

## پیش‌بینی برهم‌کنش پروتئین-پروتئین بین ویروس و انسان با استفاده از شبکه‌های دوقلوی ناهمگون

سارا محمدزاده<sup>۱</sup>، زهرا قربانعلی<sup>۲</sup>، فاطمه زارع-میرکآباد<sup>۲\*</sup>

۱- کارشناس علوم کامپیوتر، مرکز زیست‌محاسباتی CBRC، دانشکده ریاضی و علوم کامپیوتر، دانشگاه صنعتی امیرکبیر، تهران، ایران.

۲- دکتری علوم کامپیوتر، مرکز زیست‌محاسباتی CBRC، دانشکده ریاضی و علوم کامپیوتر، دانشگاه صنعتی امیرکبیر، تهران، ایران.

\* صندوق پستی: ۱۵۹۱۶۳۴۳۱۱، تهران، ایران.

f.zare@aut.ac.ir

دریافت: ۱۴۰۲/۰۹/۱۵

پذیرش: ۱۴۰۳/۱۲/۰۴

### چکیده

عفونت‌های ویروسی بیماری‌هایی هستند که توسط ویروس‌ها پس از ورود به سلول‌های میزبان و بر اثر همانندسازی به وجود می‌آیند. عفونت‌زایی، با ایجاد برهم‌کنش میان پروتئین‌های ویروس و پروتئین‌های سلول میزبان صورت می‌گیرد. از این‌رو، شناسایی این برهم‌کنش‌ها نقش بسزایی در جلوگیری، درمان و کنترل عفونت‌زایی دارد. با توجه به اینکه مطالعات آزمایشگاهی بسیار پرهزینه و زمان‌بر هستند، در سال‌های اخیر محققان با استفاده از روش‌های محاسباتی به پیش‌بینی برهم‌کنش میان پروتئین‌های ویروس و انسان می‌پردازند. گرچه این روش‌ها عملکرد مناسبی دارند، اما یکی از چالش‌های اصلی آنها به‌کارگیری نمایش مناسبی برای پروتئین‌ها است که بتواند اطلاعات ساختاری آنها را در بر داشته باشد. در این مقاله قصد داریم چارچوبی به نام PBS برای پیش‌بینی برهم‌کنش پروتئین-پروتئین بین ویروس‌ها و انسان ارائه دهیم که از توانایی ترنسفورمرها برای نمایش پروتئین‌ها استفاده می‌کند. این مدل با به‌کارگیری شبکه‌های عصبی دوقلو ناهمگون، فضای نمایش را یکپارچه‌سازی می‌کند و در نهایت برهم‌کنش میان پروتئین‌های ویروس و انسان را مدل می‌کند. چارچوب PBS با کسب امتیاز ACC برابر با ۸۱/۴۱ درصد، امتیاز AUC-ROC برابر با ۸۷/۳۵ درصد، امتیاز AUC-PR برابر با ۸۷/۷۸ درصد، امتیاز F1 برابر با ۸۱/۵۸ درصد و امتیاز Precision برابر با ۸۰/۸۴ درصد عملکرد مناسبی از خود نشان می‌دهد.

کپی‌رایت © ۲۰۲۵، انتشارات دانشگاه تربیت مدرس (TMU Press). این مقاله به‌صورت دسترسی آزاد منتشر شده و تحت مجوز بین‌المللی Creative Commons Attribution-NonCommercial 4.0 قرار دارد. بر اساس این مجوز، شما می‌توانید این مطلب را در هر قالب و رسانه‌ای کپی، بازنشر و بازآفرینی کنید و یا آن را ویرایش و بازسازی نمایید، به شرط آنکه نام نویسنده را ذکر کرده و از آن برای مقاصد غیرتجاری استفاده کنید.

همچنین، توانایی مدل در پیش‌بینی برهم‌کنش‌های بین پروتئین‌های ویروس آنفلانزا H1N1 و انسان به‌عنوان مورد مطالعاتی سنجیده می‌شود.

**کلید واژگان:** برهم‌کنش پروتئین-پروتئین، ترنسفورمر، ProtBert، شبکه‌های عصبی دوقلو، فضای نمایش.

## ۱- مقدمه

بیماری‌های ایجاد شده توسط ویروس‌ها، که عفونت‌های ویروسی نامیده می‌شوند، هر ساله هزینه‌های سنگینی بر دوش سیستم‌های مراقبت بهداشتی سراسر جهان می‌گذارند. از مثال‌های شناخته شده در این حوزه می‌توان به همه‌گیری بیماری کوید-۱۹ و ایدز اشاره کرد که به ترتیب ناشی از عفونت‌زایی ویروس‌های SARS CoV-2 و HIV هستند و علاوه بر مرگ‌ومیرهای فراوان، منجر به خسارات عظیم اجتماعی و اقتصادی شده‌اند [۱]. از این رو، یافتن روش‌های مقابله با ویروس‌ها ضروری است. یکی از لازمه‌های مقابله با ویروس‌ها، شناسایی نحوه عملکرد آنها می‌باشد. با توجه به اینکه این موجودات، زنده نیستند، برای تکمیل چرخه همانندسازی<sup>۱</sup> خود به سلول میزبان<sup>۲</sup> نیاز دارند و برای عفونت‌زایی، با سلول میزبان برهم‌کنش پروتئین-پروتئین<sup>۳</sup> تشکیل می‌دهند [۴-۲]. شناسایی زود هنگام برهم‌کنش‌های میان پروتئین‌های ویروس و پروتئین‌های سلول میزبان، می‌تواند دید مناسبی نسبت به عملکرد و انتشار آنها فراهم کند. همچنین، محققان قادر خواهند بود از این دانش برای جلوگیری و درمان بیماری‌های ناشی از ویروس‌ها بهره ببرند [۵].

برهم‌کنش‌های میان ویروس و سلول میزبان غالباً از طریق انجام آزمایش‌هایی نظیر Yeast Two-Hybrid و Affinity purification شناسایی می‌شود [۶]. با توجه به عدم مقیاس‌پذیری و زمان‌بر بودن روش‌های آزمایشگاهی، محققان به دنبال استفاده از مدل‌های محاسباتی برای حل مسئله پیش‌بینی برهم‌کنش‌های پروتئین-پروتئین بین

ویروس و انسان هستند، که به اختصار آن را HVPP<sup>۴</sup>

نامیده می‌شوند [۷].

در مطالعات اولیه محاسباتی، دانشمندان برای حل مسئله HVPP از آنالیز ویژگی‌های فیزیکوشیمیایی نظیر آب‌دوستی، حل‌پذیری، برآمدگی و دسترسی‌پذیری استفاده می‌کردند [۸ و ۹]. هرچند مدل‌های ابتدایی عملکرد قابل قبولی داشتند، اما اطلاعات ساختاری پروتئین را در نظر نمی‌گرفتند. این در صورتی است که عملکرد پروتئین‌ها مبتنی بر ساختار آنها صورت می‌پذیرد و برهم‌کنش بین پروتئین‌ها نیز وابسته به این ویژگی است. به همین دلیل، در ادامه مسیر حل مسئله HVPP، محققان به استفاده از ساختار پروتئین‌ها روی آوردند. در این راستا، پازوس<sup>۵</sup> و همکاران برای تشخیص مناطق برهم‌کنش پروتئین‌ها از ساختار اول آنها (توالی پروتئین‌ها) استفاده کردند [۱۰]. این مدل با بهره‌مندی از روش‌های هم‌ترازی توالی‌ها و تمرکز بر جهش‌های مرتبط با آنها به پیش‌بینی برهم‌کنش بین پروتئین‌ها می‌پردازد [۱۰]. همچنین، گالت<sup>۶</sup> و همکاران با استفاده از ساختار اول پروتئین‌ها، تحلیل زیست‌فعالی آنها و در نظر گرفتن فراوانی اسیدآمینه‌ها، مناطق برهم‌کنش‌های پروتئینی را پیش‌بینی کردند [۱۱]. چن<sup>۷</sup> و همکاران با ارائه مدلی به نام ECA-HVP برای پیش‌بینی برهم‌کنش میان پروتئین‌های انسانی و ویروس، چندین مازولیتی کدگذاری گوناگون من جمله AAC، DDE و MMI را برای نمایش اولیه پروتئین‌ها با یکدیگر هم‌افزا کردند [۱۲]. در سال‌های اخیر افزایش دادگان برهم‌کنش پروتئین-پروتئین بین ویروس‌ها و انسان، ظهور مدل‌های

5 Pazos  
6 Gallet  
7 Chen

1 Replication  
2 Host cell  
3 Protein-Protein Interactions  
4 Human – Virus Protein-protein interaction Prediction

یادگیری ماشین و قدرت‌مند شدن سیستم‌های کامپیوتری محققان را به استفاده گسترده از قابلیت‌های یادگیری ماشین در یافتن الگوهای برهم‌کنش پروتئینی علاقه‌مند کرده است [۱۳].

