# Solid state calculations and DFT errors

Francesca Peccati

ISCD Summer School, July 2018

# Overview

# Bibliography

*Solid state DFT*

- R. Dovesi, B. Civalleri, R. Orlando, C. Roetti, V. R. Saunders Ab initio quantum simulation in solid state chemistry, *Rev. Comput. Chem.* 21, 1-125 (2005)

*DFT error*

- Density functional theory is straying from the path toward the exact functional M. G. Medvedev, I. S. Bushmarinov, J. Sun, J. P. Perdew, K. A. Lyssenko, *Science*, 355, 6320, 49-52 (2017)
- P. Pernot, B. Civalleri, D. Presti, A. Savin Prediction uncertainty of density functional approximations for properties of crystals with cubic symmetry *J. Phys. Chem. A*, 119, 21, 5288-5304 (2015)
- G. N. Simm, J. Proppe, M. Reiher Error assessment of computational models in chemistry *Chimia* 71, 202-208 (2017)

## Contact

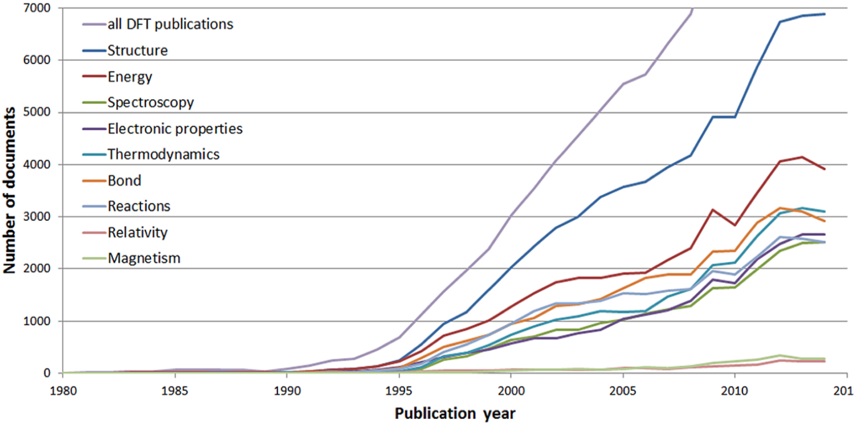
fpeccati@lct.jussieu.fr

# Solid state calculations

# Outline

# Why DFT?

- The increase in speed and computational power of computers, as well as theoretical advances, have made computational chemistry a routine validation tool in chemistry;
- Since the original applications in condensed matter phyisics, modern solid state DFT can calculate a vast range of structural, chemical, optical, spectroscopic, elastic, vibrational and thermodynamic phenomena.

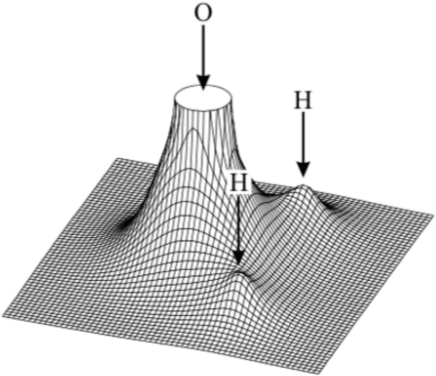


# Hohenberg-Kohn theorems

**Theorem 1**

*The ground state energy is determined completely by the electron density $\rho$. There exists a one-to-one correspondence between the electron density and the energy.*

- the integral of the density defines the number of electrons $N$;
- the cusps in the density define the position of the nuclei;
- the height of the cusps defines the corresponding nucleus.



# Hohenberg-Kohn theorems

$$\rho(\mathbf{r}_1) = \int \cdots \int |\Psi(\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_n)|^2 ds_1 d\mathbf{q}_2 d\mathbf{q}_3 \cdots d\mathbf{q}_n \tag{1}$$

**Theorem 2**
*The ground state energy can be obtained variationally: the density that minimizes the total energy* $E[\rho(\mathbf{r})]$ *is the exact ground state density.*

$$E[\rho'(\mathbf{r})] > E[\rho(\mathbf{r})] \tag{2}$$

# Hohenberg-Kohn theorems

*The ground state energy can be obtained variationally: the density that minimizes the total energy* $E[\rho(\mathbf{r})]$ *is the exact ground state density.*

$$E[\rho(\mathbf{r})] = \int \rho(\mathbf{r}) v_{ext}(\mathbf{r}) d\mathbf{r} + F[\rho(\mathbf{r})] \tag{3}$$

$$\hat{F} = \hat{T} + \hat{V}_{ee}$$

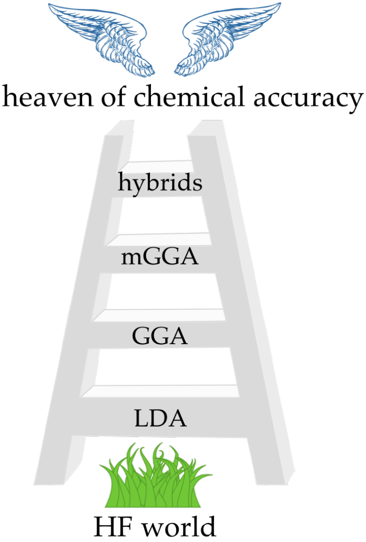
$\hat{V}_{ee}$ contains the exchange-correlation term $\hat{V}_{xc}$.

This implies that the relationship between the exchange-correlation functional and the energy is mediated by the density.

**Kohn-Sham theory (analogous to HF)** Orbitals are reintroduced ($3N$ variables) and electron correlation emerges as a separate term.

$$\rho_{approx} = \sum_{i=1}^{N} |\phi_i|^2 \tag{4}$$

# Jacob's Ladder



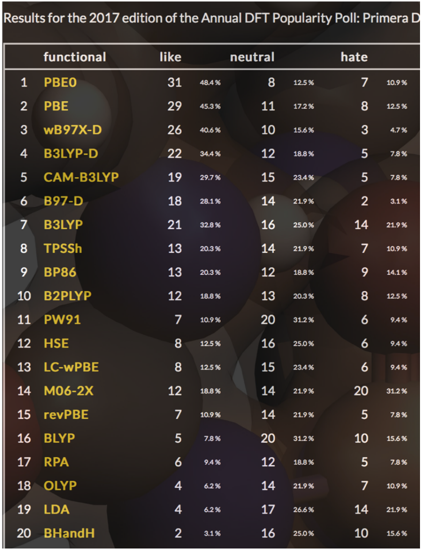
- hybrid $\epsilon_{xc}[\rho, \nabla\rho, \nabla^2\rho, E_x^{HF}]$
- meta-GGA $\epsilon_{xc}[\rho, \nabla\rho, \nabla^2\rho]$
- GGA $\epsilon_{xc}[\rho, \nabla\rho]$
- LDA $\epsilon_{xc}[\rho]$

