Gold standard finite temperature simulations of materials via machine learning

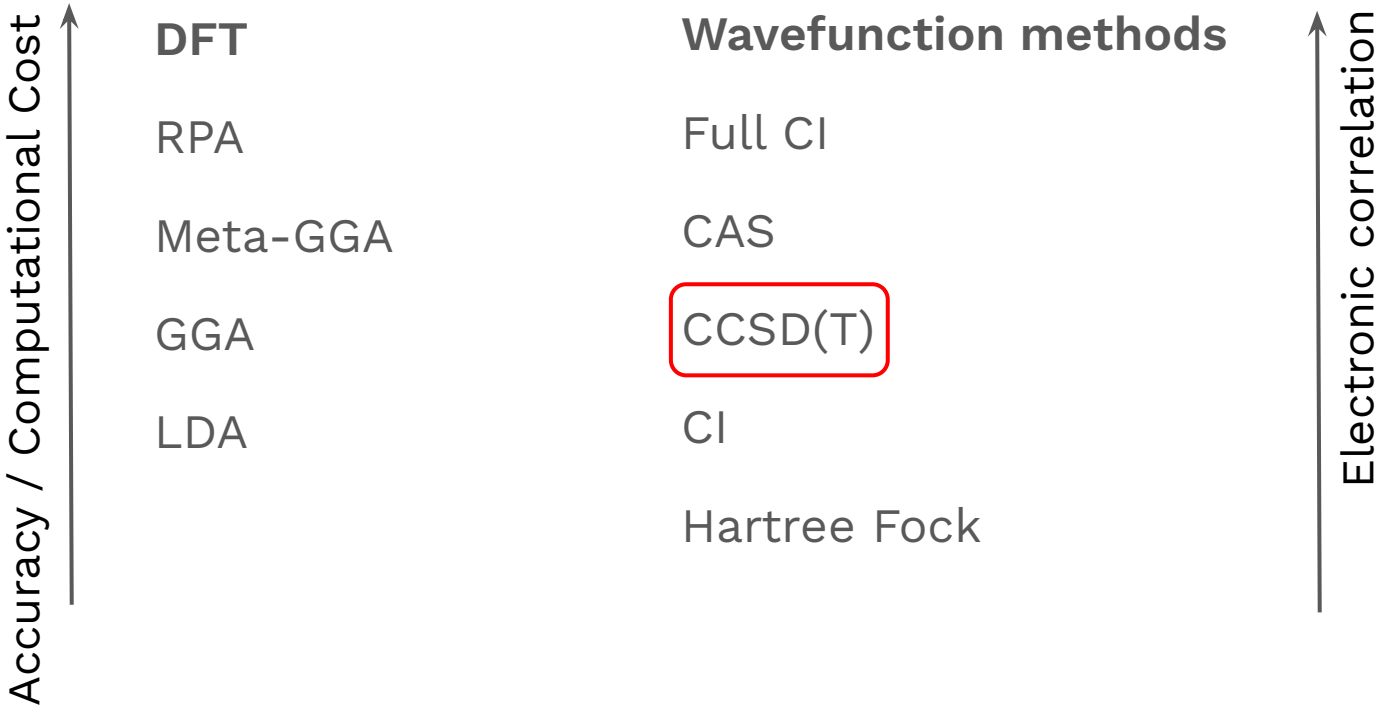
Basile Herzog[1], Alejandro Gallo[2], Mauricio Chagas da Silva[1], Andreas Irmler[2], Felix Hummel[2], Michael Badawi[1], Tomas Bucko[3], Sébastien Lebegue[1], Andreas Grüneis[2] and Dario Rocca[1]

[1] LPCT, Université de Lorraine & CNRS, Nancy (France)
[2] Institute for Theoretical Physics, TU Wien, Vienna (Austria)
[3] Comenius University in Bratislava and Slovak Academy of Sciences, Bratislava (Slovakia)

