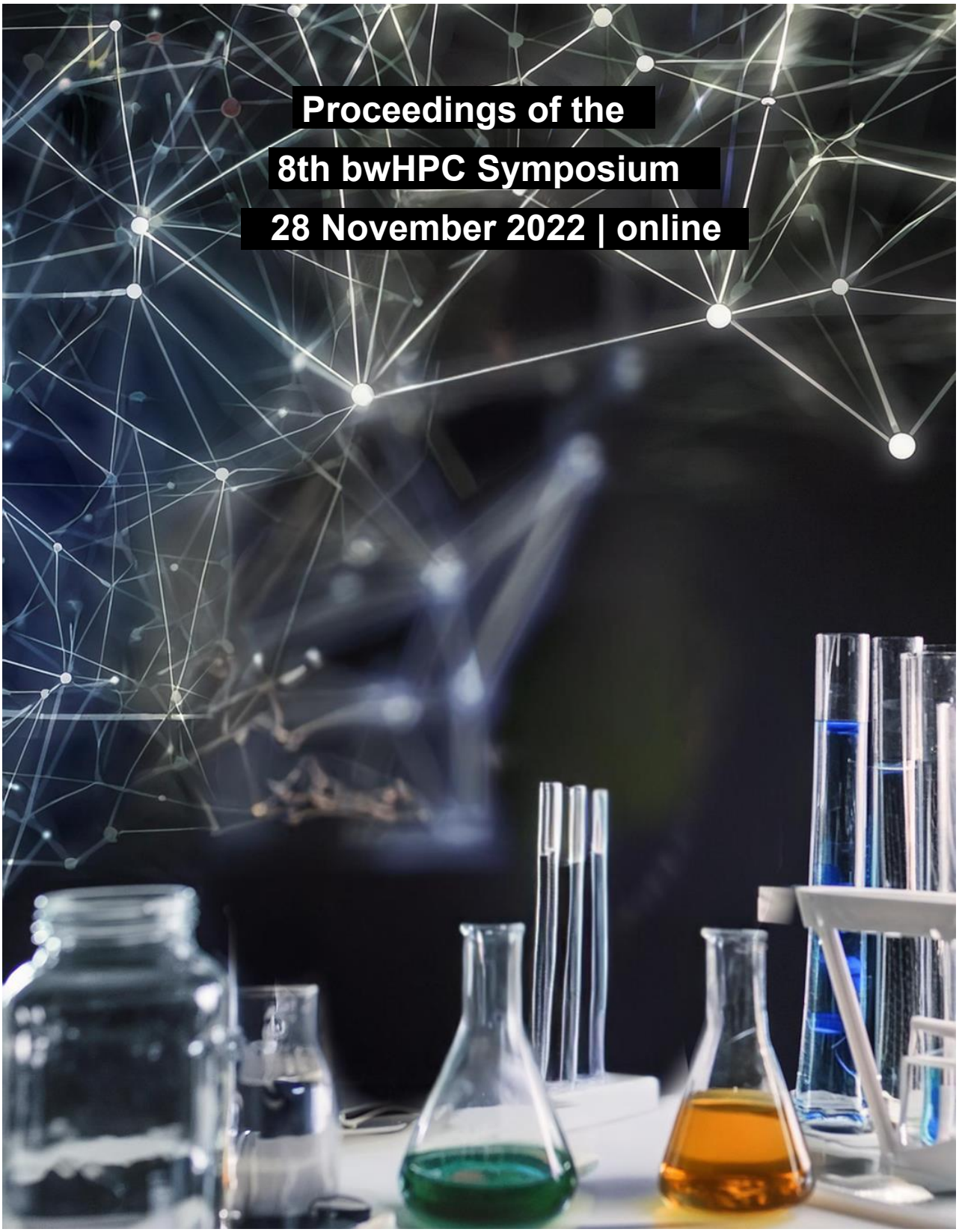
bw|HPC

HLRS

Baden-Württemberg
MINISTERIUM FÜR WISSENSCHAFT, FORSCHUNG UND KUNST



Proceedings of the 8th bwHPC Symposium

28 November 2022 | online

Cosima-Maria Weyers (editor)

Proceedings

8th bwHPC Symposium

28 November 2022

online

DOI:

Vorwort

Digitale Forschungsinfrastrukturen sind wichtige Werkzeuge, um bahnbrechende Forschungsergebnisse zu ermöglichen und damit Antworten auf gesellschaftliche Herausforderungen zu finden. Der Anspruch des Landes ist es, allen Wissenschaftlerinnen und Wissenschaftlern in Baden-Württemberg mit leistungsfähigen Forschungsinfrastrukturen optimale Rahmenbedingungen für innovative und exzellente Forschung zu bieten. An diesem Anspruch orientiert sich auch die Landesstrategie für das High Performance Computing, im Dialog mit der Spitzenforschung die notwendigen erstklassigen Infrastrukturen aufzubauen, verlässlich zu gestalten und durch geeignete Support- und Fortbildungsstrukturen zu unterstützen.

Um das gesamte Leistungsspektrum abzudecken und eine optimale Betreuung der Nutzerinnen und Nutzer sicherzustellen, leben unsere Rechenzentren eine Kultur der Zusammenarbeit. Diese Kooperationskultur ist das Ergebnis des gemeinsamen Wirkens vieler engagierter Menschen. Im Namen des Ministeriums für Wissenschaft, Forschung und Kunst danke ich allen Beteiligten für ihr außerordentliches Engagement und ihren Einsatz für die besten Lösungen.

Lassen Sie uns diesen Weg auch zukünftig gemeinsam mit landesweiten Kooperationen erfolgreich gehen. Nur durch eine vertrauensvolle Zusammenarbeit und selbstbestimmte Gestaltung können wir die Chancen der digitalen Zukunft optimal nutzen.

Wesentlich für den Erfolg ist, dass die hochschulübergreifende Zusammenarbeit auch den fortwährenden Dialog zwischen Betreibern und Nutzenden einschließt, der auf beiden Seiten das Bewusstsein für die Möglichkeiten und Bedürfnisse des anderen schärft. Für den gelebten Austausch ist das bwHPC-Symposium ein wichtiges Format. Eine starke Einbindung der Nutzerinnen und Nutzer in die Konzeption und Gestaltung der Infrastrukturdienste ist unerlässlich, um das volle Potenzial der digitalen Werkzeuge bestmöglich auszuschöpfen. Mit Blick auf die bestehenden digitalen Infrastrukturen im Land ergeben sich daraus weitergehende Anforderungen, wie z. B. die Integration von Werkzeugen zur Datenanalyse, das Management und die Archivierung von Forschungsdaten sowie die Berücksichtigung neuer methodischer Ansätze. Eine fortlaufende Weiterentwicklung der digitalen Forschungsinfrastruktur bleibt essentiell, um den geeigneten Rahmen für eine exzellente, international konkurrenzfähige Forschungslandschaft zu setzen.

Dr. Raphael Dorn
Ministerium für Wissenschaft, Forschung
und Kunst des Landes Baden-Württemberg

Grußwort zum 8. bwHPC Symposium

Das 8. bwHPC Symposium brachte erstmals die HPC Community des Landes Baden-Württemberg an das Höchstleistungsrechenzentrum der Universität Stuttgart (HLRS). Das HLRS ist ein zentraler Bestandteil des baden-württembergischen Landeskonzeptes für das Hoch- und Höchstleistungsrechnen. In seiner doppelten Rolle als Bundeshöchstleistungsrechenzentrum und als Ankerzentrum für das Land Baden-Württemberg stellt das HLRS die Schnittstelle zwischen der obersten Ebene des Höchstleistungsrechnens und dem Bereich des Hochleistungsrechnens dar. Insbesondere die Problematik der Durchlässigkeit und der Schulung stehen bei dieser Schnittstellenfunktion im Vordergrund. Aspekte des nationalen und europäischen Hoch- und Höchstleistungsrechnens und ihr Einfluss auf die wissenschaftliche Benutzergruppen waren daher ein Fokus der Veranstaltung.

Während Computersimulationen in den letzten Jahrzehnten das Hoch- und Höchstleistungsrechnen dominiert haben, sind in den letzten Jahren die Themen Large Data und Artificial Intelligence stärker in den Fokus gerückt. Beide neuen Konzepte erfordern eine flexible Reaktion der Betreiberzentren um auch in diesen Bereichen das Potential der Hard- und Software für wissenschaftliche Durchbrüche voll nutzen zu können. Im Fokus des Symposiums standen daher auch die National ForschungsDaten Infrastruktur (NFDI) sowie das Thema der Künstlichen Intelligenz.

Die Vorträge der Benutzer spiegelten im Symposium die Breite der Anwendungen der wissenschaftlichen Community in Baden-Württemberg wieder. Das Programm war zwar noch deutlich von klassischen Anwendungen geprägt doch zeigten sich spannende neue Ansätze und Beiträge zu neuen Forschungsthemen, die für das Höch- und Höchstleistungsrechnen im Land Baden-Württemberg in den kommenden Jahren sukzessive erweitern und damit verändern werden.

Prof. Michael M. Resch
Höchstleitungsrechenzentrum Stuttgart
Universität Stuttgart

Inhaltsverzeichnis

Vorträge

Using BinAC to analyze microbiome samples

Anupam Gautam[1,2] and Daniel H. Huson[1,2]

[1]Algorithms in Bioinformatics, Institute for Bioinformatics and Medical Informatics, University of Tübingen, Sand 14, 72076 Tübingen, Germany

[2]International Max Planck Research School "From Molecules to Organisms", Max Planck Institute for Biology Tübingen, Max-Planck-Ring 5, 72076 Tübingen, Germany

Abstract

In metagenomics analysis, one common approach involves aligning sequencing reads against a protein reference database and subsequently binning those reads into taxonomic and functional categories. Here we discuss the use of the DIAMOND+MEGAN pipeline for this purpose and report on work that we have done with the help of BinAC (also de.NBI) to compare the performance of DIAMOND+MEGAN using two different protein reference databases. We showed that, while NCBI-nr is more comprehensive, the AnnoTree database is a good alternative when only prokaryotic microbiome members are of interest.

1. Introduction

For metagenomics analysis, one approach is to align sequencing reads against a protein reference database and then to bin those reads into taxonomic and functional categories. This can help to answer the question of which organisms are present and what functional potential they hold. This is embodied in the DIA- MOND+MEGAN approach [1, 2], in which reads are initially aligned against the NCBI-nr protein reference database using DIAMOND [3] and then binned into taxonomic and functional categories using MEGAN [4, 5]. MEGAN provides algorithms for analysing short- and long-read reads, and allows the user to interactively explore that data using interactive plots, alignment visualizations, gene-centric assembly, and many other techniques. The current computational bottleneck of this approach is the size of the NCBI-nr database [6], which as of today contains 536,937,122 entries. Further, NCBI-nr contains protein sequences from all domains of life, while researchers working on microbiomes are often focused on the prokaryotic aspect (Bacteria and Archaea).

Because of this, we are interested in considering alternatives to the full NCBI-nr database. Hence, we explored the use of the AnnoTree [7] database, which contains protein sequences for Bacteria and Archaea, based on the annotation of genomes from the GTDB database [8]. And in our study, we found that AnnoTree is only 1/4 the size of full NCBI-nr, while showing similar alignment rates and higher assignment rates, and calculations only take half as long, as compared to using the prokaryotic part of NCBI-nr [9] (While for the dataset used for this extended abstract AnnoTree run is 3.4 times faster than the NCBI-nr run. This is due to the increased number of protein entries in the NCBI-nr database compared to the one used in the original paper. Additionally, it's worth noting that the reported time in this instance is based on real-time measurements reported by software, whereas the paper [9] presented CPU time).

2. Summary of the study

In our study [9], we used 10 samples (9 short- and 1 long-reads) from different environmental sources like seagrass, river, skin, bioreactor, etc., and different sequencing technology. We also used one mock community long-read dataset, MBARC-26 [10], consisting of 23 bacterial and 3 archaeal strains.

We setup the protein FastA file and mapping file for AnnoTree required by the DIAMOND+MEGAN pipeline from data provided on the AnnoTree download page. The NCBI-nr FastA file was downloaded from NCBI-FTP and the mapping file for MEGAN was obtained from the MEGAN download page.

We carried out DIAMOND+MEGAN runs on all the datasets, comparing both against AnnoTree (referred to as AnnoTree runs) and against the prokaryotic part of NCBI-nr (referred to as NCBI-nr runs).

On the mock community, both types of runs resulted in taxonomic profiles with very few false positives (5 for NCBI-nr run and 2 for AnnoTree). On the real metagenomics samples, DIAMOND was able to align more than 51% of reads, with nearly 1% more reads aligned by DIAMOND against the AnnoTree.

DIAMOND-generated alignment files were processed by MEGAN to assign reads to the NCBI taxonomy

[11] and the GTDB-taxonomy, the EC [12], eggNOG [13], InterPro [14], KEGG [15], and SEED [16] functional categories. For both runs, we observed similar assignment rates, with some categories (GTDB, eggNOG, KEGG) showing an overall higher assignment rate for the AnnoTree run.

We further carried out a read-wise comparison of these assignments between both runs. On the NCBI-taxonomy, the two types of runs agree to 30 to 60% of all reads, with only a small percentage of reads assigned incompatibly to different lineages. For GTDB, a taxonomy higher assignment rate (≈ 99%) was observed for the AnnoTree runs, and there was a decrease in the level of conflicting assignments for most data sets. Also, for the KEGG functional category, the AnnoTree run show a much higher assignment rate of aligned reads.

Looking at CPU time used, an NCBI-nr run takes twice as long for alignment compared to an AnnoTree run, on average (Moreover, for the dataset used in this extended abstract, an NCBI-nr run takes 4.4 times longer than an AnnoTree run, possibly due to the increased database size. The time reported here is based on real-time measurements provided by DIAMOND, while the paper [9] presented CPU time).

3. Usage of BinAC and example workflow

We used BinAC to carry out the initial pre-processing of the datasets for our study and for generating the required files needed by the DIAMOND+MEGAN pipeline. Below we show how can one use BinAC (General workflow Figure 1) to carry out a DIAMOND+MEGAN analysis.

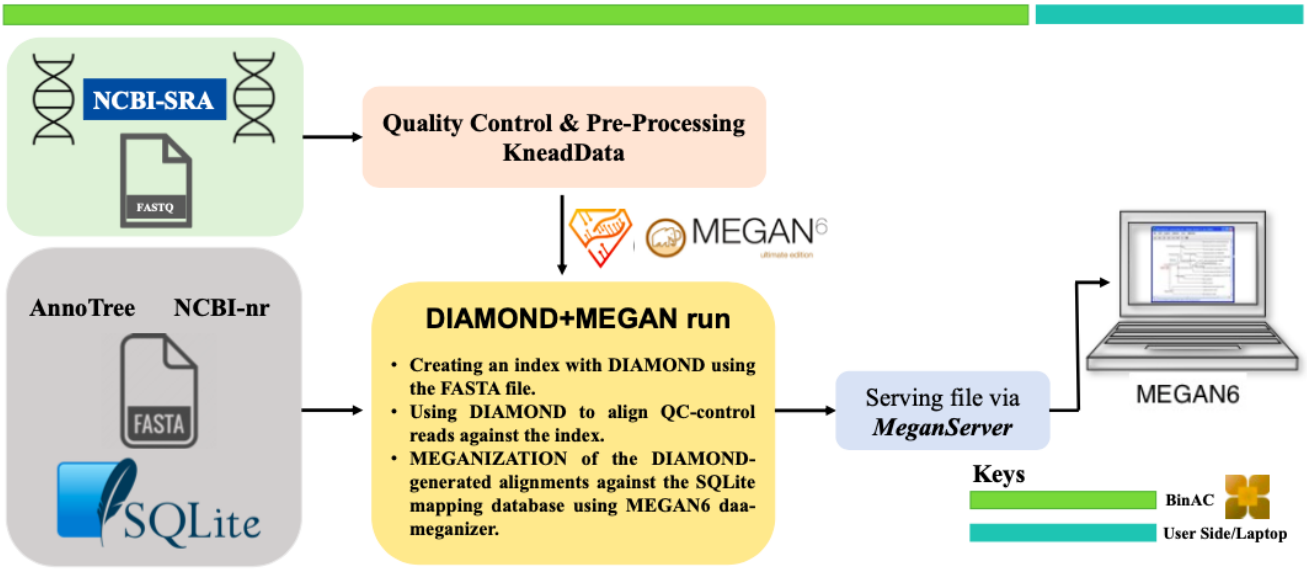


Figure 1: Workflow. The figure illustrates the general workflow for conducting a DIAMOND+MEGAN analysis on BinAC. The process begins with data retrieval from NCBI-SRA, followed by obtaining the respective databases for AnnoTree and NCBI-nr. Subsequently, quality control and preprocessing steps are implemented. The workflow then proceeds to the alignment phase using the preprocessed files and DIAMOND. The alignment results are then processed by MEGAN6 (daa-meganiser). Finally, the files are served by MeganServer, allowing for interactive analysis through the MEGAN6 GUI. The key distinguishes tasks performed on BinAC's side from those carried out on the user's laptop or end.

3.1 Data and databases

Here we use a whole genome shotgun metagenomic dataset set from the Human Microbiome Project (HMP) [17, 18] belonging to the healthy human subjects (HHS) cohort SRS011405 (https://www.ncbi.nlm.nih. gov/sra/SRX023992). It has two sequencing runs, their respective SRR ids are SRR061164 and SRR061166. Samples were downloaded using the NCBI SRA toolkit and fastq-dump. Below is the PBS script that was used for downloading the sample. In it, we indicate that we need 1 node with 1 core (nodes=1:ppn=1) and a wall-clock time of 10 hours (walltime=10:00:00). More information on the syntax of PBS can be found on the BinAC wiki pages https://wiki.bwhpc.de/e/BinAC/Quickstart_Guide. We used --split-3 to download or generate separate paired reads. The variable wd determines the working directory for file storage.

```
#PBS -N SRRDataDownload
#PBS -l nodes=1:ppn=1
#PBS -l walltime=10:00:00
#PBS -S /bin/bash
#PBS -j oe
#PBS -o AbsolutePathToLogFile
#PBS -M YourEmailAddress
```

```
wd="AbsolutePathToWrokingDirectory"
fq="AbsolutePathTosratoolkit/sratoolkit.3.0.2-ubuntu64/bin/fastq-dump"

echo ">>>>>>>>>>>>>>>>>>>> downloading data >>>>>>>>>>>>>>>>>>>>"

cd $wd
$fq --split-3 SRR061164
$fq --split-3 SRR061166
```

We submitted a short job using qsub -q short SRRDataDownload.sh. The whole PBS script is present in a file with the name SRRDataDownload.sh. It took 01:38:52 (Table 1) to download both SRR accession files.

For generating indices, a FastA file for AnnoTree was downloaded from the MEGAN-AnnoTree download page using the command wget https://software-ab.cs.uni-tuebingen.de/download/ megan-annotree/annotree.fasta.gz, and an NCBI-nr file was downloaded using command wget https: //ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz. Since NCBI-nr is large, we used a short job to down- load it. We also require nodes.dmp and names.dmp to add taxon node and taxon name information to the index. Both files are present in the taxdmp.zip, which is downloaded using the following command wget https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/taxdmp.zip. We also need prot.accession2taxid.FULL mapping information, which can be obtained using the command wget https://ftp.ncbi.nlm.nih.gov/ pub/taxonomy/accession2taxid/prot.accession2taxid.FULL.gz.

The files nodes.dmp, names.dmp, and prot.accession2taxid.FULL are only required if you want to restrict DIAMOND alignment to bacteria and archaea.

3.2 DIAMOND Index generation

Next, we generated the DIAMOND index using the above mentioned database files. DIAMOND (ver- sion:v2.0.15.153) was downloaded using the following command wget https://github.com/bbuchfink/ diamond/releases/download/v2.0.15/diamond-linux64.tar.gz.

For the NCBI-nr index we use the below mentioned PBS script, indicating that we need 1 node with 28 cores (nodes=1:ppn=28) and a wall-clock time of 14 hours (walltime=14:00:00). The variable temp determines the path to the temporary directory required by DIAMOND. We use the makedb command of DIAMOND and files are provided using the mentioned flag. We also used --taxonmap (for prot.accession2taxid.FULL), --taxonnodes (for nodes.dmp), --taxonnames (for names.dmp), and -t (for temporary directory path).

```
#PBS -N nrIndex
#PBS -l nodes=1:ppn=28
#PBS -l walltime=14:00:00
#PBS -S /bin/bash
#PBS -j oe
#PBS -o AbsolutePathToLogFile
#PBS -M YourEmailAddress

wd="AbsolutePathToWrokingDirectory"
diamond="AbsolutePathToDIAMOND/diamond"
taxdmp="AbsolutePathToDIAMOND/taxdmp/"
temp="AbsolutePathToTemporarayDirectory/"

echo ">>>>>>>>>>>>>>>>>>>> [diamond] makedb >>>>>>>>>>>>>>>>>>>>"
```

```
$diamond makedb --in ${wd}nr.gz -d ${wd}nr.dmnd \
--taxonmap ${wd}prot.accession2taxid.FULL \
--taxonnodes ${taxdmp}nodes.dmp --taxonnames ${taxdmp}names.dmp -t ${temp}
```

This is submitted using qsub -q short makeIndexnr.sh. The whole PBS script is present in a file with the name makeIndexnr.sh. It took 04:52:42 (Table 1) to generate the NCBI-nr index.

