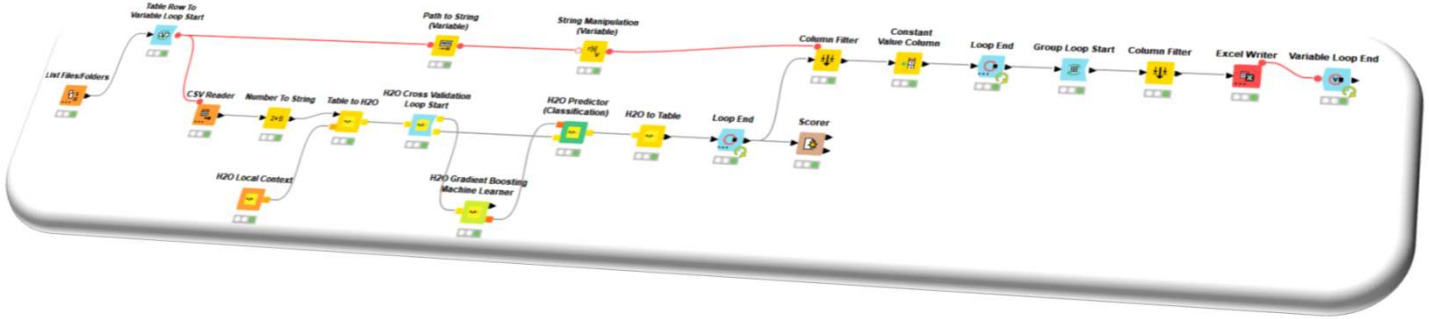
# Gépi tanulás KNIME környezetben: példák a gyógyszer- és anyagtudomány területéről

**Bajusz-Rácz Anita**

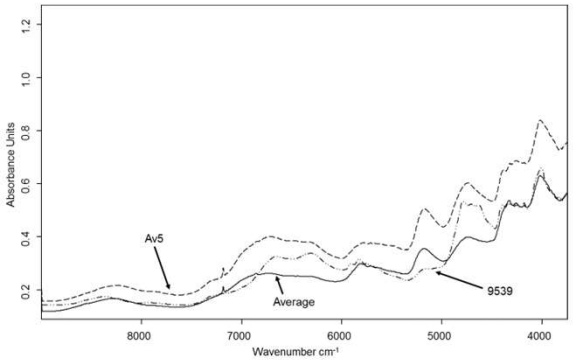
HUN-REN TTK Anyag- és Környezetkémiai Intézet

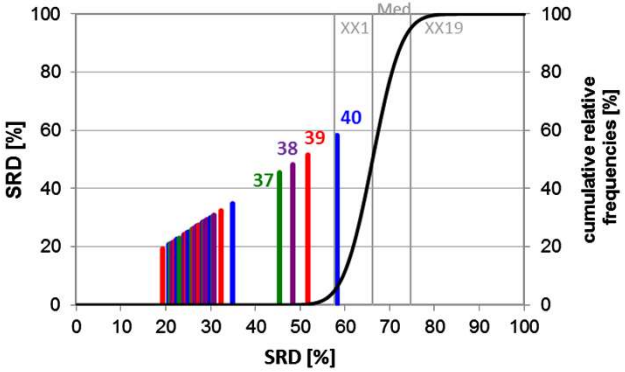


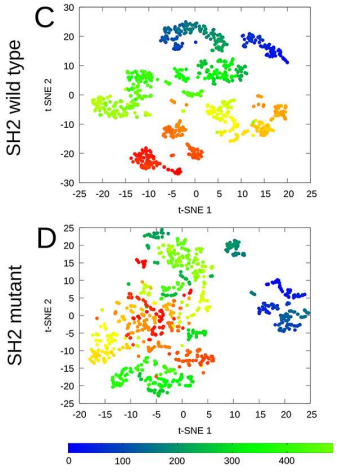

*Occam borotvája vagy gépi tanulás Workshop, Szeged, 2024. március 8.*

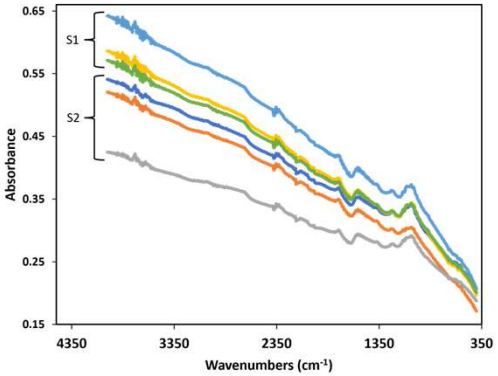
# Bevezetés

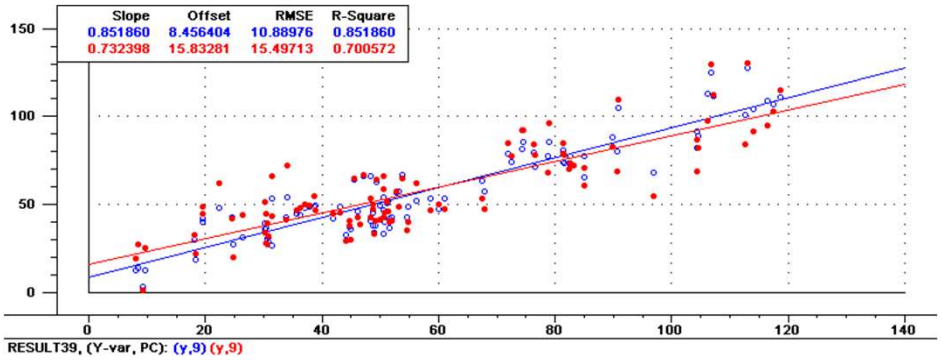
- Honnan indultam? Hova érkeztem?

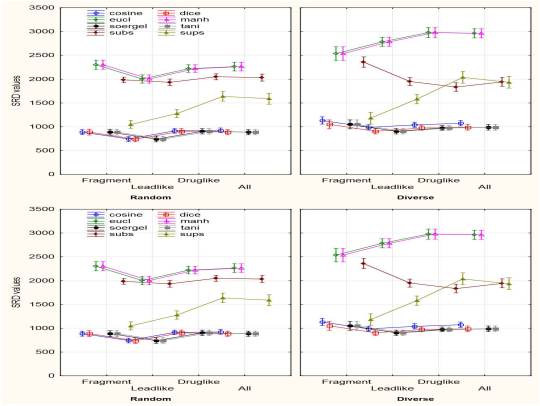


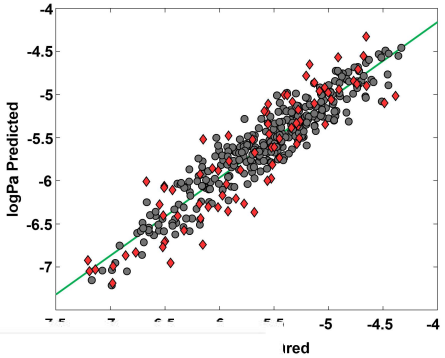



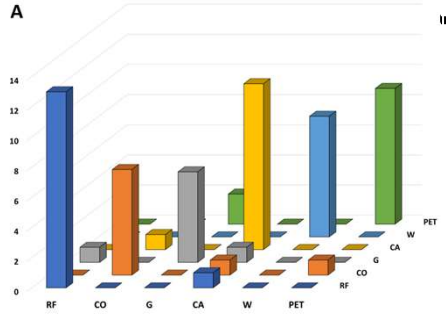




Klasszikus → ML

# Miért éppen KNIME?

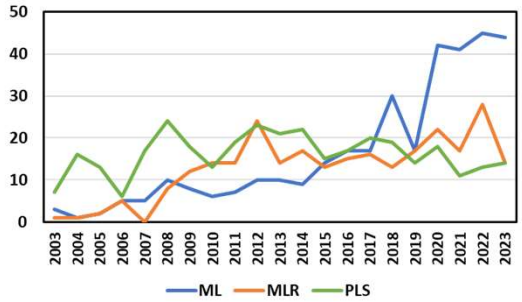
- Egyre nagyobb igény az ML módszerek használatára
- KNIME: híd a klasszikus statisztikai szoftverek és a programozás világa között
- Egyszerű használat – GUI és drag-and-drop mechanizmus
- Kiterjeszthetőség: Integrálható eszközök és algoritmusok más szoftverkörnyezetből (R, Python, Java, Weka stb.)
- Ingyenesen elérhető
- Közösségi támogatás

# Miért a gyógyszerkutatás és miért az anyagtudomány?

- „Témát lehet adni és kapni. A téma adódik. Kialakulhat. Létrejöhet és beindulhat. A témával kapcsolatban minden mozgásforma lehetséges, csak éppen a választás lehetősége kizárt.”

