# Machine Learning for Atomic and Molecular Simulations

Michele Ceriotti
https://cosmo.epfl.ch

IPAM - November 2019

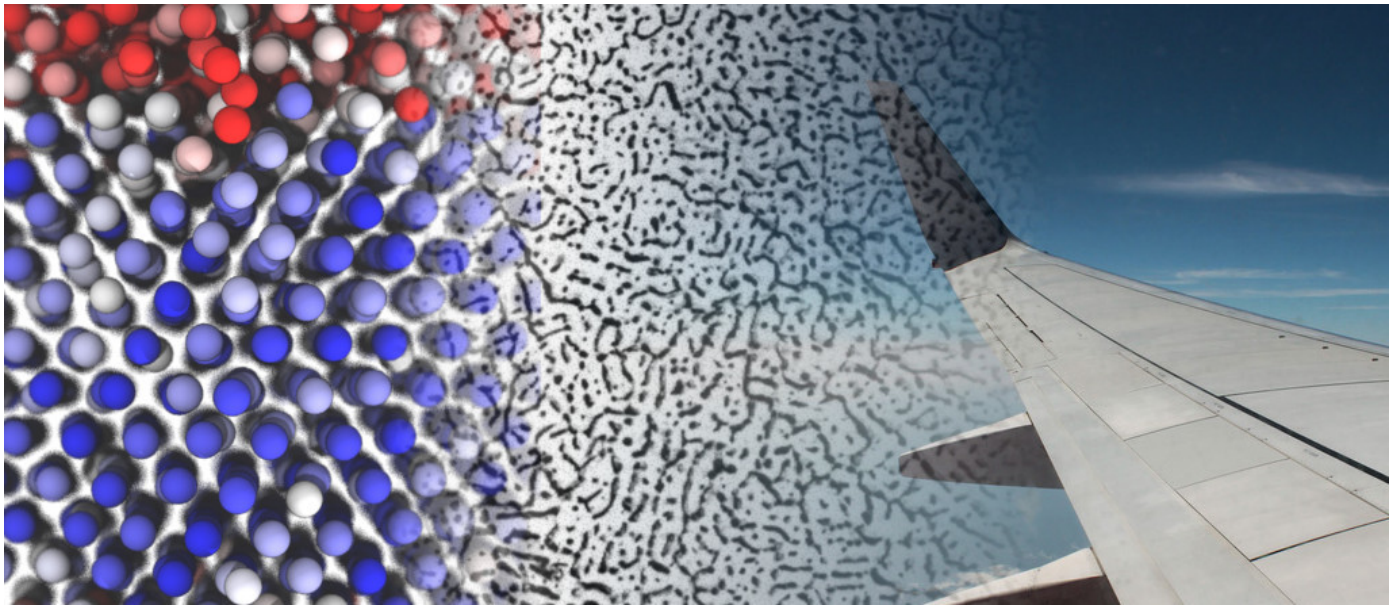
COSMO
COMPUTATIONAL SCIENCE AND MODELLING

EPFL

# Atomistic materials modelling

- Direct insights into mechanisms and structure-property relations at the atomic time and length scales
- Evaluation of microscopic quantities that enter meso-scale models
- Electronic structure calculations give first-principles estimate of stability and properties *of a given geometry* $\mathbf{q}$

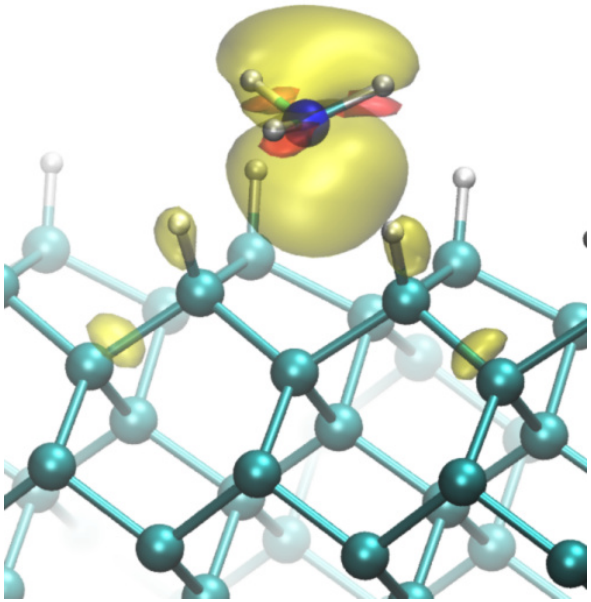
$$\hat{H}(\mathbf{q})\,|\Psi\rangle = V(\mathbf{q})\,|\Psi\rangle$$

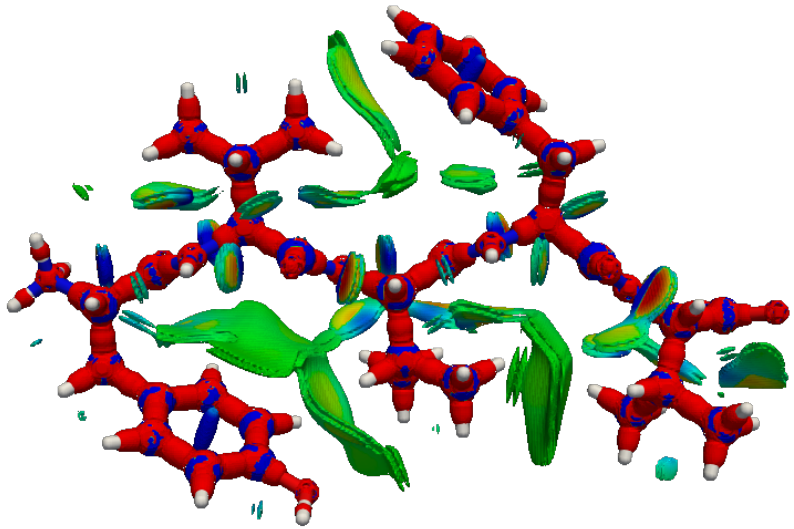


# Atomistic materials modelling

- Direct insights into mechanisms and structure-property relations at the atomic time and length scales
- Evaluation of microscopic quantities that enter meso-scale models
- Electronic structure calculations give first-principles estimate of stability and properties *of a given geometry* **q**

$$\hat{H}(\mathbf{q})\,|\Psi\rangle = V(\mathbf{q})\,|\Psi\rangle$$



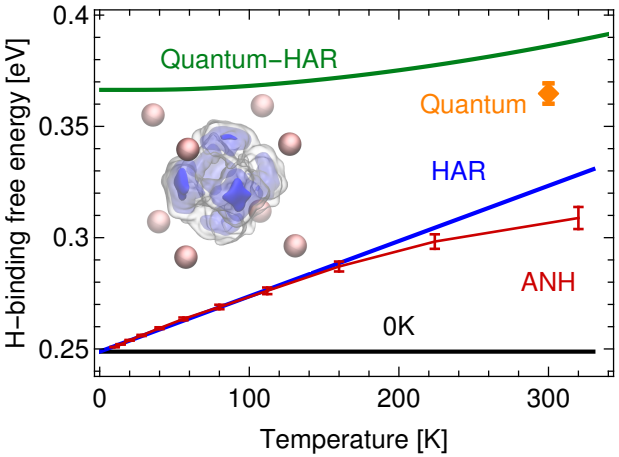


# Finite-temperature thermodynamics of materials

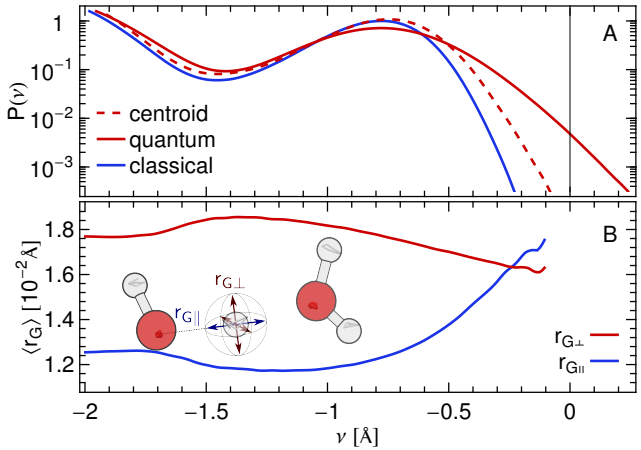
- Materials at finite temperature undergo fluctuations

$$\langle A \rangle = \int \mathrm{d}\mathbf{q} e^{-\beta V(\mathbf{q})} A(\mathbf{q})$$

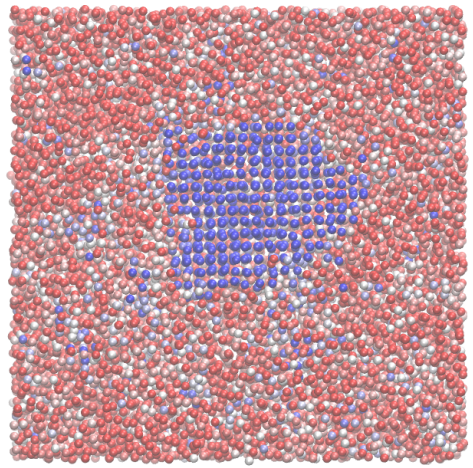
  - Sampling of anharmonic free energies
  - Nuclear quantum effects
  - Activated events and phase transitions
- All of these simulations require evaluating atomic-scale properties an enormous ($\mathcal{O}(10^6)$) number of times




Cheng, Paxton, **MC**, Phys. Rev. Lett. (2018)




**MC** et al., Chem. Rev. (2016)



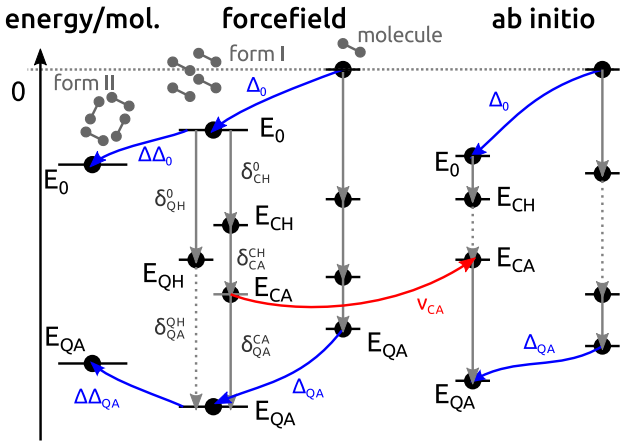
Cheng & **MC**, J. Chem. Phys. (2017)

# Finite-temperature thermodynamics of materials

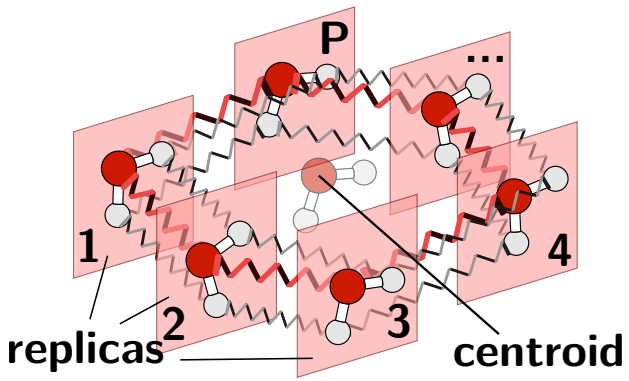
- Materials at finite temperature undergo fluctuations

$$\langle A \rangle = \int \mathrm{d}\mathbf{q} e^{-\beta V(\mathbf{q})} A(\mathbf{q})$$

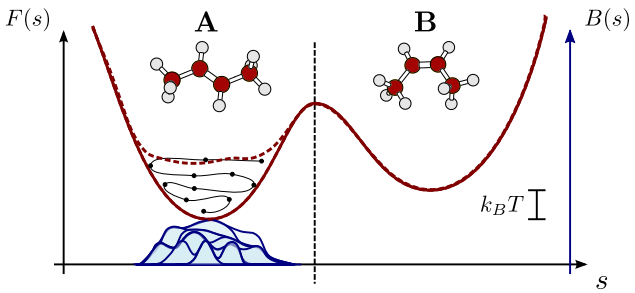
  - Sampling of anharmonic free energies
  - Nuclear quantum effects
  - Activated events and phase transitions
- All of these simulations require evaluating atomic-scale properties an enormous ($\mathcal{O}\left(10^6\right)$) number of times




Rossi, Gasparotto & **MC**, Phys. Rev. Lett. (2016)



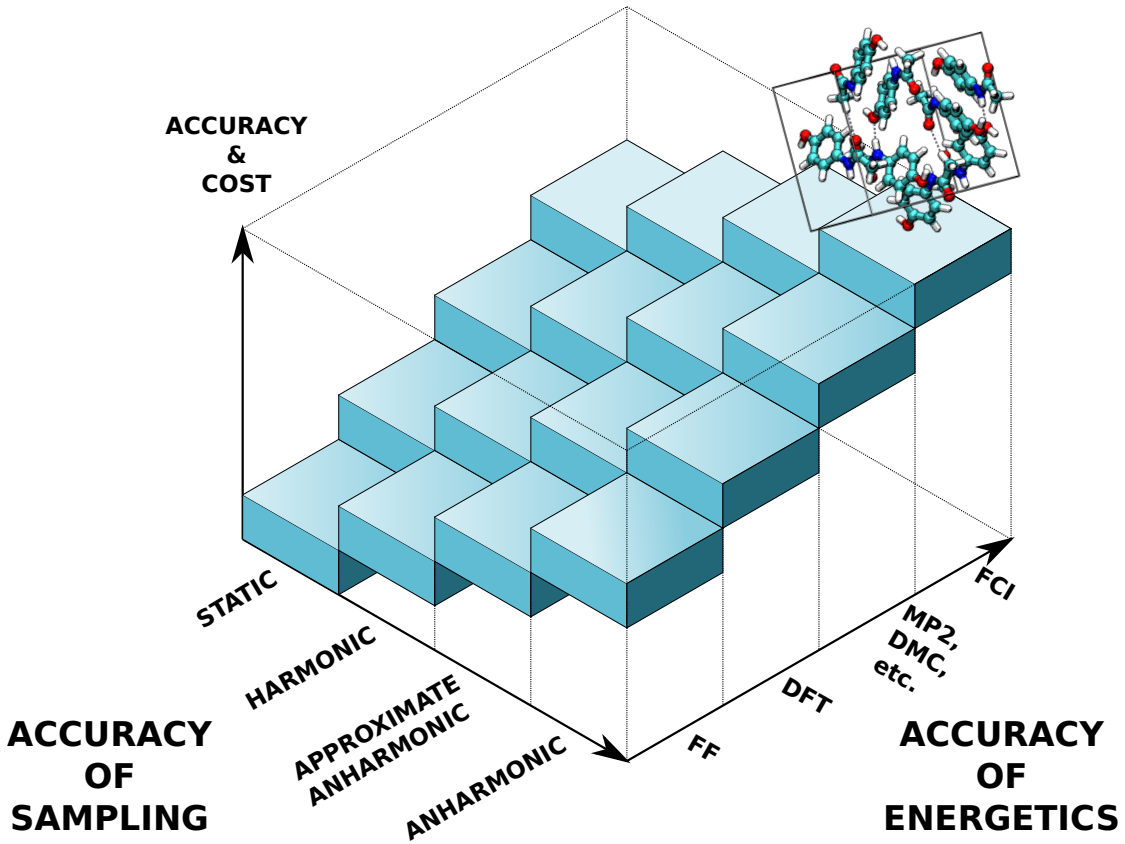

Markland & **MC**, Nat. Rev. Chem. (2018)




Cheng, Tribello, **MC**, Phys. Rev. B (2015)

# Difficult choices...

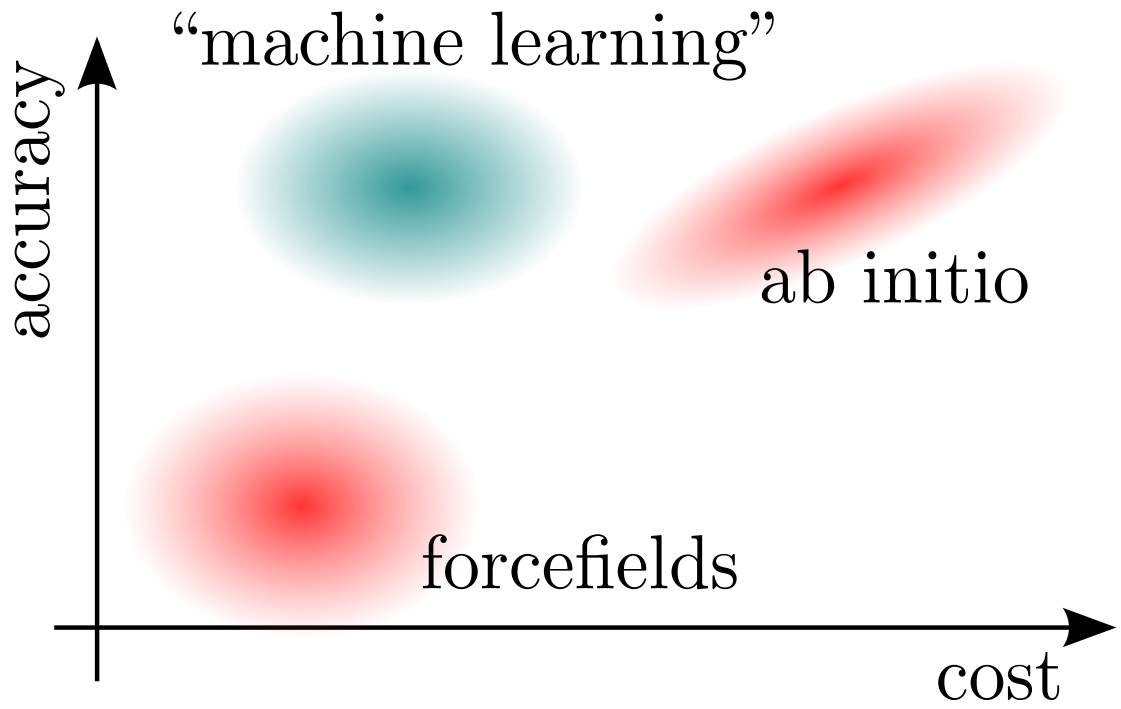
- Hard to get accurate electronic structure *and* thorough statistics
- ML potentials (& properties) as a way to beat the trade-off between inexpensive potential & good statistics / accurate & no sampling




Kapil, Engel, Rossi, **MC**, JCTC (2019)

# Difficult choices...

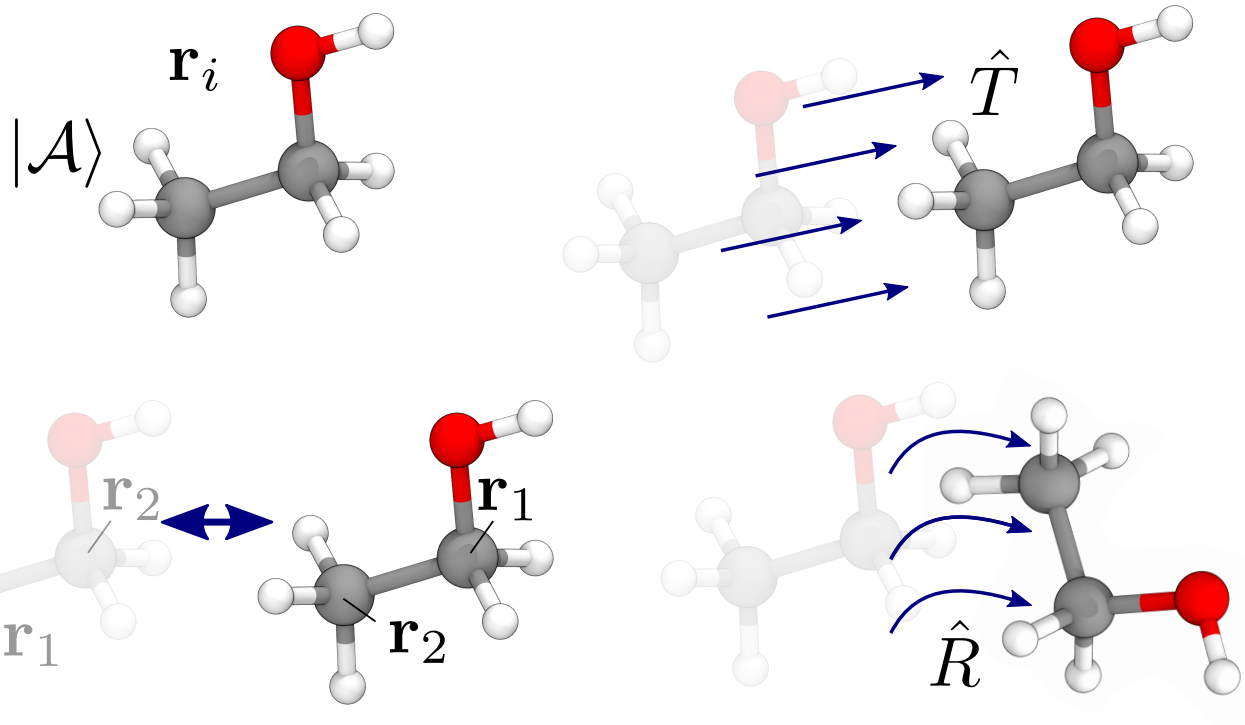
- Hard to get accurate electronic structure *and* thorough statistics
- ML potentials (& properties) as a way to beat the trade-off between inexpensive potential & good statistics / accurate & no sampling



