

دانشگاه قم

دانشکده فنی و مهندسی

پایان نامه دوره کارشناسی ارشد

رشته مهندسی فناوری اطلاعات گرایش تجارت الکترونیک

عنوان:

روشی برای باز‌هدف‌گذاری داروها با در نظر گرفتن عوارض‌جانبی

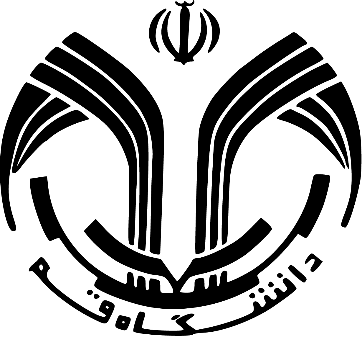
نگارنده:

سید محمد حسن میراشرفی

استاد راهنما:

دکتر امیر لکی‌زاده

تابستان/99



«صورت جلسه دفاع از پایان نامه کارشناسی ارشد»

با تأییدات خداوند متعال و با استعانت از حضرت ولی‌عصر (عجل الله تعالی فرجه الشریف) جلسه دفاعیه پایان نامه کارشناسی‌ارشد آقای/خانم: سید محمد حسن میراشرفی رشته: مهندسی فناوری اطلاعات (IT) تحت عنوان: **روشی برای باز‌هدف‌گذاری داروها با در نظر گرفتن عوارض‌جانبی** با حضور هیأت داوران در محل دانشگاه قم در تاریخ: / / تشکیل گردید.

در این جلسه، پایان‌نامه با نمره (به عدد...................، به حروف ...............................................) و با درجه: عالی بسیار خوب خوب قابل قبول مورد دفاع قرار گرفت.

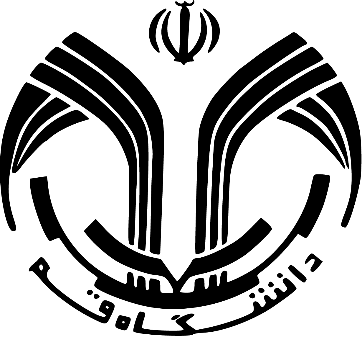
|  |  |  |  |
| --- | --- | --- | --- |
| نام و نام‌خانوادگی | سمت | مرتبه علمی | امضاء |
|  | استاد راهنما |  |  |
|  | استاد مشاور |  |  |
|  | استاد ناظر |  |  |
|  | استاد ناظر |  |  |
|  | نماینده کمیته تحصیلات تکمیلی |  |  |

**مدیر امور آموزشی و تحصیلات تکمیلی**

**نام و امضاء**

**معاون آموزشی و پژوهشی دانشکده**

**نام و امضاء**



مدیریت تحصیلات تکمیلی

«تعهد نامه اصالت پایان نامه/ رساله»

اینجانب آقای سید محمد حسن میراشرفی به شماره دانشجویی: 9514163029 دانشجوی مقطع کارشناسی ارشد دوره/دکتری رشته مهندسی فناوری اطلاعات که در تاریخ ----/--/99 از پایان نامه /رساله خود تحت عنوان " **روشی برای باز‌هدف‌گذاری داروها با در نظر گرفتن عوارض‌جانبی** " با کسب نمره و درجه دفاع نموده‌ام. بدین وسیله متعهد می شوم.

این پایان نامه / رساله حاصل تحقیق و پژوهش انجام شده توسط اینجانب بوده و در مواردی که از دستاوردهای علمی و پوژهشی دیگران (اعم از پایان نامه، کتاب، مقاله و...) استفاده نموده ام، مطابق ضوابط و رویه موجود، نام منبع مورد استفاده و سایر مشخصات آن را در فهرست مربوطه ذکر و درج کرده‌ام.

این پایان نامه / رساله قبلا برای دریافت هیچ مدرک تحصیلی (هم سطح، پایین تر یا بالاتر) در سایر دانشگاه ها و موسسات آموزش عالی ارائه نشده است.

چنانچه بعد از فراغت از تحصیل، قصد استفاده و هر گونه بهره برداری اعم از چاپ کتاب، ثبت اختراع و... از این پایان نامه را داشته باشم، از حوزه معاونت پژوهشی واحد مجوزهای مربوطه را اخذ نمایم.

چنانچه در هر مقطع زمانی خلاف موارد فوق ثابت شود، عواقب ناشی از آن را می پذیرم و دانشگاه قم مجاز است با اینجانب مطابق ضوابط و مقررات رفتار نموده و در صورت ابطال مدرک تحصیلی اینجانب هیچ گونه ادعایی نخواهم داشت.

**صحت امضای دانشجو مورد تایید است.**

**ناظر و نماینده تحصیلات تکمیلی**

**نام و نام خانوادگی:**

**تاریخ و امضاء**

**نام و نام خانواده گی متعهد:**

**تاریخ و امضاء**

تقدیم به

به تمامی کسانی که به هر شکل در جهت تعالی جامعه بشری تلاش می‌کنند.

تشکر و قدردانی

حمد و سپاس خدای را که توفیق کسب دانش و معرفت را به ما عطا فرمود. بر خود لازم می‌دانم از تمامی اساتید بزرگوار، به ویژه اساتید دوره کارشنارسی‌ارشد که در طول سالیان گذشته مرا در تحصیل علم و معرفت و فضائل اخلاقی یاری نموده‌اند تقدیر و تشکر نمایم. به‌ طور ویژه از استاد گرامی و بزرگوار جناب آقای دکتر لکی‌زاده که راهنمایی اینجانب را در انجام این پایان‌نامه تقبل نموده‌اند و با شکیبایی نهال پژوهش را پروراندند، نهایت تشکر و سپاسگزاری را دارم.

**چکیده**

تحقیق و توسعه در زمینه تولید و کشف داروهای جدید پروسه‌ای طولانی و پر‌هزینه است. گزارش‌ها نشان می‌دهد تولید یک داروی جدید از آغاز تا تولید و توزیع آن در بازار مصرف حدود 10 تا 15 سال و با صرف هزینه‌ای بالغ بر 0.8 تا 1.5 میلیارد دلار صورت می‌پذیرد. از این‌رو یافتن یک راهبرد جدید برای کشف دارو امری ضروری به نظر می‌رسد. بازهدف‌گذاری دارو به معنی شناسایی کاربردهای جدید، خارج از محدوده کاربردهای اصلی، برای داروهای تأیید شده است. این کار از طریق شناسایی ویژگی‌های داروها بر پایه محاسبات کامپیوتری و استفاده از قابلیت یادگیری ماشین و بهره‌گیری از کلان‌داده‌ها صورت پذیرفته و باعث صرفه‌جویی در زمان و کاهش هزینه و ریسک فرآیند تحقیق و توسعه می‌شود. همچنین زمینه مناسبی را برای کار بیشتر جهت پیدا‌کردن درمان بیماری‌های نادر و نیز فرصتی را برای تولید داروهای اختصاصی افراد با توجه به ویژگی‌های متابولیسمی و ... هر فرد فراهم می‌کند. پژوهش‌های پیشین، عمدتا نگاهی تک‌بعدی به مساله بازهدف‌گذاری دارو داشته‌اند و تنها با استفاده از یک یا دو منبع داده‌ای به بررسی مساله پرداخته‌اند. مطالعات جدید سعی در برطرف‌کردن این نقیصه داشته و بدین منظور، با جمع‌آوری اطلاعات از منابع مختلف و یکپارچه‌سازی آن‌ها از طریق تشکیل یک شبکه چند‌بعدی، به کشف روابط بین داروها و بیماری‌ها و در نهایت ارائه کاربردهای جدید دارویی پرداخته‌اند. این پژوهش روشی را برای باز‌هدف‌گذاری دارو مبتنی بر ساخت شبکه‌ای یکپارچه از داده‌های متنوع و ناهمگون به روش فاکتورگیری ماتریسی و با لحاظ کردن اطلاعات مربوط به عوارض جانبی داروها به عنوان یکی از معیارهای شباهت داروها ارائه می‌دهد. نتایج ارزیابی‌های مختلف نشان می‌دهد در مقایسه با بهترین روش‌ها، این روش توانسته است فرآیند بازهدف‌گذاری دارو را بر مبنای معیار‌های AUC و AUPR به ترتیب به میزان 1.13 و 14.23 درصد بهبود دهد که بیانگر کارایی رویکرد خلاصه‌سازی و کاهش ابعاد، یکپارچه‌سازی، و استفاده موثر از داده‌های مربوط به عوارض جانبی داروها در روش پیشنهادی می‌باشد.

**کلیدواژه‌ها:** بازهدف‌گذاری دارو، پیش‌بینی پیوند، فاکتورگیری ماتریسی، عوارض جانبی

فهرست مطالب

Table of Conte

[**1-** **فصل اول: مقدمه و تاریخچه** 1](#_Toc50132039)

[1 -1 مقدمه 2](#_Toc50132040)

[1 -2 بیان مسئله و ضرورت آن 3](#_Toc50132041)

[1 -3 اهداف پژوهش 6](#_Toc50132042)

[1 -4 پرسش‌های پژوهش 7](#_Toc50132043)

[1 -5 فرضیات پژوهش 7](#_Toc50132044)

[1 -6 نوآوری پژوهش 7](#_Toc50132044)

[1 -7 مفاهیم پایه زیستی](#_Toc50132045) 8

[1 -7 -1 قاعده اصلی زیست‌شناسی مولکولی](#_Toc50132046)...........................................................................................12

[1 -8 مفاهیم پایه کامپیوتری و محاسباتی... ....19](#_Toc50132047)

[1 -8 -1 پیش‌بینی پیوند..............................................................................................................................20](#_Toc50132048)

[1 -8 -2 سیستم‌های توصیه‌گر....................................................................................................................23](#_Toc50132049)

[1 -9 ساختار کلی پایان‌نامه 28](#_Toc50132050)

[**2-** **فصل دوم: مروری بر مطالعات انجام شده** 30](#_Toc50132051)

[2 -1 مقدمه 31](#_Toc50132052)

[2 -2 انواع روش‌ها از منظر منابع داده‌ای 31](#_Toc50132053)

[2 -2 -1 داده‌های مرتبط با دارو ..32](#_Toc50132054)

[2 -2 -2 داده‌های مرتبط با پروتئین ..34](#_Toc50132055)

[2 -2 -3 داده‌های مرتبط با بیماری ..36](#_Toc50132056)

[2 -2 -4 داده‌های مرتبط با ژن ..37](#_Toc50132057)

[2 -2 -5 مدل‌های ترکیبی ..39](#_Toc50132058)

[2- 3 انواع روش‌ها از منظر رویکرد محاسباتی 40](#_Toc50132059)

[2 -3 -1 یادگیری ماشین ..40](#_Toc50132060)

[2 -3 -2 تحلیل شبکه ..42](#_Toc50132061)

[2 -3 -3 یادگیری عمیق ..45](#_Toc50132062)

[2 -3 -4 متن‌کاوی ..46](#_Toc50132063)

[**3-** **فصل سوم: روش پیشنهادی** 50](#_Toc50132064)

[3 -1 مقدمه 51](#_Toc50132065)

[3 -2 مجموعه داده 52](#_Toc50132066)

[3 -3 محاسبه‌ی ماتریس‌های شباهت 54](#_Toc50132067)

[3 -4 ایجاد ویژگی‌های خلاصه‌شده برای دارو و بیماری.. 55](#_Toc50132068)

1-4-3 [گشت تصادفی با راه‌اندازی مجدد.................................................................................................55](#_Toc50132069)

[2-4-3 چارچوب کاهش ابعاد.......................................................................................................................56](#_Toc50132070)

[3-4-3 یکپارچه‌سازی اطلاعات ماتریس‌ شباهت‌های مختلف...............................................................58](#_Toc50132071)

[3 -5 فرآیند بهینه سازی DCA 59](#_Toc50132072)

[3 -6 پیش بینی از فضای دارویی بر روی فضای بیماری 61](#_Toc50132073)

[**4-** **فصل چهارم: نتایج پژوهش** 64](#_Toc50132074)

[4 -1 مقدمه 65](#_Toc50132075)

[4 -2 روش ارزیابی 65](#_Toc50132075)

[4 -2 -1 اعتبارسنجی متقاطع.....................................................................................................................65](#_Toc50132076)

[4 -2 -2 اعتبارسنجی متقاطع طبقه‌ای.....................................................................................................65](#_Toc50132077)

4 -2 -3  [معیارهای ارزیابی...........................................................................................................................67](#_Toc50132078)

[3-4 نتایج روش پیشنهادی 69](#_Toc50132079)

[4-4 مقایسه روش پیشنهادی با مطالعات پیشین 74](#_Toc50132080)

[4-5 مطالعات موردی 76](#_Toc50132081)

[**5-** **فصل پنجم: جمع‌بندی و کارهای آینده** 78](#_Toc50132082)

[5 -1 خلاصه پژوهش 79](#_Toc50132083)

[5 -2 بحث و نتیجه گیری 80](#_Toc50132084)

[**-6 منابع...................................................................................................................................**81](#_Toc50132085)

فهرست جداول

[جدول ‏1‑1 نمونه های موفق بازهدف‌گذاری دارو 5](#_Toc46353740)

[جدول ‏2‑1 تقسیم‌بندی روش‌های بازهدف‌گذاری دارو 31](#_Toc46353741)

[جدول ‏2‑2 خلاصه‌ای از انواع دیتابیس‌های موجود 32](#_Toc46353742)

[جدول ‏2‑3 مروری بر روش‌های بازهدف‌گذاری بر مبنای رویکردهای مختلف محاسباتی 48](#_Toc46353743)

[جدول ‏3‑1 انواع ویژگی‌ها در مجموعه‌داده 53](#_Toc46353744)

[جدول ‏4‑1 جدول درهم‌ریختگی . 67](#_Toc46353745)

[جدول ‏4‑1 عملکرد روش پیشنهادی بر روی ماتریس‌های شباهت مختلف.](#_Toc46353745) 70

[جدول ‏4‑3 معیارهای ارزیابی محاسبه شده برای روش پیشنهادی 72](#_Toc46353747)

[جدول ‏4‑4 مقایسه عملکرد روش‌ پیشنهادی با عملکرد روش‌های پیشین 74](#_Toc46353748)

[جدول ‏4‑5 تایید موارد بازهدف‌گذاری کشف شده توسط روش پیشنهادی در گزارش‌های مختلف 76](#_Toc46353749)

فهرست شکل‌ها

[شکل ‏1‑1 نمایی از یک سلول...................................................................................................................................................9](#_Toc46353847)

[شکل ‏1‑2 نمایش ساختار چهارگانه پروتئین 11](#_Toc46353848)

[شکل ‏1‑3 جریان اطلاعات در سیستم‌های بیولوژیکی 13](#_Toc46353849)

[شکل ‏1‑4 مراحل بیان ژن.......................................................................................................................................................14](#_Toc46353850)

[شکل ‏1‑5 نمونه‌ای از داده GO................................................................................................................................................15](#_Toc46353851)

[شکل ‏1‑6 نمایش گرافی GO...................................................................................................................................................15](#_Toc46353852)

[شکل ‏1‑7 توسعه سنتی دارو در مقابل بازهدف‌گذاری 19](#_Toc46353853)

[شکل ‏1‑8 شماتیک فاکتور‌گیری ماتریسی 27](#_Toc46353854)

[شکل ‏2‑1 بازهدف‌گذاری مبتنی بر شباهت ساختار شیمیایی دارو 34](#_Toc46353855)

[شکل ‏2‑2 بازهدف‌گذاری دارو با استفاده از ساختار پروتئینی و جایگاه اتصال 35](#_Toc46353856)

[شکل ‏2‑3 بازهدف‌گذاری دارو با استفاده از رابطه بیماری‌ها 36](#_Toc46353857)

[شکل ‏2‑5 نقشه ارتباط میان مفاهیم مختلف بیولوژیکی 39](#_Toc46353858)

[شکل ‏3‑1 نمای کلی روش پیشنهادی 51](file:///C:\Users\Mirashrafi\Downloads\Final_1.docx#_Toc46353859)

[شکل ‏4‑1 اعتبارسنجی متقاطع ۵ دسته‌ای 66](file:///C:\Users\Mirashrafi\Downloads\Final_1.docx#_Toc46353860)

فهرست نمودار‌ها

[نمودار ‏4‑1 میزان تاثیر ویژگی‌های مختلف در دقت فرآیند 70](#_Toc46353807)

[نمودار ‏4‑2 مشخصه عملکرد روش پیشنهادی 71](file:///C:\Users\Mirashrafi\Downloads\Final_1.docx#_Toc46353808)

[نمودار ‏4‑3 میزان دقت-حساسیت روش پیشنهادی 71](#_Toc46353809)

[نمودار ‏4‑4 AUPR به‌ازای مقادیر مختلف ابرپارامتر احتمال بازگشت به گره اولیه 73](#_Toc46353810)

[نمودار ‏4‑5 AUPR برای مقادیر مختلف ابرپارامتر بعد فضای پنهان در فاکتورگیری 73](#_Toc46353811)

[نمودار ‏4‑6 مقایسه عملکرد روش‌ پیشنهادی با روش‌های پیشین 73](#_Toc46353811)

فهرست علائم و اختصارات

|  |  |
| --- | --- |
| Food and Drug Administration | FDA |
| Deoxyribonucleic acid | DNA |
| Ribonucleic acid | RNA |
| messenger RNA | mRNA |
| Gene Ontology | GO |
| Collaborative filtering | CF |
| Bipartite local model | BLM |
| Support vector machine | SVM |
| Collective variational autoencoder | cVAE |
| Diffusion component analysis | DCA |
| Random walk with restart | RWR |
| Singular value decomposition | SVD |
| Broyden–Fletcher–Goldfarb–Shanno algorithm | BFGS |
| Limited-memory BFGS | L-BFGS |
| Kullback-Leibler | KL |
| Receiver operating characteristic curve | ROC |
| Area under the curve | AUC |
| Area under Precision and recall curve | AUPR |

# فصل اول: مقدمه و تاریخچه

فصل اول

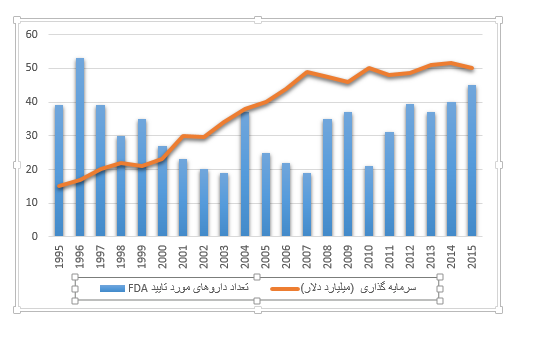
کلیات پژوهش

## مقدمه

تحقیق و توسعه در زمینه تولید و کشف دارو[[1]](#footnote-1)های جدید پروسه‌ای طولانی و پر‌هزینه است. گزارش‌ها نشان می‌دهد تولید یک داروی جدید از مرحله اول تا رساندن آن به بازار مصرف بین 10 تا 15 سال و با صرف هزینه‌ای بالغ بر 0.8 تا 1.5 میلیارد دلار صورت می پذیرد. با این حال علی‌رغم تمام سرمایه‌گذاری‌ها و پیشرفت‌های انجام‌شده در 20 سال اخیر، تعداد داروهای جدیدی که مورد تایید انجمن غذا و دارو آمریکا ([[2]](#footnote-2)FDA) قرار گرفته، چشمگیر نبوده و سالانه نزدیک به 90 درصد داروهای ارزیابی شده توسط این انجمن رد شده و از عرضه‌شان برای مصارف درمانی جلوگیری می‌گردد. طولانی‌بودن پروسه تولید دارو به علت حساسیت و اهمیت تکمیل تمام مراحل آزمایشگاهی می‌باشد. هزینه زیاد این فرآیند نیز باعث شده جهت کاهش ریسک از دست دادن سرمایه، تحقیقات محدود به چارچوب نظریه‌های کلاسیک باقی بماند[1].

از جمله راهکارهای حل این مساله پیدا‌کردن کاربردها و مصارف جدید درمانی برای داروهای موجود از طریق شناسایی ویژگی‌های داروها بر‌پایه محاسبات کامپیوتری و استفاده از قابلیت یادگیری‌ماشین و بهره‌گیری از کلان‌داده‌هاست. از این طریق می‌توان علاوه بر کاهش هزینه و ریسک، از چارچوب‌ها و پیش‌فرض‌های قطعی دانسته شده قبلی خارج شده و به سمت کشف الگوهای جدید پیش رفت. در حال حاضر هزینه‌های ورود یک داروی باز‌هدف‌گذاری‌شده به بازار 300 میلیون دلار تخمین زده می‌شود.

نمودار 1-1 بازگوکننده این واقعیت است که هزینه برای کشف داروهای جدید روند رو‌به‌رشدی داشته و این مسئله لزوم تغییر نگاه و توجه به راهبردهای جدید را برای ما روشن می‌کند[1].



نمودار ‏1‑1 سرمایه‌گذاری در توسعه دارو و تعداد داروهای مورد تأیید[1]

## بیان مسئله و ضرورت آن

بازهدف‌گذاری دارو[[3]](#footnote-3) به معنی شناسایی کاربردهای جدید، خارج از محدوده کاربرد اصلی برای دارو‌های از قبل تأیید‌شده است. این استراتژی نسبت به کشف داروی جدید برای درمان یک بیماری خاص مزایای مختلفی دارد. اول و شاید مهم‌تر از همه ریسک شکست پایین‌تر این استراتژی است. زیرا در صورت تکمیل مراحل اولیه آزمایش در مدل‌های بالینی و آزمایش‌های انسانی، داروی مدنظر به اندازه کافی ایمن بوده، و احتمالاً از نظر ایمنی در آزمایش‌های بعدی نیاز به اثبات نخواهد داشت. دوم اینکه، با این روش می‌توان مدت زمان تولید دارو را کاهش داد، زیرا بیشتر آزمایش های بالینی، ارزیابی ایمنی و در برخی موارد، تدوین فرموالسیون از قبل تکمیل شده است. و سوم، سرمایه‌گذاری کمتری مورد نیاز است. گرچه بسته به مرحله و روند توسعه، این مورد می‌تواند در موارد مختلف بسیار متفاوت باشد. این مزایا می‌توانند منجر به بازگشت کم‌خطرتر و سریع‌تر سرمایه‌گذاری در تولید داروهای بازهدف‌گذاری شده شوند[2].

همچنین در مواردی سرعت شیوع یک بیماری جدی مثل کرونا بسیار بالاست و نمی‌توان تا زمان ساخت داروی جدید منتظر ماند. یا در سناریوی دیگری یک بیماری نادر ولی خطرناک شناسایی می‌شود و به دلیل نادر‌بودن بیماری و تعداد کم افراد مبتلا شده، صرف هزینه چند میلیارد دلاری برای شرکت‌های داروسازی توجیه اقتصادی نخواهد داشت لذا پزشکان و محققان باید به بررسی اثربخشی داروهای موجود برای درمان این بیماری بپردازند.

رشد انفجاری انواع داده‌های زیستی در ابعاد بزرگ و با دقت بالا موجب شکل‌گیری زمینه تحقیقاتی با نام داروشناسی محاسباتی[[4]](#footnote-4) و ایجاد فرصتی برای تحلیل و آنالیز سیستماتیک داده‌های مختلف شده است و تحلیل این داده‌ها جهت بهبود و کاهش ریسک فرآیند توسعه داروها می‌تواند مفید واقع گردد. یکپارچه‌سازی منابع مختلف داده‌ای باعث انباشت شواهد بیشتر جهت پشتیبانی تصمیم و ارزیابی و اعتبارسنجی بهتر نتایج حاصل از مدل‌سازی نیز می‌شود. با استفاده داده‌های ذکر‌شده و روش‌های یادگیری عمیق می‌توان به بینشی جامع و همه‌جانبه از مسئله دست پیدا کرد و پیشنهادهای جدید و موثر‌تری ارائه داد.

سیلدانفیل[[5]](#footnote-5) با نام تجاری ویاگرا[[6]](#footnote-6) نمونه‌ای موفق از بازهدف‌گذاری دارو می‌باشد. سیلدنافیل ابتدا در اواخر دهه 80 میلادی برای درمان بیماری درد قفسه سینه[[7]](#footnote-7) تجویز شد. درد قفسه سینه هنگامی رخ می‌دهد که خون‌رسانی به ماهیچه‌های قلب محدود گردد. این مساله خود بدان علت روی می‌دهد که شریان‌های خون‌رسانی به قلب سخت و باریک می‌شوند. فرضیه مؤثر این بود که این دارو می‌تواند جریان خون را در ناحیه قلبی-عروقی بیمار افزایش دهد. متأسفانه دارو در طول آزمایشات بالینی فاقد کارآیی لازم بر روی این بیماری شناخته‌شد و روند توسعه آن متوقف گردید تا اینکه بیماران اثرات‌جانبی غیر‌معمول از جمله نعوظ طولانی‌مدت را گزارش کردند. بنابراین محققان تصمیم گرفتند تا دارو را برای این کاربرد جدید مورد بررسی قرار دهند و یک مطالعه گسترده در‌بردارنده 3700 مورد، اثر بخشی مولکول را تأیید کرد و دارو برای اختلال نعوظ بازهدف‌گذاری شد. ویاگرا تبدیل به یک داروی بسیار مؤثر شد و زندگی میلیون‌ها انسان را تحت‌تأثیر قرار داد. این دارو در طول اولین سال‌های عرضه خود دارای فروش سالیانه بیش از 1.5 میلیارد دلار گردید. در حال حاضر این دارو برای پرفشاری شریان ریوی نیز کاربرد پیدا کرده است. در جدول 1-1 نمونه‌های موفق دیگری نیز آورده شده است[2].

جدول ‏1‑1 نمونه های موفق بازهدف‌گذاری دارو[2]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **نام دارو** | **کاربرد اصلی** | **کاربرد جدید** | **سال تصویب** | **توضیحات** |
| **ماینوکسیدیل[[8]](#footnote-8)** | فشار خون | ریزش مو | 1988 | فروش جهانی 860 میلیون‌دلاری در سال 2016 |
| **سیلدنافیل** | درد قفسه سینه | اختلال نعوظ | 1998 | فروش آن در سال 2012 معادل با 2.05 میلیارد‌دلار بود. |
| **آتوموکستین[[9]](#footnote-9)** | بیماری پارکینسون | بیش فعالی | 2002 | دارو با نام تجاری استراترا، فروش 855 میلیون‌دلاری را در سال 2016 ثبت کرد. |
| **ریتوکسی‌مب[[10]](#footnote-10)** | انواع سرطان | روماتیسم مفصلی | 2006 | فروش سالانه دارو در سال 2015، 7 میلیارد دلار بوده است. |
| **توپیرامات[[11]](#footnote-11)** | تشنج | چاقی | 2012 | کوسیمیا حاوی توپیرامات در ترکیب با فنترمین است. |
| **فینگولیمود[[12]](#footnote-12)** | رد پیوند | ام‌‌اس | 2010 | فروش سالانه دارو 3.1 میلیارد دلار در سال 2017 بوده است. |
| **دلوکستین[[13]](#footnote-13)** | افسردگی | بی اختیاری ادرار | 2004 | دارو توسط آژانس دارویی اروپا برای بی‌اختیاری ادراری و برای درمان افسردگی و درد مزمن در ایالات متحده مورد تأیید قرار گرفته است. |
| **رالوکسیفن[[14]](#footnote-14)** | پوکی استخوان | سرطان سینه | 2007 | توسط FDA برای سرطان سینه تهاجمی مورد تأیید قرار گرفته است. فروش 237 میلیون‌دلاری در سال 2015 داشته است. |
| **داپوکستین[[15]](#footnote-15)** | تخفیف درد و افسردگی | انزال زودرس | 2012 | در انگلستان و تعدادی از کشورهای اروپایی مورد تأیید قرار گرفته است و در ایالات‌متحده در حال اخذ تأییدیه است. پیش بینی می‌شود که فروش 750 میلیون دلار را داشته باشد. |

## اهداف پژوهش

اهداف این تحقیق به صورت مختصر موارد زیر می باشد:

* یافتن راهی برای اندازه‌گیری شباهت داروها و بیماری‌ها
* یافتن نحوه یکپارچه‌سازی داده های نامتجانس زیستی
* ارائه مدلی جهت پیدا‌کردن ارتباطات از دست‌رفته و امتیاز‌دهی به آن‌ها
* یافتن موارد مصرف جدید داروهای کشف شده قبلی برای درمان بیماری‌ها با استفاده از اطلاعات عوارض‌جانبی داروها

## پرسش‌های پژوهش

این پژوهش در پی یافتن پاسخ پرسش‌های زیر می‌باشد:

سوال اصلی:

* چگونه می توان برای داروهای موجود، کاربرد درمانی جدید پیدا کرد؟

سوال فرعی:

* استفاده از اطلاعات عوارض‌جانبی دارو چه نقشی در فرآیند باز‌هدف‌گذاری می‌تواند داشته باشد؟

## فرضیات پژوهش

فرضیات این پژوهش بدین‌قرار می‌باشند:

* داروهایی با ساختار مشابه، عملکرد نزدیک به هم دارند.
* بیماری‌هایی با نشانه‌های مشابه، ارگان‌های یکسانی از بدن را درگیر می‌کنند.
* جمع‌آوری و یکپارچه‌سازی داده‌های متنوع زیستی منجر به بهبود نتایج داروهای پیشنهاد شده می‌گردد.

## نوآوری پژوهش

* نوآوری بطور خاص ارائه چارچوب جدید برای ترکیب منابع داده ای.
* استفاده از چندین پایگاه داده شامل عوارض جانبی داروها در بهبود فرآیند بازهدف‌گذاری.

