



## סיווג מנגנוני פעולה של תרופות כאטיולוגיים או כפליאטיביים בעזרת רשתות מורכבות ומודלי שפה גדולים

ליעד ברטל בית הספר למנהל עסקים אוניברסיטת בר-אילן	עדן ברמן בית הספר למנהל עסקים אוניברסיטת בר-אילן	מאור ימיני בית הספר למנהל עסקים אוניברסיטת בר-אילן	קטלין ז'גודניק בית הספר למנהל עסקים אוניברסיטת בר-אילן	אלון ברטל בית הספר למנהל עסקים אוניברסיטת בר-אילן
--	---	---	---	--

### תקציר

חברות פארמה משקיעות משאבים רבים להבנת מנגנוני הפעולה של תרופות בניסיון לפתח אסטרטגיות טיפוליות המקדמות רפואה מדויקת ואפקטיבית. עם זאת, מיעוט מחקרים ניתחו באופן שיטתי את היכולת לזהות תרופות המטפלות בגורם המחלה (מנגנונים אטיולוגיים) לבין תרופות המקלות על תסמיני המחלה (מנגנונים פליאטיביים). מאמר זה מציג מודל חדש מבוסס למידת מכונה לצורך סיווג מנגנוני הפעולה של תרופות (Mechanism of Action, MoA) כאטיולוגיות או כפליאטיביות, תוך השוואה בין שלושה מודלים של למידה עמוקה: (i) מודל שפה מאומן (SciBERT) Science Bidirectional Encoder Representations from Transformers, (ii) מודל רשת בסיסי (GNN) Baseline Graph Neural Network, ו-(iii) מודל משולב DruGNNosis-MoA הממזג GNN עם הטמעות SciBERT כתכונות קודקודים. המודל המשולב DruGNNosis-MoA הצטיין בסיווג מנגנוני פעולה של תרופות (F1-score 0.942). DruGNNosis-MoA תורם לאפיון מנגנוני תרופות למחקרים עתידיים, ובכך מקדם רפואה מותאמת ופיתוח טיפולי.

**מילות מפתח:** אטיאולוגי, גרפים לומדים, למידת מכונה, מנגנוני פעולה של תרופות, מודלי שפה, פליאטיבי.

### מבוא וסקירת ספרות

רפואה מודרנית כוללת שימוש בתרופות להקלה בסימפטומים, ומיקסום הטיפול בשיטות שאינן פולשניות (Ravikumar, et al., 2018; Yu, et al., 2020). אולם, תהליך פיתוח תרופות חדשות הינו ממושך, עתיר משאבים ונוטה לכשלים תכופים (Xue, et al., 2018; Morgan, et al., 2011). אכן, יוזמות רבות לפיתוח תרופות לא מצליחות להשיג אישור לשימוש ציבורי, ואף בעיות נוספות מתעוררות בתדירות גבוהה בשלבים מאוחרים יותר של תהליך הפקת התרופות (Ciociola, et al., 2014). אתגרים כגון מניעת תופעות לוואי, איזון יעילות מול רעילות, שיפור תרופות קיימות, והתייחסות לשיקולים משפטיים וכלכליים תורמים למורכבות תהליך פיתוח ואישור התרופה (Ravikumar, et al., 2018).

תרופות יכולות לטפל באופן ישיר בגורם המחלה, או באופן עקיף להתמקד בהקלת תסמיני המחלה

(Yildirim, et al., 2007; Lindpaintner, 2002). תהליך פיתוח תרופות ניתן לשיפור על ידי הרחבת ההבנה של שני מנגנוני פעולה: (i) אטיולוגי – התמקדות בגורמי השורש או בגורמים התורמים למחלה, וכן (ii) פליאטיבי – הקלה של תסמינים באמצעות פעולות על חלבונים שאינם מעורבים ישירות עם הגורם העיקרי, אלא מעורבים במניעת תסמיני המחלה (Yildirim, et al., 2007; Lindpaintner, 2002). הבחנה זו בין מנגנוני פעולה של תרופות יכולה להנחות פיתוח אסטרטגיות טיפול מדויקות, מזעור תופעות לוואי, אופטימיזציה של הקצאת משאבים, וסיוע בהחלטות רפואיות המותאמות אישית למטופלים (Vogt, et al., 2014).

ההבחנה בין מנגנוני פעולה אטיולוגיים לפליאטיביים כרוכה באתגרים רבים הנובעים מהטבע המורכב של תהליכי מחלה, החפיפה הפוטנציאלית בין שורש המחלה לבין הקלה על הסימפטומים, המנגנון הכפול שמציגים תרופות

נרחב ובשיטות אנליטיות מתקדמות כגון למידת מכונה. מחקרים אחרים מתמקדים לעיתים קרובות בניבוי אינטראקציות בין תרופה לחלבון מטרה, ייעוד מחדש של תרופות (He, et al., 2020), או תיאור מנגנוני תרופות ייחודיות ללא סיווג שיטתי (Vogt, et al., 2014).

חידושים טכנולוגיים יכולים לסייע בתחום גילוי ופיתוח תרופות. בנייהם אינטראקציות חלבונים מאומתים באופן ניסויי ורחב היקף, זיהוי גנים הקשורים למחלות, ומודלים נרחבים של למידת מכונה. למשל, Graph Neural Network (GNN) הינו מודל למידת מכונה המכיל את המאפיינים המובנים של רשת על ידי העברת מידע בין הצמתים (Zhou, et al., 2020). מודלי GNN יעילים באיתור תבניות מורכבות וקשרים תלויים בנתונים יחסיים. בנוסף, מודלים אלו בעלי יעילות גבוהה במשימות שונות של ניתוח רשתות, כולל סיווג צמתים, חיזוי קשרים ואשכול גרפים (Zhou, et al., 2020). התבססות על מבנה הרשת תוך שימוש במודל GNN הינו בעל פוטנציאל משמעותי לחשיפת דפוסים סמויים וקשרים מהותיים בנושא מנגנוני פעולה של תרופות, המאפשר את סיווגם לאטיולוגיים או לפליאטיביים.

במטרה להבדיל בין מנגנוני תרופות אטיולוגיים לעומת פליאטיביים, נעזרנו במספר מערכי נתונים ביו-רפואיים נרחבים, וביצענו חקירה מקיפה של 2,018 תרופות מאושרות FDA של ארה"ב. פיתחנו שלושה מודלים לסיווג: (i) מודל שפה מאומן מראש (Science-BERT (SciBERT), Beltagy, et al., 2019) לתיאורי תרופות מ-DrugBank (Wishart, et al., 2018); (ii) אימון מודל GNN מבוסס Feature Embedding של Eigenvector (Huang, et al., 2020); ו- (iii) אימון מודל GNN מתוגבר במאפיינים טקסטואליים הנגזרים ממודל SciBERT מאומן. ליתר דיוק, תוך שימוש במודל SciBERT יצרנו מאפייני צמתים (ייצוג ווקטורי של מספרים, Embeddings) של תיאורים טקסטואליים של תרופות.

מחקרנו מחדש ומבהיר את ההבחנה בין מנגנוני תרופות אטיולוגיות ופליאטיביות. זאת תוך

מסוימות, מספר מוגבל של טיפולים למחלה מסוימת, השונות והייחודיות של מטופלים, הנדירות והווריאציה של מחלות, והאופי של שירותי הבריאות (Yu, et al., 2020). מספר המחקרים אשר ביצעו סיווג מקיף של תרופות על בסיס מנגנוני פעולה אלו הינו מזער, וזה המניע למחקר שלנו.

גילוי תרופות חדשות יכול לנבוע מייעוד חדש של תרופות קיימות (Parvathaneni, et al., 2019). דבר זה כרוך בזיהוי מאפייני מנגנוניים משותפים בין תרופות שאושרו, לבין כימיקלים חדשים (Anatomical Therapeutic Chemical (ATC) classes) (Xue, et al., 2018). במערכת ה-ATC, חומרים פעילים מקובצים על סמך איברי מטרה או מערכות מסוימות, כמו כן גם לפי מאפיינים טיפוליים, תרופתיים וכימיים. ניצול קטגוריות ה-ATC מפשט זיהוי וחקירה של טיפולים תרופתיים פוטנציאליים עבור מחלות חדשות או מחלות שטרם נחקרו לעומק (Olson, et al., 2017).

