UQ for materials modelling: Motivation and perspective

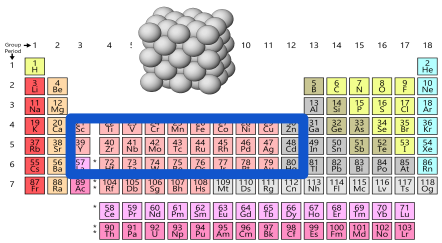
Michael F. Herbst

Applied and Computational Mathematics, RWTH Aachen University

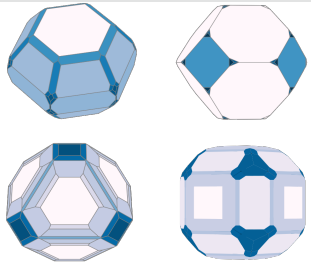
15 April 2022

Slides: `https://michael-herbst.com/talks/2022.04.15_siamuq.pdf`

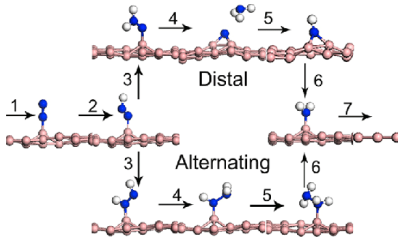
Task: Develop a new electrochemical catalyst[1]



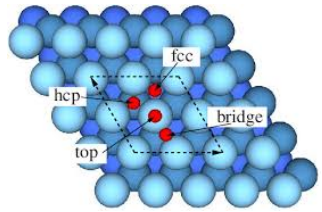
Host metal + dopants
$\simeq 30 \times 30 = 900$



Surface terminations
$\simeq 3 - 5$



Reaction intermediates
$\simeq 10$



Adsorpotion configurations
$\simeq 30$

- Combinatorial design space: $\simeq 10^5 - 10^6$ compounds
- Systematic experiments: Time and cost intensive

$\Rightarrow$ Computational screening to complement and accelerate

- Harvest curated data bases
- Data-driven methods and statistical learning

- Growing list of workflow tools and curated data

$\Rightarrow$ Millions of first-principle calculations

pymatgen FireWorks AFLOW Automatic · FLOW for Materials Discovery MATERIALSCLOUD AiiDA NOMAD NOVEL MATERIALS DISCOVERY

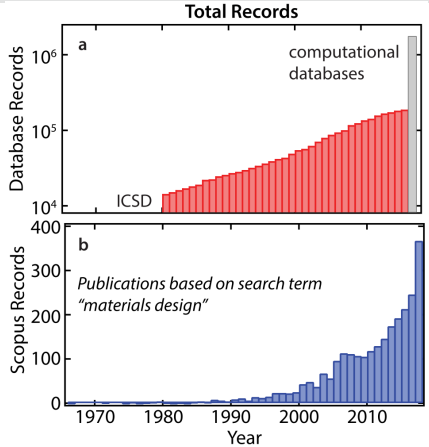
[1] https://www.cmu.edu/aced/index.html

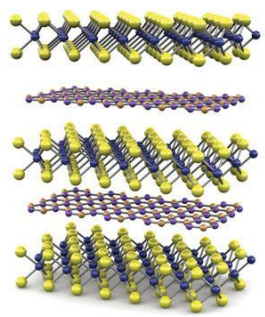
Status of high-throughput screening

- Exponentially growing impact

- Broad span of successful discoveries:[a]
  - Semiconductors[b]
  - Lithium-ion-based batteries[c]
  - Magnetic compounds[d]
  - 2D materials: Batteries[e], electronics[f]

⇒ Crucial tool to tackle 21st century challenges





---

[a] K. Alberi *et. al.* J. Phys. D, **52**, 013001 (2019).

[b] S. Luo *et. al.* WIREs Comput. Mol. Sci. **11**, e1489 (2021).

[c] L. Kahle *et. al.* Energy & Environ. Science, **13**, 928 (2020).

[d] S. Jiang *et. al.* J. Alloys Comp. **867**, 158854 (2021).

