

[Artificial Intelligence in the Life Sciences 2 (2022) 10003](https://doi.org/10.1016/j.ailsci.2022.100037)7

|  |  |  |
| --- | --- | --- |
|  | Contentslistsavailableat[ScienceDirect](http://www.ScienceDirect.com) |  |
| ArtificialIntelligenceintheLifeSciences |
| journalhomepage:[www.elsevier.com/locate/ailsci](http://www.elsevier.com/locate/ailsci) |

Viewpoint   
Deeplearningofprotein–ligandinteractions—Rememberingtheactors JürgenBajorath   
*DepartmentofLifeScienceInformaticsandDataScience,B-IT,LIMESProgramUnitChemicalBiologyandMedicinalChemistry,Rheinische*   
*Friedrich-Wilhelms-Universität,Friedrich-Hirzebruch-Allee5/6,BonnD-53115,Germany*

Oneoftheintenselyinvestigatedapplicationsofdeeplearningin drugdesignisthepredictionofcompoundpotency(affinity)basedupon three-dimensionalstructuresofprotein–ligandcomplexes.Consistently accurateligandbindingaffinitypredictionswouldrepresentamilestone eventforthefieldandputstructure-basedliganddesignonanewlevel. Forthispurpose,convolutionalneuralnetworks(CNNs)[1]withvoxel representationsofligandbindingsitesaswellasgraphneuralnetworks (GNNs)[2]includingmessagepassingneuralnetworks(MPNNs)[3]are applied.GNNs/MPNNslearndirectlyfrommoleculargraphs.Ingeneral, MPNNsarebecomingincreasinglypopularforrepresentationlearningin chemistry.ForaffinitypredictionsusingGNNs/MPNNs,protein–ligand complexstructuresaretranslatedintointeractiongraphs.

Employingtheseneuralnetworkarchitectures,avarietyofaffin-itypredictionmodelsbasedonprotein–ligandcomplexstructureshave beenreportedinrecentyears(foranup-to-datesummarysee,forex-ample,[4]).Graph-basedmodelsoftenachievehighcorrelation(atthe 80%level)betweenpredictedandexperimentallyobservedligandbind-ingaffinitiesandpredictionaccuracywithinorclosetoanorderof magnitude(10-fold).Theseobservationshavetriggeredrecurrentas-sumptionsorclaimsthatdeepneuralnetworksarecapableoflearn-ingspecificprotein–ligandinteractions.However,thiswouldalsoim-plythattheresultingmodelswouldcapture,inonewayoranother, thephysico-chemicalfoundationsoftheseinteractions.Mightthisbe conceivablebylearningfrommolecularinteractiongraphs?Regardless ofsuchprincipalconsiderations,severalobservationssuggestthatthe promisingresultsreportedforvariousprotein–ligandaffinityprediction modelsshouldbeconsideredwithcaution.

Anessentialresourceforprotein–ligandcomplexstructureswith availableexperimentalaffinitymeasurementsisthePDBbinddatabase [5]thatprovidesthebasisformanyinvestigations.However,thenum-berofhigh-resolutionstructureswithhigh-qualityaffinitymeasure-mentsislimitedandthecompositionofPDBbindisbiasedtowardspre-ferredcrystallographictargets[6,7],whichnaturallylimitsthegener-alizationabilityofpredictivemodelsderivedfromthesedata.Thisis consistentwiththefindingthatdifferenttrainingandtestdataparti-tionscansignificantlyinfluencemodelperformance[6].Ontheother hand,trainingsetsofvaryingsizeoftenyieldsimilarlyaccurateprotein–ligandinteractionmodels[4],whichiscounterintuitivefordeeplearn-ing.IthasalsobeenobservedthatCNNmodelstrainedonlyonprotein orligandrepresentationscanapproachormeettheaccuracyofmodels trainedonprotein–ligandinteractiondata[6–9].Takentogether,these

*E-mailaddress:*[bajorath@bit.uni-bonn.de](mailto:bajorath@bit.uni-bonn.de)

