Bedrijfstakonderzoek
BTO 2024.021| Januari 2024

Voorspellen van biologische verwijdering van persistente stoffen

Bedrijfstakonderzoek

KWR

Bridging Science to Practice

BTO 40

Colofon

Voorspellen van biologische verwijdering van persistente stoffen

BTO 2024.021| Januari 2024

Dit onderzoek is onderdeel van het collectieve Bedrijfstakonderzoek van KWR, de waterbedrijven en Vewin.

Opdrachtnummer
402045 -260

Projectmanager
Geertje Pronk/Jolijn van Engelenburg

Opdrachtgever
BTO - Verkennend onderzoek

Auteur(s)
Tessa Pronk, Peer Timmers, Jasper Immink, Xin Tian, Bas Wols

Kwaliteitsborger(s)
Peter van Thienen en Thomas ter Laak

Verzonden naar
Dit rapport is verspreid onder BTO-participanten.

Een jaar na publicatie is het openbaar.

Keywords

biologische afbraak, afvalwaterzuivering, QSAR, tekstmining

Jaar van publicatie
2024

Meer informatie
Tessa Pronk

T +316 15403473
E Tessa.Pronk @kwrwater.nl

PO Box 1072
3430 BB Nieuwegein
The Netherlands

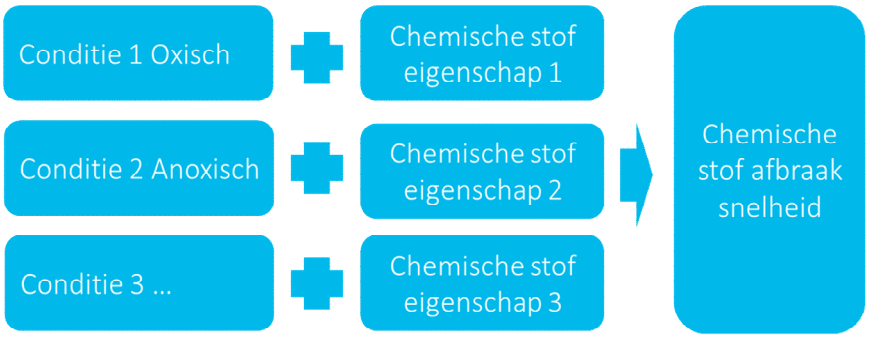
T +31 (0)30 60 69 511
E info@kwrwater.nl
I www.kwrwater.nl

Managementsamenvatting

Ideale omstandigheden voor biologische afbraak van persistente stoffen in de zuivering zijn nog niet goed te achterhalen en te voorspellen

Auteur(s) , Tessa Pronk, Peer Timmers, Jasper Immink, Xin Tian, Bas Wols

Biologische afbraak van organische microverontreinigingen (OMV's) speelt een belangrijke rol tijdens de waterzuivering. Vaak worden stoffen als persistent bestempeld, terwijl deze stoffen mogelijk onder andere omstandigheden goed biologisch afgebroken kunnen worden. Om hiervan een beter beeld te krijgen, is getracht te achterhalen welke parameters de biologische afbraak van OMV's beïnvloeden. Wanneer een relatie tussen zo'n parameter en de afbraak bekend is, is het mogelijk te voorspellen onder welke omstandigheden deze OMV's biologisch afgebroken kunnen worden (zie figuur). Deze kennis kan dan gebruikt worden om de waterzuivering aan te sturen. Vooralsnog is het lastig te achterhalen welke parameters de biologische afbraak van OMV's beïnvloeden, voornamelijk omdat de huidige gepubliceerde data in de wetenschappelijke literatuur ontoereikend is. Data wordt zeer inconsistent gerapporteerd, waardoor vergelijken van gerapporteerd onderzoek zeer lastig is en het opbouwen van een hoog kwalitatieve dataset zeer tijdrovend is. De beste oplossing voor dit probleem is het uitvoeren van meer grootschalige gecontroleerde experimenten waarbij dan een minimale set informatie moet worden gerapporteerd, bij voorkeur in een vorm die automatische tekstmining toelaat. Tekstmining van wetenschappelijke literatuur op basis van 'Large Language Models' is mogelijk succesvoller dan de regel-gebaseerde tekstmining op basis van natuurlijke spraakverwerking die in dit project is toegepast.



Doel van het onderzoek is een relatie vast te stellen tussen de biologische afbreekbaarheid van de organische microverontreinigingen (OMV's), de eigenschappen van de OMV in combinatie met condities zoals oxisch, anoxisch, methanogeen, pH, temperatuur, concentraties, HRT en redox-condities. Met een dergelijke relatie kan de degradatie van OMV's worden voorspeld onder deze condities.

Belang: Grip op biologische afbraakprocessen tijdens waterzuivering

Biologische processen spelen vaak een belangrijke rol bij de verwijdering van organische micro-verontreinigingen (OMV's) in de waterzuivering. Over de verwijdering van OMV's tijdens fysische en/of chemische processen als membraanfiltratie en actiefkoolfiltratie is steeds meer bekend. Tijdens deze processen kan de verwijdering dan ook goed worden voorspeld. Voor biologische processen in actief slib en tijdens zandfiltratie, actiefkoolfiltratie, oeverbankfiltratie en duininfiltratie is echter nog weinig bekend over verwijdering van OMV's. Hierdoor worden veel OMV's als persistent beschouwd in de huidige drink- en afvalwaterzuivering, terwijl deze stoffen mogelijk wel

biologisch kunnen worden afgebroken onder andere omstandigheden. Daardoor is het nog niet mogelijk om waterzuiveringsprocessen zo te sturen dat biologische verwijdering van (persistente) OMV's wordt gestimuleerd. Het is daarvoor belangrijk te achterhalen welke factoren sturend zijn voor verwijdering van specifieke OMV's of mengsels daarvan. Wanneer er relaties bestaan tussen de verwijdering van specifieke OMV's en factoren, kunnen die nader onderzocht worden om mogelijk de verwijdering van OMV's te voorspellen. Dergelijke informatie zou uiteindelijk kunnen worden gebruikt om de huidige zuivering te sturen.

Aanpak: Tekstmining en voorspellen van biologische afbraak

In dit project is gekeken naar diverse factoren die van belang zijn voor de biologische afbraak van OMV's, zoals pH, redox (oxische/anoxische) condities, temperatuur, zoutgehalte, nutriënten, koolstofbronnen en meer. Ook andere parameters zoals de matrix (actief slib, zand, etc.), de micro-organismen en biologische mechanismen die verantwoordelijk zijn voor de OMV-afbraak zijn bekeken.

Om voorspellingsmodellen te bouwen is eerst data nodig. Data is verzameld uit wetenschappelijke literatuur welke eerst geselecteerd werd met behulp van regel-gebaseerde tekstmining op basis van natuurlijke spraakverwerking. Daarna is handmatig de data geëxtraheerd. Zo is een kleine database opgebouwd met parameters die mogelijk sturend zijn voor verwijdering van OMV's. Deze database is gebruikt om met 'Machine learning' te onderzoeken of er relaties zijn tussen deze parameters en de verwijdering van OMV's.

Resultaten: Biologische afbraak is lastig te voorspellen door suboptimale datarapportage

Het blijkt vooralsnog lastig om te bepalen welke factoren de verwijdering van OMV's beïnvloeden. Daardoor is het voorspellen van de biologische afbraak van OMV's (nog) niet mogelijk. Dit heeft met name te maken met de uniformiteit en de uitvoerigheid van data die gerapporteerd is in de literatuur. De wetenschappelijke literatuur is inconsistent in de data die gerapporteerd worden, waardoor het haast onmogelijk is om studies onderling te vergelijken.

Toepassing: Verbeterde datarapportage is essentieel

Om de kwaliteit van de beschikbare data te vergroten, is het aan te raden experimenten rond biologische afbraak grootschalig en steeds onder vergelijkbare en gecontroleerde omstandigheden uit te voeren. Een andere aanbeveling is om uit te gaan van minimale sets informatie bij de rapportage van wetenschappelijk onderzoek, zodat rapportages consistent en vergelijkbaar worden. Ook zou onderzoek in wetenschappelijke artikelen zo gerapporteerd moeten worden, dat automatische tekstmining mogelijk wordt. Een standaard 'template' of protocol voor rapportage en/of het uploaden van ruwe data zou hiervoor een oplossing zijn, die in plaats van, of naast, de huidige leesversie van een artikel kan worden gepubliceerd. Dit maakt openbare data beter bruikbaar voor anderen. Een dergelijke praktijk zou niet alleen voor dit onderzoek, maar ook in het algemeen nodig zijn om beter relaties te vinden tussen gepubliceerde gegevens. In dit onderzoek zijn geen Large Language Models (LLM's, zoals GPT 4) geïmplementeerd, maar dergelijke LLM's kunnen in de toekomst mogelijk effectief zijn bij het extraheren van informatie uit ongestructureerde gegevens, zoals teksten.

Rapport

Dit onderzoek is beschreven in het rapport *Voorspellen van biologische verwijdering van persistente stoffen* (BTO 2024.021).

Meer informatie
Dr. Tessa Pronk
T +316 15403473
E Peer.Timmers@kwrwater.nl

PO Box 1072
3430 BB Nieuwegein
The Netherlands

Inhoud

Meer informatie
Dr. Tessa Pronk
T +316 15403473
E Peer.Timmers@kwrwater.nl

PO Box 1072
3430 BB Nieuwegein
The Netherlands

1 Inleiding

1.1 Aanleiding

In drinkwaterbronnen, zoals oppervlaktewater en grondwater, kunnen lage concentraties (ng/L - µg/L) organische microverontreinigingen (OMV's) aanwezig zijn. De concentraties van OMV's zullen naar verwachting in de toekomst toenemen in de drinkwaterbronnen door maatschappelijke- en klimaatontwikkelingen (vergrijzing, verdroging, etc.). Daarnaast hebben industriële lozingen in het verleden geleid tot hogere gehaltes van OMV's in het oppervlaktewater, wat resulteerde in tijdelijke innamestops bij enkele drinkwaterbedrijven. De verwachting is dat het aantal innamestops vanwege industriële lozingen zal toenemen in de toekomst. Ook is de kans groot dat er in de komende jaren weer "nieuwe" OMV's zullen opduiken. Dit betekent dat de drinkwatersector, naast het beschermen van de bronnen, deze stoffen ook goed moet kunnen verwijderen tijdens de drinkwaterzuivering. Drinkwaterbedrijven en waterschappen hebben steeds vaker en steeds meer te maken met stoffen waarvan een deel minder goed of niet verwijderd wordt in de zuivering, zogenaamde persistente stoffen.

Veel gebruikte zuiveringsprocessen voor het verwijderen van OMV's zijn membraanfiltratie, actieve-koolfiltratie, en (geavanceerde) oxidatieprocessen. Deze zuiveringsprocessen zijn allen in staat om OMV's grotendeels te verwijderen. Recent BTO onderzoek heeft zich gericht op het voorspellen van de zuivering van een groot deel van de OMV's voor deze drie de processen (Hofman-Caris et al., 2020a, Hofman-Caris et al., 2020b, Hofman-Caris et al., 2020c, Hofman-Caris et al., 2021). Deze voorspellingen op basis van zowel stofeigenschappen en procesmatige parameters zijn verwerkt in de online tool Aquapriori (de Vries et al., 2017). Deze tool kan de waterbedrijven helpen met het maken van keuzes in de zuivering om specifieke OMV's beter te verwijderen met bovengenoemde processen. Ook in het project 'Kennisimpuls Waterkwaliteit' zijn modellen gemaakt voor het voorspellen van OMV verwijdering met deze technieken (Pronk et al., 2023).

Nadeel van fysisch/chemische technieken is dat ze vaak energie-intensief zijn en gebruik maken van chemicaliën, of een reststroom produceren. Biologische processen in de drinkwaterzuivering vinden plaats in zand en/of multimedia snelfilters, oeverfiltratie, biologisch actieve-koolfiltratie, duinpassage en langzame-zandfiltratie. Deze worden toegepast in de meeste Nederlandse en Vlaamse drinkwaterzuiveringen om microbiologisch en chemisch stabiel en veilig drinkwater te produceren en kunnen mogelijk ook ingezet worden voor OMV-verwijdering. Hierdoor kunnen ze een alternatief vormen voor fysisch-chemische zuiveringsprocessen om OMV's te verwijderen. Daarom is de gedachte ontstaan om meer inzicht te krijgen in de potentiële biologische afbreekbaarheid van persistente OMV's: kunnen deze met andere biotische processen en onder andere condities verwijderd worden? Wanneer bekend is of stoffen mogelijk biologische afgebroken kunnen worden, kan onderzoek leiden tot het sturen van zuiveringsprocessen, waardoor de condities voor afbraak geoptimaliseerd kunnen worden.

1.2 Biologische afbreekbaarheid van OMV's

Door de toenemende aandacht voor OMV's in het milieu heeft ook het onderzoek naar de biologische afbreekbaarheid van deze OMV's een hoge vlucht genomen en neemt de hoeveelheid literatuur op dit onderwerp sterk toe. Onderzoek waar OMV-verwijdering bestudeerd wordt varieert van veldonderzoek tot lab-schaalonderzoek. Ook is er grote variëteit aan stoffen dat onderzocht wordt.

