Redes Neurais Convolucionais e Virtual Screening

Guilherme Rafael Graeff

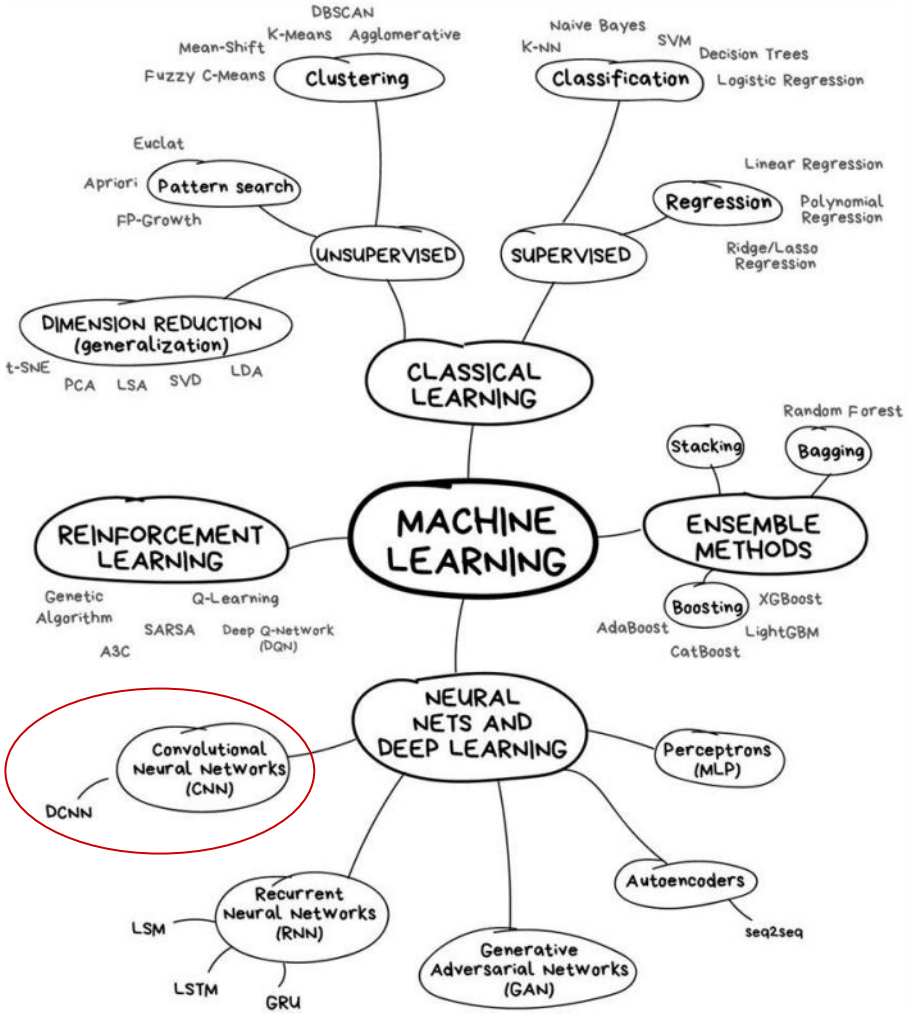
Índice

Protein-Ligand Scoring with Convolutional Neural

2017

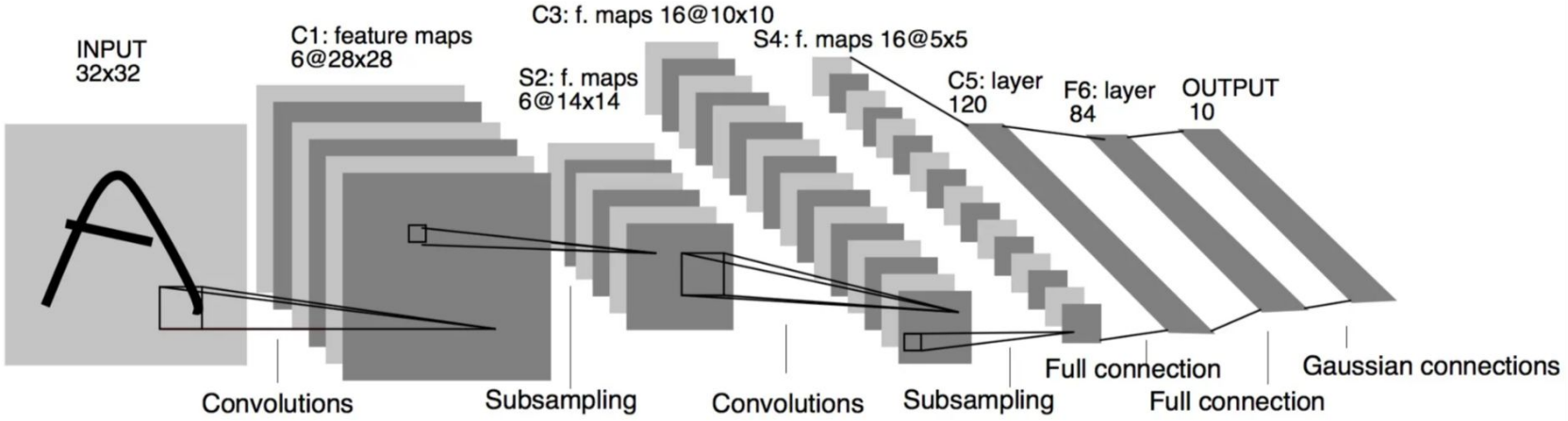
Protein-Ligand Scoring with Convolutional Neural Networks

Matthew Ragoza,[†,‡] Joshua Hochuli,[‡,¶] Elisa Idrobo,[§] Jocelyn Sunseri,[∥] and David Ryan Koes[*,∥]


MACHINE LEARNING
CLASSICAL LEARNING
UNSUPERVISED
Clustering
DBSCAN
K-Means
Agglomerative
Mean-Shift
Fuzzy C-Means
Pattern search
Euclat
Apriori
FP-Growth
DIMENSION REDUCTION (generalization)
t-SNE
PCA
LSA
SVD
LDA
SUPERVISED
Classification
Naive Bayes
K-NN
SVM
Decision Trees
Logistic Regression
Regression
Linear Regression
Polynomial Regression
Ridge/Lasso Regression
REINFORCEMENT LEARNING
Genetic Algorithm
A3C
SARSA
Q-Learning
Deep Q-Network (DQN)
ENSEMBLE METHODS
Stacking
Random Forest
Bagging
Boosting
AdaBoost
XGBoost
LightGBM
CatBoost
NEURAL NETS AND DEEP LEARNING
Convolutional Neural Networks (CNN)
DCNN
Perceptrons (MLP)
Recurrent Neural Networks (RNN)
LSM
LSTM
GRU
Autoencoders
seq2seq
Generative Adversarial Networks (GAN)