# DFT popularity poll

The annual popularity poll for density functionals:
**edition 2017**


DFT2017 poll

**organized by:**
marcel swart
f. matthias bickelhaupt
miquel duran


Results for the 2017 edition of the Annual DFT Popularity Poll: Primera D

| | functional | like | | neutral | | hate | |
|---|---|---|---|---|---|---|---|
| 1 | PBE0 | 31 | 48.4 % | 8 | 12.5 % | 7 | 10.9 % |
| 2 | PBE | 29 | 45.3 % | 11 | 17.2 % | 8 | 12.5 % |
| 3 | wB97X-D | 26 | 40.6 % | 10 | 15.6 % | 3 | 4.7 % |
| 4 | B3LYP-D | 22 | 34.4 % | 12 | 18.8 % | 5 | 7.8 % |
| 5 | CAM-B3LYP | 19 | 29.7 % | 15 | 23.4 % | 5 | 7.8 % |
| 6 | B97-D | 18 | 28.1 % | 14 | 21.9 % | 2 | 3.1 % |
| 7 | B3LYP | 21 | 32.8 % | 16 | 25.0 % | 14 | 21.9 % |
| 8 | TPSSh | 13 | 20.3 % | 14 | 21.9 % | 7 | 10.9 % |
| 9 | BP86 | 13 | 20.3 % | 12 | 18.8 % | 9 | 14.1 % |
| 10 | B2PLYP | 12 | 18.8 % | 13 | 20.3 % | 8 | 12.5 % |
| 11 | PW91 | 7 | 10.9 % | 20 | 31.2 % | 6 | 9.4 % |
| 12 | HSE | 8 | 12.5 % | 16 | 25.0 % | 6 | 9.4 % |
| 13 | LC-wPBE | 8 | 12.5 % | 15 | 23.4 % | 6 | 9.4 % |
| 14 | M06-2X | 12 | 18.8 % | 14 | 21.9 % | 20 | 31.2 % |
| 15 | revPBE | 7 | 10.9 % | 14 | 21.9 % | 5 | 7.8 % |
| 16 | BLYP | 5 | 7.8 % | 20 | 31.2 % | 10 | 15.6 % |
| 17 | RPA | 6 | 9.4 % | 12 | 18.8 % | 5 | 7.8 % |
| 18 | OLYP | 4 | 6.2 % | 14 | 21.9 % | 7 | 10.9 % |
| 19 | LDA | 4 | 6.2 % | 17 | 26.6 % | 14 | 21.9 % |
| 20 | BHandH | 2 | 3.1 % | 16 | 25.0 % | 10 | 15.6 % |

# Outline

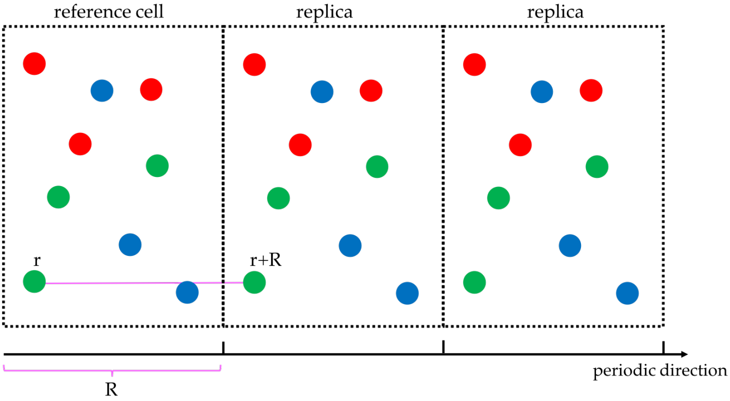
1. Solid state calculations
   - Motivation
   - Introduction to solid state calculations
   - Symmetry
   - The Hamiltonian
   - The basis set

2. DFT errors
   - Classification of DFT methods
   - Motivation
   - Introduction to error analysis
   - Performance estimators
   - A posteriori correction
   - Using HF and LDA to estimate result variability

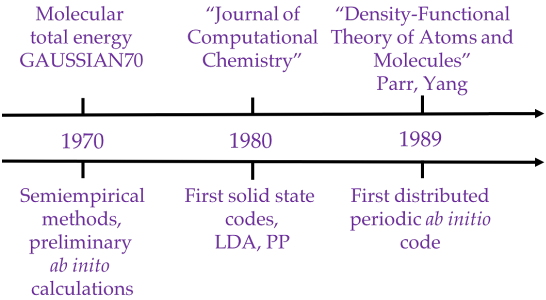
# DFT in solid state

Solid systems are usually periodic.

Periodic DFT calculations dominate the computational side of material science due to their optimal accuracy/cost ration.



# History



- solid state calculations evolved slowly and through solid state physics;
- nowadays, several *ab initio* codes are available, with a large variety of approaches.

# *Ab inito* solid state calculations

### The model

*e.g.* a point defect in a crystalline system can be simulated either with a finite cluster, or with a supercell approach.

### The Hamiltonian

DFT, which functional? LDA is still used. HF? In some cases it is still preferred.

### The basis set?

localized functions or plane waves?

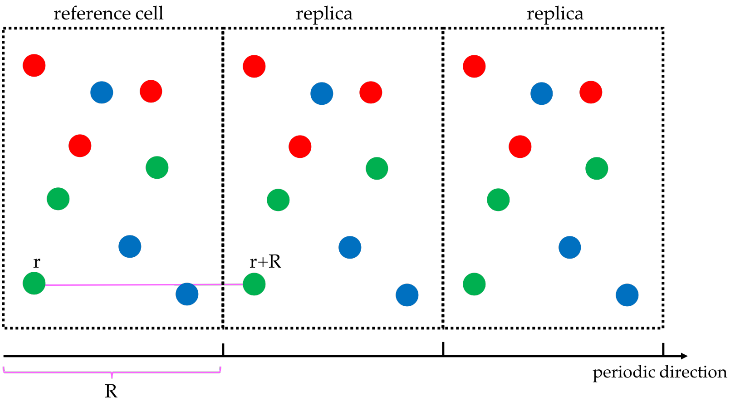
### The computational scheme

*e.g.* direct or reciprocal space representation? All-electron or pseudopotential

# Outline

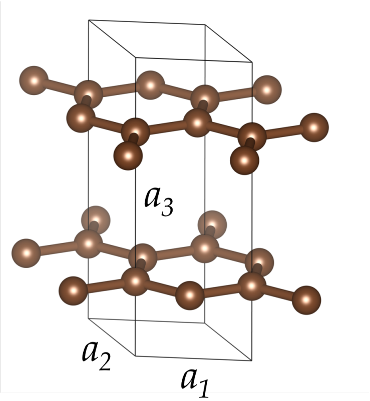
1. Solid state calculations
    - Motivation
    - Introduction to solid state calculations
    - Symmetry
    - The Hamiltonian
    - The basis set
2. DFT errors
    - Classification of DFT methods
    - Motivation
    - Introduction to error analysis
    - Performance estimators
    - A posteriori correction
    - Using HF and LDA to estimate result variability

# Translational invariance



A perfect crystal consists of a three-dimensional array of atoms, ions or molecules, a few of whch form a spatial pattern that is repeated identically throughout the crystal.

# The lattice



- a lattice is a collection of points repeated at intervals of length $a_1$, $a_2$ and $a_3$, along three non-coplanar directions;
- the constants $a_1$, $a_2$ and $a_3$ are the *lattice parameters*;
- the vectors $\mathbf{a}_1$, $\mathbf{a}_2$ and $\mathbf{a}_3$ are the basis vectors;
- lattice parameters and angles between basis vectors define the *cell parameters*;

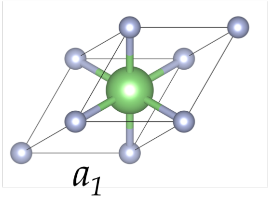
# Lattice vectors

A vector **g** joining two lattice points is a *lattice vector*. every lattice vector can be expressed by the basis vectors and three integer coefficients $n_1$, $n_2$ and $n_3$.

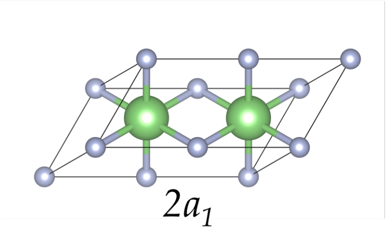
$$\mathbf{g} = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + n_3\mathbf{a}_3 \tag{5}$$

Basis vectors $\mathbf{a}_1$, $\mathbf{a}_2$ and $\mathbf{a}_3$ define the *unit cell*, which can be either *primitive* or not. Applying the lattice vectors **g** to the unit cell, the whole space is filled.

