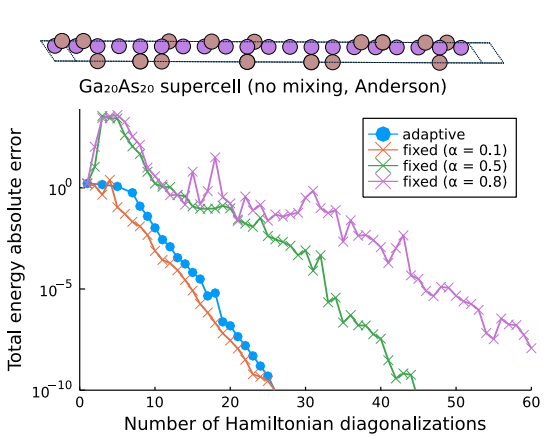
A robust and efficient line search for self-consistent field iterations

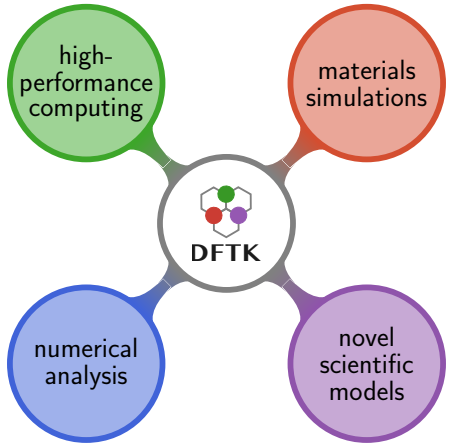
Michael F. Herbst*, Antoine Levitt§

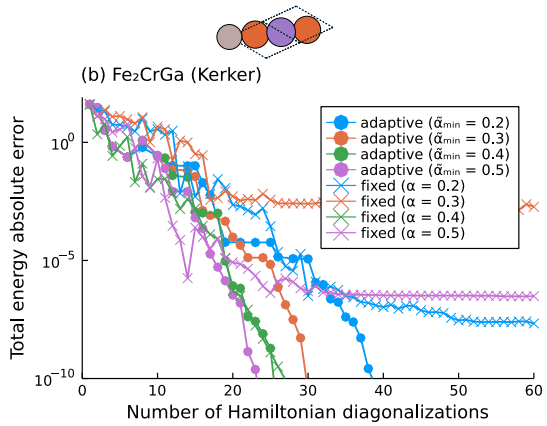
*Applied and Computational Mathematics, RWTH Aachen University
§CERMICS, École des Ponts and Inria Paris

13 January 2022

Slides: `https://michael-herbst.com/talks/2022.01.13_gdrnbody_linesearch.pdf`







Motivation: Reliable black-box electronic structure methods

- Virtual materials design $\Rightarrow$ millions of calculations:
  - design space search, data generation for surrogates, ...
- Key requirements:
  - Automation (saves human time)
  - Efficiency (saves computer time)
  - Reliability (saves computer time & human time)
- State of the art:
  - Many parameters to choose (algorithms, tolerances, models)
  - Choice by experience
  - Little error control (basically trial and error)
  - Workflow success rate: $\simeq 80 - 99\%$

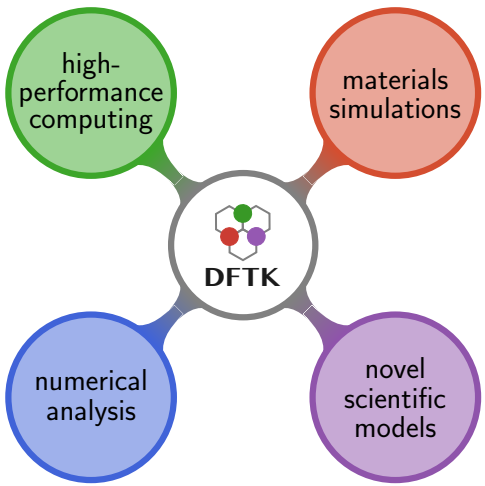
  $\Rightarrow$ Still tens of thousands of failed calculations!

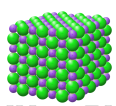
$\Rightarrow$ Robust, error-controlled algorithms:

- Ongoing effort centred around DFTK

Applied and Computational Mathematics | RWTH AACHEN UNIVERSITY

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





$H\Psi = E\Psi$

- julia code for plane-wave DFT
- Fully composable with julia ecosystem:
  - Arbitrary precision (32bit, >64bit, ...)
  - Algorithmic differentiation **new**
  - 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 > 800 electrons

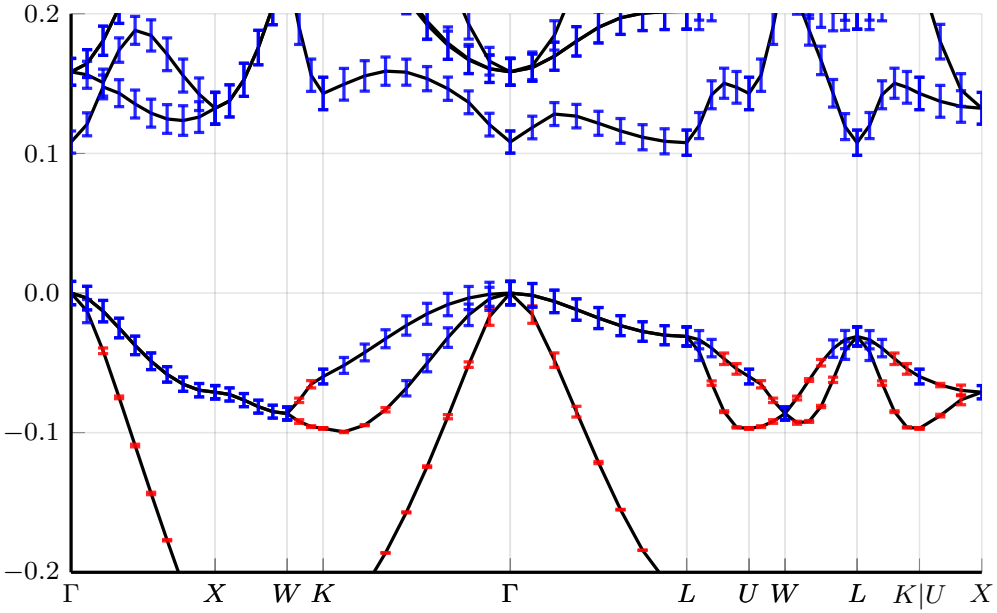
$\Rightarrow$ Build with multidisciplinary research in mind