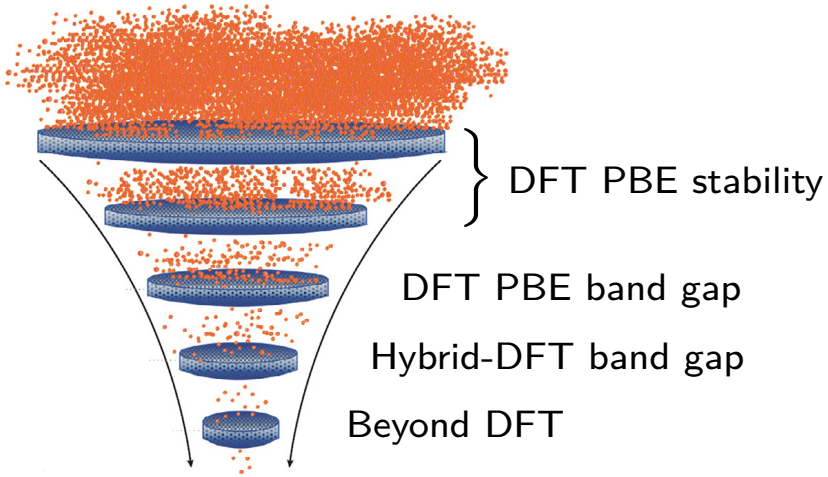
[e] A. Babak *et. al.* ACS Nano, **9**, 9507 (2015).

[f] C. Klinkert *et. al.* ACS Nano, **14**, 8605 (2020).

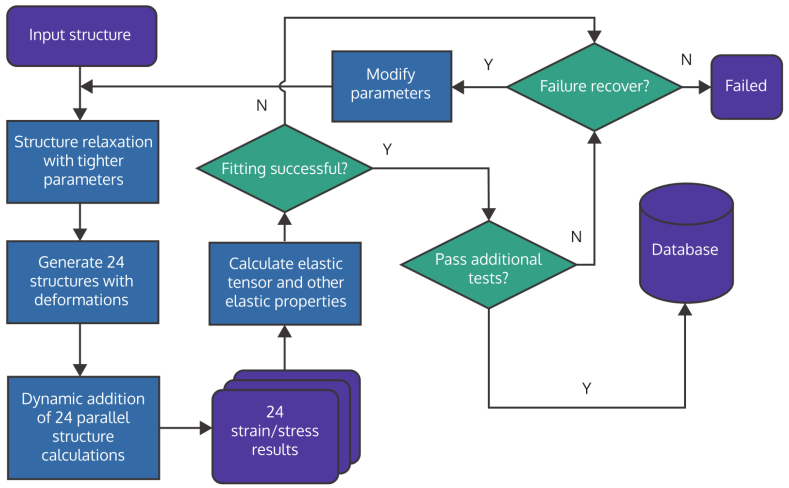
`https://www.edfenergy.com/electric-cars/batteries`;

J. Evans *Beyond graphene* Chemistry World (2014).

Sketch of high-throughput workflows



Design funnel for photovoltaic materials



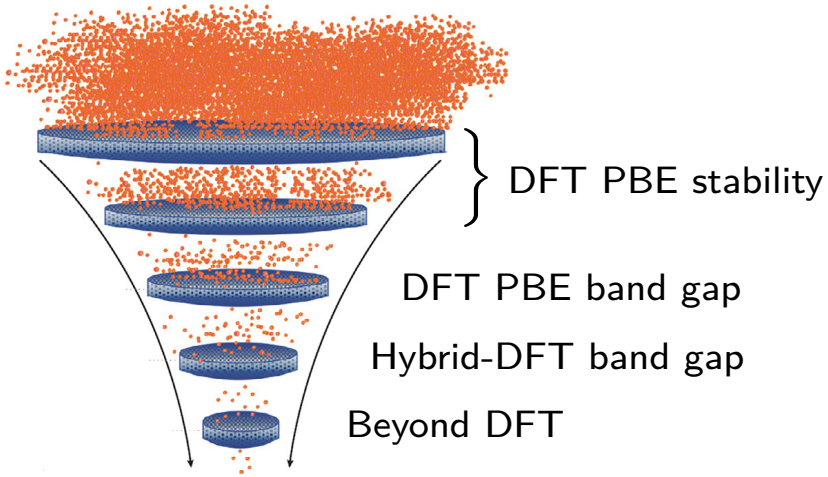
Workflow for computing elasticity tensors

