POC 1
Avaliação de quatro métodos de feature-importance para a explicabilidade de features estruturais de peptídeos geradas usando o método aCSM

André Luiz M. Dutra[1],
Orientadora: Raquel C. de Melo Minardi[1], **Coorientador: Diego C. B. Mariano**[1]

[1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais
Minas Gerais, Brasil.

cienciandre@ufmg.br

Abstract. *With the advancement of machine learning techniques, their use has shown promise in the field of structural bioinformatics, made possible by structural feature extraction techniques such as the atomic Cutoff Scanning Matrix (aCSM). This work proposes a solution to the lack of interpretability of these models by utilizing feature importance techniques associated with the classification of aCSM signatures into peptide sequence clusters in the generation of data that facilitates the construction of semantic visualizations. Tests of the method applied to cluster classifiers demonstrated high effectiveness of the suite of feature importance metrics in the qualitative interpretation of different aspects of the observed model.*

Resumo. *Com o avanço de técnicas de aprendizado de máquina, seu uso se mostrou promissor na área de bioinformática estrutural, viabilizado por técnicas de extração de features estruturais como a atomic Cutoff Scanning Matrix (aCSM). Este trabalho propõe uma solução ao problema de falta de interpretabilidade destes modelos utilizando técnicas de feature importance associadas à classificação de assinaturas aCSM em clusters de sequências de peptídeos para a geração de dados que viabilizam a construção de visualizações semânticas. Testes do método aplicado a classificadores de clusters demonstraram alta efetividade da suite de métricas de feature importance na interpretação qualitativa de diferentes características do modelo observado.*

1. Introdução

Com o crescente avanço de técnicas de aprendizado de máquina, rapidamente evidenciaram-se os benefícios de seu uso na solução de problemas em bioinformática (área de pesquisa voltada ao uso de métodos computacionais para a análise de dados biológicos), especialmente na área de bioinformática estrutural, analisando estruturas de macromoléculas como proteínas e peptídeos. Algoritmos de aprendizado de máquina são usados, por exemplo, para realizar a predição de propriedades e comportamentos de moléculas, bem como detectar padrões que apontem para seus papéis nos processos biológicos em que estão inseridos. Como exemplo, Rosa et al. apresentam quatro ferramentas baseadas em aprendizado de máquina amplamente testadas e utilizadas no

âmbito de pesquisa para realizar a predição de características farmacocinéticas in silico de moléculas [Rosa et al. 2022]. De modo a viabilizar o uso destas técnicas, algoritmos de extração de features são utilizados para converter a representação bruta da macromolécula, usualmente caracterizada por conjuntos de cadeias de micromoléculas e seus respectivos átomos, junto a suas posições tridimensionais no espaço [Ruczinski 2002], em vetores de features estruturados que podem ser utilizados de maneira eficiente em tarefas de classificação e regressão. Nesse sentido, denominam-se assinaturas estruturais vetores de features gerados levando em consideração a configuração espacial na qual a estrutura da molécula se encontra, sendo o atomic Cutoff Scanning Matrix, ou aCSM, um algoritmo voltado à extração desse tipo de assinatura [Pires et al. 2013b].

O aCSM utiliza contagens de determinados pares de átomos em diferentes limiares de distância para gerar assinaturas estruturais, tendo as assinaturas do tipo aCSM demonstrado resultados promissores em tarefas de predição e detecção de padrões. Pires et al. observaram, por exemplo, que assinaturas geradas utilizando o mCSM, uma variação local do aCSM aplicada a proteínas para uso em modelos de detecção de sítios de mutação, apresentou desempenho tão bom quanto ou melhor que o de assinaturas geradas utilizando outras técnicas [Pires et al. 2013a]. No entanto, na contramão destes resultados, o uso de assinaturas do tipo aCSM traz desafios em compreender, semanticamente, que características concretas da estrutura original da molécula ocasionaram o resultado obtido pelos modelos, uma vez que o comportamento da molécula é definido por sua estrutura espacial e as propriedades farmacofóricas de seus átomos [Lehninger et al. 2005].

Dessa maneira, este artigo abordará o problema de encontrar pistas que correlacionem os resultados de modelos que utilizam assinaturas aCSM às respectivas estruturas originais das moléculas das quais a assinatura foi extraída. É importante notar, nesse contexto, que a assinatura aCSM se comporta como um gargalo de informação entre os dados da estrutura original e os resultados obtidos nos modelos. Assim, o problema apresentado, na prática, se divide em dois: correlacionar os resultados da classificação com a assinatura estrutural utilizada e correlacionar a assinatura estrutural com a estrutura original da molécula. Buscamos, portanto, propor técnicas que, a partir das assinaturas e dos modelos nos quais elas foram usadas, gerem dados que evidenciem as correlações existentes entre a estrutura original e os resultados obtidos pelos modelos de aprendizado, de maneira que, em trabalhos futuros, estes dados possam ser utilizados para gerar visualizações com teor semântico de correlação da estrutura original aos resultados obtidos.

Um dos principais desafios ao alcance desse objetivo é a perda inerente de informação da estrutura original ocasionada pela extração da assinatura aCSM. Inevitavelmente, detalhes cruciais da complexa estrutura espacial da molécula são perdidos, tornando impossível o retorno à estrutura original exata da molécula a partir apenas da assinatura. Com isso, o problema de correlacionar a estrutura original à assinatura aCSM, um dos dois problemas mencionados anteriormente, é reduzido ao problema de inferência, a partir da assinatura aCSM, de um conjunto limitado de características concretas e abstratas da estrutura espacial da molécula. Surge assim um terceiro problema, o problema de definir quais aspectos concretos e abstratos da estrutura podem e têm relevância para serem extraídos da assinatura do tipo aCSM.

Propomos, neste artigo, uma solução para os problemas apresentados baseada em técnicas de feature importance aplicadas a modelos de classificação utilizando assinaturas

aCSM. Técnicas de feature importance definem, segundo algum critério, a porcentagem de influência de cada feature de uma assinatura na predição de determinada característica alvo por um modelo. Assim, sugerimos uma solução na qual o problema de inferência de características da estrutura original é reduzido a um problema de classificação de determinadas características da estrutura utilizando as prórpias assinaturas aCSM, e do modelo gerado são extraídas as feature importances de cada campo da assinatura. Dado que as feature importances apresentem correlações evidentes com os resultados da classificação, a comparação entre as feature importances deste modelo com as feature importances de um modelo arbitrário que use assinaturas aCSM permite uma forma consistente de correlacionar elementos da estrutura original aos resultados do modelo analisado, que é justamento o objetivo final deste artigo.

O texto deste trabalho está organizado da seguinte maneira: na seção 2, são detalhados os conceitos fundamentais para o entendimento do problema e da solução proposta. Dentre os conceitos discutidos estão o detalhamento do funcionamento do aCSM, por exemplo. Na seção 3, a solução proposta neste projeto para o tratamento dos problemas apresentados é explicada em mais detalhes. A seção 4 detalha as escolhas de implementação de testes da solução descrita na seção anterior. Na seção 5, são apresentados os resultados encontrados para a implementação, bem como sua análise no contexto do problema. Por fim, a seção 6 traz as conclusões tiradas a partir dos resultados encontrados, juntas às perspectivas de continuidade da pesquisa em trabalhos futuros.

