

[Artificia](https://doi.org/10.1016/j.ailsci.2021.100012)l[IntelligenceintheLifeSciences1(2021)100012](https://doi.org/10.1016/j.ailsci.2021.100012)

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

ResearchArticle

AutoOmics:Newmultimodalapproachformulti-omicsresearch

ChiXua,#,DenghuiLiua,#,LeiZhanga,#,ZhimengXua,#,WenjunHea,HualiangJiangb,c, MingyueZhengb,NanQiaoa,∗

a*LaboratoryofHealthIntelligence,HuaweiTechnologiesCo.,Ltd.,China*   
b*DrugDiscoveryandDesignCenter,StateKeyLaboratoryofDrugResearch,ShanghaiInstituteofMateriaMedica,ChineseAcademyofSciences,555ZuchongzhiRoad, Shanghai201203,China*   
c*ShanghaiInstituteforAdvancedImmunochemicalStudies,andSchoolofLifeScienceandTechnology,ShanghaiTechUniversity,Shanghai200031,China*

|  |  |  |
| --- | --- | --- |
| article | info | abstract |
| *Keywords:*  Multi-omics  Cancergenomics  Integrativeanalysis  Automaticmachinelearning Deeplearning | | Deeplearningisverypromisinginsolvingproblemsinomicsresearch,suchasgenomics,epigenomics,pro-teomics,andmetabolics.Thedesignofneuralnetworkarchitectureisveryimportantinmodelingomicsdata againstdifferentscientificproblems.Residualfully-connectedneuralnetwork(RFCN)wasproposedtoprovide betterneuralnetworkarchitecturesformodelingomicsdata.Thenextchallengeforomicsresearchishowto integrateinformationfromdifferentomicsdatausingdeeplearning,sothatinformationfromdifferentmolecular systemlevelscouldbecombinedtopredictthetarget.Inthispaper,wepresentanovelmulti-omicsintegration approachnamedAutoOmicsthatcouldefficientlyintegrateinformationfromdifferentomicsdataandachieve betteraccuracythanpreviousapproaches.Weevaluatedourmethodonfourdifferenttasks:drugrepositioning, targetgeneprediction,breastcancersubtypingandcancertypeprediction,andallthefourtasksachievedstate ofartperformances. |

**Introduction**

Withthedevelopmentofsequencingtechnologies,researchershave extendedthelargescalewholegenomeprofilingexperimentsfromge-nomicstoepigenetics,proteomicsandmetabolics.Toobtainwhole omicsprofilingdatafromasinglesampleorindividualismoreand morepopularinbiomedicalresearches[1].Ithelpsresearcherstoex-tractevidencesfromdifferentmolecularsystemlevels,toexploreand understandtheunderlyingbiologicalmechanisms.Forexample,incan-cerresearch,researchersneedtoconfirmevidencesfromcancercell geneexpression,genemutations,genecopynumbervariationsandgene methylationstoformaproperhypothesis,toolslikeoncoplotisdevel-opedtohelpvisualizeandanalyzethemulti-omicsdata[2].

Deeplearningisverypopularingenomicsresearchrecently[3].Pre-viousworkhaveprovedtheadvantageofDeepNeuralNetwork(DNN) againsttraditionalmachinelearningmethods,suchassupportvector machine(SVM),logisticregressionandXgboostinsingle-omicsarea[4–6].Howtoapplydeeplearninginmulti-omicsresearchisanewchal-lengingareaforresearchers.

Thesimplestwaytointegratemulti-omicsdataistoconcatenate alltheomicsdataastheinputtotheDNN.Forexample,DeepSynergy [7]concatenatesfingerprintsofmolecularstructuresofchemicalsand

∗Corresponding[author.](mailto:qiaonan3@huawei.com)

*E-mailaddress:*[qiaonan3@huawei.com](mailto:qiaonan3@huawei.com)(N.Qiao).

#Theseauthors[contributedequally](mailto:qiaonan3@huawei.com)

omicsdataofcancercelllinesdirectlyasinputfortheMulti-layerper-ception(MLP)neuronnetwork.Theproblemisthatthedatadistribution ofdifferentomicsdatavaryalot,someomicsdataevenhavedifferent datatypes,whichmakestheDNNdifficulttofitagoodmodel.

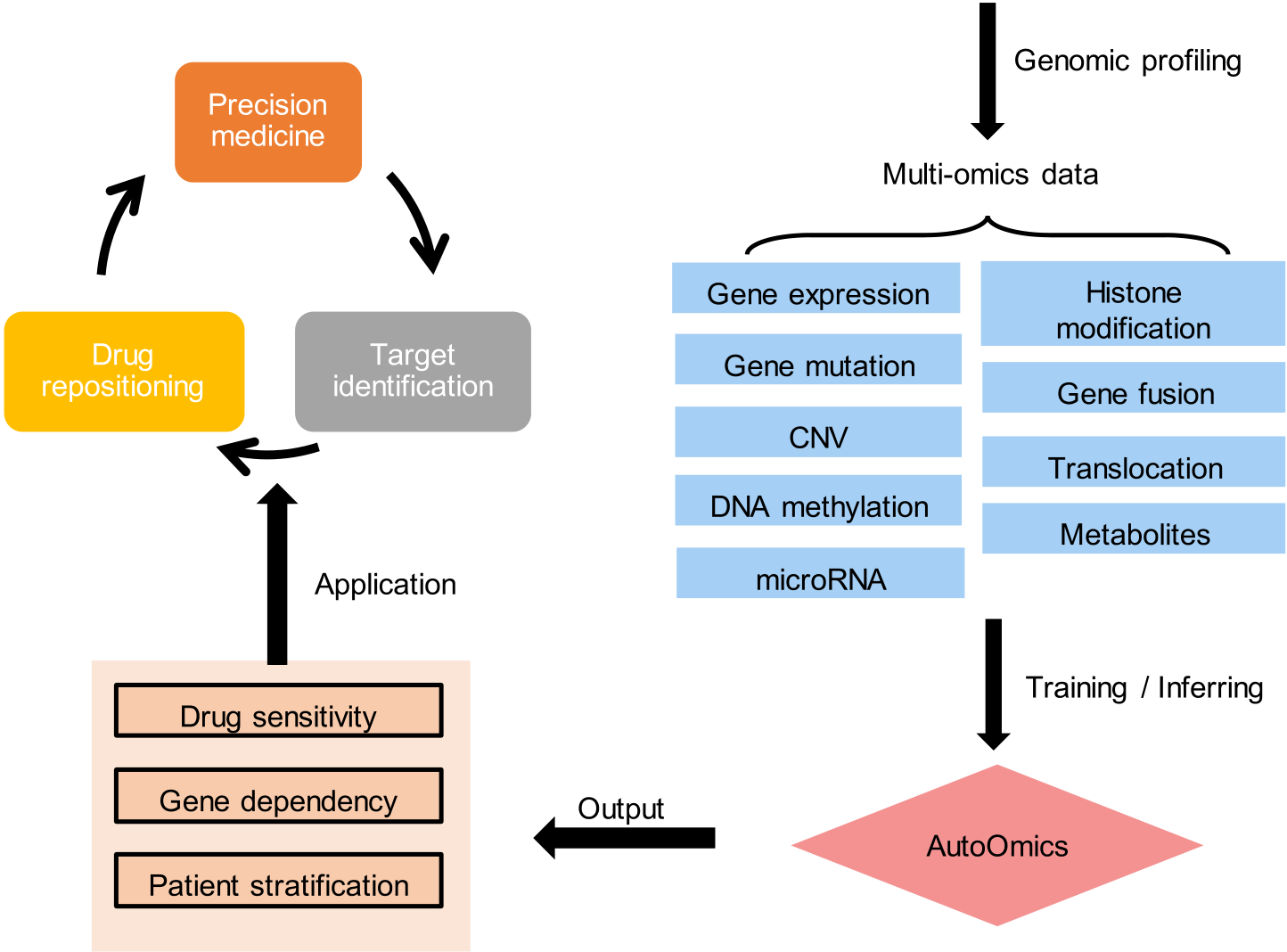
Theadvancedapproachistousesub-networksforeachomicsdata, andconcatenatetheoutputofthesub-networkstopredictthetargets. Forexample,MOLI[8]usesthreeMLPsub-networksforthreeomics data(cancercelllinegeneexpression,mutationandCNVinGDSC[9]), thenconcatenatesthesub-networkstogetherforprediction.PaccMann [5]extendsthisapproachbyusingthreesub-networkstomodelmolec-ularstructures,omicsdataandgeneinteractionnetworksseparately. Thelimitationisthatitcannotfinetuneoroptimizethesub-networks independently.

