

Onderzoeksopzet en Theorie Afstudeeropdracht

*Artificial intelligence voor het ontwikkelen van chemische producten*

Dennis de Braal | Student nr. 120152177 | HBO Informatica | Inzendcode 264S1 | *06-09-2021*

Inhoud

[1. Inleiding 2](#_Toc96451248)

[1.1 Aanleiding 2](#_Toc96451249)

[1.1.1 Dow 3](#_Toc96451250)

[1.1.2 Industrie 4.0 5](#_Toc96451251)

[1.1.3 Data Analytics & Geavanceerde Algoritmes 7](#_Toc96451252)

[1.1.4 Chemische Industrie en Data Analytics & Geavanceerde Algoritmes 7](#_Toc96451253)

[1.2 Probleemformulering 8](#_Toc96451254)

[1.2.1 Probleemstelling 9](#_Toc96451255)

[1.2.2 Doelstelling 10](#_Toc96451256)

[1.2.3 Onderzoeksvraag 11](#_Toc96451257)

[1.3 Aanpak 12](#_Toc96451258)

[2. Onderzoeksontwerp 13](#_Toc96451259)

[2.1 Onderzoeksmodel 16](#_Toc96451260)

[2.2 Technisch ontwerp 16](#_Toc96451261)

[2.2.1 Dataverzameling 16](#_Toc96451262)

[2.2.2 Data-analyse 18](#_Toc96451263)

[3. Theorie 19](#_Toc96451264)

[3.1 Paradigma’s van de Wetenschap 20](#_Toc96451265)

[3.2 Toepassingen en ML-methoden 22](#_Toc96451266)

[3.2.1 Toepassingen 22](#_Toc96451267)

[3.4 ML vereisten 23](#_Toc96451268)

[3.4.1 Data 23](#_Toc96451269)

[3.4.2 Software 23](#_Toc96451270)

[3.4.3 Hardware 23](#_Toc96451271)

[4. Literatuurlijst 24](#_Toc96451272)

# 1. Inleiding

De inleiding van een afstudeeropdracht bestaat niet uit een algemeen verhaal met daarin bijvoorbeeld een beschrijving van de organisatie, maar is een redelijk uitgebreid hoofdstuk, waarin duidelijk wordt gemaakt waar het onderzoek over gaat. Het bestaat uit de aanleiding, probleemformulering en de aanpak. Deze zaken zijn, als het goed is, al aanbod gekomen bij het plan van aanpak/ onderzoeksvoorstel en kunnen hieruit worden overgenomen. Aandachtspunt daarbij is dat het plan van aanpak vaak spreekt over het onderzoek in de toekomstige tijd, terwijl bij de onderzoeksrapportage uiteraard het onderzoek is uitgevoerd en dus in de verleden tijd wordt beschreven.

## 1.1 Aanleiding

*De aanleiding voor een onderzoek is iets wat de aandacht heeft getrokken van de onderzoeker en voor verbetering vatbaar is. Dit kan bijvoorbeeld een situatie zijn binnen een organisatie dat als probleem wordt gezien. Het onderzoek moet iets bijdragen aan de kennis over het vakgebied (al is het maar over de specifieke casus), maar vooral ook bijdragen aan de oplossing van het praktijkprobleem. Dit kan het best gegarandeerd worden als het probleem ook belangrijk wordt gevonden door anderen, bijvoorbeeld door de directie van een organisatie of binnen een beroepsvereniging.*

*Beschrijf de externe context van het probleem bondig, maar helder. Voor welke maatschappelijke, politieke, economische, technische en/of andere context is het probleem relevant?*

*Ook moet worden aangegeven voor welke doelgroep dit onderzoek bestemd is (de organisatie; de directie; de werknemers; de bedrijfstak, etc.)*

De aanleiding van dit onderzoek komt voort uit de noodzaak om deel uit te maken van de ‘Fourth Industrial Revolution’ en het implementeren van wat het Industrie 4.0 framework genoemd kan worden.

Dow heeft zichzelf als doel gesteld om het meest innovatieve chemische bedrijf ter wereld te zijn. Om dit te kunnen bereiken zal Dow Industrie 4.0 en de daarbij behorende technologieën moeten omarmen.

Als eerste volgt een beknopte beschrijving van Dow, het bedrijf waarbinnen dit onderzoek wordt uitgevoerd, en de doelgroep van dit onderzoek. In de daaropvolgende paragrafen wordt verder ingegaan op het concept Industrie 4.0, hoe het ontstaan is en wat het inhoudt, en de invloed hiervan op de chemische industrie.

Binnen het Industrie 4.0 framework zijn er veel verschillende ondersteunende technologieën. De focus van dit onderzoek ligt op het gebruik van Artificial Intelligence en Machine Learning voor het ontwikkelen van nieuwe producten. Dow heeft recentelijk een eerste versie van een door AI-/ML-technologie ondersteunde applicatie uitgebracht (PU Predictive Intelligence), wat heeft geleid tot interesse vanuit andere business groups en functies in de organisatie.

### Dow

Dit onderzoek is uitgevoerd binnen verschillende afdelingen van Dow. Dow is een chemisch bedrijf dat wereldwijd opereert en vestigingen heeft in zo’n 160 landen. Dow heeft zo’n 50.000 werknemers in dienst, wat het één van de grootste chemische bedrijven ter wereld maakt.

De uitvoering van het onderzoek is gedaan vanuit de Dow-vestiging in Terneuzen. Deze locatie is de grootste productiefaciliteit van Dow buiten de Verenigde Staten. Daarnaast is Terneuzen een zogenaamde hub-locatie, waar de administratieve werkzaamheden voor alle vestigingen in Europa, het Midden-Oosten en India worden uitgevoerd. Op deze locatie zijn ongeveer 3.500 werknemers werkzaam.

Dow is opgedeeld in verschillende business groups, gebaseerd op de product portfolio. Elk van deze business groups produceert en verkoopt een specifieke groep producten:

* **Industrial Intermediates & Infrastructure** (Industrial Solutions and Polyurethanes & Construction Chemicals)
* **Performance Materials & Coatings** (Coatings & Performance Monomers and Consumer Solutions)
* **Packaging & Specialty Plastics** (Olefins, Aromatics & Alternatives and Feedstocks & Energy)

De ambitie van Dow is “om de meest innovatieve, klantgerichte, inclusieve en duurzame ‘materials sciences company’ van de wereld te worden”. Dow wil dit bereiken door het “leveren van waardegroei en best-in-class prestaties”.

Om deze doelstelling te kunnen bereiken, is het noodzakelijk dat Dow meebeweegt met de (r)evolutie die gaande is op het gebied van digitalisering en automatisering (Industrie 4.0). In de volgende paragrafen zal dieper worden ingegaan op het Industrie 4.0-concept.

Daarnaast is het van groot belang om moderne technologieën, zoals AI en ML, te ontwikkelen en te implementeren in de organisatie. De industrie in het algemeen en de concurrentie van Dow in het bijzonder, hebben deze technologie omarmd. Om ervoor te zorgen dat Dow haar huidige positie in de industrie en de markt kan behouden en mogelijk te versterken, is het noodzakelijk de nodige aandacht aan deze onderwerpen te schenken.

Een veelgehoorde klacht, die ook vanuit klantonderzoek naar voren is gekomen, is dat Dow veel tijd nodig heeft om te kunnen reageren op specifieke klantverzoeken. Het gaat hierbij vooral om verzoeken voor producten waarvan de formulering niet (volledig) bekend is.

Om die formulering te creëren moet er extra onderzoek gedaan worden, wat gepaard gaat met laboratoriumproeven. Dit onderzoek is een kostbaar en vooral tijdrovend proces. Tijdens de duur van deze onderzoeken, kan het gebeuren dat een klant uiteindelijk overstapt naar een concurrent. Dit verlies van klanten kan grote gevolgen hebben voor de organisatie. Dow is dan ook naarstig op zoek naar manieren om dit proces te versnellen. Niet alleen op het gebied van chemische ontwikkelingen, maar ook met behulp van moderne IT-technologieën.

De doelgroep van dit onderzoek bestaat uit meerdere afdelingen binnen de organisatie.

De voornaamste doelgroep van dit onderzoek zijn de besluitvormers op het strategisch niveau binnen de verschillende business groups. Het gaat hierbij om de Research & Development-, en Marketing-afdeling van elke business group. De personen op deze posities (Leaders, Managers en Directors) zijn continu bezig met het zoeken naar manieren om ‘slimmer’ en effectiever te kunnen werken en met klanten om te gaan. Iets wat vandaag de dag steeds meer wordt bereikt door het inzetten van innovatieve technologieën.

Met behulp van de uitkomsten van dit onderzoek kunnen zij bepalen:

* Of AI/ML een geschikte oplossing is in hun specifieke omgeving en de specifieke toepassing.
* Welke voorwaarden er gelden en wat er nodig is om AI-/ML-technologie te kunnen inzetten.

Daarnaast is ook de IT-afdeling van Dow (Information Systems/ IS) en in het bijzonder het Digital Operations Center (DOC), een doelgroep van dit onderzoek.

Het DOC is een onderdeel van Information Systems dat zich richt op het digitaliseren van de verschillende productieorganisaties, door het inzetten van ‘advanced manufacturing’ en opkomende technologieën. Het DOC is als het ware de digitale Research & Development-afdeling van de verschillende business groups.

De uitkomsten van dit onderzoek biedt het DOC handvatten voor het implementeren van AI/ML-technologie voor het ontwikkelen van chemische producten.

