مهندسی ژنتیک و ایمنی زیستی
ISSN 2588-5073 چاپی
ISSN 2588-5081 الکترونیکی
دوره ۹، شماره ۲، پاییز و زمستان ۹۹
صفحه ۱۹۳-۱۸۱

https://dorl.net/dor/20.1001.1.25885073.1399.9.2.11.6

DOR:20.1001.1.25885073.1399.9.2.11.6

Genetic Engineering and Biosafety Journal
Volume 9, Number 2
2021

http://gebsj.ir/

https://ecc.isc.ac/showJournal/23064

طراحی یک مدل بیوانفورماتیکی برای پیش‌بینی فعالیت ترکیبات دارویی وکاربرد آن بر مهار تکثیر HIV و ژن BACE-1

Design of a bioinformatics model to predict drug compound properties and its application in inhibition of HIV replication and BACE-1

کریم عباسی، علی مسعودی‌نژاد*

Karim Abbasi, Ali Masoudi-Nejad*

آزمایشگاه سیستم بیولوژی و بیوانفورماتیک، مرکز تحقیقات بیوشیمی و بیوفیزیک، دانشگاه تهران، ایران.

Laboratory of system Biology and Bioinformatics (LBB), Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran.

* نویسنده مسئول مکاتبات، پست الکترونیکی: amasoudin@ut.ac.ir

(تاریخ دریافت: ۹۹/۱۱/۵ - تاریخ پذیرش: ۹۹/۱۲/۱۰)

چکیده

در این مقاله روش جدیدی برای مسئله پیش‌بینی خواص ترکیب‌های مولکولی در قدم بهینه‌سازی پیشرو در طراحی دارو ارائه می‌گردد. تعداد داده‌های برچسب شده در دسترس در قدم بهینه‌سازی پیشرو اندک است. در سال‌های اخیر این چالش مورد توجه قرار گرفته است و از تکنیک‌های یادگیری انتقالی و یادگیری عمیق برای حل آن استفاده شده است. بدین منظور از مجموعه‌داده‌های مشابه به عنوان داده‌های کمکی برای آموزش یک مدل قابل اعتماد بهره گرفته شده است. در این روش، استخراج ویژگی از ترکیب‌های مولکولی نقش اساسی در انتقال دانش از مجموعه‌داده‌های مشابه (کمکی) به مجموعه داده‌ی اصلی ایفا می‌کند. در این مقاله تاثیر استفاده از شبکه‌های پیچشی گرافی که علاوه بر در نظر گرفتن ویژگی‌های اتم‌ها، قادر به در نظر گرفتن ویژگی‌های پیوندهای مولکولی می‌باشد، سنجیده می‌گردد. برای ارزیابی روش، از دو مجموعه داده استاندارد BACE و HIV بهره گرفته شده است. نتایج بیانگر این امر است که روش پیشنهادی قادر به استخراج دانش موثرتری از مجموعه داده‌های مشابه برای انتقال به مجموعه‌داده‌ی هدف بوده است.

واژه‌های کلیدی

شبکه پیچیشی گرافی،
پیوندهای مولکولی،
یادگیر عمیق،
یادگیری انتقالی،
طراحی دارو

[DOR: 20.1001.1.25885073.1399.9.2.11.6]

مقدمه

یکی از مهم‌ترین اهداف کشف دارو پیدا کردن داروهای کاندید جدید برای بیماری‌ها است. نشان داده شده است که چرخه توسعه و کشف دارو حدود ۱۲ الی ۱۵ سال زمان نیاز دارد و هزینه آن به طور تقریبی ۲,۶ میلیارد دلار تخمین زده شده است. از این‌رو پروسه‌های کشف دارو بسیار هزینه‌بر و زمان‌بر است. همچنین نرخ بالای شکست چرخه توسعه دارو بر این قضیه دامن می‌زند. به همین دلیل روش‌های طراحی دارو مبتنی بر کامپیوتر (CADD) معرفی گردیدند که از روش‌های محاسباتی برای تسریع قدم‌های اولیه توسعه دارو استفاده می‌نماید (Leelananda and Lindert 2016; Oglic, et al. 2018). یکی از مهم‌ترین قدم‌های CADD، بهینه‌سازی ترکیبات پیشرو است که هدف آن طراحی ترکیباتی است که دارای خصوصیات ADME مناسب و سمیت کم باشند (Waring, et al. 2015; Wenzel, et al. 2019). در این قدم روش‌های محاسباتی بسیار مورد استفاده قرار می‌گیرند که بدین منظور از نمونه‌های آموزشی بهره می‌برند. با توجه به اینکه تعداد داده‌های برچسب شده در دسترس در قدم بهینه‌سازی پیشرو اندک است نیاز به طراحی یک مدل مناسب برای پیش‌بینی خصوصیات ترکیبات با داده‌های کم بسیار ضروری به نظر می‌رسد (Altae-Tran, et al. 2017; Simões, et al. 2018).

روش‌های محاسباتی دارو در سال‌های اخیر بسیار مورد توجه قرار گرفته است (Ezzat, et al. 2018). در این روش‌ها، استخراج ویژگی از داده‌ی ورودی یکی از قدم‌های مهم در پیش‌بینی خواص/فعالیت ترکیب دارویی و یا جفت پروتئین-ترکیب است. هدف از گام استخراج ویژگی تعیین دانش گویا، غیرزائد و متمایزکننده از داده‌ی خام ورودی است که گام‌های بعدی در فرآیند آموزش را تسهیل کند. در روش‌های پیشین، گام‌های استخراج ویژگی به دو دسته‌ی مهم تقسیم می‌گردد: روش‌های استخراج محور داده و غیر محور داده. تفاوت اصلی در میان این دو دسته این اصل است که در رویکرد محور داده، ویژگی‌ها به صورت خودکار از هر داده‌ی ورودی استخراج می‌شود ولی در رویکرد غیرمحور داده ویژگی‌ها برای هر ورودی از یک رویه‌ی ثابت استفاده می‌نماید. یادگیری عمیق کلاسی از الگوریتم‌های یادگیری ماشین است که فضای جدیدی از ویژگی‌ها را به صورت سلسله مراتبی آموزش می‌بینند که از سیستم بینایی انسان الهام گرفته شده است (Goodfellow, et al. 2016). در سال‌های اخیر یادگیری عمیق در حوزه‌های تحقیقاتی زیادی همچون ژنومیکس (Zou, et al. 2019)، رده‌بندی RNAهای کدکننده (Amin, et al. 2019; Asgari, et al. 2019)، پیش‌بینی ساختار دوم پروتئین (Asgari, et al. 2019)، طراحی داروهای نوین (Popova, et al. 2018)، بیوانفورماتیک (Min, et al. 2017; Masoudi-Sobhanzadeh, et al. 2019; Hooshmand, et al. 2020)، بینایی ماشین (Voulodimos, et al. 2018)، پردازش زبان طبیعی (Young, et al. 2018)، و ترجمه‌ی ماشینی (McCann, et al. 2017) مورد استفاده قرار گرفته است. یادگیری عمیق البته در پیش‌بینی فعالیت دارو و یا جفت پروتئین-ترکیب نیز مورد استفاده قرار گرفته است. یادگیری عمیق قادر است که یک نمایش سلسله مراتبی از ویژگی‌ها را به صورت خودکار آموزش ببینند که این قدرت باعث افزایش کارایی عملکرد آن در حوزه‌های متفاوت شده است. در سال ۲۰۱۲ شبکه‌های عمیق چندوظیفه‌ای برنده رقابتی شده‌اند که Kaggle برای پیش‌بینی ویژگی‌های مولکول طراحی کرده بود (Dahl 2012). رامسندار و همکارانش (Ramsundar, et al. 2015) یک چارچوب فراهم کردند که در آن قدرت پیش‌بینی شبکه‌های چندوظیفه‌ای در مقابل شبکه‌های تک‌وظیفه‌ای مورد ارزیابی قرار بگیرد. آنها یک مجموعه داده بسیار عظیم فراهم کردند که به طور تقریبی شامل ۴۰ میلیون داده بر روی ۲۰۰ پروتئین هدف زیستی است. نتایج بدست آمده نشان‌گر آن است که شبکه‌های چندوظیفه‌ای قادر به افزایش دقت چشم‌گیری نسبت به روش‌های تک‌وظیفه‌ای می‌باشند. در سال‌های اخیر یک معماری عمیق برای استخراج ویژگی‌ها از ساختارهای مولکولی ارائه شده است که شبکه‌های پیچشی گرافی (GCN) نامیده می‌شود (Duvenaud, et al. 2015; Kearnes, et al. 2016). این شبکه‌ها یک بردار ویژگی برای هر مولکول بصورت خودکار آموزش می‌بینند. GCN اطلاعات هر اتم و زیرساختار همسایگی آنرا با یکدیگر ترکیب می‌نماید تا بتواند زیرساخت‌های موثر در پیش‌بینی ویژگی‌های مولکول آموزش ببیند. روش GCN برای آموزش مناسب نیازمند تعداد زیادی نمونه آموزشی می‌باشد در حالیکه همانگونه که اشاره شد در قدم بهینه‌سازی ترکیبات پیشرو داده‌های آموزشی کمی موجود است. به همین دلیل آلتاترن و همکارانش (Altae-Tran, et al. 2017) روشی معرفی کردند که در آن از چارچوب یادگیری one-shot برای آموزش یک مدل قوی جهت استخراج ویژگی‌های مولکول در حضور داده‌های کم استفاده کرده‌اند. آن‌ها بیان کردند که در صورتی نمونه‌های آموزشی و آزمایشی از

