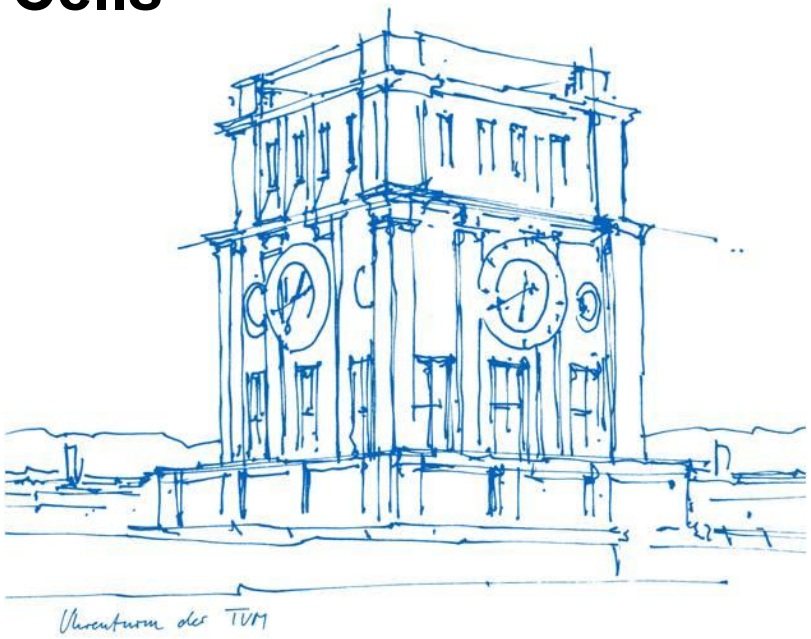
TUM Data Innovation Lab with cellasys

How to Handle Data from Living Cells

Anne Christopher, Magdalena Eberl, Sebastian Zett

Munich, August 06, 2019



Agenda

1	Introduction
2	Data Collection & Pre-Processing
3	Data Analysis & Results
4	Summary & Conclusion

Agenda

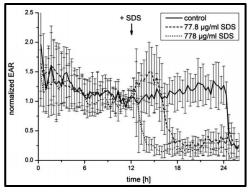
1	**Introduction**
2	Data Collection & Pre-Processing
3	Data Analysis & Results
4	Summary & Conclusion

cellasys

Analysis of Living Cells

Microphysiometry

- Novel methodology of **analyzing living cells**
- Interface of **electronic engineering and life sciences** (biology, chemistry)
- Electrochemical and optochemical sensor technology to **record cell metabolism**

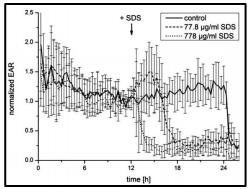


- Use of algorithms and models to **draw conclusions** from this raw data:
  → prediction models for toxicological effects
  → development of new drugs

Analysis of Living Cells

Microphysiometry

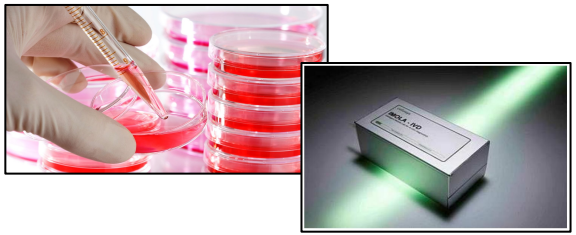
- Novel methodology of **analyzing living cells**
- Interface of **electronic engineering and life sciences** (biology, chemistry)
- Electrochemical and optochemical sensor technology to **record cell metabolism**



- Use of algorithms and models to **draw conclusions** from this raw data:
  → prediction models for toxicological effects
  → development of new drugs

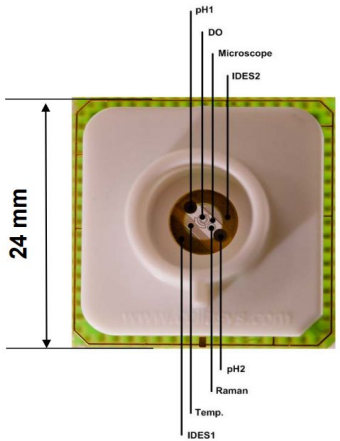
cellasys

- Founded in 2007 as TUM spin-off
- Specialized in providing systems and solutions for microphysiometry



cellasys' Solution for the Analysis of Living Cells

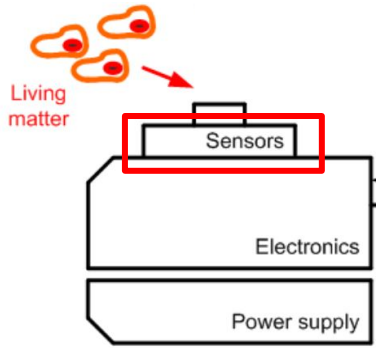
Bio-Chip



- Sensors for pH, oxygen, temperature & impedance

IMOLA-IVD*

**Intelligent Mobile Lab for in-vitro diagnostic (IMOLA-IVD)*



- Boxed in-vitro cell chip test system
- Contains the Bio-Chip

Incubator

- Assures homogeneous temperature distribution
- Contains 6 IMOLA-IVDs

Experiments Follow the ChemDef Protocol

Phase Nr.	Hours	Reference Group		Test Group
		Negative and positive control with cell culture	Blank without cell culture	4 replicates with cell culture
1	0 - 6 h	optimal medium (DMEM* + 10% FBS**)	optimal medium	optimal medium
2	6 - 12 h			test medium
3	12 - 16 h			optimal medium
4	16 - 20 h			test medium
5	20 - 24 h	toxic medium (0.2% SDS***)	toxic medium	toxic medium

** DMEM: Dulbecco's Modified Eagle Medium, **FBS: Fetal Bovine Serum, ***SDS: Sodium Dodecyl Sulfate*

Project Goals

Optimize existing approaches of preparation & analysis

Develop methods to **reduce the noise** in the data

Develop methods to **assess the validity** of the data

Project Goals

- **Optimize existing approaches** of preparation & analysis
- Develop methods to **reduce the noise** in the data
- Develop methods to **assess the validity** of the data

