Karlsruhe Institute of Technology

Challenges and Opportunities in using Workflow Technology

for Reproducible and Reusable Simulation Protocols

INSTITUTE OF NANOTECHNOLOGY

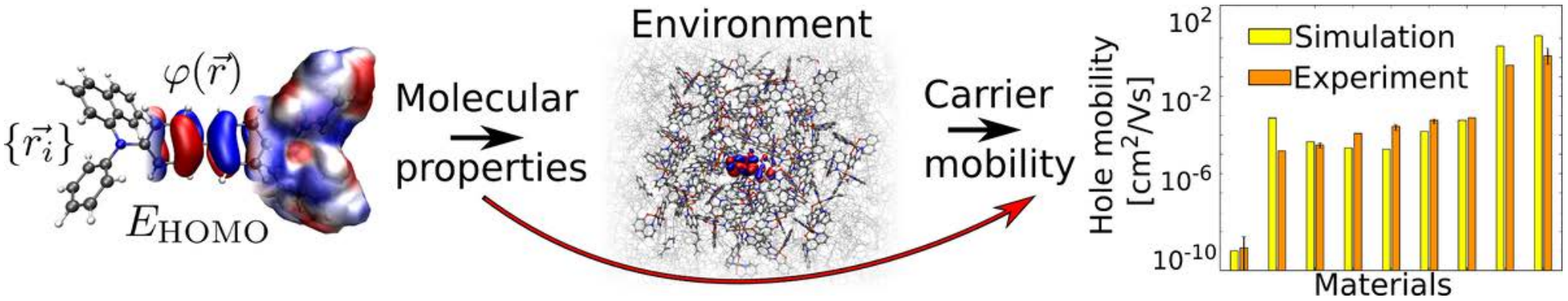
Challenges and Opportunities of Workflow Technology

- 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

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



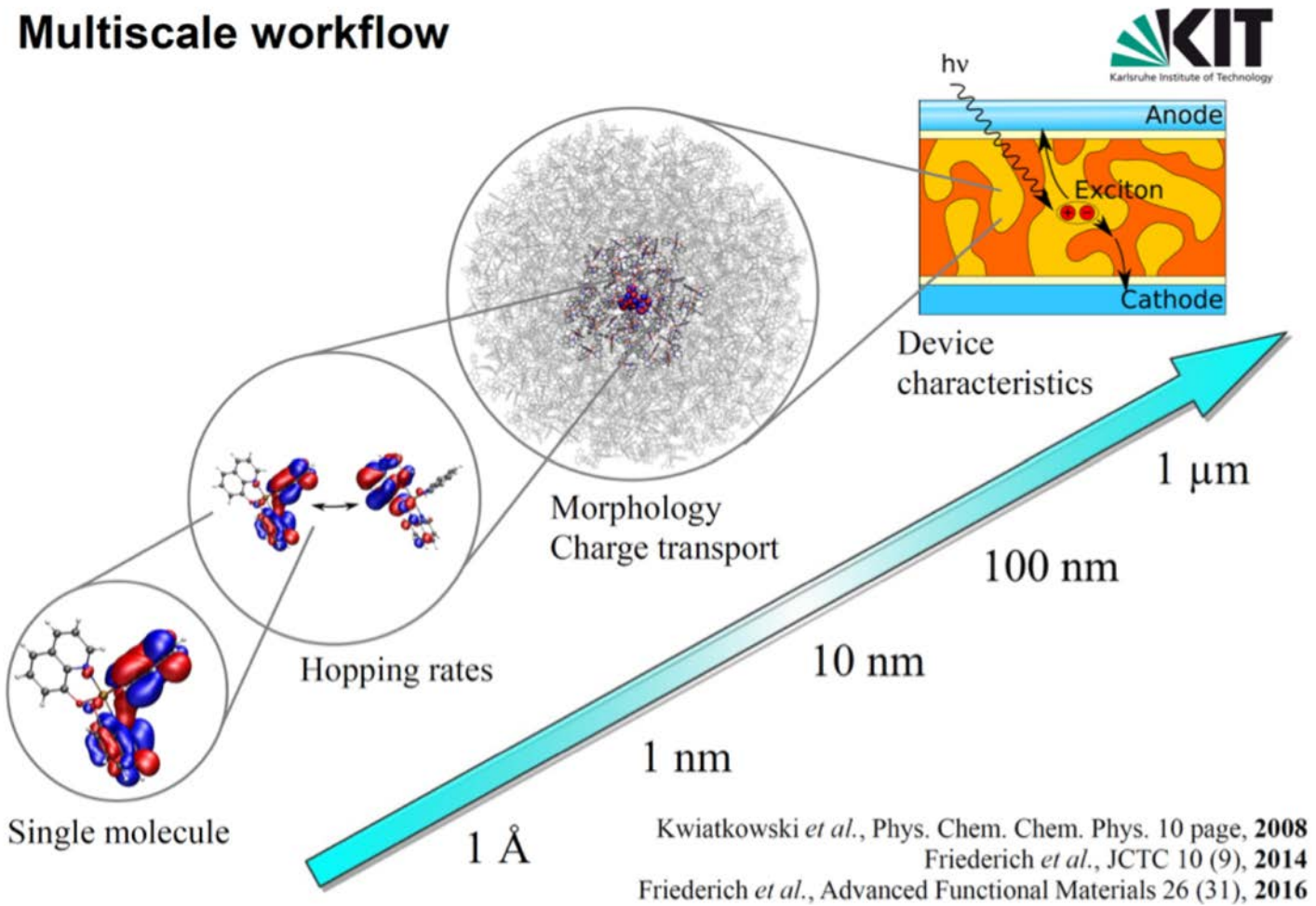
- Development and application of methods for multi-scale simulations of nanoscale materials and devices