Similarly, we used the below PBS script to generate the AnnoTree index, the main change is in the requested wall-clock time, which is one hour (walltime=01:00:00). The options --taxonmap, --taxonnodes, --taxonnames are not used.

```
#PBS -N AnnoTreeIndex
#PBS -l nodes=1:ppn=28
#PBS -l walltime=01:00:00
#PBS -S /bin/bash
#PBS -j oe
#PBS -o AbsolutePathToLogFile
#PBS -M YourEmailAddress

wd="AbsolutePathToWrokingDirectory"
diamond="AbsolutePathToDIAMOND/diamond"
temp="AbsolutePathToTemporarayDirectory/"

echo ">>>>>>>>>>>>>>>>>>>> [diamond] makedb >>>>>>>>>>>>>>>>>>>>"

$diamond makedb --in ${wd}annotree.fasta.gz -d ${wd}annotree.dmnd -t ${temp}
```

This is submitted using qsub -q short makeAnnoTreeIndex.sh. The whole PBS script is present in a file with the name makeAnnoTreeIndex.sh. It took 00:20:47 (Table 1) to generate AnnoTree index.

3.3 Data pre-processing

We used KneadData for pre-processing of FastQ files to remove any contamination, low-quality reads, etc. We installed KneadData (version:v0.12.0) via Conda and downloaded Trimmomatic (version:0.39) [19]. We used the hg37 and human contamination database to remove human reads and contamination. It was downloaded using the command kneaddata database --download human bow tie2 $DIR, where $DIR is the path to the directory where you want to store database. Below is the PBS script we use for processing sample SRR061164, another sample can be processed in the same way,

Table 1: **BinAC Submitted Job Statistics**. For each run, we report the job name, the script name, num- ber of node requested (nodes), number of core requested (ppn), wall-clock time requested WR (HH:MM:SS), and wall-clock time used by job for completion WC (HH:MM:SS).

JobName	ScriptName	nodes	ppn	WR	WC
SRRDataDownload	fordataDownload.sh	1	1	10:00:00	01:38:52
nrIndex	makeIndexnr.sh	1	28	14:00:00	04:52:42
AnnoTreeIndex	makeIndexAnnoTree.sh	1	28	01:00:00	00:20:47
kneaddata SRR061164	kneaddata SRR061164.sh	1	28	10:00:00	00:33:27
kneaddata SRR061166	kneaddata SRR061166.sh	1	28	10:00:00	00:37:47
DM SRR061164	dm nr SRR061164.sh	1	28	48:00:00	20:17:58
DM SRR061164	dm nr SRR061166.sh	1	28	48:00:00	20:33:29
DM SRR061164 Annotree	dm annotree SRR061164.sh	1	28	48:00:00	05:40:09
DM SRR061166 Annotree	dm annotree SRR061166.sh	1	28	48:00:00	06:15:07

you can also process more than one sample in a single PBS script but you need to take care off the wall-clock time parameter. In the script we indicate that we need 1 node with 28 cores (nodes=1:ppn=28) and a wall-clock time of 10 hours (walltime=10:00:00). Variables r1 and r2 store the paired-end reads file absolute path, variable trimmo contains the absolute path to the Trimmomatic installation directory and kd db contains the absolute path to the hg37 and human contamination database.

```
#PBS -N kneaddata_SRR061164
#PBS -l nodes=1:ppn=28
#PBS -l pmem=4gb
#PBS -l walltime=10:00:00
#PBS -S /bin/bash
#PBS -j oe
#PBS -o AbsolutePathToLogFile
#PBS -M YourEmailAddress

source AbsolutePathToCondaInstallationDirectory/anaconda3/etc/profile.d/conda.sh
conda activate kneaddata

wd="AbsolutPathToWrokingDirectory"
r1="AbsolutPathToReadsFile/SRR061164_1.fastq"
r2="AbsolutPathToReadsFile/SRR061164_2.fastq"
trimmo="AbsolutPathToTrimmomaticInstallationDirectory/Trimmomatic-0.39"
sample="SRR061164"
kd_db="AbsolutPathToindexDirectory"

echo ">>>>>>>>>>>>>>>>>>>> [kneaddata] pe >>>>>>>>>>>>>>>>>>>>"
kneaddata --input1 $r1 --input2 $r2 \
 --output ${wd}${sample} -t 28 \
  --run-fastqc-start --reference-db ${kd_db} --trimmomatic $trimmo \
  --trimmomatic-options=" Your Trimmomatics options " \
   --max-memory 40000m --run-fastqc-end --bypass-trf  --reorder
```

One important thing to take care of is that we have installed kneaddata in a conda environment so we need to run source AbsolutePathToCondaInstallationDirectory/anaconda3/etc/profile.d/conda.sh before activating the environment. We can specify Trimmomatic options by using flag --trimmomatic-options, for more details please check the kneadData https://huttenhower.sph. harvard.edu/kneaddata/, and Trimmomatic http://www.usadellab.org/cms/?page=trimmomatic pages. We also used --bypass-trf as this just an example, but you can proceed for the removal of tandem repeats depending on your data.

The job was submitted using qsub -q short kneaddata SRR061164.sh. The whole PBS script is present in a file with the name kneaddata SRR061164.sh. It took 00:33:27 (Table 1) for run to complete.

The generated output files were concatenated to a single file by using com- mand cat SRR061164 1 kneaddata paired 1.fastq SRR061164 1 kneaddata paired 2.fastq SRR061164 1 kneaddata unmatched 1.fastq SRR061164 1 kneaddata unmatched 2.fastq >SRR061164.fastq. There are different strategies for handling paired-end data, depend on use-case. same was done for sample SRR061166.

3.4 DIAMOND+MEGAN run

Finally, the DIAMOND+MEGAN run was carried out by using the PBS script mentioned below, and we used MEGAN Ultimate Edition (version:6.24.22, built 12 Apr 2023) for linux from MEGAN6 Download Page, downloaded using wget https://software-ab.cs.uni-tuebingen.de/download/megan6/MEGAN_Ultimate_unix_6_24_22.sh. The most efficient way to use DIAMOND is to run it on a single compute node, using all threads. In the following, the same instructions can be used to run the Community Edition of MEGAN.

We used 1 node with 28 cores (nodes=1:ppn=28) and a wall-clock time of 48 hours (walltime=48:00:00). For the NCBI-nr run DIAMOND alignment was restricted to bacteria and archaea (--taxonlist 2,2157: where 2 is for Bacteria, and 2157 is for Archaea). DIAMOND blastx command was used. The flag -f 100 instructs DIAMOND to output DAA format (DIAMOND Alignment Archive), -b for block size, -t is for the temporary directory. variable diamond stores the absolute path to the DIAMOND, megan store path to daa-meganizer, dmnd store path to index, and mdbfile store path to megan-mapping files. Generated DAA files were processed by MEGAN for binning by using daa-meganizer, which is present in the MEGAN installation tools directory, we call this pro- cess meganization. For Meganization, we used a mapping file that is obtained from the MEGAN6 download page. For NCBI-nr run we used megan-map-Feb2022-ue.db (wget https://software-ab. cs.uni-tuebingen.de/download/megan6/megan-map-Feb2022-ue.db.zip) and for the AnnoTree run we used megan-mapping-annotree-June-2021.db (wget https://software-ab.cs.uni-tuebingen.de/download/megan-annotree/megan-mapping-annotree-June-2021.db.zip).

```
#PBS -N DM_SRR061164
#PBS -l nodes=1:ppn=28
#PBS -l walltime=48:00:00
#PBS -S /bin/bash
#PBS -j oe
#PBS -o AbsolutePathToLogFile
#PBS -M YourEmailAddress

wd="AbsolutPathToWrokingDirectory"
reads="AbsolutPathToReadsFile/SRR061164.fastq"
sample="SRR061164"

diamond="AbsolutePathToDIAMOND/diamond"
megan="AbsolutePathToReadsDIAMOND/megan/tools/daa-meganizer"
dmnd="AbsolutePathTonrIndexFile/nr.dmnd"
output="AbsolutePathToOutputDirectory/"

mdbfile="AbsolutePathToMEGANMappingFile/megan-map-Feb2022-ue.db"

echo ">>>>>>>>>>>>>>>>>>>> [DIAMOND] align to nr >>>>>>>>>>>>>>>>>>>>"

$diamond blastx -p 28 -d $dmnd -q $reads -o ${output}${sample}.daa -f 100 -b 4\
-c 1 -t ${wd} --taxonlist 2,2157

echo ">>>>>>>>> [MEGAN] MEGANIZING via megan-map-Feb2022-ue.db >>>>>>>>>>>"

$megan -i ${output}${sample}.daa -mdb $mdbfile -t 28

#PBS -N DM_SRR061164_Annotree
#PBS -l nodes=1:ppn=28
#PBS -l walltime=48:00:00
#PBS -S /bin/bash
```

```
#PBS -j oe
#PBS -o AbsolutePathToLogFile
#PBS -M YourEmailAddress

wd="AbsolutePathToWrokingDirectory"
reads="AbsolutePathToFile/SRR061164.fastq"
sample="SRR061164"

diamond="AbsolutePathToReadsDIAMOND/diamond"
megan="AbsolutePathToDIAMOND/megan/tools/daa-meganizer"
dmnd="AbsolutePathTonrIndexFile/annotree.dmnd"
output="AbsolutePathToOutputDirectory/"
mdbfile="AbsolutePathToMEGANMappingFile/megan-mapping-annotree-June-2021.db"

echo ">>>>>>>>>>>>>>>>>>>> [DIAMOND] align to AnnoTree >>>>>>>>>>>>>>>>>>>>"

$diamond blastx -p 28 -d $dmnd -q $reads -o ${output}${sample}.daa -f 100 -b 4\
-c 1 -t ${wd}

echo ">>>>>>>>> [MEGAN] MEGANIZING via megan-mapping-annotree-June-2021.db >>>>>>>>>>>"

$megan -i ${output}${sample}.daa -mdb $mdbfile -t 28
```

The only difference between the PBS script for the NCBI-nr run and AnnoTree is the index, and mapping files and that we don't need the --taxonlist flag for the AnnoTree run. The jobs were submitted using qsub -q short dm nr SRR061164.sh for the NCBI-nr run and qsub -q short dm annotree SRR061164.sh for the AnnoTree run. The whole PBS script is present in the named files. Similar jobs for sample SRR061166 were submitted. In Table 2, we report on DIAMOND alignment for samples SRR061164 and SRR061166 against NCBI-nr and AnnoTree, and in Table 3, we report on meganization of the alignments against NCBI- nr and AnnoTree.

Table 2: **DIAMOND alignment for sample SRR061164, and SRR061166 against NCBI-nr and AnnoTree**. For both the NCBI-nr and AnnoTree runs on the SRR061164, and SRR061166 short-read samples, we report the number of reads, the number of aligned reads, the total number of alignments, and the real time required (in minutes, reported by DIAMOND).

Run	Sample	Reads	Aligned	Alignments	Time
NCBI-nr	SRR061164	58,473,295	47,262,780	981,477,133	1,050.40m
AnnoTree	SRR061164	58,473,295	44,676,509	705,639,919	222.06m
NCBI-nr	SRR061166	59,097,591	46,922,500	970,394,311	1,059.92m
AnnoTree	SRR061166	59,097,591	44,606,308	697,445,398	254.276m

Table 3: **Meganization of alignments against NCBI-nr and AnnoTree**. For both the AnnoTree and NCBI-nr runs on the SRR061164, and SRR061166 short-read samples, we report the number of aligned reads, and the number of reads assigned to the NCBI and GTDB taxonomies, and to the EC, EGGNOG, INTERPRO, SEED, and KEGG functional classifications. We also report the time required for meganization of each file (reported by daa-meganizer).

Run	Sample	Aligned	Taxonomy	GTDB	EC	EGGNOG	INTERPRO	SEED	KEGG	Time
NCBI-nr	SRR061164	47,262,780	43,290,410	39,524,786	8,395,046	15,138,478	15,413,493	11,663,174	15,726,592	166.668m
AnnoTree	SRR061164	44,676,509	40,654,907	40,610,772	7,659,943	16,732,698	17,057,298	11,169,402	19,115,257	117.87m
NCBI-nr	SRR061166	46,922,500	42,231,682	38,600,797	8,305,174	14,713,798	14,928,688	11,510,802	15,453,216	172.77m
AnnoTree	SRR061166	44,606,308	39,722,656	39,691,627	7,497,388	16,198,287	16,583,549	10,984,994	18,748,077	120.02m

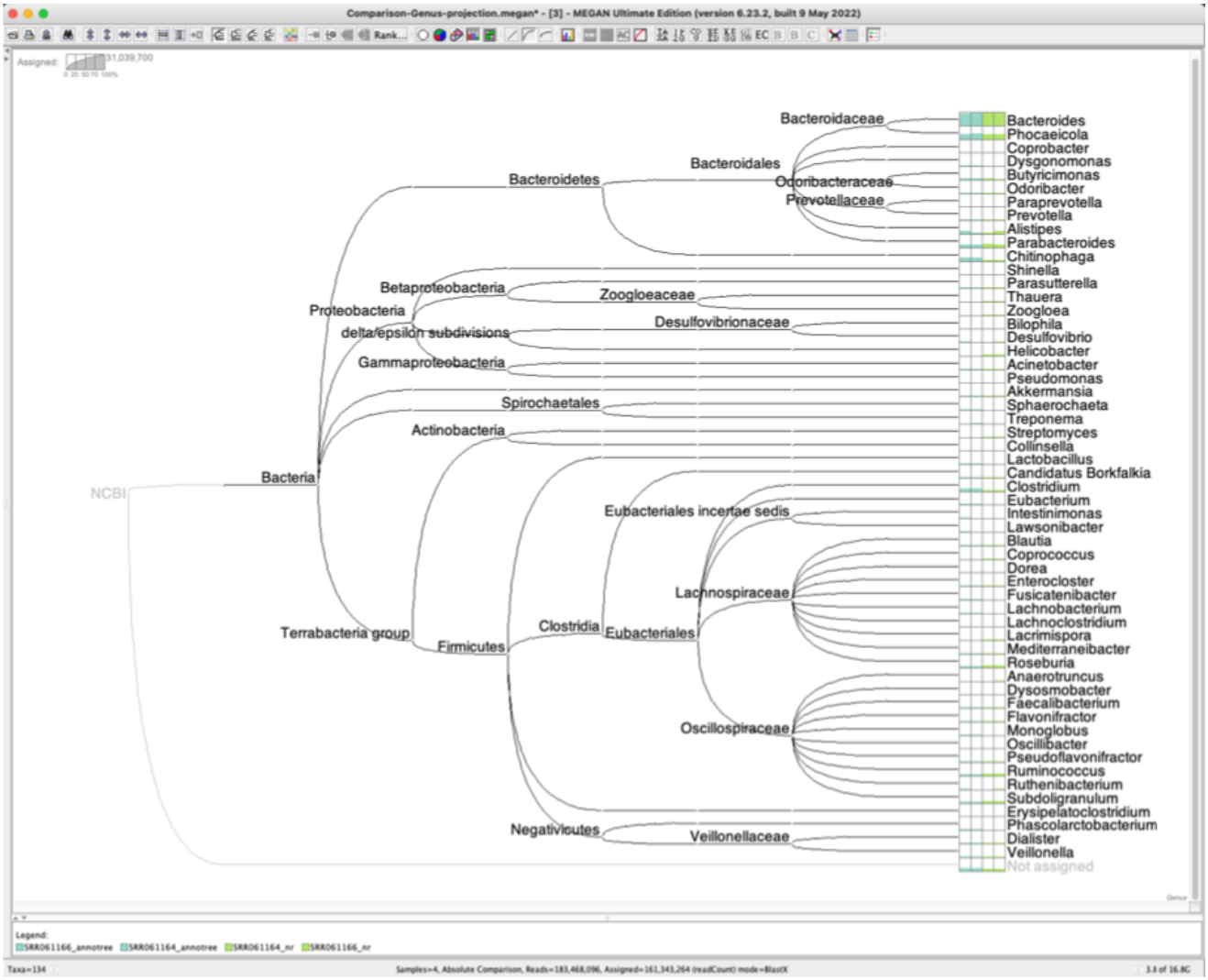


Figure 2: **Comparative View**. The NCBI taxonomy comparative view of short-read samples for SSR061164 and SSR061166 in NCBI-nr, and AnnoTree run, projected to the genus level.

We ran an instance of MeganServer [20] on BinAC to serve the meganized DAA files via HTML to an instance of MEGAN running on a personal MacBook Pro laptop. Figure 2 shows the default NCBI-taxonomy viewer of MEGAN with samples rank projected at the genus level. We used MEGAN Ultimate edition but the above commands are also applicable in the MEGAN Community edition. For more details on how to carry out the different types of analysis using MEGAN, one can follow the protocol by Gautam et al [1] or Caner et al [2]. Additionally, consult the MEGAN manual available at https://software-ab. cs.uni-tuebingen.de/download/megan6/manual.pdf.

This report uses a more recent NCBI-nr release than the one used in our paper describing the use of AnnoTree. Now, the alignment rate of NCBI-nr is higher than compared to the alignment rate using AnnoTree. A new release of AnnoTree is expected in 2024 and we anticipate that will improve the alignment rate when using AnnoTree.

Acknowledgements

The authors acknowledge support by the High Performance and Cloud Computing Group at the Zentrum für Datenverarbeitung of the University of Tübingen, the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant no INST 37/935-1 FUGG.

References

[1] Anupam Gautam, Wenhuan Zeng, and Daniel H Huson. DIAMOND+ MEGAN microbiome analysis. In *Metagenomic Data Analysis*, pages 107–131. Springer, 2023.

[2] Caner Bâgcı, Sascha Patz, and Daniel H Huson. DIAMOND+ MEGAN: fast and easy taxonomic and functional analysis of short and long microbiome sequences. *Current protocols*, 1(3):e59, 2021.

[3] Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using DIA- MOND. *Nature methods*, 12(1):59–60, 2015.

[4] Daniel H Huson, Alexander F Auch, Ji Qi, and Stephan C Schuster. MEGAN analysis of metagenomic data. *Genome research*, 17(3):377–386, 2007.

[5] Daniel H Huson, Sina Beier, Isabell Flade, Anna Górska, Mohamed El-Hadidi, Suparna Mitra, Hans- Joachim Ruscheweyh, and Rewati Tappu. MEGAN community edition-interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS computational biology*, 12(6):e1004957, 2016.

[6] Dennis A Benson, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and David L Wheeler. Gen- Bank. *Nucleic acids research*, 33(suppl 1):D34–D38, 2005.

[7] Kerrin Mendler, Han Chen, Donovan H Parks, Briallen Lobb, Laura A Hug, and Andrew C Doxey. AnnoTree: visualization and exploration of a functionally annotated microbial tree of life. *Nucleic acids research*, 47(9):4442–4448, 2019.

[8] Donovan H Parks, Maria Chuvochina, David W Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil, and Philip Hugenholtz. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature biotechnology*, 36(10):996–1004, 2018.

[9] Anupam Gautam, Hendrik Felderhoff, Caner Bâgci, and Daniel H Huson. Using AnnoTree to get more assignments, faster, in DIAMOND+MEGAN microbiome analysis. *Msystems*, 7(1):e01408–21, 2022.

[10] Esther Singer, Bill Andreopoulos, Robert M Bowers, Janey Lee, Shweta Deshpande, Jennifer Chiniquy, Doina Ciobanu, Hans-Peter Klenk, Matthew Zane, Christopher Daum, et al. Next generation sequencing data of a defined microbial mock community. *Scientific data*, 3(1):1–8, 2016.

[11] Conrad L Schoch, Stacy Ciufo, Mikhail Domrachev, Carol L Hotton, Sivakumar Kannan, Rogneda Khovanskaya, Detlef Leipe, Richard Mcveigh, Kathleen O'Neill, Barbara Robbertse, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database*, 2020, 2020.

[12] Edwin C Webb et al. Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes. (Ed. 6), 1992.

[13] Jaime Huerta-Cepas, Damian Szklarczyk, Davide Heller, Ana Hern´andez-Plaza, Sofia K Forslund, Helen Cook, Daniel R Mende, Ivica Letunic, Thomas Rattei, Lars J Jensen, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic acids research*, 47(D1):D309–D314, 2019.

[14] Alex L Mitchell, Teresa K Attwood, Patricia C Babbitt, Matthias Blum, Peer Bork, Alan Bridge, Shoshana D Brown, Hsin-Yu Chang, Sara El-Gebali, Matthew I Fraser, et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic acids research*, 47(D1):D351–D360, 2019.