# Physics-inspired atomic representations

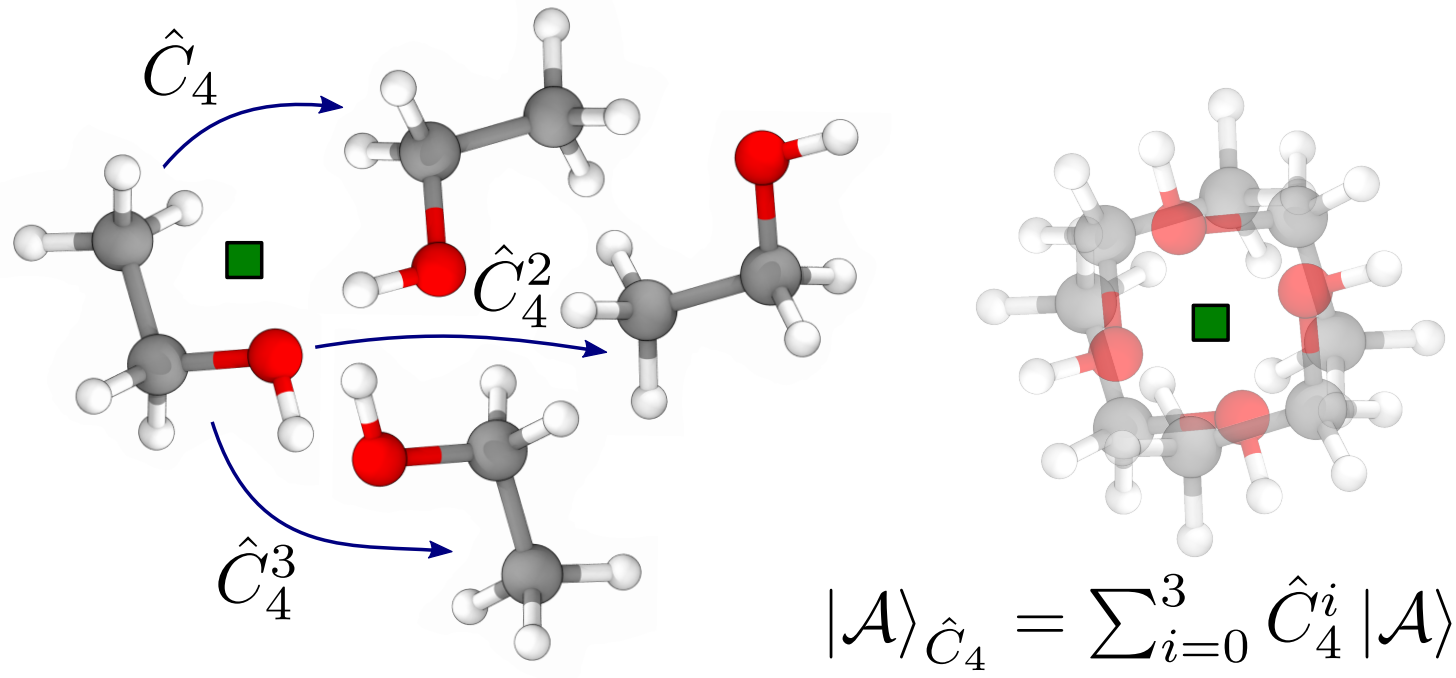
# The role of symmetry

- Structure representations should reflect basic physical symmetries: atom permutation, rigid translations, rigid rotations, inversion, . . .
- Cartesian coordinates of the atoms do not fulfill most of these
- Incorporating symmetries reduces the effective dimensionality of the problem, making regression more data efficient



# Symmetry by averaging

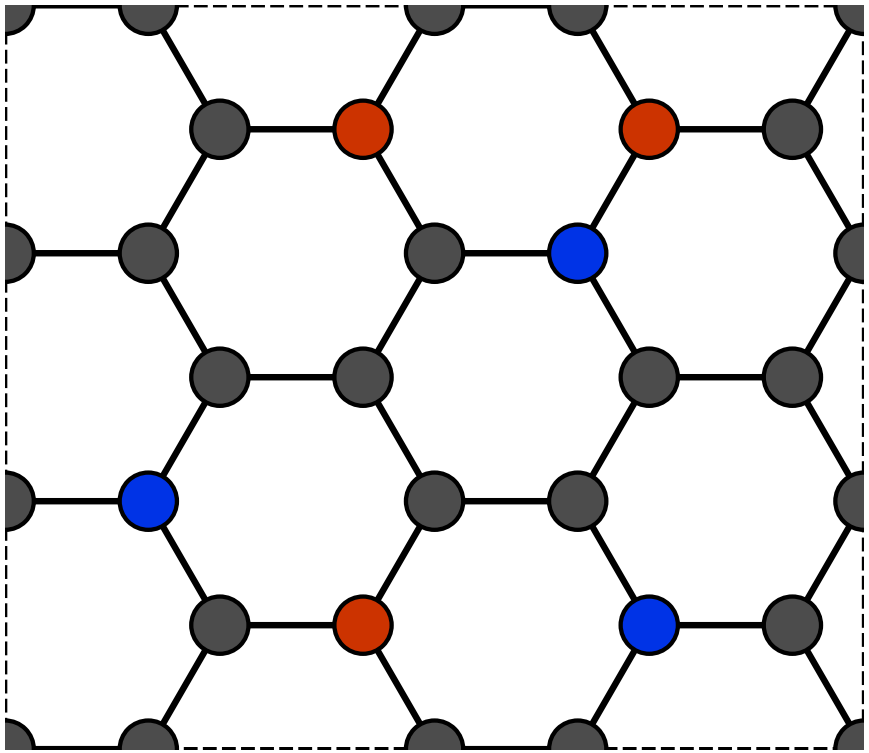
- Symmetries can be incorporated by *data augmentation* - multiple copies of each input with symmetries applied
- Symmetries can be incorporated in the structure of the model (convolutional NN, tensor models, etc.)
- Symmetries can be encoded into the representation by *Haar integration* - averaging the feature vector over the symmetry operations




$$|\mathcal{A}\rangle_{\hat{C}_4} = \sum_{i=0}^{3} \hat{C}_4^i \, |\mathcal{A}\rangle$$

# A symmetry-adapted representation for structures

- Start with an arbitrary configuration of atoms. Associate a vector to each structure that describes it, staying as abstract as possible
- A decorated atom density solves the problem of index permutations
- *Translational* symmetry can be included by averaging over $\hat{t}$, leading to a sum of atom-centered representations



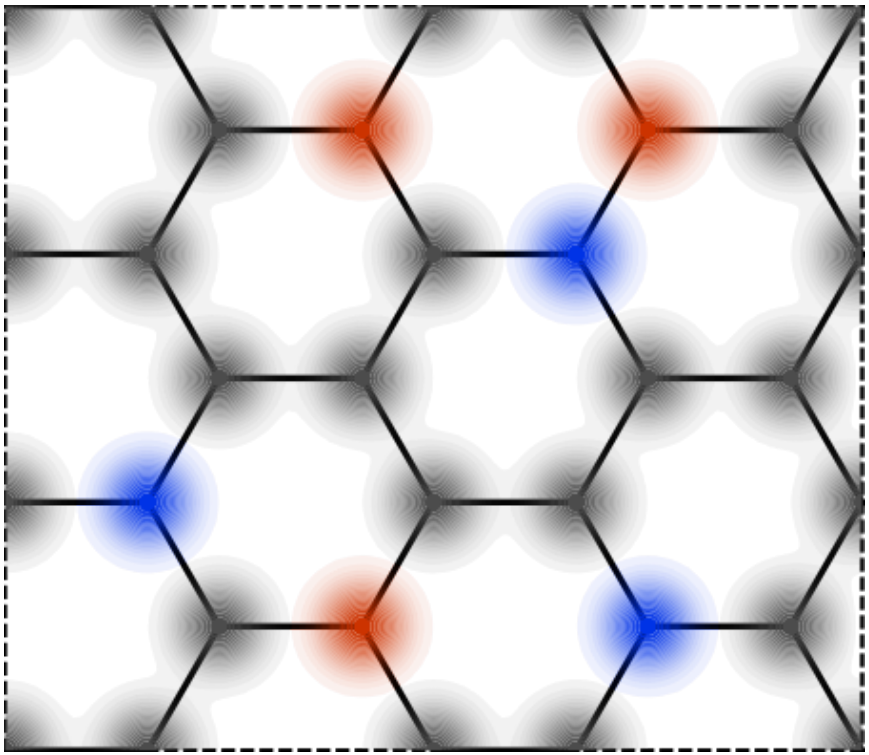
$\mathcal{A} \rightarrow$

```
C 0.00 0.00 0.00
C 0.00 1.00 0.00
B 1.00 2.00 0.00
.......
```

Willatt, Musil, **MC**, JCP (2019)

# A symmetry-adapted representation for structures

- Start with an arbitrary configuration of atoms. Associate a vector to each structure that describes it, staying as abstract as possible
- A decorated atom density solves the problem of index permutations
- *Translational* symmetry can be included by averaging over $\hat{t}$, leading to a sum of atom-centered representations



$$\langle \mathbf{r} | \mathcal{A} \rangle = \sum_i g(\mathbf{r} - \mathbf{r_i}) \, |\alpha_i\rangle$$

$|\mathrm{C}\rangle$ ■

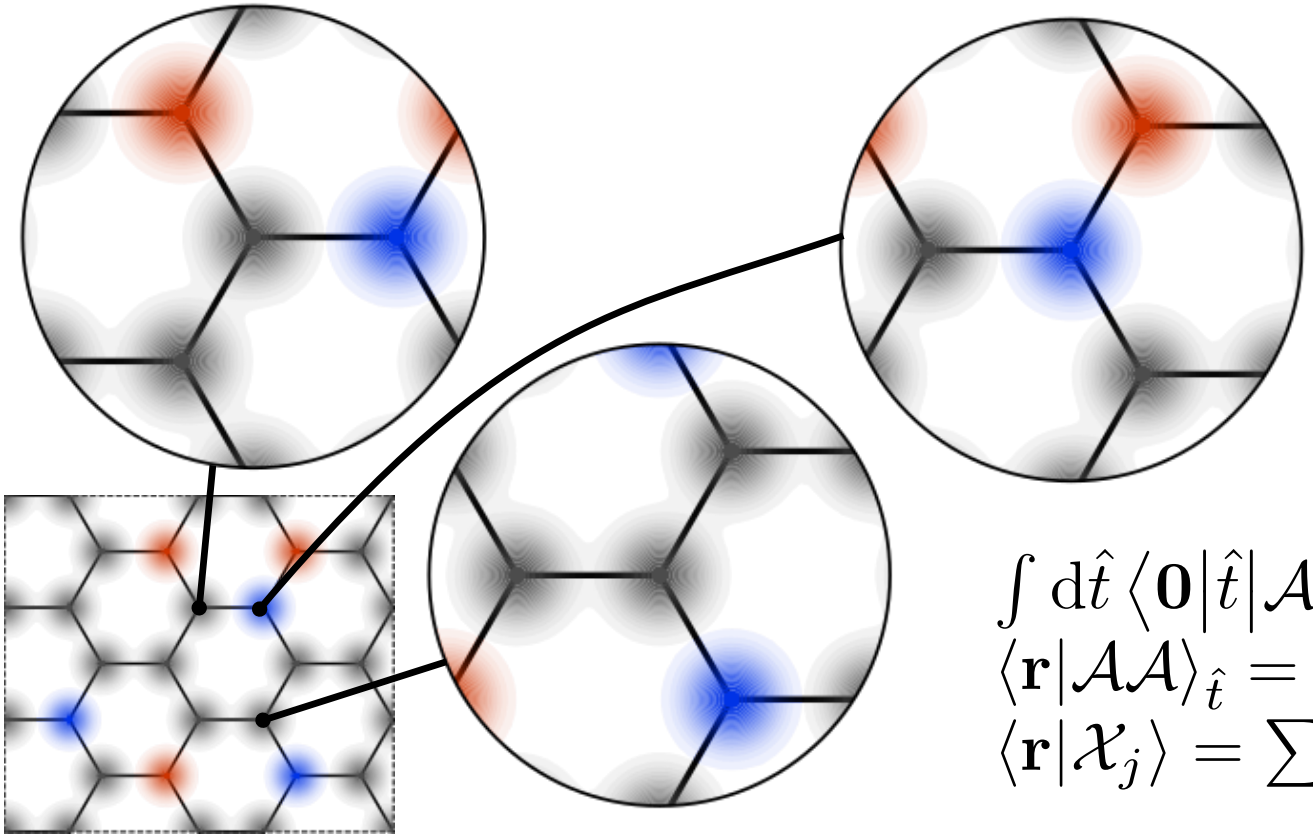
$|\mathrm{N}\rangle$ ■

$|\mathrm{B}\rangle$ ■

Willatt, Musil, **MC**, JCP (2019)

# A symmetry-adapted representation for structures

- Start with an arbitrary configuration of atoms. Associate a vector to each structure that describes it, staying as abstract as possible
- A decorated atom density solves the problem of index permutations
- *Translational* symmetry can be included by averaging over $\hat{t}$, leading to a sum of atom-centered representations

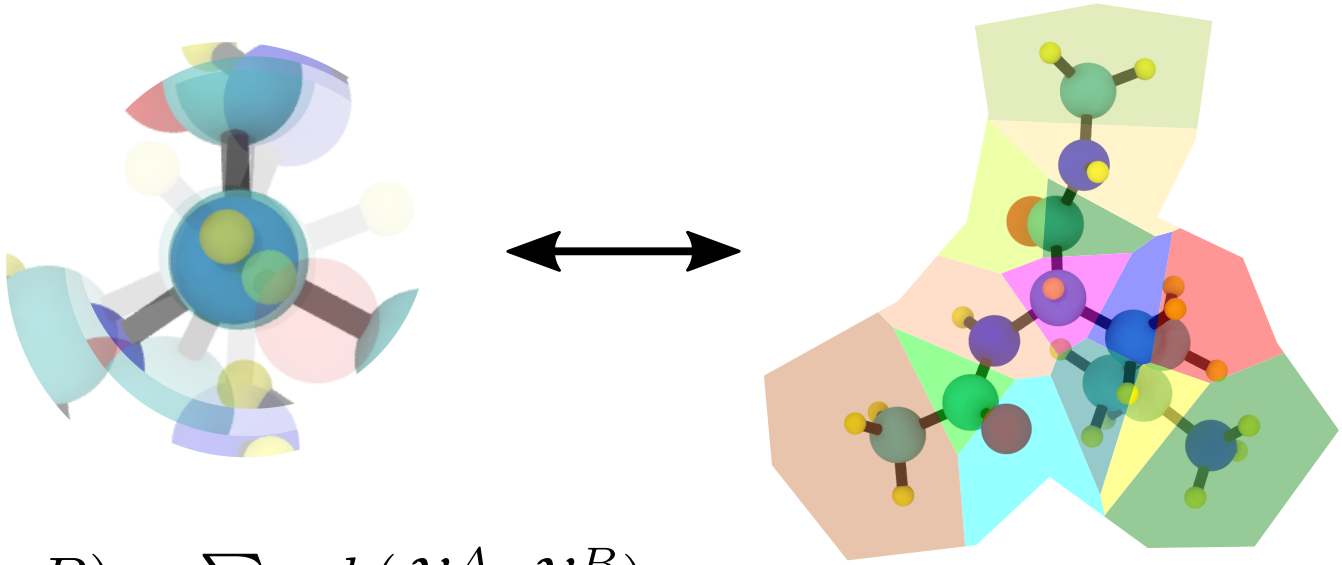


$$\int \mathrm{d}\hat{t} \left\langle \mathbf{0} \middle| \hat{t} \middle| \mathcal{A} \right\rangle \left\langle \mathbf{r} \middle| \hat{t} \middle| \mathcal{A} \right\rangle =$$
$$\left\langle \mathbf{r} \middle| \mathcal{A}\mathcal{A} \right\rangle_{\hat{t}} = \sum_j \left| \alpha_j \right\rangle \left\langle \mathbf{r} \middle| \mathcal{X}_j \right\rangle$$
$$\left\langle \mathbf{r} \middle| \mathcal{X}_j \right\rangle = \sum_i g(\mathbf{r} - \mathbf{r}_{ij}) \left| \alpha_i \right\rangle$$

Willatt, Musil, **MC**, JCP (2019)

# Additivity, nearsightedness and locality

- Only by symmetry arguments we get a representation of a structure in terms of a sum over atom-centered terms
- This form implies (for a linear model or an average kernel) an additive form of the property
- Nearsightedness principle suggests that one only needs to consider finite range of correlations



$$K(A,B) = \sum_{i,j} k(\mathcal{X}_i^A, \mathcal{X}_j^B)$$

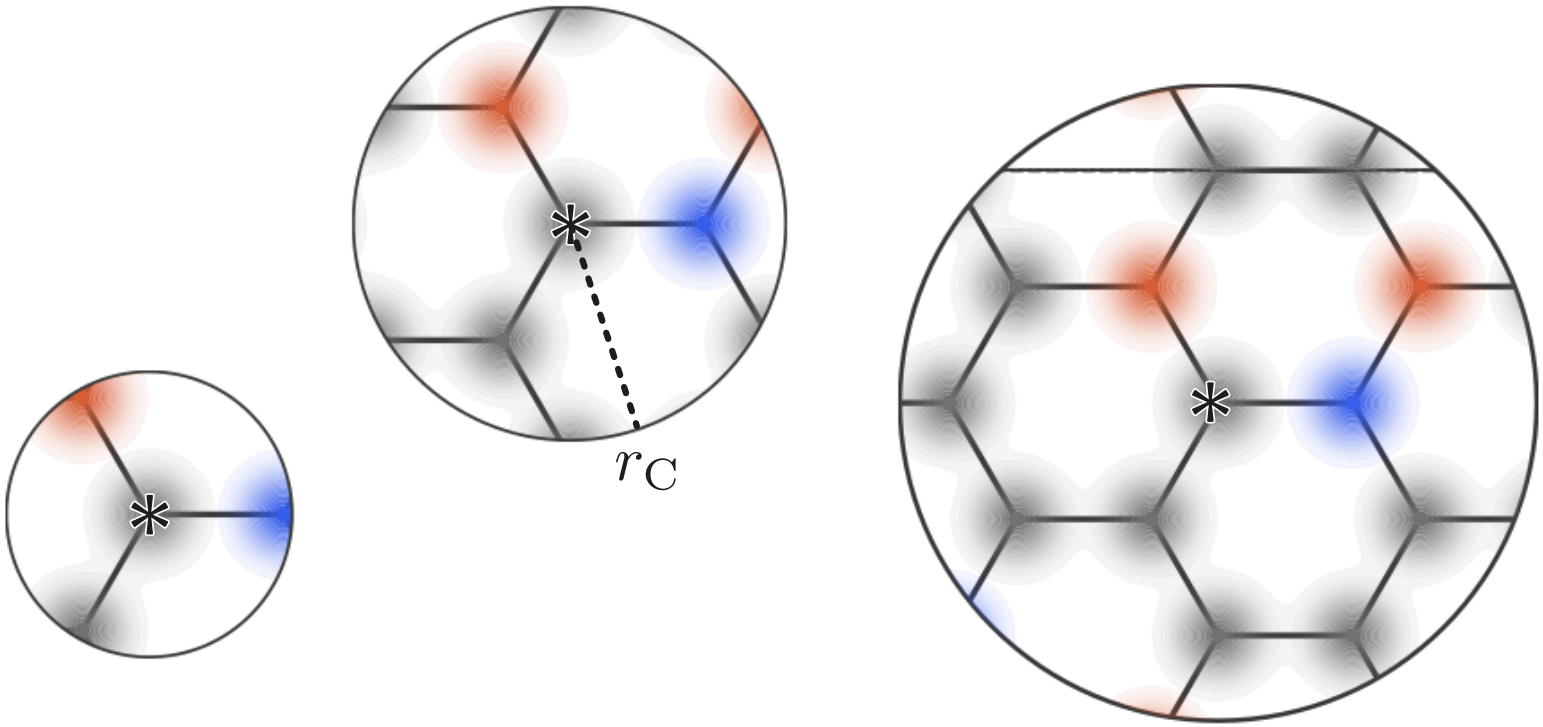
$$|\mathcal{A}\rangle = \sum_j |\mathcal{X}_j\rangle$$

$$E(A) = \sum_i \epsilon(\mathcal{X}_i^A)$$

De, Bartók, Csányi, **MC**, PCCP (2016)

# Additivity, nearsightedness and locality

- Only by symmetry arguments we get a representation of a structure in terms of a sum over atom-centered terms
- This form implies (for a linear model or an average kernel) an additive form of the property
- Nearsightedness principle suggests that one only needs to consider finite range of correlations

