POC 2
aCSM-Explanation: Uma técnica de visualização molecular voltada à explicabilidade de classificações de instâncias baseadas em assinaturas estruturais 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 proven promising in the field of structural bioinformatics, facilitated by structural feature extraction techniques such as aCSM. The high performance of signature extraction techniques, however, is marked by a low interpretability of these results at the structural level. In this work, we propose aCSM-Explanation, a molecular visualization technique aimed at explaining labels assigned using aCSM signatures. Applying the technique to datasets labeled by structural clusters, the visualizations provided visually informative results, consistently highlighting commonalities among random instances within the same cluster.*

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 aCSM. O alto desempenho de técnicas de extração de assinatura é marcado, no entanto, por uma baixa interpretabilidade destes resultados a nível estrutural. Propomos neste trabalho a aCSM-Explanation, uma técnica de visualização molecular voltada à explicabilidade de rotulações feitas usando assinaturas aCSM. Aplicando a técnica a bases rotuladas por clûsteres estruturais, as visualizações apresentaram resultados visualmente informativos, destacando consistentemente pontos em comum entre instâncias aleatórias de mesmo clûster.*

1. Introdução

Técnicas de aprendizado de máquina têm sido cada vez mais aplicadas na solução de problemas de bioinformática, área de pesquisa voltada ao uso de métodos computacionais para a análise de dados biológicos [Lesk 2020]. Modelos de aprendizado podem ser atribuídos a diferentes tarefas, como realizar a predição de propriedades bioquímicas e farmacocinéticas de moléculas [Rosa et al. 2022], bem como detectar padrões que apontem para seus papéis nos processos biológicos em que estão inseridos [Sharan et al. 2007].

Para possibilitar o uso de dados moleculares como instâncias em modelos de machine learning, é necessário converter a representação espacial da molécula, 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 representativos destas estruturas, os quais denominamos assinaturas estruturais

[Martins et al. 2023]. Para isso, são utilizadas técnicas de extração de features voltadas ao contexto específico de estruturas biomoleculares, dentre as quais se destaca o atomic Cuttof Scanning Matrix. Também denominado aCSM, o atomic Cuttof Scanning Matrix é um algoritmo de extração de assinaturas estruturais baseado em contagens categorizadas de pares de átomos de uma molécula em diferentes intervalos de distância [Pires et al. 2013b].

Embora técnicas de extração de assinaturas estruturais como o aCSM demonstrem resultados muito promissores, sendo utilizado para o treinamento de modelos de alta performance em diversos âmbitos [de Castro Barbosa et al. 2022] [Portelli et al. 2020], nenhuma das técnicas atuais possui explicabilidade dos resultados dos modelos a nível estrutural da molécula. Visualizações de instâncias voltadas à explicabilidade de features já são um campo amplamente desenvolvido para modelos baseados em imagens, tendo como estado-da-arte técnicas de visualização de importância de features a nível de píxels [Nauta et al. 2021] [Chen et al. 2019]. De maneira análoga, existe um alto interesse científico na criação de técnicas de explicabilidade, a nível de átomos, de modelos voltados a assinaturas estruturais de biomacromoléculas, uma vez que, embora o comportamento de um polipeptídeo seja definido completamente por sua estrutura espacial e as propriedades farmacofóricas de seus átomos [Lehninger et al. 2005], estas características não são trivialmente identificáveis na estrutura original da molécula.

Deste modo, propomos, neste artigo, o aCSM-explainer, uma técnica inovadora de visualização estrutural de polipeptídeos voltada à explicabilidade, a nível de átomos e ligações, de rotulações feitas a partir de assinaturas aCSM. Partindo de uma base de assinaturas aCSM, junto a seus respectivos rótulos denotando uma característica de interesse, e da estrutura original e rótulo de uma instância específica, geramos uma visualização interativa da estrutura da instância evidenciando tanto a importância de cada átomo para a classificação da molécula em seu respectivo rótulo, por meio de uma coloração usando mapas de calor, quanto a importância de cada ligação (par de átomos), por meio de projeções de arestas com transparência ponderada pela importância do par.

O texto deste trabalho está organizado da seguinte maneira: na seção 2, são detalhados os trabalhos relacionados ao problema abordado e à solução proposta neste trabalho, bem como os conceitos fundamentais abordados em cada um deles. Na seção 3, a solução proposta neste projeto para o tratamento dos problemas apresentados é explicada em mais detalhes. Na seção 4, são apresentados os resultados encontrados após a aplicação da técnica na base de dados escolhida, bem como uma análise comparativa entre as visualizações de cada instância. Por fim, a seção 5 traz as conclusões tiradas a partir dos resultados encontrados, juntas às perspectivas de continuidade da pesquisa em trabalhos futuros.

