272	3OAK	2.15	Estrutura cristalina de um complexo Spn1 (Iws1)-Spt6
273	5G15	2.06	Estrutura de Aurora A (122-403) ligada ao monobody ativador MB1 e AMPPcP
274	4MIT	2.35	Estrutura cristalina da RacC de *Entamoeba histolytica* ligada ao domínio PBD da EhPAK4
275	4M1L	2.10	Complexo de IQCG e Ca^{2+}-calmodulina
276	4Y61	3.36	A proteína 4Y61 é a estrutura cristalina do complexo entre Slitrk2 LRR1 e PTP delta Ig1-Fn1, obtida por difração de raios X, liberada em 2015 e expressa em *Homo sapiens*.
277	6F0F	2.0	Estrutura cristalina ASF1-IP2_S
278	4H6J	1.52	Identificação da Cys 255 em HIF-1 como um novo alvo para o desenvolvimento de inibidores covalentes da interação proteína-proteína do domínio PASB de HIF-1/ARNT
279	5J28	2.00	Complexo holoenzimático Ki67-PP1γ (proteína fosfatase 1, isoforma gama)
280	4PC0	2.50	Estrutura do complexo humano RbAp48-MTA1(670-711)
281	3DVU	2.50	Estrutura cristalina do complexo do homólogo M11 da Bcl-2 do herpesvírus gama murino 68 e o domínio BH3 da Beclin 1
282	6BX8	1.98	Mesotripsina humana (PRSS3) complexada com a variante do inibidor da via do fator tecidual (TFPI1-KD1-K15R-I17C-I34C)
283	4AOQ	2.00	Tripsina catiônica em complexo com o inibidor de tripsina mutado de *Spinacia oleracea* (SOTI-III) (F14A)
284	2BTF	2.55	A proteína 2BTF é a estrutura cristalina do complexo profilina-beta-actina de *Bos taurus*, resolvida por difração de raios X a 2,55 Å, liberada em 1994.
285	4PZ6	2.41	Guanililtransferase PCE1 ligada à RNA pol II CTD fosforilada em Ser2/Ser5
286	5ML9	2.35	Coestrutura cristalina do receptor Fc gama IIIA interagindo com Affimer F4, uma proteína de ligação específica que bloqueia a ligação de IgG ao receptor
287	3DXC	2.10	Estrutura cristalina do domínio intracelular da APP humana em complexo com Fe65-PTB2
288	6K06	1.75	Estrutura cristalina da importina-alfa e GM130 fosfomimética
289	1U0S	1.90	Domínio P2 da quinase quimiotática CheA em complexo com o regulador de resposta CheY da termófila *Thermotoga maritima*
290	3QBQ	2.50	Estrutura cristalina dos domínios extracelulares do complexo RANK-RANKL de camundongo
291	4NL9	1.50	Estrutura cristalina do heterodímero SAM de ANKS3-ANKS6 humano
292	4RJF	2.01	Estrutura cristalina do grampo deslizante humano a 2,0 Å de resolução

293	4C4P	2.00	Estrutura cristalina da RAB11 do tipo selvagem complexada com FIP2
294	4GN4	1.86	Obody AM2EP06 ligado à lisozima da clara de ovo de galinha
295	5H3J	1.33	Estrutura cristalina do domínio GRASP de GRASP55 complexado com a extremidade C-terminal de Golgin45
296	3ZEU	1.65	Estrutura de um heterodímero YgjD-Yeaz de *Salmonella typhimurium* ligado ao ATPγS composto 13
297	6JB2	1.50	Estrutura cristalina do nanocorpo D3-L11 mutante Y102A em complexo com a lisozima da clara de ovo de galinha
298	1XD3	1.45	Estrutura cristalina do complexo UCHL3-UBVMe
299	4HEP	1.75	Complexo do bacteriófago lactocócico TP901-1 com um nanocorpo de lhama VHH (VHH17)
300	3DXE	2.00	Estrutura cristalina do domínio intracelular da APP humana (mutante T668A) em complexo com Fe65-PTB2
301	1LZW	2.50	Base estrutural da mudança mediada por ClpS no reconhecimento de substrato de ClpA
302	5MTJ	1.95	YES1-SH2 em complexo com monobody MB(YES_1)
303	3DXD	2.20	Estrutura cristalina do domínio intracelular da APP humana (mutante T668E) em complexo com Fe65-PTB2
304	6FBX	1.64	Estrutura cristalina da proteína pró-sobrevivência NRZ do peixe-zebra complexada com BAD BH3
305	5WOS	2.45	Informações estruturais e funcionais sobre a regulação da apoptose pelo CNP058 do vírus da varíola do canário
306	2WY8	1.70	Proteína de subversão do complemento SBI-IV de *Staphylococcus aureus* em complexo com o fragmento C3d do complemento
307	5EO9	2.30	Estrutura cristalina do complexo do domínio 1 de DPR6 ligado ao domínio 1+2 de DIP-Alpha
308	4U32	1.65	Mesotripsina humana complexada com o domínio Kunitz 1 de HAI-2
309	1GRN	2.10	Estrutura cristalina do complexo CDC42/CDC42GAP/AlF_3
310	2B7C	1.80	Fator de troca de nucleotídeos de guanina Eef1Bα mutante K205A de levedura em complexo com Eef1A
311	6F3Z	2.00	Complexo de *E. coli* LolA e domínio periplásmico de LolC
312	6S8S	2.21	Estrutura expandida do domínio C-terminal de DDX6 humano em complexo com um peptídeo FDF de EDC3
313	5ELU	2.35	Inibição isoforma-específica de interações proteína-proteína dependentes de SUMO
314	4ZI2	2.20	Domínio tipo BART de BARTL1/CCDC104 em complexo com ARL3FL ligado a GPPNHP em P21 21 21
315	5WUJ	2.30	Estrutura cristalina do complexo FliF-FliG de *Helicobacter pylori*

316	3TKL	2.18	Estrutura cristalina da Rab1A ligada a GTP em complexo com o domínio em hélice de Lida de *Legionella pneumophila*
317	3KUC	1.92	Complexo de RAP1A(E30D/K31E)GDP com RAFRBD(A85K/N71R)
318	5IP4	1.81	Estrutura de raios-X do domínio C-terminal de Doublecortin humana
319	4I0C	1.95	Estrutura do anticorpo camelídeo CabHul5 em complexo com lisozima humana
320	5V5I	2.20	Protease OTU do vírus da febre hemorrágica da Crimeia-Congo ligada à variante de ubiquitina CC.1
321	2OIN	2.50	Estrutura cristalina do mutante R155K da NS3-4A do HCV
322	5H9B	2.25	*Drosophila* CaMKII selvagem em complexo com um fragmento do canal de potássio EAG e Mg^{2+}/AMPPNP
323	5EP6	1.45	Estrutura cristalina de NAP1 em complexo com TBK1
324	3G6D	3.20	A proteína 3G6D é a estrutura cristalina do complexo entre CNTO607 Fab e IL-13, obtida por difração de raios X a 3,20 Å, liberada em 2009 e expressa em *Homo sapiens* e *Escherichia coli*
325	2IY1	2.46	SENP1 (mutante) de comprimento total com SUMO1
326	5KY5	1.50	POFUT1 de camundongo em complexo com EGF(+) e GDP
327	2WP3	1.48	Estrutura cristalina do complexo Titin M10-Obscurin Like 1 IG
328	6GHO	1.79	Estrutura cristalina de SPX em complexo com YJBH
329	4CMM	1.92	Estrutura de CD47 humano em complexo com a proteína reguladora de sinalização (SIRP) alfa V1 humana
330	4AYD	2.40	Estrutura do complexo entre CCPs 6 e 7 do fator H do complemento humano e a variante 1 R106A mutante de FHbp de *Neisseria meningitidis*
331	5IMK	1.23	Nanocorpo direcionado para VSIG4 humano no grupo espacial C2
332	1E6E	2.30	Complexo redutase/adrenodoxina da mitocôndria em sistemas P450
333	1WQJ	1.60	Base estrutural para a regulação dos fatores de crescimento semelhantes à insulina (IGFs) pelas proteínas de ligação de IGF (IGFBPs)
334	5TVQ	2.35	Domínio catalítico da TDP2 de camundongo ligado a SUMO2
335	5VMO	1.70	Estrutura cristalina do complexo GIV66:BIM do iridovírus de garoupa
336	3QBT	2.00	Estrutura cristalina de OCRL1 (resíduos 540-678) em complexo com Rab8A:GPPNHP
337	4MRT	2.00	Estrutura da fosfopanteteína transferase Sfp em complexo com coenzima A e uma proteína carreadora de peptídeo
338	3KNB	1.40	Estrutura cristalina do C-terminal da titina em complexo com Obscurin-Like 1
339	2W80	2.35	Complexo proteína-proteína entre o fator H de ligação da *Neisseria meningitidis* e 123-mer. Parte do *general set* do PDBbind.
340	5HU3	1.89	*Drosophila* CaMKII-D136N em complexo com um fragmento fosforilado do canal de potássio EAG e Mg^{2+}/ADP
341	1AY7	1.70	Ribonuclease Sa em complexo com Barstar

