GW in practice with ABINIT

M. Giantomassi

Université Catholique de Louvain
Louvain-la-Neuve, Belgium

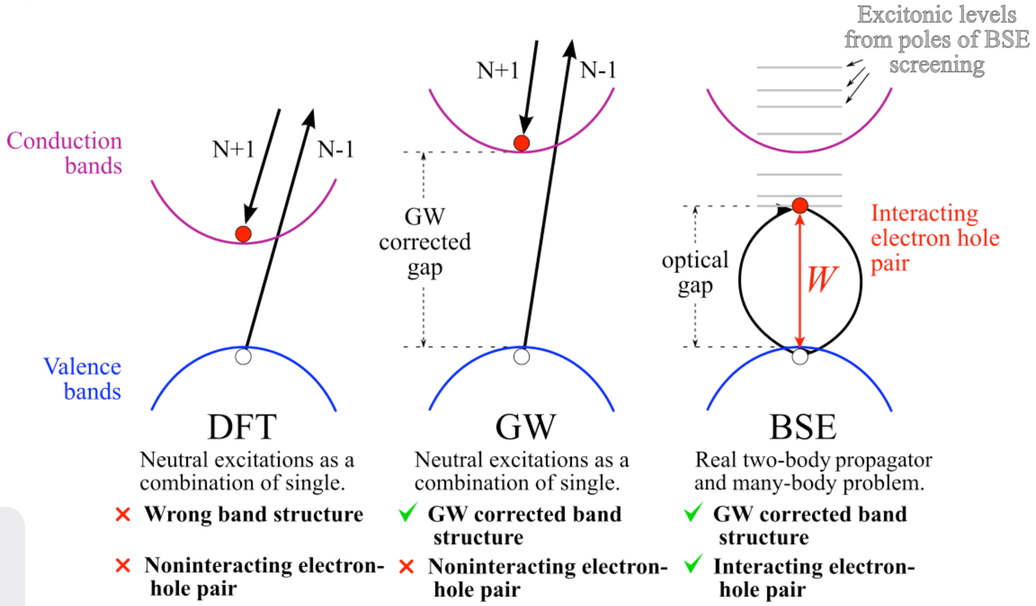
Theoretical Spectroscopy Lectures

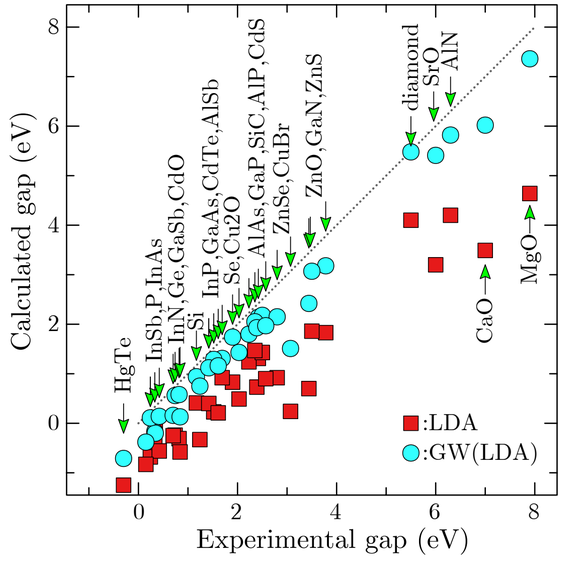
Lausanne, Switzerland

21-25 March, 2022

Why do we need *GW*?

- LDA/GGA **systematically underestimate** exp. band gaps
- Hedin's equations provide a rigorous approach to study excitation energies within MBPT
- Hedin's equations are extremely hard to solve
  - We usually employ a simplified version *e.g.* *GW* in the one-shot version (G_0W_0)
  - G_0W_0 improves bands gaps and dispersions by including screening and exchange effects beyond Kohn-Sham (KS) theory
  - G_0W_0 still undershoots gaps: lot of discussions on starting point, vertex-corrections, self-consistency, e-ph contributions, etc.
- For accurate optical properties, we need to go beyond *GW* and include e-h interaction via *e.g.* the Bethe-Salpeter equation (BSE)



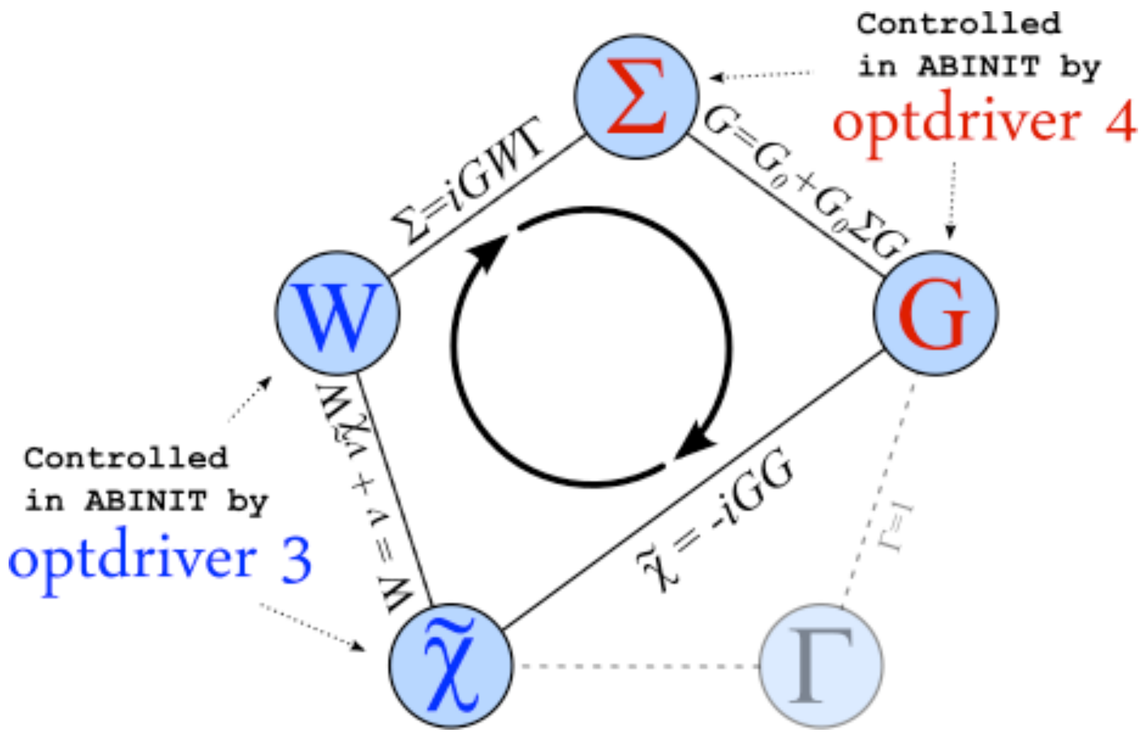


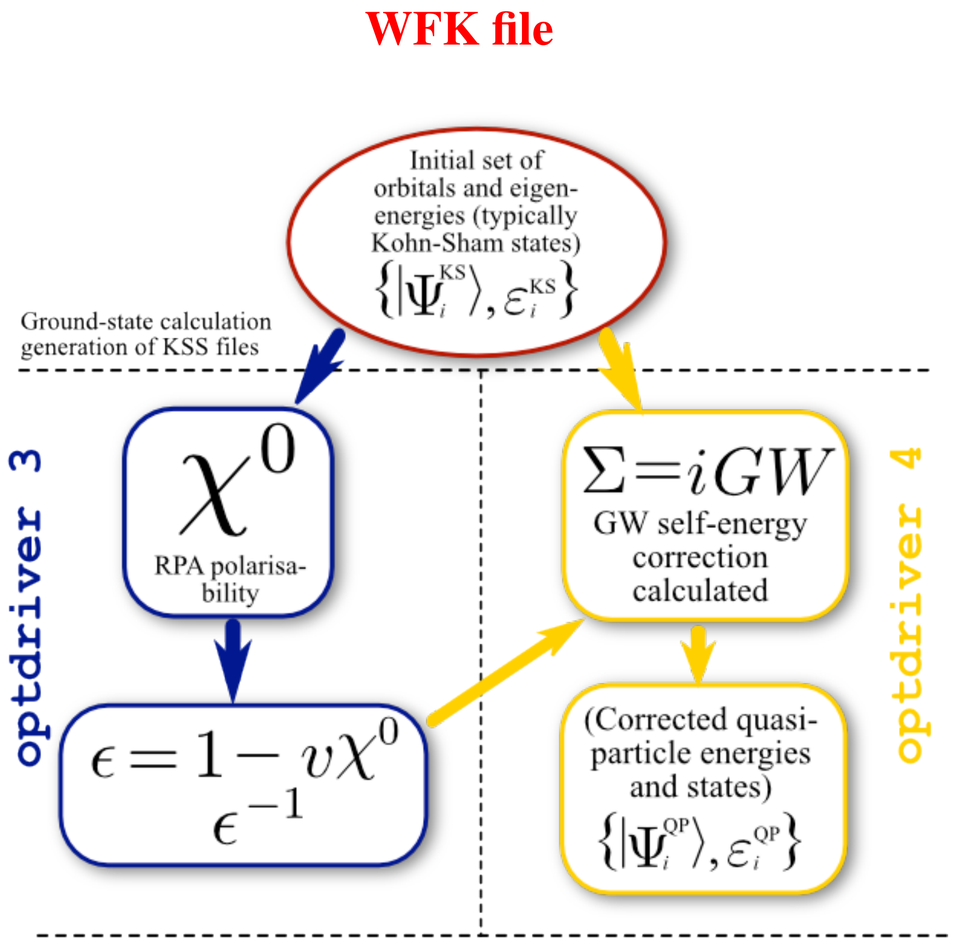
[adapted from van Schilfgaarde *et al.*, PRL **96**,

Capabilities of the ABINIT-*GW* code

- Formalism in Fourier space ($\mathbf{G}$ and ω) at $T = 0\,K$
- Norm-conserving pseudos (*recommended*) and PAW
- Semiconductors and metals
- Spin-polarized systems
- RPA with Adler-Wiser expression or Hilbert transform
- Different approximations for Σ (*HF, GW, COHSEX* ...)
- Four different plasmon-pole models (PPMs)
- Contour deformation (CD) or analytic continuation (AC) for $\Sigma(\omega)$
- Different level of self-consistency: $G^0W^0, G^0W, GW^0, GW, qpSCGW$
- Energy-only self-consistency or energies + wavefunctions
- MPI parallelism with memory distribution over bands

GW flowchart with ABINIT





- **optdriver** defines the runlevel (*screening*, *sigma*, …)
- Steps are connected via files and input variables: **getwfk** or **getwfk_filepath, getscr_filepath …**
- **gwcalctyp** defines the approximation for Σ and the self-consistency mode

Understanding gwcalctyp

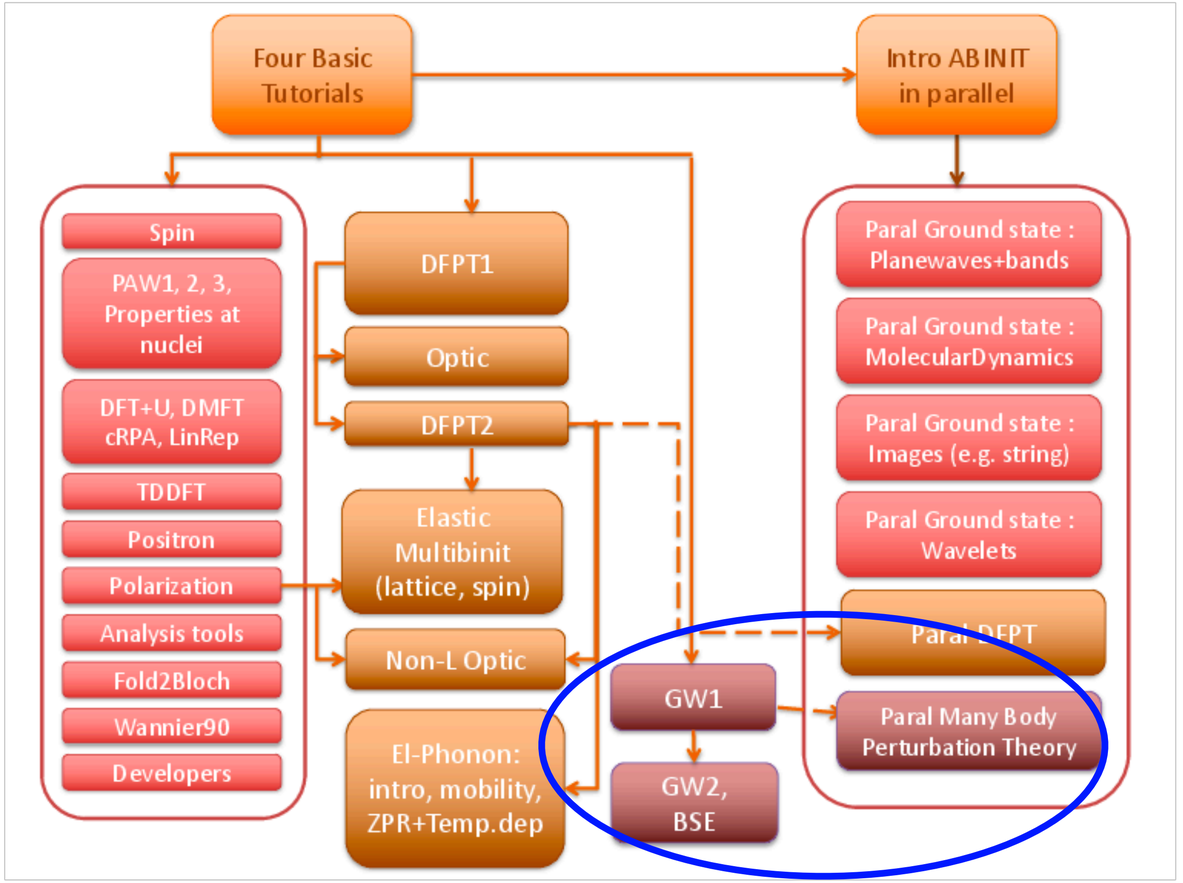
- **gwcalctyp** governs the choice between the different capabilities of the *GW* code
- It consists of two digits **AB**
    - **A** defines the self-consistency type:
        - 0 for one-shot calculation
        - 1 for updating energies only
        - 2 for updating both energies and wavefunctions
    - **B** defines the expression for Σ and how to treat the frequency convolution
        - 0 for plasmon-pole model
        - 1 for analytic continuation
        - 2 for contour-deformation
        - …
- Default value is 00 *i.e.* G_0W_0 with plasmonpole model (only two frequencies computed and stored in the SCR file)

Tutorials and documentation

https://docs.abinit.org/tutorial/

Three *GW* tutorials:

- *GW1*: convergence studies for silicon with PPM
- *GW2*: aluminium. CD and spectral function
- *Paral MBPT*: CD with Hilbert transform + technical details on the MPI implementation



All *GW* variables documented at:

https://docs.abinit.org/variables/gw/

See also the theory notes at:

https://docs.abinit.org/theory/mbt/

AbiPy examples:

Abitutorials https://github.com/abinit/abitutorials

GoWo lesson with AbiPy

How to post-process the SIGRES file

Plane-waves, symmetries and BZ

Plane-wave expansion of orbitals

- For periodic systems with lattice vectors **R**, Bloch theorem's states:

$$\psi_{n\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} u_{n\mathbf{k}}(\mathbf{r}) \quad \text{with} \quad u_{n\mathbf{k}}(\mathbf{r}+\mathbf{R}) = u_{n\mathbf{k}}(\mathbf{r})$$

