MACHINE LEARNING POTENTIALS

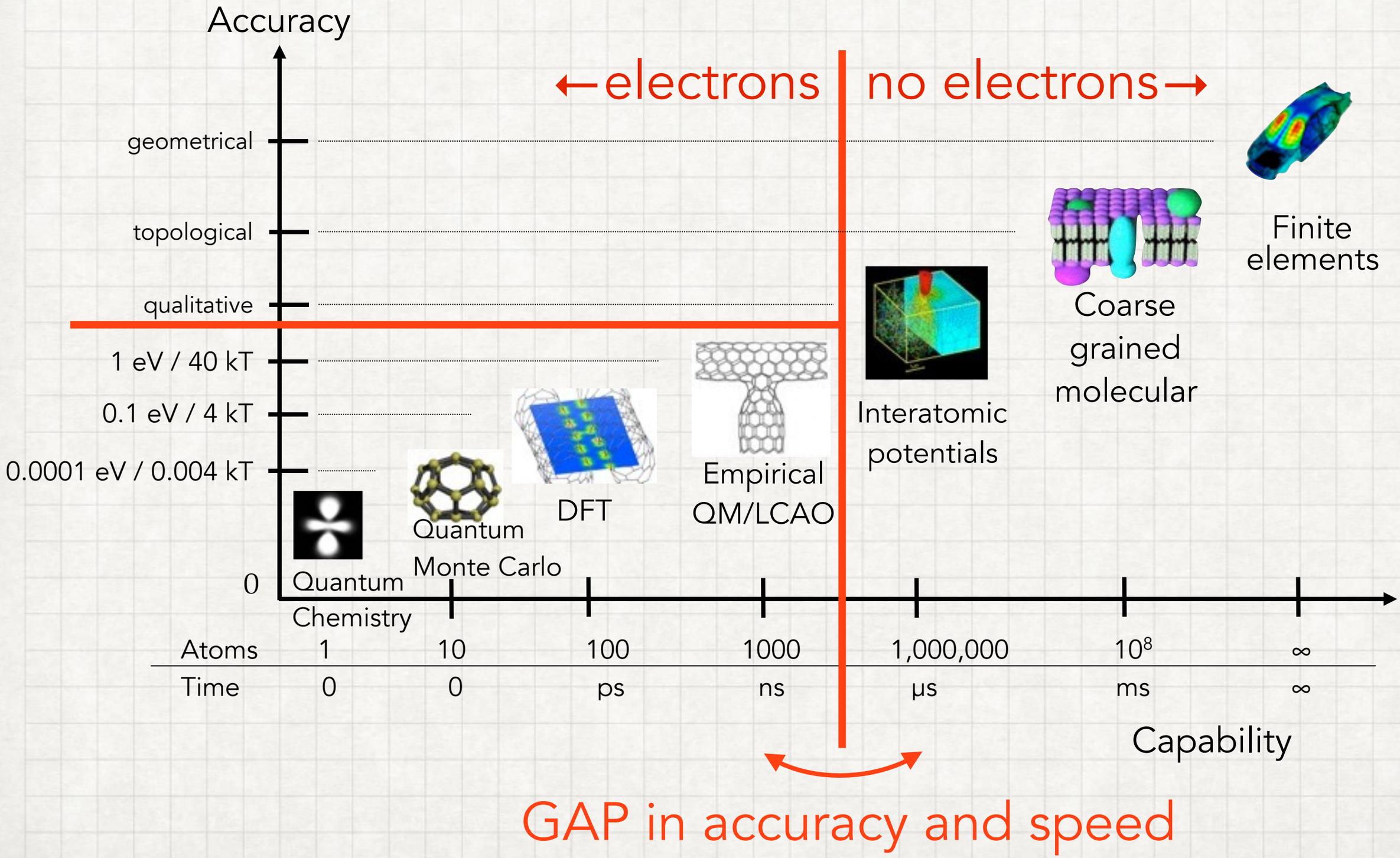
FITTING POTENTIALS AND SEARCHING CHEMICAL SPACE

Albert Bartók-Pártay

UNIVERSITY OF CAMBRIDGE

Department of Engineering

MODELLING MATERIALS



QUICK INTRO TO MACHINE LEARNING

- function fitting
- infer a function based on data
- many dimensions
- good physical model is unknown/doesn't exist
- avoid overfitting by incorporating prior assumptions (smoothness)
- **pros**:
  - automatic (sort of)
  - flexible
  - easy to expand with more data
  - error estimate
- **cons**:
  - slow
  - result is not a physical model
  - little extrapolation
- See David MacKay's book: Information Theory, Inference and Learning Algorithms

QUICK INTRO TO MACHINE LEARNING

THE PROBLEM

$f : \mathbb{R}^d \to \mathbb{R}$ unknown

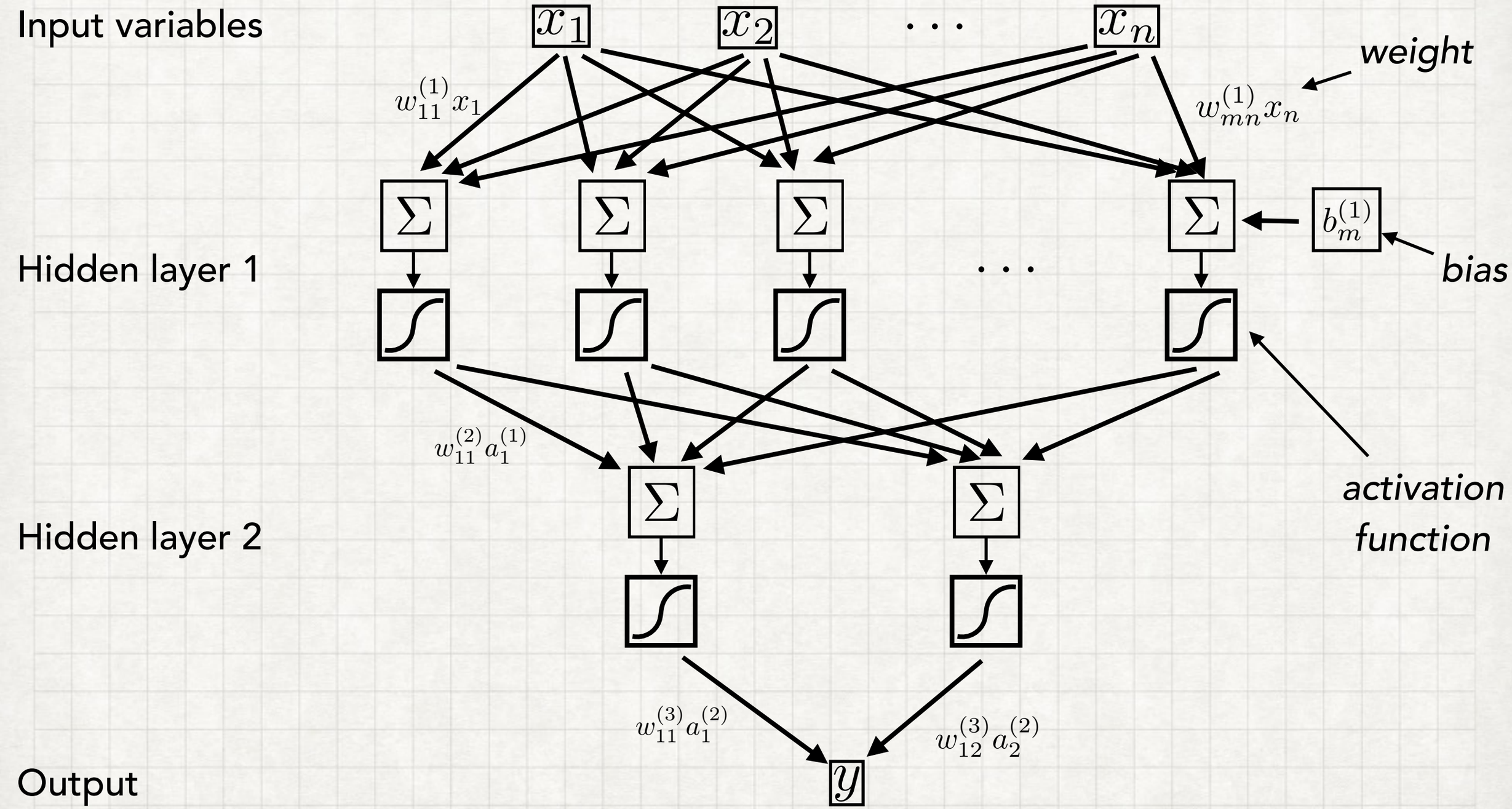
Make predictions of $f(\mathbf{x})$ given data points $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$

and (noisy) observations $\{y_1, y_2, \ldots, y_N\}$

Solutions:

- linear interpolation
- splines
- fitting polynomials
- neural networks
- Gaussian Process regression