342	4ZRK	2.32	Complexo Merlin-FERM e LATS1
343	5N88	1.70	Estrutura cristalina de um anticorpo ligado a uma proteína viral
344	6GUM	1.79	Estrutura do domínio UFD da E1 de *Arabidopsis thaliana* em complexo com E2
345	2JJT	2.30	Estrutura de CD47 humano em complexo com a proteína reguladora de sinalização (SIRP) alfa
346	2JJS	1.85	Estrutura de CD47 humano em complexo com a proteína reguladora de sinalização (SIRP) alfa
347	2G2U	1.60	Estrutura cristalina do complexo SHV-1 β-lactamase/inibidor de β-lactamase (BLIP)
348	2WWK	1.70	Estrutura cristalina do complexo Titin M10-Obscurin Like 1 IG mutante F17R
349	1UUG	2.40	Complexo da uracil-DNA glicosilase de *Escherichia coli* com seu inibidor Ugi
350	4EHQ	1.90	Estrutura cristalina do domínio de ligação da calmodulina de Orai1 em complexo com Ca^{2+}/calmodulina
351	4MI8	2.10	Estrutura cristalina do complexo da proteína Bcl-2 homóloga M11 do herpesvírus murino gama-68 com um peptídeo derivado do domínio BH3 de Beclin 1
352	1J2J	1.60	Estrutura cristalina da região N-terminal de GGA1 GAT em complexo com ARF1 na forma GTP
353	3HCT	2.10	Estrutura cristalina de TRAF6 em complexo com UBC13 no grupo espacial P1
354	3QQ8	2.00	Estrutura cristalina de P97-N em complexo com FAF1-UBX
355	3P95	1.30	Mesotripsina humana complexada com um inibidor de tripsina pancreática bovina variante (BPTI-K15R/R17D)
356	4DM6	1.90	Estrutura cristalina do homodímero do domínio LBD de RARB em complexo com TTNPB
357	6UYS	1.59	Estrutura cristalina de SUMO1 acetilado em K37 em complexo com PML-SIM fosforilado
358	3KUD	2.15	Complexo de RAS-GDP com RAFRBD(A85K)
359	5FVK	1.66	Estrutura cristalina do complexo VPS4-VFA1 de *Saccharomyces cerevisiae*
360	3HTU	2.00	Estrutura cristalina do subcomplexo humano VPS25-VPS20
361	4D0G	2.50	Estrutura de Rab14 em complexo com a proteína de acoplamento de Rab (RCP)
362	1LFD	2.10	Estrutura cristalina da proteína RAS ativa complexada com o domínio de interação de RALGDS
363	1LP1	2.30	Proteína Z em complexo com um affibody selecionado in vitro
364	4YIQ	1.85	Estrutura do heterodímero CEACAM6-CEACAM8
365	1J7D	1.85	Estrutura cristalina de HMMS2-HUBC13

366	6UYU	1.66	Estrutura cristalina de SUMO1 acetilado em K45 em complexo com PML-SIM fosforilado
367	1FFW	2.70	A proteína 1FFW é a estrutura cristalina do domínio de ligação CheY da CheA em complexo com CheY e imido difosfato, resolvida por difração de raios X a 2,7 Å, liberada em 2001 e proveniente de *Escherichia coli*. A estrutura foi depositada por Gouet et al., e a pesquisa fornece insights sobre o mecanismo de função do regulador de resposta CheY.
368	2P1L	2.50	Estrutura do complexo Bcl-XL:Beclin 1
369	5TAR	1.90	Estrutura cristalina da KRAS4B farnesilada e metilada em complexo com PDE-Delta (forma cristalina II – com região hipervariável ordenada)
370	5X4L	2.40	Estrutura cristalina do domínio UBX da UBXD7 humana em complexo com o domínio N de P97
371	2TGP	1.90	Geometria do sítio reativo e dos grupos peptídicos na tripsina, tripsinogênio e seus complexos com inibidores
372	3MJH	2.03	Estrutura cristalina de Rab5a humano em complexo com o dedo de zinco C2H2 de EEA1
373	4AQE	2.27	Estrutura cristalina do mutante associado à surdez da caderina-23 de camundongo (EC1-2 S70P) e protocaderina-15 (EC1-2) – Forma I
374	5MN2	2.35	Coestrutura cristalina do receptor Fc Gama IIIA interagindo com Affimer G3, uma proteína de ligação específica que bloqueia a ligação de IgG ao receptor
375	3WA0	2.31	Estrutura cristalina do complexo Merlin com DCAF1/VprBP
376	4C4K	1.95	Estrutura cristalina do complexo do domínio Ig M10 da titina com obscurina
377	4DG4	1.40	Mesotripsina humana S39Y complexada com inibidor de tripsina pancreática bovina (BPTI)
378	1CBW	2.60	A estrutura 1CBW refere-se ao complexo entre a quimotripsina bovina e o inibidor BPTI, resolvida por difração de raios X a 2,60 Å, liberada em 1997 e proveniente de *Bos taurus*. A pesquisa foi depositada por Hynes et al., e envolve a interação entre uma serina protease e seu inibidor.
379	3D1M	1.70	Estrutura cristalina do Sonic Hedgehog ligado ao terceiro domínio FNIII de CDO
380	2G45	1.99	Coestrutura cristalina do domínio ZnF UBP da enzima deubiquitinante isopeptidase T (ISOT) em complexo com ubiquitina
381	4APX	1.65	Estrutura cristalina da caderina-23 de camundongo (EC1-2) e protocaderina-15 (EC1-2) – Forma I
382	4MQV	1.95	Complexo cristalino de RPA32C com o N-terminal de SMARCAL1
383	3GJ8	1.82	Estrutura cristalina do complexo RanGDP-Nup153ZnF34 humano
384	5MK0	1.77	Estrutura cristalina do domínio BRO1 da proteína tirosina fosfatase HIS (HD-PTP/PTPN23) complexada com peptídeo Endofin

385	4ETW	2.05	Estrutura do complexo enzima-ACP, essencial para a síntese de biotina
386	5AJJ	1.75	Estrutura cristalina do fator de virulência F1L do vírus da varíola em complexo com o domínio BH3 de BID humano
387	5W3X	2.00	Estrutura cristalina de POPP2 em complexo com IP6, AcCoA e o domínio WRKY de RRS1-R
388	4G01	2.20	Estrutura do complexo Ara7-GDP-Ca^{2+}/VPS9a
389	3TL8	2.50	Estrutura do complexo AvrPtoB-BAK1 revelando dois domínios de interação quinase estruturalmente semelhantes em um único efetor tipo III
390	2IYB	2.35	Estrutura do complexo entre o terceiro domínio LIM de TES e o domínio EVH1 de Mena
391	5MRV	1.85	Estrutura cristalina da carboxipeptidase O humana em complexo com NVCI
392	2PTT	1.63	Estrutura do receptor de célula NK 2B4 (CD244) ligado ao seu ligante CD48
393	3GJ7	1.93	Estrutura cristalina do complexo RanGDP-Nup153ZnF12 humano
394	2X9A	2.47	Estrutura cristalina de G3P do fago If1 em complexo com seu coreceptor, o domínio C-terminal de TolA
395	4JS0	1.90	Complexo de CDC42 com o domínio CRIB-PR de IRSp53
396	4NIQ	2.30	Estrutura cristalina de VPS4 MIT-VFA1 MIM2
397	3EJH	2.10	Estrutura cristalina do par de domínios FN8-9 da fibronectina em complexo com um peptídeo de colágeno tipo I
398	2V3B	2.45	Estrutura cristalina do complexo de transferência de elétrons rubredoxina - redutase de rubredoxina de *Pseudomonas aeruginosa*
399	2B11	2.30	Estrutura cristalina do complexo proteína-proteína entre citocromo C F82W e peroxidase de citocromo C
400	4XXW	2.26	Estrutura cristalina da caderina-23 de camundongo (EC1-2) e protocaderina-15 (EC1-2) – Variante de splicing
401	4DID	2.35	Estrutura cristalina do domínio N-terminal do efetor *Salmonella* SopB em complexo com CDC42
402	1L0Y	2.50	Complexo da cadeia beta do receptor de célula T com a superantígena SpeA, em presença de zinco
403	2X2D	1.95	Complexo acetil-CyPA com o domínio N-terminal da cápside do HIV-1
404	2C7M	2.40	Resíduos 1-74 de Rabex-5 humano em complexo com ubiquitina
405	2C7N	2.10	Resíduos 1-74 de Rabex-5 humano em complexo com ubiquitina
406	3GJ4	2.15	Estrutura cristalina do complexo RanGDP-Nup153ZnF3 humano
407	4M0W	1.40	Estrutura cristalina da protease semelhante à papaína do SARS-CoV mutante C112S em complexo com ubiquitina
408	3UYO	1.83	Estrutura cristalina do complexo Monobody SH13/Abl1 SH2
409	6AKM	2.30	Estrutura cristalina do complexo SLMAP-SIKE1
410	4UF1	2.30	Proteína DPV022 do vírus *Deerpox* em complexo com Bak BH3