Voor biologische processen in de drinkwaterzuivering, zoals ze plaatsvinden in (zand en/of multimedia) snelfilters, oeverfiltratie, biologisch actieve-koolfiltratie, duinpassage en langzame-zandfiltratie is aangetoond dat zij een verscheidenheid aan OMV's kunnen verwijderen (Rattier et al., 2012, Zearley and Summers 2012, Bertelkamp et al., 2014, D'Alessio et al., 2015, Hijnen 2016, Hollender et al., 2018, Wang et al., 2021, Benner et al., 2013, Shimabuku

et al., 2019, Hedegaard and Albrechtsen 2014, Zhou et al., 2022, Timmers et al., 2023). Dit is ook aangetoond voor een breed scala aan OMV's tijdens biologische processen in de afvalwaterzuivering (Falås et al., 2016, Gonzales-Gil et al., 2019, Patureau et al., 2019). Omdat biologische processen in veel Nederlandse afval- en drinkwaterzuiveringen bestaan, zal het stimuleren of initiëren van een dergelijk proces dus (vaak) geen grote / extra investering vergen. Hierdoor zouden biologische zuiveringsprocessen een relatief goedkope, duurzame en milieuvriendelijke oplossing kunnen zijn voor de verwijdering van OMV's, mits deze toepasbaar zijn en goed werken. Om in te schatten wat er aangepast moet worden aan de huidige waterzuivering voor verbeterde OMV-verwijdering, is voor OMV's kennis nodig over de factoren die biologische afbreekbaarheid beïnvloeden, namelijk de eigenschappen van de stof (molecuulstructuur, hydrofobiciteit, etc.), de omstandigheden voor afbraak (pH, redox (oxische/anoxische) condities, temperatuur, zoutgehalte, etc.), de aanwezigheid van koolstofbronnen of nutriënten voor microbiële groei, de aanwezigheid van electronacceptoren en electrondonoren voor de biologische afbraak, de verblijftijd of contacttijd in het systeem en de aanwezigheid van de juiste micro-organismen.

Een van de uitdagingen om deze kennis over de biologische afbreekbaarheid van OMV's in kaart te brengen is het brede scala aan verschillende OMV's en de groeiende hoeveelheid literatuur die daarover gepubliceerd is. Ook ontbreekt er een standaard om de gegevens over afbreekbaarheid en de daarbij behorende condities in de wetenschappelijke literatuur te rapporteren, waardoor voor het verzamelen en sorteren van gegevens in een database veel expertise en interpretatie nodig is. Dit is een zeer arbeidsintensief proces, dat idealiter geautomatiseerd zou worden d.m.v. tekstmining of andere data-gedreven tools. Nadat er een database is opgebouwd met voldoende data, kan immers gekeken worden of er verbanden zijn tussen de afbraak van OMV's en bepaalde eigenschappen van de stoffen of de condities waarbij deze afgebroken werden. Daarnaast kan er gekeken worden of bepaalde interventies een positief of negatief effect hebben op OMV verwijdering. Mits voldoende gegevens aanwezig, kunnen op basis van de databases voorspellingen gedaan worden over de biologische afbreekbaarheid van OMV's.

1.3 Voorspellen van biologische afbreekbaarheid van OMV's

Biologische afbraak is een zeer belangrijk proces in de afval- en drinkwaterzuivering voor de verwijdering van nutriënten, maar ook van OMV's. Door de complexiteit van biologische afbraak heeft het meerwaarde om te kunnen voorspellen of een stof wel of niet biologisch afgebroken kan worden (het afbraakpotentieel) en in welke mate de stof onder bepaalde condities verwijderd kan worden per tijdseenheid (de afbraaksnelheid). Er zijn diverse methoden voor het bepalen van biologische afbraak. Een veelgebruikte methode om de biodegradatie experimenteel te bepalen is de OECD 309 methode. Deze methode bestudeert de aerobe afbraak van een stof in oppervlaktewater. De test wordt uitgevoerd door een teststof toe te voegen aan oppervlaktewater en dat voor 60 dagen onder oxische condities te incuberen. Als een stof in een afbraakstudie volgens de OECD 309 methode een halfwaardetijd (DT50) heeft die langer is dan 40 dagen, wordt deze stof als persistent beschouwd (Hofman-Caris et al., 2020d). In aanvulling op dergelijke experimentele methoden zijn er ook modellen die deze afbreekbaarheid van OMV's voorspellen. Deze worden hieronder besproken.

Een collectie van modellen voor het voorspellen van diverse stofeigenschappen is beschikbaar onder de naam OPERA (OPEn (quantitative) structure-activity Relationship Application). Opera-modellen werden al toegepast op meer dan 750.000 chemicaliën om vrij beschikbare voorspelde gegevens te produceren op het CompTox Chemistry Dashboard van de Amerikaanse Environmental Protection Agency (US EPA). Een van deze modellen voorspelt de biodegradatiehalfwaardetijd (in dagen): BIODEGRADATION_HALF_LIFE_DAYS_DAYS_OPERA_PRED. De methode maakt voornamelijk gebruik van gegevens uit de openbaar beschikbare PHYSPROP-database die bestaat uit gemeenschappelijke fysisch-chemische en milieutechnische eigenschappen. Deze datasets werden aangepast met behulp van een geautomatiseerde workflow om alleen hoogwaardige gegevens te selecteren, en de chemische structuren werden gestandaardiseerd voordat de moleculaire descriptoren werden berekend. De modelleringsprocedure is ontwikkeld op basis van de vijf principes van de Organisatie voor Economische

Samenwerking en Ontwikkeling (OECD) voor quantitative structure–activity relationship (QSAR)-modellen: een gedefinieerd eindpunt; een eenduidig algoritme; een gedefinieerd toepassingsgebied (AD); passende maatstaven voor de modelfit, robuustheid en voorspelbaarheid; en (indien mogelijk) een mechanistische interpretatie (OECD 2004). Het QSAR-concept is gebaseerd op het principe dat vergelijkbare structuren ook vergelijkbare eigenschappen hebben en daardoor een vergelijkbare, biologische activiteit of vergelijkbaar milieugedrag, zoals in dit specifieke geval biologische afbreekbaarheid, vertonen. Er werd gekozen voor een gewogen k-nearest neighbor met een minimum aantal vereiste descriptoren berekend met PaDEL, een open-source programma. De algoritmen selecteerden alleen de meest pertinente en mechanistisch interpreteerbare descriptoren (2-15, met een gemiddelde van 11 descriptoren). De afmetingen van de gemodelleerde datasets varieerden van 150 chemicaliën voor biologische afbreekbaarheidshalfwaardetijd tot 14.050 chemicaliën voor de octanol-water partitie coëfficiënt (K_{OW}). Deze laatste wordt uitgedrukt als $\log K_{OW}$. De biodegradatiehalfwaardetijd is er voor verbindingen die voornamelijk uit koolstof en waterstof bestaan (d.w.z. koolwaterstoffen). Om deze beperking te omzeilen en de gebruiker te helpen beslissen over de betrouwbaarheid van een voorspelling, is er een betrouwbaarheidsniveau-index toegevoegd die varieert van 0 tot 1 ten opzichte van de nauwkeurigheid van de voorspelling van de 5 naaste buren. Hoe hoger deze index, hoe waarschijnlijker het is dat de voorspelling betrouwbaar is. Deze maat wordt niet altijd meegenomen in de berekening, omdat deze niet wordt meegeleverd bij de opvraag van stofeigenschappen voor een batch met chemicaliën via het Chemistry Dashboard.

Andere schattingsmethoden voor afbraak zijn opgenomen in de vrij beschikbare reeks programma's die bekend staat als **EPI** (Estimation Programs Interface) Suite™. Deze reeks is ontwikkeld en wordt onderhouden door de US EPA en de Syracuse Research Corporation (SRC). Het kan gratis worden gedownload op https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface . Deels hebben deze modellen een gezamenlijke basis met OPERA; ze maken gebruik van de openbaar beschikbare PHYSPROP-database bestaande uit een set van fysisch-chemische en milieutechnische eigenschappen. De voorspellingen zijn dan ook gebaseerd op QSARs. Het voorspellingsmodel voor biologische afbraak (**Biowin**) (US EPA, 2023) is een van de meest gebruikte schattingsmethoden voor de biologische afbraak van algemene chemicaliën (Pavan and Worth 2008). Het is ontwikkeld om biodegradatie in afvalwaterzuiveringsinstallaties te voorspellen. Er zijn momenteel zeven verschillende modules opgenomen in het Biowin-programma die zowel lineaire als niet-lineaire regressiemodellen bevatten op basis van 36 vooraf geselecteerde fragmenten en molecuulgewicht. Het Biowin-model wordt gebruikt als hulpmiddel binnen de EU Technical Guidance Documents (EU TGD) om persistentie te beoordelen wanneer er geen gegevens beschikbaar zijn voor een bepaalde stof of de beschikbare gegevens moeilijk te interpreteren zijn (EC, 2003). Alle Biowin modellen geven een waarschijnlijkheid van afbraak (makkelijk of snel afbreekbaar), behalve Biowin3 en 4 die ook een snelheid aangeven. De EU TGD beveelt voor het bepalen van biologische afbreekbaarheid het gebruik van de Biowin2-modeluitvoer <0,5 of Biowin6-modeluitvoer <0,5 en Biowin3-uitvoer aan.

Biowin 3 en 4 leveren schattingen op voor de tijd die nodig is om volledige, ultieme en primaire biologische afbraak te bereiken in een typisch of "evaluatief" aquatisch milieu. Hierbij is de interpretatie van de classificatie: 5,00 -> uren 4,00 -> dagen 3,00 -> weken 2,00 -> maanden 1,00 -> langer. Het criterium om gemakkelijke afbreekbaarheid te voorspellen is of het Biowin3-resultaat in 'weken' of sneller is. Biowin3-uitkomsten in maanden (<2,2) moeten op betrouwbare wijze voor een stof kunnen bepalen of die al dan niet gemakkelijk biologisch afbreekbaar is en daarom het potentieel heeft om persistent te zijn in het aquatisch milieu (EC, 2003) (Ecetoc, 2013). Primaire biologische afbraak is de transformatie van een moederstof in een initiële metaboliet. Ultieme biologische afbraak is de transformatie van een moederverbinding in kooldioxide en water, minerale oxiden van andere elementen die in de testverbinding aanwezig zijn en nieuw celmateriaal. Vervolgens geeft de tool via elk model de beoordeling met betrekking tot de tijd die nodig is om ultieme en primaire biologische afbraak te bereiken in een typisch of "evaluatief" aquatisch milieu. De beoordelingen voor elke verbinding wordt gemiddeld, zodat een enkele waarde voor modellering wordt verkregen. De uiteindelijke of primaire beoordeling van een verbinding wordt berekend door deze waarden op te tellen voor alle fragmenten die in die verbinding aanwezig zijn. Biowin7 is een lineair model voor de waarschijnlijkheid van anaerobe afbraak (Ecetoc, 2013).

De robuustheid van de OPERA en EPISuite voorspellingsmodellen en methoden zijn vergelijkbaar ($0,7 < R^2 < 0,85$, gebaseerd op R^2's waarbij de modellen zijn toegepast op testdatasets zoals gerapporteerd door OPERA en EPI Suite) (Mansouri et al., 2018). Hier wordt de volgende R^2-definitie gebruikt:

$$R^2 = 1 - \frac{RSS}{TSS}$$

Hier is RSS de *residual sum of squares*: de verschillen tussen modelvoorspelling en eigenlijke waarde, gekwadrateerd en gesommeerd. De TSS is de total sum of squares: de verschillen tussen gemiddelde en eigenlijke waarde, gekwadrateerd en gesommeerd. De TSS is ook wel de variantie. Biowin modellen presteren beter bij het voorspellen van niet gemakkelijk biologisch afbreekbare stoffen dan het voorspellen van gemakkelijk biologisch afbreekbare stoffen (Ecetoc, 2013).

BioTransformer 3.0 is een softwareprogramma dat het metabolisme van kleine moleculen voorspelt in zoogdieren, in hun darmmicrobiota en in bodem-/aquatische microbiota. BioTransformer bestaat uit vijf onafhankelijke modules: EC-gebaseerd, CYP450, Fase II, Humaan darm microbieel en omgeving microbieel. Deze laatste is relevant voor de biologische afbraak of omvorming van stoffen zoals die in een rioolwaterzuiveringsinstallatie plaatsvindt. BioTransformer gebruikt naast een op machine-learning gebaseerde benadering om het metabolisme van kleine moleculen te voorspellen, ook een set regels van het EAWAG-BBD/PPS-systeem om de producten van microbiële afbraak in het milieu te voorspellen. Het pathway prediction system van de Universiteit van Minnesota (UM-PPS, http://umbbd.msi.umn.edu/predict/) herkent functionele groepen in organische verbindingen die potentiële doelwitten zijn van microbiële katabole reacties en voorspelt transformaties van deze groepen op basis van biotransformatieregels. De regels zijn gebaseerd op de biokatalyse/biodegradatiedatabase van de Universiteit van Minnesota (http://umbbd.msi.umn.edu) en de wetenschappelijke literatuur. De software is beschikbaar via een webapplicatie. Input is een SMILES code of SDF van een stof (precursor), en output is een lijst van stoffen die resulteren (metabolieten), inclusief reactie (bv. 'fosforylatie') en het betrokken enzym. Het geeft dus geen informatie over de afbraaksnelheid en de condities die nodig zijn voor deze afbraak.

