Mennyiségi szerkezet-hatás összefüggések retenciós indexek és biológiai aktivitás előrejelzésére

Ph.D. Értekezés

Készítette: Farkas Orsolya

Témavezető: Dr. Héberger Károly

tudományos tanácsadó

MTA Kémiai Kutatóközpont

2007

Köszönetnyilvánítás

Szeretnék köszönetet mondani mindazoknak, akik munkám során segítettek:

Témavezetőmnek, Dr. Héberger Károlynak türelméért, segítőkészségéért és szakmai tudásának átadásáért.

A kutatásban résztvevő külföldi partnereinknek, Igor. G. Zenkevichnek, John F. Kalivasnak és Forrest Stoutnak értékes munkájukért.

Dr. Jakus Juditnak és az MTA KK Biooxidációs csoport összes munkatársának.

A BME VBK Varga József Alapítványának anyagi támogatásáért.

Családomnak és barátaimnak.

TARTALOMJEGYZÉK

1.BEVEZETÉS

A kemometria viszonylag új, dinamikusan fejlődő kémiai tudományág, amely matematikai, statisztikai és formális logikai módszereket alkalmaz azért, hogy optimális kísérleteket tervezzen, valamint kémiai adatok elemzésével a lehető legtöbb információt nyerje ki. Ahogy a mérőberendezések száma növekszik, és felbontóképességük javul, úgy nő a keletkező adatok mennyisége is, ezek feldolgozása, értékelése hagyományos módszerekkel gyakran nem oldható meg.

A kemometriai módszerek közé tartoznak a mennyiségi szerkezet-hatás összefüggések (Quantitative structure activity relationship, magyar nyelvben is elfogadott rövidítése QSAR) is. A QSAR matematikai kapcsolatot ír le a molekulák szerkezeti jellemzői és valamilyen aktivitás adatuk (pl. toxicitás, receptorhoz való kötődés mértéke, stb.) között. A mennyiségi függvénykapcsolat segítségével könnyebben megérthető, hogy milyen tulajdonságok felelősek egy aktivitás létéért vagy nemlétéért, ill. mértékéért. A modell alapján költséges kísérletek nélkül új vegyületek különféle aktivitás értékeit számszerűen előre lehet jelezni. A vegyületek értékelése, valamint kívánt tulajdonságokkal rendelkező új vegyületek tervezése is lehetségessé válik. Az osztályozási módszerek hasonló célt szolgálnak, segítségükkel a molekulák, hatásuk alapján (pl. toxikus vagy nem toxikus) csoportosíthatók számított vagy könnyen mérhető tulajdonságai felhasználásával.

Doktori munkám során öt különböző kémiai ill. biológiai probléma megoldását végeztem el kemometriai módszerek segítségével. Célom kettős volt; egyrészt magának az adott problémának, gyakorlati feladatnak a megoldása, másrészt az irodalomban található módszerek összehasonlítása. Alkoholok, kén-, nitrogén- és oxigéntartalmú heterociklusos molekulák valamint zsírsav metil-észterek retenciós indexének előrejelzésére dolgoztam ki QSRR (Quantitative structure-retention relationship) modellt. A retenciós index segítséget nyújt olyan vegyületek szerkezetének azonosításánál, amelyek tömegspektruma nagyon hasonló (pl.

azonos molekulatömegű izomerek esetében). A QSRR jelentőségét igazolják korábbi összefoglalók [*Kaliszan, 1987; Kaliszan, 1997*], aktualitását pedig az is mutatja, hogy a Chemical Reviews és a Journal of Chromatography A-ban egyidejűleg (2007-ben) jelent meg összefoglaló cikk, mely ezzel a témával foglalkozik [*Kaliszan, 2007*, *Héberger, 2007*].

A retenciós indexek előrejelzésén kívül ezt a három vegyületcsoportot használtam a QSAR modellezés egyik fő lépésének, a változókiválasztásnak behatóbb tanulmányozására. Különböző kemometriai módszerek (parciális legkisebb négyzetek módszere, előre irányuló változóbevonás, párkorrelációs módszer, peremregresszió stb.) hatékonyságát hasonlítottam össze a változókiválasztásban. A változókiválasztást követően az előrejelző modelleket többszörös lineáris regressziós módszerrel építettem.

Munkám második felében antidepresszáns hatású gyógyszerjelölt molekulák osztályozását végeztem el a hERG kálium csatornához való kötődés alapján. Ez a tulajdonság egy nem kívánt mellékhatás: a kálium csatornához való kötődés befolyásolhatja a szívműködést, és hirtelen halálhoz vezethet, gyógyszerjelölt vegyületek esetén vizsgálata újabban elengedhetetlen. A már ismert hERG aktivitású antidepresszánsok felhasználásával épített mennyiségi szerkezet-hatás modellek segítségével lehetőség nyílik olyan vegyületek tervezésére, melyek hERG gátló aktivitása elhanyagolható. Hat különböző változókiválasztási és osztályozó módszert (lineáris diszkriminancia elemzés, genetikus algoritmus, előreirányuló változóbevonás, parciális legkisebb négyzetek módszere és ezek kombinációi) felhasználva konzisztens modelleket építettem, melynek alapján a mellékhatással rendelkező antidepresszánsok kiszűrhetők.

Dolgozatom harmadik részében flavonoidok antioxidáns aktivitását előrejelző QSAR modelleket mutatok be. A természetben több mint 4000 flavonoid található. Antioxidáns sajátságaikat okozó szerkezeti hatásokról sok irodalom áll rendelkezésre, ám a biológiai

aktivitás számszerű előrejelzésére alig található modell. Az antioxidáns aktivitás meghatározásához a parciális legkisebb négyzetek módszerének használtam.

A témák sokfélesége miatt a többváltozós kemometriai módszerek ismertetése után az egyes témákhoz közvetlenül kapcsolódó irodalmakat külön-külön tárgyalom. A retenciós indexekhez kapcsolódó irodalmi összefoglaló után közvetlenül következnek az ebben a témakörben elért eredmények, majd ezt követi külön-külön a biológiai jellegű problémák irodalmának feldolgozása és a kapott eredmények ismertetése.

Kaliszan R., Quantitative Structure-Chromatographic Retention Relationships, Wiley, New York, **1987**.

Kaliszan R., Structure and Retention in Chromatography - A Chemometric Approach, Harwood, Amsterdam, **1997**.

Kaliszan R. QSRR: Quantitative Structure-(Chromatographic) Retention Relationships *Chem. Rev.* **2007** 107 (7) 3212-3246.

Héberger K. Quantitative structure-(chromatographic) retention relationships *J. Chromatogr., A* **2007** 1158 (1-2) 273-305.

2. KEMOMETRIAI MÓDSZEREK

2.1. Mennyiségi szerkezet-hatás összefüggések

A mennyiségi szerkezet-hatás összefüggés számítások folyamatának lényegét a 2.1.1. ábra szemlélteti. A számítások menete három fő lépésre osztható: a *változókiválasztásra*, a *modell paramétereinek becslésére* és a *validálásra* [*Draper és Smith, 1981*; *Miller, 1990*]. A vegyületek szerkezetét különböző tulajdonságaikkal (leíró változókkal, deszkriptorokkal) jellemezhetjük, melyek számított vagy egyszerűen mérhető sajátságok. Ezekből a változókiválasztás során kiemeljük azokat, amelyek a legjobban kódolják az előre jelezni kívánt hatást, és segítségükkel építjük a matematikai modellt, ahol a hatás a függő változó, a deszkriptorok pedig a független változók. Fontos, hogy a modellek ne csak azokat a molekulákat jellemezzék, amelyeket a modellépítés során felhasználtunk, hanem új, ismeretlen, de szerkezetileg hasonló vegyületek tulajdonságait is. A modellek előrejelző képességét a validálási folyamat során ellenőrizzük. Erre azért is szükség van, hogy az adatok közti véletlen korreláció lehetőségét kizárjuk.

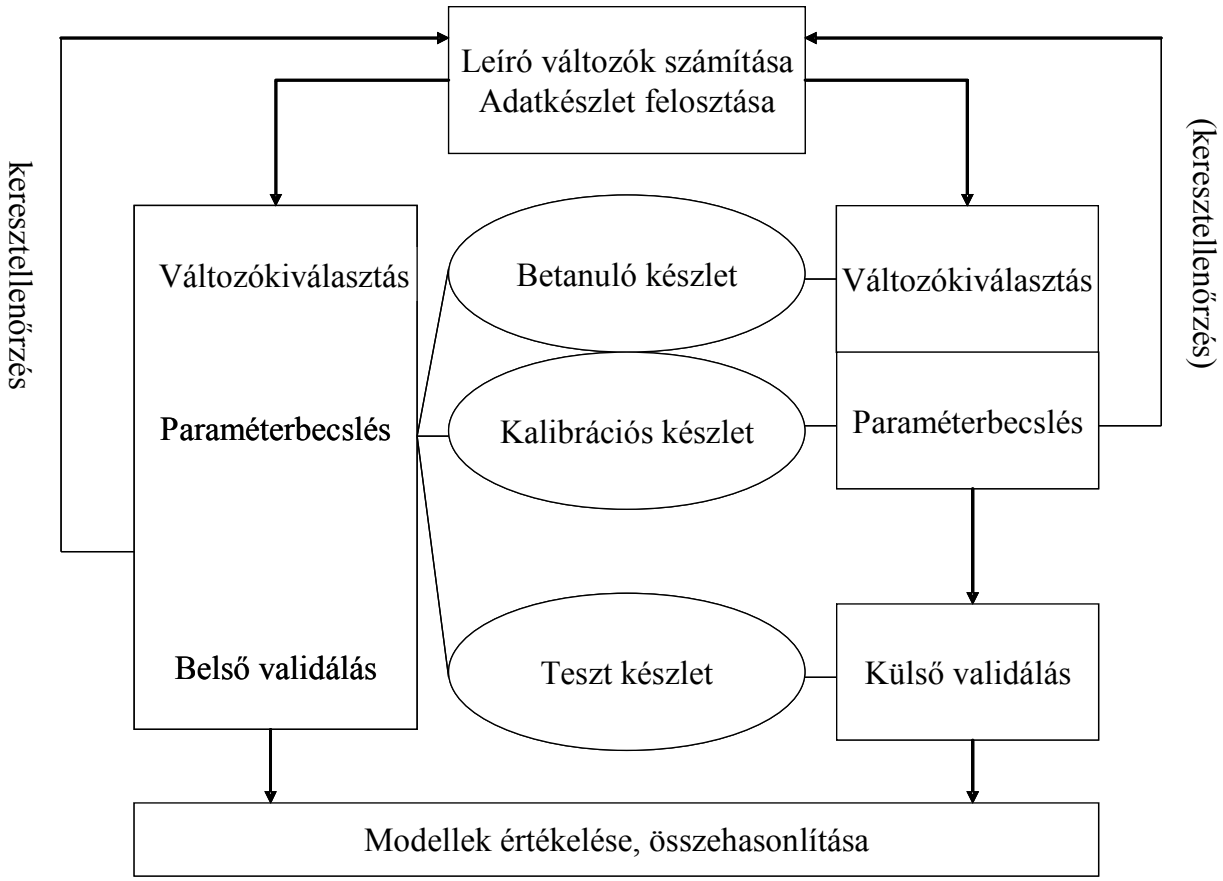
Ahhoz, hogy ezt megtehessük, az adatkészletünket már az elemzés előtt több részre kell felosztanunk, *betanuló (begyakorló), kalibrációs* ill. *teszt* készletre.

A betanuló készlet segítségével történik a független változók kiválasztása, a kalibrációs készlet felhasználásával végezzük a matematikai modell paramétereinek becslését, és a teszt készlettel ellenőrizzük, hogy modellünk valóban alkalmas-e szélesebb körű előrejelzésre.

Fontos megemlíteni, hogy a QSAR elnevezés valójában biológiai aktivitással kapcsolatos számításokra utal, mellette mennyiségi szerkezet-sajátság összefüggésekről is (quantitative structure-property relationship, QSPR) beszélünk. Retenciós indexek előrejelzése esetén a mennyiségi szerkezet-retenciós összefüggések (quantitative structure-retention

relationship, QSRR) kifejezés terjedt el. A QSAR mozaikszó azonban annyira gyakran használt, hogy használata a biológiai aktivitásokon kívül más tulajdonságok előrejelzése esetén is megszokott.

2.1.1. ábra. A QSAR modellezés folyamata – két lehetséges út



A következő néhány fejezetben részletesebben ismertetem a QSAR modellezés során dolgozatomban használt deszkriptorokat, a változókiválasztás, a paraméterbecslés és a validálás fontos szempontjait, az alkalmazott számítási módszerek jellemzőit valamint a számításhoz használt programokat.

2.1.1. Deszkriptorok

Független vagy leíró változóknak, illetve az angol megnevezés után *deszkriptoroknak* nevezik a molekulák mindazon sajátságait, melyek a molekulaszerkezethez (egyértelműen) hozzárendelhetők. A leíró változók lehetnek mért, illetve számított adatok. A módszer kialakulásának kezdetén elsőként számított deszkriptorokkal aromás vegyületek szubsztituenseinek (elektronikus, hidrofil ill. hidrofób valamint sztérikus) tulajdonságait jellemezték. Közéjük tartozik például az ún. Hammett-σ, az oktanol-víz megoszlási hányados (log*P*) és a moláris refrakció [*Livingstone, 2000*]. A QSAR modellezés fejlődésével folyamatosan nőtt az igény az egyszerűen számítható deszkriptorok iránt. A számított jellemzőknek számtalan fajtája van, tárgyalásukra a dolgozatban nincs lehetőség, így csak azokat ismertetem, amelyek számításaimban előfordulnak.

Legegyszerűbbek az ún. „nulladimenziós" szerkezeti változók, ide tartozik például a molekulatömeg, vagy atomok, funkciós csoportok, kötések száma a molekulában.

A topológiai deszkriptorok voltak az első olyan tulajdonságok, amiket bármilyen molekulából egyértelműen ki lehetett számítani, ehhez mindössze a vegyület kétdimenziós szerkezetére volt szükség. A topológiai deszkriptorok a molekulákban található elágazások mértékét, a molekulák kiterjedését (pl. hosszúkás-e inkább vagy gömbszerű) és flexibilitását jellemzik. A legismertebb és legszélesebb körben alkalmazott topológiai deszkriptorok az elágazási indexek [pl.: *Wiener, 1947; Randic, 1975*], melyek megmutatják, hogy egy molekulában egy atomhoz hány másik atom kapcsolódik. Az elágazási indexek (és általában a topológiai indexek) számításakor a hidrogénatomokat nem szokás figyelembe venni.

Az 1980-as évek végétől számítástechnika rohamos fejlődésével lehetővé vált a molekulák háromdimenziós szerkezetéhez köthető tulajdonságok számítása. A WHIM (weighted holistic invariant molecular) deszkriptorokat Todeschini és csoportja fejlesztették ki

[*Todeschini és mtsai, 1994*]. A molekulák háromdimenziós Descartes koordinátáit hatféle módon súlyozhatják (a, nem súlyozzák b, atomtömeggel c, van der Waals térfogattal d, elektronegativitással e, atomi polarizálhatósággal f, Kier-Hall elektrotopológiai indexszel súlyozzák. [*Todeschini és Consonni, 2000*]). Az így kapott koordináták számított kovariancia mátrixán (ld. Függelék) főkomponens-elemzést hajtanak végre (ld. később) és a főkomponensek mátrixából a sajátértékek segítségével származtatják a WHIM deszkriptorokat. Ezen változóknak a fizikai értelmezése nehézkes, de ettől függetlenül igen sok esetben sikerrel alkalmazták őket QSAR modellezésben; például aromás szénhidrogének toxicitásvizsgálatakor [*DiMarzio és mtsai, 2001*], fotodinamikus terápiában használt fotoszenzibilizátorok tumorellenes aktivitásának előrejelzésére [*Vanyúr és mtsai, 2003*], poliklórozott bifenilek retenciós indexének megállapítására [*Gramatica és mtsai, 2000*] és királis szulfoxidok kromatográfiás elválasztásánál [*Montanari és mtsai, 1998*].

A szintén Todeschini csoportja által kifejlesztett GETAWAY (geometry, topology and atom-weights assembly) deszkriptorok [*Consonni és mtsai, 2002a, 2002b*] használata még nem terjedt el, pedig rendszerint jobb eredményeket szolgáltatnak, mint aWHIM deszkriptorok. Ezeket a molekula Descartes-koordinátáinak kalapmátrixából (ld. Függelék) számítják. A diagonális elemek mutatják meg, hogy az egyes atomok mennyire befolyásolják a molekula alakját.

További deszkriptorok nyerhetők kvantumkémiai számításokkal. Az elektrosztatikus deszkriptorok a molekulákban található parciális töltésekről nyújtanak képet, a Kier-Hall [*Kier és Hall, 1981*] és a Galvez topológiai töltés indexekből [*Galvez és mtsai, 1994*] pedig a vegyértékelektronokról nyerhetünk információt. A geometriai deszkriptorok [*Labute, 2000*] az atomok térbeli elhelyezkedését, átlapolásaikat mutatják. A fent említett deszkriptorok mindegyike súlyozható a WHIM deszkriptorokhoz hasonló módon.

2.1.2. Az adatkészlet felosztása

A három részre történő felosztás célja, hogy a modellünk statisztikai szempontból szignifikáns és köznapi értelemben robusztus legyen. A validálás szempontjából nagyon fontos a teszt készlet elkülönítése. [*Miller, 1990; Golbraikh és Tropsha, 2002; Trophsa és mtsai, 2000*]. A szakirodalomban elterjedt az is, hogy a betanuló és a kalibrációs készletet nem különítjük el (azaz két részre osztunk). Ha kevés adatunk van, akkor a teljesen független teszt készlet kiválasztása nem valósítható meg minden esetben, a validálás fejezetnél erre a problémára részletesebben kitérek. Fontos, hogy az egyes készletek megfelelően reprezentálják a teljes adatkészletet. Bevett gyakorlatnak számít, hogy a molekulák legalább 20-30%-át „teszik félre" a validáláshoz.

2.1.3. Változókiválasztás

A rendelkezésre álló deszkriptorok közül az adott tulajdonságot kódolni képes független változók kiválasztása kulcsfontosságú lépés a QSAR során. A molekulaszerkezet leírására több ezer deszkriptor létezik, az összes lehetséges kombináció tesztelése a kombinatorikus robbanás (a számítási idő a deszkriptorok számának növekedésével exponenciálisan nő) miatt nyilvánvalóan nem lehetséges. Ugyanaz az algoritmus pedig más deszkriptorokat választhat ki legjobbnak, ha kissé eltérő kiválasztási kritériumokat alkalmazunk, és ily módon számtalan modellt kaphatunk.

Első lépésben, változók számának csökkentéseként, egyszerű módszerekkel keressük azokat a deszkriptorokat, amik redundáns információt tartalmaznak vagy a varianciájuk nagyon kicsi (konstans vagy csaknem konstans változók). Redundáns változók akkor fordulnak elő, ha két vagy több változó között nagy a korreláció mértéke (több változó korrelációja esetén multikollinearitásnak nevezik) vagy ha a független változók száma nagyobb, mint a mért

adatok mennyisége (ha a mért adatok száma n, a változók száma p, és $p > n$, akkor legalább p-n változó felírható a többi változó lineáris kombinációjaként). Ez több problémát okozhat a modellépítés során. Több módszernél (pl. többszörös lineáris regresszió (MLR), főkomponens-elemzés (PCA)) mátrixokat szorzunk és invertálunk. A korrelált változók jelenléte szingularitást (nulla vagy nullához nagyon közeli sajátértéket) okoz a mátrixban és a szinguláris mátrix nem invertálható. Korrelált változók jelenléte ezért a modell instabilitását okozhatja (ha kis változás történik a mért adatokban, a modell regressziós koefficiensei nagymértékben megváltoznak).

Más okok miatt sem érdemes a deszkriptorok közül túl sokat beépíteni a modellbe. Sok változó nagymértékű független korrelációt eredményezhet [*Topliss és Edwards, 1979*]. Ha túl sok független változónk van, lehet, hogy nagyon jó leíró-képességű modellt kapunk, de az valószínűleg "túlillesztett" lesz, azaz nem lesz alkalmas olyan molekulák tulajdonságainak előrejelzésére, amelyeket nem tartalmazott az eredeti modell. A deszkriptorok egy része ilyen esetben a mérési zajt, vagy a jelenséghez nem kapcsolható információt kódolja. Kiválasztásuk történhet egy, illetve több lépésben, de általában előszelekciót kell alkalmazni. Egyes módszerek nem képesek sok változót egyszerre kezelni, nem tudják kiválasztani a „jó" változókat, vagy a számítás ideje növekszik drasztikusan a deszkriptorok számának növekedésével [*Livingstone és Salt, 2005*].

Fontos megjegyezni, hogy a változókiválasztás folyamata gyakran egybeesik a modellépítés folyamatával, egyes módszerek a kettőt együtt hajtják végre. Vannak olyan számítási módszerek, amik csak változókiválasztásra alkalmasak, vannak, amik modellépítésre és vannak, amik mind a kettőre. Ezekről részletesebben szólok a 2.2 fejezetben.

2.1.4. Modellépítés

A modellépítés során konkrét matematikai összefüggéshez jutunk a deszkriptorok és a függő változó között. A modellek jóságát jellemző legfontosabb statisztikai paraméterek a *korrelációs koefficiens négyzete* (R^2), a *reziduális szórás* (SD) és a *teljes modell F tesztje*. A korrelációs koefficienst lineáris esetben használják, számítása:

$$R^2 = 1 - \frac{\Sigma (\hat{y}_i - \bar{y})^2}{\Sigma (y_i - \bar{y})^2} \qquad (1)$$

ahol y a független változó valódi értéke, $\bar{y}$ az egyes y-ok átlagértéke, $\hat{y}$ a független változó becsült értéke.

A reziduális szórást (residual error, standard deviation v. error) az alábbi egyenlet szerint fejezhetjük ki:

$$SD = \frac{\Sigma (y_i - \hat{y}_i)^2}{n - p - 1} \qquad (2)$$

ahol n a függő- és a p a független változók száma a modellben.

Az ún. teljes F-próba annak a feltételezésnek felel meg, hogy legalább egy paraméter a konstans tagon kívül az egyenletben nem nulla. A számított F értéket összehasonlítjuk az F táblázatban az adott szabadsági fokhoz tartozó kritikus értékkel. Ha számított érték nagyobb a kritikusnál, akkor a modell jobb leírást az adott valószínűségi szinten ahhoz képest, mintha csak a konstans tagot tartalmazná. Minél jobb a modell, annál nagyobb szám az F statisztika értéke. Az F statisztikát az alábbi képlettel számíthatjuk ki:

$$F = \frac{(SD_1 - SD_p)/p}{SD_p/(n - p - 1)} \qquad (3)$$

ahol SD_1 a csak konstans tagot, SD_p pedig p db független változót tartalmazó modellből számított SD.

Ha a becsült független változók értékét kivonjuk a mértből, megkapjuk az ún. *reziduumok* értékét. Ezeket az értékeket ábrázolhatjuk bármelyik függő változó vagy a becsült független változó függvényében. A reziduumok ábrájának alakjából fontos információt nyerhetünk a modell jóságáról, pl. vannak-e kiugró értékek vagy van-e trend, görbület az egymás utáni adatokban (az egymást követő értékek közti eltérések kisebbek vagy nagyobbak lesznek) stb.

Fontos hangsúlyozni, hogy az ebben a fejezetben tárgyalt statisztikai paraméterek kizárólag a modell leíró képességét jellemzik, és nem adnak információt annak előrejelző-képességéről. Szintén meg kell említeni a *szabadsági fok* fogalmát, amit úgy kapunk meg, hogy az objektumok számából kivonjuk a modellben levő paraméterek számát (p+1)-et. Az R^2, az SD és az F mind függenek a szabadsági foktól.

2.1.5. A modell validálása

A modellek validálására több, széles körben elfogadott módszer létezik. Megtehetjük, hogy a vegyületek egy részét a számítások megkezdése előtt félretesszük, és egyáltalán nem használjuk fel a változókiválasztás és paraméterbecslés során. A teszt készlet és a betanuló készlet elemeinek hasonlónak kell lenniük egymáshoz, és reprezentálniuk kell az egész adatkészletet. A validálásnak ezt a formáját *külső validálásnak* (external validation) nevezzük. A külső validálás jóságának mértékét jelzi az R^2_{teszt}:

$$R^2_{teszt} = 1 - \frac{\sum_{i=1}^{teszt} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{teszt} (y_i - \bar{y})^2} \qquad (4)$$

A R^2_{teszt} annál jobb, minél jobban közelít értéke egyhez. Fontos azonban megadni az előrejelzéshez tartozó reziduális szórás mértékét is vagy a PRESS (predicted error sum of squares) értéket:

$$SD_{teszt} = \frac{\sum_{i=1}^{teszt} (y_i - \hat{y}_i)^2}{n - p - 1} \quad (5)$$

$$PRESS = \sum_{i=1}^{teszt} (y_i - \hat{y}_i)^2 \quad (6)$$

Elterjedt ún. belső validálási módszer a *keresztellenőrzés* (cross-validation). Ennél az eljárásnál nem tesszük félre a számítás kezdete előtt molekulák egy részét az ellenőrzéshez (tehát nincs független teszt készlet). A változókiválasztást és a paraméterbecslést a teljes adatkészlettel végezzük el, majd a kihagyjuk a modellből a teszt készlet molekuláinak függő változó értékeit, és a többi felhasználásával jelezzük azokat előre. (Ezt a folyamatot az adatkészlet nagyságától függően többször megismételhetjük egymás után más-más felosztással, és az eredményt átlagolhatjuk.) Az így kapott független változó értékekre a szokásos módon SD-t számítottam, és SD_{CV}-nek neveztem. A külső validáláshoz hasonló módon számított paramétert pedig R^2_{CV}-nek neveztem.

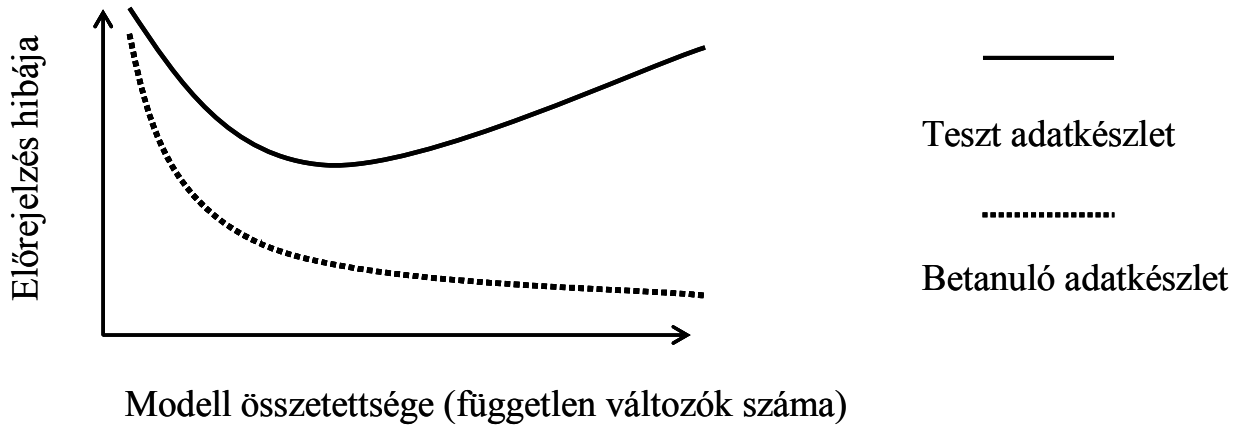
A rendelkezésre álló irodalom nem egységes abban, hogy a külső validálást vagy a belsőt célszerűbb-e használni, mindkét módszer elfogadottnak számít, és széles körben alkalmazzák. Tropsha szerint a belső validálás során kapott statisztikai paraméterek jó értéke szükséges, de nem elégséges feltétele a megbízható modellnek. Ezt az állítását valós adatokon elvégzett számításokkal igazolja [*Golbraikh és Tropsha, 2002*]. Hawkins ellenben szimulációs számításai alapján azzal érvel, hogy minél nagyobb a vegyületek száma egy készletben, annál megbízhatóbb a modell és belső validálás esetén több vegyületet használhatunk fel (tulajdonképpen az összeset) a modellépítésnél [*Hawkins, 2003*]. Így azonban megnő a túlillesztés veszélye. Baumann szerint is a külső validálás ad igazán megbízható képet arról, hogy mennyire prediktív a modell. A belső validálás módja (pl. az adatok hány %-át hagyjuk ki, hányszor ismételjük meg) nagyban függ a független változók számától, az objektumok

számától, ennek a kettőnek az arányától, sőt, a modellezési módszertől is [*Baumann, 2003; Baumann,2004*].

A hatékonyság növelésének érdekében úgy is végezhetjük a modellépítést, hogy a változókiválasztáskor és/vagy a paraméterbecsléskor keresztellenőrzéssel választjuk ki a legjobb deszkriptorkombinációt, majd a független teszt készlet segítségével külső validálással igazoljuk a modell jóságát (ld.2.1.1. ábra jobb oldali elágazása).

