

[Artificia](https://doi.org/10.1016/j.ailsci.2023.100079)l[IntelligenceintheLifeSciences4(2023)100079](https://doi.org/10.1016/j.ailsci.2023.100079)

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

Anindustrialevaluationofproteochemometricmodelling:Predicting drug-targetaffinitiesforkinases   
AstridStroobantsa,∗,LewisH.Mervinb,OlaEngkvistc,GraemeR.Robba   
a*Chemistry,Oncology,R&D,AstraZeneca,Cambridge,UK*   
b*MolecularAI,DiscoverySciences,R&D,AstraZeneca,Cambridge,UK*   
c*MolecularAI,DiscoverySciences,R&D,AstraZeneca,Gothenburg,Sweden*

|  |  |  |
| --- | --- | --- |
| article | info | abstract |
| *Keywords:*  Cheminformatics  Machine-learning  Virtualscreening  Kinaseselectivity  Proteochemometricmodelling PCM  DTA  Drugtargetaffinity | | Deeplearningproteochemometric(PCM)modelshavebeenreportedtoachieveexcellentperformancesonpublic benchmarkingdatasets.Nevertheless,numerouspapershavecastdoubtoncommonlyusedevaluationmetrics, suggestingtheydonotreflecttrueprospectivepredictiveabilities.Theaimofthisstudyistoprovideacompre-hensiveassessmentofperformanceofastate-of-the-artPCMmodelonproprietarydataandevaluateitspotential overothermodellingapproachesasavirtualscreeningtoolforkinaseinhibitors.Whilstthemodelhasbeen showntoachieveanRMSEof0.48onapublicbenchmarkingdataset,animpairedoverallperformancewas observedfortheproprietarydatasetinthisstudy,withanRMSEof0.85andaPearsonCorrelationCoefficient of0.65usingatemporalsplittingstrategy.Wehypothesisethatthemorelimitedperformancecanbeinpart attributedtoashiftinthechemicalspaceobservedovertimeinanindustrialsetting,whichisnotconsidered bythemorelenientrandomligandsplittingstrategy,morecommonlyusedonbenchmarkingdatasets.Theover-allperformanceofthePCMmodelwasstatisticallysimilartoamultitaskmodelandonlyslightlysuperiorto aKNNandrandomforestPCMmodel.Acomprehensiveanalysisofperformancewasperformedtocapturethe keychallengesfacedinthedesignofcompetitivekinaseinhibitors,whichrevealedthekeylimitationsofPCM modelling.Forexample,themodelshowedpoorpredictiveabilitiesforunderstudiedtargets,andalimitedability toassessligandselectivityandpromiscuity,withnoimprovedperformanceoveramultitaskmodelorarandom forestPCMmodel.Overall,thesefindingsrevealthatthePCMmodelassessedinthisstudydoesnotprovide significantbenefitsoverlesscomplexmodelssuchasmultitaskmodelorarandomforestPCMmodelasavirtual screeningtoolforkinaseinhibitorsinanindustrialsetting.Takentogether,thisstudyhighlightstheneedfor morerobustevaluationsofPCMmodelsbyusingstrictersplittingstrategies,moreextensivebenchmarkingand morecomprehensiveperformanceanalysisbeyondtraditionalmetrics. |

**Introduction**

Theapplicationofartificialintelligence(AI)todrugdiscoveryhas beenanincreasinglydynamicfield[1–3].Theneedtoacceleratethe drugcandidateidentificationprocesswitheffectivehigh-throughputvir-tualscreeningkeepsinspiringnumerousAIfocusedpublications[4]. Forthepredictionofdrug-targetaffinities(DTA),proteochemometric (PCM)modellinghasemergedasacomprehensiveandversatilesolu-tion[5],withnewapproachesbeingpublishedregularly.Inthisfield, thefirstmachinelearning(ML)solutionsconsistedofsingletaskquanti-tativestructureactivityrelationship(QSAR)models(includingrandom forests(RF),supportvectormachineandlogisticregression)thatrelied

onrule-basedmethodssuchasafingerprintrepresentation[6–8].How-ever,multi-targetdeeplearning(DL)approachessuchasmultitaskmod-elswereshowntooutperformthesesingletargetpredictionmethodsby leveraginglargerheterogeneousdatasourcesandcross-targetinforma-tion[9,10].PCMmodels,incontrasttomultitaskmodels,considernot justchemicalbutalsoproteininformationandcanthus,inprinciple, betterleveragecross-targetinformationtosimultaneouslygeneralizeto novelligandsandtargets.Withtheriseofdeeplearning(DL),more complexarchitecturesandrepresentationscouldbeexplored,includ-ingbutnotlimitedtolanguagerepresentations(e.g.SMILES,amino acidsequences)[11],graphs[12–14]andvoxels[15].Onbenchmarking datasets,DLPCMmodelshaveachievedexcellentperformancesontra-

*Abbreviations:*AI,artificialintelligence;DL,deeplearning;DTA,drugtargetaffinity;KNN,k-nearestneighbours;MAP,maximumachievableperformance;ML, machinelearning;MSE,meansquareerror;PCC,Pearsoncorrelationcoefficient;PCM,proteochemometric;QSAR,quantitativestructure-activityrelationships;RF, randomforest;RMSE,root-mean-squareerror;SMILES,simplifiedmolecular-inputline-entrysystem.

∗Corresponding[author.](mailto:astrid.stroobants@astrazeneca.com)

*E-mailaddress:*[astrid.stroobants@astrazeneca.com](mailto:astrid.stroobants@astrazeneca.com)(A.Stroobants).

<https://doi.org/10.1016/j.ailsci.2023.100079>  
[Received24January2023;Receivedinrevise](https://doi.org/10.1016/j.ailsci.2023.100079)dform18May2023;Accepted23May2023   
Availableonline15June2023   
[2667-3185/© 2023TheAuthors.PublishedbyElsevi](http://creativecommons.org/licenses/by-nc-nd/4.0/)erB.V.ThisisanopenaccessarticleundertheCCBY-NC-NDlicense (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

*A.Stroobants,L.H.Mervin,O.Engkvistetal.*

ditionalevaluationmetrics.Meansquareerrors(MSE)between0.2–0.3 havebeenrepeatedlyreportedontheDavisdataset[11,12,14].Never-theless,severalpapershavenowcastdoubtonthetraditionalevaluation metrics,suggestingthatregardingdrug-targetinteractionsasindepen-dententitiescanleadtoredundanciesbetweentrainingandvalidation andthustooverfitting,resultinginoverlyoptimisticresultsthatdonot reflectthemorelimitedprospectivepredictiveabilityofthemodels[16–18].Therefore,withthisstudy,weaimtoprovideamorecomprehensive assessmentofthepredictiveabilitiesofPCMmodellingonproprietary datatoassessitspotentialasavirtualscreeningtoolinanindustryset-tingoverothermodellingstrategies.

Wefocusonastate-of-the-artPCMmodel,GraphDTA[12],whichis abimodalneuralnetworkbasedonconvolution,reliantonagraphrep-resentationoftheligandandanaminoacidsequencerepresentationof theprotein.ThemodelwassubmittedtotheIDG-DREAMDrug-Kinase BindingPredictionChallenge[19],wherethemodelwasamongtheten bestperformingsolutions.Sincethen,ithasoftenbeenusedasabench-markforotherpublications.Inthisstudyweaimtofurtherevaluatethe bioactivitypredictionperformanceofGraphDTAontheproteinkinase familyinanindustrialsetting.

Proteinkinasesareanimportantyetchallengingdrugtargetclass. Theyareinterestingdrugtargetsintreatmentareassuchasoncology, neurodegenerativeandviraldiseases.Theseenzymesmodulatevarious cellular,metabolic,andsignallingpathwaysthroughphosphorylation. Abnormalitiesintheirregulationcanleadtovariousdiseases,including cancer,diabetes,andinflammatorydisorders.Thefirstkinaseinhibitor, Imatinib,receivedFDAapprovalover20yearsagoandsincethenmore than70newdrugshavebeenapproved[20–22].Mostinhibitorstarget theATPbindingsite,whichisveryconservedamongstthedifferentki-nasesubfamilies.Thesimilarityinbindingsiteleadstoselectivitychal-lengesandselectivityandresistanceareseenaskeychallengesinthe developmentofeffectivekinaseinhibitors[20].Therefore,successful high-throughputvirtualscreeningsolutionsforproteinkinasesshould enablethegeneralisationtonewtargets(e.g.proteinmutations)and compounds,aswellastheidentificationofselectivekinaseinhibitors. Thesecriteria,oftenunexploredintraditionalevaluationsofPCMmod-els,areassessedinthisstudytoformamorecomprehensiveperfor-manceevaluationforindustrialapplications.