Deze handvatten kan het DOC gebruiken om in samenwerking met de business groups, specifieke applicaties te ontwikkelen om zo het gebruik van deze technologie op grotere schaal toe te passen.

Daarnaast zijn de uitkomsten mogelijk te gebruiken in andere functies binnen de organisatie (Supply Chain, Logistiek, etc.) om inzicht te krijgen in de mogelijkheden van AI/ML en wat erbij komt kijken om deze technologie te kunnen toepassen. Echter met de kanttekening dat het onderzoek specifiek gericht is op de ontwikkeling van nieuwe producten. Mogelijk is er extra onderzoek nodig om de uitkomsten te kunnen generaliseren naar andere functies.

### 1.1.2 Industrie 4.0

De ‘Fourth Industrial Revolution’, oftewel Industrie 4.0 of 4IR, zal leiden tot een verandering in hoe wij leven en werken. Althans, dat stelt Schwab in zijn artikel over de aanstaande vierde industriële revolutie [Schw].

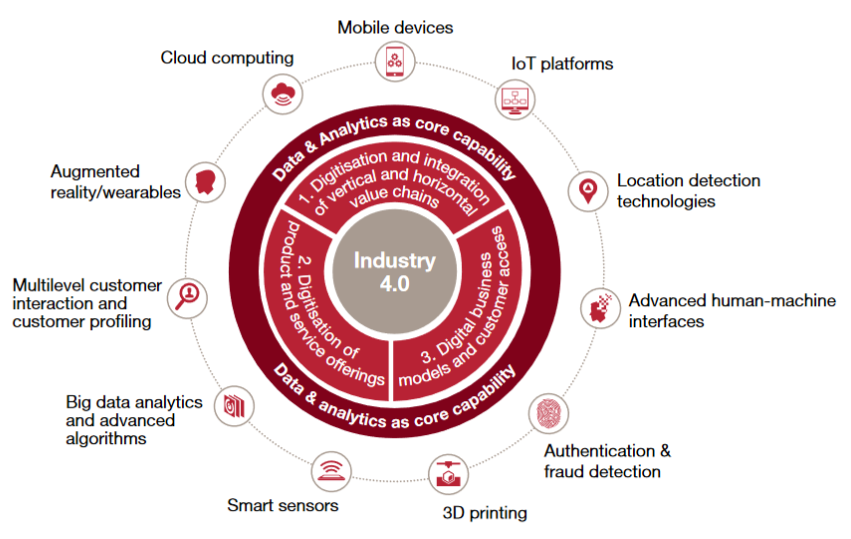
Ondanks dat er wel degelijk verschillen bestaan tussen de termen ‘Fourth Industrial Revolution’ en ‘Industrie 4.0’, worden deze door elkaar gebruikt in dit verslag. De essentie van beide concepten is het digitaliseren van de manufacturing- en automotive-industrie. Voor beide concepten gelden dezelfde technologieën, die als basis dienen voor de invulling van die digitalisering.

Na de First Industrial Revolution (water en stoom om de productieapparatuur aan te drijven), de Second Industrial Revolution (elektriciteit en massaproductie) en de Third Industrial Revolution (elektronica en informatietechnologie om productie te automatiseren), volgt nu dus de Fourth Industrial Revolution.

In deze vierde industriële revolutie wordt er voortgeborduurd op de derde revolutie met het verder digitaliseren van de verschillende industrieën. 4IR onderscheidt zich daarbij door het creëren van een fusie van verschillende technologieën, die de grens tussen de fysieke wereld en de digitale wereld doet vervagen [Schw].

Een ander belangrijk kenmerk van deze vierde revolutie is de snelheid waarmee deze zich ontwikkeld. In tegenstelling tot de voorgaande revoluties die zich op lineaire snelheid ontwikkelden, evolueert de vierde revolutie zich met exponentiële snelheid.

Geissbauer et all. vatten het concept, inclusief de ondersteunende digitale technologieën, samen in het volgende raamwerk [Geis]:



Figuur : Industry 4.0 framework and contributing digital technologies

Het framework weergegeven in figuur 1 is een weergave van de technologieën, kerncapaciteiten en de digitaliserings-onderwerpen die tezamen Industry 4.0 vormen.

Het belang voor een organisatie om mee te kunnen gaan in deze industriële revolutie is groot. Zo stellen Bauernhansl et all. in het door Frauenhofer gepubliceerde rapport over de potentiële economische voordelen van Industrie 4.0, dat de implementatie van Industrie 4.0 kan leiden tot een verlaging van [Bauer]:

* Logistieke kosten (10-30%)
* Productiekosten (10-30%)
* Kwaliteitsmanagement gerelateerde kosten (10-20%)

Het voortbestaan van een organisatie kan in gevaar komen als men deze revolutie aan zich voorbij laat gaan.

### 1.1.3 Data Analytics & Geavanceerde Algoritmes

Uit een enquête uitgevoerd door PWC blijkt dat data analytics aan de basis staat voor Industrie 4.0. Dit gaat niet alleen over de besluitvormingsprocessen en de ondersteuning daarvan door (big) data, maar ook zeker over de ontwikkeling van nieuwe producten [Geis].

De voorbeelden die worden genoemd in de volgende paragraaf zijn gebaseerd op de mogelijkheid om geavanceerde analyses uit te voeren op de data, die in de loop der jaren is verzameld.

De volledige scope van het Industrie 4.0 framework is te breed om als onderwerp van het onderzoek te dienen. In samenspraak met de opdrachtgever is besloten om het onderzoek te richten op het onderwerp ‘Data Analytics en Geavanceerde Algoritmes’.

De keuze voor dit specifieke onderdeel van het framework, komt voort uit het initiatief dat Dow is gestart op dit gebied; AI en ML voor het produceren van PU-foams. Hierover meer in de volgende paragraaf. Dit onderzoek zal zich richten op een bredere toepassing van AI en ML voor het produceren van chemische producten.

### 1.1.4 Chemische Industrie en Data Analytics & Geavanceerde Algoritmes

De chemische industrie omarmt Industrie 4.0. Uit het onderzoek van PWC blijkt dat vooral deze vorm van industrie van plan is, om aanzienlijk te investeren in verdere digitalisering van de operationele processen en de implementatie van de aan Industrie 4.0 ondersteunende technologieën [Geis].

Dit komt ook tot uiting door verschillende initiatieven, die enkele van de grootste chemische bedrijven ter wereld, welke tevens concurrenten van Dow zijn, op dit moment aan het ontwikkelen zijn.

Zo heeft BASF een simulatie-tool ontwikkeld (Ultrasim) die gebruikmaakt van verschillende Industrie 4.0-technologieën om de ontwikkeling van nieuwe producten te kunnen versnellen.

Bayer maakt gebruikt van Machine Learning (data analytics en algoritmes) om de kwaliteit van batches al in een vroeg stadium in het productieproces te kunnen bepalen en op deze manier de efficiëntie van het proces te verhogen en afval te verminderen (QuaRTZ – Quality Release Time Zero).

Ook Dow maakt gebruik van Industrie 4.0-technologieën om het proces voor het creëren van nieuwe producten te verbeteren en te versnellen. Met behulp van Artificial Intelligence en Machine Learning kan de formulering van nieuwe PU-foams bepaald worden, wat moet leiden tot een kortere time-to-market en een reductie van de productiekosten.

## 1.2 Probleemformulering

*De probleemformulering, oftewel de probleemstelling, bestaat globaal uit een doelstelling en een vraagstelling. De doelstelling heeft betrekking op welk resultaat gewenst is als eindresultaat van het onderzoek. Dit betreft een inhoudelijk doel (zoals bijvoorbeeld het oplossen van een probleem binnen een organisatie) en niet een persoonlijk doel (zoals bijvoorbeeld het behalen van een voldoende voor de afstudeeropdracht).*

*De vraagstelling heeft betrekking op de vraag die in het onderzoek beantwoord dient te worden om het doel te kunnen behalen. De vraagstelling bestaat uit een hoofdvraag (ook wel onderzoeksvraag genoemd) die vervolgens opgedeeld wordt in een aantal deelvragen. De deelvragen (meestal tussen de 3 en 5) geven samen het antwoord op de hoofdvraag waardoor het beantwoorden van de deelvragen tijdens het onderzoek automatisch leidt tot het antwoord op de hoofdvraag.*

### 1.2.1 Probleemstelling

Het ontwikkelen van nieuwe chemische producten is een tijdrovend en kostbaar proces. Door middel van trials en testen in het laboratorium wordt gezocht naar de juiste formulering om een product met de gewenste eigenschappen te ontwikkelen. Vanwege de lange duur van het proces, kan niet altijd tijdig aan de vraag van de klant worden voldaan. Het risico bestaat zelfs dat de klant haar business onderbrengt bij een concurrent.

Daarnaast is er de noodzaak om te blijven innoveren op het gebied van Research & Development en de digitalisering daarvan. Concurrenten van Dow zijn inmiddels al ver gevorderd in de digitale transformatie en Dow kan op dat gebied niet achterblijven.

Deze problematiek speelt binnen elke business group van Dow. De marketing-, IT- en R&D-functies van Dow en de business groups zijn actief bezig om een oplossing te vinden voor deze problemen.

De PU-business, onderdeel van de business groep Industrial Intermediates & Infrastructure, heeft een applicatie ontwikkeld die met behulp van AI/ML de formulering van het te ontwikkelen foam-product probeert te voorspellen.

