Masteroppgåve i kjemi

Tema: Kjemometri

# Kvalitet av MS-spektra og LC-MS data

## *Wavelet-vidareutviklingar av komponent-deteksjons-algoritmen og nye kvalitetsmål*

Nicolai Sikveland


Kjemisk institutt

Universitetet i Bergen

Oktober 2015

# Forord

Takk til rettleiar Bjørn Grung for god rettleiing.

Takk til Ph.D.-kandidat Kjersti Hasle Enerstvedt for lån av ZM2-datasettet, til Ph.D. Irene Skaar for lån av HCO-datasettet og til Ph.D.-kandidat Yingxu Zeng for lån av TAG-datasettet.

Takk til overingeniør Bjarte Holmelid for hjelp med LC-MS og konvertering av datasett.

Takk til fyrsteamanuensis Svein Are Mjøs for hjelp med konvertering av datasett.

Takk til Professor Tor Sørevik for hjelp til å forstå wavelet-emnet.

Takk til mine medstudentar for hyggelege pausar. Av medstudentane vil eg spesielt takka Håvard G. Frøysa for gode råd.

Takk til mine foreldre for all støtte.

Takk til Gud for kraft og innsyn.

## Samandrag

I dette arbeidet er CODA-algoritmen vidareutvikla ved bruk av  ulike wavelettransformasjon (WT) -algoritmer. Eit universelt kvalitetsmål for LC-MS eksisterer ikkje. Derfor vart det utarbeida tre kvalitetsmål basert på singulærverdiar til datasettet, normforhold mellom toppar frå same forbindelse og korrelasjon mellom toppar tilhøyrande same forbindelse.

## P.1 Notasjon

### P.1.1 Generell notasjon for datasett i dokumentet

*Fylgjande notasjonar vert nytta for å enkelt kunna referera til dei ulike resultata i tekst og figurar.*

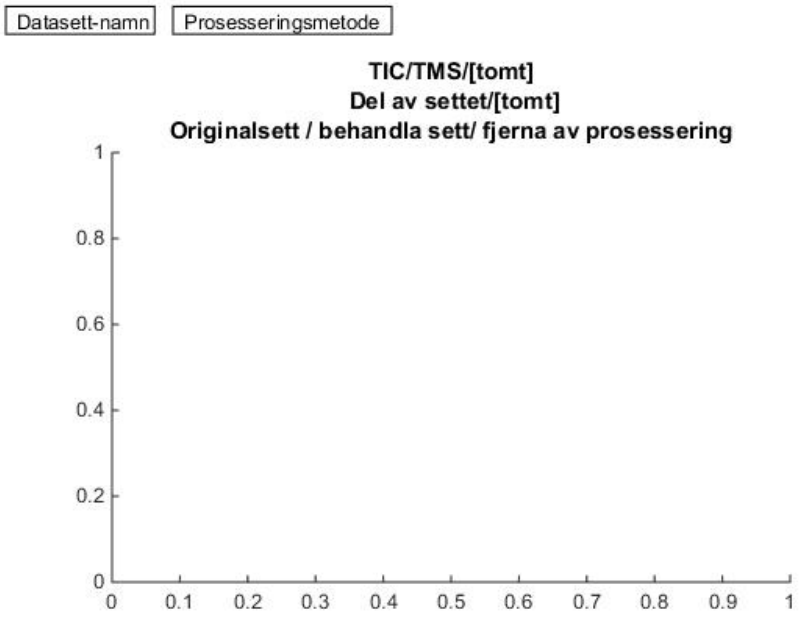
*Notasjonen er lik for alle datasetta, men det vert i tillegg informert om kva datasett det gjeld.*

*Fylgjande datasett-symbol gjeld data på 2D-format, som seinare vert introdusert i kapittel 2.1*

«Fjerna av WT og CODA» vil sei alt som vert forkasta, for ein metode som utførar både WT og CODA som ein del av prosesseringa.
«Fjerna av CODA» referer berre til CODA-delen av algoritmen, men vil for ein metode som ikkje nyttar WT på dei prosesserte data, utgjera alt som vert forkasta.

### P.1.2 Generell notasjon for figurar i dokumentet

*Figur P.1-1 (nedanfor)* viser korleis mange av figurane i resultata er merka. Ved plotting av datasett vil datasettnamnet vera i ein boks i øvre venstre hjørne av figuren. Kva metode som er nytta til prosessering vil vera oppgjeve, i ein boks, til høgre for datasett-namnet. Generell notasjon for datasetta vil også verta nytta i figurane, for å signalisera kva for eit prosess-steg ein ser på. Den generelle notasjonen vil stå som ein tittel over kvart plott, referera til kva ein plottar, og kan bestå av fleire ledd. Fyrste ledd kan f.eks. vera «TMS» (totalt massespekter). Ledd nr. to kan referera til ein spesiell del av eit sett, f.eks. «kvalitetsområde av ….» . Det siste leddet i den generelle notasjonen er av type; «Behaldt av CODA» eller «Originalt datasett», etc. Totalt vert då den generelle notasjonen; «TMS kvalitetsområde behaldt av CODA»




*Figur P.1-1: Eksempelfigur for namngjeving av datasett*

# Innhald

Forord ..... i

Samandrag ..... ii

P.1 Notasjon ..... iii

P.1.1 Generell notasjon for datasett i dokumentet ..... iii

P.1.2 Generell notasjon for figurar i dokumentet ..... iii

1 Innleiing ..... 1

1.1 Bakgrunn og motivasjon ..... 1

1.2 Oversikt over støyfjerningsmetodar ..... 1

1.3 Kvalitetsmål ..... 5

1.4 Målsetjing ..... 5

2 Teori ..... 6

2.1 LC-MS ..... 6

2.1.1 Kromatografi ..... 6

2.1.2 Væskekromatografi (LC) ..... 6

2.1.3 Massespektrometri (MS) ..... 8

2.1.4 LC-MS ..... 13

2.1.5 Støy i LC-MS ..... 15

2.2 Komponent-deteksjon-algoritmen (CODA) ..... 16

2.2.1 Intervallbasert CODA ..... 18

2.3 Kvalitetsmål i LC-MS ..... 19

2.4 Matematiske eigenskapar og transformasjonar ..... 21

2.4.1 Eigenverdiar, singulærverdiar og rang ..... 21

2.4.2 Basisar for matriser og funksjonar ..... 22

2.4.3 Wavelettransformasjon (WT) ..... 23

3 Verktøy, datasett og eksperimentelt ..... 35

3.1 Datamaskin og programvare ..... 35
3.1.1 Datamaskin-spesifikasjonar ..... 35
3.1.2 Programvare ..... 35
3.2 Datasett og datasettbehandling ..... 35
3.2.1 Datasett ..... 35
3.3 MATLAB-programmering ..... 47
4 Fordeling av massekromatografiske kvalitetsindeksar (mcq-verdiar) ..... 50
4.1 Mcq-fordeling i ideelle sett og reelle sett ..... 50
4.2 Mcq-fordeling i wavelettransformerte sett ..... 54
5 Nye CODA-versjonar ..... 60
5.1 CODA_ndWT ..... 61
5.2 CODA_CWT ..... 61
5.3 CODA_WPT ..... 62
5.4 CODA_ WPT2 ..... 63
5.5 CODA_WPTlim ..... 63
5.6 CODA_SHD_ndWT ..... 64
6 Nye kvalitetsmål ..... 65
6.1 Singulærverdi-forhold ..... 65
6.2 Toppsamanlikning ..... 66
6.2.1 Korrelasjonsbasert kvalitetsmål ..... 67
6.2.2 Normbasert kvalitetsmål ..... 68
7 Testing og diskusjon ..... 69
7.1 Validering av CODA-versjonane og fastsetjing av ideelle parameterar ..... 69
7.1.1 CODA_ndWT ..... 72
7.1.2 CODA_CWT ..... 82
7.1.3 CODA_WPT ..... 91

7.1.4 CODA_WPT2 ..... 95
7.1.5 CODA_WPTlim ..... 97
7.1.6 CODA_SHD_ndWT ..... 101
7.2 Testing av kvalitetsmål ..... 104
7.2.1 Singulærverdi-forhold - metoden ..... 104
7.2.2 Toppsamanlikning-metodane ..... 106
7.3 Samanlikning av CODA-versjonar ..... 106
7.3.1 Samanlikning av TIC ..... 106
7.3.2 Singulærverdi-metode ..... 113
7.3.3 Toppsamanlikning-baserte kvalitetsmål ..... 115
7.4 Totalsamanlikning ..... 118
8 Konklusjon ..... 120
9 Vidare arbeid ..... 121
10 Referansar ..... 122
11 Vedlegg ..... 125
11.1 Vedleggsoversikt ..... 125

# 1 Innleiing

## 1.1 Bakgrunn og motivasjon

Kjemiske analysar er i dag ein viktig del samfunnet. Dei er blant anna viktige i miljø og klimaspørsmål, medisinsk diagnostikk og produktutvikling innan ymse industriar som farmasøytisk industri, næringsmiddel-industri, petroleumsindustri, og mykje meir. I alle typar kjemisk analyse er ein opptatt av kvalitet, altså sikkerheit i målingane. Kvaliteten kan for eksempel vera relatert til nøyktighet ved kvantifisering av kjente forbindelsar eller til sikkerheit ved deteksjon av ukjente forbindelsar. Ein ting som er sikkert er at kvalitet er viktig. Ofte må ein også avgjera kva som er god kvalitet, ved å velja vekk noko. I eit deteksjonstilfellet må ein avgjer om ein helst vil risikera å ikkje detektera komponenten, eller om ein vil risikera å få ein falsk deteksjon.

LC-MS (væskekromatografi – massespektrometri) er ein analysemetode der høgt støy-nivå er eit problem. Støy er signal som ikkje har opphav i analytten, og kjem av eksempelvis elektrisk støy eller kolonnebløding. Det er ynskjeleg å fjerna støyen, og det er derfor utvikla ei rekke forskjellige støyfjerningsmetodar. Blant desse teknikkane er CODA (Component Detection Algoritm)[1], WMSM (Windowed Mass Selection Method) [2] og ulike filter-teknikkar som MF (Matched Filtration) [3, 4], GSD (Gaussian Second Derivative matched filtration) [4, 5], SG [4, 6], MEND (Matched filtration with Experimental Noise Determination) [4, 7], median filtrering [8] og ein metode introdusert av Capadona et al. basert modellering av støy vha. wavelettransformasjon [9]. Av desse metodane skiljar CODA seg ut ved at den ikkje gjer endringar i kromatogramma, men heller forkastar støyfulle kromatogram. Det har også vorte lansert fleire forslag til forbetringar av CODA-algoritmen, bla. CODA_DW Windig [10] og Sandve [11] sine intervall-oppdelings versjonar.

## 1.2 Oversikt over støyfjerningsmetodar

CODA-algoritmen sin funksjon er avhengig av to inputparameterar; ein vindaugebreidd-parameter og ei kvalitetsgrense. CODA reknar ut ein kvalitetsverdi for kvart kromatogram, som er der den beste verdien er 1 og dårlegaste er 0. Kvalitetsverdien er eit mål på likskap mellom kromatogrammet og eit glatta og sentrert kromatogram. Algoritmen fjernar alle kromatogram med kvalitet under den definerte grensa.

CODA_DW er ei vidareutvikling av CODA, som fungerer betre ved baselinje-støy i enn den original-algoritmen. Reknar ut ein dw-verdi (Durbin Watson) for kvart kromatogram, som er eit mål på «tilfeldigheit» i fordeling av data. Denne verdien er den normaliserte summen av kvadrerte verdiar for den deriverte. Den deriverte vert rekna ut som differansen mellom nabopunkt. Dw-verdien er negativt korrelatert med kvalitet for kromatogrammet, ettersom stor differanse mellom nabopunkt er karakteristisk for støy (dvs. tilfeldig signal). I CODA_DW nyttar ein dermed ei øvre dw-grense for å skilja ut mindre gode kromatogram.

CODA-versjonen CODA DER tek ein enkel derivasjon av data (som også vert gjort i DW) før utveljing av kromatogram. Versjonen er elles lik til den originale algoritmen. Derivasjonen vert gjort for å gjera CODA mindre sensitiv for bakgrunnsstøy, for å inkludera massekromatogram som inneheld både baselinje og analytisk signal. For å gjera både CODA_DW og CODA DER endå meir sensitive for desse massekromatogramma, er det foreslått å opphøga datasetta i andre eller tredje potens, ved utrekning av mcq-verdiane.

Li et al. [12] har utvikla ein støyfjerningsmetode for LC-MS-data med namn RNIE (reduce noise information entropy), som er basert på informasjonsentropi og er designa for å fjerna støy. Metoden fungerer på same måte som CODA ved at den veljar ut høgkvalitetsmassar.

Sandve [11] utvikla i si masteroppgåve fleire nye utviklingar av CODA. Den mest lovande algoritmen var, i fylgje Sandve, CODA slicehalvdyn. Tanken bak Sandve si algoritme er at ein masse kan innhalda berre støy i eit kromatografisk intervall, og signal i eit anna. Viss ein då delar opp datasettet i intervall og køyrar CODA separat på desse, vil ein kunne fjerna meir støy og/eller behalda meir signal. For å unngå kutting av signal-toppar, som kan skje dersom dei vert plassert mellom intervall, vert det berre fjerna massar frå eit mindre intervall i midten av det store intervallet. Algoritmen bevegar seg då framover med steg på storleik med det minste intervallet, slik at dei nærliggande store intervalla overlappar. Ved køyring av denne typen algoritme vil ein ikkje kunne fjerna dei forkasta del-massane, ettersom dette ville ført

til ei ufullstendig datamatrise. Forkasta verdiar vert derfor nullsett, og ferdig prosessert data vil ha same storleik som original-settet.

I Cappadona et al. [9] sin metode for støyfjerning i LC-MS data, vert det nytta wavelet-transformasjon for å karakterisera støy i kromatogramma. Metoden vart laga med fokus på å karakterisera kjemisk og stokastisk støy og fjerna denne, slik at signaltoppar med intensitet under «støygrensa» kan identifiserast. Wavelet-transformasjon (WT) er ein matematisk transformasjon som byggjar på wavelets, små lokale bølgjefunksjonar, som basisfunksjonar. Det vert fokusert på eigenskapane som udesimert diskret WT har til å dela opp eit signal i høgfrekvente og lågfrekvente ledd. WT-metoden som vert nytta er rekursiv over fleire nivå. WT startar ved å gå frå kromatogram-vektoren til å dela den opp i eit høgfrekvent ledd (d1) og eit lågfrekvent ledd (a1), der desse til saman utgjer nivå 1. Nivå 2 vert oppnådd ved å utføra WT på a1, slik at dette vert delt opp i eit høgfrekvent ledd (d2) og eit lågfrekvent ledd (a2). Nivå 3 vert då danna ved å utføra WT på a2, osv. Ein kan også gå tilbake til førre nivå ved å inverstransformera (iWT) a-leddet og d-leddet. Støyfjerningsmetoden vert basert på at mesteparten av den stokastiske støyen sit i d1-leddet, og at baselinja kan approksimerast ved a6-leddet. Metoden estimerar det stokastiske støynivået som median absolutt avvik (MAD) av d1-leddet, og deretter vert alle verdiar i d1 under støy-verdien nullsett. (Ved MAD = 0 vert støyen modellert som ei N(0;1) fordeling.) iWT vert så nytta på det uendra a1-leddet og det støybehandla d1-leddet for å få eit kromatogram utan stokastisk støy. Baselinja vert approksimert som a6-leddet til det originale kromatogrammet. Ein konstruerer så det ferdig prosesserte settet ved å interpolera mellom felles punkt i dei to støybehandla kromatogramma med eit delvis kubisk hermite polynom.

WMSM (Windowed Mass Selection Method) [2] er ein metode som fjernar støy frå kromatogramma med å sletta dei delane av kromatogramma som ein antar er støy. Dei to stega i algoritmen er å fyrst fjerna tilfeldig støy og så fjerna bakgrunnsstøy. Algoritmen nyttar to bevegelige vindauge i prosessering av kromatogramma. Brukaren veljar eit lite vindauga (w1) og eit stort (w2), der det minste definerar kor smal ein signal-topp kan vera og det store definerar kor brei ein signal-topp kan vera. Dersom ein topp er mindre enn det minste

vindauga (w1) eller større enn det største vindauga (w2) vert den rekna som støy, og fylgjeleg nullsett. I praksis kan ein sei at det minste vindauga (w1) vert brukt til fjerning av tilfeldig støy (smale toppar) og det største (w2) til fjerning av bakgrunnsstøy. Algoritmen går gjennom kromatogrammet i oppdelte delar (del-intervall), definert av vindauge-storleikane. For fjerning av tilfeldig støy er delintervalla overlappande, og forflyttar seg med eit kromatografisk punkt om gongen. Bakgrunnsfjerninga føregår på ikkje-overlappande, nærliggjande, intervall.

MF, GSD og SG er alle digitale filter, som virker ved at dei bytter ut kvar verdi med ei vekting av dei nærliggjande kromatografiske verdiane– og dermed glattar ut kromatogramma. Det vert sett på ein vektor med gjevne punktet som senterpunkt. Vektinga vert utført ved å ta indreproduktet av vektoren og eit sett med filterkoeffisientar. Det er viktig at koeffisientane liknar ein kromatografisk topp i fordeleling av intensitetar, og ikkje støy. [4] MF har koeffisientar ut frå fordelinga til ein gaussisk funksjon. GSD nyttar koeffisientane ut frå ein andrederivert gaussisk funksjon. SG utførar ein polynomdivisjon over vektoren, og set verdien for det midtre punktet i polynomdivisjonen som nytt punkt midtpunkt i vektoren. [4]

Hastings et al. [8] har introdusert ein metode kalla median-filtrering. Median-filtreringa går ut på at eit bevegeleg median-filter-vindauga går gjennom kromatogrammet. Kvar punkt vert sett lik medianen av vindauget den er ein del av. Effekten av filtreringa er ei glatting i kromatografisk retning, og dermed reduksjon av stokastisk støy.

MEND-metoden [7] finn ein støykarakteristikk ved bruk av ”tomme” massekromatogram, som igjen vert brukt til å finna transferfunksjonar H. Desse funksjonane vert nytta for å utføra massefiltrering (MF - støyfjerningsteknikk) av datasettet. Det vert så utført ”peak picking” ved bruk av score’ar for kvar topp (peak) som vert bestemt av nærliggjande verdiar i både masse- og tids-domena. Til slutt vert også kjente uønska addukt og ion fjerna.

## 1.3 Kvalitetsmål

Når ein nyttar støyfjerningsmetodar er det viktig å kunna vurdera prosesseringa. Det er derfor naudsynt med kvalitetsmål for kva som er eit bra kromatogram. Slike kvalitetsmål må også vera konstruert ut ifrå kva ein ynskjer å behalda i det prosesserte analyse-settet. Eit ynskje kan vera å fjerna all støy frå datasettet, medan eit anna kan vera å behalda alt signal. Eksempel på kvalitetsmål som vert nytta er signal til støy-forhold (S/N), COMPARELCMS_SIM [10, 13] informasjonsentropi i datasett [14] eller at kjente signal i analyseprøven er til stades i spektra/massekromatogramma.

## 1.4 Målsetjing

I dette arbeidet vert det fokusert på å fastsetja skånsame mcq-verdiar til CODA-algoritmen. Skånsame mcq-verdiar er verdiar som er låge nok til å ikkje fjerna signal frå datasetta. Ved å kombinera tankegangen frå wavelet-karakterisering av støy med CODA kan det tenkast at ein kan fastsetja ei mcq-grense som berre fjernar støy. Denne mcq-grensa vil då variera ut frå korleis støyen i eit gjeve kromatogram ser ut. Fokuset for arbeidet vert sett på å behalda alt signal i datasettet, framfor å fjerna all støy.

Det vart sett to mål for arbeidet:

1. Å utvikla nye wavelet-baserte CODA-versjonar for bruk i LC-MS –datasett.
2. Finna nye objektive kriterium for kvalitet ved støyfjerning i LC-MS - datasett. Og å testa ut eksisterande / kombinera ?

# 2 Teori

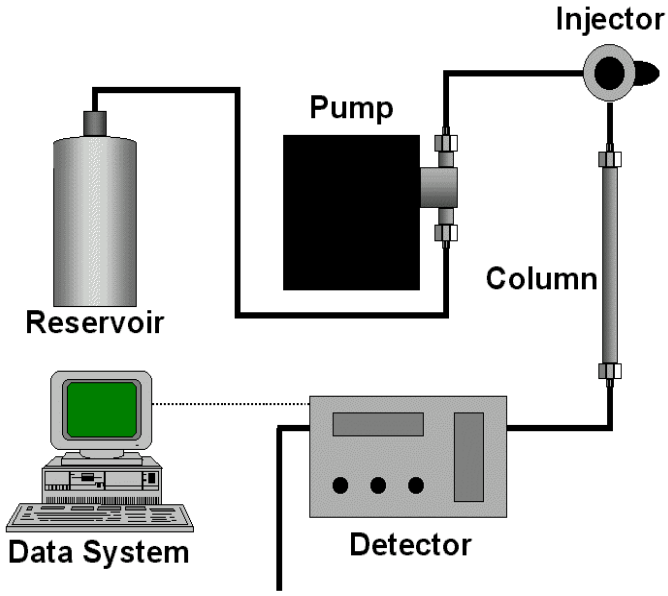
## 2.1 LC-MS

### 2.1.1 Kromatografi

Kromatografi er ein metode for å skilja ut ulike kjemiske komponentar frå ei blanding av fleire komponentar. Separasjonen i kromatografi skjer ved at komponentar vert ført gjennom/forbi ein stasjonær-fase av ein mobil-fase. Komponentane kjem gjennom den stasjonære fasen ved ulik tid, kalla retensjonstid eller elueringstid, ut ifrå kor lenge dei er i den stasjonære fasen i forhold til den mobile fasen. Separasjonmetoden kan verta brukt til analyse, isolasjon og til fastsetjing av fysiske eigenskapar. [15]

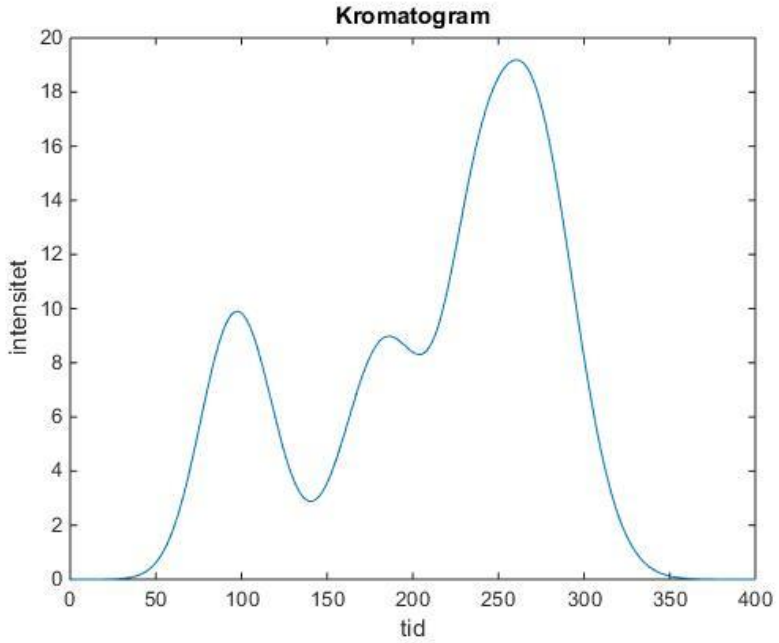
### 2.1.2 Væskekromatografi (LC)

Væskekromatografi, LC (Liquid Chromatography), er ein kromatografisk metode, der den mobile fasen er i væskeform og den stasjonære fasen er i fast- eller væske-form. Væskekromatografi vert ofte utført som kolonne-kromatografi, men vert også utført på plane overflater. Her vil det verta fokusert på High Performance kolonne-LC (HPLC). Eit HPLC-system er skissert i Figur 2.1-1. I HPLC vert det nytta trykk for å få prøven og mobilfasen effektivt gjennom kolonna. Fordi det vert nytta trykk kan kolonna vera tettpakka av små uniforme partiklar som gjer god separasjon. Ved bruk av LC som ein analysemetode, som i HPLC, vert kolonna kopla saman med ein detektor.




*Figur 2.1-1: HPLC-system [16] . HPLC (High Performance LC), ein mykje brukt LC-type. Reservoaret inneheld mobilfasen, og prøven vert ført inn i systemet ved injektoren.*

Ved bruk av detektor som er in-line-kopla til kolonna, som i HPLC, kan ein ta opp eit kromatogram. Kromatogrammet er ei kontinuerleg eller diskret kurve i eit koordinatsystem, med tid som x-akse og detektorutslag som y-akse. I Figur 2.1-2 vert det vist eit eksempel på eit kromatogram.




*Figur 2.1-2: Kromatogram med fleire forbindelsar. Dette kromatogrammet har ikkje fullstendig separerte toppar.*

I normalfase-LC er kolonna (stasjonærfasen) polar og den mobile fasen upolar, og ved reversfase er forholdet omvendt. Reversfase-LC er meir vanleg enn normalfase, men vart introdusert seinare. I LC er det eit stort utval av kolonner og mobilfasar å velja mellom, og desse kan veljast for å passa best mogleg til analyttane.

Mobilfasen kan ha lik samansetnad gjennom heile køyringa av kromatografen, men kan også endra samansetnad, f.eks. ved å nytta ei blanding av to løysemiddel av ulik polaritet og variera forholdet i ein retning. Lik samansetning av mobilfasen gjennom heile køyringa vert kalla isokratisk elusjon, og endra samansetning vert kalla gradient elusjon. Ein fordel ved å nytta gradient elusjon er kortare elusjonstid for ei blanding forbindelsar med ulik affinitet til kolonna, og toppforma til dei seinast eluerande forbindelsane vil verta betre.

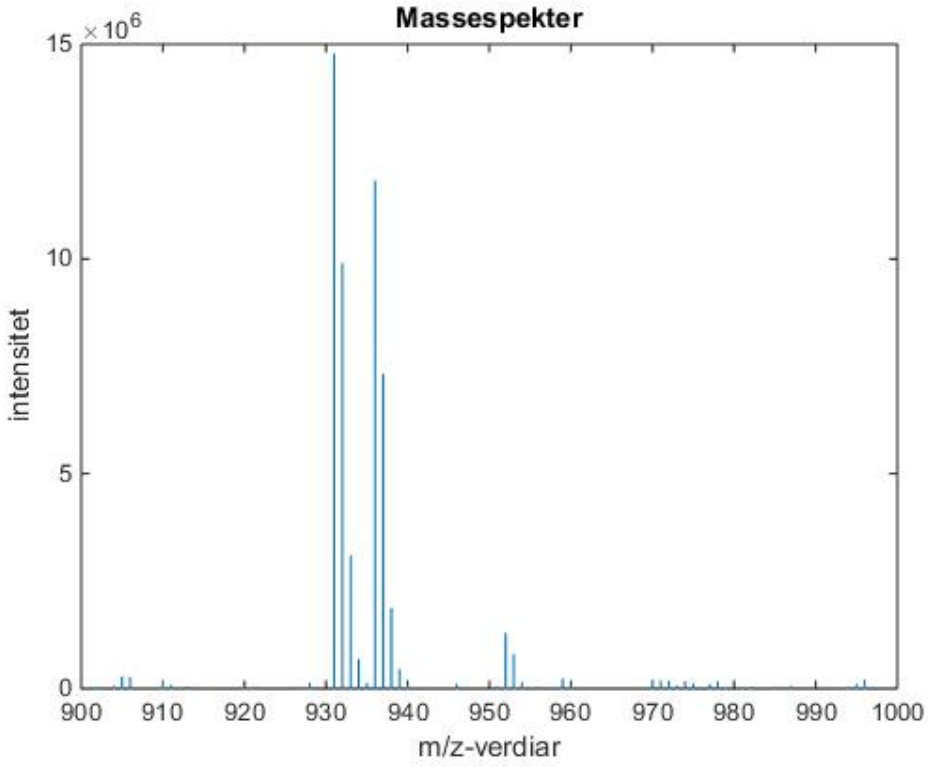
I HPLC er det, som vist i Figur 2.1-1, tilkopla ein detektor som målar eigenskapar ved analytten som kjem ut. Kva slags detektor som vert nytta kan variera. Eksempel på detektor-typar for HPLC er refraksjonsindeks-detektor, elektrokjemisk-detektor, konduktivitet-detektor, fluorisens-detektor, UV-spektroskop (med ei eller fleire bølgjelengder) og massespektrometer. Alle dei nemnde detektorane er vanlege for HPLC, og det er ingen standard-metode, ettersom dei alle har fordelar og ulemper.

[17-19]

### 2.1.3 Massespektrometri (MS)

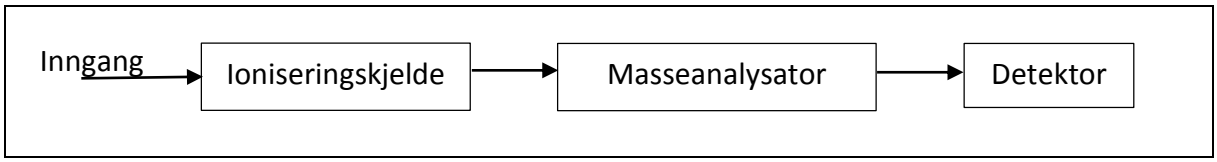
Massespektrometri (MS) er ein spektrometrisk analysemetode, som målar masse per ladning (m/z) av ioniserte molekyl og fragment av desse. Det er vanlig at instrumentet oppgjer intensiteten til den m/z -verdien med høgst førekomst som 100 % og dei andre intensitetane relativt til denne, og at m/z -verdien er molarmassen delt på ladning. Eit massespekter, som vist i Figur 2.1-3, består av m/z-verdiar langs x-aksen og intensitet langs y-aksen. Det framstår derfor oppstykka av rette vertikale linjer (som i eit histogram), kvar for ulike m/z verdiar, og ikkje kontinuerleg som i eit kromatogram. Massespektrometeret gjer analoge målingar, som vil gje eit kontinuerleg signal, men spektera vert sentroidisert innan gjevne

oppløysingsrammer for kvart instrument. Sentroidisering vil sei å diskretisera signala ved å samla signal ved nærmaste definerte diskrete m/z –verdi. Ofte forenklar ein m/z-verdi – omgrepet, ved å anta lik ladning, og snakkar heller om massar.




*Figur 2.1-3: Eksempel på massespekter-struktur. (I dette tilfellet er ikkje intensiteten oppgjeven i prosent.) Kvar linje i spekteret representerer eit ion (anten fragment-ion eller moder-ion), med ein m/z-verdi som grovt sett kan kallast ein masse viss ein antek lik ladning.*

Eit massespektrometer er bygd opp som i Figur 2.1-4 under.




*Figur 2.1-4: Oversikt over eit MS-apparat (Inspirert av figur 1.1 i [20] og figur 10.1 i [17] )*

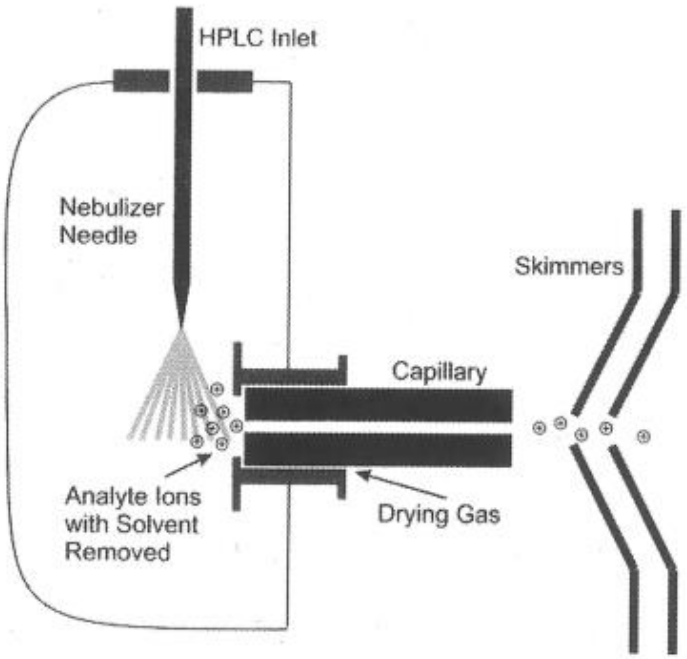
Ioniseringskjelda ioniserar forbindelsane, masseanalysatoren skiljar ion ut frå masse og ladning, og detektoren detekterar iona.

#### *2.1.3.1 Ionisering i MS*

For å måla ein m/z –verdi må analytten fyrst ha ei ladning, noko som ikkje er så vanlig i stabil form. For å oppnå lada ion vert analytten ionisert. Dette skjer vanligvis ved å fjerna eit eller fleire elektron frå "moder-molekylet" - ein oppnår altså ei plusslading. Dess fleire molekyl som vert ioniserte, dess sterkare signal vil ein få, og på denne måten kan ein også måla førekomst av kvart ion.

Det finst mange ulike ioniseringsteknikkar for MS, men ein kan grovt sett dela dei inn i to kategoriar; harde og mjuke. EI (elektron impact) er ein hard ioniseringsteknikk, medan ESI (electron spray ionisation), CI ( chemical ionisation) og APCI (atmospheric pressure chemical ionization) er eksempel på mjukare ioniseringsteknikkar . Dei harde ioniseringsteknikkane ioniserar fleire molekyl, i forhold til totalen, enn dei mjuke teknikkane. Ved bruk av harde ionisasjonsteknikkar vil ein også oppnå ein høgare grad av fragmentering, altså oppdeling av iona til mindre ion pga. ustabilitet, som er eit resultat av den høge energien som vert tilført. Fragmentering kan vera ynskja eller uynskt. Ved å sjå på dei ulike fragment-iona tilhørande eit moder-ion, får ein meir informasjon om strukturen til analytten enn ein får utan fragmentering. Fragmentering er lite gunstig dersom ein er ute etter å finna mengda av ein kjent forbindelse, ettersom det då vil vera mindre av forbindelsen som når detektoren.

ESI-metoden, vist i Figur 2.1-5, skiljar ut allereie ioniserte forbindelsar ifrå ei løysing, som f.eks. forbindelsane løyst i mobilfasen i HPLC. Iona, i løysing, går fyrst gjennom ei nål, som gjer løysinga om til aerosol (fin tåke) . Aerosolen vert utsett for ei sterk spenning og tørkegass, som gjer at iona sklijar seg frå løysemiddelet. Spenningsgradienten i tillegg til ein trykkgradient leiar iona i retning masseanalysatoren, gjennom eit kapillærrøyr, medan tørkegassen går i motsett retning. [20, 21]

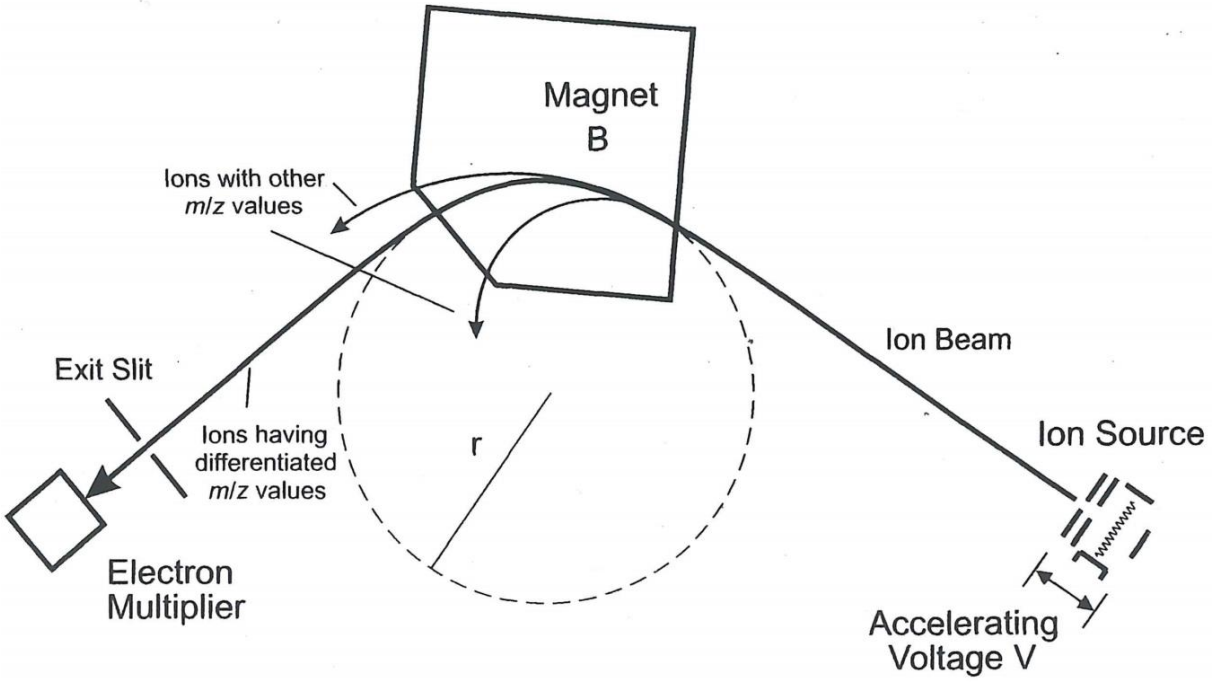



*Figur 2.1-5 : ESI (elektrospray ionisering) ( henta frå frå [20]).*

### *2.1.3.2 Masseanalysator i MS*

m/z-verdiane vert skilde ved at iona går gjennom ein masseanalysator, som også vert kalla eit massefilter. I masseanalysatoren vert ion med ulike m/z-verdiar skilde på bakgrunn av fysiske lovar. m/z-verdien vert funne ved å nytta elektrisitet eller elektriske og/eller magnetiske felt for å måla dei fysiske eigenskapane til iona. Det finnes fleire typar masseanalysatorar, som magnetisk sektor, transmisjons-kvadrupol og TOF (Time-of-Flight). Ein kan også kopla saman fleire masseanalysatorar i eit tandem MS system (MS-MS).

For å illustrera korleis ein masseanalysator fungerer, vert det her sett på ein magnetisk sektor masseanalysator. I eit magnetisk sektor MS (Figur 2.1-6) vert iona akselerert av ei spenning, og går så inn i eit magnetisk felt som står vinkelrett banen til iona. Banen til eit ion vil då vera avhengig av farten det har inn i feltet og den magnetiske krafta som verkar på ionet. Banen er derfor avhengig av masse og ladning til ionet.




*Figur 2.1-6: Magnetisk sektor (henta frå [20]). Akselererte ion kjem ut av ioniseringskjelda, går inn i den magnetiske sektoren. Ion innanfor eit bestemt m/z området treff detektoren.*

I magnetisk sektor-masseanalysatoren i Figur 2.1-6 er det berre ein av dei tre ion-straumane som har rett bane til å nå detektoren. Iona som treff detektoren går i ein bane med radius r, når den er inne i magnetfeltet. Kva slags m/z-verdiar som treff detektoren vert fastesett av Formel 2.1-1 . For denne typen magnetisk sektor kan ein sekvensielt justera feltstyrken, B, og halda spenninga, V, konstant, for å ta opp eit spektrum over eit område av m/z-verdiar. Ein annan type magnetisk sektor analysator held både feltstyrke og spenning konstante og detekterar ion over eit større radius område, og kan då av Formel 2.1-1 avgjera m/z-verdi ut frå kvar iona treff detektoren.

$$m/z = \frac{B^2 r^2}{2V} \qquad \textit{Formel 2.1-1}$$

B er styrken på det magnetiske feltet, r er radius for banen til iona og V er spenninga i ioniseringskjelda.

### *2.1.3.3 Deteksjon i MS*

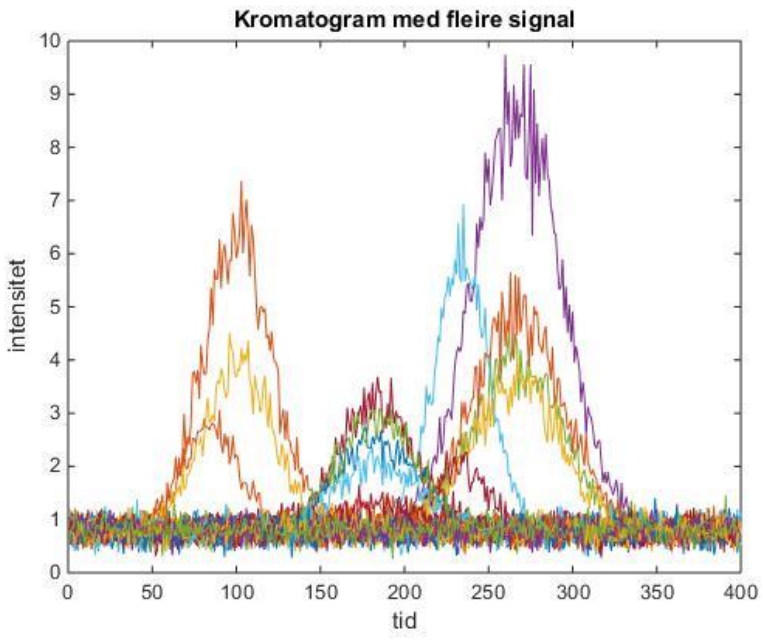
Når iona er gjennom masseanalysatoren gjenstår deteksjonen. Relativt få ion når detektoren, sjølv ved masseanalysatorar som slepp gjennom alle iona, og det er derfor nyttig med

forsterking av signalet. Til signalforsterking vert det nytta multipliseringsdetektorar, der iona kolliderar med ein overflate som enten sendar ut elektron eller foton. Desse partiklane går gjennom n kollisjonar med ein overflate, som gjer $2^n$ elektron eller foton. Signalet vert til slutt målt som straumstyrke eller lysintensitet.

[20, 22, 23]

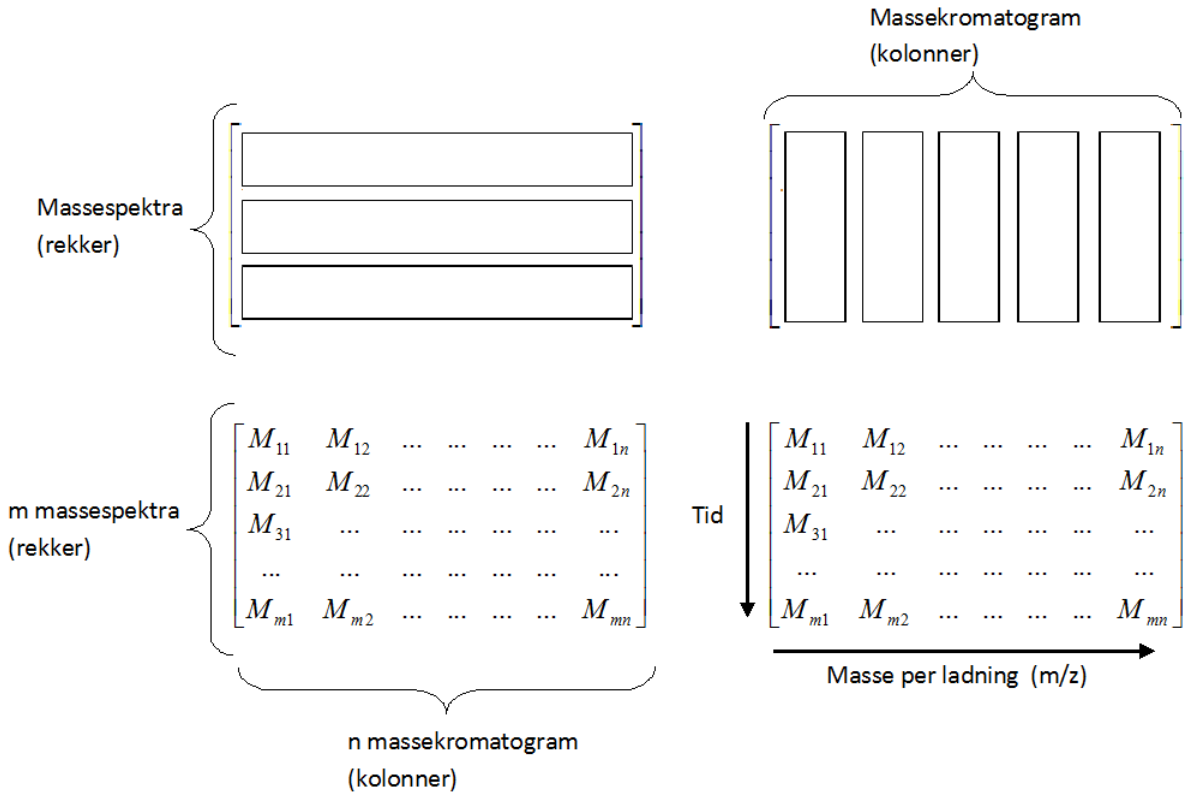
### 2.1.4 LC-MS

LC-MS (væskekromatografi-massespektrometri) er ein mykje brukt analysemetode. LC-MS består av, som namnet indikerar, ein væskekromatograf (LC), vanlegvis HPLC, og eit massespektrometer (MS) kopla saman. Her får MS rolla som detektor for LC (sjå Figur 2.1-1) og er kopla inn etter LC-kolonna. For ei kromatografisk køyring vil ein då få fleire kromatografiske profilar; ein for kvar m/z-verdi, som vert illustrert i Figur 2.1-7. Ein kromatografisk profil som tilhøyrar ein bestemt masse (m/z-verdi) kan ein kalla for eit massekromatogram. Eit massekromatogram kan også verta referert til som eit EIC (Extracted Ion Chromatogram).



*Figur 2.1-7: LC-MS - kromatogram*

Ein får ei datamatrise som resultat, til forskjell frå ein vektor, ettersom det vert teke opp fleire massespekter over tidsperioden det tek for all analytten å gå gjennom LC-kolonna. Analysedata kan lagrast i ei matrise, med tidsaksen vertikalt og masse-aksen horisontalt. Ei [m x n]-datamatrise M består altså av m massespektra (antall punkt på tidsaksen) og n kromatografiske profilar (antall massar) . Datasett-strukturen vert illustrert i Figur 2.1-8.




*Figur 2.1-8: Figuren viser forma til eit LC-MS datasett. Datasettet er ei m x n – matrise, M, der det er m rekker og n kolonner. For kvart av dei m tidspunkta vert det teke opp eit massespekter med n massar.*

For å få eit overblikk over heile datasettet kan ein summera alle massane i kvart massespekter saman, slik ein berre har ein verdi for kvar tidseining i matrisa – ein har ein vektor som er summen av alle kolonnane. Eit slikt kromatogram vert kalla eit TIC (Total Ion Current) - kromatogram. Eit TIC-kromatogram vil ha lik utsjånad som eit kromatogram der detektoren berre gjer ein verdi (intensitet) ved kvart tidspunkt, som i Figur 2.1-2.

For LC-MS er det vanlig å nytta anten ESI ( electrospray ionisation), APCI (atmospheric pressure chemical ionization) eller TSI (thermospray ionization) som ioniseringsteknikk. [17] Her vil fokuset vera på LC-ESI-MS.

#### *2.1.4.1 Datasett-storleik*

Datasett frå LC-MS kan pga. sin todimensjonale struktur, verta relativt store, mtp. lagringskapasitet og dedikert minne, i forhold til mange andre analysedata. Dette gjeld spesielt data frå høgtoppløyselege LC-MS. I slike tilfelle kan det vera nyttig med teknikkar for å redusera storleiken, viss den originale oppløysinga ikkje er naudsam. Reduksjon av storleiken kan gjerast ved å slå saman nærliggjande «desimal-massar», innanfor definerte intervall. Ofte vert massane omgjort til heiltalsmassar. Denne teknikken for samling av masser vert kalla binning (av engelsk).

### 2.1.5 Støy i LC-MS

Eit problem med LC-ESI-MS er støy i kromatogramma. Støy er signal som ikkje har opphav i analytten og den kan verta delt inn i to hovudgrupper; «tilfeldig» (stokastisk) støy og bakgrunnsstøy. Den tilfeldige støyen, kvit støy, er smale toppar i kromatogrammet, som ikkje går over mange einingar på tidsaksen. Desse toppane har ofte sitt opphav i ioniseringskjelda. Bakgrunnsstøy er ofte breie band med støy, som ligg over store deler av kromatogrammet. Banda er vanlegvis mykje breiare enn ein topp med signal. Bakgrunnsstøyen treng ikkje liggja på eit kontinuerleg intensitets-nivå, noko som ville gjort det lettare å fjerna den, men kan variera langs tidsaksen. Opphavet til bakgrunnstøy er typisk kolonneblødning, dvs. mobilfase frå væskekromatografen kjem med i massespektrometeret. [2]

Ein eigenskap som skaper ei todeling innan stokastisk støy, er om støystyrken er avhengig av signalstyrken til analytten. Støy som er avhengig av signalstyrke vert kalla heteroskedastisk, medan motparten vert kalla homoskedastisk.

Det finnes også ulike effektar med opphav i ioniseringskammeret til LC-MS spektrometeret, som oppstår ved høge ionekonsentrasjonar og kan føra til senking eller auke av TIC.

## 2.2 Komponent-deteksjon-algoritmen (CODA)

CODA (component detection algorithm) [1] er ei algoritme som vel ut dei massane (massekromatogramma) som har høgst kvalitet og forkastar dei andre. På denne måten skiljar CODA seg ut frå andre klassiske støyfjerningsalgoritmer, ved at den ikkje endrar på massekromatogramma, men veljar ut kva for nokre som inneheld kjemisk informasjon. Algoritmen fungerer slik at den bereknar ein likskapsindeks, mcq-indeks (mass chromatographic quality indeks), for kvart kromatogram. Mcq-verdien er alltid i eit intervall mellom 0 og 1, der ein høg verdi vil indikera god kvalitet, medan ein låg mcq vil indikera dårlegare kvalitet. Brukaren vel ei mcq-grense, der massekromatogram som har ein lågare mcq-verdi enn denne vert forkasta. Mcq-verdien vert rekna ut på basis av tilfeldig støy (stokastisk) og bakgrunnsstøy. Originaldata vert samanlikna med ei tilnærming til eit støyfjerna datasett, for kvar masse. Mcq-verdien vert funnen ved å laga to nye matriser av datasettet. Den fyrste matrisa, M1, er den lengdeskalerte originalmatrisa, M, som vert vist i Formel 2.2-1. Lengdeskaleringa vert gjort for kvar kolonne i settet, og den euklidske lengda (Formel 2.1-1) vert nytta.

Lengdeskalering :

$$M1_{ij} = \frac{M_{ij}}{\lambda_j} \qquad \textit{Formel 2.2-1}$$

*Datasettstrukturen M er definert i kapittel 2.1.4, og* $\lambda$ vert definert i Formel 2.2-2.

Euklidsk lengd:

$$\lambda_j = \sqrt{\sum_{i=1}^{m} M_{ij}^{\,2}} \qquad \textit{Formel 2.2-2}$$

Den andre matrisa, M2, representerer det støyfjerna datasettet. M2 vert konstruert over to steg; fyrst vert tilfeldig støy fjerna og så vert bakgrunnstøyen fjerna, samtidig som variasjonen vert sett til 1 ved standardisering. Den tilfeldige støyen vert fjerna ved glatting av

originaldatasettet. Glattinga vert utført ved å multiplisera M med ei w-diagonal matrise, $W_w$, som vert vist i Formel 2.2-3. Strukturen til glattingsmatrisa, $W_w$, vert vist i *Figur 2.2-1*.

$$M^{glatta} = \frac{1}{w} \cdot W_w \cdot M \qquad \textit{Formel 2.2-3}$$

Glattingsprosedyren førar til tap av w-1 verdiar i kvart massekromatogram, likt fordelt på starten og slutten. Som illustrert for $W_5$ i *Figur 2.2-1*, består dei (w-1)/2 fyrste og siste radene i glattingsmatrisa av nullverdiar. Viss ein ser vekk frå dei totalt fire (5-1) nullkolonnane har dei w første elementa i $W_w$, frå diagonalen og til høgre i kvar rekke, verdi 1, og resten av matrisa har verdi 0. Glattingsprosedyren fører til jamnare kurver i kromatogramma, ettersom den set kvar verdi lik snittet av dei w-1 naboverdiane i kromatogrammet (likt fordelt på kvar side av punktet) og verdien sjølv. w må vera eit (positivt) oddetal, ettersom multiplikasjonen summerer symmetrisk over kvart kromatografisk punkt og dei nærliggjande punkta.

$$W_5 = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
1 & 1 & 1 & 1 & 1 & 0 & 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & 1 & 1 & 1 & 1 & 1 & 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & 0 & 1 & 1 & 1 & 1 & 1 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
\dots & \dots & \dots & 1 & 1 & 1 & 1 & 1 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 1 & 1 & 1 & 1 & 1 & 0 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 1 & 1 & 1 & 1 & 1 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 & 0 & 0 & 0 & 0 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 & 0 & 0 & 0 & 0
\end{bmatrix}$$

*Figur 2.2-1: Glattingsmatrise, W, for w = 5*

Det andre steget i konstruksjonen av M2 er som nemnt å fjerna bakgrunnstøy og å standardisera matrisa. Bakgrunnsstøyen vert fjerna ved å sentrera kolonnane, ved å trekkja ifrå snittverdien. Sentrering i kombinasjon med standardisering vert kalla autoskallering. M2 vert då konstruert ved å autoskalera $M^{glatta}$, som vist i Formel 2.2-5. Utrekning av standardavvik vert vist i Formel 2.2-4.

Standardavvik for kolonna j:

$$s_j = \sqrt{\frac{\sum_{i=1}^{m}(M_{ij} - x_j)^2}{m-1}}$$ *Formel 2.2-4*

$x_j$ er snittverdien for kolonna

Autoskalering av $M^{glatta}$ :

$$M2_{ij} = \frac{M_{ij}^{glatta} - x_j}{s_j}$$ *Formel 2.2-5*

Mcq-verdien vert så funnen ved å kombinera M1 og M2, for kvar masse i datasettet, som vist i Formel 2.2-6.

$$mcq_j = \frac{\sum_{i=-1+w/2}^{m+1-w/2}(M1_{ij} \cdot M2_{ij})}{\sqrt{m-w}}$$ *Formel 2.2-6*

Det vert summert over m-w+1 kromatografiske verdiar, ettersom dei glatta masse-kromatogramma har w-1 mindre verdiar enn dei originale, pga. dei har fått nullstilt dei (w-1)/2 fyrste og siste verdiane.

### 2.2.1 Intervallbasert CODA

SHD-metoden vart introdusert i masteroppgåva til I. Ø. Sandve i 2011 [11].

Tanken bak SHD er at ein masse kan innhalda berre støy i eit kromatografisk intervall, og signal i eit anna. Viss ein då delar opp datasettet i intervall og køyrar CODA separat på desse, vil ein kunne fjerna meir støy og/eller behalda meir signal. For å unngå kutting av signal-toppar, som kan skje dersom dei vert plassert mellom intervall, vert det berre fjerna massar frå eit mindre intervall i midten av det store intervallet. Algoritmen bevegar seg då framover med steg på storleik med det minste intervallet, slik at dei nærliggande store intervalla overlappar. Ved køyring av denne typen algoritme vil ein ikkje kunne fjerna dei forkasta del-massane, ettersom dette ville ført til ei ufullstendig datamatrise. Forkasta verdiar vert derfor nullsett, og ferdig prosessert data vil ha same storleik som original-settet.

## 2.3 Kvalitetsmål i LC-MS

Eit problem i LC-MS er at ein ikkje har eit standardisert mål på kva som er eit godt datasett kontra eit dårlig datasett. Det finnes ulike typar kvalitetsmål, brukt av ulike aktørar. Eit problem knytt til dette er at ein kan velja det kvalitetsmålet som får eins eige resultat til å sjå best mogleg ut. Ei anna utfordring er at ein kan få problem med å tolka analysedataane sine, om ein ikkje er sikker på kva som bra kvalitet og ikkje. Med gode kvalitetsmål kan ein også i større grad nytta støyfjerningsalgoritmer, ettersom ein då har eit mål på kva ein tek vekk og kva ein sit igjen med. [4]

For å finna den optimale metoden må ein veta kva bruksområde det prosesserte datasett skal nyttast til. For eksempel kan ein vera interessert i å fjerna all mogleg støy, eller så vil ein behalda alle kjemiske signal. Dei to nemnte kriteria representerer ein konflikt. Dersom ein fjernar all støyen kan dette gå på bekostning av signal, og vil ein behalda alt signalet kan det godt vera at ein beheld litt støy, for å vera sikkar.

Det kan vera mest gunstig å finna kvalitet ut frå kromatogram, anten alle EIC eller TIC, framfor spektra – ettersom kromatogram er kontinuerlege kurver. Det er kan derfor lagast generelle reglar for kvalitet, ut frå forma til kurvene. Høg kvalitet vil, i dei fleste tilfelle, bety at datasettet inneheld mykje informasjon som også er tolkbar.

Informasjonsindeks-mål

Gong et al. [14] har utvikla eit informasjonsindeks-mål, Φ, basert på informasjonsteori. Målet er basert på Shannon-entropi [24] som kan reknast ut av Formel 2.3-1

$$H = -\int p_i \log p_i \, di \qquad \textit{Formel 2.3-1}$$

Generelt står $p_i$ for verdien til ein sannsyn-distribusjonsfunksjon, ein pdf, i punktet i. Integralet i Formel 2.3-1 er berre gjeldane for kontinuerlege data, og vert erstatta med addisjon av punkt når ein har diskrete data. For eit kromatogram med m punkt vil «i» vera definert i intervallet [1,m], og $p_i$ vil vera intensiteten ved tidspunkt «i». Informasjonsindeksen når sin maksverdi ved normalfordelte kurver, og ein høg verdi vil derfor indikera eit høgkvalitets-kromatogram, ifylgje kjelde [14]. For å ikkje gjera forskjell på kromatogram av ulik intensitet vert uttrykket

frå Formel 2.3-1 normalisert med hensyn på p. Informasjonsindeksen, for eit kontinuerleg signal, vert då uttrykt av Formel 2.3-2 .

$$\Phi = -\int \frac{p_i}{\sum p} \cdot \log \frac{p_i}{\sum p} di \qquad \textit{Formel 2.3-2}$$

I Formel 2.3-3 vert informasjonsindeksen rekna ut for det j'te kromatogrammet av LC-MS-datasettet M (, der M har struktur som i kapittel 2.1.4). Datasettet er diskret, og det vert derfor nytta addering, framfor integrasjon.

$$\Phi_j = -\sum_{i=1}^{m} \left( \frac{M_{ij}}{\sum_{i=1}^{m} M_{ij}} \cdot \log \frac{M_{ij}}{\sum_{i=1}^{m} M_{ij}} \right) \qquad \textit{Formel 2.3-3}$$

Φ er eit mål for vektorar, ikkje matriser, og ein kan då f.eks. nytta målet for kvart EIC for eit LC-MS-datasettet, eller ta snitte av alle informasjonsindeksane, for å få ein verdi som representerer heile settet.

Signal til støy - forhold (SNR – Signal Noise Ratio) [25] [4] er eit kvalitetsmål der ein finn forholdet mellom signal (S) og støy (N = noise) i kromatogrammet. Formel 2.3-4 viser ei mogleg utrekninga av SNR.

$$SNR = \frac{S}{N} \qquad \textit{Formel 2.3-4}$$

Det finnest også andre måtar å rekna ut SNR på. Ein definisjon nytta av [26] og er gjeven i Formel 2.3-5.

$$SNR = \sqrt{\frac{\sum_i s_i^2}{\sum_i n_i^2}} \qquad \textit{Formel 2.3-5}$$

$s_i$ er signal i punktet «i» og $n_i$ er støy i punktet.

Det er også andre måtar å rekna ut SNR på. Blant anna definerer EPA (United States Environmental Protection Agency) [27] SNR til å vera halvparten av SNR frå Formel 2.3-5.

Den kritiske fasen i utrekninga av SNR-verdien er å fastsetja kva som er signal, kva som er støy og kvar ein målar dette. Ein måte å definera signal-toppar på, er gaussiske toppar med høgare intensitet enn eit visst antall standardavvik. Denne grensa vert ofte kalla for deteksjonsgrensa, LOD (Limit of Detection), dersom ein ser på konsentrasjonar. Om ein ser på signalet og ikkje konsentrasjonar, er korrekt termologi «limit of decision» - ei fastsetjingsgrense . [28]

Ved eit Signalet i utrekning av SNR kan då vera avstand frå den høgste signaltoppen ned til baselinja. Ein kan definera områda utan signal til å vera støy, og estimera støynivået som standardavviket til desse områda.

Liang et al. [29] hevdar at ei limit of decision bør definerast som storleiken nettosignalet til ein komponent må ha for å kunna skiljast frå ein analytisk blank-prøve. Nettosignal er det signalet i ein komponent som er unikt, dvs. ikkje samvarierer med andre komponentar. Limit of decision vil ut ifrå denne definisjonen kunna fastsetjast ut frå kromatografiske områder utan kjemiske komponentar, nullkomponentsområder, fordi desse i praksis er analytiske blank-prøvar. For eit nullkomponentområde, kan ein for eit LC-MS-sett finna eigenverdiar til kombinasjonar av spektra i området. (Eigenverdiar vert introdusert i neste kapittel) Formel 2.3-6 viser korleis ein finn antall moglege kombinasjonar av w spektra ein kan laga av totalt m spektra.

$$\binom{m}{w} = \frac{m!}{(m-w)!\cdot w!} \qquad \textit{Formel 2.3-6}$$

For eit nullkomponentsområdet vil då m vera antall scannings-punkt. w kan velgast til å vera f.eks. 4. Liang et al. nyttar fordelinga av desse eigenverdiane av desse, til å fastsetja ei limit of decision. Limit of decision kan fastsetjast som ei grense ved f.eks. dei 5 % øvste eigenverdiane eller grensa ved ein F-fordelingstest.