[15] Minoru Kanehisa and Susumu Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[16] Ross Overbeek, Robert Olson, Gordon D Pusch, Gary J Olsen, James J Davis, Terry Disz, Robert A Edwards, Svetlana Gerdes, Bruce Parrello, Maulik Shukla, et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic acids research*, 42(D1):D206–D214, 2014.

[17] A framework for human microbiome research. *Nature*, 486(7402):215–221, 2012.

[18] Heather Huot Creasy, Victor Felix, Jain Aluvathingal, Jonathan Crabtree, Olukemi Ifeonu, James Matsumura, Carrie McCracken, Lance Nickel, Joshua Orvis, Mike Schor, et al. HMPDACC: a Human Microbiome Project Multi-omic data resource. *Nucleic Acids Research*, 49(D1):D734–D742, 2021.

[19] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

[20] Anupam Gautam, Wenhuan Zeng, and Daniel H Huson. MeganServer: facilitating interactive access to metagenomic data on a server. *Bioinformatics*, 39(3):btad105, 2023.

A Proof of Concept for High Energy Physics Data Archival of PhD and Master Theses at the University of Freiburg

Michael Böhler

Institute of Physics, University Freiburg, Germany

Abstract

This contribution describes a proof of concept for the archival of the entire life cycle of the data, the analysis code, and the necessary software stack of an analysis realised during a typical PhD or master thesis in experimental High Energy Physics. Altogether, the derived datasets, the container with the appropriate software, and the dedicated analysis code are wrapped together into a dataset and stored on an appropriate long-term storage, e.g. on the S3 storage of the bwSFS (Baden-Württemberg – Storage for Science) hosted in the computing centres of the University of Freiburg and Tübingen. A database keeps track of the individual datasets, so that based on the author's name, the thesis title, or the doi of the considered thesis, the corresponding dataset can be pulled up. Furthermore, it is described how the setup can be generalised to be utilised for the entire physics department at the University of Freiburg.

1. Introduction

A typical physics analysis for a PhD thesis, in the field of experimental High-Energy Physics (exp. HEP) at an LHC experiment like the ATLAS experiment, evaluates currently input data of the order of 40 PB. This corresponds to the recorded dataset of the LHC Program of Run 2 (2015-2018) and the corresponding simulated events. Since, according to the guidelines of the German Research Foundation [1], research data has to be archived for at least 10 years, a good compromise has to be found for storing the complete dataset and the necessary software tools used for the physics analysis of a PhD or master thesis considering the available storage capacity. The analysis workflow of the LHC experiments foresees a data analysis with an appropriate data reduction on the world wide LHC computing grid (WLCG). The fundamental idea of the archival procedure described here, is to store a list of the original datasets, the used software stack as docker container and the analysis specific software, which has been utilised to provide the reduced datasets. The reduced datasets, the software stack used for further analysis steps, as well as analysis specific software has to be archived in addition. All items listed can be bundled together, labelled with appropriate metadata and stored in the dedicated S3 bucket reserved for thesis data archival. This article is organised as follows. First the data size reduction and a suitable folder structure is discussed in detail, together with the software suite *dtool* [2], which is used for bundling the data and the source code and attaching meaningful metadata. Then the web front-end, which is implemented with the Python based web framework *Django* [3] and the hardware requirements are described. Before concluding, the generalization of the proposed setup is briefly discussed.

2. Data size

As described above, the original dataset of about 40 PB is not suitable to be stored for each thesis individually. Therefore, the thesis data archival has to start with a reduced dataset. This dataset has to be produced on the WLCG. Once an analysis specific reduced dataset is available, it has to be downloaded to the local storage system. A typical thesis, analysing ATLAS data [4] from LHC Run2 (2015-2018), is based on $\sim$ 10 TB reduced datasets. Usually, another analysis step applies the final

selection and creates further output data and files with histograms. The output data is usually of order of 5-10 GB and is used to create additional histograms or as input data for further statistical analysis such as high dimensional likelihood. This data amounts to another data volume of about 1 GB. These final analysis steps hardly contribute to the total amount of roughly 10 TB per thesis. An overview of the single steps is shown in Fig. 1.

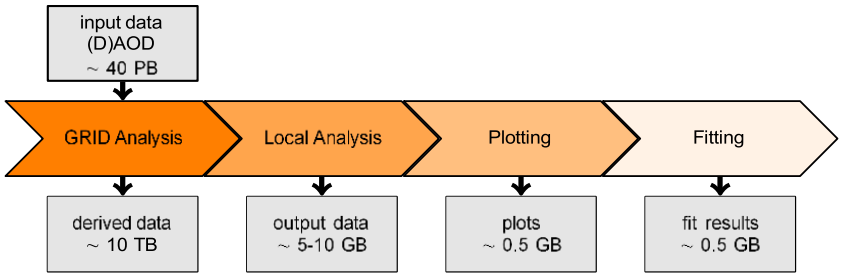


Figure 1: *Data sizes of input and output data of a typical ATLAS analysis.*

item	data size [GB]
reduced data set from WLCG	10,000
software stack - docker image	5
source code user analysis	« 1
local output data	5 - 10
plots and fit results	1
sum	10,015

Table 1: *List of the data sizes of all components of a physical analysis. The sum is dominated by the input dataset size.*

In addition to the datasets, the source code of the user analysis, and the appropriate software stack has to be archived. Table 1 summarises the total amount of data. The overall sum is still dominated by the input dataset size. The ATLAS computing model for Run-3 (2022-2025) foresees comparable data sizes, although larger amounts of data are expected, this can be achieved through more aggressive data reduction during the data processing on the WLCG.

3. Data structure

All components of the analysis have to be archived together, so that when examining a thesis, both data and source code are available and the respective result can be reproduced. The proof of concept described here is based on *dtool*, a software suite dedicated for managing scientific data. *dtool* allows to freeze a dataset and provides a command line interface to label a dataset with appropriate metadata. Only three steps are necessary to create a *dtool* dataset

1. create a so called “proto dataset”
2. add data and metadata to this “proto dataset”
3. convert the “proto dataset” into a dataset by “freezing” it

During the creation of the “proto dataset”, the metadata in the upper (grey) box in Fig. 2 is generated automatically. Then the data which should be archived has to be moved or copied to the data directory in the “proto dataset”. A proposed data structure is shown in the lower (yellow) box in Fig. 2.

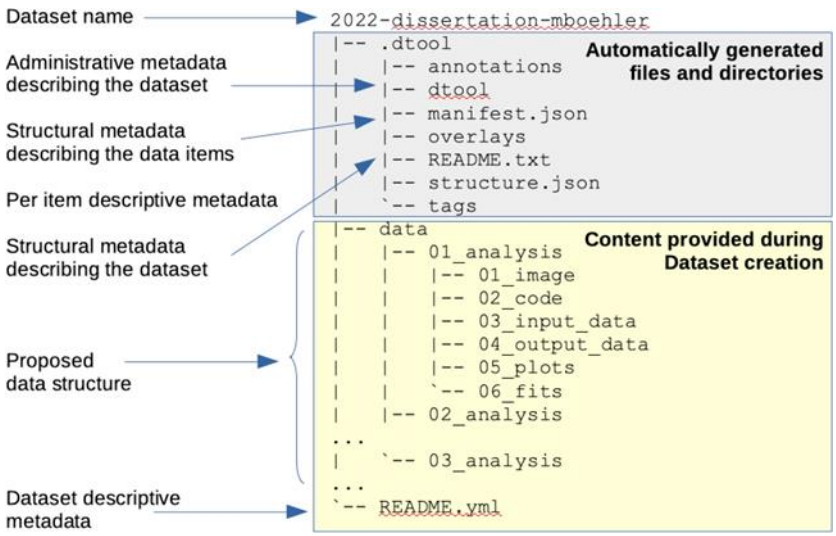


Figure 2: *Data structure of a dtool dataset. The directory called .dtool contains the hidden metadata which is created automatically by the dtool software and should not be edited by the user [2]. The README.yml file contains user defined metadata, which should describe the dataset, it can be read and modified by the Python API. The items of the dataset are stored in the directory named data. The lower (yellow) box shows the proposed structure to store all components of the analysis. Sometimes a PhD thesis consists of more than one analysis, in such a case one needs to number the analyses.*

The items in the dataset descriptive metadata are created based on a predefined template. This template can be customised in such a way that each user is able to create valid and meaningful metadata for the dataset. Figure 3 shows an example of the process of generating the metadata based on a tailor-made template of the thesis archival.

```
dtool readme interactive 2022-dissertation-mboehler
author [lastname, first name]: Boehler, Michael
title [thesis title]: Study of Higgs Properties
type [thesis type Dissertation/Master/Bachelor]: Dissertation
supervisor [last name, first name]: Schumacher, Markus
group [Experimentelle Teilchenphysik Abt. Prof. Markus Schumacher]:
doi [.]: https://doi.org/10.7717/peerj.6562
uri [.]:
date_published [2022-09-27]: 2022-09-27
date_modified [2022-09-27 ]: 2022-09-27
Updated readme
```

Figure 3: *Example of creating metadata with a tailor-made template for the thesis archival use case.*

dtool provides support for storages systems as S3, Azure, ECS S3, and iRODS. A Python Application Programming Interface (API) allows to manage the datasets and read and overwrite the metadata. These functionalities are quite convenient because the setup of the proof of concept described here, is designed to store data in the S3 data store of the bwSFS [5] storage system at the University of Freiburg and Tübingen. The setup requires to be able to adjust the metadata, because important entries as the doi or the final submission date of the thesis might be unknown when freezing the dataset.

4. Web frontend

The dataset selection as well as metadata maintenance should be able to be done centrally via a simple and intuitive web interface. The Python based web framework *Django* allows to create such a web front-end with a few lines of code. It keeps the data within any preferred database and provides a user administration for admin users (privileged permissions) and normal users (unprivileged permissions). The data from the database is published on the web front-end via so called views, following the model–template–views (MTV) architectural pattern. [6] The big advantage of using *Django* instead of any php based web front-end, is the simple integration of *dtool's* Python API. This allows to synchronise the metadata of the datasets with the data stored in the *Django* database. The initial ingest to the *Django* database is carried out by a synchronisation of the S3 storage and the database, when the user is uploading the dataset to the storage. A data steward with privileged access is allowed to maintain the metadata via the web interface (see Fig. 4). When the dataset was stored long enough (10 years), the web interface of the data steward shows the expiration and the dataset can be deleted.

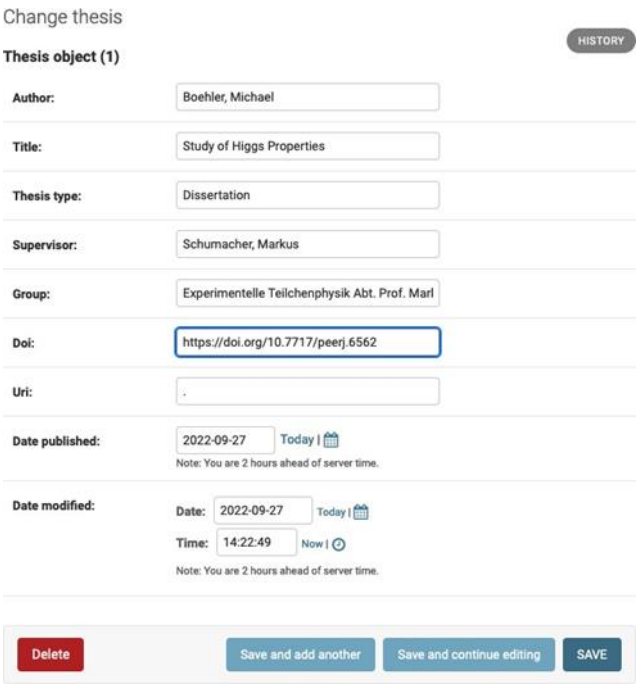


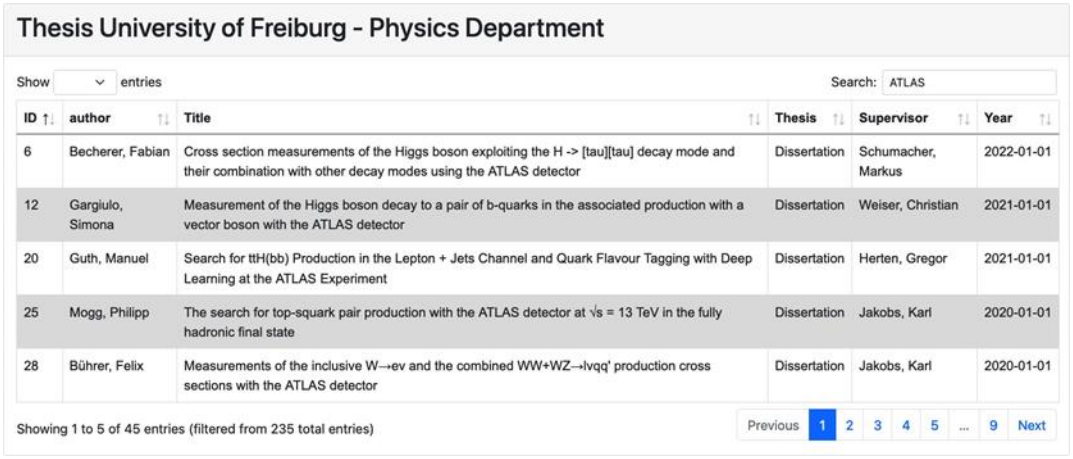
Figure 4: *Example of the Django admin view (privileged access) for metadata modification.*

Unprivileged users (e.g. all members of the University of Freiburg) are allowed to search the *Django* database via a dedicated web site. Figure 5 shows a demonstrator web page with access to a *Django* database preloaded with all PhD theses of the physics department since 2010 (data source: FreiDok[7]).

5. Hardware requirements

The web server can be installed on a small virtual

machine with 4GB RAM, 1 core and 100 GB disk space. It has to be accessible at least from the internal network of the University. The web server requires access to the S3 storage back-end, such that the metadata from S3 can be read and published on the web server and metadata modifications or dataset deletions executed by a privileged user on the web server can act on datasets stored on the S3 storage. The *Django* software and *dtool* need to be installed on the web server. Since the initial dataset creation should be performed by the user, the *dtool* package together with the template providing the appropriate metadata structure has to be present in the working environment of the user, ideally pre-installed on the used login nodes. All necessary components are summarised in the sketch in Fig. 6



Thesis University of Freiburg - Physics Department

Show entries Search: ATLAS

ID	author	Title	Thesis	Supervisor	Year
6	Becherer, Fabian	Cross section measurements of the Higgs boson exploiting the H -> [tau][tau] decay mode and their combination with other decay modes using the ATLAS detector	Dissertation	Schumacher, Markus	2022-01-01
12	Gargiulo, Simona	Measurement of the Higgs boson decay to a pair of b-quarks in the associated production with a vector boson with the ATLAS detector	Dissertation	Weiser, Christian	2021-01-01
20	Guth, Manuel	Search for ttH(bb) Production in the Lepton + Jets Channel and Quark Flavour Tagging with Deep Learning at the ATLAS Experiment	Dissertation	Herten, Gregor	2021-01-01
25	Mogg, Philipp	The search for top-squark pair production with the ATLAS detector at √s = 13 TeV in the fully hadronic final state	Dissertation	Jakobs, Karl	2020-01-01
28	Bührer, Felix	Measurements of the inclusive W→eν and the combined WW+WZ→lνqq' production cross sections with the ATLAS detector	Dissertation	Jakobs, Karl	2020-01-01

Showing 1 to 5 of 45 entries (filtered from 235 total entries) Previous 1 2 3 4 5 ... 9 Next

Figure 5: *Example view of a Django web page with access to a database preloaded with all PhD theses of the physics department since 2010.*

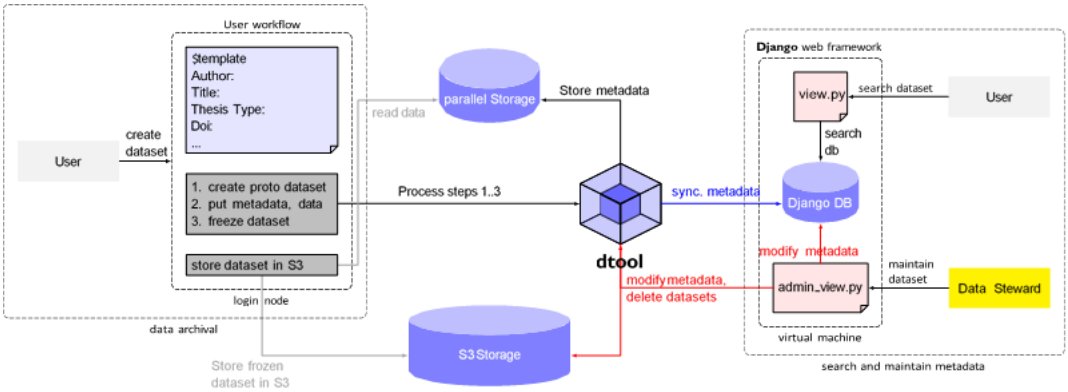


Figure 6: *Overview of the individual components of this proof of concept. The central component is the dtool software package, which provides the necessary command line interface to create meaningful metadata during the data archival process, as well as the functionality required for an automated synchronisation of the Django database and the available metadata on the S3 storage.*

6. Generalization

The success of a proof of concept is usually driven by its flexibility and scalability. Is this setup scalable to the entire physics department or even to the entire university?

Figure 7 shows the number of PhD theses of the physics department since 2010. The golden bars show the number of dissertations in exp. HEP and the grey hashed bars are stacked on top of the exp. HEP numbers showing the dissertations of other physics groups, which are registered on the publication platform FreiDok[7] of the University of Freiburg. There were 34 PhD theses submitted in the field of exp. HEP and 180 in total. Considering the 10 TB per thesis, this would amount to a total disk space of 340 TB for the exp. HEP community and 1.8 PB for the entire physics department assuming that a non HEP thesis requires the same amount disk space as a exp. HEP analysis, which is a rather conservative assumption. The number of privileged users can be estimated by the number of primary supervisors of the recorded PhD theses integrated over the last 10 years, as shown in Fig. 8. *Django* can easily handle order of 50 privileged users.

Since *Django* provides a user administration interface, each working group within the physics department can have its own privileged user, who is responsible to maintain the metadata of the thesis datasets and executes the final dataset deletion once or twice a year.

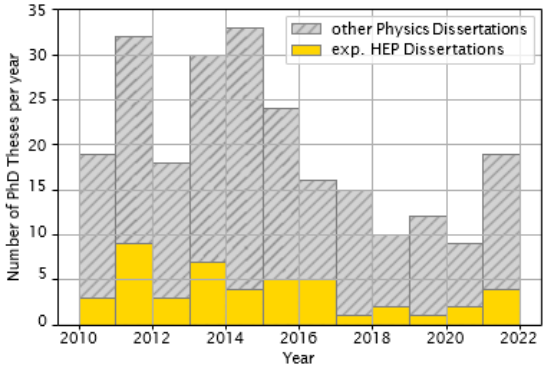


Figure 7: *Number of recorded PhD theses since 2010, exported from FreiDok [7]. There were 34 PhD the- sis submitted in the field of exp. HEP and 180 in total in the physics department.*

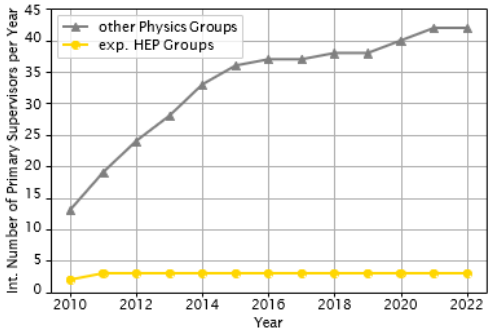


Figure 8: *Integrated number of primary theses super- visors since 2010. In total there are 42 supervisors (grey triangles), 4 of them (golden circles) belong to exp. HEP (data source FreiDok[7]).*

7. Conclusion

A rather simple setup has been presented, which fulfils the requirements to archive the datasets, software stack, and the user analysis code of exp. HEP theses. This makes use of well-established software packages: *Django* and *dtool*. The datasets are searchable in a web interface. The chosen implementation allows to both monitor the age of a given dataset and to delete it centrally via the web interface. The built-in functionality of a user management allows to scale the setup to other communities, if the capacity of the storage system is big enough.

Acknowledgements

The author acknowledge support by the state of Baden-Württemberg and the German Research Foundation (DFG) through grants INST 39/1098-LAGG and INST 39/1099-1 FUGG (bwSFS).

References

[1] Allianz Der Deutschen Wissenschaftsorganisationen Grundsätze zum Umgang mit Forschungsdaten. (Allianz der deutschen Wissenschaftsorganisationen,2010), https://doi.org/10.2312/allianzoa.019

[2] Olsson, T. & Hartley, M. Lightweight data management with dtool. *PeerJ*. 7 pp. e6562 (2019,3), http://dx.doi.org/10.7717/peerj.6562

[3] Django Software Foundation Django. (2019,5,5), https://Djangoproject.com

[4] I. Bird, *et al.* "Update of the Computing Models of the WLCG and the LHC Experiments," CERN-LHCC- 2014-014.