- Avoids two-language problem: Just julia
- Development started in 2018
- Only 7k lines of code

$\Rightarrow$ Low entrance barrier across backgrounds

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

Rigorous error analysis: First results[1]



- Fully guaranteed error bounds for band structures
- This case: Reduced Kohn-Sham model
- Captures basis set error, floating-point error, convergence error

- Recent work also considers others quantities of interest[2]:
- E.g. densities and forces

[1] M. F. Herbst, A. Levitt and E. Cancès. Faraday Discuss. **224**, 227 (2020).

[2] E. Cancès, G. Dusson, G. Kemlin, A. Levitt. *Practical error bounds for properties in plane-wave electronic structure calculations* arXiv 2111.01470.

Algorithmic differentiation: UQ & data-driven approaches

- Data-enhanced density-functional theory methods
- High-order DFPT properties
- Sensitivity analysis and inference (uncertainties)

  $\Rightarrow$ Require efficient unusual, higher-order derivatives

- Combinatorial explosion:
  - "One PhD student per property" paradigm not feasible

  $\Rightarrow$ Use algorithmic differentiation ($\approx$ automatic derivatives)

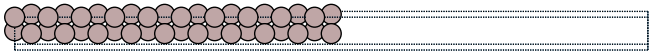
- Simple example: Stresses (Definition *vs.* julia code):[1]

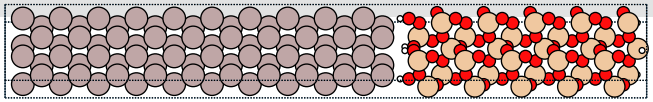
$$\text{Stress} = \frac{1}{\det(\mathbf{L})} \left.\frac{\partial E\left[P_*, (I + \mathbf{M})\,\mathbf{L}\right]}{\partial \mathbf{M}}\right|_{\mathbf{M}=0}$$

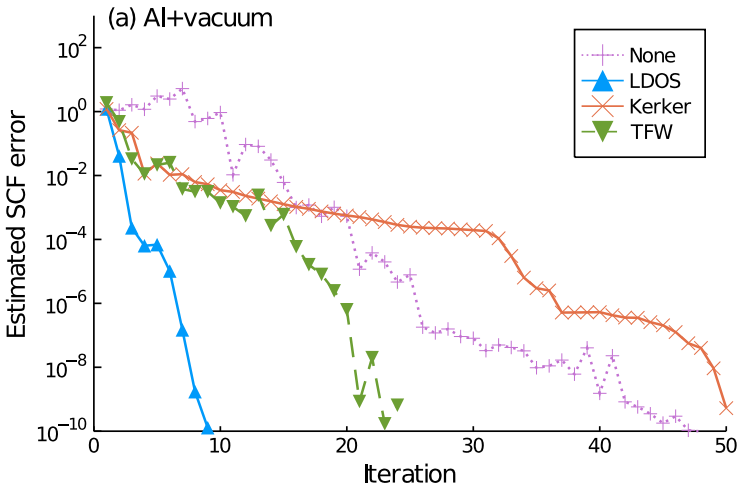
```
# Run SCF, get P*
scfres = self_consistent_field(basis)
     L = basis.model.lattice
stress = 1/det(L) * gradient(
    M -> recompute_energy(
                scfres, (I + M) * L),
    zero(L)
)
```

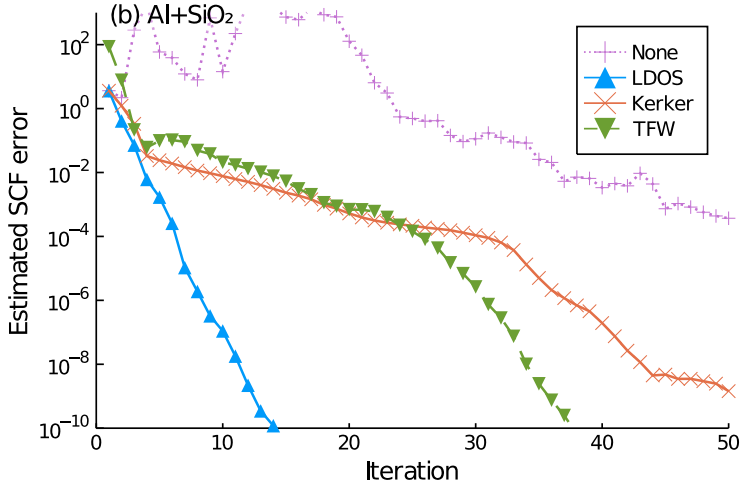
[1] Live code: `https://github.com/JuliaMolSim/DFTK.jl/blob/master/src/postprocess/stresses.jl`

Black-box algorithms: LDOS mixing[2]









- Long-standing problem: Suitable mixing for inhomogeneous systems
  - E.g. metal+insulator, catalytic surfaces, ...
- State-of-the-art: local Thomas-Fermi-von Weizsäcker mixing (TFW)[1]
- **DFTK** approach: LDOS mixing *automatically* interpolates between Kerker mixing (in the metallic region) and no mixing (insulating region)

⇒ Parameter-free and black-box

---

[1] D. Raczkowski, A. Canning, L. W. Wang, Phys. Rev. B. **64**, 121101 (2001).