[DOR: 20.1001.1.25885073.1399.9.2.11.6]

مجموعه داده‌های متفاوتی باشند آنگاه انتقال دانش به درستی انجام نمی‌گیرد و در اکثر موارد حتی بدتر از روش‌های استاندارد موجود در حوزه عمل می‌نماید. این بدین معنی است که روش پیشنهادی ارائه شده توسط آن‌ها قادر به عمومیت‌دهی دانش میان وظایف متفاوت نبوده است. عباسی و همکاران (Abbasi, et al. 2019) سوال تحقیقی "چگونه مدل‌های قابل اعتمادی برای پیش‌بینی ویژگی‌های مولکول‌های ترکیب دارویی با استفاده از داده‌های کمکی مشابه (مرجع) آموزش ببینیم؟" را مورد بررسی قرار دادند. آن‌ها برای پاسخ به این سوال فرض کردند که قطعه‌های پروتئینی مشابهی میان مجموعه داده‌های متفاوتی وجود دارد. بدین منظور از یک شبکه چند وظیفه‌ای برای یادگیری بردارهای ویژگی ترکیبات دارویی مجموعه داده‌ی مرجع (مجموعه داده‌های مشابه کمکی) استفاده کردند. این شبکه ازلایه‌های پیچشی گرافی برای استخراج ویژگی‌ها و از لایه‌های تماما متصل برای درنظر گرفتن اطلاعات برچسب استفاده می‌نماید. در قدم بعد نیز شبکه‌ای شامل لایه‌های پیچشی گرافی و تماما متصل برای استخراج ویژگی‌های داده‌های برچسب‌دار هدف استفاده می‌شود. در قدم سوم، بردارهای ویژگی مرجع و هدف به فضای جدیدی نگاشت می‌گردد به گونه‌ای که آن‌ها دارای توزیع حاشیه‌ای یکسانی باشند و قدرت تمایزپذیری داده‌های مرجع و برچسب‌دار هدف نیز حفظ گردد. در نهایت برای یادگیری بهتر فضای ویژگی برای داده‌های هدف غیربرچسب‌دار، تابع هزینه جدیدی به نام تابع هزینه معنایی نیز معرفی گردید تا بتواند شبکه بهتری آموزش ببیند. با توجه به اینکه روش عباسی و همکاران (Abbasi, et al. 2019) توانسته است در مقایسه با دیگر روش‌های موجود در حوزه موفق عمل نماید، این روش به عنوان معماری پایه در این مقاله انتخاب شده است.

سوال تحقیقی که در این مقاله مورد بررسی قرار می‌گیرد این است که اطلاعات پیوندها در ساختار مولکولی ترکیب دارویی چگونه می‌تواند در انتقال دانش میان مجموعه داده‌ها در مسئله پیش‌بینی خواص مولکولی تاثیرگذار باشد. عباسی و همکاران (Abbasi, et al. 2019) برای استخراج ویژگی از شبکه‌های کانولوشن گرافی بهره گرفته‌اند. اما، این شبکه‌ها تنها از بردار ویژگی اتم‌ها و اطلاعات همسایگی آن‌ها برای یادگیری بردار ویژگی استفاده نموده‌اند و بسیاری از اطلاعات مربوط به یال‌ها (پیوندهای مولکولی) نادیده در نظر گرفته می‌شود. در سال‌های اخیر، روش‌هایی مطرح گردیده است که سعی داشته‌اند اطلاعات یال‌های گراف را نیز در یادگیری اتوماتیک سلسله مراتبی ویژگی‌ها دخالت دهند (Velickovic, et al. 2018; Gong and Cheng 2019). در روش (Gong and Cheng 2019)، برای هر یال یک بردار ویژگی استخراج می‌گردد و سپس این اطلاعات در یادگیری بردارهای ویژگی برای ساختار گرافی ورودی تاثیر داده می‌شود. در این مقاله، برای استخراج بردار ویژگی برای هر ترکیب، از این روش بهره گرفته خواهد شد. برای ارزیابی روش پیشنهادی، این روش بر روی مجموعه‌داده‌های BACE و HIV اعمال گردیده است و نتایج به دست آمده با روش‌های موفق پیشین مورد مقایسه قرار گرفته است که بیانگر این است که اطلاعات پیوندها در ساختارهای مولکولی ترکیب‌های دارویی می‌تواند در یادگیری دانش قابل انتقال بسیار تاثیرگذار باشد.

در ادامه، در ابتدا مروری بر مهمترین کارهای ارائه شده در این حوزه انجام می‌شود. سپس فرموله بندی مسئله معرفی گردیده و جزئیات روش پیشنهادی مطرح می‌گردد. سپس آزمایشات طراحی شده ارائه گشته و نتایج آن‌ها تحلیل و بررسی می‌گردد و در انتها نتیجه‌گیری بیان می‌شود.

مروری بر کارهای پیشین

در این بخش، روش‌های پیشین در حوزه‌ی پیش‌بینی خواص ترکیبات دارویی و پیش‌بینی تعامل پروتئین-ترکیب (CPI) مورد بررسی و تحلیل قرار می‌گیرد. باقریان و همکارانش (Bagherian, et al. 2020) روش‌های محاسباتی موجورد برای پیش‌بینی تعامل پروتئین-ترکیب به شش دسته تقسیم می‌گردند که در ادامه هر کدام از این دسته ها به صورت خلاصه مورد بررسی قرار می‌گیرد.
۱) روش‌های مبتنی بر شباهت: در این روش اطلاعات شباهت میان داروها و شباهت میان پروتئین در پیش‌بینی دخالت داده می‌شود. بدین منظور معیارهای مشابهت برای محاسبه معیار شباهت بسیار مورد اهمیت واقع خواهد شد. یکی از مهمترین چالش‌های الگوریتم‌های موجود در این دسته، تعداد داده‌های با مقدار تعامل شناخته شده برای جفت‌های پروتئین-دارو می‌باشد.
۲) روش‌های مبتنی بر ویژگی: در این دسته از روش‌ها، بردارهای ویژگی برای دارو و پروتئین با استفاده از الگوریتم‌های استخراج ویژگی غیر محور داده به دست آمده و سپس به الگوریتم‌های رده‌بندی (و یا رگرسیون) همانند ماشین بردار پشتیبان، Random Forest و یا رده‌بندی‌های مبتنی بر کرنل فرستاده می‌شوند. در رویکرد غیرمحور داده ویژگی‌ها برای هر ورودی از یک رویه‌ی ثابت استفاده می‌نماید. ۳) روش‌های مبتنی بر تجزیه

[DOR: 20.1001.1.25885073.1399.9.2.11.6]

ماتریس ۴) روش‌های مبتنی بر یادگیری عمیق: یادگیری عمیق کلاسی از الگوریتم‌های یادگیری ماشین است که فضای جدیدی از ویژگی‌ها را به صورت سلسله مراتبی آموزش می‌بینند که از سیستم بینایی انسان الهام گرفته شده است (Goodfellow, et al. 2016). در این رویکرد (شناخته شده به عنوان رویکرد محور داده)، ویژگی‌ها به صورت خودکار از هر داده‌ی ورودی استخراج می‌شود. و ۵) روش‌های ترکیبی. در سال‌های اخیر یادگیری عمیق در حوزه‌های تحقیقاتی زیادی همچون ژنومیکس (Zou, et al. 2019)، رده‌بندی RNAهای کدکننده (Amin, et al. 2019; Asgari, et al. 2019)، پیش‌بینی ساختار دوم پروتئین (Asgari, et al. 2019)، طراحی داروهای نوین (Popova, et al. 2018)، بیوانفورماتیک (Min, et al. 2017; Masoudi-Sobhanzadeh, et al. 2019; Hooshmand, et al. 2020)، بینایی ماشین (Voulodimos, et al. 2018)، پردازش زبان طبیعی (Young, et al. 2018)، و ترجمه‌ی ماشینی (McCann, et al. 2017) مورد استفاده قرار گرفته است. یادگیری عمیق البته در پیش‌بینی فعالیت دارو و یا جفت پروتئین-ترکیب نیز مورد استفاده قرار گرفته است که در این بخش این روش‌ها مورد مطالعه و بررسی قرار می‌گیرند.

