

[Artificia](https://doi.org/10.1016/j.ailsci.2022.100051)l[IntelligenceintheLifeSciences2(2022)100051](https://doi.org/10.1016/j.ailsci.2022.100051)

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

Revisitingactivelearningindrugdiscoverythroughopenscience JürgenBajorath   
*DepartmentofLifeScienceInformaticsandDataScience,B-IT,LIMESProgramUnitChemicalBiologyandMedicinalChemistry,Rheinische Friedrich-Wilhelms-Universität,Friedrich-Hirzebruch-Allee5/6,BonnD-53115,Germany*

Activelearning(AL)isamachinelearning(ML)approachdesigned tominimizetheamountoftrainingdatafordevelopingpredictivemod-els[1–3].Theunderlyingideaistheiterativeselectionofmostinforma-tivetraininginstancestograduallyrefineMLmodelsandimprovetheir predictiveperformance.Forlabeledandunlabeledtrainingdata,differ-entselectionstrategieshavebeenintroduced[3],attemptingtobalance exploration(samplingofrepresentativedata)andexploitation(focus-ingonmostdesiredpredictionoutcomes).Therearerelatedapproaches attheinterfacesbetweencomputationandexperimentsuchasiterative biologicalscreening[4].Inthiscase,MLmodelsarebuilttoprioritize smallsubsetsofscreeningcompoundlibrariesforexperimentalevalua-tionandincludenewlyidentifiedhitstore-trainthemodelsforthenext roundofselection.Iterativesubsetselectionandmodelrefinementaim atidentifyingthemajorityofavailablehitswhilelimitingthenumber ofdatabasecompoundsthatareexperimentallytested.

Despitetheincreasingpopularityofdata-hungrydeeplearningap-proachesindrugdiscovery,high-qualitycompoundactivityor*invivo* propertydataareoftenlimited.ThisisamajorreasonwhyMLap-proachescapableofoperatinginsparselypopulateddataspacesare attractive.Moreover,datagenerationmightsometimesberatherex-pensiveandtime-consuming.Insuchsituations,AL-drivenpredictions mightgreatlycontributeto,forexample,focusingexperimentaldesign onthemostpromisingcompounds.Hence,althoughALhasalreadybeen consideredforabouttwodecadesinpharmaceuticalresearch[1],itcon-tinuestobeatopical–andintellectuallystimulating– approach.

InanewcontributiontoAILSCI,Thompsonetal.makeuseofAL toaddressanexpensivemodelingtask,thatis,thecalculationofrela-tivebindingfreeenergy(RBFE)oftestcompoundsasameasureofpo-tencyalterations[5].Inleadoptimization,RBFEcomputations,which datebacktothe1980s[6,7],havebecomepopularforpredictingpo-tentcompoundsascandidatesforsynthesis[8,9].Recentprogressin RBFEanalysishaslargelybeenduetoincreasingcomputationalpower, graphicsprocessingunit(GPU)computing,andadvancesinconforma-tionalsamplingprocedures[8].However,althoughRBFEcalculations cannowbecarriedoutonalargerscaletoguidecandidateselection andreducesyntheticefforts[9],thecalculationscontinuetobecompu-tationallydemandingandrepresentasubstantialcostfactorforlarger compoundlibraries,aspointedoutbyThompsonandcolleagues[5].In otherwords,RBFEanalysisisatime-consumingcomputationalexercise tolimitevenmoreexpensiveexperimentalefforts.Foragivenleadop-timizationseries,ormultipleseriespursuedinparallel,eachcompound

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

wouldbesubjectedtoRBFEcalculationstoultimatelyselectthemost likelycandidate(s)forfurtherincreasedpotency.Aslongasthepoolof potentialcandidatesremainssmall,iterativeRBFEcalculationsfollowed bysynthesismightbereadilyfeasible;ifthepoolbecomeslarge–orif sizeablecomputationallyenumeratedlibrariesareinvestigated– RBFE calculationsbecomeratherchallenging.

