CSci 79502: Machine Learning

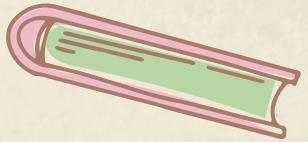
Spring 2023

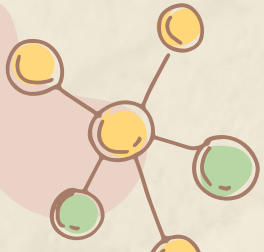
ML & DL for CPI & Binding Affinity

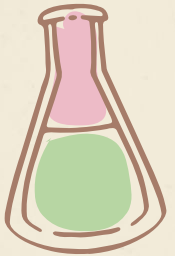
By: Roziena Badree & Mengriu Mao

Agenda

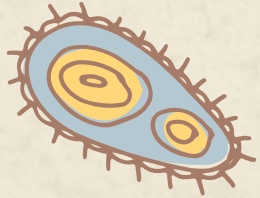
Introductory Items







What is the problem?



- The identification of compound-protein interactions (CPIs) is a crucial step in the process of drug discovery.

- The laboratory determination of CPIs is costly and time-consuming → as a result, computer science has become a promising and efficient alternative for predicting novel interactions between compounds and proteins on a large scale.

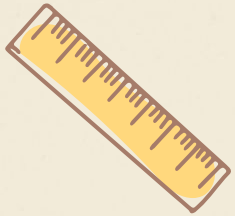
What are we aiming to do?

- Accurately predict if there is an interaction between the compound and protein (yes or no value called CPI)
- Accurately predict compound-protein binding affinity (a continuous value and the strength of the binding interaction)
- Compare the BACPI and ESM models
- Ultimately, guide model to focus on the effective sites of atoms and amino acids → increase the interpretability of the model

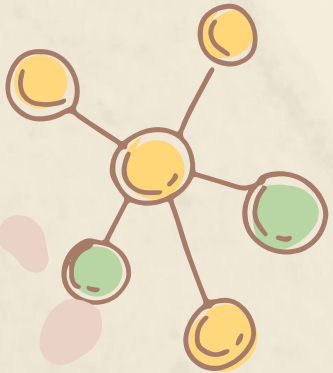
An Equal Breakdown of Work

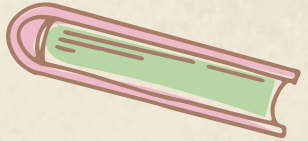
	Roziena	Mary
Codes	✓	✓
Poster	✓	✓
Presentation	✓	✓
Report	✓	✓

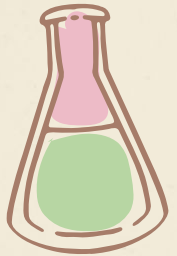
State-of-the-art & Related Work



- So far, we have only seen neural networks vs language models aka the Evolutionary Scale Model (ESM)
  - For neural networks, the best results so come from the BACPI model
    - Work done between School of Computer Science and Engineering in Hunan, Changsha, China and Old Dominion University in Norfolk, VA
  - The ESM model is a protein model trained on a masked language modeling objective → has the largest database of protein so far
    - Work is being done by Cold Spring Harbor Lab in Long Island, NY and Facebook's/Meta's researchers





Two Approaches

The Datasets

CPI Interaction	Binding Affinity
Human and C.*elegans* datasets containing positive and negative interactions	4 types of binding affinity so 4 datasets
Positive interactions: • Human = 3369 samples • C.*elegans* = 4000 samples Negative interactions: • Human = 384,916 samples • C.elegans = 88,261 samples	• IC50 = 489,280 samples • Ki = 144,525 samples • Kd = 12,589 samples • EC50 = 37,896 samples

$\pi = 3,141592$

What does each dataset look like?

