

[Artificia](https://doi.org/10.1016/j.ailsci.2023.100066)l[IntelligenceintheLifeSciences3(2023)100066](https://doi.org/10.1016/j.ailsci.2023.100066)

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

ShortCommunications   
Naturalproductssubsets:Generationandcharacterization   
AnaL.Chávez-Hernández,José L.Medina-Franco∗  
*DIFACQUIMResearchGroup,DepartmentofPharmacy,SchoolofChemistry,UniversidadNacionalAutónomadeMéxico,AvenidaUniversidad3000,MéxicoCity 04510,Mexico*

|  |  |  |
| --- | --- | --- |
| article | info | abstract |
| *Keywords:*  Artificialintelligence  Chemicalspace  Chemicalmultiverse  Chirality  Denovodesign  Deeplearning  Molecularfingerprint  Molecularrepresentation Naturalproducts | | Naturalproductsareattractivefordrugdiscoveryapplicationsbecauseoftheirdistinctivechemicalstructures, suchasanoveralllargefractionofsp3carbonatoms,chiralcenters(bothfeaturesassociatedwithstructural complexity),largechemicalscaffolds,anddiversityoffunctionalgroups.Furthermore,naturalproductsareused in*denovo*designandhaveinspiredthedevelopmentofpseudo-naturalproductsusinggenerativemodels.Public databasessuchastheCollectionofOpenNatUralProdUcTsandtheUniversalNaturalProductdatabase(UNPD) arerichsourcesofstructurestobeusedingenerativemodelsandotherapplications.Inthiswork,wereport theselectionandcharacterizationofthemostdiversecompoundsofnaturalproductsfromtheUNPDusing theMaxMinalgorithm.Thesubsetsgeneratedwith14,994,7,497,and4,998compoundsarepubliclyavailable at<https://github.com/DIFACQUIM/Natural-products-subsets-generation>.Weanticipatethatthesubsetswillbe pa[rticularlyusefulinbuildinggenerativemodelsbasedonnaturalproduct](https://github.com/DIFACQUIM/Natural-products-subsets-generation)sbyresearchgroups,particularlythose withlimitedaccesstoextensivesupercomputerresources. |

**1.Introduction**

Naturalproductsandfragmentsderivedfromnaturalproductshave beenattractiveindrugdesignanddevelopmentbecauseoftheirdistinc-tivechemicalstructures.Forexample,naturalproductshave,ingeneral, anoveralllargerdiversityoffunctionalgroupsandlargerstructural complexitythansyntheticmolecules[1–4].However,adrawbackfor somenaturalproducts,particularlythosewithsizeablestructuralcom-plexity,isthattheycanbechallengingtosynthesize.Aworkaroundfor thisissueistheso-calledpseudo-naturalproductswhicharesyntheti-callyfeasiblecompoundsgeneratedthrougha*denovo*combinationof naturalproductfragments[3].Pseudo-naturalproductsallowtheexplo-rationofunchartedareasofbiologicallyrelevantchemicalspacethat aredifferentfromthechemicalspacecoveredbythecompoundsfrom whichtheyaregenerated.

TheCollectionofOpenNatUralProdUcTs(COCONUT)andthe UniversalNaturalProductDatabase(UNPD)aretwolargecompound databases.COCONUT[5]isarguablythemostextensivepublicnatu-ralproductdatabase,with389,184uniquestructures.UNPD[6],with 153,375naturalproducts,isthesecond-largestpublicnaturalprod-uctdatabase.AdistinctivefeatureofUNPDcomparedtoCOCONUTis thatcompoundsinUNPDcontainchiralityinformation.InLatinAmer-ica,severalpublicnaturalproductsdatabasescompilethecompounds isolatedandcharacterizedfromthecountryoforigin.Examplesare

NuBBEDB[7,8],SistematX[9,10]fromBrazil;CIFPMA[11,12]from Panama;PeruNPDB[13]fromPeru;andBIOFACQUIM[14,15]from Mexico.Thelatterdatabasecontainsthestructuresof531naturalprod-ucts.

*Denovo*designgeneratesvirtualmoleculesfromscratch.Itfilters structuresgeneratedusingseveralscoringfunctionsandassessessyn-theticchemicalfeasibilitytoremovereactiveandunrealisticcompounds [16].*Denovo*designbasedongenerativealgorithmssuchasdeeplearn-inginvolvesusingneuralnetworks[17,18]anddatabaseswithmany compounds[19].Inthesourcecompounddatabases,manycompounds withthree-dimensionalinformation(*e.g.*,stereochemistry)arerelevant tobuildrobustgenerativemodels[20].However,forseveralresearch groups,itisdifficulttoaccesssupercomputerresourcestohandlemany compoundstoobtainappropriatesubsetsthatimpactthemodelpredic-tion[21,22].