2. Conceitos Fundamentais

Esta seção se dedica a esclarecer os conceitos fundamentais associados à definição do problema e à compreensão da solução proposta. Primeiramente, serão discutidos os conceitos de bioinformática estrutural e aprendizado de máquina, conceitos base para a compreensão do contexto do problema. Em seguida, será o funcionamento do método aCSM de geração de assinaturas estruturais será apresentado em detalhes, uma vez que todo o artigo se centraliza em torno da interpretação das features obtidas após sua aplicação. Por fim, serão definidas as técnicas de feture importance utilizadas na solução proposta, junto às peculiaridades de interpretação de cada uma.

2.1. Bioinformática Estrutural

A bioinformática é o campo da biologia e da ciência da informação responsável pela coleta, armazenamento, análise e difusão de dados biológicos [Lesk 2020]. Neste contexto, a bioinformática estrutural é o ramo da bioinformática focado no estudo da estrutura espacial de macromoléculas, como proteínas, peptídeos e ácidos nucleicos. Como a geometria destas moléculas é um fator determinante para a definição de seu papel fisiológico e comportamento, principalmente no caso de proteínas e peptídeos [Lehninger et al. 2005], este campo não se resume apenas ao estudo de topologias espaciais, sendo também um dos principais responsáveis pelo avanço da ciência na compreensão da fisiologia de seres vivos.

2.1.1. Proteínas e Peptídeos

Proteínas e peptídeos são biomoléculas biológicas formadas por uma ou mais cadeias de aminoácidos, ligados por meio de ligações peptídicas [2]. São moléculas profunda-

mente envolvidas na grande maioria dos processos biológicos dos seres vivos, e por isso sua pesquisa é de extrema relevância. A diferença entre peptídeos e proteínas está em seu tamanho. Peptídeos são biomoléculas menores, formados por uma única cadeia de dois a 50 aminoácidos ligados por meio de ligações peptídicas [Martins et al. 2023]. Já proteínas são macromoléculas formadas por uma ou mais cadeias cadeias com mais de 50 aminoácidos.

Devido a seu tamanho reduzido, peptídeos não conseguem formar estruturas funcionalmente complexas, como enzimas. No entanto, sua estrutura análoga à de proteínas favorece ligações a moléculas maiores. Assim, peptídeos geralmente formam complexos de ligantes com proteínas maiores, podendo se comportar como cofatores ou inibidores, por exemplo. Por este motivo, peptídeos apresentam um grande potencial de uso como fármacos.

2.2. Aprendizado de Máquina

O aprendizado de máquina é um subcampo da inteligência artificial que utiliza minimização de funções de erro para realizar detecção de padrões em dados a partir de um conjunto de dados de treino [Bishop 2016]. Algoritmos baseados em aprendizado de máquina utilizam de grandes bases de dados como treino para calibrar seus próprios parâmetros, até que as tarefas as quais eles foram designados a executar apresentem resultados pertinentes e possam ser generalizadas para dados fora da amostra utilizada.

2.2.1. Aprendizado Supervisionado

O aprendizado supervisionado é um subcampo do aprendizado de máquina focado no estudo de modelos gerados a partir de dados rotulados. Dada uma base de treino e um conjunto de rótulos, associados de modo que cada exemplo da base esteja associado a um rótulo, o objetivo de um modelo de aprendizado supervisionado é, aprender a rotular novos exemplos fora da base de treino. É denominado um problema de classificação o problema de atribuir rótulos categóricos aos dados (por exemplo, dada uma imagem, definir se ela possui ou não um cachorro). Já o problema de atribuir rótulos contínuos (por exemplo, dada a altura e o peso de uma pessoa, definir a sua taxa de gordura corporal) é denominado um problema de regressão.

2.2.2. Vetores de Features

Uma característica (feature) corresponde a uma propriedade ou característica individual e mensurável de um determinado fenômeno [Bishop 2016]. Uma assinatura ou vetor de features é, portanto, uma lista de diferentes características extraídas de um mesmo fenômeno. Assim, vetores de features constituem uma forma de representar um determinado fenômeno utilizando suas características mais relevantes para determinada tarefa, destacando-se neste contexto tarefas de aprendizado de máquina.

2.3. O método aCSM

O método aCSM consiste em um algoritmo de extração de assinaturas estruturais de proteínas e peptídeos [Pires et al. 2013b]. Ele é proposto como uma evolução do Cutoff

Scanning Matrix, o CSM, um algoritmo de extração de assinaturas de proteínas baseado em comparações entre as posições dos centroides de pares de aminoácidos da molécula [Pires et al. 2011]. Mas, diferente do CSM, o aCSM, ou atomic Cutoff Scanning Matrix, constrói a assinatura estrutural comparando posições de pares de cada átomo da molécula, seguindo determinadas regras. Esta subseção se destina a explicar em mais detalhes o processo de montagem da assinatura aCSM.

2.3.1. Dados Iniciais

O aCSM recebe como dados iniciais os dados de cada proteína ou peptídeo como arquivos em formato pdb (formato padrão de representação de proteínas e peptídeos, cunhado pelo Protein Data Bank). Nele, a molécula é representada como uma tabela na qual cada átomo é listado em uma linha. Cada linha apresenta o elemento químico do átomo, suas coordenadas x, y e z no espaço, em Ångströms, o resíduo a que o átomo pertence e a cadeia em que o resíduo está inserida, dentre outras informações [Ruczinski 2002].

2.3.2. Descrição do Método

O método aCSM consiste basicamente em tratar a molécula como um grafo completo, no qual os vértices são os átomos e as arestas são as distâncias entre eles, e calcular a matriz de distâncias do grafo. Em seguida, são definidas uma distância máxima, que determina o intervalo onde as distâncias da matriz serão analisadas, e a distância de um cutoff, que corresponde a um passo discreto nesse intervalo. O intervalo de 0 à distância máxima é então dividido em cuttofs, que são pequenos intervalos complementares de tamanho igual à distância de corte definida. Por fim, é feita uma varredura (scanning) na matriz de distâncias entre os átomos da molécula, onde, para cada intervalo de cutoff, é feita a contagem de quantos pares de átomos têm distância dentro do intervalo. A contagem é refeita em cada cutoff para diferentes categorias de pares de átomos. A listagem das contagens de cada cutoff em um vetor constitui a assinatura estrutural aCSM final da molécula. A seção (D) da figura 5 demonstra visualmente a geração do aCSM. Círculos azuis representam os cutoffs de uma molécula para os pares de um átomo específico. O processo é repetido para todos os átomos.

