دانشگاه صنعتی نوشیروانی بابل

دانشکده مهندسی برق و کامپیوتر

سمینار کارشناسی ارشد

رشته ی مهندسی کامپیوتر گرایش هوش مصنوعی

مروری بر روشهای ارتباط دارو-هدف مبتنی بر شبکه عصبی گراف

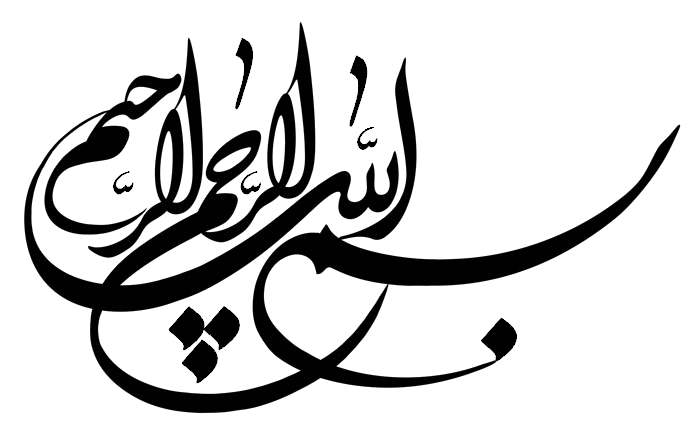
دانشجو:

**سونیا فلاحتی**

استاد راهنما:

**دکتر فاطمه زمانی**

بهار 1402



## چکیده

تولید وتوسعه داروهای جدید پرهزینه و زمان برهستند. استفاده مجدد از دارو رویکردی مهم برای تسریع فرآیند کشف دارواست. استفاده مجدد از داروها به دلیل کاهش هزینه های تحقیق و توسعه ، کاهش زمان تولید دارو، کاهش خطرات جانبی و افزایش اطمینان از اثربخشی دارو ها ، میتواند بسیار موثر و مفید باشد. به منظوراستفاده مجدد از داروها به طور موثر، دانستن اینکه کدام پروتئین ها توسط کدام داروها هدف قرار می گیرند مفید است. پیش بینی ارتباط میان دارو-هدف گام مهمی در این راستا می‌باشد.

مدل های محاسباتی بسیاری برای پیش بینی ارتباط میان دارو-هدف مطرح شده اند. مدل‌های یادگیری عمیق دارای بهترین عملکرد در پیش‌بینی ارتباط دارو-هدف هستند. در اغلب تحقیقات انجام شده در یادگیری عمیق نمایش مولکولهای دارو را به صورت توالی رشته ای در نظر می گیرند که روشی طبیعی برای نشان دادن مولکول ها نیست. زیرا هنگام استفاده ازرشته ها، اطلاعات ساختاری مولکول از بین می رود که میتواند قدرت پیش بینی یک مدل و کارایی آن را کاهش دهد.

در سالهای اخیر از شبکه عصبی گراف در پیش بینی ارتباط میان دارو و پروتئین ها استفاده شده است که کارایی بسیار خوبی در مقایسه با سایر روشهای مورد استفاده از خود نشان داده است.

مولکولها را می توان با استفاده از گراف نشان داد. شبکه عصبی گراف می تواند اطلاعات ساختاری گراف مولکولی را با استفاده از ویژگی های گره های همسایه به دست آورد و ویژگیهای ساختاری مولکولهارا بهتر به نمایش بگذارد و مستقیماً داده های گراف را پردازش کند.

پیاده سازی اولیه انجام شده نشان دهنده عملکرد مطلوب روش شبکه عصبی گراف در بررسی و پیش بینی ارتباط دارو-هدف می باشد. در این گزارش یک بررسی جامع از کاربردهای اخیر شبکه عصبی گراف در زمینه پیش بینی ارتباط دارو-هدف ارائه خواهیم کرد.

**واژه‌های کلیدی:** شبکه عصبی گراف؛ پیش بینی ارتباط دارو-هدف؛ یادگیری عمیق؛ تولید دارو؛ هوش مصنوعی؛ استفاده مجدد از دارو.

فهرست مطالب

**[چکیده ‌ح](#_Toc123596109)**

[**فهرست مطالب ‌ط**](#_Toc123596110)

[**فهرست شکل ‌ها ‌ك**](#_Toc123596111)

[**فهرست جدول ‌ها ک**](#_Toc123596112)

[فصل 1 مقدمه 1](#_Toc123596113)

[فصل 2 مفاهیم اولیه 5](#_Toc123596120)

[فصل 3 مطالعات پیشین 12](#_Toc123596126)

[فصل 4پیاده سازی 28](#_Toc123596134)

[فصل 5 نتیجه گیری 30](#_Toc123596134)

[**مراجع** 32](#_Toc123596145)

## 

## فهرست شکل‌ها

[شکل 1-1 تکنیک های هوش مصنوعی در کشف دارو: 3](#_Toc123596145)

[شکل 2-1-1 انواع مختلف نمایش دارو در مدلهای مبتنی بر DL: 6](#_Toc123596146)

[شکل 2-1-2 نمایش های مختلف پروتئین در مدلهای مبتنی بر DL: 7](#_Toc123596145)

[شکل 2-2-1معماری های مختلف یادگیری عمیق در استفاده مجدد دارو: 8](#_Toc123596146)

[شکل 3-1-1مدل DeepDTA پیشنهاد شده در سال 2018: 13](#_Toc123596145)

[شکل 3-1-2مدل پیشنهادی WideDTAدر سال 2019: 14](#_Toc123596146)

[شکل3-1-3 معماری مدل AttentionDTA: 15](#_Toc123596146)

[شکل 3-1-4معماری مدل DeepGS: 16](#_Toc123596145)

[شکل 3-1-5معماری مدل پیشنهاد شده GraphDTA: 17](#_Toc123596146)

[شکل 3-1-6نمایش دارو و پروتئین در مدل پیشنهادی DGraphDTA: 19](#_Toc123596145)

[شکل3-1-7 معماری مدل پیشنهادی DGraphDTA : 20](#_Toc123596146)

[شکل3-1-8مدل WGNN-DTA: 21](#_Toc123596145)

[شکل3-1-9معماری مدل WGNN-DTA برای گراف وزندار پروتئین: 22](#_Toc123596146)

[شکل 3-1-10مدل S2GC: 22](#_Toc123596145)

[شکل 3-1-11معماری مدل GSATDTA با استفاده از تزنسفور: 23](#_Toc123596146)

[شکل 4-1 گزارش نتایج پیاده سازی انجام شده با مدل GAT-GAT: 29](#_Toc123596146)

## فهرست جدول‌ها

[جدول 2-3-1 معرفی مجموعه داده: 11](#_Toc123596143)

[جدول 3-1-1مقایسه نتایج حاصل ازWideDTAو سایر مدلهای مطرح شده: 14](#_Toc123596144)

[جدول 3-1-2 نتایج مدل Graph-DTA روی دیتاست Davis در مقایسه با سایر مدل ها: 18](#_Toc123596143)

[جدول 3-1-3 نتایج مدل Graph-DTA روی دیتاست Kiba در مقایسه با سایر مدل ها: 18](#_Toc123596144)

[جدول 3-1-4مقایسه نتایج مدل DGraphDTA : 21](#_Toc123596143)

[جدول 3-1-5 مقایسه نتایج مدل GSATDTA با سایر مدلهای مطرح شده روی دیتاست Davis: 24](#_Toc123596144)

[جدول 3-1-6 مقایسه نتایج مدل GSATDTA با سایر مدلهای مطرح شده روی دیتاست Kiba: 24](#_Toc123596144)

[[جدول 3-1-7خلاصه ای ازپژوهش های انجام شده در پیش بینی DTA: 25](#_Toc123596144)](#_Toc123596144)

**فصل1**

**مقدمه**

**1-1 مقدمه**

هدف اولیه از کشف دارو[[1]](#footnote-1)، تولید داروهای ایمن و موثر برای بیماری های انسان است. فرایند کشف و تولید داروی جدید از مرحله ی شناسایی هدف تا رسیدن به مرحله آزمایشات بالینی بسیار زمان بر و پرهزینه است. به طور معمول حدود 10 تا 15 سال و 0.8 تا 1.5 میلیارد دلارهزینه تولید یک داروی جدید است[1].

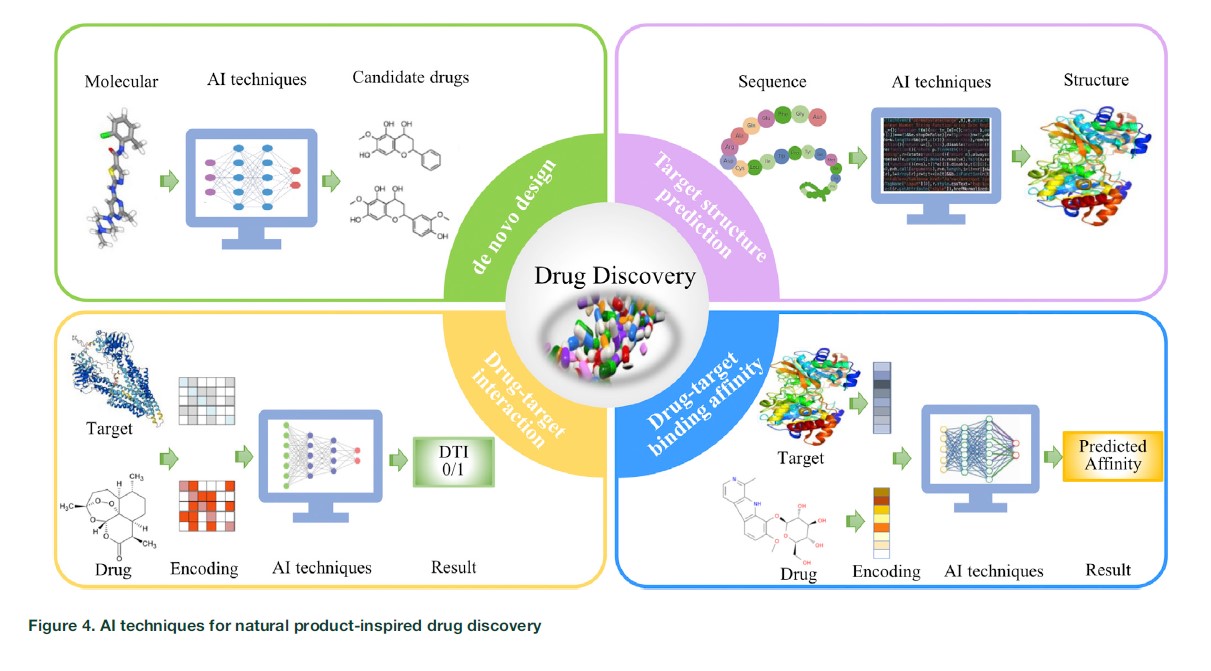
در سالهای اخیر استفاده از هوش مصنوعی[[2]](#footnote-2) (AI) در کشف دارو بسیار مورد توجه قرار گرفته است. هوش مصنوعی در فرایند تولید دارو بهبود قابل توجهی را به همراه دارد. از مزایای استفاده از هوش مصنوعی در فرایند تولید دارو میتوان به کاهش زمان و هزینه های تحقیقات ، کاهش خطا در فرایند تولید، بهبود کیفیت و کارایی دارو، افزایش سرعت و دقت در شناسایی و تولید دارو، پیش بینی خطرات و عوارض جانبی دارو، شناسایی مولکولهای دارویی جدید، شناسایی نقاط ضعف و قدرت مولکولهای دارویی ، بهبود فرآیند طراحی دارو و بهبود فرآیند ارزیابی دارواشاره کرد.