Integrate those methods into cellasys' software environment DALiA

Agenda

1	Introduction
2	**Data Collection & Pre-Processing**
3	Data Analysis & Results
4	Summary & Conclusion

Dataset Description

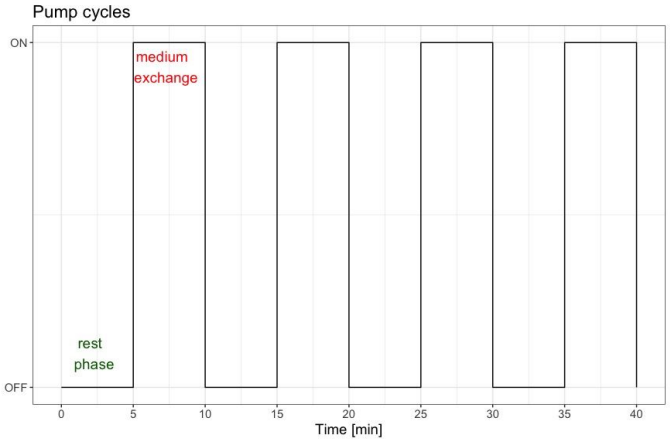
Data from **two 24h experiments**:

- Each experiment uses **6 IMOLAs**
- Data comes as .exp (text format) file
- Contains measurement recordings, air bubble detections and current configuration information
- Information about **valid and invalid IMOLAs provided** (see table)

Exp.	Valid	Invalid
1	IMOLA 1, 2, 3, 4	IMOLA 5, 6
2	IMOLA 2, 3, 5	IMOLA 1, 4, 6

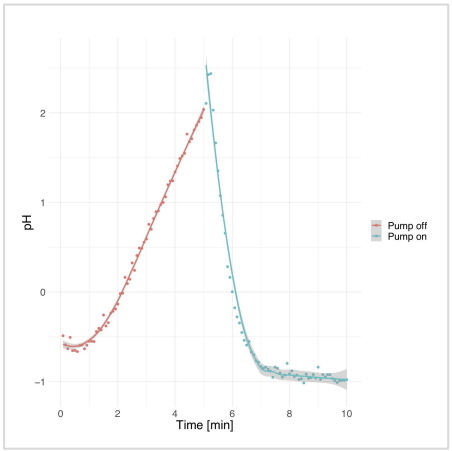
Pump Intervals

- **Fluidic system provides fresh cell culture medium** to cells on chips
- **Pump cycle:** pump switches between OFF (rest phase) and ON (medium exchange)
- 1 interval = 5 min rest + 5 min pump



Pumping Interval

- **Rest phase:** cells start to metabolize (pH increases)
- **Pump phase:** re-calibration of pH



→ pH change (slope) during rest phase as indirect measure for the **extracellular acidification rate (EAR)**

Speed-Up of the Existing Data Preparation Script “Bubble”

Steps of data preparation:

- Reads the .exp input file
- Performs formatting & restructuring steps, joins the different data
- Outputs one table per IMOLA

Implemented the routine again from scratch:

- Exploitation of R functionalities (lapply)
- Faster reading (readr)
- Parallelization of independent operations (parLapply, foreach)
- Profiling the code (profvis)

Script	Runtime
Bubble (old)	1,800 s (30 min)
Bubble (new)	180 s (3 min)
Bubble_Cluster	60 s (1 min)
Bubble_Online	6 s

10x times faster with 2 cores
30x times faster with 6 cores

Normalizing the Data for Validation

Problem:

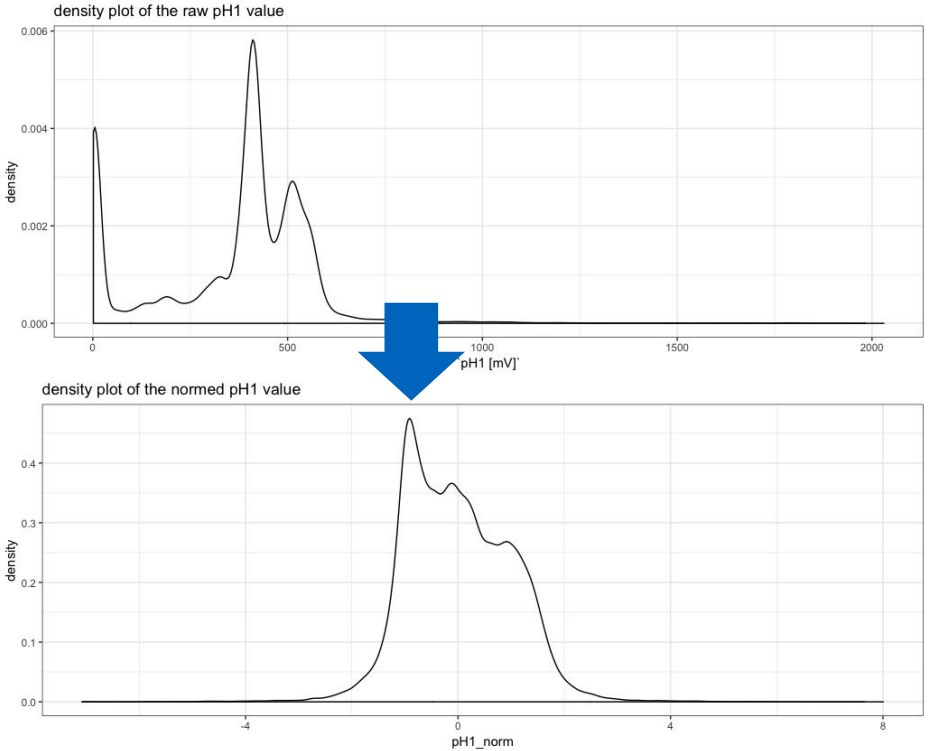
- Sensor are not calibrated and reveal drifting behavior → data not suitable for comparison

Solution:

- Normalize the data X of each interval to

$$\frac{X - \mu}{\sigma}$$

where
μ = mean of one interval
σ = standard deviation of one interval



Check for Dummy Chips and Open Circuits

- Used to **test the system's electronic components**
- Original cell culture and sensors are **replaced** by a Dummy Bio-Chip or **removed completely** (Open Circuit)
- For both cases **expected ranges** for the sensor readings are known

→ **Check** whether an IMOLA is Open Circuit, contains a Dummy Bio-Chip or real cell culture

Sensor	Dummy Bio-Chip Range	Open Circuit Range
pH	300 mV +/- 30 mV	0mV +/- 30 mV
Temperature	1500 mV +/- 150 mV	0 mV +/- 150 mV
O_2	1850 mV +/- 185 mV	2075 +/- 185 mV
Impedance_real	130 Ω +/- 13 Ω	0 Ω +/- 5 Ω
Impedance_imag.	-35 Ω +/- 5 Ω	0 Ω +/- 5 Ω