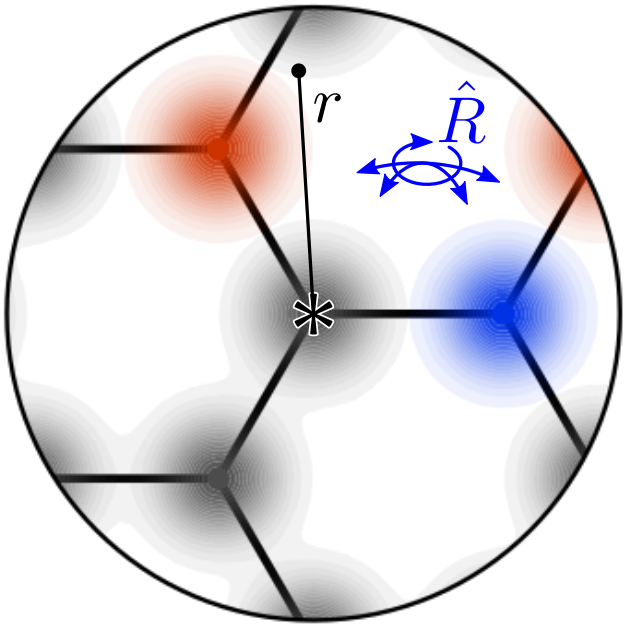


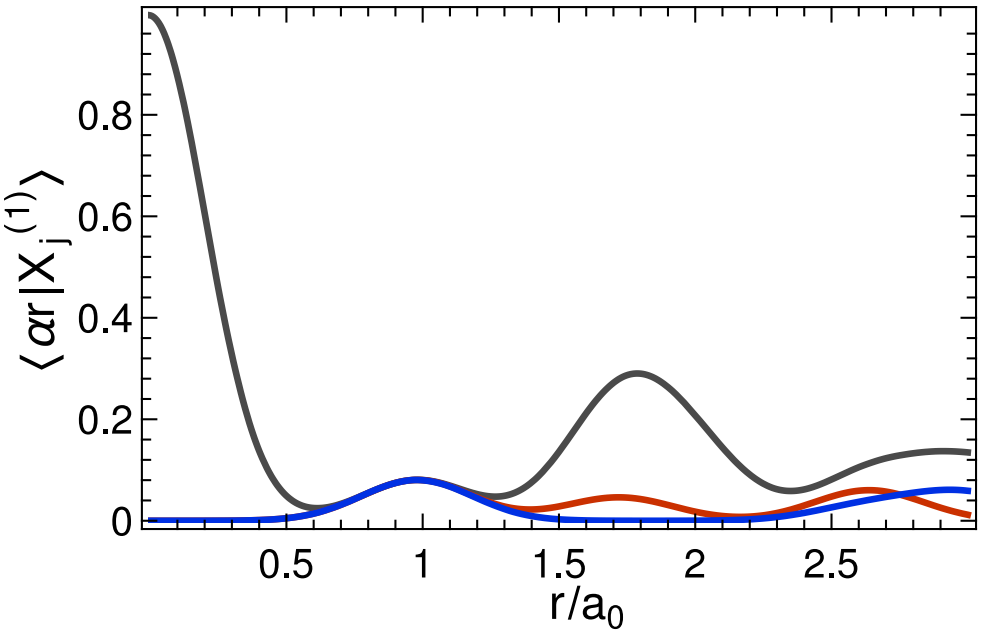

Prodan, Kohn, PNAS (2005)

# Rotations, density correlations and potentials

- Incorporating rotations leads to a hierarchy of spherical invariants
- Invariant representations are essentially the same *n*-body correlations that are used in statistical theories of liquids
- Linear models built on $\left\langle \cdot \middle| \mathcal{X}^{(n-1)} \right\rangle_{g\to\delta}$ yield *n*-body potential expansion

$$V\left(\{\mathbf{r}_i\}\right) = \sum_{ij} V^{(2)}\left(r_{ij}\right) + \sum_{ij} V^{(3)}\left(r_{ij}, r_{ik}, \omega_{ijk}\right) \dots$$



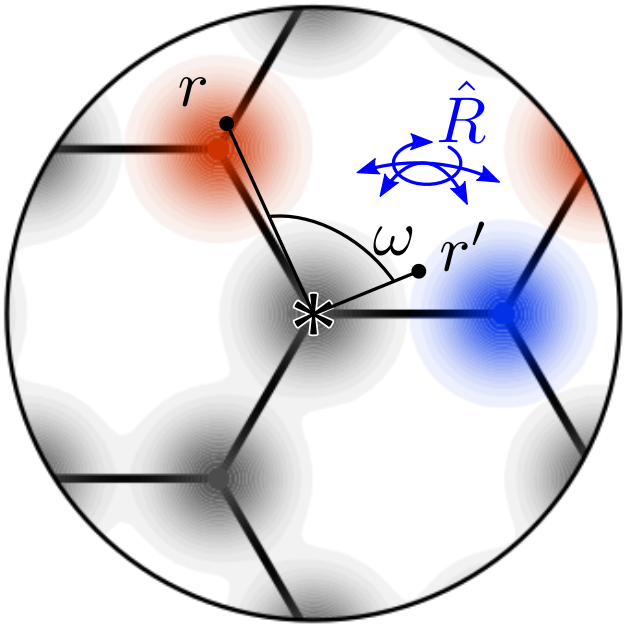

$$\left\langle \alpha r \middle| \mathcal{X}_j^{(1)} \right\rangle_{\hat{R}} = \int \mathrm{d}\hat{R} \left\langle \alpha r \hat{\mathbf{r}} \middle| \hat{R} \middle| \mathcal{X}_j \right\rangle$$

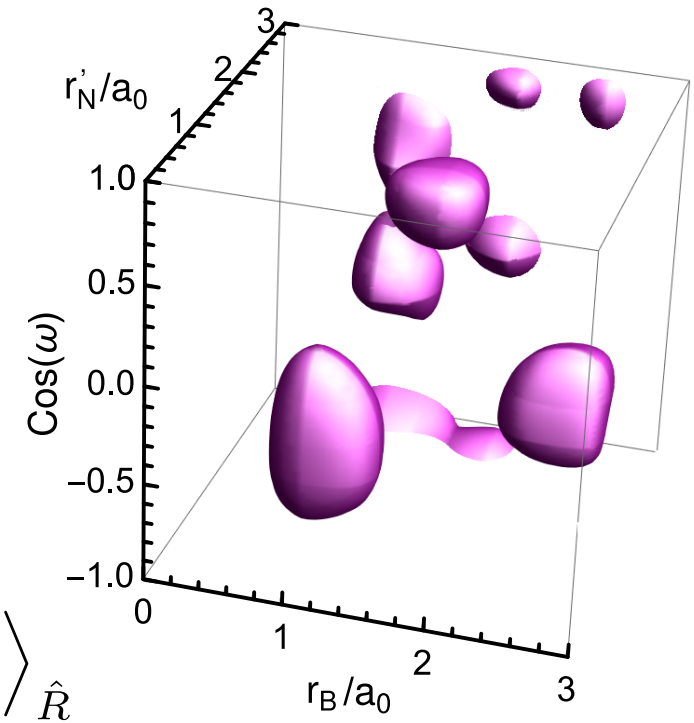
Willatt, Musil, **MC**, JCP (2019), arxiv:1807.00408

# Rotations, density correlations and potentials

- Incorporating rotations leads to a hierarchy of spherical invariants
- Invariant representations are essentially the same *n*-body correlations that are used in statistical theories of liquids
- Linear models built on $\left\langle \cdot \middle| \mathcal{X}^{(n-1)} \right\rangle_{g\to\delta}$ yield *n*-body potential expansion

$$V(\{\mathbf{r}_i\}) = \sum_{ij} V^{(2)}(r_{ij}) + \sum_{ij} V^{(3)}(r_{ij}, r_{ik}, \omega_{ijk}) \dots$$



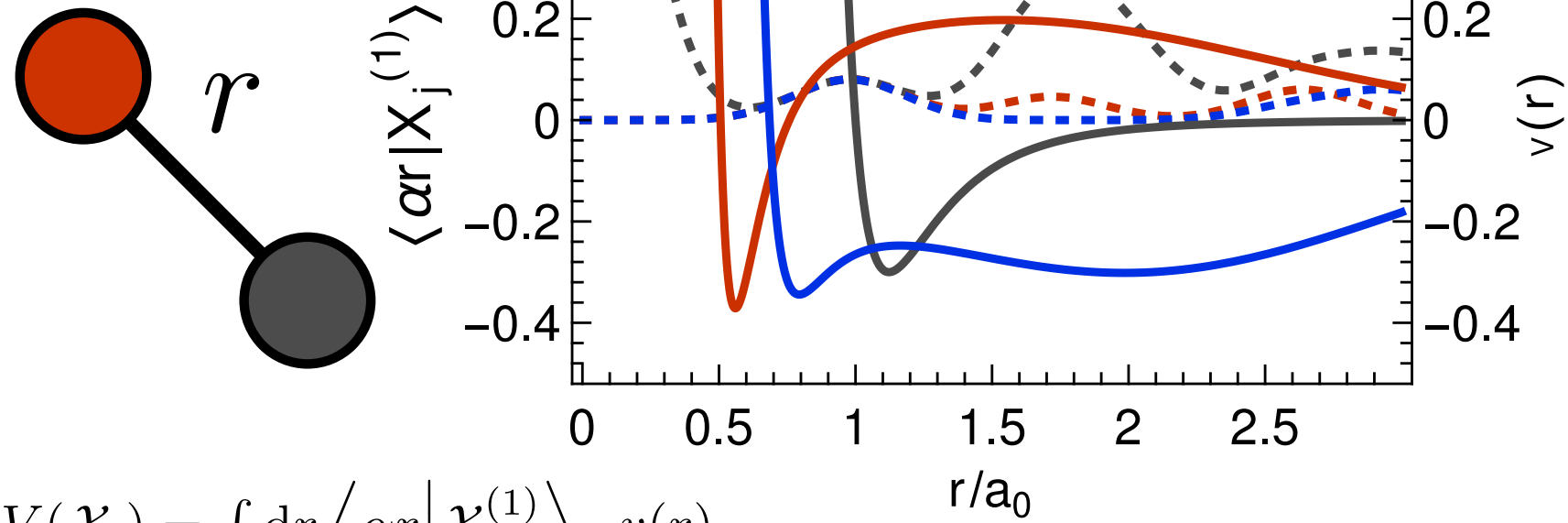

$$\left\langle \alpha r \alpha' r' \omega \middle| \mathcal{X}_j^{(2)} \right\rangle_{\hat{R}} = \int \mathrm{d}\hat{R} \left\langle \alpha r \hat{\mathbf{r}} \middle| \hat{R} \middle| \mathcal{X}_j \right\rangle \left\langle \alpha' r' \hat{\mathbf{r}}'(\omega) \middle| \hat{R} \middle| \mathcal{X}_j \right\rangle$$

Willatt, Musil, **MC**, JCP (2019); Drautz, PRB (2019); Glielmo, Zeni, De Vita, PRB (2018)

# Rotations, density correlations and potentials

- Incorporating rotations leads to a hierarchy of spherical invariants
- Invariant representations are essentially the same *n*-body correlations that are used in statistical theories of liquids
- Linear models built on $\left\langle \cdot \middle| \mathcal{X}^{(n-1)} \right\rangle_{g\to\delta}$ yield *n*-body potential expansion

$$V\left(\{\mathbf{r}_i\}\right) = \sum_{ij} V^{(2)}\left(r_{ij}\right) + \sum_{ij} V^{(3)}\left(r_{ij}, r_{ik}, \omega_{ijk}\right) \dots$$



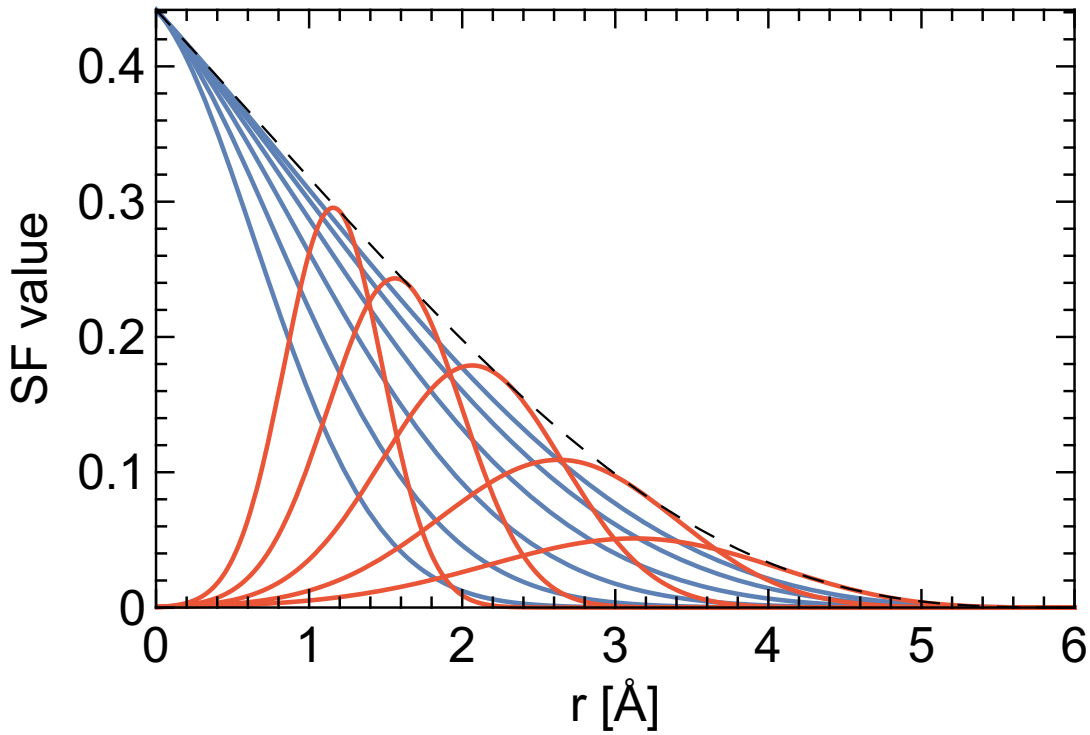

$$V(\mathcal{X}_j) = \int \mathrm{d}r \left\langle \alpha r \middle| \mathcal{X}_j^{(1)} \right\rangle_{\hat{R}} v(r)$$

Willatt, Musil, **MC**, JCP (2019); Drautz, PRB (2019); Glielmo, Zeni, De Vita, PRB (2018)

# Symmetry adapted representations & SOAP

- Most of the existing machine learning representations and kernels emerge as special cases of this framework
- Not necessary to use position basis. Radial functions and spherical harmonics $\rightarrow$ SOAP power spectrum and kernel

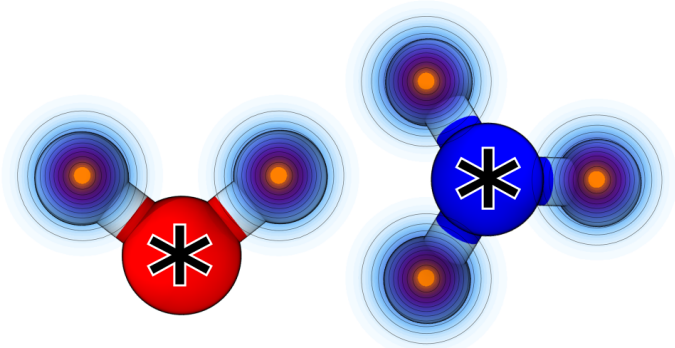
$$\langle \alpha\beta G_2 | \mathcal{X}_j \rangle = \langle \alpha | \alpha_j \rangle \int \mathrm{d}r \; G_2(r) \left\langle \beta r \middle| \mathcal{X}_j^{(1)} \right\rangle_{\hat{R}, g \rightarrow \delta}$$



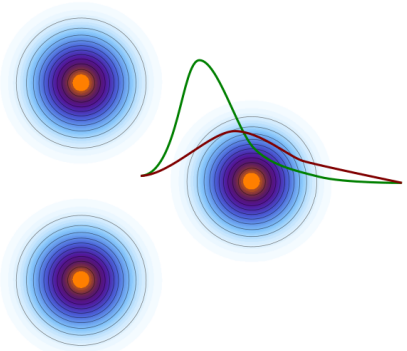

Behler, Parrinello, PRL (2007); Faber et al. JCP (2018)

# Symmetry adapted representations & SOAP

- Most of the existing machine learning representations and kernels emerge as special cases of this framework
- Not necessary to use position basis. Radial functions and spherical harmonics $\rightarrow$ SOAP power spectrum and kernel



$$\langle \mathbf{r}|\mathcal{X}_j\rangle = \psi(\mathbf{r}) = \sum_i g(\mathbf{r}-\mathbf{r}_{ij})$$

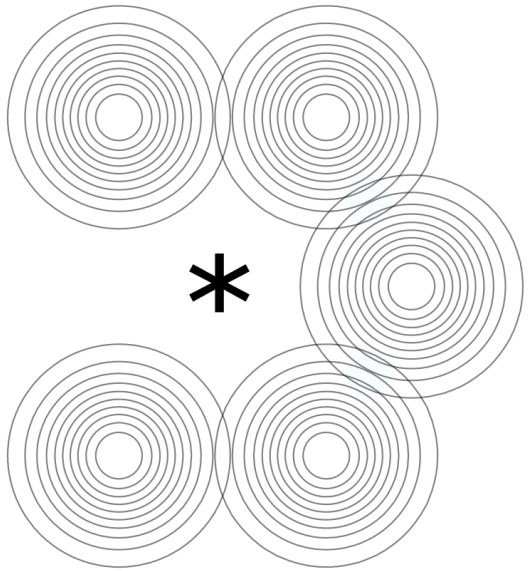


$$\langle nlm|\mathcal{X}_j\rangle = \int \mathrm{d}\mathbf{x}\, \psi(\mathbf{r}) R_n(r) Y_m^l(\hat{\mathbf{r}})$$

Bartók, Kondor, Csányi, PRB (2013); De, Bartók, Csányi, **MC**, PCCP (2016)

# Symmetry adapted representations & SOAP

- Most of the existing machine learning representations and kernels emerge as special cases of this framework
- Not necessary to use position basis. Radial functions and spherical harmonics $\rightarrow$ SOAP power spectrum and kernel



$$\left\langle nn'l\middle|\mathcal{X}^{(2)}\right\rangle_{\hat{R}} = \sum_m \langle nlm|\mathcal{X}\rangle \langle n'lm|\mathcal{X}\rangle$$

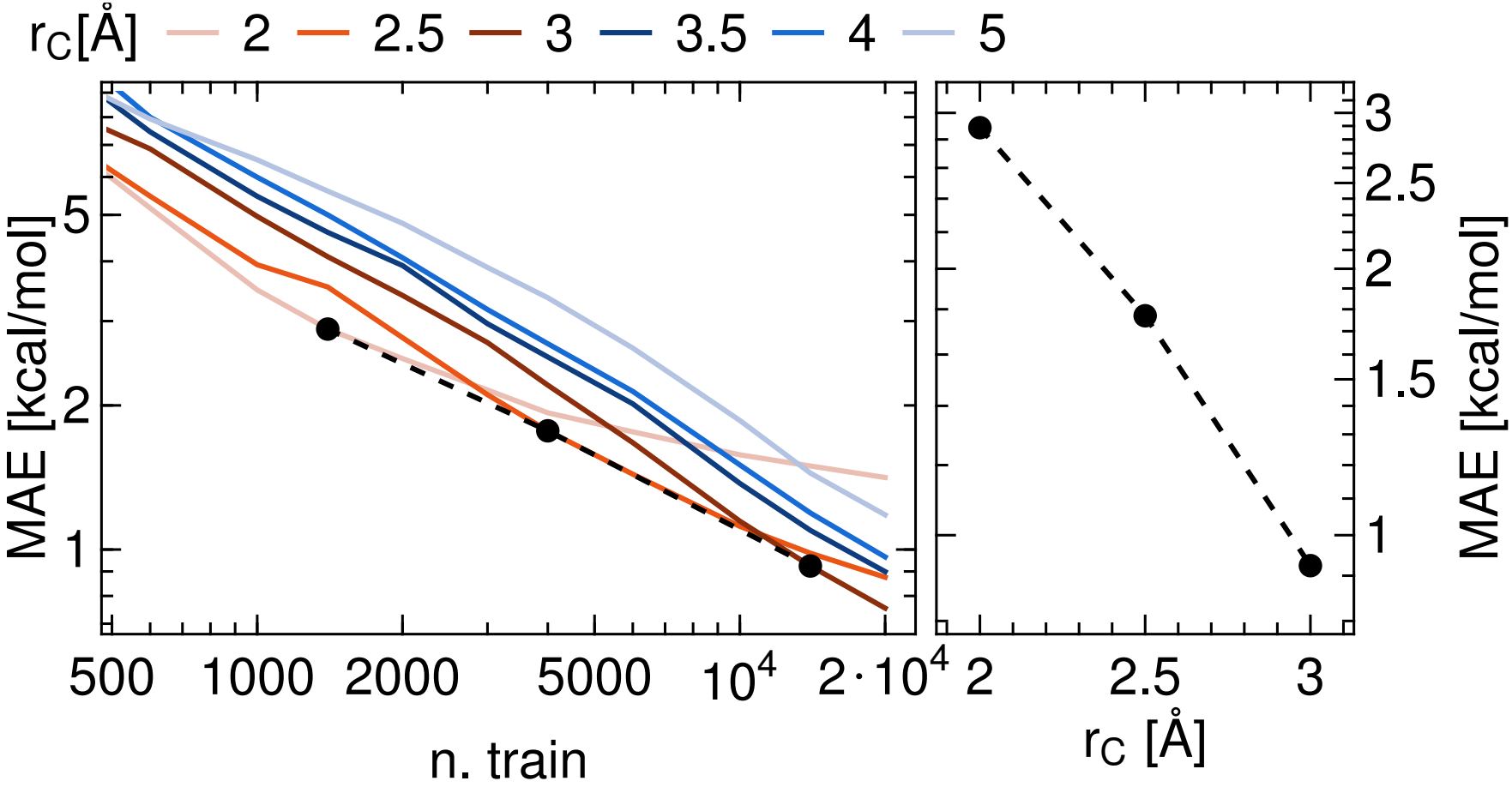
Bartók, Kondor, Csányi, PRB (2013); De, Bartók, Csányi, **MC**, PCCP (2016)

# Optimizing representations

## and what we learn in the process

# Understanding the range of interactions

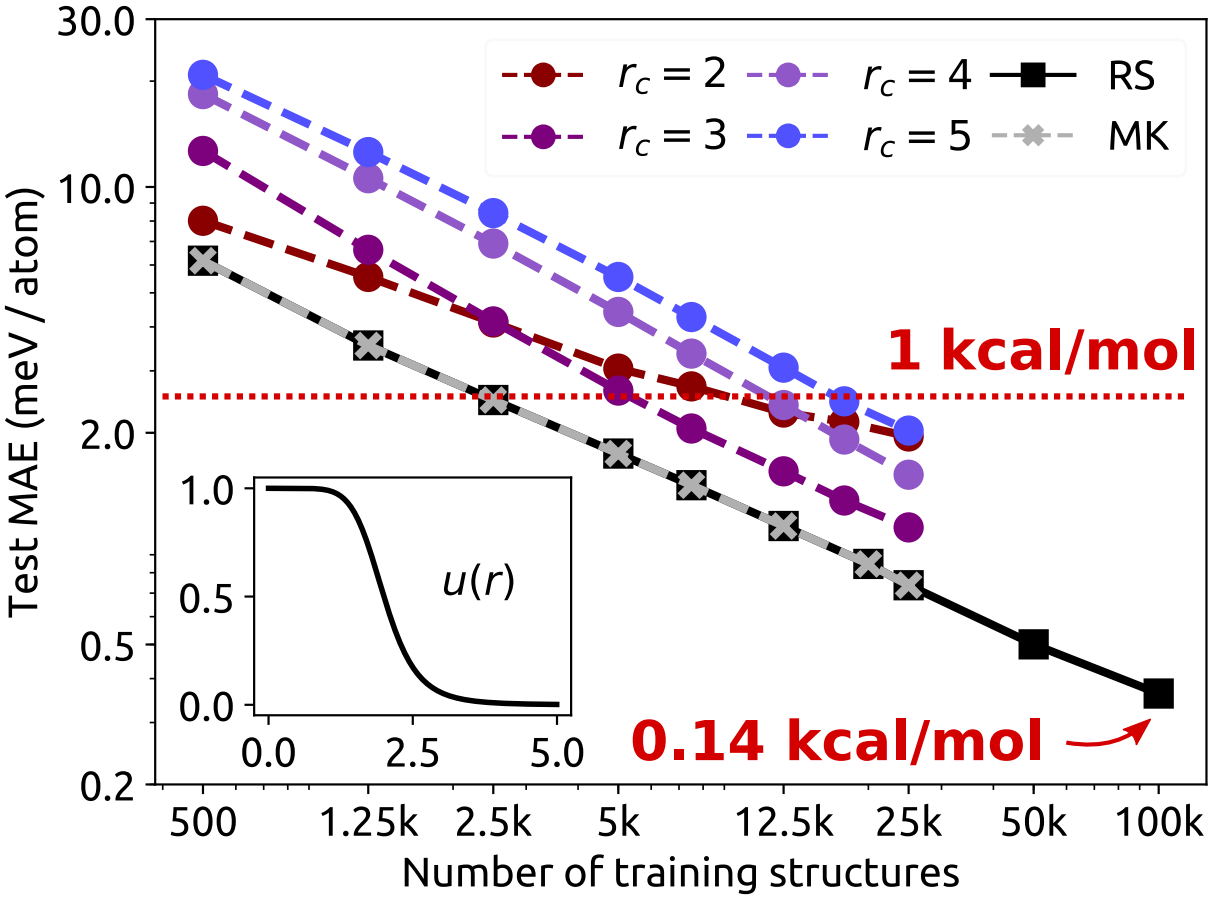
- Environment kernels can be built for different cutoff radii
- Dimensionality/accuracy tradeoff, a measure of the range of interactions
- A multi-scale kernel $K(A,B) = \sum_i w_i K_i(A,B)$ yields the best of all worlds.