NEURAL NETWORKS



NEURAL NETWORKS

$$f(x) = x$$

$$f(x) = \frac{1}{1 - \exp(x)}$$

$$f(x) = \tanh(x)$$

$$f(x) = \Theta(x)$$

$$L = \sum_i [y_i - \mathrm{NN}(\mathbf{x}_i; \mathbf{w}, \mathbf{b})]^2$$

GAUSSIAN PROCESS REGRESSION

WEIGHT-SPACE VIEW

$$y(\mathbf{x}) = \sum_i \alpha_i K(\mathbf{x}_i, \mathbf{x})$$

$$\mathbf{y} = \mathbf{K}\boldsymbol{\alpha}$$

$$L = ||\mathbf{y} - \mathbf{K}\boldsymbol{\alpha}||^2 + \lambda||\alpha||^2_{\mathbf{K}}$$

$$L = \mathbf{y}^T\mathbf{y} - 2\boldsymbol{\alpha}^T\mathbf{K}\mathbf{y} + \boldsymbol{\alpha}^T\mathbf{K}^T\mathbf{K}\boldsymbol{\alpha} + \lambda\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha}$$

$$\mathbf{0} = \nabla_{\boldsymbol{\alpha}} L = -2\mathbf{K}\mathbf{y} + 2\mathbf{K}\mathbf{K}\boldsymbol{\alpha} + 2\lambda\mathbf{K}\boldsymbol{\alpha}$$

$$\boldsymbol{\alpha} = (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y}$$

GAUSSIAN PROCESS REGRESSION

FUNCTION-SPACE VIEW

$$y(\mathbf{x}) = \sum_h w_h \phi_h(\mathbf{x}) + \varepsilon$$

$$\mathbf{w} \sim N(\mathbf{0}, \sigma_w) \qquad \varepsilon \sim N(\mathbf{0}, \sigma_\varepsilon)$$

$$\langle y_i y_j \rangle = \left\langle \sum_{hh'} w_h w'_h \phi_h(\mathbf{x}_i)\phi_h(\mathbf{x}_j) + \varepsilon_i \varepsilon_j \right\rangle =$$

$$\sum_{hh'} \sigma_w^2 \delta_{hh'} \phi_h(\mathbf{x}_i)\phi_h(\mathbf{x}_j) + \sigma_\varepsilon^2 \delta_{ij} =$$

$$\sigma_w^2 \boldsymbol{\phi}(\mathbf{x}_i)^T \boldsymbol{\phi}(\mathbf{x}_j) + \sigma_\varepsilon^2 \delta_{ij} = K(\mathbf{x}_i, \mathbf{x}_j) + \sigma_\varepsilon^2 \delta_{ij} = C(\mathbf{x}_i, \mathbf{x}_j)$$

GAUSSIAN PROCESS REGRESSION

FUNCTION-SPACE VIEW

$$P(\mathbf{y}) \propto \exp\left(-\frac{1}{2}\mathbf{y}^T\mathbf{C}^{-1}\mathbf{y}\right)$$

$$P(y \mid \mathbf{y}) \propto P([\mathbf{y}, y])$$

$$P(y \mid \mathbf{y}) \propto P([\mathbf{y}, y]) = \exp\left(-\frac{1}{2}[\mathbf{y}, y]\begin{bmatrix}\mathbf{C} & \mathbf{k} \\ \mathbf{k}^T & \kappa\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{y} \\ y\end{bmatrix}\right)$$

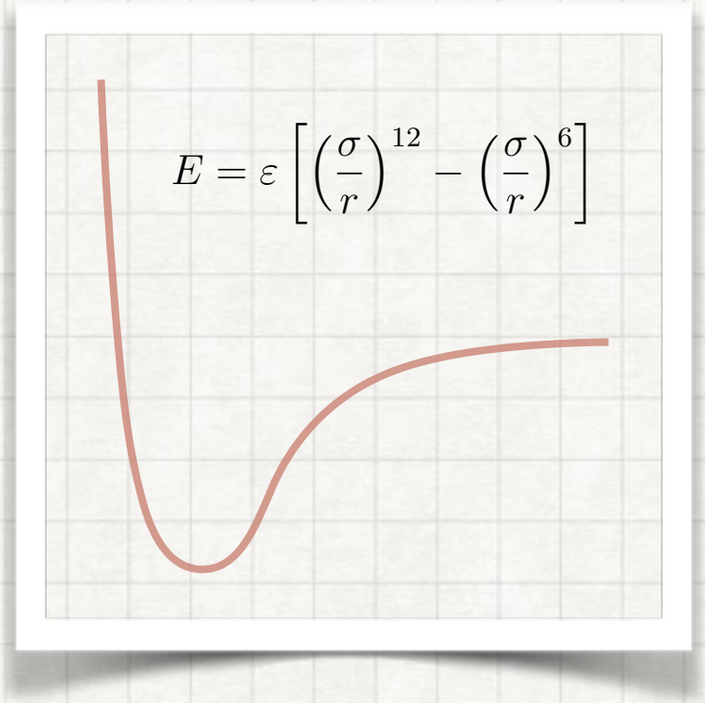
It's a Gaussian distribution with

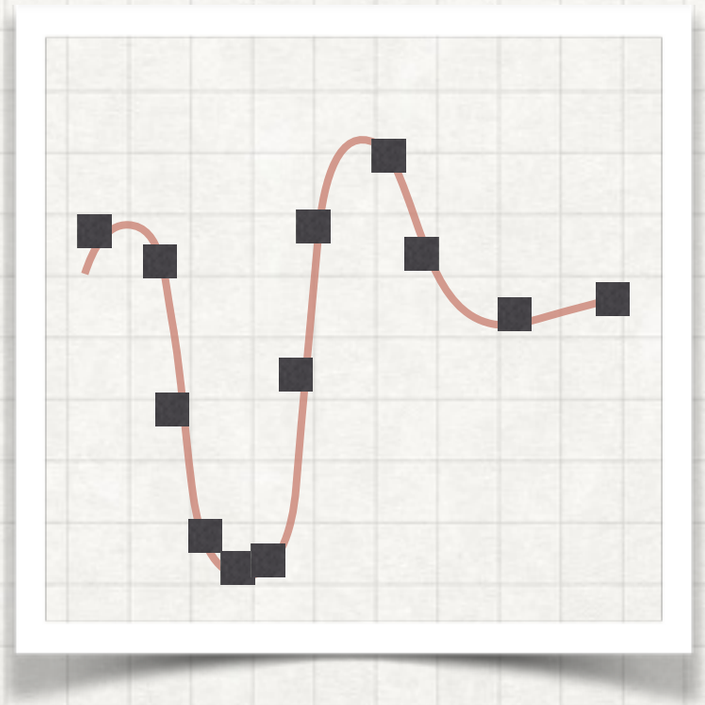
$$\hat{y} = \mathbf{k}^T\mathbf{C}^{-1}\mathbf{y}$$

$$\sigma^2_{\hat{y}} = \kappa - \mathbf{k}^T\mathbf{C}^{-1}\mathbf{k}$$

INTERATOMIC POTENTIALS

- Analytic potentials
  - fixed functional formula
  - based on physical understanding
  - few fixed parameters
  - fit to experimental and/or computational data
- Machine learning potentials
  - flexible functional form
  - no physical motivation
  - data driven - mostly computational





PES FITTING

$$E_{\text{total}} = \sum_i^{\text{atoms}} \varepsilon(\text{neighbourhood}_i) + \text{long range}$$

$$E_{\text{total}} = \sum_i^{\text{atoms}} \varepsilon(\{\mathbf{r}_{ij}\}_j^{\text{neighbours}_i}) + \text{long range}$$

$$\varepsilon(\mathbf{x}) = \sum_n \alpha_n k(\mathbf{x}_n, \mathbf{x})$$

atomic environment

- Represent a local neighbourhood configuration: “descriptor”, “fingerprint”, “feature vector”, “symmetry function”
  - Rotational, reflectional, translational and permutational invariance
  - Faithfulness - no two different configuration give the same representation
  - Continuous, differentiable and smooth
- Fits are based on comparisons of configurations
  - Gaussian kernel
  - Dot-product kernel aka linear regression
  - Neural network kernel

$$k(\mathbf{x}, \mathbf{x}') = \exp(-|\mathbf{x} - \mathbf{x}'|^2 / 2\sigma^2)$$

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{x} \cdot \mathbf{x}'$$

$$k(\mathbf{x}, \mathbf{x}') \approx V - |\mathbf{x} - \mathbf{x}'|^2$$

GAUSSIAN PROCESS

- total energy is sum of local contributions

$$\varepsilon(\mathbf{x}) = \sum_n \alpha_n k(\mathbf{x}_n, \mathbf{x})$$

- database of atomic configurations, evaluated by QM

$$E_j = \sum_i^{\text{atoms}_j} \varepsilon(\mathbf{x}_i)$$

- minimise cost function

$$L = \sum_j \left| E_j - \sum_i^{\text{atoms}_j} \varepsilon(\mathbf{x}_i) \right|^2 + \lambda ||\alpha||^2$$