(Ezésez Géza karrierje)

- A gyógyszerkutatás – választás
- Az anyagtudomány – létrejött
- Gyógyszerkutatás: QSAR, reakció tervezés stb. – több évtizedes hagyomány
- Anyagtudomány: feltörekvő trend




„machine learning”/”pls”/”mlr” AND „qsar” AND „drug design”

nature reviews drug discover

Explore content About the journal Publis

nature > nature reviews drug discovery > review artic

Review Article | Published: 11 April 2019

**Applications of machine lear development**

Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer & Shanrong Zhao

nature materials

Explore content About the journal Publish with us

nature > nature materials > perspectives > article

Perspective | Published: 18 April 2019

**Exploiting machine learning for e discovery and development**

Sean Ekins, Ana C. Puhl, Kimberley M. Zorn, Thomas R. Lane, Hickey & Alex M. Clark

nature > npj computational materials > review articles > article

Review Article | Open access | Published: 08 August 2019

**Recent advances and applications of machine learning in solid-state materials science**

Review Article | Open access | Published: 05 April 2022

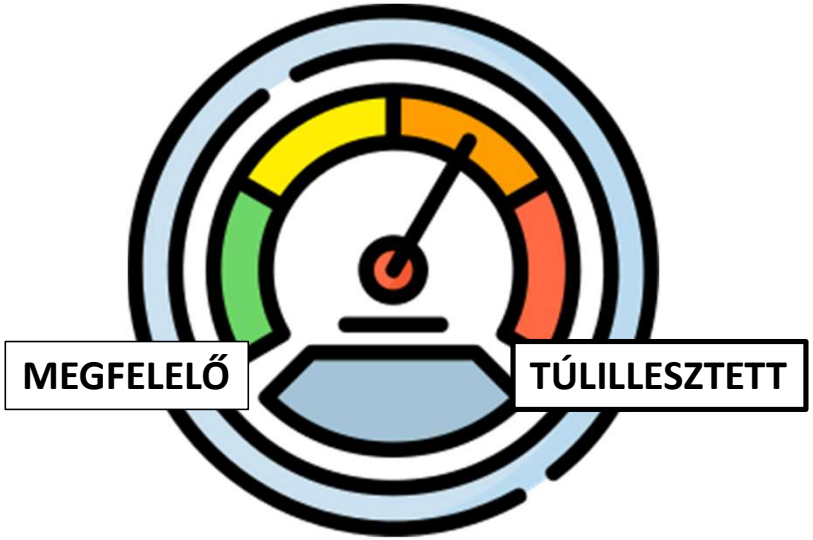
**Recent advances and applications of deep learning methods in materials science**

Kamal Choudhary, Brian DeCost, Chi Chen, Anubhav Jain, Francesca Tavazza, Ryan Cohn, Cheol Woo Park, Alok Choudhary, Ankit Agrawal, Simon J. L. Billinge, Elizabeth Holm, Shyue Ping Ong & Chris Wolverton

*npj Computational Materials* **8**, Article number: 59 (2022) | Cite this article

# Machine learning – Gépi tanulás osztályozási és regressziós feladatokra

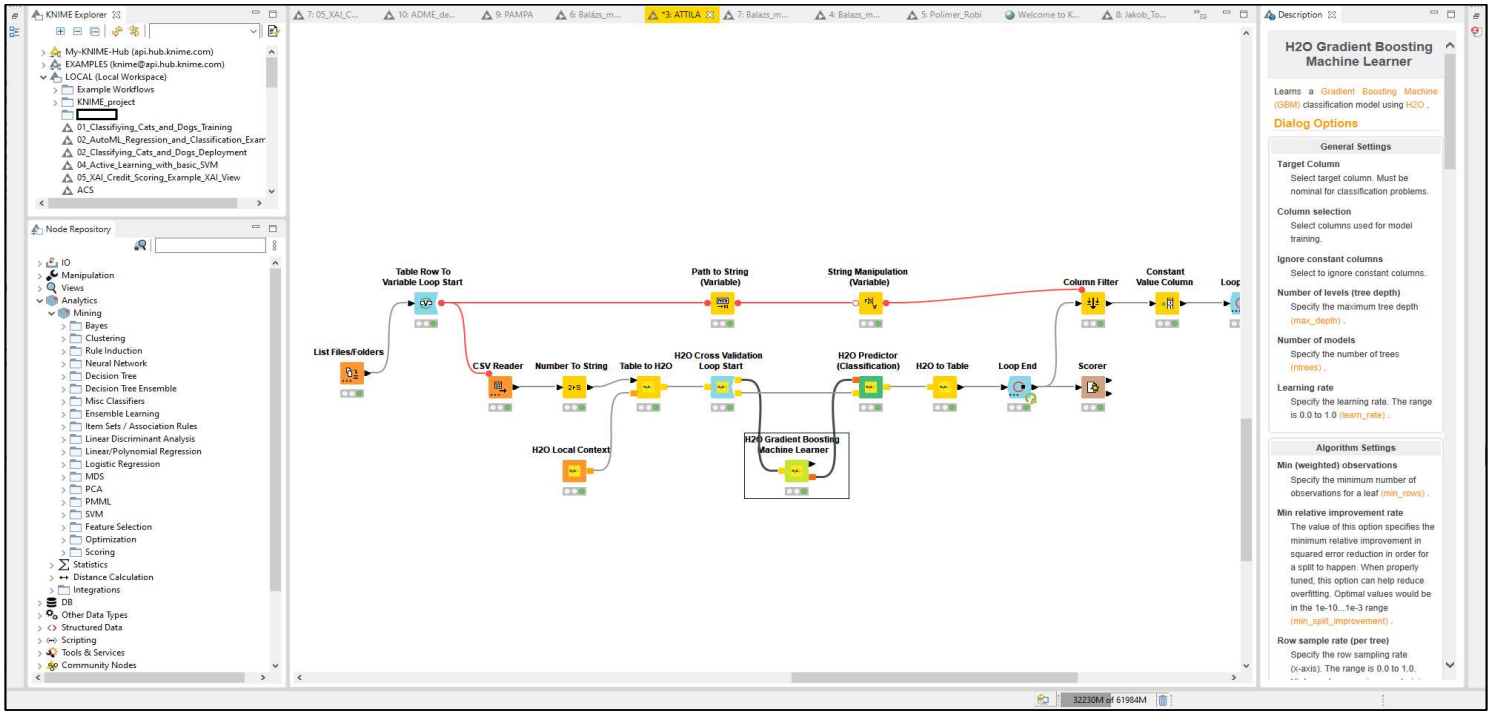
- A méret a lényeg?
  - Regresszió: igen, Klasszifikáció: nem feltétlen
- Linearitás vs. nemlinearitás
- Üres cellák?
- Dimenziók száma, dimenzió redukció
- Modell validálás és teljesítmény értékelés
  - Túlillesztett modell kiszűrése
- Leggyakrabban használt algoritmusok: ANN, Fa-alapú (XGBoost, Random forest), SVM



# A KNIME szoftverkörnyezet

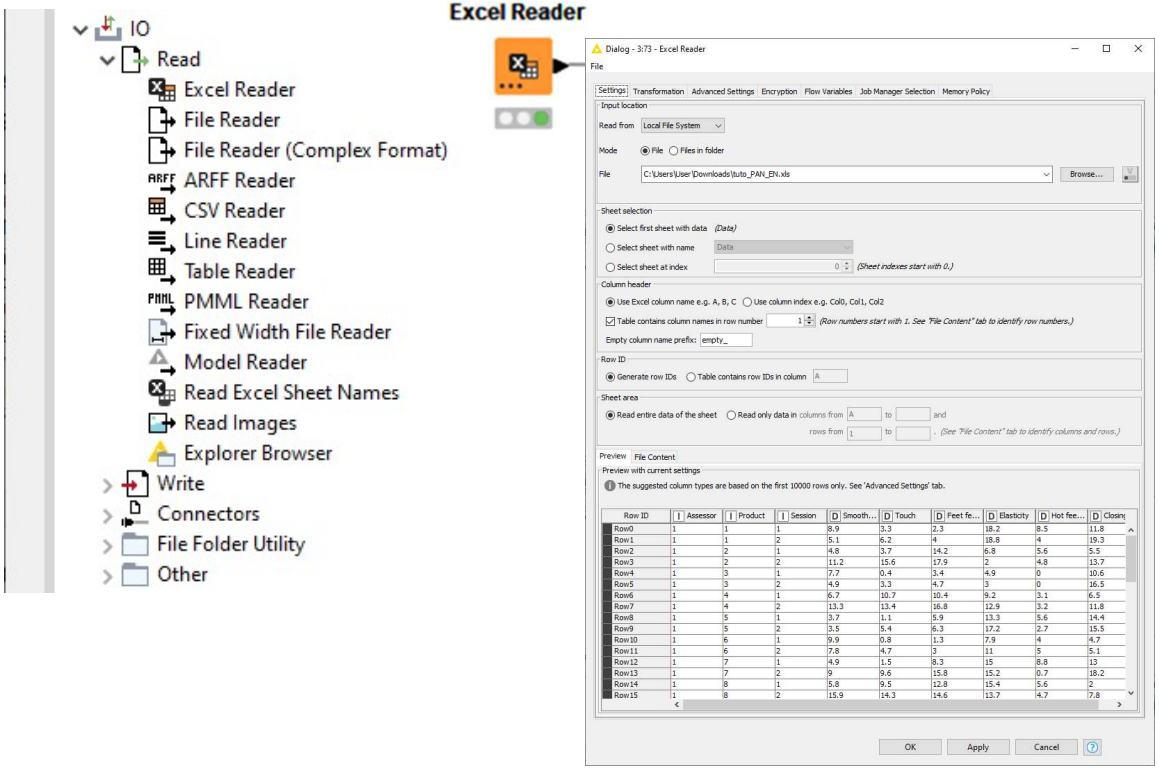
- KNIME: drag-and-drop alapú programozás
- Definíciók: Workspace, workflow, node
- A vonalak jelentik a kapcsolatot az egységek között (adatáramlás)
- Node repository
- Explorer
- Node description
- Console

