

[Artificia](https://doi.org/10.1016/j.ailsci.2021.100007)l[IntelligenceintheLifeSciences1(2021)100007](https://doi.org/10.1016/j.ailsci.2021.100007)

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

Computationalpredictionoffrequenthittersintarget-basedandcell-based assays   
ConradStorka,NeannMathaib,JohannesKirchmaira,b,c,∗  
a*UniversitätHamburg,FacultyofMathematics,InformaticsandNaturalSciences,DepartmentofInformatics,CenterforBioinformatics,20146Hamburg,Germany* b*DepartmentofChemistryandComputationalBiologyUnit(CBU),UniversityofBergen,N-5020Bergen,Norway*   
c*DepartmentofPharmaceuticalSciences,DivisionofPharmaceuticalChemistry,FacultyofLifeSciences,UniversityofVienna,1090Vienna,Austria*

|  |  |  |
| --- | --- | --- |
| article | info | abstract |
| *Keywords:*  Machinelearning  Frequenthitters  Nuisancecompounds  PAINS  Biologicalassays  High-throughputscreening | | Compoundsinterferingwithhigh-throughputscreening(HTS)assaytechnologies(alsoknownas“badlybehav-ingcompounds”,“badactors”,“nuisancecompounds” or“PAINS”)poseamajorchallengetoearly-stagedrug discovery.Manyoftheseproblematiccompoundsare“frequenthitters”,andwehaverecentlypublishedasetof machinelearningmodels(“HitDexter2.0”)forflaggingsuchcompounds.  Herewepresentanewgenerationofmachinelearningmodelswhicharederivedfromalarge,manually curatedandannotateddataset.Forthefirsttime,thesemodelscover,inadditiontotarget-basedassays,also cell-basedassays.Ourexperimentsshowthatcell-basedassaysbehaveindeeddifferentlyfromtarget-basedas-says,withrespecttohitratesandfrequenthitters,andthatdedicatedmodelsarerequiredtoproducemeaningful predictions.Inadditiontotheseextensionsandrefinements,weexploredavarietyofadditionalsetupsformod-eling,includingthecombinationoffourmachinelearningclassifiers(i.e.k-nearestneighbors(KNN),extratrees, randomforestandmultilayerperceptron)withfoursetsofdescriptors(Morgan2fingerprints,Morgan3finger-prints,MACCSkeysand2Dphysicochemicalpropertydescriptors).  Testingonholdoutdataaswellasdatasetsof“darkchemicalmatter” (i.e.compoundsthathavebeenexten-sivelytestedinbiologicalassaysbuthavenevershownactivity)andknownbadactorsshowthatthemultilayer perceptronclassifiersincombinationwithMorgan2fingerprintsoutperformothersetupsinmostcases.Thebest multilayerperceptronclassifiersobtainedMatthewscorrelationcoefficientsofupto0.648onholdoutdata.These modelsareavailableviaafreewebservice. |

**Introduction**

High-throughputscreening(HTS)assaytechnologiesareacorner-stoneofmoderndrugdiscovery.Theyallowthebiologicaltestingof largenumbersofcompoundsontargetsofinterestwithinashortpe-riodoftime[1].Amajorchallengefacedinhigh-throughputscreening isfalse-positivehitsresultingfromdifferenttypesofassayinterference [2].

Compoundscausingassayinterferencearereferredtoas“badlybe-havingcompounds”,“badactors” or“nuisancecompounds”.Manyof them,butbyfarnotall,are“frequenthitters” (i.e.compoundswhich showhigher-than-expectedhitratesinbiologicalassays).Thisisbecause notalltypesofassayinterferencearefrequentevents.Infact,many typesofassayinterferencearetriggeredonlybyspecificconditions. Importantly,notallfrequenthittersarenuisancecompounds.Quite onthecontrary:frequenthitterbehaviorcanbearesultoftruepromis-cuitymediatedby“privilegedscaffolds” [3].Privilegedscaffoldsen-

∗Corresponding[author.](mailto:johannes.kirchmair@univie.ac.at)

*E-mailaddress:*[johannes.kirchmair@univie.ac.at](mailto:johannes.kirchmair@univie.ac.at)(J.Kirchmair).

ablecompoundstobind,inaspecificmanner,toanumberofdistinct proteins.Suchcompoundscanbeparticularlyusefulinthecontextof polypharmacologyanddrugrepurposing.

Anestablishedexperimentalstrategytodiscriminategenuinehits fromfalse-positiveresultsistheuseoforthogonalandcounterscreen assays[4],butevenwithsuchanadvancedexperimentalsetupsome casesofassayinterferencemaynotbecapturedbecausetheunderlying mechanismsaremanifold.

Giventhecomplexitiesinvolvedintheconductionandanalysisof experimentalscreens,computationaltoolstoaidthediscriminationof genuinehitsfromfalseonesareinhighdemand.Today,avarietyofin silicoapproachesforcherry-pickingthemostpromisinghitsforfollow-upstudiesareatourdisposal[5–10].Wewilldiscussthesebrieflyin thecontextoftheindividualtypesofassayinterference.

Themostprominentcauseofinterferenceinbiologicalassays(bio-chemicalassaysinparticular)isrelatedtotheformationofaggre-gatesbysmallmoleculesthatengageinnonspecificinteractionswith

<https://doi.org/10.1016/j.ailsci.2021.100007>  
[Received14June2021;Receivedinrevisedfo](https://doi.org/10.1016/j.ailsci.2021.100007)rm5August2021;Accepted6August2021   
Availableonline8August2021   
2667-3185/© 2021TheAuthors.PublishedbyElsevierB.V.ThisisanopenaccessarticleundertheCCBYlicense(<http://creativecommons.org/licenses/by/4.0/>)

*C.Stork,N.MathaiandJ.Kirchmair*

biomacromolecules[5].Severalcomputationalapproacheshavebeen reportedfortheassessmentofsmallmoleculeswithregardtotheirrisk offormingcolloidalaggregates.ThesetoolsincludeAggregatorAdvi-sor[11],ChemAgg[12]andSCAMdetective[13].AggregatorAdvi-sorflagspotentialaggregatorsbasedontheirmolecularsimilaritytoa setof12,000knownaggregators,takinglogPintoaccount.ChemAgg andSCAMDetectivearemachinelearningmodelsfortheclassifica-tionofsmallmoleculesintoaggregatorsandnon-aggregators.Whereas ChemAggisbasedonaXGBoostmodel,SCAMdetectiveutilizesasetof randomforestmodels.

Asecondimportantcauseofassayinterferenceisthechemicalre-activityofcompounds,inparticularthatrelatedtoelectrophilicity [14].Chemicallyreactivecompoundsmaybindcovalentlytobiomacro-moleculesorinteractwiththeassayscreeningtechnologyinanun-desiredway.Computationalapproachesforidentifyingreactivecom-poundsaremostlybasedonsetsofruleswhichdescribesubstructures thathavebeenlinkedtochemicalreactivity[15].

Furthertypesofassayinterferencearecoveredundertheumbrella ofthewell-knownpan-assayinterferencecompounds(PAINS)concept [16].PAINSarecompoundsbasedonmolecularscaffoldsthathavebeen associatedwithvarioustypesofassayinterference.PAINSincludere-doxcyclingcompounds(e.g.toxoflavins),covalentbinders(e.g.isothia-zolonesorene-rhodanines),membranedisruptors(e.g.curcumin),metal complex-formingcompounds(e.g.hydroxyphenylhydrazones)andun-stablecompounds(e.g.phenol-sulfonamides)[17].Themolecularfrag-mentslinkedtoPAINShavebeencompiledinacollectionofseveral hundredstructuralpatterns,andthiscollectionhasbeenimplementedin variousinsilicoplatformsandsoftwarelibrariestooffermeansforflag-gingpotentiallyproblematiccompounds[18].Analternativeapproach toflaggingpotentialPAINSwasrecentlypresentedbyKoptelovetal. [19].Theyusediscriminativesubgraphminingtoidentifycharacter-isticpatternsinPAINSandnon-PAINS,andutilizethesepatterns,in combinationwithnumericaldescriptors,toderivedecisiontreemodels forPAINSprediction.

Anumberoffocusedmachinelearningmodelshavebeendevised fortheidentificationofcompoundsthatlikelycausespecifictypesof assayinterference.Forexample,LuciferaseAdvisor[20]andChemFluc [21]aremodelsforthepredictionofcompounds(luciferaseinhibitors) thatmayinterferewithluciferase-basedassays.InterPred[22]includes asetofQSARmodelsforthepredictionofluciferaseinhibitorsandaut-ofluorescencecompoundsincell-basedandtarget-basedassays.

Severalcomputationaltoolsareinexistencethatpredictfrequent hittersindependentoftheunderlyingmechanisms(genuinepromiscu-ity;varioustypesofassayinterference).Forexample,researchersatAs-traZenecahavederivedastatisticalmodelforthepredictionoffrequent hittersbasedontheirin-househistoricalbioactivitydata[23].Another statisticalmodelforthepredictionoffrequenthittersisBADAPPLE[24]. IncontrasttotheAstraZenecamodel,theBADAPPLEmodelisderived frommolecularscaffoldsratherthancompletemolecularstructures. Morerecently,machinelearninghasbeenmovedintothefocusalso inthefieldoffrequenthitterandassayinterferenceprediction.For example,HitDexter2.0[25],developedbysomeofus,predictsfre-quenthittersutilizingasetofextratreemodelsthataretrainedonlarge setsofdataextractedfromthePubChemBioassaydatabase[26].More recently,Feldmannetal.[27]reportedamachinelearningapproach forthepredictionoftruepromiscuouscompounds(multi-targetcom-pounds)inwhichtheyremovedlikelyaggregatorsandothertypesof assayinterferencecompoundsfromthetrainingsetsinanefforttowork withcleanersetsofpromiscuousandnon-promiscuouscompounds. Whereasasizablenumberofinsilicomodelsforthepredictionof frequenthittersandbadlybehavingcompoundsareatourdisposalto-day,mostofthemhaveclearlimitationswithrespecttothecoverageof mechanismsofinterferenceandassaytechnologies.Inparticular,theex-istingapproachesarefocusedon,orlimitedto,biochemical(i.e.target-based)assaysanddonotadequatelyrepresentcell-basedassays,which canbehaveverydifferentlywithrespecttoassayinterference.