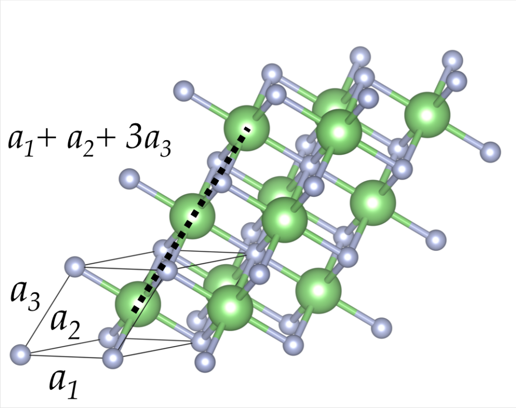
# Types of cell



primitive cell

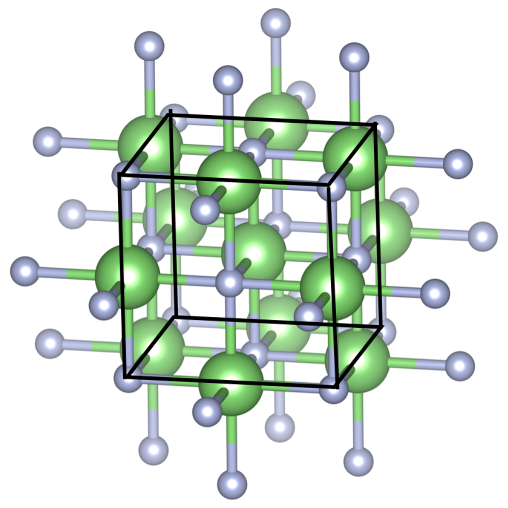


unit cell

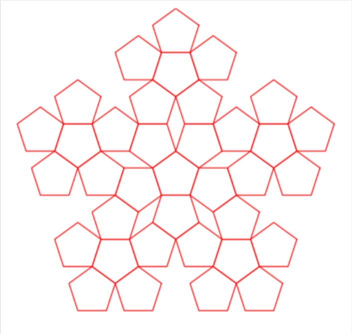


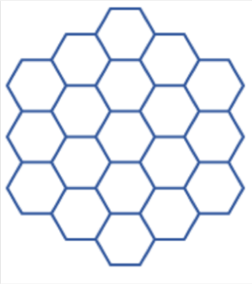
All cells containing the same number of lattice points are equivalent.

# Types of cell



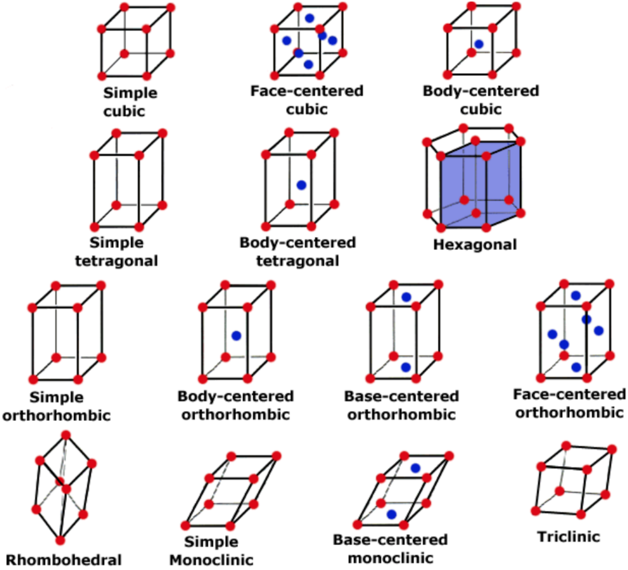
# Space filling



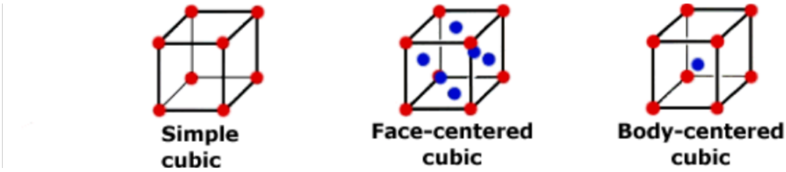


- primitive three dimensional lattices have been classified into seven crystalline systems, triclinic, monoclinic, orthorombic, tetragonal, cuubic. trigonal and hexagonal.
- these, with additional seven non-primitive lattices, consitute the set of all conceivable lattices in ordinary space.

# Bravais lattices



# Space groups



- filling the unit cell of a lattice with matter in a well-defined geometrical arrangement permits the creation of an ideal crystal;

- crystals usually exhibit point symmetry in addition to the set of translations;

- point symmetry and translational symmetry combine to form a space group (230).

# The reciprocal lattice

The reciprocal lattice is a construction of vast importance for condensed matter physics. Starting with a Bravais lattice, the reciprocal lattice is defined as:

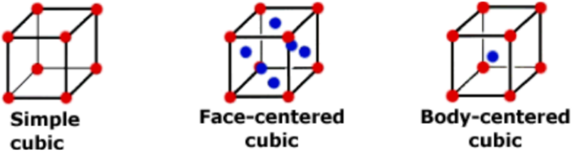
$$\mathbf{b}_i \cdot \mathbf{a}_j = 2\pi\delta_{ij} \tag{6}$$

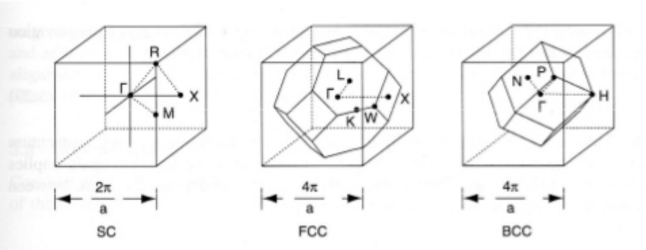
Like in direct space, any reciprocal lattice vector can be expressed as a linear combination of the basis vectors with integer coefficients.

$$\mathbf{K} = K_1\mathbf{b}_1 + K_2\mathbf{b}_2 + K_3\mathbf{b}_3 \tag{7}$$

The *first Brillouin zone* is a primitive cell in the reciprocal space and is important for the description of waves in a periodic medium.

# First Brilloin zone

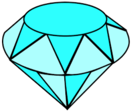




# Outline

# Periodic boundary conditions

Let us imagine a crystal composed of $N$ atoms. The number of atoms at the surface is proportional to $N^{-1/3}$. When $N$ is large, the perturbation deriving form the presence of the boundary is limited only to few surface layers, and has no influence on *bulk properties*.

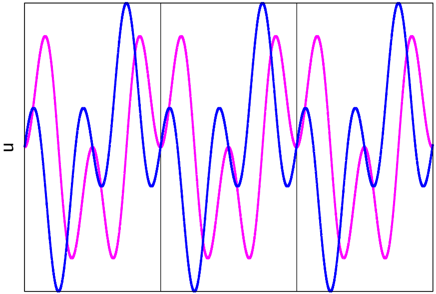


The potential energy of the crystal must be a periodic function with the same perodicity as the lattice, so that for a translation by any direct lattice vector $\mathbf{g}$, the potential energy does not change.

$$V(\mathbf{r}-\mathbf{g}) = V(\mathbf{r}) \tag{8}$$

# Bloch theorem

The *Bloch theorem* imposes conditions on the electron wavefunction of a periodic solid.



Bloch's Theorem (1928, developed to describe the conduction of electrons in solids):

$$\Psi_{n,k}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} u_{n,k}(\mathbf{r}) \tag{9}$$

where $u_{n,k}(\mathbf{r})$ satisifes the condition $u_{n,k}(\mathbf{r}) = u_{n,k}(\mathbf{r}+\mathbf{R})$ and $\mathbf{k}$ is the *crystal wave vector*.

# Periodic boundary conditions

Also the Schrödinger equation for the system must be translation invariant

$$\hat{H}(\mathbf{r})\Psi(\mathbf{r}) = E\Psi(\mathbf{r}) \tag{10}$$

$$\hat{H}(\mathbf{r}-\mathbf{g})\Psi(\mathbf{r}-\mathbf{g}) = E\Psi(\mathbf{r}-\mathbf{g}) \tag{11}$$

The correct eigenfunctions must obey Bloch theorem

$$\Phi(\mathbf{r}+\mathbf{g};\mathbf{k}) = e^{i\mathbf{k}\cdot\mathbf{g}}\Phi(\mathbf{r};\mathbf{k}) \tag{12}$$

This provides a relationship between the values of an eigenfunction at equivalent points of the lattice, which indicates that the periodicity is generally different from that of the lattice.