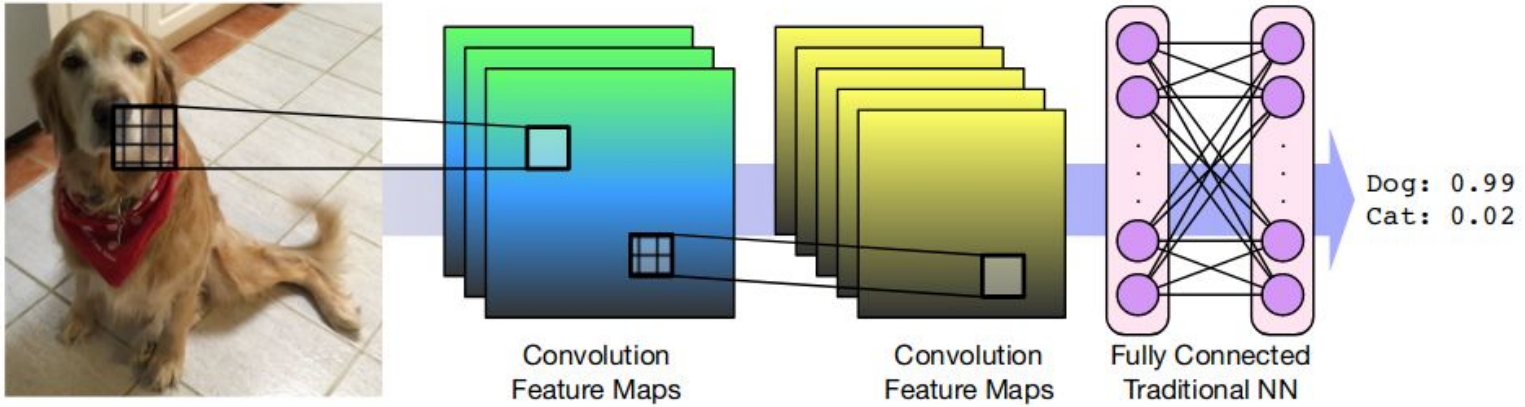
Redes Neurais Convolucionais

- Neurônios
- Camadas ocultas
- Funções de ativação
- Função de perda
- *Forward* Pass
- *Backpropagation*
- Herda as características de uma rede neural artificial
- **Camada de convolução**
- **Camada de pooling**



Redes Neurais Convolucionais

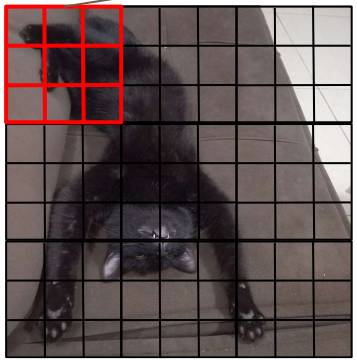
- Convolução

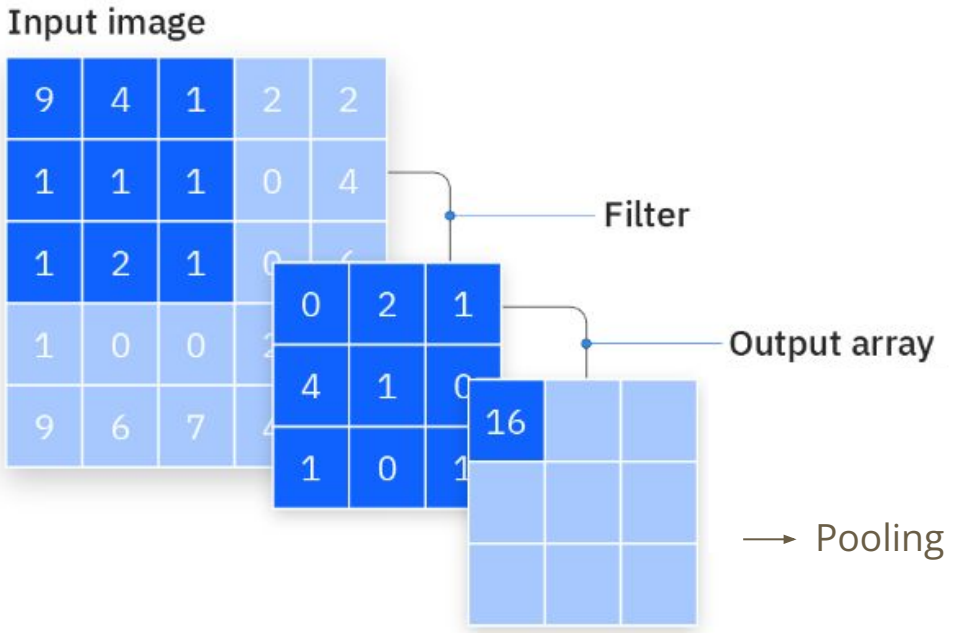


Protein–Ligand Scoring with Convolutional Neural Networks - 10.1021/acs.jcim.6b00740

