UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE FÍSICA DE SÃO CARLOS

André Gustavo Jucovski

Análise Multiparamétrica do Problema de Fases em Cristalografia de Proteínas por Aprendizado Profundo. Caso de estudo: Lisozima da Clara do Ovo de Galinha

São Carlos

2020

André Gustavo Jucovski

Análise Multiparamétrica do Problema de Fases em Cristalografia de Proteínas por Aprendizado Profundo. Caso de estudo: Lisozima da Clara do Ovo de Galinha

Trabalho de conclusão de curso apresentado ao Programa de Graduação em Física do Instituto de Física de São Carlos da Universidade de São Paulo, para obtenção do título de Bacharel em Física Biomolecular.

Orientador: Prof. Dr. Andre Luis Berteli Ambrosio

São Carlos

2020

Jucovski, André Gustavo

Análise multiparamétrica do problema de fases em cristalografia de proteínas por aprendizado profundo. Caso de estudo: Lisozima da clara do ovo de galinha / André Gustavo Jucovski; orientador Andre Luis Berteli Ambrosio – São Carlos, 2020.

25 p.

Trabalho de conclusão de curso (Graduação em Ciências Físicas e Biomoleculares) - Instituto de Física de São Carlos, Universidade de São Paulo, 2020.

1. Cristalografia. 2. Aprendizado profundo. 3. Problema de fases. 4. Difração de raios X. I. Ambrosio, Andre Luis Berteli, orient.

RESUMO

O problema de fases é notório na cristalografia de proteínas por difração de raios-X. A perda experimental de informações sobre as fases das ondas espalhadas construtivamente pelos componentes do cristal se devem à limitações tecnológicas intrínsecas aos sistemas de detecção dessa radiação. Assim, é impossibilitado o cálculo direto da função de distribuição de densidade eletrônica na cela unitária através de uma transformada de Fourier. Atualmente, há dois métodos experimentais que podem ser aplicados para contornar esse problema: (i) a quantificação seletiva do componente dispersivo (λ-dependente) do fator de espalhamento atômico ou (ii) a substituição parcial do solvente aquoso ordenado por íons mais elétron-densos (metálicos ou halogênicos). Alternativamente, informações prévias, na forma de estruturas cristalinas conhecidas que são funcionalmente relacionadas ou homólogas a componentes no cristal, podem servir como fonte de um conjunto inicial de fases. Apesar de desafiadoras, quando viáveis, as aplicações desses diferentes métodos já possibilitaram a determinação de mais de uma centena de milhares de modelos atômicos, para as mais diversas proteínas (e seus complexos). Tendo em vista a existência de uma vasta coleção de informações estruturais já disponibilizadas no banco de dados *Protein Data Bank*, propõe-se aqui uma análise multiparamétrica do problema das fases, com base em aprendizagem de máquina profunda. Nossa hipótese é a de que o extensivo mapeamento estatístico de observações sobre distribuições de fases conhecidas, como um modelo preditivo, pode permitir conclusões sobre o valor alvo de fases em conjuntos de dados ainda não resolvidos, eliminando assim a necessidade de experimentos adicionais ou de estruturas homólogas previamente conhecidas. Todavia, qualquer implementação nesse sentido permanece inédita na literatura. Assim, neste Trabalho de Conclusão de Curso, serão testadas redes profundas para um seleto grupo de estruturas de lisozimas provenientes da clara do ovo de galinhas (da sigla HEWL, do inglês *Hen egg-white lysozyme*), no sentido de se avaliar a aplicabilidade inicial da aprendizagem de máquina para a resolução do problema das fases. Devido ao seu uso extensivo como modelo nos estudos de cristalização e cristalografia, centenas de estruturas de HEWL já foram determinadas experimentalmente, possibilitando assim, sua testagem na elaboração de um modelo de aprendizagem assertivo.

Palavras-chave: Cristalografia. Aprendizado profundo. Problema de fases. Difração de raios X.

1 INTRODUÇÃO

A cristalografia tem como sua principal técnica a difração de raios X. As primeiras análises dos padrões de difração no início do século XX pelas mãos de Max von Laue e seu alunos, Walther Friedrich e Paul Knipping, a partir de cristais de compostos simples, como sais, confirmaram de modo primordial a constituição atômica da matéria e a ligação entre os átomos.(1)

Mesmo sendo demonstrado que proteínas podiam formar cristais, foram necessárias décadas para o desenvolvimento de técnicas que possibilitassem a obtenção dos padrões de difração de alta qualidade (por John Desmond Bernal e Dorothy Hodgkin(2)). Décadas ainda se passaram até a determinação bem sucedida da estrutura da primeira macromolécula, a mioglobina, em 1958 por John Kendrew e Max Perutz (3). A diferença temporal entre a análise do primeiro padrão de difração e o sucesso com uma macromolécula deveu-se ao desenvolvimento de técnicas especiais para determinação do faseamento de moléculas biológicas.

A difração de raios X fornece uma das ferramentas mais importantes para se examinar a estrutura tridimensional de macromoléculas biológicas (1). A física da difração requer que, para se resolver características da estrutura molecular, seja necessário empregar radiação com um comprimento de onda da ordem dos espaçamentos atômicos intrínsecos; os raios X possuem comprimento de onda adequado e interagem de maneira significativa com a nuvem eletrônica dos átomos. No entanto, quantitativamente, a interação entre os raios X e os elétrons é fraca, e essa radiação, bastante energética, pode causar a ionização, danificando a molécula em estudo. Portanto, é necessário examinar um vasto número de moléculas simultaneamente: isso é conseguido usando um cristal contendo milhões de cópias de uma molécula em uma rede regular, que são utilizadas como amplificadores do sinal.

Quando um cristal é exposto a um feixe de raios X, uma pequena fração da radiação incidente é espalhada elasticamente e de maneira construtiva, formando um padrão de difração. Os raios X são espalhados em todos os pontos do cristal, com uma "força" proporcional à concentração de elétrons naquele ponto. Todos os raios X espalhados ao longo de qualquer direção particular interferem uns com os outros, dando origem a características detalhadas no padrão de difração que dependem da disposição e dos tipos dos átomos no cristal (coordenadas x, y e z). A análise do padrão de difração pode, portanto, permitir que a disposição espacial dos átomos seja deduzida. Em linhas gerais, determina-se uma estrutura cristalográfica, de proteínas ou não, resolvendo-se a equação 1, denominada de distribuição espacial de densidade eletrônica:

$$\rho(xyz) = \frac{1}{V}\sum_{hkl} F(hkl)exp[-2\pi(hx + ky + lz) + i\alpha(hkl)] \tag{1}$$

Ao coletar dados de difração de raios X de um cristal, quantificamos as intensidades das ondas difratadas a partir de uma série de planos imaginários que "cortam"o cristal em todas as direções. O agrupamento destes planos imaginários em famílias, por quantidade e orientação com respeito a cela unitária do cristal, determinam as frequências espaciais das ondas difratadas (*h, k, l*). A partir destas intensidades, derivamos as amplitudes das ondas dispersas para cada família de planos (*F(hkl)*).

Praticamente todos os detectores de raios X usados na cristalografia de macromoléculas - tanto no passado (baseado em luminescência foto-estimulada ou dispositivos semicondutores de carga acoplada) quanto atualmente (contagem direta de fótons) - são bidimensionais e possuem grande área de detecção, de maneira a cobrir um grande ângulo sólido da radiação isotropicamente difratada. Todavia, independente da tecnologia, apenas o fluxo direcional (definido pelo chamado Ângulo de Bragg) de energia do campo eletromagnético da radiação espalhada é medido $(W.s^{-1}.m^{-2})$. Mais precisamente, a onda incidente, a qual possui um período da ordem $10^{-18}s$, é acumulada no detector por longo intervalo de tempo (de milisegundos até segundos) muito maior que o seu período característico. Esse fenômeno, descrito classicamente pela promediação do vetor de Poynting no tempo, resulta na perda da informação das fases α(hkl) correspondentes à cada onda incidente. Como consequência, é impossibilitada a reconstrução matemática espacial da distribuição de densidade eletrônica da proteína. Devemos, portanto, obter as fases a partir de algum conhecimento complementar da estrutura molecular.

Na cristalografia de moléculas pequenas (compostas por dezenas de átomos), algumas suposições básicas sobre a atomicidade dão origem a relações entre as amplitudes das quais a informação de fase pode ser extraída. Na cristalografia de proteínas (milhares de átomos por molécula), este e outros métodos *ab initio* só podem ser usados nos casos raros em que há dados para a resolução máxima de pelo menos 1,2 Å. Para a maioria dos casos em cristalografia de proteínas, conjuntos iniciais de fases são derivados usando as coordenadas atômicas de uma proteína estruturalmente similar (substituição molecular) ou encontrando as posições de átomos pesados que são intrínsecos à proteína ou que foram adicionados (métodos tais como MIR, MIRAS, SIR, SIRAS, MAD, SAD ou combinações destes(4).