→ materials design and discovery

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

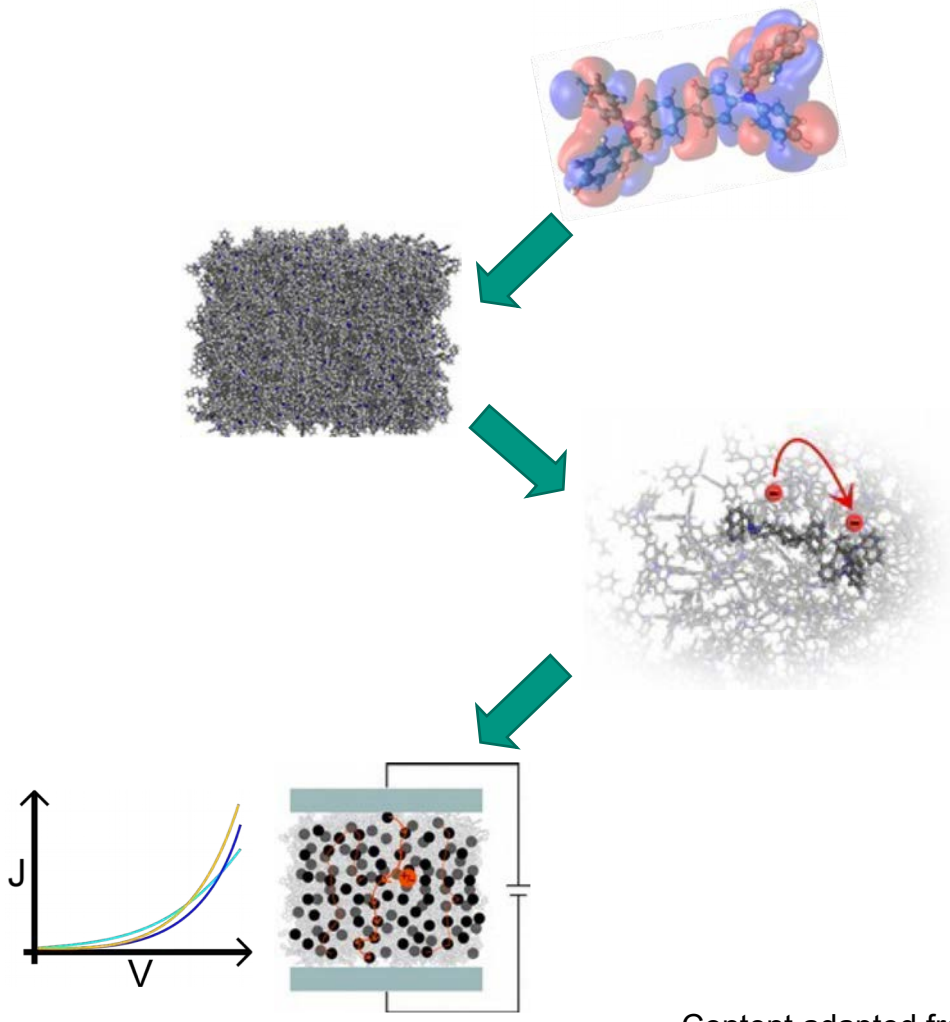
Prof. Dr.
Wolfgang Wenzel

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



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, ...



Content adapted from:

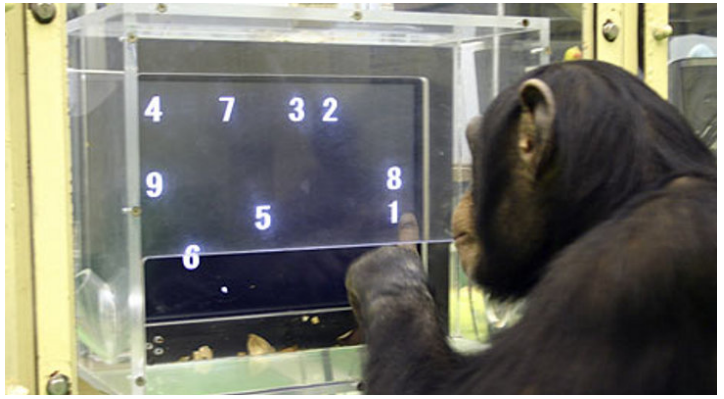
Challenges and Opportunities of Workflow Technology

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

Challenges - Integration of new employees

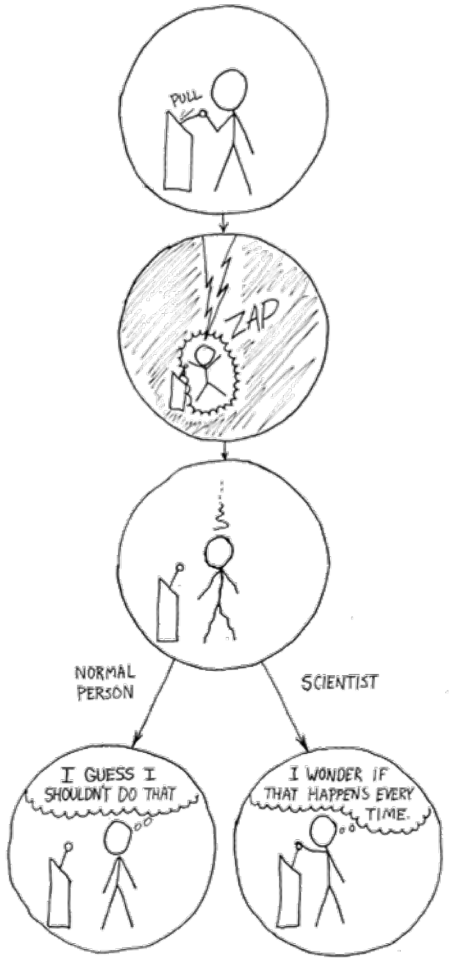
- New student in the office
  - Reproduction of a given result as first task
  - Application of the solution to another data set (Bachelor level)
  - Improvement of the given method (Master level)
  - Development of a new method (PhD Cand. level)
- Reality
  - Bachelor Thesis – 3 months
  - Barely enough time to get familiar with the work environment (command line, ssh, HPC, software)

Training of new Group Members



Content adapted from:

Challenges – Reproducability/Replicability



xkcd.com

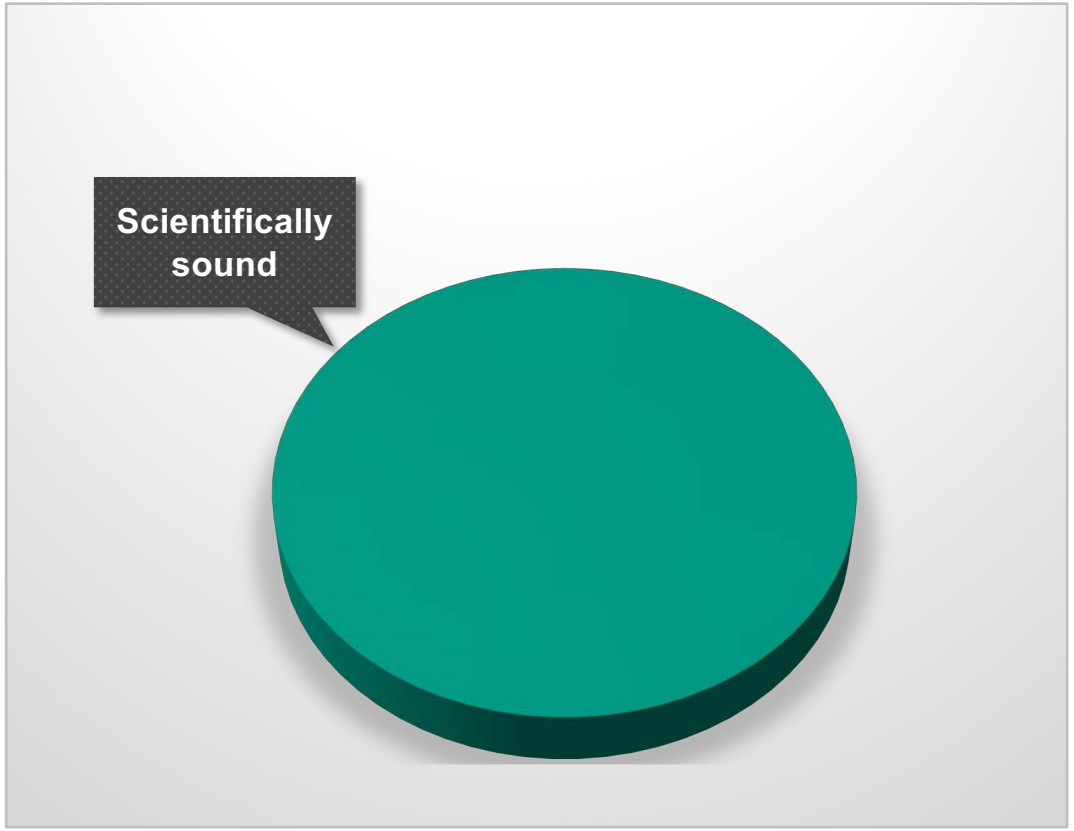
- Motivation
  - Continue ongoing projects
  - Build on work of others (Publications)
  - Use existing data for Benchmarking/Datamining (Repository)
  - Increase impact of own publications (citations)

