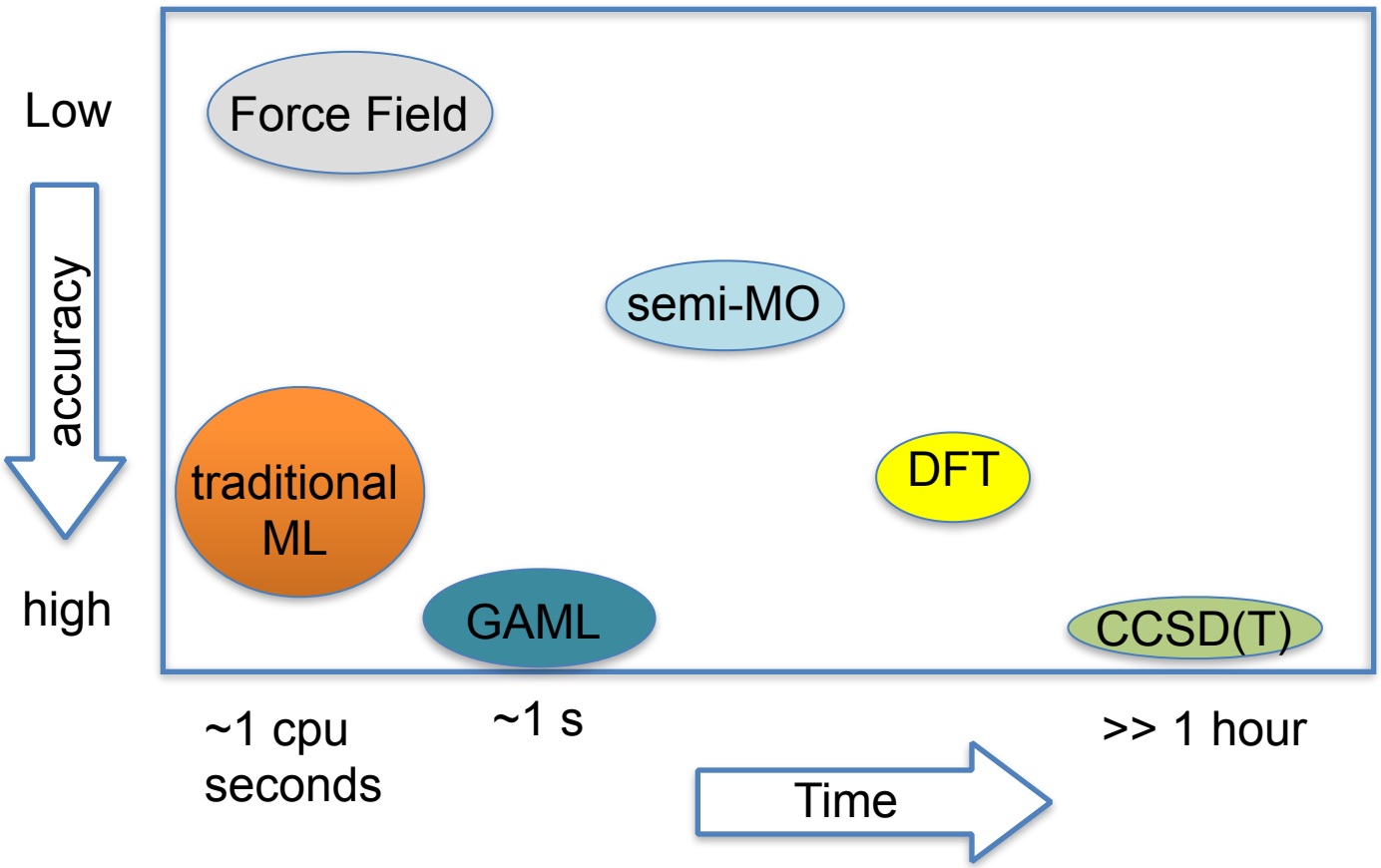
More efficient representations of compounds for machine learning models

Bing Huang and Anatole von Lilienfeld
Institute of Physical Chemistry and National Centre for Computational Design and Discovery of Novel Materials (MARVEL)
Department of Chemistry
University of Basel
Switzerland

UNI BASEL

Breaking the hex!

$$\hat{H}\psi = E\psi$$



Machine Learning - basics

feature abstraction

- Given data set {$\mathbf{X}_0$; **Y**}, learn f: **X** -> **Y** and then infer for new $\mathbf{X}_0$'

training test

for molecules, $\mathbf{X}_0$: {Z, R}, **Y**: E

- **Kernel ridge regression**

$k(\mathbf{X}_i, \mathbf{X}_j)$ **covariance**

$$Y_i^{\text{est}}(\mathbf{X}_i) = \sum_j \alpha_j \boxed{\exp\left(-\frac{d(\mathbf{X}_j, \mathbf{X}_i)}{\sigma}\right)} + b$$

length-scale of the data set

$$\min_\alpha \sum_i (Y_i^{\text{est}}(\mathbf{X}_i) - Y_i)^2 + \lambda \sum_i \alpha_i^2$$

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

noise-level

N parameters to be regressed for N molecules
+ 2 global parameters

M. Rupp, *et al.*, *PRL*, 2012

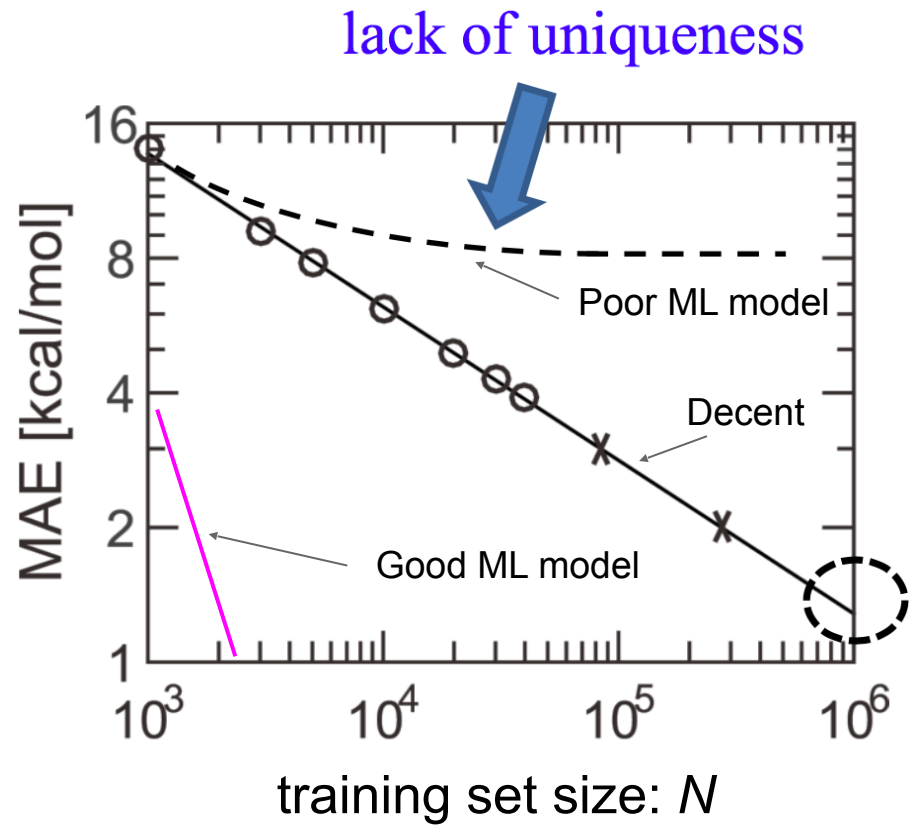
Machine Learning - basics

- More training data, better results for proper **X** (refer to **M** hereafter)

  at large *N*

  $$\log(\text{Error}) = a - b\log(N)$$

- representation (**M**) central to ML



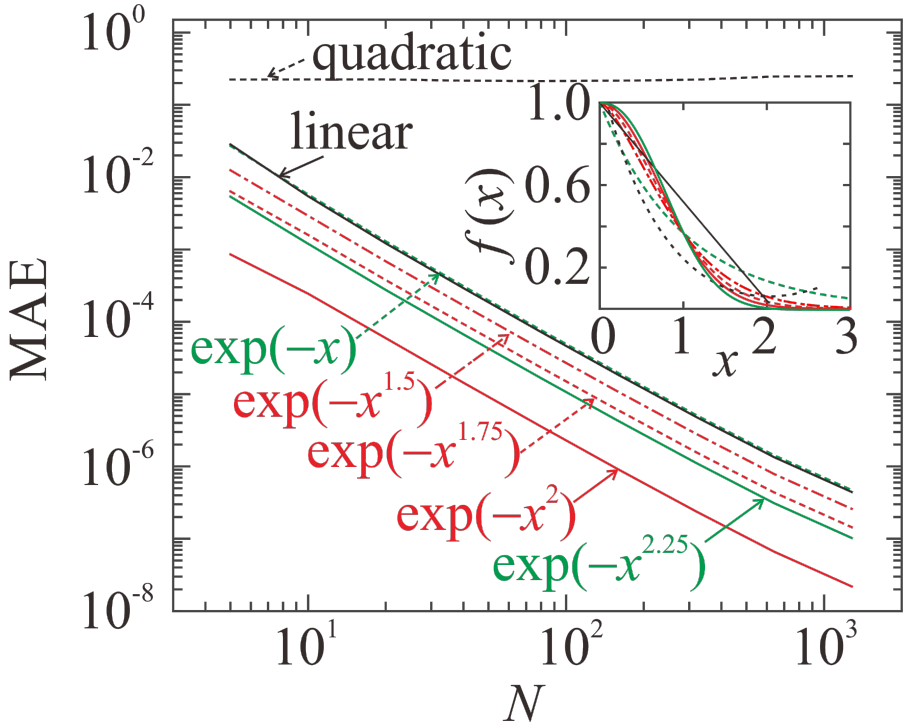
Huang, OAvL, *JCP Comm* (2016); arxiv.org/abs/1608.06194

Learning a simple 1-D function

$$f(x) = \exp(-x^2)$$

$$f^{\mathrm{est}}(x) = \sum \alpha_i k(\underbrace{ax_i + b}_{M_i}, \underbrace{ax + b}_{M})$$

For KRR, f & a*f + b
as rpsts are identical



$$\log(\mathrm{Error}) = a - b\log(N)$$

target similarity

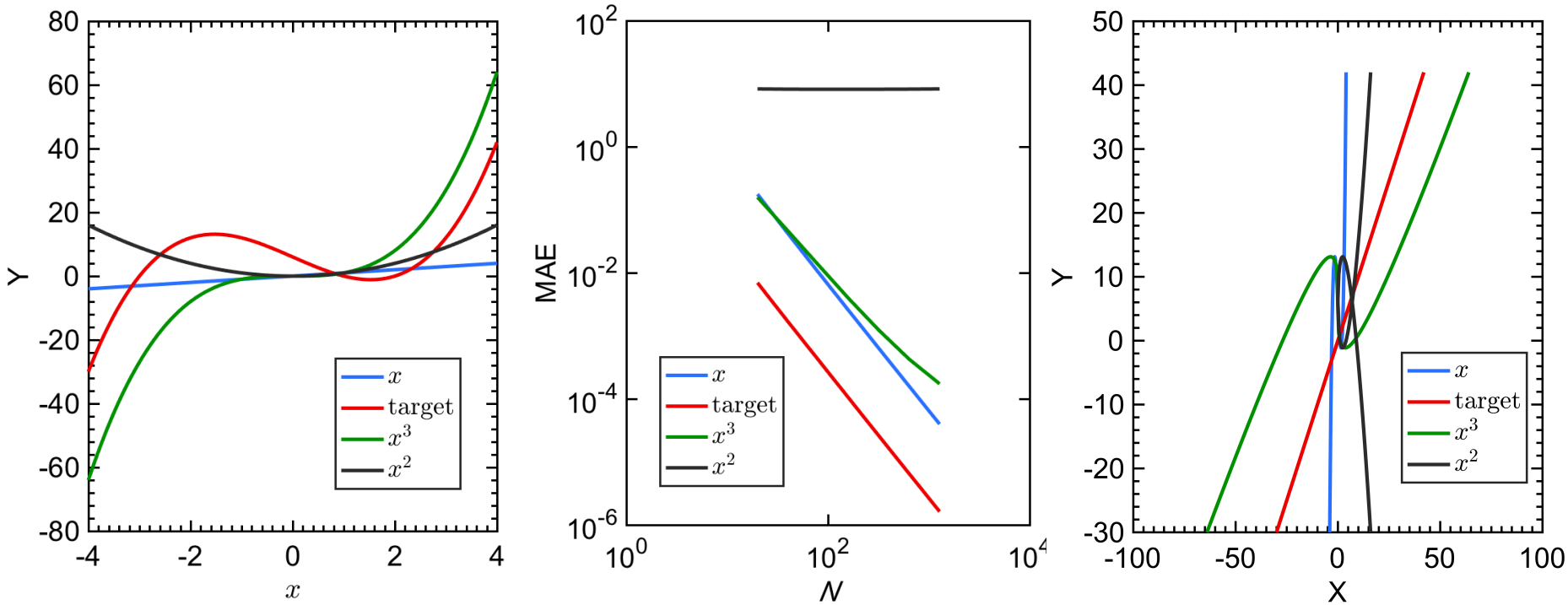
uniqueness

lack of uniqueness
→ absurd results
→ noise in training

OAvL *et al*, *IJQC* (2013)

Learning a "complicated 1-D function

target: Y = (*x*-1)(*x*-2)(*x*+3)