- regularising: factor & norm

$$||\alpha||^2 = \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha}$$

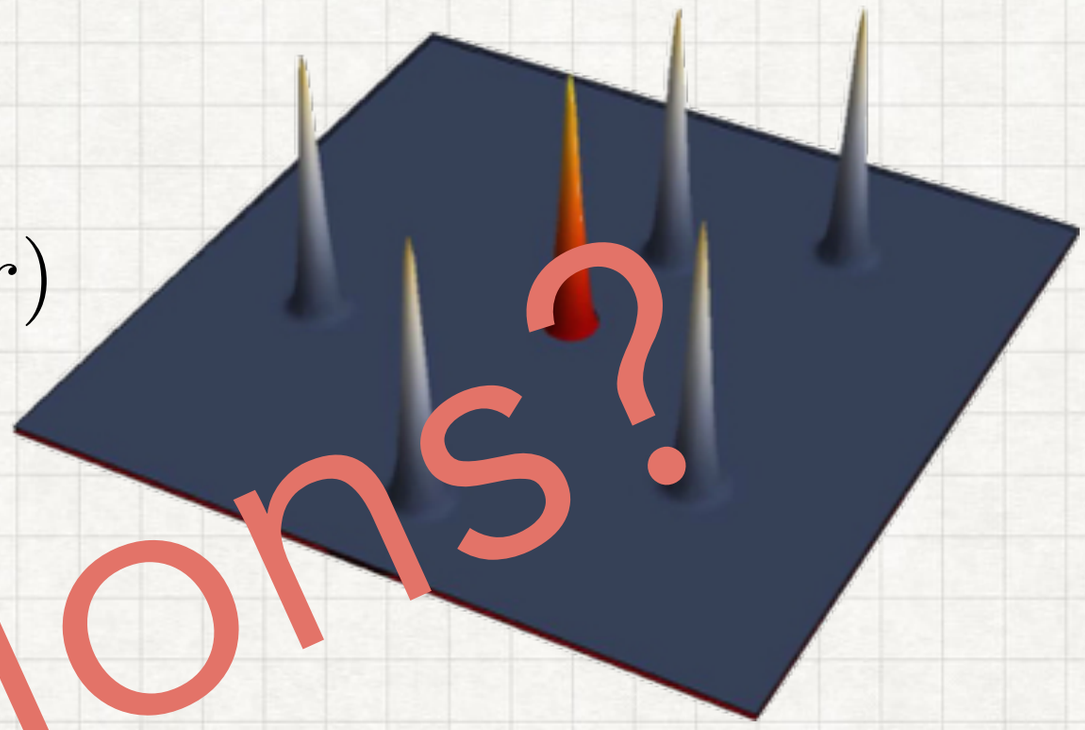
- result: closed formula for coefficients

ATOMIC DENSITY

$$\rho_i(\mathbf{r}) = \sum_j \delta(\mathbf{r} - \mathbf{r}_{ij}) f_{\text{cut}}(r)$$

$$S_{ij} = \int \mathrm{d}\mathbf{r} \rho_i(\mathbf{r}) \rho_j(\mathbf{r})$$

$$\rho_i(\mathbf{r}) = \sum_{nlm} c^{(i)}_{nlm} g_n(r) Y_{lm}(\hat{\mathbf{r}})$$



- Translational invariance
- Permutational invariance
- Expanding into a basis

POWER SPECTRUM

For simplicity, omit radial dependence

$$\rho(\mathbf{r}) = \sum_{lm} c_{lm} Y_{lm}(\hat{\mathbf{r}})$$

$$\rho(\hat{R}\,\mathbf{r}) = \sum_{lm} c_{lm} Y_{lm}(\hat{R}\,\hat{\mathbf{r}}) = \sum_{lm} c_{lm} \sum_{m'} D^l_{mm'}(\hat{R}) Y_{lm'}(\hat{\mathbf{r}})$$

$$= \sum_{lm'} \left(\sum_{m} D^l_{mm'} c_{lm} \right) Y_{lm'} \equiv \sum_{lm'} c'_{lm'} Y_{lm'}$$

$$\mathbf{c}_l \xrightarrow{\hat{R}} \mathbf{D}^l \mathbf{c}_l \qquad \text{or equivalently} \qquad \mathbf{c}'_l = \mathbf{D}^l \mathbf{c}_l$$

POWER SPECTRUM

- The Wigner D-matrices are unitary

$$\mathbf{D}^{l\,\dagger}\mathbf{D}^{l} = \mathbf{I}$$

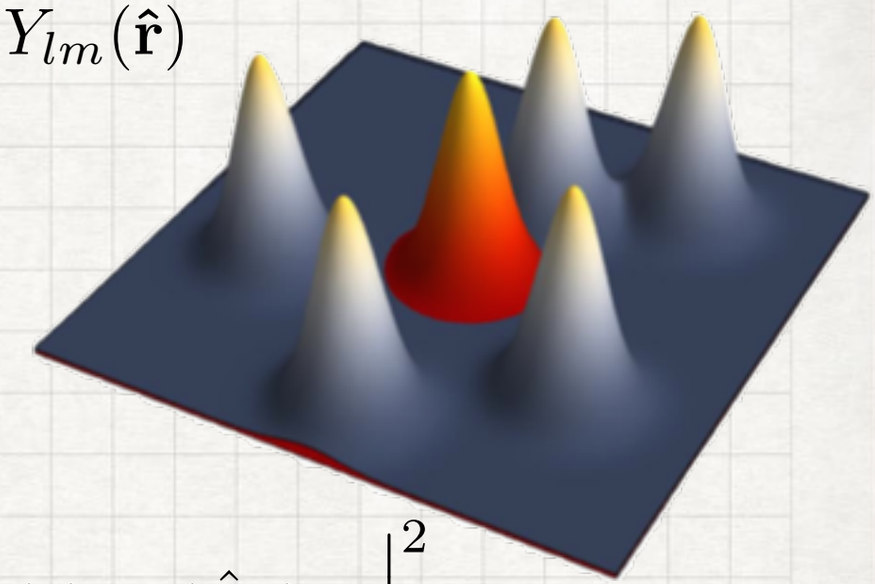
- Construct an invariant

$$p_l = \mathbf{c}_l^{\dagger}\mathbf{c}_l \overset{\hat{R}}{\rightarrow} \left(\mathbf{c}_l^{\dagger}\mathbf{D}^{l\,\dagger}\right)\mathbf{D}^{l}\mathbf{c}_l = \mathbf{c}_l^{\dagger}\mathbf{c}_l = p_l$$

  - Power spectrum elements are equivalent to Steinhardt Q_l bond-order parameters
  - Higher order invariants (bispectrum) correspond to the Steinhardt W_l bond-order parameter

SMOOTH OVERLAP OF ATOMIC POSITIONS

$$\rho_i(\mathbf{r}) = \sum_j \exp\left(-\alpha|\mathbf{r} - \mathbf{r}_{ij}|^2\right) = \sum_j \sum_{lm} c_{nlm}^{(i)j} g_n(r) Y_{lm}(\hat{\mathbf{r}})$$



- Overlap integral

$$S(\rho_i, \rho_{i'}) = \int \rho_i(\mathbf{r}) \rho_{i'}(\mathbf{r}) d\mathbf{r},$$

- Integrate over all rotations:

$$k(\rho_i, \rho_{i'}) = \int \left|S(\rho_i, \hat{R}\rho_{i'})\right|^2 d\hat{R} = \int d\hat{R} \left| \int \rho_i(\mathbf{r}) \rho_{i'}(\hat{R}\mathbf{r}) d\mathbf{r} \right|^2$$

- ...[few pages of algebra]...

$$k(\rho_i, \rho_{i'}) = \sum_{n,n',l} p_{nn'l}^{(i)} p_{nn'l}^{(i')} \qquad p_{nn'l} = \mathbf{c}_{nl}^{\dagger} \mathbf{c}_{nl}$$

- Final formula: dot-product of power spectra!

KERNELS TO COMPARE SMALL CLUSTERS

Example: water dimer

- Representation: interatomic distances
  - 6 atoms - 15 interatomic distances

$$\mathbf{x} = \{r_{ij}\}$$

- Symmetrise kernel function over symmetry group:

$$\tilde{K}(x, x') = \sum_{p \in S} K(p(x), x')$$

GP TO FIT MULTIPLE MODELS

- Example: total energy is sum of pair and triplets

$$E = \sum_{p \in \text{ pairs}} \varepsilon_p^{(2)} + \sum_{t \in \text{ triplets}} \varepsilon_t^{(3)}$$