2

*C.Stork,N.MathaiandJ.Kirchmair*  *ArtificialIntelligenceintheLifeSciences1(2021)100007*

**Table2**   
Definitionsofvaluesforthemanuallyassignedlabel“bioactivitytype”.

|  |  |
| --- | --- |
| Labelvalue | Description |
| specificbioactivity  nonspecificbioactivity  other | Assaysdesignedtomeasureaspecificbiologicalpropertysuchastheactivityofanenzyme.Cytotoxicityassaysarenotincludedin thiscategory.Counterscreenassaysareincludediftheymeasureaspecificbiologicaleffect.Anexampleofacounterscreen assignedthislabelvalueisaluciferasecounterscreenthatiscommonlyemployedtoidentifycompoundswhichcancause interferenceinluciferase-based(bioluminescence)assays  Assaysthatmeasurecellgrowth,cellviability,cytotoxicity,cellgrowthinhibition,orothernonspecificassayreadouts Assaysthatmeasurephysicochemicalprocesses(notbioactivities),DNAorRNAbinding,etc. |

**Table3**   
Datasetsizesandcompoundsremovedduringchemicalstructureprocessing.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Target-based  assaydataset | Cell-basedassay dataset | Extendedcell-based assaydataset |
| No.compoundsinthedatasetpriortochemicalstructureprocessing  No.compoundsremovedduetoinvalidSMILES  No.compoundsremovedduetolackofasingle,validactivityoutcome1  No.compoundsremovedduetopresenceofelementsuncommontodrug-likecompounds No.compoundsremovedbythemolecularweightfilter  No.compoundsinthefinaldataset | 1,545,406  1  45,184  331  10,847  1,489,043 | 1,421,472  3  23,259  381  11,120  1,386,709 | 1,858,887  9  53,984  3151  22,106  1,779,637 |

1Compoundsthatwereremovedbecauseofthelackofavalidactivityoutcomethatcanbederivedfromtherawdata(i.e.compoundswithoutasingleannotated“Active” or“Inactive” assayoutcome)

Asthelastfilteringcriterion,anyassayswithoutatleastonecom-poundmeasuredasactiveandonecompoundmeasuredasinactivewere removedfromthedataset.Foracompleteoverviewofallassaysre-movedduringdatapreparationseeTableSI\_1.

*Chemicalstructureprocessing*   
 TheSMILESnotationsofthe1,545,406compoundscoveredbythe target-basedassaydataset,the1,421,472compoundscoveredbythe cell-basedassaydataset,andthe1,858,887compoundscoveredby theextendedcell-basedassaydatasetwereretrievedfromthePub-ChemBioassaydatabaseviathePubChemPUGRESTinterface[32].The ChEMBLStructurePipeline[33](alsoknownas“ChEMBLCompound CurationPipeline”),wasutilizedto(i)neutralizechargedmolecules, (ii)removesaltandsolventcomponents,and(iii)neutralizecharged moleculesoncemore(tocovercaseswhereachargedcomponentwas removedduringstepii).Thetechnicaldescriptionofthischemicalstruc-turepreparationprocedureisreportedinRef.[33].

Anycompoundswithmolecularweightbelow180orabove900Da wereremovedfromthedataset,aswellasanycompoundscomposed ofanyelementsotherthanH,B,C,N,O,F,Si,P,S,Cl,Se,Brand I.Moleculesrepresentedbymorethanonetautomerweremergedtoa singlerepresentationusingthe“canonalize” methodimplementedinthe“TautomerEnumerator” classofRDKit[34](version2020.09.1).During thisprocedurethecompoundswererepresentedasRDKitmoleculesand wereinalaststepconvertedtocanonicalSMILES.Furtherduplicate compoundswereremovedbasedonidenticalSMILES.Foranoverview oftheremovedcompoundsseeTable3.Foralladditionaldatasetsused withinthisstudy,includingtheChEMBL23database[35],thedark chemicalmatter(DCM)datasetcompiledbyWassermannetal.[36], thedatasetofDahlinetal.[37](containingcompoundsthatareknown tocauseinterferenceinbiologicalassays),andthedatasetofBorrel etal.[22](containingcompoundsthatwereexperimentallyconfirmed tocausefalsepositivereadoutsinbioluminescenceassaysduetolu-ciferaseinhibitionand/orautofluorescence),thesamechemicalstruc-turestandardizationprocesswasperformed.SincethedatasetofBorrel etal.containsonlyCASnumbersascompoundidentifiers,theSMILES notationswerefetchedviatheChemicalIdentifierResolver[38].

*Extractionofactivitydatafromtheselectedassays*   
 Foreachoftheselectedassays,anycompoundsconsistently(i.e.one orseveraltimes)labeledas“Active” weredefinedasactive,andany compoundsconsistentlylabeledas“Inactive” weredefinedasinactive.

3

*C.Stork,N.MathaiandJ.Kirchmair*

tobetakenintoaccountforprediction(n\_neighbors)wassetto1;for theRFandtheETclassifiers,theclassweight(class\_weight)wasset to“balanced”;fortheMLPclassifier(implementedinscikit-learn),the numberofiterationswassetto1000assomeofthecalculationsdidnot convergewithinthedefault,200iterations.

Thegenerationoftheindividualmodelswasrepeatedfortentimes, withdifferentrandomstates(i.e.42to51),inordertocomputetheme-dianandthevarianceoftheperformancemetrics(detailsprovidedinthe Resultssection).Thefinalmodelsweregeneratedwithrandomstate=42 andtheapplicationofthesyntheticminorityoversamplingtechnique (SMOTEversion0.7.0)[45].

*Performancemeasurementsandvarianceestimation*

|  |
| --- |
| TheMCC(Eq.(2))wasusedastheprimarymeasureofmodelperfor-mance.TheMCCisabalancedmetricthattakesthetruepositive(TN), falsepositive(FP),truenegative(TN)andfalsenegative(FN)instances intoaccount:  *MCC*= *TP*⋅ *TN*−*FP*⋅ *FN*  (2)  TheMCCreturnsvaluesbetween-1(totaldisagreementbetweenpre-√(*TP*+*FP*)⋅ (*TP*+*FN*)⋅ (*TN*+*FP*)⋅ (*TN*+*FN*)  dictionandobservation)and+1(perfectagreement).  Theareaunderthe(receiveroperatingcharacteristic)curve(AUC) wasusedasanindicatoroftherankingperformanceofthemodels. Thetestsforstatisticalsignificancewereperformedwiththe“ttest\_rel” functionofthe“scipy.stats” module.Thevarianceintheper-formanceofthemodels(onthetestdata)wasestimatedbytestingthe modelsontenrandomlycompiledsubsets(80%)oftheoriginaltestset. |

**Results**

*Analysis,annotationandrefinementofPubChemBioassaydata*

Inordertodevelopabetterunderstandingoftherelevanceofthe dataavailablefromthePubChemBioassaydatabaseformodelingthe frequenthitterbehaviorofsmallmolecules,weconductedacompre-hensiveanalysisofthechemicalandbiologicaldata.

Withmorethan297Millionmeasuredbioactivities,thePubChem Bioassaydatabaseistheworld’slargest,publiccollectionofbioassay data[30].Itisalsooneofonlyafewdataresourcesofferingaccess toalargeamountofhigh-throughputscreeningdata.Thenumberof measuredbioactivitiesrecordedperassayvariesgreatlyacrosstheindi-vidualassaydatasets,fromasinglecompoundto646,275compounds (Table4).

Wedecidedtobaseourworkonthe1180(i.e.474+706)assaydata setscontainingmeasurementsforatleast10,000compoundsbecause thesedatasetsofferagoodtrade-off betweendataqualityandcoverage. Thevastmajorityofthesedatasetshavebeengeneratedbythemost reputableHTSfacilities(includingtheScrippsResearchInstitute,the Sanford-BurnhamMedicalResearchInstitute,TheBroadInstituteofMIT andHarvard,andtheNIH/NationalCenterforAdvancingTranslational Sciences(NCATS)),forwhichreasonahighstandardinHTScanbe expected.

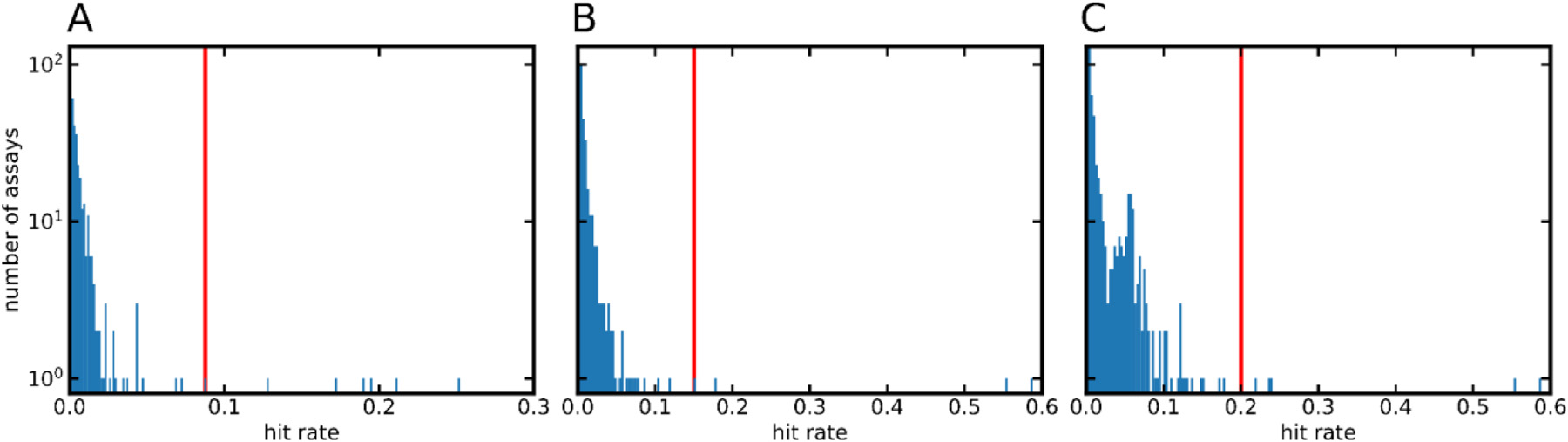
**Table4**   
SizeofthePubChemBioassaydatasets.1

|  |  |
| --- | --- |
| NumberofassaysinthePubChem Bioassaydatabase | Numberofmeasuredcompounds |
| 587,477  633,294  5082  1403  474  706 | 1  2to99  100to999  1000to9999  10,000to99,999  100,000to646,275(maximum) |

1Numbersreferringtotheraw,unprocessedPubChemBioassaydatabase.

