

[Artificia](https://doi.org/10.1016/j.ailsci.2022.100049)l[IntelligenceintheLifeSciences2(2022)100049](https://doi.org/10.1016/j.ailsci.2022.100049)

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

LIDeBTools:ALatinAmericanresourceoffreelyavailable,open-source cheminformaticsapps   
DenisN.PradaGori,Lucas N.Alberca,SantiagoRodriguez,Juan I.Alice,Manuel A.Llanos, CarolinaL.Bellera,AlanTalevi∗  
*LaboratoryofBioactiveCompoundsResearchandDevelopment(LIDeB),FacultyofExactSciences,UniversityofLaPlata(UNLP),LaPlata,BuenosAires,Argentina*

|  |  |  |
| --- | --- | --- |
| article | info | abstract |
| *Keywords:*  Cheminformatics  Opensource  Webapplications  Clustering  Decoygeneration  Decoys  Druggabilityprediction Druggability |  | Cheminformaticsisthechemicalfieldthatdealswiththestorage,retrieval,analysisandmanipulationofan increasingvolumeofavailablechemicaldata,anditplaysafundamentalroleinthefieldsofdrugdiscovery, biology,chemistry,andbiochemistry.Opensourceandfreelyavailablecheminformaticstoolsnotonlycontribute tothegenerationofpublicknowledge,butalsotoreducethetechnologicalgapbetweenhigh-andlow-tomiddle-incomecountries.Here,wedescribeaseriesofin-housecheminfor[maticsapplicationsdevelopedb](https://lideb.biol.unlp.edu.ar/)youracademic drugdiscoveryteam,whicharefreelyavailableonourwebsite(<https://lideb.biol.unlp.edu.ar/>)asWebApps andstand-aloneversions.Theseappsincludetoolsforclusteringsm[allmolecules,decoygeneratio](https://lideb.biol.unlp.edu.ar/)n,druggability assessment,classificatorymodelevaluation,anddatastandardizationandvisualization. |

**1.Introduction**

Cheminformaticsdescribestheuseofinformationtechnologytohan-dlechemicalinformation[1].Thefielditselfhasbeenintegratedwith thechemicalsciencesformanydecades.However,thetermwascoined relativelyrecently,whentheincreasingamountofchemicaldatagen-eratedwithinthedrugdiscoveryfield(e.g.,duetotheimplementa-tionofcombinatorialchemistryandhigh-throughputscreeningplat-forms)madetheuseofchemicalinformationtechnologiesincreasingly mandatory[2].Althoughcheminformaticshasawidescope,coretasks withinthecheminformaticsfieldincludethemanagementofchemical databasesanddatasets,storageandretrievalofchemicalinformation, structure-propertyrelationships,insilicoscreening,andthedesignof combinatorialandfocusedlibraries.

Asinotherareasrelatedtoinformatics,manyrelevantcheminfor-maticssoftwareweredevelopedundertheopen-sourcephilosophy.The open-sourceparadigmbeganintheinformaticsfieldandisintrinsically relatedtothenotionsofcollaborativeresearchandpublicknowledge. Itwaslaterpartiallyadoptedinotherfields,forexample,inthephar-maceuticalsector[3,4],whenitwasrealizedhowefficientthecollab-orativemodelmightbecomeinrelationtotraditionalclose-doorsand market-drivenphilosophies.Theopen-sourceapproximationhasspecial relevancetobridgetechnologicalandscientificgapsbetweenlow-and high-incomecountriesandtoaddressneglectedneedsfromthepoorest regionsoftheglobe.Thisiswellreflectedinourorganization:LIDeB (LaboratoryofBioactiveCompoundsResearchandDevelopment)isa

∗Corresponding[author.](mailto:alantalevi@gmail.com)

*E-mailaddress:*[alantalevi@gmail.com](mailto:alantalevi@gmail.com)(A.Talevi).

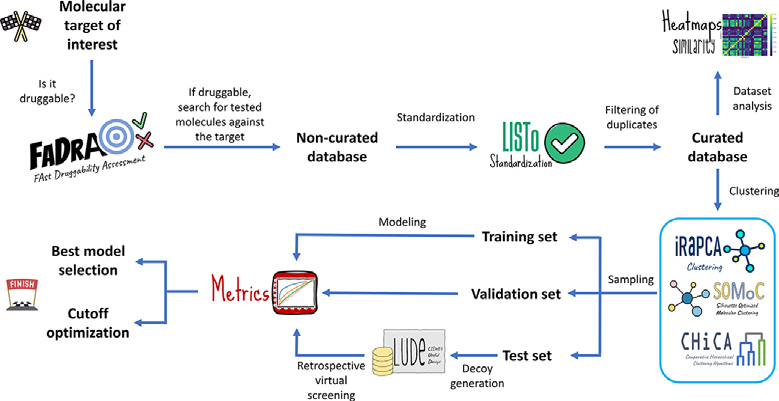
smallacademicresearchcenterdependentontheUniversityofLaPlata (UNLP,Argentina)withafocusondrugdiscoveryprojects,andapartic-ularinterestinthefieldofinfectioustropicaldiseases.Aspartofthese, wehavedevelopedpubliclyaccessiblecheminformaticsWebAppsde-ployedthroughtheStreamlitframeworkandstandalonesourcecodes. Here,weprovideanover[viewoftheseresources,which](https://lideb.biol.unlp.edu.ar/)arefreelyac-cessibleonourwebsiteat<https://lideb.biol.unlp.edu.ar/>.Theseinclude clustering,decoygenerat[ion,druggabilitypredictiona](https://lideb.biol.unlp.edu.ar/)pps,andother secondaryresourcesrelatedtochemicaldatastandardizationandvisu-alization.

Severaldatascienceandmachinelearningplatformshavebeende-velopedthusfar,enablinguserswithlimitedcodingabilitytocreate pipelinesfordataexploration,analysis,andmining.Somewell-known examplesincludeKNIME,PipelinePilot,andAlteryx.Althoughtheyof-ferfreeandopen-sourcedistributions,mostadvancedfeaturesorthird-partyapplicationsareonlyavailableundercommerciallicenses,which sometimesmakestheminaccessibletosmallresearchunitsindevelop-ingcountries.

AllapplicationswithintheLIDeBToolssuiteweredeployedasWeb AppsontheStreamlitplatform,soscientistscanusethemthrougha user-friendlyinterfaceusingcomputationalresourcesallocatedtothe cloud.Alternatively,theirstandalonedistributionsarewritteninplain Python,underanObject-OrientedProgramming(OOP)paradigmand usingopen-sourcelibraries,offeringahigherlevelofcustomizationand codereusability.Moreover,theycaneasilybepluggedintoexisting chemoinformaticpipelines.

<https://doi.org/10.1016/j.ailsci.2022.100049>  
[Received23September2022;Receivedinrev](https://doi.org/10.1016/j.ailsci.2022.100049)isedform24November2022;Accepted1December2022   
Availableonline2December2022   
2667-3185/© 2022TheAuthors.PublishedbyElsevierB.V.ThisisanopenaccessarticleundertheCCBYlicense(<http://creativecommons.org/licenses/by/4.0/>)

*D.N.P.Gori,L.N.Alberca,S.Rodriguezetal.*



**Fig.1.**Aschemeofadrugdiscoverymachinelearningcampaign,fromassess-mentoftargetdruggabilitytoretrospectivevirtualscreeningforprotocolvali-dation.Possiblestepswhereourreportedopen-sourcetoolscanbeincorporated areshown.

Ageneralworkflowofamachinelearning-baseddrugdiscovery projectisshowninFig.1,whichindicatesthepartoftheworkflow inwhicheachofourtoolsmightbeintegrated.

**2.Clusteringtools**

Clusteringofsmallmoleculesimpliessortingacollectionofchem-icalstructuresintosmallergroups(clusters)suchthatthemolecules withinaclusterdisplayahighdegreeofsimilarity(within-clustersimi-larity)comparedtotheirsimilaritytoelementsallocatedtoothergroups (between-clustersimilarity).Ideally,clusteringshouldprovidecompact clustersfarawayfromeachotherwithintherelevantchemicalspace[5]. Aclusteringstrategyrequiresasensibledecisionregardingthecluster-ingalgorithmandparametersettingstoobtainagoodapproximationof anoptimalsolution.Clustersfoundusingaclusteringprocedureoffer ahypothesisofthepossiblegroupswithinthedatabecausetheground truth(ifexistent)isgenerallyunknown[6].

LIDeBToolsincludetwoin-houseclusteringapproximations,theit-erativeRandomsubspacePrincipalComponentAnalysis(iRaPCA)ap-proachandSilhouette-OptimizedMolecularClustering(SOMoC),along withanimplementationofsomeclassicalagglomerativeclusteringal-gorithms,ComparativeHierarchicalClusteringAlgorithms(CHiCA), whichallowsaninteractiveexplorationoftheoutputdendrogramsfor aninformedchoiceofthemostsuitableclusteringmethodandtheop-timalleveltocutthetree,dependingontheuser’sneeds.