2.3.3. Seleção de Pares

Ao invés de realizar uma contagem total de pares de átomos com distância dentro do limiar do cutoff, o aCSM conta os pares segundo diferentes categorias. As categorias de pares contadas em cada cutoff são determinadas por propriedades manifestadas por cada átomo da molécula. Assim, os átomos da molécula são caracterizados em uma ou mais das seguintes categorias: aceptor, doador, aromático, hidrofóbico, negativo, neutro, positivo e sulfeto. As categorias correspondem a diferentes propriedades que os átomos da molécula podem manifestar, e não são disjuntas entre si, logo, um átomo pode pertencer a mais de uma delas. Por exemplo, átomos doadores e aceptores de hidrogênio apresentam alta eletronegatividade, e átomos em um anel aromático tendem a ter comportamento hidrofóbico. Definidas as 8 propriedades buscadas em cada átomo da molécula, em cada

cutoff são contados, para cada uma das 36 combinações de pares possíveis das 8 categorias, quantos pares de átomos da molécula possuem distância dentro do limiar do cutoff e têm cada átomo do par correspondente a pelo menos uma categoria cada da combinação de par observada. Isso gera uma assinatura com 36 campos por cutoff, onde o número de cutoffs é determinado pela distância máxima dividida pelo tamanho do cutoff, sendo ambos valores hiperparâmetros do algoritmo.

As oito categorias de átomos utilizadas pelo algoritmo para a contagem de pares foram definidas levando em consideração as quatro principais forças de estabilização da estrutura de proteínas: Ligações iônicas, ligações dissulfeto, ligações de hidrogênio e interações hidrofóbicas. Elas se caracterizam como as principais interações moleculares participantes na definição da estrutura final da molécula [Lehninger et al. 2005]. Ligações iônicas atuam ocasionando a aproximação de resíduos com átomos de cargas opostas, formando ligações iônicas entre os íons carregados em cada resíduo. Ligações dissulfeto ocorrem quando uma ligação covalente é formada entre dois enxofres das cadeias laterais de duas cisteínas, ligando-as. Ligações de hidrogênio atuam exercendo uma forte interação intermolecular entre átomos de hidrogênio e átomos de Oxigênio e Nitrogênio, todos presentes em abundância em resíduos. Por fim, interações hidrofóbicas atuam posicionando os resíduos da molécula no meio aquoso em que ela está inserida segundo sua polaridade (e, por consequência, sua hidroafinidade). Assim, resíduos apolares tendem a se aglutinar ao centro da molécula, enquanto resíduos polares tendem a se distribuir nas periferias em maior contato com o meio aquoso.

As categorias, assim, estão relacionadas a estas forças: aceptor e doador são características de átomos em ligações de hidrogênio; negatividade e positividade são características de átomos em ligações iônicas, e também denotam a polaridade do átomo, que está relacionada à sua hidroafinidade; neutro, aromático e hidrofóbico denotam átomos com pouca hidroafinidade; e sulfetos são os átomos participantes de ligações dissulfeto.

2.4. Métricas de Feature Importance

Em muitos contextos, como o de algoritmos de bioinformática, tratado neste artigo, a interpretabilidade de um modelo de aprendizado de máquina é tão importante quanto o seu desempenho de predição. Tendo isto em vista, métricas de feature importance oferecem uma maneira de interpretar os efeitos de cada feature dos dados no comportamento de uma predição. A técnica consiste em, dado um modelo de aprendizado supervisionado ou um conjunto de dados destinado ao uso em modelos de aprendizado supervisionado, definir qual é a porcentagem de importância de cada feature dos dados na predição dos rótulos, segundo algum critério.

Estes critérios podem variar essencialmente de duas formas principais. As métricas se subdividem entre gnósticas ou agnósticas ao modelo e univariadas ou multivariadas. Para técnicas agnósticas ao modelo, o resultado da métrica não depende de que modelo é utlizado no treinamento. A análise é feita apenas com as features e os rótulos, tornando sequer necessário o treinamento de um modelo para a geração das feature importances. Já métricas agnósticas dependem do modelo utilizado, seja por utilizar características de um modelo específico, ou por utilizar o modelo para classificar instâncias geradas sintéticamente para a composição da feature importance. Quanto às técnicas utilizadas neste trabalho, as duas primeiras são agnósticas e univariadas, enquanto as duas últimas são gnósticas e multivariadas.

2.5. Técnicas Univariadas vs Multivariadas

Conforme mencionado, as técnicas de geração de feature importance podem ser univariadas ou multivariadas. Técnicas univariadas não consideram correlações entre as features, tratando-as como variáveis independentes entre si. Isso faz com que features correlacionadas, em um cenário ideal, obtenham importâncias iguais e iguais às importâncias das features não correlacionadas. Isso faz com que features não necessariamente essenciais sejam valorizadas, agindo como um falso-positivo para a importância destas features. Já técnicas multivariadas levam em consideração, total ou parcialmente, as correlações entre as features. Isso faz com que features correlacionadas, em um cenário ideal, dividam entre si a importância do padrão que as relaciona. Embora estas técnicas ressaltem as features independentemente importantes, features importantes mas fortemente correlacionadas são penalizadas neste tipo de assinatura, agindo como um falso-negativo para a impotância destas features [Patel et al. 2022].

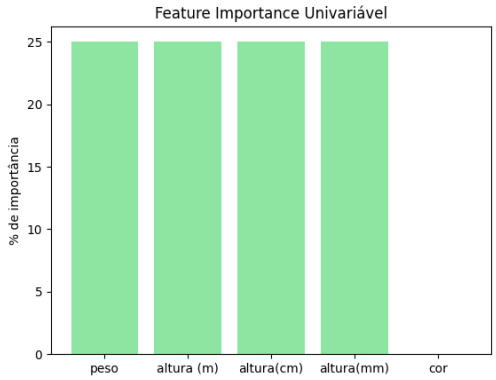

Figure 1. Exemplo fictício de uma técnica univariável de métrica de feature importances aplicada em um contexto onde o padrão é explicado igualmente pelas features ”peso” e ”altura”. Note que as três features ”altura” possuem correlação linear completa entre si.

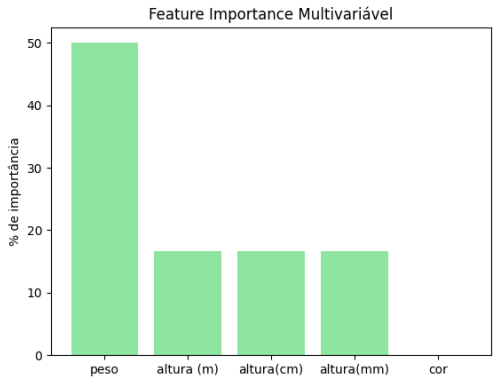

Figure 2. Exemplo fictício de uma técnica multivariávelvariável de métrica de feature importances aplicada ao mesmo contexto que o exemplificado na figura 1.