[5] Hahn, U., Von Suchodoletz, D., Quandt, M., Gloglowski, K. & Seifert, M. bwSFS - Storage for Science: Stand und Entwicklungen eines verteilten FDM-Systems. (Heidelberg University Library,2021)

[6] Kaplan-Moss, A. The Definitive Guide to Django - Web Development Done Right. (Apress,2008)

[7] FreiDok plus Universitätsbibliothek Freiburg. (2022), https://freidok.uni-freiburg.de/inst/1008, [Online; Stand 6. Oktober 2022]

Climate sensitivity and convective parameterization in the Earth system model of intermediate complexity PlaSim

Felix Pollak[1], Elisa Ziegler[2], Olga Erokhina[1], and Kira Rehfeld[2]

[1]Institute of Environmental Physics, Heidelberg University, Germany

[2]Department of Geosciences and Department of Physics, Tübingen University, Germany

Abstract

Earth system models of intermediate complexity, like the Planet Simulator (PlaSim), can be used for studying the long-term climate response to external forcing, for example, from greenhouse gas emissions. The equilibrium climate sensitivity (ECS) quantifies this response and is one of the most important parameters to project future global warming. Convective parameterization plays a key role for the hydrological cycle of atmospheric general circulation models such as PlaSim, but its influence on ECS remains uncertain. The aim of this study is to investigate the influence of two different convection schemes on the ECS in PlaSim. To achieve this goal, we implement a simple Tiedtke convection scheme (following Molteni [1]) in PlaSim and tune the model. The results show that the new scheme reduces the ECS compared to the operational Kuo scheme. Five of eight configurations yield ECS values which are within the very likely range obtained from more complex climate simulations. In addition to the convection scheme, we identify other parameters, like the horizontal heat diffusion that influence the ECS in PlaSim.

1. Introduction

Numerical climate models provide a unique toolbox to investigate future climate scenarios under global warming. In addition, they are used to investigate the behaviour of the climate system and its key mechanisms. Physics-based models are invaluable in providing the scientific base for policies tackling anthropogenic warming [2]. A downside of the most comprehensive model types, like Atmospheric General Circulation Models (AGCMs), is their high computational cost. This limits the applicability for multi-centennial or longer simulations, which are crucial for investigating the long-term climate response. In contrast, AGCMs with simpler parameterization schemes, classified as Earth System Models of Intermediate Complexity, like the Planet Simulator (PlaSim) [3, 4], have an increased computational efficiency, while still retaining a desired level of accuracy. The results depend on the chosen parameterization schemes and their level of complexity. The parameterization of clouds and deep convection is considered a large uncertainty in these models [5]. Equilibrium climate sensitivity (ECS) is a central quantity for describing how sensitive Earth's climate system is to changes in radiative forcing. It is defined as the temperature change following an instantaneous doubling of atmospheric CO_2 concentrations and after having reached a new equilibrium [6]. The IPCC 6th Assessment Report (AR6) [7] concludes that the ECS is very likely (≥ 90 %) in the interval of 2 ◦C to 5 ◦C, 33 likely (≥ 66 %) in the interval of 2.5 ◦C to 4 ◦C, virtually certain (≥ 99 %) above 1.5 ◦C and has the best estimate of 3 ◦C. Despite large efforts to reduce the uncertainty, especially on the upper bound, ECS remains 35 a major source of uncertainty in climate projections [8]. We investigate the climate sensitivity in PlaSim and its dependence on convective parameterization. To this end, we implement a new convection scheme in PlaSim and tune it to present-day climate. We compare it to the existing Kuo convection scheme [9, 10]. Moreover, we study how PlaSim's ECS depends on certain

parameters like the horizontal heat diffusion, which plays a key role in obtaining reliable simulations for present-day climate [11].

2. Theory and methods

2.1 Planet Simulator

This work uses a modified PlaSim version, called PaleoPlaSim1. It contains updates in several model compartments and was designed for long transient simulations [12]. PlaSim's dynamical core is a simplified General Circulation Model (GCM) called PUMA, which solves the wet primitive equations in the atmosphere [3, 4]. Here, all simulations apply a T21 (5.63◦×5.63◦) or T42 (2.81◦×2.81◦) spectral resolution combined with 10 vertical layers. PlaSim has a modular structure, which allows selecting subsets and coupling them to the atmospheric GCM [13]. For this work, only the mixed layer (ML) ocean and the sea ice module were considered. The ML ocean module in PlaSim uses a mixed layer model with constant thickness to simulate the sea surface temperatures (SSTs) and other related oceanographic quantities [13]. In contrast, for an atmosphere-only simulation, the SSTs are prescribed from climatology. The same holds for the sea ice distribution, which can either be prescribed from climatology or simulated by the sea ice module, which uses a thermodynamic sea ice model, based on a zero layer model [11]. Many physical processes, which are not resolved in PlaSim are included by simplified parameterizations. Further details on the model can be found in the PlaSim reference manual [4] and the PaleoPlaSim documentation [12].

2.2 Convective parameterization in PlaSim

The vertical distribution of latent heat and moisture in the atmosphere is mainly driven by cumulus convection. Therefore, it is of crucial importance to adequately describe this process within a GCM. However, the spatial extent of cumulus clouds is *O*(1 km), whereas T42/T21 GCM resolution implies *O*(500 km) grid cells. This necessitates including collective effects of an ensemble of convective clouds with parameterization schemes [5]. The deep cumulus convection in PlaSim is parameterised by a modified version of the standard Kuo convection scheme [9, 10]. Here, the main drivers for cumulus convection are moisture convergence into one grid column plus the surface evaporation from below and the temperature differences between the cloud air and the surrounding air [4, 9, 10]. Because of its simplicity, it has been widely used. However, due to some critical model deficiencies (e.g., overestimation of net heating in the lower troposphere), mass flux schemes have been implemented in operational models since the 1980s [15]. They describe cumulus convection with an ensemble of convective clouds consisting of updrafts, which lead to entrainment and detrainment of environmental air, and downdrafts proportional to the mass flux from the updrafts (Figure 1).

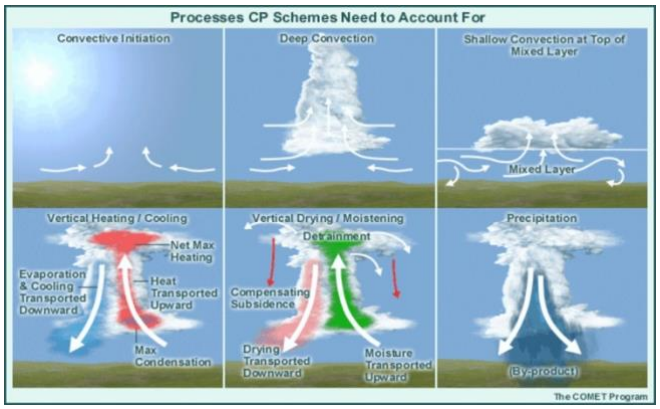


Figure 1: *The main processes occurring due to convection. These processes have to be accounted for in convective parameterization schemes. Figure from The COMET Program [14].*

This study investigates the influence of one of these convective parameterizations on the climate sensitivity, as simulated in PlaSim. Therefore, the simplified Tiedtke mass flux scheme from the SPEEDY GCM [1] was implemented. Here, the updrafts transport saturated air from the planetary boundary layer (PBL) up to the "top-of-convection" (TCN) level. Entrainment occurs between the PBL and a predefined layer, typically in the mid-troposphere. Detrainment is only allowed at the TCN level, which simplifies the calculation of the latent heat release. Thus, convective precipitation is only calculated at the TCN level [1]. The new parameters, introduced from the Tiedtke scheme, were tuned, such that PlaSim simulates a mean climatology in accordance with observations.

2.3 Parameter tuning

Parametrizations may be physically constrained. However, parameters are often uncertain and sometimes unobservable [16]. Tuning then estimates parameters which minimise the difference between observations and model simulations of a chosen set of variables.

Atmospheric components of PlaSim have been tuned in the past using an adjoint method [17]. This study uses the approach of Mehling *et al.* [18]. As a first step, a suitable objective function was chosen, which defines the tuning target. Here, the normalised mean absolute errors between ERA5 reanalysis data and some target fields were computed. Those fields consist of the 2-D precipitation and evaporation patterns. Additionally, the 2 m temperature and top of the atmosphere (TOA) fields were selected to avoid overtuning and retain a reasonable climatology. To obtain a list of the most important parameters for tuning, an initial sensitivity experiment was carried out. Here, the parameters were chosen randomly within physically motivated bounds. Subsequently, a Random Forest Regression was used to calculate the permutation importance of each parameter on the objective function [19, 20]. Finally, after identifying a set of tunable parameters, Bayesian optimization was used to identify the optimal parameters that minimise the objective function[1] [21].

[1] Tuned values: https://github.com/felyx04/tuned-params-bwClusterAbstract

2.4 Climate sensitivity

Here, we investigate three research questions regarding the ECS of PlaSim:

1. What is the influence of the convection scheme on ECS in PlaSim?
2. How does horizontal heat diffusion in the ML ocean impact ECS?
3. Which parameters impact the ECS most?

To answer the first question, we compare ECS values for both convection schemes. Therefore, an ensemble of 150 simulations was performed for each scheme. Every single simulation was run for 50 years with a CO_2 concentration of 397 ppm, restarted with a doubled concentration of 794 ppm and rerun for 50 more years. The last 10 years of both intervals were selected for time averaging the global mean surface temperature (GMST). The temperature difference between both intervals yields the ECS. Additionally, both models were run once with and once without the sea ice module.

The second question concerns the results from *Angeloni et al.* [11] who identified the horizontal heat diffusion in the ML ocean as an important process for reproducing the observed present-day climate. It reduces the cold bias for the GMST, which is present for the default PlaSim version when coupled to ML ocean and sea ice module. Their tuning of the horizontal heat diffusion coefficient K_h resulted in two different values for the Northern (NH) and Southern Hemisphere (SH). Following the same steps as above, we investigate the influence of the horizontal heat diffusion on the ECS.

Regarding the third question, each investigated parameter was randomly varied within its physically or numerically constrained bounds. The remaining parameters were set to their default values. A 100-year-long simulation with a doubling of CO_2 concentrations after 50 years was run, as described above, to obtain the corresponding ECS. This procedure was repeated a reasonable number of times to sample the whole parameter space.

All simulations were carried out with the *bwUniCluster 2.0*. With convection modifications, runtime demands remain low, 1 simulation year equals 4 to 5 min on a single core. With 16 cores 1 simulation year required 50 s machine time. The time steps for the simulation were set to 45 min.

3. Results and discussion

3.1 Evaluation of the new convection scheme and tuning

To compare the Kuo and Tiedtke convection schemes, we focus on precipitation and TOA fields. Figure 2 (a-c) shows the offset in global mean precipitation between PlaSim and ERA5 reanalysis data (period: 1979-2018). Panel d-f depict the difference in TOA radiation imbalance. Tuning the Tiedtke scheme strongly improves the agreement with ERA5. Without tuning, the TOA radiation imbalance is underestimated almost everywhere. The improvement after tuning is less pronounced for precipitation compared to TOA radiation imbalance. The strong dry bias over South America, which is already present in the default Kuo PlaSim version, is decreased in the tuned version. Furthermore, a dry bias over Sub-Saharan Africa and a wet bias over the Indian Ocean are reduced as well. The precipitation patterns of PlaSim with Kuo convection and the tuned Tiedtke scheme look similar. The dry bias over South America is more pronounced in the Tiedtke version, whereas the dry bias over Africa is stronger for the Kuo scheme. Much stronger deviations are introduced by the different convective clouds, influencing the TOA radiation imbalance. The Tiedtke scheme overestimates the TOA radiation imbalance almost everywhere over the sea. The Kuo scheme tends to underestimate it. Over land, both models tend to underestimate the fluxes. This process is more distinct for the Kuo scheme.

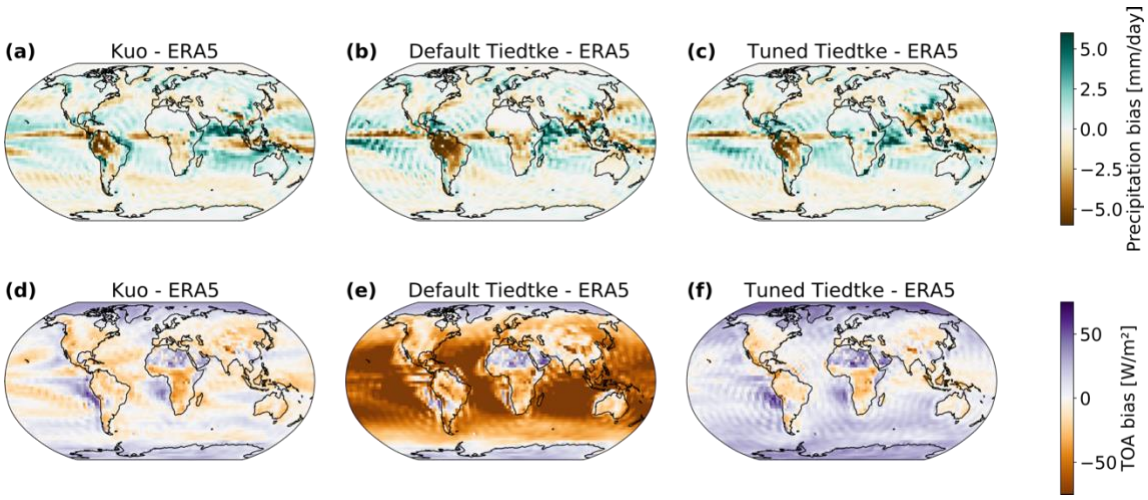


Figure 2: *Global mean annual bias of precipitation (panel a-c) and TOA radiation imbalance (panel d-f) between PlaSim simulations and ERA5 reanalysis data (averaged over period 1979-2018). Panel a and d show the results for standard Kuo convection, panel b and e for the untuned Tiedtke scheme and panel c and f for the tuned one. All simulations are performed with T42 spatial resolution and run for 100 years with boundary conditions set according to the period 1979-2018. The annual mean is obtained by averaging over the last 50 simulation years.*

3.1 Climate Sensitivity in PlaSim

After the successful tuning of the new Tiedtke scheme, we determined the ECS. Table 1 shows the ensemble mean ECS for each of the four configurations. In general, if the sea ice module in PlaSim is switched on, ECS values are higher. The direct comparison between the two different convection schemes (with the same configuration of compartments) yields larger ECS values for the Kuo compared to the Tiedtke convection scheme.

The experiments with enabled horizontal heat diffusion (#N = 5,6,7,8) show lower values for the ECS. The strongest decrease in ECS, with a value of ΔECS = 0.955 ◦C, is obtained for the Kuo model, coupled to sea ice and ML ocean compartment. In contrast, the smallest decrease of ΔECS = 0.339 ◦C is obtained for the Tiedtke model when coupled to the ML ocean.

Four parameters were found to strongly influence ECS (Figure 3): *gamma, tswr2, drhssea* and *hdiffk*. *gamma* describes how much of the precipitation reevaporates. *tswr2* affects the cloud back scattering. *drhssea* is the wetness factor for ice-free ocean, and *hdiffk* is the horizontal heat diffusion coefficient K_h. We refer to the PlaSim User's Guide [3] and the source code for further information on these parameters. The range of ECS values obtained for one parameter is always larger for the Kuo convection compared to the Tiedtke convection. Furthermore, besides *drhssea*, the ECS's dependence on the parameters looks similar for the remaining three parameters for both types of convection. Panel (b) shows that the ECS exhibits asymptotic behaviour for *tswr2* for both convection schemes. This leads to high ECS values, when approaching the asymptote from low values. By contrast, non-physically low values of −30 ◦C are reached when approaching from high *tswr2* values. These parameter values have been tested and were found to result in unrealistic climatologies.

#N	Convection scheme	Compartments	Horizontal heat diffusion	ECS [°C] (± SD)	IPCC range [7]
1	Tiedtke	Atmosphere + ML Ocean	No	1.973 ± 0.049	✓
2	Tiedtke	Atmosphere + ML Ocean + Ice	No	3.131 ± 0.062	✓
3	Kuo	Atmosphere + ML Ocean	No	2.565 ± 0.069	✓
4	Kuo	Atmosphere + ML Ocean + Ice	No	6.845 ± 0.140	✗
5	Tiedtke	Atmosphere + ML Ocean	Yes	1.634 ± 0.042	✗
6	Tiedtke	Atmosphere + ML Ocean + Ice	Yes	2.572 ± 0.043	✓
7	Kuo	Atmosphere + ML Ocean	Yes	1.955 ± 0.052	✓
8	Kuo	Atmosphere + ML Ocean + Ice	Yes	5.890 ± 0.101	✗

Figure 3: *Parameter dependence of the ECS. Each parameter was randomly varied within its predefined bounds, while leaving the other parameters at their default values. Those values were then used for a 100- year-long simulation, including a doubling of CO_2 concentrations after 50 years. The first row shows the results for the Tiedtke convection in combination with the sea ice module and the second one for the Kuo convection. Dotted vertical lines refer to the default value of each parameter. Dotted horizontal lines refer to the default ECS value for the given PlaSim configuration.*

4. Discussion and outlook

We extended the Planet Simulator by a simplified Tiedtke convection scheme and tuned it via Bayesian Optimization to present-day climate. The model tuning significantly decreased differences between the model output and ERA5 reanalysis data, especially for the TOA radiation imbalance. However, some known model deficiencies, like the dry bias along the equator, could not be removed with the new convection scheme. Future work could address this problem by using a more sophisticated Tiedtke scheme, instead of the simplified version used here.

The ECS for PlaSim changes with convection schemes, with enabled or disabled sea ice modules and with activated horizontal heat diffusion using distinct values for the NH and SH. The Tiedtke convection yields smaller ECS values compared to the Kuo convection. The sea ice module increases the ECS. Three out of four Tiedtke configurations show ECS values within the *very likely* range of the ECS as given by the AR6 [7].

The default Kuo convection coupled to the ML ocean and sea ice modules yields an ECS exceeding this range. Switching on the horizontal heat diffusion leads to a more realistic present-day climate [11] and also reduces the ECS for this configuration. However, it is still above the upper bound. This stresses the possibility of having models with high ECS values, exceeding the AR6 range, which still produce realistic climatologies [8]. Moreover, the four most influential parameters on the ECS were investigated. Overall, the Kuo scheme reacts more sensitive to changes in these parameters. Besides the *drhssea* parameter, all show a similar pattern for both convection schemes. Exploiting the whole parameter space reveals non-physical behaviour of the ECS, including negative values or an asymptote for the *tswr2* parameter, which leads to unrealistic climatologies. Therefore, future work should define a metric to exclude unrealistic parameters to obtain a more plausible range of ECS values, similar to the work done by Schneider von Deimling *et al.* [8]. Nevertheless, this study shows how sensitive the ECS of a climate model can react to even small changes in parameters. This is of special interest for model tuning, when multiple sets of parameters are possible, which yield similar global mean climate fields but distinct ECS values. Such a situation would imply the existence of distinct model configurations consistent with present-day observations but resulting in substantially different future projections and would need to be accounted for.

5. Acknowledgements

We thank Frank Lunkeit and Heather Andres for helpful discussions about PlaSim, Oliver Mehling for his expertise on model tuning of PlaSim and the whole SPACY group. Furthermore, we acknowledge support by the state of Baden-Württemberg for support from the bwHPC project and for the usage of the bwUniCluster 2.0, which is part of the High Performance Computing infrastructure. The code for the new convection scheme will be made public with the accepted version of this extended abstract.

References

[1] F. Molteni. Atmospheric simulations using a GCM with simplified physical parametrizations. I: model climatology and variability in multi-decadal experiments. *Climate Dynamics*, 20(2):175–191, January 2003.

[2] T. Stocker. Introduction to climate modelling. https://climatehomes.unibe.ch/~stocker/papers/stocker21physik2.pdf, 2021. last accessed on 30.09.2022.

[3] F. Lunkeit, S. Blessing, K. Fraedrich, H. Jansen, E. Kirk, U. Luksch, and F. Sielmann. User's Guide Ver- sion 16.0. https://www.mi.uni-hamburg.de/en/arbeitsgruppen/theoretische-meteorologie/ modelle/sources/psusersguide.pdf. last accessed on 30.09.2022.

[4] F. Lunkeit, H. Borth, M. Böttinger, K. Fraedrich, H. Jansen, E. Kirk, A. Kleidon, U. Luksch, P. Paiewonsky, S. Schubert, F. Sielmann, and H. Wan. Planet simulator reference manual version 16.0 https://www.mi.uni-hamburg.de/en/arbeitsgruppen/theoretische-meteorologie/modelle/sources/psreferencemanual-1.pdf. last accessed on 30.09.2022.

[5] M. Tiedtke. A Comprehensive Mass Flux Scheme for Cumulus Parameterization in Large-Scale Models. *Monthly Weather Review*, 117(8):1779–1800, August 1989. Publisher: American Meteorological Society Section: Monthly Weather Review.