A partir dos trabalhos pioneiros de Perutz, Kendrew, Blow, Crick e outros, foram desenvolvidos métodos de substituição isomórfica: adição de átomos mais elétron-densos aos cristais, sem perturbar a estrutura da proteína. Hoje em dia, métodos de cristalografia de pequenas moléculas podem ser usados para encontrar a subestrutura de átomos pesados e as fases para toda a proteína podem ser inicializadas a partir deste conhecimento prévio. Mais recentemente, fontes melhoradas de raios-X, detectores e softwares levaram ao uso rotineiro de espalhamento anômalo para obter informações de fase a partir de selênio incorporado ou de enxofre intrínseco. Nos melhores casos, apenas um único conjunto de dados de raios X (SAD) é necessário para fornecer as posições dos dispersores anômalos, que, juntamente com os procedimentos de modificação de densidade, podem revelar a estrutura da proteína completa. Por exigirem a realização de experimentos adicionais, estes aumentam o tempo e o custo de resolução das estruturas, também não sendo aplicáveis em diversos casos.

O conhecimento acumulado ao longo de quase 60 anos de utilização da cristalografia de proteínas está publicamente disponibilizado no banco de dados *Protein Data Bank*, ou PDB. Dados experimentais, na forma de amplitudes dos fatores de estrutura (associadas aos planos cristalinos de origem), e suas respectivas fases, juntamente com um conjunto de coordenadas atômicas, são depositados pelos cientistas, quando das respectivas publicações (majoritariamente, mas não exclusivamente). É a partir deste conhecimento disponibilizado que uma abordagem alternativa para a resolução do problemas das fases pode ser aplicada; quando a macromolécula em estudo é semelhante à outra macromolécula cuja estrutura já é conhecida, o método de substituição molecular(5) permite que as fases sejam ob-

tidas a partir da estrutura desta. O cálculo da substituição molecular envolve a determinação computadorizada de uma solução única para funções de rotação e translação; a concordância dos fatores de estrutura é usada para identificar a tradução correta. Se uma combinação de orientação e a translação corretas puder ser identificada, o modelo poderá finalmente ser usado para calcular um conjunto de fases iniciais para todos os fatores da estrutura. Mapas de densidade eletrônica podem então ser calculados usando fases da estrutura do modelo e magnitudes ponderadas da estrutura desconhecida. O mapa resultante deve ser examinado, e o modelo subsequentemente modificado, para se determinar a estrutura de interesse.

Até a data de confecção deste TCC, da ordem de 150.000 estruturas atômicas de macromoléculas biológicas, determinadas exclusivamente por cristalografia de raios X, estão disponibilizadas no PDB. Destas, quase 80.000 são únicas, de acordo com um critério de busca de aplicando-se 100% de identidade sequencial. Estimamos que no banco de dados, de maneira bastante superficial, haja da ordem de 10 bilhões de relações já determinadas entre fases, amplitudes e simetria (entre outros parâmetros tão importantes quanto) para fatores de estruturas cristalográficos. É justamente com base na disponibilidade deste enorme conjunto de informações que propomos aqui uma análise estatística multiparamétrica, por aprendizagem de máquina, para avaliar a presença (ou não) de relações matemáticas ainda desconhecidas que permitam a predição futura - computacionalmente efetiva - do problema das fases em cristalografia de proteínas.

As técnicas de aprendizado de máquina estão sendo aplicadas em um grande leque de áreas, e profissionais, assim como os cientista de dados estão sendo cobiçados por diversos setores diferentes. Com o aprendizado de máquina, podemos identificar correlações em um grande banco de dados e extrair conhecimento que outrora não seriam facilmente identificadas a partir de operações manuais, maximizando a efetividade em se tomar decisões(6).

Normalmente *Big Data* está associado ao aprendizado de máquina, sendo o termo usado para descrever imensos conjuntos de dados criados a partir de uma grande coleta e armazenamento de informações, tal como os experimentos depositados no *Protein Data Bank*. E é nesse ponto que o aprendizado de máquina entra, por ser bem adequado a conjuntos de dados complexos, com um alto número de amostras e atributos. Quanto maior o conjunto de dados, maior o poder análitico e preditivo. Devido a grande variação de aplicações, as técnicas de aprendizado de máquina podem ser encontradas em áreas diversas(7) como finanças, publicidade e no nosso caso em cristalografia.

No entanto, a análise desse grande volume de informações exige a criação de ferramentas adequadas, que sejam capazes de extrair relações ocultas nos dados para subsidiar novas descobertas científicas. Uma dessas ferramentas é o aprendizado de máquina: um conjunto de métodos matemáticos e computacionais capazes de extrair relações estatísticas ou padrões dos dados e, a partir deles, realizar predições.

2 JUSTIFICATIVA E OBJETIVOS

2.1 Justificativa e Objetivos

A impossibilidade tecnológica de se detectar experimentalmente as fases das ondas de raios X difratadas por um cristal constitui a limitação fundamental da cristalografia de macromoléculas. Atualmente, soluções para este problema tem caráter experimental (substituição isomorfa ou espalhamento anômalo), ou dependem da disponibilidade prévia de estruturas de proteínas homólogas (substituição molecular); ambas as aproximações apresentam consideráveis e reconhecidas limitações em suas aplicações(1).

A princípio, não há relações matemáticas possíveis que permitam ”adivinhar”um conjunto de fases iniciais para conjunto de reflexões de macromoléculas biológicas(1). Todavia, a aplicação de métodos computacionais de fronteira, baseados em aprendizagem de máquina, tem o potencial de identificar relações estatísticas multiparamétricas ainda desconhecidas. Assim, buscaremos incorporar informações ao modelo de aprendizado de máquina, valorizando, por exemplo, somente informações obtidas imediatamente, quando da realização do experimento, como amplitudes, índices h, k, e l, parâmetros de rede e simetria; também, utilizaremos frameworks e bibliotecas de fronteira (scikit-learn, TensorFlow, pandas), todas de código aberto e constantemente refinadas pela comunidade de aprendizado de máquina do Python, propiciando-nos maior liberdade na configuração dos modelos empregados.

2.2 Objetivo Geral

Avaliar a identificação de possíveis relações estatísticas entre conjuntos de entrada rotulados (i.e., amplitudes experimentais dos fatores de estrutura, índices h, k, l, entre outros) e um conjunto de saída (fases dos fatores de estrutura), utilizando as estruturas proteicas obtidas por cristalografia de raios X depositadas na base de dados Protein Data Bank (PDB).

Todavia, é necessária a determinação prévia de uma métrica adequada para se resolver esse problema, bem como uma codificação adequada para as fases, por se tratarem de uma grandeza periódica entre 0 e 2π radianos (circular, entre 0^o e 360^o). Portanto, para este trabalho, onde é inicialmente necessário o estabelecimento de provas de conceito matemáticas, utilizaremos exclusivamente as estruturas cristalográficas de lisozima de clara de ovo de galinha.

2.3 Objetivos Específicos

1. **Coleta de Dados e Pré-processamento:** Selecionar conjuntos de dados de trabalho (estruturas já depositadas) com base em dois critérios: (i) utilizar estruturas originalmente depositadas no PDB ou aquelas automaticamente otimizadas pelo projeto PDB-REDO; (ii) utilizar somente estruturas obtidas obtidas por radiação síncrotron, somente por fontes convencionais, ou combinadas. Obter métricas de qualidade dos dados para subsidiar a escolha do dataset;

2. **Pré-processamento:** Estabelecer conjunto de treinamento, teste e validação, levando em conta a dimensão física da base de dados. Fixar métricas de desempenho e um desempenho de referência.

3. **Treinamento dos modelos:** Otimização do hardware (paralelização) para treinamento eficiente dos modelos. Treinar modelos com indicativos instantâneos de desempenho, garantindo que o processo de aprendizagem está convergindo.

4. **Avaliação dos modelos:** Empregar as métricas de desempenho para validação dos modelos com relação ao desempenho de referência.

5. **Melhoria de desempenho:** Otimização dos hiper-parâmetros. Regularização do modelo. Otimização do uso de hardware dedicado.

6. **Avaliação no conjunto de teste:** Avaliação final do desempenho dos modelos escolhidos no conjunto de testes.

7. **Avaliação dos mapas de densidade eletrônica:** Avaliação final do desempenho dos modelos escolhidos na geração de mapas de densidade eletrônica.

3 MATERIAIS E MÉTODOS

3.1 Coleta e pré-processamento dos dados

A primeira etapa deste projeto baseia-se em coletar informações da base de dados PDB, a qual contêm resultados de mais de 150.000 experimentos cristalográficos de estruturas macromoleculares. Estes são depositados sob duas formatações diferentes, sempre aos pares: (i) um arquivo *.mtz*, contendo informações experimentais, como parâmetros cristalinos, amplitudes e incertezas das reflexões, respectivos índices de Miller (h, k, l, associados às frequências espaciais das ondas espalhadas), fases calculadas entre outros, e (ii) arquivos *.pdb*, contendo as coordenadas atômicas do modelo obtido.

Até o momento, restringimos a busca para estruturas de HEWL resolvidas no grupo espacial $P4_32_12$, totalizando 584. Assim, foi realizado o download desses arquivos utilizando o conjunto de comandos da biblioteca *urllib*. Automatizou-se a conversão dos *.mtz* originais (binários) em arquivos *.cif* (ASCII), a partir de rotinas do pacote de programas Phenix (8).