عملکرد مدل‌های یادگیری ماشین وابسته به داده‌های مثبت و منفی است. در این مسئله، برهم‌کنش‌های شناخته شده میان پروتئین‌های ویروس و انسان مجموعه داده مثبت را تشکیل می‌دهند. این درحالی است که اطلاعات مربوط به عدم وجود برهم‌کنش میان پروتئین‌ها به صورت تجربی در دسترس نیست. بنابراین، یکی از چالش‌های اصلی در مسئله HVPP روش انتخاب داده منفی است. دو رویکرد اصلی مطالعات این حوزه برای حل این چالش عبارتند از، مبتنی بر داده مثبت و مبتنی بر داده منفی. در ادامه هر یک از این روش‌ها توضیح داده می‌شوند.

در روش مبتنی بر داده مثبت، فرض بر این است که نحوه انتخاب داده منفی در عملکرد مدل تاثیرگذار است. بنابراین، برای جلوگیری از بروز خطا، تنها از نمونه‌های مثبت برای یادگیری مدل استفاده می‌کنند. هدف اصلی در این مطالعات پیش‌بینی برهم‌کنش‌های ناشناخته میان پروتئین‌های انسان و ویروس است. برای این منظور، شینگ‌لی<sup>۸</sup> و همکاران از روش‌های احتمالاتی تجزیه ماتریسی استفاده کردند [۱۴]. همچنین، نورانی<sup>۹</sup> و همکاران مدل هسته‌گذاری داده‌های متنوع را پیشنهاد کردند که تنها از داده‌های مثبت برای حل مسئله HVPP استفاده می‌کند [۱۵].

در روش‌های مبتنی بر داده منفی، محققان مسئله را به صورت یک مسئله کلاسه‌بندی دودویی تعریف می‌کنند. از این رو، سعی در ایجاد فرضیه مناسب برای در نظر گرفتن داده منفی مناسب دارند. در این راستا، عید<sup>۱۰</sup> و همکاران فرض کردند اگر دو پروتئین ویروسی با یکدیگر مشابه

باشند و پروتئین انسانی با یکی از این پروتئین‌ها برهم‌کنش داشته باشد، آن پروتئین انسانی نمی‌تواند با پروتئین ویروسی دیگر نیز به‌عنوان داده منفی جفت شود [۱۶]. در مطالعه‌ای دیگر، باسط<sup>۱۱</sup> و همکاران در ادامه مدل پیشنهادی قبلی از عدم شباهت وزن‌دار توالی‌های اسیدآمینه‌ای برای ساخت مجموعه نمونه‌های منفی استفاده کردند [۱۷]. هرچند هر دو دسته سعی در حل چالش داده منفی می‌کنند، با این وجود عدم استفاده از نمونه منفی، یادگیری مدل را دچار مشکل می‌کند و تولید نمونه منفی با روش‌های گفته شده نیز باعث سوگیری مدل می‌شود [۱۸].

یکی دیگر از چالش‌های اصلی در این مسئله، نحوه نمایش پروتئین‌ها است. لیو<sup>۱۲</sup> و همکاران برای نمایش پروتئین‌ها از روش one-hot-encoding استفاده کردند [۱۹]. برای این منظور، هر توالی به شکل یک ماتریس با ۲۲ سطر و ۱۰۰۰ ستون نشان داده می‌شود. در این نمایش سطرها نماینده ۲۲ اسیدآمینه موجود و ستون‌ها نشان‌دهنده موقعیت هر اسیدآمینه در طول توالی می‌باشند. در این صورت، اگر طول توالی یک پروتئین کوتاه‌تر از ۱۰۰۰ اسیدآمینه باشد، طول آن را با تکرار به اندازه مطلوب می‌رسانند و توالی‌های طولانی‌تر از ۱۰۰۰ اسیدآمینه را نیز برش می‌زنند. همچنین، یانگ<sup>۱۳</sup> و همکاران برای تعبیه توالی پروتئین‌ها از تکنیک doc2vec استفاده کردند [۲۰]. برای این کار هر توالی پروتئین را به‌عنوان یک جمله در نظر گرفتند که با کمک الگوریتم word2vec و doc2vec قابلیت ارائه یک نمایش مناسب از پروتئین‌ها را دارد. ژو<sup>۱۴</sup> و همکاران در مطالعه‌ای دیگر، یک روش نمایش برای خوشه‌بندی ساختار اولیه توالی‌های پروتئینی براساس اسیدآمینه توالی‌ها پیشنهاد کردند. آنها برای پیش‌بینی برهم‌کنش بین پروتئین‌ها از SVM استفاده کردند [۲۱].

<sup>12</sup> Liu  
<sup>13</sup> Yang  
<sup>14</sup> Zhou

<sup>8</sup> Shing Lee  
<sup>9</sup> Nourani  
<sup>10</sup> Eid  
<sup>11</sup> Basit

درون گونه‌ای انسان است. این دو تابع با ضریب  $\alpha$  در تابع جریمه مدل ( $\mathcal{L}$ ) به صورت زیر تاثیر می‌گذارند:

$$\mathcal{L} = \mathcal{L}_1 + \alpha \cdot \mathcal{L}_2 \quad \text{معادله ۱}$$

حالت خاصی از مدل MTT که  $\alpha$  صفر باشد، یا به عبارتی مدل تنها از شبکه برهم‌کنش پروتئین-پروتئین انسان و ویروس برای یادگیری استفاده کند، مدل STT نام دارد.

با این حال، برهم‌کنش میان پروتئین‌ها بر اساس ساختار سه بعدی آنها شکل می‌گیرد و مطالعات معرفی شده مبتنی بر ساختار اول آنها هستند. این امر به آن دلیل است که ساختار سه بعدی پروتئین‌ها و علی‌الخصوص پروتئین‌های ویروس به دلیل جهش‌های مکرر، در دسترس نیست.

مقصود از انجام این مطالعه، ارائه مدلی با نام PBS برای مسئله پیش‌بینی برهم‌کنش پروتئین‌های ویروس و انسان است که به حل چالش‌های مطرح شده می‌پردازد. برای حل چالش انتخاب داده منفی، مدل‌های یادگیری گروهی مبتنی بر رای‌گیری را به کار گرفته می‌شود. برای حل چالش نحوه نمایش پروتئین‌ها از مزایای روش‌های انتقال یادگیری و ترنسفورمرها استفاده می‌شود که توانایی درک ویژگی‌های مربوط به ساختار سوم را دارند. در ادامه برای مدل کردن مسئله، با توجه به تفاوت ذات پروتئین‌های ویروس و انسان، از مدل‌های شبکه عصبی دوقلوی ناهمگون استفاده می‌شود. شبکه‌های عصبی دوقلوی ناهمگون، یکی از مدل‌های یادگیری عمیق هستند که می‌توانند فضای ویژگی مشترک مناسبی بین ورودی‌هایی از جنس متفاوت به وجود بیاورند و بر اساس آن الگوهای وجود رابطه را شناسایی کنند.

در ادامه، ابتدا به تبیین محاسباتی مسئله می‌پردازیم، سپس روش پیشنهادی خود را شرح داده و عملکرد مدل پیشنهادی را با سایر مطالعات پیشین مقایسه می‌کنیم. در نهایت کارایی مدل PBS را در پیش‌بینی برهم‌کنش‌های پروتئینی بین ویروس H1N1 و انسان می‌سنجیم.