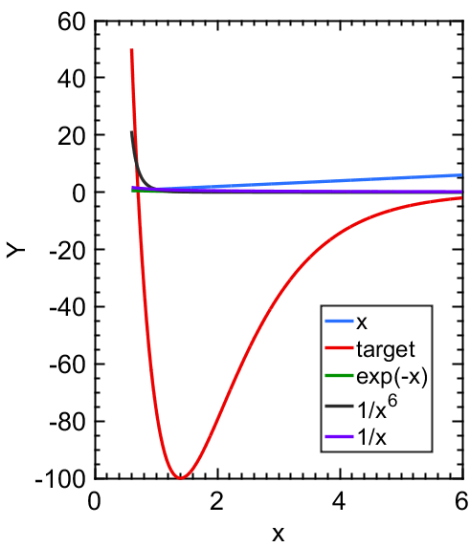


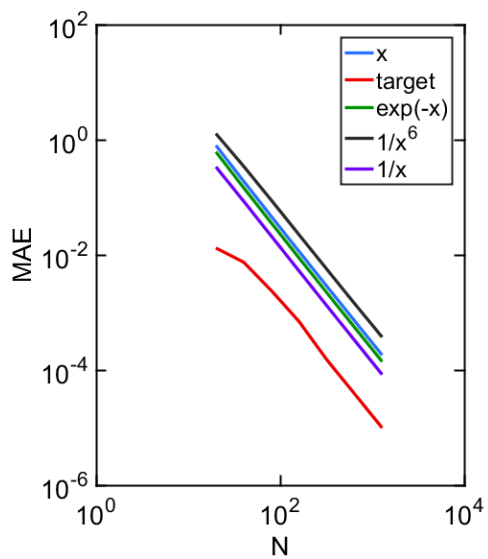
General guidelines for designing M

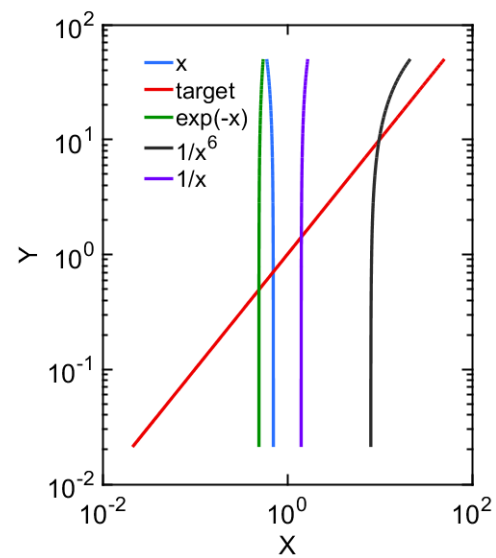
- in case you know *f* well (exact form unknown)
  use it as M
- otherwise you'd better know how *f* behaves
  use one monotonic part of *f*, refer it as *g*

best *g* minimizes $||g - f||_2$

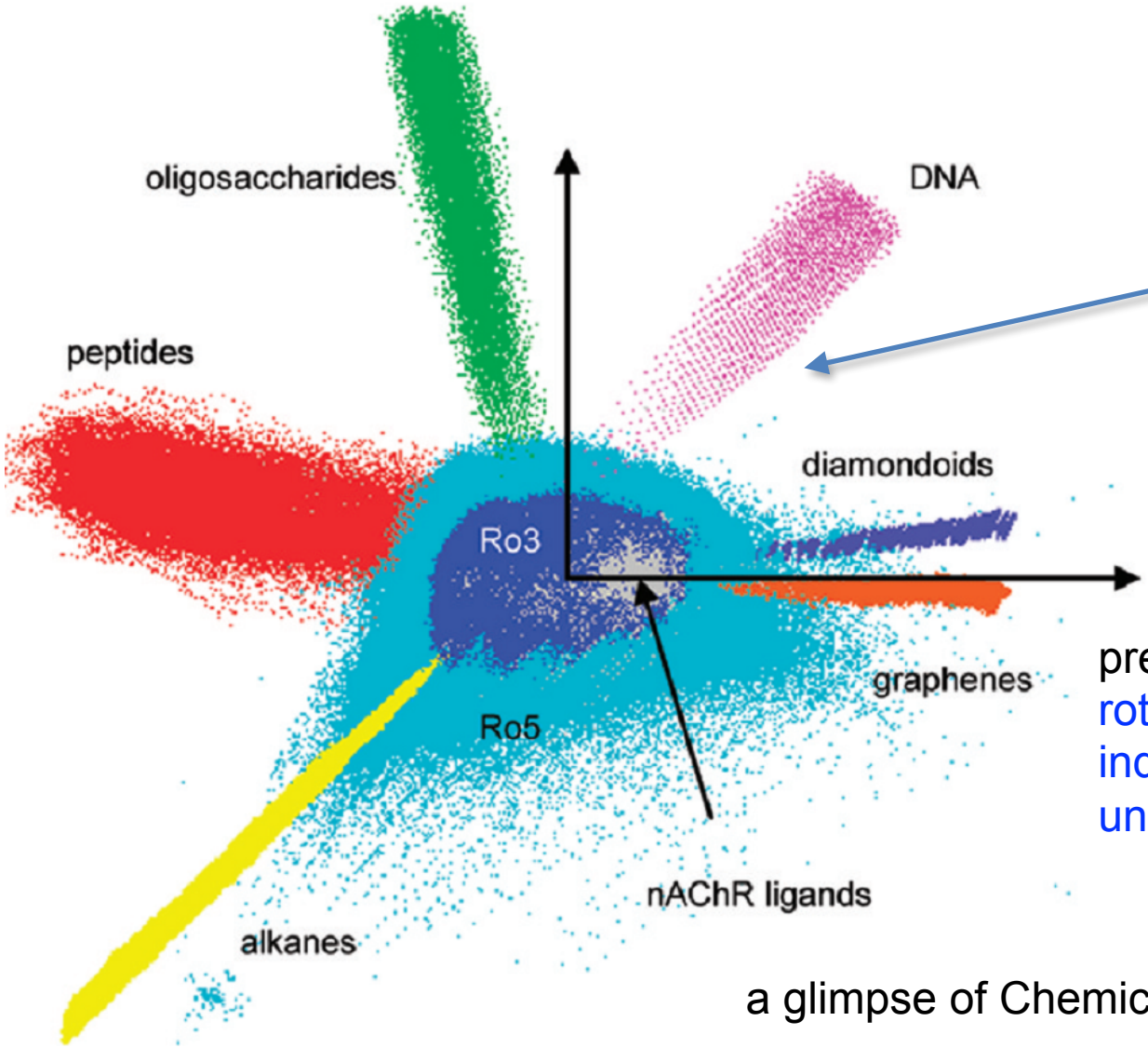
e.g., Morse potential $V(x) = -100*(2*\exp(-(x-1.4)) - \exp(-2(x-1.4)))$
Performance $V(x) > 1/x > \exp(-(x-1.4)) > x > 1/x^6 >> -(x-1.4)^2$

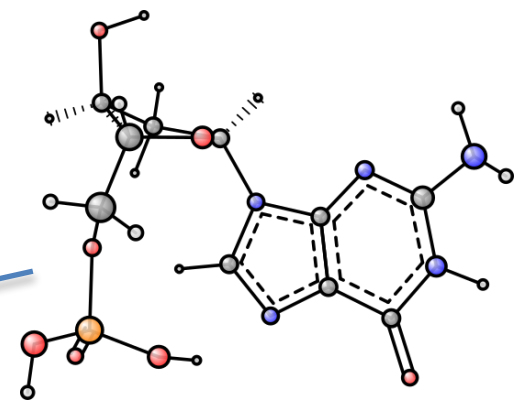






Representing molecules





X_0: $\{Z, R\}$

prerequisites:
rotation, translation invariant
index permutation invariant
unique (1 rpsts~2 molecule ⌫)

a glimpse of Chemical Compound Space (CCS)

J-L. Reymond *et al*, *ACS Chem. Neuro*. (2012)

Representing molecules

$\Psi/\rho/V$-based

atom/electron density

$\hat{H}\Psi = E\Psi$ straightforward !!

$E = E[\Psi] = E[\rho] = E[V\_ext]$

impractical

practical

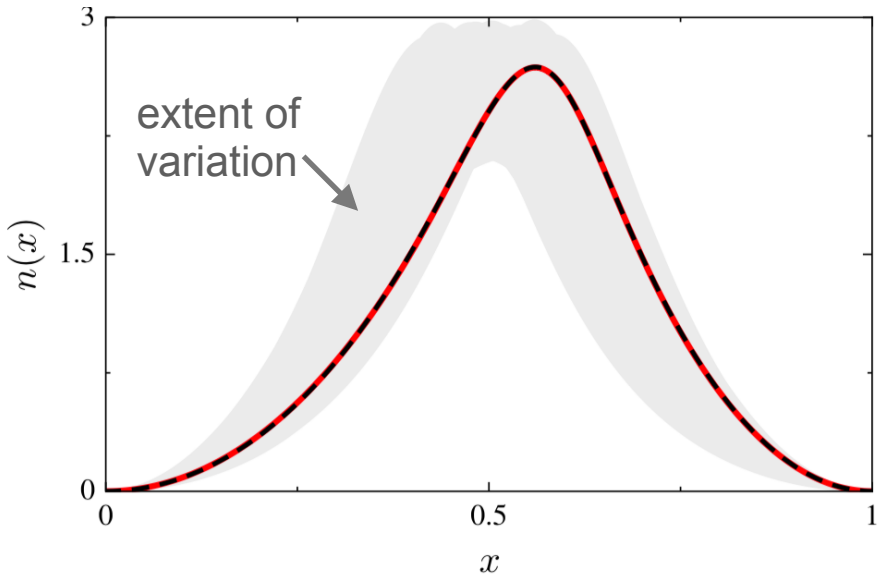
accessible

CCS: infinite dimension

projection to basis set

M: [c_1, c_2, ..., c_n]

Learning an 1-D functional

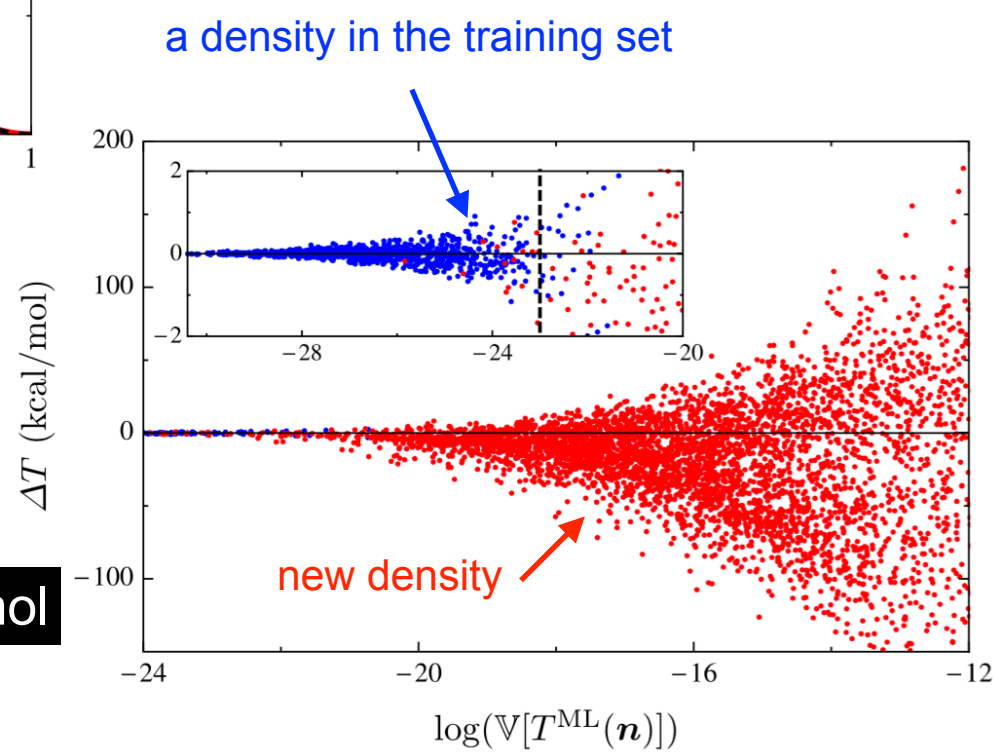


$$v(x) = -\sum_{i=1}^{3} a_i \exp[-(x-b_i)^2/(2c_i^2)].$$

noninteracting fermions in 1D

property: kinetic energy

MAE < 1.0 kcal/mol



J. C. Snyder, *et al*, *PRL* (2012)

fingerprint representations

$$P(\mathbf{r}) = \sum_I Z_I e^{-a|\mathbf{r}-\mathbf{R}_I|^2}$$

remove rotation dependence

projection to 1-D frequency domain & substitution

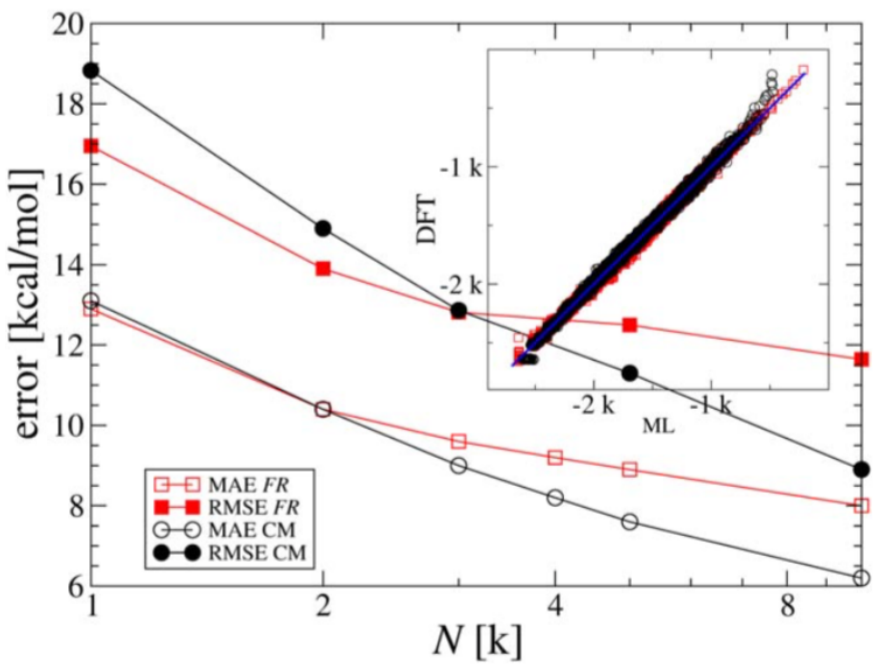
$$FR(r) = \sum_I Z_I^k \left(\cos\left[Z_I^m \sum_J Z_J^n e^{-b(r-r_{IJ})^2} \right] - 1 \right)$$

$$C(\eta) = \frac{1}{N} \sum_i^N \sum_{j \neq i} e^{-\left(\frac{r_{ij}}{\eta}\right)^2}$$

projection to 1-D η-grid

1-D discrete array

works good for Al_n systems



GDB-9 dataset

OAvL *et al*, *IJQC* (2015)

V. Botu, *et al*., *IJQC* (2015)

Representing molecules

why are fingerprint rpsts bad for molecules, but good for Al_n like systems?

