Universidad del País Vasco Euskal Herriko Unibertsitatea

ZTF-FCT
Zientzia eta Teknologia Fakultatea
Facultad de Ciencia y Tecnología

GRADU AMAIERAKO LANA

KIMIKAKO GRADUA

PROTEINEN SEKUENTZIAREN MENPEKO EREDU KIMIOINFORMATIKOAK MINIBIZIAREN KONTRA

XABIER JIMÉNEZ DE ABERÁSTURI GONZÁLEZEK AURKEZTUTAKO MEMORIA

MATRIKULA ETA DEFENTSA DATA: 2019ko ekainaren 21a

ZUZENDARIAK: Sonia Arrasate Gil eta Humberto González-Díaz

SAILA: Kimika Organikoa II Saila

AURKIBIDEA

1. SARRERA

Azken hamarkadetan gizakiak aurrera eramandako teknologia eta medikuntzaren garapena, herrialde garatuetako osasunak gorakada itzela jasatea eragin du. Dena den, oraindik badira gaixotasun batzuk zeinen kontra ez den irtenbide guztiz eraginkorrik topatu. Horietako bat minbizia izan daiteke, XXI mendeko gaixotasuna izan litekeena. Haren kontra ikerkuntza talde askok konposatu berriak sintetizatzen dituzte, baina hauek eraginkorrak direnetz ziurtatzeko zenbait froga egin behar dira. Saiakuntza horiek aurrera eramateko laguntza handikoa izan daiteke espero den emaitza aurresatea denbora eta baliabideak aurrezteko asmotan. Helburu honekin erabili daitezke esate baterako eredu kimioinformatikoak, erreminta boteretsuak baitira kimikan.

1.1. MINBIZIAREN ZERGATIAK

Esandako moduan, minbiziari XXI mendeko gaixotasuna dei diezaiokegu, funtsean, "Onkologia Medikoko Elkarte Espainiarrak" (SEOM) dio minbizi kasuen kopurua gero eta altuagoa dela mundu mailan.[1] 2012. urtean 14 milioi tumore-kasu berri topatu ziren munduan, 2018an 18 milioira igo zen jada, eta estimazioak bete ezkero, 30 milioi kasu berri ingurura hel gintezke 2040. urtean. Elkarte honen arabera ere minbizia garatzeko probabilitateak altuagoak dira gizonezkoetan emakumezkoetan baino eta esponentzialki igo egiten dira adinak aurrera egin ahala.

Ezaguna denez, kantzer motak asko eta asko dago eta haien artean arruntena kolon eta ondesteko minbizia da, behintzat Espainia mailan. Aipatzekoak ere badira biriketakoa eta gernu-maskuriako minbiziak. Gizonezkoen kasuan aparteko garrantzia dauka prostata-kantzerrak eta emakumezkoen kasuan aldiz, bularrekoak. Beste mota asko ere badaude milaka hildako eragiten dituztenak urtero, hala nola,

pankreasekoa, aho eta faringekoa, giltzurrunetakoa edota hain ezaguna den leuzemia. "Espainiako Estatistika Institutu Nazionalaren" (INE) esanetan, 2016. urtean ia 113.000 hildako eragin zituen minbiziak, eta ez litzateke harritzekoa izango zifra horrek gora egitea hurrengo urteetan.[2]

Puntu honetan argi dago kantzerra badela gaur egun benetako osasun-arazo larria, baina zeintzuk dira datu ikaragarri hauek eragiten dituzten kausak? Orokorrean, nabarmena da batez ere herrialde garatuetako biztanleriaren batez besteko adinak gorakada handia jaso duela, eta lehen azaldutako legez, askoz errazagoa da minbiziaren agerpena pertsona helduetan gazteetan baino. Hau alde batera utzita zergatiei buruz hitz egin beharrean, zuzenagoa da arrisku-faktoreei buruz hitz egitea, izan ere, oso zaila da minbiziaren jatorria zehatz-mehatz zein den jakitea kasu bakoitzean. Hartara, jarraian "Estatu Batuetako Minbiziaren Institutu Nazionalak", plazaratutako arrisku-faktore nagusiak ditugu.[3]

Lehen arrisku faktorea dieta da, gizentasuna hobeto esanda. Erlazio estua dago gizentasunaren eta minbizia sufritzeko probabilitateen artean, eta hau dela eta, oso garrantzitsua da dieta kontrolatzea eta aktibitate fisiko egokia aurrera eramatea. Dietarekin erlazionatuta, berebiziko garrantzia dauka alkoholak. Funtsean, zentzuzkoa da pentsatzea ahoko, faringeko edota gibeleko minbizia jasateko arrisku faktorea dela alkoholaren neurrigabekeria, eta hartara, komenigarria dela neurriz hartzea. Hauetaz gain, aparteko aipua merezi du tabakoak. Oraingoan bai, kausa bati buruz hitz egin dezakegu, eta ez soilik arrisku-faktore bati buruz. Izan ere, tabakoa tumore kasu gehienen jatorri nagusia da, eta ez soilik biriketako minbizian, baita leuzemian, aho, eztarri, maskuria eta gibeleko minbizietan ere.

1.2. MINBIZIAREN ETAPAK ETA TRATAMENDUA

Hasteko, azaldu beharra dago minbizia akats genetikoen ondoriozko gaixotasuna dela, DNA bikoizterakoan emandako akatsen ondorioa hain zuzen ere. Oro har, akatsak berez ematen dira DNA kateen erreplikan, baina mekanismoa hain da zehatza, non akats horiek oso kantitate txikitan ematen diren; 10^{-5} eta 10^{-9} bitarteko akats-tasarekin.[4]

Mutazio hauek behin eta berriro gertatzen dira zelulen zatiketan inolako ondorio larririk gabe. Baina batzuetan, badituzte ondorio garrantzitsuak, esaterako, protoonkogeneetan ematen direnean. Protoonkogeneak zelulen hazkundea eta ugalketa kontrolatzen dituzten geneak dira. Hauetan akatsen bat emanez gero, posible da haien zeregina egiteari uztea. Momentu horretan protoonkogene izatetik onkogene izatera pasatu egiten dira, eta zelulen bizi-zikloaren deskontrola eragiten dute tumoreei hasiera emanez.[5]

Behin mutazio horiek emanda, zelula kaltetua behin eta berriz ugaltzen hasten da, gene mutatua bere barne duten klon berriak sortuz. Zelula guzti horiek haien morfologia edota funtzio zelularra aldatzen dute, eta orduan displasia eman dela esan dezakegu. Displasiak neoplasia edo hiperplasia eragin ditzake, hau da, etengabe eta erritmo altuegian ugaltzen diren zelulen multzo handia. Multzo horri jada tumore dei dakioke. Tumoreak onberak izan daitezke baina baita gaiztoak ere. Azken hauek ez diote handitzeari uzten eta beste leku batzuetara hedatzeko joera dute metastasiari hasiera emanez, minbizia inbaditzaileari, alegia.[6]

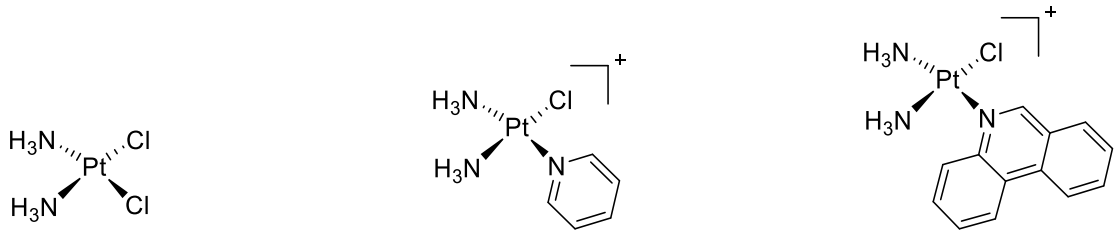
Prozesu oso honetan berebiziko garrantzia du angiogenesiak. Angiogenesia odol-hodi berrien eraketa da, aurretiazko odol-hodietatik abiatuz. Displasiarako beharrezkoa da odol-hodi berriak sortzea eratu diren zelula gaiztoak elikatzeko, eta berdina gertatzen da neoplasiarekin, tumoreak nutrienteak eta oxigenoa behar baititu

bizirik jarraitzeko. Aurreko bietan garrantzitsua bada, metastasian oinarrizkoa da angiogenesia, tumore gaiztoek linfa edo odol-hodiak erabiltzen baitituzte beste leku batzuetara zabaltzeko.[7]

Minbizia tratatzeko eta tumorearekin bukatzeko asmotan tratamendu desberdinak daude. Haietako batzuk konposatu kimiko edo biologikoak erabiltzen dituzte eta beste batzuk, aldiz beste printzipio batzuetan oinarritzen dira, hala nola erradioterapia, terapia genikoa tumorea kentzeko ebakuntza eta hezur-muinaren transplantea.

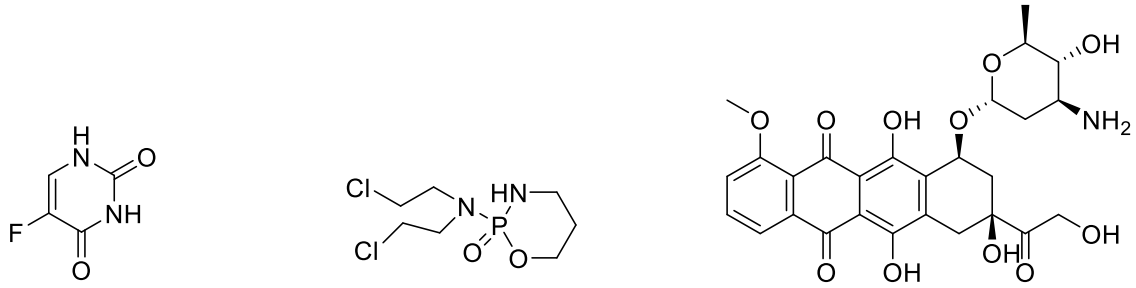
Farmakoak edo bestelako konposatuak erabiltzen dituzten teknika nagusiak hiru dira: kimioterapia, inmunoterapia edo terapia antiangiogenikoa. Lehenengo kasuan, zenbait konposatu kimiko erabiltzen dira zelula tumoralen zatiketa ekiditen dutenak. Haren arazoa da askotan zelula osasuntsuak ere erasotuak izaten direla. Soilik zelula gaiztoak suntsitzeko asmoz inmunoterapia daukagu; honako honetan medikamenduak erabiltzen dira immunitate-sistema kitzikatzeko eta sistemak berak zelula gaixotuen apoptosia eragin dezan. Azkenik, konposatu antiangiogenikoek angiogenesi tumorala eragozten dute; hau da, tumorea bizirik mantentzeko beharrezkoak diren odol-hodi berrien eraketa ekiditen dute.

Gaur egun, konposatu asko eta asko dago kantzerraren kontra jotzeko, kimikoki izaera oso desberdina dutenak. Alde batetik koordinazio konposatu metalikoak eta konposatu organometalikoak ditugu, eta arlo honetan *cis*-platinoa da errege.[8] Konposatu honek *cis*-diaminodikloroplatino(II) du izena, eta geometria karratu laua dauka. Haren deribatuak ere erabiliak edo behintzat entseatuak izan dira, hala nola, pyriplatinoa eta fenantriplatinoa, 1. Irudian ikus ditzakegunak. Hauetaz gain, gaur egun metalozenoak edo urre(I) eta urre(III)-ren konposatuen aktibitate antikantzerigenoa ere aztertzen ari da.[9,10]

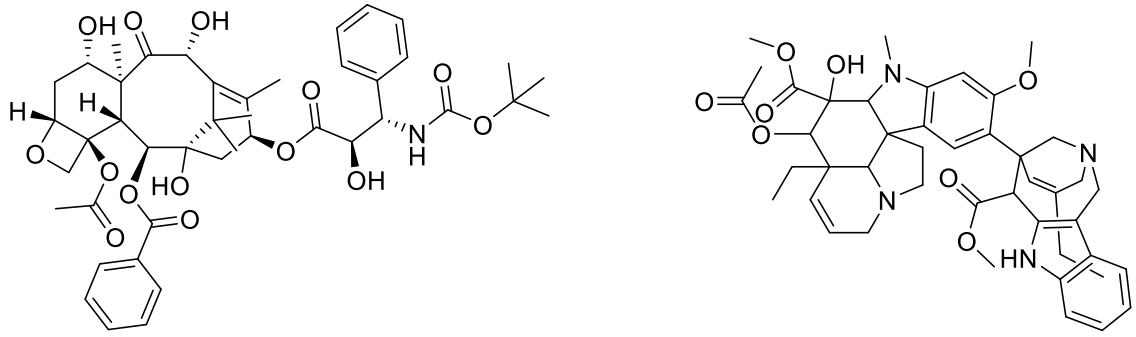


1. Irudia- *Cisplatinoaren, Pyriplatinoaren eta Fenantriplatinoaren egiturak.*

Beste alde batetik, metalik gabeko konposatu organikoak ere oso erabiliak dira minbiziaren irtenbidetzat eta askotan bat baino gehiago batera erabiliak izan dira. Esate baterako, 5-fluorouraziloa, ziklofosfamida, epirrubizina, dozetaxela edo binorelbina launaka erabiliak izan dira saiakuntza desberdinetan, eta haien egiturak 2. Irudian ikus daitezke.[11]



5-Fluorouraziloa Ziklofosfamida Epirrubizina



Dozetaxela Binorelbina

2. Irudia- *5-fluorourazilo, ziklofosfamida, epirrubizina, dozetaxel eta binorelbina konposatuen egiturak.*

1.3. EAEK EREDUEN OINARRIAK

Ikusi dugun moduan, konposatu asko eta asko izan daitezke aktiboak minbiziaren kontra, eta hasieran esandako legez, aktibitate hori aurreikusteko pisuzko modu bat eredu kimioinformatikoak dira. Hauen artean QSAR ("Quantitative Structure-Activity Relationship") ereduak edo EAEK (Egitura-Aktibitate Erlazio Kuantitatiboa) ereduak nabarmentzen dira. Eredu hauek funtsezko printzipio batean oinarritzen dira: konposatu baten egituraren eta haren propietateen arteko erlazioan. Gure kasuan iragarpenaren helburu den propietatea konposatuen aktibitate biologikoa da, eta hortik dator haien izena.[12]

Erlazio hori lehen aldiz definitu zuten 1968. urtean Crum-Brown eta Fraser ikerlariek. Haien esanetan konposatu batek sistema biologiko zehatz batean duen eragina (Φ) haren egitura kimikoaren (C) funtzio bat da. Modu honetan, (1) Ekuazioa lor daiteke.[13]

$$\Phi = f(C) \qquad (1)$$

Dena den, oraindik definitzeke geratzen diren zenbait arlo daude. Hasteko, konposatu kimikoaren egitura matematikoki definitu behar da; izan ere, ekuazioa betetzeko, egitura kuantifikagarriak diren aldagaiekin ordezkatu behar da. Hau egiteko helburuarekin, deskriptore molekularrak erabil daitezke.

1.4. DESKRIPTORE MOLEKULARRAK

Deskriptore molekularrak konposatu baten estrukturari buruzko informazioa balio numerikoen bidez ematen duten parametroak dira. Asko dira kalkula daitezkeen eta molekula baten egiturari buruzko informazioa ematen duten deskriptoreak; adibidez, masa molarra, molekularen bolumena, azalera edo momentu dipolarra. Hau dela eta, haiek sailkatzeko beharra dago, eta hiru irizpide jarrai daitezke horretarako. Alde batetik deskriptoreen balioak esperimentalki edo teorikoki kalkula daitezke, eta hortxe daukagu lehenengo irizpidea.

Esate baterako, partizio koefizientea bi moduetan kalkula daiteke: laborategian konposatu bat ^{n}oktanol eta uretan nola banatzen den neurtu daiteke haren kontzentrazioa bi disolbatzaileetan determinatuz; eta konputazionalki ere, badaude propietate hori kalkulatzen duten "software"-ak.[14] Beste alde batetik, deskriptoreak lokalak edo molekula osoari erreferentzia egiten diotenak izan daitezke. Batzuetan molekularen zatirik adierazgarrienak soilik hartuz, eta deskriptore lokalak erabiliz, nahikoa da konposatuaren propietateak aurresateko.[15]

Azkenik, deskriptoreak sailkatzeko azken modu bat ere badago haien dimentsioetan oinarrituta. 1D deskriptoreak atomo eta loturak zenbatzean lortzen direnak dira (formula enpirikoa edo molekularra, asegabetasun kopurua, masa molarra...). 2Dkoak aldiz, molekula irudikatzearen ondorioz lor daitezkeen parametroak dira (tamaina, efektu esterikoak, adarkapen-maila...). Azkenik, 3D deskriptoreak ditugu, eta hauek konputazionalki kalkulaturikoak dira (LUMO eta HOMO orbitalen energia, molekularen azalera, ionizazio potentziala eta abar).[16]

Orainaldian, oso erabiliak diren eta egituren inguruko informazio nahiko adierazgarria ematen duten bi deskriptore molekular, ^{n}oktanol/ur partizio koefizientea (logP) eta "Polar Surface Area" (PSA) dira (gainazal polarraren azalera, GPA). logP parametroak konposatu baten lipofilitatearen inguruko informazioa ematen digu, lehenago aipatutako moduan, hura kalkulatzeko metodo desberdinak daude eta haietako zein erabili den jakiteko, aurrizki bat jarri ohi da hasieran. Hala AlogP, KlogP edo PROlogP aurki ditzakegu, zeinak metodo atomikoen bidez kalkulatuak izan diren. Honek esan nahi du, konposatuaren atomo esanguratsuenak eta haien loturak eta ingurune kimikoa kontuan hartzen direla kalkulu konputazionaletan. Konputazionalki ere fragmentazio-metodoak erabilita kalkula daiteke,

molekularen zati esanguratsuenen efektu esteriko eta elektronikoak kontuan hartuz ClogP lortzeko.[17] GPA-ri dagokionez, konposatu baten gainazalaren zer azalera okupatzen duten atomo polarrek adierazten digu. Honetarako, molekula batean dauden atal polar desberdinek egiten duten kontribuzioa gehitu egiten da funtzio taldeetan oinarrituz kontribuzio hori neurtzeko.[18]