- each term is a GP itself

$$\varepsilon^{(2)}(..) = \sum_n \alpha_n^{(2)} k(..{}_n, ..)$$

- covariance is a sum pair and triplet covariances

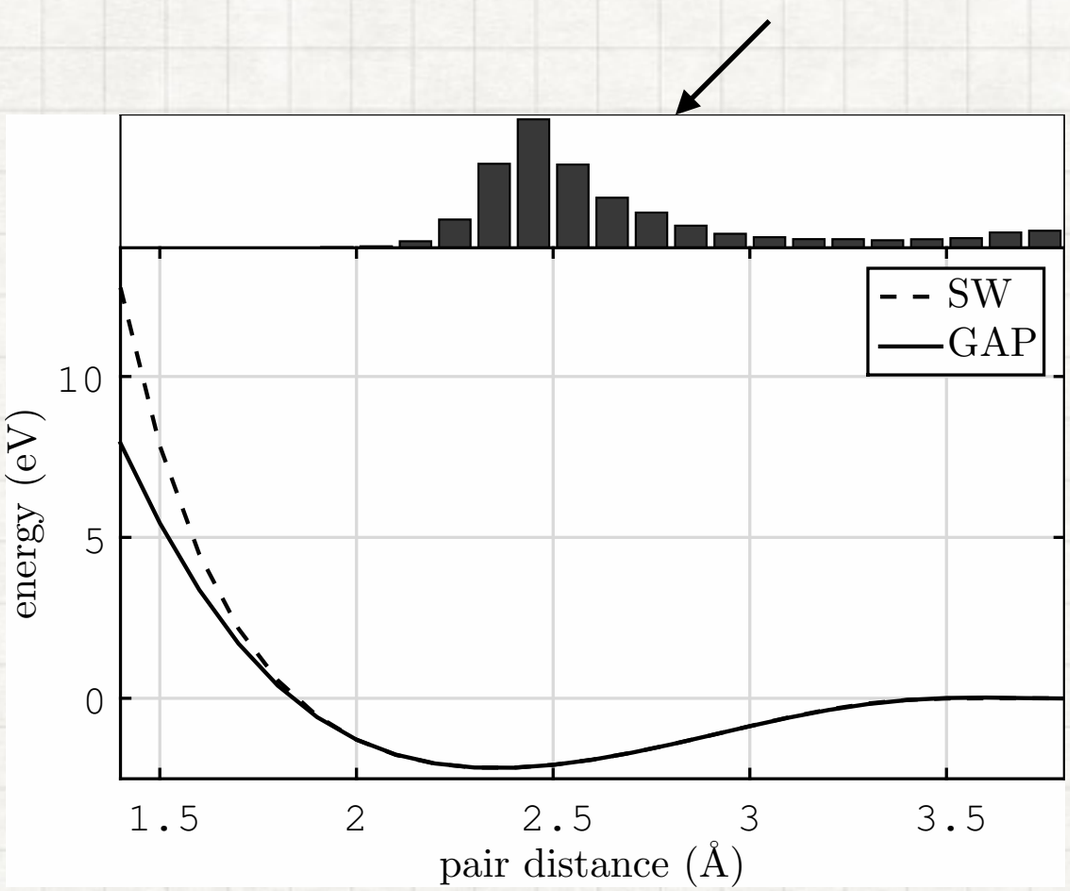
$$\varepsilon^{(3)}(.\dot{.}) = \sum_n \alpha_n^{(3)} k(.\dot{.}_n, .\dot{.})$$

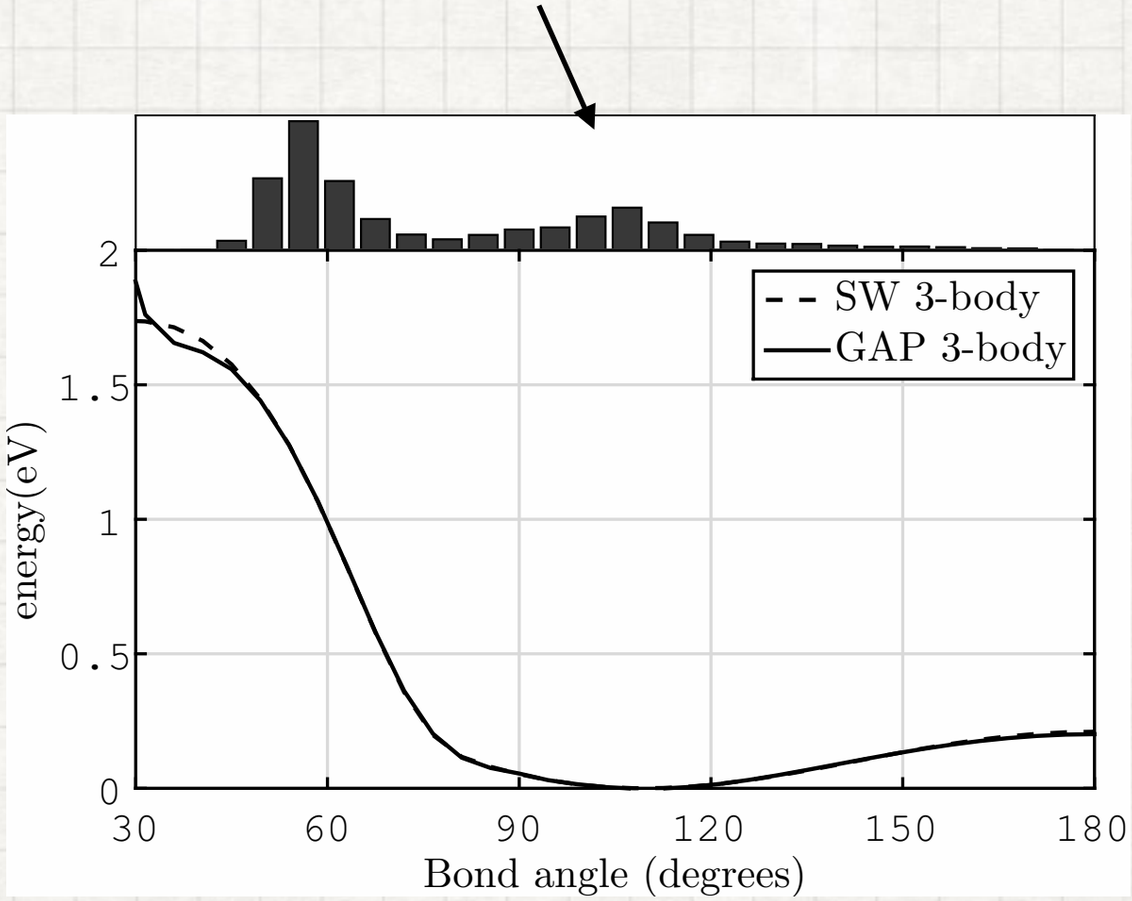
$$\langle E_N E_M \rangle = \sigma^2_{w^{(2)}} \sum_{p \in \text{pairs}_N} \sum_{q \in \text{pairs}_M} C^{(2)}(p,q) + \sigma^2_{w^{(3)}} \sum_{t \in \text{triplets}_N} \sum_{u \in \text{triplets}_M} C^{(3)}(t,u)$$

GP TO FIT MULTIPLE MODELS

Fit of a two and three-body model on Si SW-data

$$\langle E_N E_M \rangle = \sigma^2_{w^{(2)}} \sum_{p \in \text{pairs}_N} \sum_{q \in \text{pairs}_M} C^{(2)}(p,q) + \sigma^2_{w^{(3)}} \sum_{t \in \text{triplets}_N} \sum_{u \in \text{triplets}_M} C^{(3)}(t,u)$$





HOW TO GENERATE DATABASES

- Target applications: large systems
- Capability of full quantum mechanics (QM): small systems

QM MD on "representative" small systems:

sheared primitive cell → elasticity
large unit cell → phonons
surface unit cells → surface energy
gamma surfaces → screw dislocation
vacancy in small cell → vacancy
vacancy @ gamma surface → vacancy near dislocation

Iterative refinement

1. QM MD → Initial database
2. Model MD

What is the acceptable validation protocol?
How far can the domain of validity be extended?