Para ilustrar as diferenças mencionadas, observe as figuras 1 e 2. Ambas são

representações fictícias e idealizadas das feature importances de um padrão cuja previsão depende igualmente do peso e da altura do objeto, e não é influenciada pela feature cor. A figura 1 representa uma técnica univariada de geração de feature importances, enquanto a figura 2 representa uma técnica multivariada. Note também que as três features altura estão completamente correlacionadas por um fator linear. Em 1, todas as alturas recebem importâncias iguais e iguais à importância do peso. Embora a feature cor, não relacionada, tenha sido identificada como não importante, as três alturas foram beneficiadas, quando apenas uma bastava para prever o valor definido. Já em 2, o peso recebeu importância de 50%, enquanto as três alturas dividiram os 50% restantes entre si. Embora a feature peso, independente, tenha recebido a importância correta, as três alturas foram penalizadas, dando a falsa impressão de que a altura é três vezes menos importante que o peso, o que não é verdade.

É notável que ambos os casos possuem falhas, mas eles funcionam de maneira complementar. Enquanto feature importances univariadas são boas em detectar quais features não são importantes, feature importances multivariadas são boas em detectar features independentemente importantes. Um conjunto de técnicas de ambos os tipos será, assim, usado ao longo do trabalho para demonstrar a influência de cada feature do aCSM em tarefas de aprendizado supervisionado.

2.5.1. F-Statistic

A F-statistic é uma métrica de comparação entre duas distribuições. A F-statistic de um par de distribuições consiste na divisão da variância entre as distribuições pelas variâncias dentro de cada distribuição. Assim, para o cálculo das feature importances, a F-statistic é calculada entre cada feature e o rótulo, e então o vetor de F-statistics normalizado representa a feature importance [Peng et al. 2020]. Esta métrica é agnóstica ao modelo, já que não requer o uso de um classificador, e univariada, já que trata cada feature independentemente. Embora seja efetiva em detectar features não-correlacionadas com o rótulo, ela se baseia em variâncias lineares, então não vai ser efetiva em capturar a maioria dos padrões não lineares de correlação. Para além disso, o F-test, feito a partir da F-statistic, não garante a detecção de correlações, apenas de variáveis não correlacionadas.

2.5.2. Mutual Information/Maximal Information Coefficient

A informação mútua é uma métrica de comparação entre duas variáveis aleatórias discretas. A informação mútua é calculada como a redução na incerteza sobre uma das variáveis após o conhecimento da outra. Matematicamente, esse valor é representado pela diferença entre a entropia da primeira variável (grau de incerteza) e a entropia condicional da primeira variável dada a segunda (grau de incerteza após o conhecimento da segunda variável) e vice versa. Como a informação mútua só pode ser calculada entre variáveis discretas, para features contínuas é usado o Maximal Information Coefficient, que corresponde à informação mútua aplicada a limiares discretos das variáveis [Kinney and Atwal 2014].

Embora as features do aCSM sejam discretas por se tratarem de contagens inteiras, implementações do MIC são bem mais comuns, e o MIC converge para a informação

mútua quando as variáveis são discretas, por isso o MIC foi utilizado como métrica de feature importance nesse caso. Para isso, a informação mútua (ou o MIC) é calculada entre cada feature e o rótulo e o vetor normalizado representa as feature importances.

De forma semelhante à técnica anterior, esta também é agnóstica ao modelo e também é univariáel. No entanto, a vantagem dessa técnica é o fato de que a entropia contempla todos os tipos de correlação, não só as lineares. Desse modo, o MIC representa uma métrica mais fiel de feature importances univariáveis para os modelos observados.

2.5.3. Impurity Reduction Importance (Gini Importance)

A impurity reduction importance é uma métrica extraída do treinamento de modelos baseados em árvores de decisão, como árvores de decisão unitárias e random forests. Nesses modelos, a árvore possui um determinado grau de ”impureza”, que representa a distância do modelo à predição correta de um valor. As bifurcações feitas em cada nó da árvore para uma feature são escolhidas de modo a minimizar o grau de impureza da árvore. Esta métrica funciona de maneira semelhante à anterior, uma vez que a impureza pode ser vista como a incerteza (e em muitos casos a função de entropia é realmente usada como impureza nas árvores) e a impureza diminuída pela bifurcação do nó de uma feature pode ser interpretada como a diminuição da incerteza do rótulo gerado [Nembrini et al. 2018]. No entanto, ao contrário do MIC, a impurity reduction é realizada em passos pequenos e graduais, onde a cada passo uma variável é escolhida de maneira gulosa para reduzir a impureza. Assim, as variáveis influenciam umas às outras e evidenciam parcialmente relações multivariáveis. Além disso, não só a entropia é utilizada como impureza nas árvores, como várias outras métricas de impureza, e neste trabalho utilizamos o índice Gini.

A feature importance é calculada somando, assim, a redução de impureza ocasionada por cada feature em cada nó, e normalizando o vetor de feature importances. É evidente que esta métrica é gnóstica ao modelo, já que depende especificamente de uma classificação usando árvores de decisão ou random forests. Além disso, conforme mencionado, o viés do algoritmo atribui uma certa multivariabilidade às relações entre as features, mas esse comportamento depende do viés do algoritmo e da métrica de impureza utilizada.

2.5.4. Permutation Importance

Por fim, a última técnica de feature importance usada foi a permutation importance. Ela consiste em, dado um modelo e os dados rotulados, escolher pares aleatórios de features, trocar e comparar o resultado original contra o resultado da classificação da nova feature com os valores trocados. Assim, mudanças entre variáveis pouco relevantes ou muito correlacionadas não causam mudança, enquanto alterar variáveis independentemente importantes modifica o resultado drasticamente o resultado obtido pelo modelo. O grau de mudança em cada feature, normalizado, é a Permutation Importance [Altmann et al. 2010]. Com isso, a permutation importance é uma técnica gnóstica ao modelo, pois requer um classificador treinado, e multivariada, pois testa a permutação de vários pares de variáveis diferentes. Sua principal desvantagem é o desempenho. Por pre-

cisar chamar o classificador muitas vezes para cada par de features analisado, a execução do programa pode ficar muito pesada, dependendo da implementação do modelo.

3. Definição da solução e workflows do projeto

Esta seção descreverá os aspectos teóricos da abordagem tomada para a proposição da solução do problema apresentado, bem como escolhas de implementação usadas para a realização dos testes.