PublishedPCMevaluationsoftenrelyonvariouspublicbenchmark-ingdatasets.Withincreasingdataavailability,focushasshiftedonthe impactofdataqualityonperformance,consideringchallengeslikeex-perimentalerrorinbioaffinityvalues,qualityofthedatasource,sparsity ofthedrug-targetmatrix,anddatabiases[23–27].Publicdatasetshave adiversequalityrangeandaresubjecttoexperimentalerror[28,29]. Somedatabases,suchasChEMBL[30]andBindingDB[31],aggregate datafromdifferentsourceswithaffinitydataforseveralproteintar-gets.Theheterogenoususeofpublicbiochemicaldatahasbeenshown toimpactdataquality,forexample,ChEMBLversion27wasshownto haveamedianpIC50standarddeviationof∼0.37acrossreplicates[32]. Effortshavebeenmadetofacilitateandautomatethecurationandhan-dlingoftheselargedatasetsinordertotacklesomekeyissuessuchas lowqualitydataentries,reflectingout-of-distributionlearningthrough datasplitting,etc.[33,34].Proprietarydatasetshaveopportunitiesfor standardisation,nevertheless,anAstraZenecastudyevaluatingbiolog-icalassayvariabilityfromexperimentsfrom2005to2014stillhigh-lightedatwo-folddifferenceinexperimentalreproducibility,withIC50 measurementshavinglowerstandarddeviationscomparedtoKDand Kimeasurements[35].Modelstrainedonpublicdatasetsoftenpredict ligandaffinityacrossmultipleproteinfamilies,whilemodelsthatare evaluatedonsmallermorefocuseddatasets,suchastheDavis[36]and Kiba[29]datasets,oftenfocusondistinguishingaffinityscoresbetween closelyrelatedtargetsofthesameproteinfamily[6,37,38].Despitea potentialdecreaseinsparsityofthedrug-targetmatrix,thisisperhaps amorechallengingtaskasthemodelsneedtolearnmoresubtlediffer-encesbetweenproteinsandligandswithsmallerdatasources.Databi-aseshavebeenshowntogovernbenchmarkingdatasetsandhavemade

2

*A.Stroobants,L.H.Mervin,O.Engkvistetal.*

*Models*

ToevaluatePCMmodellinginanindustrialsetting,thestate-of-the-artGraphDTAmodelwasusedasareference.GraphDTAisabimodal neuralnetworkbasedonconvolutionalneuralnetworks.Inthisstudy, thegraphisomorphismnetwork(GIN)versionofGraphDTAwasap-plied.Forfurtherdetailspleaserefertotheoriginalpaper[12].The modelwasoptimizedonaMSElosswithAdamandwastrainedfor200 epochswithalearningrateof0.0005andabatchsizeof128.

TheperformanceoftheDLPCMmodelwascomparedtoseveralal-gorithmicbaselines:aKNN,amultitask(ChemProp[18])modelandtwo PCMmodelsusingRFregressionwithdifferentproteinrepresentations: onewithonehotencodingoftheprotein(PCM-OHE)andonewithZ-scales(PCM-Zscale).TheRFmodelsweregeneratedusingScikit-learn [40]andtrainedonthedescriptorsdescribedinthenextsection.The KNNmodelusedinthisstudyisbasedontheKNNinBornetal.[41]*.* TheKNNmodelreliedonthecombinationoftheTanimotosimilarity betweenligandsandtheLevenshteindistancebetweenaminoacidse-quencesoftheproteinsasadistancemetricbetweensamples,forfurther detailsrefertotheoriginalpublication[41].TheChemPropmultitask modelwastrainedusingdefaultsettingsand50epochs.

*Descriptors*

IntheoriginalGraphDTApublication,theligandsaredescribedas graphswhiletheproteinsweredescribedbytheirfullproteinamino acidsequence.Theligandrepresentationwasnotmodified,forfurther detailsrefertotheoriginalpublication[12].Sincethestandardpractice ofleveragingthefullprimarystructurehasbeenchallengedsincethe publicationofGraphDTA(betterperformancewasshownforPCMmod-elswhenfocusingontherepresentationoftheATPbindingsite[41]), thefollowingstrategywasadoptedinthisstudy:GraphDTAwasadapted toonlyconsiderthesequenceofkeyaminoacidsintheATPbindingsite [42].Inthisstudythe85keyresiduesfromthekinase–ligandinterac-tionfingerprintsandstructuredatabase(KLIFS)wereused.Thisnum-beringschemewaspreviouslydevelopedtocapturethecatalyticcleft with85residuestofacilitatethecomparisonoftheinteractionpatterns ofkinase-inhibitors,andthustoidentifydriversofkinase-inhibitorse-lectivity.The85KLIFSaminoacidswereidentifiedbasedonapublished structurallyvalidatedmultiplesequencealignmentof497humanpro-teinkinasedomainsfromModiandDunbrack[43].Allbenchmarking modelsalsoreliedonthereducedKLIFSrepresentationofthetargets. ThePCM-ZscaleRFmodelusedaZ-scale5descriptionofthe85pro-teinaminoacids.Z-scalesareprincipalpropertiesderivedfromprin-cipalcomponentanalysisofphysiochemicalpropertiesofaminoacids [44].ThecompounddescriptionforthePCMRFmodelsuseda2048-bithashedbinaryMorganfingerprintsofradius3generatedfromthe PythonRDKitmodule[45].The2048-bitencodingsizewaschosenaf-terevaluatingtheperformanceofencodingsizes512,1024and2048 (seeFig.S1fordetails).

*Splittingstrategy*

Threesplittingstrategieswereconsideredinthisstudy:arandomlig-andsplit,akinasesplitandatemporalsplit.Theligandsplitconsistsofa ten-foldcross-validationwheretheligandsaresplitbetweenfoldswith-outstratificationfortheprotein.Instead,thesplitwasstratifiedbythe meanpIC50perligand.Thissplittingstrategyremainsamongthemost commonsplitsadoptedonbenchmarkingdatasets.Thekinasesplitcon-sistsofafive-foldcross-validationwherethekinasesaresplitbetween foldswithstratificationbythemeanpIC50perkinase.Wealsoadopted atemporalsplitwhichconsistsofaneight-foldcross-validation,further detailsareprovidedinFig.1.Atemporalsplitisacommonstrategyused inindustrytoestimatetheprospectivepredictiveability.Thecompound registrationdatewasusedasatimestampandatwo-yeartimebinwas selected,roughlyequivalenttoaprojectduration.Thetrainingdatain

3

*A.Stroobants,L.H.Mervin,O.Engkvistetal.*

testsets.Foreachkinase,thepotencyvalueofeachcompoundwasran-domlyassignedtoanother,thusgeneratingrandomstructure–potency relationshipspertargetfortraining.Modelsweretrainedontheseran-dompermutationsandwereappliedtopredictarandomizedtestset. Thisreferencewasappliedasapreviousstudyhighlightedanunreal-isticallylimitedmarginbetweenrandomandbestpredictionmodels [46].Statisticalsignificanceassessmentoftherelativeperformanceof allmodelscomparedtotheMKRwasachievedusingthenon-parametric Wilcoxontest[47].Statisticalsignificancebetweentheperformance ofGraphDTAandothermodellingapproacheswasalsoassessedwith thismethod.Thealphathresholdwassetto0.05.The*p*-valueswere comparedwithalpha(*p*below0.05)andthenullhypothesiswasre-jected/accepted.

Inafollowingsectionweassessedtheperformanceforthetem-poralsplitonacommunaltestset(dataobtainedbetweenthedate range2019–2020).Thisevaluationaimedtohighlighthowmodelper-formancevariesduetotheprogressiveaccumulationoftrainingdata andtheexpansionofthechemicalspaceexploredintraining.