A 2.1.5.1 ábra szemlélteti a hiba nagyságának változását a modellben található változók számának függvényében. Látszik, hogy a független változók számának növelésével egyre jobb leíró modellt kapunk, az előrejelző képesség viszont bizonyos számú változó bevonása után romlani kezd, azaz a modell "túlillesztett" lesz. (Az ábra egy idealizált esetet mutat be, a valóságban cikk-cakkos és több minimum is lehet rajta. Egy ilyen példa található a 6.3.1. ábrán)

2.1.5.1. ábra



2.2. Számítási módszerek

2.2.1. A többszörös lineáris regresszió és a független változók számának csökkentése

A függő változó (**y**) és a *p* db független változó ($\mathbf{x_1}$, $\mathbf{x_2}$ … $\mathbf{x_p}$) közti kapcsolatot az alábbi módon írhatjuk fel:

$$\mathbf{y} = b_0 + b_1\mathbf{x_1} + b_2\mathbf{x_2} + \ldots + b_p\mathbf{x_p} \qquad \text{azaz,} \tag{7}$$

$$\mathbf{y} = \boldsymbol{b}\mathbf{X} \tag{8}$$

ahol **X** egy $n \times p$ méretű, n mért adatot és p független változót tartalmazó adatmátrix (ld. részletesebbn Függelék), **y** egy $n \times 1$ méretű vektor**,** ami a függő változót tartalmazza, ***b*** pedig $p \times 1$ méretű**,** az egyes függő változók regressziós koefficienseit tartalmazó vektor.

Célunk a ***b*** regressziós koefficiensek becslése, ami klasszikus módon a legkisebb négyzetek (LKN) módszer szerint történik. A regressziós koefficiensek az alábbi képlet segítségével becsülhetők, ahol az $\mathbf{X}^T$ az **X** transzponáltját jelöli:

$$\boldsymbol{b}_{\mathrm{LKN}} = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\mathbf{y} \tag{9}$$

A kalibrációs készlet alapján becsüljük a függvény együtthatóit. Ezeket úgy választjuk ki, hogy az eltérés a mért adat és a becsült között minimális legyen.

Ha minden változót bevonunk a modellbe, akkor a 2.1.3. fejezetben tárgyalt túlillesztés veszélye forog fenn. Ennek elkerülésének érdekében végezzük a változókiválasztást. A dimenziócsökkentést alapvetően három különböző úton érhetjük el:

- Peremregresszióval vagy a Lasso módszerrel
- Ortogonális változók használatával, azaz eredeti, összefüggő, független változók helyett azok lineáris kombinációit (látens változók) használjuk
- Megpróbáljuk a legjobb alrendszert kiválasztani a deszkriptorok közül

2.2.2 A peremregresszió és a Lasso

A *peremregresszió* (ridge regression, RR) [*Hoerl és Kennard, 1970a, 1970b*] lényege, hogy a regressziós koefficiensek becslését a következőképp hajtjuk végre:

$$\boldsymbol{b}_{RR} = (\mathbf{X}^T\mathbf{X}+k\mathbf{I})^{-1}\,\mathbf{X}^T\mathbf{y} \qquad (10)$$

ahol $0 < k < 1$, I pedig egységmátrix.

A k paraméter hozzáadásával kiküszöbölhető a numerikus instabilitás és ennek köszönhetően kisebb lesz a modell varianciája, de torzított lesz a becslés. A megfelelő k érték megállapítása a peremregresszió nyomának ábrájáról (***b*** értékek ábrázolása k függvényében) történhet. A számításokat 0 és 1 közötti k értékekkel le kell futtatni, majd az x tengelyen a k-, az y tengelyen pedig a különböző deszkriptorok regressziós koefficiens értékeit kell ábrázolni. Az a k érték a megfelelő ahol a paramétergörbék megtörnek.

A peremregresszió nyoma változókiválasztásra is használható, bár meglehetősen szubjektív módszer. Azok a deszkriptorok tekinthetők megfelelőnek, amiknek regressziós koefficiensei nullához tartanak.

A *Lasso módszer* a peremregresszióval rokon módszer [*Tibshirani, 1996*]. Abban különböznek, hogy peremregresszió esetén a 11-es, míg a Lassonál a 12-es normálegyenletet kell optimalizálni, azaz az illesztés hibáját minimalizálni [*Tikhonov, 1963*]:

$$\min \left\| Xb - y \right\|_2^2 + k\left\| Ib \right\|_2 \qquad (11)$$

$$\min \left\| Xb - y \right\|_2^2 + k\left\| Lb \right\|_1 \qquad (12)$$

ahol $\| \|_2$ az euklideszi távolságot, vagy az ún. L_2 normát jelent, $\| \|_1$ pedig abszolút-értéket, vagy L_1 normát.

Látjuk, hogy a 11-es illetve 12-es egyenlettel kifejezett hiba két részre bontható fel. Az egyenletek bal oldali tagja, a torzítás fejezi ki az eltérést a függő változó becsült és valódi értéke közt, azaz a 2-es egyenletben leírt reziduális szórásról van szó. Az L_2 norma, az egyenlet jobb oldalán található variancia tag a becsült eredmények reprodukálhatóságáról ad információt [*Forrester és Kalivas, 2004*]. Általában a torzítás csökkentésével a variancia nő és fordítva, így a számítások során a kettő közti optimalizálásról beszélünk.

2.2.3 Főkomponens elemzés és a részleges legkisebb négyzetek módszere

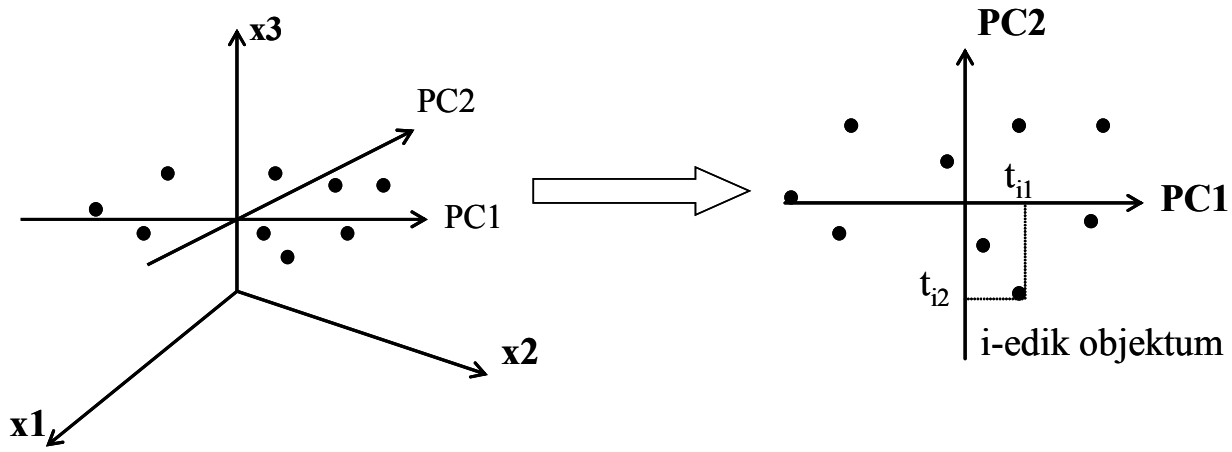
A *főkomponens-elemzés* (principal component analysis, PCA) egy adatstrukúra-elemző módszer, melynek lényege, hogy az eredetileg korrelált független változók lineáris kombinációiként új korrelálatlan független változókat hozunk létre. Az új független változókat nevezzük főkomponenseknek, ezek egymásra merőlegesek. Az első főkomponenst úgy hozzuk létre, hogy az adatokban levő variancia lehető legnagyobb részét magyarázza, a második az elsőre merőleges és a második legtöbb varianciát magyarázza és így tovább. Az eredeti változók által kifeszített teret a főkomponensek segítségével alacsonyabb dimenziójú térbe vetítjük, remélve, hogy ily módon csoportosulásokat, kiugró értékeket fedezünk fel [*Wold és mtsai, 1987; Sokváltozós adatelemzés, 2001a; Esbensen, 2002*].

Az eredeti adatmátrixot felírhatjuk két mátrix szorzataként:

$$\mathbf{X}=\mathbf{T}\mathbf{P}^{\mathbf{T}} \qquad (13)$$

ahol **X** az adatmátrix, ami *n* sorból (esetek, objektumok) és *p* oszlopból (változók, tulajdonságok) áll. **T** az ún. főkomponens mátrix, ami az **X** mátrix vektorainak vetítése egy kisebb dimenziós altérbe a $\mathbf{P}^T$ vetítési mátrix segítségével, és ez megadja az objektumok koordinátáit a **T** hipersíkban. A **T** mátrix oszlopai a főkomponens vektorok (scores), a **P** mátrix sorai pedig a főkomponens-együttható vektorok (loadings). (Az irodalmak egy része a score és a loading kifejezéseket felcserélve használja.) A főkomponensek ábráján az objektumok találhatók és különféle csoportokat, mintázatokat fedezhetünk fel. A főkomponens-együtthatók ábráján pedig a független változók jelennek meg. Egymáshoz viszonyított helyzetükből lehet hasonlóságukra következtetni és ily módon változókiválasztásra is lehet használni ezt a diagramot.

2.2.3.1. ábra Független változók és főkomponensek kapcsolata



A PCA-val azonos elven alapul a PCR, azaz a főkomponens-regresszió módszere. A PCA-nál csak független változókat használunk, a PCR esetében pedig az deszkriptorok főkomponens-elemzése után az új változókat tekintjük független változóknak, és többváltozós regressziót végzünk köztük és a függő változó közt. A PCR előnye az MLR-lel szemben, hogy

kevesebb független változóra van szükség. Ez azonban gyakran csak látszólagos előny, mert az új változók előállításához az összes régire is szükség van.

A *részleges legkisebb négyzetek módszere* (partial least squares regression, PLS) lineáris regressziós módszer [*Wold, 1995; Sokváltozós adatelemzés, 2001b*]. Megközelítésében a PCR-hoz hasonló, de fontos különbség köztük, hogy a PCR-nél csak a független változókkal dolgozunk (**X**), míg a PLS-nél a függő változókban lévő információt is használjuk (nem csak egy, hanem egyszerre több függő változónk (**Y**) is lehet). A kiindulási változók helyett ún. látens vagy rejtett változókat használunk. A számítás során arra törekszünk a két adatmátrix (függő és független változók) rejtett változói között a korreláció maximális legyen.

A PLS módszert kézzelfoghatóbbá teszi, ha az **X** és **Y** adatmátrixot a PCA-hoz hasonlóan külön-külön két mátrix szorzatára bontjuk [*Geladi, 1986*]:

$$\mathbf{X}_{N,M} = \mathbf{T}_{N,L}\,\mathbf{P}_{L,M}^{\,T} \qquad (14)$$

$$\mathbf{Y}_{N,K} = \mathbf{U}_{N,L}\mathbf{V}_{L,K}^{\,T} \qquad (15)$$

a K a függő változók száma, **T** és **U** a független ill. függő változók rejtett változók mátrixa (scores), **P** és **V** pedig a hozzájuk tartozó együttható mátrixok (loadings). A két adatmátrix azon rejtett változóit keressük, amelyek közt a korreláció a lehető legnagyobb, feltételezve, hogyközöttük lineáris kapcsolat áll fenn. Ezt belső összefüggésnek nevezzük:

$$\mathbf{u}_{l\,N,1} = q_{ll}\,\mathbf{t}_{l\,N,1} \qquad (16)$$

ahol $\mathbf{u}_l$ **Y** l-edik rejtett változója, $\mathbf{t}_l$ **X** l-edik rejtett változója, q_{ll} pedig a becsülendő l-edik regressziós együttható.

Elgondolkodtató kérdés, hogyan értelmezzük a szabadsági fokot a PLS és a PCA esetén, hiszen hiába lesz kevesebb paraméterünk a rejtett változók bevezetésével, azok legtöbb esetben a többi eredeti független változóban kódolt információt is hordozzák. Az irodalomban többféle módszer terjedt el a fenti kérdés gyakorlati megvalósításakor, melyeket jól összesít Voet munkája [*Voet, 1998*]. Tradicionálisan a PLS (vagy PCA esetén a főkomponensek) komponensek számához (A) hozzáadnak 1-et, ám ezzel általában alulbecslik a valódi szabadsági fokot. Reálisabbnak tűnhet a DF (degree of freedom, szabadsági fok) = 2A+1 megközelítés. Maga Voet egy ún. pszeudo szabadsági fokot (PDF) definiál, melyet kétféleképpen kapott (újrahelyettesítéssel és keresztellenőrzéssel) SD ill. SD_{CV} értékből számít ki, *n* pedig a független változók száma:

$$\text{PDF} = n\left(1 - \sqrt{SD / SD_{CV}}\right) \qquad (17$$

2.2.4. Alrendszer választása

A függő változóra legjobb becslést adó alrendszer kiválasztása több úton történhet [*Livingstone és Salt, 2005*]. A *legjobb alrendszer* (best subset selection, BSS) módszer lényege, hogy a modellbe bekerülő változók számát önkényesen fixálhatjuk és az ezen belül az összes létező deszkriptorkombinációval elvégezzük az MLR számítást. Ez az eljárás rendkívül időigényes és alkalmazhatósága a kombinatorikus robbanás miatt korlátozott. Az illesztés jóságának jellemzésére különféle statisztikai paramétereket használhatunk, pl. R^2, R^2_{teszt}, SD.

Az *előreirányuló változóbevonás* során a modellbe először csak a konstans tagot (tengelymetszetet) építjük be a modellbe, majd az adatsorban legelöl található deszkriptort, és részleges F-próbával összehasonlítjuk az így kapott két modellt. Ha van szignifikáns különbség a két modell közt, a másodikként bevett változót is megtartjuk, ha nincs, akkor eltávolítjuk. A

számítás kezdetekor meg kell adni a szignifikanciaszintet, aminek értéke természetesen mindig az adott problémától függ, de igen gyakran használjuk az 5%-ot. A második független változó vizsgálata után következik a harmadik változó és a többi. A módszer hátránya, hogy a változók eredeti sorrendje nagyban meghatározza, hogy mely deszkriptorok kerülnek be a végső modellbe, és a számítás könnyen egy lokális minimumba futhat, azaz nem feltétlenül találjuk meg a legjobb modellt. Ennek ellenére közkedvelt módszer, mivel használata gyors és egyszerű, és sok esetben jól használható eredményhez vezet.

A részleges F-próba számításánál a következő egyenletet használjuk a szignifikáns független változó „beléptetéséhez" (F to enter):

$$F_{\text{to enter}} = \frac{SD_p - SD_{p+1}}{(SD_{p+1})/(n - p - 1)} \qquad (18)$$

ahol SD a reziduális szórás, *p* a független változók száma, n az objektumok száma.

2.2.5. Párkorrelációs módszer

A párkorrelációs módszer (Pair-Correlation Method, PCM) [*Rajkó és Héberger, 2001; Héberger és Rajkó, 2002*] a deszkriptorokat rangsorolja, ily módon változókiválasztásra használható. Lényegét egyszerűen egy függő változó (y) és két deszkriptor (X1, X2) esetében lehet szemléltetni, de a módszer elvben korlátlan számú független változó esetén is alkalmazható. A számítás az alábbi lépések szerint történik:

1. Kiválasztunk egy adatpárt: $y_i > y_j$ $\quad i,j = 1, 2 \ldots m$ és $y_i \neq y_j$

2. Kiszámoljuk az alábbi különbségeket:

$\Delta X1 = X1_i - X1_j$ és $\Delta X2 = X2_i - X2_j$

Négyféle eredményt kaphatunk. Ha ΔX1 és ΔX2 is pozitív szám, akkor A-val, ha ΔX1 pozitív, ΔX2 negatív, akkor B-vel, ha ΔX1 negatív és ΔX2 pozitív, akkor C-vel, ha ΔX1 és ΔX2 is negatív, akkor D-vel jelöljük az adott esetet.

3. A különbségképzést minden adatpárra elvégezzük. Összeszámoljuk az A, B, C, D eseteket és kontingenciatáblázatban összegezzük. Az egyes esetek gyakoriságát k_A, k_B, k_C ill. k_D-vel jelöljük.

2.2.5.1. táblázat

	ΔX1 < 0	**ΔX1 > 0**
ΔX2 < 0	k_D	k_B
ΔX2 > 0	k_C	k_A

Ha k_B szignifikánsan nagyobb k_C-nél, akkor X1 és y közt szorosabb az összefüggés, mint X2 és y között, ha pedig k_C a nagyobb, akkor fordítva. Ennek megállapításához különféle próbákat végezhetünk, pl. Feltételes F-tesztet, χ^2-tesztet, McNemar tesztet [*Conover, 1980*].

4. Kettőnél több független változó esetén a változókat páronként kell összehasonlítani. Az összevetésnek egy változóra nézve három kimenetele lehet; „győztes” vagy „vesztes” lesz, illetve a minősítés eredménye „döntetlen”. Egyszerű rendezésnél (simple ordering) a nyerések száma döntő a rangsorban, különbségrendezésnél (difference ordering) a nyerések számából kivonjuk a vereségek számát, súlyozásos rendezésnél (significance ordering) pedig figyelembe vesszük mekkora mértékű volt a győzelem az egyes összehasonlításokkor, azaz valószínűségekkel súlyozzuk a győzelmek és vereségek különbségét.

A módszer előnye, hogy eloszlásfüggetlen, és akkor is képes különbséget tenni két független változó közt, ha a paraméteres (eloszlásfüggő) eljárások nem. Hátránya viszont, hogy előrejelzésre nem alkalmazható, és nem érzékeny az esetleges kiugró függő változó értékekre.

2.2.6. Osztályozás

Az osztályozás az ún. felügyelt csoportosítási eljárások közé tartozik, azaz a különféle osztályoknak előre ismerteknek kell lenniük. A lineáris diszkriminancia elemzés (linear discriminant analysis, LDA) kiszámítja a különböző osztályokhoz tartozás valószínűségét és olyan diszkrimináló teret (lineárisan korrelálatlan új tengelyek által kifeszített tér) keres, ahol az osztályok a lehető legjobban elkülönülnek egymástól. Majd megbecsli, hogy egy új objektum melyik osztályba tartozik a legnagyobb valószínűséggel. A molekulák szerkezetét a mennyiségi szerkezet-hatás összefüggés számításokhoz hasonlóan különféle deszkriptorok segítségével kódoljuk.

Az LDA a leggyakrabban használt eljárások közé tartozik. Ez a módszer tulajdonképpen a főkomponens-elemzéshez hasonlít, mivel az eredeti független változók helyett azok lineáris kombinációit használja fel arra, hogy létrehozza a diszkrimináló teret. Az LDA viszont a főkomponens-elemzéssel szemben nem a teljes varianciát, hanem az egyes csoportok közti varianciát igyekszik maximalizálni. Az így kapott kanonikus változók száma a csoportok számánál eggyel kevesebb. Az első kanonikus változónak van a legnagyobb szerepe az osztályok elkülönítésében, a másodiknak a második legnagyobb, stb.

Az LDA-t hasonló módon lehet változókiválasztásra használni és validálni, mint a regressziós módszereket. A helyesen besorolt objektumok aránya (%-ban kifejezve) mutatja az osztályozás hatékonyságát. Az LDA hátránya, hogy hajlamos a túlillesztésre.

A döntési fák (CART, classification and regression trees) szintén osztályozási célokat szolgálnak. Csomópontokból állnak, amelyekből elágazások indulnak ki. Azok a csomópontok, amikből már nem indul ki újabb elágazás, adják a keresett osztályozást. A döntésekből jönnek létre a szabályok, amik alapján az ismeretlen objektumok is osztályokba sorolhatók. A döntések logikai jellegűek, azaz ha egy bizonyos deszkriptor értéke kisebb vagy nagyobb egy meghatározott értéknél, akkor az objektum az elágazás egyik vagy másik végére kerül. A döntési fák értelmezése általában egyszerű és igen szemléletes képet nyújtanak.

[*Sokváltozós adatelemzés, 2001c; Breiman, 1984; Vandeginste, 1998*]

2.2.7 Genetikus algoritmus

A genetikus algoritmus (GA, genetic algorithm) fejlesztői az evolúció természetes folyamatait vették mintául ennek a minimumkeresési módszernek a megalkotásához, mely változókiválasztásra is használható. A genetikus algoritmus úgy próbálja megtalálni a lehető legjobb megoldást, hogy a többi módszerektől eltérően nem csak egy, hanem párhuzamosan sok eshetőséget vizsgál egyszerre és a számítási lépések közben folyamatosan javítja a modellt. A kiindulási állapotot úgy kell elképzelnünk, hogy különféle „kromoszómáink", bitfüzéreink vannak, melyek deszkriptor-kombinációkat jelentenek: ha egy deszkriptor szerepel a kromoszómában, akkor a kódja 1, ha nem, akkor 0. A „gének", azaz a deszkriptor kódolási szakaszok jelenléte egy kromoszómán véletlenszerű. A számítás kezdetekor rendelkezésre álló kromoszómakészlet a kiindulási populáció. Az egyes kromoszómák „rátermettségét" statisztikai célfüggvények segítségével vizsgálhatjuk (leggyakrabban R^2 vagy $R^2_{CV,}$ például MLR számítással), azaz megállapíthatjuk, hogy milyen szoros az összefüggés az adott deszkriptor-kombináció és a kódolni kívánt függő változó között. Majd a populáció tagjai között rekombináció (génkicserélődés, azaz deszkriptorok cseréje) zajlik le és mutáció (egyes gének helyettesítése más génekkel random módon) is történik és így jön létre a második populáció, melynek szintén megvizsgáljuk a rátermettségét, és a ciklus addig folytatódik, amíg az elvárt eredményt elérjük. Természetesen dolgozhatunk egyszerre több kezdeti populációval is, ez esetben a számítási folyamat a több populációban független egymástól. Bonyolult, sokváltozós problémák esetén igen jó eredményeket adhat ez a módszer.

[*Lucasius és Kateman, 1993; Lucasius és Kateman, 1994; Niculescu, 2003*].

2.3 Programok

A molekulák háromdimenziós szerkezetének optimálását a HyperChem [*HyperChem*] programmal hajtottam végre. Az optimálás a szemi-empirikus AM1 módszerrel történt. A deszkriptorokat a Dragon program segítségével számítottam [*Todeschini, 2002*]. A PCM számításokat egy Microsoft Excel makróval végeztem, melyet Rajkó Róbert készített. A Lasso számításokat Forrest Stout és John F. Kalivas végezték Matlab programmal [*Matlab*]. A 3.4.2. fejezetben található egyik PLS számítást (metil-észterek regressziós koefficienseinek varianciája) Héberger Károly végezte az Unscrambler programmal [*Unscrambler*]. A genetikus algoritmus futtatását a MobyDigs programmal végeztem [*Todeschini, 2004*]. A parciális legkisebb négyzetek módszere és diszkriminancia analízis összekapcsolásával végzett csoportosítást pedig egy házi Microsoft Excel Makróval számítottam, melyet Molnár Tamás doktoráns készített Héberger Károly útmutatása alapján. Az összes többi számítást a Statistica program felhasználásával készítettem el [*Statistica 5.5, Statistica 6*].

Irodalmak a 2. fejezethez

Baumann K. Cross-validation as the objective function for variable selection techiques *TrAC, Trends Anal. Chem.* **2003**, 22 (6), 395-406.

Baumann K. Validation tools for variable subset regression *J. Comput.-Aided Mol. Des.* **2004**, 18, 549-562.

Breiman L., Friedman J. H., Olshen R. A., Stone C. J. Classification and regression trees Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software, **1984**

Conover J., Practical Nonparametric Statistics, 2nd edition, John Wiley & Sons, New York, **1980**

Consonni V., Todeschini R., Pavan M. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors 1. Theory of novel 3D molecular descriptors *J. Chem. Inf. Comput. Sci.* **2002**, 42, 682-692.

Consonni V, Todeschini R., Pavan M., Gramatica P., Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies *J. Chem. Inf. Comput. Sci*. **2002**, 42, 693-705.

Di Marzio W., Galassi S., Todeschini R., Consolaro F. Traditional versus WHIM molecular descriptors in QSAR approaches applied to fish toxicity studies *Chemosphere* **2001**, 44, 401-406.

Draper N.R., Smith H. In *Applied Regression Analysis*, John Wiley & Sons Inc.: New York, **1981**, 412-419.

Esbensen K.H. Multivariate Data Analysis – In Practice. An Intruduction to multivariate data analysis and experimental design, **2002**, 5.kiadás, Camo Process

Forrester J.B., Kalivas J.H. Ridge Regression Optimization Using a Harmonius Approach *J. Chemom*. **2004**, *18*, 372-384.

Galvez J., Garcia R., Salabert M.T., Soler R. Charge Indexes. New Topological Descriptors. *J. Chem. Inf. Comput. Sci.* **1994**, 34, 520-552

Geladi P., Kowalski B.R. Partial least-squares regression: a tutorial. *Anal. Chim. Acta* **1986**, 185, 1-17.

Golbraikh A., Tropsha A. Beware of q2! *J. Mol. Graphics Mod*. **2002**, 20, 269-276.
Hawkins D.M., Bashak S.C., Mills D. Assessing model fit by cross-validation *J. Chem. Inf. Comput. Sci.* **2003**, 43, 579-586.

Gramatica P., Navas N., Todeschini R. 3D-modelling and prediction by WHIM descriptors. Part 9. Chromatographic relative retention time and physico-chemical properties of polychlorinated biphenyls (PCBs) *Chemom. Intell. Lab. Syst*. **1998**, 40, 53-63.

Héberger K., Rajkó R. Generalization of Pair – Correlation Method (PCM) for Nonparametric Variable Selection *J. Chemom*. **2002**, 16, 436-443.

Hoerl A.E., Kennard R.W.
a, Ridge Regression: Biased Estimation for Nonorthogonal Problems *Technometrics* **1970**,12, 55-67.

b, Ridge regression: Applications to Nonorthogonal Problems *Technometrics* **1970**, 12, 69-82.

Kier L.B., Hall L.H. Derivation and Significance of Valence Molecular Connectivity *J. Pharmacol. Sci.* **1981**, 70, 583-589.

Labute P. A widely applicable set of descriptors *J. Mol. Graphics Model*. **2000**, 18, 464-477.

Livingstone J.D. The characterization of chemical structures using molecular properties. A survey *J. Chem. Inf. Comput. Sci.* **2000**, 40, 195-209.

Livingstone D.J., Salt D.W. Variable selection – Spoilt for choice? *Reviews in Computational Chemistry, Volume 21,* szerk: Lipkowitz K.B., Larter R., Cundari T.R., Wiley, **2005**

Lucasius C.B., Kateman G. Understanding and using genetic algorithms: Part 1. Concepts, properties and context. *Chemom. Intell. Lab. Syst*. **1993**, 19, 1–33.

Lucasius C.B., Kateman G. Understanding and using genetic algorithms Part 2. Representation, configuration and hybridization *Chemom. Intell. Lab. Syst*. **1994**, 25, 99-145.

Matlab 7.0 with the Matlab Optimization Toolbox, The MathWorks, Natick, MA, USA

Miller A.J. *Subset Selection in Regression*, Chapman and Hall: London, **1990**, 43-82.

Niculescu S.P. Artificial neural networks and genetic algorithms in QSAR *J. Mol. Struct. (Teochem)* **2003**, 622, 71-83.

Rajkó R., Héberger K. Conditional Fisher's Exact Test as a Selection Criterion for Pair-Correlation Method. Type I And Type II Errors *Chemom. Intell. Lab. Syst*. **2001**, 57, 1-14.

Randic M. J. On Characterization of Molecular Branching *J. Am. Chem. Soc*. **1975**, 97, 6609-6615.

Sokváltozós adatelemzés (Kemometria) szerk: Horvai Gy., Nemzeti Tankönyvkiadó, Budapest, **2001**
a,1.3. fejezet:Faktoranalízis, főkomponens analízis és változataik 84-88 (Héberger K., Rajkó R.)
b,2.2. fejezet: Modellépítés a regressziós számítások során, korrelált változók esetén 191-192 (Héberger K., Rajkó R., Kolossváry I.)
c,1.2. fejezet: Csoportosítás (alakfelismerés) 56-66 (Borosy A.P.)

Statistica 5.5 Software Package, StatSoft Inc., Tulsa, OK, USA

Statistica 6 Software Package, StatSoft Inc., Tulsa, OK, USA
Tibshirani R. Regression shrinkage and selection via the Lasso *J. Royal. Stat. Soc B.* **1996**, 58, 267-288.

Tikhonov, A.N. Solution of Incorrectly Formulated Problems and the Reguralization Method *Sov. Math. Dokl*. **1963**, *4*, 1035-1038.

Todeschini R., Lasagni M., Marengo E. New molecular descriptors for 2D and 3D structures – Theory. *J. Chemom.* **1994**, 8, 263-272.

Todeschini R., Consonni V. *Handbook of Molecular Descriptors*; Wiley – VCH: Weinheim, **2000**

Todeschini R., Consonni V., Pavan, M. Dragon Software Version 2.1, **2002**.

Todeschini R., Ballabio D., Consonni V., Mauri A., Pavan M. MobyDigs Professional version 1.0 **2004** Milano Chemometrics and QSAR Research Group

Topliss J.G., Edwards R.P. Chance factors in studies of quantitative structure-activity relationships *J. Med. Chem.* **1979** 22 (10) 1238-1244.

Unscrambler 9.2, Camo Process As., Oslo, Norway

Vandeginste B.G.M., Massart D.L., Buydens L.M.C, Jong S.D.E., Lewi P.J., Smeyers-Verbeke J., In *Handbook of Chemometrics and Qualimetrics*: Part B; Elsevier: Amsterdam, The Netherlands, **1998** Chapter 33. Supervised pattern recognition, 207-223.

Vanyúr R., Héberger K., Kövesdi I., Jakus J. Prediction of tumoricidal activity and accumulation of photosensitizers in photodynamic therapy using multiple linear regression and artificial neural networks *Photochem. Photobiol.* **2002**, 75, 471-478.

Wiener H. Structural determination of paraffin boiling points *J. Am. Chem. Soc*. **1947**, 69, 17-20.