Deskriptore molekularrekin arazoak izan ditzakegu datu asko ditugunean esku-artean. Milioika deskriptore baldin baditugu eta haien arteko balioak oso desberdinak badira informazio oso handia daukagu datu horietan. Informazio guzti hori maneiatzeko, "Informazioaren Teoria" erabil daiteke. Shannon-ek garatutako teoria honen arabera, datu asko ditugunean informazioa magnitude fisiko bat balitz moduan trata daiteke, datuen desordenean oinarritzen den magnitudea hain zuzen ere. Hau da, Shannon-ek defendatzen du datuen entropia (Shannonen entropia, Sh) dela informazioa ahalik eta hoberen transmititzen duen magnitudea, informazio kantitatea neurtzen duen magnitudea. Shannonen entropia definitzeko (2) Ekuazioa erabili egiten da.[19]

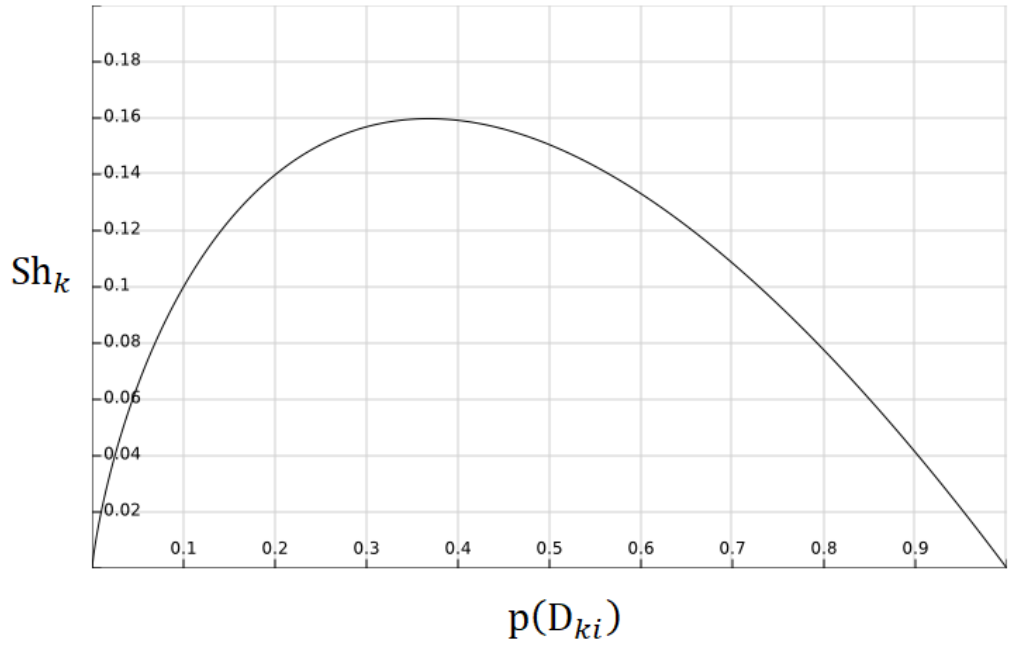
$$\mathrm{Sh_{ki}} = -\mathrm{p(D_{ki})} \cdot \log \mathrm{p(D_{ki})} \quad (2)$$

(2) Ekuazioan $p(D_{ki})$ k deskriptore batek i konposatu batean balio bat izateko probabilitatea da, eta hau kalkulatzeko modu asko egon arren, sinpleenetako bat minimo-maximo eskalatze bat erabiltzea da, (3) Ekuazioan adierazten denaren modukoa.[20]

$$\mathrm{p(D_{ki})} = \frac{\mathrm{D_{ki}} - \mathrm{D_{ki,min}}}{\mathrm{D_{ki,max}} - \mathrm{D_{ki,min}}} \qquad (3)$$

D_{ki} i konposatuaren k deskriptore molekularra da, "$D_{ki,min}$" aurkitu den deskriptore horren baliorik txikiena da eta "$D_{ki,max}$", aldiz, baliorik altuena. (2) Ekuazioari esker gainera, entropiaren balioak tarte zehatz batera mugatzen ditugu, ezin baitute edozein balio hartu probabilitateak 0 eta 1-en artean mugaturik daudelako. Funtsean, (2)

Ekuazioaren funtzioa irudikatuz gero 3. Irudia lortzen dugu eta ikusten dugu entropiaren balio guztiak 0 eta 0.16-ren artean egon behar dutela, baldin eta informazio guztia iturri bakar batetik baldin badator.



3. Irudia- *(2) Ekuazioko Shannonen entropiaren irudikapen grafikoa.*

Molekula baten egituratik zuzenean ere Shannonen entropiak lor daitezke programa batzuekin. Esate baterako, S2SNet programak ADN kateen edo proteina desberdinen Shannonen entropiak kalkula ditzake haien sekuentziatik abiatuta. Horretarako Grafo Teorian eta Markov-en kateetan oinarritzen da. Proteinen kasua orokorrean azalduta, aminoazido bakoitza sekuentzian haren ostean datorren aminoazidoarekin lotzeaz gain, bere mota berdineko hurrengo aminoazidoarekin ere lotzen du programak. Honi esker konexio matrize bat eratzen du (Markov-en matrizea). Hala, aminoazido bakoitza zenbat eta zer aminoazidori dagoen lotuta neurtzen du, eta bere mota bereko hurrengo aminoazidoraino dagoen distantzia ere neurtzen du, hots, erdian zenbat aminoazido dauden. Sare berri honetatik indize topologiko desberdinak kalkulatzen ditu, eta horietatik Shannonen entropia desberdinak kalkulatzen ditu beste hainbat

deskriptorekin batera. Kasu honetan aminoazido bakoitza informazio iturri desberdin bat denez, Shannonen balioa haien batura da eta ez dago 0.16 balio maximora mugatua.[21]

1.5. PTIA EREDUAK

Jadanik ezaguna dugu deskriptore molekularren funtzioa zein den EAEK ereduetan, baina oraindik deskriptore horien eta interesatzen zaigun aktibitate biologikoaren (v_i) arteko erlazioa zein den definitu behar dugu. Kasurik sinpleenean, baliteke aktibitatea deskriptore bakar baten funtzioa izatea, eta gainera haien arteko korrelazioa lineala izatea. Adibidez, aktibitatea konposatuaren masa molarrarekiko (D_1 deskriptorearekiko) linealki proportzionala balitz hurrengo lehenengo mailako (4) Ekuazioa lortuko genuke.[13]

$$v_i = a + b \cdot D_1 \qquad (4)$$

Hala ere, eta logikoa denez, konposatu baten aktibitatea ez dago soilik masa molarraren menpe, izan ere, masa molarra hain deskriptore zabala da, non ez den egitura zehazteko oso adierazgarria. Beraz, ekuazio konplexuagoak garatzeko beharrean gaude. Oraingoan demagun masa molarraz gain partizio koefizientea (logP = D_2) ere sartzen dugula ekuazioan, eta gainera bi aldagai hauen eta aktibitatearen arteko erlazioa ez dela lineala baizik eta karratua. Kasu honetan, (4) Ekuazioa baino konplexuagoa den (5) Ekuazioa lortuko genuke.[22]

$$v_i = a + b_1 \cdot D_1 + c_1 \cdot D_2 + b_2 \cdot D_1^2 + c_2 \cdot D_2^2 \quad (5)$$

Ikusi dugunez, modu asko dago aktibitatearen eta deskriptoreen arteko erlazioak finkatzeko, eta ikusitako ekuazioei orokorrean "Machine Learning" (ML) edo Ikasketa Automatikoko (IA) ereduak deritze. Arazoa da, gaur egun minbiziaren kontrako konposatu asko eta askoren entseguak eta sintesiak egin direla, eta "Big Data"-ren munduan barneratu garela, datu masiboen munduan, alegia.

Orokorrean, IA eredu arruntek ez dituzte ondo aurresaten konposatuen aktibitateak hainbeste datu erabiltzen direnean oinarritzat, eta beraz, horri aurre egiteko irtenbideak bilatu behar dira.[13]

Irtenbide horietako bat "Perturbation Theory and Machine Learning" (PTML) ereduak dira, Perturbazio Teoria eta Ikasketa Automatikoa (PTIA) delakoa. Hauetan ez dira soilik konposatuen deskriptore molekularrak kontuan hartzen; konposatu bakoitzaren deskriptore bakoitzaren eta baldintza berdinetan entseatutako konposatu guztien deskriptore berdinaren batezbestekoaren arteko diferentzia ere kontuan hartzen da. Diferentzia horri "Moving Average" edo Batezbesteko Higikorra deritzo (BH) eta hauek dira Perturbazio Teoriaren operadoreak. Esan dugun moduan, datu masiboekin analisiak egiterako orduan IA eredu arruntetan aurkitzen ez diren arazoei aurre egin behar zaie. Adibidez aztertzen diren konposatu guztietan ez da aktibitate biologiko berdina neurtzen, entsegu batzuetan IC_{50} neurtzen da, beste batzuetan EC_{50}, edo K_i... Gainera, entseguak ez dira beti ematen baldintza berdinetan, organismo desberdinetan frogatzen baitira, lerro-zelular desberdinetan eta abarretan. Beraz, aktibitatea ez da soilik konposatuaren egituraren araberakoa izango (deskriptore molekularren araberakoa) "boundary-conditions" edo muga-baldintza hauen menpekoa ere izango da. Hau dela eta, ezin da soilik esan konposatu bat aktiboa izango den edo ez; konposatu bati aktibitate biologiko zehatz bat neurtuz gero, eta organismo eta lerro-zelular jakin batzuetan entseatuz gero aktiboa izango denetz iragarri behar da.[24]

Zailtasun guzti hauek kontuan hartuta, PTIA ereduetan lortzen diren ekuazioen eredu orokorra ikus daiteke (6) Ekuazioan.[25]

$$f(v_{ij})_{kalk} = a_0 + a_1 \cdot f(v_{ij})_{erref} + \sum_{k=1}^{k_{max}} b_k \cdot D_{ki}(\text{farmakoa}) + \sum_{k=1,j=0}^{k_{max},j_{max}} c_{k,j} \cdot \Delta D_{ki}(\text{farmakoa}, \mathbf{c}_j) \quad (6)$$

Funtzio honek esan nahi du, "i" konposatu baten aktibitatea ($f(v_{ij})_{kalk}$) aurresan daitekeela "j" baldintzetan. Honetarako erabiltzen den lehen aldagaia $f(v_{ij})_{erref}$ da, erreferentziatzat hartzen den aktibitatearen balioa. Jarraian (4) Ekuazioan ikusitako k deskriptoreen eragin lineala dago aldagaitzat. Azkenik, lehen aipaturiko k deskriptoreen batezbesteko higikorrak daude, $\mathbf{c}_j$ muga-baldintza guztien bektorean kalkulatu egiten direnak. Haien kalkulua egiteko (7) Ekuazioa erabiltzen da, Box-Jenkinsen batezbesteko higikorren antzekoa dena.[26,27]

$$\Delta D_{ki}(\text{farmakoa}, \mathbf{c}_j) = D_{ki}(\text{farmakoa}) - < D_k(\text{farmakoa}, \mathbf{c}_j) > \quad (7)$$

Ekuazio honek adierazten du i konposatu baten k deskriptore molekular baten batezbesteko higikorra kalkulatzeko haren balioari (D_{ki}(farmakoa)) baldintza berdinetan probatuak izan diren konposatu guztien deskriptore horren batezbestekoa (<D_k(farmakoa, $\mathbf{c}_j$)>) kendu behar diogula. Kontuan hartu behar da $\mathbf{c}_j$ bektore bat dela, eta honek esan nahi du baldintza bat baino gehiago hartzen direla kontuan aldi berean, BH-ak kalkulatzeko. Beraz kasu honetan, "Multiple condition Moving Average" edo Batezbesteko Higikor Anizkoitzei (BHA) buruz hitz egin behar da.

(6) Ekuazioan azaldutako kasua eredu lineal bat da Analisi Lineal Diskriminante (ALD) ("Linear Discriminant Analysis", LDA) baten bidez eraiki daitekeena. Aipagarria da mota honetako ereduek abantaila gehigarri bat dutela. Izan ere, aktibitate biologikoa aurresaterako orduan, balio bat lortu beharko litzateke, edozein tartetan egon litekeena. Baina, ALD-ren barne-algoritmo bati esker, posible da emaitza 0 (ez-aktiboa) eta 1 (aktiboa) balioei mugatzea eredua sinpleagoa izan dadin. Horretarako, algoritmoak $f(v_{ij})_{kalk}$ balioen

Mahalanobisen distantzia erabiltzen du aktiboa eta ez-aktiboa izateko probabilitateak kalkulatzeko.[28] Bietako zein den altuagoa determinatuz, $f(v_{ij})_{aurr}$ itzultzen du, aurresandako aktibitatea, alegia (1 aktiboa bada eta 0 aktiboa ez bada).[29]

2. HELBURUAK

Egindako lan honek helburu nagusi argia dauka: minbizia mota askoren kontrako konposatuen aktibitate biologikoa aurresatea entseguetan erabilitako proteinen sekuentzien eragina kontuan hartuta. Horretarako hurrengo urratsak beteko dira:

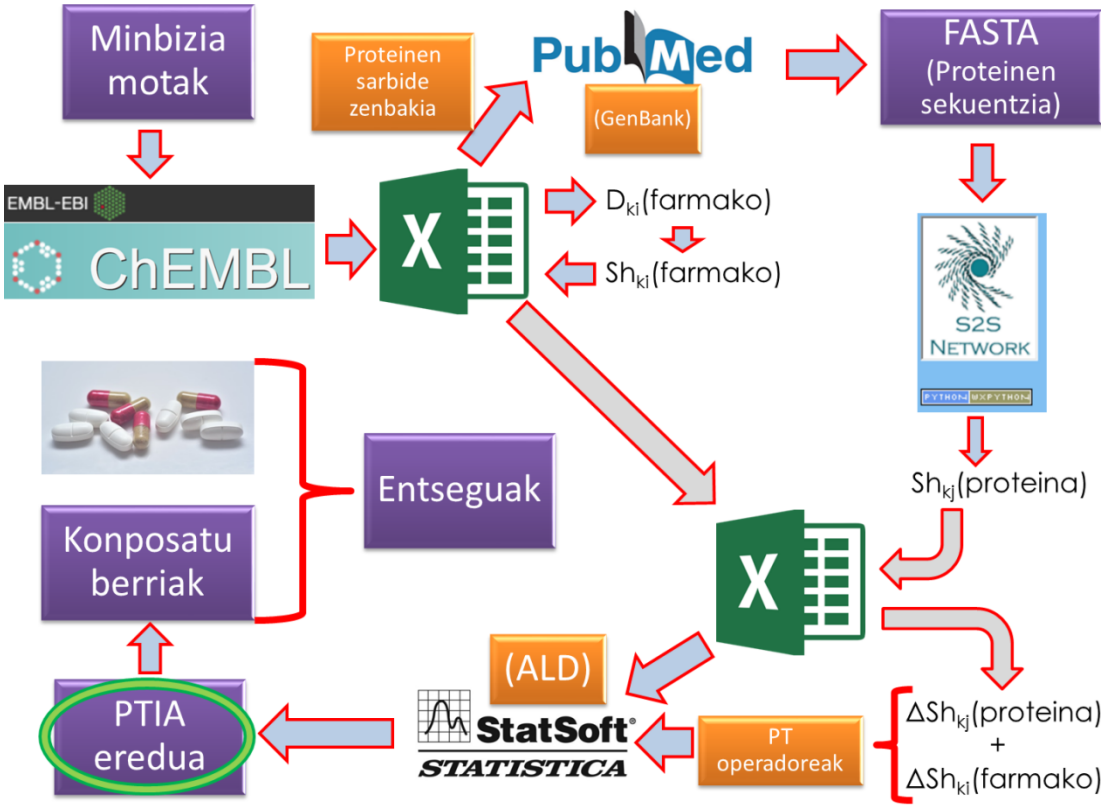
I. Lehenik, guk dakigun arte eraturik ez dagoen eta minbizia mota asko biltzen dituen datu base handia eratu nahi dugu, eta behin datu basea eratuta, minbiziaren kontrako farmako aktiboen egitura nagusiak zeintzuk diren aztertu nahi da.

II. Eredu kimioinformatiko desberdinak eratuko dira kimiometrian oinarrituta haien arteko konparaketa egiteko, eta ikusteko zeintzuk diren aktibitatearen aurreikuspenean gehien eragiten duten faktoreak. Faktore horien artean interes berezia dugu konposatuen itu diren proteinen sekuentziak zenbateko eragina duen neurtzean. Azaldu behar da helburua ez dela aktibitatearen balio numerikoa iragartzea baizik eta konposatua aktiboa izango den edo ez aurresaten duen eredua eratzea ahalik eta kasu gehien ondo iragartzeko.

III. Azkenik, gure ikerkuntza taldean (Organometalikoak sintesian) sintetizaturiko hainbat konposatu organiko berri eredutik pasatu nahi dira minbiziaren kontrako aktibitatea erakuts lezaketen aurresateko.

3. PROZEDURA

Lana aurrera eramateko jarraitu diren pausuak 4. Irudian ikus ditzakegu eskematikoki azalduta.



4. Irudia- *Lanerako jarraituko den prozeduraren eskema.*

3.1. DATU BASEAREN ERAKETA

3.1.1. Konposatuen datuen deskarga

Datu basea eratzeko eman behar den lehengo urratsa datuak bilatzea eta deskargatzea da. Honetarako ChEMBL web orrialdea erabili da.[30] Orrialde horretan hitz gako desberdinak bilatu dira (hainbat minbizia motaren izenak) modu desberdinetan eta bilaketek emaitza kopuru desberdinak eman dizkigute. Lortutako datuak Excel orrietan gorde dira, Excel dokumentu bakoitzean bilaketa bakoitzaren emaitzak gordez (bilaketa bakoitzaren informazioa eranskinetako 1.E. Taulan dago).

3.1.2. Datu-basearen osaketa

Behin konposatu guztiak deskargatuta Excel orriak batu eta dokumentu bakar batean denak elkartu behar dira. Horretarako

konprobatu behar da Excel orri guztietako zutabeak orden berdinean daudela eta bata bestearen azpian itsatsi. Basea jada eratuta egonik, hura garbitu behar da. Hau burutzeko bi urrats eman behar dira, bikoiztutako entseguak ezabatu eta konposatuen aktibitateari, GPA-ri edota AlogP-ari buruzko informaziorik ez dutenak eliminatu; horiek bi izango baitira erabiliko diren deskriptore molekularrak. Azkenik, informaziorik gabeko hutsuneak MD-rekin ("Missing Data", daturik ez) bete behar dira aurreragoko kalkuluetan akatsik ez egoteko. Behin basea garbituta 971481 entsegurekin geratu gara, eta haietatik 258839 entseguk baino ez dute proteinei buruzko informazioa (basearen adibide laburtua eranskinetako 2.E. Taulan).