ThemoresophisticatedapproachistotrainDNNseparatelyforeach omicsdata,andthenconcatenatetheembeddinglayerstogetherto makethefinalpredictions,forexample,DeepDR[6]trainsautoencoder (AE)forgeneexpressionandgenemutationseparately,thelatentspaces arethenconcatenatedandforwardedwithaMLPtomakethefinalpre-diction.

Theproposalofresidualfully-connectedneuralnetwork(RFCN)im-plementedinAutoGenome[10]haveshownusagoodframeworkto modelthesingle-omicsdata,wewanttogofurthertoextendittomulti-omicsresearch,byconsideringa.theadvantageofRFCNneuralnetwork architectures;b.theAutoMLfeaturesfromAutoGenome;c.novelmulti-

<https://doi.org/10.1016/j.ailsci.2021.100012>  
[Received2October2021;Receivedinrevised](https://doi.org/10.1016/j.ailsci.2021.100012)form6November2021;Accepted6November2021   
Availableonline18November2021   
[2667-3185/© 2021TheAuthors.PublishedbyElsevi](http://creativecommons.org/licenses/by-nc-nd/4.0/)erB.V.ThisisanopenaccessarticleundertheCCBY-NC-NDlicense (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *C.Xu,D.Liu,L.Zhangetal.* | | | | | | *ArtificialIntelligenceintheLifeSciences1(2021)100012* **Fig.1.**Integrativeanalysisofmulti-omics datafrommultiple-originsamplesandbuild-ingAImodelsforbiomedicineresearch. |
|  | | | |  |  |
|  |  |  |  | | |





modelapproachestointegratedifferentmulti-omicsdata.AutoOmicsis developedforthispurpose,usersspecifythelocationoftheinputomics dataandthelearningtargets,AutoMListhenusedtotraintheDNN modelautomatically,afteranoptimalmodelistrained,modelinterpre-tationmodulewillbeusedtoexplaintheinfluenceofeachgenetothe learningtargets.AutoOmicssupportsbothregressionandclassification tasks,itdoesn’trequiretheuserstomastertensorflow[11]orpytorch [12]tostartwith.

WeevaluatedtheperformancesofAutoOmicsonfourdifferent multi-omicsbiomedicaltasks:a.drugresponseprediction,b.genede-pendencyprediction,c.breastcancersubtypepredictionandd.pan-cancerpatientstratification,andAutoOmicsoutperformedalltheexist-ingmethods.TheresultsshowedthatAutoOmicscouldefficientlyinte-gratelargescalemulti-omicsdataandgenerateexplainableAImodels. WeenvisionAutoOmicstobecomeapopularmethodinmulti-omics research.

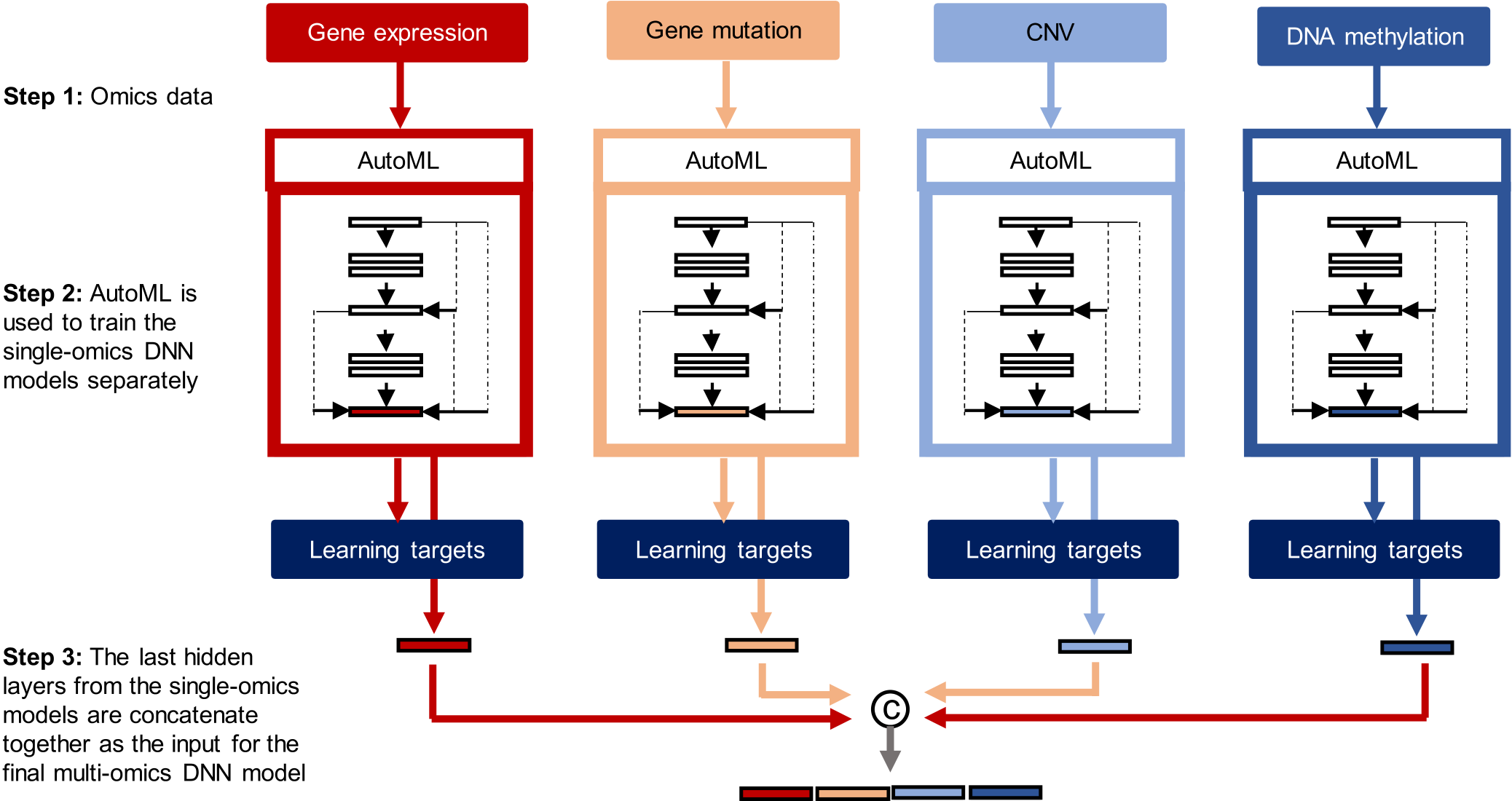
**Results**

*OverviewofAutoOmics*

Researchersfromhospitals,pharmaceuticalcompaniesandaca-demicinstitutesusuallyusepatienttissues,animalmodelsandcelllines intheirresearchtostudybiomedicalproblems.Touncoverthemolecu-larlevelmechanisms,highthroughputsequencingtechnologiesareused toprofilemultipletypesofomicsdata,suchasgeneexpression,gene mutation,copynumbervariation(CNV),DNAmethylation,microRNA andhistonemodification(Fig.1).Byanalyzingthemulti-omicsdata,re-searcherscouldformulatenewhypothesisorcreatemathematicalmod-elsforforecasting,suchasdrugsensitivityprediction,genedependency