پاهیکالا و همکارانش (Pahikkala, et al. 2014) روش‌های موجود در پیش‌بینی فعالیت جفت پروتئین-ترکیب را مورد بررسی قرار داده و چهار فاکتور مهم که در پیش‌بینی نتایج بسیار تاثیرگذار بوده‌اند را استخراج کرده‌اند. این فاکتورها شامل موارد زیر است: ۱) فرموله‌بندی مساله: پیش‌بینی می‌تواند به عنوان یک وظیفه رده‌بندی در نظر گرفته شود که برچسب فعال و یا غیرفعال را به هر ترکیب و یا جفت پروتئین-ترکیب اختصاص می‌دهد و یا به عنوان یک مساله رگرسیون در نظر گرفته می‌شود که مقدار عددی فعالیت را پیش‌بینی نماید. در بسیاری از موارد، مساله به صورت رگرسیون در نظر گرفته و یک مقدار عددی پیوسته را پیش‌بینی می‌نماید و سپس این عدد با استفاده از آستانه‌گذاری به برچسب‌های رده‌بندی تبدیل می‌گردد. ۲) ارزیابی مجموعه داده‌ها: کدام مجموعه داده برای ارزیابی سیستم مورد استفاده قرار گرفته است. به طور مثال، تنوع ترکیب‌های دارویی و یا تنوع دنباله‌های پروتئینی موجود در مجموعه داده بسیار حائز اهمیت است. ۳) رویه‌ی ارزیابی: در مرحله‌ی ارزیابی از چه تنظیماتی برای محاسبه‌ی معیار کارایی استفاده شده است. به طور مثال، از اعتبارسنجی متقاطع ساده و یا اعتبارسنجی تودرتو برای محاسبه‌ی کارایی بهره گرفته شده است. ۴) تنظیمات آزمایشات: تفاوت توزیع حاشیه‌ای داده‌های آموزشی و آزمایشی مورد بررسی قرار می‌گیرد که این حالت خود دارای چهار تنظیم متفاوت است: الف) تنظیمات warm: در این حالت پروتئین و ترکیب‌های مشاهده شده در مجموعه آزمایشی، در مجموعه آموزشی نیز حتما مشاهده شده‌اند. ب) تنظیمات پروتئین cold:در این حالت، ممکن است در مجموعه آزمایشی دنباله پروتئینی مشاهده گردد که در مجموعه آموزشی مشاهده نشده است. ج) تنظیمات ترکیب cold: در این حالت، ممکن است در مجموعه آزمایشی دنباله ترکیب کاندید دارویی مشاهده گردد که در مجموعه آموزشی مشاهده نشده است. د) در این حالت، ممکن است در مجموعه آزمایشی هر دو دنباله پروتئینی و ترکیبی مشاهده گردد که در مجموعه آموزشی مشاهده نشده است که این مورد از چالش‌برانگیزترین حالت‌ها است. عباسی و همکاران (Abbasi, et al. 2019) علاوه بر موارد ذکر شده، موارد ضروری دیگری نیز معرفی نمودند که در طراحی مدل‌های مبتنی بر یادگیری ماشین (بالاخص یادگیری عمیق) موثر است. موارد مذکور عبارت است از: ۱) فضای ویژگی ورودی شبکه ۲) شبکه استخراج ویژگی برای دنباله ترکیب و دنباله پروتئین و ۳) ترکیب توصیف‌گر ویژگی دنباله ترکیب و پروتئین.

سوباکی و همکاران (Tsubaki, et al. 2018) دنباله‌ی پروتئینی را به عنوان ورودی گرفته و به شبکه‌ی CNN برای آموزش ویژگی‌ها می‌فرستد. اوزترک و همکاران (Öztürk, et al. 2018) از شبکه‌های CNN برای آموزش بردار ویژگی برای هر دو دنباله‌ی پروتئینی و دنباله‌ی ترکیب استفاده نموده‌اند. بایستی توجه گردد که شبکه‌های CNN قدرت تفسیرپذیری بالایی دارند زیرا با رسم فیلترهای آموزش دیده و نقشه‌های ویژگی حاصل از هر لایه اطلاعات زیادی استخراج می‌گردد. شبکه‌های پیچشی گرافی به عنوان روشی برای یادگیری یک اثرانگشت برای مولکول‌ها معرفی گردیده است (Duvenaud, et al. 2015; Kearnes, et al. 2016). در GCN، ساختار گرافی مولکول به عنوان ورودی به شبکه فرستاده می‌شود که در این ساختار گرافی هر اتم با استفاده از یک راس و هر اتصال با استفاده از یک یال نمایش داده می‌شود. مهم‌ترین لایه‌ی شبکه GCN، لایه‌ی پیچشی گرافی است که نخستین بار در (Bruna, et al. 2013) معرفی گردیده است. به طور متداول، شبکه‌ی GCN شامل چندین لایه‌ی متوالی پیچشی گرافی است که زیرساختارهای محلی موثر در مولکول را به صورت سلسله مراتبی استخراج می‌نماید. در لایه پیچشی گرافی در ابتدا بردار ویژگی هر راس (اتم) به فضای جدیدی نگاشت می‌گردد و سپس با استفاده از بردارهای

[DOR: 20.1001.1.25885073.1399.9.2.11.6]