ThisCommunicationreportstheselectionandcharacterizationofthe mostdiversecompoundsfromUNPD.Thesubsetswereselectedusinga dissimilarity-basedcompoundselection(DBCS)method,theMaxMinal-gorithm[23].ThecompoundsubsetswerecharacterizedbytheNatural ProductLikeness(NPL)score[24],structuraldiversity,anddistribution inchemicalspace.ThestructuraldiversitywasassessedwiththeTani-motocoefficientandmolecularfingerprintsofdifferentdesigns.Chem-icalspaceanalysiswasperformedthroughthevisualrepresentationof thechemicalmultiverse[25]usingT-distributedStochasticNeighbor

*Abbreviations:*COCONUT,CollectionofOpenNatUralProdUcTs;DNMT,DNAmethyltransferase;ECFP,extendedconnectivityfingerprint;TMAP,TreeMAP;NPL, naturalproduct-likeness;UNPD,UniversalNaturalProductdatabase.  
 ∗Correspondenc[eauthor.](mailto:medinajl@unam.mx)

*E-mailaddress:*[medinajl@unam.mx](mailto:medinajl@unam.mx)(J.L.Medina-Franco).

<https://doi.org/10.1016/j.ailsci.2023.100066>  
[Received20December2022;Receivedinrevi](https://doi.org/10.1016/j.ailsci.2023.100066)sedform5February2023;Accepted21February2023   
Availableonline26February2023   
2667-3185/© 2023TheAuthors.PublishedbyElsevierB.V.ThisisanopenaccessarticleundertheCCBYlicense(<http://creativecommons.org/licenses/by/4.0/>)

*A.L.Chávez-HernándezandJ.L.Medina-Franco*



*ArtificialIntelligenceintheLifeSciences3(2023)100066*

**Fig.1.**MaxMinalgorithmimplementedinthiswork.Adatabase (UNPD)wassplitintoagivennumberofsubsets.Fromeachsub-set,arandomcompound(X)wasselected.Themolecularsimilar-itywascalculatedbetweenXandeachcompound(A,B,C,D,and E)fromthesubsetusingtheTanimotocoefficientandtheECFP4 fingerprint.Inthisfigure,onlyasubset(thatcontainsthecom-poundsA,B,C,D,andE)isshownforillustrativepurposes.The compoundwiththesmallestmolecularsimilaritywaschosenand deletedfromtheoriginalsubset.Theprocesswasrepeatedtogen-erateanewsubsetwiththenumberofcompoundsdesired.

|  |  |
| --- | --- |
| Embeddingt-SNE[26]andtherecentlyproposedTreeMAP(TMAP) [27].Weanticipatethatthenaturalproductsubsetsgeneratedandthor-oughlycharacterizedinthisworkwillbehelpfultothescientificcom-munityinavarietyoftasksincludingbuildinggenerativemodels,in particulartothoseresearchgroupswithlimitedaccesstolargesuper-computerresources.  **2.Materialsandmethods**  *2.1.Datasets*  Forthisstudy,differenttypesofdatabaseswithcompoundsfromvar-iousoriginswereusedincludingthreedatasetsofnaturalproducts,and adatasetfocusedonaspecificmoleculartargetofpharmaceuticalrele-vance.Naturalproductdatabasesfromdifferentsourceswereemployed tocomparetheNPLscoreoftheUNPDsubsets(describedinthe*Natu-ralproduct-likeness*section).Thedatasetfocusedonaspecificmolecu-lartargetandthenaturalproductsfromdifferentsourceswereusedto evaluateandcomparethemoleculardiversityofUNPDsubsets.Usinga focuseddatabasegivesavalueoflowstructuraldiversity,whichserves asareferencetocomparethestructuraldiversityoftheUNPDsubsets. Likewise,thechemicalstructuresfromothernaturalproductdatabases servedasreferencecompoundcollectionstoevaluatethediversityof thedatasetsnewlygenerated.Eachtypeofdatasetanditscontentsare describedbelow. | Threedatasetsofnaturalproductswereused,namely,COCONUT, UNPD,andBIOFACQUIM.COCONUT[5]had389,184uniquestruc-tures,notincludingchiralityinformation.UNPD[6]with153,372 naturalproducts,encodingtheirchirality.BIOFACQUIM[14,15]isa databasewith531naturalproductsisolatedandcharacterizedinMex-ico,andthechiralityofthecompoundsisincluded.Adatasetwith715 compoundsfocusedonDNAmethyltransferase1(DNMT1)inhibitors. DNMT1isanepigenetictargetrelevanttodrugdiscovery[28,29].The datasetofDNMT1inhibitorswasobtainedfromtheChEMBLdatabase, release31[30,31].  *2.2.Datasetstandardization*  Thepreparationofcompounds,encodedinSMILESstrings[32]was performedusingtheopen-sourcecheminformaticstoolkitRDKit [33]andMolVS[34].Compoundswithvalenceerrorsoranychemi-calelementotherthanH,B,C,N,O,F,Si,P,S,Cl,Se,Br,andI,were removed.Stereochemistryinformation,whenavailable,wasretained. Compoundswithmultiplecomponentsweresplit,andthelargestcom-ponentwasretained.Theremainingcompoundswereneutralizedand reionizedtogeneratethecorrespondingcanonicaltautomer.  *2.3.Naturalproduct-likeness*  TheNPLscore,introducedbyErtletal.[24],wascomputedfor allcompoundsinCOCONUT,UNPD,andBIOFACQUIM.TheNPLscore |

2

*A.L.Chávez-HernándezandJ.L.Medina-Franco*

**Table1**   
DescriptivestatisticsofNPLscorescomputedforCOCONUT,UNPD,andBIO-FACQUIM.

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | COCONUT | UNPD | BIOFACQUIM |
| count  mean  astd  bmin  cQ1  median  dQ3  emax | 389,184  0.89  1.38 −3.53 −0.32  0.97  2.01  4.08 | 153,372  1.81  0.96 −2.51  1.14  1.87  2.56  4.08 | 531  1.73  0.90 −0.36  1.03  1.65  2.45  3.87 |

astd:standarddeviation.

bmin:minimumvalue.

cQ1:valueunderwhich25%ofdatapointsarefoundinincreasingorder. dQ3:valueunderwhich75%ofdatapointsarefoundinincreasingorder. emax:maximumvalue.

rangesbetween−5(ifthecompoundismoresimilartoasyntheticcom-pound)and5(ifthecompoundismoresimilartoanaturalproduct).