Reproducibility

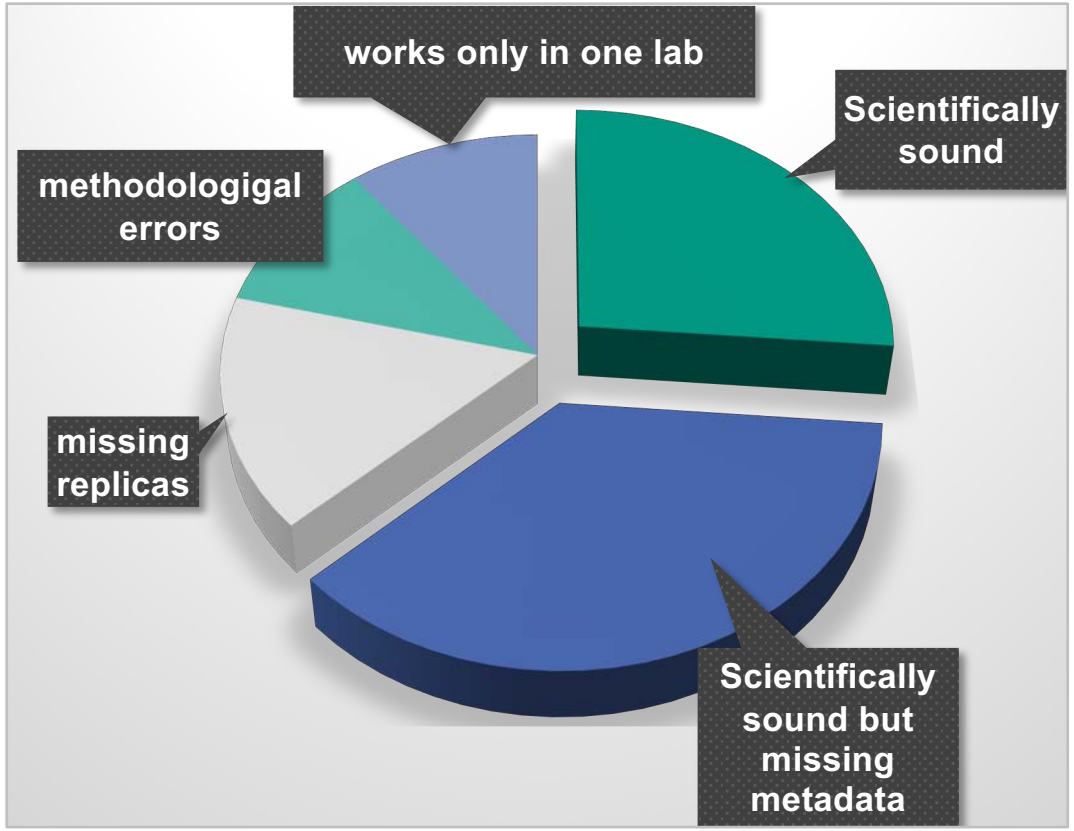
Content adapted from:

Challenges – Reproducability/Replicability

Quality of Published Data (Journal/Database)

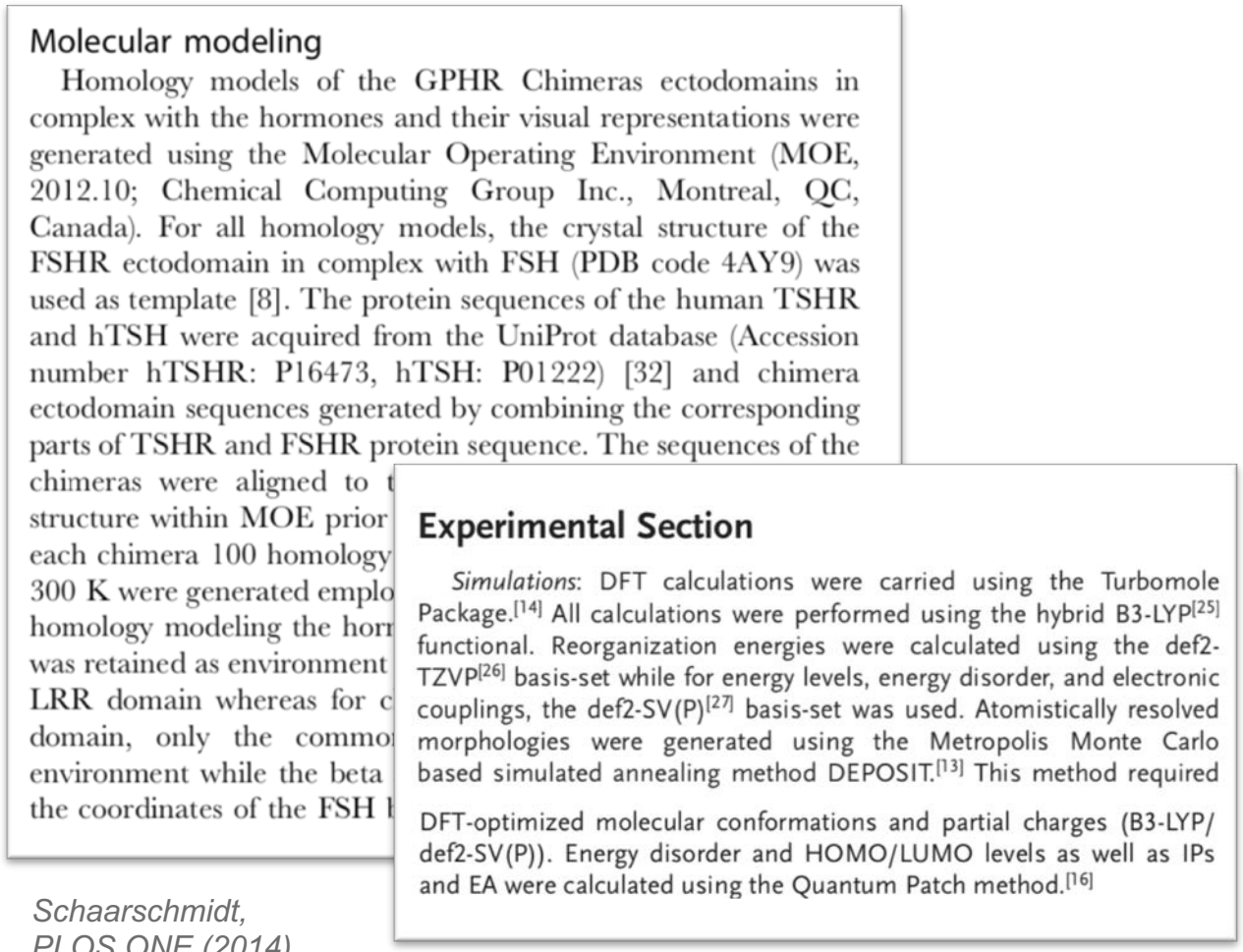


VS.



fictive data

Challenges – Reproducability/Replicability



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 t
structure within MOE prior
each chimera 100 homology
300 K were generated emplo
homology modeling the horr
was retained as environment
LRR domain whereas for c
domain, only the commo
environment while the beta
the coordinates of the FSH

Experimental Section

Simulations: DFT calculations were carried using the Turbomole Package.[14] All calculations were performed using the hybrid B3-LYP[25] functional. Reorganization energies were calculated using the def2-TZVP[26] basis-set while for energy levels, energy disorder, and electronic couplings, the def2-SV(P)[27] basis-set was used. Atomistically resolved morphologies were generated using the Metropolis Monte Carlo based simulated annealing method DEPOSIT.[13] 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)

- 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

4. Challenges - Scalability

Scalability

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

HOW LONG CAN YOU WORK ON MAKING A ROUTINE TASK MORE EFFICIENT BEFORE YOU'RE SPENDING MORE TIME THAN YOU SAVE?
(ACROSS FIVE YEARS)