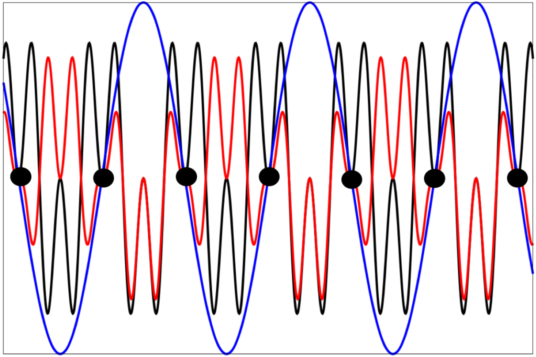
The $j$-th component of the wavevector $\mathbf{k}_j$ can be written as

$$\mathbf{k}_j = \frac{n_j}{N_j}\mathbf{b}_j \tag{13}$$

The wavevector **k** can be interpreted as a point in the reciprocal lattice, which can be written in terms of the reciprocal lattice vectors.

# Bloch function

$$\Psi_{n,k}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} u_{n,k}(\mathbf{r}) \tag{14}$$



$$e^{ix} = cos(x) + i(sinx) \tag{15}$$

# Bloch functions

$$\Phi(\mathbf{r}+\mathbf{g};\mathbf{k}) = e^{i\mathbf{k}\cdot\mathbf{g}}\Phi(\mathbf{r};\mathbf{k})$$

**r** is the position in space and **k** the wave vector. Different **k** parameters label the different solutions to Schrödinger equation.

To proof

$$\hat{H}(\mathbf{r})\Psi(\mathbf{r}) = E\Psi(\mathbf{r})$$

$$\hat{H}(\mathbf{r}-\mathbf{g})\Psi(\mathbf{r}-\mathbf{g}) = E\Psi(\mathbf{r}-\mathbf{g})$$

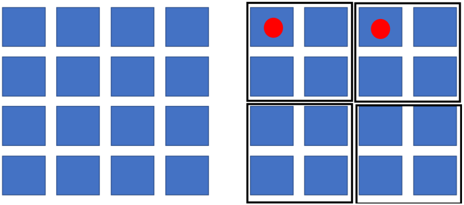
we can write

$$\hat{H}(\mathbf{r}-\mathbf{g})\Phi(\mathbf{r}-\mathbf{g};\mathbf{k}) = E(\mathbf{k})\Phi(\mathbf{r}-\mathbf{g};\mathbf{k}) \tag{16}$$

corresponding to

$$\hat{H}(\mathbf{r})e^{-i\mathbf{k}\cdot\mathbf{g}}\Phi(\mathbf{r};\mathbf{k}) = E(\mathbf{k})e^{-i\mathbf{k}\cdot\mathbf{g}}\Phi(\mathbf{r};\mathbf{k}) \tag{17}$$

# Periodic boundary conditions



Bloch functions span an infinite crystal and not decay to zero at infinity. To circumvent normalization problems that may arise form this behavior, be start by considering a finite crystal of $N$ cells, where $N = N_1 \times N_2 \times N_3$ and then let $N$ grow to infinity. So, we start from a *finite crystal*.

$$\Phi(\mathbf{r} + mN_j\mathbf{a}_j; \mathbf{k}) = \Phi(\mathbf{r}; \mathbf{k}) \tag{18}$$

According to Bloch theorem

$$\Phi(\mathbf{r} + mN_j\mathbf{a}_j; \mathbf{k}) = e^{imN_j\mathbf{k}\cdot\mathbf{a}_j}\Phi(\mathbf{r}; \mathbf{k}) \tag{19}$$

Then

$$e^{imN_j\mathbf{k}\cdot\mathbf{a}_j} = e^{imN_j\mathbf{k}_j\cdot\mathbf{a}_j} = 1 \tag{20}$$

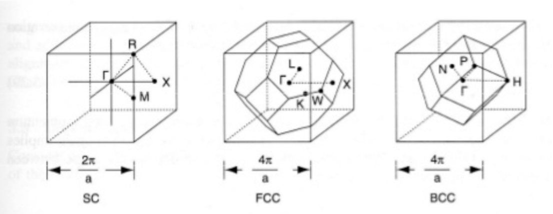
# K points sampling

The $j$-th component of the wavevector $\mathbf{k}_j$ can be written as

$$\mathbf{k}_j = \frac{n_j}{N_j}\mathbf{b}_j \tag{21}$$

The wavevector $\mathbf{k}$ can be interpreted as a point in the reciprocal lattice, which can be written in terms of the reciprocal lattice vectors

$$imN_j\frac{n_j}{N_j}\mathbf{b}_j \cdot \mathbf{a}_j = 0 \tag{22}$$



# K points sampling

With $N$ approaching to infinity, also the number of **k** points approches infinity.

In the basis of the Bloch functions, the Hamiltonian of the periodic system in block diagonal. We can sample the **k** points.

# One-electron Electrostatic Hamiltonian

We sample the Hamiltonian at a finite number of **k** points and solve the Schrödinger equation in the reciprocal space.

$$\hat{H}\Psi_n(\mathbf{r};\mathbf{k}) = E_n(\mathbf{k})\Psi_n(\mathbf{r};\mathbf{k}) \tag{23}$$

$\Psi_n(\mathbf{r};\mathbf{k})$ are the crystalline orbitals

$$\Psi_n(\mathbf{r};\mathbf{k}) = \sum_j c_{jn}(\mathbf{k})\Phi_j(\mathbf{r};\mathbf{k}) \tag{24}$$

where $c_{jn}$ are to be determined.
In the basis of Bloch functions we can write

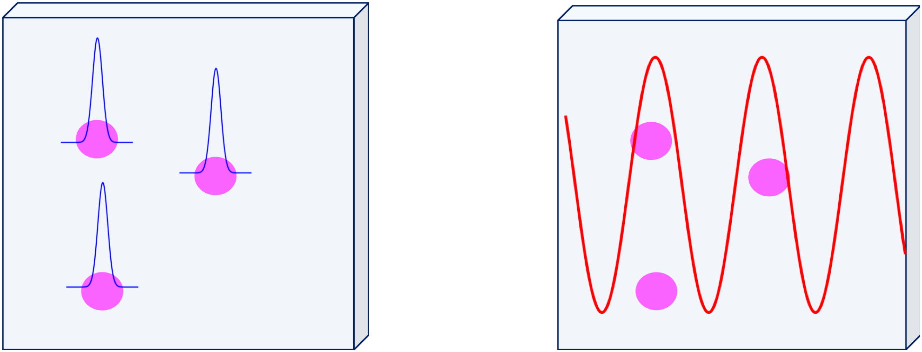
$$H(\mathbf{k})C(\mathbf{k}) = S(\mathbf{k})C(\mathbf{k})E(\mathbf{k}) \tag{25}$$

the size of all matrices is equal to the number of Bloch functions in the basis and $S(\mathbf{k})$ is the overlap matrix (non-orthogonal basis sets).

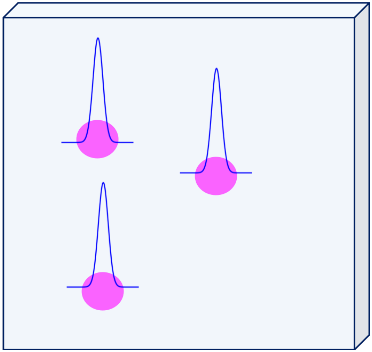
# Outline

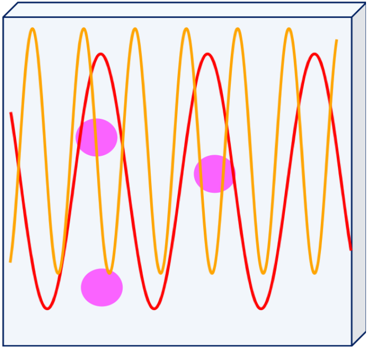
1. Solid state calculations
   - Motivation
   - Introduction to solid state calculations
   - Symmetry
   - The Hamiltonian
   - The basis set
2. DFT errors
   - Classification of DFT methods
   - Motivation
   - Introduction to error analysis
   - Performance estimators
   - A posteriori correction
   - Using HF and LDA to estimate result variability

# Gaussian functions or plane waves?



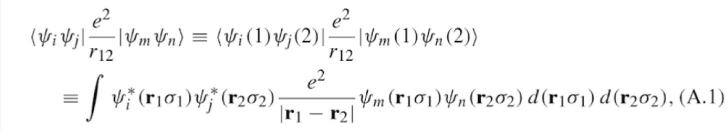
# Gaussian functions or plane waves?