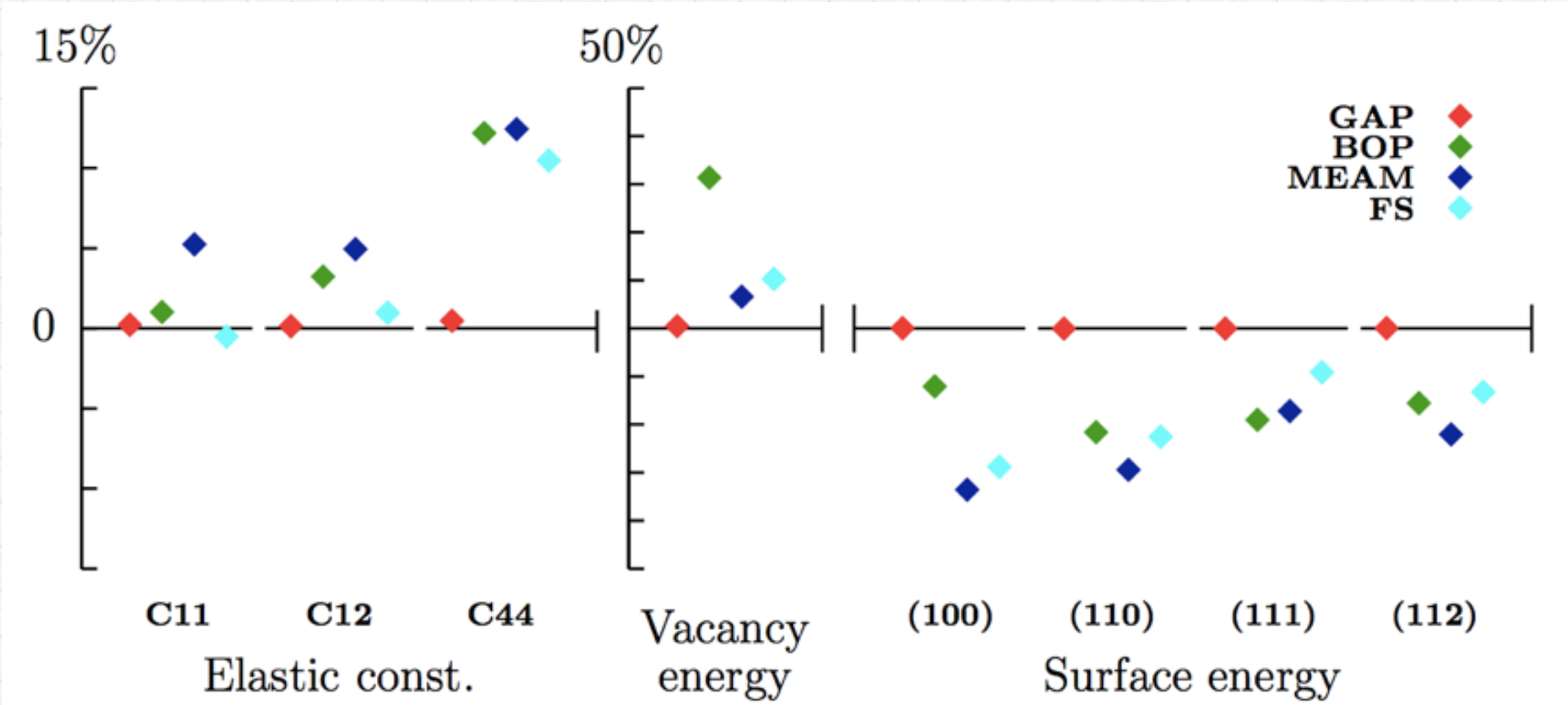
BUILDING A DATABASE FOR TUNGSTEN

	Database:	Computational cost[a] [ms/atom]	Elastic constants[b] [GPa]	Phonon spectrum[b] [THz]	Vacancy formation[c] [eV]	Surface energy[b] [eV/Å^2]	Dislocation structure[d] [Å^{-1}]	Dislocation-vacancy binding energy [eV]	Peierls barrier [eV/b]
GAP_1 :	2000 × primitive unit cell with varying lattice vectors	24.70	0.623	*0.583*	*2.855*	*0.1452*	*0.0008*		
GAP_2 :	GAP_1 + 60 × 128 atom cell	51.05	0.608	0.146	*1.414*	*0.1522*	*0.0006*		
GAP_3 :	GAP_2 + vacancy in: 400 × 53 atom cell, 20 × 127 atom cell	63.65	0.716	0.142	0.018	*0.0941*	*0.0004*		
GAP_4 :	GAP_3 + (100), (110), (111), (112) surfaces 180 × 12 atom cell (110), (112) gamma surfaces 6183 × 12 atom cell	86.99	0.581	0.138	0.005	0.0001	0.0002	*-0.960*	*0.108*
GAP_5 :	GAP_4 + vacancy in: (110), (112) gamma surface 750 × 47 atom cell	93.86	0.865	0.126	0.011	0.0001	0.0002	-0.774	*0.154*
GAP_6 :	GAP_5 + $\frac{1}{2}\langle 111\rangle$ dislocation quadrupole 100 × 135 atom cell	93.33	0.748	0.129	0.015	0.0001	0.0001	-0.794	0.112

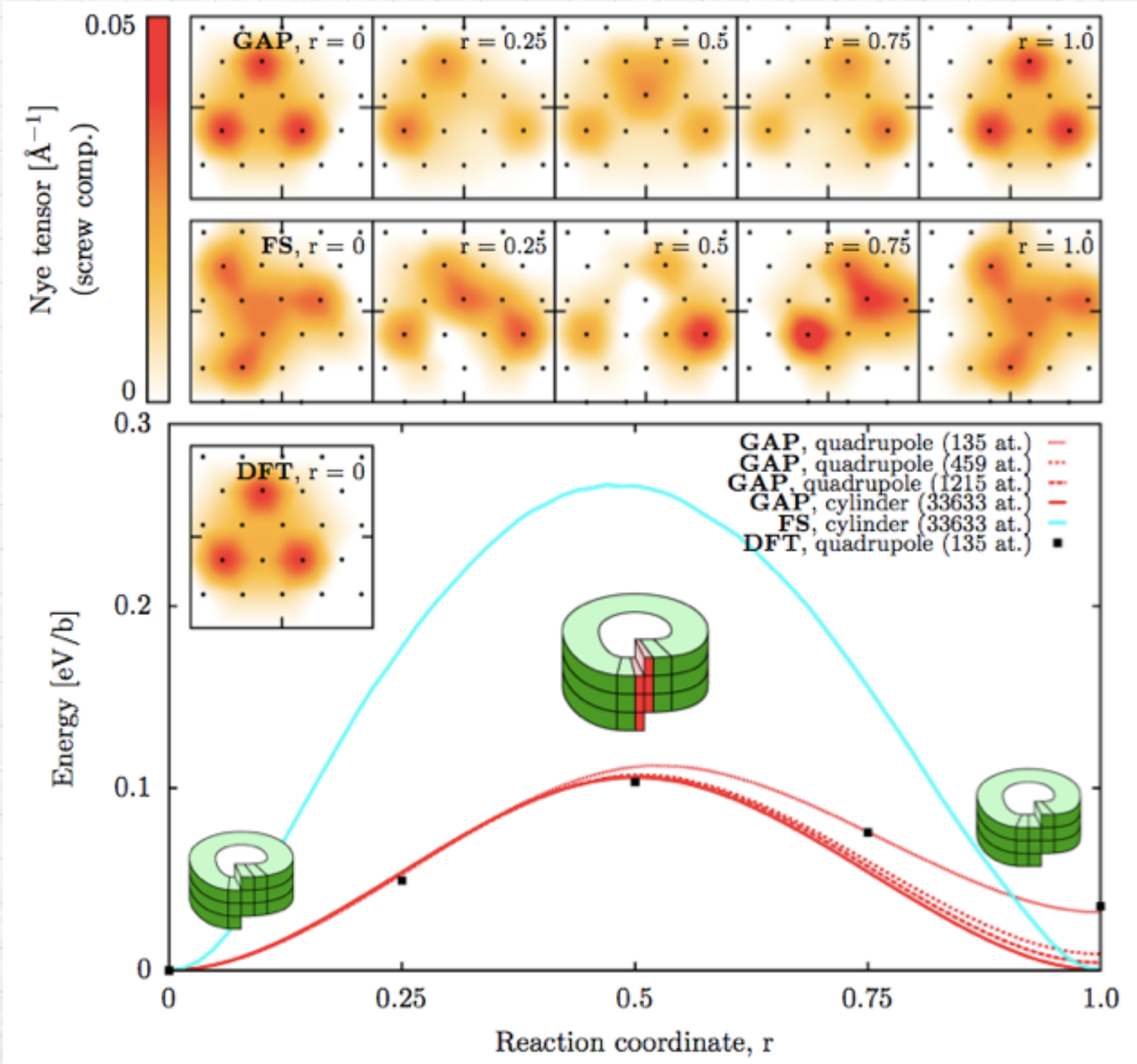
[a] Time on a single CPU core of Intel Xeon E5-2670 2.6GHz, [b] RMS error, [c] formation energy error, [d] RMS error of Nye tensor over the 12 atoms nearest the dislocation core, cf. Figure 2.