$\Psi/\rho/V$-based

$||g - f||_2$ large for molecules, small for Al_n

Representing molecules

Ψ/ρ/*V*-based

many body expansion (MBE) of total energy

E-based

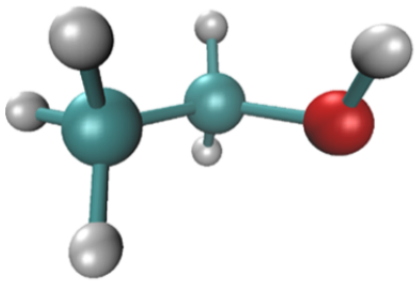
$$E(\{\mathbf{R}_I\}) = \sum_{I}^{[Z]} E^{(1)}(R_I) + \sum_{J>I}^{[Z]} E^{(2)}(R_{IJ}) + \sum_{K>J>I}^{[Z]} E^{(3)}(R_{IJ}, R_{IK}, R_{JK}) + \cdots$$

CCS: dimension is significantly reduced!!!

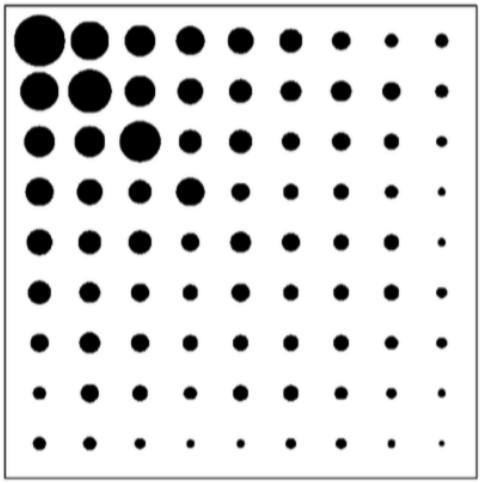
M: [{*E*(1)}, {*E*(12)}, {*E*(123)}, …]

BH, OAvL, *JCP comm.*, 2016

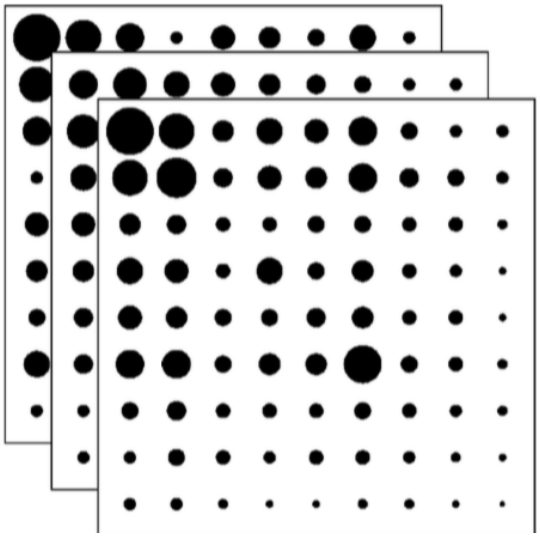
Coulomb matrix (CM)



$$C_{ij} = \begin{cases} 0.5\, Z_i^{2.4} & \forall i = j \\ \frac{Z_i Z_j}{|\mathbf{R}_i - \mathbf{R}_j|} & \forall i \neq j. \end{cases}$$



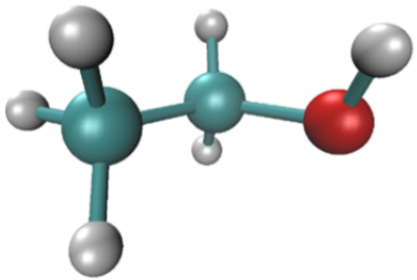
sorted CM



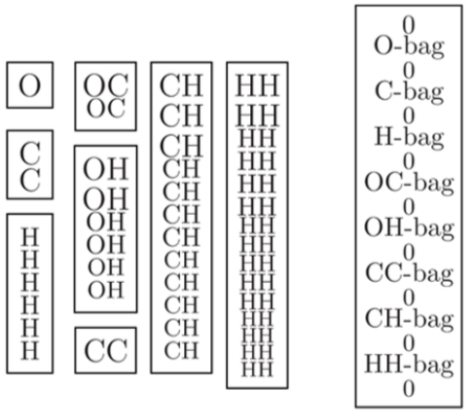
random CM

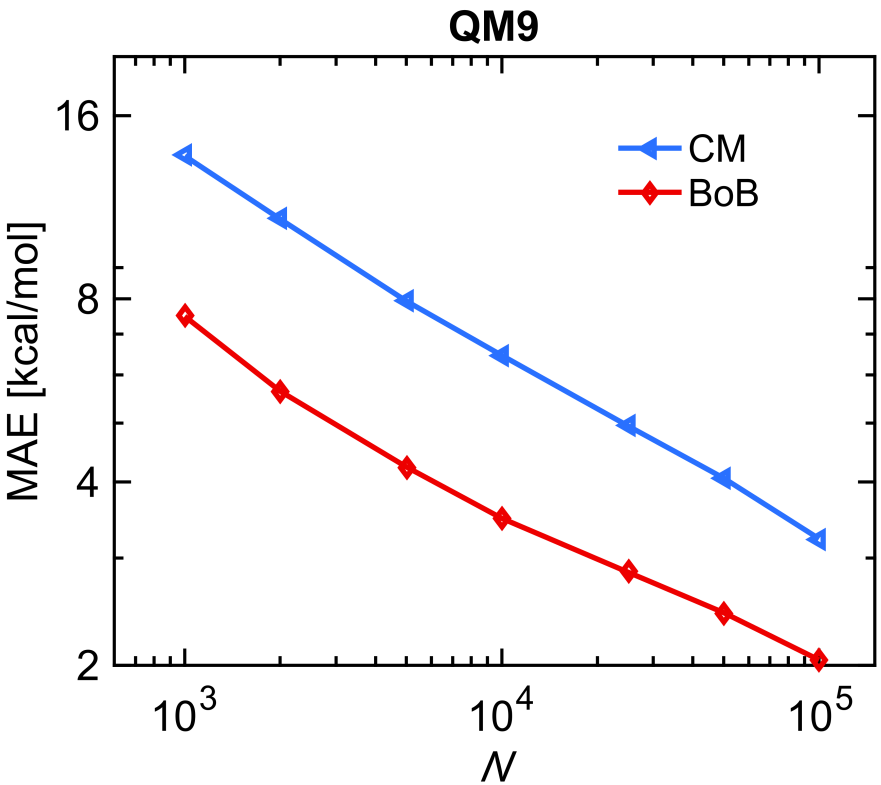
"CM", M. Rupp, *et al.*, PRL, 2012

Bag of Bonds (BoB)



	O	C	C	H	H	H	H	H	H
O	O	OC	OC	OH	OH	OH	OH	OH	OH
C	OC	C	CC	CH	CH	CH	CH	CH	CH
C	OC	CC	C	CH	CH	CH	CH	CH	CH
H	OH	CH	CH	H	HH	HH	HH	HH	HH
H	OH	CH	CH	HH	H	HH	HH	HH	HH
H	OH	CH	CH	HH	HH	H	HH	HH	HH
H	OH	CH	CH	HH	HH	HH	H	HH	HH
H	OH	CH	CH	HH	HH	HH	HH	H	HH
H	OH	CH	CH	HH	HH	HH	HH	HH	H

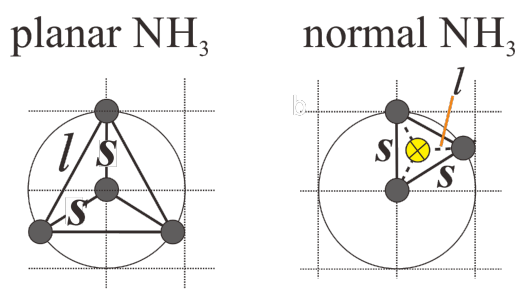




much better than CM, why??

K. Hansen, *et al.*, *JPCL*, 2015

non-uniqueness issue



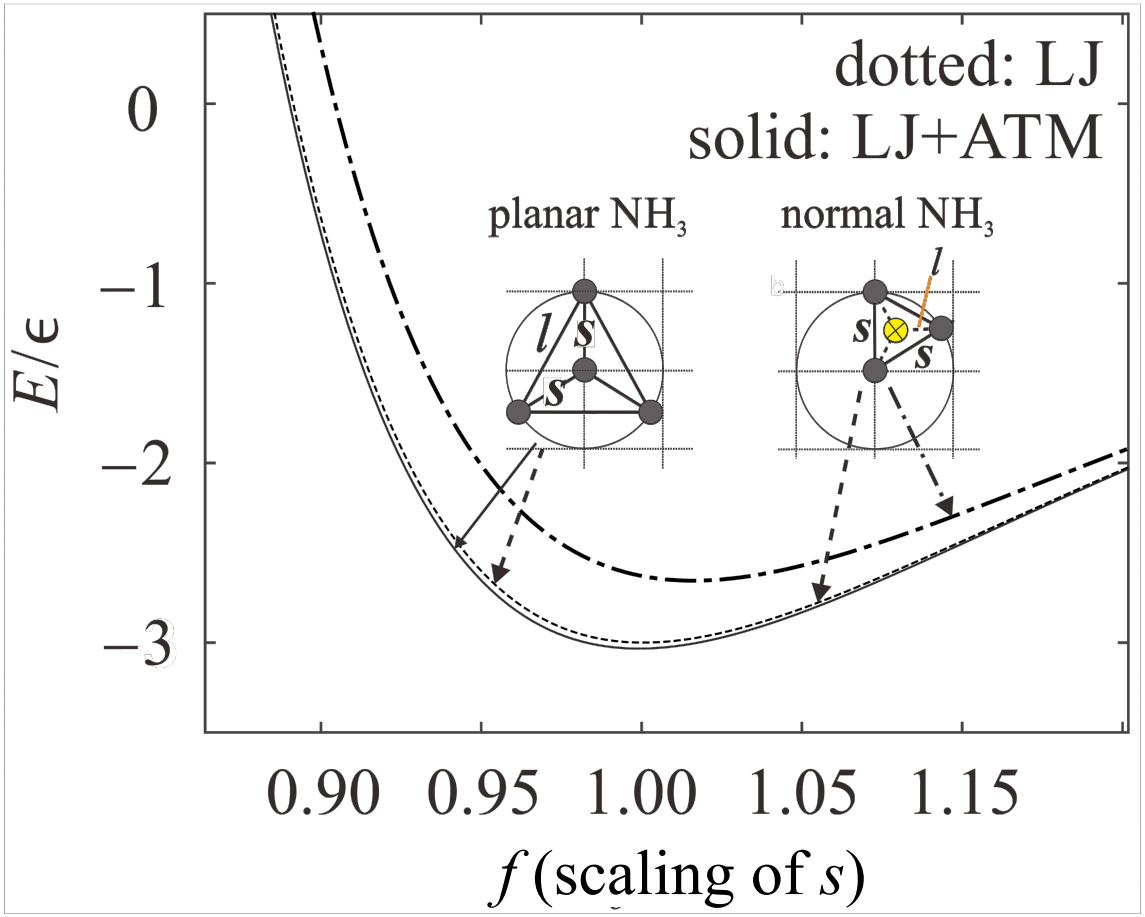
homometric molecules

same set of interatomic distance pairs

OAvL, *et al.*, IJQC, 2015

non-uniqueness issue

LJ: Lennard-Jones 2-body vdW potential
ATM: Axilrod-Teller-Muto 3-body vdW potential