- All datasets contain the simplified molecular-input line entry system (SMILES) data for the compounds
- Then, we have the amino acid sequences for the proteins
- Last, we have either the interaction data (0 for negative and 1 for positive) or the binding affinity (continuous)
- In one experiment, we feed this data into two different neural networks and in a second experiment, we feed this data into a language model to accurately generate the compound-protein representations

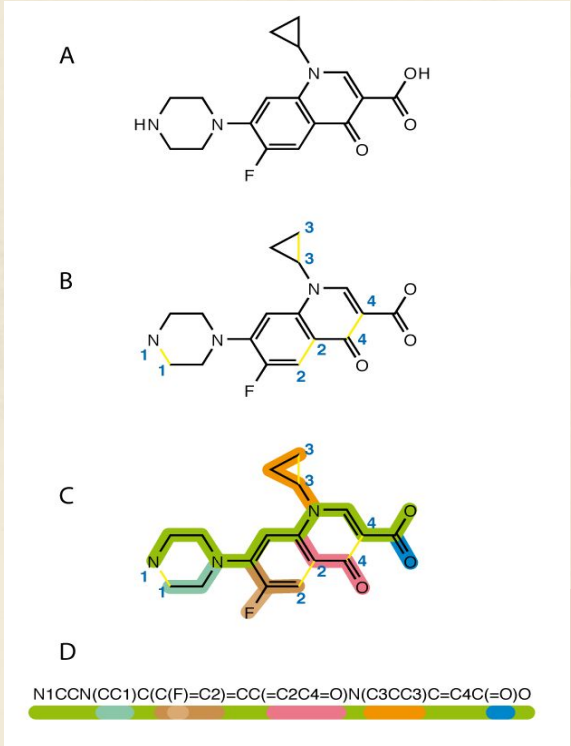
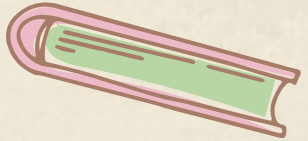
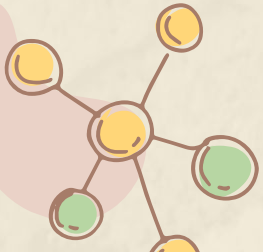
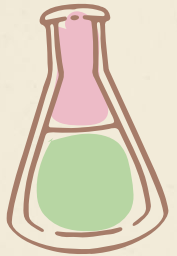


Figure: SMILES for ciprofloxacin

Approach # 1: BACPI Replication

Graph Attention Network (GAT)

- GAT is used to process the compounds into an atom structure graph
  - Used RDKit to convert the SMILES format to graph representation (G = {V,E})
    - V is the set of vertices → *v* represents the *i*th atom
    - E is the set of edges → *e* is the chemical bond between the *i*th and *j*th atoms
  - Fed G and the randomly initialized embeddings of vertices into the GAT
    - Embeddings of both source & target nodes were considered to allow the weight to depend on more than just the number of neighbors → can capture anything like structure (attention function that allows a node to tend to some neighbors more than others) → attention scores calculated by using LeakyReLU activation function → weighted matrix
  - Took the message of the neighbors, which is their raw features multiplied by this matrix & scaled it using the normalized attention mechanism → summed all scaled messages → passed through a final non-linear activation function

Fingerprinting the Compound

Used Extended Connectivity Fingerprints (ECFPs) as the second representation of the compound

- ECFPs are a class of topological fingerprints for molecular substructure characterization → describe the characteristics of substructures consisting of each atom and circular neighborhoods within a diameter range
  - We used RDKit (i.e. from rdkit import ...) to calculate the fingerprint of compounds and obtain a feature vector (the atom, adjacency, and radius analysis)

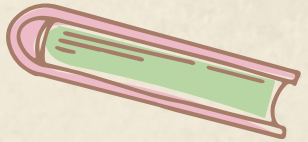
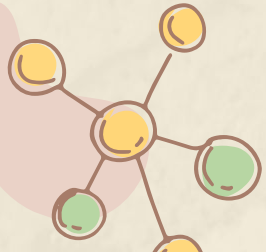
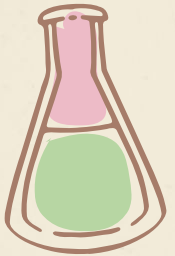
Convolutional Neural Network (CNN)

- CNN is used to process the proteins (extract local features and learn vector representations)
  - Hidden layers aka convolutional layers detected the patterns
  - Used a context window *w* to split the protein sequences into overlapping subsequences of amino acids (AAs) to improve prediction performance → set *w* = 3 so that AAs can be split into diverse subsequences (i.e. MRPSG → MRP, RPS, PSG) of set length of 3 → regarded as AA residues
  - Translated all residues into randomly initialized embeddings → Updated them through several convolutional layers with a non-linear activation function (ReLU) → Obtained final output vector for all residues along the protein sequence