IMPROVEMENT FOR TUNGSTEN

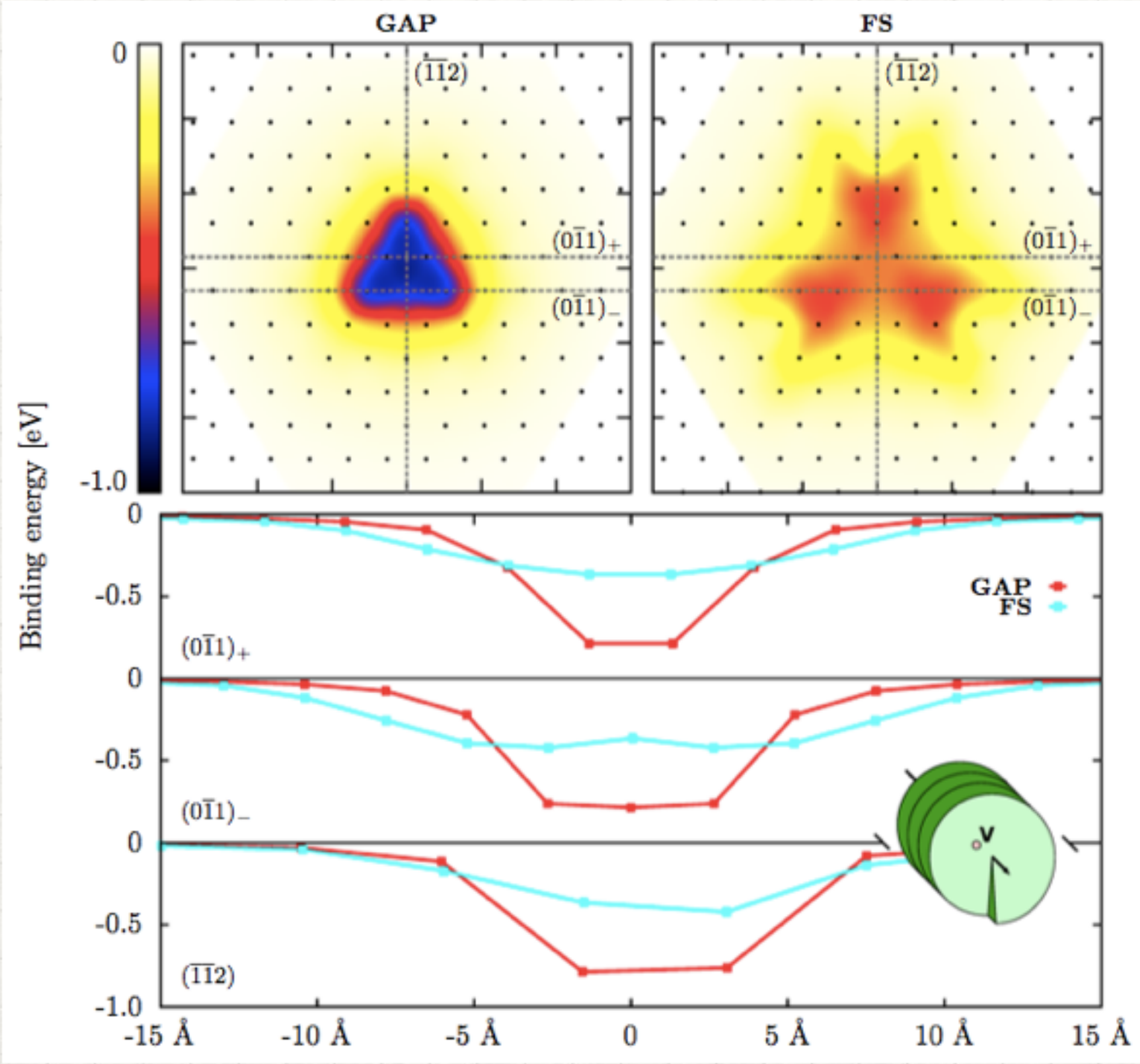
Errors relative to DFT



PEIERLS BARRIER FOR SCREW DISLOCATION GLIDE



VACANCY-DISLOCATION BINDING ENERGY

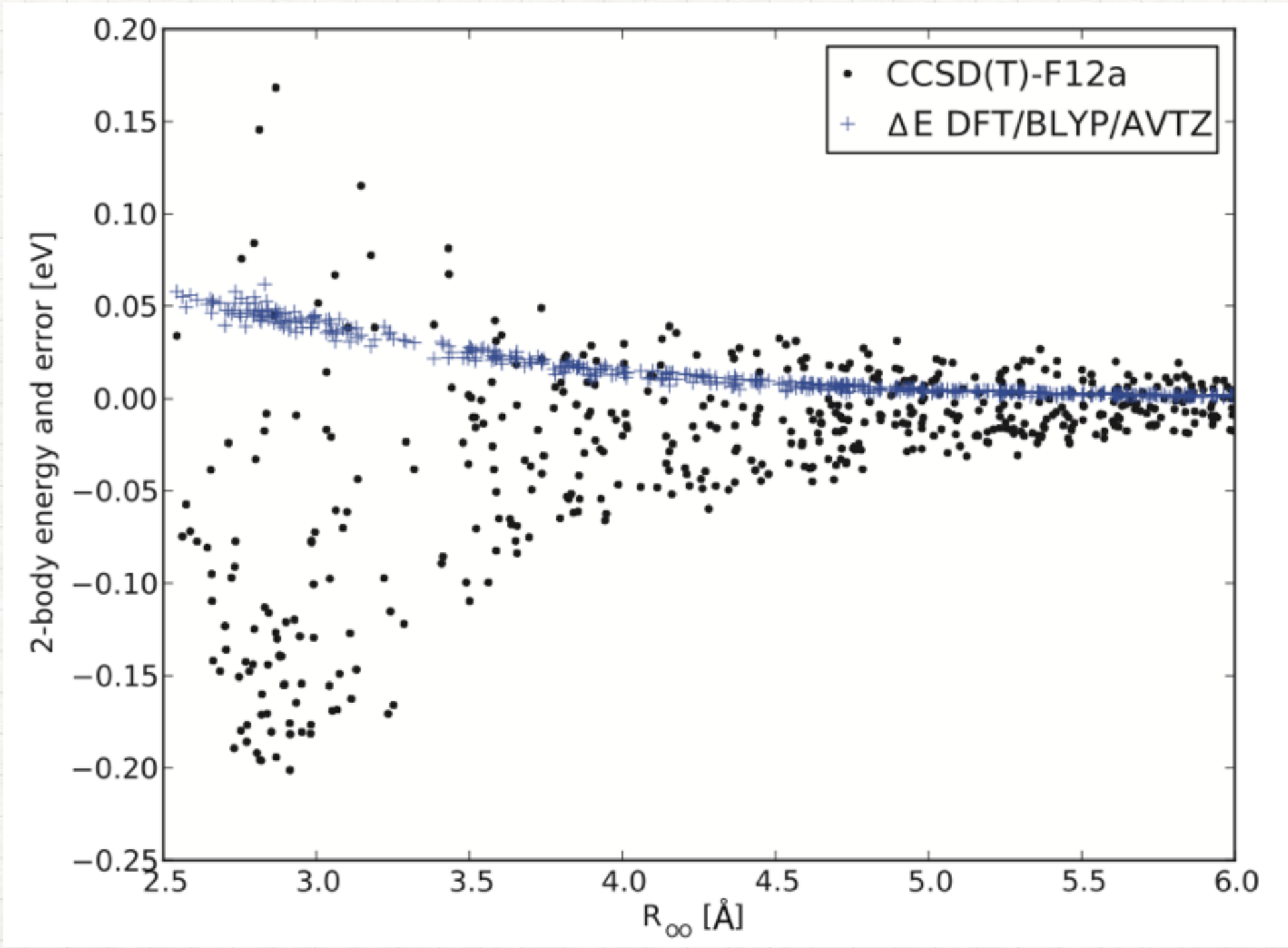


(~100,000 atoms in 3D simulation box)

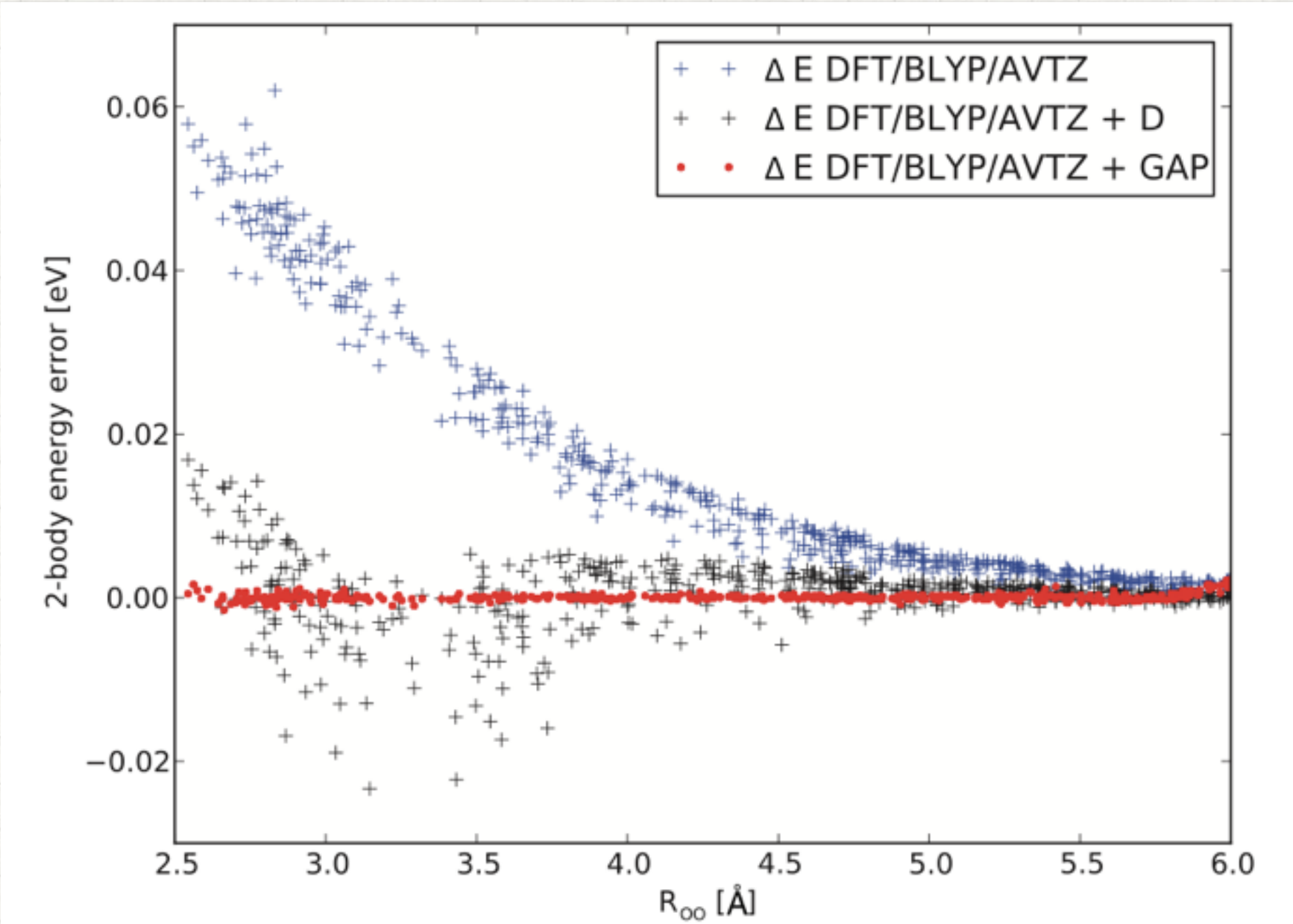
WATER MODEL

- >2-body terms described by DFT or
- >3-body terms described by SCME (with Thor Wikfeldt)
- All data generated by MD
- GAP1: DFT → MP2 AVTZ (energies and forces)
  - 6000 configs < 4.5 Å , AMOEBA bulk phase 300K
  - 1000 configs < 6.0 Å , AMOEBA bulk phase 300K
  - 2000 configs confined dimers, DFT, 4000K
- GAP2: MP2 → CCSD(T) AVDZ/f12 (only energies)
  - 800 configs < 4.5 Å, AMOEBA bulk phase 300K
  - 500 configs < 6.0 Å, AMOEBA bulk phase 300K