3.2. DATUEN TRATAMENDUA

3.2.1. Entropien kalkulua

Sarreran azaldu den moduan, hainbeste daturen informazioa hoberen erabiltzeko modua Shannonen entropia kalkulatzean datza. Gainera, horrela deskriptoreen balioak unitate eta eskala berdinean jartzea lortzen dugu. Alde batetik GPA eta AlogP gure deskriptoreen probabilitateak kalkulatu ditugu sarrerako (3) Ekuazioaren antzekoa den (8) Ekuazioa erabiliz.[31]

$$p(D_{ki}) = \frac{D_{ki} - D_{ki,min} + 0.001}{D_{ki,max} - D_{ki,min} + 0.001} \qquad (8)$$

Ekuazio honetan ere D_{ki} gure i konposatuaren k deskriptore molekularra da, baina "$D_{ki,min}$" datu-baseko deskriptore horren baliorik txikiena da 10 aldiz zatituta, eta "$D_{ki,max}$", aldiz, baliorik altuena da 10 aldiz biderkatuta. 10 aldiz balio handiagoak eta txikiagoak ezartzearen arrazoia aurresango ditugun konposatuak tarte horretan sartuko direla ziurtatzea da. Demagun sintetizatutako konposatu berri baten deskriptore batek orain arte neurtutako deskriptore maximoa baino balio altuagoa duela; kasu horretan probabilitatea 1 baino handiagoa izango litzateke eta ekuazioak zentzua izateari utziko lioke.

Azkenik, 0.001 gehitu egin dugu probabilitatearen balioa 0 ez izateko, horrek Shannonen entropiaren indeterminazioa eragingo bailuke.

Ostean Shannonen entropiak kalkulatu ditugu (2) Ekuazioari jarraituz. Hemendik aurrera gure deskriptore molekularrak Shannonen entropiak izango dira, eta beraz, deskriptoreak aipatzerakoan ez gara AlogP edota GPA-ri buruz arituko baizik eta haien entropiei buruz. Honetaz gain, Gradu Amaierako Lan honetan proteinen sekuentzien eragina aztertzen ari garenez, haien deskriptoreak ere beharrezkoak dira. Kasu honetan, Shannonen entropiak lortzeko, prozedura desberdina erabili dugu aurreko biekin konparatuta.

Lehenik proteinen sarbide zenbakiak ("Accession number") datu basetik atera ditugu eta bikoiztuta daudenak eliminatu ditugu guztira 269 proteina desberdin izanik. 269 proteina horien zenbakiak GenBank[32] datu-basean bilatu ditugu eta haien FASTA izeneko kodeak deskargatu ditugu, nukleotido sekuentziaren informazioa daukaten kodeak, alegia. Sekuentzia hori S2SNet[21] programan sartu eta honek proteina guztien Shannonen 6 entropia desberdin itzuli dizkigu sarreran azalduriko moduan. Momentu horretan espero ez genuen arazo bati aurre egin behar izan diogu. Izan ere, proteina guztien entropiak kalkulatuak izan dira birenak izan ezik: O14686 eta Q8NEZ4 sarbide zenbakia dutenak. Konturatu gara kalkulatu gabeko bi proteinek guztien artean sekuentziarik luzeenak dauzkatela, 5537 eta 4911 aminoazido hurrenez-hurren. Irtenbide baten bila, bi sekuentziak 4128 aminoazidotara laburtu ditugu, kalkulatzea posible izan den proteina luzeenaren aminoazido kopurura hain zuzen ere. Oraingoan jada proteina guztien Shannonen entropiak kalkulatu eta datu basean sartu ditugu.

Entropien zereginarekin bukatzeko azken arazo bat aurkitu dugu, baina oraingoan espero genuena. Lehen esan dugun moduan ia

milioi bat entsegu ditugu totalean, baina soilik 250000 inguru eginda daude proteina jakin bat konposatuaren itu delarik, hortaz, proteinarik gabeko entseguekin zer egin erabaki behar izan dugu. Proteinarik ez badago ezin da entropiarik kalkulatu, baina geroago erabiliko den programa estatistikoak balio numeriko bat eskatzen du lan egin ahal izateko. Hori hala izanik, proteinarik gabeko entseguetan proteinen entropiei 0 balioa ematea erabaki dugu. Hau erabakitzeko arrazoia hurrengoa izan da: proteinarik ez dagoenez, edozein parametrok balio hori hartzeko probabilitatea 0 izan behar du, eta 3. Irudian ikusten den moduan, 0 probabilitateak Shannonen entropiaren 0 balioa emateko joera dauka. Dena den, sarreran azaldu den bezala, probabilitatea ezin da 0 izan, bestela entropiaren indeterminazioa ematen da, beraz, ez da irtenbide guztiz egokia eta beste batzuk ere egon daitezke.

3.2.2. Muga baldintzak

Jada azaldu den eran, mota hauetako ereduetan aktibitatea ez da soilik konposatuaren egituraren funtzioa, beste baldintza askok ere eragina dute eta. Eragina duten eta kuantifikagarriak ez diren faktore horiei muga baldintza deritzegu eta datu basea talde desberdinetan banatuko dute. Baldintza hauek aukeratzeko orduan bi gauza izan behar dira kontuan ondorio kontrajarriak eragiten dituztenak. Alde batetik ahalik eta baldintza gehien sartu nahi ditugu ahalik eta faktore gehienak kontuan hartzeko iragarpenak egiterako orduan; baina, beste alde batetik, ezin dugu gure datu basea talde asko eta askotan banatu, bestela talde bakoitzaren entsegu kopurua oso txikia izango litzateke eta ezin izango genuke aurreikuspen onik egin hainbeste datu gutxirekin.

Badaude baldintza batzuk zeinak derrigorrez sartu behar diren ereduan, hala nola neurtutako aktibitatea eta haren unitatea, lerro

zelularra edo organismoa. Hauetaz aparteko baldintzak aukeratzeko aurreko bi faktoreak izan ditugu kontuan eta aukeratutakoak eta haien azalpen laburra jarraian aurki dezakegu:

- c_0, "Activity (units)": zer aktibitate biologiko neurtu den eta zer unitaterekin, K_i (nM), IC_{50} (nM), Potency (nM), inhibition (%)…
- c_1, "Cell name": entsegua zer lerro zelularretan burutu den adierazten du. Lerro zelularrak kodetuta datoz, esaterako, HEp-2, HEK-293T, L1210…
- c_2, "Assay organism" + "organism": kasu honetan bi baldintza konbinatu ditugu. Ikusi dugu edo bi datuetan organismoa berdina dela edo soilik bietako batean datorrela entseguaren organismoa (edo bietako batean ere ez). Hortaz ez dugu informaziorik galdu eta baldintza bat gutxiago erabili dugu. Organismo arruntenak arratoiak (*Mus musculus, Rattus norvegicus…)* eta gizakiak *(Homo sapiens)* dira.
- c_3, "Assay type": zer entsegu mota eraman den aurrera esan nahi du. Letrekin kodetuta datoz eta F ("Functional") motakoa da nagusi nahiz eta beste batzuk ere badauden (A, B…)
- c_4, "Target mapping": zein den konposatuaren itua azaltzen du, itu hori proteina bat edo beste molekularen bat izan daitekeelarik.
- c_5, "Curated by": saiakuntzaren informazio esperimentalaren fidagarritasuna neurtzen du hiru mailatan, "expert" (fidagarritasun altuena), "intermediate" (bitartekoa) eta "autocuration" (fidagarritasun mailarik baxuena).[33]

3.2.3. Batezbesteko higikorrak, muga-balioak, desiragarritasuna, irteerako aldagaia eta esperotako aktibitatea.

Entsegu guztiak behin haien taldeetan sailkatuta, batezbesteko higikorrak kalkulatu behar dira. Horretarako, talde bakoitzeko

konposatuen deskriptore molekularren Shannonen entropien batezbestekoa kalkulatu dugu eta konposatu bakoitzaren deskriptoreari dagokion batezbestekoa kendu diogu (7) Ekuazioari jarraiki. Batezbesteko higikor hauek deskriptore guztientzako kalkulatu ditugu, hau da, AlogP eta GPA-rako eta baita proteinen sei entropietarako ere. Proteinen Shannonen entropien BHA-k kalkulatzeko (ΔSh_{kj}) (7) Ekuazioaren homologoa den (9) Ekuazioa erabili da.[27]

$$\Delta \mathrm{Sh_{kj}}(\mathrm{proteina}, \mathbf{c_j}) = \mathrm{Sh_{kj}}(\mathrm{proteina}) - < \mathrm{Sh_k}(\mathrm{proteina}, \mathbf{c_j}) > \quad (9)$$

Ikusten dugun moduan batezbesteko higikorrak muga-baldintza askok eratutako taldeen barruan kalkulatu ditugu, beraz, batezbesteko higikor anizkoitzak erabili ditugu. BHA-k erabiltzeko arrazoia eta BH sinpleak (muga-baldintza bakar batean oinarriturikoak) alde batera uzteko arrazoia lan honen aitzindaria den H. Bediagak[34] egindako ikerketan daukagu, emaitza hobea lortu baitzuen BHA-k erabiliz sinpleak erabili beharrean (muga baldintzen batezbestekoen adibide laburtua eranskinetako 3.E. Taulan).

Honen ostean, "cutoff" edo muga-balioak deiturikoak ezarri behar dira. Hauek, konposatu aktiboak ez-aktiboetatik banatzen dituzten balioak dira, hau da, konposatu aktiboen eta ez aktiboen arteko limitea ezartzen dutenak. Muga-balioak neurtu diren aktibitate biologiko guztietarako finkatu behar dira, haien unitatea zein den kontuan hartuz. Orain arte egindako saiakuntzetan proposatutako balioei jarraituz nM-etan neurtzen diren aktibitate guztiei 100nM-ko muga ezarri diegu parametro farmakologikoetan oinarrituz. Taldeak burututako lan gehienetan horrela finkatu da unitate horren muga-balioa edozein aktibitaterako eta emaitza onak atera direla kontuan hartuta irizpide berarekin jarraitzea erabaki dugu.[35] Beste kontzentrazio unitate batzuetan dauden aktibitateak 100nM-ren baliokidean jarri

ditugu, 0.1 μM adibidez. Unitatea ehunekotan (%) duten aktibitateen muga balioa %70 izatea erabaki dugu hasiera batean; farmakologikoki nahiko altua den balio bat hartu dugu jakinda emaitza txarrak lortuz gero aldatu egin daitekeela. Beste unitate mota batzuk dituzten aktibitateekin (Kg, g, min^{-1}, egunak…), orokorrean, entsegu gutxiago eginda daude, eta haien muga-balioa batezbestekoarekin finkatzea erabaki dugu; hau da, entsegu horietan neurtutako konposatu guztiek erakutsitako aktibitateen batezbestekoa. Batezbestekoen unitatea 1000 baino altuagoa denean, muga-balioa 1000-n uztea erabaki dugu, litekeena delako entseguren batek aktibitate altuegia aurkeztu izana eta horrek batezbestekoa asko igo izana, beharbada entsegua gaizki egin zelako (muga balioen adibide laburtua eranskinetako 4.E. Taulan).

Muga-balioarekin oso erlazionatuta dagoen beste balioa "desirability" edo desiragarritasuna da; hau da, konposatua aktiboa izan dadin, haren aktibitatearen balioa mugatik gora edo behera egon behar duen ($d(c_0)$). Kasu honetan, aktibitateei banan-banan egokitu zaie desiragarritasuna: -1 (aktiboak muga-balioaren azpitik) edo +1 (aktiboak muga-balioaren gainetik). Adibide moduan, kontzentrazioetan -1 balioa jarri da konposatuaren kontzentrazioa zenbat eta txikiagoa izan efektu zehatz bat lortzeko orduan eta aktiboagoa delako; ehunekoetan +1 finkatu da orokorrean hildako zelulen eta abarren ehunekoari egiten diolako erreferentzia, eta ehunekoa maximizatzea komeni delako; denbora unitateei -1 dagokie tratamendu denbora ahalik eta baxuena izatea komeni delako, eta abar.

Muga balioak eta desiragarritasuna finkatuta, irteerako aldagaia defini dezakegu. Esan dugun bezala, eredu hauetan ez da gure helburua zein izango den konposatu baten aktibitatea aurresatea, baizik eta konposatu bat aktiboa izan daitekeen edo ez aurresatea.

Ondorioz, gure irteerako aldagaiak hori utzi behar du agerian, eta konposatua aktiboa izan daitekeenean 1 balioa ematea nahi dugu eta ez-aktiboa izango deneko kasuetan 0 balioa. Ereduaren erantzunak horiek izateko, datu-baseko konposatuen aktibitate behagarria edo esperimentala definitu behar dugu: $f(v_{ij})_{esp}=1$ aktibitatearen balioa muga balioaren goitik dagoenean eta desiragarritasuna +1 denean edo aktibitatearen balioa mugatik behera dagoenean eta desiragarritasuna -1 denean (aktiboak); eta $f(v_{ij})_{esp}=0$ kontrako bi kasuetan (ez-aktiboak). Zorrotzak izanda, $f(v_{ij})_{esp}=1$ izateak ez du esan nahi konposatua aktiboa denik, baizik eta guk ezarritako muga balioa baino aktibitate hobea daukala. Hemendik aurrera konposatu aktibotzat hartuko ditugu $f(v_{ij})_{esp}=1$ dutenak nahiz eta egiatan balitekeen aktiboak ez izatea.

Datuen tratamenduarekin bukatzeko azkenengo sarrera aldagai bat kalkulatu behar izan da: erreferentziazko aktibitatea. Erreferentzia hori aurretiaz esperotako aktibitatea izatea erabaki dugu. Honek edozein aurreikuspen egin aurretik konposatu bat baldintza jakinetan aktiboa izateko probabilitatea adierazten du. Hau egiteko c_0 muga-baldintza baino ez dugu kontuan izan, neurtutako aktibitatea eta haren unitatea, alegia; a aktibitate mota bakoitzean neurtutako konposatuak hartu ditugu eta aktiboen kopurua kopuru totalarekin zatitu dugu (10) Ekuazioari jarraituz.[35]

$$f(v_{ij})_{erref} = p(v_{ij} = 1)_{erref} = \frac{n_{c_0,a}(f(v_{ij}=1/c_0)}{n_{c_0,a}} \qquad (10)$$

3.3. EREDU KIMIOINFORMATIKOAREN ERAKETA

Eredua eratzeko lehenengo urratsa "training" (entrenamendu sorta) eta "validation"-erako (berrespenerako) konposatuak aukeratzea da. Izan ere, entsegu guztiak ezin dira eredua eratzeko erabili (entrenamendu sorta) batzuk eredua ondo funtzionatzen duen frogatzeko erabili behar dira (berrespena). Aurreko lanetan oinarrituz

konposatuen %75 eredua egiteko eta beste %25-a eredua balidatzeko erabiltzea erabaki dugu.[27] Hautaketa hori zoriz egin da entsegu guztien artean. Ordenean eginez gero baliteke entrenamendu sortako eta berrespeneko konposatuak haien artean oso desberdinak izatea, eta horrek eragingo luke eratutako eredua ezin balidatzea.

Hemendik aurrerako lan guztia Statistica programarekin[36] egin dugu. Hasteko, analisi diskriminante bat egiteko aukera hautatu dugu. Ondoren, gure PTIA ereduaren aldagaiak sartu ditugu, eta horretarako, (11) Ekuazioa planteatu dugu (6) Ekuazioan proteinen aldagai berriak sartuz.

$$f(v_{ij})_{kalk} = a_0 + a_1 \cdot f(v_{ij})_{erref} + \sum_{k=1}^{k_{max}} b_k \cdot Sh_{ki}(farmakoa) + \sum_{k=1}^{k_{max}} c_k \cdot \Delta Sh_{ki}(farmakoa, \mathbf{c}_j) + \sum_{k=0}^{k_{max}} d_k \cdot Sh_{kj}(prot) + \sum_{k=0}^{k_{max}} e_k \cdot \Delta Sh_{kj}(prot, \mathbf{c}_j) \quad (11)$$

(6) Ekuazioan azaldutako aldagaietaz gain (erreferentziazko aktibitatea edo aktiboa izateko aurretiazko probabilitatea, farmakoen deskriptore molekularrak eta haien batezbesteko higikor anizkoitzak) proteinei erlazionaturiko aldagaiak ere sartu ditugu. Hau da, proteinen deskriptore molekularrak (S2SNet-ean kalkulatutako Shannonen 6 entropiak) eta haien BHA-k. Farmako eta proteinen deskriptoreen BHA-k kalkulatzeko modua (7) eta (9) Ekuazioetan azalduta dago hurrenez-hurren.

Planteatutako eredu mota ikusita, sarrerako aldagaitzat hurrengoak hartu ditugu:

- Sh_1(farmako): farmakoen AlogP-aren Shannonen entropia.
- Sh_2(farmako): farmakoen GPA-ren Shannonen entropia.
- ΔSh_1(farmako): farmakoen AlogP-aren Shannonen entropiaren BHA.

- ΔSh_2(farmako): farmakoen GPA-ren Shannonen entropiaren BHA.
- Sh_k(proteina): non k 0 eta 5 artean dagoen eta proteinen Shannonen 6 entropiei egiten dien erreferentzia.
- ΔSh_k(proteina): non k 0 eta 5 artean dagoen eta proteinen Shannonen 6 entropien BHA-ei egiten dien erreferentzia.
- $f(v_{ij})_{erref}$: edozein konposatu aurresan aurretik aktiboa izateko duen probabilitatea da, (10) Ekuazioarekin kalkulatzen dena.

Lehen esan bezala irteerako aldagaitzat aldiz, aktibitate esperimentala ($f(v_{ij})_{esp}$) 0 eta 1 kodeekin aukeratu dugu. Honetaz gain, eredu lineala aukeratu dugu. Azkenik, hainbat saiakuntza egin ditugu *a priori* probabilitate desberdinekin, kalkulu-metodo desberdinak erabiliz eta urrats kopuru maximoa aldatuz.

