CRC 1456
MATHEMATICS
OF EXPERIMENT

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

DFG

Reproducible and accessible research within the CRC 1456

Goals of and experiences from the infrastructure project of the CRC 1456 ”Mathematics of Experiment”

The Infrastructure team of the CRC:
Christoph Lehrenfeld, Martin Uecker, Markus Osterhoff,
Christian Holme, Christoph Rügge
(supported by the GWDG and the Göttingen eResearch Alliance)

DMV Jahrestagung (Ready for MaRDI, am I a digital mathematician?), September 27, 2021

CRC 1456
MATHEMATICS
OF EXPERIMENT

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

DFG

Disclaimer:

- This talk does not introduce sophisticated new concepts

Reproducible and accessible research within the CRC 1456

Goals of and experiences from the infrastructure project of the CRC 1456 "Mathematics of Experiment"

The Infrastructure team of the CRC:
Christoph Lehrenfeld, Martin Uecker, Markus Osterhoff,
Christian Holme, Christoph Rügge
(supported by the GWDG and the Göttingen eResearch Alliance)

DMV Jahrestagung (Ready for MaRDI, am I a digital mathematician?), September 27, 2021

Disclaimer:

- This talk does not introduce sophisticated new concepts
- focus is on application of existing (good) ideas

Reproducible and accessible research within the CRC 1456

Goals of and experiences from the infrastructure project of the CRC 1456 "Mathematics of Experiment"

The Infrastructure team of the CRC:
Christoph Lehrenfeld, Martin Uecker, Markus Osterhoff,
Christian Holme, Christoph Rügge
(supported by the GWDG and the Göttingen eResearch Alliance)

DMV Jahrestagung (Ready for MaRDI, am I a digital mathematician?), September 27, 2021

CRC 1456
MATHEMATICS
OF EXPERIMENT

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

DFG

Disclaimer:

- This talk does not introduce sophisticated new concepts
- focus is on application of existing (good) ideas
- mostly presentation of plans

Reproducible and accessible research within the CRC 1456

Goals of and experiences from the infrastructure project of the CRC 1456 "Mathematics of Experiment"

The infrastructure team of the CRC:
Christoph Lehrenfeld, Martin Uecker, Markus Osterhoff,
Christian Holme, Christoph Rügge
(supported by the GWDG and the Göttingen eResearch Alliance)

DMV Jahrestagung (Ready for MaRDI, am I a digital mathematician?), September 27, 2021

CRC 1456
MATHEMATICS
OF EXPERIMENT

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

DFG

Disclaimer:

- This talk does not introduce sophisticated new concepts
- focus is on application of existing (good) ideas
- mostly presentation of plans
- happy for every feedback / ideas / interactions

Reproducible and accessible research within the CRC 1456

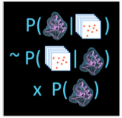
Goals of and experiences from the infrastructure project of the CRC 1456 "Mathematics of Experiment"

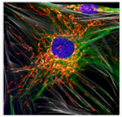
The infrastructure team of the CRC:
Christoph Lehrenfeld, Martin Uecker, Markus Osterhoff,
Christian Hohne, Christoph Ruegge
(supported by the GWDG and the Göttingen eResearch Alliance)

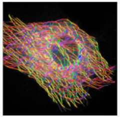
DMV Jahrestagung (Ready for MaRDI, am I a digital mathematician?), September 27, 2021

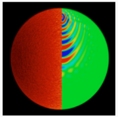
The CRC 1456 "Mathematics of Experiment" (DFG-funded)

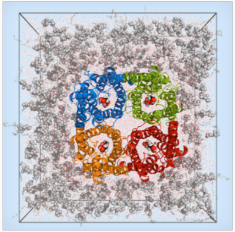
- Speaker: Thorsten Hohage
- 17 projects with 28 PIs
- each project pairs scientists from math and natural sciences

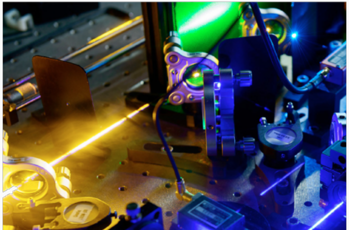


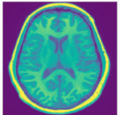








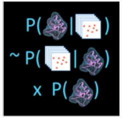


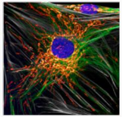


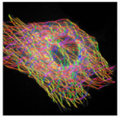
The CRC 1456 ”Mathematics of Experiment” (DFG-funded)