تکنیکهای هوش مصنوعی شامل روشهای مختلفی برای پیش بینی ارتباط میان دارو و پروتئین است. از جمله این روشها میتوان به مدلهای شبکه عصبی، SVM، روشهای رگرسیون خطی و غیرخطی، مدلهای گرافی و غیره اشاره کرد. همچنین از الگوریتمهای ژنتیک، الگوریتمهای تکاملی، الگوریتمهای شبکه عصبی برای پیدا کردن ساختار بهینه دارو و پروتئین استفاده میشود.

استفاده از تکنیک های هوش مصنوعی زمینه ی درک فضاهای بسیار پیچیده بیولوژیکی و استخراج ویژگی ها از داده های ورودیرا فراهم می کند و نتایج چشم گیری را ارائه میدهد. برخی از شرکت های بزرگ داروسازی، مانند Bayer ، Roche و Pfizer نیزفعالیت خود رابرای توسعه روش های مبتنی بر تکنیک های هوش مصنوعی در تولید و کشف دارو آغاز کرده اند. با توجه به نرخ موفقیت پایین، هزینه های بالا و سرعت آهسته کشف و توسعه داروهای جدید، استفاده مجدد از داروهای قدیمی برای درمان بیماری های جدید بسیار مقرون بصرفه است. استفاده مجدد از دارو رویکردی برای تسریع فرآیند کشف دارواست. در واقع به دنبال کشف کاربردهای جدید برای یک داروی موجود هستیم که قبلاً بررسی نشده و در حال حاضر تجویز نشده است.

در فرایند استفاده مجدد از دارو بر اساس داروهای مورد تایید FDA، نیازی به آزمایشات انسانی نیست. همچنین ریسک پایین شکست، چرخه زمانی کوتاهتر، میزان موفقیت بالاتر و سرمایه‌گذاری کمتر از مزایای استفاده مجدد از دارو هستند.

تکنیک های AI تقریبا در تمام جنبه های کشف دارو مورد استفاده قرار گرفته است. در شکل 1 چهار دسته بندی مهم را معرفی و بررسی میکنیم[1].



شکل 1-1 تکنیک های هوش مصنوعی در کشف دارو[1]

**طراحی و تولید داروهای جدید(De novo drug design)**

در سالهای اخیر تکنیک های هوش مصنوعی فرصت های جدیدی را برای طراحی و تولید داروهای جدید فراهم آورده اند و روند کشف دارو را تسریع کرده اند. چارچوب اساسی مدل‌های مبتنی بر یادگیری عمیق برای طراحی داروی جدید در سمت چپ بالای شکل 4 نشان داده شده است.

**پیش بینی ساختار هدف (Target structure prediction)**

بیشتر اهداف داروها، پروتئین ها هستند که نقش مهمی در فعالیت آنزیم ها دارند. تکنیک‌های تجربی که برای تعیین ساختار پروتئین ها پیشنهاد شده اند زمان‌بر و پرهزینه هستند.این روشهای قدیمی تنها ساختار 100000 پروتئین منحصر به فرد که تنها بخش کوچکی از پروتئین های شناخته شده را تشکیل می دهند را رمزگشایی کرده‌اند.

با رشد سریع قدرت محاسباتی و پیشرفت های ناشی از تکنیک های هوش مصنوعی، روشهای جدیدی برای پیش بینی ساختار پروتئین ها پیشنهاد شده است که در سمت راست بالای شکل 4 ارائه شده است.

**پیش بینی تعاملات دارو-هدف[[3]](#footnote-3)** (**DTI prediction**)

پیش بینی DTI نقش حیاتی در توسعه دارو ، استفاده مجدد از دارو، و شناسایی عوارض جانبی احتمالی داروها ایفا می کند. به تعاملات میان مولکولهای دارو و پروتئین های هدف اشاره دارد. فرایند پیش بینی DTI در گوشه پایین سمت چپ شکل 4 نشان داده شده است. ابتدا ترکیبات مولکولهای دارو و

پروتئین ها با استفاده از روشهای Embedding کدگذاری می شوند. سپس از آنها به عنوان ورودی در مدلهای مبتنی بر یادگیری عمیق استفاده می شود که به طور قابل توجهی فرایند کشف دارو را تسهیل می کند.

**پیش بینی ارتباط دارو-هدف[[4]](#footnote-4)** (**DTA prediction**)

برای توصیف قدرت اتصال بین یک مولکول دارو و پروتئین استفاده می شود واطلاعاتی در موردارتباط بین یک مولکول دارو و هدف ارائه می دهد. پیش‌بینی دقیق DTA در توسعه اولیه دارو و در طول فرایند استفاده مجدد از دارو بسیار مهم است.

این پیش‌بینی می‌تواند برای تعیین اینکه آیا مولکول دارو می‌تواند به یک هدف متصل شود یا خیر،استفاده شود. روش مبتنی بر توالی رشته ای[[5]](#footnote-5) و روش مبتنی بر ساختار[[6]](#footnote-6) دو روش متداول در پیش‌بینی DTAهستند.

در این گزارش به بررسی و مقایسه پژوهش های انجام شده در پیش بینی DTA میپردازیم. در فصل 2 به معرفی مدل های یادگیری عمیق و معرفی مجموعه داده ها و معیار های ارزیابی میپردازیم. در فصل 3 مروری بر پژوهش های انجام شده را خواهیم داشت. در فصل 4 خلاصه ای از پیاده سازی انجام شده و در فصل آخر نتیجه گیری از این مطالعات بیان میشود.

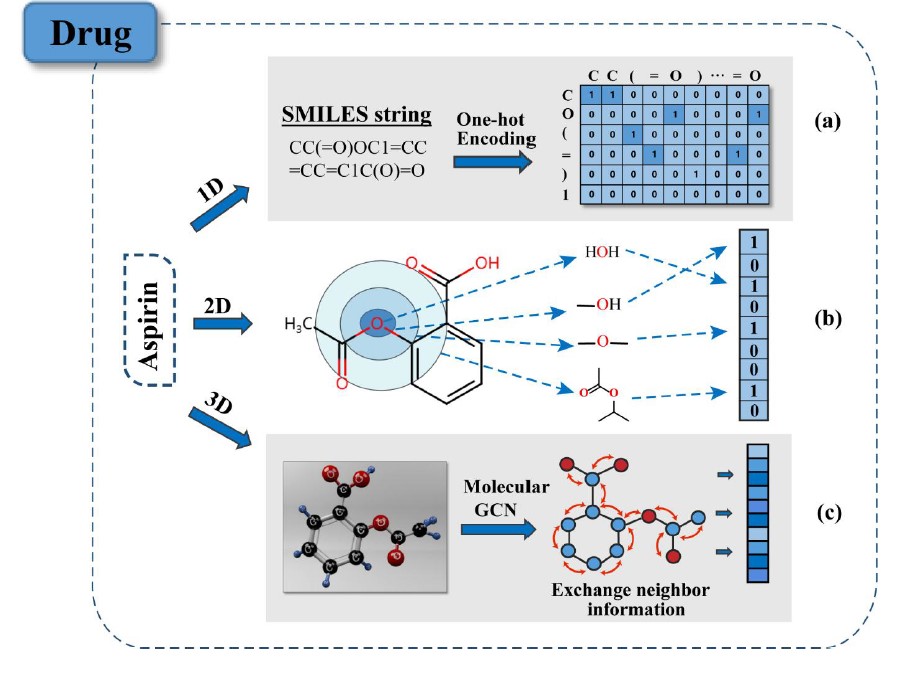
**فصل 2**

**مفاهیم اولیه**

**2-1 انواع نمایش دارو-هدف**

داده های ورودی در مدلهای مبتنی بر یادگیری عمیق[[7]](#footnote-7) (DL ) برای کشف دارو مولکول ها هستند. داروها و اهداف پروتئینی آنها مولکول هایی کوچک یا درشت هستند. برای مشخص کردن این مولکول ها، انواع مختلفی ازنمایش‌های مولکولی استفاده شده است.

بسیاری از روش‌های یادگیری ماشین[[8]](#footnote-8) ML) ) از توالی‌های ساده و ویژگی های مولکولی از پیش تعریف شده به صورت دستی استفاده میکنند. از آنجایی که نوع نمایش داده ها بر روی عملکرد مدل در پیش بینی DTA تاثیر بسزایی دارد مطالعات بسیاری در این مورد انجام شده است. بسیاری از مدل های مبتنی بر DL تمایل دارند از نمایش های سه بعدی استفاده کنند. دارو ها با نمایش های مولکولی مختلفی در شبکه های عصبی عمیق استفاده می شوند که در این بخش معرفی میشوند. انواع مختلف بازنمایی های دارویی بصورت یک بعدی ، دو و سه بعدی که درتحقیقات توسعه دارو استفاده می شوند در شکل 2 نشان داده شده است. در این شکل انواع مختلف نمایش داروی آسپرین را مشاهده میکنیم.

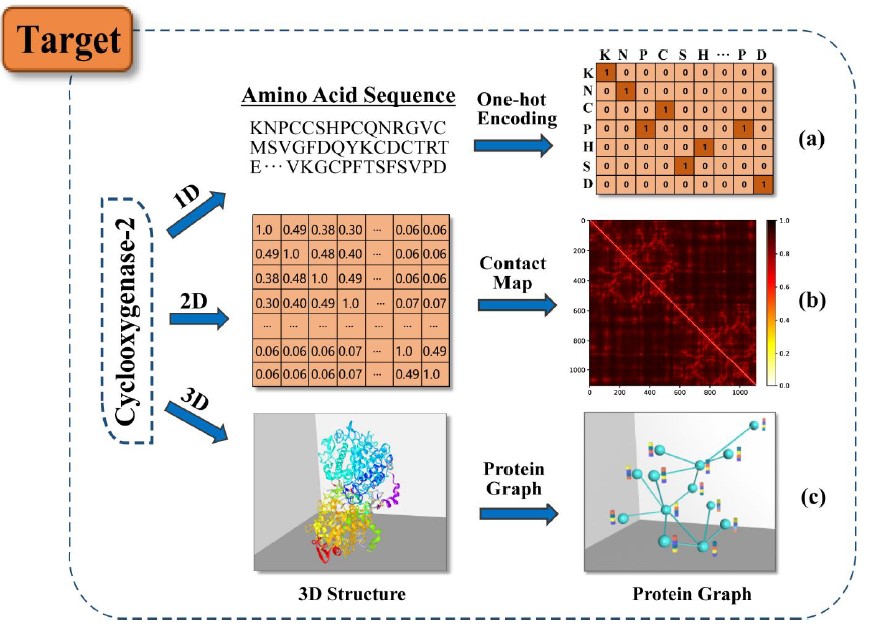


شکل 2-1-1 انواع مختلف نمایش دارو در مدلهای مبتنی بر DL [2]

SMILES[[9]](#footnote-9) یک نماد متنی برای ترجمه ساختار 3بعدی مولکولهای شیمیایی به صورت رشته است که میتواند برای پردازش در نرم افزار های رایانه ای به کار گرفته شود. رشته های SMILES دارو را می توان به یک گراف مولکولی تبدیل کرد به طوری که اتم‌ها و پیوندها را به ‌عنوان گره ها و یال های یک گراف نشان دهیم و به عنوان ورودی در ساختار GNNها استفاده کنیم.