[2] M. F. Herbst, A. Levitt. J. Phys. Condens. Matter **33**, 085503 (2021).

This work: Automatise damping selection

- DFT-like problems: $\min_{\rho} \mathcal{E}(\rho)$
- Duality of density and potential:

$$V(\rho) = \nabla_\rho E_{\mathsf{Hxc}}(\rho)$$
$$\rho(V) = \mathrm{diag}\left[f_{\mathsf{FD}}(H_0 + V)\right]$$

  with Fermi-Dirac f_{FD}, core Hamiltonian H_0, Hartree-XC energy $E_{\mathsf{Hxc}}(\rho)$

- Potential-mixing SCF procedure:

$$V_{n+1} = V_n + \alpha\, \delta V_n$$
$$\delta V_n = V(\rho(V_n)) - V_n$$

  with

- State of the art: Trial and error
  - Guess an α and make it smaller until convergence
  - Problematic in high-throughput context

$\Rightarrow$ Goal: Choose damping α *automatically*

This work: Automatise damping selection

- DFT-like problems: $\min_{\rho} \mathcal{E}(\rho)$
- Duality of density and potential:

$$V(\rho) = \nabla_\rho E_{\mathsf{Hxc}}(\rho)$$
$$\rho(V) = \mathrm{diag}\left[f_{\mathsf{FD}}(H_0 + V)\right]$$

with Fermi-Dirac f_{FD}, core Hamiltonian H_0, Hartree-XC energy $E_{\mathsf{Hxc}}(\rho)$

- Potential-mixing SCF procedure:

$$V_{n+1} = V_n + \alpha\, \delta V_n$$
$$\delta V_n = P^{-1}\left[V(\rho(V_n)) - V_n\right]$$

with mixing/preconditioner P

- State of the art: Trial and error
  - Guess an α and make it smaller until convergence
  - Problematic in high-throughput context

$\Rightarrow$ Goal: Choose damping α *automatically*

This work: Automatise damping selection

- DFT-like problems: $\min_{\rho} \mathcal{E}(\rho)$
- Duality of density and potential:

$$V(\rho) = \nabla_\rho E_{\mathsf{Hxc}}(\rho)$$
$$\rho(V) = \mathrm{diag}\left[f_{\mathsf{FD}}(H_0 + V)\right]$$

with Fermi-Dirac f_{FD}, core Hamiltonian H_0, Hartree-XC energy $E_{\mathsf{Hxc}}(\rho)$

- Potential-mixing SCF procedure:

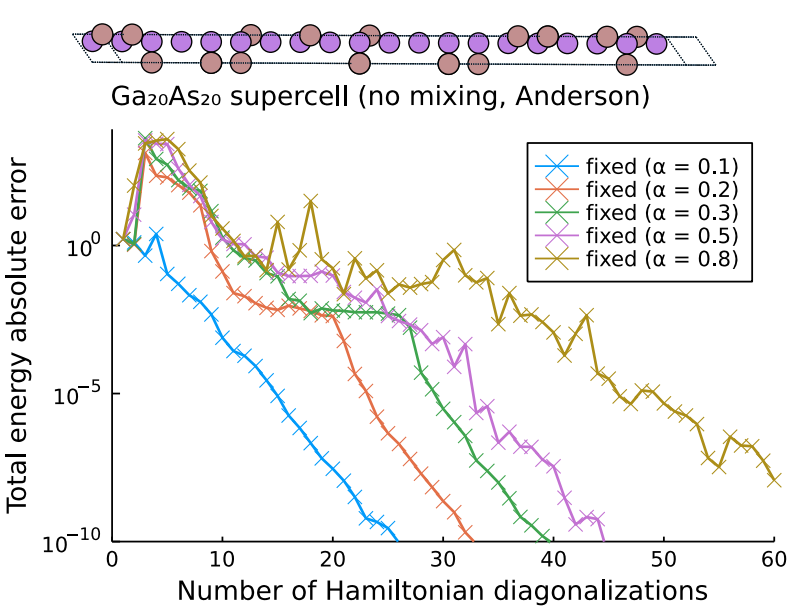
$$V_{n+1} = V_n + \alpha\, \delta V_n$$
$$\delta V_n = \mathsf{Anderson}\Big(P^{-1}\left[V(\rho(V_n)) - V_n\right]\Big)$$