- Many parameters to choose (algorithms, tolerances, models)
  - Elaborate heuristics: Failure rate $\simeq 1\%$
  - Still: Thousands of failed calculations

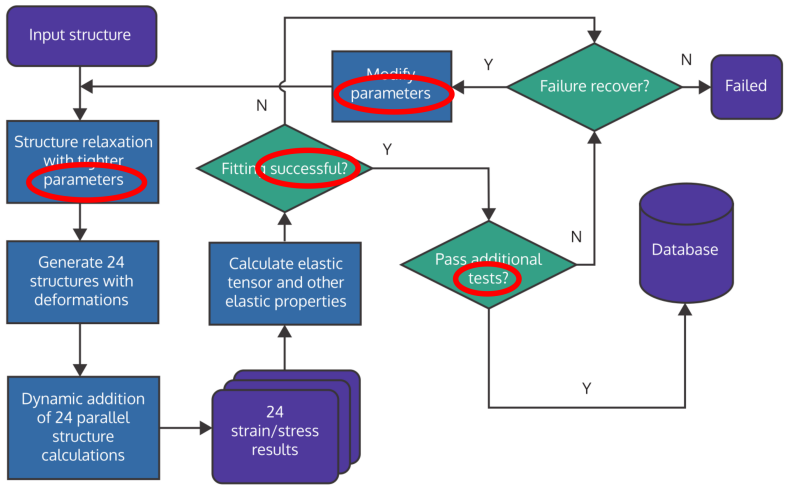
  $\Rightarrow$ Wasted resources & human attention
- Carbon footprint?
- More complex design spaces?
- Higher quantity and quality of data?

G. Hautier Comput. Mater. Sci. **164**, 108 (2019); L. Himanen *et. al.* Adv. Science **6**, 1900808 (2019).

Sketch of high-throughput workflows



Design funnel for photovoltaic materials



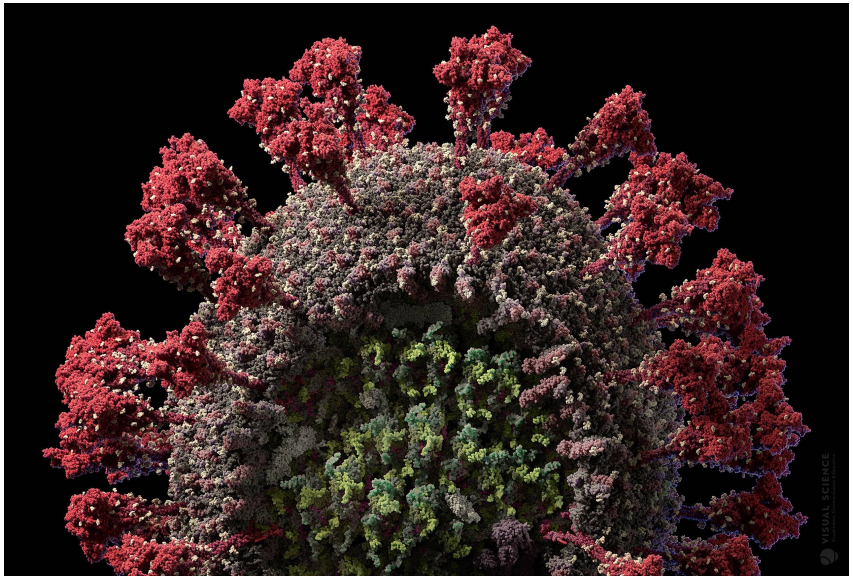
Workflow for computing elasticity tensors

- Many parameters to choose (algorithms, tolerances, models)
  - Elaborate heuristics: Failure rate $\simeq 1\%$
  - Still: Thousands of failed calculations

  $\Rightarrow$ Wasted resources & human attention
- Carbon footprint?
- More complex design spaces?
- Higher quantity and quality of data?