Agenda

1	Introduction
2	Data Collection & Pre-Processing
3	**Data Analysis & Results**
4	Summary & Conclusion

Agenda

Agenda

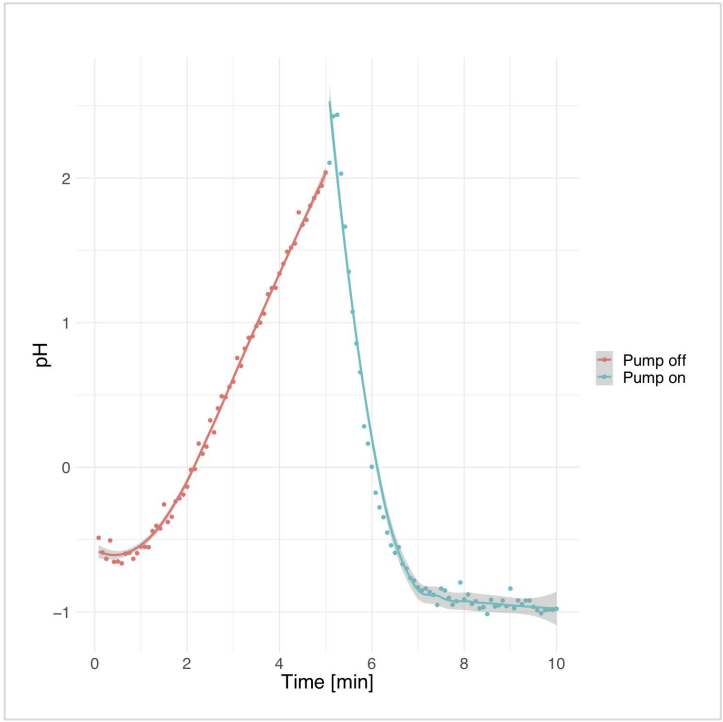
Criterion Based Validation

Why validation?

- Only curves of pH values that follow a valid curve pattern depict normal cellular metabolism
- Interpretations about cellular activity can be made only from valid pH curves

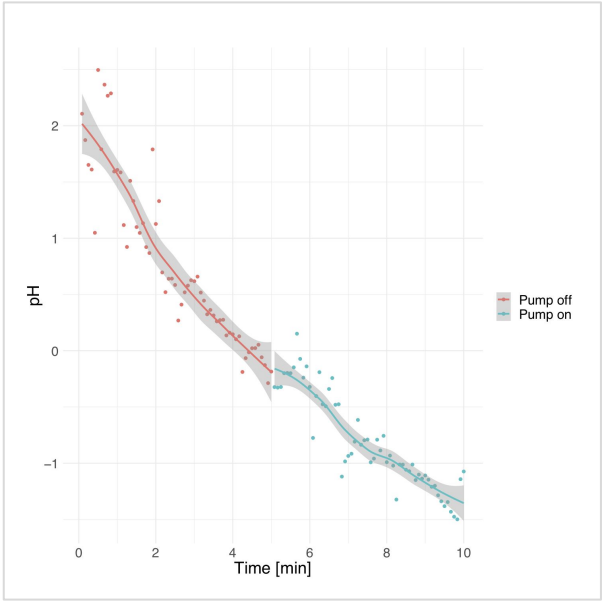
What is a good / valid pH curve?

- Curves which resemble a shark fin structure as shown in figure
- pump-off phase (rest): increasing pH value
- pump-on phase: first decreasing and then constant pH value

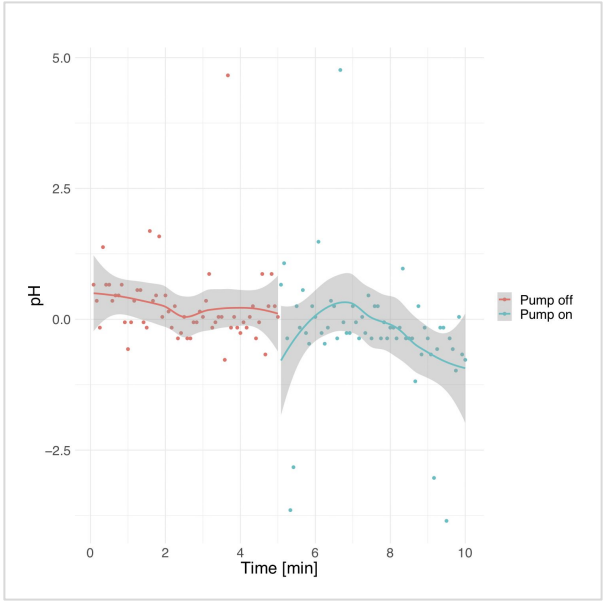


Criterion Based Validation: Invalid Intervals

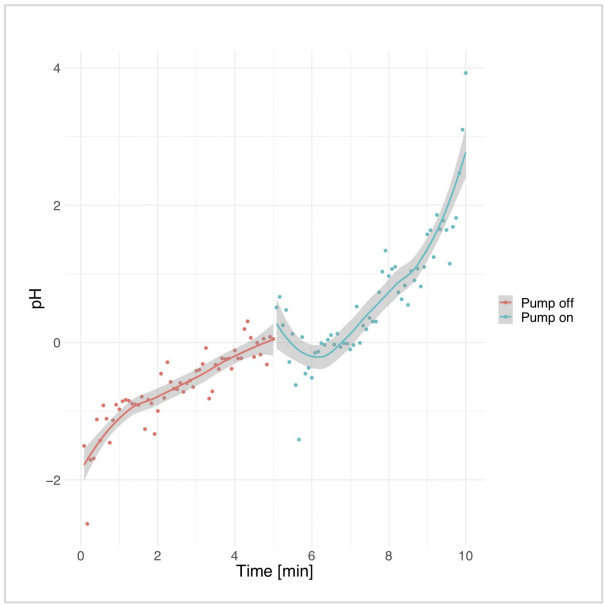
Negative pH slope in pump-off phase



High loss (MSE) in pump-on phase



Mean pH of pump-on phase higher than corresponding 75% quantile of pump-off phase



Input Data and Expected Results

What data do we have ?

- Data from two 24h experiments: Experiment_1 & Experiment_2
- Each experiment has 6 IMOLAs

What are the expected results?