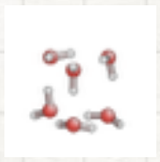
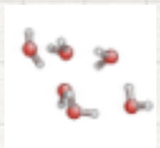
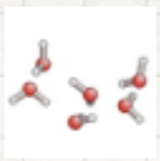
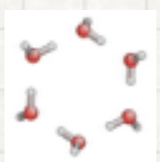
WATER DIMER ENERGY



WATER DIMER CORRECTIONS



WATER HEXAMER ENERGIES

	CCSD(T)	BLYP	DMC	Bowman	BLYP+GAP
	0.00	0.00	0.00	0.00	0.00
	0.24	-0.59	0.33	0.46	0.15
	0.70	-2.18	0.56	2.39	0.52
	1.69	-2.44	1.53	4.78	1.77

[kcal / mol]

(B3LYP optimised geometries from Tschumper)

GAP error < 0.03 kcal/mol / H_2O

ICE POLYMORPHS

Relative energies

[meV / H_2O]	Expt	DMC±5	PBE0+vdW^{TS}	BLYP+GAP±3
Ih	0	0	0	0
II	1	-4	6	-5
IX	4		3	-3
VIII	33	30	76	30
0				5

error < 0.15 kcal/mol / H_2O

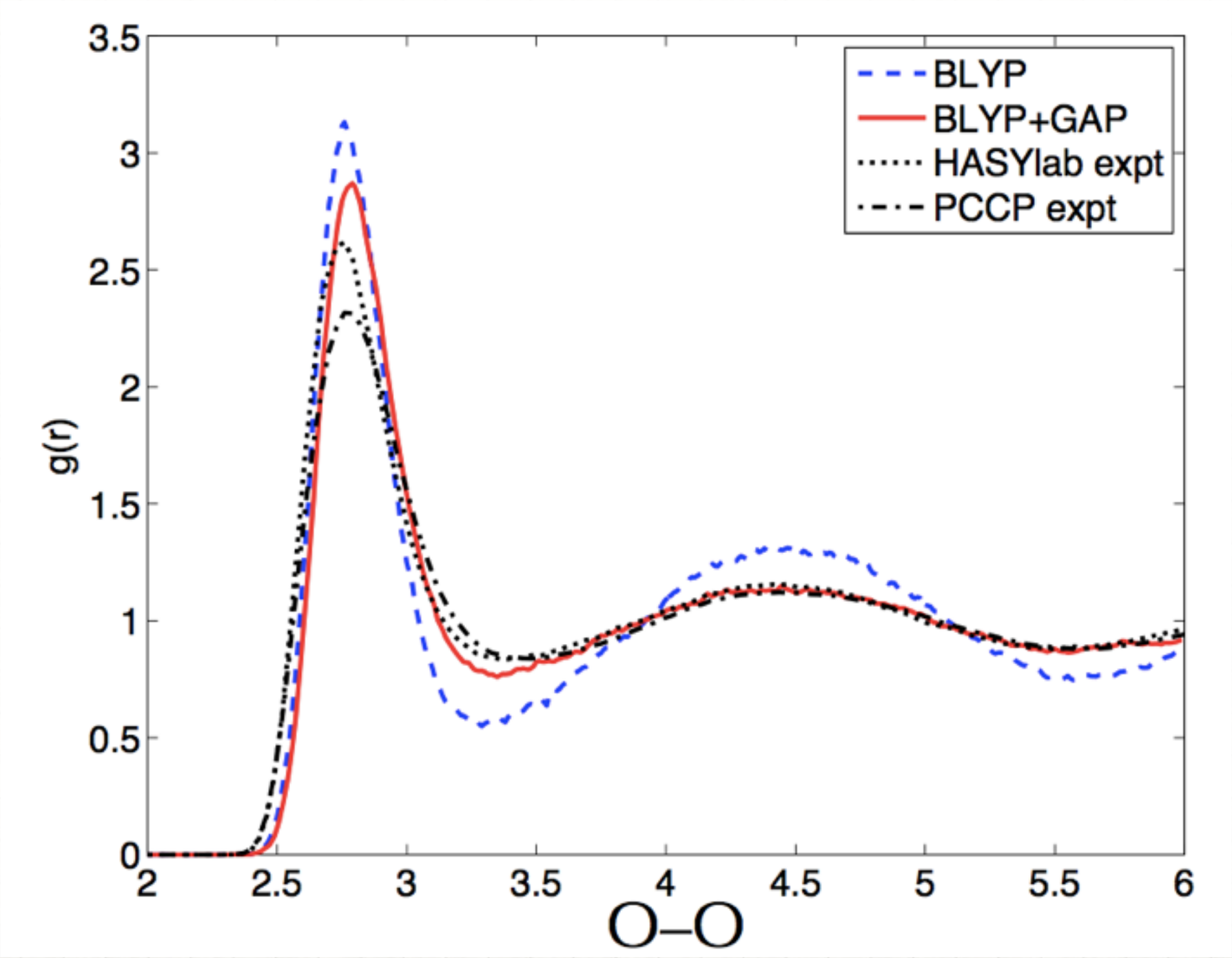
ICE POLYMORPHS

Relative volumes

[Å^3 / H_2O]	Expt	DMC	PBE0+vdW^{TS}	BLYP+GAP
Ih	0	0	0	0
II	-7.1	-7.0	-6.2	-6.7
IX	-6.4		-6.0	-6.5
VIII	-12.0	-12.2	-10.2	-11.9
0				1.0

error < 5%

LIQUID RDF


BLYP
BLYP+GAP
HASYlab expt
PCCP expt
g(r)
O–O
3.5
3
2.5
2
1.5
1
0.5
0
2
2.5
3
3.5
4
4.5
5
5.5
6

STRUCTURE MATCHING

Work with Michele Ceriotti and Sandip De

- goal: compare structures - molecules or bulk
- use local, atomic descriptors/kernels to create a global kernel (SOAP)
- compute covariance matrix for all atom pairings
- build global kernel
- use “kit” of atoms to compare different number of atoms

$$C_{ij}(\mathrm{A},\mathrm{B}) = k(\mathscr{X}_i^{\mathrm{A}},\mathscr{X}_j^{\mathrm{B}})$$

molecules

atoms in molecules

AVERAGE STRUCTURAL KERNEL

$$\bar{K}(\mathrm{A},\mathrm{B}) = \frac{1}{N^2}\sum_{ij} C_{ij}(\mathrm{A},\mathrm{B})$$
$$= \left[\frac{1}{N}\sum_{i}\mathbf{p}\left(\mathscr{X}_i^{\mathrm{A}}\right)\right]\cdot\left[\frac{1}{N}\sum_{j}\mathbf{p}\left(\mathscr{X}_j^{\mathrm{B}}\right)\right]$$

- cheap to compute
- equivalent to total energy kernel
- not sensitive

BEST MATCH STRUCTURAL KERNEL

$$\hat{K}(\mathrm{A},\mathrm{B}) = \frac{1}{N}\max_{\pi}\sum_{i} C_{i\pi_i}(\mathrm{A},\mathrm{B})$$

- find optimal permutations of the atoms (Hungarian algorithm)
- positive definiteness not guaranteed (might not be used as a distance metric)
- may be discontinuous

REGURALISED ENTROPY MATCH KERNEL

- best-match kernel

$$\hat{K}(\mathrm{A},\mathrm{B}) = \max_{\mathbf{P}\in\mathscr{U}(N,N)} \sum_{ij} C_{ij}(\mathrm{A},\mathrm{B}) P_{ij}$$

- **P** is a doubly stochastic matrix

- reformulate the problem to maximise information entropy for **P**

- use regularisation

$$\hat{K}^{\gamma}(\mathrm{A},\mathrm{B}) = \mathrm{Tr}\mathbf{P}^{\gamma}\mathbf{C}(\mathrm{A},\mathrm{B})$$