4

*C.Stork,N.MathaiandJ.Kirchmair*  *ArtificialIntelligenceintheLifeSciences1(2021)100007*

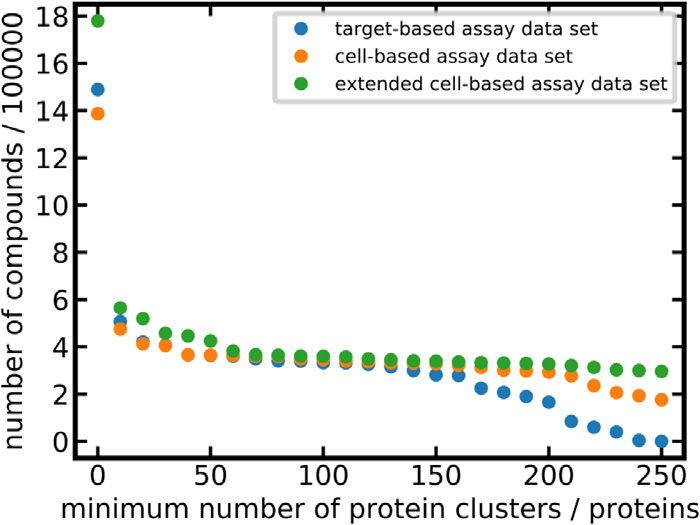


**Fig.1.**Histograms(200binseach)showingthehitratesoftheassaysincludedinthe(A)target-based,(B)cell-based,and(C)extendedcell-basedassaydatasets. Theredlinemarksthemeanhitrate+3*𝜎*.Notethatthescalesofthex-axesdifferforthethreediagrams.

**Table5**   
Compositionofthetrainingandtestsets.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Promiscuityclass | Classdefinitions | No.compoundsin  thetrainingset | No.compoundsin thetestset |
| target-basedassaydata set | HPROM1  PROM  NPROM  HPROM1  PROM  NPROM  HPROM1  PROM  NPROM | ATR*>*0.053  ATR*>*0.022  ATR*<*0.007  ATR*>*0.058  ATR*>*0.025  ATR*<*0.008  ATR*>*0.070  ATR*>*0.030  ATR*<*0.010 | 4614  20274  219061  5578  24913  226382  5135  24673  235241 | 550  2303  24483  616  2825  25427  538  2776  26398 |
| cell-basedassaydataset |
| extendedcell-based assaydataset |

1ThecompoundslabeledasHPROMareasubsetofthecompoundslabeledasPROM.



**Fig.2.**Datasetsize(numberofcompounds)asafunctionoftheminimum numberofproteinclusters(inthecaseoftarget-basedassays)orproteins(in thecaseofcell-basedassays)forwhichmeasureddataareavailable.

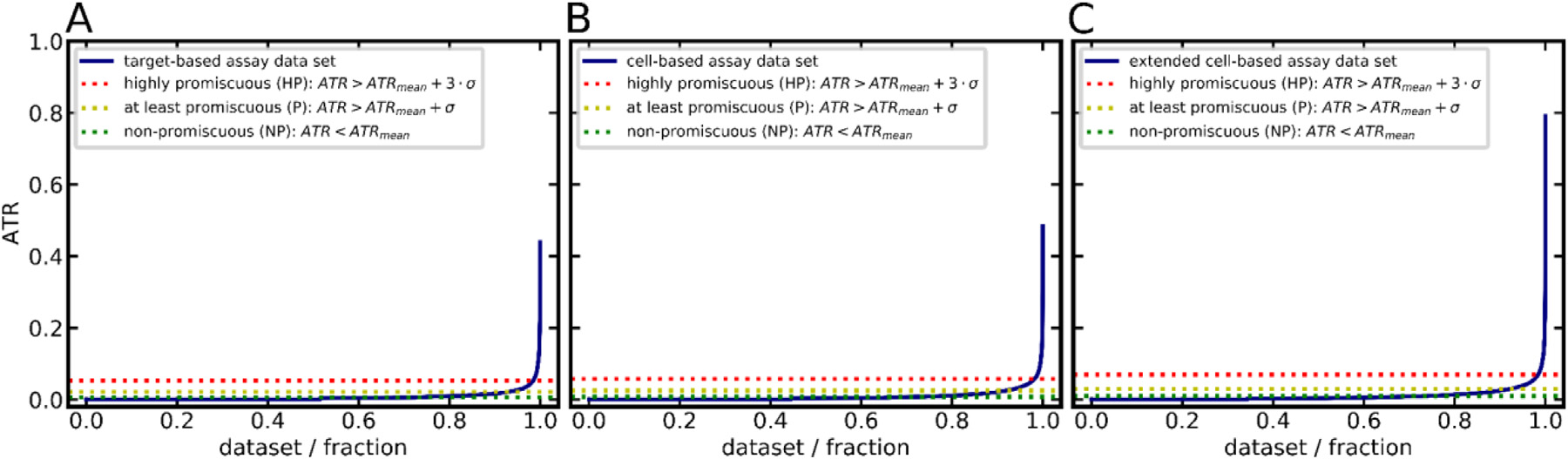
*Analysisofcompoundhitratesandassignmentofpromiscuityclasslabels*

TheATR(Eq.(1))canbeusedtoassigncategoricalpromiscuity valuestocompounds,suchas“non-promiscuous”,“promiscuous” or“highlypromiscuous”.ThesignificanceandrobustnessoftheATRde-pendsonthequalityandquantityoftheunderlyingdata:thehigherthe valueof*T*(i.e.thetotalnumberofassaysacompoundwastestedin),the morerobusttheATR.ThemainadvantageoftheATRoveralternative metricsisitsinterpretabilityasitreflectsthehitrateofacompound. Inthiswork,wesettheminimumthresholdof*T*foracompoundtobe includedinthedatasetsusedformodeldevelopmentto100,whichrep-resentsagoodbalancebetweenATRqualityandcoverage(Fig.2).This filteringprocedureresultedinasetof332,653compoundsmeasured intarget-basedassays,345,743compoundsmeasuredincell-basedas-saysdesignedtomeasureaspecificbioactivity,and360,094compounds measuredinanextendedsetofcell-basedassays.

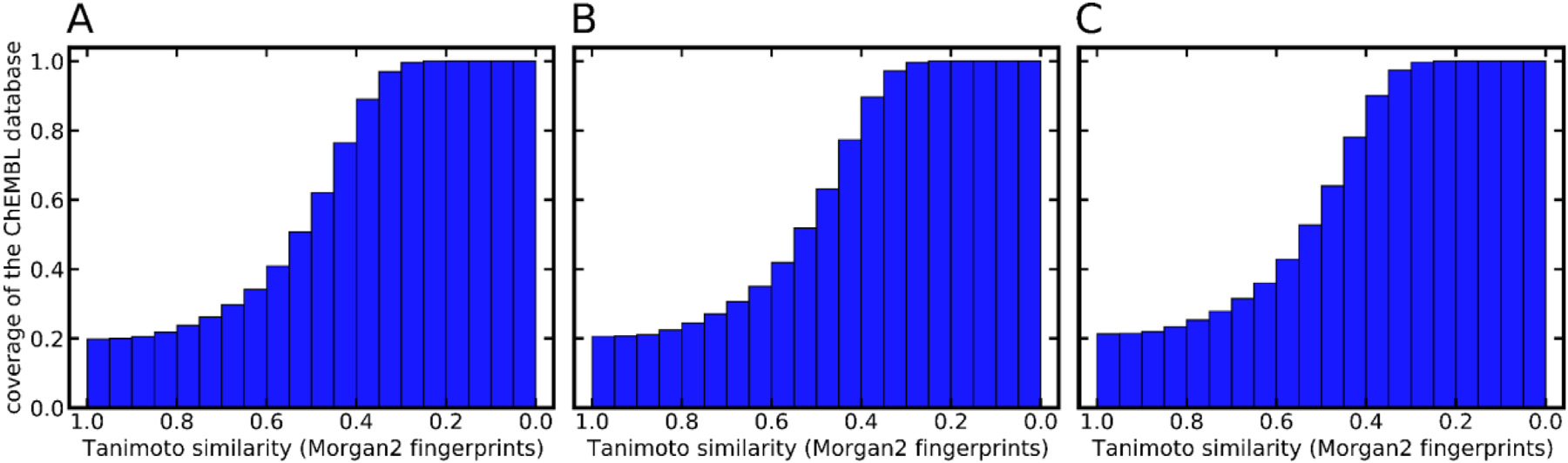
BasedontheATRthresholdsreportedinTable5,allcompoundswere assignedapromiscuitylabel:highlypromiscuous(HPROM),promiscu-

5

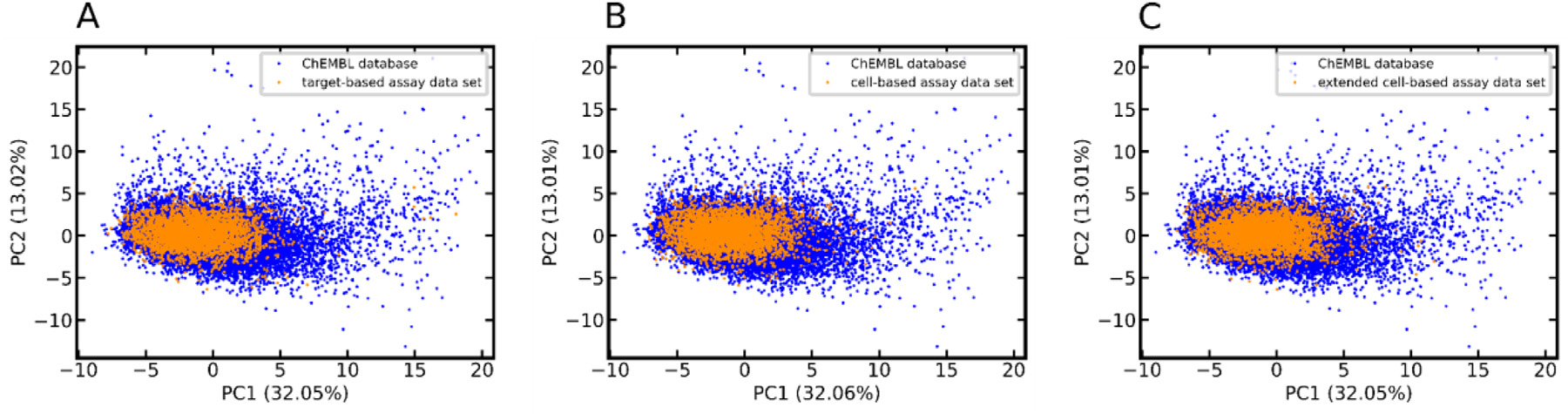
*C.Stork,N.MathaiandJ.Kirchmair*  *ArtificialIntelligenceintheLifeSciences1(2021)100007*



**Fig.3.**ATRdistributionamongcompoundsofthe(A)target-basedassaydataset,(B)cell-basedassaydataset,and(C)extendedcell-basedassaydataset.