Een klant kan in deze applicatie het type product aangeven met de bijbehorende eigenschappen die de klant wenst, en de applicatie gaat zoeken naar de chemische formulering om dat product te creëren.

Eerst zal de applicatie zoeken in een database met bekende formuleringen om te achterhalen of de formulering voor het gewenste product bekend is. Zo niet, dan zal een ML-algoritme op basis van door de klant aangegeven parameters (o.a. dichtheid van het product, omgevingstemperatuur, en vele andere), proberen een voorspelling te maken van de formulering om het product met die gewenste eigenschappen te kunnen creëren.

Het inzetten van deze toepassing van AI/ML leidt tot het sneller kunnen ontwikkelen van een formulering, met een aanzienlijke verlaging van het aantal experimenten en tests dat uitgevoerd dient te worden.

Eén van de grootste uitdagingen voor het succesvol voorspellen van de formulering, is het verzamelen en converteren van de data op basis waarvan het algoritme de voorspelling kan maken. Deze gegevens zijn veelal verspreid vastgelegd en zijn dan ook moeilijk te achterhalen. Tevens is er geen vaste structuur voor het vastleggen van die gegevens waardoor de conversie, waarbij de gegevens geschikt worden gemaakt om door een algoritme te worden gebruikt, bemoeilijkt wordt.

Daarnaast gaat het om vertrouwelijke gegevens, zogenaamde ‘Crown Jewel Data’. Deze gegevens vormen mede de basis van de concurrentiepositie van Dow en er dient dan ook voorzichtig mee om worden te gaan. Als deze gegevens in handen vallen van de concurrentie kan dit een impact hebben op Dow’s positie in de markt.

Dow heeft een strategie opgesteld (AI Strategy/ MLOps), waarin een route is uitgestippeld om te komen tot een infrastructuur, die het ontwikkelen van de AI/ML capaciteiten binnen Dow kan verbeteren en versnellen. Een ander onderdeel van die strategie gaat over het opzetten van een ‘data science workbench’ waarin geprobeerd wordt een platform te creëren, in een op het gebied van cyber security beveiligde omgeving, waarop data scientists kunnen werken met de crown jewel data.

De hiervoor beschreven toepassing is slecht één van de vele mogelijkheden die AI/ML kan bieden op het gebied van materials science. Daarnaast is de toepassing specifiek voor een bepaalde productsoort en business groep. De toepassing kan niet 1-op-1 worden ingezet binnen andere business groepen.

Het management van het Digital Operations Center en de verschillende business groups zijn op zoek naar verdere toepassingen van AI/ML op het gebied van het ontwikkelen van producten. Zij willen weten welke mogelijkheden AI/ML biedt en wat de benodigdheden zouden zijn om andere AI-/ML-toepassingen te kunnen inzetten in Dow.

Qua benodigdheden gaat het om de hoeveelheid data, die nodig is om een bepaalde AI-/ML-techniek te kunnen gebruiken en te gebruiken dataverzamelings- en dataconversie-methoden.

Daarnaast zal er ook aandacht moeten worden geschonken aan het veiligheidsaspect. Het product wordt gecreëerd op basis van een chemische reactie, waarbij er allerlei gevaren kunnen ontstaan als er in dat proces iets misgaat.

### 1.2.2 Doelstelling

Het doel van dit onderzoek is te achterhalen op welke wijze AI/ML ingezet kan worden, om het ontwikkelen van nieuwe producten in de chemische sector te bespoedigen. Het gaat hierbij om producten, die worden gecreëerd op basis van chemische formuleringen; dat wil zeggen een mengsel van verschillende chemische ingrediënten om zo een product met bepaalde eigenschappen te creëren. Het gaat vooral om producten voor verdere verwerking in andere producten. Hierbij te denken aan coatings voor o.a. verfproducten, plastics voor bijvoorbeeld de automotive-industrie, foams en nog vele andere producten en toepassingen.

Het gebruik van AI-/ML-technologieën voor het ontwikkelen en creëren van nieuwe producten zou moeten leiden tot de volgende voordelen:

* Het sneller kunnen ontwikkelen van nieuwe producten en sneller kunnen inspelen op de vraag van de klant.
* Het kunnen ontwikkelen van een grotere verscheidenheid aan producten.
* Mogelijkheid tot het simuleren van productieprocessen.

Dit onderzoek dient een overzicht op te leveren van AI-/ML-technieken en -methoden die binnen Dow zouden kunnen worden ingezet, en de daarbij behorende hardware-, software- en data-vereisten. Aan de hand van dit overzicht kan het management een besluit vormen over de manier waarop zij AI/ML willen toepassen binnen hun business groep.

### 1.2.3 Onderzoeksvraag

Om een oplossing te vinden voor de problematiek die geschetst is in de vorige paragraaf, is de volgende onderzoeksvraag opgesteld:

*Op welke wijze kan AI-/ML-technologie een bijdrage leveren aan het versnellen van de processen om nieuwe chemische producten te ontwikkelen, hierbij rekening houdend met de veiligheidsaspecten die van toepassing zijn bij deze processen?*

Deze onderzoeksvraag zal beantwoord worden aan de hand van de volgende deelvragen:

1. *Welke AI-/ML-technieken en -modellen worden er in de chemische sector gebruikt om het proces voor nieuwe chemische producten te ontwikkelen te ondersteunen?*
2. *Welke data-, hardware- en software-vereisten worden er gesteld om die technieken en modellen te kunnen gebruiken?*
3. *Welke veiligheidsrisico’s bestaan er bij het proces voor het creëren van nieuwe chemische producten? Welke relatie bestaat er tussen de geïdentificeerde risico’s en de inzet van AI/ML en hoe kunnen die risico’s worden gereduceerd?*
4. *Wat heeft Dow beschikbaar op het gebied van data, hardware en software, om de in de chemische sector gebruikte AI-/ML-technieken en -modellen, in de eigen organisatie te kunnen inzetten? Welke aanpassingen zijn er nodig in de bestaande IT-architectuur en -infrastructuur om AI/ML te kunnen inzetten binnen Dow, ter ondersteuning van de processen voor de ontwikkeling van nieuwe chemische producten?*

De relatie tussen de deelvragen kan als volgt worden uitgelegd.

Als bekend is voor welke toepassingen AI/ML ingezet kan worden in het ‘materials science’-werkveld, kan worden onderzocht welke specifieke AI-/ML-modellen en -algoritmes hiervoor gebruikt worden. Zodra dit inzichtelijk is gemaakt kan worden bepaald welke eisen er worden gesteld op het gebied van data, hardware en software. Op basis van deze eisen kan vervolgens worden onderzocht in welke mate de binnen Dow bestaande infrastructuren en architecturen m.b.t. data, hardware en software, toereikend zijn en of er eventueel aanpassingen gemaakt dienen te worden. Hierbij speelt het kunnen reduceren van eventuele risico’s op het gebied van de chemische processen een grote rol. Er dient immers gewaarborgd te worden dat de inzet van AI/ML niet leidt tot het vergroten van die risico’s.

## 1.3 Aanpak

*Tot slot van het inleidende hoofdstuk wordt de lezer een leeswijzer gegeven, waarin kort de opzet van de rest van de rapportage beschreven wordt.*

# 2. Onderzoeksontwerp

*Bij het onderzoeksontwerp wordt nader uitgewerkt hoe het onderzoek vormgegeven wordt. Dit kan worden opgedeeld in een inhoudelijk en een technisch deel.*

Dit hoofdstuk bevat de beschrijving van de opzet van het onderzoek en de motivatie daarvoor. Er wordt eerst ingegaan op het type onderzoek dat uitgevoerd is en waarom. In de paragrafen daarna volgt het onderzoeksmodel en de details van de onderzoeksmethoden.

Binnen de basisprincipes van onderzoek, de zogenaamde methodologie, kan er onderscheid gemaakt worden tussen twee typen onderzoek:

* Fundamenteel onderzoek
* Praktijkgericht onderzoek

Het belangrijkste verschil tussen beide typen onderzoek zit in het soort probleem dat met het onderzoek wordt geprobeerd op te lossen. Fundamenteel onderzoek gaat meestal om het ontwikkelen van kennis. Praktijkgericht onderzoek is gericht op het vinden van oplossingen voor praktijkproblemen (Verhoeven, 2018) [Verhoev].

Uit de probleemstelling van dit onderzoek kan worden opgemaakt dat dit onderzoek een praktijkgericht onderzoek betreft.

Daarnaast kan er een onderscheid worden gemaakt tussen kwantitatief en kwalitatief onderzoek. Kwantitatief onderzoek maakt gebruik van cijfermatige gegevens. Kenmerken over objecten, organisaties en personen worden met behulp van statistische technieken verwerkt (Verhoeven, 2018) [Verhoev]. Vaak staat bij type onderzoek centraal dat het onderzoek is te generaliseren naar een grote populatie (Saunders et al. 2019) [Sau].

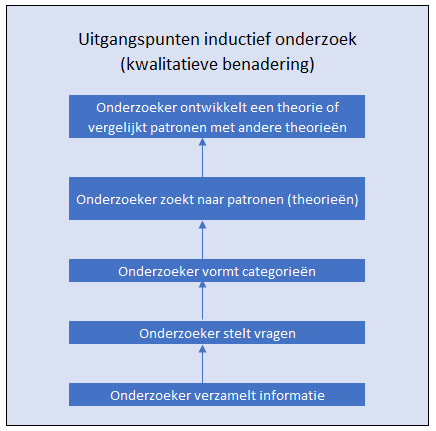
Kwalitatief onderzoek daarentegen wordt uitgevoerd met niet-cijfermatige gegevens. De gegevens worden verwerkt in alledaagse taal. Bij kwalitatief onderzoek gaat het vaak om inzicht te krijgen in een probleem. De methoden voor het verzamelen van kwalitatieve data zijn flexibeler en zijn gericht op het interpreteren van die data (Verhoeven, 2018) [Verhoev].