یانگ<sup>۱۵</sup> و همکاران در مطالعه‌ای روش LightGBM را پیشنهاد دادند که ویژگی‌های چندوجهی را برای پیش‌بینی برهم‌کنش‌های میان پروتئین‌های ویروسی و انسانی با یکدیگر تلفیق می‌کند [۲۲]. در مطالعه‌ای دیگر، دانگ<sup>۱۶</sup> و همکاران برای حل مشکل نحوه نمایش پروتئین‌ها از مزایای روش‌های انتقال یادگیری استفاده کردند و مدل MTT را پیشنهاد کرد. در این مدل برای نمایش پروتئین‌ها مدل پیش‌آموزش داده شده UniRep به کار گرفته می‌شود [۱۸]. برتری مدل UniRep این است که چالش‌های مربوط به one-hot-encoding را ندارد. همچنین، به دلیل استفاده از روش LSTM، از حافظه برخوردار است و می‌تواند میان ورودی و خروجی ارتباط برقرار کند. ورودی مدل MTT دو توالی اسیدآمینه انسان و یک توالی اسیدآمینه ویروس است، که به کمک مدل UniRep به صورت بردار عددی به نمایش در می‌آیند. بردارهای مربوط به پروتئین‌های انسان و پروتئین ویروس به دو شبکه MLP<sup>۱۷</sup> مجزا داده می‌شوند. بردارهای خروجی MLP ها  $h'$  و  $h$  نامیده می‌شوند که  $v$  نشان‌دهنده خروجی مربوط به شبکه پروتئین ویروس و  $h'$  و  $h$  خروجی مربوط به شبکه پروتئین انسان هستند. در ادامه، برای مدل کردن مسئله HVPP شبکه دیگری تعریف می‌شود که ورودی آن حاصل ضرب مولفه به مولفه دو بردار  $v$  و  $h$  (به نام  $\gamma_1$ ) برای پیش‌بینی وجود برهم‌کنش میان پروتئین‌های ویروس و انسان، و حاصل ضرب مولفه به مولفه دو بردار  $h'$  و  $h$  (به نام  $\gamma_2$ ) برای پیش‌بینی برهم‌کنش‌های پروتئینی انسان می‌باشد. دو بردار  $\gamma_1$  و  $\gamma_2$  به طور جداگانه به دو شبکه با یک لایه خطی و تابع فعال‌ساز sigmoid داده می‌شوند. تابع  $\mathcal{L}_1$  تابع جریمه شبکه مربوط به پیش‌بینی برهم‌کنش پروتئین-پروتئین انسان و ویروس را محاسبه می‌کند. همچنین، تابع  $\mathcal{L}_2$ ، نشان‌دهنده تابع جریمه شبکه برهم‌کنش پروتئینی

<sup>17</sup> MultiLayer Perceptron

<sup>15</sup> Yang

<sup>16</sup> Dang

## ۲- مواد و روش‌ها

در این قسمت ابتدا مسئله را به طور دقیق تعریف می‌شود و سپس مدل پیشنهادی (PBS) شرح داده می‌شود.

### ۱-۲ مسئله برهم‌کنش پروتئین-پروتئین بین ویروس و انسان

هدف از مسئله HVPP شناسایی برهم‌کنش‌هایی است که بین پروتئین‌های ویروس و انسان شکل می‌گیرد و منجر به عفونت‌زایی و بیماری می‌شود. بنابراین، اگر  $H = \{h_1, h_2, \dots, h_n\}$  مجموعه پروتئین‌های انسان و به تعداد  $n$  پروتئین و همچنین  $V = \{v_1, v_2, \dots, v_m\}$  مجموعه پروتئین‌های ویروس‌ها و به تعداد  $m$  پروتئین باشد، هدف بررسی وجود برهم‌کنش بین جفت پروتئین‌های  $\langle h, v \rangle$  است به‌طوری‌که:

معادله ۲

$$HVPP \langle h, v \rangle = \begin{cases} 1 & \text{اگر دو پروتئین با یکدیگر برهم‌کنش داشته باشند} \\ 0 & \text{در غیر این صورت} \end{cases}$$

چنانچه از تعریف مسئله بر می‌آید، می‌توان آن را به‌صورت یک مسئله کلاسه‌بندی دودویی در نظر گرفت.

### ۲-۲ مجموعه داده

در این مطالعه مجموعه داده مورد استفاده برای برهم‌کنش‌های شناخته شده پروتئین-پروتئین بین ویروس‌ها و انسان از مقاله [۱۸] استخراج می‌شوند. این

برهم‌کنش‌ها، که به‌عنوان داده مثبت در نظر گرفته می‌شوند، به‌صورت زیر نمایش داده می‌شوند:

معادله ۳

$$P = \{ \langle h, v \rangle \mid h \in H \text{ and } v \in V \}$$

مجموعه  $P$  حاوی جفت پروتئین‌های ویروس و انسانی است که برهم‌کنش بین آنها به‌صورت تجربی در آزمایشگاه شناسایی شده است. با توجه به در دسترس نبودن جفت پروتئین‌هایی که عدم برهم‌کنش آنها ثابت شده باشد، داده‌های منفی از بین برهم‌کنش‌های ناشناخته میان پروتئین‌های مجموعه  $H$  و  $V$  به‌صورت تصادفی و به اندازه ۱۰ برابر مجموعه داده مثبت انتخاب می‌شوند. این مجموعه  $N$  نامیده شده و به شکل زیر تعریف می‌شود:

معادله ۴

$$N = \{ \langle h, v \rangle \mid h \in H \text{ and } v \in V \text{ and } \langle h, v \rangle \notin P \}$$

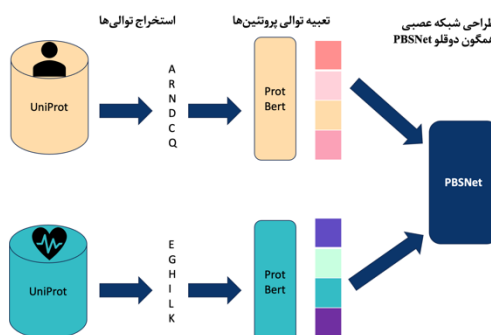
### ۳-۲ شرح چارچوب PBS

در این بخش قصد داریم به شرح چارچوب PBS بپردازیم که برای حل مسئله HVPP توسعه می‌یابد. چارچوب اصلی PBS متشکل از گام‌های زیر است که در ادامه به شرح دقیق هر یک از آنها می‌پردازیم. شکل ۱ چارچوب PBS را به‌طور تصویری نمایش می‌دهد:

استخراج توالی پروتئین‌ها

تعبیه توالی پروتئین‌ها

طراحی شبکه عصبی دوقلوی ناهمگون PBSNet



شکل ۱ چارچوب مدل PBS. در این چارچوب جفت پروتئین‌های استخراج شده ابتدا توسط ProtBert تعبیه می‌شوند و سپس برهم‌کنش آنها توسط مدل PBSNet که یک شبکه عصبی دوقلوی ناهمگون است پیش‌بینی می‌شود.

## ۲-۳-۱ استخراج توالی پروتئین‌ها

در اولین گام از چارچوب PBS لازم است برای هر یک از پروتئین‌های موجود در مجموعه داده، توالی آنها را استخراج کرد. یکی از پایگاه داده‌های مربوط به دادگان پروتئینی، که علاوه بر اطلاعات توالی‌ها سایر اطلاعات فیزیکوشیمیایی آنها را نیز دارا است، UniProt نام دارد [۲۳]. با استفاده از این پایگاه داده، اطلاعات توالی پروتئین‌ها استخراج می‌شوند.

## ۲-۳-۲ تعبیه توالی پروتئین‌ها

در گام دوم چارچوب PBS، با توجه به اینکه توالی‌های استخراج شده دارای طول‌های متفاوتی هستند، لازم است نمایشی از پروتئین‌ها ارائه شود که ضمن حفظ اطلاعات ساختاری آنها، از طول ثابت باشد. مدل‌های متفاوتی برای نمایش پروتئین‌ها وجود دارند که با استفاده از تعداد زیادی پروتئین به یادگیری ارائه تعبیه برای آنها می‌پردازند. برای انتخاب مدل مورد استفاده باید در نظر داشت یکی از چالش‌های اصلی در مسئله HVPP در دسترس نبودن ساختار سوم پروتئین‌ها است که باعث می‌شود اکثر مقالات از ساختار اول آنها استفاده کنند. با این حال، اگر هر توالی پروتئین به صورت یک جمله در زبان طبیعی در نظر گرفته شود، می‌توان از مزایای ترنسفورمرها در درک ویژگی‌های ساختاری پروتئین‌ها استفاده کرد. یکی از مدل‌های مبتنی بر ترنسفورمرها که براساس داده‌های پروتئینی آموزش دیده ProtBert نام دارد [۲۴]. چنانچه در [۲۴] نشان داده شده است، این مدل قادر به درک ویژگی‌های ساختار سوم براساس توالی پروتئین است. مدل PBS نیز برای حل چالش نحوه نمایش پروتئین‌ها و در نظر گرفتن ویژگی‌های ساختار سوم آنها از مدل ProtBert بهره می‌برد. بردارهای تولید شده توسط ProtBert از طول ۱۰۲۴ هستند و خروجی آن با  $\mathbb{P}_p$  نمایش داده می‌شود که  $p \in H \cup V$ .