2

*C.Xu,D.Liu,L.Zhangetal.*  *ArtificialIntelligenceintheLifeSciences1(2021)100012*



|  |  |
| --- | --- |
|  |  |

**Fig.2.**SchemeofAutoOmicsforautomaticmulti-omicsintegrationforAImodelconstruction.Step1:Collectingmulti-omicsdata.Takemulti-omicsdataof4 datatypesasexample,ageneexpressionmatrix,agenemutationmatrix,acopynumbervariation(CNV)matrixandaDNAmethylationmatrix.Step2:Useeach single-typeomicsdatarespectivelyasinputtotrainamodelforthelearningtargets.Fournetworkstructureclasses(MLP,ResNet,DenseNetandENAS),numbers oflayers,numbersofneuronsperlayerandhyperparametercombinations(batchsize,learningrateandoptimizers)wasperformedtosearchforamodelwiththe highestperformanceevaluationscoresasoptimalsingle-omicsmodelsforthissingle-typeomicsdata.Then,weightsofthealloptimalmodelswerefixed,andstopped updatinginthefollowingsteps.Step3:Concatenatelatentlayersfromalltheoptimalsingle-omicsmodels.Herewechoseallthelastlatentlayersandextractthe correspondingvectorvaluesintoaconcatenatedvectorastheinputforStep4.Step4:Usetheconcatenatedvectorasinputtotrainamodelforthelearningtargets again.SametotheprocessinStep2,searchforoptimalmulti-omicsmodel.Then,weightsofthisoptimalmodelwasfixed.Alltheoptimalsingle-omics(fromStep 2)andmulti-omics(fromStep4)modelswerecombinedtogetherintoawholenetwork,asthefinalAutoOmics-basedAImodel.

*Drugresponseprediction*

Newcomplexdiseasesarisealongwithchangesinlifestylesanden-vironment,creatingnewchallengesanddemandsfornewbiomedicine treatments[14–16].Althoughbillionsofdollarsandtensofyearshave beenspentonperdenovodrugR&D,thesuccessrateremainsquitelow. Themainreasoncomesfromsafetyissuesandunclearmechanismsof actionsfornewdrugcandidates[14].Drugrepositioning,discovering newusesforexistingFDA-approveddrugs,canavoidthesafetyissues andskiptoxicitytesting,shortentimecostinR&Dandincreasesuccess rate[15].Famousexamplese.g.Sildenafilinitiallyforpulmonaryarte-rialhypertensiontreatmentislaterfoundtotreaterectiledysfunction [17].

3

*C.Xu,D.Liu,L.Zhangetal.*

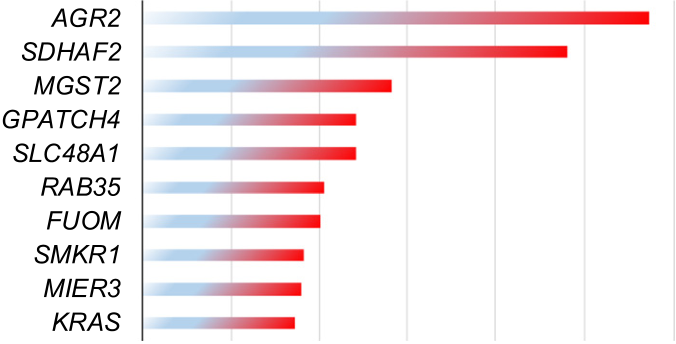
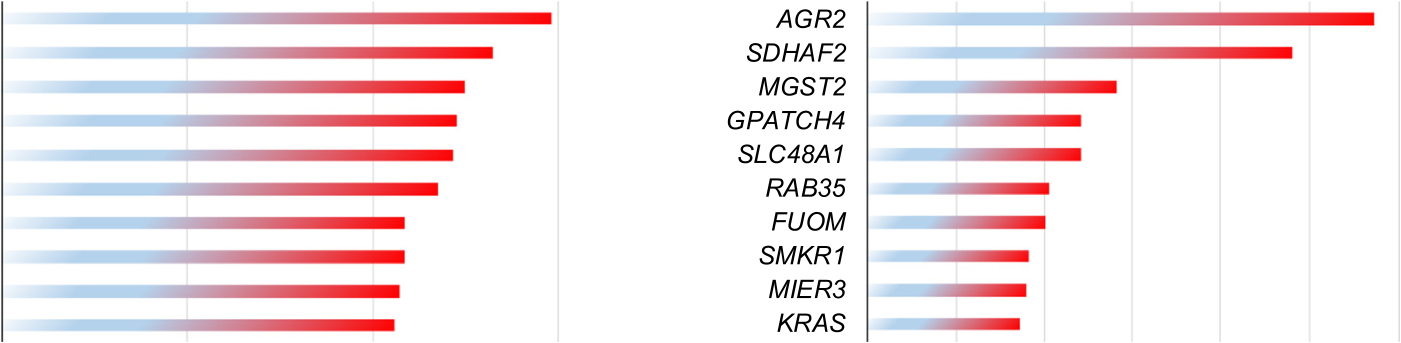
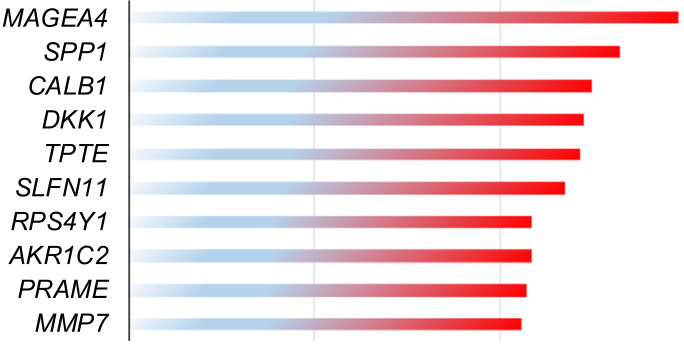
achievedthebestperformanceforbothgeneexpressionandmutation single-omicsmodelsthanotherstructures.Bothofthesetwonetworks harboredsingledenseblocks(Fig.3B)withanetgrowthrateof128 and512respectively;fortheoptimalmulti-omicsmodelpart,adirect FC(Fig.3B)ofMLPratherthanRFCN-ResNetandRFCN-DenseNetwas foundtobethebestoption(Fig.3A).Searchstrategyandsearchspaces aresummarizedinMethodsSection.Theabovethreesearchedmodels intotalcomprisedtheanticancerdrugsensitivitypredictionmodel. Meansquarederror(MSE)andspearmancorrelationcoefficient (SCC)oftheoptimalsingle-omicsmodelsofgeneexpressionwere1.532 and0.8717,and1.9and0.8463forgenemutationmodel.MSEandSCC oftheoptimalmulti-omicsmodelwereobviouslyimproved,withanob-viouslydecreasingMSE0.3266andincreasingSCC0.967(Supplemen-taryFigure3A).Weutilizedlog-transformedIC50*<*-2(approximately 0.135*𝜇*M)asstandardthreshold[19]todefinedrugsensitive(posi-tive)andresistant(negative)groups,andfoundthatthemulti-omics modelalsooutperformedsingle-omicsmodelsinareaunderthereceiver operatingcharacteristiccurve(AUROC),precision,recallandaccuracy (SupplementaryFigure3A).Ourresultsdemonstratedasignificantim-provementbymulti-omicsdataintegrationusingAutoOmicsthanonly usingsingleomicsdatafordeeplearningmodeling.