Het doel van dit onderzoek is het verkrijgen van inzicht in de manier waarop machine learning ingezet kan worden in het materials science werkveld. Hierbij speelt de context en de diepte van het onderzoek een grote rol. De kwalitatieve aanpak sluit hier het beste op aan.

De keuze voor een kwalitatieve aanpak komt ook voort uit de aard van het onderzoek. Het doel van het onderzoek is exploratief van aard. Er dient kennis gegenereerd te worden over de toepassing en inzet van machine learning in het materials science werkveld.

Vanwege deze inslag is dan ook gekozen voor een inductieve aanpak. Bij de inductieve methode wordt een theorie of model ontwikkelt op basis van het verzamelen van data. Dit in tegenstelling tot de inductieve methode waarbij een theorie of model het uitgangspunt van het onderzoek is.

Figuur 2 is een weergave van de inductieve methode zoals beschreven door Saunders et. al (2019) [Sau].



Figuur 2 - Uitgangspunten kwalitatief onderzoek. Overgenomen uit Methoden en technieken van onderzoek, Saunders, M. N. K., Lewis, P., Thornhill, A., Arnoldy, M., & Smitt, P. (2019). Methoden en technieken van onderzoek. Pearson

Naar aanleiding van de uitkomsten van de kritische literatuurstudie die is uitgevoerd in fase drie, is een model gecreëerd dat de basis heeft gevormd voor de verdere invulling en uitvoering van het onderzoek

De uitvoering van dit praktijkgerichte, kwalitatieve onderzoek is gebaseerd op het onderzoeksproces zoals beschreven door Saunders et al. (2019) [Sau] en is weergegeven in figuur 3:



Figuur 3 – Onderzoeksproces. Overgenomen uit Methoden en technieken van onderzoek, Saunders, M. N. K., Lewis, P., Thornhill, A., Arnoldy, M., & Smitt, P. (2019). Methoden en technieken van onderzoek. Pearson

De eerste twee fasen, Oriëntatie op probleem en contex,t en Formuleren onderzoeksdoelstelling en vraagstelling, zijn behandeld in het plan van aanpak. Dit hoofdstuk gaat in op de methodologische verantwoording en geeft een beschrijving van de aanpak van het onderzoek en de motivatie voor die aanpak.

De uitkomsten van fase drie, kritische literatuurstudie, zijn beschreven in hoofdstuk 3 en vormen het theoretisch kader van het onderzoek.

In paragraaf 2.1 wordt op basis van het onderzoeksmodel een globaal overzicht gegeven van de verschillende onderdelen van het onderzoek. Deze worden vervolgens uitgewerkt in paragraaf 2.2, waarbij verder wordt ingegaan op de details van de gebruikte onderzoeksmethoden.

## 2.1 Onderzoeksmodel

*In het onderzoeksmodel wordt beschreven hoe het onderzoek is opgezet. Oftewel: de manier waarop antwoorden op de deelvragen gezocht gaan worden. Hierbij wordt aangegeven welke theorie input levert, welke informatie op welke manier verzameld gaat worden, etc.*

## 2.2 Technisch ontwerp

*Vervolgens wordt dit verder uitgewerkt. Er wordt meer uitgebreid beschreven welke onderzoekstechnieken gebruikt zijn en hoe dit aangepakt is. Zijn bijvoorbeeld open of half gestructureerde interviews gehouden en op welke wijze is de selectie van de geïnterviewden tot stand gekomen?*

In deze paragraaf is beschreven welke methoden en technieken zullen worden gebruikt om de data te verzamelen en te analyseren. Er wordt eerst ingegaan op de methoden voor dataverzameling, en vervolgens komen de methoden aan bod om de verzamelde gegevens te analyseren.

### 2.2.1 Dataverzameling

De hoofdvraag is opgedeeld in een viertal deelvragen. Het antwoord op deelvragen 1 en 2 is gezocht door middel van een literatuuronderzoek. Voor beide deelvragen is een literatuuronderzoek uitgevoerd, waarbij de dataverzameling en analyse, in samenhang met elkaar zijn uitgevoerd. Deelvragen 3 en 4 zullen worden beantwoord door het houden van interviews.

#### 2.2.1.1 Literatuuronderzoek

Het doel van het literatuuronderzoek is het verschaffen van kennis en inzichten met betrekking tot het gebruik van machine learning in het materials science werkveld. De resultaten hebben geleid tot de structuur van het onderzoek en het theoretisch kader dat als basis dient voor het verdere vervolg van het onderzoek. Daarnaast heeft het literatuuronderzoek een antwoord gegeven op deelvragen 1 en 2.

De geraadpleegde literatuur is afkomstig van verschillende internet-bronnen. Er is gezocht naar wetenschappelijke artikelen via de zoekmachines Google Scholar en Arxiv. De uitwerking en resultaten van de zoektocht zijn weergegeven in ***bijlage x – Werkblad Deelvraag 1.***

De uitvoering van het literatuuronderzoek is gebaseerd op de methode Systematisch Literatuuronderzoek. Deze methode geeft een duidelijke structuur aan het onderzoek en verhoogt daarmee de betrouwbaarheid en herhaalbaarheid van het onderzoek. Systematisch literatuuronderzoek bestaat uit de volgende stappen:

* Probleemdefinitie;
* Zoekstrategie bepalen;
* Literatuur verzamelen;
* Waarderen en evalueren van de bronnen;
* Synthese.

Voor het doorlopen van de verschillende stappen van deze methode, is gebruikgemaakt van de checklist ‘Systematisch Literatuuronderzoek Protocol’ (Weber, 2011) [Weber]. ***De ingevulde checklists zijn als bijlage aan dit rapport bijgevoegd.***

Het gebruik van de checklist, waar de aanpak en uitvoering van het literatuuronderzoek in zijn beschreven, maakt het onderzoek herhaalbaar wat ten goede komt aan de betrouwbaarheid ervan.

Niet alle onderdelen van de checklist zijn uitgewerkt in het document zelf. Het onderzoeksgebied en beschrijving van het probleem zijn aan bod gekomen in hoofdstuk 1 en 2 van dit rapport.

De resultaten van de dataverzameling zijn vastgelegd in Excel wat verdere verwerking mogelijk maakt. ***De Excel werkbladen zijn als bijlage bijgevoegd aan dit rapport.***

Voor de verdere details over de uitvoering van de dataverzameling voor het literatuuronderzoek, wordt verwezen naar de beide checklists. Deze checklists bevatten de volledige details van de opzet van de literatuurstudie. Zo wordt er onder andere vermeld welke criteria zijn gehanteerd voor het selecteren van bronnen, alsmede de inhoudelijke selectie- en exclusiecriteria.

*Sneeuwbal methode; analyse referenties van de gevonden 31 artikelen*

De zoektocht heeft 156 artikelen opgeleverd waarvan de titel en samenvatting voldoen aan de selectiecriteria zoals gedefinieerd in paragraaf 7.1 van het ***document Checklist Systematisch Literatuuronderzoek Deelvraag 1***.

Deze artikelen zijn onderworpen aan een verdere analyse waarvan de resultaten zijn beschreven in paragraaf 2.2.2.

#### 2.2.1.2 Interviews

### 2.2.2 Data-analyse

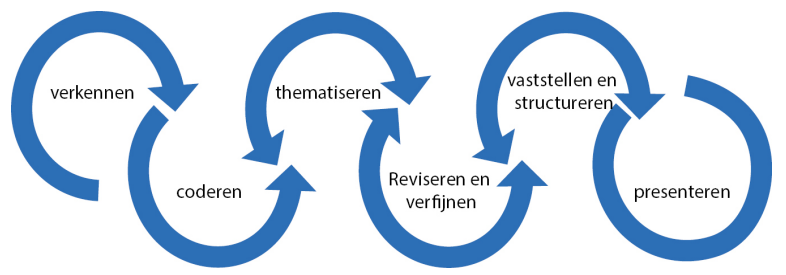
Deze paragraaf bevat een beschrijving van de data-analysemethoden die zijn gebruikt om de verzamelde data te ordenen en te analyseren.

#### 2.2.2.1 Analyse Literatuuronderzoek

Om de resultaten van de dataverzameling van het literatuuronderzoek op een ordelijke wijze te kunnen analyseren, is gebruikgemaakt van de methode ‘***Thematische Analyse’.***

Thematische analyse is een hulpmiddel om kwalitatieve data te identificeren, te analyseren en rapporteren. Met behulp van dit hulpmiddel kunnen gegevens worden geordend en geïnterpreteerd in het licht van de probleemstelling (Verhoeven, 2019) [Verhoev].

De methode bestaat uit een aantal stappen die zijn weergegeven in figuur 4.



Figuur 4 - Thematische analyse in zes stappen. Overgenomen uit Wat is onderzoek? Verhoeven, N. (2018). (6e editie). Boom Lemma.

De uitwerking van de analyse, met daarbij de gebruikte codering en structurering van de literatuur, is vastgelegd in bijlage xxx (checklist 1). Deze gegevens bieden ondersteuning voor het herhalen van het onderzoek, wat de betrouwbaarheid ervan ten goede komt.