BACPI

- Produced attentions in both directions (atom to AA-reside and AA-residue to atom)
  - First, transformed atom features, fingerprint features, and residue features into a single layer-NN (LeakyReLu)
    - Took C (content matrix of compound), U (trainable parameter matrix), and P (content matrix of protein) and aligned them (use tanh and transpose the P matrix)
      - Result was a matrix showing the interaction strength between each atom and residue and vice versa → Calculated normalized attentions in both directions using a softmax function (containing concatenation ops) → Transformed result into a single-layer NN and then predicted final binding affinity by concatenating compound, fingerprint, and protein features, applying LeakyRelu, and flattening
        - Obtained both interaction and binding affinity

Approach # 2: ESM

Evolutionary Scale Modeling (ESM)

- Once again, used ECFPs or fingerprinting to represent the compound
  - Features were atom, adjacency, and radius analysis
- For the protein, implemented transformer pre-trained model (used ESM-2) for the training data → AAs arranged in a many combinations to form structures that carry function, the same way letters form words and sentences carry meaning
  - Obtained an atom-to-residue contact map of the compound and protein → Passed this through classification and regression models to predict interaction and affinity on the test data

How does ESM2 model from Meta work?

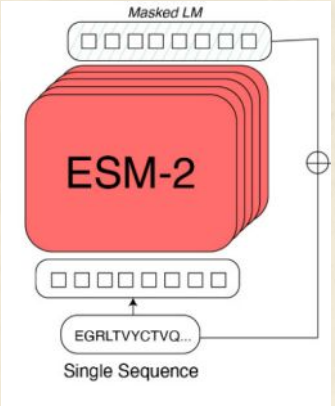
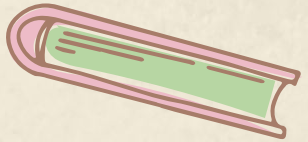
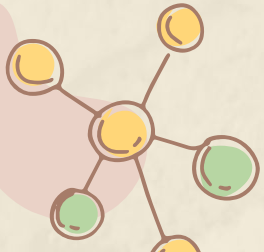
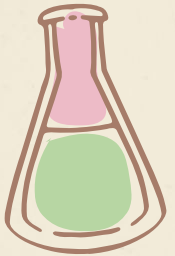


Figure: ESM2 model

- For the input, randomly dropped out amino acids → fed that into the transformer that learned how to predict the missing amino acids & gained insight into protein structure

Models & Evaluation

Models Used

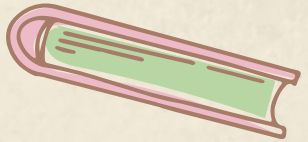
- ESM
  - Supervised Models
    - Regression: Ridge Linear Regression, Lasso Linear Regression, SVM (Gaussian RBF), RF, & Multi-layer Perceptron
    - Classification: Linear Regression, Elastic Net Regression, Random Forest, Extra Trees, SVM, Gradient Boosted, GaussianNB, Multinomial NB, Logistic Regression, Perceptron, Multi-layer Perceptron, & KNN
  - Unsupervised Models for Classification: Affinity Propagation, KMeans, Outlier, & Spectral Clustering
  - Stochastic Gradient Descent

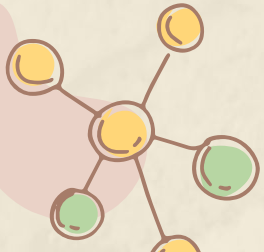
- Deep Learning Comparison with BACPI

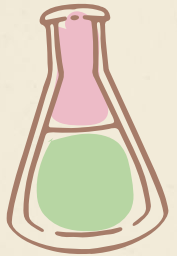
Evaluations

- For CPI (looking for 0 or 1 values (i.e. classification))
  - Used the RMSE values
    - Would also like to look at the confusion matrix, but the values returned are continuous.
      - What is the filter? For example: are all values greater than 0 indicative of an interaction (so mark as 1), even if the number is very small?
- For binding affinity (regression)
  - Used the RMSE values