HowcanoneapplyALtoreducethemagnitudeofRBFEanalysis? Fromagivencompoundlibrary,aconfinedsubsetmustbeselectedfor whichRBFEcalculationsarecarriedout.Onthebasisofthesedata,one thendevelopsanMLmodeltopredictRBFEvaluesfromchemicalstruc-ture.ThetrainedmodelisusedtopredictRBFEvaluesfortheremaining librarycompoundsandselectanothersubsetofpromisingcandidates forRBFEcalculations,theresultsofwhicharethenusedtore-buildthe MLmodelandfurtherrefinethepredictions.Thisprocess,reminiscent ofiterativescreening,iscontinueduntilapre-definednumberofcom-poundsfromsubsetshasbeeninvestigated,aimingtoidentifythemost potentlibrarycompoundsviaML.Itisbasedonthepremise,ofcourse, thatMLiscomputationallymuchlessexpensivethanRBFEanalysis. ThisschemewasinvestigatedbyThompsonetal.Startingfroma computationallibraryofcongenericcompounds,theauthorswereable todetect75ofthe100top-rankedRBFEcompoundsbysamplingofonly 6%ofthelibrary[5];animpressiveresult.Threeearlierstudies(includ-ingtwopreprints)alreadyappliedALinthecontextofRBFEanalysis (usingdistinctsystemset-ups),asdiscussedbytheauthors.Thefirst ofthesestudies[10]reportedasimilarsuccessrateinidentifyingtop-scoringcompoundsonthebasisofacomparablysmalllibrarysample. However,theworkbyThompsonandcolleaguesreachesfarbeyond theseearlierstudies,andmanyothersintheALfield,forseveralrea-sons.ForbenchmarkingtheALapproach,theauthorsrequiredRBFE data.Therefore,theygeneratedalibraryof10,000congenericcom-poundsandcomputedRBFEvaluesforallofthem.Notably,theauthors madetheentirelibrarywithRBFEdataandcustomcodegeneratedfor theiranalysispubliclyavailabletoensurefullreproducibilityandenable follow-upinvestigations;anoutstandingcontributiontoopenscience. Moreover,goingbeyondearlierstudies,Thompsonetal.systematically exploredfiveMLmethodsanddifferentmainparametersettingsforAL includingthe*(i)*selectionoftheinitialsubset,*(ii)*numberofcompounds sampledperiteration(sizeofeachsubset),and*(iii)*acquisitionfunction (forselectingtraininginstances).Especiallytheacquisitionfunctionis oftenthoughttoplayacriticalroleforAL.Theauthorsfoundthattheir ALresultswererobustandsurprisinglyinsensitivetotheuseofalter-

<https://doi.org/10.1016/j.ailsci.2022.100051>  
[Received2December2022;Accepted2Dece](https://doi.org/10.1016/j.ailsci.2022.100051)mber2022   
Availableonline5December2022   
[2667-3185/© 2022TheAuthors.PublishedbyElsevi](http://creativecommons.org/licenses/by-nc-nd/4.0/)erB.V.ThisisanopenaccessarticleundertheCCBY-NC-NDlicense (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

*J.Bajorath*  *ArtificialIntelligenceintheLifeSciences2(2022)100051*

nativeMLmethodsanddifferentparametersettings(includingrandom selectionoftheinitialdatasubsetandtheuseofagreedy(exploitation-based)acquisitionfunction).Thelargestinfluenceontheresultswas observedforthenumberofcompoundssampledperiteration,witha preferenceforsubsetscomprisingatleast60compounds[5].Theob-servedinsensitivityoftheALresultstovaryingparametersettingsis particularlyinteresting.Whiletheinsensitivitymightpartlydependon thecharacteristicsofthecompoundsystemunderinvestigation,italso indicatesthatthechoiceofacquisitionfunctionsislesscriticalforAL thanoftenassumedandthatsmalltrainingsetscanbesufficientfor generatingwell-performingMLmodels.

TheworkofThompsonetal.providesnewinsightsintoALand stronglysupportsopenscience.

**DeclarationofCompetingInterest**

Theauthorsdeclarethattheyhavenoknowncompetingfinancial interestsorpersonalrelationshipsthatcouldhaveappearedtoinfluence theworkreportedinthispaper.

**DataAvailability**

Nodatawasusedfortheresearchdescribedinthearticle.