4. EMAITZAK ETA EZTABAIDA

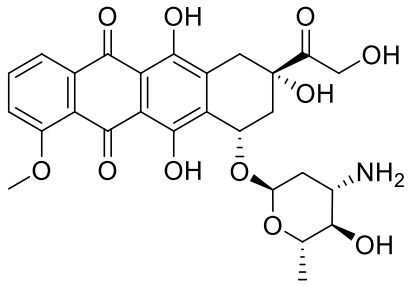
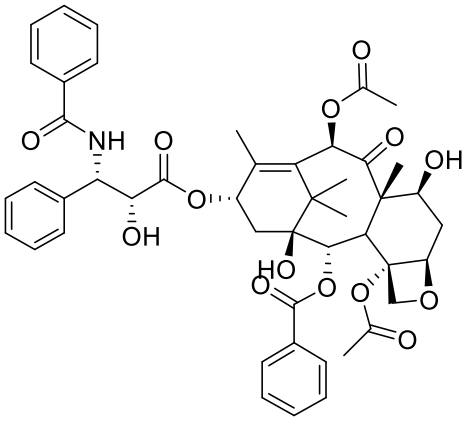
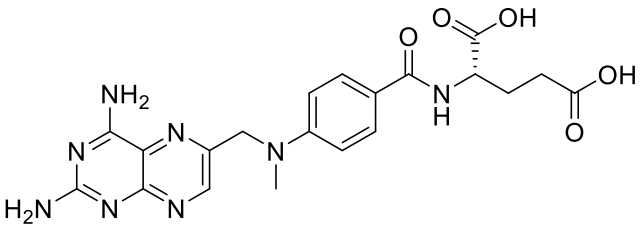
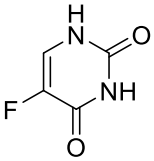
4.1. DATU-BASEA ETA EGITURA AKTIBO NAGUSIAK

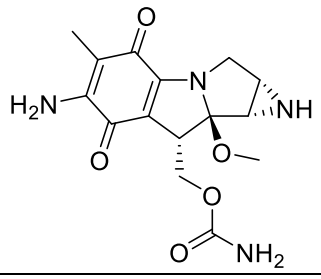
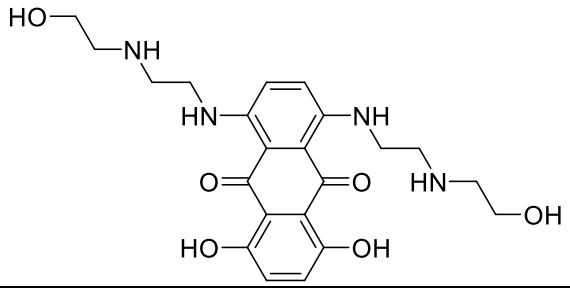
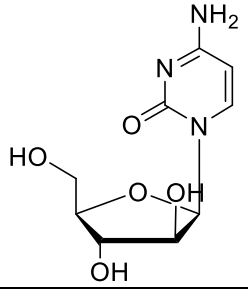
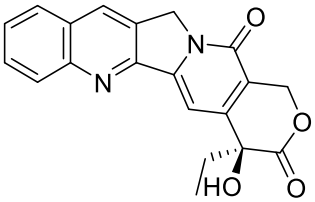
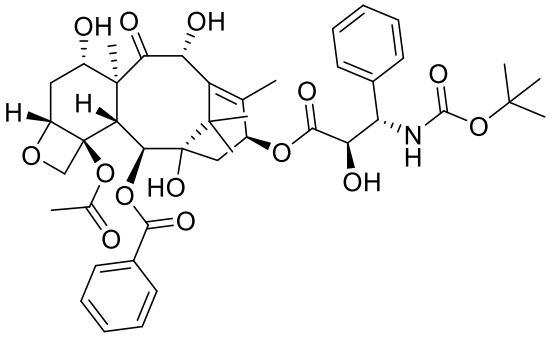
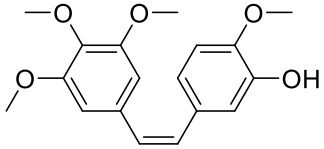
Lanaren lehen helburua erabilitako datu-basea eraikitzean lortu dugu. Izan ere, guk dakigun arte ez zegoen eratuta 17 minbizia mota desberdinen kontrako entseguak barne hartzen dituen hain datu-base handirik. Guztira 971481 entseguri buruzko informazioa bildu da, hala nola, entseguaren baldintzak, erabilitako konposatuaren aktibitatea, deskriptoreak eta abar. Dena dela, entsegu guzti horietatik soilik 258831-etan entseatu zen konposatua proteina baten aurka, eta beraz, bakarrik kasu horietan bildu ahal izan dugu proteinen Shannonen entropiei buruzko informazioa.

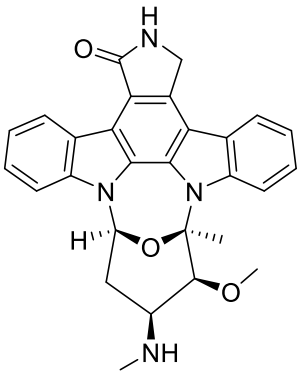
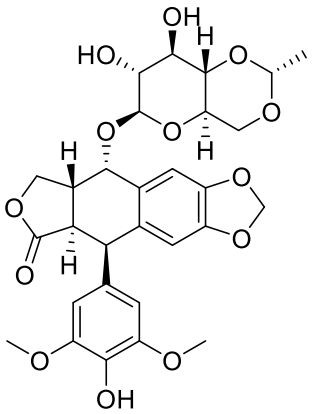
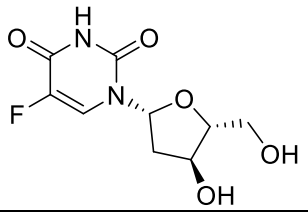
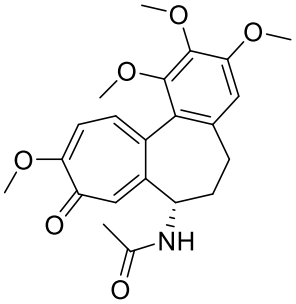
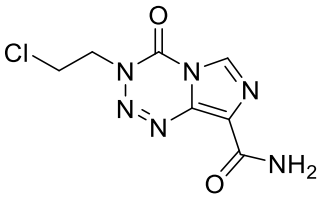
Entsegu guzti horietatik emaitza positiboak (konposatu aktiboak) emaitza negatiboak (konposatu ez-aktiboak) eman dituztenetik bereizi ditugu. Entsegu guztietatik 98122 entsegutan baino ez da konposatu aktiborik lortu, hau da, kasu guztien %10a baino ez da aktiboa guk ezarritako muga-balioekin. Gainera, entsegu positibo

horietatik bakarrik 31455 konposatu aitzindari desberdin topatu dira. Honek esan nahi du aktiboak diren konposatu guztiak 31455 oinarrizko egitura horietatik deribatuak izan zirela. Datu base osoan dauden konposatu aktibo guztien artean, gehien errepikatzen diren konposatu aitzindarien egiturak 1. Taulan adierazten dira haien izen, fase kliniko eta agerraldi kopuruarekin batera.

1. Taula- *Konposatu aktiboen aitzindari errepikatuenen egiturak eta fase klinikoak.*

Izena	Egitura	n*	Fasea
Doxorubizina		479	Onartua eta erabilgarria
Plaklitaxela (taxola)		284	Onartua eta erabilgarria
Metotrexatoa		209	Onartua eta erabilgarria
5-fluorouraziloa		156	Onartua eta erabilgarria

Mitomizina		139	Onartua eta erabilgarria
Mitoxantrona		134	Onartua eta erabilgarria
Zitarabina		119	Onartua eta erabilgarria
Kamptotezina		117	Fase preklinikoan
Dozetaxela		109	Onartua eta erabilgarria
Konbretastatin A-4		99	Fase preklinikoan

Estaurosporina		90	Fase preklinikoan
Etoposida		81	Onartua eta erabilgarria
Floxuridina		79	Onartua eta erabilgarria
Koltxizina		78	Onartua eta erabilgarria
Mitozolomida		78	Fase preklinikoan

*n: egitura bakoitzaren konposatu eratorriak aktibo suertatu diren entsegu kopurua.

Ikusi ditugun egituren artean batzuk konplexuagoak eta besteak sinpleagoak dira; baina kasu gehienetan zenbait ezaugarri errepikatu egiten dira. Funtsean 5 dira gehien ikusten diren taldeak: eraztun

aromatikoak, heterozikloak, aminak, alkohol taldeak (eta eterrak) eta karboniloak. Orokorrean talde hauek polarrak dira eta konposatua entseguaren itu diren molekulekin elkarrekintzan sartzeko aukerak igotzen dituzte. Aipatzekoa da oxigeno eta nitrogenoaz gain ia ez dagoela heteroatomorik eta soilik halogenoren bat agertzen dela noizean behin. Honetaz gain, esan behar dugu, 15 egitura hauek badutela zerikusi handia sarreran ikusitako 2. Irudiko egiturekin. Izan ere, bi egitura errepikatzen dira (5-fluorouraziloa eta dozetaxela) eta doxorrubizina eta epirrubizina oso oso antzekoak dira. Binorelbinak ere 1. Taulako egiturekiko antza handia dauka heteroziklo eta eraztun aromatikoak, aminak, alkohol, eter eta karboniloak azaltzen baitira. Beste aldetik, ziklofosfamidak ez du hainbesterako antzik (fosforoa ez da agertzen 15 egituretan, eta kloroa ez da oso ohikoa) baina egitura aktiboak oso desberdinak izan daitezke haien artean, eta 1. Taulan soilik egitura batzuk aurkeztu dira; horrek ez du esan nahi datu-basean ziklofosfamidaren antzekoagoak diren egitura aktiborik ez dagoenik.

4.2. PTIA EREDUAK

Eredua eraikitzeko bi aukera proposatu ditugu hasiera batean. Gure datu basean proteinari buruzko informazioa duten eta ez duten entseguak daudela, eta gure helburua proteinen eragina kontuan hartzea dela kontuan hartuta, bi eredu egitea erabaki dugu; lehenengoa, non entsegu guztiak barne hartzen ditugun, eta bigarren bat non erabilitako entseguak soilik proteinen informazioa dutenak diren.

4.2.1. Eredua proteina eta proteina gabeko entseguak kontuan hartuta

Eredu honetarako datu baseko 971481 datuak erabili ditugu. Statistica programan hurrengo faktoreak finkatu ditugu: kalkulu metodoa ("forward stepwise"), urrats kopurua (5) eta *a priori* probabilitatea (0.5

aktiboa izateko eta 0.5 ez-aktiboa izateko). Lortutako funtzioa (12) Ekuazioan adierazita dago:

$$
\begin{aligned}
f(v_{ij})_{kalk} = \; & +22{,}8449301038927 \\
& +16{,}8949570478744 \cdot f(v_{ij})_{erref} \\
& -174{,}693989693652 \cdot Sh_1(\text{farmakoa}) \\
& -42{,}3993842349664 \cdot Sh_2(\text{farmakoa}) \qquad (12) \\
& +66{,}3738016943425 \cdot \Delta Sh_2(\text{farmakoa}, \mathbf{c}_j) \\
& -0{,}439208314290454 \cdot \Delta Sh_4(\text{proteina}, \mathbf{c}_j) \\
& n = 971481 \qquad \chi^2 = 298650{,}36 \qquad p < 0.05
\end{aligned}
$$

(12) Ekuazioaren arabera, konposatuen aktibitate biologikoa 5 aldagairen menpekoa da, erreferentziazko balioaren, bi deskriptoreren eta bi batezbesteko higikorren menpekoa. Termino bakoitza modu honetan definitzen da:

- $f(v_{ij})_{erref}$: jada azalduta dagoen moduan, konposatu bati aktibitate jakin bat neurtzerakoan aktiboa izateko aurretiazko probabilitatea da.
- Sh_1(farmakoa) = Sh_{AlogP}(farmakoa): ^{n}oktanol/ur partizio koefizientearen Shannonen entropia da.
- Sh_2(farmakoa) = Sh_{GPA}(farmakoa): GPA-ren edo konposatuen gainazal polarraren azaleraren Shannonen entropiari dagokio.
- ΔSh_2(farmakoa, $\mathbf{c}_j$) = ΔSh_{GPA}(farmakoa, $\mathbf{c}_j$): c_0, c_1, c_2, c_3, c_4 eta c_5 muga baldintzetako GPA-ren batezbesteko higikor anizkoitza da.
- ΔSh_4(proteina, $\mathbf{c}_j$): c_0, c_1, c_2, c_3, c_4 eta c_5 muga baldintzetan proteinaren 4. Shannonen entropiaren BHA da.

- c_1, c_2, c_3, c_4, c_5: prozeduran zehaztutako muga baldintzak dira (aktibitatea, lerro zelularra, organismoa, entsegu mota, itua eta fidagarritasuna).

Eredu honetan proteinen eragina sartu egiten da aktibitatea aurresateko orduan. 5. aldagairik garrantzitsuena proteinen deskriptore baten BHA da, beraz, proteina baten 4. entropia batezbestekotik zenbat aldentzen den garrantzitsua da konposatu baten aktibitatea iragartzeko. Dena den, aldagai hori baino garrantzitsuagoak dira farmakoaren lipofilitatea (AlogP) eta farmakoaren gainazal polarra, baita hura batezbestekotik zenbat aldentzen den. Bestalde, X^2 erabili da p balioa kalkulatzeko, eta $p<0.05$ izanik, esan dezakegu 0 (konposatu ez aktiboak) eta 1 (konposatu aktiboak) taldeak estatistikoki nabarmenki bananduta daudela.

Bi taldeak bata bestetik nabarmenki bananduta daude, baina nahiko bananduak dauden edo ez jakiteko, sentikortasun, espezifikotasun eta zehaztasunari erreparatu behar diegu. Sentikortasuna (Sn) kalkulatzeko $f(v_{ij})_{esp}=1$ duten konposatuak erabiltzen dira (aktiboak), 1 moduan iragarritakoak zati benetan 1 diren guztiak. Espezifikotasuna (Sp) berdin kalkulatzen da baina ez-aktiboak diren konposatuekin eta zehaztasuna (Ac) konposatu guztiekin kalkulatzen da, hau da, asmatutakoak zati kopuru totala. Parametro estatistiko hauek bi multzotarako kalkulatzen dira, prozeduran aipatutako entrenamendu sorta eta berrespenerako. Parametroak 2. Taulan daude kalkulatuta.

2. Taula- *Egindako lehen ereduko parametro estatistikoak.*

Parametroa	Entrenamendu sorta	Berrespena
%Sp	88,9	88,9
%Sn	76,7	76,5
%Ac	87,7	87,6

Parametro hauetatik ondorioak ateratzeko hiru puntu izan behar ditugu kontuan: zehaztasunak %75 eta %95 artean egon behar du eredua egokitzat hartzeko; sentikortasun eta espezifikotasunaren artean oreka egon behar da, ezin da bietako bat oso altua eta bestea oso baxua izan, horrek suposatuko bailuke soilik aktiboak edo ez-aktiboak iragarri ahal izatea baina ez biak batera; azkenik, entrenamendu sortako eta berrespeneko multzoen arteko parametroak oso antzekoak izan behar dira eredua ondo eraiki dela eta ondo balioztatu dela ziurtatzeko.

Gure lehenengo eredu honek 3 betebehar horiek betetzen ditu, beraz, erabilgarria den eredutzat hartu daiteke. Dena den, ereduan zenbait aldaketa sartu ditugu parametroak hobetzen saiatzeko edo eredua sinplifikatzeko aldagai kopurua murriztuz. Horretarako, urrats kopurua eta *a priori* probabilitateak aldatu ditugu eredu berriak eratzeko. Lortutako emaitzak jarraian agertzen dira laburbilduta 3. Taulan.

3. Taula- *Datu guztiekin eratutako eredua hobetzeko egindako saiakerak.*

Eredua	**Urrats kopurua**	***a priori* probabilitatea**	**Parametroak entrenamendu sortan**	**Parametroak berrespenean**
1	4	p(0)=0,5 p(1)=0,5	%Sp=88,9 %Sn=76,7 %Ac=87,7	%Sp=88,9 %Sn=76,5 %Ac=87,6
2	5	p(0)=0,6 p(1)=0,4	%Sp=89,8 %Sn=74,6 %Ac=88,3	%Sp=89,8 %Sn=74,5 %Ac=88,4
3	5	p(0)=0,3 p(1)=0,7	%Sp=88,5 %Sn=77,8 %Ac=87,4	%Sp=88,5 %Sn=77,7 %Ac=87,4

Lehen saiakeran aldagai bat gutxiago sartu da eredua sinpleagoa eginez eta parametroak ia ez dira aldatu. Baina kendutako aldagaia proteinen Shannonen entropiaren BHA izan da, eta beraz, eredu

honek ez du proteinen eragina kontuan hartzen eta ez du gure helburu nagusia betetzen. 2. kasuan, 0 izateko *a priori* probabilitatea aldatu dugu, eta zehaztasuna igo arren, sentikortasunaren beherakada egon da, hortaz, ereduak txarrerantz egin du, espezifikotasun eta sentikortasunaren arteko oreka galdu baitugu. Azkenik, 1 izateko *a priori* probabilitatea handitu dugu, eta zehaztasunak pixka bat behera egin du. Guretzat azken aldaketa honek ez du zentzu handirik, nahi duguna konposatu aktibo eta ez-aktiboak ondo bereiztea delako. Aldiz, aktiboa izan daitekeen edozein konposatu entseatzea helburu duenarentzat zentzuko aldaketa izan daiteke, aktiboak diren konposatu gehien-gehienak ondo iragarriko baititu ez aktiboak diren konposatu asko ere aktibo moduan aurresanez. Aldaketa hauetaz gain beste bi saiakera egin ditugu, kalkulu metodoa aldatuz ("backward stepwise", "forward entry", "backward removal"…) eta prozeduran aktibitate bakoitzarentzat ezarritako muga-balioak aldatuz, baina parametro estatistikoak okertzea baino ez dugu lortu. Hau ikusita, datu guztiak erabiliz lortutako eredurik onena hasieran egindakoa dela ondorioztatu dugu.

4.2.2. Eredua soilik proteinaren sekuentzia duten entseguak kontuan hartuta

Bigarren eredu mota honetarako datu baseko 258839 datu erabili ditugu, soilik proteinen sekuentzia gordeta daukatenak. Aurreko kasuan bezala eredu desberdinak frogatu ditugu. Oraingoan, erabilitako kalkulu metodo bakarra "forward stepwise" delakoa izan da, aurrekoan emaitza onenak eman dituenak, alegia. Statistica programan *a priori* probabilitateak eta urrats kopuruak aldatu ditugu, eta eredu bakoitzean lorturiko parametro estatistikoak 4. Taulan bildu ditugu.

