Exercise 4 - Hackathon Prep

3D Data Processing in Structural Biology

Shira Gelbstein, Daniel Levin, Imri Shuval and Ron Levin

הצהרה: אנו, כל חברי הקבוצה החתומים מטה, עברנו על התרגיל והפתרון המוגש במלואם.
ש.ג. , ד.ל. , א.ש. , ר.ל.

1) תוצאות שינויי השכיבה ממנה מוציאים את הembedding:

Layer 5 AUC = 0.83

Layer 9 AUC = 0.89

Layer 15 AUC = 0.81

Layer 20 AUC = 0.74

Layer 36 AUC = 0.81

ניתן לראות שדווקא בשכבות הנמוכות יותר, הפיצ'רים יותר אינדקטיביים עבור זיהוי NES.

2) לא תמיד, מכיוון שאם נגדיל את מספר השכבות/מימד הפיצ'רים אבל לא נגדיל את כמות הדאטה הרשת שלנו עלולה לעשות overfit לדאטה האימון.

3) המדד מחשב את המרחק בין כל נקודה (embedding) לסנטרואיד של הנקודות המייצגות פטטידים חיוביים, ולסנטרואיד של השליליים.

בכך הנוסחה "מענישה" על מרחק מהסנטרואיד של החיוביים, ו"מתגמלת" על מרחק של הנקודה מהסנטרואיד של השליליים. בכך היא נותנת לנו מדד טוב שמאפשר סיווג של הפטטידים.

4) התוצאה הכי טובה הגיעה ל-AUC של 0.97 עם:

Embedding size: 2560

Batch size: 32

Epochs: 50

Learning rate: 1e-3

Hidden dimension: 512

Dropout: 0.4

Embedding layer: 9

5)

a. ניתן להשתמש במידע מתוך המבנים שמהם אנחנו מפיקים את הembedding, כמו התאמה של "מנעול ומפתח" בין המוטיב הקצר וחלבון עליו הוא חובר, או ניתן לחשב את עוצמת האינטראקציה בין שתי הchains על ידי פונקציית אנרגיה ייעודית של כלים אחרים, לדוגמה ddG filter של Rosetta.

b. אופציה 1: ניתן להשתמש בCNN כדי לעבד את המבנים התלת מימדיים ולהשתמש בזה, בנוסף ל-embeddings של ה-ESM, כקלט ל-classifier.

אופציה 2: ניתן להמיר מידע מתוך הרצפים של כ-20 עמדות באתר שבו אנחנו חושדים שקיים NES Motiv ולאחר מכן להשתמש ב Random Forest כדי לבצע את הקלסיפיקציה. מאחר ועל פי מידע מקדים במוטיבים ידועים ישנן עמדות מסויימות שבהם קיימים חומצות אמינו עם מספר מוגבל של זהויות על מנת שיתקבל NES motive.

6) אמנם לא ניתן להפריד באופן מושלם, אך ניתן לראות שאם נמתח קו אופקי סביב ה-0 כן נקבל הפרדה סבירה. אלגוריתם k-means נכשל לגמרי במציאת קלאסטרים משמעותיים.

7)