Outline

- Introduction & Motivation
- Machine Learning Perturbation Theory (MLPT)
- Adsorption enthalpy of CO_2 in Protonated Chabazite
- Possible limitations and solution: Machine Learning Monte-Carlo (MLMC)
- Conclusion

Methods vs computational cost



Coupled Cluster Theory

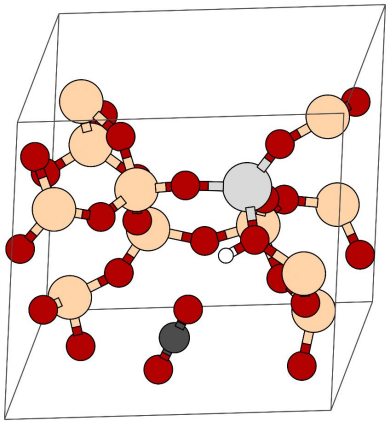
- CCSD: $|\Psi\rangle = e^{T_1+T_2}|HF\rangle$
- T_1,T_2: single and double substitutions
  triples are treated perturbatively
- exponential ansatz: allows systematics inclusion of highest degree of correlations (Taylor expansion)

- Gold standard method in the quantum chemistry community: reaches chemical accuracy for most applications
- $O(N^6)$ complexity

Liao, Ke, and Andreas Grüneis. 2016. “Communication: Finite Size Correction in Periodic Coupled Cluster Theory Calculations of Solids.” *The Journal of Chemical Physics* 145 (14): 141102.

Case study: adsorption enthalpy of CO_2 in protonated chabazite

$$\Delta H = \langle E(CO_2@HCHAB)\rangle - \langle E(HCHAB)\rangle - \langle E(CO_2)\rangle - k_B T$$



- *Ab-initio* Molecular Dynamics (AIMD) - 200K steps

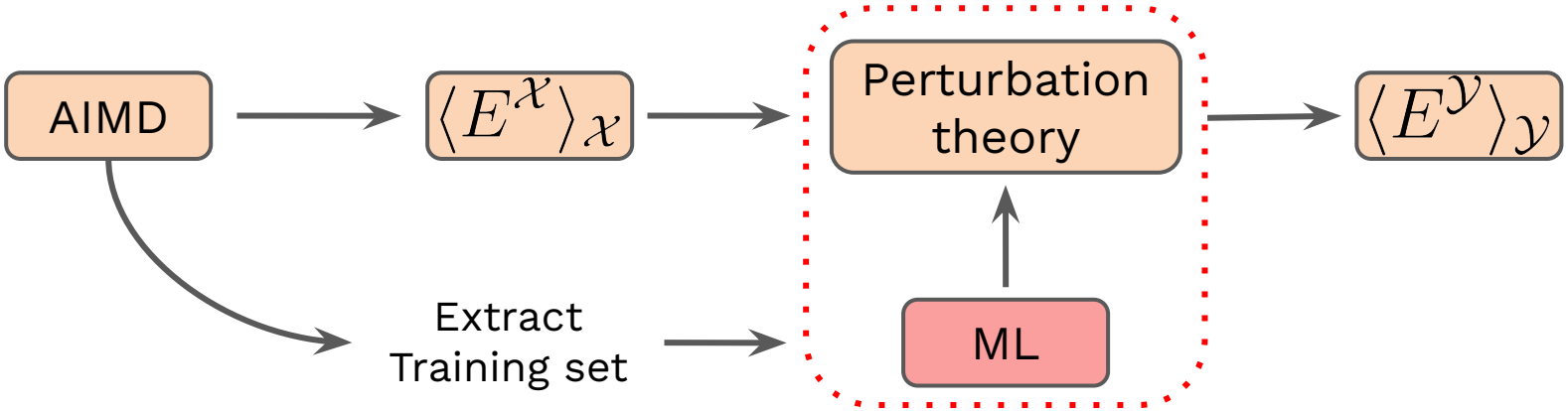
-CCSD(T) ⟶ 10B CPU Hours ⟶ 1000 human years