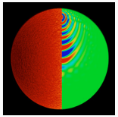
- Motivation: experimental data are increasingly indirect, noisy measurements

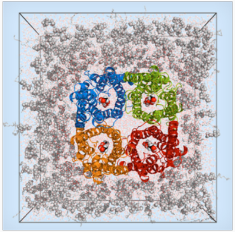
- Speaker: Thorsten Hohage
- 17 projects with 28 PIs
- each project pairs scientists from math and natural sciences

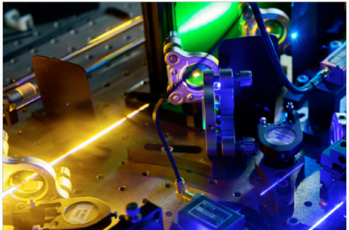


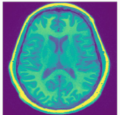








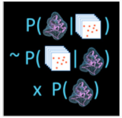


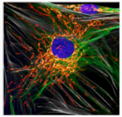


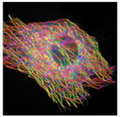
The CRC 1456 "Mathematics of Experiment" (DFG-funded)

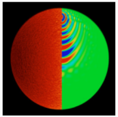
- Motivation: experimental data are increasingly indirect, noisy measurements
- Challenges: (geom.) nonlinearities, incomplete information, complex dependency structures...

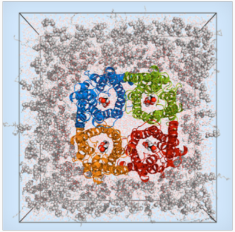
- Speaker: Thorsten Hohage
- 17 projects with 28 PIs
- each project pairs scientists from math and natural sciences

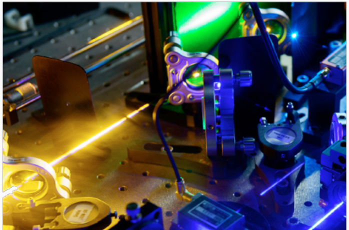


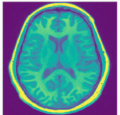








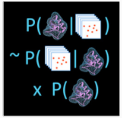


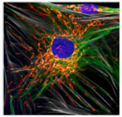


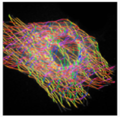
The CRC 1456 "Mathematics of Experiment" (DFG-funded)

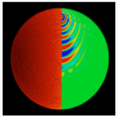
- Motivation: experimental data are increasingly indirect, noisy measurements
- Challenges: (geom.) nonlinearities, incomplete information, complex dependency structures...
- Bottleneck: extracting quantitative information from large data sets.

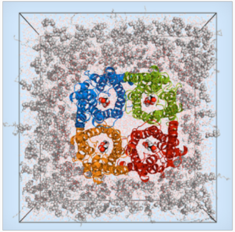
- Speaker: Thorsten Hohage
- 17 projects with 28 PIs
- each project pairs scientists from math and natural sciences

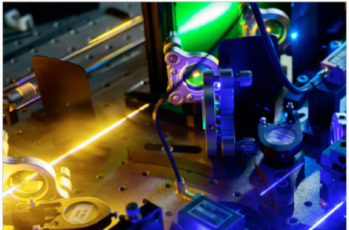


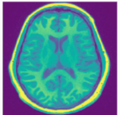








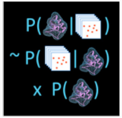


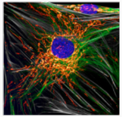


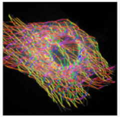
The CRC 1456 "Mathematics of Experiment" (DFG-funded)

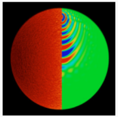
- Motivation: experimental data are increasingly indirect, noisy measurements
- Challenges: (geom.) nonlinearities, incomplete information, complex dependency structures...
- Bottleneck: extracting quantitative information from large data sets.
- Goal: develop mathematical theory and tools to extract maximal quantitative information from experimental data