G. Hautier Comput. Mater. Sci. **164**, 108 (2019); L. Himanen *et. al.* Adv. Science **6**, 1900808 (2019).

Typical sizes and time-scales (1)

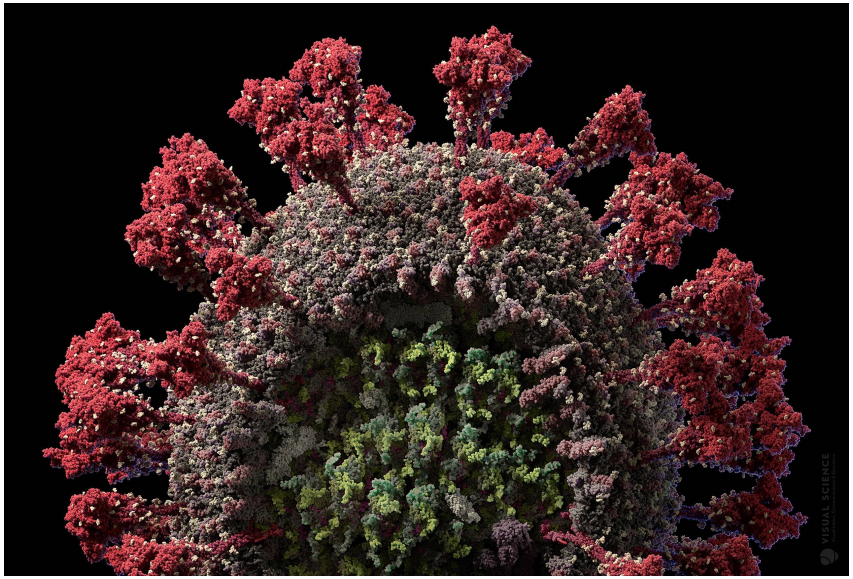


- Coronavirus: A molecule of interest
- Understand binding virus $\leftrightarrow$ ACE2 receptor:[1]
  - Simulation time: $4\mu s$
  - Step size: $2fs$

  $\Rightarrow$ 2 Million timesteps
- What's the typical size for a time step?

[1] J. Williams *et. al.* Proteins **90**, 1044 (2022)

Typical sizes and time-scales (1)

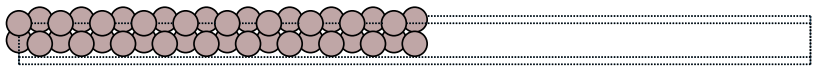


- Coronavirus: A molecule of interest
- Understand binding virus $\leftrightarrow$ ACE2 receptor:[1]
  - Simulation time: $4\mu s$
  - Step size: $2fs$

  $\Rightarrow$ 2 Million timesteps
- What's the typical size for a time step?

[1] J. Williams *et. al.* Proteins **90**, 1044 (2022)

Typical sizes and time-scales (2)

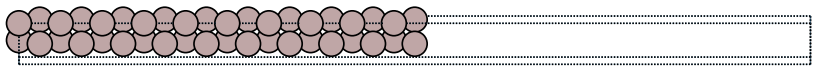


- Aluminium surface

- Density-functional theory (DFT) (standard *ab initio* method)
- 40 atoms $\simeq$ 10 minutes
- 1000 atoms $\simeq$ 24 hours

- How about the coronavirus?
- . . . and DFT is a *cheap* first-principle method

Typical sizes and time-scales (2)