המושגים אטיולוגי ופליאטיבי ביחס למנגנוני פעולה של תרופות זכו לתשומת לב מועטה במחקר (Yildirim, et al., 2007; Lindpaintner, 2002). עם זאת, חוקרים סיפקו תובנות בסיסיות על מושגים אלו כחלק מדיונים רחבים יותר. Lindpaintner ושותפיו סיפק דוגמאות המבהירות את ההבחנה בין מנגנונים אלו, בעוד Yildirim ושותפיו הרכיבו רשתות המעריכות את הקירבה בין תרופות למחלות לפי אינטראקציות חלבונים, ואף חשפו העשרה של תרופות פליאטיביות (Yildirim, et al., 2007; Lindpaintner, 2002). בנוסף, החוקרים הציעו להשתמש במרחק הקצר ביותר בין חלבון המטרה של תרופה לבין חלבון תוצר של גן פגום כאינדיקטור לסיווג מנגנוני תרופות על סמך השלבים המולקולריים ביניהם. לטענתם, מרחקי רשת קצרים בין חלבונים אלו מעידים על מנגנון אטיולוגי, בעוד שמרחקים ארוכים מציינים מנגנון פליאטיבי.

מחקרים רבים (Vogt, et al., 2014) בתחום מחקר ופיתוח תרופות התעלמו מההבחנה החיונית של מנגנוני פעולה אטיולוגיים ופליאטיביים. יחד עם זאת מחקרים אלו לא השתמשו במערך נתונים

מידע על הקשרים בין גנוטיפ לפנוטיפ של מחלה בבני אדם נאסף ממערך הנתונים Online Mendelian Inheritance in Man (OMIM) Morbid Map (Amberger, et al., 2015), אשר נאסף בראשון למאי, 2023. מערך הנתונים מכיל תקציר מפת גנים אנושיים של OMIM, הכולל 8,156 מחלות ו-16,358 גנים הנמצאים בקישוריות למחלה. מחלות מיוצגות על ידי שם הפנוטיפ שלהן, בעוד גנים מיוצגים על ידי סמלי גנים, כלומר שמותיהם.

## שיטות

### תיג מנגנוני פעולה של תרופות

בניתוח מידע על רישומי התרופות ב-DrugBank, התמקדנו בשדות 'Identification' ו-'Pharmacology' כדי לסווג ידנית כל תרופה כבעלת מנגנון אטיולוגי, פליאטיבי או שניהם. כל תרופה תויגה על ידי שני סטודנטים בעלי תואר ראשון בביולוגיה חישובית. במקרה של תוויות סותרות, הסיווג נקבע ע"י המחברת השנייה. הערכה ידנית זו הניבה מערך נתונים מתויג במנגנון הפעולה של כל תרופה. למשל, עבור מחסור בוויטמין מסוים, התערבויות כגון טיפול חלופי בנטילת קפסולות ויטמינים או הפרשה מושרית של הוויטמין החסר סווגו כאטיולוגיות, בעוד שתרופות להקלה על תסמיני המחסור סווגו כפליאטיביות.

### אימון SciBERT

אימנו מודל שפה SciBERT (Beltagy, et al., 2019) לסיווג תרופות כאטיולוגיות או כפליאטיביות באמצעות התיוגים שהכנו ותיאורי תרופות מ-DrugBank. נתוני ההכשרה כללו טקסט מקטעי ה-'Pharmacodynamics' וה-'Mechanism of Action' ב-DrugBank, תוך השמטה ידנית של אזכורים מפורשים לסיווג התרופה. זאת בכדי להבטיח שהמודל ילמד מההקשר הכללי ולא מרמזים מפורשים. הדאטה חולק ל-90% סט אימון ו-10% סט מבחן, כאשר 10% מסט האימון נוצל לסט אימות. ביצענו 10-fold Cross-Validation (CV) כדי לאמן את מודל השפה SciBERT, והערכנו את ביצועיו בין היתר

שימוש בטכניקות למידת מכונה מתקדמות לניתוח מערכי נתונים ביו-רפואיים נרחבים ומגוונים. החקירה השיטתית והמקיפה של מנגנונים אלו לא רק מבטיחה שינוי פרדיגמה בהבנת פעולות התרופות, אלא גם מעידה על פוטנציאל לחדש את תהליך פיתוח התרופות בשלבים מוקדמים, לכוון בחירות טיפול מותאמות אישית, ולייעל אסטרטגיות ייעוד מחדש של תרופות קיימות.

## נתונים

### דאטה של תרופה-מטרה (Target)

באמצעות מידע על תרופות וגן היעד שלהם מ-DrugBank גרסה 5.1.10 (Wishart, et al., 2018), אספנו מידע על תרופות מאושרות FDA המכוונים כלפי גנים אנושיים, למעט אלו המשפיעות על גנים רטרו-ויראליים. תהליך זה הביא ל-2,018 תרופות הקשורות ל-2,232 גני מטרה. התרופות מיוצגות על ידי מספר מזהה כימי, וגנים מיוצגים על ידי שמות ומספר מזהה חלבוני מ-GenBank (Benson, et al., 2012).

### דאטה של אינטראקציית חלבון-חלבון (PPI)

אספנו אינטראקציות חלבון-חלבון (Protein-Protein Interaction) של בני אדם מאומתים על ידי ניסויים ממסד הנתונים BioGRID (Oughtred, et al., 2021) גרסה 4.4.217, ללא כפילויות (בשם 'non-redundant'). אנו מתייחסים לכל אינטראקציה בין חלבון א' לחלבון ב' כדו-כיוונית. מיזגנו אינטראקציות זהות שהתגלו במערכות ניסוי שונות או שפורסמו במקורות שונים. חלבונים מיוצגים על ידי סמלי הגנים הרשמיים שלהם, כלומר שמות הגנים. ההתייחסות לגנים ולחלבונים במאמר זה הינה זהה. כדי להבטיח קוהרנטיות וקישוריות במערך הנתונים, כללנו רק גנים אנושיים תוך החסרת אינטראקציות עם חלבונים רטרו-ויראליים. מערך הנתונים הסופי כולל 781,272 אינטראקציות חלבון-חלבון בקרב 19,893 חלבונים מהגנום האנושי.

### דאטה של מחלה-גן

## פיתוח Graph Neural Network (GNN)

מאחר ומודלי GNN מצטיינים בסיווג צמתים על ידי למידת מאפיינים מורכבים ממבני גרפים, אימנו שני מסווגי GNN על הרשת ההטרוגנית: Baseline GNN ו-DruGNNosis-MoA. בכדי לשפר את יכולת הלמידה של שני סוגי ה-GNN, שילבנו שתי שכבות SAmple-and-aggreGatE (SAGE) Convolution layers (SAGEConv) (Hamilton, et al., 2017) מספריית PyG (Fey, et al., 2019). נעזרנו בפונקציית האקטיבציה Rectified Linear Unit (ReLU) (Shang, et al., 2016) עבור זיהוי תבניות לא לינאריות, ופונקציית Sigmoid (Vlačić, et al., 2022) לחיזויים בינאריים, תוך שימוש באופטימיזציה Adam (Kingma, et al., 2014) להתאמות משקל במהלך אימון. ארכיטקטורה זו מהווה את הבסיס לפיתוח שני סוגי ה-GNNs.

**מודל Baseline GNN:** בהינתן רשת הטרוגנית, חישבנו את הערכים העצמיים (Eigenvector) לכל צומת (Huang, et al., 2020). ערכים אלו לוכדים תכונות אינפורמטיביות עבור כל צומת. לאחר מכן נרמלנו את הערכים העצמיים (לפי ציון Z) לאנליזות הבאות.