3.1. Workflow do Projeto

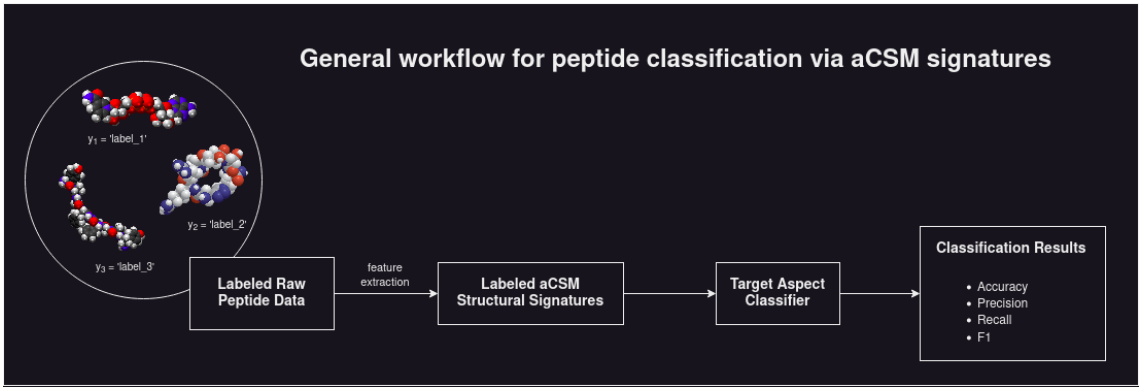

Figure 3. Workflow geral de classificação de peptídeos usando assinaturas aCSM

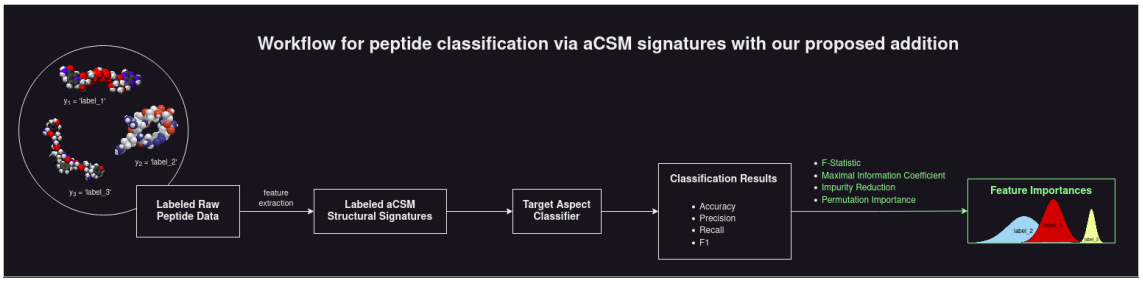

Figure 4. Workflow geral de classificação de peptídeos usando assinaturas aCSM com a nossa proposta de adição marcada em verde

A figura 3 apresenta o workflow geral utilizado para o treinamento de classificadores de peptídeos utilizando assinaturas do tipo aCSM. Se trata de um modelo sequencial simples no qual os dados dos peptídeos rotulados são coletados, e a partir deles assinaturas aCSM são extraídas. Com as assinaturas aCSM, o classificador é treinado para a identificação de um aspecto alvo específico, representado nos rótulos, e ao fim são feitos testes de desempenho no modelo com um conjunto de testes, e métricas de teste são apresentadas como resultado. Pequenas variações deste workflow são usadas em todos os artigos que apresentam e validam as assinaturas do tipo CSM e derivados [Pires et al. 2011][Pires et al. 2013a][Pires et al. 2013b][Martins et al. 2023], como por exemplo é apresentado na seção (A) da figura 5.

Propomos neste artigo uma pequena alteração ao workflow de classificação original, conforme destacado em verde na figura 4. A ideia consiste em, ao fim do treinamento do classificador, coletar métricas de feature importance da classificação realizada.

Dado um classificador arbitrário que utilize como dados de entrada assinaturas aCSM, a coleta das feature importances nos permite atribuir interpretabilidade às influências de cada feature da assinatura estrutural nos resultados da classificação. Deste modo, dado que as feature importances se mostrem um método consistente de identificar padrões da classificação dentre as features da assinatura, um dos três subproblemas apresentados na introdução será solucionado, o problema de correlacionar a assinatura aCSM aos resultados dos classificadores.

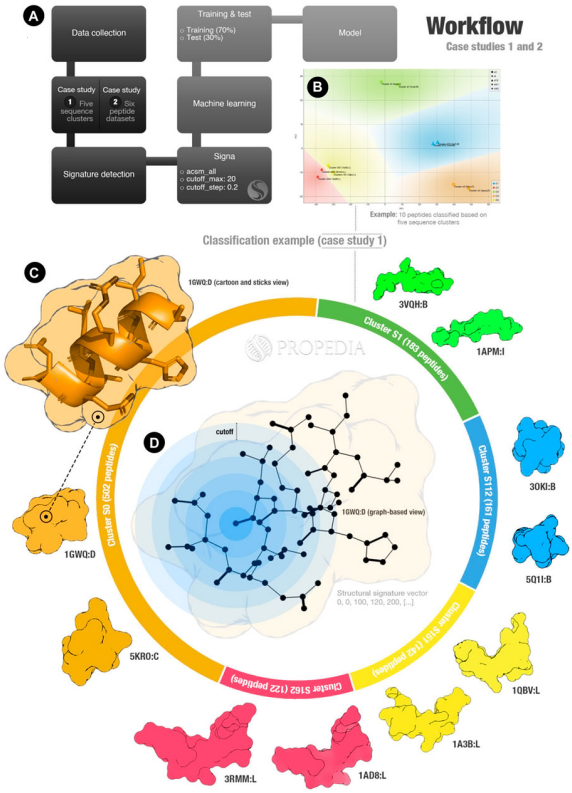

Figure 5. Workflow de validação das assinaturas de peptídios do Propedia. (A) Workflow geral do processo. (B) Exemplo de representação de clusters para o estudo de caso 1 (eixos x e y representam PCA1 e PCA2). As cores representam as regiões para classificação dos clusters de sequências (geradas usando o Orange Data Mining). (C) Infográfico representa o primeiro exemplo do estudo de caso. (D) Representação da assinatura estrutural baseada em grafos. Cuttoffs são representados por círculos azuis. Figura gerada usando ChimeraX e Open-source PyMOL.

No entanto, sobram os problemas de correlacionar a estrutura aCSM a aspectos da estrutura original, e o problema de definir os aspectos da estrutura original com os quais a assinatura será correlacionada. Estes problemas são especialmente desafiadores pelo fato de não existir um conjunto canônico de aspectos da estrutura original relevantes em todas as classificações. Mesmo que um conjunto desse formato pudesse ser definido, ele variaria e teria que ser reformulado para cada classificação feita. Indo além, há ainda grandes chances de que estas características sequer sejam conhecidas, já que a necessidade do uso de classificadores implica que não se tenha muitas pistas sobre que partes da estrutura original são responsáveis pelo comportamento buscado na classificação.

A abordagem destes problemas foi inspirada em um modelo de teste, descrito

na figura 5, usado para validar as assinaturas do tipo aCSM, apresentado no artigo de publicação da base de dados Propedia v.2.3, uma base contendo milhares complexos peptídeo-proteína, dentre outros dados [Martins et al. 2023]. Nele, as assinaturas foram utilizadas para classificar peptídeos em clusters de sequências já presentes na versão anterior do Propedia.

Estes clusters foram feitos combinando várias técnicas de alinhamento de sequências a técnicas de agruamentos de dados, agrupando todos os peptídeos da base em conglomerados contendo trechos de sequência análogos entre si [Martins et al. 2021]. Os clusters são representados por um centroide na forma de uma sequência que melhor se alinha a todos os elementos do cluster. Todos os peptídeos da base têm estrutura resolvida manualmente (com cristalografia de raios x, por exemplo), e em quase todos os clusters a sequência centroide corresponde a um peptídeo da base, com estrutura resolvida.