**References**

[2][RekerD.Practicalconsiderationsforactivemachinelearningindrugdiscovery.Drug](http://refhub.elsevier.com/S2667-3185(22)00021-6/sbref0002)  [DiscovTodayTechnol2019;32:73–9.](http://refhub.elsevier.com/S2667-3185(22)00021-6/sbref0003)

[3][YuJ,LiX,ZhengM.Currentstatusofactivelearningfordrugdiscovery.ArtifIntell](http://refhub.elsevier.com/S2667-3185(22)00021-6/sbref0003)  [LifeSci2021;1:100023.](http://refhub.elsevier.com/S2667-3185(22)00021-6/sbref0004)

[4][BajorathJ.Integrationofvirtualandhigh-throughputscreening.NatRevDrugDis-](http://refhub.elsevier.com/S2667-3185(22)00021-6/sbref0004) [cov2002;1:882–94.](http://refhub.elsevier.com/S2667-3185(22)00021-6/sbref0004)

[5][ThompsonJ,WaltersWP,FengaJA,PabonNA,XuH,GoldmanBB,MoustakasD,](http://refhub.elsevier.com/S2667-3185(22)00021-6/sbref0004) [SchmidtM,YorkF.Optimizingactivelearningforfreeenergycalculations.Artif](http://refhub.elsevier.com/S2667-3185(22)00021-6/sbref0005) [IntellLifeSci2022;2:100050.](http://refhub.elsevier.com/S2667-3185(22)00021-6/sbref0006)

[6][KollmanPM.Molecularmodeling.AnnRevPhysChem1987;38:303–16.](http://refhub.elsevier.com/S2667-3185(22)00021-6/sbref0006)

[7][BeveridgeDL,DicapuaFM.Freeenergyviamolecularsimulation](http://refhub.elsevier.com/S2667-3185(22)00021-6/sbref0006)[:applica-tionstochemicalandbiomolecularsystems.AnnRevBiophysBiophysChem 1989;18:431–92.](http://refhub.elsevier.com/S2667-3185(22)00021-6/sbref0007)

[8][AbelR,WangL,HarderED,BerneBJ,FriesnerRA.Advancingdrugdiscovery](http://refhub.elsevier.com/S2667-3185(22)00021-6/sbref0007)  [throughenhancedfreeenergycalculations.AccChemRes2017;50:1625–32.](http://refhub.elsevier.com/S2667-3185(22)00021-6/sbref0009)

[9][SchindlerCEM,BaumannH,BlumA,BöseD,BuchstallerHP,BurgdorfL,CappelD, CheklerE,CzodrowskiP,DorschD,EguidaMKI,FollowsB,FuchsT,GrädlerU, GuneraJ,JohnsonT,JorandLebrunC,KarraS,KleinM,KnehansT,KoetznerL, KrierM,LeiendeckerM,LeuthnerB,LiL,MochalkinI,MusilD,NeaguC,Ripp-mannF,SchiemannK,SchulzR,SteinbrecherT,TanzerEM,UnzueLopezA,Via-cavaFollisA,WegenerA,KuhnD.Large-scaleassessmentofbindingfreeenergy calculationsinactivedrugdiscoveryprojects.JChemInfModel2020;60:5457–74.](http://refhub.elsevier.com/S2667-3185(22)00021-6/sbref0009) [10][KonzeKD,BosPH,DahlgrenMK,LeswingK,Tubert-BrohmanI,BortolatoA,](http://refhub.elsevier.com/S2667-3185(22)00021-6/sbref0009) [RobbasonB,AbelR,BhatS.Reaction-basedenumeration,activelearning,and freeenergycalculationstorapidlyexploresyntheticallytractablechemicalspace andoptimizepotencyofcyclin-dependentkinase2inhibitors.JChemInfModel 2019;59:3782–93.](http://refhub.elsevier.com/S2667-3185(22)00021-6/sbref0010)

[[1]WarmuthMK,LiaoJ,RätschG,MathiesonM,PuttaS,LemmenC.Activelearning](http://refhub.elsevier.com/S2667-3185(22)00021-6/sbref0001)   
[withsupportvectormachinesinthedrugdiscoveryprocess.JChemInfComputSci](http://refhub.elsevier.com/S2667-3185(22)00021-6/sbref0001)   
[2003;43:667–73.](http://refhub.elsevier.com/S2667-3185(22)00021-6/sbref0001)

2