2. Referencial Teórico

Esta seção se dedica a esclarecer os conceitos fundamentais associados à definição do problema e à compreensão da solução proposta, bem como analisar trabalhos que abordam problemas semelhantes ou propõem soluções semelhantes à proposta neste artigo.

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.

Proteínas e peptídeos são biomoléculas biológicas formadas por uma ou mais cadeias de aminoácidos, ligados por meio de ligações peptídicas. São moléculas profundamente 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.

O método atomic Cuttof Scanning Matrix, ou aCSM, consiste em um algoritmo de extração de assinaturas estruturais de proteínas e peptídeos, apresentando resultados promissores em diversos âmbitos, desde a predição de resistência a antibióticos em proteínas de membranas de bactérias [Portelli et al. 2020], à pesquisa de novos anti-virais à base de plantas contra os vírus da dengue e do zika-vírus [de Castro Barbosa et al. 2022]. A técnica também possui múltiplas variantes efetivas em diferentes contextos, como assinaturas voltadas à detecção de mutações, por exemplo [Pires et al. 2013a].

O aCSM constrói a assinatura estrutural realizando contagens categorizadas de pares de átomos de uma molécula. Dada uma molécula, com as posições tridimensionais de seus átomos, o aCSM define um número fixo de intervalos de distância disjuntos, denominados cuttofs, e realiza contagens de arestas dentro de cada intervalo. Além dos cuttofs, o aCSM categoriza as arestas conforme os pares de contatos que elas formam. Estes pares dizem respeito a oito propriedades farmacofóricas que cada um de seus átomos adjacentes pode performar não excluisvamente (aceptor, doador, hidrofóbico, aromático, positivo, negativo, neutro, sulfeto). Assim, uma aresta pode se enquadrar em 36 categorias possíveis (aceptor-aceptor, aceptor-doador ...). Para cada cuttof definido, são contadas as arestas dentro do limite do cuttof correspondentes a cada uma das 36 categorias, sendo o conjunto das contagens o vetor de features resultante[Pires et al. 2013b].

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]. Aplicando o MIC entre cada feature e a coluna de rótulos, calculamos o quão correlata a feature é aos rótulos da classificação, servindo assim como uma métrica de importância de features da assinatura.

Aprendizado de Máquina Interpretável, também denominado Inteligência Artificial Explicável (XAI), se define como qualquer método que promova a extração e representação de conhecimento relevante de um modelo de aprendizado de máquina em relação a relações contidas nos dados ou aprendidas pelo modelo, sendo um conhecimento considerado como relevante se ele proporcionar compreensão para um público específico sobre um problema escolhido. No contexto de XAI, se destacam as técnicas de explicabilidade voltadas a modelos baseados em imagens, em especial de redes convolucionais. Neste contexto, se destacam modelos visualmente explicáveis com precisão a nível de píxels, como o ProtoPNet [Chen et al. 2019] e o ProtoTree [Nauta et al. 2021], como técnicas de XAI consolidadas no âmbito de modelos de imagem.

Por fim, visualizações semelhantes à proposta foram apresentadas em trabalhos anteriores, porém apenas no contexto específico do problema de detecção de sítios de ligação [Gainza et al. 2019] [Rodrigues and Ascher 2022]. O problema de detecção de sítios de ligação consiste em encontrar a região na superfície de uma proteína onde existe potencial de conexão com uma determinada molécula ligante. Deste modo, de maneira semelhante à visualização resultante deste trabalho, soluções deste problema também se caracterizam como mapas de calor representando probabilidades sobre os átomos da molécula. No entanto, no caso do problema de detecção de sítios, o mapa é o resultado dos modelos de aprendizado para a solução do problema, enquanto a solução proposta neste trabalho se propõe a atribuir interpretabilidade a modelos de classificação arbitrários.

3. Solução Proposta

Esta seção descreverá os aspectos teóricos e práticos da abordagem tomada para a proposição da solução do problema apresentado. A seção de Workflow de Geração da Visualização detalhará os passos necessários para se gerar as visualizações propostas em um contexto de classificação de instâncias. A seção de Base de Dados descreverá a base de instâncias utilizada como teste. A seção de pré-processamento dos dados descreverá as etapas de pré-processamento aplicadas às instâncias para a geração das visualizações, e a seção de Viosualização das Instâncias descreverá de maneira teórica e prática as decisões tomadas quanto à maneira de demonstrar visualmente os dados pré-processados para cada instância.