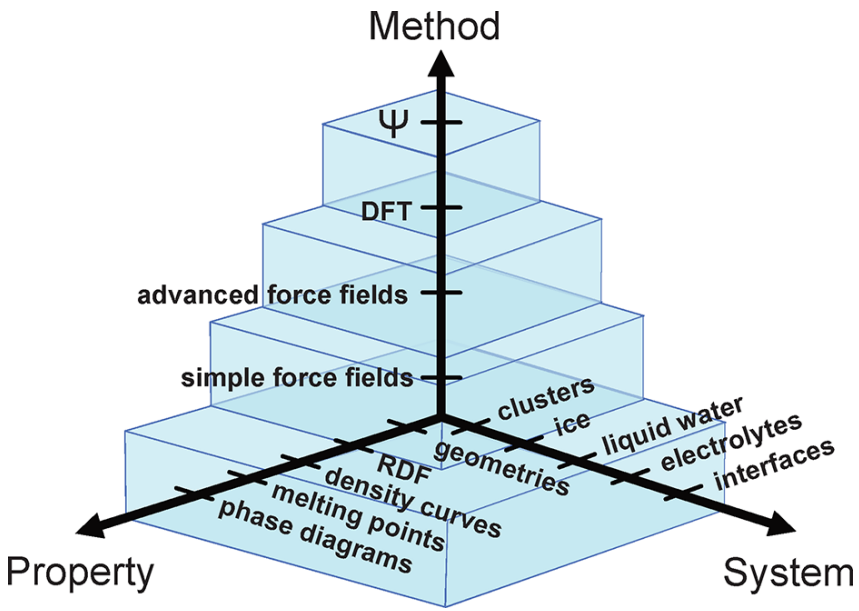


- Aluminium surface

- Density-functional theory (DFT) (standard *ab initio* method)
- 40 atoms $\simeq$ 10 minutes
- 1000 atoms $\simeq$ 24 hours

- How about the coronavirus? $\simeq 10^6$ years
- . . . and DFT is a *cheap* first-principle method

Typical sizes and time-scales (3)



[1] J. Behler Chem. Rev. **121** 10037 (2021).

Modelling stages

- First principle data acquisition
  - Usually density-functional theory
- Generation of interatomic potentials
  - Machine-learned interatomic potentials (MLIP)
  - Feature generation & molecular representation[1]
  - SOAP[2], SNAP[3], ACE[4], molecular graphs ...
  - Surrogate generation based on data (energy, force, stresses)
  - Least-squares fit
  - Neural networks[5]
- Molecular dynamics
  - Use potential to propagate dynamics in time

---

[1] F. Musil *et. al.* Chem. Rev. **121**, 9759 (2021)

[2] A. Bartók *et. al.* Phys. Rev. B **87**, 184115 (2013).

[3] A. Thompson *et. al.* J. Comput. Phys. **285**, 316 (2015).

[4] R. Drautz Phys. Rev. B **99**, 014104 (2019).

[5] J. Behler Chem. Rev. **121** 10037 (2021).

Sources of uncertainties

- First principle data acquisition
  - DFT model appropriate?
  - Numerical parameters converged?
- Generation of interatomic potentials
  - Data pool representative?
  - Completeness of the feature / molecular representation?
  - Fitting accurate & transferable?
- Molecular dynamics
  - Time step sufficiently small?
  - Dynamics equilibrated?
  - Trajectory captures event of interest?
  - Sampling sufficient?

Challenges and open problems

- Local versus global picture
  - DFT error in reaction energies $\leftrightarrow$ atomisation energies
  - Force error $\rightarrow$ Error in MD quantities
- Dependencies between the stages
  - Structure selection $\leftrightarrow$ MD quantity of interest
  - Feature generation $\leftrightarrow$ DFT numerical parameters
- Unification of analytical and statistical approaches
  - Numerical error *and* model error in DFT
- Identification of suitable model problems
  - Rigorously treatable, but relevant
- Overcoming interdisciplinary barriers
  - Standard software & methods: Huge & non-trivial
  - Development and integration of mathematical results?

What's coming in this minisymposium

- Efforts for interdisciplinary software (this talk)
- Potential Uncertainty (D. Foster)
- UQ for Long-Time Properties (D. Perez)
- Uncertainties in employing MLIP (K. Sargsyan, M. Ceriotti)
- Atomic Cluster Expansion (C. Ortner)
- Data set selection for learning (M. Sachs, C. van der Oord)

Join our lunch break (at 12noon EDT)