*2.1.iRaPCA*

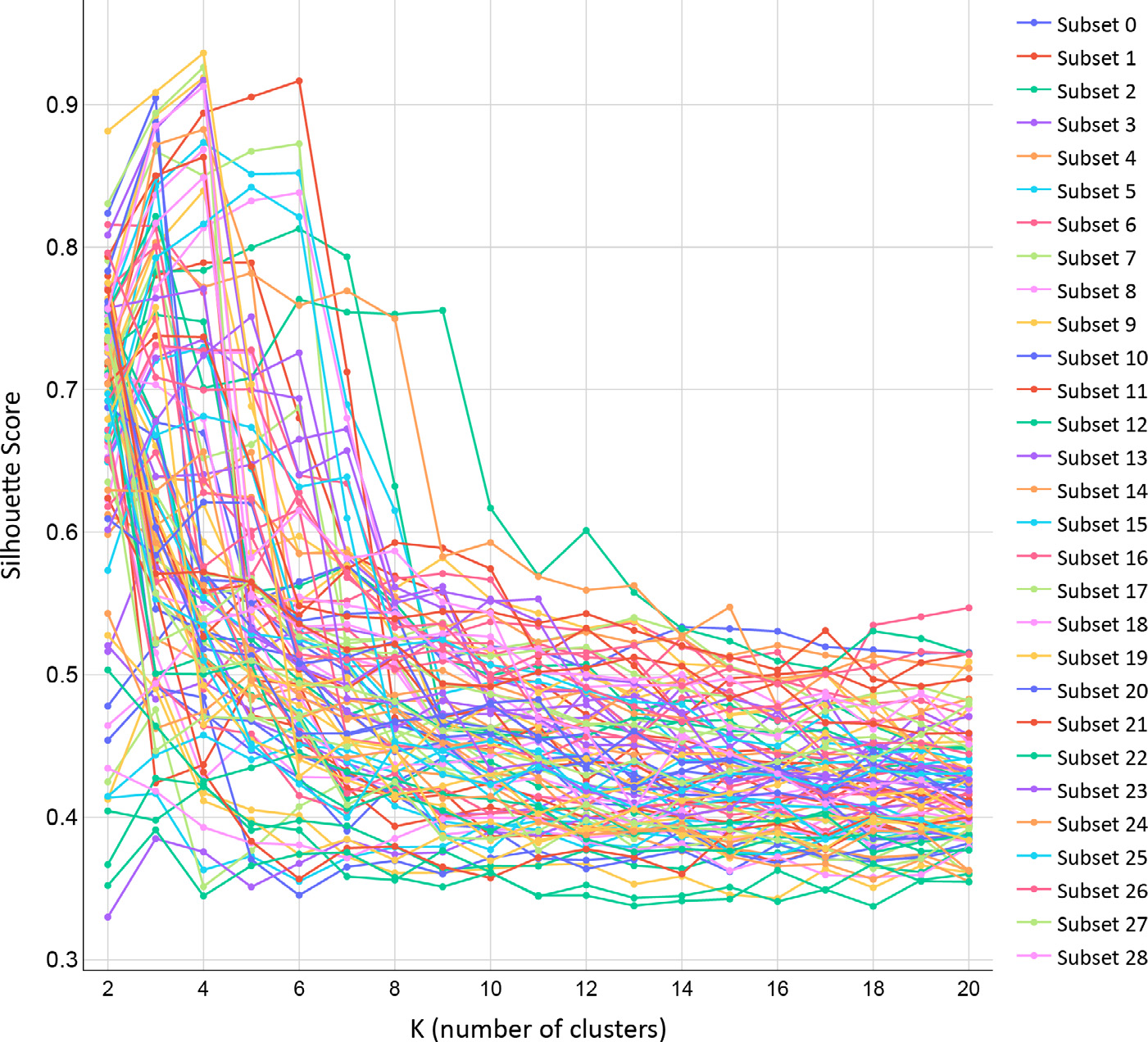
iRaPCAclusteringisbasedonaniterativecombinationofrandom subspace(featurebagging),dimensionalityreductionbyPrincipalCom-ponentAnalysis(PCA)[7]andtheK-meansalgorithm[8].TheiRaPCA algorithmcanbedescribedasasequentialcombinationoffoursteps: *inputandencoding,dimensionalityreduction,actualclustering*and,option-ally,*iteration*.ToadaptiRaPCAtoaparticularuser’sneeds,thealgo-rithmincludescustomizablehyperparameters.

Inthe*inputstep*,themoleculestobeclusteredshouldbeprovidedas a.csvfile,whereeachlinecorrespondstoamoleculeinSMILESformat. AfteroptionalstandardizationemployingMolVS[9],Mordred[10]is usedtocompute1613moleculardescriptorsforeachmolecule.Alter-natively,usersmayprovidetheirownpoolofmoleculardescriptorsas tab-delimited.txtfile.

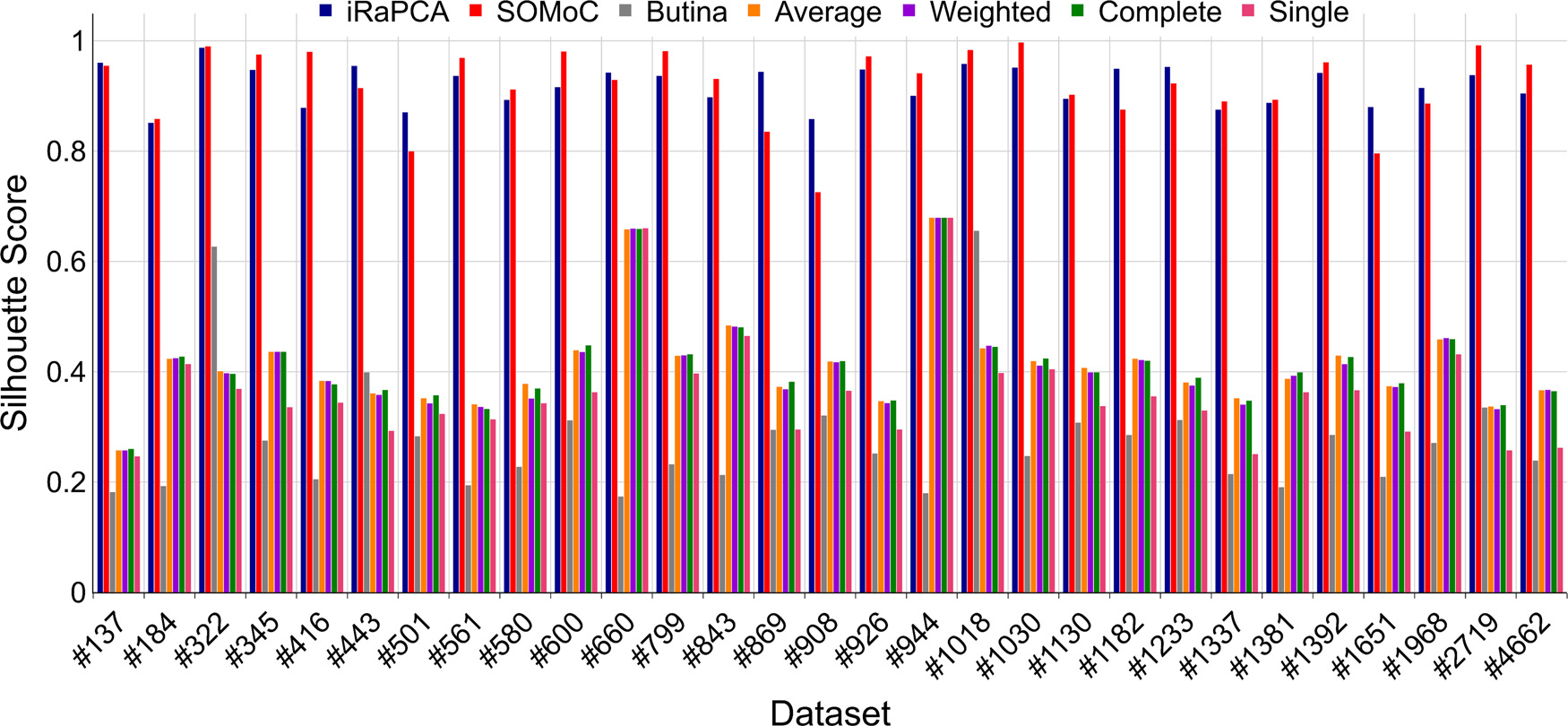
Subsequently,the*dimensionalityreduction*stepstartswiththere-movalofdescriptorswithlowinformationcontent,and100random subsetsof200descriptorsarecreatedusingarandomsubspace/feature baggingapproximation[11],followedbytheremovalofhighlycorre-lateddescriptorswithineachsubset.Importantly,theusermaydefine ahighernumberofrandomsubsetsinthestandaloneversionofthe script.Subsequently,PCAisconductedoneachsubset;bydefault,only thefirsttwoprincipalcomponentsareconsideredforthenextstepsof theprocedure.Inthefinalclusteringstep,theK-meansalgorithmisap-pliedtothePCspacebysystematicallyvaryingthenumberofclusters

2

*D.N.P.Gori,L.N.Alberca,S.Rodriguezetal.*



*ArtificialIntelligenceintheLifeSciences2(2022)100049* **Fig.2.**Exampleofsilhouettescorevs.Kplotgener-atedbyiRaPCA,forthedefault100randomlychosen descriptorsubsets.



**Fig.3.**PerformancecomparisonoftheiRaPCAandSoMOCalgorithmsversusButinaandseveralagglomerativehierarchicalclusteringalgorithmsacross29bench-markdatasets.Thesewereorderedandlabeledinthefigureaccordingtotheirsize(i.e.,bythenumberofcompoundsineachdataset).

*encoding,dimensionalityreduction,*and*clustering*.Itshyperparameters,as inthecaseofiRaPCA,canalsobecustomized.