Visando a construção de um conjunto estruturado em tabelas, foram utilizadas duas bibliotecas customizadas desenvolvidas anteriormente, *CIFParser* e *PDBParser*, as quais se baseiam em expressões regulares para a extração de informações disponíveis. Deste modo possibilitando a conversão para o formato de *Dataframe* da biblioteca *Pandas*, o qual é extensivamente aplicado na resolução de problemas por aprendizado de máquina por ter excelente integração com bibliotecas importantes do segmento, tais como *numpy* e *scikit-learn*.

Vale a pena ressaltar que, neste momento inicial, optamos por realizar escalonamentos e/ou normalização das amplitudes advindas de conjuntos de dados diferentes, através da inclusão do fator de estrutura normalizado E, normalização entre 0 e 1, normalização por Z-score, e por mínimos e máximos.

	index_h	index_k	index_l	FOBS	SIGFOBS	FMODEL	PHI	FOM	RESOL	E	SIGE	length_a	length_b	length_c	angle_alpha	angle_beta	angle_gamma	volume	MATTHEWS	ID	norm_FOBS	min_max_FOBS	z_FOBS
0	0	0	4	304.80	8.20	205.500	180	0.65300	9.33	0.34	0.01	78	78	37	90	90	90	231220.61	NaN	2bpu	0.204000	0.201200	1.2040
1	0	0	8	239.80	5.13	10.016	-180	0.02992	4.67	0.25	0.01	78	78	37	90	90	90	231220.61	NaN	2bpu	0.160400	0.157600	0.7515
2	0	0	12	923.50	17.02	1128.000	0	1.00000	3.11	1.23	0.02	78	78	37	90	90	90	231220.61	NaN	2bpu	0.617700	0.616700	5.5100
3	0	0	16	810.50	17.97	734.500	0	1.00000	2.33	1.72	0.04	78	78	37	90	90	90	231220.61	NaN	2bpu	0.542000	0.540500	4.7230
4	0	0	20	208.90	8.72	221.000	180	1.00000	1.87	0.79	0.03	78	78	37	90	90	90	231220.61	NaN	2bpu	0.139800	0.136800	0.5360
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9664231	65	10	3	16.17	7.81	28.880	74	0.90700	1.20	0.62	0.30	79	79	37	90	90	90	237710.42	2.07	5kxr	0.011610	0.009080	-0.6226
9664232	65	11	0	9.14	5.43	9.055	180	0.46500	1.20	0.36	0.21	79	79	37	90	90	90	237710.42	2.07	5kxr	0.006560	0.004017	-0.6880
9664233	65	11	1	9.38	4.24	7.760	-1	0.40480	1.20	0.37	0.17	79	79	37	90	90	90	237710.42	2.07	5kxr	0.006737	0.004190	-0.6860
9664234	65	11	2	12.12	6.06	21.530	-85	0.82230	1.20	0.47	0.23	79	79	37	90	90	90	237710.42	2.07	5kxr	0.008705	0.006160	-0.6606
9664235	66	1	1	11.18	4.96	13.050	48	0.65700	1.20	0.43	0.19	79	79	37	90	90	90	237710.42	2.07	5kxr	0.008026	0.005486	-0.6694

9664236 rows × 23 columns

Figura 1: Exemplo de *dataframe* contendo informações após extração e análise com bibliotecas customizadas obtido através da biblioteca *pandas*.

Fonte: Elaborada pelo autor.

3.2 Escolha de atributos e conjuntos para aprendizado

Por conter um grande número de dados faltantes *NaN*, optou-se por retirar o atributo *MATTHEWS* ao invés de substitui-los pela mediana. A partir desse ponto define-se esse *dataframe* como base para todos os modelos de aprendizagem contidos no trabalho.

É necessário para o aprendizado de máquina a divisão do *dataframe* em um conjunto de treino e outro de teste, considerando que tem-se informações de 584 estruturas e que o objetivo central da aplicação deste trabalho no futuro é propiciar a inserção do conjunto de reflexões de uma estrutura qualquer, resgatando as fases de seu fator de estrutura, optou-se por repartir o *dataframe* utilizando o número de estruturas e não uma porcentagem N de reflexões.

Portanto, definiu-se como conjunto de teste as estruturas *6ac2* e *6bo1*, resultando num conjunto de treino de 582 estruturas. O total de reflexões para cada conjunto é apresentado na Tabela 1.

Tabela 1: Quantidade de reflexões por conjunto

Treino	9664236
6ac2	33735
6bo1	32342

Fonte: Elaborada pelo autor.

3.3 Subconjuntos para testes iniciais

A intenção nessa etapa era analisar a importância do *ID* de cada estrutura no aprendizado, todavia para uma boa codificação desse atributo foi necessário dividir a identificação original em 4 novas, representando cada digito, e por se tratar de dados categóricos, aplicou-se a função *pd.get_dummies* convertendo os valores categóricos para *dummy* (leva apenas o valor de 0 ou 1, para indicar a presença do atributo), deste modo possibilitando o treino. Contudo como código de identificação do PDB é composto por um caractere numérico $(0-9)$ seguido por três caracteres que podem ser tanto numéricos quanto alfabéticos $(A-Z)$, após a conversão o *dataframe* pode ganhar até 118 novos atributos, acarretando num enorme acréscimo de memória utilizada tornando o processo computacional de treino ainda mais custoso, e até impossibilitando o aprendizado em máquinas comuns.

Como alternativa restringiu-se o conjunto de teste para apenas 36 estruturas reduzindo consideravelmente o número de reflexões como mostrado na Tabela 2, desta forma viabilizando comparações entre modelos contendo ou não o código de identificação.

Tabela 2: Quantidade de reflexões conjuntos de treino.

36 Estruturas	776282
582 Estruturas	9664236

Fonte: Elaborada pelo autor.

3.4 Modelos

Tendo em vista o sucesso e a grande utilização em outros problemas de aprendizado de máquina com dados estruturados, foram utilizados três algorítimos de aprendizado principais, todos eles construídos em cima do ambiente propiciado pela biblioteca *Scikit-Learn*:

a) SGDRegressor: É um algorítimo presente na própria *Scikit-Learning*, este apresenta um modelo linear que é ajustado focando a minimização da perda empírica regularizada com o SGD,

que vem do inglês *Stochastic Gradient Descent*, o qual significa que o gradiente de perda é estimado por vez a cada amostra e o modelo é atualizado ao longo do caminho com uma taxa de aprendizagem.

b) LightGBM: É um *framework* de *boosting* de gradiente, desenvolvido pela *Microsoft*, que baseia seu treinamente em algoritimos de arvore. Desenvolvido para apresentar alta performance em grande conjuntos de dados, um menor consumo de memória e uma rápida velocidade de treinamento.

c) XGBoost(9):, é baseado na técnica de *tree boosting*, a qual utiliza estimadores CART (*Classification and Regression Trees*) que são treinados para ajustar-se ao erro preditivo do estimador precedente.

Agora, entrando na confecção dos modelos em si, para o presente projeto, em especial, foram desenvolvidos 6 modelos, dos quais os dois primeiros servem como controle, visando demonstrar a eficácia dos demais.

3.4.1 Otimização

Objetivando a otimização dos modelos, e por consequência uma melhora da automação dos algorítimos, foi aplicado o conceito de *Pipelines*, o qual tem o propósito de criar uma estrutura que executa métodos de transformações e um ajuste, o estimador final(10).

Além disso definiu-se uma distribuição de hiper parâmetros para cada estimador, os quais serão explicitados neste trabalho, em associação ao *RandomizedGridSearch* que tem como função procurar os melhores hiper parâmetros através de repetitivos ajustes e validações cruzadas.

Deve-se salientar que em modelos que utilizam a codificação *Bitternion*(11) foi necessária a utilização do *MultiOutputRegressor* para adaptar o estimadores a trabalhar com uma saída dupla. Configurando deste modo um estimador para o seno e outro para o cosseno do ângulo do fator de estrutura.

3.4.2 Controle

Nos modelos a seguir, utilizou-se *PHI* como alvo de predição, ou seja, uma saída simples considerando o o ângulo da fase.

- *SGDRegressor_36*: Uma simples aplicação do SGDRegressor para o conjunto de dados com 36 estruturas, neste momento ainda considerávamos a codificação do *ID*, seguindo a seguinte distribuição de hiper parâmetros:

Tabela 3: Hiper parâmetros para o SGDRegressor.

sgdregressor__penalty	'l1'
sgdregressor__max_iter	1500
sgdregressor__alpha	uniform(loc=0.0, scale=1.0)
sgdregressor__l1_ratio	uniform(loc=0.0, scale=1.0)
sgdregressor__learning_rate	'constant'

Fonte: Elaborada pelo autor.

- *XGBoost_36*: Foi utilizado o método de regressão presente na sua API, XGBRegressor para o conjunto de dados com 36 estruturas, considerando os *ID* e também realizando uma validação cruzada aleatória com 3 interações com os seguintes hiper parâmetros:

Tabela 4: Hiper parâmetros para o XGBRegressor.

xgbregressor__min_child_weight	0.5
xgbregressor__max_depth	20
xgbregressor__n_estimators	100
xgbregressor__alpha	10

Fonte: Elaborada pelo autor.