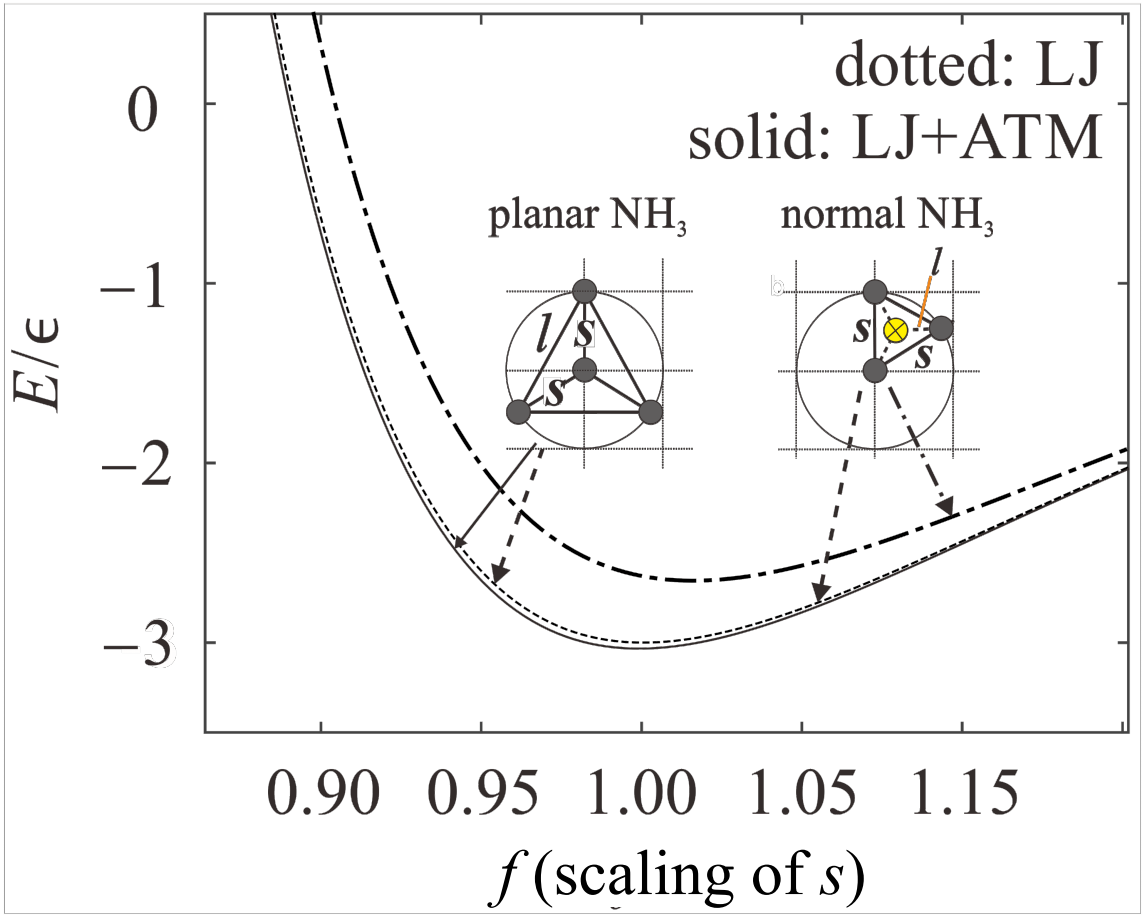
homometric molecules

same set of interatomic distance pairs

BH, OAvL, *JCP comm.*, 2016

non-uniqueness issue

LJ: Lennard-Jones 2-body vdW potential
ATM: Axilrod-Teller-Muto 3-body vdW potential



homometric molecules

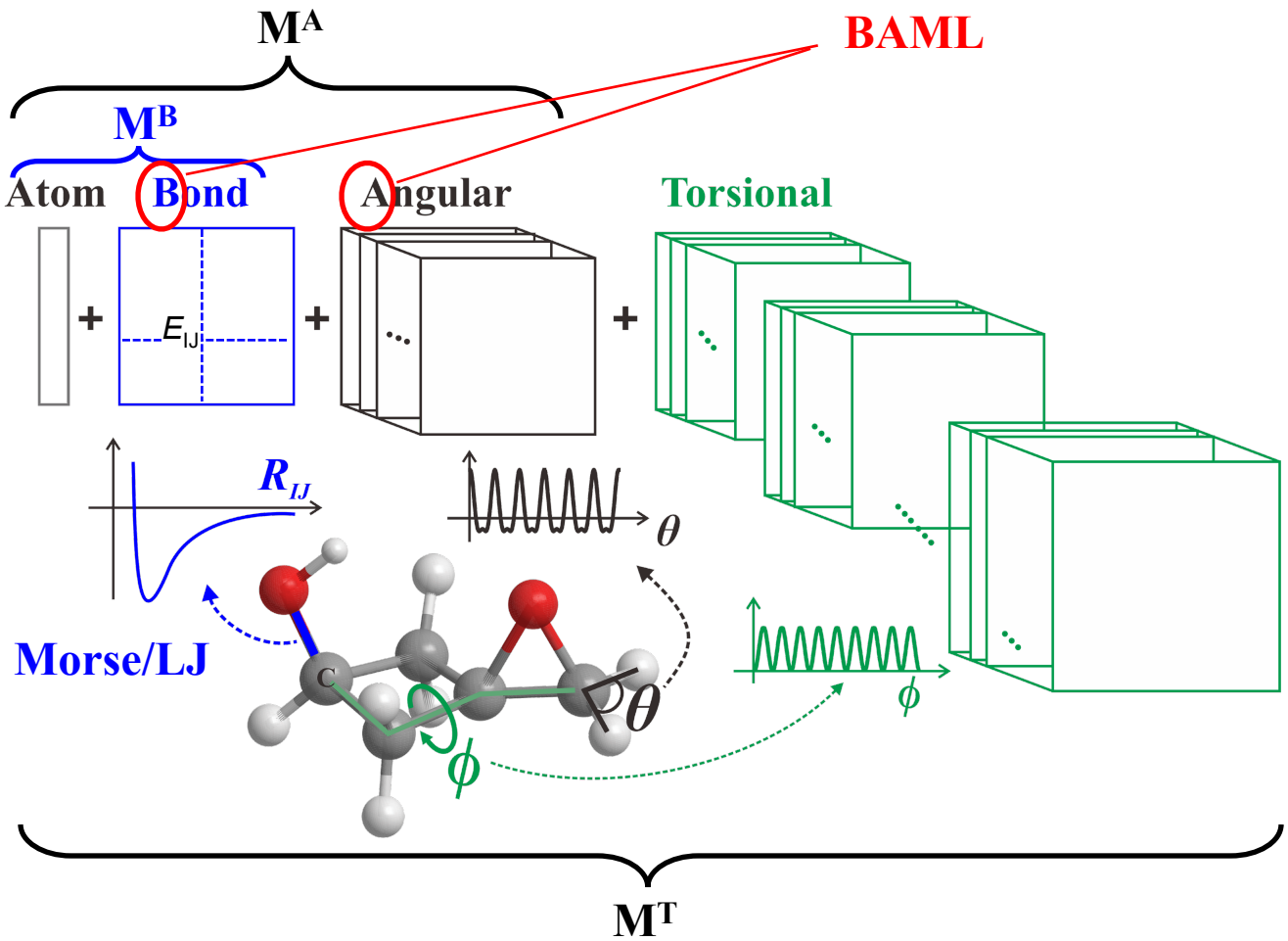
same set of interatomic distance pairs

2-body interaction is not enough!

BH, OAvL, *JCP comm.*, 2016

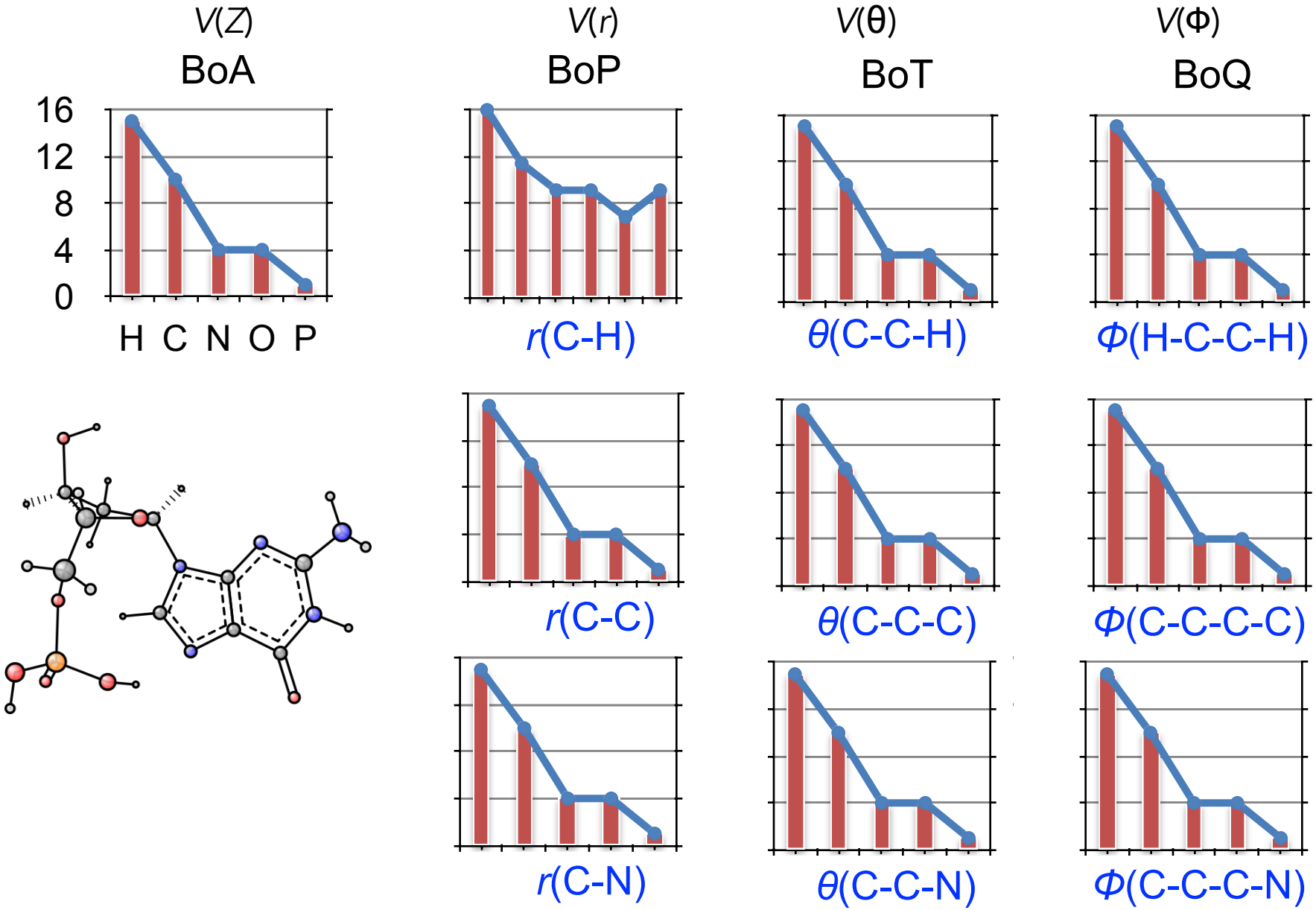
BAML

bags of Universal force field (UFF) contributions

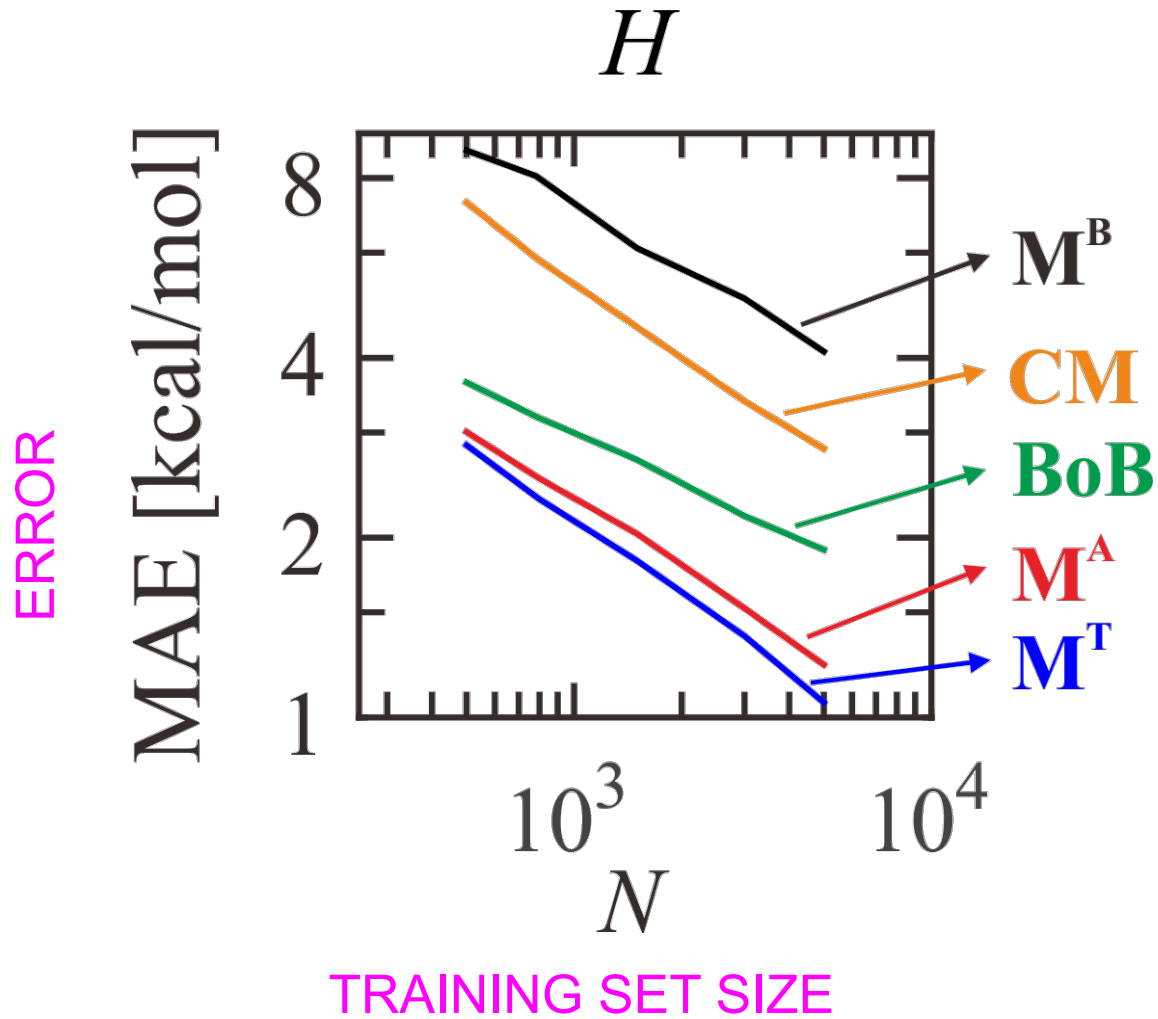


BH, OAvL, *JCP comm.*, 2016

BAML


V(Z)
BoA
16
12
8
4
0
H C N O P
V(r)
BoP
r(C-H)
V(θ)
BoT
θ(C-C-H)
V(Φ)
BoQ
Φ(H-C-C-H)
r(C-C)
θ(C-C-C)
Φ(C-C-C-C)
r(C-N)
θ(C-C-N)
Φ(C-C-C-N)