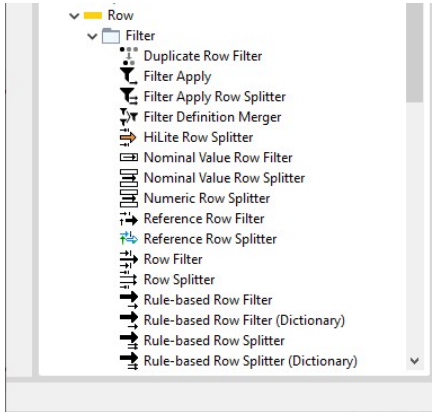

**vs.**

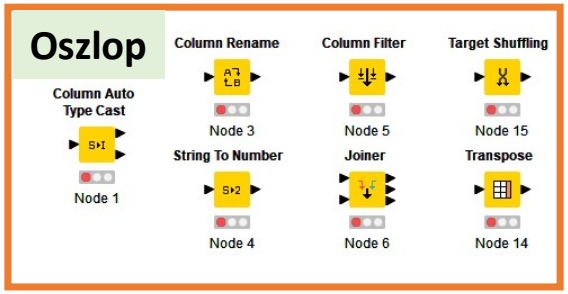


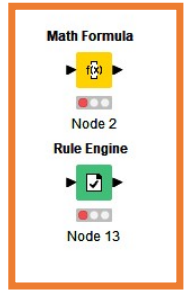
# Legfontosabb Node egységek - adatelőkezelés