HOW MUCH TIME YOU SHAVE OFF	HOW OFTEN YOU DO THE TASK: 50/DAY	5/DAY	DAILY	WEEKLY	MONTHLY	YEARLY
1 SECOND	1 DAY	2 HOURS	30 MINUTES	4 MINUTES	1 MINUTE	5 SECONDS
5 SECONDS	5 DAYS	12 HOURS	2 HOURS	21 MINUTES	5 MINUTES	25 SECONDS
30 SECONDS	4 WEEKS	3 DAYS	12 HOURS	2 HOURS	30 MINUTES	2 MINUTES
1 MINUTE	8 WEEKS	6 DAYS	1 DAY	4 HOURS	1 HOUR	5 MINUTES
5 MINUTES	9 MONTHS	4 WEEKS	6 DAYS	21 HOURS	5 HOURS	25 MINUTES
30 MINUTES		6 MONTHS	5 WEEKS	5 DAYS	1 DAY	2 HOURS
1 HOUR		10 MONTHS	2 MONTHS	10 DAYS	2 DAYS	5 HOURS
6 HOURS				2 MONTHS	2 WEEKS	1 DAY
1 DAY					8 WEEKS	5 DAYS

xkcd.com

Challenges - Competence Drain

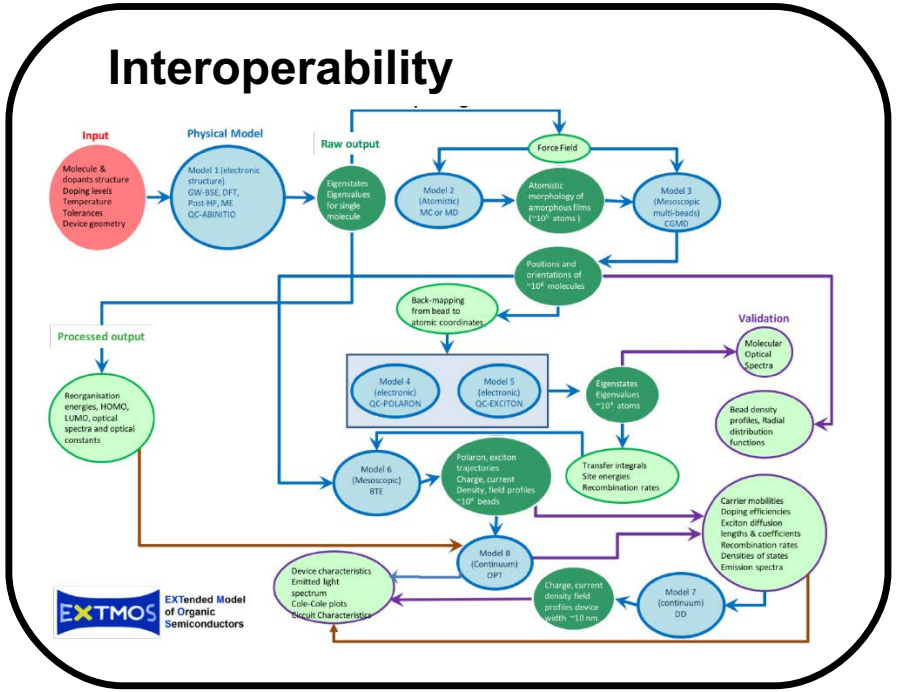
- Highly specialized employee (PhD cand., etc.) implements method
  - Results in highly specialized hard to use software tool
- Employee leaves
  - Knowledge about usage of software tool leaves with employee
  - Usage/Maintenance/Support of software tool hard to impossible


```
//Peter wrote this, nobody knows what it does, don't change it!
```

Content adapted from:

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)

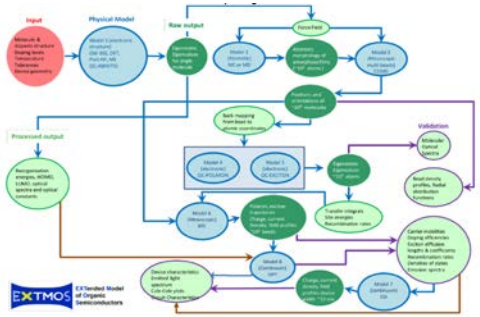



```
//When I wrote this, only God and I understood what I was doing
//Now, God only knows
```

Content adapted from:

Challenges – Scientific Group

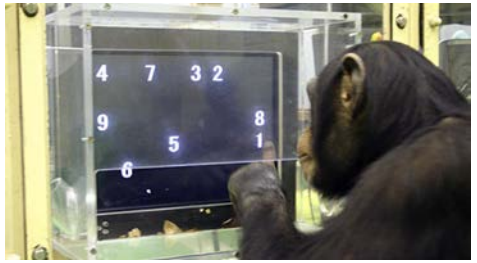
Interoperability



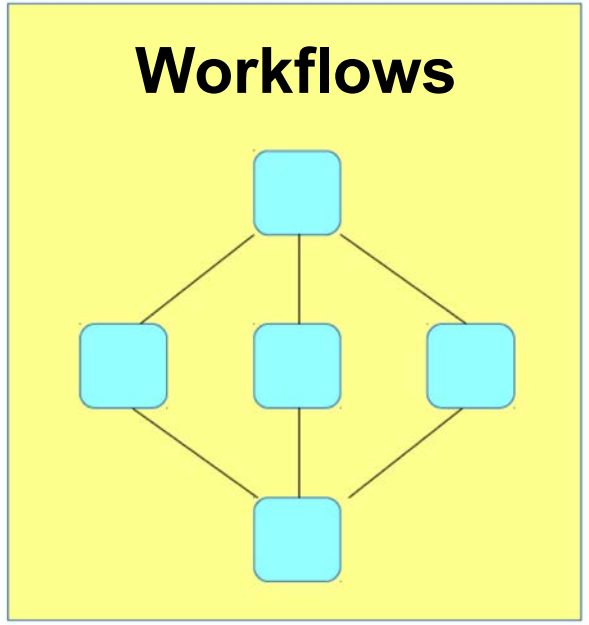
Reproducibility

Scalability

Training of new group members



Workflows



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

Workflows

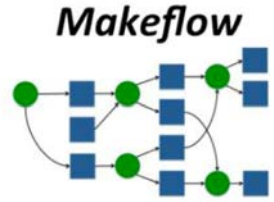
Focus	Components	Abstraction Level	Backend
• Automation • Speedup • Provenance • Innovation	• Applications • (Web)services/Utilities • Libraries • Data • Control Elements	• Command Line • Script based • Workflow file • GUI	• HPC Clusters • Cloud/Grid Resources • Workstation

https://github.com/MD-Studio/MDStudio/

A workflow represents the coordinated execution of repeatable computational steps while accounting for dependencies and concurrency of tasks.