## مفاهیم پایه زیستی

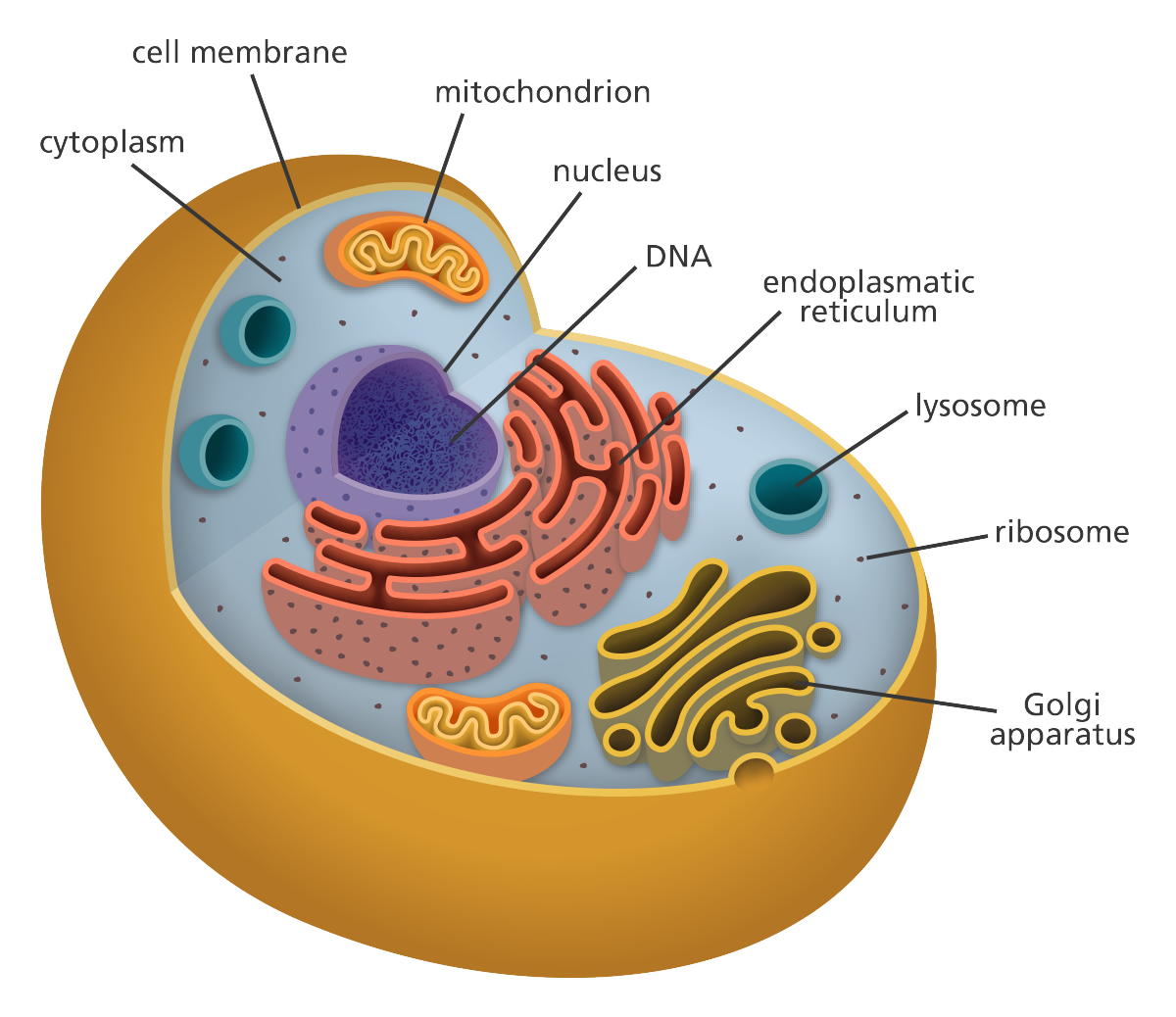
زیست‌شناسی[[16]](#footnote-16) علم شناخت حیات است. زیست‌شناسی مولکولی و سلولی[[17]](#footnote-17)، مطالعه زیست‌شناسی در سطح مولکول‌هاست. زیست‌شناسی مولکولی، علم استنباط برهم‌کنش‌های مولکولیِ فعالیت‌های بیولوژیکی در بین سیستم‌های مختلف درون‌سلولی است که شامل ارتباطات بین DNA[[18]](#footnote-18)، [[19]](#footnote-19)RNA، پروتئین و بیوسنتز و برهم‌کنش[[20]](#footnote-20) ملکول‌های زیستی با یکدیگر در سلول می‌باشد. به‌علاوه، چگونگی تنظیم این برهم‌کنش‌ها نیز مورد بررسی قرار می‌گیرد[3].

بسیاری از داده‌های زیست‌شناسی مولکولی کمی می‌باشند و به تازگی کارهای زیادی در رابطه با ارتباط زیست‌شناسی مولکولی با علم کامپیوتر در بیوانفورماتیک[[21]](#footnote-21) و آمار زیستی انجام گرفته‌است.

بیوانفورماتیک دانش استفاده از علوم کامپیوتر و آمار و احتمالات در شاخه زیست‌شناسی مولکولی است. بیوانفورماتیک یک دانش بین‌رشته‌ای است که شامل روش‌ها و نرم‌افزارهایی برای فهم اطلاعات زیستی است. بیوانفورماتیک همچنین در شبیه‌سازی و مدل‌سازی DNA، RNA، پروتئین و پیوندهای زیست مولکولی کمک‌کننده است[4].

به منظور درک بهتر از چگونگی تغییر فعالیت‌های سلولی نرمال در بیماری‌های مختلف، باید اطلاعات زیستی ترکیب شوند تا تصویری جامع از این فعالیت‌ها شکل بگیرد. از این‌رو بیوانفورماتیک در زمینه تجزیه و تحلیل و تفسیر انواع مختلفی از داده‌ها به کمک ما می‌آید. این داده‌ها شامل توالی نوکلئوتیدها و اسیدهای آمینه و ساختارهای پروتئینی است.

سلول[[22]](#footnote-22) یا یاخته واحد اساسی ساختاری، عملکردی و بیولوژیکی همه‌ی موجودات شناخته‌شده و کوچک‌ترین واحد زنده است. همه سلول‌ها از سه بخش اصلی غشای پلاسمایی، سیتوپلاسم و هسته تشکیل شده‌اند[5].



شکل ‏1‑1 نمایی از یک سلول[6]

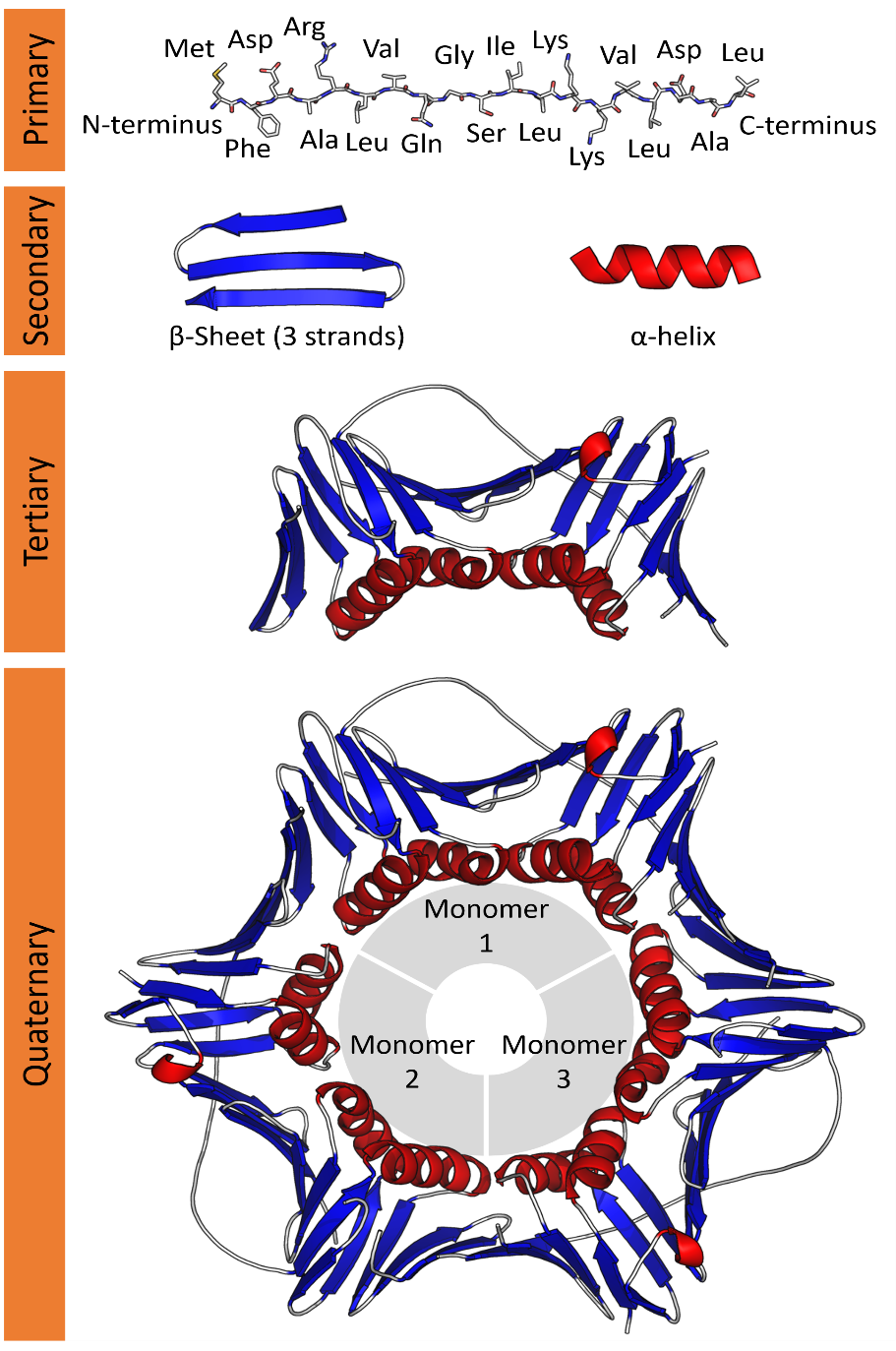
پروتئین‌ها[[23]](#footnote-23) به عنوان بزرگ‌ترین درشت‌مولکول‌های موجود در بدن، از جمله مهم‌ترین اجزای بدن بوده که نقش‌های بسیار حیاتی و مهمی را در بدن ایفا می‌کنند. رفتار سلولی و تمام فعالیت‌هایی که در سلول انجام می‌شود بر عهده پروتئین‌ها است. همه پروتئین‌ها با هم برهم‌کنش دارند و تقریباً می‌توان گفت که همه پروتئین‌ها اثر خود را با همکاری پروتئین‌های دیگر در سلول اعمال می‌کنند و هیچ پروتئینی نیست که در یاخته به تنهایی عمل کند.

اسیدهای‌آمینه[[24]](#footnote-24) الفبای پروتئین را تشکیل می‌دهند. به طور کلی پروتئین‌ها از تعدادی اسید‌آمینه تشکیل شده‌اند که در یک رشته به صورت متوالی قرار می‌گیرند. ترتیب و ماهیت اسیدهای آمینه ویژگی‌های هر پروتئین را مشخص می‌کند. ترتیب این اسیدهای آمینه در پروتئین توسط ژن[[25]](#footnote-25) مشخص می‌شود[4].

در سطح مولکولی، کلیه مکانیسم‌های زیستی سلول‌ها توسط پروتئین‌ها انجام می‌شود. پروتئین‌ها در ارتباط با یکدیگر به‌طور دقیق و بسیار کنترل‌شده‌ای وظایف خود را انجام می‌دهند. اساس مولکولی اغلب بیماری‌ها، بروز نقص یا تداخل در کارکرد عادی پروتئین‌هایی که از طریق آن‌ها مکانیسم‌های زیستی سلولی انجام می‌شود، تعریف می‌گردد. اگر کار پروتئین‌ها از حالت طبیعی خارج شود، امکان بروز بیماری‌های مختلف وجود خواهد‌داشت.

زمانی که دو یا بیش از دو پروتئین در نتیجه نیروهای الکترواستاتیک یا رویدادهای شیمیایی در تماس فیزیکی با یکدیگر قرار می‌گیرند پیوند پروتئین–پروتئین[[26]](#footnote-26) رخ می‌دهد. این گونه پیوندها غالباً عملکردهای زیستی را رقم می‌زنند. بسیاری از فرآیندهای مولکولی مهم در سلول مانند تکثیر DNA و حتی سوخت و ساز خود پروتئین‌ها توسط کمپلکس‌های بزرگ مولکولی انجام می‌شوند که این کمپلکس‌ها از تعداد زیادی پروتئین‌های سازمان‌دهی شده که در پیوند با یک‌دیگر هستند، ساخته شده‌اند. ناگفته پیداست که اغتشاش در شبکه برهم‌کنش پروتئین‌ها می‌تواند منشأ اختلالات زیستی و بیماری‌های مختلف نظیر آلزایمر و سرطان باشد.

عملکرد هر پروتئین وابسته به نوع ساختارش[[27]](#footnote-27) است. به لحاظ ظاهری هر پروتئین دارای چهار نوع ساختار است. ساختار اول پروتئین مشخص کننده تعداد و نوع ترتیب قرار‌گرفتن آمینو‌اسیدها در زنجیر پلی‌پپتیدی است. ساختار دوم یک پروتئین از الگوهای سه‌‌بعدی منظم و تکراری، مانند هلیکس‌‌ها، لوپ‌‌ها یا صفحات بتا تشکیل می‌‌شود. ساختار سوم پروتئین به ما نشان می‌‌دهد که چطور این ساختارهای منظم در فضا در جوار هم چیده شده‌‌اند تا یک پروتئین یا زیرواحدی از یک پروتئین ساخته شود. در نهایت، ساختار چهارم نشان می‌‌دهد که چگونه زنجیره‌‌های آمینواسیدهای مختلف این زیرواحدها برای تشکیل دادن یک کمپلکس پروتئینی فعال آرایش می‌‌یابند. در بیشتر مطالعات از ساختار سوم آن که نمایش حالت سه‌بعدی‌ای است که پروتئین بعد از پیچش به خود می‌گیرد استفاده می‌شود و دومین پروتئین[[28]](#footnote-28) نام دارد[4].



شکل ‏1‑2 نمایش ساختار چهارگانه پروتئین[7]

دزوکسی ریبونوکلئیک اسید یا بطور خلاصه DNA نوعی اسید‌نوکلئیک است و دارای دستورالعمل‌های ژنتیکی‌ای است که برای کارکرد و توسعه‌ی بیولوژیکی موجودات زنده و ویروس‌ها مورد استفاده قرار می‌گیرد. در حقیقت اطلاعات لازم برای ایجاد یک موجود در مولکول DNA ذخیره شده است. نقش اصلی مولکول DNA ذخیره‌سازی طولانی‌مدت اطلاعات ژنتیکی و دستوری است. پیام‌های ژنتیکی موجود در مولکول DNA در نهایت برای مواردی چون ساخت پروتئین و مولکول‌های RNA در یاخته، مورد استفاده قرار می‌گیرند.

ریبونوکلئیک اسید یا بطور خلاصه RNA همراه با DNA و پروتئین، سه مولکول درشت اصلی می‌باشد که برای همه‌ی گونه‌های شناخته‌شده‌ی زیستی، ضروری است. RNA هم مانند DNA دارای زنجیره‌ی بلندی می‌باشد که شامل اجزای سازنده‌ای به نام نوکلئوتیدها است.

RNA وظیفه حمل اطلاعات از DNA به سیتوپلاسم را دارد. به این نوع RNA که اطلاعات را از DNA به ریبوزوم‌ها حمل می‌کند، RNA پیام‌بر[[29]](#footnote-29) می‌گویند. نوعی دیگر RNA ناقل[[30]](#footnote-30) است که اسیدهای‌آمینه را به ریبوزوم منتقل می‌کند تا ریبوزوم، اسیدهای‌آمینه را بر اساس اطلاعات موجود در mRNA کنار یکدیگر ردیف کند[5].

تفاوت‌های DNA و RNA :

• DNA در ذخیره و RNA در انتقال اطلاعات وراثتی نقش دارد.

• مولکول DNA دو رشته‌ای ولی مولکول RNA تک‌رشته‌ای است.

• DNA برعکس RNA از هسته سلول خارج نمی‌شود. بطور دقیق‌تر جایگاه DNA در هسته بوده اما، RNA هم در هسته یافت می‌شود و هم در سیتوپلاسم.

توالی نوکلئیک اسید[[31]](#footnote-31) به بررسی برای تعیین دقیق ترتیب و آرایش جای گرفتن نوکلئوتیدها در DNA یا RNA گفته می‌شود. توالی‌یابی DNA شامل هر روش یا فناوری‌ای است که ترتیب چهار باز آدنین، گوانین، سیتوزین و تیمین را در یک رشته از DNA تعیین می‌کند.

اطلاعات موجود در مولکول DNA در نهایت برای مواردی چون ساخت پروتئین و مولکول‌های RNA در یاخته، مورد استفاده قرار می‌گیرد. بخش‌هایی از DNA که اطلاعات ژنتیکی یک صفت را با خود حمل می‌کنند ژن نامیده می‌شوند.

### قاعده اصلی زیست‌شناسی مولکولی[[32]](#footnote-32)

همان‌طور که گفته شد نقش پروتئین‌‌ها برای ساخت سلول و کارکرد آن حیاتی است و رفتار سلول و تمام فعالیت‌هایی سلول از طریق آن انجام می‌گیرد. قاعده اصلی زیست‌شناسی مولکولی توضیحی برای جریان اطلاعات ژنتیکی در یک سیستم بیولوژیکی است به این صورت تعریف می‌شود که DNA باعث ساخت RNA شده و RNA پروتئین را می‌سازد. در حقیقت این قاعده توضیحی برای درک انتقال توالی اطلاعات بین DNA، RNA و پروتئین است.

برای ساخت پروتئین در یاخته مراحل زیر باید انجام شود:

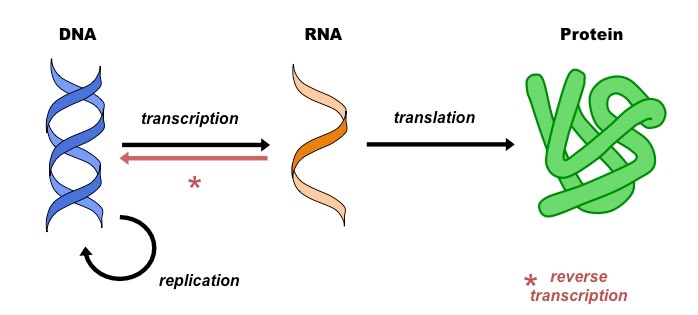
• DNA باید رونویسی[[33]](#footnote-33) شود.

• mRNA از هسته خارج می‌شود و وارد سیتوپلاسم می‌شود.

• در ریبوزوم، mRNA به پروتئین ترجمه[[34]](#footnote-34) می‌شود.

فرآیند ساخته‌شدن mRNA از روی DNA، رونویسی نامیده می‌شود. این فرآیند یک رشته RNA پیام‌رسان تولید می‌کند.

در فرآیند ترجمه از روی رشته RNA پیا‌م‌رسان بالغ، پروتئین تولید می‌شود. فرآیند ترجمه به کمک پروتئینی بزرگ و پیچیده به نام ریبوزوم انجام می‌شود. طی این فرآیند رشته RNA وارد ریبوزوم می‌شود و ریبوزوم رمز‌های رشته را سه حرف به سه حرف می‌خواند و با اضافه‌کردن آمینواسید مربوطه یک دنباله از آمینواسید‌ها می‌سازد که پروتئین را تشکیل می‌دهد.

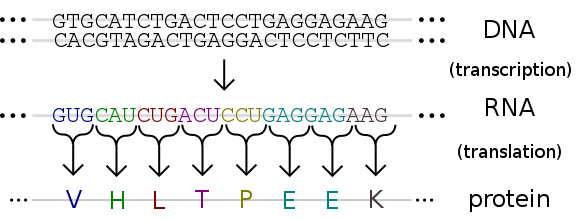


شکل ‏1‑3 جریان اطلاعات در سیستم‌های بیولوژیکی[8]

ژن دنباله‌ای از نوکلوئوتیدهاست که در‌برگیرنده اطلاعات لازم جهت تولید مولکول‌های RNA و یا پروتئین‌های لازم‌ برای سلول هستند. هر ژن در بخشی از DNA سلول‌ وجود دارد. درون سلول‌ها و طی فرآیند رونویسی، ژن‌ها به مولکول‌های RNA تبدیل می‌شوند که یا به شکل مستقیم در سلول استفاده می‌شوند و یا دربرگیرنده اطلاعاتی جهت تولید پروتئین هستند و طی فرآیند ترجمه، پروتئین مربوط به آن‌ها ساخته می‌شود[3].

ژن‌ها تمامی صفات سلول‌ را کنترل می‌کنند و عملکرد سلول‌ها به کمک ژن‌ها و پروتئین‌های ساخته شده از روی آن‌ها تعیین می‌شود. این ژن‌ها از پدر و مادر به ارث می‌رسند و ممکن است به مرور در اثر فرآیند تقسیم سلولی و یا در پیوند سلول‌ با محیط بیرونی دچار تغییر شوند. بیان ژن[[35]](#footnote-35) فرآیندی است که در آن اطلاعات درون ژن استفاده می‌شود تا یک محصول کاربردی از آن بدست آید. ژن‌ها باید به گونه‌ای که تنها هنگامی که سلول به آن‌ها نیاز دارد بیان شوند. یک سلول میزان بیان یک ژن را بر اساس شرایط محیطی (مانند دما، مواد اولیه موجود و ...) و شرایط داخلی (مانند متابولیسم، چرخه سلولی و ...)‌ و از همه مهم‌تر کاربردش در یک ارگانیسم پیچیده تعیین می‌کند. به عنوان مثال تمامی سلول‌های بدن ما، DNA و به تبع آن ژن‌های یکسانی دارند، اما تفاوت در میزان بیان ژن‌ها به جهت عملکرد متفاوت باعث می‌شود که بطور مثال سلول‌های چشم ما سلول‌هایی حساس به نور باشند و در طرف مقابل سلول‌های روی پوست، سلول‌هایی مقاوم و با عملکرد متفاوت باشند.

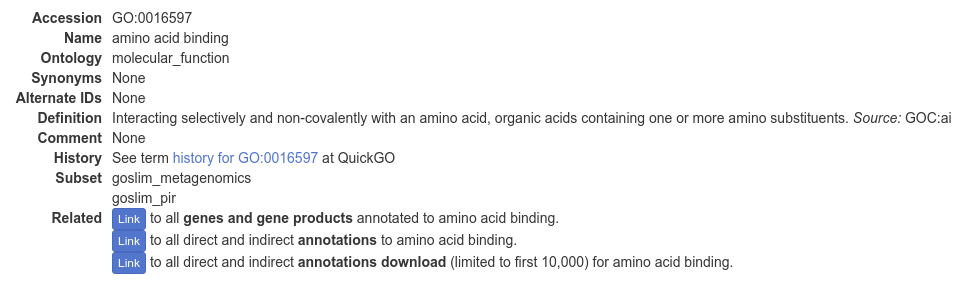
بیان ژن‌ها شامل دو مرحله می‌باشد. ابتدا از روی ژن‌ها، RNAها رونویسی می‌شوند و سپس به پروتئین ترجمه می‌شوند.



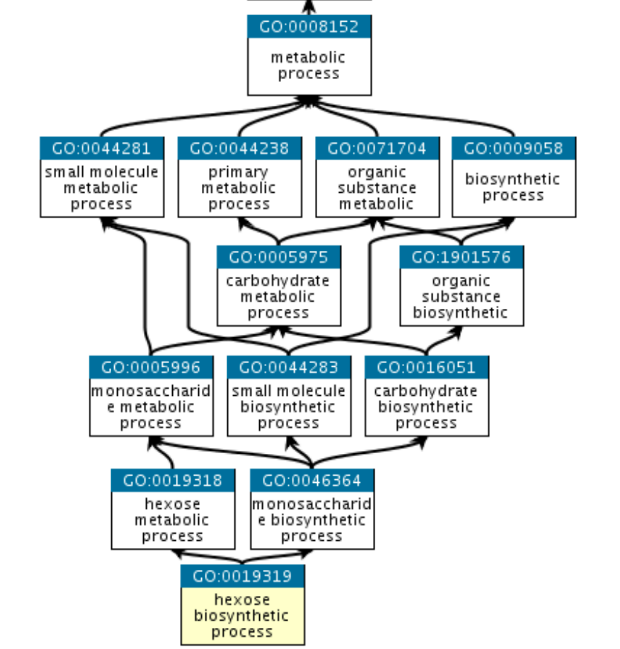
شکل ‏1‑4 مراحل بیان ژن[9]

هستان‌شناسی ژن[[36]](#footnote-36) یا به صورت خلاصه GO یک پروژه گروهی در بیوانفورماتیک برای استفاده از یک نمایش و هستان‌شناسی واحد برای بیان ویژگی‌های ژن‌ها و محصولات ژنتیکی است. این پروژه اطلاعاتی ساختاریافته و قابل پردازش از عملکرد ژن‌ها و محصولات ژنتیکی ارائه می‌دهد. امروزه داده‌های پروژه GO به طور گسترده در علوم زیستی استفاده می‌شوند[10].

در نمایش گرافی اصطلاحات GO به صورت گره‌هایی در یک شبکه به هم متصل هستند که این اتصالات گره‌های پدر و پسر را مشخص می‌کنند. بنابراین ساختار هستان‌شناسی GO یک گراف جهت‌دار بی‌دور بوده که در آن هر راس گراف یک اصطلاح GO و هر یال گراف یک رابطه بین دو اصطلاح GO از یک یا دو دامنه متفاوت است.



شکل ‏1‑5 نمونه‌ای از داده GO[11]



شکل ‏1‑6 نمایش گرافی GO[12]

در زیست‌شناسی هر نوع حالت غیر‌عادی در موجود زنده که عملکرد بدن را مختل کند، بیماری[[37]](#footnote-37) تعریف می‌شود. در حقیقت بیماری به ناهنجاری‌ای در بدن یا روان گفته می‌شود که به علت ناراحتی، اختلال عملکرد یا تنش در بیمار یا سایر افراد مرتبط با او ایجاد می‌گردد. اساس مولکولی اغلب بیماری‌ها بروز نقص یا تداخل در عملکرد طبیعی پروتئین‌هایی است که مکانیسم‌های زیستی سلول را انجام می‌دهند. هر بیماری با شماری علائم و نشانه‌های ویژه شناخته و آشکار می‌شود.

دارو[[38]](#footnote-38) به ماده‌ای گفته می‌شود که با اثر بر گیرنده‌ای[[39]](#footnote-39) خاص در داخل، خارج یا دیواره سلول باعث شروع یا مهار عملکردی خاص می‌گردد و قدرت اثر دارو نیز با میزان و تعداد این پیوند نسبت مستقیم دارد. بطور خلاصه دارو به مواردی اطلاق می‌شود که جهت تشخیص، درمان، بهبودی، تشخیص و یا پیشگیری از یک بیماری بکار می‌رود. داروها ممکن است به صورت خوراکی (قرص و شربت)، مالیدنی (پماد و قطره)، استنشاقی (از راه تنفس) یا تزریقی (آمپول) تهیه شده باشند. جذب دارو زمانی انجام می‌گیرد که از طریق کبد وارد جریان خون شود.

داروشناسی[[40]](#footnote-40) شاخه‌ای از داروسازی و زیست‌شناسی بوده که به مطالعه واکنش متقابل داروها (یا مواد شیمیایی) بر موجودات زنده می‌پردازد. داروشناسی شامل دو بخش فارماکودینامیک[[41]](#footnote-41) و فارماکوکینتیک[[42]](#footnote-42) است[13].

فارماکودینامیک اثر دارو بر بدن را توضیح می‌دهد. فعالیت دلخواه داروها به این صورت فهرست می‌شود:

* شکستن غشای سلولی
* واکنش شیمیایی
* بر همکنش با آنزیم‌های پروتئینی
* بر همکنش باپروتئین‌های ساختاری
* بر همکنش باپروتئین‌های حامل
* بر همکنش با کانال‌های یونی
* پیوند لیگاندی با گیرنده ها (هورمونی، انتقال دهنده عصبی)

واکنش‌های شیمیایی بدین صورت است که یک دارو با گیرنده‌های خود در غشای سلول یا داخل سلول میزبان به صورت مکمل بوده و در آن قرار می‌گیرد و پس از جفت شدن مولکول دارو و گیرنده، پیغامی به سلول داده می‌شود که بسته به ماهیت مولکول دارو، ممکن است عملکرد سلول را تغییر داده یا ثابت نگه دارد. همچنین می‌تواند باعث ترشح هرمون‌ها و آنزیم‌ها شود.