ویژگی راس‌های همسایه، بردار ویژگی نهایی راس به‌روزرسانی می‌گردد. سوباکی و همکارانش (Tsubaki, et al. 2018) از شبکه GCN برای استخراج ویژگی ترکیب دارویی در پیش‌بینی فعالیت جفت پروتئین-ترکیب استفاده کرده‌اند. آلتاتران و همکارانش (Altae-Tran, et al. 2017) یک چارچوب جدیدی برای کشف دارو با استفاده از تکنیک یادگیری one-shot در حضور تعداد داده‌های کم ارائه داده‌اند. بدین منظور آن‌ها از شبکه‌های GCN با تکنیک بهبود تکراری LSTM برای استخراج ویژگی‌های معنادار بهره گرفته‌اند. گااو و همکارانش (Gao, et al. 2018) نیز از شبکه GCN برای استخراج ویژگی ترکیب دارویی بهره گرفته‌اند. وو و همکارانش (Wu, et al. 2018) مدلی به نام MoleculeNet ارائه داده‌اند که در آن الگوریتم‌های استخراج ویژگی مولکولی متفاوت و الگوریتم‌های یادگیری ماشین بر روی مجموعه وسیعی از مجموعه داده‌های در دسترس اعمال گردیده‌اند و نتایج آن‌ها با یکدیگر مقایسه گشته است. آن‌ها نتیجه گرفتند که روش‌های مبتنی بر گراف (همانند شبکه GCN) در مقایسه با روش‌های دیگر بهتر عمل نموده و قادر به استخراج ویژگی‌های معنادارتری هستند. البته شبکه‌ی GCN در مقایسه با شبکه‌ی CNN تفسیرپذیری کمتری دارد. زپوپه و همکارانش (Pope, et al. 2019) شبکه GCN را به گونه‌ای اصلاح کردند که با استخراج گروه‌های عملکردی موثر قدرت تفسیرپذیری بالاتری داشته باشند. شبکه‌های عصبی بازگشتی (RNN) یک کلاسی از شبکه‌های عصبی است که یک دنباله‌ی زمانی را به عنوان ورودی می‌گیرد. در طی مرحله‌ی آموزش، این شبکه اطلاعات دنباله را در گام‌های زمانی به خاطر می‌سپارد تا بتواند در تصمیم‌گیری استفاده نماید. در (Chakravarti and Alla 2019)، ابتدا الگوریتم استخراج ویژگی MLNCT بر روی ترکیب اعمال گشته و سپس به شبکه دو جهته LSTM فرستاده می‌شود. فوشه و همکارانش (Fooshee, et al. 2018) نمایش SMILES مولکول را به عنوان ورودی به شبکه‌ی LSTM می‌فرستند تا جریان الکترونی را پیش‌بینی کنند. ون و همکارانش (Wen, et al. 2017) در ابتدا با استفاده از توصیف‌گر ECFP و توصیف‌گر PSC به ترتیب دنباله‌های ترکیب و پروتئین را توصیف کرده و سپس این توصیف‌گرها با یکدیگر الحاق شده و به شبکه باور عمیق فرستاده می‌گردند. شبکه‌های باور عمیق یک کلاسی از شبکه‌های عصبی هستند که توسط انباشته کردن ماشین بولتزمن محدود شده (RBM) به وجود آمده‌اند. همچنین استراتژی‌هایی برای ترکیب توصیف‌گرهای پروتئین و ترکیب ارائه شده است. به صورت متداول، این توصیف‌گرها به آسانی به یکدیگر الحاق شده تا بتوانند بردار توصیف‌گر نهایی را تشکیل دهند. در (Pham and Le 2019)، از عملگر الحاق برای ترکیب بردارهای ویژگی پروتئین و ترکیب استفاده شده است. اگرچه، در سال‌های اخیر، از مکانیسم توجه به عنوان روشی برای ترکیب توصیف‌گرهای ترکیب و دنباله پروتئینی استفاده گردیده است. مکانیسم توجه برای نخستین بار در (Bahdanau, et al. 2015) معرفی گشته است. سوباکی و همکاران (Tsubaki, et al. 2018) از مکانیسم توجه برای محاسبه‌ی قدرت اتصال یک ترکیب دارو با زیرساختارهای پروتئینی استفاده کرده است. سپس مجموع وزن‌دار ویژگی‌های زیرساختارهای پروتئینی با ضرایب توجه محاسبه شده و سپس این بردار به بردار ویژگی ترکیب الحاق شده و بردار توصیف‌گر نهایی را تشکیل می‌دهد. در نهایت بردار ویژگی حاصل به لایه‌های پیش‌بینی فرستاده می‌شود. حسن و همکاران (Hassan, et al. 2018) از الگوریتم استخراج ویژگی BINANA برای محاسبه‌ی توصیف‌گر جفت پروتئین-ترکیب استفاده کرده‌اند و سپس این توصیف‌گر به لایه‌های پیش‌بینی فرستاده می‌شود. بایستی دقت گردد که توصیف‌گر BINANA جفت پروتئین-ترکیب را به صورت همزمان توصیف می‌نماید.

عباسی و همکاران در (Abbasi, et al. 2020) روشی برای یادگیری یک مدل قوی برای پیش‌بینی تعامل جفت پروتئین-ترکیب غیر برچسب‌دار هدف ارائه دادند. بدین منظور آن‌ها برای آموزش مدل قوی‌تر از مجموعه داده‌های مرجع مشابه بهره گرفتند. در قدم اول، یک شبکه استخراج ویژگی برای مجموعه داده‌ی مرجع آموزش می‌بینند. با توجه به اینکه در این حالت، شبکه دارای دو ورودی دنباله پروتئین و دنباله ترکیب است، شبکه استخراج ویژگی دارای دو زیرشبکه موازی با یکدیگر خواهد بود که در روش پیشنهادی اطلاعات هر دو دنباله با استفاده از تکنیک مکانیسم توجه دو طرفه پیشنهادی ترکیب می‌گردد. سپس در قدم دوم، شبکه استخراج ویژگی برای مجموعه داده‌ی هدف با استفاده از تکنیک تطبیق دامنه آموزش می‌بینند. بدین منظور از روش تطبیق دامنه رقابتی استفاده شده است. برای جلوگیری از انتقال دانش منفی، تکنیک تطبیق دامنه به گونه‌ای اصلاح شده است که جفت پروتئین-ترکیب‌هایی از مجموعه داده‌ی مرجع که شباهت بیشتری به داده هدف دارند تاثیر بیشتری در پیش‌بینی داشته باشند. در قدم سوم، جفت پروتئین-ترکیب‌های هدف غیربرچسب‌دار به شبکه استخراج ویژگی هدف فرستاده می‌گردند و خروجی بردار ویژگی به شبکه پیش‌بینی داده‌های مرجع فرستاده می‌شود تا برچسب نمونه هدف پیش‌بینی گردد. چون در گام دوم، از تکنیک تطبیق دامنه استفاده

شده است در نتیجه توزیع بردارهای ویژگی داده‌های هدف و مرجع به یکدیگر نزدیک است و می‌توان به راحتی از شبکه پیش‌بینی داده‌های مرجع استفاده کرد.

مواد و روش‌ها

روش پیشنهادی

در این بخش، روش پیشنهادی با جزئیات مورد بررسی قرار می‌گیرد. بدین منظور در ابتدا فرموله‌بندی مسئله ذکر می‌گردد و سپس معماری کلی روش پیشنهادی مورد بررسی قرار می‌گیرد.

فرموله‌بندی مساله

در این بخش در ابتدا صورت دقیق مساله بیان می‌گردد. فرض کنید چندین وظیفه مرجع و یک وظیفه هدف در اختیار دارید. مجموعه داده مرجع شامل $|S|$ وظیفه است. داده‌های موجود در kامین وظیفه مرجع با استفاده از $S^{(k)} = \{x_i^{(s,k)}, y_i^{(s,k)}\}_{i=1}^{m^k}$ نمایش داده می‌شود که $x_i^{(s,k)} \in \chi^S$ بیانگر iامین نمونه از kامین وظیفه مرجع است و $y_i^{(s,k)} \in \mathcal{Y}^{S,k}$ بیانگر برچسب متناظر $x_i^{(s,k)}$ است. داده‌ای $x_i^{(s,k)}$ یک ترکیب دارویی است که در آزمون تجربی مورد بررسی قرار گرفته و برچسب متناظر آن $y_i^{(s,k)}$ نتیجه آزمایش متناظر است. همچنین مجموعه داده هدف شامل دو مجموعه مستقل است: مجموعه داده هدف برچسب‌دار و مجموعه داده هدف غیربرچسب‌دار که به ترتیب با استفاده از $T^{\ell} = \{x_i^{(t,\ell)}, y_i^{(t,\ell)}\}_{i=1}^{n^{\ell}}$ و $T^{u} = \{x_i^{(t,u)}\}_{i=1}^{n^{u}}$ نمایش داده می‌شود. پارامترهای n^{ℓ} و n^{u} به ترتیب نشانگر تعداد داده‌های برچسب‌دار و غیربرچسب‌دار هدف می‌باشند. همچنین، iامین نمونه هدف برچسب‌دار و غیربرچسب‌دار به ترتیب با استفاده از $x_i^{(t,\ell)} \in \chi^T$ و $x_i^{(t,u)} \in \chi^T$ نمایش داده می‌شود. بایستی دقت گردد که تعداد نمونه‌های هدف برچسب‌دار بسیار کم است ($n^{u} \gg n^{\ell}$) . شمای کلی روش در شکل ۱ نشان داده شده است. روش از یک شبکه چند وظیفه‌ای برای یادگیری بردارهای ویژگی ترکیبات دارویی مجموعه داده‌ی مرجع استفاده می‌نماید (شکل ۱-الف). در این قدم، تابع هزینه نظارتی برای آموزش شبکه مورد استفاده قرار می‌گیرد. این شبکه ازلایه‌های پیچشی گرافی که قادر است اطلاعات مربوط به پیوندهای مولکولی را دخیل نماید برای استخراج ویژگی‌ها و از لایه‌های تماما متصل برای درنظر گرفتن اطلاعات برچسب استفاده می‌شود.

