# Advanced programming Group Assignment Bevindingen Week 1 – Bryan van der Ven

## Welke manieren om moleculen te beschrijven zijn er in Rdkit?

Fysisch-chemische eigenschappen: molecuulgewicht, topologish polair oppervlak (TPSA), octanol-water-coefficient (logP)

Structurele eigenschappen: aantallen van specifieke substructuren

Binary fingerprints: vectoren die de aanwezigheid van specifieke substructuren binnen het molecuul weergeven. Veel voorkomende binary fingerprints zijn Morgan finger prints, MACCS keys, Topological Torsion fingerprints. Het gebruiken van binary fingerprints is waardevol voor clusteren van grote hoeveelheden typen moleculen

RDkit heeft ook de mogelijkheid om 3D eigenschappen van moleculen te vergelijken en ook electrostatische eigenschappen. Echter, deze worden niet benoemd in de beschrijving van de opgave op Canvas, dus het kan zijn dat deze misschien wat lastiger zijn om mee te werken. Maar misschien wel waardevol om te kijken of ze gebruikt worden in soortgelijk onderzoek, want misschien dat we er dan toch ook naar kunnen kijken?

## Andere tools naast RDkit?

Er zijn er best wel veel. Maar ik denk dat de meest waardevolle waar we misschien naar willen kijken heet ChemDes (<https://github.com/ifyoungnet/ChemDes>) Deze combineert meerdere pakketten (onder andere ook RDkit!) en levert in totaal 3679 moleculaire descriptoren en 59 typen fingerprints. Maar omdat ons project relatief kort is en de begeleiding van dit vak RDkit voorstelt denk ik dat het ook geen kwaad kan om daar vooralsnog mee verder te gaan.

## Hoe kunnen we deze manieren gebruiken in de voorspellng?

Nadat we de SMILES correct hebben ingeladen, denk ik dat het belangrijk is dat we eerst een uitgbereide EDA uitvoeren gebaseerd op een paar veel voorkomende descriptoren en fingerprints die worden gebruikt in soortgelijke onderzoeken en/of PCA.

## Zijn er al bekende methodes?

In de vergelijkbare onderzoeken die ik heb bekeken komt het er eigenlijk op neer dat een machine learning techniek (zoals random forest variable importance measures, K-nearest neighbor, logistic regression, support vector macine) wordt gebruikt om een uiteindelijke keuze te maken welke descriptoren en fingerprints er uiteindelijk gebruikt gaan worden voor de voorspellingen van inhibitors. Ook wordt er vaak ongeveer tussen de 2 en 6 verschillende soorten fingerprints gebruikt, dus ik denk dat het wel belangriijk is dat we die sowieso gaan gebruiken.

## Hoe nu verder?

Mij maakt het niet heel erg veel uit welke taken door mij uitgevoerd moeten worden. Ik denk dat het voor de volgende meeting belangrijk is dat we een duidelijk idee hebben hoe onze data eruit ziet en wat precies onze aanpak gaat zijn (vooral lettend op wat er van ons verwacht wordt in het format van de paper).

## Bronnen gebruikt

<https://xinhaoli74.github.io/blog/rdkit/2021/01/06/rdkit.html>

<https://github.com/ifyoungnet/ChemDes>

<https://ieeexplore.ieee.org/document/8950239>

<https://www.mdpi.com/2227-9717/9/11/2074>