Redes Neurais Convolucionais

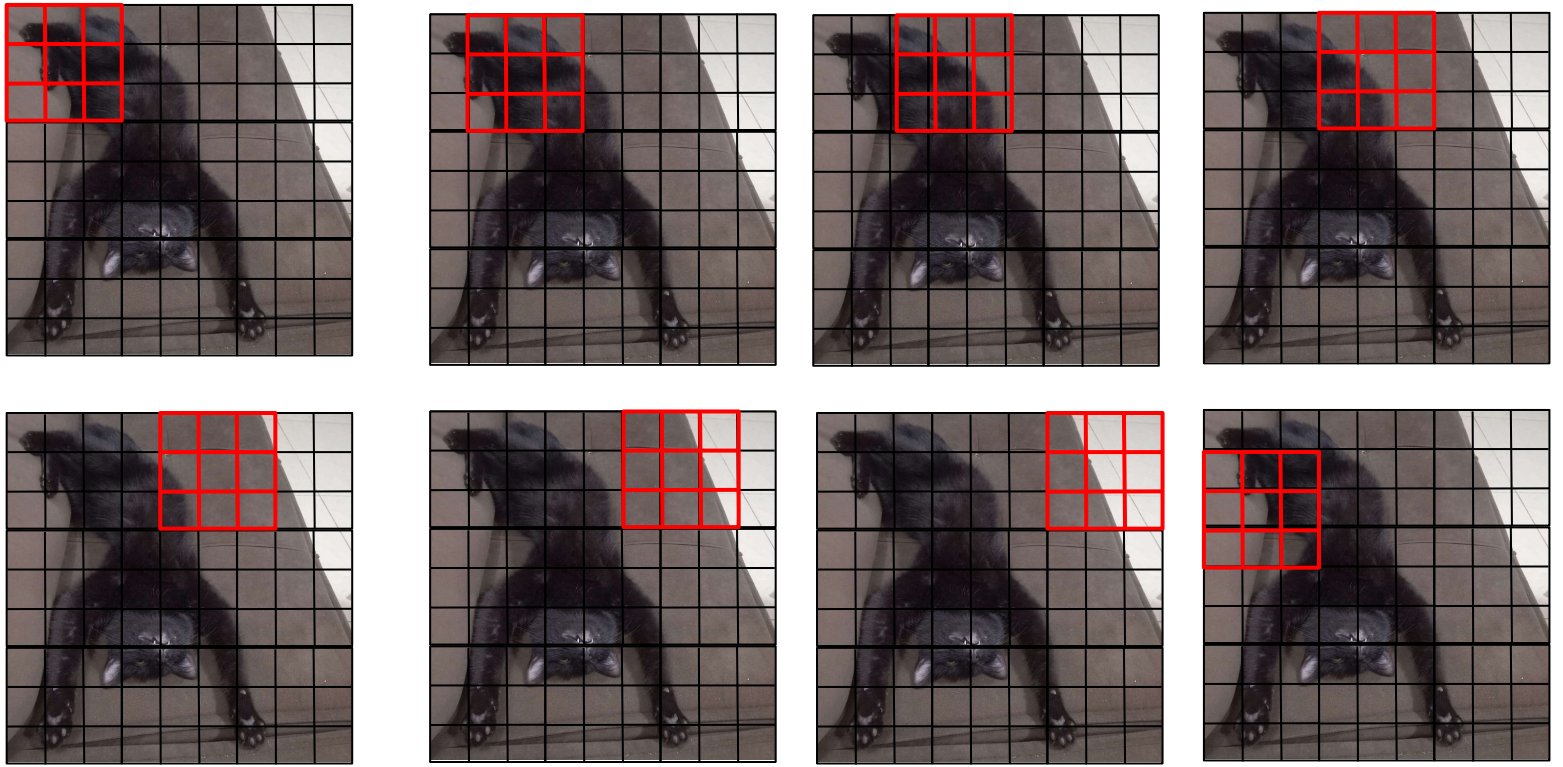
- Kernel/Filter





https://www.ibm.com/br-pt/topics/convolutional-neural-networks

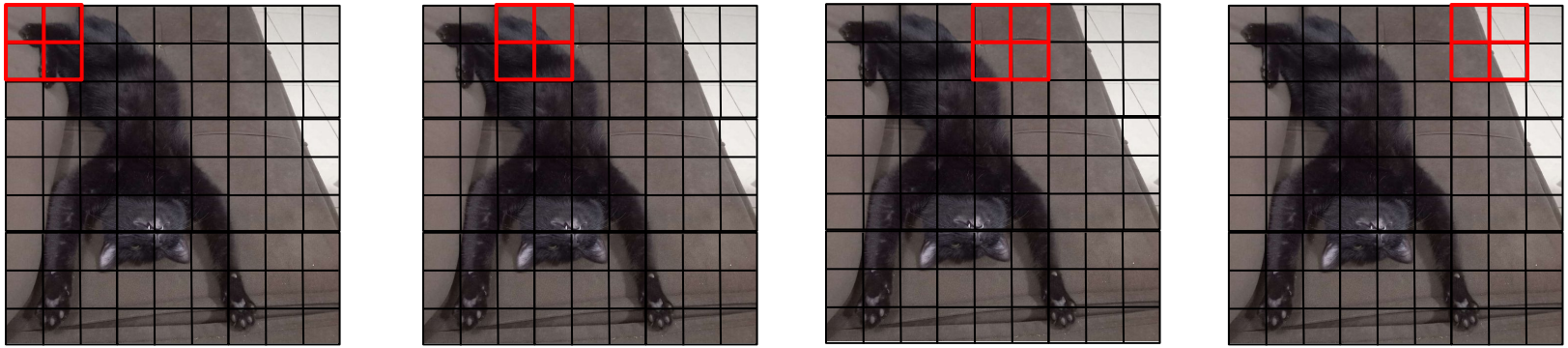
Redes Neurais Convolucionais



Redes Neurais Convolucionais

Pooling

- Average pooling, Max pooling, Avg-TopK pooling...



Redes Neurais Convolucionais

Transfer learning, os pesos dos filtros podem ser inicializados a partir de outra rede.

Backpropagation, atualização dos pessos do kernel (https://www.youtube.com/watch?v=z9hJzduHToc)

Virtual Screening

Triagem virtual busca explorar uma variedade de compostos.

Docking

Prevê a orientação preferencial do ligante, em relação ao alvo

Scoring Function

Prevê aproximadamente a afinidade da ligação na etapa de Docking

Actives/Decoys/Alvo

Actives - exemplo positivo, interage com o alvo

Decoy - grupo controle negativo

Alvo - estrutura alvo - proteína

Proteína e ligante

https://doi.org/10.2210/pdb3F3E/pdb

Descrição dos conjuntos de Dados

- Se busca avaliar o modelo treinado de diferentes formas, além do conjunto de dados utilizado para o treinamento (DUD-E), foram utilizados outros dois conjuntos para **validação do modelo** na tarefa de *virtual screening*.

- São dados estruturais da proteína e dos ligantes, os dados passam por um processo de atracamento.