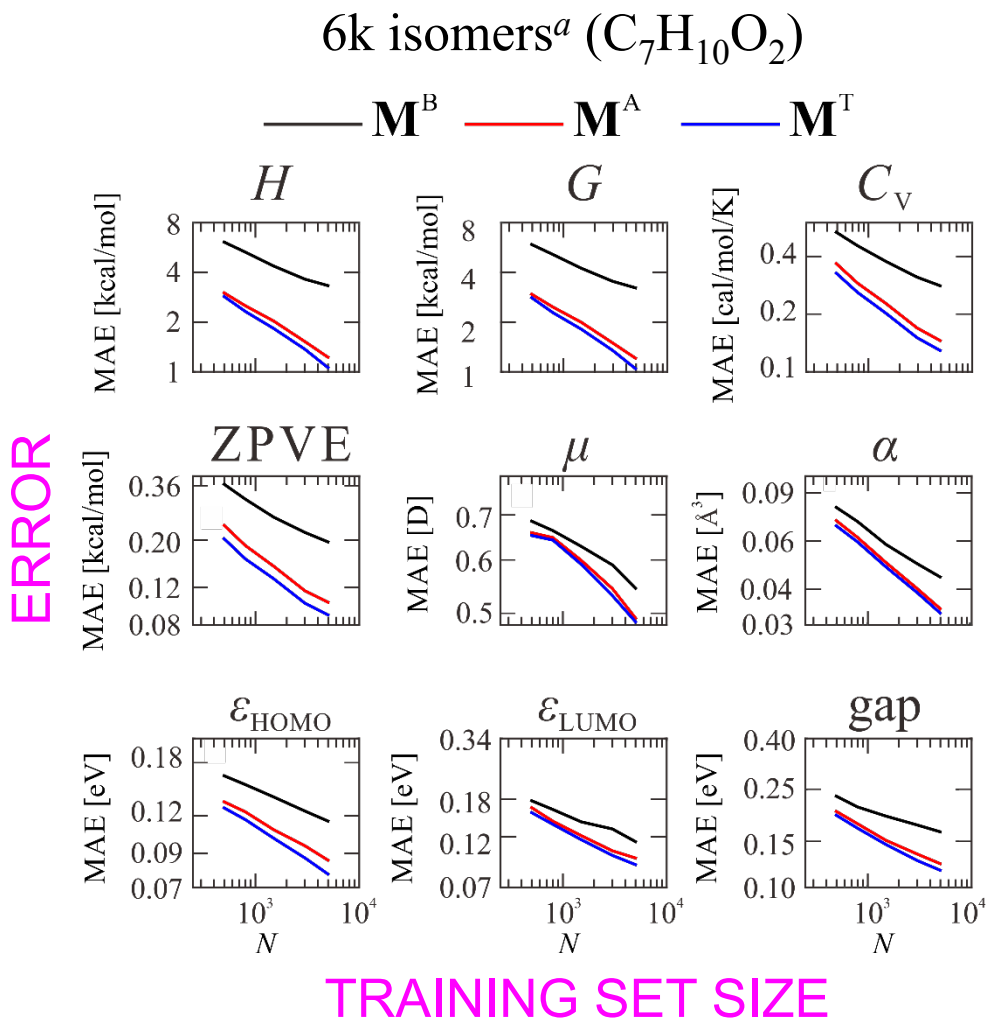
BAML



database: 6k isomers
($C_7H_{10}O_2$)

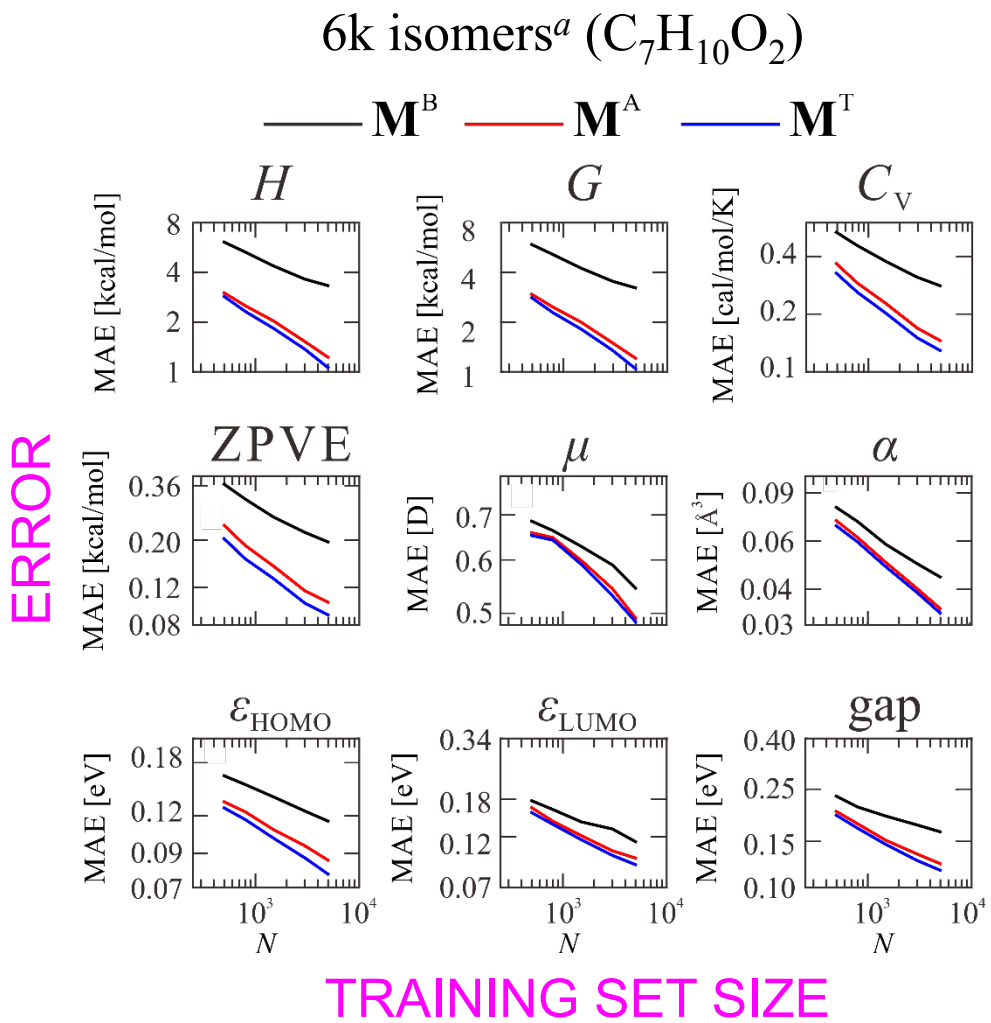
BH, OAvL, *JCP comm.*, 2016

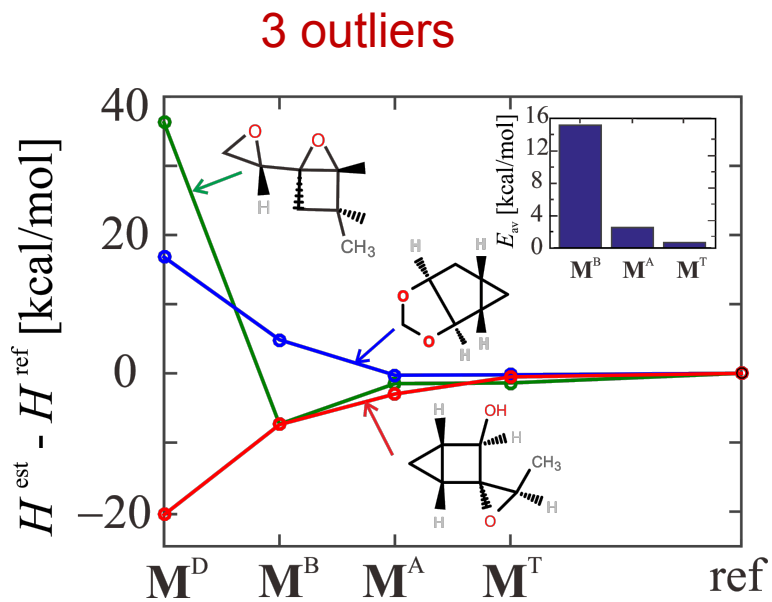
BAML



BH, OAvL, *JCP comm.*, 2016

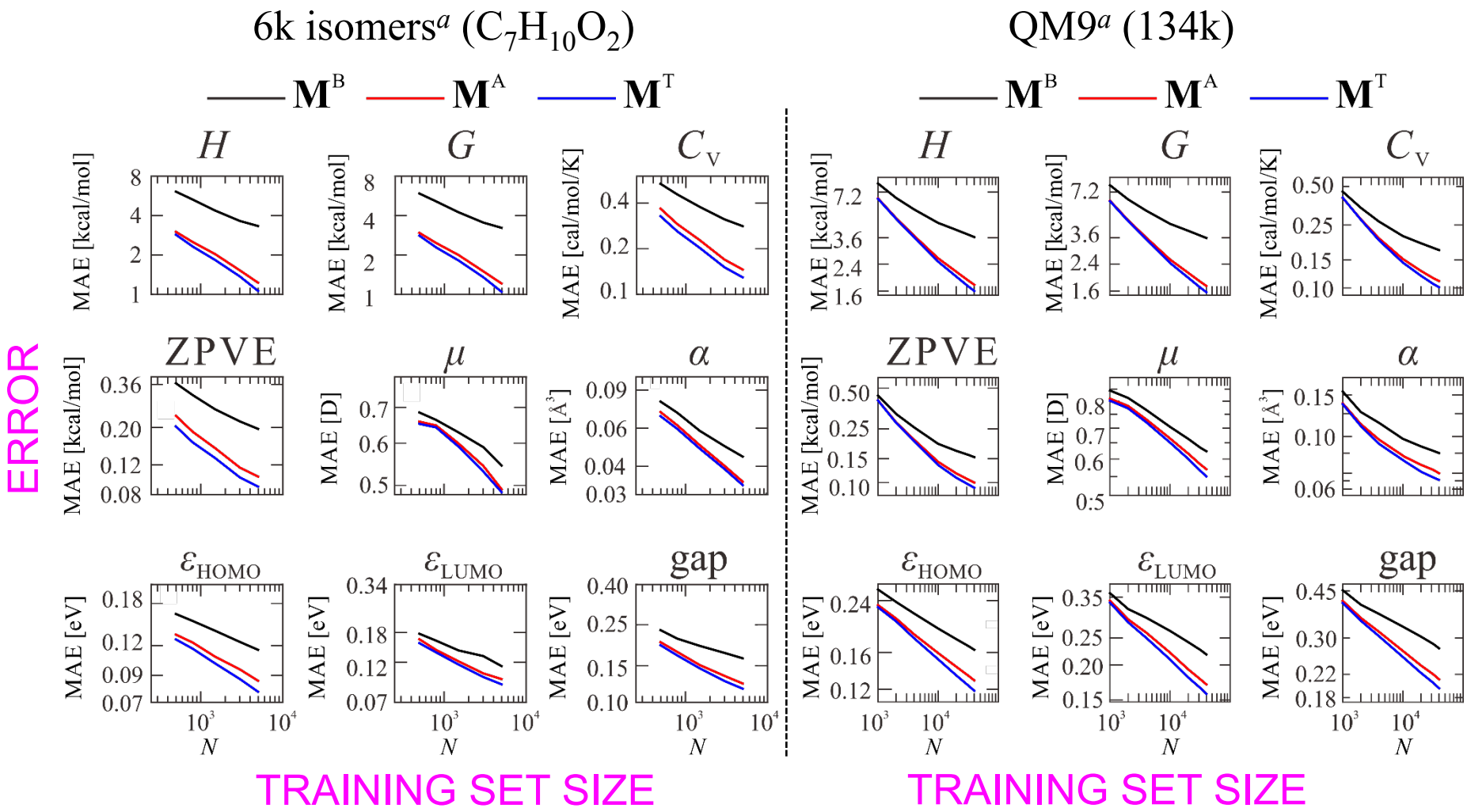
BAML





BH, OAvL, *JCP comm.*, 2016

BAML



BH, OAvL, *JCP comm.*, 2016

BAML

Comparison QM7b database (size: 7211)

MAE (5k out-of-sample)

	BAML	BoB	SOAP[a]	CM[b]	accuracy[b]
E (PBE0)/eV	0.05	0.08	0.04	0.16	0.15, 0.23, 0.09-0.22
α (PBE0)/ Å^3	0.07	0.09	0.05	0.11	0.05-0.27, 0.04-0.14
HOMO (GW)/eV	0.10	0.15	0.12	0.16	-
LUMO (GW)/eV	0.11	0.16	0.12	0.16	-
IP (ZINDO)/eV	0.15	0.20	0.19	0.17	0.20, 0.15
EA (ZINDO)/eV	0.07	0.17	0.13	0.11	0.16, 0.11
E_{1st}* (ZINDO)/eV	0.13	0.21	0.18	0.13	0.18, 0.21