Het iteratieve karakter van deze methode, heeft een positieve invloed op de kwaliteit van het onderzoek. De kennis die wordt opgedaan tijdens de uitvoering van een iteratie, kan worden toegepast in een daaropvolgende iteratie.

Voor de resultaten omtrent de analyse van de literatuur, wordt verwezen naar hoofdstuk 3 van dit document.

#### 2.2.2.2 Analyse Interviews

# 3. Theorie

*De theorie binnen een praktijkgericht onderzoek biedt het kader om de praktijk te kunnen analyseren en conclusies hieraan te kunnen verbinden. De gebruikte theorieën en ook de beschrijving hiervan staan in dienst van het vinden van een antwoord op de gestelde deelvragen en daarmee dus op de hoofdvraag.*

*Het is daarom ook de bedoeling om de literatuur niet als een opsomming van verschillende theorieën te presenteren, maar gesorteerd naar bijvoorbeeld deelvragen waarbij niet de complete theorieën worden besproken, maar slechts de elementen die centraal staan bij het beantwoorden van de vragen.*

## 3.1 Materials Science

Materialen zijn overal rondom ons en hebben een grote invloed op hoe de mens haar leven leidt, en de maatschappij zich als geheel ontwikkelt. Vanuit historisch perspectief is de vooruitgang van de mensheid sterk verbonden me haar vermogen om materialen, te produceren en aan te passen. Deze relatie is zelfs zo sterk aanwezig, dat de eerste beschavingen worden genoemd naar de materiaalsoort die kenmerkend was voor die tijd; stenen tijdperk, bronzen tijdperk en ijzeren tijdperk.

In de hedendaagse tijd zijn de verschillende drankverpakkingen voor frisdranken, een mooi voorbeeld om de diversiteit aan materialen te illustreren. Zo zijn frisdranken te verkrijgen in blikjes (metaal), glazen flessen (keramiek) en plastic flessen (polymeren). Elk van deze verpakkingen heeft een set onderscheidende eigenschappen, wat de gebruikte materiaalsoort uitermate geschikt maakt voor het bewaren, en goedhouden van frisdrank.

Materials science is een samenkomst van verschillende wetenschappen en technieken. Het is een multidisciplinair werkveld dat elementen bevat van o.a. metaalbewerking, keramische techniek, scheikunde en natuurkunde [Shack]. Het doel van de materials science discipline is het ontdekken van nieuwe van nieuwe materialen, en het verbeteren van bestaande materialen. Men tracht dit doel te bereiken door het onderzoeken en leren begrijpen van de structuur en samenstelling van materialen. [Ask]. Materials science bestaat uit een aantal componenten die zijn weergegeven in figuur 5.



Figuur 5 - Componenten van Materials Science. Overgenomen uit Jr., W. C. D., & Rethwisch, D. G. (2013). Materials Science and Engineering: An Introduction (9de editie). Wiley.

Deze componenten kunnen als volgt worden uitgelegd. [Cal] [Ask]

Processing oftewel verwerking slaat op de handelingen die nodig zijn om materialen te vervaardigen uit chemicaliën (middels een proces dat synthese wordt genoemd), en het fabriceren van materialen tot bruikbare objecten.

De structuur van een materiaal omvat de ordering van de interne componenten, en wordt onderverdeeld in 3 lagen: [Cal]

* **(Sub-)Atomische structuur**: de wijze waarop de atomen en moleculen zijn georganiseerd ten opzichte van elkaar. Sub-atomische structuur gaat over de interactie tussen een enkele atoom en de nuclei.
* **Microstructuur**: betreft de structuur van een materiaal dat kan worden waargenomen door middel van microscopische vergroting. De microstructuur bestaat uit een grote groep opeengestapelde atomen.
* **Macrostructuur**: deze laag gaat om de structurele elementen die met het blote oog kunnen worden waargenomen.

De structuur en de wijze van verwerking hebben een hele grote invloed op de eigenschappen van het materiaal. Door middel van het aanpassen van de structuur of het verwerkingsproces, kan er invloed uitgeoefend worden op de uiteindelijke eigenschappen van het materiaal.

De belangrijkste eigenschappen van een vast materiaal kunnen worden ingedeeld naar de volgende categorieën:

* Mechanische eigenschappen.
* Elektrische eigenschappen.
* Thermische eigenschappen.
* Magnetische eigenschappen.
* Optische eigenschappen.
* Chemische eigenschappen.

De prestaties van een materiaal (performance) kunnen worden gedefinieerd als de functie van de verschillende eigenschappen die het materiaal bezit. Eigenschappen en prestaties zijn sterk aan elkaar verbonden. Het zijn immers de eigenschappen van een materiaal, die de prestaties ervan bepalen.

Een andere belangrijke indeling die in materials science wordt gemaakt, is die van het soort materiaal. Voor vaste materialen wordt uitgegaan van vijf soorten. Deze indeling is gemaakt op basis van de chemische samenstelling en atomische structuur van de verschillende materialen binnen elke soort. [Ask] [Cal]

* **Metalen**; materialen die bestaan uit één of meer metalen elementen (bijvoorbeeld ijzer, koper, aluminium, goud, etc.), en bevat vaak een toevoeging van een kleine hoeveelheid niet-metalen elementen. Deze soort materiaal is terug te vinden in objecten als zilverwaren, munten, sieraden, enz.
* **Keramieken**; een chemische verbinding (compound) tussen metaal- en niet-metaal-elementen. Enkele bekende voorbeelden van keramische materialen zijn onder andere glas, porselein en baksteen.
* **Polymeren**; De materialen in deze soort zijn overwegend organische materialen, die worden geproduceerd door een proces dat polymerisatie wordt genoemd. Polymerisatie voegt enkelvoudige moleculen (monomeren) samen tot een keten van zichzelf herhalende moleculen (polymeren). Polymeren zijn terug te vinden in onder andere rubber, allerlei soorten plastics en vele andere toepassingen, van kogelvrije vesten, Liquid Crystal Displays’s (LCD’s) tot kleren en koffiemokken.
* **Composieten**; een composiet bestaat uit een samenvoeging van twee of meer individuele materialen uit de hiervoor genoemde categorieën. Het idee achter het ontwikkelen van een composiet, is het mixen van de eigenschappen van de individuele materialen. De bekendste toepassingen van composiet zijn glasvezel en met koolstofvezel (carbon fiber) versterkte polymeren die worden gebruikt bij het bouwen van vliegtuigen.
* **Geavanceerde materialen**; deze soort bestaat uit onder andere halfgeleiders, biomaterialen en nano-materialen. Deze materialen worden gebruikt in high-tech toepassingen zoals computers, ruimtevaart en luchtvaart. In veel gevallen worden de geavanceerde materialen ingezet om de eigenschappen van bestaande materialen en producten sterk te verbeteren.

## 3.2 Paradigma’s van de Wetenschap

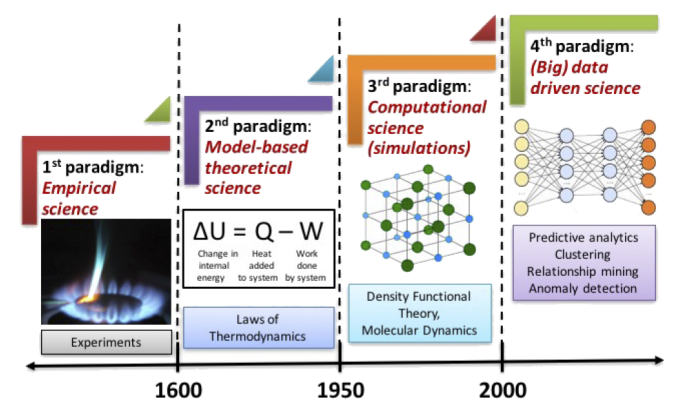
Het materials science-werkveld is in de loop der jaren geëvolueerd tot een wetenschap die wordt gedreven door data. Lange tijd was deze wetenschap puur empirisch, gebaseerd op experimenteren en natuurlijke verschijnselen, en overeenkomend met het stenen en metalen tijdperk (brons, ijzer en staal). In de eeuwen daarna kwam het volgende paradigma op, gebaseerd op de wetten van de fysica en theoretische modellen en generalisatie. Dit tweede paradigma was tot de jaren ’50 van de twintigste eeuw de voornaamste vorm in de wetenschap.

Door de opkomst van computers werd het derde paradigma, ‘computational science’, populair. Door de toegenomen rekenkracht werd het mogelijk om zeer complexe modellen te analyseren en op te lossen. In het veld van computationele scheikunde, zijn vooral de methoden ‘density functional theory (DFT)’ en ‘molecular dynamics (MD)’ zeer bekend. [Agra] [Das] [Schled]

Het toepassen van deze paradigma’s in de verschillende takken van de wetenschap, theorieën, experimenten en berekeningen, hebben geleid tot het genereren van een gigantische hoeveelheid data. Deze hoeveelheid data heeft de wetenschap verder doen evolueren naar het vierde paradigma; ‘(big) data driven science’.

Het kunnen verwerken en analyseren van de gigantische hoeveelheid data wordt mede mogelijk gemaakt door de snelle ontwikkelingen op het gebied van de informatica (opslag en rekenkracht). [Agra] [Das] [Schled]

Figuur 5 geeft een overzicht van de evolutie van de wetenschap en de vier genoemde paradigma’s.