Dados: DUD-E

A tarefa de *Virtual Screening* faz uso do *Databese of Usefull Decoys - Enhenced* (DUD-E).

102 alvos(proteínas),

20000 moléculas ativas e mais de

um milhão de moléculas do tipo *decoy.*

São geradas poses de ligantes para moléculas **ativas** e **decoy** utilizando *Docking with Smina* utiliza a função de score do **Autodock Vina**.

A molécula é posicionada na posição de referência do alvo

Caixa que possui 8Å centrada à este ligante de referência. Caso não exista um ligante de referência, *script*.

Então é selecionada a pose com o melhor ranking por afinidade (KCal/mol)definido pela função de score do Autodock Vina.

22.645 exemplos positivos e **1.407.141 exemplos negativos**. O valor expressivo de exemplos negativos se dá pelo maior número de *decoy* presente no conjunto de dados.

Dados: Independent Test Set

O trabalho opta por avaliar a **acurácia** da **classificação** com um conjunto de dados de teste independente, garantindo que dados utilizados no treinamento não estejam presentes no teste.

Um conjunto foi gerado através do *ChEMBL* por Riniker e Landrum, seguidos por Heikamp e Bajorath.

O outro conjunto de dados é um subconjunto dos dados presentes em *maximum unbiased validation* (MUV) dataset que é baseado nos dados de bioatividade presentes no PubChem.