*2.4.Selectionofsetsofdiversecompounds*

Dissimilarity-basedcompoundselection(DBCS)methodsallowthe identificationofdiversecompounddatasetsbydirectlycalculatingdis-tancesordissimilaritiesbetweenthechemicalstructures[23].Among theDBCSmethods,theMaxMinalgorithmreducesthenumberofcom-poundschoosingthemoleculeswiththelargestdiversityfromtheorigi-naldatabases.ThreesubsetsofUNPDweregeneratedusingtheMaxMin algorithmasfollows(Fig.1):Theinitial153,372compoundsfromUNPD weresplitintothreeways:(A)thirtyrandomsubsetsof5,000com-poundseach;(B)fifteenrandomsubsetsof10,000compoundseach;and (C)tenrandomsubsetsof15,000compoundseach.Anewdiverseset with500compoundswasselectedfromeachoneusingMaxMin[23]. First,arandomcompoundwaspickedfromeachsubset.Thebinary similaritybetweenthecompoundselected(queryset)andtheremain-ingcompounds(targetset)wascalculated.Anewcompoundwasse-lectedfromthetargetsetifthishadthelowestsimilarityvalueand thenremovedfromthetargetset.Theiterationprocesscontinuedun-tilthenumberofcompoundsdesiredsetto500wasreached.Intotal, threediversesubsetsfromUNPDweregenerated:UNPD-A(15,000com-pounds),UNPD-B(7,500),andUNPD-C(5,000).Forthediversityselec-tion,weusedtheTanimotocoefficientandtheextendedconnectivity fingerprint(ECFP)[35]of1024-bitsanddiameter4(ECFP4).Forthedi-versitysetcalculations,weusedanE5–2670v1processor,16cores,and 64GbytesofRAM).Themaximumcalculationtimeforeachinitialsub-setofthenaturalproductsubsetswere:UNPD-A(15,000compounds), 19,989.21s;UNPD-B(7,500compounds),102,569.61s,andUNPD-C (5,000compounds),209,241.25s.

*2.5.Structuraldiversity*

ThestructuraldiversityofthreeUNPDsubsets,UNPD,BIOFAC-QUIM,andDNMT1,wascomparedthroughthedistributionofthepair-wisesimilarityvaluesgeneratedwiththeTanimotocoefficientusing threemolecularfingerprintsMolecularACCesSystem(MACCS)Keys (166-bits)[36],extendedconnectivityfingerprint(ECFP)[35]of1024-bitswithdiameter6(ECFP6)anddiameter4(ECFP4).

*2.6.Chemicalmultiversevisualization*

Thechemicalmultiverse[25]ofthethreeUNPDsubsetswascom-paredtothechemicalmultiverseoftheentireUNPDcollection.Achem-icalmultiverseisagroupofnumericalvectorsthatdifferentlydescribe asetofmoleculesdependingonthemolecularrepresentation[25].So, eachchemicalspaceisanM-dimensionalcartesianspaceinwhichcom-poundsarelocatedbyasetofMdescriptors.Eachtypeofmolecular

3

*A.L.Chávez-HernándezandJ.L.Medina-Franco*  *ArtificialIntelligenceintheLifeSciences3(2023)100066*

**Table2**   
SummaryofthestructuraldiversityofthethreeUNPDsubsets,BIOFACQUIM,andDNMT1datasets.Thenumberofinitialanduniquecompoundsfromeach databaseisindicated.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | Compounds | Uniquecompounds | MACCSkeys(166-bits)a | ECFP4a | ECFP6a |
| UNPD-A  UNPD-B  UNPD-C  UNPD  BIOFACQUIM DNMT1 | 15,000  7,500  5,000  153,372  531  714 | 14,994  7,497  4,998  153,372  531  714 | 0.341  0.346  0.356  0.43  0.447  0.417 | 0.091  0.094  0.092  0.111  0.119  0.119 | 0.077  0.08  0.079  0.094  0.099  0.1 |

aThestructuraldiversityisreportedasthemedianvalueofthedistributionofthepairwisecomparisonusingtheTanimotocoefficientandmolecular fingerprints(MACCSkeys,ECFP4andECFP6).Afulldiversityassessmentispresentedatthecumulativedistributionfunctionsofthepairwisesimilarityvalues inFig.2.

**Table3**   
Descriptivestatisticsofthenumberofhydrogenbonddonorsandacceptors(HBD,HBA)oftheUNPDsubsetsandtheentireUNPD.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Property | HBD |  |  |  | HBA |  |  |  |
| Dataset | UNPD-A | UNPD-B | UNPD-C | UNPD | UNPD-A | UNPD-B | UNPD-C | UNPD |
| count  mean  astd  bmin  cQ1  median  dQ3  emax | 14,994  2.51  3.17  0.00  0.00  2.00  3.00  36.00 | 7,497  2.6  3.10  0.00  1.00  2.00  3.00  33.00 | 4,998  2.69  3.40  0.00  0.00  2.00  4.00  33.00 | 153,372 3.15  3.55  0.00  1.00  2.00  4.00  36.00 | 14,994  5.58  4.95  0.00  2.00  4.00  7.00  53.00 | 7,497  5.65  4.80  0.00  3.00  5.00  7.00  53.00 | 4,998  5.94  5.46  0.00  2.00  5.00  8.00  52.00 | 153,372  7.05  5.68  0.00  3.00  5.00  9.00  54.00 |

astd:standarddeviation.

bmin:minimumvalue.

cQ1:valueunderwhich25%ofdatapointsarefoundinincreasingorder.   
dQ3:valueunderwhich75%ofdatapointsarefoundinincreasingorder.   
emax:maximumvalue.

**Table4**   
Descriptivestatisticsofthenumberofrotatablebonds(RB)andLogPvaluesoftheUNPDsubsetsandtheentireUNPD.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Property | RB |  |  |  | LogP |  |  |  |
| Dataset | UNPD-A | UNPD-B | UNPD-C | UNPD | UNPD-A | UNPD-B | UNPD-C | UNPD |
| count  mean  astd  bmin  cQ1  median  dQ3  emax | 14,994  4.74  6.02  0.00  1.00  3.00  6.00  59.00 | 7,497  5.34  6.66  0.00  1.00  3.00  7.00  63.00 | 4,998  4.81  5.69  0.00  1.00  3.00  6.00  59.00 | 153,372 5.97  6.08  0.00  2.00  4.00  8.00  63.00 | 14,994  2.94  3.02 −18.53  1.46  2.87  4.32  24.43 | 7,497  2.98  3.00 −14.10  1.40  2.79  4.24  23.11 | 4,998  2.94  3.13 −18.16  1.43  2.90  4.45  21.63 | 15,372  2.94  3.12 −20.82  1.33  2.96  4.58  25.12 |

astd:standarddeviation.

bmin:minimumvalue.

cQ1:valueunderwhich25%ofdatapointsarefoundinincreasingorder.   
dQ3:valueunderwhich75%ofdatapointsarefoundinincreasingorder.   
emax:maximumvalue.

**Table5**   
Descriptivestatisticsofthetopologicalsurfacearea(TPSA)andmolecularweight(MW)oftheUNPDsubsetsandtheentireUNPD.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Property | TPSA |  |  |  | MW |  |  |  |
| Dataset | UNPD-A | UNPD-B | UNPD-C | UNPD | UNPD-A | UNPD-B | UNPD-C | UNPD |
| count  mean  astd  bmin  cQ1  median  dQ3  emax | 14,994  90.78  82.74  0.00  40.46  69.67  112.05  877.36 | 7,497  93.58  80.35  0.00  46.53  74.60  116.20  927.43 | 4,998  94.78  90.04  0.00  39.10  69.92  119.61  900.36 | 153,372 114.63  93.94  0.00  54.37  85.97  145.91  927.43 | 14,994  371.94  196.43  16.04  246.31  330.29  445.60  1,887.28 | 7,497  369.57  194.20  16.04  248.32  328.45  444.48  1,889.30 | 4,998  388.43  210.40  17.03  252.28  342.47  466.74  1,875.31 | 153,372 450.55  228.63  16.04  302.28  396.20  534.66  1,907.36 |

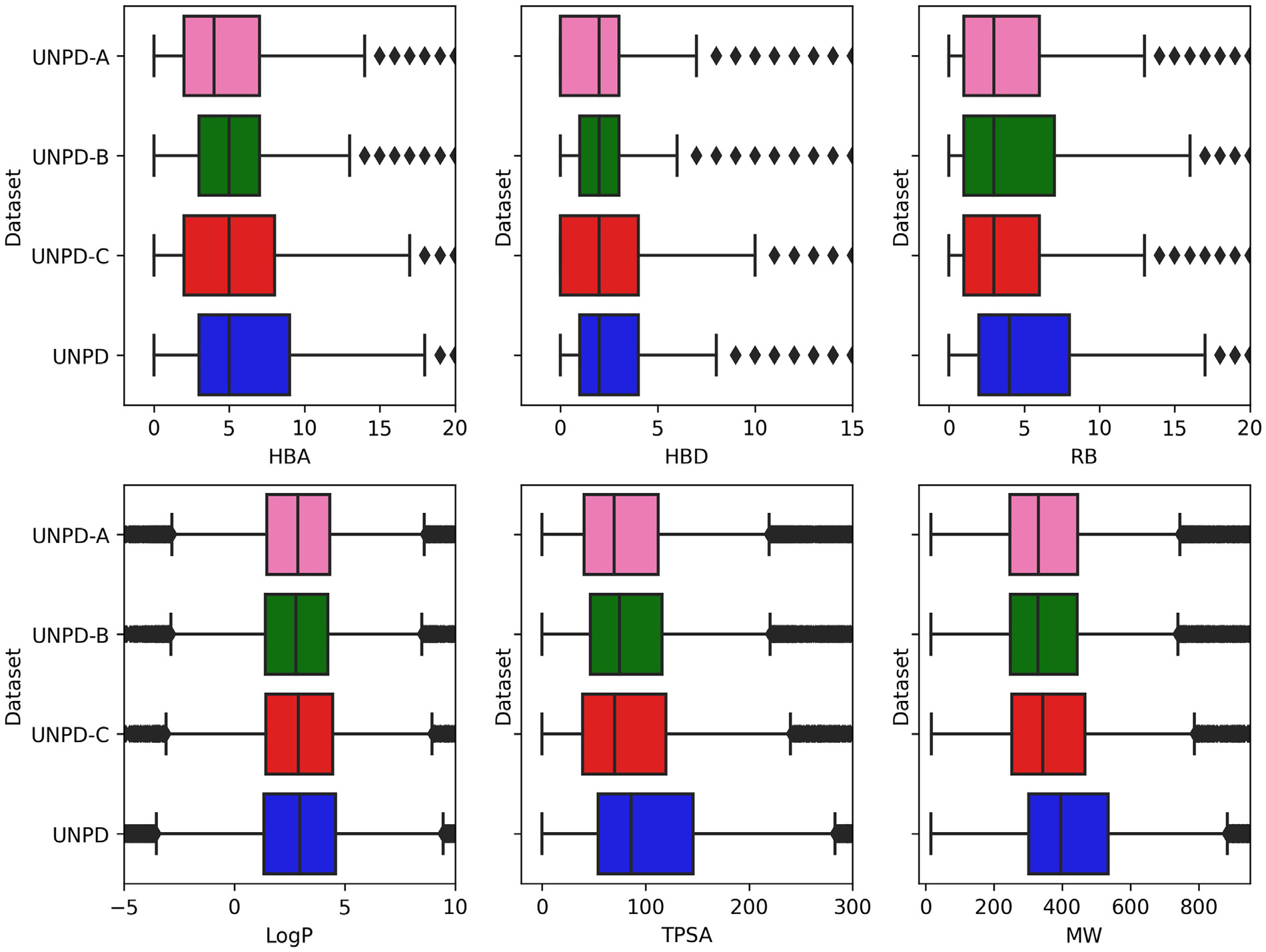
astd:standarddeviation.

bmin:minimumvalue.

cQ1:valueunderwhich25%ofdatapointsarefoundinincreasingorder.   
dQ3:valueunderwhich75%ofdatapointsarefoundinincreasingorder.   
emax:maximumvalue.

4

*A.L.Chávez-HernándezandJ.L.Medina-Franco*  *ArtificialIntelligenceintheLifeSciences3(2023)100066*



**Fig.3.**Box-whiskerplotsofsixpropertiesofpharmaceuticalrelevance:hydrogenbonddonors(HBD),hydrogenbondacceptors(HBA),topologicalpolarsurface

area(TPSA),numberofrotatablebonds(RB),molecularweight(MW),andpartitioncoefficientoctanol/water(LogP).Thedatasetsarerepresentedindifferent

colors,UNPD-A(pink),UNPD-B(green),UNPD-C(red),andUNPD(blue).Blackdiamondsrepresentoutliers.

fingerprintgeneratesadistinctchemicalspace,andthesetsofchem-icalspacescomprisethechemicalmultiverse.Forthiswork,thevi-sualrepresentationofthechemicalmultiversewasdoneusingt-SNE [26]andTMAPs[27]asvisualizationmethods,andMACCSkeys(166 bits),ECFP4,andECFP6asmolecularrepresentations.

Thet-SNEgeneratesplotsthatorganizecompounds.Similarcom-poundsformclustersanddissimilarcompoundsaredistantfromeach other.TMAPallowsvisualizingofmanychemicalcompoundsthrough thedistancebetweenclusters.LocalSensitiveHashing(LSH)allowseach compoundtobegroupedhierarchicallyaccordingtocommonsubstruc-turesusingmolecularfingerprints.Thenumberofnearestneighbors, *k*=50,andthefactorusedbytheaugmentedqueryalgorithm,kc=10, wereusedtodeveloptheTMAPgraphs.Inpreviousstudies,weused TMAPtodescribethemoleculardiversityofnaturalproductssuchas COCONUT[2],BIOFACQUIM[15],andafocusedlibraryofHIV-1pro-teaseinhibitorsusingnaturalfragmentsfromCOCONUT[37].

**3.Resultsanddiscussion**

*3.1.Natural-product-likeness*

Table1summarizesthedescriptivestatisticsforNPLscorescalcu-latedforBIOFACQUIM,UNPD,andCOCONUT.TheCOCONUT,UNPD, andBIOFACQUIMhadNPLscorevaluesintherangeof[−3.53,4.08], [−2.51,4.08],and[−0.36,3.87],respectively.Asanticipated,natural productcompoundsfromCOCONUT,UNPD,andBIOFACQUIMhad NPLscores*ca.*4.Only25%ofCOCONUT´scompounds(min=−3.53, Q1=−0.32)hadNPLscores*ca.*−3.53,and25%ofUNPD´scompounds

5

*A.L.Chávez-HernándezandJ.L.Medina-Franco*  *ArtificialIntelligenceintheLifeSciences3(2023)100066*



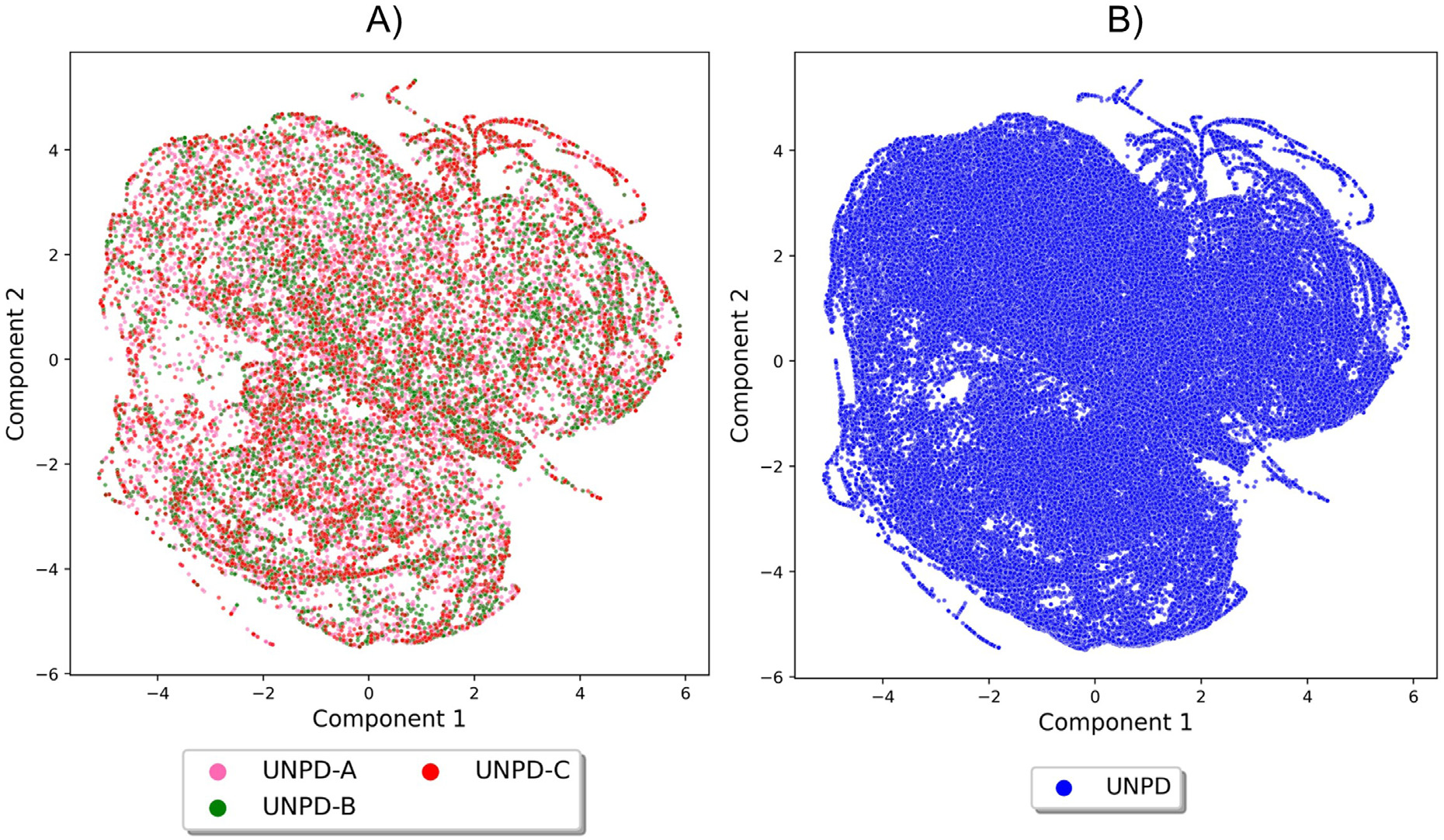
**Fig.4.**ChemicalmultiversevisualizationofUNDPsubsetsusingTMAPsandMACCSkeys(166-bits),ECFP4andECFP6,asmolecularrepresentations.Thedatasets

arerepresentedasUNPD-A(pink),UNPD-B(green),andUNPD-C(red).Theoverlapbetweennaturalproductsubsetsisrepresentedingray.

versedatasetandhadthelowestmediansimilarityvalue(0.091, 0.077)calculatedwithECFP4andECFP6,respectively,followedby theUNPD-B(0.094,0.08),andUNPD-C(0.092,0.079);andtherefer-encedatabases:UNPD(0.111,0.094),BIOFACQUIM(0.119,0.099)and DNMT1(0.119,0.1).ThestructuraldiversitycalculatedwithMACCS keys(Table2andFig.2)indicatedthatUNPDsubsetshadthesame trend,beingtheUNPD-Athemostdiverse(median=0.341)followed byUNPD-B(median=0.346),andUNPD-C(median=0.356);followed byUNPD(median=0.43),BIOFACQUIM(median=0.447),andDNMT1 (median=0.417).ForUNPD,thedatasetgeneratedfromalargernum-berofsubsets(thirty*versus*fifteenortensubsets)wasthemostdiverse, *i.e.*,UNPD-AaccordingtotheECFP4,ECFP6,andMACCSkeysfinger-prints.

6

*A.L.Chávez-HernándezandJ.L.Medina-Franco*  *ArtificialIntelligenceintheLifeSciences3(2023)100066*



**Fig.5.**ChemicalspacevisualizationofUNPDandUNPDsubsetsusingt-SNEbasedontheproperties:hydrogenbonddonors,hydrogenbondacceptors,topological

polarsurfacearea,numberofrotatablebonds,molecularweight,andpartitioncoefficientoctanol/water.ThedatasetsarerepresentedinscatterpointsasUNPD-A

(pink),UNPD-B(green),UNPD-C(red),andUNPD(blue).PanelAshowstheUNPDsubsets(A-C),andpanelBshowstheentireUNPD.

B:LogP*<*=4.24,UNPD-C:LogP*<*=4.45,andUNPD:LogP*<*=4.58.Re-gardingMWandTPSA,75%ofUNPD’scompounds(TPSA*<*=145.91 andMW*<*=534.66)weremorediversethan75%ofUNPDsubsets (A-C):UNPD-A’scompounds(TPSA*<*=112.05,MW*<*=445.60);UNPD-B’scompounds(TPSA*<*=116.20,MW*<*=444.48),andUNPD-C’scom-pounds(TPSA*<*=116.61,MW*<*=466.74).RegardingRB,UNPD’scom-pounds(RB*<*=8.0)andUNPD-B’scompounds(RB*<*=7.0)weremoredi-versethanUNPD-A’scompounds(RB*<*=6.0)andUNPD-C’scompounds (RB*<*=6.0).RegardingHBAandHBD,UNPD’scompounds(HBD*<*=4.0, HBA*<*=9.0)andUNPD-C’scompounds(HBD*<*=4.0,HBA*<*=8.0)were morediversethanUNPD-A(HBD*<*=3.0,HBA*<*=7.0)andUNPD-B (HBD*<*=4.0,HBA*<*=7.0).Overall,theUNPD-Cdatasetwasthemostdi-verseintermsofthepropertiesofpharmaceuticalrelevanceafterthe entireUNPD.

*3.4.Visualizationofthechemicalspaceandchemicalmultiverse*

AsdescribedinSection2.6,achemicalmultiverseisconceptualized asagroupofmolecularrepresentationsthateachdescribeacompound dataset(incontrasttoachemicalspacethatisdefinedbyonlyonesetof descriptors).Fig.4showsavisualrepresentationofthechemicalmul-tiverseoftheUNPDdatasetsusingTMAPsandMACCSkeys(166-bits), ECFP4,andECFP6asmolecularrepresentations.Inthisstudy,thechem-icalmultiverseiscomprisedofthreemolecularfingerprints:onebased onstructuralkeys,MACCSkeys(Fig.4A);andthehashedmolecular fingerprint,ECFP4andECFP6(Figs.4BandC).Thedatasetsarerepre-sentedinscatterpointswithdifferentcolorsasUNPD-A(pink),UNPD-B (green),andUNPD-C(red).Theoverlapbetweendifferentdatasetsis depictedingray.TheTMAPgeneratedwithMACCSkeysshowsfewer clusterswithmorecompoundsthanECFP4andECFP6.Fornaturalprod-ucts,theTMAPgeneratedwithECFP6showsthatthecompoundsofthe UNPDsubsetsaremoreevenlydistributedwithrespecttotheTMAPs constructedwithECFP4andMACCSkeys.Thechemicalmultiverserep-resentedbyECFP6ismoreaccurateindescribingthestructuraldiversity thanECFP4becauseECFP6encodesmolecularfragmentsinmoredetail

7

*A.L.Chávez-HernándezandJ.L.Medina-Franco*

fromsmalldatasetsofcompoundswithdiversestructuresanddiverse propertiesofpharmaceuticalrelevancehereingenerated;andaddtothe distinctivestructuralcomplexityanddiversityofthenaturalproductsas alargerfractionofsp3carbonatomsandchiralcenters[3,4].

*ArtificialIntelligenceintheLifeSciences3(2023)100066* **Acknowledgments**   
 A.L.C.-H.isthankfultoCONACyT,Mexico,forthePh.D.scholarship number847870.

**4.Conclusions**  **Supplementarymaterials**

Wereporttheselectionandcharacterizationofthethreesubsets withthemostdiversecompoundsfromUNPDusingtheMaxMinalgo-rithm.Threesubsetswith14,994,7497,and4998compoundsselected fromtheUNPDcontainthemoststructurallydiversenaturalproducts. UnlikecompoundsintheCOCONUTdatabase,moleculesinUNPDare annotatedwithchirality.Thestructuraldiversityofcompoundsisnot affectedbythenumberofsubsetsderivedfromtheoriginaldatabase fromwhichanewdatabaseisgenerated,andananalysisofthechem-icalmultiversesupportsthatUNPDsubsetscontainthemostdiverse molecules.Duringthestudy,wealsoconcludedthatthevisualization ofthechemicalspacedescribedwithECFP6ismoreaccuratetode-scribethestructuraldiversityofcompoundscomparedwithECFP4and MACCSkeys(166bits).Thenaturalproductsubsetshadalargediversity ofchemicalcompoundswithdifferentstructuralfeaturesandproperties ofpharmaceuticalrelevance.TheNPLscoresupportsthatthechemical structuresofnaturalproductsareverydifferentanddiverseasdefined bythethresholdoftheNPLscore,andasexpected,naturalproducts (COCONUT,UNPD,andBIOFACQUIM)hadNPLscoresvaluescloseto 5.

Asignificantperspectiveofthisworkisthatthenaturalproductsub-setsderivedfromtheUNPDcanbeusedtodevelopgenerativemodels thatusedeeplearningalgorithmsandrequirethemostdiversecom-pounds,suchas*denovo*design.Thenaturalproductssubsetscanalso beusedtodeveloppredictivemodels;forvirtualscreening;andrefer-encedatabasesforevaluatingthestructuraldiversityorsimilarityto aspecificsubset,amongotherapplications.Thepublicavailabilityof thenaturalproductsubsetscansavecostlycomputationalresourcesfor researchgroupswithlimitedaccessibilitytosupercomputermeans.

**Supplementarymaterial**

TheMaxMinalgorithmandstructuraldiversityimplementedin Pythonlanguage,theinteractiveTMAPs,andthethreesubsetsgen-[eratedfromUNPDwith14,994,7,497,and4,998compoundswith stereochemicalinformation,arepubliclyavailableatGitHub:https: //github.com/DIFACQUIM/Natural-products-subsets-generation.](https://github.com/DIFACQUIM/Natural-products-subsets-generation)

**Funding**

AuthorsthanktheDirecciónGeneraldeCómputoydeTecnologías deInformaciónyComunicación(DGTIC),UNAM,forthecomputational resourcestouseMiztlisupercomputeratUNAMunderprojectLANCAD-UNAM-DGTIC-335;andtheinnovationspaceUNAM-HUAWEIthecom-putationalresourcestousetheirsupercomputerunderprojectNo.7“*De-sarrolloyaplicacióndealgoritmosdeinteligenciaartificialparaeldiseñode fármacosaplicablesaltratamientodediabetesmellitusycáncer*”.Wealso thankthefundingofDGAPA,UNAM,ProgramadeApoyoaProyectosde InvestigacióneInnovaciónTecnológica(PAPIIT),grantNo.IN201321.

**DeclarationofCompetingInterest**

Theauthorsdeclarethattheyhavenoknowncompetingfinancial interestsorpersonalrelationshipsthatcouldhaveappearedtoinfluence theworkreportedinthispaper.

**Dataavailability**

AlldataispublicatGitHub.

8

*A.L.Chávez-HernándezandJ.L.Medina-Franco*

[[27]ProbstD,ReymondJ-L.Visualizationofverylargehigh-dimensionaldatasetsas](http://refhub.elsevier.com/S2667-3185(23)00010-7/sbref0027)  [minimumspanningtrees.JCheminform2020;12:12.](http://refhub.elsevier.com/S2667-3185(23)00010-7/sbref0028)

[[28]Prado-RomeroDL,Medina-FrancoJL.Advancesintheexplorationoftheepigenetic](http://refhub.elsevier.com/S2667-3185(23)00010-7/sbref0028)  [relevantchemicalspace.ACSOmega2021;6:22478–86.](http://refhub.elsevier.com/S2667-3185(23)00010-7/sbref0028)

[[29]ConeryAR,RocnikJL,TrojerP.Smallmoleculetargetingofchromatinwritersin](http://refhub.elsevier.com/S2667-3185(23)00010-7/sbref0028)  [cancer.NatChemBiol2022;18:124–33.](http://refhub.elsevier.com/S2667-3185(23)00010-7/sbref0029)

[[30]MendezD,GaultonA,BentoAP,ChambersJ,DeVeijM,FélixE,MagariñosMP,](http://refhub.elsevier.com/S2667-3185(23)00010-7/sbref0029) [MosqueraJF,MutowoP,NowotkaM,Gordillo-MarañónM,HunterF,JuncoL, MugumbateG,Rodriguez-LopezM,AtkinsonF,BoscN,RadouxCJ,Segura-Cabr-eraA,HerseyA,LeachAR.ChEMBL:towardsdirectdepositionofbioassaydata. NucleicAcidsRes2019;47:D930–40.](http://refhub.elsevier.com/S2667-3185(23)00010-7/sbref0030)

[[31]DaviesM,NowotkaM,PapadatosG,DedmanN,GaultonA,AtkinsonF,BellisL,](http://refhub.elsevier.com/S2667-3185(23)00010-7/sbref0030) [OveringtonJP.ChEMBLwebservices:streamliningaccesstodrugdiscoverydata](http://refhub.elsevier.com/S2667-3185(23)00010-7/sbref0031) [andutilities.NucleicAcidsRes2015;43:W612–20.](http://refhub.elsevier.com/S2667-3185(23)00010-7/sbref0032)

[[32]WeiningerD.SMILES,achemicallanguageandinformationsystem.1.Introduction](http://refhub.elsevier.com/S2667-3185(23)00010-7/sbref0032)  [tomethodologyandencodingrules.JChemInfComputSci1988;28:31–6.](http://refhub.elsevier.com/S2667-3185(23)00010-7/sbref0032)

[[33]RDKit,(n.d.).https://www.rdkit.org(accessed08January082022).](http://refhub.elsevier.com/S2667-3185(23)00010-7/sbref0032)

[34]MolVS,(n.d.)[.](https://www.rdkit.org)<https://molvs.readthedocs.io/en/latest/>(accessed08accessedJan- uary2022).

9