Inthe*inputstep*,themoleculesareprovidedagaininSMILESnota-tionasa.csvfile,butinsteadofcalculatingmoleculardescriptors,the molecularrepresentationsareencodedintoEState1molecularfinger-prints,calculatedbyRDKit,withafixed-lengthfingerprintcontaining79 features[25].TheUMAPalgorithm,anonlineardimensionalityreduc-tiontechniquebasedonRiemanniangeometryandalgebraictopology, isthenappliedtoreducethehyperdimensionalspaceofthefingerprints, retainingthemostinformativefeaturesoftheprocess.Thesizeofthe embeddedspaceisdeterminedbasedonthesizeofthedataset.Inthe

3

|  |  |
| --- | --- |
| *D.N.P.Gori,L.N.Alberca,S.Rodriguezetal.* | *ArtificialIntelligenceintheLifeSciences2(2022)100049* **Fig.4.**Exampleofadendrogramgeneratedby CHiCA. |

*2.3.CHiCA*

CHiCAimplementsdiverseclassicalhierarchicalagglomerativeclus-teringapproaches.Despiteexhibitingmodestperformanceincompari-sontoiRaPCAandSOMoC,asshowninFig.3,CHiCAallowstheuser, byusinginteractivegraphs,anintuitiveandvisualchoiceofthenum-berofclusterstoconsider.MoleculesarecharacterizedbyMorganfin-gerprintswithuser-definedradioandlength(bydefault,2and1024, respectively)[26].Itispossibletoincludespecial,fuzzyfeaturesinthe fingerprints,suchashydrogen-bonddonors,acceptors,aromaticrings, halogenatoms,andacidicorbasicgroups.Asinourotherclustering apps,theinputmoleculesareprovidedinSMILESnotationthrough.csv files,andadistancematrixisobtained.CHiCAallowsthemeasurement ofthepairwisedistancebetweenmoleculesusingdifferentmetrics(e.g., Jaccard,Euclidean,Cosine)byemployingSciPy[20].TheSciPy’sclus-teringpackageisthenusedtoperformhierarchicalclustering.Different agglomerativemethodsareavailableforthispurpose.Bydefault,single-linkageisused.

Theclustersareplottedinaninteractivedendrogram(Fig.4)that displaysthecompositionofeachclusterbydrawingthelinksbetween anon-singletonclusteranditschildren.Thetopofalinkindicatesa clusteranditstwolegsspecifywhichclustershavebeenmerged,while thelengthofthelegsreflectsthedistancebetweenthechildclusters. Thisdendrogramallowstheusertovisuallyselectadistancecutoff for clusteringmoleculesbasedontheirdistributioninthegeneratedclusters andthenumberofoutliers.

Additionally,aninteractivescatterplotofdistancevssilhouette scorevsthenumberofmoleculesisshown(Fig.5).Thisplotallows theusertoselectanoptimalcutoff forclusteringbasedonthisscore orthenumberofoutliers,numberofsmallclusters(definedasclusters withmorethanonemoleculeandlessthan5%ofthetotalmolecules by)ornumberofdenseclusters(definedasclusterswithmorethan5% ofthetotalmoleculesbydefault).

Oncetheoptimalcutoff hasbeenidentified,CHiCAisre-runwiththis distancevalue,generatingfouroutputfiles:*Clustering\_assignations.csv* whichspecifieswhichmoleculeshavebeenassignedtowhichcluster accordingtothechosencutoff value;*Cluster\_distribution.csv*,whichspeci-fiesthenumberofobtainedclustersandtheirsizes;*Validations.csv*which containsthevaluesoftheCVIs;and*Settings.csv*,whichcontainsthese-lectedparametersintherun.

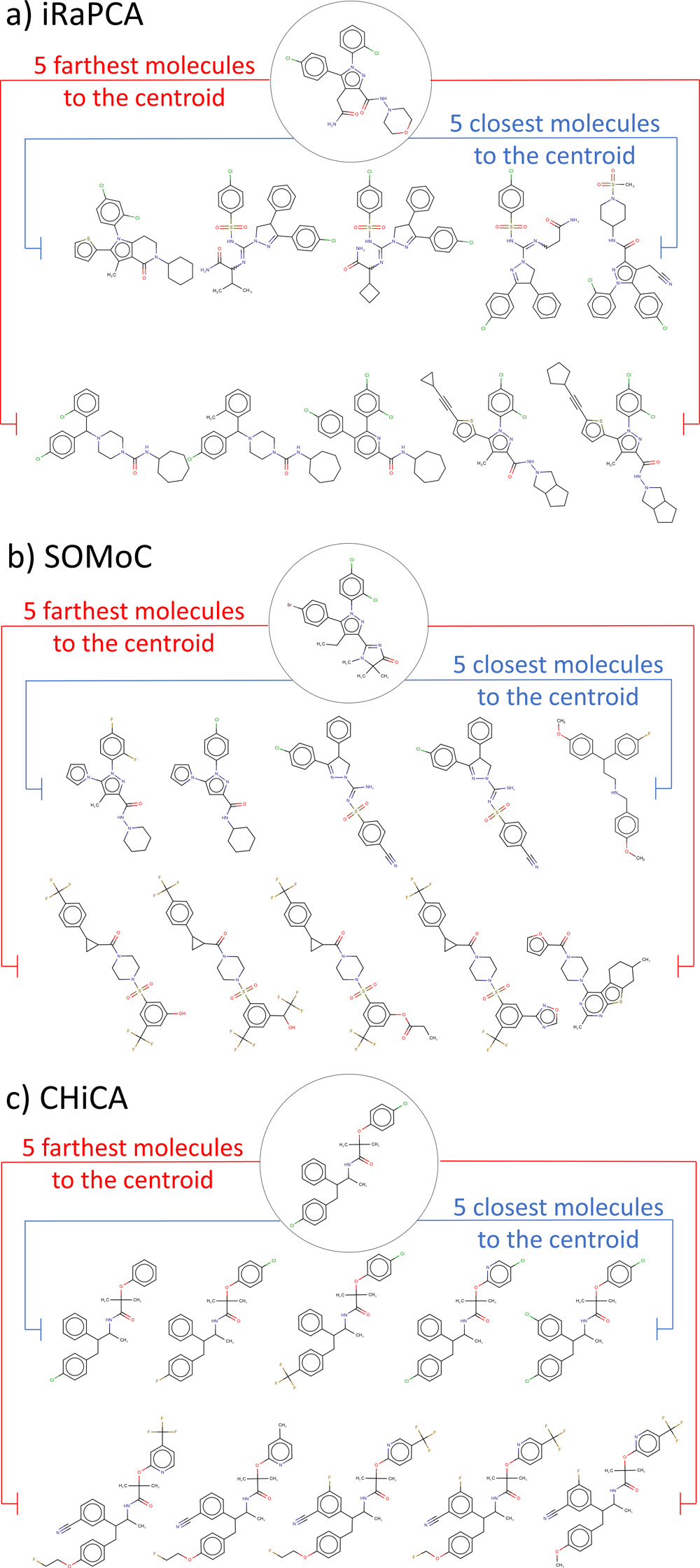
Themolecularstructuresofthegroupcentroidforthemostpopu-latedclusterobtainedusingiRaPCA,SOMoCandCHiCAforthehuman cannabinoidreceptor1dataset(Hum\_can\_rec)areshowninFig.6.

**3.Decoygeneration**

Anyvirtual(or,forthecase,wet)screenisconfrontedwithanin-trinsicclassimbalance:thenumberofactivecompoundsinachemical libraryissignificantlyexceededbythenumberofinactivecompounds.

4

*D.N.P.Gori,L.N.Alberca,S.Rodriguezetal.*



**Fig.6.**Thefigureshowsthecentroidsandtheclosestandfarthestneigh-borsfromthecentroids(withinthemostpopulatedcluster)obtainedforthe“Hum\_can\_rec” datasetusingiRaPCA,SOMoCandCHiCA(weightedlinkage) methods.

senonsometheoreticalbasistoassurearathersmallprobabilityofmis-labeling;theDirectoryofUsefulDecoysEnhanced(DUD-E)[30],with arecentlyreleaseimprovedversion[31]ispossiblythemostpopular onlinedecoy-generationtooltovalidatestructure-basedscreeningpro-tocols,althoughsomeotherdecoygeneratorshavealsobeenreported, includingmethodsthatrelyonhardmachinelearningapproaches[32–34].Althoughundoubtedlyuseful,DUD-Ewasattimeslimitedbythe capacityofitsonlineservertoreturnputativedecoysinashorttime, sometimesdemandingseveraldaystocompleteatask(thiswasparticu-

5

|  |  |
| --- | --- |
| *D.N.P.Gori,L.N.Alberca,S.Rodriguezetal.* | *ArtificialIntelligenceintheLifeSciences2(2022)100049*  **Fig.7.**GeneralworkflowforLUDe. |

**Table1**   
NumberofdecoysgeneratedineachstepoftheLUDeworkflowforfiveactivecompoundsintheFABP4datasetofDUD-E.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| QueryMolecules | Selectedbymatching physicochemical  properties | PasstheTcfilter peractive | PassthefMCS filter | Passthe  Frameworkfilter | Non-duplicated decoysper  active: | Passthecross-checkTC filteragainstallqueries submitted |
| **Query1**  **Query2**  **Query3**  **Query4**  **Query5** | 800  1000  1000  1000  800 | 582  641  624  755  642 | 582  640  623  754  622 | 582  640  623  754  622 | 582  625  586  523  515 | 217  286  285  185  182 |

Tc:Tanimotocoefficient.