- If we define the reciprocal lattice with lattice vectors **G**, such that:

$$e^{i\mathbf{G}\cdot\mathbf{R}} = 1$$

the periodic part of the Bloch's function can be written:

$$u_{n\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{G}} u_{n\mathbf{k}}(\mathbf{G}) e^{i\mathbf{G}\cdot\mathbf{r}}$$

where the coefficients $u_{nk}(\mathbf{G})$ are obtained by a Fourier transform:

$$u_{n\mathbf{k}}(\mathbf{G}) = \frac{1}{\Omega} \int_{\Omega} u_{n\mathbf{k}}(\mathbf{r}) e^{-i\mathbf{G}\cdot\mathbf{r}} d\mathbf{r}$$

Unit cell

- The basis set is truncated such that $\frac{\|\mathbf{k}+\mathbf{G}\|^2}{2} < E_{\text{cut}}$

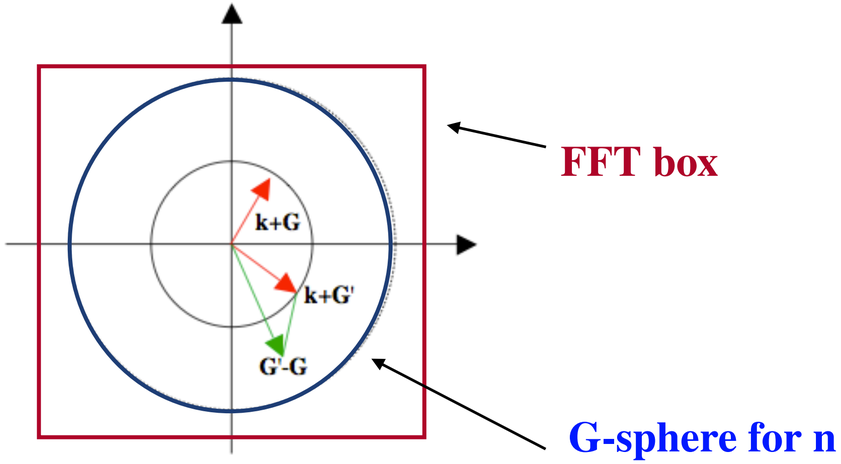
Convolution theorem

- Density associated with one eigenfunction: $n_{b\mathbf{k}}(\mathbf{r}) = u^*_{b\mathbf{k}}(\mathbf{r})\, u_{b\mathbf{k}}(\mathbf{r})$

- In Fourier space: $$n_{b\mathbf{k}}(\mathbf{r}) = \Big(\sum_{\mathbf{G}} u^*_{b\mathbf{k}}(\mathbf{G}) e^{-i\mathbf{G}\mathbf{r}}\Big)\Big(\sum_{\mathbf{G}'} u_{b\mathbf{k}}(\mathbf{G}') e^{i\mathbf{G}'\mathbf{r}}\Big) = \sum_{\mathbf{G}\mathbf{G}'} \left[u^*_{b\mathbf{k}}(\mathbf{G}) u_{b\mathbf{k}}(\mathbf{G}')\right] e^{i(\mathbf{G}'-\mathbf{G})\mathbf{r}}$$

convolution

- The radius of the **G**-sphere for $n(\mathbf{G})$ is **twice** the radius of the **G**-sphere used for the wavefunctions



Take-home message:

- Product in **r**-space —> convolution in **G**-space
- The FFT mesh should enclose the sphere of radius 2 * G_{max} to treat the convolution exactly

Plane-wave expansion of two-point functions

- Infinite system simulated with Born-von-Karman (BvK) periodic boundary conditions *i.e.* (N_1, N_2, N_3) supercell of volume $V = N\Omega$ with $N = N_1 N_2 N_3$
- All our MBPT functions are defined in the BvK supercell
- $G, \tilde{\chi}, W$ are invariant if we translate **both** $\mathbf{r}$ and $\mathbf{r}$' by $\mathbf{R}$, that is: $G(\mathbf{r}_1, \mathbf{r}_2) = G(\mathbf{r}_1 + \mathbf{R}, \mathbf{r}_2 + \mathbf{R})$
- This implies the following Fourier expansion:

$$f(\mathbf{r}_1, \mathbf{r}_2) = \frac{1}{V} \sum_{\substack{\mathbf{q} \\ \mathbf{G}_1\mathbf{G}_2}} e^{i(\mathbf{q}+\mathbf{G}_1)\cdot\mathbf{r}_1} f_{\mathbf{G}_1\mathbf{G}_2}(\mathbf{q})\, e^{-i(\mathbf{q}+\mathbf{G}_2)\cdot\mathbf{r}_2}$$

$$f_{\mathbf{G}_1\mathbf{G}_2}(\mathbf{q}) = \frac{1}{V} \iint_V e^{-i(\mathbf{q}+\mathbf{G}_1)\cdot\mathbf{r}_1} f(\mathbf{r}_1, \mathbf{r}_2)\, e^{i(\mathbf{q}+\mathbf{G}_2)\cdot\mathbf{r}_2}\, \mathrm{d}\mathbf{r}_1\, \mathrm{d}\mathbf{r}_2$$

where the $\mathbf{q}$-points belong to the BZ mesh that is dual to the BvK supercell: $(\frac{1}{N_1}, \frac{1}{N_2}, \frac{1}{N_3})$

- If G is expanded with cutoff energy ecut then $\tilde{\chi} = GG$ will have components up to 4 ecut!

Crystal symmetries and k-points

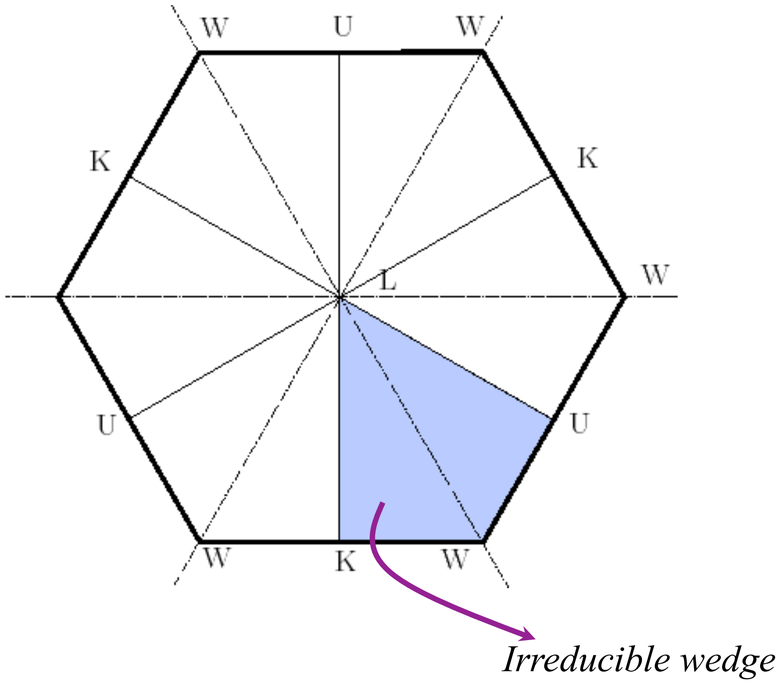
- Wavefunctions and eigenvalues in the full Brillouin zone (BZ) can be reconstructed from the **irreducible wedge** (IBZ)
- S = Rotation, **t** = fractional translation

$$\begin{cases} \epsilon_{\mathcal{S}\mathbf{k}} & = & \epsilon_{\mathbf{k}} \\ u_{\mathcal{S}\mathbf{k}}(\mathbf{r}) & = & e^{-i\mathcal{S}\mathbf{k}\cdot\mathbf{t}}\, u_{\mathbf{k}}(\mathcal{S}^{-1}(\mathbf{r}-\mathbf{t})) \\ u_{\mathcal{S}\mathbf{k}}(\mathbf{G}) & = & e^{-i(\mathcal{S}\mathbf{k}+\mathbf{G})\cdot\mathbf{t}}\, u_{\mathbf{k}}(\mathcal{S}^{-1}\mathbf{G}) \end{cases}$$

Spatial symmetries

$$\begin{cases} \epsilon_{n\mathbf{k}} & = & \epsilon_{-\mathbf{k}} \\ u_{n\mathbf{k}}(\mathbf{r}) & = & u^{\dagger}_{n-\mathbf{k}}(\mathbf{r}) \\ u_{n\mathbf{k}}(\mathbf{G}) & = & u^{\dagger}_{n-\mathbf{k}}(-\mathbf{G}) \end{cases}$$

Time-reversal symmetry



Take-home message:

- $u_{n\mathbf{k}}(\mathbf{G})$ are computed and stored only for **nkpt** **k**-points in the IBZ
- The higher the number of symmetries **nsym**, the faster the calculation
- Space group is automatically detected, all symmetries are used by default (**kptopt 1**)
- ABINIT may find less symmetries than expected if lattice and positions are not given with enough digits

NSCF run to generate the WFK file

Generation of the WFK file with empty bands

- Expensive step so we recommend to:
  - start immediately with reasonably large **nband** (>> 10 nband_occ)
  - perform initial convergence studies
  - generate new WFK if nband is not enough
- Use **nbdbuf** to save a lot of time as high-energy states converge slowly
- The WFK defines the BZ sampling and the list of **k**-points where QP corrections can be computed

```
getden_filepath "prefix_DEN"
iscf -2          # NSCF
tolwfr 1e-18
nband 1200  # occ + empty
nbdbuf 120   # ~10% of nband
nstep 100    # default is too small
```

NB: only the first nband - nbdbuf states are converged within **tolwfr**

Best practices:

- First compute the KS bands to locate the CBM/VBM then select the **k**-mesh (**ngkpt, nshiftk, shiftk**) accordingly
- Don't use datasets to run *GW* in a single run. Split everything and optimize the number of MPI procs for each step
- The default eigensolver (conjugate gradient, CG) cannot use more than **nkpt * nsppol** MPI cores
- Use **paral_kgb 1** (LOBPCG solver) if ncores > nkpt * nsppol …

LOBPCG eigensolver

- More scalable than CG: parallelized over **k**-points, bands, FFT, spins
- More difficult to configure (**npkpt, npband, npfft, bandpp**)
- Good news: memory scales with all MPI levels
- Use **autoparal 1** to let ABINIT find a good configuration for given number of MPI cores

```
getden_filepath "prefix_DEN"
iscf -2          # NSCF
tolwfr 1e-18
nband 1200  # occ + empty
nbdbuf 120   # ~10% of nband
nstep 100     # default is too small

paral_kgb 1
autoparal 1  #  Only for GS!
```

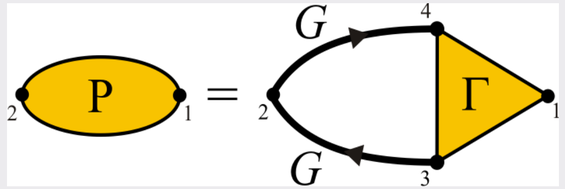
https://docs.abinit.org/tutorial/paral_gspw/

Best practices:

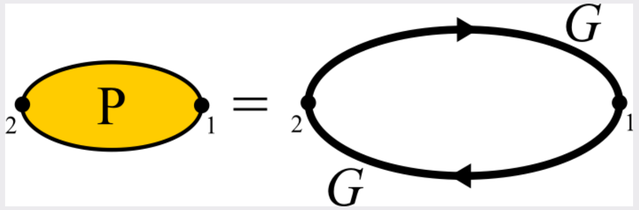
- Use an even number of MPI processors when **nsppol** 2
- Ideally, the number of MPI cores should be proportional to **nkpt * nsppol**
- Avoid prime numbers for **nband** as **npband** must divide nband
- **npband** and **bandpp** can affect the SCF convergence. Increasing bandpp usually makes the algorithm more stable

RPA in Fourier space

Hedin's irreducible polarizability



RPA polarizability



RPA in $\mathbf{r}/\omega$ space

- In the **Random Phase Approximation** (RPA), the irreducible polarizability reads:

$$\tilde{\chi}(\mathbf{r}_1, \mathbf{r}_2, \omega) = \frac{-i}{2\pi} \int G(\mathbf{r}_1, \mathbf{r}_2, \omega - \omega') G(\mathbf{r}_2, \mathbf{r}_1, \omega)\, e^{-i\eta\omega'} d\omega'$$

- Usually, we start from the Kohn-Sham G^0:

$$G^0(\mathbf{r}_1, \mathbf{r}_2, \omega) = \sum_{n\mathbf{k}} \frac{\psi_{n\mathbf{k}}(\mathbf{r}_1)\psi^*_{n\mathbf{k}}(\mathbf{r}_2)}{\omega - \varepsilon_{n\mathbf{k}} + i\eta\, \mathrm{sgn}(\varepsilon_{n\mathbf{k}} - \mu)}$$

Lehmann representation of the time-ordered G

where μ is the chemical potential and η a positive infinitesimal (**zcut**, default 0.1 eV)

- In $\mathbf{r}/\omega$ space, one obtains the so-called Adler-Wiser (AW) expression

$$\tilde{\chi}^0(\mathbf{r}_1, \mathbf{r}_2, \omega) = \sum_{n\mathbf{k}, m\mathbf{k}'} (f_{n\mathbf{k}} - f_{m\mathbf{k}'}) \frac{\psi_{n\mathbf{k}}(\mathbf{r}_1)\psi^*_{m\mathbf{k}'}(\mathbf{r}_1)\psi_{n\mathbf{k}}(\mathbf{r}_2)^*\psi_{m\mathbf{k}'}(\mathbf{r}_2)}{\omega - \varepsilon_{n\mathbf{k}} + \varepsilon_{m\mathbf{k}'} - i\eta\, \mathrm{sgn}(\varepsilon_{n\mathbf{k}} - \varepsilon_{m\mathbf{k}'})}$$

- Finally, we transform to reciprocal-space …

RPA polarizability in Fourier space

- In semiconductors with time-reversal symmetry, one obtains:

Controlled by **nkpt,kptopt,...**

Number of bands by **nband**

$$\chi^0{}_{\mathbf{G}_1\mathbf{G}_2}(\mathbf{q},\omega) = \frac{2}{V}\sum_{\mathbf{k}}^{\mathrm{BZ}}\sum_{cv} M^{cv}_{\mathbf{G}_1}(\mathbf{k},\mathbf{q})\left[M^{cv}_{\mathbf{G}_2}(\mathbf{k},\mathbf{q})\right]^{\dagger}$$

$$\times\left\{\frac{1}{\omega+\varepsilon_{c\mathbf{k}-\mathbf{q}}-\varepsilon_{v\mathbf{k}}-i\eta}-\frac{1}{\omega-\varepsilon_{c\mathbf{k}-\mathbf{q}}+\varepsilon_{v\mathbf{k}}+i\eta}\right\}$$

Matrix size given by **ecuteps**

Frequency grids(real, imaginary and complex) by **nfreqre, nfreqim, ...**

Small shift by **zcut**

with the oscillator matrix elements M given by:

Determined by **ecuteps**

$$M^{b_1b_2}_{\mathbf{G}}(\mathbf{k},\mathbf{q}) \equiv \langle \mathbf{k}-\mathbf{q}, b_1|e^{-i(\mathbf{q}+\mathbf{G})\cdot\mathbf{r}}|\mathbf{k}, b_2\rangle = \sum_{\mathbf{G}'} u^{\dagger}_{\mathbf{k}-\mathbf{q}b_1}(\mathbf{G}')u_{\mathbf{k}b_2}(\mathbf{G}+\mathbf{G}')$$

computed with FFT

Determined by **ecutwfn**

- By default, **ecutwfn** == **ecut.** Smaller values make the calculation faster but less accurate
- **fftgw** select the FFT mesh for M. **21** (default) treats the convolution exactly, **11** is faster but less accurate (aliasing effects)

Typical screening input for G_0W_0 with PPM

- **q**-mesh automatically defined by the **k**-mesh found in the WFK file
- Two frequencies: 0 and **ppmfreq** along the imaginary axis
- **ppmfreq** set to Drude plasma frequency if not specified
- Parameters to converge: **nband** and **ecuteps.**
- NB: These two parameters are **coupled,** see next slide

```
getwfk_filepath "prefix_WFK"
optdriver 3
ecuteps 6
nband 1000 # occ + empty
```