Assim, partindo do preceito da modelagem por homologia de que trechos de sequência análogos tendem a denotar trechos de estrutura análogos [Lesk 2020], podemos utilizar os clusters como o aspecto da estrutura original a ser inferido da estrutura aCSM. Embora não possamos inferir a estrutura original completa do peptídeo de origem da assinatura, a definição do cluster a que ele pertence permite uma visualização do centroide, que certamente possui trechos de sequência e de estrutura análogos aos da estrutura original do peptídeo. Para além disso, a definição do cluster a que a molécula da assinatura oferece outras opções de moléculas do cluster para realizar a comparação e, assim, definir elementos da estrutura que podem estar correlacionados com os trechos da assinatura responsáveis por determinados comportamentos. Por fim, Martins et al. demonstraram que as assinaturas aCSM apresentam bom desempenho na tarefa de classificação dos peptídeos nos clusters do propedia, conforme mencionado na figura 5 [Martins et al. 2023].

Com isso, temos todos os elementos necessários para montar a solução completa. O problema de decidir que aspectos da estrutura original correlacionar à assinatura é resolvido utilizando os próprios clusters de peptídeos do Propedia como rótulos. E, sabendo que a inferência dos clusters pode ser feita com bom desempenho utilizando classificação com as assinaturas aCSM, o problema de correlacionar as informações dos clusters à assinatura aCSM é equivalente ao problema de correlacionar os resultados de um classificador arbitrário às features do aCSM, que já foi tratado utilizando o workflow apresentado na figura 4, onde a correlação é feita utilizando técnicas de feature importance.

Desse modo, treinando um classificador de clusters de peptídeos e coletando suas feature importances, criamos uma interface de comparação de dados com as feature importances de outros classificadores arbitrários. A comparação entre as feature importances de ambos permite a definição de partes da assinatura importantes tanto para a inferência de um cluster com determinadas características estruturais quanto para a inferência do resultado do classificador analisado, criando assim uma linha de correlação indireta entre elementos da estrutura original e os resultados da classificação. Embora essa inferência não seja trivialmente realizada comparando apenas os valores das feature importances dos dois modelos, a coleta destas informações constituem as bases para a geração, em trabalhos futuros, de visualizações de dados que evidenciem de formas mais diretas estas relações, o que é exatamente o objetivo deste trabalho.

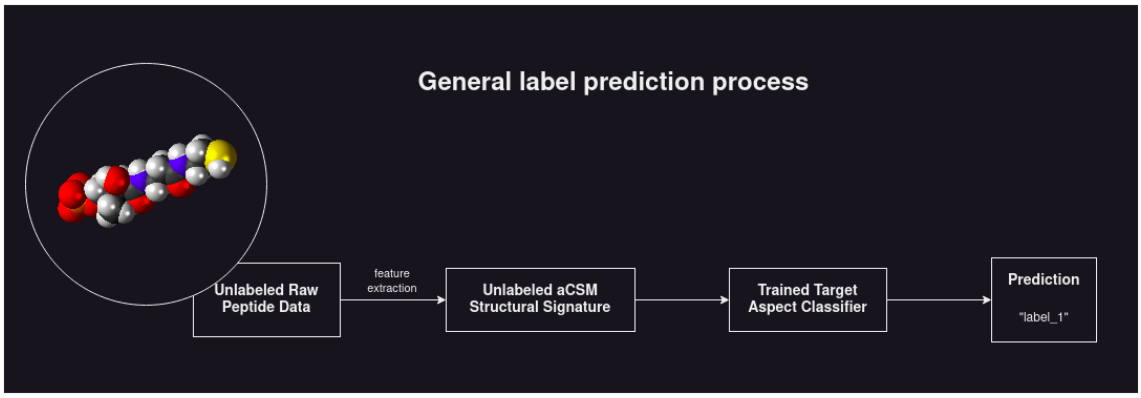

Figure 6. Processo geral de classificação de peptídeos não-rotulados

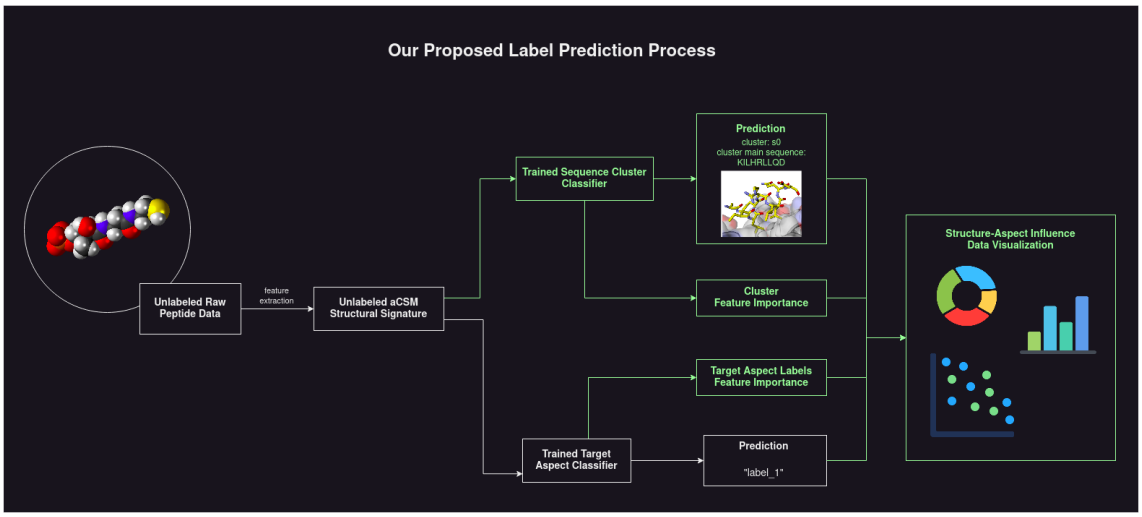

Figure 7. Processo proposto de classificação de peptídeos não-rotulados, com a geração de visualizações correlacionais da classificação feita. Adições ao processo original estão marcadas em verde

Dessa maneira, dado o processo identificado na figura 6, descrevendo o processo padrão de predição de um determinado aspecto alvo de um peptídeo não rotulado utilizando classificadores pré-treinados utilizando o workflow geral definido na figura 3, propomos um novo processo, no qual as etapas da solução descritas são adicionadas. Após coletada a assinatura aCSM, ela é classificada tanto pelo classificador do aspecto alvo quanto pelo classificador de clusters proposto, sendo ambos treinados utilizando o workflow modificado apresentado na figura 4, que realiza também a coleta das feature importances. Assim, tendo as informações da predição do aspecto alvo, da predição do cluster da sequência original e as feature importances de ambas, visualizações de dados correlacionando estes dados, as quais serão definidas em trabalhos futuros, podem ser geradas, apresentando informações semânticas sobre o papel da estrutura da molécula na classificação do aspecto alvo.