## 2.4 Matematiske eigenskapar og transformasjonar

### 2.4.1 Eigenverdiar, singulærverdiar og rang

Eigenverdiar [30] og eigenvektorar er definert frå den fylgjande likninga:

$M x = \lambda x$ *Formel 2.4-1*

Formel 2.4-1 er M ei n x n matrise, x er ein eigenvektor som ikkje er ein nullvektor, og $\lambda$ er eigenverdien til M. Alle x og $\lambda$, som er moglege for M vert kalla respektivt eigenvektorar og eigenverdiar. Eigenvektorar er lineært uavhengige til kvarandre, og representerer derfor ulike eigenskapar. Fordi kvar eigenvektor har ein eigenverdi vil også kvar eigenverdi representera ulike eigenskapar. Problemet med eigenverdiar er at dei ikkje kan finnast for alle typar matriser, ettersom matrisa må vera kvadratisk. Det i mange tilfelle derfor vera betre å finna singulærverdiane til matrisa, ettersom desse gjer mykje av den same informasjonen og dei kan finnast for alle matriser.

Ei m x n matrise, M, med kolonnerang r kan faktoriserast som vist i Formel 2.4-2. At matrisa har kolonnerang r vil seia at r av kolonnane er lineært uavhengige. Matrisa kan faktoriserast som vist i Formel 2.4-2 :

$M = U\Sigma V^T$ *Formel 2.4-2*

Dette vert kalla singulærverdi-dekomponering, SVD (Singular Value Decomposition). U er ei ortogonal m x m matrise og V er ei ortogonal n x n matrise. Σ er ei diagonalmatrise der dei fyrste r diagonalelementa er positive og i rangert rekkefølgje frå størst til minst. Desse r diagonalelementa i Σ vert kalla singulærverdiar. Antall singulærverdiar med verdi ulik null vert kalla den matematiske rangen til matrisa. Denne rangen er den same som ein finn av antall eigenverdiar. I ei m x n matrise med full rang vil antall singulærverdiar vera den minste verdien av m eller n. [30, 31]

I eit ideelt LC-MS datasett (utan støy) vil antall singulærverdiar som ikkje er null, tilsvara kor mange kjemiske komponentar som er analysert. [32] Antall kjemiske komponentar vert kalla kjemisk rang. Når det er støy tilstades vil antall singulærverdiar over null, matematisk rang, vera høgare enn kjemisk rang for datasettet. Ved lite støy vil likevel dei ”ekte” verdiane skilja seg ut ved at dei er klart større enn dei andre singulærverdiane.

### 2.4.2 Basisar for matriser og funksjonar

Basisar [30, 33] kan sjåast på som sett av enkle ledd som ein ved å summera, kan spenna meir kompliserte objekt av. Basisar vert for eksempel nytta for vektorar og funksjonar. Felles for

dei to typane basisar er at dei består av lineært uavhengige ledd (vektorar eller funksjonar). Lineær uavhengigheit kan forklarast ut frå Formel 2.4-3, der c-ane er skalere koeffisientar, v-ane er vektorar av lik lengd og **0** er ein nullvektor. Vektorane er lineært avhengige dersom uttrykket berre har ei løysing (dvs. at alle c = 0).

$$c_1v_1 + c_2v_2 + \ldots + c_pv_p = \mathbf{0} \qquad \textit{Formel 2.4-3}$$

Fordi ledda i ein basis er lineært uavhengige, vil kvart objekt berre kunna representerast på ein måte av basisen. Dette vert kalla unik representasjon. I Formel 2.4-4 og Formel 2.4-5 vert vektoren v og funksjonen f(t) representert av basis-vektorane $v_i$ og basis-funksjonane $f_i(t)$, der $c_i$ er skalare koeffisientar.

$$v = \sum c_i v_i \qquad \textit{Formel 2.4-4}$$

$$f(t) = \sum c_i f_i(t) \qquad \textit{Formel 2.4-5}$$

Eit eksempel på basis for vektoren $\begin{bmatrix}3\\1\end{bmatrix}$ er eit sett av vektorane $\begin{bmatrix}1\\0\end{bmatrix}$ og $\begin{bmatrix}0\\1\end{bmatrix}$, der den fyrste basis-vektoren vert gonga med 3 og den andre med 1.

$$\begin{bmatrix}3\\1\end{bmatrix} = 3\begin{bmatrix}1\\0\end{bmatrix} + \begin{bmatrix}0\\1\end{bmatrix}$$

Ortogonalitet mellom basis-vektorar har ein dersom indreproduktet mellom dei er null. Dersom vektorane også er normaliserte, har ein ein ortonormal basis. Ortogonale eller ortonormale basisar er ofte ynskja.

### 2.4.3 Wavelettransformasjon (WT)

Transformasjon [34] av data vil sei å representera data på ein ny måte. Ein får fram andre eigenskapar av data, noko som kan vera nyttig i fleire samanhengar. Ein kan også transformera, gjera endringar i data og så tilbaketransformera (inverstransformering) til den originale forma.

Wavelet-transformasjon (WT) er ei matematisk transformering som kan blant anna vert nytta til komprimering, analyse og støyfjerning av signal. WT vert nytta for både 1D- og 2D-data, men her vil det verta fokusert på 1D-transformasjon.

WT har likskapstrekk med den meir allment kjente teknikken fouriertransformasjon (FT), då begge byggjar på bruk av basisfunksjonar til å utføra analyse av datasett. FT nyttar sinus og cosinus som basisfunksjonar og i WT vert wavelets nytta som basisfunksjonar. WT kan beskrivast som; «å finna koeffisientar som er indreproduktet av eit signal og ein waveletfamilie.» [35] Wavelets betyr små bølgjer [36], og heiter dette fordi dei er lokaliserte bølgjer. At dei er lokaliserte betyr at dei berre er ulike null i eit lokalt område. [34]

Det finnes fleire ulike wavelet-transformasjonsmetodar, men ein kan grovt sett dela opp i to hovudgrupper; DWT (diskret wavelet transformasjon) og CWT (kontinuerleg wavelet-transformasjon), der diskret WT vert nytta for diskrete signal og kontinuerlege WT helst vert nytta for funksjonar.

Ein waveletfamilie, også kalla ein wavelet basis, kan presenterast som fylgjer :

$$\psi^{a,b}(x) = |a|^{-1/2}\psi(\frac{x-b}{a}), \quad a,b \in R \ , \ a > 0 \qquad \textit{Formel 2.4-6}$$

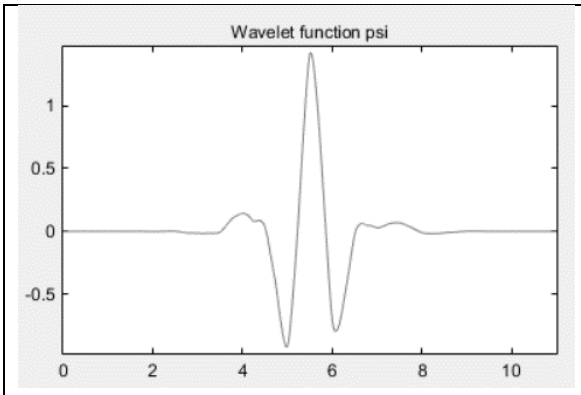
I formelen over er ein heil waveletfamilie representert. $\psi$ (psi) er moderwaveleten, som er unik for kvar waveletfamilie. Moderwaveleten er ein funksjon defineret av eit sett wavelet-filter-koeffisientar. Alle dotter-waveletane , $\psi^{a,b}(x)$ , for fastsette a og b, vert definerte ut ifrå moder-waveleten [34]. Dotter-waveletane sin storleik og plassering avhengar av skaleringsparameteren a og forskyvingsparameteren b. Skaleringsparameteren avgjer både amplitude og breidd av waveleten, medan forskyvingsparameteren avgjer (endring i) plassering på x-aksen.

Skaleringsparametaren, a, og forskyvingsparameteren, b, vert ofte kalla for tidsskaleringsparameterane, viss ein har eit datasett der x er tid.

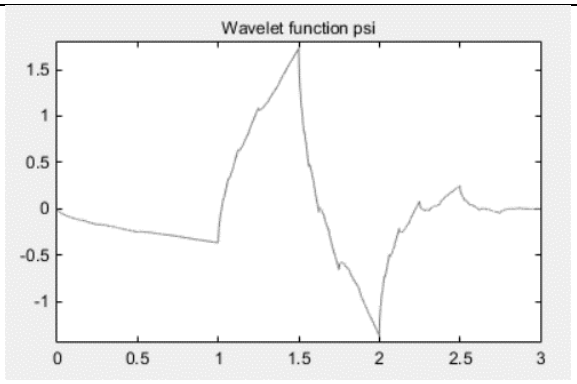
Ein kan dela wavelets inn i både familiar og storfamiliar. Ein wavelet familie er som vist definert ut frå moderwaveleten. Moderwaveletane kan igjen høyra til ein storfamilie. Wavelets innan same storfamilie vil ha liknande namngjeving men ulikt nummer, som f.eks.

*Daubechies 1* og *Daubechies 4*, der *Daubechies 1* og *Daubechies 4* begge er wavelet-familiar. Forskjellen innad i storfamiliane er at waveletane er bygd på opp av eit ulikt antall wavelet-filter-koeffisientar [37], der antall koeffisientar vert bestemt ut frå nummeret i namngjevinga. Waveletfamiliar med høgare nummer vil innehalda dei same wavelet-filter-koeffisientane som dei med lågare nummer, i tillegg til nokre fleire.

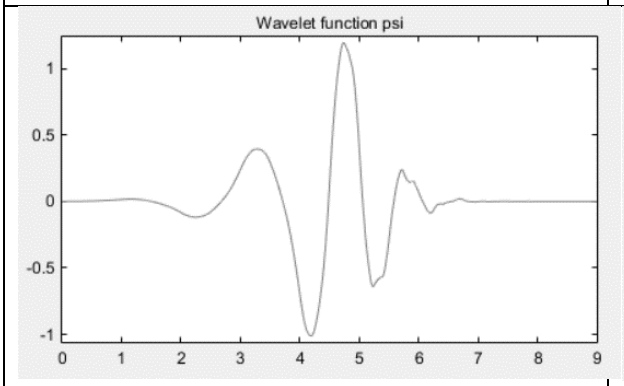
I Figur 2.4-1 er nokre wavelets frå ulike familiar plotta.



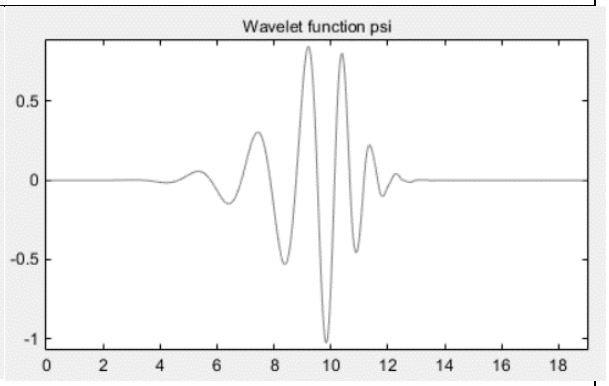

*Symlet 6*



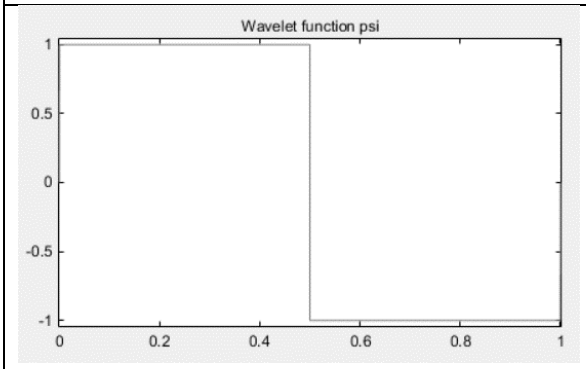

*Daubechies 2*



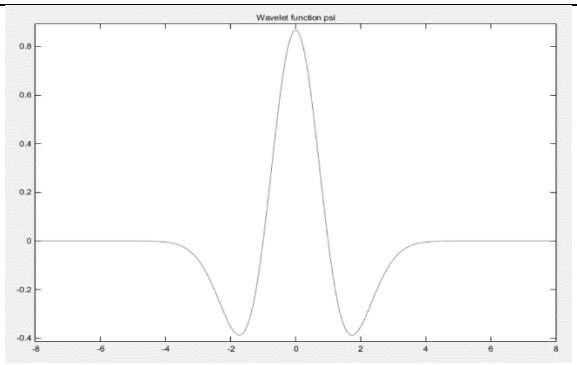

*Daubechies 5*




*Daubechies 10*




*Haar* eller *Daubechies 1*




*Mexican hat*

*Figur 2.4-1: Grafisk framstilling av nokre wavelet-moder-funksjonar [38] [39]*

Det finnes eit stort antall moderwavelets, noko som er gunstig ved signalanalyse. For å kunna dekomponera eit signal på ein bra måte, er det ein føresetnad at waveletane liknar på signalet, eller den delen av signalet ein vil analysera.

### 2.4.3.1 Kontinuerlig wavelettransformasjon (CWT)

Kontinuerleg WT (CWT) [35, 40]- transformeringa, også kalla analyse av funksjonen, vert utført som i Formel 2.4-7 og Formel 2.4-8

$$(W\ f)(a,b) = \langle f, \psi_{a,b} \rangle = |a|^{-1/2} \int_{x \in R} f(x)\, \psi\left(\frac{x-b}{a}\right) dx \qquad \textit{Formel 2.4-7}$$

$$(W\ f)(a,b) = \int f(x)\, \psi^{a,b}(x)\, dx \qquad \textit{Formel 2.4-8}$$

Vinkelparentesar står for indreprodukt, W er alle waveletane i ein waveletfamilie og f er signalet. Formelen er gjeven ved fastsette a og b – parameterar.

Inverstransformasjonen (iCWT), også kalla syntese av funksjonen, vert utført som i Formel 2.4-9.

$$\textit{iCWT}: \quad f(x) = \int_{a=0}^{\infty} \int_{b=-\infty}^{\infty} (w\ f)\,(a,b)\, \psi^{a,b}(x) \frac{da\ db}{a^2} \qquad \textit{Formel 2.4-9}$$

Her vert f(x) uttrykt som ein superposisjon av wavelets. w er vekter kalla waveletkoeffisientar. f er signal som skal prosesserast, x er fastsatt, og a og b varierer.

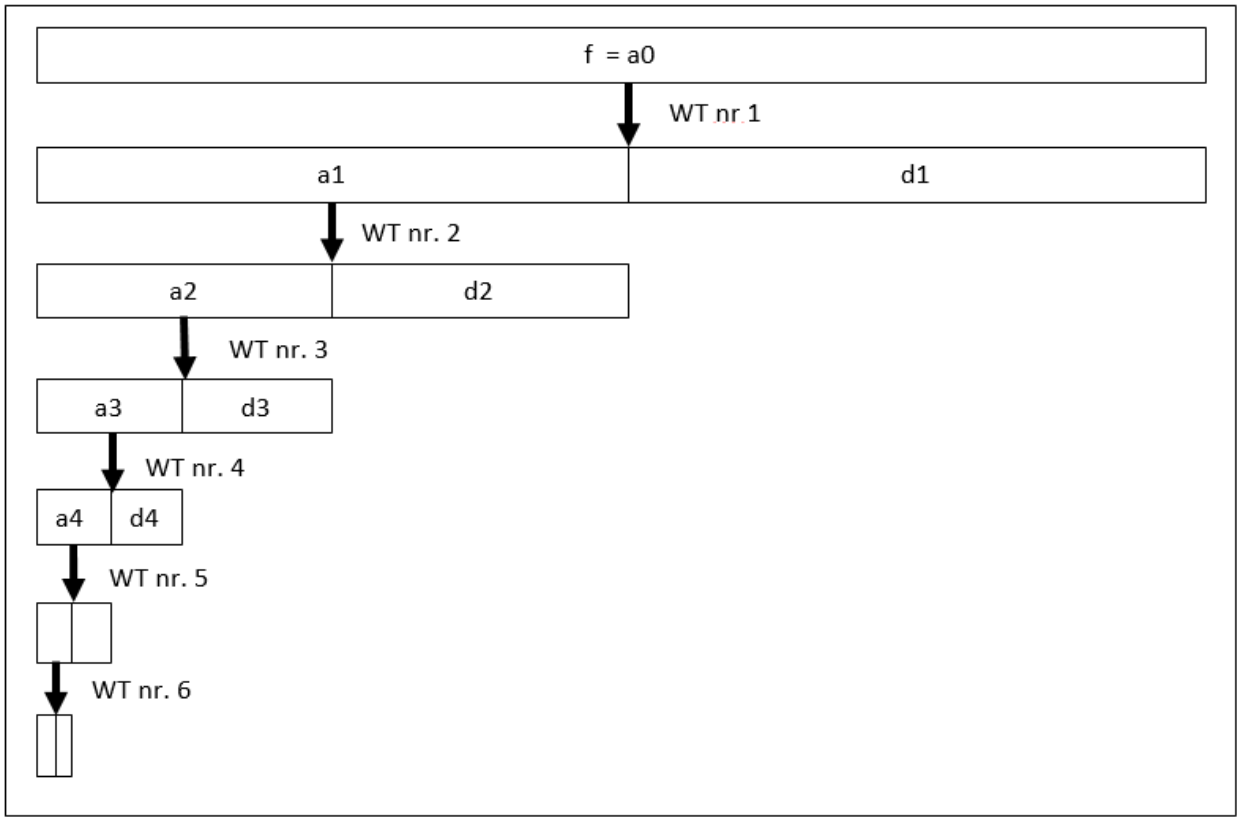
Ein familie av wavelets er som oftast ein ortogonal basis [33]. Ortogonale basisar er gunstige, blant anna fordi det gjer utrekningar mindre krevjande ved at ein får færre ledd.

### 2.4.3.2 Diskret wavelettransformasjon (DWT)

Som nemnd har CWT og DWT ulike nytteområder og føresetnadar . Ved CWT er både a, b og signalet som vert analysert kontinuerlege, medan for DWT er dei alle diskrete. Når ein arbeidar med empiriske datasett, vil desse aldri innehalda eit uendeleg antall verdiar, og datasetta er derfor diskrete. DWT vil då vera det naturlege valet.

For DWT har det vorte vanleg å nytta eit dekomponeringsskjema som i Figur 2.4-2, som er eit oktavband. Dekomponeringa vert ofte kalla ei trealgoritme. Vanlegvis er a og b definerte som; $a = 2^j$ og $b = k\,2^j$ . [35] (For a og b er j oktaven.) Dette betyr at data-vektoren, f, må ha lengd $2^p$, der $p \geq J$, som er den største j, for at dekomponeringa skal gå opp. Dersom ein vektor f ikkje har passande lengd er det vanleg å leggja til nullverdiar symmetrisk i starten og slutten av f. Desse verdiane vert kalla dummy-verdiar [41]. Trealgoritmen nyttar «high pass» og «low pass» filter, definert ut frå wavelet-basisen, slik at datavektoren vert delt opp i eit «high pass» d-ledd (detaljledd) og eit «low pass» a-ledd (approksimert ledd). a-leddet inneheld dei grove lågfrekvente konturane i vektoren, medan d-leddet viser dei høgfrekvente «spisse» konturane. Her kan f sjåast på som a-leddet ved nivået j = 0. Algoritmen fungerer rekursivt, på den måten at den først dekomponerer f ( = $a^0$) til $a^1$ og $d^1$, og så dekomponerer $a^1$ til $a^2$ og $d^2$ osv. For kvart nytt nivå vert a- og d- ledda halvert i lengd, ettersom dei totalt sett har lengda til a-leddet i førre nivå.

[37]




*Figur 2.4-2: D*WT-dekomponering av signalvektoren f ved trealgoritmen

«Low pass» og «high pass» filtera er skaleringssekvensar og wavelet-sekvensar. Ein waveletsekvens er ein diskretisert waveletfunksjon (som fylgjar av dei fastsette a og b), og ein skaleringssekvens er motparten til denne. Skaleringssekvensen er bygd opp av dei same wavelet-filterkoeffisientane, men har annan rekkefylgje og forteikn for desse. [37]

For DWT er waveletkoeffisientar definert som indreprodukt av signalvektoren og ein wavelet-sekvens, der sekvens vert nytta om ein dotter-wavelet på diskret form. Waveletkoeffisientane finn ein i detaljledda. Dette vil sei at å rekna ut indreprodukta mellom eit a-ledd i nivået j og waveletsekvensane er å finna d-leddet i nivået j+1. Skaleringskoeffisientar er koeffisientane som høyrar til i a-ledda, og er indreprodukta av skaleringssekvensane og a-koeffisientane i førre nivå.

Utrekning av DWT, m.a.o. analyse av signalvektoren, vert vist i Formel 2.4-10 og Formel 2.4-11.

$$DWT\{f[n];2^j,k2^j\} = c_{j,k} = \sum_n f[n]\, h_j^*\left[n - 2^j k\right] \qquad \textit{Formel 2.4-10}$$

$$b_{j,k} = \sum_n f[n]\, g_j^*\left[n - 2^j k\right] \qquad \textit{Formel 2.4-11}$$

Der f [n] er eit punkt i datavektoren f,

$c_{j,k}$ er wavelet-koeffisientar,

$b_{j,k}$ er skalerings-koeffisientar,

$h_j\left[n - 2^j k\right]$ er analyse-waveletsekvensane,

$g_j\left[n - 2^j k\right]$ er analyse-skaleringssekvensane.

Ettersom k er definert innanfor eit gjeve intervall og steg-storleiken aukar med ein faktor av 2 for kvart nivå ($b = 2^j k$), vil ein få halvert antall waveletsekvensar (dottar-wavelets) og dermed antall koeffisientar (indreprodukt) for kvart nivå.

Formel for inverstransformering (iDWT), altså syntese av signal-vektoren, vert vist i Formel 2.4-12.

$$iDWT: \quad f[n] = \sum_{j=1}^{J} \sum_{k \in Z} c_{j,k} \tilde{h}[n - 2^j k] + \sum_{k \in Z} b_{J,k} \tilde{g}_J[n - 2^J k] \qquad \textit{Formel 2.4-12}$$

J er antall nivå i trealgoritmen.

$\tilde{h}[n - 2^j k]$ er syntese-waveletsekvensane

$\tilde{g}_J[n - 2^J k]$ er syntese-skaleringssekvensane, for nivå J som er den høgste j

Sekvensane vert konstruert rekursivt for j>1 som vist under, i Formel 2.4-13, Formel 2.4-14 og Formel 2.4-15.

$$g_1[n] = g[n] \ , \ h_1[n] = h[n] \qquad \textit{Formel 2.4-13}$$

$$g_{j+1}[n] = \sum_k g_j[k] \, g[n - 2k] \qquad \textit{Formel 2.4-14}$$

$$h_1[n] = \sum_k h[k] \, g[n - 2k] \qquad \textit{Formel 2.4-15}$$

[35]

Ein alternativ måte å representera DWT på er matriseform, ettersom signalet er diskret.

DWT vert då utført som i Formel 2.4-16 og Formel 2.4-17 og iDWT som i Formel 2.4-18

$a^j = G \, a^{j-1}$ *Formel 2.4-16*

$d^j = H\, a^{j-1}$ *Formel 2.4-17*

*iDWT :* $a^{j-1} = G^* a^j + H^* d^j$ *Formel 2.4-18*

Her er H og G matriser med analyse-waveletsekvensane og analyse-skaleringssekvensane i kolonnane.

H* og G* ( der * betyr konjugert) er matriser med syntese-waveletsekvensane og syntese-skaleringssekvensane i kolonnane.

For ei matrise A med ortonormale basisar gjeld ; $A^* = A^T$ [37]. Ved DWT er waveletane ortonormale til kvarandre [42] . Dette betyr at ein kan skriva om uttrykket for iDWT ( i Formel 2.4-18) som fylgjar i Formel 2.4-18:

*iDWT :* $a^{m-1} = G^T a^m + H^T d^m$ *Formel 2.4-19*

Dette gjer at både DWT og iDWT kan verta effektivt utført, ettersom å transponera ei matrise treng lite utrekning. Det betyr også at rekonstruksjonen er utan tap, forutan avrundingsfeil. [35]

For iDWT har det her vore snakk om ein fullstendig rekonstruksjon, ved å nytta alle a- og d-ledda for dei underliggjande nivåa, men det er også mogleg å utelata nokre ledd ved å ikkje ta desse med i addisjonen. Ein vil då kunna få f.eks. a2 med same lengd utgangsvektoren f ( = $a^0$).

### *2.4.3.3 Udesimert wavelettransformasjon (ndWT)*

ndWT (nondecimated WT) [43-46] er ein spesialversjon av DWT, og er ei «Á Trout» WT-algoritme. Á Trout er fransk for hol, og betyr her at skaleringssekvensane i ndWT får konstant lengd for alle nivåa, ved å fylla inn med nullar mellom wavelet-filter-koeffisientane. Dette gjer også at kvar dekomponeringsvektor, $a^j$ eller $d^j$, vil ha lik lengd som den originale vektoren, $a^0$ ( = f). Ettersom $a^j$ og $d^j$ har lik lengd, kan $d^j$ finnast på fylgjande måte:

$d^j = a^j - a^{j-1}$ *Formel 2.4-20*

Transformasjonen vert kalla skift-invariant, fordi alle skaleringssekvensane har lik lengd. At ndWT er skift-invariant gjer transformasjonen uortogonal, og inverstransformasjonen kan derfor ikkje finnast som for iDWT (Formel 2.4-19). Transformasjonane kan, likevel, lett inverterast. Ved addisjon av eit gjeve a-ledd, d-leddet til same nivå og alle d-ledda av lågare nivå finn ein den inverterte. Formel 2.4-21 viser inversjon til eit gjeve nivå j, medan Formel 2.4-22 viser fullstendig inversjon.

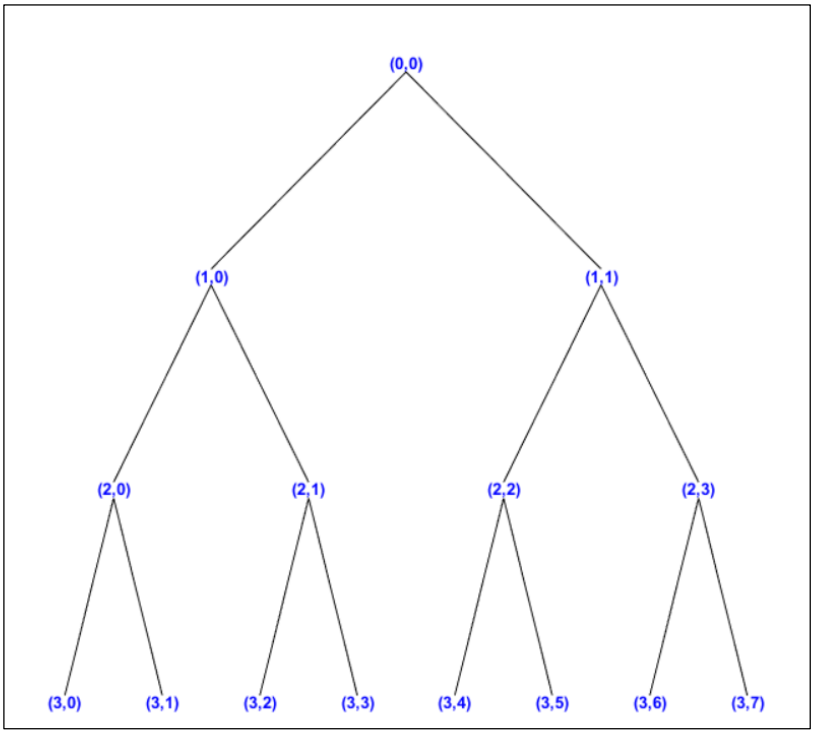
$$a^j = a^J + \sum_{i=j+1}^{J} d^i$$ *Formel 2.4-21*

$$a^0 = a^J + \sum_{j} d^j$$ *Formel 2.4-22*

,der $a^0$ er den originale vektoren, og alle a - og d - ledd har lik lengd som $a^0$

### *2.4.3.4 Wavelet packet transformasjon (WPT)*

WPT er ein type DWT der trealgoritmen har vorte utvida til å gjelda heile det binære treet, som vil sei at også detaljledda vert dekomponert til a- og d- ledd. Dekomponeringsskjemaet for WPT vert vist i Figur 2.4-3. For WPT vert det meir ryddig å referera til j- (nivå) og k- (koeffisient-vektor-nummer) verdiar enn d og a. Dei forskjellige koeffisient-vektorane vert ofte kalla noder, og vert då namngjevne på forma node(j,k). node(1,1) og node(1,2) vil då vera a1 og d1 ut frå DWT - oppsettet. Ut frå det binære treet vil ein då ha $2^j$ noder for kvart nivå j .




*Figur 2.4-3: WPT-tre: Tre-struktur som visar dekomponeringmønsteret for wavelet packet transformasjon. Figuren er henta frå MATLAB. [39] Dekomponeringa fylgjar same system som for DWT, med a-ledd til venstre og d-ledd til venstre, men for WTP vert også d-ledd dekomponert til nye a- og d- ledd (som vert høgre og venstre under d-ledda).*

WPT er gunstig dersom ein vil sjå grundigare på detaljledda, ettersom ein kan dekomponera desse vidare. Om ein vil gjera endringar i koeffisientane, før inverstransformering, har ein ved WPT moglegheita til behandla fleire eigenskapar i data ( les noder) separat. For WPT er det ofte aktuelt å finna den «beste basisen» for datavektoren, der ein basis er ein kombinasjon av ikkje overlappande noder. At noder ikkje har overlapp betyr at dei ikkje ligg direkte over/under kvarandre i dekomponeringsskjemaet; dvs. (3,4) overlappar med (2,2) og (1,1), men overlappar ikkje med (3,5) og (2,3). Den beste basisen kan fastsetjast ut frå ulike kriterium. Eit kriterium kan vera å finna basisen som inneheld mest mogleg informasjon, fordelt på eit minst mogleg antall koeffisientar med verdi over ei grense. Dette kriteriet vert gjerne kalla informasjonskriteriet. Den beste basisen kan vera relevant å finna f.eks. dersom ein vil komprimera eit signal eller fjerna støy. Når ein har definert og funne den beste basisen, kan ein velja ut koeffisientar ein vil behalda og f.eks. nullsetja dei andre. [37]

#### *2.4.3.5 Wavelettransformasjon i kromatografi*

Ved dataanalyse er målet ofte å trekkja fram spesifikke kvalitetar ved datasettet. Ved bruk av waveletanalyse vil val av både wavelet-algoritme, grad av dekomponering og type wavelet ha innverknad på kva som vert framheva. Viss ein f.eks. tenkar seg at d1-leddet inneheld støy og signal, vil WPT vera ein meir logisk måte å modellera støyen med, enn f.eks. DWT, ettersom WPT dekomponerar d1-leddet (node 1,2) inn i fleire ledd. Når ein skal velja wavelets, basis, er det gunstig at desse liknar på forma til det ein søkar etter. Wavelets som vert trukke fram i artiklar som gode til modellering av kromatografisk data er eksempelvis *Coiflet 1*, forkjellige *Daubechies* wavelets, forkjellige *Symlets* og *Mexican hat*. [9, 47-53]

# 3 Verktøy, datasett og eksperimentelt

## 3.1 Datamaskin og programvare

### 3.1.1 Datamaskin-spesifikasjonar

Prosessor: 4 kjerna Intel Core i7 -4712 MQ

Minnekort: 8 GB DDRL L

Skjermkort: NVIDIA GeForce 840M med 2 GB dedikert minne

Operativsystem: Windows 8.1

### 3.1.2 Programvare

MATLAB (Matrix Laboratory) R2014a og R2014 b frå Mathworks vart nytta som programmeringsmiljø. Wavelet Toolbox (v. 4.14) – tillegget til MATLAB vart også nytta.

Dei to fylgjande programma vart nytta til konvertering av datasettet.

MSConvert frå proteowizard 3.0.7127 (http://proteowizard.sourceforge.net/)

Chrombox D 12-09b frå (www.chrombox.org)

## 3.2 Datasett og datasettbehandling

### 3.2.1 Datasett

I oppgåva er det nytta fleire datasett, der eit er simulert og dei resterande er reelle data. Ein oversikt over datasetta vert vist i Tabell 3-1.

*Tabell 3-1: Oversikt over datasetta nytta i arbeidet*

| Namn | Opphav | Dimensjon (m x n) | Prosessering | Masse-oppløysing (m/z) |
|---|---|---|---|---|
| sim3 | simulert | 400x40 | | |
| HCO | LC-ESI-MS | 2571 x 1004 | binnining, baselinje-subtraksjon | 1 |
| ZM2 | LC-ESI-MS | 722 x 701 | kutta (trunkert), binning | 1 |

| | | | | |
|---|---|---|---|---|
| TAG | LC-ESI-MS | 839 x 1502 | binning | 1 |

*Tabelltekst til Tabell 3-1: Prosessering refererar her til endringar som har vorte gjort i data etter opptak, forutan sentroidisering og konvertering av filtype. Under dimensjon er det m antall retensjonstider og n antall m/z-verdiar.*

#### *3.2.1.1 Datasettbehandling*

sim3- datasette er i utgangspunktet konstruert med ei baselinje bestående av verdien 2 addert til kvar verdi. Denne vert trukke i frå, før settet vart nytta.

ZM2- settet vart konverter som vist i Vedlegg 42, og vart i tillegg binna ved bruk av HCO-settet vart i tillegg binna matlab-funksjonen *binning,* oppgjeven i kapittel 3.3

Settet vart i tillegg trunkert (kutta) slik at det startar med retensjonspunkt 780.

HCO-settet vart konvertert som vist i Vedlegg 43. Det vart forsøkt å subtrahera baselinja ved å trekkja frå verdien $2.119*10^5$ frå kvart punkt i datasettet.

TAG-settet vart tildelt i binna MATLAB format. Anna info om konverteringa er ikkje kjent.

#### *3.2.1.2 sim3-settet*

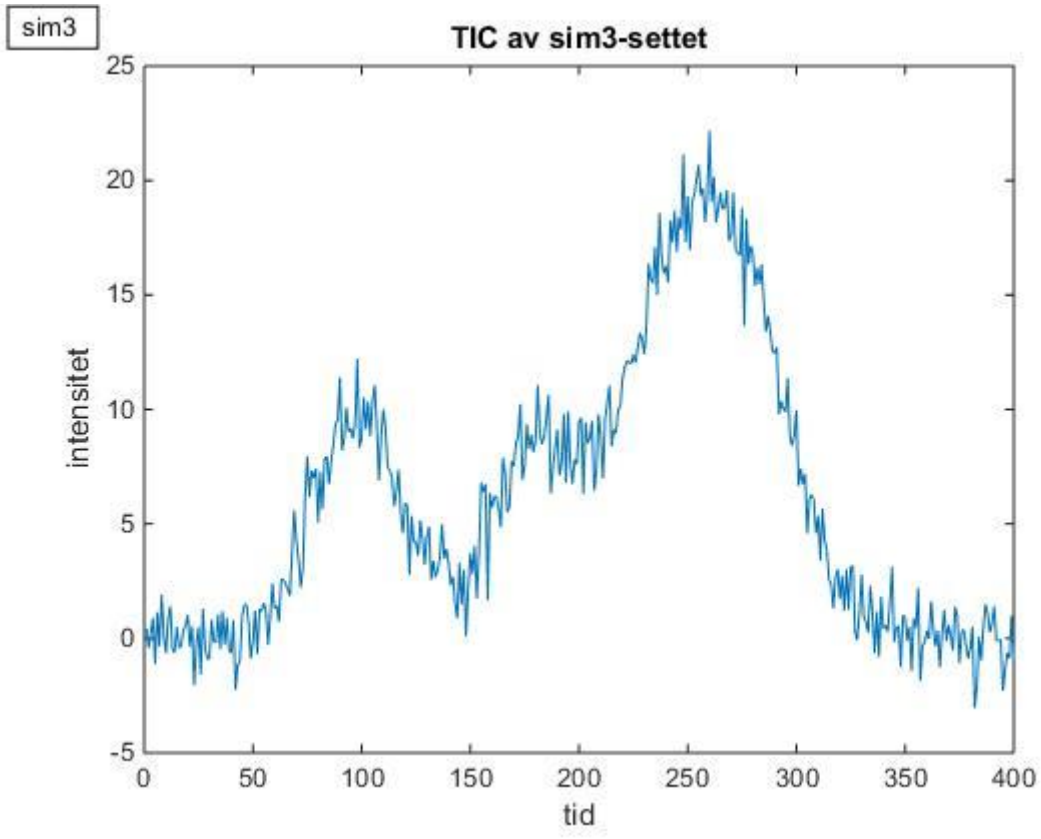
Det simulerte datasettet sim3 er konstruert som produktet av fem basisspektra og fem gaussiske kurver. Dei fem basispektera er samansette av reelle spektra, medan kurvene som representerer kromatogram er tillagte kurver med ulik breidd, høgd og posisjon. Totalt har settet 14 massar over støynivået og ein masse som druknar i støyenss. Tabell 3-2 viser ein oversikt over toppane i sim3-settet, og TIC av settet vert vist i Figur 3.2-1.

*Tabell 3-2: Oversikt over toppar og antall m/z-verdiar i kvar topp for sim3-datasettet*

| Topp nr. | Toppunkt (tidsakse) | m/z-verdiar (antall) |
|---|---|---|
| 1 | 84 | 1 |
| 2 | 100 | 2 |
| 3 | 184 | 5 (8) |

| | | |
|---|---|---|
| 4 | 234 | 2 |
| 5 | 267 | 4 |

*Tabelltekst til Tabell 3-2: topp nr. 3 har 5 ikkje overlappande massar (m/z-verdiar) av normal størrelse, av dei resterande tre er ein stor nok til å visa i EIC og to er under støynivået. Ein av dei to massane under støynivået og den siste av dei tre overlappar med høgare toppar i forbindelse 2.*




*Figur 3.2-1: TIC-representasjon av sim3-settet.*

For å laga datasettet vert spektera og kurvene kombinert ved Formel 3.2-1. Kurvene vert lagra i ei kromatogrammatrise, C, der kvar kurve er ein kolonnevektor. Spektera vert lagra i ei spektermatrise, S, der kvart spekter er ein radvektor. LC-MS-settet vert i formelen representert som M. M har like mange rader som lengda av kurvene, og like mange kolonner som lengda spektera.

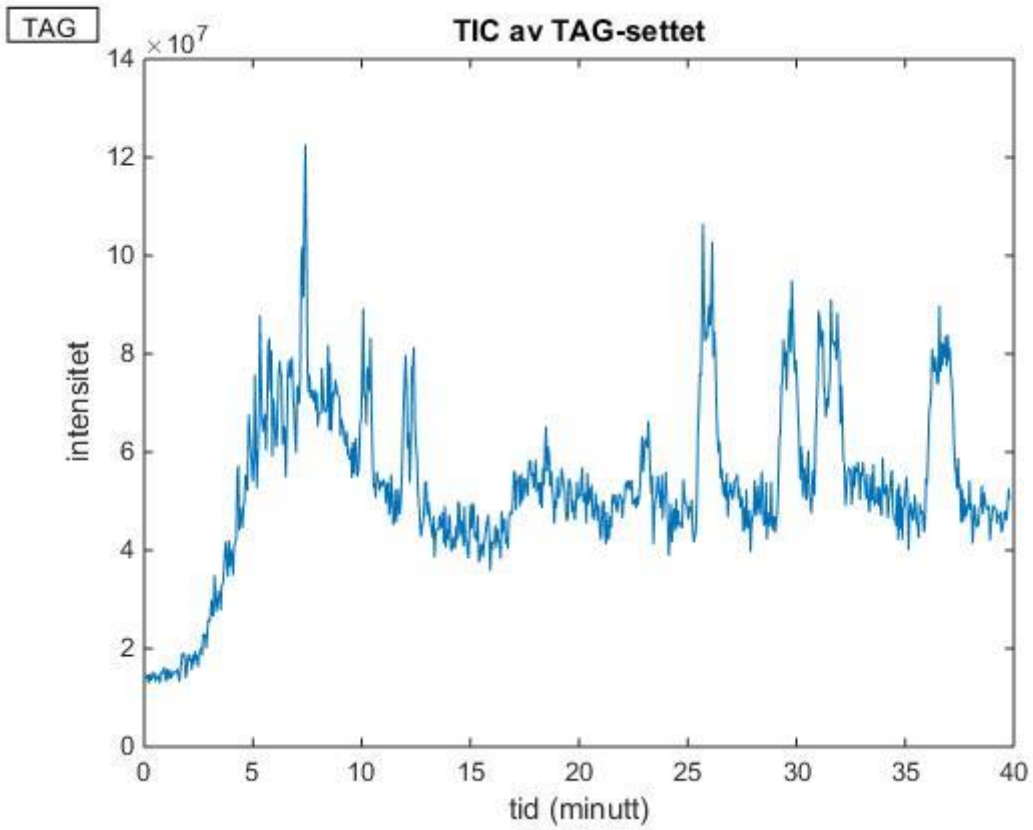
$$M = C \cdot S \qquad \textit{Formel 3.2-1}$$

Settet vert så pålagt støy bestående av to ledd, eit homoskedastisk ledd og eit heteroskedastisk ledd. Den heteroskedastiske støyen er spesifikk for kvar masse og er berre

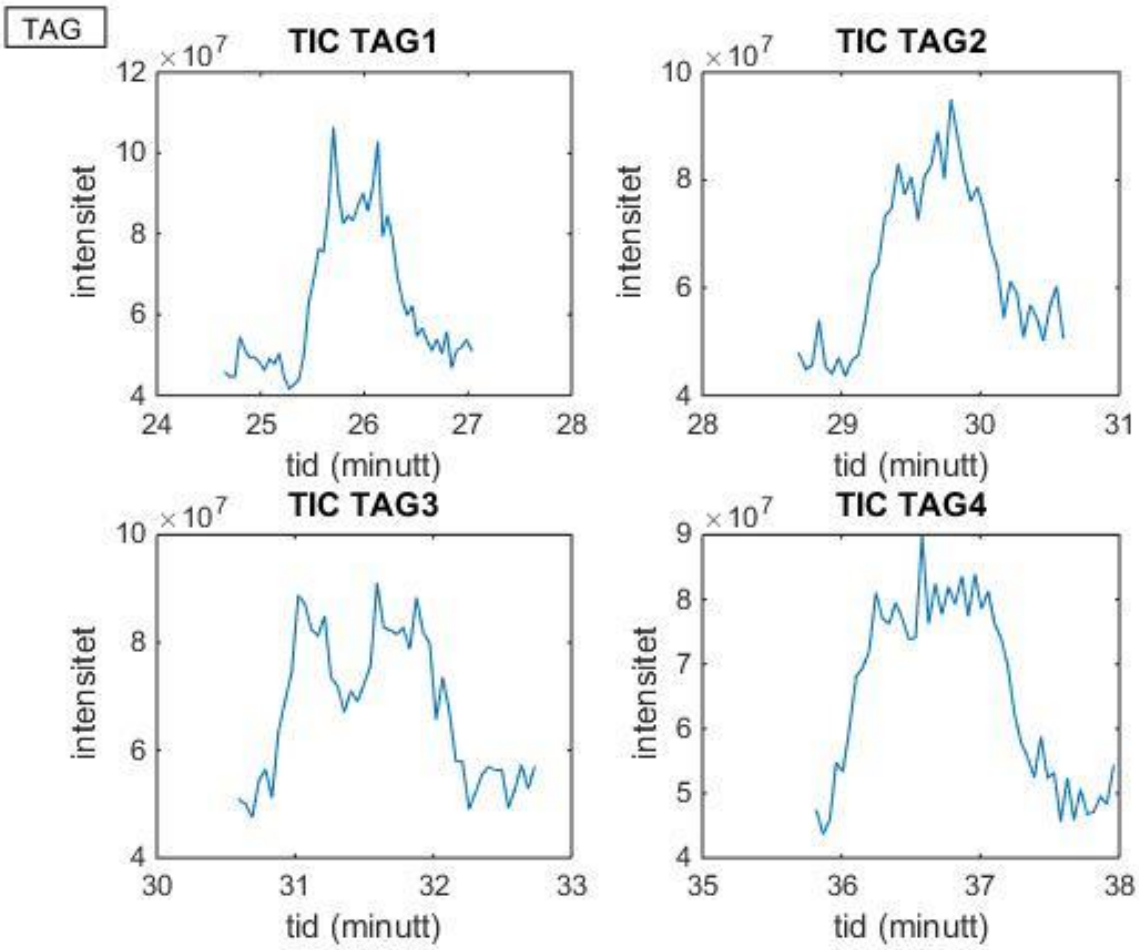
avhengig av den opprinnelege signalverdien til den spesifikke massen. Sjå i kapittel 3.3 for å finna skriptet «make_dataset3», som visar i detalj korleis datasettet vart konstruert.

### *3.2.1.3 TAG-settet*

TAG-settet [54] er eit LC-MS-sett av fem triacylglyserider (TAG) i blanding. TIC av settet vert presenter i Figur 3.2-2. Settet vert delt opp fire under sett; TAG1-4




*Figur 3.2-2: TIC av TAG-settet.*




*Figur 3.2-3: TIC av TAG1, TAG2, TAG3 og TAG4*

Settet vart teke opp på LC-MS av typen Agilent 1100 series LC/MSD trap, SL model, med ESI-ionisering.

*Tabell 3-3 : MS-opptaktsparameterar for TAG –settet*

| | |
|---|---|
| Analysator | Ion Trap QqQ |
| Ionekjelde | ESI+ |
| Smart View med oppløysing (m/z/sec) | 13000 (FWHM/m/z = 0.6 - 0.7) |
| Capillary exit voltage (V) | Kontrollert av Smart View-funksjon |
| Skimmer voltage (V) | |
| Lens voltage (V) | |
| Octapole voltages (V) | |

*Tabell 3-4: LC-gradientprogram for TAG-settet*

| tid [min] | %A | %B | %C |
|---|---|---|---|
| 0 | 90 | 0 | 10 |
| 5 | 65 | 30 | 5 |
| 20 | 90 | 0 | 10 |
| 25 | 65 | 30 | 5 |
| 55 | 90 | 0 | 10 |
| | | | |
| A = isopropanol:ammoniumacetat 90:10 [V/V]<br>B = aceton<br>C = acetonitril<br>Flow = 0.2 mL/min<br>Deteksjons-bølgjelengd = 254 nm<br>Temp. = 40 °C | | | |

Av settet er det vald ut to tidsområder:

*Tabell 3-5: Utvalde tidsområder av settet TAG*

| Område | Plassering i datasett (ret. Pkt.) | Område-klassifisering |
|---|---|---|
| Q | 750 til 839 | 1 toppklynge |
| K | 600 til 700 | 2 toppklynger |

*Tabell 3-6 : Oppdeling i undersett av TAG – etter kjente komponentar*

| TAG-område | frå (ret. pkt.) | til (ret. pkt.) | frå (min) | til (min) |
|---|---|---|---|---|
| TAG1 | 520 | 570 | 24,66 | 27,03 |
| TAG2 | 605 | 645 | 28,7 | 30,6 |
| TAG3 | 645 | 690 | 30,6 | 32,73 |
| TAG4 | 755 | 800 | 35,82 | 37,96 |

TAG-forbindelsane vart ionisert i ESI og kjem derfor på formen $[M + NH_4]^+$ og $[M + Na]^+$, der M er forbindelsen. I Tabell 3-7 er det definert fire områder i TAG-settet, der alle inneheld kjent TAG-signal. Datasetta inneheld signal for TAG med masse som M, men på forma $[M+NH_4]^+$ og $[M+Na]^+$

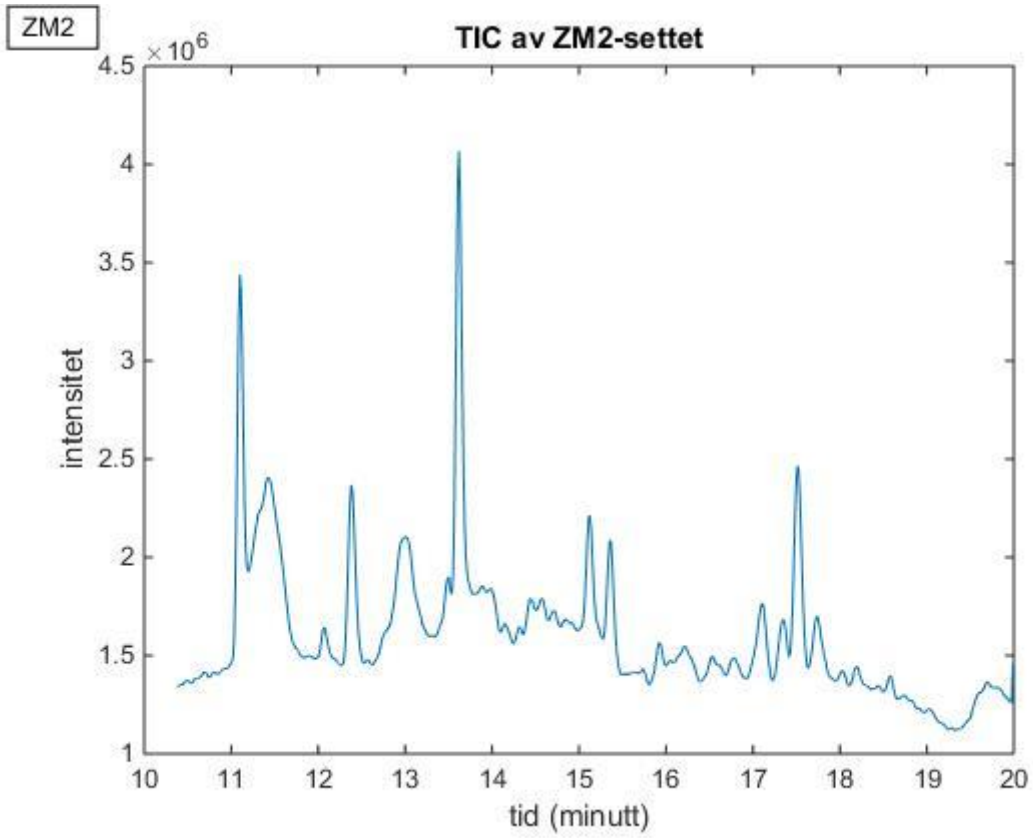
*Tabell 3-7: Oversikt over TAG-forbindelsar, og m/z-verdiar for moderion og forløperion, i TAG-områda i TAG-settet*

| TAG-område | M+ (m/z) | $[M+NH_4]^+$ (m/z) | $[M+Na]^+$ (m/z) |
|---|---|---|---|
| TAG1 | 913 | 931 | 936 |
| TAG2 | 941 | 959 | 964 |
| TAG3 | 889 | 907 | 912 |
| TAG4 | 917 | 935 | 940 |

*Tabelltekst til Tabell 3-7: Områda inneheld ein TAG-komponent kvar, utanom TAG3 som inneheld to komponentar, som får lik m/z-verdi ved den aktuelle oppløysinga. TAG4 inneheld i utgangspunktet .., men pga. binning [] viser massen … i staden.*

#### *3.2.1.4 ZM2-settet*

Settet (med originalnamn «ZM2, xad 10-16, Prep 1-11_17.02.2015_ flavo90_50_2_2») er lånt av Ph.D.-kanditat Kjersti Enerstvedt, ved Kjemisk Institutt UiB. Den analyserte prøva er eit oppreinska ekstrakt av eit flavanoid frå ei marin kjelde, med nokre forureiningar. TIC av settet er plotta i Figur 3.2-4. Flavanoidet har masse 287 (m/z-verdi) og har retensjonstida 11.1 minutt. Toppane som er tilstades ved dei andre retensjonstidene stammar antakeleg frå forureiningar i prøven eller i løysemiddelet.




*Figur 3.2-4: TIC representasjon av ZM2-settet.*

ZM2-settet vart teke opp på eit LC-MS-instrument som består av ein Agilent 1200 series LC modul og eit Agilent 6420A QqQ massespektrometer i MS2 modus, som nyttar ESI-ionisering. Operativsystemet er Mass Hunter Workstation frå Agilent. Kolonna som er nytta er av typen Agilent ZORBAX SB-C18, RRHT; 2.1 x 50 mm, 1.8 µm.

*Tabell 3-8: LC-gradientprogram for ZM2-settet*

| tid [min] | %B |
|---|---|
| 0 | 10 |
| 1 | 10 |
| 2 | 12 |
| 3 | 15 |
| 6 | 20 |
| 8 | 35 |
| 10 | 45 |
| 12 | 50 |
| 15 | 60 |
| 18 | 10 |
| 20 | 10 |
| | |
| A = H2O + 1% HCOOH<br>B = ACN<br>Flow = 0,3 mL/min<br>Inititaltrykk = ca. 200 bar | |

*Tabell 3-9: MS-opptaktsparameterar for ZM2 –settet*

| | |
|---|---|
| Ioniseringskjelde | ESI+ |
| Fragmentor (V) | 80 |
| Nålespenning (V) | 4000 |
| Masseområde (m/z) | 100-800 |

#### *3.2.1.5 HCO-settet*

Settet [55] er ein til dels oppreinska fraksjon av eit råekstrakt, av blad av planta Hemigraphis Colorata. Råekstraktet som prøven til HCO-settet er teken ut av, har tre definerte kjente forbindelsar, som er antocyaniner. Desse vert presenter i tabell 1 i artikkelen «Purple anthocyanin colouration on lower (abaxilal) leaf surface of Hemigraphis colorata (Acanthaceae)» til Skaar et al. [55]. Dei tre toppane (**2**, **1**, **3**) som vart funne i artikkelen vert i Tabell 3-10 (under) gjeve med plassering for HCO settet og lik nummerering som i artikkelen. Ettersom det er desse tre forbindelsane ein har kunnskap om i settet, vil dei frå no av verta referert til som kvalitetstoppar.

*Tabell 3-10: Kvalitetstoppar for 2012-utgåve av HCO-settet*

| Forbindelse | Toppunkt (min) | M+ (m/z) | F+ (m/z) |
|---|---|---|---|
| **2** | 8,9 | 463,1203 | 301,0659 |
| **1** | 9,4 | 639,1572 | 301,0624 |
| **3** | 10,6 | 653,1721 | 301,0633 |

*Tabelltekst for Tabell 3-10: Toppunkt, start- og slutt- elusjonstider, funnen masse av moderiona (M+) og eit fragment (F+) for kvar forbindelse ut frå artikkelen. [55] (Dette vil ikkje sei at det berre er eit fragment for kvar forbindelse.)*

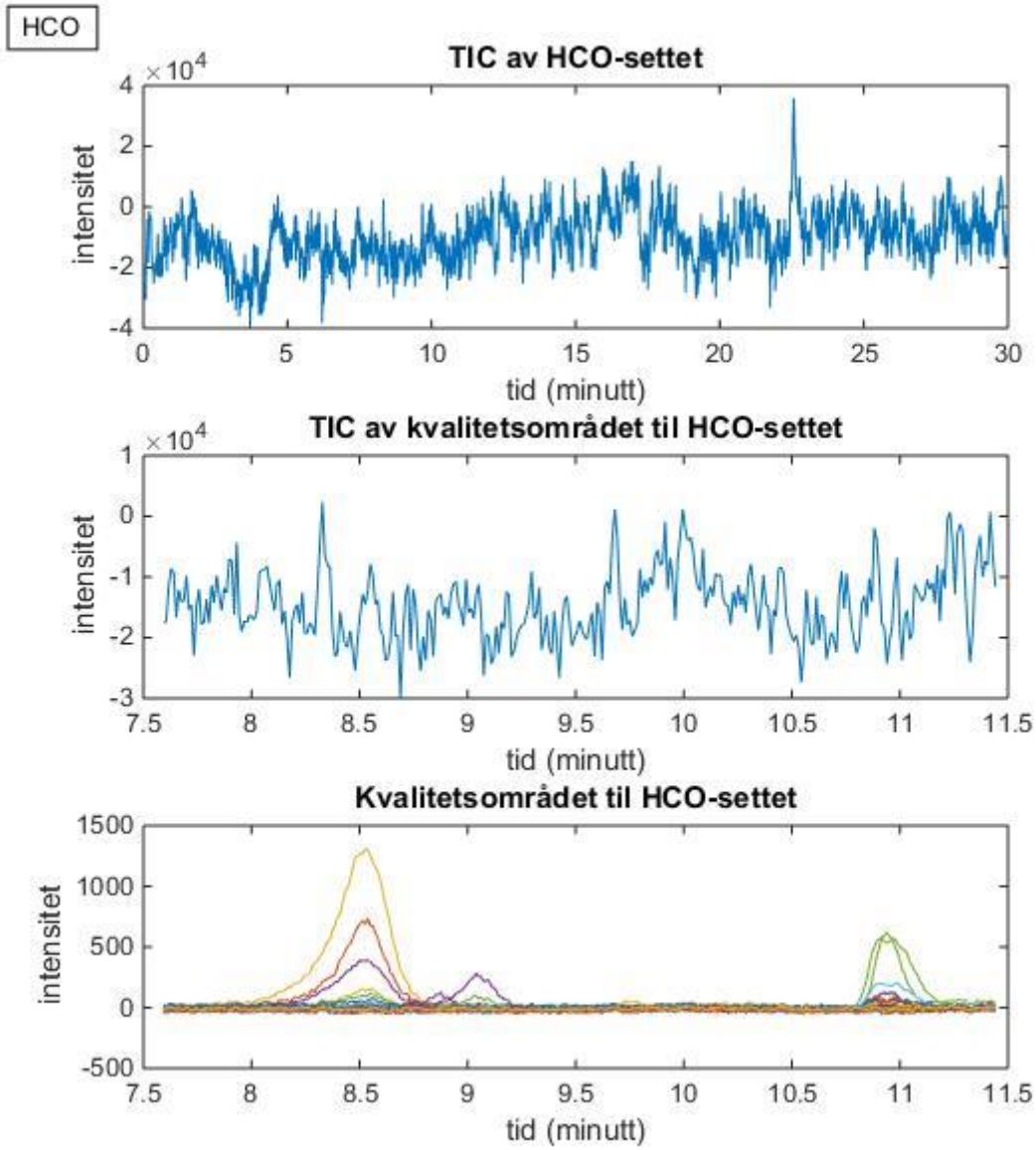
Det originale HCO-settet, som er nytta i artikkelen til Skaar et al., vart teke opp i 2012, men dette var ikkje lenger lagra digitalt. Ein meir oppreinska fraksjon av ekstraktet vart analysert i 2013 på same instrument, med like instrument-innstillingar. Resultata for kvalitetstoppane er gjevne i Tabell 3-11. For det nye settet har retensjonstidene forandra seg litt, men dei originale moderiona til toppane **2** og **1**, samt fragmentet til **1** er framleis tilstades. Fordi forbindelse **3** i 2012-settet vart danna under opparbeidinga av prøvane, og at det finnest ein fjerde topp etter topp **3** i begge setta som gjer utsjånaden til dei to kvalitetsområda svært lik, vert det her antatt at topp **3** i 2012-settet er danna av topp **3** i 2013-settet. Topp **2**, **1** og **3** vil frå no av referera til toppane/forbindelsane i 2013-settet, ettersom det er desse som er nytta.

*Tabell 3-11: Kvalitetstoppar for 2013-utgåve av HCO-settet*

| Forbindelse | Toppunkt (min) | M+ (m/z) | F+ (m/z) |
|---|---|---|---|
| **2** | 8,5 | 463 | 301 |
| **1** | 9,1 | 639 | |
| **3** | 11,0 | 647 | 325 |

*Tabelltekst til Tabell 3-11: Settet har oppløysing 1 pga. binning av massar. Det er denne utgåva av HCO-settet som vert nytta i arbeidet.*

I Figur 3.2-5 er TIC av heile settet og kvalitetsområdet plotta, i tillegg til alle EIC til kvalitetsområdet. Det framgår av plotta at settet inneheld mykje støy.




*Figur 3.2-5: HCO-settet (2013-versjon) ; (øvst) TIC av settet, (midtarst) TIC av kvalitetsområdet og (nedst) EIC av kvalitetsområdet. Plottet er av settet etter binning og baselinjesubtraksjon.*

HCO-settet vart teke opp ved eit AccuTOF T100LC –instrument - med operativsystemet MassCenter ( v.1.3.4) med Datamanager. Kolonna var ei Agilent Zolvax Eclipse SB18 50mm x 2.1 mm x 1.8 µm Tabell 3-12 og Tabell 3-13 er relevante opptaksparameterar for respektivt LC og MS til settet oppgjeve.