Wold S., Esbensen K., Geladi P., Principal Component Analysis, *Chemom. Intell. Lab. Syst*. **1987**, 2, 37–52.

Wold S. PLS for Multivariate Linear Modeling. *Chemometric Methods for Molecular Design,* szerk: Waterbeemd H.,VCH, Weinheim, **1995**,195-218.

3. VÁLTOZÓKIVÁLASZTÁSI MÓDSZEREK HATÉKONYSÁGÁNAK ÖSSZEHASONLÍTÁSA RETENCIÓS INDEXEK ELŐREJELZÉSÉBEN

3.1. Mi okoz nehézséget a változókiválasztási módszerek hatékonyságának összehasonlításában?

A 2.1.3. fejezetben már szó esett arról, hogy a változókiválasztás kulcsfontosságú részfolyamat a QSAR modellezésben. A modellek jóságának és egyben az egyes módszerek hatékonyságának összehasonlítására a 2.1.4. fejezetben ismertetett statisztikai paramétereket alkalmazom. A 2.1.5. fejezetben található 2.1.5.1. ábrán jól látszik, hogy minél több független változót veszünk be a modellbe, annál jobb lesz annak leíróképessége, ám túl sok deszkriptor bevonása a modell túlillesztéséhez és előrejelző képességének romlásához vezet. Azt is említettem a 2.2.3 fejezetben, hogy bizonyos módszerek esetén, mint pl. a látens változókat használó PLS vagy PCA, a szabadsági fok nem definiálható egyértelműen, és ez a módszerek összehasonlítását megnehezíti. Továbbá, a modellben levő független változók számának meghatározásakor érdemes szem előtt tartani az ún. minimálelvet vagy Occam borotvájának elvét, miszerint két (közel) azonos teljesítőképességű modell közül az egyszerűbbet kell választani.

Ha tehát korrekt módon akarjuk két módszer teljesítőképességét összehasonlítani, figyelembe kell venni az azokkal létrehozott modellek szabadsági fokát is. Ezért olyan modelleket építettem, amelyekben a független változók száma megegyezik. Az alkoholok Kováts-indexének előrejelzésekor öt különféle változókiválasztási módszert használtam, majd azonos számú változót tartalmazó MLR modelleket építettem. A zsírsav metil-észterek retenciós indexének előrejelzésekor továbbfejlesztettem a változókiválasztási módszerek helyes összehasonlításának módját. Egyrészt hasonlóképpen építettem modelleket, mint az alkoholok esetén, másrészt ún. optimális modelleket is építettem (nem rögzítettem a deszkriptorok számát

a modellekben és nem feltétlenül MLR-rel, hanem ha lehetőség volt rá, az aktuális (pl. PLS, stb.) módszerrel építettem a modelleket) és a kétféle módon kapott modellek statisztikai paramétereit összehasonlítottam.

3.2. A Kováts-féle retenciós index

Modelljeinkben a függő változó a Kováts-féle retenciós index volt [*Kováts, 1958*]. Ez egy speciális, a homológ sorokon belül a retenciós idő logaritmusa és a szénatomszám linearitására épülő relatív retenciós adat, mely izoterm körülmények között az alábbi módon definiálható:

$$\mathrm{RI}_x = 100\frac{\lg t'_{Rx} - \lg t'_{Rn}}{\lg t'_{Rn+1} - \lg t'_{Rn}} + 100n$$

Így egy tetszőleges komponens retenciós ideje két normál alkán retenciója alapján logaritmikus interpolációval határozható meg. A t'_{Rx} az ismeretlen, a t'_{Rn} és $t'_{\mathrm{Rn+1}}$ az n és n+1 szénatomszámú n-alkán redukált retenciós ideje ($t'_{\mathrm{R}} = t_{\mathrm{R}} - t_{\mathrm{M}}$, ahol t_{R} a bruttó retenciós idő, t_{M} pedig az oszlop holt ideje, rendszerint metánnal határozzák meg).

Gázkromatográfiás analízissel nagymennyiségű (retenciós) adatot nyerhetünk szerkezetileg igen változatos vegyületekről, a Kováts-index pedig kiváló reprodukálhatósága miatt jól alkalmazható függő változóként a változókiválasztási módszerek összehasonlításánál. Az alkoholok retenciójának mechanizmusa apoláris állófázison jól ismert, ami szintén a modellek korrekt összehasonlítást segíti.

3.3 Alkoholok Kováts- indexének előrejelzése

3.3.1. Alkoholok retenciós index előrejelzésének irodalma, célkitűzés

Az alkoholok viszonylag nagy dipólusmomentummal rendelkeznek, ezért köztük és az állófázis közt jóval komplexebb kölcsönhatások lépnek fel, mint pl. szénhidrogének esetén. A molekulák méretét és alakját leíró deszkriptorokon kívül a polaritást jellemző tulajdonságokat is be kell építeni a modellekbe. Bermejo a forráspontot, moláris refrakciót és elágazási indexeket használta alkoholok Kováts-indexének előrejelzésére [*Bermejo és mtsai, 1987*], mások [*Bergman és mtsai, 1991*] ezeket a deszkriptorokat a dipólusmomentummal egészítették ki. Az olyan fizikai paraméterek használata, mint a forráspont, kiváló előrejelzést tesznek lehetővé, de a vegyületek nagy részénél nincs rá irodalmi adat. Elágazási indexek felhasználásával [*Guo és mtsai, 2000*] is készült alkoholok retencióját leíró modell. Junkes és munkatársai [*Junkes és mtsai, 2003*] egy általuk kifejlesztett ún. szemiempirikus topológiai index alkalmazásával hoztak létre előrejelzésre alkalmas modelleket.

Mivel alkoholok retenciós indexének előrejelzésére már megfelelő modellek állnak rendelkezésre, elsődleges célom nem a minél jobb modellek építése volt, hanem a különféle változókiválasztási módszerek összehasonlítása. Másodlagos feladatomnak tekintettem a viszonylag új WHIM deszkriptorok alkalmazhatóságának vizsgálatát retenciós indexek előrejelzésében.

3.3.2. A számítás menete

Összesen 44 egyenes láncú és elágazó telített alkohol OV-1 állófázison mért Kováts-indexét használtam fel a modellépítéshez [*Pias és mtsai, 1975; Zhang és mtsai, 1996*]. Mivel a WHIM deszkriptorok a molekula háromdimenziós szerkezetéről szolgáltatnak információt,

elvégeztem az alkoholok geometriájának optimálását. A számított WHIM és a nulladimenziós deszkriptorok száma 109 volt. (A Dragon programot úgy állítottam be, hogy automatikusan eltávolította azokat a deszkriptorokat, melyek közt a korreláció mértéke nagyobb, mint 0,99.) Majd a független változók számát főkomponens-elemzéssel csökkentettem. A főkomponens-együtthatók ábrájának értékelése vizuális módon történt. Jól látható volt, hogy a deszkriptorok egy része a retenciós index körül helyezkedik el, míg más deszkriptorok jóval távolabb találhatók. Ezért az ábrán egy kört húztam a Kováts-indexet jelképező pont köré (r=0,4 egység) és az azon belül található (erősen korrelált) deszkriptorokat választottam ki a további számításokhoz.

Öt különböző módszert, peremregressziót (RR), részleges legkisebb négyzetek módszerét (PLS), páronkénti korrelációs módszert (PCM), előreirányuló változóbevonást (FS) valamint a legjobb alrendszer regressziót (BSS) használtam változókiválasztásra, és minden esetben többszörös lineáris regresszióval építettem modellt. A modellek négy illetve három változót tartalmaztak. (Az irodalom szerint ennyi független változó elég az alkoholok retenciójának modellezésére, és az egyszerű értelmezhetőség miatt sem érdemes ennél többet használni.) A változókiválasztáskor az összes vegyületet felhasználtam, a modell előrejelző képességének tesztelésére pedig *n* elem kihagyásos keresztellenőrzést végeztem. Az adatkészletet két részre osztottam, a betanuló (és egyben kalibrációs) készlet a molekulák 67, míg a teszt készlet a 33%-ukat tartalmazta. Négy különböző betanuló-teszt készlet kombinációt használtam. Az előrejelzés jóságát a négy készletkombinációval végzett számításokból kapott négy R^2_{CV} és SD_{CV} adat átlagával jellemeztem. Ezeken a statisztikai paramétereken kívül a reziduumok ábráját is tanulmányoztam a modellek értékelésekor.

A változószelektálást a peremregresszióval a következőképpen hajtottam végre. Kilenc különböző *k* értéket beállítva (0,001-0,6 között) kiszámítottam a regressziós egyenletek meredekségét, majd a regresziós egyenletek becsült paramétereit a *k* értékek függvényében,

azaz a peremregresszió nyomának az ábráján. Az ábra segítségével kiválasztottuk azt a 4 deszkriptort, amelyek a nullához tartottak és legmeredekebben futottak felé.

A PLS esetében a négy ill. három legnagyobb regressziós koefficiensű deszkriptort választottam ki. Az előreirányuló változóbevonáskor 5%-os szignifikanciaszintet állítottam be, és az első négy illetve három legkisebb *p* értékkel rendelkező független változót választottam ki. A BSS esetén a legnagyobb R^2 értékkel rendelkező három illetve négy deszkriptort tartalmazó modelleket választottam. A PCM esetén a súlyozásos rendezéssel rangsoroltam a változókat, és értelemszerűen az első négyet illetve hármat választottam ki közülük.

3.3.3. Eredmények és következtetések

A 3.3.3.1. táblázatban láthatók a négy illetve három független változót tartalmazó modellek statisztikai paraméterei. A 2. fejezetben tárgyalt statisztikai paramétereken kívül bevezettem egy újat, egy ún. „egyensúlyi" paramétert (*b*), mely az SD és az SD_{CV} hányadosa, vagyis arról ad képet, mennyire „kiegyensúlyozott" az adott modell a leíró és előrejelző képességét tekintve. Ha *b* értéke közel van 1-hez, akkor a modell előrejelző képessége hasonló a leíróképességéhez. Ha a *b* érték 0,67-hoz közeli, az azt jelenti, hogy a modell kiegyensúlyozottsága megfelel annak az elvárásnak, ami a jelen esetben az adatkészlet felosztásából következik (a molekuláknak 67%-a a betanuló készlet tagja). Ha a *b* érték ennél kisebb, az azt jelzi, hogy a modell leíró és előrejelző képessége között valószínűleg csekély az összefüggés.

A 3.3.3.1. táblázat adataiból látszik, hogy az RR 4 és RR 3 modellek statisztikai paraméterei nem tekinthetők túl jónak; az SD értékek magasak, az R^2 és a R^2_{CV} értékek pedig a legalacsonyabbak az összes modell közül. Hasonló mondható el a PLS modellekről is. Az 3.3.3.1. ábrán egymás alatt bemutatom a PLS 4 modell mért és számított RI értékei közti kapcsolatot mutató ábráját (a) (RR 4 modellé hasonló) és az RR4 modell reziduum ábráját (b)

(PLS 4 modellé is hasonlóan néz ki). Már a mért és számított RI értékek ábráján látszik, hogy az adatsor két végén található (legkisebb és legnagyobb RI-jű) alkoholok számított és mért Kováts-indexe közt egyirányú az eltérés, míg az adatsor közepén található vegyületek esetén az eltérés a másik irányú. Látványosabban ugyanezt mutatja a reziduumok U-alakú képe (3.3.3.1. b ábra).

A PCM modellek leíróképessége jelentősen jobb az RR és a PLS modellekhez képest, a statisztikai paraméterek alapján megfelelőnek tekinthető. Ám a reziduumok ábrája rendellenes képet tár elénk, valamint az előrejelző képességet jellemző statisztikai paraméterek értéke sem jó, így ezek a modellek sem megfelelők ismeretlen RI-k becslésére.

Az FS és BSS modellek leíró és előrejelző statisztikai paraméterei megfelelőnek tekinthetők, ezek a modellek megbízható becslésre alkalmasak alkoholok retenciós indexének előrejelzésében. Az 3.3.3.2. ábrán egymás mellett mutatom a PCM 4 és az FS 4 modell két-két ábráját, az egyiken a számított retenciós index értékek láthatók a kísérleti értékek függvényében (a), a másikon pedig a reziduumok függvényében (b). Jelentős különbség az (a) ábrák közt nem látható, a (b) ábrákon viszont jól látszik a két modell közti különbség. Míg a PCM 4 modell reziduumábrája rendellenes U-alakú, azaz ez a modell a jó statisztikai paraméterei ellenére sem adekvát az RI-leírásban, addig az FS 4 modell reziduumábrája normális képet mutat.

A *b* paraméter az RR4, az FS és a BSS modell kiegyensúlyozottságát mutatja. A PCM modellek esetén értéke kis kiegyensúlyozatlanságra utal, a modellek kevésbé jól teljesítenek az előrejelzésben, mint az elvárható lenne a leíróképességük alapján.

A 3.3.3.1 táblázat a modellekben található deszkriptorokat is feltünteti. A deszkriptorok rövidítésének jelentése a Függelékben található, az F1 táblázatban. Látszik, hogy a molekulatömeg és a méretet valamint a polaritást egyaránt jellemző átlagos elektrotopológiai állapot kulcsszerepet játszik a retenciós viselkedés leírásában. A modelleket azonban ki kell

3.3.3.1. táblázat

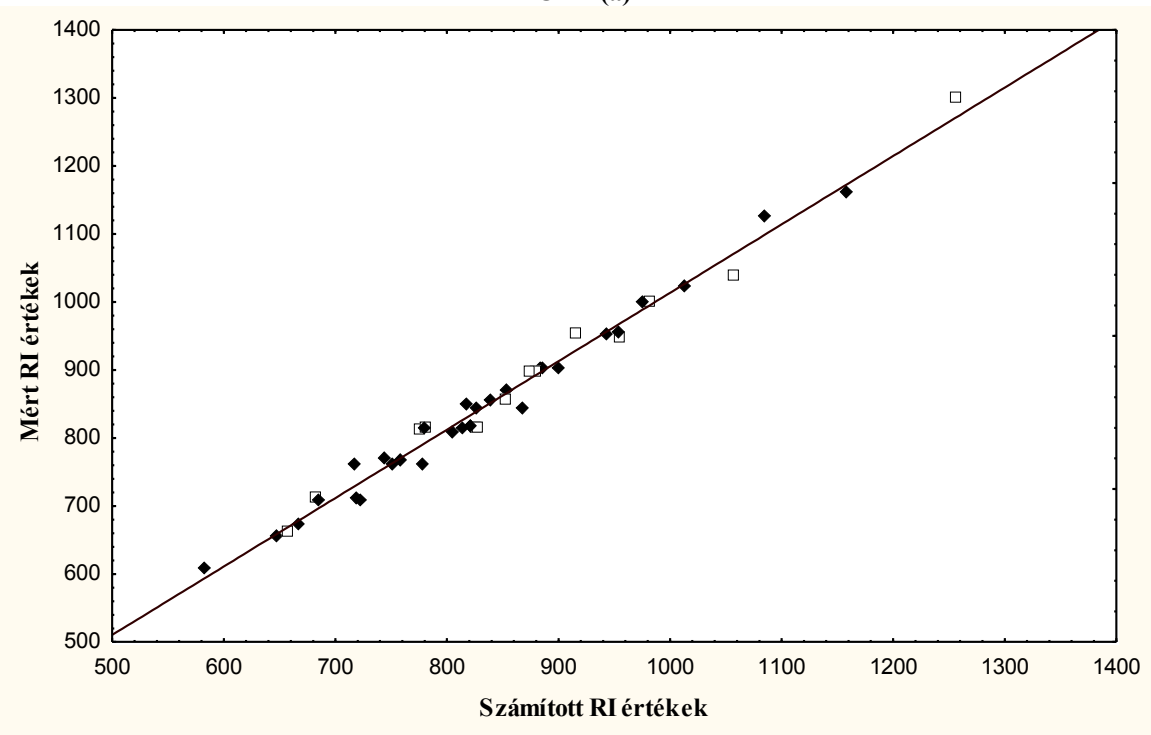
Az alkoholok RI számítására alkalmazott modellek statisztikai paraméterei és a modellekben található deszkriptorok

Változókiválasztási módszer	Deszkriptorok a modellben	R^2	F	SD	$\overline{R}^2_{CV}$	SD_{CV}	*b*
RR 4	**MW, P2u, P2m, P1u**	0,9442	164,92	33,95	0,8193	50,23	0,68
RR 3	**MW, P2u, P2m**	0,9433	222,10	33,76	0,8406	44,65	0,76
PLS 4	**AMW, Ms, P1u, G1m**	0,9457	169,97	33,47	0,9455	33,52	1,00
PLS 3	**AMW, Ms, P1u**	0,9456	232,21	33,06	0,9426	31,84	1,04
PCM 4	**MW, AMW, Ms, Vm**	0,9819	531,56	19,29	0,9398	36,99	0,52
PCM 3	**MW, AMW, Ms**	0,9816	712,85	19,23	0,9397	35,17	0,55
FS 4	**Ms, Vu, L1s, Tv**	0,9883	824,11	15,53	0,9796	22,94	0,68
FS 3	**Ms, Vu, L1s**	0,9826	753,42	18,71	0,9659	29,56	0,63
BSS 4	**Ms, L1u, L1e, Vu**	0,9902	985,60	14,22	0,9801	22,40	0,63
BSS 3	**Ms, As, Vm**	0,9838	810,34	18,05	0,9707	29,49	0,61

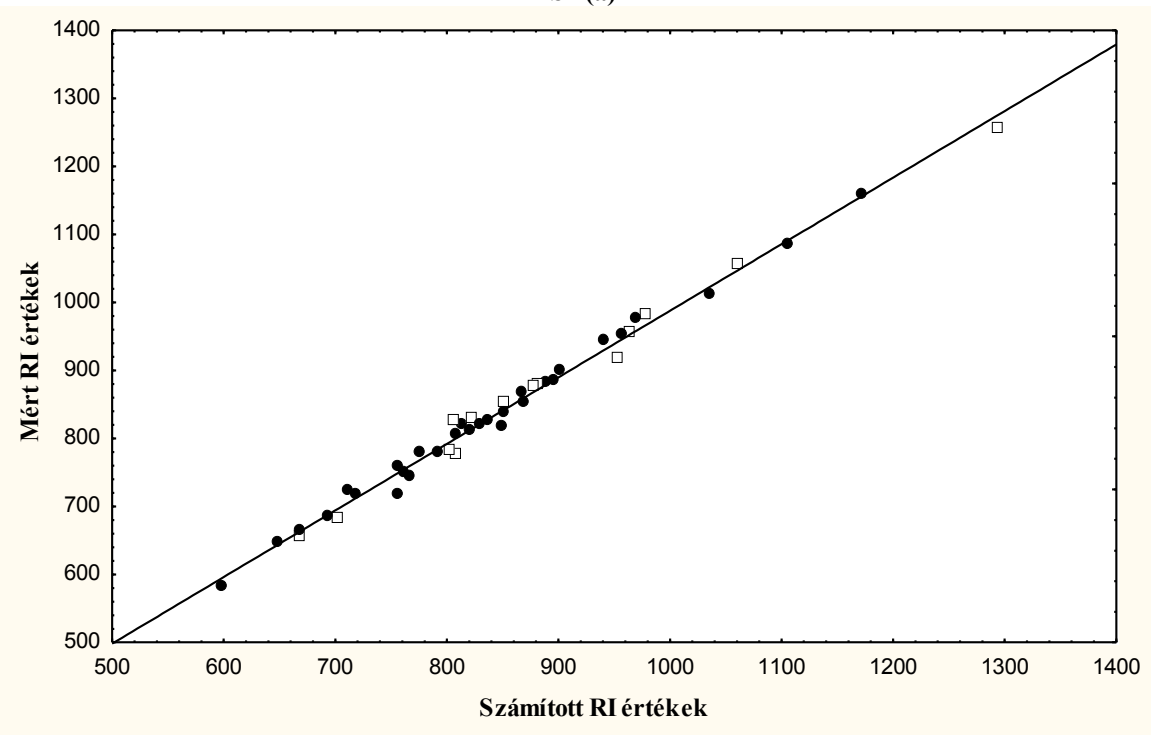
3.3.3.2 ábra

PCM4 és FS 4 modellek összehasonlítása

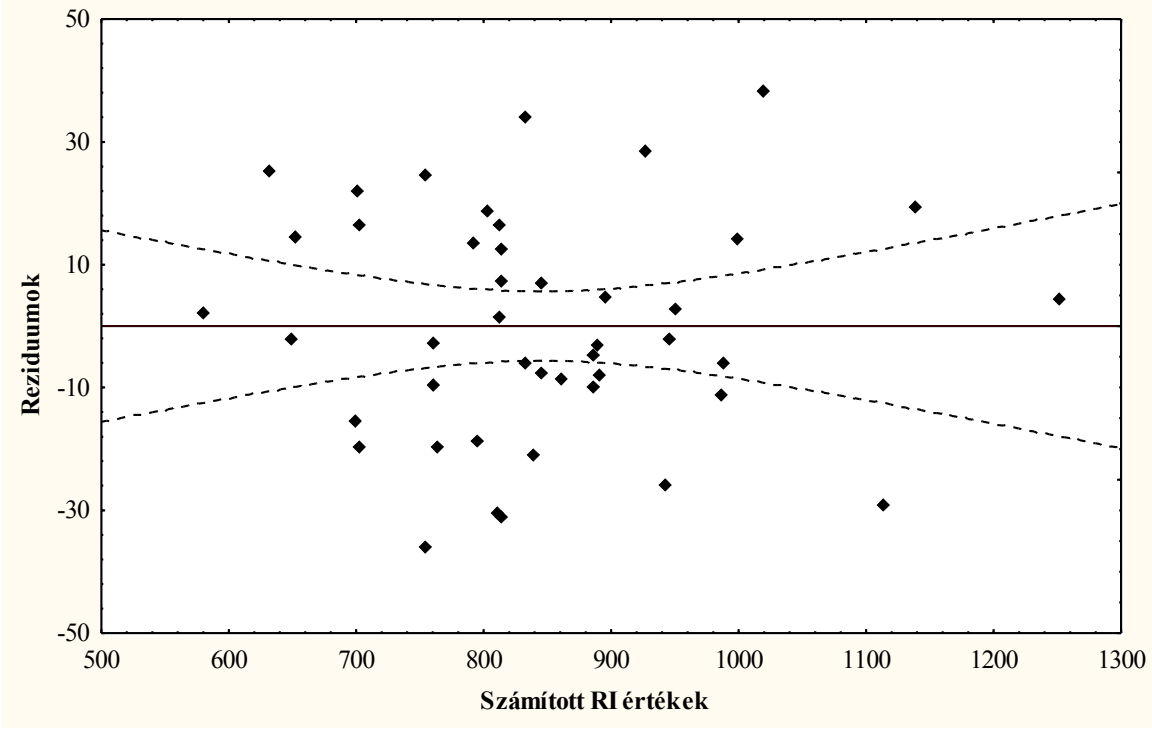
PCM4 (a)



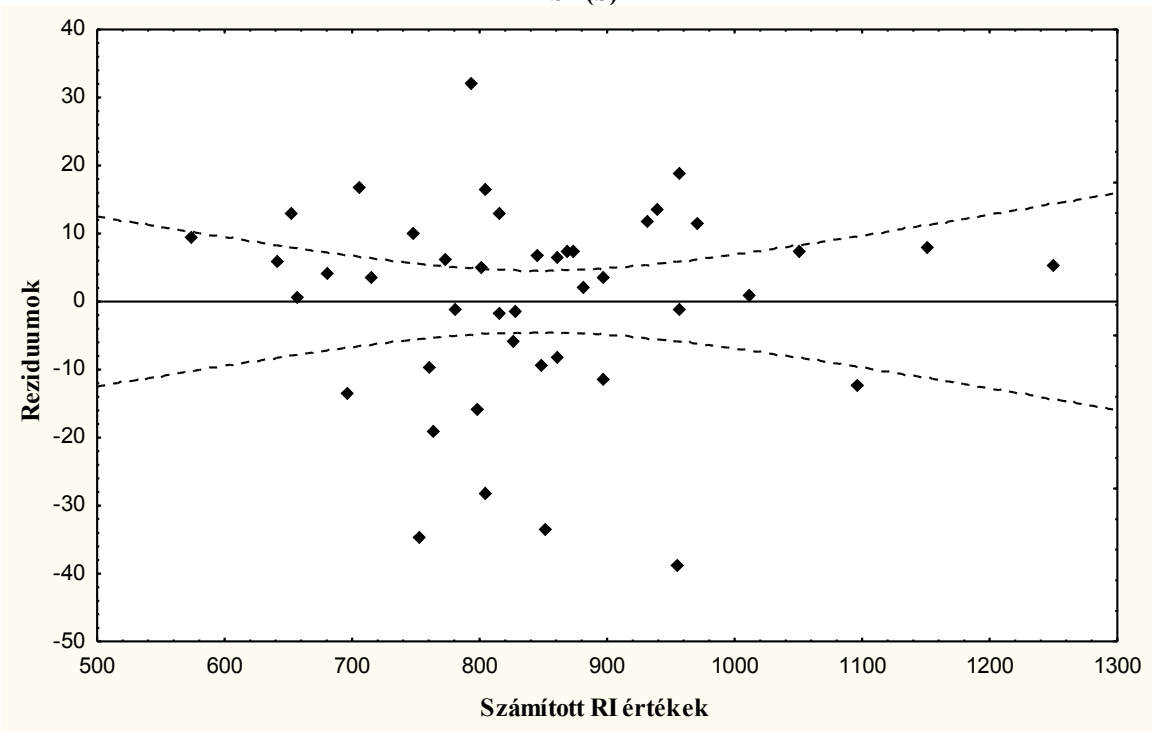
FS4 (a)



PCM4 (b)



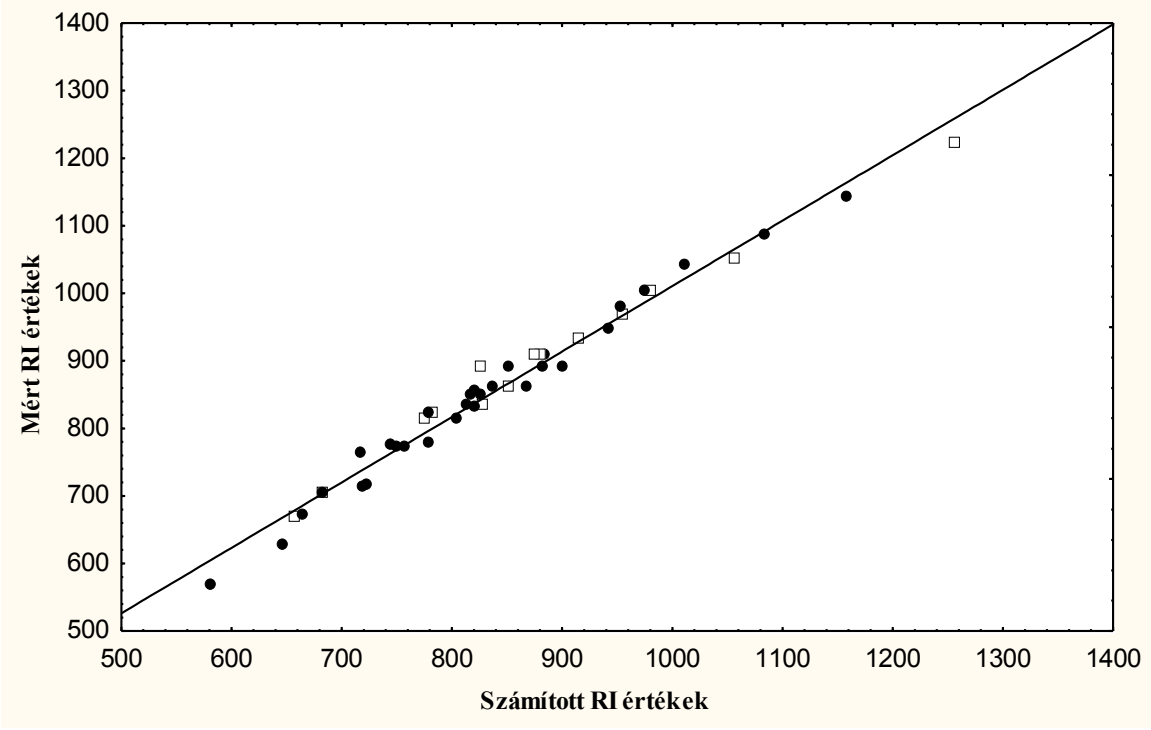
FS4 (b)



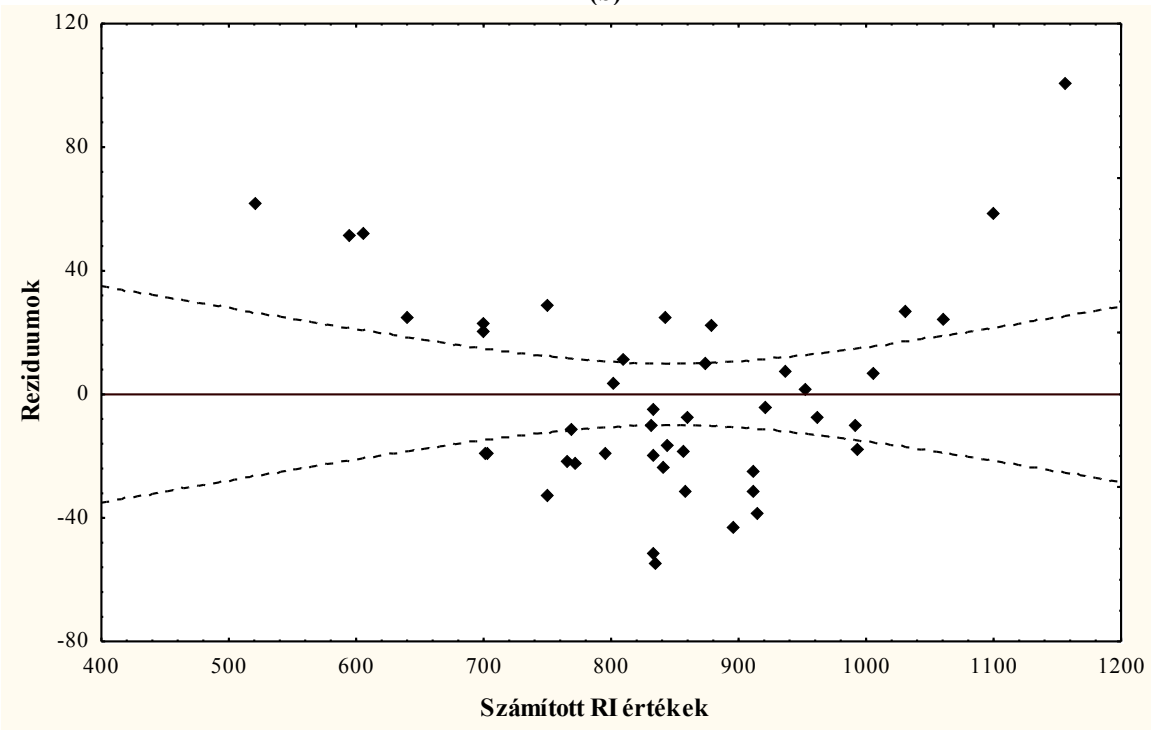
3.3.3.1. ábra

A PLS 4 modell számított és mért RI értékei (a), és az RR 4 modell reziduum-ábrája (b)

PLS 4 (a)



RR4 (b)



egészíteni más deszkriptorokkal, amelyek a molekulák alakját, kompaktságát, méretét és szimmetriáját jellemzik. Erre a WHIM deszkriptorok alkalmasnak bizonyultak.