3.1. Workflow de Geração da Visualização

Como a aCSM é uma técnica de extração de vetores de features, as assinaturas geradas utilizando esta técnica usualmente passam por um workflow padrão de treinamento e classificação. Este workflow se caracteriza por, dada uma base de instâncias e rótulos, coletar as assinaturas aCSM das instâncias, em seguida treinar um classificador arbitrário com a base de assinaturas e rótulos geradas e, por fim, coletar métricas de desempenho, conforme pode ser observado nas etapas em preto da figura 1. Variações deste workflow básico são utilizadas em outros trabalhos utilizando assinaturas do tipo CSM, incluindo o próprio trabalho de publicação da aCSM [Pires et al. 2013b] [Martins et al. 2023] [Pires et al. 2011] [Pires et al. 2013a]. Em contextos de uso prático, existe ainda um segundo worflow, após o treinamento do modelo, no qual, para cada instância não-rotulada e fora da base de treino, sua assinatura aCSM é coletada e utilizada para rotular a instância utilizando o modelo treinado [Portelli et al. 2020] [de Castro Barbosa et al. 2022] [Sharan et al. 2007].

Neste contexto, a aCSM-Explainer se propõe como uma técnica de explicabilidade de rotulações aplicável independentemente do classificador utilizado e do tipo de rotulação caracterizada na base. Desse modo, propomos uma alteração dos workflows padrão de treino e aplicação de modelos de classificação mencionados sem modificar nenhum dos passos usualmente aplicados, apenas adicionando novas etapas que possibilitem a geração das visualizações de instâncias paralelamente ao worflow original de classificação das mesmas.

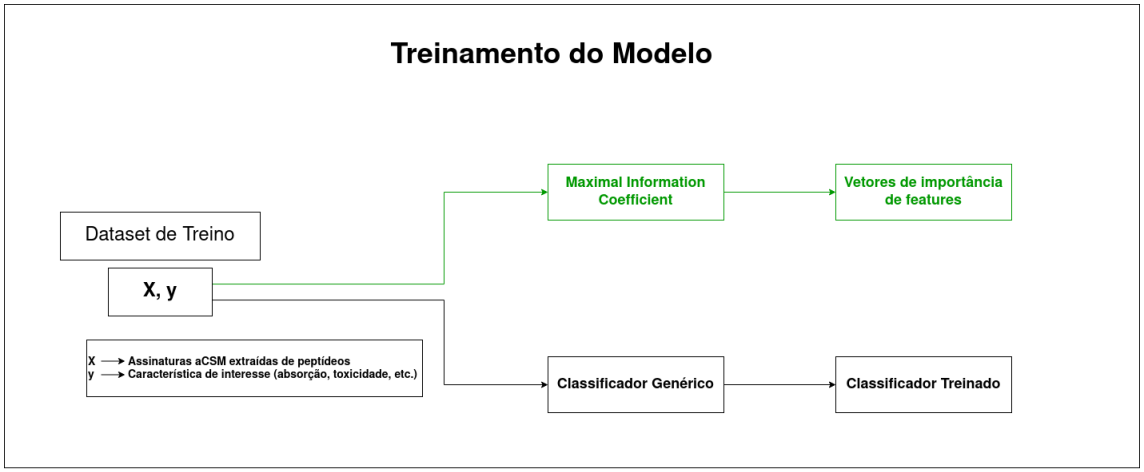


Figure 1. Workflow de treino proposto para a aplicação da aCSM-Explanation. Em preto, os passos presentes em um workflow padrão de treino de classificadores baseados em assinaturas aCSM. Em verde, a etapa extra de cálculo de importâncias necessária para a aplicação da técnica proposta.

Na figura 1, apresentamos o workflow de treinamento das instâncias proposto neste trabalho. Dado um dataset de treino, constituído de pares assinatura-rótulo, propomos, em adição ao workflow orignial de treinar um classificador genérico com a base, uma etapa de cálculo de importâncias de features a partir da base utilizando a métrica de Maximal Information Coefficient. esta métrica gera um conjunto de porcentagens representando a importância de cada feature na classificação da coluna de labels. São calculadas tanto as importâncias da classificação geral quanto as importâncias da classificação binária com relação a cada rótulo da base, e estes dados são armazenados junto ao modelo treinado. Note que a etapa de cálculo de importâncias não depende do modelo utilizado, dando liberdade de escolha ao usuário quanto ao modelo de classificador utilizado.

Já na figura 2, temos o workflow de classificação de uma instÇancia não rotulada utilizando o modelo treinado anteriormente. Em preto, temos o workflow padrão, no qual a assinatura da molécula é coletada, passada pelo classificador e o resultado é tomado como o novo rótulo da molécula. A este workflow adicionamos duas etapas, nas quais coletamos a label gerada pelo classificador, selecionamos o vetor de importâncias relativo à label gerada e utilizamos estas importâncias para gerar a visualização final proposta, a aCSM-Explanation da instância, na qual são ressaltadas quais regiões e ligações da molécula foram mais relevantes para o resultado da classificação. Por se tratar de uma técnica agnóstica ao modelo, visualizações de instâncias previamente rotuladas, incluindo instâncias da própria base de treino, também podem ser geradas dessa mesma maneira a partir de seus rótulos, não dependendo necessariamente do resultado de um classificador.