**Fig.4.**CumulativecoverageofthecompoundsincludedintheChEMBLdatabasebythecompoundsincludedinthe(A)target-basedassaydataset(B)cell-based assaydataset,and(C)extendedcell-basedassaydataset.



**Fig.5.**PCAoftheChEMBLdatabaseandthe(A)target-basedassayset,(B)cell-basedassayset,and(C)extendedcell-basedassayset.ThePCAisderivedfromthe 442Dmolecularpropertydescriptors(seeTableS1inRef.[46])implementedinMOE.Forthesakeofclarity,only1%ofthedatapoints(randomlyselected)are visualized.Thenumbersinparenthesesreportthevarianceexplainedbytherespectiveprincipalcomponent(PC).

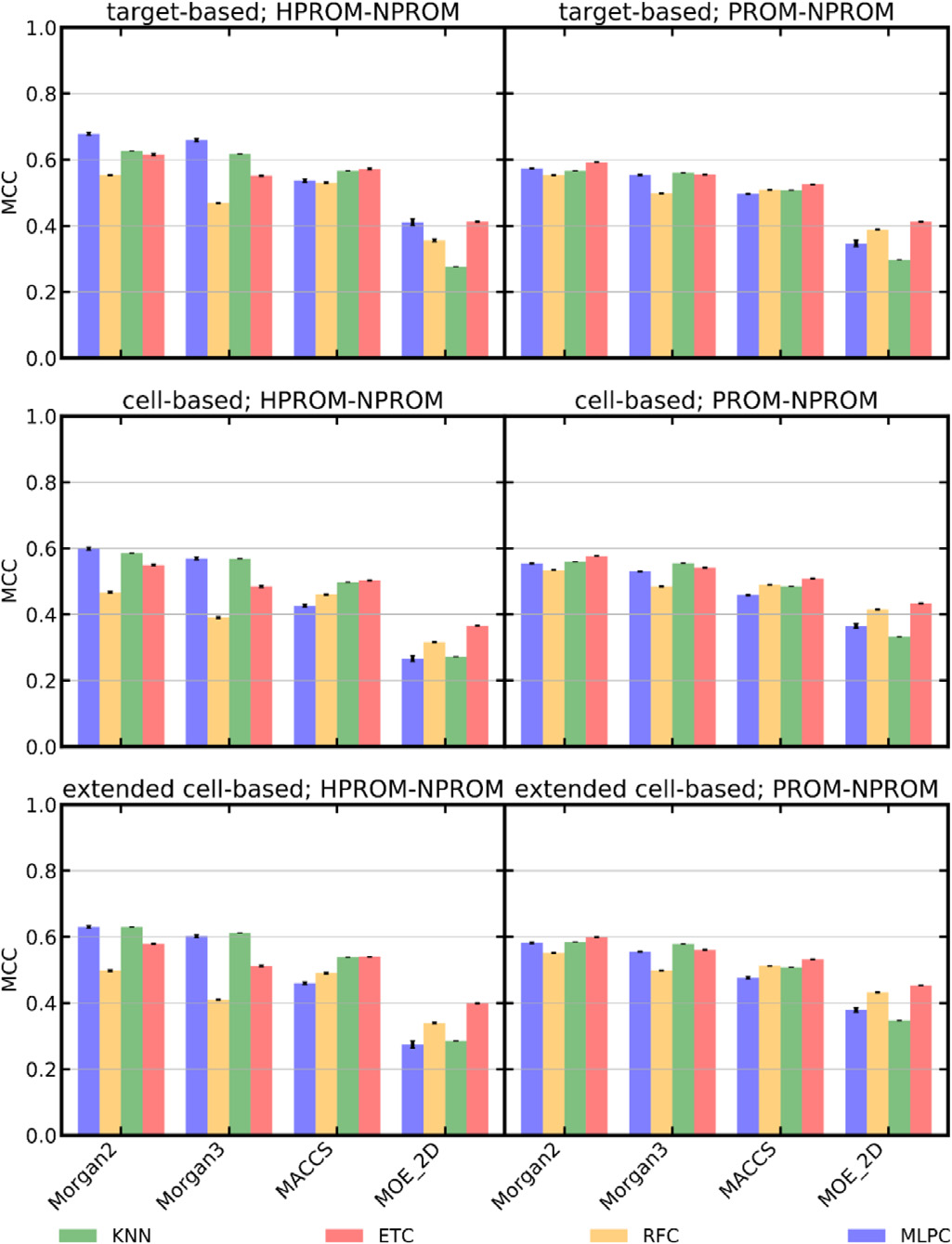
50%ofthecompoundsintheChEMBLdatabaserepresentedbyatleast onecompoundintherespectivetrainingsetwithaTanimotocoefficient of0.5orhigher.

ThePrincipalComponentAnalysis(PCA)scatterplotspresentedin Fig.5showthattheareasinchemicalspacethataremostdenselypop-ulatedwiththecompoundsfromtheChEMBL23databasearealsowell representedbytheassaydatasetsusedformodeltraining.However, thereareasignificantnumberofcompoundsincludedintheChEMBL databasethatarechemicallydistinctfromthoserepresentedbythe trainingsets.TheseareinparticularcompoundswithPC1valuesgreater than10,whichaccountfor2.5%ofthetotalnumberofcompounds oftheChEMBLdatabase.Visualinspectionofthesecompoundsreveals thattheyareunusuallylarge,withmolecularweightbetween575and 900Da.

Thetarget-basedandthecell-basedassaydatasets(trainingdata only)haveanoverlapof180,278compounds(representing75%ofthe target-basedand72%ofthecell-basedassaydataset,respectively). Only13,045(7%)ofthesecompoundshavecontradictingpromiscuity labels(withHPROMtreatedasasubsetofPROM).Atfirstsightthelevel ofagreementbetweenreadoutsfromtarget-basedandcell-basedassays

6

*C.Stork,N.MathaiandJ.Kirchmair*



**Fig.6.**Performance(quantifiedasMCC)ofmachinelearningmodelstrainedon differenttypesofdescriptors.Thevarianceofthetenexperiments(eachusinga distinctrandomseedbetween42to51;seeMaterialsandMethodsfordetails) isindicatedbyerrorbars.

ET,RF,MLP)andfoursetsofdescriptors(i.e.Morgan2andMorgan3fin-gerprints,eachof1024bitsinlength,MACCSkeys,andthecompleteset of2062DphysicochemicalpropertydescriptorsimplementedinMOE, referredtoas“MOE\_2D”)withina10-foldcross-validationframework. Foreachsetuptenofthesecross-validationexperimentswereperformed usingdistinctrandomseeds.Thisallowed,foreachsetup,thecalcula-tionofastandarddeviationthatisindependentofthecross-validation. AsshowninFig.6,thetaskofdiscriminatingHPROMfromNPROM compounds(MCCsofupto0.679)issimplerthanthatofdiscriminat-ingPROMfromNPROMcompoundsacrossthethreeassaydatasets (theMCCofthebestHPROM-NPROMclassifier,0.679,issignificantly higherthanthatofthebestPROM-NPROMclassifier,0.599;p-value 2.48× 10−12).ThisisexpectedbecauseofthelargerATRmarginbe-tweentheHPROMandtheNPROMclass(marginof3*𝜎*)thanbetween thePROMandtheNPROMclass(marginof1*𝜎*).Nosubstantialdiffer-encesinmodelperformancewereobservedwithrespecttothetypeof assaymodeled:thebestsetupsyieldedcomparableMCCsforthetarget-basedassayset(MCCs0.679and0.592forHPROM-NPROMandPROM-NPROMclassification,respectively),cell-basedassayset(MCCs0.602 and0.577,respectively),andextendedcell-basedassayset(MCCs0.631 and0.599,respectively).

Thedifferencesinmodelperformancethatcanbeattributedtothe modelalgorithmsarerathersmall,onaverage0.104inMCC.Themax-imumdifferenceinMCCobservedforanymodeltrainedonidentical input(i.e.samedatasetandsamedescriptorset)was0.224.Overall, theMLPclassifiersperformedbestinHPROM-NPROMclassification(the MCCofthebestMLPclassifier,0.679,issignificantlyhigherthanthat ofthesecond-bestmodel,aKNNmodelthatobtainedanMCCof0.630;

7

*C.Stork,N.MathaiandJ.Kirchmair*  *ArtificialIntelligenceintheLifeSciences1(2021)100007*

**Table6**   
Overviewofhyperparametersoptimizedduringgridsearchwithina10-foldcross-validationframework.1

|  |  |  |
| --- | --- | --- |
| Classifier | Parameter | Values |
| KNN  RF,  ET  MLP | n\_neighbors(numberofneighborsconsidered)  n\_estimators(numberoftrees)  max\_features(featurestakenintoaccountforbestsplitsearch) hidden\_layer\_sizes(numberofperceptronsperlayer)2  hidden\_layer\_sizes(numberhiddenlayer)2  activation(activationfunction) | 1,**3**,5,10  50,100,**200**,300,400,500 ‘sqrt’,‘none’,**‘0.2’**,‘0.4’,‘0.6’,‘0.8’50,100,**250**,500  **1**,2,3,4,5 **‘relu’**,‘tanh’,‘logistic’ |

1Thehyperparametervaluesindicatedinboldarethoseweidentifiedasmostsuitableformodelbuilding.These valueswereusedforthegenerationofthefinalmodels.

2hidden\_layer\_sizesacceptstwovalues:oneforthenumberofperceptronsperlayerandoneforthenumberof hiddenlayers.