*Tabell 3-12 LC-gradientprogram for HCO-settet*

| min | %B |
|---|---|
| 0 | 10 |
| 1 | 10 |
| 3 | 18 |
| 6 | 18 |
| 8 | 35 |
| 13,75 | 45 |
| 14,5 | 65 |
| 20,75 | 65 |
| | |
| A = H2O (+ 0.1 % TFA)<br>B = ACN (+ 0.1 % TFA)<br>Flow = 0.2 mL/min<br>Inititaltrykk = 208 bar<br>Temp = 25 °C | |

*Tabell 3-13 : MS-opptaksparameterar for HCO-settet*

| | |
|---|---|
| Ionization Mode | ESI+ |
| Needle Volt (V) | 1000 |
| Desolvating Chamber Temp (°C) | 250 |
| Orifice1 Volt (V) | 80 |
| Orifice2 Volt (V) | 5 |
| Ring Lens Volt (V) | 10 |
| Orifice1 Temp (°C) | 80 |
| Ion Guide RF Volt (V) | 2000 |
| Ion Guide Bias Volt (V) | 26 |
| Pusher Bias Volt (V) | -0,79 |
| Reflectron Volt (V) | 970 |
| Detector Volt (V) | 2300 |

## 3.3 MATLAB-programmering

I dette kapittelet vert programma som vart programmert i arbeidet presentert. Presentasjon av metodane som programma utførar kjem i kapittel 5.

I dette arbeidet vart programma i Tabell 3-14 nytta. Alle programma er programmert i forbindelse med arbeidet, med unntak av *GAUSSGEN*, *LINTRANS* og *binning* som er skrivne av Bjørn Grung. Totalt, forutan dei lånte programma, består programma av ca. 1300 linjer

Programma er tilgjengelige frå fylgjande URL-adresse:

https://www.dropbox.com/sh/qaeme9r3qfmcexd/AAD5Hbm_l53-DGfXnsXDR9_Za?dl=0

For å køyra programma må ein ha tilgang til MATLAB sin Wavelet Toolbox 4.14

Blant funksjonane i toolboxen er *ndwt*, *wpt* og *cwt*. For *ndwt* og dei andre diskrete WT er ikkje Mexican hat waveleten implementert i MATLAB.

*cwt*-funksjonen i MATLAB kan gje CWT for diskrete data. Dette går ein ut ifrå er ein approksimasjon, ettersom definisjonen på CWT er at analysedata (x) og skaleringsparameterar (a og b) er kontinuerlige. [35]

*Tabell 3-14: Program nytta i oppgåva*

| Namn på program | Namn i dokumentet |
|---|---|
| **CODA-versjonar** | |
| CODA | |
| CODA_WT2 | CODA_ndWT |
| CODA_WT3 | CODA_CWT |
| CODA_WPT | |
| CODA_WPT2 | |
| CODA_WPTlim | |
| CODA_slichalvdyn_WT2 | CODA_SHD_ndWT |
| | |
| **Hjelpeprogram til CODA** | |
| sim_MCQ | |
| CODA_slichalvdyn | CODA_SHD |
| CODA_null | |
| WT | |
| iWT | |
| TIC | |
| WPT_basis | |
| WPT_matrix | |
| CWT | |
| thresh_h | |
| | |
| **Kvalitetsprogram og hjelpeprogram** | |
| singular_fraction_2 | singulærverdiforhold |
| kvalitet_korr_mult_topp_2 | korrelasjonbasert kvalitetsmål |
| kvalitet_korrelasjon_2 | |
| kvalitet_euNorm_mult_topp_2 | normbasert kvalitetsmål |
| kvalitet_euNorm_2 | |
| finn_toppar_2 | |
| k_storste | |
| kol_normalisering | |
| finn_indeks | |
| | |
| **Konstruksjon av datasett** | |
| make_dataset3 | |
| 5_basis_norm_spektra.mat | |
| make_XCMS_dataset | |
| noise_generator | |
| GAUSSGEN | |
| LINTRANS | |
| | |
| **Binning** | |
| binning | |

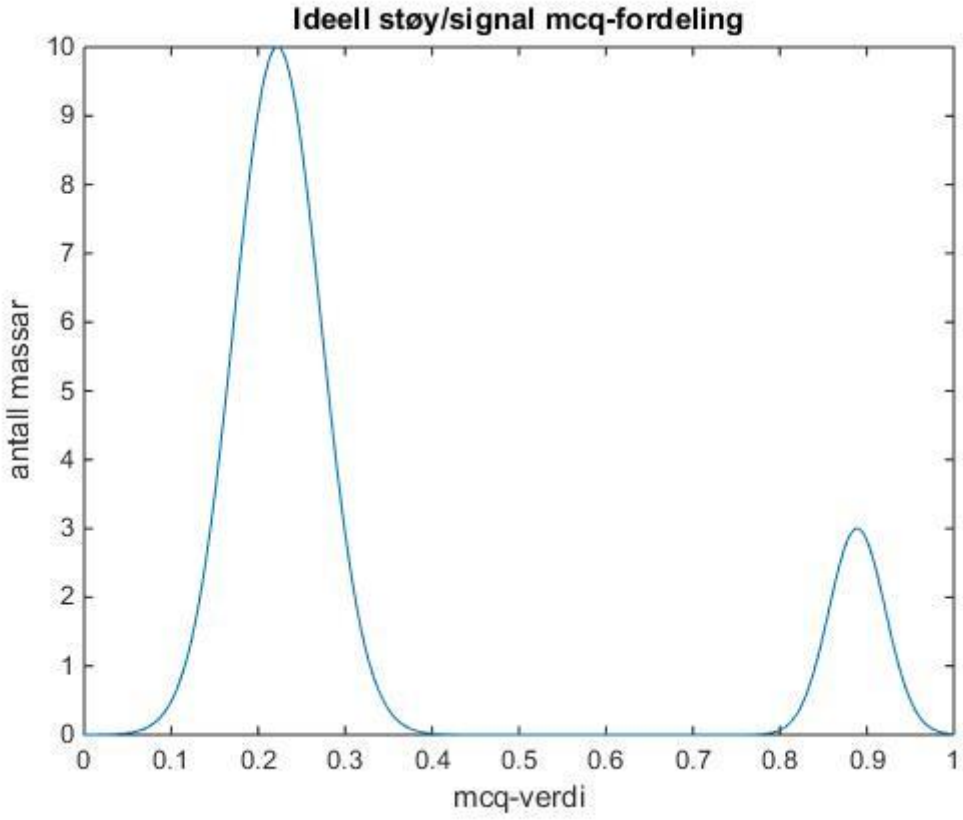
*Tabelltekst til* Tabell 3-14*: Programma er ei samling av funksjonar og skript. «Namn i dokumentet» er berre oppgjeve dersom metoden er nemnd i dokumentet og programmet har eit anna namn. 5_basis_norm_spektra.mat er ei fil med 5 massespektra, som er nytta til å konstruera datasett.*

Programmet sim_MCQ vart nytta til å rekna ut mcq-verdiar til alle CODA-versjonane. Det har i etterkant av arbeidet vorte avdekka ein feil i programmet, som gjer kvart punkt i den glatta matrisa frå Formel 2.2-3 består av snittet av seg sjølv og dei w-1 neste punkta, i staden for snittet av w verdiar med punktet sjølv som sentrum.

# 4 Fordeling av massekromatografiske kvalitetsindeksar (mcq-verdiar)

## 4.1 Mcq-fordeling i ideelle sett og reelle sett

Som tidligare nemnd reknar CODA ut ein mcq-verdi (mass chromatografic quality), for kvar masse i eit LC-MS datasett, og forkastar så massar under ei satt mcq-grense. Fordeling av mcq-verdiar til eit datasett vil då optimalt sett ha eit klart skilje mellom høg-kvalitetsmassar og låg-kvalitetsmassar, slik at ein lett kan fastsetja ei mcq-grense. I Figur 4.1-1 er mcq-fordelinga for eit «ideellt» sett gjeve.

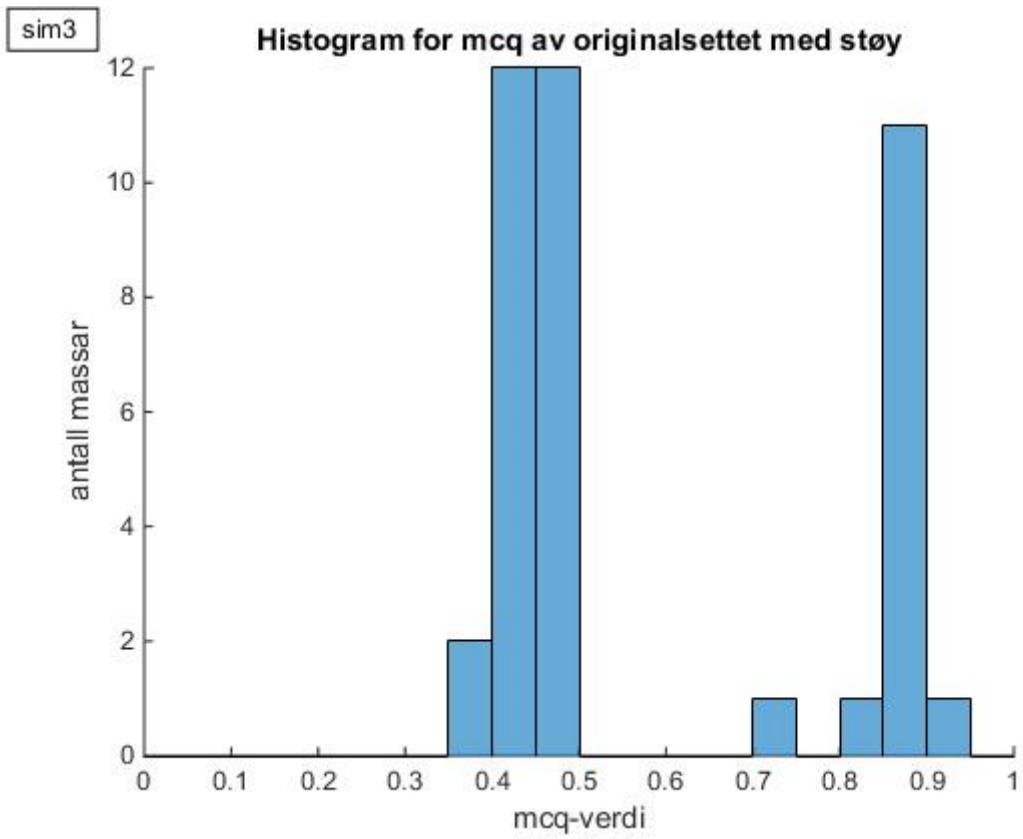



*Figur 4.1-1: «Ideell» fordelingskurve for mcq-verdiar for eit ikkje-signal / signal datasett . Då reknar ein med at signal vil få høge mcq-verdiar og områder med lite kjemisk signal få låge mcq-verdiar.*

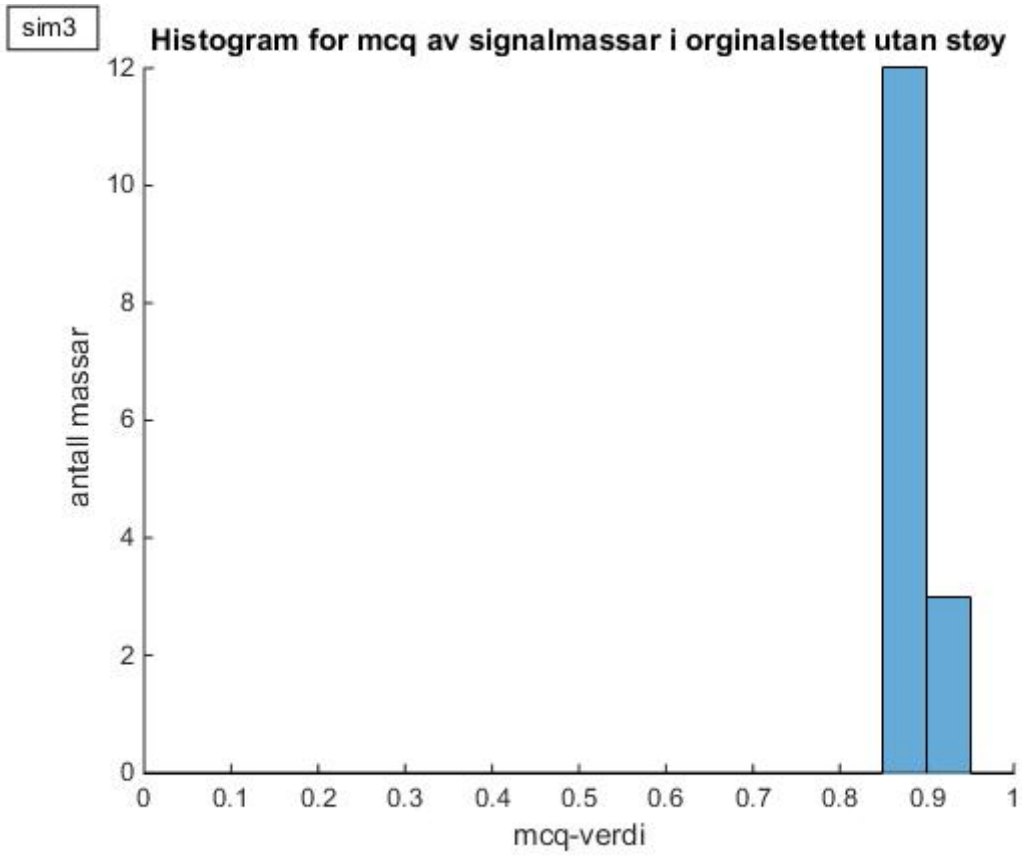
I figuren over held støyen held seg fint separert frå det faktiske kjemiske signalet, som held seg lengre oppe på skalaen. Eksempelet er i mange tilfelle ikkje sant for faktisk analysedata,

fordi massekromatogramma av låg kvalitet ofte ligg nærmare massekromatogramma av høg kvalitet, som gjer det vanskeleg å velja ut ei mcq-grense for CODA.

Det simulerte settet sim3 sine mcq-verdiar er plotta i Figur 4.1-2. Settet har berre 40 massar der 14 av desse er analytisk signal, i tillegg til ein signalmasse under støygrensa. Det ser ut til å vera eit skilje mellom signal og støy i fordelinga. Ut frå fordelinga til sim3 utan støy som er vist i Figur 4.1-3, ser det ut som om alle massane i M over mcq = 0.5 er signal, medan alle under, med unntak av ein, er støymassar.

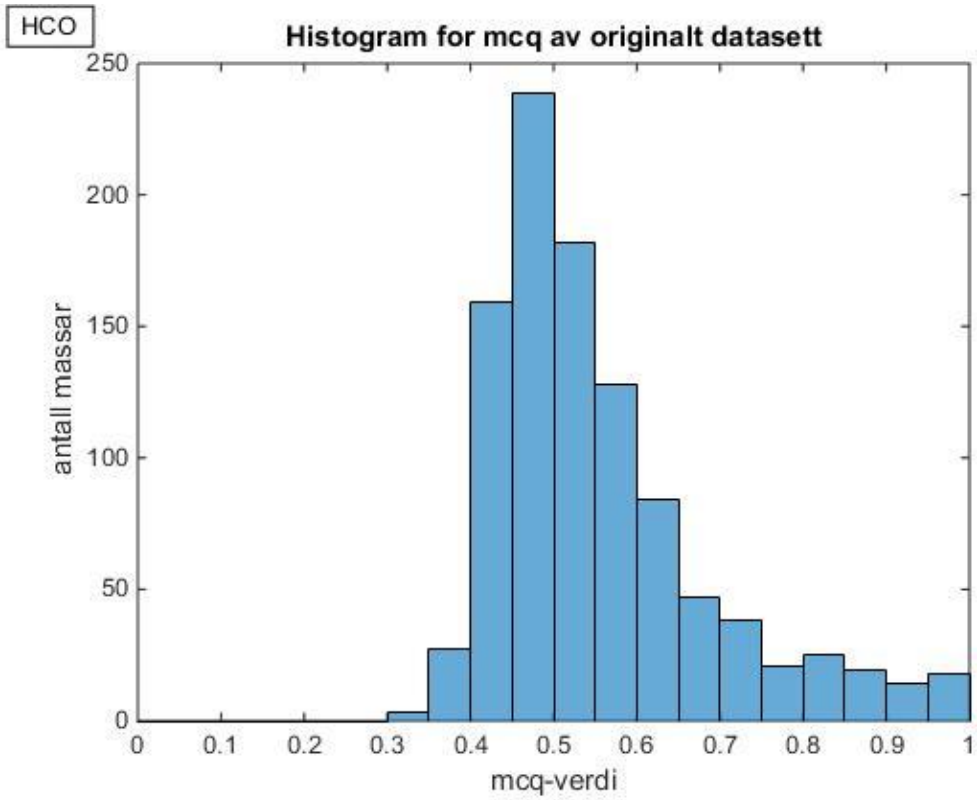



*Figur 4.1-2: Histogram av mcq-verdi-fordelinga til datasettet sim3, der mcq er berekna for det originale settet (med pålagt støy).*



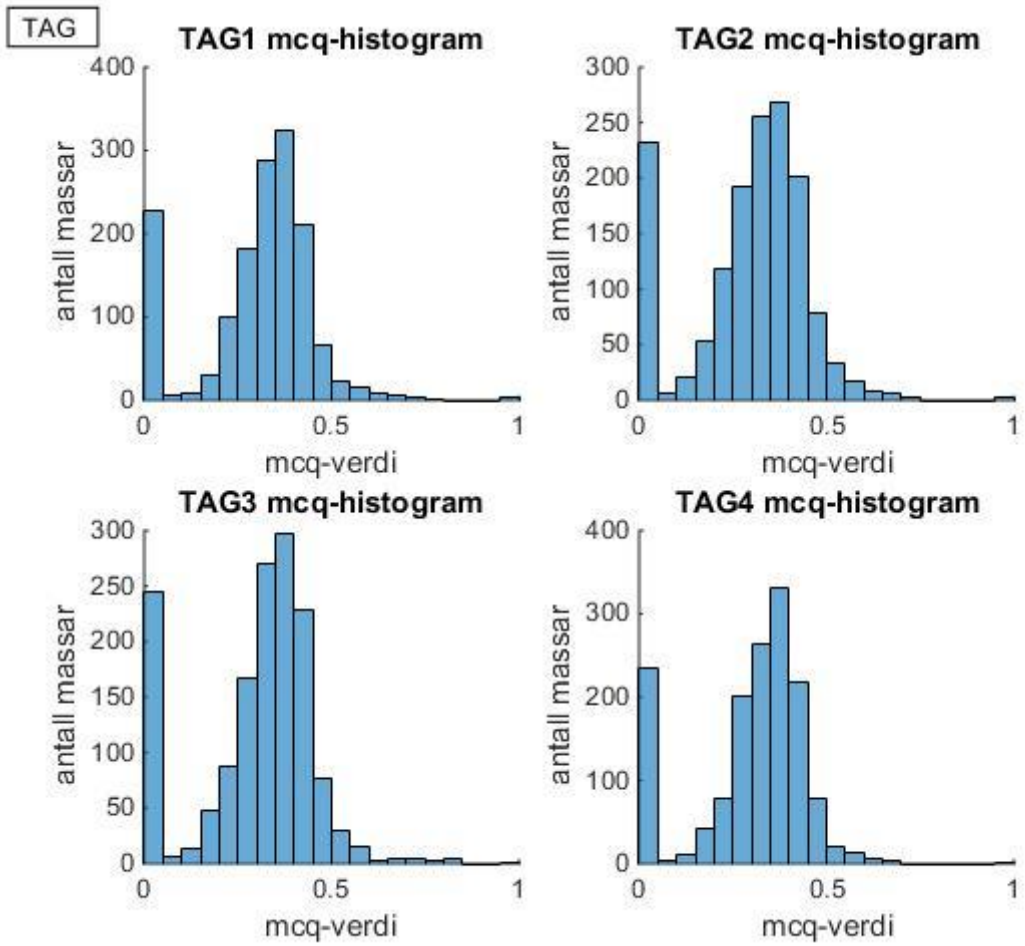

*Figur 4.1-3: Histogram av mcq-verdi-fordelinga til datasettet sim3, der mcq er berekna for signalmassar i det originale settet utan pålagt støy*

Fordelinga av dei 1004-massane til HCO-settet, i Figur 4.1-4, visar få teikn til kva mcq-grensa går, men det kan tenkast at støyen er normalfordelt rundt mcq = 0.48 og alt over 0.7 er signal. Dette synleggjer behovet for eit mål for kvar grensa går, i større grad enn for sim3-settet.




*Figur 4.1-4: Histogram av mcq-verdi-fordelinga til datasettet HCO*

Mcq-fordelingane til dei fire TAG-setta, i Figur 4.1-5, viser tilnærma normalfordeling, som stemmer godt med at setta inneheld relativt få TAG-massar i forhold til antall massar.

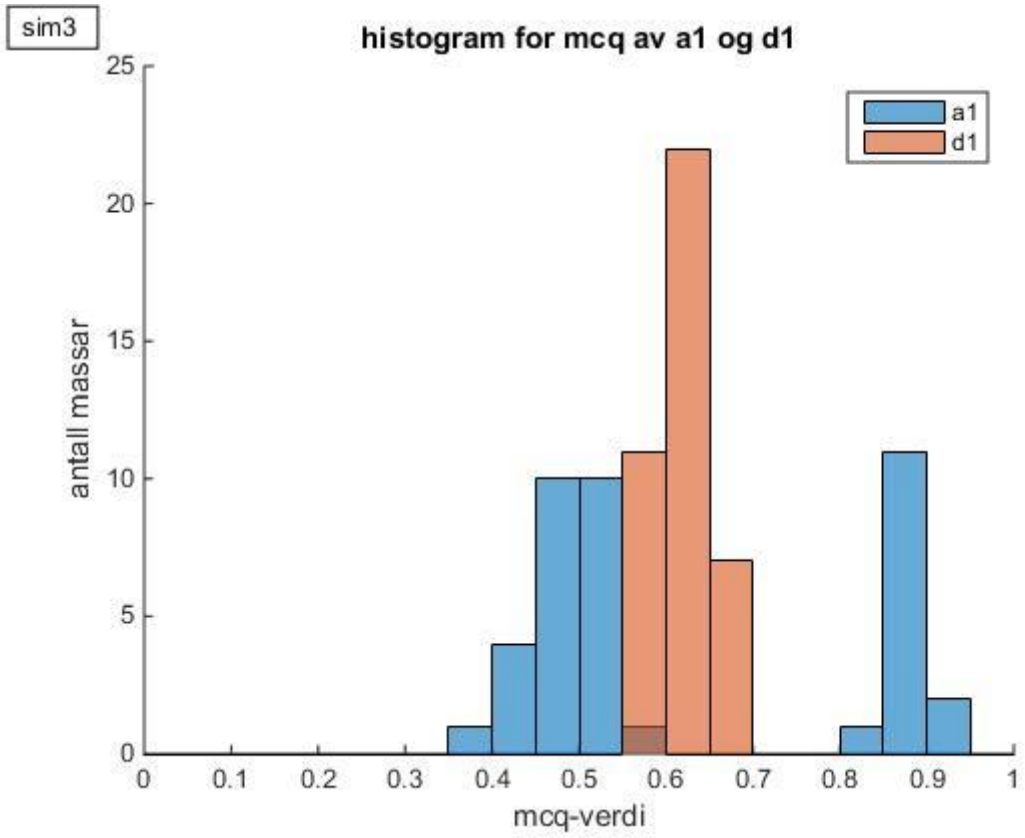



*Figur 4.1-5: Histogram av mcq-verdi-fordeling i TAG1, TAG2, TAG3 og TAG4*

ZM2-settet sine mcq-fordelingar er vist i Vedlegg 1. Settet ser ut til å ha fleire fordelingar.

## 4.2 Mcq-fordeling i wavelettransformerte sett

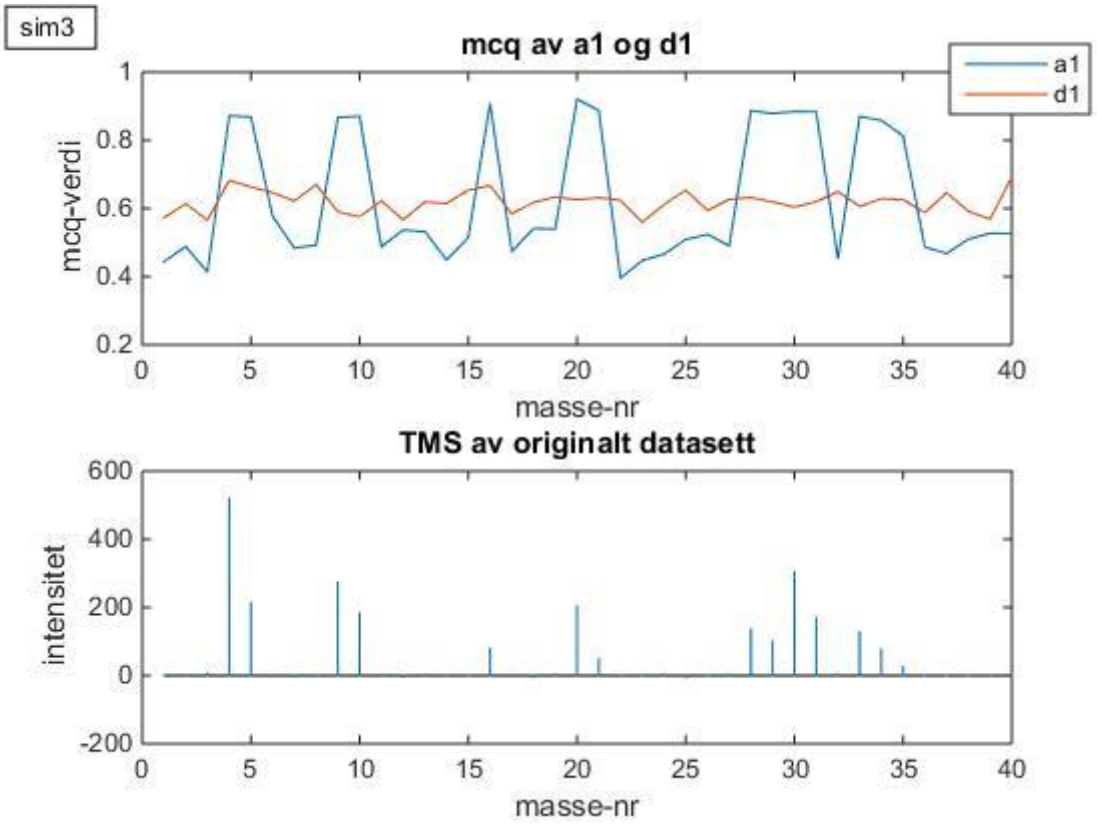
Ein forsøker her å sjå om wavelettransformasjon (WT) kan gje ekstra informasjon om kvar mcq-grensa bør gå. For wavelettransformasjonane i dette kapittelet vert det nytta udesimert WT med dekomponeringsnivå 1 og symlet 5 waveleten som moderwavelet. Udesimerte WT dekomponerer massekromatogramma til komponentar med lik lengd som massekromatogramma, som nemnd i teorien. Ved dekomponering til a1 og d1 vert det gått ut frå at a1 vil innehalda den kjemiske informasjonen og at d1 berre inneheld stokastisk støy. Alle mcq-verdiar er rekna ut med vindaugebreidd (w frå Formel 2.2-3) lik 5. Ein ser fyrst på sim3-settet, i Figur 4.2-1.




*Figur 4.2-1: Histogram over fordeling av mcq-verdiar for a1 og d1 av det simulerte datasettet sim3.*

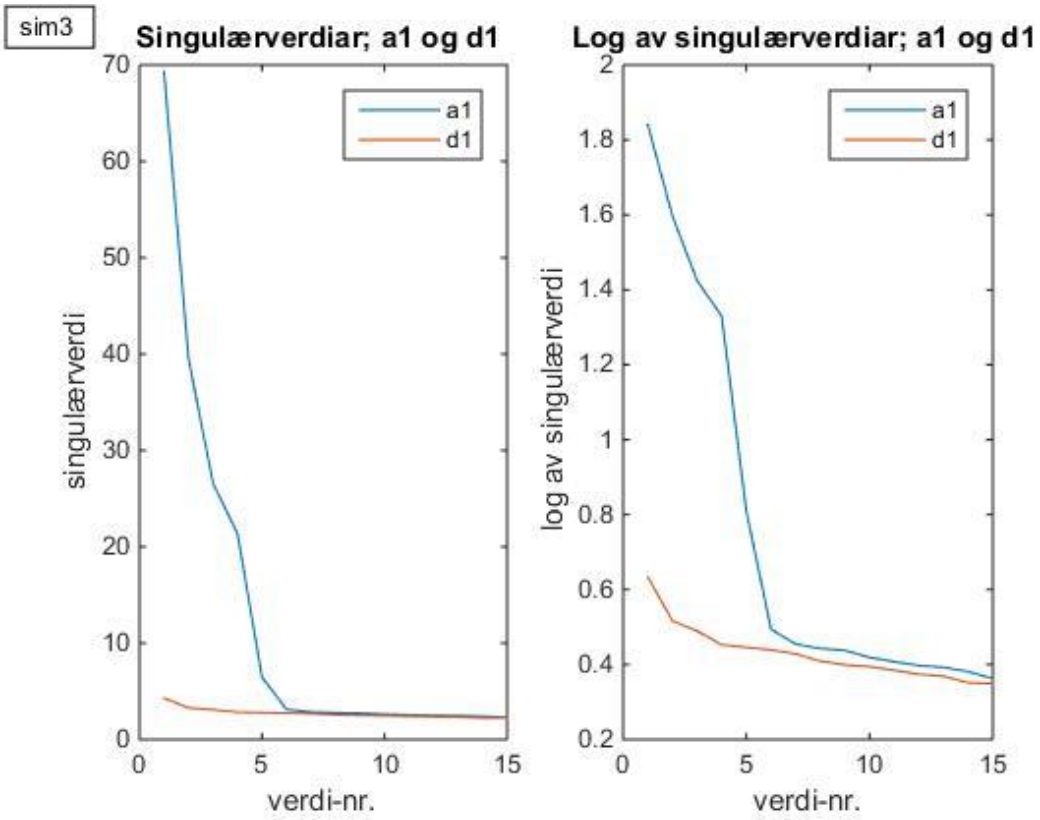
Av histogrammet ser ein at det er 14 a1-massar som har høgare mcq-verdi enn d1. Det er vert då naturleg å gå ut frå at desse er signal, viss ein ser fordelinga i samanheng med histogrammet til originalsettet utan støy (i Figur 4.1-3). I histogrammet av det støyfrie settet er det 15 massar med høg mcq-verdi, så det støykvelte signalet i signalmasse nr. 15 er framleis kvelt i a1. Det er viktig å hugsa på at alt høgfrekvent data ikkje er fjerna i a1. Ved fleire dekomponeringar (dvs. d2, d3, osv) vert det fjerna meir støy, som kunne vert gunstig for den siste massen, men det kan også gå på bekostning av det analytiske signalet i settet.

Figur 4.2-2 består både eit mcq-plott og TMS av sim3, og bekreftar at a1 har høgare mcq-verdi enn d1 ved signaltoppane. Det kjem også fram at d1 ikkje har høgare mcq ved signaltoppane enn elles, som indikerer at a1 inneheld den kjemiske informasjonen.



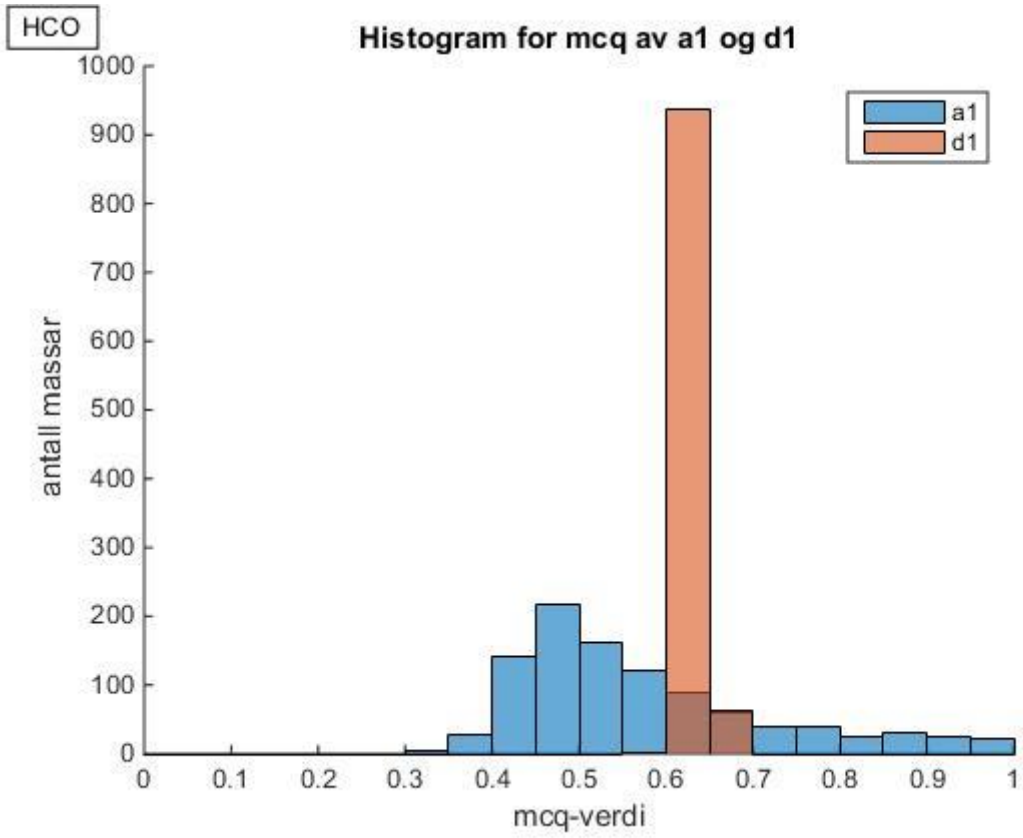

*Figur 4.2-2: Fordeling av mcq-verdiar for a1 og d1, ut frå plassering , av det simulerte datasettet sim3*

Av singulærverdi-plotta i Figur 4.2-3 viser at det er mellom 4 og 5 kjemiske komponentar (kjemisk rang) i a1 og ingen i d1. Ettersom den fyrste komponenten i sim3-settet i stor grad overlappar med den andre komponenten, gjer det meining at det er 4 store singulærverdiar og ein mellomstor.



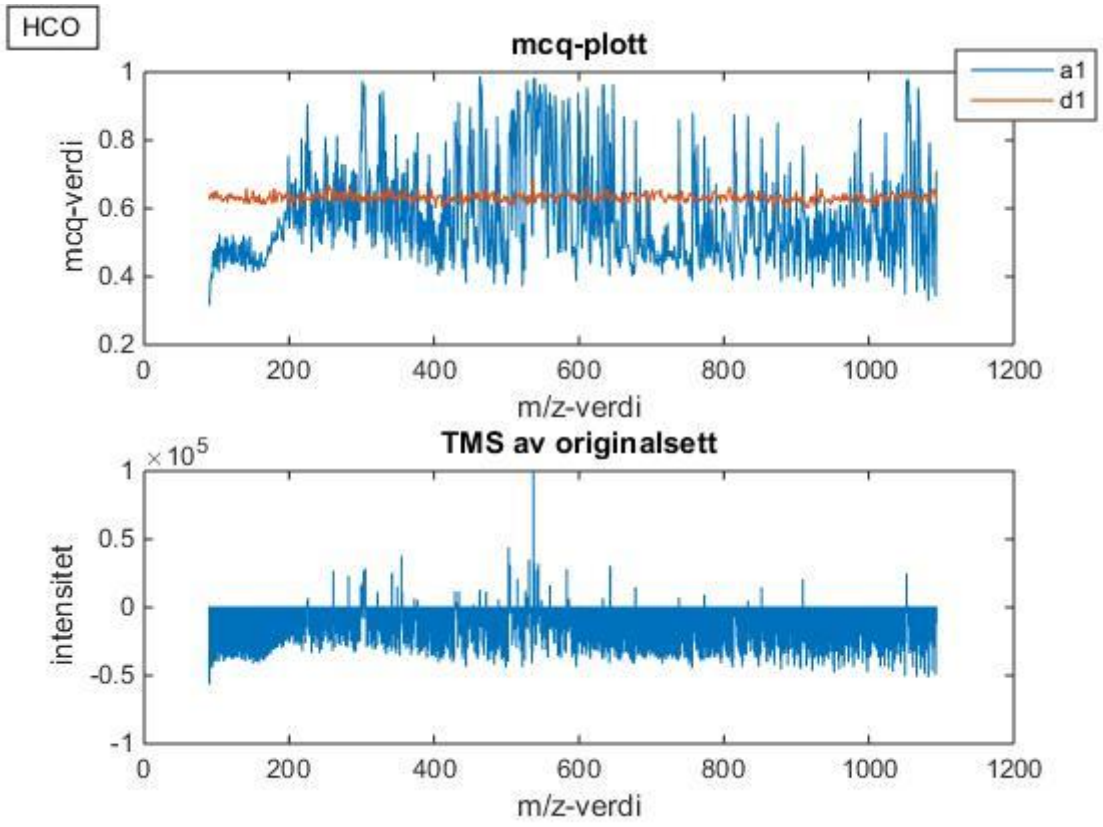

*Figur 4.2-3: Dei 15 største singulærverdiane og logaritmen av dei, for a1 og d1 av sim3-settet.*

For mcq-verdiane av det dekomponerte HCO-settet, i Figur 4.2-4, kan ein sjå at d1 har eit smalt band av mcq-verdiar rett under 0.7. Teorien frå Figur 4.1-4 om at støy-kromatogramma i HCO er tilnærma normalfordelte på mcq-aksen, vert vidare bekrefta.




*Figur 4.2-4: Histogram av mcq-verdiar for a1 og d1 av HCO-settet*

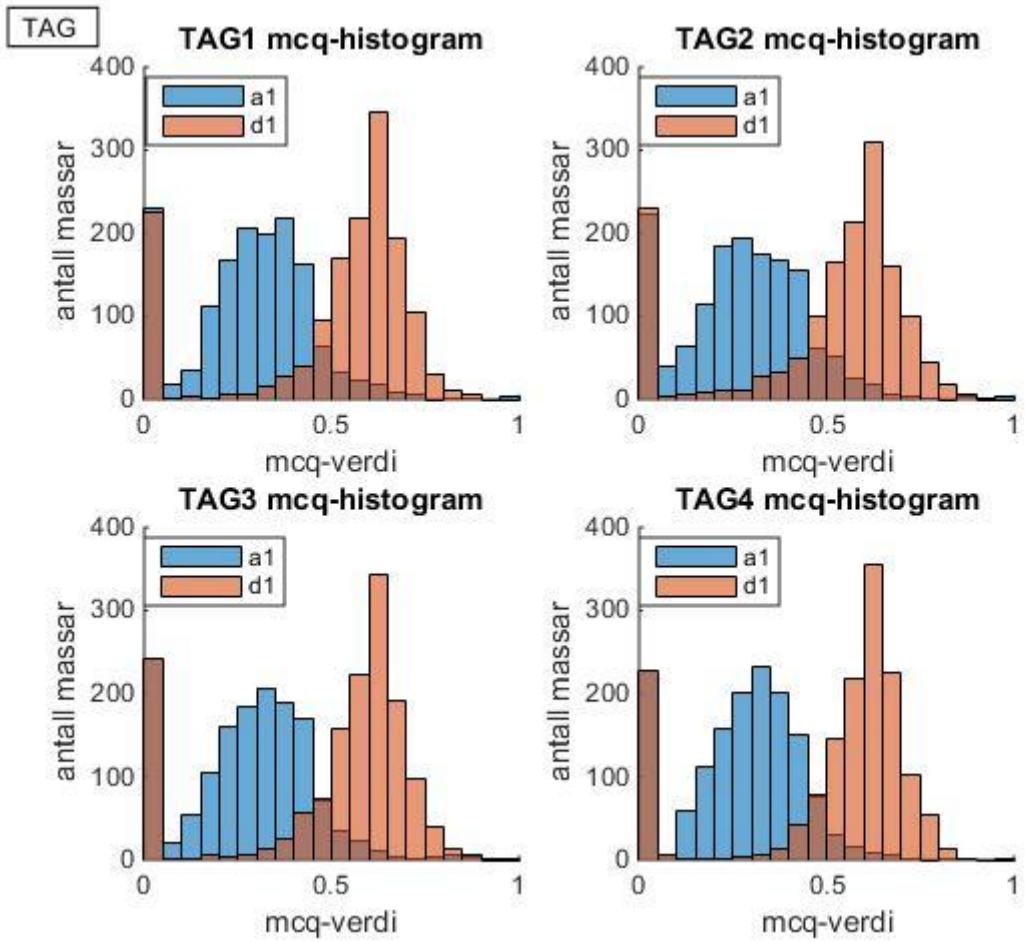
Plottet av mcq-verdiar av a1 og d1 mot TMS, i Figur 4.2-5, gjer lite ekstra informasjon, pga. størrelsen på settet og dei store intensitetsforskjellane.




*Figur 4.2-5: Plot av mcq-verdi mot massar, og massespektra for HCO-settet.*

I Vedlegg 3 er det lagt ved eit plott av dei 30 største singulærverdiane til HCO-settet. Av plottet ser det ut som om settet har kjemisk rang 9. Dei 18 største singulærverdiane til a1 og d1 er plotta i Vedlegg 4. Av det sistnemnde plotte kan det sjå ut som at a1 har rang 9 og d1 har rang 1. d1 sin signifikante singulærverdi er på storleik med den minste signifikante singulærverdien til a1. Ettersom a1 ikkje har miste rang og mcq-verdiane til d1 ligg i eit smalt sjikt, er det mogleg d1 sin singulærverdi har opphav i heteroskedastisk støy.

Figur 4.2-6 viser histogram over mcq-verdiane til a1 og d1 ledda til TAG1-TAG4-setta. Alle histogramma viser den same trenden, ved at nokre få a1-massar har høgare mcq-verdi enn alle d1-massane. Trenden stemmer godt overeins med at det her er vald ut spesifikke TAG-toppar og at massespektera er store.




*Figur 4.2-6: Histogram av mcq-verdi-fordeling til a1 og d1 av TAG1, TAG2, TAG3 og TAG4. Alle histogramma har ein liten blå topp nær mcq = 1, som er lite synleg pga. storleiken.*

Fordelinga av mcq-verdiane til a1 og d1 av ZM2 er vist i Vedlegg 2. ZM2 er eit komplekst sett, med fleire fordelingar, men det ser ut som om den eine a1-fordelinga er lågare enn toppunktet til d1-fordelinga og det er også nokre verdiar høgare enn d1.

# 5 Nye CODA-versjonar

I dette arbeidet har det vorte utvikla fleire nye versjonar av CODA. Målet er å automatisk fastsetja ei ideell mcq-grense for eit datasett. Fokuset er på å fjerna støy utan å fjerna noko signal, som vil sei at ein heller beheld litt støy enn å risikera å fjerna signal. For å fastsetja mcq-grensa nyttar CODA-versjonane ulike wavelettransformasjonar til å estimera støyen. Alle

metodane med unntak av CODA_WPTlim utførar i tillegg ein wavelettransformasjon på kromatogramma, før dei vert behandla med CODA, som har som hensikt å fjerna støy.

## 5.1 CODA_ndWT

Metoden er eit forsøk på automatisering av CODA-algoritmen, og har i tillegg eit støyfjerningssteg. Algoritmen startar med å dekomponera kvart kromatogram med ndWT til fyrste nivå, a1 og d1, av same lengd som dei originale kromatogramma. Mcq-indeksane vert så funne for både a1 og d1. Detaljleddet (d1) vert nytta som eit estimat for støy, ettersom stokastisk støy ofte vert karakterisert som spisse, høgfrekvente toppar. Metoden har fire måtar å fastsetja mcq-grensa på, der alle er ut frå mcq-verdiane til støyestimatet. Mcq-grensa kan verta sett til å vera snitt av medianen og minimum, medianen, snitt av medianen og maksimum eller maksimum av mcq-vektoren til støyestimatet. Denne mcq-grensa vert nytta i den klassiske delen av CODA-algoritmen. Det er det approksimerte leddet a1 som vert køyrt gjennom CODA, ettersom d1 vert rekna som støy. Det CODA_ndWT behandla settet består då av a1-kromatogramma som har mcq-verdi høgare eller lik mcq-grensa.

## 5.2 CODA_CWT

Metoden nyttar kontinuerleg WT (CWT) i kombinasjon med CODA til å velja ut høgkvalitets-kromatogram. CWT vert, som nemnd i teorien, vanlegvis nytta for kontinuerlege data, men ettersom det finnest ein MATLAB-metode (sjå kap. 3.3) for diskrete data vert denne nytta. Det vert nytta ein type wavelet, med ein tilhøyrande skaleringsfaktor, til å approksimera støy og ein wavelet kombinert med ein gjeven skaleringsfaktor (kjent som «a» i frå teori-kapittelet) til å approksimera signalet. Dette vert gjort for kvart kromatogram i datasettet. På same måte som for CODA_ndWT vert støyapproksimasjonane nytta til å finna ei mcq-grense for ei CODA-utveljing av massar frå dei approksimerte kromatogramma. Ettersom CWT-behandla signal kan avvika i intensitet frå originaldata, vert det prosesserte datasettet skalert slik at den høgste toppen har lik intensitet som den høgste i originalsettet. Metoden nullsett også alle negative verdiar i det prosesserte settet, ettersom kromatografiske toppar er positive av natur og at CWT er ekstra utsett for å få negative verdiar.

## 5.3 CODA_WPT

CODA_WPT vert laga for å få ein meir skånsam metode enn CODA_ndWT og CODA_CWT. I denne metoden vert WPT (Wavelet Packet Transformasjon) nytta for å dekomponera datamatrisa, M. Brukaren kan sjølv velja kor mange nivå M skal dekomponerast til. Den beste basisen for datasettet vert estimert ut frå TIC-vektoren, for eit antall trulege wavelets, ved hjelp av eit entropikriterium. Entropikriteriet baserar seg på at data med høgt informasjonsinnhald har liten entropi, medan data med høg entropi har lite informasjonsinnhald. Lågt informasjonsinnhald vil her bety at alle koeffisientane i ein node har stort sett dei same verdiane. [37] Basisen med den lågaste entropien, altså høgast informasjonsinnhald, vert definert som den beste basisen. I programmet er det lagt inn ei liste av wavelets, som kan tenkast å likna det kromatografiske signalet (sjå teori). Waveletane er gjevne i tabell 5.1 nedanfor.

*Tabell 5-1: Wavelets lagt inn i CODA_WPT, CODA_WPT2 og CODA_WPTlim*

| Familie | Nummer |
|---|---|
| Coiflets | 1, 2, 4, 5 |
| Daubechies | 4, 12 |
| Symlets | 4, 5, 6 |

*Tabelltekst til Tabell 5-1 : Tabellen visar 9 forskjellige wavelets. F.eks. Coiflet 1 og Symlet 5*

Den beste basisen og entropiverdien for denne vert rekna ut for kvar av waveletane, og den waveleten med minst entropiverdi vert vald. Node-nummera for den beste basisen vert lagra, og kvar av kolonnane (EIC) i M vert så transformert, dvs. koeffisientane vert funne, med hensyn på desse nodane (beste basis, som vart funnen av TIC). Kvar av nodene til kolonnane vert så terskla. Ein kan velja om tersklinga skal vera hard eller mjuk. Ved hard terskling vert koeffisientar med lågare absoluttverdi enn terskelgrensa nullsett, og for mjuk terskling vert dei større nærliggjande koeffisientane skalert ned i tillegg, for å få ein mjukare overgang. Terskelverdien vert utleia seinare. Etter terskling vert iWPT utført. mcq-grensa vert definert som mcq av node(1,2) (d1-leddet) til TIC, delt på antall masser. CODA vert så utført for det inverstransformerte (iWPT) datasettet.

Terskel-verdien vert definert av Formel 5.3-1:

$$Terskelgrense = amp \cdot th \qquad \text{Formel 5.3-1}$$

*th* er ei terskelgrense som vart introdusert av Donoho et al [37, 56] , og *amp* er ein skaleringsparameter (skalar) som vert vald av brukaren i tilfelle tersklinga ikkje er god nok.

*th* er ei grense basert på antall punkt og skalert etter medianen av punkta. I denne metoden vel ein å fastsetja *th* ut frå d1-leddet (dvs. node(1,2) ) til TIC. Det vert fyrst rekna ut ein skaleringsverdi *s*, av Formel 5.3-2:

$$s = median(|d1|/n) \qquad \text{Formel 5.3-2}$$

I formelen står *n* for antall massar. Det vert delt på n fordi terskelen vert nytta for EIC og ikkje summen av dei. Det vert så rekna ut ein verdi S av Formel 5.3-3:

$$S = m \cdot \log_2(m) \qquad \text{Formel 5.3-3}$$

I formelen er *m* antall retensjonstider. *th*-verdien vert rekna ut av Formel 5.3-4:

$$th = \sqrt{2 \cdot \log_2(S)} \cdot s \qquad \text{Formel 5.3-4}$$

## 5.4 CODA_ WPT2

Algoritmen fungerer på same måte som CODA_WPT, med unntak av fastsetjing av mcq-grensa. I denne versjonen vert støyen modellert som iWPT (inverstransformert) til verdiane som vert fjerna ved terskling av koeffisientane. Mcq-grensa vert så fastsett på same måte ut frå dette støy-estimatet, som det vart gjort i CODA_ndWT og CODA_CWT .

## 5.5 CODA_WPTlim

CODA_WPTlim er ein CODA-versjon der fastsetjing av mcq-grensa er avhengig av ein «grensemodus», som betyr at grensa kan verta fastsett som i anten CODA_WPT eller CODA_WPT2. Metoden nyttar berre WPT til fastsetjing av mcq-grense, dvs. det prosesserte

datasettet har berre mista støymassar i CODA. Ingen WT, glatting eller liknande er gjort med dei prosesserte massane.

## 5.6 CODA_SHD_ndWT

Programmet er basert på CODA_ndWT og den intervallbaserte CODA-versjonen [11] (frå no av referert til som SHD, som står for «slice halvdynamisk»). I metoden CODA_SHD_ndWT vert fyrst mcq-grensa fastsett som for ndWT-metoden i tillegg til at d1-ledda vert fjerna. a1-ledda vert så prosesserte som ved SHD-metoden med den fastsette mcq-grensa. Grunngjevinga for å kombinera CODA_SHD med CODA_ndWT kan vera at ein ynskjer ei meir skånsam og effektiv algoritme enn det CODA_ndWT er åleine, eller at ein vil fastsetja mcq-verdien automatisk for CODA_SHD og fjerna meir stokastisk støy.

# 6 Nye kvalitetsmål

I dette kapittelet vert det gjennomgått kva for nokre kvalitetsmål som vart utvikla i dette arbeidet

## 6.1 Singulærverdi-forhold

Forholdet mellom ”ekte ” (analytisk signal) og ”falske” (støy) singulærverdiar kan vera eit mål på kvaliteten på datasetta.

Når ein har funne singulærverdiane for eit datasett kan ein finna eit uttrykk for forholdet mellom verdiane som stammar frå dei kjemiske komponentane versus støyen. I dette arbeidet vart fylgjande formel komen fram til og nytta:

$$F = \frac{\sum_{i=1}^{k} s_i}{\sum_{i=k+1}^{m} s_i} \qquad \textit{Formel 6.1-1}$$

der F = singulærverdiforhold

s = singulærverdi-vektoren

m = lengda av singulærverdi-vektoren

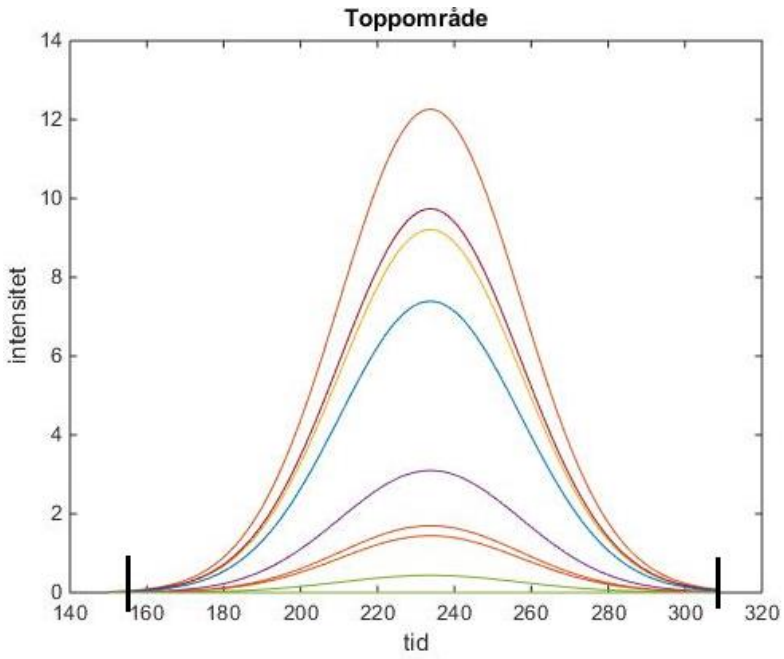
Singulærverdi 1 til k er store singulærverdiar og singulærverdi k+1 er den fyrste av dei mindre singulærverdiane. For uttrykket for F vert det nytta to kriterium for å avgjera kva verdi som er den siste verdien i teljaren, $s_k$ . Dersom eit av dei er oppfylt for i, så er $s_k = s_i$ . Det fyrste kriteriet er oppfylt dersom $s_i > 4 \cdot s_{i+1}$ . Det andre kriteriet er oppfylt for $s_1 > 10 \cdot s_{i+1}$.

Ved utrekning av singulærverdi-forhold (F) for støybehandla datasett og fjerna verdiar ved støybehandling, vert plasseringa av $s_k$ avgjort berre ut frå originalmatrisa, slik at F - verdiane for originalmatrisa og den prosesserte matrisa er samanliknbare. Algoritmen reknar også ut F for signal fjerna i prosesseringa basert på dei same vilkåra; noko som kan indikera om det er fjerna analytisk signal i prosesseringa.

## 6.2 Toppsamanlikning

Det vart utvikla to mål, basert på likskap til massar i same kromatografiske toppområde. Ein går ut frå a priori kunnskap om nokre faktiske forbindelsar i datasettet, der kvar forbindelse har minst eit fragment-ion i tillegg til moder-ionet. Programmet tek inn posisjon, dvs. retensjonsstart og retensjonsslutt for kvar utvald toppområde. Eit eksempel på eit toppområde er vist Figur 6.2-1 under. Eit toppområde er her definert som eit kromatografisk område med fragment som kjem frå same moderion. Toppområda vert fastsett til å vera gjeldane frå fyrste topp startar til siste topp sluttar, og kvar topp får då same lengd. Dei minst intense av toppane kan då få med områder som er under støynivået. Toppar i same toppområdet vert antatt å ha liknande form og same lengd dersom ein trekk vekk all støy, ettersom fragmentering fyrst skjer i massespektrometeret, dvs. at konsentrasjonsforholdet mellom fragmenta vil halda seg konstant under retensjonen.

Eit absolutt krav for samanlikningane er at dei valde massane ikkje vert fjerna frå settet, av CODA, då det ikkje vil vera mogleg å finna likskap, eller den må definerast som null. Det er også naudsynt at kvart toppområde må innehalda minst to massar, då ein elles ikkje vil kunna samanlikna. Nøyaktighet i val av startpunkt og sluttpunkt for elusjon av toppen er ikkje veldig viktig, ettersom ein ser på forskjell mellom det originale settet og det prosesserte settet, med det same området. Presisjon er derimot viktig, som vil sei at ein ser på det same toppområdet for begge setta, både for same støyfjerningsmetode og for forskjellige støyfjerningsmetodar.




*Figur 6.2-1: Toppområde med markert start og sluttpunkt.*

### 6.2.1 Korrelasjonsbasert kvalitetsmål

Det vart utvikla eit program som reknar ut ein kvalitetsindeks basert på korrelasjon mellom kromatografiske toppar. Ettersom topp-vektorane, frå fragmenta, har lik lengd, kan ein rekna ut korrelasjonen mellom dei. Korrelasjon vert funnen av Formel 6.2-1 (nedanfor) . Fordi korrelasjon er eit mål på likskap mellom to vektorar, finn ein korrelasjonen mellom kvar to-topps-kombinasjon av utvalde toppar, og så snittet av korrelasjonane. Ideen er at ein vil ha mindre korrelasjon mellom toppane når ein har mykje støy, dersom denne er stokastisk fordelt. Kvaliteten for kvart toppområde vert, rekna ut separat. Ved utrekning av korrelasjonsmålet for eit datasett tek ein snittet av kvaliteten til kvart område. Områda vert vekta likt uavhengig av kor mange utvalde fragment-/moder- ion kvart område har. Korrelasjonsmålet vil då ha ein verdi mellom 0 og 1, der 1 vert rekna som best kvalitet.

Formel for å rekna ut korrelasjon mellom to vektorar:

$$korr_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}$$ *Formel 6.2-1*

*, der X og Y er vektorar med lik lengd n*

### 6.2.2 Normbasert kvalitetsmål

Funksjonen har mange likskapar med den korrelasjonsbaserte metoden, men er ulik i måten likskapen mellom toppane i eit toppområde vert målt. I denne metoden vert toppane fyrst normalisert, med hensyn på euklidsk lengd (norm 2). Dette vil sei at kvar topp-vektor (masse) har lengd 1 (unit length). Likskapen vert så målt som normen til snittvektoren av massane. Dette vil gje kvalitetsmålet, lengda, ein verdi mellom 0 og 1, der 0 vil tilsvara minimal og 1 maksimal likskap mellom toppane. Likskap vert som tidlegare nemnd positivt korrelert med kvalitet. Nedanfor er formelen for utrekning av euklidsk lengd vist.

$$Euklidsk\ leng = norm2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$ *Formel 6.2-2*

*, der x er ein vektor med n element.*

# 7 Testing og diskusjon

## 7.1 Validering av CODA-versjonane og fastsetjing av ideelle parameterar

Denne seksjonen har som formål å finna parameterar der CODA-metodane fjernar støy og beheld signal, og til ein viss grad visa at metodane fjernar støy og beheld signal. Ein ser på kromatogram og spekter, før og etter behandling, og differansen (dvs. fjerna data).

Her vert setta sim3, HCO og TAG nytta mest, og ZM2 vert til dels nytta. sim3 er nyttig fordi ein har full kjennskap til støy og signal. HCO vert nytta fordi det er eit reelt sett som har nokre klart definerte forbindelsar (sjå Tabell 3-11) . Testinga av HCO har størst fokus på det kjente området i settet, frå no av kalla «kvalitetsområdet», som vil sei at ein ser mest på endringar i dette området, men prosesseringa vert alltid gjort på heile settet. TAG er i likskap med HCO eit reelt sett, med kjente forbindelsar (sjå Tabell 3-7). ZM2-settet har ein liten kunnskap om, men det kan vera interessant å sjå korleis metodane behandlar eit komplekst sett.

Ved plotting av kromatogram og massespektra for metodane vert det gjort eit namngjevingsskilje (også nemnd i notasjonskapittelet P.1): «Fjerna av WT og CODA» vil sei alt som vert forkasta, for ein metode som utførar både WT og CODA som ein del av prosesseringa. «Fjerna av CODA» referer berre til CODA-delen av algoritmen, men vil for ein metode som ikkje nyttar WT på dei prosesserte data, utgjera alt som vert forkasta.

Av wavelet-typane er det *Mexican hat* som liknar mest på ein kromatografisk profil, men ettersom denne berre var tilgjengeleg for kontinuerleg WT i MATLAB vert den berre nytta i CWT-metoden til å transformera til støyredusert signal. Støy-modelleringa i CWT vert gjort med *Biortogonal 2.8* waveleten. *Mexican hat* vert skalert med 12 og *Biortogonal 2.8* vert skalert med 1.

Det er eit mål å ha minst mogleg variasjon i input-parameterane. For grensekriteriet er derfor utgangspunktet vert å setja det lik medianen til støy-estimatet.

For ndWT metoden er *Symlet 5* waveleten vald, ut ifrå at formlikskapen til ein kromatografisk profil. Alle WPT-metodane finn beste wavelettre ut frå dekomponeringsnivåa 1 til 5. For alle setta som vert vel WPT-metodane ut *Coiflet 5* waveleten som den best eigna, av waveletane i Tabell 5-1 (kap. 5.3). Tersklinga vert vald til å vera hard, i alle tilfella, ettersom ein då ikkje

endrar på storleiken til dei behaldte verdiane. Skaleringsparameteren (amp frå Formel 5.3-1) som trengs for å få optimal terskling i WPT, dvs. glatte kurver, er sett ulikt for dei ulike setta; 20 for sim3, 10 for HCO, 10 for ZM2 og 400 for TAG1-4. Grense-modusen til WPTlim vert alltid vald til å vera lik WPT2, ettersom WPT-alternativet ikkje fungerer godt i praksis. For SHD_ndWT-metoden vert ulike intervallstorleikar nytta for dei ulike setta, pga. datasetta sin ulike storleik og natur. Intervallstorleikane vert presentert i Tabell 7-1.