- Filterelés, szelektálás, vágás, módosítás

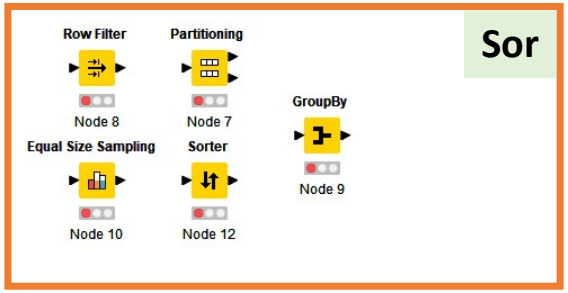


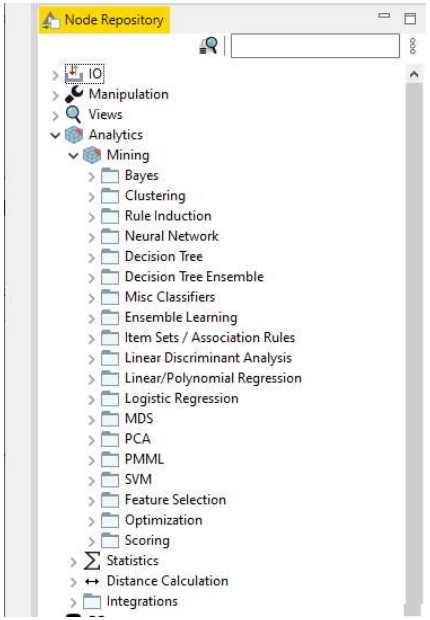



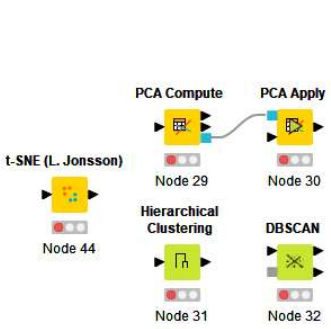

Matematikai műveletek

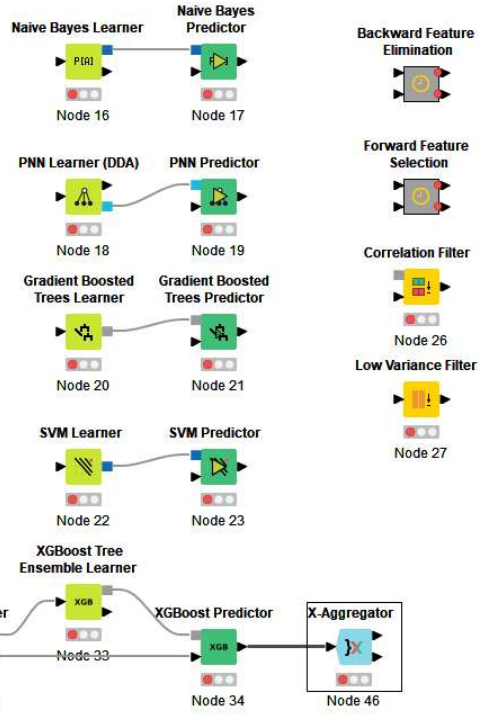


# Legfontosabb Node egységek - Modellezés

- Beépített egység csomagok

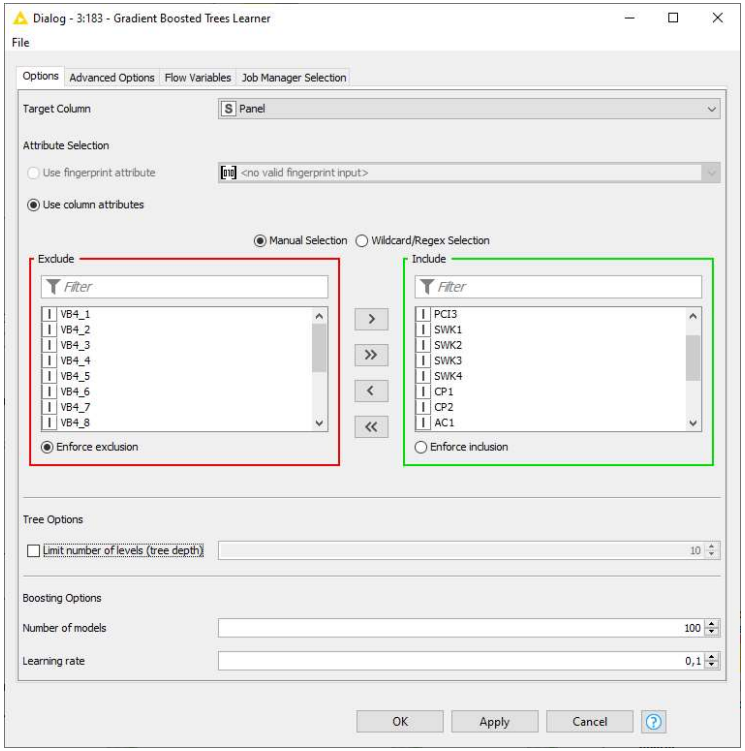


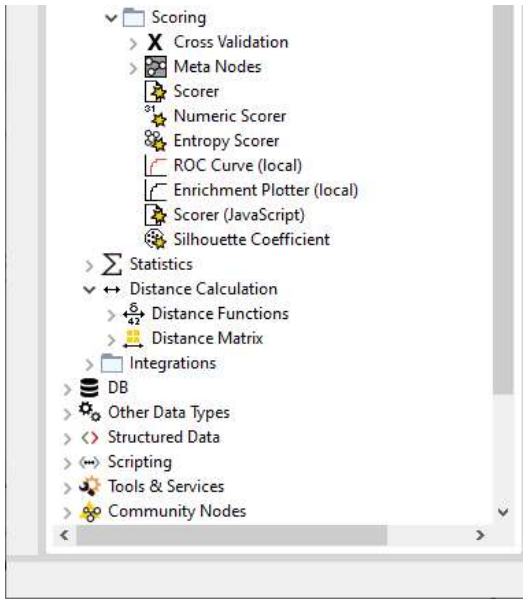


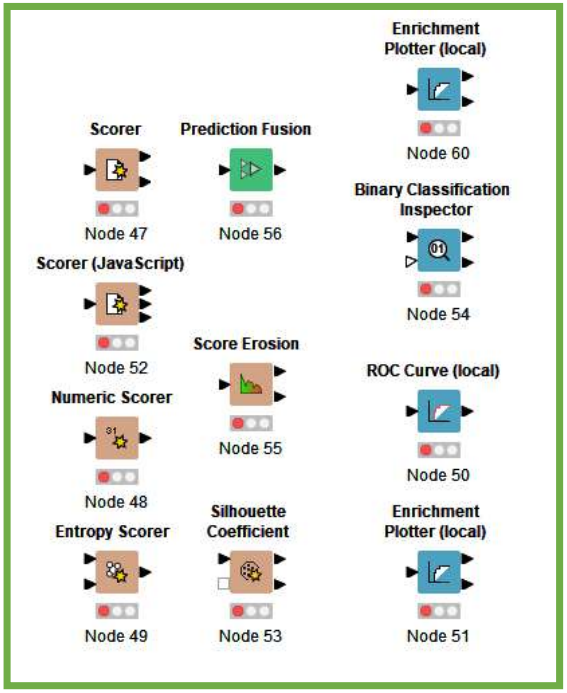


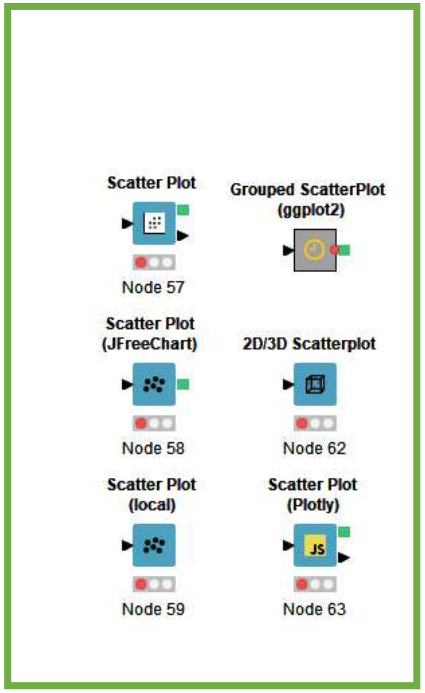


# Legfontosabb Node egységek - Kiértékelés

- Teljesítmény paraméterek, valószínűségi értékek transzformálása, vizualizáció, exportálás

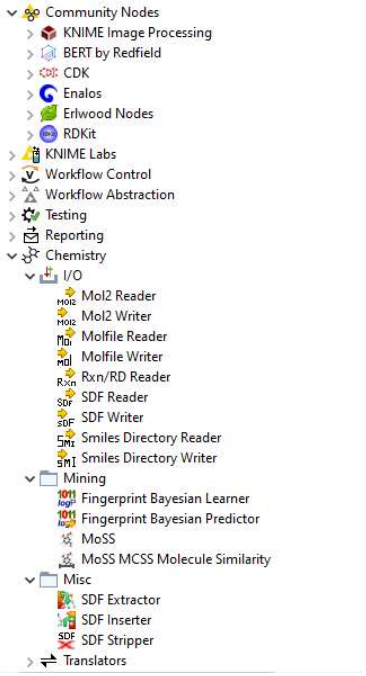


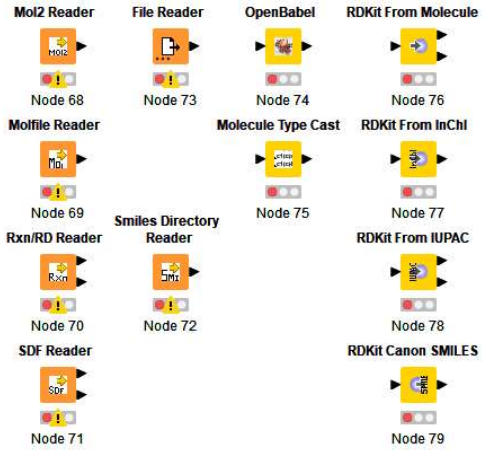



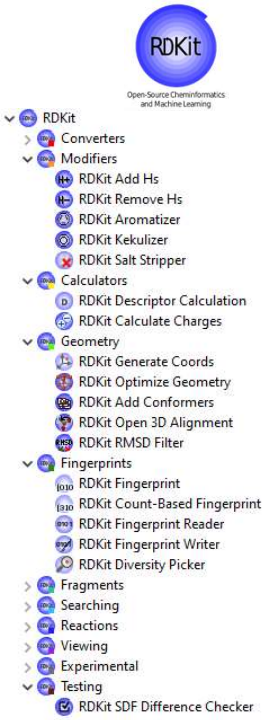



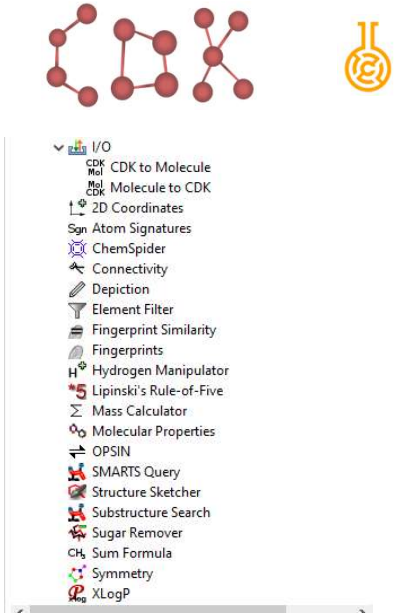

# KNIME a gyógyszerkutatásban (QSAR/QSPR) - adatelőkészítés, adat generálás

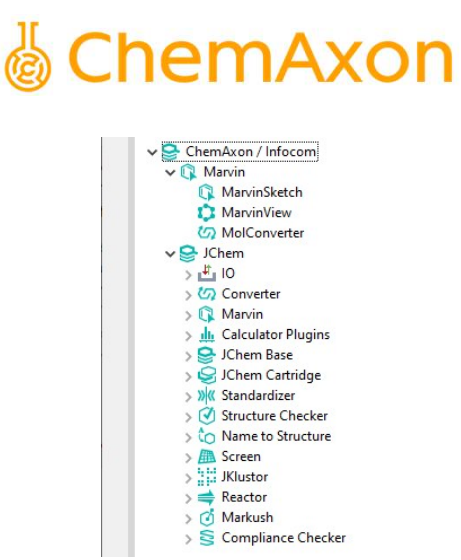
- Gyógyszerkutatásban, keminformatikában használható node-ok
- Közösségi (Community) node csomagok



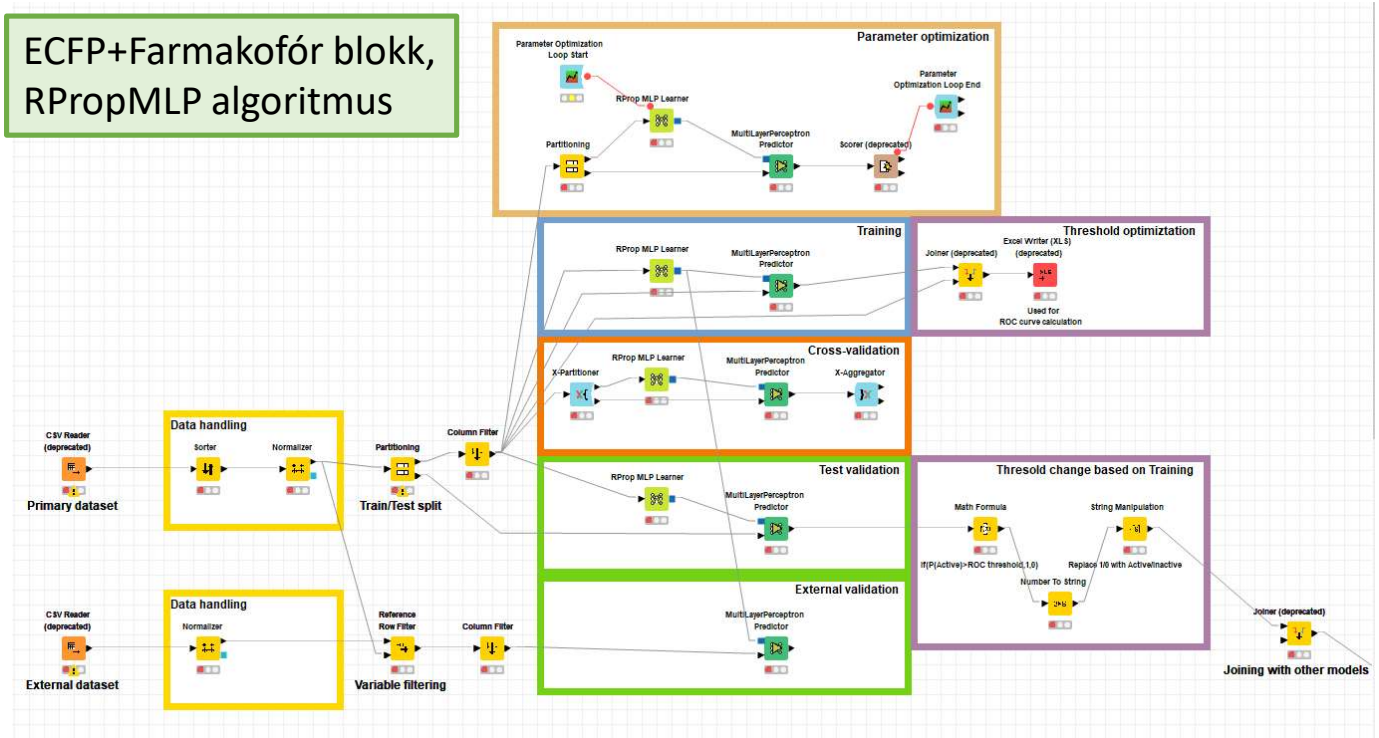



# ADME modellezés KNIME környezetben

- Cytochrome P450 (CYP) 2C9 izoenzimjén való aktivitás
- 45000 különböző molekula
- XGBoost, RPropMLP algoritmus
- 1D,2D,3D deszkriptorok, ECFP + Farmakofór ujjlenyomatok, Kölcsönhatás ujjlenyomat + Dokkolás

