

[Artificia](https://doi.org/10.1016/j.ailsci.2023.100064)l[IntelligenceintheLifeSciences3(2023)100064](https://doi.org/10.1016/j.ailsci.2023.100064)

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

Perspective   
Exploringchemicalspace— Generativemodelsandtheirevaluation MartinVogt   
*DepartmentofLifeScienceInformatics,b-it,LIMESProgramUnitChemicalBiologyandMedicinalChemistry,RheinischeFriedrichWilhelms-Universität, Friedrich-Hirzebruch-Allee,5-6,Bonn53115,FederalRepublicofGermany*

|  |  |  |
| --- | --- | --- |
| article | info | abstract |
| *Keywords:*  Artificialintelligence  Chemicalspace  Chemicalspaceexploration Deepneuralnetworks  Generativemodels  InverseQSAR/QSPR | | Recentadvancesinthefieldofartificialintelligence,specificallyregardingdeeplearningmethods,haveinvig-oratedresearchintonovelwaysfortheexplorationofchemicalspace.Comparedtomoretraditionalmethods thatrelyonchemicalfragmentsandcombinatorialrecombinationdeepgenerativemodelsgeneratemolecules inanon-transparentwaythatdefieseasyrationalization.However,thisopaquenaturealsopromisestoexplore unchartedchemicalspaceinnovelwaysthatdonotrelyonstructuralsimilaritydirectly.Theseaspectsandthe complexityoftrainingsuchmodelsmakesmodelassessmentregardingnovelty,uniqueness,anddistributionof generatedmoleculesacentralaspect.Thisperspectivegivesanoverviewofcurrentmethodologiesforchemical spaceexplorationwithanemphasisondeepneuralnetworkapproaches.Keyaspectsofgenerativemodelsin-cludechoiceofmolecularrepresentation,thetargetedchemicalspace,andthemethodologyforassessingand validatingchemicalspacecoverage. |

**1.Introduction**

Thechemicalspace(CS)ofsmallmoleculesreferstotheuniverse ofallconceivablechemicallystablesmallmolecules[1].Combinatorial estimatesputitssizeinexcessof1060molecules[2].Thisestimatecan easilyvarybyarbitraryordersofmagnitudedependingonthedefinition ofwhatconstitutesa“small” molecule.However,themainsignificance ofthisnumberisthatitissolargethat(a)onlyanalmostinfinitesimal fractionofCScaneverberealized,synthesized,andexplored*invitro*or *invivo*andlikewise(b)onlya“slightly” largerinfinitesimalfractioncan beexplored*insilico*regardlessofthecomputationalresourcesavailable noworinthefuture.Generativemodels(GMs)aimtoexploreCSnot byexhaustiveenumerationbutbyrandomlysamplingandexploring itsproperties.“Generative” referstothecentralaspectofsuchmodels aimingtolearnprobabilitydistributionsfromtrainingdata,whichare thenusedtoguidethegenerationofnovelmolecules.

SmallmoleculeCScanbedescribedalgorithmicallybyusingcom-binatorialapproachesthatsystematicallyexplorechemicalstructural graphsofincreasingsize.Whileexhaustiveenumerationinthiswayis infeasibleingeneral,ithasbeenrealizedformoleculesoflimitedsize (upto11,13,and17heavyatoms)andlimitedatomtypes(C,N,O, S,andhalogens)yieldingCSsof26.4million,977million,and166bil-lionmolecules,respectively[3–5].Thenumbersresultingfromthese impressiveeffortsdemonstratetheirexponentialgrowthandtheulti-matefutilityofextendingthisapproachtoevenjustmoderatelylarger molecules.Thus,beyondacertainsizesuchexhaustiveeffortshaveto

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

bereplacedbyexploratoryeffortsthatwillonlybeabletosampleatiny fractionofCSunderconsideration.

ThecreationofarbitrarychemicalstructuresexploringCSis,inprin-cipleatleast,“easy” asanyrandomchemicalgraphstructurethatfol-lowsbasicvalencerulescanbeconsideredalegalmolecule.Assuch, asimplebrute-forceapproachtosamplingCScanbeenvisionedthat (1)generatesrandommathematicalgraphs(oflimitedsize)(2)labels verticesandedgesrandomlywithatomtypesandbondorders,and(3) filterstheresultingchemicalgraphstructurestoremovechemicallynon-sensicalstructures.Alternatively,andmorecloselyreflectingtheap-proachtakenbymanycurrentdeeplearningapproaches,therandom graphgenerationprocesscanbesubstitutedbythegenerationofran-domstringstobeinterpretedaslinearrepresentationsofmoleculeslike SMILES[6].Suchasimplisticapproachhasitsanalogyinthefigura-tivemonkeyrandomlyhittinglettersonthekeyboardofatypewriter aspopularizedbytheinfinitemonkeytheorem[7].Thistheorem,when appliedtoCSguaranteesthatitisinprinciplepossibletoexhaustively exploreCSthiswaybecauseanymoleculehasacertain(althoughmin-imal)probabilityofbeinggenerated.Atthesametime,theprocessout-linedaboveisnotpracticalbecausetheprobabilityofgeneratingvalid molecularrepresentationsisextremelylow.However,iftherandom generationprocesscanbedirectedtomainlygeneratevalidmolecular representations,andprobabilitiescanbeadjustedtofocusonreason-ablechemicalstructures,theoutlinedworkflowisnotonlyfeasible,but isalsotheprincipleonwhichmostcurrentGMsrely.

Whiletheabilityto“randomlygeneratevalidchemicalstructures”isaprerequisiteforasuccessfulGMitisnotameaningfulcriterionto

<https://doi.org/10.1016/j.ailsci.2023.100064>  
[Received30December2022;Receivedinrevi](https://doi.org/10.1016/j.ailsci.2023.100064)sedform1February2023;Accepted2February2023   
Availableonline4February2023   
2667-3185/© 2023TheAuthors.PublishedbyElsevierB.V.ThisisanopenaccessarticleundertheCCBYlicense(<http://creativecommons.org/licenses/by/4.0/>)

*M.Vogt*  *ArtificialIntelligenceintheLifeSciences3(2023)100064*

judgeitsquality*perse*.Forinstance,forgenericCSexplorationitwould bepreferablethateverymoleculeoftherelevantCShasroughlythe samechanceofbeinggeneratedorthebiastowardscertainmoleculesis atleastreasonable(e.g.,largermoleculesarelesslikelytobegenerated thansmallerones).Ontheotherhand,oftenCSisexploredwithcertain goalsinmind.Justas*invitro*and*invivo*explorationofCSinleadopti-mizationcampaignsaimstooptimizeendpointslikepotencyorADMET properties[8],computationalCSexplorationcanbeconductedwithre-specttocertaingoalsfocusingonmoleculeswithfavorableproperties forendpointsofinterest[9].

LeadoptimizationcampaignsusuallyfocusonexploringCSaround oneorafewscaffolds.However,computationalapproachesarenotlim-itedinthiswayandcanemphasizeexplorationbasedonrelevantphysic-ochemicalandbiologicalproperties.

Thus,toassesstheusefulnessofGMsthefollowingaspectsneedto beconsidered:

•Validityofgeneratedrepresentations.

•Noveltyofgeneratedmolecules.

•Syntheticaccessibilityanddrug-likenessofgeneratedmolecules.•ExtenttowhichthecoveredCSisrepresentativeandrelevantforthe intendedpurpose.

Someoftheseaspectslikevalidityandnoveltycanbeeasilyquan-tifiedusingstatisticalmetrics.Furthermore,computationalmetricslike thesyntheticaccessibilityscore[10]andthequantitativeestimateof drug-likeness[11]cangiveindicationshowreasonableandpractical proposedstructuresare.ThedegreetowhichaCSisrepresentativecan beassessedbydeterminingdistributionsofstructuralandphysicochem-icalquantities,whilethebiologicalrelevanceofaCSischaracterizedby theextenttowhichitisenrichedwithmoleculespossessingfavorable propertieslikeatarget-specificactivity.ForGMstargetingafocusedCS, e.g.,CScontainingmoleculeswithspecificbiologicalactivitiesorsyn-theticallyaccessibleCS,traditionalbenchmarkapproachesarenotap-plicableastheCStobecoveredbyGMscanbeextremelylarge.Thus, itshouldnotbeexpectedthataGMisabletorecreatemoleculesfroma hold-outtestset.Onthecontrary,recreationofknownmoleculesmight beanindicationthattheCScoveredisrelativelysmallandoverpopu-latedwithmoleculesfromalreadyexploredCS.

GMsthatpredatethedeeplearningrevolutionofthelastyears mainlyusedfragment-basedapproachesandbyrecombiningfragments ensuredthattherandomlygeneratedstructures(afterfiltering)repre-sentreasonablemolecules[12–14].Bydesign,theCScoveredbysuch GMsisclearlydefined(and,atthesametime,limited)bythefragment librariesused.ThismightguaranteethattheCScoveredisrepresenta-tiveregardingstructuralcomposition,forinstanceifafragmentlibrary wasderivedfromareferencesetofrelevantmolecules,butthisdoesnot (necessarily)extendtobiologicalproperties.

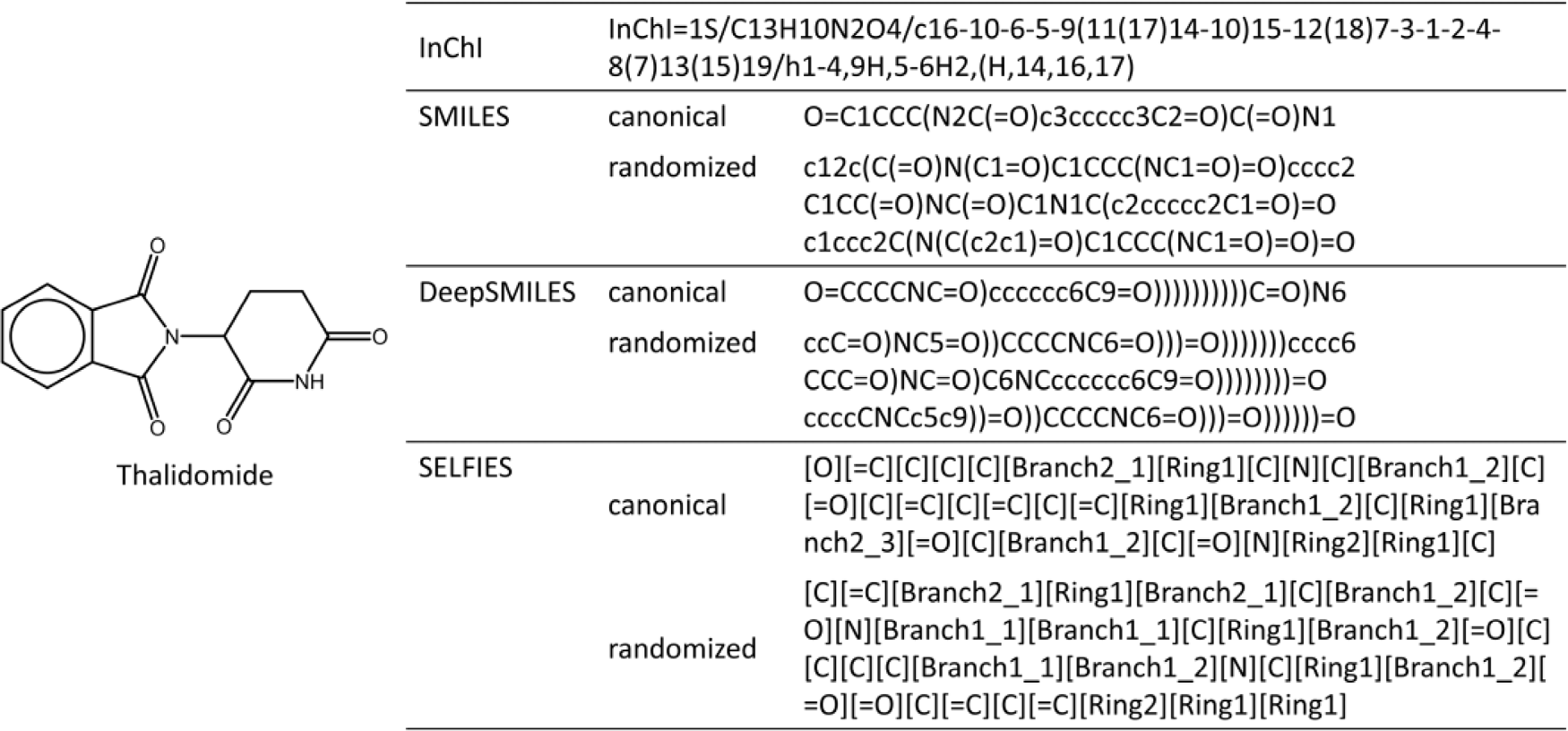
OneoftheexpectedadvantagesofCSexplorationusingdeepneural networks(DNNs)inso-calleddeepGMsistheirpotentialforabstracting fromagivenstructuralcontextandgenerateandexploreCSnotbased onstructuralsimilarityalonebutinsteadbasedonmultipleparameters thatarerelevantforendpointstobeoptimized[15,16].Inessence,this isalsothecentralquestionoftheinverseQSAR/QSPRproblem[17–20]andisoneofthedrawingpowersofapplyingadvancedartificialin-telligence(AI)methodstotheexplorationofCS.ArchitecturesofDNNs donotincorporateanychemicalknowledge,allofwhichhastobeac-quiredduringtrainingandisencodedinthelearnableparametersofthe network.However,thisaspectalsomakesithardertoassesswhethera trainedDNNfulfillstheseexpectations.Thisperspectiveextendsapre-viousreviewonCSexplorationmethodologies[21]byadditionallyfo-cusingontheproblemofassessingandvalidatingCScoveredbyGMs.

**2.Moleculargraphrepresentationsforgenerativemodels**

TherepresentationofmoleculesandtheirprocessinginGMsplayan importantroleinhowCSiscovered.Inthisregard,acrucialdistinction

2

*M.Vogt*  *ArtificialIntelligenceintheLifeSciences3(2023)100064*



**[Fig.1.](https://www.rdkit.org)**[Molecularrepre](https://www.rdkit.org)sentations.Forthalidomide,severallinearrepresentationsareshown.CanonicalandthreerandomizedSMILESweregeneratedusingRDKit

(<https://www.rdkit.org>).DeepSMILESandSELFIESweredeterminedonthebasisoftheSMILESrepresentations.DuetotheirlengthonlyonerandomizedSELFIES

i[sshown.](https://www.rdkit.org)

(seeFig.1).ThisensuresthatanyGMgeneratingSELFIES,will,byde-signonlyproducevalidmolecules.InSELFIES,symbolshavetobein-terpretedinacontext-dependentmanner,e.g.,thesamesymbolcanin-dicateaspecificelement,thesizeofasidechain,orthesizeofaring. ThecomplexsemanticsofSELFIEShastheeffectthatsmallchangesin thesyntaxcanleadtostructurallyhighlydissimilarmolecules.Thus, whileamodelbasedonSELFIESwouldgenerateonlyvalidrepresen-tationsitmightbemuchhardertolearnstructuralpropertiesfroma referenceCSwiththegoalofexpandingit.Thisobservationhasbeen confirmedshowingthatindeedSELFIES-basedGMswillgenerate100% validmolecularrepresentations,howevertheCScoveredtendstodiffer morestronglyfromareferencetrainingsetregardingmetricscharacter-izingstructuralandphysicochemicalproperties[22].

**3.Architecturesofgenerativemodels**

*3.1.Conventionalapproachesbasedonprobabilisticmethods*

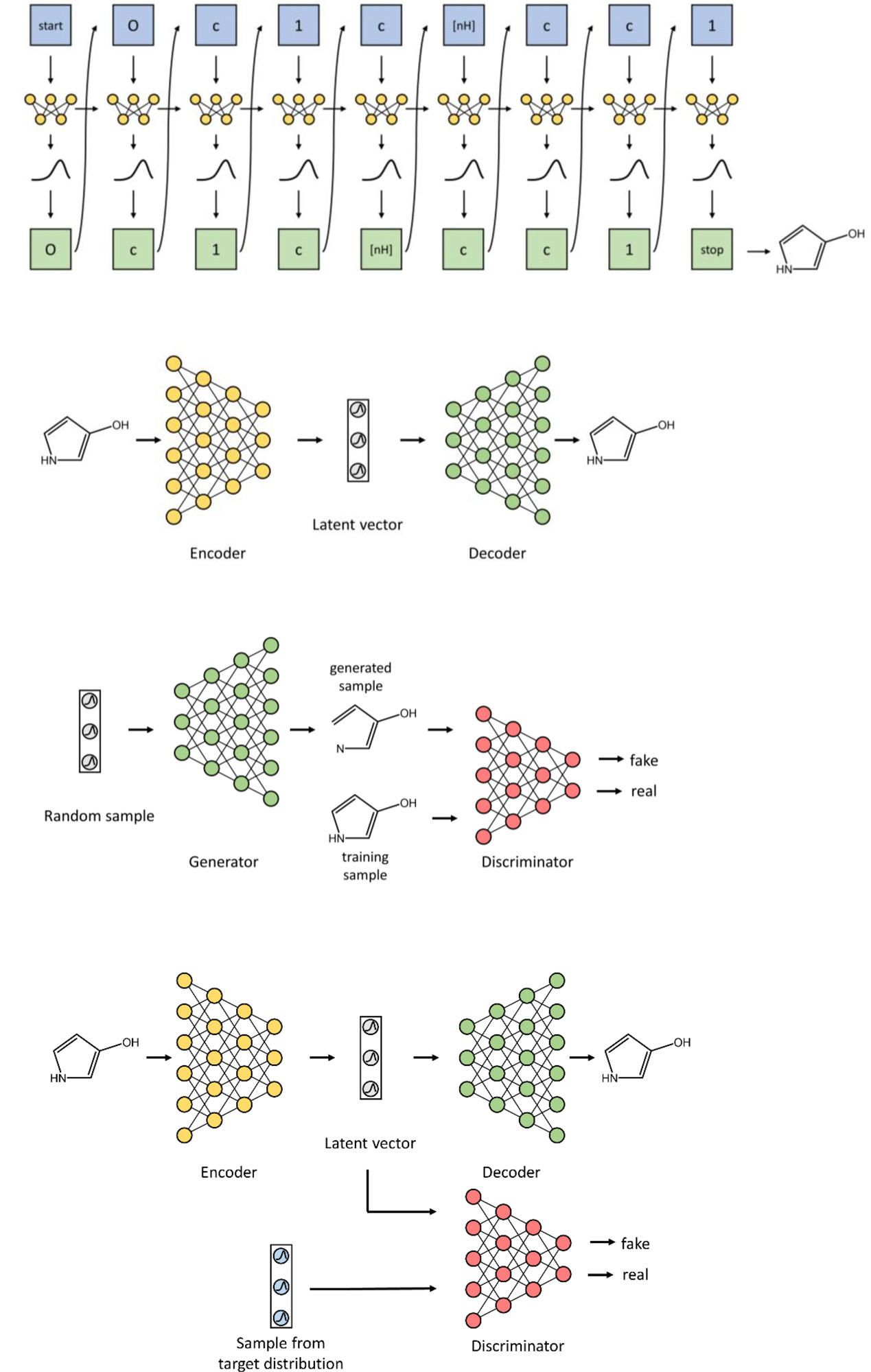
Approachesfrommachinelearningandutilizationofoptimization algorithmshavebeenincorporatedinGMsforCSexplorationthatpre-datetheriseofdeeplearninginchemoinformatics[12].Mostcommon amongthesearefragment-basedmethodsrelyingonfragmentrecom-binations[33–38]thatcanbeeitherexploredcombinatoriallyorusing singleormulti-parameteroptimizationstrategies[39,40].Thesemeth-odstypicallyemployprobabilisticmethodswiththeaimtogenerate chemicallyreasonablemoleculesandoptimizethesewithrespecttoone ormoreendpoints.

Popularprobabilisticmethodsusedinfragment-basedGMsinclude geneticorevolutionaryoptimizationalgorithms(GAs)[38–43],ant colonyoptimization[44],Markovchains[35],orMonte-Carlotree search[43].Bydesign,allthesemethodsarewellsuitedforfocusedor goaldirectedCSexplorationcampaignsbecausetheiriterativeproba-bilisticnaturecanguideexplorationtowardsCSwithfavorableprop-ertiesaspartofasingleormulti-objectiveoptimizationstrategy. Fragment-based*denovo*designmethodsarenotlimitedtoexplorationof localCSandcanalsobeadoptedforgenericsamplingofCS[38]putting theemphasisonexploringreasonablesyntheticallyaccessibleCSwhere thedistributionofgeneratedmoleculesmimicsthatofareferencesetto whichthemodelsareadapted.

Probabilisticmethodsarenotlimitedtofragment-basedapproaches. E.g.,ChemGE[42]usesapoolofchromosomesrepresentingrules

3

*M.Vogt*  *ArtificialIntelligenceintheLifeSciences3(2023)100064*



**Fig.2.**DNNarchitecturesforgenerativemodels.Fourdifferent architecturesfordeepgenerativemodelsaredisplayedschemati-cally.(a)Recurrentneuralnetwork,(b)variationalautoencoder, (c)generativeadversarialnetwork,(d)adversarialautoencoder.

|  |  |
| --- | --- |
| expressionsormatchingringclosuresymbolsintheirprobabilitydistri-butions.  TrainedRNNscangeneratemolecularrepresentationsbysampling sequencesofsymbols*s*1*,...,st*−1*,st,...*wheretheprobabilityofeachsym-bolisdeterminedbythelearneddistributions.Severalmoleculargen-erativenetworkshavebeensuggestedbasedonthisprinciple[46–51]. Frequently,however,RNNsareusedasmodulesinotherDNNarchitec-tures.E.g.,thelearnedhiddenstatesofRNNscanbeusedasthebasis forencodingmoleculesinaso-calledlatentspaceaspartofvariational autoencoders(VAEs)[15,29].Theyhavealsobeenusedasgenerators inGANs[54]orautoencoder-basedGANs[55–57]insteadofmorecon-ventionalconvolutionalneuralnetworks.  *3.2.2.Variationalautoencoders*   Theideaofautoencodersistotakeamolecularrepresentationand mapitontoalatentcontinuousvectorspaceinsuchawaythattheorigi-nalmoleculecanbereconstructed.Tothisend,anautoencoderrequires | twoDNNs(SeeFig.2b),onetoencodetherepresentationasalatentvec-torandonetodecodethelatentvector,recreatingtheoriginalrepresen-tation(oranalternativerepresentationforthesamemolecule[29]).The latentspaceisarelativelylow-dimensionalmostlynon-redundantvec-torspace.Becausetheprojectionmappingofmoleculesintothisspace hastobelearnedandisnotpredefinedbythearchitectureitispossible totrainautoencodersinsuchawaythatneighborhoodsinlatentspace reflectsimilaritiesofpropertiesorbiologicalactivitiesofinterest.Thus, samplingfromneighborhoodsofmoleculesinlatentspacecanbeused tosamplenovelmoleculeswithsimilarpropertiesthusexpandingthe localCS[15,16,29,58,59].  Whilebasicautoencodersmightbeabletorecreateinputswithouter-rortheyarenotsufficienttoguaranteethatlatentspacerepresentations havemeaningfulsimilarityrelationshipsandthusarenotwellsuited forgenerativepurposes.VAEscanbeseenasBayesianequivalentsof autoencodersandinsteadoflearningafixedrepresentation,VAEslearn amultivariateGaussiandistributioninlatentspace.Thishelpstoen- |

4

*M.Vogt*  *ArtificialIntelligenceintheLifeSciences3(2023)100064*

surecontinuity,sothatneighborhoodsinlatentspacereflectmolecular similarity[15].

*3.2.3.Generativeadversarialnetworks*   
 AGANconsistsoftwocomponents,ageneratorandadiscriminator [27].Basedonatrainableprobabilitydistribution,thegeneratorwill generatemolecularrepresentationswhilethediscriminatorwillclassify arepresentationaseither“real” or“fake” aimingtodistinguish“real”molecularrepresentationsfromareferencetrainingsetfrom“fake” ones producedbythegenerator(seeFig.2c).Thetwonetworkshaveop-positegoals,hencethename“adversarial”;whilethediscriminator’s taskistoaccuratelydistinguishrealfromgeneratedsamples,thegen-erator’sgoalistomakethistaskashardaspossibleforthediscrimi-nator.Thegeneratoranddiscriminatornetworkaretrainedsimultane-ously,andlearningisachievedbythecompetitivenatureofthepro-cess.WhiledifferentGANarchitectureshavebeenproposedforCSex-ploration[54,60],forinstanceusingRNNsasgenerators[54],theyare alsooftencombinedwithautoencodersinwhatareknownasadversarial autoencoders(AAEs).Here,ageneratorsamplesfromthelatentspace oftheautoencoderwhileadiscriminatordistinguishesrealfromfake latentvectors.Independentofthediscriminator,thedecoderisusedto generatethemolecularrepresentationfromthelatentvector[55–57,61] (seeFig.2d).

*3.3.Trainingandgoal-directedchemicalspaceexplorationfordeep generativemodels*

Althoughithasbeenrecentlyshownthatitispossibletosuccessfully trainDNNsonrelativelysmalltrainingsetsofaround10,000molecules [22],i.e.,suchmodelsareabletogeneratevalidmolecularrepresen-tationsatanacceptablerate,thesemodelsmightnotbegoodenough toaccuratelyreflecttheCSofreferencesets.However,thefocusedex-plorationofCSandtheoptimizationofsampledmoleculesregarding specific(biological)propertieslikeactivityisacentraltaskinchemoin-formatics.Typically,thedataavailableforthesetasksisoftenatbest oneorderofmagnitudelessthanrequiredtotrainadeepGM.Forex-ample,evenforthemostprominenttargets,onlyafewthousandactive moleculesarereportedindatabaseslikeChEMBL[62].Twopopularap-proachestoaddressthisissuearetransferlearningandreinforcement learning[27].Theyrelyonfirsttrainingamodelonalargegenericdata setcoveringamoregeneralCS,e.g.,bioactivemolecules,andthentun-ingthemodeltowardsaregionofinterestusingamuchsmallertraining setrelevantforthetaskathand.Inthecaseoftransferlearningthisis donebyretrainingapre-trainedmodelusingafocusedreferencedata setofinterest[46,47,50,51],whileinreinforcementlearning,theprob-abilisticdistributionunderlyingthegenerativeprocessistunedtowards atargetCS[49,55,58].ThisapproachiswellsuitedforGANs[54,60,63] andAAEs[55–57,61]buthasalsobeenappliedtoRNNs[64,65].

Consideringthatneighborhoodrelationshipsinlatentspacerepre-sentationscanbemodifiedduringtraining,theycanalsobetunedto reflectsimilaritiesofchemicalpropertiesofinterest.Gómez-Bombarelli etal.[15]andColbyetal.[16]modifiedthelossfunctionoftheVAEto includeregressionerrorswithrespecttomolecularpropertiesinorder tocontrollatentspacerepresentationsandallowoptimizationofsuch propertiesinlatentspace.InAAEs,thedistributionofmoleculesinla-tentrepresentationiscontrolleddirectlybythediscriminator,whichhas beenshowntoimprovethecontinuityoflatentspacerepresentations [66].Polykovskiyetal.[56]andHongetal.[61]combinedmolecular representationswithmolecularpropertiesresultinginlatentspacerep-resentationsthatallowconditionalgenerationofmoleculeswithdesired properties.Here,latentspaceofmolecularrepresentationistrainedto beindependentofspecificmolecularpropertieswhilethedecoderis augmentedwithapropertyvectortogeneratemoleculeswithdesired properties.

5

*M.Vogt*  *ArtificialIntelligenceintheLifeSciences3(2023)100064*

GMswereusedtosampleonebillioncompoundseachandthebestre-sultingmodelsachievedacoverageof39%.Followingtheargument above,acoverageofabout63%wouldbethetheoreticallybestperfor-mancethatcouldbeexpectedassumingperfectcoverageandidentical probabilitiesforeachmolecule.GiventhattheGMsalsogenerateda significantportionofmoleculesnotcoveredbyGDB-13thisindicated thattheGMsareindeedabletocoverasignificantpartofthewhole GDB-13.

Whilevalidity,novelty,anduniquenesscanconfirmthataGMwill indeedexplorenovelCS,thesemetricsarenotsufficienttocharacterize itregardingrelevantbiologicalandphysicochemicalproperties.Beyond purelyrandomsamplingofCS,theenvisionedutilityofGMsliesintheir abilitytogeneratenovelcompoundshavingpropertiesthatarecharac-terizedbyareferencesetusedfortrainingeitherdirectlyoraspartofre-trainingtofocusthecoveredCS.TheabilityofGMstoadequatelycover thetargetedCScanbeassessedbycomparingthepropertydistributions ofthetrainingsetmoleculesandthesampledmolecules.Forpractical applicationsthisapproachislimitedtopropertiesthatcanbeeasilycom-puted.Theseincludesimpledescriptorslikemolecularweightoratom composition,topologicaldescriptorsliketheBertztopologicalcomplex-ity[71],orphysicochemicalpropertiesliketheoctanol-waterpartition coefficientlogP(O/W)forwhichcomputationalmodelsexist[22,67,68] butalsoextendstomorecomplexdescriptorslikethesyntheticacces-sibilityscore[10]andthequantitativeestimateofdrug-likeness[11]. Thesimilaritybetweenreferenceandsampledistributionsisquantified usingmetricsliketheKullback–Leiblerdivergence,theJensen–Shannon distanceortheWassersteindistance[22].Structuralsimilaritycanbe assessedbasedonscaffoldandfragmentdistributions[68]andonTan-imotosimilaritiesofstructuralfingerprints[22,67,68]andcanalsobe usedtodetectissueslikemodecollapseorstrongbiasesinthetrained models.

Whilethesedistribution-baseddescriptorscanbeusedtodetermine howrepresentativeaCSiswithrespecttostructureandphysicochem-icalproperties,theydonotaddressbiologicalproperties.TheFréchet ChemNetDistance(FCD)[72]hasbeendevelopedwiththisinmind. TheconceptwasinspiredbytheFréchetinceptiondistance[73].How-ever,insteadofrelyingontheInceptionv3[74]networkforimageclas-sification,itreliesonaDNN,termedChemNet,thathasbeentrainedto predictbioactivitiesofmoleculesformorethan6000assays.Theacti-vationsoftheneuronsofthefinalhiddenlayerofthenetworkencode therelevantfeaturesforthepredictionsoftheoutputlayer.Forasetof moleculesthedistributionoftheseactivationsaremodeledasamulti-variateGaussiandistribution.Thedistributionsofareferencesetofreal moleculesandasampleofgeneratedmoleculesarecomparedusingthe Fréchetdistance.Whilethequalityofthisdistancewillbelimitedbythe accuracyofChemNetandthequalityandgeneralityoftheassaydata, theconceptisveryelegantfortheassessmentofthebiologicalrelevance ofaCS.

Undertheassumptionthatlargertrainingdatasetsshouldbeable toimprovetheperformanceofGMs,Skinnideretal.[22]studiedthe correlationbetweenthemetricsdiscussedhereandtrainingsetsize. Thefoundaveryhighcorrelationbetweenvalidityandtrainingsetsize (usingSMILES).FCDalsoshowedhighcorrelation,whilemoststruc-turalandphysicochemicaldescriptorsshowedgoodtosatisfactorycor-relations.Whilesmallertrainingsetsweresufficienttogeneratevalid representationsatareasonablerate,othermetricsimprovedsignifi-cantlywithlargertrainingsets.ThisindicatedthatitwaseasierforGMs tolearnthesyntaxoftherepresentations,butmoreinformationwas requiredtoaccuratelyreflecttheCSofthereferencesets.Somewhat counter-intuitively,trainingsetsizeswereonlyslightlycorrelatedwith uniquenessandnegativelycorrelatedwithnovelty.Apossibleexplana-tionmightbethataGMthatonlylearnedthegrammarandisnot(yet) restrictedbythepropertiesofthereferencespacemightbelesslikely toreproducetrainingmoleculeswhilegeneratinguniquemoleculesata similarlevel.

6

*M.Vogt*  *ArtificialIntelligenceintheLifeSciences3(2023)100064*

**Dataavailability**

Nodatawasusedfortheresearchdescribedinthearticle.

**References**

[[1]](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0002)[KirkpatrickP,EllisC.Chemicalspace.Nature2004;432:823823.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0001)

[[2]](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0002)[BohacekRS,McMartinC,GuidaWC.Theartandpracticeofstr](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0001)[ucture-baseddrug](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0002)  [design:amolecularmodelingperspective.MedResRev1996;16:3–50.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0003)

[[3]FinkT,BruggesserH,ReymondJL.Virtualexplorationofthesmall-moleculechem-](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0003) [icaluniversebelow160Daltons.AngewChemIntEd2005;44:1504–8.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0003)

[[4]BlumLC,ReymondJL.970milliondruglikesmallmoleculesforvirtualscreeningin](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0003)  [thechemicaluniversedatabaseGDB-13.JAmChemSoc2009;131:8732–3.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0005)

[[5]RuddigkeitL,vanDeursenR,BlumLC,etal.Enumerationof166billionorganic smallmoleculesinthechemicaluniversedatabaseGDB-17.JChemInfModel 2012;52:2864–75.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0005)

[[6]WeiningerD.SMILES,achemicallanguageandinformationsystem.1.Introduction](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0005)  [tomethodologyandencodingrules.JChemInfModel1988;28:31–6.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0006)

[[7]BorelE.Lamécaniquestatiqueetl’irréversibilité.JdePhysiqueThéoriqueetAp-](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0006) [pliquée1913;3:189–96.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0008)

[[8]WermuthCG,AldousD,RaboissonP,etal.Thepracticeofmedicinalchemistry.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0008)  [AcademicPress;2015.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0008)

[[9]VogtM.Howdoweoptimizechemicalspacenavigation?ExpertOpinDrugDiscov](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0008)  [2020;15:523–5.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0010)

[[10]ErtlP,Schuffenhaue](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0010)r[A.Estimationofsyntheticaccessibilityscoreofdrug-like moleculesbasedonmolecularcomplexityandfragmentcontributions.JCheminf 2009;1:8.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0010)

[[11]BickertonGR,PaoliniGV,BesnardJ,etal.Quantifyingthechemicalbeautyofdrugs.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0010)  [NatChem2012;4:90–8.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0011)

[[12]SchneiderG,FechnerU.Computer-baseddenovodesignofdrug-likemolecules.Nat](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0011)  [RevDrugDiscov2005;4:649–63.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0012)

[[13]HartenfellerM,ZettlH,WalterM,etal.DOGS:Reaction-drivendenovodesignof](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0012)  [bioactivecompounds.PLoSComputBiol2012;8:e1002380.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0014)

[[14]YonchevD,BajorathJ.Integratingcomputationalleadoptimizationdiagnosticswith](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0014)  [analogdesignandcandidateselection.FutureSciOA2020;6:FSO451.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0014)

[[15]Gómez-BombarelliR,WeiJN,DuvenaudD,etal.Automaticchemicaldesignusinga](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0014) [data-drivencontinuousrepresentationofmolecules.ACSCentralSci2018;4:268–76. [16]ColbySM,NuñezJR,HodasNO,etal.Deeplearningtogenerateinsilicochemi-calpropertylibrariesandcandidatemoleculesforsmallmoleculeidentificationin](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0016) [complexsamples.AnalChem2019;92:1720–9.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0017)

[[17]BaskinI,GordeevaE,DevdarianiR,etal.Methodologyoftheinverseproblemso-lutionforthestructurepropertyrelationincaseoftopologicalindices.DoklAkad](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0017) [NaukSSSR1989;307:613–17.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0018)

[[18]BrüggemannR,PudenzS,CarlsenL,etal.TheuseofHassediagramsasapotential](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0018)  [approachforinverseQSAR.SARQSAREnvironRes2001;11:473–87.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0019)

[[19]MiyaoT,FunatsuK.Findingchemicalstructurescorrespondingtoasetofcoordi-](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0019) [natesinchemicaldescriptorspace.MolInform2017;36:1700030.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0019)

[[20]Sanchez-LengelingB,Aspuru-GuzikA.Inversemoleculardesignusingmachine](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0019) [learning:generativemodelsformatterengineering.Science2018;361:360–5.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0021)

[[21]VogtM.Usingdeepneuralnetworkstoexplorechemicalspace.ExpertOpinDrug](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0021)  [Discov2021;17:297–304.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0021)

[[22]SkinniderMA,StaceyRG,WishartDS,etal.Chemicallanguagemodelsenablenav-](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0021)[igationinsparselypopulatedchemicalspace.NatMachIntel2021;3:759–70.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0022)

[[23]JinW,BarzilayR,JaakkolaT,DyJ,KrauseA.Junctiontreevariationalautoencoder](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0022) [formoleculargraphgeneration.In:Proceedingsofthe35thinternationalconference onmachinelearning,80.PMLR;2018.p.2323–32.ProceedingsofMachineLearning Research10–15Jul.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0023)

[[24]YouJ,LiuB,YingR,etal.Graphconvolutionalpolicynetworkforgoal-directed](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0023) [moleculargraphgeneration.In:NIPS’18:Proceedingsofthe32ndinternationalcon-ferenceonneuralinformationprocessingsystems;2018.p.6412–22.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0024)

[[25]LiY,ZhangL,LiuZ.Multi-objectivedenovodrugdesignwithconditionalgraph](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0024)  [generativemodel.JCheminf2018;10:33.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0025)

[[26]MercadoR,RastemoT,LindelöfE,etal.Graphnetworksformoleculardesign.Mach](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0025)  [LearnSciTechnol2021;2:025023.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0026)

[[27]GoodfellowI,BengioY,CourvilleA.Deeplearning.Cambridge,MA:TheMITPress;](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0026)  [2017.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0027)

[[28]HellerSR,McNaughtA,PletnevI,etal.InChI,theIUPACinternationalchemical](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0027)  [identifier](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0029).[JCheminf2015;7:23.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0029)

[[29]BjerrumE,SattarovB.Improvingchemicalautoencoderlatentspaceandmolecular denovogenerationdiversitywithheteroencoders.Biomolecules2018;8:131.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0029)

[[30]Arús-PousJ,JohanssonS,PrykhodkoO,etal.RandomizedSMILESstringsimprove](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0029) [thequalityofmoleculargenerativemodels.ChemRxiv2019.doi:10.26434/chem-rxiv.8639942.v2.](https://doi.org/10.26434/chemrxiv.8639942.v2)

[[31]O’BoyleN,DalkeA.DeepSMILES:anadaptationofSMILESforusein](https://doi.org/10.26434/chemrxiv.8639942.v2) [machine-learningofchemicalstructures.ChemRxiv2018.doi:10.26434/chem-rxiv.7097960.v1.](https://doi.org/10.26434/chemrxiv.7097960.v1)

[[32]KrennM,HäseF,NigamA,etal.Self-referencingembeddedstrings(SELF-](https://doi.org/10.26434/chemrxiv.7097960.v1)[IES):a100%robustmolecularstringrepresentation.MachLearnSciTechnol 2020;1:045024.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0032)

[[33]WangR,GaoY,LaiL.LigBuilder:amulti-purposeprogramforstructure-baseddrug](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0032)  [design.JMolModel2000;6:498–516.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0033)

[[34]ChéronN,JastyN,ShakhnovichEI.OpenGrowth:anautomatedandrationalalgo-](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0033) [rithmforfindin](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0035)g[newproteinligands.JMedChem2015;59:4171–88.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0035)

[[35]](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0034)[KutchukianPS,LouD,ShakhnovichEI.FOG:Fragmentoptimizedgrowthalgorithm](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0035)

7

*M.Vogt*  *ArtificialIntelligenceintheLifeSciences3(2023)100064*

[[72]PreuerK,RenzP,UnterthinerT,etal.FréchetChemNetdistance:Ametricforgener-ativemodelsformoleculesindrugdiscovery.JChemInfModel2018;58:1736–41. [73]HeuselM,RamsauerH,UnterthinerT,etal.GANstrainedbyatwotime-scale](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0072) [updateruleconvergetoalocalNashequilibrium.AdvNeuralInfProcessSyst 2017;30:6627–38.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0073)

[[74]SalimansT,GoodfellowI,ZarembaW,LeeD,SugiyamaM,LuxburgU,etal.Im-](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0073)[provedtechniquesfortrainingGANs.AdvNeuralInfProcessSyst2016;29:2234–42. [75]SattarovB,BaskinII,HorvathD,etal.Denovomoleculardesignbycombining deepautoencoderrecurrentneuralnetworkswithgenerativetopographicmapping. JChemInfModel2019;59:1182–96.](http://refhub.elsevier.com/S2667-3185(23)00008-9/sbref0075)

8