[6] R. Knutti, M. A. A. Rugenstein, and G. C. Hegerl. Beyond equilibrium climate sensitivity. *Nature Geoscience*, 10(10):727–736, October 2017.

[7] IPCC. *Climate Change 2021: The Physical Science Basis. Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2021.

[8] T. Schneider von Deimling, H. Held, A. Ganopolski, and S. Rahmstorf. Climate sensitivity estimated from ensemble simulations of glacial climate. *Climate Dynamics*, 27(2-3):149–163, August 2006.

[9] H. L. Kuo. On Formation and Intensification of Tropical Cyclones Through Latent Heat Release by Cumulus Convection. *Journal of the Atmospheric Sciences*, 22(1):40–63, January 1965. Publisher: American Meteorological Society Section: Journal of the Atmospheric Sciences.

[10] H. L. Kuo. Further Studies of the Parameterization of the Influence of Cumulus Convection on Large- Scale Flow. *Journal of the Atmospheric Sciences*, 31(5):1232–1240, July 1974. Publisher: American Meteorological Society Section: Journal of the Atmospheric Sciences.

[11] M. Angeloni, E. Palazzi, and J. von Hardenberg. Evaluation and climate sensitivity of the PlaSim v.17 Earth SystemModel coupled with ocean model components of different complexity. preprint, Climate and Earth system modeling, October 2020.

[12] A. Parnell, H. Andres, O. Erokhina, et al. Paleoplasim 1.0 – description, tuning and validation of a transient version of the PlanetSimulator, a GCM of intermediate complexity. in prep.

[13] K. Fraedrich, H. Jansen, E. Kirk, U. Luksch, and F. Lunkeit. The Planet Simulator: Towards a user friendly model. *Meteorologische Zeitschrift*, 14(3):299–304, July 2005.

[14] The COMET Program/MetEd. Nwp essentials: Precipitation and clouds. https://www.meted.ucar.edu/education_training/lessons/1157. last accessed on 14.09.2023.

[15] H. Park and S.-Y. Hong. An Evaluation of a Mass-Flux Cumulus Parameterization Scheme in the KMA Global Forecast System. *2*, 85(2):151–169, 2007.

[16] T. Mauritsen, B. Stevens, E. Roeckner, T. Crueger, M. Esch, M. Giorgetta, H. Haak, J. Jungclaus, D. Klocke, D. Matei, U. Mikolajewicz, D. Notz, R. Pincus, H. Schmidt, and L. Tomassini. Tuning the climate of a global model: TUNING THE CLIMATE OF A GLOBAL MODEL. *Journal of Advances in Modeling Earth Systems*, 4(3), March 2012.

[17] G. Lyu, A. Köhl, I. Matei, and D. Stammer. Adjoint-Based Climate Model Tuning: Application to the Planet Simulator. *Journal of Advances in Modeling Earth Systems*, 10(1):207–222, January 2018.

[18] O. Mehling, E. Ziegler, H. Andres, M. Werner, and K. Rehfeld. Parameterization dependence of the hydrological cycle in a general circulation model of intermediate complexity. Technical Report EGU21- 1328, Copernicus Meetings, March 2021. Conference Name: EGU21.

[19] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, October 2001.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duch- esnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[21] E. Schulz, M. Speekenbrink, and A. Krause. A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, 85:1–16, August 2018.

The Dynamics of Adult Neurogenesis in the Dentate Gyrus of the Hippocampus

Aadhar Sharma[1], Stefan Rotter[1]

[1]Bernstein Center Freiburg & Faculty of Biology, University of Freiburg, Germany

Abstract

As a rule, the adult brain does not generate new cells. However, in some animal species including humans, neurons are generated and integrated into the dentate gyrus of the hippocampus throughout the animal's lifetime. Since the dentate acts as the gateway to the hippocampus, the addition of new cells might affect hippocampal functions such as memory formation and contextual pattern separation. We have developed a biologically constrained model that allows us to study the dynamics of adult neurogenesis in the dentate gyrus more comprehensively than ever before. Through large-scale simulations, we were able to confirm the long-standing hypothesis that age-dependent properties of newborn cells are crucial for their successful integration into pre-existing networks. We also found that low rates of adult neurogenesis guarantee stable growth and prevent pathologies. Finally, our findings support previous experimental observations that newborn cells seem to compete with mature cells for synaptic input. The resulting synaptic redistribution might have interesting implications for the computations performed by hippocampal circuits.

1. Introduction

The foundation of the brain's neuronal architecture is laid during early development, but the cellular composition of the healthy adult brain no longer changes significantly. Nevertheless, in some species, new neurons are constantly being born in specialised regions throughout the animal's entire lifespan [1]. One of the two known neurogenic niches resides in the dentate gyrus and generates its principal neurons known as granule cells. Located within the hippocampal formation, the dentate acts as a regulatory gateway for the flow of information through it. Therefore, it is widely believed that adult neurogenesis in the dentate gyrus potentially contributes to various hippocampal functions, including memory formation, spatial navigation, and contextual pattern separation, among others [2, 3]. However, the precise role of newly generated neurons and their impact on the computational properties of the hippocampus remains unclear.

Newborn granule cells exhibit characteristic morphological and electrophysiological features [3-6]. As these adult-born cells gradually mature, their properties progressively resemble those of fully mature cells. However, despite the convergence of age-dependent properties upon maturation, it cannot be assumed that adult-born cells partake in the same computations as mature cells. Interestingly, adult-born cells compete for synaptic input with mature cells, leading to a redistribution of synaptic connections [7]. This not only highlights the high degree of plasticity within the dentate network, but also suggests that adult-born cells may be involved in distinct computational processes.

We hypothesised that the various processes involved in maturation, specifically the convergence of age-dependent cell properties, play a role in the integration of adult-born neurons into the pre-existing networks. To investigate the role of these parameters in integration dynamics and computational properties, we have developed a computational model of adult neurogenesis in the dentate gyrus. Following a maturation process that closely resembles experimental observations [4-6,11], adult-born granule cells are introduced into a simplified hippocampal network where plastic connections are established under the control of homeostatic structural plasticity [8,9].

Results indicate that adult-born cells gradually mature, establish plastic connectivity, and robustly integrate into the dentate circuits avoiding runaway growth. Age-dependent properties play a crucial role in facilitating this integration. Additionally, our analysis of network dynamics indicates that pathological states resembling epilepsy may emerge if large numbers of cells are rapidly added. The detailed study of connectivity further suggests that, although network integration has been achieved, newborn cells keep competing with mature cells for synaptic resources. Such synaptic redistribution holds implications for the computations performed in hippocampal circuits.

2. Methods

2.1 Network properties

The goal is to investigate the integration dynamics of adult-born granule cells into the circuits of the dentate gyrus. To achieve this, the hippocampal networks are simplified into two populations of excitatory-inhibitory networks, which represent the principal cells, diverse interneurons, and complex connectivity motifs of the entorhinal cortex (EC) and dentate gyrus (DG).

Within EC and DG, static connections are established randomly between the excitatory and inhibitory subpopulations (solid arrows in Fig. 1(a)). In contrast, plastic connectivity (dashed arrows) is established through homeostatic structural plasticity, where connectivity is regulated such that neurons fire at their specified firing rate set point. For each excitatory population (EC_E, DG_{GC}, and DG_{abGC}), the recurrent connections are made structurally plastic. Similarly, reciprocal connections between the adult-born (DG_{abGC}) and developmentally-born granule cells (DG_{GC}) are also structurally plastic. The plastic connections originating from the EC_E and leading to the granule cell subpopulations, known as the perforant pathway, form the main input to the DG. In our model, each subpopulation receives additional stimulation from a Poisson drive that resembles the background activity within the EC and DG.

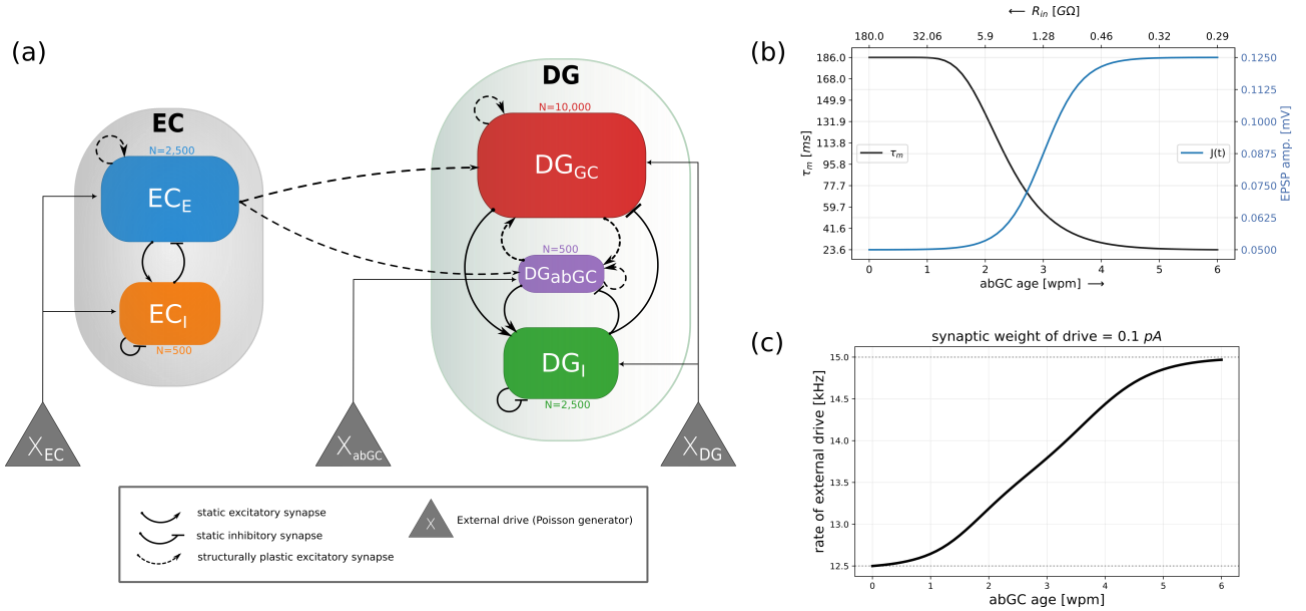


Figure 1: (a) *The EC—DG network comprises two excitatory-inhibitory populations. Solid lines indicate static connections while dashed lines represent plastic connectivity. The numbers indicate the respective neuron counts in each subpopulation.* ***(b)*** *The age-dependent profiles for membrane time constant (τ_m) and synaptic weight represent the response amplitude of unitary inputs. The top x-axis indicates that input resistance (R_{in}) also varies with age.* ***(c)*** *As adult-born cells undergo morphological growth, it is expected that their background inputs will increase as well. Therefore, the membrane capacitance profile for adult-born cells is adjusted to represent their age-dependent background input.*

2.2 Modelling adult neurogenesis

As mentioned before, adult-born cells exhibit distinct age-dependent properties. Newborn cells exhibit properties such as large input resistance, slow membrane time constant, and small depolarising synaptic currents [4-6,11]. Additionally, their early connectivity is limited to receiving excitatory and inhibitory inputs solely from local interneurons, but they gradually integrate into the full network over time [12]. We model these age-dependent functional properties by updating the membrane time constant, synaptic weight, and the rate of the external drive of adult-born cells as they mature (see Fig. 1(b, c)). These age-dependent profiles closely align with experimental findings [4,11].

At the onset of the simulation, the DG_{abGC} population is absent from the network and the rest of the EC–DG network self-organises through homeostatic structural plasticity. This creates a stable plastic network which resembles the pre-existing EC–DG circuits. At this stage, newborn granule cells are introduced into the network. Initially, these newborn cells have parameters that correspond to zero weeks post mitosis (wpm). As the simulation progresses and new dynamic equilibria are established, newborn cells are incrementally aged with an arbitrary resolution and provided parameters which correspond to the new age. This maturation process continues until adult-born cells are six weeks old, indicating full maturity, and have parameters that are identical to those of developmentally born cells (DG_{GC}).

2.3 Simulation setup

We simulated networks comprising 12 000 to 35 500 leaky integrate-and-fire neurons with structural plasticity using the NEST simulator [8,10]. The NEST kernel supports multithreaded and distributed simulations for cluster computers. Due to the very high computational demands of the project, running all simulations and subsequent analyses on a large cluster computer like the bwForCluster NEMO was an indispensable requirement. A typical job involving 16 000 neurons that requested 80 processors and 320 GB of main memory took approximately two days to complete. We stored simulation results in NEMO workspaces as compressed HDF5 files which, taken all together, occupy 2 TB of secondary storage. The process of structural plasticity creates the primary computational load involving millions of primitives per neuron. Additionally, simulating adult neurogenesis requires multiple maturation stages, which results in exceedingly long simulation times. Given these constraints and requirements, the bwForCluster NEMO plays a pivotal role in enabling our current and related future work.

3. Results & Discussions

Prior to the onset of adult neurogenesis, the activity-dependent self-organisation of the base EC–DG network is finalised (see Fig. 2(a)). The firing of EC_E and DG_{GC} occurs at their designated set points, demonstrating stable connectivity dynamics. The introduction of DG_{abGC} at time 0, indicating the onset of adult neurogenesis, perturbs the existing equilibria. However, the structural plasticity homeostat eventually establishes new equilibria. Adult-born cells are matured in increments of half a week until six weeks. Firing rate homeostasis eventually occurs at each maturation stage and the network dynamics remains stable. Interestingly, the structural plasticity homeostat over-corrects the maturation perturbation at 3.5 wpm by transiently deleting all perforant pathway synapses to the adult-born cells. Nonetheless, increased maturation resolutions resolve this behaviour and the dynamics exhibit smooth stabilisation.

Analysis of activity dynamics reveals that mature adult-born cells fire in an asynchronous-irregular regime, similar to DG_{GC}. Furthermore, the distributions of firing rate, irregularity, cytosolic calcium concentration, and membrane potentials between DG_{GC} and mature DG_{abGC} do not display significant differences (not shown). Finally, the perforant pathway connectivity formed by mature adult-born cells is at levels comparable to that of DG_{GC}. The convergence of dynamics suggests that adult-born cells have been successfully integrated into the EC–DG network in a smooth and robust manner.

To investigate the importance of age-dependent parameters for network integration, we performed experiments where the parameters of adult-born cells do not vary with age. The absence of age-dependent parameters led to network dynamics resembling pathological states. Therefore, we think that a gradual and sufficiently slow maturation of the functional properties makes an important contribution to the successful integration of adult-born cells into the DG circuits.

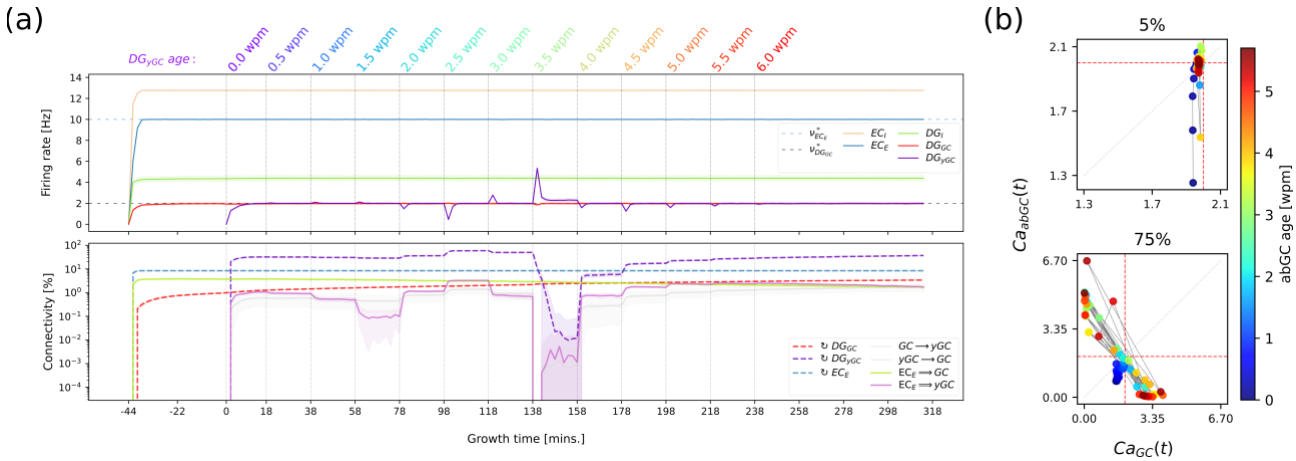


Figure 2: (a) *Dynamics of adult neurogenesis. The top panel illustrates the activity dynamics, while the bottom panel represents the connectivity dynamics. Adult-born cells are introduced at time 0 and age by 0.5 weeks until they reach 6 weeks of age. Firing rate homeostasis is observed at each maturation stage, and after 258 minutes, the activity dynamics fully stabilise at the set point. Although the connectivities of adult-born (yGC) and developmentally-born (GC) cells vary throughout the stages, they eventually reach comparable equilibria.* ***(b)*** *The cytosolic calcium concentration of developmentally-born (x-axis) and adult-born (y-axis) is normalised to the firing rate set point at 2 Hz. The colour of the data points represents normalised calcium levels after the onset of neurogenesis. In the 5% network, both populations smoothly equilibrate at their joint set point. However, in the 75% network, strong switching dynamics are observed as the calcium levels of the two populations fluctuate around the joint set point.*

The rate of adult neurogenesis in an animal is influenced by various factors, but typically, fewer than 10% of all granule cells in the dentate gyrus are adult-born [13-15]. To investigate the impact of different rates on network dynamics, we conducted simulations with varying proportions of adult-born to developmentally-born cells. Our findings revealed that networks with more than 10% adult-born cells exhibit increasing non-linear switching dynamics [16]. In contrast, networks with lower rates (<10%) barely showed any switching (see Fig. 2(b)). Switching dynamics indicate a competitive interaction between the two cell populations, where one population transiently fires at very high rates while suppressing the other. With increasing relative sizes, the duration of these transient states also exponentially increases. These dynamics bear a resemblance with epilepsy, providing a potential explanation for the observed low rates of adult neurogenesis in the dentate gyrus.

Granule cells are known to lack direct recurrent connections and instead receive indirect feedback through excitatory interneurons called mossy cells. Therefore, to investigate whether there is a redistribution of perforant pathway synapses between adult-born and developmentally-born granule cells, we disabled the direct feedback for adult-born cells. Unsurprisingly, in the absence of feedback, adult-born cells formed more perforant pathway synapses (not shown). Furthermore, we observed that no synapse redistribution occurs within the developmentally-born population where direct recurrence is allowed (not shown). This suggests that the only way adult-born cells can form perforant pathway connections is by appropriating the synapses of developmentally-born cells. While this experiment predicts the appropriation of existing perforant pathway synapses by adult-born cells, further research

is required to definitively support this claim. Nevertheless, the competition-based synaptic redistribution has implications for the computational properties of the hippocampal circuits.

4. Conclusion

Our experiments provide evidence of the robust integration of adult-born cells into the existing EC—DG network. The functional integration of these cells is significantly influenced by their age-dependent properties. When neurogenesis occurs at low enough rates, adult-born cells smoothly integrate into the dentate gyrus. However, the rapid addition of a large number of cells into the DG leads to the emergence of pathological states. Finally, we predict that a competitive synapse redistribution takes place between adult-born and developmentally-born cells. Our findings align with experimental observations and contribute valuable insights into the dynamics of adult neurogenesis.

Acknowledgement:

We thank Prof. Josef Bischofberger from the University of Basel for generously sharing experimental data and providing us with invaluable feedback and support. We also acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant no INST 39/963-1 FUGG (bwForCluster NEMO).