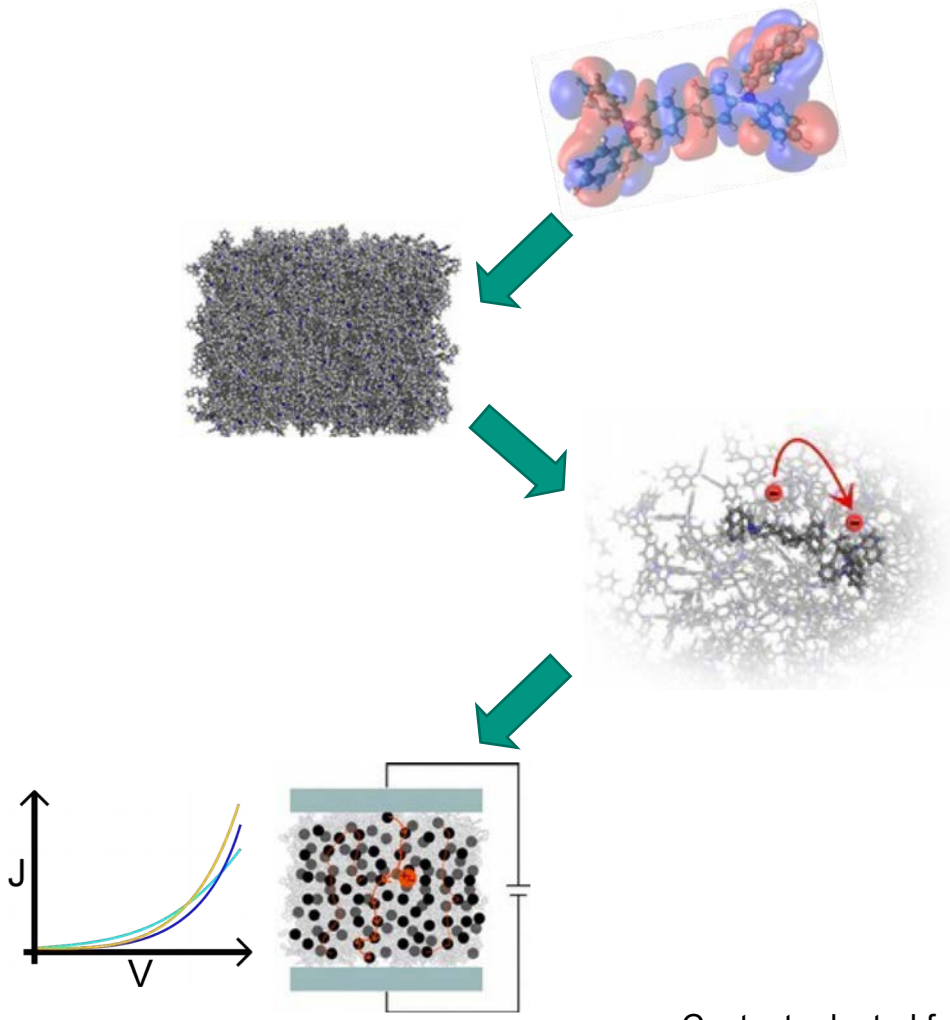
Workflows

Focus	Components	Abstraction Level	Backend
• Automation • Speedup • Provenance • Innovation	• Applications • (Web)services/Utilities • Libraries • Data • Control Elements	• Command Line • Script based • Workflow file • GUI	• HPC Clusters • Cloud/Grid Resources • Workstation



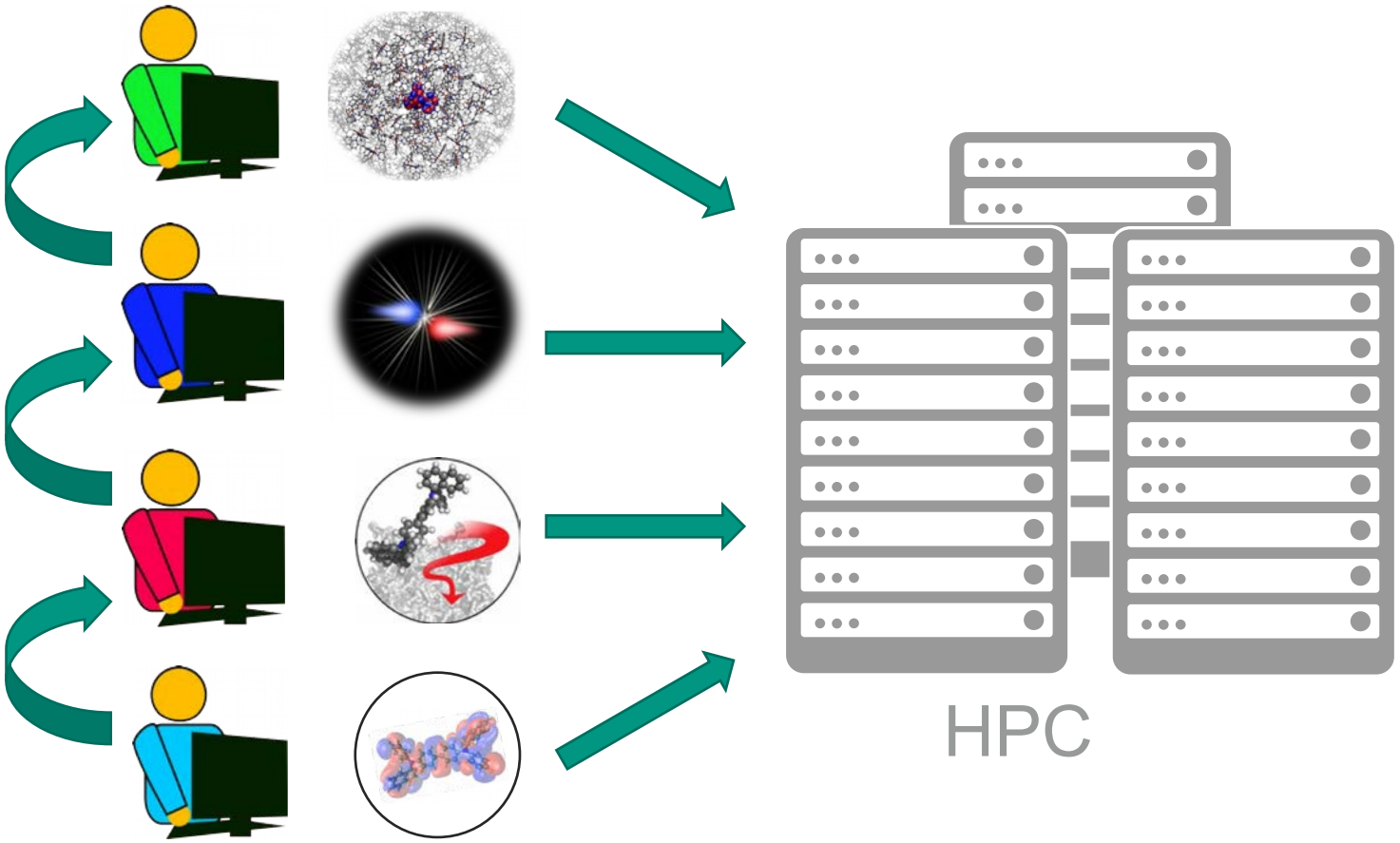
The multiscale simulation workflow for Organic Electronics

- 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, ...



Content adapted from:

The multiscale simulation workflow for Organic Electronics



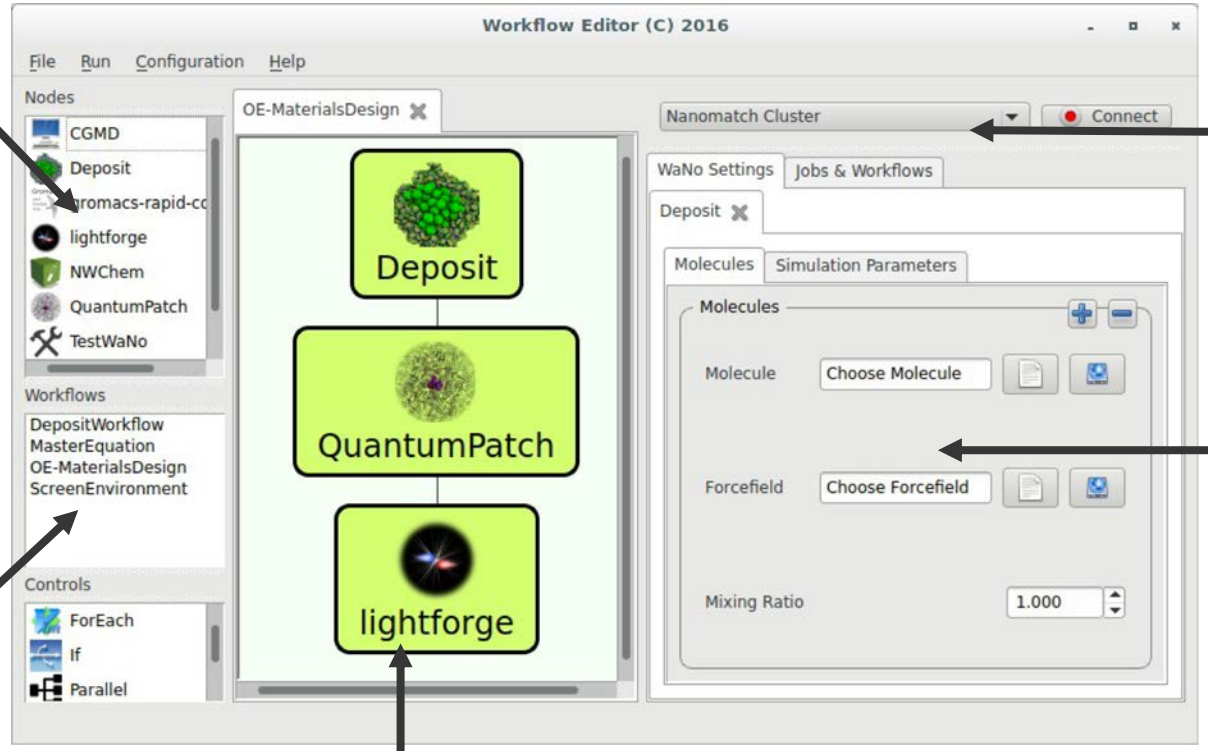
- High level of complexity of multiscale modeling:
  - Input preparation
  - Data transfer to computing resources
  - Job submission and monitoring,

→ Traditional approach to multiscale modeling

Content adapted from:

Translation enabled by SimStack

Embedded Scientific modules = „WaNos“



Connect to remote computational resources

Define input files and parameters for each module

Saved workflows for reproducible multiscale simulations

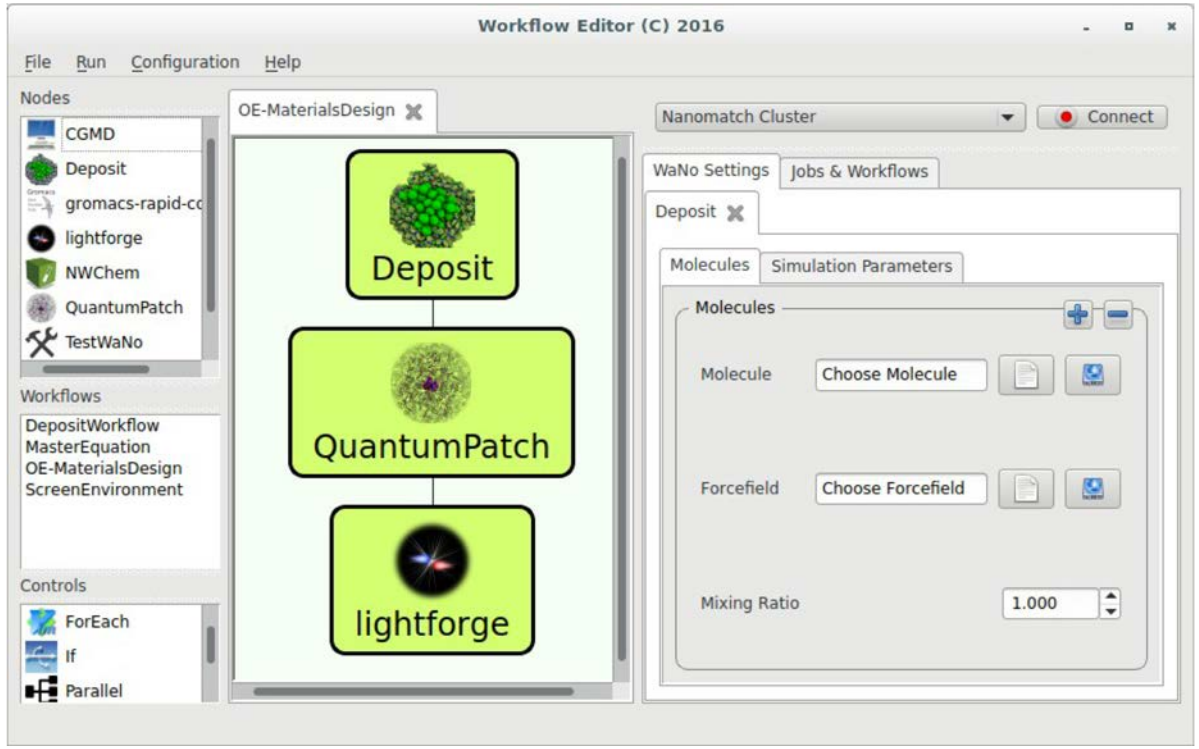
Construct a workflow by drag & drop

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)



Content adapted from:

Workflows @INT - Example Deposit

Deposit

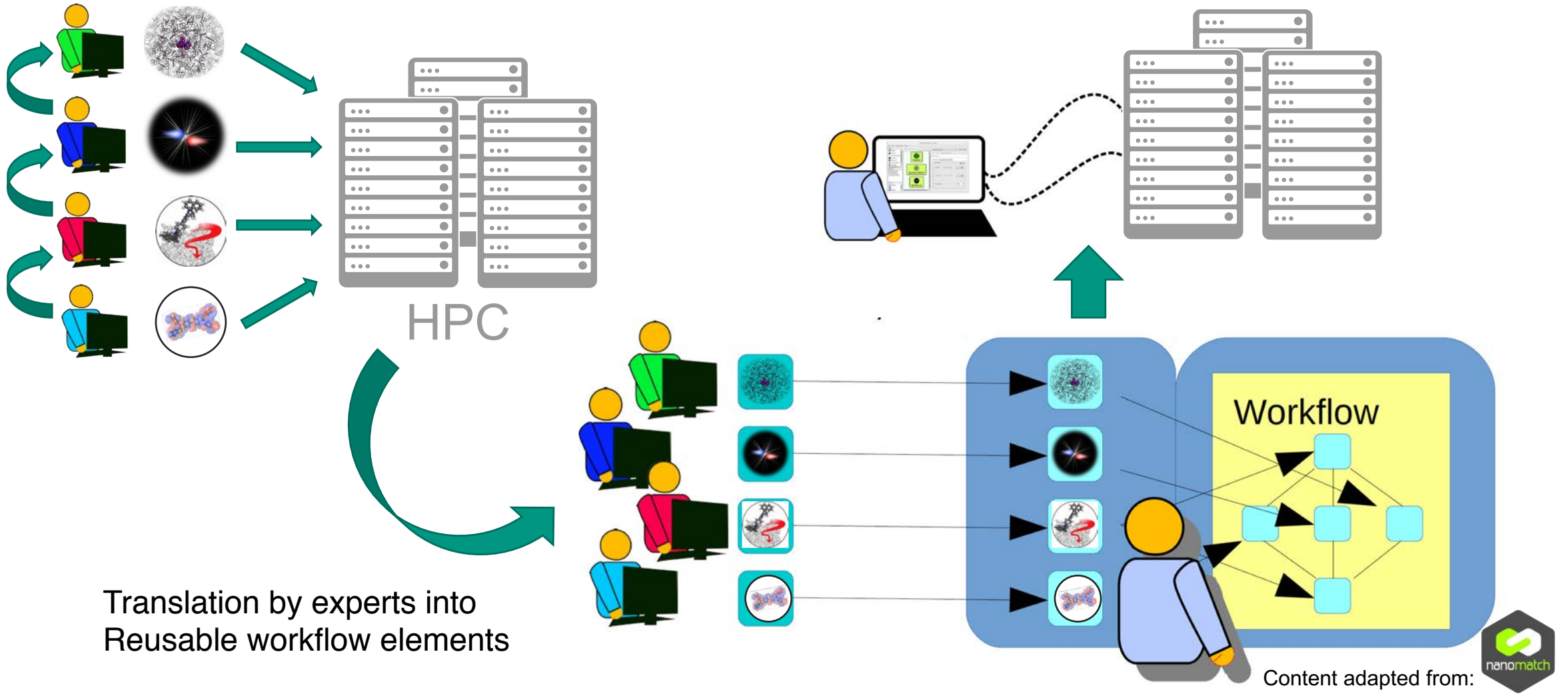
- Deposit is a Monte-Carlo tool to generate organic thin-film morphologies
  - Complex input language (roughly 80 parameters up front)
  - Minimum number of input files: 2
  - Minimum number of parameters, which you have to actually know: 7