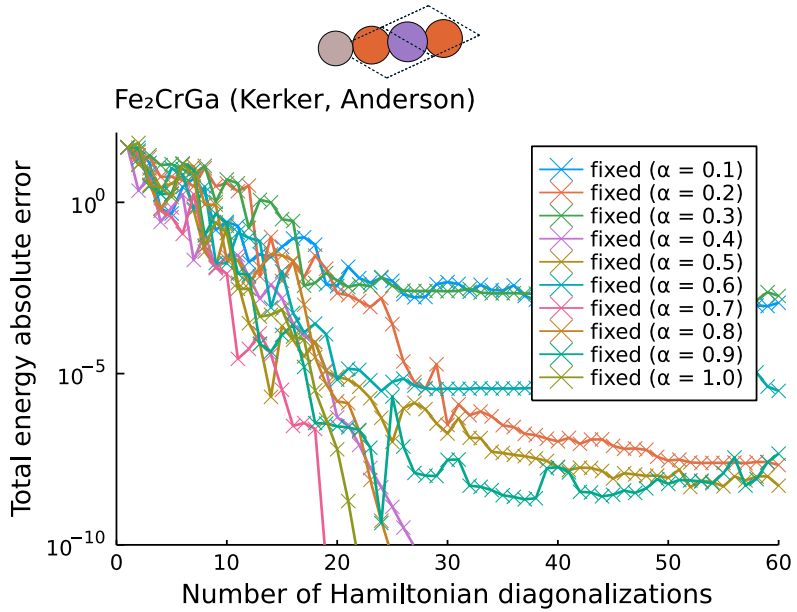
with mixing/preconditioner P, Anderson/DIIS acceleration

- State of the art: Trial and error
  - Guess an α and make it smaller until convergence
  - Problematic in high-throughput context

$\Rightarrow$ Goal: Choose damping α *automatically*

Guessing α can be hard (1)





- PBE functional / Anderson acceleration
- Non-linearities
- Anderson stagnation
- Unsuitable mixing

Guessing α can be hard (2)

System	Mixing	damping α									
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Al_{40} supercell	None	**38**	40	40	39	44	50	49	×	76	×
Al_{40} surface	None	**46**	48	50	49	51	60	61	66	89	×
$Ga_{20}As_{20}$ supercell	None	**26**	33	40	42	45	44	70	70	65	76
CoFeMnGa	Kerker	×	×	×	×	28	**21**	24	28	22	22
Fe_2CrGa	Kerker	×	×	×	27	×	×	**19**	25	×	22
Fe_2MnAl	Kerker	×	48	×	×	×	20	21	17	16	**15**
$FeNiF_6$	Kerker	×	×	×	×	×	×	×	23	22	**21**
Mn_2RuGa	Kerker	×	×	×	×	37	24	23	**22**	23	23
Mn_3Si	Kerker	×	×	×	×	26	30	22	**20**	×	×

- Number of iterations until energy converged to 10^{-10}
  - "×" is no convergence in 100 iterations
- Note: We focus on the 10% difficult cases
  - Metal supercells or surfaces without suitable mixing
  - Transition-metal compounds
  - Anderson issues due to non-linearities

$\Rightarrow$ Selecting suitable α can be challenging

Adaptive damping[1]: Key ideas

- Dual minimisation problem in V for DFT:

$$\min_V \mathcal{I}(V) = \min_V \mathcal{E}(\rho(V))$$

- Potential-mixing:
  - $V_n \rightarrow$ Find search direction δV_n
  - $V_{n+1} = V_n + \alpha\, \delta V_n$
- **Theorem** in discrete setting: Convergence if α small enough[1]
  - Caveat: Proof does not cover Anderson-type acceleration

$\Rightarrow$ Construct cheap model for $\mathcal{I}(V_n + \alpha\, \delta V_n)$

$\Rightarrow$ Use backtracking line search to fix α

- Start from trial damping $\tilde{\alpha}$
- Accept good steps (energy or SCF residual decreases)
- Otherwise: Shrink α and try again

[1] M. F. Herbst, A. Levitt. *A robust and efficient line search for self-consistent field iterations* arXiv 2109.14018.

Adaptive damping[1]: Key ideas

- Dual minimisation problem in V for DFT:

$$\min_V \mathcal{I}(V) = \min_V \mathcal{E}(\rho(V))$$

- Potential-mixing:
  - $V_n \rightarrow$ Find search direction δV_n
  - $V_{n+1} = V_n + \alpha\, \delta V_n$
- **Theorem** in discrete setting: Convergence if α small enough[1]
  - Caveat: Proof does not cover Anderson-type acceleration

$\Rightarrow$ Construct cheap model for $\mathcal{I}(V_n + \alpha\, \delta V_n)$

$\Rightarrow$ Use backtracking line search to fix α

- Start from trial damping $\tilde{\alpha}$
- Accept good steps (energy or SCF residual decreases)
- Otherwise: Shrink α and try again

[1] M. F. Herbst, A. Levitt. *A robust and efficient line search for self-consistent field iterations* arXiv 2109.14018.

Constructing the model (1)

- Use Taylor ...

$$\mathcal{I}(V_n + \boldsymbol{\alpha}\,\delta V_n) = \mathcal{I}(V_n) + \boldsymbol{\alpha}\left\langle \nabla\mathcal{I}_{|V=V_n} \middle| \delta V_n \right\rangle + \frac{\boldsymbol{\alpha}^2}{2}\left\langle \delta V_n \middle| \nabla^2\mathcal{I}_{|V=V_n}\,\delta V_n \right\rangle + O(\boldsymbol{\alpha}^3\,\|\delta V_n\|^3)$$

- ... and some algebra:

$$\nabla\mathcal{I}_{|V=V_n} = \chi_0(V_n)\Big[V(\rho(V_n)) - V_n\Big]$$

$$\nabla^2\mathcal{I}_{|V=V_n} = \chi_0(V_n)\Big[K(\rho(V_n))\,\chi_0(V_n) - 1\Big] + \chi_0'(V_n)\Big[V(\rho(V_n)) - V_n\Big]$$

where

- Hartree-XC kernel $K(\rho) = \nabla^2_\rho\, E_{\mathsf{Hxc}}(\rho)$
- Independent-particle susceptibility $\chi_0(V) = \rho'(V)$

Constructing the model (1)

- Use Taylor ...