411	5JZE	2.47	Protease do domínio OTU do vírus Erve em complexo com ISG15 de camundongo
412	5OYL	2.25	Estrutura cristalina da proteína G do VSV (VSV G CR2)
413	5YR0	1.90	Estrutura do complexo de domínio em hélice de Beclin1-UVRAG
414	2CCL	2.03	A estrutura 2CCL corresponde ao mutante S45A, T46A do complexo tipo I coesina-dockerina da celulosoma de *Clostridium thermocellum*, resolvida por difração de raios X a 2,03 Å, e foi liberada em 2007. Foi depositada por Carvalho et al. e envolve a interação entre proteínas da adesão celular. A expressão do mutante foi realizada em *Escherichia coli*.
415	1F3V	2.00	Estrutura cristalina do complexo entre o domínio N-terminal de TRADD e o domínio TRAF de TRAF2
416	5XOC	2.40	Estrutura cristalina do complexo SMAD3-FOXH1 humano
417	3VUX	1.70	Estrutura cristalina de A20 ZF7 em complexo com ubiquitina linear, forma II
418	5KY4	1.47	Estrutura cristalina da POFUT1 de camundongo em complexo com NOTCH1 EGF26 e GDP
419	4K1R	1.63	Estrutura cristalina do domínio catalítico SST2 de *Schizosaccharomyces pombe* em complexo com ubiquitina
420	3VEP	2.50	Estrutura cristalina de SIGD4 em complexo com seu regulador negativo RSDA
421	3QC8	2.20	Estrutura cristalina do domínio UBX de FAF1 em complexo com o domínio N de P97/VCP, revelando o motivo de toque e giro FCISP do domínio UBX
422	5G1X	1.72	Estrutura cristalina da quinase Aurora-A em complexo com N-MYC
423	4H5S	1.70	Estrutura do complexo NECL-2 e CRTAM
424	1X1Y	1.90	Interação mediada por água em uma interface proteína-proteína
425	4LLO	2.00	Estrutura do complexo domínio EAG-CNBHD do canal EAG1 de camundongo
426	4PW9	2.49	Estrutura cristalina do complexo de transferência de elétrons entre uma sulfito desidrogenase e um citocromo tipo C de *Sinorhizobium meliloti*
427	2R9P	1.40	Mesotripsina humana complexada com inibidor de tripsina pancreática bovina (BPTI)
428	6ASR	2.36	Complexo do domínio UBM2 de REV1 com ubiquitina
429	1T63	2.07	Estrutura cristalina do domínio de ligação ao ligante do receptor de andrógeno com DHT e um peptídeo derivado do coativador GRIP1 NR BOX3
430	1WQ1	2.50	Complexo Ras-RasGAP
431	3B08	1.70	Estrutura cristalina do domínio HOIL1-L-NZF de camundongo em complexo com di-ubiquitina linear
432	4D0N	2.10	Domínio RhoGEF de AKAP13 (AKAP-LBC) em complexo com RhoA
433	4BWQ	2.10	Estrutura cristalina de U5-15kd em complexo com PQBP1

434	2V8S	2.22	Complexo do domínio HABC de VTI1B com o domínio ENTH de EPSINR
435	5XIU	1.80	Estrutura cristalina do domínio UDM2 de RNF168 em complexo com diubiquitina ligada a Lys63
436	1T0P	1.66	Base estrutural do reconhecimento de ICAM por integrina αLβ2 revelada na estrutura do complexo dos domínios de ligação de ICAM-3 e αLβ2
437	1T5Z	2.30	Estrutura cristalina do domínio de ligação ao ligante do receptor de andrógeno (LBD) com DHT e um peptídeo derivado do coativador fisiológico ARA70
438	3G9W	2.17	Estrutura cristalina de Talin2 F2-F3 em complexo com a cauda citoplasmática da integrina β1D
439	4XKL	2.10	Estrutura cristalina do domínio ZF2 de NDP52 em complexo com mono-ubiquitina
440	3R85	1.95	Estrutura cristalina do domínio BH3 de SOUL humano em complexo com Bcl-XL
441	4A49	2.21	Estrutura do complexo fosfoTyr371-C-Cbl-UbcH5b
442	3GJ5	1.79	Estrutura cristalina do complexo RanGDP-Nup153ZnF4 humano
443	5FZT	2.10	Estrutura cristalina de R7R8 em complexo com um fragmento de DLC1
444	3GJ3	1.79	Estrutura cristalina do complexo RanGDP-Nup153ZnF2 humano
445	2OOB	1.90	Estrutura cristalina do domínio UBA da ligase de ubiquitina Cbl-b em complexo com ubiquitina
446	2QHO	1.85	Estrutura cristalina do domínio UBA da ligase de ubiquitina EDD em complexo com ubiquitina
447	3BZD	2.30	Manipulação do processo de dobramento e ligação acoplados impulsiona a maturação da afinidade em um complexo proteína-proteína
448	1S0W	2.30	A proteína 1S0W é uma beta-lactamase inibidora de Escherichia coli e Streptomyces clavuligerus, cristalizada por difração de raios X, com mutações introduzidas e depositada em 2004
449	1WRD	1.75	Estrutura cristalina do domínio GAT de TOM1 em complexo com ubiquitina
450	1S1Q	2.00	Domínio UEV de TSG101 em complexo com ubiquitina
451	5TL7	2.44	Estrutura cristalina da protease semelhante à papaína do SARS-CoV em complexo com o domínio C-terminal do ISG15 de camundongo
452	3N06	2.00	Antagonista mutante do receptor de prolactina humano H27A em complexo com o domínio extracelular do receptor de prolactina humano
453	3MZG	2.10	Estrutura cristalina de um antagonista do receptor de prolactina humano em complexo com o domínio extracelular do receptor de prolactina humano
454	3N0P	2.10	Antagonista mutante do receptor de prolactina humano H30A em complexo com o domínio extracelular do receptor de prolactina humano
455	3NCB	2.10	Antagonista mutante do receptor de prolactina humano H180A em complexo com o domínio extracelular do receptor de prolactina humano

456	3NCC	2.50	Antagonista do receptor de prolactina humano em complexo com o domínio extracelular mutante H188A do receptor de prolactina humano
457	2SGP	1.80	A entrada 2SGP descreve a estrutura cristalina do complexo entre o domínio 3 da variante PRO 18 do inibidor ovomucoide de peru e a proteína Streptomyces griseus B. O complexo é um heterodímero com atividade inibidora de protease serina.
458	4MYW	1.95	A proteína 4MYW é um complexo cristalino de Hsp90 com um inibidor, resolvido a 1,95 Å, relevante para estudos de desenho de fármacos.
459	1SGE	1.80	Estrutura cristalina da variante Glu18 do inibidor OMTKY3 em complexo com SGPB a pH 6.5. Proteína de *Streptomyces griseus* e *Meleagris gallopavo*. Método: Difração de raios X.
460	3BK3	2.70	Complexo proteína-proteína entre o domínio VWC1 da Crossveinless 2 (CV-2) e BMP-2. Parte do *general set* do PDBbind.
461	1PPF	1.80	A entrada 1PPF descreve a estrutura cristalina do complexo entre a elastase leucocitária humana e o terceiro domínio do inibidor de ovo de peru. Os organismos envolvidos são Homo sapiens e Meleagris gallopavo.
462	1P69	3.10	A entrada 1P69 apresenta a estrutura do complexo entre o adenovírus humano 12 e o receptor celular CAR, com uma mutação P417S, obtida por cristalografia de raios X. O sistema de expressão foi Escherichia coli BL21(DE3).
463	2B10	2.80	A entrada 2B10 descreve a estrutura cristalina do complexo entre o citochrome c mutante F82S e a citochrome c peroxidase de *Saccharomyces cerevisiae*, obtida por cristalografia de raios X. O sistema de expressão utilizado foi Escherichia coli.
464	1FCC	2.50	A proteína 1FCC do PDB SKEMI é um complexo entre o anticorpo Fab D1.3 e a lisozima de galinha. Esse complexo é amplamente estudado como modelo de interação antígeno-anticorpo, fornecendo insights sobre especificidade e afinidade molecular
465	2GYK	1.60	A entrada 2GYK descreve a estrutura cristalina do complexo entre o domínio DNase da colicina E9 e a proteína de imunidade mutante IMME9 (D51A), obtida por cristalografia de raios X. O sistema de expressão utilizado foi Escherichia coli BL21(DE3).
466	4EKD	2.71	A entrada 4EKD descreve a estrutura de RGS2 humano em complexos com a proteína mutante Galpha-q(R183C) de camundongo, obtida por cristalografia de raios X. O sistema de expressão utilizado foi Trichoplusia ni e Escherichia coli.
467	1EAW	2.93	A entrada 1EAW descreve a estrutura do complexo entre MTSP1 (matriptase) humano e BPTI (aprotinina) de boi, determinada por cristalografia de raios X. O sistema de expressão utilizado foi Saccharomyces cerevisiae e Escherichia coli.