[a] S. De, *et al.*, *PCCP*, 2016
[b] G. Montavon, *et al.*, *NJP,* 2013

BH, OAvL, *JCP comm.*, 2016

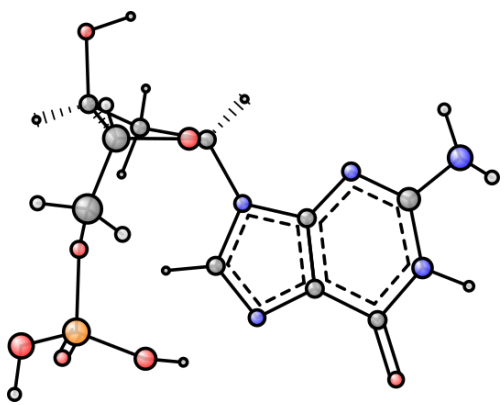
HDAD

$V(r) = r$

$V(\theta) = \theta$

$V(\Phi) = \Phi$

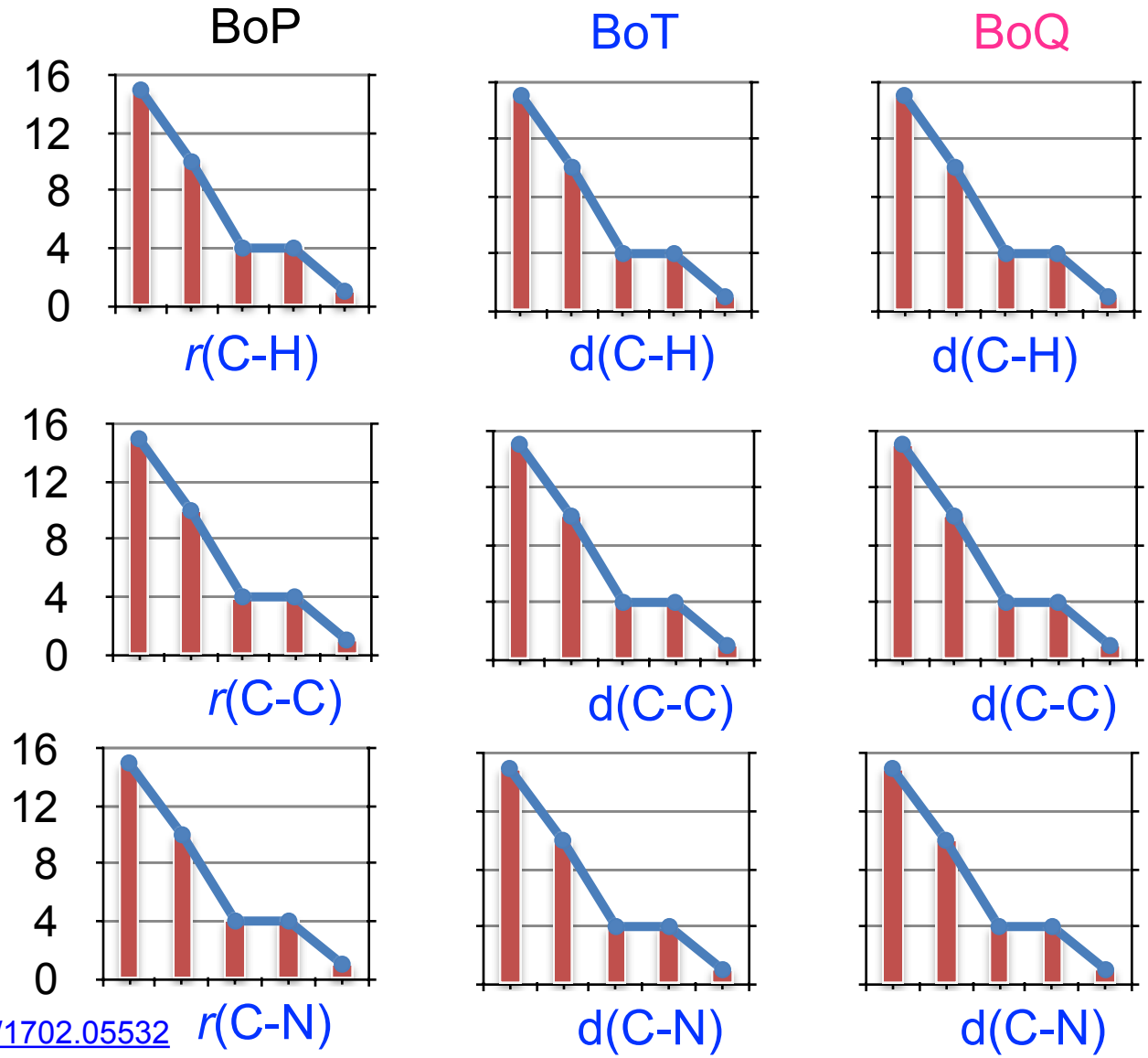
shortcoming:
force prediction



Histogram of Distance

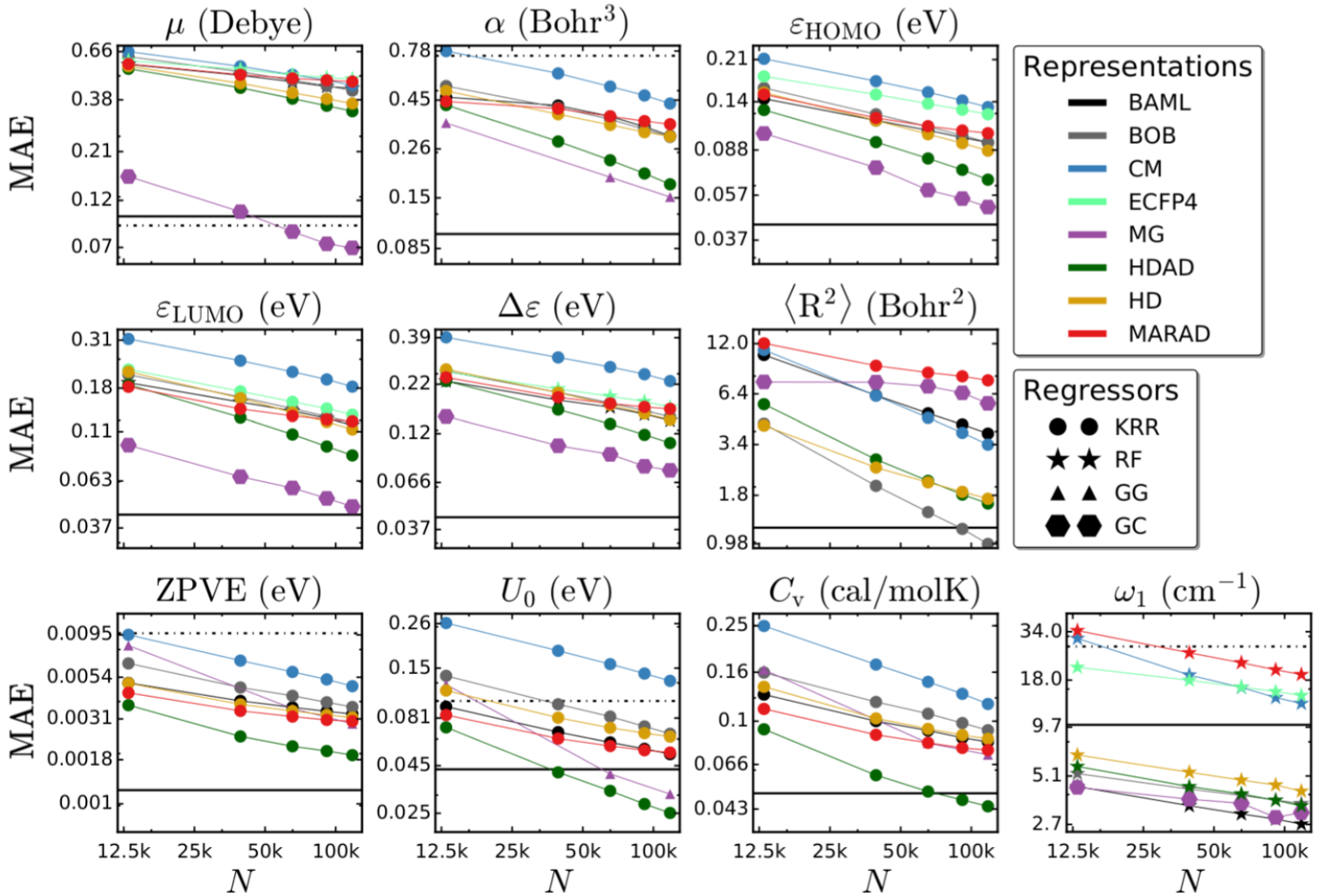
Histogram of Angles

Histogram of Dihedral angles



F. Faber, *et al.*, 2017, arxiv.org/abs/1702.05532

HDAD



F. Faber, *et al.*, 2017, arxiv.org/abs/1702.05532

Why is BAML worse than HDAD?

* empirical force field terms fails to describe reality in many cases

* uniqueness might also be an issue
    * e.g., a slighted deviated Morse potential may cause uniqueness issue

Bear in mind once again:

* be cautious to use the target function as representations!

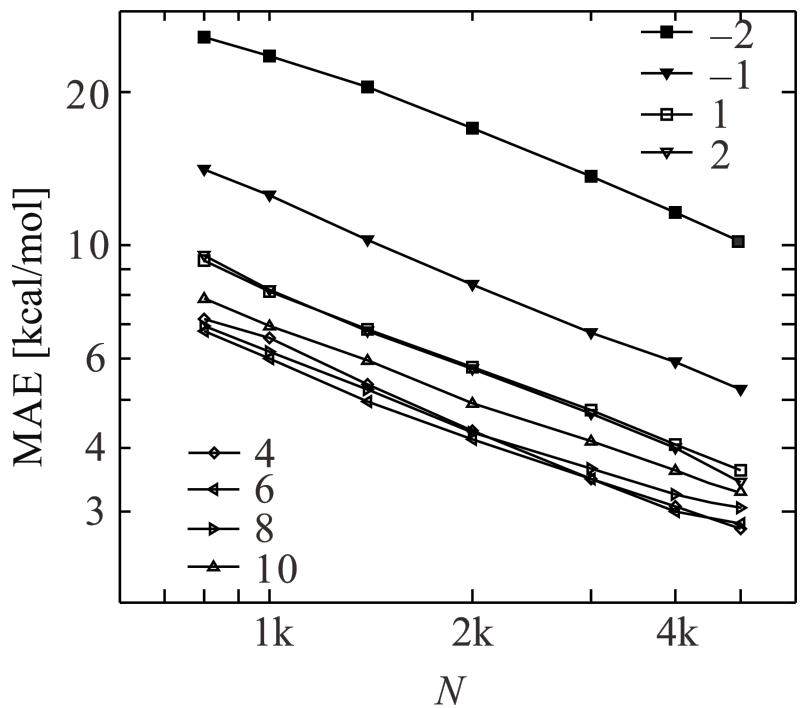
Improving the physics

E(2) = $Z_i Z_j$/R^n

QM7b dataset (size:7211)
property: enthalpy (H)

$$H^{\text{est}}(\text{CM}^{(n)}) = \sum \alpha_i k(\text{CM}^{(n)}, \text{CM}_i^{(n)})$$

Coulomb force: good as a rpst for bonding, bad for dispersion



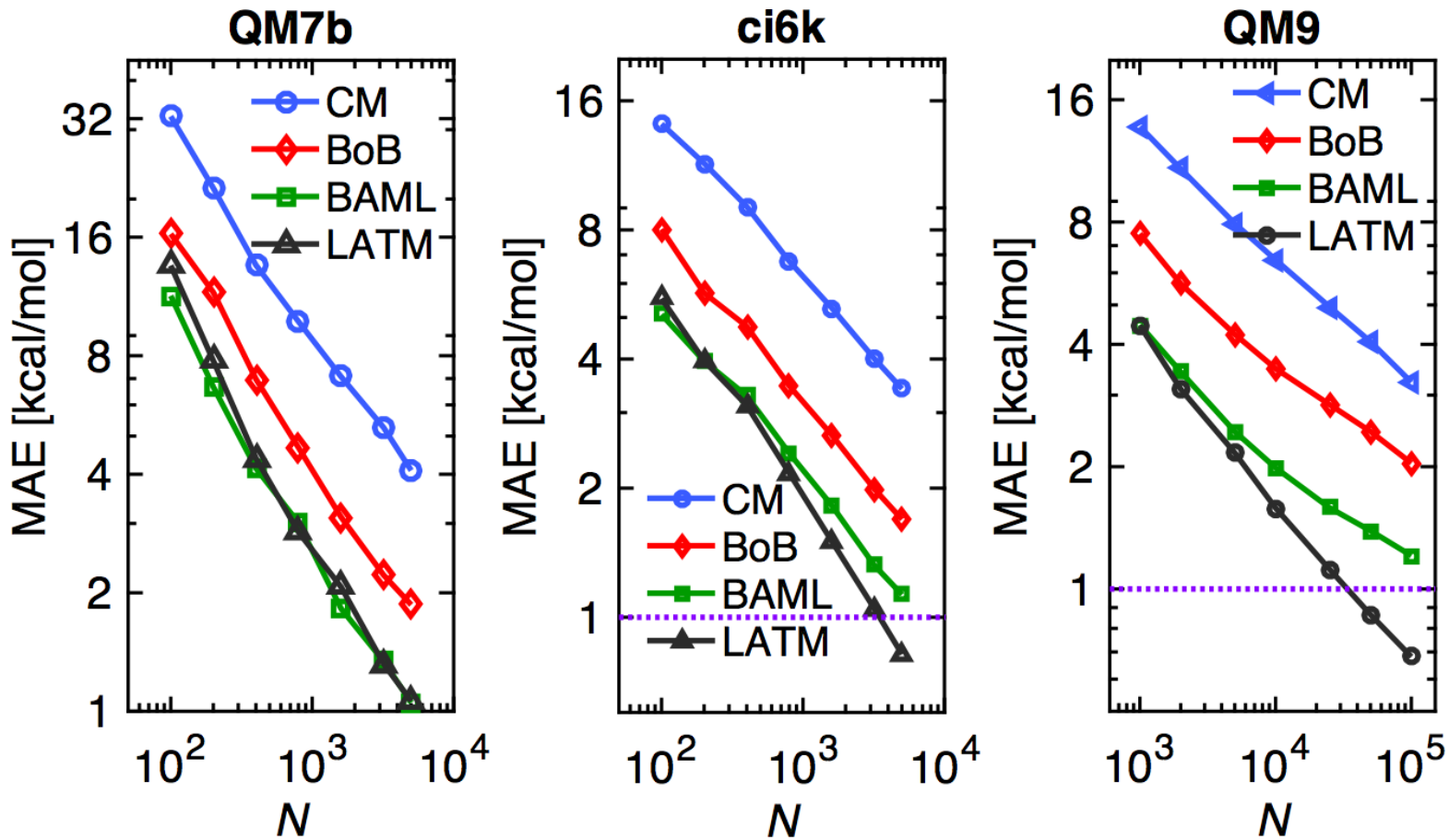
London force: good for dispersion, decent for bond

as a comprise, London wins!!