## ۲-۳-۳ طراحی شبکه عصبی دوقلو ناهمگون

در سومین گام از چارچوب PBS و پس از استخراج بردار تعبیه پروتئین‌ها، پیش‌بینی برهم‌کنش میان آنها باید مدل شود. در این بخش، هدف ما ایجاد فضای پنهان مشترکی بین پروتئین‌های انسان و ویروس‌ها است که در آن پروتئین‌هایی که با یکدیگر برهم‌کنش دارند، به یکدیگر نزدیک و در غیر این صورت از یکدیگر فاصله بگیرند. یکی از مدل‌هایی که در این حوزه کاربرد دارد، شبکه‌های عصبی دوقلو هستند. شبکه‌های عصبی دوقلو از دو شبکه با وزن‌ها و لایه‌های به اشتراک گذاشته شده برای استخراج بردارهای تعبیه در فضای ویژگی پنهان استفاده می‌کنند که می‌تواند ارتباط یا عدم ارتباط بین دو ورودی را درک کند. با توجه به اینکه در مسئله HVPP، پروتئین‌های ویروس و انسان از دو طبیعت مختلف نشأت گرفته‌اند با الهام از مطالعه [۲۵]، شبکه PBSNet را براساس شبکه‌های عصبی دوقلو ناهمگون و مطابق با شکل ۲ پیشنهاد می‌شود.

به طور کلی، شبکه پیشنهادی از دو شاخه متمایز  $f(\mathbb{P}_h)$  و  $g(\mathbb{P}_v)$  به ترتیب برای هدایت تعبیه پروتئین‌های انسان ( $h \in H$ ) و پروتئین‌های ویروس‌ها  $v \in V()$  در جهت مدل کردن برهم‌کنش میان آنها ساخته می‌شود. وظیفه این دو شبکه بهبود نمایش ارائه شده توسط ProtBert و ایجاد فضای نمایش یکپارچه برای پروتئین‌های ویروس و انسان است. خروجی‌های این دوشبکه در ادامه به چندین لایه چگال برای پیش‌بینی برهم‌کنش بین پروتئین‌ها وارد می‌شوند. در ادامه به شرح جزئیات می‌پردازیم.

شبکه‌های  $f(\mathbb{P}_h)$  و  $g(\mathbb{P}_v)$  بردارهای به دست آمده از ProtBert را به عنوان ورودی دریافت می‌کنند و سپس به صورت یک تابع غیرخطی این ورودی را به فضای نهان مشترک بین ویروس‌های انسان و ویروس هدایت می‌کنند به طوری که:  $f(\mathbb{P}_h): R^{1024} \rightarrow R^k$  و  $g(\mathbb{P}_v): R^{1024} \rightarrow R^k$ . در این جا  $k$ ، بُعد بردار تعبیه پروتئین‌ها در فضای اشتراکی است. معماری شبکه  $f(\mathbb{P}_h)$  از دولا به چگال با تابع فعال‌ساز ReLu، که اغلب در طراحی شبکه‌های عمیق

تابع، تفاوت توزیع احتمالاتی دو شبکه را محاسبه می‌کند. هرچه مقدار واگرایی کمتر باشد، به این معنا است که توزیع دو شبکه به یکدیگر شبیه‌تر است. تابع واگرایی KL-divergence به صورت زیر محاسبه می‌شود و هدف ما کمینه کردن آن است.

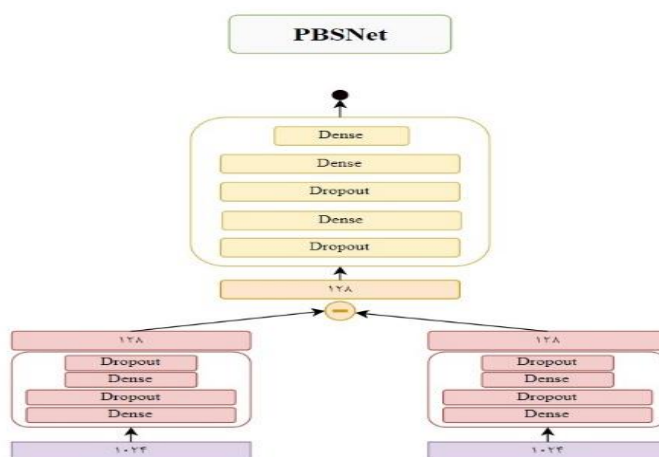
معادله ۵

$$D_{KL}(f, g) = \int \log\left(\frac{f(\mathbb{P}_h)}{g(\mathbb{P}_v)}\right)$$

تا به اینجا، برای دو پروتئین اولیه انسان و ویروس، به برداری از ابعاد پایین‌تر و مبتنی بر فضای اشتراک نهان دو گونه انسان و ویروس دست یافته‌ایم. در ادامه این معماری، برای کشف الگوی وجود یا عدم وجود برهم‌کنش بین این پروتئین‌ها، خروجی‌های حاصل از دو شبکه را با یکدیگر مقایسه کرده و تفاوت آنها را به صورت قدر مطلق محاسبه می‌شود.

استفاده می‌شود، ساخته شده است. لایه اول این شبکه دارای ۵۱۲ نورون و لایه دوم آن دارای ۲۵۶ نورون می‌باشد. همچنین، برای جلوگیری از بیش‌برازش بعد از هر لایه، یک لایه Dropout از مرتبه ۰/۲ نیز قرار داده می‌شود. خروجی نهایی این شبکه برداری با نام  $\mathbb{P}_h$  و از بعد  $R^k$  است. معماری شبکه  $g(\mathbb{P}_v)$  به طور مشابه با  $f(\mathbb{P}_h)$  تعریف می‌شود و بردار  $\mathbb{P}_v$  را از ابعاد  $R^k$  تولید می‌کند. با به کارگیری آزمایش و خطا مقدار  $k$  در این مطالعه برابر ۱۲۸ در نظر گرفته می‌شود. این دو شبکه به صورت مستقل از یکدیگر به روزرسانی گشته و متناسب با پروتئین‌های انسان یا ویروس بهینه‌سازی می‌شوند.

با توجه به هدف ما در ایجاد یک فضای ویژگی مشترک برای پروتئین‌های انسان و ویروس، لازم است در نهایت توزیع خروجی این دو شبکه به یکدیگر نزدیک باشد. بنابراین، تابع هدف<sup>۱۸</sup> این دو شبکه را برابر تابع واگرایی کولبک-لایبلر<sup>۱۹</sup> یا KL-divergence قرار می‌دهیم. این



شکل ۲ مدل PBSNet برای پیش‌بینی برهم‌کنش بین دو پروتئین انسان و ویروس. این شبکه از دو شاخه متمایز برای هدایت تعبیه پروتئین‌های انسان و ویروس در فضای نهان اشتراکی تولید شده در ادامه و پس از گذشتن از چند لایه چگال برای پیش‌بینی برهم‌کنش بین دو پروتئین ورودی استفاده می‌شود.

<sup>19</sup> Kullback-Leibler divergence

<sup>18</sup> Loss function

بنابراین، بردار  $\mathbb{P}_u = |\mathbb{P}_h - \mathbb{P}_v|$  ساخته می‌شود. با توجه به نحوه تعریف دو شبکه  $f(\mathbb{P}_h)$  و  $g(\mathbb{P}_v)$  که هدف آنها نزدیک کردن بردارهای اولیه پروتئین‌هایی که با یکدیگر برهم‌کنش دارند و دور کردن بردارهای پروتئین‌های با برهم‌کنش نامشخص در فضای اشتراکی بود، این بردار حاوی اطلاعات مربوط به وجود برهم‌کنش است. بنابراین، پروتئین‌هایی که با یکدیگر برهم‌کنش دارند در فضای یکپارچه در فاصله نزدیکی به یکدیگر قرار دارند. برای شناسایی وجود برهم‌کنش، بردار  $\mathbb{P}_u$  از دو لایه چگال با تابع فعال‌ساز ReLU عبور می‌کند. لایه اول این شبکه از ۶۴ نورون، لایه دوم آن ۳۲ نورون تشکیل می‌شوند. برای جلوگیری از بیش‌برازش میان لایه‌های چگال، لایه‌های Dropout از مرتبه ۰/۲ قرار می‌گیرد. در انتها، بردار نهایی با بعد ۳۲ به یک تک نورون با استفاده از تابع فعال‌سازی sigmoid می‌رود که یک خروجی بین صفر و یک تولید می‌کند. این خروجی احتمال وجود برهم‌کنش بین جفت پروتئین انسان و ویروس داده شده را نشان می‌دهد. اگر این عدد بزرگ‌تر از ۰/۵ باشد نشان‌دهنده این است که مدل برای جفت پروتئین ورودی، وجود برهم‌کنش و در غیراین صورت عدم وجود برهم‌کنش را پیش‌بینی کرده است.