در قدم بعد نیز شبکه‌ای شامل لایه‌های پیچشی گرافی و تماما متصل برای استخراج ویژگی‌های داده‌های برچسب‌دار هدف استفاده می‌شود. بردارهای ویژگی مرجع و هدف به فضای جدیدی نگاشت می‌گردد به گونه‌ای که آن‌ها دارای توزیع حاشیه‌ای یکسانی باشند و قدرت تمایزپذیری داده‌های مرجع و برچسب‌دار هدف نیز حفظ گردد. در نهایت برای یادگیری بهتر فضای ویژگی برای داده‌های هدف غیربرچسب‌دار تابع هزینه معنایی مورد استفاده قرار می‌گیرد تا بتواند شبکه بهتری آموزش ببیند.

در ادامه شبکه پیچشی گرافی که قادر است اطلاعات یال‌های گراف ورودی را علاوه بر توصیف‌گرهای نود و اطلاعات همسایگی دخیل نماید توضیح داده می‌شود. این شبکه به اختصار شبکه پیچشی گرافی وزن‌دار شده (EGCN) (Gong and Cheng 2019) نامیده می‌شود.

شبکه پیچش گرافی وزن‌دار شده

فرض کنید که یک مولکول ترکیب با N اتم در اختیار داریم. ماتریس X به ابعاد $N \times F$ بیانگر توصیف‌گر اتم‌های موجود در گراف مولکول می‌باشد. بدین منظور X_{ij} بیانگر مقدار ویژگی jام از اتم iام است و $X_i \in \mathbb{R}^F$ بیانگر بردار ویژگی F بعدی برای iامین اتم است. همچنین، ماتریس E به ابعاد $N \times N \times P$ بیانگر بردار ویژگی متناظر با هر یال (پیوند در ساختار مولکولی) است. با توجه اینکه ممکن است که میان بسیاری از جفت اتم‌ها هیچ پیوندی وجود نداشته باشد، بردار ویژگی متناظر با آن بردار صفر در نظر گرفته می‌شود. یکی از مهمترین ویژگی‌ها توصیفگر EGCN که در این مقاله مورد استفاده قرار گرفته است این مطلب است که برای هر یال، یک بردار ویژگی در نظر گرفته شده است. در حالیکه در اکثر روش‌های پیشین، از یک ویژگی دودویی به عنوان توصیف‌گر استفاده شده بود. در این مقاله از ویژگی‌های همانند رتبه پیوند، نوع جفت اتم، حضور اتم در حلقه، نوع حلقه به عنوان ویژگی‌های اولیه یال‌ها (پیوندهای مولکولی) استفاده شده است. برای تشکیل ماتریس ویژگی اتم‌ها برای هر ترکیب دارویی، بسته‌ی RDKit (Landrum 2006) برای استخراج ویژگی‌های اتم مورد استفاده قرار می‌گیرد. ویژگی‌های استخراج شده شامل نوع اتم، درجه آزادی، implicit valences، formal charges، hybridization of spins، تعداد الکترون‌های رادیکال و تعداد پیوندهای همسایه در گراف می‌باشد.

شبکه‌ی EGCN یک شبکه‌ی رو به جلوی چند لایه است. بدین منظور از بالانویس l برای اشاره به شماره‌ی لایه‌ها استفاده خواهد شد. در الگوریتم EGCN، برای اجتناب از مقیاس ویژگی‌های

[DOR: 20.1001.1.25885073.1399.9.2.11.6]

متفاوت لبه، در ابتدا این ویژگی‌ها نرمال می‌گردند. بدین منظور از روش زیر استفاده خواهد شد:

$$\tilde{E}_{ijp} = \frac{\hat{E}_{ijp}}{\sum_{k=1}^{N} \hat{E}_{ikp}} \quad (۱)$$

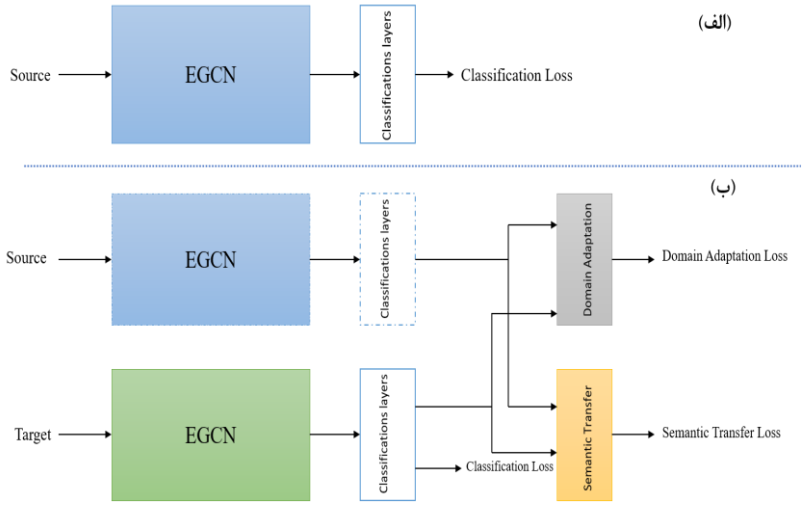
$$E_{ijp} = \sum_{k=1}^{N} \frac{\tilde{E}_{ikp}\tilde{E}_{jkp}}{\sum_{v=1}^{N} \hat{E}_{vkp}} \quad (۲)$$

در رابطه‌ی (۱)، نماد $\hat{E}$ بیانگر ماتریس ویژگی پیوندهای ترکیب مولکولی می‌باشد. پس از اعمال روابط (۱) و (۲) تضمین می‌گردد که هر درایه ماتریس نرمالیزه شد مثبت بوده و جمع درایه‌های بردارهای ویژگی در طی سطر و یا ستون یک می‌باشد.

در لایه‌ی پیچشی گرافی، ماتریس‌های X^{l-1} و E به عنوان ورودی به لایه lام شبکه فرستاده می‌شود. ماتریس ویژگی مربوط به اتم‌ها به صورت زیر به روزرسانی می‌گردد:

$$X^{l} = \sigma\left[\|_{p=1}^{P}\left(E_{.p}X^{l-1}W^{l}\right)\right] \quad (۳)$$

که σ بیانگر تابع فعال‌سازی غیرخطی است، نماد || بیانگر عملگر الحاق است. ماتریس وزن W^{l} به ابعاد $F^{l-1} \times F^{l}$ می‌باشد که مجهول لایه lام می‌باشد و در طی مرحله آموزش به دست می‌آید. ماتریس $E_{.p}$ بیانگر یک ماتریس $N \times N$ است که برای هر همسایگی pامین ویژگی را در اختیار قرار می‌دهد. بایستی دقت گردد که در رابطه‌ی (۳) اطلاعات اتم‌هایی که در همسایگی هم قرار دارند برای به روزرسانی بردارهای ویژگی اتم‌ها مورد استفاده قرار می‌گیرد. در (Gong and Cheng 2019)، علاوه بر رابطه‌ی (۳)، رابطه‌ی دیگری نیز برای به روز رسانی ارائه داده‌اند که در این حالت ماتریس ویژگی پیوندها نیز در هر لایه به روزرسانی می‌گردد. بدین منظور آنها از مکانیسم توجه (attention mechanism) استفاده نموده‌اند.



شکل ۱-شمای کلی روش پیشنهادی- در این روش از شبکه EGCN که قادر به دخالت اطلاعات پیوندهای مولکولی است برای استخراج ویژگی استفاده شده است.

Figure 1- The overall schematic of the proposed approach. the network EGCN is utilized to incorporate the molecular bond knowledge into the feature extraction step.

نتایج و بحث

در این بخش، آزمایشاتی که برای ارزیابی روش پیشنهادی طراحی گردیده است، بیان گردیده و نتایج آن مورد بررسی و تحلیل قرار می‌گیرد. برای ارزیابی روش پیشنهادی، دو مجموعه داده استاندارد BACE و HIV انتخاب گردیده است. مجموعه داده‌ی HIV شامل یک آزمون می‌باشد که قدرت ۴۱۱۲۷ ترکیب دارویی را برای مهار تکثیر HIV آزمایش می‌کند. این مجموعه توسط سازمان برنامه درمانی دارویی (DTP) تهیه شده و برچسب ترکیب‌های دارویی موجود در آن به سه دسته غیرفعال، فعال و نسبتاً فعال تقسیم می‌شود. مجموعه داده‌ی BACE شامل یک آزمون با ۱۵۲۲ ترکیب دارویی برای مهار ژن(BACE-1) β-secretase-1 می‌باشد.