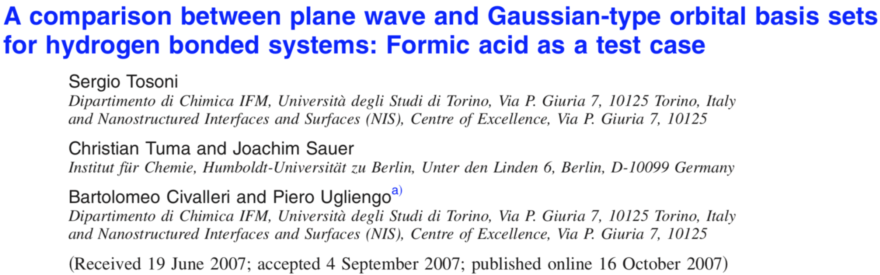
# Pros and cons of Gaussians and plane waves

- plane waves are cheaper but require more memory (huge basis sets);
- plane wave basis sets are defined by a single parameter, Gaussian basis sets are more flexible;
- from a programming point of view, plane waves are simpler to treat;
- with plane waves we do not have basis set superposition error;
- with Gaussian functions we can perform all-electron calculations;
- due to extremely high computational costs when using delocalized basis sets exact (Fock-) exchange is rarely calculated with plane waves.



$$\begin{aligned}\langle \psi_i \psi_j | \frac{e^2}{r_{12}} | \psi_m \psi_n \rangle &\equiv \langle \psi_i(1)\psi_j(2) | \frac{e^2}{r_{12}} | \psi_m(1)\psi_n(2) \rangle \\ &\equiv \int \psi_i^*(\mathbf{r}_1\sigma_1)\psi_j^*(\mathbf{r}_2\sigma_2) \frac{e^2}{|\mathbf{r}_1 - \mathbf{r}_2|} \psi_m(\mathbf{r}_1\sigma_1)\psi_n(\mathbf{r}_2\sigma_2)\, d(\mathbf{r}_1\sigma_1)\, d(\mathbf{r}_2\sigma_2), \quad \text{(A.1)}\end{aligned}$$

# Pros and cons of Gaussians and plane waves



### A comparison between plane wave and Gaussian-type orbital basis sets for hydrogen bonded systems: Formic acid as a test case

Sergio Tosoni
*Dipartimento di Chimica IFM, Università degli Studi di Torino, Via P. Giuria 7, 10125 Torino, Italy and Nanostructured Interfaces and Surfaces (NIS), Centre of Excellence, Via P. Giuria 7, 10125*

Christian Tuma and Joachim Sauer
*Institut für Chemie, Humboldt-Universität zu Berlin, Unter den Linden 6, Berlin, D-10099 Germany*

Bartolomeo Civalleri and Piero Ugliengo[a)]
*Dipartimento di Chimica IFM, Università degli Studi di Torino, Via P. Giuria 7, 10125 Torino, Italy and Nanostructured Interfaces and Surfaces (NIS), Centre of Excellence, Via P. Giuria 7, 10125*

(Received 19 June 2007; accepted 4 September 2007; published online 16 October 2007)

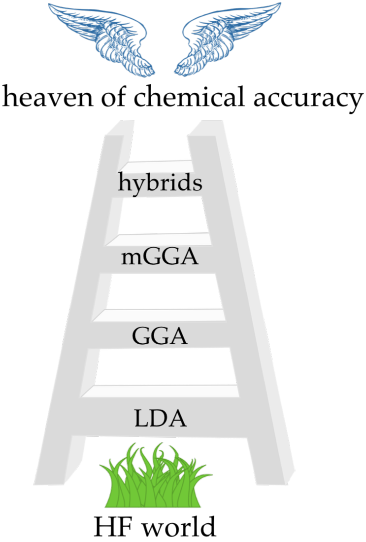
# Summary of solid state DFT

- Solid state calculations are expensive;
- Symmetry is important;
- In 3D we have less choice of functionals than in 0D.

# DFT errors

# Outline

# Jacob's Ladder



- hybrid $\epsilon_{xc}[\rho, \nabla\rho, \nabla^2\rho, E_x^{HF}]$
- meta-GGA $\epsilon_{xc}[\rho, \nabla\rho, \nabla^2\rho]$
- GGA $\epsilon_{xc}[\rho, \nabla\rho]$
- LDA $\epsilon_{xc}[\rho]$

# Hohenberg-Kohn theorems: implications

*The ground state energy can be obtained variationally: the density that minimizes the total energy $E[\rho(\mathbf{r})]$ is the exact ground state density.*

$$E[\rho(\mathbf{r})] = \int \rho(\mathbf{r}) v_{ext}(\mathbf{r}) d\mathbf{r} + F[\rho(\mathbf{r})] \tag{26}$$

$$\hat{F} = \hat{T} + \hat{V}_{ee}$$

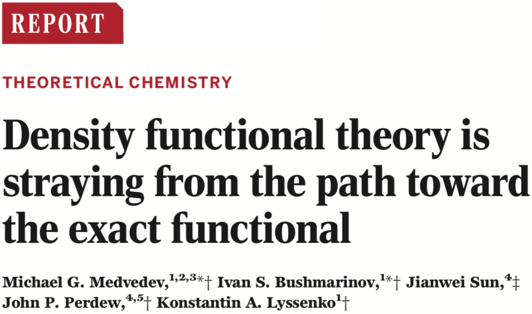
$\hat{V}_{ee}$ contains the exchange-correlation term $\hat{V}_{xc}$.

This implies that the relationship between the exchange-correlation functional and the energy is mediated by the density.

# DFT functional design

When designing a density functional, what is normally done is to tune the formulas so that energy differences for some chosen systems are as close to known targets as possible.

This approach overlooks the fact that the reproduction of exact energy is not a feature of the exact functional, unless the input electron density is exact as well.



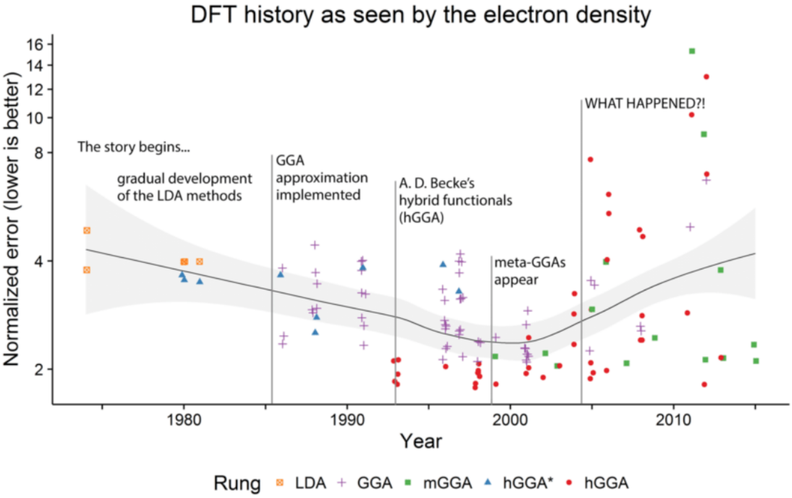
REPORT

THEORETICAL CHEMISTRY

## Density functional theory is straying from the path toward the exact functional

**Michael G. Medvedev,[1,2,3]*† Ivan S. Bushmarinov,[1]*† Jianwei Sun,[4]‡ John P. Perdew,[4,5]† Konstantin A. Lyssenko[1]†**

# How do we evaluate the quality of a DFT functional?



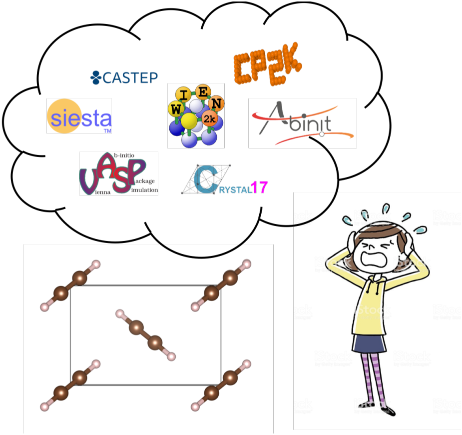
*Science 355, 6320, 49-52*

# Outline

1. Solid state calculations
   - Motivation
   - Introduction to solid state calculations
   - Symmetry
   - The Hamiltonian
   - The basis set
2. DFT errors
   - Classification of DFT methods
   - Motivation
   - Introduction to error analysis
   - Performance estimators
   - A posteriori correction
   - Using HF and LDA to estimate result variability

# Practical aspects

If ranking is a concern for density functional approximation designers to assess the overall performance of new developments, it is less practically useful for end users.