Experiment	Valid	Invalid
1	IMOLA 1 IMOLA 2 IMOLA 3 IMOLA 4	IMOLA 5 IMOLA 6
2	IMOLA 2 IMOLA 3 IMOLA 5	IMOLA 1 IMOLA 4 IMOLA 6

Results of the Criterion Based Validation

Experiment_1

IMOLA	High MSE	Negative Rest	Mean Pump	Valid IMOLA [%]
1	4	8	6	88.811189
2	3	16	34	74.825175
3	5	3	10	90.909091
4	3	4	29	76.923077
5	16	87	64	4.195804
6	47	72	32	18.881119

Experiment_2

IMOLA	High MSE	Negative Rest	Mean Pump	Valid IMOLA [%]
1	22	39	28	44.755245
2	2	2	4	95.804196
3	1	4	4	94.405594
4	58	94	2	24.475524
5	2	2	4	95.804196
6	8	120	4	3.496503

If a 50% benchmark is set for number of valid intervals for an IMOLA to be valid:

- IMOLA 1,2,3, and 4 are valid from Experiment_1
- IMOLA 2,3 and 5 are valid from Experiment_2

Results from criterion based validation matches expected results!

Expected Results		
Exp	**Valid**	**Invalid**
1	IMOLA 1,2,3,4	IMOLA 5,6
2	IMOLA 2,3,5	IMOLA 1,4,6

Agenda

Validation Using Clustering of Functional Data

What is 'Functional Data Analysis (FDA)'?

- 'FDA' deals with the analysis of curves or functions
- Curves are estimated from data as being linear combinations of basis functions (with the assumption that they are intrinsically smooth)

What is 'Clustering'?

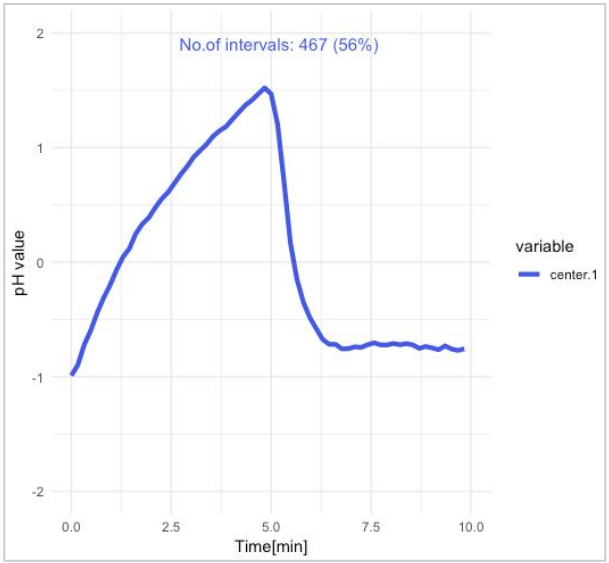
- 'Clustering' is a technique that groups together a set of data objects, such that the objects within the same cluster are more similar (w.r.t. a distance metric) than to those objects in other clusters

How did we use 'FDA' and 'Clustering'?

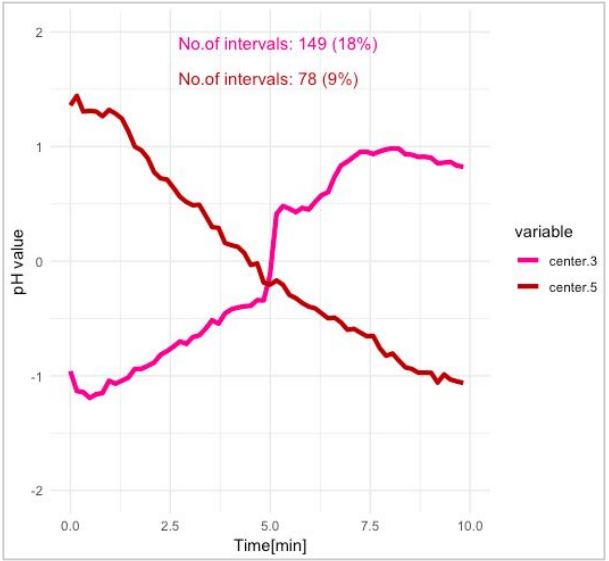
- We aim to do validation, i.e classifying intervals as being (in)valid based on the shape of the pH curves
- We estimate the functional data (curves) using splines which are polynomial curves
- We use a clustering algorithm to find similar functional data (similar curves)
- We validate the intervals using the cluster patterns observed

Observed Cluster Patterns

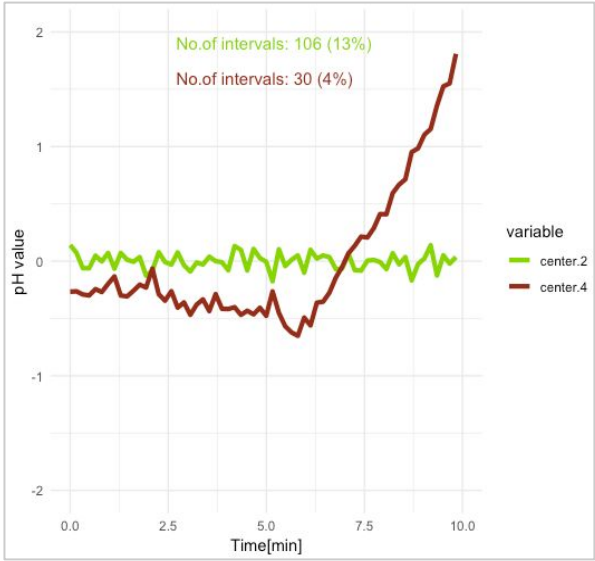
Valid pH curve



Invalid pH curves (increasing & decreasing pH)



Invalid pH curves (constant pH & pH pump off < pH pump on)



Results for Experiment_1

IMOLA	1	2	3	4	5	Valid IMOLA [%]
1	99	3	5	1	7	86.086957
2	107	1	29	5	1	74.825175
3	129	2	7	3	2	90.209790
4	111	0	29	2	1	77.622378
5	5	27	58	7	46	3.496503
6	16	73	21	12	21	11.188811

- Intervals belonging to cluster 1 are considered valid
- If a 50% benchmark is set for number of valid intervals for an IMOLA to be valid:

  → IMOLA 1, 2, 3 and 4 are valid

Results for Experiment_1

IMOLA	1	2	3	4	5	Valid IMOLA [%]
1	99	3	5	1	7	86.086957
2	107	1	29	5	1	74.825175
3	129	2	7	3	2	90.209790
4	111	0	29	2	1	77.622378
5	5	27	58	7	46	3.496503
6	16	73	21	12	21	11.188811

- Intervals belonging to cluster 1 are considered valid
- If a 50% benchmark is set for number of valid intervals for an IMOLA to be valid:

  → IMOLA 1, 2, 3 and 4 are valid

Results from validation based on clustering matches the expected results!