4. Taula- *Soilik proteinaren sekuentzia duten datuekin egindako eredua eratzeko egindako saiakerak.*

Eredua	Urrats kopurua	*A priori* probabilitatea	Parametroak entrenamendu sortan	Parametroak berrespenean
1	5	p(0)=0,5 p(1)=0,5	%Sp=92,0 %Sn=81,8 %Ac=91,0	%Sp=92,0 %Sn=81,2 %Ac=90,9
2	4	p(0)=0,5 p(1)=0,5	%Sp=92,0 %Sn=81,9 %Ac=91,0	%Sp=92,0 %Sn=81,3 %Ac=90,9
3	3	p(0)=0,5 p(1)=0,5	%Sp=92,0 %Sn=81,8 %Ac=91,0	%Sp=91,9 %Sn=81,3 %Ac=90,9
4	4	p(0)=0,6 p(1)=0,4	%Sp=93,1 %Sn=79,8 %Ac=91,8	%Sp=93,1 %Sn=79,4 %Ac=91,7
5	4	p(0)=0,4 p(1)=0,6	%Sp=91,8 %Sn=82,4 %Ac=90,8	%Sp=91,7 %Sn=82,0 %Ac=90,8

Lehenengo, bigarren eta hirugarren saiakerak oso antzekoak dira, baina aldagai kopuru desberdinarekin eta amaieran 4 aldagaikoa aukeratu dugu. Honen arrazoia hurrengoa da. 3 aldagaiko ereduan ez da proteinen deskriptorerik sartzen, eta beraz, gure helburua betetzen ez duenez, baztertuta utzi dugu. Hau dela eta, proteinaren sekuentzia kontuan hartzen duen eta aldagairik kopuru baxuena duen eredua hartu dugu. Aurreko kasuan bezala, *a priori* probabilitateak aldatzea ez du eragin handirik izan. 0-aren probabilitatea igotzean zehaztasuna eta espezifikotasuna igotzen dira baina sentikortasunak nabarmen egiten du behera eta ondorioz, sentikortasun eta espezifikotasunaren arteko oreka galtzen da. 1-en probabilitateak igotzean, aldiz, zehaztasuna galdu egiten da, eta lehen azaldu bezala gure helburuetarako ez da zentzu handiko aldaketa bat, naturan probabilitate altuagoa baitago konposatu bat ez-aktiboa izateko

aktiboa izateko baino. Hau guztia kontuan hartuta 2. saiakerarekin geratu gara eta lortutako funtzioa (13) Ekuazioan daukagu:

$$
\begin{aligned}
f(v_{ij})_{kalk} = \; & -6{,}26531641261611 \\
& +22{,}4884848514776 \cdot f(v_{ij})_{erref} \\
& -70{,}0426856775957 \cdot Sh_2(\text{farmakoa}) \\
& +93{,}7287842105343 \cdot \Delta Sh_2(\text{farmakoa}, \mathbf{c}_j) \\
& -0{,}583335679197273 \cdot \Delta Sh_4(\text{proteina}, \mathbf{c}_j) \\
& n = 258839 \qquad \chi^2 = 118744{,}56 \qquad p < 0.05
\end{aligned}
\tag{13}
$$

(13) Ekuazioaren arabera, konposatuen aktibitate biologikoa 4 aldagairen menpekoa da, erreferentziazko balioaren, deskriptore bakar baten eta bi batezbesteko higikorren menpekoa, hauen esanahia honako hau izanda:

- $f(v_{ij})_{erref}$: konposatu bati aktibitate zehatz bat neurtzean aktiboa izateko aurretiazko probabilitatea da.
- Sh_2(farmakoa) = Sh_{GPA}(farmako): GPA-ren edo konposatuen gainazal polarraren azaleraren Shannonen entropia da.
- ΔSh_2(farmakoa, $\mathbf{c}_j$) = ΔSh_{GPA}(farmako): c_0, c_1, c_2, c_3, c_4 eta c_5 muga baldintzetako GPA-ren batezbesteko higikor anizkoitza.
- ΔSh_4(proteina, $\mathbf{c}_j$): c_0, c_1, c_2, c_3, c_4 eta c_5 muga baldintzetan proteinaren 4. Shannonen entropiaren BHA da.
- c_1, c_2, c_3, c_4, c_5: prozeduran zehaztutako muga baldintzak dira (aktibitatea, lerro zelularra, organismoa, entsegu mota, itua eta fidagarritasuna).

Eredu berri honetan ere proteinen eragina sartu da aktibitatea iragartzeko. 4. aldagairik garrantzitsuena proteinen 4. Shannonen entropiaren BHA da, hortaz, 4. entropia hori batezbestekotik zenbat urruntzen den arabera, aktibitatea ere aldatuko da. Eredu honetan,

konposatuen lipofilitatea ez da faktore erabakigarria aktibitatea aurresateko, eta ondorioz, ez da (13) Ekuazioan sartzen. Aldiz, farmakoen gainazal polarraren azalerak proteinaren deskriptorea baino garrantzitsuagoa izaten jarraitzen du. Honetaz gain, X^2 erabili da berriro ere p balioa kalkulatzeko, eta $p<0.05$ denez 0 eta 1 taldeak esanguratsuki bananduta daude estatistikoki (jada ezaguna zena zehaztasun handia lortu delako eredu honetan).

4.2.3. Eratutako bi ereduen arteko konparaketa

Aurreko ataletan bi eredu desberdin eraiki dira, eta bi horien arteko konparaketa egiteko momentua da. Ereduen egokitzapenari gehien eragiten dieten faktoreen arteko alderaketa 5. Taulan ageri da.

5. Taula- *Eratutako bi eredu nagusien arteko konparaketa.*

		Eredua proteina eta proteina gabeko entseguekin	**Eredua soilik proteinadun entseguekin**
n (entsegu kopurua)		971481	258839
Aldagai kopurua		5	4
Proteinaren aldagaia		ΔSh_4(proteina)	ΔSh_4(proteina)
***A priori* probabilitatea**		p(0)=0,5 ; p(1)=0,5	p(0)=0,5 ; p(1)=0,5
Espezifikotasuna (%Sp)	**Entrenamendu sorta**	88,9	92,0
	Berrespena	88,9	92,0
Sentikortasuna (%Sn)	**Entrenamendu sorta**	76,7	81,9
	Berrespena	76,5	81,3
Zehaztasuna (%Ac)	**Entrenamendu sorta**	87,7	91,0
	Berrespena	87,6	90,9

Faktore deigarriena parametro estatistikoei dagokie, izan ere, kasu guztietan espezifikotasun, sentikortasun eta zehaztasun altuagoa lortzen dugu soilik proteinaren inguruko informazioa duten entseguak erabilita. Honek garrantzi handia dauka, izan ere, esan nahi du

bigarren ereduak ehuneko handiagoan asmatzen duela konposatu bat aktiboa den edo ez, eta beraz, aktibitatea ondo aurresateko aukera gehiago ditugula.

Bestalde, *a priori* probabilitateak berdinak dira eta ereduaren funtzioan sartzen den proteinaren aldagaia ere berdina da (proteinen 4. Shannonen entropiaren BHA). Hala ere, aldagai horrek aktibitatea iragartzeko duen inportantzia ez da berdina, eta eragina altuagoa da bigarren ereduan. Funtsean, bigarren ereduan 4. aldagaia da proteinari dagokiona, eta lehenengo ereduan, aldiz bosgarrena. Hau guztiz zentzuzkoa da, izan ere, proteinak garrantzi handiagoa izan behar du hari buruzko informazioa duten entseguetan, haren informazioa ez duten entseguetan baino. Hau dela eta, lehenengo ereduan gutxienez bost aldagai sartzera behartuta gaude gure helburua betetzeko, eta honek (12) Ekuazioa (13) Ekuazioa baino konplexuagoa izatea eragiten du. Azkenik erabilitako entsegu kopurua dugu, hau zalantza barik altuagoa da lehenengo ereduan, datu base osoa erabili baita. Honek eredua zabalagoa eta konposatu mota gehiagorentzat izatea eragin dezake, baina faktore hau ez da garrantzitsuegia aurrekoekin alderatuz gero. Ondorioz, esan dezakegu, gure helburuetarako eredu zuzenagoa dela bigarrena, nahiz eta bien arteko diferentzia oso handia ez izan.

4.3. AURRETIK EGINDAKO EREDUEN ETA GURE EREDUEN ARTEKO KONPARAKETA

Aurreko atalean lortu ditugun bi ereduak beste lan batzuetan egindako ereduekin konparatu dira. Konparatutako gainerako eredu guztiak minbiziaren kontrako ereduak dira, baina gure lanarekiko desberdintasunak dituzte zenbait alorretan. Desberdintasun eta antzekotasun horiek 6. Taulan dauzkagu adierazirik.

6. Taula- *Gure bi ereduen eta aurretik eratutako minbiziaren kontrako beste ereduen arteko konparaketa.*

Minbizi mota	Proteinen sekuentzia	Soilik proteinen sekuentziako entseguak	Kantzer mota desberdin kopurua	Erabilitako metodoa	Ereduaren aldagai kopurua	Entsegu kopurua	Sn(%)	Sp(%)	Erreferentzia
Bularrekoa	Ez	-	1	ALD	>10	24285	>90	>90	37,38
Maskurikoa	Ez	-	1	ALD	E.E.*	E.E.	>90	>90	39
Garunekoa	Ez	-	1	ALD	E.E.	E.E.	>90	>90	40
Kolonekoa	Ez	-	1	ALD	>10	1237 ("training")	>90	>90	41
Bularrekoa	Ez	-	1	ALD	>10	2272	>85	>95	42
Prostatakoa	Ez	-	1	ALD	>10	1250 ("training")	>85	>95	43
Mota anitz	Ez	-	>10	ALD	>10	87701	>70	90	25
				ALD	3		>70	>90	
				SNA*	4		>80	>80	
Mota anitz	Bai	Ez	>15	ALD	5	971481	>75	>85	Lan hau
	Bai	Bai	>15	ALD	4	258839	>80	>90	Lan hau

*SNA: Sare Neuronal Artifizialak; E.E.: Ez Eskuragarri.

6. Taula honetan lehendabizi ikusten duguna zera da, guk dakigun arte proteinaren sekuentzia konposatuen aktibitate biologikoa aurresateko kontuan hartzen duten eredu bakarrak Gradu Amaierako Lan honetan sortutakoak direla. Honetaz gain, lehen garatu den moduan, bi eredu sortu dira bata entsegu guztiekin eta bestea soilik proteinadun entseguekin. Gainera, gure bi ereduek minbizi mota kopuru handiena hartzen dute barne, argitaratuta dauden gehienak minbizi mota bakar batean zentratzen baitira. Honetaz aparte, gainerakoekin konparatuz, nahiko aldagai gutxi erabili ditugu gure funtzioa osatzeko, beraz, eredu nahiko sinpleak eratu ditugu. Egia da minbizi mota anitzetarako garatutako ereduetako bi oraindik sinpleagoak direla, baina bi horietan espezifikotasun eta sentikortasun baxuenak daude.[25]

Bestalde, kasu guztietan batean izan ezik analisi lineal diskriminantea erabili da. Hau erabili ez denean sare neuronal artifizialak erabili dira, baina metodo hau aurretik egindako ALD ereduen parametro estatistikoak hobetzeko baino ez zen egin. Parametro estatistikoak aipatuta, gure bi ereduak ez dira Sn eta Sp altuenak dituztenak, baina honek arrazoi logiko bat dauka. Lehenengo 6 ereduetan minbizia mota bakar bat lantzen da, eta entsegu gutxiago dago, hortaz, pentsatzekoa da konposatu aktibo guztiak antzekoagoak izatea gure datu basean baino. Hala ere, eredu horiek kantzer mota bakar batera daude mugatuta gureak askoz zabalagoak diren bitartean. Funtsean, minbizi mota askoren kontra dauden beste hiru ereduak gureak baino espezifikotasun eta sentikortasun baxuagoa dute, hortaz, gure ereduak oso eredu erabilgarriak direla ondoriozta dezakegu.

Azkenik, lan honetan garatutako ereduak berezi eta berritzaile egiten dituen azken faktore bat ere badaukagu; izan ere, emaitzen lehenengo atalean esandako moduan, guk dakigun arte, minbiziaren

kontra eraiki den datu-baserik handiena erabili dugu, gainerako lanetan erabilitako entseguak baino askoz kasu gehiago dituenak.

4.4. LABORATEGIAN SINTETIZATURIKO KONPOSATU BERRIEN AKTIBITATEAREN IRAGARPENA

Lan honen azken helburua gure ikerketa taldeko laborategietan sintetizatutako konposatu organiko berriak eredutik pasatzean datza. Honetarako konposatu berrien GPA kalkulatu dugu Dragon softwarearekin[16] eta datu horiek (13) Ekuazioan ordezkatu ditugu, parametro estatistiko onenak dituen ereduan, alegia. Modu horretan konposatu berri bakoitzaren $f(v_{ij})_{kalk}$ kalkulatu dugu eta aktiboa izateko probabilitatea lortzeko minimo-maximo normalizazio bat erabili dugu datu base osoan lortutako $f(v_{ij})_{kalk}$ balio maximo eta minimoekin.[20] Hala, konposatu bakoitza baldintza zehatzetan eta proteina jakinen kontra aktiboa izateko probabilitatea kalkulatu da eta emaitzak 7. Taulan agertzen dira. Taula honetan kolorea zenbat eta gorriagoa izan orduan eta txikiagoa da aktiboa izateko probabilitatea eta zenbat eta berdeagoa izan, orduan eta probabilitate handiagoa dugu.

7. Taula- *Gure taldean sintetizatutako konposatu organiko berriak aktiboak izateko probabilitateak.*

Muga-baldintzak	**c_0: GI_{50} (nM) c_1: SK-MEL-2 c_2: Homo sapiens c_3: F c_4: Proteina anitz c_5: Autocuration**		**c_0: Inhibition (%) C_1: - c_2: Homo sapiens c_3: F c_4: Proteina c_5: Autocuration**	
Proteina / Konposatua*	**O00255**	**O75496**	**Q03164**	**O75496**
AC520	0,1423	0,1654	0,4795	0,5421
AC534	0,1449	0,1679	0,4821	0,5446
AC538	0,1463	0,1693	0,4835	0,5460
AC539	0,1463	0,1693	0,4835	0,5460
AC606	0,1419	0,1649	0,4791	0,5416
AD17	0,1419	0,1649	0,4791	0,5416

CSA114f2	0,1418	0,1648	0,4790	0,5416
CSA117f2	0,1491	0,1721	0,4863	0,5489
CSA130f2	0,1468	0,1698	0,4840	0,5466
CSA133f2	0,1468	0,1698	0,4840	0,5466
CSA136f2	0,1418	0,1648	0,4790	0,5416
CSA137f3	0,1418	0,1648	0,4790	0,5416
CSA140f2	0,1481	0,1712	0,4853	0,5479
CSA146f2	0,1418	0,1648	0,4790	0,5416
CSA147f2	0,1418	0,1648	0,4790	0,5416
CSA151f2	0,1418	0,1648	0,4790	0,5416
CSA155f1	0,1418	0,1648	0,4790	0,5416
CSA197f2.1	0,1418	0,1648	0,4790	0,5416
CSA210f2	0,1418	0,1648	0,4790	0,5416
CSA230f2	0,1534	0,1765	0,4906	0,5532
CSA241f2	0,1395	0,1625	0,4767	0,5393
CSA243f3	0,1395	0,1625	0,4767	0,5393
CSA246f2	0,1380	0,1610	0,4752	0,5378
CSA249f2	0,1434	0,1664	0,4806	0,5431
IB001	0,1373	0,1603	0,4745	0,5371
IB0011	0,1343	0,1573	0,4715	0,5341
IB0012	0,1343	0,1573	0,4715	0,5341
IB002	0,1475	0,1706	0,4847	0,5473
IB003	0,1475	0,1706	0,4847	0,5473
IB004	0,1343	0,1573	0,4715	0,5341
IB005	0,1343	0,1573	0,4715	0,5341
IB006	0,1426	0,1657	0,4798	0,5424
IB007	0,1343	0,1573	0,4715	0,5341
IB008	0,1475	0,1706	0,4847	0,5473
IB009	0,1401	0,1631	0,4773	0,5399
IB010	0,1475	0,1706	0,4847	0,5473
MM001	0,1423	0,1654	0,4795	0,5421
MM002	0,1423	0,1654	0,4795	0,5421
MM003	0,1492	0,1723	0,4864	0,5490
MM004	0,1446	0,1676	0,4818	0,5444
MM005	0,1419	0,1649	0,4791	0,5416
MM006	0,1449	0,1679	0,4821	0,5446
MM007	0,1449	0,1679	0,4821	0,5446
MM008	0,1423	0,1654	0,4795	0,5421
MM009	0,1369	0,1599	0,4741	0,5366

*Carlos Santiago, Mikel Martínez, Asier Carral eta Iratxe Barbolla, doktoretza tesiak garatzen, Leioan UPV-EHU (konposatu bakoitzaren egitura eta deskriptoreak eranskinetako 5.E. Taulan).

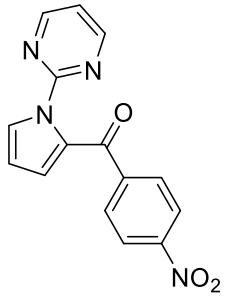
Argi dago azken bi zutabeetako muga baldintzak konposatu bat aktiboa izateko egokiagoak direla. Honen arrazoia neurtutako aktibitate motan dago; izan ere, inhibizioa (ehunekotan) neurtzean aurretiaz konposatu bat aktiboa izateko probabilitatea altuagoa da, hau da, ereduan agertzen den $f(v_{ij})_{erref}$ altuagoa da azken bi zutabeetan. Inhibizioak konposatu batek entzima baten (edo beste proteina baten) aktibitatea ekiditeko duen gaitasunari dagokio, eta GI_{50}, aldiz, zelula tumoralen ugaritzea %50-ean inhibitzeko konposatuaren beharrezko kontzentrazioa da (nM-tan).[44,45] SK-MEL-2 lerro zelularra gizonezkoek pairatutako melanomako lerro zelular bat da eta bigarren zutabeko entsegua, berriz, ez zen lerro zelular jakin batean egin.[46] Gainerako baldintzak berdinak dira bi entsegu motetan, biak gizakietan egin ziren, itua proteinak ziren, fidagarritasun maila baxuena da bietan eta F entsegu mota erabili zen. F entsegu mota hauek "functional" izeneko entseguak dira, konposatuaren eragin biologikoa neurtzen dutenak.[30]

Bestalde, argi geratzen den beste efektu bat proteinarena da. Izan ere, O75496 itua denean, konposatua aktiboa izateko probabilitatea altuagoa da O00255 edo Q03164 proteinak itu direnean baino. Honen arrazoia hiru proteinen 4. Shannonen entropiaren balioan dago. O75496-ren balioa hiruretatik baxuena da, beraz, BHA ere baxuena da, eta haren koefizientea (13) Ekuazioan negatiboa denez, zenbat eta baxuagoa izan BHA orduan eta altuagoa da $f(v_{ij})_{kalk}$ eta, hortaz, aktiboa izateko probabilitatea ere altuagoa da.