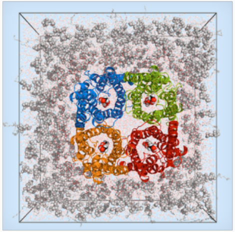
- Speaker: Thorsten Hohage
- 17 projects with 28 PIs
- each project pairs scientists from math and natural sciences

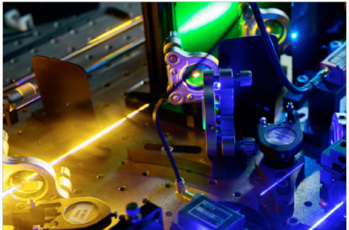


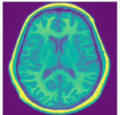












Diversity in research data (data and algorithms/methods)

Different types of data sources:

- X-ray tomography
- molecular dynamics simulations
- MRI scans
- Dopplergrams
- ...

Different types of algorithms:

- Bayesian optimization (MCMC)
- Optimal transport
- Inverse Problems
- Numerics of PDEs (FEM)
- ...

heterogeneous environments:

different communities, different software frameworks, different data formats, different data repositories, different scienfitic culture, ..

Example: Software within the CRC

Open source software of CRC members

- Berkeley Adv. [C] Recon. Toolbox: MRI imaging (M. Uecker)
- Netgen/NGSolve [C++/python]: FEM (C. Lehrenfeld)
- ProxToolbox [python]: nonlinear optimization (R. Luke)
- GROMACS [C++] : molecular dynamics (H. Grubmüller and B. L. de Groot)
- transport [R]: optimal transport (D. Schuhmacher)
- otfinference [R]: inference for optimal transport (A. Munk)
- MultiScaleOT [C++/python]: numerical optimal transport (B. Schmitzer)
- FDRSeg [R]: step function estimation (H. Li)
- DataJoint [matlab] : framework for scientific databases and data pipelines (A. Ecker)
- HoloHomoToolbox [matlab] : toolbox for holographic tomography (T. Salditt)
- ISD [python]: Bayesian modeling of biomolecular structures (M. Habeck)
- ...

Goals of the INF project

1. Support CRC members with Reproducible Research (highest priority!)

2. Facilitate software collaboration between several projects:
   - algorithmic interfaces
     (e.g. couple optimization solver of group X with a forward solver of group Y)
   - data exchange
     (e.g. apply algorithm from group X on data from group Y)

3. Flexible exchange of data sets and algorithms from one (interactive) platform

Support with Reproducible Research

CRC commitment for Reproducible Research

- Open Source:
  Source code for all methods shall be published with an open source license.
- Open Data (FAIR):
  Datasets obtained shall be made accessible and reusable.
- Reproducible Research:
  Publications shall be published alongside everything necessary for reproduction.
  [data, source code, description meta data, dependency description, containers, ... details: see e.g. Max Horn's talk]

Measures for Reproducible Research I

- Two-day training workshops on general principles of reproducible research (more to come, especially domain-specific workshops)
  ⇝ raising the awareness for the importance of reproducible research and available techniques and tools
- Data (and software) policies of the CRC (decided on in June 2021)
  ⇝ minimal standards for data quality, documentation, accessibility, persistence[1]

[1] `https://www.uni-goettingen.de/en/647064.html`

Measures for Reproducible Research II

- Setup of fallback data repository for long-term storage with citable identifiers (dataverse / DataCite DOI).
  ⇝ Offer persistent and accessible storage solution where needed
  (often domain-specific repositories are prefered (zenodo, mridata, ...))
- We are setting up binder-like jupyter-instances for ‘LiveDocs’
  ⇝ Better accessibility of software and storage and simpler ways to present results

Software Interaction

Interaction of algorithms / data sets within the CRC i

We have expertise and developments in the following 4 categories:

(i) Measurement data u^{obs}

(ii) Non-linear inverse problems / optimization

$$\text{Find } c \text{ s.t. } \left\| F(c) - u^{\text{obs}} \right\| \to \min!$$

(iii) Forward problem F

Given c, evaluate $F(c)$, e.g. as the solution of a PDE, ...

(iv) Bayesian algorithms

Given measured data u estimate uncertainty of reconstructions c

$$P(c|u) \propto P(u|c)P(c)$$

Interaction of algorithms / data sets within the CRC ii

The status:

Algorithms within the CRC are typically developed and/or tested in a narrow application range,

e.g. one algorithm from (ii) is combined with only one forw. prob. in (iii).