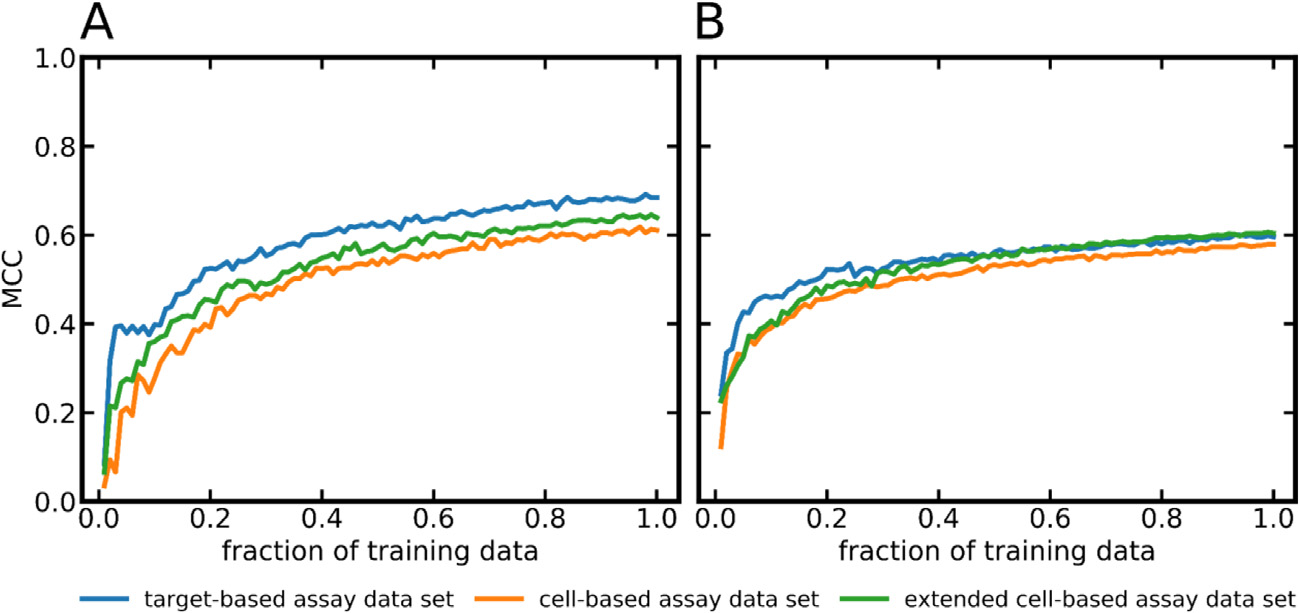
**Table7**   
Cross-validationandtestsetperformanceofthebestmodelsofdifferenttypes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data | Classification | Machine  learning  algorithm | Cross-validationperformance1 | Testsetperformance |

MCC2AUC2BalancedaccuracySensitivitySpecificityMCCAUCBalancedaccuracySensitivitySpecificity

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| target-basedassaydata set | HPROM-NPROM | KNN  ET  RF  MLP  KNN  ET  RF  MLP  KNN  ET  RF  MLP  KNN  ET  RF  MLP  KNN  ET  RF  MLP  KNN  ET  RF  MLP | 0.6240.8430.733 0.6300.9640.734 0.5880.9640.695 0.6860.9460.796 0.5870.8440.745 0.5970.9280.746 0.5780.9290.718 0.5990.9070.777 0.5710.8270.704 0.5720.9500.697 0.5140.9470.651 0.6110.9290.754 0.5660.8450.74  0.5930.9250.747 0.5720.9250.717 0.5790.9100.77  0.6000.8420.721 0.5990.9560.708 0.5370.9560.662 0.6390.9390.764 0.5860.8540.749 0.6180.9350.757 0.5900.9340.725 0.6070.9210.783 | 0.469  0.469  0.392  0.595  0.506  0.504  0.445  0.578  0.41  0.395  0.303  0.512  0.501  0.511  0.445  0.571  0.443  0.417  0.325  0.532  0.516  0.529  0.459  0.593 | 0.998  0.998  0.999  0.997  0.984  0.986  0.991  0.975  0.998  0.998  0.999  0.996  0.979  0.983  0.989  0.969  0.998  0.999  0.999  0.997  0.981  0.985  0.990  0.972 | 0.3760.9090.871 0.5080.9660.677 0.4840.9650.677 0.6480.9490.798 0.4120.8640.816 0.5180.9100.721 0.5180.9080.731 0.5800.8990.768 0.3380.8990.857 0.5310.9400.692 0.5200.9320.692 0.5760.9150.767 0.4130.8600.809 0.5510.9110.743 0.5430.9100.747 0.5610.9010.764 0.3400.8950.858 0.5270.9440.683 0.5190.9430.686 0.5670.9210.753 0.4290.8710.819 0.5650.9230.742 0.5540.9200.748 0.5870.9100.781 | 0.818  0.357  0.358  0.601  0.822  0.464  0.489  0.562  0.812  0.387  0.387  0.541  0.834  0.513  0.525  0.562  0.798  0.368  0.374  0.511  0.835  0.506  0.523  0.596 | 0.924  0.997  0.996  0.995  0.809  0.977  0.973  0.974  0.902  0.997  0.996  0.992  0.783  0.973  0.968  0.965  0.919  0.998  0.997  0.994  0.804  0.978  0.973  0.967 |
| PROM-NPROM | |
| cell-basedassaydatasetHPROM-NPROM | |
| PROM-NPROM | |
| extendedcell-based assaydataset | HPROM-NPROM |
| PROM-NPROM | |

1TheoptimizedhyperparametersarereportedinTable6.   
2ThevarianceisreportedinTableSI\_3.



**Fig.**  **7.**Performance (quantified as MCC) of the

hyperparameter-optimizedMLPclassifierstrainedonthe

target-based,cell-basedandextendedcell-basedassaydata

setasafunctionoftrainingsetsize.(A)HPROM-NPROMclas-

sifiers,(B)PROM-NPROMclassifiers.Foreachdatapointthe

variancewascalculatedfromtencalculations(withdifferent

randomseeds;seeMaterialsandMethodsfordetails).Because

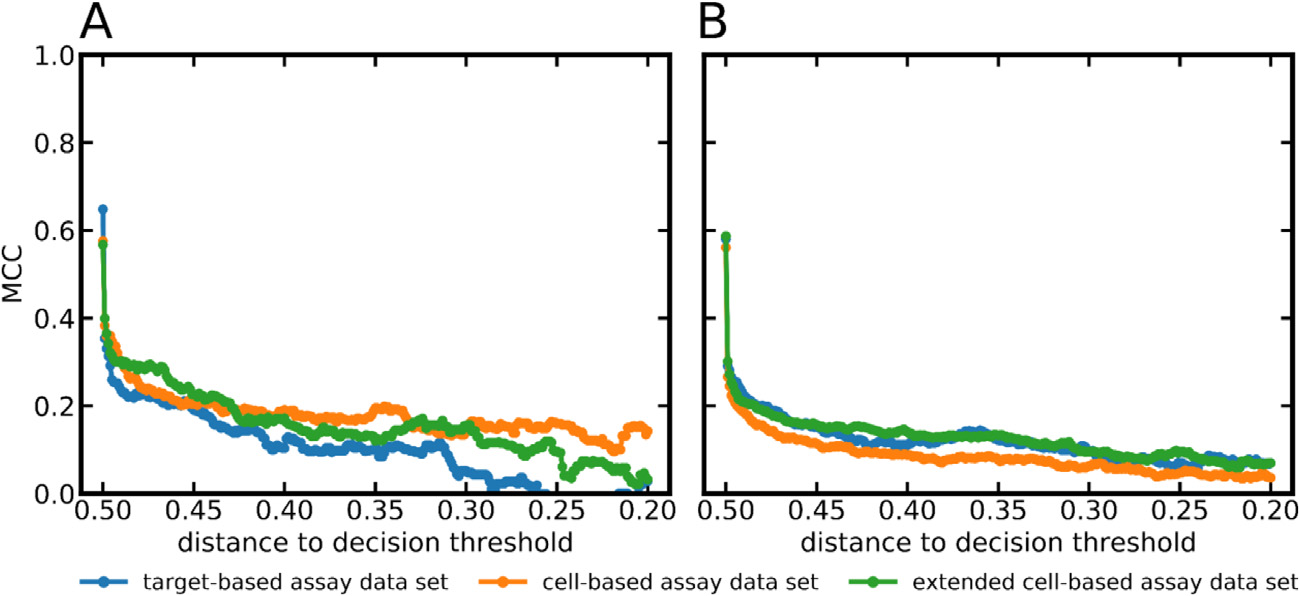
thevariancevalueswerewithintherangeof1.4× 10−6to

5.9× 10−4theyarenotvisualizedinthesegraphs.

|  |  |
| --- | --- |
| distinctareasinthechemicalspaceandhencecontributetotheexten-sionoftheapplicabilitydomainofthemodel.  *Evaluationofthefinalmachinelearningmodels*  Atotalof24finalmodelsofdifferenttypes(i.e.modelstrained onthefulltrainingset,balancedwithSMOTE;seeMaterialsand | Methodsfordetails)weretestedontheholdoutdata(i.e.10%of thedatathatwassetasidepriortomodelbuilding).The24mod-elsresultfromthecombinationofthreedifferenttrainingsets(i.e. target-based,cell-basedandextendedcell-basedassaydataset),four machinelearningalgorithms(KNN,ET,RF,MLP),andtwodiffer-enttypesofclassification(i.e.HPROM-NPROMandPROM-NPROM). AllofthesemodelsarebuiltonMorgan2fingerprintsandutilizethe |

8

*C.Stork,N.MathaiandJ.Kirchmair*



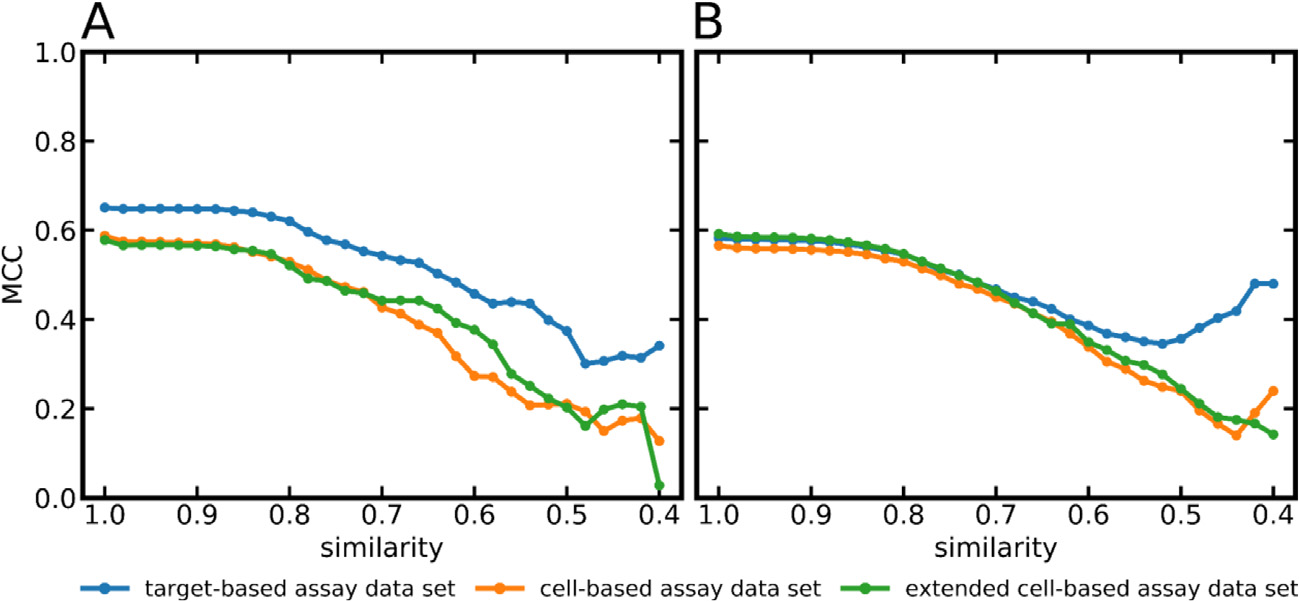
*ArtificialIntelligenceintheLifeSciences1(2021)100007*