Next,theoverallcross-validationandtestperformancewasalsoeval-uatedforasubsetofkinasestoassesstheabilityofthemodelstopredict ligandaffinityforunderstudiedkinases.Wefocusedonthe18under-studiedkinaseswhichhavelessthan1000datapointsinthetrainingset inthelasttemporalfold(Fold8).

Theperformancewasalsoevaluatedonaperligandbasisbycalcu-latingtheconcordanceindexforeachligandandthemodelsensitivity andspecificity;therebyassessingifPCMmodellingcanbeusedtoassess ligandpromiscuity.Theconcordanceindex(CI)wascalculatedusing the*lifelines*pythonpackage[48].Theassessmentofmodelsensitivity andspecificitywasachievedviatransformingtheaffinitiestoabinary representationindicatingwhetheraligandwasactive(pIC50equalor above6)orinactive(pIC50smallerthan6)foraspecifictarget.ApIC50 thresholdof6isbelievedtobeadequateforseparatinghitsandweaker off-targetinteractions,similarpIC50cut-offshavebeenproposedinprior kinaseinhibitorclassificationstudies[49].Hence,thisbecomesaclas-sificationproblemforwhichthesensitivityandspecificitycanbecalcu-latedforeachofthe927ligandsthatweretestedacrossmorethan25 kinases.Thisanalysiswascarriedoutoverthevalidationresultssince onlytwoligandsinthetestweretestedacrossmorethan25kinases. Statisticalsignificancewasassessedusingthenon-parametricWilcoxon test.Duetothelargetestnumber(927ligands),thealphathresholdwas correctedusingtheBonferronicorrection.The*p*-valueswerecompared withthecorrectedalpha(*p*below5.10−5)andthenullhypothesiswas rejected/accepted.Inaddition,toassessselectivity,themodels’ability topredictdifferencesinaffinitiesperligandacrosssimilartargetswas assessedbyconcentratingonfourselectivityexamples:selectivitybe-tween(1)CDK1andCDK2,(2)SIK2andSIK3,(3)IGF1RandFGFR1, (4)CLK2andCLK3.Toachievethis,wesoughttorelatethePCCbetween thedeltainexperimentalaffinitiesandthedeltainpredictedaffinities foreachpairoftargets.

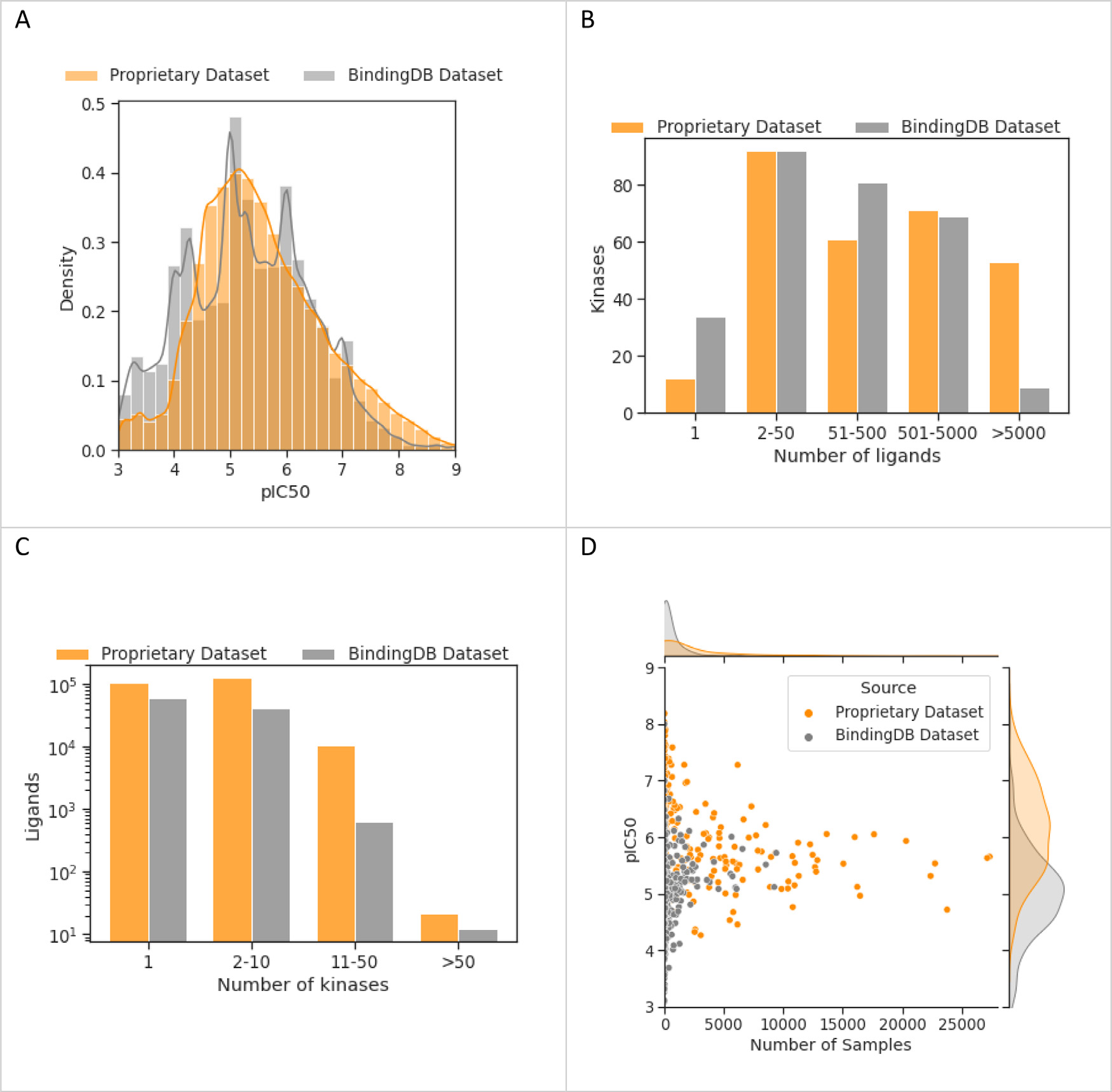
**Results&discussion**

*Datasetcharacteristicsandlimitations*

Inthisfirstsectionwesoughttoanalysethecharacteristicsofthe proprietarydatasetandcomparethemtotheBindingDB[31]dataset usedinarecentPCMpaperfocusedonthekinasetargetclass[41];an overviewisprovidedinFig.2.Priorworkhasshownthatbenchmark-ingdatasetscanbebiasedtowardsactivecompounds[50],presumably duetoadrivetopublishpositiveresults.Nevertheless,theoveralldata distributionprofileofthecuratedin-housedatasetlargelyresemblesthe distributionoftheBindingDBkinasesubset,withamedianpIC50of5.4 and5.3respectively.Thein-housedatasetcontainsfivetimesmoredata (742,533vs.158,900).Inthein-housedataset,53kinaseshavemore than5000ligandstestedagainstthem,versus9forpublicdata,respec-tively.Thisanalysishighlightsthatcertainkinasesdominatethedataset

4

*A.Stroobants,L.H.Mervin,O.Engkvistetal.*  *ArtificialIntelligenceintheLifeSciences4(2023)100079*



**Fig.2.DatasetcharacteristicscomparisonofkinaseinhibitordatafromaproprietarysourceandBindingDB.A)**DistributionofpIC50scoresintheproprietary dataset(*N*=742,533)andBindingDB(*N*=185,988).**B)**Histogramofthenumberofligandsscreenedperkinase.Morekinaseshavebeenscreenedforover5000 ligandsintheproprietarydataset.**C)**Histogramofthenumberofkinasesscreenedperligand.Moreligandshavebeenscreenedforover10kinaseintheproprietary dataset.**D)**UnderstudiedkinasestendtohavemoreactivesamplesintheproprietarydatasetandmoreinactivesamplesinBindingDB.