Figuur 6 - De 4 Paradigma's van de wetenschap; Empirisch, Theoretisch, Computationeel en Data-gedreven. Overgenomen uit Perspective: Materials informatics and big data: Realization of the “fourth paradigm” of science in materials science door Agrawal, A., & Choudhary, A. (2016). APL Materials, 4(5), 053208. https://doi.org/10.1063/1.4946894

De hoeveelheid data die voortkomt uit het vormen van theorieën, uitvoeren van experimenten en simulaties, kan worden getypeerd als ‘big data’.

Big data kan worden beschreven aan de hand van de 5V’s van big data [Schled] [Agra]:

* Volume; de hoeveelheid data, de grootte van de dataset.
* Velocity; de snelheid waarmee de data gegenereerd wordt.
* Variety; de verschillende datatypen (zowel gestructureerd als ongestructureerd).
* Veracity; de accuraatheid en betrouwbaarheid van de data.
* Value; het vermogen om bruikbare informatie te kunnen afleiden uit de data.

Eén van de grootste uitdagingen met betrekking tot big data is het kunnen analyseren van die data. De opkomst van artificial intelligence en machine learning hebben een grote invloed in de mate waarop de data succesvol kan worden geanalyseerd, en er bruikbare informatie uit te kunnen achterhalen. [Kit]

In de volgende paragrafen wordt een algemene introductie gegeven over machine learning en de toepassingen ervan binnen het materials science werkveld.

## 3.3 Machine Learning in Materials Science

Het proces voor het ontwikkelen van nieuwe chemische producten, het onderwerp van deelvraag 1, bevat een breed scala aan handelingen en activiteiten. Alle vier de componenten waaruit materials science uit bestaat (zie beschrijving in paragraaf 3.1), spelen een rol in de ontwikkeling van nieuwe producten. Elk van deze componenten bestaat uit een groot aantal activiteiten en processen, die elk een aandeel hebben in de ontwikkeling van een product.

Om een idee te geven over de enorme hoeveelheid mogelijkheden, die er zijn om AI/ML in te zetten voor elk van de componenten van materials science, volgen enkele voorbeelden voor het gebruik van AI/ML in het synthese proces, wat een onderdeel is van het verwerkingsproces. De genoemde onderzoeken zijn slechts enkele voorbeelden. Er zijn nog vele andere onderzoeken alleen al voor dit ene specifieke sub-proces van materials science.

Materialen worden gevormd door het samenvoegen van verschillende elementen (op molecuul-niveau). De samenstelling van de verbonden elementen, vormen de structuur van het materiaal. De elementen worden verbonden door een proces dat synthese wordt genoemd. In de synthese reageren atomen op elkaar, en worden zo verbonden of verbroken. De chemische reactie is afhankelijk van een groot aantal factoren, waaronder de atomische structuur van een molecuul, type en hoeveelheid katalysator, het type reactie en de reactiesnelheid [Poul] [Ask].

Voor veel van deze factoren wordt actief onderzocht hoe AI/ML een bijdrage kan leveren. De Almeida et al. (2019) hebben onderzoek gedaan naar het gebruik van ML om de syntheseweg te ondersteunen [Alm]. Segler et al. (2018) gebruiken in hun onderzoek ML juist om de omgekeerde syntheseweg te vinden.

Een andere benadering is gekozen door Raccuglia et al. (2016), die ML hebben ingezet om te bepalen of een reactie al dan niet succesvol zal zijn [Racc]. Ook voor het zoeken naar de elementen die het beste gebruikt kunnen worden, kan ML worden ingezet, zoals aangetoond door Yuan et al. (2018) [Yuan]. Of het gebruik van ML om te voorspellen welke hoeveelheid product of materiaal een reactie zal opleveren, zoals is onderzocht door Żurański et al. (2021) [Zur].

Een grote uitdaging in het bestuderen van de literatuur over dit onderwerp, is dan ook het verzamelen van publicaties, die een breed perspectief bieden op het gebruik van AI/ML in het materials science domein. En niet in gaan op slechts een klein onderdeel daarvan.

In de volgende paragraaf wordt een overzicht gegeven van verschillende toepassingen van AI/ML in het materials science. Hierbij is een indeling gehanteerd die eenvoudig is terug te herleiden naar de componenten waaruit materials science bestaat. Daarnaast wordt ingegaan op de motivatie achter de samenstelling van het overzicht.

### 3.3.1 ML-Toepassingen en Technieken

Uit een analyse van 175 publicaties over het gebruik en de ontwikkeling van AI/ML in materials science, zijn zo’n 35 publicaties overgebleven die verwijzen naar een serie onderzoeken, over verschillende toepassingen van AI/ML binnen dit domein.

### 3.3.2 Indeling Toepassingen

*Tabel xx* geeft een overzicht van de ML-technieken en toepassingen, ingedeeld naar vijf categorieën. Deze categorieën bestaan uit de vier componenten van de materials science discipline, aangevuld met de categorie ‘methoden en technieken’.

Voor een beschrijving van de vier componenten (verwerking, structuur, eigenschappen en prestaties) wordt verwezen naar paragraaf 3.1.

De categorie ‘methoden en technieken’ bestaat uit toepassingen op het gebied van computationele scheikunde. Dit is een onderdeel van de chemie, waarbij gebruik wordt gemaakt van computers om theoretische scheikunde te ondersteunen en verbeteren. [Jens]

Theoretische scheikunde is het gebruikmaken van wiskundige methoden, gecombineerd met de wetten van de fysica, om chemische problemen en fenomenen te bestuderen. [Jens] [Cram]

Bekende voorbeelden van computationele scheikunde zijn onder andere Density Functional Theory (DFT) [Bur] en Molecular Dynamics (MD) [Att 1-28].

Een diepgaande beschrijving van computationele scheikunde valt buiten de scope van dit document. Geïnteresseerde lezers wordt verwezen naar het werk van Jensen (2017) en Cramer en Cramer (2013). [Jens] [Cram]

Door deze indeling kan op duidelijke wijze de relatie worden gelegd naar de verschillende componenten waaruit materials science bestaat. De toevoeging van methoden en technieken komt voort uit de opkomst van het gebruik van AI/ML, dat wordt ingezet ter ondersteuning en verbetering van de computationele methoden en technieken.

## 3.4 Machine Learning Workflow

Om te kunnen bepalen wat er nodig is om de in de vorige paragraaf gepresenteerde ML-technieken en -modellen te kunnen inzetten (zie deelvraag 2 van dit onderzoek), is het van belang om inzicht te krijgen hoe deze technieken en modellen zijn geïmplementeerd.

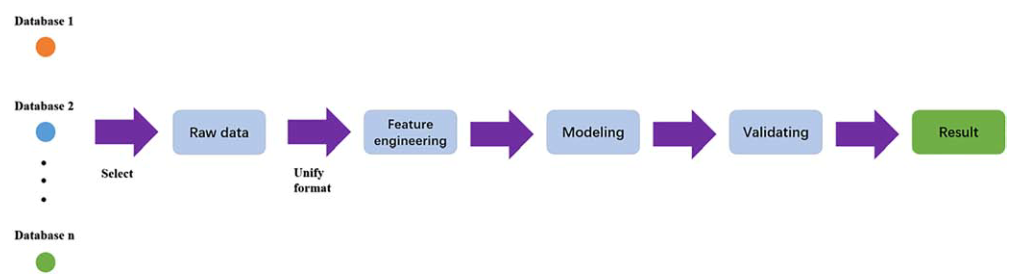
In de volgende paragrafen wordt eerst ingegaan op het proces voor het ontwikkelen, trainen en valideren van een ML-model (de zogenaamde ML-workflow). Uit deze beschrijving kan in algemene termen worden opgemaakt, wat benodigd is om een ML-model in de praktijk te kunnen inzetten.

Op basis van deze informatie is bepaald welke gegevens er verzameld moesten worden, om een antwoord te kunnen geven op deelvraag 2. De resultaten van deze dataverzameling worden gepresenteerd in paragraaf 3.x.x.

Het proces voor het ontwikkelen van een ML-model voor het materials science-domein, kan grofweg worden opgedeeld in drie stappen [Das]:

* Dataverzameling en -representatie.
* Selecteren en valideren van het (ML-)model.
* Optimaliseren van het (ML-)model.

Dit proces kan worden weergegeven zoals de workflow gegeven in figuur 6 [Wei].



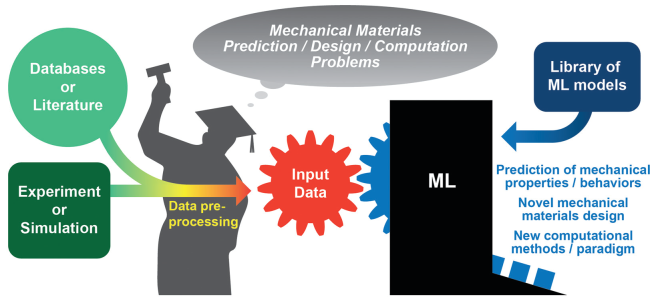
Figuur 6 - Machine Learning Workflow. Overgenomen uit Machine learning in materials science door Wei, J., Chu, X., Sun, X., Xu, K., Deng, H., Chen, J., Wei, Z., & Lei, M. (2019). InfoMat, 1(3), 338–358. https://doi.org/10.1002/inf2.12028

In deze workflow is de stap dataverzameling- en representatie, verder gedetailleerd in de stappen ‘raw data’ en ‘feature engineering’.

De stappen ‘modeling’ en ‘validating’ staan voor het selecteren en ontwikkelen van het model, en de evaluatie en optimalisatie daarvan [Wei].