- Machine learning thermodynamic perturbation theory (MLPT)

a “cheap” AIMD (e.g. PBE) + only a few number (~100) of CCSD(T) calculations ⟶ 1 human week

B. Chehaibou, M. Badawi, T. Bucko, T. Bazhirov, and D. Rocca, JCTC (2019)

Machine Learning Thermodynamic Perturbation Theory



1. AIMD is performed using the cheap theory $\mathcal{X} \rightarrow \langle E^{\mathcal{X}} \rangle_{\mathcal{X}}$
2. $\Delta E_i = E_i^{\mathcal{X}} - E_i^{\mathcal{Y}}$ is learned on Ntrain~100 configurations evenly spaced from the MD
3. Perturbation theory is applied to $\{E_i^{\mathcal{X}}\} \rightarrow \langle E^{\mathcal{Y}} \rangle_{\mathcal{Y}}$

Thermodynamic Perturbation Theory

- Computationally cheap Hamiltonian (e.g. PBE)

$$H^{\mathcal{X}} = T + V^{\mathcal{X}}$$

- Average in canonical ensemble of $H^{\mathcal{X}}$

$$\langle O \rangle_{\mathcal{X}}$$

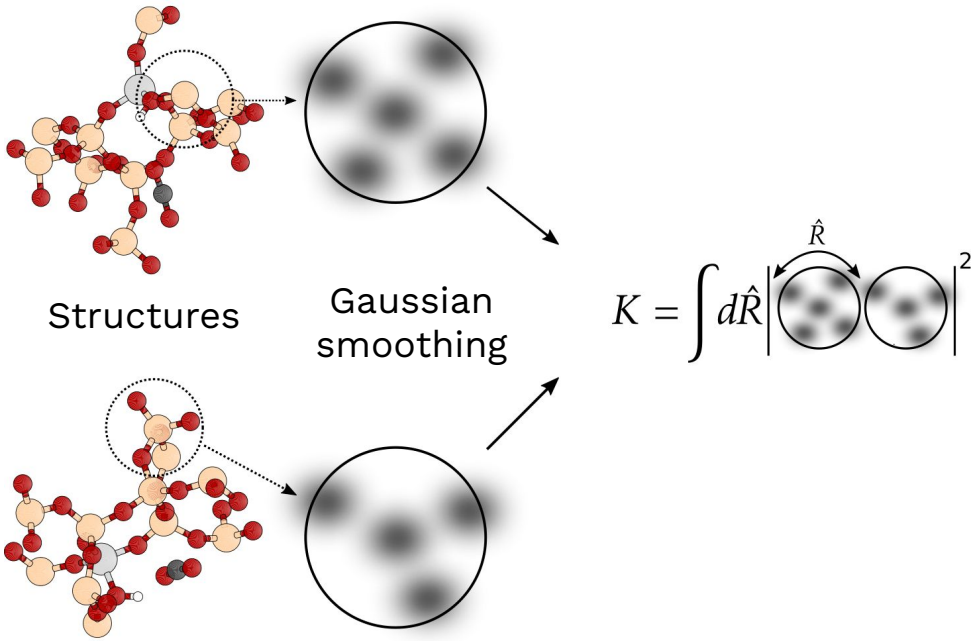
- Computationally expensive Hamiltonian (e.g. CCSD(T))

$$\begin{aligned} H^{\mathcal{Y}} &= H^{\mathcal{X}} + V^{\mathcal{Y}} - V^{\mathcal{X}} \\ &= H^{\mathcal{X}} - \Delta V \end{aligned}$$

- Average in canonical ensemble of $H^{\mathcal{Y}}$

$$\langle O \rangle_{\mathcal{Y}} = \frac{\langle O e^{\beta \Delta V} \rangle_{\mathcal{X}}}{\langle e^{\beta \Delta V} \rangle_{\mathcal{X}}}$$