NB: gwcalctyp defaults to 00

Best practices:

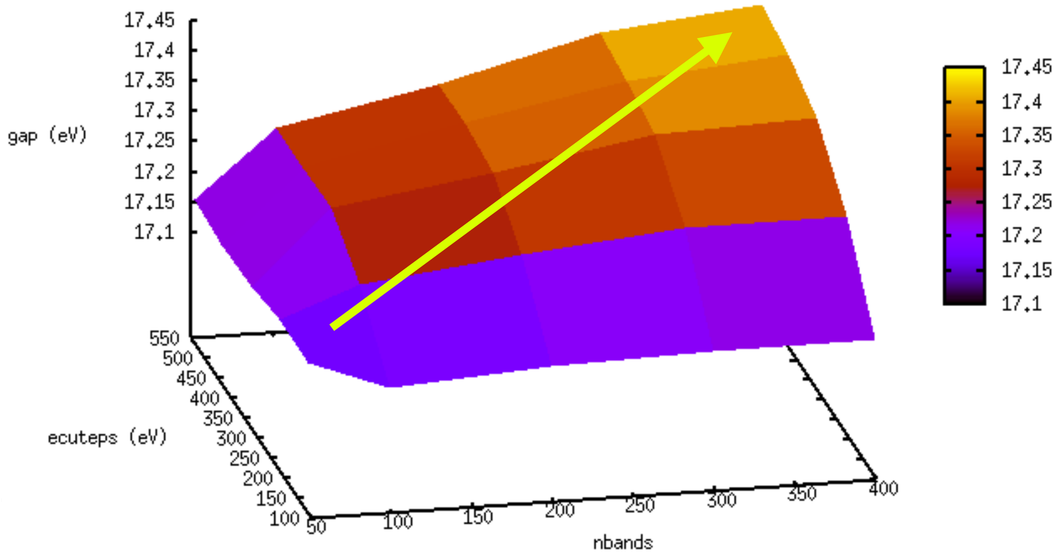
- Fix ecuteps to a reasonable value (e.g. 6-8 Ha) and generate multiple SCR files with different nband to reduce the number of computation. Then ecuteps can be changed in the sigma part
- Remember that nband cannot be greater than the number of bands stored in the WFK file
- For optimal distribution, the number of MPI cores should divide (nband - nband_occ)

M. J. van Setten et al. PRB B **2017**, *96*, 155207

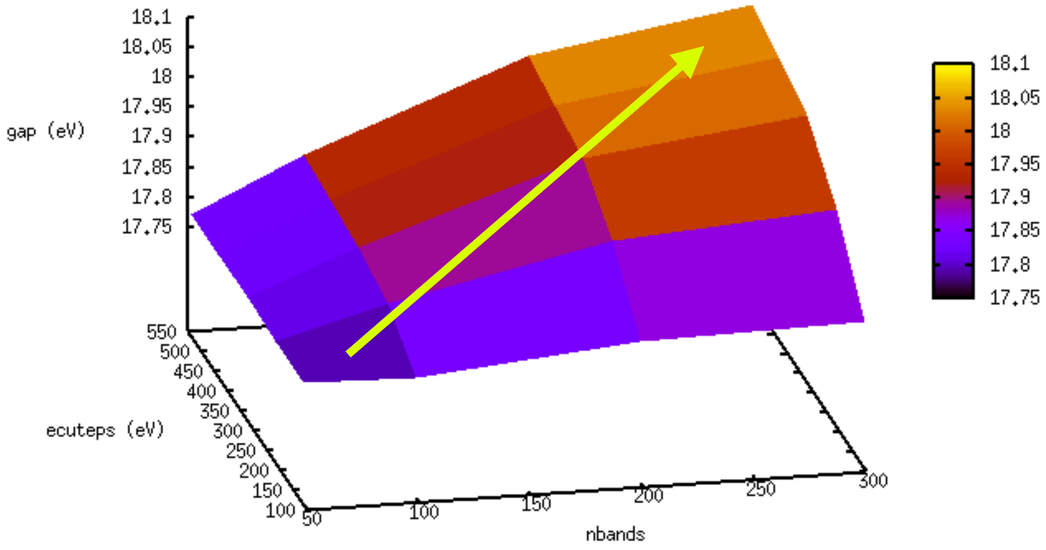
Cross convergence in gold

- Matrix elements with large |**G**| or |**G**'| get contributions from transitions involving high-energy states
- We need to increase both **nband** and **ecuteps**
- Increasing one parameter while keeping the other one fixed might lead to unconverged results (see figures)
- Fortunately, the convergence study wrt (**nband, ecuteps**) and the **k**-mesh can be partly decoupled

2 × 2 × 2 k-point



14 × 14 × 14 k-point



Best practices:

- Start with a coarse **k**-mesh e.g. 4x4x4
- Converge QP gaps wrt (**nband, ecuteps**)
- Finally, fix (**nband, ecuteps**) and converge the BZ sampling

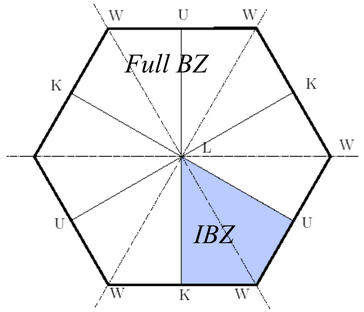
Symmetries in the RPA code

- The polarizability is invariant under all the spatial symmetries of the crystal:

$$\chi^0(\mathbf{r}_1, \mathbf{r}_2) = \chi^0\left(\mathscr{R}^{-1}(\mathbf{r}_1 - \mathbf{t}), \mathscr{R}^{-1}(\mathbf{r}_2 - \mathbf{t})\right)$$

- In **G**-space, one obtains the following equation:

$$\chi^0{}_{\mathbf{G}_1\mathbf{G}_2}(\mathscr{R}\mathbf{q}) = e^{i\mathbf{t}\cdot(\mathbf{G}_2-\mathbf{G}_1)}\chi^0{}_{\mathscr{R}^{-1}\mathbf{G}_1\ \mathscr{R}^{-1}\mathbf{G}_2}(\mathbf{q})$$



that allows one to reconstruct the matrix in the BZ from the IBZ

- Additional tricks can be used for $M^{cv}(\mathbf{k}, \mathbf{q})$ so that we don't need to sum over the BZ

Take-home message:

- $\chi^0(\mathbf{q})$ is computed only for the **q**-points in the IBZ
- For given **q**, the BZ integration can be restricted to the $\mathrm{IBZ}_{\mathbf{q}}$ defined by the *little-group* of **q** (**symchi** 1, default)
- High-symmetry **q**-points require less computing time

Hedin's equation for W in Fourier space

- In **r**/t space, the screened interaction W fulfils the integral equation (in matrix notation):

$$W = v + v\tilde{\chi}W$$

for which one obtains:

$$W = (1 - v\tilde{\chi})^{-1}v = \varepsilon^{-1}v$$

where we have introduced the dielectric matrix ε

$$\varepsilon(1,2) = \delta(1,2) - \int v(1,3)\tilde{\chi}(3,2)\,\mathrm{d}3$$

$$(1) \equiv (\mathbf{r_1}, t_1)$$

$$\delta(12) = \delta(\mathbf{r_1} - \mathbf{r_2})\,\delta(t_1 - t_2)$$

- In Fourier space, one obtains:

$$\varepsilon_{\mathbf{G}_1\mathbf{G}_2}(\mathbf{q};\omega) = \delta_{\mathbf{G}_1,\mathbf{G}_2} - v(\mathbf{q},\mathbf{G}_1)\tilde{\chi}_{\mathbf{G}_1\mathbf{G}_2}(\mathbf{q};\omega)$$

- The SCR file stores the (symmetrized) inverse dielectric matrix: $\varepsilon^{-1}_{\mathbf{G}_1\mathbf{G}_2}(\mathbf{q};\omega)\frac{|\mathbf{q}+\mathbf{G}_1|}{|\mathbf{q}+\mathbf{G}_2|}$

Long wavelength limit

- In semiconductors, the head and the wings of $\tilde{\chi}_{\mathbf{G},\mathbf{G}'}$ go to zero for $|\mathbf{q}| \to 0$

$$\underbrace{\lim_{\mathbf{q}\to 0} \tilde{\chi}_{00}(\mathbf{q},\omega) = 0}_{\text{Head}} \qquad \underbrace{\lim_{\mathbf{q}\to 0} \tilde{\chi}_{\mathbf{G}0}(\mathbf{q},\omega) = 0 \qquad \lim_{\mathbf{q}\to 0} \tilde{\chi}_{0\mathbf{G}}(\mathbf{q},\omega) = 0}_{\text{Wings}}$$

- At the level of the symmetrized ε, this leads to a $\frac{0}{0}$ form for $|\mathbf{q}| \to 0$
- The limit is finite but the value depends on the direction $\hat{\mathbf{q}}$
- To compute the limit we need $\lim_{\mathbf{q}\to 0} \langle b_1, \mathbf{k}-\mathbf{q} | e^{-i\mathbf{q}\cdot\mathbf{r}} | b_2, \mathbf{k}\rangle$
- The Taylor expansion of $e^{-i\mathbf{q}\cdot\mathbf{r}}$ leads to:

$$\langle b_1, \mathbf{k}-\mathbf{q} | e^{-i\mathbf{q}\cdot\mathbf{r}} | b_2, \mathbf{k}\rangle = \delta_{b_1 b_2} - i\mathbf{q}\cdot\langle b_1, \mathbf{k} | \mathbf{r} | b_2, \mathbf{k}\rangle + \mathcal{O}(q^2)$$

ill defined in periodic systems

- The matrix elements of $\mathbf{r}$ are therefore expressed in terms of the $[H, \mathbf{r}]$ commutator:

$$\langle b_1, \mathbf{k} | \mathbf{r} | b_2, \mathbf{k}\rangle = \frac{\langle b_1, \mathbf{k} | [H, \mathbf{r}] | b_2, \mathbf{k}\rangle}{\varepsilon_{b_1\mathbf{k}} - \varepsilon_{b_2\mathbf{k}}} \quad \text{for } b_1 \neq b_2$$

$$H = T + V_{Hxc}(\mathbf{r}) + V_{loc}(\mathbf{r}) + \underbrace{V_{NL}(\mathbf{r}, \mathbf{r}')}_{\text{pseudopotential part}}$$

- The final expression reads:

CPU demanding

$$\langle b_1, \mathbf{k}-\mathbf{q} | e^{-i\mathbf{q}\cdot\mathbf{r}} | b_2, \mathbf{k}\rangle \underset{\mathbf{q}\to 0}{\approx} -i\mathbf{q}\cdot\frac{\langle b_1, \mathbf{k} | \nabla - [V_{NL}, \mathbf{r}] | b_2, \mathbf{k}\rangle}{\varepsilon_{b_2\mathbf{k}} - \varepsilon_{b_1\mathbf{k}}} \quad \text{for } b_1 \neq b_2$$

Take-home message:

- The commutator $[V_{NL}, \mathbf{r}]$ is important for optical properties or for *GW* calculations in large cells. Less critical for *GW* in bulk systems
- This term is included by default (**inclvkb** 2). Use 0 to ignore it
- Heads and wings converge fast with **nband** and slowly with the number of **k**-points
- Randomly shifted **k**-meshes are usually used to converge the macroscopic dielectric function with/wo local-field (LF) effects:

$$\varepsilon_M^{\mathrm{LF}}(\omega) = \lim_{\mathbf{q}\to 0} \frac{1}{\varepsilon_{00}^{-1}(\mathbf{q},\omega)} \qquad \varepsilon_M^{\mathrm{NLF}}(\omega) = \lim_{\mathbf{q}\to 0} \varepsilon_{00}(\mathbf{q},\omega)$$

- In *GW*, one usually uses the same high-symmetry **k**-mesh both in screening and sigma to reduce **nkpt**

RPA with Hilbert transform

https://docs.abinit.org/tutorial/paral_mbt/

- Compute spectral function for several ω' (fast thanks to the Dirac δ):

$$\tilde{\chi}^{\mathcal{S}}_{\mathbf{G}_1\mathbf{G}_2}(\mathbf{q},\omega') = \frac{2}{V}\sum_{\substack{\mathbf{k}\\ b_1 b_2}} \left[f(\varepsilon_{b_1\mathbf{k}-\mathbf{q}}) - f(\varepsilon_{b_2\mathbf{k}})\right] \times M^{b_1 b_2}_{\mathbf{G}_1}(\mathbf{k},\mathbf{q})\left[M^{b_1 b_2}_{\mathbf{G}_2}(\mathbf{k},\mathbf{q})\right]^{\dagger} \mathrm{sign}(\omega')\,\delta(\omega' - \varepsilon_{b_1\mathbf{k}-\mathbf{q}} + \varepsilon_{b_2\mathbf{k}})$$

Number of spectral function frequencies set by **nomegasf**

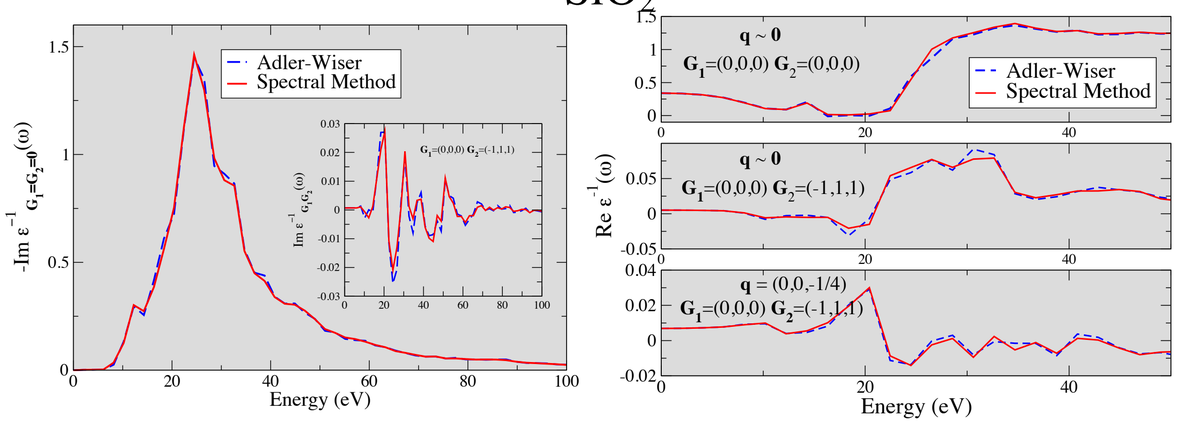
Approximation for delta function set by **spmeth**

- Use Hilbert transform (HT) to obtain the polarizability:

$$\chi^0{}_{\mathbf{G}_1\mathbf{G}_2}(\mathbf{q},\omega) = \int_0^{+\infty} \tilde{\chi}^{\mathcal{S}}_{\mathbf{G}_1\mathbf{G}_2}(\mathbf{q},\omega')\left(\frac{1}{\omega-\omega'+i\delta} - \frac{1}{\omega+\omega'-i\delta}\right)\mathrm{d}\omega'$$

- Much more efficient than the AW expression when several frequencies are needed
- Wall-time almost independent on N_ω

SiO_2



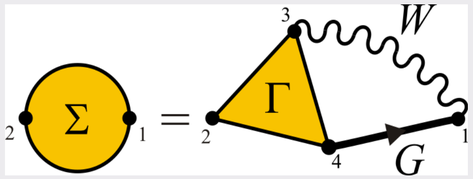
How to compute $\tilde{\chi}(\omega)$ for several frequencies

```
getwfk_filepath "prefix_WFK"
optdriver 3
ecuteps 6
nband 1000          # occ + empty

gwcalctyp   2       # Contour-deformation IMPORTANT

freqremax  40 eV    # Max real frequency (linear mesh)
nfreqre    20       # Number of real ω
nfreqim     5       # Number of imaginary ω.
                    # Gauss-Legendre mesh from 0 to +∞

# Activate spectral function
spmeth      1
nomegasf  100 # Number of points for the spectral function.
```

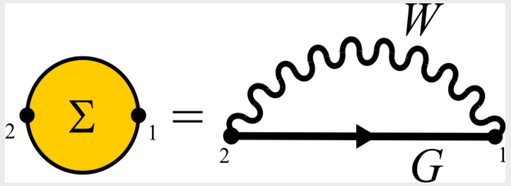
The size of the SCR file scales as npweps^2 * nqibz * nomega. This array is not MPI-distributed in the sigma part!