**מודל DruGNNosis-MoA:** בהינתן רשת הטרוגנית, מודל זה משלב תכונות טקסטואליות (הטמעות - embeddings) עבור כל צומת. כדי ליצור תכונות אלו, יישמנו את מודל ה-SciBERT המאומן על תיאורי התרופות מ-DrugBank, סמלי גנים מ-BioGRID, ושמות מחלות מ-OMIM. לאחר מכן ביצענו הפחתת מימדים לפלט הטקסטואלי מ-SciBERT למימד של 128 קואורדינטות על ידי יישום Principal Component Analysis (PCA). דבר זה ייעל את הקלט עבור מודל ה-DruGNNosis-MoA, וממקד את המודל בהיבטים הרלוונטיים ביותר של הנתונים.

## אימון Graph Neural Network (GNN)

את צמתי התרופות חילקנו לסט אימון (90%) ולסט מבחן (10%). ביצענו 10-fold Cross-Validation (CV) על סט האימון, ותוך כדי שמנו

על ידי F1-score. המודל הטוב ביותר אשר נבחר בזמן ה-CV, נמדד על סט המבחן.

## מידול רשתות

כדי לחשוף את היחסים המורכבים בין אלמנטים ביולוגיים וכימיים מגוונים, הרכבנו ארבע רשתות באמצעות ספריית NetworkX של פייתון (Hagberg, et al., 2008).

## הרכבת רשתות

**רשת תרופה-מטרה:** רשת המקשרת בין תרופות לגני המטרה שלהן על בסיס המידע מ-DrugBank. רשת זו מעוצבת כגרף דו-צדדי, בה צמתים תואמים לתרופות או לגני מטרה. קשתות מציינות מקרים בהם תרופה מכוונת לחלבון תוצר של גן מסוים.

**רשת אינטראקציות חלבון:** רשת המקשרת בין חלבונים אנושיים במערכת ביולוגית על בסיס המידע ב-BioGRID. חלבונים מיוצגים על ידי הגנים התואמים להם, ומופיעים ברשת כצמתים. שני חלבונים קשורים זה לזה אם דווחה אינטראקציה ביניהם.

**רשת מחלה-גן:** רשת המקשרת בין מחלות לגנים, כלומר פנוטיפ לגנוטיפ, על בסיס המידע ב-OMIM Morbid Map. הרשת מעוצבת כגרף דו-צדדי, בה צמתים מסמנים מחלות או גנים, וצלעות מציינות אסוציאציות ידועות בין מחלה לגן.

**רשת הטרוגנית:** מייזגנו את שלושת הרשתות לעיל בכדי להרכיב את הרשת ההטרוגנית. צמתי הרשת מייצגים תרופות, גנים ומחלות. לכל תרופה קיים תיוג המציין אם מנגנון הפעולה שלה היא אטיולוגית או פליאטיבית. סוגי הצלעות ברשת כוללות: תרופה-גן מטרה (מיקוד), גן-גן (אינטראקציות חלבון), ותרופה-גן (אסוציאציה). כדי להתאים את הנתונים למודל סיווג בינארי, השמטנו מהרשת תרופות המציגות מנגנון כפול של אטיולוגי ופליאטיבי. אנו משתמשים ברשת ההטרוגנית בשלבים הבאים לצורך פיתוח שני מודלי GNN, על ידי יישום ספריית PyTorch Geometric (PyG) עם אובייקט ה-HeteroData (Fey, et al., 2019).

המודל הנבחר הציג את הביצועים הטובים ביותר, הוערך על סט המבחן תוך שימוש באותם מדדים.

תרשים 1 מציג את מהלך שיטות העבודה בתצורה ויזואלית.

### תוצאות

התחלנו בתיוג ידני של מנגנוני פעולה של תרופות כאטיולוגיות או כפליאטיביות. באמצעות סט תיוגים זה, אימנו את מודל השפה SciBERT לסווג תרופות על בסיס מנגנוני הפעולה שלהן. לאחר מכן בנינו רשתות ביולוגיות ואימנו שני מודלי GNN בעלי קונפיגורציות שונות. לבסוף הערכנו את הביצועים של שלושת המודלים. בפרק התוצאות נבדוק את תוצאות הניתוחים.

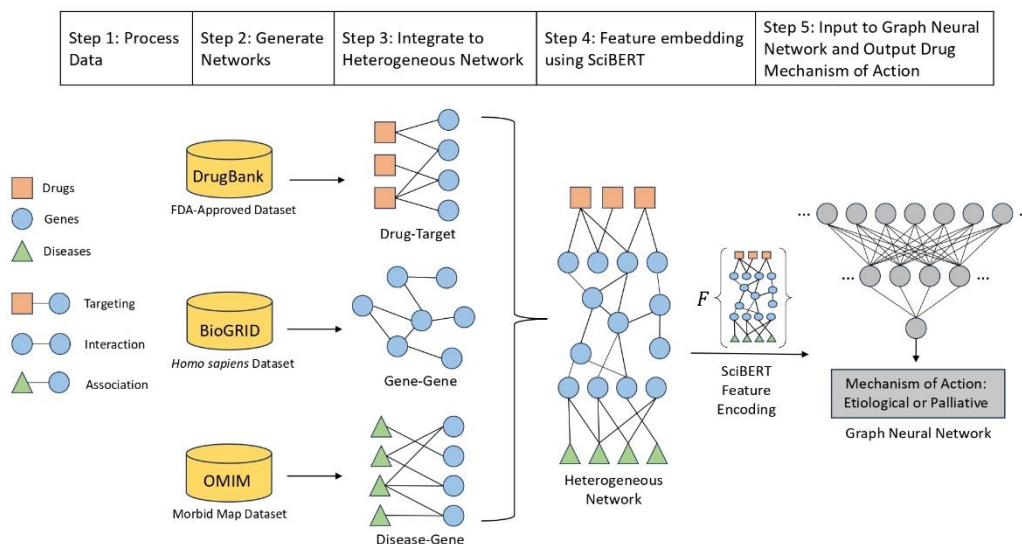
### תיוג מנגנוני פעולה של תרופות

חלק מהתרופות שנותחו (77, 3.816%) הראו מנגנון כפול אטיולוגי ופליאטיבי בהתבסס על המחלות בהן הן מטפלות. אלו הושמטו מהנתונים בכדי לאפשר סיווג בינארי. לדוגמה, מתוטרקסט (Methotrexate) פועלת כתרופה אטיולוגית

בצד 10% מימנו עבור סט האימות שהתחלף בכל שלב. החלוקה הראשונית ושלב ה-CV נערכו באופן מרובד (stratified), דבר המבטיח חלוקה מאוזנת ופרופורציונלית של מחלקות התיוגים, ואף מפחית את הסיכון לאימון מוטה של המודל. יישמנו גישת סיווג טרנסדוקטיבית המאפשרת לכל סט לגשת למבנה המלא של הרשת ההטרוגנית. עם זאת, כל סט מוגבל בכך שאינו יכול להבחין בתיוגים של צמתים מסט אחר.

### הערכת המודל

כדי להבטיח השוואה הוגנת בין מודל ה-SciBERT המאומן; מודל ה-Baseline GNN; ומודל ה-DruGNNosis-MoA במונחים של סיווג מנגנוני פעולה, הפעלנו אותו תהליך של 10-fold CV עבור הערכת הביצועים שלהם. המדדים הממוצעים דווחו לכל דגם של המודל בשלבי ה-10-fold CV, והם כוללים: Recall, F1-score, Precision, ו-AUC (Area Under the Curve). הכרעת המודל בעל הביצועים הטובים ביותר התקבל על ידי השוואת הממוצעים של ה-F1 score וה-Roc-AUC בין שלושת המודלים.



**תרשים 1.** תרשים זרימה של שיטות העבודה שיושמו במחקר זה כדי לבנות את מודל ה-DruGNNosis-MoA. המחשת השלבים המשמשים לסיווג תרופות כבעלי מנגנון אטיולוגי או פליאטיבי. בשלב 1, אספנו ועיבדנו מערכי נתונים הכוללים מידע על תרופות, גנים ומחלות. בשלב 2, הרכבנו שלוש רשתות נפרדות עבור אינטראקציות תרופה-מטרה, חלבון-חלבון ומחלה-גן. בשלב 3, מיזגנו רשתות אלו ליצירת רשת הטרוגנית מקיפה, המכילה מגוון סוגי צמתים וצלעות. בשלב 4, המרנו טקסט לתכונות צמתים תוך שימוש במודל SciBERT (Beltagy, et al., 2019). לבסוף, בשלב 5, שילבנו את תכונות הצמתים עם ה-GNN כדי לבצע את משימת הסיווג, הבחנת כל צומת תרופה כאטיולוגית או כפליאטיבית.