ECFP+Farmakofór blokk, RPropMLP algoritmus

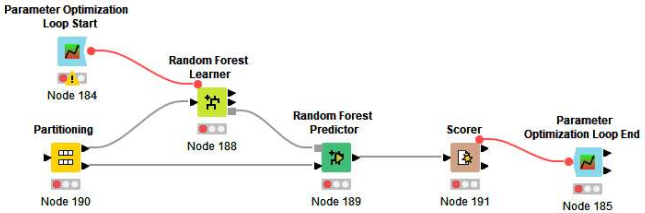


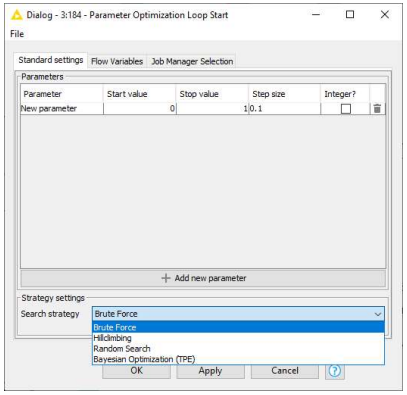

- Valószínűségi limitértékek optimálása
- Konszenzus voksolás a hat modell alapján
- AUC: 0.85 (Belső teszt), 0.84 (Külső teszt)

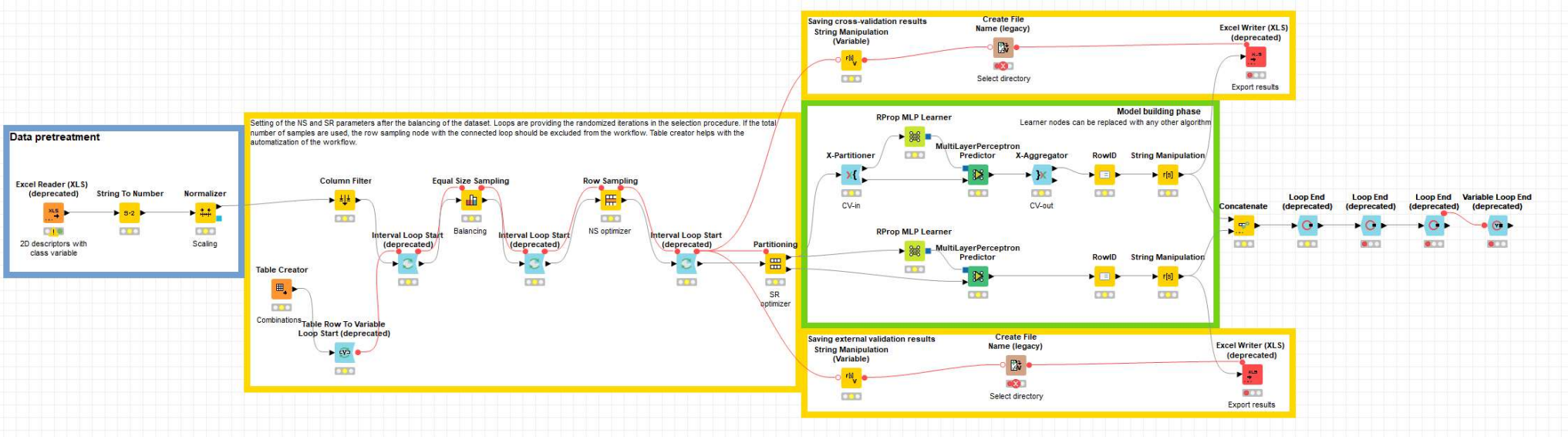
*Rácz, A., Keserű, G.M. Large-scale evaluation of cytochrome P450 2C9 mediated drug interaction potential with machine learning-based consensus modeling. J Comput Aided Mol Des **34**, 831–839 (2020). https://doi.org/10.1007/s10822-020-00308-y*

# Automatizálhatósági lehetőségek

- Probléma specifikus loop egységek
- A loop start – loop end node-ok
- Paraméter optimálás
- Iteratív modell futtatások



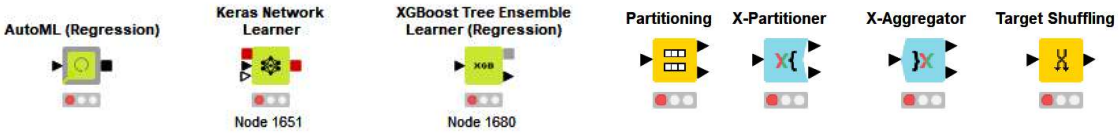




*Rácz, A.; Bajusz, D.; Héberger, K. Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification. Molecules* ***2021****, 26, 1111. https://doi.org/10.3390/molecules26041111*

# Anyagtudomány és spektrumok KNIME környezetben

- **Osztályozás**
  - Polietilén minták osztályozása adalékanyag szerint
  - Szén nanostruktúrák osztályozása a kiindulási alapanyag szempontjából
- **Regresszió**
  - Polimer minták merevségének meghatározása
- Spektrum adatelőkezelés: előzetesen kivitelezhető
- Outlier szelekció csak numerikus szinten elérhető
- Modellezés: beépített egységek mellett – Keras, Sklearn, R stb. csomagok



# Osztályozás

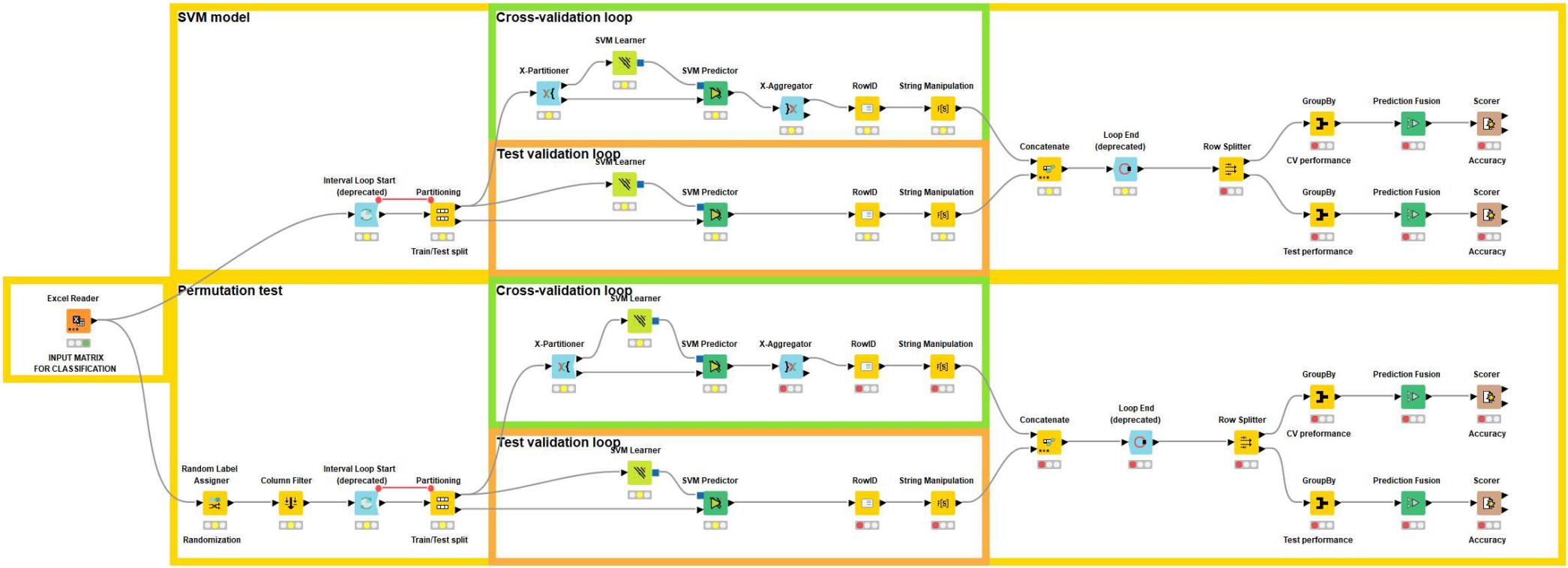
## Szén nanostruktúrák kiindulási alapanyag szerinti osztályozása

- 62 minta
- ATR-IR spektrum (1867 változó) – alapvonal korrekció + deriválás + SNV
- 6 csoport