Afterperformingtheworkflowforallquerymolecules,themolec-ularfingerprintsoftheresultinglistofdecoysarecomparedtothose ofallquerycompounds,andonlydecoyswithaTanimotosimilarity below0.2(bydefault)toanyqueryareretained.Thisensuresthatthe decoysarenotonlydifferentfromthequerytheywerederivedfrom butalsofromanyoftheknownactivecompoundsusedtogenerate them.Upto50decoysperquerycompoundareretrievedbydefault, althoughthisnumbercanalsobecustomized.Thefollowingoutput filesaregenerated:Generated\_decoys.csv,whichcontainsallthede-coysinSMILESnotation;Decoys\_analysis.csv,whichprovidesatable containingasummaryofthenumberofmoleculesthatpassedeachsuc-cessivefilter;andDecoys\_setting.csv,containingasummaryofallthe settingsusedintherun.Asanexample,Table1presentsthenumber ofdecoysobtainedwhenusingtheactivecompoundsincludedinthe FattyAcidBindingProtein4(FABP4)subsetfromDUD-Easqueriesfor LUDe.Thisdatasetcontains47FABP4inhibitors.Onlyfiveofthesehave beenincludedinthetable,forillustrativepurposes.Forinstance,LUDe found800moleculesthatwerephysicochemicallysimilartoquery1. Amongthese,582passedthethreefiltersthatensuretopologicaldis-similarityfromthequery.Ofthese,only217moleculesweredissimilar totheremainingqueries.Bydefault,fiftydecoyswerethenrandomly selectedforquery1,andthesameanalysiswasrepeatedforeveryother query.Itisimportanttoemphasizethatbecause2Dsimilaritymethods areusedtoselectdecoysforeachquery(excludingdecoyswithhigh 2Dsimilaritytoanyquery),LUDedecoysshouldnotbeusedtoassess theperformanceofinsilicoscreensbasedon2Dsimilarityapproxima-tions/molecularfingerprints.Thisisageneralwarningforanymethod thatusesa2Dsimilaritycriteriontoexcludepotentialdecoysfrom thelist.

TheperformanceofLUDewascomparedwiththatofDUD-E across102targetsubsetsofactivecompoundsandtheircorrespond-i[ngDUD-E-generateddecoys,whi](http://dude.docking.org/subsets)chareavailableontheDUD-Ewebsite (<http://dude.docking.org/subsets>).Foreachsetofactivecompounds,a

6

|  |  |
| --- | --- |
| *D.N.P.Gori,L.N.Alberca,S.Rodriguezetal.* | *ArtificialIntelligenceintheLifeSciences2(2022)100049*  **Fig.8.**A)DOEscoresoftheoriginalDUD-Eset (blue)comparedwiththeLUDegeneratedde-coys(red)across102targets.Thetargetswith evenindiceswerenotlabeledonthex-axisow-ingtospacelimitations.B)Meandoppelganger scoresoftheoriginalDUD-Eset(blue)com-paredwiththeLUDegenerateddecoys(red) across102pharmacologicaltargets.Thetargets withevenindicesarenotlabeledonthex-axis owingtospaceconstraints. |

thisscore,FCFP6fingerprintsweregeneratedusingRDKit[19]forev-eryknownactivecompoundanddecoy,andforeachofthe102phar-macologicaltargets;thesimilaritybetweeneachdecoyandeachac-tivecompound(query)wasassessedusingtheTanimotocoefficient. TheDoppelgangerscoreofadecoyisdefinedasthemaximumvalue oftheTanimotocoefficientobtainedinthisway,acrossallactivecom-poundsusedasqueries.Foreachtarget,wereportthemeanDoppel-gangerscoreoveralldecoysandthemaximumstructuralsimilaritybe-tweenanactiveandadecoy.Fig.8BshowsthemeanDoppelganger scoreoveralldecoysforeachofthe102targets.TheDoppelganger scorewaslowerfor85%ofthetargetsforLUDedecoysthanforDUD-E decoys,withanaverageDoppelgangerscoreacrossthetargetsof0.23 forLUDedecoysand0.25forDUD-Edecoys.Thispossiblyreflectsthe effectoftheadditionalfiltersimplementedinLUDetoensuretopolog-icaldissimilaritybetweenthedecoysandqueries(fMCS,Muckoscaf-fold).ThemaximumaverageDoppelgangerscorepertargetwasvery similarforLUDeandDUD-E(0.36and0.37,respectively).Thissug-geststhatthechanceoffalse-negativedecoystendstobereducedfor LUDe.

**4.Druggabilitypredictiontool**

Druggabilityreferstotheabilityofagivenproteintobindwitha highaffinitytosmalldrug-likemolecules[41].Assessingthedrugga-bilityofpotentialpharmacologicaltargetspriortoinitiatingatarget-focuseddrugdiscoverycampaigniscrucial.*FastDruggabilityAssessment* (FaDrA)isadruggabilitypredictionwebapplicationbasedonfourlin-earclassifiersthatcandiscriminatedruggablefromnon-druggabletar-getsfromcompleteproteomeswithinafewminutes,withacceptable accuracy,basedonlyontheproteinsequence.

7

*D.N.P.Gori,L.N.Alberca,S.Rodriguezetal.*

**Table2**   
ResultsoftherandomizationtestandtheLGOcross-validationforthefourbest models.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy(s.d.)1 | Accuracy  Randomizationtest (s.d.)1 | Accuracy  LGO  (s.d.)1 |
| **260**  **361**  **424**  **763** | 0.820(0.066) 0.808(0.069) 0.814(0.068) 0.801(0.069) | 0.499(0.150)  0.500(0.134)  0.503(0.129)  0.499(0.128) | 0.723(0.110) 0.712(0.113) 0.758(0.110) 0.744(0.104) |

1s.d.indicatesStandardDeviation.

**Table3**   
Performanceofthebestmodelsintheexternalvalidation.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy(s.d.)1 | Recall(s.d.)1 | Precision(s.d.)1 |
| **260**  **361**  **424**  **763** | 0.803(0.066) 0.802(0.070) 0.818(0.069) 0.803(0.068) | 0.939(0.059) 0.757(0.108) 0.818(0.095) 0.908(0.072) | 0.744(0.070) 0.839(0.085) 0.825(0.082) 0.756(0.074) |

1s.d.standsforStandardDeviation.

topredictclasslabelsoftheremovedproteins.Theprocedurewasre-peated1000times,witheachofthetrainingsetexamplesremovedat leastonce.

Incontrast,intherandomizationtesttheclasslabelwasrandom-izedacrosstheproteinscomprisingthetrainingset.Thetrainingset withtherandomizedvariablewasusedtotrainnewmodels,fromthe descriptorselectionstep.Thisprocedurewasrepeated1000timesfor eachdescriptorsubset.Randomizedmodelsareexpectedtohavepoor accuracycomparedwithrealmodels.

TheresultsofbothtestsarepresentedinTable2.Theresultssug-gestsomedegreeofoverfittingforeachclassifier,butverylowchances ofspuriouscorrelation(asseenasthemeanaccuracyintherandom-izationtest,practicallyidenticaltotheexpectedaccuracyforrandom classification,i.e.,0.5forbalancedsets).

Theequationsofthefourbestmodelsandthedefinitionsofthede-scriptorsincludedinthemhavebeenincorporatedintheSupplementary Information.Thepredictivepoweroftheselectedmodelswasfurther examinedbycalculatingtheaccuracy,recallandprecisionoverthe66 proteinsequencesthatcomposethetestset(Table3).

Finally,thepredictionsofthefourmodelswerecombinedinameta-classifiertoreducethenumberoffalsepositives,thatis,proteinsthat arepredictedtobedruggablebutareactuallynon-druggable.Themeta-classifierconsidersthepredictionsofthefourindividualmodelsandthe assessmentoftheapplicabilitydomain(AD)usingtheleverageapproach and3d/nasthecut-off value,wheredisthenumberofdescriptorsinthe correspondingmodelandnisthenumberofcompoundsinthetraining set.Finally,meta-classifierpredictionisperformedusingthedecision treeshowninFig.9.

Table4comparestheperformanceofeachindividualclassifier, intermsofaccuracy,withthatofthemeta-classifier.Itcanbeob-servedthatthemeta-classifierobtainsthebestaccuracywithinthenon-druggableclass;itisalsotheschemethatleadstoless“non-conclusive”results.

FaDrAallowsdruggabilitypredictionofcompleteproteomeswithin afewminutes.Theappreceivesa“.fasta” fileasinput.Thisfilemay

*ArtificialIntelligenceintheLifeSciences2(2022)100049*

includeasimplesequence,afewproteinsequences,oracompletepro-teome.Afterreadingthefile,thenecessarydescriptorsforthemodels arecalculatedusingPyBioMedandthefourmodelsareappliedtoeach protein.Inaddition,theapplicabilitydomainisdetermined.Apiechart isdisplayed(Fig.10)showingthepercentageof“druggable” and“non-druggable” proteins,andthepercentageof“non-conclusive” results.Fi-nally,a.CSVfileisgeneratedthatsummarizesthepredictionsforthe fourindividualmodelsandthemeta-classifier,aswellaswhethereach proteinsequencebelongstotheADsofthemodels.Forexample,the completeproteomeof*Escherichiacoli*(GenomeassemblyASM584v2) wasruninFaDrA.4298proteinswereloadedofwhich1978(46%)were predictedas“non-druggable”,2084(48.5%)werepredictedas“drug-gable”,and236(5.5%)generatednon-conclusiveresults.Table5shows thepredictionsforthefirsttenproteinsintheE.coliproteome.