3.4.3 Bitternion

Para os demais modelos, utilizou-se a codificação proposta para o ângulo de fase do fator de estrutura. Deste modo foi possível descrever um alvo como uma tupla do seno e do cosseno do *PHI*, tal como explicitado na equação 2.

$$f(\phi) = [\cos\phi, \sin\phi] \tag{2}$$

- *LightGBM*: Foi utilizado o método de regressão presente na sua API, LGBMRegressor para três casos distintos:
  - *LightGBM_36_ID*: Conjunto de treino com 36 estruturas e considerando a codificação de *ID*.
  - *LightGBM_36_SID*: Conjunto de treino com 36 estruturas e não considerando a codificação de *ID*.
  - *LightGBM_582_SID*: Conjunto de treino com 582 estruturas e não considerando a codificação de *ID*.

  Para todos os casos utilizando os mesmo hiper parâmetros, os quais foram definidos pelo processo de *boosting* com o *RandomizedGridSearch* com uma distribuição pré definida, que são descritos na Tabela 5 além de serem submetidos a uma validação cruzada com 5 interações.

Tabela 5: Hiper parâmetros para o LGBMRegressor.

LGBMRegressor__boosting_type	dart
LGBMRegressor__learning_rate	0.9730
LGBMRegressor__max_depth	50
LGBMRegressor__n_estimators	150
LGBMRegressor__number_leaves	60
LGBMRegressor__reg_alpha	0.9060
LGBMRegressor__reg_lambda	0.4343

Fonte: Elaborada pelo autor.

- *XGBoost*: Utilizamos o método de regressão presente na sua API, XGBRegressor para dois casos diferentes:
  - *XGBoost_36_ID*: Conjunto de treino com 36 estruturas e considerando codificação de ID.
  - *XGBoost_582_SID*: Conjunto de treino com 582 estruturas e desconsiderando ID.

  Em todos os casos após análises em uma distribuição de parâmetros, foram definidos os hiper parâmetros descritos na Tabela 6, além de ter sido realizada uma validação cruzada com 3 interações.

Tabela 6: Hiper parâmetros para o XGBRegressor.

XGBRegressor__min_child_weight	0.5
XGBRegressor__alpha	10
XGBRegressor__max_depth	20
XGBRegressor__n_estimators	100

Fonte: Elaborada pelo autor.

4 RESULTADOS E DISCUSSÃO

4.1 Avaliação dos Modelos

Para a avaliação dos modelos para a predição das estruturas de teste, é necessário considerar a distância média entre os ângulos preditos e aferidos, todavia apenas considerando apenas esta métrica é provável chegar em conclusões errôneas, então mutuamente foi analisado a distribuição desses erros, focando em uma grande amostragem de casos onde o $\Delta PHI < 15^{o}$.

4.1.1 Controle

4.1.1.1 SGDRegressor

Este regressor serve como referência para os demais modelos presentes no trabalho, nele foram aplicados ambos os conjuntos de treinamento considerando as identificações, e tendo como objetivo a predição do ângulo do fator de estrutura PHI para todas as reflexões, é possível observar na Figura 2 que os valores preditos para todas as amostras é o mesmo, isso indica que houve *overfitting* no aprendizado.

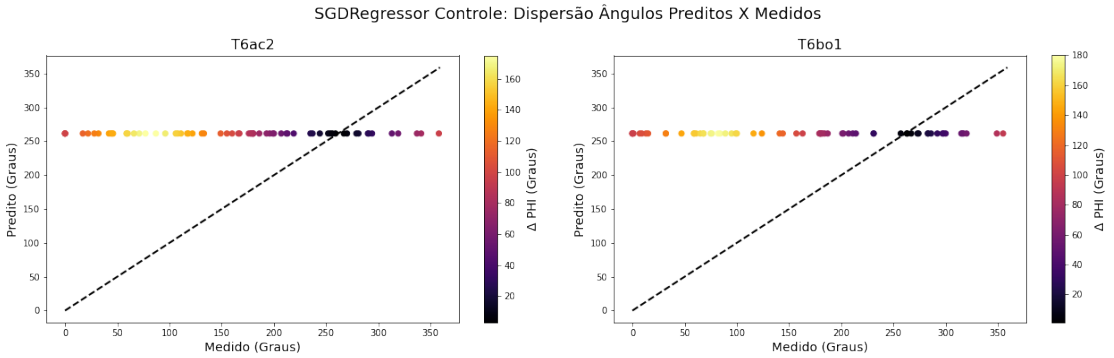


Figura 2: Dispersão de ângulos preditos e medidos para o SGDRegressor
Fonte: Elaborada pelo autor.

A partir da Tabela 7, é possível notar um alto valor de erro para todas as métricas utilizadas, associando com o histograma (Figura 3) da distância entre ângulos preditos e medidos, é observável um grande pico próximo de 80^{o} e 100^{o}, o qual evidencia a hipótese de *overfitting* durante o aprendizado.

Tabela 7: Métricas para SGDRegressor, onde RMSE(*Root Mean Squared Error*), MAE(*Mean Absolute Error* e AADE(*Average Angle Distance Error*)

Conjunto	RMSE (º)	MAE (º)	AADE (º)
T6ac2	138	113	89
T6bo1	138	113	89

Fonte: Elaborada pelo autor.

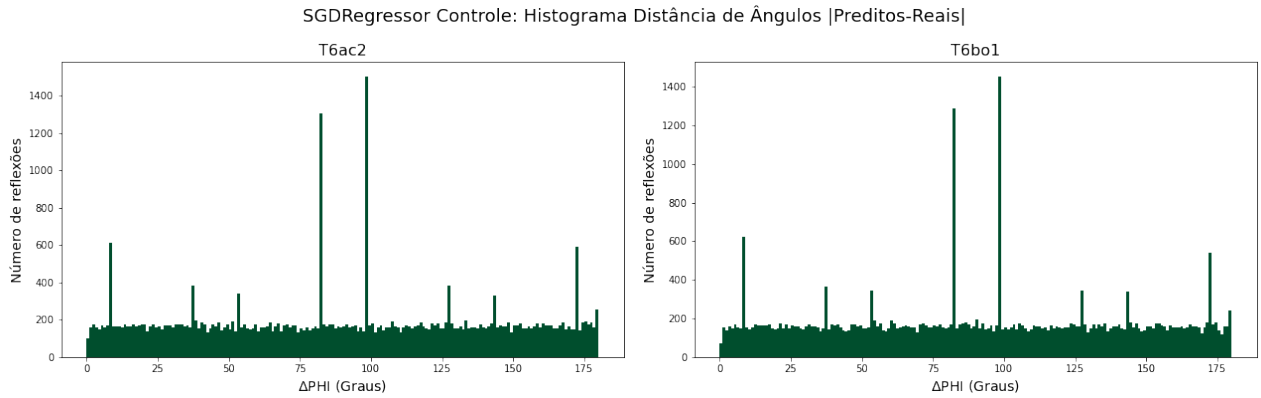


Figura 3: Histograma da distância de ângulos preditos e medidos para SGDRegressor
Fonte: Elaborada pelo autor.

4.1.1.2 XGBoost

Partindo neste momento para um regressor mais robusto, optou-se pelo *XGBRegressor* para avaliar se foi obtida uma melhora. Levando em conta novamente a predição do ângulo de fator de estrutura sem nenhuma codificação, concluise que ocorreu novamente um *overfitting*, todavia foi obtida um reta com uma distribuição em uma pequena amplitude conforme a Figura 4.

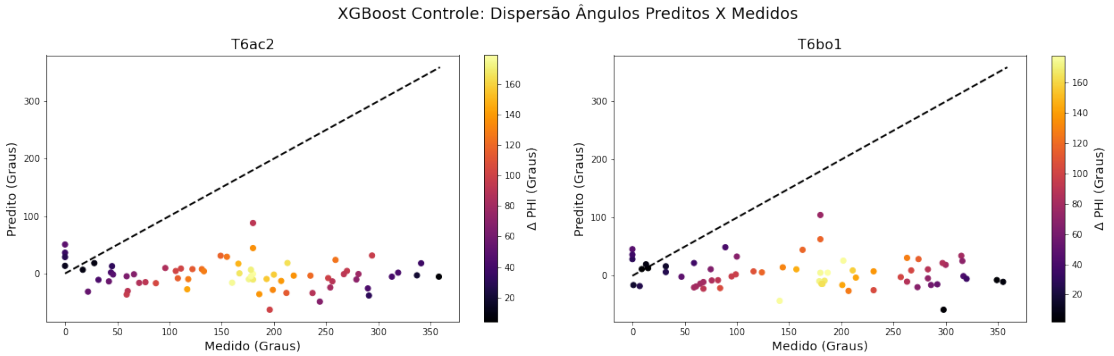


Figura 4: Dispersão de ângulos preditos e medidos para o XGBRegressor
Fonte: Elaborada pelo autor.