Figuur 7 is een vergelijkbare weergave van dit proces maar dan specifieker ingaand op het materials science domein. Hierbij kan worden opgemerkt dat de drie voornaamste componenten van dit proces bestaan uit [Guo]:

* Een goed georganiseerde dataset met materiaal-data, achterhaald uit literatuur en bestaande databases, of verkregen door het uitvoeren van experimenten en simulaties.
* Een ML-model dat in staat is om de data-representatie van materialen en moleculen te kunnen ontleden.
* Een duidelijk gedefinieerd probleem dat door het ML-model dient te worden opgelost.



Figuur 7 - Schematische weergave van de ML-workflow in Materials Science. Overgenomen uit Artificial intelligence and machine learning in design of mechanical materials door Guo, K., Yang, Z., Yu, C. H., & Buehler, M. J. (2021). Materials Horizons, 8(4), 1153–1172. https://doi.org/10.1039/d0mh01451f

In de volgende paragrafen wordt verder ingegaan op de verschillende stappen in dit proces.

### 3.4.1 Data Verzameling

De eerste stap in het proces is het verzamelen van data. Vanuit een machine learning-oogpunt kan de data ingedeeld worden in twee categorieën:

* Training data; te gebruiken om het model te trainen/laten leren.
* Test data; te gebruiken om het getrainde model te testen met data van buiten de training set.

De benodigde training data is afhankelijk van de manier waarop het ML-model getraind gaat worden. Voor ‘supervised learning’ bestaat de training data uit een set input-waarden en bijbehorende output-waarden (ook wel gelabelde data genoemd).   
Als de training data alleen input-waarden bevat, kan ‘unsupervised learning’ gebruikt worden om te proberen bepaalde patronen te herkennen in de data set of de data te clusteren in categorieën.

Als er de beschikking is over een dataset bestaande uit input-waarden en slecht een gelimiteerde hoeveelheid output-waarden, kan er gebruikt gemaakt worden van ‘semi-supervised learning’.

Op het moment is ‘supervised learning’ de meest volwassen methode om een ML-model te trainen. [But]

Op het gebied van materials science kan de data worden onderverdeeld in vier categorieën [Xi] [Wei]:

* Materiaaleigenschappen (fysieke eigenschappen of gegevens over de structuur van het materiaal).
* Data over de chemische reactie van grondstoffen.
* Grafische data of afbeeldingen (bijvoorbeeld een afbeelding van de oppervlakte van een materiaal.
* Data uit literatuur (tekstuele data).

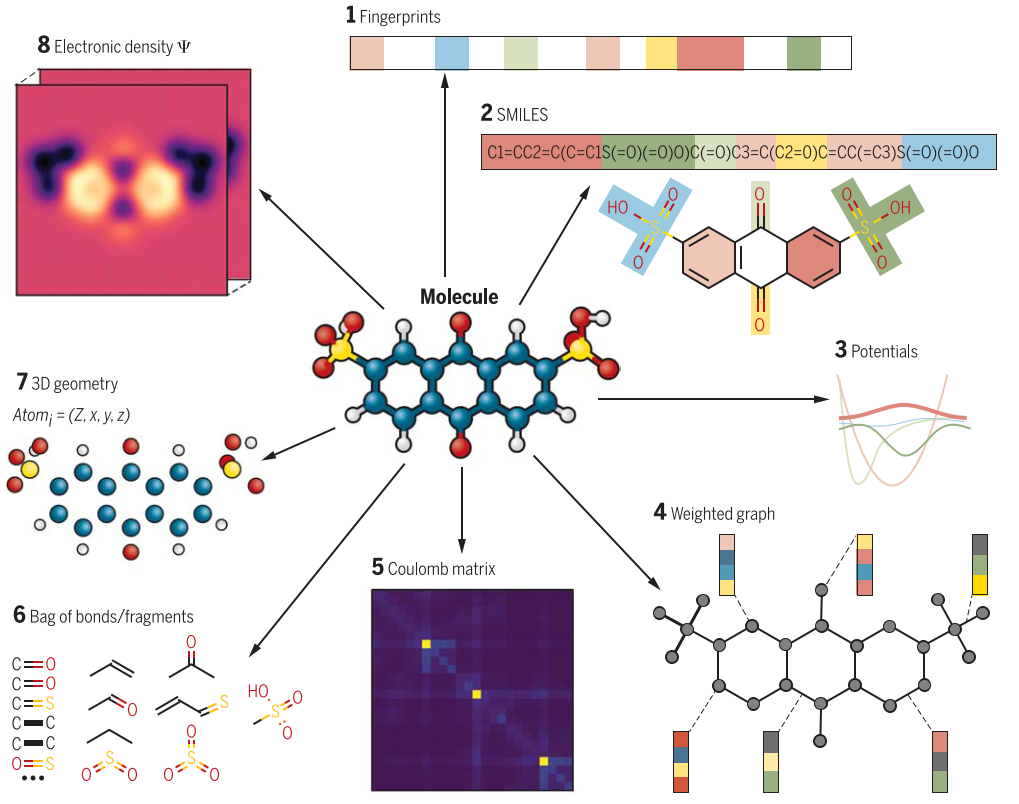
Data kan worden verkregen uit een aantal bronnen [Sha] [Guo] [Das] [Wei]:

* Data voortgekomen uit experimenten en simulaties (zoals DFT en MD).
* Online-databases met gegevens over materialen (zoals AFLOW, MATDAT, Material Project, en anderen).
* Literatuur; met behulp van text-mining technieken kunnen grote aantallen tekstuele bronnen worden doorzocht op bruikbare gegevens.

### 3.4.2 Data Representatie

Deze stap in het proces wordt ook wel ‘Feature Engineering’ genoemd. Feature Engineering heeft als doel om kenmerken (features) te onttrekken uit de ruwe data, om het ML-algoritme te kunnen laten werken. [Wei]

Zie figuur 8 voor de veelgebruikte weergaven van een molecuul, die door een ML-algoritme gebruikt kunnen worden. [Wei]



Figuur 8 - Weergave van een molecuul. Overgenomen uit Machine learning in materials science door Wei, J., Chu, X., Sun, X., Xu, K., Deng, H., Chen, J., Wei, Z., & Lei, M. (2019). InfoMat, 1(3), 338–358. <https://doi.org/10.1002/inf2.12028>. Copyright 2018, The American Association for the Advancement of Science

Voor de traditionele ML-methoden is het vaak nodig om de kenmerken manueel te selecteren en te verwerken. Dit is echter een arbeidsintensief proces welke veel tijd kost, en waarvoor een grote mate van domeinkennis vereist is. Het gebruik van deep learning-technieken voor het selecteren van de kenmerken is aan opmars bezig, en zou wel eens de trend kunnen worden op dit gebied. [Wei]

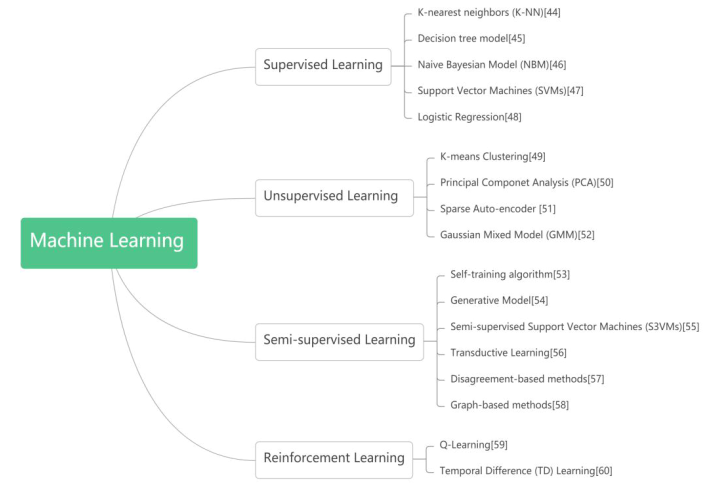
Enkele voorbeelden deep learning-technieken, die veelbelovende prestaties laten zien vergelijkbaar met of zelfs beter dan manuele methoden, zijn MEGNet, SchNet, en CGCNN. [Mor]

### 3.4.3 Model Building

Zodra de data is verzameld en verwerkt tot input voor een ML-algoritme, dient er gekozen te worden welk(e) ML-model(len) er gebruikt gaan worden.

Het probleem dat opgelost dient te worden, in combinatie met welke data en in welke vorm er beschikbaar is, bepalen welk(e) ML-algoritme(s) er gebruikt kan worden. Ook bestaat de mogelijkheid om meerdere algoritmes ingezet worden, een methode die ‘bagging’ of ‘stacking’ genoemd wordt. Hiermee kan mogelijk een meer robuust algemeen model gecreëerd worden. [But]

Een algemeen onderscheid dat gemaakt kan worden tussen de verschillende ML-algoritmes, is dat tussen ‘supervised learning’, ‘unsupervised learning’, ‘semi-supervised learning’ en ‘reinforcement learning’. Figuur 9 geeft een weergave van de veelgebruikte algoritmes per categorie. [Xi]



Figuur 9 - Indeling en bijbehorende algoritmes per soort model. Overgenomen uit Functional Nanomaterials Design in the Workflow of Building Machine-Learning Models door Xi, Z. (2021). arXiv e-prints, arXiv-2108.

Een ander onderscheid dat gemaakt kan worden tussen de verschillende algoritmes, is dat tussen regressie- en classificatie-algoritmes.