نمایش پروتئین ها در شکل 3 دیده میشود. رشته های پروتئین از 20 آمینو اسید استاندارد تشکیل شده‌اند که هر کدام از

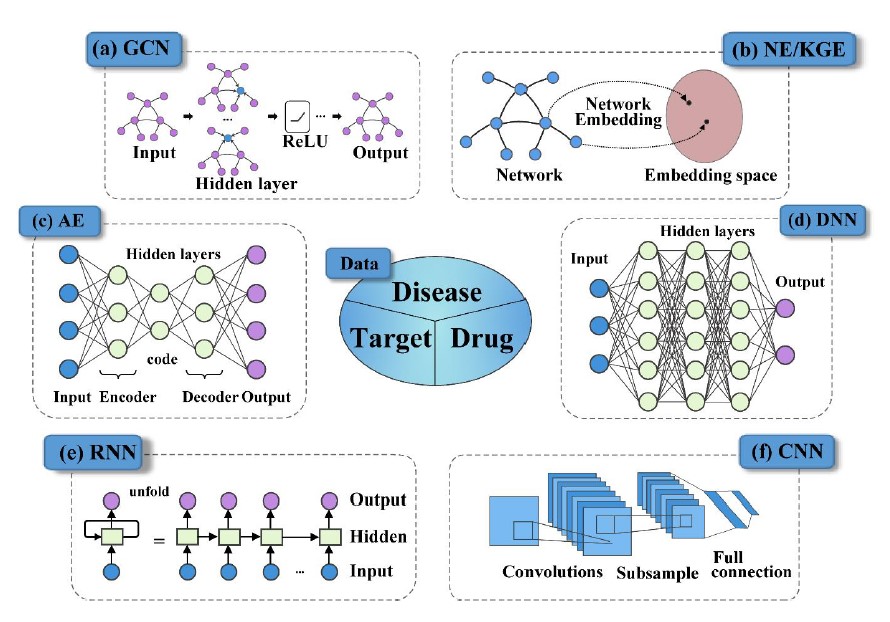
آنها را می توان به سادگی رمزگذاری کرد. از Map Contact برای نمایش دو بعدی پروتئین استفاده میشود و ساختار 3بعدی آنها بصورت گراف برای پردازش توسط GNN بکار می رود.



شکل 2-1-2 نمایش های مختلف پروتئین در مدلهای مبتنی بر DL [2]

2-2**معرفی مدل‌های یادگیری عمیق**

مدل های DL را می توان بر اساس هدف، عملکرد، روش یادگیری و ساختار آنها طبقه بندی کرد. معمولا در توسعه دارو از ترکیب دو یا چند مدل برای دستیابی به نتایج بهتر استفاده میشود. در شکل زیر معماری های مختلف مدلهای DL مشاهده میشود. انواع مختلفی از مدل های DL وجود دارد، اما فقط تعدادی از آنها در این بخش توضیح داده میشود.

****

شکل 2-2-1معماری های مختلف یادگیری عمیق در استفاده مجدد دارو[2]

**2-2-1پرسپترون‌های چندلایه[[10]](#footnote-10) (MLP)**

پرسپترون چند لایه رایج ترین ساختار شبکه عصبی است. MLP ها از یک لایه ورودی و یک لایه خروجی تشکیل شده‌اند که تمامی آن‌ها به هم متصل شده‌اند. این شبکه‌ها دارای تعداد لایه ورودی و خروجی یکسانی هستند اما ممکن است چندین لایه پنهان نیز داشته باشند.

**2-2-2شبکه‌های عصبی کانولوشنی (CNN)[[11]](#footnote-11)**

CNNها چندین لایه دارند که ویژگی‌ها را پردازش و از داده‌ها استخراج می‌کنند و به طور گسترده‌ برای شناسایی تصاویر ماهواره‌ای، پردازش تصاویر پزشکی، پیش بینی سری‌های زمانی و در فرایند توسعه دارو مورد استفاده قرار می‌گیرند.

**2-2-3شبکه‌های حافظه کوتاه مدت طولانی[[12]](#footnote-12) (LSTM)**

نوعی شبکه عصبی بازگشتی هستند که می‌توانند وابستگی‌های طولانی مدت را یاد بگیرند و به خاطر بسپارند. این شبکه‌ها اطلاعات را در طول زمان حفظ می‌کنند و در پیش بینی سری‌های زمانی مفید هستند زیرا ورودی‌های قبلی را به خاطر می‌آورند. LSTM ‌ها دارای یک ساختار زنجیره‌ای هستند که در آن، چهار لایه در تعامل با هم، به روشی منحصر به فرد ارتباط برقرار می‌کنند. LSTM‌ها معمولاً برای تشخیص گفتار و توسعه دارو نیز استفاده می‌شوند.

**2-2-4شبکه‌های عصبی بازگشتی[[13]](#footnote-13) (RNN)**

 RNNها، دارای اتصالاتی هستند که حلقه‌هایی جهت­‌دار را تشکیل می‌دهند، به خروجی LSTM اجازه می‌دهد تا به عنوان ورودی به فاز فعلی تغذیه شود. خروجی LSTM، ورودی مرحله فعلی می‌شود و می‌تواند ورودی‌های قبلی را با توجه به حافظه داخلی خود حفظ کند . RNN ‌ها معمولاً برای زیرنویس تصاویر، تجزیه و تحلیل سری‌های زمانی، پردازش زبان طبیعی، تشخیص دست خط و ترجمه ماشین و و توسعه دارو نیز استفاده می‌شوند.

**2-2-5شبکه‌های مولد متخاصم[[14]](#footnote-14) (GAN)**

GAN‌ها، از مدلهای  یادگیری عمیق  مولد هستند که نمونه‌های جدیدی از داده‌ها را شبیه به داده‌های آموزشی ایجاد می‌کنند. GAN دو بخش اصلی دارد:

یک مولد (generator) که یاد می‌گیرد داده‌های جعلی تولید کند و یک متمایزگر یا تشخیص دهنده (discriminator) ، که از این اطلاعات نادرست درس می‌گیرد.

استفاده از GAN‌ها طی یک دوره زمانی بسیار افزایش یافته است. توسعه دهندگان بازی‌های ویدئویی، از GAN برای ارتقای بافت‌های با وضوح پایین و دو بعدی در بازی‌های ویدئویی قدیمی با بازآفرینی آن‌ها در وضوح k4 یا بالاتر، از طریق آموزش تصاویر استفاده می‌کنند. GAN‌ها به تولید تصاویر واقعی و شخصیت‌های کارتونی، ایجاد عکس از چهره انسان و ارائه اشیاء سه بعدی کمک می‌کنند.

**2-2-6 اتوانکدرها[[15]](#footnote-15) یا خودرمزگذارها**

خودرمزگذارها ([Autoencoders](https://www.simplilearn.com/other-forms-of-deep-learning-tutorial))، نوع خاصی از شبکه عصبی پیشخور[[16]](#footnote-16) هستند که در آن‌ها، ورودی و خروجی یکسان است. خودرمزگذارها شبکه‌های عصبی آموزش دیده‌ای هستند که داده‌ها را از لایه ورودی به خروجی همانندسازی (replicate) می‌کنند. از خودرمزگذارها برای اهدافی مانند اکتشافات دارویی، پیش بینی محبوبیت و پردازش تصویر استفاده می‌شود. یک خودرمزگذار از سه قسمت اصلی تشکیل شده است: رمزگذار، رمز و رمزگشا.

خودرمزگذارها به گونه‌ای ساختار یافته‌اند تا ورودی را دریافت کرده و آن را به یک بازنمایی متفاوت تبدیل کنند. سپس سعی می‌کنند ورودی اصلی را تا آنجا که ممکن است، به درستی بازسازی کنند.

**2-2-7شبکه عصبی گراف (GNN) [[17]](#footnote-17)**

گراف‌ها ساختارهای داده‌ای قوی هستند که شامل روابط بین اشیا هستند.کاربرد های زیادی در بینایی ماشین وپردازش زبان طبیعی و پزشکی و نوسعه دارو دارند. GNN‌هابرای انجام وظایف پیش‌بینی در سطح گره، سطح لبه و سطح گراف مفید هستند.گره‌ها و یال ‌های گراف ساختار یک مولکول را نشان می‌دهند. گره ها اتم هستند و یال ها پیوندهای شیمیایی هستند. از گراف برای تجزیه و تحلیل ساختار مولکول‌ها استفاده می شود.

**2-2-8مکانیسم توجه[[18]](#footnote-18) در یادگیری عمیق**

مکانیسم توجه در یادگیری عمیق تکنیکی است که برای بهبود عملکرد یک شبکه عصبی با اجازه دادن به مدل برای تمرکز روی مهم‌ترین داده‌های ورودی در حین تولید پیش‌بینی استفاده می‌شود.این کار با وزن دادن به داده های ورودی انجام می شود تا مدل برخی از ویژگی های ورودی را بر سایرین اولویت دهد. در نتیجه، مدل می‌تواند پیش‌بینی‌های دقیق‌تری را تنها با در نظر گرفتن مهم‌ترین متغیرهای ورودی تولید کند. مکانیسم توجه اغلب در کارهای پردازش زبان طبیعی مانند ترجمه ماشینی استفاده می شود، جایی که مدل باید به بخش های مختلف عبارت ورودی توجه کند تا معنای آن را به طور کامل درک کند و ترجمه مناسب ارائه دهد.

**2-2-9ترنسفورمر[[19]](#footnote-19)**

شبکه عصبی ترنسفورمر یک مدل یادگیری عمیق است که با مکانیسم توجه به خود[[20]](#footnote-20)، به‌طور متفاوتی اهمیت هر بخش از داده‌های ورودی را وزن می‌کند. عمدتاً در زمینه‌های پردازش زبان طبیعی و بینایی ماشین استفاده می‌شود.

ترنسفورمرها لزوماً داده‌ها را به ترتیب پردازش نمی‌کنند. ترنسفورمر در سال 2017 توسط تیمی در Google Brain معرفی شد و از آن زمان به مدلی پیشرفته برای بسیاری از مشکلات NLP تبدیل شده اند. Self-attention به مدل اجازه می دهد تا اهمیت و ارتباط هر بخش از داده های ورودی را بیاموزد و کل ورودی را به عنوان یک عنصر در یک زمان پردازش کند.

به عنوان مثال، اگر داده ورودی یک جمله باشد، self-attention می تواند به مدل کمک کند تا تشخیص دهد کدام کلمات برای درک معنای جمله اهمیت بیشتری دارند. این ویژگی اجازه می‌دهد تا موازی سازی بیشتر از شبکه های عصبی بازگشتی  باشد و بنابراین زمان آموزش را کاهش می‌دهد.

**2-3مجموعه داده**

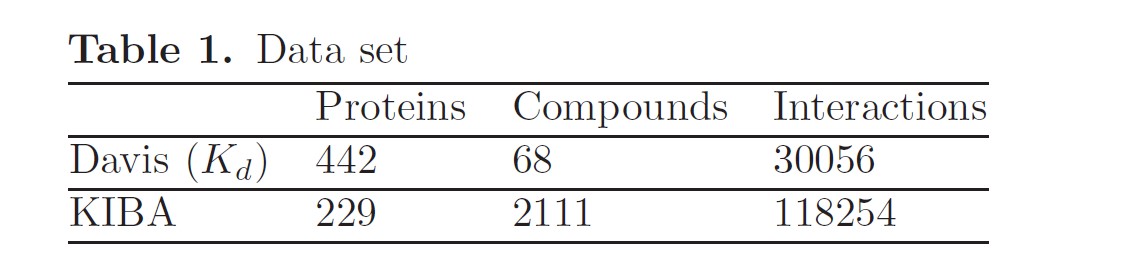
مجموعه داده های معیاری که برای ارزیابی عملکرد مدل پیش بینی DTA استفاده می شود شامل دو مجموعه داده Davis و KIBAاست که در اکثر مقالات مورد استفاده قرار گرفته است.در ادامه بطور مختصر هر دو مجموعه داده معرفی میشوند.

**2-3-1معرفی مجموعه داده ها**

Davis یک مجموعه داده از Kinase است که توسط (دیویس و همکاران سال2011) معرفی شد شامل 25046 جفت دارو- هدف است.

مجموعه داده Kiba توسط (تانگ و همکاران سال2014) ارائه شده است. رشته های SMILES ازهر دو مجموعه داده از پایگاه داده PubChem و رشته های پروتئین از پایگاه داده پروتئین UniProt استخراج شده است.

مشخصات هردو مجموعه داده در جدول زیر آورده شده است:



جدول 2-3-1 معرفی مجموعه داده

**2-4معیار های ارزیابی مدل پیشنهادی**

برای ارزیابی مدل از concordance index(CI) و mean square error (MSE) استفاده میشود.

