**תרגיל 4: עיבוד תלת-מימדי בביולוגיה מבנית**

1. שינויchosen\_embedding\_layer משפיע על איכות הייצוג. השכבות המוקדמות של הרשת לוכדות מידע מקומי, בעוד שכבות מאוחרות לוכדות מידע פונקציונלי-גלובלי, ולכן שכבה 33 (המאוחרת ביותר) לרוב נותנת AUC גבוה יותר.
2. לא, מודל גדול יותר לא תמיד משפר ביצועים. לפעמים הוא מייצר embedding עשיר מדי או רועש לתפקיד פשוט יחסית כמו סיווג פפטידים בינארי.
3. הציון משקף אם הפפטיד "קרוב יותר" לחיוביים או לשליליים במרחב הembedding. אם הוא קרוב לחיוביים ורחוק משליליים אזי היחס יהיה חיובי, ולהפך. זו דרך פשוטה ונורמליזציה רכה להבדיל בין הקבוצות.
4. הAUC הטוב ביותר שהושג: 0.89, עם ההיפרפרמטרים:  
   embedding\_size=1280,  
    hidden\_dim=128,  
    dropout=0.2,  
   batch\_size=64,  
   epochs=50,  
   learning\_rate=1e-3
5. א. קלטים נוספים שיכולים לשפר ביצועים:  
   ממוצע pLDDT של המבנה (אם קיים).  
   מאפיינים פיזיקלים וכימיים של הפפטידים.  
   ב. ארכיטקטורות או אלגוריתמים חלופיים:   
   **לעשות**
6. הt-SNE הראה שהקלאסים נפרדים חלקית אך לא לגמרי. KMeansמצליח לזהות קלאסטר די ברור של פפטידים חיוביים, אך עם חפיפות – כלומר הקלאסטרים לא מושלמים אך כן משמעותיים.
7. לפי התוצאות שקיבלנו,pLDDT הוא מדד טוב יותר להפרדה בין הקבוצות:  
   עבור COM distance: AUC 0.47, לא מופרד בכלל.  
   עבור mean pLDDT: 0.76 AUC, הפרדה בינונית-טובה.  
   לכן, יציבות מבנית (pLDDT) מבדילה בין פפטידים יותר טוב מאשר מרחק ממרכז מסה (COM distance) של פפטיד ייחוס.