### ۳- نتایج

در این بخش ابتدا پایگاه‌داده انتخابی و نحوه ارزیابی مدل برای داده تست بررسی می‌شود. سپس، معیارهای ارزیابی استفاده شده را معرفی می‌کنیم و عملکرد UniRep و ProtBert را بر روی شبکه PBSNet ارزیابی می‌کنیم. در ادامه چارچوب پیشنهادی را با مدل‌های پیشین مقایسه می‌کنیم. در نهایت صحت مدل را در پیش‌بینی برهم‌کنش‌های بین پروتئین ویروس H1N1 و انسان سنجیده می‌شود.

#### ۳-۱ پایگاه‌داده

برای استخراج برهم‌کنش‌های شناخته شده بین پروتئین‌های انسان و ویروس‌ها از مجموعه داده معرفی

شده در مقاله [۱۸] استفاده می‌کنیم. این مجموعه داده شامل ۶۵۱ پروتئین است که ۶۵۱ پروتئین آن متعلق به ویروس‌ها و ۴۰۰۰ پروتئین آن از گونه انسانی هستند و شامل ۱۱۲۳۹ برهم‌کنش میان پروتئین‌های ویروس و انسان است. این مجموعه داده را مجموعه داده ابتدایی می‌نامیم.

با توجه به اینکه چارچوب پیشنهادی PBS تعبیه اولیه پروتئین‌ها را براساس مدل ProtBert به‌دست می‌آورد و این مدل روی توالی‌هایی با طول کم‌تر از ۱۰۲۴ اسید آمینه آموزش دیده است، کیفیت تهیه تولیدشده برای توالی‌های بزرگ‌تر مطلوب نیست. به همین دلیل توالی‌های طولانی از مجموعه داده ابتدایی را حذف کرده و مجموعه داده جدیدی تشکیل می‌دهیم که آن را مجموعه داده ثانویه می‌نامیم. مجموعه داده ثانویه شامل ۴۳۶۰ پروتئین است که ۶۱۳ پروتئین آن پروتئین ویروس و ۳۷۴۷ پروتئین انسانی هستند و شامل ۹۲۹۳ برهم‌کنش است.

برای مقایسه منصفانه مدل PBS با مدل‌های پیشین، تقسیم‌بندی داده آموزشی و تست به‌صورت مشابه با مقاله [۱۸] انجام شده است.

#### ۳-۲ ارزیابی مدل برای مجموعه داده تست

یکی از مشکلاتی که سایر مطالعات به‌دلیل عدم وجود داده منفی با آن روبه‌رو هستند، تعریف داده منفی است. برای این منظور مشابه با مقاله [۱۸]، ۱۰ برابر داده‌های مثبت، داده منفی تصادفی انتخاب می‌شود. در این صورت، ۱۰ بسته داده منفی وجود دارد که تعداد جفت پروتئین‌های هر بسته به اندازه مجموع داده مثبت می‌باشد.

با توجه به عدم توازن داده‌های مثبت و منفی، باید در نظر داشت که مدل‌های یادگیری ماشین به سمت داده غالب، در اینجا داده‌های منفی، منحرف می‌شوند [۲۶]. برای حل این چالش از مدل یادگیری مبتنی بر رای‌گیری استفاده می‌کنیم. برای آموزش و ارزیابی مدل، ابتدا داده‌های مجموعه تست کنار گذاشته شده و سپس داده‌های مثبت با

معادله ۶

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

تعاریف مربوطه به TP، FN، FP و TN در جدول ۱ آمده است.

معیار AUC-ROC با استفاده از حد آستانه‌های مختلف، مساحت زیر نمودار مرتبط با نسبت نرخ مثبت صحیح (TPR) به نرخ خطای مثبت (FPR) را اندازه‌گیری می‌کند به طوری که:

معادله ۷

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

معیار AUC-PR، مساحت زیر نمودار نسبت امتیاز Precision را به Recall نمایش می‌دهد به طوری که:

معادله ۸

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

جدول ۱ تعاریف مربوط به TP، TN، FP و FN.

| علامت اختصاری | مفهوم                                                                                                  |
|---------------|--------------------------------------------------------------------------------------------------------|
| TP            | تعداد برهم‌کنش‌هایی که در مطالعات تجربی وجود دارند و مدل پیشنهادی نیز توانسته آنها را درست شناسایی کند |
| TN            | تعداد برهم‌کنش‌هایی که شناسایی نشده‌اند و مدل پیشنهادی نیز آنها را شناسایی نکرده است                   |
| FP            | تعداد برهم‌کنش‌هایی شناسایی نشده‌اند اما مدل پیشنهادی برهم‌کنش آنها را شناسایی کرده است                |
| FN            | تعداد برهم‌کنش‌هایی که در مطالعات تجربی وجود دارند اما مدل پیشنهادی نتوانسته آنها را شناسایی کند       |

هر یک از دسته‌های داده منفی، یک دسته داده آموزشی را تشکیل می‌دهند. از این دسته‌های داده آموزشی به صورت مجزا برای آموزش مدل استفاده می‌شود. پس از آنکه مدل با هر یک از دسته‌ها آموزش دید، از آن برای پیش‌بینی مجموعه تست استفاده می‌شود. در نهایت برای مجموعه تست، ۱۰ دسته خروجی تولید می‌شود و پیش‌بینی نهایی براساس رأی‌گیری بر روی تمام خروجی‌ها اعلام می‌شود. اگر برهم‌کنش بین یک جفت پروتئین انسان و ویروس از مجموعه تست در بیش از ۵ دسته آموزشی مثبت اعلام شود، خروجی مدل ۱ و در غیر این صورت ۰ خواهد بود.

### ۳-۳ معیارهای ارزیابی

برای ارزیابی کارایی مدل نیاز به معیارهایی است تا بتوان نتایج آن را با سایر مطالعات نیز مقایسه کرد. از این رو، در این مطالعه نتایج بر اساس ۴ معیار متفاوت سنجیده می‌شوند: دقت<sup>۲۰</sup> (ACC)، مساحت زیر نمودار مشخصه عملکرد گیرنده<sup>۲۱</sup> (AUC-ROC)، مساحت زیر نمودار فراخوان-دقت<sup>۲۲</sup> (AUC-PR) و امتیاز F1. در ادامه به معرفی معیارهای ارزیابی می‌پردازیم. در مسئله‌های طبقه‌بندی دودویی مقدار ACC طبق فرمول زیر محاسبه می‌شود:

<sup>22</sup> Area under the precision-recall curve

<sup>20</sup> Accuracy

<sup>21</sup> Area under receiver operating characteristic curve

معیار F1-score میانگین هارمونیک میان امتیازهای Precision و Recall است. این معیار به صورت زیر محاسبه می شود:

معادله ۹

$$F1 - score = \frac{Precision * Recall}{Precision + Recall}$$

در مسئله HVPP، هدف ارائه لیستی از برهم کنش های احتمالی به زیست شناسان برای بررسی های آتی و کاهش فضای جست و جو است. با توجه به اینکه آزمایش های شناسایی برهم کنش بین پروتئین ها زمان بر و پرهزینه می باشد، تعداد برهم کنش های مثبت کاذب این لیست باید تا حد امکان محدود باشد. به عبارت دیگر، مدل هایی که

جدول ۲ جزئیات امتیازات کسب شده در هر دسته را نشان می دهد.

### ۳-۵ ارزیابی نمایش پروتئین ها

یکی از چالش های اصلی در مسئله HVPP نحوه نمایش پروتئین ها است. برای غلبه بر این چالش مدل PBS از مدل پیش آموزش دیده ProtBert (PB) بهره می برد. این

در معیار Precision عملکرد مناسبی دارند، کارایی بهتری در این مسئله از خود نشان می دهند.