4. Implementação da solução

Dada a solução apresentada na seção anterior, a implementação e validação da mesma será feita treinando e testando um classificador de clusters e observando suas feature

importances, conforme o workflow da figura 4. A validação da solução depende tanto da validação do workflow proposto, demonstrando que as feature importances são um método efetivo em apontar correlações entre os dados da assinatura e os dados da classificação, quanto da geração do classificador de clusters com feature importances, que é parte essencial do processo definido na figura 7 para a análise de novas classificações. Como o classificador de clusters utiliza o workflow proposto, sua geração valida ambos os elementos necessários.

4.1. Base de Dados

A base de dados utilizada, conforme mencionado na seção anterior, foi a base Propedia v.2.3. A base é uma atualização da base Propedia, que contém complexos peptídeo-proteína conglomerados em clusters de sequência (utilizados na solução deste artigo), iterfaces e sítios de ligação [Martins et al. 2021]. A atualização 2.3, além de aumentar consideravelmente a base, alcançando cerca de 49.300 peptídeos, gerou também assinaturas aCSM para os peptídeos da base [Martins et al. 2023], que foram essenciais para a realização deste trabalho.

Neste contexto, como forma de replicar os resultados obtidos no artigo do Propedia v2.3, os mesmos dados de classificação usados por eles nos testes de validação das assinaturas foram utilizados para o treinamento do classificador de clusters implementados: os 1112 peptídeos dos 5 clusters mais populados da base. As assinaturas dos peptídeos da base foram coletadas considerando tamanho máximo de intervalo de 20 em Ångströms, e tamanho do cutoff de 0,2 em Ångströms. Assim, com 36 contagens de pares por cutoff, temos assinaturas de 3600 features.

4.2. Modelos Treinados

Como dois dos métodos de feature importance utilizados são gnósticos ao modelo utilizado, implementações de diferentes modelos para cada um deles foram necessárias. No caso da impurity reduction importance, que requer o uso de um modelo baseado em árvores, foi implementado um classificador baseado no algoritmo Random Forest. Já para o caso da permutation importance, que funciona com qualquer modelo, foi utilizado o modelo que apresentou o melhor desempenho no artigo do Propedia 2.3, o Gradient Boosting. Também como forma de replicar os resultados obtidos, tanto o Gradient Boosting quanto o Random Forest foram implementados com os mesmos iperparâmetros do artigo orignial, uma vez que ambos apresentaram um bom desempenho na classificação dos clusters.

Os modelos foram treinados tanto para a classificação multiclasse dos cinco clusters quanto para a classificação binária de cada cluster. Dessa forma, as feature importances gerais e as de cada classe puderam ser coletadas para análise. A partir dos dois modelos, as feature importances foram coletadas e os resultados serão analisados na seção seguinte.

5. Resultados

As figuras 8 e 9 apresentam os resultados das feature importances encontradas ao fim do workflow de classificação dos clusters. Cada plot representa um dos quatro métodos de extração de feature importances. O eixo y representa a porcentagem de importância, e no

eixo x está distribuída a assinatura aCSM. Conforme mencionado na seção 4.1, as assinaturas têm 3600 features, distribuídas em 36 padrões de pares contados em 100 cutoffs. Para melhor evidenciar os padrões, a assinatura foi reorganizada agrupando as features por padrões de pares. Assim, cada setor do eixo x representa as contagens nos 100 cutoffs do par denotado pelo padrão apresentado, sendo o a categoria de cada átomo do par representado por uma sigla e as siglas explicadas na legenda à direita.

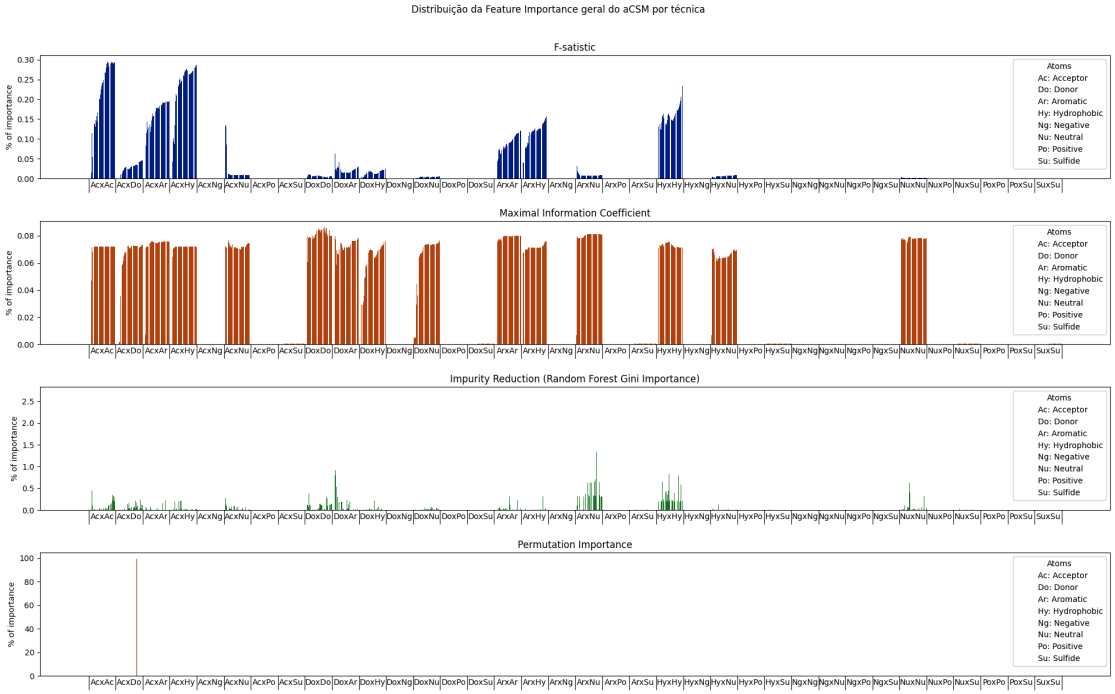

Figure 8. Gráfico de barras apresentando os resultados das feature importances da classificação geral nos cinco clusters. O eixo x representa as labels do aCSM agrupadasd em conjuntos de 100 cuttoffs, e o eixo y apresenta a porcentagem de importância de cada feature.