Voor afbraak in rioolwaterzuiveringsinstallaties wordt het model SimpleTreat 4.0 (Struijs, 2014) gebruikt. SimpleTreat is een beoordelingstool voor het lot van stoffen in afvalwaterzuiveringsinstallaties. De tool houdt rekening met de belangrijkste processen zoals vervluchtiging, menging, adsorptie en afbraak. Naast het gebruik als onderzoeksinstrument, wordt SimpleTreat 4.1 gebruikt om stoffen te beoordelen in overeenstemming met REACH (Registration, Evaluation, Authorization and restriction of CHemicals) en andere regelgeving omtrent stoffen in de EU, zoals biociden, geneesmiddelen en gewasbeschermingsmiddelen. Het afbraak-deel in dit model is relevant. Voor biologische afbraak in actief slib wordt een eerste orde kineticamodel gebruikt, gebaseerd op de halfwaardetijd. De halfwaardetijd is geschat uit OECD/EU gestandaardiseerde biodegradatietests OECD 301, 302 en 310 (https://www.umweltbundesamt.de/en/publikationen/application-of-simpletreat-40-in-european-substance) . Voor anaerobe vergisting is een tweede model beschikbaar. Anaerobe biologische afbraak is vooral relevant voor chemische stoffen met een hoge verdelingscoëfficiënt tussen vaste stof en water (Struijs, 2014). De halfwaardetijd voor eliminatie van de chemische stof in de anaerobe vergister samen met de concentratie, wordt gebruikt om de chemische flux via vaste stoffen die de bodem bereikt te berekenen. Dit resulteert in een anaerobe reductiefactor (ARF) als gevolg van anaerobe biologische afbraakprocessen in de gistingstank. De factor is afhankelijk van de gemeten verblijftijd in de anoxische tank en is een functie van de gemeten halfwaardetijd.

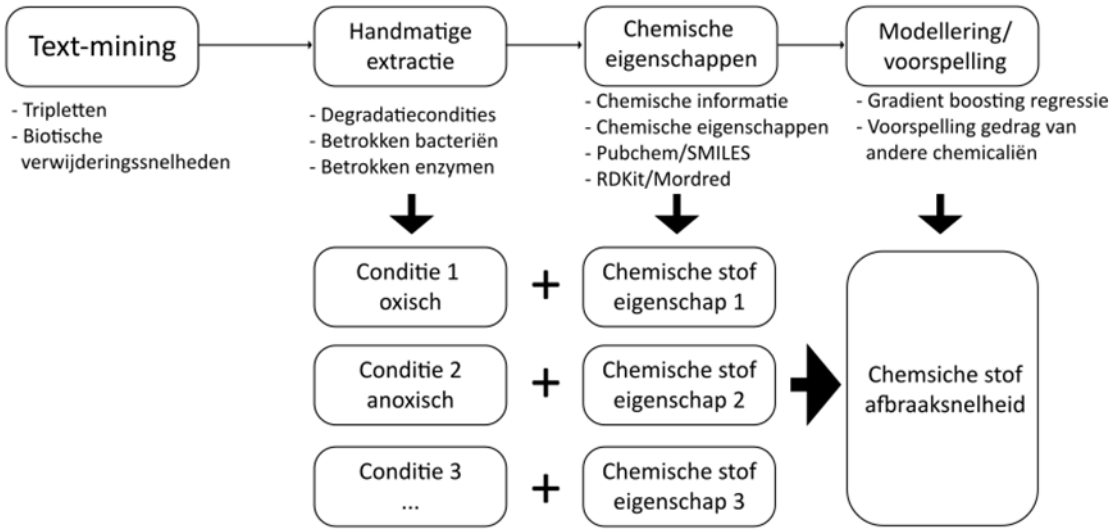
Ook in EPISuite is een berekening voor totale verwijdering in een rioolwaterzuiveringsinstallatie beschikbaar. Dit is de 'removal in waste water treatment', opgesplitst in biodegradatie, adsorptie (voor een groot deel afhankelijk van logK_{OW}), en vervluchtiging (Henry's constante). In de standaardmodus is de biodegradatie zeer laag, vanwege een instelling op een standaard zeer hoge halfwaardetijd voor stoffen, wat dit model dus niet bruikbaar maakt voor biodegradatie (Straub et al., 2022).

Samenvattend: met biodegradatievoorspellingsoftware OPERA Half-life, Biowin, SimpleTreat en BioTransformer, kan worden voorspeld of stoffen biologisch afbreekbaar zijn en wat de mogelijke transformatieproducten zijn. Biowin kan voorspellen of een stof snel of langzaam afgebroken wordt onder oxische condities en BioTransformer geeft informatie over transformatieproducten. Deze tests en software zijn alle gebaseerd op aerobe en snelle afbraak en geven geen informatie over de afbraaksnelheid en de condities die nodig zijn voor deze afbraak. Dit betekent dat wanneer een stof als persistent en niet-afbreekbaar wordt bevonden door deze modellen, dat dit niet altijd werkelijk het geval hoeft te zijn. De processen in de afval-en drinkwaterzuivering waar biologische afbraak een rol speelt treden namelijk niet alleen op onder zuurstofrijke condities, maar ook onder zuurstof-limiterende en zuurstofloze condities. Denk hierbij aan denitrificatie in de afvalwaterzuivering of anoxische condities in grondwater en tijdens duininfiltratie en oeverbankfiltratie. Daarnaast kan het zijn dat de verwijdering pas na 28 dagen optreedt, wat mogelijk is tijdens grondwaterextractie en bij duininfiltratie of oeverbankfiltratie. De resultaten van Biowin voorspellingen zijn dus indicatief voor de persistentie van de stof in het milieu, maar geven niet de gehele afbraakpotentie van een stof weer in alle systemen en condities relevant voor de (drink)waterzuivering. Er is al eerder aangetoond dat stoffen zoals gabapentine, 1-H benzotriazool, die als persistent bevonden waren met OECD biodegradatietests (Hofman-Caris et al., 2020d), wel degelijk tijdens snelle zandfiltratie, duininfiltratie of oeverfiltratie goed verwijderd worden (Timmers et al., 2023; Timmers et al., 2022; van de Grift et al., 2020). SimpleTreat biedt wel een berekening voor anaerobe afbraak, gebaseerd is op halfwaardetijd en verblijftijd. Ook Biowin7 voorspelt de waarschijnlijkheid van anaerobe afbraak, op basis van QSAR.

Bovenstaande informatie laat zien dat er dus nog een lacune is in de voorspellingstools en databases betreft biologische afbraak van stoffen die niet ‘ready biodegradable’ zijn, d.w.z. de stoffen die verwijderd worden onder niet-zuurstofrijke condities en niet binnen 28 dagen. Deze stoffen worden door Biowin als persistent bevonden, maar worden in de realiteit mogelijk wel verwijderd tijdens bepaalde waterzuiveringsprocessen en in het milieu. Daarnaast helpen de huidige voorspellingstools niet om te weten te komen onder welke condities een stof beter verwijderd wordt. Om de waterzuivering te kunnen sturen op verbeterde biologische OMV-verwijdering, is het belangrijk om kennis te vergaren over het afbraakpotentieel en in welke mate een stof onder bepaalde condities verwijderd kan worden (i.e. de afbraaksnelheid).

1.4 Doel/opzet van dit onderzoek

In dit project is een verkenning gedaan om systematisch kennis te vergaren over de biologische afbraak van OMV’s via tekstmining en andere data-tools. Daarnaast is verkend of een database kan worden opgebouwd met deze kennis en of er verbanden te vinden zijn tussen afbraak van OMV’s en bepaalde factoren uit de literatuur. Ook is verkend hoe deze kennis te gebruiken voor voorspellingsmodellen, met behulp van quantitative structure–activity relationship (QSAR), ofwel kwantitatieve structuur-activiteit relaties, om de eigenschappen die biologische afbraak dicteren in kaart te brengen. Een schematische weergave van dit onderzoek is weergegeven in Figuur 1.

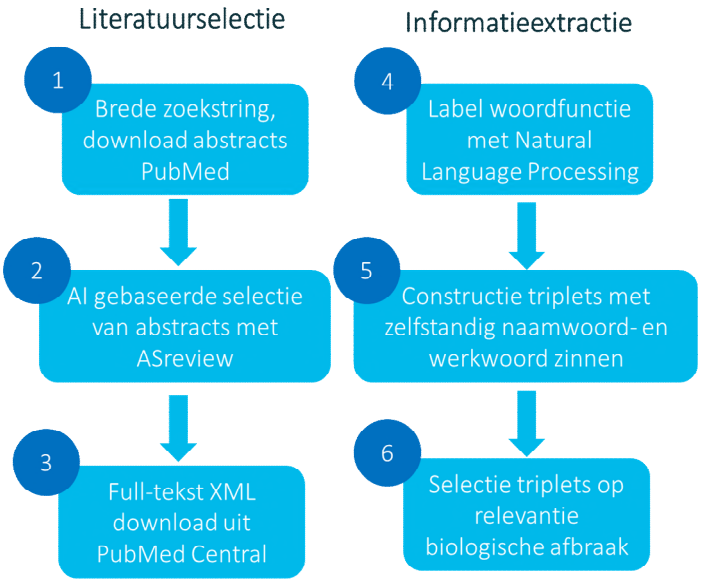


Figuur 1: Schematische opzet van dit project.

2 Methoden

2.1 Tekstmining voor biologische afbraaksnelheden in de waterzuivering

Op basis van data kan een zogenaamd 'Quantitative Structure Activity Relation' (QSAR) model gemaakt worden rond biologische afbraak. In zo'n model worden stofeigenschappen (o.a. directe eigenschappen zoals massa, dimensies, ladingverdeling, structuur en afgeleide eigenschappen zoals wateroplosbaarheid, vluchtigheid) gekoppeld aan zuiveringsefficiëntie onder verschillende omstandigheden via een statistisch model. Voor stofeigenschappen zijn verschillende openbare databases beschikbaar, verder beschreven in Paragraaf 2.3. Voor zuivering van stoffen onder verschillende omstandigheden is dit niet het geval. Om een database samen te stellen met gegevens over afbraak van chemische stoffen onder verschillende omstandigheden in de zuivering, is gekeken in wetenschappelijke artikelen. Vanwege de grote omvang van wetenschappelijke literatuur, is tekstmining potentieel een efficiënte manier om informatie te extraheren. Abstracts van artikelen hebben in de regel een toegankelijk formaat en bevatten vaak de belangrijkste resultaten van een studie. Desondanks levert het doorzoeken van 'full-tekst' artikelen naar verwachting meer concrete getallen rond afbraaksnelheden op dan het doorzoeken van abstracts. In een abstract is nu eenmaal beperkt ruimte voor resultaten. Daarom is er gezocht in 'PubMed Central'; een database met 9 miljoen gearchiveerde full-tekst artikelen (t/m juni 2022). Informatie is geëxtraheerd in een aantal volautomatische stappen, hieronder uitgelegd in Figuur 2.



Figuur 2. Stappen om informatie rond biologische afbraak te extraheren uit veel literatuur.

1. Brede selectie van relevante artikelen. In een eerste stap zijn door middel van een zoekstring (zie Bijlage I) mogelijk relevante artikelen verzameld. De zoekstring is zo gemaakt dat het zoveel mogelijk relevante artikelen bevat betreft biologische afbraak van organische microverontreinigingen.

2. De abstracts van deze hits zijn vervolgens ingeladen in de open source systematische review tool AS Review. In deze tool kan, per abstract, aangegeven worden of deze relevant is of niet. In deze tool loopt op de achtergrond een deeplearningmodel mee om een inschatting van de relevantie van onbeoordeelde

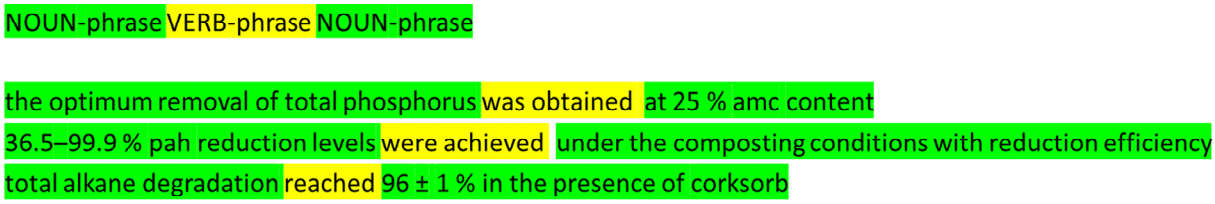
abstracts te maken. Als het model voldoende is getraind, geeft AS Review dit aan. Vervolgens kunnen de meest relevant geschatte abstracts worden geselecteerd.