`https://michael-herbst.com/siamuq-lunch`

Zoom ID: 92967096593
Passcode: 639327

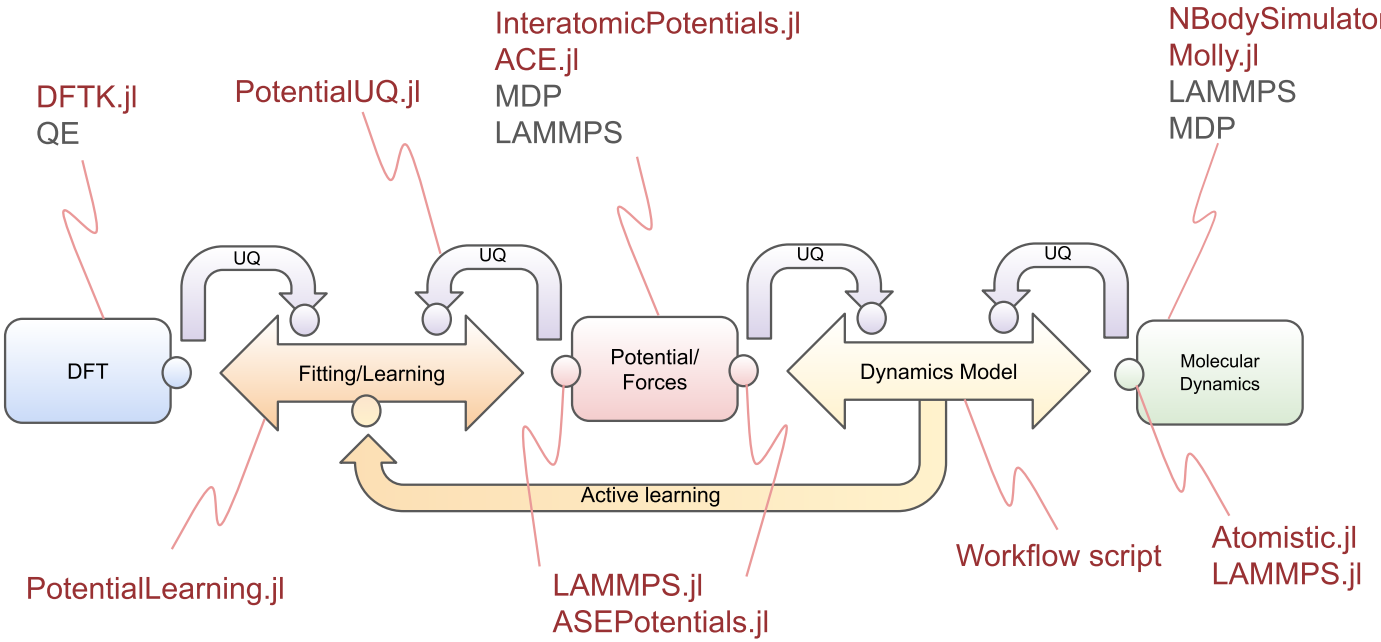
Interdisciplinary field ⇒ Multidisciplinary community

- **Mathematicians:** Toy models and unphysical edge cases
- **High-performance person:** Exploit hardware specialities
- **Scientist:** Design new models, not tweak numerics
- **Practitioner:** Reliable, black-box code, high-level interface

- State-of-the-art first-principle codes:
  - Difficult problem ⇒ Complex codes
  - Hard-coded details: Workflow, algorithms, optimisations
  - Huge code bases: 1M lines and beyond
  - Non-standard input syntax and API
  - Two-language problem: Algorithmic code hardly accessible

⇒ Innovations might not cross community boundaries

A first-principles pipeline for cross-disciplinary research



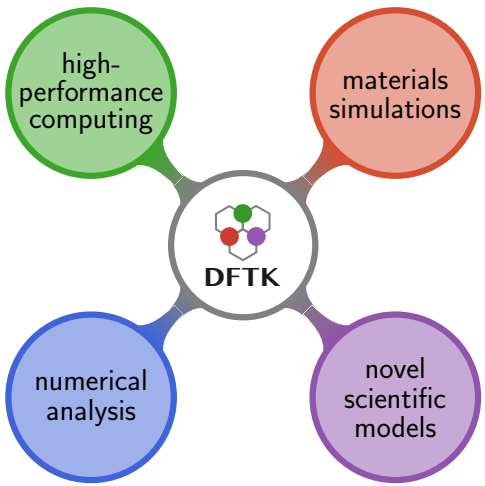
- Centred around the julia language & ecosystem
  - Pure-julia variants for key components
  - Interfaces to state-of-the-art equivalents

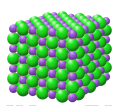
  $\Rightarrow$ Switch between flexibility & features

- But why julia?

  $\Rightarrow$ Let's take a look at DFTK

Density-functional toolkit[1] — https://dftk.org





$H\Psi = E\Psi$

- julia code for density-functional theory (DFT)
- Fully composable with julia ecosystem:
  - Arbitrary precision (32bit, >64bit, ...)
  - Algorithmic differentiation
  - Numerical error control
- Supports mathematical developments *and* scale-up to relevant applications
- *i.e.* reduced problems for rigorous analysis (1D, analytic potentials) *and* DFT on > 1000 electrons

$\Rightarrow$ Build with multidisciplinary research in mind

- Avoids two-language problem: Just julia
- Development started in 2019
- Only 7k lines of code

$\Rightarrow$ Low entrance barrier across backgrounds

[1] MFH, A. Levitt and E. Cancès. JuliaCon Proc., **3**, 69 (2021).