והן בתסמינים, מהוות אחוז קטן בכל המחלקות (7.746% - 1.587%). הדבר מצביע על קיום תפקידי תרופה מגוונים בתחומים טיפוליים שונים.

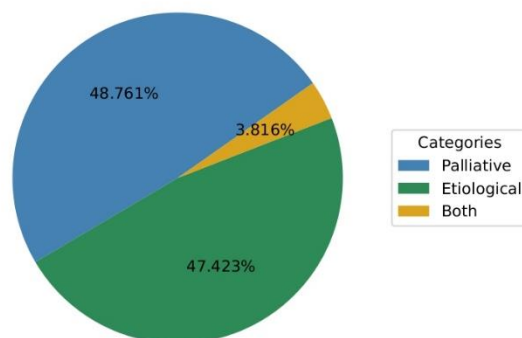
### אימון SciBERT

באמצעות תיוג התרופות מהסעיף הקודם, אימנו את מודל ה-SciBERT לסיווג תרופות. עבור הזנת טקסט למודל, השתמשנו בתיאורי תרופות מ-DrugBank; סמלי גנים מ-BioGRID; ושמות מחלות מ-OMIM. בהתאם להנחיות (Devlin, et al., 2018), נעזרנו במחלקת ה-Trainer מבית Hugging Face (Wolf, et al., 2019) עם הפרמטרים הבאים: אופטימיזציה AdamW עם דעיכה (decay) במשקל של  $2e^{-3}$  השונה מ-Adam; סטנדרטי בגישה הייחודית שלו ברגולציית הדעיכה; קצב למידה (learning rate) של  $2e^{-5}$ ; גודל מקבץ (batch size) של 16; ושלושה מחזורי אימון (epochs).

מודל SciBERT ייצג כל טקסט כווקטור מספרי (ווקטור הטמעות) בגודל 768. לכל סוג צומת (תרופות, גנים ומחלות) הרכבנו מטריצה בעלת מספר שורות כמספר הצמתים מסוג זה, ומספר עמודות של 768 כגודל הווקטור. מכאן ביצענו Principal Component Analysis (PCA) להפחתת ממדים מ-768 ל-128. תוכן הווקטורים המופחתים של כל סוג צומת שולב ברשת ההטרוגנית שלנו כתכונה של הצומת. אינטגרציה זו שיפרה משמעותית את יכולות הלמידה של ה-

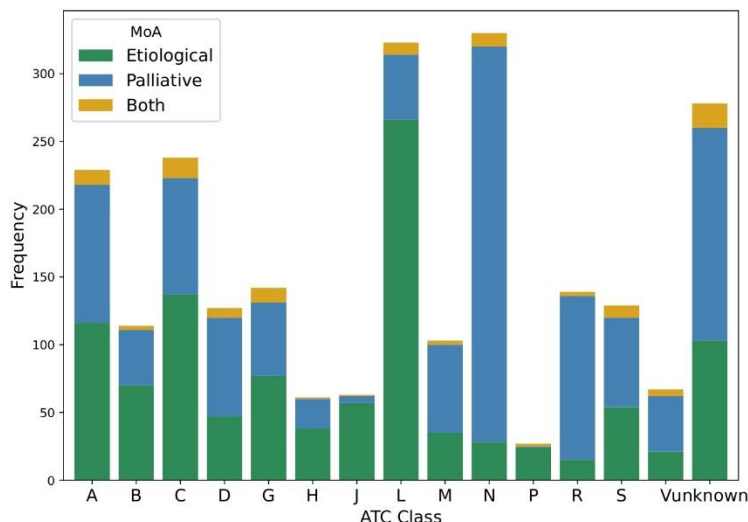
בטיפול בסרטן על ידי עיכוב חלוקת תאים, ופועלת כתרופה פליאטיבית בטיפול בדלקת פרקים על ידי הפחתת אותות המעודדים דלקת. התרופות הנותרות מציגות חלוקה כמעט שווה של מנגנונים אטיולוגיים (957, 47.423%) או פליאטיביים (984, 48.761%) (תרשים 2). זאת בניגוד למחקר של Yildirim, et al., (2007) אשר דיווחו על חוסר איזון במנגנונים. סיבות אפשריות להבדל זה כוללות שינויים פוטנציאליים במגמות עיצוב תרופות במשך 17 שנים, ושיטות סיווג שונות הנובעות מאי חשיפת שיטות ההליך המחקרי.

תרופות הקשורות למספר מחלקות ATC מצביעות על פוטנציאל הייעוד מחדש שלהן על ידי שימוש בהן בקטגוריות טיפוליות שונות. ניתחנו כיצד מנגנוני פעולה של תרופות מחולקים בין קטגוריות ה-ATC. הניתוח שלנו חשף התפלגויות ברורות בתוך כל מחלקה (תרשים 3). במחלקת 'מערכת העצבים' ('Nervous System'), מרבית התרופות הינן פליאטיביות (88.484%, 292), המתייחסות לתסמינים כמו כאב והפרעות פסיכיאטריות, אך לא לגורמים השורשיים של המחלה. לעומת זאת, במחלקת 'נוגדי זיהום' (



**תרשים 2.** התפלגות מנגנוני פעולה של תרופות. מספר התרופות האטיולוגיות והפליאטיביות מאוזן, כאשר 47.423% (957) הם אטיולוגיים ו-48.761% (984) הם פליאטיביים. רק אחוז קטן (3.816%, 77) מהתרופות מציגות את שני המנגנונים.

רוב התרופות הינן אטיולוגיות (90.476%, 57), המכוונות ספציפית לגורם הזיהום כמו חיידקים ונגיפים. תרופות בעלות מנגנון כפול, המטפלות הן בשורש



**תרשים 3.** התפלגות מנגנוני פעולה של תרופות על פני מחלקות אנטומיות של טיפולים כימיקלים (Anatomical Therapeutic Chemical (ATC) classes). לכל מחלקת ATC יש התפלגות ייחודית של מנגנוני תרופות. מחלקות מסוימות, למשל 'L', 'J', ו-'P', מכילות בעיקר תרופות בעלות מנגנונים אטיולוגיים. לעומת זאת, מחלקות כמו 'N' ו-'R' מורכבות בעיקר מתרופות בעלות מנגנונים פליאטיביים. המחלקות הנותרות מפגינות מעין איזון בין תרופות עם מנגנונים אטיולוגיים ופליאטיביים. הגדרות מחלקות ATC:

A: Alimentary tract and metabolism; B: Blood and blood forming organs; C: Cardiovascular system; D: Dermatologicals; G: Genito urinary system and sex hormones; H: Systemic hormonal preparations, excluding sex hormones and insulins; J: Antiinfectives for systemic use; L: Antineoplastic and immunomodulating agents; M: Musculo-skeletal system; N: Nervous system; P: Antiparasitic products, insecticides and repellents; R: Respiratory system; S: Sensory organs; V: Various; Unknown: Drug entries that do not have an ATC class assigned to them in DrugBank.

1, מצביע על שכיחות צמתים עם חיבור יחיד. דרגות גבוהות אינן נפוצות, ממצא המתיישר עם מיעוט צמתי ה-hubs אשר קובעים את רמת הקישוריות של הרכיב.

רשת אינטראקציות חלבון: רשת זו מכילה 19,893 צמתים של חלבונים המחוברים ביניהם על ידי 781,272 קשתות, למעט 3,344 אינטראקציות עצמיות. היא מורכבת מארבעה רכיבי קשירות כאשר שלושה מהם הינם חלבונים מבודדים. כלומר, למעט שלושה חלבונים, ניתן למצוא מסלול מחלבון אחד לכל חלבון אחר. הדרגה השכיחה ביותר היא 1, ממצא המצביע על כך שלמרות הקישוריות בין החלבונים, לרוב הם מסועפים ומוחזקים על ידי קשר אחד.

