Challenges and Opportunities in using Workflow Technology for Reproducible and Reusable Simulation Protocols
Challenges and Opportunities of Workflow Technology

- Multiscale Materials Modelling and Virtual Design @KIT
- Challenges in Materials Modelling
- Opportunities of Workflow Technology
- Challenges of Workflow Technology
Challenges and Opportunities of Workflow Technology

- Multiscale Materials Modelling and Virtual Design @KIT
- Challenges in Materials Modelling
- Opportunities of Workflow Technology
- Challenges of Workflow Technology
Development and application of methods for multi-scale simulations of nanoscale materials and devices

- nanoscale electronics
  - single molecule electronics
  - organic electronics
- carbon based systems
  - graphene and carbon nanotubes
- Workflows
  - Rapid Prototyping with Graphical Workflow editor (SimStack)

Materials design and discovery
Multiscale Materials Modelling and Virtual Design
AG-Wenzel (INT)

![Multiscale workflow diagram]

- **Single molecule**
- **Morphology**
- **Charge transport**
- **Hopping rates**
- **Device characteristics**
- **Resolution scales**
  - 1 Å
  - 10 nm
  - 100 nm
  - 1 µm

References:
- Friederich et al., JCTC 10 (9), 2014
- Friederich et al., Advanced Functional Materials 26 (31), 2016
Multiscale simulation in Organic Electronics

1. Single molecule parametrization (QM)
   - Geometry optimization
   - Customized force-fields

2. Generation of atomistic morphologies
   - Molecules parametrized on quantum mechanical level
   - Simulation of physical vapor deposition

3. Calculation of charge hopping rates
   - Full quantum mechanical electronic structure analysis
   - Electronic couplings, reorganization and orbital energies

4. Charge transport simulations
   - Time resolved charge carrier/exciton dynamics
   - IVs, IQEs, carrier balance, quenching, ...
Challenges and Opportunities of Workflow Technology

- Multiscale Materials Modelling and Virtual Design @KIT
- Challenges in Materials Modelling
- Opportunities of Workflow Technology
- Challenges of Workflow Technology
Challenges - Integration of new employees

- New student in the office
  - Reproduction of a given result as first task
  - Application of the solution to another data set (Bachelor level)
  - Improvement of the given method (Master level)
  - Development of a new method (PhD Cand. level)

- Reality
  - Bachelor Thesis – 3 months
  - Barely enough time to get familiar with the work environment (command line, ssh, HPC, software)

Training of new Group Members

Content adapted from:
Challenges – Reproducability/Replicability

- **Motivation**
  - Continue ongoing projects
  - Build on work of others (Publications)
  - Use existing data for Benchmarking/Datamining (Repository)
  - Increase impact of own publications (citations)

Reproducibility

Content adapted from: xkcd.com
Challenges – Reproducability/Replicability

Quality of Published Data (Journal/Database)

Scientifically sound

Scientifically sound but missing metadata

works only in one lab

methodological errors

missing replicas

fictive data

VS.
Challenges – Reproducability/Replicability

- Preparation of input data
- Missing steps/customizations
  - e.g. Data sources
- Hardware
- Compiler options
- Software Versions

→ For Reproducibility, all important aspects of a simulation need to be captured

**Molecular modeling**

Homology models of the GPHR Chimeras ectodomains in complex with the hormones and their visual representations were generated using the Molecular Operating Environment (MOE, 2012.10; Chemical Computing Group Inc., Montreal, QC, Canada). For all homology models, the crystal structure of the FSHR ectodomain in complex with FSH (PDB code 4AY9) was used as template [8]. The protein sequences of the human TSHR and hTSH were acquired from the UniProt database (Accession number hTSHR: P16473, hTSH: P01222) [32] and chimera ectodomain sequences generated by combining the corresponding parts of TSHR and FSHR protein sequence. The sequences of the chimeras were aligned to the structure within MOE prior to each chimera homology modeling. For each chimera 100 homology modeling simulations were generated employing the aforementioned method. The structure with the lowest energy was then selected and used as a template for LRR domain whereas for other domains, only the common core was retained as environment while the beta structures and the coordinates of the FSHR were preserved.

**Experimental Section**

Simulations: DFT calculations were carried out using the TurboMole Package. All calculations were performed using the hybrid B3-LYP functional. Reorganization energies were calculated using the def2-TZVPP basis-set while for energy levels, energy disorder, and electronic couplings, the def2-SV[p] basis-set was used. Atomistically resolved morphologies were generated using the Metropolis Monte Carlo based simulated annealing method DEPOSI.[33] This method required DFT-optimized molecular conformations and partial charges (B3-LYP/def2-SV(p)). Energy disorder and HOMO/LUMO levels as well as IPs and EA were calculated using the quantum patch method.[16]

Schaarschmidt, PLOS ONE (2014)

Friederich, Advanced Materials (2017)
4. Challenges - Scalability

Requirements

- Availability of Resources
- Automatization of the simulation protocol
- Homogeneous input data

![Scalability Chart](xkcd.com)
Challenges - Competence Drain

- Highly specialized employee (PhD cand., etc.) implements method
  - Results in highly specialized hard to use software tool

- Employee leaves
  - Knowledge about usage of software tool leaves with employee
  - Usage/Maintenance/Support of software tool hard to impossible

//Peter wrote this, nobody knows what it does, don't change it!
Challenges - Interoperability

- Software needs to be inter-operable
  - Multi-scale environments
  - Data-sharing

- Software Development
  - Software needs to be usable by other groups
  - Software aggregates need to be
    - prepared
    - shared
    - archived
    - presented (Deliverables)

---

//When I wrote this, only God and I understood what I was doing
//Now, God only knows