**5.Othertools**

InadditiontothepreviouslydiscussedWebApps,LIDeBToolsalso includethreeothersecondarytoolsthatemployavailablepackages andalgorithmswithminorchangesandareimplementedwithauser-friendlyinterface.TheseareHeatmap-Similarity,LIDeB’sStandardiza-tionTool(LISTo),andMetrics.

TheHeatmap-SimilarityWebAppbuildsaheatmapofmolecular similaritiesusingRDKit.Theseplotsofintermolecularsimilarityallow forfastvisualinspectionofthemoleculardiversityofchemicaldatasets. Theinputsfortheapparetwo.txtfiles(whichcanbeidenticalordif-ferent),whereeachlinecorrespondstoamoleculeinSMILES.Theal-gorithmconstructsasimilaritymatrixbetweentwoloadedsetsofcom-pounds,computedasaTanimotosimilaritycoefficientusingMorgan fingerprintswithauser-definedradius(from1to3,2bydefault)and bitlength(1024bydefault).Thematrixisplottedasaheatmapwith acolorscaleindicatingsimilarity,andtheresultingplotscanbedown-loadedas.pngfiles.

LISToisastandardizationWebAppthathelpstoautomaticallystan-dardizecollectionsofchemicalstructuresthatmaypresentdifferent, non-homogeneousmolecularrepresentations.Thissimplestandardiza-tionisusefulforensuringahomogeneousformatofmoleculesbefore calculatinganyconformation-independentmoleculardescriptor.Inthis processeachmolecule,submittedina.txtfileinSMILESnotation,passes throughaseriesofstandardizationstepsthatareincludedinMolVS.By default,LISToretainsonlytheparentfragmentofthemolecule(the largestorganiccovalentunitinthemolecule),removesallstereochem-icalinformationfromtetrahedralcentersanddoublebonds,mantains theunchargedversionofthefragmentparent,replacesatomswiththe mostabundantisotope,disconnectsmetalsfromorganicatoms,ap-pliesaseriesoftransformationstocorrectcommondrawingerrors, andremoveshydrogenatoms.TheWebAppalsoreturnsthecanoni-caltautomer,whichisauniquerepresentation,selectedwithascor-ingfunctionfromallpossibletautomersthatcouldbegeneratedfrom amolecule;importantly,thecanonicaltautomerisnotnecessarilythe mostenergeticallyfavorable[47].Alogfilewithadditionalinformation regardingpotentialproblemsforspecificstandardizationactionscan bedownloaded.Inaddition,thestandardizedmoleculescanbevisual-izedintheWebAppthroughtheinteractivechemicalviewermols2grid (<https://github.com/cbouy/mols2grid>).

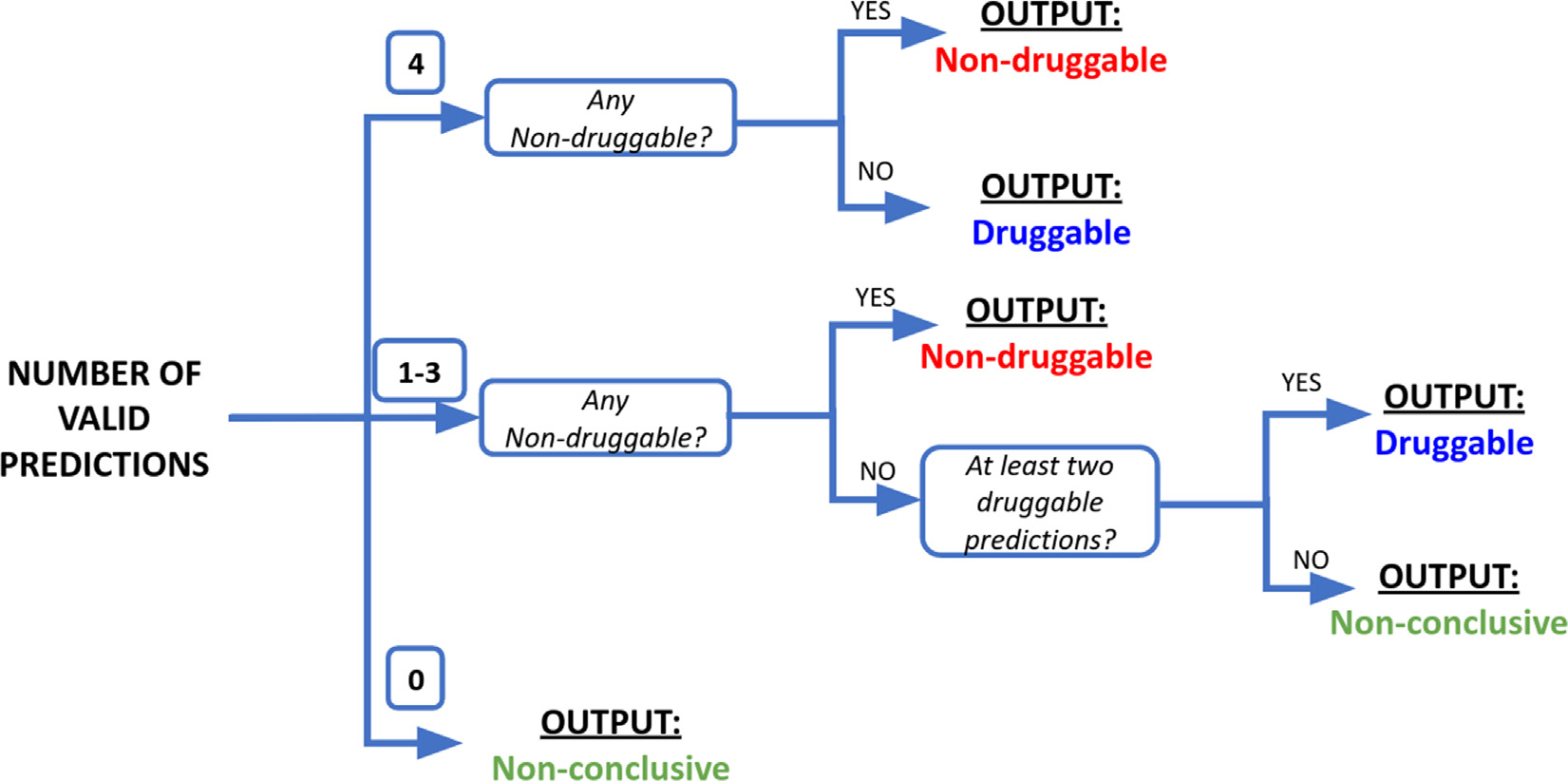
[Finally,theMetricsWebAppeva](https://github.com/cbouy/mols2grid)luatestheperformanceofclas-sificatorymodelsbycalculatingdifferentmetricsimplementedin

**Table4**   
Accuracybycategoryforeachselectedmodelandthemeta-classifierinthetestset.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Model260 | Model361 | Model424 | Model763 | Meta-classifier |
| **Druggableaccuracy**  **Non-druggableaccuracy**  **Numberof“Non-conclusives”** | 0.848  0.593  9 | 0.800  0.700  6 | 0.879  0.643  5 | 0.933  0.667  8 | 0.727  0.900  3 |

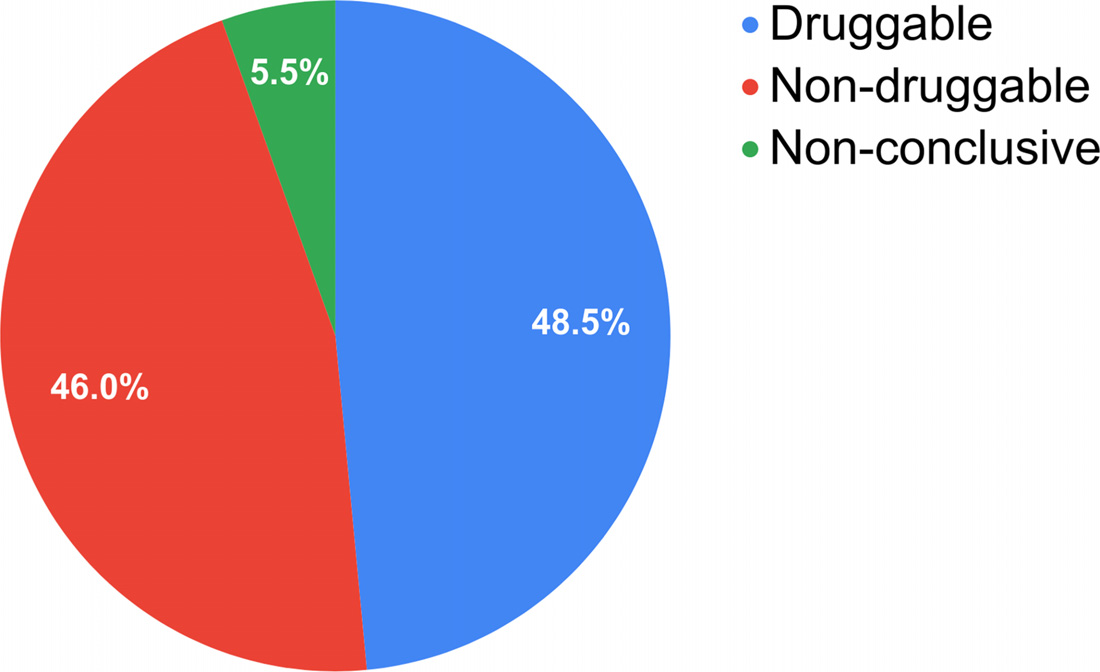
8

*D.N.P.Gori,L.N.Alberca,S.Rodriguezetal.*  *ArtificialIntelligenceintheLifeSciences2(2022)100049*



**Fig.9.**Decisiontreeformeta-classifierdruggabilityassessment.Foreachprotein,thepredictionsoffourindividualclassifiersareprovidedalongwiththeirrespective

ADpredictions.Dependingonhowmanypredictionsareconsideredvalid(i.e.,insidetheAD),theclassassignmentfollowsdifferentcriteria.