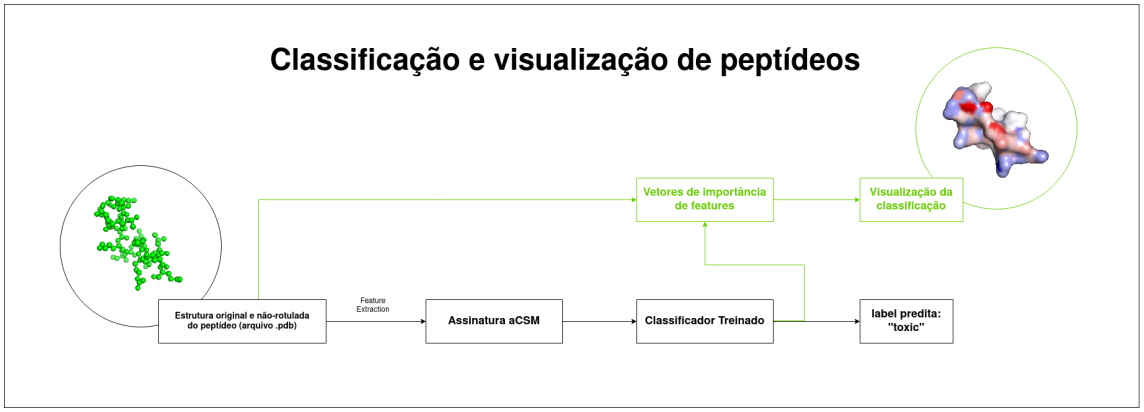


Figure 2. Workflow de classificação e geração da aCSM-Explanation para uma instância não-rotulada. Em preto, os passos presentes em um workflow padrão de classificação de instâncias. Em verde, as etapas extras necessárias para a geração da visualização proposta.

Estes workflows sumarizam as etapas necessárias para a geração das visualizações propostas neste trabalho.

3.2. Base de Dados

A base de dados utilizada 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, 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.

Esta base foi escolhida por já possuir as assinaturas calculadas e apresentar, em trabalhos anteriores, um bom desempenho em tarefas de classificação [Martins et al. 2023], diminuindo, assim, viéses nos resultados finais relativos à base utilizada e a detalhes de aplicação do aCSM. No trabalho citado, Martins et al. aplicaram as instâncias a tarefas de classificação nos clûsteres de sequência do próprio Propedia v2.3. Estes clusters foram gerados 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]. Deste modo, partindo do preceito da modelagem por homologia de que trechos de sequência análogos tendem a denotar trechos de estrutura análogos [Lesk 2020], esperamos poder comparar visualizações de instâncias de mesmos clûsteres e encontrar estruturas semelhantes ressaltadas nas visualizações.

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.

3.3. Pré-Processamento de Dados

Nesta seção, serão descritas todas as operações de pré-processamento e geração de dados aplicadas às instâncias da base de dados coletadas. Isso inclui o cálculo inicial de importâncias, estruturas de dados utilizadas para representação das instâncias e a geração das importâncias por elemento estrutural utilizadas diretamente nas visualizações. Note que, neste caso, a geração das assinaturas aCSM não foi descrita, uma vez que a base utilizada já possui as assinaturas calculadas.

3.3.1. Cálculo das Importâncias de Features

Conforme mencionado na seção 3.1, o cálculo das importâncias de cada feature foi feito utilizando a métrica de Maximal Information Coefficient. Esta métrica calcula, para uma base de dados rotulada, a quantidade múltua de informação entre a coluna de rótulos e cada uma das colunas de features, gerando ao fim um vetor de porcentagens representando a porcentagem de importância de cada feature na classificação da coluna de labels.

O processo é feito uma vez para a coluna original de labels, obtendo a importância total da classificação da base, e uma vez para cada label diferente da base. Isso é feito substituindo a coluna de labels original por uma coluna binária representando, para uma label x, quais instâncias correspondem a x e quais não correspondem. O vetor de importâncias gerado aplicando o MIC em relação à nova coluna gerada representa as importâncias de cada feature na classificação da label x. Com isso, conseguimos calcular separadamente as importâncias de classificação de cada label.

A escolha da MIC em detrimento de outras técnicas de cálculo de importância de features se deu principalmente devido ao seu forte poder de representatividade dos dados. Trabalhos anteriores avaliando o desempenho de quatro diferentes técnicas de importância de features aplicadas a assinaturas aCSM demonstraram que, dentre as quatro técnicas testadas, ela era a única capaz de detectar todas as features importantes para uma classificação. As demais técnicas testadas ou apresentaram um forte viés de modelo ou ressaltaram features presentes em poucas instâncias da base [Dutra 2023].