CI میزان احتمال شباهت بین مقادیر واقعی و پیش بینی شده را اندازه گیری می کند، فرض کنید δi و bi به ترتیب بیانگر مقدار واقعی و مقدار پیش‌بینی‌شده نمونه i باشند. بنابراین CI به صورت زیر تعریف می شود:

(1)

MSE نشان دهنده تفاوت بین مقادیر پیش بینی شده و مقادیر واقعی است در طول آموزش، یک مدل تلاش می کند تا اختلاف بین مقدار واقعی و پیش بینی شده را کاهش دهد.

(MSE) به عنوان loss function استفاده میشود، در MSE ، P بردار پیش بینی است و Y مربوط به بردار خروجی های واقعی است و n تعداد نمونه ها را نشان می دهد:

(2)

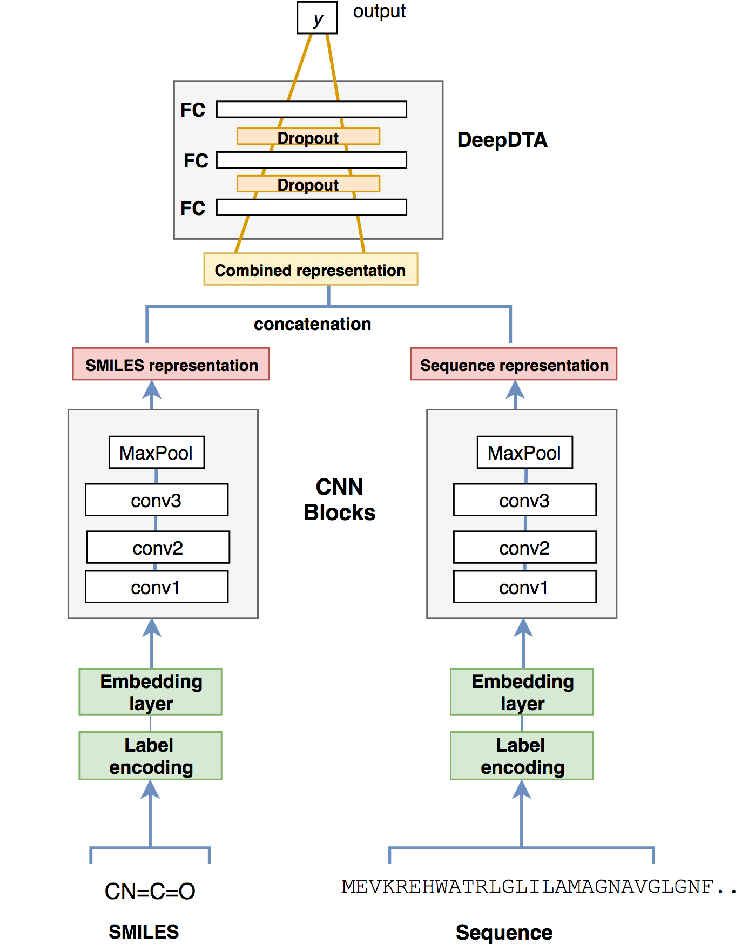
**فصل3**

**مرورمطالعات پیشین**

**3-1 مرور مطالعات انجام شده**

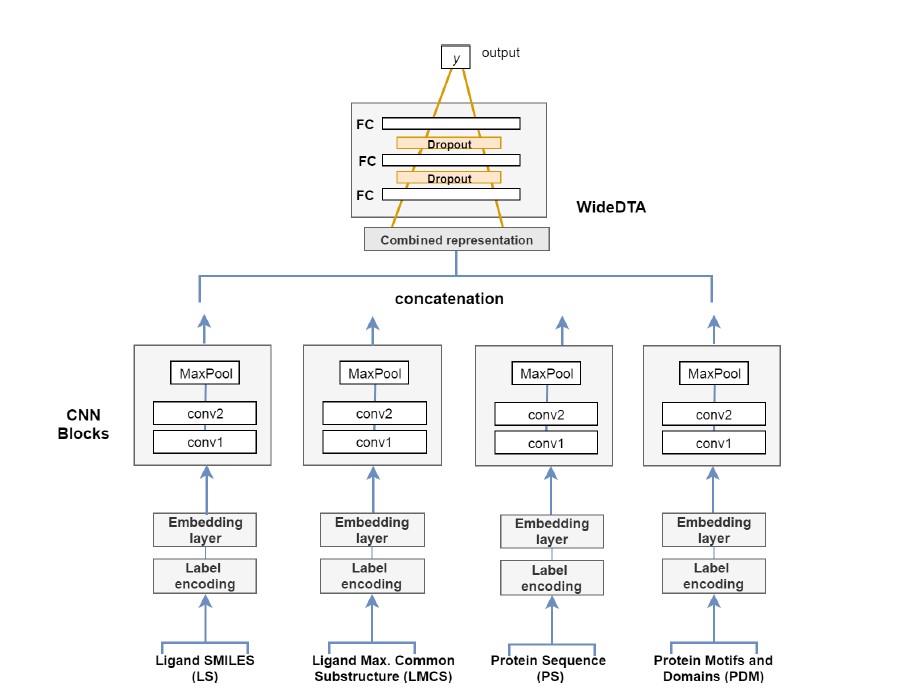
لی و همکاران در سال2015 یک روش مبتنی بر جنگل تصادفی[[21]](#footnote-21) را پیشنهاد کردند[3]. این مدل RF همان سال در KronRLS، نیز مورد استفاده قرار گرفت که از similarity matrixبرای بهبود عملکرد پیش‌بینی DTA استفاده کرد[4].در سال 2017مدل SimBoost مطرح شد که یک بردار ویژگی را با هر جفت دارو-هدف مرتبط می کند از جفت‌هایی با ویژگی‌های اتصال مشاهده‌شده، یک مدل رابرای یادگیری آموزش می‌دهد[5].

در سال 2018 مدل DeepDTA [6] معرفی شد، اولین مدل مبتنی بر یادگیری عمیق است که از ساختار 3بعدی داده ها استفاده نکرده است و از رشته های SMILES یک بعدی برای نمایش مولکول های دارو استفاده کرده است. این مدل از شبکه های عصبی کانولوشنCNN) [[22]](#footnote-22)) برای یادگیری و نمایش رشته های پروتئین و رشته های SMILES استفاده کرد. عملکرد DeepDTA هنگام استفاده از دو بلوک CNN برای استخراج ویژگی های داروها و پروتئین ها در مقایسه با KronRLS و الگوریتم های SimBoost به طور قابل توجهی افزایش یافته است. در ادامه نمایش‌های حاصل از بلوک های CNN راباهم ترکیب کرده تا به یک لایه کاملاً متصل وارد شوند و DTA را پیش‌بینی می کند.

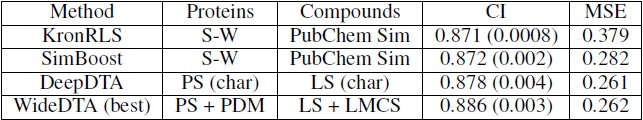


شکل 3-1-1مدل DeepDTA پیشنهاد شده در سال 2018[6]

WideDTA [7]در سال 2019 معرفی شده است توسعه ای از DeepDTA است. ازدو لایه یک بعدیCNN برای بازنمایی داروها و پروتئین ها استفاده می کند. در WideDTAتوالی SMILES و پروتئین به‌عنوان کلمات نظرگرفته شده اند و به ترتیب با یک دنباله هشت کاراکتری و یک دنباله سه کاراکتری ، نشان داده شده اند. این مدل چهار قطعه متنی مختلف از اطلاعات مربوط به پروتئین ها و لیگاندها را ترکیب می کند. در جدول 3-1-1 نتایج این مدل در مقایسه با DeepDTA ودو مدل قبلی ارائه شده است.



شکل 3-1-2مدل پیشنهادی WideDTAدر سال 2019[7]



جدول 3-1-1مقایسه نتایج حاصل ازWideDTAو سایر مدلهای مطرح شده[7]

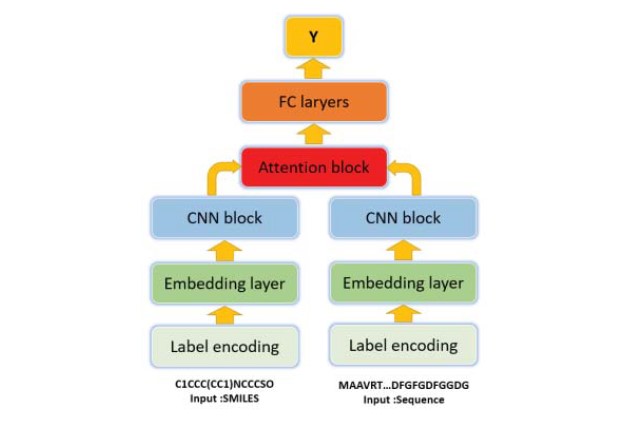
DeepAffinity [8] یک مدل یادگیری عمیق نیمه نظارت‌شده [[23]](#footnote-23)که ازترکیب RNN وCNN برای پیش‌بینی DTA استفاده می‌کند. از داده‌های بدون برچسب و برچسب‌گذاری‌شده، برای رمزگذاری مشترک مولکول های دارو و پروتئین وهمینطور از مکانیسم‌ توجه برای تمرکز برروی بخشهای مهم پروتئین و دارو استفاده میکند تا پیش بینی دقیق تری را ارائه دهد.

مدلی مبتنی بر یادگیری عمیق ، به نام AttentionDTA [9]، از مکانیسم توجه برای پیش‌بینی DTA استفاده می‌کند. متفاوت از مدل‌هایی است که از ساختارهای سه‌بعدی برای دارو-هدف یا نمایش گرافی داروها و پروتئین‌ها استفاده می‌کنند، با استفاده از مکانیسم توجه روی بخشهایی از رشته های دارو و پروتئین که هنگام پیش‌بینی میل اتصال آنها مهم هستند، تمرکز میکند.

این مکانیسم به مدل اجازه می دهد تا بازنمایی های مؤثرتری از دارو و پروتئین را به دست آورد.دو بلوک کانولوشن 3 لایه جداگانه برای بازنمایی دارو و پروتئین بکار گرفته و سپس یک بلوک توجه در ادامه قرار گرفته است.

AttentionDTA چندین مزیت دارد. اول، پیش‌بینی میل پروتئینی دارو را به عنوان یک مساله رگرسیون در نظر می‌گیرد، که از تأثیر انتخاب نمونه منفی بر مدل جلوگیری می‌کند و اطلاعات کاربردی و مفیدتری ارائه می‌دهد. دوم، مکانیسم توجه تجسم[[24]](#footnote-24) خوبی را ارائه می‌کند و به ما امکان می‌دهد مدل را به خوبی تجزیه و تحلیل کنیم.

در مقایسه با DeepDTA نتایج نشان می دهد که مدل مبتنی بر توجه می تواند به طور موثرتری اطلاعات و ویژگیهای دارو و پروتئین را استخراج کند.



شکل3-1-3 معماری مدل AttentionDTA [9]

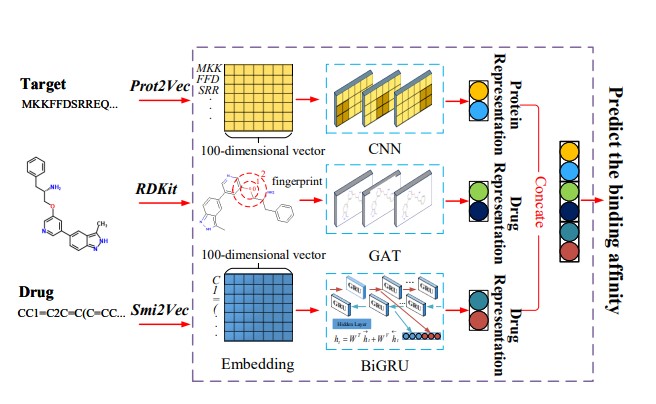
GANsDTA [10]یک مدل نیمه نظارت شده[[25]](#footnote-25)جدید مبتنی بر شبکه‌های متخاصم مولد (GANs)[[26]](#footnote-26) برای پیش‌بینی DTA بین رشته های SMILES و پروتئین است. مدل GANsDTA می‌تواند با استفاده از دو GAN بر روی داده‌های برچسب‌دار و بدون برچسب، ویژگی‌های پروتئین‌ها و داروها را استخراج کند.