Learning curve

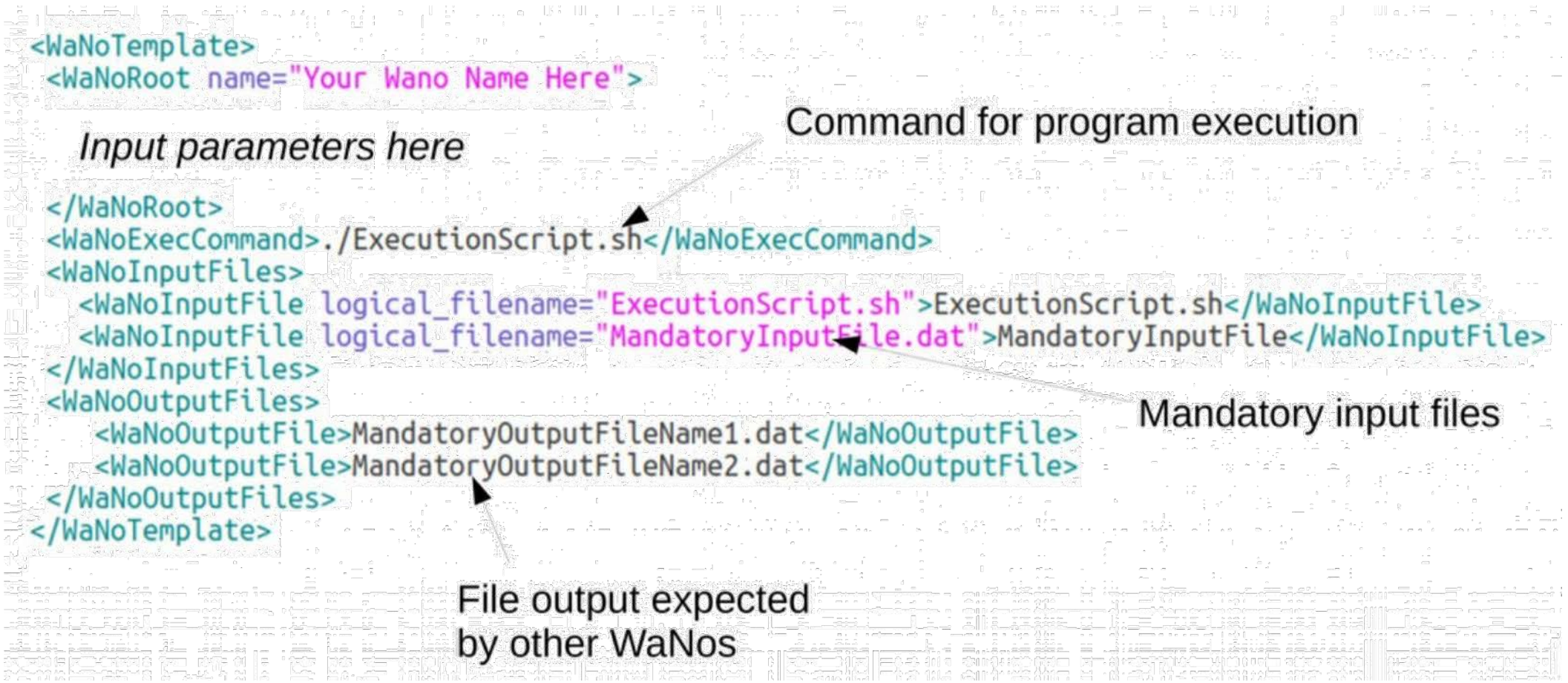
- Learn bash
- Find out about the important parameters
- Learn ssh
- Scp files over
- Learn specific qsub commands
- Write (or at least adapt) a submission script
- …

Content adapted from:

The multiscale simulation workflow for Organic Electronics



Incorporation of Modules
(Workflow Active Nodes - WaNos)




```
<WaNoTemplate>
 <WaNoRoot name="Your Wano Name Here">
   Input parameters here
 </WaNoRoot>
 <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>
</WaNoTemplate>
```

Content adapted from:

Incorporation of Modules
(Workflow Active Nodes - WaNos)

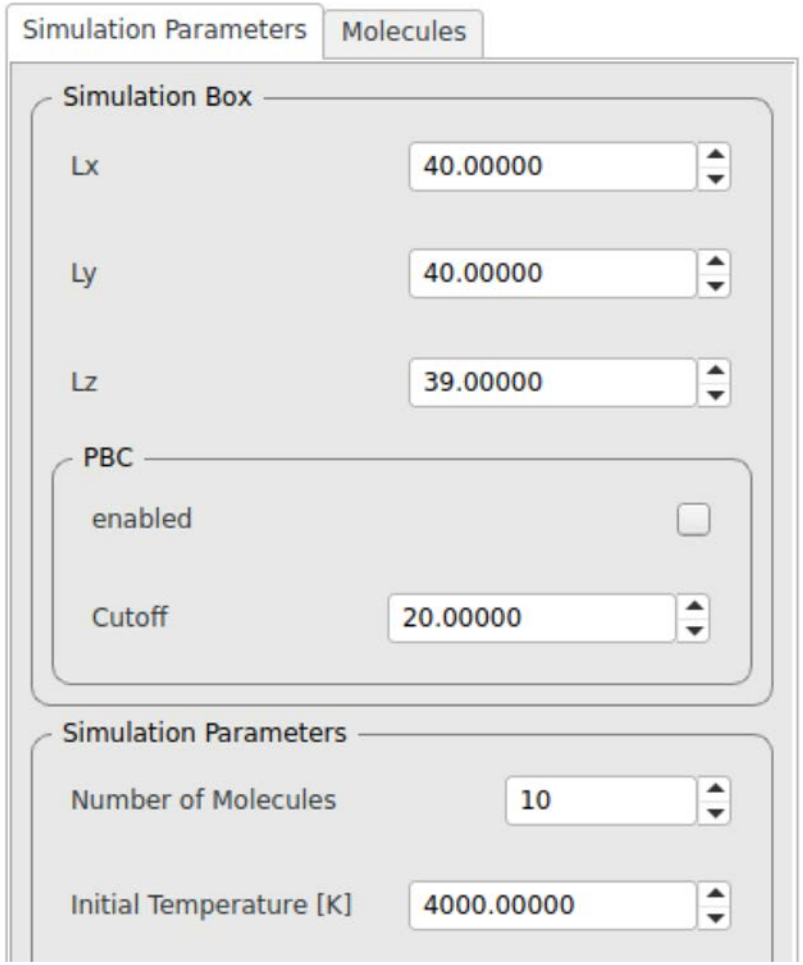
```
<WaNoTemplate>
 <WaNoRoot name="Random Number Generator">

  <WaNoFloat name="This is a float field">0.0</WaNoFloat>

  <WaNoInt name="This is an int field">1</WaNoInt>

  <WaNoChoice name="Choice of options">
    <Entry id="0" chosen="True">Standard option</Entry>
    <Entry id="1">Second option</Entry>
    <Entry id="2">Third option</Entry>
  </WaNoChoice>

  <WaNoDropDown name="maybe dropdown is better...">
    <Entry id="0" chosen="True">Standard option</Entry>
    <Entry id="1">Second option</Entry>
    <Entry id="2">Third option</Entry>
  </WaNoDropDown>

  <WaNoString name="String">"Writing letters..."</WaNoString>


  <WaNoBool name="I like beer">True</WaNoBool>

  <WaNoBox name="Group your parameters in a box">

   <WaNoBool name="Neglect box content">False</WaNoBool>

   <WaNoInt name="Another optional int">3</WaNoInt>

   <WaNoDropDown name="And another drop down">
    <Entry id="0" chosen="True">Standard option</Entry>
    <Entry id="1">Second option</Entry>
    <Entry id="2">Third option</Entry>
  </WaNoDropDown>

  </WaNoBox>
```