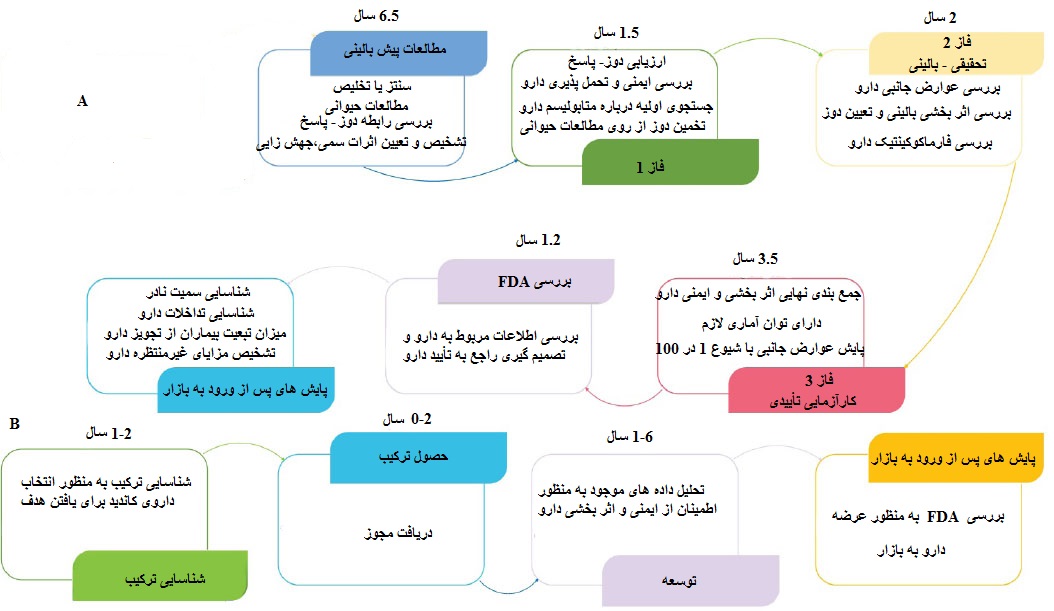
فارماکوکینتیک، تأثیر بدن بر دارو را بررسی می‌کند و دارای چهار مرحله جذب، توزیع، متابولیزم و دفع می‌باشد. مرحله جذب، فرآیند ورود و گذشتن دارو از مسیر تجویز شده به درون جریان خون و پلاسما است و میزان آن بسته به عوامل مختلفی است. توزیع، نفوذ و انتشار مواد دارویی به داخل بافت‌ها و مایع بین‌سلولی است. متابولیزم، شکسته‌شدن مواد دارویی به اجزای عمل‌کننده است. دفع نیز فرآیند پاکسازی و حذف باقی‌مانده‌ی مواد دارویی از بدن و خارج نمودن آن است.

هر دارو بعنوان یک ترکیب شیمیایی دارای ساختاری شیمیایی[[43]](#footnote-43) است. ساختار شیمیایی شامل هندسه‌ی مولکولی، ساختار الکترونی و ساختار کریستالی مولکول است. هندسه مولکولی به‌آرایش فضایی اتم‌ها در یک مولکول و نحوه‌ی چیدمان پیوندهای شیمیایی اتم‌ها باهم اشاره می‌کند. ساختار الکترونی توصیف اشغال اوربیتال‌های مولکولی یک مولکول توسط الکترون‌ها است. ساختار کریستالی به ساختار و جهت‌گیری اتم‌ها، یون‌ها، یا مولکول‌های داخل بلور جامد یا مایع گفته می‌شود.

طراحی داروها برای تحت تأثیر قراردادن یک هدف بیولوژیکی خاص تقریبا غیرممکن است. به این معنی که داروها علاوه بر داشتن اثرات شناخته شده[[44]](#footnote-44) که برای آن هدف خاص و از پیش تعیین‌شده تولید شده‌اند، می‌توانند اثرات ناخواسته‌ای[[45]](#footnote-45) نیز از خود در بدن به جاگذارند. اثرات ناخواسته داروها خود منجر به ایجاد عوارض‌جانبی[[46]](#footnote-46) پیش‌بینی نشده برای مصرف‌کننده می‌شود. با این حال گرچه فعالیت‌های غیرقابل پیش‌بینی هدف‌های اضافی معمولاً ناخواسته و مضر است اما گاهی اوقات می توانند منجر به کشف موارد مصرف دارویی جدید شود. برای مثال سیلدنافیل که برای درمان نوعی بیماری درد قفسه سینه توسعه یافته بود، اثر جانبی آن در داوطلب‌های انسانی منجر به تغییر در حوزه درمانی آن گردید[14].

بطور کلی توسعه دارو بعد از انجام مطالعات اولیه تا گرفتن مجوز قانونی و عرضه آن به بازار شامل چند فاز است. فاز اولیه موسوم به مطالعات میکرودوز انسانی، دارو به 15-10 فرد سالم تجویز می‌شود تا داده‌های اولیه جذب، توزیع، متابولیزم و دفع انسانی جمع‌آوری شود. در فاز بعدی ایمنی دارو سنجیده شده و اثرات فیزیولوژیک دارو روی بدن مورد بررسی قرار می‌گیرد. در این فاز دارو را روی 100-20 فرد مورد ارزیابی قرار می‌دهند و مدت آن معمولاً از چند ماه تا حدود یک سال و نیم طول می‌کشد. افراد مورد ارزیابی در این فاز معمولاً داوطلب های سالم هستند. گرچه اگر یک بیماری تهدیدی برای زندگی به شمار رود، ممکن است از بیماران واقعی نیز استفاده شود. در فاز بعدی اثر‌بخشی دارو را مورد ارزیابی قرار می‌دهند و اثرات جانبی و دیگر جنبه‌های ایمنی، تعیین‌شده و دوره درمان را در چند صد بیمار مبتلا بررسی می‌کند. این مطالعات معمولاً یک اطمینان اولیه از اثر‌بخش بودن دارو در برابر بیماری ایجاد می‌کنند و زمان آن به طور معمول از 1 تا 3 سال طول می‌کشد. در نهایت یک آزمایش گسترده‌تر با چند هزار بیمار انجام می‌گیرد که اثر‌بخشی نهایی دارو را تعیین می‌کند و واکنش‌های جانبی حاصل از استفاده طولانی مدت آن را مورد پایش قرار می‌دهد. همچنین دارو را با داروهای مشابه که پیشتر در بازار موجود بوده‌اند مورد مقایسه قرار می‌گیرد. کنترل‌های علمی مناسبی در نظر گرفته می‌شود تا نتیجه‌گیری‌های معنادار به لحاظ آماری راجع به اثربخشی دارو حاصل شود. این فاز معمولاً برای تکمیل نیاز به 6-2 سال دارد.

همان طور که در شکل 1-7 مشاهده می‌کنید، راهبردهای قدیمی توسعه دارو معمولاً شامل کشف، مطالعات پیش بالینی، بررسی ایمنی، پژوهش های بالینی و بررسی سازمان غذا و دارو و پایش ایمنی پس از ورود به بازار بوده است. در مقابل بازهدف‌گذاری تنها در بردارنده چهار مرحله شناسایی ترکیب، حصول ترکیب، توسعه و پایش ایمنی پس از عرضه به بازار است[1].



شکل ‏1‑7 توسعه سنتی دارو در مقابل بازهدف‌گذاری[1]

## مفاهیم پایه کامپیوتری و محاسباتی

مساله پیدا‌کردن مصرف جدید درمانی برای داروها را می‌توان در قالب مساله کلی پیش‌بینی‌پیوند[[47]](#footnote-47) در شبکه و سیستم‌های توصیه‌گر[[48]](#footnote-48) مدل‌سازی کرد. در سیستم‌هاي توصیه‌گر، دو مجموعه از کاربران و اقلام وجود دارد که ارتباطات بین اعضای این دو مجموعه را می‌توان در قالب یک گراف دو‌بخشی مدل کرد. هر کاربر یک مجموعه از اقلام را توسط مقادیري امتیازدهی می‌کند و پیشنهادهای سیستم براساس پیش‌بینی امتیاز کاربر برای کالاهایی است که کاربر آن‌ را امتیازدهی نکرده است. به همین شکل در شبکه‌های اجتماعی نیز در هر لحظه فرد با مجموعه‌ای از افراد پیوند دارد و عدم وجود پیوند بین دو فرد لزوماً به معنی مرتبط‌نبودن آن دو فرد نیست. پس سیستم توصیه‌گر مثلاً بنا بر تعداد دوستان مشترک، ایجاد پیوند به افراد را پیشنهاد می‌دهد. هر چه تعداد دوستان مشترک بیشتری باشد و در حلقه‌های اجتماعی یکسانی فعالیت کنند، با احتمال بیشتری می‌توانند با یکدیگر پیوند داشته باشند. پیوندها در شبکه‌های زیستی[[49]](#footnote-49) و به طور خاص شبکه دارو-بیماری را نیز می‌توان بر همین پایه مدل‌سازی کرد.

سیستم‌های توصیه‌گر نرم‌افزارهای مؤثری هستند که برای حل مشکل سربار[[50]](#footnote-50) بر روی اینترنت استفاده می‌شوند. این سیستم‌ها می‌توانند به طور خودکار اقلام را به کاربران هدف بر اساس مشاهده اطلاعات ترجیحات کاربر، رفتار خرید، رفتار ارزیابی و ... توصیه کنند. این کار به عوامل متعددی نظیر امتیاز کاربر به مجموعه‌ای از اقلام بر اساس سطح رضایت‌مندی، موردپسندها و غیرپسندهای آن‌ها، سن، جنسیت، شغل، منطقه یا محل، جامعه و ... کاربران بستگی دارد. برخی از وب‌سایت‌هایی که از یک موتور توصیه برای فیلتر‌کردن انتخاب‌ها استفاده می‌کنند، آمازون، نت‌فلیکس و ... هستند. در این وب‌سایت‌ها، مشتری احتمالاً توصیه‌های مربوط به محصولات مرتبط و یا اضافی را دریافت خواهد کرد.

### پیش‌بینی پیوند

هر سیستم متشکل از تعدادی المان مجزا که ارتباطاتی با هم دارند را می‌توان جزء شبکه‌های پیچیده[[51]](#footnote-51) به حساب آورد. شبکه‌های اجتماعی، زیستی و ‌‌‌‌‌اطّلاعاتی ‌می‌توانند توسط شبکه‌های پیچیده و پویا توصیف شوند. در این شبکه‌ها، گره‌ها[[52]](#footnote-52) نشان‌دهنده موجودیت‌ها و پیوند‌ها[[53]](#footnote-53) نشان‌دهنده ارتباطات بین موجودیت‌ها هستند. هنگام ایجاد تقریب توپولوژیکی از یک سیستم پیچیده، به علّت محدودیت زمان و فضا، وجود بعضی خطا‌ها یا پیوند‌های زائد اجتناب‌ناپذیر است؛ همزمان بعضی پیوند‌های بالقوّه کشف‌نشده نیز وجود خواهد داشت. بنابراین در این شبکه‌ها، مسئله‌‌ای به نام مسئله ‌‌پیش‌بینی پیوند مطرح ‌می‌شود و امکان وجود پیوند میان دو گره، براساس ویژگی‌های گره‌ها و دیگر پیوند‌های مشاهده‌شده در شبکه بررسی می‌گردد. پیوند‌هایی که ‌‌پیش‌بینی ‌می‌شوند، ‌می‌توانند پیوند‌های گمشده[[54]](#footnote-54) یا پیوند‌های آینده باشند[15].

کاهش هزینه و افزایش میزان موفقیت آزمایشات در شبکه‌های زیستی، کشف مکانیزم بیماری‌ها و روش‌های درمان آن‌ها در شبکه‌های ژن-بیماری[[55]](#footnote-55)، پیشنهاد دوستان بالقوّه به کاربران در شبکه‌های اجتماعی آنلاین، پیشنهاد کالا به مشتریان در شبکه‌های کاربر-آیتم[[56]](#footnote-56) و کشف ارتباط مخفی بین مجرمان از کاربردهای پیش‌بینی پیوند هستند.

الگوریتم‌های پیش‌بینی پیوند را می‌توان به سه دسته اصلی روش‌های مبتنی بر شباهت[[57]](#footnote-57)، روش‌های آمار و احتمالی[[58]](#footnote-58) و روش‌های الگوریتمی[[59]](#footnote-59) تقسیم کرد[16].

در رویکرد مبتنی بر شباهت، همه‌ی زوج گره‌ها بر اساس یک شاخص شباهت، رتبه‌بندی ‌می‌شوند و در زوج گره‌هایی که بیش‌ترین شباهت را دارند، احتمال به وجود آمدن پیوند بیش‌تر خواهد بود. همه‌ی فعّالیت‌های یک کاربر به وسیله یک بردار نشان داده شده و شباهت بین دو کاربر را می‌توان از کسینوس بین بردار‌های فعّالیت دودویی متناظرشان بدست آورد.

رویکرد مبتنی بر شباهت خود شامل دو دسته از شاخص‌ها می‌شود:

* شاخص‌های شباهت مبتنی بر ویژگی‌های گره که از ‌‌‌اطّلاعات مربوط به گره‌ها از قبیل پروفایل‌شان در شبکه‌‌های ‌‌اجتماعی آنلاین برای بیان شباهت میان گره‌ها استفاده می‌کنند.
* شاخص‌های شباهت مبتنی بر ساختار که برای محاسبه شباهت بین دو گره از ‌‌‌‌‌اطّلاعات ساختاری (توپولوژیکی) شبکه استفاده می‌کنند.

شاخص‌های شباهت مبتنی بر ساختار خود ‌می‌توانند از سه دیدگاه مورد توجّه قرار بگیرند: شاخص‌های محلّی[[60]](#footnote-60)، سراسری[[61]](#footnote-61) و شبه‌محلّی[[62]](#footnote-62).

شاخص‌های محلّی از ‌‌‌‌‌اطّلاعات همسایگی گره‌ها استفاده ‌می‌کنند و ساختار کل شبکه تأثیری در محاسبه‌ی پیش‌بینی پیوند نخواهد داشت. شاخص‌های سراسری به ‌‌‌‌‌اطّلاعات توپولوژیکی سراسری نیاز دارند. شاخص‌های شبه‌محلّی نیاز به ‌‌‌‌‌اطّلاعات توپولوژیکی سراسری ندارند امّا از ‌‌‌‌‌اطّلاعات بیش‌تری نسبت به شاخص‌های محلّی استفاده ‌می‌کنند. از شاخص‌های شباهت مبتنی بر ساختار محلّی، میتوان به جاکارد[[63]](#footnote-63)، اتصال ترجیحی[[64]](#footnote-64)، همسایگی مشترک[[65]](#footnote-65) و آدامیک آدار[[66]](#footnote-66) اشاره کرد. شاخص کاتز[[67]](#footnote-67)، SimRank و LHN2 از جمله شاخص‌های شباهت مبتنی بر ساختار سراسری هستند. شاخص مسیرِ محلّی[[68]](#footnote-68)، پیاده‌رویِ‌‌ تصادفی محلّی[[69]](#footnote-69) نیز جزء شاخص‌های شباهت مبتنی بر ساختار شبه‌محّلی به حساب می‌آیند. مزیّت‌ شاخص‌های محلّی، زمان اجرای پایین و مقیاس‌پذیری بالای آن‌هاست و عیب آن‌ها، میدان دید محدودشان است که ‌‌‌‌باعث کاهش دقّت‌شان می‌شود. به طور معکوس،‌ شاخص‌های سراسری، میدان دید وسیع‌تر، ‌‌‌‌دقّت بالاتر و به همین نسبت، زمان اجرای بالاتری هم دارند. شاخص‌های شبه‌محّلی هم از لحاظ دقّت و زمان اجرا، بین شاخص‌های محلّی و سراسری قرار دارند[17].

معیارهای شباهت زیادی جهت محاسبه میزان شباهت بین جفت گره‌ها در شبکه‌ وجود دارند. ضریب جاکارد[[70]](#footnote-70) یکی از پرکاربردترین و معمول‌ترین معیارهایی است که به عنوان معیار شباهت بین گره‌ها استفاده می‌گردد. برای محاسبه آن از رابطه زیر استفاده می‌شود.

|  |  |
| --- | --- |
| (1) |  |

در این رابطه صورت کسر برابر مجموعه همسایه‌هاي مشترک گره y ,x و مخرج کسر مجموعه کل همسایه‌هاي گره y ,x را نشان می دهد. JC هم نشان‌دهنده‌ي شاخص جاکارد می‌باشد. شاخص جاکارد همواره مقداري بین صفر و یک دارد. هر‌چه مقدار شاخص به 1 نزدیک‌تر باشد احتمال اینکه پیوندی بین دو گره y ,x ایجاد شود، بیشتر خواهد بود.

دسته‌ی دوم از الگوریتم‌های پیش‌بینی پیوند، روش‌های آمار و احتمالی نام دارند. در این روش‌ها فرض می‌شود که شبکه دارای ساختاری معین است. بنابراین برای شبکه مدلی که با ساختار آن متناظر باشد در نظر می‌گیرند. پس از انتخاب مدل، با توجه به ساختار شبکه و با استفاده از روش‌های آماری، پارامترهای این مدل تخمین زده می‌شوند. سپس با استفاده از پارامترهای تخمین زده شده، برای تمام پیوندهای ناموجود شبکه، احتمال شکل‌گیری پیوند محاسبه می‌شود. در نهایت می‌توان با استفاده از مقادیر احتمال پیوندها، مانند آنچه در روش‌های مبتنی بر مشابهت انجام می‌شود، پیوندهای پیشنهادی را رتبه‌بندی کرد.

سومین دسته از الگوریتم‌های پیش‌بینی پیوند، روش‌های الگوریتمی هستند. در این دسته از روش‌های مبتنی بر شباهت و روش‌های آمار و احتمالی، فرآیند پیش‌بینی پیوند بر اساس امتیازدهی به پیوندهای ناموجود شبکه انجام نمی‌شود. این دسته از روش‌ها با بهره‌گیری از الگوریتم‌های یادگیری نظارت‌شده[[71]](#footnote-71)، تکنیک‌های بهینه‌سازی[[72]](#footnote-72) و فاکتورگیری ماتریس[[73]](#footnote-73) برای حل مسائل تلاش می‌کنند[18].

در الگوریتم‌های پیش‌بینی پیوند نظارت‌شده، معمولاً بخشی از پیوندهای شبکه عمداً کنار گذاشته می‌شود. سپس با ارزیابی عملکرد الگوریتم در پیش‌بینی پیوندهای کنار گذاشته‌شده، پارامترهای شبکه تنظیم می‌شوند. الگوریتم‌های پیش‌بینی پیوند نظارت‌شده با تنظیم پارامتر در فرآیند پیش‌بینی پیوند، می‌توانند ساختار مخفی‌شده در شبکه را کشف کنند و در نتیجه رفتار هوشمندانه‌ای داشته باشند.

### سیستم‌های توصیه‌گر

سیستم توصیه‌گر یک راه‌حل کلیدی و مؤثر برای غلبه بر انبوه اطلاعات در دسترس کاربران است. یک سیستم توصیه‌گر با ارائه‌ی پیشنهاد به کاربر کمک می‌کند تا در میان حجم عظیم اطلاعات، سریع‌تر به هدف خود نزدیک شود. در سیستم‌های توصیه‌گر تلاش می‌شود که با حدس‌زدن شیوه‌ی تفکر کاربر از طریق اطلاعاتی که از نحوه‌ی رفتار او با کاربران مشابه آنها وجود دارد، مناسب‌ترین و نزدیک‌ترین مورد به سلیقه‌ی کاربر شناسایی و پیشنهاد شود.

روش‌هاي موجود براي سیستم‌هاي توصیه‌گر به سه گروه تقسیم می‌شوند[19]:

* روش‌هاي مبتنی بر محتوا[[74]](#footnote-74)
* روش‌هاي پالایش مشارکتی(اجتماعی)[[75]](#footnote-75)
* روش‌های ترکیبی[[76]](#footnote-76)

#### روش‌هاي مبتنی بر محتوا

هسته اصلی روش‌هاي بر پایه محتوا شناسایی ویژگی‌هاي مشترك بین اقلامی است که کاربر به آنها امتیاز بالا داده و توصیه اقلام جدید براي این کاربر با استفاده از این ویژگی‌هاست.

سیستم‌هاي بر پایه محتوا اطلاعات، محتواي اقلام را به صورت بردارهاي ویژگی نشان می‌دهند و از پروفایل کاربران براي توصیه اقلام جدید استفاده می‌کنند، به این صورت که می‌توان یک کالایي را که بردار ویژگی آن شبیه بردار پروفایل کاربر است به آن کاربر پیشنهاد داد. شباهت را اکثراً توسط شباهت کسینوسی[[77]](#footnote-77) یا کمترین فاصله اندازه‌گیري می‌کنند.

این دسته از روش‌ها که به‌طور کامل بر پایه محتوا هستند، با چالش‌هاي محدودیت اطلاعات و بایاس شدن روبرو هستند. محدودیت اطلاعات از آن‌جا ناشی می‌شود که سیستم‌هاي توصیه‌گر اطلاعات کمی در مورد کاربران و محتواي اقلام دارند. دلایل زیادي براي کمبود این اطلاعات وجود دارد: براي نمونه، ممکن است یک کاربر اطلاعات شخصی خودش را به خاطر حریم خصوصی ارائه نکند. یا ممکن است به دست آوردن محتواي واقعی اقلامی مانند موزیک و عکس سخت یا هزینه‌بر باشد و یا ممکن است محتواي اطلاعاتی یک کالا براي تشخیص مشخصات آن کافی نباشد. بایاس شدن سیستم توصیه‌گر به این معنا است که یک سیستم توصیه‌گر بر پایه محتوا، همیشه اقلامی را توصیه می‌کند که شبیه به اقلامی است که کاربر قبلا به آنها امتیاز بالا داده است. راهکاري که در بعضی مقالات براي این مسئله ارائه شده است به این صورت است که به اقلامی که خیلی شبیه هستند مقداري تصادفی را اضافه ‌کنند.

#### روش‌هاي پالایش مشارکتی

پالایش مشارکتی یکی از روش‌هايی است که با تحلیل امتیازات کاربران سعی می‌کند امتیاز بعدي را بر اساس کاربران مشابه پیش‌بینی کند یعنی به کاربر آیتمی توصیه خواهد شد که دیگران در گذشته با تمایل‌ها و ترجیح‌های مشابه او، این اقلام را پسندیده‌اند. در این روش، خود کالا اهمیتی ندارد و بر اساس انتخاب کاربران دیگر و انتخاب‌های گذشته‌ی آن کاربر، پیشنهادهای جدیدی به او ارائه می‌شود. در این روش ماتریسی به نام ماتریس امتیاز تشکیل می‌شود که عموماً بسیار تنک و خلوت است. در نتیجه روش‌هاي پالایش مشارکتی با چالش‌هاي نظیر خلوت بودن داده‌ها و مقیاس پذیري مواجه هستند. از طرف دیگر روش‌هاي پالایش مشارکتی برخی از محدودیت‌هاي روش‌هاي مبتنی بر محتوا را رفع کرده‌اند. براي نمونه، اگر محدودیتی در اطلاعات براي کالا موجود باشد، روشهاي پالایش مشارکتی می‌توانند با استفاده از بازخوردي که از کاربران دیگر می‌گیرند این کالا را توصیه کنند. به بیان دیگر، پیشنهاد‌هاي تولید‌شده توسط روش‌هاي پالایش مشارکتی بر مبناي کیفیت اقلامی است که همسایه‌هاي کاربر مورد نظر امتیازدهی کرده است، نه بر اساس محتواي اقلام.

روش‌ پالایش مشارکتی خود در حالت کلی به دو دسته بر پایه حافظه[[78]](#footnote-78) و بر پایه مدل[[79]](#footnote-79) تقسیم می‌شود. این روش‌ خود به دو صورت مبتنی بر کاربر و مبتنی بر کالا پیاده‌سازي می‌شود. بطور کلی در روش‌هاي بر پایه حافظه از امتیازات موجود به ‌صورت مستقیم براي تخمین امتیازات نامعلوم بهره‌برداری می‌شود. روش‌هاي بر پایه حافظه‌ی مبتنی بر کاربر، ارزیابی امتیاز یک کاربر به یک کالا را با استفاده از امتیازهاي داده شده به این کالا توسط کاربرانی که الگوي امتیازدهی مشابهی(همسایه‌ها) با کاربر مورد نظر دارند، انجام می‌دهد. همسایه‌هاي یک کاربر معمولاً کاربرانی هستند که امتیازات آن‌ها بر روي اقلام بیشترین همبستگی را با امتیاز کاربر مورد نظر داشته باشد. روش‌هاي بر پایه حافظه‌ی مبتنی بر کالا، امتیاز یک کاربر به یک کالا را بر مبناي امتیازات داده شده توسط کاربر به اقلام مشابه با کالاي مورد نظر تخمین می‌زنند. در روش‌هاي مبتنی بر کالا، دو کالا زمانی مشابه هستند که چندین کاربر به آنها امتیاز مشابه داده باشند.

برخلاف روش‌هاي بر پایه حافظه که به‌صورت مستقیم از امتیازات ذخیره شده براي تخمین استفاده می‌کنند، روش‌هاي بر پایه مدل از این امتیازات براي یادگیري یک مدل پیش‌بینی‌کننده امتیازات نامعلوم استفاده می‌کنند. ایده اصلی در روش‌هاي بر پایه مدل، مدل‌کردن پیوندهای کاربران و اقلام با فاکتورهایي است که ویژگی‌هاي نهفته کاربران و اقلام را در سیستم نشان می‌دهد. این مدل در ابتدا با استفاده از اطلاعات موجود آموزش داده می‌شود تا در مرحله بعدي براي تخمین امتیازات کاربران به اقلام جدید از آن استفاده شود. روش‌هاي مدل‌گراي زیادي براي توصیه اقلام وجود دارد که می‌توان به خوشه‌بندي بیزین[[80]](#footnote-80) و تحلیل معناي پنهان[[81]](#footnote-81) اشاره کرد.

#### روش‌های ترکیبی

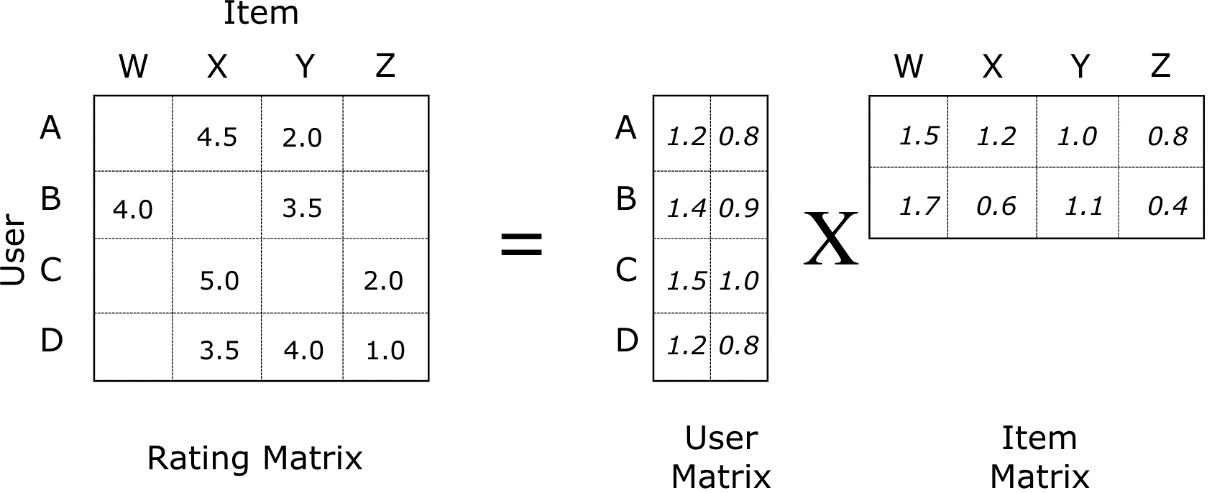
هر دو سیستم‌های توصیه‌گر مبتنی بر محتوا و سیستم‌های توصیه‌گر پالایش مشارکتی دارای مشکلاتی می‌باشند، به همین دلیل، روش‌های ترکیبی برای سیستم‌های توصیه‌گر ارائه شده‌اند که پالایش مبتنی بر محتوا و پالایش مشارکتی را با هم ترکیب می‌کنند و به صورت زیر دسته‌بندی می‌شوند[20].

* پیاده‌سازی روش پالایش مشارکتی و مبتنی بر محتوا به صورت جداگانه و ترکیب پیش‌بینی‌های آن‌ها.
* ترکیب برخی از ویژگی‌های مبتنی بر محتوا در یک روش مشارکتی.
* ترکیب برخی از ویژگی‌های مشارکتی در یک روش مبتنی بر محتوا.
* ایجاد یک مدل واحد کلی که شامل هر دو ویژگی‌های مبتنی بر محتوا و مشارکتی باشد.

همچنین تنک بودن داده‌های امتیازدهی کاربر-آیتم یکی از موانع عمده برای دست‌یابی به دقتی بالا در پیش‌بینی امتیازها محسوب می‌شود. برای غلبه بر این مشکل، پژوهشگران روش‌های ترکیبی را برای طراحی سیستم‌های توصیه‌گر پیشنهاد کرده‌اند. یکی از این روش‌ها به کارگرفتن اطلاعات کمکی در کنار داده‌های امتیازدهی است. یكی از روش‌های مهم استفاده شده در سیستم‌های پیشنهاددهنده، روش مبتنی بر فاکتورگیری ماتریس است. در این روش، با استفاده از امتیازات داده شده‌ی کاربران به برخی از آیتم‌ها، برای هر کاربر و هر آیتم، یک بردار ویژگی استخراج می‌شود که طول این بردار برابر با ویژگی‌های نهان[[82]](#footnote-82) کاربر یا آیتم است. ویژگی‌های نهان آیتم‌ها علاوه بر مقادیر به‌دست آمده از ماتریس امتیازها به محتوای آیتم‌ها نیز وابسته است. خطا و زمان اجرای این روش، وابسته به مقدار اولیه ویژگی‌های نهان می‌باشد[21].