این مدلهای یادگیری عمیق داروها را به شکل string نمایش میدادند که با ساختار طبیعی مولکولها متفاوت است. این روش‌ها نشان داده‌اند که شبکه‌های عمیق بهتر از الگوریتم‌های یادگیری ماشین سنتی می‌توانند عمل کنند.

هنگام استفاده ازرشته ها، اطلاعات ساختاری مولکولها از بین می رود ، که می تواند قدرت پیش بینی یک مدل و همچنین عملکرد آنرا مختل کند. در سالهای اخیر شبکه عصبی گراف (GNN) به طور گسترده در زمینه های مختلف استفاده شده است. ازیک گراف متشکل از گره ها و یال ها به عنوان ورودی GNN استفاده می شود و هیچ محدودیتی برای اندازه گراف ورودی وجود ندارد ، که قابلیتی انعطاف پذیر برای استخراج اطلاعات مولکول ها را ارائه می دهد.

شبکه کانولوشن گراف[[27]](#footnote-27) (GCN) و شبکه توجه گراف[[28]](#footnote-28) (GAT) مدل‌های پرکاربرد GNN هستند و به تدریج در توسعه دارو، مانند پیش‌بینی خواص دارو و همینطور پیش بینی DTA استفاده شده‌اند.

[11]DeepGS روش دیگری است که از شبکه های عصبی عمیق[[29]](#footnote-29) برای استخراج ویژگی های رشته های SMILES دارو و رشته های پروتئین استفاده می‌کند. از تکنیکهای Embedding پیشرفته مانند prot2vec و Smi2vec برای رمزگذاری رشته های SMILES و آمینو اسید های پروتئین استفاده میکند. از آنجایی که ساختار دارو را با گراف نمایش داده است با استفاده از GAT به استخراج ویژگیهای گراف دارو و با استفاده از BiGRU به اطلاعات شیمیایی آنها دستیابی پیدا می کند.

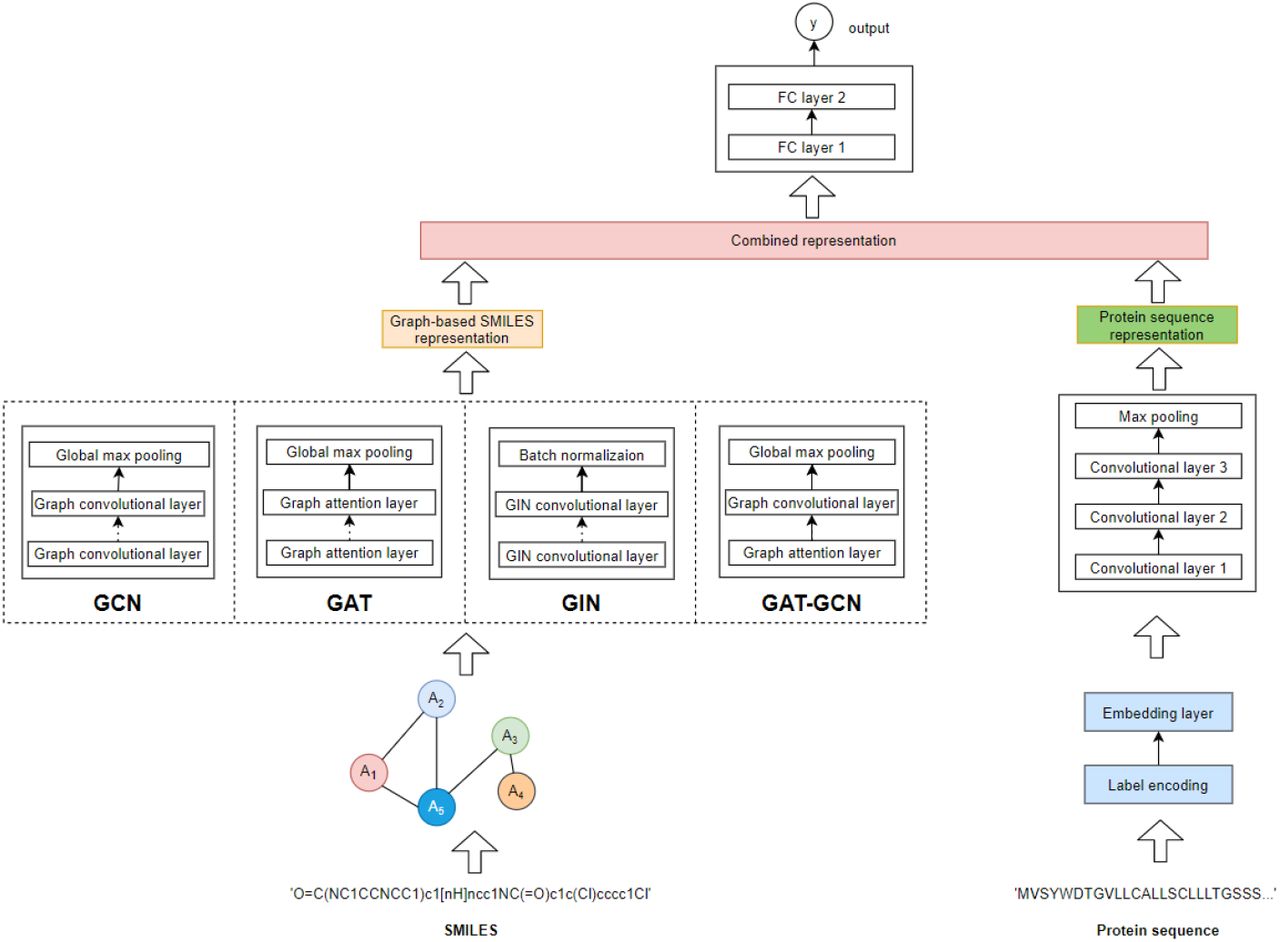


شکل 3-1-4معماری مدل DeepGS []

DTA [12]Graph- در سال 2020 نمایش گراف را برای استفاده از اطلاعات ساختار مولکولها ی دارو معرفی کرد و از GNN برای پیش بینی DTA استفاده کرده است. نتایج تأیید می‌کند که در مدل‌های یادگیری عمیق برای پیش‌بینی DTA، نمایش داروها به صورت گراف می تواند منجر به بهبود عملکرد مدل شود. این مدل ترکیبات دارویی را به عنوان گرافی از فعل و انفعالات بین اتم در نظر گرفته و مدل را حول این مفهوم پیاده سازی میکند. برای توصیف یک گره در گراف، از مجموعه ای از ویژگی های اتمی استخراج شده از DeepChem (رامسوندار و همکاران، 2019) استفاده می کند. هر گره ، یک بردار ویژگی باینری چند بعدی است که شامل پنج مورد اطلاعات زیر است:

نماد اتم، تعداد اتم های مجاور، تعداد هیدروژن های مجاور، implicit value اتم و aromatic structure**.** رشته های SMILES را بااستفاده از RDKit و استخراج ویژگی های اتمی به گراف مولکولی متناظرش تبدیل میکند. یک مدل پیش‌بینی DTA را بر اساس ترکیبی از شبکه های عصبی گراف و CNN پیشنهاد می‌کند. برای پروتئین ها از رشته ای ازکاراکترهای ASCII استفاده می کند و چندین لایه CNN 1 بعدی را روی آنها اعمال میکند.

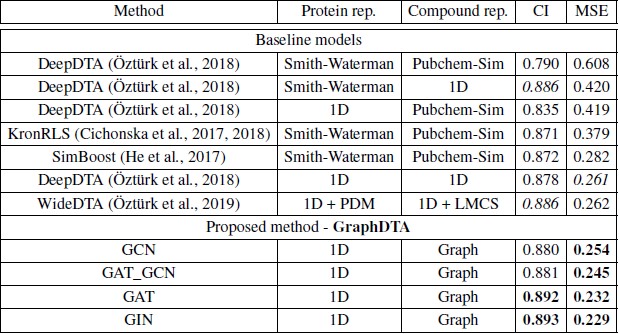
برای دارو از 4مدل مختلف GNN شاملGCN ، GAT ، GIN و ترکیبی از GAT-GCN استفاده میکند.

****

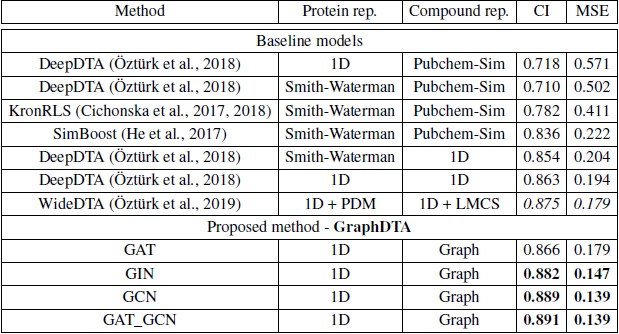
شکل 3-1-5معماری مدل پیشنهاد شده GraphDTA [12]

GraphDTA در 3 مرحله کار می کند : ابتدا کد SMILESیک دارو به یک گراف مولکولی تبدیل می‌شود و یک الگوریتم یادگیری عمیق، نمایش گراف را می‌آموزد. در همین حال، رشته پروتئین کدگذاری و تعبیه شده است و با عبور از چندین لایه کانولوشن 1 بعدی ویژگی هایش بدست می آید. در نهایت، دو بردار بدست آمده با هم ادغام میشوند و از چندین لایه کاملاً متصل عبور می کنند تا خروجی را تخمین بزنند.

نتایج حاصل از این مدل روی مجموعه داده های Kiba وDavis در جدول های زیر مشاهده می شود.