O75496 proteinak "Geminin" izena du eta zelulen ugalketa inhibitzen duen 25 kDa-eko eta 209 aminoazidoko proteina da, beharbada, hori da datu base osoan entsegu gehien jasan dituen proteina izatearen arrazoia.[47] O00255-k "Menin" du izena, eta sistema endokrinoko minbizietan agertzen den 610 aminoazidoz osatutako proteina da.[48] Azkenik Q03164 histona-lisina-*N*-metiltransferasa 2A da, 3969

aminoazido ditu eta DNA eraldatzen du leuzemia eraginez gizakietan.[49]

Iragarritako 45 konposatuetatik aktiboena GPA-rik handiena duen konposatua da, CSA230f2 kodea daukan konposatua hain zuzen ere, 93.6 $Å^2$-ko GPA duena. (13) Ekuazioan ikusten dugunez GPA-ren kasuan bi aldagai kontrajarri ditugu. Alde batetik, GPA zenbat eta altuagoa izan orduan eta altuagoa da haren Shannonen entropia (7. Taulako konposatuetan), eta honen koefizientea negatiboa denez, aktiboa izateko probabilitateak behera egingo du. Baina, beste aldetik, GPA-ren entropiaren BHA-k koefiziente positiboa du eta hau dela eta, Sh_{GPA}(farmako) zenbat eta altuagoa izan orduan eta probabilitate altuagoa egongo da konposatu bat aktiboa izan dadin. Bi faktore kontrajarri horietatik koefiziente altuena daukana batezbesteko higikorra da, koefiziente positiboa duena, alegia. Ondorioz, GPA-ren entropia igo ahala aktiboa izateko probabilitatea igotzen da aldagaien kolinealtasuna alde batera uzten baldin badugu. CSA230f2 konposatu honen egitura 5. Irudian daukagu.



5. Irudia- *CSA230f2 konposatuaren egitura.*

Emaitzen lehen atalean ikusi dugun moduan, minbiziaren kontra aktiboak diren egiturak nahiko desberdinak izan daitezke haien artean. Hala ere, ezaugarri nagusi orokor batzuk azpimarratu ditugu eta molekula honek horietako batzuk betetzen ditu: eraztun aromatikoak ditu, heterozikloak, karbonilo bat, amina tertziario bat… Orokorrean, eta lehen esan bezala ituarekiko elkarrekintzak

ahalbidetzen dituzten funtzio talde polarrak daude, esaterako, karboniloa, pirrola eta pirimidina taldeak. Gainera 1. Taulan arrunta ez den talde bakarra du, nitro taldea hain zuzen ere, ez baitu halogenorik edo nitrogenoa eta oxigenoa ez diren bestelako heteroatomorik.

Atal honi amaiera emateko, esan behar da 5. Irudiko konposatuak aktiboa izateko probabilitate altuena daukala, baina horrek ez du esan nahi konposatu aktiboena izango denik. Funtsean, konposatu guztiek antzeko probabilitateak dituzte aktiboak izateko, nahiko egitura antzekoak izanik (heteroziklo eta eraztun aromatikoetan oinarrituak) eta heteroatomoak berdinak izanik (nitrogenoa eta oxigenoa batez ere, baina baita halogenoren bat egitura batzuetan) deskriptore molekularren balioak antzekoak dira eta (GPA-ren balioa, alegia). Gainera, probabilitate horiek kalkulatzeko ez da Mahalanobisen distantzia kalkulatu ALD-ak egiten duen moduan, horren ordez, gure datu-baseko $f(v_{ij})_{kalk}$ balio maximo eta minimoak erabili dira. Beraz, probabilitateen kalkulua pixka bat alda liteke Statistica programaren kalkulutik.

5. ONDORIOAK

Laburbilduz, Gradu Amaierako Lan honetan honako ondorio hauek lortu ditugu:

I. Minbiziaren kontrako konposatuen aktibitatea proteinen sekuentziaren arabera aurresateko bi eredu kimioinformatiko eraiki ditugu. Batean sortutako datu baseko kasu guztiak hartu dira kontuan eta bestean soilik proteinen sekuentzia duten kasuak. Dena dela, bi ereduetan kasu asko eta asko zegoen, eta espezifikotasun, sentikortasun eta zehaztasun oso onak lortu direnez bi ereduekin, PTIA ereduak datu masiboetako baseekin lan egiteko oso eredu egokiak direla ondorioztatu dugu.

II. Gure bi ereduak aurretik egindako beste batzuekin konparatzean ikusi dugu gureak nahiko berritzaile eta bereziak direla; izan ere, proteinen sekuentzien eragina kontuan hartzen duten lehenak dira eta gainera, entsegu gehien eta minbizi mota gehien dituztenak dira.

III. Azkenik, laborategian gure ikerkuntza-taldeak sintetizaturiko 45 konposatu berri eredutik pasa ditugu eta ikusi dugu inhibizioa ehunekotan neurtuz gero eta "Geminin" proteina entseguaren itua izanik, aktiboak izateko probabilitateak 0.5 baino altuagoak direla. Gainera, probabilitate altueneko konposatua GPA altuenekoa dela azaldu dugu eta haren egitura datu basean gehien agertzen diren egitura aktiboekin alderatzean, ohartu gara, zenbait ezaugarri komunean dituztela; hala nola, heterozikloak, eraztun aromatikoak eta karboniloak bezalako talde polarrak.

6. BIBLIOGRAFIA

1. *Las cifras del cáncer en España 2019;* Sociedad Española de Oncología Médica, Depósito Legal: M-3800-2019, https://seom.org/publicaciones/el-cancer-en-espanyacom (sarrera 2019/06/11)

2. *Defunciones según la Causa de Muerte 2016;* Instituto Nacional Estadística, Notas de Prensa, 2017/12/21

3. National Cancer Institute: Risk Factors for Cancer https://www.cancer.gov/about-cancer/causes-prevention/risk (sarrera 2019/05/23)

4. Nachman, M.W.; Crowell, S.L. Estimate of the Mutation Rate per Nucleotide in Humans. *Genetics.* **2000**, 156, 297-304

5. Weinstein, B.; Joe, A. Oncogene Addiction. *Cancer Res.* **2008**, 68, 3077-3080

6. Meza, R.; Jeon, J.; Moolgavkar, S.H.; Luebeck, E.G. Age-specific incidence of cancer: Phases, transitions, and biological implications. *PNAS*. **2008**, 105, 16284-16289

7. Folkman, J.; Tumor angiogénesis therapeutic implications. *The New England Journal of Medicine*. **1971**, 285, 1182-1186

8. Johnstone, T.C.; Park, G.Y.; Lippard, S.J. Understanding and Improving Platinum Anticancer Drugs – Phenanthriplatin. *Anticancer Research*. **2014**, 34, 471-476

9. Gasser, G.; Ott, I.; Metzler-Nolte, N. Organometallic Anticancer Compounds. *J. Med. Chem.* **2011**, 54, 3-25

10. Yeo, C.I.; Ooi, K.K.; Tiekink, E.R.T. Gold-Based Medicine: A Paradigm Shift in Anti-Cancer Therapy? *Molecules*. **2018**, 23, 14-23

11. Joensuu, H.; Bono, P.; Kataja, V.; Alanko, T.; Kokko, R.; Asola, R.; Utriainen, T.; Turpeeniemi-Hujanen, T.; Jyrkkiö, S.; Möykkynen, K.; Helle, L.; Ingalsuo, S.; Pajunen, M.; Huusko, M.; Salminen, T.; Auvinen, P.; Leinonen, H.; Leinonen, M.; Isola, J.; Kellokumpu-Lehtinen, P.L. Fluorouracil, Epirubicin, and Cyclophosphamide With Either Docetaxel or Vinorelbine, With or Without Trastuzumab, As Adjuvant Treatments of Breast Cancer: Final Results of the FinHer Trial. *Journal of Clinical Oncology*. **2009**, 27, 5685-5692

12. Zhang, L.; Tan, J.; Han, D.; Zhu, H. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug Discovery Today*. **2017**, 22, 1680-1685

13. Crum-Brown, A.; Fraser, T.R. On the connection between chemical constitution and physiological action. *Trans. R. Soc. Edinburgh.* **1868**; 25, 151-203.

14. Todeschini, R.; Consonni, V. Handbook of Molecular Descriptors; Wiley-VCH, Weinheim, Germany, **2008**.

15. Ponce, Y.M. Total and local (atom and atom type) molecular quadratic indices: significance interpretation, comparison to other molecular descriptors, and QSPR/QSAR applications. *Bioorg. Med. Chem.* **2004**, 12, 6351-6369

16. Mauri, A.; Consonni, V.; Pavan, M.; Todeschini, R. Dragon software: an easy approach to molecular descriptor calculations. *MATCH Commun. Math. Comput. Chem.* **2006**, 56, 237-248

17. Mannhold, R.; Dross, K. Calculation Procedures for Molecular Lipophilicity: a Comparative Study. *Quant. Struct.-Act. Relat.* **1996**, 15, 403-409

18. Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000**, 43, 3714-3717

19. Shannon, C.E. A Mathematical Theory of Communication, *The Bell System Technical Journal*. **1948**, 27, 379-423

20. Aksoy, S.; Haralick, R.M. Feature normalization and likelihood-based similarity measures for image retrieval. *Pattern Recognition Letters*. **2001**, 22, 563-582

21. Munteanu, C.R.; Magalhaes, A.L.; Duardo-Sánchez, A.; Pazos, A.; González-Díaz, H. S2SNet: A Tool for Transforming Characters and Numeric Sequences into Star Network Topological Indices in Chemoinformatics, Bioinformatics, Biomedical, and Social-Legal Sciences. *Current Bioinformatics*. **2013**, 8, 429-437

22. Manallack, D.T.; Ellis, D.D.; Livingstone, D.J. Analysis of Linear and Nonlinear QSAR Data Using Neural Networks. *J. Med. Chem.* **1994**, 37, 3758-3767

23. Ferreira da Costa, J.; Silva, D.; Caamaño, O.; Brea, J.M.; Loza, M.I.; Munteanu, C.R.; Pazos, A.; García-Mera, X.; González-Díaz, H. Perturbation theory/machine learning model of ChEMBL data for dopamine targets: docking, synthesis, and assay of new l-prolyl-l-leucyl-glycinamide peptidomimetics. *ACS Chemical Neuroscience*. **2018**, 9, 2572-2587

24. González-Díaz, H.; Arrasate, S.; Gómez-SanJuan, A.; Sotomayor, N.; Lete, E.; Besada-Porto, L.; Ruso, J.M. General Theory for Multiple Input-Output Perturbations in Complex Molecular Systems. 1. Linear QSPR Electronegativity Models in Physical, Organic, and Medicinal Chemistry. *Current Topics in Med. Chem*. **2013**, 13, 1713-1741

25. Bediaga, H.; Arrasate, S.; González-Díaz, H. PTML Combinatorial Model of ChEMBL Compounds Assays for Multiple Types of Cancer. *ACS Combinatorial Science*. **2018**, 20, 621-632

26. Box, G.E.P.; Pierce, D.A.; Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models. *Journal of the American Statistical Association*. **1970**, 65, 1509-1526

27. Blay, V.; Yokoi, T.; González-Díaz, H. Perturbation Theory−Machine Learning Study of Zeolite Materials Desilication. *J. Chem. Inf. Model*. **2018**, 58, 2414−2419

28. De Maesschalck, R.; Jouan-Rimbaud, D.; Massart, D.L. The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*. **2000**, 50, 1-18

29. Huberty, C.J.; Discriminant Analysis. *Review of Educational Research*. **1975**, 45, 543-598

30. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A.P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L.J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M.P.; Overington, J.P.; Papadatos, G.; Smit, I.; Leach, A.R. The ChEMBL database in 2017. *Nucleic Acids Research*, **2017**, 45, D945–D954

31. Herrera-Ibatá, D.M.; Pazos, A.; Orbezogo-Medina, R.A.; Romero-Durán, F.J.; González-Díaz, H. Mapping chemical structure–activity information of HAART-drug cocktails over complex networks of AIDS epidemiology and socioeconomic data of U.S. counties. *BioSystems*. **2015**, 3564, 1-15

32. Benson, D.A.; Karsch-Mizrachi, I.; Lipman, D.J.; Ostell, J.; Wheeler, D.L. GenBank. *Nucleic Acids Research.* **2005**, 33, D34-D38

33. Cartwright, H. Artificial Neural Networks; Humana Press, New York, **2015**

34. Bediaga, H. Eredu kimioinformatikoak: minbiziaren kontrako konposatuen diseinua. GrAL, Euskal Herriko Unibertsitatea, Leioa, **2018**

35. Nocedo-Mena, D.; Cornelio, C.; Camacho-Corona, M.R.; Garza-González, E.; Waksman de Torres, N.; Arrasate, S.; Sotomayor, N.; Lete, E.; González-Díaz, H. Modeling Antibacterial Activity with Machine Learning and Fusion of Chemical Structure Information with Microorganism Metabolic Networks. *J. Chem. Inf. Model.* **2019**, 59, 1109−1120

36. StatSoft, Inc. (2001). STATISTICA (data analysis software system), version 6. www.statsoft.com

37. Speck-Planche, A.; Cordeiro, M.N. Erratum to: Fragment-based in silico modeling of multi-target inhibitors against breast cancer-related proteins. *Mol. Divers.* **2017**, 21, 525.

38. Speck-Planche, A.; Cordeiro, M.N. Fragment-based in silico modeling of multi-target inhibitors against breast cancer-related proteins. *Mol. Divers*. **2017**, 21, 511-523.

39. Speck-Planche, A.; Kleandrova, V.V.; Luan, F.; Cordeiro, M.N. Unified multi-target approach for the rational in silico design of anti-bladder cancer agents. *Anticancer Agents Med. Chem*. **2013**, 13, 791-800.

40. Speck-Planche, A.; Kleandrova, V.V.; Luan, F.; Cordeiro, M.N. Chemoinformatics in multi-target drug discovery for anti-cancer therapy: in silico design of potent and versatile anti-brain tumor agents. *Anticancer Agents Med. Chem*. **2012**, 12, 678-685.

41. Speck-Planche, A.; Kleandrova, V.V.; Luan, F.; Cordeiro, M.N. Rational drug design for anti-cancer chemotherapy: multi-target QSAR models for the in silico discovery of anti-colorectal cancer agents. *Bioorg. Med. Chem*. **2012**, 20, 4848-4855.

42. Speck-Planche, A.; Kleandrova, V.V.; Luan, F.; Cordeiro, M.N. Chemoinformatics in anti-cancer chemotherapy: multi-target QSAR model for the in silico discovery of anti-breast cancer agents. *Eur. J. Pharm. Sci.* **2012**, 47, 273-279.

43. Speck-Planche, A.; Kleandrova, V.V.; Luan, F.; Cordeiro, M.N. Multi-target drug discovery in anti-cancer therapy: fragment-based approach toward the design of potent and versatile anti-prostate cancer agents. *Bioorg. Med. Chem*. **2011**, 19, 6239-6244.

44. Cleland, W.W. The kinetics of enzyme-catalyzed reactions with two or more substrates or products. *Biochim. Biophys. Acta.* **1963**, 67,173-187

45. Merschjohann, K.; Steverding, D.; In vitro trypanocidal activity of the anti-helminthic drug niclosamide. *Experimental Parasitology*. **2008**, 118, 637–640

46. Fogh, J.; Fogh, J.M.; Orfeo, T. One Hundred and Twenty-Seven Cultured Human Tumor Cell Lines Producing Tumors in Nude Mice. *J. Natl. Cancer Inst*. **1977**, *59*, 221-226

47. McGarry, T.J.; Kirschner, M.W. Geminin, an Inhibitor of DNA Replication, Is Degraded during Mitosis; *Cell*. **1998**, 93, 1043-1053

48. Chandrasekharappa, S.C.; Guru, S.C.; Manickam, P.; Olufemi, S.E.; Collins, F.S.; Emmert-Buck, M.R.; Debelenko, L.V.; Zhuang, Z.; Lubensky, I.A.; Liotta, L.A.; Crabtree, J.S.; Wang, Y.; Roe, B.A.; Weisemann, J.; Bogusky, M.S.; Agarwal, S.K.; Kester, M.B.; Kim, Y.S.; Heppner, C.; Dong, Q.; Spiegel, A.M.; Burns, A.L.; Marx, S.J. Positional Cloning of the Gene for Multiple Endocrine Neoplasia-Type 1. *Science*. **1997**, 276, 404-407

49. Tkachuk, D.C.; Kohler, S.; Cleary, M.L.; Involvement of a Homolog of Drosophila Trithorax by 11q23 Chromosomal Translocations in Acute Leukemias. *Cell*. **1992**, 71, 691-700

Eranskinak

1.E. Taula- *ChEMBL orrialdean egindako bilaketen informazioa.*

ChEMBL	Bilaketa mota	Data	Ordua	Artikulu kopurua	Entsegu kopurua
Bladder_cancer	Assays*	29/01/2019	11:12	10	60
Bladder_cancer	Targets*	29/01/2019	11:24	0	0
Blastoma	Assays	01/02/2019	15:01	667	19325
Blastoma	Targets	01/02/2019	15:35	23	856
Brain_cancer	Assays	29/01/2019	11:20	7	31
Brain_cancer	Targets	29/01/2019	11:25	0	0
Breast_cancer	Assays	29/01/2019	11:04	1012	9473
Cancer	Assays	28/01/2019	12:45	8691	104828
Carcinoma	Assays	01/02/2019	14:45	4988	89521
Carcinoma	Targets	01/02/2019	15:27	0	0
Cell_skin_cancer	Assays	30/01/2019	10:15	0	0
Cervical_cancer	Assays	31/01/2019	10:09	18	105
Colorectal_cancer	Assays	29/01/2019	11:27	4	5
Germ_cell_tumor	Assays	01/02/2019	15:05	0	0
Head_cancer	Assays	31/01/2019	10:30	0	0
Kidney_cancer	Assays	30/01/2019	10:18	2	2
Leukemia	Assays	29/01/2019	10:01	10815	333784
Leukemia	Targets	29/01/2019	10:48	43	94483
Lung_cancer	Assays	31/01/2019	10:01	647	5740
Lymphoma	Assays	01/02/2019	14:57	469	2670
Lymphoma	Targets	01/02/2019	15:32	22	8148
Melanoma	Assays	30/01/2019	10:16	2480	353079
Neck_cancer	Assays	31/01/2019	10:05	8	11
Pancreas_cancer	Assays	30/01/2019	10:01	2	5
Pancreatic_cancer	Assays	31/01/2019	10:27	13	27
Prostate_cancer	Assays	29/01/2019	11:41	324	3325
Sarcoma	Assays	01/02/2019	14:50	1300	36576
Sarcoma	Targets	01/02/2019	15:28	28	25190
Skin_cancer	Assays	30/01/2019	10:11	10	18
Testicular_cancer	Assays	31/01/2019	10:17	1	7
Totala	-	-	-	31584	1087269