$$\mathcal{I}(V_n + \alpha\,\delta V_n) = \mathcal{I}(V_n) + \alpha \left\langle \nabla\mathcal{I}_{|V=V_n} \middle| \delta V_n \right\rangle + \frac{\alpha^2}{2} \left\langle \delta V_n \middle| \nabla^2\mathcal{I}_{|V=V_n}\,\delta V_n \right\rangle + O(\alpha^3 \,\|\delta V_n\|^3)$$

- ... and some algebra:

$$\nabla\mathcal{I}_{|V=V_n} = \chi_0(V_n)\Big[V(\rho(V_n)) - V_n\Big]$$

$$\nabla^2\mathcal{I}_{|V=V_n} = \chi_0(V_n)\Big[K(\rho(V_n))\,\chi_0(V_n) - 1\Big] + \chi_0'(V_n)\underbrace{\Big[V(\rho(V_n)) - V_n\Big]}_{\simeq 0 \text{ near convergence}}$$

where

  - Hartree-XC kernel $K(\rho) = \nabla^2_\rho\, E_{\text{Hxc}}(\rho)$
  - Independent-particle susceptibility $\chi_0(V) = \rho'(V)$

Constructing the model (1)

- Use Taylor ...

$$\mathcal{I}(V_n + \alpha\,\delta V_n) = \mathcal{I}(V_n) + \alpha \left\langle \nabla\mathcal{I}_{|V=V_n} \middle| \delta V_n \right\rangle + \frac{\alpha^2}{2} \left\langle \delta V_n \middle| \nabla^2\mathcal{I}_{|V=V_n}\,\delta V_n \right\rangle + O(\alpha^3 \,\|\delta V_n\|^3)$$

- ... and some algebra:

$$\nabla\mathcal{I}_{|V=V_n} = \chi_0(V_n)\Big[V(\rho(V_n)) - V_n\Big]$$

$$\nabla^2\mathcal{I}_{|V=V_n} \simeq \chi_0(V_n)\Big[K(\rho(V_n))\,\chi_0(V_n) - 1\Big]$$

where

  - Hartree-XC kernel $K(\rho) = \nabla^2_\rho\, E_{\mathsf{Hxc}}(\rho)$
  - Independent-particle susceptibility $\chi_0(V) = \rho'(V)$

Constructing the model (2)

- Using the self-adjointness of χ_0 this yields:

$$\varphi_n(\alpha) = \mathcal{I}(V_n) + \alpha \langle V(\rho(V_n)) - V_n | \chi_0(V_n)\delta V_n \rangle + \frac{\alpha^2}{2}\Big[\langle \chi_0(V_n)\delta V_n | K(\rho(V_n))\chi_0(V_n)\delta V_n \rangle - \langle \delta V_n | \chi_0(V_n)\delta V_n \rangle \Big]$$

- Problem: Applying χ_0 at *each SCF step* is too expensive!
- Additional approximation:

$$\alpha\chi_0(V_n)\delta V_n = \rho(V_n + \alpha\delta V_n) - \rho(V_n) + O(\alpha^2 \|\delta V_n\|^2)$$

- Now: Expensive step is $\rho(V)$ (involves diagonalisation)
  - $\rho(V_n)$ known (needed to construct δV_n)
  - $\rho(V_n + \alpha\delta V_n) = \rho(V_{n+1})$ (if accepted) $\Rightarrow$ diagonalisation reuse

  $\Rightarrow$ More efficient than standard line search strategies (e.g. Armijo)

Constructing the model (2)

- Using the self-adjointness of χ_0 this yields:

$$\varphi_n(\alpha) = \mathcal{I}(V_n) + \alpha \langle V(\rho(V_n)) - V_n | \chi_0(V_n)\delta V_n \rangle + \frac{\alpha^2}{2}\Big[\langle \chi_0(V_n)\delta V_n | K(\rho(V_n))\chi_0(V_n)\delta V_n \rangle - \langle \delta V_n | \chi_0(V_n)\delta V_n \rangle \Big]$$

- Problem: Applying χ_0 at *each SCF step* is too expensive!
- Additional approximation:

$$\alpha\chi_0(V_n)\delta V_n = \rho(V_n + \alpha\delta V_n) - \rho(V_n) + O(\alpha^2 \|\delta V_n\|^2)$$

- Now: Expensive step is $\rho(V)$ (involves diagonalisation)
  - $\rho(V_n)$ known (needed to construct δV_n)
  - $\rho(V_n + \alpha\delta V_n) = \rho(V_{n+1})$ (if accepted) $\Rightarrow$ diagonalisation reuse

$\Rightarrow$ More efficient than standard line search strategies (e.g. Armijo)

Adaptive damping algorithm

Input: Current iterate V_n, search direction δV_n, trial damping $\tilde{\alpha}$
Output: Damping α_n, next iterate V_{n+1}

Initialise $\alpha_n \leftarrow \tilde{\alpha}$
loop
    Make tentative step $V_{n+1} = V_n + \alpha_n \delta V_n$
    Compute $\rho(V_{n+1}), \mathcal{I}(V_{n+1})$ (the expensive step)
    if accept V_{n+1} (details follow) **then**
        break
    else
        Build the coefficients of the model φ_n
        if model φ_n is good (details follow) **then**
            $\alpha_n \leftarrow \arg\min_\alpha \varphi_n(\alpha)$
            Scale α_n to ensure $|\alpha_n|$ is strictly decreasing
        else
            $\alpha_n \leftarrow \frac{\alpha_n}{2}$
        end if
    end if
end loop

Notes and details

- Step acceptance: Energy or residual decrease

$$\mathcal{I}(V_{n+1}) < \mathcal{I}(V_n) \quad \text{or} \quad V(\rho(V_{n+1})) - V_{n+1} < V(\rho(V_n)) - V_n$$