We choose the functional based on *availablity, cost and uncertainty*.

# The virtual measurement

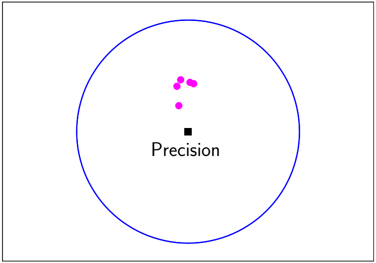
The definition of prediction uncertainty for computational chemistry has been formalized in the *virtual measurement framework*.

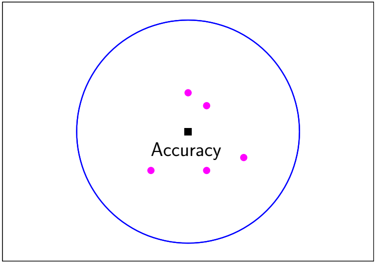
The interest of the virtual measurement framework is to define a *statistical approach* in agreement with international standards for the calculation of measurement uncertainty.

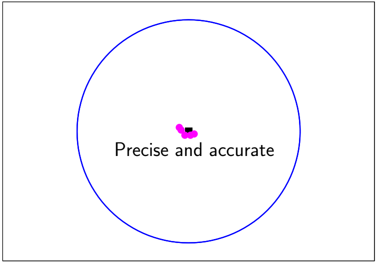
It is currently used in some specific applications: the estimation of the prediction uncertainty for scaled harmonic and anharmonic frequencies. and zero-point energy.

# Outline

# Precision and accuracy



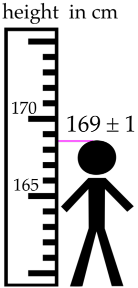




# Uncertainty

## Definition

Uncertainty: non-negative parameter characterizing the dispersion of the quantity values being attributed to a measurand



## What is the relationship between precision and uncertainty?

The $\pm$ is the uncertainty, and it informs the reader of the precision of the value.

# Error sources in calculations

**Numerical errors**

Finite precision, truncation.

**Discretization errors**

Basis set, grid.

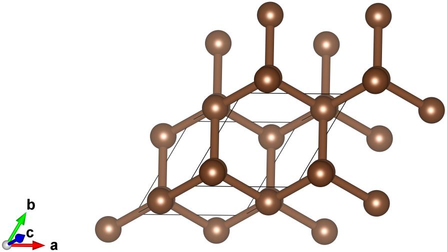
**Parametrization errors**

DFT parameters.

**Approximation errors**

From Born-Oppenheimer onwards.

# How much do these errors affect the final result?



C A4, diamond. Value of the crystallographic cell parameter *a* in Å.

expt: 3.5668 Å

| normal | k | loose tol | param | basis |
|---|---|---|---|---|
| 3.5675 | 3.5704 | 3.5670 | 3.5575 | 3.5866 |

# Outline

1. Solid state calculations
   - Motivation
   - Introduction to solid state calculations
   - Symmetry
   - The Hamiltonian
   - The basis set

2. DFT errors
   - Classification of DFT methods
   - Motivation
   - Introduction to error analysis
   - Performance estimators
   - A posteriori correction
   - Using HF and LDA to estimate result variability

# Definitions

We use a *benchmrak set* as a reference for the calculation accuracy.

We call *error* the difference between the value of a property $c_{m,s}$, calculated for system $s$ with a method $m$, and the corresponding reference value $o_s$.

$$\epsilon = c_{m,s} - o_s \tag{27}$$

$$E_m = \{e_{m,s}; s = 1, N_s\} \tag{28}$$

*mean absolute error*

$$MAE = \frac{1}{N_s}\sum_{s=1}^{N_s} |e_{m,s}| \tag{29}$$

*mean absolute deviation*, a measure of dispersion

$$MAD = \frac{1}{N_s}\sum_{s=1}^{N_s} |e_{m,s} - \bar{E}_m| \tag{30}$$

# Definitions

*mean signed error*, a measure of location

$$MSE = \bar{E}_m = \frac{1}{N_s} \sum_{s=1}^{N_s} e_{m,s} \tag{31}$$

*root-mean-square error*

$$RMSE = \sqrt{\frac{1}{N_s} \sum_{s=1}^{N_s} e_{m,s}^2} \tag{32}$$

*root-mean-square deviation*

$$RMSD = \sqrt{\frac{1}{N_s} \sum_{s=1}^{N_s} (e_{m,s} - \bar{E}_m)^2} \tag{33}$$

*root-mean-square error*

$$RMSE^2 = RMSD^2 + MSE^2 \tag{34}$$

# Performance estimators

$$RMSE^2 = RMSD^2 + MSE^2 \tag{35}$$

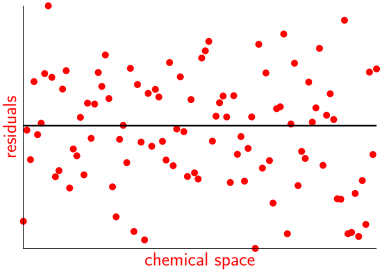
In an error set is affected by a *constant* (system-independent) contribution, then the *MSE* estimates the mean value of the systematic error and the *RMSD* the standard deviation of the remaining (*random*) error.

In the case of negligible contribution of the reference data uncertainty, the *RMSD* and uncertainty on *MSE* could be combined to generate a prediction uncertainty.

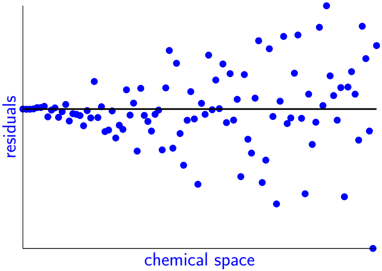
Density functional approximations do not produce only constant systematic error, so besides correcting the systematic error addition corrections are necessary to access the random contribution of the errors.

# Homo- and heteroscedasticity

## Homoscedasticity



## Heteroscedasticity



# Uncertainty estimation

In the accepted approach to uncertainty estimation, *it is assumed that the result of a measurement has been corrected for all recognized significant systematic effects and that every effort has been made to identify such effects.*

This is a key point in computational chemistry, where most errors in the result are systematic as a consequence of the various approximations introduced in the computational model. How do these errors combine?

To estimate uncertainty, we compare computational results with reference data (benchmarks).

We can build a model to assess prediction uncertainty.

- *internal calibration*;
- *a posteriori correction*.

# Outline

1 Solid state calculations
  - Motivation
  - Introduction to solid state calculations
  - Symmetry
  - The Hamiltonian
  - The basis set

2 DFT errors
  - Classification of DFT methods
  - Motivation
  - Introduction to error analysis
  - Performance estimators
  - A posteriori correction
  - Using HF and LDA to estimate result variability

# General scheme

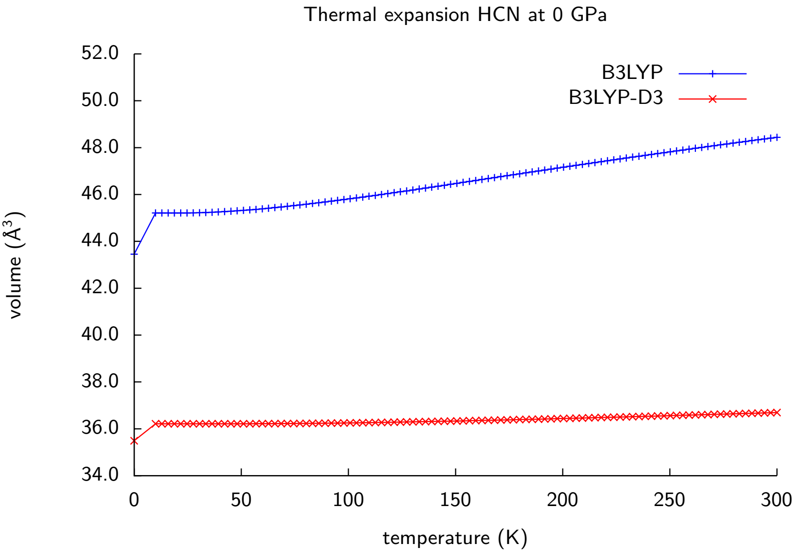
- build and validate a statistical model of the errors from the benchmark set;
- evaluate the uncertainties of the parameters involved in the model;
- propagate the uncertainties of the parameters in the calibration model to the prediction model;
- validate the prediction model.