Os conjuntos são filtrados

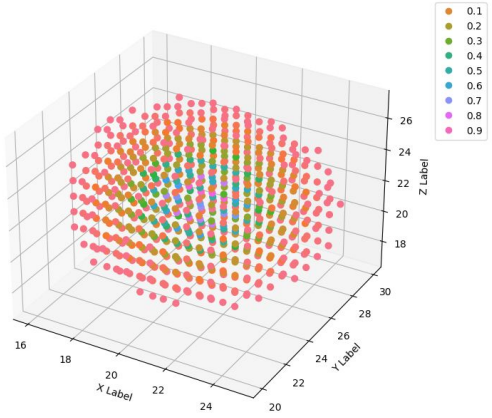
- Remoção de qualquer alvo que possuam 80% ou mais de identidade de sequência com um alvo de treinamento.
- Complexo de ligação contém o alvo disponível no *Protein Data Bank*. Estas estruturas são utilizadas para gerar poses atracadas em um sítio de ligação conhecido.
- Dentre outros filtros

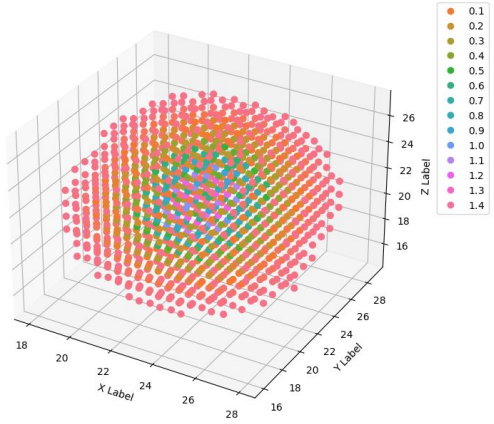
Após filtrar os dados, os conjuntos de testes independentes para a tarefa de V*irtual Screening* consiste de **13 alvos** provenientes de Riniker e Landrum ChEMBL, e **9 alvos provenientes do conjunto MUV.**

Reflexões Sobre a Aplicação

$$A(d,r) = \begin{cases} e^{-\frac{2d^2}{r^2}} & 0 \le d < r \\ \frac{4}{e^2 r^2} d^2 - \frac{12}{e^2 r} d + \frac{9}{e^2} & r \le d < 1.5r \\ 0 & d \ge 1.5r \end{cases}$$

A biblioteca utilizada para realizar a transformação dos dados em um *grid* possui ligações em Python através do pacote de código aberto *ibmolgrid*. Este *grid* é um array multidimensional, este array provê uma distribuição contínua da entrada.





Atom type é representado como uma distribuição de densidade ao redor do centro. Cada Átomo é uma função A(d, r) onde ‘d’ é a distância do até centro do átomo e ‘r’ é o raio de van der Waals:

```
1 6.05  4kqp/4kqp_rec_0.gninatypes 4kqp/4kqp_min_0.gninatypes
1 6.05  4kqp/4kqp_rec_0.gninatypes 4kqp/4kqp_docked_0.gninatypes
0 1.14  3jya/3jya_protein.pdb 3jya/3jya_ligand.sdf
0 1.23  1bzc/1bzc_protein.pdb 1bzc/1bzc_ligand.sdf
1 2.11  1nc3/1nc3_protein.pdb 1nc3/1nc3_ligand.sdf
0 -1.72 3kgp/3kgp_protein.pdb 3kgp/3kgp_ligand.sdf
1 1.9   1z9g/1z9g_protein.pdb 1z9g/1z9g_ligand.sdf
```

Reflexões Sobre a Aplicação

- A técnica em questão utiliza o grid de densidade gerados a partir da estrutura. Isto é uma maneira abrangente de representar a estrutura tridimensional de uma interação proteína e ligante.

- Este tipo de modelo consegue aprender as principais caracteríticas, relativa ao atracamento, da interação entre proteína e ligante.

- Na tarefa de Virtual Screening, o modelo é treinado e otimizado 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.