**Table5**   
Outputtableforthepredictionofthefirst10proteinsin*E.coli*proteome.

**Fig.10.**PiechartforFaDrApredictionofthe*E.coli*proteome.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ProteinID | PREDICT1 | AD1 | PREDICT2 | AD2 | PREDICT3 | AD3 | PREDICT4 | AD4 | CLASSIFICATION |
| **NP\_414542.1 NP\_414543.1 NP\_414544.1 NP\_414545.1 NP\_414546.1 NP\_414547.1 NP\_414548.1 NP\_414549.1 NP\_414550.1 NP\_414551.1** | Druggable  Druggable  Druggable  Druggable  Non-druggable Non-druggable Druggable  Druggable  Druggable  Druggable | NO  YES  YES  YES  NO  YES  YES  YES  YES  YES | Non-druggable Druggable  Druggable  Druggable  Non-druggable Druggable  Druggable  Druggable  Druggable  Druggable | NO  YES  YES  YES  NO  YES  YES  YES  YES  YES | Non-druggable Druggable  Druggable  Druggable  Druggable  Druggable  Druggable  Druggable  Druggable  Druggable | NO  YES  YES  YES  NO  YES  YES  YES  YES  YES | Druggable  Druggable  Druggable  Druggable  Druggable  Druggable  Druggable  Druggable  Non-druggable Druggable | NO  YES  YES  YES  NO  YES  YES  YES  YES  YES | Non-conclusive Druggable  Druggable  Druggable  Non-conclusive Non-druggable Druggable  Druggable  Non-druggable Druggable |

scikit-learn[48].Theinputisa.txtfilewithtwocolumns,named“class”and“score”,whereeachlinecorrespondstoamolecule;the“class” col-umnshouldindicatetheactiveandtheinactivecompounds(class1and 0respectively)andthe“score” columnthescorewhichwasgivenby theclassificationmodelthatwasemployed.Theoutputincludesdif-ferentperformancemetrics(accuracy,balancedaccuracy,F-measure, precision,recallandMatthewscorrelationcoefficient)andaconfusion matrixfortheuser-selectedscorethresholdtosplitactiveandinac-tivecompounds(bydefaultissetto0.5).Moreover,thisthresholdcan bemodified,andinthiscasetheresultswillbeupdatedimmediately. Moreover,theROCandPrecision-Recallcurvesareplotted(Fig.11a) andtheirareasunderthecurvearepresentedinatablealongwiththe

9

*D.N.P.Gori,L.N.Alberca,S.Rodriguezetal.*  *ArtificialIntelligenceintheLifeSciences2(2022)100049*



**Fig.11.**a)ExampleofROCandPRcurvesgeneratedbyMetricsWebApp;b)ExamplesofPPVsurfacesgeneratedbyMetricsWebApp.

3Dplot,orbyanalyzingthetablewiththeSe,Sp,andPPVvaluesfor differentscorecutoff values,itispossibletochooseanadequatescore threshold,associatedwiththedesiredPPVrange,toselectpredicted hitsinprospectivevirtualscreeningexperiments[49].Importantly,the Se/SpratiosobservedintheROCcurveforeachcut-off valueareused tobuildthegraph,assumingthatsuchratioswillremainapproximately thesameacrossdifferentdatasets.

**6.Conclusions**

Wepresentedanddiscussedanarrayofcheminformaticstools [developedinouracademicdrugdiscoverylaboratory.Theseopen-sourceresourcesarefreelyavailableasonlineWebApps(https://lideb. biol.unlp.edu.ar/)andasstandaloneversions(https://github.com/ LIDeB/).Theyweredevelopedusingpubliclyavailableopen-sourcere-](https://lideb.biol.unlp.edu.ar/)[sourcesandinternalprogramming.](https://github.com/LIDeB/)

Theopensource/open-softwaremodelcontributestothedevelop-mentofpublicknowledgeandbridgesthetechnologicalgapbetween low-tomiddle-incomecountriesandhigh-incomecountries.Although suchagapexistsinpracticallyallfieldsofscienceandtechnology,it ispossiblynarrowerwithintheinformaticcommunity,becauseofthe persistenteffortsofglobaldeveloperstomakevaluableresourcesfreely availabletothepublic.

**DeclarationofCompetingInterest**

Theauthorsdeclarethattheyhavenoknowncompetingfinancial interestsorpersonalrelationshipsthatcouldhaveappearedtoinfluence theworkreportedinthispaper.

**Dataavailability**

Havesharedthelinktomycode.

**Acknowledgments**

[TheauthorsarefundedbyUNLP(IncentivosUNLPandSubsidiopara JóvenesInvestigadoresUNLP2019y2021)andAgenciaNacionalde PromociónCientífic](https://doi.org/10.13039/501100003074)a[yTecnológica(ANPCyT),viagrantsPICT2017-0643,PICT2019-1075,PICT2019-00984andPICT-CATI-2021-00073.](https://doi.org/10.13039/501100003074) A.T.,L.N.A.andC.L.B.aremembersoftheResearchCareerfromthe ArgentinianNationalResearchCouncil(CONICET,Argentina).Theau-thorsgratefullyacknowledgethesupportofStreamlit.

**Supplementarymaterials**

Supplementarymateri[alassociatedwiththisarticlec](https://doi.org/10.1016/j.ailsci.2022.100049)anbefound,in theonlineversion,atdoi:[10.1016/j.ailsci.2022.100049](https://doi.org/10.1016/j.ailsci.2022.100049).

**References**

[1]XuJ,Ha[glerA.Chemoinfor](https://doi.org/10.3390/70800566)maticsanddrugdiscovery.Molecules2002;7(8):566– [600.doi:10.3390/70800566.](http://refhub.elsevier.com/S2667-3185(22)00019-8/sbref0002)

[2][LeonisG,MelagrakiG,AfantitisA.Opensourcechemoinformaticssoftwareinclud-ingKNIMEanalyticsplatform.In:LeszczynskiJ,editor.Handbookofcomputational chemistry.Dordrecht:SpringerNetherlands;2016.p.1–30.](http://refhub.elsevier.com/S2667-3185(22)00019-8/sbref0002)

[3][BhardwajA,ScariaV,RaghavaGPS,LynnAM,ChandraN,BanerjeeS,Raghunan-](http://refhub.elsevier.com/S2667-3185(22)00019-8/sbref0002)dananMV,PandeyV,TanejaB,YadavJ,DashD,BhattacharyaJ,MisraA,KumarA, RamachandranS,ThomasZ,BrahmachariSK.Opensourcedrugdiscovery– Anew paradigmofcollaborati[veresearchintuberculosisdr](https://doi.org/10.1016/j.tube.2011.06.004)ugdevelopment.Tuberculosis 2011;91(5):479–86.doi[:10.1016/j.tube.2011.06.004.](https://doi.org/10.1016/j.tube.2011.06.004)

10

*D.N.P.Gori,L.N.Alberca,S.Rodriguezetal.*

[4]ÅrdalC,RøttingenJ-A.Opensource[drugdiscoveryinpractice:ac](https://doi.org/10.1371/journal.pntd.0001827)asestudy.PLoS [NeglTropDis2012;6(9):e1827.doi:10.1371/journal.pntd.0001827.](http://refhub.elsevier.com/S2667-3185(22)00019-8/sbref0005)

[[5]TanP-N,SteinbachM,KumarV.Introductiontodatamining.1steditor.USA:Ad-](http://refhub.elsevier.com/S2667-3185(22)00019-8/sbref0005) [dison-WesleyLongmanInc;2005.](http://refhub.elsevier.com/S2667-3185(22)00019-8/sbref0005)

[[6]Rivera-BorrotoOM,Marrero-PonceY,García-delaVegaJM,Grau-ÁbaloRC.Com-](http://refhub.elsevier.com/S2667-3185(22)00019-8/sbref0005)parisonofcombinatorialclusteringmethodsonpharmacologicaldatasetsrepre-sentedbymachinelea[rning-selectedrealm](https://doi.org/10.1021/ci2000083)oleculardescriptors.JChemInfModel [2011;51:3036–49.doi](http://refhub.elsevier.com/S2667-3185(22)00019-8/sbref0007)[:10.1021/ci2000083.](https://doi.org/10.1021/ci2000083)

[[7]Jolliff](http://refhub.elsevier.com/S2667-3185(22)00019-8/sbref0007)e[I.Principalco](http://refhub.elsevier.com/S2667-3185(22)00019-8/sbref0007)[mponentanalysis.2](https://doi.org/10.1021/ci2000083)[ndeditor.NewYork:Springer-Verlag;](http://refhub.elsevier.com/S2667-3185(22)00019-8/sbref0007)  [2002.](http://refhub.elsevier.com/S2667-3185(22)00019-8/sbref0007)