### ۳-۴ ارزیابی عملکرد مدل PBS

برای ارزیابی مدل پیشنهادی، از روش ارزیابی متقابل ۵ برابری<sup>۲۳</sup> استفاده شده و مجموعه داده اولیه را گرفته می شود. در این مدل ارزیابی، مجموعه داده ها به ۵ دسته تقسیم می شوند. هر بار یکی از دسته ها به عنوان مجموعه تست در نظر گرفته شده و سایر داده ها به عنوان مجموعه آموزشی استفاده می شوند. مدل PBS در این ارزیابی به طور متوسط امتیاز ACC برابر با ۸۱/۴۱ درصد، امتیاز AUC-ROC برابر با ۸۷/۳۵ درصد، امتیاز AUC-PR برابر با ۸۷/۷۸ درصد، امتیاز F1 برابر با ۸۱/۵۸ درصد و امتیاز Precision برابر با ۸۰/۸۴ درصد کسب کرده است.

درحالی است که مدل MTT به عنوان یکی از مدل هایی که در این مطالعه مورد بررسی قرار می گیرد، از نمایش UniRep (UR) استفاده می کند. مدل UniRep مبتنی بر LSTM و مدل ProtBert براساس ترنسفورمرها عمل می کند. طبق مطالعه [۲۴]، تعبیه ایجاد شده توسط مدل ProtBert ویژگی های ساختار سوم را نیز درک می کند.

جدول ۲ عملکرد مدل PBS با ارزیابی متقابل ۵ برابری.

| امتیاز  | Precision | F1    | AUC-PR | AUC-ROC | ACC   |
|---------|-----------|-------|--------|---------|-------|
| فولد ۱  | ۸۰/۵۴     | ۸۱/۰۷ | ۸۷/۵۷  | ۸۷/۰۶   | ۸۰/۹۵ |
| فولد ۲  | ۸۱/۲۲     | ۸۱/۱۴ | ۸۶/۹۱  | ۸۶/۹۳   | ۸۱/۱۶ |
| فولد ۳  | ۸۱/۵۶     | ۸۲/۲۲ | ۸۹/۱۶  | ۸۷/۹۷   | ۸۲/۰۸ |
| فولد ۴  | ۷۹/۵۹     | ۸۱/۳۵ | ۸۷/۳۳  | ۸۶/۹۷   | ۸۰/۹۳ |
| فولد ۵  | ۸۱/۳۳     | ۸۲/۰۸ | ۸۷/۹۴  | ۸۷/۸۵   | ۸۱/۹۲ |
| میانگین | ۸۰/۸۴     | ۸۱/۵۸ | ۸۷/۷۸  | ۸۷/۳۵   | ۸۱/۴۱ |

**۳-۶ مقایسه عملکرد روش پیشنهادی و مدل‌های پیشین**  
در این مطالعه از مجموعه داده مقاله [۱۸] استفاده می‌کنیم. در مقاله [۱۸] دو مدل MTT و STT برای حل مسئله HVPP معرفی می‌شود که در مقدمه به آنها پرداختیم. در این زیربخش عملکرد مدل PBS را علاوه بر این دو مدل، با نتایج حاصل از مطالعات [۱۶، ۲۰ و ۲۱] و با استفاده از مجموعه داده آموزش و تست یکسان مقایسه می‌کنیم. نتایج حاصل از این مقایسه مطابق با جدول ۴ است. چنانچه در جدول ۴ مشخص است، مدل PBS با به‌کارگیری ProtBert برای به‌دست آوردن نمایش اولیه پروتئین‌ها، به کارگیری شبکه PBSNet برای پیش‌بینی برهم‌کنش‌ها و رای‌گیری برای تصمیم‌گیری روی مجموعه تست توانسته عملکرد بهتری نسبت به سایر مدل‌ها در معیارهای مختلف کسب کند.

برای ارزیابی اینکه کدام نمایش عملکرد مناسب‌تری دارد، به ازای مجموعه تست و آموزش یکسان، یکبار از نمایش UniRep و یکبار از نمایش PrtoBert برای ارزیابی چارچوب PBS استفاده می‌کنیم. نتایج حاصل از این مقایسه در جدول ۳ آمده است. بنابر نتایج حاصله، مدل در صورت استفاده از ProtBert برای مجموعه داده اولیه در حدود ۲۰ درصد Precision، ۳۰ درصد امتیاز F1، ۲۰ درصد امتیاز AUC-PR، ۱۷ درصد امتیاز AUC-ROC و ۱۵ درصد دقت بهتری نسبت به زمان استفاده از UniRep کسب می‌کند. به‌طور مشابه هنگام استفاده از مجموعه داده ثانویه و مدل ProtBert عملکرد مدل در هر یک از معیارهای ارزیابی به‌طور متوسط ۲۰ درصد بهتر از حالت UniRep است. این تفاوت عملکرد، نشان‌دهنده آن است که مدل ProtBert نمایش مناسب‌تری نسبت به UniRep ارائه می‌دهد.

جدول ۳ مقایسه عملکرد مدل PBS با نمایش UniProt و Protbert.

| داده                | مدل | Precision | F1    | AUC-PR | AUC-ROC | ACC   |
|---------------------|-----|-----------|-------|--------|---------|-------|
| مجموعه داده ابتدایی | UR  | ۷۱/۹۲     | ۳۳/۱۳ | ۶۴/۵۴  | ۶۳/۲۴   | ۵۶/۵۶ |
|                     | PB  | ۹۰/۲۴     | ۶۳/۱۴ | ۸۴/۰۶  | ۸۰/۶۹   | ۷۱/۶۵ |
| مجموعه داده ثانویه  | UR  | ۶۵/۶۰     | ۳۴/۰۹ | ۶۱/۲۷  | ۵۸/۴۹   | ۵۵/۴۸ |
|                     | PB  | ۸۸/۸۶     | ۶۸/۴۱ | ۸۳/۹۶  | ۸۰/۲۴   | ۷۴/۰۲ |

جدول ۴ مقایسه عملکرد چارچوب PBS با سایر مطالعات حوزه HVPP.

| مدل  | Precision | F1    | AUC-PR | AUC-ROC | ACC   |
|------|-----------|-------|--------|---------|-------|
| PBS  | ۹۰/۲۴     | ۶۳/۱۴ | ۸۴/۰۴  | ۸۰/۶۹   | ۷۱/۶۵ |
| MTT  | ۷۶/۶۸     | ۷۶/۶۸ | ۸۶/۶۷  | ۷۲/۵۵   | ۷۷/۶۸ |
| STT  | ۷۵/۵۹     | ۷۵/۵۹ | ۷۲/۶۸  | ۷۹/۲۵   | ۷۵/۵۹ |
| [۲۰] | ۷۰/۶۵     | ۷۰/۷۸ | ۸۰/۵۲  | ۷۶/۷۵   | ۷۰/۷۲ |
| [۱۶] | ۵۸/۹۲     | ۵۸/۹۲ | ۶۲/۵۹  | ۶۱/۸۵   | ۵۸/۹۳ |
| [۲۱] | ۶۷/۰۹     | ۶۷/۰۹ | ۷۹/۵۴  | ۷۴/۷۲   | ۶۷/۰۹ |

این برتری به دلیل در نظر گرفتن ویژگی های ساختار سوم و بهینه کردن نمایش اولیه با استفاده از شبکه های عصبی دوقلوی ناهمگون حاصل می شود. بر این اساس، مدل پیشنهادی توانسته است در معیار Precision به طور میانگین ۲۰ درصد و معیار AUC-PR را به طور متوسط ۱۰ درصد بهبود ببخشد.

### ۳-۷ نمونه مطالعاتی

یکی از مهم ترین قابلیت های مدل های آموزش دیده، استفاده از آنها در حل مسائل دنیای واقعی است. برای اینکه توانایی مدل PBS را در این شرایط بسنجیم، ویروس H1N1 را به عنوان نمونه مورد مطالعه انتخاب می کنیم و برهم کنش بین پروتئین های این ویروس و پروتئین های انسانی را به عنوان مجموعه تست قرار می دهیم. در این حالت مجموعه تست شامل ۴۰۰ پروتئین انسانی، ۲۵ پروتئین ویروس H1N1، ۵۵۱ برهم کنش شناخته شده و ۶۳۶ برهم کنش ناشناخته است. پس از آموزش مدل با

استفاده از تمام مجموعه داده، به غیر از داده تست، از آن برای پیش بینی برهم کنش های مجموعه تست استفاده می کنیم. امتیازهای حاصل از این آزمایش در جدول ۵ قابل مشاهده است. با توجه به معیارهای به دست آمده، به نظر می رسد مدل عملکرد مناسبی از خود نشان می دهد.