GW self-energy and QP energies

Hedin's self-energy



GW self-energy



The self-energy

- The self-energy is formally defined by:

$$G^{-1} = G_0^{-1} - \Sigma$$

- The poles of G are moved in energy by Σ compared to G_0.
- Σ includes all the interaction effects that are not captured by the mean-field G_0
- The equation can also be written as:

$$G = G_0 + G_0 \Sigma G$$

which is known as the **Dyson equation**.

- Many implementations, however, do not solve the Dyson equation directly
- From the Dyson equation, indeed, one can derive an effective single-particle problem referred to as the **quasiparticle equation** that is "easier" to solve…

Quasiparticle equation (I)

- Rather than computing G from Σ, we try to solve the **quasiparticle equation**:

$$\left[-\frac{1}{2}\nabla^2 + V_{\text{ext}}(\mathbf{r}) + V_{\text{H}}(\mathbf{r})\right]\phi^{\text{QP}}_{n\mathbf{k}}(\mathbf{r}) + \int \Sigma(\mathbf{r},\mathbf{r}',\omega = \varepsilon^{\text{QP}}_{n\mathbf{k}})\,\phi^{\text{QP}}_{n\mathbf{k}}(\mathbf{r}')d\mathbf{r}' = \varepsilon^{\text{QP}}_{n\mathbf{k}}\phi^{\text{QP}}_{n\mathbf{k}}(\mathbf{r})$$

- In general, the eigenvalues $\epsilon^{\text{QP}}_{n\mathbf{k}}$ are complex.
- The real part can be interpreted as the QP energy
- The imaginary part is related to the lifetime of the QP excitation due to e-e scattering:

$$\text{Im}\,\varepsilon_{n\mathbf{k}}(\omega) = 1/\tau^{\text{QP}}$$

- Note, however, that other scattering processes contribute to the lifetime: phonons, impurities, defects, etc.

Quasiparticle equation (II)

- The similarity between the quasiparticle equation:

$$\left[-\frac{1}{2}\nabla^2 + V_{\text{ext}}(\mathbf{r}) + V_{\text{H}}(\mathbf{r})\right]\phi^{\text{QP}}_{n\mathbf{k}}(\mathbf{r}) + \int \Sigma(\mathbf{r},\mathbf{r}',\omega=\varepsilon^{\text{QP}}_{n\mathbf{k}})\,\phi^{\text{QP}}_{n\mathbf{k}}(\mathbf{r}')d\mathbf{r}' = \varepsilon^{\text{QP}}_{n\mathbf{k}}\phi^{\text{QP}}_{n\mathbf{k}}(\mathbf{r})$$

  and the Kohn-Sham one:

$$\left[-\frac{1}{2}\nabla^2 + V_{\text{ext}}(\mathbf{r}) + V_{\text{H}}(\mathbf{r})\right]\phi^{\text{KS}}_{n\mathbf{k}}(\mathbf{r}) + V_{\text{xc}}(\mathbf{r})\phi^{\text{KS}}_{n\mathbf{k}}(\mathbf{r}) = \varepsilon^{\text{KS}}_{n\mathbf{k}}\phi^{\text{KS}}_{n\mathbf{k}}(\mathbf{r})$$

  suggests to treat the difference between Σ and V_{xc} as a perturbation with respect to the KS calculation

- In fact, the approximation $\phi^{\text{QP}}_{n\mathbf{k}}(\mathbf{r}) \simeq \phi^{\text{KS}}_{n\mathbf{k}}(\mathbf{r})$ is reasonable for many materials so that we can write:

$$\epsilon^{\text{QP}}_{n\mathbf{k}} = \epsilon^{\text{KS}}_{n\mathbf{k}} + \left\langle \phi^{\text{KS}}_{n\mathbf{k}} \middle| \Sigma(\mathbf{r},\mathbf{r}',\omega=\epsilon^{\text{QP}}_{n\mathbf{k}}) - V_{\text{xc}}(\mathbf{r}) \middle| \phi^{\text{KS}}_{n\mathbf{k}} \right\rangle$$

Linearized QP equation

- Since Σ depends on ω, we have to solve the **non-linear equation**:

$$\epsilon_{n\mathbf{k}}^{\mathrm{QP}} = \epsilon_{n\mathbf{k}}^{\mathrm{KS}} + \left\langle \phi_{n\mathbf{k}}^{\mathrm{KS}} \middle| \Sigma(\mathbf{r}, \mathbf{r}', \omega = \epsilon_{n\mathbf{k}}^{\mathrm{QP}}) - V_{\mathrm{xc}}(\mathbf{r}) \middle| \phi_{n\mathbf{k}}^{\mathrm{KS}} \right\rangle$$

In practice, one assumes small QP corrections and Σ is linearized around the KS energy:

$$\left\langle \Sigma(\omega = \epsilon_{n\mathbf{k}}^{\mathrm{QP}}) \right\rangle = \left\langle \Sigma(\omega = \epsilon_{n\mathbf{k}}^{\mathrm{KS}}) \right\rangle + \left(\epsilon_{n\mathbf{k}}^{\mathrm{QP}} - \epsilon_{n\mathbf{k}}^{\mathrm{KS}} \right) \left\langle \frac{\partial \Sigma(\omega)}{\partial \omega} \bigg|_{\omega = \epsilon_{n\mathbf{k}}^{\mathrm{KS}}} \right\rangle$$

- Defining the renormalization factor $Z_{n\mathbf{k}}$ as:

$$Z_{n\mathbf{k}} = \left(1 - \frac{\partial \left\langle \phi_{n\mathbf{k}}^{\mathrm{KS}} \middle| \Sigma(\omega) \middle| \phi_{n\mathbf{k}}^{\mathrm{KS}} \right\rangle}{\partial \omega} \right)^{-1}$$

the linearized QP energy is given by:

$$\epsilon_{n\mathbf{k}}^{\mathrm{QP}} = \epsilon_{n\mathbf{k}}^{\mathrm{KS}} + Z_{n\mathbf{k}} \left\langle \phi_{n\mathbf{k}}^{\mathrm{KS}} \middle| \Sigma(\mathbf{r}, \mathbf{r}', \omega = \epsilon_{n\mathbf{k}}^{\mathrm{KS}}) - V_{\mathrm{xc}}(\mathbf{r}) \middle| \phi_{n\mathbf{k}}^{\mathrm{KS}} \right\rangle$$

- Z gives the area underneath the QP-peak in the spectral function *(0.7-1.0 if the QP picture is justified)*

Exchange and correlated self-energies

- The GW self-energy in ω-space is given by the convolution integral:

$$\Sigma(\mathbf{r}_1, \mathbf{r}_2; \omega) = \frac{i}{2\pi} \int G(\mathbf{r}_1, \mathbf{r}_2; \omega + \omega') W(\mathbf{r}_1, \mathbf{r}_2; \omega') e^{i\omega'\delta^+} \, \mathrm{d}\omega' .$$

- W is rewritten as bare Coulomb v plus correlated screening $W^c(\omega)$:

$$W(\omega) = v + \ (\varepsilon^{-1} - 1)v = v + W^c(\omega)v$$

Long-ranged, *Short-ranged*

- W^c is short-ranged because $\varepsilon^{-1}_{\mathbf{G}_1\mathbf{G}_2} \approx \delta_{\mathbf{G}_1\mathbf{G}_2}$ for large $\mathbf{G}$

- Σ therefore consists of two terms: exchange *(Fock operator)* and correlation

$$\Sigma(\mathbf{r}_1, \mathbf{r}_2; \omega) \equiv \Sigma_x(\mathbf{r}_1, \mathbf{r}_2) + \Sigma_c(\mathbf{r}_1, \mathbf{r}_2; \omega)$$

Static and Hermitian *non-Hermitian, Frequency-dependent*

Exchange + correlated self-energy

- ABINIT computes $\Sigma_{nm\mathbf{k}}(\omega) = \langle n\mathbf{k} \,|\, \Sigma(\omega) \,|\, m\mathbf{k} \rangle$ for a subset of KS states
- We never compute $\Sigma(\mathbf{r}_1, \mathbf{r}_2; \omega)$
- Diagonal terms $\Sigma_{n\mathbf{k}}(\omega)$ are enough for G_0W_0
- Computing GW band structures is not an easy task!

```
--- !SelfEnergy_ee
kpoint     : [   0.000,    0.000,    0.000, ]
spin       : 1
KS_gap     :    2.544
QP_gap     :    3.110
Delta_QP_KS:    0.567
data: !SigmaeeData |
     Band      E0 <VxcDFT>   SigX SigC(E0)      Z dSigC/dE  Sig(E)    E-E0       E
        2   4.418 -11.332 -13.262   1.352   0.766  -0.305 -11.775  -0.443   3.975
        3   4.418 -11.332 -13.262   1.352   0.766  -0.305 -11.775  -0.443   3.975
        4   4.418 -11.332 -13.262   1.352   0.766  -0.305 -11.775  -0.443   3.975
        5   6.961 -10.028  -5.550  -4.316   0.766  -0.305  -9.904   0.124   7.085
        6   6.961 -10.028  -5.550  -4.316   0.766  -0.305  -9.904   0.124   7.085
        7   6.961 -10.028  -5.550  -4.316   0.766  -0.305  -9.904   0.124   7.085
...
```

*The **k**-point must be in the WFK file*

Two mutually exclusive methods to specify the states

Explicit:

nkptgw: number of **k**-point in $\Sigma_{n\mathbf{k}}$

kptgw: **k**-point list

bdgw: range of n indices for each **k**-point

*Automatic with **qprange** = num*

0 → QP corrections only for the fundamental and the optical gap

+num → QP corrections for all the **k**-points in the IBZ. Include num bands above and below ϵ_F

-num → QP corrections for all the **k**-points in IBZ. Include all occupied states and num empty states.

Exchange part (I)

Coulomb interaction

$$\Sigma_x = iGv \longrightarrow \Sigma_x(\mathbf{r}_1, \mathbf{r}_2) = -\sum_{\mathbf{k}}^{\mathrm{BZ}} \sum_{\nu}^{\mathrm{occ}} \Psi_{n\mathbf{k}}(\mathbf{r}_1)\Psi^{\dagger}_{n\mathbf{k}}(\mathbf{r}_2)\, v(\mathbf{r}_1, \mathbf{r}_2)$$

Fock operator

- The matrix elements of Σ_x in the Bloch basis-set are computed in **G**-space:

Controlled by **nkpt,kptopt,...**

$$\langle b_1\tilde{\mathbf{k}}|\hat{\Sigma}_x|b_1\tilde{\mathbf{k}}\rangle = -\frac{4\pi}{V}\sum_{\nu}^{\mathrm{occ}}\sum_{\mathbf{q}}^{\mathrm{BZ}}\sum_{\mathbf{G}}\frac{|M_{\mathbf{G}}^{b_1 b_1}(\tilde{\mathbf{k}},\mathbf{q})|^2}{|\mathbf{q}+\mathbf{G}|^2}$$

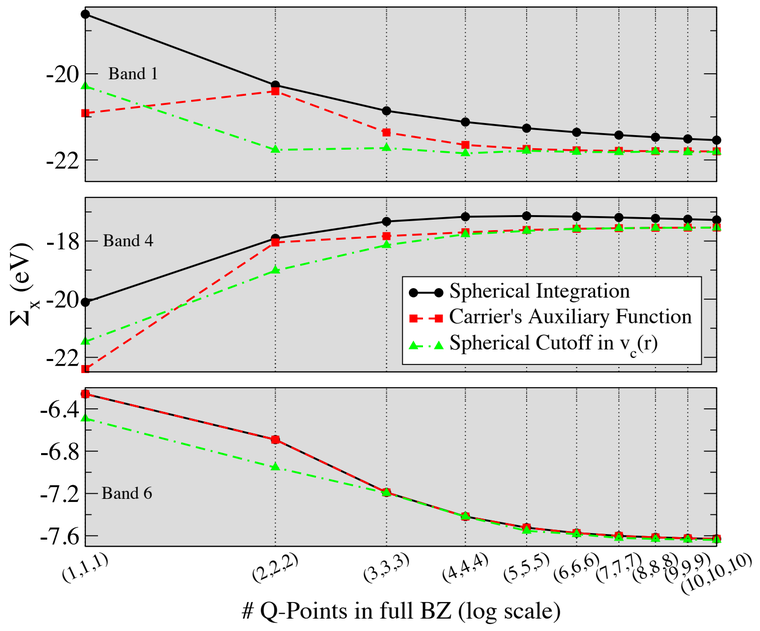
Number of bands by **nband**

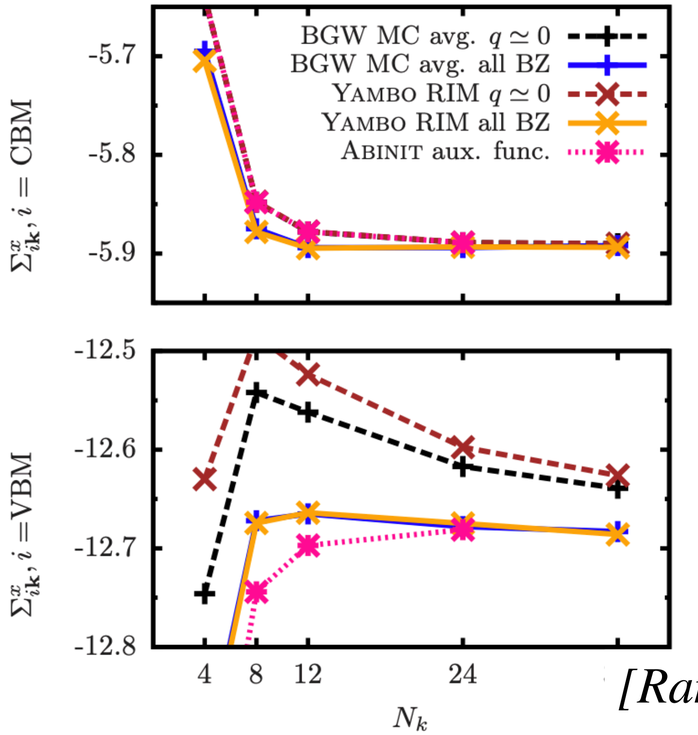
Controlled by **ecutsigx**

- Fast as only occupied states are involved
- Slow converge with respect to **ecutsigx** due to the long-range Coulomb interaction v
- The Coulomb singularity in **q**-space hinders the convergence with the **q**-point sampling ...

gw_icutcoul

- **gw_icutcoul** selects the technique to integrate the Coulomb divergence
- Spherical integration (**gw_icutcoul** 3)
  - approximate the mini-box around Γ with a sphere
  - perform the integration analytically (**gw_icutcoul** 3)
- Default is the auxiliary function by *[Carrier2007]* (**gw_icutcoul** 6)
  - We add and remove a function $f(\mathbf{q})$ with the same asymptotic behaviour as $v(\mathbf{q})$ for $\mathbf{q} \to 0$
  - The integral over the BZ of $f(\mathbf{q})$ has an analytical expression
  - The integrand $\Sigma_x(\mathbf{q}, \ldots) - f(\mathbf{q})$ is smooth and can be integrated with a coarse $\mathbf{q}$-mesh
  - Better than the spherical integration but can be improved…
- Note that the Monte-Carlo (MC) approach (**gw_icutcoul** 16) converges faster than the auxiliary function integration technique





