Relatório: *Protein-Ligand Scoring with Convolutional Neural Networks* - Avaliação da tarefa de Virtual Screening

Guilherme R. Graeff [1]

[1] *Structural Bioinformatics and Computational Biology Lab - SBCB, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil.*

Correspondence*:
Guilherme Rafael Graeff
guilherme.graeff@ufrgs.br

1 INTRODUÇÃO

Este trabalho faz parte da avaliação da disciplina de Algoritmos Para A Bioinformática E Biologia Computacional. Compreende uma apresentação sobre uma aplicação de Redes Neurais Convolucionais para a tarefa de *Virtual Screening* [5]. Explorando a abordagem utilizada para a representação dos dados de estrutura molecular, analisando os *datasets* escolhidos para o treinamento do modelo e verificando os resultados obtidos pelo trabalho.

O trabalho referência possui dois conjuntos de treinamento com objetivos distintos: um para a predição de pose e o outro para *virtual screening*. Este relatório analisa o desenvolvimento do modelo treinado nos dados referentes ao *Virtual Screening*. A aplicação de algoritmos está presente em diversas etapas daquele trabalho, seja na preparação dos dados, aplicação de ferramentas, desenvolvimento e avaliação do modelo ou até mesmo em etapas de automatização do processo. Esta integração entre algoritmo e problema de pesquisa biológico adquire complexidade de ambos os campos do conhecimento, contribuindo para o objetivo deste relatório, explicar esta aplicação especifica.

O restante do texto está organizado da seguinte forma: a próxima seção apresenta a área de Aprendizado de Máquina e *Virtual Screening*. Na seção 3 são apresentados os problemas da pesquisa . A descrição dos dados utilizados esta presente na Seção 4. A seção 5 está dedicada a apresentação do modelo e à otimização do mesmo. Em seguida, são discutidos os resultados do trabalho na seção 6. A Seção 7 conclui o relatório.

2 CONTEXTUALIZAÇÃO

Esta seção apresenta conceitos presentes no trabalho, acompanhado de uma breve explicação importante para a compreensão da técnica utilizada.

2.1 Aprendizado de Máquina - Redes Neurais Convolucionais

Rede Neural Artificial [1] é uma técnica de Aprendizado de Máquina capaz de compreender padrões presentes em determinado conjunto de dados, possui conceitos inspirados em neurônios biológicos, possui também a capacidade de aprender a partir de uma função de erro. O objeto de estudo do trabalho minimiza a perda logística multinomial utilizando uma variante da descida gradiente estocástica (SGD) e *backpropagation* para o treinamento. Uma Rede Neural Convolucional (CNN) [3] é uma rede neural que possui camadas de convolução e camadas *pooling*, esta técnica é amplamente utilizada para o reconhecimento de imagens no campo da Visão Computacional[3]. A convolução consegue captar informação conformacional do dado, como por exemplo arestas de objetos em uma imagem, o mesmo se

aplica quando o dado possui mais dimensões. Já a camada de *pooling* é capaz de reduzir a dimensionalidade do dado para que operações sejam realizadas neste 'menor' espaço, ou seja, esta seria a entrada para a rede neural totalmente conectada no fim da arquitetura da rede. A convolução consegue capturar as características que definem o modelo, então não é necessária a extração de *features* relevantes do modelo.

2.2 Atracamento molecular com Smina

O algoritrmo utilizado para *Docking* molecular foi o *Smina*[2], uma implementação derivada do Autodock Vina [7]. Neste contexto, dois conceitos são importantes para a identificação de um exemplo positivo e um exemplo negativo, são chamadas *decoy* aqueles que são exemplos negativos pois são moléculas que não possuem afinidade com o receptor, ao contrário das moléculas ativas que possuem afinidade com o receptor. Um 'Alvo' por sua vez é a molécula receptora do ligante, esta fica estática durante o *docking*. O que são alvos? O que são Actiive Molecules? O que são decoys?

Falar sobre o Autodock Vina Scoring Function utilizado através do Smina.

3 DESAFIOS NA APLICAÇÃO DE CNNS AO VIRTUAL SCREENING

Reforça o objetivo principal do trabalho, explorar o campo de estudos

Utilizar algoritmos de *machine learning* como função de score recebeu destaque em pesquisas, explorar novos métodos se fez necessário para o avanço da pesquisa na área.