توصیه‌گرهايی که مبتنی بر فاکتورگیري ماتریسی هستند با تبدیل کاربران و اقلام به فضاي ویژگی‌هاي نهفته[[83]](#footnote-83)، هرکدام از درایه‌هاي نامعلوم در ماتریس امتیازات را با استفاده از ضرب‌داخلی بردارهاي ویژگی متناظر که از امتیازات معلوم استنتاج شده‌اند، تعیین می‌کند.

مقداردهی اولیه ماتریس‌ها زمانی که ابعاد مسئله بزرگ باشد و محدودیت‌هاي زیادي بر روي ماتریس‌هاي فاکتور وجود داشته باشد، پیچیده‌تر هم خواهد شد. راهکار پیشنهادي استفاده از ماتریس شباهت بین کاربران و اقلام در مقدار‌دهی اولیه به جای استفاده از مقادیر تصادفی است.



شکل ‏1‑8 شماتیک فاکتور‌گیری ماتریسی[22]

نتیجه هر بار ضرب ماتریس‌های ویژگی کاربر و اقلام در یکدیگر باید به ماتریس اصلی امتیازات نزدیک‌‌تر شود. برای این‌کار یک تابع هزینه تعریف می‌شود که هر بار این فاصله را بوسیله کسر ماتریس بدست‌آمده حاصل‌ضرب از ماتریس اولیه اندازه می‌گیرد. با توجه به فاصله بدست‌آمده هر بار ماتریس‌های اولیه‌ی ویژگی با توجه به قوانین بروزرسانی، آپدیت می‌گردند و دوباره عمل ضرب مذکور تکرار می‌گردد. با توجه به اینکه پیدا کردن کمینه عمومی براي این مسئله غیرممکن است، ضرب مذکور تا زمانی که حاصل تفریق و فاصله ماتریس بدست آمده با ماتریس اولیه در دو گام پشت سر‌هم از مقدارتعیین شده‌ای کمتر گردد، ادامه می‌یابد.

دقت و همگرایی در این روش معمولاً خیلی وابسته به مقداردهی اولیه ماتریس‌های ویژگی است. اگر مقداردهی اولیه به‌صورت درست انجام نشود، همگرایی به کندي صورت می‌گیرد و حتی در بعضی موارد ممکن است به راه‌حل نادرست بیانجامد.

از جمله پژوهش‌های انجام شده در این زمینه، در رابطه با روش تجزیه مقادیر منفرد[[84]](#footnote-84)‌(SVD) است. چندین روش فاکتورگیري ماتریسی بر روي سیستم‌هاي توصیه‌گر پیاده‌سازي شده است که از آن جمله می‌توان به تحلیل معناي نهفته احتمالی (PLSA)، فاکتور‌گیری ماتریسی حاشیه بیشینه[[85]](#footnote-85)، بیشینه انتظار براي فاکتورگیري ماتریسی[[86]](#footnote-86) و فاکتورگیري ماتریسی منظم[[87]](#footnote-87) اشاره کرد.

## ساختار کلی پایان‌نامه

این پژوهش در 5 فصل به بررسی موضوع بازهدف‌گذاری دارو می‌پردازد که سرفصل مطالب بدین شرح می‌باشد:

* **فصل اول:** این فصل ابتدا به بیان مساله بازهدف‌گذاری دارو و ضرورت انجام تحقیق پرداخته و اهداف و فرضیات پژوهش در آن بررسی می‌شود. در ادامه مفاهیم پایه زیستی، کامپیوتری و محاسباتی مربوطه گفته شده و نحوه مدل‌سازی مسئله در دنیای کامپیوتر و ارتباط آن با مفاهیم تحلیل‌داده واکاوی می‌گردد.
* **فصل دوم:** در این فصل مطالعه‌ای بر روی پژوهش‌های پیشین در زمینه بازهدف‌گذاری دارو انجام شده و رویکرد‌های مختلف و منابع داده‌ای متفاوت معرفی می‌گردد.
* **فصل سوم:** در این فصل جزئیات روش پیشنهادی برای بازهدف‌گذاری دارو با استفاده از اطلاعات عوارض جانبی، شرح داده شده و روش ارائه شده شامل استفاده از تکنیک‌های گشت تصادفی[[88]](#footnote-88) برای ترکیب ویژگی‌های مختلف دارویی و بیماری، روش فاکتور‌گیری ماتریسی[[89]](#footnote-89) شرح داده می‌شود.
* **فصل چهارم:** در این فصل نتایج حاصل از ارزیابی روش پیشنهادی گزارش شده و نتایج آن با روش‌های پیشین مقایسه می‌شود و روی برخی از ارتباطات پیش‌بینی شده توسط روش پیشنهادی، بررسی موردی انجام می‌گیرد.
* **فصل پنجم:** این فصل در‌خصوص دستاوردهای بدست آمده بحث و نتیجه‌گیری کرده و جهت کارهای آتی پیشنهاداتی را ارائه داده است.

# فصل دوم: مروری بر مطالعات انجام شده

فصل دوم

تاریخچه پژوهش

## مقدمه

هدف این فصل بررسی پیشینه پژوهش و مروری بر روش‌ها و رویکردهای مختلفی است که تاکنون در مساله بازهدف‌گذاری دارو ارائه شده است. پژوهش‌های صورت گرفته در این خصوص را می‌توان از دو منظر مورد بررسی قرار داد. یکم از منظر نوع داده مورد استفاده قرار داده شده و دوم از منظر رویکرد محاسباتی به کار گرفته شده، که بر این اساس عناوین جدول 2-1 تنظیم شده است. در ادامه به بررسی تفصیلی این روش‌ها می‌پردازیم.

جدول ‏2‑1 تقسیم‌بندی روش‌های بازهدف‌گذاری دارو[23]

|  |  |
| --- | --- |
| باز‌هدف‌گذاری از منظر نوع داده | باز‌هدف‌گذاری از منظر رویکرد محاسباتی |
| * داده‌های مرتبط با دارو * داده‌های مرتبط با پروتئین * داده‌های مرتبط با بیماری * داده‌های مرتبط با ژن | * تحلیل شبکه[[90]](#footnote-90) * یادگیری ماشین[[91]](#footnote-91) * یادگیری عمیق[[92]](#footnote-92) * متن‌کاوی[[93]](#footnote-93) |

## انواع روش‌ها از منظر منابع داده‌ای

با توسعه سریع تکنیک‌های مرتبط با محاسبات زیستی، بانک‌های اطلاعاتی مختلفی از دارو مثل ChemBank, OMIM, Pubmed, DrugBank و بانک‌های اطلاعاتی ژنومی گسترده‌ای نیز مانند MIPS, GenBank و ... ایجاد شده‌اند. در جدول 2-2 خلاصه‌ای از انواع دیتابیس‌های مهم منتشر شده و برخی اطلاعات درباره آن‌ها مشاهده می‌شود.[23]

جدول ‏2‑2 منابع داده ای مختلف در بازهدف گذاری دارو [1]

|  |  |  |
| --- | --- | --- |
| آدرس اینترنتی | نام | زمینه |
| https://pubchem.ncbi.nlm.nih.gov/ | PubChem | ساختار شیمیایی دارو |
| https://www.uniprot.org/ | UniProt | ساختار سوم پروتئین |
| http://sideeffects.embl.de/ | SIDER | عوارض‌جانبی دارو |
| https://www.drugbank.ca/ | DrugBank | پروتئین‌های هدف داروها |
| http://stitch.embl.de/ | STITCH | ارتباطات ترکیبات شیمیایی و پروتئین |
| https://string-db.org/cgi/input.pl | STRING | پیوندهای پروتئین‌ها با یکدیگر |
| https://www.pharmgkb.org/ | PharmGKB | ارتباطات دارو-بیماری |
| https://omim.org/ | OMIM | متون مقالات مرتبط با بیماری |
| http://bio-annotation.cn:18080/DincRNAClient/ | DincRNA | ژن‌های مرتبط هر بیماری |
| https://phewascatalog.org/ | PheWAS | بانک اطلاعات فنوم |
| https://www.ebi.ac.uk/gwas/ | GWAS | بانک اطلاعات ژنوم |
| https://clue.io/cmap/ | CMap | ارتباطات دارو-ژن |
| https://www.ebi.ac.uk/interpro/ | InterPro | بیان ژن |
| https://ftp.ncbi.nlm.nih.gov/genbank/ | GenBank | توالی ژن |

### داده‌های مرتبط با دارو

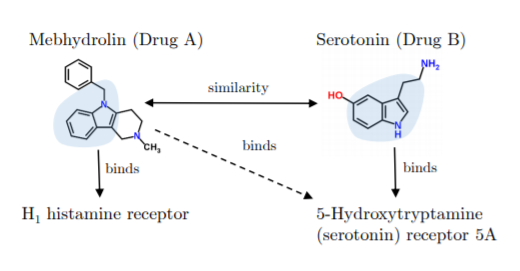
روش‌هایی که از داده‌های مرتبط با دارو استفاده می‌کنند از اصل شباهت[[94]](#footnote-94) ویژگی این‌چنین بهره می‌گیرند که داروهایی با ساختار شیمیایی مشابه فعالیت‌های بیولوژیکال یکسانی را انجام می‌دهند. روش‌های مبتنی بر اثرانگشت مولکولی[[95]](#footnote-95) و روش‌های مبتنی بر لیگاند[[96]](#footnote-96) در این دسته قرار می‌گیرند.

در روش‌های مبتنی بر اثرانگشت مولکولی هر دارو در یک بردار دودویی با مقادیر 0 و 1 کدگذاری می‌شود که صفر و یک به ترتیب به معنی وجود/عدم وجود یک زیرساختار شیمیایی خاص در ترکیب دارو است. در مقاله وانگ[[97]](#footnote-97) و همکاران روشی برای بازهدف‌گذاری دارو براساس اندازه‌گیری شباهت دو بردار دارو ارائه شده است[24].

لیگاند، در بیوشیمی و فارماکولوژی به ماده‌ای گفته می‌شود که در پیوند و ترکیب با یک بیومولکول یا یک گیرنده، هدف بیوشیمیایی خاصی را دنبال کند. بطور خاص یک لیگاند به عنوان ماشه برای فعال‌سازی یک عمل بیوشیمی به پروتئین هدف متصل شده و با آن یک ترکیب شیمیایی جدید ولی قابل برگشت را ایجاد می کند. در پژوهش کیسر[[98]](#footnote-98) و همکاران لیگاندهای شناخته شده بر اساس شرکت‌کننده‌های اتصال به هدف معلوم خود و خصوصیات شیمیایی با استفاده از الگوریتم‌های یادگیری‌ماشین بدون‌نظارت خوشه‌بندی شدند[25].

شولر[[99]](#footnote-99) و همکاران در سال 2019 نیز در روشی ترکیبی و با استفاده از تابع کرنل[[100]](#footnote-100) همزمان از شباهت ساختار شیمیایی و نیز شباهت لیگاند‌های شناخته شده روشی را برای بازهدف‌گذاری اراده می‌دهد[26].

با این حال در عمل مشخص شده که تغییرات اندک بر روی ساختار مولکول، می‌تواند منجر به پیامدهای بیولوژیکی بسیار متفاوت شود. برخی از ترکیبات دارویی قبل از اینکه به لحاظ فارماکولوژیکی فعال شده باشند، با واکنش سلول، دستخوش تغییرات قرار می‌گیرند. از این‌رو تکیه صرف بر ساختار شیمیایی‌ای که در پایگاه‌های داده ثبت شده، به لحاظ علمی رویکرد دقیقی نمی‌باشد و ریسک بالایی وجود دارد مبنی بر این که پیش بینی‌ها اثر‌بخشی تجربی ضعیفی را برای مورد مصرف جدید ابراز نمایند. همچنین در شکل ‏2‑1 شمای کلی این روش آورده شده است.



شکل ‏2‑1 بازهدف‌گذاری مبتنی بر شباهت ساختار شیمیایی دارو[27]

### داده‌های مرتبط با پروتئین

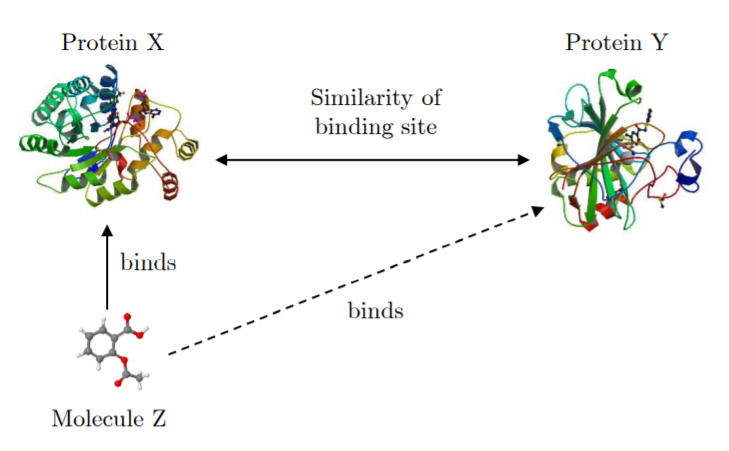
بیشتر مولکول‌های کوچک زیست‌فعال[[101]](#footnote-101) اثرات خود را با بر‌هم‌کنش با پروتئین‌ها اعمال می‌کنند. این بر‌هم‌کنش را می‌توان با استفاده از مدل‌سازی کامپیوتری ساختار سه‌بعدی پروتئین هدف و دارو تحلیل کرد. این روش موسوم به داکینگ مولکولی[[102]](#footnote-102) بوده و در راهبردهای کشف دارو به کار می‌رود. این روش با شناسایی و بهینه‌سازی میل اتصال به جایگاه فعال هدف کمک می‌کند تا قدرت داروی توسعه یافته را افزایش دهد. به سبب توان روش داکینگ مولکولی، فرصت‌های بازهدف‌گذاری دارو بر این اساس زیاد بوده که موجب محبوبیت این روش نیز شده ‌است. در حال حاضر مشخص گردیده که بیشتر ترکیبات با بیش از یک پروتئین واکنش می‌دهند. از این‌رو شناسایی پروتئین‌های هدف‌های شناخته‌نشده‌ی یک ترکیب دارویی با مشاهده و غربال ساختارهای سه‌بعدی پروتئین‌های موجود در پایگاه‌داده‌ها هدف اصلی این روش است. اگر هدف‌های اضافی پیش‌بینی‌شده، مرتبط با بیماری باشند، آن‌گاه دارو می‌تواند بر این اساس بازهدف‌گذاری شود. بر این اساس بسیاری از مقالات اخیر بر جایگاه‌های اتصال[[103]](#footnote-103) متمرکز شده و مشابهت‌های نسبی آن‌ها را مورد مقایسه قرار داده‌اند[28].

به طور خلاصه این روش متشکل از چند گام است: استخراج جایگاه اتصال از ساختار سه‌بعدی توالی‌های پروتئین، شناسایی جایگاه‌های اتصال مشابه در طول پروتئوم با استفاده از جستجوی الگوریتمی و در نهایت تحلیل دستی داکینگ به منظور حصول اطمینان از امکان برهم کنش فیزیکی.

همان‌گونه که در شکل 2-2 آمده است ساختار سه بعدی پروتئین‌ها و جایگاه‌های اتصال مربوط به آن‌ها می‌توانند از طریق یک تابع امتیاز‌دهی مورد مقایسه قرار گیرند. بر این اساس، فرض شده که جایگاه‌های اتصال مشابه می‌توانند به لیگاند مشابه متصل شوند. برای نمونه با علم به این که پروتئین X یک جایگاه اتصال مشابه با دیگری در پروتئین Y دارد و این که مولکول Z با پروتئین X پیوند برقرار می‌کند، می‌توان نتیجه‌گیری کرد که مولکول Z نیز با پروتئین Y پیوند برقرار کند.

اروچی[[104]](#footnote-104) و همکاران در پژوهشی در سال 2019 با استفاده از این روش به بازهدف‌گذاری دارویی برای هدف پروتئینی Pim-1[[105]](#footnote-105) پرداختندکه به عنوان عامل سرطان پروستات و لوسمی حاد مغز استخوان شناخته‌‌شده است[29].

اما این روش خود با چالش‌هایی نیز روبرو است: اول اینکه اطلاعات ساختار سه بعدی باید موجود باشد در صورتی که در حال حاضر پایگاه‌های‌داده هنوز نمی‌توانند کل پروتئوم را پوشش دهد و دسترسی به ساختار سه بعدی برخی پروتئین‌ها امکان‌پذیر نیست. چالش بعدی تشخیص خودکار جایگاه اتصال است، به ویژه در زمانی که ساختار پروتئین بدون لیگاند کریستالیزه می‌شود.



شکل ‏2‑2 بازهدف‌گذاری دارو با استفاده از ساختار پروتئینی و جایگاه اتصال[27]

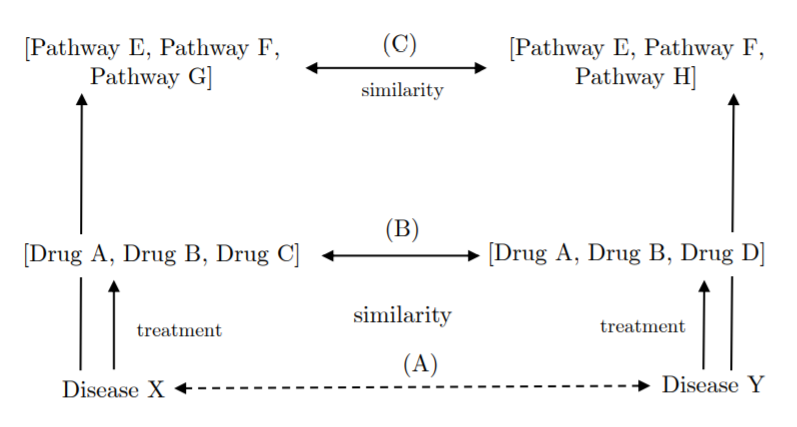
در نهایت از آن جایی که این روش‌ها تعداد زیادی از جایگاه‌های غیر‌صحیح را تولید می‌کند، اعتبار‌سنجی دستی و آزمایشگاهی تنها راهکار برای ارزیابی پیش‌بینی‌ها می‌باشند[30].

روش‌های مبتنی بر پروتئین احتمالاً دقیق‌ترین روش برای بررسی برهم‌کنش حقیقی بین یک دارو و هدف پروتئینی هستند. روش های داکینگ یک تصویر دقیق سطح پایین از کمپلکس بیوشیمیایی ارائه می‌کنند، اما در بخش مدل‌سازی با چالش‌هایی روبرو هستند. شناسایی پروتئین‌های هدف اضافی لزوماً همواره منتهی به فرصت‌های بازهدف‌گذاری نمی‌شوند و نتایج باید در یک حوزه بیولوژیکی گسترده‌تر تفسیر شوند.

### داده‌های مرتبط با بیماری

در گذشته، بیماری‌ها بر اساس علت بیماری‌زایی مانند عفونت یا اختلال بیولوژیکی مشاهده‌شده همچون رشد بدون کنترل سلول دسته‌بندی می‌شدند. از آن جایی که بیماری‌های مشابه با روش مشابه، درمان می‌شوند، توصیف بهتر ارتباط موجود بین بیماری‌زایی می‌تواند فرضیه‌های بازهدف‌گذاری دارو را پایه‌ریزی کند. شکل ‏2‑3 بازهدف‌گذاری دارو با استفاده از رابطه بیماری‌ها[27]ساخت شبکه رابطه‌ی بین بیماری‌ها را نشان می‌دهد.

گروهی از محققان بیماری را با استفاده از فهرستی از داروهای مورد استفاده در درمان و موارد مصرف تأیید‌نشده از دارو تعریف کرده‌اند. با وجودی که ممکن است سهل انگارانه به نظر برسد، اما این کار با نمونه‌های موفق پشتیبانی شده و به طور متداول در کارآزمایی‌های بالینی اجرا می‌شود. پژوهشگران یک مورد مصرف دارویی برای دو بیماری مشابه در نظر گرفتند تا داروی



شکل ‏2‑3 بازهدف‌گذاری دارو با استفاده از رابطه بیماری‌ها[27]

تجویز شده برای یکی از آن‌ها را برای درمان دیگری پیشنهاد دهند. با استفاده از 700 بیماری و 2000 دارو، بیش از 150000 رابطه جدید تولید شده است. جالب توجه این که موارد مصرف جدید با داده های بالینی آزمایشی مطابقت دارند و کاربرد پیش‌بینی شده‌ی جدید اغلب توسط پزشکان گزارش شده است. در مواردی نیز داروهای مشابه برای برخی از رابطه‌ها فاقد شواهد بالینی هستند که باید مطالعات بیشتری در موردشان صورت گیرد[31].

اولین کار صورت گرفته با استفاده از این متد پژوهش اگروال[[106]](#footnote-106) و همکاران در سال 2009 است که از طریق بررسی مسیر بیولوژیکی[[107]](#footnote-107) ژن‌های بیماری‌زا به بررسی الگوهای مشترک بیولوژیکی بیماری‌ها پرداخته است[32].

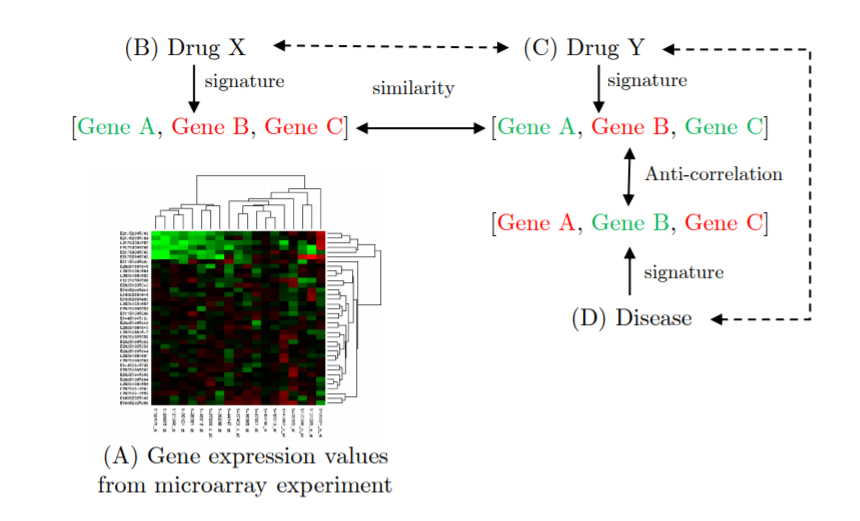
### داده‌های مرتبط با ژن

ارگان‌های زنده موجودات را می‌توان با مشاهده عملکرد بیان ژنی آن‌ها در یک حالت خاص درک کرد. بر اساس حالت ارگان زنده، ژن‌های مشخصی بیشتر و یا کمتر بیان می‌شوند و این ژن‌ها از طریق بررسی تعداد نسبی مولکول‌های mRNA رونویسی شده، قابل شناسایی هستند. تمایز بیان ژن را می‌توان دلیلی بر اثر مولکولی خاص دانست که تحت عنوان امضای بیان ژن[[108]](#footnote-108) شناخته می‌شود. یکی از موفق‌ترین نمونه‌های این روش پایگاه‌داده CMap[[109]](#footnote-109) است. این پایگاه که نتیجه تحقیق لمب[[110]](#footnote-110) و همکاران است در سال 2006 بوجود آمده و بر اساس این ایده که عملکرد دارو را می‌توان با بررسی پروفایل بیان ژن حاصل از مصرف آن دارو بر روی یک سیستم بیولوژیکی بدست آورد، کار می‌کند[33].

در شکل 2-4 در قسمت A نمونه‌ای از نتیجه به دست آمده از یک آزمایش بیان ژن مشاهده می‌شود. برخی از ژن‌های شناسایی شده بیشتر از حد مشخصی‌اند (رنگ سبز) و برخی دیگر نیز تعداد کمتری دارند (رنگ قرمز). B و C داده‌های بیان ژن از نقشه ارتباط هستند که امضایی را فراهم می‌آورند که می‌تواند داروها را به جنبه عملکردی‌شان ارتباط دهد. برای نمونه داروی X و Y مشابه در نظر گرفته می‌شوند چون مقادیر قابل ملاحظه‌ای از ژن‌های مرتبط را به اشتراک گذاشته‌اند.

در این‌جا فرض بر این است که ترکیبات با امضاهای مشابه به لحاظ عملکردی مرتبط هستند، چرا که آن‌ها عملکرد سلول‌ها را در یک حالت مشابه تغییر می‌کنند. بر این مبنا محققان مجموعه‌ای از داروها را شناسایی کردند که هدف‌های مشخص پروتئینی و مکانیسم عمل را به اشتراک می‌گذارند. نکته جالب توجه این که این ترکیبات شباهتی را در سطح امضا نشان می‌دهند و پروفایل بیان ژنی مشابه دارند. اما با این حال به لحاظ بالینی برای درمان بیماری‌های متفاوتی مورد استفاده قرار می‌گیرند.

مطالعه دادلی[[111]](#footnote-111) و همکاران در سال 2011 نمونه‌ای موفق از این رویکرد تحقیقی بوده که توانسته است کاربرد درمانی داروی تاپیرامات[[112]](#footnote-112) را برای بیماری التهاب روده[[113]](#footnote-113) نشان دهد. در جایی که این دارو سابق بر این به عنوان داروی ضد تشنج شناخته می‌شد[34].



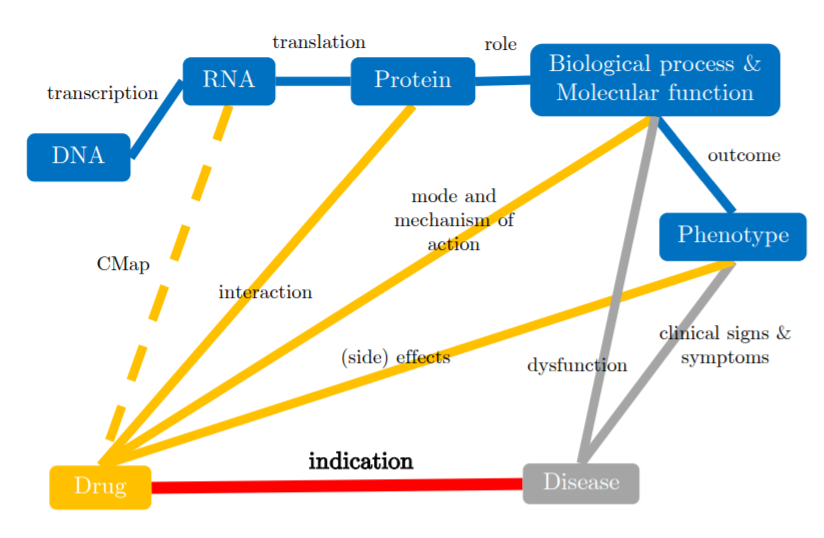
شکل ‏2‑4 بازهدف گذاری دارو با استفاده از بیان ژن[27]

با این حال این رویکرد نیز خود با چالش‌هایی روبرو است از جمله اینکه اولاً، در این روش پروفایل بیان ژن دارو یا بیماری باید در دسترس باشد. CMap فهرست نسبتاً کوچکی را ارائه می‌کند و بسیار کوچک‌تر از آن است که نماینده همه داروهای مورد تأیید و آزمایشگاهی باشد. این مورد ترکیباتی که می‌توانند مورد بررسی قرار گیرند را محدود می‌کند.

ثانیاً، پروفایل‌های بیان ژن خود به طور قابل بحثی یک حالت بیماری یا یک پاسخ دارویی را تعریف می‌کند، بافت‌ها در CMap گنجانده نشده‌اند و این منبع تنها براساس ثبت پاسخ سلول‌های سرطانی توسعه یافته و اساساً با همه انواع بیماری مرتبط نیست[27].

### مدل‌های ترکیبی

از آنجایی که داده‌های بیولوژیک از طرفی در برخی موارد ناکامل و محدود هستند و از طرف دیگر مفاهیم مکملی را بیان می‌کنند لذا می‌توان از ترکیب منابع‌داده‌ای مختلف بهره گرفت که علاوه بر بالا‌رفتن دقت فرآیند باعث ارائه پیش‌بینی‌هایی منحصر به فرد نیز بشوند. پژوهش‌ حاضر نیز بازهدف‌گذاری دارو را از این منظر مورد بررسی قرار داده است. در شکل ‏2‑5 نقشه ارتباط میان مفاهیم مختلف بیولوژیکی[27] آمده است.



شکل ‏2‑5 نقشه ارتباط میان مفاهیم مختلف بیولوژیکی[27]

## انواع روش‌ها از منظر رویکرد محاسباتی

روش‌هایی که تا اینجا مورد بحث قرار گرفتند، هر کدام مبتنی بر یکی از مفاهیم بیولوژیکی خاص بودند. حال از منظر رویکرد محاسباتی مورد استفاده قرار گرفته در بازهدف‌گذاری به بررسی انواع الگوریتم‌ها اعم از رویکردهای آماری، الگوریتم‌های یادگیری‌ماشین و ... را می‌پردازیم. تقریباً با هر کدام از رویکردهای گفته شده می‌توان به دلخواه یکی از انواع منابع داده‌ای که در بخش قبل ذکر گردید، را انتخاب کرد و به بررسی موضوع باز‌هدف‌گذاری پرداخت.