جدول 3-1-2 نتایج مدل Graph-DTA روی دیتاست Davis در مقایسه با سایر مدل ها[12]

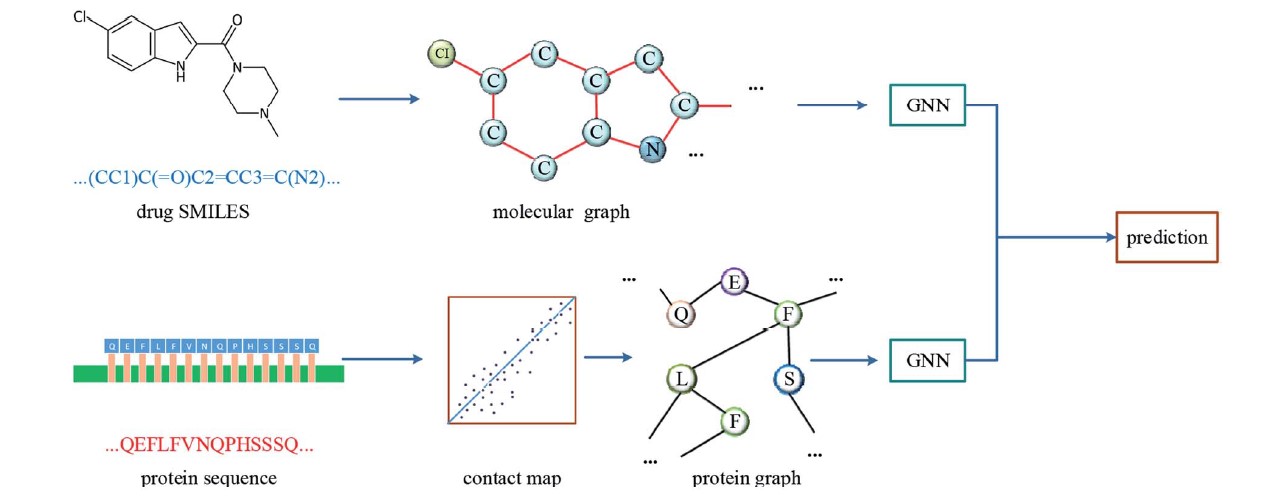


جدول 3-1-3 نتایج مدل Graph-DTA روی دیتاست Kiba در مقایسه با سایر مدل ها[12]

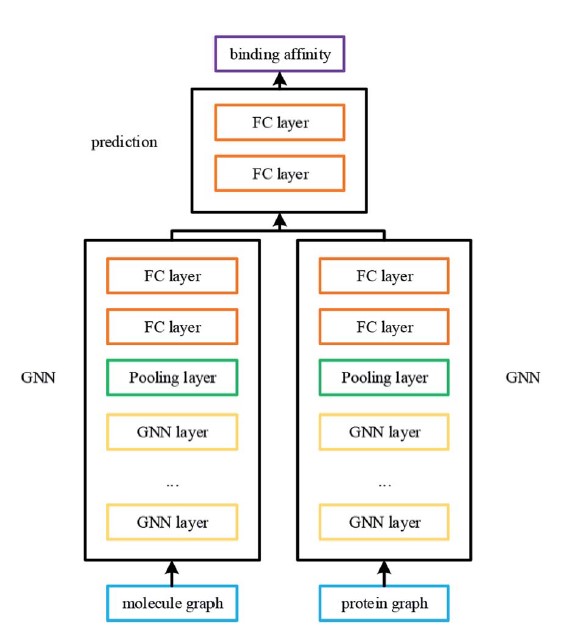
نتایج نشان می دهد که GraphDTA نسبت به سایر مدلهای مطرح شده عملکرد بهتری در پیش بینی DTAداشته است. اما چالشی که مطرح میشود این است که در GraphDTA، CNN برای به دست آوردن ویژگی‌های پروتئینی از طریق رشته ها مورد استفاده قرار گرفت و گرافی برای هر پروتئین ایجاد نشد.

DGraphDTA [13] مدلی است که علاوه بر دارو، پروتئین ها را هم به صورت گراف در نظر میگیرد و برای پیش بینی DTA پیشنهاد شده است.

گراف پروتئین بر اساس خروجی contact map ساخته شده است که می‌تواند ویژگی‌های ساختاری پروتئین را با توجه به توالی رشته های آن پیش‌بینی کند. این مدل هر دو گراف دارو و گراف پروتئین را ترکیب می کند تا دقت پیش بینی DTA را تا حد زیادی بهبود بخشد. استفاده از GNN و contact map می‌تواند تأثیر مثبتی بر دقت پیش‌بینی ارتباط دارو-هدف داشته باشد. در دو شکل زیر چگونگی تشکیل گراف های دارو و هدف و معماری این مدل نشان داده شده است.

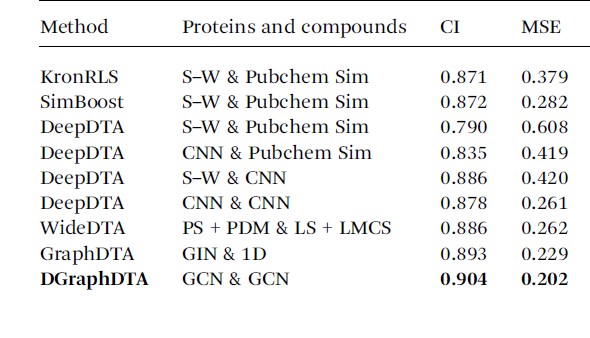


شکل 3-1-6نمایش دارو و پروتئین در مدل پیشنهادی [13]DGraphDTA



شکل3-1-7 معماری مدل پیشنهادی [13] DGraphDTA

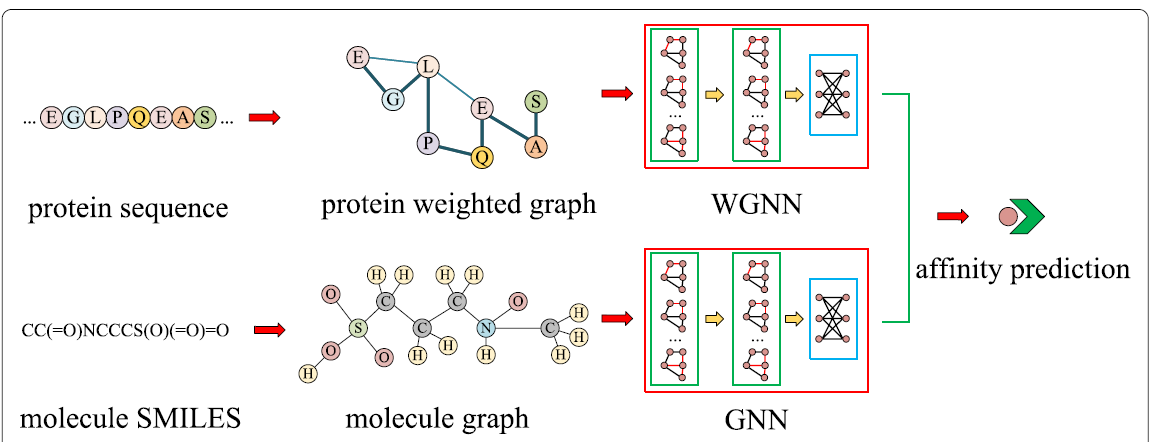
این روش بر روی مجموعه داده‌های Kiba و Davis آزمایش شد و قابلیت تعمیم[[30]](#footnote-30) قوی را نشان داد و از سایر روش‌ها بهتر عمل کرد.همانطور که در دو جدول زیر مشاهده می شود در مقایسه با DeepDTA، WideDTA و GraphDTA این مدل پیشنهادی با GCN های سه لایه بهبود عملکرد قابل توجهی دارد.



جدول 3-1-4مقایسه نتایج مدل [13] DGraphDTA

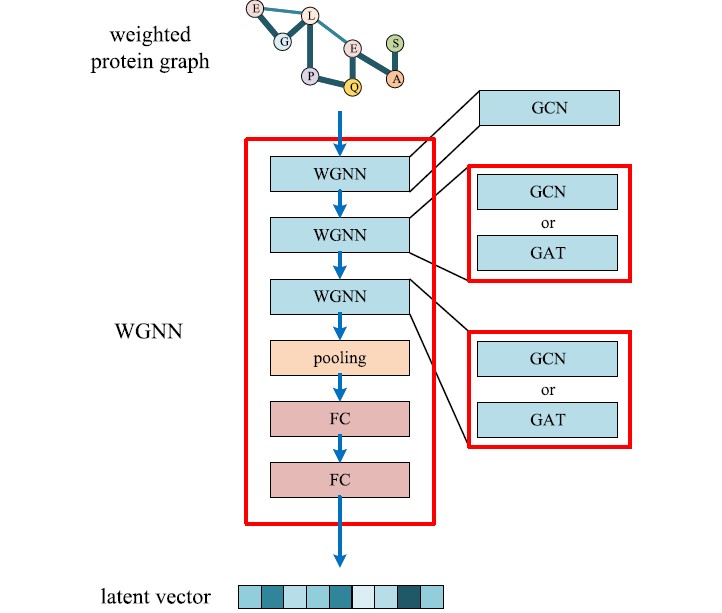
WGNN‑DTA [14] یک مدل گراف وزندار است که می تواند برای پروتئین ها گراف وزندار مناسبی ایجاد کند و آزمایش های مختلف نشان می دهد این مدل برای پیش بینی DTA سریعتر از مدل DGraphDTA است و دقت بالایی دارد.

ورودی ها ی مدل رشته پروتئین و رشتهSMILES است. یک گراف وزنی برای پروتئین ساخته شده است که می تواند به طور جامع اطلاعات ساختار پروتئین را شرح دهد و به جای مقدار باینری که در DGraphDTA استفاده می شود، از احتمال تعامل به عنوان وزن یال های گراف استفاده می شود.



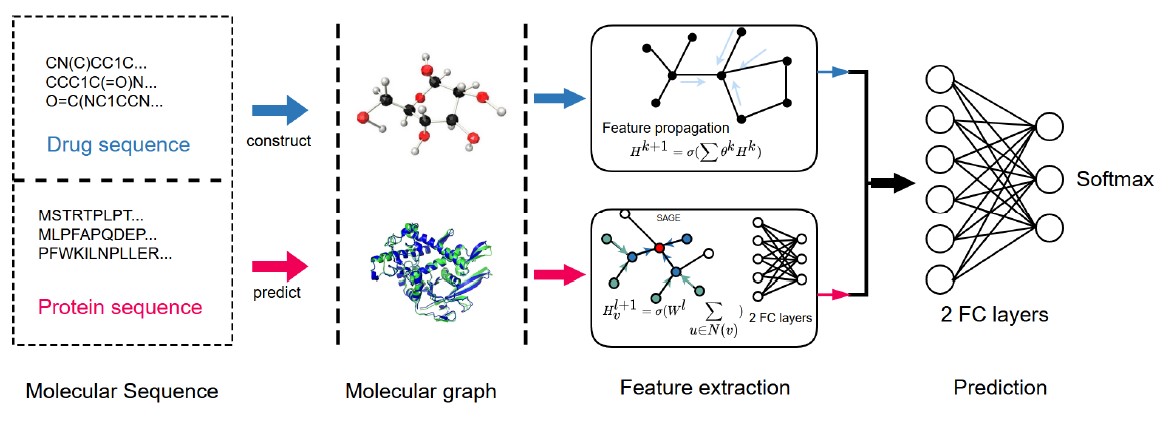
شکل3-1-8مدل WGNN-DTA ] 14 [

از دو نوع GNN، (GCN) و (GAT) برای استخراج بردارهای نهفته گراف های پروتئین و دارو ، و پیش‌بینی DTAبر اساس بردارهای نهفته استفاده میکند.



شکل3-1-9معماری مدل WGNN-DTA برای گراف وزندار پروتئین ] 14 [

S2GC [15] مدل مبتنی بر یادگیری عمیق که از GNNبرای پیش بینی DTAاستفاده میکند. رشته های پروتئین و رشته های SMILES دارو به گراف تبدیل میشوند. برخلاف سایرمدل ها که از GCN های 2 و3 لایه استفاده میکنند این مدل از تعداد لایه های خیلی بیشتر حدود 16لایه برای GCN استفاده کرده است.

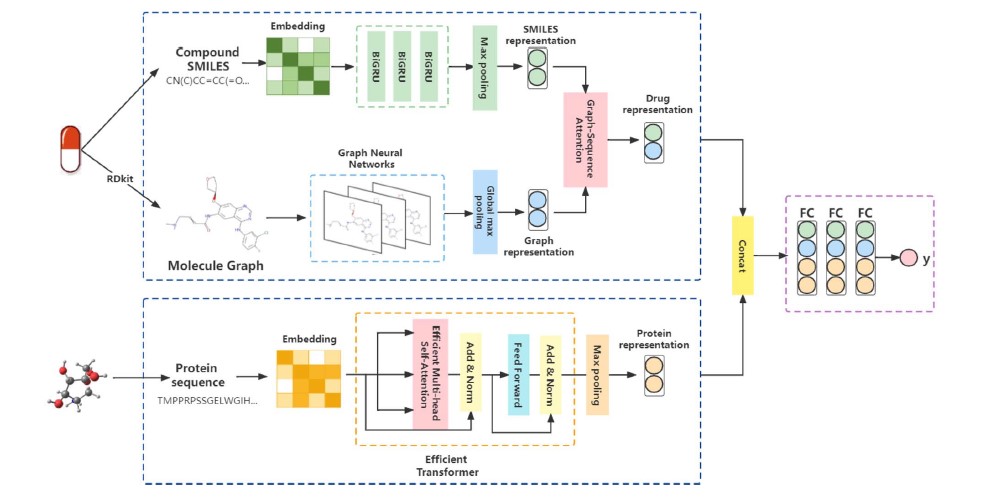


شکل 3-1-10مدل S2GC [15]

FusionDTA [16].با استفاده از BiLSTM و Multi-head linear attention ، DTA را پیش بینی میکند.