با توجه به اینکه، یکی از مهم‌ترین چالش‌های مورد بحث در این مقاله، تعداد کم داده‌های آموزشی برچسب خورده در دسترس است، دو دسته تنظیمات بر روی تعداد داده‌های برچسب‌دار در نظر گرفته شده است: ۱) دسته اول: شامل ۱۰ مورد ترکیب مثبت و ۱۰

[DOR: 20.1001.1.25885073.1399.9.2.11.6]

مورد ترکیب منفی، ۲) دسته دوم: شامل ۱ مورد ترکیب مثبت و ۱ مورد ترکیب منفی می‌باشد.

همچنین بقیه داده‌های موجود در مجموعه داده‌ی هدف به عنوان داده بدون برچسب در نظر گرفته می‌شود. این روال ده بار برای هر مجموعه داده هدف تکرار می‌شود و در نهایت میانگین و انحراف معیار ROC-AUC برای هر وظیفه گزارش شده است.

معماری سیستم کدگذار ویژگی در داده‌های مرجع و هدف شامل لایه پیچشی گرافی با ابعاد فضای پنهان ۳۲، لایه ادغام، لایه پیچشی گرافی با ابعاد فضای پنهان ۳۲ و لایه ادغام می‌باشد. همچنین لایه‌های شبکه‌ی رده‌بند شامل لایه تجمیع گرافی، لایه تماما متصل با تابع فعال‌سازی relu و لایه تماما متصل با تابع فعال‌سازی softmax می‌باشد. اشتراک گذاری سیستم کدگذار ویژگی میان مجموعه داده های هدف برچسب دار و غیربرچسب دار به صورت سخت انجام شده است. به عبارت دیگر لایه‌های کدگذار ویژگی به صورت مشترک در داده های هدف برچسب دار و غیربرچسب دار مورد استفاده قرار می‌گیرد.

در این مقاله، روش پیشنهادی بر روی کامپیوتری با مشخصات Intel(R) Core(TM) i7-7700HQ CPU، NVIDIA GeForce GTX 1070 8Gb GDDR5 و 32 GB DDR4 RAM اجرا شده است. برای پیاده‌سازی روش پیشنهادی از پایتون ۳,۶ و کتابخانه‌های تنسورفلو و کراس بهره گرفته شده است به طور میانگین هر تکرار در روش پیشنهادی حدود چهار صدم ثانیه می‌باشد.

در این بخش، مجموعه‌ای از آزمایشات برای بررسی انتقال دانش میان مجموعه‌داده‌های بیوفیزیک شامل HIV و BACE صورت پذیرفته است. در این حالت دو مجموعه آزمایش صورت پذیرفته است: در بخش اول مجموعه داده‌ی HIV به عنوان مجموعه داده‌ی مرجع در نظر گرفته شده است و BACE به عنوان مجموعه داده‌ی هدف در نظر گرفته شده است، در بخش دوم نیز بالعکس در نظر گرفته شده است. در ادامه هر کدام از این آزمایشات با جزئیات مورد بررسی و تحلیل قرار می‌گیرند.

انتقال دانش از مجموعه داده‌ی HIV به مجموعه داده‌ی BACE. نتایج به دست آمده در این حالت در جدول ۱ نمایش داده شده است. روش پیشنهادی توانسته است در مقایسه با دیگر روش‌های پایه بهتر عمل نماید. در مقایسه با روش (Abbasi, et al. 2019)، در معیار ROC-AUC در دو حالت متفاوت از تعداد نمونه‌های آموزشی در دسترس به ترتیب ۲٪ و ۱٪ بهبود داشته است.

در جدول ۲، روش پیشنهادی با دیگر روش‌های موفق موجود در حوزه شامل LogReg، KernelSVM، XGBoost، RF، Influence (IRV)، Relevance Voting، Multitask، Bypass، GC و Wave مورد مقایسه قرار گرفته است. بایستی دقت گردد که در روش‌های مذکور، داده‌های موجود در هر آزمون به سه دسته‌ی مجزا شامل: آموزش، اعتبارسنجی و آزمایشی به ترتیب با نسبت‌های ۸۰، ۱۰ و ۱۰ تقسیم می‌گردد. روش پیشنهادی با روش‌های مذکور در دو حالت متفاوت مورد مقایسه قرار گرفته است: ۱) برای داشتن مقایسه‌ای عادلانه، مدل بر روی مجموعه آموزشی یکسانی با روش‌های مذکور آموزش دیده و بر روی مجموعه آزمایشی یکسانی ارزیابی شده است. ۲) روش پیشنهادی با تنظیمات داده‌های کم (تنظیمات اصلی روش پیشنهادی) مورد ارزیابی قرار گرفته است.

همانگونه که در جدول ۲**Error! Reference source not found.** نشان داده شده است روش پیشنهادی در حالت اول کارایی قابل قبولی نسبت به دیگر روش‌های قابل مقایسه دریافت کرده است. همچنین روش پیشنهادی در مقایسه با (Abbasi, et al. 2019) نتایج قابل مقایسه‌ای را بدست آورده است. در حالت دوم (تعداد داده‌های آموزشی کم)، روش پیشنهادی از روش Weave بهتر عمل نموه است و نتایج قابل مقایسه‌ای با روش‌های GC، IRV و Logreg دریافت کرده است. بایستی دقت شود که روش‌های Weave، GC، IRV و Logreg با ۱۱۴۲ داده آموزش دیده و بر روی ۱۴۳ داده آزمایش گردیده‌اند. روش پیشنهادی در این حالت بر روی ۲۰ نمونه آموزش دیده و بر روی ۱۴۰۷ نمونه آزمایش گردیده است. نتایج جدول ۲ بیانگر این امر است که استفاده از شبکه پیچشی گرافی وزن‌دار شده یالی قادر به انتقال دانش بیشتری نسبت به شبکه پیچشی گرافی استاندارد است.

یادگیری انتقالی از مجموعه داده‌ی BACE به مجموعه داده‌ی HIV. نتایج به دست آمده در این بخش در جدول ۳ آورده شده است. روش پیشنهادی در مقایسه با روش (Abbasi, et al. 2019)، در معیار ROC-AUC در دو تنظیم یاد شده به ترتیب ۳٪ و ۱٪ بهبود داشته است.

[DOR: 20.1001.1.25885073.1399.9.2.11.6]

جدول ۱- مقایسه روش پیشنهادی با روش‌های پایه بر روی مجموعه داده BACE. در این آزمایش مجموعه داده HIV به عنوان مجموعه داده مرجع استفاده شده است. نمادهایی ۱۰+/۱۰-بیانگر ۱۰ نمونه آموزشی مثبت و ۱۰ نمونه آموزشی منفی در مجموعه داده هدف برچسب‌دار بوده است. به همین منوال ۱+/۱- بیانگر یک نمونه آموزشی مثبت و یک نمونه آموزشی منفی در مجموعه داده هدف برچسب‌دار بوده است.

Table 1- Comparison of our approach and other approaches on BACE dataset. In this experiment, the ROC-AUC score is reported. 10+/10- indicates that there are ten positive and ten negative samples in the labeled target data. 1+/1- indicates that there are one positive and one negative sample in the labeled target data.

روش	۱۰+/۱۰-	۱+/۱-
RF (100 trees)	0.65±0.06	0.52±0.03
SVM	0.46±0.07	0.39±0.03
GraphConv (GC)	0.55±0.03	0.47±0.06
Multitask(Source+labeled target data)	0.62±0.03	0.54±0.02
Abbasi, et al 2019	0.69±0.03	0.59±0.03
روش پیشنهادی	0.71±0.02	0.60±0.04

جدول ۲- مقایسه روش پیشنهادی با نه روش موفق موجود بر روی مجموعه داده BACE. در این آزمایش روش‌های مورد مقایسه بر روی ۸۰ درصد داده‌های آموزش دیده و بر روی ۱۰ درصد آزمایش گردیدند. این آزمایش ۱۰ بار تکرار شده و میانگین و انحراف معیار گزارش شده است.

Table 2- Comparison of our approach with nine states-of-the-art models on BACE dataset. The comparable models are trained with 80% of compounds per each task and are tested on 10% of compounds.