Content adapted from:
Challenges – Scientific Group

- Competence Drain
- Interoperability
- Reproducibility
- Scalability
- Training of new group members

Content adapted from:
Challenges and Opportunities of Workflow Technology

- Multiscale Materials Modelling and Virtual Design @KIT
- Challenges in Materials Modelling
- Opportunities of Workflow Technology
- Challenges of Workflow Technology
A workflow represents the coordinated execution of repeatable computational steps while accounting for dependencies and concurrency of tasks.

https://github.com/MD-Studio/MDStudio/
Focus
• Automation
• Speedup
• Provenance
• Innovation

Components
• Applications
• (Web)services/Utilities
• Libraries
• Data
• Control Elements

Abstraction Level
• Command Line
• Script based
• Workflow file
• GUI

Backend
• HPC Clusters
• Cloud/Grid Resources
• Workstation
The multiscale simulation workflow for Organic Electronics

1. Single molecule parametrization (QM)
   - Geometry optimization
   - Customized force-fields

2. Generation of atomistic morphologies
   - Molecules parametrized on quantum mechanical level
   - Simulation of physical vapor deposition

3. Calculation of charge hopping rates
   - Full quantum mechanical electronic structure analysis
   - Electronic couplings, reorganization and orbital energies

4. Charge transport simulations
   - Time resolved charge carrier/exciton dynamics
   - IVs, IQEs, carrier balance, quenching, ...
The multiscale simulation workflow for Organic Electronics

Traditional approach to multiscale modeling

High level of complexity of multiscale modeling:
- Input preparation
- Data transfer to computing resources
- Job submission and monitoring,

[Diagram of workflow]

Content adapted from:
Translation enabled by SimStack

Embedded Scientific modules = „WaNos“

Saved workflows for reproducible multiscale simulations

Connect to remote computational resources

Define input files and parameters for each module

Construct a workflow by drag & drop

Content adapted from:
Translation enabled by SimStack

… a generic workflow platform conquering complexity

- Open to arbitrary software modules
- Rapid prototyping: 30min to include new modules, 1 h to construct functional workflows
- Maximal reusability and scalability
- Module interoperability:
  - Fully automated, file based data transfer between modules
  - Schema based data transfer in development
  - Compatible with OWL ontologies (e.g. EMMO, once developed)
Deposit

Deposit is a Monte-Carlo tool to generate organic thin-film morphologies

- Complex input language (roughly 80 parameters up front)
- Minimum number of input files: 2
- Minimum number of parameters, which you have to actually know: 7

Learning curve

- Learn bash
- Find out about the important parameters
- Learn ssh
- Scp files over
- Learn specific qsub commands
- Write (or at least adapt) a submission script
- …
The multiscale simulation workflow for Organic Electronics

Translation by experts into Reusable workflow elements

Content adapted from:
Incorporation of Modules (Workflow Active Nodes - WaNos)

```xml
<wanoTemplate>
  <wanoRoot name="Your Wano Name Here">
    Command for program execution
    <wanoExecCommand>/ExecutionScript.sh</wanoExecCommand>
    <wanoInputFiles>
      <wanoInputFile logical_filename="ExecutionScript.sh">ExecutionScript.sh</wanoInputFile>
      <wanoInputFile logical_filename="MandatoryInputFile.dat">MandatoryInputFile</wanoInputFile>
    </wanoInputFiles>
    <wanoOutputFiles>
      <wanoOutputFile> MandatoryOutputFileName1.dat </wanoOutputFile>
      <wanoOutputFile> MandatoryOutputFileName2.dat </wanoOutputFile>
    </wanoOutputFiles>
  </wanoRoot>
<wanoTemplate>
```

Input parameters here

Mandatory input files

File output expected by other WaNos

Content adapted from:
Incorporation of Modules (Workflow Active Nodes - WaNos)

Content adapted from:
Deposit is a Monte-Carlo tool to generate organic thin-film morphologies.

- Complex input language (roughly 80 parameters up front)
- Minimum number of input files: 2
- Minimum number of parameters, which you have to actually know: 7
The multiscale simulation workflow for Organic Electronics

1. Single molecule parametrization (QM)
   - Geometry optimization
   - Customized force-fields

2. Generation of atomistic morphologies
   - Molecules parametrized on quantum mechanical level
   - Simulation of physical vapor deposition

3. Calculation of charge hopping rates
   - Full quantum mechanical electronic structure analysis
   - Electronic couplings, reorganization and orbital energies

4. Charge transport simulations
   - Time resolved charge carrier/exciton dynamics
   - IVs, IQEs, carrier balance, quenching, ...

Content adapted from:
Connection to Compute Resources

- Middleware actively developed in FZ Jülich
- Handles User Authentication
- Can connect to all common schedulers
- Handles data transfer between workflow steps
- Setup in < 1h with installer provided by Nanomatch

Content adapted from:
Challenges and Opportunities of Workflow Technology

- Multiscale Materials Modelling and Virtual Design @KIT
- Challenges in Materials Modelling
- Opportunities of Workflow Technology
- Challenges of Workflow Technology
Reproducability of Workflows

What needs to be captured?

- Software, Tools, Scripts
- WaNos
- Workflow templates
- Executed Workflows
- Author

How to capture it

- Containerization (e.g. udocker)
  - https://github.com/indigo-dc/udocker
- Versioning
- Tags
- Repositories
Interoperability of workflow frameworks

It is unlikely that one workflow framework will meet all the needs of the community

→ Shared data Formats and/or Converters will be required
Acknowledgements

- Prof. Dr. Wolfgang Wenzel
- Dr. Ivan Kondov (SCC)
- Dr. Timo Strunk
- Dr. Tobias Neumann

http://www.nanomatch.com
http://www.simstack.eu

Thank You!

joerg.schaarschmidt@kit.edu