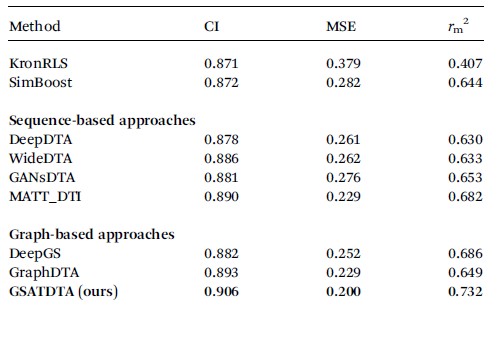
Multiscale-DTA [17] از Multi-scaleCNNs و مکانیسم توجه برای پیش بینی DTA استفاده میکند. ابتدا رشته های دارو و پروتئین با روشهایEmbedding کد گذاری میشوند . سپس با عبور از دو بلوک CNN سه لایه و مکانیسم توجه ویژگی های چند مقیاسی و جامع مولکول های دارو و هدف استخراج میشوند وبعد از ترکیب با هم وارد لایه های FCمیشوند تا DTA را پیش بینی کند.

در سال 2022 مدل جدیدی به نام GSATDTA [18] معرفی شد.این مدل از مکانیسم توجه به خود در ساختار ترنسفورمر برای استخراج ویژگی های رشته های پروتئین استفاده میکند. برای دارو 2 مرحله را برای استخراج ویژگی ها به کار می برد. ابتدا بعد از Embedding رشته های SMILES دارو از 3 لایه BiGRU و یک لایه ادغام در ادامه استفاده میکند.در مرحله بعدی بعد از تبدیل ساختار دارو به گراف از GNN ها برای پردازش آن استفاده میکند در نهایت ویژگیهای بدست آمده از مکانیسم توجه عبور میکنند تا به نمایش نهایی برسد. Graph–sequence attention به شبکه اجازه میدهد تا روی مرتبط ترین بخش داده های ورودی تمرکز کند و دقیقترین اطلاعات و ویژگیها را از رشته های SMILES و گراف دارو استخراج کند. معماری و ساختار این در شکل زیر مشاهده میشود.

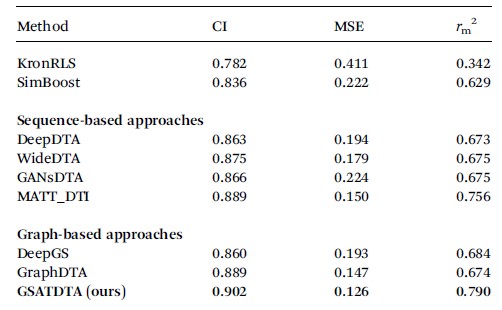


شکل 3-1-11معماری مدل GSATDTA با استفاده از تزنسفور [18]

دراغلب مدل هایی که مطرح شدند از CNN برای استخراج و نمایش ویژگی های رشته های پروتئین استفاده می شود، و یکسری از ارتباطات میان رشته های بزرگ و طولانی آمینو اسیدهای پروتئین که فاصله زیادی از هم دارند نادیده گرفته میشوند با استفاده از ترنسفورمر این long-distance relationships در رشته های آمینو اسید قابل شناسایی و دستیابی است. multi-head self-attention در ترنسفومر با استفاده از انکدر و دیکدر ویژگیهای رشته های پروتئین را از موقعیت های مختلف رشته استخراج میکند و در نهایت پیش بینی دقیقتری از DTA ارائه خواهد شد. نتایج عملکرد این مدل در دو جدول زیر قابل مشاهده است.



جدول 3-1-5 مقایسه نتایج مدل GSATDTA با سایر مدلهای مطرح شده روی دیتاست Davis [18]



جدول 3-1-6 مقایسه نتایج مدل GSATDTA با سایر مدلهای مطرح شده روی دیتاست Kiba [18]

در جدول زیر خلاصه ای از پژوهش های انجام شده در پیش بینی DTA ذکر شده است.

جدول 3-1-7 خلاصه ای ازپژوهش های انجام شده در پیش بینی DTA

|  |  |  |
| --- | --- | --- |
| مقاله | روش | توضیحات |
| KronRLS  [2015] | از جنگل تصادفی و  Similarity matrix برای پیش بینی DTA استفاده کرد. | اولین مدل برای پیش‌بینی DTA بدون استفاده از داده‌های ساختاری سه‌بعدی است. یک روش مبتنی بر ML است که از ماتریس‌های شباهت هدف-هدف و شباهت دارو-دارو به عنوان ویژگی برای پیش‌بینی مقادیرDTA استفاده می‌کند. |
| SimBoost  [2017] | یک روش ماشین لرنینگ است  از gradient boosting  برای پیش بینی DTAاستفاده میکند. | این مدل از ماتریس های شباهت و ویژگی های ساخته شده به دست آمده از شباهت هدف-هدف، شباهت دارو-دارو، و شبکه های دارو-هدف برای پیش بینی DTA استفاده میکند.  SimBoost عملکرد پیش بینی را در مقایسه با روش KronRLS بهبود بخشید. |
| DeepDTA  [2018] | یک مدل مبتنی بر یادگیری عمیق است که از شبکه های عصبی کانولوشنCNN)) استفاده کرد. | یکی از اولین روش‌های پیش‌بینی DTA است که از دو CNN اعمال شده بر روی رشته های SMILES دارو و توالی‌های آمینو اسیدی پروتئین ها برای پیش بینی DTA استفاده کرد. DeepDTA بهتر از KronRLS و الگوریتم های SimBoost عمل کرده است. |
| WideDTA  [2019] | یک مدل مبتنی بر یادگیری عمیق است از اطلاعات متنی و بیولوژیکی دارو و هدف برای پیش‌بینی DTA استفاده می‌کند. | مدل WideDTA توسعه ای از DeepDTA است. از CNN برای بازنمایی داروها و پروتئین ها استفاده می کند. در WideDTAرشته های SMILES و پروتئین به‌عنوان کلمات نظرگرفته شده اند و به ترتیب با یک دنباله هشت کاراکتری و یک دنباله سه کاراکتری ، نشان داده شدند.  WideDTA چهار قطعه متنی مختلف از اطلاعات مربوط به پروتئین ها و لیگاندها را ترکیب می کند. |
| DeepAffinity  [2019] | یک مدل یادگیری عمیق نیمه نظارت‌شده که ازترکیب RNN و CNN برای پیش‌بینی DTA استفاده می‌کند. | از داده‌های بدون برچسب و برچسب‌گذاری‌شده، برای رمزگذاری مشترک مولکولهای دارو و پروتئین وهمینطور از مکانیسم‌ توجه برای تمرکز برروی بخشهای مهم پروتئین و دارو استفاده  میکند. |
| AttentionDTA  [2019] | مدل مبتنی بر یادگیری عمیق، که از مکانیسم توجه برای پیش‌بینی DTA استفاده می‌کند. | نتایج نشان می‌دهد که مدل مبتنی بر توجه می‌تواند به طور موثر ویژگی‌های پروتئین مربوط به اطلاعات دارویی و ویژگی‌های دارویی مرتبط با اطلاعات پروتئین را برای پیش‌بینی DTA استخراج کند. |
| GANsDTA  [2020] | مدل نیمه نظارت شده جدید مبتنی بر شبکه‌های متخاصم مولد (GANs) برای پیش‌بینی DTA است. | مدل GANsDTA می‌تواند با استفاده از دو GAN بر روی داده‌های برچسب‌دار و بدون برچسب، ویژگی‌های پروتئین‌ها و داروها را استخراج کند. |
| DeepGS  [2020] | مدل مبتنی بر یادگیری عمیق از ترکیب CNN و GATو بلوک های BiGRU استفاده میکند. | از GAT برای استخراج ویژگیهای گراف دارو و ازBiGRU برای دستیابی به اطلاعات شیمیایی آنها استفاده میکند.با تکنیکهای Embedding برای رمزگداری رشته های SMILES و آمینو اسید های پروتئین استفاده میکند. |
| GraphDTA  [2020] | نمایش گراف را برای مولکولها دارو معرفی کرد و از ترکیب انواع GNN و CNN برای پیش بینی DTA استفاده کرده است. | تبدیل رشته های SMILES دارو به گراف می تواند منجر به بهبود عملکرد مدل شود. GraphDTA از چهار نوع شبکه عصبی گراف از جمله GCN، GAT، GIN، و یک معماری ترکیبی GAT-GCN، برای پیش بینی DTA استفاده کرده است.  نتایج نشان می دهد که GraphDTA نسبت به سایر مدلهای مطرح شده عملکرد بهتری در پیش بینی DTAداشته است. |
| DGraphDTA  [2020] | با استفاده از GNNهای مختلف روی گراف دارو و گراف پروتئین کار میکند. | مدلی است که علاوه بر دارو، پروتئین ها را هم به صورت گراف در نظر میگیرد.با استفاده از GNN، DGraphDTA می تواند رشته های پروتئین رابر اساس contact map به گراف تبدیل کند و می‌تواند ویژگی‌های ساختاری پروتئین را با توجه به رشته های آنها پیش‌بینی کند. |
| WGNN-DTA  [2022] | با استفاده از GNNهای مختلف روی گراف دارو و گراف وزندارکه برای پروتئین میسازد DTA را پیش بینی میکند. | یک گراف وزنی برای پروتئین ساخته شده که می تواند به طور جامع اطلاعات ساختار پروتئین را شرح دهد و به جای مقدار باینری که در DGraphDTA استفاده می شود، از احتمال تعامل به عنوان وزن یال های گراف استفاده می شود.از (GCN) و (GAT) برای استخراج بردارهای نهفته گراف های پروتئین و دارو، و پیش‌بینی DTAبر اساس بردارهای نهفته استفاده میکند. |
| S2GC  [2022] | مدلی مبتنی بر یادگیری عمیق که از GNNبرای پیش بینی DTAاستفاده میکند. | رشته های پروتئین و رشته های SMILES دارو به گراف تبدیل میشوند. برخلاف سایرمدل ها که از GCN های 2 و3 لایه استفاده میکنند این مدل از تعداد لایه های خیلی بیشتر حدود 16لایه برای GCN استفاده کرده است. |
| FusionDTA  [2022] | مدل مبتنی بر یادگیری عمیق با استفاده از BiLSTM و multi-head linear attention  DTA را پیش بینی میکند. | از multi-head linear attention  Mechanism برای جمع آوری اطلاعات مهم بر اساس وزن های توجه استفاده کرده است. |
| Multiscale-DTA  [2022] | ازmulti-scaleCNNs ومکانیسم توجه برای پیش بینی DTAاستفاده میکند. | ابتدا رشته های دارو و پروتئین با روشهایEmbedding کد گذاری میشوند . سپس با عبور از دو بلوک CNN سه لایه و مکانیسم توجه ویژگی های چند مقیاسی و جامع مولکول های دارو و هدف استخراج میشوند وبعد از ترکیب با هم وارد لایه های FCمیشوند تا DTA را پیش بینی کند. |
| GSAT-DTA  [2022] | از ترنسفورمر برای پروتئین و BiGRU وGNN برای دارو استفاده کرده است. | Graph–sequence attention  به شبکه اجازه میدهد تا روی مرتبط ترین بخش داده های ورودی تمرکز کند و دقیقترین اطلاعات و ویژگیها را از رشته های SMILES و گراف دارو استخراج کند.  multi-head self-attention در ترنسفومر با استفاده از انکدر و دیکدر ویژگیهای رشته های پروتئین را از موقعیت های مختلف رشته استخراج میکند و در نهایت خروجی DTA تخمین زده میشود. |