3. Extraheren van relevante informatie. Met behulp van de identificatienummers voor artikelen, 'Pubmed ID' en (PMID) nummers, die geadministreerd zijn bij de abstracts zijn vervolgens de full-tekst XML versies gedownload van PubMed Central. Deze XML files zijn te doorzoeken via de programmeertaal 'R' via de pakketten 'XML' en 'xml2'. Per artikel zijn eerst alle tekstparagrafen geselecteerd; dit is in de XML files namelijk aangegeven met een code. Het doel van het doorzoeken was om 'triplets' te selecteren die bestaan uit een zelfstandig naamwoord, werkwoord en zelfstandig naamwoord waarin een feit staat dat gerelateerd is aan verwijdering van een chemische stof.

4. Om de triplets te kunnen extraheren zijn de zinnen in de tekstparagrafen met behulp van Natural Language Processing (NLP) gelabeld met hun functie (b.v. 'werkwoord', 'zelfstandig naamwoord', 'bijvoeglijk voornaamwoord', etc.) via het pakket 'OpenNLP'.

5. Woorden zoals voorzetsels, bijwoorden en betrekkelijke voornaamwoorden die bij een zelfstandig naamwoord horen zijn op basis van hun labels automatisch aan dit zelfstandig naamwoord toegevoegd tot een 'zelfstandig naamwoordzin' ontstaat. Hetzelfde werd gedaan voor werkwoorden om tot een 'werkwoordzin' te komen. Per zin is vervolgens geselecteerd op de aanwezigheid van de volgorde: zelfstandig naamwoordzin, werkwoordzin, zelfstandig naamwoordzin (een 'triplet'). Een voorbeeld hiervan is 'total alkane degradation (**zelfstandig naamwoord1**) reached (**werkwoord**) 96 ± 1 % in the presence of corksorb (**zelfstandig naamwoord2**)'. Zie voor meer voorbeelden van triplets Figuur 3.

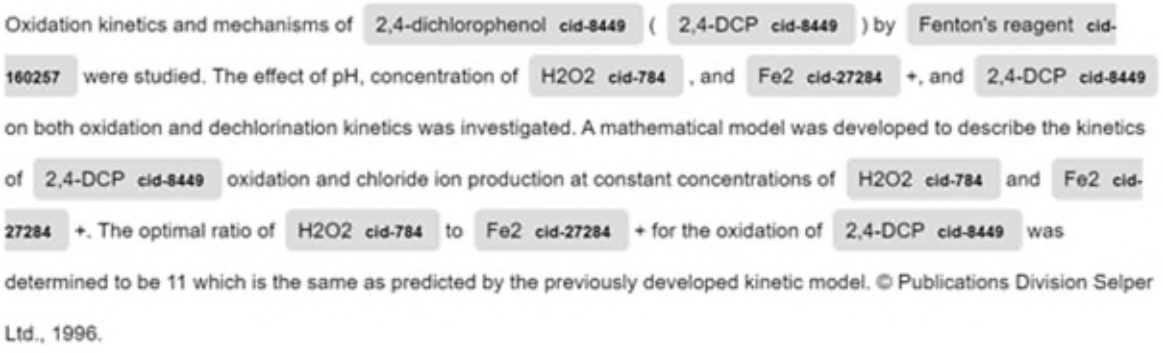
6. Een selectie is vervolgens gedaan op de triplets die te relateren zijn aan degradatie door te selecteren op woorden of woorddelen ('dt50', 'reduc', 'removal'). Een extra selectie om daadwerkelijke afbraakpercentages te verkrijgen is gedaan op de aanwezigheid van '%', behalve bij triplets die rond dt50 zijn gebouwd. Het is in de triplets niet altijd duidelijk om welke stof het gaat. Soms staat de stof zelf niet in het triplet. Om dit te verbeteren is op de titels van de artikelen waar de triplets uit kwamen een techniek toegepast die 'named entity recognition' (zie Figuur 4) heet om chemicaliën te duiden. Hierbij wordt met machine learning eerst bepaald of het waarschijnlijk is dat in de zin naar een chemische stof verwezen wordt, en vervolgens worden de woorden in de zin gekoppeld aan namen van chemicaliën in PubChem. Vanwege een lange benodigde rekentijd is dit uitgevoerd op titels, niet op het abstract of de full-tekst.



Figuur 3. Voorbeeld van 'triplets' rond biologische afbraak. Kleuren zijn slechts ter illustratie; alleen de tekst wordt geproduceerd in de methode (zie Figuur 2 voor de stappen hierin).

Waarom geen geavanceerde concept mining door middel van machine learning?

De meest geavanceerde techniek voor concept mining (het herkennen van concepten zoals 'afbraak') ten tijde van de uitvoering van dit project rond taalmodellen vereist handmatige labelen van een groot aantal teksten, training van taalmodellen en vervolgens toepassing van het getrainde model op nieuwe teksten (Figuur 4). De eerste stap, handmatige labelen van teksten, is een tijdrovend proces. Zo kostte het met de hand labelen van chemische verbindingen in wetenschappelijke papers meer dan 80 uur in een ander BTO-project Zeer Zorgwekkende Stoffen (ZZS). Voor dit huidige project zouden meerdere concepten herkend moeten worden rond de afbraak, stoffen, en omstandigheden van de gemeten resultaten in de literatuur, en de daadwerkelijke waarden zouden, waar deze bij elkaar horen, gekoppeld aan elkaar geadministreerd moeten worden. Vanwege budget- en tijdsbeperkingen konden we deze stap niet implementeren in het huidige project. Het BTO-project ZZS toont de verbeterde effectiviteit van modellen rond Generative Pre-trained Transformers (GPT). Deze modellen gebruiken 'deep learning' om tekst te genereren. GPT en vergelijkbare technologieën kunnen ook efficiënt informatie uit teksten extraheren. De ZZS-studie toonde aan dat deze technologieën in sommige gevallen met meer dan 70% nauwkeurigheid informatie uit teksten kunnen halen. Ook al is dit een veelbelovende nieuwe ontwikkeling, data-kwaliteit is heel belangrijk voor het maken van goede verklarende modellen voor biologische afbraak. Daardoor is ook de meest geavanceerde tekstminingtechniek nog steeds niet toereikend om zonder handmatige correcties en controles in de literatuur zelf dit soort pluriforme data te extraheren.



Figuur 4: Schematische weergave van het labelen van tekst uit wetenschappelijke literatuur voor het herkennen van concepten door op machine learning gebaseerde tekstminingtools ('named entity recognition'). In dit voorbeeld worden chemische stoffen gelabeld met een PubChem Compound Identifier (cid).

2.2 Databasesjabloon voor afbraak van stoffen onder verschillende omstandigheden

De triplets geven inzicht in de aanwezigheid van daadwerkelijke getallen rond biologische afbraak in de wetenschappelijke artikelen, en de stoffen waarover het gaat. Hiervoor hoeft de full-tekst van deze artikelen dus niet doorgelezen te worden. De triplets zijn voldoende. Deze informatie is gebruikt om artikelen te selecteren voor

handmatige dataextractie voor het opbouwen van een database. Voor de dataextractie is een sjabloon gemaakt met verschillende typen informatie die experts bij KWR als belangrijk hebben gemarkeerd. In dit sjabloon zijn de volgende parameters toegevoegd die, voor zover mogelijk, handmatig uit de literatuur geëxtraheerd waren:

1. CAS-nummer
2. Compound (Stof)
3. Mixture Effect (Mix effect): is in het artikel 1 OMV bestudeerd of een mengsel van OMV's?
4. Removal (Verwijdering): Trad er verwijdering op van de stof(fen)? Ja of nee
5. electron donor: welke electrondonor was toegevoegd? Dit kan de OMV zelf zijn (metabolische afbraak) of een andere donor (co-metabolische afbraak)
6. Electron acceptor: welke electron acceptor was toegevoegd? Dit kan bijvoorbeeld zuurstof of nitraat zijn
7. Experiment type: wat voor soort experiment was het? Bijv full-scale of lab-scale reactor
8. Condition (Conditie): Welke condities heerste in het experiment? Oxisch/anoxisch/methanogeen etc.
9. Metabolism (Metabolisme): Is het metabolisme van OMV-afbraak bekend? Metabool, co-metabool, heterotroof, autrotroof, etc.
10. Stimulant/addition (Addititie): Was er een stimulant toegevoegd om OMV afbraak te verbeteren? Bijvoorbeeld bacteriën of voeding voor bacteriën of andere stoffen
11. Temperature (Temperatuur) (°C)
12. pH (Zuurgraad)
13. Redox potential (Redoxpotentiaal)
14. Biomass/Matrix (Biomassa): welke biomassa was gebruikt? Bijvoorbeeld actief slib, etc.
15. Removal percentage (Verwijderingspercentage) (%)
16. Difference Removal % (Verschil in verwijdering): wanneer een stimulant was gebruikt, wat was het verschil in verwijdering t.o.v. de controle?
17. Removal rate (Verwijderingssnelheid (aantal/tijdseenheid)): snelheid waarmee een stof wordt verwijderd
18. Removal time (Periode voor verwijdering): hoelang duurde het voordat de stof verwijderd was
19. Start concentration pollutants (Startconcentratie): wat was de concentratie van de stof(fen) op t=0?
20. DT50 or t1/2: halfwaardetijd
21. Difference DT50 (Verschil DT50): wanneer een stimulant was gebruikt, wat was het verschil t.o.v. de controle?
22. Rate constant (Snelheidsconstante) (k)
23. Difference Rate constant (Verschil snelheidsconstante): wanneer een stimulant was gebruikt, wat was het verschil t.o.v. de controle?
24. Michaelis-Menten constant (Km)
25. Organism (Organisme): welke organisme was verantwoordelijk voor OMV verwijdering?
26. Pathway (Metabole route): welke metabole route was verantwoordelijk voor OMV verwijdering?
27. Enzymes (Enzym): welk enzym was verantwoordelijk voor OMV verwijdering?
28. Transformation products (Transformatieproducten): welke transormatieproducten werden gevormd?
29. PaperDOI: de permanente weblink naar het bijbehorende wetenschappelijk artikel

Van de resultaten uit de vorige stap waar een selectie van artikel uit gekomen is, zijn een gelijk aantal van 10 artikelen per OMV uitgekozen en verdeeld over verschillende onderzoekers om data manueel te extraheren en in de database te zetten.

2.3 QSAR voor eigenschappen die afbraak voorspellen onder verschillende omstandigheden in de zuivering

Een QSAR is, zoals in Paragraaf 1.3 beschreven, een type model waar de chemische structuur van een molecuul wordt verbonden aan de eigenschappen en functionaliteit van het molecuul (zoals bijvoorbeeld de biologische enzymatische activiteit). Dit gebeurt door deze moleculaire eigenschappen te vangen in allerlei moleculaire parameters (zogenaamde descriptoren) en regressietechnieken toe te passen om de uitwerkingen in getallen te voorspellen. Zo kan een QSAR aan de hand van dit model voorspellingen doen over de activiteit van gegeven moleculen.

In dit onderzoek werden deze QSAR's ontwikkeld aan de hand van datasets verkregen uit de tekstmining. De eerste stap was het bepalen van de *chemische descriptoren* voor de gebruikte chemische stoffen. Met chemische descriptoren worden gekwantificeerde eigenschappen van het molecuul bedoeld: voorbeelden zijn het aantal carboxyl- of aminegroepen, of de fysieke grootte van het molecuul. Deze descriptoren werden berekend aan de hand van de betreffende molecuulstructuur: door middel van het Python-package Mordred werden per molecuul ruim 1000 chemische descriptoren berekend. Daarnaast werden ook empirische descriptoren in het model meegenomen, welke voornamelijk bepaald zijn uit de literatuur. Als *doelparameter* (de te voorspellen parameter in de QSAR) werd de verwijderingsefficiëntie genomen, zoals verkregen met behulp van de tekstmining. Met behulp van de bovengenoemde regressietechnieken kan een QSAR aan de hand van een reeks van chemische descriptoren van een molecuul één of meerdere doelparameters voorspellen.

De QSAR-modelbouw werd door het Python-package scikit-learn gefaciliteerd. De uitvoering is als volgt. Allereerst werd een dataset gebouwd waarin alle descriptoren werden gekoppeld aan hun bijbehorende doelparameter. Vervolgens werd deze dataset opgedeeld in een train- en een testset, met een verhouding van 80%-20%. Dit doet men om 20% van de data achter te houden om de voorspelbaarheid van het model te testen op data die niet gebruikt is om het model te trainen. Het regressiemodel dat de doelparameters voorspelt is een Machine Learning-model dat met behulp van statistische methoden de doelparameters voorspelt. In alle gevallen was dit de Gradient Boosting Regressor (GBR). Vervolgens werden uit de lijst van chemische descriptoren de meest waardevolle chemische descriptoren geselecteerd – waardevol betekent hier de voorspellende kracht die elke chemische parameter heeft om de doelparameter te voorspellen. Dit werd gedaan met de *relevantie*-waarde (Engels: *importance)*: hoe hoger de relevantie-waarde van een chemische descriptor, hoe relevanter deze is voor de voorspelling van de doelparameter.
Een aantal parameters met de hoogste relevantie-waarde werden geselecteerd en als uiteindelijke descriptoren in de modelontwikkeling gebruikt, terwijl de rest niet werd meegenomen. Dit werd gedaan omdat een kleinere hoeveelheid parameters vaak tot betere modellen leidt (te veel parameters resulteerde in overfitting van de resultaten), en er werd gekozen om in alle modellen de belangrijkste 15 parameters in de uiteindelijke modelbouw mee te nemen.