É possível observar na tabela 8 um aumento para duas das três métricas, uma explicação para isso é a condição do nosso alvo ser circular, deste modo não considerando que ângulos como 350^o e 10^o estão apenas a 20^o de distância, tal comportamento não ocorre com a distância média dos ângulos (AADE, do inglês *Average Angle Distance Error*), demonstrando uma melhora, ainda que mínima, de 2^o.

Tabela 8: Métricas para XGBRegressor, onde RMSE(*Root Mean Squared Error*), MAE(*Mean Absolute Error* e AADE(*Average Angle Distance Error*)

Conjunto	RMSE (º)	MAE (º)	AADE (º)
T6ac2	203	173	87
T6bo1	202	172	87

Fonte: Elaborada pelo autor.

Neste modelo, o histograma (Figura 5) apresenta uma distribuição próxima para todos os ΔPHI, embora não seja o ideal, por não apresentar os picos em ângulo aleatórios, mostra um avanço no tratamento perante o *overfitting*.

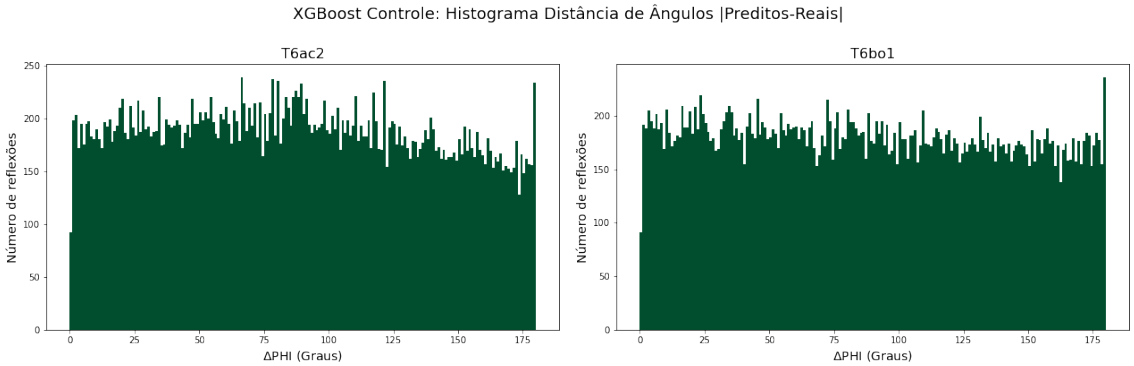


Figura 5: Histograma da distância de ângulos preditos e medidos para XGBRegressor
Fonte: Elaborada pelo autor.

4.1.2 36 Estruturas

4.1.2.1 LightGBM

A partir deste momento, foi introduzido o regressor LGBMRegressor além de começar a utilização uma nova codificação para o alvo do projeto, aplicando o conceito de *Biternions*, deste modo o PHI é descrito agora como uma tupla de seno e cosseno do ângulo do fator de estrutura. Nota-se, pelo gráfico (Figura 6), que há valores mais distribuídos para a estrutura T6bo1, evidenciando uma melhora no sistema de predição desenvolvido neste projeto. Embora exista *overfitting* para a estrutura T6ac2, é possível lapidar o método com intuito de evitá-lo.

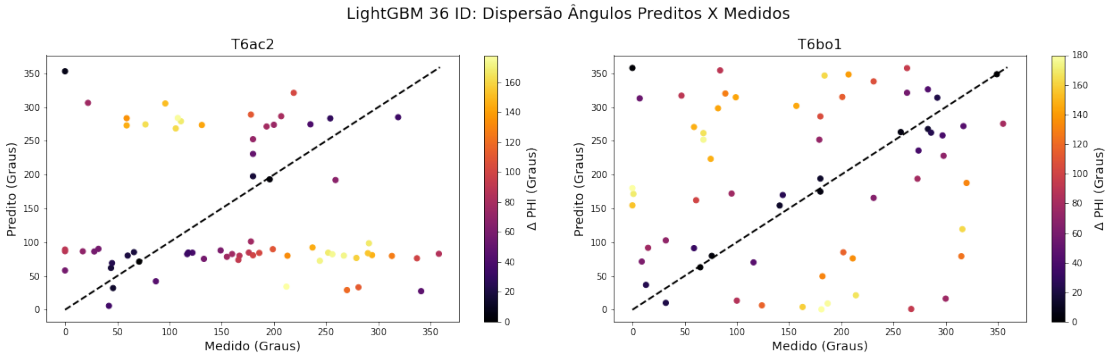


Figura 6: Dispersão de ângulos preditos e medidos para o LGBMRegressor
Fonte: Elaborada pelo autor.

A partir da Tabela 9, é possível ver que o RMSE e MAE continuam altos enquanto o AAD se mantêm estável. Mas é possível ver a partir do histograma (Figura 7) começou a ser observável uma tendência de picos maiores nos extremo, no extremo perto de 0^o é a nossa meta, todavia é possível ver uma grande presença perto de 180^o.

Tabela 9: Métricas para LightGBM, onde RMSE(*Root Mean Squared Error*), MAE(*Mean Absolute Error* e AADE(*Average Angle Distance Error*)

Conjunto	RMSE (º)	MAE (º)	AAD (º)
T6ac2	147	118	86
T6bo1	159	127	87

Fonte: Elaborada pelo autor.

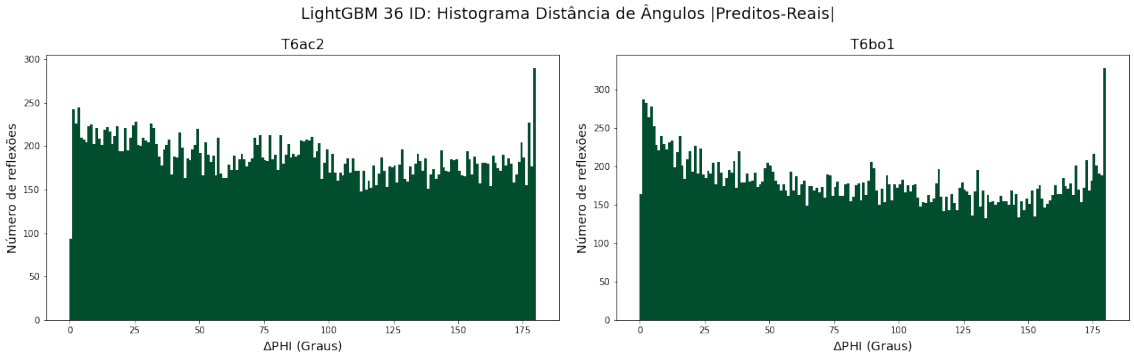


Figura 7: Histograma da distância de ângulos preditos e medidos para LGBMRegressor

Fonte: Elaborada pelo autor.

4.1.2.2 XGBoost

Seguindo o mesmo tratamento que o modelo anterior, o XGBRegressor consegue lidar melhor com a estrutura T6ac2, de acordo com o gráfico de dispersão mostrado na Figura 8, não existe mais um *overfitting* predizendo 100^{o} para uma grande fração da amostra.

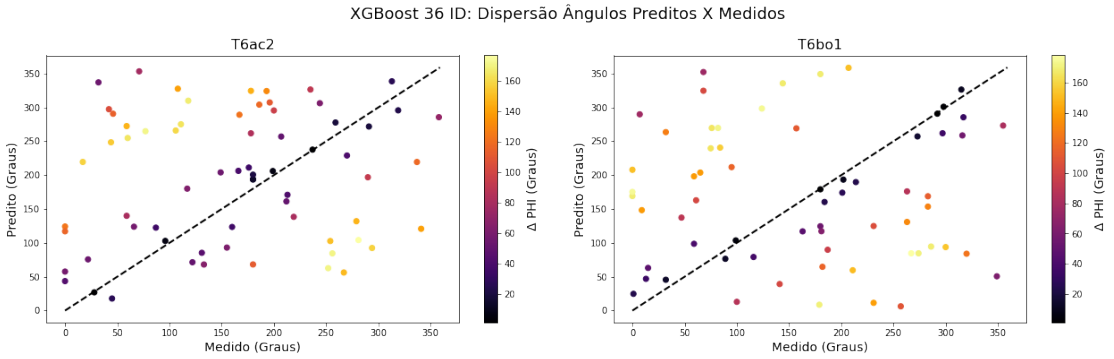


Figura 8: Dispersão de ângulos preditos e medidos para o XGBRegressor

Fonte: Elaborada pelo autor.

E levando em comparação o XGBregressor com o alvo sendo PHI, pode-se ver uma incrível melhora no modelo - com PHI descrito por *Bitternions*, mesmo que o erro associado média das distância dos ângulos tendo um decréscimo relativamente pequeno como vista na Tabela 10, obteve-se uma grande melhora considerando o histograma (Figura 9) adquirindo um formato exponencial decrescente.

Tabela 10: Métricas para XGBRegressor, onde RMSE(*Root Mean Squared Error*), MAE(*Mean Absolute Error* e AADE(*Average Angle Distance Error*)

Conjunto	RMSE (º)	MAE (º)	AADE (º)
T6ac2	141	109	78
T6bo1	146	115	83

Fonte: Elaborada pelo autor.