**Fig.8.**Performance(quantifiedasMCC)of(A)theHPROM-NPROMMLPclassifiersand(B)thePROM-NPROMMLPclassi-fiersasafunctionofthedistanceofthepredictedclassproba-bilitytothedecisionthreshold(thedecisionthresholdapplied toallmodelsinthisstudyis0.5).Foreachdatapointthevari-ancewasestimatedbyrunningthemodelsontenrandomly selectedsubsetsofthetestdata(seeMaterialsandMethodsfor details).Becausethevariancevalueswerewithintherangeof 9.6× 10−6to4.5× 10−3theyarenotvisualizedinthesegraphs.

|  |  |
| --- | --- |
| hyperparameterssetsoptimizedduringthepreviouscross-validation experiments.  *Modelperformanceonthetestset*   TheaverageMCCobtainedbythe24modelsontherespectivetest setswas0.507,whichis0.087lowerthaninthecross-validationsce-nario(Table7).Overall,thedecreaseinperformance(onthetestset comparedtocross-validation)wasmorepronouncedfortheHPROM-NPROMclassifiers(averagedeclineinMCC0.103)thanthePROM-NPROMclassifiers(averagedecline0.070).Thesteeperdropinperfor-manceobservedfortheHPROM-NPROMclassifiersislikelyrelatedto thefactthatthenumberofcompoundsrepresentingtheactiveclassis muchlowerfortheHPROM-NPROMtrainingset(approximately5000 compounds)thanforthePROM-NPROMtrainingset(approximately 23,000compounds).ThebestMCCamongallHPROM-NPROMclas-sifierswasobtainedbytheMLPclassifiertrainedonthetarget-based assaydataset(MCC0.648).Thebest-performingPROM-NPROMclassi-fierwastheMLPclassifiertrainedontheextendedcell-basedassaydata set(MCC0.587).  Importantly,asubstantialdecreaseinperformancewasobservedfor theHPROM-NPROMKNNclassifiersforallthreeassaydatasets(for example,theKNNclassifierofthetarget-basedassaydataset;three nearestneighbors;cross-validationMCC0.624;testsetMCC0.376)and alsothePROM-NPROMKNNclassifiersforallthreeassaydatasets(for example,theKNNclassifierofthetarget-basedassaydataset;three nearestneighbors;cross-validationMCC0.587;testsetMCC0.421).This declineintheperformancemayberelatedtomodeloverfitting.  IncontrasttotheobservationsmadefortheKNN,theMCCvalues obtainedbytheRFandETclassifiersremainedstable.Themaximum declineinMCCobservedforthesemodelswas0.122.TheMLPclassifiers showedthemostrobustperformanceacrossthethreedatasetsandthe twotypesofclassifications(i.e.HPROM-NPROMandPROM-NPROM), withamaximumdeclineinMCCof0.072.ForthisreasonthesesixMLP classifierswereselectedtoformtheHitDexter3setofmachinelearning modelsandtheywereinvestigatedfurtherregardingtheirapplicability domains.  *Predictionsuccessasafunctionofthedistanceofthepredictedclass probabilityfromthedecisionthreshold*   Commonly,adirectlyproportionalrelationshipisobservedbetween thereliabilityofclassassignmentsandthedistancesbetweenthepre-dictedclassprobabilitiesandthedecisionthreshold.Thisholdstruealso fortheHitDexter3models.Fig.8showsthatclassassignmentsbased onpredictedprobabilitiescloseto0orcloseto1.0(thiscorrespondsto adistancetothedecisionthresholdofapproximately0.5asweapply adecisionthresholdof0.5inallcases)areparticularlyreliable(MCC valuesofupto0.648)fortheHitDexter3models.TheMLPclassifiers differentiatingPROMandNPROMcompoundsforthethreedatasetsre-portpredictedclassprobabilitiesgreaterthan0.95orsmallerthan0.05 foronaverage97%ofthecompoundsinthetestset. | *Predictionsuccessasafunctionofthedistanceofthetestcompoundstothe trainingset*   Itisexpectedthattestcompoundsthatarestructurallydissimilar fromthoserepresentedbythetrainingdataposegreaterchallengesto themodelthanthosethatarestructurallyrelated.Fig.9showsthatthe HitDexter3modelsperformwellforcompoundsrepresentedbyatleast onemoleculeinthetrainingsetthatisstructurallyrelated(i.e.havingat leastonecompoundinthetrainingsetforwhichthepairwiseTanimoto coefficientbasedonMorgan2fingerprintsisatleast0.7).Predictionsfor compoundsthataremoredissimilartothoserepresentedinthetrain-ingdataarelessreliableandshouldbeconsideredwiththenecessary caution.  *Predictionsuccessasafunctionoftheapplieddecisionthreshold*   Inthecurrentcontext,thedecisionthresholdappliedtoaclassifier decidesonwhenacompoundisclassifiedasafrequenthitterorasa non-promiscuouscompound.Thedefaultvalueforthedecisionthresh-oldis0.5.Therearesomeusecaseswhereadifferentdecisionthreshold maybepreferred.Forexample,incaseswherethedetectionoffrequent hittersisapriority(i.e.prioritizationofsensitivityoverspecificity),a lowerdecisionthresholdmayresultinbetterpredictions.Fig.10vi-sualizestheeffectsofchangesinthedecisionthresholdontheMCC, balancedaccuracy,sensitivityandspecificity.Thefactthatthecurves remainfairlystableuntilthedecisionthresholdapproachesextremeval-ues(i.e.valuescloseto0.0or1.0)indicatesthattheclassifiersproduce clearpredictionsformostcompounds.Incaseswheresensitivityisof primaryimportance,usersareadvisedtoconsideranycompoundswith predictedprobabilitiesgreaterthan0.0aspotentialfrequenthitters.  *Predictingfrequenthittersincell-basedassayswithmodelstrainedondata fromtarget-basedassaysandviceversa*   Compoundsmaybehavedifferentlyintarget-basedandcell-based assays,inparticularalsowithregardtotheirassayinterferenceandfre-quenthitterbehavior.Inordertoobtainabetterunderstandingofthe relevanceandvalueofdedicatedmodelsforthepredictionoffrequent hittersintarget-basedandcell-basedassays,wecomparedtheperfor-manceoftheMLPclassifiersontestdataofthesameassaydomainto theirperformanceontestdataoftheotherassaydomain(i.e.classifiers trainedontarget-basedassaydataweretestedoncell-basedassaytest setandviceversa).  AsshowninFig.11,thePROM-NPROMMLPclassifierstrainedand testedondatafromthesameassaydomainclearlyoutperformedthose trainedontheotherdomain.Thegraphsalsoindicatethatthedifference inperformanceisnottheresultofdifferencesinthechemicalspacecov-eredbytheindividualdatasets:evenfortestcompoundsthatarestruc-turallycloselyrelatedtothoserepresentedbythetrainingset,models trainedontarget-basedassaydatadonotperformwelloncell-based assaydataandviceversa.  Forthecell-basedassaytestdata,theMCCofthePROM-NPROM MLPclassifiertrainedontarget-basedassaydatawasjust0.189(vs. |

9

*C.Stork,N.MathaiandJ.Kirchmair*



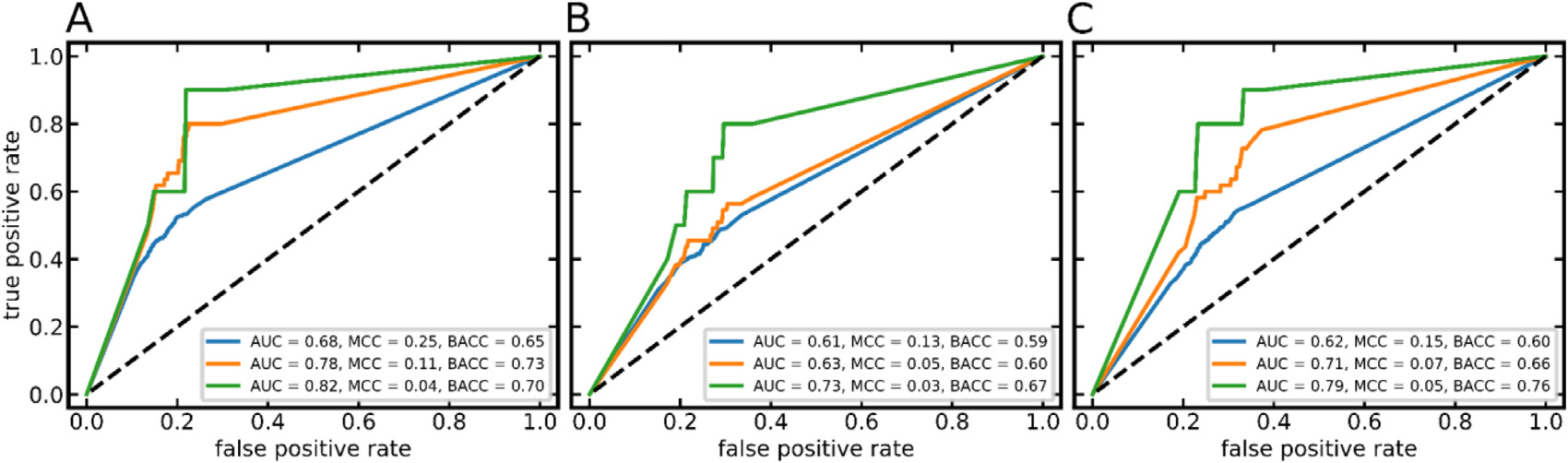
*ArtificialIntelligenceintheLifeSciences1(2021)100007*