DFTK: Bringing mathematical research to the applications

- Self-adapting black-box DFT methods[a,b]
- Numerical analysis of DFT[c]
- Practical error bounds[d,e]

- Exploring algorithmic differentiation:
  - "Automatic response": Phonons & higher-order properties
  - Full AD-able simulation pipeline: DFT, potentials, MD
- Uncertainty quantification all the way: DFT, potentials, MD
- Approximate computing on modern GPUs

- Outreach and teaching
  - Community building: julia-based first-principle ecosystem
  - Mathematics of computational chemistry

Growing user base:

MIT
KU The University of Kansas
ASU Arizona State University
Carnegie Mellon University
École des Ponts ParisTech
Inria
UBFC Université Bourgogne Franche-Comté
서울대학교 Seoul National University
RWTH Aachen University
TUM

[a] MFH, A. Levitt. J. Phys. Condens. Matter **33**, 085503 (2021).

[b] MFH, A. Levitt. J. Comput. Phys. **459**, 111127 (2022).

[c] E. Cancès, G. Kemlin *et. al.* J. Matrix Anal. Appl., **42**, 243 (2021).

[d] MFH, A. Levitt, E. Cancès. Faraday Discus. **223**, 227 (2020).

[e] E. Cancès, G. Dusson *et. al.* arxiv 2111.01470v1.

Example: Routine computation of DFT model sensitivities

- DFT sensitivities require unusual, higher-order derivatives
- Combinatorial explosion:
  - "One PhD student per derivative" paradigm not feasible

  $\Rightarrow$ Use algorithmic differentiation $(\approx$ automatic derivatives$)$
- Illustration: DFT problem yields fixed-point density ρ_{SCF}

$$0 = \mathrm{diag}\left[\mathbb{1}_{(-\infty,\varepsilon_F]}(H_{a\theta}(\rho_{\mathrm{SCF}}))\right] - \rho_{\mathrm{SCF}}$$

  via a nested iterative scheme (self-consistent field)
- DFT Hamiltonian $H_{a\theta}$ depends on numerous parameters, e.g.
  - a: Lattice constant
  - θ: DFT model parameters
- Defines implicit function $\rho_{\mathrm{SCF}}(a, \theta)$

Computing stress sensitivities

- Model sensitivity of stress term $S(a,\beta,\kappa) = \frac{\partial \mathcal{E}(\rho_{\mathrm{SCF}}(a,\theta))}{\partial a}$:

$$\frac{dS}{d\theta} = \frac{\partial S}{\partial \rho_{\mathrm{SCF}}} \frac{\partial \rho_{\mathrm{SCF}}}{\partial \theta} \tag{1}$$

- Computed by implicit differentiation (response theory):

$$\frac{\partial \rho_{\mathrm{SCF}}}{\partial \theta} = [1 - \chi_0 K]^{-1} \chi_0 \frac{\partial H_{a\theta}}{\partial \theta}$$

- Parameters appear in innermost layer (model definition)
  - Each DFT model: Different derivatives $\frac{\partial H_{a\theta}}{\partial \theta}$ (can be horrible)
  - Each quantity of interest: Different sensitivity derivative (1)

  $\Rightarrow$ Combinatorial explosion

- Opportunity of algorithmic differentiation (AD):
  - Generic framework for DFT derivatives / response properties
  - Saves manual coding: Request gradient (1), AD delivers

  $\Rightarrow$ New properties/derivatives by non-DFT experts!