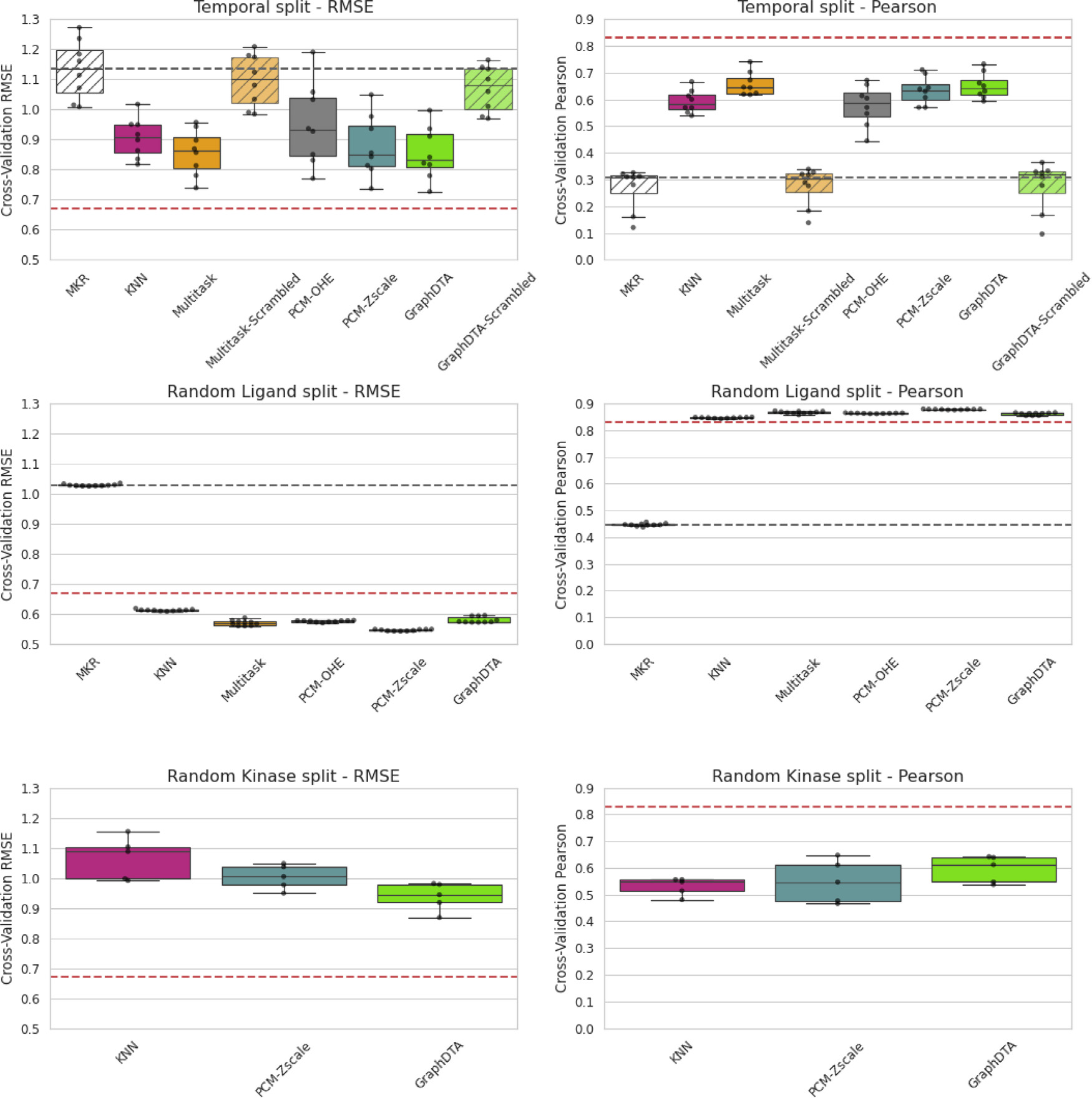
achievesanRMSEof0.94andaPCCof0.59whichisaweakerperfor-mancecomparedtoresultsobtainedforthetemporalsplit.

Thedifferenceinperformancebetweentheligandsplittingstrategies (randomandtemporal)islikelytobeinpartduetothetemporalsplit alsoevaluatingproteingeneralisabilityasprogressivelydatabecomes availablefornewkinasesovertime.Additionally,thedifferenceinper-formancecanalsobeattributedtoanincreasedsimilaritybetweenthe ligandsinthetrainingandvalidationsetsintheligandsplit,whilea largershiftinchemicalspacebetweentrainingandvalidationisex-pectedforatemporalsplit.Indeed,theaverageTanimotosimilarity acrossfoldsinthetemporalsplitis0.50±0.02whileintherandom ligandsplittheaveragescoreis0.68±0.001.Outsideofanindustrial setting,otherligandsplittingstrategiescanbeadoptedtobetterassess theperformanceofPCMuponbiggershiftsinthechemicalspace,such asascaffoldsplitorevenaclusteredsplitwhereligandsareclustered basedonsimilarity.

Thevariabilitybetweenfoldsisgreaterinthetemporalsplitcom-paredtotherandomsplit,withastandarddeviationof0.05versus 0.005,respectively,whichcanbepartiallyexplainedbyvariabilityin thedataqualityofthevalidationsetandthescaleofthechemicalshift

5

*A.Stroobants,L.H.Mervin,O.Engkvistetal.*  *ArtificialIntelligenceintheLifeSciences4(2023)100079*



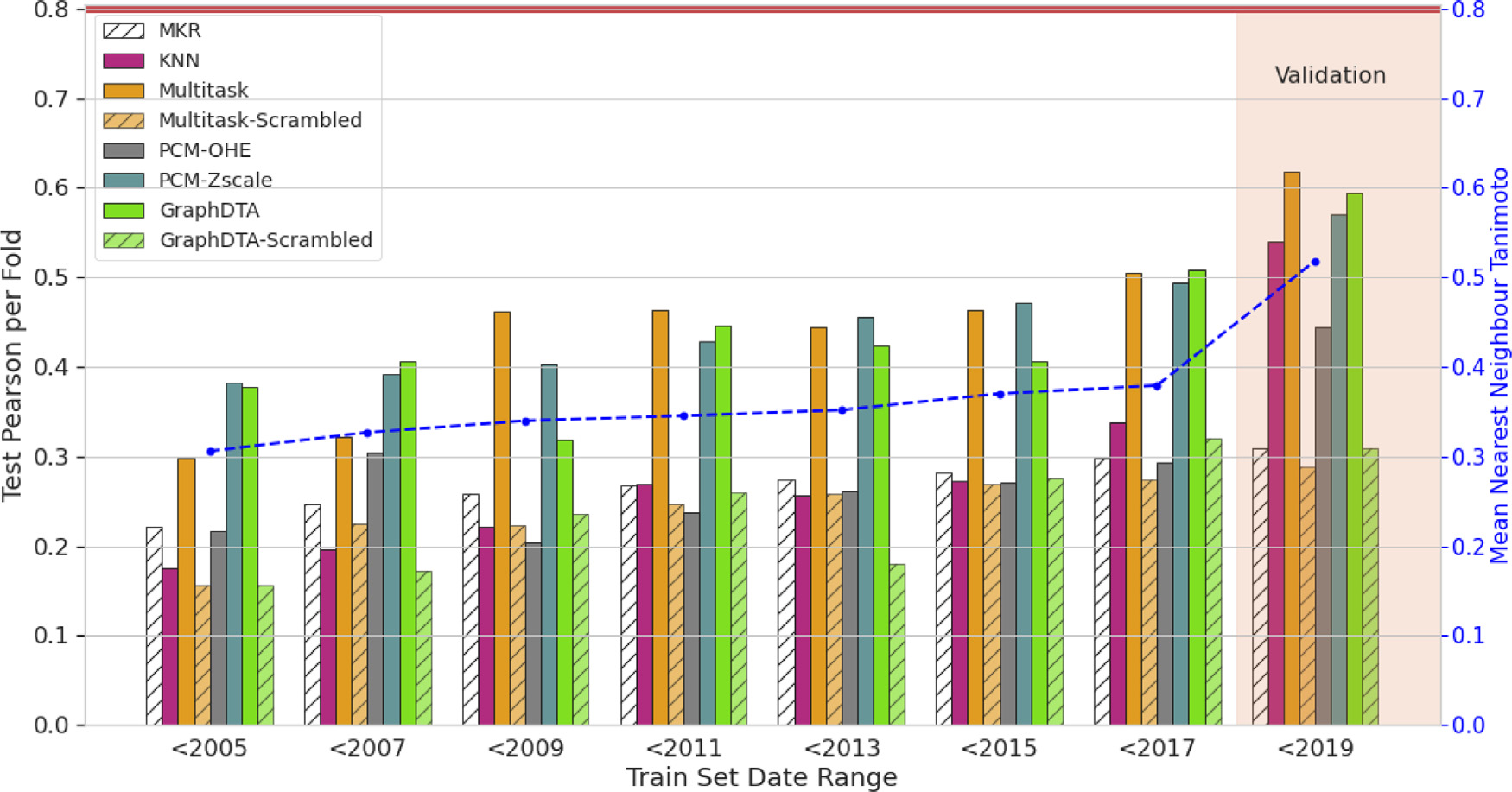
**Fig.3.Distributionofcross-validationperformanceresultsacrossfolds**.Performanceisreportedusingtwometrics,PCC(right)andRMSE(left),andacross threesplittingstrategies:thetemporal(upper),randomligand(middle),andkinase(lower)splittingstrategies.Acrossthedifferentsplitingstrategies,different relativemodelperformancecanbeobserved:intherandomligandsplitPCM-ZscaleisthebestperformingmodelwhileGraphDTAisthebestperformingmodelin thetemporalandkinasesplit.MAPisrepresentedbythereddottedlineswhilethemedianperformanceachievedwiththeMKRisshownbytheblackdottedlines. TherandompredictionmodelsforGraphDTAandthemultitaskmodelareindicatedbythehashedboxes.Allmodelsperformstatisticallybetterthanassigningthe medianpotencyperkinase,whichhighlightstheirpredictiveabilitiesbeyondthelearningofthepIC50biasperkinase.Theexactnumericalscorescanbefoundin TablesS1–S3.AsummaryofkeystatisticalanalysisontheseresultsisprovidedinTableS4.