### یادگیری ماشین

همان طور که می‌دانیم در یادگیری ماشین، به جای برنامه‌نویسی همه چیز، داده‌ها به یک الگوریتم عمومی داده می‌شوند و این الگوریتم است که براساس داده‌هایی که به آن داده شده منطق خود را می‌سازد. در این بین استفاده از الگوریتم‌هایی مانند ماشین بردار پشتیان[[114]](#footnote-114)، کمترین مربعات[[115]](#footnote-115) و رگرسیون لجستیک[[116]](#footnote-116) جزء قدیمی‌ترین روش‌های بکار گرفته شده در این حوزه به‌شمار می‌آیند.

بلیک‌لی[[117]](#footnote-117) و همکاران در سال 2009 روشی بر پایه یادگیری ماشین نظارت‌شده به نام مدل محلی دوگانه[[118]](#footnote-118) ارائه دادند که برای پیش‌بینی از ماشین بردار پشتیبان به عنوان طبقه‌بند محلی[[119]](#footnote-119) استفاده می‌کند[35]. این روش بازهدف‌گذاری را در قالب یک مساله طبقه‌بندی دودویی فرموله کرده است. بدین معنی که برای ارتباط داشتن زوج‌ها، کلاس 1 و عدم ارتباط زوج‌ها کلاس 0 را در نظر گرفته است. در این روش برای پیش‌بینی ارتباط هر زوج دارو-پروتئین دو پیش‌بینی مستقل از هم از هر طرف انجام می‌شود. یک‌بار با توجه به شباهت و روابط داروها با هم پیش‌بینی ارتباط با یک دارو با یک پروتئین خاص انجام می‌شود و در مرتبه دوم با توجه به شباهت پروتئین‌ها و روابط پروتئین‌ها با هم ارتباط/عدم ارتباط یک زوج دارو-پروتئین انجام می‌گیرد و در نهایت این نتایج با هم ترکیب شده و طبقه‌بند تصمیم می‌گیرد. هر چند این پژوهش در زمان خودش به دقت خوبی نسبت به سایر روش‌ها رسید اما برای پیش‌بینی در مورد دارو یا پروتئین‌هایی که تا به حال ارتباطی برای‌شان کشف نشده(تمامی ستون‌های یک سطر 0 باشد) کارا نبوده و نمی‌تواند بدرستی عمل بکند.

در بیشتر پژوهش‌های انجام شده پیوندهای شناخته‌شده بعنوان نمونه‌های مثبت و سایر پیوندها بعنوان نمونه‌های منفی جهت آموزش به مدل داده می‌شوند. در صورتی که نمی‌توان صرفاً عدم گزارش شدن پیوند بین زوج‌ها بعلت محدودیت دانش و ... را به عنوان عدم ارتباط قطعی لحاظ کرد. پس دادن این نمونه‌ها بعنوان نمونه‌های منفی به مدل موجب بایاس شدن طبقه‌بند می‌گردد. برای مقابله با این چالش تاکنون راه‌حل های متفاوتی ارائه شده است. از جمله در پژوهشی مربوط به سال 2017 پنگ[[120]](#footnote-120) و همکاران روشی به نام NDTISE توسعه دادند که در آن ابتدا به بررسی عمیق‌تر نمونه‌های بدون برچسب (زوج‌هایی که ارتباط بین آن‌ها گزارش نشده) پرداختند[36]. این نمونه‌ها به دو دسته تقسیم شد و نمونه‌های مشکوک[[121]](#footnote-121) از نمونه‌هایی که با قطعیت بیشتر می‌توان درباره عدم ارتباط آن‌ها نظر داد جدا شد. در مرحله بعد تنها از مجموعه دوم نمونه‌هایی به عنوان نمونه‌های منفی به مدل داده شد و در نهایت یک طبقه‌بند پشتیبان وظیفه تصمیم‌گیری را انجام داد. نتایج بدست آمده حاکی از بهبود دقت این روش نسبت به روش‌هایی که به صورت تصادفی نمونه‌های گزارش نشده را بعنوان نمونه منفی به مدل می‌دهند، بود.

از جمله پژوهش‌هایی که با استفاده از روش کمترین مربعات به حل مساله بازهدف‌گذاری پرداخته‌اند، پژوهشی است که چن[[122]](#footnote-122) و همکاران در 2017 انجام داده‌اند[37]. در مدل ارائه‌شده به نام FRMSL با بهره‌گیری از متد کمترین مربعات منظم،[[123]](#footnote-123) بهترین برازش برای تابع نگاشت فضای دارو به فضای بیماری را پیدا می‌کند. همچنین در این روش توابع هسته[[124]](#footnote-124) خطی برای ترکیب شباهت‌های دارو و بیماری استفاده شده است.

اما در پژوهشی دیگر کیم[[125]](#footnote-125) و همکاران در سال 2019 بطور موازی چند مدل پیش‌بینی را با استفاده از رگرسیون لجستیک و ماشین بردار پشتیبان با توابع هسته خطی و غیرخطی پیاده‌سازی کردند[38]. در این پژوهش برای هر زوج دارو-بیماری یک امتیاز ارتباط به ازای هر ویژگی(4 نوع شباهت دارو و 3 نوع شباهت بیماری) به عنوان متغیر وابسته محاسبه شده و توسط ماشین مدل می‌شود. در نهایت دقت روش با طبقه بند رگرسیون لجستیک از ماشین بردار پشتیبان با تابع هسته خطی بیشتر شد.

### تحلیل شبکه

تحلیل شبکه یکی از استراتژی‌هایی که به طور گسترده در سال‌های اخیر برای بازهدف‌گذاری مورد استفاده قرار گرفته است. مهم‌ترین تفاوت این رویکرد با رویکرد پیشین در نظر گرفتن ویژگی‌های توپولوژیکی است. امکان استفاده از معیارهای شباهت محلی و سراسری وجود دارد که این امر باعث حفظ اطلاعات کمی و کیفی شبکه شده و یکپارچه‌سازی شباهت‌ها را بطور موثرتری اجرا می‌کند. در این رویکرد معمولاً یک شبکه نامتجانس با چند سطح تعریف می‌شود که در هر سطح آن ویژگی‌های یک موجودیت خاص(دارو، بیماری و پروتئین) وجود دارد.

این استراتژی خود به چهار شیوه پیاده‌سازی می‌شود:

* خوشه‌بندی[[126]](#footnote-126): در این روش با استفاده از الگوریتم‌های خوشه‌بندی با توجه به ساختار توپولوژیک شبکه ابتدا زیرشبکه‌ها یا به عبارتی خوشه‌های جالب توجه شناسایی شده و سپس الگوریتم در این خوشه‌ها هدف‌گذاری جدید برای دارو را شناسایی می‌کند.

در پژوهش لو[[127]](#footnote-127) و همکاران که در سال 2016 برای شناسایی داروی جدید جهت بهبود درمان نوعی خاصی از سرطان ریه[[128]](#footnote-128) انجام شده از خوشه بندی k-means استفاده شده است[39]. در این پژوهش ابتدا گروه داروهایی با ساختار شیمیایی یکسان با داروهای از پیش تایید شده برای درمان سرطان ریه که روی ژن‌های این بیماری اثرگذاری دارند شناسایی گردید و سپس ادامه مطالعه بر روی آن گروه پیگیری گردید. همچنین برای چالش خوشه‌هایی که روی هم قرار گرفته‌اند و با هم گره‌های مشترک دارند نیز راه‌حلی ارائه شده است.

* انتشار برچسب[[129]](#footnote-129): ایده اصلی این روش انتشار دادن اطلاعات از گره منبع به گره‌های بدون برچسب بصورت لایه لایه در تمام شبکه است. این روش از نظر نحوه و میزان انتشار خود به دو دسته محلی و عمومی تقسیم می‌شود.

یان[[130]](#footnote-130) و همکاران در پژوهش خود در سال 2016 روشی به نام LPMIHN بر پایه انتشار برچسب توسعه دادند که در آن دو بار فرآیند انتشار برچسب بصورت مستقل انجام می‌شود[40]. در این روش بطور مثال برای یک داروی مورد جستجو ابتدا یک بار در شبکه شباهت دارویی فرآیند انتشار صورت می‌گیرد و در مرتبه دوم نیز این بار با توجه به برهم‌کنش‌های کشف‌شده در شبکه شباهت پروتئینی انتشار برچسب در سطح شبکه شباهت پروتئین انجام می‌شود و امتیازات پیوند احتمالی بروز می‌گردد. در نهایت زوج‌هایی که امتیاز ارتباط‌شان از مقدار آستانه از پیش تعیین شده بیشتر باشد بعنوان کاندید انتخاب می‌گردند.

از آنجایی که انتشار اطلاعات بصورت محلی اطلاعات محدودی از شبکه را وارد می‌کند امکان انجام پیش‌بینی اشتباه بیشتر بوده و به همین خاطر بیشتر سعی در استفاده همزمان و متوازن از روش محلی و سراسری دارند.

در پژوهش شهرضا[[131]](#footnote-131) و همکاران در 2017 روش انتشار برچسب برای پیش‌بینی ارتباطات دارو-پروتئین هدف، دارو-بیماری و بیماری-پروتئین با نام Heter-LP ارائه شده که در آن از تلفیق منابع اطلاعاتی مختلفی در سطوح مختلف زیستی برای یافتن روابط جدید استفاده می‌شود[41]. از مهم‌ترین مزایای Heter-LP نسبت به روش‌های قبلی می‌توان به تلفیق موثر داده‌های ورودی، عدم نیاز به نمونه‌های منفی و استفاده از ویژگی‌های محلی و سراسری شبکه در کنار هم اشاره کرد. انتشار برچسب در این روش شامل یک بخش نگاشت و یک بخش تکرار شونده دیگر برای تعیین برچسب‌های نهایی کل رئوس شبکه می‌باشد.

* فاکتورگیری ماتریسی: به منظور توسعه الگوریتم‌های کارآمدتر و دقیق تر، تکنیک‌های متنوعی از زمینه‌های مختلف در حوزه‌های زیست پزشکی گنجانیده شده است که از جمله آن‌ها می توان به روش فاکتورگیری ماتریسی اشاره کرد. این روش یک ماتریس داده را به ماتریس‌هایی با ابعاد پایین تبدیل کرده در‌حالی‌که خصوصیات توپولوژیکی ماتریس داده اصلی در آن‌ها نیز پنهان نگه داشته شود. براي تخمین ماتریس‌هاي ویژگی نیاز به یک تابع هزینه است تا هر بار فاصله بین ماتریس اصلی و ماتریس تخمین زده شده را مشخص کند. فاکتورگیری ماتریسی در بین تمامی روش‌های بازهدف‌گذاری کمترین وابستگی را به داشتن نمونه‌های منفی دارد. فاکتورگیری ماتریسی خود به شیوه‌های مختلفی پیاده‌سازی می‌شود گرچه می‌توان ادعا کرد بیشتر این تفاوت‌ها در تابع هزینه‌ی بکارگیری شده است.

در پژوهش کدیا[[132]](#footnote-132) و همکاران در سال 2020 روشی بر پایه فاکتورگیری ماتریسی غیرمنفی ارائه شده که در آن روی ماتریس‌های تجزیه‌شده‌ی ویژگی در حین عملیات ضرب و بروزرسانی محدودیت غیرمنفی بودن اعمال شده است[42]. علاوه بر آن در این روش جهت مقدار‌دهی اولیه ماتریس‌های ویژگی علاوه بر ماتریس شباهت، کوتاه‌ترین مسیر بین گره‌ها در هر سطح از شبکه بطور مجزا حساب شده و تاثیر آن در مقداردهی اعمال می‌شود.

در پژوهشی دیگر مونگیا[[133]](#footnote-133) و همکاران در سال 2020 جهت مقداردهی اولیه ماتریس‌های ویژگی دارو و بیماری با استفاده از الگوریتم K-نزدیک‌ترین همسایه[[134]](#footnote-134) تنها اطلاعات K شبیه‌ترین عناصر را وارد پروفایل ویژگی هر یک کردند. این روش برای کمینه‌سازی تابع هزینه از منظم‌سازی گراف لاپلاسی[[135]](#footnote-135) استفاده می‌کند[43].

در پژوهش لو[[136]](#footnote-136) و همکاران در سال 2017 روشی برای بازهدف‌گذاری بر اساس برهم‌کنش‌های دارو-پروتئین با یکپارچه‌سازی چهار نوع شباهت دارویی و سه نوع شباهت پروتئینی ارائه شد[44]. در این روش در مرحله اول علاوه بر شباهت محلی با استفاده از گشت تصادفی، شباهت سراسری نیز استفاده شده است. علاوه بر آن از آنجایی که نزدیکی هندسی در فضای بردار ویژگی ممکن است ارتباط عملکردی را منعکس کند، از یک رویکرد تکمیل ماتریس برای دستیابی به یک ماتریس نشان‌دهنده‌ی بردارهای ویژگی ابعاد پایین نیز استفاده شده است.

* تعبیه(برازش) گراف[[137]](#footnote-137): تعبیه در یادگیری ماشین به معنای نگاشت از فضای ورودی به یک فضای جدید و قابل هضم برای مساله است به نحوی که اطلاعات مفید و مهم ورودی حفظ شود. از آنجا که معمولاً اطلاعات مورد نیاز برای بازهدف‌گذاری به شکل گراف و شبکه هستند و کار کردن با این داده‌ها به خاطر تفاوت از نظر مقیاس، ویژگی و موضوع سخت و در برخی موارد غیر‌ممکن است، استفاده از روش‌های تعبیه می‌تواند کمک فراوانی به حل مساله بازهدف‌گذاری دارو کند. تعبیه گراف رویکردی برای تبدیل گره‌ها، لبه‌ها و ویژگی‌های گراف به فضای بردار (یک بعد پایین) عددی در عین اینکه حداکثر خواص مانند ساختار گراف و اطلاعات آن حفظ شود. استفاده از تعبیه گراف راهی مناسب برای روبرو‌شدن با چالش خلوت بودن ماتریس‌های زیستی نیز به شمار می‌رود. این روش به عنوان یک روش مجزای بازهدف‌گذاری دارو مطرح نیست و بیشتر بعنوان راهی برای پیش‌پردازش داده با هدف کاهش ابعاد شناخته می‌شود و از سایر روش‌های یادگیری‌ماشین، یادگیری عمیق و یا پیش‌بینی‌پیوند باید در کنار آن استفاده کرد.

یو[[138]](#footnote-138) و همکاران در پژوهشی در سال 2020 بسته جامعی به نام BioNEV طراحی کردند که با استفاده از آن دو دسته وظیفه پیش‌بینی پیوند و طبقه‌بندی چند کلاسه در شبکه‌های زیستی قابل اجراست[45]. در بخش پیش‌بینی پیوند سه شبکه مورد علاقه پژوهشگران یعنی دارو-بیماری، برهم‌کنش‌های پروتئینی، تداخلات دارویی و در بخش طبقه‌بندی نیز پیش‌بینی کارکرد پروتئین و طبقه‌بندی انواع عبارات پزشکی پیاده‌سازی شده است.

### یادگیری عمیق

یادگیری عمیق زیر شاخه‌ای از یادگیری ماشینی و بر مبنای مجموعه‌ای از الگوریتم‌ها است که در تلاشند تا مفاهیم انتزاعی سطح بالا در داده را مدل نمایند. یادگیری عمیق به سرعت در بسیاری از زمینه‌های علم و فناوری مانند پردازش تصویر، متن، تشخیص صدا، فیلم و ... موفقیت‌های چشم‌گیری داشته است. به‌تازگی از روش‌های یادگیری عمیق برای بازهدف‌گذاری دارو نیز استفاده شده است. با استفاده از روش‌های یادگیری عمیق ویژگی‌های ساختاری یک شبکه و وضعیت توپولوژیکال یک گره در سطح شبکه را بعنوان یک ویژگی استخراج کرد و از آن برای محاسبه شباهت دو گره همسان بهره برد. در این روش می‌توان ویژگی‌های هر نمونه دارو، بیماری، ژن، پروتئین و ... بعنوان یک بردار دودویی به صورت خام و بدون دادن هیچ اطلاعات اضافی‌ای به مدل داد و یا ابتدا از روش‌های دیگر بازنمایی اطلاعات مفید را استخراج و سپس به عنوان ورودی به مدل داد.

deepDR نام مدلی است که ژنگ[[139]](#footnote-139) و همکاران در سال 2019 برای بررسی بازهدف‌گذاری داروها ارائه دادند[46]. در این مدل ابتدا از گشت تصادفی برای بازنمایی شبکه‌ها استفاده می‌شود و سپس محققین از یک خودرمزگذار[[140]](#footnote-140) برای یکپارچه‌سازی و فشرده‌سازی داده‌ها بهره بردند. مدل در نهایت برای پیش‌بینی پیوند‌های جدید از cVAE[[141]](#footnote-141) استفاده کرده و میزان AUC=0.90 و AUPR=0.92 می‌گردد.

در مطالعه‌ای دیگر اوزتورک[[142]](#footnote-142) و همکاران در سال 2018 با ارائه مدل DeepDTA و با استفاده از شبکه‌های عصبی پیچشی[[143]](#footnote-143) به بررسی پیوندهای دارو-پروتئین پرداختند[47]. ایشان ابتدا روی دو نوع داده توالی پروتئین و رشته SMILES پروتئین را تکنیک رمزگذاری برچسبی[[144]](#footnote-144) را پیاده کردند. سپس هر دوی داده‌ها را جهت نگاشت ویژگی به طور موازی به لایه‌های کانولوشنی می‌دهند. نتایج در لایه ادغام[[145]](#footnote-145) نمونه‌کاهی شده و ویژگی‌های استخراج شده به یکدیگر متصل[[146]](#footnote-146) شده و جهت پیش‌بینی نهایی به سه لایه تماماً متصل[[147]](#footnote-147) داده می‌شود.

### متن‌کاوی

متن‌کاوی، فرآیند استخراج اطلاعات با کیفیت از متن است. اطلاعات با‌کیفیت، به‌طور معمول از فهم الگوها و گرایش‌ها از طریق معانی و به وسیله‌ی یادگیری الگوهای آماری حاصل می‌شود. مقالات، کتب و بطور کلی متون منتشرشده در حوزه‌های پزشکی و داروسازی منبع غنی و عظیمی از اطلاعات درخصوص دارو و بیماری به حساب آمده و زمینه مناسبی را جهت داده‌کاوی آن فراهم کرده است.

مدل ABC سوانسون[[148]](#footnote-148) محبوب‌ترین متدولوژی روش‌های برپایه متون است. بنا بر این اصل اگر یک مطالعه نشان دهد که بیماری A به دلیل عامل B ایجاد شده است و در عین حال مطالعه دیگر گزارش کند که داروی C که برای بیماری دیگری استفاده می‌شود فعال کننده عامل B است، می‌توان داروی C را برای بیماری A بازهدف‌گذاری کرد.

در پژوهشی که در سال 2019 برای کشف دارویی جهت درمان ریزش موی ناشی از شیمی‌درمانی بوسیله ژانگ[[149]](#footnote-149) و همکاران صورت گرفت ابتدا با انجام یک سری جستجو با واژه‌های کلیدی مرتبط مجموعه متونی از پایگاه‌داده‌های زیستی گردآوری شد. با استفاده از واژه‌نامه تمامی ژن‌های مرتبط شناسایی و ارتباط‌شان نیز در قالب شبکه بازنمایی شد. گره‌های بزرگتر تعداد تکرار بیشتری در متون داشتند و یال‌های ضخیم‌تر نیز بواسطه دفعات تکرار مرتبط بودن ژن‌ها در عبارات نشان‌دهنده وزن بیشتر پیوند بودند. بعد از پیداشدن ژن‌های بیماری‌زای موثر و شبیه به یکدیگر، تمامی داروهای مرتبط از پایگاه داده DGIdb[[150]](#footnote-150) استخراج و مطالعات بعدی صرفاً روی این دسته انجام گردید[48]. در پایان یکبار دیگر درقالب جدول 2-3 بطور خلاصه به مرور روش‌های شرح داده شده می‌پردازیم.

جدول ‏2‑3 مروری بر روش‌های بازهدف‌گذاری بر مبنای رویکردهای مختلف محاسباتی

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| مطالعه | سال | مجموعه داده | نوع روش | خلاصه |
| بلیک‌لی و همکاران[35] | 2009 | ساختار شیمیایی دارو و نقشه توالی پروتئین | یادگیری‌ماشین – SVM | بازهدف‌گذاری از طریق بررسی ارتباط دارو-پروتئین بر اساس شباهت محلی با استفاده از SVM  AUC=97 , AUPR=78 |
| پنگ و همکاران[36] | 2017 | سه نوع ویژگی پروتئین و یک نوع ویژگی دارو | یادگیری‌ماشین – SVM | بازهدف‌گذاری از طریق بررسی ارتباط دارو-پروتئین. استفاده از بردار ویژگی به جای شباهت. جداسازی نمونه مشکوک از منفی قطعی  AUC=94, F-measure=92 |
| چن و همکاران[37] | 2017 | چهار نوع شباهت دارو و سه نوع شباهت بیماری | یادگیری ماشین – کمترین مربعات | بهره‌گیری از کمترین مربعات منظم جهت یافتن بهترین برازش برای تابع نگاشت فضای دارو به فضای بیماری  AUC=85 |
| کیم و همکاران[38] | 2019 | چهار نوع شباهت دارو و سه نوع شباهت بیماری | یادگیری ماشین – رگرسیون لجستیک | بازهدف‌گذاری از طریق بررسی ارتباط دارو-بیماری  وزن دهی به شباهت‌ها و استفاده از تابع هسته غیرخطی  AUC=87 , AUPR=25 |
| لو و همکاران[39] | 2016 | چهار نوع شباهت دارو | تحلیل شبکه – خوشه بندی | شناسایی داروی جدید جهت بهبود درمان نوعی خاصی از سرطان ریه  استفاده از خوشه‌بندی جهت یافتن داروهایی با ساختار یکسان |
| یان و همکاران[40] | 2016 | ساختار شیمیایی دارو و نقشه توالی پروتئین | تحلیل شبکه – انتشار برچسب محلی | بازهدف‌گذاری از طریق بررسی ارتباط دارو-پروتئین  انجام دو بار فرآیند انتشار برچسب بصورت مستقل از هر سطح بطور مجزا  AUC=99 , AUPR=92 |
| شهرضا و همکاران[41] | 2017 | دو نوع شباهت دارو و دو نوع شباهت بیماری | تحلیل شبکه – انتشار برچسب سراسری | بازهدف‌گذاری از طریق بررسی ارتباط دارو-بیماری  عدم وابستگی به نمونه‌های منفی و استفاده از ویژگی‌های محلی و سراسری در کنار یکدیگر  AUC=82 , AUPR=84 |
| کدیا و همکاران[42] | 2020 | سه نوع شباهت دارو و یک نوع شباهت بیماری | تحلیل شبکه – فاکتورگیری ماتریسی غیرمنفی | بازهدف‌گذاری از طریق بررسی ارتباط دارو-بیماری  تاثیردادن کوتاه‌ترین مسیر بین گره‌ها در هر سطح از شبکه جهت مقدار‌دهی اولیه ماتریس‌های ویژگی علاوه بر ماتریس شباهت.  AUC=93 |
| مونگیا و همکاران[43] | 2020 | دو نوع شباهت دارو و یک نوع شباهت بیماری | تحلیل شبکه – فاکتورگیری ماتریسی | بازهدف‌گذاری از طریق بررسی ارتباط دارو-بیماری  هنگام مقداردهی اولیه تنها اطلاعات K شبیه‌ترین عناصر را وارد پروفایل ویژگی هر یک می‌کند.  AUC=98 , AUPR=75 |
| لو و همکاران[44] | 2017 | چهار نوع شباهت دارو و سه نوع شباهت پروتئین | تحلیل شبکه – فاکتورگیری ماتریسی | بازهدف‌گذاری از طریق بررسی ارتباط دارو-پروتئین  استفاده از گشت تصادفی جهت یکپارچه‌سازی شباهت و PCA جهت بردارهای ویژگی ابعاد پایین  AUC=91 , AUPR=92 |
| یو و همکاران[45] | 2020 | بسته به داده متفاوت | تحلیل شبکه – تعبیه گراف | طراحی پکیج جامع بازهدف‌گذاری بر پایه تعبیه گراف به نام BioNEV  دو دسته وظیفه پیش‌بینی پیوند و طبقه‌بندی چند کلاسه در شبکه‌های مختلف زیستی قابل اجراست. |
| ژنگ و همکاران[46] | 2019 | نه نوع شباهت دارو و یک ماتریس ارتباط دارو-بیماری | یادگیری عمیق | بازهدف‌گذاری از طریق بررسی ارتباط دارو-بیماری  استفاده از از یک خودرمزگذار برای یکپارچه‌سازی و فشرده‌سازی داده‌ها  AUC=90 , AUPR=92 |
| اوزتورک و همکاران[47] | 2018 | دو نوع شباهت پروتئینی | یادگیری عمیق | بازهدف‌گذاری از طریق بررسی ارتباط دارو-پروتئین  استفاده از شبکه‌های عصبی کانولوشنی جهت نگاشت ویژگی و لایه‌های تماماً متصل جهت پیش‌بینی نهایی  AUPR =78 |
| ژانگ و همکاران[48] | 2019 | مجموعه متون بیماری | متن‌کاوی | شناسایی ژن‌های مرتبط با بیماری از طریق مقایسه با واژه‌نامه و میزان ارتباط با تحلیل فراوانی در متن |

# فصل سوم: روش پیشنه

# ادی

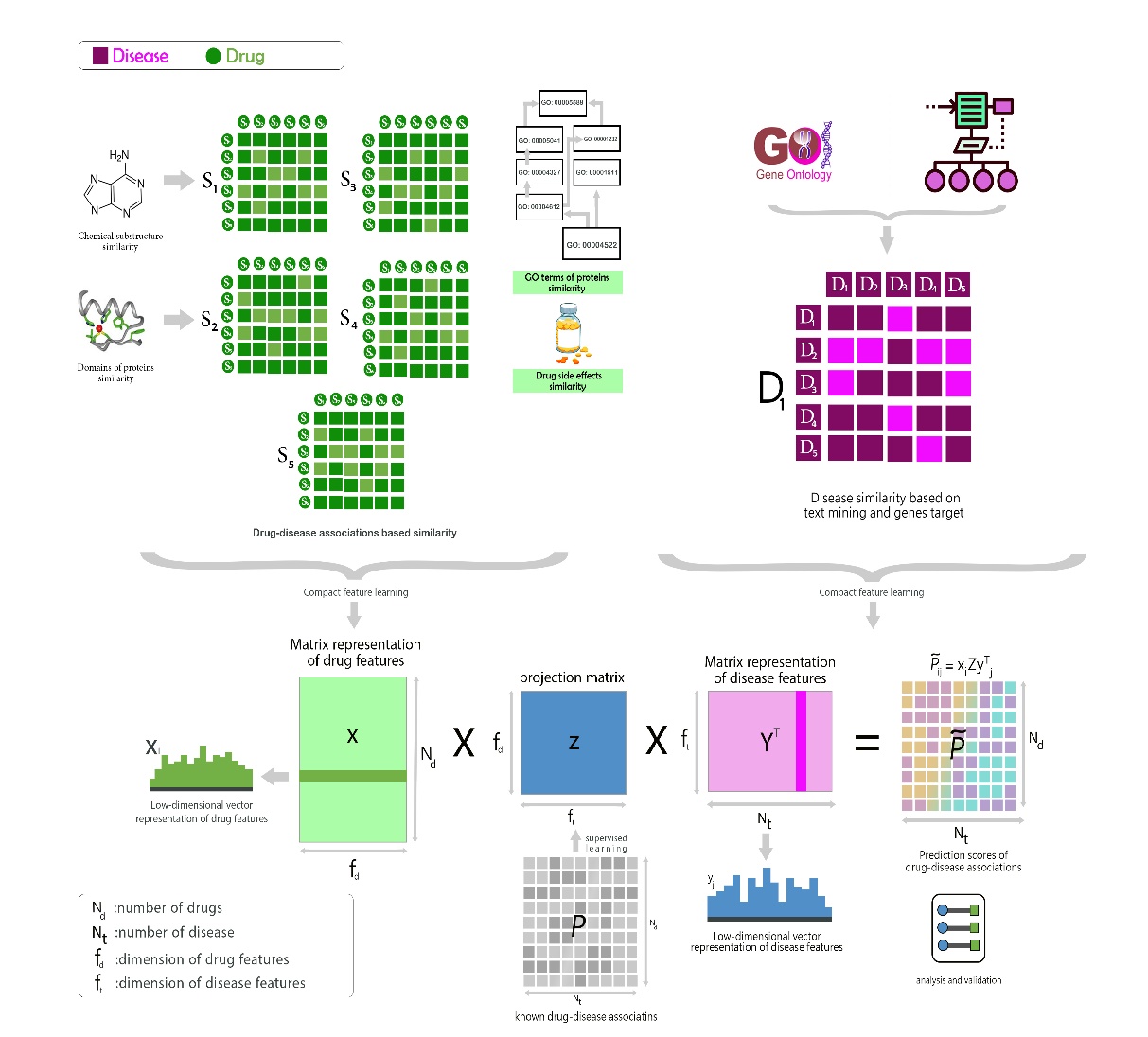
فصل سوم

روش پیشنهادی

## مقدمه

در این فصل جزئیات روش پیشنهادی برای بازهدف‌گذاری دارو با استفاده از عوارض‌جانبی (DRSE)، شرح داده می‌شود. روش ارائه شده در این مطالعه شامل استفاده از تکنیک‌های گشت تصادفی[[151]](#footnote-151) برای ترکیب ویژگی‌های مختلف دارویی و بیماری، تجزیه مؤلفه انتشار[[152]](#footnote-152) (DCA) برای استخراج ویژگی‌های مهم‌تر بوده و در نهایت از روش فاکتور‌گیری ماتریسی[[153]](#footnote-153) برای پیش‌بینی نهایی استفاده می‌کند. شکل 3-1 نمای کلی روش پیشنهادی را نشان می‌دهد.