## Polietilén minták adalékanyag szerinti osztályozása

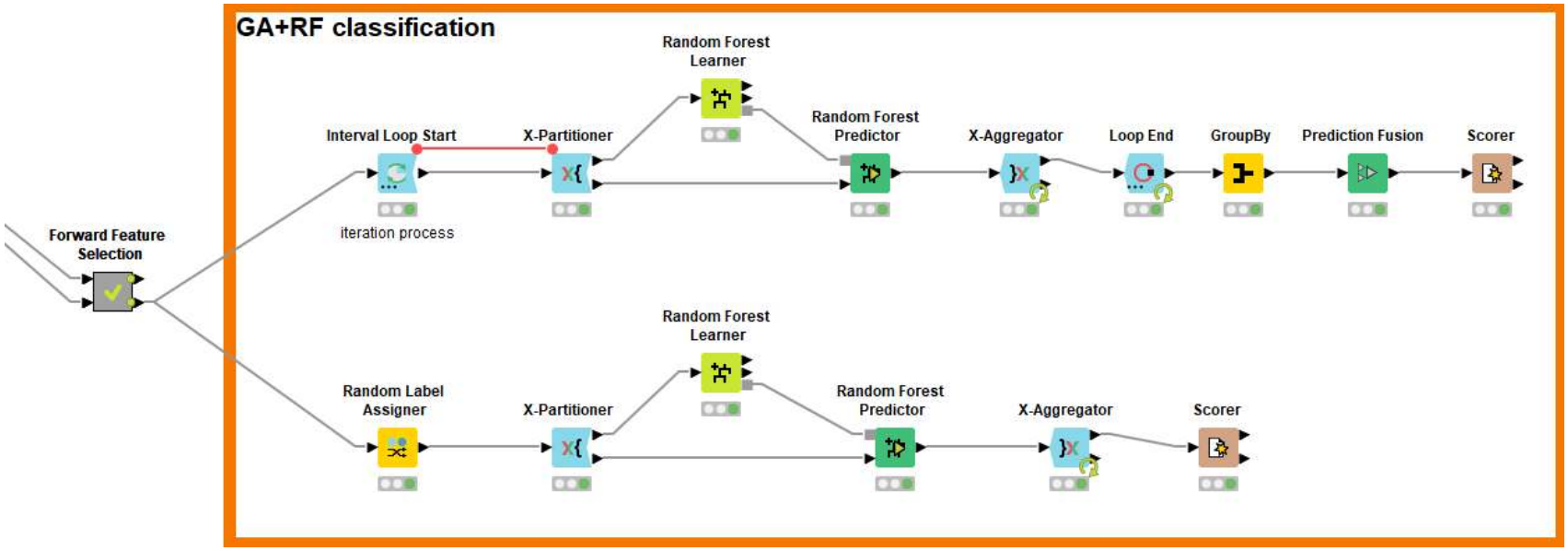
- 36 minta
- ATR-IR spektrum (3733 változó) – standardizálás
- 4 csoport

- **Szén nanostruktúrák kiindulási alapanyag szerinti osztályozása**
  - SVM algoritmus
  - Double-CV iteratív módon
  - Permutációs teszt – túlillesztés elkerülése
  - BACC: 0.865 (CV)




*A Rácz, K László, S Klébert: Qualitative and quantitative chemometric modelling of nanostructured carbon samples based on infrared spectroscopy. Carbon, 2024*

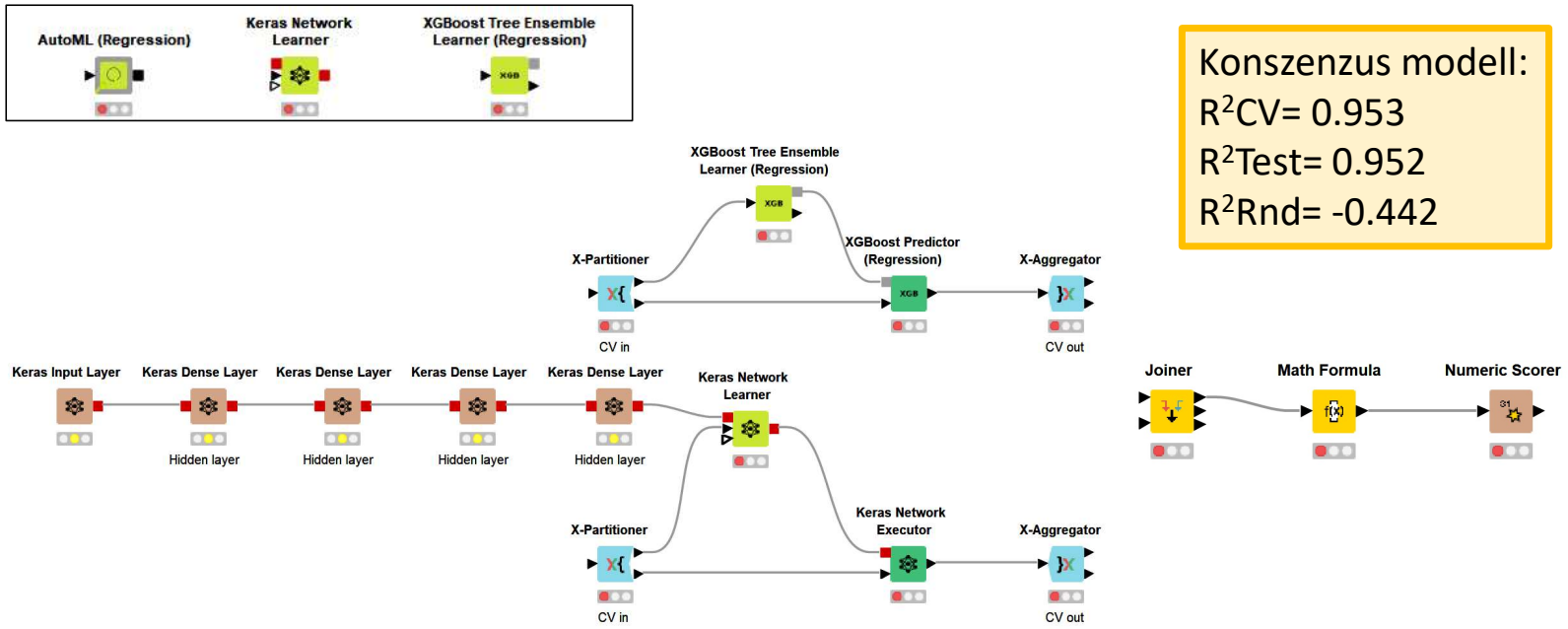
- **Polietilén minták adalékanyag szerinti osztályozása**
  - Genetikus algoritmus
  - Random Forest
  - 3-fold CV 10x iterációban
  - Permutációs teszt használata
  - BACC: 0.917




*A Rácz, D Tátraaljai, S Klébert: Determination of melt flow index and polymer additives in polyethylene based on IR spectra and multivariate modeling. Materials Today Chemistry, 2023*

# Regressziós modell konszenzus értékeléssel

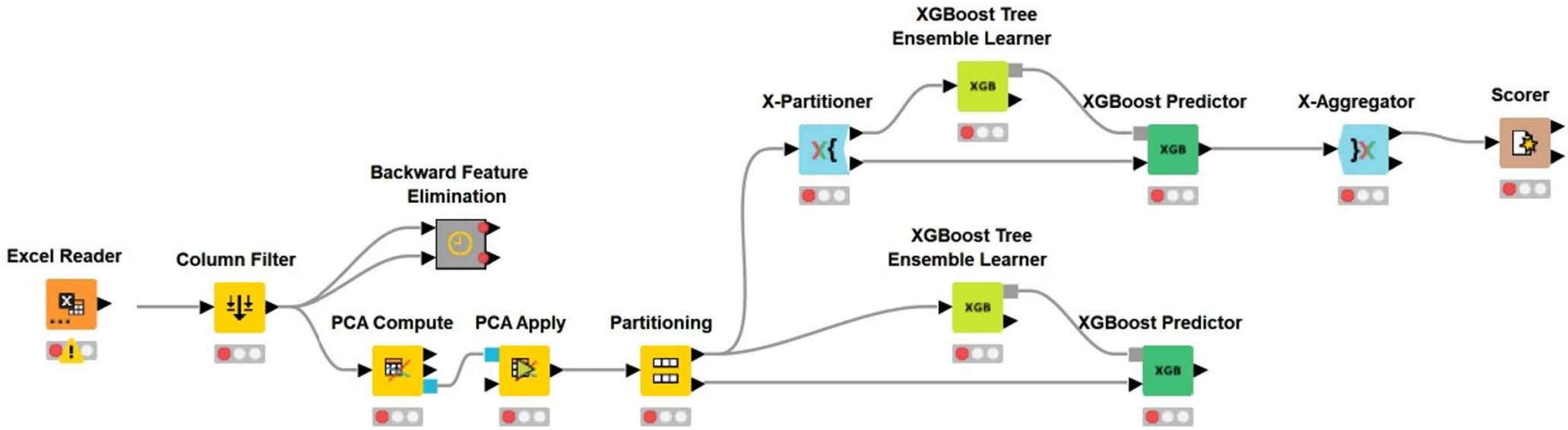
- Polimer minták merevségének predikciója
  - 195 minta ATR-IR spektruma
  - Y: merevség (modulus)
  - Spektrum előkezelés: Deriválás+SNV – 3451 változó