- order N^2 if using the Sinkhorn algorithm, efficient

$$\mathbf{P}^{\gamma} = \arg\min_{\mathbf{P}\in\mathscr{U}(N,N)} \sum_{ij} P_{ij}\left(1 - C_{ij} + \gamma \ln P_{ij}\right)$$

$$E(\mathbf{P}) = -\sum_{ij} P_{ij} \ln P_{ij}.\ \mathbf{P}^{\gamma}$$

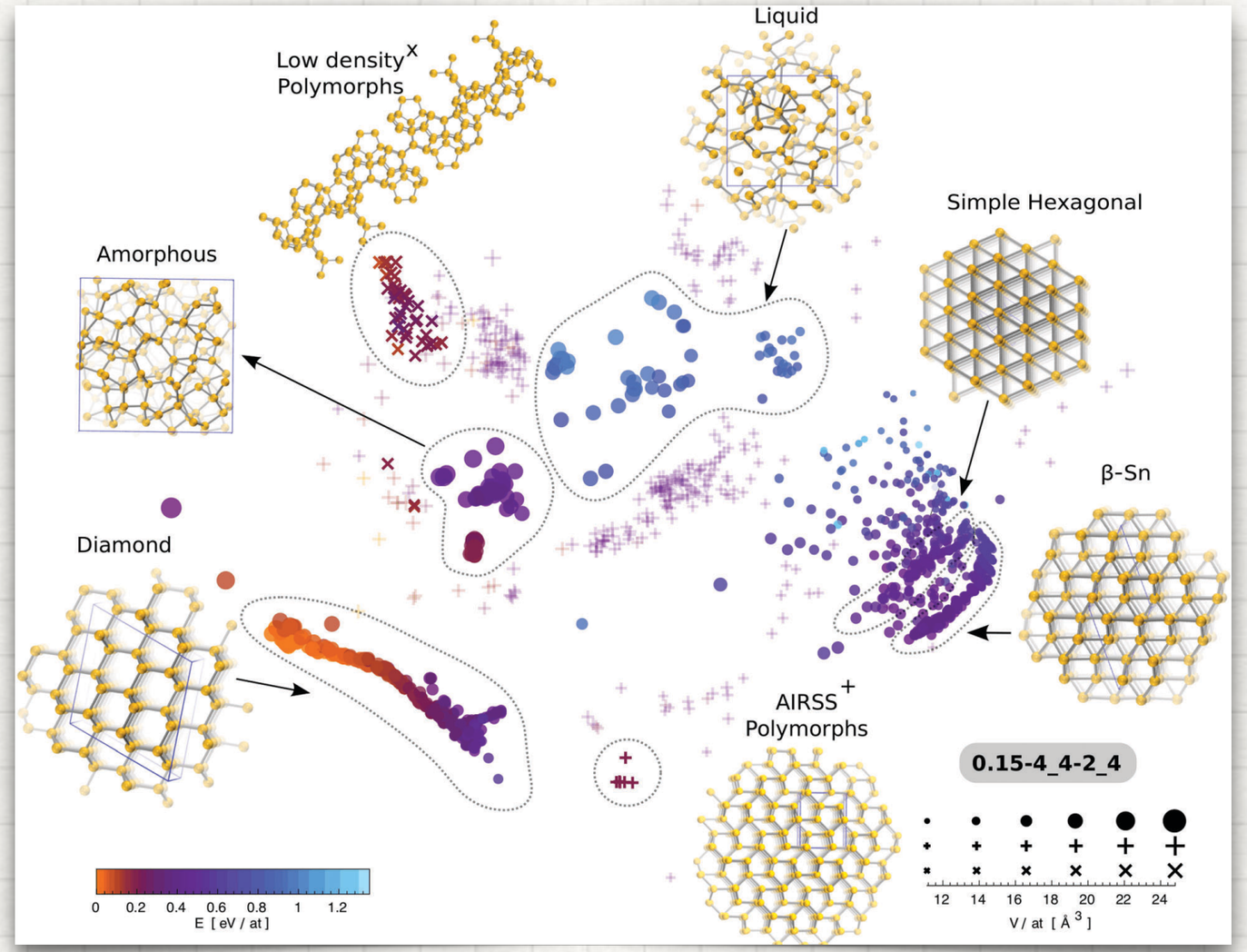
- γ in the limits of 0 and ∞ defaults to the best-match and average kernel

PERMUTATIONAL STRUCTURAL KERNEL

$$\breve{K}(\mathrm{A},\mathrm{B}) = \frac{1}{N!}\sum_{\pi}\prod_{i} C_{i\pi_i}(\mathrm{A},\mathrm{B}) = \mathrm{perm}\mathbf{C}(\mathrm{A},\mathrm{B})$$

- sum over all permutations
- combinatorial time
- limited use

LANDSCAPE OF SILICON POLYMORPHS

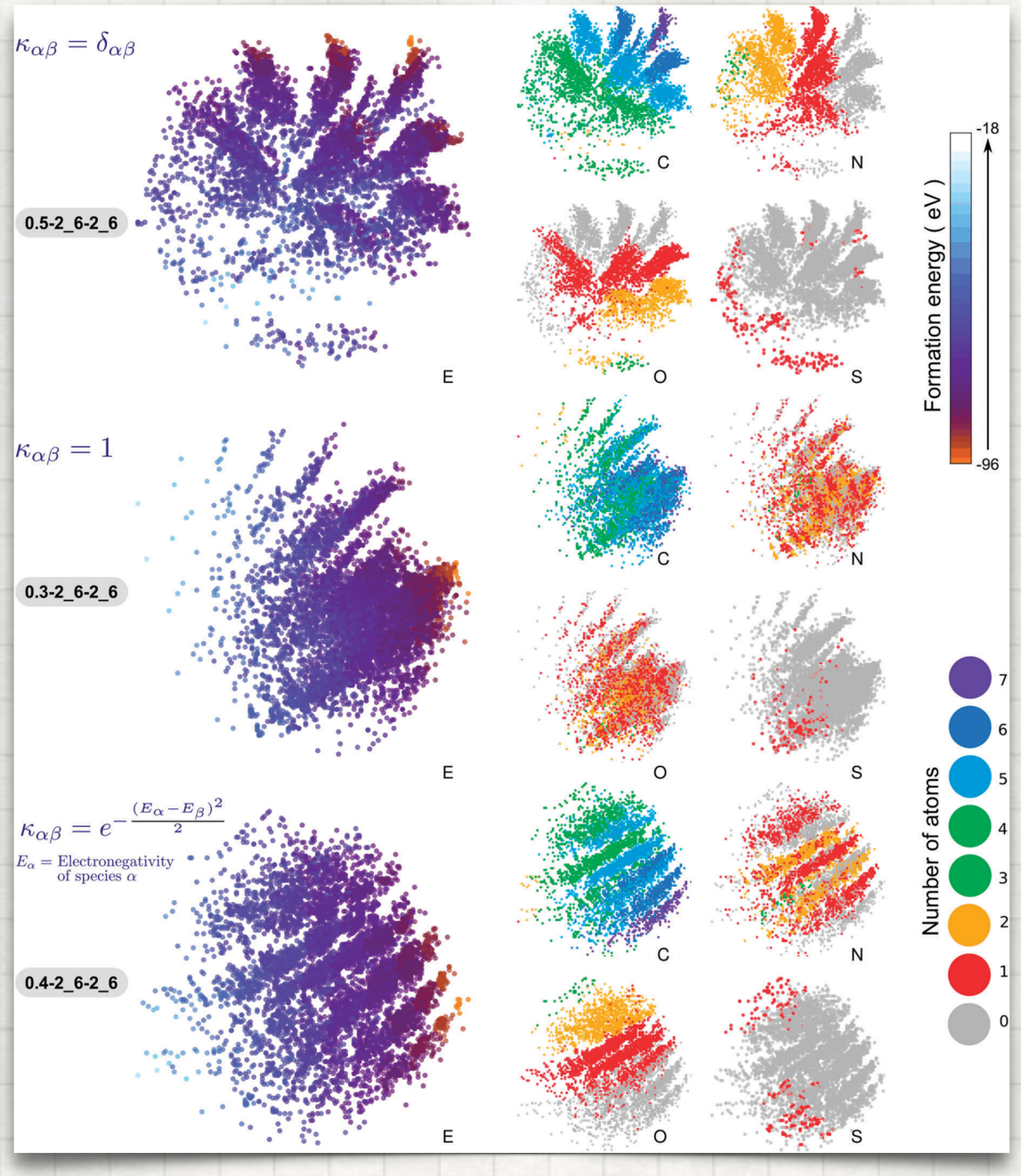


- sketch map (see Ceriotti's other work) of silicon structures
- best-match kernel
- 5 Å cutoff

MULTISPECIES APPLICATION

$$\tilde{k}(\mathscr{X},\mathscr{X}') = \int \mathrm{d}\hat{R} \left| \int \mathrm{d}\mathbf{r} \sum_{\alpha\alpha'} \kappa_{\alpha\alpha'} \rho^{\alpha}_{\mathscr{X}}(\mathbf{r}) \rho^{\alpha'}_{\mathscr{X}'}(\hat{R}\mathbf{r}) \right|^2$$

- sketch-map of organic molecules (QM7b)
- different
  - number of atoms
  - composition
  - bonding
  - conformation
- density overlaps of species



ACKNOWLEDGMENTS

- Dario Alfè
- Michele Ceriotti
- Gábor Csányi
- Sandip De
- Mike Gillan
- Ferenc Huszár
- Fred Manby
- Alan Nichol
- Mike Payne
- Peter Pinski
- Thor Wikfeldt

The Leverhulme Trust

UNIVERSITY OF CAMBRIDGE
Department of Engineering

Magdalene College
Department of Engineering
The Leverhulme Trust
The Isaac Newton Trust