The aim:

Combine (some) algorithms and data sets as needed for the CRC projects (and beyond),

e.g. combine algorithm from (ii) with many/all(?) forw. prob. in (iii).
or combine forw. problem in (iii) with many/all(?) optim. solvers in (ii).[2]

[2] less general than in S. Rave's talk.

Interaction of algorithms / data sets within the CRC

[3] https://github.com/regpy/binder-ngsolve-bart

Interaction of algorithms / data sets within the CRC

Interfaces:

- Identify classes of algorithms / data
- For each class of algorithm / data type adopt common software interface

[3] `https://github.com/regpy/binder-ngsolve-bart`

Interaction of algorithms / data sets within the CRC

Interfaces:

- Identify classes of algorithms / data
- For each class of algorithm / data type adopt common software interface

$\rightsquigarrow$ **allows to test/compare methods in a larger context**

[3] `https://github.com/regpy/binder-ngsolve-bart`

Interaction of algorithms / data sets within the CRC

Interfaces:

- Identify classes of algorithms / data
- For each class of algorithm / data type adopt common software interface

$\rightsquigarrow$ **allows to test/compare methods in a larger context**

Disclaimer:

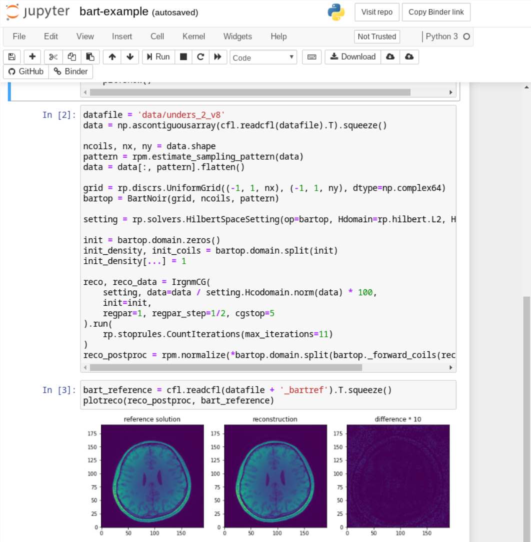
This will most certainly be only possible for a small set of involved packages.

Implemented prototypical couplings in `regpy` with `bart` (MRI reconstructions) with `NGSolve` (PDE solver) and MCMC methods.[3]

[3] https://github.com/regpy/binder-ngsolve-bart

Example of a LiveDoc with data from dataverse:

mybinder.org/v2/gh/hcmh/binder-ngsolve-bart/dataverse

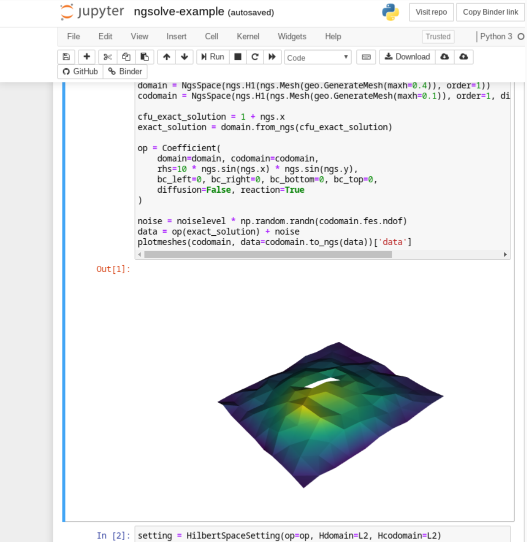


In [2]:

```
datafile = 'data/unders_2_v8'
data = np.ascontiguousarray(cfl.readcfl(datafile).T).squeeze()

ncoils, nx, ny = data.shape
pattern = rpm.estimate_sampling_pattern(data)
data = data[:, pattern].flatten()

grid = rp.discrs.UniformGrid((-1, 1, nx), (-1, 1, ny), dtype=np.complex64)
bartop = BartNoir(grid, ncoils, pattern)

setting = rp.solvers.HilbertSpaceSetting(op=bartop, Hdomain=rp.hilbert.L2, H

init = bartop.domain.zeros()
init_density, init_coils = bartop.domain.split(init)
init_density[...] = 1

reco, reco_data = IrgnmCG(
    setting, data=data / setting.Hcodomain.norm(data) * 100,
    init=init,
    regpar=1, regpar_step=1/2, cgstop=5
).run(
    rp.stoprules.CountIterations(max_iterations=11)
)
reco_postproc = rpm.normalize(*bartop.domain.split(bartop._forward_coils(rec
```