Machine Learning model : Descriptor & KRR



- Smooth Overlap of Atomic Positions (SOAP)

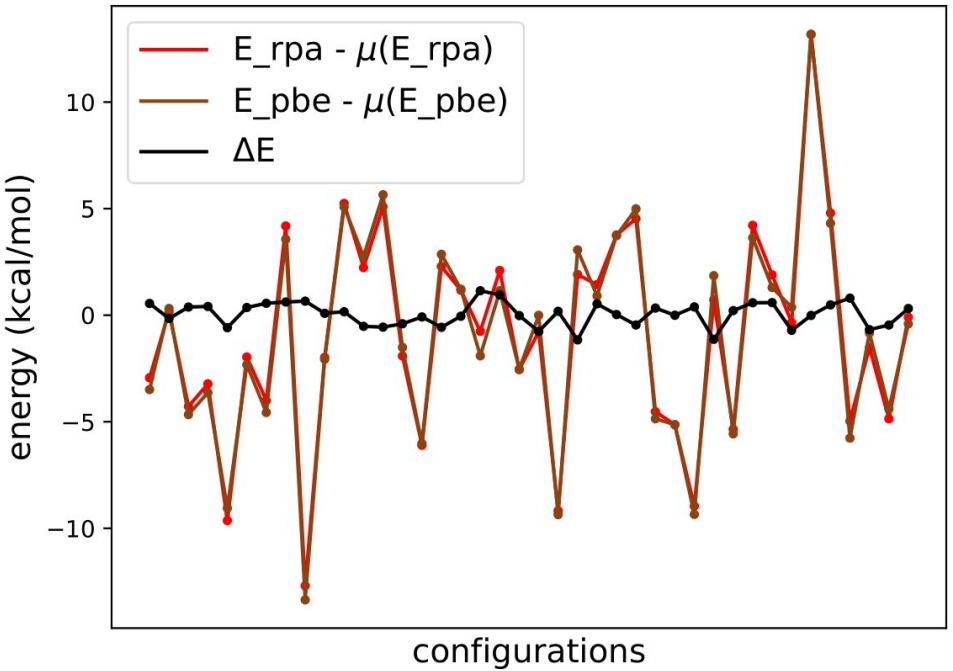
- Kernel Ridge Regression (KRR)

$$\alpha = (K_{train} + \lambda I)^{-1} y_{train}$$

$$y_{pred} = K_{pred} \alpha$$

Albert P. Bartók, Risi Kondor, and Gábor Csányi, PRB 2013
Sandip De, Albert P. Bartók, Gábor Csányi, and Michele Ceriotti, PCCP 2016

Machine Learning model : Δ-Machine Learning



ΔE: smooth function, easy to learn

Small training set:
Ntrain=100/Nval=10

System	HCHAB	CO2@HCHAB
RMSE (kcal/mol)	0.50	0.63

R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, JCTC (2015)
B. Chehaibou, M. Badawi, T. Bucko, T. Bazhirov, and D. Rocca, JCTC (2019)

Results : CO_2@HCHAB

Method	PBE-D2	MP2	CCSD(T)	Exp.
Enthalpy (kcal/mol)	-9.72	-9.50	-7.69	-8.41

- Good agreement with experiment, within chemical accuracy
- First finite temperature result at this level of theory

Possible sources of error

$$\langle E^{\mathcal{Y}} \rangle_{\mathcal{Y}} = \frac{\langle E^{\mathcal{Y}} e^{\beta \Delta E} \rangle_{\mathcal{X}}}{\langle e^{\beta \Delta E} \rangle_{\mathcal{X}}} = \sum_i^M \frac{w_i E_i^{\mathcal{Y}}}{\sum_j w_j}$$

$$w_i = \exp\left(\beta \Delta E_i\right)$$

Production and target methods might have insufficient statistical overlap: averages in target space would be dominated by few configurations.