Bartók, De, Poelking, Kermode, Bernstein, Csányi, **MC**, Science Advances (2017) *[data: QM9, von Lilienfeld&C]*

# Understanding the range of interactions

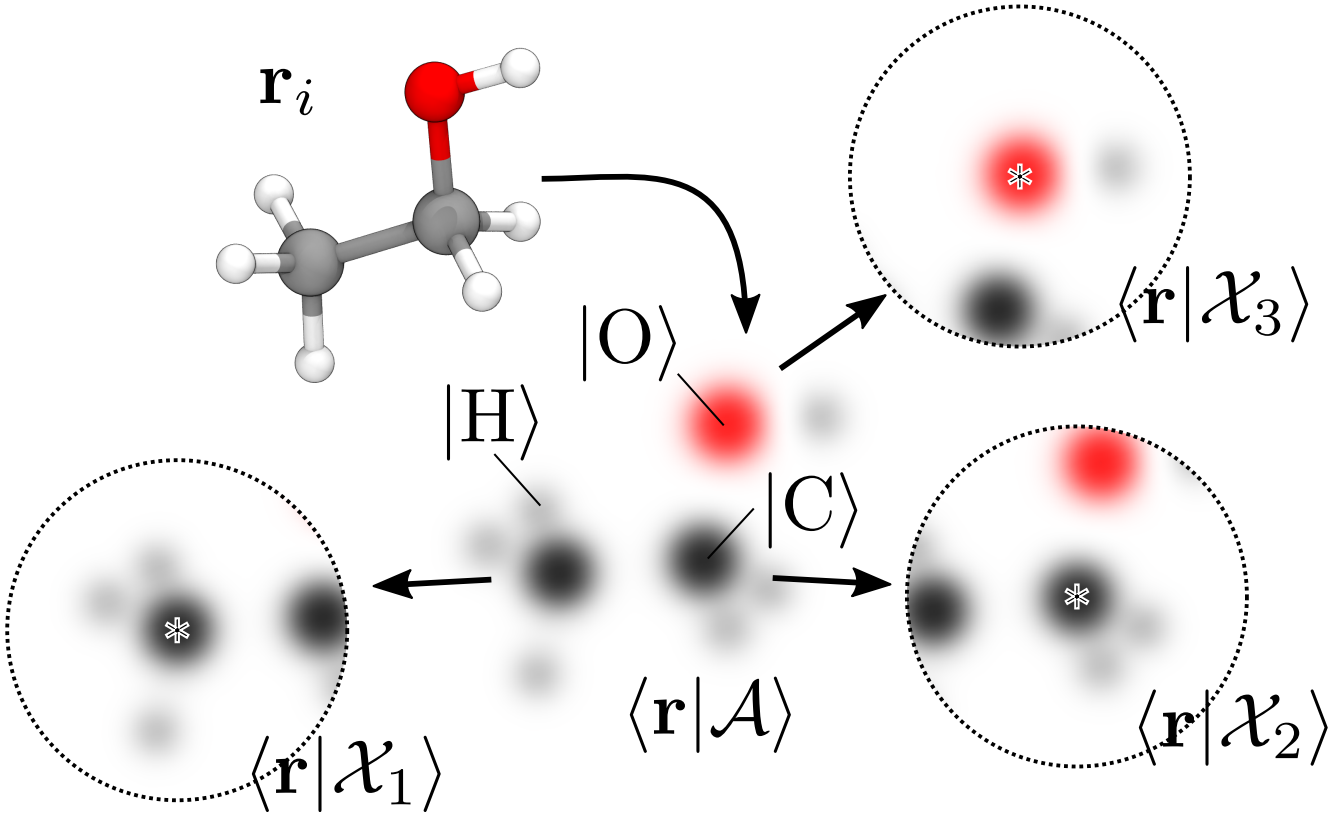
- Environment kernels can be built for different cutoff radii
- Dimensionality/accuracy tradeoff, a measure of the range of interactions
- A multi-scale kernel $K(A,B) = \sum_i w_i K_i(A,B)$ yields the best of all worlds.




Willatt, Musil, **MC**, PCCP (2018)

# A data-driven periodic table of the elements

- How to learn with multiple species? Decorate atomic Gaussian with elemental kets $|\mathrm{H}\rangle$, $|\mathrm{O}\rangle$, . . .
- Expand each ket in a finite basis, $|\alpha\rangle = \sum_J u_{\alpha J} |J\rangle$. Optimize coefficients
- Dramatic reduction of the descriptor space, more effective learning . . .
- . . . and as by-product get a data-driven version of the periodic table!



# A data-driven periodic table of the elements

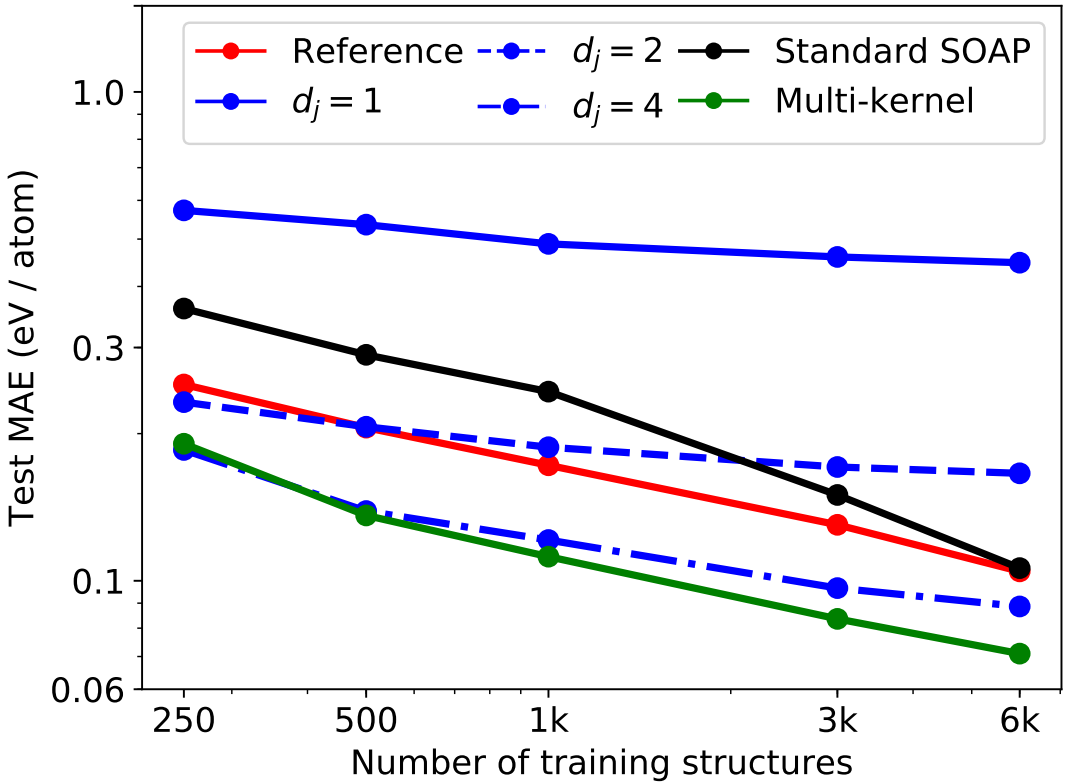
- How to learn with multiple species? Decorate atomic Gaussian with elemental kets $|\mathrm{H}\rangle$, $|\mathrm{O}\rangle$, . . .
- Expand each ket in a finite basis, $|\alpha\rangle = \sum_J u_{\alpha J} |J\rangle$. Optimize coefficients
- Dramatic reduction of the descriptor space, more effective learning . . .
- . . . and as by-product get a data-driven version of the periodic table!

$$|\mathrm{H}\rangle = 0.5\,|🔥\rangle + 0.1\,|🪨\rangle + 0.2\,|💧\rangle$$
$$|\mathrm{C}\rangle = 0.2\,|🔥\rangle + 0.8\,|🪨\rangle + 0.3\,|💧\rangle$$
$$|\mathrm{O}\rangle = 0.1\,|🔥\rangle + 0.1\,|🪨\rangle + 0.6\,|💧\rangle$$

Empedocles et al. (ca 360BC). Metaphor courtesy of Albert Bartók

# A data-driven periodic table of the elements

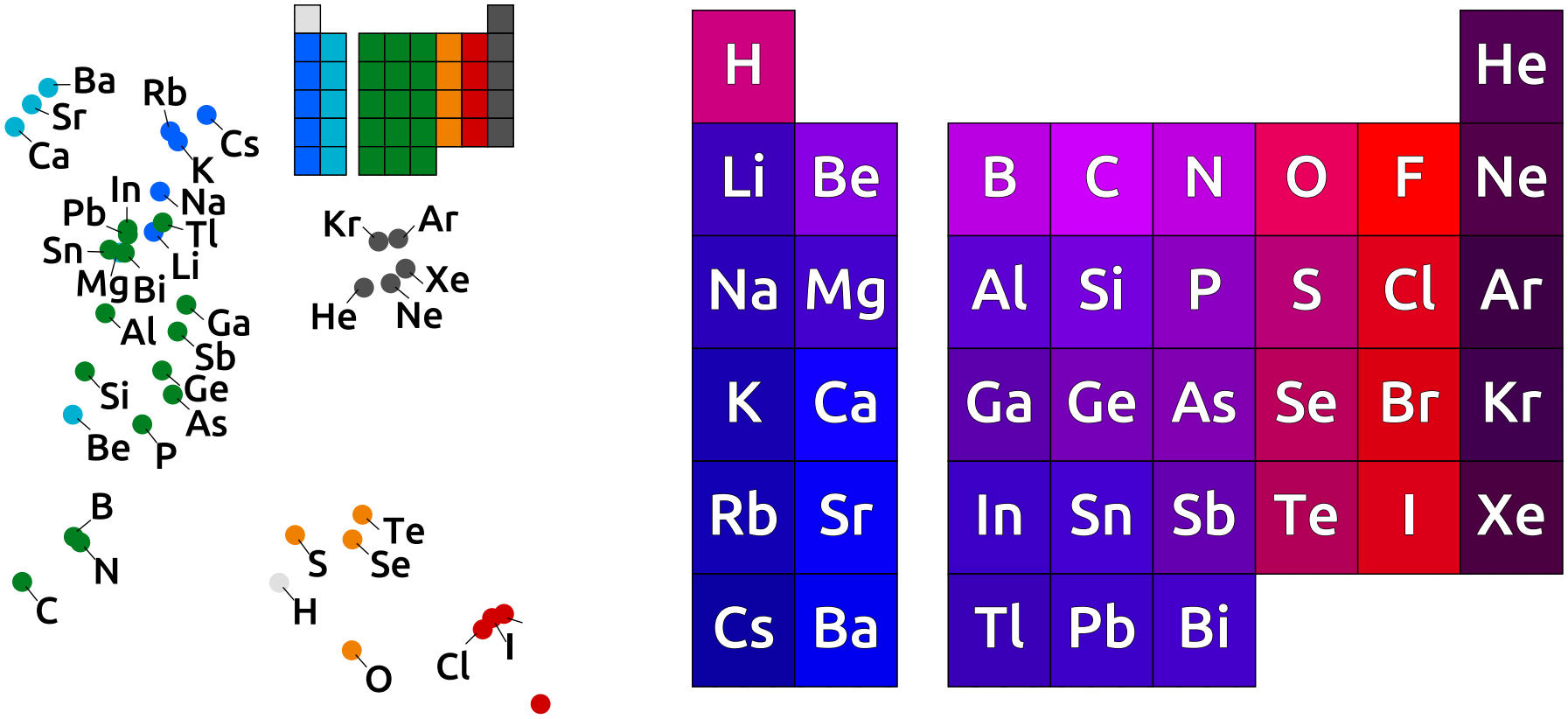
- How to learn with multiple species? Decorate atomic Gaussian with elemental kets $|\mathrm{H}\rangle$, $|\mathrm{O}\rangle$, . . .
- Expand each ket in a finite basis, $|\alpha\rangle = \sum_J u_{\alpha J} |J\rangle$. Optimize coefficients
- Dramatic reduction of the descriptor space, more effective learning . . .
- . . . and as by-product get a data-driven version of the periodic table!




Willatt, Musil, **MC**, PCCP (2018); *[data: Elpasolites, von Lilienfeld&C]*

# A data-driven periodic table of the elements

- How to learn with multiple species? Decorate atomic Gaussian with elemental kets $|\mathrm{H}\rangle$, $|\mathrm{O}\rangle$, . . .
- Expand each ket in a finite basis, $|\alpha\rangle = \sum_J u_{\alpha J} |J\rangle$. Optimize coefficients
- Dramatic reduction of the descriptor space, more effective learning . . .
- . . . and as by-product get a data-driven version of the periodic table!

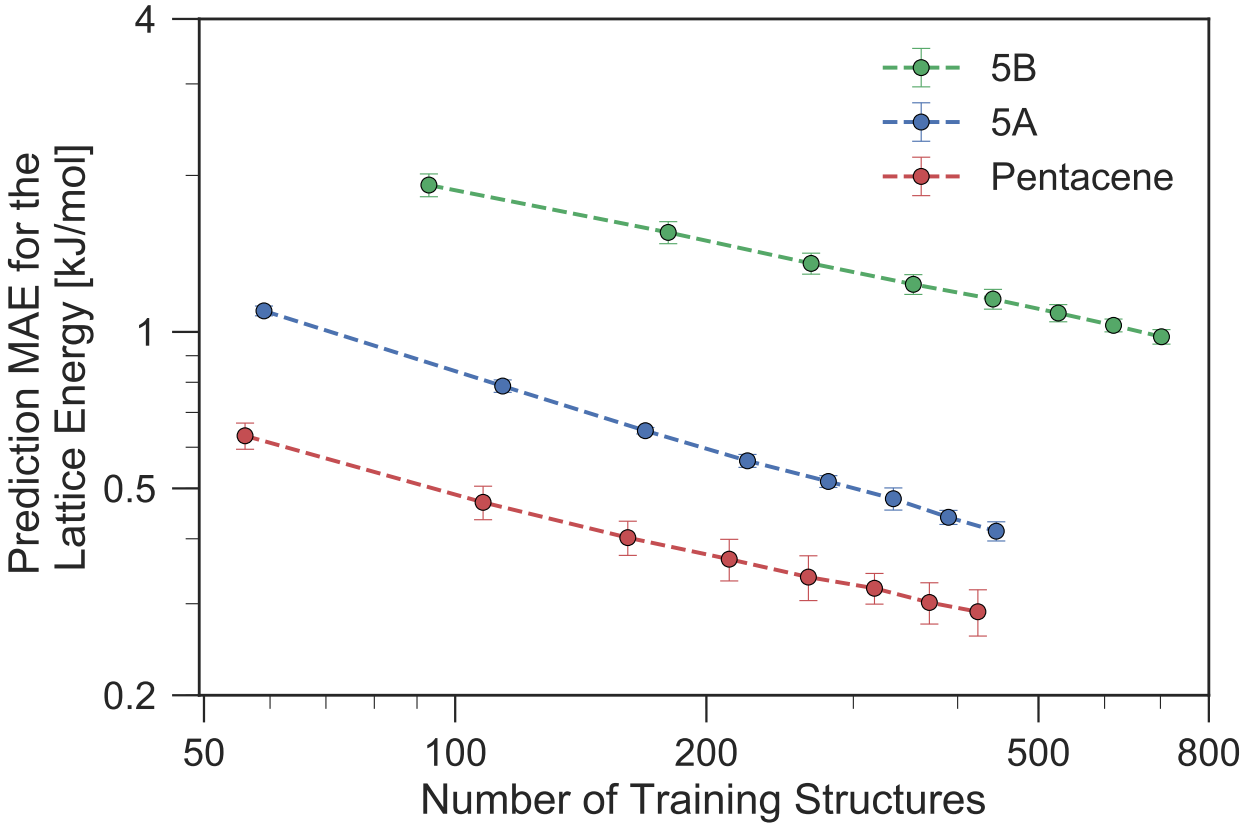


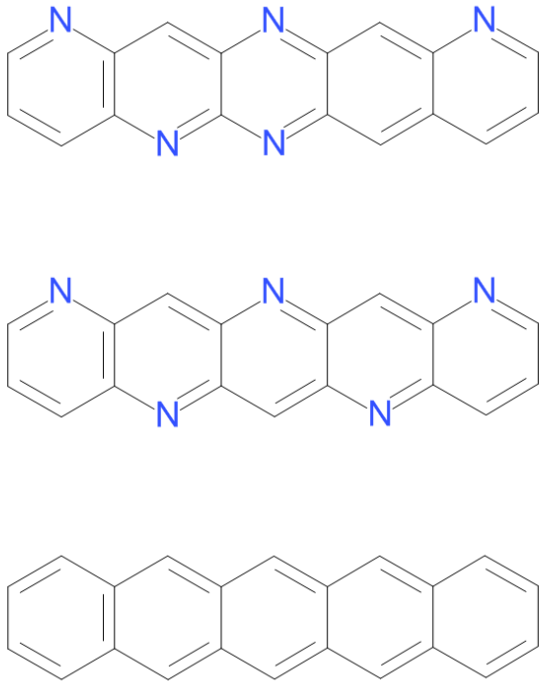

Willatt, Musil, **MC**, PCCP (2018); *[data: Elpasolites, von Lilienfeld&C]*

# A representation for all seasons

# Accurate predictions for molecular crystals

- Substituted pentacenes - model systems for molecular electronics
- Easily achieve sub-kcal/mol accuracy, with REMatch-SOAP kernels