In [3]:

```
bart_reference = cfl.readcfl(datafile + '_bartref').T.squeeze()
plotreco(reco_postproc, bart_reference)
```



```
domain = NgsSpace(ngs.H1(ngs.Mesh(geo.GenerateMesh(maxh=0.4)), order=1))
codomain = NgsSpace(ngs.H1(ngs.Mesh(geo.GenerateMesh(maxh=0.1)), order=1, di

cfu_exact_solution = 1 + ngs.x
exact_solution = domain.from_ngs(cfu_exact_solution)

op = Coefficient(
    domain=domain, codomain=codomain,
    rhs=10 * ngs.sin(ngs.x) * ngs.sin(ngs.y),
    bc_left=0, bc_right=0, bc_bottom=0, bc_top=0,
    diffusion=False, reaction=True
)

noise = noiselevel * np.random.randn(codomain.fes.ndof)
data = op(exact_solution) + noise
plotmeshes(codomain, data=codomain.to_ngs(data))['data']
```

Out[1]:

In [2]:

```
setting = HilbertSpaceSetting(op=op, Hdomain=L2, Hcodomain=L2)
```

Interaction platform LiveDoc

where CRC-software/data comes together

The vision

Use flexible interfaces to exchange algorithms and data from one platform:

- Simple scripting language to define the combination of different tools:

  ```
      Forward problem W
  +   Optimization solver X
  +   Bayesian solver Y
  +   data set Z
  ```

- Access through web interface (binder / jupyter)
- Maybe triggers remote computation (server / cluster)
- Delivers result data ± visualization

The vision

Use flexible interfaces to exchange algorithms and data from one platform:

- Simple scripting language to define the combination of different tools:

```
    Forward problem W
+   Optimization solver X
+   Bayesian solver Y
+   data set Z
```

- Access through web interface (binder / jupyter)
- Maybe triggers remote computation (server / cluster)
- Delivers result data ± visualization

Two key components:

Component 1: Data catalogue

Relevant data shall be organized through a unified interface that gives

- access to shared and public domain data for CRC members
- access to public domain data for everyone
- especially access to all CRC-related data sets

The `data catalogue` is not the storage, but rather the database (+ interface).

Our current implementation:
We use the CRC1456 `dataverse` of `gro.data`[4] as the catalogue where data can be stored or links to other data repositories are stored.

[4] https://data.goettingen-research-online.de/dataverse/crc1456

Component 2: LiveDoc

Platform with access to data (through data catalogue) and algorithms allows to

Component 2: LiveDoc

Platform with access to data (through data catalogue) and algorithms allows to

- access shared/public data/algorithms for CRC members/everyone

Component 2: `LiveDoc`

Platform with access to data (through data catalogue) and algorithms allows to

- access shared/public data/algorithms for CRC members/everyone
- combine different algorithms from within CRC interactively

Component 2: LiveDoc

Platform with access to data (through data catalogue) and algorithms allows to

- access shared/public data/algorithms for CRC members/everyone
- combine different algorithms from within CRC interactively
- combine algorithms with other external tools (e.g. TensorFlow)

Component 2: `LiveDoc`

Platform with access to data (through data catalogue) and algorithms allows to

- access shared/public data/algorithms for CRC members/everyone
- combine different algorithms from within CRC interactively
- combine algorithms with other external tools (e.g. TensorFlow)
- use automated (regression) testing:
  - combinations of algorithms and data sets define regression tests
  - regression tests are triggered regularly (e.g. on software updates)