*Tabell 7-1: Intervallstorleikar for SHD-metoden*

| | Lite intervall | Stort intervall |
|---|---|---|
| sim3 | 23 | 201 |
| HCO | 101 | 301 |
| ZM2 | 23 | 101 |

I Tabell 7-2 til Tabell 7-8 vert grensekriteriet, mcq-grensa, antall massar fjerna og antall massar behaldt gjeve for kvart av datasetta og for kvar metode. SHD_ndWT-metoden vert ikkje testa på TAG-setta ettersom desse allereie er små intervall. Grensekriteria har her to ulike verdiar for metodane. «Median» betyr at grensa vert sett til medianen av mcq-verdiane til støyestimatet, medan «min-median» er snittet av den minste og medianen av mcq-verdiane til støyestimatet. WPT-metoden har som nemnd berre eit grensekriterium. For SHD_ndWT-metoden vert ein masse rekna som behaldt dersom den vert behaldt på minst eit intervall.

*Tabell 7-2: Grensekriterium, mcq-grenser, antall massar behaldt og antall massar fjerna, for køyring av CODA-metodane for sim3 - datasettet.*

| sim3 | ndWT | CWT | WPT | WPT2 | WPTlim | SHD_ndWT |
|---|---|---|---|---|---|---|
| Grense-kriterium | median | median | - | median | median | Median |
| mcq-grense | 0.6209 | 0.5848 | 0.4178 | 0.4598 | 0.4598 | 0.6209 |
| Massar behaldt | 14 | 40 | 14 | 14 | 23 | 15 |
| Massar fjerna | 26 | 0 | 26 | 26 | 17 | 25 |

*Tabell 7-3: Grensekriterium, mcq-grenser, antall massar behaldt og antall massar fjerna, for køyring av CODA-metodane for HCO – datasettet.*

| HCO | ndWT | CWT | WPT | WPT2 | WPTlim | SHD_ndWT |
|---|---|---|---|---|---|---|
| Grense-kriterium | median | median | - | median | median | median |
| mcq-grense | 0.6320 | 0.5870 | 0.4675 | 0.5378 | 0.5378 | 0.6320 |
| Massar behaldt | 267 | 1004 | 950 | 950 | 438 | 779 |
| Massar fjerna | 737 | 0 | 54 | 54 | 566 | 225 |

*Tabell 7-4: Grensekriterium, mcq-grenser, antall massar behaldt og antall massar fjerna, for køyring av CODA-metodane for TAG1-datasettet.*

| TAG1 | ndWT | CWT | WPT | WPT2 | WPTlim |
|---|---|---|---|---|---|
| Grense-kriterium | median | median | - | median | median |
| mcq-grense | 0.5873 | 0.5559 | 0.2527 | 0.3376 | 0.3376 |
| Massar behaldt | 42 | 754 | 14 | 13 | 740 |
| Massar fjerna | 1459 | 747 | 1487 | 1488 | 761 |

*Tabell 7-5: Grensekriterium, mcq-grenser, antall massar behaldt og antall massar fjerna, for køyring av CODA-metodane for TAG2-datasettet.*

| TAG2 | ndWT | CWT | WPT | WPT2 | WPTlim |
|---|---|---|---|---|---|
| Grense-kriterium | min-median | median | - | median | median |
| mcq-grense | 0.2479 | 0.5485 | 0.2931 | 0.3258 | 0.3258 |
| Massar behaldt | 870 | 561 | 14 | 14 | 747 |
| Massar fjerna | 631 | 940 | 1487 | 1487 | 754 |

*Tabell 7-6: Grensekriterium, mcq-grenser, antall massar behaldt og antall massar fjerna, for køyring av CODA-metodane for TAG3-datasettet.*

| TAG3 | ndWT | CWT | WPT | WPT2 | WPTlim |
|---|---|---|---|---|---|
| Grense-kriterium | min-median | median | - | median | median |
| mcq-grense | 0.2859 | 0.5540 | 0.4769 | 0.3331 | 0.3331 |
| Massar behaldt | 785 | 752 | 6 | 9 | 748 |
| Massar fjerna | 716 | 749 | 1495 | 1492 | 753 |

*Tabell 7-7: Grensekriterium, mcq-grenser, antall massar behaldt og antall massar fjerna, for køyring av CODA-metodane for TAG4-datasettet.*

| TAG4 | ndWT | CWT | WPT | WPT2 | WPTlim |
|---|---|---|---|---|---|
| Grense-kriterium | min-median | median | - | median | median |
| mcq-grense | 0.2044 | 0.5640 | 0.4727 | 0.3385 | 0.3385 |
| Massar behaldt | 1072 | 669 | 8 | 9 | 745 |
| Massar fjerna | 429 | 832 | 1493 | 1492 | 756 |

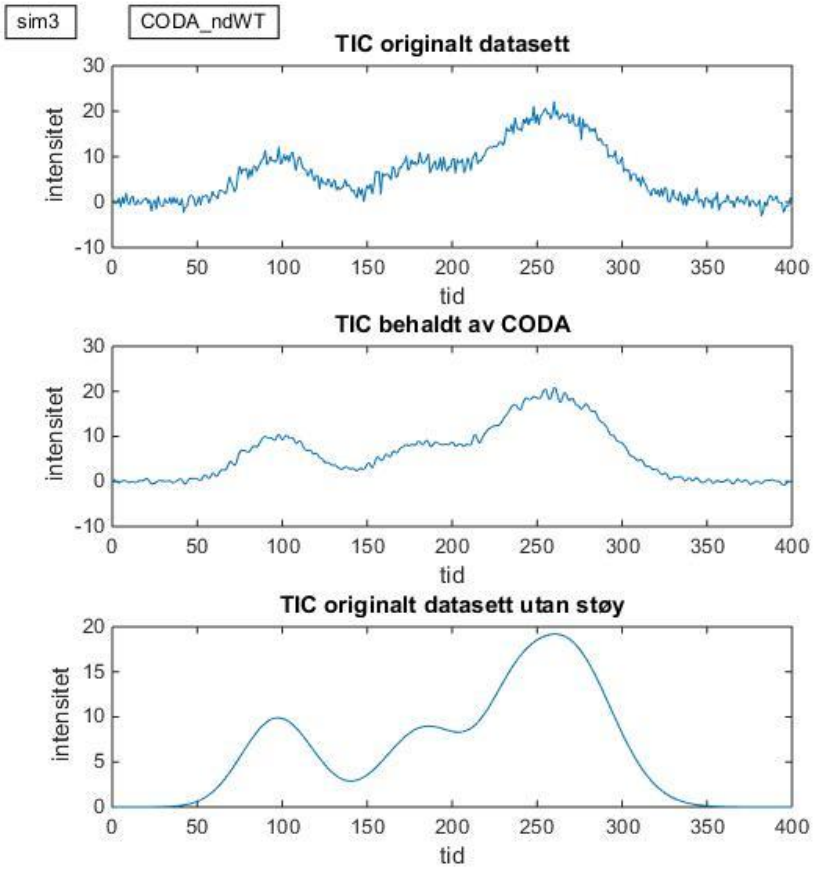
*Tabell 7-8: Grensekriterium, mcq-grenser, antall massar behaldt og antall massar fjerna, for køyring av CODA-metodane for ZM2-datasettet.*

| ZM2 | ndWT | CWT | WPT | WPT2 | WPTlim | SHD_ndWT |
|---|---|---|---|---|---|---|
| Grense-kriterium | median | median | - | median | median | min-median |
| mcq-grense | 0.6592 | 0.3192 | 0.4555 | 0.4102 | 0.4102 | 0.3540 |
| Massar behaldt | 152 | 701 | 311 | 347 | 347 | 532 |
| Massar fjerna | 549 | 0 | 390 | 354 | 354 | 169 |

### 7.1.1 CODA_ndWT

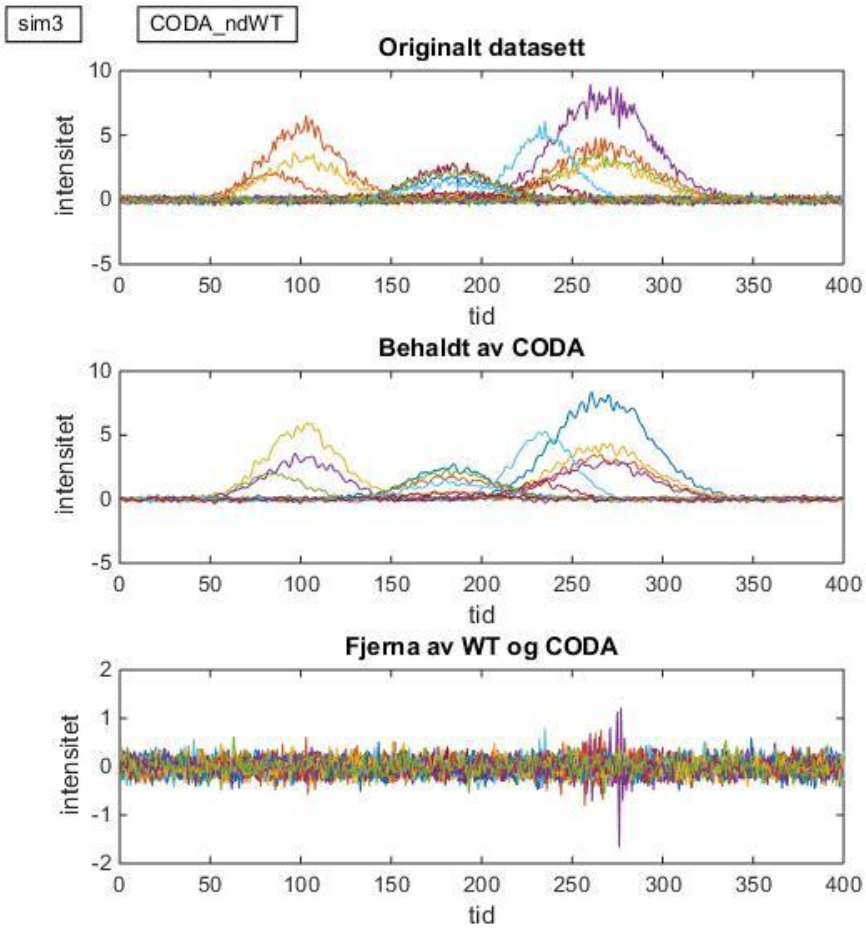
#### *7.1.1.1 sim3-settet*

For settet sim3 fungerer CODA_ndWT bra til å fjerna stokastisk støy. Av Figur 7.1-1 er det tydeleg at det prosesserte settet har behaldt forma som det originalt hadde, at mykje stokastisk («tilfeldig støy») er fjerna, men også at det framleis er litt støy igjen.




*Figur 7.1-1: TIC for (øvst) originalt datasett, (midtarst) CODA_ndWT-prosessert datasettet og (nedst) det originale datasettet utan pålagt støy, av settet sim3*

I Figur 7.1-2, der massekromatogramma for originalt sett, behaldte og forkasta data er plotta, er det tydeleg at det analytiske signalet i massekromatogramma er bevart. Det forkasta signalet inneheld ingen likskap med toppane, og alt ser ut til å vera likt fordelt rundt tids-aksen, som er karakteristisk for stokastisk støy. Mellom 250 og 300 på tidsaksen har det vorte fjerna støy med høgare intensitet enn elles. Dette er pga. at noko av støyen er heteroskedastisk, altså avhengig av signalstyrken til det analytiske signalet.

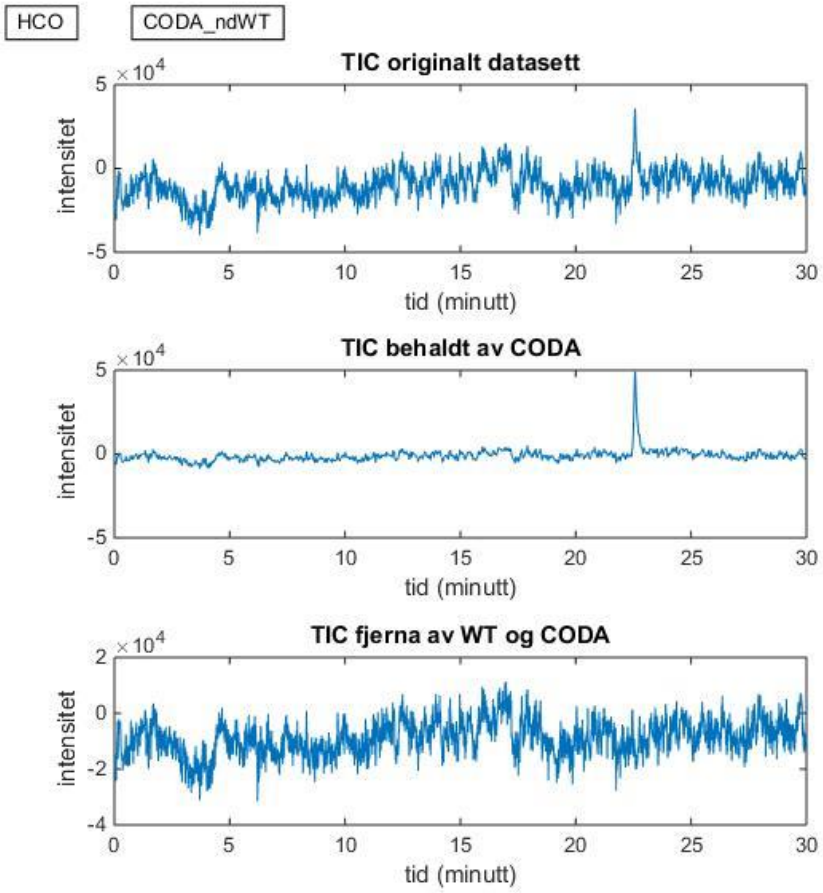



*Figur 7.1-2: Alle massekromatogram for (øvst) originalt datasett, (midtarst) CODA_ndWT-prosessert datasettet og (nedst) det som vert fjerna av metoden, av settet sim3*

Vedlegg 5 visar massekromatogramma (EIC) som vert fjerna av CODA-delen av algoritmen, etter WT er utført. 14 av dei totalt 40 EIC vert behaldt, og dei 26 fjerna EIC inneheld tydeleg støy. At WT-delen av algoritmen også fjernar støy frå EIC som inneheld analytisk signal synleggjer at både WT og CODA er nyttige. I Vedlegg 6 er det lagt ved tre TMS (totalt massespekter). Dei tre TMS er ikkje signifikant forskjellige (ut frå utsjånaden), og ein kan konkludera med at massespektera i liten grad er påverka av støyen, ettersom den er liten for kvar enkeltmasse. TMS-observasjonane sikrar også at ingen signifikant massar er fjerna.

#### *7.1.1.2 HCO -settet*

For HCO-settet kan ein sjå av TIC i Figur 7.1-3 at CODA_ndWT-algoritmen fjernar mykje støy frå TIC, som gjer at analytiske signal kjem meir til syne. Det er likevel berre den høgste signaltoppen som visar godt igjen i TIC, og også ved inn-zooming på det definerte kvalitetsområdet (sjå Vedlegg 7) er det vanskeleg å sjå småtoppane (kvalitetstoppane **2**, **1** og **3**).

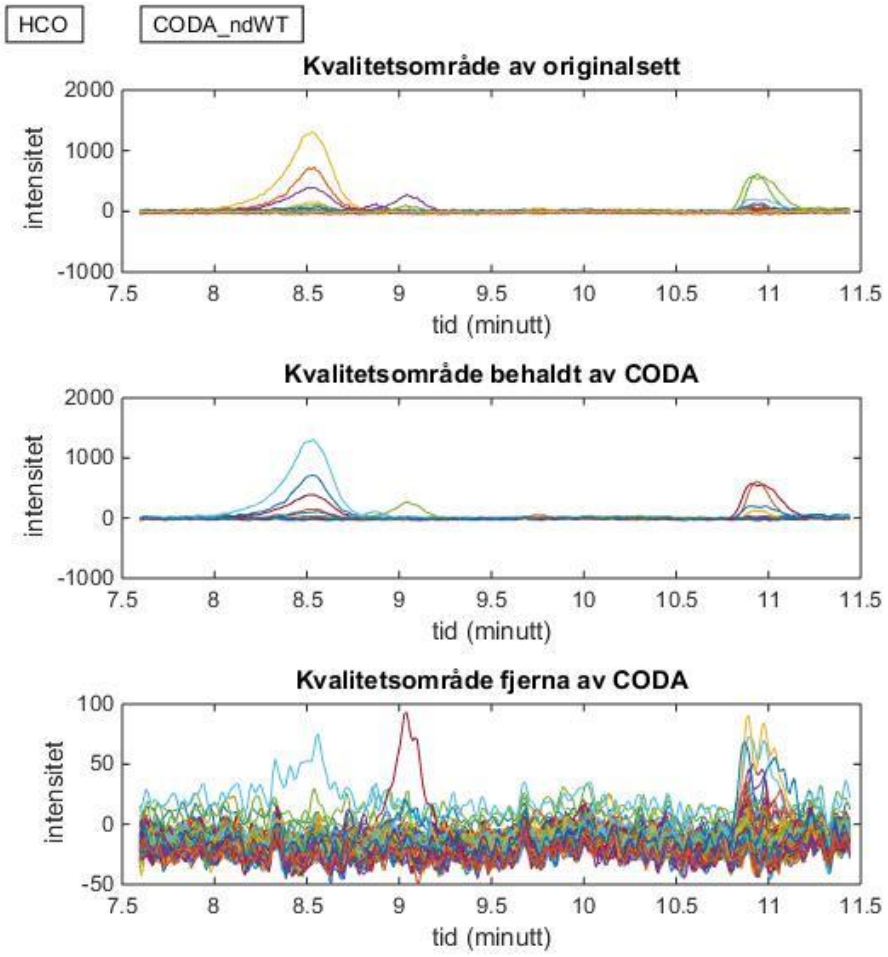



*Figur 7.1-3: TIC (Total Ion Current/Chromatogram) for HCO-settet; (øvst) originalt datasett, (midtarst) CODA_ndWT-prosessert datasettet og (nedst) det som vert fjerna av metoden*

Av

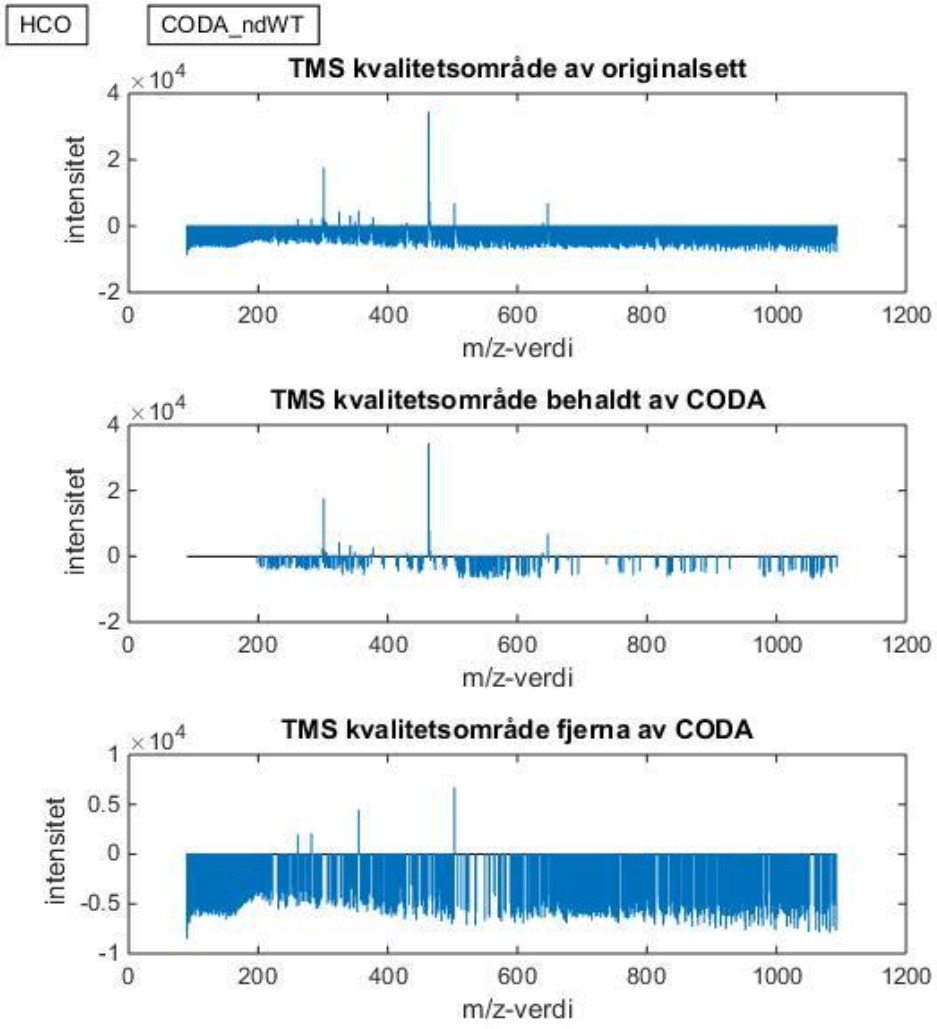
Figur 7.1-4 er det tydeleg at den eine massen tilhøyrande forbindelse **1** ( ved ca. 9 minutt) har vorte fjerna. Her er mcq-grensa satt til å vera medianen til mcq av d1 til EIC, men den gjevne massen vert også fjerna dersom ein nyttar det lågaste grensekriteriet (snittet av den minste mcq og medianen til mcq av d1). At massen vert fjerna av CODA kan ha noko med storleiken på settet å gjera, ettersom massen kan vera opphav til støy i andre delar av settet. Det er også

mogleg at det vart fjerna analytt-signal ved forbindelse **2** (ved 8.5 min.) og forbindelse **3** (ved 11 min.) .




Figur 7.1-4: EIC av kvalitetsområdet for HCO-settet; (øvst) originalt datasett, (midtarst) CODA_ndWT-prosessert datasettet og (nedst) det som vert fjerna av CODA-delen i metoden

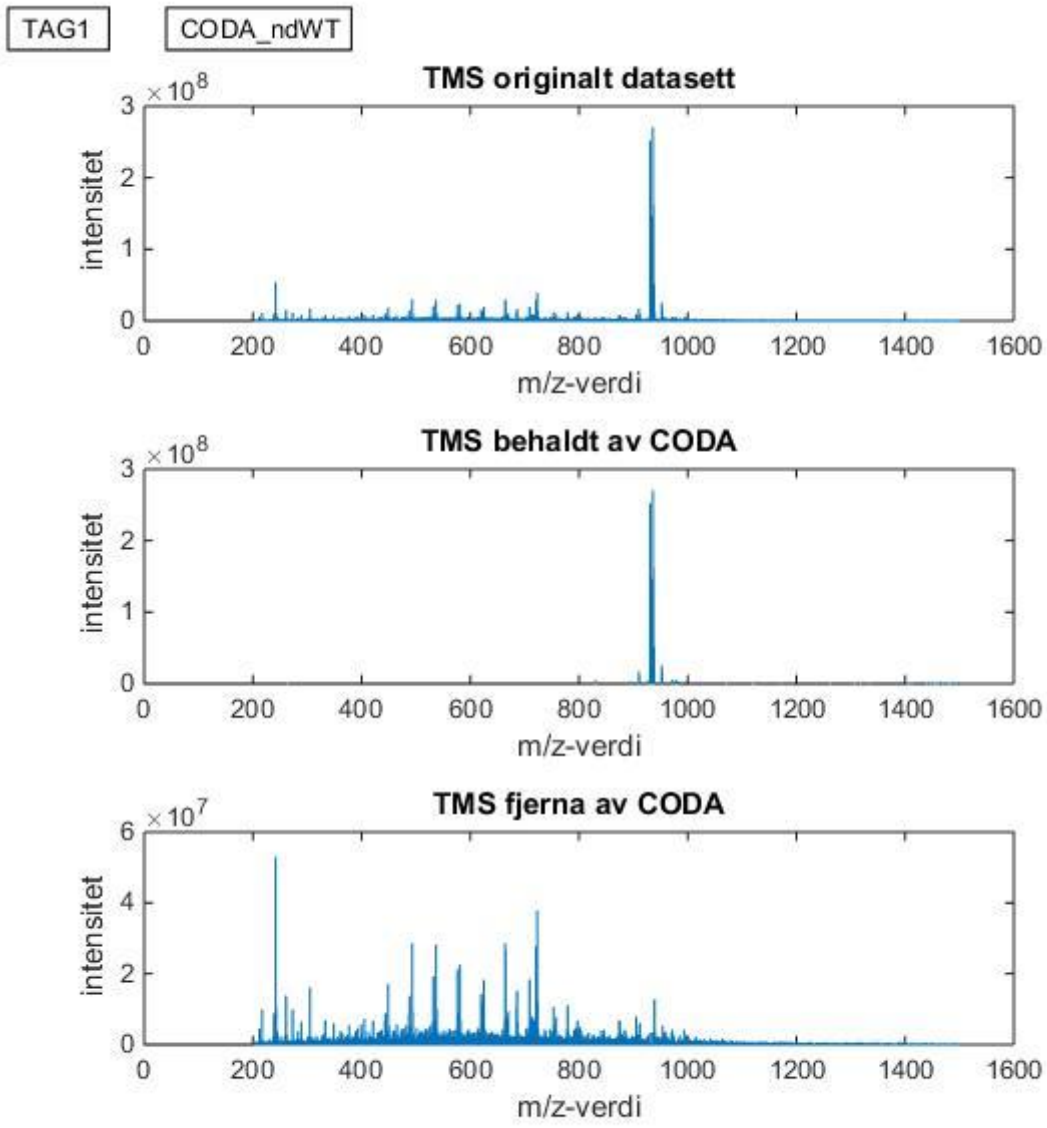
Spektra i Figur 7.1-5 visar at dei mest instense signala i kvalitetsområdet har vorte behaldt, men at nokre signal har gått tapt. Av dei sistnemnde kan det også vera at mange av dei er heteroskedastisk støy.




*Figur 7.1-5: TMS av kvalitetsområdet for HCO-settet; (øvst) originalt datasett, (midtarst) CODA_ndWT-prosessert datasettet og (nedst) det som vert fjerna av metoden.*

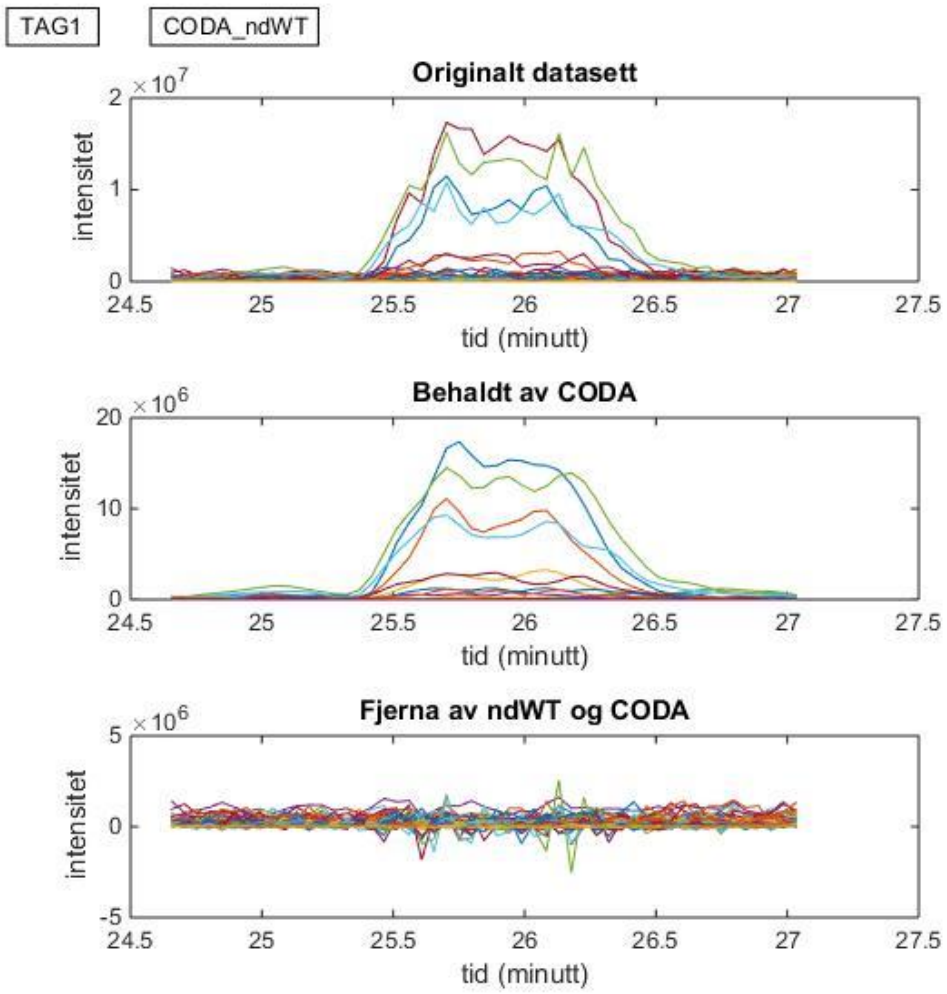
### *7.1.1.3 TAG-setta*

Ved prosessering av TAG1-settet vert 42 av dei totalt 1501 massane behaldt. TMS for settet er vist i Figur 7.1-6 viser at det er dei store massane som vert behaldt. Dei mest intense massane som vert behaldt er i hovudsak Na- og $NH_4$- forløperiona ( frå Tabell 3-7) , og ulike fragment/isotopar av desse. At desse er tilstades peikar mot at metoden er skånsam nok.




*Figur 7.1-6: TMS av TAG1-settet; (øvst) originalt datasett, (midtarst) CODA_ndWT-prosessert datasettet og (nedst) det som vert fjerna av CODA-delen i metoden.*

I Figur 7.1-7 er alle massekromatogramma plotta for det originale settet, det behandla settet og forkasta data. Det kjem ikkje fram nokon markant toppstruktur i dei fjerna EIC, samtidig som at toppane det behandla datasettet er finare enn i originalen. Dette forsterkar inntrykket om at metoden er tilstrekkeleg skånsam.



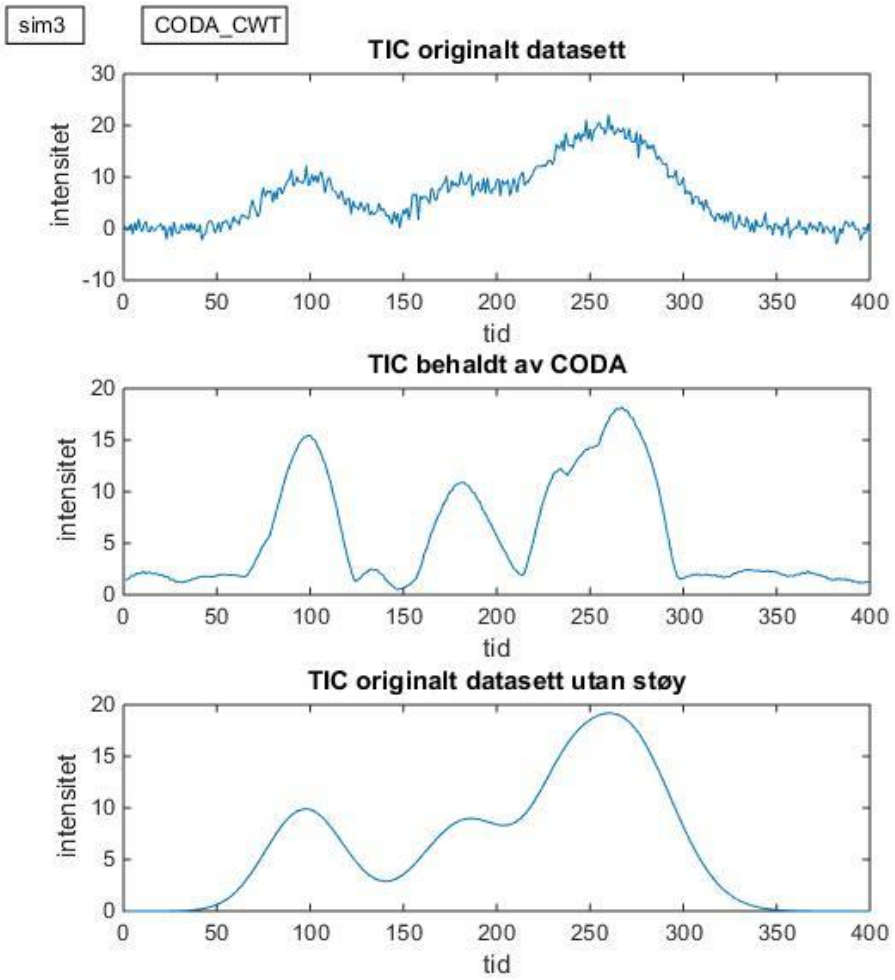

*Figur 7.1-7: EIC av TAG1-settet; (øvst) originalt datasett, (midtarst) CODA_ndWT-prosessert datasettet og (nedst) det som vert fjerna av metoden.*

EIC for TAG2-TAG4 - setta er vist i Vedlegg 8 - Vedlegg 10. For dei tre setta vert det fjerna signalmassar når grensekriteriet er medianen av støyestimatet, grunna at profilane er lange og flate. Grensekriteriet vert følgjeleg nedjustert til det lågaste alternativet, ettersom ein vil behalda alt analytisk signal.

### 7.1.2 CODA_CWT

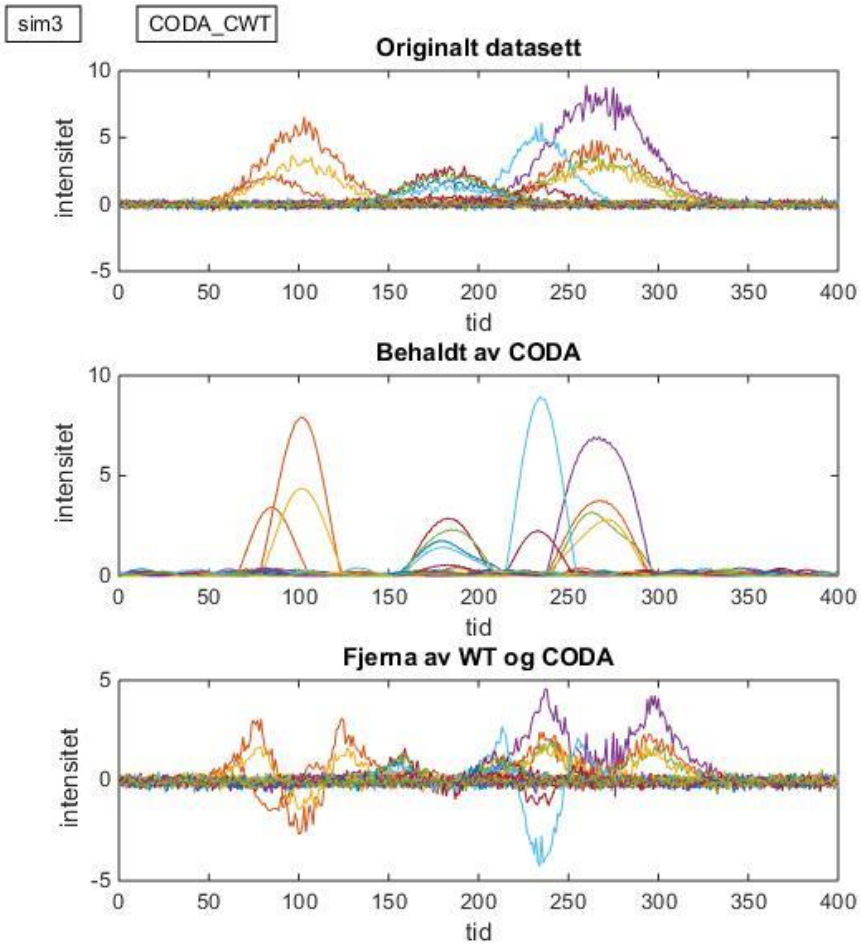
#### *7.1.2.1 sim3-settet*

Den tredje toppen (ved ca. 250) i TIC av det CODA_CWT-prosesserte sim3-settet (i Figur 7.1-8) framstår som mindre reell enn TIC av det originale sim3 (og sim3 utan støy). Avstanden mellom toppane er også større enn den er i originalen, ettersom toppbreidda har minka. Toppane har i tillegg vorte glattare, som vil sei at den stokastiske støyen har vorte minska.




*Figur 7.1-8: TIC av sim3-settet; (øvst) originalt datasett, (midtarst) CODA_CWT-prosessert datasett og (nedst) det som vert fjerna av metoden.*

I Figur 7.1-9 kjem det fram kvifor TIC av den tredje toppen er mindre gaussisk. Den tredje toppen inneheld signal frå to ulike analyttar, og når toppbreidda og intensiteten for desse vert endra kjem forskjellane meir fram.



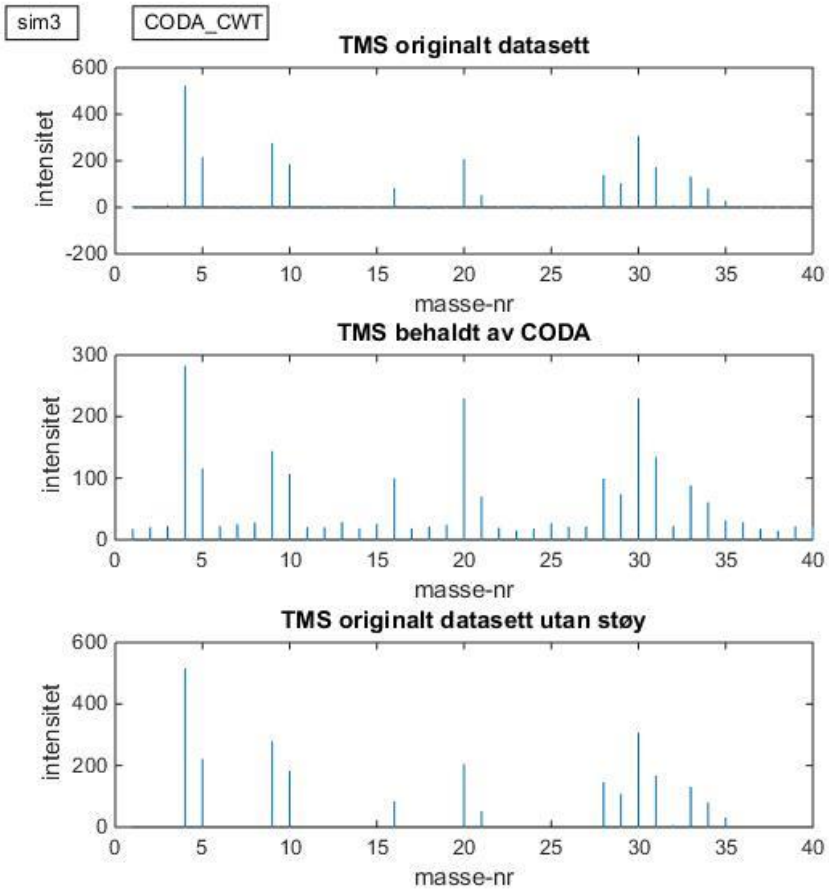

*Figur 7.1-9: Alle EIC av sim3-settet; (øvst) originalt datasett, (midtarst) CODA_CWT-prosessert datasettet og (nedst) det som vert fjerna av metoden.*

Det klart at CWT-algoritmen er røff med dei kromatografiske toppane. Forutan reduksjon av toppbreidda, kan ein sjå på i det midtre plottet i Figur 7.1-9, over, at det siste toppområdet har fått litt forvriding i to av massane. Dette er også tilfellet for toppområdet mellom 150 og 200 på tidsaksen. Det har også skjedd endringar i intensiteten til toppane og forholdet mellom intensitetatane. Metoden skalerer settet, som nemnt i kap. 5.2, slik at høgste topp i har lik intensitet som høgste topp i det originale settet, men det endra forholdet mellom toppane

kjem av WT. Dei endra intensitetane og det endra intensitetsforholdet gjer metoden særs lite eigna, dersom ein vil finna konsentrasjonar av analyttar.

Alle endringane i datasettet kjem av transformasjonsstega før CODA, ettersom ingen av dei 40 massane vert fjerna av CODA. Dersom ein hevar mcq-grensekriteriet til det høgste kriteriet (frå nr 1 til 3 på ein skala frå 0 til 3) vert heller ingen massar fjerna. CODA_CWT har vanskeleg for å forkasta massar, ettersom bruk av *Mexican hat* - waveletar gjer alt signal om til tilnærma gaussisk form. Waveletane er i tillegg skalert for å framheva signal, slik at dei gaussiske toppane får stor breidd. Innverknaden på støy kan ein sjå ved å samanlikna kurvestrukturen i det behandla datasettet med nullkomponentområdene det originale datasettet (sjå vedlegg 8-11)

TMS (Figur 7.1-10) for det behandla settet er lågare enn for det originale, noko som er venta, ettersom mange av massane får mindre breidd. I tillegg er det tydeleg at dei transformerte støymassane vert meir synlege, noko som også skjer pga. CODA_CWT nullsett verdiar under null, pga. naturen til «Mexican hat»-waveleten.

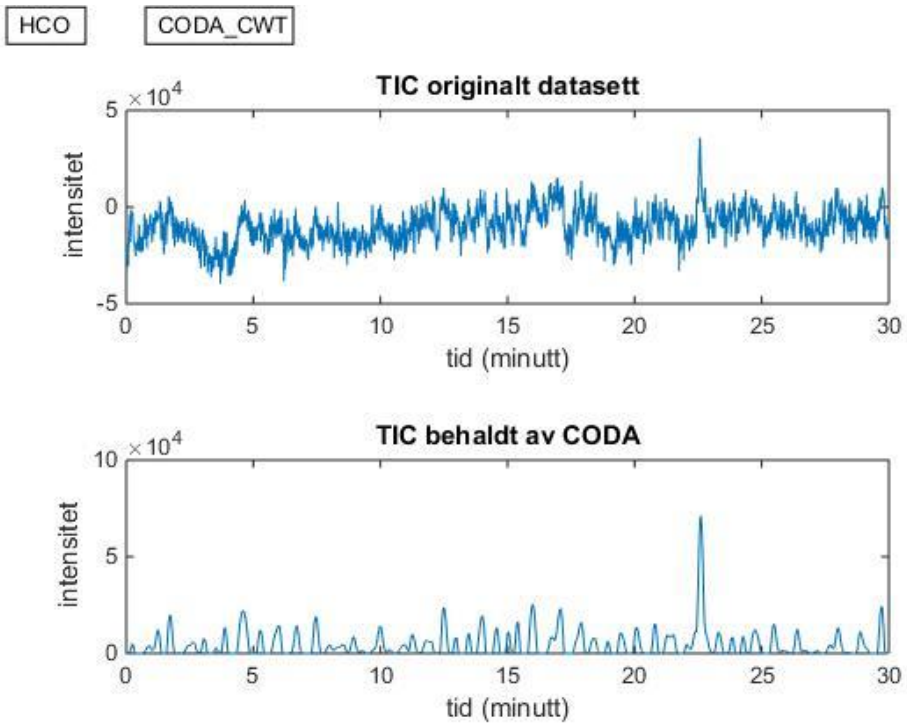



*Figur 7.1-10: TMS av sim3-settet; (øvst) originalt datasett, (midtarst) CODA_CWT-prosessert datasettet og (nedst) originalt datasett utan støy.*

### *7.1.2.2 HCO-settet*

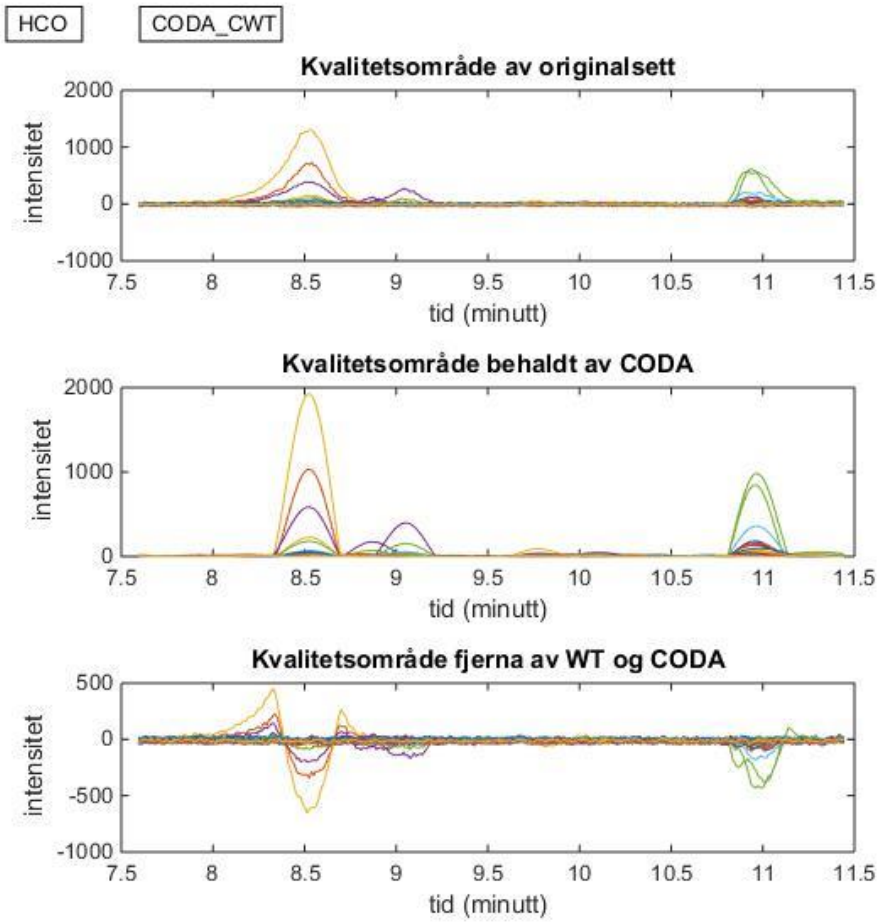
Ved CODA_CWT-prosessering av HCO-settet vert ein del av støyen i TIC fjerna, som gjer at toppen ved ca. 23 minutt kjem betre til syne. Det er derimot umogleg å skilja toppane i kvalitetsområdet frå toppar som vert addert opp av transformerte støymassar, bla. fordi CODA

ikkje fjernar nokon av dei 1004 massane (,noko som heller ikkje skjer viss ein aukar mcq-kriteriet).




*Figur 7.1-11: TIC av HCO-settet; (øvst) originalt datasett, (midtarst) CODA_CWT-prosessert datasettet og (nedst) originalt datasett utan støy.*

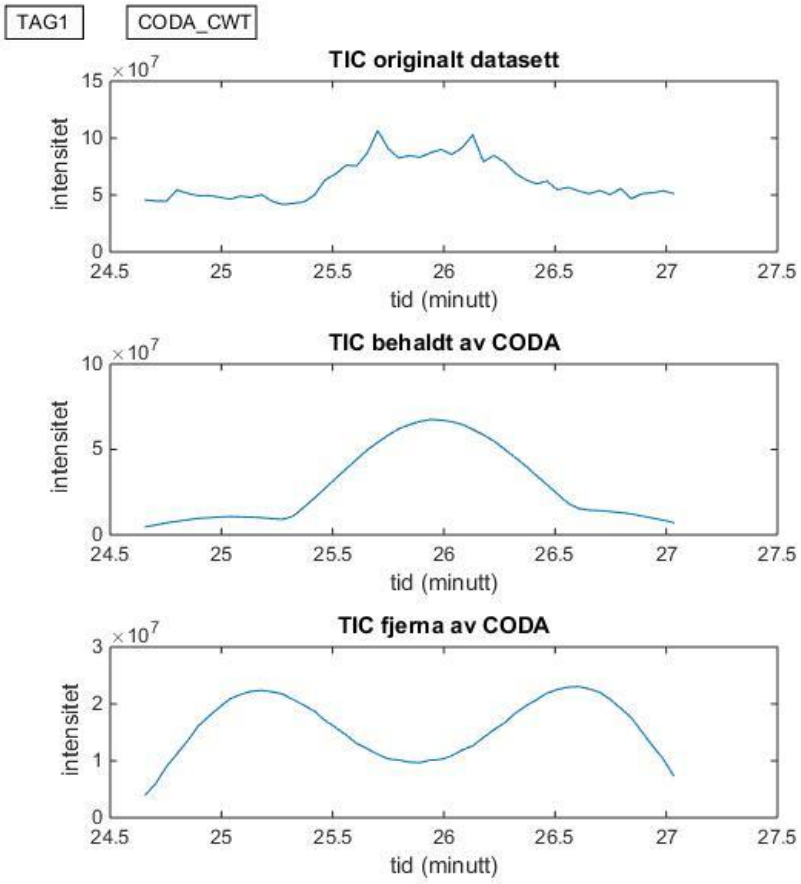
Det fyrste ein ser ved inspeksjon av EIC av kvalitetsområdet i Figur 7.1-12, er at CWT-algoritmen får glatta ut mykje støy og gjort toppane meir symmetriske. Ein annan ting er at intensitetane vert endra, og at det kjem fleire små toppar/kurver til syne. Av TIC-representasjonen (i Vedlegg 12) av det behandla settet kan ein sjå kvalitetstoppane, i motsetnad til i TIC av originalsettet. Problemet i TIC av det behandla settet er dei store ekstratoppane, som vert addert opp av mange små toppar som ein kan sjå i Figur 7.1-11. Ein av desse toppane er ein intens topp rundt 10 minutt, som har høgare intensitet enn kvalitetstoppane.




*Figur 7.1-12: EIC av kvalitetsområdet til HCO-settet; (øvst) originalt datasett og (nedst) CODA_CWT-prosessert datasettet.*

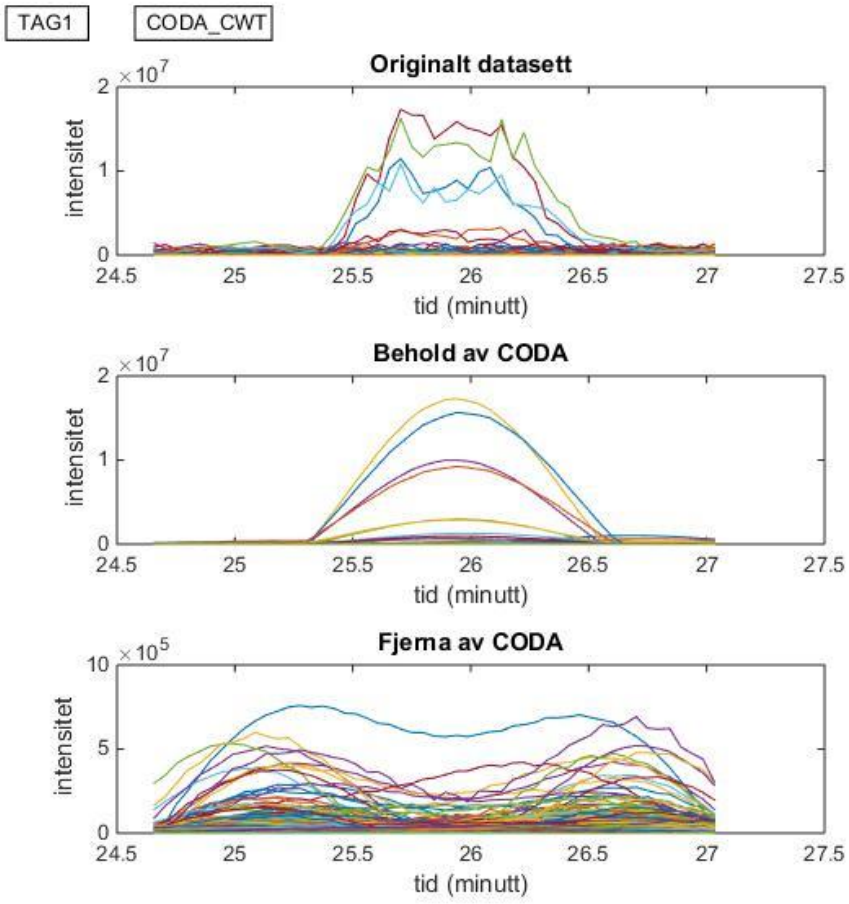
### *7.1.2.3 TAG-setta*

For både TAG1, TAG2, TAG3 og TAG4 fungerer CODA_CWT bra, som vil sei at CODA-delen av algoritmen fjernar støy i tilstrekkeleg grad, slik at TIC liknar på dei analytiske profilane i toppområda, og ikkje støyen. TIC for TAG1 er vist i Figur 7.1-13. (TIC og TMS for TAG2-TAG4 ligg i Vedlegg 13 til Vedlegg 18 )

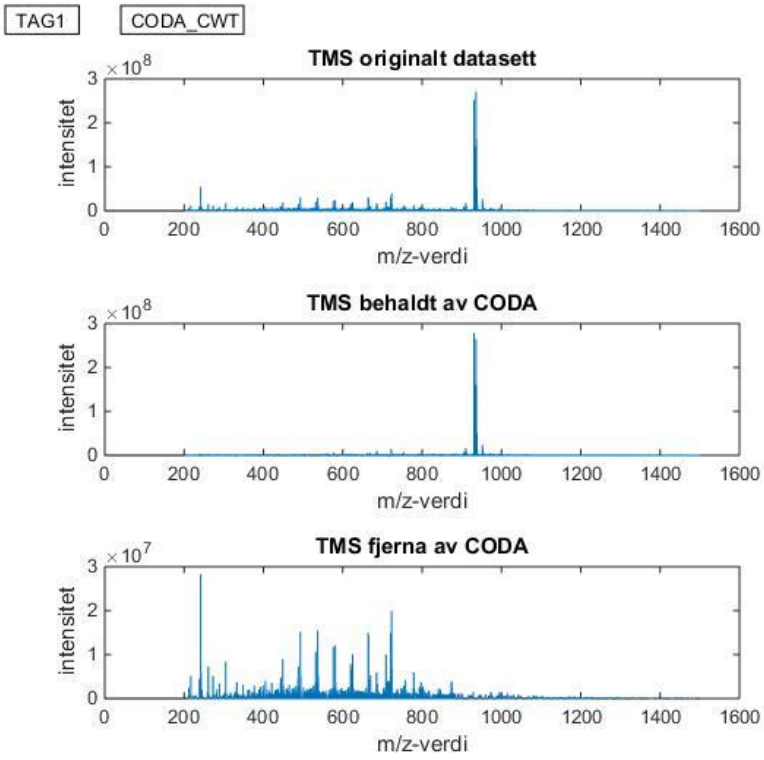



*Figur 7.1-13: TIC av TAG1-settet; (øvst) originalt datasett, (midtarst) CODA_CWT-prosessert datasettet og (nedst) signal fjerna av CODA-delen i metoden.*

Figur 7.1-14 viser settet som EIC, og viser forma til behalde og fjerna data. Ut frå plotta ser det ut som at det analytiske signalet er behaldt, og det vert bekrefta av TMS i Figur 7.1-15.




*Figur 7.1-14: EIC av TAG1-settet; (øvst) originalt datasett, (midtarst) CODA_CWT-prosessert datasettet og (nedst) signal fjerna av CODA-delen i metoden.*

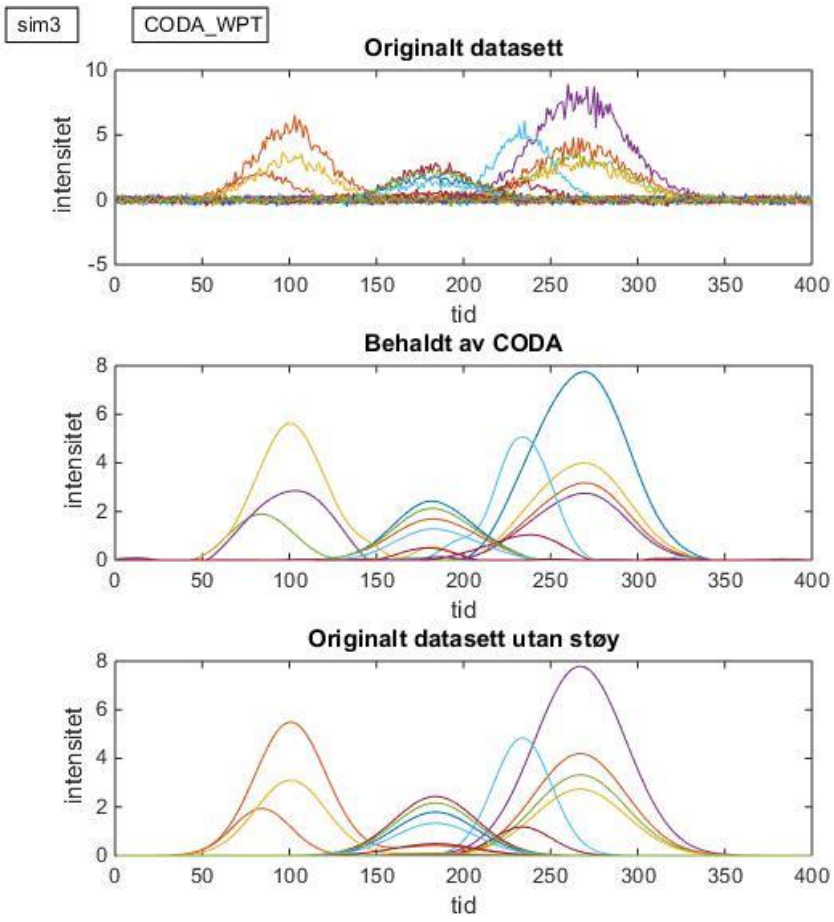



*Figur 7.1-15:* TMS av TAG1-settet; (øvst) originalt datasett, (midtarst) CODA_CWT-prosessert datasettet og (nedst) signal fjerna av CODA-delen i metoden.

### 7.1.3 CODA_WPT

#### *7.1.3.1 sim3-settet*

Ved bruk av WPT vert det prosesserte settet nesten identisk med det originale settet utan pålagt støy, som vert vist i Figur 7.1-16.




*Figur 7.1-16:* EIC av sim3-settet; (øvst) originalt datasett, (midtarst) CODA_WPT-prosessert datasettet og (nedst) originalt datasett utan støy.

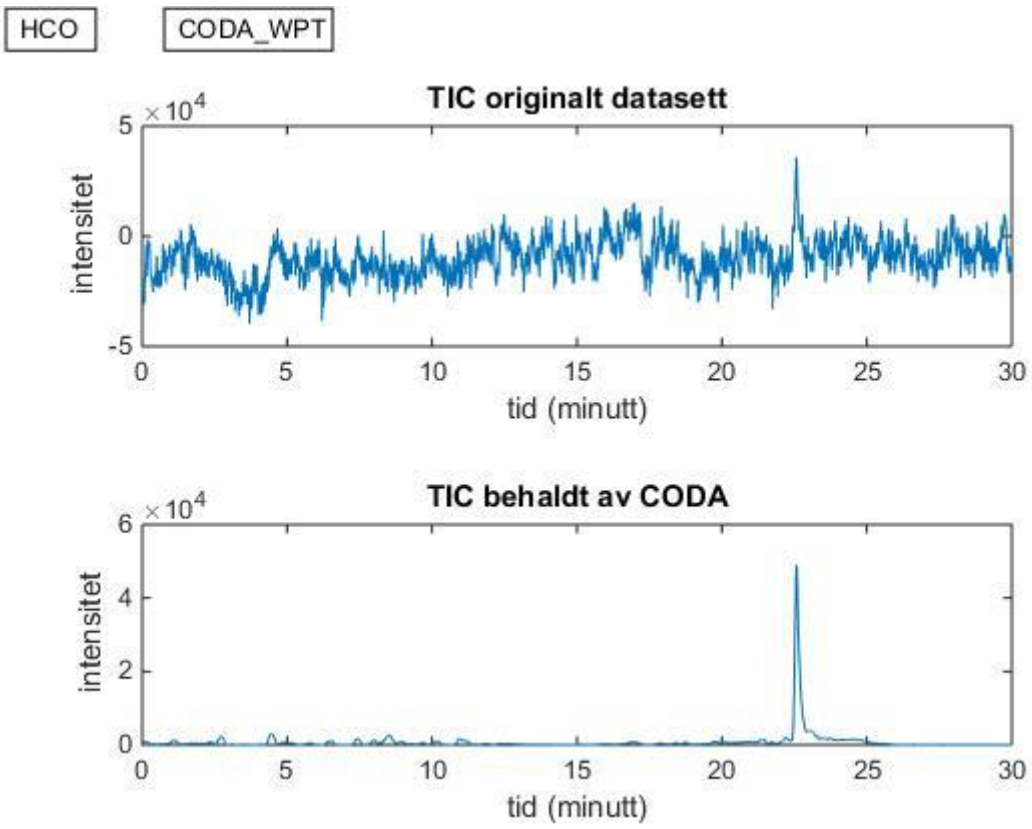
Nokre av dei kromatografiske profilane har vorte litt forvrengte i forhold til forma dei har i originalen utan støy. Eksempel på dette er at den nest høgste profilen ved tida 100 har vorte rundare i toppen, og at den høgste profilen ved tida 240 har hale mot venstre. Desse små forskyvingane har truleg med at støyen har endra på forma til kurvene. Det er også litt støy gjenverande heilt i starten av kromatogrammet.

CODA-delen av CODA_WPT fjernar berre nullsette massar for sim3-settet. Ettersom nesten all den pålagte støyen er fjerna, fungerer CODA som ynskjeleg.

TIC er vist i Vedlegg 19, og bekreftar at det prosesserte settet i stor grad er likt originalen utan støy.