$$I_w = \frac{(M-N)}{M} \in [0, 0.5]$$

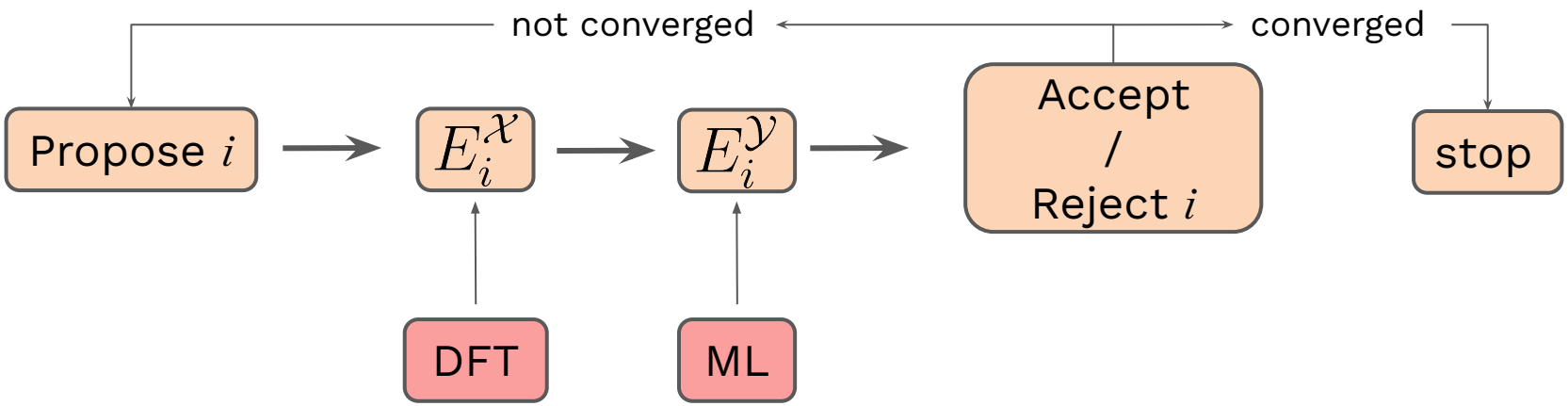
N,M such that $\frac{\sum_i^N w_i}{\sum_j^M w_j} \geq 0.5$

$I_w = 0$: MLPT will fail

Solution: Machine Learning Monte-Carlo (MLMC)

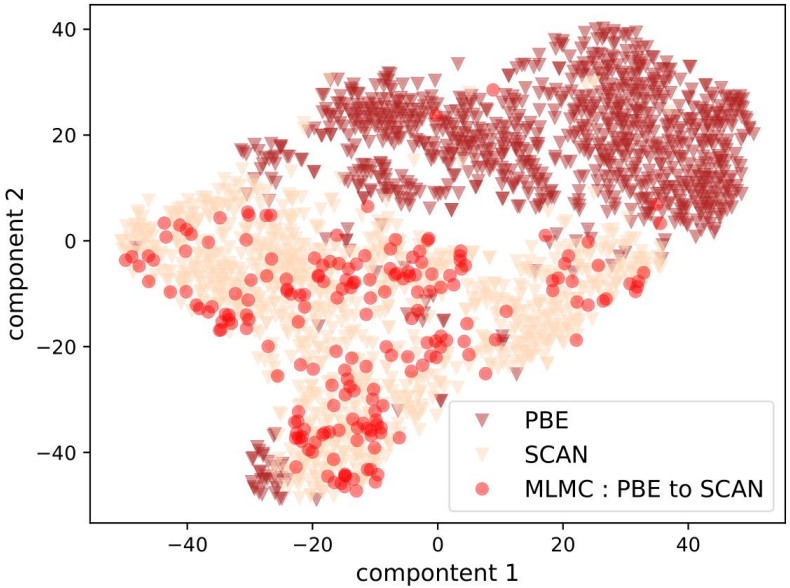
B. Herzog, M. Chagas da Silva, ... & D. Rocca, JCTC (2022).