Na het opdelen van de dataset, werd de trainingset gebruikt om het model te trainen. Dit begon door de data te normaliseren en herschalen rond nul. Dit verbetert de efficiëntie van het modelbouwproces zonder in te boeten aan modelkwaliteit. Hierna werd een hyperparameter-optimalizatiestap uitgevoerd d.m.v. een grid search. Dit houdt in dat alle variabelen die het GBR-model beschrijven (hyperparameters) op gestructureerde wijze worden gevarieerd en de set hyperparameters die het best voorspellende model voortbrengen wordt gevonden. De hyperparameters die geoptimaliseerd werden waren de hoeveelheid *submodellen* (jargon: *estimators*), de maximale diepte van de GBR, en de *leersnelheid* (Engels: *learning rate)*.
De kwaliteit van een model werd bepaald door tienvoudige kruisvalidatie: de doelparameter dataset werd op tien verschillende manieren verdeeld in model-train gedeelte (75%) en een validatiegedeelte (25%). Dit betekent dat de originele data is opgesplitst in 60% model-traingedeelte, 20% validatiegedeelte en 20% testgedeelte. Een model wordt dan getraind op het model-train gedeelte en de prestatie op het grid searchwordt opgeslagen. De prestatie

van de kruisvalidatie wordt dan bepaald door het gemiddelde te nemen van tien verschillende validatieprestaties. De kruisvalidatie geeft de kwaliteit van ieder punt in de grid search aan en de beste combinatie van hyperparameterwaarden wordt gekozen aan de hand van de verkregen R^2 van het model op de validatieset. Ook werd uiteindelijke modelkwaliteit gekwantificeerd met behulp van de R^2-waarde op de testset.

Dit onderzoek is toegespitst op het vinden van chemische descriptoren die mogelijk een effect kunnen hebben op biotische verwijdering. Daarom is één van de belangrijkste modeloutputs de relevantie van de chemische descriptoren in het uiteindelijke model. Omdat deze waarde aangeeft hoeveel effect de descriptor heeft op de voorspelling van het uiteindelijke model, zou de ideale output van onze modelleringssoftware het volgende zijn:

- een model dat goede voorspellingen doet op de training-data én op de testdata (*lage systematische afwijking/variantie of bias/variance*) kan voorspellen welke stoffen verwijderd zullen worden;

- één of meerdere chemische parameters die een duidelijke hoge relevantie hebben binnen dat model.

De gegenereerde relevantiewaardes voor chemische descriptoren dienen als een kwantificering van hoe veelbelovend een descriptor is voor de (biotische) verwijdering van het molecuul.

Om een goed beeld te krijgen van de chemische parameters voor biotische verwijdering, worden een aantal verschillende modellen gebouwd om de voorspellende kracht van de data en de descriptoren in verschillende omstandigheden te testen. In volgende paragrafen zal dieper ingegaan worden op de verschillende modeltypes. De volgende verschillende modellen zijn gebouwd:

- Variërende doelparameter (verwijderingsefficiëntie). In het databestand staan drie verschillende verwijderingsefficiënties getabelleerd: (i) rate constants; (ii) halfwaardetijden; (iii) combinatie van tijdsspanneen verwijderingspercentage. Er wordt een model gebouwd voor elk van deze drie opties, met een gefitte verwijderingssnelheden voor de derde optie zoals hieronder beschreven in paragraaf 2.3.1.

- Modellen specifiek voor de toestand. Er worden drie modellen gebouwd, waar steeds alleen de data wordt meegenomen die: (i) oxische toestand beschrijft; (ii) methanogene toestand beschrijft; (iii) anoxische, non-methanogene toestand beschrijft.

- Vergelijkingen van enkele experimenten. Er worden modellen gebouwd waar alleen data van één enkel experiment wordt meegenomen. Dit betekent dus identieke experimentele condities, in plaats van over een breed scala aan experimenten.

2.3.1 Doelvariabele

In de resultaten van de tekstmining waren drie verschillende indicatoren van zuiveringsefficiëntie te vinden. De dataset werd opgesplitst in drie verschillende datasets voor modeltraining. Dit werd gedaan om de vergelijkbaarheid binnen één dataset te vergroten, en een dataset te verkrijgen waar de parameters beter voorspellend zijn dan de binnen de grote, diverse dataset.
Om te beginnen bevatte de database voor 78 verbindingen gepubliceerde halfwaardetijden. Daarnaast bevatte deze voor 105 verbindingen de verwijderingssnelheid (proportionaliteitsconstante of afbraakconstante voor de betreffende chemische reactie). Voor het grootste gedeelte van de waarnemingen (391 waarnemingen) was een verwijderingspercentage in combinatie met een tijd waarin die verwijdering plaatsvond beschikbaar. Omdat de getabelleerde tijden onderling significant verschilden, werd een eerste-orde chemische reactievergelijking gefit aan deze waarden, met de vorm:

$$C_t = C_0 * e^{-kt},$$

met C_t de concentratie van de chemische stof op tijdstip t, C_0 de concentratie van de chemische stof op tijdstip $t = 0$, en k de snelheidsconstante van de chemische reactie. Hier is het belangrijk om te vermelden dat een eerste-orde chemische reactie impliceert dat de microbiologische componenten in overmaat aanwezig zijn. Dit is niet altijd juist, maar omdat in de publicaties vaak niet meer informatie beschikbaar was, is dit de best haalbare beschrijving van een verwijderingsefficiëntie.

2.3.2 Toestand

Er bestaan drie verschillende zuurstofcondities in de gebouwde database. Om beter inzicht te krijgen welke parameters belangrijk zijn in elk van deze condities, wordt voor elk van deze condities een apart model gebouwd. Allereerst zijn er oxische condities – aanwezigheid van zuurstof in het systeem. Anoxische condities zijn systemen waar geen zuurstof aanwezig is. Een specifiek geval van anoxische condities in deze studie zijn methanogene condities: een anoxische omgeving die zo gereduceerd is dat voornamelijk methanogene processen domineren. Methanogene processen zijn microbiologische processen waarbij methaan als afbraakproduct wordt gevormd tijdens de afbraak van stoffen. Methaanvorming is dominant onder specifieke redoxcondities en onder de aanwezigheid en concentraties van specifieke electrondonoren en acceptoren.

2.3.3 Enkele-experimentvergelijkingen

Omdat in de tekstdata een grote variëteit tussen de experimentele condities bestond, werden ook modellen gebouwd waar de verwijderingsefficiëntie tussen verschillende stoffen in hetzelfde experiment met elkaar werden vergeleken. Hiervoor werden de experimenten van 1 studie gebruikt, zodat omstandigheden vergelijkbaar zijn (Falås *et al., 20*16), en mogelijk scherpere conclusies kunnen worden getrokken voor deze experimentele condities.

2.4 Resultaten Tekstmining

De zoekstring leverde 4749 abstracts op. Bij AS review bleek dat het algoritme voor prioritering van abstracts zo krachtig is dat de selectie zich te snel toespitste op een enkele stof(groep). Daarom moest het aantal verschillende abstracts met het label ‘goed’ dat bij aanvang toegevoegd wordt vergroot worden. Na prioritering in AS review bleven er 3079 abstracts over. Niet alle abstracts hoefden hiervoor gelezen te worden, wat een tijdswinst opleverde. In 673 van de vervolgens gedownloade volledige artikelen (full tekst) zaten ‘triplets’ die aan de selectiecriteria (zie Figuur 5) voldeden. In totaal werden er 2872 triplets geëxtraheerd. Na inspectie bleek dat niet alle triplets van goede kwaliteit waren. Voor sommige triplets was niet duidelijk over welke stof het ging, en/of welke omstandigheid. Zie Tabel 1 voor enkele voorbeelden van gevonden triplets met variabele interpreteerbaarheid. De context die aanwezig was in de verschillende triplets per artikel was niet altijd voldoende (zie Figuur 5). Ook werd (mogelijk meer gestructureerde) informatie in tabellen met deze aanpak niet geëxtraheerd. Het bleek derhalve onmogelijk om op grote schaal informatie direct en automatisch uit de triplets over te nemen in een database, omdat de triplets zeer verschillend waren in taal- en woordgebruik en context nodig was uit de rest van het artikel.

1626	trip	the removal efficiency	was decreased to	to around 80 % at 40 h hrt	28330191	Degradatic
1627	trip	the removal efficiency	was decreased to	to around 50 %	28330191	Degradatic
1628	trip	the effect of peptone on the removal rate of 4-cp	studied	the effect of peptone on the degradation of 4-cp in	28330191	Degradatic
1629	trip	the bioreactor	showed	99 % removal of 20 mg/l of initial 4-cp present in the influ	28330191	Degradatic
1630	trip	the bioreactor	showed	an increase in removal efficiency from 60 to 76 % in the pı	28330191	Degradatic
1631	trip	table 1	summarizes	the effect of peptone on the removal rate of 4-cp by the r	28330191	Degradatic
1632	trip	the bioreactor	showed	greater than 99 % removal efficiency	28330191	Degradatic
1633	trip	the bioreactor	showed	greater than 99 % removal of 4-cp at 24 h hrt	28330191	Degradatic
1634	trip	kargi and konya (2007)	had shown	the removal of 4-cp up to 800 mg/l of loading rate with 9(	28330191	Degradatic
1635	trip	the bioreactor	had achieved	99.8 % removal for higher loading rate of 400 mg/l/day	28330191	Degradatic
1636	trip	5effect of loading rate on the volumetric removal	using	different bioreactorsbioreactorcompoundconcentration (	28330191	Degradatic
1637	trip	the bioreactor	showed	an excellent removal efficiency for 4-cp throughout the o	28330191	Degradatic

Figuur 5. Triplets uit een artikel dat afbraak beschrijft in een bioreactor van stof 4-cp onder verschillende omstandigheden.

Tabel 1. Voorbeelden van triplets met informatie over afbraak van stoffen in de zuivering.

Interpreteerbaarheid	*Zelfstandig naamwoordzin*	*Werkwoordzin*	*Zelfstandig naamwoordzin*
Goed	*a 95 % removal of dinoseb (2-sec-buthyl-2,6-dinitrophenol) in methanogenic conditions*	*was obtained from*	*from a soil*
Goed	*that triphenyl ester opes*	*can be successfully removed*	*(>70 %) by anaerobic degradation in stps*
Redelijk	*the optimum removal of total phosphorus*	*was*	*obtained at 25 % amc content*
Redelijk	*36.5–99.9 % pah reduction levels*	*were*	*achieved under the composting conditions with reduction efficiency*
Slecht	*that chryseomonas luteola (42 % removal) and pseudomonas aeruginosa (40.5 % removal)*	*had*	*the ability*
Slecht	*the removal efficiencies*	*dropped*	*to 53.2 %*

De triplets zijn uiteindelijk gebruikt als alternatieve samenvatting van de inhoud van het artikel rond afbraak, om artikelen aan te wijzen met een diverse groep stoffen die vervolgens handmatig verwerkt werden om de database te vullen.

Om een weloverwogen selectie van artikelen te verkrijgen voor handmatige extractie, is ook de chemische naam van de stof(fen) waar het artikel over gaat geëxtraheerd uit de titel of abstract. Daarnaast zijn deze stoffen geclusterd op basis van moleculaire eigenschappen. Dit is voornamelijk gedaan om verschillende stoffen en groepen van stoffen te selecteren uit de literatuur, om een zo breed mogelijk scala aan verschillende soorten stoffen mee te nemen in de handmatige extractie voor de database.

Uiteindelijk is er een bestand gecreëerd waar de meest relevante artikelen zijn samengebracht, met de volgende informatie:

1. PMCID (PubMed central ID nummer);
2. PMID (PubMed ID nummer);
3. DOI (‘Digital Object Identifier’, de permanente link naar het artikel);
4. Triplet informatie
5. Title (Titel);
6. Abstract (Abstract);
7. asreview_ranking (Plaats in de ranking van de tool AS-review);
8. Chemical-names (Herkende chemische namen);
9. Chemical-cids (PubChem's ID voor chemische stof),;

10. Chemical-names-abst;
11. Chemical-cids-abst;
12. cids;
13. cluster;

Uiteindelijk zijn er 40 hoogkwalitatieve artikelen uit de literatuur-selectie gekozen waaruit handmatig de data geëxtraheerd is, waarbij zoveel mogelijk verschillende stoffen meegenomen zijn. Een groter aantal artikelen was niet mogelijk in verband met de grote hoeveelheid tijd die het kost om manueel deze data te extraheren.

2.5 QSAR-model met betrekking tot stofeigenschappen op verzamelde data

In deze sectie worden de resultaten van de drie verschillende type modellen besproken.

2.5.1 Overkoepelende modellen gefit op doelvariabele

Allereerst werden drie verschillende modellen gebouwd voor de drie verschillende doelvariabelen. Dit resulteerde in modellen met R^2-waarde zoals in Tabel 2.

Tabel 2. Overzicht modelkwaliteiten waar alle tekstdata wordt gebruikt, gegeven dat het de juiste doelvariabele bevat.