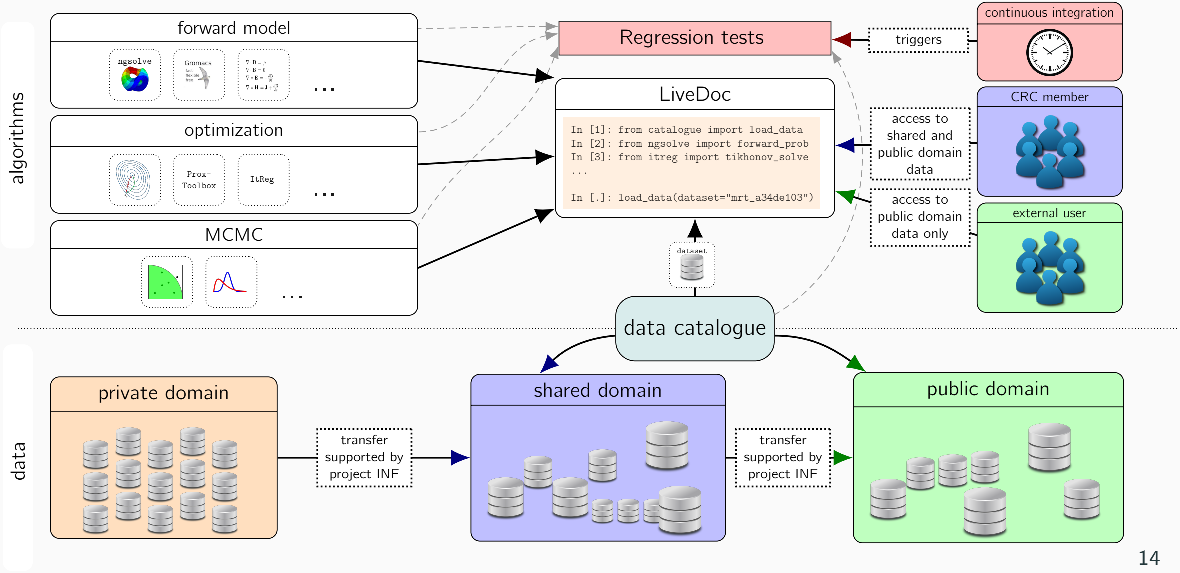
  ⇝ identifies deficiencies in the (evolving) interface design and bugs in your software

Component 2: `LiveDoc`

Platform with access to data (through data catalogue) and algorithms allows to

- access shared/public data/algorithms for CRC members/everyone
- combine different algorithms from within CRC interactively
- combine algorithms with other external tools (e.g. TensorFlow)
- use automated (regression) testing:
    - combinations of algorithms and data sets define regression tests
    - regression tests are triggered regularly (e.g. on software updates)
    - ⇝ identifies deficiencies in the (evolving) interface design and bugs in your software
- facilitates construction of academic examples for teaching (outreach):
    - use the pool of methods and data sets for teaching demos
    - combine textbook version of one algorithm with “black-box”es for the others
    - work on “realistic” show cases while being able to focus on one problem

Schematic of the LiveDoc/Data catalogue


algorithms
forward model
ngsolve
Gromacs
fast
flexible
free
$\nabla \cdot \mathbf{D} = \rho$
$\nabla \cdot \mathbf{B} = 0$
$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$
$\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t}$
...
optimization
Prox-Toolbox
ItReg
...
MCMC
...
Regression tests
triggers
continuous integration
LiveDoc
CRC member
access to shared and public domain data
external user
access to public domain data only
dataset
data catalogue
data
private domain
transfer supported by project INF
shared domain
transfer supported by project INF
public domain

```
In [1]: from catalogue import load_data
In [2]: from ngsolve import forward_prob
In [3]: from itreg import tikhonov_solve
...

In [.]: load_data(dataset="mrt_a34de103")
```

Summary

- None of the ideas or technologies is new.
- The problem is in the implementation, especially in this heterogeneous setup.
- We have a long road ahead of us:
    - Mostly only proof-of-concept realizations of the advanced stuff so far
    - Most important and most time consuming is the support for the subgroups for the "basic" things
- Personal situation is difficult: We have an open PostDoc position!

Summary

- None of the ideas or technologies is new.
- The problem is in the implementation, especially in this heterogeneous setup.
- We have a long road ahead of us:
  - Mostly only proof-of-concept realizations of the advanced stuff so far
  - Most important and most time consuming is the support for the subgroups for the "basic" things
- Personal situation is difficult: We have an open PostDoc position!

Thank you for your attention!