Molekylære egenskaber:

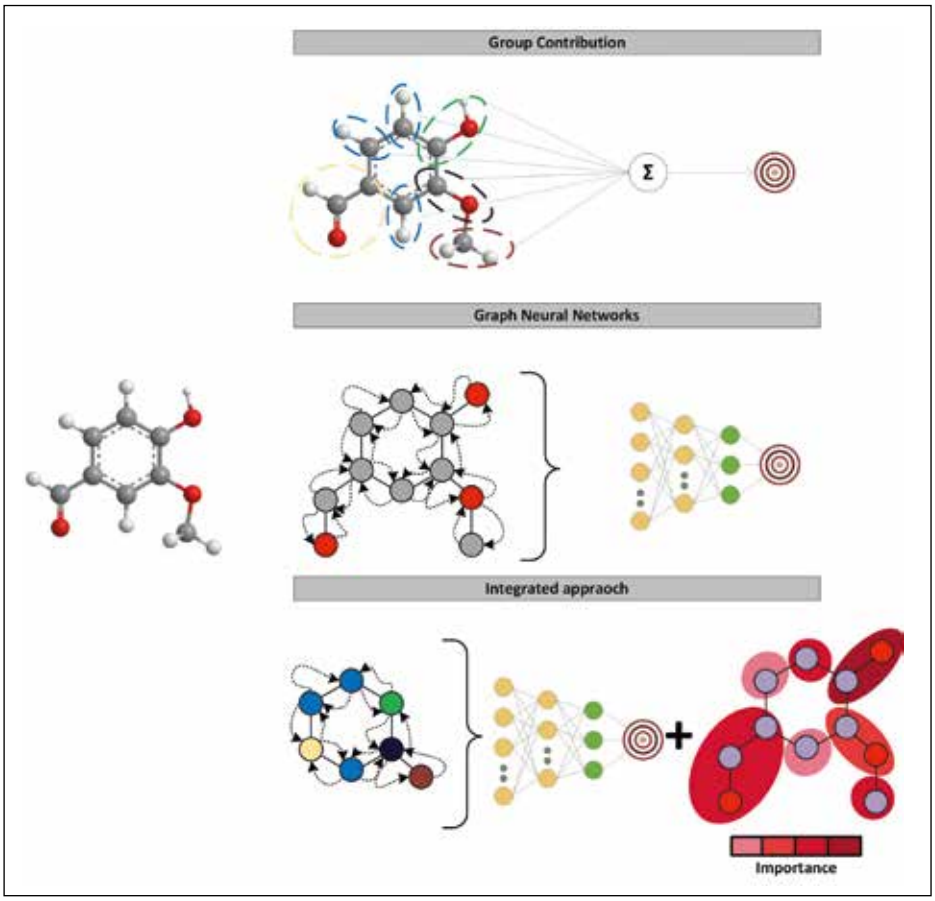
Kunstig intelligens baner vejen for nye og bedre beregningsmetoder

Kunstig intelligens gør det muligt at udvikle ekspertsystemer, der er i stand til at udtrække vigtige strukturelle oplysninger fra molekylet og lære at korrelere disse til forskellige egenskaber af interesse.

Af Adem Rosenkvist Nielsen Aouichaoui, Jens Abildskov og Gürkan Sin, Process and Systems Engineering center (Prosys), DTU Kemiteknik

En ofte brugt kemivits lyder som følger: ”Don’t trust molecules, they make up everything (de udgør alt)”. I den vittighed gemmer der sig en universel kendsgerning: Molekyler er byggestenene i alt omkring os, både levende og livløse, naturlige og syntetiske. Kemiske produkter er vidt udbredte og anvendes til forskellige formål i en lang række industrier som for eksempel sundhedspleje, energisystemer, overflade- og afgrødebeskyttelse. Processerne til fremstilling og oprensning af disse kemikalier involverer også andre produkter såsom reaktanter og tilsætningsstoffer for at muliggøre adskillelsen og oprensningen af de ønskede produkter.