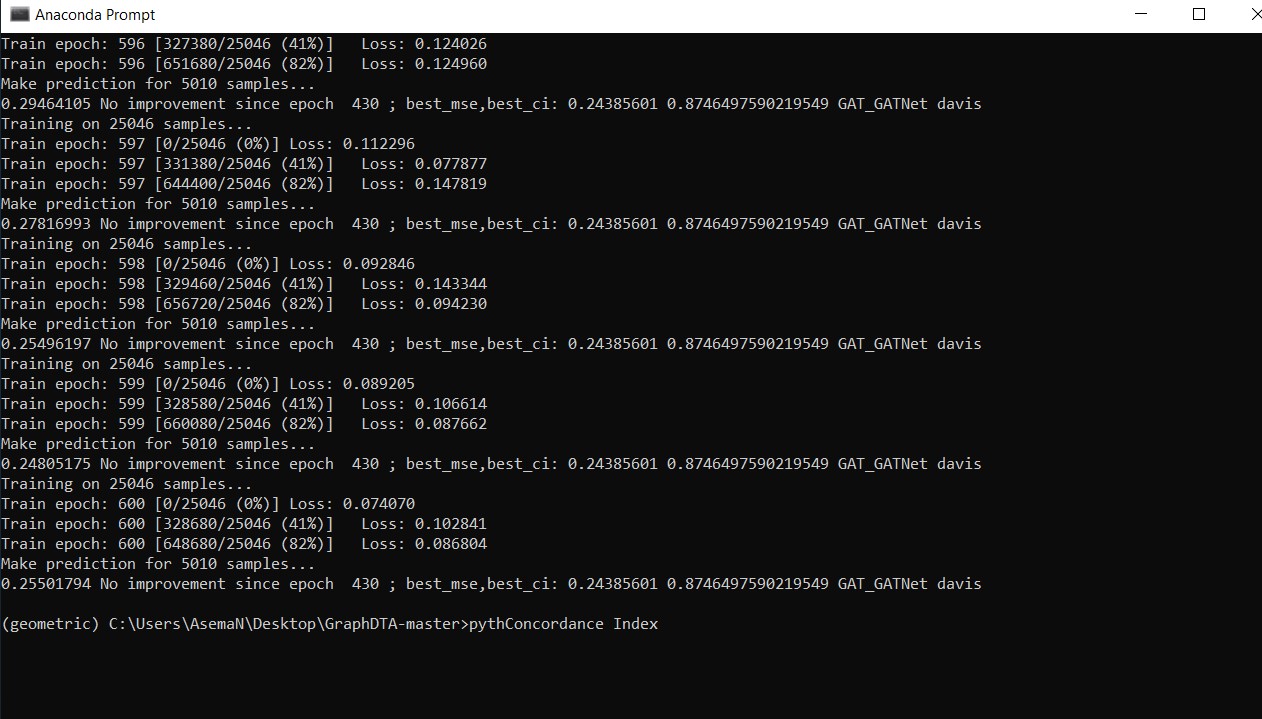
**فصل 4**

**پیاده سازی**

**4-1 پیاده سازی انجام شده**

با اجرای کد مربوط به مدل مطرح شده GraphDTA و مشاهده نتایج مدلهای مختلف GNN که بر روی نمایش گراف دارو استفاده شده است. در تحقیقات قبلی از GCN برای نمایش گراف دارو استفاده شده است و بردار های ویژگی در سطح گره های گراف استخراج شده اند و برای تخمین DTA به کار رفته اند [12]. لازم به ذکر است در GCN پس از آخرین لایه یک لایه ادغام max pooling اضافه می شود. بر خلاف GCNها ، GAT یک معماری مبتنی بر توجه را برای نمایش گره های پنهان در یک گراف با استفاده از مکانیسم توجه پیشنهاد می کند. در این مقاله از یک مدل ترکیبی GAT-GCN استفاده شده که گراف شبکه عصبی با یک لایه GAT شروع می شود وگراف دارو را به عنوان ورودی می گیرد و از لایه GCN بعدی عبور می کند. هر لایه توسط یک تابع ReLU فعال می شود.

معماری GAT-GCN به صورت GAT-GAT تغییر داده شد که در آن GCN با GAT جایگزین شده است. از آموزش مدل تحت این معماری جدید (با 600 epoch) روی مجموعه داده Davis به نتایج CI=0.874 و MSE=0.243 رسیدیم که با نتایج گزارش شده در [12] CI=0.881،MSE=0.245 قابل مقایسه می باشد.



شکل 4-1 گزارش نتایج پیاده سازی انجام شده با مدل GAT-GAT

**فصل5**

**نتیجه گیری**

کاربرد یادگیری عمیق در کشف دارو شامل مدل ها و الگوریتم های پیش بینی است که می تواند منجر به کاهش هزینه و زمان این فرایند شود. استفاده مجدد از دارو، فرآیند شناسایی کاربردهای جدید برای داروهای موجود بر روی بیماری های جدید است. این یک رویکرد مقرون به صرفه برای کشف و توسعه دارو است.

پیش‌بینی ارتباط دارو-هدف (DTA) گام مهمی دراستفاده مجدد از دارو است که می‌تواند به تسریع فرایند توسعه دارو کمک قابل توجهی کند. در سالهای اخیر از شبکه‌ عصبی گراف (GNN) برای پیش‌بینی DTA استفاده شده است. داده های غیر اقلیدسی مانند ساختار مولکولهای دارو و ساختارهای پروتئین را می توان به طور موثر با استفاده از گراف نشان داد. GNN امکان مطالعه تعاملات و تداخلات دارویی و کشف آنتی‌بیوتیک‌های جدید را فراهم می‌کند. شناسایی تداخلات دارو-هدف می تواند فرآیند کشف دارو را سرعت ببخشد و در نتیجه توسعه دارو را تسهیل کند. GNN می تواند فعل و انفعالات پیچیده بین داروها و اهداف را با مدل سازی آنها به عنوان گراف به تصویر بکشد.

طبق مطالعات انجام شده به منظور بهبود دقت پیش بینی DTA ، انتخاب یک مدل GNN مناسب در مرحله اول بسیار مهم است. شبکه‌های کانولوشن گراف (GCN) و شبکه‌ های توجه گراف (GAT) دو مدل GNN محبوب هستند که برای پیش‌بینی ِDTA شده‌اند. دو معماری GCN و GAT و تعداد لایه های مختلفی که برای آنها در نظر گرفته میشود از پرکاربردترین معماری ها هستند که به نتایج خوبی هم دست یافته اند.

طبق مطالعات انجام شده و مقایسه نتایج ارائه شده، استفاده از ترنسفورمرها و مکانیسم توجه در استخراج ویژگی های دارو و پروتئین و بهبود عملکرد مدل در پیش بینی DTA تاثیر زیادی دارد.

**مراجع**

[1] Chen, Wei, et al. "Artificial intelligence for drug discovery: Resources, methods, and applications." *Molecular Therapy-Nucleic Acids* (2023).

[2] Pan, Xiaoqin, et al. "Deep learning for drug repurposing: Methods, databases, and applications." *Wiley interdisciplinary reviews: Computational molecular science* 12.4 (2022): e1597.

[3] Li, Hongjian, et al. "Low-quality structural and interaction data improves binding affinity prediction via random forest." Molecules 20.6 (2015): 10947-10962.

[4] Pahikkala, Tapio, et al. "Toward more realistic drug–target interaction predictions." Briefings in bioinformatics 16.2 (2015): 325-337.

[5] He, Tong, et al. "SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines." Journal of cheminformatics 9.1 (2017): 1-14.

[6] Öztürk, Hakime, Arzucan Özgür, and Elif Ozkirimli. "DeepDTA: deep drug–target binding affinity prediction." Bioinformatics 34.17 (2018): i821-i829.

[7] Öztürk, Hakime, Elif Ozkirimli, and Arzucan Özgür. "WideDTA: prediction of drug-target binding affinity." arXiv preprint arXiv:1902.04166 (2019).

[8]Karimi, Mostafa, et al. "DeepAffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks." Bioinformatics 35.18 (2019): 3329-3338.

[9]Bae, Haelee, and Hojung Nam. "GraphATT-DTA: Attention-Based Novel Representation of Interaction to Predict Drug-Target Binding Affinity." *Biomedicines* 11.1 (2022): 67.

[10]: Zhao, Lingling, et al. "GANsDTA: Predicting drug-target binding affinity using GANs." Frontiers in genetics 10 (2020): 1243.

[11] Lin, Xuan. "DeepGS: Deep representation learning of graphs and sequences for drug-target binding affinity prediction." arXiv preprint arXiv:2003.13902 (2020).

[12] Nguyen, Thin, et al. "GraphDTA: predicting drug–target binding affinity with graph neural networks." Bioinformatics 37.8 (2021): 1140-1147.

[13] Jiang, Mingjian, et al. "Drug–target affinity prediction using graph neural network and contact maps." RSC advances 10.35 (2020): 20701-20712.

[14] Jiang, Mingjian, et al. "Sequence-based drug-target affinity prediction using weighted graph neural networks." BMC genomics 23.1 (2022): 1-17

[15] Ma, Dong, Shuang Li, and Zhihua Chen. "Drug-target binding affinity prediction method based on a deep graph neural network." *Mathematical Biosciences and Engineering* 20.1 (2023): 269-282.

[16] Yuan, Weining, Guanxing Chen, and Calvin Yu-Chian Chen. "FusionDTA: attention-based feature polymerizer and knowledge distillation for drug-target binding affinity prediction." *Briefings in Bioinformatics* 23.1 (2022): bbab506.

[17] Chen, Haoyang, et al. "MultiscaleDTA: A multiscale-based method with a self-attention mechanism for drug-target binding affinity prediction." *Methods* 207 (2022): 103-109.

[18] Yan, Xiangfeng, and Yong Liu. "Graph–sequence attention and transformer for predicting drug–target affinity." *RSC advances* 12.45 (2022): 29525-29534..

1. Drug discovery [↑](#footnote-ref-1)
2. Artificial Intelligence [↑](#footnote-ref-2)
3. Drug-Target Interaction [↑](#footnote-ref-3)
4. Drug-Target binding affinity [↑](#footnote-ref-4)
5. Sequence-based method [↑](#footnote-ref-5)
6. Structure-based method [↑](#footnote-ref-6)
7. Deep Learning [↑](#footnote-ref-7)
8. Machine Learning [↑](#footnote-ref-8)
9. Simplified Molecular Input Line Entry System [↑](#footnote-ref-9)
10. Multi-Layer Perceptron [↑](#footnote-ref-10)
11. Convolutional Neural Network [↑](#footnote-ref-11)
12. Long short-term memory [↑](#footnote-ref-12)
13. Recurrent Neural Network [↑](#footnote-ref-13)
14. Generative Adversarial Network [↑](#footnote-ref-14)
15. Autoencoder [↑](#footnote-ref-15)
16. Feedforward Neural Network [↑](#footnote-ref-16)
17. Graph Neural Network [↑](#footnote-ref-17)
18. Attention-Based Model [↑](#footnote-ref-18)
19. Transformer Neural Network [↑](#footnote-ref-19)
20. Self-Attention [↑](#footnote-ref-20)
21. Random forest [↑](#footnote-ref-21)
22. Convolutional Neural Network [↑](#footnote-ref-22)
23. Semisupervised [↑](#footnote-ref-23)
24. visualization [↑](#footnote-ref-24)
25. semi-supervised [↑](#footnote-ref-25)
26. Generative adversarial networks [↑](#footnote-ref-26)
27. Graph Convolutional Network [↑](#footnote-ref-27)
28. Graph Attention Network [↑](#footnote-ref-28)
29. Deep Neural Network [↑](#footnote-ref-29)
30. Generalizability [↑](#footnote-ref-30)