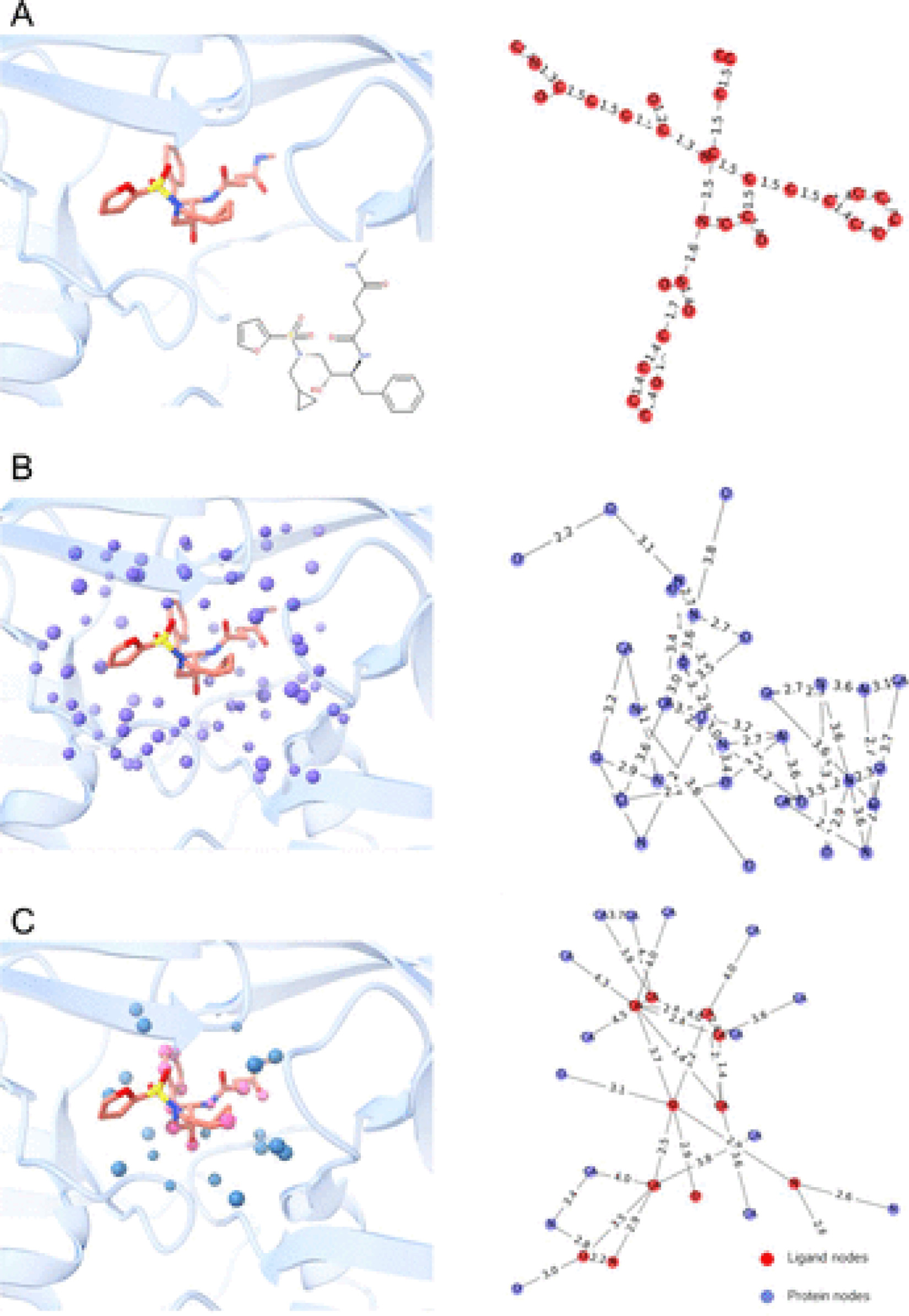
findingsindicatethatCNNandGNNaffinitypredictionmodelsmight primarilymemorizetrainingdatainformationfromcomplexstructures, ratherthanlearnspecificprotein–ligandinteractions.Similarmemo-rizationeffectshavebeenobservedfordeepcompoundclassification modelsdisplayinglimitedgeneralizationability[10].

Themostrecentscientificallyrigorousinvestigationofprotein–ligandinteractionmodelshasemployedanMPNNarchitecturetolearn fromdifferentgraphrepresentationsofprotein–ligandcomplexesfrom PDBbind[4].Foragivencomplexstructure,Rognanandcolleaguesgen-eratedgraphrepresentationsoftheligand(L),protein(P),andprotein–ligandinteractions(I)usingdifferentlydefinednodesandedges,asillus-tratedinFig.1[4].Forexample,proteinbindingsiteresidueswererep-resentedasnodesannotatedwithinteraction-relevantchemicalprop-ertyinformation.Furthermore,ininteractiongraphs,differentnodes representinginteractionsitesformedbyligandatomsandbindingsite residues,respectively,werecombinedandedgesaccountedfornon-covalentshort-rangeinteractions(annotatedwiththeirdistances). Fromthethreegraphrepresentations,seventrainingconstellations weregeneratedincludingtheindividualgraphs(L,P,I),threepair-wisecombinations(PL,PI,LI),andthetriplet(PLI)combiningallthree graphs.ThisframeworkwasusedtoderivedifferentMPNNmodels, whichwerethenappliedtopredicttheligandaffinitiesofcomplexstruc-turesfromdifferenttestsets.Predictiveperformancewasquantifiedby calculatingPearson’scorrelationcoefficient(RP)forpredictedandex-perimentalaffinitiesaswellastherootmeansquareerror(RMSE). Themodelsweregenerallyfoundtobepredictive,withRPvaluesof above0.6andmaximally∼0.8.Importantly,modelsbasedonlyonthe LorPgraphsweremoreaccuratethanthemodeltrainedontheinter-actiongraph(Imodel),withtheligand-basedmodelperformingbest. TheaccuracywasfurtherincreasedbythePLmodel,theperformance ofwhichwasverysimilartothePLImodeltrainedonallthreegraphs. Thesefindingsconclusivelydemonstratedthataffinitypredictionsusing MPNNsdidnotdependonlearningspecificprotein–ligandinteractions, letalonetheunderlyingphysics.