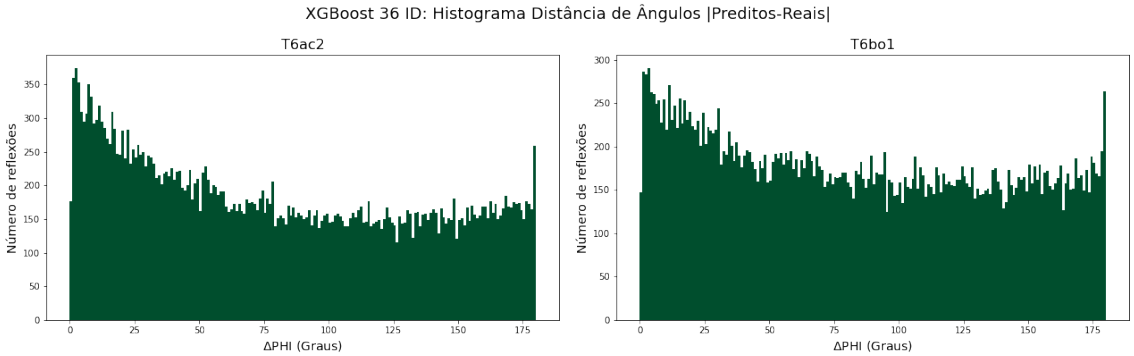


Figura 9: Histograma da distância de ângulos preditos e medidos para XGBRegressor

Fonte: Elaborada pelo autor.

4.1.3 Sem ID

Com o objetivo de obter o máximo de eficiência de aprendizado, é necessário utilizar todo o conjunto de treino, entretanto do modo inicial dele, o qual consta com uma codificação para os IDS, este processo torna-se excessivamente custoso do ponto de vista computacional.

Como alternativa, foi realizado um treinamento com o *dataset* de 36 estruturas, porém excluindo os atributos de IDs. Deste modo, quando fosse expandido para as 582 estruturas a redução do seu tamanho, e por consequência da memória necessária e tempo de processamento, implicará em um conjunto com $1/7$ da dimensão em relação ao arquivo original.

Esse procedimento foi realizado no LGBMRegressor, pois era o modelo que apresentava mais *overfitting*, deste modo se melhorasse o pior regressor, com certeza seria induzido ao melhor. Além do fato de ser um algoritmo com velocidade superior de execução considerando o XGBRegressor.

4.1.3.1 LightGBM

É notável, pelo gráfico de dispersão (Figura 10, que foi obtida uma melhora no modelo, já que no caso no qual foi considerado os identificadores o *overfitting* para a estrutura T6ac2 está muito mais claro. Neste momento há uma melhor distribuição, mesmo que por hora erroneamente.

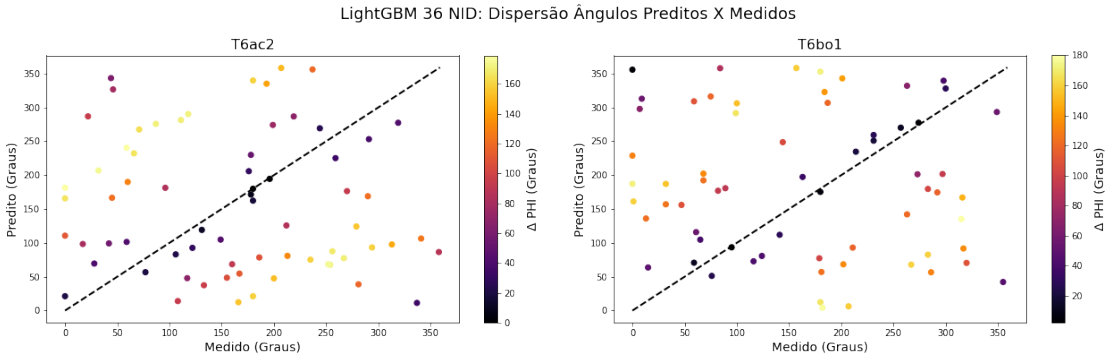


Figura 10: Dispersão de ângulos preditos e medidos para o LGBMRegressor
Fonte: Elaborada pelo autor.

Além disso, foi obtida uma melhora no ADDE (Tabela 11) e no histograma (Figura 11), desta forma concluí-se que a remoção da identificação, torna o modelo mais eficiente tanto em questões de eficiência quanto em acurácia.

Tabela 11: Métricas para LGBMRegressor, onde RMSE(*Root Mean Squared Error*), MAE(*Mean Absolute Error* e AADE(*Average Angle Distance Error*)

Conjunto	RMSE (º)	MAE (º)	AADE (º)
T6ac2	147	117	85
T6bo1	151	120	84

Fonte: Elaborada pelo autor.

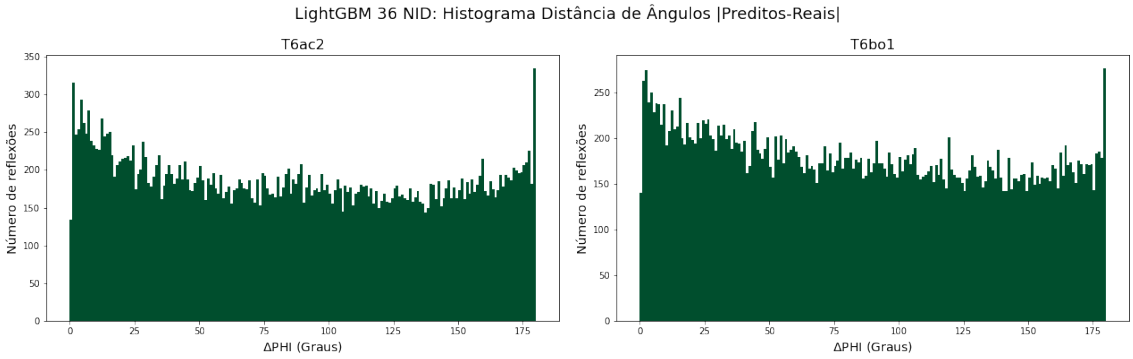


Figura 11: Histograma da distância de ângulos preditos e medidos para LGBMRegressor
Fonte: Elaborada pelo autor.

4.1.4 582 Estruturas

Com o sucesso da otimização do conjunto de treino com a remoção dos atributos de identificação, é possível expandi-lo para 582 estruturas sem prejudicar o aprendizado. Deste modo para os próximos dois modelos, o conjunto de treino é o de 582 estruturas, além de utilizar a codificação de *Bitternion* para o ângulo do fator de estrutura.

4.1.4.1 LightGBM

Note que, mesmo que hajam pontos nos extremos, considerando a linha de equidade entre valores de predição e valores medidos (Figura 12), com a utilização da ferramenta de mapa de cores

da biblioteca *matplotlib*, é possível evidenciar que o ΔPHI é pequeno. Não aparenta ter *overfitting*, e possui grande quantidade de pontos próximos a linha.

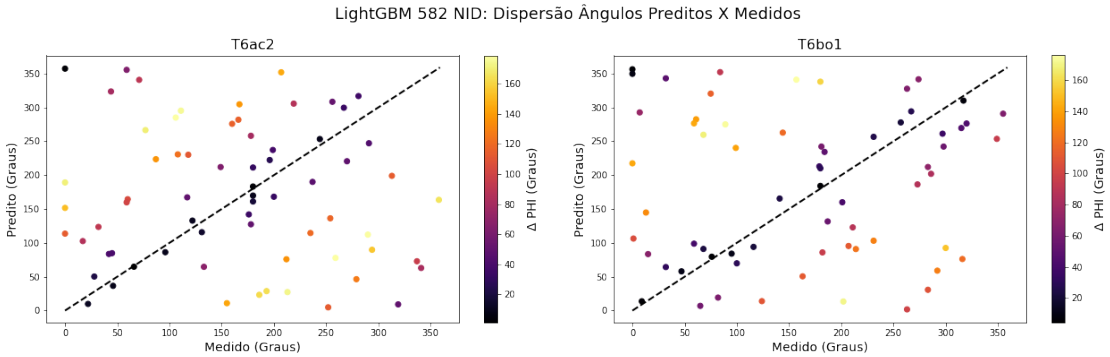


Figura 12: Dispersão de ângulos preditos e medidos para o LGBRegressor
Fonte: Elaborada pelo autor.

Foi obtido também um valor equivalente aos modelos anteriores de AADE (Tabela 12), isso se deve ao fato que o histograma (Figura 13) ainda não apresenta uma configuração ótima, mesmo que esteja começando a tomar forma. Ainda é necessário ajustes para maximizar a eficiência.

Tabela 12: Métricas para LGBMRegressor, onde RMSE(*Root Mean Squared Error*), MAE(*Mean Absolute Error* e AADE(*Average Angle Distance Error*)

Conjunto	RMSE (º)	MAE (º)	AADE (º)
T6ac2	148	116	83
T6bo1	146	114	80

Fonte: Elaborada pelo autor.