WecomparedtheAutoOmicsbaseddrugresponsepredictionmodel withpopularexistingmodels-DeepDR[6],PaccMann[5]andMOLI [8].DeepDRreportsMSEas1.96intheoriginalpaper[6].Were-producedthenetworkarchitectureofDeepDRandachievedaMSEas 1.8793andF1as0.7283usingthesamedataforAutoOmics,which wassignificantlyoutperformedbyAutoOmics(MSE0.3266,F10.8907, Fig.3C).ForPaccMann,werandomlyqueried28drugsforIC50pre-dictionsfromitswebservertocomparewithAutoOmicspredictions, andfoundthatAutoOmicsshowedhigherAUROC,areaunderpre-cisionrecallcurve(AUPRC)andaccuracyinbothcell-wise(AUROC 0.998vs.0.702,AUPROC0.854vs.0.792,accuracy0.982vs.0.768) anddrug-wise(AUROC0.987vs.0.708,AUPROC0.764vs.0.742,ac-curacy0.978vs.0.705)levels(Fig.3D).ForMOLI,ittrainedresponse predictionmodelsfor4drugs(Paclitaxel,Gemcitabine,Cetuximab,Er-lotinib)usingGDSCdata,andevaluatedinapatient-derivedxenograft (PDX)micedataset[20].ThePDXdataincludesgeneexpression,mu-tationandcopynumbervariation(CNV)profilesfor399miceanduses tumorsizereductionasdrugresponseindexto63drugtreatments.To comparewithMOLI,wepredictedresponsesforthe4drugsbyinputting thePDXmicegeneexpressionandmutationintotheAutoOmicsmodel andobtainedpredictionsforthe4drugs.TheresultsshowedthatAu-toOmicsachievedhigherAUPRCvalueforthe4drugsthanMOLI(Pa-clitaxel0.616vs.0.24,Gemcitabine0.558vs.0.49,Cetuximab0.771vs. 0.11,Erlotinib0.7vs.0.33,Fig.3E).

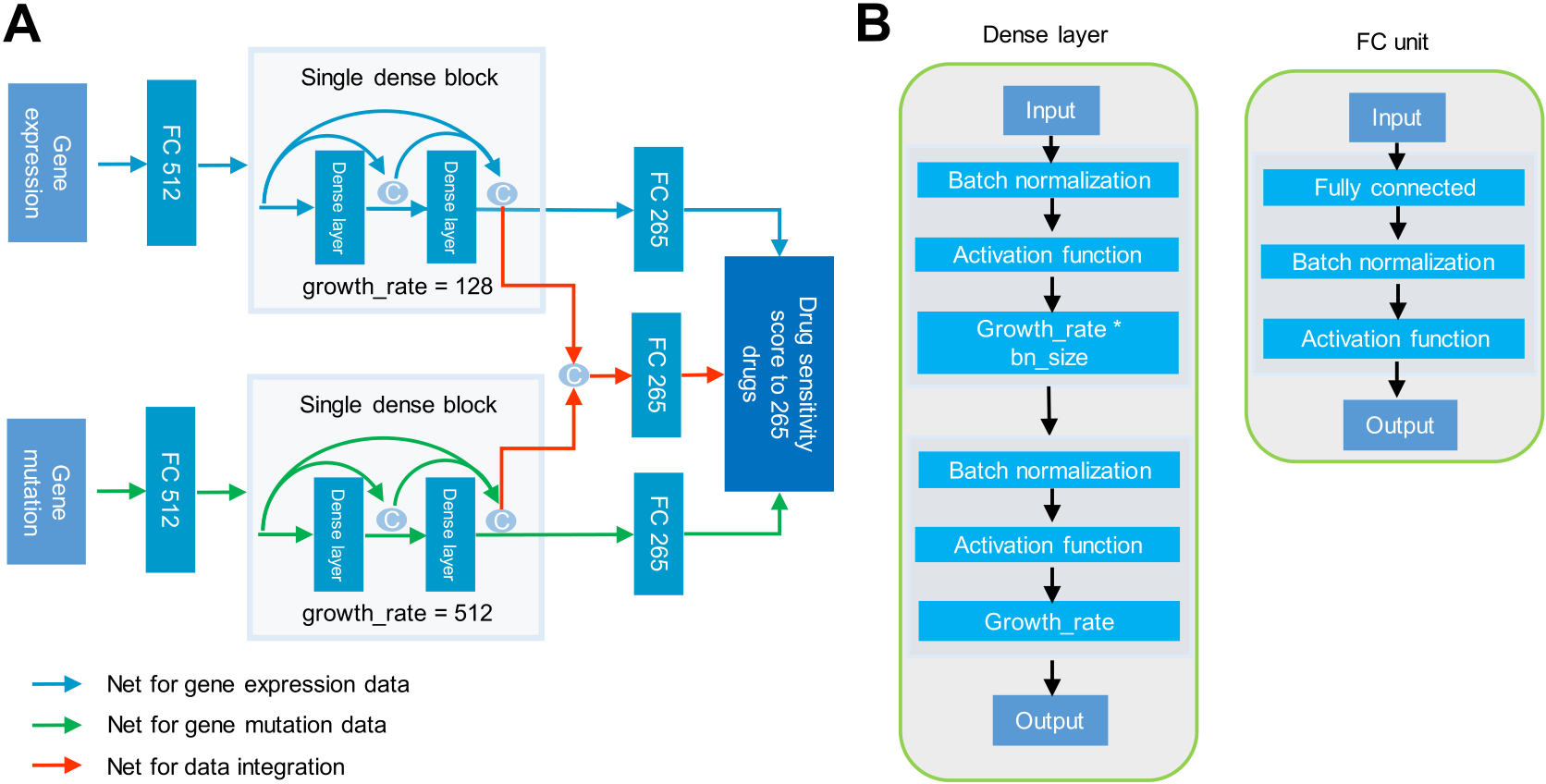
Inaddition,wecomparedAutoOmicswithotherstrategiesofomics dataintegration:concatenatedrawinput+MLP(F1:0.707),VAEla-tent+MLP(F1:0.728),VAErecon+MLP((F1:0.733),andrawin-put+AutoGenome(F1:0.725),geneexpression+AutoGenome(F1: 0.7315)andgenemutation+AutoGenome(F1:0.708),andwefound thatAutoOmicsshowedthehighestF1scoreas0.891amongall(Sup-plementaryFigure3B).Interestingly,theF1scoresweresortedas:Au-toOmics*>*AutoGenomeforgeneexpression*>*AutoGenomeforgene expressionandmutationconcatenation*>*AutoGenomeforgenemu-tation(F1:0.8907*>*0.7315*>*0.7253*>*0.7075).Itindicatedthat integratingmulti-omicsrawinputdatadirectlycausedneutralization ofgood-performancedata(geneexpression,F1:0.7315)andpoorly-performancedata(genemutation,F1:0.7075),leadingtoworsepredic-tions(F1:0.7253)thanthatofthegood-performanceone.Incontrast, AutoOmicslargelypromotedtheperformance(F1:0.8907)betterthan usingeachsingledata(SupplementaryFigure3B).