References

1. Fred H. Gage, Adult neurogenesis in mammals. *Science* 364, 827-828(2019)
2. T. Nakashiba et al., Cell 149, 188–201, issn: 0092-8674 (2012)
3. S. M. Miller, A. Sahay, Nature Neuroscience 22, 1565–1575, issn: 1097-6256 (2019)
4. C. Schmidt-Hieber, P. Jonas, J. Bischofberger, Nature 429, 184–187, issn: 0028-0836 (2004)
5. J. T. Gonçalves, S. T. Schafer, F. H. Gage, Cell 167, 897–914, issn: 0092-8674 (2016)
6. M. Lodge, J. Bischofberger, Behavioural Brain Research 372, 112036, issn: 0166-4328 (2019)
7. E. W. Adlaf et al., eLife 6, e19886 (2017)
8. S. Diaz-Pier, M. Naveau, M. Butz-Ostendorf, A. Morrison, Frontiers in Neuroanatomy 10, 57 (2016)
9. J. V. Gallinaro, S. Rotter, Scientific Reports 8, 3754, eprint: 1706 . 02912 (2018)
10. T. Fardet et al., NEST 2.20.0, Jan. 2020, (https://doi.org/10.5281/zenodo.3605514)
11. L. Li, S. Sultan, S. Heigele, C. Schmidt-Salzmann, N. Toni, J. Bischofberger, eLife 6, e23612 (2017)
12. T. Hainmueller, M. Bartos, Nature Reviews Neuroscience 21, 153–168, issn: 1471-003X (2020)
13. van Praag H, Kempermann G, Gage FH. Nat Neurosci. 1999 Mar; 2(3):266-70. PMID: 10195220. (1999)

14. A. Denoth-Lippuner, S. Jessberger, Nature Reviews Neuroscience 22, 223–236, issn: 1471-003X (2021)

15. P. Andersen, R. Morris, D. Amaral, T. Bliss, J. O'Keefe, The Hippocampus Book, isbn: 9780195100273 (2006)

16. F. Lagzi, S. Rotter, PLoS ONE 10, e0138947 (2015)

Planned Missing Data in Social Surveys: Evaluating Strategies Regarding Their Design and Imputation

Julian Axenfeld

Deutsche Institut für Wirtschaftsforschung (DIW) Berlin

Abstract

Surveys are facing pressures to shorten questionnaires: Long questionnaires are associated with low response rates, poor response quality, and are particularly considered inappropriate for the increasingly popular online mode. This is why survey designs with planned missing data, such as split questionnaire designs, are becoming more and more common in large-scale social surveys: They help reducing survey length by administering varying components of the whole questionnaire to each respondent. However, imputation may be needed to obtain reasonably analysable data with such a design. Yet, these data can be difficult to impute due to common features of social survey data, such as low correlations, predominantly categorical data, and relatively small sample sizes available to support imputation models with many potential predictor variables. In this extended abstract, I will discuss findings from a series of Monte Carlo simulation studies in which split questionnaire designs are simulated using real social survey data from the German Internet Panel and subsequently imputed. Estimates based on the imputed data are compared to population benchmarks to determine their accuracy. In the course of these studies, several different strategies regarding the design of the planned missing data and their imputation are examined in their effects on the accuracy of estimates.

1 Introduction and Theory

Long questionnaires in social surveys can lead to problems with data quality, such as low response rates and high breakoff (see for example Galesic et al. 2009) as well as increased measurement error (Peytchev and Peytcheva 2017). Furthermore, declining response rates (de Heer & de Leeuw 2002; de Leeuw et al. 2018) and the high financial costs of traditional survey modes make more and more surveys move to an online mode of data collection. Online surveys are comparatively cheap, but also require particularly short questionnaires. According to online survey respondents themselves, questionnaires should not exceed 25 minutes (Revilla and Höhne 2020). However, this would mean that moving a conventional (e.g., face-to-face) survey to an online mode may force researchers to drop lots of survey items that are potentially highly important to substantive researchers.

Split questionnaire designs (SQDs, Raghunathan and Grizzle 1995) could be a way to reconcile the need to collect data on many items with the need to keep questionnaires short for the individual respondent. SQDs entail that all survey items are allocated to (mutually exclusive) modules. Some items that are deemed especially important can be allocated to a core module, which is presented to all respondents alike. Each respondent further receives a randomly assigned subset of at least two of the other modules. Thereby, the questionnaire length for each respondent is reduced while still collecting data on all items and all bivariate relations between items.

However, this also yields considerable amounts of planned missing data for all respondents and for most variables. This means that conventional, simple procedures to deal with missing data, such as listwise or pairwise deletion, may not work with this type of data for many kinds of analyses, since it would yield datasets that are empty or have very low case numbers.

In consequence, Raghunathan and Grizzle (1995) propose to use multiple imputation (MI; Rubin 1987; van Buuren 2018) to complete data from SQDs and thereby make them analysable. By using multiple

imputation, each missing value is replaced with a number of *m* plausible values based on an imputation model. For the imputation model, one needs to specify a suitable imputation method and a set of relevant predictors. Finally, MI yields us multiple *m* imputed datasets, which are analysed separately with consequential estimates being pooled thereafter by using Rubin's Rules (Rubin 1987).

Ideally, the imputation of social survey data from an SQD would be performed before providing the data to substantive researchers. This would remove burden from individual researchers who might not be sufficiently familiar with multiple imputation or lack the computational resources to impute the data themselves. However, this also means the challenge for this study is to impute all variables in the survey and all relations between them at once rather than only a small fraction that is relevant for some specific research objective.

There are various other issues that may make imputing social survey data from SQDs especially challenging. First, presenting each respondent only a fraction of the whole questionnaire means that large amounts of data are missing by design and need to be imputed. This entails not only that the imputation models need to rely on a relatively small number of available cases. It also means that the imputations will have a big impact on the eventual substantive analyses of the data, implying that poor imputations could lead to invalid and unreliable inference from the data. Second, many variables need to be considered in the imputation model, especially with respect to the low case numbers. In theory, predictor sets in imputation models should at least cover all analysis variables (here: potentially all variables, since the analysis objective is unknown) as well as variables highly correlated to the imputed variable (Meng 1994; van Buuren 2018). Thus, we would need to include all variables in the data as predictors. However, estimating such complex imputation models accurately may be difficult considering that usually sample sizes are limited. Third, survey data in general often shows a tendency towards low correlations, suggesting potentially weak imputation models despite many predictor variables. Finally, variables from social surveys are often categorical rather than continuous, which restricts the number of imputation models available for this kind of data.

2 Data and methods

To determine which imputation strategies may work with planned missing data in a social survey, various methods were examined in a simulation study conducted via bwHPC (for the full detailed results see Axenfeld et al. 2022a).

First, the multivariate normal model (as included in the *Amelia* package in R; R Core Team 2021; Honaker et al. 2011) and Bayesian linear regressions (as included in the *mice* package (van Buuren and Groothuis-Oudshoorn 2011) may allow to construct relatively simple regression models, in which only one parameter per predictor variable needs to be estimated. However, these methods may be a poor fit for the many non-continuous variables in social survey data, for which the linearity and normality assumptions may not hold. Furthermore, imputed values would turn out continuous rather than discrete-categorical as observed in the real data. Amelia offers an option to transform imputed values to a discrete scale, but it remains unclear how this affects correlation estimates.

Second, generalised linear models for categorical variables (also included in *mice*; e.g., logistic regression) may fit the imputed variables' measurement level better, but previous research suggests that these methods in particular do not work well with a large number of predictors when sample sizes are limited (van Buuren 2018; White et al. 2011).

Third, predictive mean matching (PMM; also included in *mice*) is a technique that matches a missing data point with observed data points based on the predicted values for the imputed variable in a regression model, drawing one of the matched observed values as imputation. Compared to linear

regressions, this method is more robust for violations of the linearity and normality assumptions. Thus, it may be useful for imputing variables that are at least ordinal.

Finally, classification trees (also included in *mice*) are a simple decision-tree technique that imputes data from dividing the data into smaller and smaller subregions based on binary splits in the predictor space and drawing an observed value from the final, smallest subregion as imputed value. Unlike the other methods, this method has no assumptions at all regarding linearity or normality. However, limited sample sizes of real-world surveys allow only a certain number of consecutive splits before case numbers drop too low to allow for further splits. This suggests that classification trees make an implicit selection of predictor variables, and the relations of the imputed variable to the implicitly excluded predictor variables may be lost.

Furthermore, three different strategies regarding the definition of predictor sets were tested. A first approach is to always include all variables as predictors in the imputation model (i.e., unrestricted predictor sets). This may be the most valid approach from a theoretical perspective, but may also mean too complex imputation models in practice.

A second approach is to remove variables from the predictor set that are not correlated with the imputed variable (i.e., restricted predictor sets). This follows the idea that if these variables indeed have no relation to the imputed variable, nothing will be lost by excluding them.

A final approach is to reduce the dimensionality of the predictor space through partial-least-squares (PLS) regression (Robitzsch et al. 2016; included in the *miceadds* package, Robitzsch and Grund 2021). This is a technique similar to principal components analysis, reducing the predictor space to a few uncorrelated components that predict the imputed variable.

We also examined three strategies to design modules for an SQD: random modules, modules each covering a single survey topic, or modules each covering all different topics (for the full detailed results, see Axenfeld et al. 2022b). While single-topic modules may be considered useful from a questionnaire design perspective, they also imply that most highly correlated variables will always be both observed or both missing. The imputation, however, would likely benefit from highly correlated variables being allocated to different modules so that they can be used as predictors for the imputation model.

For the simulation study, data from wave 37 and 38 of the German Internet Panel (GIP; Blom et al. 2016a,b; 2019a,b; for further information see Blom, Gathmann & Krieger 2015, Blom et al. 2017, Cornesse et al. 2021) were used. The GIP is an online panel survey based on the German general population (size of the population dataset in this study: n=4,061). The 61 items included in this study were all categorical (mostly ordinal or binary). They covered socio-demographic characteristics and the sampling cohort (which were allocated to the core module) and on the respondents' membership in organizations, Big-Five personality traits, the influence of lobbying over EU politics, and domestic and party politics.

The simulation was conducted in several batch jobs on bwUniCluster 2.0 with over a thousand runs executed in parallel on 36 nodes of the former multiple_e queue (28 simulation runs per node). Each simulation run entailed several steps:

1. Draw 2,000 cases from our population dataset.
2. Simulate planned missing data i.e., drop observations. No missing data is generated on the 11 sociodemographic core items, while 40% missing data is created on the remaining 50 survey items by allocating them to five modules and assigning three of them to each respondent.
3. Impute the planned missing data (m=20 drawn after 10 iterations). To keep resource usage manageable, different batch jobs were used for each imputation method.

4. Calculate Spearman correlations between all variables based on the imputed data and compare them to population benchmarks.

The average deviations of the estimates from the population benchmarks (i.e., average Monte Carlo biases) were used to evaluate the different imputation strategies.

3 Results and discussion

The results indicate that several established imputation strategies perform poorly, especially (ordinal) logistic regression models and classification trees, destroying a considerable amount of correlations between variables. Correlations were preserved surprisingly well with Bayesian linear regressions and with *Amelia*. However, one should still bear in mind that the regression assumptions may be invalid and the imputed values will not be discrete. When continuous imputations were transformed to a discrete scale via *Amelia*, correlation estimates turned out less accurate. Predictive mean matching also resulted in moderate biases when using unrestricted predictor sets. Results with predictive mean matching could be improved a lot by reducing complexity in the imputation model, either using restricted predictor sets or partial-least-squares regression. Furthermore, the analysis showed that single-topic modules overall lead to inferior estimates compared to random modules or modules each covering all topics. Random modules and modules covering all topics showed a very similar performance, indicating that there might be only little potential for optimising modules for imputation beyond a random allocation of items to modules.

4 Acknowledgements

This work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) [project numbers: BL 1148/1–1, BR 5869/1–1, WO 739/20–1]. The Monte Carlo simulations were run on the High Performance Computing facilities of the state of Baden-Württemberg (bwHPC). This work uses data from the German Internet Panel (GIP) funded by the DFG through the Collaborative Research Center (SFB) 884 “Political Economy of Reforms” (SFB 884) [Project-ID: 139943784].

References

Axenfeld, J. B., Bruch, C., and Wolf, W. (2022a). General-purpose imputation of planned missing data in social surveys: Different strategies and their effect on correlations. *Statistics Surveys*, 16, 182-209.

Axenfeld, J. B., Blom, A. G., Bruch, C., and Wolf, W. (2022b) Split Questionnaire Designs for Online Surveys: The Impact of Module Construction on Imputation Quality. *Journal of Survey Statistics and Methodology*, 10, 1236-1262.

Blom, A. G., Bossert, D., Funke, F., Gebhard, F., Holthausen, A. and Krieger, U.; SFB 884 "Political Economy of Reforms" Universität Mannheim (2016a) German Internet Panel, Wave 1 - Core Study (September 2012). Cologne: GESIS Data Archive. ZA5866 Data file Version 2.0.0, https://doi.org/10.4232/1.12607.

Blom, A. G., Bossert, D., Gebhard, F., Funke, F., Holthausen, A. and Krieger, U.; SFB 884 "Political Economy of Reforms" Universität Mannheim (2016b) German Internet Panel, Wave 13 - Core Study (September 2014). Cologne: GESIS Data Archive. ZA5924 Data file Version 2.0.0, https://doi.org/10.4232/1.12619.

Blom, A. G., Fikel, M., Friedel, S., Höhne, J. K., Krieger, U., Rettig, T. and Wenz, A.; SFB 884 "Political Economy of Reforms", Universität Mannheim (2019a) German Internet Panel, Wave 37 - Core Study

(September 2018). Cologne: GESIS Data Archive. ZA6957 Data file Version 1.0.0, https://doi.org/10.4232/1.13390.

—(2019b) German Internet Panel, Wave 38 (November 2018). Cologne: GESIS Data Archive. ZA6958 Data file Version 1.0.0, https://doi.org/10.4232/1.13391.

Blom, A. G., Gathmann, C. and Krieger, U. (2015) Setting up an online panel representative of the general population: The German Internet Panel. Field Methods, 27, 391-408.

Blom, A. G., Herzing, J. M. E., Cornesse, C., Sakshaug, J. W., Krieger, U. and Bossert, D. (2017) Does the recruitment of offline households increase the sample representativeness of probability-based online panels? Evidence from the German Internet Panel. Social Science Computer Review, 35, 498-520.

De Heer, W. and De Leeuw, E. (2002) Trends in household survey nonresponse: A longitudinal and international comparison. Survey Nonresponse, 41, 41-54.

De Leeuw, E., Hox, J. and Luiten, A. (2018) International nonresponse trends across countries and years: An analysis of 36 years of Labour Force Survey data. Survey Insights: Methods from the Field. https://surveyinsights.org/?p=10452.

Galesic, M. and Bosnjak, M. (2009) Effects of questionnaire length on participation and indicators of response quality in a web survey. Public Opinion Quarterly, 73, 349-360.

Honaker, J., King, G. and Blackwell, M. (2011) Amelia II: A Program for Missing Data. Journal of Statistical Software, 45, 1-47.

Meng, X. L. (1994). Multiple-imputation inferences with uncongenial sources of input. Statistical Science, 9, 538-558.

Peytchev, A. and Peytcheva, E. (2017) Reduction of measurement error due to survey length: Evaluation of the split questionnaire design approach. Survey Research Methods, 11, 361-368.

Raghunathan, T. E., & Grizzle, J. E. (1995). A split questionnaire survey design. Journal of the American Statistical Association, 90(429), 54-63.

Revilla, M. and Höhne, J. K. (2020) How long do respondents think online surveys should be? New evidence from two online panels in Germany. International Journal of Market Research, 62, 538-545.

Robitzsch, A., and Grund, S. (2021) miceadds: Some Additional Multiple Imputation Functions, Especially for 'mice'. R package version 3.11-6.

Robitzsch, A., Pham, G., & Yanagida, T. (2016). Fehlende Daten und Plausible Values. In Breit, S., & Schreiner, C. (eds.) Large-Scale Assessment Mit R: Methodische Grundlagen Der Österreichischen Bildungsstandardüberprüfung, chapter 8, pp. 259-293. Facultas.

Rubin, D. B. (1987) Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons.

Van Buuren, S. and Groothuis-Oudshoorn, K. (2011) mice: Multivariate imputation by chained equations in R. Journal of Statistical Software, 45, 1-67.

Van Buuren, S. (2012) Flexible Imputation of Missing Data. Boca Raton: CRC press.

White, I. R., Royston, P. and Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine* 30:377-399.

Smallholder adaptation through agroforestry: Agent-based simulation of climate variability in Ethiopia

Habtamu Yismaw, Christian Troost, Thomas Berger

Department of Land Use Economics (490d), University of Hohenheim

Abstract

In this research, we employed household-level micro-simulation to assess the economic viability of acacia woodlot investments by smallholder farmers in Ethiopia. The decision-making process of these households integrates factors such as woodlot investment, crop production, and food security, set within an unstable environment marked by yield and price fluctuations and limited financial access. Using the agroeconomic agent-based modelling software MPMAS, we modelled this decision problem for a representative sample of 72 farms. Using bwHPC (specifically bwForCluster), we simulated household decisions over a decade, emphasising food security, income, and land use implications. Our first findings hint at considerable long-term economic benefits of Acacia woodlots, demonstrating robustness against price shocks.

1. Introduction

In recent decades, the upper Nile basin of Ethiopian Highlands has witnessed an unprecedented expansion of Acacia decurrens (green wattle) woodlots [1], [2]. Predominantly cultivated by smallholder farmers on former cropland, this tree species was initially introduced as a versatile solution for short-rotation forestry to counteract deforestation, cater to the escalating firewood demand, and enhance soil fertility [3]. Acacia decurrens is a leguminous, nitrogen fixing tree that helps reduce soil acidity and erosion and improve nitrogen and organic matter contents when grown over repeated cycles [2], [4]–[6]. Preliminary net present value (NPV) analyses validate the immediate economic advantage of adopting the acacia-based agroforestry system for regional smallholders, elucidating its widespread adoption [7]. In addition, this expansion led to the establishment of a charcoal value chain which brought off-farm employment opportunities for farmers and reduced seasonal migration for casual labour work in other parts of the country [6].

Nonetheless, while the immediate economic benefits are evident, the long-term sustainability and robustness of smallholder investments in acacia remains a subject of debate. Establishment of a woodlot commits farmer resources for a number of years. Continued woodlot expansion may lead to long-term declines of acacia product prices and emerging threats through acacia diseases; additionally, farmers grapple with weather and market risks for annual food crops and livestock they produce, all while still relying on their own production for food security. This study seeks to understand the farmers' capacity to navigate potential future disruptions in yield, price, and disease in the long term, and whether acacia investments fortify their resilience or inadvertently create detrimental dependencies. To this end, we employ farm-level microsimulation to simulate farming household production decisions and their repercussions across a decade under diverse scenarios.

2. Data and Methods

Smallholders investment in acacia is at least a four-year investment decision [7]. Seedlings of acacia are planted along with annual crops in the first year, and trees grow with pasture in the following years. Earliest at the end of the fourth year, farmers cut the trees, make charcoal and sell it to distributors or final consumers [6].

Investment in acacia – akin to livestock - hence constitutes a multi-year planning problem. Given the uncertainty of crop yields, prices and other shocks, multiple possible outcomes (states of nature) and their intertemporal effects have to be considered in farm investment planning. We conceptualised the agent decision problem as a mixed-integer programming (MIP) whole-farm model [8]–[10] with 15 years planning horizon and risk management. This MIP problem was used to parameterise the decision module of the agent-based simulation package - Mathematical Programming-based Multi-Agent Systems (MPMAS) [11], [12] for a sample of 72 farm agents representing smallholder farming households in the study area in the Awi Zone of the Amhara region. The district is the main acacia growing hot spot in the country. Information to initialize each agent was obtained from a comprehensive farm household survey collected in 2018 in Ethiopia's northwestern highlands. At the beginning of each simulation year of a 10 year simulation run[2], each farm agent solves its own version of the MIP reflecting their resources and available production and investment options at that point of time. Then MPMAS calculates production outcomes of the year depending on exogenous input on weather-specific yields, prices or the occurrence of plant diseases. The combined outcomes of decisions and their consequences determine the starting conditions for the agent decisions at the beginning of the next season (i.e. the model is recursive-dynamic [9]).

The farm agents' main goals in the decision problem are to satisfy their minimum demand for food covering other non-food minimum expenditures under all states of nature. Once these main goals are fulfilled, they maximize discretionary income[3]. Farm agents' decisions are constrained by their resource base, household labour, production and off-farm employment options, market access, human and social capital, and cultural constraints on labour and consumption. The decision model captures both the ex-ante decision situation where the agent plans for the coming season and beyond, and the ex-post decision situation where the agent has to cope with the potentially adverse outcomes of the previous cropping season. Time series data on yield and prices was collected from Ethiopia's central statistical agency (CSA). Land use decisions simulated by the model were validated against land use observations in the survey and by interactive validation experiments with local experts (similar to [8]).

To evaluate the economic sustainability of acacia production, we compared simulations for different expected price scenarios. Price shock scenarios were defined as combinations of percentage point changes of expected prices of tree products and crops and compared with a no price shock scenario. In addition, scenarios with and without the occurrence of acacia seedling disease shocks are simulated. Simulations were repeated varying parameters to reflect uncertainty in input data and model assumptions [13]. A total of 16 parameters and variables were selected for uncertainty analysis comprising crop and tree product yields, livestock output, output prices, input prices and financial parameters (interest rates, inflation rates and discount rate). Using the Sobol' quasi-random sequence an experimental design of 40 repetitions was defined to reflect the global uncertainty space defined by these 16 factors [9], [14].

[2] Simulations were conducted over a span of 10 simulation years. At the start of each simulation year, agents plan for a forthcoming 15-year term. Agent planning has a 15-year horizon to consider expected costs and benefits over the lifecycle of all investments (incl. perennials, oxen, forestry). A simulation time span of 10 years was considered sufficient to observe the actual outcomes of two acacia cycles of 4 years each.

[3] Discretionary income is household income after food and essential non-food expenditures are covered.