3.3.2. Estrutura de Dados de Representação das Instâncias

A representação da estrutura das instâncias de treino do modelo são fornecidas em um formato denominado PDB, definido pelos arquivos utilizados no Protein Data Bank, a maior base de dados públicos de polipetídeos existente [Burley et al. 2017]. A representação é feita de maneira hierárquica, na qual cada molécula é representada por uma estrutura, que pode possuir vários modelos de representação, que por sua vez possuem múltiplas cadeias, cada uma contendo múltiplos resíduos que, por fim, possuem átomos, com suas respectivas posições tridimensionais [Ruczinski 2002].

Embora a representação PDB seja extremamente descritiva quanto às subestruturas da molécula, ela se mostrou pouco eficiente para o cálculo dos valores necessários para a geração da visualização proposta. Assim, de maneira a facilitar o cálculo destes valores, geramos uma nova estrutura de dados de representação das instâncias baseada em

grafos. A representação em grafo é a mais adequada para o problema, visto que a aCSM também é uma assinatura baseada em grafos.

Deste modo, a estrutura foi definida para cada molécula como um grafo ponderado em vértices e arestas. no grafo, os vértices correspondem aos átomos da molécula e o grafo é completo, ou seja, existe uma aresta entre quaisquer dois pares de vértices. Tanto os vértices quanto as arestas são podenrados por pesos. As arestas possuem dois pesos: um correspondente à distância física entre seus vértices adjacentes, calculada a partir das posições tridimensionais de cada átomo, e outra correspondente à importância da aresta. Já os vértices possuem apenas um peso, correspondente à importância do vértice.

As importâncias dos vértices e arestas representam, respectivamente, as importâncias de átomos e importâncias de ligações, constituindo, assim, o cerne dos dados gerados para a montagem da visualização proposta. Seus valores são calculados a partir das importâncias de features calculadas utilizando o MIC, conforme descrito na seção anterior. Os detalhes sobre como o cálculo de cada uma destas importâncias é feito serão descritos nas duas seções seguintes.

3.3.3. Cálculo das Importâncias de Arestas

O cálculo das importâncias de arestas foi feito a partir das importâncias de features calculadas utilizando o MIC. Existe uma correlação quase direta entre as importâncias de features e as importâncias de arestas. Para compreendê-la, devemos considerar o significado semântico das features.

Conforme descrito na seção 2, o aCSM consiste em uma contagem categorizada de arestas em intervalos de distância. Assim, definidos os intervalos, denominados cuttofs, temos que, para cada cuttof, contamos o número de arestas dentro do cuttof correspondente a um de 36 pares de contatos possíveis entre átomos. A contagem de arestas para cada par de contatos dentro de cada intervalo constitui uma feature da assinatura.

Por outro lado, para cada aresta da molécula, temos como informação sua distância e os pares de contatos que seus átomos adjacentes formam. Deste modo, podemos selecionar o cuttof no qual a aresta está inserida e tirar a média das importâncias do cuttof para cada par de contatos possível. De maneira a selecionar apenas os contatos relevantes para a aresta selecionada, selecionamos apenas as importâncias de features correspondentes a pares de contatos que a aresta constitui, somando importãncia 0 para demais contatos. Com isso, temos uma maneira consistente de calcular a importância de cada aresta do modelo de grafos definido, que serão usadas para a geração da visualização de ligações.

3.3.4. Cálculo das Importâncias de Vértices

As importâncias de vértices, por sua vez, são calculadas a partir das importâncias de arestas. Dado que a assinatura aCSM é baseada em contagens de arestas, é natural que as importâncias de vértices sejam derivadas secundariamente. Para cada vértice, sua importância é definida como uma média das importâncias de suas arestas adjacentes. Por se tratar de um grafo completo, todos os vértices tomarão média do mesmo número de

arestas. Deste modo, calculamos de maneira consistente a importância de cada vértice do modelo de grafos definido, que serão usadas para a geração da visualização de átomos.

3.4. Visualização das Instâncias

Nesta seção, será descrita em detalhes a geração da visualização resultante da aCSM-Explanation. A visualização se caracteriza como uma visualização híbrida das importâncias de vértices e arestas feita utilizando a ferramenta PyMOL. Ao longo das seções, a ferramenta utilizada e a forma como foi feita a representação de cada importância serão detalhados, bem como será explicada a dinâmica complementar sobre a qual ambas as visualizações operam.

3.4.1. Ferramenta de Visualização

As visualizações finais foram geradas utilizando o PyMOL, uma ferramenta de visualização científica molecular voltada a representações de biomacromoléculas, especialmente de polipeptídeos e ácidos nucleicos. A ferramenta fornece múltiplas representações para uma mesma molécula ressaltando diferentes atributos semânticos da mesma, bem como ferramentas que permitem adicionar novos elementos às estruturas, o que foi essencial para a geração de diferentes representações ressaltando tanto átomos quanto ligações. O PyMOL representa a atual ferramenta estado-da-arte em representações moleculares de alta precisão [DeLano et al. 2002].