Whenanalyzingimportanceoffeaturescontributingtothefinal prediction,welistedtopgeneexpressionsandgenemutationsranked bySHAPvaluesthatshowedhighestimportancetoall265drugre-sponsepredictions(Figs.3Fand4G).Functionalenrichmentanalysis showedthattop-50rankedgeneexpressionswereenrichedindoxoru-

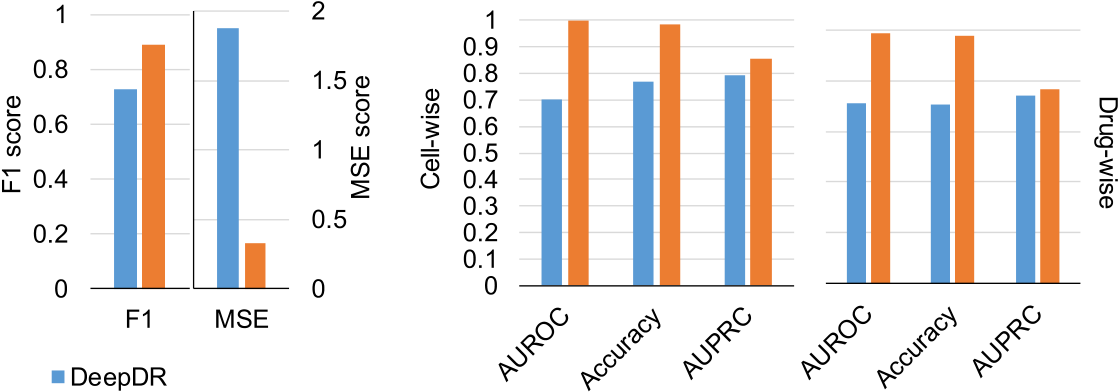
4



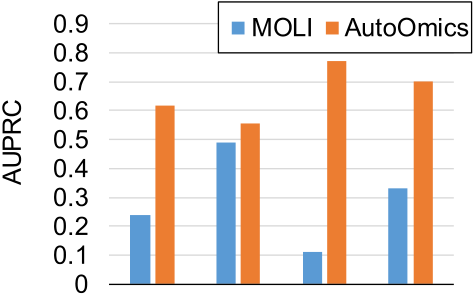
*C.Xu,D.Liu,L.Zhangetal.*  *ArtificialIntelligenceintheLifeSciences1(2021)100012*

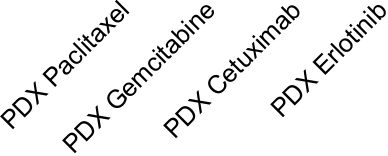


|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |



|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |





**Fig.3.**AutoOmics-baseddrugresponsepredictionmodel.

(A)IllustrationofAutoOmics-baseddrugsensitivitypredictionmodelusingGDSCcancercelllinegeneexpressionandmutationprofilestopredictsensitivityscores to265anticancerdrugs.