Improving the physics

Atoms + **L**ondon + **A**xilrod-**T**eller-**M**uto **(LATM)**

$$E^{(2)}(\mathbf{R_I}, \mathbf{R_J}) = -\frac{C_{6_{IJ}}}{R_{IJ}^6} \qquad E^{(3)}(\mathbf{R_I}, \mathbf{R_J}, \mathbf{R_K}) = C_{9_{IJK}} \frac{3\cos[\phi_I]\cos[\phi_J]\cos[\phi_K] + 1}{R_{IJ}^3 R_{IK}^3 R_{JK}^3}$$



HB, OAvL, *to be submitted* (2017)

extending *E*-based approach

build rpst based on decomposition of any extensive property:

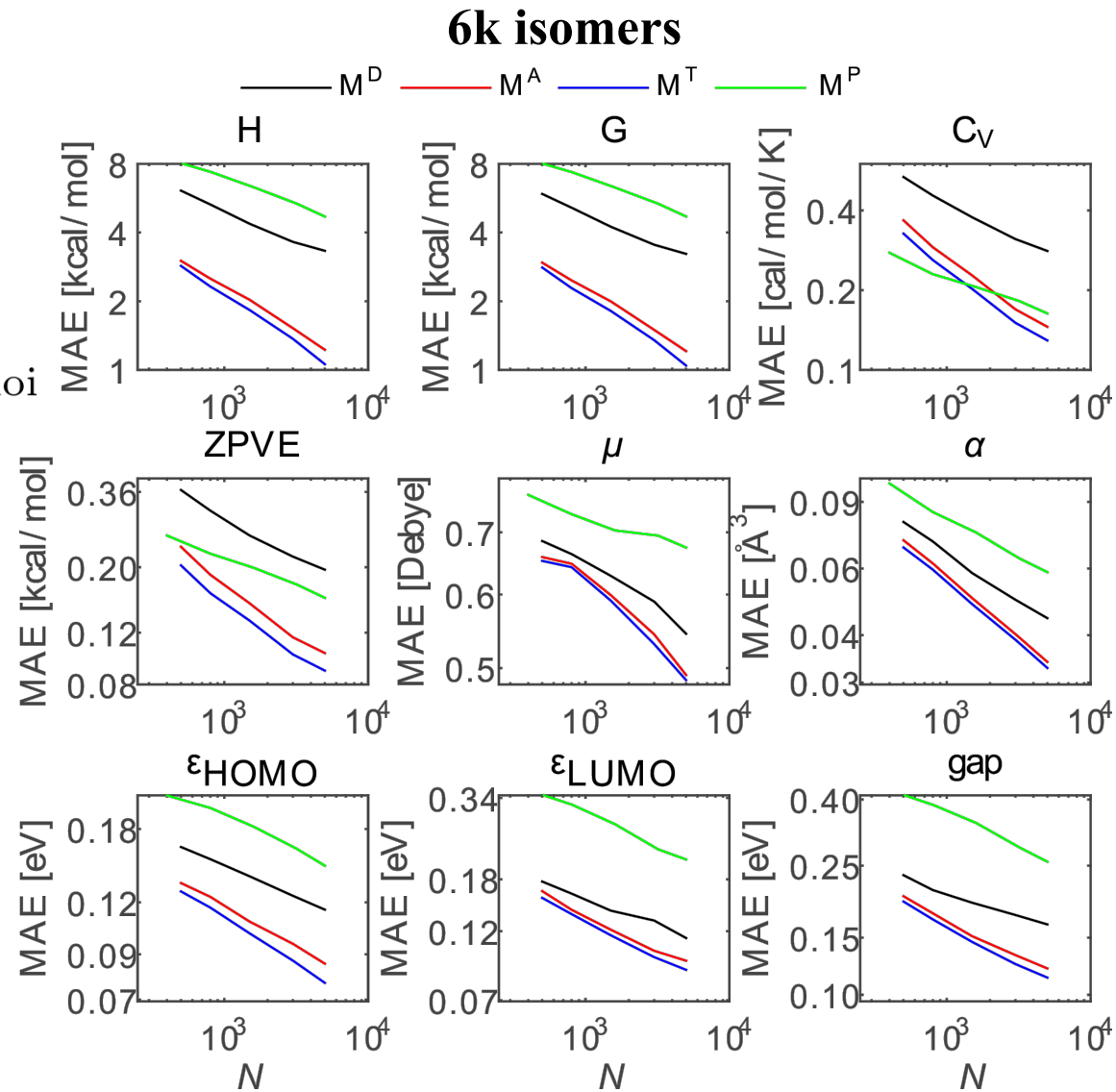
e.g., polarizability model

$$\alpha = \sum_I \alpha_I = \sum_I c_I V_I^{\text{voronoi}}$$

$$\mathbf{M}^{\mathrm{P}} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_n \end{bmatrix}$$

Only ONE-body!!

BH, OAvL, *JCP comm.*, 2016



Categorizing M

I

II

III

global representations

$\Psi/\rho/V$-based

continuous

discrete

E-based

atomic representations

Go Atomic

$$\varepsilon_i = \varepsilon(\mathbf{d}_i, \mathbf{w}) = \sum_h w_h \phi_h(\mathbf{d}_i);$$

$$\langle \varepsilon_i \varepsilon_j \rangle = \left\langle \sum_{hh'} w_h w_{h'} \phi_h(\mathbf{d}_i) \phi_{h'}(\mathbf{d}_j) \right\rangle = \sum_{hh'} \langle w_h w_{h'} \rangle \phi_h(\mathbf{d}_i) \phi_{h'}(\mathbf{d}_j)$$
$$= \sigma_w^2 \sum_h \phi_h(\mathbf{d}_i) \phi_h(\mathbf{d}_j)$$

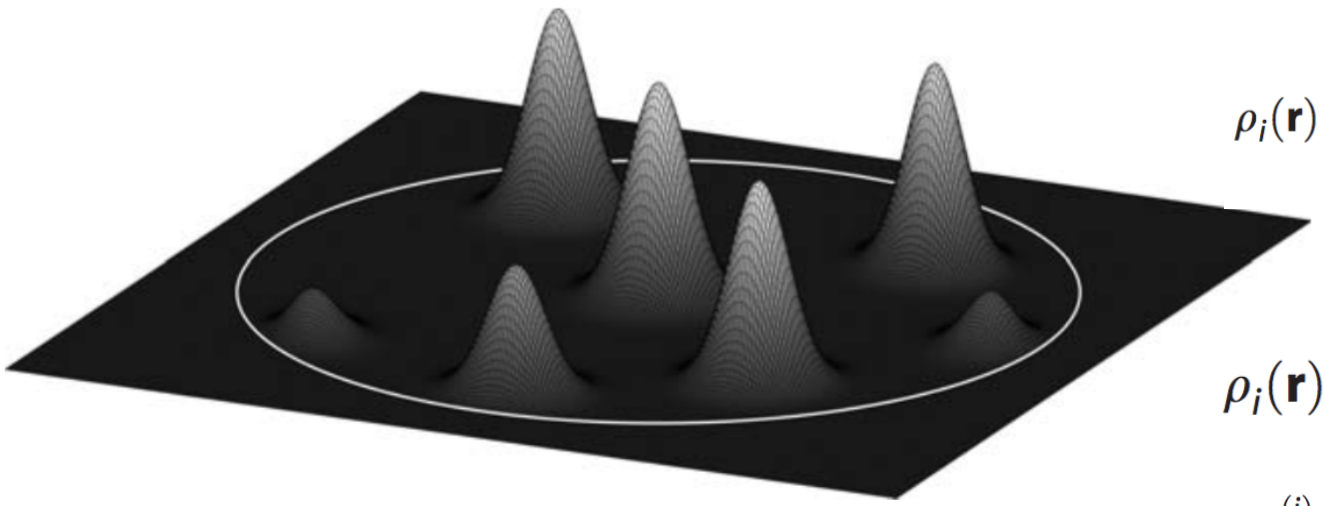
covariance

$$\langle E_N E_M \rangle = \left\langle \sum_{i \in N} \varepsilon(\mathbf{d}_i) \sum_{j \in M} \varepsilon(\mathbf{d}_j) \right\rangle = \left\langle \sum_{i \in N} \sum_{j \in M} \sum_{hh'} w_h w_{h'} \phi_h(\mathbf{d}_i) \phi_{h'}(\mathbf{d}_j) \right\rangle$$
$$= \sum_{i \in N} \sum_{j \in M} \sum_{hh'} \langle w_h w_{h'} \rangle \phi_h(\mathbf{d}_i) \phi_{h'}(\mathbf{d}_j) = \sigma_w^2 \sum_{i \in N} \sum_{j \in M} \sum_h \phi_h(\mathbf{d}_i) \phi_h(\mathbf{d}_j)$$
$$= \sigma_w^2 \sum_{i \in N} \sum_{j \in M} C(\mathbf{d}_i, \mathbf{d}_j)$$

$$Y_i^{\text{est}}(\mathbf{X}_i) = \sum_j \alpha_j \boxed{\exp\left(-\frac{d(\mathbf{X}_j, \mathbf{X}_i)}{\sigma}\right)} + b$$

P. R. Bartok *et al*, *IJQC* (2015)

Smooth Overlap of Atomic Positions (SOAP)



$$\rho_i(\mathbf{r}) \equiv \sum_j^{\text{neigh.}} \exp\left(-\frac{|\mathbf{r}-\mathbf{r}_{ij}|^2}{2\sigma_{\text{atom}}^2}\right)$$

projection to basis set

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

$$p_{nn'l}^{(i)} \equiv \frac{1}{\sqrt{2l+1}} \sum_m c_{nlm}^{(i)} (c_{n'lm}^{(i)})^*$$

$$C'(\rho_i, \rho_j) = \sum_{n,n',l} p_{nn'l}^{(i)} p_{nn'l}^{(j)}$$

A serious problem of SOAP:
for large r_cutoff, how to distinguish
two very different atoms around centre?

application: simple crystals so far

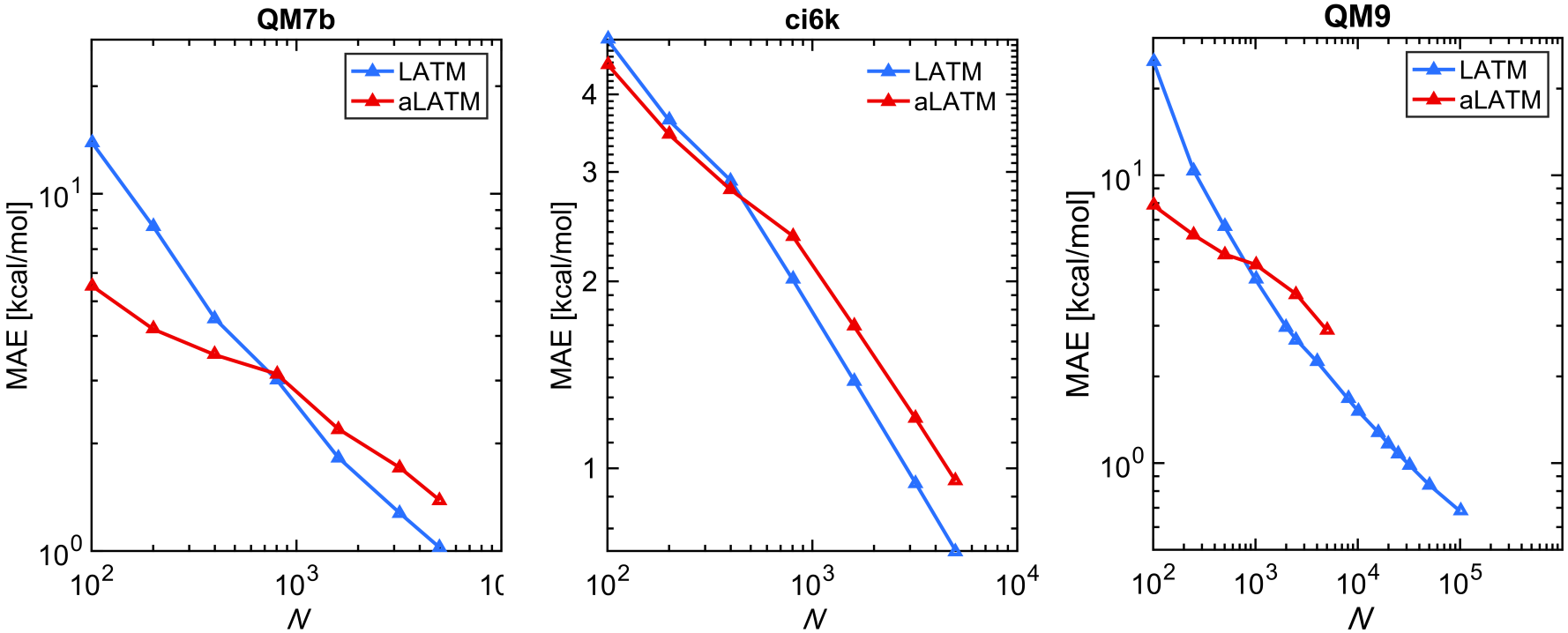
Fix SOAP for molecules by RE-Match, glory lost as an atomic rpst

works best with a small r_cutoff !!