[Rangel2020]

Correlated part

- The matrix elements of $\Sigma_c(\omega)$ are given by:

Controlled by **nkpt,kptopt,...**

Fixed by **ecuteps**

$$\langle b_1\tilde{\mathbf{k}}|\hat{\Sigma}_c(\omega)|b_2\tilde{\mathbf{k}}\rangle = \frac{i}{2\pi V}\sum_{\mathbf{q}}^{\mathrm{BZ}}\sum_{\mathbf{G}_1\mathbf{G}_2}\sum_{n=1}^{+\infty}\left[M^{nb_1}_{\mathbf{G}_1}(\tilde{\mathbf{k}},\mathbf{q})\right]^{\dagger} M^{nb_2}_{\mathbf{G}_2}(\tilde{\mathbf{k}},\mathbf{q})\,\tilde{v}_{\mathbf{G}_1\mathbf{G}_2}(\mathbf{q})\,J^{n\tilde{\mathbf{k}}-\mathbf{q}}_{\mathbf{G}_1\mathbf{G}_2}(\mathbf{q},\omega)$$

Number of bands by **nband**

Frequency dependent term (convolution)

where the expression for J depends on the integration technique (CD, PPM, AC)

- Correlation is short-ranged (can use **ecuteps** << **4 ecut**)
- The most CPU demanding part of the sigma code: double sum over **G** + computation of J
- Slow convergence wrt **nband,** especially for absolute QP energies. <u>Gaps are easier to converge</u>
- $J(\omega)$ involves a complicated numerical integration. Possible approaches:
  1. plasmon-pole models (fast but approximated)
  2. numerical integration techniques (demanding but accurate)

Plasmonpole models

$$\Sigma_c(\mathbf{r}_1, \mathbf{r}_2; \omega) = \frac{i}{2\pi} \int G(\mathbf{r}_1, \mathbf{r}_2; \omega + \omega') \quad \underbrace{W^c(\mathbf{r}_1, \mathbf{r}_2; \omega')}_{\text{Analytic model}} \quad e^{i\omega'\delta^+} \, \mathrm{d}\omega' .$$

Plasmon-pole models

- **Main idea**: approximate the imaginary part of $\varepsilon^{-1}(\omega)$ with a delta-peak (*plasmon resonance*)

$$\mathrm{Im}\,\epsilon^{-1}_{\mathbf{G}_1\mathbf{G}_2}(\mathbf{q},\omega) = A_{\mathbf{G}_1\mathbf{G}_2}(\mathbf{q})\left[\delta(\omega-\omega_{\mathbf{G}_1\mathbf{G}_2}(\mathbf{q})) - \delta(\omega+\omega_{\mathbf{G}_1\mathbf{G}_2}(\mathbf{q}))\right] \xrightarrow{\textit{Kramers-Kronig}} \mathrm{Re}\,\epsilon^{-1}_{\mathbf{G}_1\mathbf{G}_2}(\mathbf{q},\omega) = \delta_{\mathbf{G}_1\mathbf{G}_2} + \frac{\Omega^2_{\mathbf{G}_1\mathbf{G}_2}(\mathbf{q})}{\omega^2 - \tilde{\omega}^2_{\mathbf{G}_1\mathbf{G}_2}(\mathbf{q})}$$

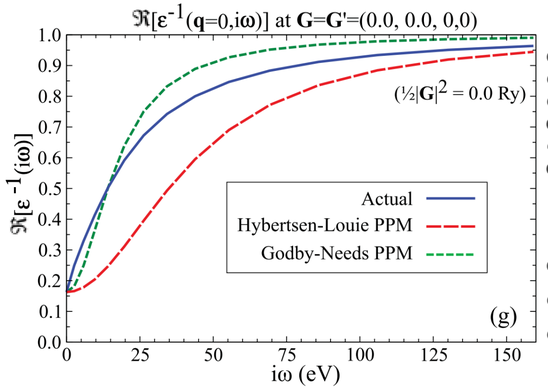
Amplitude of the peak *Plasmon frequency*

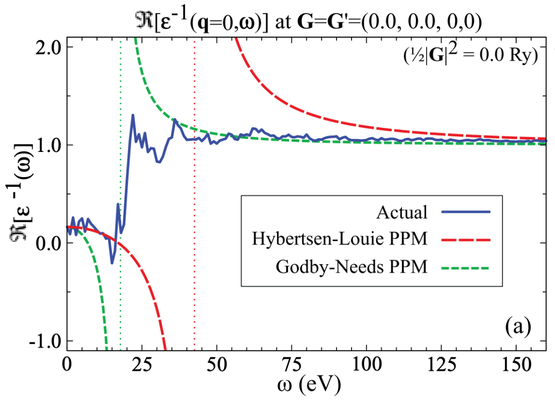
- The two parameters are fitted so to reproduce *ab-initio* results at selected frequencies:

Most commonly used models

ppmodel 1: Reproduce $\varepsilon^{-1}(\omega=0)$ and $\varepsilon^{-1}(i\omega_{plasma})$ (Godby-Needs, **default**)

ppmodel 2: Reproduce $\varepsilon^{-1}(\omega=0)$ and *f*-sum rule. (Hybertsen-Louie)





Other models

ppmodel 3 : Spectral decomposition of ε^{-1} PRB **47** 15931 (1993)

ppmodel 4 : Approximation to the reducible polarizability PRB **34**, 5390 (1986)

Plasmon-pole model: pros and cons

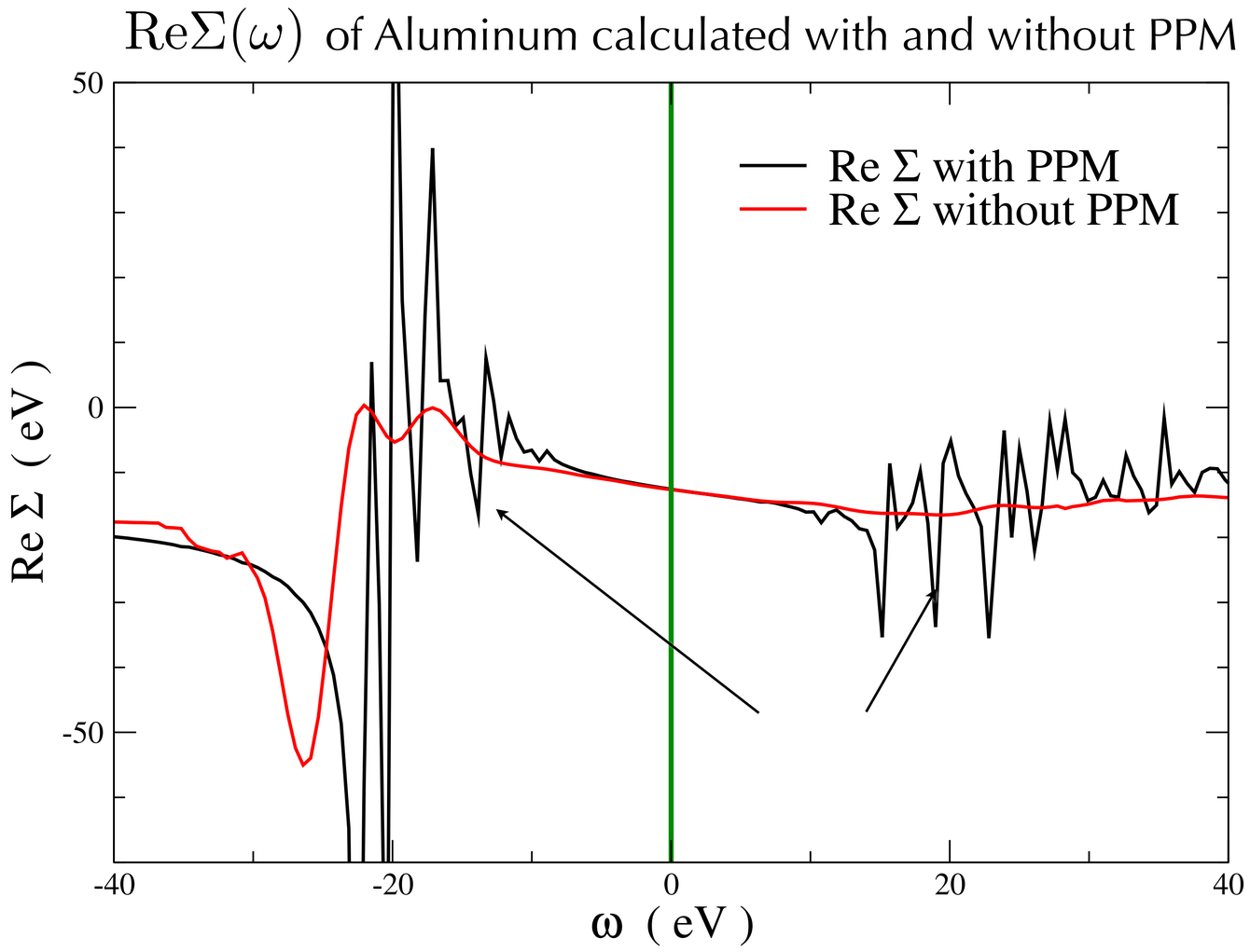
- Accurate QP energies close to the gap (error $\approx 0.1 - 0.2$ eV)
- Ideal tool for initial convergence studies
- Very efficient both in term of CPU and memory as the convolution integral has the analytical expression:

ppmodel=1,2

$$\mathcal{J}^s_{\mathbf{G}_1\mathbf{G}_2}(\mathbf{q},\omega) = \Omega^2_{\mathbf{G}_1\mathbf{G}_2}(\mathbf{q}) \int \frac{e^{i\omega'\delta}\,\mathrm{d}\omega'}{\Big(\omega + \omega' - \epsilon_s + i\eta\,\mathrm{sign}(\epsilon_s - \mu)\Big)\Big(\omega'^2 - (\tilde{\omega}_{\mathbf{G}_1\mathbf{G}_2}(\mathbf{q}) - i\eta)^2\Big)}$$

- Questionable in systems with *d*- or *f*-electron systems as one usually finds multiple peaks in $\mathfrak{I}\varepsilon^{-1}$
- Im $\Sigma(\omega)$ is a sum of delta peaks. No satellites in spectral function or electron lifetimes
- QP corrections for low- and high-energy states are qualitatively wrong. . .

Plasmon-pole model breakdown



WARNING

- Avoid PPMs for computing band widths in metals
- Avoid PPMs for self-consistent *GW* calculations, especially when updating the wavefunctions

Typical input file for Σ with PPM

- Remember: **ecutsigx** <= **4*ecut**
- Need to converge wrt **ecuteps** and **nband.**
- Cannot use more npweps (nband) than what is stored in the input SCR (WFK) files.
- Mind the default values of **gwcalctyp, ppmodel, zcut**

```
getwfk_filepath "prefix_WFK"
getscr_filepath "prefix_SCR"
ecut 8
ecuteps 4
ecutsigx 32
nband 1200

nkptgw 1
kptgw 0 0 0
bdgw 2 4
```

Best practices:

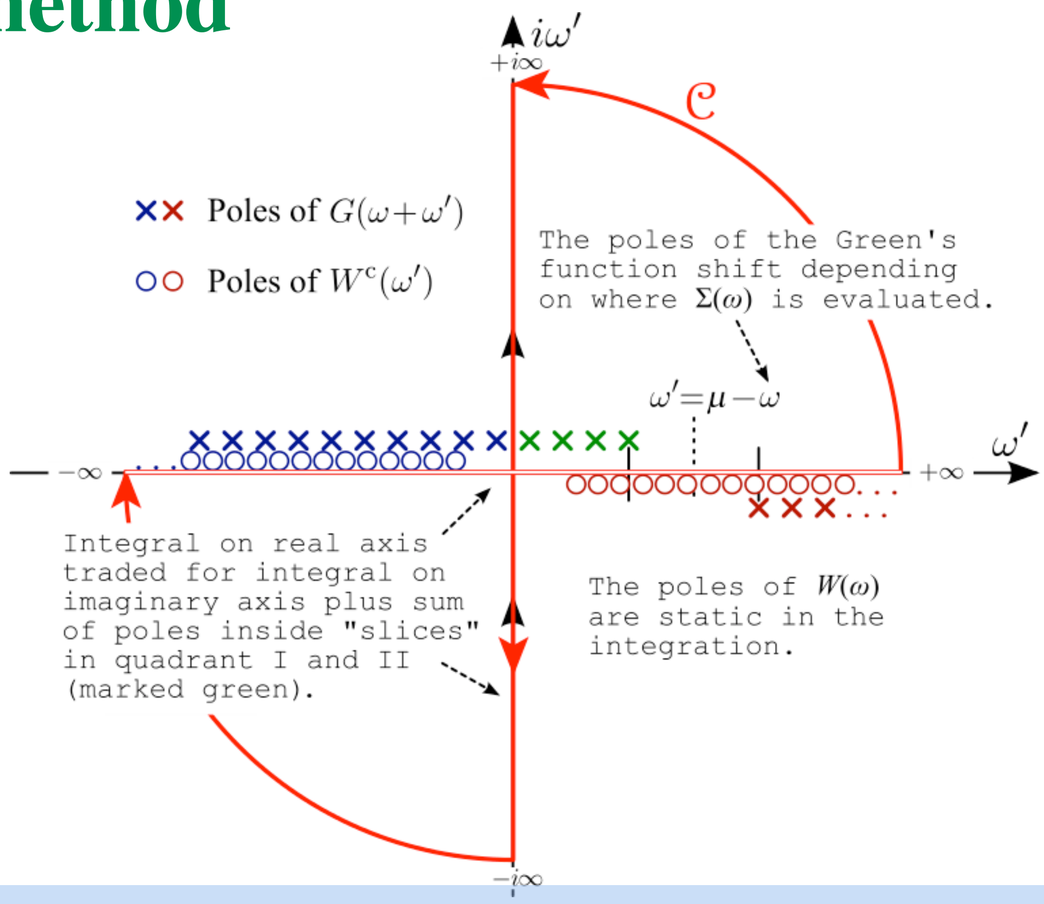
- Converge first wrt **ecutsigx** (fast, you may want to use the max value directly)
- Use number of MPI cores that divides nband for efficiency reasons
- The sigma part is faster than RPA but it requires more memory, especially when one increase ecuteps
- Again, avoid single input with datasets, split the calculation and customize resources

Contour deformation method

Phys. Rev. B **67**, 155208 (2003)

CD integration method

- The numerical integration along the real-axis is challenging because of the poles of G and W (strong oscillations)
- In the CD technique, the integral along the real axis is transformed into an integral over the red contour (Cauchy's residue theorem)
- The integral along the imaginary axis requires less points (smooth integrand as we are far from the poles)
- The sum over the z_p poles of the integrand inside the contour can be evaluated exactly
- More expensive than PPM, especially at the level of memory



$$\Sigma_c(\omega) = \frac{i}{2\pi}\left\{2\pi\, i \sum_{z_p}^{\mathcal{C}} \lim_{z\to z_p} G(z)\, W^c(z)\,(z-z_p) - \int_{-\infty}^{+\infty} G(\omega+i\omega')\, W^c(i\omega')\, \mathrm{d}(i\omega')\right\}$$