Bár elsődleges célom nem az volt, hogy az irodalomban található modelleknél jobbat hozzak létre, elvégeztem az azokkal való összehasonlítást. Guo munkája nem tartalmazta a modellek R^2_{CV} értékét. Ennek számítását a cikk adataiból elvégeztem, és R^2_{CV} = 0,9920,

valamint SD_{CV} = 21.9 értéket kaptam. Nála a teszt készletben a vegyületek száma mindössze 6 volt. Az általam kapott FS 4 és BSS 4 modellek SD_{CV} értéke a szabadsági fokokat is figyelembe véve hasonlónak mondható. Junkesék modelljének R^2_{CV} értéke kiváló, ők azonban a szemiempirikus topológiai index számításához magukat a mért RI-adatokat is felhasználják.

Az 1990-es években volt egy tudományos vita arról, hogy melyik módszer a legjobb az előrejelzésre. Frank és Friedman [*Frank és Friedman, 1993*] szimulációs számítások alapján arra következtettek, hogy a regressziós koefficiensek „összenyomásával" dolgozó peremregresszió a leghatékonyabb módszer, és felülmúlja a látens változókat használó PLS-t és PCR-t, valamint a BSS-t is. Később Friedman [*Hastie és mtsai, 2001*] egy valós példán (ahol prosztatarák kialakulásában szerepet játszó tényezők kiválasztása volt a cél) is hasonlókat állapított meg: az RR és a PLS hasonló eredményt adnak, de míg az RR a *k* paraméter (tetszőleges) változtatásával folyamatosan csökkenti a regressziós koefficienseket, addig a PLS diszkrét lépésekben, és ez kissé instabillá teszi a PLS-t. A BSS pedig szintén diszkrét lépésekben csökkenti a regressziós koefficienseket, és gyakran nem tudja elkerülni a „túllövésnek" (overshoot) problémát. Friedmanékkal azonban nem mindenki ért egyet. Wold [Wold*, 1993*] szerint a PLS, mint látens változókkal operáló módszer számos kémiai probléma megoldásánál hasznosabbnak bizonyult, mint az RR, különösen, akkor, amikor erősen korreláló független változók voltak a modellben. Míg az RR igazi optimumát még vizuális értékeléssel is nehéz, ha nem lehetetlen meghatározni, a PLS optimum viszonylag egyszerűen adódik: keresztellenőrzést kell végzni, és a legkisebb SD_{CV}-t eredményező PLS modellt kell választani.

Mivel az irodalomban ez említett módszereket nemcsak paraméterbecslésre, hanem változókiválasztásra is használják érdemes volt megnézni teljesítőképességüket a modellezés ezen aspektusában, még akkor is ha ezért eredményeim nem hasonlíthatók össze közvetlenül a fent említett szerzők eredményeivel. Érdekes azonban megvizsgálni, hogy van-e különbség a teljesítőképességükben attól függően, hogy modellépítésre vagy változókiválasztásra

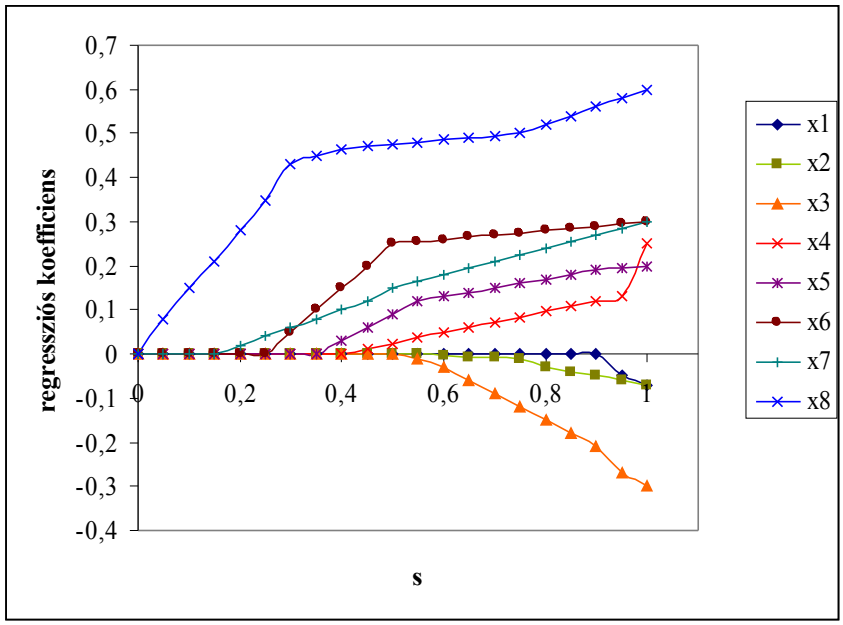
használjuk őket. Eredményeim azt mutatják, hogy a PLS valamivel jobbnak bizonyult a változókiválasztásban a peremregressziónál, ám a legjobb eredményt az alkoholok retenciójának előrejelzésénél a „hagyományos” módszerek, azaz az FS illetve a BSS adták.

Friedman 2001-ben kiadott könyvében a peremregresszió mellett a hasonló Lasso módszer lehetőségeit is bemutatja a változókiválasztásban. A két módszer közti különbséget a már említett prosztatarák előrejelzéses példával lehet érzékeltetni (3.3.3.3. ábra, Friedman alapján). A peremregresszió nyomának ábráján a *k* paraméter variálásakor az egyes változók regressziós koefficiense változik a *k* változtatásával (csökkentésével csökken), de nem lesz olyan *k* érték, ahol egy vagy több változó nulla, vagy nullához közeli értéket vesz fel (azaz kiesik), csak amikor *k* is nulla lesz. A Lasso módszer ezzel szemben jól alkalmazható a változókiválasztásra, mert ott a *k*-nak megfelelő paraméter változtatásakor a független változók közül a kevésbé jelentősek „belefutnak” a nullába, vagy nullához nagyon közeli értékbe, így egyértelműen kihagyhatók a modellből. Egy másik jelentős összefoglaló szerzői [*Draper és Smith, 1998*] is hasonló álláspontot képviselnek. Ezért későbbi munkám során, a zsírsav metil-észterek modellezésénél, a peremregresszió helyett a Lasso módszert (is) alkalmazom változókiválasztásra.

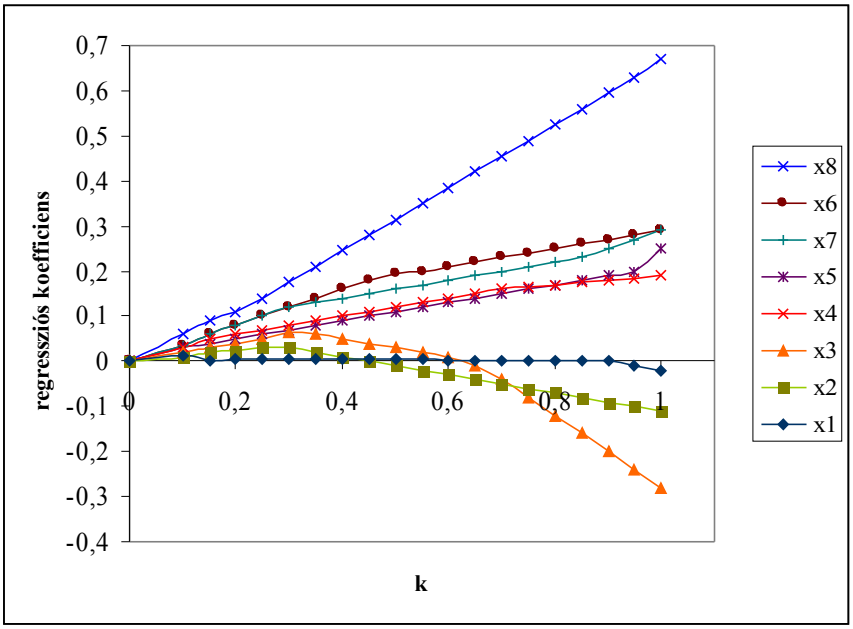
3.3.3.3. ábra

A peremregresszió és a Lasso hatékonyságának összehasonlítása a változókiválasztásban

a, Nyolc független változó regressziós koefficiens értékei különféle összenyomási faktor (s) értékeknél, Lasso módszer



B, Nyolc független változó regressziós koefficiens értékei különféle *k* értékeknél, peremregresszió



3.3.4. Összefoglalás

- A legjobb módszer a független változók kiválasztására - ellentétben az elvárásokkal - a legjobb alrendszer módszer (BSS) és az előreirányuló változóbevonás (FS) volt. Ezekkel a módszerekkel sikerült alkoholok Kováts-indexének előrejelzésére alkalmas modelleket építenem.

- Elsőként építettem olyan modelleket, amikben alakkal és mérettel összefüggő WHIM indexek szerepelnek (nulladimenziós deszkriptorokkal kiegészítve), az alkoholok Kováts-retenciós indexének leírására és előrejelzésére.

- A PCM módszer alkalmazásával jó leíróképességű modellt sikerült létrehoznom. A modellt viszont előrejelzésre nem célszerű alkalmazni.

- Megállapítottam, hogy - az irodalomtól eltérően - a PLS módszer felülmúlta a peremregressziót az RI előrejelzésben, de a más esetekben hatékonyan alkalmazott peremregresszió és PLS módszer nem vezettek elfogadható eredményre a modellépítés során.

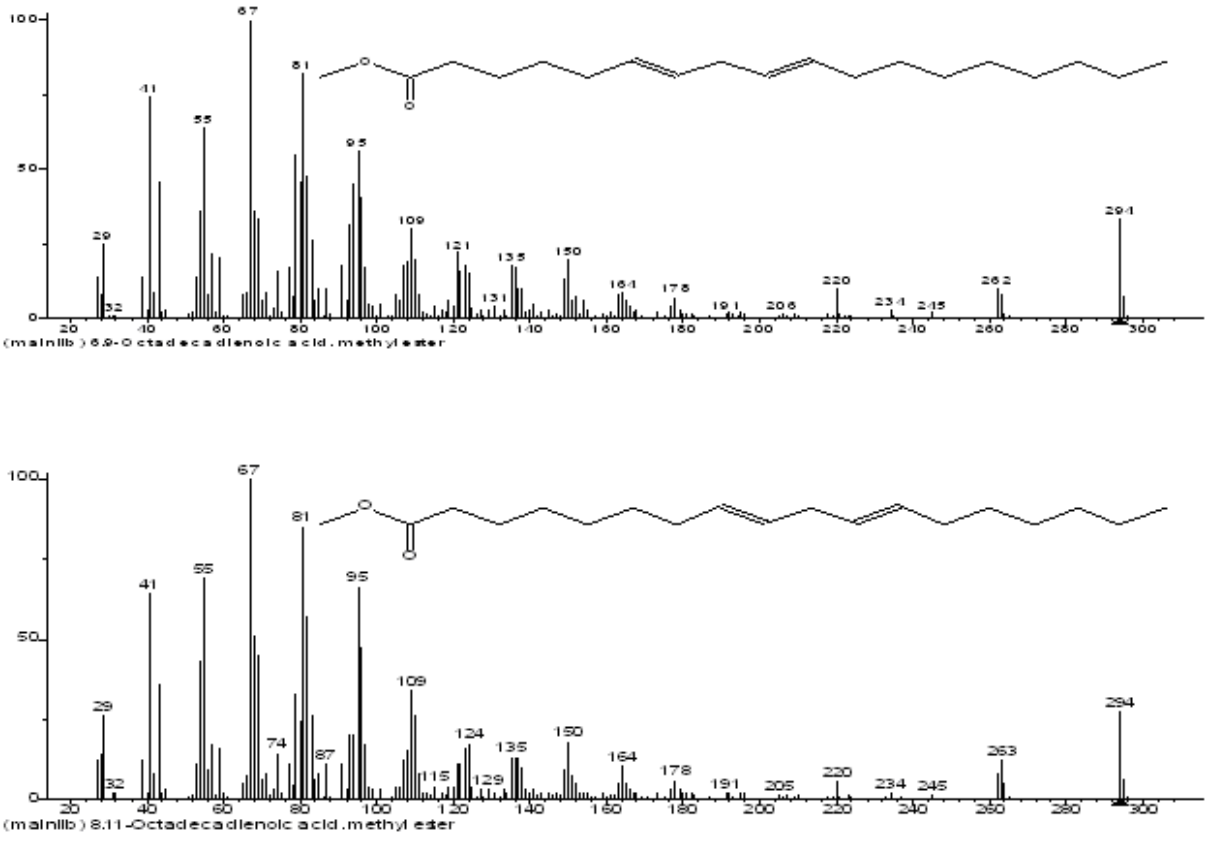
3.4. Zsírsav metil-észterek Kováts-indexének előrejelzése

3.4.1. Zsírsav metil-észterek

A zsírsav metil-észterek a GC-MS vizsgálatok fontos tárgyai közé tartoznak. A metil-észterek analitikai szempontból a biológiai eredetű, nem vagy kevéssé illékony trigliceridek bomlásakor keletkező karbonsavak azonosításánál jelentősek [*Wretensjö és mtsai, 1990; Spitzer, 1991; Liu és mtsai, 1996; Doneanu, 1998*].

3.4.1.1. ábra

A 6,9 oktadiénsav (a) és a 8,11 oktadiénsav (b) metil-észterének tömegspektruma



A növényi és állati mintákból származó vegyületek szénatomszáma általában 12-24 között mozog. Több alcsoportja van a metil-észtereknek, akad köztük egyszeresen és

többszörösen telítetlen, egyenes és elágazó láncú vegyület. Összesen néhány száz vegyületről van szó, de csak kisebb hányaduknak határozták meg a retenciós idejét kísérleti úton [NIST, 2005]. A vegyületek azonosításához az MS spektrumon kívül mindenképp szükség van retenciós adatra is, mert az izomerek spektruma sok esetben megkülönböztethetetlen egymástól. Ezt az esetet jól szemlélteti a 6,9 oktadiénsav és a 8,11 oktadiénsav metil-észtereinek tömegspektruma.

Azzal együtt, hogy a metil-észterek nem túl bonyolult szerkezetűek, többségüknek mégsem számították ki a retenciós indexét. Megemlítendő, hogy a zsírsav-metil észterek azonosításakor használatos egy másik, a retencióval kapcsolatos jellemző az ekvivalens lánchosszúság, melynek előrejelzésére több modell is született [*Mjos és Grahl-Nielsen, 2006 és a benne lévő hivatkozások*].

A számításokkal célom elsősorban az volt, hogy továbbfejlesszem a változókiválasztási módszerek összehasonlításának helyes módját, valamint, hogy működőképes QSRR modelleket hozzak létre zsírsav metil-észterek retenciós mechanizmusának jobb megértésére és az RI előrejelzésére. Célom továbbá 37 ismeretlen retenciós indexű metil-észter RI-jének számítása volt. Ezek különféle zsírokban, olajokban fordulnak elő, és azok összetételének elemzésekor segítséget nyújtnak az összetevők azonosításában.

3.4.2. A számítás menete

Összesen 130 db zsírsav metil-észter retenciós indexe állt rendelkezésemre. Az adatok részben az irodalomból [*NIST, 2005*], részben a szentpétervári egyetemen dolgozó kutatópartnerünk méréseiből származtak. A méréseket Biochrom-1 gázkromatográfiás készüléken hajtották végre, az alkalmazott állófázis OV-101 volt. A metil-észterek közt lineáris

és elágazásokat tartalmazó vegyületek is voltak, egyes ill. kettős kötéseket is tartalmaztak és több cisz-transz izomer pár is megtalálható volt.

154 deszkriptor került a kiindulási készletbe, ezek között voltak „nulladimenziós" deszkriptorok (pl. moláris tömeg, kettős kötések száma, forgatható kötések száma, stb.) valamint kétdimenziós topológiai deszkriptorok és flexibilitási indexek.

Az adatkészletet véletlen módon három részre osztottam; a molekulák 34%-a került a betanuló, 33%-a a kalibrációs és 33%-a a teszt készletbe. A teszt készleten külső validálást végeztem. A változókiválasztás eredményességét és a modellek leíróképességét a korrelációs együtthatóval, a reziduális szórással és F teszttel jellemeztem, valamint vizsgáltam a reziduumok lefutását is. Az előrejelző képességet a R^2_{teszt}-tel és a teszt készletre vonatkozó reziduális szórás számításával ellenőriztük. A modell hibájának torzítás részét leíró (ld. 12. egyenlet a 2.2.2. fejezetben) reziduális szórás mellett a varianciatagot megadó L_2 normát is kiszámítottuk.

Öt különböző változókiválasztási módszer alkalmaztunk, ezek a következők voltak: párkorrelációs módszer, előreirányuló változóbevonás, módosított legjobb alrendszer kiválasztása, részleges legkisebb négyzetek módszere és a Lasso módszer.

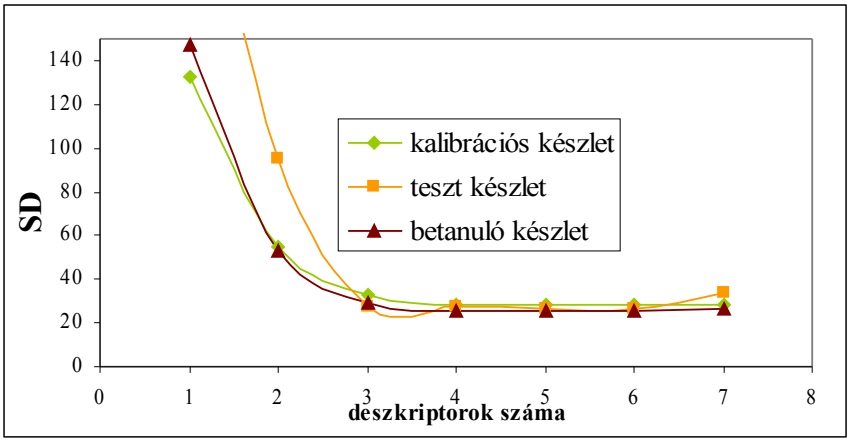
A modellépítést kétféleképpen hajtottam végre. Első esetben a deszkriptorok számát minden modellben 5-ben határoztam meg (mert az előreirányuló változóbevonás során, 5%-os szignifikanciaszint beállításakor 5 változót választott ki az algoritmus). A modellépítést többszörös lineáris regresszióval végeztem. A páronkénti korrelációs módszer esetén a sorbarakott változók közül az első 5-tel építettem MLR modellt. Az összes alrendszer regressziót hatalmas számításigénye miatt módosítottuk (mBSS). Kiválasztottam a legjobb három független változót tartalmazó modellt, majd ezeket a deszkriptorokat kihagyva hétszer lefuttattam a programot (természetesen mindig kihagyva az újonnan kapott legjobb 3 független változót), majd az így kigyűjtött 21 deszkriptorból válogattam ki a legjobbakat, 5-ben limitálva

a számukat. A PLS esetén számításakor kiszámítottuk a regressziós koefficiensek varianciáját is, és az 5 legjobb értékkel rendelkező deszkriptort tettem bele az MLR modellbe. A Lasso módszernél a regressziós koefficiensek nagysága szerint rangsoroltam a változókat és az első 5-öt használtam fel a modellben.

A második esetben, az ún. optimális modellezéskor a különböző módszereknél más-más utat kellett követnem. A párkorrelációs módszer csak rangsorolja a változókat, ezért az MLR-t használtam itt is regressziós módszernek. Különböző számú (1-7) deszkriptort tartalmazó modellek közül a betanuló készleten számított SD értékek (SD_{bet} és $SD_{kalibr)}$ alapján választottam ki az optimális számú (4) független változót tartalmazó modellt (ld. 3.4.2.1. ábrán). Az SD értékek 4-nél több változó bevonása után csak elhanyagolható mértékben csökkentek. A FS esetén az optimális modell értelemszerűen 5 változót tartalmaz. Az mBSS esetén hasonló utat követtünk, mint a PCM esetén, itt az optimális modell 7 változót tartalmazott. A PLS esetén 66 deszkriptor bizonyult szignifikánsnak Martens módszere szerint, a betanuló készleten végrehajtott keresztellenőrzés segítségével pedig megállapítottam, hogy az optimális PLS komponens szám hét. A modellt utána PLS-sel építettem. A Lasso módszer esetén a nullától lényegesen 24 deszkriptor regressziós koefficiense különbözött.

3.4.2.1. ábra

A független változók optimális számának kiválasztása a PCM módszer esetében



3.4.3. Eredmények és következtetések

3.2.3.1. A modellekben található független változók

A modellekbe bekerülő független változókat az 3.2.3.1.1. és a 3.2.3.1.2 táblázat tartalmazza. A deszkriptorok nevének magyarázata a Függelékben található, az F2 táblázatban.

A PCM módszer számos deszkriptort, pontosabban az egyszerű rendezés 115, a különbség- és a súlyozásos rendezés pedig 56-ot talált szignifikánsnak. Az első 5 legjobb deszkriptor mindhárom módszer esetén ugyanaz volt, méghozzá egy elektronegativitással súlyozott Wiener-típusú index, a teljes információtartalom indexe, a Z-súlyozott távolság mátrix legnagyobb sajátértéke, a molekulatömeg és a Broto-Moreau index.

Az optimális Lasso model megtalálása az L-görbe segítségével történt (ld. 3.2.3.1.1. ábra). Az ábrán kör jelöli az optimális modellt, ami 24 független változót tartalmaz, többek között a molekulatömeget, a forgatható kötések számát és a Kier-féle flexibilitási indexet.

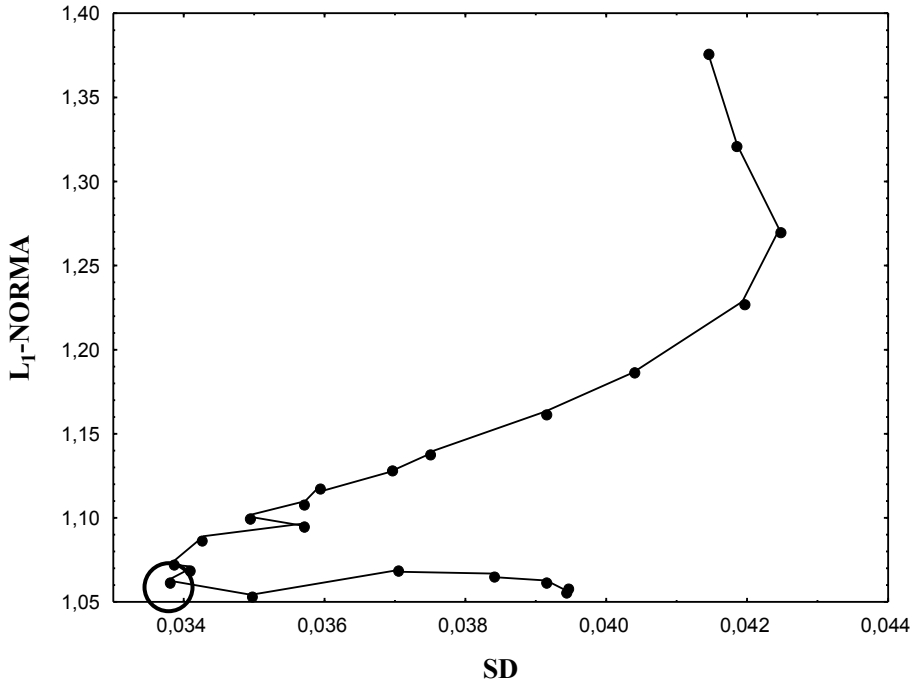
Az előreirányuló változóbevonás során a kettős kötések száma, a forgatható kötések száma, a Narumi és a négyzetes topológiai index, valamint még egy forgathatósággal kapcsolatos index találtatott szignifikánsnak.

A PLS a molekulatömeget, a Z-súlyozott távolság mátrix legnagyobb sajátértékét, valamint három szimmetriával kapcsolatos topológiai indexet választotta ki.

A módosított összes alrendszer módszer szerint a forgatható kötések száma, valamint topológiai indexek voltak azok a deszkriptorok, melyek a legszorosabb korrelációt mutatják az RI-vel.

3.2.3.1.1. ábra

Az optimális Lasso módszer kiválasztása



Megállapítható, hogy a zsírsav-metil észterek szerkezete jól jellemezhető topológiai indexek, valamint néhány nulladimenziós deszkriptor segítségével. A gázkromatográfiás retenció egy igen összetett folyamat, melyben többfajta intermolekuláris kölcsönhatásnak van szerepe. Az apoláris állófázisokon diszperziós erők a meghatározók. Esetünkben a molekulák poláris(abb) része, a metoxikarbonil fragmentum ugyanaz, ezért polaritással kapcsolatos jellemzőknek nincs kitüntetett szerepe az elválasztásban. A retenciós index leírásánál a topológiai és az ún. térkitöltési („bulkiness”) paramétereké a meghatározó szerep, melyek információt adnak a molekulák hosszáról és alakjáról, a szénlánc elágazásának mértékéről. A kettős kötések száma különbséget tesz az azonos lánchosszú és kb. azonos elágazási fokú vegyületek közt. A flexibilitási jellemzők és a forgatható kötések száma további megkülönböztetést tesz lehetővé. A molekulatömeg közismerten szoros összefüggésben áll a

3.2.3.1.1. táblázat

Ötváltozós modellek statisztikai paraméterei és a modellekben található deszkriptorok

Változó-kiválasztási módszer	Deszkriptorok a modellben	R^2_{bet}	F_{bet}	SD_{bet}	R^2_{kal}	F_{kal}	SD_{kal}	R^2_{teszt}	F_{teszt}	SD_{teszt}	L_2
PCM 5	Whete , **Eig1Z** , TIC0, MW, ATS1v	0,99837	4774	25,60	0,99781	3466	28,18	0,99773	37149	28,74	3590
Lasso 5	**RBN**, MPC10, Eig1Z , **MW** , PHI	0,99891	7176	20,88	0,99803	3844	24,29	0,99505	18710	42,44	49
FS 5	HNar, RBF, **nDB** , **RBN**, **QIndex**	0,99871	6018	22,80	0,99810	6018	26,26	0,99792	45365	25,58	880
mBSS 5	**RBN, piPC03,** PCD**, GATS3v,** QIndex	0,99933	11692	16,36	0,99847	4976	23,53	0,97381	4368	90,72	464
PLS 5	**MW**, Eig1Z, TIC3, **TIC4,** TIC 5	0,99762	3270	30,92	0,99764	3213	29,27	0,97511	4883	87,43	15

3.2.3.1.2. táblázat

Optimális modellek statisztikai paraméterei és a modellekben található deszkriptorok

Változó-kiválasztási módszer	Deszkriptorok a modellben	R^2_{bet}	F_{bet}	SD_{bet}	R^2_{kal}	F_{kal}	SD_{kal}	R^2_{teszt}	F_{teszt}	SD_{teszt}	L_2
PCM opt	**Whete , Eig1Z, TIC0, MW**	0,99834	6003	25,52	0,99772	4260	28,42	0,99764	36181	27,04	171
Lasso opt	24 deszkriptor, Ld. függelék	0,99840	26753	24,18	0,99868	31820	20,81	0,99831	31151	22,78	384
FS opt	HNar, RBF, **nDB**, **RBN**, **QIndex**	0,99871	6018	22,80	0,99810	6018	26,26	0,99792	45365	25,58	880
mBSS opt	**Whete, TIC0, Eig1z, piPC03, piID , PCD, GATS3v**	0,99979	12345	13,46	0,99846	3344	24,26	0,99786	47235	25,94	407
PLS opt	7 PLS komponens	0,99939	707000	14,88	0,99885	363000	19,48	0,99863	31180	20,54	14,2

retenció mértékével, ám az FS 5 modellből látszik, hogy szerepeltetése nem feltétlenül szükséges a jó előrejelzéshez, más vele korreláló, jobb deszkiptorok átveszik, felülmúlják szerepét.