شکل 3-1 نمای کلی روش پیشنهادی



روش پیشنهادی در گام‌های زیر خلاصه می‌شود:

* جمع‌آوری ویژگی‌های مختلف
* محاسبه ماتریس شباهت از روی ویژگی‌ها با استفاده از تابع جاکارد
* محاسبه ماتریس احتمالات گشت تصادفی برای هر ماتریس شباهت
* خلاصه‌سازی ماتریس‌های حاصل از گشت‌تصادفی و ایجاد یک ماتریس ترکیبی با ابعاد پایین[[154]](#footnote-154) برای دو مجموعه‌ی دارو و بیماری
* استفاده از فاکتورگیری ماتریس برای پیش‌بینی احتمال پیوندهای دارو و بیماری

## مجموعه داده

بخشی از داده‌های استفاده شده در این پژوهش، از مطالعه لیانگ[[155]](#footnote-155) و همکاران دریافت شده است[49]. محققین این پژوهش داده‌های مورد نیاز را از پایگاه‌داده‌های معتبری چون PubChem[[156]](#footnote-156)، InterPro[[157]](#footnote-157)، UniProt[[158]](#footnote-158) و [[159]](#footnote-159)DincRNA دریافت و پاک‌سازی کرده‌اند. این مجموعه ‌داده شامل سه نوع ماتریس ویژگی برای داروها، یک ماتریس شباهت تجمیع شده و ماتریس پیوندهای دارو-بیماری می‌باشد:

* ویژگی ساختار شیمیایی که از پایگاه داده PubChem دریافت شده است[50].
* ویژگی هستان‌شناسی ژن‌های مرتبط با پروتئین‌های هدف که از پایگاه داده InterPro دریافت شده است[51].
* ویژگی ساختار سوم و کروی کمپلکس پروتئین‌های هدف که از پایگاه داده UniProt ‌دریافت شده است[52].
* ویژگی‌ ژن‌های مرتبط هر بیماری که از پایگاه‌داده DincRNA دریافت شده است[53].
* ویژگی معنایی بیماری‌ها که بوسیله متن‌کاوی مقالات موجود در وب و مرتبط با هر بیماری استخراج شده است.
* ماتریس پیوندهای دارو-بیماری

در پژوهش ما علاوه بر مجموعه داده بالا و در جهت بهبود عملکرد فرآیند بازهدف‌گذاری، از ویژگی چهارمی نیز به نام عوارض‌جانبی دارو استفاده گردید. ابتدا مجموعه کلی داروهای قرار داده شده در پایگاه‌داده SIDER[[160]](#footnote-160) دریافت شد و در گام بعدی صرفاً ویژگی‌ مربوط به داروهای موجود از میان داده‌ها استخراج گردید و به فرمت قابل بهره‌برداری تبدیل گردید[54].

جدول 3-1 انواع ویژگی‌ها در مجموعه‌داده

|  |  |  |  |
| --- | --- | --- | --- |
| نوع گراف | تعداد راس طرف اول | تعداد راس طرف دوم | تعداد یال |
| دارو - ساختار شیمیایی | ۷۶۳ | ۶۲۳ | ۹۶۲۰۲ |
| دارو – هستان‌شناسی ژن مرتبط با پروتئین‌های هدف | ۷۶۳ | ۴۴۴۷ | ۵۵۹۱۸ |
| دارو - ساختار کروی کمپلکس پروتئین‌های هدف | ۷۶۳ | ۱۴۲۵ | ۷۵۹۳ |
| بیماری – ژن | ۶۸۱ | 386 | - |
| دارو - بیماری | ۷۶۳ | ۶۸۱ | ۳۰۵۱ |
| دارو - عوارض جانبی | ۷۶۳ | ۳۰۴۶ | ۹۷۳۰ |

در ماتریس‌های ویژگی به تعداد دارو/بیماری سطر، به تعداد ویژگی‌های خاص (به عنوان مثال تعداد پروتئین‌های هدف) ستون داشته و مقادیر هر درایه نیز صفر یا یک (به عنوان مثال به معنای اینکه هر پروتئین، هدف این دارو قرار می‌گیرد (۱) یا خیر (۰)) است. به طور مثال در مورد ماتریس ویژگی دارو-عوارض جانبی نیز ما به تعداد دارو‌ها سطر داشته و لیستی از تمام عوارض جانبی همه داروها شامل عطسه، سرفه و غیره را به‌عنوان ستون‌های ماتریس داریم. مقادیر یک و صفر هر درایه به ترتیب مشخص کننده داشتن/نداشتن یک عارضه جانبی خاص برای یک داروی خاص می‌باشد. در این حالت می‌توان مجموعه‌‌داده‌ها را به‌صورت گراف‌ دو‌بخشی مدل کرد که خلاصه‌ای از آن در جدول 3-1 آورده شده است.

شایان ذکر است درخصوص بیماری‌ها دو ماتریس ویژگی ذکر شده به صورت مستقیم در اختیار نبوده و از ماتریس شباهت تجمیع شده حاصل از دو گونه ویژگی مذکور مقاله لیانگ و همکاران استفاده شده است[49].

علاوه بر ویژگی‌های بالا، ماتریس شباهت بر اساس ماتریس پیوندهای دارو-بیماری نیز محاسبه و در نظر گرفته شده است. بخش زیر به چگونگی محاسبه ماتریس شباهت بر اساس ویژگی‌های گفته شده می‌پردازد.

## محاسبه‌ی ماتریس‌های شباهت

ویژگی‌های گفته شده در بالا، با استفاده از تابع جاکارد با فرمول زیر به ماتریس شباهت تبدیل می‌شوند.

|  |  |
| --- | --- |
| (2) |  |

دو بردار ویژگی A و B را در نظر بگیرید که هر کدام شامل n درایه با مقادیر ۰ با ۱ هستند.

* به معنای تعداد درایه‌هایی است که در هر دو بردار A و B مقدار ۱ دارد.
* به معنای تعداد درایه‌هایی است که در بردار A برابر ۰ و در B برابر ۱ است.
* به معنای تعداد درایه‌هایی است که در A برابر ۱ و در B برابر ۰ است.

تابع جاکارد برای هر جفت دارو، بردار دودویی ویژگی مربوط به آن دارو‌ها را دریافت می‌کند و با استفاده از فرمول بالا میزان شباهت آن‌ها را محاسبه می‌کند. به عنوان مثال برای محاسبه شباهت دو دارو و بر اساس ویژگی پروتئین‌های هدف، سطر ام و سطر ام ماتریس ویژگی پروتئینی به عنوان ورودی به تابع جاکارد داده می‌شود.

## ایجاد ویژگی‌های خلاصه‌شده برای دارو و بیماری

به محاسبه ویژگی‌های با ابعاد کمتر از روی ویژگی‌‌های اولیه، یادگیری ویژگی‌های خلاصه[[161]](#footnote-161) می‌گویند. روش پیشنهادی از تجزیه و DCA[55] برای یادگیری و ایجاد نمایش‌های برداری با ابعاد کوچکتر از ویژگی‌های اصلی ماتریس‌های شباهت، استفاده می‌کند. این الگوریتم که اخیراً توسعه یافته، ترکیبی از روش انتشار شبکه[[162]](#footnote-162) (گشت تصادفی با راه‌اندازی مجدد) و روش کاهش ابعاد[[163]](#footnote-163) بوده و قادر است خصوصیات توپولوژیکی یک شبکه را نیز استخراج کند[56].

### گشت تصادفی با راه‌اندازی مجدد

گشت تصادفی با راه‌اندازی مجدد[[164]](#footnote-164) (RWR)، یک الگوریتم انتشار شبکه است که به‌طور گسترده برای تجزیه و تحلیل داده‌های پیچیده‌ی شبکه بیولوژیکی استفاده شده است[57]–[61]. برخلاف روش‌های گشت تصادفی معمولی، RWR یک احتمال از پیش تعریف شده‌ای برای راه‌اندازی مجدد در گره اولیه در هر تکرار را معرفی می‌کند که می‌تواند الگوی اتصال توپولوژیکی محلی و سراسری در شبکه را در نظر بگیرد تا به طور کامل از روابط مستقیم یا غیرمستقیم زمینه‌ای بین گره‌ها بهره ببرد. به بیان دقیق‌تر ، فرض کنید ماتریس مجاورت یک شبکه شباهت داروها (یا بیماری‌ها) باشد. ما نیز ماتریس دیگری را تعریف می‌کنیم، که در آن هر عنصر احتمال انتقال از گره‌ به گره را نشان می‌دهد، یعنی:

|  |  |
| --- | --- |
| (3) |  |

سپس فرض کنید بردار توزیع با درایه باشد که در آن هر عنصر احتمال بازدید یک را پس از تکرار گشت تصادفی با شروع از گره را ذخیره می‌کند. بنابراین RWR از گره را می‌توان به‌صورت زیر تعریف کرد:

|  |  |
| --- | --- |
| (4) |  |

در این رابطه، نشان‌دهنده بردار پایه استاندارد بعدی است که و . هم‌چنین به معنای احتمال راه‌اندازی مجدد از پیش تعریف شده است، که در واقع تأثیر نسبی بین اطلاعات توپولوژیکی محلی و سراسری را در فرآیند انتشار را کنترل می‌کند. همچنین به ازای مقادیر بالاتر تاثیر ساختارهای محلی شبکه زیادتر می‌شود. پس از ثابت شدن روند تکرار، می‌توانیم توزیع ایستا[[165]](#footnote-165) را به‌دست آوریم، که ما به آن «حالت انتشار[[166]](#footnote-166)» برای گره (یعنی) می‌گوییم[55]. به طور شهودی، عنصر ام حالت انتشار، که با نشان داده می‌شود، بیان‌گر احتمال شروع آن است که گشت تصادفی با راه‌اندازی مجدد از گره شروع شود و در حالت ایستا در گره پایان یابد. دو گره دارای حالت انتشار مشابه، به طور کلی نشان‌دهنده‌ی آن است که آن‌ها نسبت به سایر گره ها در شبکه دارای موقعیت‌های مشابهی هستند و بنابراین احتمالاً عملکردهای مشابهی دارند.

به بیان ساده، گشت تصادفی با راه‌اندازی مجدد، برای هر ماتریس شباهت یک گراف ایجاد می‌کند که وزن‌ یال‌ها میزان شباهت است، به عنوان مثال می‌تواند شباهت بین داروها (یا بیماری‌ها) باشد. گشت تصادفی از هر دارو (یا بیماری) شروع شده و با احتمال شباهت به گره‌های دیگر می‌رود و با احتمال مشخصی در هر مرحله به گره شروع باز‌می‌گردد. در نهایت بعد از پایان کار گشت تصادفی و رسیدن به حالت تعادل، به ازای هر ماتریس شباهت، یک ماتریس احتمالاتی که سطر‌ها و ستون‌های آن داروها (یا بیماری‌ها) هستند حاصل می‌شود. به عنوان مثال سطر ام این ماتریس نشان می‌دهد که چقدر احتمال دارد که از دارو (یا بیماری) ام شروع کنیم و در حالت تعادل، در داروهای (یا بیماری‌های) دیگر قرار بگیریم.

گشت تصادفی برای هر 5 ماتریس شباهت دارویی اجرا می‌شود و نتایج حاصل از گشت تصادفی که در این مرحله شامل اطلاعات ساختاری شبکه نیز هست، به‌دست می‌آید. ابعاد هر یک از این ماتریس‌ها، مشابه ابعاد ماتریس‌های اولیه یعنی تعداد دارو در تعداد دارو و یا تعداد بیماری در تعداد بیماری است.

### چارچوب کاهش ابعاد

حالت‌های انتشار ناشی از فرآیند RWR فوق الذکر، ممکن است به دلیل کیفیت نسبتاً پایین و ابعاد زیاد داده‌های زیستی کاملاً دقیق نباشد. فقدان اطلاعات یا اطلاعات اشتباه در شبکه، حتی به مقدار کم هم می‌تواند به طرز قابل توجهی بر نتایج فرآیند انتشار اثر بگذارد[62]. علاوه بر این، به طور کلی استفاده از داده با ابعاد بالا به عنوان ویژگی‌های ورودی کار مناسبی نیست. برای پرداختن به این مسئله، روش پیشنهادی از یک تکنیک کاهش ابعاد، به نام تجزیه و تحلیل مؤلفه انتشار DCA، استفاده می‌کند تا ابعاد فضای ویژگی‌ها را کاهش داده و ویژگی‌های مهم توپولوژیکی حالت انتشار را به‌دست آورد. به منظور کاهش نویز و کاهش ابعاد، DCA هر حالت انتشار را با یک مدل لجستیک چندجمله‌ای[[167]](#footnote-167) بر اساس یک بردار فضای پنهان[[168]](#footnote-168) که ابعاد آن بسیار کمتر از بردار اصلی بعدی ( تعداد ویژگی‌های اصلی است) و نمایانگر حالت‌های انتشار است، نشان می‌دهد. به طور خاص، احتمال اختصاص داده شده به گره در حالت انتشار گره اکنون به طور زیر مدل می‌شود:

|  |  |
| --- | --- |
| (5) |  |

که . به ویژگی زمینه گره و به ویژگی گره گفته می‌شود و هر دو، ویژگی‌های توپولوژیکی شبکه را توصیف می‌کنند. اگر و در فضا، هم‌سو باشند و ضرب داخلی آن‌ها عدد بزرگی ‌شود، به این معنا است که در یک گشت تصادفی با شروع از گره ، احتمال بازدید مکرر گره زیاد است. DCA مجموعه‌ای از حالت‌های مشاهده شده را به عنوان ورودی در نظر می‌گیرد. در این روش با استفاده از آنتروپی نسبی (واگرایی کولبک-لیبر[[169]](#footnote-169)KL) به دنبال کاهش ورودی به برداری با ابعاد کمتر است به طوری که توزیع ورودی با خروجی دارای کمترین تفاوت باشد. این عمل برای تمام ها و ها توسط کمینه‌سازی تابع زیر انجام می‌شود:

|  |  |
| --- | --- |
| (6) |  |

که نشان‌گر واگرایی KL بین دو توزیع است. چارچوب DCA برای بهینه‌سازی این تابع از یک روش شبه-نیوتن استاندارد به نام L-BFGS[63] استفاده می‌کند.

### یکپارچه‌سازی اطلاعات ماتریس‌ شباهت‌های مختلف

به دلیل اینکه ما در این مطالعه از چند نوع ماتریس شباهت برای دارو‌ها استفاده می‌کنیم به منظور در نظر گرفتن تمام ماتریس‌های شباهت در فرآیند کاهش ابعاد، چارچوب کاهش ابعاد فوق برای ادغام چندین ماتریس شباهت دارویی گسترش داده می‌شود. فرض کنیم که ماتریس شباهت دارویی داریم:

برای ترکیب شبکه‌ها با اطلاعات متنوع (ماتریس‌های شباهت بر اساس ویژگی‌های مختلف) ابتدا RWR بر روی شبکه‌ها به طور جداگانه اجرا شده و سپس حالت‌‌های انتشار شبکه برای هر نود در شبکه ، از طریق فرمول زیر به‌دست می‌آید:

|  |  |
| --- | --- |
| (7) |  |

این فرمول مشابه فرمول (۴) می‌باشد؛ با این تفاوت که به ازای هر متغیر است. در واقع، بردار ویژگی زمینه گره برای هر شبکه با هر گره با و بردارهای ویژگی در بین تمام شبکه‌ها مشترک هستند. در نهایت، DCA تابع هدف زیر را بهینه می‌کند:

|  |  |
| --- | --- |
| (8) |  |

این تابع را نیز می‌توان با روش شبه نیوتن L-BFGS حل کرد[63]. اگرچه میزان واگرایی هر کدام از شبکه‌ها در عملکرد تابع هدف فوق دارای وزن‌های برابر است، اما برای تأکید بر اهمیت نسبی هر کدام از شبکه‌ها، می‌توان وزن‌های متفاوتی را لحاظ کرد.

برای اینکه چارچوب DCA برای شبکه‌های بزرگ بیولوژیکی مقیاس‌پذیرتر باشد، در روش پیشنهادی از یک نوع DCA به نام clusDCA[64] استفاده می‌‌کنیم که دارای یک تابع هدف جایگزین است و می‌تواند براساس تجزیه مقدارهای منفرد[[170]](#footnote-170) (SVD) بهینه‌سازی شود. به طور خلاصه، برای هر دارو یا بیماری، clusDCA قادر به یادگیری یک بردار با ابعاد کم است که با یک روش مناسب، تفاوت توزیع حالات انتشار با خروجی را به حداقل می رساند و توزیع بردار خروجی مدل را تحت نُرم L2 در فضای لگاریتمی یاد می‌گیرد[64].

در این مطالعه احتمال راه‌اندازی مجدد برای گشت تصادفی برابر با ۰.۵ در نظر گرفته شده است. هم‌چنین با توجه به نتایج آزمایش‌ها تحقیقات پیشین[44] مقادیر ابعاد ویژگی‌های پنهان برای دارو و برای بیماری در نظر گرفته شده است که برابر با 10٪ و 15٪ ابعاد اصلی می‌باشد. علاوه بر این، بیشینه تعداد تکرارهای روش پییشنهادی ۲۰ تکرار تعیین شده است.

## فرآیند بهینه سازی DCA

برای سادگی کار، در این بخش فرآیند بهینه‌سازی DCA را برای یک شبکه ورودی نشان می‌دهیم. بهینه‌سازی DCA برای شبکه‌های متعدد، با تعمیم این روش صورت می‌گیرد. تابع زیر را در نظر بگیرید:

|  |  |
| --- | --- |
| (9) |  |

ما ابتدا فرمول را بر حسب و بر اساس تعریف KL و بیان می‌کنیم، یعنی:

|  |  |
| --- | --- |
| (10) |  |

که نشان‌دهنده آنتروپی است. حال گرادیان این تابع هدف را با توجه به پارامترهای محاسبه می‌کنیم:

|  |  |
| --- | --- |
| (11) |  |
| (12) |  |

این تابع هدف را می‌توان با استفاده از روش شبه نیوتن L-BFGS استاندارد حل کرد تا بردارهای ابعاد پایین محاسبه گردد. در طول آزمایشات ما، مقادیر اولیه بردارهای   
با مقادیر تصادفی یکنواخت در بازه مقداردهی شدند.

همان‌طور که در پیش‌تر ذکر شد، برای اینکه روش DCA برای شبکه‌های بزرگ بیولوژیکی مقیاس‌پذیرتر باشد، ما از یک رویکرد مبتنی بر فاکتورگیری ماتریس کارآمدتر برای تجزیه حالت‌های انتشار و به دست آوردن بردار‌های ابعاد پایین، استفاده می‌کنیم. طبق تعریف داریم:

|  |  |
| --- | --- |
| (13) |  |

عبارت اول در معادله فوق مطابق با تقریب بعد پایین ، و عبارت دوم یک عامل نرمال‌سازی است که تضمین می‌کند توزیع به خوبی تعریف شده باشد. محدودیتی که الزام می‌کند که مجموع درایه‌های برابر یک باشد را با حذف عبارت دوم، آرام‌سازی[[171]](#footnote-171) می‌کنیم، یعنی:

|  |  |
| --- | --- |
| (14) |  |

علاوه بر این، به جای به کمینه کردن آنتروپی نسبی میان حالت انتشار اصلی و تقریبی، از مجموع مربعات خطا به عنوان تابع هدف استفاده می‌کنیم:

|  |  |
| --- | --- |
| (15) |  |

این تابع هدف را می‌توان با SVD بهینه سازی کرد. برای جلوگیری از پیشامد لگاریتم صفر، یک ثابت مثبت کوچک به اضافه می‌کنیم و محاسبه ماتریس حالت انتشار لگاریتم که با نماد *L* نشان می‌دهیم، به صورت زیر محاسبه می‌شود:

|  |  |
| --- | --- |
| (16) |  |

که ، و از کنار هم گذاشتن به‌دست آمده است. سپس، ماتریس را با استفاده از SVD به سه ماتریس تجزیه می‌کنیم:

|  |  |
| --- | --- |
| (17) |  |

برای به‌دست آوردن بردارهای ابعاد پایین و با ابعاد ، به راحتی اولین بردارهای تکین در و اولین مقادیر تکین در را انتخاب می‌کنیم. به طور دقیق تر، فرض کنید ماتریسی باشد که در آن، هر سطر نمایان‌گر بردار ویژگی با ابعاد پایین مربوط به هر گره در شبکه باشد و ماتریسی باشد که در آن، هر ردیف بردار مربوطه از ویژگی‌های زمینه را نشان ‌دهد. بنابراین، و را می‌توان به صورت زیر محاسبه کرد:

|  |  |
| --- | --- |
| (18) |  |

برای ادغام داده‌های شبکه ناهمگن، بهینه‌سازی DCA تک شبکه‌ای که در بالا توضیح داده شد را به نوع چند شبکه تعمیم می‌دهیم. فرض کنید مجموعه‌ای از ماتریس‌های حالت انتشار لگاریتمی بر اساس مجموعه‌ای از حالت‌های انتشار مبتنی بر شبکه‌ ورودی باشد. بنابراین، تابع هدف زیر را بهینه می‌کنیم:

|  |  |
| --- | --- |
| (19) |  |

که بردار ویژگی به گره در شبکه اختصاص دارد و ویژگی‌های گره در تمام شبکه‌ مشترک است. این تابع هدف نیز می‌تواند توسط SVD بهینه شود.

## پیش بینی از فضای دارویی بر روی فضای بیماری

ما از بردار‌های ابعاد پایین ویژگی‌های دارویی و بیماری که از یادگیری ویژگی‌های خلاصه به‌دست آمده‌اند، برای پیش‌گویی پیوندهای جدید بیماری و دارو استفاده می‌کنیم. با توجه به این شهود که نزدیکی هندسی در فضای بردار ویژگی ممکن است ارتباط عملکردی را منعکس کند، ما از یک رویکرد تکمیل ماتریس برای دستیابی به یک ماتریس نشان‌دهنده‌ی بردارهای ویژگی ابعاد پایین استفاده کرده‌ایم که ویژگی‌های فضای دارو را بر فضای بیماری تصویر می‌کند[65]. بردارهای ویژگی دارویی پیش‌گویی شده از لحاظ هندسی به ارتباطات شناخته شده با بیماری مرتبط خود نزدیک هستند.

به زبان ریاضی، ما از برای نشان دادن ماتریس ویژگی‌های دارویی (که هر سطر ماتریس، نشان‌دهنده‌ی ویژگی داروی است) و از () برای نشان دادن ماتریس ویژگی‌های بیماری (که هر سطر ماتریس، نشان‌دهنده‌ی ویژگی بیماری است) استفاده می‌کنیم. لازم به ذکر است که و به ترتیب نشان‌دهنده‌ی تعداد دارو و تعداد بیماری‌هاست. فرض کنید نشان‌دهنده ماتریس ارتباطات دارو و بیماری‌ها باشد؛ به طوری که نشان‌دهنده وجود پیوند بین دارو و بیماری و نشان‌دهنده عدم چنین پیوندی باشد. برای یادگیری ماتریس پیش‌بینی ، یک تابع دوسویه[[172]](#footnote-172) تعریف شده است تا پیوندهای بیماری-دارویی ناشناخته در (یعنی آن مقادیر صفر در ماتریس ارتباطات) را پیش‌گویی کند. به بیان دقیق‌تر، تابع دوسویه به صورت زیر تعریف می‌شود:

|  |  |
| --- | --- |
| (20) |  |

که مخفف ماتریس پیوند دارو- بیماری است و از مرحله یادگیری ویژگی‌های خلاصه‌شده (به عنوان مثال، انتشار شبکه و فرآیندهای کاهش ابعاد شبکه) و ماتریس طرح[[173]](#footnote-173) است که باید آموزش داده شود. سپس از فرمول زیر استفاده می‌کنیم تا مقدار احتمال پیوند بین جفت دارو و بیماری را محاسبه کنیم.

|  |  |
| --- | --- |
| (21) |  |

هر چه عدد بزرگ‌تر باشد، به این معنا است که داروی با احتمال بیشتری با بیماری پیوند دارد.

ماتریس طرح‌ دارای ابعاد است، به طور معمول بین آن دسته از بردار‌های دارو یا بیماری‌ که از نظر هندسی در فضا نزدیک هستند، هم‌بستگی قابل‌توجهی وجود دارد که در نتیجه می‌تواند تعداد پارامترهای مؤثر مورد نیاز برای مدل‌سازی اثرات متقابل دارو را کاهش دهد. به همین دلیل، ما یک محدودیت رتبه پایین[[174]](#footnote-174) روی اعمال می‌کنیم تا فقط تعداد کمی از فاکتور‌های پنهان از طریق تجزیه مرتبه پایین که محاسبه شوند. این محدودیت رتبه پایین نه تنها مشکل بیش برازش[[175]](#footnote-175) را حل می‌کند بلکه از نظر محاسباتی روند بهینه‌سازی را بهبود می‌دهد[66].

مسئله‌ی بهینه‌سازی با محدودیت رتبه پایین در ماتریس طرح اولیه ، یک مسئله NP-سخت[[176]](#footnote-176) است. یک آرام‌سازی استاندارد از محدودیت درجه پایین، به حداقل رساندن نرم ردیابی (یعنی مجموع مقادیر منفرد) ماتریس است، که معادل کمینه‌سازی نرم‌های فروبنیوس[[177]](#footnote-177) است. بنابراین، با حل کردن مسئله بهینه‌سازی زیر می‌توان فاکتورگیری Z به G و H را انجام داد:

|  |  |
| --- | --- |
| (22) |  |

که در آن پارامتر ، پارامتر منظم‌سازی[[178]](#footnote-178) است که تعادل بین کمینه‌سازی مربعات خطا در جفت پیوندهای شناخته شده و نرم فروبنیوس را کنترل می‌کند. مسئله بهینه‌سازی را می‌توان با کمینه‌سازی متناوب[[179]](#footnote-179) حل کرد[65].

روش پیشنهادی با نرم‌افزار متلب نسخه R1028b بر روی کامپیوتری با 7 پردازنده، هشت گیگابایت حافظه‌اصلی و سیستم عامل اوبونتو اجرا شده است.

# فصل چهارم: نتایج پژوهش

فصل چهارم

نتایج پژوهش

## مقدمه

در این فصل نتایج حاصل از ارزیابی روش پیشنهادی (DRSE) گزارش می‌شود و همچنین نتایج آن با روش‌های پیشین مقایسه می‌شود. همچنین درباره‌ی برخی از ارتباطات پیش‌بینی شده توسط روش پیشنهادی بررسی موردی انجام شده است.

## روش ارزیابی

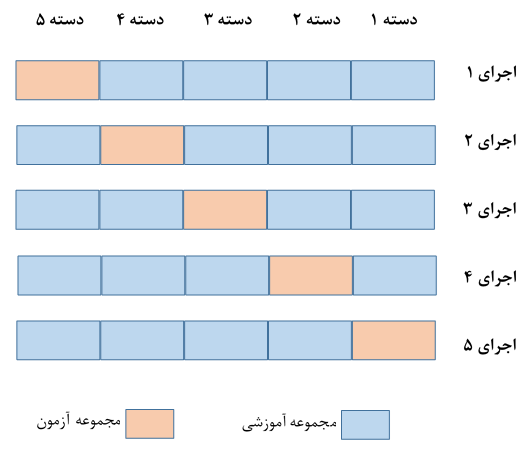
در این مطالعه، برای ارزیابی مدل از روش اعتبارسنجی متقاطع طبقه‌ای[[180]](#footnote-180) و معیارهای اصلی دسته‌بندی استفاده می‌کنیم. این مفاهیم در بخش‌های زیر به تفصیل توضیح داده شده‌اند.

### اعتبارسنجی متقاطع

معمولا برای ارزیابی از اعتبارسنجی متقاطع[[181]](#footnote-181) ۵ دسته‌ای استفاده می‌شود. در این روش داده‌ها به ۵ دسته مساوی تقسیم می‌شوند و روش پیشنهادی با ۴ دسته از داده‌ها آموزش می‌بیند و با یک دسته دیگر مورد آزمون قرار می‌گیرد. معیار ارزیابی برای دسته آزمون محاسبه می شود. این فرآیند تکرار می‌شود تا جایی که تمام بخش‌ها یک بار به عنوان داده آموزشی لحاظ شوند. در نهایت میانگین، معیارهای محاسبه شده در تمام دسته‌های آزمون، به عنوان عملکرد کلی مدل در نظر گرفته می‌شود. روند کلی اعتبارسنجی متقاطع ۵ دسته‌ای در شکل 4-1 به نمایش درآمده است. لازم به ذکر است با توجه به اینکه تعداد نمونه‌های با برچسب صفر و یک در مجموعه‌داده یکسان نیست از همین رو، از تعمیم خاصی از اعتبارسنجی متقاطع، به نام اعتبارسنجی متقاطع طبقه‌ای در این مطالعه استفاده شده است.