Contribution from poles inside $\mathscr{C}$

Integral over imag axis

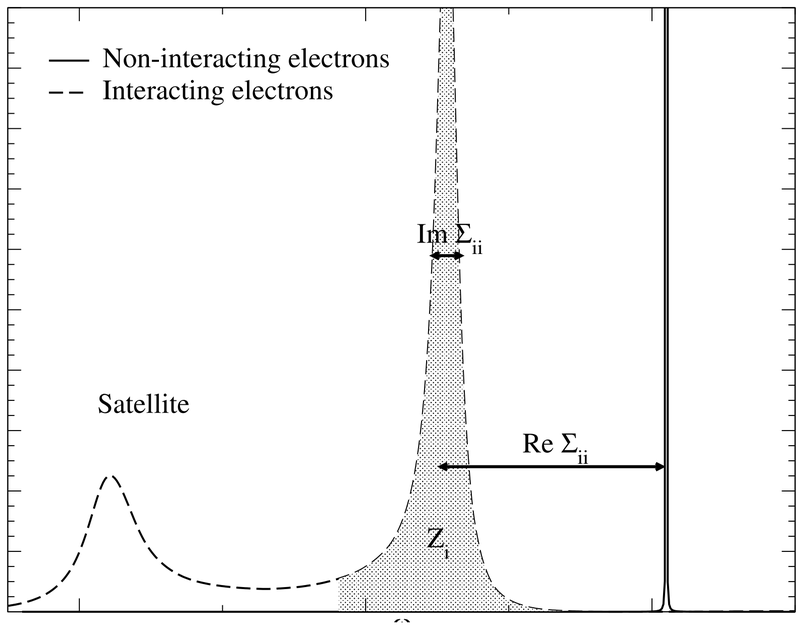
Remember to use gwcalctyp 2 both in screening and sigma

Spectral function with CD

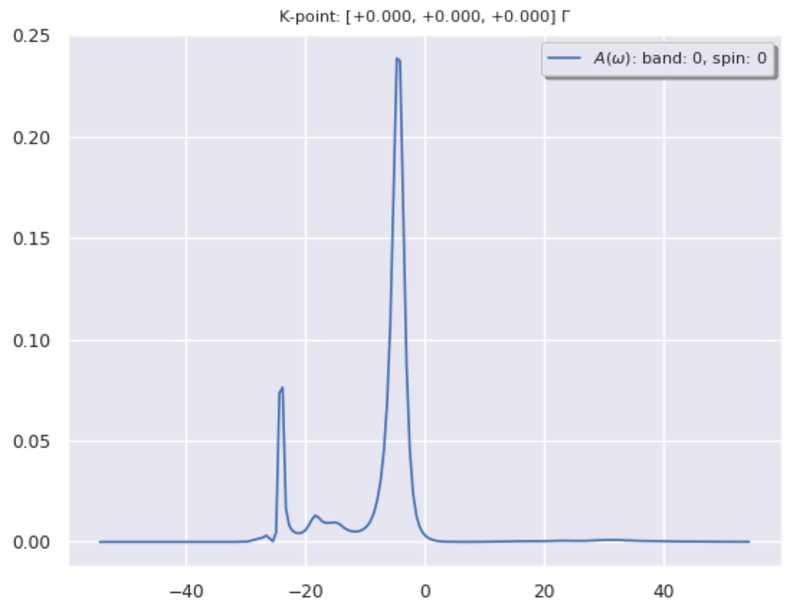
Spectral function

$$A_{n\mathbf{k}}(\omega) = \frac{1}{\pi} |\, \Im G_{n\mathbf{k}}(\omega) \,|$$

$$A_{n\mathbf{k}}(\omega) = \frac{1}{\pi} \frac{|\, \Im \Sigma_{n\mathbf{k}}(\omega) \,|}{(\omega - \varepsilon_{n\mathbf{k}} - \Re \Sigma_{n\mathbf{k}}(\omega))^2 + \Im \Sigma_{n\mathbf{k}}(\omega)^2}$$



- **nfreqsp**: Number of ω-points for the spectral function
- **freqsp**: Maximum real frequency
- Final results saved in the **_SIG** file
- See lesson GW2 (spectral function of Aluminium)



GW with pseudopotentials

Pseudopotentials in a nutshell

- A pseudopotential (PP) mimics the interaction seen by *valence* electrons due to the *core* electrons and the nucleus
- By construction, the PP reproduces the atomic energies and the valence wavefunctions of the all-electron (AE) atom outside a certain radius
- AE valence wavefunctions are replaced by pseudized orbitals that are easier to describe in Fourier space
- Advantages of PPs:
    - Much smaller cutoff energy
    - Less electrons involved in the calculation (frozen core approximation)
- Drawback of PPs:
    - Cannot reproduce the nodal shape of AE orbitals
    - Cannot account for core relaxation effects

- In *GW* codes based on PPs, we only compute the **valence part of the self-energy.** Many-body effects due to core electrons are treated at the KS level and imported from the atomic environment

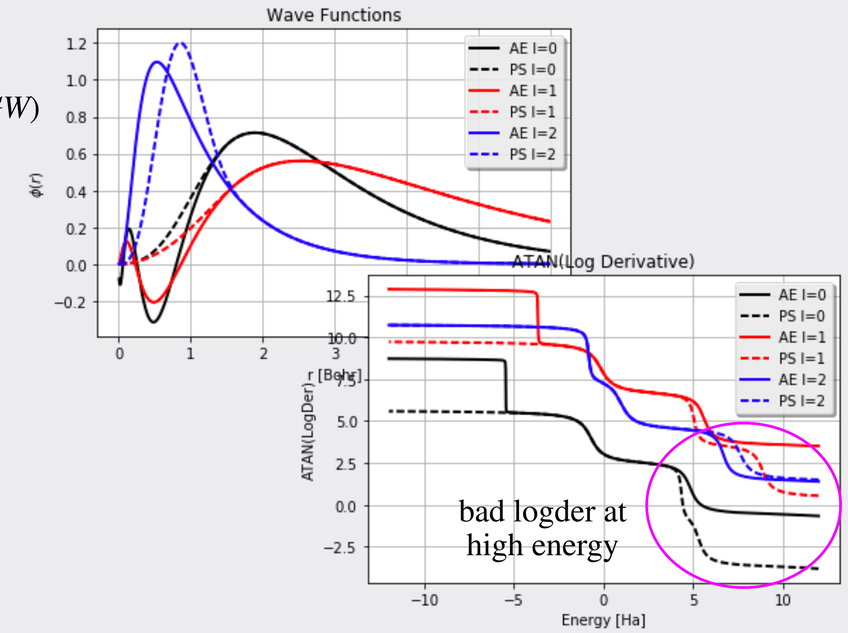
GW with pseudopotentials

Important things to know when using pseudos for *GW*:

- The matrix elements of Σ_X are sensitive to the nodal shape of the orbitals
- Pseudos for *GW* calculations have to be generated carefully:
  - pseudized orbitals should be close to the all-electron ones (not always possible)
  - need multiple projector per *l*-channel to have good scattering properties in the empty region (no ghost)
  - shallow semi-cores should be included, **possibly the full *n*-shell** (*e.g. 3spd* in Ga instead of *3d* only)

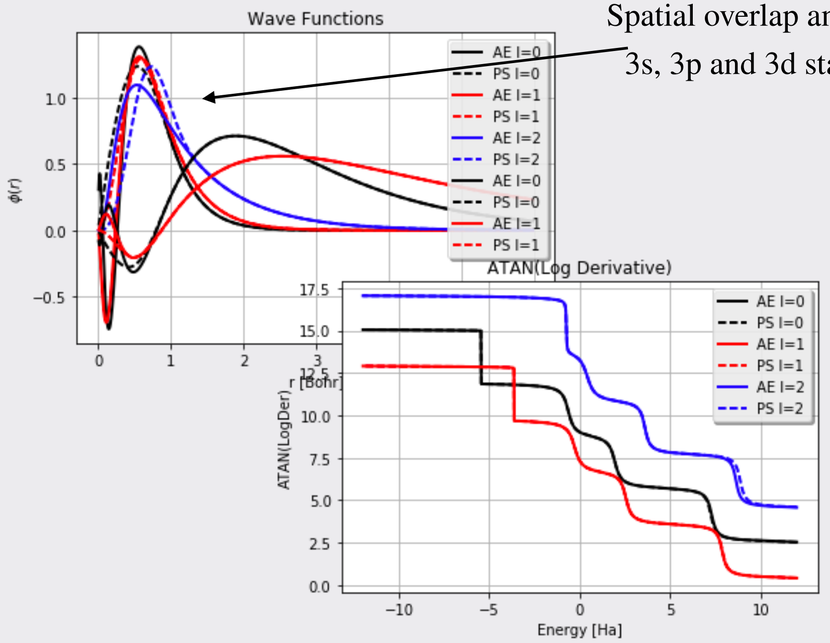
Ga-3d

(OK for GS, bad for *GW*)



Ga-3spd

(OK for *GW*)

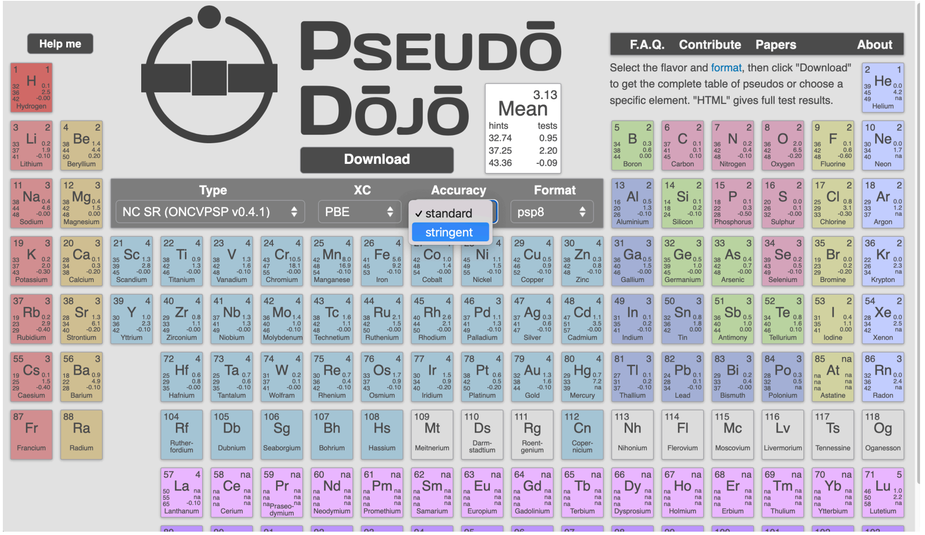


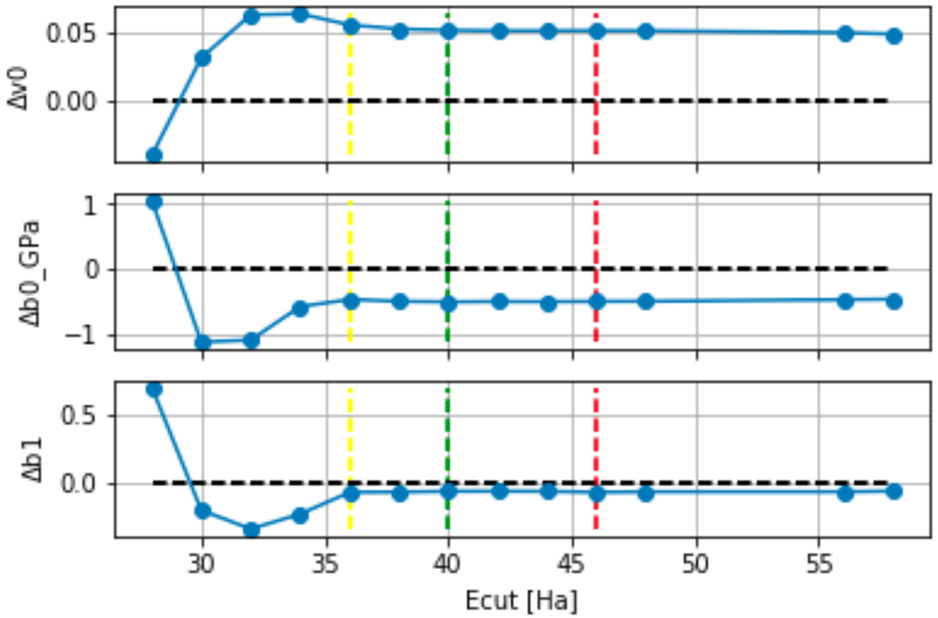
Spatial overlap among
3s, 3p and 3d states

The PseudoDojo project

http://www.pseudo-dojo.org/

- NC and PAW pseudos
- LDA, PBE, PBE-sol
- Scalar and relativistic pseudos with SOC
- Two tables:
  1. standard for GS applications
  2. stringent for *GW* or more accurate GS studies
- Stringent pseudos have:
  - **more electrons** in valence
  - **smaller core radi**
  - **more expensive** in terms of ecut
- For some elements e.g. Si there is no difference between the standard and the stringent version
- All pseudos validated by comparing with AE results
- Hints for ecut are provided *(low/normal/high)*





GW band structures

GW band structures

https://docs.abinit.org/tutorial/gw1/#7-how-to-compute-gw-band-structures

- *GW* corrections can be computed only for the **k**-points in the WFK file (**k**-mesh)
- *GW* band structures (**k**-path) require some sort of interpolation technique
- Three methods available:

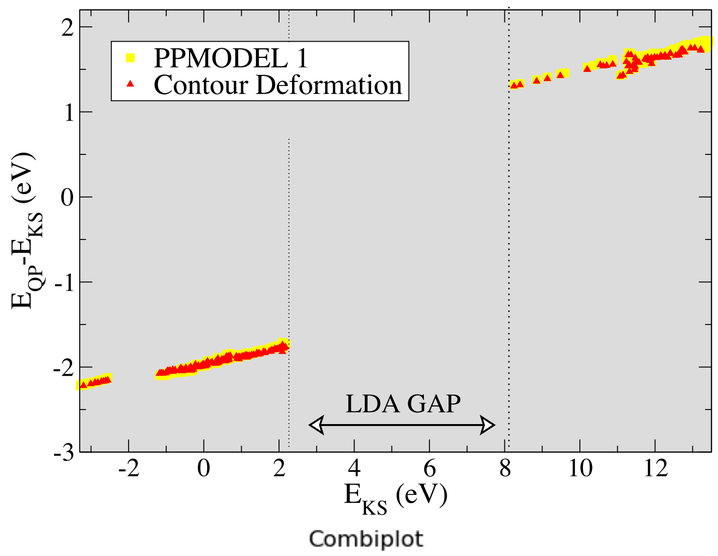
  1. ***Energy-dependent scissors operator***
     - fit QP corrections as a function of the KS eigenvalues: $\epsilon^{QP} = \epsilon^{KS} + \Delta(\epsilon^{KS})$
     - Easy to implement but rather crude, see this **AbiPy example**
  2. ***Wannier interpolation***
     - Accurate but much more complex (requires wannier90 and *maximally-localized* Wannier functions)
     - See tests/wannier90/t03.in and Phys. Rev. B 79, 045109, (2009)
  3. ***Star-function interpolation***
     - Less accurate than Wannier but much easier to use *(same method as in Bolztrap)*
     - Possible instabilities in the presence of band crossings

- In all methods, QP corrections for all **k**-points in the IBZ and all the relevant bands are needed.