Et kemisk produkts anvendelsesmuligheder går hånd i hånd med dets forskellige egenskaber, både som ren komponent og som en del af en blanding. Disse egenskaber kan variere fra termiske, fysiske og miljømæssige til toksikologiske og sikkerhedsrelaterede egenskaber. De bestemmes normalt i laboratorier, der kræver en række forskellige forsøgsopstillinger, måleinstrumenter, procedurer og ekspertise. Det kræver omfattende arbejde at fastlægge disse egenskaber, især når man tager den enorme størrelse af det kemiske design-



Figur 1. Metoder til forudsigelse af molekylers egenskaber.
Øverst: gruppebidragsmetoden, hvor molekylets egenskab er lig summen af gruppernes bidrag. Midten: Graph Neural Networks, hvor molekylet er beskrevet som en Graph, hvor knudepunkterne repræsenterer atomer. Dette bruges som input til en Deep Neural Network for at beregne egenskaben. Nederst: en kombineret tilgang, hvor grupper anses som knudepunkter, hvilket muliggør visualiseringen af de mest betydningsfulde funktionelle grupper.

rum i betragtning. Et studie viste, at der teoretisk set er over 166 milliarder mulige organiske forbindelser, der består af op til 17 atomer af nitrogen (N), oxygen (O), svovl (S) og halogener (fluor (F), chlor (Cl), brom (Br), jod (I)) [1]. Det viser, hvor stor en opgave det ville være at bestemme alle molekylers egenskaber eksperimentelt under alle forhold.

Udnyttelse af det kendte til at udforske det ukendte

For kemiingeniører er det vigtigt at kende molekylers egenskaber ved procesdesign, opstilling af masse- og energibalancer samt processimulering. En vigtig anvendelse er produktdesign, som går ud på at identificere molekyler med særlige egenskaber og funktioner. Sådanne designproblemer kan omfatte identifikation af nye miljøvenlige kølemidler eller opløsningsmidler til oprensning af bioprodukter. Det har derfor været vigtig praksis at udvikle matematiske modeller, der kan forudsige egenskaber ud fra molekylets strukturelle information. En populær modeltype er gruppebidragsmodellen.

Modellens input er forekomsten af et sæt foruddefinerede understrukturer og funktionelle grupper, som hver især har en koefficient, der bidrager til den samlede egenskab gennem et lineært forhold [2]. Et eksempel på denne tilgang er illustreret i figur 1. Disse modeller er blevet rost for deres nøjagtighed og fortolkningsevne på trods af deres enkelhed, men deres lineære natur lykkes ikke altid med at beskrive egenskaber med ikke-lineær opførsel og molekyler, der udviser "nærmeste-nabo-effekter". Sidstnævnte skyldes i høj grad, at sådanne modeller ikke tager højde for den geometriske relation mellem grupperne.

Kunstig intelligens: Nye tilgange til modellering af molekylers egenskaber

Den hurtige udvikling i regnekraft og adgang til beregningsressourcer har i høj grad understøttet fremkomsten af kunstig intelligens (artificial intelligens, AI). Som så mange andre applikationer og områder har forskere forsøgt at udnytte dette til at overkomme ulemperne ved klassiske tilgange til modellering af kemiske egenskaber. Machine learning (ML), en underkategori af AI, er et sæt af algoritmer, der kan korrelere en række inputs til et givet output uden eksplicit at blive programmeret til det.

Ud over blot at korrelere inputs til et givet output, er nogle ML-metoder også i stand til at ekstrahere og lære nye repræsentationer ud fra forskellige inputformater. Det vil sige, i stedet for det besværlige konventionelle arbejde med at definere og identificere grupperne i et molekyle, kan dette outsources til en ML-algoritme. Graph Neural Networks er en type neuralt netværk, der opererer på en graf repræsentation som input og er en populær model inden for repræsentationslæring [3]. Grafer er i stand til at beskrive mange objekter og fænomener som forsyningskæder, sociale netværk og smittesporing. Grafer er også en populær 2D-repræsentation af molekylet: Knudepunkterne svarer til atomer, mens kanterne svarer til kemiske bindinger. Dette gør Graph Neural Networks til en intuitiv tilgang til at forudsige molekylers egenskaber. Diverse information relateret til molekylet kan inkorporeres i knudepunkterne og kanterne i en graf. Nogle af disse informationer er typen af atomer og kemiske bindinger (som en binær variabel), antallet af bindinger og hydrogenatomer (som en heltalsvariabel) samt information om chiralitet og hybridisering. Et eksempel på en molekylær graf kan ses i figur 1.

