בס"ד

1. ככל שהשכבה יותר נמוכה היא תלכוד מאפיינים יותר מקומיים, בעוד ככל שנעלה בשכבות הן יתפסו מידע שכולל הקשר יותר רחב שיכול ללמד גם על הפונקציונליות והמבנה של הפפטידיט. לכן הAUC עולה ככל שהשכבות עולות (עד לשכבה מסויימת שיכולה לגרום כבר לאובר-פיטינג).
2. לא בהכרח. אם הדאטה בייס קטן יחסית, מודל גדול מידי יכול לגרום לאובר-פיטינג. ספציפית במקרה שלנו ראינו שembedding של 640 השיג את התוצאה הטובה ביותר.
3. בגלל שהציון הזה נותן ציון גבוה לפפטידיטים שדומים יותר לחיוביים וציון שלילי (נמוך) לכאלו שיותר דומים לשליליים, וככה אפשר להגדיר טרשולד שיסווג מי חיובי ומי שלילי. ככל שהניקוד גבוה יותר הפפטיד נראה יותר כמו פפטיד חיובי, וזה עוזר להבחין בין הקבוצות. הפונקצייה הלוגריתמית log1p מונעת ערכים קיצוניים מדי.
4. embedding_size-640
   layer-6
   test_size-0.3
   batch_size-32
   epochs-50
   lr-0.01
   hidden_dim-64
   dropout-0.2
   AUC: 0.9442
5. א. אפשר להוסיף פיצ'רים שמשקפים תכונות פיזיקליות-בילוגיות כמו הרכב חומצות האמינו (מה החלוקה בין הסוגים), אורך, משקל, מטען המולקולה. בנוסף אפשר במקום להשתמש רק בembedding הסופי, אפשר לייצר מפות אטנשיין, שיעזרו ללכוד תלויות רחבות.
   ב. אפשר להשתמש בקונבולוציה (כמו שעשינו בתרגיל הקודם), לפעמים random forest נותן תוצאות טובות עבור דאטה קטן וכן אפשר להשתמש ברשתות עם attention ויש מצב שזה יצליח טוב יותר.
6. אפשר להפריד בצורה גסה אבל כן יש יחסית הרבה חריגות. אפשר לראות שהK-mean לא מצליח לסווג טוב- הוא מחלק לשתי קבוצות לאורך, בעוד שחלוקה יותר הגיונית הייתה מרכז כחול ופריפריה כתום.
7. ממוצע הPLDDT הוא מבחין טוב יותר מCOM. קיבלנו שהAUC של הCOM הוא 0.47, שזה אפילו פחות טוב מניחוש רנדומלי. לעומת זאת, קיבלנו שהACU של הPLDDT הוא 0.76, כלומר הרבה יותר מאפשר לסווג. גם בגרפים של הboxplot אפשר לראות שבPLDDT החציון של הפפטידיטים החיוביים גדול יותר משל השליליים, לעומת הCOM שבו הציונים די זהים ויש הרבה יותר חריגות.