Star-function interpolation with AbiPy

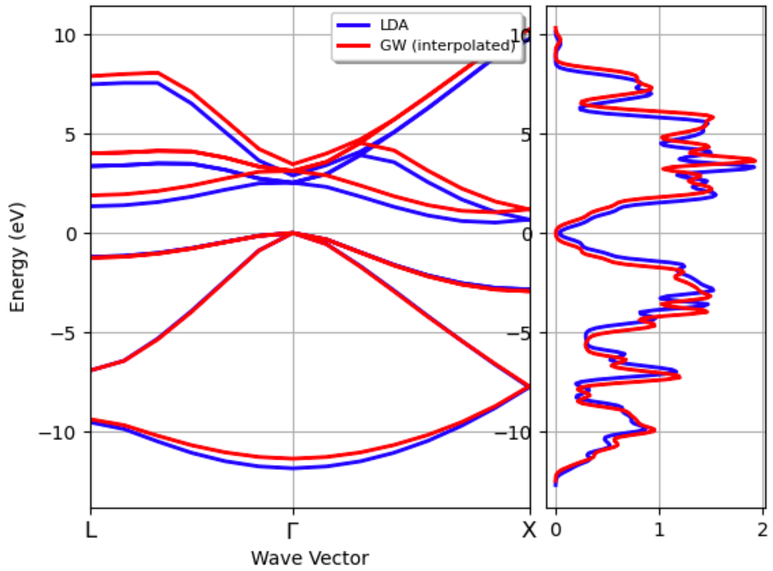
- Good compromise between accuracy and easiness of use
- Can interpolate either QP energies or QP corrections (*recommended*)
- In brief:
  1. Compute KS band structure along a high-symmetry **k**-path
  2. Run *GW* for all **k**-points in the IBZ and the relevant bands
  3. Use SIGRES.nc and GSR.nc to interpolate QP corrections

Corrections are smooth hence easier to interpolate



- Python example:

```
# Get quasiparticle results from the SIGRES.nc database.
sigres = abiopen(abidata.ref_file("si_g0w0ppm_nband30_SIGRES.nc"))

# Read the KS band energies computed on the k-path
with abiopen(abidata.ref_file("si_nscf_GSR.nc")) as gsr_nscf:
    ks_ebands_kpath = gsr_nscf.ebands

# Interpolate the QP corrections.
# The QP energies are returned in r.qp_ebands_kpath
r = sigres.interpolate(lpratio=5, ks_ebands_kpath=ks_ebands_kpath)

r.qp_ebands_kpath.plot(with_gaps=True)
```



Thank you for your attention

Supplemental material

Additional tricks

- Perform the SCF run only once with a converged **k**-mesh and **tolvrs** ~ 1e-8, save the DEN file and use **getden_filepath** to read it
- To speedup the NSCF part, use **getwfk_filepath** to initialise the KS orbitals from a previous WFK should you need to increase **nband** and/or change the BZ sampling
- Hate repeating the structure in each input file over and over again?
  Read it from file with the **structure** variable:

```
structure ”abivars:text_file"
structure “abifile:out_GSR.nc"
structure "poscar:POSCAR"
```

- Most of the wall-time in the *GW* code is spent in the FFT part
- Use optimized BLAS/LAPACK and FFT libs (FFTW3 or MKL-DFTI)
- Life is much easier if intel-MKL is available

https://docs.abinit.org/tutorial/compilation/#how-to-compile-abinit-on-a-cluster-with-the-intel-toolchain-and-modules

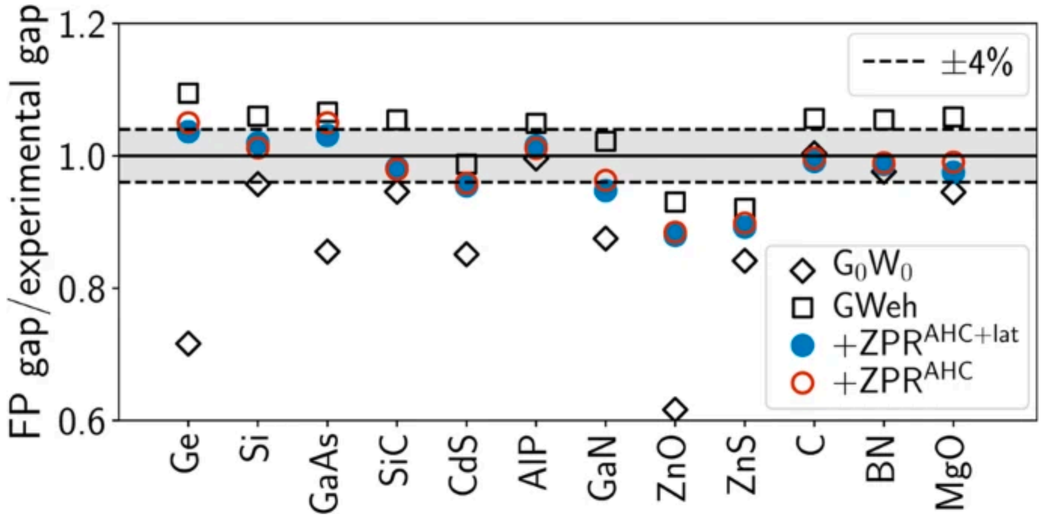
Last but not least: mind the phonons

Electron-phonon self-energy

- All the MBPT equations presented so far assume **clamped** atoms (*no phonons*)
- In a more realistic approach, one should include a $\Sigma_{e-ph}(T)$ self-energy to account for electron-phonon (e-ph) scattering
- $\Sigma_{e-ph}(T)$ leads to temperature-dependent QP energies and band gaps due to phonon population (this is expected according to experiments)
- Even at $\mathrm{T} = 0$, e-ph enters into play due to zero-point renormalization effects
- This contribution is not small, especially in polar semiconductors in which the e-ph interaction diverges as $1/|\mathbf{q}|$ for polar modes (Frohlich interaction)
- Even in non-polar materials such as diamond, there is a significant correction to the QP gap

The quest for the experimental gap

- G_0W_0 underestimates bands gaps
- Including electron-hole interaction via vertex corrections (GW_{eh}) improves over G_0W_0
- GW_{eh} usually overshoots experiments although it is considered the most accurate e-e method available nowadays
- With the exception of problematic systems (ZnO and ZnS), correcting GW_{eh} with e-ph interaction leads to better results



First-principles results without electron–phonon interaction8 are based on G
(diamonds) or on GWeh (squares). Blue full circles are the GWeh results to wh
AHC ZPRg and lattice ZPRlatg have been added, while empty red circles are
GWeh results to which only AHC ZPRg was added. GWeh usually overshoo
while adding the electron–phonon interaction gives better global agreement.
numerical values in Supplementary Table III.

Take-home message:

- On the one hand, we need to go beyond GW by including vertex corrections for the e-e interaction
- On the other hand, we cannot ignore the contribution due to e-ph interaction

Electron-phonon with ABINIT

- We are still far from a unified *ab-initio* treatment of e-e and e-ph interaction
- With Abinit v9.6.2, one can compute, on top of KS ingredients (one-shot method), the following e-ph induced quantities:

1. Temperature dependent QP energies and QP gaps
2. E-ph induced spectral functions (polaron satellites)
3. Phonon-limited mobilities
4. Superconducting properties in metals within the isotropic Eliashberg formalism

https://docs.abinit.org/tutorial/eph_intro/

https://docs.abinit.org/tutorial/eph4zpr/

https://docs.abinit.org/tutorial/eph4mob/

Self-consistency methods

Self-consistency in the *GW* approximation

- If the initial DFT band structure if not adequate, one should update QP energies or wavefunctions in the self-consistent cycle

- Different kind of self-consistency are possible: G_oW, GW_o, GW

- Full self-consistent *GW* calculations do not improve over G_oW_o. Moreover Σ is not hermitian and energy dependent

- The **Quasiparticle Self-consistent GW (QSGW)** overcomes these problems:
-

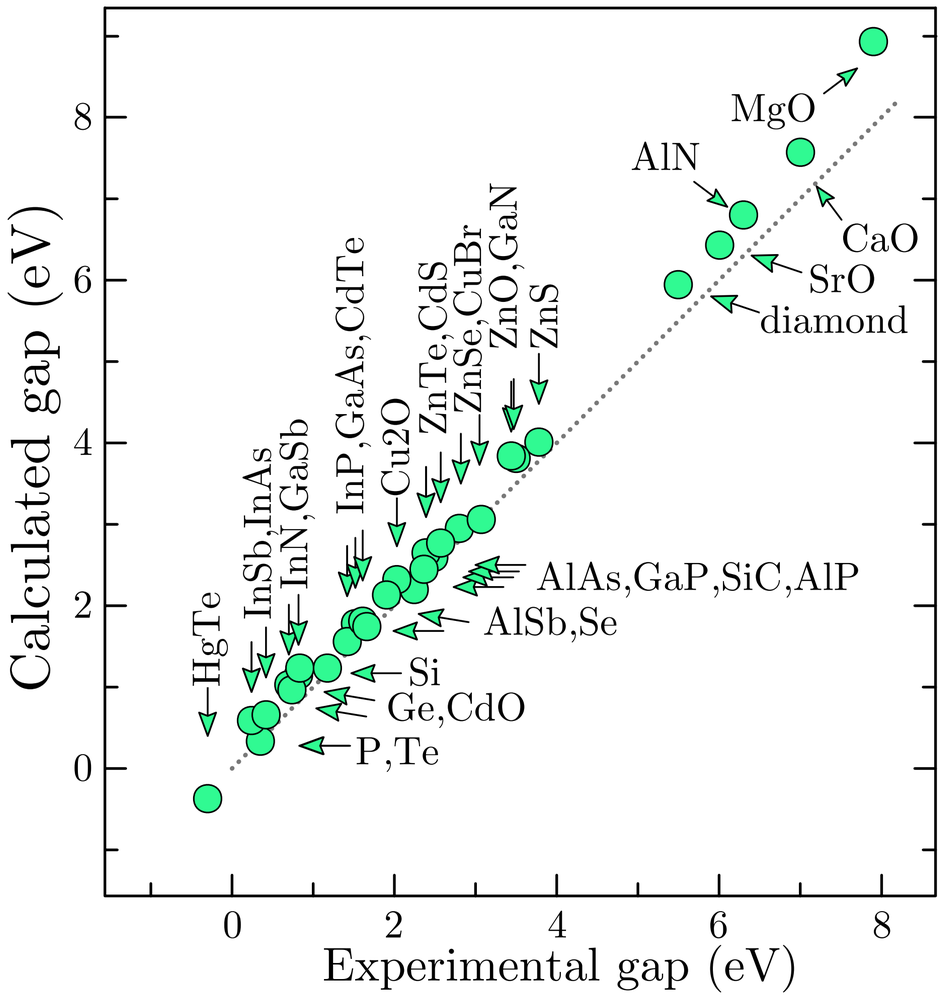
where $\Re$ means that one only retains the hermitian part of the matrix.

$$\langle \phi_i | \Sigma | \phi_j \rangle = \frac{1}{2} \Re \left[\langle \phi_i | \Sigma(\varepsilon_i) | \phi_j \rangle + \langle \phi_i | \Sigma(\varepsilon_j) | \phi_j \rangle \right]$$

S. V. Faleev, M. van Schilfgaarde, and T. Kotani, Phys. Rev. Lett. **93**, 126406 (2004)

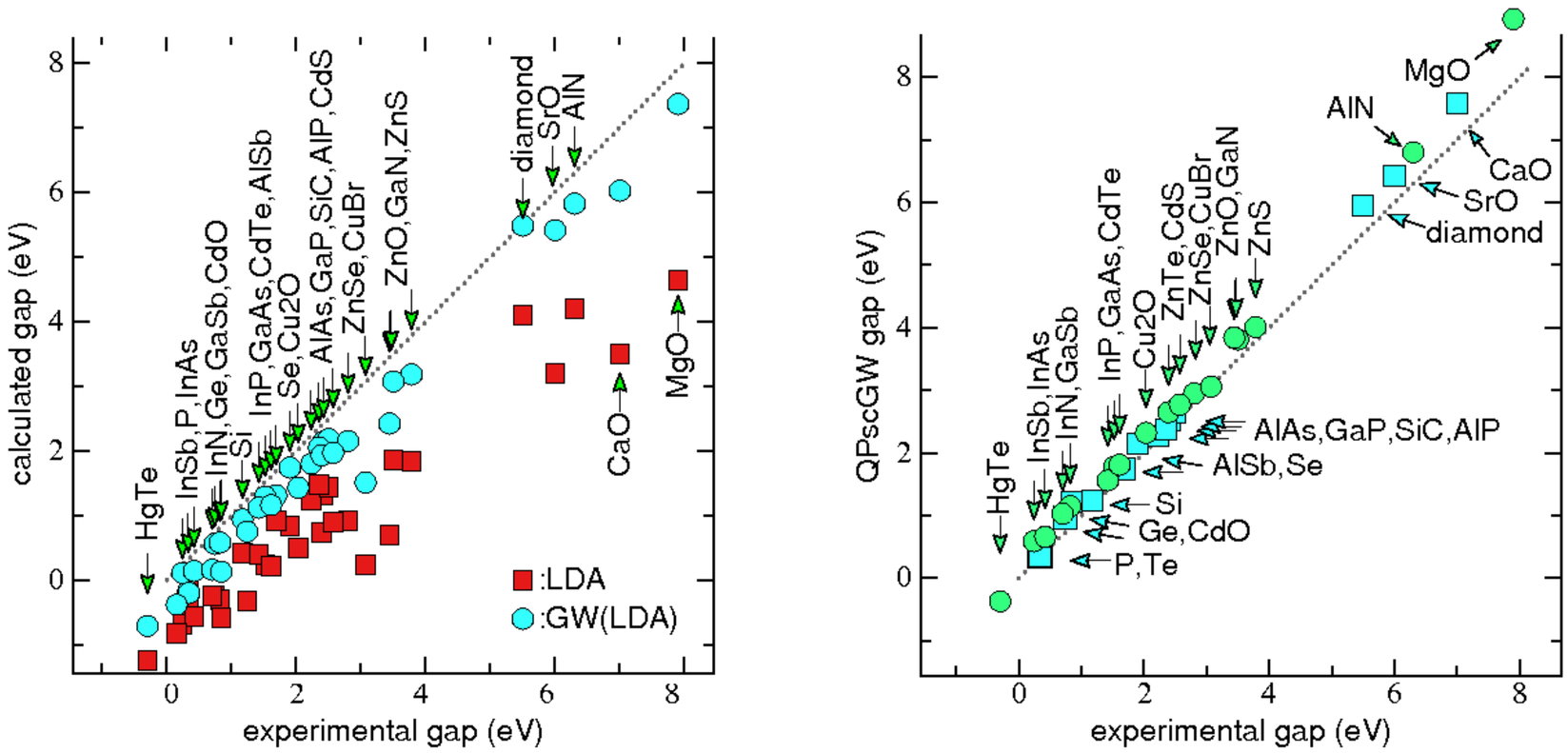
Self-consistency in the *GW* approximation

- The QS*GW* band gap is slightly bigger than the experimental one.