Konszenzus modell:
$R^2$CV= 0.953
$R^2$Test= 0.952
$R^2$Rnd= -0.442

# KNIME workflow építése lépésről lépésre

- Az adat importálástól a modellezésig

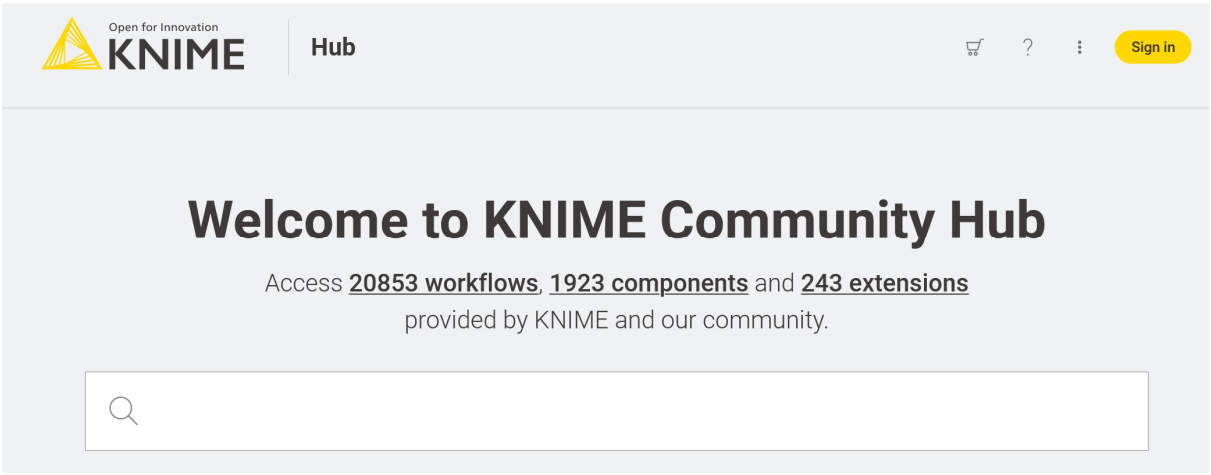


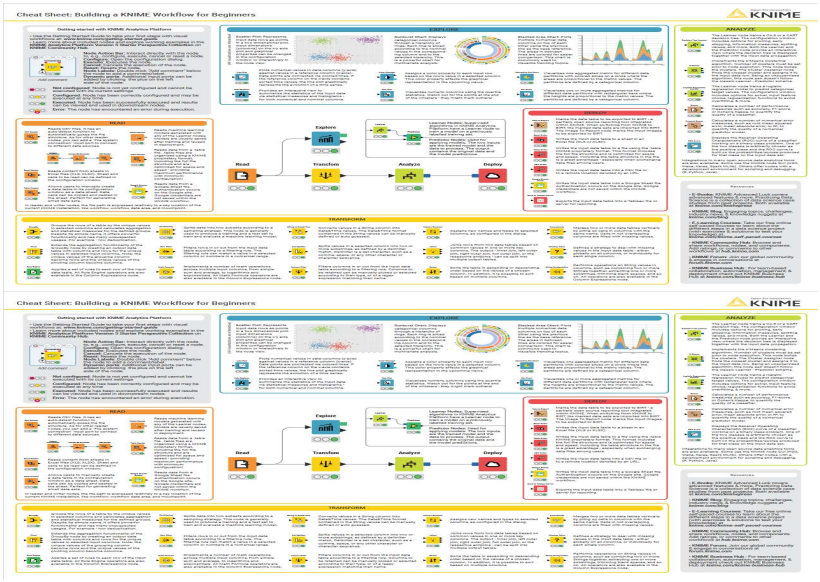
# KNIME: kihívások és limitációk

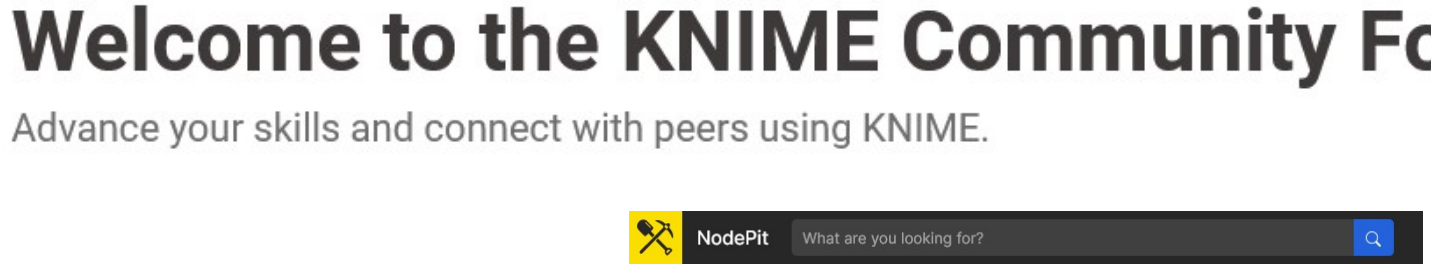
- Bonyolult regressziós workflow-k
- Blokkonkénti változó szelekció hiánya
- Spektrum adatelőkezelés
- Memória/disc igényes
- Bonyolultan az egyszerűt
- Vizualizáció
- Batch processing vs. Real time analysis

# Hasznos oldalak a KNIME használatához

- CheatSheets: https://www.knime.com/knimepress#cheat-sheets
- KNIME Hub, KNIME community forum, NodePit