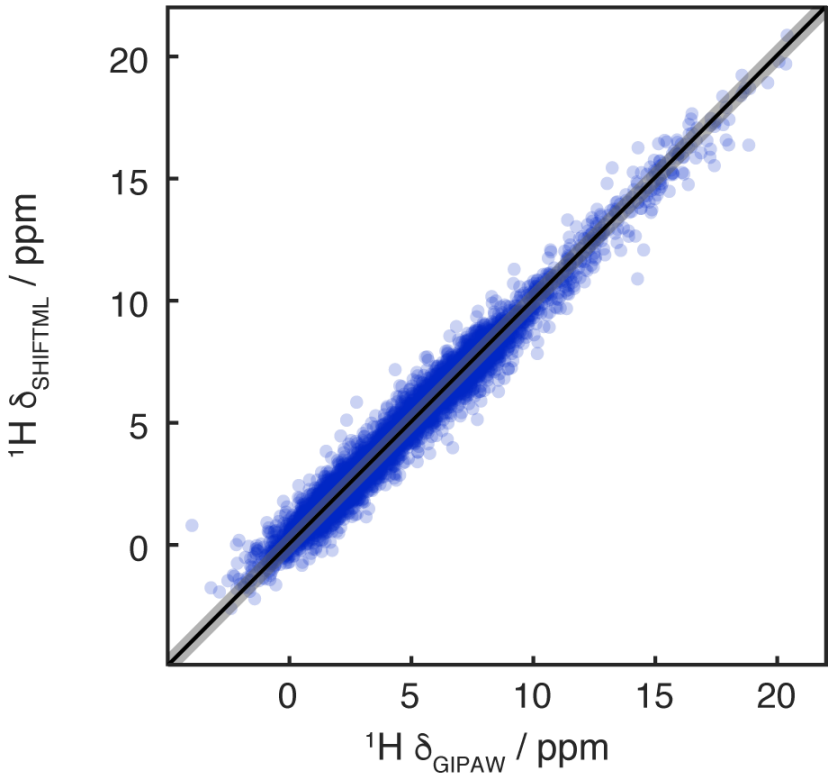


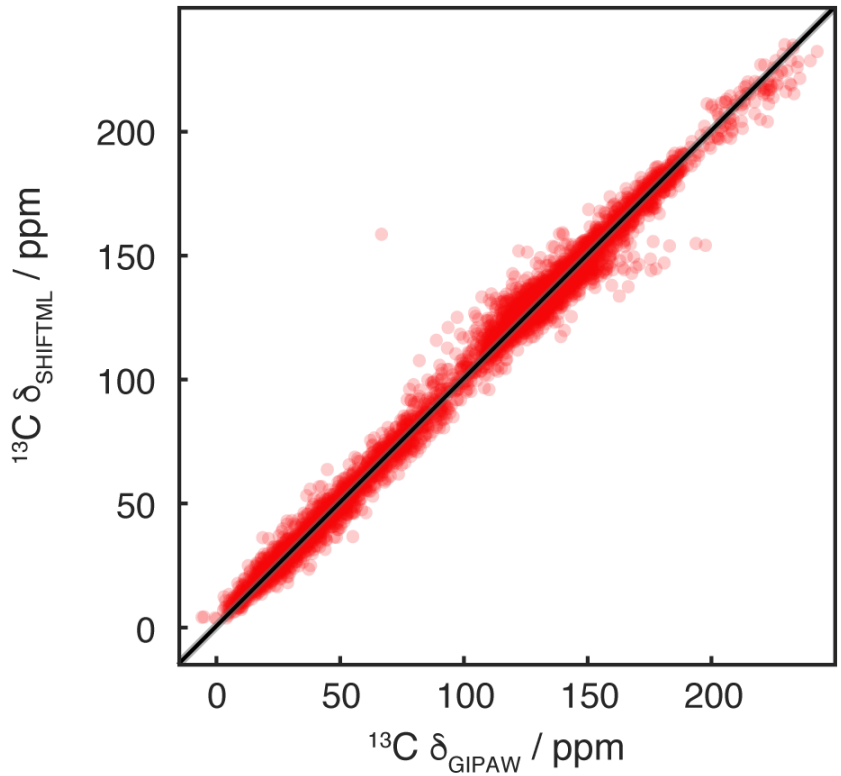




Musil, De, Yang, Campbell, Day, **MC**, Chemical Science (2018) *[data: G.Day, J.Yang]*

# More than interatomic potentials

- Solid-state NMR relies on GIPAW-DFT to determine crystal structure of molecular materials
- Train a ML model on 2000 CSD structures, predict chemical shieldings with DFT accuracy (RMSE H: 0.5, C: 5, N: 13, O: 18 ppm)
- Precise enough to do structure determination!

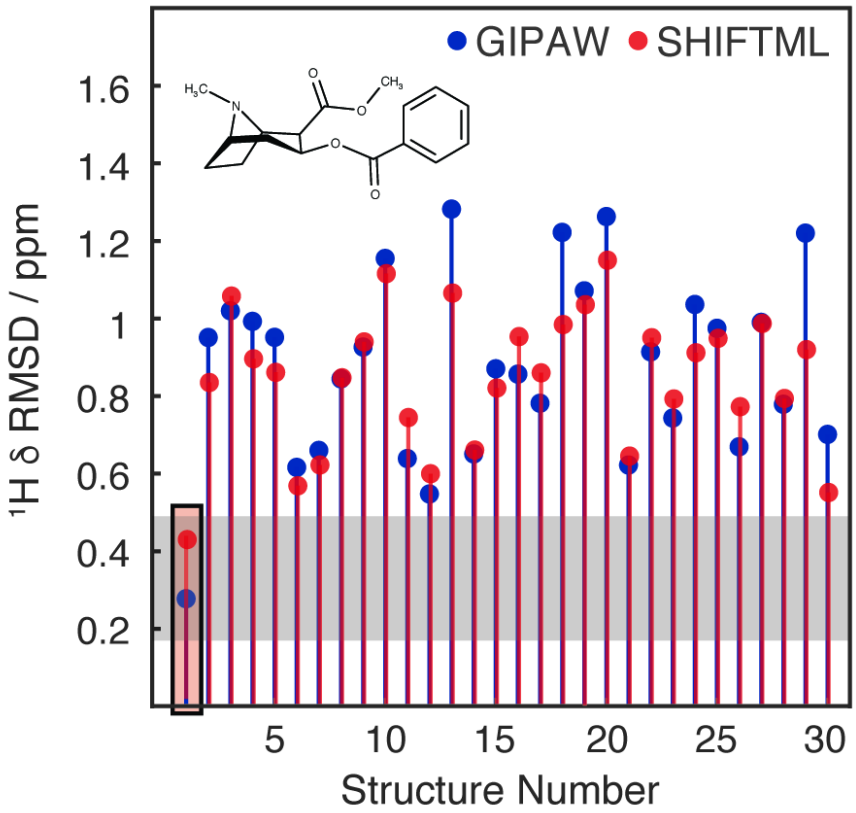


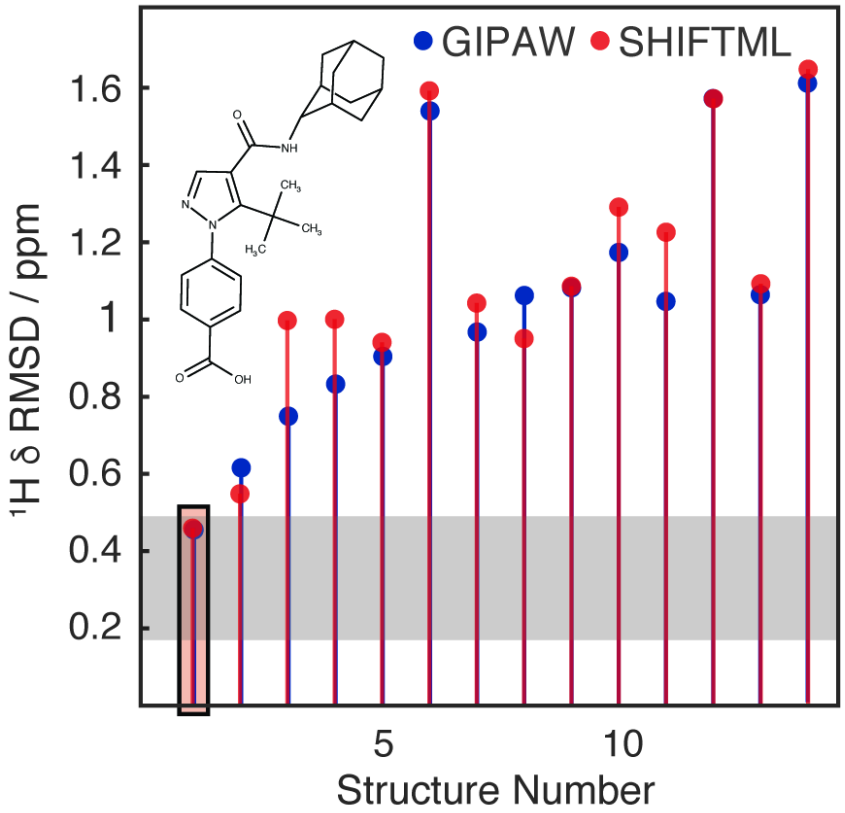




Paruzzo, Hofstetter, Musil, De, **MC**, Emsley, Nature Comm. (2018); **http://shiftml.org** *[data: CSD-500]*

# More than interatomic potentials

- Solid-state NMR relies on GIPAW-DFT to determine crystal structure of molecular materials
- Train a ML model on 2000 CSD structures, predict chemical shieldings with DFT accuracy (RMSE H: 0.5, C: 5, N: 13, O: 18 ppm)
- Precise enough to do structure determination!



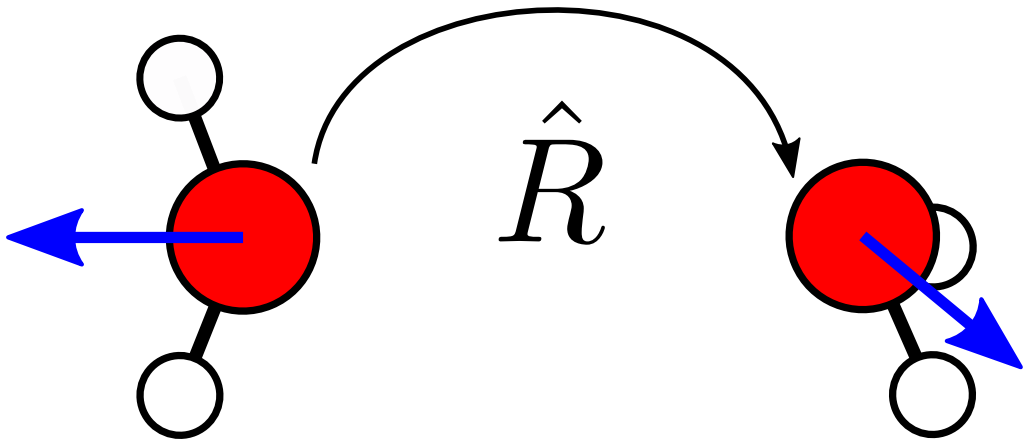




Paruzzo, Hofstetter, Musil, De, **MC**, Emsley, Nature Comm. (2018); **http://shiftml.org** *[data: CSD-500]*

**There are more things in heaven and earth, Horatio, than those transforming like a scalar**

# Machine-learning for tensors

- Vectors and tensors are ubiquitous in the description of crystalline materials ($\alpha_{xy}$, $\mu_x$, . . . )
- Formulate density-based symmetrized features that learn efficiently by incorporating the geometric transformations of tensors

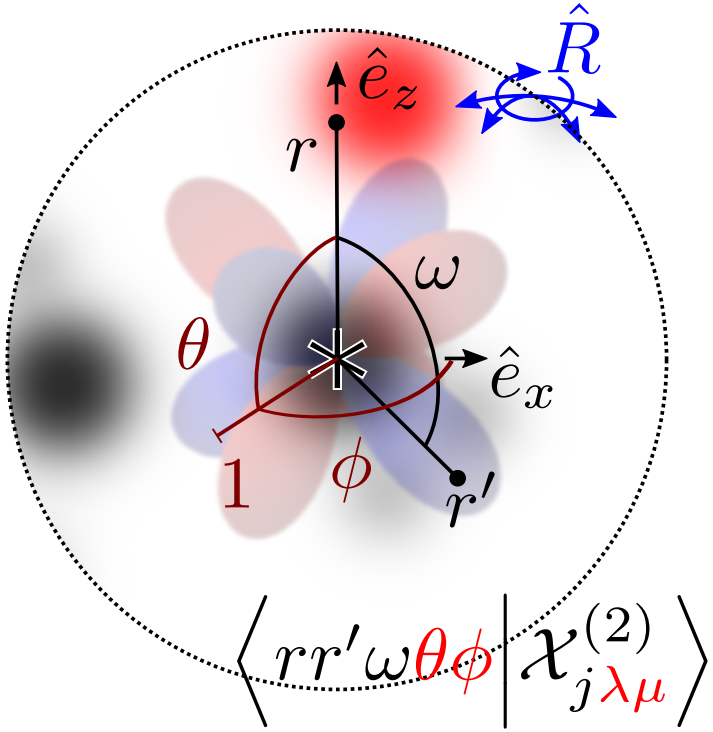



Glielmo, Sollich, & De Vita, PRB (2017); Grisafi, Wilkins, Csányi, & **MC**, PRL (2018)

# Machine-learning for tensors

- Vectors and tensors are ubiquitous in the description of crystalline materials ($\alpha_{xy}$, $\mu_x$,. . . )
- Formulate density-based symmetrized features that learn efficiently by incorporating the geometric transformations of tensors

$$T^{\mu}_{\lambda} \rightarrow \left| \mathcal{X}^{(2)}_{j} \lambda\mu \right\rangle$$

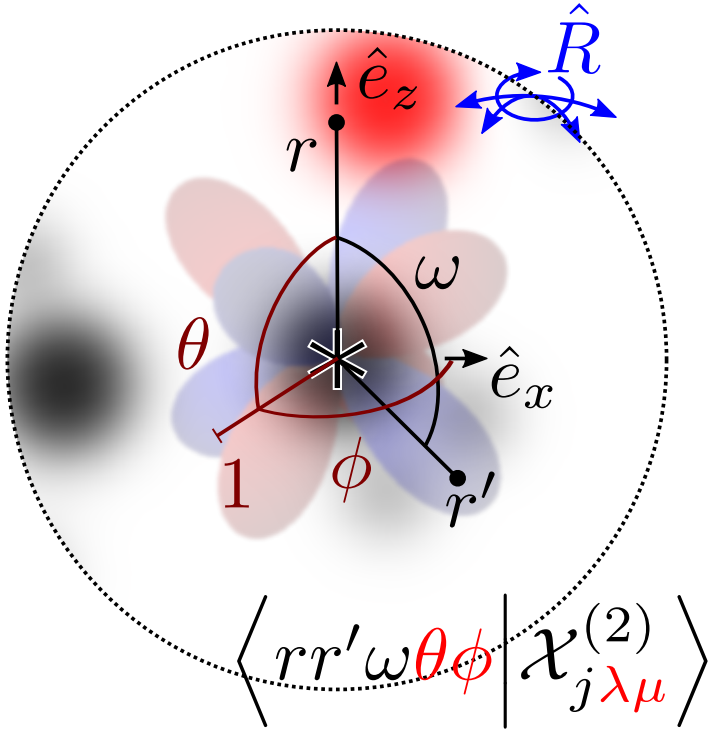



Willatt, Musil, & **MC**, JCP (2019); Grisafi, Wilkins, Csányi, & **MC**, PRL (2018)

# Machine-learning for tensors

- Vectors and tensors are ubiquitous in the description of crystalline materials ($\alpha_{xy}$, $\mu_x$,. . . )
- Formulate density-based symmetrized features that learn efficiently by incorporating the geometric transformations of tensors

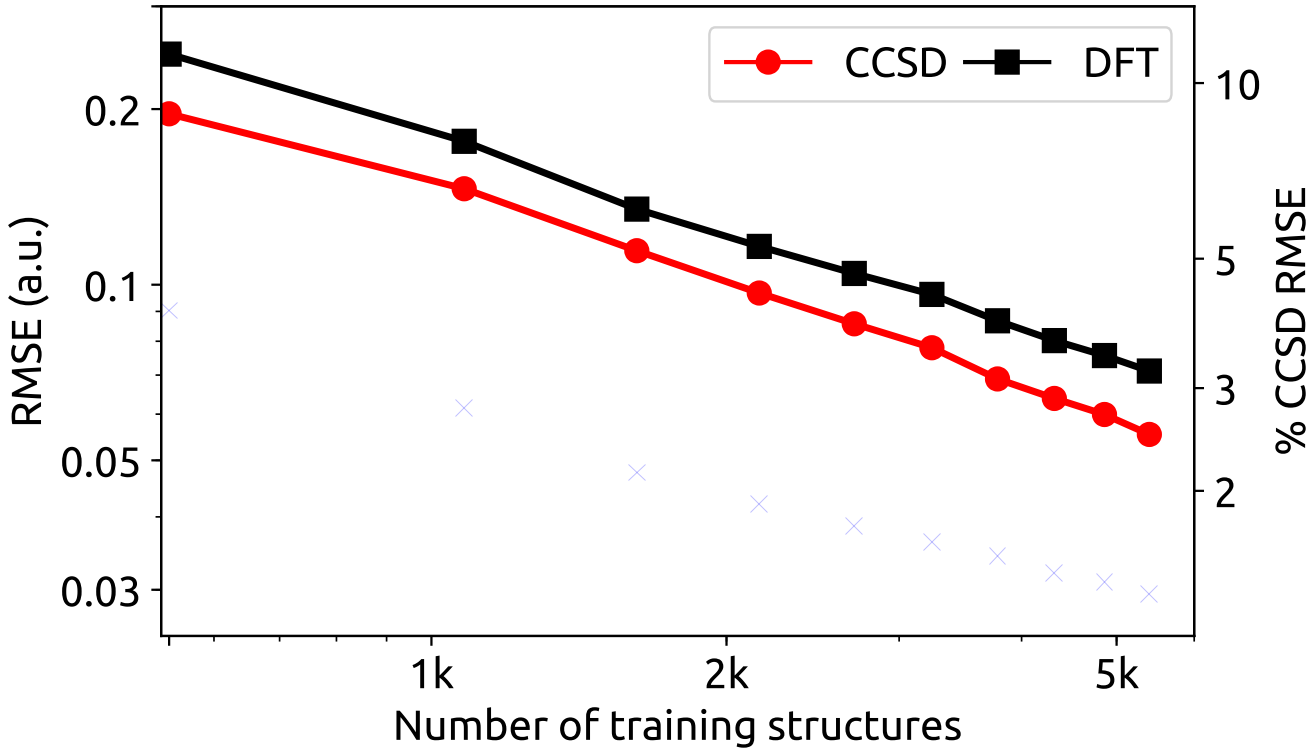
$$\int \mathrm{d}\hat{R}\, \langle \mathbf{r}|\, \hat{R}\, |\mathcal{X}_j\rangle\, \langle \mathbf{r}'|\, \hat{R}\, |\mathcal{X}_j\rangle\, \langle \mathbf{r}''|\, \hat{R}\, |\lambda\mu\rangle \rightarrow \left\langle rr'\omega\,\theta\phi \middle| \mathcal{X}_j^{(2)}\,\lambda\mu \right\rangle$$




Willatt, Musil, & **MC**, JCP (2019); Grisafi, Wilkins, Csányi, & **MC**, PRL (2018)

# Molecular polarizabilities at the CCSD level

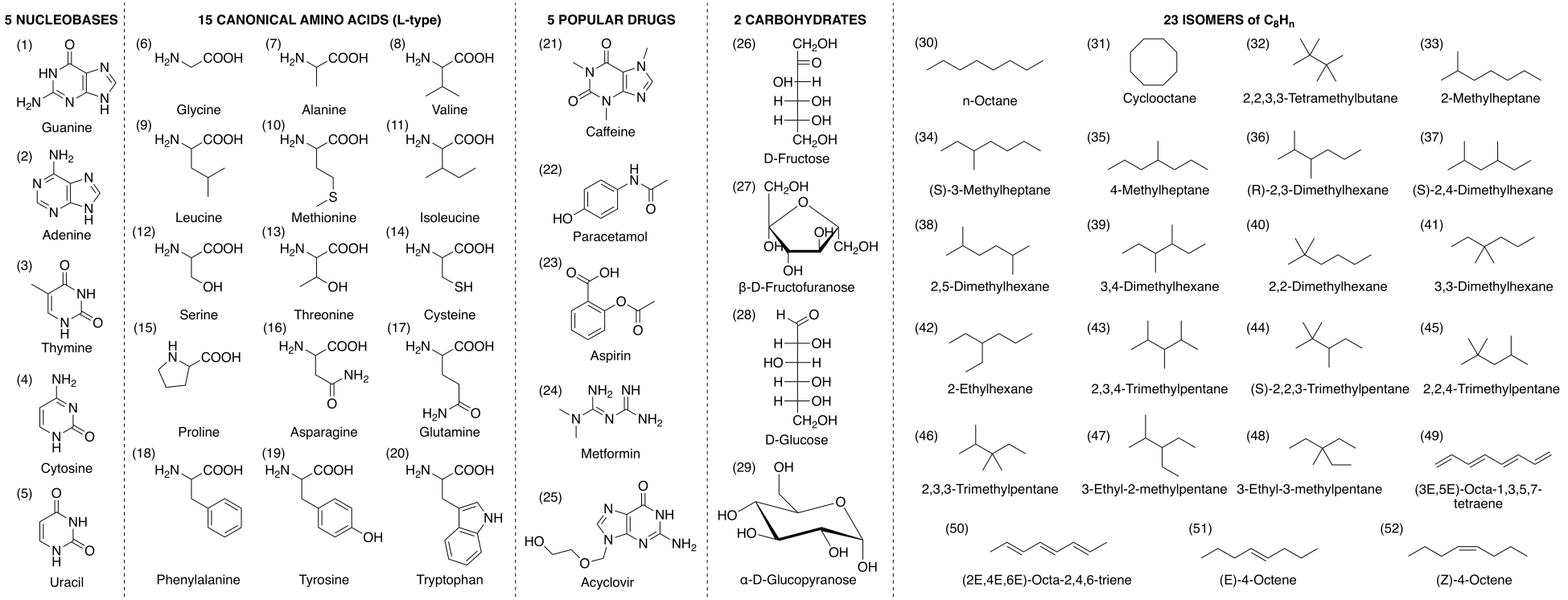
- DFT is not very accurate for the dielectric response. Train a ML model (AlphaML) on a dataset of small molecules with CCSD accuracy
- The model can extrapolate to much large compounds (up to aciclovir $C_8H_{11}N_5O_3$) with better-than-DFT accuracy