רשת מחלה-גן: רשת זו מכילה 8,156 צמתים של מחלות, 16,358 צמתים של גנים, ו-27,692 קשתות המחברות ביניהם. היא מורכבת מ-5,430 רכיבי

GNN בפיתוח תכונות חזקות ומייצגות יותר עבור הצמתים.

### בניית רשתות

סיכום כללי של הרשתות שנותחו בפרויקט זה מוצג בטבלה 1.

רשת תרופה-מטרה: רשת זו מכילה 2,018 צמתים של תרופות, 2,232 צמתים של חלבוני מטרה, ו-10,250 צלעות המחברות ביניהם. הרשת מורכבת מ-129 רכיבי קשירות, כאשר 75 מתוכם הינם זוגות של תרופה ומטרה. הבחנו כי קיים מספר קטן של צמתים בעלי קישוריות גבוהה (hubs), ואילו מספר רב של צמתים בעלי קישוריות נמוכה (Barabási, 2009; Barabási, et al., 2003). רכיב הקשירות הגדול ביותר ברשת כולל 1,797 צמתי תרופות, 2,027 צמתי חלבונים, ו-9,916 קשתות. יש לציין כי הדרגה הנפוצה ביותר ברכיב זה היא

טבלה 1. מאפייני הרשתות שנותחו.

מספר צלעות	מספר צמתים	רשת
10,250	2,018 תרופות 2,232 גנים	תרופה - מטרה
781,272	19,893 חלבונים	חלבון - חלבון
27,692	8,156 מחלות 16,358 גנים	מחלה - גן
822,086	1,941 תרופות 31,524 גנים 8,156 מחלות	הטרוגנית*

למגוון צר של חלבוני מטרה המשמשים כמטרות טיפול למחלות ולתסמינים, ואף מגבילה את מגוון אפשרויות הטיפול.

תרשים 4 ממחיש תת-גרף של הרשת ההטרוגנית המציג את התרחשותן של תרופות "follow-on" הקשורות לגנים של סרטן מעי הגס (ירוק). בתת-גרף זה ישנם 48 גנים (כחול) הקשורים לסרטן מעי הגס, חלקם עם קשרים לגנים אחרים. תת-קבוצה קטנה של גנים אלו מהווים מטרות טיפול של תרופות (כתום), והיא חושפת התרחשות מקבץ תרופות המתמקדות במטרות מסוימות ובודדות. ניתן לראות כי רוב התרופות הללו מסווגות כאטיולוגיות, דבר התואם לממצאים שלנו כי רוב התרופות המיועדות לטיפול בסרטן פועלות למניעת חלוקת תאים ועיכוב גדילתם. מודלי ה-GNN שלנו, אשר נועדו לסווג תרופות כאטיולוגיות או כפליאטיביות, בנויים על רשת הטרוגנית זו.

### אימון Graph Neural Network (GNN)

מבחינת ארכיטקטורה, מודל ה-Baseline GNN ומודל ה-DruGNNosis-MoA שפיתחנו חולקים מספר מרכיבים. הם בנויים משתי שכבות SageConv, פונקציית אקטיבציה ReLU, אופטימיזציית Adam עם שיעורי למידה של  $5e^{-4}$  למודל Baseline GNN ו- $6e^{-4}$  למודל DruGNNosis-MoA, פונקציית Sigmoid על הפלט, ולבסוף טכניקת gradient clipping עם סף

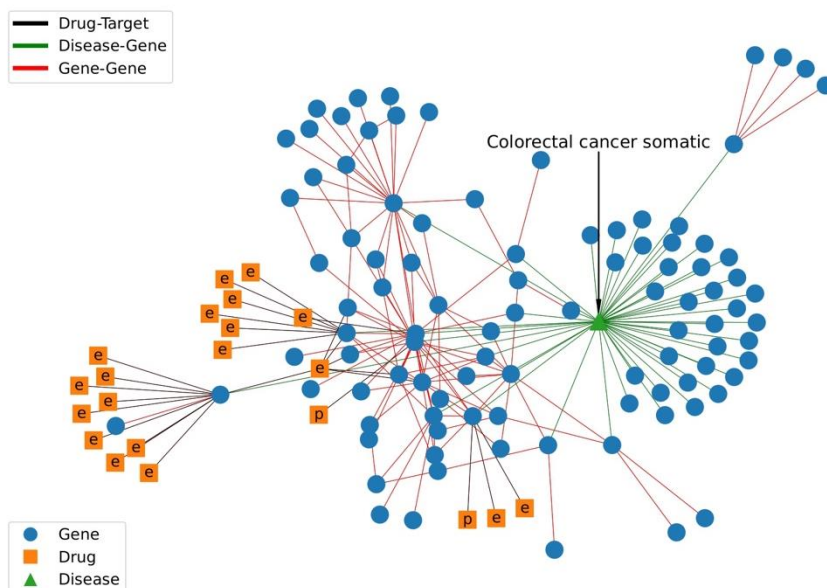
קשירות, כאשר 1,065 מתוכם הינם זוגות עצמאיים של מחלה וגן. כל רכיב קשירות כולל סט ייחודי של מחלות וגנים הקשורים זה לזה. רכיב הקשירות הגדול ביותר כולל 265 צמתי מחלות, 379 צמתי גנים, ו-1,493 קשתות. הדרגה השכיחה ביותר היא 2, כלומר רוב המחלות והגנים ברכיב זה קשורים לשני גנים/מחלות אחרים.

רשת הטרוגנית: מכילה 1,941 צמתים של תרופות (לאחר השמטת תרופות בעלות מנגנון כפול של אטיולוגי ופליאטיבי), 31,524 צמתים של גנים, ו-8,156 צמתים של מחלות. סוגי הקשתות מתחלקים על פי: 9,778 קשרים בין תרופות לגנים, 784,616 קשרים בין חלבונים, ו-27,692 קשרים בין מחלות לגנים, סך הכל 822,086 קשתות ברשת.

עבור שני דגמי ה-GNN שפותחו, חישבנו את מטריצות מאפייני הצמתים הבאות: (i) תכונות Eigenvector עבור מודל ה-Baseline GNN; ו-(ii) תכונות טקסטואליות (פלט מודל ה-SciBERT המאומן) עבור מודל DruGNNosis-MoA. בשני המודלים הרשתיים יש תיוג אטיולוגי או פליאטיבי לצמתי התרופות.

ניתוח טופולוגי חשף התרחשות משמעותית של תרופות "follow-on". אלו תרופות המכוונות לחלבונים המהווים מטרות לתרופות קודמות. מגמה זו משקפת אסטרטגיה רווחת בתעשיית התרופות לרכז מאמצים בחלבוני המטרה. כתוצאה מכך ישנה הצטברות של תרופות המטפלות במאגר מצומצם יחסית של חלבונים (Aronson, et al., 2020; DiMasi, et al., 2011; DiMasi, et al., 2004). אסטרטגיה זו גורמת





**תרשים 4.** תת-גרף של הרשת ההטרוגנית. מחלת סרטן המעי הגס (משולש ירוק) קשורה באופן ישיר ועקיף לגנים רבים (נקודות כחולות). למטרות הדמיה, בחרנו את כל הגנים המקושרים ישירות, וכמה גנים מקושרים בעקיפין, לסכום של 50 גנים. חלק מהגנים הללו הינם מטרות יעד של תרופות (ריבועים כתומים). כל תרופה מסומנת כבעלת מנגנון אטיולוגי ('e') או פליאטיבי ('p'). ניתן לראות תרופות "follow-on", שהן תרופות המכוונות לאותו גן כמו תרופות אחרות. רשת זו מראה את קיומם של יעדי תרופות פוטנציאליים רבים לסרטן המעי הגס, אך רק מעטים נבחרו לטיפול תרופתי עד כה. בנוסף, תת-גרף זה מגלה דומיננטיות של תרופות אטיולוגיות, התואם לממצאים שלנו לגבי מחלקת 'L' בקטגוריות ATC.

DruGNNosis-MoA משתמש בתכונות המופקות מניתוחי טקסט.