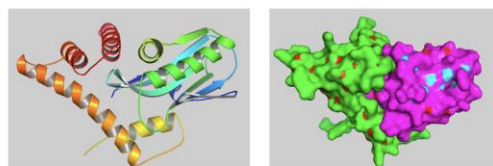
برای بررسی بیشتر کارایی مدل PBS روی مورد مطالعاتی انتخابی، تصمیم گرفتیم از ابزارهای مشاهده مجازی<sup>۲۴</sup> استفاده کنیم. برای این منظور ۴ جفت پروتئین ویروس-انسان را در نظر گرفتیم. مطابق پیشنهاد مدل این ۴ جفت پروتئین باید برهم کنش داشته باشند، اما در پایگاه داده این برهم کنش شناسایی نشده است. اطلاعات توالی این پروتئین ها را در ابزار HDock بارگذاری کردیم [۲۷]. ابزار HDock می تواند با دریافت ساختار توالی و براساس مدل سازی مبتنی بر همولوژی، ساختار سه بعدی آنها را براساس پایگاه داده بانک پروتئینی PDB بازسازی کند و در نهایت به صورت پویا، اطلاعات ۱۰۰ حالت برتر داک شدن این دو پروتئین را خروجی دهد.

جدول ۵ نتایج عملکرد مدل PBS برای پیش بینی برهم کنش بین پروتئین های ویروس و انسان.

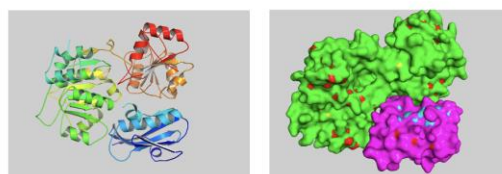
| داده | Precision | F1    | AUC-PR | AUC-ROC | ACC   |
|------|-----------|-------|--------|---------|-------|
| H1N1 | ۸۰/۴۳     | ۷۰/۴۴ | ۸۳/۴۳  | ۸۲/۹۰   | ۷۶/۷۷ |

جدول ۶ بررسی امتیاز داک شدن دو پروتئین انسانی و ویروس H1N1 که توسط PBS دارای احتمال بالای برهم کنش ثبت شده اند.

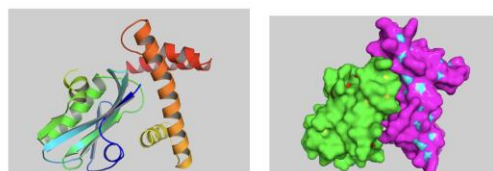
| امتیاز قابلیت اطمینان | میزان انرژی پیوند | احتمال وجود برهم کنش براساس مدل PBS | پروتئین ویروس H1N1 | پروتئین انسانی |
|-----------------------|-------------------|-------------------------------------|--------------------|----------------|
| ٪ ۹۰/۸۷               | -۲۶۴/۸۷           | ۹۷/۹۲                               | Q0HD54             | Q93079         |
| ٪ ۷۸/۷۱               | -۲۱۵/۳۷           | ۹۶/۸۹                               | Q82506             | P51991         |
| ٪ ۸۵/۶۴               | -۲۳۹/۳۰           | ۹۶/۳۰                               | Q0HD54             | Q9NR30         |
| ٪ ۸۲/۴۷               | -۲۲۷/۴۴           | ۹۴/۶۸                               | P03496             | Q99879         |



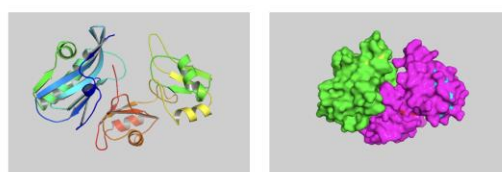
**Q0HD54-Q93079**



**Q0HD54-Q9NR30**



**P03496-Q99879**



**Q82506-P51991**

شکل ۲ تصاویر حاصل از داک کردن پروتئین‌های انسان و ویروس H1N1 که نتایج کمی آنها در جدول ۶ آمده است.

در این مطالعه به پیش‌بینی برهم‌کنش بین پروتئین‌های انسانی و پروتئین‌های ویروس پرداختیم. یکی از مهم‌ترین چالش‌هایی که مطالعات این حوزه با آن رو به رو هستند، نحوه نمایش پروتئین‌ها است. با توجه به اینکه برهم‌کنش پروتئین‌ها وابسته به ساختار سه بعدی آنها است و غالباً اطلاعات ساختار سوم برای تمام پروتئین‌ها موجود نیست، اکثر مطالعات از ساختار اول استفاده می‌کنند. برای حل این چالش ما از روش ProtBert استفاده کردیم که یک مدل مبتنی بر ترنسفورمرها است و ویژگی‌های ساختار سوم را نیز درک می‌کند. در ادامه برای مدل کردن مسئله شبکه عصبی دوقلوی ناهمگونی با نام PBSNet طراحی

برای هر جفت ورودی، امتیازهای حاصل از داک شدن در جدول ۶ قرار دارد. با توجه به امتیازهای به دست آمده، به نظر می‌رسد مجموعه برهم‌کنش‌های پیش‌بینی شده توسط مدل با احتمال بالایی امکان وجود برهم‌کنش بین آنها وجود دارد. بنابراین، می‌توان نتایج پیشنهادی مدل را در مسائل دنیای واقعی نیز در نظر گرفت. نتایج حاصل از داک کردن در شکل ۳ به طور خلاصه تصویر شده‌اند.

#### ۴- نتیجه‌گیری

- [4] G. Gerold, J. Bruening, B. Weigel, and T. Pietschmann, "Protein Interactions during the Flavivirus and Hepacivirus Life Cycle," *Mol Cell Proteomics*, vol. 16, no. 4 suppl 1, pp. S75–S91, Apr. 2017, doi: 10.1074/MCP.R116.065649.
- [5] S. Sadegh et al., "Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing," *Nat Commun*, vol. 11, no. 1, Dec. 2020, doi: 10.1038/S41467-020-17189-2.
- [6] S. Li et al., "Comprehensive characterization of human–virus protein-protein interactions reveals disease comorbidities and potential antiviral drugs," *Comput Struct Biotechnol J*, vol. 20, pp. 1244–1253, Jan. 2022, doi: 10.1016/J.CSBJ.2022.03.002.
- [7] N. Goodacre, P. Devkota, E. Bae, S. Wuchty, and P. Uetz, "Protein-protein interactions of human viruses," *Semin Cell Dev Biol*, vol. 99, pp. 31–39, Mar. 2020, doi: 10.1016/j.semcd.2018.07.018.
- [8] L. Young, R. L. Jernigan, and D. G. Covell, "A role for surface hydrophobicity in protein-protein recognition," *Protein Science*, vol. 3, no. 5, pp. 717–729, 1994, doi: 10.1002/PRO.5560030501/TITLE/A\_ROLE\_FOR\_SURFACE\_HYDROPHOBICITY\_IN\_PROTEIN\_PROTEIN\_RECOGNITION.
- [9] S. Jones and J. M. Thornton, "Analysis of protein-protein interaction sites using surface patches," *J Mol Biol*, vol. 272, no. 1, pp. 121–132, Sep. 1997, doi: 10.1006/jmbi.1997.1234.
- [10] F. Pazos, M. Helmer-Citterich, G. Ausiello, and A. Valencia, "Correlated mutations contain information about protein - protein interaction," *J Mol Biol*, vol. 271, no. 4, pp. 511–523, Aug. 1997, doi: 10.1006/jmbi.1997.1198.
- [11] X. Gallet, B. Charleaux, A. Thomas, and R. Brasseur, "A fast method to predict protein interaction sites from sequences," *J Mol Biol*, vol. 302, no. 4, pp. 917–926, Sep. 2000, doi: 10.1006/JMBI.2000.4092.
- [12] H. M. Chen, J. X. Liu, D. Liu, G. F. Hao, and G. F. Yang, "Human-virus protein-protein interactions maps assist in revealing the pathogenesis of viral infection," *Rev Med Virol*, vol. 34, no. 1, p. e2517, Jan. 2024, doi: 10.1002/RMV.2517.
- [13] S. S. Raj and S. S. V. Chandra, "Significance of Sequence Features in

کردیم. شبکه‌های عصبی دوقلو قابلیت ایجاد فضای ویژگی مشترک بین دو ورودی را دارند و الگوریتم‌های مربوط به وجود رابطه را به خوبی درک می‌کند. با توجه به اینکه طبیعت ورودی‌های مسئله وابسته به دو نوع متفاوت انسان و ویروس بود، شبکه‌های دوقلو ناهمگون را به کار گرفتیم. پس از بررسی و مقایسه مدل با سایر مطالعات، مدل پیشنهادی با امتیاز ACC برابر با ۸۱/۴۱ درصد، امتیاز AUC-ROC برابر با ۸۷/۳۵ درصد، امتیاز AUC-PR برابر با ۸۷/۷۸ درصد، امتیاز F1 برابر با ۸۱/۵۸ درصد و امتیاز Precision برابر با ۸۰/۸۴ درصد در ارزیابی ۵ برابری متقابل عملکرد بهتری نسبت به سایر مدل‌ها دارد. همچنین، برای ارزیابی مدل در رویارویی با مسائل دنیای واقعی به‌طور خاص روی برهم‌کنش‌های بین پروتئین‌های ویروس H1N1 و انسان مطالعه کردیم.