*Assays: bilaketak entseguetan eginak, *Targets: bilaketak ituetan egina

2.E. Taula- *Eratutako datu-baseko adibide laburtua.*

1	Konposatuaren izena	f (v_{ij})	d (c0)	Muga-balioa	v_{ij}	c0= ACTIVITY (UNITS)	c1=CELL NAME	c2=ASSAY ORGANISM	c3= ASSAY TYPE	c4=TARGET MAPPING	c5=CURATED BY	Sh (PSA)	DSh (PSA)	DSh4 (Prot)	Prot. Zbk.
2	(-)-7-EPI-DEOXYNUPHARIDINE	0	1	70,00	23,00	Inhibition (%)	B16	Mus musculus	B	Protein	Autocuration	0,0085	-0,0202	0,0109	P10721
3	(-)-7-EPI-DEOXYNUPHARIDINE	0	1	70,00	20,90	Inhibition (%)	B16	Mus musculus	B	Protein	Autocuration	0,0085	-0,0202	0,0109	P10721
4	(-)-7-EPI-DEOXYNUPHARIDINE	0	1	70,00	25,30	Inhibition (%)	B16	Mus musculus	B	Protein	Autocuration	0,0085	-0,0202	0,0109	P10721
5	(-)-7-EPI-DEOXYNUPHARIDINE	0	1	70,00	5,10	Inhibition (%)	B16	Mus musculus	B	Protein	Autocuration	0,0085	-0,0202	0,0109	P10721
6	(-)-ANTOFINE	1	-1	100,00	36,00	IC50 (nM)	KB	Homo sapiens	F	Non-molecular	Autocuration	0,0142	-0,0192	0,0000	MD
7	(-)-ANTOFINE	1	-1	100,00	36,00	GI50 (nM)	KB	Homo sapiens	F	Non-molecular	Autocuration	0,0142	-0,0119	0,0000	MD
8	(-)-ANTOFINE	1	-1	100,00	25,00	IC50 (nM)	DU-145	Homo sapiens	F	Non-molecular	Autocuration	0,0142	-0,0141	0,0000	MD
9	(-)-ANTOFINE	1	-1	100,00	25,00	GI50 (nM)	DU-145	Homo sapiens	F	Non-molecular	Autocuration	0,0142	-0,0138	0,0000	MD
10	(-)-ANTOFINE	1	-1	100,00	22,00	IC50 (nM)	A549	Homo sapiens	F	Non-molecular	Autocuration	0,0142	-0,0201	0,0000	MD
11	(-)-ANTOFINE	1	-1	100,00	22,00	GI50 (nM)	A549	Homo sapiens	F	Non-molecular	Autocuration	0,0142	-0,0146	0,0000	MD
12	(-)-ANTOFINE	1	-1	100,00	25,00	IC50 (nM)	MD	Homo sapiens	F	Protein	Autocuration	0,0142	-0,0189	0,0828	Q03164
13	(-)-EPICATECHIN	0	-1	100,00	14125,40	Potency (nM)	MD	Homo sapiens	F	Non-molecular	Autocuration	0,0380	0,0105	0,0000	MD
14	(-)-EPICATECHIN	0	-1	0,10	2,58	ID50 (uM)	MD	Moloney murine leukemia virus	F	Protein	Autocuration	0,0380	-0,0012	0,0000	Q03164
15	(-)-EPICATECHIN	0	-1	0,10	2,65	ID50 (uM)	MD	Moloney murine leukemia virus	F	Protein	Autocuration	0,0380	-0,0012	0,0000	Q03164
16	(-)-GOSSYPOL HEXAMETHYL ETHER	0	-1	100,00	80000,00	IC50 (nM)	HeLa	Homo sapiens	F	Non-molecular	Autocuration	0,0281	-0,0024	0,0000	MD
17	(-)-GOSSYPOL HEXAMETHYL ETHER	0	-1	100,00	80000,00	IC50 (nM)	HeLa	Homo sapiens	F	Non-molecular	Autocuration	0,0281	-0,0024	0,0000	MD
18	(-)-GOSSYPOL HEXAMETHYL ETHER	0	-1	100,00	80000,00	IC50 (nM)	U87	Homo sapiens	B	Protein	Autocuration	0,0281	-0,0214	0,0000	P42336
19	(-)-GOSSYPOL HEXAMETHYL ETHER	0	-1	100,00	80000,00	IC50 (nM)	U87	Homo sapiens	B	Protein	Autocuration	0,0281	-0,0214	0,0000	P42336
20	(-)-GOSSYPOL HEXAMETHYL ETHER	0	-1	100,00	80000,00	IC50 (nM)	MD	Homo sapiens	F	Protein	Autocuration	0,0281	-0,0050	0,0828	Q03164

21	(-)-GOSSYPOL HEXAMETHYL ETHER	0	-1	100,00	80000,00	IC50 (nM)	MD	Homo sapiens	F	Protein	Autocuration	0,0281	-0,0050	0,0828	Q03164
22	(-)-ISOLARICIRESINOL	0	-1	100,00	22810,00	IC50 (nM)	XF498	Homo sapiens	F	Non-molecular	Autocuration	0,0351	0,0023	0,0000	MD
23	(-)-ISOLARICIRESINOL	0	-1	100,00	20190,00	IC50 (nM)	A549	Homo sapiens	F	Non-molecular	Autocuration	0,0351	0,0008	0,0000	MD
24	(-)-ISOLARICIRESINOL	0	-1	100,00	24880,00	IC50 (nM)	SK-MEL-2	Homo sapiens	F	Multiple proteins	Autocuration	0,0351	0,0082	0,0000	O00255
25	(-)-ISOLARICIRESINOL	0	-1	100,00	30000,00	IC50 (nM)	SK-OV-3	Homo sapiens	B	Protein	Autocuration	0,0351	0,0040	-0,0435	P09769
26	(-)-PINORESINOL	0	-1	100,00	13320,00	IC50 (nM)	XF498	Homo sapiens	F	Non-molecular	Autocuration	0,0291	-0,0037	0,0000	MD
27	(-)-PINORESINOL	0	-1	100,00	30000,00	IC50 (nM)	A549	Homo sapiens	F	Non-molecular	Autocuration	0,0291	-0,0052	0,0000	MD
28	(-)-PINORESINOL	0	-1	100,00	26600,00	IC50 (nM)	SK-MEL-2	Homo sapiens	F	Multiple proteins	Autocuration	0,0291	0,0022	0,0000	O00255
29	(-)-PINORESINOL	0	-1	100,00	30000,00	IC50 (nM)	SK-OV-3	Homo sapiens	B	Protein	Autocuration	0,0291	-0,0020	-0,0435	P09769
30	(-)-STEGANACIN	1	-1	14,01	0,30	ED50 (ug ml-1)	KB	Homo sapiens	F	Non-molecular	Autocuration	0,0350	0,0055	0,0000	MD
31	(-)-U-50488	0	-1	100,00	15848,90	Potency (nM)	MD	Homo sapiens	F	Non-molecular	Autocuration	0,0114	-0,0161	0,0000	MD
32	(-)-U-50488	0	-1	100,00	39810,70	Potency (nM)	MD	Homo sapiens	F	Non-molecular	Autocuration	0,0114	-0,0161	0,0000	MD
33	(-)-U-50488	0	-1	100,00	39810,70	Potency (nM)	MD	Homo sapiens	F	Non-molecular	Autocuration	0,0114	-0,0161	0,0000	MD
34	(-)-U-50488 CHLORIDE	0	-1	100,00	631,00	Potency (nM)	SW480	Homo sapiens	F	Non-molecular	Autocuration	0,0114	-0,0158	0,0000	MD
35	(-)-U-50488 CHLORIDE	0	-1	100,00	29092,90	Potency (nM)	MD	Homo sapiens	T	Non-molecular	Autocuration	0,0114	-0,0134	0,0000	MD
36	(-)-USNIC ACID	0	-1	100,00	39810,70	Potency (nM)	MD	Homo sapiens	F	Non-molecular	Autocuration	0,0406	0,0132	0,0000	MD
37	(+)-ALPHA-TOCOPHERYL SUCCINATE	0	-1	100,00	368000,00	IC50 (nM)	MCF7	Homo sapiens	F	Non-molecular	Autocuration	0,0278	-0,0057	0,0000	MD
38	(+)-ALPHA-TOCOPHERYL SUCCINATE	0	-1	100,00	368000,00	IC50 (nM)	MCF7	Homo sapiens	F	Non-molecular	Autocuration	0,0278	-0,0057	0,0000	MD
39	(+)-CC-1065	0	1	70,00	62,00	ILS (%)	P388	Mus musculus	F	Non-molecular	Intermediate	0,0600	0,0157	0,0000	MD
40	(+)-CC-1065	1	-1	73,90	0,10	OD (mg kg-1 day-1)	P388	Mus musculus	F	Non-molecular	Intermediate	0,0600	0,0299	0,0000	MD
41	(+)-CC-1065	0	-1	20,21	100,00	Relative IC50 (-)	L1210	Mus musculus	F	Non-molecular	Autocuration	0,0600	0,0286	0,0000	MD
42	(+)-CC-1065	1	-1	649,76	0,00	IC50 (ug.mL-1)	L1210	Mus musculus	F	Non-molecular	Autocuration	0,0600	0,0179	0,0000	MD

43	(+)-CC-1065	0	-1	100,00	3000,00	IC50 (nM)	L1210	Mus musculus	F	Non-molecular	Autocuration	0,0600	0,0254	0,0000	MD
44	(+)-CC-1065	1	-1	100,00	0,07	ID50 (nM)	L1210	Mus musculus	B	Homologous protein	Autocuration	0,0600	0,0297	0,0000	P49327
45	(+)-CRYPTOSPORIPSIN	0	-1	100,00	1595,88	GI50 (nM)	UACC-257	Homo sapiens	F	Non-molecular	Autocuration	0,0251	-0,0042	0,0000	MD
46	(+)-CRYPTOSPORIPSIN	0	-1	100,00	2546,83	GI50 (nM)	M19-MEL	Homo sapiens	F	Non-molecular	Autocuration	0,0251	-0,0037	0,0000	MD
47	(+)-CRYPTOSPORIPSIN	0	-1	100,00	2004,47	GI50 (nM)	Malme-3M	Homo sapiens	F	Non-molecular	Autocuration	0,0251	-0,0036	0,0000	MD
48	(+)-CRYPTOSPORIPSIN	0	-1	100,00	2238,72	GI50 (nM)	LOX IMVI	Homo sapiens	F	Non-molecular	Autocuration	0,0251	-0,0033	0,0000	MD
49	(+)-CRYPTOSPORIPSIN	0	-1	100,00	3655,95	GI50 (nM)	MOLT-4	Homo sapiens	F	Non-molecular	Autocuration	0,0251	-0,0033	0,0000	MD
50	(+)-CRYPTOSPORIPSIN	0	-1	100,00	1840,77	GI50 (nM)	SR	Homo sapiens	F	Non-molecular	Expert	0,0251	-0,0073	0,0000	MD
51	(+)-CRYPTOSPORIPSIN	0	-1	100,00	3126,08	GI50 (nM)	M14	Homo sapiens	F	Non-molecular	Autocuration	0,0251	-0,0036	0,0000	MD
52	(+)-CRYPTOSPORIPSIN	0	-1	100,00	912,01	GI50 (nM)	CCRF-CEM	Homo sapiens	F	Non-molecular	Autocuration	0,0251	-0,0035	0,0000	MD
53	(+)-CRYPTOSPORIPSIN	0	-1	100,00	1698,24	GI50 (nM)	HL-60	Homo sapiens	F	Non-molecular	Autocuration	0,0251	-0,0035	0,0000	MD
54	(+)-CRYPTOSPORIPSIN	0	-1	100,00	868,96	GI50 (nM)	K562	Homo sapiens	F	Non-molecular	Autocuration	0,0251	-0,0084	0,0000	MD
55	(+)-CRYPTOSPORIPSIN	0	-1	100,00	2280,34	GI50 (nM)	RPMI-8226	Homo sapiens	F	Unassigned	Autocuration	0,0251	-0,0076	0,0000	MD
56	(+)-CRYPTOSPORIPSIN	0	-1	100,00	2722,70	GI50 (nM)	SK-MEL-2	Homo sapiens	F	Multiple proteins	Autocuration	0,0251	-0,0037	0,0000	O00255
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
971475	ZOMEPIRAC SODIUM	0	-1	100,00	29092,90	Potency (nM)	SW480	Homo sapiens	F	Non-molecular	Autocuration	0,0238	-0,0035	0,0000	MD
971476	ZONISAMIDE	0	-1	100,00	1412,50	Potency (nM)	SW480	Homo sapiens	F	Unassigned	Autocuration	0,0319	-0,0003	0,0000	MD
971477	ZOXAZOLAMINE	0	-1	100,00	11220,20	Potency (nM)	MD	Homo sapiens	F	Non-molecular	Expert	0,0215	-0,0047	0,0000	MD
971478	ZSTK-474	0	-1	100,00	68610,00	CC50 (nM)	Huh-7	Homo sapiens	F	Non-molecular	Autocuration	0,0148	-0,0102	0,0000	MD
971479	ZUCAPSAICIN	0	-1	100,00	31622,80	Potency (nM)	MD	Homo sapiens	F	Non-molecular	Autocuration	0,0235	-0,0039	0,0000	MD
971480	ZUCLOPENTHIXOL	1	-1	78,95	15,00	RF (-)	P388	Mus musculus	F	Non-molecular	Intermediate	0,0127	-0,0066	0,0000	MD
971481	ZUCLOPENTHIXOL	0	-1	3,49	15,00	MDR ratio (-)	MCF7-DOX	Homo sapiens	B	Protein	Autocuration	0,0127	0,0054	0,0000	P10721
971482	ZUCLOPENTHIXOL	1	-1	3,49	2,60	MDR ratio (-)	MCF7-DOX	Homo sapiens	B	Protein	Autocuration	0,0127	0,0054	0,0000	P10721

3.E. Taula- *Muga baldintzen batezbestekoen adibide laburtua.*

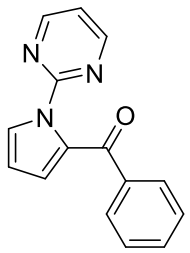
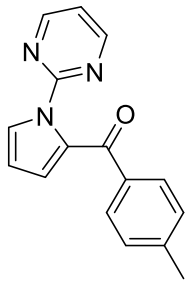
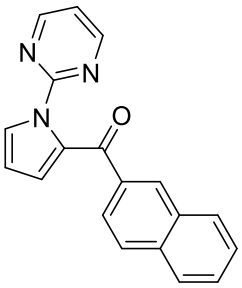
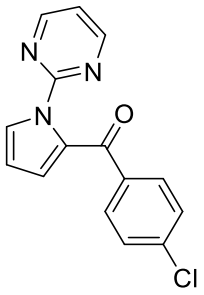
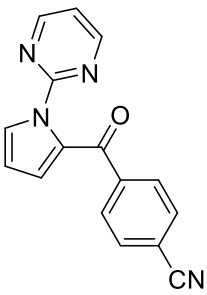
c0,c1,c2,c3,c4,c5	$n_j(c_j)$	<Sh(ALOGP)>	<Sh(GPA)>	<Sh0(Prot)>	<Sh1(Prot)>	<Sh2(Prot)>	<Sh3(Prot)>	<Sh4(Prot)>	<Sh5(Prot)>
Potency (nM)MDHomo sapiensFNon-molecularAutocuration	117117	0,158722	0,027487	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000
Potency (nM)SW480Homo sapiensFNon-molecularAutocuration	47247	0,158750	0,027238	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000
GI50 (nM)M14Homo sapiensFNon-molecularAutocuration	38349	0,158766	0,028668	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000
GI50 (nM)MOLT-4Homo sapiensFNon-molecularAutocuration	37650	0,158753	0,028430	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000
GI50 (nM)CCRF-CEMHomo sapiensFNon-molecularAutocuration	37191	0,158772	0,028632	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000
GI50 (nM)SK-MEL-2Homo sapiensFMultiple proteinsAutocuration	36856	0,158769	0,028830	6,423247	6,410625	6,415089	6,411363	6,412761	6,411170
GI50 (nM)HL-60Homo sapiensFNon-molecularAutocuration	35658	0,158765	0,028581	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000
GI50 (nM)LOX IMVI Homo sapiensFNon-molecularAutocuration	34885	0,158747	0,028366	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000
GI50 (nM)UACC-257Homo sapiensFNon-molecularAutocuration	33538	0,158812	0,029263	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000
GI50 (nM)Malme-3MHomo sapiensFNon-molecularAutocuration	33119	0,158760	0,028720	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000
GI50 (nM)SK-MEL-28Homo sapiensFProteinAutocuration	31289	0,158841	0,028904	6,028253	6,011612	6,017702	6,012736	6,015083	6,012903
GI50 (nM)UACC-62Homo sapiensFProteinAutocuration	30540	0,158876	0,029528	5,347108	5,317319	5,329042	5,319617	5,324318	5,319885
GI50 (nM)K562Homo sapiensFProteinAutocuration	26894	0,158648	0,026687	5,347108	5,317319	5,329042	5,319617	5,324318	5,319885
GI50 (nM)SK-MEL-5Homo sapiensFProteinAutocuration	25795	0,158756	0,027399	5,976351	5,957748	5,963845	5,958351	5,960113	5,958048
GI50 (nM)SRHomo sapiensFNon-molecularAutocuration	21005	0,158633	0,026534	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000
IC50 (nM)MDHomo sapiensFNon-molecularAutocuration	18986	0,158646	0,033984	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000
GI50 (nM)M19-MELHomo sapiensFNon-molecularAutocuration	13723	0,158784	0,028760	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000
Inhibition (%)MDHomo sapiensFNon-molecularAutocuration	12413	0,158644	0,031822	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000
GI50 (nM)RPMI-8226Homo	12027	0,159012	0,032708	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000