Regressie-algoritmes gaan op zoek naar de relatie tussen input X en output Y, op basis van gelabelde trainingsdata. Voorbeelden van dit soort algoritmes zijn Linear Regression, Logistic Regression en Support Vector Regression. [Schled] [Guo]

Classificatie algoritmes trachten de dataset in te delen naar verschillende (al dan niet vooraf gedefinieerde) categorieën. Voorbeelden hiervan zijn onder andere Logistics Regression en Support Vector Machines. [Schled]

Tevens bestaan er ook ML-algoritmes die in staat zijn de regressie- en classificatie-taak te combineren, zoals Decision Tree en Gradient Boosting. [Guo]

Een ander soort algoritme is dat van Artificial Neural Networks (ANN). Dit soort algoritme is gebaseerd op de werking van het menselijk brein. Een ANN heeft een input layer en een output layer, met daartussen één of meer hidden layers. Deze hidden layers bestaan uit activeringsfuncties (zoals een sigmoid function) die de input een bepaald gewicht geven en een output genereren voor de volgende laag in het netwerk. Als een netwerk bestaat uit meerdere hidden layers, spreekt men van een Deep Neural Network.

Een ANN is één van best werkende algoritmes voor Machine Learning, maar er is veel data nodig om het algoritme te trainen, en zijn ook rekenkundig intensief wat de nodige middelen vraagt. Er zijn echter al enkele succesvolle initiatieven waarbij een netwerk ook goed presteert met kleinere datasets. Eén van de voorbeelden hiervan is Elemnet, dat beter presteert dan conventionele ML-modellen, op basis van een dataset van slechts 4000 verbindingen. [Das]

Een groot voordeel van een ANN is dat deze in staat is om features te achterhalen uit de dataset. Dit neemt de noodzaak weg van de complexe taak om features te identificeren en ontwikkelen. [Das]

### 3.2.4 Model Evaluation

De laatste stap in het proces is het valideren van het getrainde ML-model. Met behulp van een test- of validatie-dataset worden de prestaties van het model ingeschat. Er bestaan drie bronnen van fouten (errors) waar rekening mee moet worden gehouden [But]:

* Model bias; het model bevat verkeerd aangeleerde aannames waardoor de juiste onderliggende relaties in de data niet worden gevonden.
* Model variance; het model levert verschillende outputs afhankelijk van welke training-dataset, of deel van de training-dataset wordt ingevoerd.
* Onherleidbare fouten.

Als een model niet goed presteert komt dit vaak door een hoge bias (underfitting) of een hoge variance (overfitting).

Underfitting geeft vaak aan dat er onvoldoende kwalitatief goede data is gebruikt voor het trainen van het model, wat ervoor zorgt dat het model niet in staat is om de relatie in de data te kunnen vinden.  
Overfitting houdt in dat het model te complex is geworden, vaak voortgekomen uit een grote toename van het aantal parameters. [But] [Das]

Er zijn verschillende meeteenheden om de prestatie van het model in uit te drukken [Das]:

* Loss function.
* Mean absolute error.
* Learning rate.
* Mean absolute relative error.
* Coefficient of determination R2.
* Receiver operating characteristic curve (ROC).

De meestgebruikte methoden om een model te evalueren zijn k-fold cross-validation en leave-one-out-cross-validation (LOOCV).

De betekenis en werking van deze meeteenheden en methoden is buiten de scope van dit onderzoek. Voor verdere informatie wordt verwezen naar het werk van [But] [Das] [Wei].

# 4. Literatuurlijst

[Agra] Agrawal, A., & Choudhary, A. (2016b). Perspective: Materials informatics and big data: Realization of the “fourth paradigm” of science in materials science. *APL Materials*, *4*(5), 053208. <https://doi.org/10.1063/1.4946894>

[Alm] De Almeida, A. F., Moreira, R., & Rodrigues, T. (2019). Synthetic organic chemistry driven by artificial intelligence. *Nature Reviews Chemistry*, *3*(10), 589–604. https://doi.org/10.1038/s41570-019-0124-0

[Ask] Askeland, D. R., Fulay, P. P., & Wright, W. J. (2010). *The Science and Engineering of Materials* (6de editie). CL Engineering.

[Att] Attig, N. (2004). *Computational Soft Matter: from Synthetic Polymers to Proteins*. NIC.

[Baum] Baum, Z. J., Yu, X., Ayala, P. Y., Zhao, Y., Watkins, S. P., & Zhou, Q. (2021). Artificial Intelligence in Chemistry: Current Trends and Future Directions. *Journal of Chemical Information and Modeling*, *61*(7), 3197–3212. <https://doi.org/10.1021/acs.jcim.1c00619>

[Bur] Burke, K. (2012). Perspective on density functional theory. *The Journal of Chemical Physics*, *136*(15), 150901. <https://doi.org/10.1063/1.4704546>

[But] Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., & Walsh, A. (2018). Machine learning for molecular and materials science. *Nature*, *559*(7715), 547–555. https://doi.org/10.1038/s41586-018-0337-2

[Cal] Jr., W. C. D., & Rethwisch, D. G. (2013). *Materials Science and Engineering: An Introduction* (9de editie). Wiley.

[Cram] Cramer, C. J., & Cramer, C. J. (2013). *Essentials of Computational Chemistry* (2de editie). Wiley.

[Das] Das, S., Pegu, H., Sahu, K. K., Nayak, A. K., Ramakrishna, S., Datta, D., & Swayamjyoti, S. (2020). Machine learning in materials modeling—fundamentals and the opportunities in 2D materials. In *Synthesis, Modeling, and Characterization of 2D Materials, and Their Heterostructures*(pp. 445-468). Elsevier.

[Guo] Guo, K., Yang, Z., Yu, C. H., & Buehler, M. J. (2021). Artificial intelligence and machine learning in design of mechanical materials. *Materials Horizons*, *8*(4), 1153–1172. https://doi.org/10.1039/d0mh01451f

[Jens] Jensen, F. (2017). *Introduction to Computational Chemistry* (3rd Edition). Wiley.

[Kit] Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 205395171452848. <https://doi.org/10.1177/2053951714528481>

[Mor] Morgan, D., & Jacobs, R. (2020). Opportunities and Challenges for Machine Learning in Materials Science. *Annual Review of Materials Research*, *50*(1), 71–103. <https://doi.org/10.1146/annurev-matsci-070218-010015>

[Poul] Poulson, T., & Walter, L. (2012). *Introduction to Chemistry*. Van Haren Publishing.

[Racc] Raccuglia, P., Elbert, K. C., Adler, P. D. F., Falk, C., Wenny, M. B., Mollo, A., Zeller, M., Friedler, S. A., Schrier, J., & Norquist, A. J. (2016). Machine-learning-assisted materials discovery using failed experiments. *Nature*, *533*(7601), 73–76. https://doi.org/10.1038/nature17439

[Sau] Saunders, M. N. K., Lewis, P., Thornhill, A., Arnoldy, M., & Smitt, P. (2019). *Methoden en technieken van onderzoek*. Pearson.

[Schled] Schleder, G. R., Padilha, A. C. M., Acosta, C. M., Costa, M., & Fazzio, A. (2019). From DFT to machine learning: recent approaches to materials science–a review. *Journal of Physics: Materials*, *2*(3), 032001. <https://doi.org/10.1088/2515-7639/ab084b>

[Seg] Segler, M. H. S., Preuss, M., & Waller, M. P. (2018). Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*, *555*(7698), 604–610. https://doi.org/10.1038/nature25978

[Sha] Sha, W., Guo, Y., Yuan, Q., Tang, S., Zhang, X., Lu, S., Guo, X., Cao, Y. C., & Cheng, S. (2020). Artificial Intelligence to Power the Future of Materials Science and Engineering. *Advanced Intelligent Systems*, *2*(4), 1900143. https://doi.org/10.1002/aisy.201900143

[Shack] Shackelford, J. F. (2022). *Introduction to Materials Science for Engineers* (8ste editie). PEARSON INDIA.

[Verhoev] Verhoeven, N. (2018). *Wat is onderzoek?* (6de editie). Boom Lemma.

[Weber] Weber, M. (2011) *Handboek ontwerpgericht wetenschappelijk onderzoek. Wetenschap met effect* (pp.177-202) Chapter: 11. Boom Lemma. Editors: Joan van Aken, Daan Andriessen

[Wei] Wei, J., Chu, X., Sun, X., Xu, K., Deng, H., Chen, J., Wei, Z., & Lei, M. (2019b). Machine learning in materials science. *InfoMat*, *1*(3), 338–358. <https://doi.org/10.1002/inf2.12028>

[Xi] Xi, Z. (2021). Functional Nanomaterials Design in the Workflow of Building Machine-Learning Models. *arXiv preprint arXiv:2108.13171.*

[Yuan] Yuan, R., Liu, Z., Balachandran, P. V., Xue, D., Zhou, Y., Ding, X., Sun, J., Xue, D., & Lookman, T. (2018). Accelerated Discovery of Large Electrostrains in BaTiO 3 ‐Based Piezoelectrics Using Active Learning. *Advanced Materials*, *30*(7), 1702884. https://doi.org/10.1002/adma.201702884

[Zur] ŻUrański, A. M., Martinez Alvarado, J. I., Shields, B. J., & Doyle, A. G. (2021). Predicting Reaction Yields via Supervised Learning. *Accounts of Chemical Research*, *54*(8), 1856–1865. https://doi.org/10.1021/acs.accounts.0c00770