3. Simulation on bwHPC

Simulations were performed on bwForCluster MLS&WISO that is part of the Tier 3 of Baden-Württemberg's high-performance compute cluster infrastructure (bwHPC). The model implementation and simulation design offer various levels at which parallelization can be exploited: The IBM CPLEX library, which is used by MPMAS to solve individual MIP problems of farm agents, allows multi-threaded processing in the branch-and-cut algorithm. MPMAS allows solving the decision problems of several agents in parallel using openMPI [15]. And finally, the HPC cluster can be used to perform many simulation runs (i.e. scenario-uncertainty repetition combinations) in parallel. For the simulations presented here, only the last form of parallelization was used, proving most efficient at acceptable overall runtimes.

4. Results and discussion

Scenarios with and without investment in acacia are compared first to show the contribution of investment in acacia to income and land-use shares. Simulation results show that acacia takes more than 40% of land-use share in the area per year on average. Results also show that with investment in acacia agents achieve 8% higher per-capita discretionary incomes each year on average (Fig. 1). As a result, agents are able to fulfil the food and non-food expenditures better with acacia than without acacia, except for a few agents with smaller farm sizes in the first couple of simulation periods.

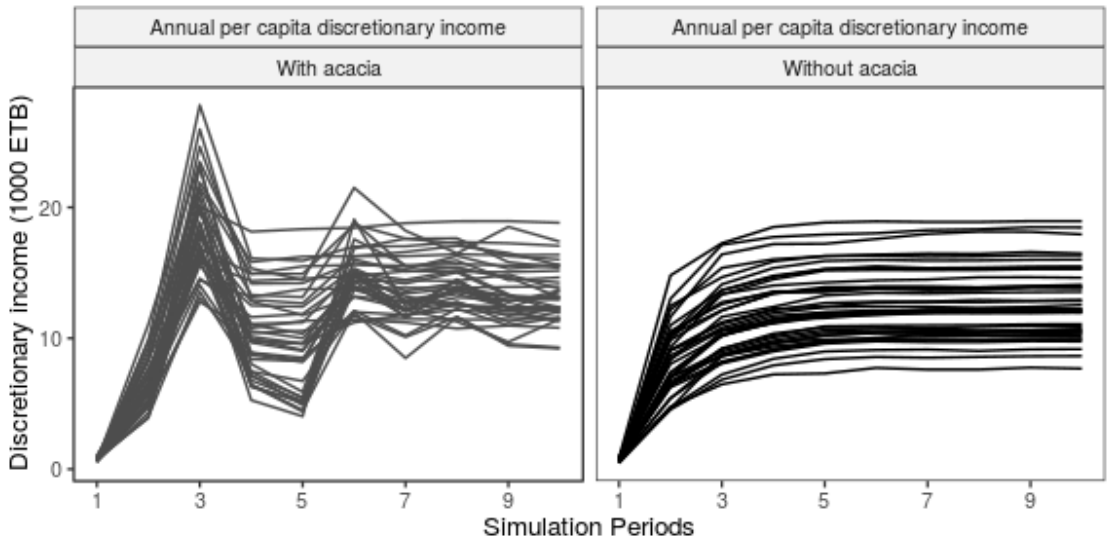


Figure 1: *Average annual per capita discretionary income in 40 model repetitions*

The effect of shocks on annual per capita discretionary income is presented in Figure 2. Occurrence of a single year outbreak of acacia seedling disease in the fourth period only reduces average per capita discretionary income by 2.7%. As the yield from investment in acacia would be realised four years after the seedling was affected, the average discretionary income decreases by 31% four years later. In the fourth year, when potato late blight and acacia seedling disease coincide, annual per capita discretionary income decreased by 27% in the shock year compared to the baseline scenario and an additional decrease of another 27% after four years because of late effects of acacia seedling disease.

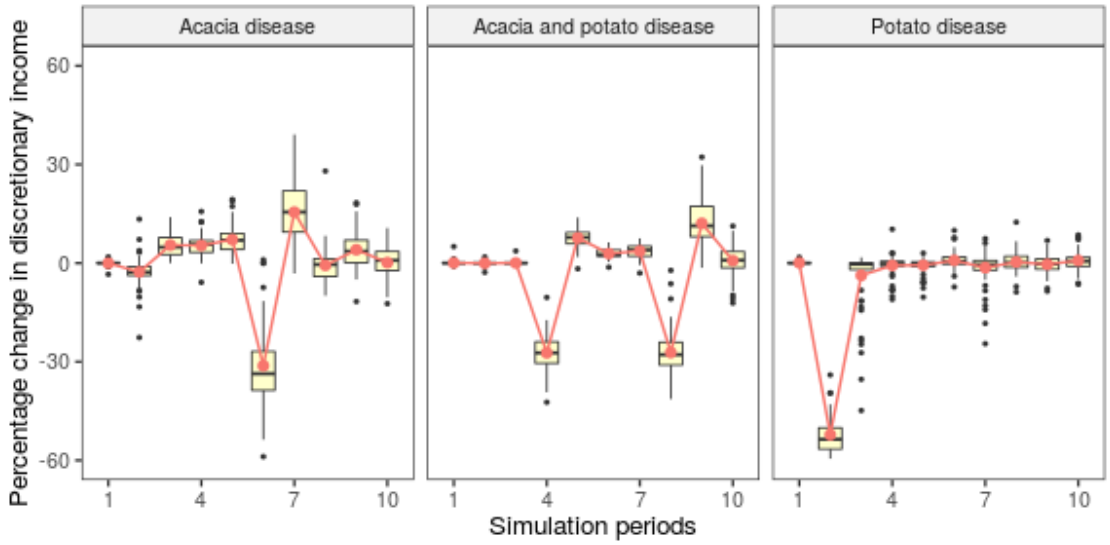


Figure 2: *Effects of shocks on discretionary income*

To understand the effect of price shocks, the percentage change of aggregated land-use share of each price scenario from the baseline is presented. First, acacia charcoal price was decreased by 50%. As a result, land-use share of acacia decreases by 26.5% on average. Whereas the aggregated land-use share of potatoes and wheat increases by 13.5% and 14.3%, respectively. The aggregated land-use shares of bamboo, teff and barley is still meager in this scenario and there is no change from the baseline. With the 50% decrease in expected charcoal price, agents would shift their acacia woodlots into croplands dominated by potatoes and wheat. Figure 3 shows the substitution of land-use from trees to crops.

Second, the expected prices of tree products were decreased (acacia charcoal and bamboo culms) by 50% simultaneously to see the effect on aggregated land-use shares in the future. The combined effect of price decrease in bamboo and acacia has a similar effect as the 50% decrease in charcoal prices alone. Aggregated land-use share of acacia decreases by 26.6% whereas potato and wheat share increase by 13.4% and 14.3% respectively. Provided that farmers can grow bamboo only in the farmstead, it is plausible to see if there is no significant change in aggregated land-use share irrespective of the price stimulus from bamboo.

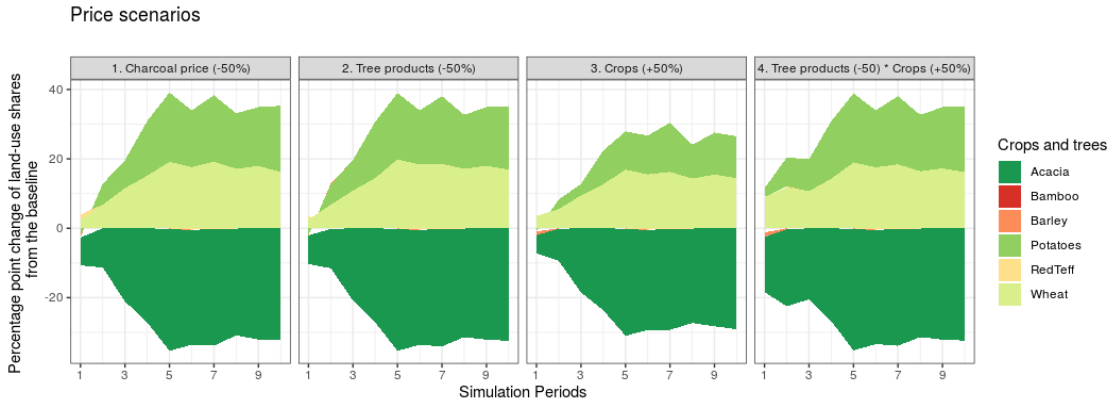


Figure 3: *Percentage point difference of land-use share for all price scenarios from the baseline*

Third, all expected prices of crops were increased by 50%, keeping other prices at the baseline value. Agents shift from forestry dominated to crop-dominated production system as a result – similar to the

results obtained in the 50% decrease in charcoal's expected price. Accordingly, the aggregated land-use share of acacia decreases by 23%, whereas it increases by 8.5% and 12.4% for potatoes and wheat, respectively. Finally, the expected price of crops was increased by 50% and decreased the expected price of tree products by 50% to see the agents' land-use decisions. The aggregated land-use of acacia decreases by 28.3%, while that of potatoes and wheat increased by 14.5% and 15%, respectively.

5. Conclusions

This study applied household level micro-simulation to analyse the economic sustainability of acacia woodlot investments by smallholder farmers in Ethiopia against shocks. The frequent and intense crop and tree diseases in the area – potato late blight and acacia seedling disease were introduced as shocks in the model. Simulation results show that both potato late blight and single-year acacia seedling disease reduce annual per-capita discretionary income significantly and force the resource-limited agents to fall short in meeting essential non-food expenditures. The trade-off in agents' land-use decisions between trees and crops by agents shows that agent still prefer to plant trees rather than crops as an ex-ante planning strategy for shocks.

The second shock that was examined was the effect of long-run expected price changes on agent land-use decisions. Simulation results show that agents are highly responsive to changes in expected prices in the long run, except for the expected price of bamboo. In cases where there is a decrease in the expected price of charcoal or an increase in the expected price of crops or both, simulation results show that agents will go back to potatoes and wheat-dominated production system instead of the acacia-dominated production system as in the baseline. This suggests that while acacia offers considerable benefits under prevailing economic conditions, farmers are not locked-in by their investment and are able to shift production fairly quickly, given the relatively short maturation time of acacia.

6. Acknowledgements

The authors acknowledge support by the state of Baden-Württemberg through bwHPC and by DFG through grant INST 35/1134-1 FUGG for the provision of high-performance computing resources. This research was part of a PhD Excellence Scholarship funded by DAAD via the EXCEED-Higher Education in Development Cooperation program and the Food Security Center at the University of Hohenheim.

References

[1] M. Wondie and W. Mekuria, "Planting of Acacia decurrens and Dynamics of Land Cover Change in Fagita Lekoma District in the Northwestern Highlands of Ethiopia," *mred*, vol. 38, no. 3, pp. 230–239, Aug. 2018, doi: 10.1659/MRD-JOURNAL-D-16-00082.1.

[2] T. Amare *et al.*, "Remediation of acid soils and soil property amelioration via Acacia decurrens-based agroforestry system," *Agroforest Syst*, vol. 96, no. 2, pp. 329–342, Feb. 2022, doi: 10.1007/s10457-021-00721-8.

[3] J. Sawyer, "Plantations in the Tropics: Environmental Concerns," 1993. Accessed: May 02, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Plantations-in-the-Tropics%3A-Environmental-Concerns-Sawyer/dd4e2be5a13fc4cec9436b243432c4c4e30fc304

[4] M. Kindu, G. Glatzel, Y. Tadesse, and A. Yosef, “Tree Species Screened on Nitosols of Central Ethiopia: Biomass Production, Nutrient Contents and Effect on Soil Nitrogen,” *Journal of Tropical Forest Science*, vol. 18, no. 3, pp. 173–180, 2006.

[5] M. L. Berihun *et al.*, “Exploring land use/land cover changes, drivers and their implications in contrasting agro-ecological environments of Ethiopia,” *Land Use Policy*, vol. 87, p. 104052, Sep. 2019, doi: 10.1016/j.landusepol.2019.104052.

[6] Z. Nigussie *et al.*, “The impacts of Acacia decurrens plantations on livelihoods in rural Ethiopia,” *Land Use Policy*, vol. 100, p. 104928, Jan. 2021, doi: 10.1016/j.landusepol.2020.104928.

[7] Z. Nigussie *et al.*, “Economic and financial sustainability of an Acacia decurrens-based Taungya system for farmers in the Upper Blue Nile Basin, Ethiopia,” *Land Use Policy*, vol. 90, p. 104331, Jan. 2020, doi: 10.1016/j.landusepol.2019.104331.

[8] J. Mössinger, C. Troost, and T. Berger, “Bridging the gap between models and users: A lightweight mobile interface for optimized farming decisions in interactive modeling sessions,” *Agric. Syst.*, vol. 195, p. 103315, Jan. 2022, doi: 10.1016/j.agsy.2021.103315.

[9] T. Berger, C. Troost, T. Wossen, E. Latynskiy, K. Tesfaye, and S. Gbegbelegbe, “Can smallholder farmers adapt to climate variability, and how effective are policy interventions? Agent-based simulation results for Ethiopia,” *Agric Econ*, vol. 48, no. 6, pp. 693–706, Jul. 2017, doi: 10.1111/agec.12367.

[10] P. Schreinemachers and T. Berger, “Land-use decisions in developing countries and their representation in multi-agent systems,” *Journal of Land Use Science*, vol. 1, pp. 29–44, 2006.

[11] P. Schreinemachers and T. Berger, “An agent-based simulation model of human–environment interactions in agricultural systems,” *Environmental Modelling & Software*, vol. 26, no. 7, pp. 845–859, Jul. 2011, doi: 10.1016/j.envsoft.2011.02.004.

[12] T. Berger and C. Troost, “Agent-based Modelling of Climate Adaptation and Mitigation Options in Agriculture,” *J Agric Econ*, vol. 65, no. 2, pp. 323–348, Jan. 2014, doi: 10.1111/1477-9552.12045.

[13] C. Troost and T. Berger, “Dealing with Uncertainty in Agent-Based Simulation: Farm-Level Modeling of Adaptation to Climate Change in Southwest Germany,” *Am J Agric Econ*, vol. 97, no. 3, pp. 833–854, Jan. 2015, doi: 10.1093/ajae/aau076.

[14] S. Tarantola, W. Becker, and D. Zeitz, “A comparison of two sampling methods for global sensitivity analysis,” *Computer Physics Communications*, vol. 183, no. 5, pp. 1061–1072, May 2012, doi: 10.1016/j.cpc.2011.12.015.

[15] C. Troost and T. Berger, “Advances in probabilistic and parallel agent-based simulation: Modelling climate change adaptation in agriculture,” *Proc. 8th iEMSs Congress*, Jan. 2016, [Online]. Available: https://scholarsarchive.byu.edu/iemssconference/2016/Stream-B/12/

Universal Dynamics at the Lowest Temperatures

Ido Siovitz, Philipp Heinen, Niklas Rasch, Stefan Lannig, Yannick Deller, Helmut Strobel, Markus Oberthaler, and Thomas Gasenzer
Kirchhoff-Institut für Physik, Im Neuenheimer Feld 227, Universität Heidelberg, Germany

Abstract

High-performance graphical processing units (GPU) are used for the repeated parallelised propagation of non-linear partial differential equations on large spatio-temporal grids. The main challenge results as a combination of the requirement of large grids for exploring scaling over several orders of magnitude, both in space and time, and the need for high statistics in averaging over many runs, in computing correlation functions for highly fluctuating quantum many-body states. With our simulations, we explore the dynamics of complex quantum systems far from equilibrium, with the aim of classifying their universal characteristics such as scaling exponents near non-thermal fixed points. Our results are strongly relevant for the development of synthetic quantum systems when exploring the respective physics in the laboratory.

1. Introduction

The characterization of complex physical structures, in particular when being far from equilibrium and exhibiting strong correlations, poses a long-standing challenge to physics. Of particular interest are universal aspects of equilibrium as well as non-equilibrium systems, which imply that experiments on one system can be relevant for the understanding of entirely different ones. An increasing interest lies also on the question to what extent complex structures could be used for developing new technologies for physical computation. A primary task is to develop theoretical and experimental tools for precisely controlling many-particle systems and observing their dynamics. Basic strategies comprise extending the range of predictive methods by studying prototypical models numerically. These models allow mutual benchmarks in terms of mathematical analysis, numerical simulations, as well as measurements on well-controlled synthetic quantum systems, in our context ultracold atomic gases. Such systems typically comprise a few thousand to million atoms cooled to temperatures in the nano Kelvin regime, where they exhibit quantum properties and in particular undergo a transition to a so-called Bose-Einstein condensate, forming a nearly incompressible superfluid that flows without friction. During recent years, we have developed a strong focus on far-from-equilibrium dynamics as well as on strong correlations in such systems. Our specific research includes the development and application of parallelised computing techniques, which ensure reaching the necessary statistical precision and system sizes for the characterisation of the considered dynamical phenomena and properties. Exemplary topics include the classification of universal quantum dynamics near non-thermal fixed points of the time evolution, the precise characterisation of strongly correlated equilibrium and non-equilibrium states, as well as long-time evolution and the approach to thermal equilibrium.

2. Simulations of ultracold superfluid quantum gases

In the following we give a brief introduction to the main aspects of universal dynamics near a non-thermal fixed point, for the example of a three-component ultracold atomic gas and discuss the challenges posed by statistical simulations of the highly fluctuating and correlated system, which we perform in a highly parallelised manner on clusters of graphical processing units (GPU).

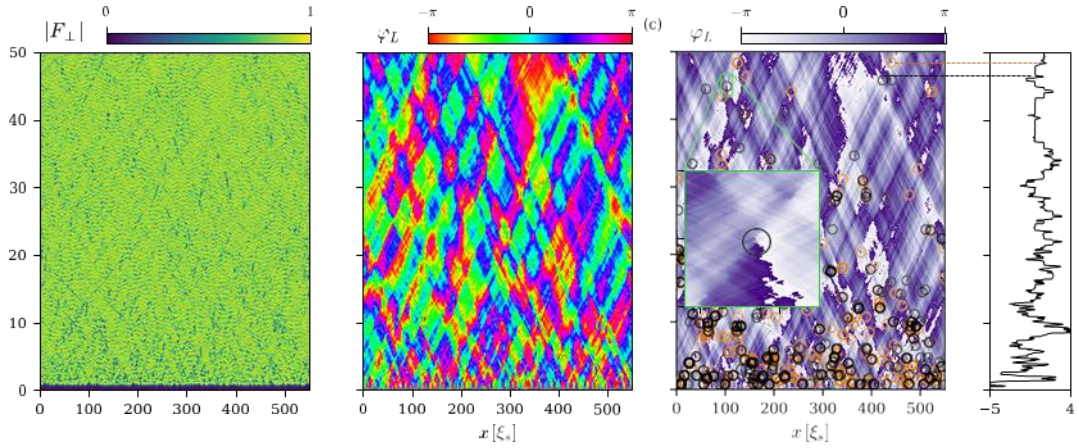


Figure 1: Space-time evolution of a spin $\mathrm{F}_\perp(t,x) = (F_x, F_y)$ of length $|F_\perp| = (F_x^2 + F_y^2)^{1/2}$ and orientation angle $\varphi\mathrm{L} = \arg(F_x + \mathrm{i}F_y)$ in the $F_x - F_y$-plane. Starting with an equilibrated but fluctuating system and suddenly changing some parameter, here an external magnetic field, introduces short-wave-length excitations which subsequently lead to the formation of textures in the transversal spin after a few characteristic time scales t_s. These spin textures, i.e., domains of different orientation (panel b), with overall constant spin length (a) are separated by kink-like defects and travel with a roughly uniform velocity in either direction. The size of the spin textures grows in time and goes together with a decreasing frequency of the occurrence of vortex-type defects in space and time (c) each of which changes the overall winding number $Q_w(t) = \int \mathrm{d}x\, \partial_x \varphi_\mathrm{L}(t,x)$ of the Larmor phase φ_L (d) along the system with a periodic boundary condition. See [20, 26] for more details.

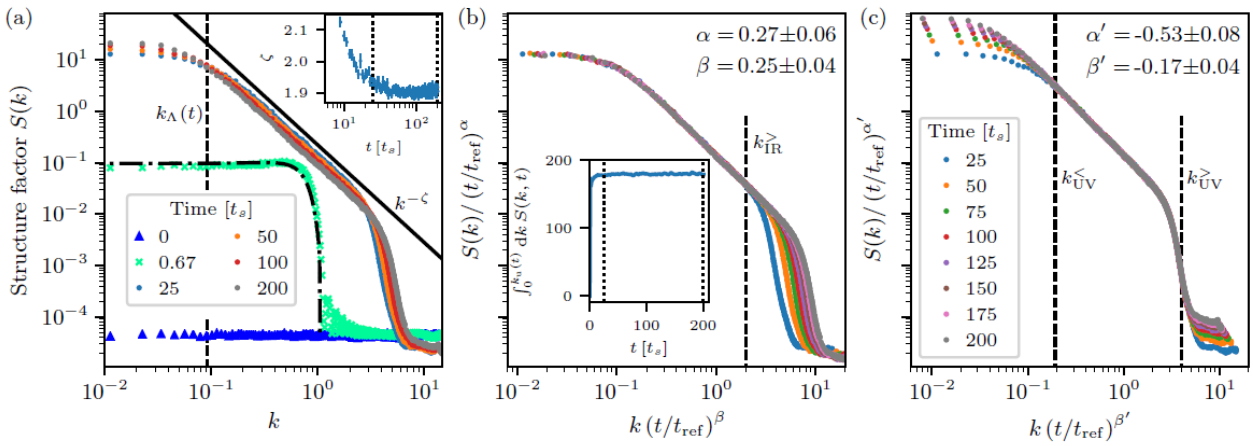
2.1 Universal dynamics far from equilibrium

Quantum many-body systems driven far out of equilibrium can show universal dynamical behaviour [1–25]. This includes, e.g., systems suddenly quenched to a non-equilibrium state and relaxing back to equilibrium. Universality typically means that correlations between fields evaluated at in general different space (and time) points show a simple form independent of the particular details of the system, defined only by symmetries of the underlying model [19]. They are commonly characterised by power laws in momentum space and frequency, and also their change in time reduces to a rescaling defined by a few universal exponents only. Examples include the evolution of ensembles of topological excitations and their quantum turbulent dynamics, phase-ordering processes, and other phenomena described by so far unknown non-thermal fixed points [9, 11, 19]. The goal is to understand which types of such dynamics there are, i.e., to find and characterise the universality classes of non-thermal fixed points, in order to learn about possible dynamics of similar type in entirely different system, ranging, e.g., from structure formation in the early universe, following the big bang, via

astrophysical objects such as neutron stars, to dynamics of ocean waves and microscopic phenomena in the solid state.