3.4.2. Visualização das Importâncias de Arestas

Conforme explicado na seção 3.3.3, as informações de importâncias de arestas geradas representam, na visualização, as importâncias de ligações na molécula. A visualização destes dados foi feita por meio de plots diretos de arestas sobre a estrutura da molécula. Cada aresta é representada por um cilindro (denominado stick) amarelo conectando fisicamente os átomos que ela liga.

A informação da importância de cada aresta é representada por meio da opacidade dos cilindros: quanto mais opaco, mais importante a aresta, e quanto mais transparente, menos importante, com arestas de importância nula não sendo mostradas. A representação do PyMOL utilizada neste caso foi a representação sticks and spheres, uma vez que ela permite a representação das ligações, na forma dos cilindros, ao mesmo tempo que traz uma representação dos átomos, no qual a importância de vértices também pode ser aplicada de maneira complementar.

As vantagens desta representação se dão principalmente pelo fato de que existe uma correlação quase direta entre as arestas da molécula e as features da assinatura, o que é especialmente relevante no contexto deste trabalho, onde se deseja explicar as features da assinatura aCSM. No entanto, as representações baseadas apenas em arestas tendem a ser muito densas e, com isso, pouco informativas, considerando que o grafo gerado é completo em arestas. Para contornar este problema, algumas medidas foram aplicadas à parte da visualização dedicada às arestas.

Primeriamente, as importâncias são mapeadas a um intervalo de opacidade [0, 1] de maneira que, dada uma opacidade máxima x, as importâncias são transformadas

linearmente para que a aresta mais importante tenha importância x, e as demais tenham importância linearmente relativa à maior. Isso permite que controlemos, com x, a densidade geral da visualização, ao mesmo tempo que a variância entre as importâncias é mantida.

Além disso, a visualização também permite personalização, de modo que, selecionando um par específico de contatos (conforme definido pela assinatura aCSM), apenas as arestas que correspondem a este contato são mostradas. Isso é especialmente efetivo para trazer informações complementares à importância de vértices, demonstrando se tipos de ligações específicos ocorrem entre regiões de alta importância, por exemplo. Um exemplo deste caso pode ser visto na figura 5(b).

3.4.3. Visualização das Importâncias de Vértices

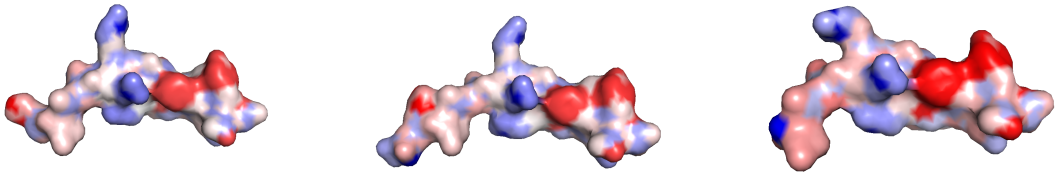
As importâncias de vértices, por sua vez, representam a importância de cada átomo na classificação do rótulo. Sua visualização é mais simples e direta, uma vez que a maioria das representações moleculares do PyMOL já apresentam a representação de cada átomo da molécula. Deste modo, as importâncias de vértices são representadas na molécula por meio de uma coloração em mapas de calor. Isso permite que a informação de importância seja representada de maneira consistente em cada átomo para qualquer representação que permita a sua visualização.

A coloração dos mapas de calor foi feita utilizando uma interpolação entre as cores azul, branco e vermelho: azul representa importâncias de valor 0, branco importâncias de valor 0.5 e vermelho as importâncias de valor 1. De modo semelhante às importâncias de arestas, às importâncias de vértices também é aplicada uma transformação linear para o intervalo [0,1], de maneira que o maior valor corresponda à importância 1 e os demais tenham importãncia proporcional à maior importância. Por fim, a representação de superfície foi utilizada para a visualização apenas de vértices, uma vez que ela ressalta melhor as diferenças de importãncias entre regiões.

As vantagens e desvantagens desta representação demonstram o quanto ela é complementar à visualização baseada em arestas apresentada anteriormente. Como vantagens, ela é concisa e simples, por se tratar apenas de uma coloração da molécula, permitindo que as próprias representações padrão da molécula sejam usadas. Além disso, ela permite que regiões mais ou menos importantes da molécula sejam identificadas, possibilitando seu uso para a identificação de subestruturas de interesse na molécula. Como desvantagens, ela não traz informações sobre as ligações das quais os átomos das regiões demarcadas mais importantes participam. É evidente, deste modo, que a visualização em arestas apresenta justamente as informações sobre ligações das quais a visualização de vértices carece, enquanto a representação em vértices possui uma representação visualmente concisa pela qual o usuário possa se guiar ao observar o decorrer das arestas. Deste modo, ambas as visualizações se comportam de modo complementar e, juntas, podem evidenciar uma série de informações importantes sobre a molécula.