### הערכת המודל

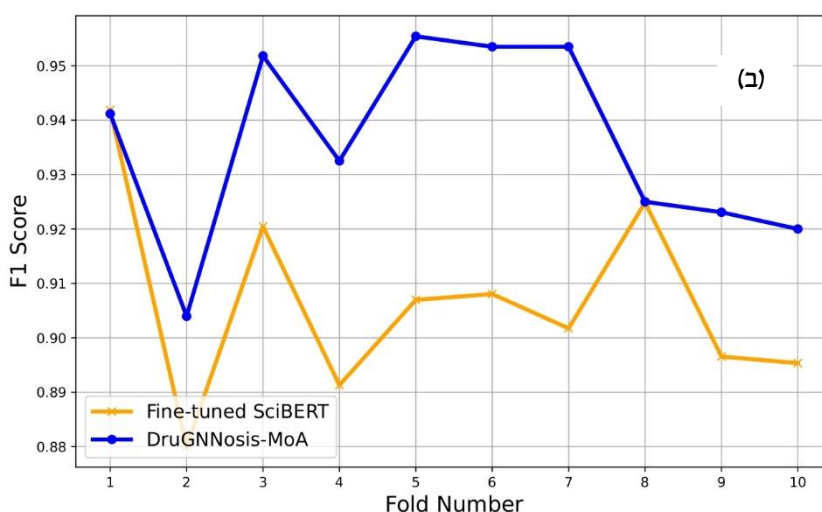
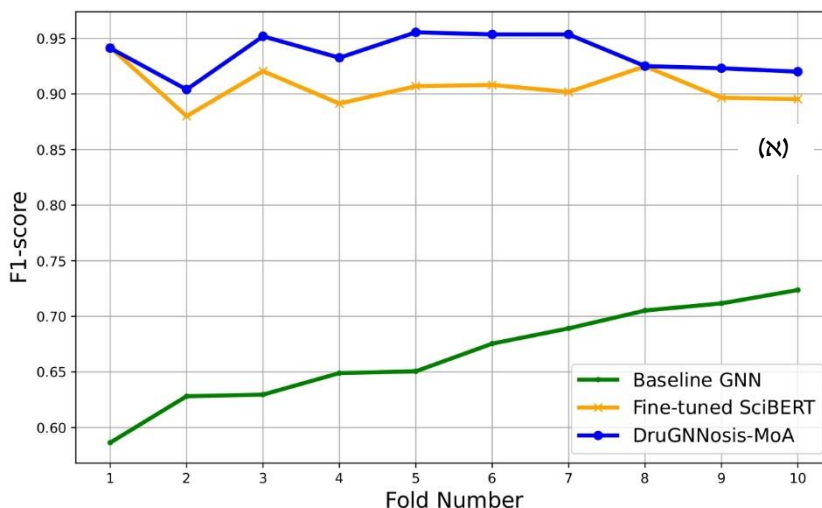
כדי להשוות בין שלושת המודלים שפותחו (DruGNNosis, Baseline GNN, SciBERT), ביצענו תהליך של 10-fold Cross Validation. עבור כל fold, חישבנו את הערכים הבאים: accuracy, F1 score, precision, recall, ROC-AUC (Area Under the Curve). לאחר מכן, חישבנו את הציון הממוצע ואת סטיית התקן עבור כל מטריצה (טבלה 2).

תחילה, הערכנו את הביצועים של ה-F1 score ב-10 fold CV על ידי יצירת גרפים השוואתיים עבור שלושת המודלים (תרשים 5א). מכאן ערכנו בדיקה מפורטת יותר של מגמות הביצועים של שני המודלים המובילים; SciBERT ו-DruGNNosis-MoA (תרשים 5ב). תרשים 5 מצגי השוואה בין כל fold ב-F1-score. מודל ה-DruGNNosis-MoA מצגי ציוני F1-score גבוהים יותר בכל fold בהשוואה למודל ה-SciBERT למעט folds מספר 1 ו-8, בהם

מוגדר של 0.5 כדי למנוע gradient explosion ולהבטיח אימון יציב ויעיל.

עם זאת, עבור מודל DruGNNosis-MoA, גישה זו התבררה כבלתי מספקת במניעת overfitting, מה שהוביל לשילוב אסטרטגיית dropout נוספת לאחר השכבה הראשונה, ומוגדרת בהסתברות של 0.3. אסטרטגיית ה-dropout פועלת על ידי כיבוי אקראי של חלק מהצמתים במהלך האימון, דבר העוזר להפחית overfitting על ידי מניעת התאמות מורכבות בנתוני האימון.

מודל ה-DruGNNosis-MoA נזקק ל-30 epochs בלבד, ואילו מודל ה-Baseline GNN דרש 100 epochs. יתרה על כן, מאפייני הצמתים בעלי תפקיד מפתח באימון GNN ומשפיעים משמעותית על ההתנהגות ועל הביצועים של המודל (Mahmoud, et al., 2023; Duong, et al., 2019). הבדל נוסף בין המודלים טמון באפייני הצמתים שלהם: מודל ה-Baseline GNN משתמש בתכונות הנגזרות ממטריצות הערכים העצמיים (EigenVector), ואילו מודל ה-

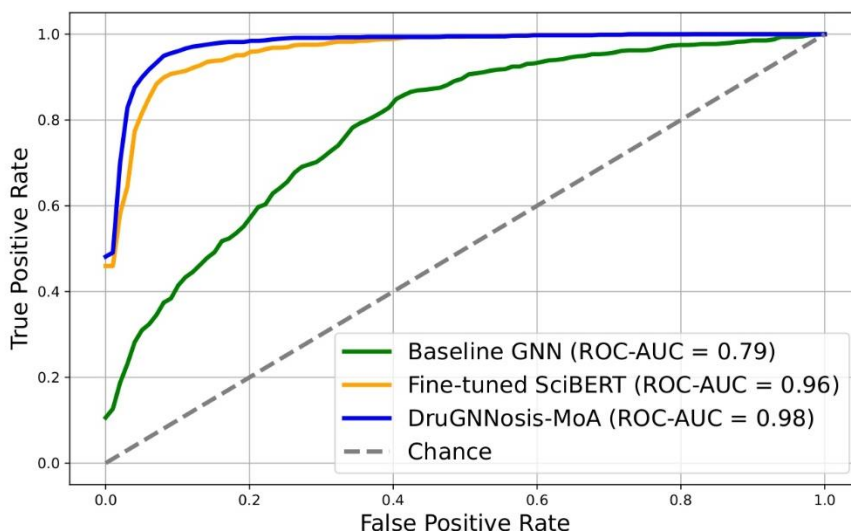


**תרשים 5.** הערכת ביצועים. (תרשים א5) השוואת F1-scores של שלושה מודלים שונים. סקירה מקיפה של ביצועי F1-score של שלושה מודלים שפותחו במחקר זה כדי לסווג תרופות כבעלות מנגנון פעולה אטיולוגי או פליאטיבי. התוצאות של מודל ה-Baseline GNN מאופיינות ב-F1-score ממוצע של 0.664 וסטיית תקן של 0.043. מבין שלושת המודלים, מודל ה-Baseline GNN מציג את הביצועים הנמוכים ביותר. הקו הכחול מייצג את מודל ה-DruGNNosis-MoA, הקו הכתום מייצג את מודל SciBERT (Beltagy, et al., 2019), והקו הירוק מייצג את מודל ה-Baseline GNN. (תרשים ב5) השוואת F1-scores בין מודלים SciBERT ו-DruGNNosis-MoA. כל נקודה מייצגת את ה-F1-score שהתקבל עבור ה-fold המסוים ב-cross validation. ניתן לראות כי מודל ה-DruGNNosis-MoA בעל ביצועים טובים יותר לאורך כל fold, למעט 1 ו-8 בהם הם שווים.