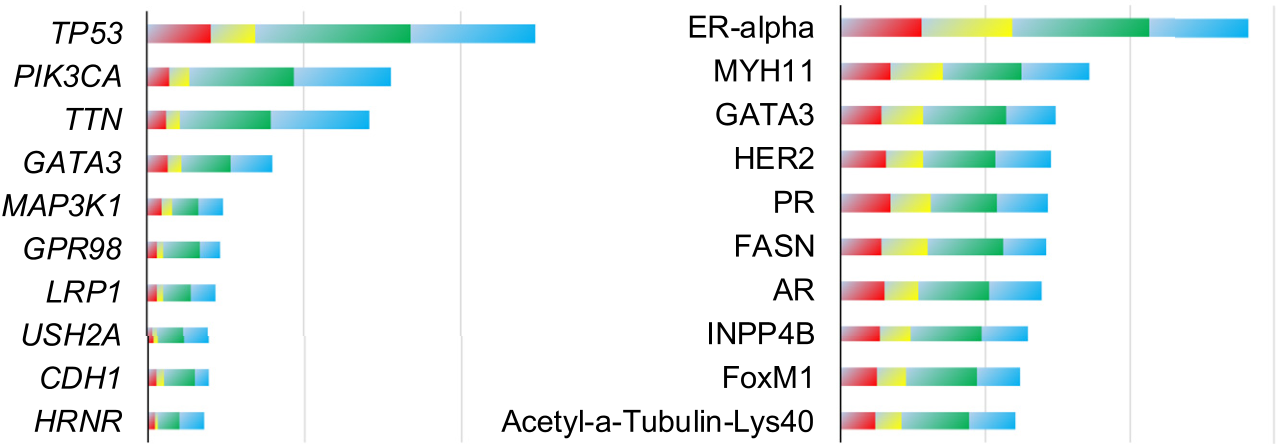
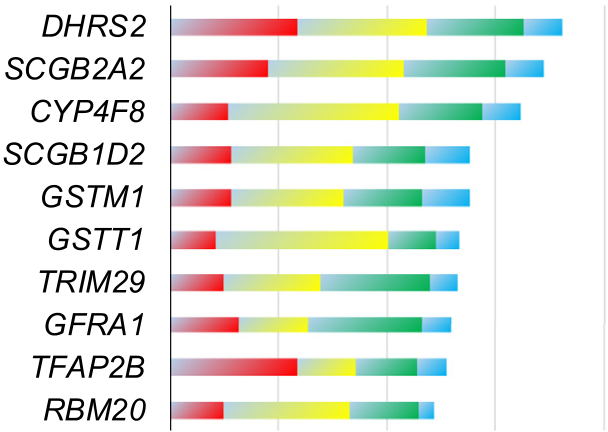
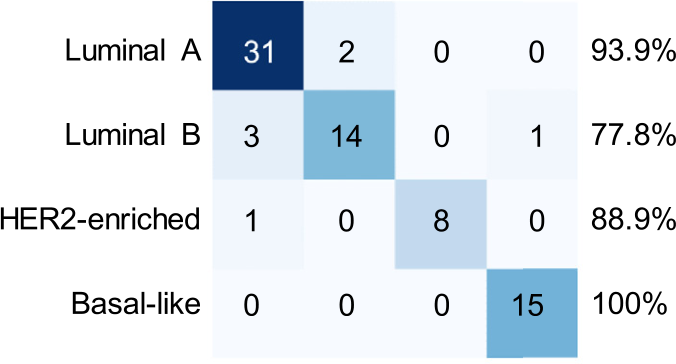
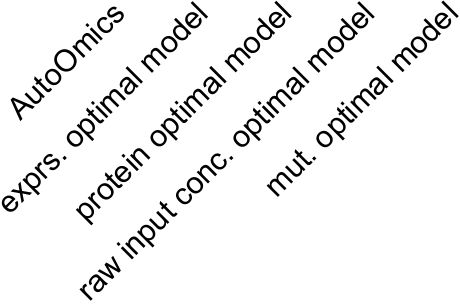
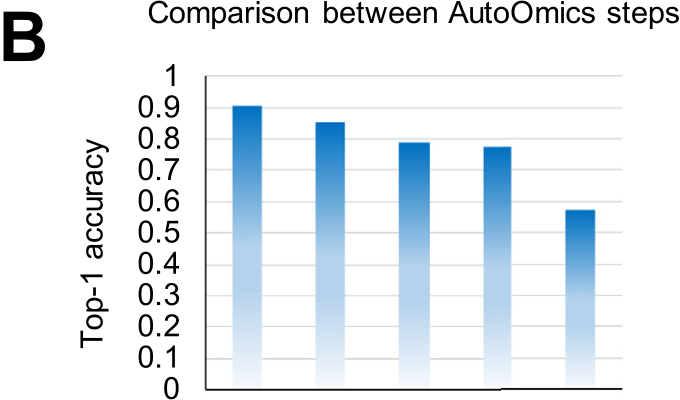
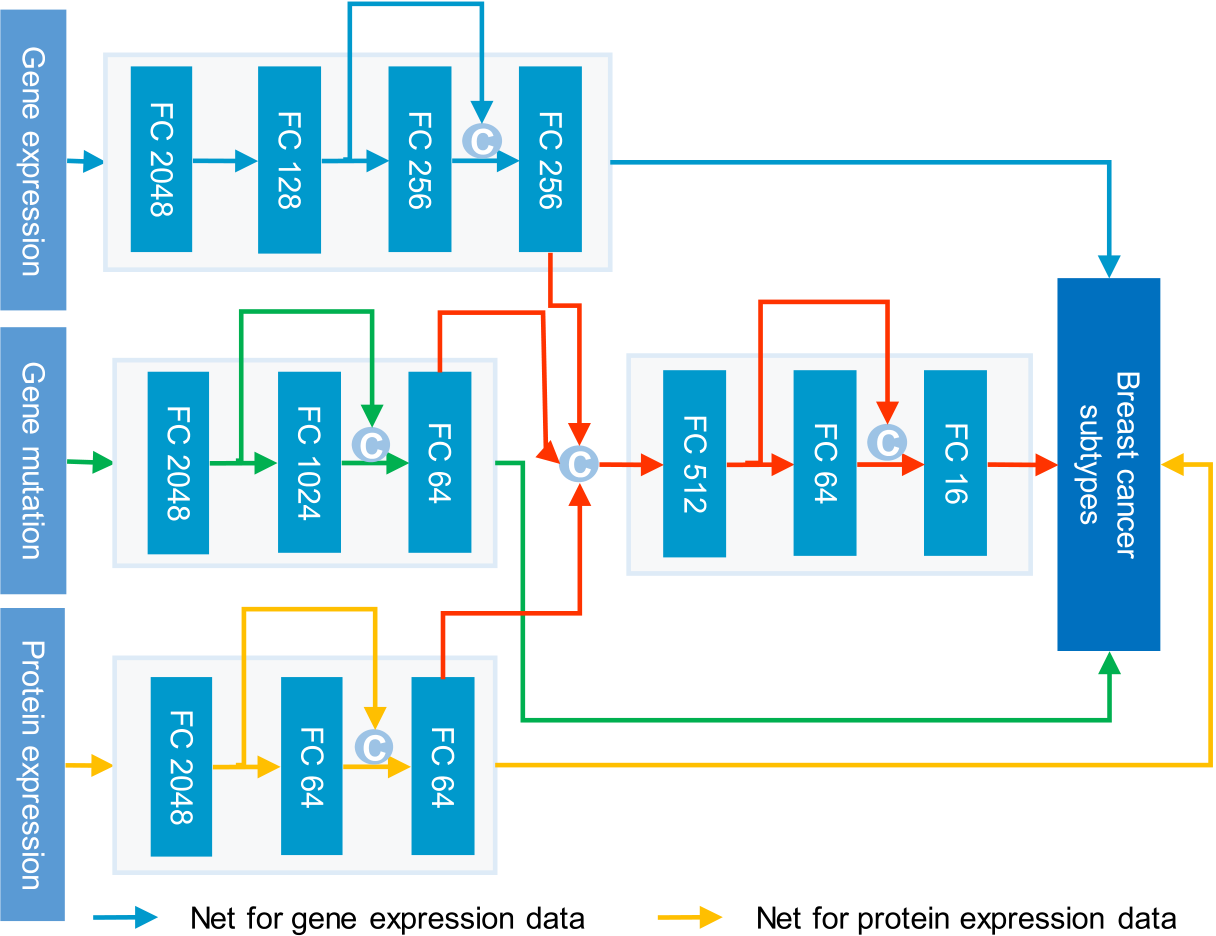
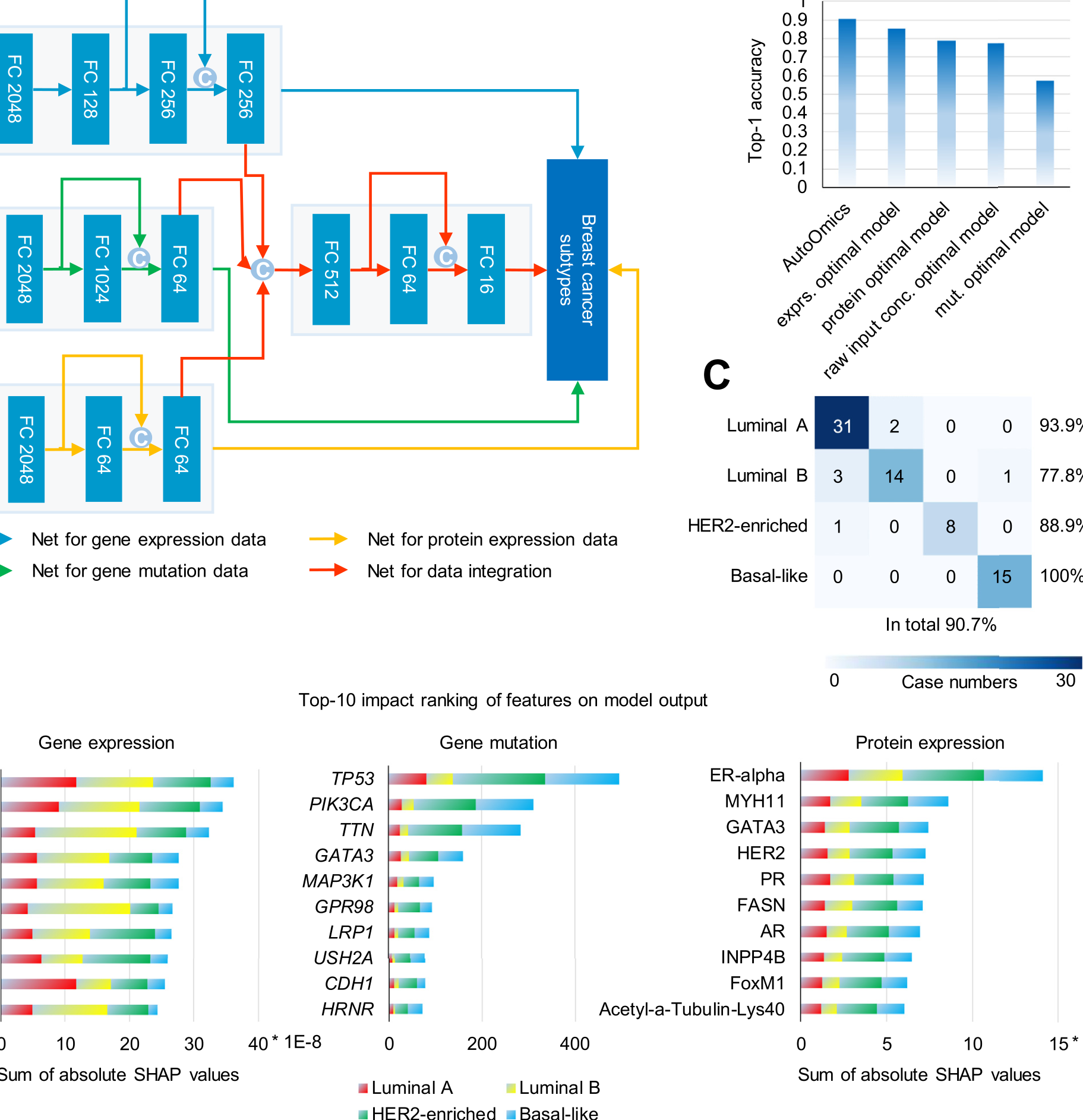
(B)IllustrationofAutomicsbasicnetunits– denselayerunitsforDenseNetandFCunits.

(C)ComparisonofAutoOmicsbaseddrugsensitivitymodeltoDeepDRusingGDSCdata.GDSCIC50valueswereusedasgolden-standardpositives.F1scoreand MSEscorewereusedforevaluationcomparison.

(D)ComparisonofAutoOmicsbaseddrugsensitivitymodeltoPaccMannusingGDSCdata.WeusedSMILESof28drugsand936celllinessharedintrainingsetsof bothourmodelandPaccMannforcomparison.GDSCIC50valueswereusedasgolden-standardpositives.Evaluationcomparisonwereperformedinbothcell-wise anddrug-wise.