A kalibrációs készleten történő paraméterbecsléskor nem minden deszkriptor volt szignifikáns. Ezeket a 3.2.1.1. táblázatban vastag betűvel kiemeltem. Az egyes készletekre vonatkozó statisztikai paramétereket alsó indexszel láttam el (bet: betanuló, kal: kalibrációs, teszt: teszt készletre vonatkozik).

3.2.3.2. Egyenlő számú változót tartalmazó modellek

Az 3.2.3.1.1 táblázat adataiból látható, hogy a modellek R^2_{bet} és F_{bet} értékei meglehetősen nagyok, köztük kicsi, (de szignifikáns) különbség található. Ez azt mutatja, hogy a különböző deszkriptorkombinációk többé-kevésbé ugyanazt a szerkezeti információt hordozzák és erősen korreláltak. A legjobb SD_{bet} és SD_{kal} értéket az mBSS 5 modell, míg a legrosszabbakat a PLS 5 modell esetében kaptuk.

A teszt készletre kapott statisztikai paraméterek vizsgálatakor kitűnik, hogy a legjobb R^2_{teszt} és SD_{teszt} értéket a FS 5 és a PCM 5 modellekre kaptuk. Az SD_{teszt} érték csak ennél a két modellnél elfogadható. A nagy SD_{teszt} érték a Lasso 5, az mBSS 5 és a PLS 5 modellek esetén azt jelzi, hogy ezek nem alkalmasak retenciós index előrejelzésre. A Lasso és a PLS módszer rosszabb teljesítményét az okozta, hogy öt deszkriptor igen messze van attól, amit optimálisnak nevezhetünk ezeknél a módszereknél (ld. 3.2.3.3. fejezet, melyben bemutatom az optimális PLS és Lasso modelleket).

Az euklideszi norma értékek vizsgálatakor fordított trend figyelhető meg. Amelyik modellnél az SD_{teszt} érték alacsony volt, ott az L_2-norma értéke nagy. A legrosszabb értéket a PCM 5 modellnél kaptam, legjobb a PLS 5 és a Lasso 5 modellé lett. Az előzetes várakozással ellentétben az mBSS módszer ebben a formájában nem volt hatékony a

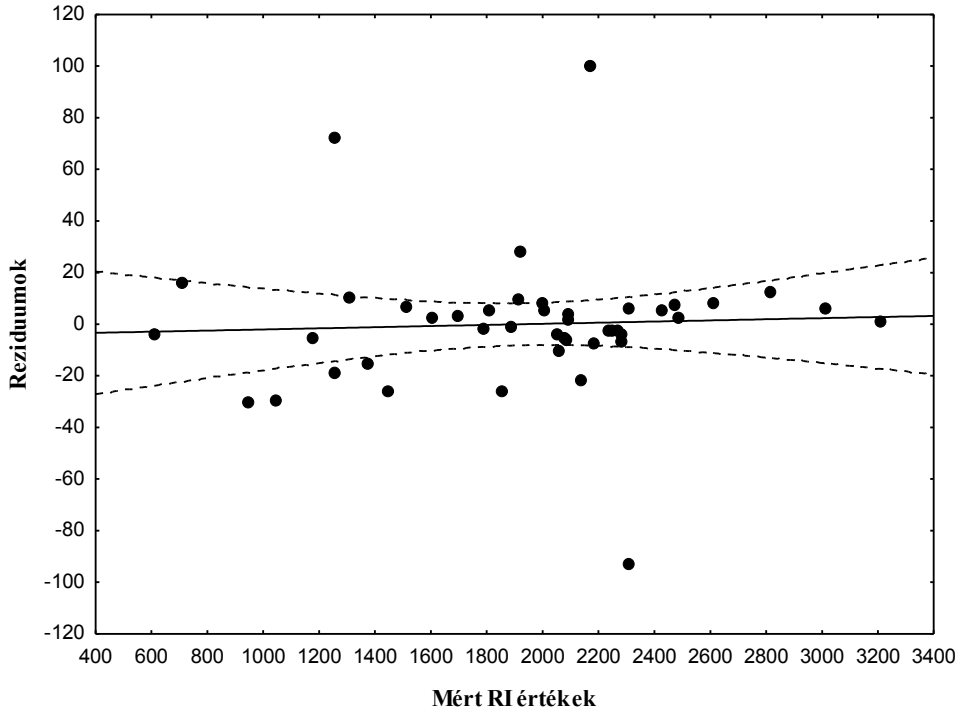
változókiválasztásban, a számítási ideje pedig még ebben az egyszerűsített formájában is jóval több, mint a többi módszeré. Szokatlan, hogy az FS 5 adja a legjobb modellt, mivel ismert, hogy a FS sok változó esetén kevéssé hatékony, gyakran lokális minimumot talál meg. Ez a példa is azt erősíti meg, hogy modellezéskor nem szabad egy módszerrel dolgozni, hanem érdemes többet kipróbálni és az eredményeket összehasonlítani.

Elvégeztem a modellek reziduumvizsgálatát, valamint a számított retenciós indexek ábrázolását a mért retenciós indexek függvényében. A reziduumok ábrája néhány kiugró értéket leszámítva adekvát képet mutatott. Mivel minden modell esetén azonosak voltak a kiugró értékek (3 db), valószínűsíthető, hogy a mérési adatokban volt a hiba. Ezt alátámasztja az is, hogy a kiugró értékű molekulákkal homológnak tekinthető kisebb vagy nagyobb szénatomszámú vegyületek nem „lógtak ki” a modellből.

Példaként a PCM 5 modell reziduumábráját mutatom be (3.2.3.2.1. ábra)

3.2.3.2.1. ábra

A PCM 5 modellel számított RI értékek a reziduumok függvényében



3.2.3.3. Optimális számú változót tartalmazó modellek, és a kétfajta modellezés összehasonlítása

Az optimális modellek statisztikai paraméterei a 3.2.3.1.2. táblázatban láthatók. Az SD_{bet}, SD_{kal} és SD_{teszt} értékek egyaránt a PLS opt modell esetében voltak a legjobbak, de az összes modell esetén elfogadhatók. A PLS modell jó eredménye nem meglepő, ha figyelembe vesszük, hogy a 7 pls komponens 66 deszkriptorról hordoz információt. A Lasso opt modell adta a második legjobb SD_{teszt} értéket, ahol a független változók száma szintén nagy a többi modellhez képest.

A PCM modell négy független változót tartalmaz. Az mBSS modell esetén a deszkriptorok optimális számának 7-et választottam, több változó bevezetése már nem csökkentette számottevően az SD_{bet} és az SD_{kalibr} értékét.

Az L_2-norma értéke a PLS opt modell esetén volt a legjobb. Érdekes módon nem a Lasso modell az, ahol az L_2-norma értéke a második legalacsonyabb, hanem a PCM opt modell.

A könnyebb összehasonlíthatóság érdekében a modellek előrejelző képességének hibáját leíró két statisztikai paramétert, az SD_{teszt}-et és az L_2-normát két egymás alatti ábrán (3.2.3.3.1. a és b) tüntettem fel, és egymás mellett ábrázoltam az öt változót tartalmazó és optimális modelleket. A PCM modellek esetén mindkét modell SD_{teszt} értéke jó volt, ráadásul emellett PCM opt modell L_2-norma értéke is igen kedvező volt. Érdekes, hogy míg a PCM 4 modellben minden deszkriptor szignifikáns volt a kalibrációs készleten is, a mindössze egy plusz változó bevonásával kapott PCM 5 modellben már csak egy deszkriptor volt az. A PCM opt modell nemcsak a jó statisztikai paraméterei miatt tekinthető megfelelő előrejelző modellnek, hanem kis változószáma miatt könnyen értelmezhető.

Megállapíthatjuk, hogy a PLS oly módon történő alkalmazása, hogy kizárólag változókiválasztásra használjuk és MLR-rel építünk modellt, nem kifizetődő retenciós indexek előrejelzése esetén. Az optimális modell kiváló eredményeket adott, nem véletlen a PLS módszer népszerűsége. Nem szabad azonban megfeledkezni arról, hogy az optimális modell valójában 66 független változót használ fel, a nagy deszkriptorszám és a látens változók miatt a modell gyakorlatilag nem értelmezhető.

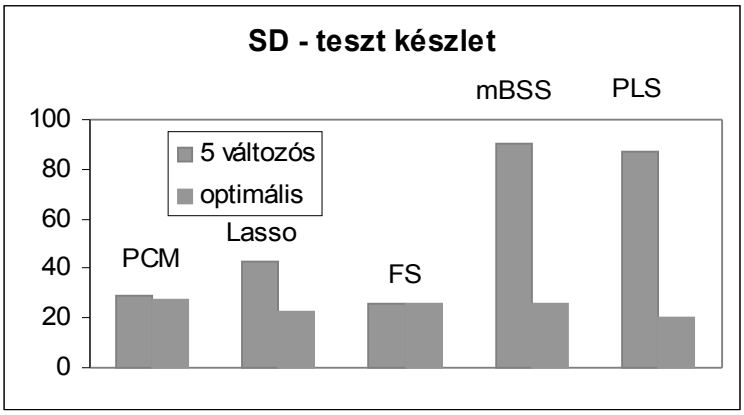
A Lasso mószerről hasonló megállapításokat tehetünk, mint a PLS-ről. Változókiválasztásra önmagában nem érdemes használni, a 24 változót tartalmazó optimális modell fizikai interpretációja pedig meglehetősen nehézkes.

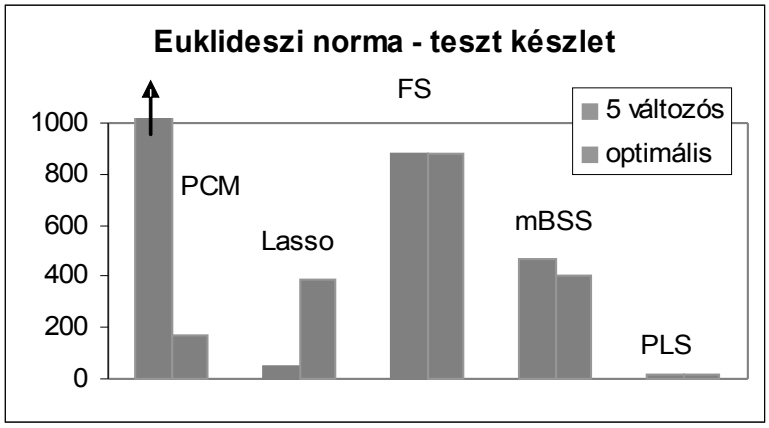
A hét változót tartalmazó mBSS modell megfelelő statisztikai paraméterekkel rendelkezik, jól értelmezhető, ám számításigénye még ebben a leegyszerűsített formájában is joval nagyobb, mint a többi módszeré.

A FS egyszerűen értelmezhető modellt adott, melynek SD teszt értéke megfelelő. Bár az L_2-norma értéke a többi módszerhez viszonyítva nagy, egyszerűsége és gyorsasága mégis igazolja használhatóságát.

3.2.3.3.1. ábra

Az SD és az L_2-norma összehasonlítása ötváltozós és optimális modellek esetén





3.2.3.4. Ismeretlen retenciós indexek számítása

A felépített modellek segítségével lehetővé vált olyan metil-észterek retenciós indexének kiszámítása, melyekre nem található adat az irodalomban. A Függelékben található F3 táblázatban levő 37 vegyület retenciós indexét jeleztem előre. Az előrejelzéshez ún. „konszenzusmodellezést" vagy egyetértési modellezést alkalmaztam, több modellel számítottam ki a retenciós indexek értékét, majd a kapott eredményeket átlagoltam. A három legjobb modellt választottam ki az előrejelzéshez, az FS 5 (FS opt), a PCM opt és a PLS opt modelleket.

3.4.4. Összefoglalás

- Sikerült zsírsav metil-észterek retenciós mechanizmusának leírására és retenciós index előrejelzésére alkalmas modelleket létrehoznom. A modellek könnyen értelmezhetők, csak kétdimenziós deszkriptorokat tartalmaznak, ezért a számítások rövid idő alatt elvégezhetők.
- A zsírsav metil-észterek példáján igazoltam, hogy minimálelv alkalmazásakor a Lasso és a PLS módszer nem vezetnek jó eredményre. Az MLR modellek az optimális modellezés esetén is praktikusabbak olyan szempontból, hogy kevesebb változót tartalmaznak.

- Igazoltam, hogy jelen esetben az MLR modellek nem működnek jól olyan változókkal, amiket PLS vagy Lasso módszer választott ki.

- Bizonyítottam, hogy a párkorrelációs módszer kiválóan alkalmazható változókiválasztásra metil-észterek RI előrejelzésekor.

- Megállapítottam, hogy ha a minimálelvet nem alkalmazzuk, a PLS módszer a legalkalmasabb az észterek retenciós indexének előrejelzésére.

- Összesen 37, eddig ismeretlen retenciós indexű zsírsav metil-észter Kováts-indexét számítottam ki, mely jelentősen könnyíti ezeknek a vegyületeknek az azonosítását.

- Megállapítottam, hogy a metil-észterek retenciós mechanizmusának leírásában fontos szerepet játszanak a topológiai deszkriptorok és egyéb terjedelmet jellemző deszkriptorok valamint a flexibilitást jellemző sajátságok.

Irodalmak a 3.fejezethez

Bergman G., Götze H. J., Hermann A., Zinn, P. Application of target factor analysis to gas chromatography. Reproduction, prediction and classification *Chromatographia* **1991**, 32, 259-264.

Bermejo J., Guillén M. D. Prediction of Kováts retention index of saturated alcohols on stationary phases of different polarity *Anal. Chem.* **1987**, 59, 94-97.

Doneanu C., Radulescu V., Efstatiade D., Rusu V., Covaci A. Capillary GC/MS characterization of fatty acids from indigenous silkworm oil *J. Microcol. Sep.* **1998**, 9 (1), 37-41.

Draper N.R., Smith H. *Applied Regression Analysis*, John Wiley & Sons Inc.: New York, **1998**.

Frank I. E., Friedman J. H. A Statistical view of some chemometrics regression tools. *Technometrics* **1993**, 35, 109-135.

Hastie T., Tibshirani R., Friedman, J. Discussion: A comparison of the selection and shrinkages methods In *The Elements of statistical learning. Data mining, interference and prediction* Springer – Verlag: New York, 2001; pp 68-75.

Guo W., Li Y., Zheng X. M. The predicting study for chromatographic retention index of saturated alcohols by MLR and ANN *Talanta* **2000**, 51, 479-488.

Junkes B. S., Amboni R. D. M.C., Yunes R. A., Heinzen V. E. F. Prediction of the chromatographic retention of saturated alcohols on stationary phases of different polarity applying the novel semi-empirical topological index *Anal. Chim. Acta* **2003**, 477, 29-39.

Kováts E. Gas-chromatographische Charakterisierung organischer Verbindungen. Teil 1 : Retentionindices aliphatischer Halogenide, Alkohole, Aldehyde und Ketone *Helv. Chim. Acta* **1958**, 41, 1915-1932.

Liu Y.D., Longmore R.B., Fox J.E.D. Separation and identification of ximenynic acid isomers in the seed oil of *Santalum spicatum* as their 4,4-dimethyloxazoline derivatives *J. Am. Oil Chem. Soc.* **1996,** 73 (12), 1729-1731.

Mjøs S.A., Grahl-Nielsen O. Prediction of gas chromatographic retention of polyunsaturated fatty acid methyl esters *J. Chromatogr., A* **2006**, 1110, 171-180.

NIST MS Data Center. NIST Standard Reference Database Number 69, June 2005. NIST, Gaithersburg, MD 20899 (http://webbook.nist.gov)

Pias J. B., Gasco L. *J. Chromatogr.* **1975**, 104, 1 D 14 Table 885.

Spitzer GC-MS characterization (Chemical lonization and electron impact modes) of the methyl esters and oxazoline derivatives of cyclopropenoid fatty acids *J. Am. Oil Chem. Soc.* **1991**, 68 (12), 963-969.

Zhang X., Lu, P. Unified equation between Kovats indices on different stationary phases *J. Chromatogr., A* **1996**, 731/1-2, 187-199.

Wretensjö I., Svensson L., Christie W.W. Gas chromatographic–mass spectrometric identification of the fatty acidsin borage oil using the picolinyl ester derivatives *J Chromatogr.* **1990**, 521 89-97.

Wold, S. Discussion: PLS in chemical practice. *Technometrics* **1993**, 35, 136-139.

4. HETEROCIKLUSOS VEGYÜLETEK RETENCIÓS INDEXÉNEK LEÍRÁSA ÉS ELŐREJELZÉSE

A 3. fejezetben bemutatottaktól eltérően a heterociklusos vegyületek vizsgálatakor elsődleges célom a kén- oxigén- és nitrogéntartalmú heterociklusos molekulák Kováts-féle retenciós indexének előrejelzése volt, másodlagos feladatomnak tekintettem az alkalmazott (nem változókiválasztási) módszerek hatékonyságának összehasonlítását.

A telített gyűrűs vegyületek retenciós tulajdonságainak modellezésekor több problémába ütközünk, melyek más vegyületcsoportoknál (pl. alkánok, alkének, poliaromás szénhidrogének) nem jelentkeznek. Az egyik problémát a gyűrűk specifikus termodinamikai sajátságai okozzák; ez jól jellemezhető azzal, hogy a retenciós indexek különbsége az ugyanannyi szénatomot tartalmazó egyenes láncú ill. ciklikus alkánok és a heterociklusok közt nem állandó, hanem a szénatomszám növekedésével nő [*Golovnya és Grigorjeva, 1986*]. További probléma, hogy a hattagúnál nagyobb gyűrűk számos konformációban létezhetnek a gyűrű méretétől, valamint szubsztituensek számától és helyzetétől függően. Ugyancsak nehézséget jelent, hogy számtalan izomerjük létezik, számuk a szénatomszám növekedésével sokkal gyorsabban nő, mint a nyílt láncú vegyületek esetében, valamint az, hogy számos heterociklusos vegyület szintézise során egyszerre keletkezik a cisz és transz izomer ismeretlen arányban.

Az irodalomban additív sémákon alapuló retenciós index számításokra találtam példát heterociklusos vegyületek esetén. Létezik egy nemlineáris egyenletet, ami tartalmazza a molekula szénatomszámát, valamint különféle koefficienseket, amelyek az elválasztásnál alkalmazott hőmérséklettől és egyéb GC-beállításoktól függenek [*Golovnya és Grigorjeva, 1986*]. Ez az összefüggés jól használható retenciós index számítására alkoholok, észterek, aminok és aromás kéntartalmú homológ vegyületsorok esetében. Az általam vizsgált molekulák

szerkezete azonban ezeknél jóval összetettebb. Néhány példa található még az irodalomban kén- illetve nitrogéntartalmú policiklusos aromás szénhidrogének retenciójának számítására topológiai deszkriptorok alkalmazásával [*Hu és mtsai, 2005; Schade*, *2006*], valamint olyan adatkészletek felhasználásával épített RI modellekre, amelyekben O, N vagy S atomot tartalmazó, de nem csak gyűrűs, hanem nyílt láncú vegyületek is előfordulnak [*Santiuste, 2000; Safa, 2005*]. További irodalmak fellelhetők nitrogéntartalmú heterociklusok retenciójának meghatározására folyadékkromatográfiás elválasztáskor [*Polyakova és mtsai, 2006*]

4.1 Heterociklusos vegyületek

Az vizsgált adatkészlet 164 molekulát tartalmazott, melyből 54 oxigén-, 85 nitrogén-, 29 pedig kéntartalmú volt. A szénatomok száma a gyűrűkben 3 és 9 között váltakozott.

A vegyületek retenciós adatai az irodalomból származnak [*NIST, 2005; Saedtler, 1986*]. Előfordult, hogy egy vegyületről több retenciós adat is rendelkezésre állt, ez esetben az átlaguk került az adatsorba.

4.2. Deszkriptorok

A molekukák szerkezetén a háromdimenziós deszkriptorok használata miatt geometriaoptimálást hajtottam végre, és ezzel együtt kiszámítottam a vegyületek totálenergiáját is. A modellépítés során három különböző deszkriptorkészlettel dolgoztam. A készleteket *A*, *B* és *C* jelöléssel láttam el. Az *A* csoport a korábban az alkoholok retenciójának előrejelzésekor már sikerrel alkalmazott WHIM és nulladimenziós, a *B* a GETAWAY deszkriptorokat, a *C* csoport pedig topológiai indexeket, molekulafragmentumokat, funkciós csoportokat és molekuláris utak számát jellemző deszkriptorokat tartalmazott. Minden deszkriptorcsoporttal kétféleképpen végeztem el a számítást. Először csak a fent megadott független változókat

használtam a modellezéskor, majd minden csoportba bevettem további független változónak a forráspontot (ezeket a modelleket FP jelöléssel láttam el). A forráspont a retenciót egyik legjobban leíró fizikai-kémiai sajátság [*Héberger, 1989; Héberger, 1990; Zenkevich, 1997; Zenkevich, 1998*]. Az FP modellekbe a totálenergiát is bevettem, azaz azt az energiamennyiséget, ami ahhoz szükséges, hogy egy molekulának az elektronjait és atommagjait egymástól végtelenre eltávolítsuk. Ez a deszkriptor a molekulatömeggel (melyről tudjuk, hogy szoros összefüggésben áll a retenció mértékével) ellentétben izomerek esetén is különböző eredményt ad. A forráspont minden esetben irodalmi adat volt [*Beilstein, 1986*], a totálenergiát pedig AM 1 szemiempirikus módszerrel számítottam.

4.3 Számítások

Az adatkészletet három részre osztottam; a vegyületek 44%-a a betanuló, 44%-a a kalibrációs és 12%-a a teszt készletbe került. A változókiválasztásra először BSS-t használtam (legnagyobb R^2-tel rendelkező modellt választottam), a független változók számát 4-ben határoztam meg, majd MLR-rel építettem a modelleket. A változókiválasztás a betanuló, a modell paramétereinek becslése a kalibrációs, a külső validálás pedig a teszt készlettel történt.

A változókiválasztást és a paraméterbecslést PLS-sel is elvégeztem. Itt azokat a deszkriptorokat válogattam be, melyek regressziós koefficiense a legnagyobb volt. Az *A* deszkriptorkészlet esetén 2000, a *B* és *C* készlet esetén 500 feletti értékű változókat válogattam be. Ezután a modellépítés és az előrejelzés több elem kihagyásos keresztellenőrzéssel történt a kalibrációs és teszt készlet felhasználásával. A PLS komponensek számát az határozta meg, hogy az SD_{cv} hol volt a legalacsonyabb. Ez a minimum az *A* deszkriptorcsoportot használó modell esetén 8, a *B* csoport esetén 9, a *C* csoport esetében pedig 11 PLS komponens felhasználásakor adódott.

A modellek jóságát az MLR modellek esetén az R^2, SD, R^2_{teszt} és SD_{teszt} paraméterek, PLS modellek esetén az R^2_{CV} és SD_{CV} paraméterek, valamint a vizsgált vegyületek számított és mért RI ábráinak alapján állapítottam meg.

4.4. Az MLR és a PLS modellek értékelése

A 4.4.1. táblázat tartalmazza a számított MLR modellek statisztikai paramétereit. Látható, hogy ezek minden esetben jobbak, azaz a betanuló és a kalibrációs készletre vonatkozó R^2, valamint a R^2_{teszt} értékek nagyobbak, az SD értékek pedig mindhárom készletre vonatkozva kisebbek, abban az esetben, amikor a forráspont is szerepel a modellben.

A legnagyobb R^2_{kal} és legkisebb SD_{kal} érték az MLR *A* FP modellnél található, ezt követi az MLR *C* FP és a MLR *B* FP modell. Az R^2_{kal} és a SD_{kal} értékek közt ennél a 3 modellnél nem található jelentős különbség, ami arra utal, hogy az *A*, *B* és *C* készletben található deszkriptorok, feltehetőleg a forráspont miatt a retenció leírására többé-kevésbé egyformán alkalmasak. Ezeknek a statisztikai paramétereknek értéke elfogadható. Az R^2_{teszt} értékek szintén jónak tekinthetők, a legmagasabb az MLR *B* FP modellé. Továbbá, ugyanennek a modellnek a legkisebb az SD_{teszt} értéke. Az MLR *A* BP és az MLR *C* BP modellek SD_{teszt} értéke viszont meglehetősen nagy. Megállapítható, hogy ez utóbbi két modell, kiváló leíróképessége ellenére sem alkalmazható előrejelzésre. Egyedül a GETAWAY deszkriptorokat és a forráspontot is tartalmazó modell képes megfelelően leírni és előrejelezni a vizsgált heterociklusos vegyületek retenciós indexét.

A 4.4.1. ábrán az MLR *B* FP modell alapján számított retenciós index értékei láthatók a kísérleti értékek függvényében. A diagramon (egyébként mindhárom modell esetén) látszik egy enyhe konvex görbület, a tengelymetszet pedig additív hibát tartalmaz. A görbület mértéke és az additív hiba az MLR *B* BP modell esetén elfogadhatóan kicsi.

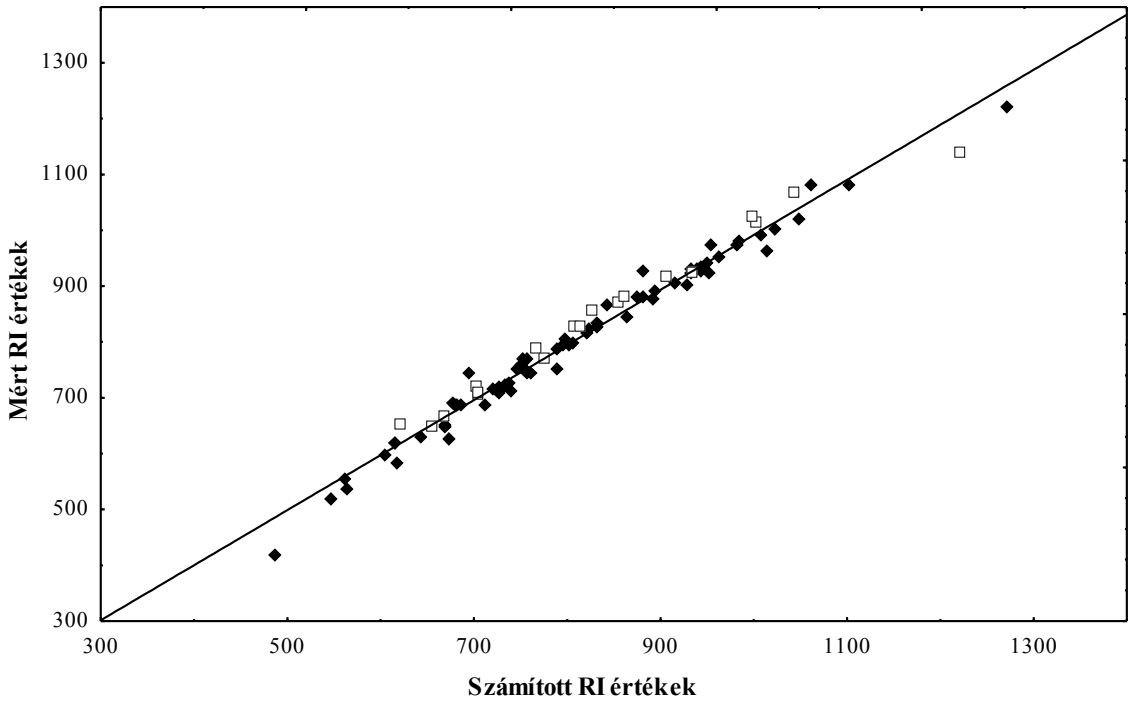
4.4.1. táblázat

Az MLR modellek statisztikai paraméterei

Modell	Deszkriptorok	R^2_{bet}	F_{bet}	SD_{bet}	R^2_{kal}	F_{kal}	SD_{kal}	R^2_{teszt}	SD_{teszt}
MLR *A* FP	BP, TE, DP14,	0,9910	1847,6	16,05	0,9916	1931,3	14,10	0,9284	44,25
MLR *A*	E3u MW, Ms, RBN,	0,9662	480,0	31,10	0,9437	272,8	36,60	0,9009	52,04
MLR *B* FP	SP07 BP, TE, H5m,	0,9923	2148,1	14,89	0,9888	1442,0	16,30	0,9692	29,05
MLR *B*	R2e+ MW, H8m, R3v,	0,9859	1171,9	20,11	0,8925	134,9	50,60	0,8107	71,93
MLR *C* FP	R4v+ BP, TE, GMTIV,	0,9914	1922,6	15,72	0,9899	1599,3	15,48	0,9452	38,40
MLR *C*	X2Av Jhetv, Jhete, X0Av, X1v	0,9572	374,8	35,03	0,9535	333,5	33,27	negatív	negatív

4.4.1. ábra

Az MLR *B* FP modell számított retenciós index értékei láthatók a kísérleti értékek függvényében



A forráspontot nem tartalmazó MLR modellek R^2_{bet} és R^2_{kal} értékei alapján megállapítható, hogy összefüggés létezik a retenciós indexek és a használt deszkriptorok közt, de ez jóval kevésbé szoros, mint a forráspontot is tartalmazó modellek esetében. Az SD_{teszt} értékek igen nagyok, így a modellek megbízható előrejelzésre alkalmatlanok. Érdekes, hogy közülük az MLR *C* a legjobb leíróképességű modell, de a validáláskor negatív R^2_{teszt} értéket kaptam eredményül. A rossz eredménynek oka lehet például az, hogy a kiválasztott deszkriptorok közt nincs olyan, ami a molekulák méretét jellemezné.