**Fig.9.**Performance(quantifiedasMCC)of(A)theHPROM-NPROMMLPclassifiersand(B)thePROM-NPROMMLPclassi-fiersasafunctionofthestructuralsimilarity(measuredasTan-imotosimilarityonMorgan2fingerprints)betweenthecom-poundsinthetestandthetrainingsets.Foreachdatapoint thevariancewasestimatedbyrunningthemodelsontenran-domlyselectedsubsetsofthetestdata(seeMaterialsandMeth-odsfordetails).Becausethevariancevalueswerewithinthe rangeof1.9× 10−6to2.9× 10−3theyarenotvisualizedin thesegraphs.

|  |  |
| --- | --- |
| **Fig.10.**PerformanceoftheHitDexter3modelsasafunctionoftheselected decisionthreshold.  0.561fortheclassifiertrainedonthecell-basedassaydata;anycom-poundspresentinthetrainingandinthetestsetaredisregardedinthe calculationoftheseMCCvalues).Likewise,forthetarget-basedassay testdatatheMCCforthePROM-NPROMMLPclassifiertrainedoncell-basedassaydatawasjust0.235(vs.0.580fortheclassifiertrainedon thetarget-basedassaydata).Theseresultsshowthattarget-basedand cell-basedassaysclearlybehavedifferentlyandthatdedicatedmodels arerequiredtoadequatelypredicttheirbehavior.  *Modelperformanceondarkchemicalmatter*   WetestedtheHitDexter3modelsalsoonthedarkchemicalmat-ter(DCM)datasetcompiledbyWassermannetal.TheDCMdataset consistsof135,489compoundswhichhavebeentestedinatleast100 target-basedandcell-basedassayswithoutasinglepositiveassayout- | **Fig.11.**Performance(quantifiedasMCC)ofMLPclassifiersasafunctionof thepairwisesimilaritybetweenthetestcompoundanditsnearestneighborin thetrainingset(measuredasTanimotocoefficientderivedfromMorgan2fin-gerprints).(A)HPROM-NPROMMLPclassifiertrainedonthetarget-basedassay dataset,(B)PROM-NPROMMLPclassifiertrainedonthetarget-basedassaydata set,(C)HPROM-NPROMMLPclassifiertrainedonthecell-basedassaydataset, (D)PROM-NPROMclassifiertrainedonthecell-basedassaydataset.Foreach datapointthevariancewasestimatedbyrunningthemodelsontenrandomly selectedsubsetsofthetestdata(seeMaterialsandMethodsfordetails).Because thevariancevalueswerewithintherangeof9.2× 10−6to6.2× 10−3theyare notvisualizedinthesegraphs.  come.Thesecompoundsarenotnecessarilywithoutactivityonanypro-teinbuttheyareunlikelyfrequenthitters.  InthetestoftheHitDexter3modelsontheDCMdataset,any testcompoundsalsopresentinthetrainingsetoftherespectivemod-elsweredisregarded(leaving24,111to37,711DCMcompoundsfor testing,dependingontheindividualtrainingset).Thetarget-based, cell-basedandextendedcell-basedHPROM-NPROMMLPmodelscor-rectlyassigned99.0%,98.6%and98.7%oftheDCMcompoundstothe NPROMclass.Incomparison,thepercentageofcorrectassignmentsof thePROM-NPROMmodelswere95.4%,93.7%and93.6%,respectively. Thisresultcorroboratesthevalidity(inparticularthespecificity)ofthe models.  *Modelperformanceonknownbadactors*   TotestthecapacityofthesixHitDexter3modelstoidentifybadac-torsinbiologicalassays,weranthemodelsontworecentlypublished |

10

*C.Stork,N.MathaiandJ.Kirchmair*  *ArtificialIntelligenceintheLifeSciences1(2021)100007*



**Fig.12.**ROCcurvesobtainedwiththeHitDexter3PROM-NPROMclassifierstrainedon(A)target-basedassaydata,(B)cell-basedassaydata,and(C)extended

cell-basedassaydata,andtestedonthedatasetofBorreletal.ThecompoundsofthetestsetwereannotatedasfrequenthittersaccordingtoDefinition1(blue

curves),Definition2(orangecurves)andDefinition3(greencurves).

datasetscontainingexperimentallyconfirmedbadactors(cave:badac-torsarenotnecessarilyfrequenthitters;HitDexter3isdesignedtoiden-tifyfrequenthitters).Asinallpreviousexperiments,wedisregardedall compoundspresentinthesetestsetsthatarealsopartofthetraining dataoftheindividualmodels.

ThefirstdatasetisfromtheworkofDahlinetal.[37].Thisdataset consistsof1139compoundsthatareknowntocausefalsereadoutsin varioustypesofbiologicalassays.Forthe891to1002testcompounds notrepresentedinthetrainingsetoftheindividualmodels,thetarget-based,cell-basedandextendedcell-basedHPROM-NPROMMLPclassi-fiersassigned24.1%,25.5%and23.0%ofallcompoundstotheHPROM class.ThemodelsdistinguishingPROMandNPROMcompoundsflagged 40.3%,39.3%and40.3%aspromiscuous,respectively.Becausebadac-torsarenotnecessarilyfrequenthitters(andviceversa),thepercentages ofcompoundsreportedbyourmodelsasPROMorHPROMarewithin theexpectedrange.

TheseconddatasetisfromtheworkofBorreletal.[22].Thisdata setcontains8947compounds,891ofwhichhavebeenobservedtocause falsepositivereadoutsinbioluminescenceassaysduetoluciferaseinhi-

bition(inoneoutofoneassay)orautofluorescence(inoneorseveral outof24assays),and8056compoundsthatareconfirmedtobehave benignintheseassays.Weexploredthreewaysoftranslatingthemea-surementsrecordedwiththeseinterferenceassaysinto“frequenthitter data"”:Compoundswerelabeledasfrequenthitteriftheyproduced

**Definition1.**afalse-positivesignalinatleastoneassay(luciferaseas-sayorassaytotestforautofluorescence).

**Definition2.**afalse-positivesignalintheluciferaseassayANDatleast oneofthe(24)assaysetupstotestforautofluorescence.

**Definition3.**afalse-positivesignalintheluciferaseassayANDatleast nineofthe(24)assaysetupstotestforautofluorescence.

Allothercompoundswerelabeledasnon-promiscuous.

AsshowninFig.12,theHitDexter3modelsreachedAUCvalues ofupto0.82(PROM-NPROMMLPclassifiertrainedonthetarget-based assaydataset,incombinationwithDefinition3),whichconfirmsthe abilityofthemodelstorankbadactorsearlyinarank-orderedlistof compounds.TheMCCandbalancedaccuracyindicatemoderateperfor-

**Table8**   
PerformanceoftheHitDexter2.0andHitDexter3machinelearningmodelsontheDCMdatasetsandtheknownbadactorsdatasetofDhalinetal.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| HitDexterversion | trainingset | testset | classification | numberof  compounds  intestset1 | numberof  compounds | fractionof  compounds |

correctlyclassifiedasDCMorbadactors4

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| HitDexter2.0 HitDexter2.0 HitDexter3  HitDexter3  HitDexter3  HitDexter2.0 HitDexter2.0 HitDexter3  HitDexter3  HitDexter3  HitDexter2.0 HitDexter2.0 HitDexter3  HitDexter3  HitDexter3  HitDexter2.0 HitDexter2.0 HitDexter3  HitDexter3  HitDexter3 | PSAdata2  CDRAdata3  target-basedassaydata  cell-basedassaydata  extendedcell-basedassaydata PSAdata2  CDRAdata3  target-basedassaydata  cell-basedassaydata  extendedcell-basedassaydata PSAdata2  CDRAdata3  target-basedassaydata  cell-basedassaydata  extendedcell-basedassaydata PSAdata2  CDRAdata3  target-basedassaydata  cell-basedassaydata  extendedcell-basedassaydata | DCM  DCM  DCM  DCM  DCM  DCM  DCM  DCM  DCM  DCM  KnownBadActors[37] KnownBadActors[37] KnownBadActors[37] KnownBadActors[37] KnownBadActors[37] KnownBadActors[37] KnownBadActors[37] KnownBadActors[37] KnownBadActors[37] KnownBadActors[37] | HPROM-NPROM HPROM-NPROM HPROM-NPROM HPROM-NPROM HPROM-NPROM PROM-NPROM  PROM-NPROM  PROM-NPROM PROM-NPROM PROM-NPROM HPROM-NPROM HPROM-NPROM HPROM-NPROM HPROM-NPROM HPROM-NPROM PROM-NPROM  PROM-NPROM  PROM-NPROM  PROM-NPROM  PROM-NPROM | 20,894  42,567  37,711  30,967  24,327  20,872  41,587  37,421  30,875  24,111  974  963  1002  987  965  910  896  941  906  891 | 20,806  42,341  37,317  30,529  24,015  20,472  40,080  35,695  28,942  22,572  140  140  241  252  222  304  330  379  356  359 | 0.996  0.995  0.990  0.986  0.987  0.981  0.964  0.954  0.937  0.936  0.144  0.145  0.241  0.255  0.230  0.334  0.368  0.403  0.393  0.403 |

1Anycompoundspresentinthetrainingsetwereremovedfromthetestset.

2Primaryscreeningassaydata.

3Confirmatorydose-responseassaydata.

4NotethatHitDextermodelsarenotdesignedtoidentifyalldifferentkindsofbadactorsbutrathertoidentifyfrequenthitters(ofwhichasignificantportionare infactbadactors).

11

*C.Stork,N.MathaiandJ.Kirchmair*

mancebut,again,HitDexter3isdesignedtopredictfrequenthitters, anditcanbeexpectedthatasubstantialproportionofthecompounds observedtocausefalse-positivereadoutsintheseinterferenceassays willbehavebenigninotherassaytypesandsetups.

*ModelperformancecomparedtoHitDexter2.0*   
 Thesetofmachinelearningmodelsdevelopedinthisworktoform HitDexter3differfromtheHitDexter2.0modelsinseveralways.For HitDexter3,

•dedicatedmodelsfortarget-basedandcell-basedassaysweredevel-opedwhereastheprevioussetofmodelsonlycovertarget-based assays.

•fourdifferentmachinelearningalgorithms(KNN,ET,RFandMLP) insteadofjustETwereexplored.ThisledtothefindingthatMLP classifiersperformbest.

•theminimumnumberofdatapointsrequiredtocalculatetheATR hasbeenincreasedfrom50to100.ThisresultsinmorerobustATRs (basedonwhichtheclasslabels,i.e.HPROM,PROMandNPROM, areassigned).