روش‌ها	**تعداد ترکیب‌های آموزشی به ازای هر وظیفه**	**تعداد ترکیب‌های آزمایش به ازای هر وظیفه**	**ROC-AUC**
Logreg			0.78±0.01
KernelSVM			0.86±0.00
XGBoost			0.85±0.00
RF			0.87±0.01
IRV			0.84±0.00
Multitask	۱۲۱۱	۱۵۱	0.82±0.01
Bypass			0.83±0.01
GC			0.78±0.01
Weave			0.81±0.00
(Abbasi, et al. 2019)			0.91±0.01
روش پیشنهادی			0.91±0.02
(Abbasi, et al. 2019)	۲۰(۱۰+/۱۰-)	۱۴۹۳	0.69±0.03
(Abbasi, et al. 2019)	۲(۱+/۱-)	۱۵۱۱	0.59±0.03
روش پیشنهادی	۲۰(۱۰+/۱۰-)	۱۴۹۳	0.71±0.02
روش پیشنهادی	۲(۱+/۱-)	۱۵۱۱	0.60±0.04

جدول ۳- مقایسه روش پیشنهادی با روش‌های پایه بر روی مجموعه داده HIV. در این آزمایش مجموعه داده HIV به عنوان مجموعه داده مرجع استفاده شده است. نمادهایی ۱۰+/۱۰-بیانگر ۱۰ نمونه آموزشی مثبت و ۱۰ نمونه آموزشی منفی در مجموعه داده هدف برچسب‌دار بوده است. به همین منوال ۱+/۱- بیانگر یک نمونه آموزشی مثبت و یک نمونه آموزشی منفی در مجموعه داده هدف برچسب‌دار بوده است.

Table 3- Comparison of our approach and other approaches on HIV dataset. In this experiment, the ROC-AUC score is reported. 10+/10- indicates that there are ten positive and ten negative samples in the labeled target data. 1+/1- indicates that there are one positive and one negative sample in the labeled target data.

روش	۱۰+/۱۰-	۱+/۱-
RF (100 trees)	0.52±0.06	0.56±0.07
SVM	0.43±0.09	0.41±0.05
GraphConv (GC)	0.50±0.06	0.45±0.03
Multitask(Source+labeled target data)	0.58±0.02	0.55±0.02
(Abbasi, et al. 2019)	0.66±0.04	0.61±0.04
روش پیشنهادی	0.69±0.02	0.62±0.04

[DOR: 20.1001.1.25885073.1399.9.2.11.6]

مشابه آزمایشات قبلی، روش پیشنهادی با روش‌های LogReg، KernelSVM، XGBoost، Influence Relevance Voting (IRV)، Multitask، Bypass، GC و Wave در جدول ۴ مورد مقایسه قرار گرفته است. همانگونه که در جدول ۴ نشان داده شده است روش پیشنهادی در تنظیمات یکسان نسبت به (Abbasi, et al. 2019) ۱٪ بهبود داشته است. در تنظیمات تعداد داده‌های کم، روش کارایی قابل مقایسه‌ای در مقایسه با روش‌های LogReg، IRV، Multitask، Bypass و Weave کسب کرده است همچنین نسبت به روش (Abbasi, et al. 2019) دقت بالاتری بدست آمده است. نتایج به دست آمده بیانگر این امر است که روش با موفقیت دانش ارزشمندی را از مجموعه داده‌ی BACE به مجموعه داده‌ی HIV انتقال داده است.

جدول ۴- مقایسه روش پیشنهادی با نه روش موفق موجود بر روی مجموعه داده HIV. در این آزمایش روش‌های مورد مقایسه بر روی ۸۰ درصد داده‌های آموزش دیده و بر روی ۱۰ درصد آزمایش گردیدند. این آزمایش ۱۰ بار تکرار شده و میانگین و انحراف معیار گزارش شده است.

Table 4- Comparison of our approach with nine states-of-the-art models on HIV dataset. The comparable models are trained with 80% of compounds per each task and are tested on 10% of compounds.

روش‌ها	تعداد ترکیب‌های آموزشی به ازای هر وظیفه	تعداد ترکیب‌های آزمایش به ازای هر وظیفه	ROC-AUC
Logreg	۳۲۹۰۲	۴۱۱۲	0.70±0.02
KernelSVM			0.79±0.00
XGBoost			0.76±0.00
IRV			0.74±0.00
Multitask			0.70±0.04
Bypass			0.69±0.03
GC			0.76±0.02
Weave			0.70±0.04
(Abbasi, et al. 2019)			0.86±0.04
روش پیشنهادی			0.87±0.04
(Abbasi, et al. 2019)	۲۰(۱۰+/۱۰-)	۴۱۱۰۷	0.66±0.04
(Abbasi, et al. 2019)	۲(۱+/۱-)	۴۱۱۲۵	0.61±0.04
روش پیشنهادی	۲۰(۱۰+/۱۰-)	۴۱۱۰۷	0.69±0.02
روش پیشنهادی	۲(۱+/۱-)	۴۱۱۲۵	0.62±0.04

نتیجه‌گیری کلی

در این مقاله، هدف اصلی، طراحی یک مدل برای پیش‌بینی فعالیت یک ترکیب دارویی در آزمون با تعداد کم داده‌های برچسب‌خورده‌ی در دسترس و دخالت دادن پیوند مولکلی بین اتم‌ها است. بدین منظور فرض می‌گردد که هر مولکول شامل زیرساختارهای محلی است که در پیش‌بینی فعالیت ترکیب دارویی تاثیر بسزایی دارند. اما با توجه به چالش تعداد کم داده‌های آموزشی در دسترس، شبکه EGCN قادر به آموزش صحیح نخواهد بود. بنابراین از یک چارچوب استفاده شده است که هدف اصلی آن انتقال دانش از مجموعه داده‌ی مرجع (مجموعه داده‌ی کمکی) به مجموعه داده‌ی هدف است. با توجه به اینکه آزمون‌های مرجع و هدف برای گروه‌های پروتئین و ترکیبات دارویی متفاوتی با توزیع‌های متفاوت انجام شده است، از یک تطبیق دامنه رقابتی برای نگاشت توصیف‌گر داده‌های دامنه مرجع و هدف به یک فضای جدید با توزیع یکسان بهره گرفته شده است. این امر موجب می‌گردد که مجموعه یکسانی از قطعه‌ها که در داده‌های هر دو دامنه تاثیرگذار هستند، آموزش ببینند. برای ارزیابی مدل، از مجموعه داده‌های دسته بیوفیزیک استفاده گردید. نتایج آزمایشات بیانگر این حقیقت بود که استفاده از بردارهای ویژگی پیوندهای مولکولی قادر به استخراج زیرساختارهای موثرتری برای انتقال دانش میان مجموعه داده‌ها بوده است.

[DOR: 20.1001.1.25885073.1399.9.2.11.6]

منابع

Abbasi, K., A. Poso, J. Ghasemi, M. Amanlou and A. Masoudi-Nejad. 2019. Deep Transferable Compound Representation across Domains and Tasks for Low Data Drug Discovery. *Journal of chemical information and modeling* **59**(11): 4528-4539.

Abbasi, K., P. Razzaghi ,A. Poso, M. Amanlou, J. B. Ghasemi and A. Masoudi-Nejad. 2020. DeepCDA: Deep Cross-Domain Compound-Protein Affinity Prediction through LSTM and Convolutional Neural Networks. *Bioinformatics*.

Altae-Tran, H., B. Ramsundar, A. S. Pappu and V. Pande. 2017 .Low data drug discovery with one-shot learning. *ACS central science* **3**(4): 283-293.

Amin, N., A. McGrath and Y. P. P. Chen. 2019. Evaluation of deep learning in non-coding RNA classification. *Nature Machine Intelligence* **1**(5): 246.

Asgari, E., P. C. Münch ,T. R. Lesker, A. C. McHardy and M. R. Mofrad. 2019. Ditaxa: Nucleotide-pair encoding of 16s rrna for host phenotype and biomarker detection. *Bioinformatics* **35**(14): 2498-2500.

Asgari, E., N. Poerner, A. McHardy and M. Mofrad. 2019. DeepPrime2Sec: Deep Learning for Protein Secondary Structure Prediction from the Primary Sequences. *bioRxiv*: 705426.