Megállapítható, hogy a 3D deszkriptorokat tartalmazó modellek jobb leírást adtak, mint a kétdimenziósak. Ez valószínűleg azért van így, mert a vizsgált heterociklusos vegyületek különféle konformációkat vehetnek fel, és ez a térszerkezet nagyban befolyásolja a retenciót. Ezek a 3D, jelen esetben a GETAWAY deszkriptorok sem elégségesek

önmagukban ahhoz, hogy megfelelő leírást adjanak a retencióról, hanem szükséges a forráspont beillesztése a modellbe.

Ha az egyes modellek statisztikai paramétereit más szemszögből hasonlítom össze, további megállapításokhoz juthatok. Látható, hogy az R^2_{bet} és R^2_{kal}, valamint az SD_{bet} és az SD_{kal} értékek közt nincs jelentős különbség (kivéve az MLR *B* modell esetében), ez a modellek stabilitását mutatja. A modellbe bekerült deszkriptorok megbízhatósága mind a hat modell esetében 99,9% feletti érték volt. Ezeket az értékeket összehasonlítottam a kalibrációs készlettel kapott modellek esetében. Az egyes deszkriptorok megbízhatósága 90% körülire csökkent le a forráspontot is tartalmazó modellek esetében, pl. a TE, azaz a totálenergia az MLR *A* BP modellnél 92%-ra, a TE az MLR *B* FP modellnél 90%-ra, és az X2Av az MLR *C* FP modellnél szintén 90%-ra. Ezek az értékek még elfogadhatónak tekinthetők. Viszont az R2e+ deszkriptoré 54%-ra csökkent az MLR *B* FP modell esetében, az MLR *C* FP modellnél pedig a TE megbízhatósága csökkent mindössze 3(!)%-ra. Ez azt mutatja, hogy ezeket a deszkriptorokat nem szükséges a modellbe építeni (ronthatják is annak jóságát) valószínűleg az adatkészlet felosztásából adódóan kerültek bele. Ez utóbbi két deszkriptor nélkül újra elvégeztem a paraméterbecslést, és azt az eredményt kaptam, hogy nem különböznek jelentősen az előzőktől, de Occam borotvájának elvét figyelembe véve mindenképp egyszerűbb, használhatóbb modellhez jutottam. Az újraszámolt modellek statisztikai paraméterei: MLR *B* FP: $R^2_{kal} = 0{,}9881$, $SD_{kal} = 16{,}70$; MLR *C* FP: $R^2_{kal} = 0{,}9899$, $SD_{kal} = 15{,}36$.

A PLS-sel történő változókiválasztás nem vezetett használható eredményre. A PLS módszer egyik esetben sem választotta ki az előzetes tudásom szerint mindenképp fontos szerepet játszó deszkriptorokat, azaz az FP-t vagy a molekulatömeget a kiindulási készletből. PLS módszer alkalmazásakor alakkal és hozzáférhetőséggel összefüggő WHIM deszkriptorok

kerültek be a modellekbe, valamint egy nulladimenziós deszkriptor az *A* csoportból míg a *C* csoportból elágazási és információs indexek kerültek be.

4.4.2. táblázat

A PLS modellek eredményei

Modell	PLS komponensek száma	Deszkriptorok száma	R^2_{CV}	SD_{CV}
PLS *A* FP	8	13	0,9856	22,38
PLS *B* FP	9	14	0,9781	29,01
PLS *C* FP	11	14	0,4195	166,69

A 4.4.2. táblázat mutatja a modellekbe bekerülő deszkriptorok számát, az azokból képzett PLS komponensek számát, valamint a statisztikai paramétereket. A modellépítést úgy végeztem el, hogy minden modellbe utólag bevettem a forráspontot és a totálenergiát. A nagy R^2_{CV} és alacsony SD_{CV} értékekből látszik, hogy a PLS FP modell alkalmas a retenciós index előrejelzésére. Ezek az értékek az MLR modellek ugyanezen értékeinél is jobbak (bár közvetlenül nem hasonlíthatók össze, mivel másképp végeztem a modell validálását). Nem szabad azonban megfeledkezni arról, hogy ez a modell jóval több független változót tartalmaz, mint az MLR modellek. A számított és a mért értékek ábrája is azt mutatja, hogy a model megfelelő, a meredekség értéke közel egy, a tengelymetszeté pedig nulla. A PLS *B* FP modell esetében is megfelelőek az R^2_{CV} és SD_{CV} értékek. A PLS *C* modell nem használható retenció előrejelzésére, az SD_{CV} túl nagy, az R^2_{CV} túl kicsi.

Azok a PLS modellek, amik nem tartalmazzák a forráspontot és a totálenergiát, rossz eredményt adnak. A PLS *A* modell esetében nagyon alacsony R^2_{CV} értéket (0,3818) és nagyon

magas SD értéket kaptam. A PLS *B* és *C* modellek esetében negatív R^2_{CV} értéket kaptam, így ezek a modellek alkalmatlanok az előrejelzésre.

4.5. Összefoglalás

- Sikerült megbízható modelleket építeni heterociklusos vegyületek retenciós indexének előrejelzésére. Ezek alkalmasak új vegyületek RI-jének számítására és az irodalmi adatok tesztelésére, adatbázisok szűrésére.
- Megállapítottam, hogy a forráspont mindenképpen szükséges, de önmagában nem elég a heterociklusos vegyületek retenciójának leírásához. A modelleket ki kell egészíteni számított deszkriptorokkal. A háromdimenziós független változók, mint pl. a GETAWAY deszkriptorok hasznosak az előrejelzésben, de az előrejelzés kétdimenziós deszkriptorokkal is megvalósítható.
- Igazoltam, hogy a GETAWAY deszkriptorok alkalmazhatók az RI-előrejelzésben. Ezek a deszkriptorok jobbnak bizonyultak a WHIM deszkriptoroknál.
- A PLS a szokásos tapasztalattal ellentétben jelen esetben sem bizonyult alkalmasnak változókiválasztási módszernek. Ha az általa kiválasztott vegyületeket kiegészítjük a forrásponttal, korrekt, előrejelzésre alkalmas modelleket kaphatunk.
- Megállapítottam, hogy jelen esetben a BSS hasznosnak bizonyult a változókiválasztásnál. Összességében a legjobb eredményt a két különböző módszer, azaz a BSS és a PLS kombinálása jelentette.

Irodalmak a 4. fejezethez

Beilsteins Handbuch der Organisher Chemie, Springer – Verlag, Berlin Heidelberg, 1986

Golovnya R.V., Grigoryeva D.N. Violation of the linearity principle of additivity of sorption energy in chromatography *J. High. Resolut. Chromatogr.* **1986**, 9, 584-589.

Jalali-Heravi M., Fatemi M.H. Prediction of thermal conductivity detection response factors using an artificial neural network *J. Chromatogr., A* **2000**, 897, 227-235.

Héberger K. Empirical correlations between gas-chromatographic retention data and physical or topological properties of solute molecules *Anal. Chim. Acta* **1989**, 223, 161– 174.

Héberger K. Discrimination between linear and non-linear models describing retention data of alkylbenzenes in gas-chromatography *Chromatographia* **1990**, 29, 375– 384.

Hu R., Liu H.X., Zhang R.S., Xue C.X., Yao X.Y., Liu M.C., Hu Z.D., Fan B-T. QSPR prediction of GC retention indices for nitrogen-containing polycyclic aromatic compounds from heuristically computed molecular descriptors *Talanta* **2005**, 68, 31-39.

Polyakova Y., Long M.J., Kyung H.R. Linear regression based QSPR models for the prediction of the retention mechanism of some nitrogen containing heterocycles *J. Liq. Chromatogr. Relat. Technol*. **2006**, 29, 533-552.

The Saedtler Standard Gas Chromatography Retention Index Library, Vol 1-4, Saedtler-Heyden, 1986

Safa F., Hadjmohammadi M.R., Use of topological indices of organic sulfur compounds in. quantitative structure-retention relationship study *QSAR Comb. Sci*. **2005**, 24, 1026-1032.

J.M. Santiuste Relationship between GLC retention data and topological indices for a wide variety of solutes on five stationary phases of different polarity *Chromatographia* **2000**, 52, 225-232.

T. Schade, J.T. Andersson, Speciation of alkylated dibenzothiophenes through correlation of structure and gas chromatographic retention indexes. *J. Chromatogr., A* **2006**, 1117, 206-213.

Zenkevich I.G. Quality control of GC retention indices data base for polymer sorbent Porapack Q *Process Control Qual*. **1997**, 9, 67–78.

Zenkevich I.G. Reciprocally unambiguous conformity between GC retention indices and boiling points within two- and multidimensional taxonomic groups of organic compounds *J. High Resolut. Chromatogr.* **1998**, 21, 565– 568.

5. ANTIDEPRESSZÁNS-JELÖLT VEGYÜLETEK OSZTÁLYOZÁSA hERG CSATORNA GÁTLÓ AKTIVITÁSUK ALAPJÁN

5.1 *A hERG K^+-csatornák és az antidepresszánsok okozta szívritmuszavar kapcsolata, antidepresszánsok hERG csatorna gátló aktivitásának előrejelzése – irodalmi összefoglaló*

Régóta ismert, hogy a pszichiátriában alkalmazott gyógyszerek szívritmuszavarokat, súlyosabb esetben akár hirtelen szívhalált is okozhatnak. Az egyre szaporodó megfigyelések, és az ezeket összefoglaló, valamint hatásmechanizmusra irányuló tanulmányok sora arra mutat, hogy az antidepresszánsok ezen mellékhatásai az új gyógyszerjelölt vegyületek szempontjából is meghatározó jelentőségűek [*Glassman 2001, Haddad 2002, DePonti 2000,* www.qtdrugs.org].

Munkám 2004-2005-ben készült, ezért az irodalmi áttekintés az eddig terjedő időszakot fogja át, ám a téma továbbra is igen aktuális, erre utal az attól kezdve megjelent publikációk sokasága [*Yoshida és Niwa, 2006; Seierstad, 2006; Coi és mtsai, 2006; Song és Clark, 2006; Leong, 2007; Du és mtsai, 2007*].

A ritmuszavar mechanizmusának hátterében a gyógyszerek által a *human ether-a-go-go* (hERG) gén által kódolt szív késői egyenirányító K^+ csatornára gyakorolt gátló hatás áll. Erre a csatornára, mely a szívritmus-szabályozó gyógyszerek célpontja, sok fajta és igen különböző szerkezetű vegyület is hathat [*DePonti, 2000, Haverkamp, 2000, Tamargo, 2000, Yap, 2000*]. A korábban már említett antidepresszánsokon kívül antihisztaminok és mikrobaellenes szerek eme K^+ csatornára gyakorolt hatását is leírták. A csatorna egy integrált membránfehérje, ezért a homológia modellek és a hagyományos dokkolási módszerek alkalmazhatósága az egyes molekulák kötődésének meghatározásában korlátozott. A hERG

gátló aktivitás kísérleti meghatározása időigényes és drága [*Cavero, 2001*], valamint az értéke nagyban függ a kísérlet körülményeitől (ezt a modellezésnél is figyelembe kell venni, ha különböző forrásokból dolgozunk). Ilyen körülmények között a hERG gátló aktivitás meghatározására a QSAR modellezés tűnik járható útnak [*Bains, 2004*].

A hERG K^+ csatornák szerkezetéről és a hozzájuk kötődni képes vegyületek jellegzetességeiről sok tanulmány született. Bains és munkatársai 124 változatos szerkezetű antipszichotikum és más hatású vegyületek tanulmányozása során állapították meg a következőket [*Bains, 2004*]: A hERG K^+ csatornához való kötődés mértéke pozitívan függ össze a molekulában található amin $N(R_1R_2R_3)$ csoport jelenlétével, ahol R_1=H vagy CH_2, R_2=CH_2 vagy $(CH_2)_2$, R_3=$(CH_2)_3$ vagy $(CH_2)_4$, legalább egy aromás gyűrű jelenlétével, öttagú, nitrogént tartalmazó heterociklus jelenlétével, deprotonálható csoport (főleg COOH) hiányával, valamint hidrogénkötést létesítő oxigén hiányával.

Ezt a leírást kiegészíti Mitcheson modellje és Pearlstein farmakofór modellje. [*Mitcheson, 2000, Pearlstein, 2003*]. Pearlstein és munkatársai szertindol analógokon és sokféle inhibitoron végeztek tanulmányokat, farmakofór és QSAR analízist is készítettek. Az általuk készített homológia modell azt sugallja, hogy a csatorna belseje aromás oldalláncokból áll, melyek közül kiemelt szerepe van a Phe656-nak és a Tyr652-nek. A kötődés a gyógyszermolekula és a csatorna között akkor valósul meg, ha az inhibitor hossztengelye párhuzamos a csatorna hossztengelyével. Ekkor a gyógyszermolekula hidrofób része (optimális esetben aromás gyűrű) π-kölcsönhatásba (π-stacking) kerül a Phe656 oldalláncal, a gyógyszeren található bázikus nitrogén pedig π-kationos kölcsönhatást hoz létre a Tyr652 oldallánccal.

Keserű megállapította [*Keserű, 2003*], hogy a kötődés erőssége a csatorna állapotától (aktivált vagy inaktív) is függhet, és emiatt bizonytalanságot jelent, ha csak egy farmakofór modellel dolgozunk. A bizonytalanságok kiküszöbölésére QSAR modellezést ajánl, ő maga

lépésenkénti lineáris regressziót végzett 68 antipszichotikum és egyéb gyógyszer felhasználásával. A számított log*P*-t, moláris refrakciót, parciális negatív töltésfelületet és két Wolsurf deszkriptort tartalmazó (W2 polarizálhatóság és D3 hidrofobicitás) modell leíróképessége jó, de előrejelző képessége csak mérsékelten mondható jónak ($R^2 = 0.5625$). A vegyületeken diszkriminancia-elemzést is végzett. Modellje az aktív vegyületek 83%-át és az inaktív molekulák 87%-át osztályozta helyesen. A modell alkalmazhatóságát gátolja, hogy a két Volsurf deszkriptor számítása nehézkes, ezért Keserű a hologram QSAR technikát és PLS módszert is felhasználta újabb, hatékonyabb modell építéséhez. A végső modell a potenciálisan aktív modelkulák 81%-át sorolta be helyesen és az inaktívoknak csak 18%-át sorolta be helytelenül.

Buyck és munkatársai [*Buyck, 2002*] fiziko-kémiai deszkriptorok felhasználásával és döntési fák segítségével jeleztek előre hERG gátló aktivitást. A végső modell három deszkriptort tartalmazott – számított log*P*-t, számított moláris refrakciót és a legbázisosabb nitrogén pK_a értékét. Megállapították, hogy savas csoport jelenléte és nagyobb topológiai felület esetén jobban kötődött a vegyület a hERG csatornához. Hasonló következtetésre jutott Roche és munkatársai [*Roche, 2002*], akik neuronhálók segítségével osztályoztak hERG blokkoló vegyületeket. A tesztvegyületek közül a nem blokkoló molekulák 93%-át, a blokkolóknak pedig 71%-át sikerült besorolni.

Aronov és munkatársai farmakofór modellezéssel és kétdimenziós topológiai deszkriptorok segítségével osztályoztak hERG K^+ csatorna blokkolókat és összehasonlították a kétfajta módszer hatékonyságát [*Aronov, 2004*]. A legjobb farmakofór modell a csatornához kötődni képes vegyületek 60%-át, a hERG K^+-hoz nem kötődőknek a 90%-át, az összes vegyületet tekintve pedig a 84%-át sorolta be helyesen a betanuló készletből, a teszt készletből pedig a 79%-ukat. A topológiai modell a molekulák 82%-át osztályozta helyesen, a blokkolóknak 47%-át, a nem blokkolók 91%-át sikerült megfelelően besorolnia.

Schneider kétdimenziós CAT (chemically advanced template search) módszerével a vegyületek 80%-át sorolta jó csoportba. Az inaktív vegyületek mindössze 9%-át sorolta be az aktívak közé, viszont az aktív vegyületek 59%-át inaktívnak mondta a modell. A topológiai és farmakofór modell kombinálásával készült újabb modelljében javult az aktív molekulák osztályozásának helyessége: 71%-ukat sikerült megfelelően besorolni, viszont a nem blokkoló vegyületek esetében az arány kissé rosszabb lett, csak 85%-ukat sikerült helyesen osztályoznia. Az összes vegyületet tekintve 82% volt a besorolás pontossága [*Schneider, 1999*] .

Megállapíthatjuk, hogy az irodalomban fellelhető modellek a felhasznált antipszichotikumok kb. 80%-át tudták helyesen osztályozni. Ha a modell helyesen sorolta be az aktív vagy az inaktív vegyületek több mint 90%-át, akkor a másik csoport elemeinek besorolása csak kevéssé volt sikeres. A hERG gátló aktivitás kódolásában fontos szerepet játszottak a gyógyszervegyületek aromás, hidrofób és bázikus tulajdonságait jellemző deszkriptorok.

A következő alfejezetekben a Laboratorios Del Dr. Esteve gyógyszercég (Avda. Mare de Déu de Montserrat, 221 - 08041 Barcelona, Spanyolország) megbízásából végzett vizsgálatokat ismertetem, melyek keretében antidepresszáns-jelölt molekulák osztályozására készítettem modelleket.

5.2. Antidepresszáns-jelölt vegyületek osztályozása hERG aktivitásuk alapján

5.2.1. Adatok

Az antidepresszáns-jelöltek molekulamechanikai módszerrel optimált szerkezetét és hozzájuk tartozó maradék hERG aktivitás kísérleti értékét az Esteve bocsátotta rendelkezésre. A számításokat az ő igényeik szerint két szakaszban végeztem. Először 502 vegyület

osztályozását, majd további 280-ét végeztem el. A molekulák a szelektív szerotonin újrafelvétel gátlók közé tartoznak, (5HT receptorcsaládra hatnak) szerkezetük igen változatos, szabadalmi okokból bemutatásuk nem lehetséges.

Az 502 vegyületet három csoportra osztottam fel hERG aktivitásuk alapján. Az egyes csoportba tartoztak azok az antidepresszáns-jelöltek, amelyek maradék hERG aktivitása 0,8-nál nagyobb és nincsen szívre irányuló káros mellékhatásuk. A hármas csoportba tartoztak azok a vegyületek, melyeknek maradék hERG aktivitása 0,5-nél kisebb volt - ezek blokkolták a hERG K^+ csatornát és nem kívánt mellékhatásokat okoztak. A második csoport egy átmeneti csoport volt, ide tartoztak azok a molekulák, melyek maradék hERG aktivitása 0,5 és 0,8 közé esett.

Az 502 molekulát két részre osztottam fel, a betanuló és kalibrációs készletbe az adatok kétharmada, azaz 335, a teszt készletbe 167 vegyület került.

5.2.2. Deszkriptorok és számítási módszerek

A Dragon program segítségével az osztályozni kívánt vegyületeknek 1481 szerkezeti sajátságát számoltam ki. Az erősen korreláló deszkriptorokat kiszűrtem, így számuk 849-re csökkent. A független változók között szerepeltek az irodalmi részben említett nulladimenziós, topológiai deszkriptorok, funkciós csoportok, Galvez topológiai töltés indexek, WHIM és GETAWAY deszkriptorok, aromás jelleget kifejező indexek, geometriai deszkriptorok, különféle molekulafragmentumok, valamint a log*P* és a moláris refrakció.

5.2.3 Modellek az első számítási folyamatban

Minden egyes modell két lépésben készült. A betanuló készlet segítségével végeztem a változókiválasztást, majd a szelektált deszkriptorok segítségével osztályoztuk a vegyületeket

hERG gátló aktivitásuk szerint. Az osztályozás jóságát a helyesen besorolt vegyületek százalékban megadott aránya jellemzi. Később a teszt készletbe tartozó molekulákat is osztályoztam a kész modellekkel. A legjobb modellek az alábbiak voltak; az osztályozás jóságát a betanuló készlet vegyületeire az 5.2.3.1., a tesztvegyületekre az 5.2.3.2. osztályozási mátrix táblázat mutatja be.

1. *CART - LDA*: Döntési fák segítségével történt a változókiválasztás, majd lineáris diszkriminancia elemzéssel az osztályozás.
2. *GA - LDA*: Genetikus algoritmussal történt a változókiválasztás, majd lineáris diszkriminancia elemzéssel az osztályozás.
3. *FS - LDA*: előreirányuló változóbevonással történt a változókiválasztás, majd lineáris diszkriminancia elemzéssel az osztályozás
4. *PLS DA*: Parciális legkisebb négyzetek módszerének diszkriminancia elemzéssel kombinált osztályozást végeztem, változókiválasztás nélkül.
5. *CART - PLS DA*: A deszkriptorok mennyiségét döntési fák felhasználásával csökkentettük, majd parciális legkisebb négyzetek módszerének diszkriminancia elemzéssel kombinált verziójával történt az osztályozás.
6. *FS - CART*: Előreirányuló változóbevonással segítségével történt a változókiválasztás, majd döntési fákkal az osztályozás.

A osztályozási mátrixok értékeiből látszik, hogy sikerült olyan hatékonyságú modelleket építeni, mint amilyeneket az irodalom tanulmányozása során megismertem. A legjobb modell az volt, amelyben a változókiválasztást genetikus algoritmussal hajtottam végre és lineáris diszkriminancia elemzéssel osztályoztam; itt az 1-es és a 3-as csoport elemeinek besorolása a betanuló és a teszt készlet esetében is 80%-os vagy afeletti eredménnyel valósult meg.

5.2.3.1. táblázat

Helyesen besorolt antidepresszánsok (%)

Betanuló készlet (n=335)

Változókiválasztás	***CART***	***GA***	***FS***	***nincs***	***CART***	***FS***
Osztályozás	***LDA***	***LDA***	***LDA***	***PLS DA***	***PLS DA***	***CART***
1. csoport	73,38	89,92	86,33	88,49	74,82	69,78
2. csoport	20,00	31,11	40,00	0,000	0,000	0,000
3. csoport	80,13	86,75	86,09	84,77	80,13	86,09
Összesen	**69,25**	**80,60**	**80,00**	**74,92**	**67,16**	**67,76**

5.2.3.2. táblázat

Helyesen besorolt vegyületek (%)

Teszt készlet (n=167)

Változókiválasztás	***CART***	***GA***	***FS***	***nincs***	***CART***	***FS***
Osztályozás	***LDA***	***LDA***	***LDA***	***PLS DA***	***PLS DA***	***CART***
1. csoport	78,08	79,45	78,08	83,56	80,82	65,75
2. csoport	9,091	4,54	18,18	0,000	0,000	0,000
3. csoport	76,39	87,50	83,33	81,94	80,55	83,33
Összesen	**68,26**	**73,05**	**72,46**	**71,86**	**70,06**	**64,67**

Feltűnő azonban, hogy a 2-es „átmeneti” csoport elemeinek besorolásánál csak igen szerény eredményt sikerült elérni, egyik modell sem tudja ezt a csoportot egyértelműen elkülöníteni az 1-es, illetve a 3-as csoporttól. Az irodalomban is csak két csoportra való felosztást tapasztaltam: hERG csatorna gátló és nem gátló vegyületekről volt szó. Ezért a következő számításokat elvégeztem úgy is, hogy a vegyületeket csak két csoportra osztottam fel. Eszerint a 0,80 feletti maradék hERG aktivitás értékkel rendelkező vegyületek tartoztak az inaktív csoportba (1), a 0,80 alattiak pedig az aktív csoportba (2).

Mivel 849 deszkriptort használtam, nagyon valószínű, hogy ugyanazt a sajátságot több független változó kódolja. Van azonban néhány olyan deszkriptor, ami több jó osztályozást eredményező modellben is előfordult, ilyenek a molekulatömeg, a kettős kötések-, a

karbonilcsoportok-, az alifás tioketon csoportok-, és az imidcsoportok száma, a geometriai távolságok összege az oxigén-oxigénkötésekben, számos topológiai deszkriptor, néhány GETAWAY és MoRSE deszkriptor.

5.2.4. Végső modellek három vegyületosztályra történő felosztásnál

A 782 vegyületből az ismert hERG aktivitású 502 molekula képezte a betanuló és kalibrációs készlet alapját. A 280 tesztvegyület hERG aktivitása nem ismert. Az antidepresszánsok szerkezete igen változatos volt, de volt köztük egy nagyobb, 102 ismert és 91 ismeretlen hERG aktivitású vegyületből álló csoport, melyre külön is elvégeztem az osztályozást, ennek eredménye az 5.2.6. fejezetben olvasható. A számítási folyamat végén a hatféle módszerrel kapott eredményeket együttesen vettem figyelembe, vagyis ún. „konszenzusmodellezéssel" vagy „megegyezési modellezéssel" adtam becslést arra, hogy a tesztvegyületek melyik osztályba tartozhatnak a legnagyobb valószínűséggel.

A változókiválasztás az egyes módszereknél az alábbiak szerint zajlott: A döntési fák esetében kimerítő kereséssel végeztük a felosztást, és a célfüggvény a félreosztályozási hiba volt. Az egyes deszkriptorok „fontosságát" a hERG aktivitás leírásában egy nulla és 100 közötti skála mutatta. A legfontosabb változó 100-as értéket kapott, ilyenből összesen 20 db volt, ezekkel végeztem el a modellépítést. Az itt és az összes többi módszerrel kiválasztott deszkriptorok nevét terjedelmi okok miatt itt nem ismertetem, de a Függelékben megtalálhatók.

A genetikus algoritmust két kiindulási populációval futtattam le, egy populáció 50 egyedből állt. Kiválasztási indikátorként a R^2_{CV}-t használtam. A MobyDigs programmal többszörös lineáris regressziót végeztem minden egyes keletkező deszkriptor-kombinációval és az így keletkező modellek közül azt tekintettem legjobbnak, amelynek a R^2_{CV} értéke a legnagyobb volt. A legjobb modell az eredeti 849 deszkriptorból alig 29-et tartalmazott.

Az FS üzemmódban történő használatakor 3%-os szignifikanciaszintet állítottam be, ekkor az előrefele irányuló változóbevitel 33 deszkriptort talált szignifikánsnak.

A PLS módszer diszkriminancia analízissel kombinált verzióját a következőképpen valósítottam meg: Amikor két csoportra (inaktív és aktív) osztottam fel a vegyületeket, akkor két ún. „dummy" (vagy segéd) függő változót használtam. A függő változók értéke 1, ha a vegyület az adott osztályba tartozik; ugyanennél a vegyületnél a másik függő változó értéke pedig nulla lesz. Ha három csoportra osztottam (inaktív, átmeneti és aktív), akkor három függő változóval dolgoztam. Ilyenkor a legkisebb távolság (maximális valószínűség) módszerét alkalmaztam: A három előrejelzett segédváltozó közül legnagyobb értékűekhez egyest, a többihez nullát rendelve.

A változókiválasztás után következett az osztályozás. Az így kapott végső hat modell osztályozási mátrixait az 5.2.4.1. táblázat mutatja.

Látható, hogy nagyjából ugyanolyan hatékonyságú modelleket kaptam, mint az első számítási folyamat során. Közülük kiemelkedik a GA-LDA és az FS-LDA modell, melyek az összes vegyület 75-80%-át képesek helyesen besorolni, az osztályozás hatékonysága az 1-es és 3-as osztály esetén pedig 80% felett van. A legjobb eredményt az FS-LDA módszerrel sikerült elérni az 1-es osztály esetén – a hERG csatornára nem ható molekulák csaknem 90%-át sikerült helyesen besorolni. A 2-es csoport osztályozása ezekkel a modellekkel sem volt sikeres.