Giventhefactthatthetrainingandtestsetsutilizedforthedevel-opmentandvalidationoftheHitDexter3andHitDexter2.0models differ,a1:1comparisonofmodelperformanceisdifficult.Formodels ofthesametype(e.g.HPROM-NPROMclassifier),differencesinMCCs onthetestdatawereintherangeof-0.035to+0.015(cell-basedmod-elsnotincludedastheyarenotavailableinHitDexter2.0).Alsoonthe DCMdatasets(theDCMdatasetsusedfortestingdifferintheircom-positionbecauseoftheremovalofanycompoundsthatarealsopresent inthetrainingsetoftherespectivemodel),themodelsbehavesimi-larly,withtheHitDexter3PROM-NPROMclassifier(fortarget-based assays)assigning5%tothePROMclassandtherespectiveHitDexter 2.0modelsassigning2%to4%oftheDCMcompoundstothePROM class(Table8).Onthesetofknownbadactors[37],thepercentageof compoundspredictedasfrequenthittersis40%forHitDexter3(PROM-NPROMclassifiertrainedontarget-basedassaydata)and33%to37% forHitDexter2.0(PROM-NPROMclassifiers;Table8).

Overall,theseresultsindicatethattheperformanceoftheHitDexter 3andHitDexter2.0machinelearningmodelsiscomparable.TheHit Dexter3modelsperformabitbetteronthesetofknownbadactors. Finally,theadditionofdedicatedmodelsforpredictingacompound’s behaviorincell-basedassaysisanimportantadvantageofHitDexter3 overHitDexter2.0.

**Conclusions**

Inthisworkwepresentthedevelopment,refinementandvalidation ofnewmodelsforthepredictionoffrequenthittersinbiologicalassays. Themodelsaretrainedonamanuallycuratedassaydatasetthatwe extractedfromthePubChemBioassaydatabaseand,forthefirsttime, thesemodelscovercell-basedassaysinadditiontotarget-basedassays. Furtheradditionsincludetheexplorationoffoursetsofdescriptorswith additionalmachinelearningalgorithmssuchasKNNandMLP,andthe useofmorerobustATRs(calculatednowonaminimumof100distinct assayscomparedto50previously).

TheMLPclassifiersturnedouttoobtainthebestclassificationper-formanceandrobustnessinmostcases,withMCCsofupto0.648indis-criminatingHPROMfromNPROMcompounds,andMCCsofupto0.580 indiscriminatingPROMfromNPROMcompounds.Usecasesthatre-quiremodelswithhighsensitivityorhighspecificitycanbeapproached byadjustingthedecisionthresholdappliedinclassification.

TestsoftheMLPclassifiersonDCMcompoundsandsetsofknown badactorscorroborategoodperformanceofthemodels:themodelscor-rectlyidentified94to99%ofallcompoundsoftheDCMdatasetas non-promiscuousandflaggedupto40%oftheknownbadactorsas frequenthitters(becausebadactorsarenotnecessarilyfrequenthitters thisnumberisinlinewiththeexpectationsforagoodmodel).

12

*C.Stork,N.MathaiandJ.Kirchmair*

[[16]BaellJB,HollowayGA.Newsubstructurefilter](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0016)s[forremovalofpanassayinterference compounds(PAINS)fromscreeninglibrariesandfortheirexclusioninbioassays.J MedChem2010;53:2719–40.](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0016)

[[17]BaellJ,WaltersMA.Chemistry:chemicalconartistsfoildrugdiscovery.Nature](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0016)  [2014;513:481–3.](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0017)

[[18]BaellJB,NissinkJWM.Sevenyearitch:pan-assayinterferencecompounds(PAINS)](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0017)  [in2017-utilityandlimitations.ACSChemBiol2018;13:36–44.](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0018)

[[19]M.Koptelov,A.Zimmermann,P.Bonnet,R.Bureau,B.Crémilleux.PrePeP:atool](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0018) fortheidentificationandcharacterizationofpanassayinterferencecompounds.Pro-ceedingsofthe24thACMSIGKDDinternationalconferenceonknowledgediscovery [&datamining,p.462–71.](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0020)

[[20]GhoshD,KochU,HadianK,SattlerM,TetkoIV.Luciferaseadvisor:high-accu-racymodeltoflagfalsepositivehitsinLuciferaseHTSassays.JChemInfModel 2018;58:933–42.](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0020)

[[21]YangZ-Y,DongJ,YangZ-J,LuA-P,HouT-J,CaoD-S.Structuralanalysisand](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0020) [identificationoffalsepositivehitsinLuciferase-basedassays.JChemInfModel](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0021) [2020;60:2031–43.](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0022)

[[22]BorrelA,HuangR,SakamuruS,XiaM,SimeonovA,MansouriK,etal.High-through-](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0022) [putscreeningtopredictchemical-assayinterference.SciRep2020;10:3986.](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0022)

[[23]MNissinkJW,BlackburnS.Quantificatio](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0022)n[offrequent-hitterbehaviorbasedonhis-](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0022) [toricalhigh-throughputscreeningdata.FutureMedChem2014;6:1113–26.](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0023)

[[24]YangJJ,UrsuO,LipinskiCA,SklarLA,OpreaTI,BologaCG.Badapple:promiscuity](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0023)  [patternsfromno](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0024)[isyevidence.JCheminform2016;8:29.](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0025)

[[25]StorkC,ChenY,ŠíchoM,KirchmairJ.HitDexter2.0:machine-learningmodelsfor](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0025)  [thepr](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0026)[e](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0025)[dictionoffreque](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0026)[n](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0025)[thitters.JChemI](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0026)[nfModel2019;59:1030–43.](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0025)

[[26]](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0025)[KimS,ChenJ,ChengT,GindulyteA,HeJ,](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0026)[HeS,etal.PubChemin2021:newdata](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0025) [contentandimprovedwebinterfaces](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0026).[NucleicAcidsRes2021;49:D1388–95.](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0026)

[[27]FeldmannC,YonchevD,StumpfeD,BajorathJ.Systematicdataanalysisanddi-](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0026)[agnosticmachinelearnin](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0027)g[revealdifferencesbetweencompoundswithsingle-and](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0027) [multitargetactivity.Mol](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0028)Pharm[2020;17:4652–66.](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0028)

[[28]KimS,ThiessenPA,BoltonEE,ChenJ,FuG,GindulyteA,etal.PubChemsubstance](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0028)  [andcompounddatabases.NucleicAcidsRes2016;44:D1202–13.](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0028)

[[29]KimS,ChenJ,ChengT,GindulyteA,HeJ,HeS,etal.PubChem2019update:](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0028)  [improvedacces](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0029)s[tochemicaldata.NucleicAcidsRes2019;47:D1102–9.](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0029)

[[30]PubChem.https://pubchem.ncbi.nlm.nih.gov/(accessedJune3,2021).](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0029)

*ArtificialIntelligenceintheLifeSciences1(2021)100007*

[31][NCBIResourceCoordinatorsDatabaseresourcesoftheNationalCenterforBiotech-](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0031) nology[Information](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0032).[NucleicAcidsRes2016;44:D7–19.](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0032)

[32][KimS,ThiessenPA,ChengT,YuB,BoltonEE.AnupdateonPUG-REST:RESTfulin-](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0032)terface[forprogrammaticaccesstoPubChem.NucleicAcidsRes2018;46:W563–70.](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0032) [33][PatríciaBentoA,HerseyA,FélixE,LandrumG,GaultonA,AtkinsonF,etal.An](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0032) [opensourcechemicalstructurecurationpipelineusingRDKit.JCheminform 2020;12:1–16.](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0033)

[34][RDKit:Open-sourcecheminformatics;http://www.rdkit.org/(accessedJune3,](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0033)  2021).

[35][ChEMBL23.http://www.ebi.ac.uk/chembl/(accessedJune3,2021).](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0036)

[36][WassermannAM,LounkineE,HoepfnerD,LeGoff](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0036) G[,KingFJ,StuderC,etal.Dark chemicalmatterasapromisingstartingpointfordrugleaddiscovery.NatChem Biol2015;11:958–66.](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0036)

[37][DahlinJL,AuldDS,RothenaignerI,HaneyS,SextonJZ,NissinkJWM,etal.Nuisance](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0036)  [compoundsincellularassays.CellChemBiol2021;28:356–70.](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0037)

[38][NCICADDGroup,NationalCancerInstitu](https://cactus.nci.nih.gov/chemical/structure)[te.ChemicalIdentifie](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0037)r[Resolver.](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0037)  [https://cactus.nci.nih.gov/chemical/structure(accessedJuly26,2021).](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0039)

[39][RogersD,HahnM.Extended-connectivityfingerprints](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0039).[JChemInfModel](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0039)  [2010;50:742–54.](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0039)

[40][MorganHL.Thegenerationofauniquemachinedescriptionforchemicalstructures-](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0039)at[echniquedevelopeda](https://doi.org/10.1021/c160017a018)tchemicalabstractsservice.JDoc1965;5:107–13. [doi](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0041)[:10.1021/c160017a018.](https://doi.org/10.1021/c160017a018)

[41][Coc](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0041)[kPJA,AntaoT,Chang](https://doi.org/10.1021/c160017a018)[JT,ChapmanBA,CoxCJ,DalkeA,etal.Biopython: freelyavailablePythontoolsforcomputationalmolecularbiologyandbioinformat-](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0041)[ics.Bioinformatics2009;25:1422–3.](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0042)

[42][FuL,NiuB,ZhuZ,WuS,LiW.CD-HIT:acceleratedforclusteringthenext-generation](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0042)  [sequencingdata.Bioinformatics2012;28:3150–2.](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0042)

[43][GarretaR,MoncecchiG.Learningscikit-learn:machinelearninginPython.Packt](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0042)  [PublishingLtd;2013.](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0043)

[44][https://www.chemcomp.com/MOE-Molecular\_Operating\_Environment.htm.](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0043) [45][ChawlaNV,BowyerKW,HallLO,KegelmeyerWP.SMOTE:syntheticminori](https://www.chemcomp.com/MOE-Molecular_Operating_Environment.htm)[t](https://doi.org/10.1613/jair.953)yover- [samplingtechnique.JArtifIntellRes2002;16:321–57.doi](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0046)[:10.1613/jair.953.](https://doi.org/10.1613/jair.953)

[46][StorkC,WagnerJ,FriedrichN-O,deBruynKopsC,Šícho](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0046)[M,KirchmairJ.H](https://doi.org/10.1613/jair.953)[itDex-ter:amachine-learningmodelforthepredictionoffrequenthitters.ChemMedChem 2018;13:564–71.](http://refhub.elsevier.com/S2667-3185(21)00007-6/sbref0046)

13