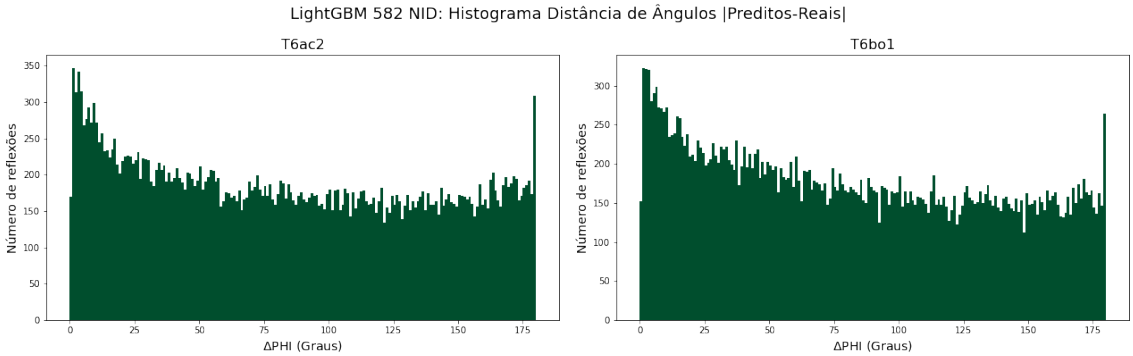


Figura 13: Histograma da distância de ângulos preditos e medidos para LGBMRegressor
Fonte: Elaborada pelo autor.

4.1.4.2 XGBoost

Por último, é possível observar a grande eficiência do modelo, uma parte significativa de predições seguem a reta identidade, na Figura 14, e apresentam um $\Delta PHI < 40^{o}$. Apesar de termos predições erradas com $\Delta PHI > 100^{o}$, considerando o ponto de partida do projeto, é possível afirmar que as modificações aplicadas até então parecem bem condizentes para que seja alcançado o objetivo deste projeto, dando os primeiros passos para a solução computacional do problema de fases.

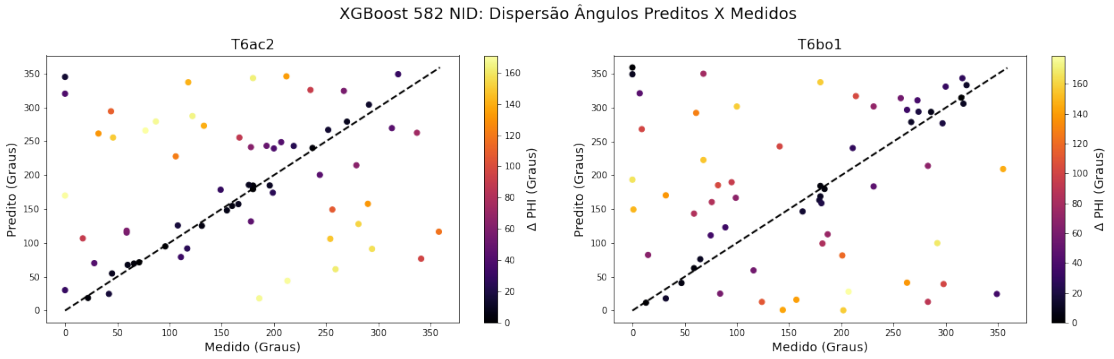


Figura 14: Dispersão de ângulos preditos e medidos para o XGBRegressor
Fonte: Elaborada pelo autor.

Há uma grande diminuição do valor de erro associado a distância média dos ângulos preditos e medidos (Tabela 13), uma redução da ordem de $\pm 15^{o}$ e o histograma (Figura 14) obtido é definitivamente animador, já apresenta um formato exponencial decrescente bem definido, apenas com um pequeno aumento perto de 180^{o}, que de certa forma era esperado por tratarmos com ângulos.

Tabela 13: Métricas para XGBRegressor, onde RMSE(*Root Mean Squared Error*), MAE(*Mean Absolute Error* e AADE(*Average Angle Distance Error*)

Conjunto	RMSE (º)	MAE (º)	AADE (º)
T6ac2	137	95	67
T6bo1	133	100	70

Fonte: Elaborada pelo autor.

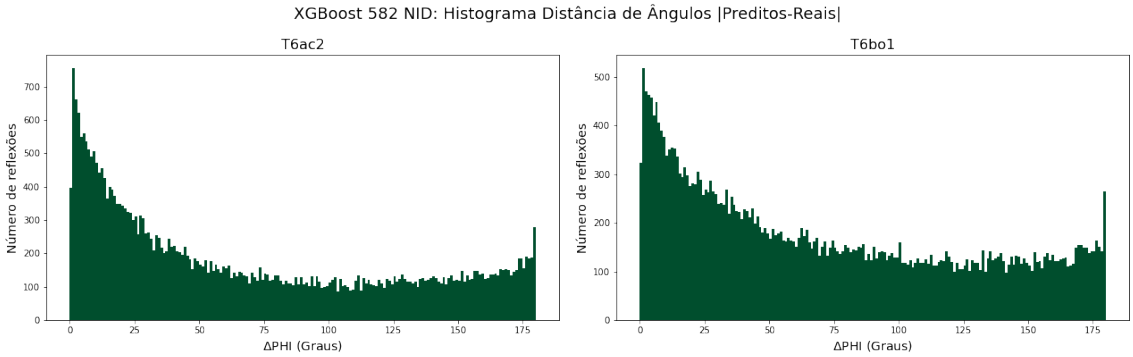


Figura 15: Histograma da distância de ângulos preditos e medidos para XGBRegressor
Fonte: Elaborada pelo autor.

4.2 Reconstrução espacial da densidade eletrônica

De maneira a se quantificar objetivamente a qualidade do conjunto de fases obtido pelo modelo XGBoost 582 NID, torna-se necessária a implementação de uma métrica numérica adequada, diferente daquela que mede o desvio médio da fase predita com relação à fase esperada. Neste sentido, pelo momento optamos por avaliar a qualidade visual dos mapas de densidade eletrônica do tipo mF_o, onde F_o é o fator de estrutura calculado para cada (hkl), através de amplitudes experimentais e fases do XGBoost, ponderadas pela respectiva figura de mérito m. O cálculo do mapa foi feito pelo programa FFT(12) do pacote CCP4.

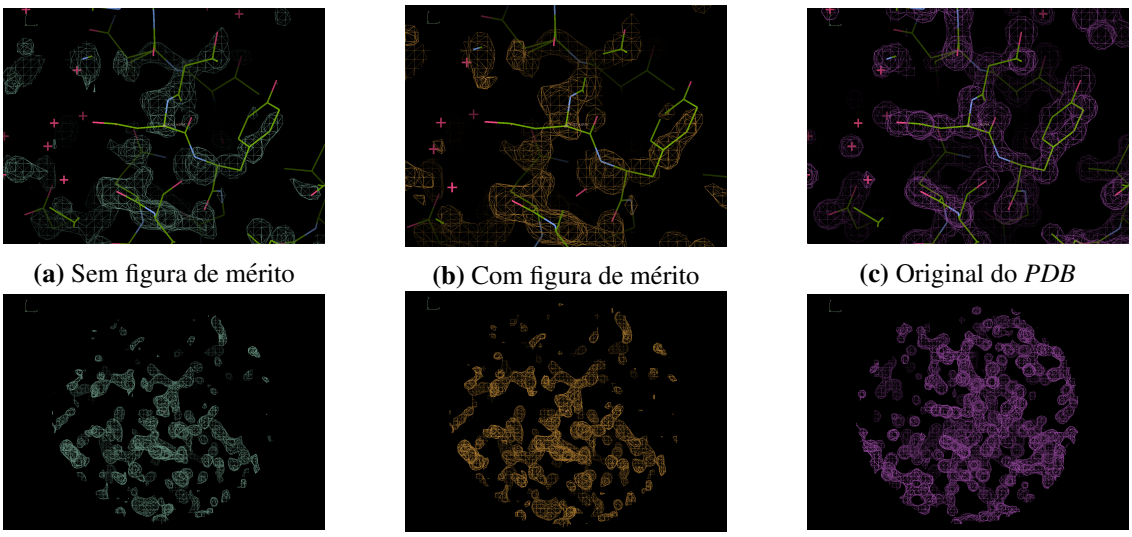
Três mapas foram gerados para cada conjunto de teste: (i) um mapa com figura de mérito idealizada igual a 1, (ii) um mapa com figura de mérito igual ao cosseno do erro médio do modelo XGBoost, e um mapa controle, com as fases e figuras de mérito originais do conjunto teste (depositadas no PDB).

A partir da obtenção do modelo preditivo XGBoost 582 NID, objetivou-se elaborar um processo para a reconstrução espacial da densidade eletrônica para ambas as estruturas de teste. Este procedimento baseia-se na seleção de 5 atributos significativos obtidos após o experimento de cristalografia em associação com o valor predito para ângulo do fator de estrutura, e por último adotou-se duas vertentes: a primeira com apenas esses valores, a segunda apresentou uma estimativa para a figura de mérito de cada reflexão, adotando-se como método de aquisição o valor do cosseno do AADE.

Obtendo-se o *dataset* final, é necessária a conversão destes dados, resgatando novamente um arquivo do tipo *.cif*. Realizou-se esse processo para as estruturas *6ac2* e *6bo1*, assim totalizando 6 arquivos de informação cristalográfica, já que considerou-se também os arquivos originais para fins comparativos.