Expected Results		
Exp	**Valid**	**Invalid**
1	IMOLA 1,2,3,4	IMOLA 5,6

FDA: Decision Making for New Data

What do we have?

- Pre-defined cluster patterns identified from data of Experiment_1 that represents commonly observed pH curve patterns

What do we want?

- For new data from other experiments, we need to validate the intervals

What do we do?

- Estimate the curves from the new data
- For each curve, find the closest pre-defined cluster and assign it to that cluster
- Classify the curve as being valid/invalid according to the cluster it is assigned to

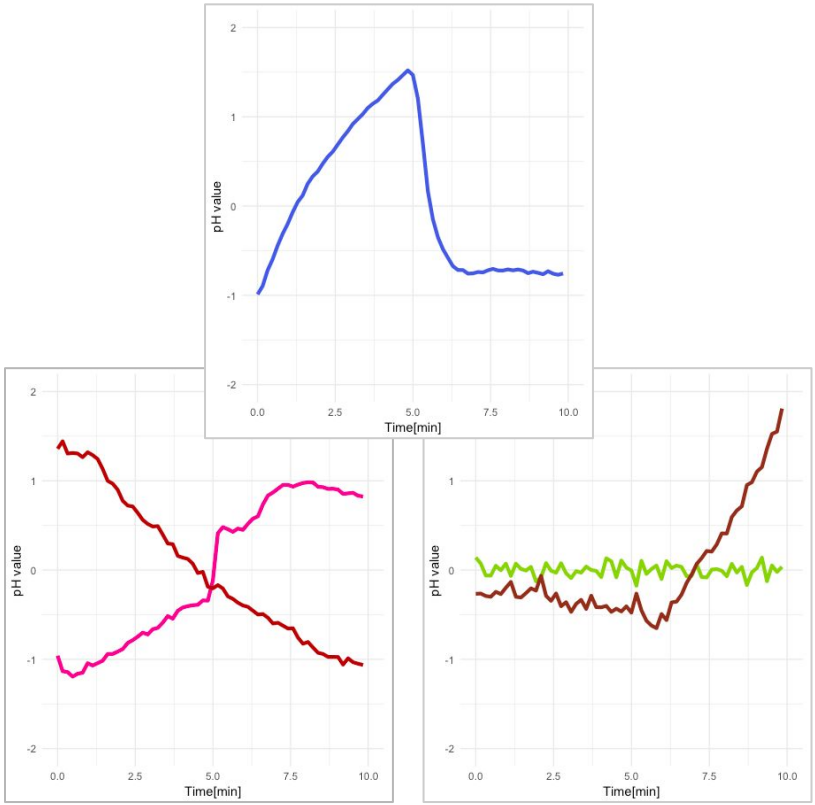


Figure: Pre-defined Cluster Patterns

Results for Experiment_2

IMOLA	1	2	3	4	5	Valid IMOLA [%]
1	42	31	20	7	37	30.656934
2	132	3	5	2	1	92.307692
3	123	14	1	1	4	86.013986
4	11	124	0	5	3	7.692308
5	134	2	3	1	3	93.706294
6	0	15	3	4	109	0.000000

- Intervals belonging to cluster 1 are considered valid
- If a 50% benchmark is set for number of valid intervals for an IMOLA to be valid:

  → IMOLA 2,3 and 5 are valid

Results for Experiment_2

IMOLA	1	2	3	4	5	Valid IMOLA [%]
1	42	31	20	7	37	30.656934
2	132	3	5	2	1	92.307692
3	123	14	1	1	4	86.013986
4	11	124	0	5	3	7.692308
5	134	2	3	1	3	93.706294
6	0	15	3	4	109	0.000000

- Intervals belonging to cluster 1 are considered valid
- If a 50% benchmark is set for number of valid intervals for an IMOLA to be valid:

  → IMOLA 2,3 and 5 are valid

Results from validation based on clustering matches the expected results!

Expected Results		
Exp	**Valid**	**Invalid**
2	**IMOLA 2,3,5**	**IMOLA 1,4,6**

Comparison of Both Validation Techniques

Experiment_2

Expected Results		
Exp	Valid	Invalid
2	IMOLA 2,3,5	IMOLA 1,4,6

Criterion Based Validation			
IMOLA	Invalid	Valid	Valid IMOLA [%]
1	75	62	45 %
2	6	137	96 %
3	8	135	94 %
4	108	35	24 %
5	6	137	96 %
6	126	5	3%

Validation Based on FDA			
IMOLA	Invalid	Valid	Valid IMOLA [%]
1	95	42	31 %
2	11	132	92 %
3	20	123	86 %
4	132	11	8 %
5	9	134	94 %
6	131	0	0 %

Both validation techniques show promising results!

Agenda

Fourier Transformation

Idea of Fourier Transformation:

- Convert a signal x from its original domain (in our case: time in sec to a frequency domain and vice versa

$$X_k = \sum_{t=0}^{N-1} x_t \exp\left(-i \frac{2\pi}{N} tk \right)$$

- Amplitude: Modulus(X_k)
  Phase: Argument(X_k)