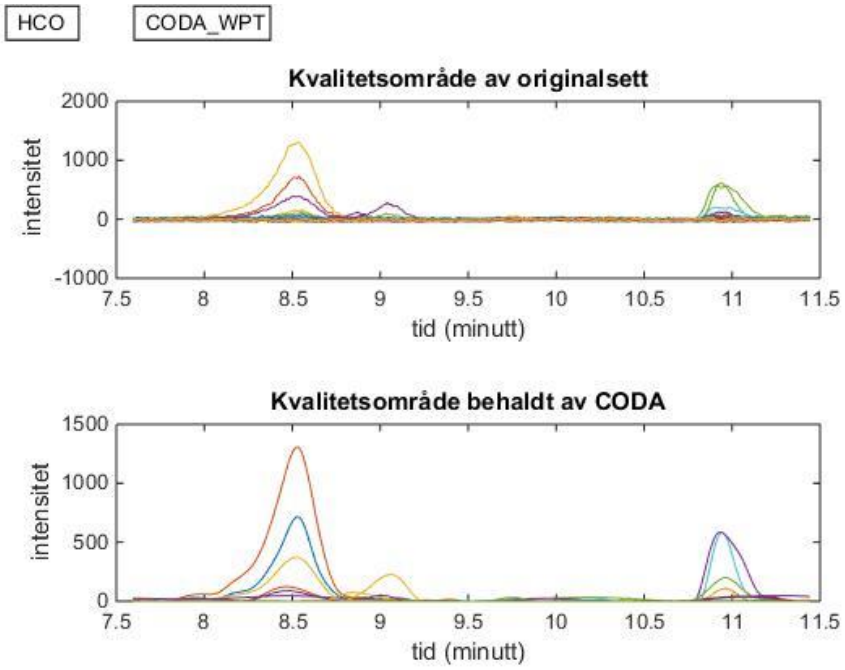
#### *7.1.3.2 HCO-settet*

CODA-delen av WPT-versjonen fjernar, som for sim3, berre nullsette massar for HCO-settet. Dette vil sei at alt som er att ved rekonstruksjon av den beste basisen, har høgare mcq-verdi enn grensa og vert derfor ikkje fjerna. TIC av originalsettet og det prosesserte settet vert vist i Figur 7.1-17




Figur 7.1-17: *TIC av HCO-settet; (øvst) originalt datasett og (nedst) CODA_WPT-prosessert datasettet.*

Av TIC ser støyfjerninga ved WPT-terskling ut til å vera veldig bra, ettersom nesten all bakgrunnstøyen er vekke. Om ein ser på kvalitetsområdet i Figur 7.1-18 er tersklinga skånsam med toppane, i tillegg til at støyen vert fjerna. Som for sim3 vert det også for WPT av HCO litt forskyvingar på forma til nokre av massane, viss ein tek utgangspunkt i gaussisk form - der kvar forbindelse har forskjellige fragment/massar med toppunkt direkte under kvarandre. Forskyvingar kan ein sjå for topp **2** (ved 8.5 min.) frå den fjerde massen og nedover, for den andre massen til topp **1** ( ved 9.1 min.). Det er også skeivhet i massane til topp **3** (ved 11.0 min.), der den tredje og fjerde massen har toppunkt meir til høgre enn den første og andre.




Figur 7.1-18: *EIC av HCO-settet; (øvst) originalt datasett og (nedst) CODA_WPT-prosessert datasettet.*

Av Vedlegg 20 kan ein sjå at TIC av kvalitetsområdet er reinare enn for CWT-metoden, men det er også for WPT-metoden eit problem med småtoppar som vert addert opp til store toppar.

#### *7.1.3.3 TAG-setta*

For TAG-setta fjernar CODA-delen av WPT-metoden meir enn nullsette kromatogram. I Vedlegg 21 til Vedlegg 24 er TIC for TAG1 og TAG2, og EIC for TAG3 og TAG4. Det er tydeleg av plotta at CODA-fjernar støy-massar for TAG1 og TAG2, medan signal-massar vert fjerna i TAG3 og TAG4. Signalmassane vert fjerna fordi profilane vert for flate i toppen, og dermed får liten mcq-verdi når gjennomsnittet vert trekt ifrå i utrekninga. WPT-metoden har ingen input-parameter for grensekriterium, ettersom mcq-grensa vert fastsett som mcq til TIC av d1. Det vert derfor ikkje prøvd med andre mcq-grenser.

### 7.1.4 CODA_WPT2

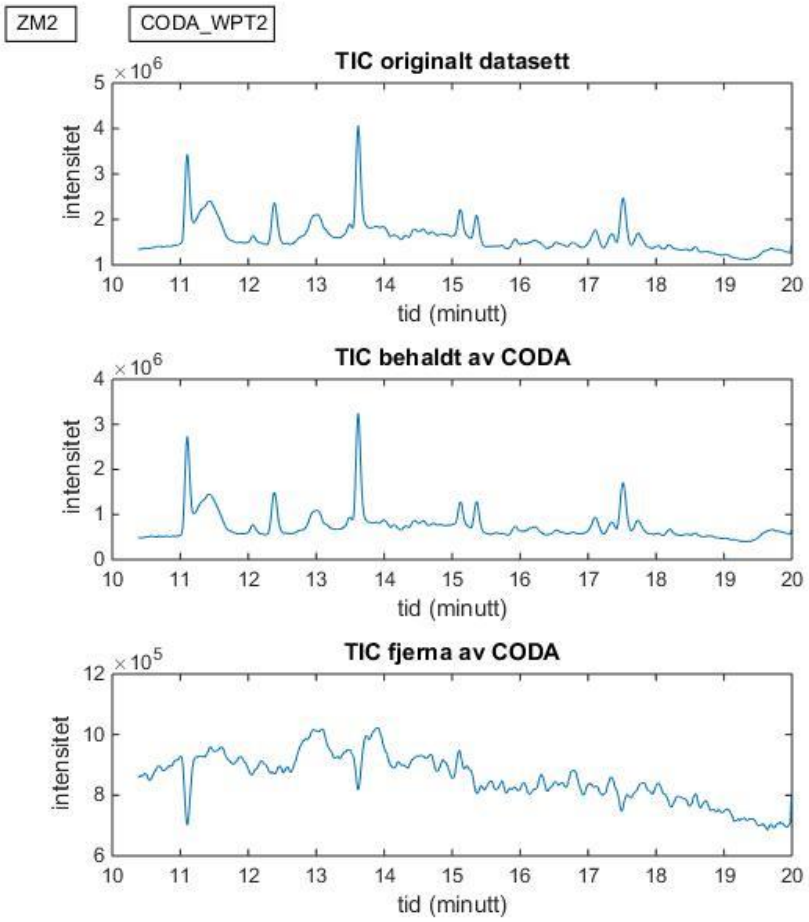
CODA_WPT2 fjernar, som CODA_WPT, ikkje anna enn nullsette kromatogram i CODA-delen. Output-matrisene for CODA_WPT2 vert derfor lik som for CODA_WPT for både sim3-datasettet og HCO-datasettet, ettersom berre mcq-grense-utrekninga er annleis.

#### *7.1.4.1 TAG-setta*

For TAG3 og TAG4 vert analytt-massane som vert fjerna for WPT-metoden, behaldt i WPT2-metoden. Dette kjem av at mcq-grensa er lågare i WPT2-metoden, for dei gjevne setta. EIC av setta er lagt ved i Vedlegg 25 og Vedlegg 26. Det kan diskuterast om det er positivt at massane vert inkludert eller ikkje. Med tanke på utsjånaden til toppane burde dei kanskje ha vorte forkasta, dersom ein var ute etter høgkvalitetsprofilar. Men ettersom målet i dette prosjektet er å inkludera alt mogleg analytisk signal, er det best å behalda toppane.

#### *7.1.4.2 ZM2-settet*

Ved bruk av CODA_WPT og CODA_WPT2 for datasettet ZM2 vart CODA-delen også nyttig som vist i Figur 7.1-19 under.




*Figur 7.1-19:* TIC av ZM2-settet; (øvst) originalt datasett, (midtarst) CODA_WPT2-prosessert datasettet og (nedst) signal fjerna av CODA-delen i metoden.

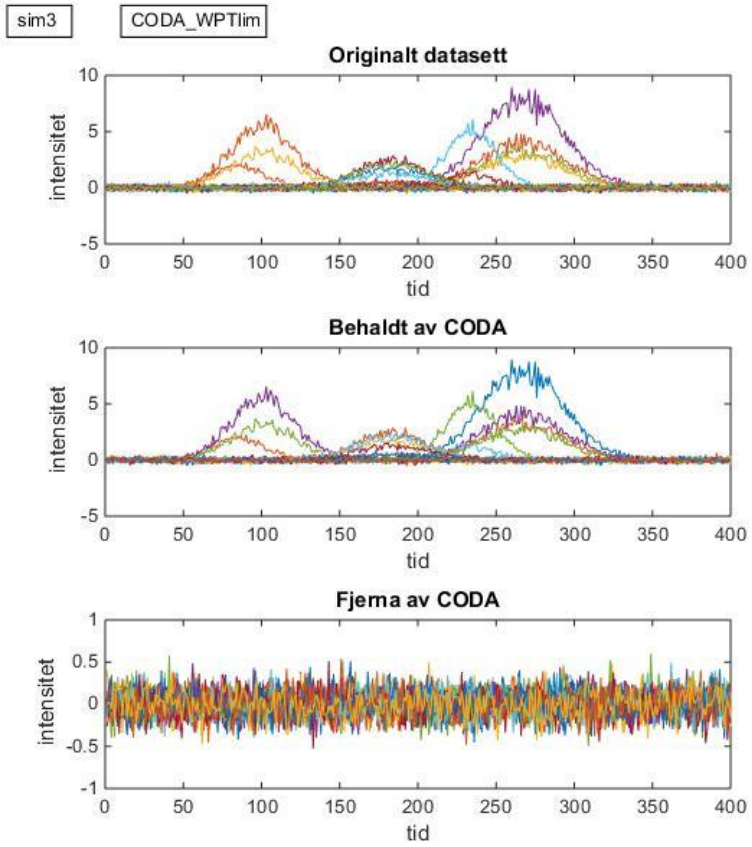
For TIC av ZM2 vert signala bakgrunnen/baselinja signifikant redusert av CODA. I Vedlegg 27 er EIC plotta, og av det fjerna signalet er det tydeleg at mange av massane har eit lågt signalet har eit tilnærma stasjonært signal over heile retensjonsområdet. Dei nemnde massane vert ikkje fjerna ved terskling, men dei får ein låg mcq-verdi i CODA-delen og vert derfor fjerna. Det er også mange minimale toppar blant dei fjerna massane, men fordi desse er låge i forhold til baselinja vert dei fjerna - ettersom dei får lite innverknad på mcq-verdiane.

### 7.1.5 CODA_WPTlim

CODA-versjonen har som nemnd i kap. 5.5 to moglege kriterium for å velja ut mcq-grensa. Grensekriteriet som er likt som for CODA_WPT2 viser mest potensial og vert derfor fokusert på.

#### *7.1.5.1 sim3-settet*

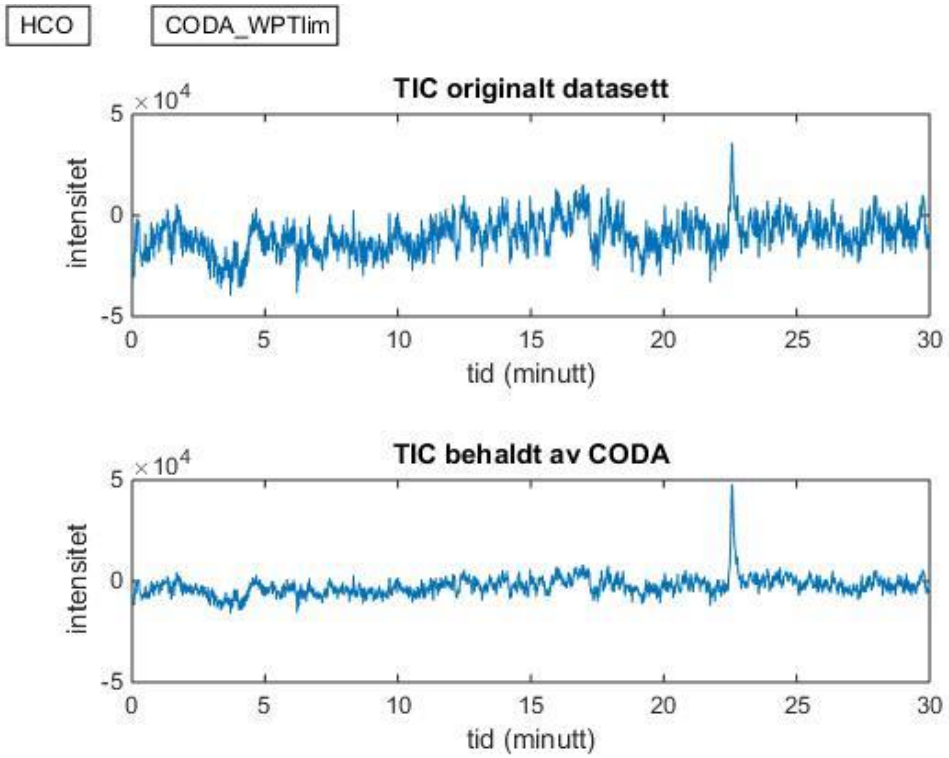
I *Figur 7.1-20* er EIC plotta for det CODA_WPTlim-prosesserte sim3-settet. WPT-delen i algoritmen vert berre nytta til å finna mcq-grensa og dei behaldte kromatografiske profilane er derfor uendra frå originalen. 17 av dei 40 massane vert fjerna, der ingen av desse inneheld nokre kromatografiske profilar (som vert bekrefta av TMS i Vedlegg 28). Ettersom støyen har vorte sentrert rundt 0, som er urealistisk for eit ekte datasett, har fjerninga av dei 17 massekromatogramma liten effekt på TIC av settet. (TIC er lagt ved i Vedlegg 29 )




Figur 7.1-20: *:* EIC av sim3-settet; (øvst) originalt datasett, (midtarst) CODA_WPTlim-prosessert datasettet og (nedst) signal fjerna av CODA-delen i metoden.

### 7.1.5.2 HCO-settet

For HCO-settet fjernar CODA_WPTlim 566 av dei 1004 massane, som førar til ei glattare baselinje i TIC, som er vist i Figur 7.1-21. Og av Vedlegg 30 der EIC er plotta, kan ein sjå at kvalitetstoppane er behaldt.

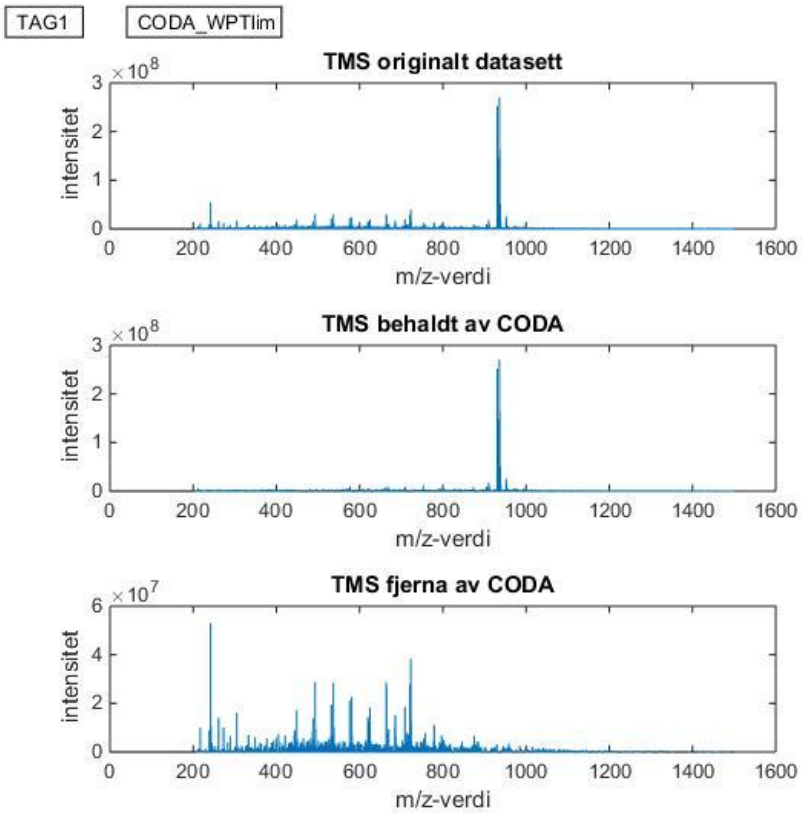



*Figur 7.1-21:* TIC av HCO-settet; (øvst) originalt datasett og (nedst) CODA_WPTlim-prosessert datasettet.

Det er tydeleg at det framleis er mykje støy igjen i datasettet. TIC av kvalitetsområdet er lagt ved Vedlegg 31, og der er det umogleg å sjå kvalitetstoppane. Det er ut frå observasjonane, mogleg at mcq-grensa burde vera høgare.

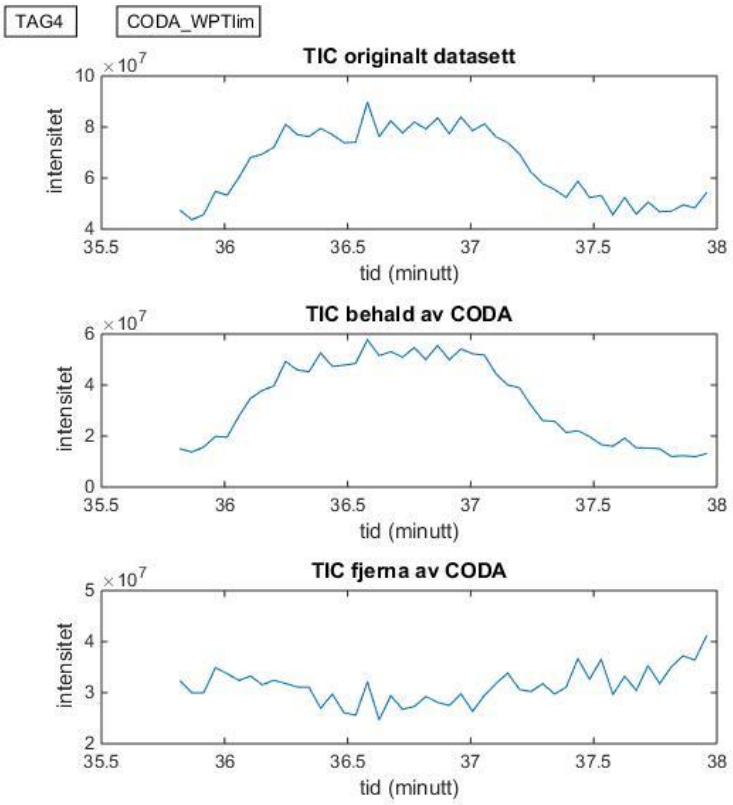
### 7.1.5.3 TAG-setta

For TAG1 vert 761 av dei 1501 massane forkasta, noko som gjer merkbare endringar i TMS av settet, i Figur 7.1-22. Av plotta kan ein sjå at store delar av støymassane er fjerna, og at analytten er bevart.

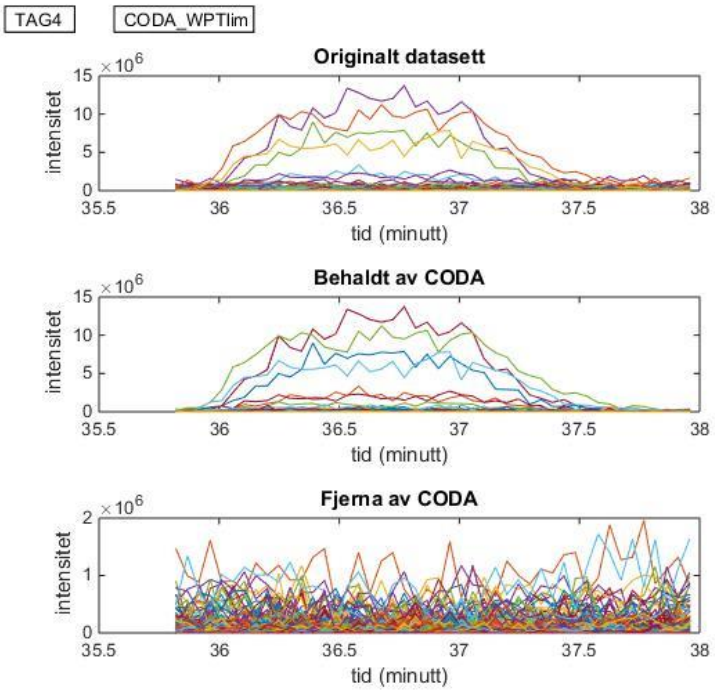



*Figur 7.1-22:* TMS av TAG1-settet; (øvst) originalt datasett, (midtarst) CODA_WPTlim-prosessert datasettet og (nedst) signal fjerna av metoden.

Den same trenden er gjeldane for dei andre TAG-setta. I Figur 7.1-3 og Figur 7.1-4 er TIC og EIC av TAG4 plotta. I TAG4 vert 756 av 1501 fjerna, og det kjem fram av TIC at baselinja har vorte meir enn halvert. EIC viser at ingen signifikant informasjon har gått tapt. TIC og EIC for TAG1-3 er lagt ved i Vedlegg 32 til Vedlegg 37.




*Figur 7.1-23: TIC av TAG4*-settet; (øvst) originalt datasett, (midtarst) CODA_WPTlim-prosessert datasettet og (nedst) det som vert fjerna av metoden.



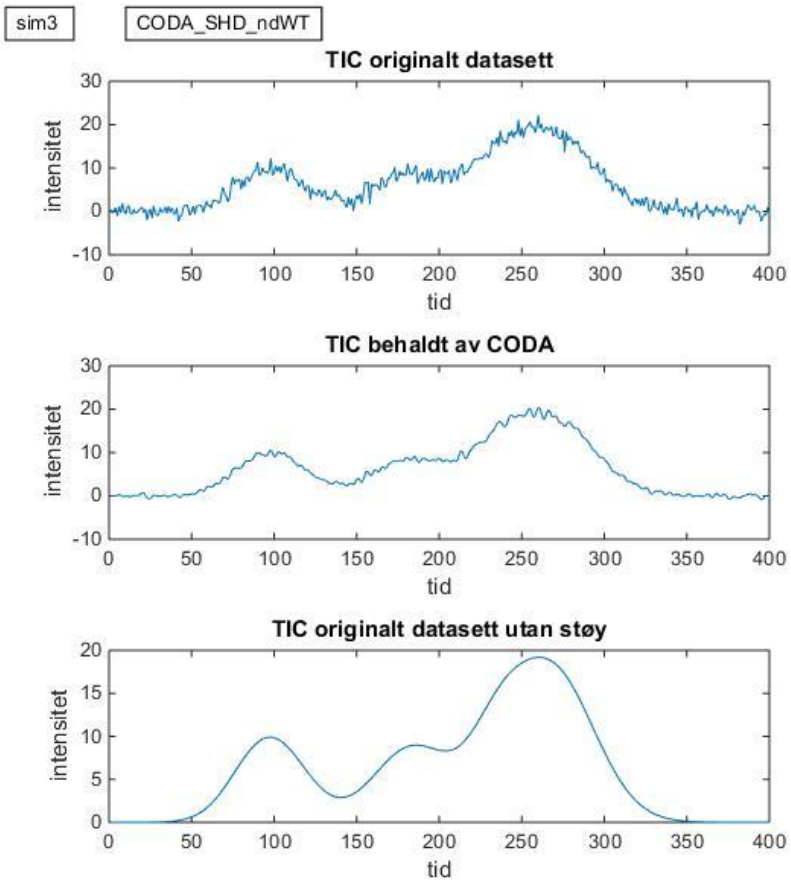

*Figur 7.1-24: EIC av TAG3*-settet; (øvst) originalt datasett, (midtarst) CODA_WPTlim-prosessert datasettet og (nedst) det som vert fjerna av metoden.

### 7.1.6 CODA_SHD_ndWT

CODA_SHD_ndWT vert testa for å sjekka om å dela opp i intervall aukar følsomheiten til ndWT-versjonen av CODA.

#### *7.1.6.1 sim3 -settet*

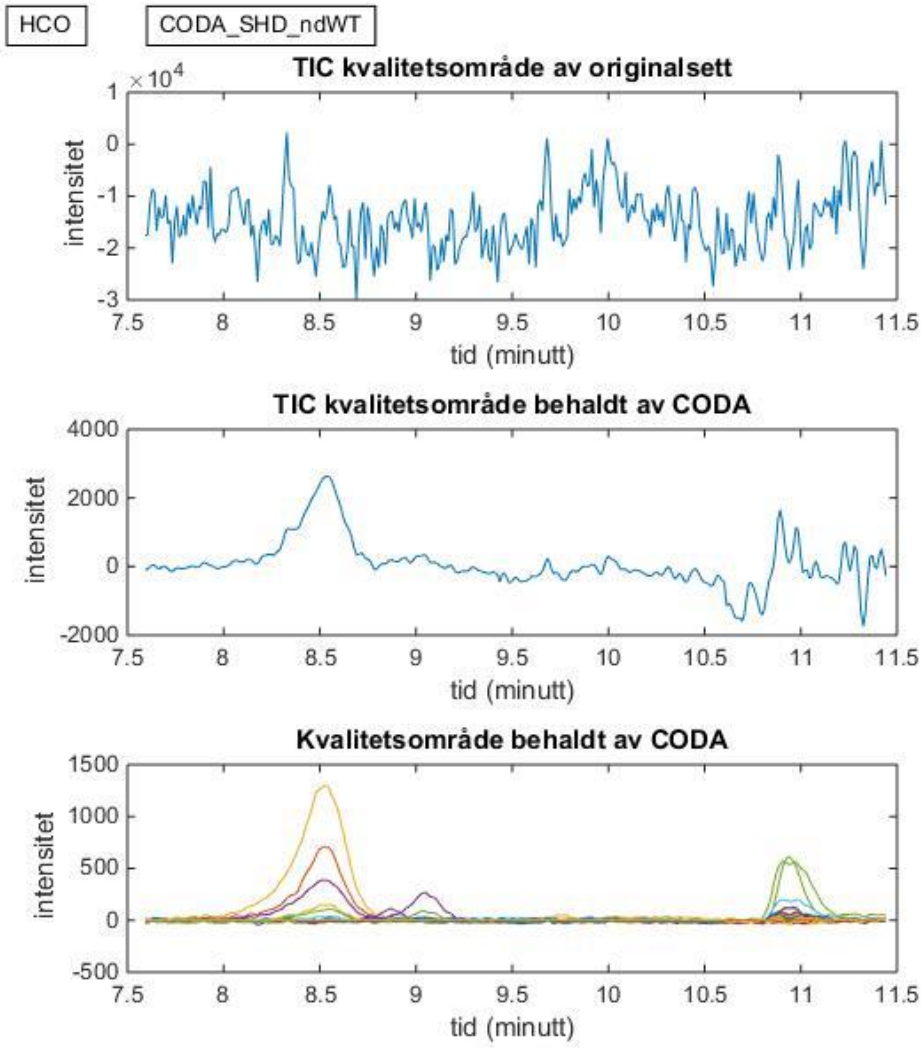
TIC av sim3 prosessert med SHD_ndWT-metoden er avbilda i Figur 7.1-25. TIC ser omtrent likt ut som for ndWT-metoden. Av dei 40 massane vert 15 behaldt i minst eit intervall og dei resterande 25 vert fjerna, medan for ndWT-metoden vart 14 massar behaldt og resten fjerna. EIC og TMS er lagt ved i Vedlegg 38 og Vedlegg 39, og viser at ingen signifikant informasjon har gått tapt.




*Figur 7.1-25: TIC av sim3*-settet; (øvst) originalt datasett, (midtarst) CODA_SHD_ndWT-prosessert datasettet og (nedst) det som vert fjerna av metoden.

#### *7.1.6.2 HCO-settet*

For HCO-settet fungerer CODA_SHD_ndWT veldig bra som baselinjefjerningsmetode for TIC. Av Vedlegg 40 kan ein sjå at TIC av settet vert mykje glattare enn i originalsettet. Det er framleis ein del ujamnheiter mellom 15 og 19 minutt, og nokre mindre ujamnheiter frå 11 til 15, men elles i settet er prosesseringa god. I Figur 7.1-26 er kvalitetsområdet plotta for TIC av det originale settet, TIC av det prosesserte settet og EIC av det prosesserte settet. Toppen **2** ved 8.5 viser godt igjen i TIC, og det kan sjå ut til at topp **3** ved 11.0 min også viser, men den ligg i starten av eit litt støyfullt område. Topp **1** ved 9.1 minutt viser ikkje igjen i TIC. Om ein fokuserer på EIC kan ein sjå at, til skilnad frå CODA_ndWT, har ikkje CODA_SHD_ndWT fjerna den nest mest intense massen til topp **1** . SHD_ndWT-algoritmen er altså meir skånsam enn ndWT-algoritmen. Når ein reknar ut mcq-verdiar for eit mindre intervall er det mindre sjanse for at signalet druknar i støyen, som dei ulike resultata for topp **1** stadfestar. Av dei fjerna EIC, i Vedlegg 41, er det også tydeleg at nokre massar berre vert fjerna innan visse intervall, då det f.eks. er eit skift ved 11 minutt.




*Figur 7.1-26: Kvalitetsområdet i HCO*-settet; (øvst) TIC av originalt datasett, (midtarst) TIC av CODA_SHD_ndWT-prosessert datasettet og (nedst) EIC av det prosessere datasettet.

### *7.1.6.3 TAG-setta*

For CODA_SHD_ndWT er det ikkje aktuelt å testa på TAG1-TAG4 ettersom desse er utplukka toppområder og algoritmen er laga for å ta inn sett bestående av fleire toppområder.

## 7.2 Testing av kvalitetsmål

### 7.2.1 Singulærverdi-forhold - metoden

Testing av metoden går i hovudsak ut på å finna kriterium for å fastsetja den siste store singulærverdien ($s_k$) for datasetta, der k er den kjemiske rangen til settet. Kriteria vart, som oppgjeve i kap. 6.1; $s_i = s_k$ dersom $s_i > 4 \cdot s_{i+1}$ eller $s_1 > 10 \cdot s_{i+1}$ . Desse kriteria er baserte på testing på datasetta Q og K, som er undersett av TAG (sjå Tabell 3-5) . I Figur 7.2-1 (under) og Vedlegg 44 viser TIC ved CODA på setta. Av desse datasetta kan ein sjå at det prosesserte settet får høgare F-verdi enn originalsettet, og at det fjern settet har låg verdi . Samtidig er det ingen struktur i TIC av dei fjerna massane og det ser ikkje ut som om noko analytisk signal er fjerna i det prosesserte settet. Vedlegg 44 viser resultatet for K, som visar den same trenden.

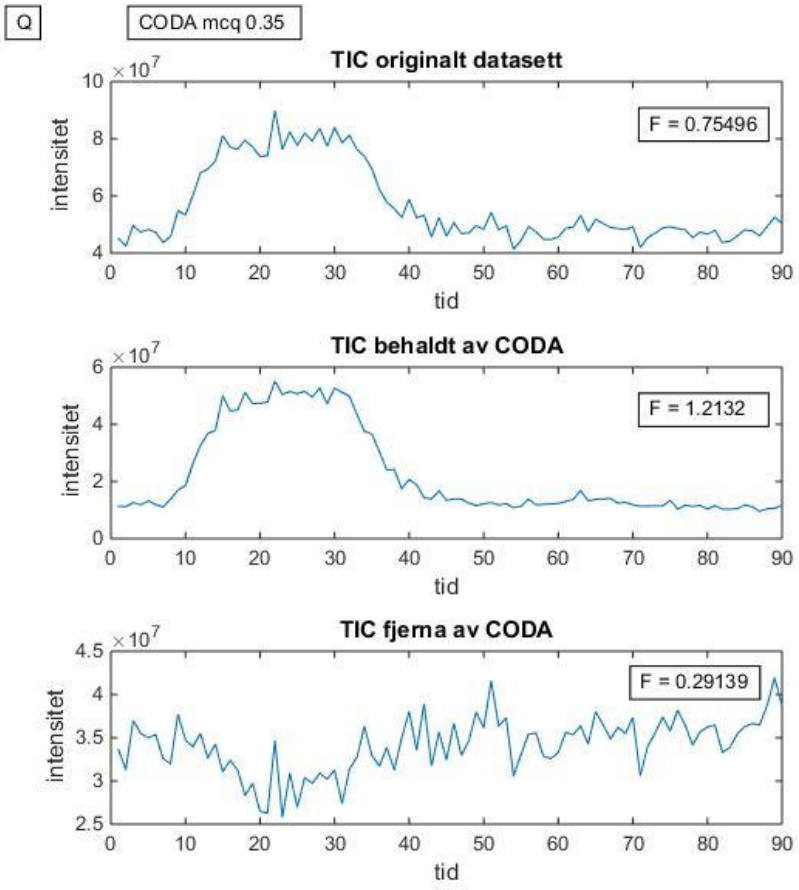
Det er kjent at sim3-settet har 5 ulike forbindelsar, men her veljar algoritmen å setja $s_k$ som verdi nr. 4. Forbindelsen i sim3 med minst intensitet er til gjengjeld berre eit ion (masse), og den femte singulærverdien med verdi 6.9 er likare til nr. 6 som er 4.9 enn til nr. 4 som er 20.7. Singulærverdi nr. 4 vart diskutert i kapittel 4.2, der singulærverdiane og logaritmen av dei er plotta i Figur 4.2-3. Der vart trukke fram at komponent nr. 2 i settet i stor grad overlappar med variasjonen i komponent nr. 1. Observasjonane for sim3-settet avdekker at målet er sårbart når ein har forbindelsar som i stor grad overlappar, og når nokre forbindelsar har ein høg samla intensitet av fragmenta samanlikna med andre.

Ved testing på HCO og ZM2 setta er resultata mindre bekreftande. HCO-settet har som nemnd ein topp som er mykje større enn dei andre komponentane, som igjen slår ut på singulærverdiane (Vedlegg 3). Algoritmen vel ut verdi nr. 7 som siste verdi, men ut frå grafen til singulærverdiane er det 9 komponentar i settet.

For ZM2-settet vert det valt ut 9 singulærverdiar, men i dette settet er det store forskjellar i intensitet mellom dei største og minste toppane, og det er eit veldig komplekst sett. Av plot

av singulærverdiar (Vedlegg 45) kan det sjå ut som om alt frå 6 til over 40 av dei 701 verdiane kan vera signifikante.

Når singulærverdi-forhold-metoden unngår å velja ut ein eller fleire signifikante singulærverdiar, dvs. forbindelsar som ein vil behalda, kan resultatet vera alvorleg. Det prosesserte settet kan få høgare kvalitetsindeks enn originalsettet ved å fjerna dei signifikante forbindelsane. Desse forbindelsane vil igjen, enda opp i det fjerna settet, som kan få ein høgare verdi enn viss forbindelsane ikkje vart fjerna i prosesseringa. Viss signifikante forbindelsar vert fjerna, så kan dette vera pga. for strenge prosesseringskrav, f.eks. for høg mcq_grense. Dette kan igjen føra til at fleire støymassar endar opp i det fjerna settet, som igjen truleg vil senka det fjerna settet sin kvalitetsindeks, ettersom mesteparten av dei vil enda opp under brøkstreken i Formel 6.1-1. Derfor er det ikkje sikkert at det fjerna settet sin kvalitetsindeks vil auka, sjølv om den inneheld signalet til signifikante forbindelsar.



*Figur 7.2-1: Område Q ved CODA for mcq = 0.35 . Singulærverdi-forholdet F er gjeve for (øvst) det originale datasettet, (midtarst) det som vert behaldt av CODA og (nedst) det som vert fjerna av CODA.*

### 7.2.2 Toppsamanlikning-metodane

Dei to algoritmane, basert på toppsamanlikning, har mindre behov for testing enn singulærverdi-metoden, ettersom inputverdiane er få.

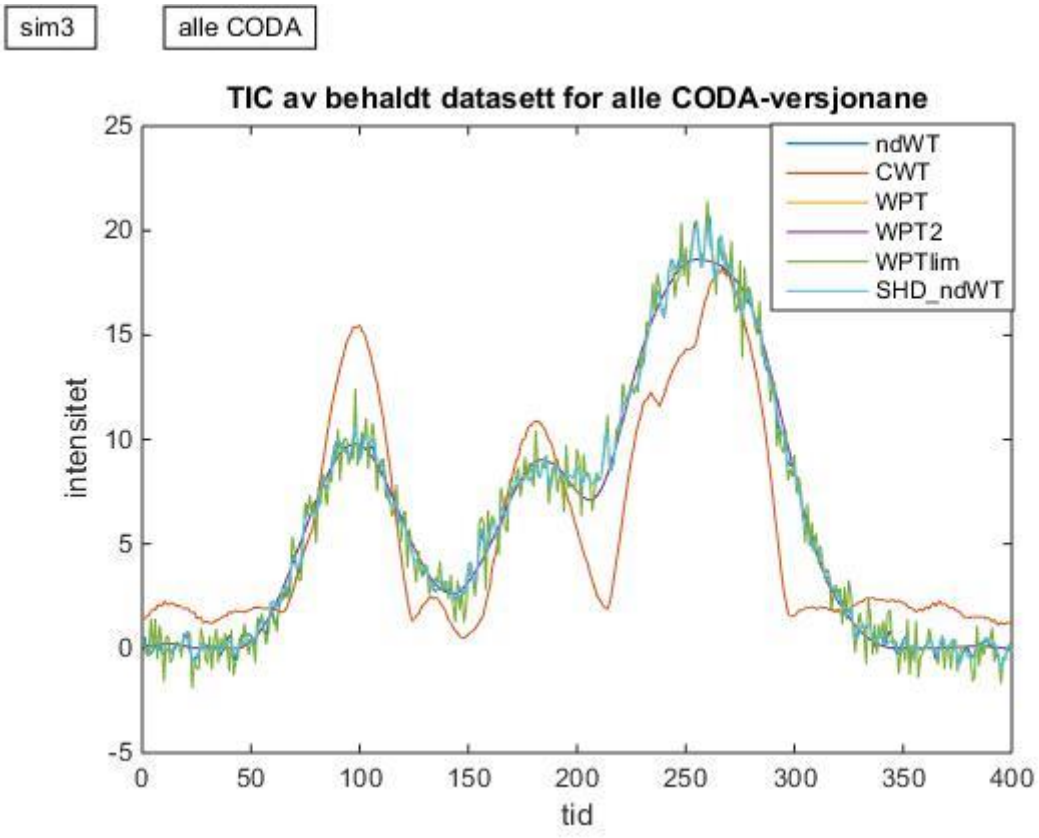
## 7.3 Samanlikning av CODA-versjonar

### 7.3.1 Samanlikning av TIC

I dette kapittelet vert det gjort samanlikningar av TIC til metodane. Det vert også samanlikna mot CODA med mcq-grensa 0.85, som har vorte nemnd av Windig [10] som ei fornuftig grense.

#### *7.3.1.1 sim3-settet*

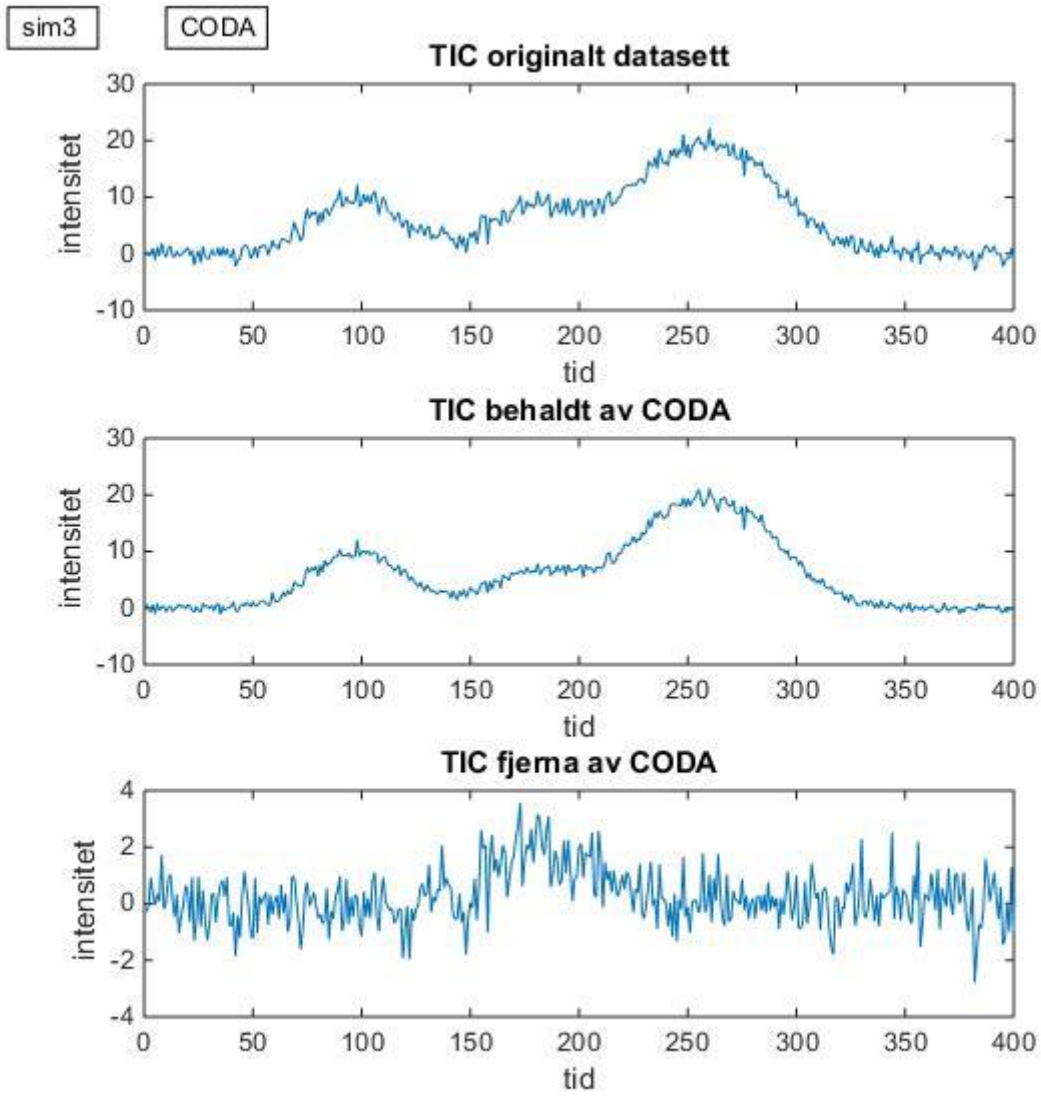
Figur 7.3-1 er det gjort ei grafisk samanlikning av dei nye CODA-versjonane for sim3-settet.




*Figur 7.3-1: Samanlikning av TIC til det behandla datasettet for kvar CODA-versjon, for sim3.*

Av figuren er det klart at det er CODA_CWT som skiljar seg mest ut. Som nemnt tidlegare ( i kap. 7.1) minskar CWT toppbreidda til dei kromatografiske profilane, som forklarar endringa i fasongen til toppklynga rundt 250 på tidsaksen. Viss ein ser vekk frå CODA_CWT, kan ein dela dei resterande resultata i to; WPT-behandla og ikkje-WPT-behandla datasett. Dei to WPT-behandla setta er nærmast støyfrie, medan dei ndWT- og SHD_ndWT- behandla framleis inneheld ein del støy.

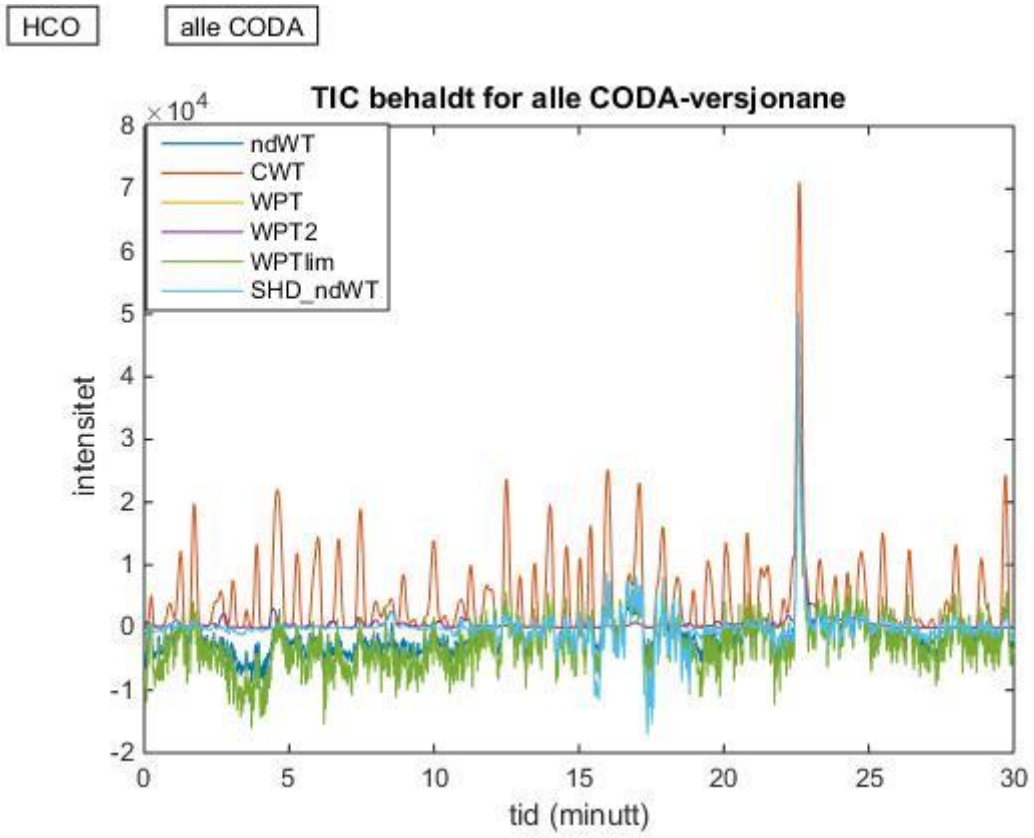
Figur 7.3-2 viser resultatet for CODA med mcq = 0.85. Det er tydeleg av TIC til dei fjerna massane at det har vorte fjerna analytisk signal frå den midtre av dei tre tydelege toppane i TIC. Av valideringskapittelet (7.1) veit ein at dei nyutvikla beheld alt av analytt-massar for settet, og dei utkonkurrerer dermed CODA med mcq = 0.85.



*Figur 7.3-2: TIC av CODA med mcq = 0.85*

### *7.3.1.2 HCO-settet*

Ei grafisk samanlikning av CODA-metodane er vist for HCO-settet i Figur 7.3-3.




*Figur 7.3-3: Samanlikning av TIC av kvar CODA-versjon, for HCO-settet*

Det er stor forskjell på TIC representasjonen av det prosesserte settet for dei ulike CODA-versjonane. Ut frå denne figuren ser CODA_WPTlim og CODA_ndWT ganske misslykka ut, men viss ein ser på TIC av original-settet har den «verste» av dei, CODA_WPTlim, fjerna over halvparten av støy-summen. (Summen er negativ.) Parameterane, og metodane generelt, er ikkje satt med mål om å fjerna all støy, men for å fjerna støy og behalda alt signal, så det vert feil å rekna resultata frå WPTlim- og ndWT–versjonane som feilslåtte.

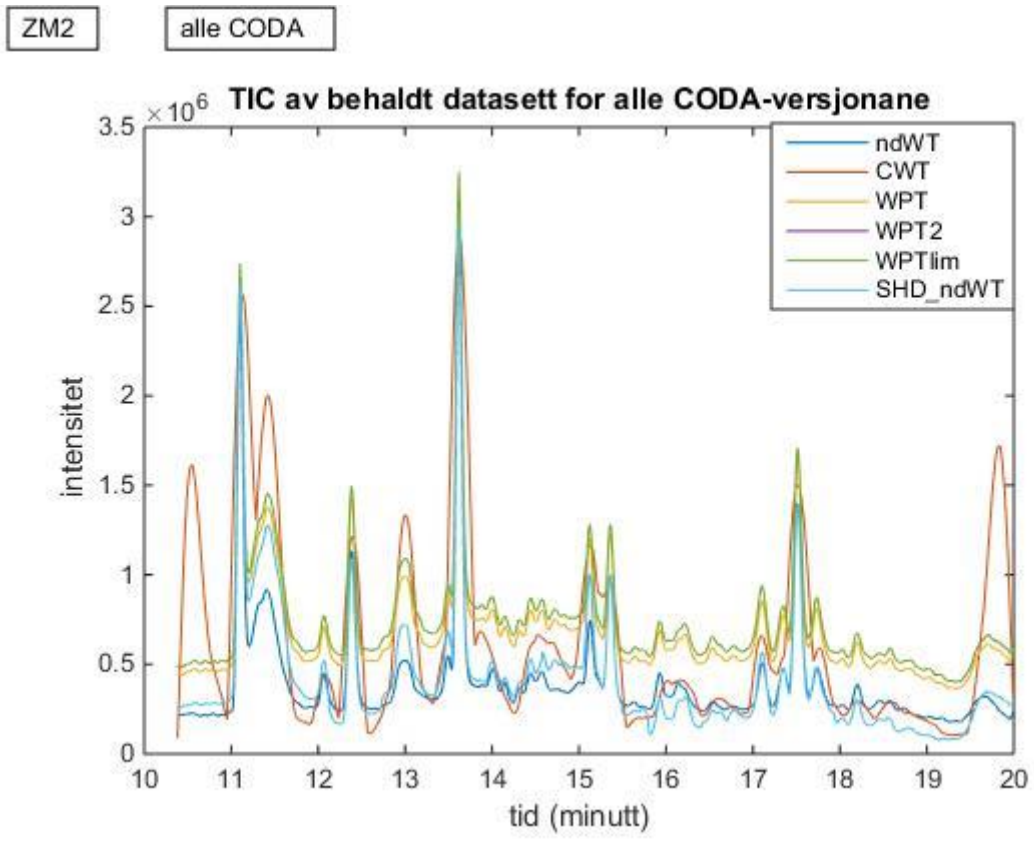
CODA_CWT gjer dei klarast definerte toppane for dei kjente signal-massane, men legg også til store toppar som ikkje høyrar heime i originaldata (,viss ein ser på plott av M) . CODA_WPT2 og CODA_WPT, som overlappar i plottet, ser ut til å gje den beste representasjonen, fordi både

topp 2 og 3, ved tidene 8.5 og 11.0 (Tabell 3-11),visar igjen i TIC. Også CODA_SHD_ndWT visar topp 2 og 3, men topp 3 er i større grad støyfull enn for WPT-algoritmane.

For TIC av CODA-prosesserte data med mcq-grensa 0.85, i Vedlegg 46, er mesteparten av støyen fjerna. Viss ein ser på kvalitetsområdet (Vedlegg 47) er toppen ved 8.5 minutt synleg, men ved inspeksjon av EIC (Vedlegg 48) kjem det fram at metoden fjernar massar frå kvalitetstoppane.

#### *7.3.1.3 ZM2-settet*

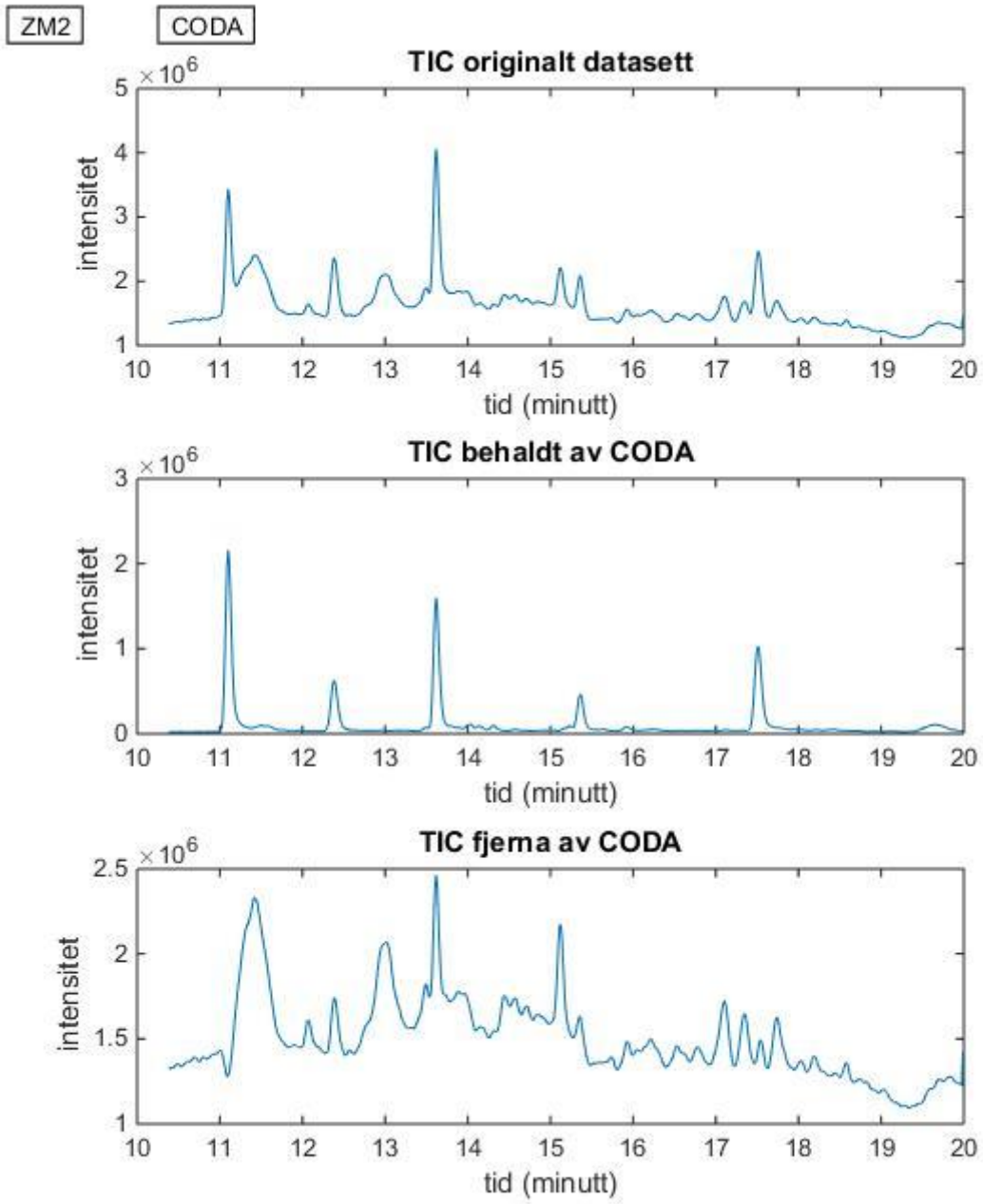
ZM2-settet er av ein annan natur enn sim3 og HCO, då mykje av støyen er baselinje-støy som gjerne ligg i massar som også inneheld signal. Det visar seg at støyfjerningseigenskapane til CODA-versjonane skiljar seg meir for dette settet enn for dei andre. TIC for dei ulike metodane er vist i Figur 7.3-4.




*Figur 7.3-4: Samanlikning av TIC av kvar CODA-versjon, for ZM2-settet. TIC til WPT2-metoden er lite synleg i plottet, ettersom det overlappar med WPT.*

CODA_SHD_ndWT ser ut til å gje best resultat, som ikkje er uventa, ettersom den delar opp massane i tidsintervall. ndWT-metoden fjernar nesten like mykje baselinje som SHD_ndWT-metoden fordi det lågaste mcq-kriteriet her er nytta for SHD_ndWT-metoden. WPTlim, WPT og WPT2 fjernar særs lite støy, og skiljar seg her frå dei andre CODA-versjonane. CODA_CWT skiljar seg ut, som før, med å summera opp store signal der dei andre metodane ikkje har signal i TIC.

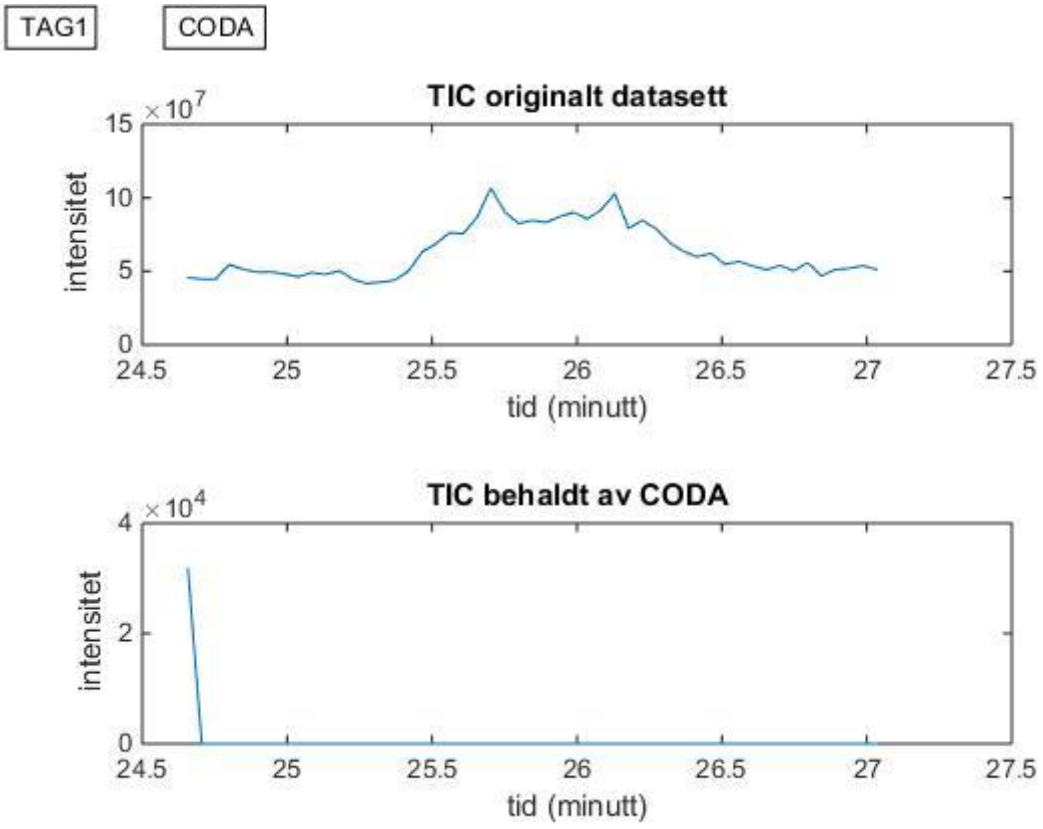
CODA med mcq på 0.85 i Figur 7.3-5 gjer eit oversikteleg TIC, men fjernar store delar av datasettet. Dette førar til at dei nye CODA-versjonane er å føretrekka, dersom ein vil fjerna støy utan å ta vekk analytisk signal.




*Figur 7.3-5: TIC ZM2-settet; (øvst) originalt datasett og (nedst) prosessert av CODA med mcq = 0.85.*

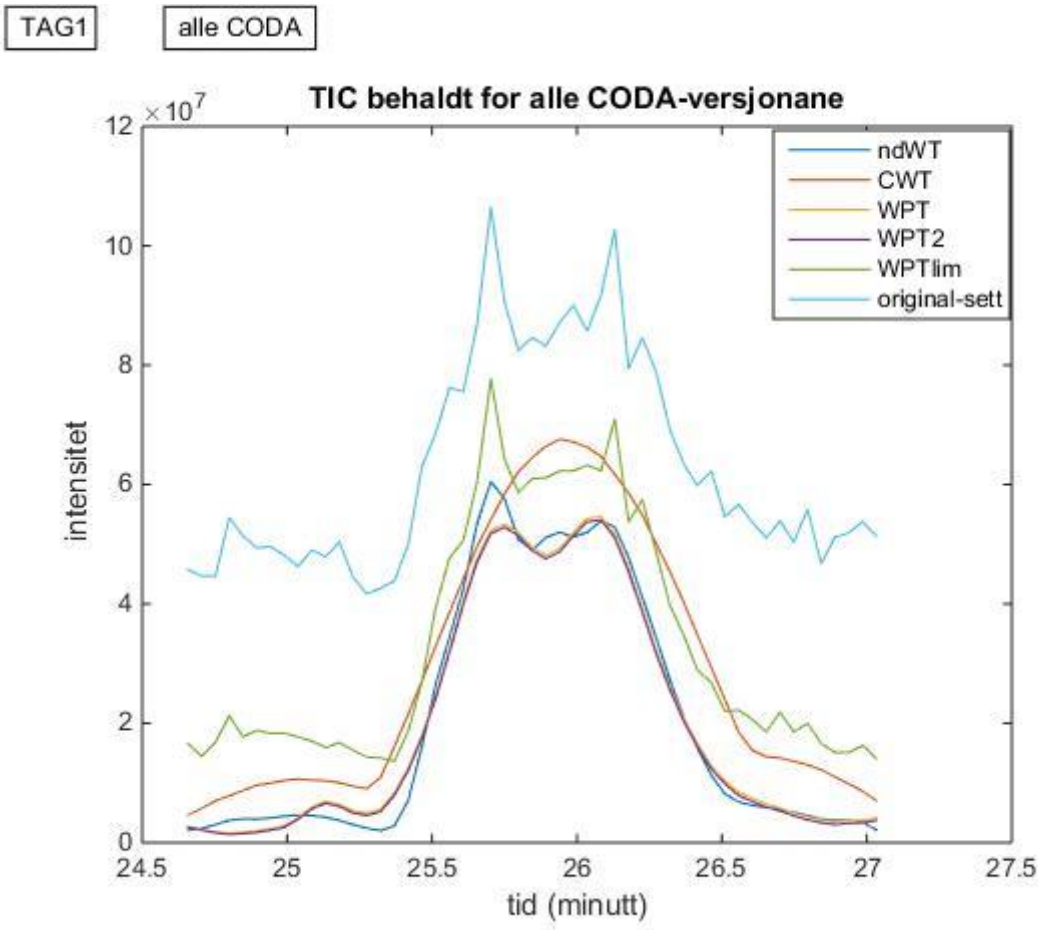
### *7.3.1.4 TAG-setta*

I Figur 7.3-6 er TIC for det originale og det CODA prosesserte TAG1-settet for mcq 0.85 plotta. Det er tydeleg at mcq-verdien er sett for høgt, ettersom toppsignalet er vekke. Det same vert observert for TAG2, TAG3 og TAG4 , i Vedlegg 49- Vedlegg 51.