### اعتبارسنجی متقاطع طبقه‌ای

اعتبارسنجی متقاطع طبقه‌ای ۵ دسته‌ای، حالت اصلاح‌شده اعتبارسنجی متقاطع ۵ دسته‌ای است و برای مواردی که داده‌های با برچسب صفر و یک نامتوازن هستند، روش ارزیابی مناسب‌تری است و از بروز بیش‌برازش جلوگیری می‌کند. تفاوت اعتبارسنجی متقاطع طبقه‌ای با اعتبارسنجی متقاطع در آن است که در مدل طبقه‌ای، وقتی داده‌ها به ۵ دسته تقسیم‌ می‌شوند، تقسیم داده‌ها به نحوی صورت می‌گیرد که نسبت داده‌های با برچسب صفر به داده‌های با برچسب یک در تمام دسته‌ها تقریبا برابر باشد. بنابراین، داده‌های با برچسب صفر به صورت یکنواخت در بین دسته‌ها توزیع می‌شوند؛ همین امر برای داده‌های با برچسب یک نیز صادق است.



شکل 4-1 اعتبارسنجی متقاطع ۵ دسته‌ای

### معیارهای ارزیابی

با‌توجه به این‌که مسئله مورد بررسی در این مطالعه، دسته‌بندی است، از معیارهای رایج دسته‌بندی برای ارزیابی آن استفاده می‌شود. برای محاسبه معیارهای دسته‌بندی از یک جدول درهم‌ریختگی[[182]](#footnote-182) مانند جدول 4-1 می‌شود. چهار کمیت توسط این جدول معرفی می‌شود:

* موارد مثبت صحیح[[183]](#footnote-183)، تعداد زوج دارو-بیماری‌هایی است که مدل، آن‌ها را به عنوان زوج دارای پیوند پیش‌بینی کرده و آن دارو و بیماری واقعا با هم ارتباط داشته‌اند.
* موارد مثبت کاذب[[184]](#footnote-184)، نشان‌دهنده تعداد زوج دارو-بیماری‌هایی است که به عنوان زوج دارای پیوند پیش‌بینی شده‌اند اما واقعا ارتباطی بین این زوج گزارش نشده است.
* موارد منفی صحیح[[185]](#footnote-185) تعداد زوج دارو-بیماری‌هایی است که واقعا ارتباطی بین این زوج گزارش نشده است و مدل، هم پیوندی بین آن‌ها پیش‌بینی نکرده است.
* موارد منفی کاذب[[186]](#footnote-186)، تعداد زوج دارو-بیماری‌هایی است که واقعا با هم ارتباط داشته‌اند اما مدل، پیوندی بین آن‌ها پیش‌بینی نکرده است.

جدول 4-1 درهم‌ریختگی

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **ارتباطات واقعی** | |
|  |  | **پیوند** | **عدم پیوند** |
| **پیش‌بینی مدل** | **پیوند** | TP | FP |
| **عدم پیوند** | FN | TN |

معیار‌های اساسی روش‌های دسته‌بندی، شامل معیارهای زیر است:

|  |  |  |  |
| --- | --- | --- | --- |
| (23) |  | | |
| (24) |  | | | |
| (25) |  | | | |
| (26) | |  |

به معنای درصد پیوند و عدم پیوندی است که به درستی شناسایی شده‌اند. ، به معنای آن است که از بین پیوندهای پیش‌بینی شده، چه درصدی از آن صحیح پیش‌بینی شده‌اند؛ درحالی‌که به معنای درصدی از پیوندهای واقعی است که توسط مدل شناسایی می‌شوند. با توجه به آن‌که دقت و حساسیت گاها اثر معکوس بر هم دارند، استفاده از معیار که میانگین هارمونیک آن‌ها است، معقولانه‌تر به نظر می‌رسد. لازم به ذکر است که مقادیر محاسبه شده برای ، ، و وابسته به مقدار حد آستانه هستند.

ذکر این نکته نیز حائز اهمیت است که وقتی پیوند بین یک جفت دارو-بیماری صفر گذاشته می‌شود، به معنای آن است که تاکنون شاهدی برای ارتباط بین آن‌ها وجود نداشته است؛ درحالی‌که ممکن است آن‌ها واقعا به هم مرتبط باشند. بنابراین در این نوع مسئله، نمی‌توان تعداد منفی صحیح و مثبت کاذب را به دقت شمارش کرد. روند آموزش همواره نیاز به هر دو نوع داده صفر و یک دارد، بنابراین برخی از زوج‌هایی که در روند آموزشی مدل صفر در نظر گرفته می‌شوند، ممکن است واقعا صفر نباشند.

با توجه به آ‌ن‌که مقادیر معیارهای فوق به مقدار حد آستانه وابسته است، در ارزیابی مدل‌ها از دو معیار جامع‌تر نیز استفاده می‌شود. یکی AUC که سطح زیر نمودار مشخصه عملکرد[[187]](#footnote-187) است. نمودار مشخصه عملکرد، نموداری است که محور افقی آن FPR[[188]](#footnote-188) و محور عمودی آن TPR[[189]](#footnote-189) است که به صورت زیر تعریف می‌شوند.

|  |  |
| --- | --- |
| (27) |  |
| (28) |  |

معیار دیگر، AUPR است که سطح زیر نمودار دقت-حساسیت[[190]](#footnote-190) است که محور افقی آن حساسیت و محور عمودی آن دقت است. در واقع این دو معیار، عملکرد مدل را مستقل از مقدار حد آستانه ارزیابی می‌کنند. در مواری که نسبت نمونه‌های صفر و یک برابر نباشد، معیار AUPR سنجش عادلانه‌تری برای ارزیابی مدل است.

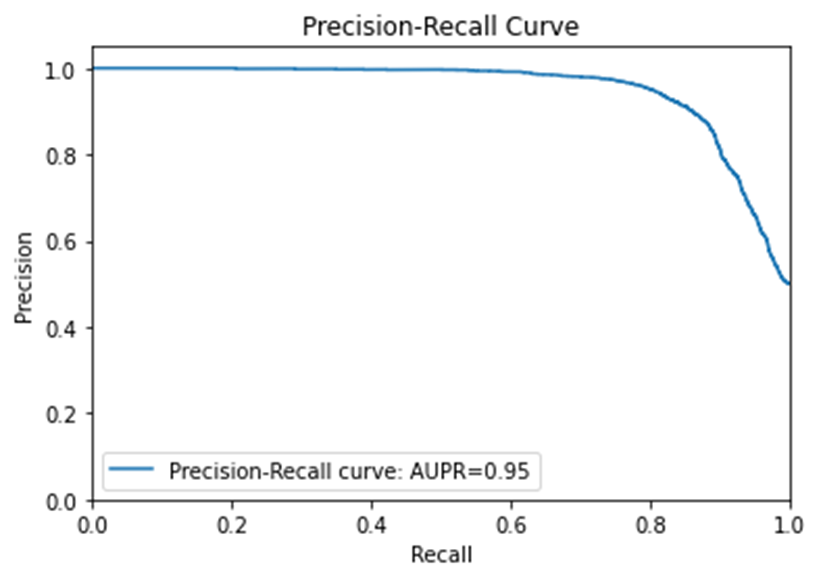
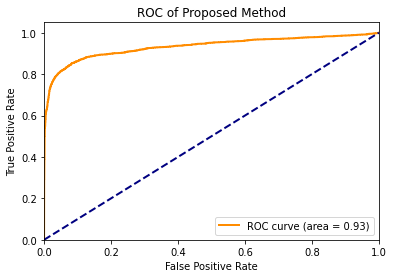
## نتایج روش پیشنهادی

به منظور یک ارزیابی دقیق روش پیشنهادی در دو حالت مورد بررسی قرار گرفت. در حالت اول، روش پیشنهادی بر روی تک تک منابع داده‌ای(ماتریس‌های شباهت) بصورت مجزا اجرا شد. در حالت بعد روش پیشنهادی بر روی ترکیبی از‌ ماتریس‌های شباهت اجرا و مورد نتایج مورد بررسی قرار گرفت. جدول 4-2 نتایج را برای هر ویژگی نشان می‌دهد. همان‌طور که در این جدول نشان داده شده است، ماتریس شباهت‌های ارتباطات دارو و بیماری و هستان‌شناسی ژن از دیگر ویژگی‌ها بهتر عمل کرده‌اند. علاوه بر این، ترکیب ویژگی‌های مختلف منجر به بهترین نتیجه شده است و نیز کارایی نحوه خلاصه‌سازی و کاهش بعد بوسیله DCA و یکپارچه‌سازی و تجمیع شباهت‌ها بوسیله‌ RWR را اثبات می‌کند. حالت ترکیب ویژگی‌ها + شباهت بر اساس ماتریس ارتباطات دارو و بیماری منجر به بهترین مقدار AUC و AUPR شده است.

جدول 4-2 ارزیابی روش پیشنهادی بر روی ماتریس‌های شباهت مختلف

|  |  |  |
| --- | --- | --- |
| **AUC** | **AUPR** | **نام روش** |
| ۸۳.۶۰ | ۸۷.۲۸ | ماتریس شباهت بر اساس ساختار شیمیایی: D1 |
| ۸۴.۴۲ | ۸۸.۴۵ | ماتریس شباهت بر اساس پروتئین‌های هدف: D2 |
| ۸۵.۳۰ | ۸۹.۱۳ | ماتریس شباهت بر اساس هستان‌شناسی ژن: D3 |
| ۸۰.۵۸ | ۸۵.۴۲ | ماتریس شباهت بر اساس عوارض جانبی: D4 |
| ۹۲.۰۷ | ۹۳.۴۷ | شباهت بر اساس ماتریس ارتباطات دارو و بیماری: D5 |
| 92.87 | 94.53 | D1+D2+D3+D5 |
| **۹۳.۲۳** | **۹۴.۸۳** | **D1+D2+D3+D4+D5 (Integrated)** |

نمودار 4-1 میزان تاثیر ویژگی‌های مختلف در دقت فرآیند



نمودار 4-2 مشخصه عملکرد روش پیشنهادی

نمودار 4-3 دقت-حساسیت روش پیشنهادی

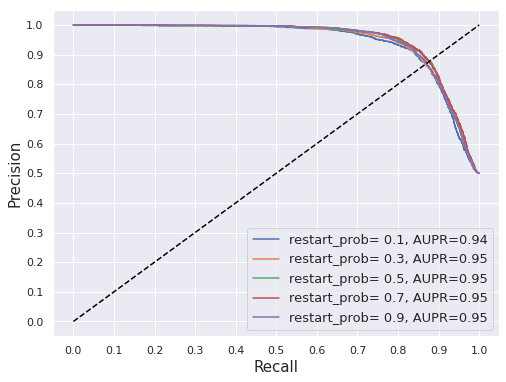
برای روش پیشنهادی میزان مشخصه عملکرد و سطح زیر آن یا AUC در نمودار 4-2 نشان داده شده است. هم‌چنین در نمودار 4-3 دقت-حساسیت روش و سطح زیر نمودار یا AUPR نشان داده شده است. این دو نمودار نماینگر عملکرد بسیار خوب مدل پیشنهادی هستند و نزدیک بودن سطح زیر این دو نمودار به عدد ۱ نشان می‌دهد این روش تقریبا برای تمام مقادیر آستانه به‌خوبی عمل کرده است. علاوه بر این، مقدار سایر معیارهای ارزیابی برای روش پیشنهادی محاسبه شده و در جدول 4-3 آمده است.

جدول 4-3 معیارهای مختلف ارزیابی محاسبه شده برای روش پیشنهادی

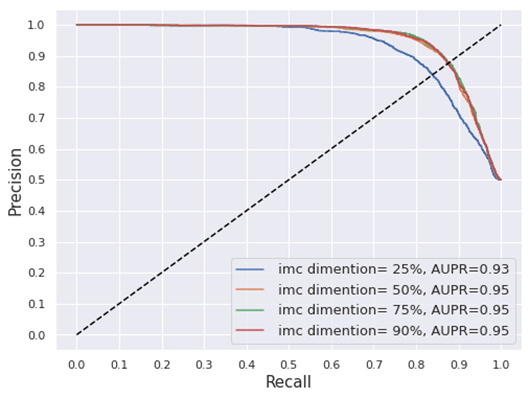
|  |  |
| --- | --- |
| مقدار | **نام معیار** |
| ۰.۸۸ | Accuracy |
| ۰.۸۹ | Precision |
| ۰.۸۸ | Sensitivity |
| ۰.۸۸ | F1 score |

از این جدول می‌توان نتیجه گرفت که ۸۸ درصد از جواب‌هایی که روش پیشنهادی خروجی می‌دهد، صحیح هستند (Accuracy  =۰.۸۸). هم‌چنین از بین جفت دارو-بیماری‌هایی که مدل آن‌ها را به عنوان نمونه مثبت پیش‌بینی کرده، ۸9 درصدشان به‌درستی شناسایی شده‌اند (Precision  =۰.۸9) و از بین تمام زوج دارو-بیماری‌های مثبتی که وجود داشته، توانسته است ۸8 درصدشان را بازیابی کند (Sensitivity  =۰.۸8). در نهایت، مقدار بالای معیار جامعی مانند F1 score بیانگر کارآمدی روش پیشنهادی است.

به منظور بررسی تاثیر مقادیر مختلف ابرپارامتر‌های احتمال بازگشت به گره ابتدایی در گشت‌تصادفی و بعد فضای پنهان در عمل فاکتور‌گیری، این دو ابر‌پارامتر به ترتیب به‌ازای مقادیر {۰.۱، ۰.۳، ۰.۵، ۰.۷، ۰.۹} و {۲۵، ۵۰، ۷۵، ۹۰} اجرا شدند که در نمودارهای 4-4 و 4-5 میزان AUPR آن‌ها نشان داده شده است. در سناریوی اجرا شده جهت یکپارچه‌سازی ابتدا به ازای مقادیر مختلف بازگشت به گره ابتدایی شباهت‌ها تجمیع شدند و بهترین نتیجه به ازای قراردادن احتمال بازگشت برابر ۰.۵ بدست آمد. در گام دوم مقدار احتمال بازگشت به میزان ۰.۵ ثابت نگه داشته شد و سپس خروجی کار به ازای مقادیر مختلف ابر‌پارامتر بعد فضای پنهان سنجیده‌شد. با توجه به این نمودارها می‌بینیم که مقادیر مختلف ابر‌پارمتر تاثیر چندانی در تغییر عملکرد روش پیشنهادی نداشته است و این حاکی از عملکرد ثابت و خوب روش پیشنهادی است که نسبت به تغییر مقادیر ابرپارامتر‌ها زیاد حساس نمی‌باشد.



نمودار 4-4 AUPR به‌ازای مقادیر مختلف ابرپارامتر احتمال بازگشت به گره اولیه



نمودار 4-5 AUPR به ازای مقادیر مختلف ابرپارامتر بعد فضای پنهان در فاکتورگیری

## مقایسه روش پیشنهادی با مطالعات پیشین

در جدول 4-4 نتیجه ارزیابی روش پیشنهادی در مقایسه با تعدادی از بهترین روش‌های موجود آمده است. نتایج نشان می‌دهد که روش پیشنهادی عملکرد بسیار بهتری نسبت به سایر روش‌های پیشین دارد و به خوبی توانسته است ارتباطات دارو-بیماری‌های موجود را پیش‌بینی کند. مقدار بالای هم AUC و هم AUPR در روش پیشنهادی در مقایسه با مقادیر بسیار پایین AUPR و مقادیر نسبتا پایین‌تر AUC نشان می‌دهد که روش پیشنهادی عملکرد بسیار موفق‌تری در یافتن پیوندهای دارو-بیماری داشته باشد.

جدول 4-4 مقایسه عملکرد روش‌ پیشنهادی با عملکرد روش‌های پیشین

|  |  |  |
| --- | --- | --- |
| **AUC** | **AUPR** | **نام روش** |
| ۷۲.۷ | ۳.۰ | TL\_HGBI [67] |
| ۸۵.۱ | ۴.۴ | MBiRW [68] |
| ۶۳.۸ | ۶ | SCMFDD [69] |
| ۹۲.۰ | ۲۴.۳ | DisDrugPred [70] |
| 74.75 | 76.89 | BLM [35] |
| 75.7 | 69.3 | Graph Embedding Matrix Factorization – GraRep [71] |
| 80.4 | 80.6 | Graph Embedding Random walk based - struc2vec [72] |
| 77.4 | 75.2 | Graph Embedding Neural network based - SDNE [73] |
| **۹۳.۲۳** | **۹۴.۸۳** | **DRSE (proposed)** |

نمودار 4-6 مقایسه عملکرد روش‌ پیشنهادی با روش‌های پیشین

روش پیشنهادی به ترتیب معیار‌های AUC و AUPR را 1.13 و 14.23 درصد نسبت به روش‌های پیشین بهبود داده است. از آن‌جایی که معیار AUPR معیار مناسب‌تری برای ارزیابی مسأله ماست روش پیشنهادی بسیار بهتر از روش‌های گذشته عمل کرده است. اهمیت معیار AUPR به نسبت سایر معیار‌ها در این است که این معیار، سنجه‌ی عادلانه‌تری برای ارزیابی این مسأله بوده و زیاد بودن تعداد صفر‌ها در ماتریس ارتباطات که صفر در آن صرفاً نشان‌دهنده نبود ارتباط شناخته شده می‌باشد، در افزایش درصد‌ها تأثیری ندارد. معیار AUC نسبت به تعداد زیاد صفر‌ها در ماتریس ارتباطات حساس است و زیاد بودن این صفر‌ها می‌تواند منجر به افزایش ناعادلانه این معیار گردد. بهبود معیار AUPR توسط روش پیشنهادی می‌تواند نشان‌دهنده عملکرد خوب مدل در برگرداندن برچسب‌های یک باشد که باعث اعتماد به مدل در پیش‌گویی ارتباطات شناخته نشده می‌گردد. این بهبود نسبت به سایر روش‌ها می‌تواند به دلیل استفاده کردن از انواع مختلف ویژگی‌ها و ترکیب‌ آن‌ها باشد. همانطور که در جدول 4-4 نیز نشان داده شد، افزایش ویژگی‌های مختلف منجر به بهبود نتایج شده است. همچنین روش پیشنهادی در این مطالعه به دلیل استفاده‌کردن از گشت تصادفی و تجزیه ماتریسی با رویکردی کارا ویژگی‌های مختلف را به شکل مناسبی با یکدیگر ترکیب می‌کند که این شیوه ترکیبی به دلیل استفاده از فضای پنهان می‌تواند ویژگی‌های توپولوژیکی را نیز در پیش‌گویی نهایی دخیل کند. تمام این موارد در کنار یکدیگر باعث بهبود 14.23 درصدی در معیار AUPR گردیده است.

## مطالعات موردی

به دلیل این‌که در این مسأله صفر‌های موجود در ماتریس ارتباطات دارو و بیماری صرفاً به معنای ناشناخته‌بودن ارتباط(و نه عدم ارتباط قطعی) است و ممکن است این صفر‌ها در آینده تبدیل به یک شوند، ما در این بخش به مطالعه موردی برخی از مثبت‌های کاذب پیش‌بینی شده توسط روش پیشنهادی می‌پردازیم. جدول 4-5 شواهد مطالعات موردی پیوندهای جدید بازهدف‌گذاری کشف شده توسط روش پیشنهادی در گزارش‌های مختلف را نشان می‌دهد.

جدول 4-5 تایید موارد بازهدف‌گذاری کشف شده توسط روش پیشنهادی در گزارش‌های مختلف

|  |  |  |
| --- | --- | --- |
| **نام دارو** | نام بیماری | منبع |
| سفپودوکسیم[[191]](#footnote-191) | عفونت استافیلوکوک[[192]](#footnote-192) | [74] |
| لووستاتین[[193]](#footnote-193) | هیپرلیپیدمی نوع دو[[194]](#footnote-194) | [75] |
| تری‌فلوپرازین[[195]](#footnote-195) | اسکیزوفرنی و پارانویا | [76] |
| سفدی‌نیر[[196]](#footnote-196) | عفونت پوستی[[197]](#footnote-197) | [77] |
| نابیلون[[198]](#footnote-198) | تهوع و استفراغ بعد از بیهوشی[[199]](#footnote-199) | [78] |
| پیپراسیلین[[200]](#footnote-200) | عفونت کلبسیلا[[201]](#footnote-201) | [79] |
| پروکائین آمید[[202]](#footnote-202) | فیبریلاسیون بطنی[[203]](#footnote-203) | [80] |
| فلکاینید[[204]](#footnote-204) | فیبریلاسیون بطنی | [81] |
| ارتاپنم[[205]](#footnote-205) | عفونت استافیلوکوک | [82] |

این موارد از جمله ارتباطاتی هستند که در ماتریس ارتباطات اصلی صفر بودند اما مدل به خوبی آن‌ها را یک پیش‌گویی کرده است. به عنوان مثال ارتباط داروی تری‌فلوپرازین با بیماری اسکیزوفرنی با وجود صفر‌بودن در ماتریس ارتباطات اولیه، به خوبی یک تشخیص داده شده است. تری‌فلوپرازین از گروه فنوتیازین‌هاست که عملکرد اصلی این دارو در مهار گیرنده‌های [دوپامینی](https://fa.wikipedia.org/wiki/دوپامین) در سلول‌های مغزی است. این دارو از طریق بستن گیرنده‌های دوپامینی در سیستم اعصاب مرکزی اثرات ناشی از تحریك دوپامین را کنترل می‌كند. هم‌چنین این دارو موجب انسداد گیرنده‌های آلفا می‌شود و فعالیت ناشی از هیستامین و سروتونین را خنثی می‌سازد[76]. مورد دیگر داروی نابیلون برای درمان تهوع است. در مطالعات اخیر، نابیلون به عنوان یک مکمل پیشگیری از تهوع و استفراغ پس از عمل، شناخته شده است. این دارو پیش‌تر اثربخشی خود را به طور بالینی در درمان تهوع و استفراغ مرتبط با شیمی‌درمانی نشان داده است[78].

# فصل پنجم: جمع‌بندی و کارهای آینده

فصل پنجم

جمع‌بندی و کارهای آینده

## خلاصه پژوهش

پیش‌بینی پیوندهای دارو و بیماری اهمیت زیادی دارد، چرا که برخی از دارو‌های مورد استفاده در بیماری‌های می‌توانند برای بیماری‌های دیگر نیز مفید عمل کنند. این موضوع می‌تواند در کاهش هزینه‌های طراحی و تولید دارو و کاهش زمان نیز تاثیرگذار باشد. بطور مثال در مواردی که سرعت شیوع یک بیماری جدید شبیه کرونا بسیار بالاست و نمی‌توان تا زمان ساخت داروی جدید منتظر ماند. یا در سناریوی دیگری یک بیماری نادر ولی خطرناک شناسایی می‌شود و به دلیل نادر‌بودن بیماری و تعداد کم افراد مبتلا شده، صرف هزینه چند میلیارد دلاری برای شرکت‌های داروسازی توجیه اقتصادی نخواهد داشت لذا پزشکان و محققان باید به بررسی اثربخشی داروهای موجود برای درمان این بیماری بپردازند. به دلیل اهمیت این موضوعات، امروزه بیوانفورماتیک و زیست‌شناسی محاسباتی با بهره‌گیری از علم تحلیل‌داده می‌توانند با بررسی داده‌ها و محاسبات مختلف پیوندهای دارو-بیماری را پیش‌بینی کنند.

سیلدانفیل با نام تجاری ویاگرا نمونه‌ای موفق از بازهدف‌گذاری دارو می‌باشد. سیلدنافیل ابتدا در اواخر دهه 80 میلادی برای درمان بیماری درد قفسه سینه تجویز شد. دارو در طول آزمایشات بالینی فاقد کارآیی لازم بر روی این بیماری شناخته‌شد و روند توسعه آن متوقف گردید تا اینکه بیماران اثرات‌جانبی غیر‌معمول از جمله نعوظ طولانی‌مدت را گزارش کردند. بعد از این بود که محققان با انجام تحقیقات مجدد دارو را برای درمان اختلال نعوظ بازهدف‌گذاری کردند. با پذیرش این فرضیه که دارو‌های مشابه می‌توانند عمل‌کرد مشابهی داشته باشند و همچنین بیماری‌های مشابه می‌توانند درمان‌های مشابهی داشته باشند می‌توان از شباهت دارو‌ها و بیماری‌ها برای باز‌هدف‌گذاری دارویی استفاده کرد. در این مطالعه، روشی برای پیش‌بینی پیوندهای دارو و بیماری با استفاده از ویژگی‌های مختلف دارویی مثل ساختار شیمیایی، عوارض جانبی، پروتئین‌های هدف و غیره و همچنین یک ماتریس ترکیبی شباهت بیماری‌ها ارائه شد.

روش پیشنهادی در این پژوهش از گشت تصادفی و خلاصه‌سازی ویژگی به همراه فاکتورسازی ماتریس برای پیش‌بینی استفاده می‌کند. به عبارت دیگر این روش شامل گام‌های زیر است:

• ترکیب منابع داده‌ای مختلف اعم از ویژگی‌های دارو، پروتئین و بیماری

• محاسبه ماتریس شباهت از روی ویژگی‌ها با تابع جاکارد

• محاسبه ماتریس احتمالات گشت تصادفی برای هر ماتریس شباهت

• خلاصه‌سازی ماتریس‌های حاصل از گشت تصادفی و ایجاد یک ماتریس ترکیبی با ابعاد پایین

• استفاده از فاکتورگیری ماتریسی برای پیش‌بینی احتمال پیوندهای دارو و بیماری

## بحث و نتیجه گیری

نتایج حاصل توسط ارزیابی تقاطعی 5 دسته‌ای با معیار‌های AUC و AUPR اندازه‌گیری شد. این معیار‌ها به ترتیب برابر با 93.23 و 94.83 درصد هستند. میزان کارایی هر ماتریس شباهت نیز در افزایش دقت پیش‌بینی بررسی شد و از میان شباهت‌ها، شباهت ساختاری و شباهت هستان‌شناسی به ترتیب نتایج بهتری از دیگر ویژگی‌ها کسب کردند. بررسی‌ها نشان داد که استفاده از ترکیب ویژگی‌های مختلف می‌تواند به افزایش دقت بیانجامد و کارایی نحوه خلاصه‌سازی و یکپارچه‌سازی انجام شده را اثبات کرد. برای تحلیل‌های بیشتر مطالعات موردی بر روی پیش‌بینی‌های به مثبت کاذب انجام شد. ارزیابی‌ها نشان می‌دهند که روش پیشنهادی با در نظر گرفتن عوارض جانبی و دیگر ویژگی‌ها توانسته است به خوبی تعاملات ناشناخته را پیش‌بینی کند.

به عنوان کار‌های آینده می‌توان به پژوهش‌های زیر پرداخت:

* استفاده از سایر منابع داده‌ای ژنومیک و ترنسکریپتومیک بدن انسان
* استفاده از دیگر توابع کرنل جهت یکپارچه‏سازی شباهت‏ها
* بررسی آنتروپی ماتریس‌های شباهت قبل از در نظر گرفتن آن‌ها و کنار گذاشتن ماتریس‌هایی که ممکن است دارای نویز بالایی باشند.
* با توجه به اینکه تاکنون شدت خطر عوارض جانبی داروها برای بدن در جایی ثبت نشده در صورت ایجاد شدن این مجموعه‌داده می‌توان آن را به عنوان عاملی در جهت کاهش خطر بازهدف‌گذاری در نظر گرفت.

**منابع**

**منابع فارسی**

[5] ب. باباعباسی, بیوانفورماتیک سلولی و مولکولی, vol. 1. 1395.

[16] م. حسني سيمكي, س. سريزدي, and ح. نظام آبادي پور, “روشي براي پياده سازي سريع الگوريتم هاي پيش بيني پيوند مبتني بر مشابهت محلي,” سومين كنفرانس ملي در مهندسي كامپيوتر، فناوري اطلاعات و پردازش داده ها. دانشگاه پيام نور, تهران, 1397, [Online]. Available: https://www.civilica.com/Paper-CITCOMP03-CITCOMP03\_137.html.

[17] ا. سلطاني مشكور, “بهينه سازي پيش بيني پيوند در شبكه هاي اجتماعي,” كنفرانس ملي كاربرد فناوري هاي نوين در علوم و مهندسي، برق و كامپيوتر و IT. دانشگاه ايوانكي, تهران, 1396, [Online]. Available: https://www.civilica.com/Paper-TESCONF01-TESCONF01\_045.html.

[20] ف. قراگزلو, “بررسي چالش هاي سيستم هاي توصيه گر پالايش مشاركتي,” ا. معاونت توسعه فناوري رسانه سازمان صدا و سيما, تهران, 1397, [Online]. Available: https://www.civilica.com/Paper-MTEC15-MTEC15\_044.html.

[21] م. حسين زاده اقدم, “ارايه روشي براي سيستم هاي توصيه گر با استفاده از فاكتورگيري ماتريسي,” كنفرانس ملي فناوري هاي نوين در مهندسي برق و كامپيوتر. جهاد دانشگاه اصفهان, اصفهان, 1396, [Online]. Available: https://www.civilica.com/Paper-PCCO01-PCCO01\_224.html.

**منابع انگلیسی**

[1] H. Xue, J. Li, H. Xie, and Y. Wang, “Review of drug repositioning approaches and resources,” *Int. J. Biol. Sci.*, vol. 14, no. 10, p. 1232, 2018.

[2] S. Pushpakom *et al.*, “Drug repurposing: progress, challenges and recommendations,” *Nat. Rev. Drug Discov.*, vol. 18, no. 1, pp. 41–58, 2019.

[3] J. Zlatanova and K. van Holde, *Molecular biology: structure and dynamics of genomes and proteomes*. Garland Science, 2015.

[4] H. H. Rashidi and L. K. Buehler, *Bioinformatics basics: applications in biological science and medicine*. CRC press, 1999.