Doelvariabele	R^2-waarde test/train	Aantal waarnemingen in databestand	Parameters met hoogste relevantie
Getabelleerde rate constants	R^2-value training set: -13,13 R^2-value testing set: -0,11	105 waarnemingen	Modelkwaliteit te laag
Getabelleerde halfwaardetijden	R^2-value training set: 0,00 R^2-value testing set: 0,11	78 waarnemingen	Modelkwaliteit te laag
Gefitte rate constants	R^2-value training set: 0,20 R^2-value testing set: 0,19	391 waarnemingen	ATSC6s, GATS1s, Estate_VSA1

Bij alle drie de modellen wordt een R^2-waarde gevonden van lager dan 0.2 voor de testset-voorspellingen. Dit geeft aan dat de voorspellende waarde van het model erg gelimiteerd is, en dat het model slechts weinig voorspellende waarde heeft gevonden in de gehele dataset. De negatieve R^2-waarden betekenen dat het model minder goede voorspellingen doet dan een model dat het gemiddelde van de dataset voorspelt. Dit is voornamelijk te verklaren doordat de modellen alle data meenemen. De omstandigheden bepalen of een stof wél of niet verwijderd wordt, en omdat alle data op een hoop wordt gegooid, wordt er in dit model weinig rekening met omstandigheden gehouden. Daarnaast kan het zijn dat er een grote onzekerheid bestaat in de doelvariabelen, waardoor de biofysische werkelijkheid en de gemeten data uiteen loopt. Tenslotte is het aantal waarnemingen relatief klein, en een verbreding van de dataset zou lage R^2-waarden kunnen verbeteren.

Voor de beste van de drie modellen werden ook de parameters met de hoogste relevantie geïdentificeerd. Dit waren de ATSC6s, de GATS1s, en de EState_VSA1. Dit zijn drie vrij abstracte parameters, waarvan het moeilijk te interpreteren is in hoeverre deze waarden betekenis hebben voor de biotische verwijdering. De interpretatie is te vinden in de volgende referenties: Moreau *et al.*, 1980; Anselin *et al.,* 2018; Lapute *et al.,* 2000.

2.5.2 Specifieke modellen voor methanogene, aerobe en anaerobe processen

De totale dataset werd opgesplitst in negen verschillende groepen, tweemaal een verdeling in drie: (i) in getabelleerde verwijderingssnelheden, getabelleerde halfwaardetijden, en gefitte verwijderingssnelheden (zoals in de vorige paragraaf); en (ii) in oxische, anoxische en methanogene condities. Dit leidt tot negen verschillende modellen, waarvan de resultaten in Tabel 3 te vinden zijn.

Tabel 3. Overzicht van de resultaten van verschillende modellen gesplitst in doelparameter en chemische condities.

	Oxische condities	Anoxische condities	Methanogene condities
Getabelleerde rate constants	R^2-waarde testset: -0,27 Weinig waarnemingen	R^2- waarde testset: 0,63 72 waarnemingen	R^2- waarde testset: -0,01 Weinig waarnemingen
Getabelleerde halfwaardetijden	R^2-waarde testset: -1,14 Te weinig waarnemingen	R^2- waarde testset: -0,24 Weinig waarnemingen	Geen waarnemingen
Gefitte rate constants	R^2- waarde testset: 0,19 173 waarnemingen	R^2- waarde testset: 0,34 181 waarnemingen	R^2- waarde testset: 0,26 46 waarnemingen

In deze tabel komen enkele modellen naar voren die redelijke of soms goede voorspellingen kunnen doen. In het geval van de getabelleerde rate constants en anoxische condities is de R^2-waarde 0.63, wat een redelijk goede voorspellende waarde is. De gevonden parameters met hoogste relevantie voor dit model zijn weer vrij abstracte parameters: de MATS5v, de AATSC5v en de MID_N. De eerste twee parameters zijn autocorrelatieparameters, en de MID_N is het moleculaire ID op de stikstof-atomen (een “vorm”-parameter van stikstofgroepen in het molecuul – een parameter gebaseerd in grafentheorie). (Yap *et al.,* 2010) Andere modellen met R^2-waarden boven 0.2 werden ook nader geanalyseerd, en ook hier werden voornamelijk abstracte parameters gevonden. Deze zijn de ATSC6 en GATS1s voor gefitte verwijderingssnelheid/oxisch en ATSC6dv, MATS4dv voor gefitte verwijderingssnelheid/methanogeen. Daarnaast werden voor gepubliceerde verwijderingssnelheid/anoxisch veel abstracte parameters met lage relevantie gevonden. Hoewel dit abstracte parameters zijn, kunnen deze wel degelijk een voorspellende waarde hebben voor biotische verwijdering. Echter, vanwege de eerder genoemde redenen kunnen deze parameters ook naar voren zijn gekomen door andere oorzaken die minder te maken hebben met biotische verwijdering.

2.5.3 Enkel-experimentvergelijking

In deze paragraaf worden modellen beschreven die gebouwd zijn aan de hand van een dataset beschreven in Falås *et al., 20*16. Hier werd de verwijderingsefficiëntie van een reeks OMV’s gemeten waar alle stoffen in een mengsel aanwezig zijn. Het eerste model werd gebouwd op basis van de selectie met de volgende condities: mengsel van OMV’s, full-scale plant, continu proces, oxische condities, geactiveerd slib. De doelvariabele hier was de gefitte verwijderingssnelheid. Dit resulteerde in 59 waarnemingen en hieruit kwam een R^2-waarde op de validatieset (25% van de training data afgesplitst gedurende de hyperparameter-search stap) van 0.25, en op de testset van 0.38. De parameter met de hoogste relevantie hier was de AMID_O-parameter, die de gemiddelde molecular-ID parameter van de zuurstofgroepen op het molecuul zijn, en verder kwamen voornamelijk autocorrelatieparameters naar boven.

Een volgende selectie had als criteria: mengsel van OMV’s, continu proces, anoxische condities, geactiveerd slib. Het aantal waarnemingen was 53, de R^2-validatie was 0.32 en de R^2-test was 0.34. Belangrijkste parameters waren de MID_O (molecular ID zuurstofgroepen), de ATSC2i en AATSC2i (autocorrelatorwaarden) en SssO (Kier-Hall som van ss-zuurstoftypen, Hall *et al.,*1995).

Bij dit laatste experiment werd nabehandeld met een anoxische nabehandeling. Daarom runnen we de bovenstaande selectie nogmaals, maar met gefitte verwijderingssnelheden gevonden na de anoxische nabehandeling. De resultaten van dit model zijn te interpreteren als een antwoord op de vraag: “Welke chemische descriptoren beschrijven stoffen waarvoor geldt dat verwijdering gebaat is bij een anoxische nabehandeling?”. De R^2- validatie was 0.59 en de R^2-test was 0.70. Veruit de belangrijkste parameter hier was de BCUTc-1l, de gewogen waarde van de Burden matrix. Dit wordt uitgelegd in de referenties van Pearlman *et al.*, 1999, Burden *et al.*, 1989 en Burden *et al.*, 1997.

Tenslotte werd nog een model gebouwd aan de hand van de data geselecteerd op basis van de parameters: OMV mengsel, continu proces, anoxische standalone reactor, co-metabolische afbraak, en geactiveerd slib. Het aantal waarnemingen was 29, de R^2- validatie was 0.38 en de R^2-test is 0.91. De parameters met hoogste output waren ATSC7se, ATSC7pe (beide autocorrelatorparameters) en AXp-1d, een waarde die de sigma-electronenconfiguratie in het molecuul kwantificeert.

Door het filteren van de data op vergelijkbaarheid in het experiment kregen we een betere onderlinge vergelijkbaarheid van de parameters in de data. Dit komt doordat deze kleinere datasets geen data bevat waar de verwijdering door andere processen bewerkstelligd kon zijn, en dat alle verwijdering in deze dataset in dezelfde omgeving gebeurt. Echter betekent het ook dat de lage hoeveelheid waarnemingen ervoor zorgt dat de betrouwbaarheid van het model vermindert en mogelijk ervoor zorgt dat het model door toeval overfit. De voorspellingen zijn in het algemeen wel goed, en daarom is het interessant om dieper in de verschillende chemische descriptoren te duiken om te kijken in hoeverre hun invloed ook door fysisch of biochemisch begrip verklaard kan worden.

3 Discussie en aanbevelingen

Dit werk heeft als doel om antwoorden te bieden op vragen met betrekking tot het gebruik van machinelearning en QSAR-methoden: Kunnen voorspellingen gedaan worden over de microbiële afbraak van verschillende chemicaliën onder verschillende omstandigheden met behulp van deze methoden? En zo ja, welke methoden tonen het meeste potentieel?

3.1 Tekstmining

Voor de dataverzameling is, bij gebrek aan een bestaande database met deze data, naar wetenschappelijke literatuur gekeken. In dit project werd een systematische aanpak gevolgd om relevante literatuur te verzamelen met behulp van een goed ontworpen 'PubMed'-zoekreeks. Dit werd gevolgd door het gebruik van 'AS Review' voor selectie, het maken van triplets en het clusteren van stoffen op basis van moleculaire eigenschappen voor een uitgebreide selectie met diverse stoffen. De resulterende database bleek van waarde bij de initiële screening van artikelen voor handmatige data-extractie, waarbij een gevarieerde reeks stoffen en groepen uit literatuur met essentiële informatie werd verzameld. Bovendien is deze literatuurselectie zelf mogelijk een waardevolle database, die eenvoudig bijgewerkt kan worden met recente literatuur door de stappen van dit onderzoek te herhalen. Deze systematische aanpak bespaart aanzienlijke tijd, voornamelijk door het automatisch filteren van literatuur die cruciale informatie bevat, wat normaal pas na het lezen ervan duidelijk wordt. Bovendien zorgt het voor objectiviteit in de initiële selectie, wat toekomstige projecten ten goede kan komen. Volledige automatisering van een op informatie gebaseerde database blijft echter momenteel een uitdaging en vereist handmatige tussenkomst. In Paragraaf 3.2 wordt ingegaan op opties om dataextractie te verbeteren.

In een ander project, 402045/233 'Impact van zeer zorgwekkende stoffen in het milieu' (in voorbereiding) werden 'Large language models' (LLM's) gebruikt om de snelheidsconstanten van chemische reacties te vinden in abstracts. LLM's konden hierbij nauwkeurig deze informatie vinden. Het gebruik van LLM's voor het scannen van feiten uit wetenschappelijke artikelen is een alternatief ten opzichte van de methode die in dit huidige onderzoek is gebruikt. Maar, er moet worden opgemerkt dat er kosten en tijd zijn verbonden aan het 'trainen' van LLM's om de goede informatie te vinden, zie de details in project - 402045/233 (in voorbereiding). Bovendien is de uitdaging voor data extractie in het huidige onderzoek groter omdat er meerdere omstandigheden en getallen verzameld en juist gekoppeld moeten worden vanuit een compleet artikel, in plaats van een enkele snelheidsconstante aan een chemische reactie uit een abstract.

Feitelijke informatie verkrijgen uit literatuur is tijdrovend, en de beschikbaarheid van data kan ontoereikend zijn. Het gebruik van Large Language Models (LLMs) zoals ChatGPT zou kunnen helpen om efficiënt informatie te extraheren uit teksten, waardoor het verzamelen van data wordt vergemakkelijkt. Echter, uitdagingen met betrekking tot de toegankelijkheid van artikelen en veiligheidszorgen in verband met LLMs vragen om verder onderzoek.

3.2 Verbeteren van datakwaliteit en -rapportage

3.2.1 Andere experimentele opzet

De gevonden data in de literatuur over afbraak van stoffen was onderling niet goed vergelijkbaar omdat er veel verschillen tussen uitgevoerde experimenten waren. Als in verschillende artikelen afbraaksnelheden of percentages voor dezelfde stof werd gegeven was dit onder zeer verschillende omstandigheden. Bijvoorbeeld 'full-scale' versus

'batch' experimenten, een andere matrix of concentratie, wel of geen herhaalde toevoeging van stoffen, een andere pH, etc. Ook ontbrak er juist vaak informatie over deze omstandigheden. Het feit dat er veelal verschillende experimentduur werd gehanteerd is opgelost door een afbraaksnelheid te berekenen vanuit gegeven afbraakpercentages. Maar, het is een feit dat de biologische afbraak zeer complex is en in grote mate afhangt van omstandigheden en de microbiële gemeenschap, inclusief hoe lang deze zich heeft kunnen specialiseren op de chemische vervuiling. Aanwezigheid van andere soorten nutriënten en omstandigheden bepalen in hoeverre deze specialisatie kan plaatsvinden. Voor het maken van voorspellingsmodellen is behoefte aan meer grootschalige experimenten met veel verschillende stoffen, onder gecontroleerde omstandigheden. Om de omstandigheden te kunnen vergelijken tussen experimenten zouden deze gestandaardiseerd moeten worden, waar mogelijk.