5.2.4.1. táblázat

Helyesen besorolt vegyületek (%)

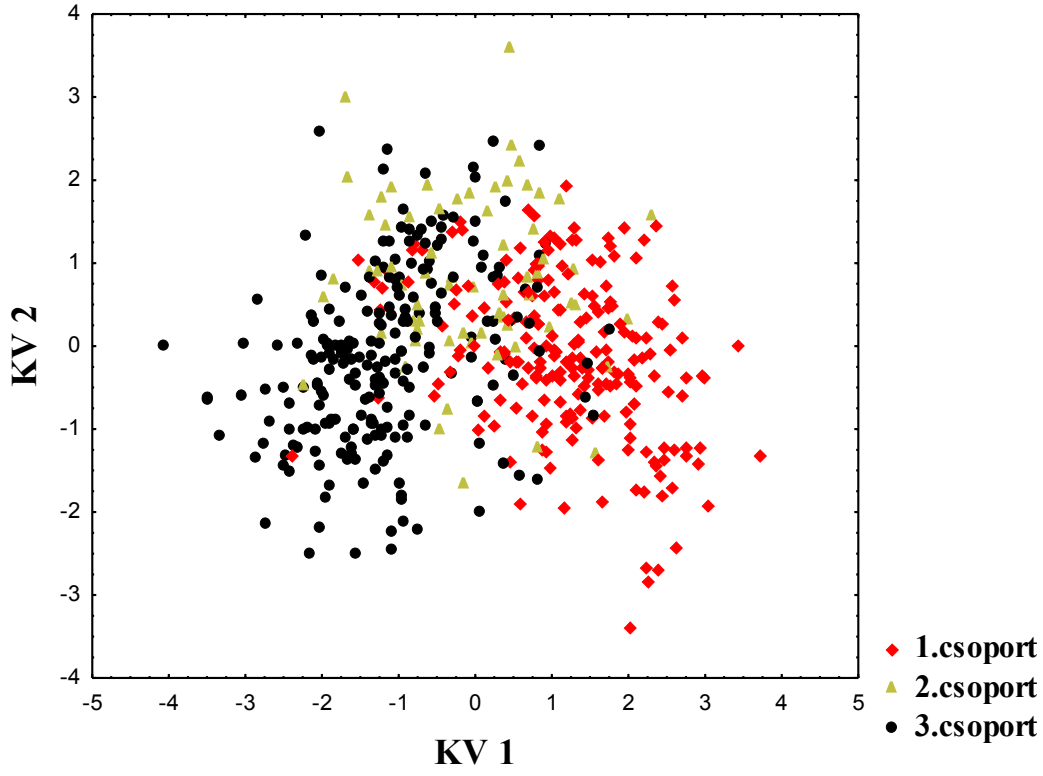
Betanuló készlet (n=502)

Változókiválasztás	***CART***	***GA***	***FS***	***nincs***	***CART***	***FS***
Osztályozás	***LDA***	***LDA***	***LDA***	***PLS DA***	***PLS DA***	***CART***
1.csoport	78,30	87,74	89,62	76,88	76,89	71,70

2. csoport	7,46	17,91	23,88	0,00	0,00	0,00
3. csoport	79,82	82,96	83,86	78,03	75,78	84,75
Összesen	**69,52**	**76,30**	**78,29**	**67,13**	**66,14**	**67,93**

5.2.4.1. ábra

Osztályozás GA-LDA módszerrel



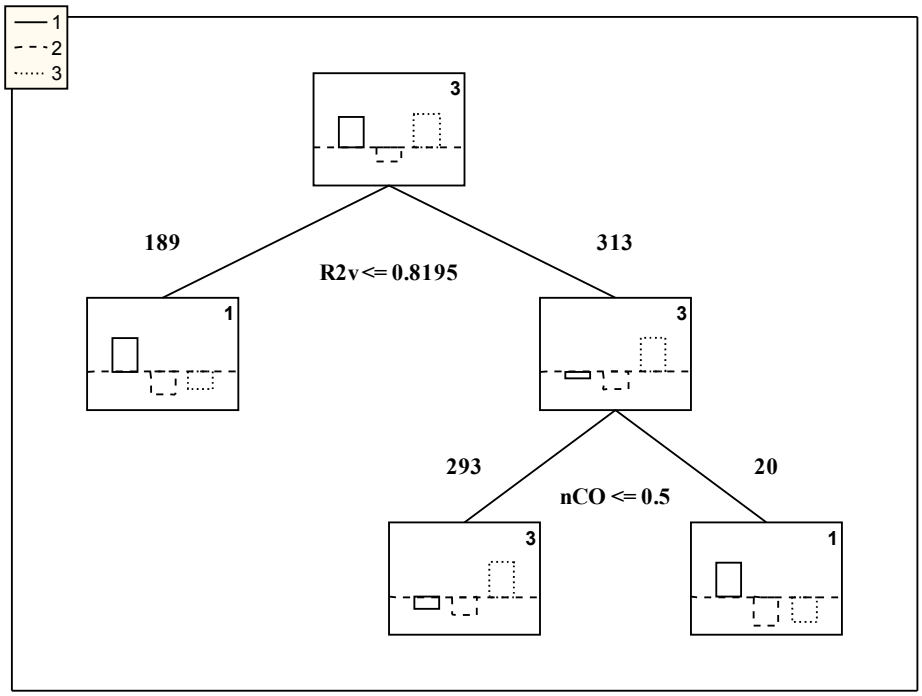
Az 5.2.4.2. ábrán jól követhető, hogy döntési fa segítségével hogyan osztályozhatjuk egyszerűen és érthetően a vegyületeket. Az FS-CART modell mindössze két független változót tartalmaz, egy van der Waals térfogattal súlyozott GETAWAY deszkriptort (R2v) és a ketocsoportok számát (nCO).

Az első felosztáskor az R2v deszkriptor segítségével két csoportra osztom a vegyületeket: 189-et az 1-es csoportba, 313-at pedig a 3-as (illetve vegyes, 2-es és 3-as) csoportba. Valójában nem mind a 189 vegyület tartozik az 1-es csoportba, ezt jelzik a lefelé mutató oszlopok (pl. a CART itt most nem ismertetett eredménytáblájából az is kiderül, hogy 135 vegyület tartozik valóban az 1-es csoportba, 22-a 2-esbe és 32 a 3-as csoportba). A 313 db

vegyes csoportba tartozó vegyület közül a következő osztáskor az nCO deszkriptor felhasználásával 293-at teszek a 3-as csoportba, 20-at azonban visszateszek az 1-es csoportba. A végső felosztás ábráin (legalsó sor) is látszik, hogy nem sikerült az összes vegyületet jól besorolni.

5.2.4.2. ábra

Osztályozás FS-CART módszerrel



5.2.5. Végső modellek két vegyületosztályra történő felosztásnál

A két csoporttal kapott végső (hat) modell osztályozási mátrixait az 5.2.5.1. táblázat tartalmazza. A modellek ugyanazokat a független változókat tartalmazzák, mint a három csoportos felosztásnál számított modellek.

5.2.5.1. táblázat

Helyesen besorolt vegyületek (%)

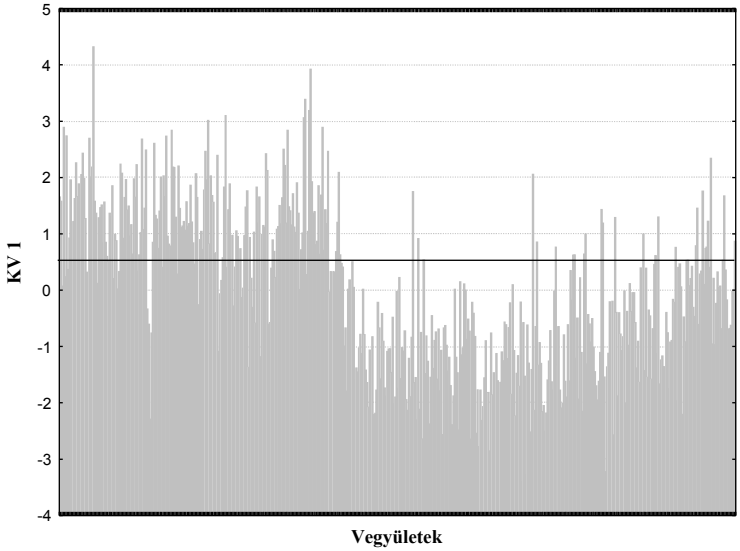
Betanuló készlet (n=502)

Változókiválasztás	***CART***	***GA***	***FS***	***nincs***	***CART***	***FS***
Osztályozás	***LDA***	***LDA***	***LDA***	***PLS DA***	***PLS DA***	***CART***
1. csoport	72,51	85,30	83,89	69,34	65,10	72,04
2. csoport	82,82	86,60	85,91	82,07	80,34	80,41
Összesen	**78,49**	**86,06**	**85,06**	**76,69**	**73,90**	**76,89**

A CART-LDA, GA-LDA és FS-LDA modellek esetén hasonlóan jól sikerült elkülöníteni a hERG csatornához nem kötődő vegyületeket, mint a 3 osztályos modellek esetén. Az összevont csoport elemeit pedig hasonlóan eredményesen sikerült besorolnom, mint a 3 osztályos modelleknél a blokkoló vegyületeket (eredetileg is a 3. osztály tagjai). A legjobb, mindkét csoportra 80% feletti besorolást adó modellek ismét a GA-LDA és az FS-LDA voltak. Az 5.2.5.1. ábra szemlélteti a két vegyületcsoport elkülönülését. A diagram bal oldalán találhatók az 1-es csoport tagjai (ekkor a kanonikus változó értéke 0,5-nél nagyobb a helyesen besorolt molekulák esetén), a jobb oldalán pedig a 2-es csoportba tartozó vegyületek láthatók.

5.2.5.1. ábra

Osztályozás FS-LDA módszerrel



A PLS DA modellek a blokkolók kiválasztásában hatékonyak voltak, de az inaktív vegyületeknek „csak” közel 70%-át sikerült besorolniuk.

5.2.6 A legnagyobb antidepresszáns alcsoport osztályozása

Az alcsoportban ismert maradék hERG aktivitású vegyületek, azaz a betanuló készlet tagjainak száma 102 volt, a tesztvegyületeké 91. Ugyanazokat a változókiválasztási és osztályozási módszereket alkalmaztam, mint a teljes adatkészlet esetén, de a túlillesztés elkerülése érdekében az LDA modelleknél független változók számában további csökkentést hajtottam végre. Előreirányuló változóbevonást alkalmaztam, a szignifikanciaszintet 5%-ra állítottam be.

A modellekbe bekerült deszkriptorok kiválóan kódolják az 1-es és 3-as csoportba tartozó vegyületek hERG aktivitásért felelős jellemzőit. A CART-LDA, a GA-LDA, az FS - LDA és a PLS DA modellek esetében a blokkolók és az inaktív vegyületek több mint 94%-át sikerült helyesen osztályozni. Az átmeneti csoportot viszont egyáltalán nem tudták osztályozni a modellek. A PLS DA szintén jó eredményt adott az 1-es és a 3-as csoportra, de a CART-tal tovább szűkített PLS DA model az inaktív vegyületeket már nem sorolta be jó hatékonysággal.

5.2.4.1. táblázat

Helyesen besorolt vegyületek (%)

Betanuló készlet (n=102)

Változókiválasztás	***CART***	***GA***	***FS***	***nincs***	***CART***	***FS***
Osztályozás	***LDA***	***LDA***	***LDA***	***PLS DA***	***PLS DA***	***CART***
1.csoport	94,12	94,11	94,12	94,12	58,82	71,70
2. csoport	0,00	0,00	0,00	0,00	0,00	0,00
3. csoport	94,94	98,73	96,20	94,94	94,94	84,75
Összesen	**89,22**	**92,16**	**90,20**	**89,22**	**83,33**	**67,93**

5.2.6.2. táblázat

Helyesen besorolt vegyületek (%)

Betanuló készlet (n=91)

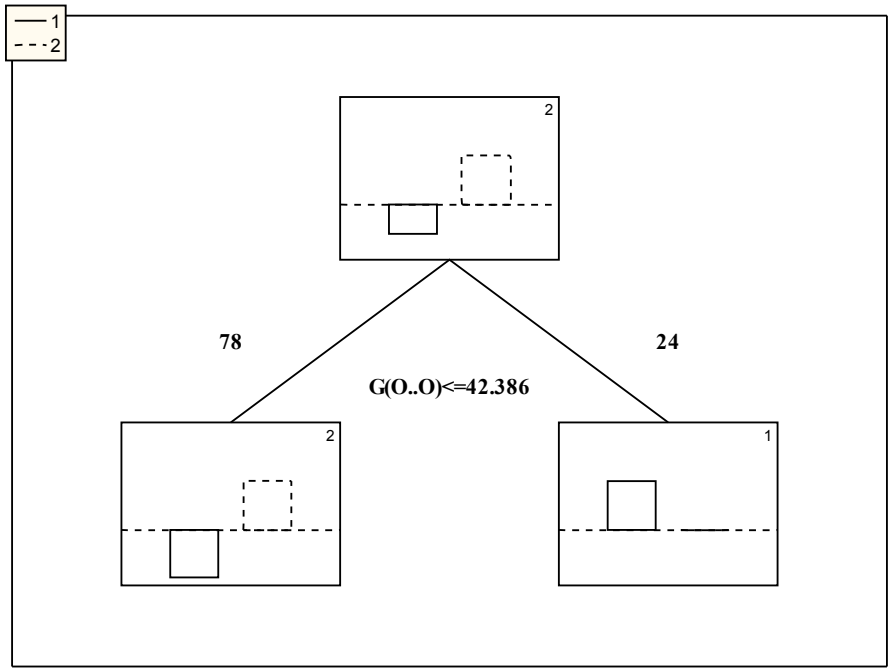
Változókiválasztás	***CART***	***GA***	***FS***	***nincs***	***CART***	***FS***
Osztályozás	***LDA***	***LDA***	***LDA***	***PLS DA***	***PLS DA***	***CART***
1. csoport	94,12	94,12	94,12	94,12	58,82	94,12
2. csoport	94,12	96,47	96,47	94,12	95,29	90,59
Összesen	**94,12**	**96,08**	**96,08**	**94,12**	**89,21**	**91,18**

Mivel ezek a modellek jóval kevesebb független változót tartalmaztak, mint a teljes vegyületkészlet modelljei, több lehetőség van a független változók bemutatására is, természetesen a látens változókat tartalmazó PLS modelleket leszámítva. A CART-LDA modell két deszkriptort tartalmazott. Ezek közül az egyik, a G(O..O) az oxigén-oxigén kötésekben az atomok közti geometriai távolságok összege, mind a négy modellben megtalálható. Sőt, az FS-CART modell mindössze ezt az egy deszkriptort tartalmazza és csak ennek a felhasználásával az 1-es csoport 71,7%-át és a 3-as csoport 84,7%-át sikerült helyesen osztályozni. Ezt az egyszerű modellt az 5.2.6.1. ábra mutatja be. A GA-LDA modellben a közös deszkriptoron kívül szerepel még az aromás kötések száma, egy Wiener típusú index, egy MoRSE deszkriptor és a topológiai távolságok összege az oxigén-oxigénkötésekben. Az FS-LDA modellben az elágazási és egy GETAWAY deszkriptor található, valamint a tioketon csoportok száma.

A két csoportot tartalmazó modellek esetén hasonlóan jó eredményeket kaptam. A CART-PLS DA modell 89%-ot, a többi modell az antidepresszánsoknak több mint 90%-át tudta helyesen osztályozni.

5.2.6.1. ábra

Osztályozás FS CART módszerrel a legnagyobb alcsoport esetében

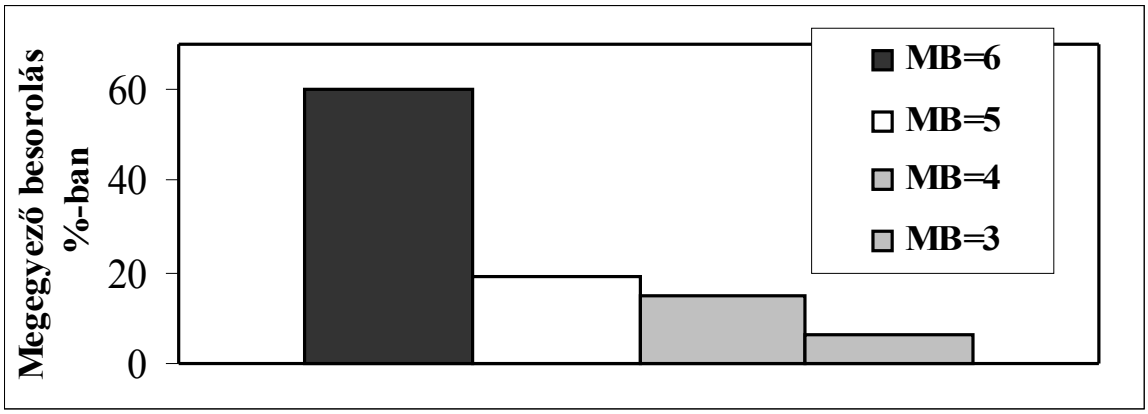


5.2.7. Megegyezési modell felállítása

A hatféle modell felállítása után végeztem egy összehasonlítást, hogy az egyes vegyületek csoportba való besorolása mennyire egyezett meg a különböző módszerek használata esetén. Az 502 betanuló vegyület esetén az ugyanolyan osztályozást eredményező találatok számát az 5.2.7.1. ábra mutatja. Látható, hogy a hat modell a vegyületek 60,4%-át ugyanúgy sorolta be, további 18,9%-ot pedig hatból öt modell osztályozott ugyanúgy. Úgy vélem, hatból öt megegyező besorolás esetén az antidepresszánsokról megbízhatóan előrejelezhető, hogy melyik osztályba tartoznak. Ha a hat és öt megegyező besorolást kapott molekulák számát összegezzük, látszik, hogy az antidepresszánsok több mint 80%-ára adódott a megegyezési modell szerint megbízható előrejelzés. Az esetek 14,5%-ában hatból négy modell ugyanúgy csoportosította a vegyületeket, és a molekulák maradék 6,2%-nál pedig csak három modell adott ugyanolyan eredményt. Ez utóbbi esetben az antidepresszánsok besorolása nem tekinthető megbízhatónak.

5.2.7.1. ábra

A megegyező besorolások száma az antidepresszánsok csoportba sorolásakor



MB : megegyező besorolás

5.3. Összefoglalás

- Sikerült hERG gátló és nem gátló antidepresszánsok osztályozására alkalmas modelleket létrehozni. A modellek szerkezetileg igen változatos vegyületek esetében is jó (80% feletti), szerkezetileg egy családba tartozó vegyületek között pedig kiváló (90% feletti) elkülönítést tettek lehetővé.
- A legjobb modellek, amelyekkel az antidepresszánsokat hERG aktivitásuk alapján csoportosítani lehetett, az előreirányuló változóbevonás illetve a genetikus algoritmus lineáris diszkriminancia elemzéssel összekapcsolt változatai voltak.
- Megmutattam, hogy CART módszer használatával egyszerű és könnyen értelmezhető modell hozható létre az antidepresszánsok osztályozására. Az összes vegyület csoportosításához két független változó szükséges, a karbonilcsoportok száma a molekulában és egy GETAWAY deszkriptor. Ezzel a GETAWAY deszkriptorok használatának jogosultságát is megmutattam a fenti osztályozás

témakörében. A legnagyobb alcsoport osztályozásához egy deszkriptor, az oxigén-oxigén kötésekben lévő geometriai távolságok összege, elégséges.

- Megegyezési modellezéssel igazoltam, hogy az általam létrehozott modellekkel megbízhatóan lehet csoportosítani antidepresszánsokat hERG aktivitásuk alapján.

Irodalom az 5. fejezethez

Aronov A.M., Goldman B.B. A model for identifying hERG K^+ channel blockers *Bioorg. Med. Chem. Lett.* **2004**, 12, 2307-2315.

Bains W., Basman A., White C. HERG binding specificity and binding site structure: evidence fro fragment-based evolutionary computing SAR study *Prog. Biophys. Mol. Biol.* **2004**, 86, 205-233.

Barnett, A.A. Safety concerns oveantipsychotic drug, sertindol. *Lancet* **1996**, 348, 256-57.

Buyck, C., Tollenaere, J., Engels, M., De Clerck, F. An in silico model for detecting potential hERG blocking. The 14th European Symposium on QSAR, 8-13 September **2002**, Bournemouth, UK

Cavero, I., Crumb, W. Native and cloned ion channels from human heart: laboratory models for evaluating the cardiac safety of new drugs *Eur. Heart J*. **2001**, 3(K), K53-K63.

Coi A., Massarelli I., Murgia L., Saraceno M., Calderone V., Bianucci A. M. Prediction of hERG potassium channel affinity by the CODESSA approach *Bioorg. Med. Chem.* **2006,** 14 (9), 3153-3159.

De Ponti, F. QT-interval prolongation by non-cardiac drugs: lessons to be learned from recent experience. *Eur. J. Clin. Pharmacol.* **2000**, 56, 1-18.

Du L., Li M., You Q., Xia L. A novel structure-based virtual screening model for the hERG channel blockers *Biochem. Biophys. Res. Commun*. **2007,** 355, 4, 889-894.

Glassman, A.H., Bigger J.T. Jr. Antipsychotic drugs: prolonged QTc interval, torsades de pointes, sudden death *Am. J. Psychiatry* **2001**, 158, 1774-82.

Haddad P.M., Anderson I.M. Antipsychotic-related QTc prolongation,torsade de pointes and sudden death *Drugs* **2002**, 62, 1649-71.

Haverkamp W. The potential for QT prolongation and proarrhythmia by non-antiarrhythmic drugs: clinical and regulatory implications. Report on a policy conference of the European Society of Cardiology. *Eur. Heart J.* **2000**, 21, 1216-1231.

Kelly H.G., Fay J.E., Lavery S.G. Thioridazine hydrochloride (Melleril): its effects on electrocardiogram and a report of two fatalities with electrocardiographic abnormabilities. *Can. Med. Assoc. J.* **1963**, 89, 546-554.

Keserű Gy. M. Prediction of hERG potassium channel affinity by traditional and hologram QSAR techniques *Bioorg. Med. Chem. Lett.* **2003**, 13, 2773-2775.

Leong M.K. A Novel Approach Using Pharmacophore Ensemble/Support Vector Machine (PhE/SVM) for Prediction of hERG Liability *Chem. Res. Toxicol.,* **2007**, 20 (2), 217 -226.

Mitcheson J.S., Chen J., Lin M., Culberson C., Sanguinetti M.C. A structural basis for drug-induced long QT syndrome *Proc. Natl. Acad. Sci. U.S.A.* **2000**, 97, 12329-12333.
Pearlstein R., Vaz R., Kang J., Chen X., Preobrazhenskaya M., Shchekotikhin A., Korolev A., Lysenkova A., Miroshnikova O., Hendrix J., Rampe D. Characterization of hERG potassium channel inhibition using CoMSiA 3D QSAR and homology modeling approaches *Bioorg. Med. Chem. Lett.* **2003**, 13, 1829-1835.

Roche O., Trube G., Zuegge J., Pflimlin P., Alanine A., Schneider G. A virtual screening method for prediction of the hERG potassium channel liability of compound libraries *Chem. Biol. Chem.* **2002**, 3, 455-459.

Schneider G., Neidhart W., Giller T., Schmid G., „Scaffold-hopping" by topological pharmacophore search: a contribution to virtual screening. *Angew. Chem., Int. Ed.* **1999**, 38, 2894-2896.

Seierstad M., Agrafiotis D. K. A QSAR model of hERG binding using a large, diverse, and internally consistent training set *Chem. Biol. Drug Des.* **2006**, 67 (4), 284-296.

Song M.H., Clark M
Development and evaluation of an in silico model for hERG binding
J. Chem. Inf. Model. **2006**, 46 (1), 392-400.

Tamargo J. Drug-induced torsade de pointes: from molecular biology to bedside. *Jpn. J. Pharmacol.* **2000**, 83, 1-19.

Vandenberg J.I., Walker B.D., Campbell T.J. hERG K+ channels: friend and foe. *Trends Pharmacol. Sci.* **2001**, 22, 240-246.

Yap Y.G., Camm J. Risk of torsades de pointes with non-cardiac drugs. Doctors need to be aware that many frugs can cause qt prolongation. *Brit. Med. J.* **2000**, 320, 1158-1159.

Yoshida K., Niwa T. Quantitative structure-activity relationship studies on inhibition of hERG potassium channels *J. Chem. Inf. Model*. **2006**, 46 (3), 1371-1378.

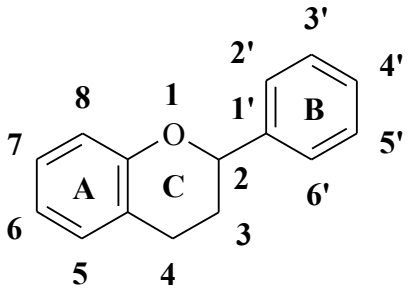
6. FLAVONOIDOK ANTIOXIDÁNS AKTIVITÁSÁNAK ELŐREJELZÉSE

6.1 Flavonoidok jelentősége

A polifenolok csoportjába tartozó flavonoidok (ld. 6.1. ábra) számos zöldségben és gyümölcsben megtalálhatók és egészségre gyakorolt kedvező hatásuk régóta köztudott. Ismert, hogy a flavonoidok rákellenes [*Block, 1992; Elangovan és mtsai, 1994; Middleton és mtsai, 2000*], vírusellenes [*Selway, 1986*] és gyulladáscsökkentő hatással [*Gabor, 1986; Middleton, 1998*] rendelkeznek, és leírták azt is, hogy csökkentik a szív-és érrendszeri betegségek kockázatát [*Facino és mtsai, 1999; Herteg és mtsai, 1993; Mazur és mtsai, 1999*].

6.1. ábra

Flavonoidok általános szerkezete



Ezek a jótékony hatások, aktivitások általában a flavonoidok antioxidáns vagy szabadgyökfogó a kapacitásával hozhatók kapcsolatba. A természetben előforduló flavanoidszármazékok száma meghaladja a 4000-et, antioxidáns hatásuk mértéke azonban nagyban különbözik egymástól. A QSAR modellezés alkalmas arra, hogy segítségével a leghatékonyabb antioxidánsokat kiválogassuk a nagymennyiségű flavonoid közül. Az összefüggés fordítva is igaz; nagy számuk és számos pozitív biológiai hatásuk miatt a flavonoidok kedvelt tárgyai a QSAR modellezésnek. Az irodalomban számos példa található

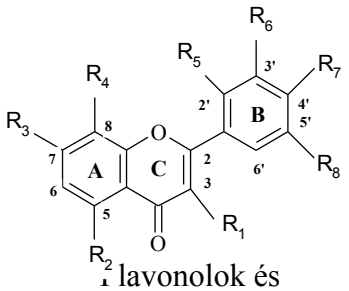
flavonoidok különféle biológiai aktivitásának sikeres előrejelzésére, pl. HIV-ellenes [*Alves és mtsai, 1999 és 2001*], tumorellenes [*Sergediene és mtsai, 1999; Moriani és mtsai, 2001*] aktivitás számítására. Különféle enzimekre gyakorolt hatásuk előrejelzése pl. citokróm P-450 [*Moon és mtsai, 2000*], p5(lck) protein tirozin kináz [*Nikolowska és mtsai, 2000; Oblak és mtsai, 2000; Stefanic-Petek és mtsai, 2000*] is megvalósult. A flavonok antioxidáns hatásának leírásáról számos szerkezet-hatás összefüggés született, de ezek közül csak kevés, ami valóban mennyiségi. Ezek szerint a flavonoidok antioxidáns aktivitása nagyban függ a molekulában található hidroxilcsoportok számától és elhelyezkedésétől. Nagyban növeli az antioxidáns kapacitást, ha a B-gyűrűn 3'és 4' pozícióban két hidroxilcsoport helyezkedik el (katechol szerkezet), ha telítetlenség van a C-gyűrűben, és oxocsoport található a 4-es helyzetben [*Rice-Evans és mtsai, 1996; Harborne és mtsai, 2000; Heim és mtsai, 2002*]. Egy tanulmány szerzői számított paramétereket, köztük képződéshőt, HOMO és LUMO energiákat és a hidroxilcsoportok számát használták gyökfogó kapacitás előrejelzésére [*Lien és mtsai, 1999*]. Egy másik kutatócsoportcsoport a leíráshoz különféle indikátorváltozókat definiált, amiket a hidroxilcsoportok száma és helyzete alapján számítottak [*Amić és mtsai, 2003*]. Megemlítendő, hogy ezek a munkák csak leírást adtak az antioxidáns aktivitásra, de nem jelezték előre azt.

Munkám során célom az volt, hogy mennyiségi meghatározásra és előrejelzésre alkalmas modellt építsek flavonoidok antioxidáns aktivitásának meghatározására, valamint hogy biológiai aktivitásuk és szerkezeti sajátságaik alapján csoportosítsam a flavonoid-származékokat. A modellépítéskor szempont volt, hogy egy gyorsan számítható, egyszerűen kezelhető modellt hozzak létre.

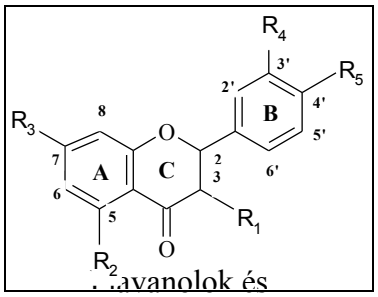
6.2. Vizsgált vegyületek és számítások

Adatkészletem 36 flavonoidot foglalt magába, antioxidáns aktivitás adataikat az irodalomból vettem [*Burda és Oleszek, 2001*]. A vegyületeim többfajta flavonoidcsaládba tartoztak, voltak köztük flavonolok flavonok, flavanonok, izoflavonok, dihidroflavonolok és szénhidrát származékaik, biflavanonok és izoflavonok (ld. 6.2. táblázat).

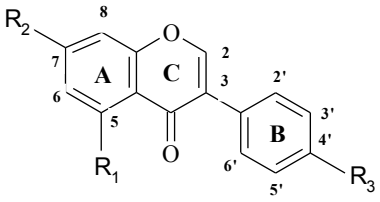
6.2. A vizsgált flavonoidcsoportok általános szerkezete



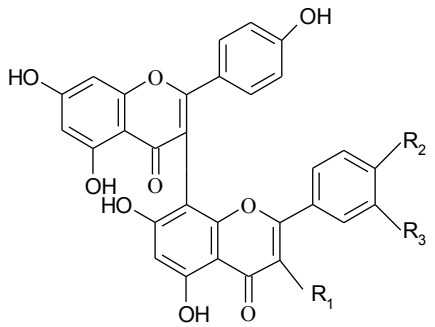
Flavonolok és flavonok



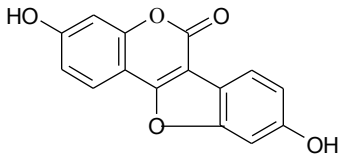
Flavanolok és dihidroflavonolok



Izoflavonok



Biflavanolok



Kumesztrol

A vizsgált flavonoidok neve és szerkezete megtalálható a Függelékben. Az antioxidáns aktivitásukat a fenti forrás szerzői az alapján jellemezték, hogy mennyire tudják meggátolni a termikus oxidációt β-karotin - linolsav modellrendszerben [23].

Független változóként konstitutív deszkriptorokat, topológiai és elágazási indexeket használtam. A flavonoidok nagy mérete és bonyolult szerkezete miatt a számítási időt jelentősen lerövidíti a kétdimenziós deszkriptorok használata. Összesen 147 deszkriptort

számítottam, majd a PLS módszert használtam modellépítésre. A PLS komponensek optimális számát keresztellenőrzéssel állapítottam meg. A flavonoidok 70%-a került a betanuló (és egyben kalibrációs) készletbe, a teszt készletbe pedig a maradék 30%-uk. A készletbe osztás a viszonylag sokfajta szerkezet és ahhoz képest viszonylag kisszámú rendelkezésre álló vegyület miatt nem random módon történt, hanem figyelve arra, hogy minden készletbe kerüljön minden flavonoid alcsoportból és szénhidrátszármazékból. A modell prediktív erejének jellemzésére R^2_{CV} és SD_{CV} értéket számítottam. A modell további elemzésére adott a PLS komponensek ábrája is lehetőséget adott.