Graph Neural Networks fungerer meget på samme måde som den menneskelige hjerne: Informationen i knudepunkterne (neuronerne) transporteres langs kanterne (synapserne). Denne proces kaldes "message passing": Atomerne sender deres naboatomer signaler om den information, der findes i dem. Disse oplysninger bruges derefter til at opdatere knudepunkternes oplysninger, og modellen giver derfor mulighed for at inkludere information om den relative position af atomerne i molekylet. Ved at gentage denne operation bliver knudepunkterne mere og mere opmærksomme på, hvad der befinder sig på længere afstand til dem. Grafrepræsentationen transformeres derefter til en vektorrepræsentation og bruges som input i et Deep Neural Network. Denne struktur gør det muligt at bruge back-propagation af fejl til at justere repræsentationen, så den passer til opgaven og de anvendte data. Det gør algoritmen ekstremt fleksibel og i stand til at udtrække information, der er relevant for den aktuelle opgave, uden behov for specifik menneskelig indgriben. Selvom dette har resulteret i state-of-the-art nøjagtighed, er modellerne notorisk kendt for at være black-box, det vil sige, at det ikke er klart, hvilken læring modellerne opnår under deres træning [3].

Integrering af fundamentale principper med Machine Learning

På den ene side har vi de klassiske tilgange, som er transparente, og på den anden side har vi de nye ML-baserede tilgange, som er geometribevidste og i stand til at beskrive ikke-lineære tendenser i egenskaberne. I vores arbejde har vi kombineret de to tilgange for at få det bedste fra de to verdener og gøre modellerne mere udbredte og anvendelige [4]. Til det formål udviklede vi en model, der betragter funktionelle grupper som en mindre graf med atomerne som knudepunkter, og de mindre grafer forbindes derefter for at danne selve molekylet. Denne hierarkiske repræsentation af molekylet fører til state-of-the-art modeller med høj nøjagtighed samt det ekstra aspekt af fortolkelighed. Disse modeller er i stand til at fremhæve og rangere vigtigheden af undergrupperne i molekylet, hvilket i mange tilfælde er i overensstemmelse med viden fra gruppebidragsmodeller (figur 1).

Dette viser den vellykkede integration mellem kemividen og datadrevne ML-tilgange, og at de to kan komplementere hinanden og at det hele er større end summen af dets dele. Værktøjet kan dermed i fremtiden bruges til at accelerere identifikationen af potentielle nye molekyler til forskellige anvendelser, samt at dirigere forskere til kun eksperimentelt at bestemme egenskaberne af de mest lovende kandidater. Dette kan resultere i hurtigere procesudvikling, der kan fremme den grønne omstilling.

E-mail:
Adem Rosenkvist Nielsen Aouichaoui:
arnaou@kt.dtu.dk

Referencer

1. J.-L. Reymond, "The Chemical Space Project," *Acc. Chem. Res.*, vol. 48, no. 3, pp. 722-730, Mar. 2015
2. A.S. Hukkerikar, B. Sarup, A. Ten Kate, J. Abildskov, G. Sin, and R. Gani, "Group-contribution+ (GC+) based estimation of properties of pure components: Improved property estimation and uncertainty analysis," *Fluid Phase Equilibria*, vol. 321, pp. 25-43, May 2012,
3. A.R.N. Aouichaoui, F. Fan, S.S. Mansouri, J. Abildskov, and G. Sin, "Combining Group-Contribution Concept and Graph Neural Networks Toward Interpretable Molecular Property Models," *J. Chem. Inf. Model.*, vol. 63, no. 3, pp. 725-744,
4. A.R.N. Aouichaoui, F. Fan, J. Abildskov, and G. Sin, "Application of interpretable group-embedded graph neural networks for pure compound properties," *Comput. Chem. Eng.*, vol. 176, p. 108291, Aug. 2023