Wilkins, Grisafi, Yang, Lao, DiStasio, **MC**, PNAS (2019); *[data: 10.24435/materialscloud:2019.0002/v2]*

# Molecular polarizabilities at the CCSD level

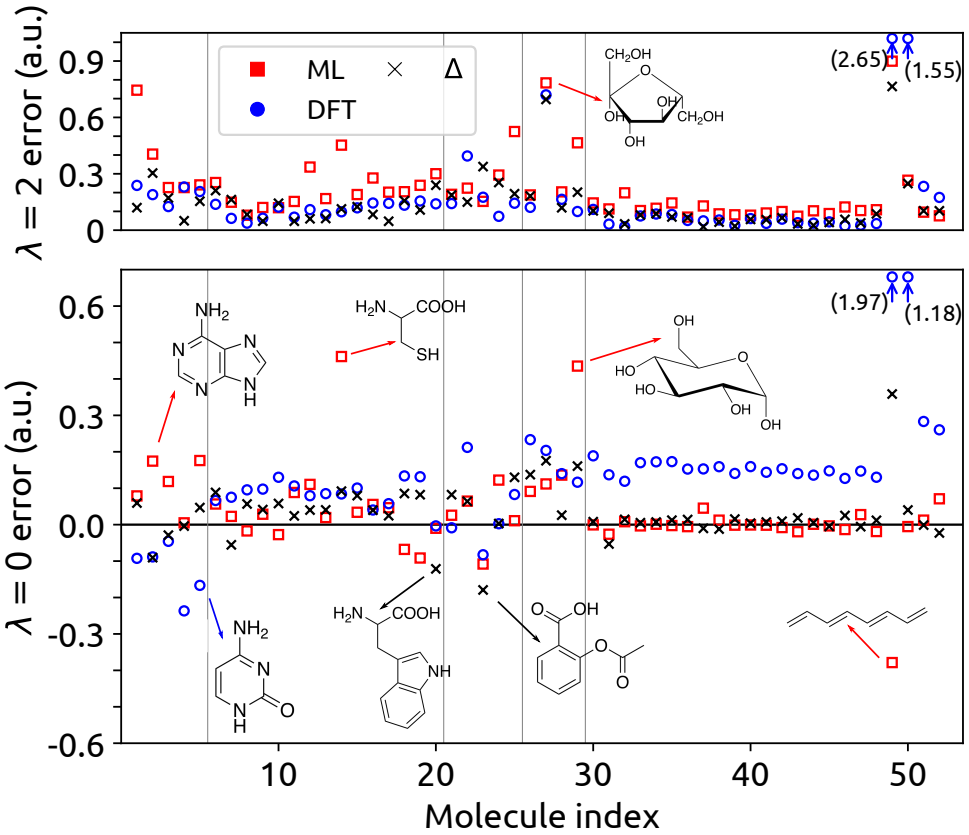
- DFT is not very accurate for the dielectric response. Train a ML model (AlphaML) on a dataset of small molecules with CCSD accuracy
- The model can extrapolate to much large compounds (up to aciclovir $C_8H_{11}N_5O_3$) with better-than-DFT accuracy




Wilkins, Grisafi, Yang, Lao, DiStasio, **MC**, PNAS (2019); **`http://alphaml.org`**

# Molecular polarizabilities at the CCSD level

- DFT is not very accurate for the dielectric response. Train a ML model (AlphaML) on a dataset of small molecules with CCSD accuracy
- The model can extrapolate to much large compounds (up to aciclovir $C_8H_{11}N_5O_3$) with better-than-DFT accuracy

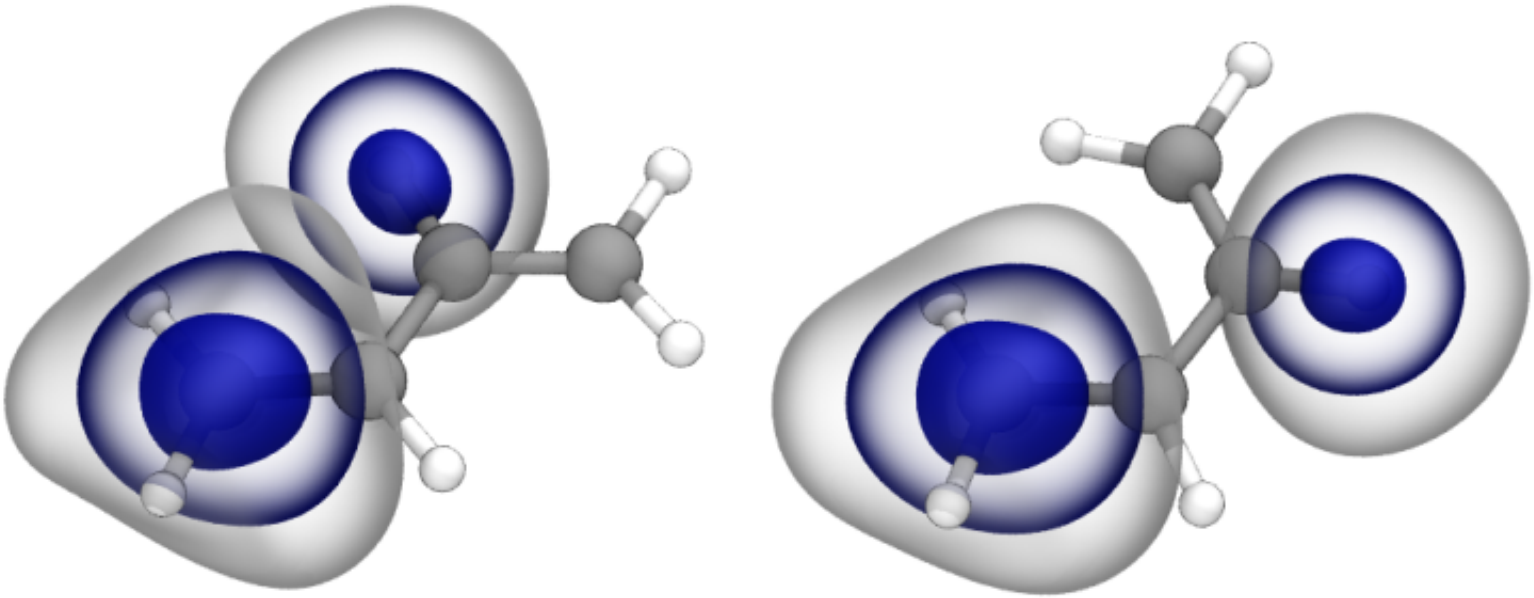



| Method | RMSE |
|---|---|
| CCSD/ML | 0.304 |
| CCSD/DFT | 0.573 |

Wilkins, Grisafi, Yang, Lao, DiStasio, **MC**, PNAS (2019); **http://alphaml.org**

# A transferable model of the electron density

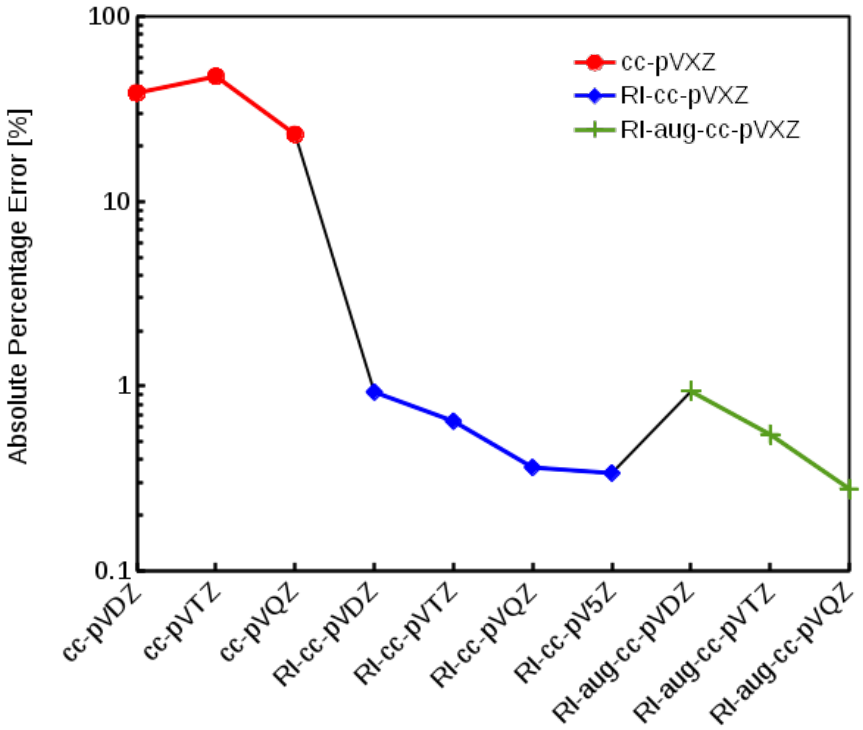
- Write the charge density in atom-centered components.
- Expand on an atomic basis $\phi_k \equiv R_n Y_l^m \rightarrow$ tensorial learning of coefficients
- Transferable enough to predict the density of polypeptides



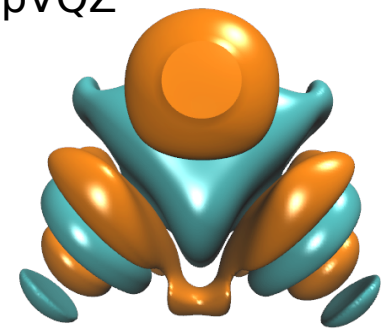
Grisafi, Wilkins, Meyer, Fabrizio, Corminboeuf, **MC**, ACS Central Science (2019)

# A transferable model of the electron density

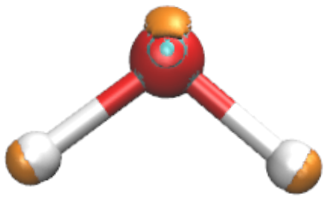
- Write the charge density in atom-centered components.
- Expand on an atomic basis $\phi_k \equiv R_n Y_l^m \rightarrow$ tensorial learning of coefficients
- Transferable enough to predict the density of polypeptides




cc-pVQZ



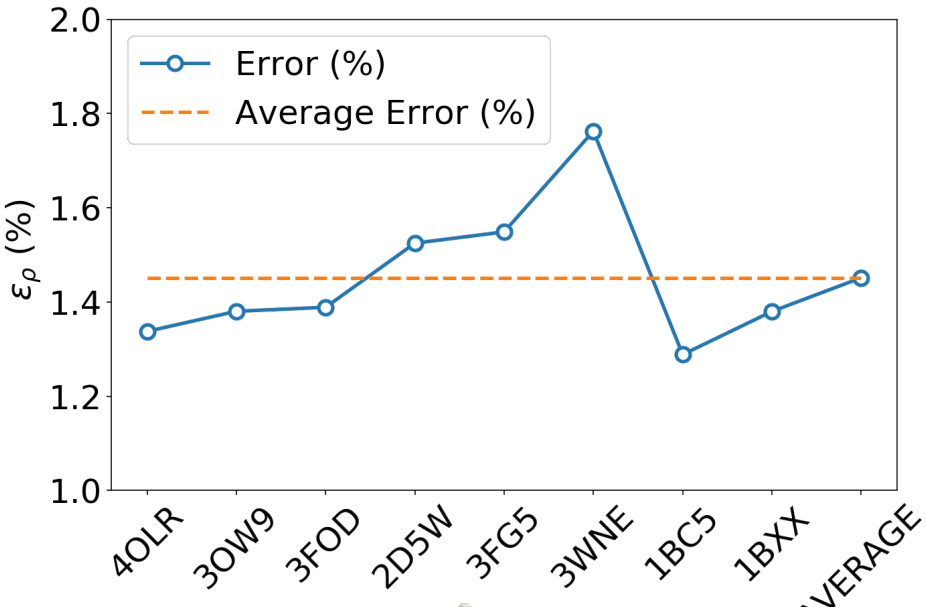
RI-cc-pVQZ

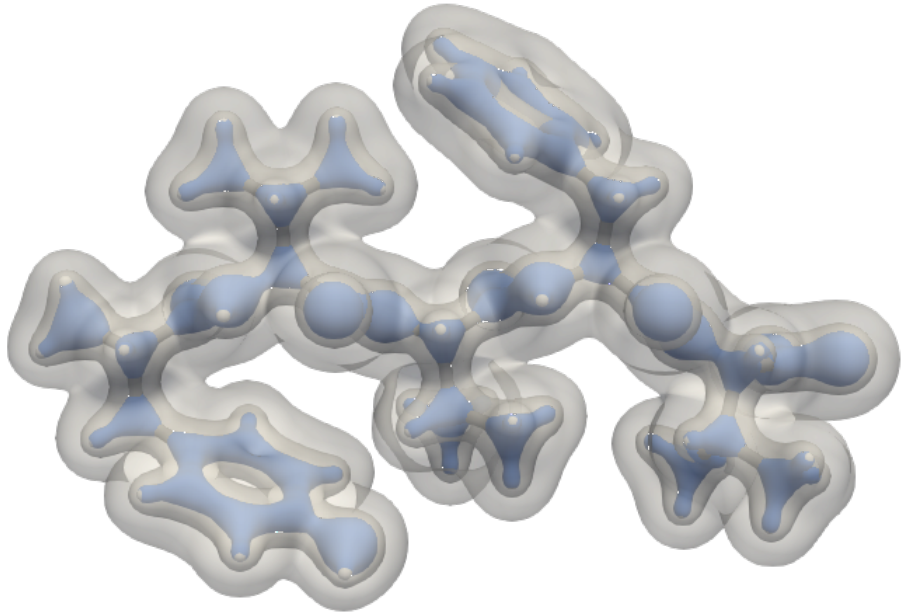


Grisafi, Wilkins, Meyer, Fabrizio, Corminboeuf, **MC**, ACS Central Science (2019)

# A transferable model of the electron density

- Write the charge density in atom-centered components.
- Expand on an atomic basis $\phi_k \equiv R_n Y_l^m \rightarrow$ tensorial learning of coefficients
- Transferable enough to predict the density of polypeptides



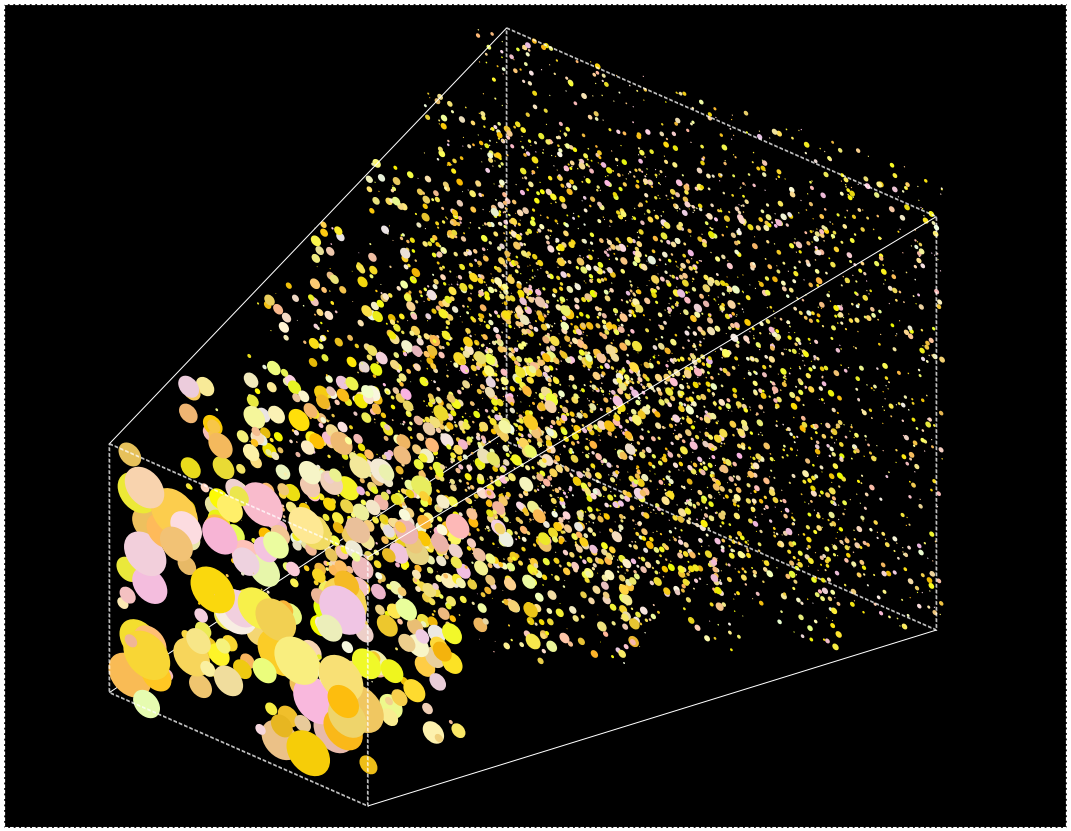



Meyer, Grisafi, Fabrizio, **MC**, Corminboeuf, Chem. Sci., (2019)

# The charged elephant in the other room

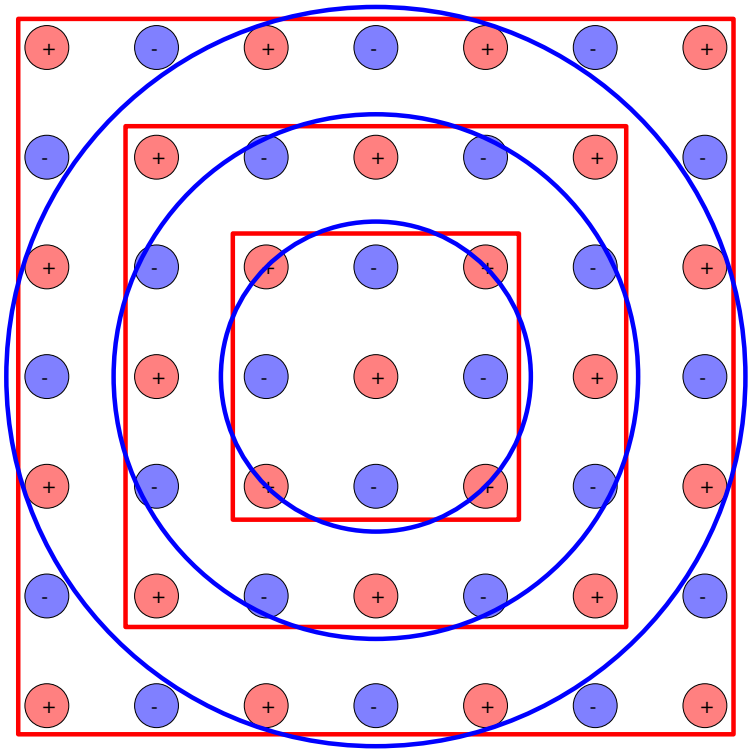
# The night sky paradox

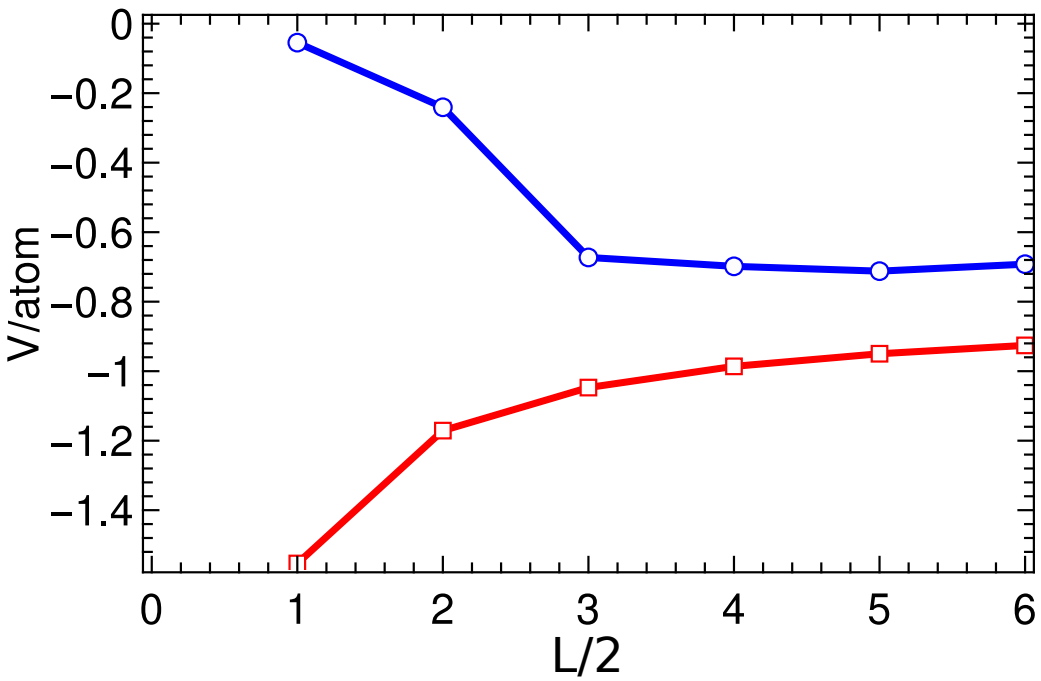
- Light intensity decays as $1/r^2$, but number of stars increases as $r^2$, so each layer contribute a constant intensity: the night sky should be bright!
- For the same reason electrostatic interactions, that decay as $1/r$, are a challenge (also) for ML