Theauthorsfurtherextendedtheiranalysis.Forexample,theMPNN modelswerefoundtodisplaylimitedgeneralizationpotential,butwere insensitivetoreductionoftrainingsetsize.Inaddition,simple“memo-rizationbaseline” modelsweregeneratedpredictingtheaffinityofcom-plexesbasedonaveragesforthemostsimilarligandsorproteinsfrom thetrainingset.Interestingly,theligandbaselinemodelnearlyreached theaccuracyoftheImodel.Takentogether,theseresultsshowedthat

<https://doi.org/10.1016/j.ailsci.2022.100037>  
[Received25May2022;Accepted25May202](https://doi.org/10.1016/j.ailsci.2022.100037)2   
Availableonline26May2022   
[2667-3185/© 2022TheAuthors.PublishedbyElsevi](http://creativecommons.org/licenses/by-nc-nd/4.0/)erB.V.ThisisanopenaccessarticleundertheCCBY-NC-NDlicense (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

*J.Bajorath*  *ArtificialIntelligenceintheLifeSciences2(2022)100037*



**Fig.1.Graphrepresentationsofproteins,ligands,andtheirinteractions**.Basedonthestructureofaprotein–ligandcomplex,graphrepresentationsencodethe (**A**)ligand(rednodes,edges),(**B**)protein’sligandbindingsite(bluenodes,edges),and(**C**)protein–ligandinteractions(blue/rednodes,edges).Forfurtherdetails, seeVolkovetal.[4].ThefigurewasreprintedwithpermissionfromVolkov,M.;Turk,J.A.;Drizard,N.;Martin,N.;Hoffmann,B.;Gaston-Mathé;Rognan,D.Onthe frustrationtopredictbindingaffinitiesfromprotein-ligandstructureswithdeepneuralnetworks.JMedChem2022.doi:10.1021/acs.jmedchem.2c00487.inpress. Copyright2022AmericanChemicalSociety.

2

*J.Bajorath*  *ArtificialIntelligenceintheLifeSciences2(2022)100037*

theaffinitypredictionsweremostlydrivenbymemorizingpatternsfrom trainingdata,withligandsimilarityrelationshipsplayingamajorrole. Theauthorsalsoshowedthatincreasingthecomplexityofinteraction graphsbyincludingmoreinteractionsoverlongerdistancesfurtherin-creasedthepredictionaccuracyoftheImodel.Thisobservationled totheconclusionthataninteractionmodeldeprivedofadditionalpro-teinorligandcontextinformationshouldprovideareasonablebasisfor furtherexploringtheabilityofdeepneuralnetworkstolearnprotein–ligandinteractions.Likeothersbefore,theauthorsalsoemphasizedthat thecurrentsparsityofhigh-qualityanddiverseprotein–ligandcom-plexdatarepresentsamajorlimitationforthefurtherdevelopmentof protein–ligandaffinitymodels.

Forcompoundpotencypredictionsandstructure-baseddrugdesign, insightsprovidedbycarefulstudiesliketheonebyRognanandcol-leaguesareoffundamentalrelevance,puttingputativemethodological advancesintoscientificperspective,raisingawarenessofpotentialover-interpretation,andbalancingexpectations.Moreover,fordeeplearning acrossthelifesciences,investigationsdemonstratinglimitationsofcur-rentapproachesandpotentialcaveatsormisinterpretationofresultsare asimportantforthefurtherdevelopmentofthefieldasmethodological breakthroughs.

**DeclarationofCompetingInterest**

Theauthorsdeclarethattheyhavenoknowncompetingfinancial interestsorpersonalrelationshipsthatcouldhaveappearedtoinfluence theworkreportedinthispaper.

3