[[8]LloydS.LeastsquaresquantizationinPCM.IEEETransInfTheory1982;28:129–37.](http://refhub.elsevier.com/S2667-3185(22)00019-8/sbref0007)  [doi](https://molvs.readthedocs.io/en/latest)[:10.1109/TIT.1982.1056489.](https://doi.org/10.1109/TIT.1982.1056489)

[[9]M.](https://molvs.readthedocs.io/en/latest)[Swain,MolVS:molecule](https://doi.org/10.1109/TIT.1982.1056489)[ValidationandStandardization,https://molvs.](https://molvs.readthedocs.io/en/latest)  [readthedocs.io/en/latest,2019(accessedJuly2022).](https://molvs.readthedocs.io/en/latest)

[[10]MoriwakiH,TianY-S,KawashitaN,TakagiT.Mordred:amoleculardescriptorcal-](https://molvs.readthedocs.io/en/latest) [culator.JCheminformatics2018;10:4.doi](http://refhub.elsevier.com/S2667-3185(22)00019-8/sbref0011)[:10.1186/s13321-018-0258-y.](https://doi.org/10.1186/s13321-018-0258-y)

[[11]SuttonC,SindelarM,McCallumA.](http://refhub.elsevier.com/S2667-3185(22)00019-8/sbref0011)Feature[bagging:preventingweight](https://doi.org/10.1186/s13321-018-0258-y)[undertrain-](http://refhub.elsevier.com/S2667-3185(22)00019-8/sbref0011) [inginstructureddiscriminativelearning.CIIRTechRepIR-4022005:1–7.](http://refhub.elsevier.com/S2667-3185(22)00019-8/sbref0011)

[[12]RousseeuwPJ.Silhouettes:agraphicalaidtotheinterpretationandvalida-tionofclusteranalysis.JComputApplMath1987;20:53–65.doi:10.1016/ 0377-0427(87)90125-7.](https://doi.org/10.1016/\penalty -\@M 0377-0427(87)90125-7)

[[13]DunnJC.Well-SeparatedClustersandOptimalFuzzyPartitions.JCybern](https://doi.org/10.1016/\penalty -\@M 0377-0427(87)90125-7)  1974;4:95–104.doi[:10.1080/01969727408546059.](https://doi.org/10.1080/01969727408546059)

[14]DaviesDL,Bouldin[DW.AClust](https://doi.org/10.1080/01969727408546059)[erSeparationMeasure.IEEETr](https://doi.org/10.1109/TPAMI.1979.4766909)ansPatternAnal MachIntell1979;1(2):224–7.doi[:10.1109/TPAMI.1979.4766909.](https://doi.org/10.1109/TPAMI.1979.4766909)

[15]CalinskiT,Hara[baszJ.Adendr](https://doi.org/10.1080/03610927408827101)[itemethodforclusteranalysis.](https://doi.org/10.1109/TPAMI.1979.4766909)CommunStatist 1974;3:1–27.doi[:10.1080/03610927408827101.](https://doi.org/10.1080/03610927408827101)

[16]Cortes-CirianoI.[Benchmarkingthepredictivepo](https://doi.org/10.1080/03610927408827101)[werofligandefficien](https://doi.org/10.1021/acs.jcim.6b00136)cy[in](https://doi.org/10.1021/acs.jcim.6b00136)dicesin QSAR.JChemInfModel2016;56:1576–87.doi[:10.1021/acs.jcim.6b00136.](https://doi.org/10.1021/acs.jcim.6b00136)

[17]ButinaD.Unsuperviseddatabaseclusteringbase[dondaylight’sfingerprin](https://doi.org/10.1021/acs.jcim.6b00136)t[a](https://doi.org/10.1021/acs.jcim.6b00136)ndTan-imotosimilarity:afastandautomatedwa[ytoclustersmallan](https://doi.org/10.1021/ci9803381)dlargedatasets.J [ChemInfComputSci1999;39:747–50.doi](http://refhub.elsevier.com/S2667-3185(22)00019-8/sbref0018)[:10.1021/ci9803381.](https://doi.org/10.1021/ci9803381)

[[18]GreenacreM,PrimicerioP.Multivariatean](http://refhub.elsevier.com/S2667-3185(22)00019-8/sbref0018)[alysisofecologicald](https://doi.org/10.1021/ci9803381)[ata.BBVAFounda-](http://refhub.elsevier.com/S2667-3185(22)00019-8/sbref0018) [tion;2013.](http://refhub.elsevier.com/S2667-3185(22)00019-8/sbref0018)

[[19]RationalDiscoveryLLC,RDKit:open-SourceCheminformaticsandMachine](http://refhub.elsevier.com/S2667-3185(22)00019-8/sbref0018) [LearningSoftware,](http://www.rdkit.org)Open-SourceCheminformaticsandMachineLearning, [http://www.rdkit.org,](http://www.rdkit.org)2006,(accessedJune2022).

[20][VirtanenP,Gommers](http://www.rdkit.org)R,OliphantTE,HaberlandM,ReddyT,etal.SciPy1.0:funda-me[ntalalgorithmsforscientifi](https://doi.org/10.1038/s41592-019-0686-2)c[co](https://doi.org/10.1038/s41592-019-0686-2)mputinginPython.NatMethods2020;17:261–72. doi[:10.1038/s41592-019-0686-2.](https://doi.org/10.1038/s41592-019-0686-2)

[21]McI[nnesL,HealyJ,MelvilleJ.U](https://doi.org/10.1038/s41592-019-0686-2)MAP:uniformma[nifoldapproximation](https://doi.org/10.21105/joss.00861)andprojec- [tion.JOpenSourceSoftware2018;3(29):861.doi](http://refhub.elsevier.com/S2667-3185(22)00019-8/sbref0022)[:10.21105/joss.00861.](https://doi.org/10.21105/joss.00861)

[[22]AllaouiM,Kherfi ML](http://refhub.elsevier.com/S2667-3185(22)00019-8/sbref0022),[CherietA.Considerably](http://refhub.elsevier.com/S2667-3185(22)00019-8/sbref0022)[improvingclustering](https://doi.org/10.21105/joss.00861)[algorithms usingUMAPdimensionalityreductiontechnique:acomparativestudy.In:El MoatazA,MammassD,MansouriA,NouboudF,editors.ImageSignalPro-cess.Cham,Germany:SpringerInternationalPublishing;2020.p.317–25.ISBN: 978-3-030-51935-3.](http://refhub.elsevier.com/S2667-3185(22)00019-8/sbref0022)

[[23]ReynoldsD.GaussianMixtureModels.In:LiSZ,JainA,editors.Encyclope-](http://refhub.elsevier.com/S2667-3185(22)00019-8/sbref0022)[diaofbiometrics.Boston,MA:SpringerUS;2009.p.659–63.doi:10.1007/ 978-0-387-73003-5\_196.](https://doi.org/10.1007/\penalty -\@M 978-0-387-73003-5_196)

[[24]EvangelidisGD,Kounades-BastianD,HoraudR,PsarakisEZ.Agenerativemodelfor](https://doi.org/10.1007/\penalty -\@M 978-0-387-73003-5_196) thejointregistrationofmultiplepointsets.In:Comput.Vis.– ECCV2014.Lecture NotesinComputerScienc[e,8695.Cham,Germany:Sprin](https://doi.org/10.1007/978-3-319-10584-0_8)gerInternationalPublish-ing;2014.p.109–22.doi[:10.1007/978-3-319-10584-0\_8.](https://doi.org/10.1007/978-3-319-10584-0_8)

[25]HallLH,KierLB.TheElec[trotopologicalState:anAtomI](https://doi.org/10.1007/978-3-319-10584-0_8)[n](https://doi.org/10.1021/ci00028a014)dexforQSAR.JChemInf ComputSci1995;35:1039–45.doi[:10.1021/ci00028a014.](https://doi.org/10.1021/ci00028a014)

[26]MorganHL.Thegenerationof[auniquemachine](https://doi.org/10.1021/ci00028a014)descriptionforchemical structures—atechn[iquedevelopedatche](https://doi.org/10.1021/c160017a018)micalabstractsservice.JChemDoc 1965;5:107–13.doi[:10.1021/c160017a018.](https://doi.org/10.1021/c160017a018)

[27]TruchonJF,Bayly[CI.Evaluatingvirtuals](https://doi.org/10.1021/c160017a018)creeningmethods:goodandbadmet-rics[forthe“earlyreco](https://doi.org/10.1021/ci600426e)gnition” problem.JChemInfModel2007;47(2):488–508. doi[:10.1021/ci600426e.](https://doi.org/10.1021/ci600426e)

[28]Li[D,JiangK,TengD](https://doi.org/10.1021/ci600426e),WuZ,LiW,TangY,WangR,LiuG.Discoveryof NewEstrogen-relatedreceptor*𝛼* agonistsviaacombinationstrategybasedon sha[pescreeningandensembl](https://doi.org/10.1021/acs.jcim.1c00662)edocking.JChemInfModel2022;62(3):486–97. doi[:10.1021/acs.jcim.1c00662.](https://doi.org/10.1021/acs.jcim.1c00662)

*ArtificialIntelligenceintheLifeSciences2(2022)100049*

[29]KaplanAL,StrachanRT,BrazJM,CraikV,SlocumS,ManganoT,AmaboV, O’DonnellH,LakP,BasbaumAI,RothBL,ShoichetBK.Structure-basedde-signofachemicalprob[esetforthe5-HT5Aserotonin](https://doi.org/10.1021/acs.jmedchem.1c02031)receptor.JMedChem 2022;65(5):4201–17.doi[:10.1021/acs.jmedchem.1c02031.](https://doi.org/10.1021/acs.jmedchem.1c02031)