# Using benchmarks

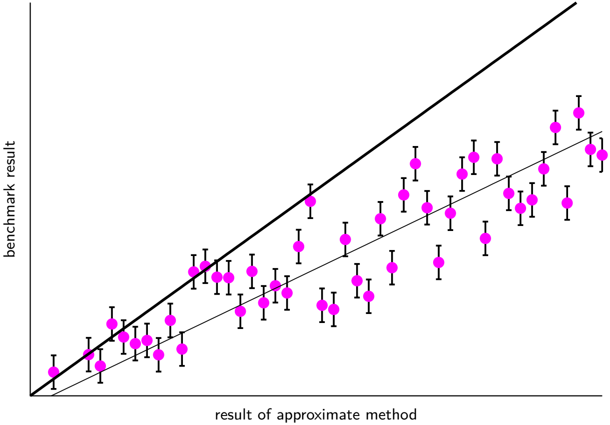
The experimental data commonly used as benchmark is not necessarily the best, exact reference to be used.

- the calculated quantities do not necessarily correspond to experimental data;
- the the theoretical method *is not supposed to provide the quantity analyzed* (*e.g.* bad gaps/Kohn Sham orbital energies);
- experimental values are subject to factors that are not taken into account;
- the inclusion of systems in the benchmark is subject to data availability.

# Thermal expansion



# Error components



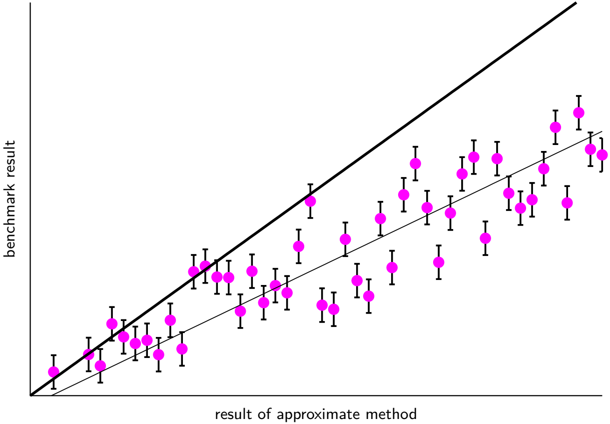
For an adequate approximate model, the data would scatter around $y = x$. Here, we have a constant deviation.

# Systematic and random errors

- *systematic error*: points are grouped around a line that is not the identity line;
- the remaining error is the *random contribution*, which is not the same as the random error in experimental measurements.

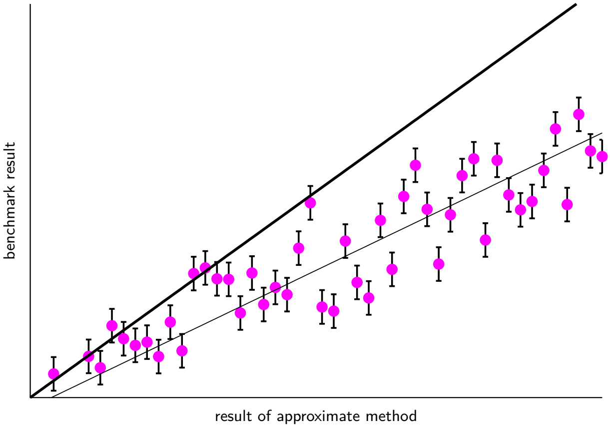
We do not have a truly random processes in the sense of repeated calculations with a model chemistry for the same system, but calculations on different systems.

# *A posteriori* correction



We can fit a linear calibration function through the data, in what is called an *a posteriori* correction. This function connects the benchmark with the calculated data. There still remains a scatter of data around the calibration line.

# Random model inadequacy



The scatter aroud the calibration line seems random, but the residuals are significantly larger than the uncertainty on benchmark results. This is a symptom of *random model inadequacy*, and implies that the uncertainty of the approximate methods exceeds the uncertainty of the benchmark

# Random error

We will consider a linear model

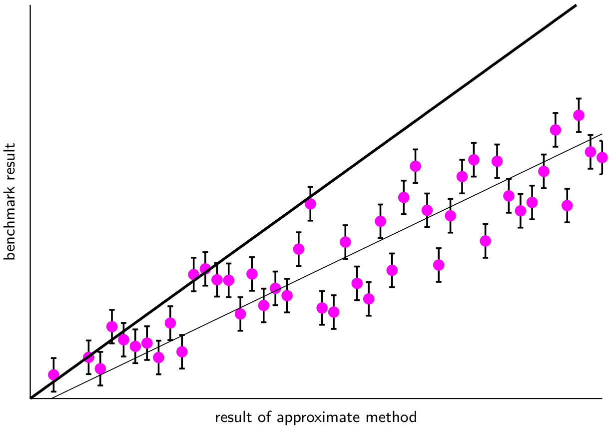
$$o_s = c_{m,s} + \epsilon_s (s = 1, N_s) \tag{36}$$

where $\epsilon_s$ are independent random variables of mean 0 and known standard deviation $u_s$. $\epsilon_s$ are random variables whose realizations are the errors $e_{m,s}$.

In most cases this model is invalid in the sense that values calculated by the DFAs are not compatible with the reference data within their uncertainty range.

To get a valid calibration model we have to account for the structure of the error set. We transform the calculated values according to the calibration line.

# Random model inadequacy



# Transformation

$$o_s = f_m(c_{m,s}; \theta_m) + \epsilon_s \tag{37}$$

$\theta_m$ os the set of parameters defining $f_m$. Here it is important not to overfit data, *i.e.* we will always find a high degree polynomial fitting all the errors in the set, but in that case we will not be able to generalize.

After calibration we look at the residuals

$$r_{m,s} = o_s - f_m(c_m; \theta_m) \tag{38}$$

And compare them to the reference data uncertainties

$$\chi^2 = \sum_{s=1}^{N_s} (\frac{r_{m,s}}{u_s})^2 \tag{39}$$

# Validation

We compare $\chi^2$ with the number of degrees of freedom $n_{df} = N_s - N_\theta$, where $N_\theta$ is the number of parameters in $f_m$.

If the residual errors present a random-like pattern, we can introduce a new stochastic term $\delta_m$, to describe the errors in excess of reference data uncertainty, which we refer to method inadequacy.

$$o_s = o_s = f_m(c_{m,s};\theta_m) + \epsilon_s + \delta_m \tag{40}$$

$\delta_m$ is a random variable od mean 0 and unknown standard deviation $d_m$.

$d_m^2$ can chosen as the difference between the variance of the residual errors and and the mean variance of reference data. Whith this choice, the corrected calculated values and reference data are compatible through the combination of their respective error bars.

# Prediction

Estimation of a new value of a property given the calculated result $c*$ (out of benchmark).

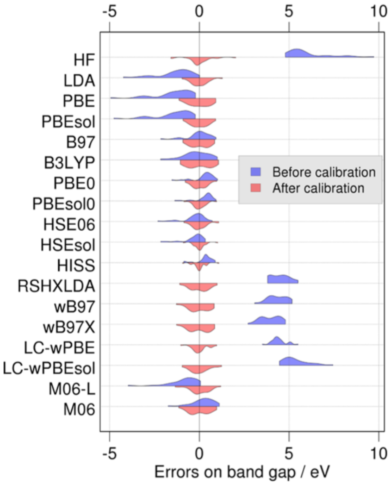
$$p_m(c*) = f_m(c*; \hat{\theta}) + \hat{\delta}_m \tag{41}$$

$\hat{\theta}$ is the set of optimal parameters.