468	1SGY	1,80	A entrada 1SGY descreve a estrutura do complexo entre o mutante TYR 18 do domínio 3 do inibidor de ovo de peru e proteína B da Streptomyces griseus, obtida por cristalografia de raios X. A expressão foi realizada em Escherichia coli.
469	2OMU	1.80	complexo proteína-proteína HEC1 com INLA G194S+S Y369S, resolvida por cristalografia de raios-X
470	2B0Z	2.70	Complexo proteína-proteína entre citocromo c peroxidase e F82I citocromo c. Parte do *general set* do PDBbind.
471	2OMZ	1.60	A estrutura do complexo InlA Y369A/hEC1 de *Listeria monocytogenes* com E-caderina humana foi determinada por cristalografia de raios-X , mostrando um complexo heterodimérico crucial para a adesão celular.
472	2OMT	2.00	Complexo proteína-proteína entre HEC1 e INLA G194S+S. Parte do *general set* do PDBbind.
473	1CHO	1.80	Estrutura do complexo entre a alfa-quimotripsina bovina e o terceiro domínio do ovomucoide de peru
474	4HRN	2.65	Estrutura de proteína sintética ligada ao domínio extracelular de HER2, com potencial efeito citotóxico em células cancerígenas
475	1P6A	2.90	Estrutura do domínio 1 do receptor Coxsackievirus e Adenovírus (CAR D1) em complexo proteína-proteína
476	3EQS	1.65	Estrutura da E3 ubiquitina-proteína ligase Mdm2 em complexo proteína-ligante.
477	1MI5	2.50	Complexo TCR/pMHC envolvendo HLA-B8 e peptídeo EB com LC13 TCR. Método: SPR.
478	4J2L	3.15	Complexo proteína-proteína envolvendo o domínio AXH de Ataxina-1 (ATXN1)
479	1XGT	2.10	Estrutura do mutante Y33L do anticorpo HyHEL-63 complexado com lisozima de ovo de galinha.
480	1OHZ	2.20	Complexo cohesina-dockerina do celulossomo de *Clostridium thermocellum*
481	1EZU	2.40	Ecotin Y69F, D70P ligado à tripsina D102N
482	3U82	3.16	Complexo proteína-proteína com glicoproteína D do vírus herpes simplex
483	1BVK	2.70	Estrutura do complexo entre o anticorpo humanizado anti-lysozyme e lysozyme. Método: Difração de raios-X
484	1M9E	2.20	Estrutura da glutationa S-transferase complexada com o ligante VWW
485	2J0T	2.54	Estrutura cristalográfica da MMP-1 em complexo com TIMP-1
486	1AKJ	2.65	Complexo entre HLA-A2 e o coreceptor CD8. Organismo: Homo sapiens, HIV. Método: Difração de raios X.
487	3CPH	2.90	Estrutura cristalográfica do complexo Sec4-Rab-GDI. Organismo: *Saccharomyces cerevisiae*. Método: Difração de raios X

488	1B6C	2.60	Estrutura cristalográfica do domínio citoplasmático do receptor tipo I de TGF-β em complexo com FKBP12. Organismo: *Homo sapiens*. Método: Difração de raios X.
489	1F47	1.95	Estrutura do protein ZipA/M185 em complexo com o ligante 17-mer. Organismo: *Escherichia coli*. Método: Difração de raios X.
490	1SBN	1.95	Estrutura do protein ZipA/M185 em complexo com o ligante 17-mer. Organismo: *Escherichia coli*. Método: Difração de raios X.
491	1DQJ	2.00	Estrutura do anticorpo anti-lisozima HyHEL-63 complexo com lisozima de clara de ovo de galinha. Organismos: *Mus musculus* e *Gallus gallus*. Método: Difração de raios X.
492	1AK4	2.36	Estrutura cristalina da ciclofilina humana ligada ao domínio amino-terminal da cápside do HIV-1. Organismos: *Homo sapiens* e HIV-1.. Método: Difração de raios X.
493	1FR2	1.60	Estrutura cristalina do domínio DNase E9 com uma proteína de imunidade mutante IM9 (E41A). Organismo: *Escherichia coli*. Método: Difração de raios X.
494	1DVF	1.90	Estrutura cristalina do complexo Fv-Fv idiotipo-anti-idiotipo. Organismo: *Mus musculus*. Método: Difração de raios X.
495	1CT0	1.80	Estrutura cristalina do variante OMTKY3-SER18I do OMTKY3 em complexo com SGPB. Organismos: *Streptomyces griseus*, *Meleagris gallopavo*. Método: Difração de raios X.
496	2OMX	1.70	Complexo proteína-proteína com o nome da proteína HEC1 e o ligante INLA S192N G194S+S. Subconjunto PDBbind: conjunto geral.
497	5M2O	1.26	A estrutura do terceiro cohesin ScaB de *Ruminococcus flavefaciens* em complexo com um dockerin do grupo 1. A proteína foi expressa em *Escherichia coli* BL21(DE3). Método experimentaL: Difração de raios X.
498	2J1K	2.30	A estrutura do cabeçalho da fibra CAV-2 em complexo com a região D1 do coxsackievirus e adenovirus receptor (CAR D1). A proteína foi extraída de *Homo sapiens* e do *Canine adenovirus 2*. Método: Difração de raios X
499	1CT2	1.65	Estrutura cristalina da variante P1 do OMTKY3, OMTKY3-THR18I, em complexo com SGPB. A proteína foi extraída de *Streptomyces griseus* e *Meleagris gallopavo*. Método: Difração de raios X.
500	1M10	3.10	Estrutura cristalográfica do complexo entre a glicoproteína Ib alfa e o domínio A1 do fator de von Willebrand. Classificação: coagulação sanguínea. Organismo: *Homo sapiens*.
501	2KSO	-	Complexo EphA2:SHIP2 SAM:SAM. Classificação: ligação proteica. Organismo: *Homo sapiens*. Método: RMN em solução.
502	4YFD	3.25	Estrutura cristalográfica do complexo PTP delta Ig1-Fn2 com IL-1RAcP. Organismo: *Mus musculus*. Método: Difração de raios X.

503	1KAC	2.60	Estrutura cristalográfica do domínio *knob* do adenovírus 12 em complexo com o domínio 1 do receptor celular CAR. Organismos: *Human adenovirus 12, Homo sapiens*. Método: Difração de raios X.
504	1SIB	2.40	Estruturas cristalinas refinadas do subtilisina novo em complexo com eglinas selvagens e duas mutantes. Organismos: *Bacillus subtilis*, *Hirudo medicinalis*. Método: Difração de raios X.
505	4KRL	2.85	Nanobody/VHH domain 7D12 em complexo com o domínio III da região extracelular do EGFR. Método: Difração de raios X. pH 6.0.
506	2DVW	2.30	A oncoproteína Gankyrin, envolvida na regulação do ciclo celular e degradação proteica, forma um complexo com a ATPase S6 do proteassomo 26S, sendo essencial para a ubiquitinação de proteínas.
507	1REW	1.86	A estrutura refinada do complexo da proteína morfogenética óssea 2 (BMP-2) com seu receptor tipo IA foi determinada por difração de raios-X, revelando detalhes da interação envolvida na sinalização óssea.
508	2NU1	1.80	Estruturas moleculares de complexos da SGPB com variantes aromáticas P1 do inibidor OMTKY3, determinadas por difração de raios-X, destacando interações enzimáticas em Streptomyces griseus.
509	4L3E	2.56	Estrutura do complexo entre o receptor de células T de alta afinidade DMF5 e o MHC de Classe I HLA-A2 ligado ao peptídeo MART-1(26-35)(A27L), determinada por difração de raios-X.
510	1XQS	2.90	Estrutura cristalina do domínio central da HspBP1 complexado com um fragmento do domínio ATPase da Hsp70, elucidando sua interação na regulação da chaperona Hsp70.
511	4E6K	2.00	Estrutura cristalina do bacterioferritina (BfrB) de *Pseudomonas aeruginosa* em complexo com a ferredoxina associada à bacterioferritina (Bfd), revelando detalhes da interação envolvida no metabolismo do ferro.
512	1FSS	3.00	Estrutura cristalina da acetilcolinesterase complexada com a toxina fasciculina-II, elucidando a inibição enzimática por peptídeos de *Dendroaspis angusticeps*.
513	1Z7X	1.95	Estrutura cristalina do inibidor de ribonuclease humana complexado com a ribonuclease I, revelando detalhes da inibição enzimática e regulação da degradação de RNA.
514	5K39	1.98	Estrutura do complexo de dockerina do coesina tipo II HE de *Clostridium thermocellum*, determinada por difração de raios-X, revelando detalhes estruturais cruciais para a adesão e organização de complexos proteicos.
515	2AK4	2.50	Estrutura cristalina do TCR SB27 em complexo com o peptídeo de 13 aminoácidos HLA-B*3508, revelando a interação entre o receptor de células T e o MHC.