Content adapted from:

Workflows @INT - Example Deposit

Deposit

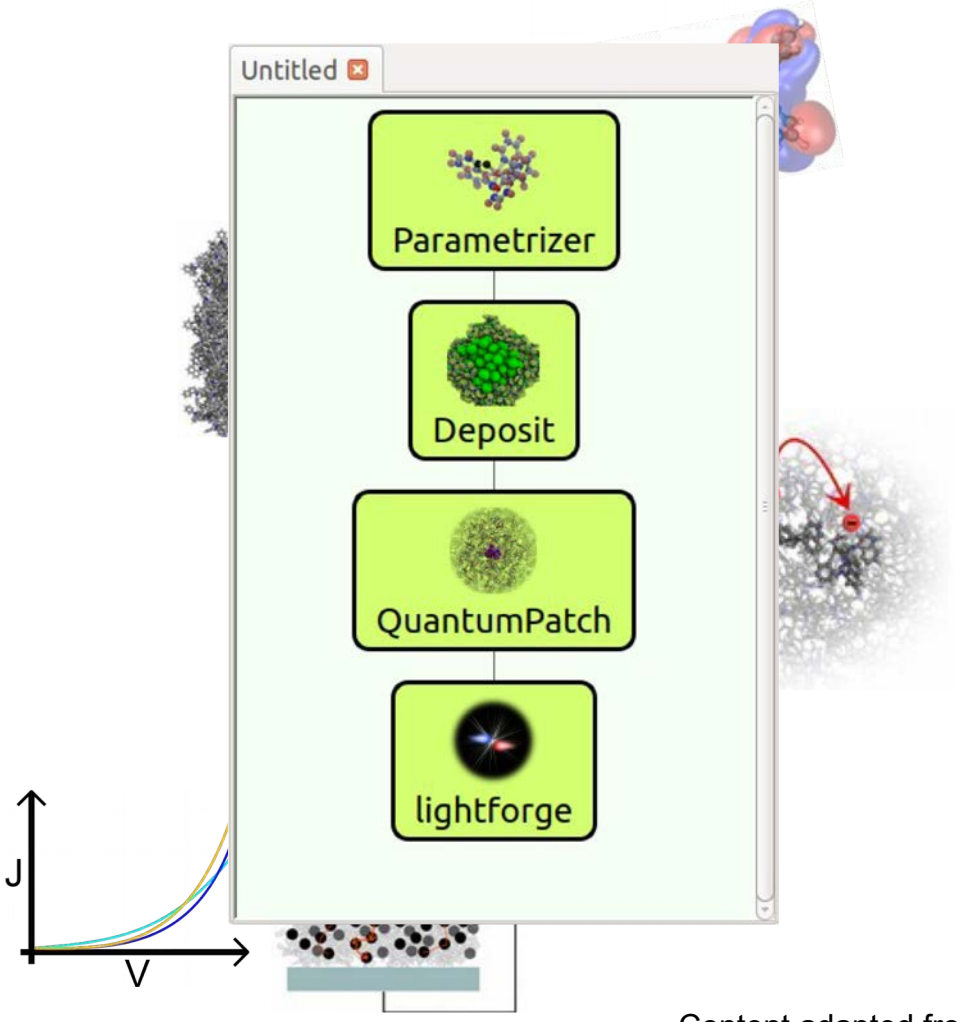
- Deposit is a Monte-Carlo tool to generate organic thin-film morphologies
  - Complex input language (roughly 80 parameters up front)
  - Minimum number of input files: 2
  - Minimum number of parameters, which you have to actually know: 7



Content adapted from:

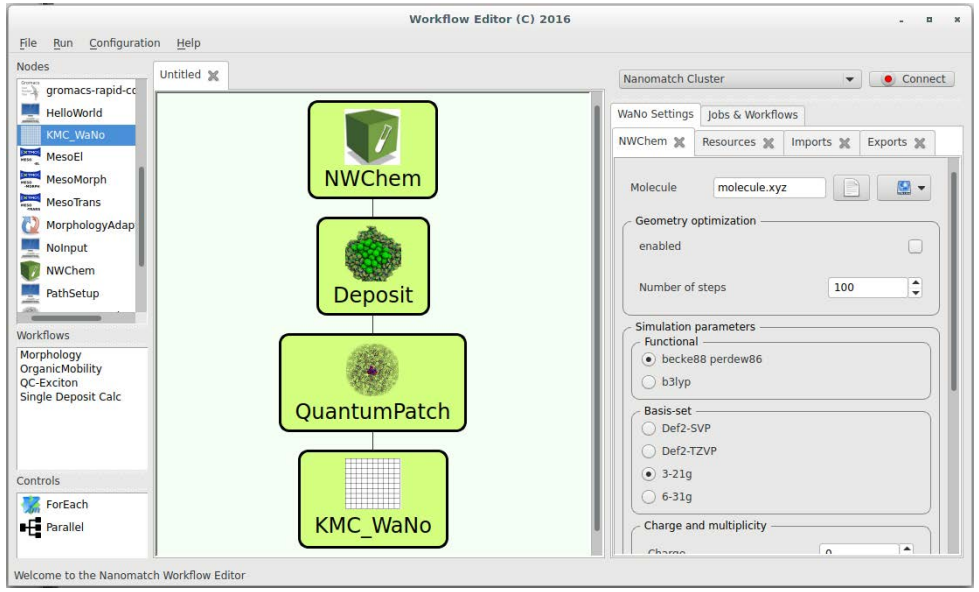
The multiscale simulation workflow for Organic Electronics

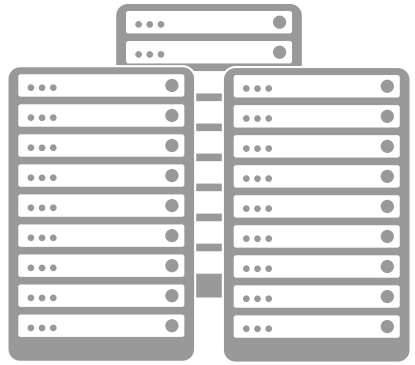
- 1. Single molecule parametrization (QM)
  - Geometry optimization
  - Customized force-fields
- 2. Generation of atomistic morphologies
  - Molecules parametrized on quantum mechanical level
  - Simulation of physical vapor deposition
- 3. Calculation of charge hopping rates
  - Full quantum mechanical 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





HPC

- 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)

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!

EXTMOS EXTended Model of Organic Semiconductors

joerg.schaarschmidt@kit.edu