*Figur 7.3-6: TIC TAG1-settet; (øvst) originalt datasett og (nedst) prosessert av CODA med mcq = 0.85.*

Dei nyutvikla CODA-versjonane fungerer mykje betre enn vanlig CODA, for TAG-setta. I Figur 7.3-7 er TAG1 plotta som eksempel. For TAG1 vert ingen av analyttmassane fjerna for nokon av metodane, samtidig som baselinja vert redusert betraktelig for alle metodane. Her gjer CWT-metoden best form for signalet, medan ndWT-, WPT- og WPT2- metodane fjernar mest støy. TAG2 får heller ingen fjerna signal for nokon av dei signifikante forbindelsane. TAG3 og TAG4 mistar nokre lågkvalitets-profilar for WPT-metoden, men mister elles ingen signifikante massar.




*Figur 7.3-7: TIC for TAG1;  alle CODA-versjonane og originalsettet.*

### 7.3.2 Singulærverdi-metode

Kvalitetsindeksane vert forsøkt tolka i kombinasjon med tidlegare resultat, for setta sim3, HCO og ZM2.

Kvalitetsindeksane for testing på sim3-settet vert gjeve i *Tabell 7-9*:

*Tabell 7-9:  F-verdiar for originalsett, behaldt sett, sett fjerna av CODA og sett fjerna av både WT og CODA,  for dei ulike CODA-versjonane for settet sim3*

| sim3 | F-verdi | | |
|---|---|---|---|
| | Behaldt | Fjerna av CODA | Fjerna av WT og CODA |
| ndWT | 5,85 | 0,25 | 0,16 |
| CWT | 2,70 | - | 0,66 |
| WPT | 14,39 | - | 0,18 |

| | | | |
|---|---|---|---|
| WPT2 | 14,39 | - | 0,18 |
| WPTlim | 2,41 | 0,37 | 0,37 |
| SHD_ndWT | 6,04 | 0,21 | 0,16 |
| | F-verdi | | |
| M | 1,37 | | |

*Tabelltekst til Tabell 7-9: Rutene utan verdi er grunna at CODA ikkje fjernar kromatogram for dei gjevne metodane.*

CODA-versjonane WPT og WPT2 gjer som venta best resultat for sim3, ettersom dei fjernar omtrent all støy, og beheld signalet. Det gjer meiningar at F indikerar dårlig kvalitet i CWT, ettersom algoritmen fjernar mykje av breidda til dei kromatografiske profilane. SHD_ndWT og ndWT har omtrent like F-verdiar, som og samsvarar med tidlegare resultat. CODA_WPTlim gjer dårlegast resultat for det som er behaldt, som kan ha med at ingen WT er utført å gjera.

Kvalitetsindeksane for HCO-settet er gjeve i Tabell 7-10:

*Tabell 7-10:* F-verdiar for originalsett, behaldt sett, sett fjerna av CODA og sett fjerna av både WT og CODA, for dei ulike CODA-versjonane for *settet HCO*

| HCO | F-verdi | | |
|---|---|---|---|
| | Behaldt | Fjerna av CODA | Fjerna av WT og CODA |
| ndWT | 0,97 | 0,19 | 0,10 |
| CWT | 1,25 | - | 0,21 |
| WPT | 2,76 | - | 0,11 |
| WPT2 | 2,76 | - | 0,11 |
| WPTlim | 0,46 | 0,14 | 0,14 |
| SHD_ndWT | 0,72 | 0,18 | 0,12 |
| | F-verdi | | |
| M | 0,26 | | |

Av Tabell 7-10 vert WPT og WPT2 rekna som dei beste CODA-versjonane for HCO-settet, fordi desse har høgst F verdi for det behaldte settet og er blant versjonane med lågast F for det fjerna settet. Dette stemmar godt overeins med dei øvrige resultata for settet der WPT og WPT2 ser ut til å fjerna all støy og behalda signalet. CWT-versjonen, sin verdi for det behaldte er ikkje samanliknbare, ettersom settet har ein ny struktur. CWT har den høgaste F-verdien for settet som er fjerna av WT og CODA, som gjer meining, ettersom det er kjent at settet inneheld mykje av signalet, grunna differansen mellom det prosesserte settet og det originale.

Kvalitetsindeksane for ZM2-settet er gjeve i Tabell 7-10:

*Tabell 7-11: F-verdiar for originalt datasett, data behaldt av CODA, data fjerna av CODA og data fjerna av WT og CODA av dei ulike CODA-versjonane for settet ZM2*

| ZM2 | F-verdi | | |
|---|---|---|---|
| | Behaldt | Fjerna av CODA | Fjerna av WT og CODA |
| ndWT | 2,43 | 0,74 | 0,69 |
| CWT | 1,89 | - | 1,38 |
| WPT | 1,79 | 0,93 | 0,87 |
| WPT2 | 1,74 | 1,00 | 0,93 |
| WPTlim | 1,73 | 0,99 | 0,99 |
| SHD_ndWT | 1,54 | 0,86 | 0,81 |
| | F-verdi | | |
| M | 1,45 | | |

For ZM2-settet er det lite forskjell mellom F-verdiane til dei ulike metodane. At variasjonen er liten kan ha med at settet er stort, med mange store singulærverdiar, at dei største singulærverdiane alltid vert behaldt og at dei neste som regel også vert behald. ndWT-metoden har høgast F-verdi, som kan relaterast til at metoden beheld berre 152 av dei originalt 701 massane, medan dei andre metodane alle beheld over 300 massar. (jf. Tabell 7-8)

### 7.3.3 Toppsamanlikning-baserte kvalitetsmål

For desse kvalitetsmåla vert berre transformasjonsdelen av CODA-versjonane vurdert, ettersom toppsamanlikningsmetodane baserar seg på endring i utvelde signalmassar, der desse må vera til stades i både det originale settet og det prosesserte. Det vert fokusert på sim3-settet og kvalitetsområdet til HCO. Dei utvaldte toppane for sim3 og HCO vert gjevne i Tabell 7-12 og Tabell 7-13, repektivt.

*Tabell 7-12: Utvalde kvalitetstoppar for sim3*

| Toppområder (nr.) | Elusjon-start (rad nr.) | Elusjon-stopp (rad nr.) | Antall massar |
|---|---|---|---|

| | | | |
|---|---|---|---|
| 1 | 46 | 150 | 2 |
| 2 | 140 | 228 | 4 |
| 3 | 180 | 334 | 4 |

*Tabelltekst til Tabell 7-12: Nummereringa for toppområder er ikkje lik som for nummerering av toppar i Tabell 3-2*

*Tabell 7-13: Utvalde kvalitetstoppar for HCO*

| Toppområde (nr.) | elusjon-start (minutt) | elusjon-stopp (minutt) | antall massar |
|---|---|---|---|
| **2** | 7.96 | 8.84 | 5 |
| **3** | 10.79 | 11.19 | 5 |

*Tabelltekst for Tabell 7-13: For denne tabellen har nummereringa for HCO-settet (Tabell 3-11) vorte nytta. Topp **1** har ikkje vorte teken med ettersom den mistar 1 av 2 massar for CODA_ndWT. Det er vald ut 5 toppar for kvart område. Dette er fleire enn dei to iona som er kvalitetssikra for kvart område, men ein går ut frå at alle høyrar til forbindelsane pga. forma til toppane.*

Fordi toppsamanlikningsmåla berre vert nytta til å avgjera kvaliteten på WT vert ikkje WPTlim metoden teken med, ettersom den ikkje utførar nokon transformasjon på settet. Det vert heller ikkje testa dobbelt for like WT. WPT2 har som kjent den same WT som WPT, og SHD_ndWT har same WT som ndWT. Toppsamanlikningsmåla for sim3 vert gjeve i Tabell 7-14.

*Tabell 7-14: Kvalitetsverdiar for originalt datasett og data behaldt av CODA for dei ulike WT for settet sim3*

| sim3 | Behaldt datasett | |
|---|---|---|
| | norm _kval | korr_kval |
| ndWT | 0,9974 | 0,9692 |
| CWT | 0,9969 | 0,9868 |
| WPT | 0,9993 | 0,9915 |
| | Originalt datasett | |
| | norm _kval | korr_kval |
| | 0,9942 | 0,9357 |

WPT transformasjonen er den beste transformasjonsmetoden ut frå både det normbaserte og det korrelasjonsbaserte målet, for sim3 settet. ndWT-transformasjonen har forøvrig høgare verdi enn CWT for norm-målet og lågare for korrelasjons-målet. Av Vedlegg 54 og

Vedlegg 55 ser ein at både toppområde 2 og 3 har skeivheiter i toppane ved CWT, noko som antakeleg har større utslag for det norm-baserte målet enn for det korrelasjonsbaserte.

Toppsamanlikningsmåla for HCO-settet er gjeve i Tabell 7-15.

*Tabell 7-15: Kvalitetsverdiar for originalt datasett og data behaldt av CODA for dei ulike WT for settet HCO*

| HCO | Behaldt datasett | |
|---|---|---|
| | norm _kval | korr_kval |
| ndWT | 0,9862 | 0,9686 |
| CWT | 0,9990 | 0,9952 |
| WPT | 0,9905 | 0,9634 |
| | Originalt datasett | |
| | norm _kval | korr_kval |
| | 0,9844 | 0,9612 |

Ved samanlikning av toppsamanlikningsmål for HCO-settet viser det seg at CWT-transformasjonen gjer høgast verdi for begge måla. CWT-metoden nyttar som kjent ein *Mexican hat*-wavelet, og gjer toppane tilnærma gaussiske. Som nemnt, i gjennomgangen av CWT metoden, har CWT med *Mexican hat*-waveleten ein tendens til å gje negative verdiar på sidene av den kromatografiske toppen (sjå waveletformen i Figur 2.4-1), og negative verdiar vert påfølgjande nullsett. Ettersom CWT endrar minskar breidda på toppane, endar ein om med ei rekke av nullsette verdiar på kvar side av toppane. (jf. Vedlegg 53 og Vedlegg 55) Desse verdiane vil ikkje påvirka toppsamanlikningsmåla, ettersom norm og korrelasjon ikkje vert påverka av ekstra nullverdiar, viss dei er gjeldane for alle involverte vektorar. Derfor fungerer kvalitetsmåla som om toppområdet var smalare, enn det definerte området, i utgangspunktet. Dette er gjeldane for både HCO og sim3 setta.

Som det vert illustrert av CWT-drøftinga over, er ein svakheit for toppsamanlikningsmetodane er at dei ikkje stillar krav til likskap mellom det originale toppområdet og det nye toppområdet, forutan at dei må vera plasserte i same intervall .

Kvaliteten til WPT- og ndWT- transformasjonane til HCO vert fordelt motsett for dei to kvalitetsmåla. For norm-målet har WPT høgast verdi, og for korrelasjonsmålet har ndWT høgst verdi. Av testinga for WPT og ndWT på HCO kjem det fram det ved WPT oppstår forskyving av toppunktet til enkelte av dei små massane ( jf. Vedlegg 56 og Vedlegg 57), men at dette ikkje skjer for ndWT (jf. Vedlegg 58 og Vedlegg 59 ). Av desse massane er 4 nytta i kvalitetstesten,

som kan sjå ut som at for HCO er korrelasjonsmålet meir sensitivt for skeivhet i toppane enn norm-målet. Dette er motsett trend av det ein såg for sim3-settet og CWT. Ein meir truleg teori er at korrelasjonmålet er meir sensitivt for ekstra toppar, som kjem fram av Vedlegg 57. Det visar seg at toppområde 2 i HCO ved WPT har ein ekstra småtopp til venstre for hovudtoppen, for to av massane. Det er lite truleg at WPT er skyldig i å konstruera toppen for dei to massane. Det er meir sannsynleg at dei anten har vorte lagt til to av massane til kvalitetstoppen ved binning.

## 7.4 Totalsamanlikning

Basert på mcq-fordelinga til a1 og d1 for dei ulike setta, kan ei mcq-grense definert ut frå ein posisjon i d1 i minst kallast ein god peikepinn på kvar mcq-grensa bør setjast.

CODA_ndWT visar potensial, ved at den plukkar ut ei mcq-grense, basert på vindauge-storleik for CODA, wavelettype og grense-kriteriet som kan ta fire verdiar. Det er framstår som enklare å bestemma wavelettype og grensekriterium, enn å velja ut mcq-verdi manuelt. Å fjerna d1-leddet frå datasettet før CODA vert utført visar seg å vera lurt, ettersom ein då separerar (a1-) verdiane over og under d1. Å trekkja frå d1-leddet, dvs. berre nytta a1, er også ein skånsam transformasjon av data, som i dette arbeidet ikkje har vist seg å endra symmetri, eller andre viktige eigenskapar, i toppområder som består av fragment frå same forbindelse.

CODA_ndWT kan få problem for større sett, med mange retensjonstider og med store forskjellar for toppintensitetar. I desse tilfella kan ein risikera å mista massar, som observert for HCO-settet. I dette tilfellet er det gunstigare å kombinera algoritmen med CODA_SHD som vart introdusert av Sandve [11]. CODA_SHD_ndWT viste gode resultat for alle setta, og fjerna ikkje massen som ndWT-algoritmen fjerna for HCO-settet. Algoritmen viser også gode resultat for fjerning av støy i TIC, slik at dei signifikante komponentane vert synlege.

CODA_WPT og CODA_WPT2 har vist seg å vera gode til å fjerna all støy ved å finna den beste wavelet-basisen, terskla den og så å tilbake-transformera. Kombinasjonen av WPT og CODA kan for sim3 og HCO virka unødvendig, ettersom WPT fjernar all støyen. For ZM2-settet får

ein nytte av kombinasjonen av CODA-delen av algoritmane i tillegg til WPT, der den CODA fjerna store delar av baselinja. TAG-setta vert også ekstra reduserte av CODA-delen til WPT-metodane. For TAG3-4 vert mcq-grensa til CODA_WPT for høg, men CODA_WPT2 fungerer bra. Eit problem som kan oppstå ved WPT-transformasjonane er at WPT ikkje behandlar hovudstrukturen skånsamt nok, anten ved terskling eller utveljing av basis, som kan føra til små forskyvingar i massane, som førar til asymmetri i toppområda.

CODA_WPTlim baserar seg berre på mcq-grense-utveljinga frå CODA_WPT2 (, ettersom CODA_WPT sitt kriterie fungerte dårleg). CODA_WPTlim gjer ok resultat for HCO, sim3, ZM2 og TAG-setta, men visar seg å vera den minst effektive metoden med hensyn på støyfjerning, som kjem fram av TIC-samanlikninga for HCO, ZM2 og TAG1.

Ein ugunstig ting med WPT-versjonane er at ein må velja ein skaleringsparameteren (amp) manuelt, ettersom den automatiske terskelverdien th (jf. Formel 5.3-1) ikkje er stor nok for nokre av setta.

CODA_CWT, med *Mexican hat*-waveleten, har vist seg å vera den minst pålitelege algoritmen, av fleire grunnar. For det fyrste er vert toppane si kromatografiske breidd kutta ned. Den andre grunnen er at intensitet-forholdet mellom toppar vert forskyve i forhold til original-data. Tredje grunnen er at alt data vert omgjort til signal-liknande toppar, slik at dei adderer seg opp i TIC og førar til store toppar som ikkje er signal.

Singulærverdi-forhold kvalitetesmålet har ei ugunstig utveljingsalgoritme, dersom ein har ein eller fleire massar med stor intensitet i forhold til andre signalmassar. Her kunne det ha vorte utarbeida eit meir robust kriterium, f.eks. basert på når kurva slakkar ut, for å få med alle store verdiar store singulærverdiar. Det vil gje større utslag i F med ein feilplassert stor singulærverdi, enn ein feilplassert liten verdi. Verdiane til kvalitetsmålet gjer til dels meining for sim3 og HCO, men er lite nyttig for ZM2-settet. Eit anna problem knytta til kvalitetsmålet er at ein singulærverdi kan visa overlappande trekk frå liknande komponentar. Dette kan føra

til at likskap mellom komponentar har noko å sei, sjølv om dette ikkje naudsamt har noko med kvaliteten på settet å gjera.

Dei toppsamanlikning-baserte kvalitetsmåla er berre eigna til å sjekka støyfjerningsmetodar som ikkje fjernar massar, og vart derfor nytta til å sjekka kvaliteten av wavelet-transformasjonane, i dette arbeidet. Det normbaserte målet ser ut til å vera meir sensitivt for asymmetri enn korrelasjonsmålet. Metodane stillar ikkje krav til likskap mellom originalt datasett og prosessert datasett, som er ein svakheit. Viss ein utelukkar CWT-metoden og resultat for korrelasjonsmålet av WPT, der WPT avdekkar ein binningfeil for HCO, så vert WPT rekna som den beste wavelet-transformasjonen av WPT og ndWT.

# 8 Konklusjon

Det vart laga seks WT-videreutviklingar av CODA; CODA_ndWT, CODA_CWT, CODA_WPT, CODA_WPT2, CODA_WPTlim og CODA_SHD_ndWT.

CODA_SHD_ndWT, som er ein hybrid mellom ndWT-versjonen og Sandve [11] sin SHD-versjon er den mest lovande ut basert på brukarvennlighet, effektivitet og skånsomheit.

Det vart konstruert og testa tre kvalitetsmål for LC-MS data. Eit av måla er basert på forholdet i singulærverdifordelinga i datasettet, medan dei andre to samanliknar topplikskap for toppar frå same forbindelse, ved korrelasjon og norm-likskap.

Singulærverdi-metoden visar seg å vera lite robust, ved bruk på store sett med mange massar av ulik intensitet. Metoden kan i tillegg ha ein svakheit ved at ein singulærverdi kan representera fleire signal.

Dei to toppsamanlikningsmetodane passar seg best til samanlikning av transformasjonsmetodane nytta i arbeidet, og kan ikkje nyttast til å samanlikna CODA. Ut frå toppsamanlikningsmetodane er WPT ein betre transformasjon enn ndWT.

# 9 Vidare arbeid

Testa fleire av metodane på same måte som WPT vart testa i WPTlim,utan faktisk wavelet-dekomponering av data, dvs. WT berre vert nytta til å finna mcq-grensa. Velja wavelets som liknar på støy ( som gjort for CWT), og ikkje berre som liknar på signal.

Eksperimentera med kvalitetsmålet i CODA - f.eks nytta median-filtrering i mcq-fastsetjing.

For WPT : kunne vurdert nytt mål for mcq-fastsetjing (/bestemmelse), og fleire må for beste basis

Testa nye kriterium for singulærverdi-forhold-metoden med f.eks. finna når verdiane glattar ut, i staden for å avhenga av den største verdien eller eit stort hopp i verdiar.

Skaleringsparameter (Amp-parameter) uoptimalt, dette bør optimaliserast. Utgangspunktet var at metoden skulle vera parameterfri – pga. brukervennlighet osv.

s

# 10 Referansar


1. Windig, W., Phalp, J. M., Payne, A. W. , *A Noise and Background Reduction Method for Component Detection in Liquid Chromatography/Mass Spectrometry.* Analytical Chemistry, 1996. **68**(20): p. 3602-3606.
2. Fleming, C.M., Kowalski, B. R., Apffel, A., Hancock, W. S., *Windowed mass selection method: a new data processing algorithm for liquid chromatography–mass spectrometry data.* Journal of Chromatography A, 1999. **849**(1): p. 71-85.
3. van Rijswick, M.H.J., *Adaptive program for high precision off-line processing of chromatograms.* Chromatographia, 1974. **7**(9): p. 491-501.
4. Fredriksson, M., Petersson, P., Jörntén-Karlsson, M., Axelsson, B., Bylund, D., *An objective comparison of pre-processing methods for enhancement of liquid chromatography–mass spectrometry data.* Journal of Chromatography A, 2007. **1172**(2): p. 135-150.
5. Danielsson, R., Bylund, D. , Markides, K. E., *Matched filtering with background suppression for improved quality of base peak chromatograms and mass spectra in liquid chromatography–mass spectrometry.* Analytica Chimica Acta, 2002. **454**(2): p. 167-184.
6. Savitzky, A., Golay, M. J. E., *Smoothing and Differentiation of Data by Simplified Least Squares Procedures.* Analytical Chemistry, 1964. **36**(8): p. 1627-1639.
7. Andreev, V.P., Rejtar, T., Chen, H. S., Moskovets, E. V., Ivanov, A. R., Karger, B. L., *A universal denoising and peak picking algorithm for LC-MS based on matched filtration in the chromatographic time domain.* Analytical Chemistry, 2003. **75**(22): p. 6314-6326.
8. Hastings, C.A., Norton, S. M., Roy, S., *New algorithms for processing and peak detection in liquid chromatography/mass spectrometry data.* Rapid Communications in Mass Spectrometry, 2002. **16**(5): p. 462-467.
9. Cappadona, S., Levander, F., Jansson, M., James, P., Cerutti, S., Pattini, L., *Wavelet-based method for noise characterization and rejection in high-performance liquid chromatography coupled to mass spectrometry.* Analytical Chemistry, 2008. **80**(13): p. 4960-4968.
10. Windig, W., *The use of the Durbin-Watson criterion for noise and background reduction of complex liquid chromatography/mass spectrometry data and a new algorithm to determine sample differences.* Chemometrics and Intelligent Laboratory Systems, 2005. **77**(1-2): p. 206-214.
11. Sandve, I.Ø., *Preprosessering av XC-MS data, Masteroppgave i kjemometri.* Universitetet i Bergen 2011.
12. Li, Y.F., Qu, H. B., Cheng, Y. Y., *An entropy-based method for noise reduction of liquid chromatography-mass spectrometry data.* Analytica Chimica Acta, 2008. **612**(1): p. 19-22.
13. Windig, W., Smith, W. F., Nichols, W. F., *Fast interpretation of complex LC/MS data using chemometrics.* Analytica Chimica Acta, 2001. **446**(1-2): p. 467-476.
14. Gong, F., Liang, Y., Xie, P., Chau, F., *Information theory applied to chromatographic fingerprint of herbal medicine for quality control.* Journal of Chromatography A, 2003. **1002**(1–2): p. 25-40.
15. Poole, C.F., *Chromatography*. 2014, McGraw Hill Education: AccessScience
16. *Getting Started in HPLC - Section 1C. HPLC Instruments*. 2000; downloaded 7.5.13]. Available from: http://www.lcresources.com/resources/getstart/1c01.htm.
17. Miller, J.M., *CHROMATOGRAPHY Concepts and Contrasts* 2ed. 2005, Hoboken, New Jersey , U.S. : John Wiley & Sons, Inc. .
18. Harris, D.C., *Quantitative chemical analysis*. 2010, New York: Freeman. p. 595-628.
19. Dorsey, J.G., *Liquid chromatography*. 2012, McGraw-Hill Education: AccessScience.com.
20. Smith, R.M., *Understanding mass spectra: a basic approach*. 2004, Hoboken, N.J.: Wiley-Interscience. XVIII, 372 s. : ill.

21. Gaskell, S.J., *Electrospray: Principles and practice.* Journal of Mass Spectrometry, 1997. **32**(7): p. 677-688.
22. Harris, D.C., *Quantitative chemical analysis*. 2010, New York: Freeman. p. 616-617.
23. Pavia, D.L., Lampman, G.M. , Kriz, G.S. , Vyvyan, J. R. , *Introduction to Spectrometry*. 4 , International Edition ed. 2009, Washington, US: BROOKS/COLE , CENGAGE Learning.
24. Shannon, C.E., Weaver, W., *The Mathematical Theory of Communication*. 1949: Univ of Illinois Press.
25. Wells, G., Prest, H., Russ, C. W. IV. . *Why use Signal-To-Noise as a Measure of MS Performanse When it is Often Meaningless?* 2011; Available from: http://www.chem.agilent.com/Library/technicaloverviews/Public/5990-8341EN.pdf.
26. Paatero, P., Hopke, P. K., *Discarding or downweighting high-noise variables in factor analytic models.* Analytica Chimica Acta, 2003. **490**(1–2): p. 277-289.
27. Rubin, J.I., Brown, S. G., Wade, K. S., Hafner, H. R., *APPORTIONMENT OF PM2.5 AND AIR TOXICS IN DETROIT, MICHIGAN*. 2006, United States Environmental Protection Agency (EPA): http://www.epa.gov/airtrends/specialstudies/2007_detroit_appx_e.pdf.
28. Nix, A.B.J., Wilson, D. W., *Assay detection limits: concept, definition, and estimation.* European Journal of Clinical Pharmacology, 1990. **39**(3): p. 203-206.
29. Liang, Y.Z., Kvalheim, O. M., Hoskuldsson, A., *DETERMINATION OF A MULTIVARIATE DETECTION LIMIT AND LOCAL CHEMICAL RANK BY DESIGNING A NONPARAMETRIC TEST FROM THE ZERO-COMPONENT REGIONS.* Journal of Chemometrics, 1993. **7**(4): p. 277-289.
30. Lay, D.C., *Linear algebra and its applications*. Vol. 3. 2006, Boston: Pearson education. mi oversetjing.
31. Marcus, M., *Matrix theory*. McGraw-Hill Education: AccessScience.
32. Lorber, A., *Quantifying chemical composition from two-dimensional data arrays.* Analytica Chimica Acta, 1984. **164**(0): p. 293-297.
33. Strang, G., *Wavelets.* American Scientist, 1994. **82**(3): p. 250-255.
34. Strang, G., Truong, N., *Wavelets and Filter Banks*. 1996, USA: Wellesley-Cambridge Press. xix-3.
35. Rioul, O., Duhamel, P., *Fast algorithms for discrete and continuous wavelet transforms.* Information Theory, IEEE Transactions on, 1992. **38**(2): p. 569-586.
36. Chui, C.K., *Wavelet Analysis And Its Applications*. United Kingdom Edition ed. Vol. 1. 1992, London: Academic Press, Inc /Academic Press Limited. 3.
37. Walczak, B., Massart, D. L., *Noise suppression and signal compression using the wavelet packet transform.* Chemometrics and Intelligent Laboratory Systems, 1997. **36**(2): p. 81-94.
38. *Nokre waveletfunksjonar*, in *Matlab . Wavelet toolbox main menu -> display : wavelet display . 12 refinements*.
39. *MATLAB - Matrix Laboratory* 2014, Mathworks Inc. .
40. Mittermayr, C.R., Nikolov, S. G., Hutter, H., Grasserbauer, M., *Wavelet denoising of Gaussian peaks: A comparative study.* Chemometrics and Intelligent Laboratory Systems, 1996. **34**(2): p. 187-202.
41. Tveit, K., *Deteksjon og estimering av heteroscedastisk støy vha. wavelettransformasjon , Hovedfagsoppgave i kjemometri.* Universitetet i Bergen, 1998.
42. Daubechies, I., *Wavelets*. 2014, AT&T Bell Laboratories: accessscience.com.
43. Shensa, M.J., *THE DISCRETE WAVELET TRANSFORM - WEDDING THE A TROUS AND MALLAT ALGORITHMS.* Ieee Transactions on Signal Processing, 1992. **40**(10): p. 2464-2482.
44. Wegner, F.V., Both, M., Fink, R. H. A., *Automated detection of elementary calcium release events using the a trous wavelet transform.* Biophysical Journal, 2006. **90**(6): p. 2151-2163.
45. Combes, J.M., Grossman, A. , Tchamitchian, Ph. , *Wavelets: Time-Frequency Methods and Phase Space* 1989, Berlin: Springer, IPTI
46. Mallat, S.G., *A Theory for Multiresolution Signal Decomposition: The Wavelet Representation.* IEEE Trans. Pattern Anal. Mach. Intell., 1989. **11**(7): p. 674-693.

47. Andersson, F.O., Kaiser, R., Jacobsson, S. P., *Data preprocessing by wavelets and genetic algorithms for enhanced multivariate analysis of LC peptide mapping.* Journal of Pharmaceutical and Biomedical Analysis, 2004. **34**(3): p. 531-541.
48. Sanchez-Ponce, R., Guengerich, F. P., *Untargeted Analysis of Mass Spectrometry Data for Elucidation of Metabolites and Function of Enzymes.* Analytical Chemistry, 2007. **79**(9): p. 3355-3362.
49. Tautenhahn, R., Bottcher, C., Neumann, S., *Highly sensitive feature detection for high resolution LC/MS.* BMC Bioinformatics, 2008. **9**(1): p. 504.
50. Cappadona, S., Nanni, P., Benevento, M., Levander, F., Versura, P., Roda, A., Cerutti, S., Pattini, L., *Improved Label-Free LC-MS Analysis by Wavelet-Based Noise Rejection.* Journal of Biomedicine and Biotechnology, 2010.
51. Chen, H.-P., Liao, H.-J., Huang, C.-M., Wang, S.-C., Yu, S.-N., *Improving liquid chromatography–tandem mass spectrometry determinations by modifying noise frequency spectrum between two consecutive wavelet-based low-pass filtering procedures.* Journal of Chromatography A, 2010. **1217**(17): p. 2804-2811.
52. Zhang, W., Chang, J., Lei, Z., Huhman, D., Sumner, L. W., Zhao, P. X., *MET-COFEA: A Liquid Chromatography/Mass Spectrometry Data Processing Platform for Metabolite Compound Feature Extraction and Annotation.* Analytical Chemistry, 2014. **86**(13): p. 6245-6253.
53. Bari, M.G., Ma, X., Zhang, J., *PeakLink: a new peptide peak linking method in LC-MS/MS using wavelet and SVM.* Bioinformatics, 2014. **30**(17): p. 2464-2470.
54. Zeng, Y.X., Araujo, P., Du, Z. Y., Nguyen, T. T., Froyland, L., Grung, B., *Elucidation of triacylglycerols in cod liver oil by liquid chromatography electrospray tandem ion-trap mass spectrometry.* Talanta, 2010. **82**(4): p. 1261-1270.
55. Skaar, I., Adaku, C., Jordheim, M., Byamukama, R., Kiremire, B., Andersen, Ø. M., *Purple anthocyanin colouration on lower (abaxial) leaf surface of Hemigraphis colorata (Acanthaceae).* Phytochemistry, 2014. **105**(0): p. 141-146.
56. L., D.D., *Progress in Wavelet Analysis and Applications* ed. S.R. Y.Mayer 1993, France: Editions Frontiers ;.

# 11 Vedlegg

## 11.1 Vedleggsoversikt

Liste over vedlegg:

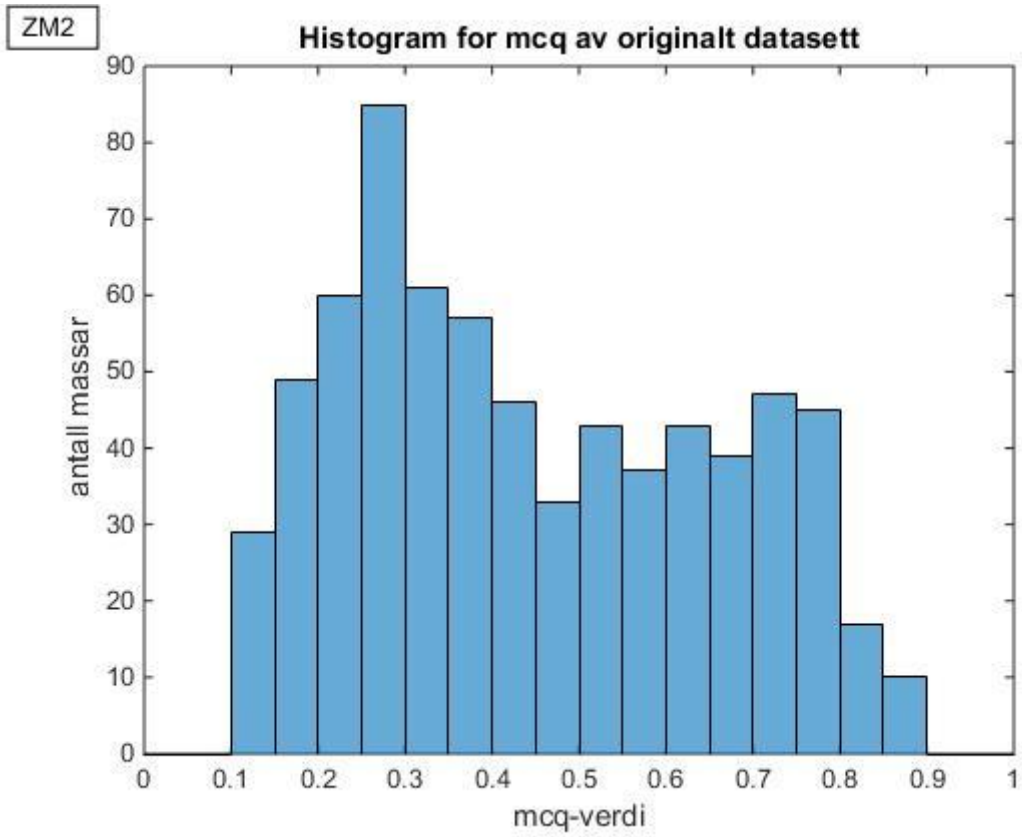
Vedlegg 1: Mcq-fordeling for ZM2-settet ..... 129
Vedlegg 2: Mcq-fordeling for a1 og d1 av ZM2-settet ..... 130
Vedlegg 3: 30 største singulærverdiar og log av singulærverdiar for HCO-settet ..... 131
Vedlegg 4: 18 største singulærverdiar og log av singulærverdiar for a1 og d1 av HCO-settet ..... 132
Vedlegg 5: Massekromatogram for sim3-datasettet; (øvst) originalt datasett, (midtarst) behaldt ved CODA_ndWT og (nedst) fjerna av CODA i metoden ..... 133
Vedlegg 6: TMS for sim3-datasettet; (øvst) originalt datasett, (midtarst) behaldt ved CODA_ndWT og (nedst) originalt datasett utan støy ..... 134
Vedlegg 7: TIC for kvalitetsområdet av HCO-settet; (øvst) originalt datasett og (nedst) behaldt av CODA_ndWT ..... 135
Vedlegg 8:EIC av TAG2-settet; (øvst) originalt datasett, (midtarst) CODA_ndWT-prosessert datasettet og (nedst) det som vert fjerna av metoden. ..... 136
Vedlegg 9: EIC av TAG3-settet; (øvst) originalt datasett, (midtarst) CODA_ndWT-prosessert datasettet og (nedst) det som vert fjerna av metoden. ..... 137
Vedlegg 10: EIC av TAG4-settet; (øvst) originalt datasett, (midtarst) CODA_ndWT-prosessert datasettet og (nedst) det som vert fjerna av metoden. ..... 138
Vedlegg 11: Alle EIC for nullkomponentsområdet i sim3 ved tida 350 til 400; (øvst) originalt datasett og (nedst) behaldt av CODA_CWT ..... 139
Vedlegg 12: TIC for kvalitetsområdet av HCO-settet; (øvst) originalt datasett, (midtarst) behaldt av CODA_CWT og (nedst) fjerna av metoden ..... 140
Vedlegg 13: TIC av TAG2-settet; (øvst) originalt datasett, (midtarst) CODA_CWT-prosessert datasettet og (nedst) signal fjerna av CODA-delen i metoden. ..... 141
Vedlegg 14: TIC av TAG3-settet; (øvst) originalt datasett, (midtarst) CODA_CWT-prosessert datasettet og (nedst) signal fjerna av CODA-delen i metoden. ..... 142
*Vedlegg 15: TIC av TAG4-settet; (øvst) originalt datasett, (midtarst) CODA_CWT-prosessert datasettet og (nedst) signal fjerna av CODA-delen i metoden.* ..... 142

Vedlegg 16: *TMS av TAG2-settet; (øvst) originalt datasett, (midtarst) CODA_CWT-prosessert datasettet og (nedst) signal fjerna av CODA-delen i metoden.* ..... 143
Vedlegg 17: TMS av TAG3-settet; (øvst) originalt datasett, (midtarst) CODA_CWT-prosessert datasettet og (nedst) signal fjerna av CODA-delen i metoden. ..... 143
Vedlegg 18: TMS av TAG4-settet; (øvst) originalt datasett, (midtarst) CODA_CWT-prosessert datasettet og (nedst) signal fjerna av CODA-delen i metoden. ..... 144
Vedlegg 19: TIC av sim3-settet; (øvst) originalt datasett, (midtarst) CODA_WPT-prosessert datasettet og (nedst) originalt datasett utan støy. ..... 146
Vedlegg 20: *Kvalitetsområdet av HCO-settet; (øvst) TIC av originalt datasett, (midtarst) TIC av CODA_WPT-prosessert datasettet og (nedst) CODA_WPT-prosessert datasettet.* ..... 147
Vedlegg 21: TIC av *TAG1-settet; (øvst) originalt datasett, (midtarst) CODA_WPT-prosessert datasettet og (nedst) det som vert fjerna av CODA-delen av metoden.* ..... 148
Vedlegg 22: TIC av TAG2-settet; (øvst) originalt datasett, (midtarst) CODA_WPT-prosessert datasettet og (nedst) det som vert fjerna av CODA-delen av metoden. ..... 148
Vedlegg 23: EIC av TAG3-settet; (øvst) originalt datasett, (midtarst) CODA_WPT-prosessert datasettet og (nedst) det som vert fjerna av CODA-delen av metoden. ..... 149
Vedlegg 24: EIC av TAG4-settet; (øvst) originalt datasett, (midtarst) CODA_WPT-prosessert datasettet og (nedst) det som vert fjerna av CODA-delen av metoden. ..... 149
Vedlegg 25: EIC av TAG3-settet; (øvst) originalt datasett, (midtarst) CODA_WPT2-prosessert datasettet og (nedst) det som vert fjerna av CODA-delen av metoden. ..... 150
Vedlegg 26: EIC av TAG4-settet; (øvst) originalt datasett, (midtarst) CODA_WPT2-prosessert datasettet og (nedst) det som vert fjerna av CODA-delen av metoden. ..... 150
Vedlegg 27: EIC av ZM2-settet; (øvst) originalt datasett, (midtarst) CODA_WPT2-prosessert datasettet og (nedst) det som vert fjerna av CODA-delen av metoden. ..... 151
Vedlegg 28: TMS av sim3-settet; (øvst) originalt datasett, (midtarst) CODA_WPTlim-prosessert datasettet og (nedst) originalt datasett utan støy. ..... 152
Vedlegg 29: TIC av sim3-settet; (øvst) originalt datasett og (nedst) CODA_WPTlim-prosessert datasettet. ..... 152
Vedlegg 30: EIC av kvalitetsområdet til HCO-settet; (øvst) originalt datasett, (midtarst) CODA_WPTlim-prosessert datasettet og (nedst) det som vert fjerna av metoden. ..... 153
Vedlegg 31: TIC av kvalitetsområdet til HCO-settet; (øvst) originalt datasett, (midtarst) CODA_WPTlim-prosessert datasettet og (nedst) det som vert fjerna av metoden. ..... 154

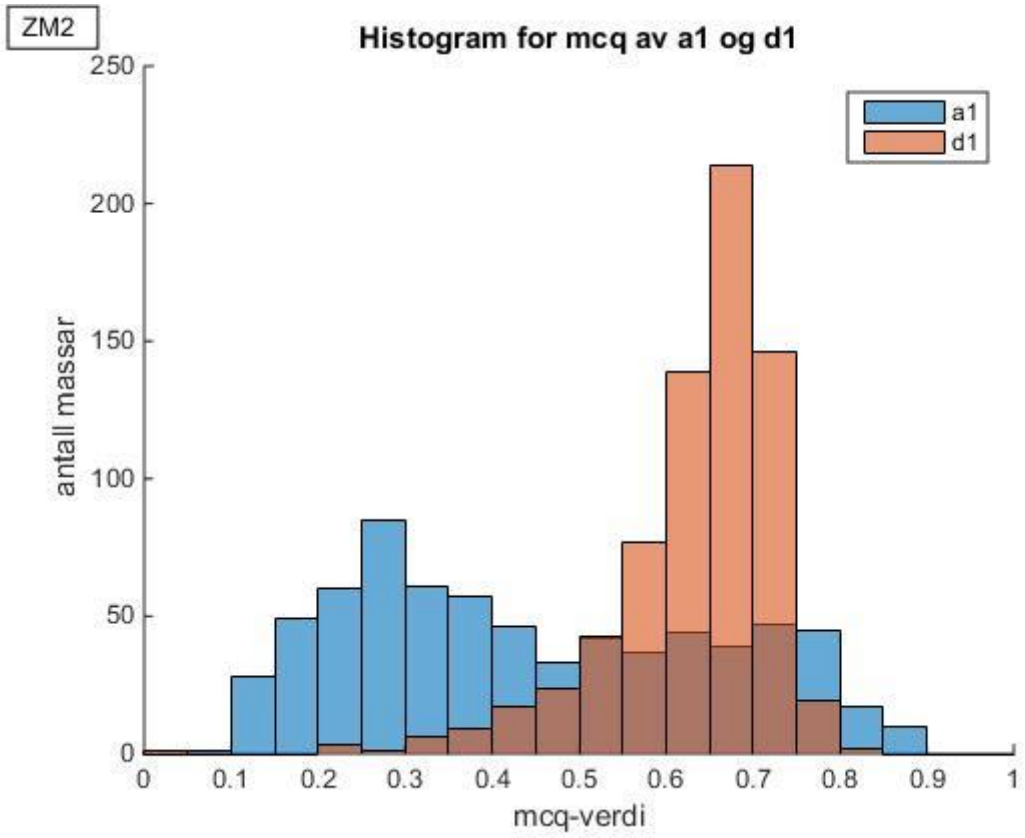
Vedlegg 32: TIC av TAG1-settet; (øvst) originalt datasett, (midtarst) CODA_WPTlim-prosessert datasettet og (nedst) det som vert fjerna av metoden. ..... 155
Vedlegg 33: EIC av TAG1-settet; (øvst) originalt datasett, (midtarst) CODA_WPTlim-prosessert datasettet og (nedst) det som vert fjerna av metoden. ..... 156
Vedlegg 34: TIC av TAG2-settet; (øvst) originalt datasett, (midtarst) CODA_WPTlim-prosessert datasettet og (nedst) det som vert fjerna av metoden. ..... 157
Vedlegg 35: EIC av TAG2-settet; (øvst) originalt datasett, (midtarst) CODA_WPTlim-prosessert datasettet og (nedst) det som vert fjerna av metoden. ..... 158
Vedlegg 36: TIC av TAG3-settet; (øvst) originalt datasett, (midtarst) CODA_WPTlim-prosessert datasettet og (nedst) det som vert fjerna av metoden. ..... 159
Vedlegg 37: EIC av TAG3-settet; (øvst) originalt datasett, (midtarst) CODA_WPTlim-prosessert datasettet og (nedst) det som vert fjerna av metoden. ..... 159
Vedlegg 38: TIC av sim3-settet; (øvst) originalt datasett, (midtarst) CODA_SHD_ndWT-prosessert datasettet og (nedst) det som vert fjerna av metoden. ..... 160
Vedlegg 39: TMS av sim3-settet; (øvst) originalt datasett, (midtarst) CODA_SHD_ndWT-prosessert datasettet og (nedst) det som vert fjerna av metoden. ..... 161
Vedlegg 40: TIC av HCO-settet; (øvst) originalt datasett og (nedst) CODA_SHD_ndWT-prosessert datasettet. ..... 162
Vedlegg 41: HCO-settet; signal som vert fjerna av CODA-delen av CODA_SHD_ndWT ..... 162
Vedlegg 42: Konvertering av ZM2-fil til matlab-formatet «.mat» ..... 163
Vedlegg 43: Konvertering av HCO-fil til matlab-formatet «.mat» ..... 165
Vedlegg 44: Område K ved CODA for MCQ = 0.40 . Singulærverdi-forholdet F er gjeve for (øvst) originalt sett , (midtarst) behaldt av CODA og (nedst) fjerna av CODA ..... 166
Vedlegg 45: Singulærverdiar for ZM2-settet ..... 167
Vedlegg 46: TIC for HCO; (øvst) originalsett, (midtarst) prosessert sett og (nedst) fjerna av CODA. ..... 168
Vedlegg 47: TIC av kvalitetsområdet for HCO; (øvst) originalsett, (midtarst) prosessert sett og (nedst) fjerna av CODA. ..... 169
Vedlegg 48: EIC av kvalitetsområdet for HCO; (øvst) originalsett, (midtarst) prosessert sett og (nedst) fjerna av CODA. ..... 170
Vedlegg 49:TIC av TAG2-settet; (øvst) originalt datasett og (nedst) datasett behaldt av CODA med mcq = 0.85 ..... 171

Vedlegg 50: TIC av TAG3-settet; (øvst) originalt datasett og (nedst) datasett behaldt av CODA med mcq = 0.85 ..... 172
Vedlegg 51: TIC av TAG4-settet; (øvst) originalt datasett og (nedst) datasett behaldt av CODA med mcq = 0.85 ..... 173
Vedlegg 52: Utvald område nr. 2 for toppsamanlikning av sett3, før og etter CODA_ndWT 174
Vedlegg 53: HCO kvalitetstopp 2 ved CODA_CWT ..... 175
Vedlegg 54: sim3 kvalitetstopp 2 ved CODA_CWT ..... 176
Vedlegg 55: sim3 kvalitetstopp 3 ved CODA_CWT ..... 177
Vedlegg 56: HCO – kvalitetstopp 3 ved CODA_WPT ..... 178
Vedlegg 57: HCO – kvalitetstopp 2 ved CODA_WPT ..... 179
Vedlegg 58: HCO – kvalitetstopp 2 ved CODA_ndWT ..... 180
Vedlegg 59: HCO – kvalitetstopp 3 ved CODA_ndWT ..... 181

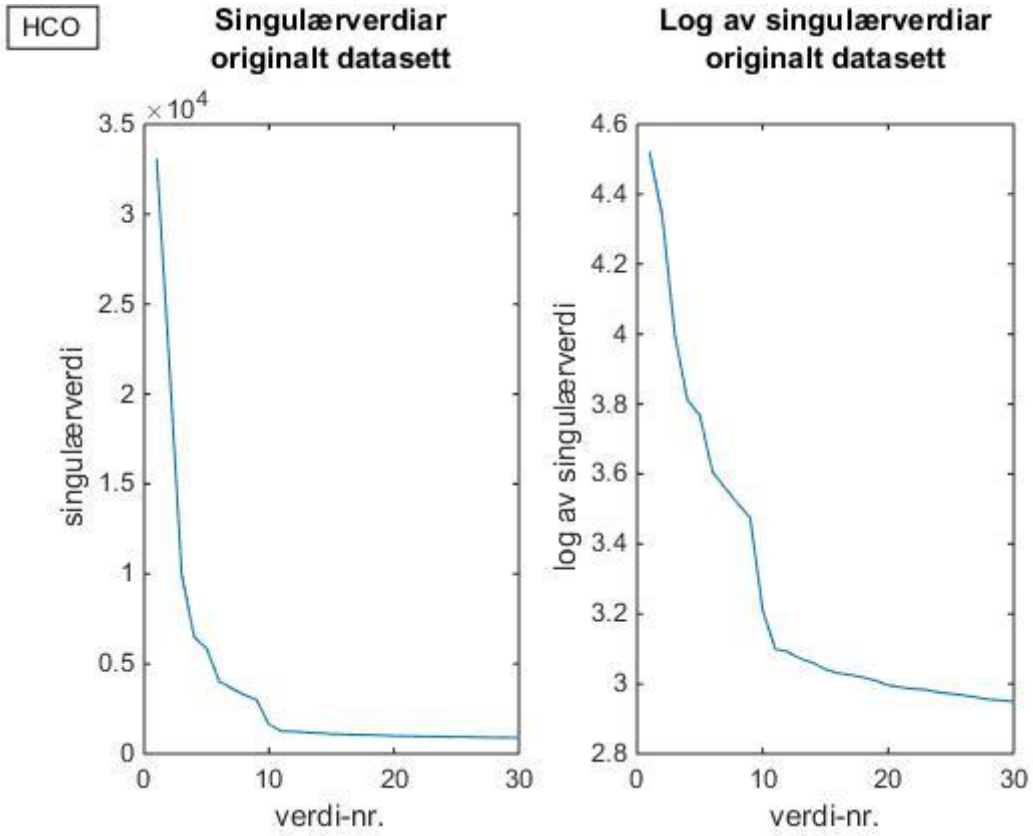
*Vedlegg 1: Mcq-fordeling for ZM2-settet*



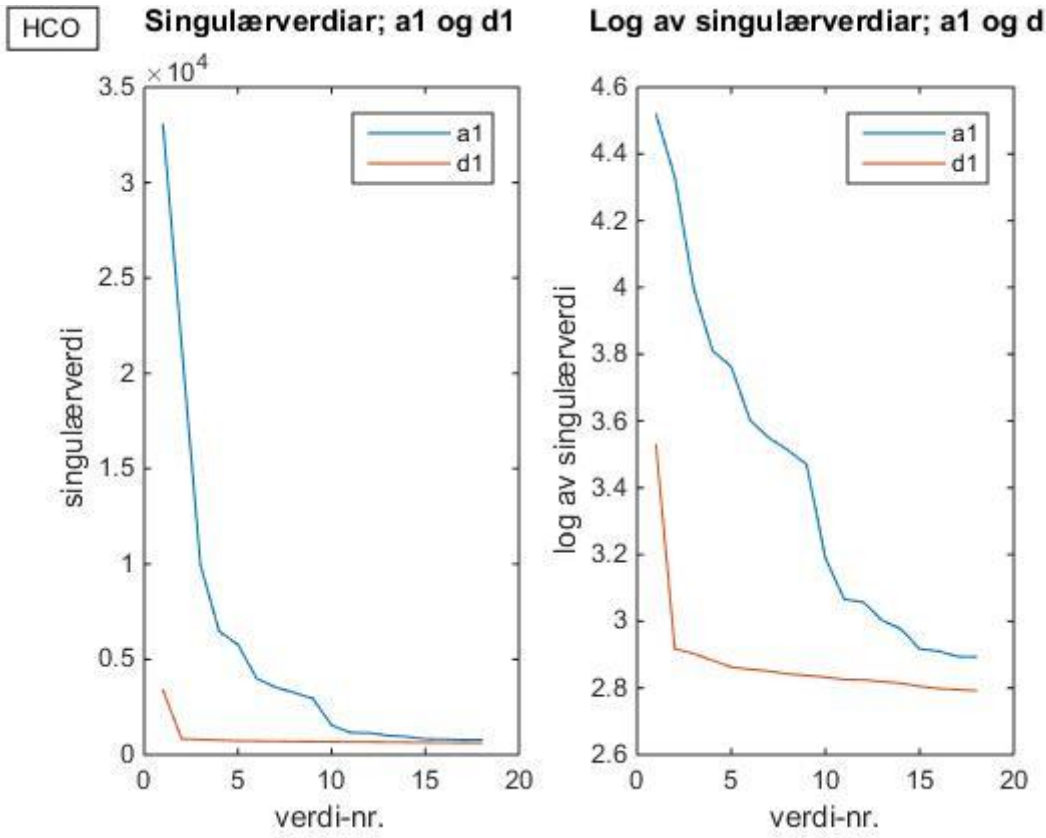
*Vedlegg 2: Mcq-fordeling for a1 og d1 av ZM2-settet*



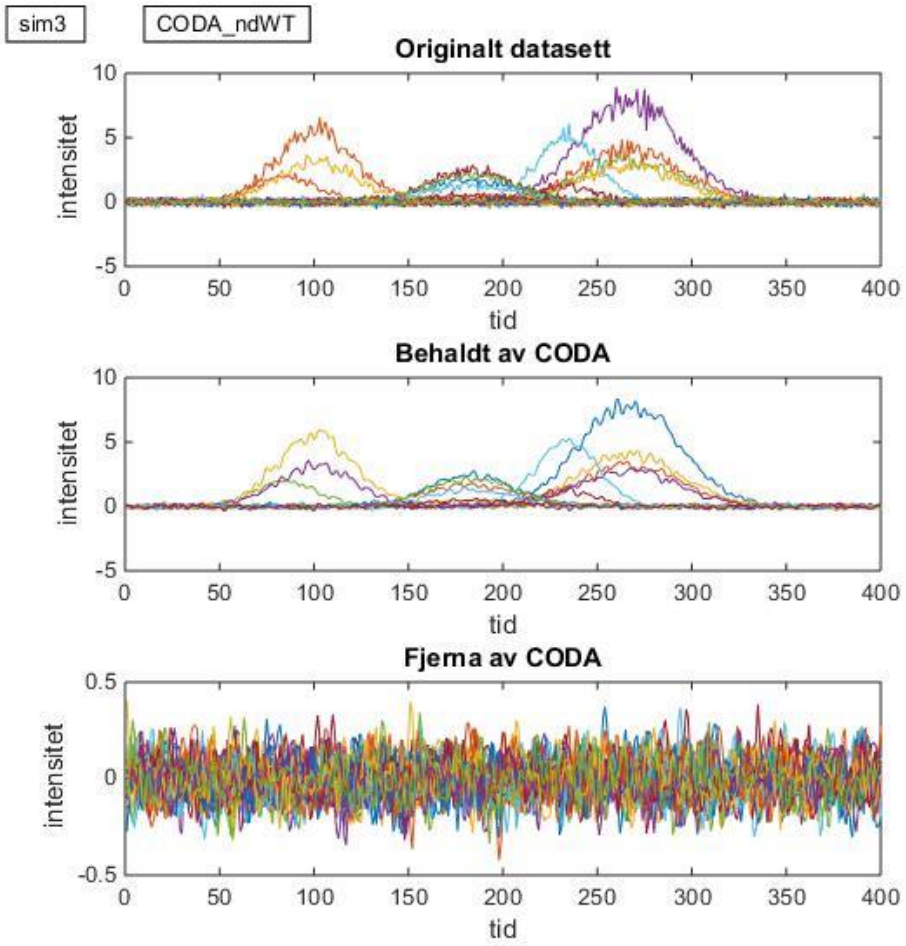
*Vedlegg 3: 30 største singulærverdiar og log av singulærverdiar for HCO-settet*



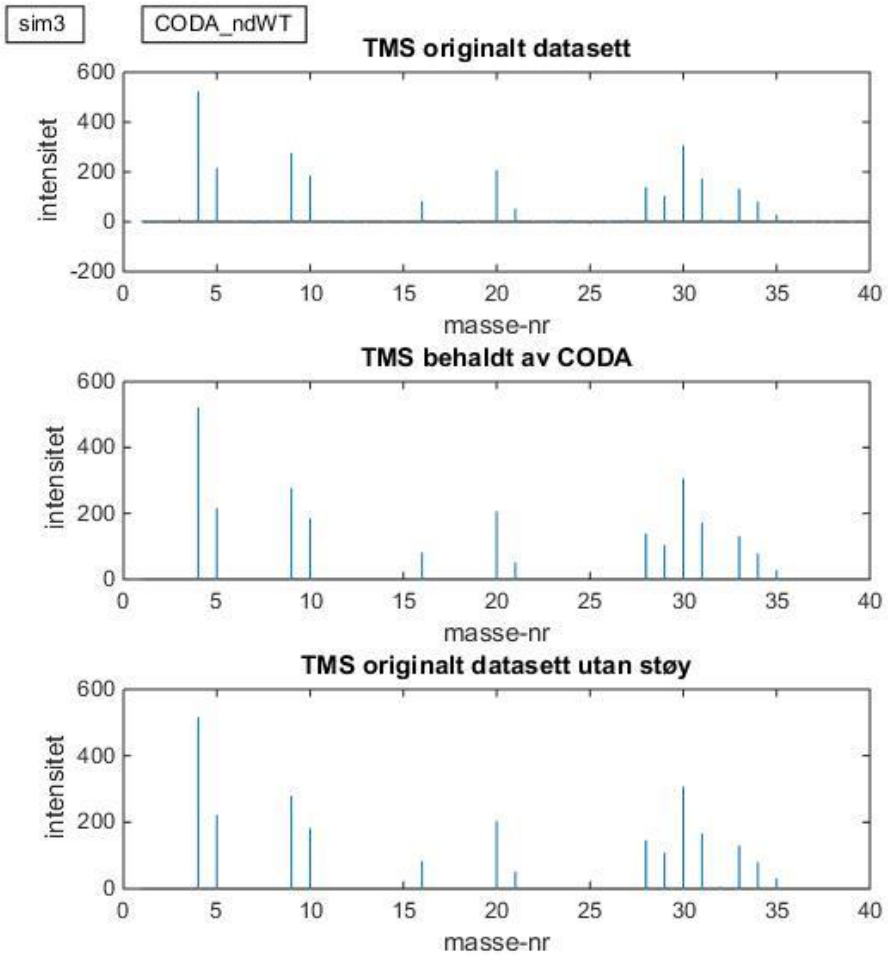
*Vedlegg 4: 18 største singulærverdiar og log av singulærverdiar for a1 og d1 av HCO-settet*



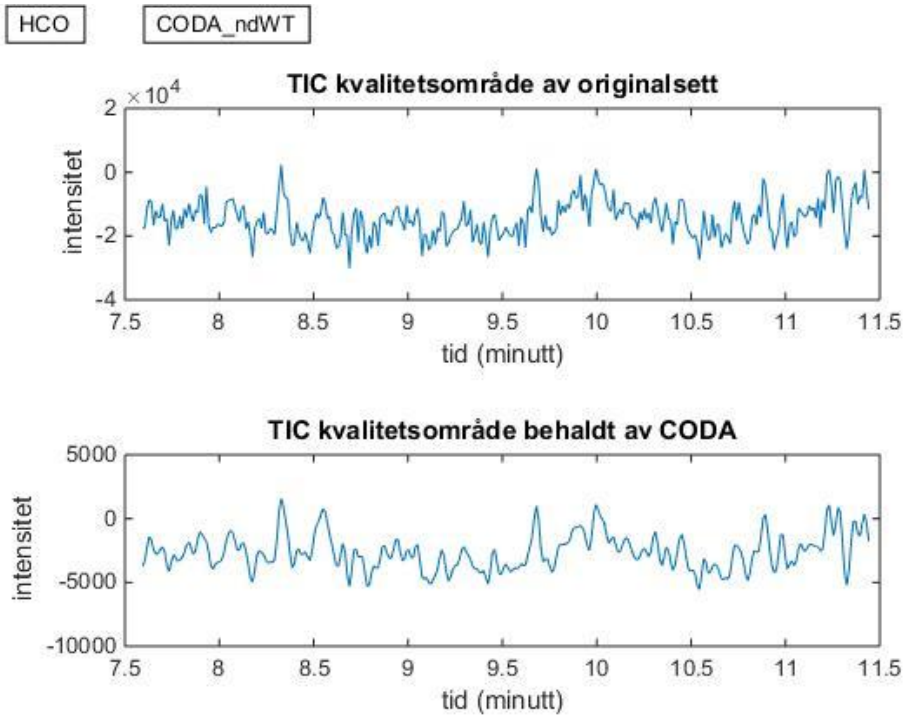
*Vedlegg 5: Massekromatogram for sim3-datasettet; (øvst) originalt datasett, (midtarst) behaldt ved CODA_ndWT og (nedst) fjerna av CODA i metoden*