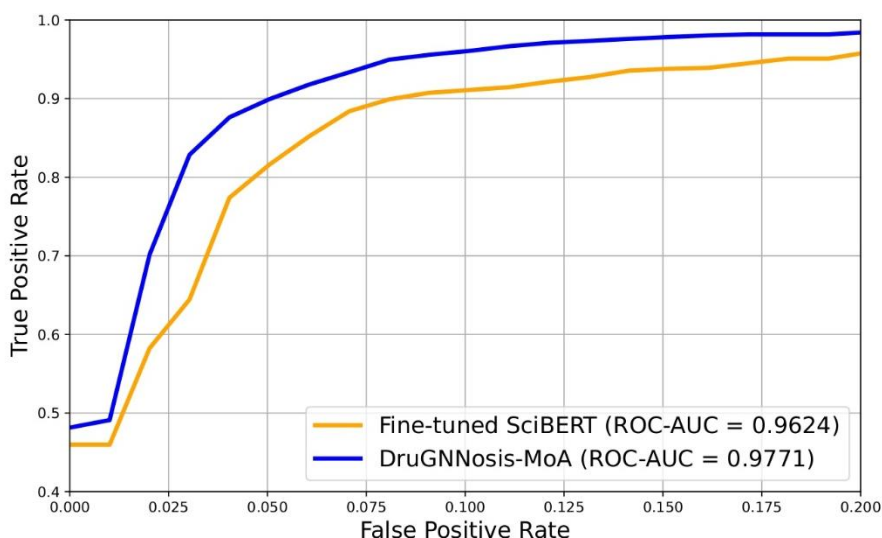
עקומה זו מדגימה את הביצועים העולים בעקבות של מודל ה-DruGNNosis-MoA בסיווג תרופות. המודל משיג ממוצע מרשים ב-AUC של 0.977, הנתמך ברמת היציבות הגבוהה ביותר, הנובעת מערך סטיית התקן הנמוך ביותר של 0.001. תוצאות אלו מצביעות על ההבדל בביצועים בין

הביצועים שווים. יש לציין כי כמחצית מה-folds, מודל ה-DruGNNosis-MoA משיג ביצועים מצוינים ב-F1-score, המגיע עד ל-0.955. דפוס זה מצביע על מגמה ב-DruGNNosis-MoA מצליח יותר מ-SciBERT במשימת סיווג תרופות כאטיולוגיות או כפליאטיביות (תרשים ב.5).

לאחר ההבחנה כי מודל ה-DruGNNosis-MoA מפגין ביצועים מעולים ב-F1-score, הערכנו את יעילותו באמצעות עקומת ROC-AUC. או משרטטים את עקומות ה-ROC-AUC הממוצעת של שלושת המודלים להשוואה ביניהם בתרשים 6.



**תרשים 6.א.** עקומת ה-ROC משווה את הביצועים של שלושה מודלים על ידי 10-fold cross validation: מודל Baseline GNN (כחול), מודל SciBERT (כתום) ומודל DruGNNosis-MoA (כחול). כל קו מייצג את ממוצע ה-ROC Area Under Curve (AUC) עבור מודל מסוים, ובכך מציין את שיעור ה-True Positive מול שיעור ה-False Positive. הקו המקווקו מייצג את הביצועים של מסווג אקראי לצורך בקרה.



**תרשים 6.ב.** עקומת ה-ROC משווה את הביצועים של שני מודלים על ידי 10-fold cross validation: מודל SciBERT (כתום) ומודל DruGNNosis-MoA (כחול). כל קו מייצג את ממוצע ה-ROC Area Under Curve (AUC) עבור מודל מסוים, ובכך מציין את שיעור ה-True Positive (נע בין 0.4 ל-1) מול שיעור ה-False Positive (נע בין 0 ל-0.2).

סטיית התקן הגבוהה ביותר שנצפתה הינה 0.029 ב-precision, בעוד שסטיית התקן הנמוכה ביותר היא 0.001 במדד ה-ROC-AUC. יש לציין כי מדד זה הינו הנמוך ביותר מבין כלל המ (א) והמודלים.

המודלים בהתאמה ליעילותם במטלת סיווג תרופות.

ככלל אנו רואים כי מודל ה-DruGNNosis-MoA מצטיין על פני המודלים האחרים במונחים של: average accuracy, F1-score, precision, recall, and ROC-AUC, כפי שניתן לראות בטבלה 2. בהערכת מודל DruGNNosis-MoA,

טבלה 2. תוצאות ה- Cross Validation עבור שלושת המודלים.

מודל	מטריקה	ממוצע	סטיית תקן
Baseline GNN	Accuracy	0.699	0.034
	F1 score	0.664	0.043
	Precision	0.715	0.051
	Recall	0.622	0.045
	ROC-AUC	0.787	0.03
SciBERT	Accuracy	0.906	0.018
	F1 score	0.906	0.018
	Precision	0.895	0.028
	Recall	0.918	0.012
	ROC-AUC	0.961	0.012
DruGNNosis-MoA	Accuracy	0.938	0.017
	F1 score	0.935	0.017
	Precision	0.931	0.029
	Recall	0.941	0.022
	ROC-AUC	0.977	0.001

התכונות והקשרים הסמויים ברשת, אנו מקבלים הבנה של יחסי הגומלין בין תרופות, גנים ומחלות.

בניגוד לממצאים של Yildirim, et al., (2007) שדיווחו על שכיחות תרופות פליאטיביות, המחקר שלנו (17 שנים מאוחר יותר בשנת 2024), מגלה איזון בין מנגנונים אטיולוגיים ופליאטיביים בקרב תרופות. דבר זה מצביע על שינוי פוטנציאלי בפיתוח תרופות לאורך זמן לטובת טיפולים המכוונים לגורמים השורשיים של מחלות, או מציע תוצאה שונה עקב הבדלים במתודולוגיות הסיווג. בעוד שהגישה של Yildirim, et al., (2007) מבוססת מרחק, הממצאים שלנו מצביעים על היעדר קורלציה משמעותית בין מנגנון הפעולה לבין המרחק בין גן מטרה לגן אסוציאטיבי למחלה. אבחנה זו מיוחסת לשימוש שלנו בנתונים נרחבים ועדכניים מאשר בפרסום הקודם.

מחקרנו חשף תת-קבוצה של תרופות בעלות מנגנונים אטיולוגיים ופליאטיביים, המשתרעים על מחלקות ATC רבות. הממצא של תרופות דו-מנגנוניות מאתגר את הסיווג הבינארי של מנגנוני תרופות.

בעת פירוש התוצאות, חיוני להכיר במגבלות השיטות של מחקר זה. התיוג הידני של מנגנוני תרופות כאטיולוגיים לעומת פליאטיביים הוא לפעמים ברור ומובהק, ולעיתים סובייקטיבי. בסיווג מנגנוני פעולה של תרופות התייחסנו לקבוצת המחלות של תרופה ולא למחלה אחת ספציפית. במקרים מסוימים (כגון פלאק

DruGNNosis- MoA על סט המבחן שלנו, עליו המודל השיג F1- score מרשים של 0.942 וערך ROC-AUC של 0.975. תוצאות אלו מחזקות את ההצלחה העקבית והגוברת של מודל זה בסיווג תרופות.

### מסקנות וסיכום

זיהוי מנגנוני פעולה של תרופות כאטיולוגיים או כפליאטיביים הינו בעל חשיבות רבה בתחומי פיתוח תרופות ובחירת הטיפול במחלות. הבחנה זו של אופן תפקוד התרופות יכולה לשמש בסיס לקידום הפיתוח של תרופות חדשות ושינוי ייעוד של תרופות. לימוד מנגנוני תרופות נחקר רבות, אך הסיווג השיטתי לפי פעולות אטיולוגיות ופליאטיביות לא נחקר בצורה מספקת. הבחנה זו בין מנגנוני התרופות מעוררת שאלה בסיסית: האם מטרת פיתוח תרופה הינה לטפל בגורמים למחלה, ו/או להקל על תסמינים ולשפר את איכות החיים של המטופלים. ההבחנה בין מנגנוני תרופות ובין המטרות הטיפוליות שלהן חיונית לקידום טיפולים רפואיים.