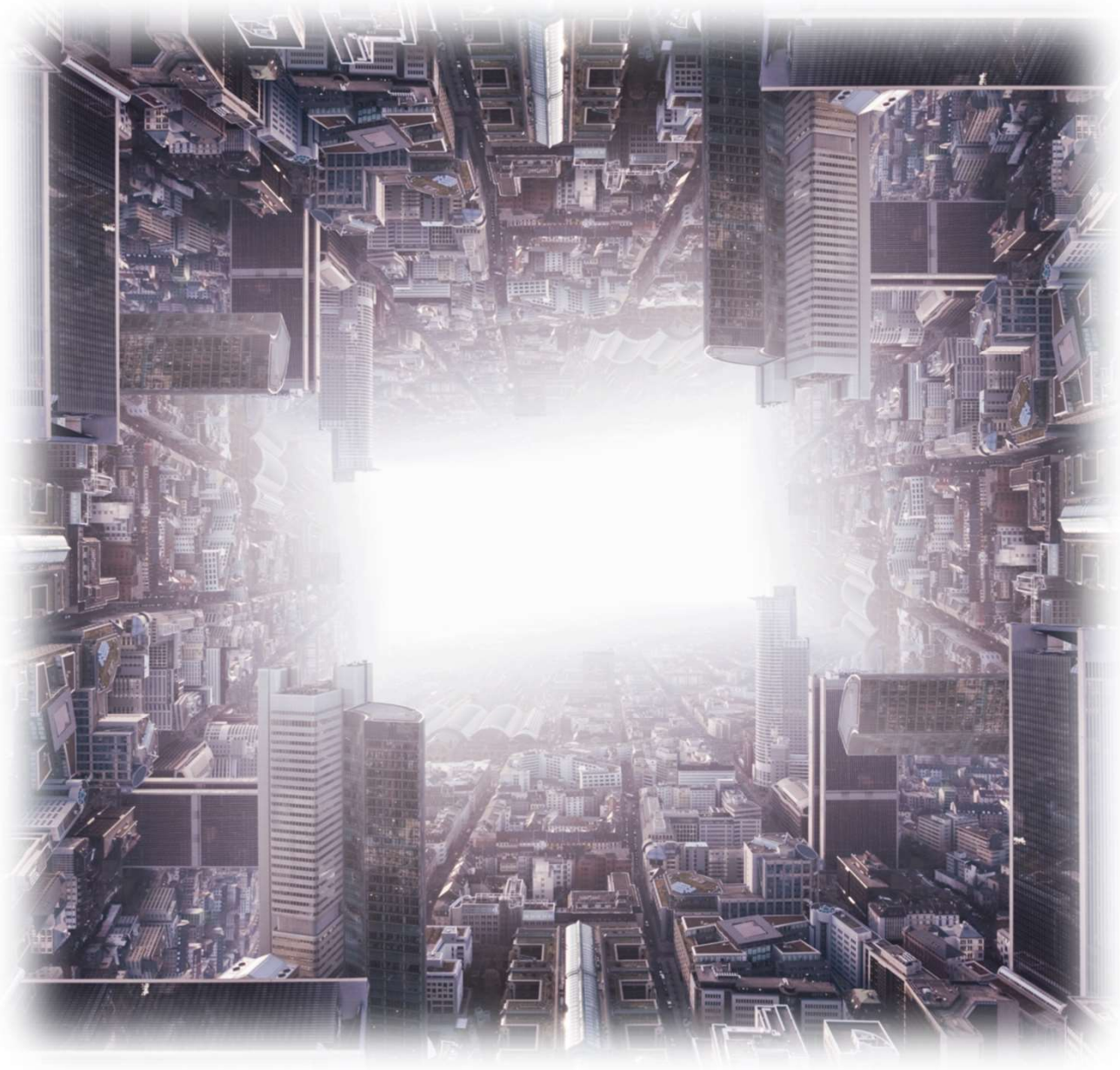


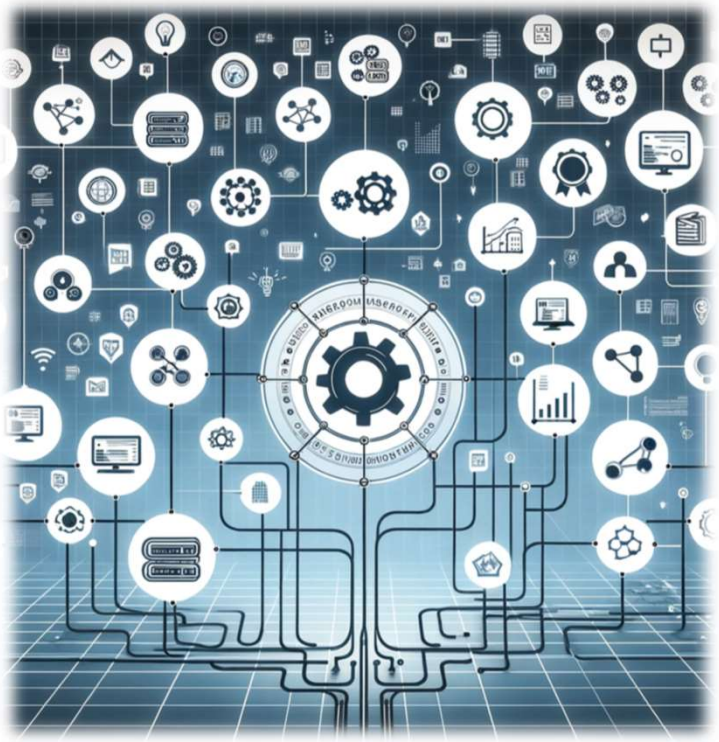
# KNIME: jövőbeli lehetőségek

- Felhő alapú adatelemzés és modellezés
- Automatikus ML modellezés
- ML interpretációs lehetőségek
- Regressziós modellezés bővítése

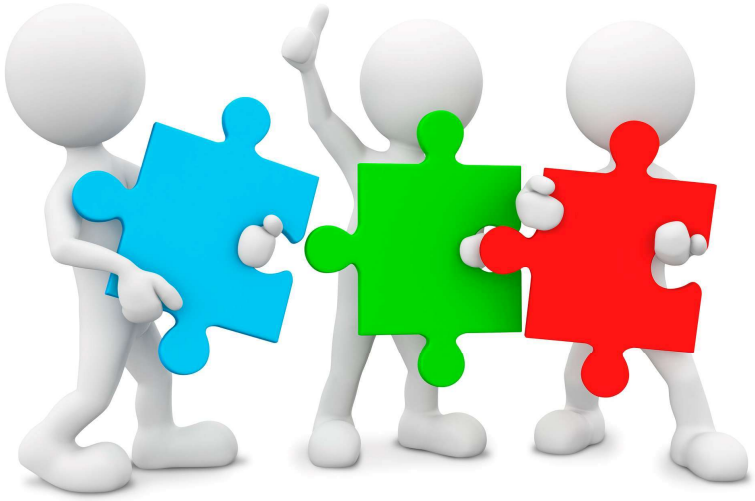


# Összefoglalás

- Vizuális megjelenés = logikusabb, átláthatóbb a programot nem ismerő számára is
- Egyre több integráció, „community package”, ML algoritmus
- Nagyon jól használható keminformatikai és analitikai feladatokban
- Spektrum adatoknál „hybrid”-ként kezelhető
- Klasszifikációs algoritmusok száma végtelen
- A határait csak a számítógépes erőforrás jelenti



# Köszönetnyílvánítás



- Héberger Károly
- Bajusz Dávid
- Keserű György Miklós
- Klébert Szilvia
- Tátraaljai Dóra
- László Krisztina
- Várdai Róbert
- Plazmakémiai Kutatócsoport

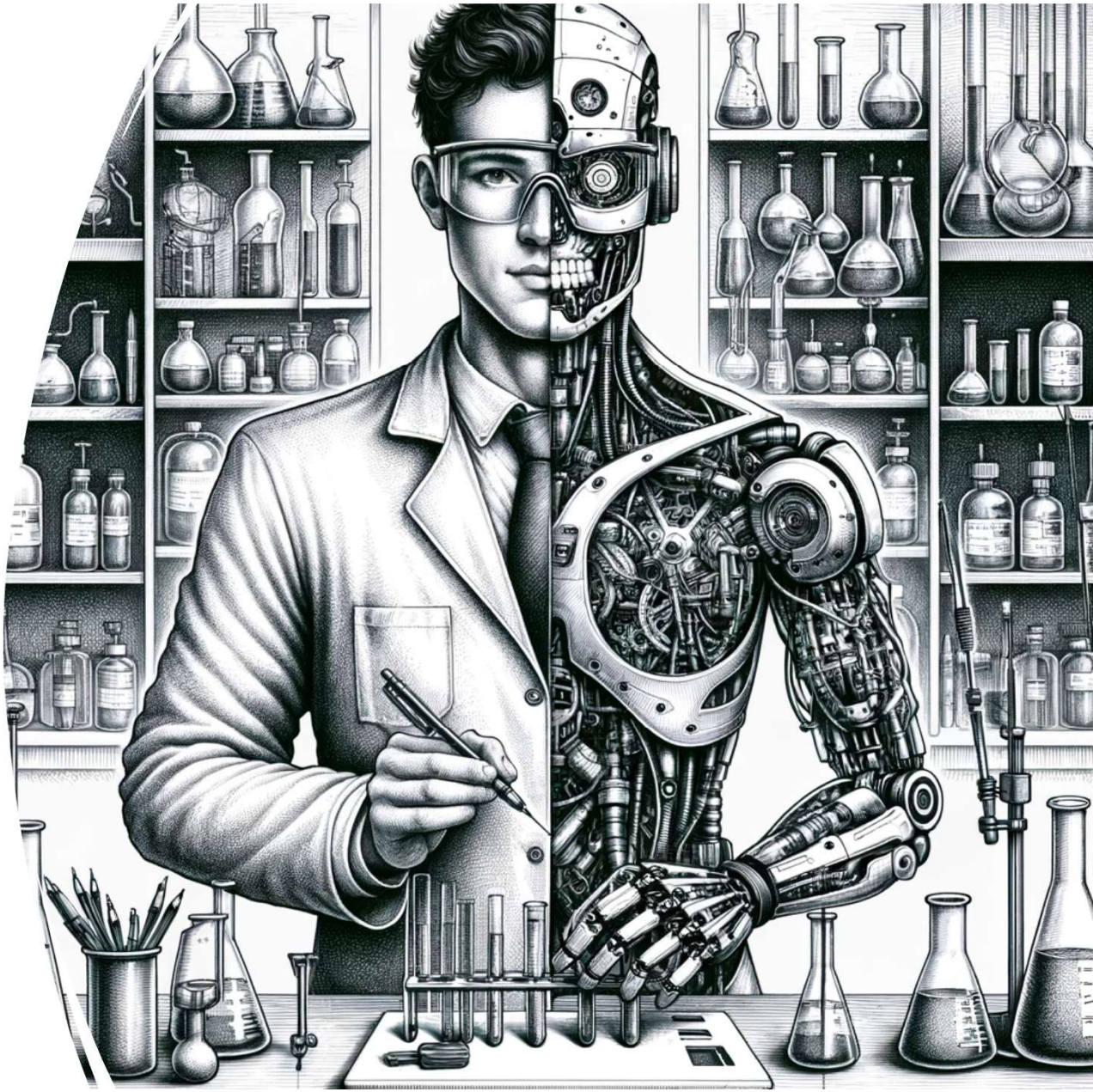
---


- OTKA K134260
- A Bolyai János Kutatási Ösztöndíj támogatásával készült.
- AZ INNOVÁCIÓS ÉS TECHNOLÓGIAI MINISZTÉRIUM ÚNKP-23-5 KÓDSZÁMÚ ÚJ NEMZETI KIVÁLÓSÁG PROGRAMJÁNAK A NEMZETI KUTATÁSI, FEJLESZTÉSI ÉS INNOVÁCIÓS ALAPBÓL FINANSZÍROZOTT SZAKMAI TÁMOGATÁSÁVAL KÉSZÜLT.

NEMZETI KUTATÁSI, FEJLESZTÉSI ÉS INNOVÁCIÓS HIVATAL

Új Nemzeti Kiválóság Program

# Köszönöm a figyelmet!



*„Draw me a half robot half human chemist in pen draw style”*