[46].Despitethesereportsandresults,KNN,RFandmultitaskmodels arenotoftenconsideredasbaselineswhenassessingDLPCMmodels, andhencetheperformancesobtainedhereprovidestrongevidencethat suchsimplecontrolsshouldbemoreroutinelyincludedinPCMmodel evaluations.

Inthissectionwehaveprovidedempiricalevidencethatthetem-poralsplittingstrategyprovidesamorerealisticassessmentoftrue prospectiveperformanceasitsimultaneouslyconsidersashiftinthe chemicalandproteinspace.Wehavealsodemonstratedthatwitha temporalsplitthestate-of-the-artPCMmodelperformssimilarlytoa multitaskmodelonaproprietarydataset.Wealsohighlighttheimpor-tanceofconsideringothermodellingapproaches,suchasKNN,RFand multitaskmodels,asbenchmarksfortheevaluationofDLPCMmodels. WecallattentiontotheremaininggaptoMAP.Thissuggeststhatthere isahypotheticalwindowtobeexploitedbybetterperformingmodels inthefuture.

6

*A.Stroobants,L.H.Mervin,O.Engkvistetal.*  *ArtificialIntelligenceintheLifeSciences4(2023)100079*



**Fig.4.Temporalsplittestresultsindicatetime-dependentperformancebehaviortowardmorerecentdata.**Theperformanceofmodelsimproveswhen trainedonprogressivelymoredata(fromlefttorightontheTrainSetDateRange),expandingthechemicalspaceexploredintraining,leadingtohigherTanimoto similaritybetweentrainandtestsetcompounds.ThemeanTanimotosimilarityacrossallfoldsisindicatedbythebluelinewhiletheMAPisindicatedbythered line.TherandompredictionmodelsforGraphDTAandthemultitaskmodelareindicatedbythehashedbars.

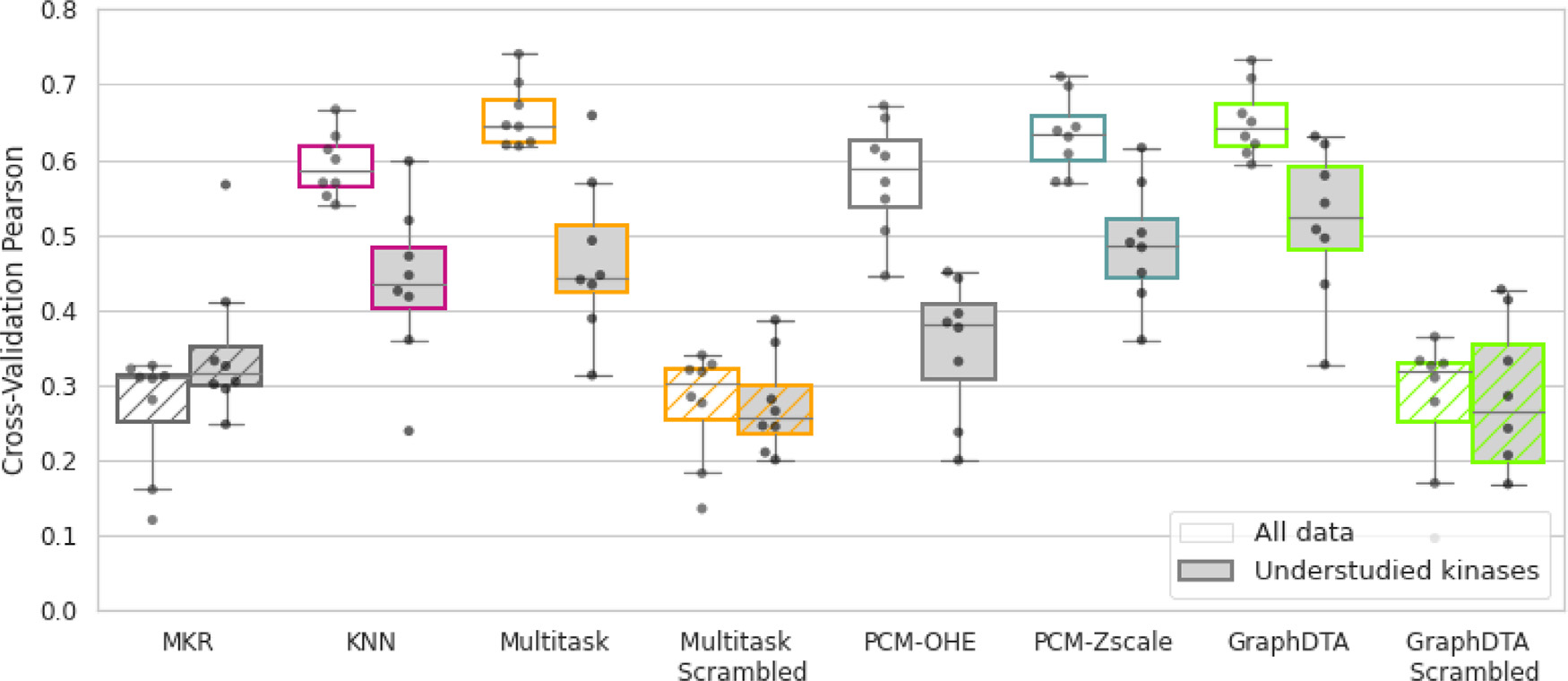
idation,whentrainedonalldataavailable,GraphDTAandthemul-titaskmodelachievedcomparativelythehighestperformance,outper-formingtheKNNandtheRFPCMmodelsbyarelativelysmallmargin of0.05.However,duringtesting,theKNNachievesdifferingperfor-mancetoGraphDTA,wheretheaforementionedoutperformstheKNN modelbyalargermarginof0.20onthefirstfold(“*<*2005″).Thisfold istrainedonamuchsmallersubsectionofthedata(morethan2times smaller)andthechemicalspaceexploredisfurtherawayfromthetest set(averageTanimotosimilarityof0.31±0.08versus0.52±0.17for validation).ThishighlightsthattheKNNmodelhaslimitedpredictivity whentrainedonsmallerdatasetswithlargershiftsinchemicalspace betweentrainingandtesting.Ontheothercontrary,GraphDTAandthe multitaskmodelachievesomepredictiveperformancebeyondtheMKR inconditionswhenthetestedmoleculesarefurtherawayfromtraining chemicalspace.Thesuperiorperformanceintheseconditionssuggests theyarebetterequippedathandlinglargershiftsinchemicalspace. Tosummarise,despitethecomparableoverallperformancesbe-tweentheKNN,themultitaskandGraphDTAhighlightedduringcross-validation,differencesinperformancecanbeobservedwhenassessing theeffectofprogressiveaccumulationoftrainingdataandtheexpan-sionofthechemicalspaceexploredintraining.GraphDTAandthemul-titaskmodeloutperformtheKNNinconditionswherethereisamore pronouncedshiftinthechemicalspaceexploredbetweentrainingand testing,suggestinghaveawiderdomainofapplicability.