Com o auxilio do software cristalográfico CCP4 (*Collaborative Computational Project Number 4*), implementou-se o a conversão do arquivo de informação cristalográfica para um arquivo de dados de reflexão *.mtz*. Com essa informação foi possível, através de uma transformada rápida de Fourier, a reconstrução do mapa espacial da densidade eletrônica.

Realizou-se uma análise qualitativa através do software *Wincoot*, de modo a assegurar um caminho para concluir o ciclo completo do processo de predição do ângulo do fator de estrutura. Métricas avançadas de validação dos mapas de densidade eletrônica serão desenvolvidas no futuro.



(a) Sem figura de mérito **(b)** Com figura de mérito **(c)** Original do *PDB*

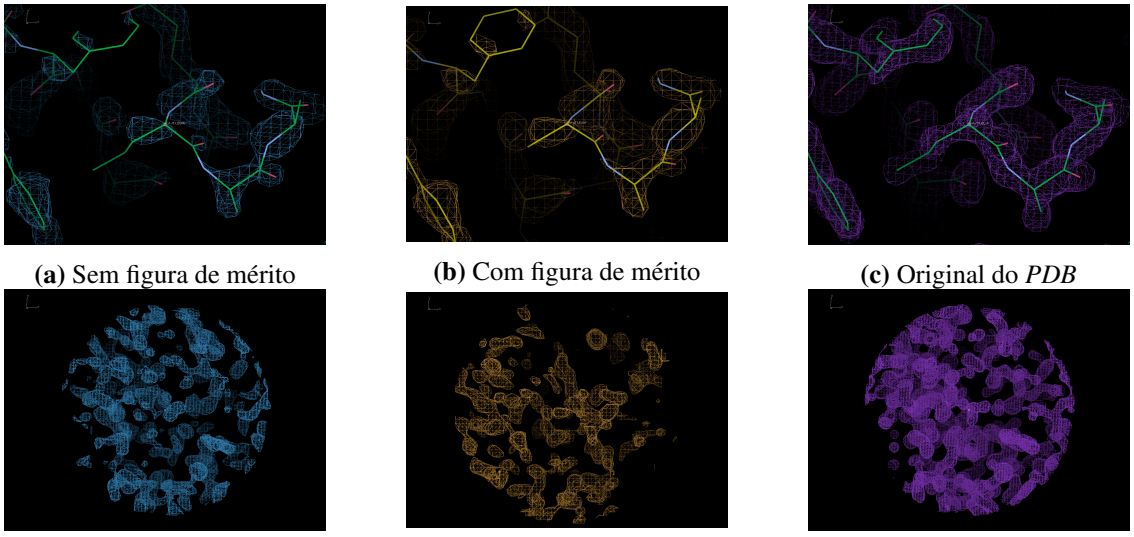
(d) Sem figura de mérito **(e)** Com figura de mérito **(f)** Original do *PDB*

Figura 16: Representação dos mapas de densidade eletrônica do tipo mF_o da estrutura 6ac2, contornados a 1σ. Visualização de um resíduo: (a), (b) e (c); Visualização total: (d), (e) e (f).

Fonte: Elaborada pelo autor.

Observa-se na Figura 16 e Figura 17 representações da densidade eletrônica associadas com a estrutura da proteína 6ac2 e 6bo1, respectivamente, na forma de bastões. É possível observar que ambos os mapas gerados por predição (com e sem a figura de mérito), mesmo com falhas e considerando

o intervalo de erro das predições, são condizentes com os mapas disponíveis no



(a) Sem figura de mérito **(b)** Com figura de mérito **(c)** Original do *PDB*

(d) Sem figura de mérito **(e)** Com figura de mérito **(f)** Original do *PDB*

Figura 17: Representação dos mapas de densidade eletrônica do tipo mF_o da estrutura 6bo1. Visualização de um resíduo: (a), (b) e (c); Visualização total: (d), (e) e (f)

Fonte: Elaborada pelo autor.

5 CONCLUSÕES E PERSPECTIVAS

Neste projeto demostrou-se técnicas para driblar as barreiras primordiais do tratamento do problema de fases da cristalografia de forma computacional. Tornou-se possível a extração de conhecimento através das informações presentes no *Protein Data Base*, utilizou-se bibliotecas customizadas previamente desenvolvidas, para a confecção de conjuntos de dados para o aprendizado de máquina.

O primeiro gargalo considerável foi a questão de como tratar o ângulo alvo da predição, visto que possui um comportamento circular no intervalo $[-\pi, \pi]$, temos como solução a utilização da codificação *Bitternion*, na qual é predita uma tupla contendo o seno e cosseno do ângulo do fator de estrutura. Discutiu-se a importância da quantidade de informação presente no conjunto de treino, e como a não utilização dos atributos de identificação da proteína afetava este fator, já que sua presença não melhorava a acurácia, no entanto, possui correlação com a extensão do *dataframe*, tornando inviável trabalhar com grandes amostras.

Imediatamente iniciou-se a procura por parâmetros que maximizassem a acurácia dos dois regressores principais (XGBRegressor e LGBMRegressor) e, consequentemente, minimizando a média da distância entre ângulos preditos e aferidos. Conclui-se nesta etapa que o modelo XGBoost 582 NID possui a melhor predição, com um histograma com dados distribuídos majoritariamente em ΔPHI pequenos, além de apresentar um AADE de $\approx 67^{o}$ para 6ac2 e $\approx 70^{o}$ para 6bo1.

Com algumas predições para as duas estruturas 6ac2 e 6bo1, foi feito a conversão dos dados tabulares para o formato de informação cristalográfica (.cif), os quais são convertidos para .mtz com o auxilio do software CCP4. Por fim, os mapas de densidade eletrônicas são gerados após uma transformada rápida de Fourier (fft). A partir de análises visuais foi obtido um resultado equivalente entre mapas com e sem figura de mérito, e, além disso, comparando com o mapa depositado no PDB para cada estrutura foi encontrado semelhanças condizentes com o grau de acurácia da predição.

Neste projeto, portanto, foi possível contribuir com o inicio do tratamento computacional para o problema de fases da cristalografia. Realizando um ciclo completo, com a obtenção de dados, aprendizado e geração do mapa de densidade eletrônica das estruturas. Ademais, como próximo passo, deve-se implementar novas métricas de perda, confecção de modelos com novos regressores e refinar hiper parâmetros. Contudo, para a análise de mapas faz-se necessário um estudo mais profundo que vise a determinação de indicadores quantitativos e de um valor mais adequado para estimativa da figura de mérito de cada reflexão, assegurando a eficácia do processo. Adicionalmente, pode-se ainda estudar a viabilidade na implementação de métodos de melhoria das fases, como achatamento de solvente, entre outras.

REFERÊNCIAS

1 RUPP, B. *Biomolecular crystallography*: principles, practice, e application to structural biology. New York: Garland Science, 2010. ISBN: 9780815340812.

2 HOWARD, J. A. K. Dorothy Hodgkin e her contributions to biochemistry. *Nature Reviews Molecular Cell Biology*, v.4.n.11, p.891-896,2003. ISSN: 1471-0080. DOI:10.1038/nrm1243.

3 KENDREW, J. C. *et al*. Structure of myoglobin. *Nature*, v.185 n.4711, p.422-427, 1960.

4 TAYLOR, G. L. Introduction to phasing. *Acta Crystallographica* D: biological crystallography, v.66. p.325-338,2010.DOI: 10.1107/S0907444910006694.

5 ARGOS, P.; ROSSMAN, M. G. Molecular replacement method. *In:* LADD, M. F. C.; PALMER, R. A.(eds) *Theory and practice of direct methods in crystallography.* Palmer. Boston, MA: Springer US, 1980, p. 361–417. DOI: 10.1007/978-1-4613-2979-4 10.

6 VASILEV, I. *et al. Python deep learning:* exploring deep learning techniques, neural network architectures with PyTorch, Keras, e TensorFlow.[S.l.] Packt Publishing Ltda., 2019. ISBN: 9781789348460.

7 TARCA, A. L. *et al*. Machine learning e its applications to biology. *PLOS Computational Biology*, v. 3. n.6.p. 1–11.2007. DOI: 10.1371/journal.pcbi.0030116.

8 ADAMS, P. D. *et al*. PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallographica* D, v.58. n.11, p.1948-54,2002.

9 CHEN, T.; GUSTRIN, C. *XGBoost*: a scalable tree boosting system. San Francisco: ACM, 2016. DOI: 10.1145/ 2939672.2939785.

10 GERON, A. *Hands-on machine learning with scikit-learn and tensorflow:* concepts, tools, e techniques to build intelligent systems. 2nd. ed. Gravenstein Highway North:O'Reilly Media, 2017. ISBN: 9781491962299.

11 PROKUDIN, S.;GEHLER, P. V.;NOWOZIN, S. *Deep directional statistics: pose estimation with uncertainty quantification.* Disponivel em:https://arxiv.org/pdf/1805.03430.pdf. Acesso em: 05.03.2021.

12 IMMIRZI, A. *Crystallographic computing techniques*. Munksgaard: F.R.Ahmed,1966. p.399.