[30]MysingerMM,CarchiaM[,IrwinJJ,ShoichetBK.Directo](https://doi.org/10.1021/acs.jmedchem.1c02031)ryofusefuldecoys,en-hanced(DUD-E):betterlig[andsanddecoysfor](https://doi.org/10.1021/jm300687e)betterbenchmarking.JMedChem 2012;55(14):6582–94.doi[:10.1021/jm300687e.](https://doi.org/10.1021/jm300687e)

[31]SteinRM,YangY,Balius[TE,O’MearaMJ,Lyu](https://doi.org/10.1021/jm300687e)J,YoungJ,TangK,ShoichetBK, IrwinJJ.Property-unma[tcheddecoysindockingbe](https://doi.org/10.1021/acs.jcim.0c00598)nchmarks.JChemInfModel 2021;61(2):699–714.doi[:10.1021/acs.jcim.0c00598.](https://doi.org/10.1021/acs.jcim.0c00598)

[32][ImrieF,BradleyAR,Dea](https://doi.org/10.1093/bioinformatics/btab080)[neCM.Generatingproperty](https://doi.org/10.1021/acs.jcim.0c00598)[-matcheddecoymoleculesus-ingdeeplearning.Bioinformatics2021;37(15):2134–41.doi:10.1093/bioinformat-ics/btab080.](https://doi.org/10.1093/bioinformatics/btab080)

[33][WangL,PangX,LiY,ZhangZ,TanW.RADER:aRApidDEcoyRetrievertofacili-](https://doi.org/10.1093/bioinformatics/btab080)tate[decoybasedassessmentofvirtua](https://doi.org/10.1093/bioinformatics/btw783)lscreening.Bioinformatics2017;33(8):1235–7. doi[:10.1093/bioinformatics/btw783.](https://doi.org/10.1093/bioinformatics/btw783)

[34]Cer[eto-Massagué A,GuaschL,Vall](https://doi.org/10.1093/bioinformatics/btw783)sC,MuleroM,PujadasG,Garcia-Vallvé S.

[DecoyFinder:aneasy-to-usepythonGUIapplicationforbuildingtarget-specifi](https://doi.org/10.1093/bioinformatics/\penalty -\@M bts249)c [decoysets.Bioinformatics2012;28(12):1661–2.doi:10.1093/bioinformatics/ bts249.](https://doi.org/10.1093/bioinformatics/\penalty -\@M bts249)

[35][NichollsA.Whatdoweknowandwhendoweknowit?JComputAidedMolDes](https://doi.org/10.1093/bioinformatics/\penalty -\@M bts249)  2008;22:239–55.doi[:10.1007/s10822-008-9170-2.](https://doi.org/10.1007/s10822-008-9170-2)

[36]IrwinJJ.Community[benchmarksforvirtualscre](https://doi.org/10.1007/s10822-008-9170-2)ening.JComputAidedMolDes 2008;22:193–9.doi[:10.1007/s10822-008-9189-4.](https://doi.org/10.1007/s10822-008-9189-4)

[37]O’BoyleNM,Ban[ckM,JamesCA,Morley](https://doi.org/10.1007/s10822-008-9189-4)C,VandermeerschT,Hutchi-son[GR.Openbabel:ano](https://doi.org/10.1186/1758-2946-3-33)penchemicaltoolbox.JCheminform2011;3:33. doi[:10.1186/1758-2946-3-33.](https://doi.org/10.1186/1758-2946-3-33)

[38]Vog[elSM,BauerMR,Boeckler](https://doi.org/10.1186/1758-2946-3-33)FM.DEKOIS:demandingevaluationkitsforobjective insilicoscreening-aversatiletoolforbenchmarkingdo[ckingprogramsand](https://doi.org/10.1021/ci2001549)scoring functions.JChemInfModel2011;51(10):2650–65.doi[:10.1021/ci2001549.](https://doi.org/10.1021/ci2001549)

[39]YangY,ChenH,NilssonI,MuresanS,EngkvistO.[Investigationofth](https://doi.org/10.1021/ci2001549)erela-tionshipbetweentopolog[yandselectivityfor](https://doi.org/10.1021/jm1008456)druglikemolecules.JMedChem 2010;53(21):7709–14.doi[:10.1021/jm1008456.](https://doi.org/10.1021/jm1008456)

[40]BemisGW,MurckoMA.T[hepropertiesofknow](https://doi.org/10.1021/jm1008456)[ndrugs.](https://doi.org/10.1021/jm9602928)1.Molecularframeworks. JMedChem1996;39:2887–93.doi[:10.1021/jm9602928.](https://doi.org/10.1021/jm9602928)

[41]HopkinsAL,GroomCR.Targetana[lysis:aprioriassessm](https://doi.org/10.1021/jm9602928)entofdruggability.Small molecule— proteininteractions.ernstscheringresearchfoundationworkshop.

Wa[ldmannH,KoppitzM,editor](https://doi.org/10.1007/978-3-662-05314-0_2)s,Berlin,Heidelberg:Springer;2003.vol42. doi[:10.1007/978-3-662-05314-0\_2.](https://doi.org/10.1007/978-3-662-05314-0_2)

[42]Gau[ltonA,HerseyA,NowotkaM,](https://doi.org/10.1007/978-3-662-05314-0_2)BentoAP,ChambersJ,MendezD,MutowoP, AtkinsonF,BellisLJ,Cibrián-UhalteE,DaviesM,DedmanN,KarlssonA,Magar-iñosMP,OveringtonJP,PapadatosG,SmitI,Leac[hAR.TheChEMBLdata](https://doi.org/10.1093/nar/gkw1074)basein 2017.NucleicAcidsRes2017;45(D1):D945–54.doi[:10.1093/nar/gkw1074.](https://doi.org/10.1093/nar/gkw1074)

[43]HatosA,Hajdu-SoltészB,MonzonAM,PalopoliN,Á[lvarezL,Aykac-FasB,B](https://doi.org/10.1093/nar/gkw1074)assotC, BenítezGI,BevilacquaM,ChasapiA,ChemesL,DaveyNE,Davidović R,DunkerAK, ElofssonA,GobeillJ,FoutelN,SudhaG,GuharoyM,HorvathT,PiovesanD.DisProt: intrinsi[cproteindisorderann](https://doi.org/10.1093/nar/gkz975)otationin2020.NucleicAcidsRes2020;48(D1):D269–76.doi[:10.1093/nar/gkz975.](https://doi.org/10.1093/nar/gkz975)

[44]Schmid[tkeP,BarrilX.U](https://doi.org/10.1093/nar/gkz975)nderstandingandpredictingdruggability.Ahigh-throughputmethodfor[detectionofdrug](https://doi.org/10.1021/jm100574m)bindingsites.JMedChem 2010;53(15):5858–67.doi[:10.1021/jm100574m.](https://doi.org/10.1021/jm100574m)

[45]SaelL,LiB,LaD,FangY,[RamaniK,Rustamov](https://doi.org/10.1021/jm100574m)R,KiharaD.Fastproteintertiary structu[reretrievalbasedong](https://doi.org/10.1002/prot.22030)lobalsurfaceshapesimilarity.Proteins2008;72:1259–73.doi[:10.1002/prot.22030.](https://doi.org/10.1002/prot.22030)

[46]DongJ,[YaoZJ,ZhangL,Luo](https://doi.org/10.1002/prot.22030)F,LinQ,LuAP,ChenAF,CaoDSJ.PyBioMed:apython libraryforvariousmolecularrepresentationsof[chemicals,proteinsandDNA](https://doi.org/10.1186/s13321-018-0270-2)sand theirinteractions.Cheminform2018;10:16.doi[:10.1186/s13321-018-0270-2.](https://doi.org/10.1186/s13321-018-0270-2)

[47]SitzmannM,IhlenfeldtWD,Nicklaus[MC.Tauto](https://doi.org/10.1007/s10822-010-9346-4)[merisminlargedatabases.JCo](https://doi.org/10.1186/s13321-018-0270-2)mput AidedMolDes2010;24:521–51.doi[:10.1007/s10822-010-9346-4.](https://doi.org/10.1007/s10822-010-9346-4)

[48]PedregosaF,VaroquauxG,Gramf[ortA,MichelV,ThirionB,](https://doi.org/10.1007/s10822-010-9346-4)GriselO,Blon-delM,PrettenhoferP,WeissR,DubourgV,VanderplasJ,PassosA,CournapeauD, BrucherM,PerrotM,Duchesnay[É.Scikit-learn:machineLear](https://doi.org/10.5555/1953048.2078195)ninginPython.JMach LearnRes2011;12:2825–30.doi[:10.5555/1953048.2078195.](https://doi.org/10.5555/1953048.2078195)

[49]AlbercaLN,ChuguranskySR,Ál[varezCL,TaleviA,Salas-Sar](https://doi.org/10.5555/1953048.2078195)duyE.Insilicoguided drugrepurposing:discoveryofnewcom[petitiveandnon-competitive](https://doi.org/10.3389/fchem.2019.00534)inhibitorsof Falcipain-2.FrontChem2019;7:534.doi[:10.3389/fchem.2019.00534.](https://doi.org/10.3389/fchem.2019.00534)

11