با وجود اینکه مدل PBS با استفاده از ProtBert ویژگی‌های ساختار سوم را در نظر می‌گیرد، همچنان از ساختارهای سوم پروتئین بهره نمی‌برد. با توجه به پیشرفت پایگاه داده‌ها، ساختار سوم‌های بیشتری از ویروس‌ها موجود هستند که می‌توانند در حل این مسئله مفید باشند. علاوه بر این استفاده از خواص بیوفیزیکی و بیوشیمیایی پروتئین‌ها نیز ممکن است عملکرد مدل‌ها را بهبود بخشد.

## ۵- منابع

- [1] E. Petersen et al., "Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics," *Lancet Infect Dis*, vol. 20, no. 9, pp. e238–e244, Sep. 2020, doi: 10.1016/S1473-3099(20)30484-9.
- [2] G. A. Smith and L. W. Enquist, "Break ins and break outs: viral interactions with the cytoskeleton of Mammalian cells," *Annu Rev Cell Dev Biol*, vol. 18, pp. 135–161, 2002, doi: 10.1146/ANNUREV.CELLBIO.18.012502.105920.
- [3] P. M. Jean Beltran, K. C. Cook, and I. M. Cristea, "Exploring and Exploiting Proteome Organization during Viral Infection," *J Virol*, vol. 91, no. 18, Sep. 2017, doi: 10.1128/JVI.00268-17.

- [20] X. Yang, S. Yang, Q. Li, S. Wuchty, and Z. Zhang, "Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method," *Comput Struct Biotechnol J*, vol. 18, pp. 153–161, Jan. 2019, doi: 10.1016/J.CSBJ.2019.12.005.
- [21] X. Zhou, B. Park, D. Choi, and K. Han, "A generalized approach to predicting protein-protein interactions between virus and host," *BMC Genomics*, vol. 19, no. 6, pp. 69–77, Aug. 2018, doi: 10.1186/S12864-018-4924-2/TABLES/8.
- [22] X. Yang et al., "Multi-modal features-based human-herpesvirus protein-protein interaction prediction by using LightGBM," *Brief Bioinform*, vol. 25, no. 2, Jan. 2024, doi: 10.1093/BIB/BBAE005.
- [23] A. Bateman et al., "UniProt: the universal protein knowledgebase in 2021," *Nucleic Acids Res*, vol. 49, no. D1, pp. D480–D489, Jan. 2021, doi: 10.1093/NAR/GKAA1100.
- [24] A. Behjati, F. Zare-Mirakabad, S. S. Arab, and A. Nowzari-Dalini, "Protein sequence profile prediction using ProtAlbert transformer," *Comput Biol Chem*, vol. 99, Aug. 2022, doi: 10.1016/j.compbiolchem.2022.107717.
- [25] Z. Ghorbanali, F. Zare-Mirakabad, N. Salehi, M. Akbari, and A. Masoudi-Nejad, "DrugRep-HeSiaGraph: when heterogenous siamese neural network meets knowledge graphs for drug repurposing," *BMC Bioinformatics* 2023 24:1, vol. 24, no. 1, pp. 1–31, Oct. 2023, doi: 10.1186/S12859-023-05479-7.
- [26] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, Nov. 2016, doi: 10.1007/S13748-016-0094-0/TABLES/1.
- Classification of Protein-Protein Interactions Using Machine Learning," *Protein Journal*, vol. 43, no. 1, pp. 72–83, Feb. 2024, doi: 10.1007/S10930-023-10168-8/METRICS.
- [14] B. Y. S. Li, L. F. Yeung, and G. Yang, "Pathogen host interaction prediction via matrix factorization," *Proceedings - 2014 IEEE International Conference on Bioinformatics and Biomedicine, IEEE BIBM 2014*, pp. 357–362, Dec. 2014, doi: 10.1109/BIBM.2014.6999185.
- [15] E. Nourani, F. Khunjush, and S. Durmuş, "Computational prediction of virus-human protein-protein interactions using embedding kernelized heterogeneous data," *Mol Biosyst*, vol. 12, no. 6, pp. 1976–1986, May 2016, doi: 10.1039/C6MB00065G.
- [16] F. E. Eid, M. Elhefnawi, and L. S. Heath, "DeNovo: virus-host sequence-based protein-protein interaction prediction," *Bioinformatics*, vol. 32, no. 8, pp. 1144–1150, Apr. 2016, doi: 10.1093/BIOINFORMATICS/BTV737.
- [17] A. H. Basit, W. A. Abbasi, A. Asif, S. Gull, and F. U. A. A. Minhas, "Training host-pathogen protein-protein interaction predictors," *J Bioinform Comput Biol*, vol. 16, no. 4, Aug. 2018, doi: 10.1142/S0219720018500142.
- [18] T. N. Dong, G. Brogden, G. Gerold, and M. Khosla, "A multitask transfer learning framework for the prediction of virus-human protein-protein interactions," *BMC Bioinformatics*, vol. 22, no. 1, Dec. 2021, doi: 10.1186/S12859-021-04484-Y.
- [19] W. Liu-Wei, Ş. Kafkas, J. Chen, N. J. Dimonaco, J. Tegnér, and R. Hoehndorf, "DeepViral: prediction of novel virus-host interactions from protein sequences and infectious disease phenotypes," *Bioinformatics*, vol. 37, no. 17, pp. 2722–2729, Sep. 2021, doi: 10.1093/BIOINFORMATICS/BTAB147.

# Prediction of virus-human protein-protein interactions using heterogeneous siamese neural networks

Sara Mohammadzadeh<sup>1</sup>, Zahra Ghorbanali<sup>2</sup>, Fatemeh Zare-Mirakabad<sup>2,\*</sup>

1. Bachelor's degree in Computer Science, Computational Biology Research Center (CBRC), department of Mathematics and Computer science, Amirkabir University of Technology, Tehran, Iran.

2. Ph.D. in Computer Science, Computational Biology Research Center (CBRC), department of Mathematics and Computer science, Amirkabir University of Technology, Tehran, Iran.

f.zare@aut.ac.ir

Receipt: 2023/12/06

Accepted: 2025/02/22

## Abstract

Viral infections represent pathological conditions arising from the intrusion of viruses into host cells and their replication. The onset of infection is intricately tied to the interplay between viral and host cell proteins. Thus, elucidating these protein-protein interactions plays a pivotal role in the encompassing prevention, treatment, and control of viral infections. Given traditional laboratory experimentation's prohibitively high costs and time-intensive nature, researchers have increasingly turned to computational approaches for predicting human-virus protein-protein interactions. Despite the performance of these computational approaches, a challenge persists in the need for an effective protein representation that adequately captures their structural intricacies. In this paper, we present PBS, a novel model for the prediction of protein-protein interactions between viruses and humans. PBS leverages the transformers to effectively represent proteins. The model unified the latent space for human and virus proteins through the implementation of heterogeneous siamese neural networks. The model achieves an accuracy score of 81.41%, an area under the ROC curve score of 87.35%, an area under the precision-recall curve score of 87.78%, an F1 score of 81.58%, and a precision score of 80.84%. These metrics collectively underscore the satisfactory performance of the PBS model. Furthermore, we assess the model's predictive capabilities in discerning interactions between proteins associated with the H1N1 influenza virus and human proteins.

**Keywords:** protein-protein interactions, transformer, ProtBert, siamese neural networks.