Preview: Algorithmic differentiation in DFT practice

- Lattice constant: Optimal size of the unit cell:

$$a_* = \arg\min_a \; \mathcal{E}\left[\rho_{\mathrm{SCF}}(a, \theta)\right] \;\Rightarrow\; S(a_*, \theta) = 0$$

- How sensitive is a_* for a system? $\Rightarrow$ Need $\frac{da_*}{d\theta}$
- Annoyances for derivation and implementation:
  - Nested iterative methods (eigensolver, SCF, lattice optimisation)
  - Second-order derivatives ($\frac{\partial S}{\partial \theta} = \frac{\partial^2 \mathcal{E}}{\partial \theta \partial a}$) and $\frac{\partial S}{\partial a} = \frac{\partial^2 \mathcal{E}}{\partial a^2}$)
  - Minimise effort to try a new DFT model
- Today: First work in progress results

Lattice constant sensitivities of silicon

(Å)	a_*	κ	$\frac{da_*}{d\kappa}$	β	$\frac{da_*}{d\beta}$
expmnt.	5.421				
PBEsol	5.449	0.804	0.713	0.0375	0.0058
PBE	5.461	0.804	0.550	0.0667	0.0194
APBE	5.465	0.804	0.482	0.0790	0.0269
PBEmol	5.467	0.804	0.456	0.0838	0.0301
XPBE	5.466	0.920	0.603	0.0706	0.0184
rev-PBE	5.467	1.245	0.744	0.0667	0.0099

```
function compute_energy(a, theta)
    model = Model(a, PbeExchange(theta), ...)
    scf(model).energies.total
end
optimise_lattice(theta) = optimise(a -> compute_energy(a, theta))
sensitivities = ForwardDiff.jacobian(optimise_lattice, [kappa, beta])
```

- Generic framework: Flexible in model or targeted QoI
  - Solver derivatives automatically computed & chained as needed
  - Builds on DFTK architecture & julia AD tools
- Ongoing: Adjoint-mode AD implementation:
  - All parameter sensitivities by a single response problem

  $\Rightarrow$ Routine computation of model sensitivities

  $\Rightarrow$ Machine-learned solid-state DFT models

Conclusion and outlook

- High-throughput screening
  - Millions of calculations
  - Efficiency and reliability need to improve

  $\Rightarrow$ Need for error control and UQ

- Multi-scale simulation workflows
  - Top-top-bottom dependencies
  - Goal-oriented uncertainty analysis
  - Combination of statistical and analytical approaches

- Interdisciplinary research softwares
  - julia language: High-level & fast
  - Ongoing construction of UQ-enabled pipeline
  - DFTK: Routine access to model sensitivities

Applied and Computational Mathematics | RWTH AACHEN UNIVERSITY

Opportunities to learn more . . .

cecam workshop *(with G. Csányi, G. Dusson, Y. Marzouk)*:

"Error control in first-principles modelling"

- **20–24 June 2022** (hybrid & CECAM-HQ, Lausanne)

⇒ `https://www.cecam.org/workshop-details/1115`

DFTK school 2022 *(with E. Cancès, A. Levitt)*:

"Numerical methods for DFT simulations"

- **29–31 August 2022** at Sorbonne Université, Paris
- Centred around  **DFTK** and its multidisciplinary philosophy
- Grounds-up introduction of electronic structure theory, mathematical background, numerical methods, implementation
- Applications in method development & simulations

⇒ `https://school2022.dftk.org` Deadline: 30th April

Acknowledgements

https://michael-herbst.com/talks/2022.04.15_siamuq.pdf

École des Ponts

Antoine Levitt

Eric Cancès

RWTH

Benjamin Stamm

Markus Towara

MIT

Valentin Churavy

Jeremiah DeGreeff

Dallas Foster

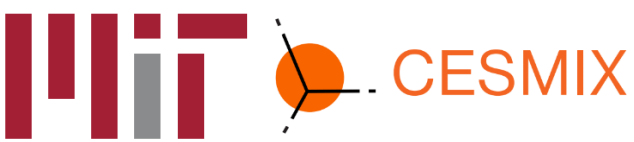
Emmanuel Luján

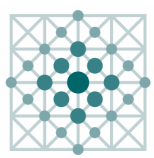
TU Berlin

Niklas Schmitz

all DFTK contributors

Summer of code





Applied and Computational Mathematics

RWTH AACHEN UNIVERSITY

Questions?

https://michael-herbst.com/talks/2022.04.15_siamuq.pdf


mfherbst


herbst@acom.rwth-aachen.de


https://michael-herbst.com/blog

https://dftk.org
https://school2022.dftk.org

https://www.cecam.org/workshop-details/1115

https://michael-herbst.com/learn-julia