6.3. Eredmények és értékelésük

A 6.3.1. ábrán látható a keresztellenőrzés eredménye, amely szerint 13 PLS használata esetén a legkisebb az SD_{CV} illetve a PRESS értéke. Az ábráról azonban az is látszik, hogy a PRESS értékek görbéjén több lokális minimum található; az egyik 2 PLS komponensnél, a másik pedig 8-nál. A 8 PLS komponens alkalmazásakor adódó szórásérték nem nagyobb jelentősen a 13 PLS komponensnél adódónál, de a szabadsági fok az előbbi esetben valószínűleg jelentősen kisebb. Ezért célszerűbbnek tartottam a modellben csak 8 PLS komponenst használni.

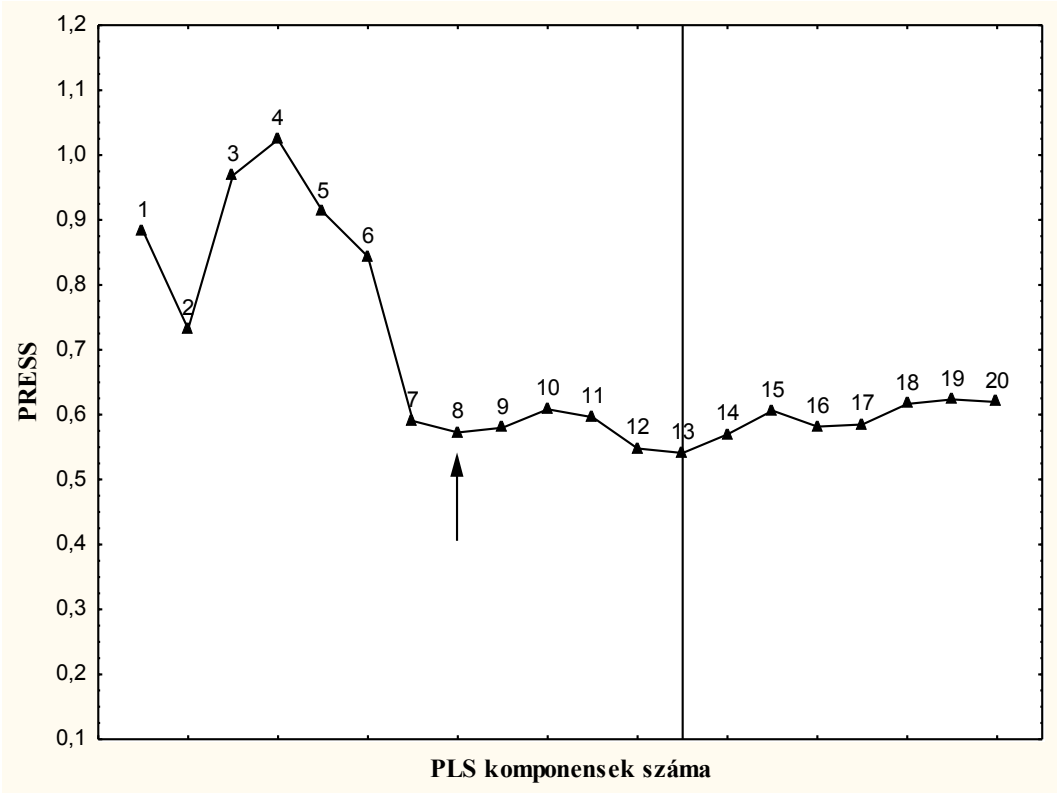
A statisztikai paraméterek értéke azt mutatja, hogy a modell alkalmas a flavonoidok antioxidáns aktivitásának előrejelzésére. Az R^2_{CV} értéke 0,9205, a PRESS értéke pedig 0,5531 volt.

Az egyes változók regressziós koefficienseinek értéke szerint az alábbi független változóknak van a legnagyobb szerepe az antioxidáns aktivitás leírásában: négy darab átlagos elágazási indexnek, valamint két közepes topológiai töltés indexnek. Érdekes, hogy a

konstitutív deszkriptoroknak (pl. az OH csoportok számának) nincs szignifikáns szerepe a modellben.

6.3.1. ábra

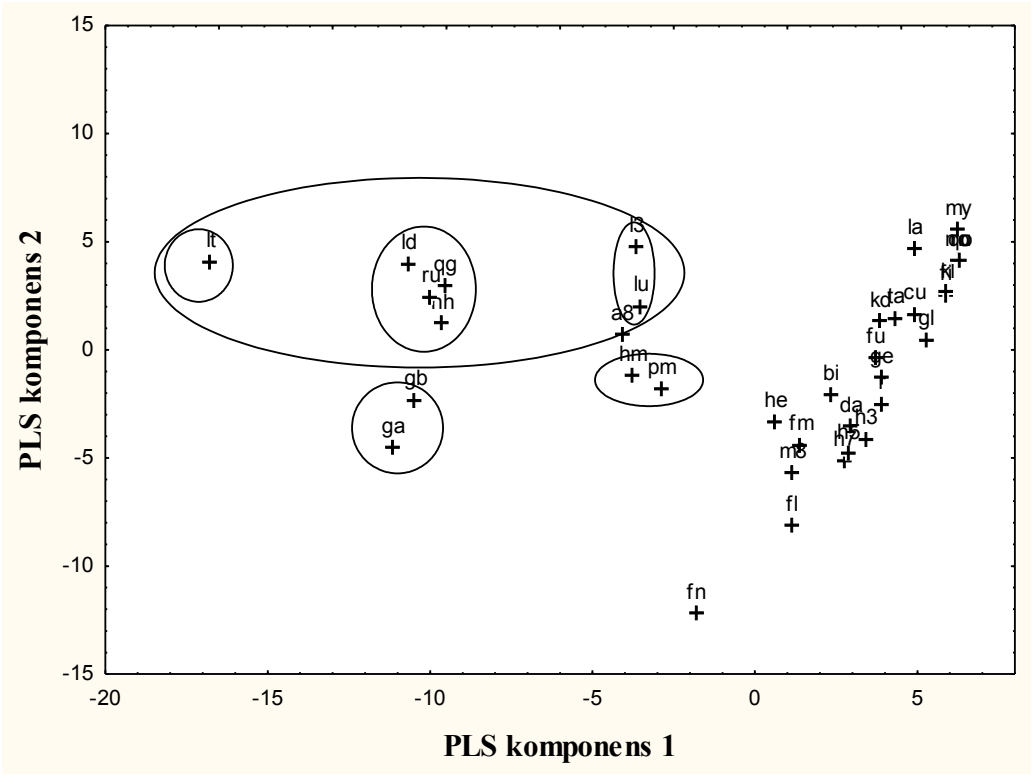
A reziduális szórás nagysága a PLS komponensek függvényében



A 6.3.2. ábrán a PLS komponensek ábrája látható, melyen megfigyelhető az egyes flavonoidcsoportok elkülönülése. A szénhidrátszármazékok (lt, ld, ru, qg, nh, l3, lu, a8 rövidítéssel ellátva, ld. Függelék) jól elkülönülnek a nem szénhidrátszármazékoktól. További különálló csoportot alkotnak a biflavonoidok (gb, ga) és a sok metoxicsoportot tartalmazó flavonoidok (hm, pm).

6.3.2. ábra

PLS komponensek ábrája – flavonoidok csoportosítása



4. Összefoglalás

- Sikerült modellt építeni igen különböző szerkezetű flavonoidok antioxidáns aktivitásának előrejelzésére. A PLS modell egyszerűen, változókiválasztás nélkül létrehozható, és az előrejelzésen kívül alkalmas lehet különféle flavonoidcsoportok osztályozására is.

Irodalmak a 6. fejezethez

Alves C.N., Pinheiro J.C., Camargo A.J., Ferreira M.M.C., Romero R.A.F., da Silva A.B.F. A multiple linear regression and partial least squares study of flavonoid compounds with anti-HIV activity. *J. Mol. Struct.* **2001**, 541, 81-88.

Alves C.N., Pinheiro J.C., Camargo A.J., de Souza A J., Carvalho R B., da Silva A.B.F. A multiple linear regression and partial least squares study of flavonoid compounds with anti-HIV activity. *J. Mol. Struct*. **1999**, 491, 123-131.

Amić D., Davidovic-Amić D., Bešlo D., Trinajstić N. Structure-radical scavenging activity relationships of flavonoids. *Croat. Chem. Act.* **2003**, 76(1), 55-61.

Block G. A. Role for antioxidants in reducing cancer risk. *Nutr. Rev.* **1992**, 50, 207-213.
Burda S., Oleszek W. Antioxidant and antiradical activities of flavonoids. *J. Agric. Food Chem*. **2001**, 49, 2774-2779.

Elangovan V., Sekar N., Govindasamy S. Chemoprotective potential of dietary bioflavonoids against 20-methylcholanthrene-induced tumorgenesis. *Cancer Lett.* **1994**, 87, 278-284.

Facino R.M., Carini M., Aldini G., Berti F., Rossoni G., Bombardelli E., Morazzoni P. Diet enriched with procyanidins enhances antioxidant activity and reduces myocardial post-ischemic damage in rats. *Life Sci*. **1999**, 64, 943-949.

Gabor M. Anti-inflammatory and anti-allergic properties of flavonoids. *Prog. Clin. Biol. Res.* **1986**, 213, 471.

Harborne J.B., Williams C.A. Advances in flavonoid research since 1992. *Phytochemistry* **2000**, 55, 481-504.

Heim K.E., Tagliaferro A.R., Bobilya D.J. Flavonoid antioxidants: chemistry, metabolism and sructure-activity relationships. *J. Nutr. Biochem.* **2002**, 13, 572-584.

Herteg M.G.L., Feskens E.J.M., Hollman P.C.H., Katan M.B., Kromhout D. Dietary flavonoids and risk of coronary heart disease. *Lancet* **1993**, 342, 1007-1011.

Lien E.J., Ren S.J., Bui H.Y.H., Wang R.B. Quantitative structure-activity relationship analysis of phenolic antioxidants. *Free Radical Biol. Med.* **1999**, 26(3-4), 285-297.

Mazur A., Bayle D., Lab C., Rock E., Rayssiguier Y. Inhibitory effect of procyanidin-rich extracts on LDL-oxidation in vitro. *Atherosclerosis* **1999**, 149, 421-422.

Middleton Jr. E. Effect of plant flavonoids on immune and inflammatory cell function. *Adv. Exp. Med. Biol.* **1998**, 439, 175.

Middleton E.J., Kandaswami C., Theoharides T.C. The effects of plant flavonoids on mammalian cells: implications for inflammation, heart disease, and cancer. *Pharmacol. Rev.* **2000**, 52, 673-651.

Moriani M.Z., Galati G., O'Brien P.J. Comparative quantitative structure toxicitz relationships for flavonoids evaluated in isolated rat hepatocytes and HeLa tumor cells. *Chem.-Biol. Interact*. **2002**, 139, 251-264.

Moon T., Chi M.H., Kim D.H., Yoon C.N., Choi Y.K. Quantitative structure-activity relationships (QSAR) study of flavonoid derivatives for inhibition of cytochrome P450 1A2. *Quant. Struct. Act. Relat.* **2000**, 19, 257-263.
Nikolovska-Coleska Z., Suturkova L., Dorevski K., Krbavcic A., Solmajer T. Quantitative structure-activity relationship of flavonoid inhibitors of p56(lck) protein tyrosine kinase: A classical/quantum chemical approach. *Quant. Struct. Act. Relat.* **1998**, 17(1), 7-13.

Oblak M., Randic M., Solmajer T .Quantitative structure-activity relationship of flavonoid analogues. 3. Inhibition of p56(lck) protein tyrosine kinase. *J. Chem. Inf. Comput. Sci.* **2000**, 40(4), 994-1001.

Rice-Evans C.A., Miller N.J., Paganga G. Structure-antioxidant activity relationships of flavonoids and phenolic acids. *Free Radical Biol. Med*. **1996**, 20, 933-956.

Selway J.W. Antiviral activity of flavones and flavans. *Prog. Clin. Biol. Res*. **1986,** 213, 521.

Sergediene E., Jönsson K., Szymusiak H., Tyrakowska B., Rietjens I.M.C.M., Cenas N. Prooxidant toxicity of polyphenolic antioxidants to HL-60 cells: description of quantitative structure-activity relationships. *FEBS Lett.* **1999**, 462, 392-396.

FÜGGELÉK

Adatok elrendezése, adatmátrix:

Az adatokat sorokba és oszlopokba szokás rendezni, a kapott táblázatot pedig mátrixnak tekinteni. Ezt a rendezett táblázatot adatmátrixnak nevezik.

$$\mathbf{X_{NP}} = \begin{bmatrix} x_{11} & \dots & x_{1p} & \dots & x_{1P} \\ \dots & & & & \\ x_{n1} & \dots & x_{np} & \dots & x_{nP} \\ \dots & & & & \\ x_{N1} & \dots & x_{Np} & \dots & x_{NP} \end{bmatrix}$$

ahol n = 1…N; p = 1…P

Megegyezés szerint az adatmátrix egy sora egy objektum (dolgozatomban különféle vegyületek) tulajdonságait tartalmazza, a mátrix egy oszlopa pedig egy-egy tulajdonság (dolgozatomban a deszkriptorok) értékeiből áll. A sorokat szokás objektumvektoroknak, az oszlopokat pedig tulajdonságvektoroknak nevezni.

Kovarianciamátrix:

A tulajdonságokat külömböző mértékegységben mérjük, számértékük nagyságrendekkel különbözhetnek. Az eltérő lépték időnként számítástechnikai vagy valószínűségszámítási okokból hátrányos lehet, ezért léptékváltásra lehet szükség. A centrálás a léptékváltás egy módja, ami konstans eltolást jelent az eredeti skálán úgy, hogy a tulajdonságvektorok elemeinek számtani közepe nulla legyen. A kovarianciamátrixot a centrált adatokból számítják:

$$\mathbf{C} = \frac{1}{N - 1} X_c^T X_c$$

ahol X_c a centrált adatokat tartalmazó adatmátrix.

Kalapmátrix:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T)^{-1}\mathbf{X}^T$$

F 1 táblázat

Alkoholok RI előrejelzésénél használt deszkriptorok

Rövidítés	Deszkriptor*
MW	molekulatömeg
AMW	átlagos molekulatömeg
Ms	elektrotopológiai index
Vm	atomtömegekkel súlyozott totál méret index
P1U	az első főkomponens irányába mutató mérettel összefüggő WHIM index, súlyozatlan
P2u	a második főkomponens irányába mutató, alakkal összefüggő WHIM index, súlyozatlan
P2m	a második főkomponens irányába mutató, alakkal összefüggő WHIM index , atomtömeggel súlyozva
G1m	az első főkomponens irányába mutató, szimmetriával összefüggő WHIM index, atomtömeggel súlyozva
L1u	az első főkomponens irányába mutató alakkal összefüggő WHIM index, súlyozatlan
L1e	az első főkomponens irányába mutató mérettel összefüggő WHIM index, elektronegativitással súlyozva
Vu	V totál méret index, súlyozatlan
As	A total méret index, elektrotopológiai sajátsággal súlyozva

*A deszkriptorok részletesebb ismertetése megtalálható: Todeschini R. Consonni, V. In *Handbook of Molecular Descriptors* Wiley – VCH: Weinheim, 2000.

F 2 táblázat

Zsírsav metil-észterek RI előrejelzésénél használt deszkriptorok

Rövidítés	Deszkriptor neve
ATS1v	A topológiai szerkezet 1-es távolságú Broto-Moreau-féle autokorrelációs indexe, van der Waals térfogattal súlyozva
ATS3m	A topológiai szerkezet 3-as távolságú Broto-Moreau-féle autokorrelációs indexe, atomtömeggel súlyozva
BEHe5	A Burden-féle kapcsolódási mátrix 5.legnagyobb sajátértéke, elektronegativitással súlyozva
BEHm1	A Burden-féle kapcsolódási mátrix 3. legnagyobb sajátértéke, atomtömeggel súlyozva
BEHm8	A Burden-féle kapcsolódási mátrix 8. legnagyobb sajátértéke, atomtömeggel súlyozva
BELm1	A Burden-féle kapcsolódási mátrix legkisebb sajátértéke, atomtömeggel súlyozva
BELm5	A Burden-féle kapcsolódási mátrix 5. legkisebb sajátértéke, atomtömeggel súlyozva
BELp6	A Burden-féle kapcsolódási mátrix 5. legkisebb sajátértéke, atomi polarizálhatósággal súlyozva
CIC1	Kiegészítő információtartalom index (elsőrendű szomszédsági szimmetria)
Eig1Z	Z-súlyozott távolság mátrix legnagyobb sajátértéke
GATS3v	3-as távolságú Geary-féle autokorreláció, van der Waals térfogattal súlyozva
GATS6v	6-os távolságú Geary-féle autokorreláció, van der Waals térfogattal súlyozva
GGI1	Topológiai töltés index, elsőrendű
HNar	Narumi topológiai index
IVDM	A hidrogén hiányos gráf szigma elektronszámának közepes információtartalma

IDDE	A (gráf) távolság megosztásekvivalens osztályokkal definiált közepes információtartalma
JGI2	Másodrendű közepes topológiai töltés index
Lop	Lopping-index
MPC10	Az (önmagába vissza nem térő) tizedrendű molekuláris utak száma
MW	Molekulatömeg
MWC06	Hatodrendű molekuláris utak száma
nDB	Kettős kötések száma
PCD	Többszörös (önmagába vissza nem térő) molekuláris utak különbsége
PHI	Kier flexibilitási index
piPC03	Az (önmagába vissza nem térő)harmadrendű többszörös molekuláris utak száma
piPC04	Az (önmagába vissza nem térő) negyedrendű többszörös molekuláris utak száma
piPC07	Az (önmagába vissza nem térő) hetedrendű többszörös molekuláris utak száma
piPC10	Az (önmagába vissza nem térő) tizedrendű többszörös molekuláris utak száma
QIndex	Négyzetes Zágráb-index
RBF	Forgatható kötések részaránya
RBN	Forgatható kötések száma
SCBO	Hagyományos kötésrend összege (H-hiányos gráfban)
TIC0	Teljes információtartalom index
TIC3	Harmadrendű szomszédos szimmetriát figyelembevevő teljes információtartalom index
TIC4	Negyedrendű szomszédos szimmetriát figyelembevevő teljes információtartalom index
TIC5	Ötödrendű szomszédos szimmetriát figyelembevevő teljes információtartalom index
VEA1	A kapcsolódásokat reprezentáló szomszédsági mátrixból származtatott sajátvektor koefficienseinek összege
VEA2	A kapcsolódásokat reprezentáló szomszédsági mátrixból származtatott átlagos sajátvektor koefficienseinek összege

Whete	Az elektronegativitással súlyozott távolságmátrix Wiener-típusú indexe
Xt	A teljes szerkezet kapcsolódási indexe

A Lasso opt modellben felhasznált deszkriptorok:

pilD, Eig1, RBN, MW, SCBO, nCq, PCD, PHI, Whete, piPC04, MPC010, piPC03, MWC06, BAC, piPC07, piPC10, CIC1, Lop, TIC0, BEHm8, BELm5, HNar, Me, TIC

F 3 táblázat

Metil-észterek Kováts-indexének előrejelzése

RI_{becs}: RI_{PCMopt}, RI_{FS5} és RI_{PLSopt} átlaga

Az észter savjának alkillánca	RI_{PCMopt}	RI_{FS5}	RI_{PLSopt}	RI_{becs}
4-metil-pentén	873	901	852	875
8-nonén	1247	1196	1176	1206
4-decén	1306	1287	1304	1299
10-undecén	1453	1394	1397	1415
9-dodecén	1508	1488	1473	1490
4-tetradecén	1702	1684	1706	1697
5-tetradecén	1702	1684	1696	1694
9(E)-tetradecén	1702	1685	1690	1693
7,10,13-hexadekatrién	1873	1858	1860	1864
2-hexadecén	1898	1887	1963	1916
11(E)-hexadecén	1898	1887	1881	1889
2,6,10-trimetil-tridecén	1814	1832	1895	1847
4,8,12,15-oktadekatetrén	2050	2043	2034	2042
9(Z),11(E),13(E),15(Z)-oktradekatetrén	2050	2044	2086	2060
8(E),10(E),12(Z)-oktadekatrién	2065	2054	2102	2074
8(E),10(E),12(E)-oktadekatrién	2065	2054	2102	2074
9 (Z),11 (Z),13 (Z)-oktadekatrién	2065	2055	2108	2076
9 (E),11 (E),13(E)-oktadekatrién	2065	2055	2108	2076
9(Z),11(E),13(E)-oktadekatrién	2065	2055	2108	2076
9(E),12(E),15(E) –oktadekatrién (linolén)	2065	2060	2071	2065
12(E)-oktadecén	2097	2089	2085	2090
9(E)-oktadecén	2097	2087	2090	2091
15(Z)-oktadecén	2097	2094	2080	2090
15(E)-oktadecén	2097	2094	2080	2090
12(Z)-oktadecén	2097	2089	2085	2090
12(E)-oktadecén	2097	2089	2085	2090
7(E)-oktadecén	2097	2087	2086	2090
6(E)-oktadecén	2097	2088	2091	2092
11(E)-oktadecén	2097	2088	2083	2089
10(Z),13(Z)-nonadekadién	2179	2172	2187	2179
10-metiloktadecén	2177	2176	2171	2175
11(E),14(E)-ikozadién	2276	2274	2283	2278
11(E)-ikozén	2292	2290	2279	2287
3,7,11,15-tetrametil-hexadecén	2186	2213	2300	2233
4,8,12,15,19-dokozapentén	2403	2426	2414	2415
10,13-dokozadién	2469	2473	2475	2472
15-dokozén	2491	2494	2480	2488

F 4

Antidepresszánsok osztályozásánál használt deszkriptorok:

Deszkriptorok a CART-LDA modellben:

MW, nDB, SMTI, XMOD, ECC, piPC05, SRW08, BEHv5, DP01, G(O..O), RDF025m, RDF120v, RDF040p, Mor01m, Mor01v, Am, HTv, R1v, C-024, MLOGP

Deszkriptorok GA-LDA modellben:

nDB, nAB, nH, Pol, Whete, Whetp, IVDM, TIC3, IC4, STN, T(O..O), T(O..S), BEHe5, BEHe7, GGI5, SP04, RDF120u, RDF120e, Mor07m, Mor08v, Mor10v, Mor05e, HATSe, nCS, nCO, nN(CO)2, C-009, C-025, N-072

Deszkriptorok az FS-LDA modellben: MW, nR05, Jhetp, X2, X5v, CIC1, STN, D/Dr10, T(F..F), MWC06, GGI3, GATS1v, GATS5p, Qtot, G(O..O), G(O..S), RDF040m, RDF085m, Mor14m, Mor02v, Mor08v, Mor21v, Mor02e, Mor09p, HIC, R2m, R1v, R2v, R8e, nCS, nCO, nN(CO)2, nHDon

Ezek közül a korábban (F1, F2 táblázatban) **még nem ismertetett típusú deszkriptorok**:

nAB: aromás kötések száma

nCS: tioketocsoportok száma

nCO: ketocsoportok száma

nN(CO)2: imidek száma

nHDon: H-donor atomok száma

C-009: R-CH-R fragmentum

C-025: CHRX2 fragmentum

N-072: RCON fragmentum

SMTI: Schultz-féle topológiai index

XMOD: módosított Randic-féle index

SRW8: az önmagába visszatérő nyolcadrendű utak száma

DP01: 1-es számú molekuláris Randic-profil

HTv: H-totál index, van der Waals térfogattal súlyozva

Pol: polaritásszám

Mor01v, Mor02v, Mor08v, Mor10v, Mor21v: 3D-MoRSE deszkriptorok, van der Waals térfogattal súlyozva

Mor01m, Mor07m, Mor14m: 3D-MoRSE deszkriptorok, molekulatömeggel súlyozva

Mor02e, Mor05e: 3D-MoRSE deszkriptorok, elektronegativitással súlyozva

Mor09p: 3D-MoRSE deszkriptor, atomi polarizálhatósággal súlyozva

RDF120u: Radiális eloszlásfüggvény deszkriptor, súlyozatlan

RDF025m, RDF085m: Radiális eloszlásfüggvény deszkriptor, atomtömegekkel súlyozva

RDF120v: Radiális eloszlásfüggvény deszkriptor, van der Waals térfogatokkal súlyozva

RDF040p: Radiális eloszlásfüggvény deszkriptor, polarizálhatósággal súlyozva

G (O..O): Az O-O geometriai kötéstávolságok összege

G(O..S): Az O-S geometriai kötéstávolságok összege

T(O..O): Az O-O topológiai kötéstávolságok összege

T(O..S): Az O-S topológiai kötéstávolságok összege

R1v, R2v: Getaway deszkriptorok, van der Waals térfogattal súlyozva

R2m: Getaway deszkriptorok, molekulatömeggel súlyozva

R8e: Getaway deszkriptor, elektronegativitással súlyozva

F 6

Flavonolok és flavonok

Szám	Név	Röv.	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	Antioxidáns aktivitás (%)
1	kaempferol	kl	OH	OH	OH	H	H	H	OH	H	65.3
2	galangin	gl	OH	OH	OH	H	H	H	H	H	64.9
3	quercetin	qu	OH	OH	OH	H	H	OH	OH	H	63.6
4	robinetin	ro	OH	H	OH	H	H	OH	OH	OH	61.7
5	fisetin	fi	OH	H	OH	H	H	OH	OH	H	61.6
6	kampferid	kd	OH	OH	OH	H	H	H	OMe	H	60.0
7	3-hidroxiflavon	h3	OH	H	H	H	H	H	H	H	59.4
8	laricitrin	la	OH	OH	OH	H	H	OH	OH	OMe	28.5
9	laricitrin 3'-*O*-glikozid	l3	OH	OH	OH	H	H	O-glu*	OH	OMe	26.2
10	miricetin	my	OH	OH	OH	H	H	OH	OH	OH	18.4
11	3,5,7,3',4',5'-hexametoxiflavon	hm	OMe	OMe	OMe	H	H	OMe	OMe	OMe	2.6
12	3,5,7,3',4'-pentametoxiflavon	pm	OMe	OMe	OMe	H	H	O-glu	OMe	OH	1.1
13	laricitrin 3,3'-*O*-diglikozid	ld	O-glu	OH	OH	H	H	OH	OH	OMe	1.1
14	quercetin 3-*O*-glikozid-7-*O*-ramnozid	qg	O-glu	OH	O-rha*	H	H	O-glu	OH	H	-6.2
15	laricirin 3,7,3'-*O*-triglikozid	lt	O-glu	OH	O-glu	H	H	OH	OH	OMe	-6.2
16	rutin	ru	O-rut*	OH	OH	H	H	H	OH	H	-10.2
17	morin	mo	OH	OH	OH	H	OH	H	OH	H	63.5
18	flavon	fl	H	H	H	H	H	H	H	H	-1.5
19	5-hidroxiflavon	h5	H	OH	H	H	H	H	H	H	-4.0
20	7-hidroxiflavon	h7	H	H	OH	H	H	H	H	H	0.0
21	krizin	cr	H	OH	OH	H	H	H	H	H	-20.8
22	8-methoxiflavon	m8	H	H	OMe	H	H	H	H	H	-29.2
23	apigenin 8-*C*-glikozid	a8	H	OH	OH	glu	H	H	OH	H	-29.6
24	luteolin 7-*O*-glikozid	lu	H	OH	O-glu	H	H	OH	OH	H	-25.3

Flavanonok és izoflavonolok

Szám	Név	Röv.	R_1	R_2	R_3	R_4	R_5	Antioxidáns aktivitás
25	flavanon	fn	H	H	H	H	H	-23.0
26	naringin	nh	H	OH	O-neohesp*	H	OH	47.4
27	heszperitin	he	H	OH	OH	OH	OMe	4.7
28	fustin	fu	OH	H	OH	OH	OH	-23.4
29	taxifolin	ta	OH	OH	OH	OH	OH	-16.8

Biflavanonok

Szám	Név	Röv.	R_1	R_2	R_3	Antioxidáns aktivitás
30	GB-1	gb	OH	OH	H	-30.1
31	GB-1a	ga	H	OH	H	-16.9

Izoflavonok

Szám	Név	Röv.	R_1	R_2	R_3	Antioxidáns aktivitás
32	daidzein	da	H	OH	OH	32.9
33	formononetin	fm	H	OH	OMe	-20.4
34	genisztein	ge	OH	OH	OH	-24.6
35	biochanin A	bi	OH	OH	OMe	-20.4

Szám	Név	Röv.	Antioxidáns aktivitás
36	kumesztrol	cu	38.7

*: O-Glu: glükóz, O-rha: ramnóz, O-rut: rutin, O-neohesp: neoheszperidin
A cukrok a glikozidos OH csoportjukon kapcsolódnak a flavonoidokhoz.

NYILATKOZAT

Alulírott Farkas Orsolya .kijelentem, hogy ezt a doktori értekezést magam készítettem és abban csak a megadott forrásokat használtam fel. Minden olyan részt, amelyet szó szerint, vagy azonos tartalomban, de átfogalmazva más forrásból átvettem, egyértelműen, a forrás megadásával megjelöltem.

aláírás

Budapest, 2007.10.23.