[6] “What is a cell?,” [Online]. Available: https://www.yourgenome.org/facts/what-is-a-cell.

[7] “Protein structure,” [Online]. Available: https://en.wikipedia.org/wiki/Protein\_structure.

[8] “Central Dogma,” [Online]. Available: https://ib.bioninja.com.au/standard-level/topic-2-molecular-biology/27-dna-replication-transcri/central-dogma.html.

[9] “Genetics.” https://en.wikipedia.org/wiki/Genetics.

[10] T. K. Attwood, A. Gisel, N.-E. Eriksson, and E. Bongcam-Rudloff, “Concepts, historical milestones and the central place of bioinformatics in modern biology: a European perspective,” *Bioinformatics-trends Methodol.*, vol. 1, pp. 1–31, 2011.

[11] “amino acid binding.” http://amigo.geneontology.org/amigo/term/GO:0016597.

[12] “Functional analysis - Data bases - Gene Ontology.” https://biocorecrg.github.io/CRG\_Bioinformatics\_for\_Biologists/functional\_analysis.html.

[13] A. J. Trevor, B. G. Katzung, S. B. Masters, and M. Kruidering-Hall, *Pharmacology examination & board review*. McGraw-Hill Medical New York, 2010.

[14] D. G. Rudmann, “On-target and off-target-based toxicologic effects,” *Toxicol. Pathol.*, vol. 41, no. 2, pp. 310–314, 2013.

[15] L. Lü and T. Zhou, “Link prediction in complex networks: A survey,” *Phys. A Stat. Mech. its Appl.*, vol. 390, no. 6, pp. 1150–1170, 2011.

[18] V. Martínez, F. Berzal, and J.-C. Cubero, “A survey of link prediction in complex networks,” *ACM Comput. Surv.*, vol. 49, no. 4, pp. 1–33, 2016.

[19] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, *Recommender systems: an introduction*. Cambridge University Press, 2010.

[22] M. F. Cetin and S. Ayvaz, “A Negative Similarity Based Hybrid Recommender System Using Apache Spark,” in *Proceedings of the 2019 3rd International Conference on Advances in Artificial Intelligence*, 2019, pp. 166–172.

[23] G. Jin and S. T. C. Wong, “Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines,” *Drug Discov. Today*, vol. 19, no. 5, pp. 637–644, 2014.

[24] Y. Wang, S. Chen, N. Deng, and Y. Wang, “Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data,” *PLoS One*, vol. 8, no. 11, p. e78518, 2013.

[25] M. J. Keiser *et al.*, “Predicting new molecular targets for known drugs,” *Nature*, vol. 462, no. 7270, pp. 175–181, 2009.

[26] J. Schuler and R. Samudrala, “Fingerprinting CANDO: Increased Accuracy with Structure-and Ligand-Based Shotgun Drug Repurposing,” *ACS omega*, vol. 4, no. 17, pp. 17393–17403, 2019.

[27] S. Croset, “Drug repositioning and indication discovery using description logics.” University of Cambridge, 2014.

[28] V. J. Haupt and M. Schroeder, “Old friends in new guise: repositioning of known drugs with structural bioinformatics,” *Brief. Bioinform.*, vol. 12, no. 4, pp. 312–326, 2011.

[29] H. Arrouchi, W. Lakhlili, and A. Ibrahimi, “Re-positioning of known drugs for Pim-1 kinase target using molecular docking analysis,” *Bioinformation*, vol. 15, no. 2, p. 116, 2019.

[30] F. A. Kruger, R. Rostom, and J. P. Overington, “Mapping small molecule binding data to structural domains,” in *BMC bioinformatics*, 2012, vol. 13, no. S17, p. S11.

[31] A. P. Chiang and A. J. Butte, “Systematic evaluation of drug–disease relationships to identify leads for novel drug uses,” *Clin. Pharmacol. Ther.*, vol. 86, no. 5, pp. 507–510, 2009.

[32] Y. Li and P. Agarwal, “A pathway-based view of human diseases and disease relationships,” *PLoS One*, vol. 4, no. 2, p. e4346, 2009.

[33] J. Lamb *et al.*, “The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease,” *Science (80-. ).*, vol. 313, no. 5795, pp. 1929–1935, 2006.

[34] J. T. Dudley *et al.*, “Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease,” *Sci. Transl. Med.*, vol. 3, no. 96, pp. 96ra76-96ra76, 2011.

[35] K. Bleakley and Y. Yamanishi, “Supervised prediction of drug–target interactions using bipartite local models,” *Bioinformatics*, vol. 25, no. 18, pp. 2397–2403, 2009.

[36] L. Peng *et al.*, “Screening drug-target interactions with positive-unlabeled learning,” *Sci. Rep.*, vol. 7, no. 1, pp. 1–17, 2017.

[37] H. Chen and J. Li, “A flexible and robust multi-source learning algorithm for drug repositioning,” in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2017, pp. 510–515.

[38] E. Kim, A. Choi, and H. Nam, “Drug repositioning of herbal compounds via a machine-learning approach,” *BMC Bioinformatics*, vol. 20, no. 10, p. 247, 2019, doi: 10.1186/s12859-019-2811-8.

[39] J. Lu *et al.*, “Identification of new candidate drugs for lung cancer using chemical–chemical interactions, chemical–protein interactions and a K-means clustering algorithm,” *J. Biomol. Struct. Dyn.*, vol. 34, no. 4, pp. 906–917, 2016.

[40] X.-Y. Yan, S.-W. Zhang, and S.-Y. Zhang, “Prediction of drug–target interaction by label propagation with mutual interaction information derived from heterogeneous network,” *Mol. Biosyst.*, vol. 12, no. 2, pp. 520–531, 2016.

[41] M. L. Shahreza, N. Ghadiri, S. R. Mousavi, J. Varshosaz, and J. R. Green, “Heter-LP: A heterogeneous label propagation algorithm and its application in drug repositioning,” *J. Biomed. Inform.*, vol. 68, pp. 167–183, 2017.

[42] G. Ceddia, P. Pinoli, S. Ceri, and M. Masseroli, “Matrix Factorization-based Technique for Drug Repurposing Predictions,” *IEEE J. Biomed. Heal. Informatics*, 2020.

[43] A. Mongia, E. Chouzenoux, and A. Majumdar, “Computational prediction of Drug-Disease association based on Graph-regularized one bit Matrix completion,” *bioRxiv*, 2020.

[44] Y. Luo *et al.*, “A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information,” *Nat. Commun.*, vol. 8, no. 1, pp. 1–13, 2017.

[45] X. Yue *et al.*, “Graph embedding on biomedical networks: methods, applications and evaluations,” *Bioinformatics*, vol. 36, no. 4, pp. 1241–1251, 2020.

[46] X. Zeng, S. Zhu, X. Liu, Y. Zhou, R. Nussinov, and F. Cheng, “deepDR: a network-based deep learning approach to in silico drug repositioning,” *Bioinformatics*, vol. 35, no. 24, pp. 5191–5198, 2019.

[47] H. Öztürk, A. Özgür, and E. Ozkirimli, “DeepDTA: deep drug–target binding affinity prediction,” *Bioinformatics*, vol. 34, no. 17, pp. i821–i829, 2018.

[48] N. Zhang, W. Xu, S. Wang, Y. Qiao, and X. Zhang, “Computational Drug Discovery in Chemotherapy-induced Alopecia via Text Mining and Biomedical Databases,” *Clin. Ther.*, vol. 41, no. 5, pp. 972–980, 2019.

[49] X. Liang *et al.*, “LRSSL: predict and interpret drug–disease associations based on data integration using sparse subspace learning,” *Bioinformatics*, vol. 33, no. 8, pp. 1187–1196, Jan. 2017, doi: 10.1093/bioinformatics/btw770.

[50] Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang, and S. H. Bryant, “PubChem: a public information system for analyzing bioactivities of small molecules,” *Nucleic Acids Res.*, vol. 37, no. suppl\_2, pp. W623–W633, 2009.

[51] A. Mitchell *et al.*, “The InterPro protein families database: the classification resource after 15 years,” *Nucleic Acids Res.*, vol. 43, no. D1, pp. D213–D221, 2015.

[52] U. Consortium, “The universal protein resource (UniProt) in 2010,” *Nucleic Acids Res.*, vol. 38, no. suppl\_1, pp. D142–D148, 2010.

[53] L. Cheng, Y. Hu, J. Sun, M. Zhou, and Q. Jiang, “DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function,” *Bioinformatics*, vol. 34, no. 11, pp. 1953–1956, 2018.

[54] M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork, “The SIDER database of drugs and side effects,” *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1075–D1079, 2016.

[55] H. Cho, B. Berger, and J. Peng, “Diffusion component analysis: unraveling functional topology in biological networks,” in *International Conference on Research in Computational Molecular Biology*, 2015, pp. 62–64.

[56] H. Cho, B. Berger, and J. Peng, “Compact integration of multi-network topology for functional analysis of genes,” *Cell Syst.*, vol. 3, no. 6, pp. 540–548, 2016.

[57] M. Cao *et al.*, “New directions for diffusion-based network prediction of protein function: incorporating pathways with confidence,” *Bioinformatics*, vol. 30, no. 12, pp. i219–i227, 2014.

[58] S. Köhler, S. Bauer, D. Horn, and P. N. Robinson, “Walking the interactome for prioritization of candidate disease genes,” *Am. J. Hum. Genet.*, vol. 82, no. 4, pp. 949–958, 2008.

[59] S. Navlakha and C. Kingsford, “The power of protein interaction networks for associating genes with diseases,” *Bioinformatics*, vol. 26, no. 8, pp. 1057–1063, 2010.

[60] C.-S. Liao, K. Lu, M. Baym, R. Singh, and B. Berger, “IsoRankN: spectral methods for global alignment of multiple protein networks,” *Bioinformatics*, vol. 25, no. 12, pp. i253–i258, 2009.

[61] X. Chen, M. X. Liu, and G. Y. Yan, “Drug-target interaction prediction by random walk on the heterogeneous network,” *Mol. Biosyst.*, vol. 8, no. 7, pp. 1970–1978, 2012, doi: 10.1039/c2mb00002d.

[62] M. Kim and J. Leskovec, “The network completion problem: Inferring missing nodes and edges in networks,” in *Proceedings of the 2011 SIAM International Conference on Data Mining*, 2011, pp. 47–58.

[63] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, “Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization,” *ACM Trans. Math. Softw.*, vol. 23, no. 4, pp. 550–560, 1997.

[64] S. Wang, H. Cho, C. Zhai, B. Berger, and J. Peng, “Exploiting ontology graph for predicting sparsely annotated gene function,” *Bioinformatics*, vol. 31, no. 12, pp. i357–i364, 2015.

[65] N. Natarajan and I. S. Dhillon, “Inductive matrix completion for predicting gene–disease associations,” *Bioinformatics*, vol. 30, no. 12, pp. i60–i68, 2014.

[66] H.-F. Yu, P. Jain, P. Kar, and I. Dhillon, “Large-scale multi-label learning with missing labels,” in *International conference on machine learning*, 2014, pp. 593–601.

[67] W. Wang, S. Yang, X. Zhang, and J. Li, “Drug repositioning by integrating target information through a heterogeneous network model,” *Bioinformatics*, vol. 30, no. 20, pp. 2923–2930, 2014.

[68] H. Luo *et al.*, “Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm,” *Bioinformatics*, vol. 32, no. 17, pp. 2664–2671, 2016.

[69] W. Zhang *et al.*, “Predicting drug-disease associations by using similarity constrained matrix factorization,” *BMC Bioinformatics*, vol. 19, no. 1, pp. 1–12, 2018.

[70] P. Xuan, Y. Cao, T. Zhang, X. Wang, S. Pan, and T. Shen, “Drug repositioning through integration of prior knowledge and projections of drugs and diseases,” *Bioinformatics*, vol. 35, no. 20, pp. 4108–4119, 2019.

[71] S. Cao, W. Lu, and Q. Xu, “Grarep: Learning graph representations with global structural information,” in *Proceedings of the 24th ACM international on conference on information and knowledge management*, 2015, pp. 891–900.

[72] L. F. R. Ribeiro, P. H. P. Saverese, and D. R. Figueiredo, “struc2vec: Learning node representations from structural identity,” in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 385–394.

[73] D. Wang, P. Cui, and W. Zhu, “Structural deep network embedding,” in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, pp. 1225–1234.

[74] Y. C. Liu, W. K. Huang, and D. L. Cheng, “Antibacterial activity of cef podoxime in vitro,” *Chemotherapy*, vol. 43, no. 1, pp. 21–26, Jan. 1997, doi: 10.1159/000239530.

[75] R. Carmena, G. Roederer, H. Mailloux, S. Lussier-Cacan, and J. Davignon, “The response to lovastatin treatment in patients with heterozygous familial hypercholesterolemia is modulated by apolipoprotein E polymorphism,” *Metabolism*, vol. 42, no. 7, pp. 895–901, 1993, doi: 10.1016/0026-0495(93)90066-W.

[76] R. Macdonald and T. P. Shields Watts, “Trifluoperazine dihydrochloride (‘stelazine’) in paranoid schizophrenia,” *Br. Med. J.*, vol. 1, no. 5121, pp. 549–550, Feb. 1959, doi: 10.1136/bmj.1.5121.549.

[77] H. S. Sader, J. M. Streit, T. R. Fritsche, and R. N. Jones, “Potency and spectrum reevaluation of cefdinir tested against pathogens causing skin and soft tissue infections: A sample of North American isolates,” *Diagn. Microbiol. Infect. Dis.*, vol. 49, no. 4, pp. 283–287, Aug. 2004, doi: 10.1016/j.diagmicrobio.2004.04.015.

[78] D. N. Levin, Z. Dulberg, A.-W. Chan, G. M. T. Hare, C. D. Mazer, and A. Hong, “A randomized-controlled trial of nabilone for the prevention of acute postoperative nausea and vomiting in elective surgery,” *Can. J. Anesth. Can. d’anesthésie*, vol. 64, no. 4, pp. 385–395, 2017.

[79] T. Pillay, D. G. Pillay, M. Adhikari, and A. W. Sturm, “Piperacillin/tazobactam in the treatment of Klebsiella pneumoniae infections in neonates,” *Am. J. Perinatol.*, vol. 15, no. 1, pp. 47–51, Jan. 1998, doi: 10.1055/s-2007-993898.

[80] D. T. Markel *et al.*, “Procainamide and survival in ventricular fibrillation out-of-hospital cardiac arrest,” *Acad. Emerg. Med.*, vol. 17, no. 6, pp. 617–623, 2010, doi: 10.1111/j.1553-2712.2010.00763.x.

[81] R. H. Falk, “Flecainide-induced ventricular tachycardia and fibrillation in patients treated for atrial fibrillation,” *Ann. Intern. Med.*, vol. 111, no. 2, pp. 107–111, 1989, doi: 10.7326/0003-4819-111-2-107.

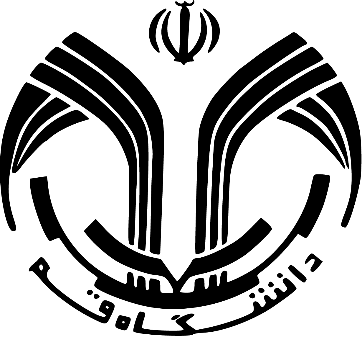
[82] R. M. Gesser, K. A. McCarroll, and G. L. Woods, “Efficacy of ertapenem against methicillin-susceptible Staphylococcus aureus in complicated skin/skin structure infections: Results of a double-blind clinical trial versus piperacillin-tazobactam,” *Int. J. Antimicrob. Agents*, vol. 23, no. 3, pp. 235–239, 2004, doi: 10.1016/j.ijantimicag.2003.07.013.

**Abstract**

Experimental methods in the field of drug development and discovery are time-consuming and costly. Researches indicate that designing a new drug from the first stage to its delivery to the consumer market takes between 10 and 15 years. Besides, this process has a cost of 0.8 to 1.5 billion dollars. Therefore, finding a new strategy for drug discovery is necessary. Drug repurposing is to identify new indications for approved drugs in addition to their main indications. The aim is to identify drug-disease associations by computational methods and machine learning algorithms using multiple types of data. Finding new indications for an approved drug by computational methods can reduce the R & D process's cost and risk. It has also created a good ground for more work to find a treat for rare diseases and also an opportunity to produce specific drugs for people according to the metabolic characteristics and ... each person.

In this study, we present a method for drug re-purposing based on an integrated network of diverse and heterogeneous data. The method predicts drug-disease associations by a matrix factorization algorithm using the side-effect features of drugs. The proposed method improved AUC and AUPR criteria by 1.13 and 14.23% compared to the previous methods, respectively. The integration of different types of drug and disease features has led to high performance. It has indicated the efficiency of the process in data abstraction, dimension reduction, and similarity integration.

**Keywords:** Computational pharmacology, Drug repurposing, Biological networks, System biology, Drug side effect



The University of Qom  
 Faculty of Technical and Engineering

Thesis  
 for degree of Master of Science (MSC) In Information Technology Engineering

Title:

A method for drug repurposing considering side effect

Supervisor:

DR. Amir Lakizadeh

By:

Seyed Mohammad Hassan Mirashrafi

summer, 2020

1. . Drug discovery [↑](#footnote-ref-1)
2. . U.S. Food and Drug Administration [↑](#footnote-ref-2)
3. . Drug repurposing [↑](#footnote-ref-3)
4. . Computational pharmacology [↑](#footnote-ref-4)
5. . Sildenafil [↑](#footnote-ref-5)
6. . Viagra [↑](#footnote-ref-6)
7. . Angina [↑](#footnote-ref-7)
8. . Minoxidil [↑](#footnote-ref-8)
9. . Atomoxetine [↑](#footnote-ref-9)
10. . Rituximab [↑](#footnote-ref-10)
11. . Topiramate [↑](#footnote-ref-11)
12. . Fingolimod [↑](#footnote-ref-12)
13. . Duloxetine [↑](#footnote-ref-13)
14. . Raloxifene [↑](#footnote-ref-14)
15. . Dapoxetine [↑](#footnote-ref-15)
16. . Biology [↑](#footnote-ref-16)
17. . Molecular & cell biology [↑](#footnote-ref-17)
18. . Deoxyribonucleic acid [↑](#footnote-ref-18)
19. . Ribonucleic acid [↑](#footnote-ref-19)
20. . Interaction [↑](#footnote-ref-20)
21. . Bioinformatics [↑](#footnote-ref-21)
22. . Cell [↑](#footnote-ref-22)
23. . Protein [↑](#footnote-ref-23)
24. . Amino acid [↑](#footnote-ref-24)
25. . Gene [↑](#footnote-ref-25)
26. . Protein–protein interaction [↑](#footnote-ref-26)
27. . Protein Structure [↑](#footnote-ref-27)
28. . Protein domain [↑](#footnote-ref-28)
29. . mRNA [↑](#footnote-ref-29)
30. . tRNA [↑](#footnote-ref-30)
31. . Nucleic acid sequence [↑](#footnote-ref-31)
32. . Central dogma of molecular biology [↑](#footnote-ref-32)
33. . Transcription [↑](#footnote-ref-33)
34. . Translation [↑](#footnote-ref-34)
35. . Gene expression [↑](#footnote-ref-35)
36. . Gene Ontology [↑](#footnote-ref-36)
37. . Disease [↑](#footnote-ref-37)
38. . Drug [↑](#footnote-ref-38)
39. . Receptor [↑](#footnote-ref-39)
40. . Pharmacology [↑](#footnote-ref-40)
41. . Pharmacodynamics [↑](#footnote-ref-41)
42. . Pharmacokinetics [↑](#footnote-ref-42)
43. . Chemical structure [↑](#footnote-ref-43)
44. . On-target effect [↑](#footnote-ref-44)
45. . Off-target effect [↑](#footnote-ref-45)
46. . Side effect [↑](#footnote-ref-46)
47. . Link prediction [↑](#footnote-ref-47)
48. . Recommender system [↑](#footnote-ref-48)
49. . Biological networks [↑](#footnote-ref-49)
50. . Overload [↑](#footnote-ref-50)
51. . Complex networks [↑](#footnote-ref-51)
52. . Nodes [↑](#footnote-ref-52)
53. . Edges [↑](#footnote-ref-53)
54. . Missing link [↑](#footnote-ref-54)
55. . Diseases-gene networks [↑](#footnote-ref-55)
56. . User-item bipartite network [↑](#footnote-ref-56)
57. . Similarity-based methods [↑](#footnote-ref-57)
58. . Probabilistic and Statistical Methods [↑](#footnote-ref-58)
59. . Algorithmic Methods [↑](#footnote-ref-59)
60. . Local [↑](#footnote-ref-60)
61. . Global [↑](#footnote-ref-61)
62. . Quasi-local [↑](#footnote-ref-62)
63. . Jaccard [↑](#footnote-ref-63)
64. . Preferential Attachment (PA) [↑](#footnote-ref-64)
65. . Common Neighbors (CN) [↑](#footnote-ref-65)
66. . Adamic Adar (AA) [↑](#footnote-ref-66)
67. . Katz [↑](#footnote-ref-67)
68. . Local Path (LP) [↑](#footnote-ref-68)
69. . Local Random Walk (LRW) [↑](#footnote-ref-69)
70. . Jaccard index [↑](#footnote-ref-70)
71. . Supervised Learning [↑](#footnote-ref-71)
72. . Optimization Techniques [↑](#footnote-ref-72)
73. . Matrix Factorization [↑](#footnote-ref-73)
74. . Content-based [↑](#footnote-ref-74)
75. . Collaborative filtering(CF) [↑](#footnote-ref-75)
76. . Hybrid recommender systems [↑](#footnote-ref-76)
77. . Cosine similarity [↑](#footnote-ref-77)
78. . Memory-based approches [↑](#footnote-ref-78)
79. . Model-based approches [↑](#footnote-ref-79)
80. . Bayesian Clustering [↑](#footnote-ref-80)
81. . Latent Semantic Analysis [↑](#footnote-ref-81)
82. . Latent features [↑](#footnote-ref-82)
83. . Latent feature space [↑](#footnote-ref-83)
84. . Singular Value Decomposition (SVD) [↑](#footnote-ref-84)
85. . Maximum Margin Matrix Factorization [↑](#footnote-ref-85)
86. . Expectation Maximization for Matrix Factorization [↑](#footnote-ref-86)
87. . Regularized Matrix Factorization (RMF) [↑](#footnote-ref-87)
88. . Random walk [↑](#footnote-ref-88)
89. . Matrix factorization [↑](#footnote-ref-89)
90. . Network-based approaches [↑](#footnote-ref-90)
91. . Machine learning-based methods [↑](#footnote-ref-91)
92. . Deep learning-based methods [↑](#footnote-ref-92)
93. . Text mining-based approaches [↑](#footnote-ref-93)
94. . Guilt-by-association principle [↑](#footnote-ref-94)
95. . Molecular fingerprint [↑](#footnote-ref-95)
96. . Ligand [↑](#footnote-ref-96)
97. . Wang [↑](#footnote-ref-97)
98. . Keiser [↑](#footnote-ref-98)
99. . Schuler [↑](#footnote-ref-99)
100. . Kernel [↑](#footnote-ref-100)
101. . Bioactive [↑](#footnote-ref-101)
102. . Molecular docking [↑](#footnote-ref-102)
103. . Binding site [↑](#footnote-ref-103)
104. . Arrouchi [↑](#footnote-ref-104)
105. . Serine/threonine-specific protein kinase [↑](#footnote-ref-105)
106. . Agarwal [↑](#footnote-ref-106)
107. . Biological pathway [↑](#footnote-ref-107)
108. . Gene expression signature [↑](#footnote-ref-108)
109. . https://clue.io/cmap [↑](#footnote-ref-109)
110. . Lamb [↑](#footnote-ref-110)
111. . Dudley [↑](#footnote-ref-111)
112. . Topiramate [↑](#footnote-ref-112)
113. . Inflammatory bowel [↑](#footnote-ref-113)
114. . SVM [↑](#footnote-ref-114)
115. . Least squares [↑](#footnote-ref-115)
116. . Logistic regression [↑](#footnote-ref-116)
117. . K Bleakley [↑](#footnote-ref-117)
118. . bipartite local model (BLM) [↑](#footnote-ref-118)
119. . local classifier [↑](#footnote-ref-119)
120. . L Peng [↑](#footnote-ref-120)
121. . Ambiguous samples [↑](#footnote-ref-121)
122. . H Chen [↑](#footnote-ref-122)
123. . Regularized least squares [↑](#footnote-ref-123)
124. . Kernel function [↑](#footnote-ref-124)
125. . E Kim [↑](#footnote-ref-125)
126. . Network clustering [↑](#footnote-ref-126)
127. . J Lu [↑](#footnote-ref-127)
128. . SCLC [↑](#footnote-ref-128)
129. . Label propagation [↑](#footnote-ref-129)
130. . XY Yan [↑](#footnote-ref-130)
131. . ML Shahreza [↑](#footnote-ref-131)
132. . G Ceddia [↑](#footnote-ref-132)
133. . A Mongia [↑](#footnote-ref-133)
134. . K−nearest neighbor [↑](#footnote-ref-134)
135. . Laplacian graph regularization [↑](#footnote-ref-135)
136. . Y Luo [↑](#footnote-ref-136)
137. . Network embedding [↑](#footnote-ref-137)
138. . X Yue [↑](#footnote-ref-138)
139. . X Zeng [↑](#footnote-ref-139)
140. . Autoencoder [↑](#footnote-ref-140)
141. . Collective variational autoencoder [↑](#footnote-ref-141)
142. . H Öztürk [↑](#footnote-ref-142)
143. . Convolutional neural network [↑](#footnote-ref-143)
144. . Label encoding [↑](#footnote-ref-144)
145. . Pool [↑](#footnote-ref-145)
146. . Concat [↑](#footnote-ref-146)
147. . Fully connected [↑](#footnote-ref-147)
148. . Swanson’s ABC model [↑](#footnote-ref-148)
149. . N Zhang [↑](#footnote-ref-149)
150. . http://www.dgidb.org/ [↑](#footnote-ref-150)
151. . Random walk [↑](#footnote-ref-151)
152. . Diffusion component analysis [↑](#footnote-ref-152)
153. . Matrix factorization [↑](#footnote-ref-153)
154. . low-dimensional vector representation [↑](#footnote-ref-154)
155. . Xujun liang [↑](#footnote-ref-155)
156. . https://pubchem.ncbi.nlm.nih.gov/ [↑](#footnote-ref-156)
157. . https://www.ebi.ac.uk/interpro/ [↑](#footnote-ref-157)
158. . https://www.uniprot.org/ [↑](#footnote-ref-158)
159. . http://bio-annotation.cn:18080/DincRNAClient/#/Home [↑](#footnote-ref-159)
160. . http://sideeffects.embl.de/ [↑](#footnote-ref-160)
161. . Compact feature learning [↑](#footnote-ref-161)
162. . Network propagation [↑](#footnote-ref-162)
163. . Dimension Reduction [↑](#footnote-ref-163)
164. . Random walk with restart [↑](#footnote-ref-164)
165. . Stationary distribution [↑](#footnote-ref-165)
166. . Diffusion state [↑](#footnote-ref-166)
167. . Multinomial logistic model [↑](#footnote-ref-167)
168. . Latent vector [↑](#footnote-ref-168)
169. . Kullback-Leibler (KL) divergence [↑](#footnote-ref-169)
170. . Singular value decomposition [↑](#footnote-ref-170)
171. . Relax [↑](#footnote-ref-171)
172. . Bilinear function [↑](#footnote-ref-172)
173. . projection matrix [↑](#footnote-ref-173)
174. . Low-rank constraint [↑](#footnote-ref-174)
175. . Overfitting [↑](#footnote-ref-175)
176. . NP-hard [↑](#footnote-ref-176)
177. . Frobenius norm [↑](#footnote-ref-177)
178. . Regularization parameter [↑](#footnote-ref-178)
179. . Alternating minimization [↑](#footnote-ref-179)
180. . Stratified cross validdation [↑](#footnote-ref-180)
181. . Cross-validation [↑](#footnote-ref-181)
182. . Contingency matrix [↑](#footnote-ref-182)
183. . True positive [↑](#footnote-ref-183)
184. . False positive [↑](#footnote-ref-184)
185. . True negative [↑](#footnote-ref-185)
186. . False negative [↑](#footnote-ref-186)
187. . Receiver operating characteristic (ROC) curve [↑](#footnote-ref-187)
188. . True positive rate [↑](#footnote-ref-188)
189. . True negative rate [↑](#footnote-ref-189)
190. . Precision and recall curve [↑](#footnote-ref-190)
191. . Cefpodoxime [↑](#footnote-ref-191)
192. . Staphylococcal infection [↑](#footnote-ref-192)
193. . Lovastatin [↑](#footnote-ref-193)
194. . Hyperlipoproteinemia Type II [↑](#footnote-ref-194)
195. . Trifluoperazine [↑](#footnote-ref-195)
196. . Cefdinir [↑](#footnote-ref-196)
197. . soft-tissue infections [↑](#footnote-ref-197)
198. . nabilone [↑](#footnote-ref-198)
199. . Postoperative Nausea and Vomiting [↑](#footnote-ref-199)
200. . Piperacillin [↑](#footnote-ref-200)
201. . Klebsiella Infections [↑](#footnote-ref-201)
202. . Procainamide [↑](#footnote-ref-202)
203. . Ventricular fibrillation [↑](#footnote-ref-203)
204. . Flecainide [↑](#footnote-ref-204)
205. . Ertapenem [↑](#footnote-ref-205)