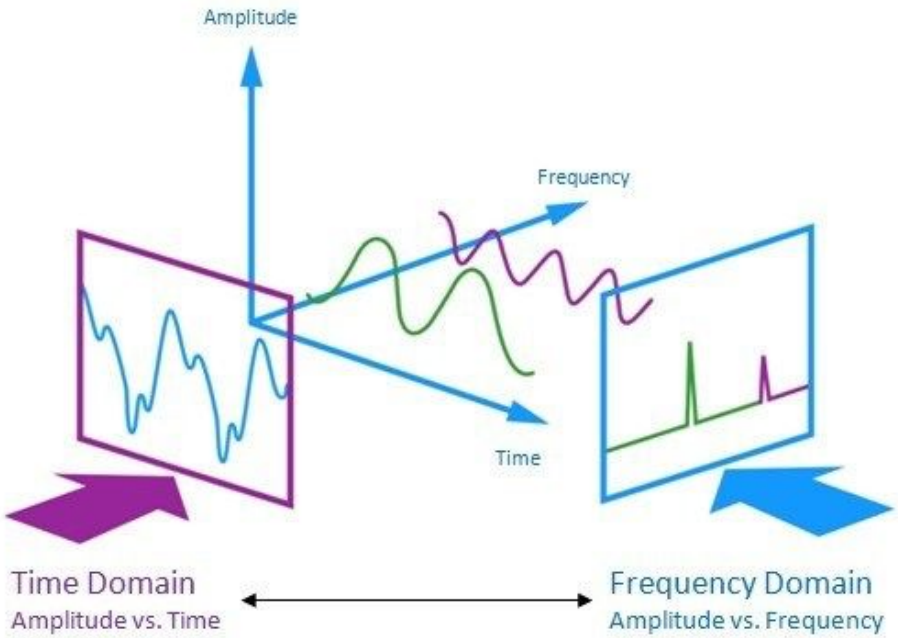
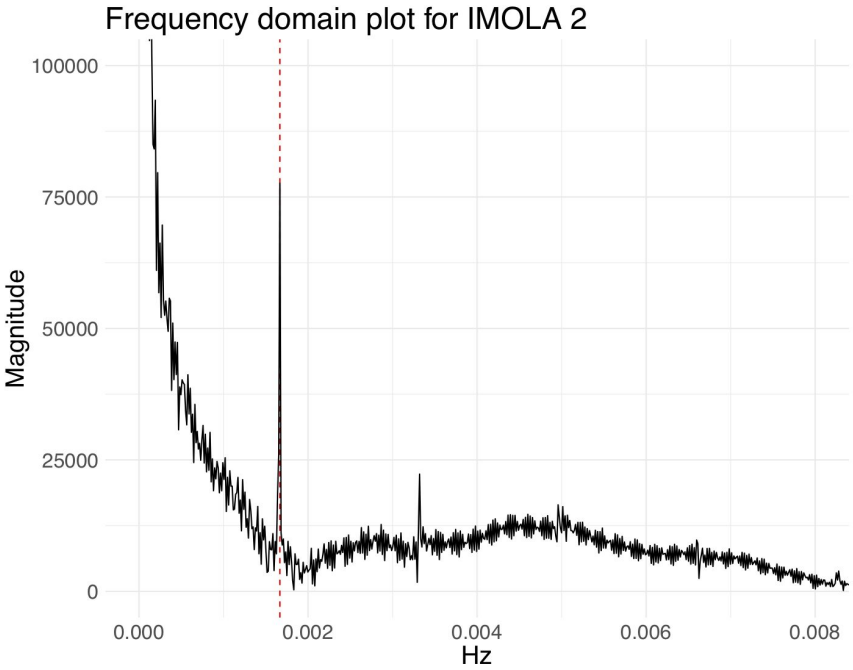


Figure: Idea of Fourier Transformation (Source)

Fourier Transformation



- Plot the **frequency domain** of a valid IMOLA to see if there is a periodic behaviour in the data
- A peak is observed at 0.00167 Hz (corresponds to **fluid cycle frequency** of 1/600s)
- All important microphysiometric information must be **stored in the frequency domain** of the fluid cycle frequency

Filter around the pump cycle frequency to eliminate the noise

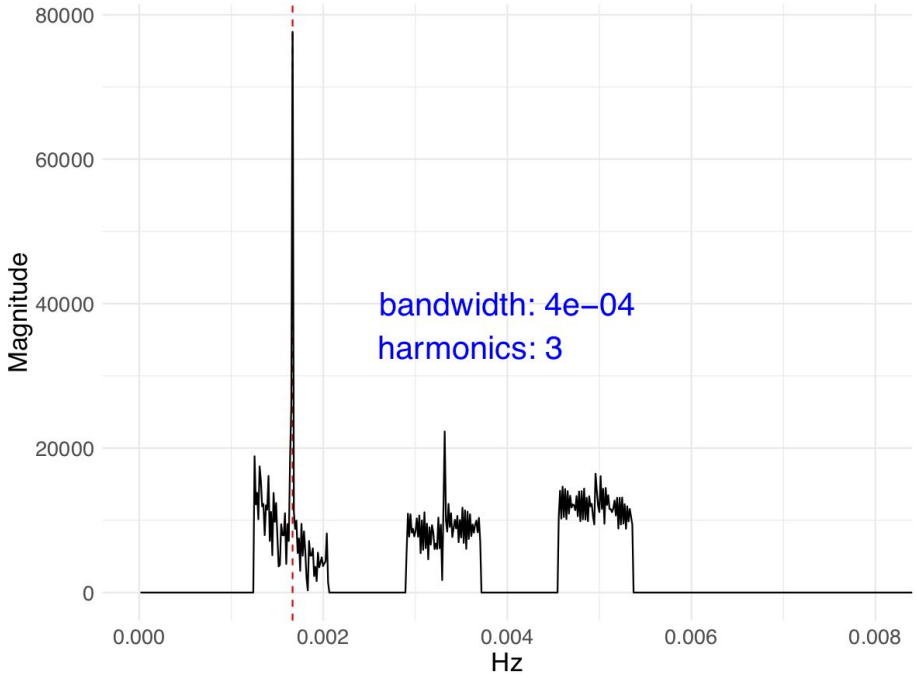
Fourier Filtering

Idea of Fourier Filtering:

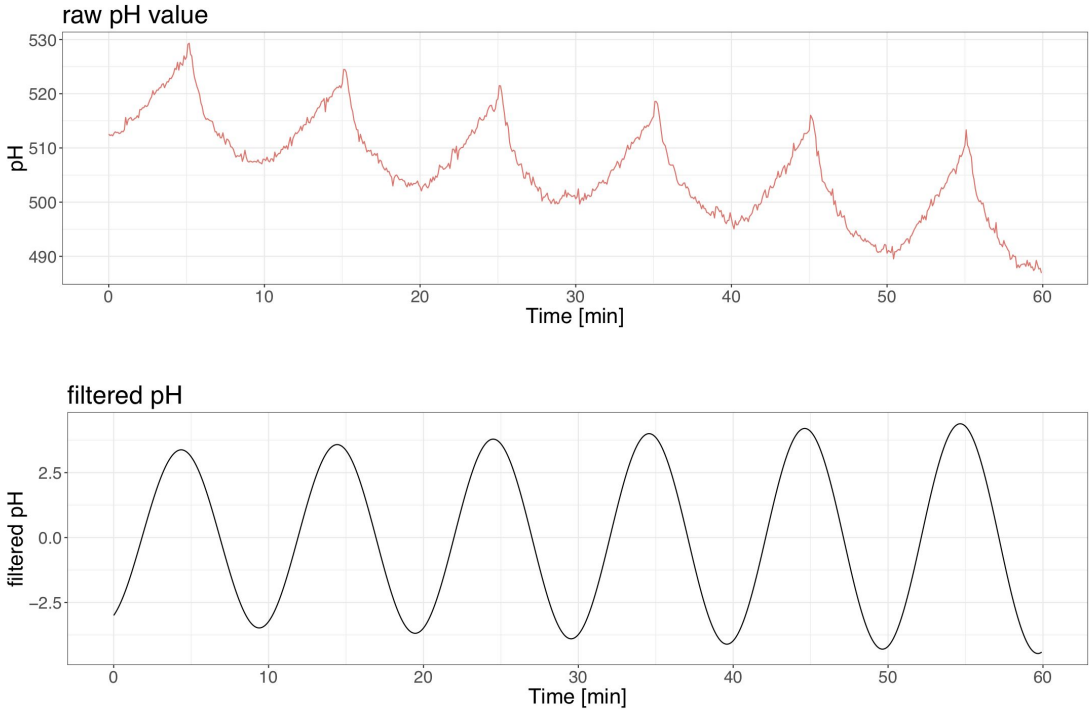
- Set the unwanted frequencies to zero
- Apply the inverse Fourier transformation to the filtered values:

$$x_t = \frac{1}{N} \sum_{k=0}^{N-1} F(X_k) \exp\left(i \frac{2\pi}{N} tk\right)$$

  where F is a filter that filters only the required frequencies
- Calculate the real part of x_t to get the filtered pH values



Fourier Filtering: Example



Evaluation of the Filters

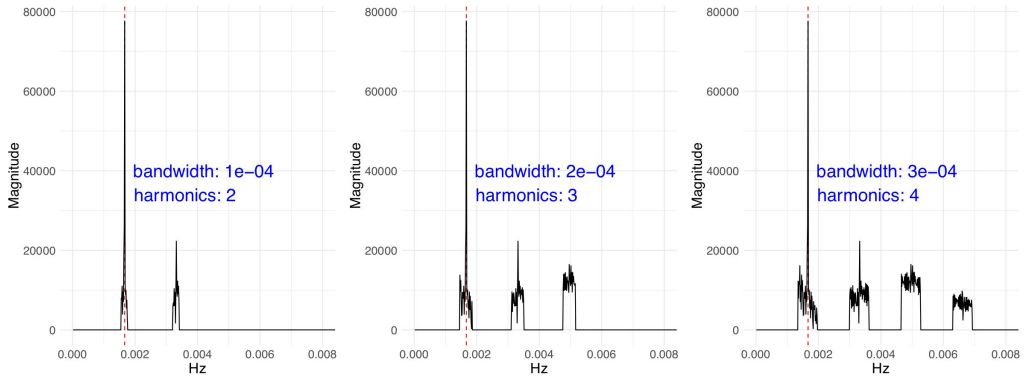
Idea of evaluation:

- Calculate the slope for each interval **before** filtering and **after** filtering
- **Normalize** the slopes to get comparable values
- Calculate the **difference** between the two slopes to see how much the filter affects the slope

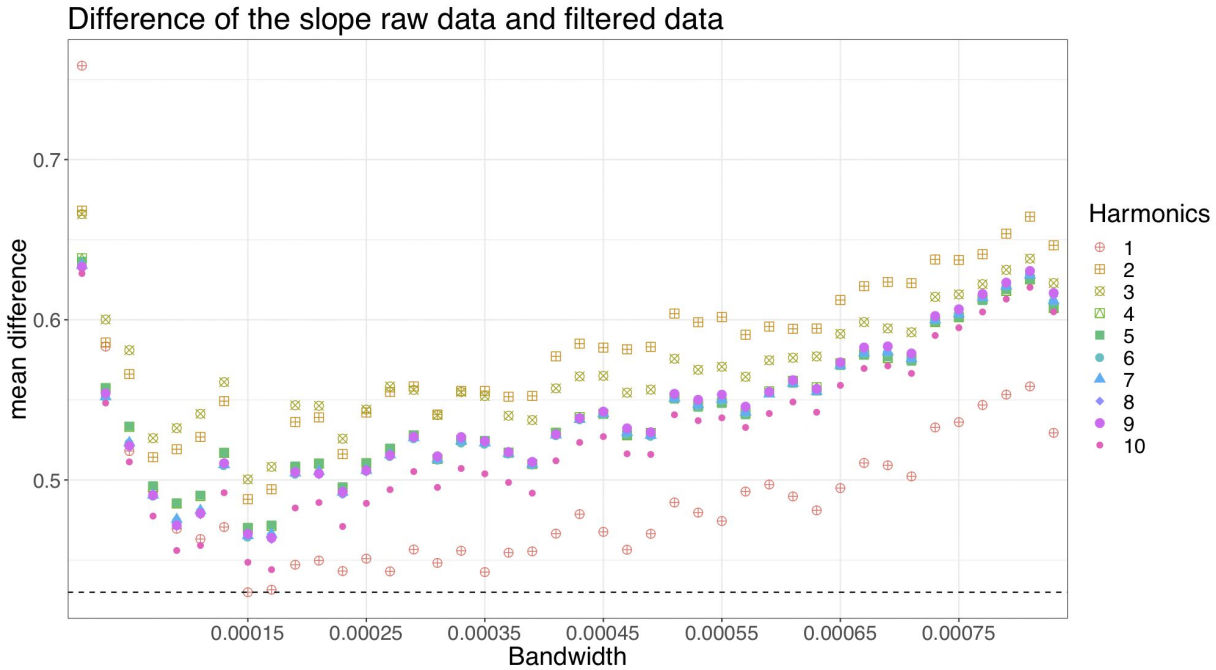
Best case: slope doesn't change
(no frequency corresponding to the cell activity was filtered out)

Varying the bandwidth and number of harmonics:

- Vary the number of harmonics from 1 to 10 and the bandwidth from 0.01 mHz to 0.83 mHz
  → 420 combinations



Evaluation of the Filters



We can conclude that the best results are obtained when using:

- **Harmonics = 1** and
- **Bandwidth = 0.15 mHz**

Agenda

1	Introduction
2	Data Collection & Pre-Processing
3	Data Analysis & Results
4	**Summary & Conclusion**

Disregarded Validation Criteria

Apart from the pH data, we also analysed the following data:

- **Air bubble detections**
  → Either too many or too few detections

Disregarded Validation Criteria

Apart from the pH data, we also analysed the following data:

- **Air bubble detections**
  → Either too many or too few detections

- **Sensor ranges**
  → No clear indication on impact on validity

Disregarded Validation Criteria

Apart from the pH data, we also analysed the following data:

- **Air bubble detections**
  → Either too many or too few detections

- **Sensor ranges**
  → No clear indication on impact on validity

- **Impedance data**
  → Recordings of impedance sensors did not fully match the expected behavior

Disregarded Validation Criteria

Apart from the pH data, we also analysed the following data:

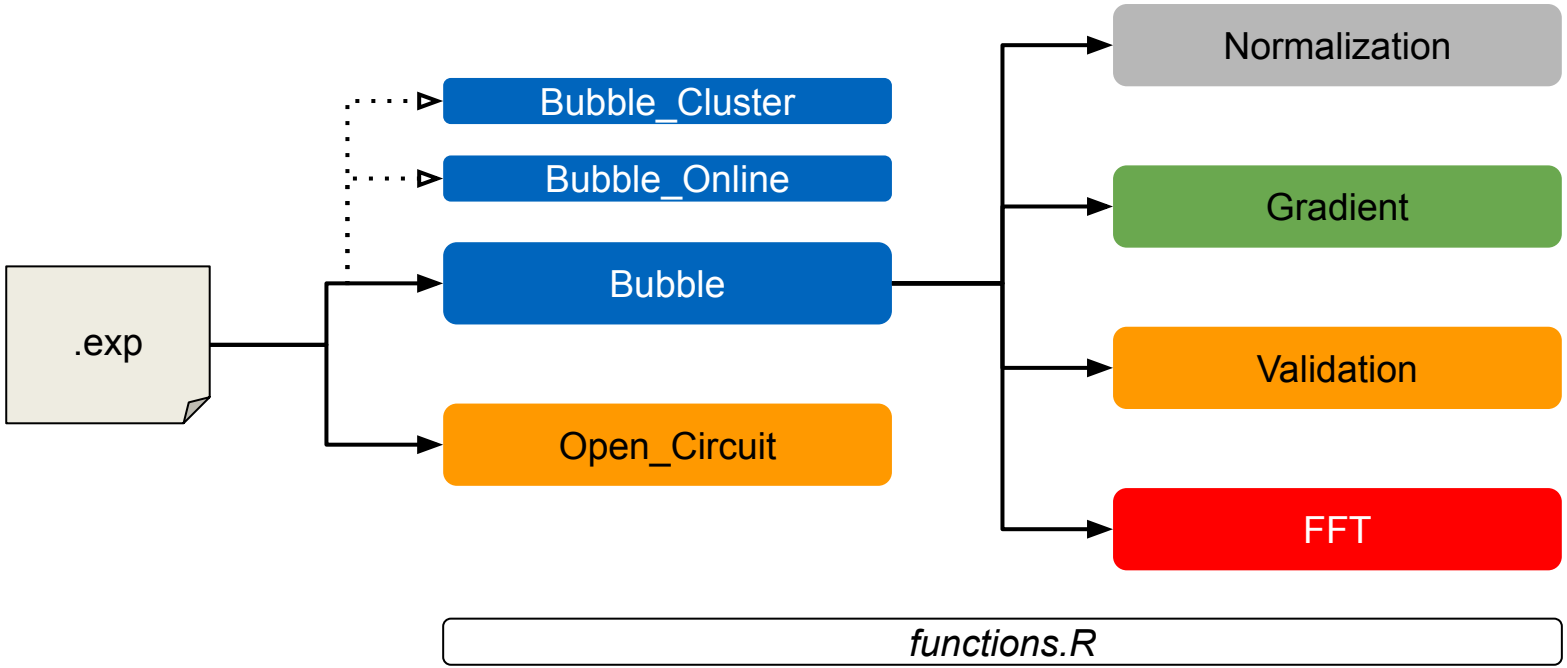
- **Air bubble detections**
  → Either too many or too few detections

- **Sensor ranges**
  → No clear indication on impact on validity

- **Impedance data**
  → Recordings of impedance sensors did not fully match the expected behavior

- **Temperature data**
  → Temperature data did not reveal further indication on cell activity

Summary of the Scripts' Workflow



Project Goals

- **Optimize existing approaches** of preparation & analysis
- Develop methods to **reduce the noise** in the data
- Develop methods to **assess the validity** of the data

Integrate those methods into cellasys' software environment DALiA

cellasys

Project Goals

Optimize existing approaches of preparation & analysis

- Parallelization accelerated data preparation by factor 10
- Provided cluster version to speed-up by factor 30
- Developed tools to assess chip type (Open Circuit, Dummy)

Develop methods to **reduce the noise** in the data

Develop methods to **assess the validity** of the data

Integrate those methods into cellasys' software environment DALiA

cellasys

Project Goals

Optimize existing approaches of preparation & analysis
- Parallelization accelerated data preparation by factor 10
- Provided cluster version to speed-up by factor 30
- Developed tools to assess chip type (Open Circuit, Dummy)

Develop methods to **reduce the noise** in the data
- Fast Fourier Transformation and Filtering
- Reduced unwanted noise while keeping signal from cells

Develop methods to **assess the validity** of the data

Integrate those methods into cellasys' software environment DALiA

Project Goals

Optimize existing approaches of preparation & analysis
- Parallelization accelerated data preparation by factor 10
- Provided cluster version to speed-up by factor 30
- Developed tools to assess chip type (Open Circuit, Dummy)

Develop methods to **reduce the noise** in the data
- Fast Fourier Transformation and Filtering
- Reduced unwanted noise while keeping signal from cells

Develop methods to **assess the validity** of the data
- Criterion and FDA-Based Validation
- Results match expectation extremely well

Integrate those methods into cellasys' software environment DALiA

Project Goals

Optimize existing approaches of preparation & analysis
- Parallelization accelerated data preparation by factor 10
- Provided cluster version to speed-up by factor 30
- Developed tools to assess chip type (Open Circuit, Dummy)

Develop methods to **reduce the noise** in the data
- Fast Fourier Transformation and Filtering
- Reduced unwanted noise while keeping signal from cells

Develop methods to **assess the validity** of the data
- Criterion and FDA-Based Validation
- Results match expectation extremely well

Integrate those methods into cellasys' software environment DALiA

- Ensured that all scripts can be run from RStudio and DALiA Analytics

TUM Data Innovation Lab with cellasys

How to Handle Data from Living Cells

Anne Christopher, Magdalena Eberl, Sebastian Zett

Munich, August 06, 2019