[adapted from van Schilfgaarde et al., PRL **96**, 226402 (2006)]

Quasi-particle SCGW (II)



Left: LDA and G_0W_0 results for the band gap. Right: *QPSGW* results

- Much more CPU demanding than *one-shot GW* due to the off-diagonal terms $\langle i|\Sigma|j\rangle$
- The QP corrections must be calculated for **all k-points** and **all occupied states**
- Check whether the chosen KS basis set is flexible enough

Quasi-particle SCGW (I)

PRL **96** 226402 (2006)

Faleev's approach to the self-consistency issue:

$$\hat{h}_{QPGW} = T + v_H[n](\mathbf{r}_1) + v_{\text{ext}}(\mathbf{r}_1) + \tilde{\Sigma}(\mathbf{r}_1, \mathbf{r}_2)$$

where $\tilde{\Sigma}$is a static and Hermitian approximation to the true GW self-energy

$$\tilde{\Sigma}^{\mathbf{k}}_{ij} \equiv \frac{1}{2}\,\text{Herm}\left[\Sigma^{\mathbf{k}}_{ij}(\epsilon_{i\mathbf{k}}) + \Sigma^{\mathbf{k}}_{ij}(\epsilon_{j\mathbf{k}})\right]$$

Equations are solved self-consistently

$$\begin{tikzcd} \hat{h}_{QPGW} \arrow[r] & \left\{\Psi^{\text{QP}}, E^{\text{QP}}\right\} \arrow[r] & n^{\text{QP}}, \tilde{\Sigma} \arrow[ll, bend left] \end{tikzcd}$$

QP states are expanded in terms of KS states (QPS file)

$$|\Psi^{\text{QP}}_{m\mathbf{k}}\rangle = \sum_n U^{\mathbf{k}}_{nm} |\Psi^{\text{KS}}_{n\mathbf{k}}\rangle$$

- *Important variables: getqps, gwcalctype*

GW self-energy in ω-domain

- In ω-space, the *GW* self-energy is given by the convolution:

$$W_{\mathbf{k}}(\mathbf{G},\mathbf{G}',\omega') = v_{\mathbf{k}}(\mathbf{G},\mathbf{G}')\epsilon^{-1}_{\mathbf{k}}(\mathbf{G},\mathbf{G}',\omega')$$

$$\Sigma(\mathbf{r}_1,\mathbf{r}_2;\omega) = \frac{i}{2\pi}\int G(\mathbf{r}_1,\mathbf{r}_2;\omega+\omega')W(\mathbf{r}_1,\mathbf{r}_2;\omega')e^{i\omega'\delta^+}\,\mathrm{d}\omega'\,.$$

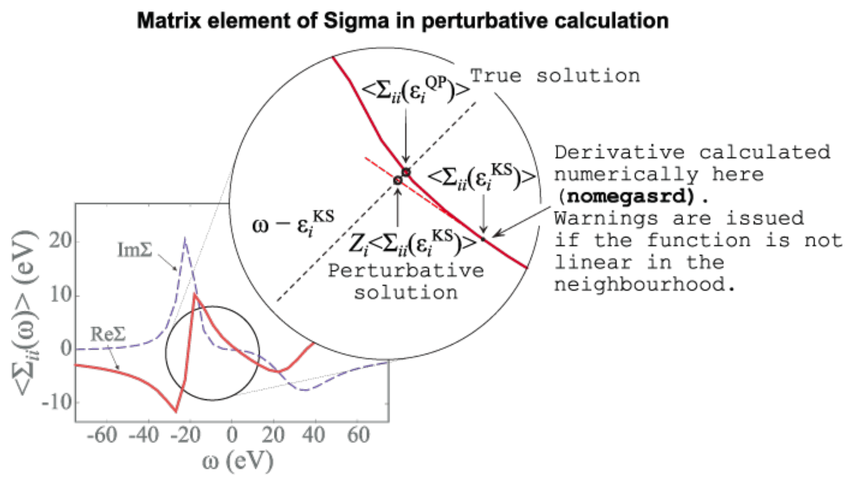
- Using $W = v + (\varepsilon^{-1}-1)v$, we rewrite Σ as exchange (x) + correlation (c):

$$\Sigma(\mathbf{r}_1,\mathbf{r}_2;\omega) \equiv \underbrace{\Sigma_x(\mathbf{r}_1,\mathbf{r}_2)}_{\text{Long range}} + \underbrace{\Sigma_c(\mathbf{r}_1,\mathbf{r}_2;\omega)}_{\text{Short range}}$$

- ABINIT computes $\Sigma^{\mathbf{k}}_{nm}$ in the KS basis. G^0W^0 energies are then obtained via the linearized QP equation:

$$\epsilon^{\mathrm{QP}} = \epsilon^{\mathrm{KS}} + Z\langle\Psi^{\mathrm{KS}}|\Sigma(\epsilon^{\mathrm{KS}}) - v_{\mathrm{XC}}|\Psi^{\mathrm{KS}}\rangle\,.$$

$$Z \equiv \left[1 - \langle\Psi^{\mathrm{KS}}|\frac{\partial\Sigma(\epsilon)}{\partial\epsilon^{\mathrm{KS}}}|\Psi^{\mathrm{KS}}\rangle\right]^{-1}$$



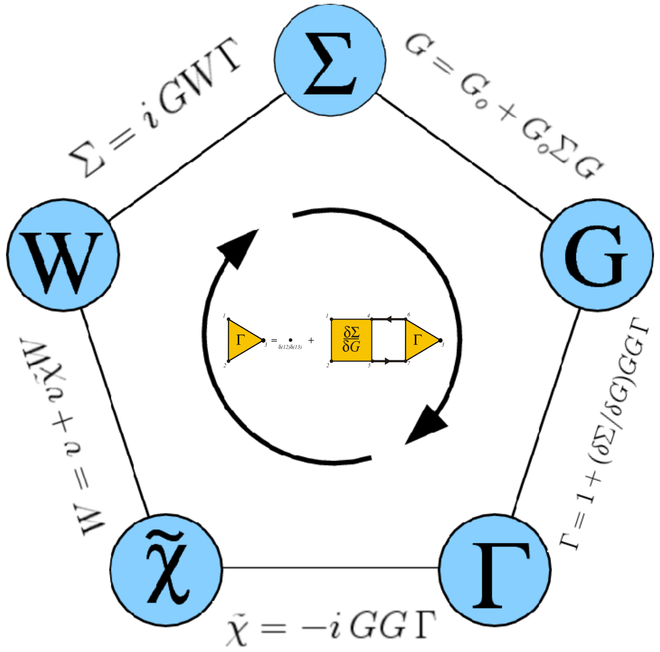
$(1) \equiv (\mathbf{r_1}, t_1)$

$\delta(12) = \delta(\mathbf{r_1} - \mathbf{r_2})\,\delta(t_1 - t_2)$

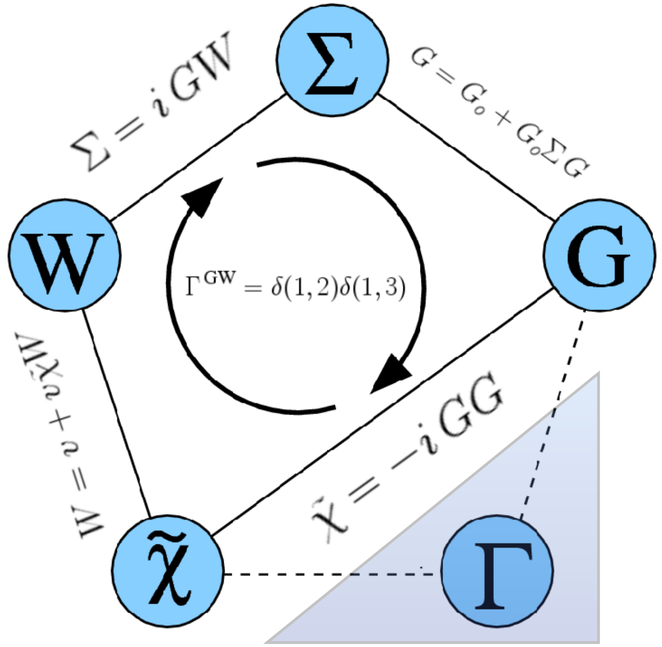
GW equations in r/t space

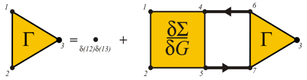
(matrix notation)

Hedin's equations

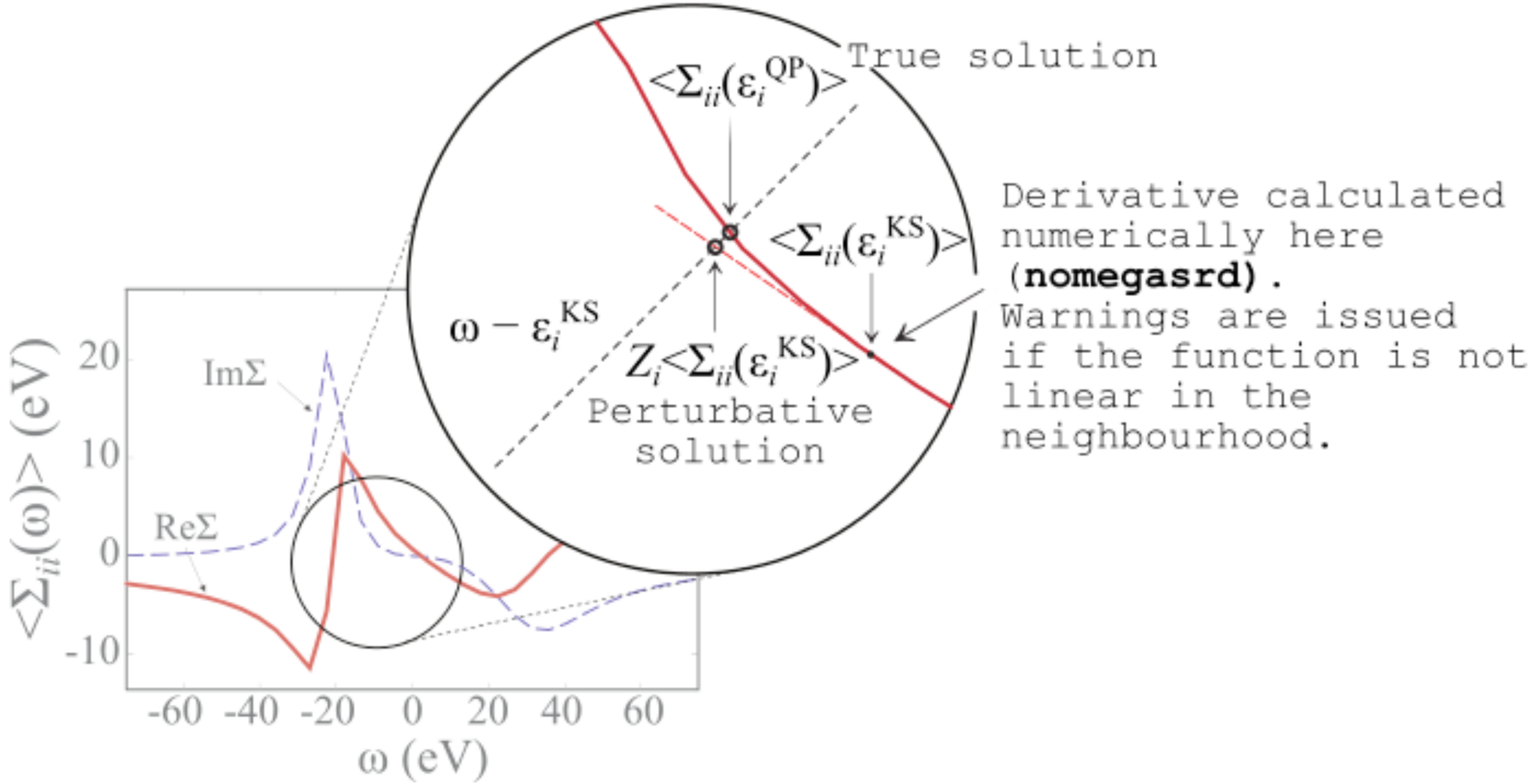


The GW approximation



- The *GW* equations are much "easier" to solve as we completely bypass the vertex equation: 
- Nevertheless, that are several technical aspects that should be considered:
  - representation: **r**-space vs **G**-space, frequency-space vs (imaginary) time, *etc*.
  - basis set expansion, integration techniques, convergence studies
  - self-consistency: one-shot, partial/full *GW* consistency or effective QP Hamiltonian

Matrix element of Sigma in perturbative calculation



- Z gives the area underneath the QP-peak in the spectral function
- Values in the ~[0.6, 1] range indicate a well-defined QP excitation
- Z is evaluated with finite differences:
  - **nomegasrd** defines the number of ω-points to sample Σ(ω)
  - **omegasrdmax** is the maximum distance from the reference KS energy
  - default values are provided

Core-valence partitioning in *GW*

Partition of Σ within the frozen-core approximation

$$G = G_{\text{core}} + G_{\text{valence}} \qquad \tilde{\chi} = \tilde{\chi}_{\text{core}} + \tilde{\chi}_{\text{valence}}$$

$$\Sigma = \underbrace{iG_c\,W + iG_v\,W_v\tilde{\chi}_c\,W_v}_{\text{Approximated with the KS PP}} + \underbrace{iG_v\,W_v}_{\text{Calculated}}$$

- The first two terms are treated at the KS level assuming an atomic environment
- Core relaxation effects are always neglected
- The XC valence-valence contribution is partly subtracted assuming

$$v_{\text{xc}}{}^{\text{core-val}} \approx v_{\text{xc}}[n_{\text{val}} + n_{\text{core}}] - v_{\text{xc}}[n_{\text{val}}]$$

therefore yielding

$$E^{\text{QP}} = \epsilon^{\text{KS}} + \langle\Psi|iG_v\,W_v(E^{\text{QP}}) - v_{\text{xc}}[n_{\text{val}}]|\Psi\rangle$$

Input files for G_0W_0 with CD

TODO

Screening part

- **nfreqre**: Number of real ω
- **nfreqim**: Number of imaginary ω
- **freqremax**: Spacing of the ω-mesh

```
getwfk_filepath "prefix_WFK"
ecut 8
ecuteps 4
nband 1200
gwcalctyp 2  # Important
```

```
getwfk_filepath "prefix_WFK"
getscr_filepath "prefix_SCR"
ecut 8
ecuteps 4
ecutsigx 32
nband 1200
nkptgw 1
kptgw 0 0 0
bdgw 2 4
```

- Need to converge wrt **ecuteps** and **nband**
- It is pointless to have **ecutsigx** bigger than **4*ecut**

Technical details on the RPA implementation

- The screening part represents the main bottleneck of G_0W_0
- The number of **G** vectors (**npweps**) in the matrix depends on **ecuteps** and the unit cell volume
- The cost quickly increases with **nkpt** (let's say nkbz^2 if symmetries are not exploited). Many *GW* calculations for (small) bulk systems are done with ~8x8x8 **k**-grids.
- The size of the SCR file scales as npweps^2 * nqibz * nomega. This array is not MPI-distributed in the sigma part.
- *GW* calculations with large unit cells and/or low-symmetry systems are hard (lot of memory)
- In other words, one should find a reasonably small value of ecuteps that leads to reasonable compromise between accuracy and computational cost.
- By default, ABINIT computes the RPA expression for two frequencies. This is enough for plasmonpole applications but not for more advanced integration techniques (CD or AC)

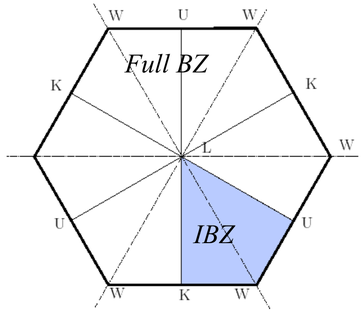
Symmetries in the SIGMA code

- The polarizability is invariant under all the spatial symmetries of the crystal:

$$\chi^0(\mathbf{r}_1, \mathbf{r}_2) = \chi^0\left(\mathscr{R}^{-1}(\mathbf{r}_1 - \mathbf{t}), \mathscr{R}^{-1}(\mathbf{r}_2 - \mathbf{t})\right)$$

- In **G**-space, one obtains the following equation:

$$\chi^0{}_{\mathbf{G}_1\mathbf{G}_2}(\mathscr{R}\mathbf{q}) = e^{i\mathbf{t}\cdot(\mathbf{G}_2-\mathbf{G}_1)}\chi^0{}_{\mathscr{R}^{-1}\mathbf{G}_1\ \mathscr{R}^{-1}\mathbf{G}_2}(\mathbf{q})$$



that allows one to reconstruct the matrix in the BZ from the IBZ

- Additional tricks can be used for the oscillator matrix elements so that we don't need to sum over the BZ

Take-home message:

- $\chi^0(\mathbf{q})$ is computed only for the **q**-points in the IBZ
- For given **q**, the BZ integration can be restricted to the $\mathrm{IBZ}_{\mathbf{q}}$ defined by the *little-group* of **q** (**symchi** 1, default)
- High-symmetry **q**-points require less computing time

Crystal symmetries and k-points (I)

- Seitz notation for the symmetry operations of the crystal:

3×3 real orthogonal matrix	$\{\mathbf{S} \mid \mathbf{v}(S)\}$	vector
	$\{S_{\alpha\beta} \mid v_\alpha(S)\}$	
rotation symrel	translation tnons	

smaller than any primitive vector of the crystal

- Applied to the equilibrium position vector of atom κ relative to the origin of the cell $\boldsymbol{\tau}_\kappa$, this symmetry transforms it as:

$$\{\mathbf{S} \mid \mathbf{v}(S)\}\,\tau_\kappa = \mathbf{S}\tau_\kappa + \mathbf{v}(S) = \tau_{\kappa'} + \mathbf{R}^a$$

where $\mathbf{R}^a$ is a translation vector of the crystal.