Machine Learning Monte-Carlo



1. Energy of the proposed configuration is computed using the cheap theory
2. Correction to the expensive theory is done using the ML model of the previous MLPT
3. Sampling is done in the expensive configurational space until convergence

Example: MLPT estimate of SCAN energy from PBE



ΔE(MLPT)	ΔE(MLMC)
-4.38	0.64

- CH_4 in protonated chabazite
- Case study with DFT functionals SCAN & PBE. MD reference is known
- MLPT $\longrightarrow$ I_w=0, bad overlap, big deviation compared to reference
- MLMC trajectory lies in correct configurational space, good agreement with reference

Conclusion

- The MLPT method allows computation of highly accurate thermodynamic property at CCSD(T) level for the first time
- Monte-Carlo resampling is currently running to confirm this result
- Future directions: other porous materials applications, surface adsorption problems, activation energies

Acknowledgements

Group leaders: Dario Rocca*, Sébastien Lebegue & Andreas Grüneis



Collaborators: Mauricio Chagas da Silva, Alejandro Gallo, Andreas Irmler, Felix Hummel

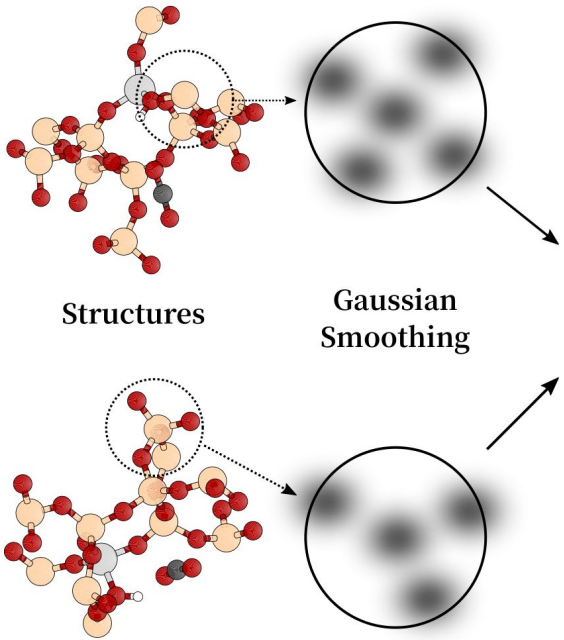
Michael Badawi for the porous materials applications and computational time

Tomáš Bučko for the fruitful discussions and collaboration that led to the development of MLPT

* dario.rocca@univ-lorraine.fr

SOAP Kernel

$$\rho^{\chi_a}(\boldsymbol{r}) = \sum_{i \in \chi_a} \exp((\boldsymbol{x}_i - \boldsymbol{r})/2\sigma^2)$$



$$K = \int d\hat{R} \left| \quad \right|^2$$

$$K(\chi_a, \chi'_b) = \int d\hat{\boldsymbol{R}} \left| \int d\boldsymbol{r} \rho^{\chi_a}(\boldsymbol{r}) \rho^{\chi'_b}(\hat{\boldsymbol{R}}\boldsymbol{r}) \right|^2$$

$$K(\chi, \chi') = \sum_{a=1}^{N_\chi} \sum_{b=1}^{N_{\chi'}} \frac{K(\chi_a, \chi'_b)}{\sqrt{K(\chi_a, \chi_a) K(\chi'_b, \chi'_b)}}$$

$$\alpha = (K_{train} + \lambda I)^{-1} y_{train}$$

$$y_{pred} = K_{pred} \alpha$$

Electronic Correlation

Chemical bonds

Non covalent interactions

Free energy profiles

Strong correlations ?