*Performanceonunderstudiedkinases*

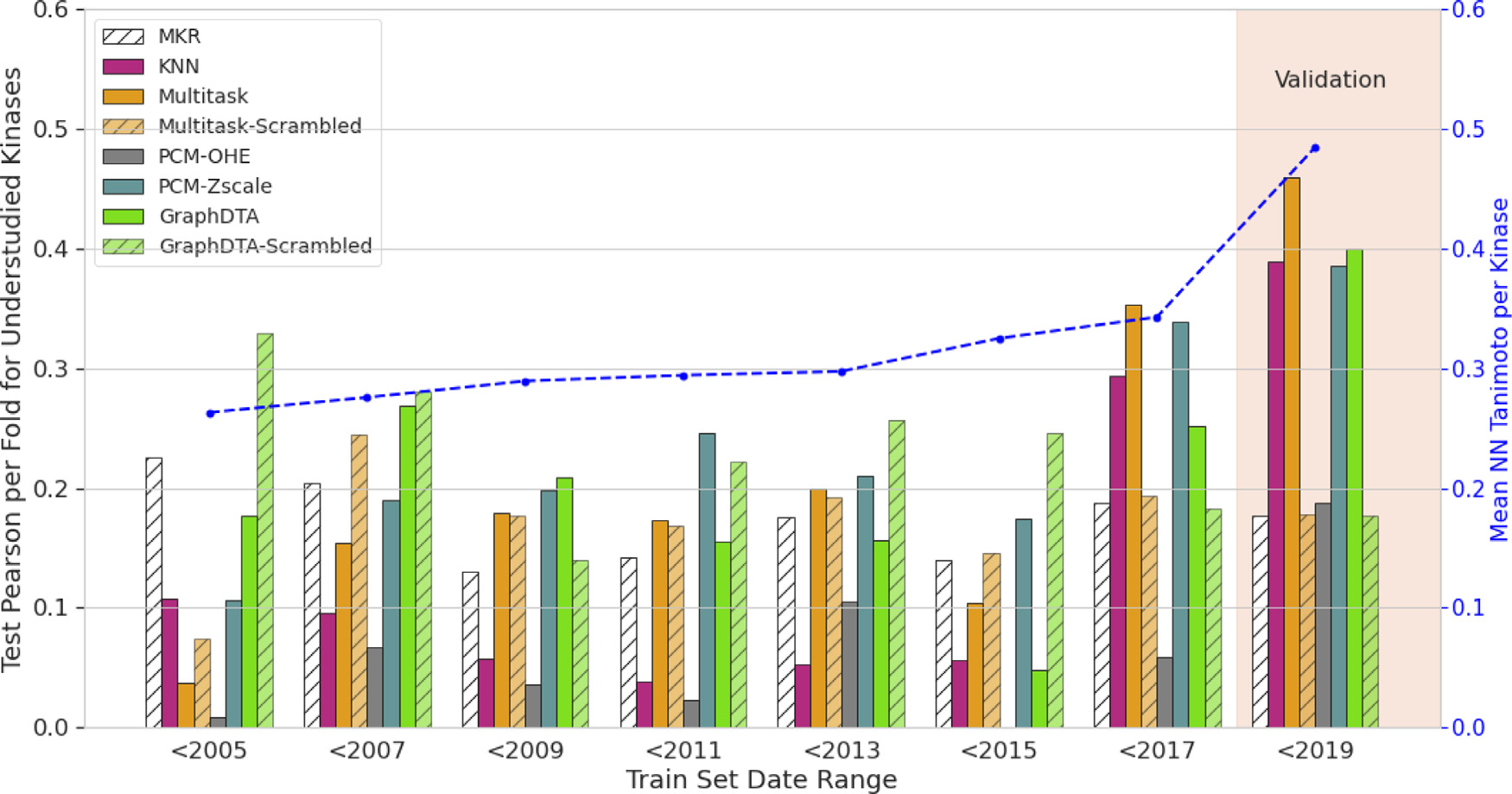
Inthissectionwesoughttoassesstheabilityofthemodelstopredict ligandaffinityforunderstudiedkinasesasdescribedinthemethodsec-tion.Thisaspectisofvitalimportanceinindustrytoenablethedesignof competitiveinhibitorsfornoveltargetsthatcaneitherbewildtypepro-teinsthatremainmainlyunexploredormutatedproteinsinthecontext ofdrugresistance.Inthisvein,Fig.5highlightsthecross-validationper-formanceofthemodelsonunderstudiedkinases.Forallmodels(KNN, GraphDTA,Multitask,PCM-OHEandPCM-Zscale),theperformanceis considerablylowerwhenonlyunderstudiedkinasesareconsideredbut somepredictiveabilityremains,withPCCssignificantlyabovethatof

7

*A.Stroobants,L.H.Mervin,O.Engkvistetal.*  *ArtificialIntelligenceintheLifeSciences4(2023)100079*



**Fig.5.Decreasedperformanceforunderstudiedkinasesintemporalcross-validation.**TheperformanceofGraphDTAisstatisticallycomparabletothemultitask modelwhenconsideringalltargetsandonlyunderstudiedkinases.AllmethodsexceptforthemultitaskandPCM-OHEmodelsarestatisticallysuperiortotheMKR forunderstudiedkinases.GraphDTAandthemultitaskmodelsignificantlyoutperformedtheirrespectiverandompredictionmodels(indicatedbythehashedboxes), highlightingthatsomepredictiveabilityisretainedforunderstudiedkinases.



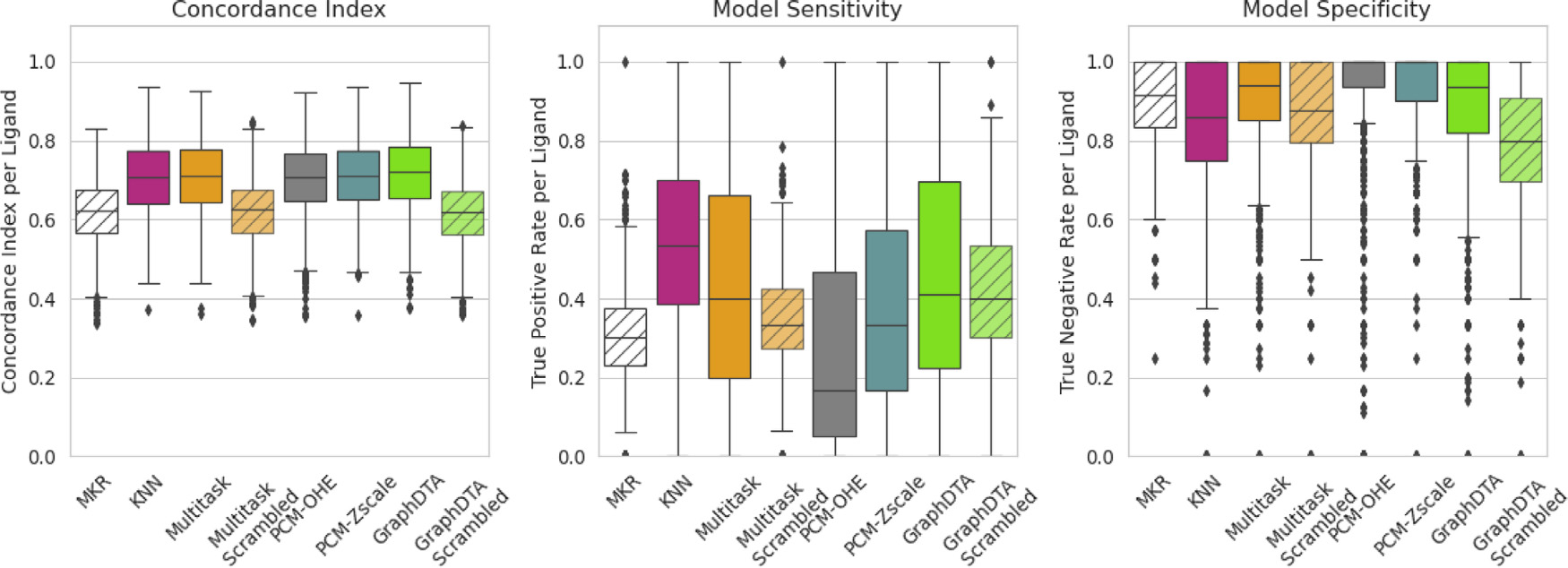
**Fig.6.Temporalperformanceinpredictingaffinityforunderstudiedkinases.**Theperformanceisassessedfor18understudiedkinasesinthetestset.Themean nearestneighbourTanimotosimilarityperkinase(blue)isalsopresentedacrossthefolds.Forthefirstsixfolds(*<*2015),poormodelperformanceisobservedwith resultscomparabletothatoftheMKR(hatchedwhitebar)andrandomprediction(hatchedbars)highlightingalackofpredictiveabilitywhennosimilarcompounds arefoundinthetrainingset.