(E)IndependentvalidationofAutoOmicsbaseddrugsensitivitymodelinaPDXmicedatasetandcomparisontoMOLI.IndependentvalidationofAutoOmicsina micePDXdataset.WetestedtheGDSC-IC50-trainedAutoOmicsmodelonamicePDXdata,whichusestumorsizereductionaslearningtargets.Heretumorsize reductionquantitywasusedasgroundtruthsforperformanceevaluation.4drugssharedbetweenourmodelandMOLIwereusedforcomparison   
(F)Top-10geneexpressionrankedbyimpactonmodeloutputforAutoOmicsbaseddrugsensitivitymodel   
(G)Top-10genemutationrankedbyimpactonmodeloutputforAutoOmicsbaseddrugsensitivitymodel.

5



|  |  |  |
| --- | --- | --- |
| *C.Xu,D.Liu,L.Zhangetal.* |  | *ArtificialIntelligenceintheLifeSciences1(2021)100012* |
|  |
|  | | |





**Fig.4.**AutoOmics-basedcancersubtypingpredictionmodel.

(A)IllustrationofAutoOmics-basedcancersubtypingpredictionmodelusingTCGAbreastcancerpatientgeneexpression,genemutationandproteinexpression profilestopredictfourbreastcancersubtypes.

(B)PerformancecomparisonbetweenAutoOmicsstepsusingtop-1accuracy.Totaldatawererandomlyseparatedin7:2:1ratiofortraining,evaluationandtest datasets.Thetestdatasetwasusedfortop-1accuracycalculation.

(C)ConfusionmatrixusingpredictionsofAutoOmics-basedcancersubtypingpredictionmodel.Truelabelsareintherowandpredictionsareinthecolumn.Top-1 accuracyforeachofthefourbreastcancersubtypesandtotaltop-1accuracyareindicated.

(D)Featureimportanceanalysis.Top-10featuresrankedbyimpactonmodeloutputareperformedforgeneexpression(left),genemutation(middle)andprotein expression(right).

itisnecessarytotakeadvantageofpatients’multi-omicsdataforcancer subtyping.

TheCancerGenomeAtlas(TCGA)databaseincludessixtypesof omicsdatainpatientindividuallevelformorethan20cancertypes [25].HereweimplementedAutoOmicstobuildaclassificationmodel forbreastcancersubtypepredictionusingpatients’geneexpression, genemutationandproteinexpressionprofilesasfeatures,andfocused onfoursubtypesofPAM50-profiling-testbased,luminalA,luminalB,

6

*C.Xu,D.Liu,L.Zhangetal.*

ofthe256-neuronlayer(Fig.4A).Itachievedtop-1accuracy0.8533 forbreastcancersubtyping(Fig.4B),betterthanpublishedgraphdeep learningbasedmodelusingthesameoriginofdataset(withanaccu-racyof0.8319)[27].Foroptimalsingle-omicsmodelsofgenemutation andproteinexpression,AutoOmicsreturnedbothENASmodelswith threelayersandoneskipconnections(Fig.4A),showingtop-1accu-racyas0.573and0.787respectively,lowerthanthatofgeneexpression (Fig.4B).ThenAutoOmicslinkedallthethreemodelsandagaingen-eratedanoptimalmodelfordataintegrationasanENAS-basedthree-FC-layernetworkwith512,64and16neurons(Fig.4A),withanim-provedtop-1accuracyas0.907(LuminalA:0.939,LuminalB:0.778, HER2-enriched:0.889,Basal-like:1),significantlybetterthaneachof thesingle-omicsmodel(Fig.4Band4C).Itoutperformedapublished approachusingSMO-MKL,whichachieveda0.798averageaccuracyof anytwoimmunohistochemistrymarkerbasedsubtypes[28].Whendi-rectlyconcatenatingthreeomicsdatatogether,thetop-1accuracywas 0.773,betterthanthatofgenemutationandworsethangeneexpression andproteinexpression(Fig.4B).

ThenweanalyzedimportancescoresrepresentedbySHAPvalues foreachfeaturetothemodeloutputs.SHAPvalueingeneexpression showedthat*RBM20*geneexpressioncontributedmoreinluminalB subtypethanothersubtypes(Fig.4D).Thefindingisaccordantwith thepublicresultswhichprovesthat*RBM20*geneexpressioniscorre-latedwithPDCD4-AS1/PDCD4,atumorsuppressorinTNBCcelllines [29]andasubsetofTNBCpotentiallybenefitstherapytargetinglu-minalsubtype’stypicalpathways[30].Inotherexamples,theSHAP valueof*TFAP2B*showsitisanimportantgenefeatureinbreastcan-cer(Fig.4D).ItisprovenbytheHMANPROTEINATLASwhichshows *TFAP2B*(ENSG00000008196)isacancer-relatedgeneanditsexpres-sionishighestinbreastcancerpathologydataandenrichedinbreast cancer.Besides,SHAPvalueof*TFAP2B*ishighestinluminalA,which isinagreementwiththat*TFAP2B*isassociatedwithWNT/ß-catenin pathwayinluminalbreastcancer,anditsencodingproteinAP-2tran-scriptionfactorregulatesluminalbreastcancergenes[31].

Inproteinexpressionlevel,basedonSHAPvalues,ER-alpharanked top1amongallproteinexpressionandshowedimportantcontribution toallfoursubtypesinbreastcancer(Fig.4D).Thisphenomenonisin accordancewiththeresearchofestrogenreceptorswhichisaveryim-portantmarkerforprognosisandamarkerthatispredictiveofresponse toendocrinetherapyinbreastcancer[32].ThelossofERexpression portendsapoorprognosisand,inasignificantfractionofbreastcan-cers,thisrepressionisaresultofthehypermethylationofCpGislands withintheER-alpha[33].

Tosystematicallyanalyzethereliabilityoffeatureimportancere-sults,weusedSHAP-rankingtopfeaturestorepresentsamplesand checkedinner-andintra-subtypesimilarityofsamples.Itshowedthat comparedtousingallfeaturesrepresentingsamples,SHAP-top-ranked featurescouldbetterclustersamplesofthesamesubtypestogetherand distinguishbetweenintrasubtypes,whichwasquantifiedbysilhouette score(SupplementaryFigures5–8).Thescoresachievedhighestusing topfeatures(top70,40and10forgeneexpression,genemutationand proteinexpression),andthendecreasedwhenmorebottom-rankedfea-tureswereincluded(SupplementaryFigure5).Itdemonstratedthattop-rankedfeaturesbySHAPvaluesplayedadirectroletoimprovebreast cancersubtypeclassification.

Additionally,webuilta24cancertypepredictionmodelusing TCGApancanceromicsdata.AutoOmicstraineda6-FC-layerand4-skip-connectionENASnetworkforgeneexpressionprofileswithtop-1accu-racy0.963,anda4-FC-layerand2-skip-connectionENASnetworkfor genemutationprofileswithtop-1accuracy0.681(SupplementaryFig-ure9Aand9B).AutoOmicslinkedthetwonetworksusinga3-FC-layer and1-skip-connectionENASnetwork,achievinganimprovedtop-1ac-curacy0.973(SupplementaryFigure9Aand9B).Forcomparison,we alsoperformedstackedensemblelearningtolinkthetwonetworksby usingtheirsoftmaxtargetlayersratherthanlasthiddenlayersasanew networkinput.WetestedbothENASandMLPfortheensemblelearning

7

*C.Xu,D.Liu,L.Zhangetal.*

showedlowaccuracyandENAS-basedrandomly-generatednetworksig-nificantlyoutperformedthem.ItimplythatENAScanachievealarger searchablespacethanRFCN-DenseNet,wherethestructureofthelatter onecanonlybeextendedinafixedmanner.