Bagherian, M., E. Sabeti, K. Wang, M. A. Sartor, Z. Nikolovska-Coleska and K. Najarian. 2020. Machine learning approaches and databases for prediction of drug-target interaction: a survey paper. *Briefings in bioinformatics*.

Bahdanau, D., K. Cho and Y. Bengio. (2015). *Neural machine translation by jointly learning to align and translate*. International Conference on Learning Representations (ICLR).

Bruna, J., W. Zaremba, A. Szlam and Y. LeCun. 2013. Spectral networks and locally connected networks on graphs. *arXiv. 2013:1312-6203*.

Chakravarti, S. K. and S. R. M. Alla. 2019. Descriptor Free QSAR Modeling Using Deep Learning With Long Short-Term Memory Neural Networks .*Frontiers in Artificial Intelligence* **2**.

Dahl, G. 2012. Deep learning how I did it: Merck 1st place interview. *Online article available from*

Duvenaud, D., D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik and R. Adams. (2015). *Convolutional networks on graphs for learning molecular fingerprints*. Advances in neural information processing systems.

Ezzat, A., M. Wu, X. L. Li and C. K. Kwoh. 2018. Computational prediction of drug-target interactions using chemogenomic approaches: an empirical survey. *Brief Bioinform* **8**.

Fooshee, D., A. Mood, E. Gutman, M. Tavakoli, G. Urban, F. Liu, N. Huynh, D. Van Vranken and P. Baldi. 2018. Deep learning for chemical reaction prediction. *Molecular Systems Design & Engineering* **3**(3): 442-452.

Gao, K. Y ,.A. Fokoue, H. Luo, A. Iyengar, S. Dey and P. Zhang. (2018). *Interpretable Drug Target Prediction Using Deep Neural Representation*. IJCAI.

Gong, L. and Q. Cheng. (2019). *Exploiting edge features for graph neural networks*. IEEE Conference on Computer Vision and Pattern Recognition.

Goodfellow, I., Y. Bengio and A. Courville. 2016. *Deep learning*, MIT Press.

Hassan, M. M., D. C. Mogollón, O. Fuentes and S. Sirimulla. 2018. DLSCORE: A Deep Learning Model for Predicting Protein-Ligand Binding Affinities.

Hooshmand, S. A., S. A. Jamalkandi, S. M. Alavi and A. Masoudi-Nejad. 2020. Distinguishing drug/non-drug-like small molecules in drug discovery using deep belief network. *Molecular Diversity*: 1-12.

Kearnes, S., K. McCloskey, M. Berndl, V. Pande and P. Riley .۲۰۱۶ .Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design* **30**(8): 595-608.

Kearnes, S., K. McCloskey, M. Berndl, V. Pande and P. Riley. 2016. Molecular Graph Convolutions: Moving Beyond Fingerprints. *J. Comput.-Aided Mol. Des.* **30**: 595−608.

Landrum, G. 2006. RDKit: Open-source cheminformatics.

Leelananda, S. P. and S. Lindert. 2016. Computational methods in drug discovery. *Beilstein journal of organic chemistry* **12**: 2694-2718.

Masoudi-Sobhanzadeh, Y., H. Motieghader and A. Masoudi-Nejad. 2019. FeatureSelect: a software for **feature** selection based on machine learning approaches. *BMC bioinformatics* **20**(1): 170.

McCann, B., J. Bradbury, C. Xiong and R. Socher. (2017). *Learned in Translation: Contextualized word vectors*. Advances in Neural Information Processing Systems.

Min, S., B. Lee and S. Yoon. 2017. Deep learning in bioinformatics. *Briefings in bioinformatics* **18**(5): 851-869.

Oglic, D., S. A. Oatley, S. J. Macdonald, T. Mcinally, R. Garnett, J. D. Hirst and T .Gärtner. 2018. Active Search for Computer-aided Drug Design. *Molecular informatics* **37**: 1700130.

Öztürk, H., A. Özgür and E. Ozkirimli. 2018. DeepDTA: deep drug-target binding affinity prediction. *Bioinformatics* **34**(17): i821-i829.

Pahikkala, T., A. Airola, S. Pietilä, S. Shakyawar, A. Szwajda, J. Tang and T. Aittokallio. 2014. Toward more realistic drug-target interaction predictions. *Briefings in bioinformatics* **16**(2): 325-337.

Pham, H. N. and T. H. Le. 2019. Attention-based Multi-Input Deep Learning Architecture for Biological Activity

[DOR: 20.1001.1.25885073.1399.9.2.11.6]

Prediction: An Application in EGFR Inhibitors. *arXiv preprint arXiv:1906.05168*.

Pope, P. E., S. Kolouri, M. Rostami, C. E. Martin and H. Hoffmann. (2019). *Explainability Methods for Graph Convolutional Neural Networks*. IEEE Conference on Computer Vision and Pattern Recognition.

Popova, M., O. Isayev and A. Tropsha. 2018. Deep reinforcement learning for de novo drug design. *Science advances* **4**.(۷)

Ramsundar, B., S. Kearnes, P. Riley, D. Webster, D. Konerding and V. Pande. 2 .۰۱۵Massively multitask networks for drug discovery. *arXiv preprint arXiv*:.*02072*.

Simões, R. S., V. G. Maltarollo, P. R. Oliveira and K. M. J. F. i. p. Honorio. 2018. Transfer and multi-task learning in QSAR modeling: advances and challenges. **9**: 74.

Tsubaki, M., K. Tomii and J. Sese. 2018. Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics* **35**(2): 309-318.

Velickovic, P., G. Cucurull, A. Casanova and A. Romero. (2018). *Graph Attention Networks*. International Conference onLearning Representations.

Voulodimos, A., N. Doulamis, A. Doulamis and E. Protopapadakis. 2018. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*.

Waring, M. J., J. Arrowsmith, A. R. Leach, P. D. Leeson, S. Mandrell, R. M. Owen, G. Pairaudeau, W. D. Pennie, S. D. Pickett and J. Wang. 2015. An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nature reviews Drug discovery* **14**: 475.

Wen ,M., Z. Zhang, S. Niu, H. Sha, R. Yang, Y. Yun and H. Lu. 2017. Deep-learning-based drug-target interaction prediction. *Journal of proteome research* **16**(4): 1401-1409.

Wenzel, J., H. Matter and F. Schmidt. 2019. Predictive Multitask Deep Neural Network Models for ADME-Tox Properties: Learning from Large Datasets. *Journal of chemical information.*

Wu, Z., B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing and V. Pande. 2018. MoleculeNet: a benchmark for molecular machine learning. *Chemical science* **9**(2): 513-530.

Young, T., D. Hazarika, S. Poria and E. Cambria. 2018. Recent trends in deep learning based natural language processing. *IEEE Computational intelligenCe magazine* **13**(3): 55-75.

Zou, J., M. Huss, A. Abid, P. Mohammadi, A. Torkamani and A. Telenti. 2019. A primer on deep learning in genomics. *Nature genetics* **51**(1): 12-18.

Genetic Engineering and Biosafety Journal
Volume 9, Number 2
2021

Design of a bioinformatics model to predict drug compound properties and its application in inhibition of HIV replication and BACE-1

Karim Abbasi, Ali Masoudi-Nejad*

Laboratory of system Biology and Bioinformatics (LBB), Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran.
*Corresponding Author, Email: amasoudin@ut.ac.ir

Abstract

In this paper, a new method for the problem of predicting the compound molecule properties in the lead optimization step in drug design is presented. In the lead optimization step, the amount of available biological data on small molecule compounds is low. In recent years, this challenge has been considered and transfer learning and deep learning techniques have been used to solve it. For this purpose, similar data sets have been used as auxiliary data to learn a reliable model. In this method, compound feature extraction plays an essential role in transferring knowledge from similar (auxiliary) data sets to the target data set. In this paper, the effect of using Edge weighted Graph Convolutional Network (EGCN) is assessed which able to consider the feature vector of the compound bond as well as the atom feature vector. To evaluate the method, we have applied the proposed approach on BACE and HIV datasets. The obtained results show that the proposed method is able to extract more efficient knowledge from similar data sets to transfer to the target data set.

Keywords: Edge weighted Graph Convolutional Network, Molecular bonds, Deep learning, Transfer learning, Drug Design.

[DOR: 20.1001.1.25885073.1399.9.2.11.6]