מודל DruGNNosis-MoA אותו פיתחנו נעזר בקלט טקסטואלי המשפר משמעותית את ביצועיו יחסית למודל ללא ניתוח טקסטים. המודל משלב נתונים מגוונים עם יחסים מורכבים בין הקודקודים ובכך מסייע בחשיפת קשרים נסתרים עם פוטנציאל לקדם פיתוח תרופות. על ידי חשיפת

ברורה יותר בין מנגנוני תרופות אטיולוגיים לפליאטיביים. דבר זה יתרום להבנה עמוקה יותר של הקשרים הבסיסיים והקשרים המורכבים הקשורים למחלה ספציפית. כתוצאה מכך, ניתן לשפר את תהליך בחירת הטיפולים, תוך התחשבות בתופעות הלוואי האפשריות של בחירת גן מטרה, בהתבסס על חיבורי הרשת.

הגישה האנליטית המוצגת במחקר שלנו תורמת להתקדמות הערכה שיטתית של מנגנוני הפעולה של תרופות מנקודת מבט אטיולוגית לעומת פליאטיבית. אסטרטגיה מבוססת למידת מכונה מציגה את הפוטנציאל לשפר את הדיוק והיעילות של מאמצי פיתוח תרופות עתידיים אשר יביאו לשיפור הטיפול, חיסכון כספי לחברות הפארמה ושיפור הטיפול לחולים

פסוריאזיס חמור/מתון; או אסטמה ואסטמה כרונית), המחלות המרובות קשורות קשר הדוק, בעוד שבמקרים אחרים, ההפרעות הטרוגניות.

תהליך פיתוח תרופות מאופיין בשיעור כישלון גבוה (Xue, et al., 2018). לגישה שלנו פוטנציאל לסייע למאמצי פיתוח תרופות ולחקור גנים שטרם הוכיחו את היכולת הטיפולי שלהם. גישה זו יכולה להפחית את השכיחות של תרופות "follow-on" ולזהות יעדים חדשים בהתבסס על המבנה והקשרים של הרשת. ניתן להפחית את עלויות פיתוח התרופות על ידי מיקוד משאבים בגנים ספציפיים במקום בדיקה של אפשרויות רבות.

התמקדות במחלקה ספציפית של ATC בתוך הרשת מספקת תובנות חשובות לגבי היחסים והחיבורים בתוך אותה מחלקה. גישה מרוכזת זו מאפשרת ניתוח מקיף יותר ויכולת להבחין בצורה

## ביבליוגרפיה

- Ravikumar, B., & Aittokallio, T. (2018). Improving the efficacy-safety balance of polypharmacology in multi-target drug discovery. *Expert Opinion on Drug Discovery*, *13*(2), 179-192.
- Yu, L., Shi, Y., Zou, Q., Wang, S., Zheng, L., & Gao, L. (2020). Exploring drug treatment patterns based on the action of drug and multilayer network model. *International Journal of Molecular Sciences*, *21*(14), 5014.
- Xue, H., Li, J., Xie, H., & Wang, Y. (2018). Review of drug repositioning approaches and resources. *International Journal of Biological Sciences*, *14*(10), 1232-1246.
- Morgan, S., Grootendorst, P., Lexchin, J., Cunningham, C., & Greyson, D. (2011). The cost of drug development: A systematic review. *Health Policy*, *100*(1), 4-17.
- Ciociola, A. A., Cohen, L. B., Kulkarni, P., Kefalas, C., Buchman, A., Burke, C., ... & FDA-Related Matters Committee of the American College of Gastroenterology. (2014). How drugs are developed and approved by the FDA: Current process and future directions. *Official Journal of the American College of Gastroenterology*, *109*(5), 620-623.
- Ravikumar, B., & Aittokallio, T. (2018). Improving the efficacy-safety balance of polypharmacology in multi-target drug discovery. *Expert Opinion on Drug Discovery*, *13*(2), 179-192.
- Yıldırım, M. A., Goh, K. I., Cusick, M. E., Barabási, A. L., & Vidal, M. (2007). Drug-target network. *Nature Biotechnology*, *25*(10), 1119-1127.
- Lindpaintner, K. (2002). The impact of pharmacogenetics and pharmacogenomics on drug discovery. *Nature Reviews Drug Discovery*, *1*(6), 463-469.
- Vogt, I., Prinz, J., & Campillos, M. (2014). Molecularly and clinically related drugs and diseases are enriched in phenotypically similar drug-disease pairs. *Genome Medicine*, *4*(7), 1-17.
- Yu, L., Shi, Y., Zou, Q., Wang, S., Zheng, L., & Gao, L. (2020). Exploring drug treatment patterns based on the action of drug and multilayer network model. *International Journal of Molecular Sciences*, *21*(14), 5014.

- Parvathaneni, V., Kulkarni, N. S., Muth, A., & Gupta, V. (2019). Drug repurposing: A promising tool to accelerate the drug discovery process. *Drug Discovery Today*, *24*(10), 2076-2085.
- Olson, T., & Singh, R. (2017). Predicting anatomic therapeutic chemical classification codes using tiered learning. *BMC Bioinformatics*, *18*(8), 1-13.
- He, S., Wen, Y., Yang, X., Liu, Z., Song, X., Huang, X., & Bo, X. (2020). PIMD: An integrative approach for drug repositioning using multiple characterization fusion. *Genomics, Proteomics & Bioinformatics*, *18*(5), 565-581.
- Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., ... & Sun, M. (2020). Graph neural networks: A review of methods and applications. *AI Open*, *1*, 57-81.
- Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. *arXiv Preprint*.
- Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., ... & Wilson, M. (2018). DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Research*, *46*(D1), D1074-D1082.
- Huang, Q., He, H., Singh, A., Lim, S. N., & Benson, A. R. (2020). Combining label propagation and simple models out-performs graph neural networks. *arXiv Preprint*.
- Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2012). GenBank. *Nucleic Acids Research*, *41*(D1), D36-D42.
- Oughtred, R., Rust, J., Chang, C., Breitkreutz, B. J., Stark, C., Willems, A., ... & Tyers, M. (2021). The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science*, *30*(1), 187-200.
- Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., & Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research*, *43*(D1), D789-D798.
- Hagberg, A., Swart, P., & Chult, D. S. (2008). Exploring network structure, dynamics, and function using NetworkX (No. LA-UR-08-05495; LA-UR-08-5495). Los Alamos National Lab. (LANL), Los Alamos, NM (United States).
- Fey, M., & Lenses, J. E. (2019). Fast graph representation learning with PyTorch Geometric. *arXiv Preprint*.
- Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems*, *30*.
- Shang, W., Sohn, K., Almeida, D., & Lee, H. (2016, June). Understanding and improving convolutional neural networks via concatenated rectified linear units. In *International Conference on Machine Learning* (pp. 2217-2225). PMLR.
- Vlačić, V., & Bölcskei, H. (2022). Neural network identifiability for a family of sigmoidal nonlinearities. *Constructive Approximation*, *55*(1), 173-224.
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv Preprint*.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv Preprint*.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv Preprint*.
- Barabási, A. L. (2009). Scale-free networks: A decade and beyond. *Science*, *325*(5939), 412-413.
- Barabási, A. L., & Bonabeau, E. (2003). Scale-free networks. *Scientific American*, *288*(5), 60-69.
- Aronson, J. K., & Green, A. R. (2020). Me-too pharmaceutical products: History, definitions, examples, and relevance to drug shortages and essential medicines lists. *British Journal of Clinical Pharmacology*, *88*(11), 2114-2122.
- DiMasi, J. A., & Faden, L. B. (2011). Competitiveness in follow-on drug R&D: A race or imitation? *Nature Reviews Drug Discovery*, *10*(1), 23-27.
- DiMasi, J. A., & Paquette, C. (2004). The economics of follow-on drug research and development: Trends in entry rates and the timing of development. *Pharmacoeconomics*, *22*, 1-14.
- Mahmoud, A. M., Desuky, A. S., Eid, H. F., & Ali, H. A. (2023). Node classification with graph neural network based centrality measures and feature selection. *International Journal of Electrical and Computer Engineering*, *13*(2), 2114.

- Duong, C. T., Hoang, T. D., Dang, H. T. H., Nguyen, Q. V. H., & Aberer, K. (2019). On node features for graph neural networks. *arXiv Preprint*.
- Andrade, L. F., Sermet, C., & Pichetti, S. (2016). Entry time effects and follow-on drug competition. *The European Journal of Health Economics*, *17*, 45-60.