P. R. Bartok *et al*, *PRB* (2013) S. De, *et al.*, *PCCP*, 2016

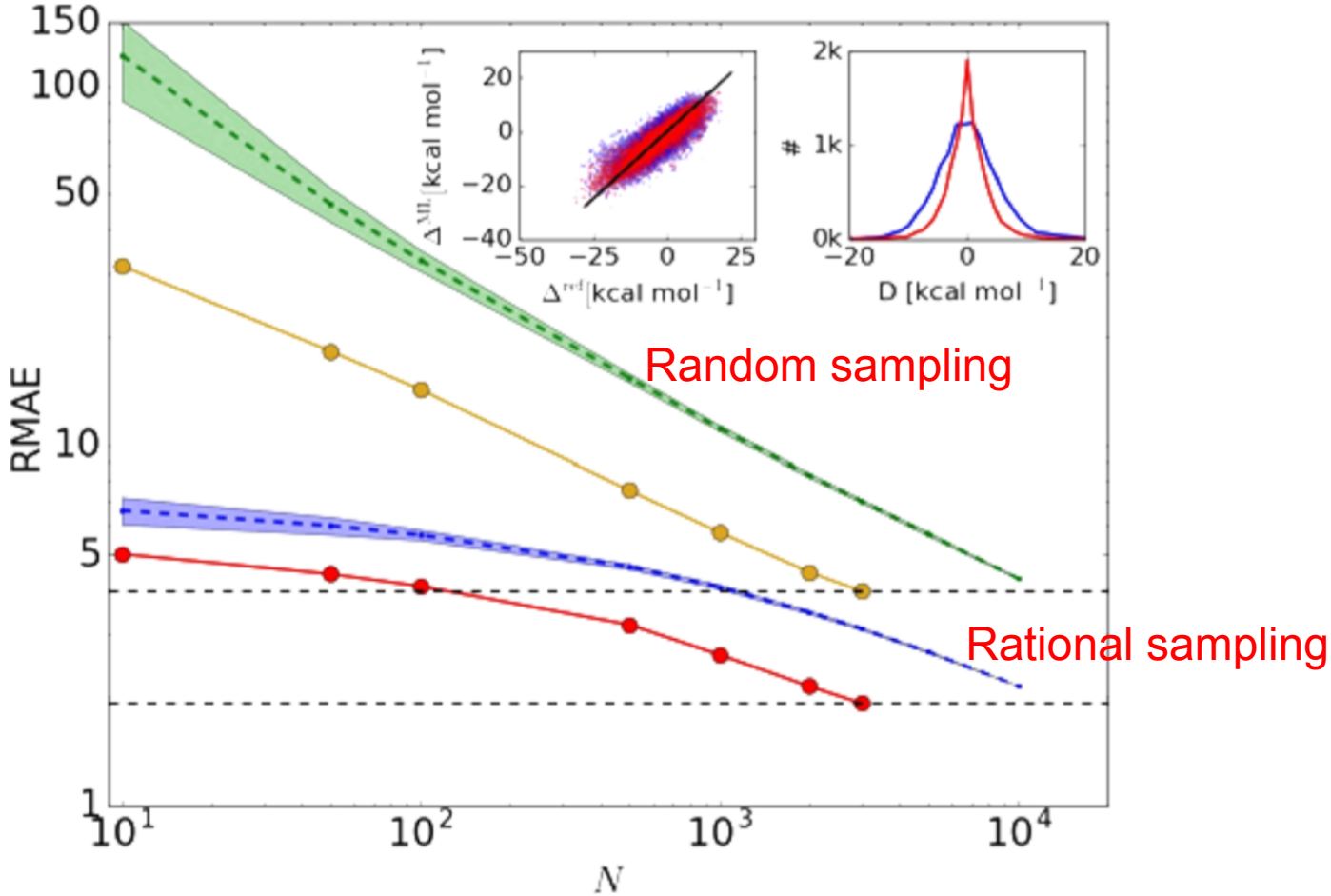
aLATM

MBE-based approach: more natural to define atomic rpst
1. includes 2-, 3-body interactions
2. both decay with r



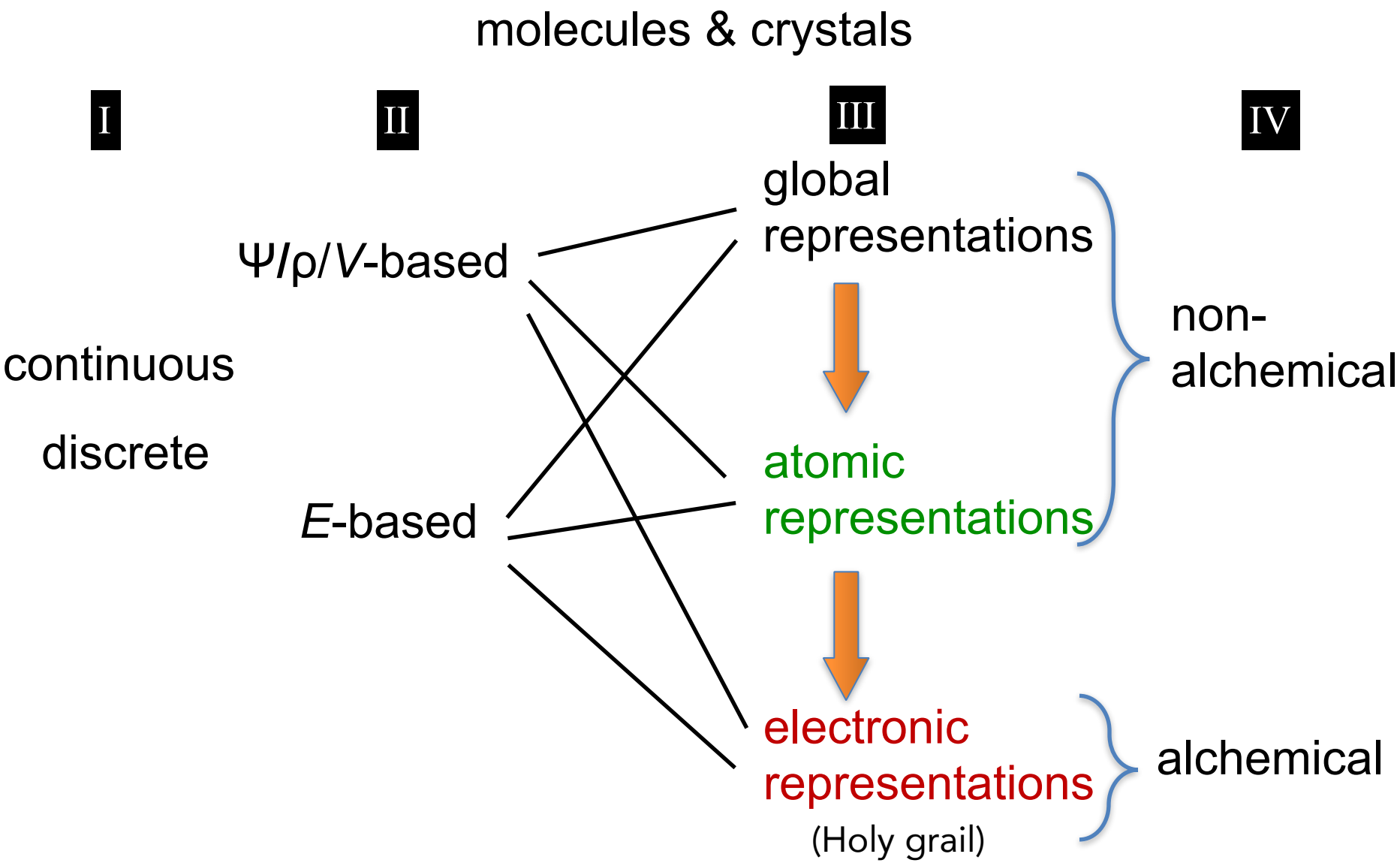
HB, OAvL, *to be submitted* (2017)

why aLATM is bad at larger *N*?



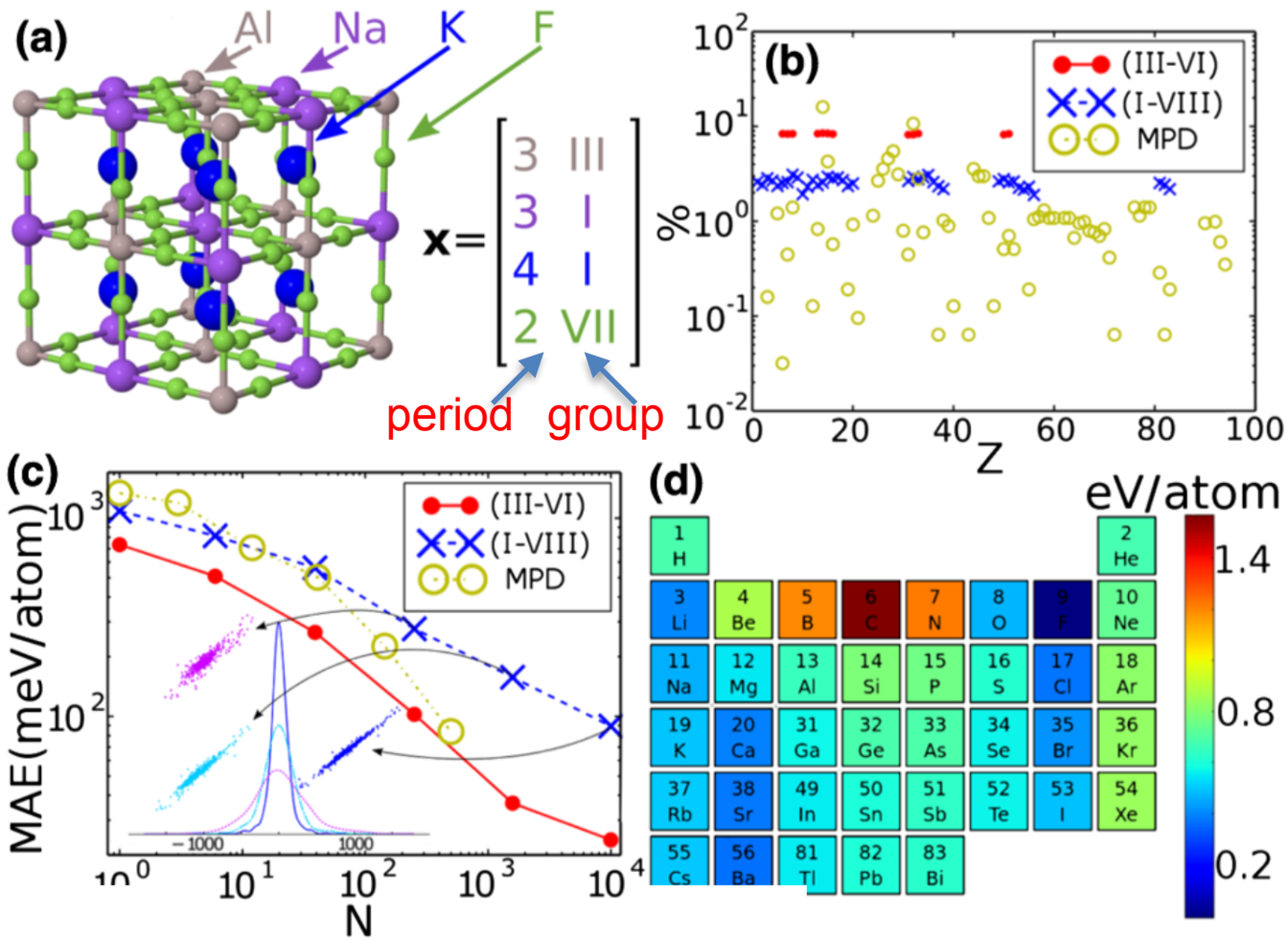
N. J. Browning, *et al.,* JPCL (2017)

Categorizing M



Go Electronic

overall 2*M* Elpasolite ABC_2D_6 Crystals



F. Faber, *et al*, *PRL* (2016)

Conclusions and Outlook

1. Almost all rpsts in literature were categorised

2. Two general approaches for rational design of rpst
    a. Schrödinger equation: ρ/Ψ/V_ext
    b. many-body expansion

3. Two general principles for rational design of rpst
    a. uniqueness (necessary for convergence)
    b. similarity to target reduces off-set of LC

4. MBE based rpst (e.g., BAML, LATM, HDAD) offer
    a. Meaning
    b. Simplicity
    c. Accuracy

5. and is generally better than ρ/Ψ/V_ext based approach

6. There is great potential for electronic rpst to beat everything else

Acknowledgements:

Prof. Dr. O. Anatole von Lilienfeld

FNSNF
SWISS NATIONAL SCIENCE FOUNDATION

MARVEL
NATIONAL CENTRE OF COMPETENCE IN RESEARCH

Research at Google

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
UNITED STATES AIR FORCE

Institute for Pure & Applied Mathematics
IPAM
University of California, Los Angeles

U.S. DEPARTMENT OF ENERGY
INCITE
LEADERSHIP COMPUTING