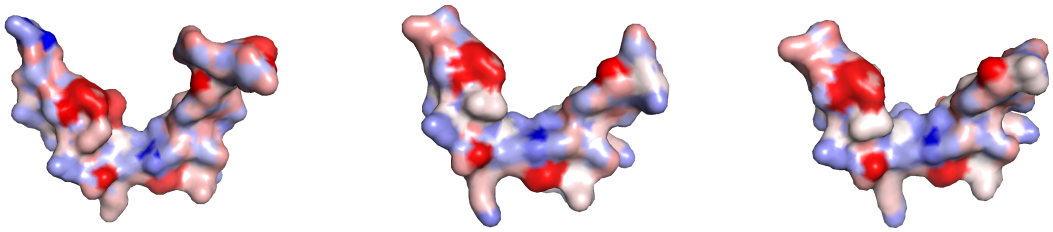
4. Resultados

De modo a testar a qualidade das visualizações geradas pela técnica aCSM-Explanation descrita, foram selecionadas, para cada clûster, amostras aleatórias da base, e as



(a) Peptídeo 3l9l _ C (b) Peptídeo 4yxr _ I (c) Peptídeo 6erv _ B

Figure 3. aCSM-Explanations de três peptídeos com respeito à importância de vértices de sua classificação no cluster s1



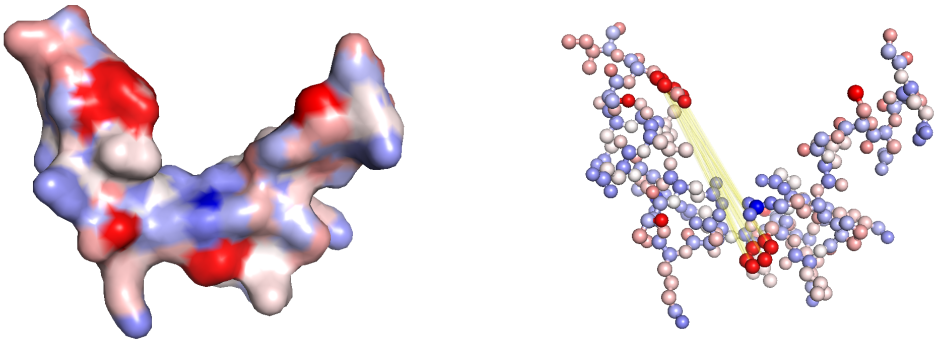
(a) Peptídeo 1ucy _ J (b) Peptídeo 2c93 _ A (c) Peptídeo 3si4 _ L

Figure 4. aCSM-Explanations de três peptídeos com respeito à importância de vértices de sua classificação no cluster s162

visualizações foram comparadas entre si. Por se tratarem de clústeres de alinhamento de sequências, é esperado que, para moléculas de um mesmo clúster, subestruturas em comum entre as moléculas estejam majoritariamente representadas como regiões importantes em todas elas. De modo semelhante, é esperado que subestruturas particulares de cada molécula sejam representadas com menos importância.

Na imagem 3, temos visualizações de três peptídeos da base pertencentes ao clúster s1: 3191_C, 4yxr_I e 6erv_B. As visualizações apresentam apenas a importância de vértices de cada molécula com respeito à classificação em seu clúster, com as três sendo apresentadas em representação de superfície molecular. É evidente que, para as três, as subestruturas em vermelho vivo, que representam os átomos mais importantes para a classificação de cada uma no clúster s1, representam subestruturas comuns a todas elas, uma vez que são praticamente idênticas entre si nas três moléculas. Isso se evidencia em especial nas regiões em vermelho vivo à direita de cada molécula, que são praticamente idênticas nos três casos. Por outro lado, regiões que diferem entre si em cada molécula, como a estrutura em forma de ponta na região superior-central de cada uma das moléculas, se apresentam na cor azul, indicando baixa importância.

O mesmo pode ser visto para instâncias do clúster s162. Na imagem 4, temos três instâncias da base pertencentes ao clúster s162: 1ucy_J, 2c93_A e 3si4_L. De modo muito semelhante às amostras da imagem 3, também podemos ver destacadas em vermelho vivo regiões representando subestruturas comuns às três moléculas, e em azul regiões variáveis entre elas. No entanto, neste caso há uma região proeminentemente vermelha, e por conseguinte importante, no canto superior-direito da imagem 4(a), embora a mesma região não possa ser visualizada nas demais três amostras. É importante ressaltar, neste caso, que as três amostras selecionadas não são necessariamente representativas de toda a base de peptídeos pertencentes ao clúster s1. É possível que a região destacada represente



(a) vertex importance (b) vertex + edge importance

Figure 5. aCSM-Explanations do peptídeo 2c93_A com respeito a sua classificação no cluster s162. Em 5(a), visualização de importância de vértices. Em 5(b), visualização mista apresentando, além da importância de vértices, importância de arestas apenas para pares de ligação do tipo aromático-aromático.

uma subestrutura comum a mais amostras do clûster não visualizadas neste trabalho.