3.2.2 Minimale dataset

Het opstellen van een 'minimale informatieset' voor biologische afbraakgegevens wordt aanbevolen, met vermelding van types, eenheden en formaten om gestructureerde inzichten te vergemakkelijken. Auteurs die onderzoek publiceren over biotische afbraak zouden idealiter een minimale dataset moeten opnemen met cruciale experimentele parameters. Bovendien kan het aanpassen van data-rapportageformaten om geautomatiseerde tekstextractie te vergemakkelijken gunstig zijn. Gestandaardiseerde sjablonen of formulieren voor publicaties die machine-leesbare informatie bevorderen, vergelijkbaar met 'Nanopublicaties', zouden data-toegankelijkheid voor meta-analyses kunnen verbeteren. Ook kan er grote winst gehaald worden wanneer auteurs alle ruwe data als supplement of zelfstandig datapakket in een publieke archief uploaden, zodat deze data gebruikt kan worden door derden voor meta-analyses.

Een voorstel voor een minimale dataset houdt naar onze mening de volgende parameters in:

1. Start en eindconcentratie stof(fen): wat was de concentratie van de stof die is toegevoegd/gemeten aan het begin en einde van het experiment?

2. Incubatietijd: hoeveel tijd zat er tussen de metingen van de stof(fen) tijdens het experiment? Hoelang was het experiment geïncubeerd?

3. (Redox) condities: wat was de conditie waarin het experiment plaatsgevonden heeft (anoxisch/oxisch/methanogeen/nitrificerend)?

4. Operationele parameters: temperatuur, HRT, SRT, pH, redox potentiaal.

5. Biomassa/Matrix: Wat voor biomassa of matrix was gebruikt in het experiment (actief slib, reincultuur bacteriën, etc.)?

3.2.3 Aanbevelingen uit het NWO CER-CEC project

Een universitair NWO onderzoeksprogramma liep parallel aan dit project, genaamd 'Cost-effective removal of contaminants of emerging concern in urban waste water treatment plants" (CER-CEC). In dit project bestudeerden onderzoekers o.a. de voorspelbaarheid via modellen van verwijdering van opkomende stoffen in afvalwaterzuivering. Dit is een meerjarig onderzoek van 3 PhD studenten geweest, en dit onderzoek is in vergelijking met het hier beschreven onderzoek veel dieper ingegaan op details. In het CER-CEC onderzoek werd ook geconcludeerd dat het moeilijk is om een goede voorspelling te doen met de op dit moment beschikbare data. Een reviewartikel van Rios-Miguel et al., 2023 schetst aanbevelingen, zoals relevante parameters en langetermijnstrategieën om de ontwikkeling van QSAR-modellen in het veld te stroomlijnen.

De auteurs bevelen ook aan om bestaande open repository databases aan te vullen met meer metadata, standaardisatie van OMV-verwijderingsmetingen, stroomlijnen van microbiologische sequencingmethoden en

identificatie van cruciale microbiologische mechanismen. Deze alomvattende aanpak zou bestaande modellen kunnen valideren en betere modellen mogelijk maken.

3.3 Modellen voor de voorspelling van afbraak

Met behulp van de afbraakdatabase zijn modellen geconstrueerd met verschillende QSAR-methoden om zuiveringsrendementen voor ongeziene stoffen en omstandigheden te voorspellen. Deze modellen maakten het mogelijk om conclusies te trekken over belangrijke parameters voor microbiologische afbraak.

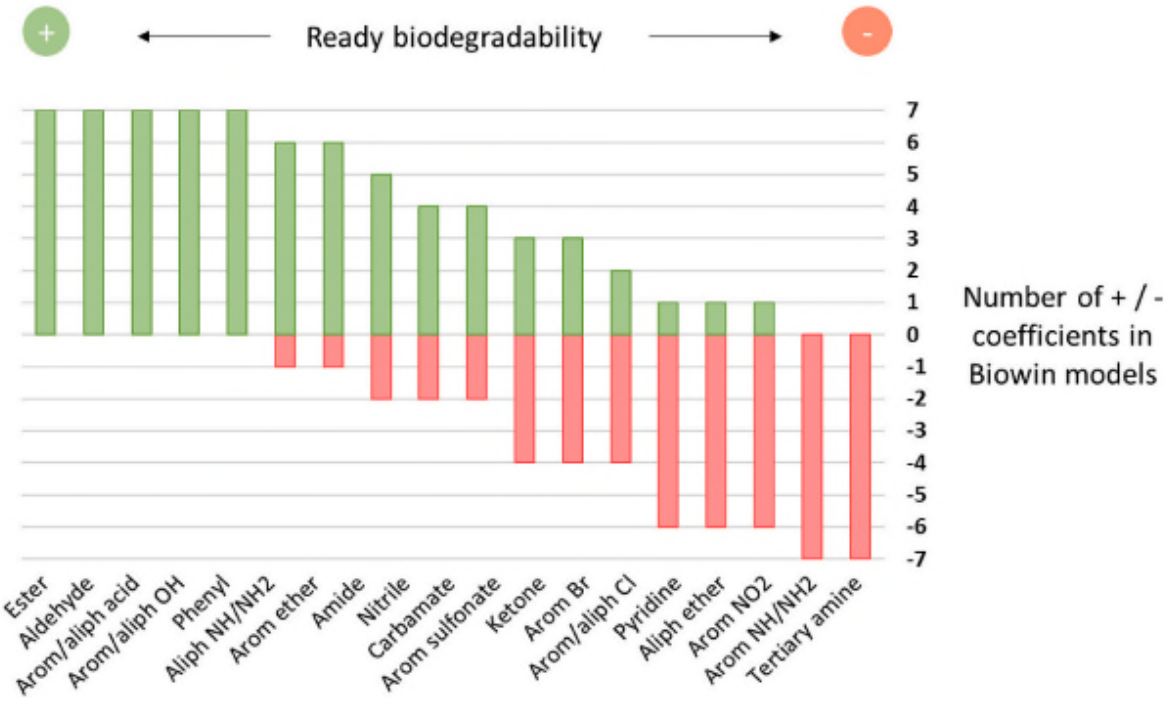
Na toepassing van de bovengenoemde techniek kwamen verschillende conclusies naar voren. Ten eerste ontbreken de verkregen modellen vaak aan hoge nauwkeurigheid in voorspellingen als gevolg van verschillen tussen datapunten uit experimenten tussen studies. Het standaardiseren van experimentele beschrijvingen zou de vergelijkbaarheid kunnen vergroten en daarmee de nauwkeurigheid van QSAR-modellen verbeteren. De nadruk op een minimale set informatie in gepubliceerd onderzoek, zoals benadrukt in Paragraaf 3.2, verdient overweging.

Gebrek aan data blijft een uitdaging. In dit onderzoek leidde dit tot twee hoofdproblemen: modellen die gebaseerd zijn op beperkte waarnemingen zijn mogelijk niet representatief voor het volledige parameterspectrum, en voorspellende precisie lijdt hieronder. Dit kan worden aangepakt door een parameterselectiemethode vergelijkbaar met dit onderzoek toe te passen met bootstrapping voor betrouwbaardere schattingen. Bovendien kunnen permutatietesten dienen als een extra controle om de betrouwbaarheid van modellen te beoordelen.

De toegepaste QSAR-methode steunt op berekende parameters door Python-pakketten zoals RDKit en Mordred. Hoewel deze parameters waardevolle moleculaire informatie bieden, is de relevantie ervan voor voorspellende modellering onderwerp van nader onderzoek. Het lijkt dat mechanistische modellen vooralsnog de voorkeur hebben zolang data beperkt is.

Overweging om meer complexe machinelearningmodellen, zoals neurale netwerken, naast traditionele Gradient Boosting Regressor (GBR) modellen te gebruiken, zou kunnen leiden tot verbeterde voorspellingen. Neurale netwerken (in staat tot zogenaamde ‘Deep Learning’) excelleren in het vastleggen van gecombineerde invloeden van inputparameters en kunnen meerdere doelparameters tegelijk voorspellen, mogelijk beter dan de GBR-modellen in dit werk.

In het project CER-CEC is gekeken naar de chemische groepen in stoffen die biodegradabiliteit positief of negatief beïnvloeden (Rios-Miguel et al., 2023). Hierbij is gebruik gemaakt van de Biowin modellen (US EPA, 2023).



Figuur 6: Indicatie welke chemische groepen biodegradabiliteit positief of negatief beïnvloeden. Berekend door het BIOWin model, aangepast van Rios-Miguel et al. 2023.

Een overzicht van de functionele groepen van stoffen die een positieve of negatieve invloed hebben op biotische degradatie is te vinden in Figuur 6. Dit werd bepaald door Biowin te laten berekenen welke stoffen sneller of langzamer werden afgebroken onder oxische condities (‘ready biodegradability’), en dus niet onder andere condities die in de waterzuivering aanwezig zijn. De opbrengsten kunnen wel helpen met vervolgonderzoek: het begrijpen van chemische degradatiepaden is cruciaal. Hier kan BioTransformer ook een goede toevoeging aan bieden: door te combineren wat te weten is over de afbraakproducten en welke chemicaliën goed afbreken, kan een goed begrip opgebouwd worden welke reactiepaden vaak voorkomen. Daarnaast is het een duidelijke indicatie en overzicht van wat men kan verwachten voor biotische afbraak van chemicaliën in het algemeen.

Het verkennen van tools zoals Biowin en BioTransformer voor datasetuitbreiding en als input voor QSAR-modellen kan gunstig zijn. Echter, vanwege beperkingen met betrekking tot het in beschouwing nemen van specifieke omstandigheden en anoxische afbraak (denitrificerende, methanogene, sulfaat-reducerende, ijzer-reducerende condities, etc.) kunnen ze de hoofdvraag van dit onderzoek, de afbraak onder verschillende omstandigheden, niet beantwoorden.

4 Conclusies

Uit het hierboven beschreven onderzoek kunnen de volgende conclusies getrokken worden:

1) Voor het maken van betrouwbare en goed voorspellende modellen voor biologische afbraak onder verschillende omstandigheden is veel data nodig. Er is data te vinden in wetenschappelijke literatuur. Deze data is niet uniform, onvolledig en onvoldoende vergelijkbaar tussen studies. Dat bemoeilijkt de extractie van deze data, en de benodigde integratie. Dit betekent dat er op dit moment geen goede modellen gemaakt kunnen worden rond de voorspelling van OMV afbraak onder verschillende omstandigheden in de zuivering. Nieuwe grootschalige, gestandaardiseerde experimenten en rapportage met een vaste minimale set aan informatie rond biologische afbraak kan dit probleem opheffen. Een toegewijde repository kan deze data bij elkaar brengen.

2) Er zijn onvoldoende betrouwbare resultaten gegenereerd op basis van de beperkte dataset die in dit project is samengesteld. Een optie is om eerst de focus te leggen op het begrijpen van biologische afbraak onder gangbare omstandigheden. Sequensen van het DNA van microbiële gemeenschappen om deze te karakteriseren en het mechanistisch beschrijven per type afbraakproces via microbiële afbraakmechanismen zijn daarbij logische opties.

3) De hier toegepaste tekstminingbenadering is door handmatige stappen vooralsnog te arbeidsintensief om op te schalen. LLM's kunnen mogelijk een rol spelen om dit probleem op te lossen. De hier toegepaste machine learning techniek zou vervangen kunnen worden door meer geavanceerde ‘Deep Learning’ techniek.

5 Referenties

Anselin, L. (2018). A Local Indicator of Multivariate Spatial Association: Extending Geary's c. Geographical Analysis 51(2), 133-150. https://doi.org/10.1111/gean.12164

Aronson, D., et al. (2006). Estimating biodegradation half-lives for use in chemical screening. Chemosphere 63(11): 1953-1960. https://doi.org/10.1016/j.chemosphere.2005.09.044

Benner, J., D. E. Helbling, H.-P. E. Kohler, J. Wittebol, E. Kaiser, C. Prasse, T. A. Ternes, C. N. Albers, J. Aamand, B. Horemans, D. Springael, E. Walravens and N. Boon (2013). Is biological treatment a viable alternative for micropollutant removal in drinking water treatment processes? Water Research 47(16): 5955-5976. https://doi.org/10.1016/j.watres.2013.07.015

Bertelkamp, Reungoat et al., 2014, Bertelkamp C., J. Reungoat, E.R. Cornelissen, N. Singhale, J. Reynissonf , A.J. Cabog , J.P. van der Hoek, A.R.D. Verliefde. (2014). Sorption and biodegradation of organic micropollutants during river bank filtration: A laboratory column study. Water Research 52: 231-214. https://doi.org/10.1016/j.watres.2013.10.068

Burden, F. R. (1989). Molecular identification number for substructure searches. Journal of Chemistry and Informational and Computational Science, 29(3): 225-227. https://pubs.acs.org/doi/abs/10.1021/ci00063a011

Burden, F. R. (1997). A Chemically Intuitive Molecular Index Based on the Eigenvalues of a Modified Adjacency Matrix. Journal of Chemistry and Informational and Computational Science, 16(4): 301-314. https://doi.org/10.1002/qsar.19970160406

D'Alessio, M., B. Yoneyama, M. Kirs, V. Kisand and C. Ray. (2015). Pharmaceutically active compounds: Their removal during slow sand filtration and their impact on slow sand filtration bacterial removal. Science of The Total Environment 524-525: 124-135. https://doi.org/10.1016/j.scitotenv.2015.04.014