- φ_n is approximate $\Rightarrow$ Only use if prediction error

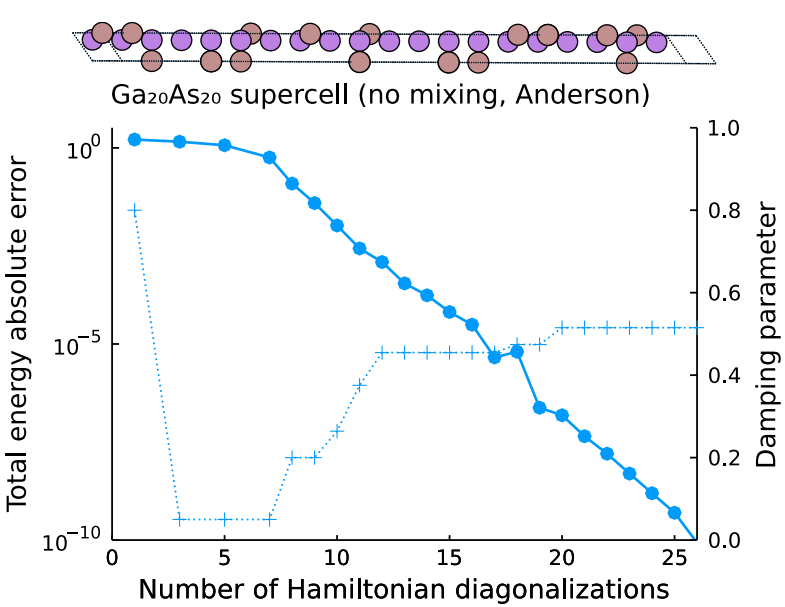
$$\frac{|\mathcal{I}(V_n + \boldsymbol{\alpha_n}\delta V_n) - \varphi_n(\boldsymbol{\alpha_n})|}{|\mathcal{I}(V_n + \boldsymbol{\alpha_n}\delta V_n) - \mathcal{I}(V_n)|} \quad \text{is small}$$

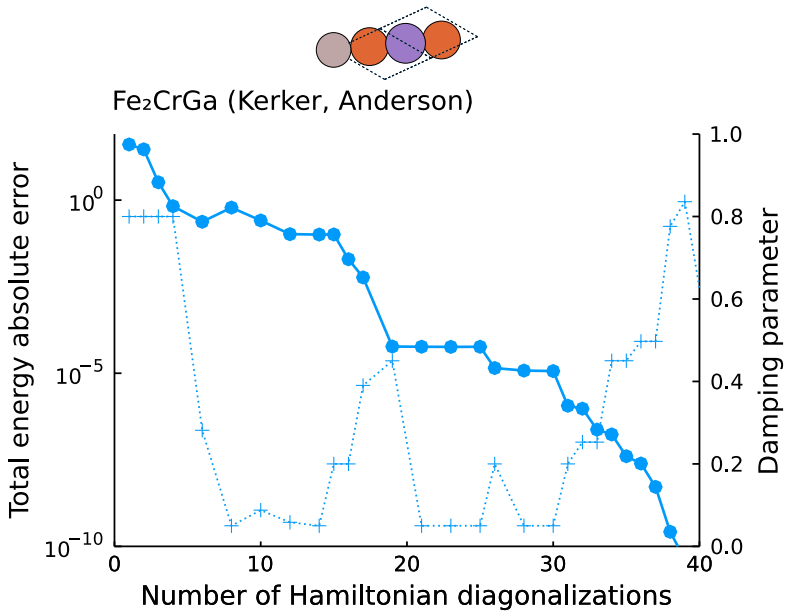
- Adaptive damping costs:
  - Only one diagonalisation per line search iteration
  - Good trial damping $\tilde{\boldsymbol{\alpha}}$ is crucial

$\Rightarrow$ Dynamical trial damping $\tilde{\boldsymbol{\alpha}}$ adjustment:

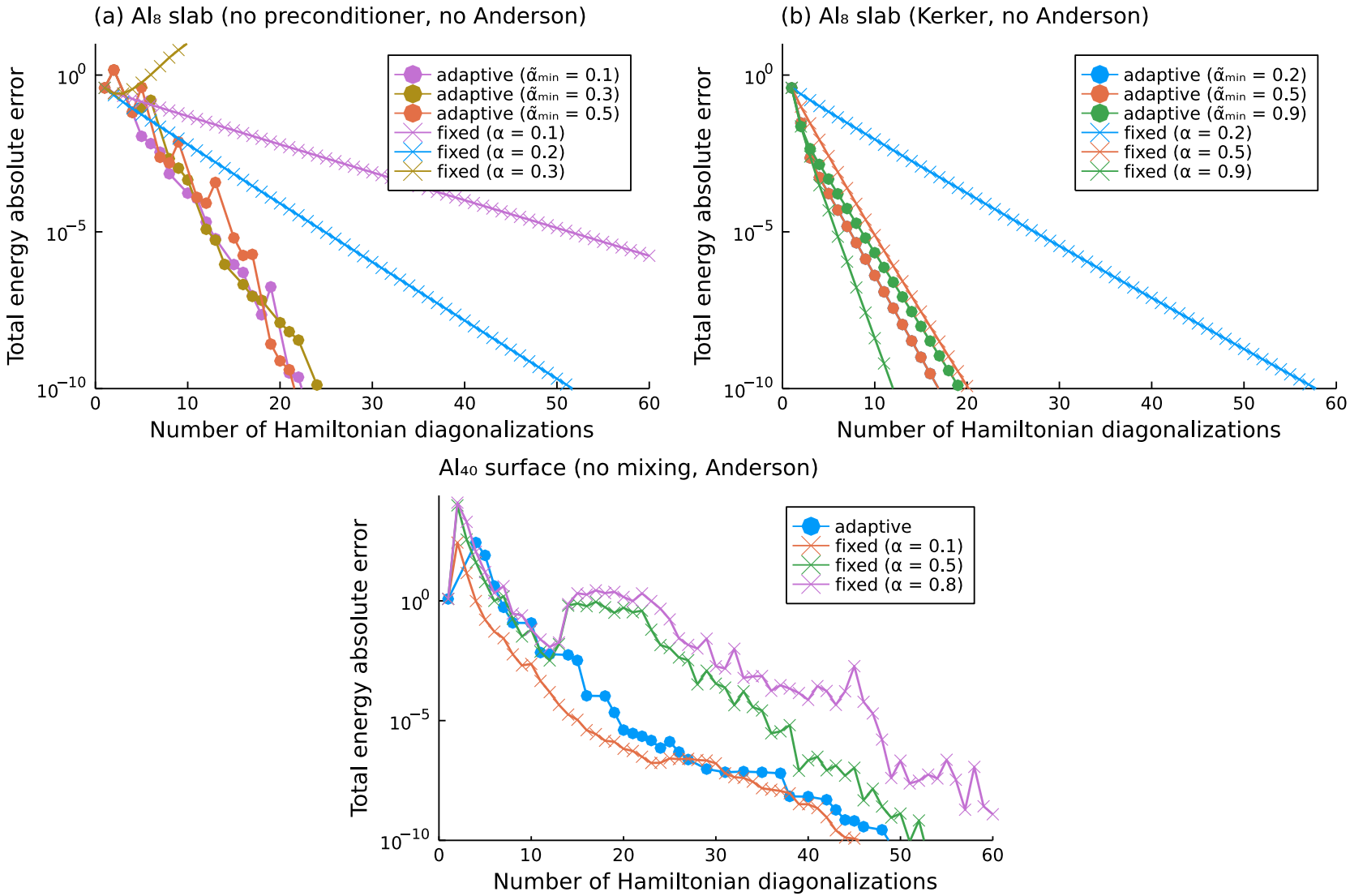
- Generally reuse damping from previous SCF step (shrinks $\boldsymbol{\alpha_n}$)
- If immediately successful ($\boldsymbol{\alpha_n} = \tilde{\boldsymbol{\alpha}}$), use φ_n to grow $\tilde{\boldsymbol{\alpha}}$
- Ensure $\boldsymbol{\alpha_n} \geq \tilde{\alpha}_{\text{min}} = 0.2$ (otherwise stagnation)