# The night sky paradox

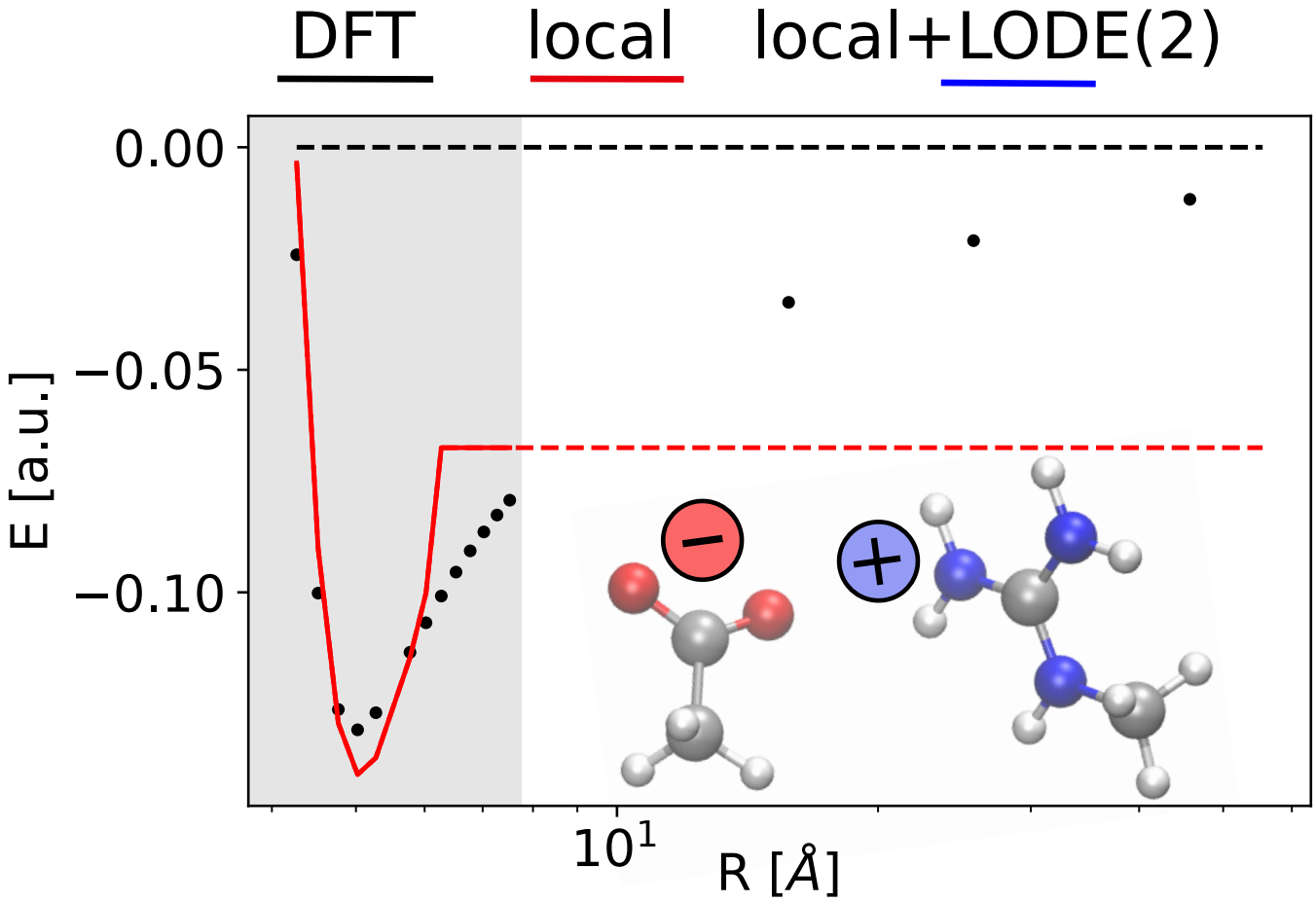
- Light intensity decays as $1/r^2$, but number of stars increases as $r^2$, so each layer contribute a constant intensity: the night sky should be bright!
- For the same reason electrostatic interactions, that decay as $1/r$, are a challenge (also) for ML





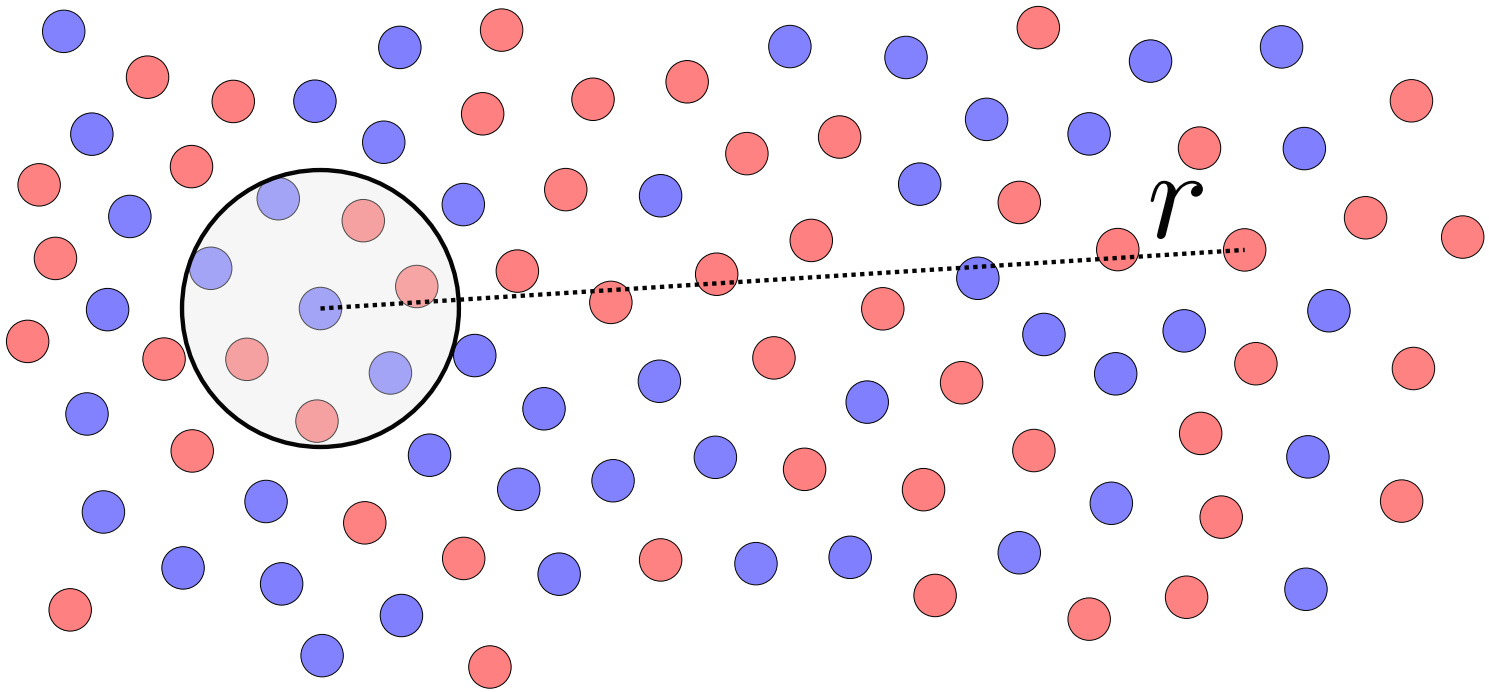
# The night sky paradox

- Light intensity decays as $1/r^2$, but number of stars increases as $r^2$, so each layer contribute a constant intensity: the night sky should be bright!
- For the same reason electrostatic interactions, that decay as $1/r$, are a challenge (also) for ML



# Long-distance equivariant representation

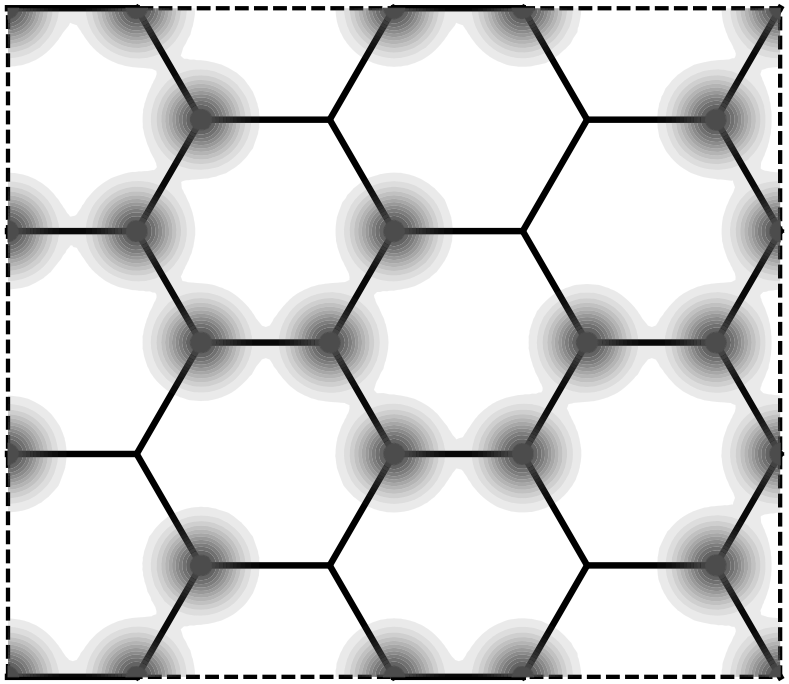
- Idea: get a local representation that reflects long-range correlations, with proper asymptotics
  1. Define an atom-density potential $\langle \alpha \mathbf{r} | \mathcal{V}^p \rangle = \int \langle \alpha \mathbf{r}' | \mathcal{A} \rangle \, / \, |\mathbf{r}' - \mathbf{r}|^p \, \mathrm{d}\mathbf{r}$. One channel per atomic species
  2. Do the usual gig: symmetrize, decompose, learn!
- Can be computed efficiently in reciprocal space



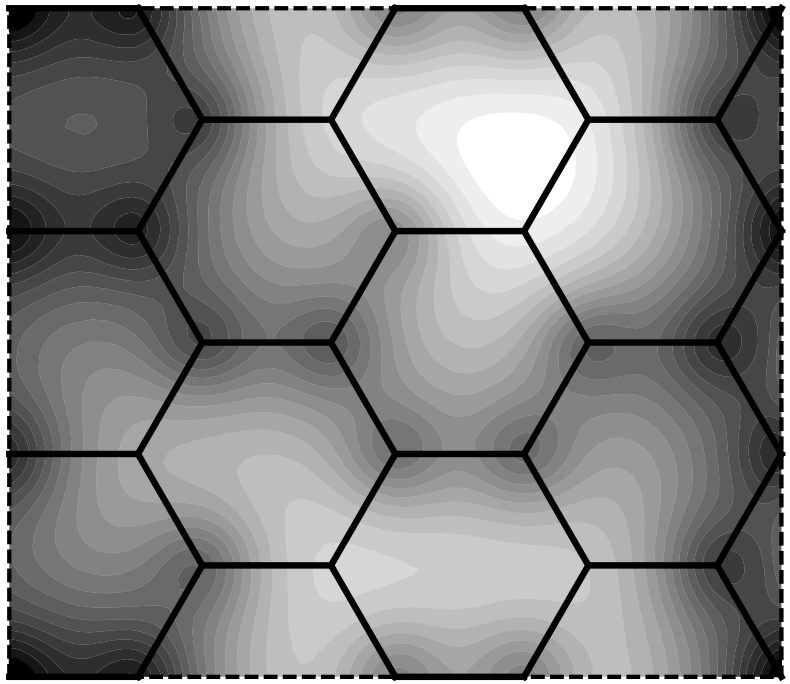

Grisafi, **MC**, arxiv:1909.04512

# Long-distance equivariant representation

- Idea: get a local representation that reflects long-range correlations, with proper asymptotics
  1. Define an atom-density potential $\langle \alpha \mathbf{r} | \mathcal{V}^p \rangle = \int \langle \alpha \mathbf{r}' | \mathcal{A} \rangle \, / \, |\mathbf{r}' - \mathbf{r}|^p \, \mathrm{d}\mathbf{r}$. One channel per atomic species
  2. Do the usual gig: symmetrize, decompose, learn!
- Can be computed efficiently in reciprocal space



$$\langle \mathbf{r} | \mathcal{A} \rangle = \sum_i g(\mathbf{r} - \mathbf{r_i}) \, |\alpha_i\rangle$$

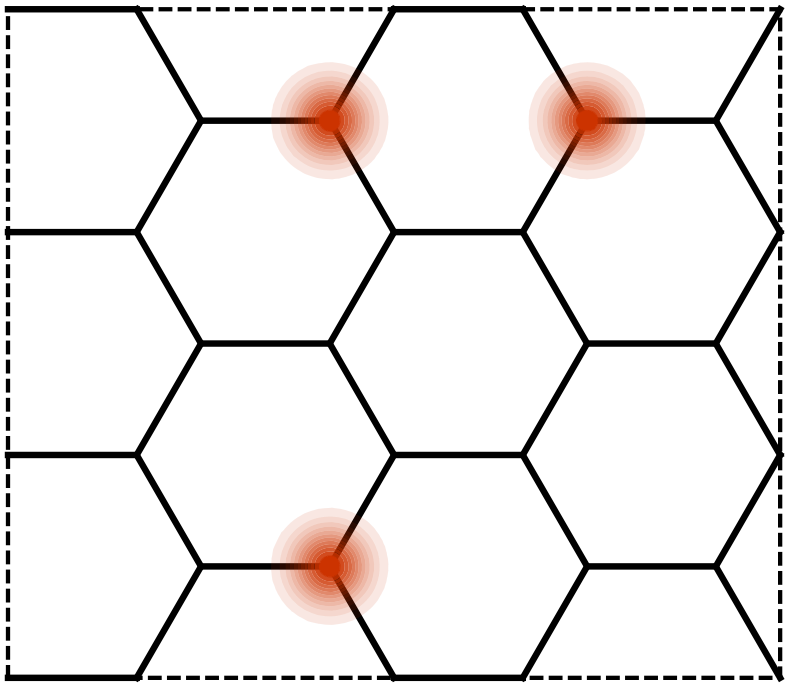


$$\left\langle \mathbf{r} \middle| \mathcal{V}^1 \right\rangle = \int \langle \mathbf{r}' | \mathcal{A} \rangle \, / \, |\mathbf{r}' - \mathbf{r}| \, \mathrm{d}\mathbf{r}'$$

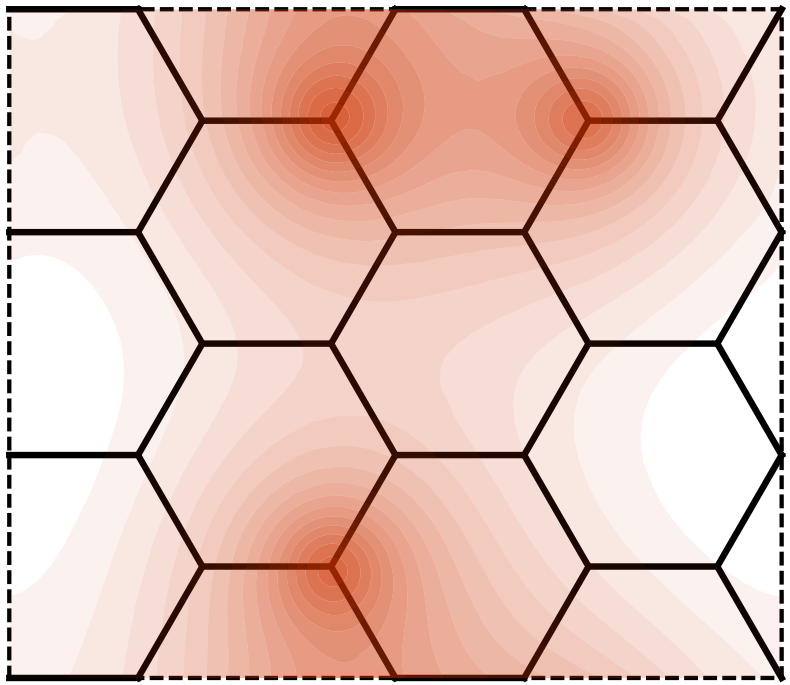
Grisafi, **MC**, arxiv:1909.04512

# Long-distance equivariant representation

- Idea: get a local representation that reflects long-range correlations, with proper asymptotics
  1. Define an atom-density potential $\langle \alpha \mathbf{r} | \mathcal{V}^p \rangle = \int \langle \alpha \mathbf{r}' | \mathcal{A} \rangle \, / \, |\mathbf{r}' - \mathbf{r}|^p \, \mathrm{d}\mathbf{r}$. One channel per atomic species
  2. Do the usual gig: symmetrize, decompose, learn!
- Can be computed efficiently in reciprocal space



$$\langle \mathbf{r} | \mathcal{A} \rangle = \sum_i g(\mathbf{r} - \mathbf{r_i}) \, | \alpha_i \rangle$$

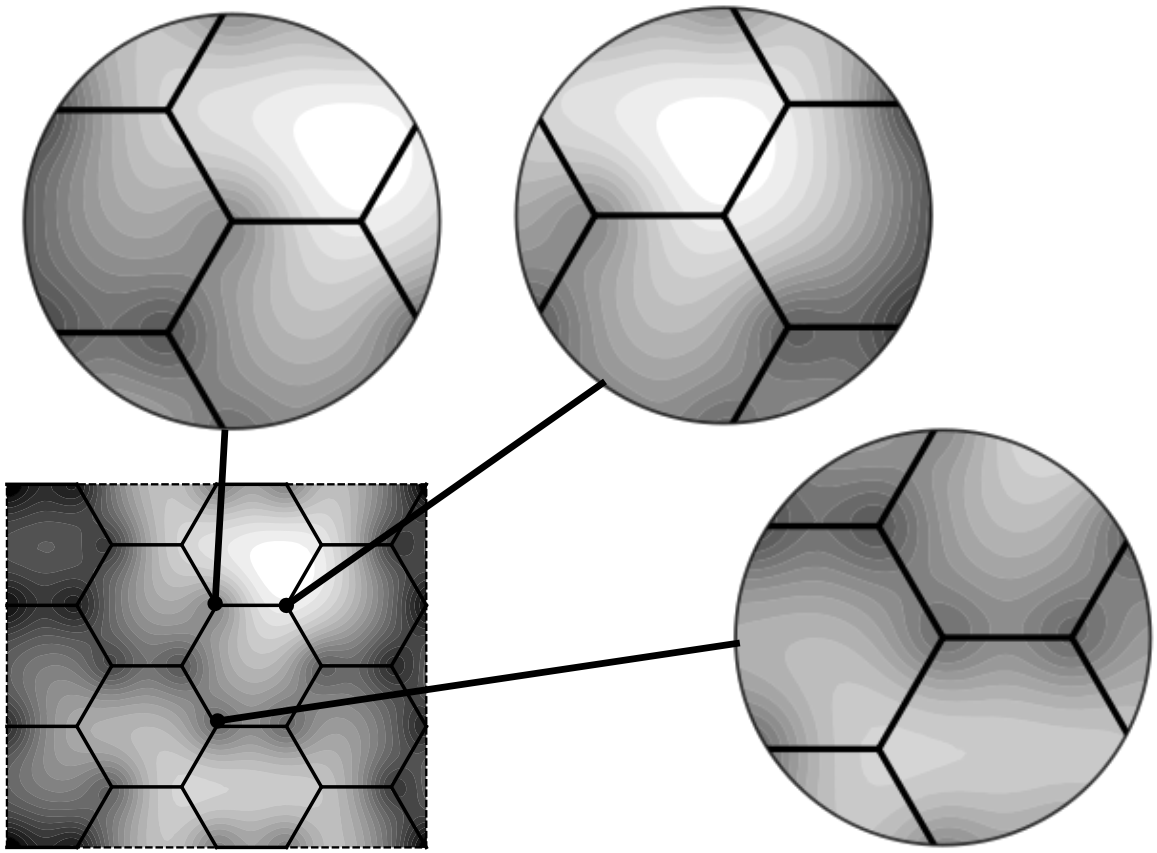


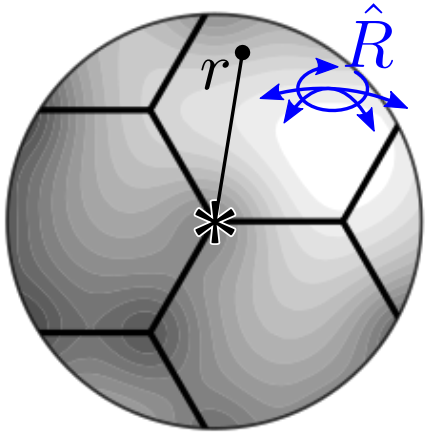
$$\left\langle \mathbf{r} \middle| \mathcal{V}^1 \right\rangle = \int \langle \mathbf{r}' | \mathcal{A} \rangle \, / \, |\mathbf{r}' - \mathbf{r}| \, \mathrm{d}\mathbf{r}'$$

Grisafi, **MC**, arxiv:1909.04512

# Long-distance equivariant representation

- Idea: get a local representation that reflects long-range correlations, with proper asymptotics
  1. Define an atom-density potential $\langle \alpha \mathbf{r} | \mathcal{V}^p \rangle = \int \langle \alpha \mathbf{r}' | \mathcal{A} \rangle \, / \, |\mathbf{r}' - \mathbf{r}|^p \, d\mathbf{r}$. One channel per atomic species
  2. Do the usual gig: symmetrize, decompose, learn!
- Can be computed efficiently in reciprocal space