Djoumbou Feunang Y., Fiamoncini J., de la Fuente A.G., Manach C., Greiner R., Wishart D.S. (2019). BioTransformer: A Comprehensive Computational Tool for Small Molecule Metabolism Prediction and Metabolite Identification; Journal of Cheminformatics 11:2; https://doi.org/10.1186/s13321-018-0324-5

EC (2003) Technical guidance document in support of Commission directive 93/67/EEC on risk assessment of new notified substances and Commission Regulation (EC) no. 1488/94 on risk assessment for existing chemicals and Directive 98/8/EC of the European Parliament and of the Council concerning the placing of biocidal products on the market. Office for Official Publications of the European Communities, Luxembourg, Luxembourg. https://op.europa.eu/en/publication-detail/-/publication/ef333513-33a4-4c1e-a9b6-c17fa054b586

Ecetoc (2013) Environmental exposure assessment of ionisable organic compounds. Technical report 123. https://www.ecetoc.org/technical-report-123/estimated-partitioning-property-data/computational-methods/biodegradation/

Falås, P., A. Wick, S. Castronovo, J. Habermacher, T. A. Ternes and A. Joss. (2016). Tracing the limits of organic micropollutant removal in biological wastewater treatment. Water Res 95: 240-249. https://doi.org/10.1016/j.watres.2016.03.009

Fate of emerging and priority micropollutants during the sewage sludge treatment – Part 2: Mass balances of organic contaminants on sludge treatments are challenging. Waste Management 125, 122-131, https://doi.org/10.1016/j.wasman.2021.02.034

Gonzalez-Gil L., Daniel Krah, Ann-Kathrin Ghattas, Marta Carballa, Arne Wick, Lissa Helmholz, Juan M. Lema, Thomas A. Ternes. (2019). Biotransformation of organic micropollutants by anaerobic sludge enzymes, Water Research 152, 202-214. https://doi.org/10.1016/j.watres.2018.12.064

Hall, L. H., Kier, L. B. (1995). Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. Journal of Chemical Informational Computational Science, 35(6): 1039-1045. https://doi.org/10.1021/ci00028a014

Hedegaard, M. J. and H.-J. Albrechtsen (2014). Microbial pesticide removal in rapid sand filters for drinking water treatment – Potential and kinetics. Water Research 48: 71-81. https://doi.org/10.1016/j.watres.2013.09.024

Hijnen, W., Hofman-Caris, R., Bertelkamp, C. (2016). Pyrazool - inventarisatie full-scale data en verkennend experimenteel onderzoek. BTO 2016.203(s), KWR Water Research Institute, Nieuwegein, The Netherlands.

Hofman-Caris, C. H. M. and B. A. Wols. (2020b). Voorspelling en validatie van de verwijdering van organische microverontreinigingen uit water; deel 2: oxidatieve processen, BTO 2020.063, KWR Water Research Institute, Nieuwegein, The Netherlands.

Hofman-Caris, C. H. M., B. A. Wols, D. Vries, M. Korevaar and W. Siegers. (2020a). Voorspelling en validatie van de verwijdering van organische microverontreinigingen uit water; deel 1: stofselectie, BTO 2020.056, KWR Water Research Institute, Nieuwegein.

Hofman-Caris, C. H. M., B. A. Wols, D. Vries, M. Korevaar, D. Harmsen and E.R. Cornelissen. (2020c). Voorspelling en validatie van de verwijdering van organische microverontreinigingen uit water; deel 3: membraanprocessen, BTO 2020.066, KWR Water Research Institute, Nieuwegein, The Netherlands.

Hofman-Caris, C. H. M., B. A. Wols, D. Vries, M. Korevaar, W. Siegers. (2021). Voorspelling en validatie van de verwijdering van organische microverontreinigingen uit water; deel 4: filtratie over aktieve kool, BTO 2021.050, KWR Water Research Institute, Nieuwegein, The Netherlands.

Hofman-Caris, C. H. M., Sattler, D., Classen, D. (2020d). Persistence of gabapentin, 1Hbenzotriazole, diglyme, DTPA, 1,4-dioxane, melamine and urotropin in surface water; testing of chemicals according to the OECD 309 guideline. KWR 2020.118, KWR Water Research institute, Nieuwegein, The Netherlands

Hollender, J., J. Rothardt, D. Radny, M. Loos, J. Epting, P. Huggenberger, P. Borer and H. Singer. (2018). Comprehensive micropollutant screening using LC-HRMS/MS at three riverbank filtration sites to assess natural attenuation and potential implications for human health. Water Research X 1: 100007. https://doi.org/10.1016/j.wroa.2018.100007

Lapute, P. (2000). A widely applicable set of descriptors. Journal of Molecular Graphical Modelling 18(4-5): 464-477. https://doi.org/10.1016/S1093-3263(00)00068-1

Mansouri, K., et al. (2018). OPERA models for predicting physicochemical properties and environmental fate endpoints. Journal of cheminformatics 10(1): 1-19. https://doi.org/10.1186/s13321-018-0263-1

Moreau, G.; Broto, P. (1980). The Autocorrelation of a Topological Structure: a New Molecular Descriptor. Nouveau Journal de Chimie Françoise, 4(6): 359-360. http://pascal-francis.inist.fr/vibad/index.php?action=getRecordDetail&idt=PASCAL8040406871

Moriwaki, H., Tian, Y.-S., Kawashita, N., Takagi, T. (2018). Mordred: a molecular descriptor calculator. Journal of Cheminformatics, 10: 4. https://doi.org/10.1186/s13321-018-0258-y

OECD (2004). The Report from the Expert Group on (Quantitative) Structure-Activity Relationships [(Q) SARs] on the Principles for the Validation of (Q) SARs. OECD SERIES ON TESTING AND ASSESSMENT number 49. https://one.oecd.org/document/ENV/JM/MONO(2004)24/En/pdf

Patureau D., R. Mailler, N. Delgenes, A. Danel, E. Vulliet, S. Deshayes, R. Moilleron, V. Rocher, J. Gasperi. (2021). Fate of emerging and priority micropollutants during the sewage sludge treatment - Part 2: Mass balances of organic contaminants on sludge treatments are challenging. Waste Manag, 125, 122-131. https://doi.org/10.1016/j.wasman.2021.02.034

Pavan, M. and A. P. Worth. (2008). Review of estimation models for biodegradation. QSAR & Combinatorial Science 27(1): 32-40. https://doi.org/10.1002/qsar.200710117

Pearlman, R. S., Smith, K. M. (1999). Metric Validation and the Receptor-Relevant Subspace Concept. Journal of Chemistry and Informational and Computational Science, 39(1), 28-35. https://doi.org/10.1021/ci980137x

Pronk, T.E., Fischer, A., van den Berg, A.E.T., Hofman, C.H.M. (2023) Prioritization of micropollutants based on removal effort in drinking water purification treatment. Water Quality Research Journal 58 (3): 184–198. https://doi.org/10.2166/wqrj.2023.032

Rattier, M., J. Reungoat, W. Gernjak, A. Joss and J. Keller. (2012). Investigating the role of adsorption and biodegradation in the removal of organic micropollutants during biological activated carbon filtration of treated wastewater. Journal of Water Reuse and Desalination 2(3): 127-139. https://doi.org/10.2166/wrd.2012.012

Rios-Miguel, A. B., van Bergen T. J. H. M., Zillien, C., Ragas, A. M. J., van Zelm, R., Jetten, M. S. M., Hendriks, A. J., Welte, C. U. (2023). Predicting and improving the microbial removal of organic micropollutants during wastewater treatment: A review. Chemosphere, 333: 138908. https://doi.org/10.1016/j.chemosphere.2023.138908

Shimabuku, K. K., T. L. Zearley, K. S. Dowdell and R. S. Summers (2019). Biodegradation and attenuation of MIB and 2,4-D in drinking water biologically active sand and activated carbon filters. Environmental Science: Water Research & Technology 5(5): 849-860. https://pubs.rsc.org/en/content/articlelanding/2019/ew/c9ew00054b

Straub, J.O., Le Roux, J.,Tedoldi, D. (2022) Are newer pharmaceuticals more recalcitrant to removal in wastewater treatment? Sustainable Chemistry and Pharmacy 30: 100834. https://doi.org/10.1016/j.scp.2022.100834

Struijs J. (2014) SimpleTreat 4.0: a model to predict fate and emission of chemicals in wastewater treatment plants: Background report describing the equations. RIVM Report 601353005. RIVM, Bilthoven The Netherlands. https://www.rivm.nl/publicaties/simpletreat-40-a-model-to-predict-fate-and-emission-of-chemicals-in-wastewater

Timmers P.H.A., T. Slootweg, A. Knezev, M. van der Schans, L. Zandvliet, A. Reus, D. Vughs, L. Heijnen, T. Knol, J. El Majjaoui, P. van der Wielen, P.J. Stuyfzand, K. Lekkerkerker-Teunissen (2022). Improved drinking water quality after adding advanced oxidation for organic micropollutant removal to pretreatment of river water undergoing dune infiltration near The Hague, Netherlands, Journal of Hazardous Materials, Volume 429, https://doi.org/10.1016/j.jhazmat.2022.128346

Timmers P.H.A., Siegers W., Ferreira M.L., van der Wielen P.W.J.J. (2023). Bioremediation of rapid sand filters for removal of organic micropollutants during drinking water production. Water Res, 249, 120921 https://doi.org/10.1016/j.watres.2023.120921

US EPA (2023) Estimation Programs Interface Suite™ for Microsoft® Windows, v 4.11 or insert version used]. United States Environmental Protection Agency, Washington, DC, USA

Van de Grift B. Timmers P.H.A. (2021). Verwijdering van OMV's tijdens bodempassage in het infiltratiebekken De Lange Vlieter. BTO 2020.206(s), KWR Water Research institute, Nieuwegein, The Netherlands.

Vries D., Wols B., Korevaar M.W. & Vonk E. (2017) AquaPriori: a priori het verwijderingsrendement bepalen. KWR 2017.027 https://www.kwrwater.nl/projecten/aquapriori/ KWR Water Research institute, Nieuwegein, The Netherlands.

Wang, J., D. de Ridder, A. van der Wal and N. B. Sutton (2021). Harnessing biodegradation potential of rapid sand filtration for organic micropollutant removal from drinking water: A review. Critical Reviews in Environmental Science and Technology 51(18): 2086-2118. https://doi.org/10.1080/10643389.2020.1771888

Xingfeng Y., Deling F., Wen G., Jining L., Lili S., Zhi Z., Linjun Z., Guixiang J. (2021) Aerobic and Anaerobic Biodegradability of Organophosphates in Activated Sludge Derived From Kitchen Garbage Biomass and Agricultural Residues. Front. Bioeng. Biotechnol. 9. https://doi.org/10.3389/fbioe.2021.649049

Yap, C. W. (2010). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. Computational Chemistry, 32:1466-1474. https://doi.org/10.1002/jcc.21707

Zearley, T. L. and R. S. Summers. (2012). Removal of Trace Organic Micropollutants by Drinking Water Biological Filters. Environmental Science & Technology 46(17): 9412-9419. https://doi.org/10.1021/es301428e

Zhou, J., D. Wang, F. Ju, W. Hu, J. Liang, Y. Bai, H. Liu and J. Qu (2022). Profiling microbial removal of micropollutants in sand filters: Biotransformation pathways and associated bacteria. Journal of Hazardous Materials 423: 127167. https://doi.org/10.1016/j.jhazmat.2021.127167

I Bijlage I: search string

Search string gebruikt voor een PubMed search:

1. biodegradation

biodegradation OR bioremediation OR “microbiological degradation” OR “microbiological consumption” OR “microbial consumption” OR “microbial remediation” OR “microbial degradation” OR (microorganism* AND degradation) OR bioaugmentation OR “biological remediation” OR “biological degradation” OR biotransformation

2. chemicals

micropollutants OR micropollutant OR “organic micropollutant” OR “organic micropollutants” OR OMP OR OMPs OR “Emerging micropollutants” OR EMP OR EMPs OR pharmaceuticals OR pharmaceutical OR “personal care products” OR PPCPs OR “pharmaceutical active compounds ” OR PhAC OR PhACs OR “detergents” OR “steroid hormones” OR “industrial chemicals” OR pesticides OR herbicides OR “organic contaminants” OR “organic contaminant”

3. water

wastewater OR “drinking water” OR groundwater OR “surface water” OR “water treatment” OR “water purification”

4. bacterial culture

((Bacterial OR bacteria OR microorganism OR microbial OR microbe OR microorganisms OR axenic OR enrichment) AND (Isolate OR strain OR culture)

5. Experiment

“Batch experiment” OR Bioreactor OR Column OR Mesocosm OR Mesocosms OR “Lab experiment” OR “Laboratory experiment” OR Lab-scale OR “laboratory scale” OR Pilot-scale OR Sludge OR “Sand filter” OR slurry

6. metabolism

Metabolism OR metabolic OR co-metabolism OR co-metabolic OR catabolic OR catabolism OR nitrification OR denitrification OR iron-reduction OR iron-oxidation OR respiration OR respiring OR “nitrate reduction” OR “ammonium oxidation” OR “sulfate reduction”

7. conditions/processes

oxic OR anoxic OR anaerobic OR aerobic OR nitrifying OR denitrifying OR sulfate-reducing OR iron-reducing OR iron-oxidizing OR “nitrate reducing” OR “ammonia oxidizing”