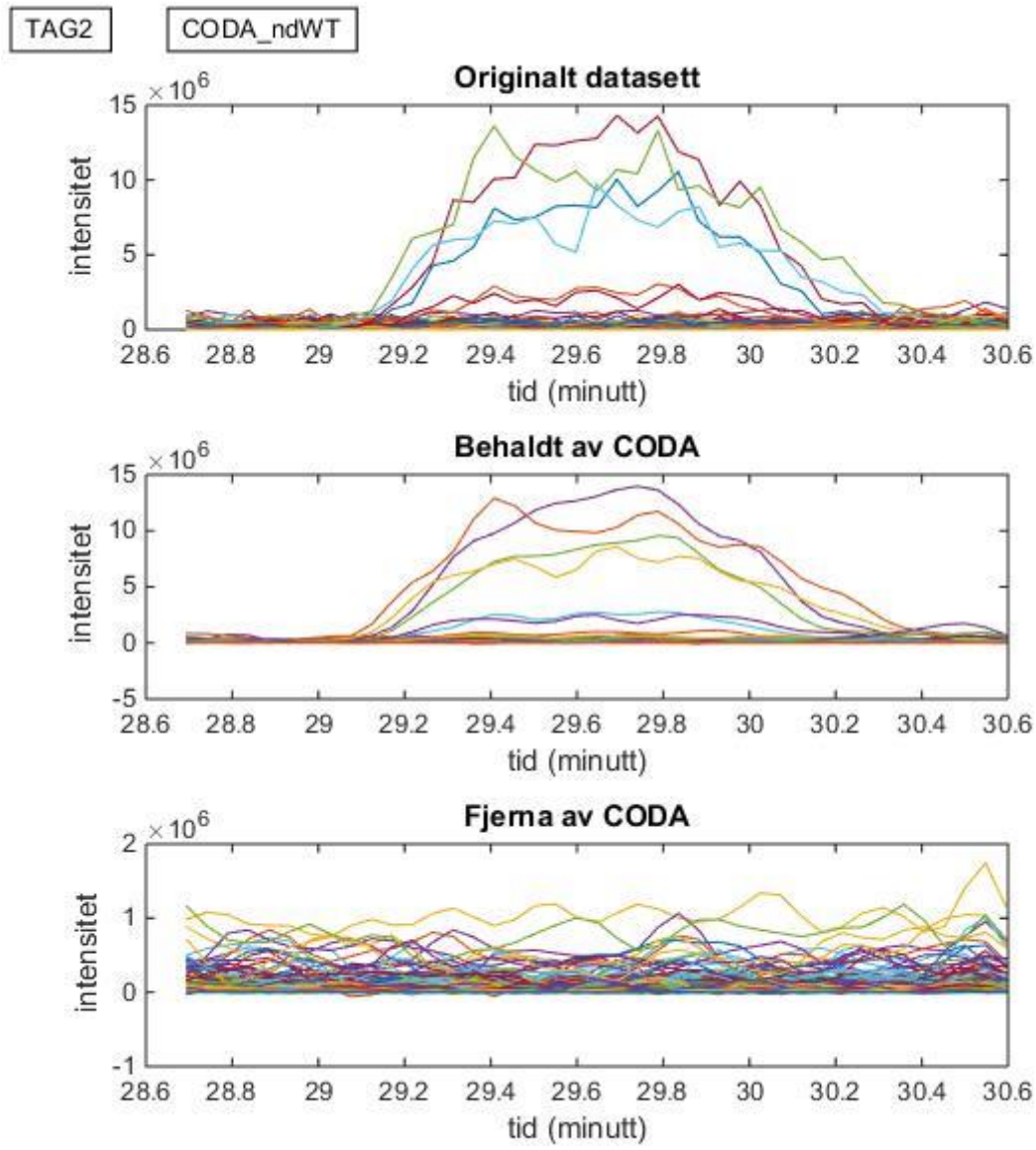
*Vedlegg 6: TMS for sim3-datasettet; (øvst) originalt datasett, (midtarst) behaldt ved CODA_ndWT og (nedst) originalt datasett utan støy*



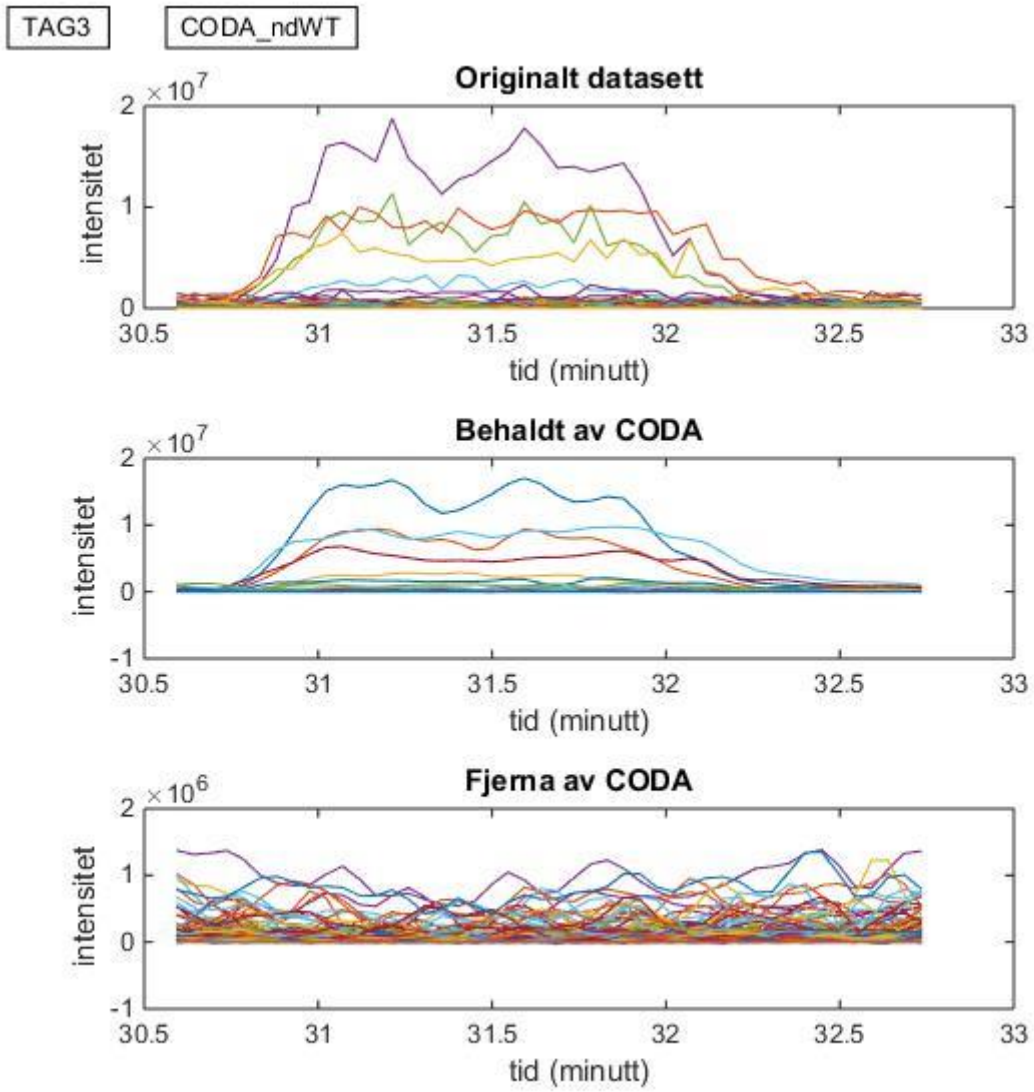
*Vedlegg 7: TIC for kvalitetsområdet av HCO-settet; (øvst) originalt datasett og (nedst) behaldt av CODA_ndWT*



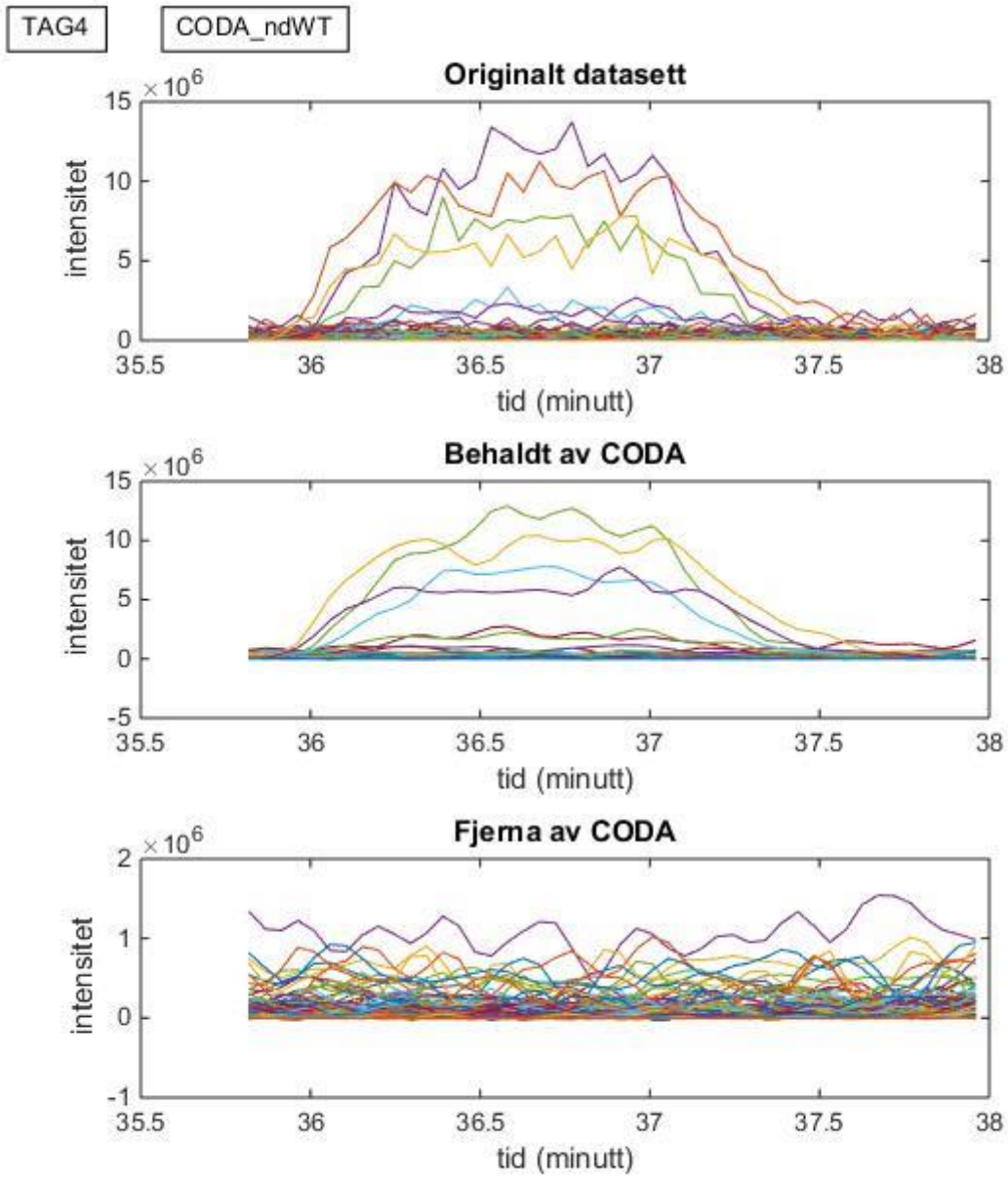
*Vedlegg 8:EIC av TAG2-settet; (øvst) originalt datasett, (midtarst) CODA_ndWT-prosessert datasettet og (nedst) det som vert fjerna av metoden.*



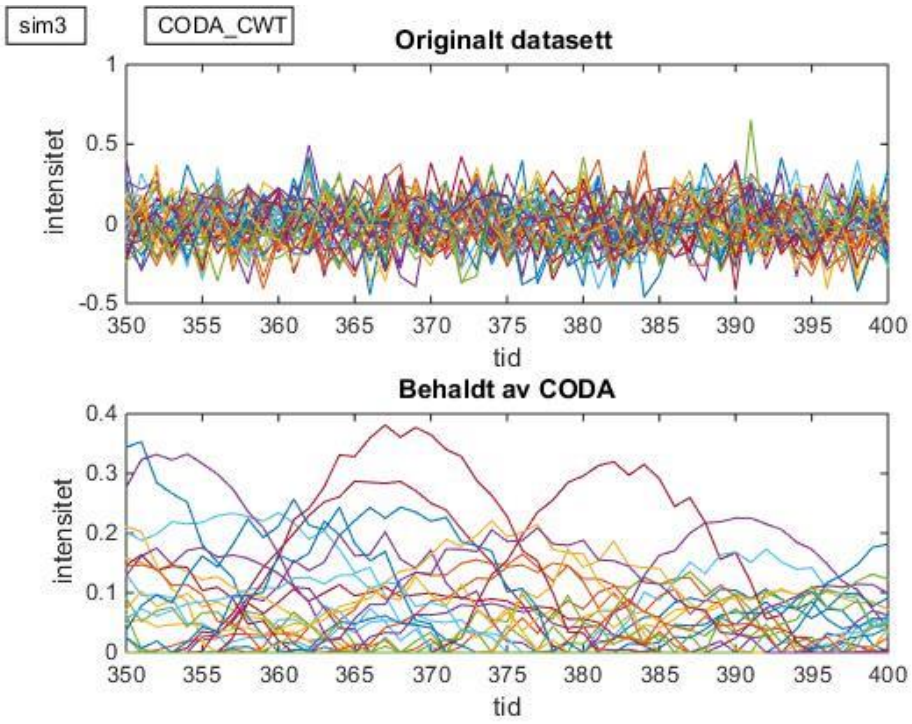
*Vedlegg 9: EIC av TAG3-settet; (øvst) originalt datasett, (midtarst) CODA_ndWT-prosessert datasettet og (nedst) det som vert fjerna av metoden.*



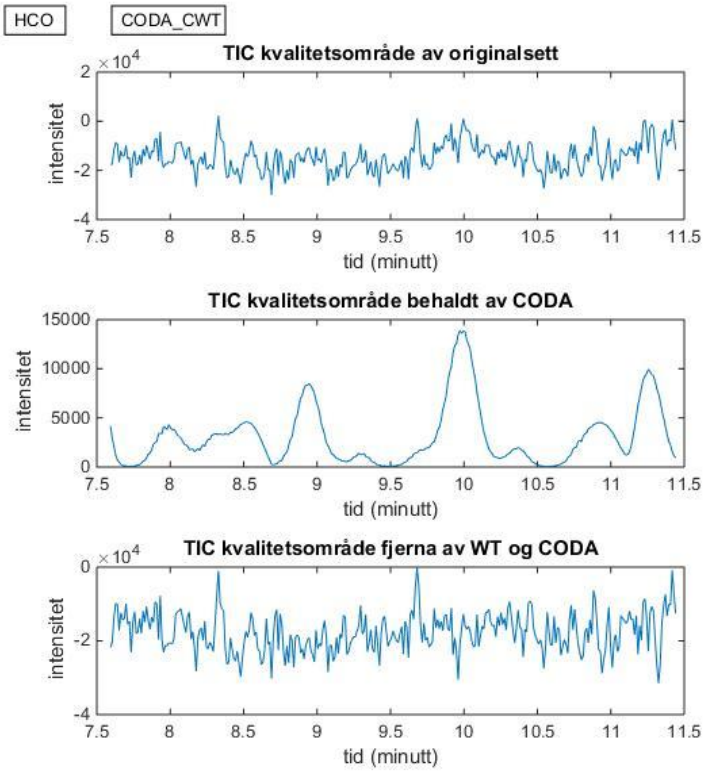
*Vedlegg 10: EIC av TAG4-settet; (øvst) originalt datasett, (midtarst) CODA_ndWT-prosessert datasettet og (nedst) det som vert fjerna av metoden.*



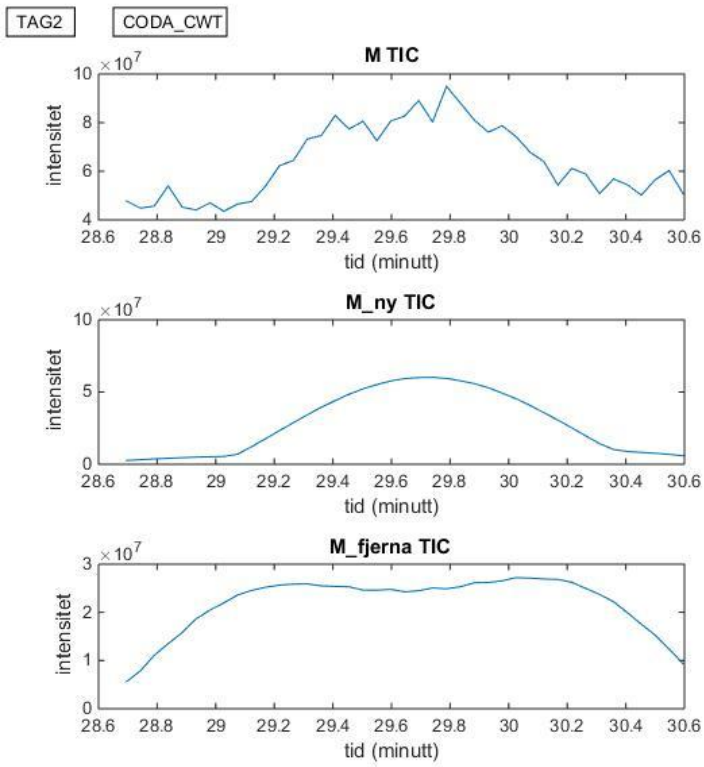
*Vedlegg 11: Alle EIC for nullkomponentsområdet i sim3 ved tida 350 til 400; (øvst) originalt datasett og (nedst) behaldt av CODA_CWT*



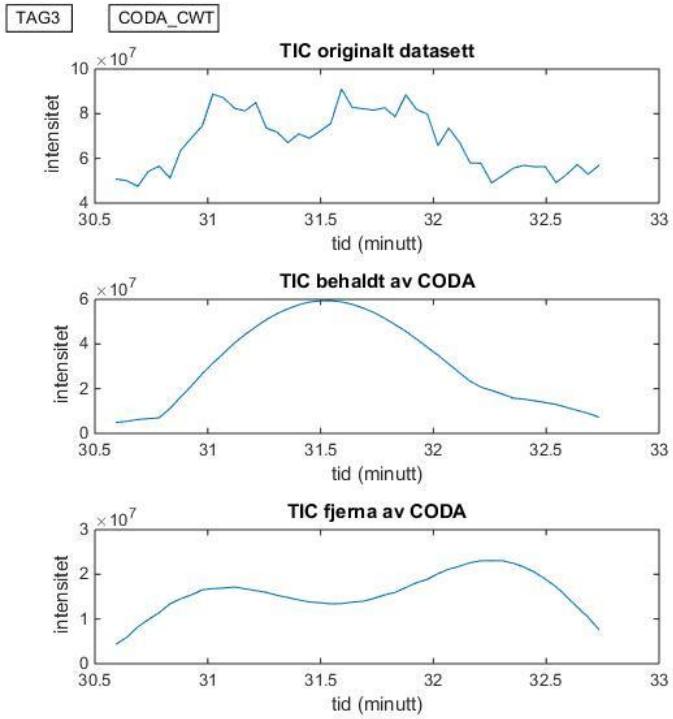
*Vedlegg 12: TIC for kvalitetsområdet av HCO-settet; (øvst) originalt datasett, (midtarst) behaldt av CODA_CWT og (nedst) fjerna av metoden.*



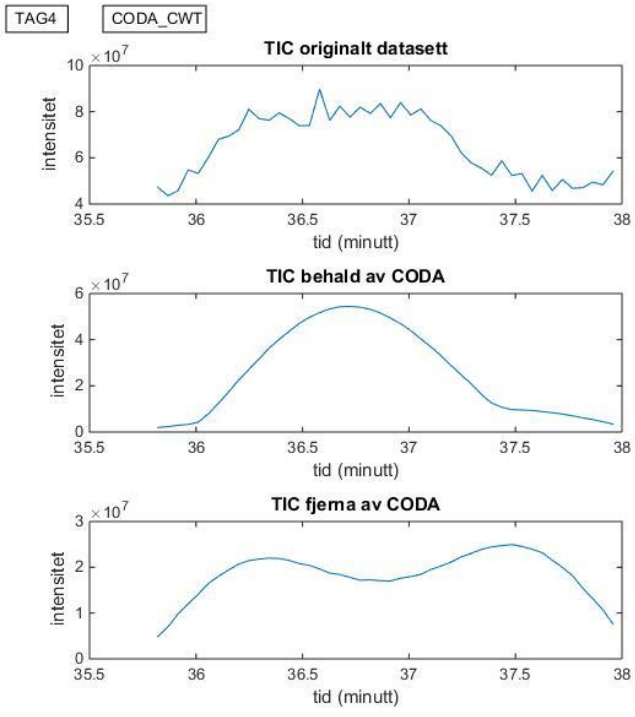
*Vedlegg 13: TIC av TAG2-settet; (øvst) originalt datasett, (midtarst) CODA_CWT-prosessert datasettet og (nedst) signal fjerna av CODA-delen i metoden.*



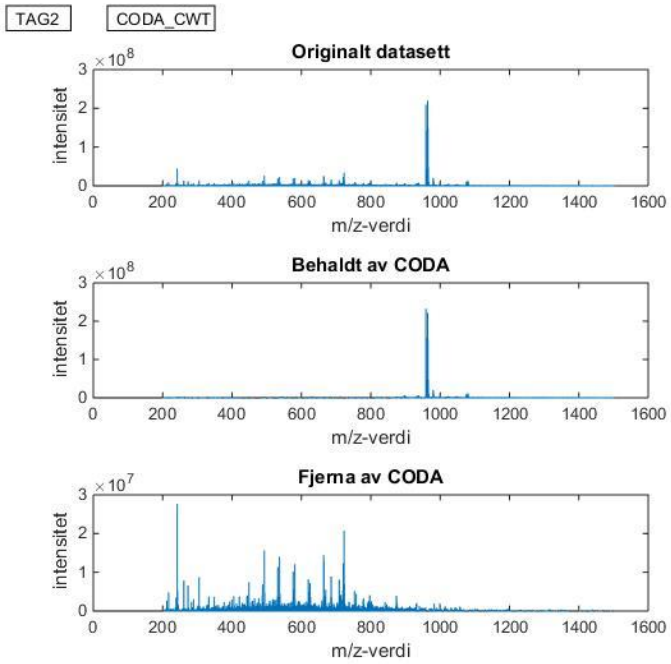
*Vedlegg 14: TIC av TAG3-settet; (øvst) originalt datasett, (midtarst) CODA_CWT-prosessert datasettet og (nedst) signal fjerna av CODA-delen i metoden.*



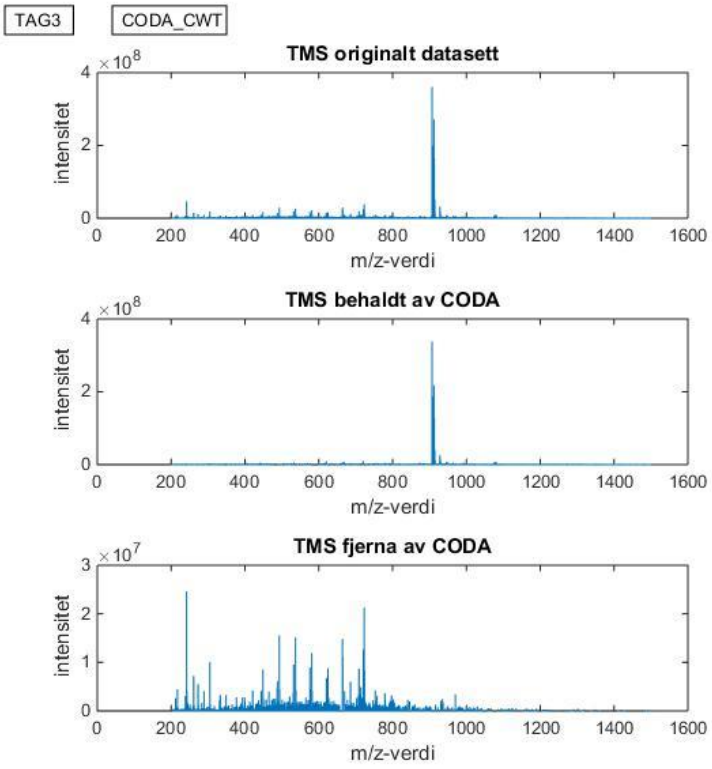

*Vedlegg 15: TIC av TAG4-settet; (øvst) originalt datasett, (midtarst) CODA_CWT-prosessert datasettet og (nedst) signal fjerna av CODA-delen i metoden.*



Vedlegg 16: *TMS av TAG2-settet; (øvst) originalt datasett, (midtarst) CODA_CWT-prosessert datasettet og (nedst) signal fjerna av CODA-delen i metoden.*

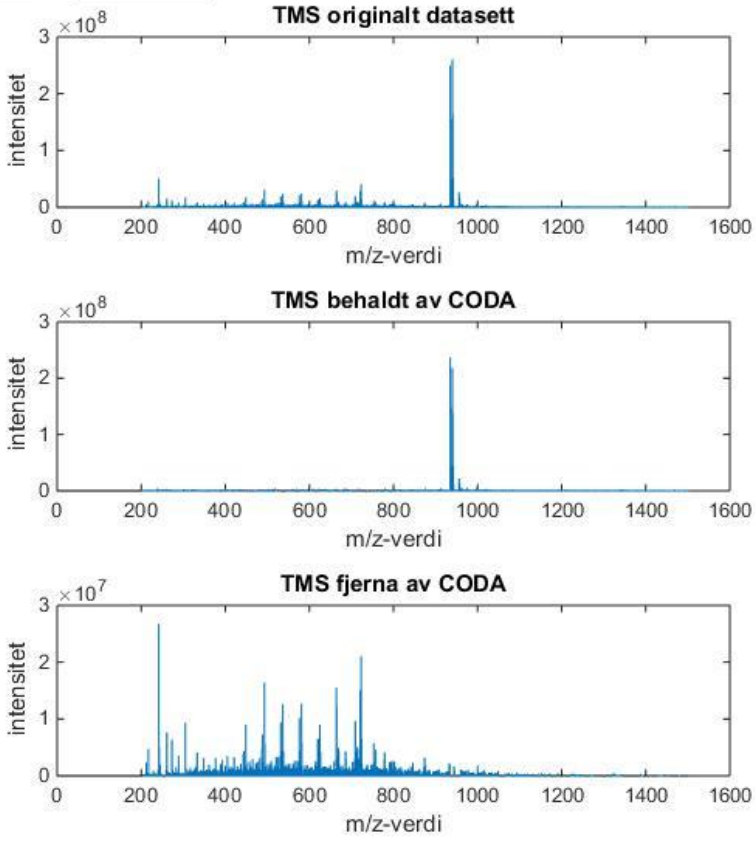



*Vedlegg 17:* TMS av TAG3-settet; (øvst) originalt datasett, (midtarst) CODA_CWT-prosessert datasettet og (nedst) signal fjerna av CODA-delen i metoden.

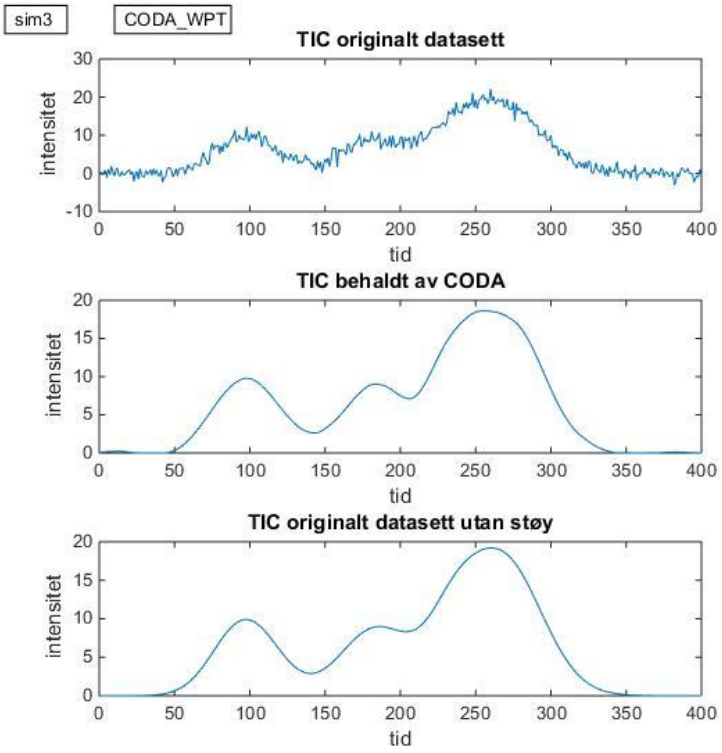



*Vedlegg 18:* TMS av TAG4-settet; (øvst) originalt datasett, (midtarst) CODA_CWT-prosessert datasettet og (nedst) signal fjerna av CODA-delen i metoden.

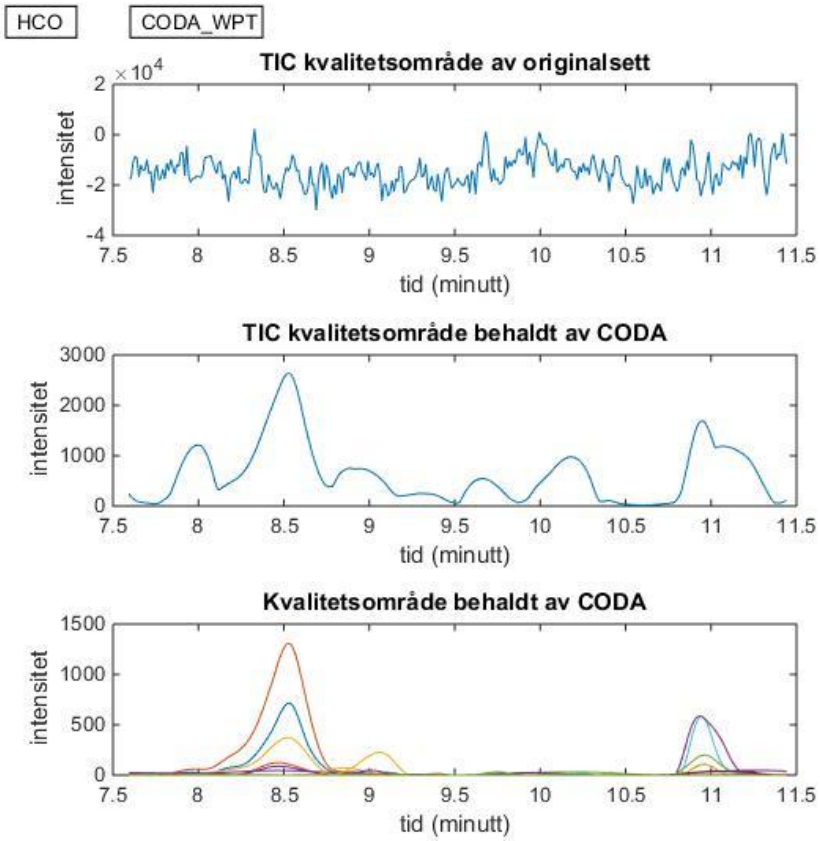
TAG4 CODA_CWT



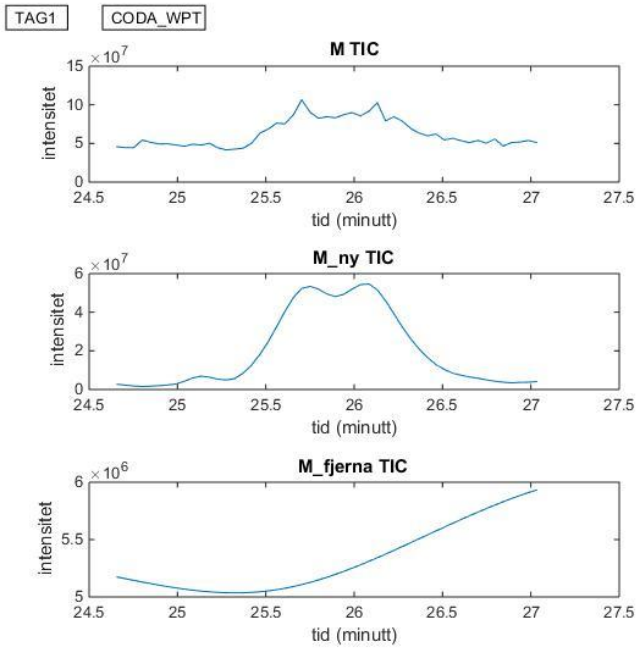
*Vedlegg 19:* TIC av sim3-settet; (øvst) originalt datasett, (midtarst) CODA_WPT-prosessert datasettet og (nedst) originalt datasett utan støy.



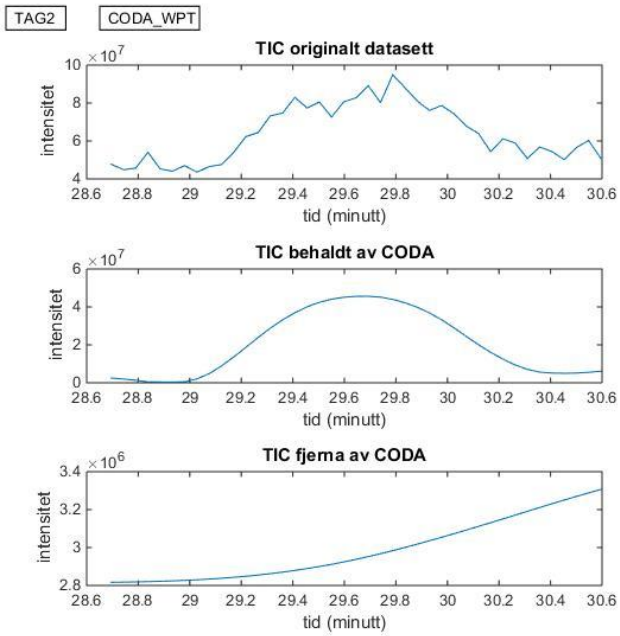
Vedlegg 20: *Kvalitetsområdet av HCO-settet; (øvst) TIC av originalt datasett, (midtarst) TIC av CODA_WPT-prosessert datasettet og (nedst) CODA_WPT-prosessert datasettet.*



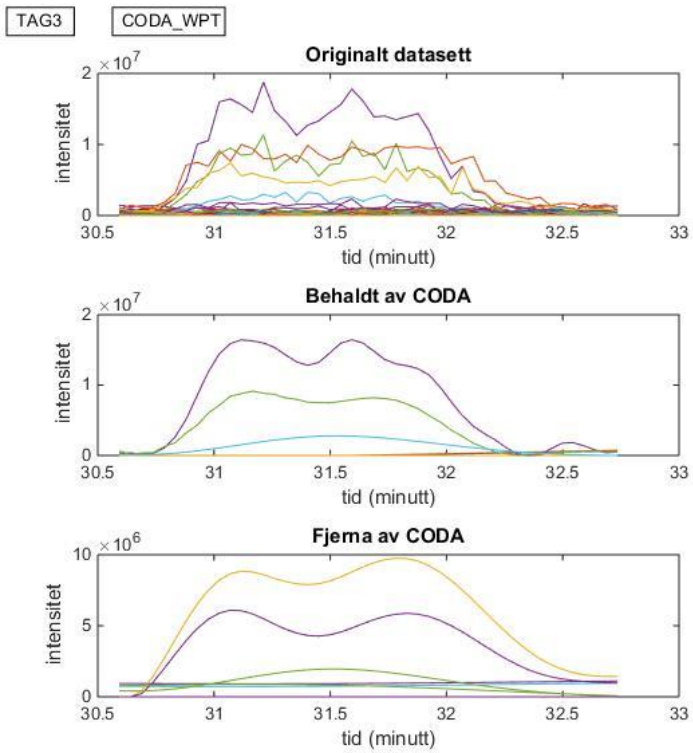
Vedlegg 21: TIC av *TAG1-settet; (øvst) originalt datasett, (midtarst) CODA_WPT-prosessert datasettet og (nedst) det som vert fjerna av CODA-delen av metoden.*



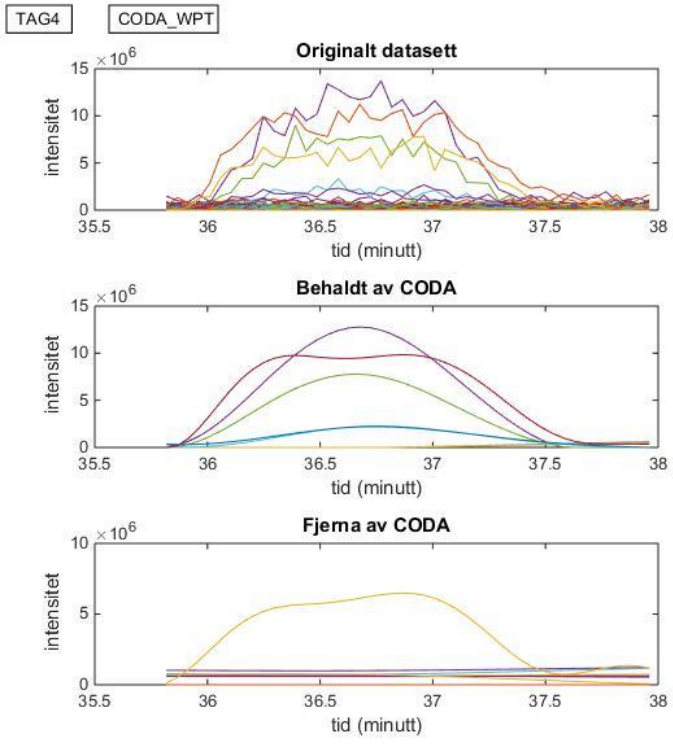

*Vedlegg 22: TIC av* TAG2-settet; (øvst) originalt datasett, (midtarst) CODA_WPT-prosessert datasettet og (nedst) det som vert fjerna av CODA-delen av metoden.



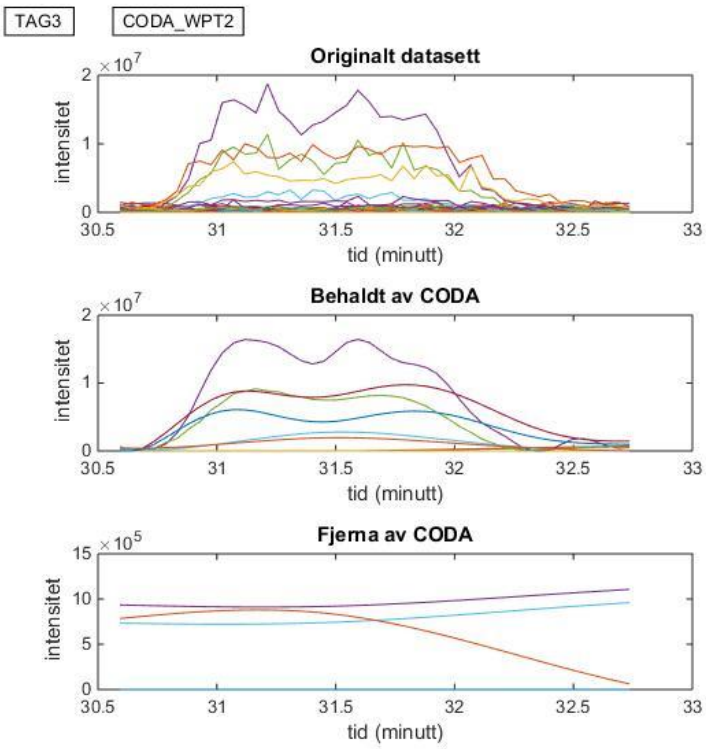
*Vedlegg 23: EIC av* TAG3-settet; (øvst) originalt datasett, (midtarst) CODA_WPT-prosessert datasettet og (nedst) det som vert fjerna av CODA-delen av metoden.



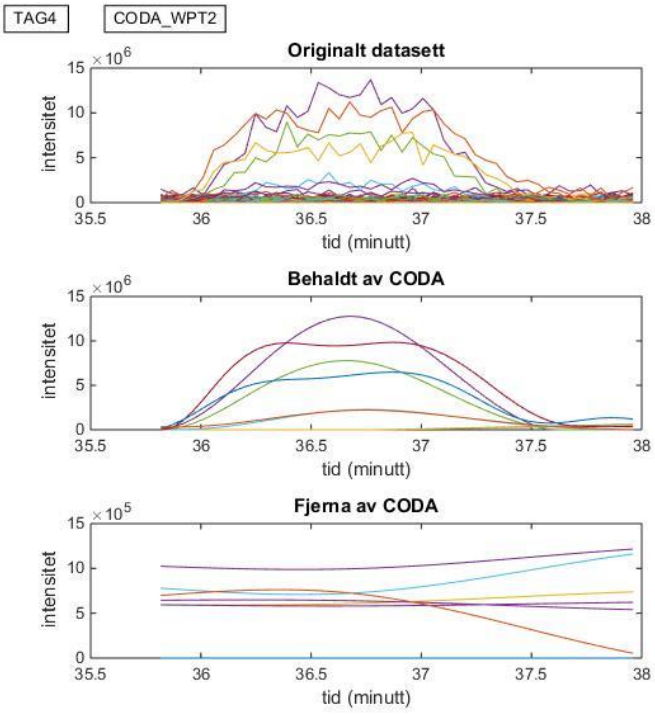

*Vedlegg 24: EIC av* TAG4-settet; (øvst) originalt datasett, (midtarst) CODA_WPT-prosessert datasettet og (nedst) det som vert fjerna av CODA-delen av metoden.



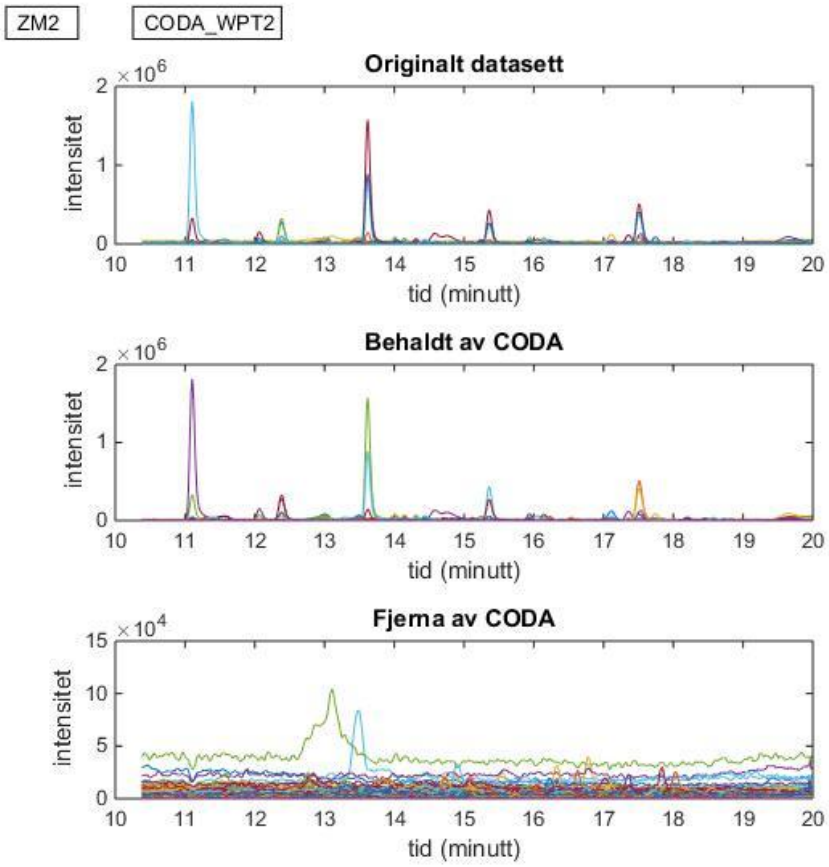
*Vedlegg 25: EIC av* TAG3-settet; (øvst) originalt datasett, (midtarst) CODA_WPT2-prosessert datasettet og (nedst) det som vert fjerna av CODA-delen av metoden.



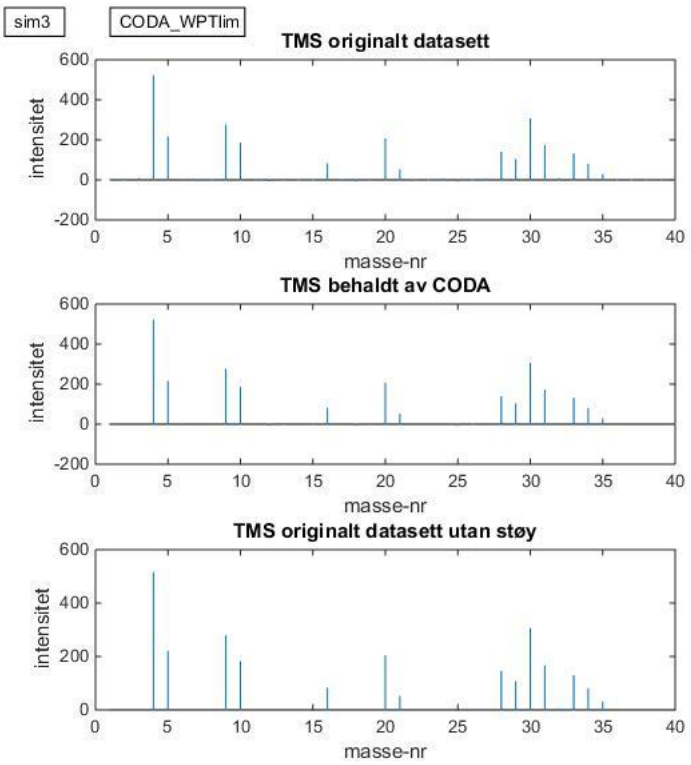

*Vedlegg 26: EIC av* TAG4-settet; (øvst) originalt datasett, (midtarst) CODA_WPT2-prosessert datasettet og (nedst) det som vert fjerna av CODA-delen av metoden.



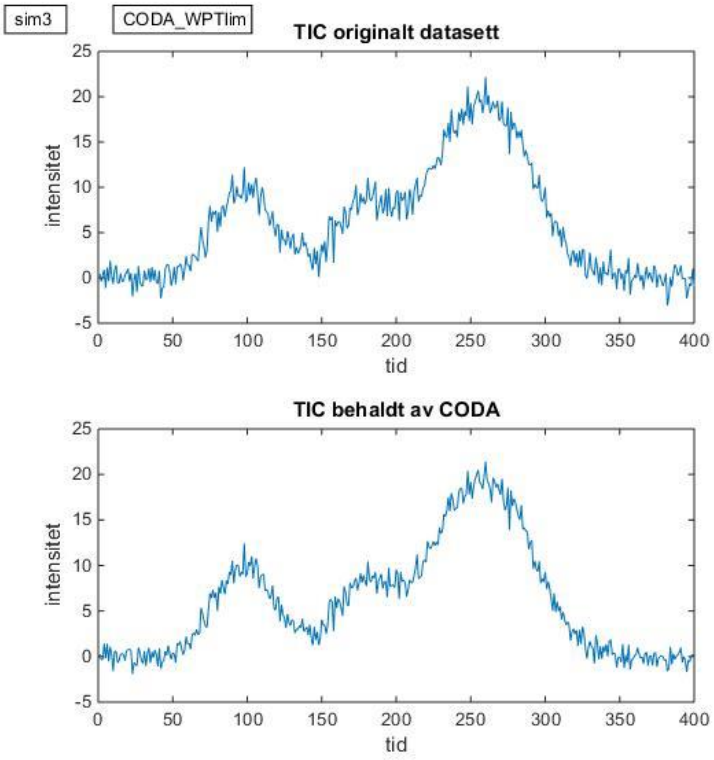
*Vedlegg 27: EIC av* ZM2-settet; (øvst) originalt datasett, (midtarst) CODA_WPT2-prosessert datasettet og (nedst) det som vert fjerna av CODA-delen av metoden.



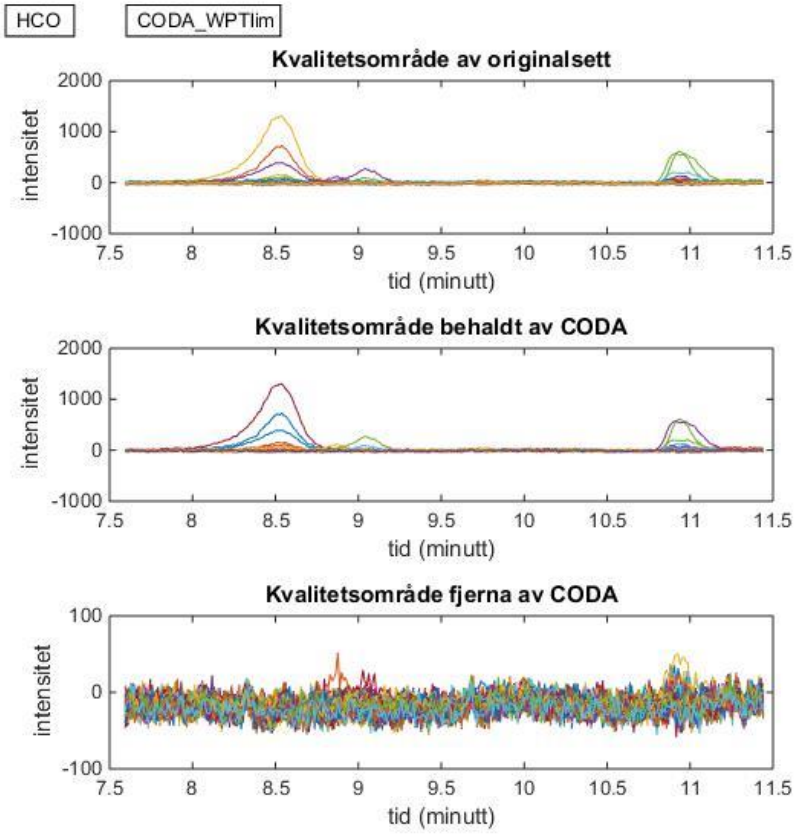
*Vedlegg 28: TMS av* sim3-settet; (øvst) originalt datasett, (midtarst) CODA_WPTlim-prosessert datasettet og (nedst) originalt datasett utan støy.



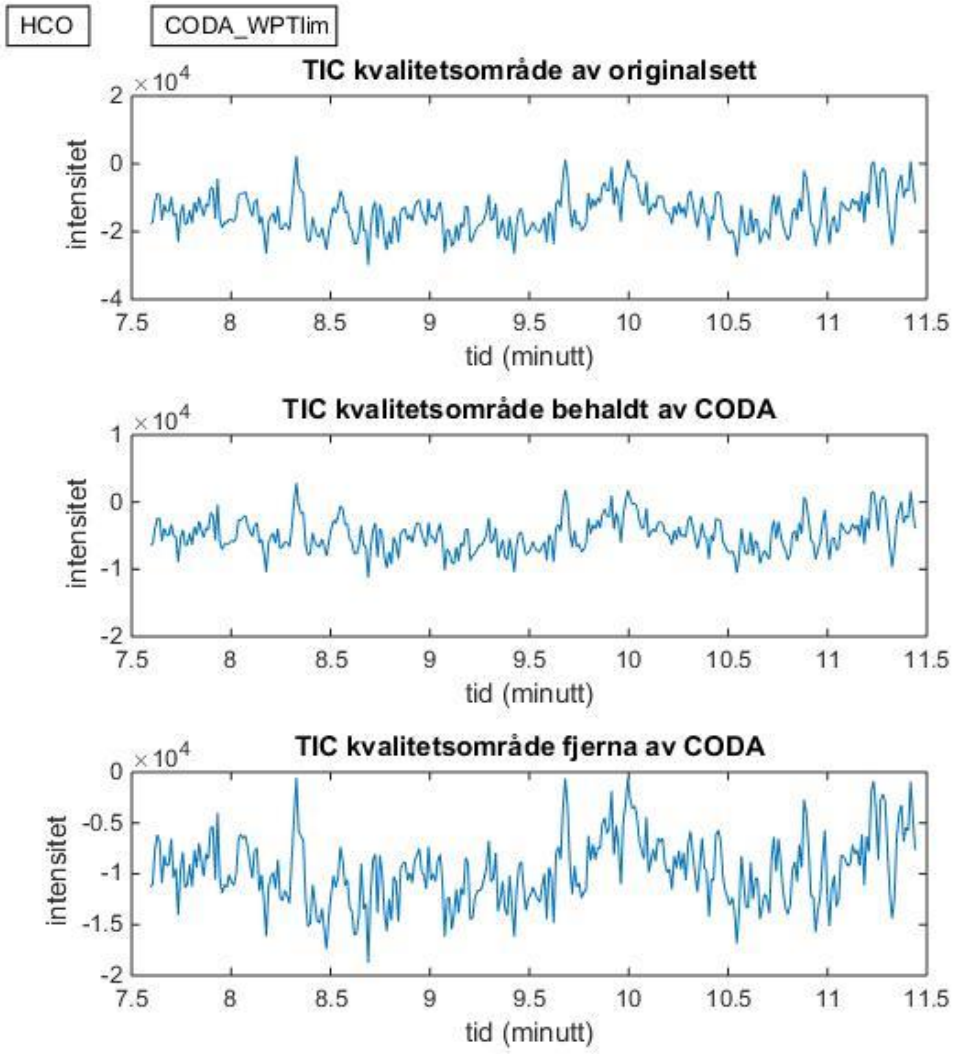

*Vedlegg 29: TIC av* sim3-settet; (øvst) originalt datasett og (nedst) CODA_WPTlim-prosessert datasettet.



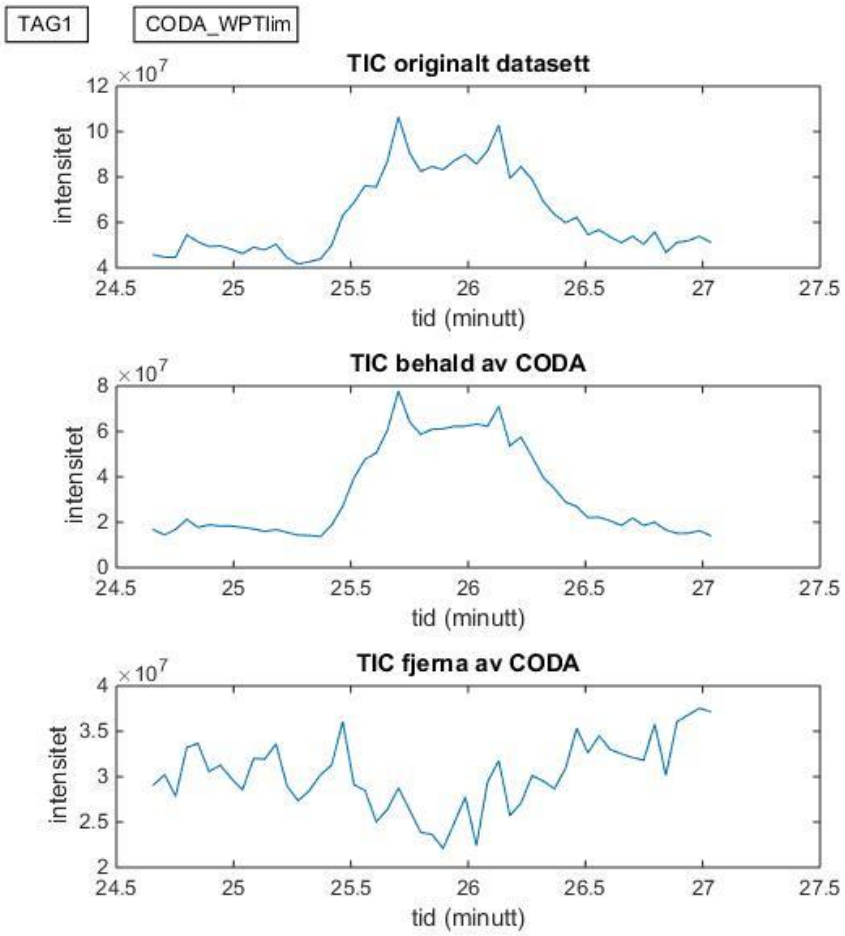
*Vedlegg 30: EIC av kvalitetsområdet til* HCO-settet; (øvst) originalt datasett, (midtarst) CODA_WPTlim-prosessert datasettet og (nedst) det som vert fjerna av metoden.



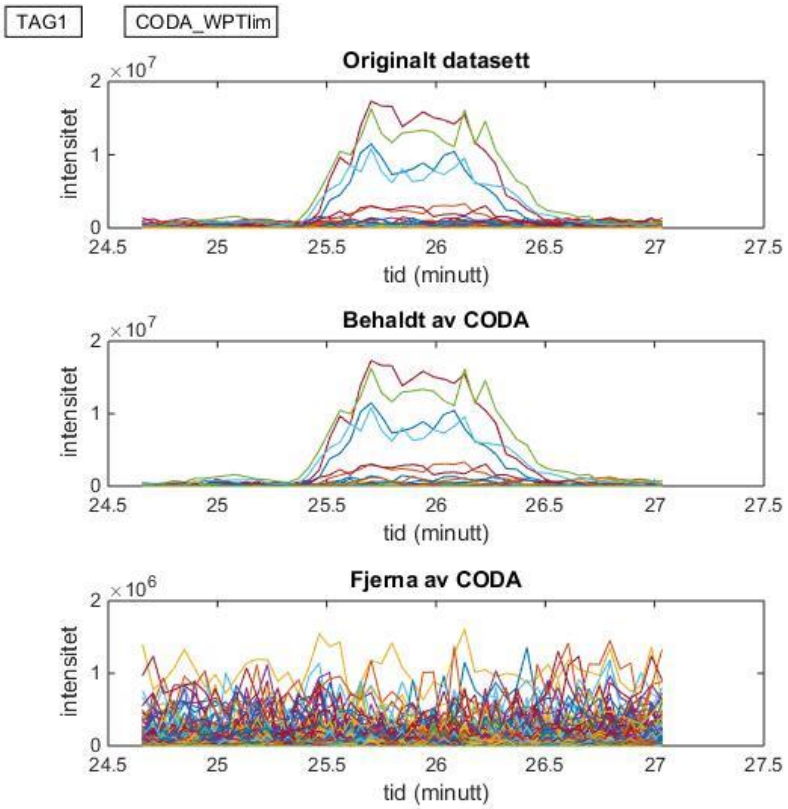
*Vedlegg 31: TIC av kvalitetsområdet til* HCO-settet; (øvst) originalt datasett, (midtarst) CODA_WPTlim-prosessert datasettet og (nedst) det som vert fjerna av metoden.



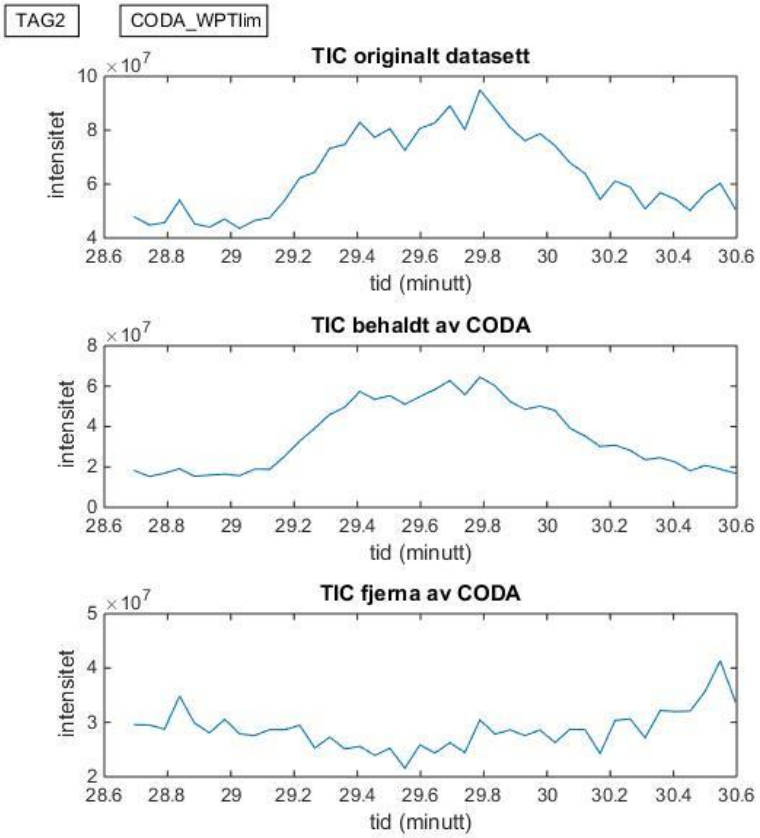
*Vedlegg 32: TIC av TAG1*-settet; (øvst) originalt datasett, (midtarst) CODA_WPTlim-prosessert datasettet og (nedst) det som vert fjerna av metoden.



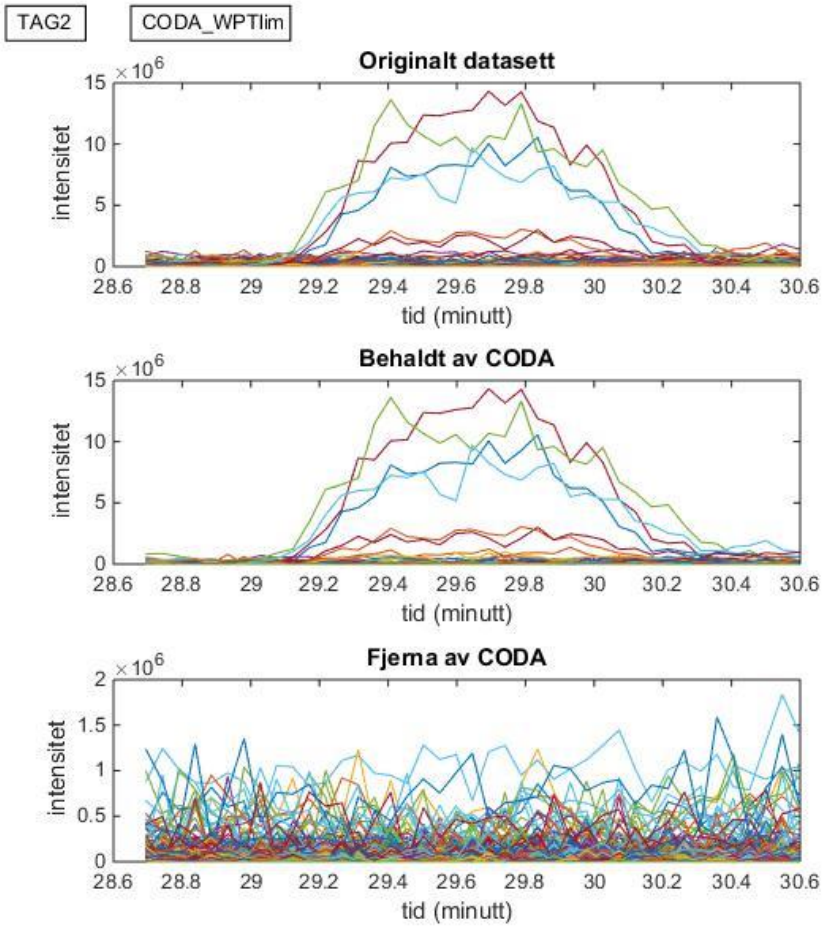
*Vedlegg 33: EIC av TAG1*-settet; (øvst) originalt datasett, (midtarst) CODA_WPTlim-prosessert datasettet og (nedst) det som vert fjerna av metoden.