Consider, for instance, a one-dimensional Bose-Einstein condensate, which consists of bosonic atoms that can move along the *x* direction, and which can be in three different internal electronic states. Its state can be described by a spin vector field $\mathbf{F}(t,x) = (F_x, F_y, F_z)(t,x)$, which encodes the local density of atoms and the mean orientation of **F** representing the atoms' three-component quantum mechanical spin-1 state.

Fig. 1 shows an example of such a field $F_\perp \equiv F_x + \mathrm{i}F_y = |F_\perp| \exp\{\mathrm{i}\varphi_\mathrm{L}\}$ separated into spin



length $|F_\perp|$ and orientation angle φ_L between the vector (F_x, F_y) and the F_x-axis [20, 26], initially fluctuating weakly about a vanishing mean $<\mathbf{F}(0,x)>\equiv 0$. Both the length and orientation grow and thereby fluctuate in time. While (a) shows that the length rather quickly assumes a fairly constant value, (b) exhibits large patches of equal orientation, separated by sharp jumps in the angle. The jumps propagate with a fairly constant velocity forward and backward along the spatial coordinate, which is here measured in units of the so-called spin coherence length ξ_s. On average, the interval lengths of co-alignment of the spin slowly increase. Such a phenomenon is typically called domain coarsening and known to be associated with power-law evolution [27, 28]. Hence, there is a correlation length scale ℓ_Λ governing

Figure 2: Scaling evolution of the structure factor $S(k,t)$. (a) Following a fast initial growth of modes up to a certain maximum, the form of $S(k,t)$ approaches a universal shape, which only slowly evolves and rescales in time, Eq. (1), separately in the (b) IR regime of momenta, towards lower *p*, with $\beta \simeq 0.25, \propto\simeq 0.27$, thereby conserving the integral under the curve (see inset), and in the (c) UV, where a different set of (negative) exponents implies that the scaling is towards higher momenta. The shape of the curve becomes universal, in particular also exhibiting a constant power-law fall off $\sim k^{-\zeta}$ (see inset of panel a). The rescaling of the IR scale $k_\Lambda(t) \sim t^{-\beta}$ reflects the coarsening seen in Fig. 1, i.e., a self-similar growth of the spin-orientation pattern characterised by a correlation length $\ell_\Lambda \sim k_\Lambda^{-1}$. The exponents together with the scaling form f_s define the universality class the underlying non-thermal fixed point belongs to. Figure taken from Ref. [20].

the size of the patches of equal colour, which grows in time as $\ell_\Lambda(t) \sim t^{\beta}$, with the exponent being $\beta \simeq 1/4$ as found in [20]. More recent investigations indicate that the coarsening process goes along with the appearance of vortex-like defects in space and time (Fig. 1c) [26], which give rise to jumps of the winding number $Q_w = \int \mathrm{d}x \partial_x \varphi_\mathrm{L}(x)$ and which can be associated with

caustics appearing in the chaotic wave propagation, similar to rogue waves on ocean surfaces [29] (Fig. 1d).

Taking averages over many such evolutions $\mathbf{F}(t,x)$, starting from said randomly fluctuating initial configurations, we typically compute a so-called structure factor, e.g., $S(t,k) = (F_\perp(t,k) * F_\perp(t,k))$, of the Fourier transform $F_\perp(t,k)$ to momenta k, of the spin orientation $F_\perp(t,k) = (F_x + \mathrm{i}F_y)(t,k)$ in the $F_x - F_y$-plane. This quantity encodes the time evolving spectrum of excitations exemplarily shown in Fig. 1. Fig. 2 depicts the resulting averaged structure factor: Starting from a constant distribution over all momenta (blue triangles), during the early evolution, a strong buildup of excitations, up to a maximum momentum $k_Q \simeq 1$ set by the quench, is seen (green crosses), which, at late times, goes over into a universal form with a plateau in the infrared (IR), a power-law fall off $\sim k^{-\zeta}$ at larger momenta, and a steep fall-off at the ultraviolet (UV) end. This distribution shows scaling evolution according to

$$S(t,k) = (t/t_{\mathrm{ref}})^{\alpha} f_{\mathrm{s}}([t/t_{\mathrm{ref}}]^{\beta} k).$$

Here f_{s} represents the universal scaling function, i.e., the shape of $S(t,k)$ (considered separately in the IR, $k < k_{\mathrm{IR}}^{>}$, and UV, $k_{\mathrm{UV}}^{<} < k < k_{\mathrm{UV}}^{>}$), which depends only on the momentum k, t_{ref} is a reference time within the range of times where scaling prevails, and the scaling exponents characterise the IR (panel b) or UV (c) rescaling, with values indicated inside the respective panels. In the IR, they are related by $\propto = d\beta$, ensuring the momentum integral over $S(t,k)$ to be conserved in time.

2.2 Highly parallelised statistical simulations of the quantum field dynamics

The systems' dynamics is subject to non-linear coupled differential equations for multi-component quantum fields $\psi_m(t,x)$, in one or more spatial dimensions, which for the above 3-component system

$$\mathrm{i}\hbar\,\partial_t \begin{pmatrix} \psi_1 \\ \psi_0 \\ \psi_{-1} \end{pmatrix} = \left[-\frac{\hbar^2}{2m}\Delta + q + (c_0 + c_1)\rho + c_1 \begin{pmatrix} -2|\psi_{-1}|^2 & \psi_{-1}^*\psi_0 & 0 \\ \psi_{-1}\psi_0^* & -|\psi_0|^2 - q/c_1 & \psi_0^*\psi_1 \\ 0 & \psi_0\psi_1^* & -2|\psi_1|^2 \end{pmatrix} \right] \begin{pmatrix} \psi_1 \\ \psi_0 \\ \psi_{-1} \end{pmatrix}, \tag{2}$$

take the form where $\hbar, m, q$, and $c_{0,1}$ are real-valued constants and the fields $\psi_m(t,x)$ are complex and define the local density of atoms $\rho = \sum_{m=0,\pm1}|\psi_m|^2$. As in quantum mechanics, the three-component field defines a spinor, which encodes the mean orientation and strength of the angular momentum vector **F**, for which examples were shown in Figs. 1 and 2. To solve the equations, we use a split-step Fourier method, in which the operator in (2) in square brackets is split into three parts, each of which we use to propagate in one time step from t_n to t_{n+1} separately: The kinetic part involving the Laplacian $\Delta = \sum_{i=1}^{d}\partial_i^2$, which is readily diagonal in momentum space; the diagonal interaction part, which includes all terms except those in the off-diagonal elements of the matrix, and the off-diagonal interaction part, which accounts for component mixing. The first step involves two Fourier transforms $\mathcal{F}$, with a phase evolution over time step $\Delta t = t_{n+1} - t_n$ sandwiched in between, while the second and third steps involve a phase and matrix multiplication directly in position space,

$$\vec{\psi}'(x_j, t_{n+1}) \equiv (\psi_1, \psi_0, \psi_{-1})^T(x_j, t_{n+1}) = \mathcal{F}^{-1}\{e^{-\mathrm{i}\Delta t|k|^2/2}\mathcal{F}[\vec{\psi}(x_j, t_n](k)\}, \tag{3}$$

$$\vec{\psi}''(x_j, t_{n+1}) = \exp\{-\mathrm{i}[V_m(x_j) + f_m(\,|\psi_0(x_j,t_n)|^2, |\psi_1(x_j,t_n)|^2, \psi_{-1}(x_j,t_n)|^2)]\Delta t\}\vec{\psi}'(x_j, t_{n+1}), \tag{4}$$

$$\vec{\psi}(x_j, t_{n+1}) = \frac{1}{\lambda^2}\begin{pmatrix} |a|^2\cos(\lambda c_1 \Delta_t) + |b|^2 & -\mathrm{i}a\lambda\sin(\lambda c_1 \Delta t) & ab\cos(\lambda c_1 \Delta t) - 1 \\ -\mathrm{i}a^*\lambda\sin(\lambda c_1 \Delta t) & \lambda^2\cos(\lambda c_1 \Delta t) & -\mathrm{i}b\lambda\sin(\lambda c_1 \Delta t) \\ a^*b^*\cos(\lambda c_1 \Delta t) - 1 & -\mathrm{i}b^*\lambda\sin(\lambda c_1 \Delta t) & |a|^2 + |b|^2\cos(\lambda c_1 \Delta t) \end{pmatrix}\vec{\psi}''(x_j, t_{n+1}). \quad (5)$$

Here, V_m and f_m include the potential and diagonal non-linear parts of the operator in (2) in square brackets, and $\lambda = \sqrt{|a|^2 + |b|^2}$,with $a = \left[\psi_{-1}^*\psi_0 + \tilde{\psi}_{-1}\tilde{\psi}_0\right]/2, b = [\psi_0^*\psi_1 + \tilde{\psi}_0\tilde{\psi}_1]/2$, and where $\tilde{\psi}_m = \psi_m - \mathrm{i}S_{mn}\psi_n$ with *S* being the off-diagonal part of the matrix in (2).

The above propagation needs to be performed on sufficiently large spatial and temporal grids, many times in sequence, starting from slightly fluctuating initial conditions, in order to accumulate sufficient statistics for evaluating the average correlations such as shown in Fig. 2 and to take into account quantum fluctuations. This requires computational resources exceeding by far local PC clusters. As an example, in our 1D simulations, we chose a spatial grid of 4096 lattice points, each of which needs to hold 3 complex numbers, which results in a 3 × 2 × 4096 = 24 k grid of at least double-precision numbers. We start as many runs simultaneously on each GPU (Nvidia V100 und A100), as its RAM captures, and one such total run for the example case described above is over typically ca. 1.5M time steps. It lasts between 120 and 150s wall time, and thus it takes ~ 11 hrs for 250 runs, or for 1000 runs on a machine with 4 GPUs, typically needed for achieving well averaged distributions. As compared with computations on CPU clusters alone, the described combination has an advantage of roughly a factor of 10 for our purposes. 2D and 3D systems require smaller grids and less runs to average well, so far up to 256^3 × 3 × 2 in three dimensions, with larger grids being possible for systems with less internal components. All data evaluation is done in parallel on CPUs, and the code has been developed in C++ with CUDA and OpenMP parallelisation.

3. Discussion and Conclusions

Architectures of high-performance graphical processing units (GPU) are used for the repeated parallelised propagation of non-linear partial differential equations on large spatio temporal grids. The main challenge consists of a combination of several bottleneck-like limitations: On the one side, studies of universal dynamics require the evaluation of the time-evolving field configurations within large volumes *and* with high resolution, in order to span as many orders of magnitude in spatial extent as possible. This requirement results from the goal to detect self-similar scaling behaviour with sufficient precision. On the other hand, the considered dynamics far from thermal equilibrium gives rise to strong fluctuations and catastrophe-type events such as rogue waves and caustics, as well as non-linear excitations like topological defects. As a result, a sufficiently high temporal resolution is needed to capture their short-time characteristics, while typically only long evolution times give rise to scaling behaviour. Moreover, due to the strong fluctuations, many runs, starting from slightly different initial conditions, are required in order to achieve sufficient averaging statistics in evaluating correlation functions, i.e., moments of the underlying probability distributions.

The exemplary results shown here demonstrate the power of the parallelised evaluation of our simulations, which allows a speedup by up to a factor 10 as compared to standard OpenMP parallelisation on CPUs. So far, this has been considered to be of higher relevance than the limitations set by the RAM size of available high-end GPUs. Possible future extensions include a parallelisation of the lattice representation of a single system over several GPUs. Such an extension is technically feasible but represents a severe limitation for the type of split-step Fourier propagation we employ in propagating the differential equations. Techniques

developed for the grid-based numerical solution of partial differential equations such as DUNE [30] could help overcoming the RAM limitations but likely require to avoid split-step Fourier in order to reduce data exchange between different GPU units. This poses a challenge for the evolution of PDEs of the non-linear Schrödinger type, which involve topological defect solutions which are most easily captured within the chosen approach.

Even harder challenges prevail in computing full quantum fluctuation properties of many-body systems, both in and, even more so, out of equilibrium. Technically similar techniques can be used, when extended to twice as large configuration spaces, to determine properties of highly correlated quantum systems, e.g., near equilibrium phase transitions [31]. In summary, the high-performance computing facilities provided by bw|HPC represent a decisive element in efforts to explore, model, and understand fundamental characteristics of the physics of complex quantum systems, in view of near-future developments in controlling such systems in a tailored way in the laboratory.

Acknowledgements

The authors acknowledge support by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG), through SFB 1225 ISOQUANT (Project-ID 273811115), grant GA677/10-1, and under Germany's Excellence Strategy – EXC 2181/1 – 390900948 (the Heidelberg STRUCTURES Excellence Cluster), by the state of Baden-Württemberg through bwHPC and DFG through grants INST 35/1134-1 FUGG, INST 35/1503-1 FUGG, INST 35/1597-1 FUGG, and 40/575-1 FUGG.

References

[1] M. Prüfer, P. Kunkel, H. Strobel, S. Lannig, D. Linnemann, C.-M. Schmied, J. Berges, T. Gasenzer, and M. K. Oberthaler, Observation of universal quantum dynamics far from equilibrium, Nature **563**, 217–220 (2018).

[2] C. Eigen, J. A. P. Glidden, R. Lopes, E. A. Cornell, R. P. Smith, and Z. Hadzibabic, Universal Prethermal 131 Dynamics of Bose Gases Quenched to Unitarity, Nature **563**, 221–224 (2018).

[3] S. Erne, R. Bücker, T. Gasenzer, J. Berges, and J. Schmiedmayer, Universal dynamics in an isolated one-dimensional Bose gas far from equilibrium, Nature **563**, 225–229 (2018).

[4] G. Gauthier, M. T. Reeves, X. Yu, A. S. Bradley, M. A. Baker, T. A. Bell, H. Rubinsztein-Dunlop, M. J. Davis, and T. W. Neely, Giant vortex clusters in a two-dimensional quantum fluid, Science **364**, 1264–1267 (2019).

[5] S. P. Johnstone, A. J. Groszek, P. T. Starkey, C. J. Billington, T. P. Simula, and K. Helmerson, Evolution of large-scale flow from turbulence in a two-dimensional superfluid, Science **364**, 1267–1271 (2019).

[6] N. Navon, C. Eigen, J. Zhang, R. Lopes, A. L. Gaunt, K. Fujimoto, M. Tsubota, R. P. Smith, and Z. Hadzibabic, Synthetic dissipation and cascade fluxes in a turbulent quantum gas, Science **366**, 382–385 (2019).

[7] J. A. P. Glidden, C. Eigen, L. H. Dogra, T. A. Hilker, R. P. Smith, and Z. Hadzibabic, Bidirectional dynamic scaling in an isolated Bose gas far from equilibrium, Nature Phys. **17**, 457–461 (2021).

[8] A. D. García-Orozco, L. Madeira, M. A. Moreno-Armijos, A. R. Fritsch, P. E. S. Tavares, P. C. M. Castilho, A. Cidrim, G. Roati, and V. S. Bagnato, Universal dynamics of a turbulent superfluid Bose gas, Phys. Rev. A **106**, 023314 (2022).

[9] J. Berges, A. Rothkopf, and J. Schmidt, Non-thermal fixed points: Effective weak-coupling for strongly correlated systems far from equilibrium, Phys. Rev. Lett. **101**, 041603 (2008).

[10] J. Schole, B. 149 Nowak, and T. Gasenzer, Critical Dynamics of a Two-dimensional Superfluid near a Non-Thermal Fixed Point, Phys. Rev. A **86**, 013624 (2012).

[11] A. Piñeiro Orioli, K. Boguslavski, and J. Berges, Universal self-similar dynamics of relativistic and nonrelativistic field theories near nonthermal fixed points, Phys. Rev. D **92**, 025041 (2015).

[12] L. A. Williamson and P. B. Blakie, Universal Coarsening Dynamics of a Quenched Ferromagnetic Spin-1 Condensate, Phys. Rev. Lett. **116**, 025301 (2016).

[13] M. Karl and T. Gasenzer, Strongly anomalous non-thermal fixed point in a quenched two-dimensional Bose gas, New J. Phys. **19**, 093014 (2017).

[14] R. Walz, K. Boguslavski, and J. Berges, Large-*N* kinetic theory for highly occupied systems, Phys. Rev. D **97**, 116011 (2018).

[15] I. Chantesana, A. Piñeiro Orioli, and T. Gasenzer, Kinetic theory of nonthermal fixed points in a Bose gas, Phys. Rev. A **99**, 043620 (2019).

[16] A. N. Mikheev, C.-M. Schmied, and T. Gasenzer, Low-energy effective theory of nonthermal fixed points in a multicomponent Bose gas, Phys. Rev. A **99**, 063622 (2019).

[17] C.-M. Schmied, A. N. Mikheev, and T. Gasenzer, Prescaling in a Far-from-Equilibrium Bose Gas, Phys. Rev. Lett. **122**, 170404 (2019).

[18] A. Mazeliauskas and J. Berges, Prescaling and far-from-equilibrium hydrodynamics in the quark-gluon plasma, Phys. Rev. Lett. **122**, 122301 (2019).

[19] C.-M. Schmied, A. N. Mikheev, and T. Gasenzer, Non-thermal fixed points: Universal dynamics far from equilibrium, Int. J. Mod. Phys. A **34**, 1941006 (2019).

[20] C.-M. Schmied, M. Prüfer, M. K. Oberthaler, and T. Gasenzer, Bidirectional universal dynamics in a spinor Bose gas close to a nonthermal fixed point, Phys. Rev. A **99**, 033611 (2019).

[21] C. Gao, M. Sun, P. Zhang, and H. Zhai, Universal Dynamics of a Degenerate Bose Gas Quenched to Unitarity, Phys. Rev. Lett. **124**, 040403 (2020).

[22] M. T. Wheeler, H. Salman, and M. O. Borgh, Relaxation dynamics of half-quantum vortices in a two dimensional two-component Bose-Einstein condensate, EPL **135**, 30004 (2021).

[23] L. Gresista, T. V. Zache, and J. Berges, Dimensional crossover for universal scaling far from equilibrium, Phys. Rev. A **105**, 013320 (2022).

[24] J. F. Rodriguez-Nieva, A. Piñeiro Orioli, and J. Marino, Universal prethermal dynamics and self-similar relaxation in the two-dimensional Heisenberg model, PNAS **119**, e2122599119 (2021).

[25] P. Heinen, A. N. Mikheev, C.-M. Schmied, and T. Gasenzer, Non-thermal fixed points of universal sine-Gordon coarsening dynamics, arXiv:2212.01162 [cond-mat.quant-gas] (2022).

[26] I. Siovitz, S. Lannig, Y. Deller, H. Strobel, M. K. Oberthaler, and T. Gasenzer, Instantons in real time characterize universal dynamics in a one-dimensional spinor Bose gas, unpublished (2023).

[27] A. J. Bray, Theory of phase-ordering kinetics, Adv. Phys. **43**, 357–459 (1994).

[28] L. F. Cugliandolo, Coarsening phenomena, Comptes Rendus de Phys. **16**, 257 (2015).

[29] M. Onorato, S. Residori, U. Bortolozzo, A. Montina, and F. Arecchi, Rogue waves and their generating mechanisms in different physical contexts, Physics Reports **528**, 47–89 (2013).

[30] P. Bastian, M. Blatt, A. Dedner, N.-A. Dreier, C. Engwer, R. Fritze, C. Gräser, C. Grüninger, D. Kempf, R. Klöfkorn, M. Ohlberger, and O. Sander, The Dune framework: Basic concepts and recent developments, Computers & Mathematics with Applications **81**, 75–112 (2021).

[31] P. Heinen and T. Gasenzer, Complex Langevin approach to interacting Bose gases, Phys. Rev. A **106**, 191 063308 (2022).