Damping parameter adaptation

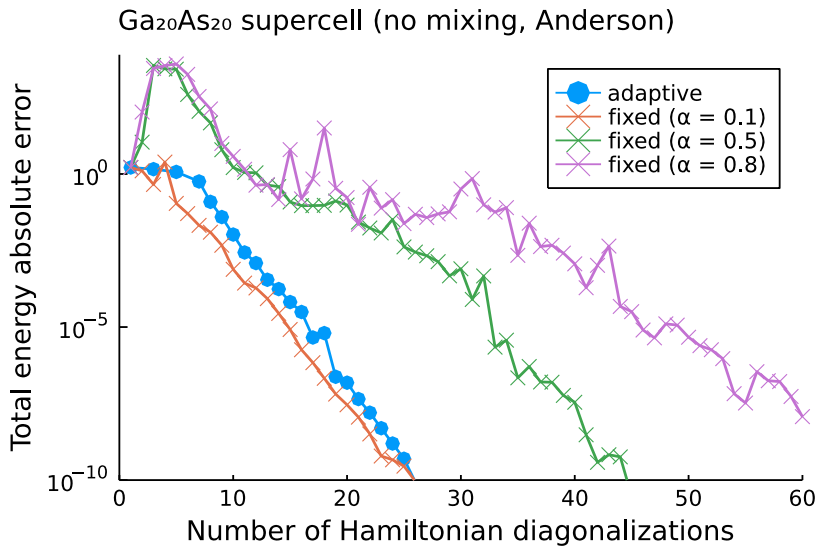




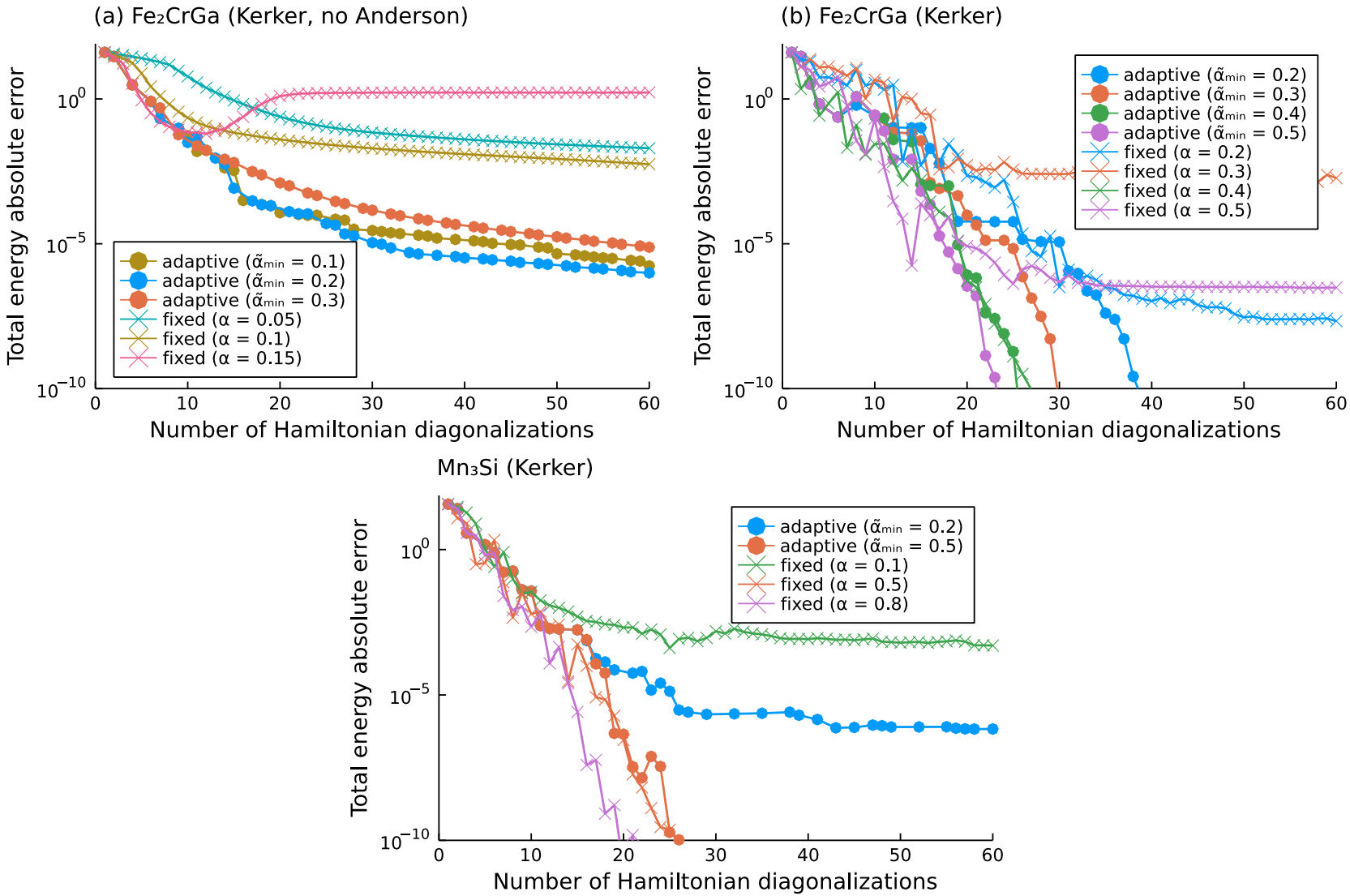
Aluminium slab and surface (simple systems)



GaAs slab (non-linear system)



Fe_2CrGa and Mn_3Si (Heusler systems)



Results overview

System	Precond.	fixed damping α										adaptive damping
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	
Al_8 slab	Kerker†	×	58	37	27	21	16	13	**11**	12	18	17
Al_8 slab	None†	×	**52**	×	×	×	×	×	×	×	×	24
Al_{40} slab	Kerker	19	15	14	12	**11**	12	12	12	12	12	12
Al_{40} slab	None	**38**	40	40	39	44	50	49	×	76	×	44
Al_{40} surface	None	**46**	48	50	49	51	60	61	66	89	×	49
$Ga_{20}As_{20}$ slab	None	**26**	33	40	42	45	44	70	70	65	76	26
CoFeMnGa	Kerker	×	×	×	×	28	**21**	24	28	22	22	30
Fe_2CrGa	Kerker	×	×	×	27	×	×	**19**	25	×	22	39
Fe_2MnAl	Kerker	×	48	×	×	×	20	21	17	16	**15**	34
$FeNiF_6$	Kerker	×	×	×	×	×	×	×	23	22	**21**	24
Mn_2RuGa	Kerker	×	×	×	×	37	24	23	**22**	23	23	36
Mn_3Si	Kerker	×	×	×	×	26	30	22	**20**	×	×	×

Number of diagonalisations

- Aluminium systems:
  - No user-chosen parameter, overhead for "easy" cases small
  - Adaptive damping less susceptible to non-matching preconditioner
- Gallium arsenide:
  - Interference of strong non-linearity with Anderson reduced
- Heusler systems:
  - Manually selecting α very challenging $\Rightarrow$ Overhead of adaptive reasonable
  - Albeit mismatching design principle, stability improvements

Conclusion

- Manual damping selection features trial and error
- Alternatives for plane-wave discretisations not straightforward
  - E.g. standard line searches / ODA / EDIIS
- **Adaptive damping** improves stability at no/little extra cost
  - No user intervention needed
  - Combinable with existing effective SCF techniques
  - Small overhead for "easy cases" (safeguard mechanism)
  - Related to schemes with proved convergence guarantees
  - Ready-to-use implementation in  DFTK
- Limitations of Anderson acceleration poorly understood:
  - Manual damping selection challenging
  - Adaptive damping improves stability
  - Further research ongoing

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`

Acknowledgements

https://michael-herbst.com/talks/2022.01.13_gdrnbody_linesearch.pdf

CERMICS

Antoine Levitt

Eric Cancès

EPFL

Marnik Bercx

Nicola Marzari

RWTH

Benjamin Stamm

all DFTK contributors

Summer of code

Applied and Computational Mathematics

RWTH AACHEN UNIVERSITY

Questions?

https://michael-herbst.com/talks/2022.01.13_gdrnbody_linesearch.pdf

Preprint arXiv 2109.14018

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

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

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

mfherbst

https://michael-herbst.com/blog

herbst@acom.rwth-aachen.de

Applied and Computational Mathematics | RWTH AACHEN UNIVERSITY