516	2VIS	3.25	Estrutura cristalina do mutante de hemaglutinina do vírus influenza (A/X-31(H3N2)) com a substituição de Thr 131 por Ile, complexado com um anticorpo neutralizante, revelando mecanismos de escape imunológico.
517	1MQ8	3.3	Estrutura cristalina do domínio I da alfaL em complexo com ICAM-1, revelando a interação crucial para a adesão celular no sistema imunológico, determinada por difração de raios-X.
518	3BTF	1.80	Estruturas cristalinas dos complexos entre a beta-tripsina bovina e dez variantes P1 do inibidor BPTI, elucidando a interação entre a enzima hidrolase e seu inibidor, determinada por difração de raios-X.
519	4JFF	2.43	Estrutura do complexo entre o receptor de células T (TCR) e um ligante de 10-mer, pertencente ao conjunto geral da base de dados PDBbind, sem classificação específica de EC.
520	1SGQ	1.90	Estrutura cristalina do variante Gly 18 do inibidor de ovo de peru, terceiro domínio, complexado com a proteinase B de *Streptomyces griseus*, revelando a interação entre a serina protease e seu inibidor.
521	3SZK	3.01	Estrutura cristalina da met-hemoglobina humana complexada com o primeiro domínio NEAT da IsdH de *Staphylococcus aureus*, revelando a interação entre a hemoglobina humana e proteínas bacterianas envolvidas na captação de ferro.
522	1MLC	2.50	Estrutura cristalina do anticorpo monoclonal Fab D44.1, gerado contra a lisozima da clara de ovo de galinha, complexado com a lisozima, revelando os detalhes da interação anticorpo-antígeno.
523	1SGD	1.80	Estrutura cristalina do variante Asp 18 do inibidor de ovo de peru, terceiro domínio, complexado com a proteinase B de *Streptomyces griseus* a pH 6,5, elucidando a interação entre a serina protease e seu inibidor.
524	1CSO	1.90	Estrutura cristalina do variante OMTKY3-ILE18I, complexado com a SGPB, revelando a interação entre o inibidor de hidrolase OMTKY3 e a serina protease SGPB, determinada por difração de raios-X.
525	1FQJ	2.02	Estrutura cristalina do complexo heterotrimérico do domínio RGS do RGS9, a subunidade gama da fosfodiesterase e a subunidade alfa do quimérico GT/I1, revelando detalhes sobre a sinalização celular mediada por proteínas, determinada por difração de raios-X.
526	4O27	3.18	Estrutura cristalina do complexo MST3-MO25 com o motivo WIF, revelando os detalhes estruturais da interação envolvida na sinalização celular humana.

APÊNDICE B – IMPLEMENTAÇÃO DO SUPER LEARNER E MODELOS INDIVIDUAIS

B.1 Treinamento dos Modelos Individuais

O código a seguir implementa e avalia diversos modelos de regressão utilizando validação cruzada k-fold (k=10) para prever a energia livre de Gibbs de interação proteína-proteína. Os modelos são treinados e avaliados com base em métricas como coeficiente de determinação (R^2), correlação de Pearson (r) e erro quadrático médio (RMSE). Após a avaliação, os modelos são treinados com todos os dados e armazenados para uso posterior.

```
import pandas as pd
import numpy as np
import pickle
from sklearn.model_selection import KFold
from sklearn.metrics import r2_score, mean_squared_error
from scipy.stats import pearsonr
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression, ElasticNet
from sklearn.neighbors import KNeighborsRegressor
from sklearn.tree import DecisionTreeRegressor
from sklearn.svm import SVR
from sklearn.ensemble import AdaBoostRegressor, BaggingRegressor,
RandomForestRegressor, ExtraTreesRegressor
import xgboost as xgb

# Carregar os dados de treinamento
dados_train = pd.read_csv('v_ofc_train_dados_pbee.csv')
X_train = dados_train.drop(columns=['pdb', 'database', 'partner1',
'partner2', 'dG_exp']).values
y_train = dados_train["dG_exp"].values

# Carregar os dados de teste
dados_test = pd.read_csv('v_ofc_test_dados_pbee.csv')
X_test = dados_test.drop(columns=['pdb', 'database', 'partner1', 'partner2',
'dG_exp']).values
y_test = dados_test["dG_exp"].values
```

```
# Função para calcular as métricas de avaliação
def calcular_metricas(y_real, y_pred):
    r2 = r2_score(y_real, y_pred)
    rmse = np.sqrt(mean_squared_error(y_real, y_pred))
    r, _ = pearsonr(y_real, y_pred)
    return r2, r, rmse

# Lista de modelos de regressão e hiperparâmetros otimizados
modelos = {
    "Regressao_Linear": LinearRegression(fit_intercept=False,
positive=False),
    "Elastic_Net": ElasticNet(max_iter=1000000),
    "KNN_Regressor": KNeighborsRegressor(n_neighbors=10, p=1,
weights="distance"),
    "Decision_Tree_Regressor": DecisionTreeRegressor(max_depth=10,
max_features="sqrt", min_samples_leaf=8, min_samples_split=5),
    "SVR": SVR(gamma='scale'),
    "AdaBoost_Regressor": AdaBoostRegressor(learning_rate=0.5,
n_estimators=200),
    "Bagging_Regressor": BaggingRegressor(bootstrap=True, max_features=0.5,
max_samples=1.0, n_estimators=100),
    "Random_Forest_Regressor": RandomForestRegressor(max_features="log2",
min_samples_leaf=1, min_samples_split=2, n_estimators=1000),
    "Extra_Trees_Regressor": ExtraTreesRegressor(n_estimators=1000),
    "XGBoost": xgb.XGBRegressor(objective='reg:squarederror')
}

# Validação cruzada K-fold
kf = KFold(n_splits=10, shuffle=True, random_state=42)

# Dicionário para armazenar métricas de cada modelo
resultados = {}

for nome, modelo in modelos.items():
    r2_scores, r_scores, rmse_scores = [], [], []

    for train_idx, val_idx in kf.split(X_train):
        X_k_train, X_k_val = X_train[train_idx], X_train[val_idx]
        y_k_train, y_k_val = y_train[train_idx], y_train[val_idx]
```

```
        modelo.fit(X_k_train, y_k_train)
        y_k_pred = modelo.predict(X_k_val)

        r2, r, rmse = calcular_metricas(y_k_val, y_k_pred)
        r2_scores.append(r2)
        r_scores.append(r)
        rmse_scores.append(rmse)

    # Média das métricas de validação cruzada
    resultados[nome] = {
        "R²": np.mean(r2_scores),
        "r": np.mean(r_scores),
        "RMSE": np.mean(rmse_scores)
    }

    # Treinar o modelo final com todo o conjunto de treinamento
    modelo.fit(X_train, y_train)

    # Salvar o modelo treinado
    with open(f'{nome}.pkl', 'wb') as f:
        pickle.dump(modelo, f)

# Avaliação no conjunto de teste final
print("\nDesempenho no conjunto de teste:")
for nome, modelo in modelos.items():
    y_pred_test = modelo.predict(X_test)
    r2, r, rmse = calcular_metricas(y_test, y_pred_test)
    print(f"{nome}: R² = {r2:.4f}, r = {r:.4f}, RMSE = {rmse:.4f}")
```

Fonte: O autor (2024)

B.2 Implementação do Super Learner

O Super Learner é treinado utilizando predições fora da amostra dos modelos individuais como entrada para um meta-modelo, o qual faz a predição final.

```
from math import sqrt
from numpy import hstack, vstack, asarray
import pandas as pd
from sklearn.model_selection import KFold
from sklearn.metrics import r2_score, mean_squared_error
```

```
from sklearn.linear_model import LinearRegression, Ridge
from sklearn.neural_network import MLPRegressor
from sklearn.ensemble import ExtraTreesRegressor
from xgboost import XGBRegressor
from sklearn.linear_model import ElasticNet
import joblib
import pickle
import matplotlib.pyplot as plt
from scipy.stats import pearsonr

# Carregar os dados de treinamento
train_data = pd.read_csv('dados_s_outliers.csv')
X = train_data.drop(columns=['pdb', 'database', 'partner1', 'partner2',
'dG_exp']).values
y = train_data['dG_exp'].values

# Lista de nomes dos modelos base
model_names = [
       'Regressao_Linear', 'Elastic_Net', 'SVR', 'Decision_Tree_Regressor',
'KNN_Regressor',
       'AdaBoost_Regressor', 'Bagging_Regressor', 'Random_Forest_Regressor',
'Extra_Trees_Regressor', 'XGBoost'
]

# Função para carregar os modelos base
def load_models():
    models = []
    for name in model_names:
        try:
            model = joblib.load(f'{name}.pkl')
            models.append((name, model))
        except Exception as e:
            print(f"Erro ao carregar {name}: {e}")
    return models

# Obter predições fora da amostra sem re-treinar os modelos
def get_out_of_fold_predictions(X, y, models, n_splits=10):
    meta_X, meta_y = list(), list()
    kfold = KFold(n_splits=n_splits, shuffle=True, random_state=42)

    for train_ix, test_ix in kfold.split(X):
```

```
        fold_yhats = list()
        test_X, test_y = X[test_ix], y[test_ix]
        meta_y.extend(test_y)

        for name, model in models:
            try:
                yhat = model.predict(test_X)  # Apenas predição
                fold_yhats.append(yhat.reshape(len(yhat), 1))
            except Exception as e:
                print(f"Erro ao predizer com {name}: {e}")

        meta_X.append(hstack(fold_yhats))

    return vstack(meta_X), asarray(meta_y)

# Carregar modelos base
models = load_models()

# Obter predições fora da amostra
meta_X, meta_y = get_out_of_fold_predictions(X, y, models)
print('Meta ', meta_X.shape, meta_y.shape)

# Salvar os dados de entrada do meta-modelo para reuso
pd.DataFrame(meta_X).to_csv('meta_X.csv', index=False)
pd.DataFrame({'meta_y': meta_y}).to_csv('meta_y.csv', index=False)

# Lista de meta-modelos com hiperparâmetros otimizados
meta_models = {
    'LinearRegression': LinearRegression(),
    'ElasticNet': ElasticNet(alpha=00.1, l1_ratio=0.5, max_iter=100000),
     'XGBoost': XGBRegressor(n_estimators=100, learning_rate=0.1, max_depth=3,
random_state=42),
    'ExtraTrees': ExtraTreesRegressor(n_estimators=100, random_state=42),
    'Ridge': Ridge(alpha=0.01),
          'MLP':  MLPRegressor(hidden_layer_sizes=(50,),  activation='relu',
solver='lbfgs', max_iter=500, random_state=42)
}

# Treinar e salvar cada meta-modelo
for name, meta_model in meta_models.items():
    meta_model.fit(meta_X, meta_y)
```

```
with open(f'super_learner_{name}.pkl', 'wb') as f:
    pickle.dump(meta_model, f)
print(f'Meta-modelo {name} treinado e salvo!')
```

Fonte: O autor (2024)

APÊNDICE C – TABELA C1: ESTRUTURAS REMOVIDAS E SEU RESPECTIVO MÉTODO DE IDENTIFICAÇÃO

	Método de identificação	
ID PDB	**IQR ∩ IForest**	**Gaussiana (I.C. 99,7%)**
1mi5	X	
3cph	X	
4yfd	X	
3tsr	X	
1dfj	X	
3onw	X	
1bvn	X	
2omx	x	
1dqj	x	
4y61	x	
2vis	x	
1kxv	x	
2wwx	x	
2kso	x	
4o27	x	x
4did	x	
1ezu	x	
4jff	x	
2b0z	x	
1akj	x	
4z80	x	
1xgt	x	
2omz	x	
4krl	x	
1ava	x	
6bxc	x	
1z7x	x	
1mlc	x	
1e6e	x	
2omt	x	
3zeu	x	

2vir	x		
2b7c	x		
2omu	x		
2omw	x		
3d1m			x
2ptt			x

Fonte: O autor (2024)

APÊNDICE D – DESEMPENHO DOS MODELOS POR BASE DE DADOS

	Teste set completo			Pdbind			Benchmark			Skempi		
Modelos	r	R^2	RMSE	r	R^2	RMSE	r	R^2	RMSE	r	R^2	RMSE
AdaBoost	0.61	0.35	2.13	0.63	0.36	2.10	0.63	0.31	2.29	0.20	-0.06	2.35
Bagging	0.64	0.39	2.06	0.66	0.41	2.01	0.71	0.37	2.19	0.07	-0.35	2.65
D. Tree	0.33	-0.17	2.86	0.30	-0.25	2.93	0.65	0.28	2.34	0.52	-0.10	2.40
ElasticNet	0.59	0.34	2.14	0.56	0.30	2.19	0.87	0.74	1.42	0.61	-0.03	2.32
Extra Trees	0.70	0.48	1.91	0.71	0.49	1.88	0.75	0.45	2.04	0.36	0.05	2.23
KNN	0.62	0.38	2.09	0.62	0.38	2.07	0.54	0.27	2.35	0.54	0.28	1.93
R. Forest	0.67	0.44	1.98	0.69	0.46	1.93	0.65	0.34	2.25	0.08	-0.13	2.43
Reg. Linear	0.65	0.41	2.04	0.62	0.37	2.08	0.87	0.69	1.53	0.65	0.24	1.99
SVR	0.15	0.02	2.62	0.10	0.00	2.63	0.38	-0.03	2.81	0.65	0.19	2.06
XGBoost	0.66	0.40	2.05	0.68	0.41	2.01	0.60	0.31	2.29	0.51	0.11	2.15
SL Elasticnet	0.69	0.47	1.92	0.70	0.48	1.89	0.67	0.38	2.18	0.51	0.21	2.03
SL ET	0.69	0.47	1.92	0.70	0.48	1.89	0.68	0.39	2.16	0.50	0.20	2.04
SL Reg. Linear	0.62	0.38	2.09	0.62	0.38	2.07	0.54	0.28	2.35	0.54	0.28	1.93
SL MLP	0.70	0.48	1.91	0.71	0.49	1.88	0.71	0.41	2.12	0.47	0.16	2.09
SL Ridge	0.69	0.47	1.92	0.70	0.48	1.89	0.67	0.38	2.17	0.51	0.21	2.03
SL XGBoost	0.63	0.39	2.07	0.63	0.39	2.05	0.55	0.29	2.33	0.52	0.27	1.95

Fonte: O Autor (2024)

APÊNDICE E – COMPARAÇÃO ENTRE OS MODELOS COM E SEM OUTLIERS NOS DADOS DE TREINAMENTO

	Dados s/outliers			Dados c/ outliers		
Modelos	**r**	**R²**	**RMSE**	**r**	**R²**	**RMSE**
AdaBoost	0.59	0.35	2.14	0.61	0.35	2.13
Bagging	0.67	0.44	1.98	0.64	0.39	2.06
D. Tree	0.48	0.10	2.50	0.33	-0.17	2.86
ElasticNet	0.57	0.32	2.18	0.59	0.34	2.14
Extra Trees	0.67	0.43	1.99	0.70	0.48	1.91
KNN	0.61	0.37	2.10	0.62	0.38	2.09
R. Forest	0.66	0.43	2.00	0.67	0.44	1.98
Reg. Linear	0.61	0.32	2.17	0.65	0.41	2.04
SVR	0.16	0.02	2.62	0.15	0.02	2.62
XGBoost	0.63	0.36	2.11	0.66	0.40	2.05
SL Elasticnet	0.67	0.45	1.97	0.69	0.47	1.92
SL ET	0.67	0.44	1.97	0.69	0.47	1.92
SL Reg. Linear	0.61	0.37	2.10	0.62	0.38	2.09
SL MLP	0.68	0.45	1.96	0.70	0.48	1.91
SL Ridge	0.67	0.45	1.97	0.69	0.47	1.92
SL XGBoost	0.62	0.37	2.09	0.63	0.39	2.07

Fonte: O autor (2024)

APÊNDICE F – RANKING DAS VARIÁVEIS POR IMPORTÂNCIA SHAP NORMALIZADA E ACUMULADA

Variável	Importância SHAP	Importância Normalizada	Importância Acumulada
ifa_packstat	0.29	9.37	9.37
dslf_fa13	0.23	7.48	16.84
hbond_bb_sc	0.21	6.81	23.65
Ref	0.18	5.72	29.38
ifa_hbonds_int	0.17	5.36	34.74
Cms	0.16	5.31	40.05
ifa_nres_int	0.10	3.17	43.23
hbond_lr_bb	0.08	2.71	45.93
ifa_side1_score	0.07	2.36	48.30
lk_ball_bridge_uncpl	0.06	2.00	50.30
ifa_dG_separated	0.06	1.98	52.28
ifa_dSASA_hphobic	0.06	1.98	54.26
ifa_dSASA_polar	0.06	1.95	56.21
ifa_dSASA_int	0.06	1.87	58.09
fa_intra_atr_xover4	0.06	1.81	59.90
p_aa_pp	0.05	1.72	61.62
lk_ball_bridge	0.05	1.70	63.32
interaction_energy	0.05	1.70	65.02
Ômega	0.05	1.66	66.68
ifa_side2_normalized	0.05	1.57	68.25
hbond_sr_bb	0.05	1.54	69.79
ifa_delta_unsatHbonds	0.05	1.49	71.28
pro_close	0.05	1.47	72.76
rama_prepro	0.04	1.39	74.15
ifa_dG_cross	0.04	1.38	75.53
fa_intra_rep_xover4	0.04	1.37	76.89
fa_intra_sol_xover4	0.04	1.34	78.24
hbond_sc	0.04	1.32	79.56
fa_sol	0.04	1.27	80.83
fa_dun_rot	0.04	1.26	82.09
ifa_side2_score	0.04	1.25	83.34
fa_dun_semi	0.04	1.21	84.55
fa_intra_elec	0.04	1.18	85.74

lk_ball_iso	0.04	1.16	86.89
ifa_dG_separated_dSASAx100	0.03	1.04	87.93
ifa_side1_normalized	0.03	1.03	88.96
fa_rep	0.03	1.01	89.97
lk_ball	0.03	0.99	90.96
ifa_nres_all	0.03	0.97	91.92
ifa_dG_cross_dSASAx100	0.03	0.94	92.86
ifa_hbond_E_fraction	0.03	0.93	93.80
fa_elec	0.03	0.85	94.64
fa_atr	0.03	0.85	95.49
ifa_per_residue_energy_int	0.03	0.83	96.32
hxl_tors	0.03	0.82	97.15
total_score	0.02	0.81	97.96
ifa_sc_value	0.02	0.80	98.76
fa_dun_dev	0.02	0.63	99.38
ifa_complex_normalized	0.02	0.62	100.00

Fonte: O Autor (2025)

APÊNDICE G - TABELA CONTENDO R^2, r e RMSE PARA O CONJUNTO DE TREINAMENTO E TESTE DOS MODELOS INDIVIDUAIS E METAMODELOS

Modelo	Conjunto Treinamento			Conjunto Teste		
	R^2	r	RMSE	R^2	r	RMSE
Reg. Linear	0,379 ± 0,000	0,616 ± 0,000	2,131 ± 0,000	0,406 ± 0,000	0,653 ± 0,000	2,036 ± 0,000
ElasticNet	0,330 ± 0,000	0,577 ± 0,000	2,214 ± 0,000	0,342 ± 0,000	0,594 ± 0,000	2,144 ± 0,000
KNN_Regressor	1,000 ± 0,000	1,000 ± 0,000	0,000 ± 0,000	0,376 ± 0,000	0,620 ± 0,000	2,087 ± 0,000
Decision Tree	0,546 ± 0,052	0,738 ± 0,035	1,822 ± 0,105	0,001 ± 0,272	0,407 ± 0,160	2,636 ± 0,356
SVR	0,065 ± 0,000	0,266 ± 0,000	2,616 ± 0,000	0,019 ± 0,000	0,150 ± 0,000	2,618 ± 0,000
AdaBoost	0,585 ± 0,010	0,794 ± 0,009	1,742 ± 0,022	0,335 ± 0,028	0,593 ± 0,022	2,156 ± 0,045
Bagging	0,902 ± 0,006	0,976 ± 0,002	0,848 ± 0,025	0,408 ± 0,037	0,650 ± 0,031	2,034 ± 0,064
Random Forest	0,907 ± 0,002	0,981 ± 0,001	0,823 ± 0,009	0,444 ± 0,011	0,679 ± 0,008	1,970 ± 0,020
Extra Trees	1,000 ± 0,000	1,000 ± 0,000	0,000 ± 0,000	0,473 ± 0,008	0,693 ± 0,006	1,919 ± 0,015
XGBoost	1,000 ± 0,000	1,000 ± 0,000	0,002 ± 0,000	0,401 ± 0,000	0,662 ± 0,000	2,046 ± 0,000
SL_Reg. Linear	1,000 ± 0,000	1,000 ± 0,000	0,000 ± 0,000	0,449 ± 0,000	0,672 ± 0,000	1,962 ± 0,000
SL_ElasticNet	1,000 ± 0,000	1,000 ± 0,000	0,002 ± 0,000	0,469 ± 0,000	0,688 ± 0,000	1,926 ± 0,000
SL_XGBoost	1,000 ± 0,000	1,000 ± 0,000	0,026 ± 0,000	0,390 ± 0,000	0,631 ± 0,000	2,064 ± 0,000
SL_ExtraTrees	1,000 ± 0,000	1,000 ± 0,000	0,000 ± 0,000	0,464 ± 0,002	0,685 ± 0,002	1,935 ± 0,004
SL_Ridge	1,000 ± 0,000	1,000 ± 0,000	0,001 ± 0,000	0,468 ± 0,000	0,688 ± 0,000	1,927 ± 0,000
SL_MLP	1,000 ± 0,000	1,000 ± 0,000	0,002 ± 0,001	0,474 ± 0,000	0,692 ± 0,000	1,917 ± 0,001

APÊNDICE H - TESTE DE PERMUTAÇÃO PARA AS MÉTRICAS DOS MODELOS INDIVIDUAIS.

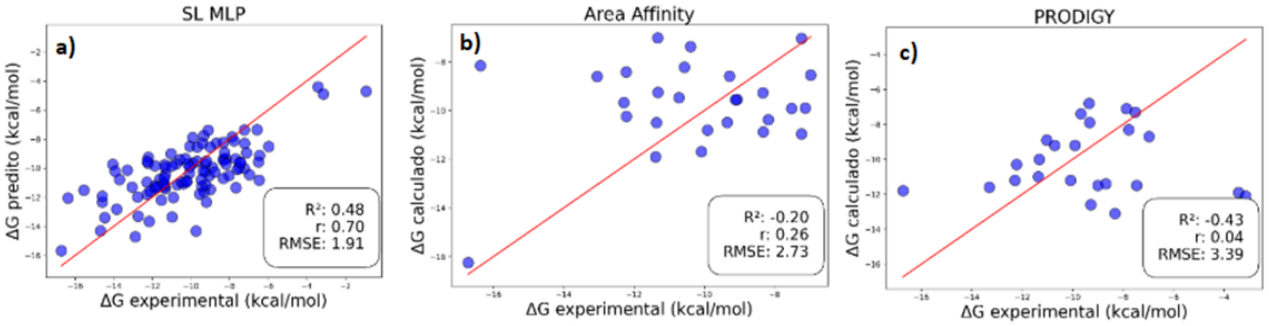
Modelo	RMSE_obs.	p_valor_RMSE	R2_obs	p_valor_R2	Pearson_obs	p_valor_Pearson
Regressao_Linear	2,04E+16	0,000	0,406	0,000	0,653	0,000
Elastic_Net	2,14E+15	0,000	0,342	0,000	0,594	0,000
KNN	2,09E+14	0,000	0,376	0,000	0,620	0,000
Decision Tree	2,81E+16	0,004	-0,128	0,004	0,272	0,008
SVR	2,62E+16	0,067	0,019	0,067	0,150	0,116
AdaBoost_Regressor	2,15E+16	0,000	0,340	0,000	0,599	0,000
Bagging_	1,96E+15	0,000	0,449	0,000	0,682	0,000
RF	1,97E+16	0,000	0,447	0,000	0,680	0,000
ET	1,93E+15	0,000	0,469	0,000	0,690	0,000
XGBoost	2,05E+16	0,000	0,401	0,000	0,662	0,000

APÊNDICE I - RESUMO SIMPLIFICADO ESTILO NOTA DE IMPRENSA

Pesquisador do DQF/UFPE contribui para avanço na predição sobre a interação entre proteínas com inteligência artificial

A dissertação de mestrado do pesquisador Manassés F. Nascimento Filho, desenvolvida no Programa de Pós-Graduação em Química da Universidade Federal de Pernambuco (UFPE), com apoio financeiro da bolsa CAPES, contribuiu para o aprimoramento de métodos que ajudam a prever como proteínas interagem entre si, etapa fundamental para o desenvolvimento de novos medicamentos, vacinas e terapias. O trabalho foi orientado pelo prof. Dr. Roberto D. Lins, e coorientado pelo Dr. Elton J. F. Chaves.

A pesquisa concentrou-se na avaliação de diferentes modelos de regressão aplicados à predição da afinidade entre proteínas, ou seja, à estimativa da energia envolvida na interação entre essas moléculas. O estudo teve como objetivo apoiar o desenvolvimento de um metamodelo computacional mais preciso e eficiente, baseado na abordagem *Super Learne*r (SL). Esse metamodelo foi incorporado ao PBEE (*Protein Binding Energy Estimator*), *software* criado pelo grupo de pesquisa para auxiliar na triagem molecular, com aplicações em identificação de compostos, avaliação de terapias com anticorpos, análise de mutações em antígenos e design de proteínas. Esse esforço resultou na publicação do artigo "*Estimating Absolute Protein–Protein Binding Free Energies by a Super Learner Model*", em 2025, no prestigiado periódico *Journal of Chemical Information and Modeling* (ACS), que apresentou um método com acurácia sem precedentes e baixo custo computacional em comparação com outras abordagens de aprendizado de máquina disponíveis.



O estudo traz benefícios potenciais para diversas áreas da saúde e da biotecnologia, como o combate a doenças e o desenvolvimento de novas terapias, reforçando a contribuição da UFPE para a ciência e inovação no Brasil e no mundo.

ANEXO A – EXEMPLO DE SCRIPT XML PARA ROSETTA E TABELA AA1 CONTENDO OS DESCRITORES

O RosettaScript foi utilizado para calcular os descritores empregados neste trabalho. A seguir, apresenta-se um exemplo de script XML utilizado no software Rosetta. Esse script delineia um pipeline para calcular as propriedades definidas na Tabela AA1.

O script inicia com a definição das funções de pontuação a serem utilizadas, que atribuem pesos diferentes a termos energéticos relacionados às interações proteicas. Em seguida, são configurados os seletores de resíduos para identificar as cadeias proteicas de interesse, denominadas "*partner1*" e "*partner2*". Também são definidas operações de tarefa para inicializar parâmetros a partir da linha de comando.

A próxima seção do script introduz métricas simples que avaliam aspectos específicos da estrutura proteica, como a energia de interação entre resíduos selecionados das cadeias "*partner1*" e "*partner2*". Na seção de filtros, uma série de filtros é definida para avaliar as propriedades estruturais e energéticas das interações. Esses filtros abrangem cálculos de complementaridade de forma, contatos de superfície molecular, buracos na interface e mudanças na energia livre (DDG).

Os "movers" são definidos para realizar minimização e análise da interface. Isso inclui minimização com parâmetros variados e diferentes níveis de detalhamento, além da análise da interface para avaliar características como empacotamento e ligação por hidrogênio.

Por fim, os "movers" e filtros são adicionados em uma sequência específica para executar a análise e o refinamento desejados, que neste caso incluem minimização, cálculo de métricas, análise da interface e aplicação de filtros para selecionar estruturas que atendam aos critérios predefinidos. Em resumo, o script XML representa um plano detalhado para análise e manipulação da estrutura proteica utilizando o Rosetta, delineando etapas voltadas à avaliação das interações entre cadeias proteicas, melhoria da qualidade estrutural e seleção de conformações que atendam aos critérios predefinidos.

```
<ROSETTASCRIPTS>
```

```
	<SCOREFXNS>
		<ScoreFunction name="beta" weights="beta_nov16"/>
	</SCOREFXNS>
	<RESIDUE_SELECTORS>
		<Chain name="partner1" chains="A"/>
		<Chain name="partner2" chains="B"/>
	</RESIDUE_SELECTORS>
	<TASKOPERATIONS>
		<InitializeFromCommandline name="init"/>
	</TASKOPERATIONS>
	<SIMPLE_METRICS>
		<InteractionEnergyMetric name="ie" residue_selector="partner1"
residue_selector2="partner2" scorefxn="beta"/>
	</SIMPLE_METRICS>
	<FILTERS>
		<ShapeComplementarity name="sc" residue_selector1="partner1"
residue_selector2="partner2" confidence="0"/>
		<ContactMolecularSurface name="cms" distance_weight="0.5"
target_selector="partner1" binder_selector="partner2 confidence="0"/>
		<InterfaceHoles name="holes" jump="1" confidence="0"/>
		<Ddg  name="ddg_filter1"  scorefxn="beta"  jump="1"  chain_num="2"
	repeats="1"	repack="0"	repack_bound="0"	repack_unbound="0"
	threshold="99999"
confidence="0"/>
	</FILTERS>
	<MOVERS>
		<MinMover name="min1" scorefxn="beta" jump="1" max_iter="50000"
tolerance="0.0001" cartesian="0" bb="0" chi="1" bb_task_operations="init"
chi_task_operations="init"/>
		<MinMover name="min2" scorefxn="beta" jump="1" max_iter="50000"
tolerance="0.0001" cartesian="0" bb="1" chi="1" bb_task_operations="init"
chi_task_operations="init"/>
		<InterfaceAnalyzerMover	name="ifa"	scorefxn="beta"
	interface="A_B" packstat="1" interface_sc="1" tracer="1"
scorefile_reporting_prefix="ifa"/>
		<RunSimpleMetrics name="iesum" metrics="ie"/>
	</MOVERS>
	<PROTOCOLS>
		<Add mover_name="min1"/>
		<Add mover_name="min2"/>
		<Add mover_name="iesum"/>
```

```
            <Add mover_name="ifa"/>
            <Add filter_name="ddg_filter1"/>
            <Add filter_name="ddg_filter2"/>
            <Add filter_name="cms"/>
            <Add filter_name="sc"/>
            <Add filter_name="holes"/>
      </PROTOCOLS>
</ROSETTASCRIPTS>
```

Fonte: Script desenvolvido pelo Drº Elton J. F. Chaves, adaptado para este trabalho.

Tabela AA1 Termos de Pontuação do Rosetta e suas Descrições

Termo de Pontuação do Rosetta (Descritor)	**Descrição**
dslf_fa13	Potencial geométrico para ligações dissulfeto.
fa_atr	Termo atrativo de Lennard-Jones entre átomos em resíduos diferentes.
fa_dun_dev	Energia interna dos rotâmeros da cadeia lateral, baseada nas estatísticas de Dunbrack, considerando a variação dos ângulos χ em relação à rotamericidade.
fa_dun_rot	Energia interna dos rotâmeros da cadeia lateral, baseada nas estatísticas de Dunbrack, considerando a frequência dos rotâmeros.
fa_dun_semi	Contraparte do fa_dun_rot para aminoácidos semi-rotaméricos.
fa_elec	Potencial eletrostático de Coulomb com um dielétrico dependente da distância.
fa_intra_atr_xover4	Atração de Lennard-Jones dentro do próprio resíduo, contada para pares de átomos além da relação de torção.
fa_intra_elec	Interação de Coulomb dentro do próprio resíduo, contada para pares de átomos além da relação de torção.
fa_intra_rep_xover4	Repulsão de Lennard-Jones dentro do próprio resíduo, contada para pares de átomos além da relação de torção.
fa_intra_sol_xover4	Solvatação LK dentro do próprio resíduo, contada para pares de átomos além da relação de torção.

fa_rep	Termo repulsivo de Lennard-Jones entre átomos em resíduos diferentes.
fa_sol	Energia de solvatação segundo o modelo Lazaridis-Karplus.
hbond_bb_sc	Energia de ligação de hidrogênio entre cadeia lateral e esqueleto principal.
hbond_lr_bb	Ligações de hidrogênio esqueleto-esqueleto distantes na sequência primária.
hbond_sc	Energia de ligação de hidrogênio entre cadeias laterais.
hbond_sr_bb	Ligações de hidrogênio esqueleto-esqueleto próximas na sequência primária.
hxl_tors	Preferência de torção do grupo hidroxila para Ser/Thr/Tyr, substituindo yhh_planarity (que cobre apenas L- e D-Tyr).
lk_ball	Contribuição anisotrópica para a solvatação.
lk_ball_bridge	Bônus de solvatação devido a pontes de moléculas de água, medido pelo sobreposição das ‘bolas’ de dois átomos polares interagentes. Suporta tipos de resíduos arbitrários.
lk_ball_bridge_uncpl	Mesmo que lk_ball_bridge, mas o valor é desacoplado da energia livre (dGfree), ou seja, um bônus constante. Suporta tipos de resíduos arbitrários.
lk_ball_iso	Mesmo que fa_sol; ver acima. Suporta tipos de resíduos arbitrários.
Ômega	Ângulo diedro ômega no esqueleto principal. Restrição harmônica de planaridade com desvio padrão de ~6°. Suporta α-aminoácidos, β-aminoácidos e oligouréas. Em oligouréas, ambas as ligações amida (‘mu’ e ‘omega’ no Rosetta) são restringidas à planaridade.
p_aa_pp	Probabilidade do aminoácido em φ/ψ.
pro_close	Energia de fechamento do anel da prolina e energia do ângulo ψ do resíduo precedente. Suporta prolina D- ou L-, além de oligouréas D- ou L-prolina.
rama_prepro	Termo de preferência de torção do esqueleto que considera se o aminoácido precedente é uma prolina ou não.

Ref	Energia de referência para cada aminoácido. Equilibra a energia interna dos aminoácidos e desempenha um papel no design de proteínas.
total_score	Soma das pontuações individuais dos resíduos no Rosetta.
Shape complementarity (sc)	Calcula a complementaridade de forma segundo Lawrence & Coleman.
InterfaceHoles (holes)	Diferença na pontuação de buracos entre as conformações ligada e não ligada.
Ddg	Calcula a energia de ligação do complexo.
InteractionEnergyMetric	Métrica para medir a energia de interação de curto e longo alcance entre resíduos usando dois conjuntos de seletores de resíduos.
complex_normalized	Energia média de um resíduo em todo o complexo.
dG_cross	Energia de ligação da interface calculada com termos de energia entre interfaces, em vez de separar fisicamente a interface. Pode ser impreciso devido a dependências ambientais de alguns termos de energia, incluindo ligações de hidrogênio e solvatação.
dG_cross/dSASAx100	dG_cross dividido por dSASA, multiplicado por 100.
dG_separated	Diferença na energia do Rosetta quando as cadeias que formam a interface são separadas versus quando estão complexadas: a energia de ligação. Calculado separando (e opcionalmente reempacotando) as cadeias.
dG_separated/dSASAx100	Energia de ligação separada por unidade de área de interface, multiplicada por 100 para ajustar as unidades no arquivo de pontuação. A escala por dSASA normaliza para interfaces maiores.
dSASA_hphobic	Parte apolar da energia de solvatação no modelo FACTS, parametrizada para aproximar a área solvente acessível atômica, em Å².
dSASA_int	Área solvente acessível enterrada na interface, em Å².

dSASA_polar	Parte polar da energia de solvatação quando as cadeias que formam a interface são separadas versus quando estão complexadas.
delta_unsatHbonds	Número de ligações de hidrogênio insatisfeitas e enterradas na interface.
hbond_E_fraction	Fração da energia da interface (dG_separated) devida a ligações de hidrogênio entre as interfaces.
hbonds_int	Total de ligações de hidrogênio entre interfaces encontradas.
nres_all	Número total de resíduos em todo o complexo.
nres_int	Número de resíduos na interface.
Packstat	Estatística de empacotamento do Rosetta para a interface (0 = ruim, 1 = perfeito).
per_residue_energy_int	Energia média de cada resíduo na interface.
sc_value	Complementaridade de forma.
side1_normalized	Energia média por resíduo de um dos lados da interface.
side1_score	Energia de um dos lados da interface.
side2_normalized	Energia média por resíduo do outro lado da interface.
side2_score	Energia do outro lado da interface.

Fonte: Adaptado, (CHAVES et al., 2023)