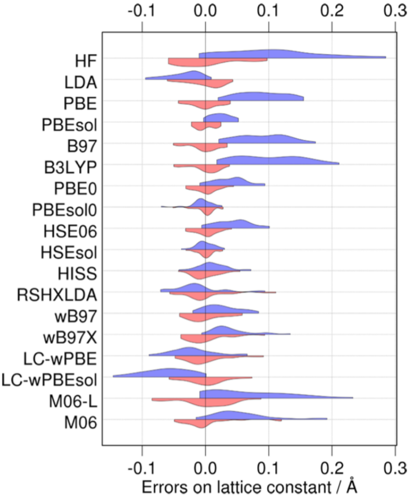
$$u^2_{p_m}(c*) = u^2_{f_m}(c*; \hat{\theta}) + d^2_m \tag{42}$$

$u^2_{f_m}(c*; \hat{\theta})$ is the parametric uncertainty on the value of the function $f_m$ t $c*$.

# Applications - band gap



## Applications - cell parameters

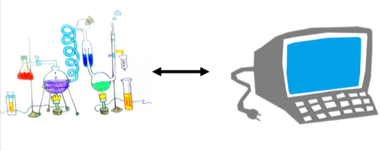


# Summary of *a posteriori* calibration

- *A posteriori* correction is a complex but accurate method to evaluate uncertainty on calculated data;
- To be statistically significant, benchmark sets must be large;
- The treatment depends a lot on the quality of reference data.

# Outline

# Introduction



experimental cell parameters

```
Sample: msl416031, T = 25 C,
P = 0.0 kbar
_database_code_amcsd 0003397
_chemical_formula_sum 'Na Cl'
_cell_length_a 5.6401
_cell_length_b 5.6401
_cell_length_c 5.6401
```

calculated cell parameters

```
LDA 5.40796345
PBE 5.64017502
BLYP 5.69647294
B3LYP 5.64936094
M062X 5.43723219
PBE0 5.60182968
HF 5.71159899
```

Tuning simulation parameters we can match perfectly experimental values...

... but should we?

# Error sources in calculations

**Numerical errors**

Finite precision, truncation.

**Discretization errors**

Basis set, grid.

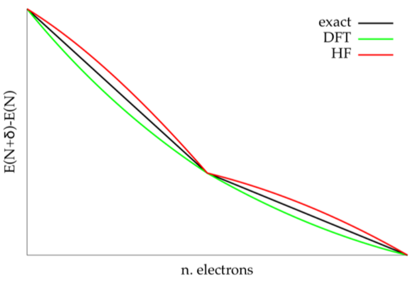
**Parametrization errors**

DFT parameters.

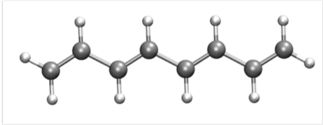
**Approximation errors**

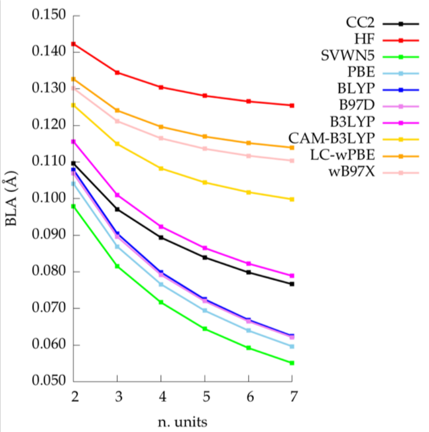
From Born-Oppenheimer onwards.

# The delocalization error

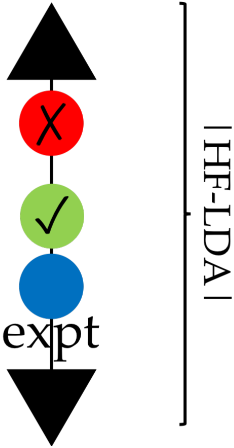


- positive curvature: unphysically strong electron delocalization (standard DFT functionals);
- negative curvature: too strong localization, (hybrids with large fractions of HF exchange and HF).

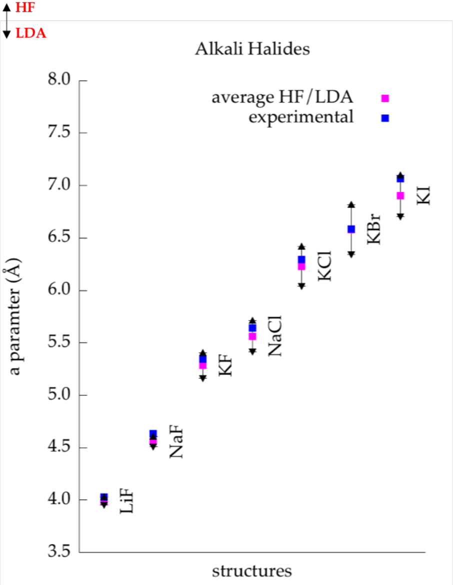


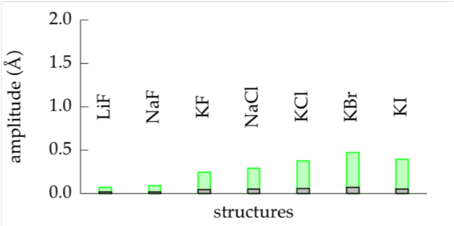


# The idea

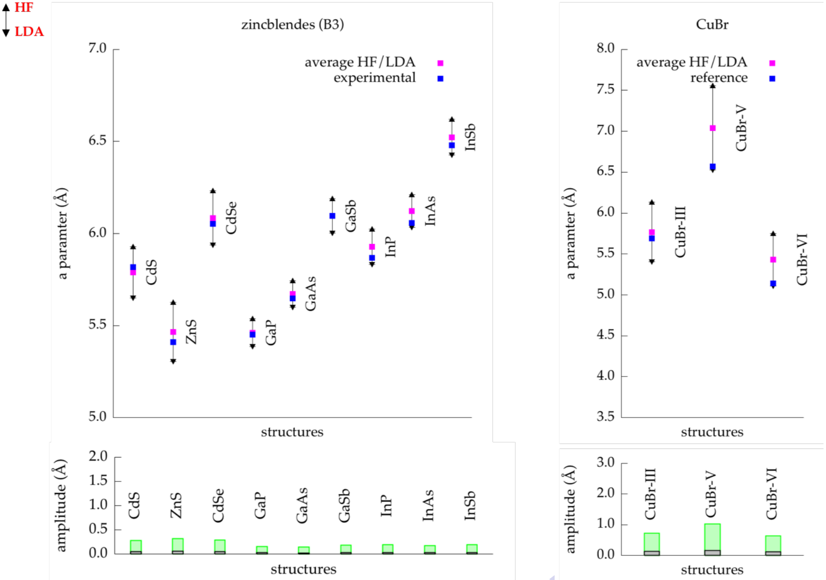


# Alkali halides

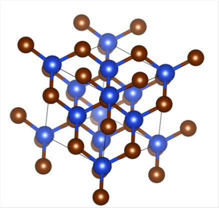




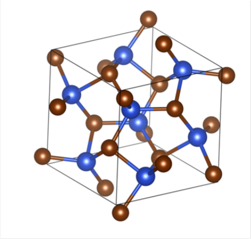
# More structures



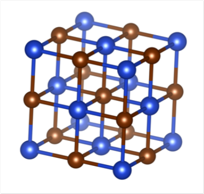
# Copper bromide



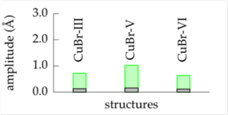
CuBr III
216



CuBr V
205



CuBr VI
225



- The HF-LDA separation is larger for CuBr than for "well-behaved" structures;
- For CuBr-V, the HF and LDA parameters are more than 1 Å apart.

# Summary of HF/LDA error bar

- If the delocalization error is the main error source, the HF/LDA separation is a measure of the variability that the choice of DFT functional introduces in our results.