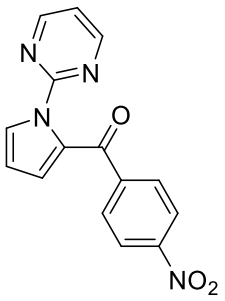
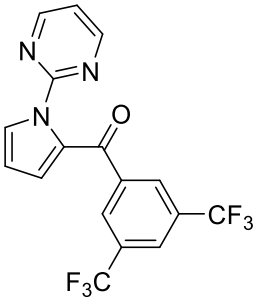
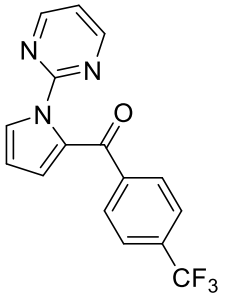
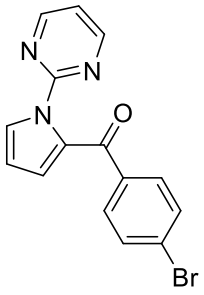
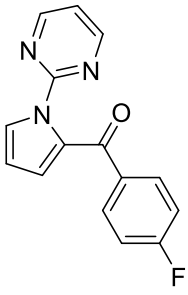
sapiensFUnassignedAutocuration									
GI50 (nM)K562Homo sapiensFNon-molecularAutocuration	10036	0,159044	0,033454	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000
GI50 (nM)RPMI-8226Homo sapiensFNon-molecularIntermediate	7752	0,158735	0,026589	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000
GI50 (nM)RPMI-8226Homo sapiensFMultiple proteinsAutocuration	6920	0,158342	0,025807	6,423247	6,410625	6,415089	6,411363	6,412761	6,411170
GI50 (nM)SK-MEL-5Homo sapiensBProteinAutocuration	6838	0,158755	0,030641	6,767054	6,757436	6,760387	6,757454	6,758221	6,756955
EC50 (nM)MDMDBMultiple proteinsAutocuration	6481	0,158726	0,024549	6,118097	6,104070	6,109155	6,105116	6,106587	6,104984
GI50 (nM)RPMI-8226Homo sapiensFNon-molecularAutocuration	5877	0,158825	0,028149	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000
CC50 (nM)Huh-7 Homo sapiensFNon-molecularAutocuration	5605	0,158621	0,024998	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000
Potency (nM)MDHomo sapiensTNon-molecularAutocuration	5604	0,158624	0,024862	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000
Ki (nM)MDHomo sapiensFNon-molecularAutocuration	5562	0,158669	0,033111	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000
Potency (nM)SW480Homo sapiensFNon-molecularIntermediate	5445	0,158863	0,030319	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000
Potency (nM)SW480Homo sapiensFNon-molecularExpert	5162	0,158835	0,029710	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000
GI50 (nM)SK-MEL-28Homo sapiensBProteinAutocuration	5024	0,158295	0,026007	7,010301	7,002101	7,004291	7,001810	7,002216	7,001166
EC50 (nM)MDMDFProteinAutocuration	4701	0,158747	0,029772	8,286521	8,281411	8,282163	8,280450	8,280417	8,279628
GI50 (nM)SRHomo sapiensFNon-molecularExpert	4522	0,159032	0,032350	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000
GI50 (nM)SRHomo sapiensFNon-molecularIntermediate	4468	0,159051	0,032950	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000
Potency (nM)MDHomo sapiensFNon-molecularExpert	3574	0,158620	0,026207	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000
Potency (nM)MDHomo sapiensFNon-molecularIntermediate	3470	0,158612	0,025457	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000
GI50 (nM)UACC-62Homo sapiensFNon-molecularIntermediate	3183	0,158351	0,026069	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000
Inhibition (%)MDHomo sapiensFNon-molecularIntermediate	2890	0,158730	0,032279	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000
GI50 (nM)UACC-62Homo sapiensFNon-molecularExpert	2888	0,158393	0,026358	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000

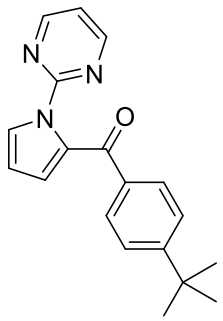
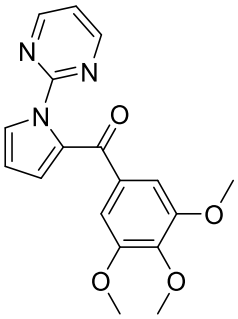
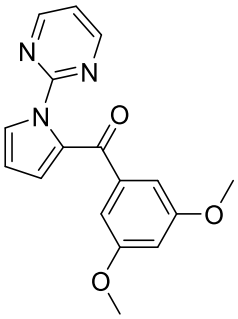
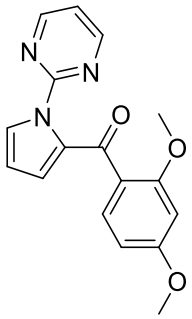
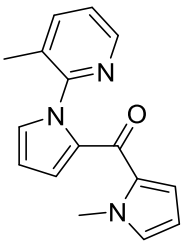
4.E. Taula- *Aktibitateen muga balioen adibide laburtua.*

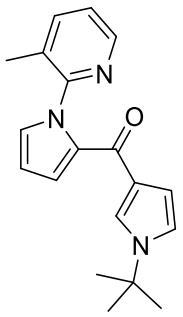
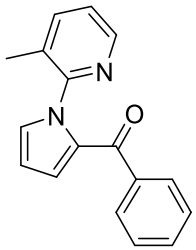
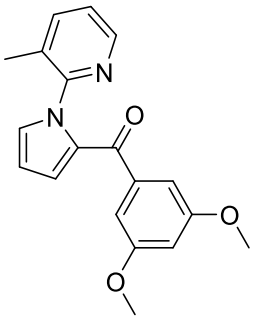
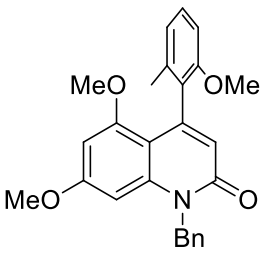
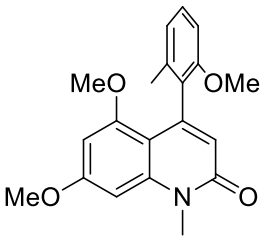
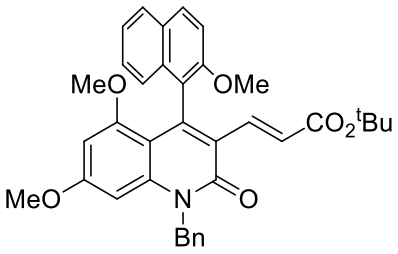
c0=ACTIVITY (UNITS)	n_j	$<v_{ij}>$	d(c0)	Muga balioa	$n_j(f(v_{ij})=1)$	$f(v_{ij})$erref
GI50 (nM)	558223	759852748,1	-1	100,0	14271	0,03
Potency (nM)	190010	19447,0	-1	100,0	5909	0,03
IC50 (nM)	65520	204104866270,0	-1	100,0	20241	0,31
Inhibition (%)	51156	69,3	1	70,0	21181	0,41
EC50 (nM)	17049	61130,9	-1	100,0	966	0,06
Ki (nM)	7880	330743,9	-1	100,0	2763	0,35
Activity (%)	7181	67,7	1	70,0	3532	0,49
CC50 (nM)	6166	55118,5	-1	100,0	25	0,00
TGI (nM)	5596	91002070756234,5	-1	100,0	19	0,00
T/C (%)	5230	144,1	1	70,0	4497	0,86
LC50 (nM)	5187	1904814570614,0	-1	100,0	30	0,01
IC50 (ug.mL-1)	4673	649,8	-1	649,8	4632	0,99
Kd (nM)	3697	160687,8	-1	100,0	746	0,20
Residual activity (%)	2943	76,0	1	70,0	1998	0,68
AC50 (nM)	2683	10585,2	-1	100,0	31	0,01
ILS (%)	2667	79,5	1	70,0	946	0,35
ED50 (ug ml-1)	1750	14,0	-1	14,0	1227	0,70
ED50 (uM)	1554	39,2	-1	0,1	100	0,06
Survivors (-)	1077	2,2	1	2,2	347	0,32
Ratio (-)	1072	96,3	1	96,3	67	0,06
T/C (-)	935	95,7	1	95,7	577	0,62
Log 1/D50 (-)	776	4,1	-1	4,1	418	0,54
Activity (-)	694	2349792,4	1	1000,0	95	0,14
ID50 (ug ml-1)	662	178,0	-1	178,0	495	0,75
GI (%)	600	49,4	1	70,0	120	0,20
GI50 (ug.mL-1)	565	4,2	-1	4,2	422	0,75
ID50 (uM)	556	59,2	-1	0,1	82	0,15
Control (%)	538	173,3	1	70,0	372	0,69
ID50 (nM)	485	215,5	-1	100,0	369	0,76
OD (mg kg-1)	476	82,0	-1	82,0	380	0,80
Weight change (%)	475	0,3	1	70,0	0	0,00
DOSE (mg.kg-1)	435	57,1	-1	57,1	313	0,72
Log 1/C (-)	391	6,9	1	6,9	244	0,62
EC50 (ug.mL-1)	361	23239,7	-1	1000,0	353	0,98
TGI (%)	347	53,4	1	70,0	142	0,41
ED50 (nM)	346	710,0	-1	100,0	230	0,66
Cures (-)	346	1,5	1	1,5	98	0,28
ID50 (M)	336	0,5	-1	0,5	311	0,93
Efficacy (%)	329	62,1	1	70,0	122	0,37
LD50 (uM)	296	329,7	-1	0,1	2	0,01
MED (mg kg-1)	285	2,7	-1	2,7	217	0,76
MST (day)	281	40,8	-1	40,8	220	0,78
ED50 (g ml-1)	269	15,0	-1	15,0	96	0,36
Growth (%)	260	1,6	1	70,0	29	0,11
Optimal dose (mg kg-1)	250	176,8	-1	176,8	160	0,64
Average weight change (g)	231	1,8	-1	1,8	199	0,86
Survival (%)	214	63,4	1	70,0	109	0,51

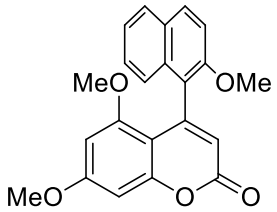
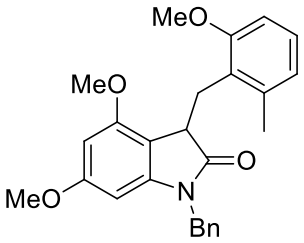
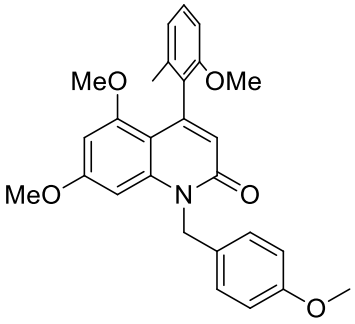
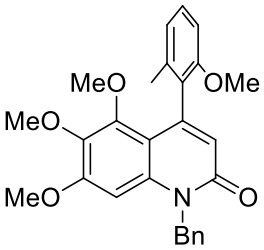
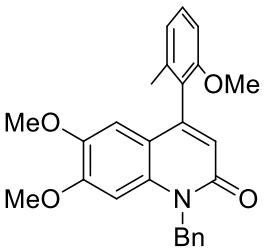
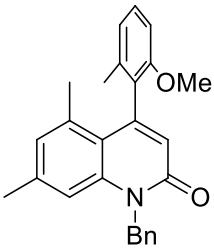
5.E. Taula- *Organometalikoak sintesian ikerkuntza taldeko laborategian sintetizatutako konposatu organiko berrien egiturak eta deskriptoreak.*

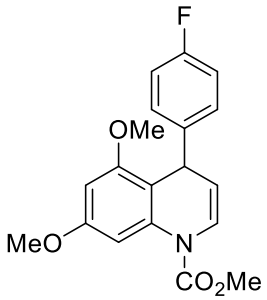
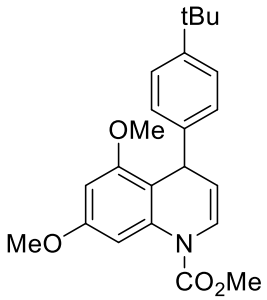
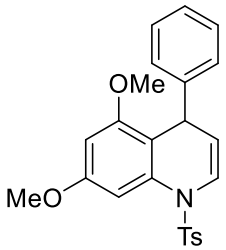
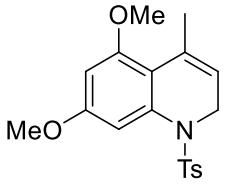
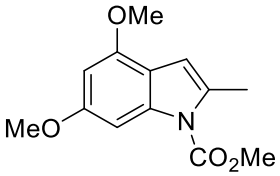
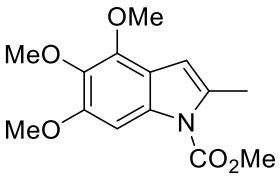
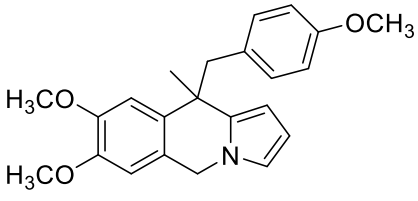
Kodea	**Ikertzailea**	**Egitura**	**GPA (Å^2)**	**Sh$_{GPA}$**
CSA 210f2	Carlos Santiago		47,78	0,02010
CSA 151f2	Carlos Santiago		47,78	0,02010
CSA 136f2	Carlos Santiago		47,78	0,02010
CSA 197f2.1	Carlos Santiago		47,78	0,02010
CSA 140f2	Carlos Santiago		71,57	0,02746

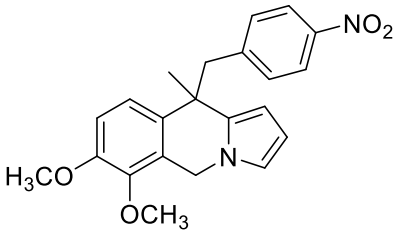
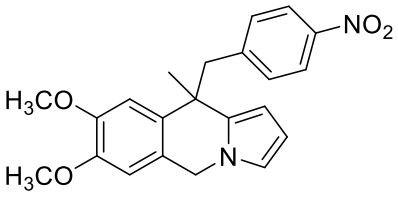
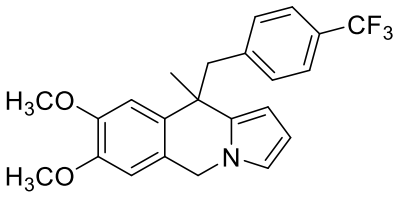
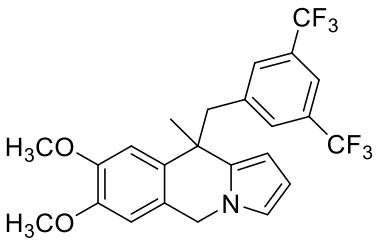
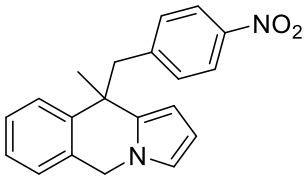
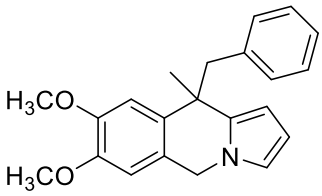
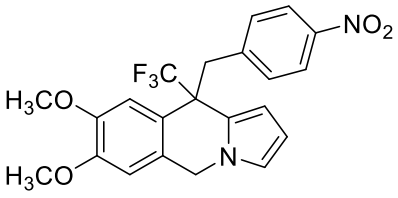
CSA 230f2	Carlos Santiago		93,6	0,03361
CSA 114f2	Carlos Santiago		47,78	0,02010
CSA 147f2	Carlos Santiago		47,78	0,02010
CSA 155f1	Carlos Santiago		47,78	0,02010
CSA 146f2	Carlos Santiago		47,78	0,02010

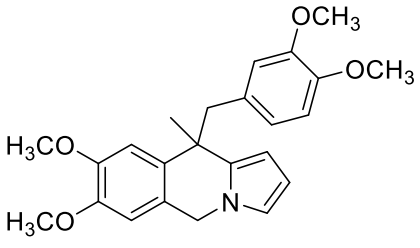
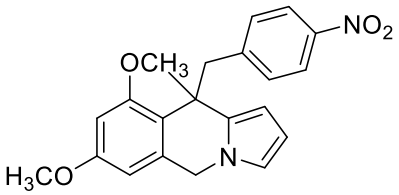
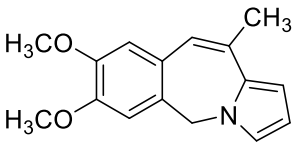
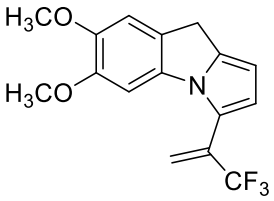
CSA 137f3	Carlos Santiago		47,78	0,02010
CSA 117f2	Carlos Santiago		75,47	0,02859
CSA 130f2	Carlos Santiago		66,24	0,02588
CSA 133f2	Carlos Santiago		66,24	0,02588
CSA 243f3	Carlos Santiago		39,82	0,01741

CSA 241f2	Carlos Santiago		39,82	0,01741
CSA246f2	Carlos Santiago		34,89	0,01568
CSA 249f2	Carlos Santiago		53,35	0,02190
MM 001	Mikel Martinez		49,69	0,02072
MM 002	Mikel Martinez		49,69	0,02072
MM 003	Mikel Martinez		75,99	0,02874

MM 004	Mikel Martinez		57,9	0,02333
MM 005	Mikel Martinez		48	0,02017
MM 006	Mikel Martinez		58,92	0,02365
MM 007	Mikel Martinez		58,92	0,02365
MM 008	Mikel Martinez		49,69	0,02072
MM 009	Mikel Martinez		31,23	0,01435

AC 606	Asier Carral		48	0,02017
AD 17	Asier Carral		48	0,02017
AC 539	Asier Carral		64,22	0,02527
AC 538	Asier Carral		64,22	0,02527
AC 520	Asier Carral		49,69	0,02072
AC 534	Asier Carral		58,92	0,02365
IB001	Iratxe Barbolla		32,62	0,01486

IB002	Iratxe Barbolla		69,21	0,02676
IB003	Iratxe Barbolla		69,21	0,02676
IB004	Iratxe Barbolla		23,39	0,01137
IB005	Iratxe Barbolla		23,39	0,01137
IB006	Iratxe Barbolla		50,75	0,02107
IB007	Iratxe Barbolla		23,39	0,01137
IB008	Iratxe Barbolla		69,21	0,02676

IB009	Iratxe Barbolla		41,85	0,01811
IB010	Iratxe Barbolla		69,21	0,02676
IB0011	Iratxe Barbolla		23,39	0,01137
IB0012	Iratxe Barbolla		23,39	0,01137