Discussion of Results







CPI Predictions

C. elegans	
Model Name	**RMSE**
Extra Trees	0.189474
Random Forest	0.211261
Gradient Boosted	0.273941
SVM	0.297722
Logistic Regression	0.322886
KNN	0.360259953
ElasticNet Regression	0.372044
Multi-Layer Perceptron	0.407379
GaussianNB	0.409982
Local Outlier Factor	0.50318407
MultinomialNB	0.518796
Perceptron	0.705601
Kmeans	0.777395731
Linear Regression	1778409083
Spectral Clustering	-
Affinity Propagation	-

Human	
Model Name	**RMSE**
Extra Trees	0.228321
Random Forest	0.242715
SVM	0.291737
Gradient Boosted	0.307586
Logistic Regression	0.339053
Linear Regression	0.363745
ElasticNet Regression	0.372946
KNN	0.385137992
Perceptron	0.408501
Multi-Layer Perceptron	0.408501
MultinomialNB	0.455701
Local Outlier Factor	0.496150446
GaussianNB	0.777464
Kmeans	0.834972822
Spectral Clustering	-
Affinity Propagation	-

- BACPI results are still pending because the codes are still running

Binding Affinity Predictions

Model Name	IC50	EC50	Ki	Kd
BACPI	0.74	0.78	0.8	1.08
Random Forest	0.792437	-	0.989753	1.184066
MLP Regressor	0.917291	-	1.083637	1.259279
Ridge Linear Regression	0.949984	-	3.034943	2.285426
Lasso Linear Regression	1.013284	-	1.422459	1.367706
Support Vector Machine (Gaussian RBF)	1.186531	-	1.198874	1.421241

Note: The results for the EC50 dataset are missing because the codes are still running.

Discussion of Results

- The ESM Extra Trees model was the best for predicting CPI interaction so far (still waiting on full BACPI results)
- The BACPI model outperforms the language processing models for binding affinity
    - We did not include molecular adjacency in the fingerprinting analysis of our compounds so a second run of the codes with molecular adjacency would most likely change our results (an initial run was done and codes were removed due to time constraints → RMSE values were slightly lower)

Conclusion

Challenges, Lessons Learned & Future Work

- Challenges
  - Each code takes an extremely long time to run (between 2 to 9 hours) → makes modifications very difficult
  - A lot of research is needed to understand the computations used to generate each model and our time is limited
- Lessons Learned
  - Start the written work sooner → you don't realize the intricacy of the material until you need to explain it in writing
- Future Work
  - Run the affinity codes on the EC50 dataset
  - Test ESM model with adjacency
  - Test other ESM models
  - Try ESM with a NN
  - Try AlphaFold (more accurate than ESM but also much slower)

Thank you for listening.

Please let us know if you have any questions.

References

[1] Ferruz, Noelia & Hocker, Birte. (2022). Towards Controllable Protein design with Conditional Transformers.

[2] Karimi, M., Wu, D., Wang, Z., & Shen, Y. (2019). DeepAffinity: interpretable deep learning of compound-protein affinity through unified recurrent and convolutional neural networks. Bioinformatics (Oxford, England), 35(18), 3329–3338. https://doi.org/10.1093/bioinformatics/btz111.

[3] Min Li, Zhangli Lu, Yifan Wu, and YaoHang Li. BACPI: a bi-directional attention neural network for compound–protein interaction and binding affinity prediction. Bioinformatics, 38(7), March 2022, pp. 1995–2002, https://doi.org/10.1093/bioinformatics/btac035.

[4] Lin, Zeming, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, et al. “Evolutionary-Scale Prediction of Atomic Level Protein Structure with a Language Model.” bioRxiv, 2022. https://doi.org/10.1101/2022.07.20.500902.

[5] Toutain, P. L.; Bousquet-Melou, A. (2002-12-14). Free Drug Fraction vs. Free Drug Concentration: A Matter of Frequent Confusion. Journal of Veterinary Pharmacology and Therapeutics. Wiley inc. 25 (6): 460–463. https://doi.org/10.1046/j.1365-2885.2002.00442.x.

[6] Whitford, David. 2013. Proteins: Structure and Function. J. Wiley & Sons.