*Performanceinthecontextofligandpromiscuityandselectivity*

AsdiscussedintheIntroductionsection,akeychallengeinthede-velopmentofeffectivekinaseinhibitorsisligandselectivity.Inthisvein, wefurthersoughttoevaluatethemodelsinthecontextofassessingki-naseinhibitorselectivity.High-throughputvirtualscreeningsolutions forproteinkinasesshouldenabletheidentificationofselectivekinase inhibitorsandthereforetheyshouldbesensitiveandselectiveenough topredictdifferencesinaffinitiesperligandacrossseveraltargets.For ligandsevaluatedoveratleast25targets,thesensitivityandselectiv-ityofthemodelswasassessedaswellasthemodel’sabilitytopredict ligandaffinityacrosstargetsusingtheconcordanceindex,asdescribed intheMethodsection.TheresultshighlightedinFig.7,demonstrate thateventhoughallmodelsarestatisticallyoutperformingtheMKR,the abilityofthemodelstocorrectlyreflectrelativeligandaffinityacross

8

*A.Stroobants,L.H.Mervin,O.Engkvistetal.*  *ArtificialIntelligenceintheLifeSciences4(2023)100079*



**Fig.7.Performanceonthepredictionofligandpromiscuity.**TheabilityofmodelstoassesspromiscuityisassessbyevaluatingtheConcordanceindexandthe models’sensitivityandspecificityonaperligandbasis(*N*=927).AllmodelsshowamoderateperformanceandthatsignificantlyoutperformsasimpleMKR.

**Table1**   
**Performanceinpredictingligandspecificity.**TheperformanceisevaluatedwiththePCCbetween theexperimentalandpredicteddeltainaffinitiesforfourpairsofsimilartargets.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| PCCbetweenpredictedandexperimentalpIC50delta | Multitask | PCM-OHE | PCM-Zscale | GraphDTA |
| Δ(CDK1-CDK2)  Δ(SIK2-SIK3)  Δ(IGF1R-FGFR1)  Δ(CLK2-CLK3) | 0.44  0.47  0.55  0.50 | 0.43  0.33  0.55  0.41 | 0.46  0.37  0.60  0.49 | 0.41  0.48  0.48  0.54 |

CDK2,(2)SIK2andSIK3,(3)IGF1RandFGFR1,(4)CLK2andCLK3.To achievethis,wesoughttorelatethePCCbetweentheexperimentaland predicteddeltainaffinitiesforeachpairoftargets,whichisshownin Table1.ThemultitaskmodeloutperformedGraphDTAintwooutofthe fourexamples,whiletheoppositewastruefortheothertwoexamples. ThePCM-Zscalewasthebestperformingmodelintwoexampleswhilst itwasamongstthetwoworstperformingmodelsfortheothertwoex-amples.ThevariabilityinrelativeperformanceindicatesthatGraphDTA doesnotoffersuperiorpredictivityoveramultitaskmodelfortheas-sessmentofligandselectivityforthetworelatedtargetsanalysedhere, andhighlightstheimportanceofbenchmarkingortriallingdifferental-gorithmsfordifferentkinaseprojects.Forallfourexamples,GraphDTA achievedaPCCof0.5betweenthetrueandpredicteddeltasinpIC50, whichhighlightsarelativelylimitedperformanceinpredictingrelative deltaaffinitiesbetweentwokinases.

TherelativelypoorpredictiveabilitiesofPCMmodelsinthecontext ofkinasepromiscuityandselectivityhighlightedinthisstudy,indicate thatthismodellingtechniqueprovidesonlyalimitedbenefitinanin-dustrialsettingwheredesigningselectivekinaseinhibitorsremainsa keychallenge.SincethePCM-ZscalemodeloutperformedthePCM-OHE modelinallfourexamples,webelievethatbetterproteindescriptors couldfacilitatethelearningoftheinteractionsbetweenproteinandlig-andfeaturesandshouldthusinprincipleimprovethesepredictiveabil-ities.

**Conclusion**

Inthisstudy,wehaveprovidedacomprehensiveassessmentofthe predictiveabilitiesofastate-of-the-artPCMmodelonproprietarydata toassessthepotentialofPCMmodellingasavirtualscreeningtoolin anindustrysetting.Theperformanceofthestate-of-the-artGraphDTA PCMmodelonproprietarydataappearsmorelimitedcomparedtothe performancesuggestedintheoriginalpublication,withameanPCCof 0.65.Wepartiallyattributedtheoverlyoptimisticperformancetothe verylenientrandomligandsplittingstrategyemployedintheoriginal publicationwhichdidnotreflecttheshiftinchemicalspaceobserved overtimeinanindustrialsetting.Wethereforerecommendtheuseof

9

*A.Stroobants,L.H.Mervin,O.Engkvistetal.*

**Acknowledgements**

A.S.isafellowoftheAstraZenecaPostdoctoralResearchPro-gramme.

**DeclarationofCompetingInterest**

Theauthorsdeclarethefollowinginterest(s):A.S.receivedfund-ingfromAstraZeneca;L.H.M.,O.E.,andG.R.R.areemployeesofAs-traZeneca.

**Dataavailability**

Thedatathathasbeenusedisconfidential.

**Supplementarymaterials**

Supplementaryma[terialassociatedwiththisarticlec](https://doi.org/10.1016/j.ailsci.2023.100079)anbefound,in theonlineversion,at[doi:10.1016/j.ailsci.2023.100079](https://doi.org/10.1016/j.ailsci.2023.100079).

**References**

[[1]GawehnE,HissJA,SchneiderG.Deeplearningindrugdiscovery.MolInform](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0001)  [2016;35(1):3–14.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0001)

[[2]LipinskiCF,etal.Advancesandperspectivesinapplyingdeeplearningfordrug](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0001)  [designanddiscovery.FrontRobotAI2019;6:108.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0002)

[[3]AskrH,etal.Deeplearningindrugdiscovery:anintegrativereviewandfuture](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0002)  [challenges.ArtifIntellRev2023;56(7):5975–6037.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0004)

[[4]KimberTB,ChenY,VolkamerA.DeepLearninginVirtualScreening:recentAppli-](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0004) [cationsandDevelopments.IntJMolSci2021;22(9):4435.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0004)

[[5]BongersBJ,IjzermanAP,VanWestenGJP.Proteochemometrics– recentde-](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0004)[velopmentsinbioactivityandselectivitymodeling.DrugDiscovToday:Technol](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0005) [2019;32-33:89–98.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0006)

[[6]GiblinKA,etal.Prospectivelyvalidatedproteochemometricmodelsforthepre-dictionofsmall-moleculebindingtobromodomainproteins.JChemInfModel 2018;58(9):1870–88.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0006)

[[7]SharPA,etal.Pred-binding:large-scaleprotein–ligandbindingaffini](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0006)ty[prediction.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0006)  [JEnzymeInhibMedChem2016;31(6):1443–50.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0007)

[[8]Cortés-CirianoI,BenderA,MalliavinT.PredictionofPARPinhibitionwithpro-](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0007)[teochemometricmodellingandconformalprediction.MolInf2015;34(6–7):357–66. [9]Rodríguez-PérezR,BajorathJ.Multitaskmachinelearningforclassifyinghighlyand weaklypotentkinaseinhibitors.ACSOmega2019;4(2):4367–75.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0009)

[[10]SturmN,etal.Industry-scaleapplicationandevaluationofdeeplearningfordrug](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0009)  [targetprediction.JCheminform2020;12(1):26.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0011)

[[11]ÖztürkH,ÖzgürA,OzkirimliE.DeepDTA:deepdrug–targetbindingaffini](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0011)ty[predic-](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0011) [tion.Bioinformatics2018;34(17):i821–9.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0012)