Na figura 8, são apresentados os resultados das feature importances da classificação geral nos cinco clusters, na forma de gráficos de barras. Neste caso, as métricas multivariadas não são interessantes, já que apresentam features relevantes na classificação das cinco categorias, o que não nos permite distinguir entre elas. No entanto, as métricas univariadas apresentam resultados interessantes sobre a importância das features no modelo. Como foi discutido, as métricas univariadas nos permitem detectar com certeza e descartar as features pouco importantes. Surpreendentemente, a importância das features foi completamente condicionada ao par de átomos observado. Se um par de átomos não tem importância, ele não terá independente do cuttoff observado. Analisando a MIC, que é mais fiel que a f-statistic, podemos ver que, para cada par de átomos, ou ele não apresentou importância em nenhum cuttof, ou ele apresentou importância na maioria dos cuttofs, se comportando de maneira cíclica na sequência de cuttoffs. A ausência de importância em pares específicos apresentados nos demonstra um padrão muito importante: todos os pares com 0 importância possuem ou um átomo positivo, ou um átomo negativo ou um átomo de sulfeto. Assim, fazendo o paralelo com as forças estabilizadoras de proteínas mencionadas anteriormente, essa base provavelmente não é formada por

nenhum resíduo eletrostaticamente carregado, como resíduos ligados a íons, por exemplo. Além disso, a ausência de pares com sulfetos denota a ausência de cisteínas na base utilizada. Apenas com os dados da feature importance já podemos inferir dados relativamente específicos sobre os resíduos dos peptídeos apenas a partir da assinatura.

Figure 9. Gráfico de linhas apresentando os resultados das feature importances da classificação binária de cada um dos cinco clusters. O eixo x representa as labels do aCSM agrupadasd em conjuntos de 100 cuttoffs, e o eixo y apresenta a porcentagem de importância de cada feature.

Observe agora a figura 9. Ela apresenta as feature importances dos modelos para a classificação binária de cada característica, na forma de gráficos de linhas. Conforme esperado, o mesmo padrão de importância observado nos dois primeiros gráficos da figura 8 aparece nos dois primeiros gráficos da figura 9. As mesmas features importantes ou não no primeiro caso assim o são no segundo, mas dessa vez odemos observar as nuances entre os diferentes clusters. Já na assinatura MIC é possível ver uma tendência de regiões diferentes a se correlacionar mais com tipos de pares diferentes. Pares relacionados a aceptores, aromáticos e hidrofóbicos tiveram maior importância na classificação do cluster s112, por exemplo. Avançando para as importâncias multivariadas, a gini importance foi capaz de capturar as peculiaridades das importâncias de alguns clusters, mas ainda gerou muita sobreposição e ruído, provavelmente devido ao seu viés associado. No entanto, a permutation importance dividiu de maneira certeira os cinco clusters em suas cinco features mais relevantes. Note que essa classificação, conforme discutido anteriormente, não significa que as demais features não sejam importantes. No entanto, as features que de fato são marcadas como muito importantes são, além de relevantes, únicas entre os dados observados. Ou seja, a permutation importance apontou para cinco pontos que caracterizam unicamente cada um dos cinco clusters classificados, com pouco

perigo de correlação entre si. Portanto, as feature importances foram capazes de identificar padrões únicos de cada classe, se mostrando assim como promissoras fontes de dados para visualizações de correlações entre as assinaturas do tipo aCSM e os resultados de modelos que as utilizam.

6. Conclusão e Perspectivas Futuras

Com este trabalho, podemos concluir que as feature importances se mostram como uma forma promissora de atribuir interpretabilidade às features do tipo aCSM utilizadas em modelos de classificação. A suite diversa de técnicas de feature importance permitiu, em diferentes aspectos e contextos, interpretar tanto padrões dos clusteres classificados quanto padrões estruturais gerais dos dados usados para análise. Associando este resultado aos workflows propostos ao longo do trabalho, esperamos, em trabalhos futuros, atingir o objetivo principal desta pesquisa de tornar acessíveis interpretações estruturais semânticas das clsasificações de proteínas. Com isso, no trabalho seguinte pretendemos trabalhar na interpretação das feature importances definidas neste trabalho. Isto inclui buscar formas de interpretar biologicamente cada campo da aCSM com relação às forças estabilizadoras nas quais eles foram baseados, bem como propor visualizações que evidenciem padrões nas comparações entre os pares de vetores de feature importance apontados. Por fim, pretendemos disponibilizar a geração dessas visualizações de maneira acessível a pesquisadores de diferentes áreas por meio de uma ferramenta web.

References

Altmann, A., Toloşi, L., Sander, O., and Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347.

Bishop, C. M. (2016). *Pattern recognition and machine learning*. Springer New York.

Kinney, J. B. and Atwal, G. S. (2014). Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111(9):3354–3359.

Lehninger, A. L., Nelson, D. L., and Cox, M. M. (2005). *Lehninger Principles of biochemistry*. W.H. Freeman.

Lesk, A. M. (2020). *Introduction to bioinformatics*. Oxford University Press.

Martins, P., Mariano, D., Carvalho, F. C., Bastos, L. L., Moraes, L., Paixão, V., and Cardoso de Melo-Minardi, R. (2023). Propedia v2.3: A novel representation approach for the peptide-protein interaction database using graph-based structural signatures. *Frontiers in Bioinformatics*, 3.

Martins, P. M., Santos, L. H., Mariano, D., Queiroz, F. C., Bastos, L. L., Gomes, I. d., Fischer, P. H., Rocha, R. E., Silveira, S. A., de Lima, L. H., and et al. (2021). Propedia: A database for protein–peptide identification based on a hybrid clustering algorithm. *BMC Bioinformatics*, 22(1).

Nembrini, S., König, I. R., and Wright, M. N. (2018). The revival of the gini importance? *Bioinformatics*, 34(21):3711–3718.

Patel, D. T., Honest, N., Vyas, P., and Patel, A. (2022). Univariate and multivariate filtering techniques for feature selection and their applications in field of machine learning.

Applying Data Science and Learning Analytics Throughout a Learner's Lifespan, page 73–93.

Peng, G., Nourani, M., Harvey, J., and Dave, H. (2020). Feature selection using f-statistic values for eeg signal analysis. *2020 42nd Annual International Conference of the IEEE Engineering in Medicine amp;amp; Biology Society (EMBC)*.

Pires, D. E., Ascher, D. B., and Blundell, T. L. (2013a). Mcsm: Predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, 30(3):335–342.

Pires, D. E., de Melo-Minardi, R. C., da Silveira, C. H., Campos, F. F., and Meira, W. (2013b). Acsm: Noise-free graph-based signatures to large-scale receptor-based ligand prediction. *Bioinformatics*, 29(7):855–861.

Pires, D. E., de Melo-Minardi, R. C., dos Santos, M. A., da Silveira, C. H., Santoro, M. M., and Meira, W. (2011). Cutoff scanning matrix (csm): Structural classification and function prediction by protein inter-residue distance patterns. *BMC Genomics*, 12(S4).

Rosa, R. S., Esteves, M. E., Bulla, A. C., and Silva, M. L. (2022). Preditores farmacocinéticos e toxicológicos in silico para via oral: Conheça e análise admetox. *BIOINFO 02 - Revista Brasileira de Bioinformática e Biologia Computacional*, page 83–97.

Ruczinski, I. (2002). Introduction to protein data bank format (lecture notes). Disponível em: `https://www.biostat.jhsph.edu/~iruczins/teaching/260.655/links/pdbformat.pdf`. Acesso em: 20/06/2023.