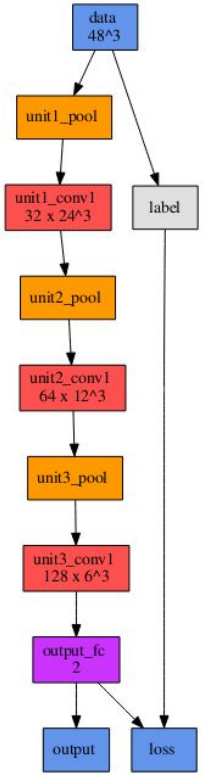


Figure 7: The network architecture of our final model.

Protein–Ligand Scoring with Convolutional Neural Networks - 10.1021/acs.jcim.6b00740

Discussão dos resultados

Análise isolada, apenas com os dados do DUD-E, resultando em CNN *scoring* supera o Vina com AUC de 0.85 contra 0.68 porém isto é dependente deste *dataset*, o modelo **não é tão generalista**.

Análise combinada dos conjuntos de dados(CSAR) utilizados no treinamento do modelo referente a tarefa de **predição de pose**. Estes testes evidenciam que há diferença ao utilizar um ou outro modelo para classificar as amostras.

Proporção 2:1 de dados presentes no conjunto DUD-E e CSAR. AUC de 0.83. Problema com *outlier (amostra vai muito bem em um modelo e muito mal no outro)*.

Discussão dos resultados

Com os modelos treinados então são utilizados ambos DUD-E quanto a combinação dele com o CSAR para analisar os resultados. O modelo é então avaliado nos conjuntos de testes ChEMBL e MUV. Estes *datasets* são mais desafiadores, para o

ChEMBL

- Vina: 0.67
- 2:1 DUD-E/CSAR CNN: 0.64
- DUD-E CNN: 0.78

Average cross-validation

- Vina: 0.71
- 2:1 DUD-E/CSAR CNN: 0.80
- DUD-E CNN: 0.86.

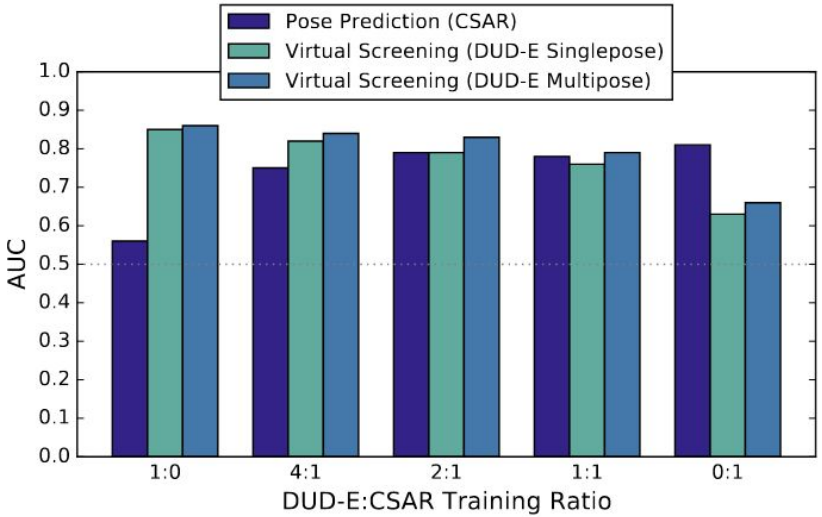
MUV:

- Vina: 0.55
- 2:1 DUD-E/CSAR CNN 0.50
- DUD-E CNN 0.52.

O modelo traz a reflexão sobre a influência da construção e escolha do conjunto de dados, neste caso específico por conta da diferença na avaliação dos modelos, pode se concluir que a construção do dado pode influenciar, então diferentes abordagens para a construção dos *decoys* podem trazer diferentes resultados.