[[12]NguyenT,etal.GraphDTA:predictingdrug–targetbindingaffini](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0012)ty[withgraphneu-](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0012) [ralnetworks.Bioinformatics2020;37(8):1140–7.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0013)

[[13]JiangM,etal.Drug–targetaffini](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0013)ty[predictionusinggraphneuralnetworkandcon-](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0013) [tactmaps.RSCAdv2020;10(35):20701–12.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0014)

[[14]VoitsitskyiT,etal.3DProtDTA:thedeeplearningmodelfordrug-targetaffini](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0014)ty[pre-dictionbasedontheresidue-levelproteingraphs.RSCAdv2023;13(15):10261–72. [15]WangS,etal.MCN-CPI:multiscaleconvolutionalnetworkforcompound-protein](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0014) [interactionprediction.Biomolecules2021;11(8).](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0016)

[[16]TorrisiM,etal.Improvingtheassessmentofdeeplearningmodelsinthecontextof](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0016)  [drug-targetinteractionprediction.bioRxiv2022;04:20.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0017)

[[17]WallachI,HeifetsA.Mostligand-basedclassificatio](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0017)n[benchmarksrewardmemo-](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0017) [rizationratherthangeneralization.JChemInfModel2018;58(5):916–32.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0018)

[[18]YangK,etal.Analyzinglearnedmolecularrepresentationsforpropertyprediction.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0018)  [JChemInfModel2019;59(8):3370–88.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0019)

[[19]CichońskaA,etal.Crowdsourcedmappingofunexploredtargetspaceofkinase](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0019)  [inhibitors.NatCommun2021;12(1):3307.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0020)

[[20]CohenP,CrossD,JännePA.Kinasedrugdiscovery20yearsafterimatinib:progress](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0020)  [andfuturedirections.NatRevDrugDiscov2021;20(7):551–69.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0020)

*ArtificialIntelligenceintheLifeSciences4(2023)100079*

[21][CohenP.Proteinkinases–themajordrugtargetsofthetwenty-firs](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0021)t[century?NatRev](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0021)  [DrugDiscov2002;1(4):309–15.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0022)

[22][CohenP,AlessiDR.Kinasedrugdiscovery– what’snextinthefield](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0022)?[ACSChem.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0022)  [Biol.2013;8(1):96–104.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0023)

[23][Lopez-delRioA,Picart-ArmadaS,Perera-LlunaA.Balancingdataondeep learning-basedproteochemometricactivityclassification.JChemInfModel](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0023) [2021;61(4):1657–69.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0024)

[24][ChenL,etal.HiddenbiasintheDUD-Edatasetleadstomisleadingperformanceof](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0024) [deeplearninginstructure-basedvirtualscreening.PLoSONE2019;14(8):e0220113.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0025) [25][VolkovM,etal.Onthefrustrationtopredictbindingaffiniti](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0025)es[fromprotein–ligand structureswithdeepneuralnetworks.JMedChem2022;65(11):7946–58.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0025)

[26][SundarV,ColwellL.Theeffec](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0025)t[ofdebiasingproteinligandbindingdataongeneral-](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0025) [ization.JChemInfModel2020;60(1):56–62.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0027)

[27][YangJ,ShenC,HuangN.Predictingorpretending:artificia](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0027)l[intelligenceforpro-tein-ligandinteractionslackofsufficientlylargeandunbiaseddatasets.FrontPhar-macol2020;11:69.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0027)

[28][PapadatosG,etal.Activity,assayandtargetdatacurationandqualityinthe](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0027)  [ChEMBLdatabase.JComputAidedMolDes2015;29(9):885–96.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0029)

[29][TangJ,etal.Makingsenseoflarge-scalekinaseinhibitorbioactivitydatasets:a](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0029) [comparativeandintegrativeanalysis.JChemInfModel2014;54(3):735–43.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0030)

[30][BentoAP,etal.TheChEMBLbioactivitydatabase:anupdate.NuclAcidRes](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0030)  [2014;42(Databaseissue):D1083–90.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0031)

[31][GilsonMK,etal.BindingDBin2015:apublicdatabaseformedicinalchem-istry,computationalchemistryandsystemspharmacology.NuclAcidRes 2016;44(D1):D1045–53.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0031)

[32][MervinLH,etal.ProbabilisticRandomForestimprovesbioactivitypredictionsclose](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0031) [totheclassificationthresholdbytakingintoaccountexperimentaluncertainty.J Cheminform2021;13(1):62.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0032)

[33][JiY.,etal.DrugOOD:out-of-Distribution(OOD)DatasetCuratorandBenchmark](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0032) forAI-aidedDrugDiscoveryAFocusonAffinityPredictionProblemswithNoise [Annotations.arXiv;2022.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0034)

[34][BéquignonOJM,etal.Papyrus:alarge-scalecurateddatasetaimedatbioactivity](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0034)  [predictions.JCheminform2023;15(1):3.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0035)

[35][KramerC,etal.Acomprehensivecompanydatabaseanalysisofbiologicalassay](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0035)  [variability.DrugDiscovToday2016;21(8):1213–21.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0036)

[36][DavisMI,etal.Comprehensiveanalysisofkinaseinhibitorselectivity.NatBiotech-](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0036) [nol2011;29(11):1046–51.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0036)

[37][SubramanianV,etal.3Dproteochemometrics:usingthree-dimensionalinformation](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0036) [ofproteinsandligandstoaddressaspectsoftheselectivityofserineproteases.Med-chemcomm2017;8(5):1037–45.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0037)

[38][SubramanianV,etal.Predictiveproteochemometricmodelsforkinasesderivedfrom](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0037)  [3Dproteinfield-baseddescriptors.Medchemcomm2016;7(5):1007–15.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0038)

[39][SheridanRP.Time-splitcross-validationasamethodforestimatingthegoodnessof](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0038)  [prospectiveprediction.JChemInfModel2013;53(4):783–90.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0040)

[40][PedregosaF,etal.Scikit-learn:machineLearninginPython.JournalofMachine](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0040)  [LearningResearch2012:12.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0041)

[41][BornJ,etal.Activesitesequencerepresentationsofhumankinasesoutperformfull sequencerepresentationsforaffinitypredictionandinhibitorgeneration:3Deffects](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0041) [ina1Dmodel.JChemInfModel2022;62(2):240–57.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0042)

[42][vanLindenOP,etal.KLIFS:aknowledge-basedstructuraldatabasetonavigateki-](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0042) [nase-ligandinteractionspace.JMedChem2014;57(2):249–77.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0043)

[43][ModiV,DunbrackRL.Astructurally-validatedmultiplesequencealignmentof497](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0043)  [humanproteinkinasedomains.SciRep2019;9(1):19790.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0044)

[44][SandbergM,etal.Newchemicaldescriptorsrelevantforthedesignofbiologically activepeptides.amultivariatecharacterizationof87aminoacids.JMedChem](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0044) [1998;41(14):2481–91.](https://www.rdkit.org/)

[45]*[RDKit:open-sourcecheminformatics](https://www.rdkit.org/)*[.cited2022;Availablefrom:https://www.rdkit.](https://www.rdkit.org/)  [org/.](https://www.rdkit.org/)

[46][JanelaT,BajorathJ.Simplenearest-neighbouranalysismeetstheaccuracyofcom-](https://www.rdkit.org/)[poundpotencypredictionsusingcomplexmachinelearningmodels.NatMachIntell](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0046) [2022.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0047)

[47][ConoverWJ.Onmethodsofhandlingtiesinthewilcoxonsigned-ranktest.JAm](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0047)  [StatAssoc1973;68(344):985–8.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0048)

[48][Davidson-PilonC.lifelines:survivalanalysisinPython.JOpenSourceSoftw](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0048)  [2019;4:1317.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0049)

[49][SorgenfreiFA,FulleS,MergetB.Kinome-wideprofilin](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0049)g[predictionofsmall](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0049)  [molecules.ChemMedChem2018;13(6):495–9.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0050)

[50][CáceresEL,MewNC,KeiserMJ.Addingstochasticnegativeexamplesintoma-chinelearningimprovesmolecularbioactivityprediction.JChemInfModel 2020;60(12):5957–70.](http://refhub.elsevier.com/S2667-3185(23)00023-5/sbref0050)

10