Takingallabove,AutoOmicsisabrandnewintegrationmethodfor multi-omicsresearch.Itcomprisesnovelnetworkunitsandnetwork architecturesspecificallydesignedforgenomicsdata.Anditalsoin-tegratesbuilt-innovelmulti-omicsintegrationmethodforefficiently takingadvantagesofmulti-omicsdata.Besides,AutoOmicsalsopro-videsbiologicalexplanationforpredictedresultsbySHAP,whichcan helpbiologisttodiscoverinterestingbiologicalmarkersforresearch. AutoOmicscansurelyspeedupbioinformaticsandgenomicsstudyand aidindissectingimportantfindingsforbiologicalresearchers.

**Methods**

**Hyper-parametersearch.**Hyper-parametersearchmethodrefers toourpreviousdescribedapproachinAutoGenome[10].Thehyper-parametersinsearchspacearelearningrate,totalbatchsize,momen-tum,weightdecay,numberoflayersinneuralnetworksandnumberof neuronsineachlayers.

**RFCN-ResNetSearchSpace.**Searchspaceareasfollowings.1)The numberofblocksforResNet,defaultvalueis[1,2,3,5,6].2)The numberofneurosineachlayer,defaultvalueis[8,16,32,64,128, 256,512,1024,2048].3)Thedrop-outratioofthefirstlayercompared withtheinputlayer,selectedfrom[0.6,0.8,1.0]   
 **RFCN-DenseNetSearchSpace.**Searchspaceareasfollowings.1) TheblocksstructureforDenseNet,defaultvalueis[2,3,4,3,4,5].2) Thegrowthrateofneurosineachblock,defaultvalueis[8,16,32, 64,128,256,512,1024,2048].3)Thedrop-outratioofthefirstlayer comparedwiththeinputlayer,selectedfrom[0.6,0.8,1.0].

**EfficientNeuralArchitectureSearch.**ENASsearchmethodrefers toourpreviousdescribedapproachinAutoGenome[10].Searchspace areasfollowings.1)Thenumberofneuronsinfromthe2rdlayertothe lastlayer,selectedfrom[16,32,64,128,256,512,1024,2048].2)The connectionrelationshipbetweendifferentlayers.

**Datacollectionandpreprocessing.**Fordrugsensitivityandgene dependencypredictiontasks,wedownloadedgeneexpressionand mutationdata,drugres[ponsedataandCRISPR-based](https://www.cancerrxgene.org)genedepen-[dencydatafromGDSC(https://www.c](https://depmap.org/portal/download/)[ancerrxgene.org](https://www.cancerrxgene.org))andDepMap (<https://depmap.org/portal/download/>[)webresources.](https://www.cancerrxgene.org)Geneexpres-s[ionprofil](https://depmap.org/portal/download/)e[includes1018cancercelll](https://depmap.org/portal/download/)inesand17,418genes.Values werelog2-transformed.Genemutationdatacovers974cancercelllines. Wesetdiscretevaluesof1and0indicatingsomaticmutatedorwide-typestatusandremovedsilentmutationcases,remainingunionsetof 19,350genesforanalysis.Drugresponsedataincludeslog-transformed IC50valuesfor990cancercelllinesrepresentingresponseto265sin-gleanticancerdrugtreatments.K-nearest-neighboralgorithmswasper-formedtofillinmissingvaluesfortheIC50responsedatabyRfunction knn.Genedependencydatacovers558cancercelllinesandresponseto CRISPRperturbationof17,634humangenes.Tofocusoncancerprior-itygenetargets[22],610geneCRISPRcaseswereremainedforanalysis. CelllinesweremappedusingidentifiersofCatalogueOfSomaticMu-tationsInCancer(COSMIC)betweendatasets,thusremained936cell linesfordrugsensitivitytaskand324forgenedependencytask. Forcancersubtypingpredictiontask,breastcancerpatients’geneex-pressio[n,mutationandproteinexpressi](https://gdac.broadinstitute.org/)ondataweredownloadedfrom TCGA(<https://gdac.broadinstitute.org/>),wherefeaturenumberswere 20,531[genes,16,806somaticmutated](https://gdac.broadinstitute.org/)genesand226proteinsrespec-tively.Geneexpressionvalueswerelog2-transformedandsilentmuta-tionswereremovedfromgenemutationdata.PAM50-basedsubtypes forpatientsweredownloadedfrompublishedpaper[35].396patients sharedbetweenthefeaturedataandsubtypedatawereusedforcancer subtypingpredictiontask.Forpancancertypeprediction,5780patient sampleswithgeneexpressionandsomaticmutatedgeneprofileswere usedformodeling.

8

*C.Xu,D.Liu,L.Zhangetal.*

[10]LiuD,etal.A[utoGenome:an](https://doi.org/10.1101/842526)AutoMLtoolforgenomicresearch.bioRxiv [2019;842526.doi:10.1101/842526.](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0011)

[[11]AbadiM,etal.TensorFlow:asystemforlarge-scalemachinelearning.In:12th USENIXSymposiumonOperatingSystemsDesignandImplementation(OSDI16).](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0011) [USENIXAssociation;2016.p.265–83.](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0012)

[[12]PaszkeA,etal.PyTorch:animperativestyle,high-performancedeeplearningli-brary.In:WallachH,etal.,editors.Advancesinneuralinformationprocessingsys-tems32.CurranAssociates,Inc.;2019.p.8026–37.](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0012)

[[13]LundbergSM,LeeS-I,etal.Aunifie](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0012)d[approachtointerpretingmodelpredictions.](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0012)

[In:GuyonI,e](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0013)t[al.,editors.Advancesinneuralinformationprocessingsystems30.](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0013) [CurranAssociates,Inc.;2017.p.4765–74.](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0014)

[[14]AshburnTT,ThorKB.Drugrepositioning:identifyinganddevelopingnewusesfor](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0014)  [existin](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0015)[gdrugs.NatRevDrugDiscov2004;3:673–83.](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0014)

[[15]](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0014)[WuZ,](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0015)[WangY,ChenL.Network-baseddrugrepositioning.MolBiosyst](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0014)  [2013;9](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0015)[:1268–81.](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0016)

[[16]XuC,etal.Accuratedrugrepositioningthroughnon-tissue-specifi](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0016)c[coresignatures](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0016)  [fromcancertranscriptomes.CellRep2018;25:523–35e5.](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0016)

[[17]LangtryHD,MarkhamASildenafi](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0016)l[.Drugs1999;57:967–89.](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0016)

[[18]YangW,eta](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0018)[l.Genomicsofdrugsensitivityincancer(GD](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0017)[SC):aresourceforthera-](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0018) [peuticb](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0018)[i](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0019)[omark](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0018)[erdiscoveryincancercells.NucleicAcidsRes2013;41:D955–61. [19]ChangY,etal.Cancerdrugresponseprofil](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0019)e[scan(CDRscan):adeeplearning](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0019)  [modelthatpredictsdrugeffectivenessfromcancergenomicsignature.SciRep](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0019)  [2018;8:8857.](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0020)

[[20]GaoH,etal.High-throughputscreeningusingpatient-derivedtumorxenograftsto](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0020)  [predictclinical](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0020)[triald](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0021)[rugresponse.NatMed2015;21:1318–25.](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0020)

[[21]McFarlandJM,](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0020)[etal.](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0021)[Improvedestimationofcancerdependenciesfromlarge-scale](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0020) [RNAiscreensusingmodel-basednormalizationanddataintegration.NatCommun](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0021) [2018;9:1–13.](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0022)

[[22]BehanFM,etal.PrioritizationofcancertherapeutictargetsusingCRISPR–Cas9](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0022)  [screens.Nature2019;568:511–16.](http://refhub.elsevier.com/S2667-3185(21)00012-X/sbref0022)

9