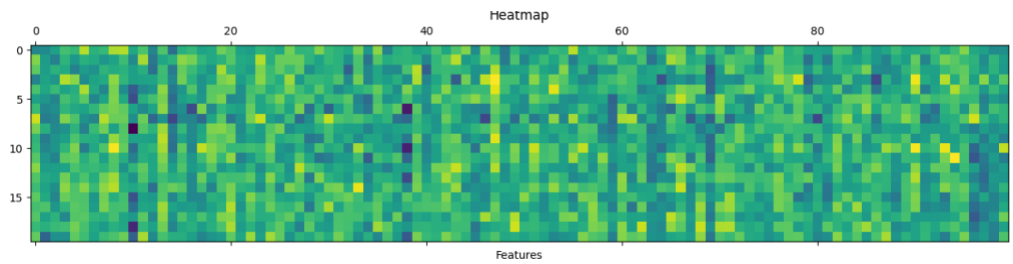
$$AUC\ (COMs) = 0.47$$

$$AUC(pLDDT) = 0.76$$

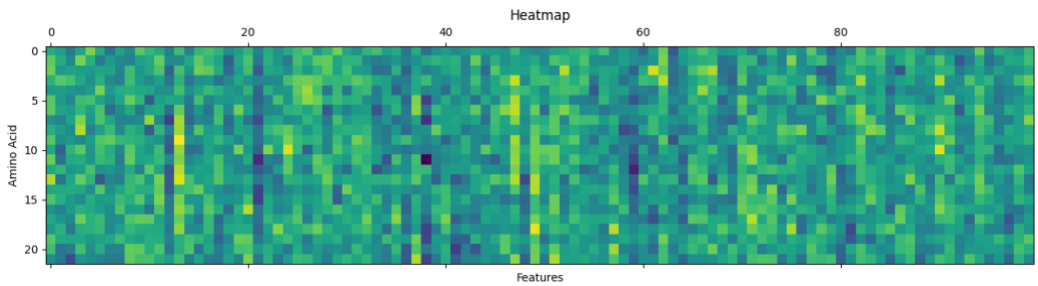
מכך אנחנו מסיקים שמרכזי המסה (COM) של החלבונים אינם אינדיקטיביים לגבי האם הפפטיד חיובי או שלילי, לעומת הplddt שמתקבל מהפרדיקציות של AlphaFold שנותנות לנו אינדיקציה יחסית חזקה.

Plots – ex4.py:

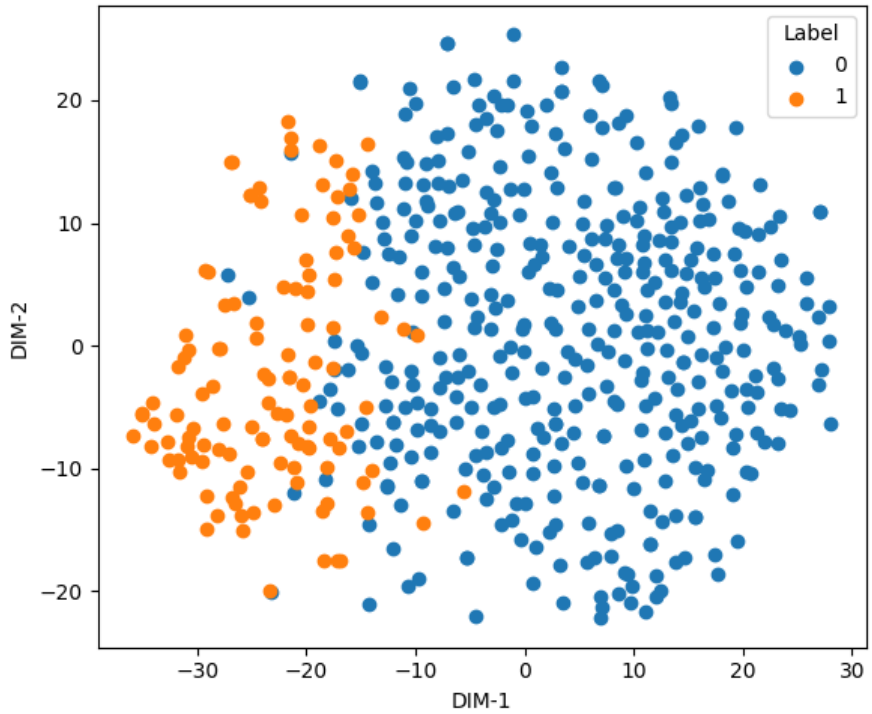
Embedding Heatmaps – Positive:



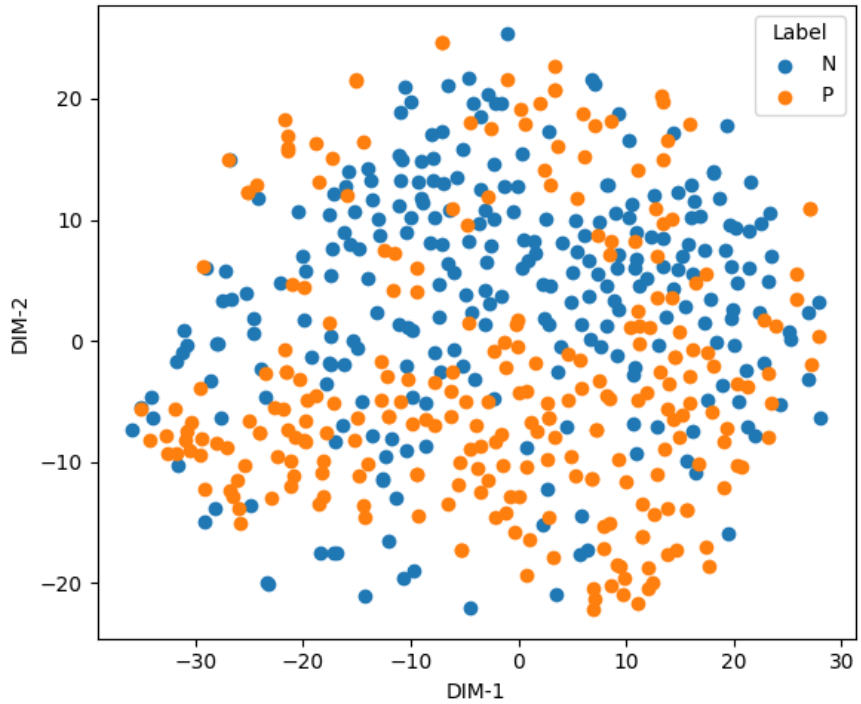
Negative:



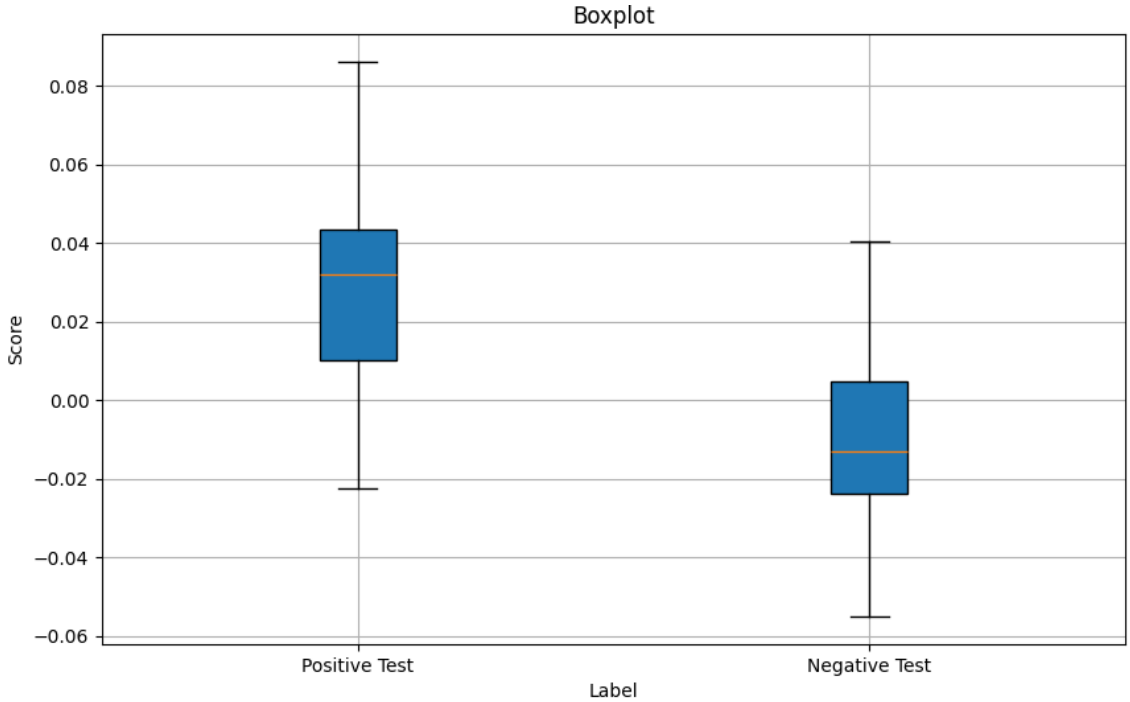
2D K-means:

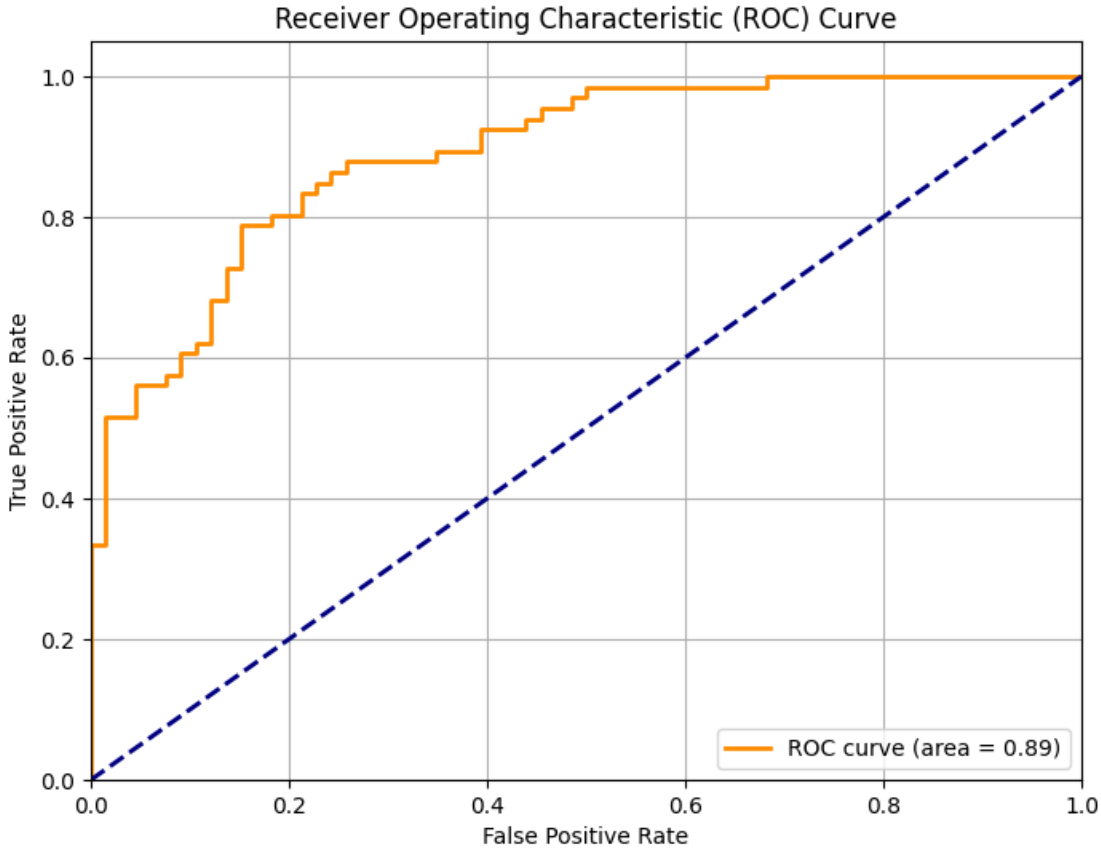


2D labels:

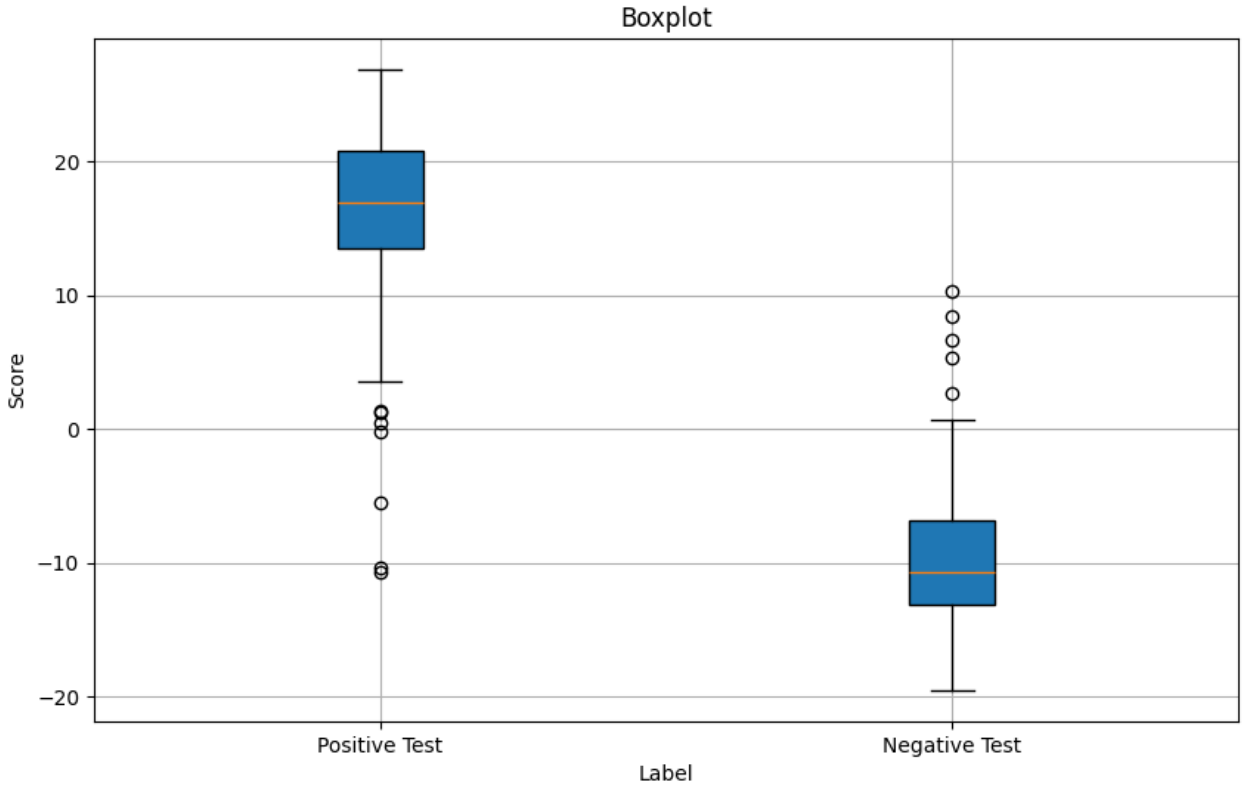


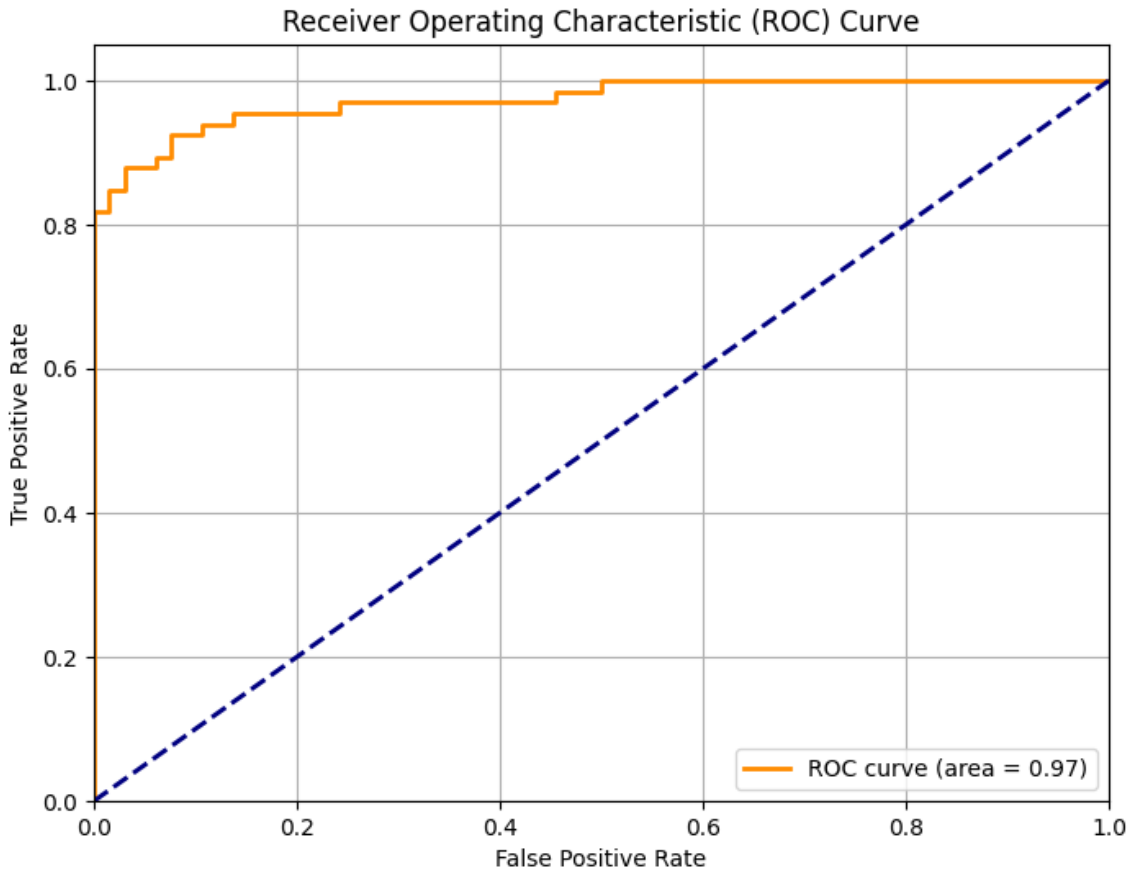
Baseline boxplot & ROC curve:





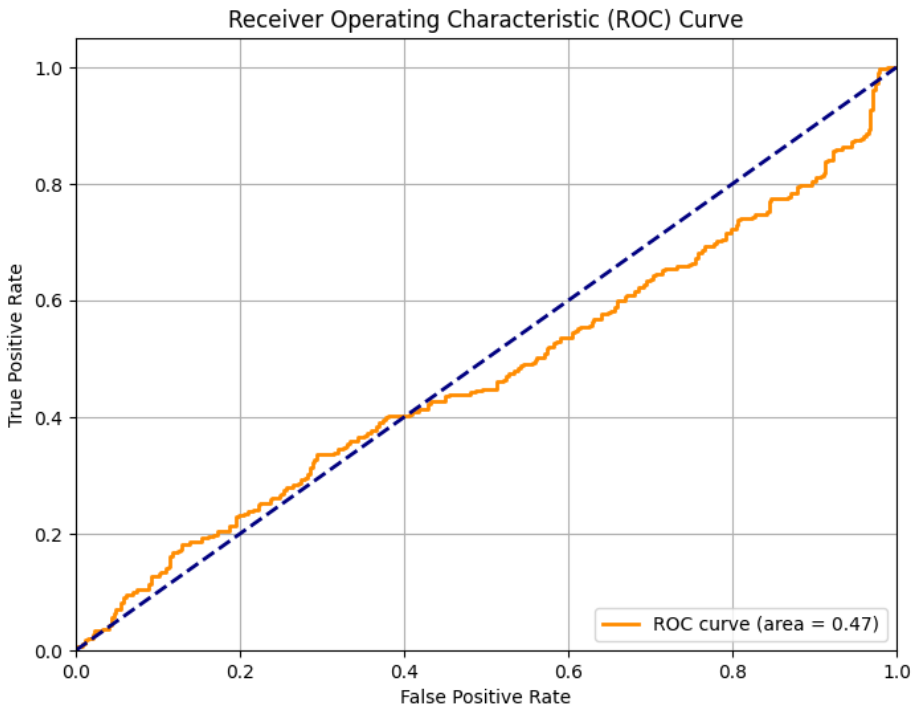
NN boxplot & ROC curve:





Plots – structure_analysis.py:

Center of Mass (COM) ROC Curve:



pLDDT ROC curve: