IKER GAZTE
NAZIOARTEKO IKERKETA EUSKARAZ

VI. IKERGAZTE
NAZIOARTEKO IKERKETA EUSKARAZ

2025eko maiatzaren 28, 29 eta 30a
Bilbo, Euskal Herria

ANTOLATZAILEA:
Udako Euskal Unibertsitatea (UEU)

ZIENTZIAK ETA NATURA ZIENTZIAK

Molekulen propietateen iragarpena sare neuronalen bitartez datu-urritasun egoeretan

Amaia Elizaran Mendarte
eta *Gustavo Ariel Schwartz*

101-108 or.
https://dx.doi.org/10.26876/ikergazte.vi.05.12

Molekulen propietateen iragarpena sare neuronalen bitartez datu-urritasun egoeretan

Amaia Elizaran Mendarte[1], Gustavo Ariel Schwartz[1]
[1]*Materialen Fisika Zentroa (CFM, CSIC-UPV/EHU). M. Lardizabal pasealekua, 5, Donostia*
amaia.elizaran1@gmail.com

Laburpena

Molekulen egitura kimikotik abiatuta haien propietateak iragartzen dituen sare neuronal errepikari (RNN) bat aurkezten da lan honetan. Egitura kimikoak SMILES errepresentazio molekularrarekin kodifikatzen dira sareko sarrera datu bezala erabiltzeko. Oro har, sare neuronal artifizialek emaitza onak ematen dituzte datu askorekin entrenatzen direnean, baina arazoak izaten dituzte datu-urritasun egoeretan. Beraz, lan hau datu-urritasun egoeretan zentratua dago eta horri aurre egiteko hainbat bide aztertzen ditu. Gure hipotesiaren arabera, algoritmoa antzekoak diren datuekin entrenatuz gero, emaitzak hobetu egingo lirateke. Horrela, datuen arteko mota ezberdinetako antzekotasunak aztertu dira; hala nola, SMILES-en arteko antzekotasunei, bai eta ezaugarri bektoreen arteko antzekotasunei erreparatu zaie.

Hitz gakoak: RNN, SMILES, datu-urritasuna.

Abstract

We present a Recurrent Neural Network (RNN) that predicts molecular properties only based on the molecular structure. The SMILES representations of the molecular structures are fed into the algorithm as an input. In general, Artificial Neural Networks work well when they have plenty of input data available, but they perform poorly under data scarcity scenarios. In this work, we specially focus on giving a solution to the problem of data scarcity and we have analyzed different approaches to tackle it. Our hypothesis is that training the model with similar data will improve the results. The analyzed similarities are of distinct nature. On the one hand, we have considered string similarities of the SMILES encodings. On the other hand, we have computed the similarities of the feature vectors.

Keywords: Recurrent Neural Network, SMILES, data scarcity.

1. Sarrera eta motibazioa

Azken urteetan sare neuronalen agerpenak jakintzaren hainbat arlotan eragin handia izan du; itzulpengintzatik hasita kimika konputazionaleraino emankortasuna handia izan da. Kimika konputazionalaren aurrerapenak Deep Learning-arekin (DL) bateratuz, molekulen sintesirako prozedura berriak diseinatzeko bidea ireki du. Izan ere, molekula berrien sintesirako metodologia klasikoan, lortu nahi den molekularen ezaugarriak jakinik, ideien sortze eta ebaluazioan denbora asko igarotzen da; gainera, dirua xahutzen da egindako hipotesiak okerrak diren kasuetan. Aldiz, DL-a baliatuz egitura eta propietateen arteko ezaugarrien azterketa (QSPR, bere ingelesezko hizkiengatik, *Quantitative Structure-Property Relationships*) asko hobetu da azken hamarkadetan (Walters, W.P. eta Barzilay, R., 2021). Horrela, dirua eta denbora xahutzea gutxitu egin da, izan ere, sare neuronalekin probatu baitaiteke hipotesia diren egitura kimikoek behar ditugun ezaugarri fisikoak eskaintzen dituzten.

Lan honetan molekulen egitura kimikotik abiatuz haien propietate fisikoak iragartzen dituen sare neuronal errepikari bat aurkezten da. Molekulen egitura kimikoak SMILES (*Simplified Molecular Input Line Entry System*) errepresentazio molekularrean kodifikatu dira. SMILES kodifikazioan molekulak *string* batean errepresentatzen dira, beraz, egokiak dira RNN-en sarrera bezala. Hain zuzen ere, RNN-en berezitasuna datu sekuentzialak prozesatzeko gaitasuna baita, denbora serieak edo testua esaterako. Horregatik, esan genezake lan honetan NLP (*Natural Language Processing*) hizkuntza prozesamenduko algoritmoetan garatzen diren tekniken analogia erabiltzen dugula molekulen egitura kimikoa aztergai.

Horrez gain, fokua datu-urritasun egoeretan DL-aren bitartez QSPR metodoa hobetzean dago. Jakina da adimen artifizialeko aplikazio guztietan datuen garrantzia ezinbestekoa dela. Alde

batetik, algoritmoek patroiak ikasteko datu asko behar dituzte eta beste alde batetik, datu horien kalitatea bermatuta egon behar da modeloen ikaste prozesua fidagarria izan dadin. Hala ere, DL-eko aplikazio askok datu-urritasuna edo desegokitasunaren arazoari aurre egin behar diote (Alzubadi, L. Et al.). Hori dela eta, datu-urritasuna gainditzeko bi metodo aurkezten dira lan honetan.

Aipatutako metodo horiek, gure datu-baseko molekulen antzekotasunean oinarritzen dira. Gure hipotesia ondorengoa da: modeloak entrenamendu fasean prozesatzen dituen datuak antzekoak badira, entrenamenduko datuen antzekoa den egitura molekular bati (test molekula) dagozkion iragarpen onargarriak egiteko gai izango da. Aitzitik, test molekularen antzekoak ez diren datuekin entrenatuz gero, datu urritasunaren ondorioz, propietateen iragarpena ez da hain zehatza izango.

2. Arloko egoera eta ikerketaren helburuak

Lan honen oinarria bi ikerkuntza arlotan kokatu daiteke. Alde batetik, DL-aren bidezko QSPR, molekulen egituraren eta propietateen arteko azterketa egongo litzateke eta beste alde batetik, sare neuronalen optimizazioa eta datu-urritasun egoeretan eman daitezkeen urratsak egongo lirateke. Literaturan aipatutako bi arloetako argitalpen andana aurki daitezke baina ez dugu biak erlazionatzen dituen artikulu asko aurkitu.

DL-aren bidezko QSPR-i dagokionez, hainbat motatako ikerketak daude. Hasteko, DL-ko aplikazioetan molekulen egituraren kodifikazio mota ezberdinen inguruko artikuluak aurki genitzake eta haien ezberdintasun eta konparaketak ere bai. SMILES errepresentazio molekularraren lehen agerpena esaterako, 1988koa da David Weininger-en eskutik. Honek, algoritmoentzako ulerkorra eta gizakiontzako zentzuzkoa izateko egitura molekularraren kodifikazio bat aurkeztu zuen (Weininger, D., 1988). Ordutik, hainbat izan dira SMILES kodea hobetzeko egin diren ekarpenak, hala nola, DeepSMILES (O'Boyle, N.M. eta Dalke, A., 2018) eta SMILES2vec (Garrett B.G et al., 2018). Bestalde, SMILES-ez gain beste kodifikazio mota batzuk ere topatzen ditugu, hauek ere DL-eko sareetako sarrera bezala erabiltzeko egitura kimikoak. Garrantzitsuena, NLP-ean oinarritzen den Mol2Vec da (Jaeger. S et al., 2018).

Beste alde batetik, datu-urritasunari soluzioak emateko azterketan zentratutako ikerketak ere anitzak dira. Azpimarragarria da hainbat soluzio ezberdin existitzen direla, mota askotako datu-urritasunari aurre egiteko. Esaterako, *Transfer Learning*-a (TL) azken urteetan gailentzen ari den metodoa da, nahiz eta beste batzuk ere existitzen diren: *Self-supervised Learning* (SSL), *Physics-informed Neural Network* (PINN) eta abar (Alzubadi. L. et al., 2023).

Azkenik, DL bidezko QSPR eta datu-urritasun egoera uztartzen dituen ikerketa bat aipatu beharra dago. Chang. R. et al.-ek 2022an argitaratutako lanean datu-urritasuna gainditzeko *Mixture of Experts* (MoE) metodoa proposatzen dute. MoE-en hainbat sare neuronal entrenatzen dira propietate ezberdinak iragartzeko, horietako bakoitzari hasierako datu ezberdinak emanik. Ondoren, aurrez entrenatutako algoritmoak berriro entrenatzen dira desiratutako propietatea aurresateko.

Testu honetan aurkezten den lanean, DL bidezko QPSR-an datu-urritasunaren ondorioak deuseztatzeko bestelako metodo bat proposatzen dugu. Ez dugu literaturan horrelako metodologiarik aipatzen duen adibiderik aurkitu.

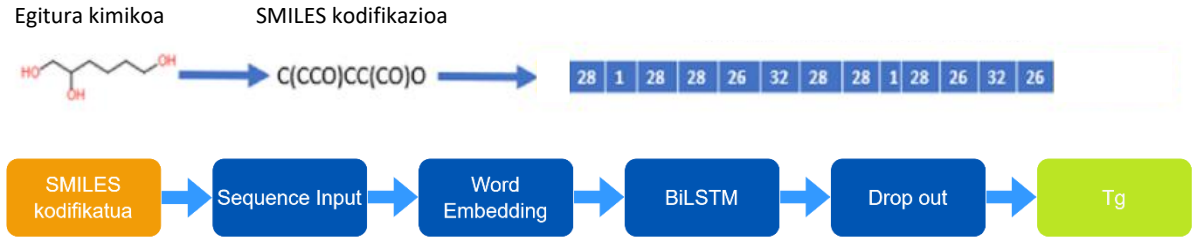
Horrenbestez, ikerketa honen helburu nagusia molekulen egituraren eta haien propietate fisiko-kimikoen arteko iragarpenerako DL modelo bat garatzea da, zeinak datu-urritasun egoeran emaitza onargarriak ematen dituen. Helburu hori lortzeko, bi metodo proposatzen dira, biak datu-baseko molekulen arteko ezaugarrien antzekotasunean oinarrituta. Lehenengo metodoarekin, SMILES errepresentazioak propietateak ondorioztatzeko orduan zenbaterainoko garrantzia duen ikusi nahi da. Bigarren metodoarekin, molekulen egituraren eta propietateen artean egon daitezkeen erlazio ez linealak kalkulatzerakoan (DL modeloarekin) sortzen diren bektore espazioetako bektoreen antzekotasunaren eragina aztertu nahi da.

3. Ikerketaren muina

Ikerketa honetan, RNN bat erabili da molekulen egituratik abiatuz haien beira-trantsizioko tenperatura (*Tg*) iragartzeko. Horretarako, molekulen egitura SMILES errepresentazio bitartez ematen zaio RNN-ari molekula bakoitzaren *Tg* balioekin batera entrenamendu faserako. Normalean, DL aplikazio guztietan datu-basearen multzo bat gordetzen da entrenamendu multzotik kanpo, ondoren, modeloak nola aurresaten duen aztertu (testatu) ahal izateko. Entrenamenduko multzotik kanpo uzten den parte horri test multzoa deritzo.

1. Irudian modeloaren eskema orokorra eta egitura molekularren kodifikazioa aurkezten dira. Egitura kimikoak SMILES bidez errepresentatuta daude datu-basean. Datu-base gehienetan ondo dokumentatua dagoen errepresentazio molekular bat da, beraz, zuzenean lortu ahal izan dira SMILES kodeak. Nolanahi ere, algoritmoak ez ditu karaktere motako sarrerak onartzen, beraz, SMILES kodeko karaktere bakoitzari zenbaki bat esleitu zaio algoritmoak irakur dezan. Irudiaren behealdekoa RNN-aren geruza ezberdinen eskema bat da. Geruza hauetan esanguratsuena BiLSTM da. Horrek RNN-ak datu sekuentzialak prozesatzea ahalbidetzen baitu, "memoria" moduko bat gehituz algoritmoari. Guzti horrekin, kodifikatutako SMILES-etik hasi eta molekula bakoitzari dagokion *Tg* balioaren iragarpena lortzen da.

1. Irudia. Egitura molekularren kodifikazioa eta RNN-aren eskema orokorra. (Borredon, C. et al. (2023))



Behin erabilitako sare neuronalaren arkitektura azaldurik, datu-urritasuna gainditzeko planteatu diren metodoak aurkezten dira. Sarreran aipatutako hipotesian oinarrituta, metodo horietako bakoitzaren helburua elkarren antzekoak diren polimeroen datu-multzo txikiagoak sortzea da. Datu-multzo berriak erreferentziazko molekula batekiko datu-base osoko gainontzeko polimeroek duten antzekotasunean oinarritzen dira. Hau da, lehenik edozein molekula hautatzen da, erreferentziazko molekula. Gero antzekotasun baldintza finkatu eta baldintza hori betetzen duten datu-baseko beste polimeroak hartzen dira, horrela antzekoak diren polimeroen datu-multzoa sortuz. Ondoren, datu-multzo txikiago horiekin sarea entrenatu eta erreferentziazko molekula test datu bezala erabiltzen da. Kasu honetan modeloaren zuzentasuna molekula bakarrarekin testatzen dugu.

Atal honetan ondorengo antolaketa edukiko du artikuluak: 3.1 atalean erabilitako datu-basearen xehetasunak zehaztu dira. 3.2 eta 3.3 ataletan datu-urritasuna gainditzeko proposatutako bi metodoen garapena egin da.

3.1 Datu-basea

DL-eko edozein aplikaziotan funtsezko osagarria dira erabiltzen diren datuak. Horien kopurua handia behar izateaz gain, datuen kalitatea ere funtsezkoa da. Gainera, lan honetan erabiltzen den Deep Learning paradigma ingelesez *supervised-learning* bezala ezagutzen dena da. *Supervised-learning* motako modeloetan datuak (SMILES egitura molekularrak) eta haiekin erlazionaturiko emaitzak (iragarri nahi den propietatea, *Tg*) binaka sartu behar dira; horrela, modeloak entrenamendu fasean molekulen eta dagokien propietatearen balioa erlaziona ditzan.

Hori kontuan izanik, erabilitako datu-basea SMILES egitura molekularrez eta bakoitzarekin erlazionatutako *Tg* balioaz osatuta dago. Guztira 189 datu sarrera ditugu datu-basean eta horiek

guztiak polimeroak dira. *Tg*-ren balioa 197 K eta 457 K artekoa da, datu guztiak 260 K-eko tarte batean daude, honek garrantzia du emaitzak aurkezteko aukeratutako metrika estatistikoarekin.

3.2 SMILES-en antzekotasuna

Lehenengo metodoari "SMILES-en antzekotasuna" deritzo, izan ere, datu-multzoak sortzerako orduan finkatu den antzekotasun baldintza SMILES errepresentazioarekin zuzenean erlazionatuta baitago. Erabilitako antzekotasuna ondo ulertzeko, beharrezkoa da NLP-n erabiltzen den *Edit Distance* izeneko metrika aurkeztea. *Edit Distance* metrikak bi *string*-en arteko desberdintasuna zenbatekoa den kuantifikatzen du, hau da, bi karaktere segida berdinak izateko egin beharreko aldaketak (ezabatu, gehitu edo ordezkatu) zenbatzen ditu.

Metrika horretan oinarrituta ondorengo baldintzak jarri ditugu datu-multzo txikiagoak sortzeko:

1. Erreferentziazko molekula bezala datu-baseko molekula guztiak hartu dira banaka. Ondorioz, 189 datu-multzo sortu dira.
2. Datu-multzo bakoitzean izan beharreko polimero kopurua gutxienez 60 da.
3. *Edit Distance* bakoitza desberdina izango da. 60 molekula antzeko izateko erreferentziazko molekula bakoitzak behar duen *Edit Distance* txikiena hartu da 1etik 25erako tarte batean.

Hiru baldintza hauek kontuan izanik sortutako datu-multzo bakoitzarekin modeloa entrenatu da eta ondoren bakoitzari dagokion erreferentziazko test molekularekin testatu da.

Modeloaren emaitzak hurrengo moduan aurkeztuko dira: datu-baseko polimero bakoitzerako *Tg* balioa zein den jakina denez, konparaketa bat egiten da benetako balioa (*ground truth, y)* eta iragarritako balioaren (y_p) artean. Emaitzak MAPE (*Mean Absolute Percentaje Error*) metrika estatistikoaren bitartez aurkezten dira, (1) adierazpena.

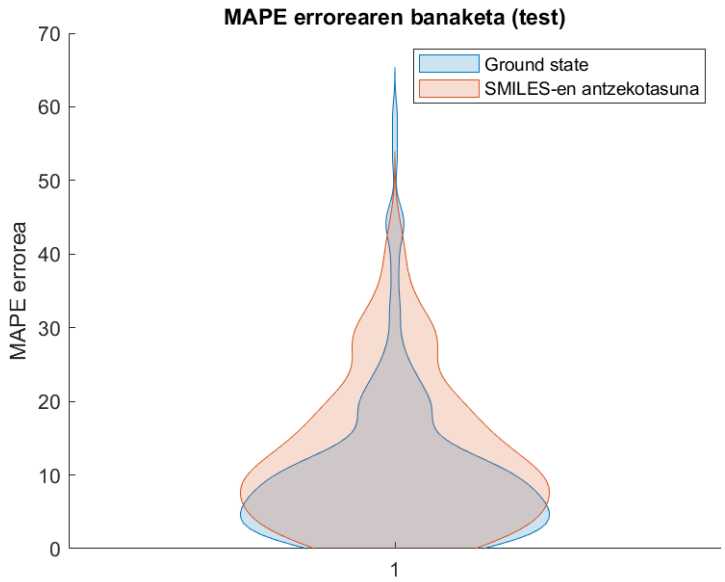
$$(1)\ \text{MAPE} = 100\ \frac{1}{}\sum_{=1} \left| \frac{}{} \right|$$

3.1 atalean aipatu bezala, metrika hau erabiltzea egokia da balioak 260 K-eko tarte batean daudelako.

Bestalde, aipatu beharra dago "SMILES-en antzekotasuna" metodotik lortutako emaitzak *Ground State* egoerako emaitzekin konparatu direla. *Ground State* egoeran ere 189 datu-multzoren entrenamendua egin da, erreferentziazko molekula bat kanpoan utziz, baina multzo hauek datu baseko beste 188 polimeroekin osatuta egon dira kasu guztietan. Beraz, erreferentziazko polimero bakoitzerako beste guztiekin entrenatu da modeloa eta ondoren testatu.

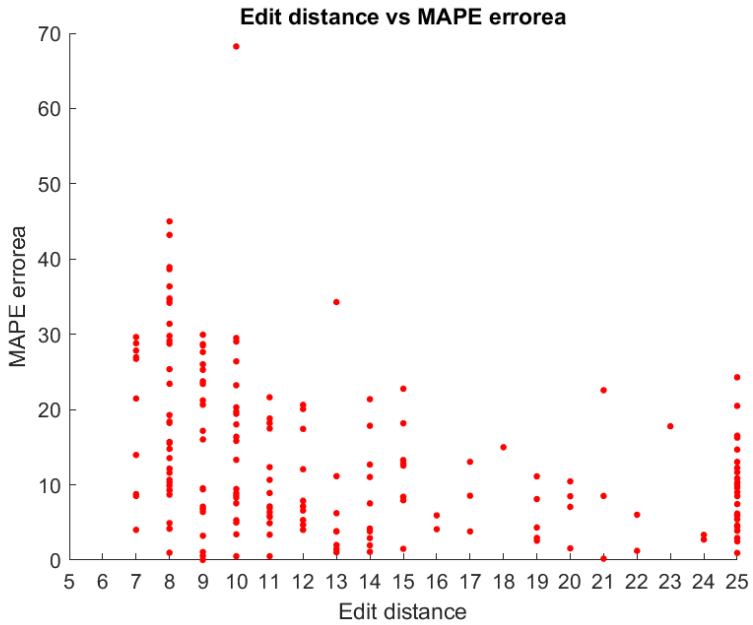
2. Irudian bi egoeretatik lortutako MAPE errorearen distribuzioa ikusten da. "SMILES-en antzekotasuna" metodoarekin lortzen dugun errorearen banaketaren mediana *Ground State*-rekin lortzen duguna baino handiagoa da.

2. Irudia. SMILES-en antzekotasuna eta *Ground State* egoeren emaitzen errorearen banaketa



Emaitza hau ikusirik, erreferentziazko molekula bakoitzari esleitutako *Edit Distance*-a eta bakoitzarekin lortutako errorea erlazionatzen dituen grafikoa aurkezten da, 3. Irudia. Grafiko hau irudikatzearen helburua banaketa baxuagorik ez lortzearen arrazoia aztertzea da.

3. Irudia. Molekula bakoitzarekin lortutako MAPE errorea multzo bakoitzari egokitutako *Edit Distance*-aren menpe



3. Irudian ikusten da errorearen minimoa *Edit Distance* metrika [15, 24] artean dagoenean dela. Bestalde, ikus daiteke tarte horretan aurkitzen diren puntuak urriagoak direla *Edit Distance* txikiagoa denean baino.

3.4 RNN barneko espazio bektorialeko bektoreen antzekotasuna

Hurrengo metodoa ulertzeko sare neuronalen arkitekturak haien barneko geruzetan burutzen dituen eragiketen zantzu batzuk aipatu beharra dago. Edozein motatako sare neuronalen abantaila

sarrerako datuen eta irteerakoen arteko erlazio ez linealak detektatzea eta haiek iragarpen egokietarako erabiltzea da. Erlazio ez lineal hauek, sare neuronalen geruza ezberdinetako emaitzei ez linealtasuna bermatzen duten aktibazio funtzioak aplikatzearen ondorioz atzematen dira. Geruza ezberdinetako balioak, pisuak deritzenak, minimizazio algoritmo baten bitartez kalkulatzen dira sarearen entrenamendu fasean. Izan ere, geruza bakoitzeko neuronetara sarrerako datuen informazioa heltzen da eta amaieran lortzen den iragarpena, *ground truth* (benetako balioa) balioarekin konparatzen du minimizazio algoritmo horrek. Horrela, entrenamendu faseko hurrengo etapan pisuak birmoldatuz, benetako balioarekin egingo duen hurrengo konparazioan desberdintasuna txikiagoa izan dadin.

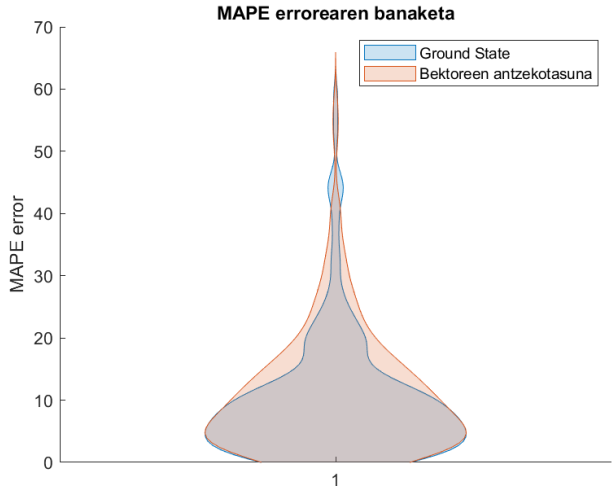
Honetan guztian algoritmoaren hiperparametroek garrantzia handia dute. Hiperparametroak aldatzerakoan algoritmoaren barneko pisuen eguneratzea era ezberdinean egiten da. Horregatik kontuan eduki behar da entrenamendu bakoitzean lortutako pisuak hiperparametroen menpekoak direla.

Erabilitako RNN-aren kasuan, BiLSTM geruzako pisuei aplikatzen zaien aktibazio funtzioa tangente hiperbolikoa da, [-1,1] tarteko emaitzak ematen dituena. Espazio bektorialeko bektoreen antzekotasuna metodoan, BiLSTM geruzako aktibazio ondorengo [-1,1] tarteko balioak dituzten bektoreak erabiltzen ditugu. Bektore hauen dimentsioa BiLSTM geruzan dagoen neurona kopurua da eta entrenamendu fasea amaitzerakoan geratzen diren balioak hartu dira, hau da, optimoenak.

Laburbilduz, entrenamendua burutzerakoan BiLSTM geruzako pisuei aktibazio funtzioa aplikatu ondoren geratutako pisuen bektorea da abiapuntua. Erreferentziazko molekula bakoitzarekin entrenatzerakoan balio ezberdineko bektoreak lortzen dira eta haien arteko antzekotasuna da aztertzen dena. Bektore horiek sarrerako egitura kimikoaren eta benetako *Tg* balioaren arteko ez linealen informazioa gordetzen dute.

Beraz, erreferentziazko molekula bakoitzarentzat lortutako bektorea, beste guztiekin lortutako bektoreekin konparatu da eta distantzia euklidear minimoa duten bektoreekin sortzen da datu-multzo berria. Kasu honetan datu-multzo berri bakoitzerako gutxieneko polimero kopurua 40 izan da. Gainera, lehenengo aldaeran bezala, 189 datu-multzo berri sortu ditugu, bana erreferentziazko molekula bakoitzarentzako. Emaitzak “SMILES-en antzekotasuna” atalean bezala aztertu dira oraingoan ere. Lortutako datu-multzo bakoitzarekin modeloa berriro entrenatu da erreferentziazko molekula test bezala erabiliz kasu bakoitzean.

4. Irudian ikus daiteke bektoreen antzekotasuna aldaerarekin lortutako MAPE errorearen banaketa *Ground State*-arenaren antzekoa dela nahiz eta mediana handiagoa duen.



4. Irudia. *Ground State* eta bektoreen antzekotasuna aldaeren MAPE errorearen banaketa

4. Ondorioak

Lan honetan DL bidezko QSPR aplikazioetako datu-urritasunaren ondorio kaltegarriak ekiditeko metodo berri bat aurkeztu da. Hain zuzen ere, test molekularen antzekoak diren molekulekin sortutako datu-multzoekin modeloaren entrenamendua egitea proposatzen da, honek izan ditzakeen bi aldaera posible aztertzearekin batera. Ondoren, aldaeretako bakoitzari dagozkion ondorioak aurkezten dira.

Lehenengo metodoaren emaitzei dagokienez, [15, 24] tartean dauden *Edit Distance*-a duten datu-multzoetatik lortutako emaitza hobea da, algoritmoak kasu hauetan ikas ditzakeen ezaugarriak gehiago direlako. Esaterako, 60 polimero antzeko izateko *Edit Distance* 9 baldin bada, orduan multzoan dauden molekula guztiak elkarren antzekoagoak dira eta ondorioz, modeloak ikasteko ezaugarri ezberdin gutxiago jasotzen ditu. Beraz, aurrez entrenatu ez duen polimero bat modeloan sartzerakoan, haren iragarpen gaitasuna txikiagoa da.

Horregatik, "SMILES-en antzekotasuna" metodoak ez ditu esperotako emaitzak ematen. Azken finean, emaitza onargarriak lortzeko 15 eta 24 arteko aldaketa kopurua behar dituzten molekula-multzoak beharrezkoak direlako. Aldaketa kopurua hain handia izanik, gutxi gora-behera SMILES-en luzeraren heren bat (luzeenak 66 karaktere dituelako), emaitza onargarriak ematen dituzten SMILES-en multzoetan dauden polimeroen SMILES-ak ez dira hain antzekoak.

Hurrena, "bektoreen arteko antzekotasuna" aldaerarekin lortutako emaitzetatik ateratako ondorioak aurkezten dira. Kasu honetan, argi ikusten da emaitzak "SMILES-en antzekotasuna" aldaeran lortutako emaitzak baino hobeak direla, 3. Irudian bi banaketak ia osotasunean gainezarrita daudela ikusten da. Gainera, aipagarria da datu-multzo bakoitzean datu-base osoaren % 21,16 dugula (40 SMILES-*Tg* bikote) eta beraz datu-urritasun egoera are handiagoa dela. Honekin, bektoreen arteko antzekotasuna datu-urritasun egoerei aurre egiteko aldaera ona dela frogatzen da.

Hala ere, kontuan izan behar da bektore hauek espazio bektorial konkretu baten parte direla, hain zuzen ere, algoritmoak unean uneko hiperparametroen menpeko espazio bektorialak ahalbidetzen ditu eta beraz, baliteke espazio bektorial hau gure aplikaziorako egokiena ez izatea.

Amaitzeko, ikerketa honetan aurkeztutako metodoa eta emaitzak teknikoak diren arren, aipatu beharra dago lan honen ildoko ikerketek aplikazio zuzen bat dutela eta abantaila handiak ekar ditzakeela QSPR klasikoarekin alderatuz. Labur esanda, dirua eta baliabideak xahutu gabe molekulen sintesia egitea ahalbidetzea da helburu nagusia. Gainera, material berrien garapenak ingurugiroan duen inpaktu negatiboa gutxituko litzateke molekulen egitura kimikotik soilik abiatuz haien propietateak iragartzea ahalbidetzen duen teknika honekin.

5. Etorkizunerako planteatzen den norabidea

Lan honetan laburbildu diren ondorioei erreparatuz hainbat hobekuntza planteatzen dira etorkizunean norabide berean jarraitu ahal izateko.

Hasteko, "bektoreen arteko antzekotasuna"-ri dagokionean, lehenik bektore horien espazio bektoriala optimizatzea beharrezkoa da. Optimizazioan zentratuz, ildo beretik joanda askoz emaitza hobeak lortzea posible litzateke.

Bigarrenez, "SMILES-en arteko antzekotasuna"-ren bidetik aurrera ez jarraitzeko erabakia hartu da, emaitzek antzekotasuna adierazten duen aldaera ona ez dela erakusten dutelako.

Horrez gain, datu-urritasuna gainditzeko beste proba batzuk proposatu berri ditugu. Esaterako, Transfer Learning-a aplikatzea MSTL (*Multi Source Transfer Learning*) modelo egitura bat diseinatuz.

6. Erreferentziak

Alzubaidi, L., Bai, J., Al-Sabaawi, A. et al. (2023). A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. *J Big Data* **10**, 4. https://doi.org/10.1186/s40537-023-00727-2

Borredon, C. et al. (2023). On the dynamics of polymers and biomolecules thourgh the use of machine learning algorithms. PhD Thesis. Centro de Física de Materiales (CSIC-UPV/EHU).

Chang, R., Wang, YX. & Ertekin, E. (2022). Towards overcoming data scarcity in materials science: unifying models and datasets with a mixture of experts framework. *npj Comput Mater* **8**, 242. https://doi.org/10.1038/s41524-022-00929-x

Garrett, B. G. et al. (2018) SMILES2vec: An Interpretable General-Purpose Deep neural Network for Predicting Chemical Properties. arXiv: 1712.02034. https://doi.org/10.48550/arXiv.1712.02034

Jaeger, S. et al. (2018). Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. Journal of Chemical Information and Modelling. 58, 27-35. 10.1021/acs.jcim.7b00616

O'Boyle N, Dalke A. (2018). DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures. ChemRxiv. 10.26434/chemrxiv.7097960.v1

Walters, W.P., Barzilay, R. (2021). Applications of Deep Learning in Molecule Generation and Molecular Property Prediction. Accounts of Chemical Research, 54(2), 263-270. 10.1021/acs.accounts.0c00699

Weininger, D. (1988). SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. Journal of Chemical Information and Computer Sciences 28, 31-36. 10.1021/ci00057a005

7. Eskerrak eta oharrak

Eskerrak eman nahi dizkiet nire tesiko zuzendariei Gustavo Ariel Schwartz eta Silvina Cerveny, doktoretza Materialen Fisika Zentroan egiteko aukera emateagatik. Horrez gain, nahiz eta eurek euskaraz ez dakiten Ikergazte2025 kongresuan nire parte hartzea bultzatzeagatik. Bestalde, CSIC-i eskerrak ematen dizkiot PID2023-146348NB-I00 proiektua PREP2023-001205 kontratuarekin diruz laguntzeagatik.

Azkenik, aipatu beharra dago lan hau Claudia Borredonen (2023) doktoretza tesitik eratorria dela. Lan honetan erabilitako RNN algoritmoa eta datu-basea doktoretza tesitik hartu ditugu.