Discussão dos resultados

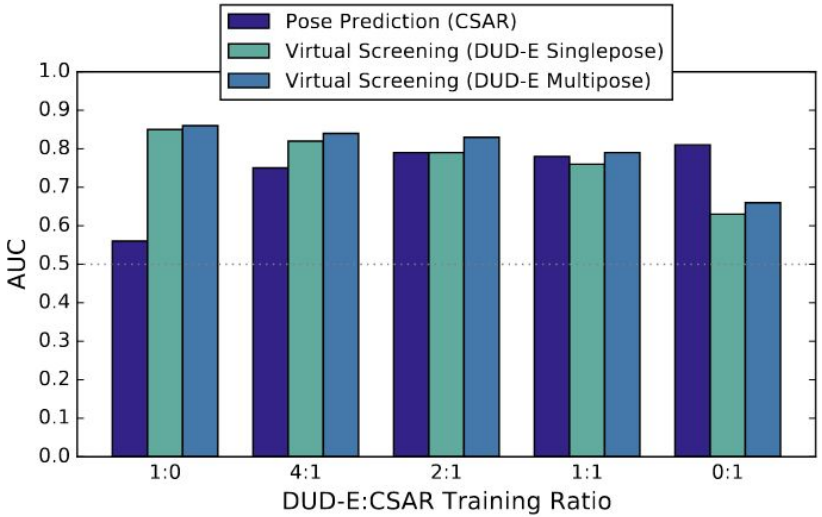
“Ideally the CNN models learn a generalizable model of protein-ligand binding from the training data. However, our models’ ability to generalize beyond the task inherent in the training data, while present, is limited (e.g. Figure 13).”



Discussão dos resultados

“That is, training to classify poses and active/inactive compounds does not generalize to the regression problem of binding affinity prediction.”

O modelo traz a reflexão sobre a influência da construção e escolha do conjunto de dados, neste caso específico por conta da diferença na avaliação dos modelos.

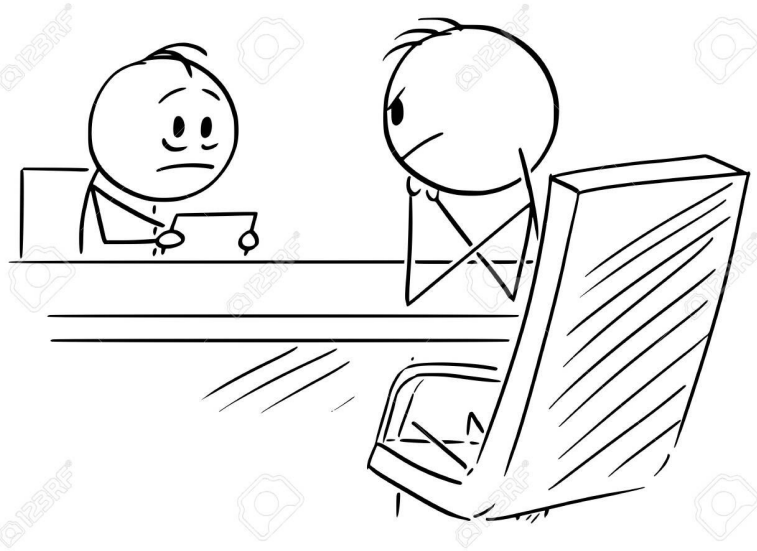


Conclusão

A ideia inicial do trabalho seria realizar a reprodução do artigo selecionado, não foi possível por conta da complexidade da tarefa.

- Complexidade
- Multidisciplinaridade
- Ferramentas

Aumentou o entendimento individual da área, melhor compreensão dos tipos de dados, prática e etc.



Referências

Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J., & Koes, D. R. (2017). Protein–Ligand Scoring with Convolutional Neural Networks. *Journal of Chemical Information and Modeling*, 57(4), 942-957. doi: 10.1021/acs.jcim.6b00740

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. *Nature*, 521, 436-444. doi:10.1038/nature14539

Mysinger, M. M., Carchia, M., Irwin, J. J., & Shoichet, B. K. (2012). Directory of Useful Decoys, Enhanced (DUD•E): Better Ligands and Decoys for Better Benchmarking. *Journal of Medicinal Chemistry*. doi: 10.1021/jm300687e

Sunseri, J., & Koes, D. R. (2019). libmolgrid: GPU Accelerated Molecular Gridding for Deep Learning Applications. *arXiv preprint arXiv:1912.04822*. URL: https://arxiv.org/abs/1912.04822