Por fim, na imagem 5 temos um detalhamento das visualizações geradas para o peptídeo 2c94_A, o mesmo representado na imagem 4(b). A imagem 5(a) contém a mesma visualização observada anteriormente para a molécula em 4(b), enquanto na imagem 5(b) temos uma variação da visualização ressaltando, além das importâncias de vértices, as importâncias de arestas da classificação. A visualização foi personalizada para mostrar apenas arestas representando pares de contato do tipo aromático-aromático. A visualização de arestas evidencia, em amarelo, que todos os átomos importantes da molécula são ligados entre si por contatos do tipo aromático-aromático. Deste modo, é evidente o caráter complementar da visualização categorizada de importâncias de arestas à importância de vértices apresentanda, com ambas constituindo a aCSM-Explanation como uma técnica consistente de atribuição de interpretabilidade aos resultados de classificações feitas utilizando a assinatura aCSM.

5. Conclusão e Perspectivas Futuras

Com este trabalho, propomos a aCSM-Explainer como uma técnica inovadora e efetiva de visualização explicável de rotulações aplicadas a moléculas. Amostras de mesmos clûsteres evidenciaram visualmente, por meio da técnica, características comuns às amostras, bem como características particulares de cada amostra, de maneira consistente. Visualizações de arestas possibilitaram a representação visual de informações complementares sobre os tipos de ligação priorizados pelas assinaturas. Além disso, a técnica se consolida como a primeira a oferecer visualizações explicáveis de moléculas compatível com classificadores arbitrários e em contextos arbitrários de classificação.

Como perspectivas futuras, pretendemos validar a ferramenta com novas bases de dados. Embora a técnica tenha sido testada apenas com peptídeos, ela também é compatível com polipeptídeos de tamanhos maiores, tornando o seu potencial uso em proteínas um objeto interessante de pesquisa. Também existe um interesse em observar o comportamento das visualizações em bases que utilizam outras rotulações que não os clûsters estruturais, de modo a evidenciar padrões estruturais menos trivialmente associados aos rótulos. Por fim, dado o caráter multidisciplinar da pesquisa, é de interesse

de pesquisadores de áreas pouco envolvidas com a computação, como pesquisadores exclusivamente da área de biologia molecular, a criação de uma ferramenta de interface amigável para a aplicação da aCSM-Explainer em suas bases de dados, sem a necessidade de programação, tornando esta também uma importante perspectiva de continuidade do objetivo da pesquisa.

References

Burley, S. K., Berman, H. M., Kleywegt, G. J., Markley, J. L., Nakamura, H., and Velankar, S. (2017). Protein data bank (pdb): the single global macromolecular structure archive. *Protein crystallography: methods and protocols*, pages 627–641.

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., and Su, J. K. (2019). This looks like that: Deep learning for interpretable image recognition. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

de Castro Barbosa, E., Alves, T. M., Kohlhoff, M., Jangola, S. T., Pires, D. E., Figueiredo, A. C., Alves, A., Calzavara-Silva, C. E., Sobral, M., Kroon, E. G., and et al. (2022). Searching for plant-derived antivirals against dengue virus and zika virus. *Virology Journal*, 19(1).

DeLano, W. L. et al. (2002). Pymol: An open-source molecular graphics tool. *CCP4 Newsl. Protein Crystallogr*, 40(1):82–92.

Dutra, A. (2023). Poc 1: Proposta de visualização de assinaturas estruturais de peptídeos geradas usando o método acsm.

Gainza, P., Sverrisson, F., Monti, F., Rodolà, E., Boscaini, D., Bronstein, M. M., and Correia, B. E. (2019). Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, 17(2):184–192.

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).

Nauta, M., van Bree, R., and Seifert, C. (2021). Neural prototype trees for interpretable fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14933–14943.

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).

Portelli, S., Myung, Y., Furnham, N., Vedithi, S. C., Pires, D. E., and Ascher, D. B. (2020). Prediction of rifampicin resistance beyond the rrdr using structure-based machine learning approaches. *Scientific Reports*, 10(1).

Rodrigues, C. H. and Ascher, D. B. (2022). Csm-potential: Mapping protein interactions and biological ligands in 3d space using geometric deep learning. *Nucleic Acids Research*, 50(W1).

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.

Sharan, R., Ulitsky, I., and Shamir, R. (2007). Network-based prediction of protein function. *Molecular Systems Biology*, 3(1).