Falar especificamente sobre o problema de docking talvez, da scoring function, por que utilizar esta técnica?

4 CARACTERÍSTICAS E PREPARAÇÃO DOS CONJUNTOS DE DADOS

Há mais de um conjunto de dados utilizado no trabalho, por conta das diferentes tarefas que se busca realizar, este trabalho apenas explica os dados utilizados para a tarefa de *Virtual Screening.*

4.1 Dados de treinamento

A tarefa de *Virtual Screening* faz uso do Databese of Usefull Decoys - Enhenced (DUD-E)[4]. São 102 alvos(proteínas), 20000 moléculas ativas(exemplos positivos) e mais de um milhão de moléculas do tipo *decoy*(exemplo negativo). Este banco de dados não possui a cristalografia da pose dos ligantes, embora possua uma referência do complexo disponível.

4.1.1 Gerando poses para treinamento

São geradas poses de ligantes para moléculas ativas e *decoy* utilizando *Docing with Smina*, *Smina* utiliza a função de score do *Autodock Vina* [2]. A molécula é posicionada na posição de referência do alvo, o *docking* acontece em uma caixa que possui 8Å centrada à este ligante de referência. Caso não exista um ligante de referência, então é utilizado um *script* que define a posição da caixa.

Os ligantes são atracados com a referência utilizando os argumentos padrão do *smina* para os parâmetros *exhaustiveness* e *sampling*. Então é selecionada a pose com o melhor ranking definido pela função de score do *Autodock Vina*. O tamanho final dos dados de treinamento são de 22.645 exemplos positivos e 1.407.141 exemplos negativos. O valor expressivo de exemplos negativos se dá melo maior número de *decoy* presente no conjunto de dados.

4.2 Dados de validação

Ainda

5 APRESENTAÇÃO DA TÉCNICA

Utilizar Funções de Score baseadas em Redes Neurais Convolucionais é uma maneira compreensiva de representar a estrutura tridimensional de uma interação proteína e ligante. A técnica em questão utiliza um *grid* de densidade de átomos [5] gerados através da estrutura, a biblioteca *libmolgrid* [6] é responsável pela transformação dos dados. Este tipo de modelo consegue aprender as principais características, relativas ao atracamento, da interação entre proteína e ligante [5]. Para a tarefa de *Virtual Screening* é treinada e otimizada para diferenciar ligantes e não-ligantes conhecidos. A técnica abordada no trabalho demonstra competitividade em relação a outras funções de *scoring*. O modelo foi otimizado utilizando a técnica de *clustered cross-validation*

O trabalho também apresenta uma maneira de visualizar os resultados obtidos

6 DISCUSSÃO DE RESULTADOS

Resultados aqui, comparações retiradas do próprio trabalho, detalhadas e comentadas

7 CONCLUSÃO

Este trabalho Conclui o relatório com uma reflexão simples, sobre o processo de aprendizagem, se referindo a dificuldade na realização do trabalho, sobre a mudança de planos, e sobre a apresentação e compreensão do conteúdo.

REFERENCES

[1]J. Feldman and R. Rojas. *Neural Networks: A Systematic Introduction*. Springer Berlin Heidelberg, 2013.

[2]David Ryan Koes, Matthew P Baumgartner, and Carlos J Camacho. Lessons learned in empirical scoring with smina from the csar 2011 benchmarking exercise. *Journal of chemical information and modeling*, 53(8):1893—1904, August 2013.

[3]Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[4]Michael M. Mysinger, Marco Carchia, John J. Irwin, and Brian K. Shoichet. Directory of Useful Decoys, Enhanced (DUD•E): Better Ligands and Decoys for Better Benchmarking. *Journal of Medicinal Chemistry*, July 2012.

[5]Matthew Ragoza, Joshua Hochuli, Elisa Idrobo, Jocelyn Sunseri, and David Ryan Koes. Protein–ligand scoring with convolutional neural networks. *Journal of Chemical Information and Modeling*, 57(4):942–957, 2017. PMID: 28368587.

[6]Jocelyn Sunseri and David Ryan Koes. libmolgrid: Gpu accelerated molecular gridding for deep learning applications, 2019.

[7]Oleg Trott and Arthur J. Olson. Autodock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2):455–461, 2010.