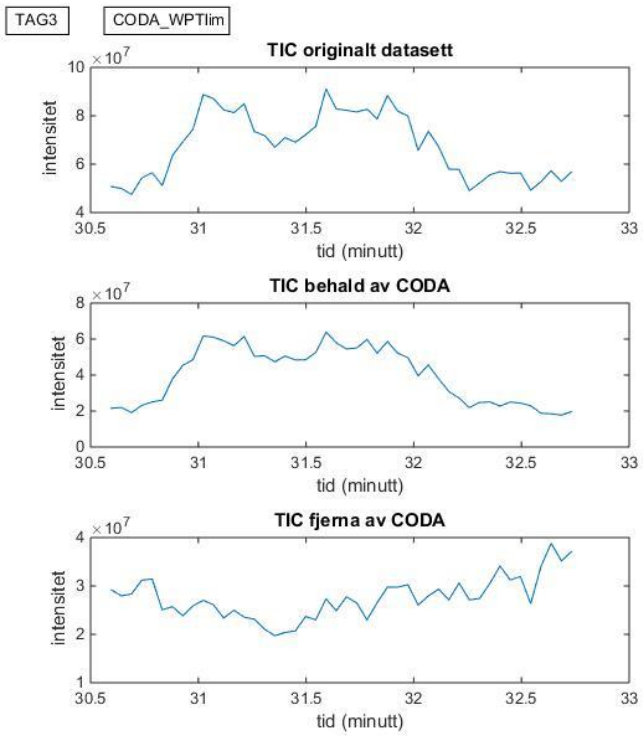
*Vedlegg 34: TIC av TAG2*-settet; (øvst) originalt datasett, (midtarst) CODA_WPTlim-prosessert datasettet og (nedst) det som vert fjerna av metoden.



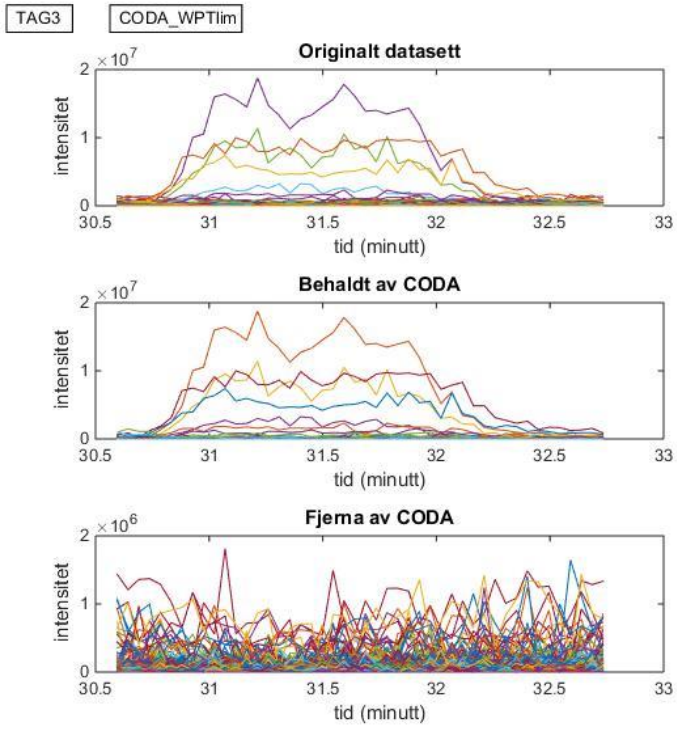
*Vedlegg 35: EIC av TAG2*-settet; (øvst) originalt datasett, (midtarst) CODA_WPTlim-prosessert datasettet og (nedst) det som vert fjerna av metoden.



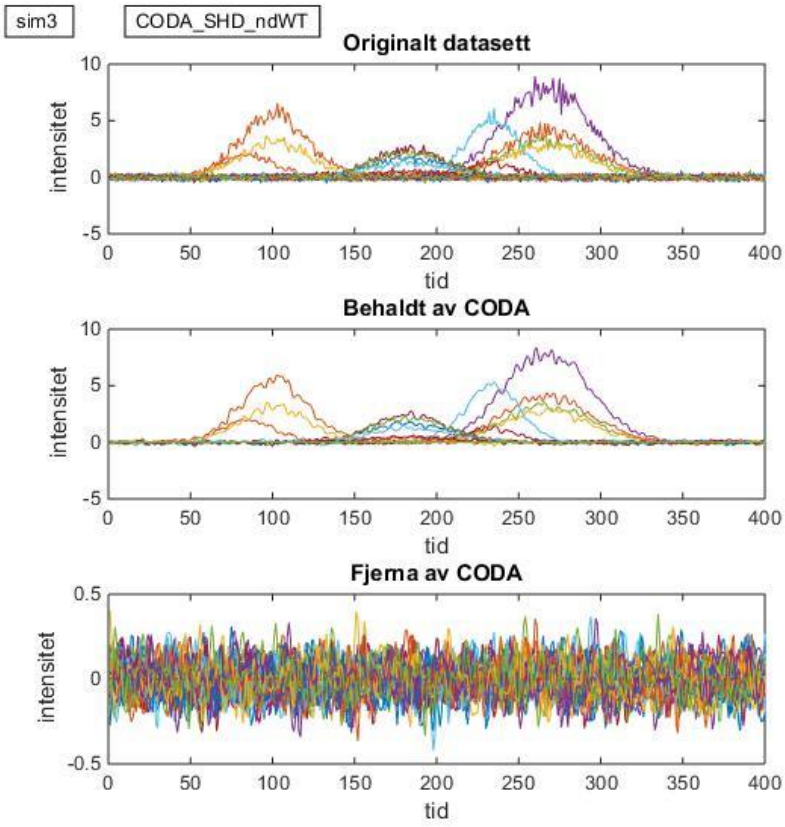
*Vedlegg 36: TIC av TAG3*-settet; (øvst) originalt datasett, (midtarst) CODA_WPTlim-prosessert datasettet og (nedst) det som vert fjerna av metoden.



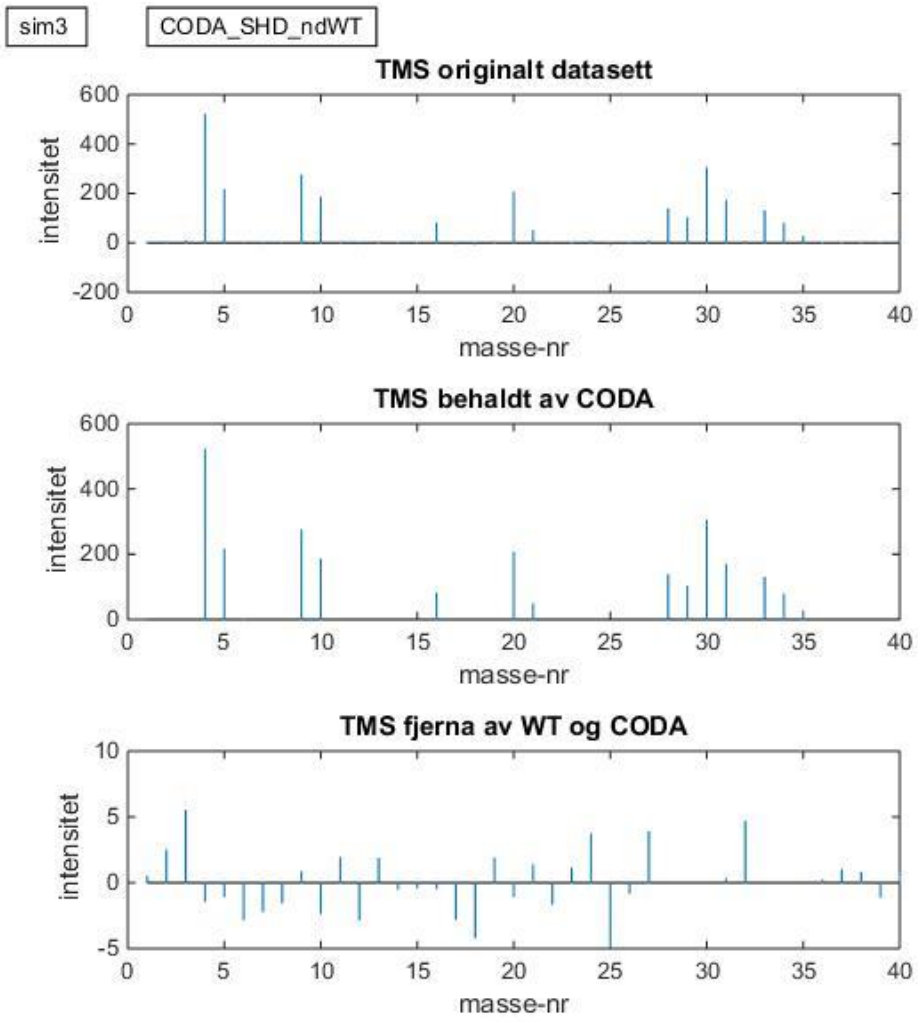

*Vedlegg 37: EIC av TAG3*-settet; (øvst) originalt datasett, (midtarst) CODA_WPTlim-prosessert datasettet og (nedst) det som vert fjerna av metoden.



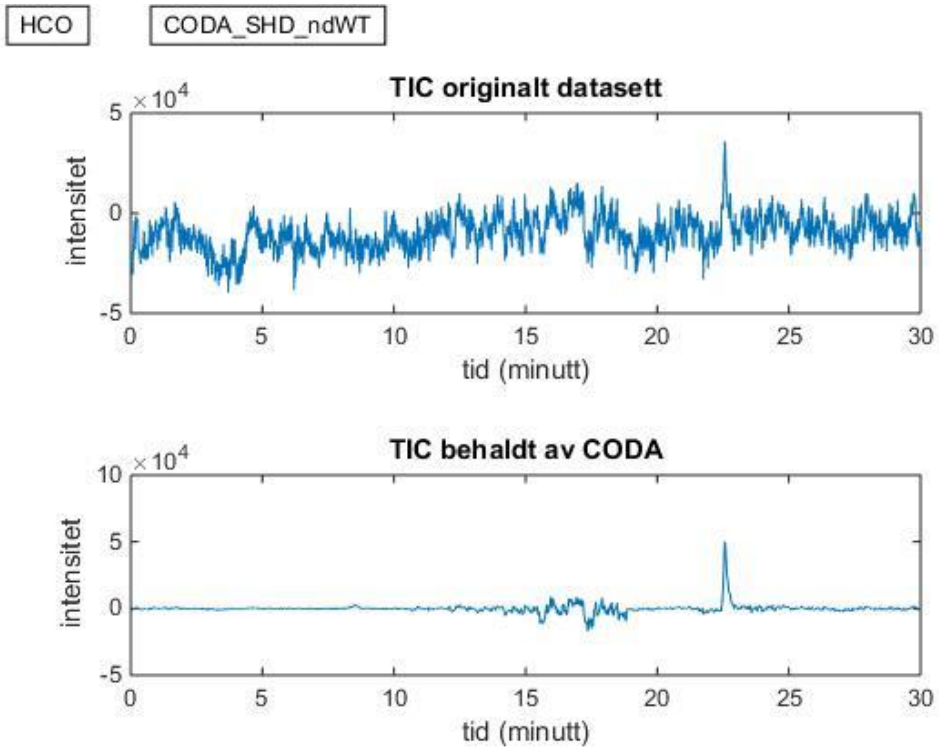
*Vedlegg 38: TIC av sim3*-settet; (øvst) originalt datasett, (midtarst) CODA_SHD_ndWT-prosessert datasettet og (nedst) det som vert fjerna av metoden.



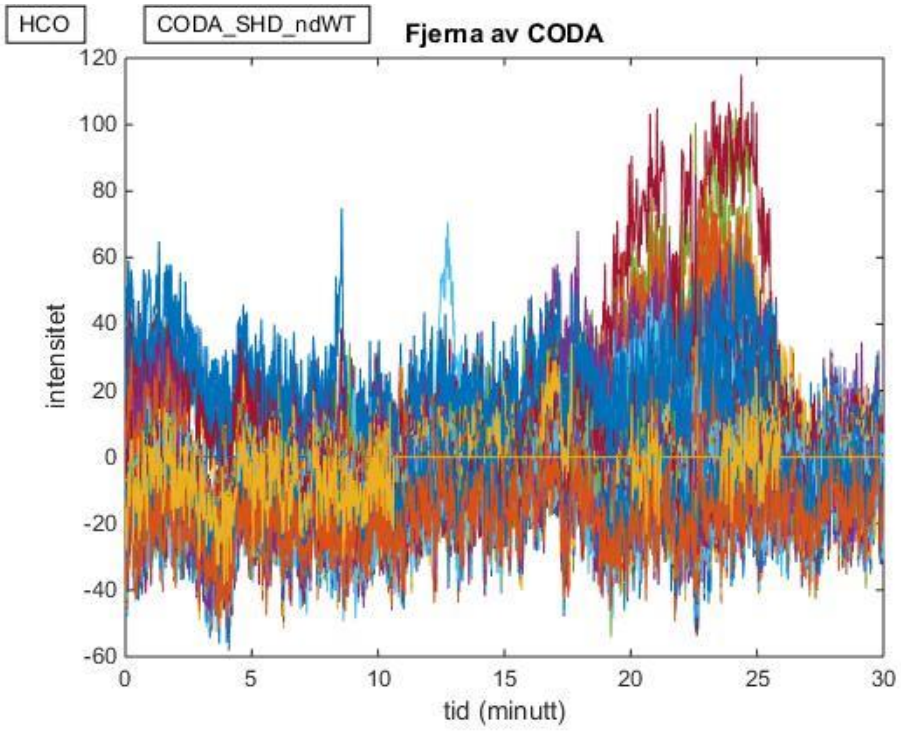
*Vedlegg 39: TMS av sim3*-settet; (øvst) originalt datasett, (midtarst) CODA_SHD_ndWT-prosessert datasettet og (nedst) det som vert fjerna av metoden.



*Vedlegg 40: TIC av HCO*-settet; (øvst) originalt datasett og (nedst) CODA_SHD_ndWT-prosessert datasettet.




*Vedlegg 41: HCO-settet; signal som vert fjerna av CODA-delen av CODA_SHD_ndWT*



Vedlegg 42:  Konvertering av ZM2-fil til matlab-formatet «.mat»

Vedlegget er ein litt modifisert epost frå Svein Are Mjøs

%%% Datastruktur

Når du åpner den i Matlab får du fire variabler: Boxset, Databox,

Info, Metset.

Dataene ligger i Databox. Dette er en 1x1 struct med en masse felt. De

viktige er:

- X: Signalene

- rt: Retensjonstidene

- ions: Massene

For å plotte ionekromatogrammene for filen skriver du f.eks:

"plot(Databox(1).rt,Databox(1).X)"

For å plotte spektrene ved hver retensjonstid skriver du

"plot(Databox(1).ions,Databox(1).X')" - merk at X er transponert

%%% Konvertering av data frå instrumentdata til mat-fil

Her er konverteringsparametrene:

1) Konvertering fra MassHunter til mzXML:

Brukte MSConvert fra proteowizard 3.0.7127

(http://proteowizard.sourceforge.net/)

- 32bit presisjon,

- ingen zlib-pakking,

- 'TPP compatibility' av,

- 'Write index' på

2) Konvertering fra mzXML til Matlab:

Brukte Chrombox D 12-09b (www.chrombox.org)

- Resolution: 1 (amu)

- Mass offset: 0 (amu)

- Mass Win: 100 (%)

- Ingen filtre for bakgrunnsfjerning

- Konvertert i Matlab R2013a

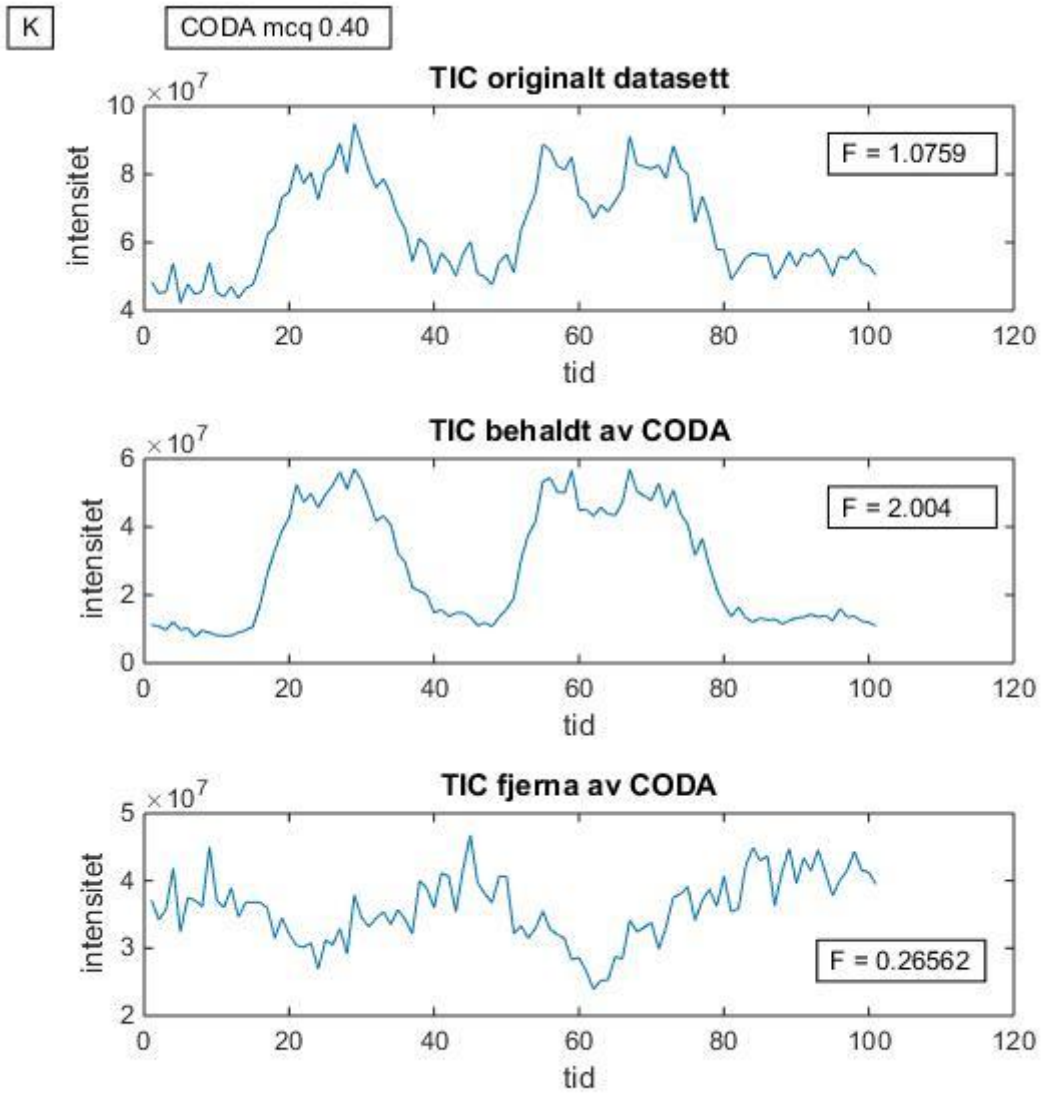
*Vedlegg 43: Konvertering av HCO-fil til matlab-formatet «.mat»*

Data-settet låg på 7rw-fil i instrument-datamaskinen. Dette vart konvertert frå profildata til centroid-data, og deretter henta ut som netCDF-format i ei cdf-fil

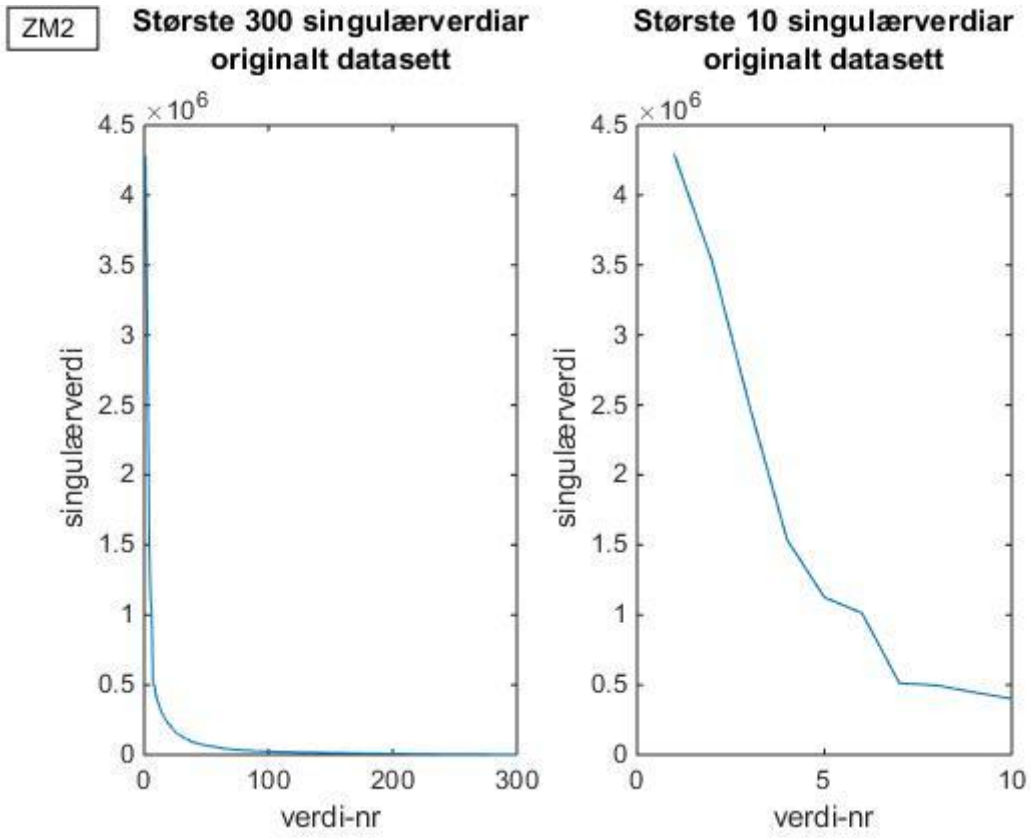
Fila vart konvertert frå cdf-format til mat-format ved bruk av matlab-funksjonen readms.m (skriven av Bjørn Grung) . Funksjonsdokumentasjonen er lagt ved under.

```
%%%  dokumentasjon start

%readms(filename,resolution)
%
%This function reads a NetCDF file and saves the data as a MATLAB file with
%the same name.
%
%filename - the name of the NetCDF file. The MATLAB file will have the same
%name.
%resolution is an optional parameter. If left unspecified, or as 0, the
%instrumental resolution will be used. Use 1 to get data with a mass
%resolution of 1.
%
%B Grung Feb 16th 2015

%%% dokumentasjon slutt
```

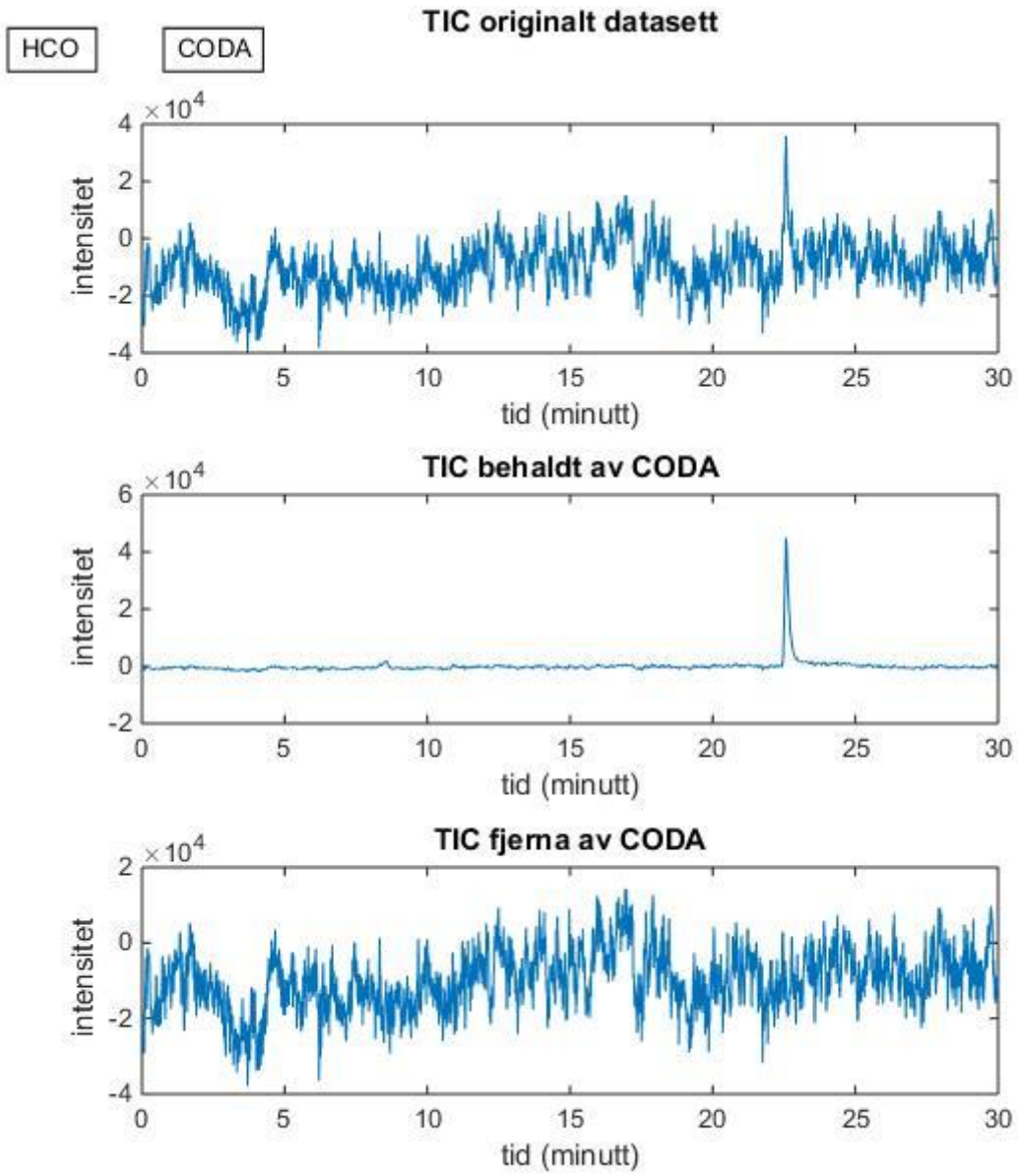
*Vedlegg 44: Område K ved CODA for MCQ = 0.40 . Singulærverdi-forholdet F er gjeve for (øvst) originalt sett , (midtarst) behaldt av CODA og (nedst) fjerna av CODA*



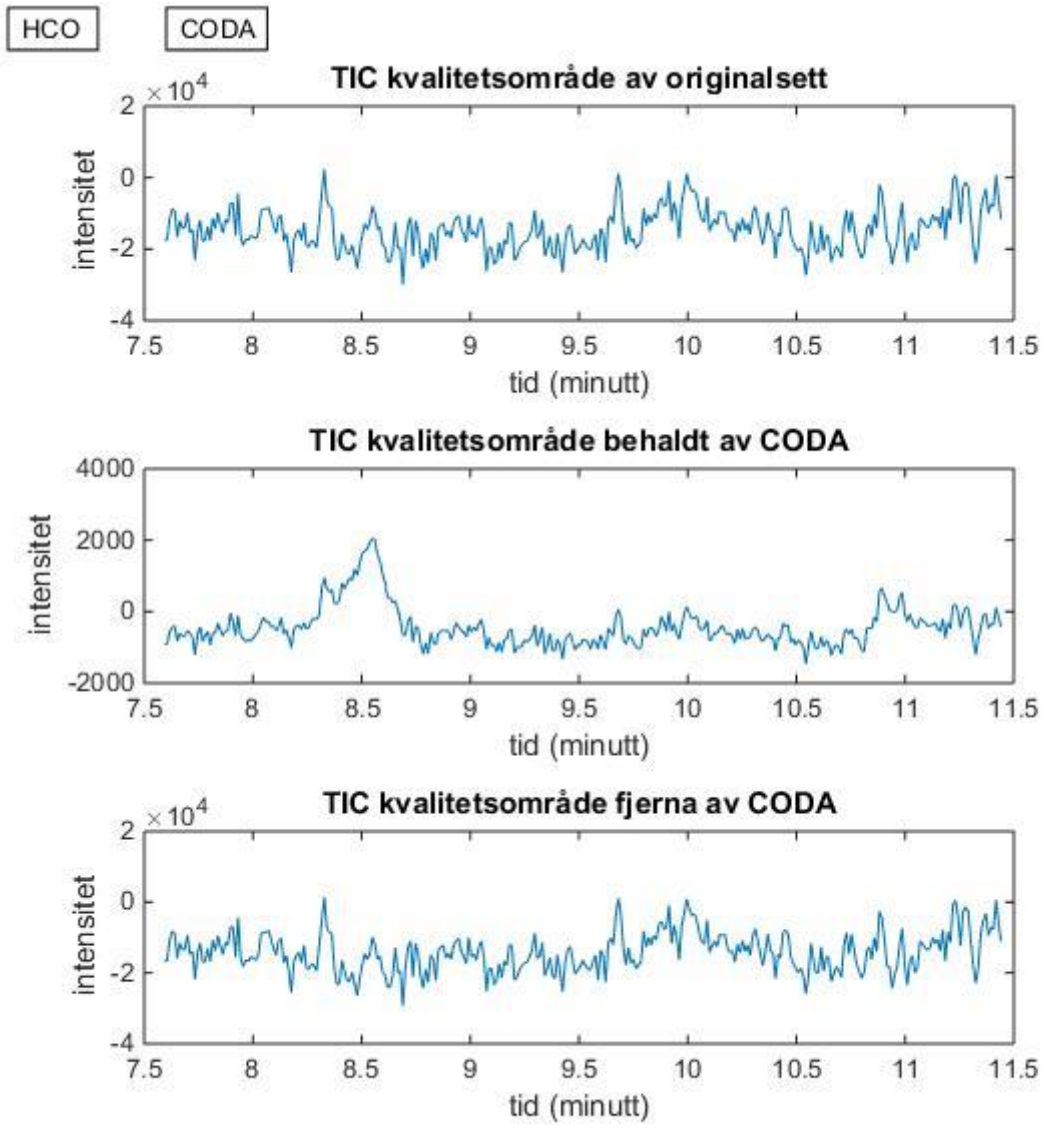
*Vedlegg 45: Singulærverdiar for ZM2-settet*



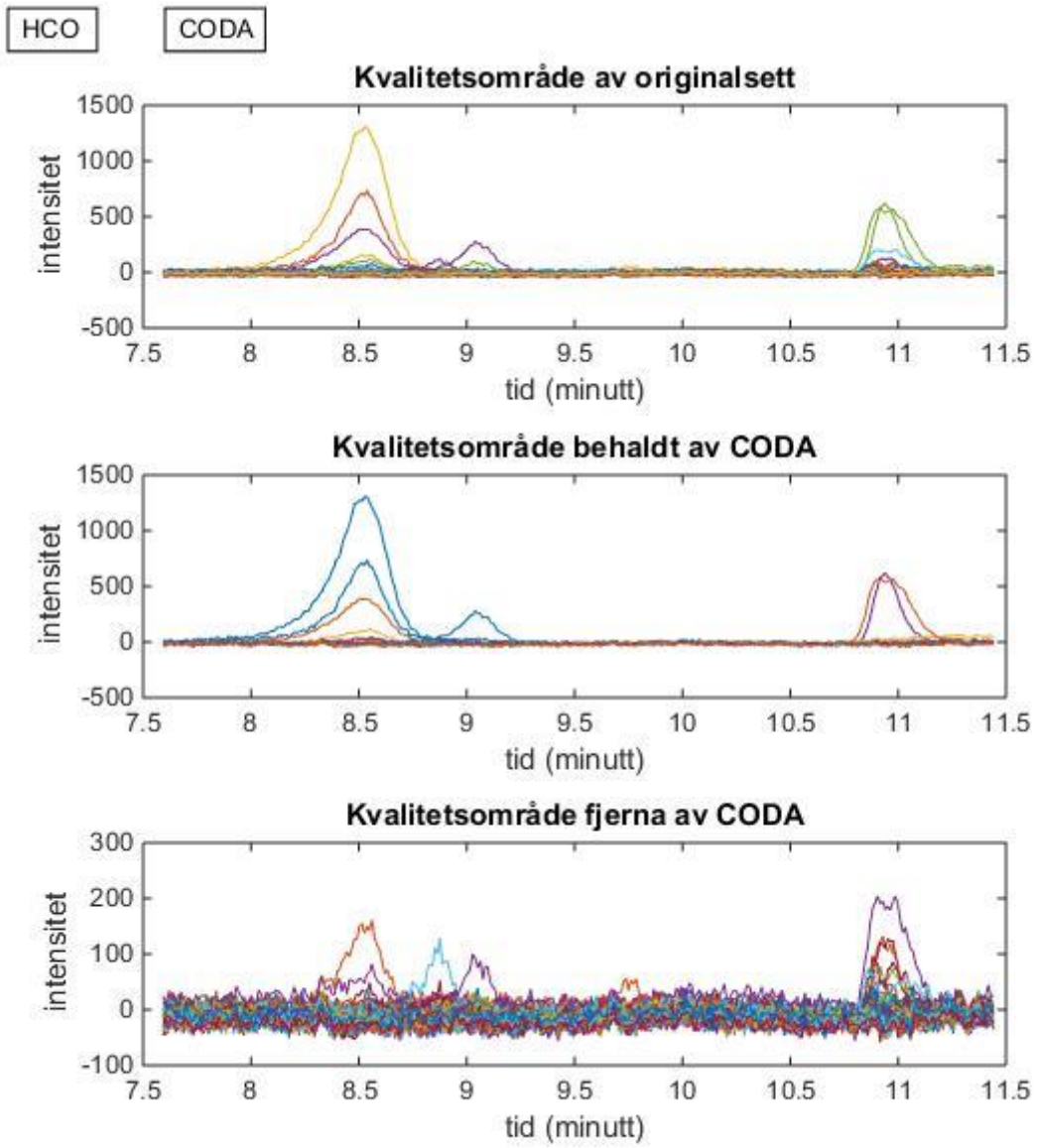
*Vedlegg 46: TIC for HCO; (øvst) originalsett, (midtarst) prosessert sett og (nedst) fjerna av CODA.*



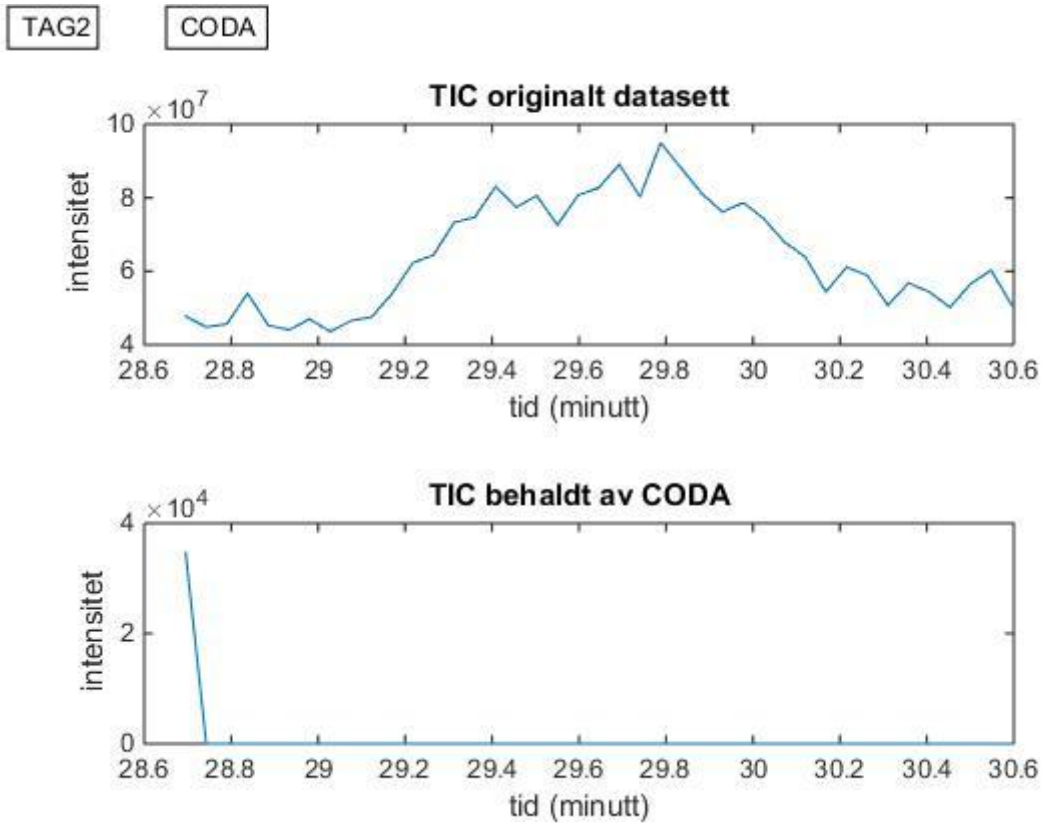
*Vedlegg 47: TIC av kvalitetsområdet for HCO; (øvst) originalsett, (midtarst) prosessert sett og (nedst) fjerna av CODA.*



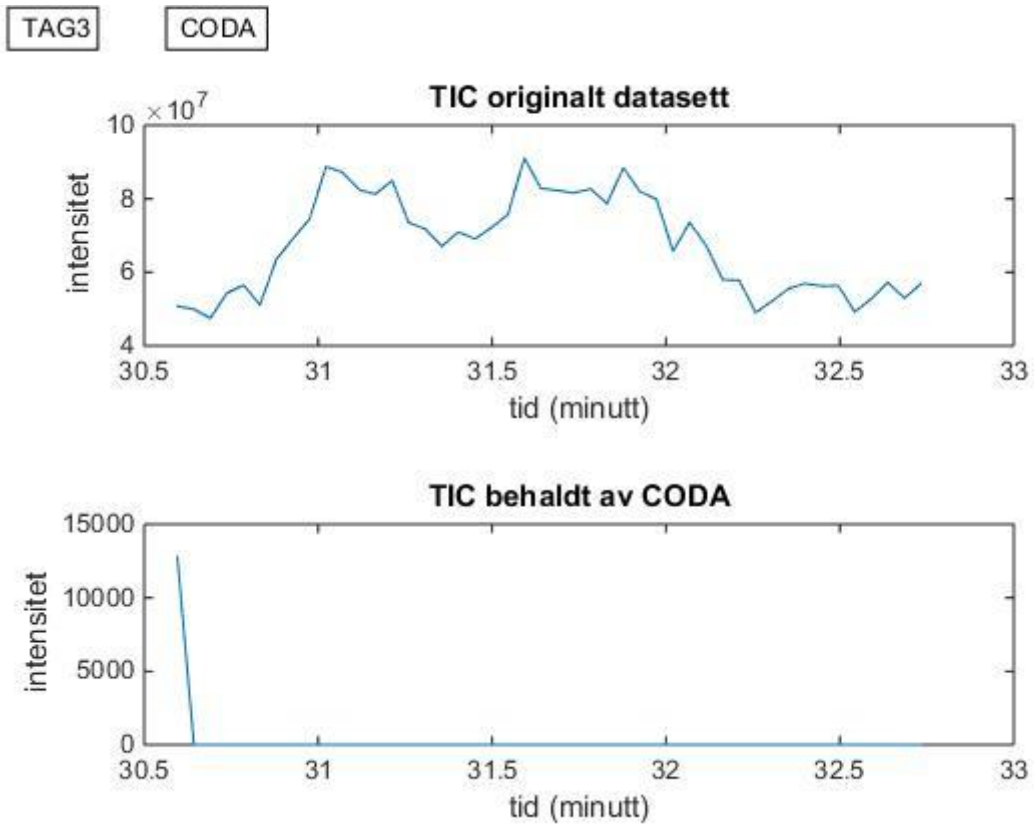
*Vedlegg 48: EIC av kvalitetsområdet for HCO; (øvst) originalsett, (midtarst) prosessert sett og (nedst) fjerna av CODA.*



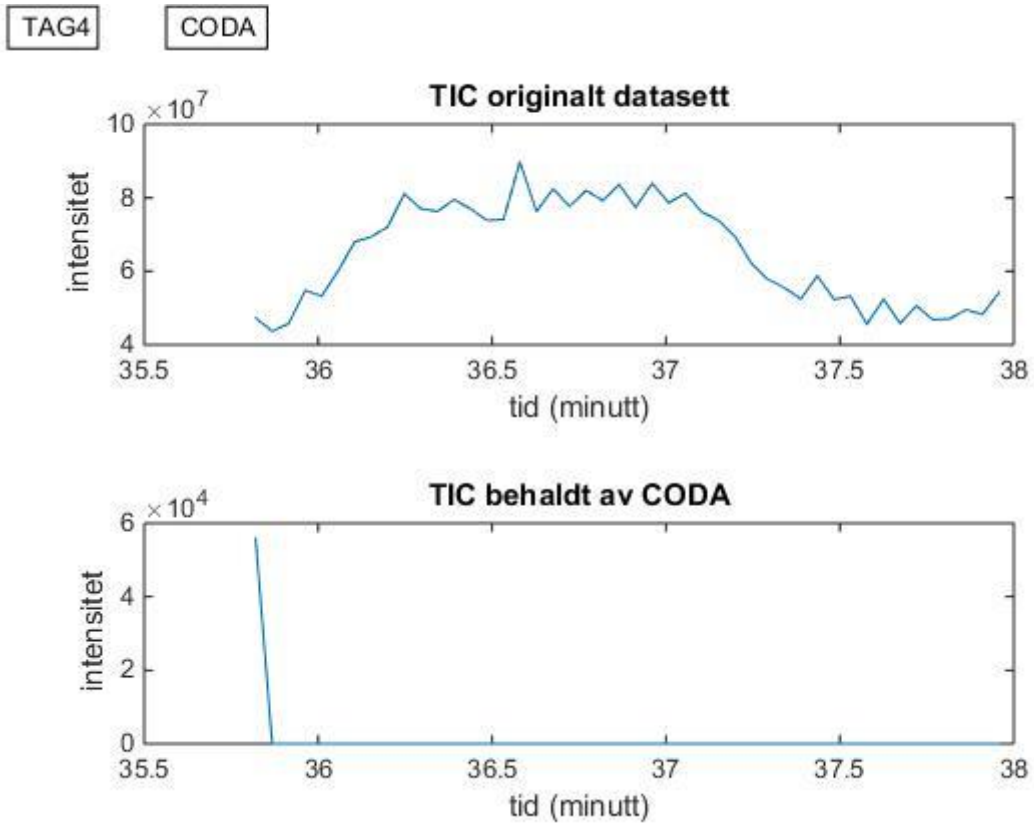
*Vedlegg 49:TIC av TAG2-settet; (øvst) originalt datasett og (nedst) datasett behaldt av CODA med mcq = 0.85*



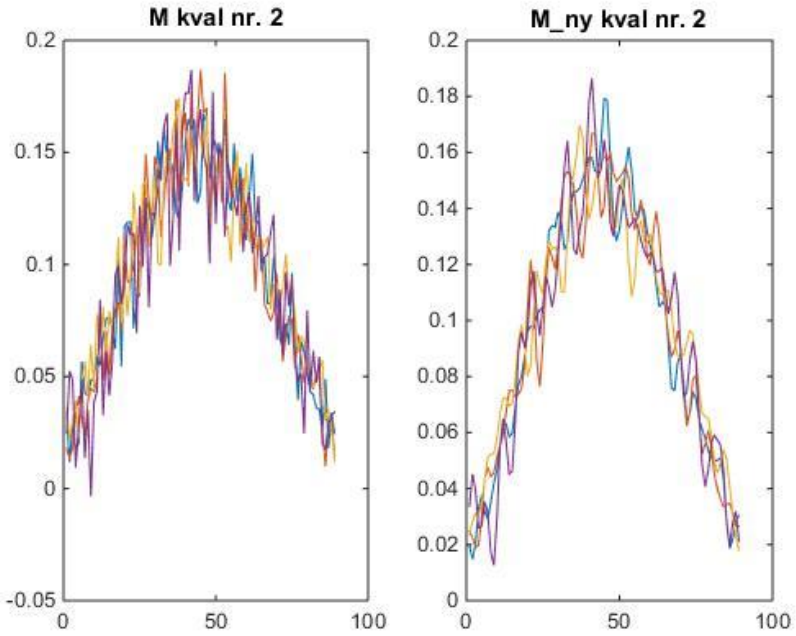
*Vedlegg 50: TIC av TAG3-settet; (øvst) originalt datasett og (nedst) datasett behaldt av CODA med mcq = 0.85*



*Vedlegg 51: TIC av TAG4-settet; (øvst) originalt datasett og (nedst) datasett behaldt av CODA med mcq = 0.85*

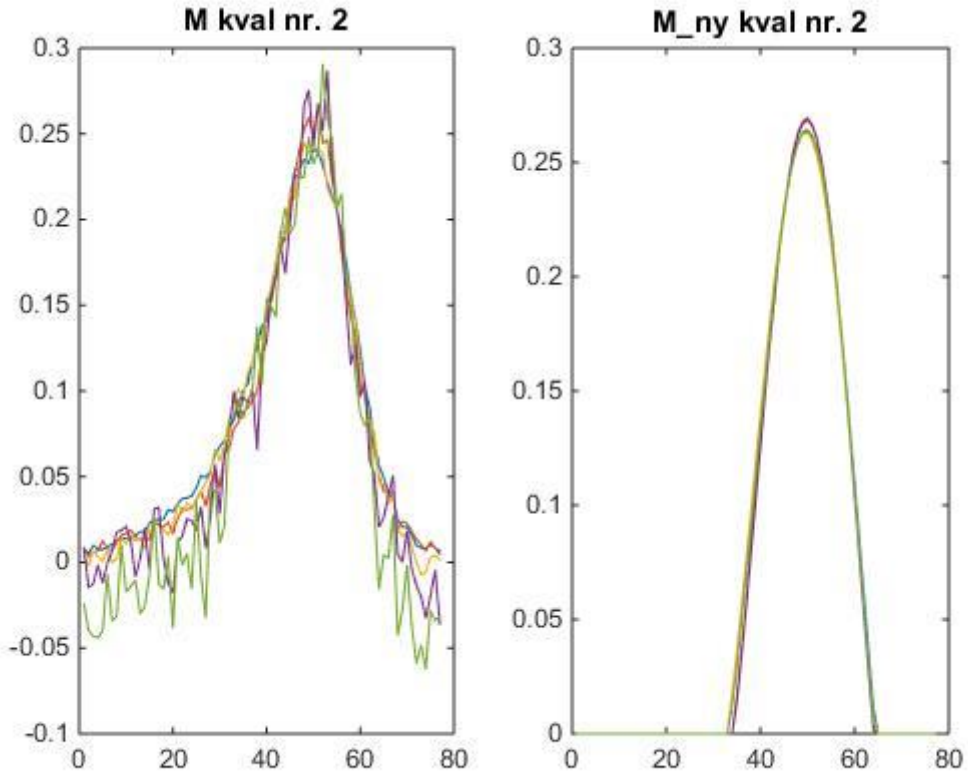


Vedlegg 52: Utvald område nr. 2 for toppsamanlikning av sett3, før og etter CODA_ndWT



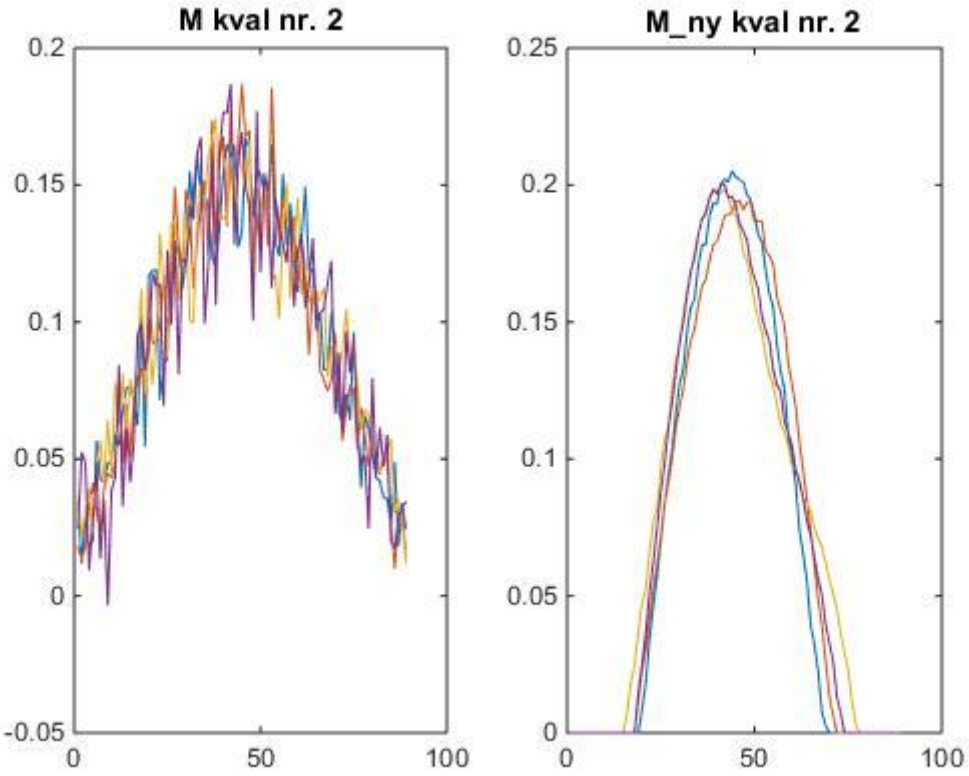
*Vedlegg 53: HCO kvalitetstopp 2 ved CODA_CWT*

Her er M og M_ny det originale datasettet og det nye.



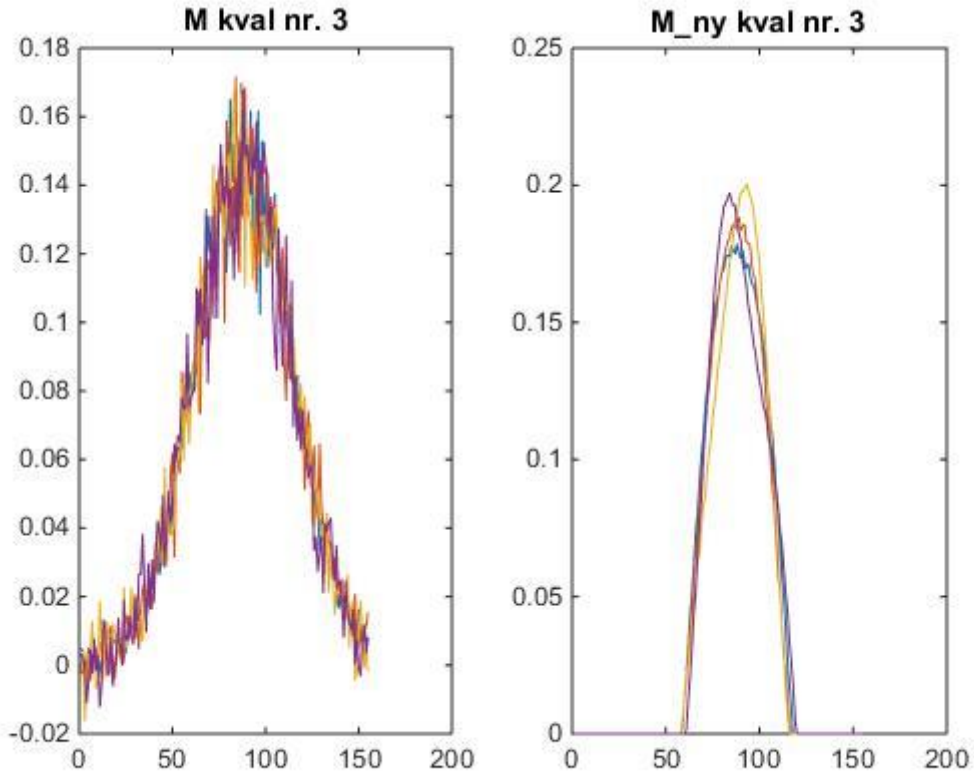
*Vedlegg 54: sim3 kvalitetstopp 2 ved CODA_CWT*

*Her er M og M_ny det originale datasettet og det nye*



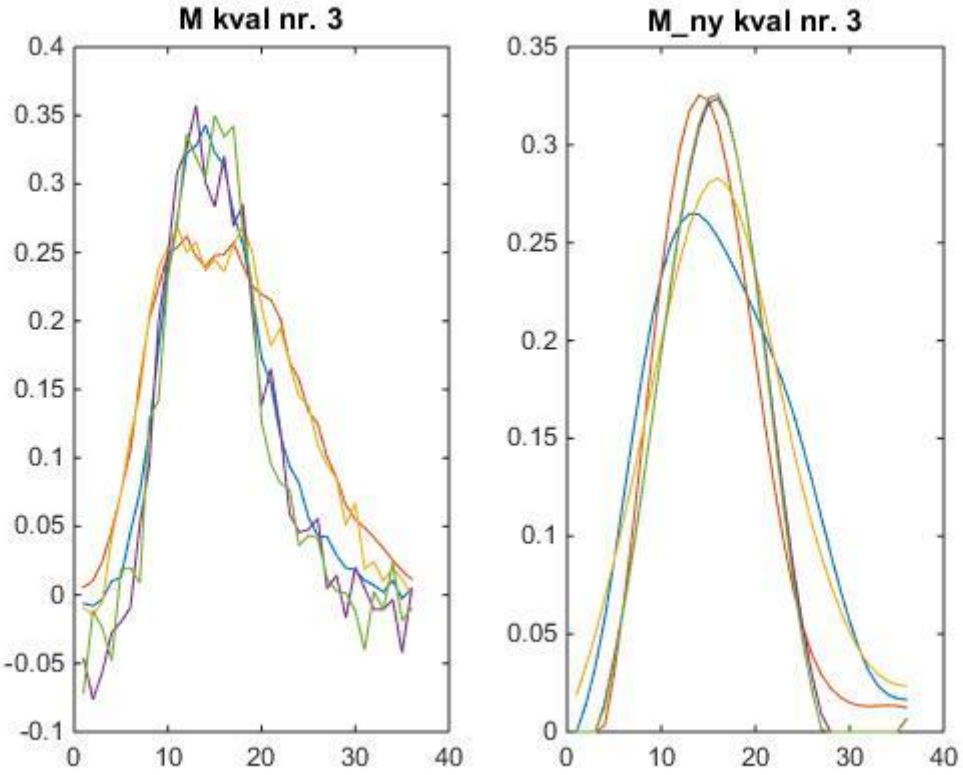
*Vedlegg 55: sim3 kvalitetstopp 3 ved CODA_CWT*

Her er M og M_ny det originale datasettet og det nye.



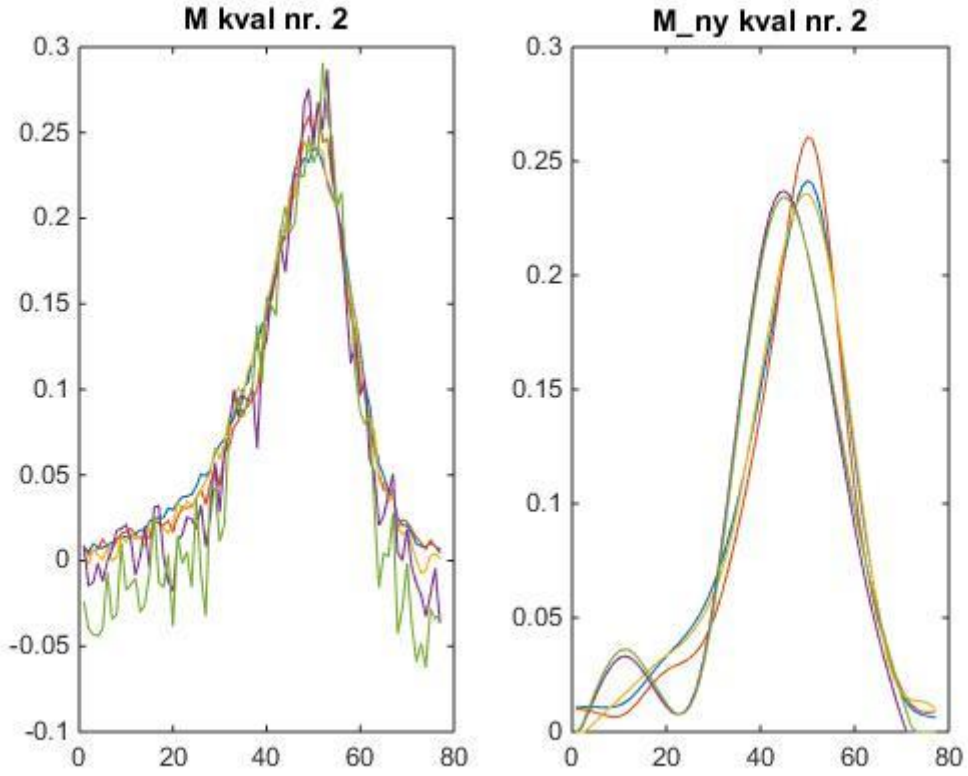
*Vedlegg 56: HCO – kvalitetstopp 3 ved CODA_WPT*

Her er M og M_ny det originale datasettet og det nye.



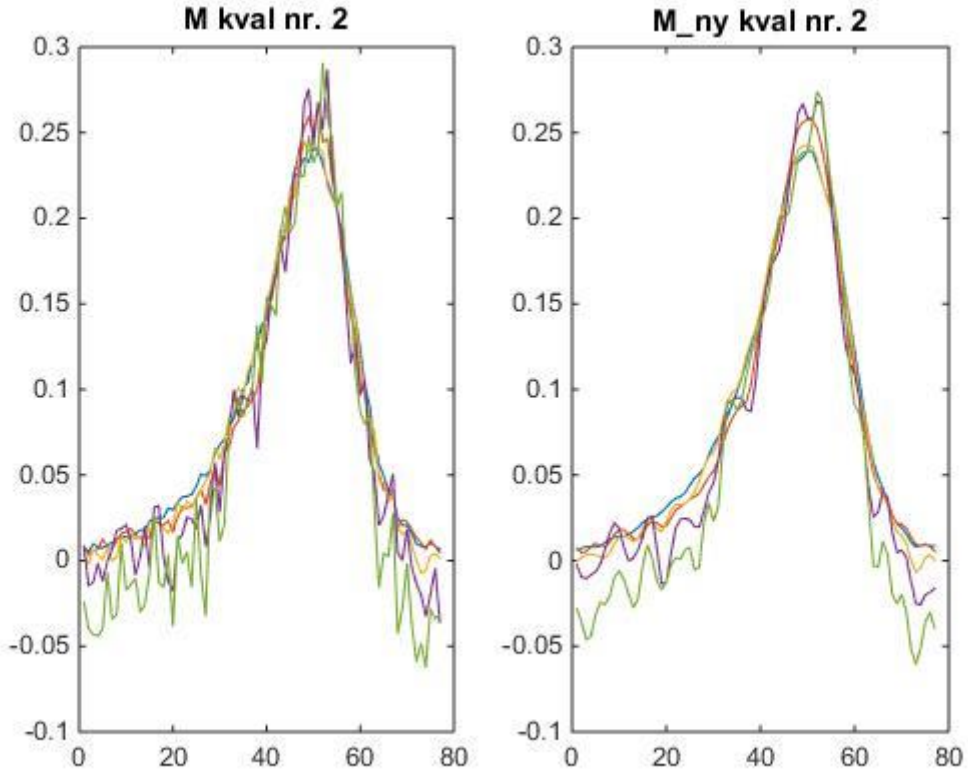
*Vedlegg 57: HCO – kvalitetstopp 2 ved CODA_WPT*

Her er M og M_ny det originale datasettet og det nye.



*Vedlegg 58: HCO – kvalitetstopp 2 ved CODA_ndWT*

Her er M og M_ny det originale datasettet og det nye.



*Vedlegg 59: HCO – kvalitetstopp 3 ved CODA_ndWT*

Her er M og M_ny det originale datasettet og det nye.