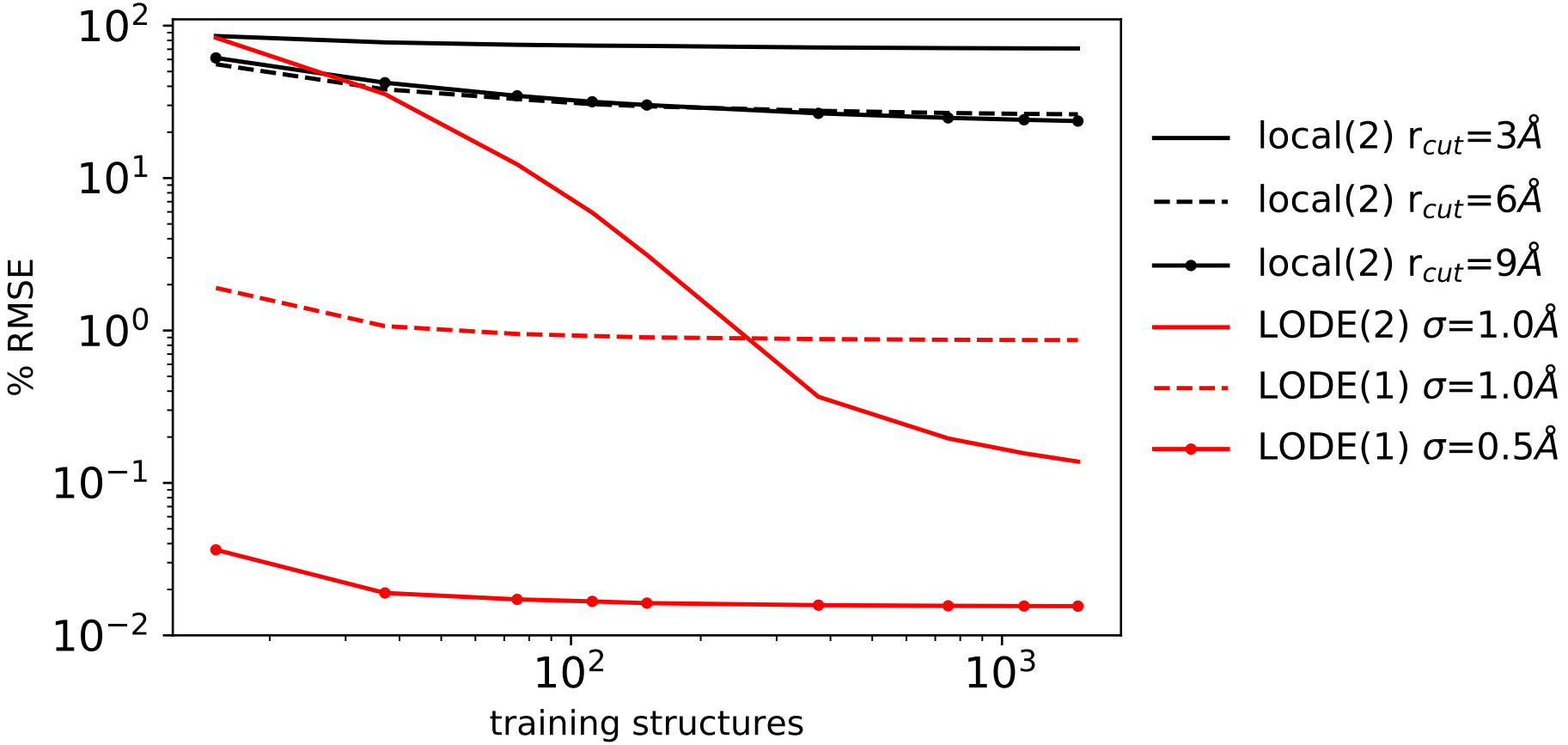





Grisafi, **MC**, arxiv:1909.04512

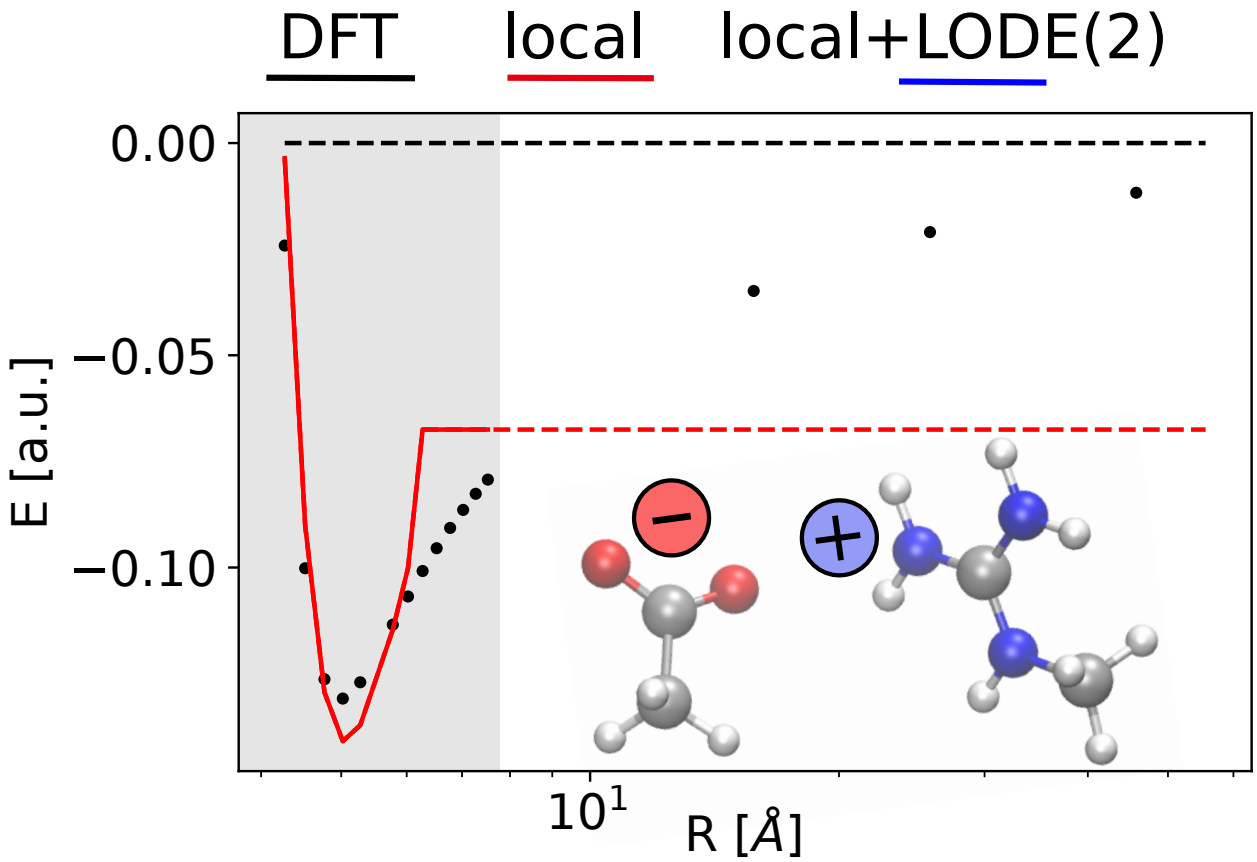
# Learning electrostatics

- A system with purely electrostatic interactions: a random gas of ions
- Extending cutoff in a local description is not efficient.
- Formal link between $\left\langle r \middle| \nu_j^{1\,(1)} \right\rangle$ and $1/r$ potential. Only error is due to the density smoothing



# Predicting binding curves for charged molecules

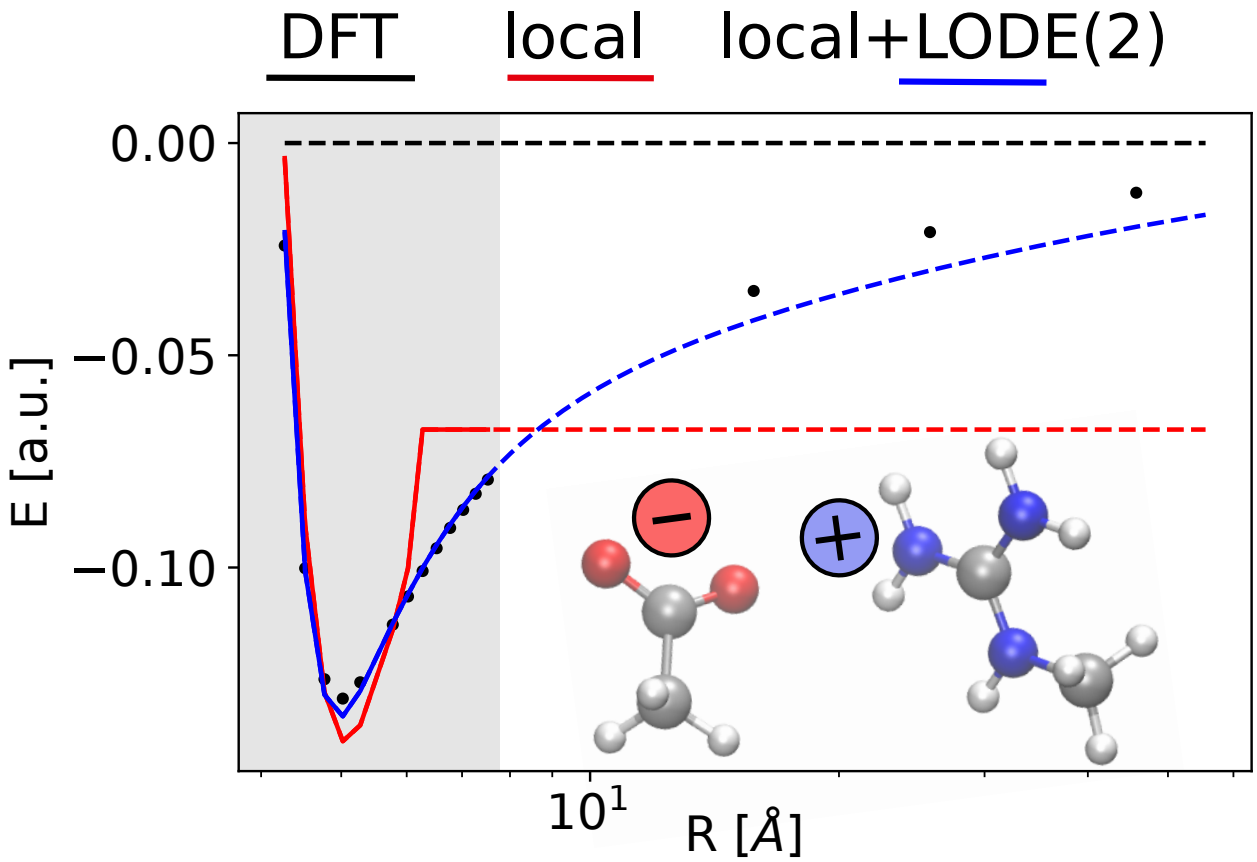
- A challenging test: rigid-molecule binding curves of charged dimers from the BioFragmentsDB
- Train on ~600 dimers, separations <8Å; test on ~60 dimers, up to $>$ 50Å
- Local ML alone fails, but SOAP+LODE combination extrapolates greatly for both monopole-monopole and monopole-dipole interactions




Grisafi, **MC**, arxiv:1909.04512

# Predicting binding curves for charged molecules

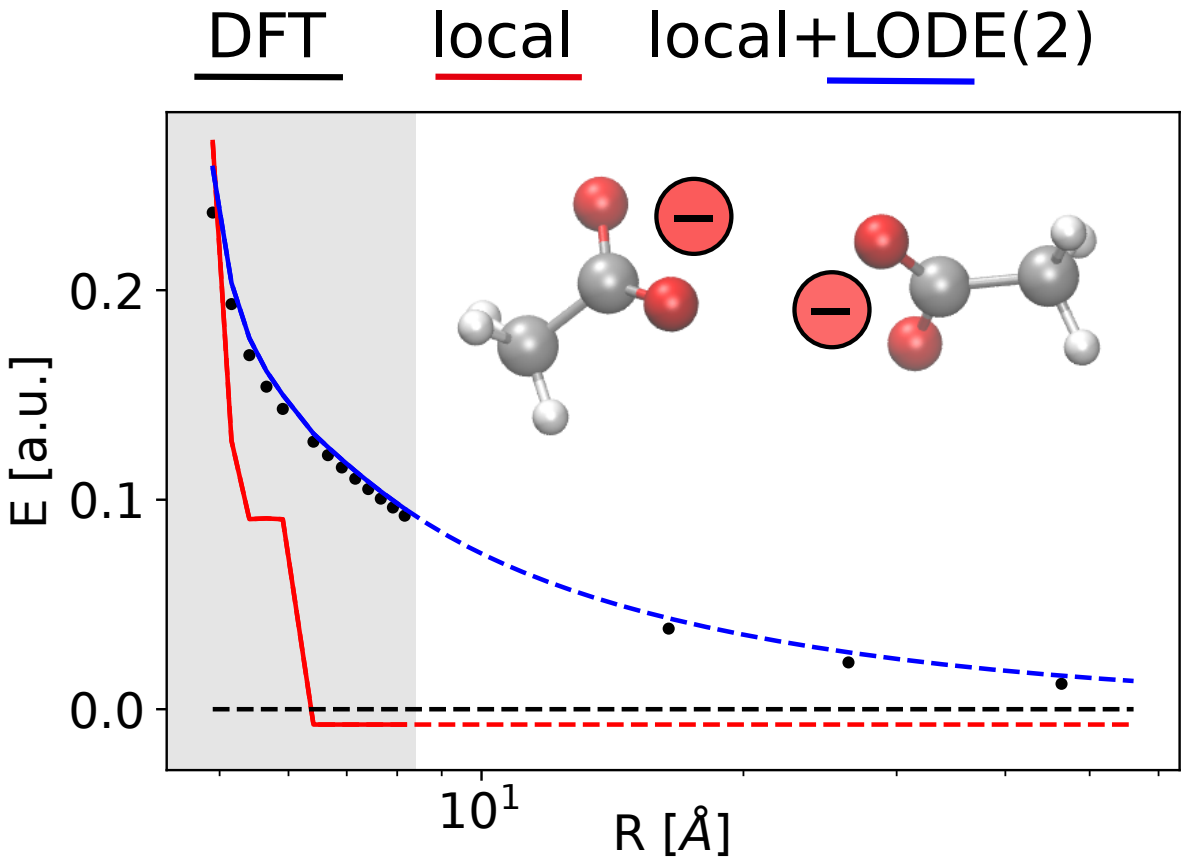
- A challenging test: rigid-molecule binding curves of charged dimers from the BioFragmentsDB
- Train on ~600 dimers, separations <8Å; test on ~60 dimers, up to $>$ 50Å
- Local ML alone fails, but SOAP+LODE combination extrapolates greatly for both monopole-monopole and monopole-dipole interactions




Grisafi, **MC**, arxiv:1909.04512

# Predicting binding curves for charged molecules

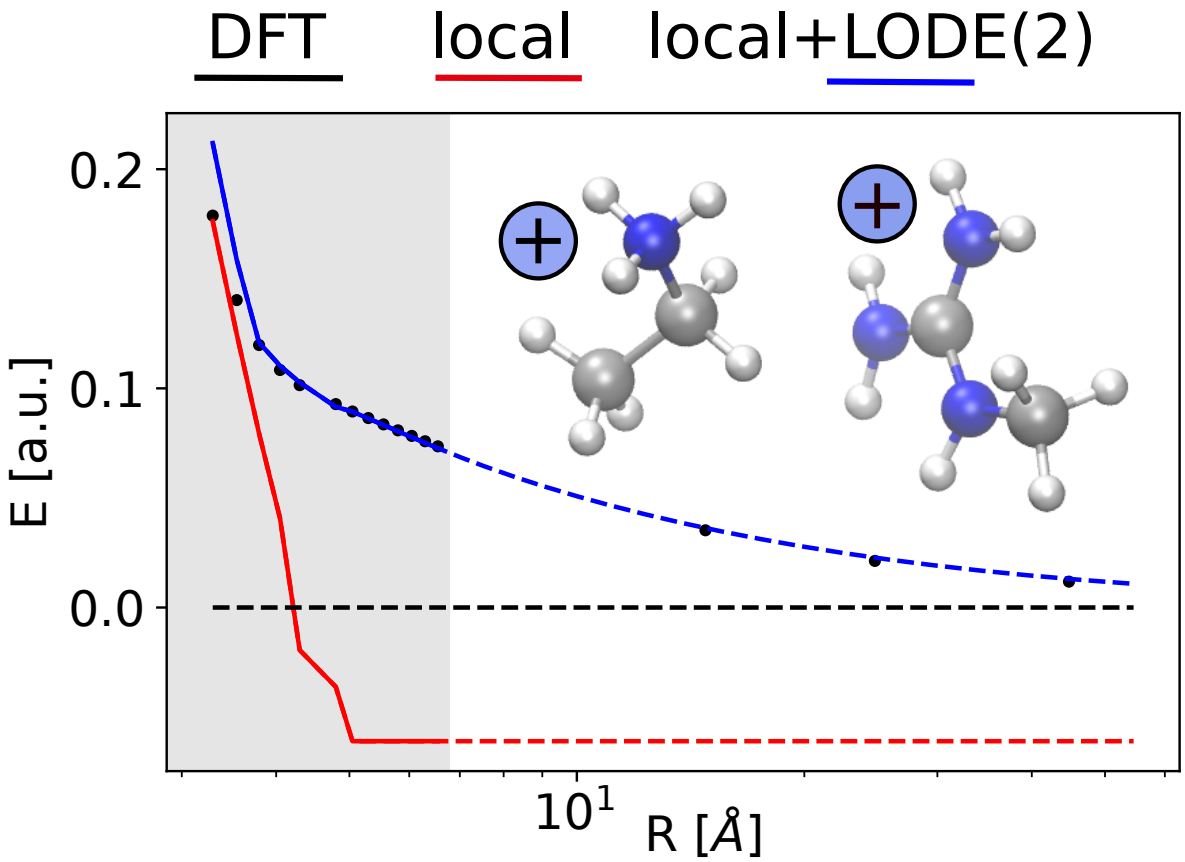
- A challenging test: rigid-molecule binding curves of charged dimers from the BioFragmentsDB
- Train on ~600 dimers, separations <8Å; test on ~60 dimers, up to $>$ 50Å
- Local ML alone fails, but SOAP+LODE combination extrapolates greatly for both monopole-monopole and monopole-dipole interactions




Grisafi, **MC**, arxiv:1909.04512

# Predicting binding curves for charged molecules

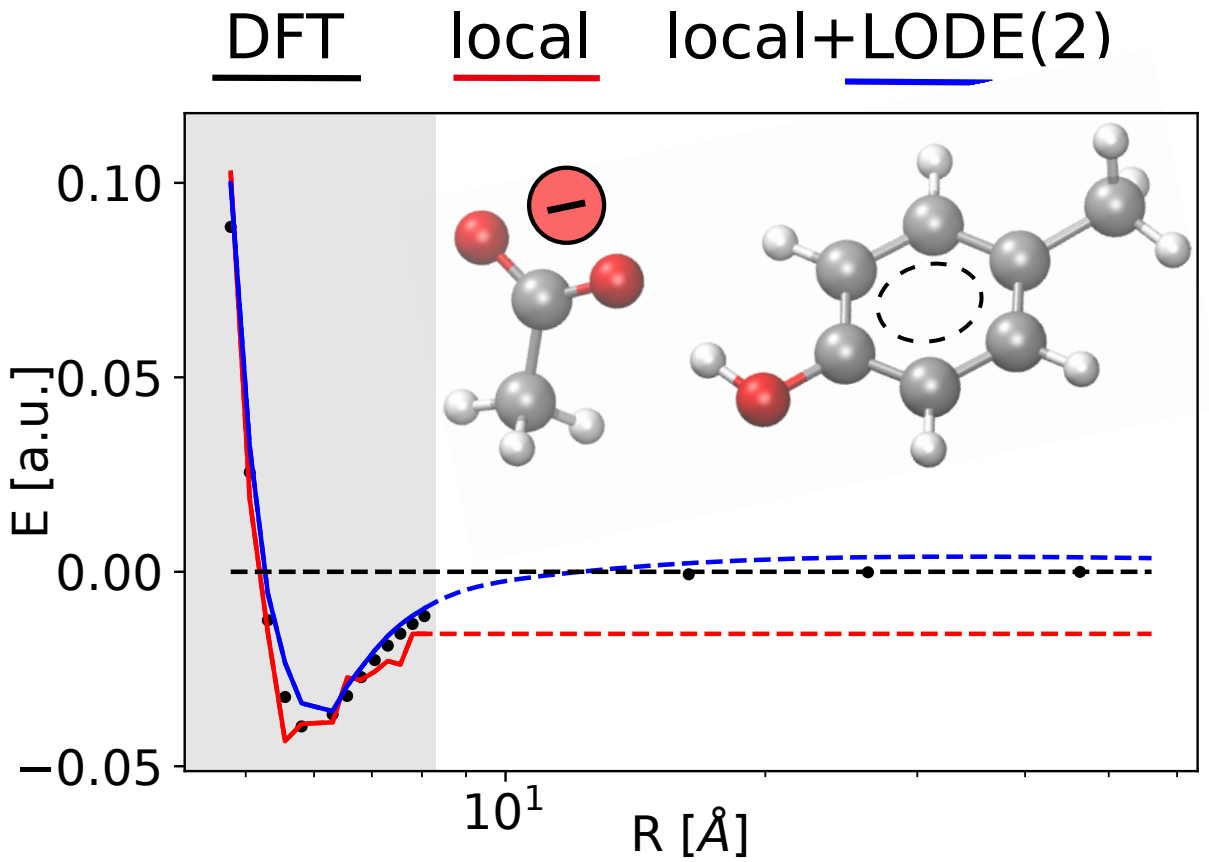
- A challenging test: rigid-molecule binding curves of charged dimers from the BioFragmentsDB
- Train on ~600 dimers, separations <8Å; test on ~60 dimers, up to $>$ 50Å
- Local ML alone fails, but SOAP+LODE combination extrapolates greatly for both monopole-monopole and monopole-dipole interactions




Grisafi, **MC**, arxiv:1909.04512

# Predicting binding curves for charged molecules

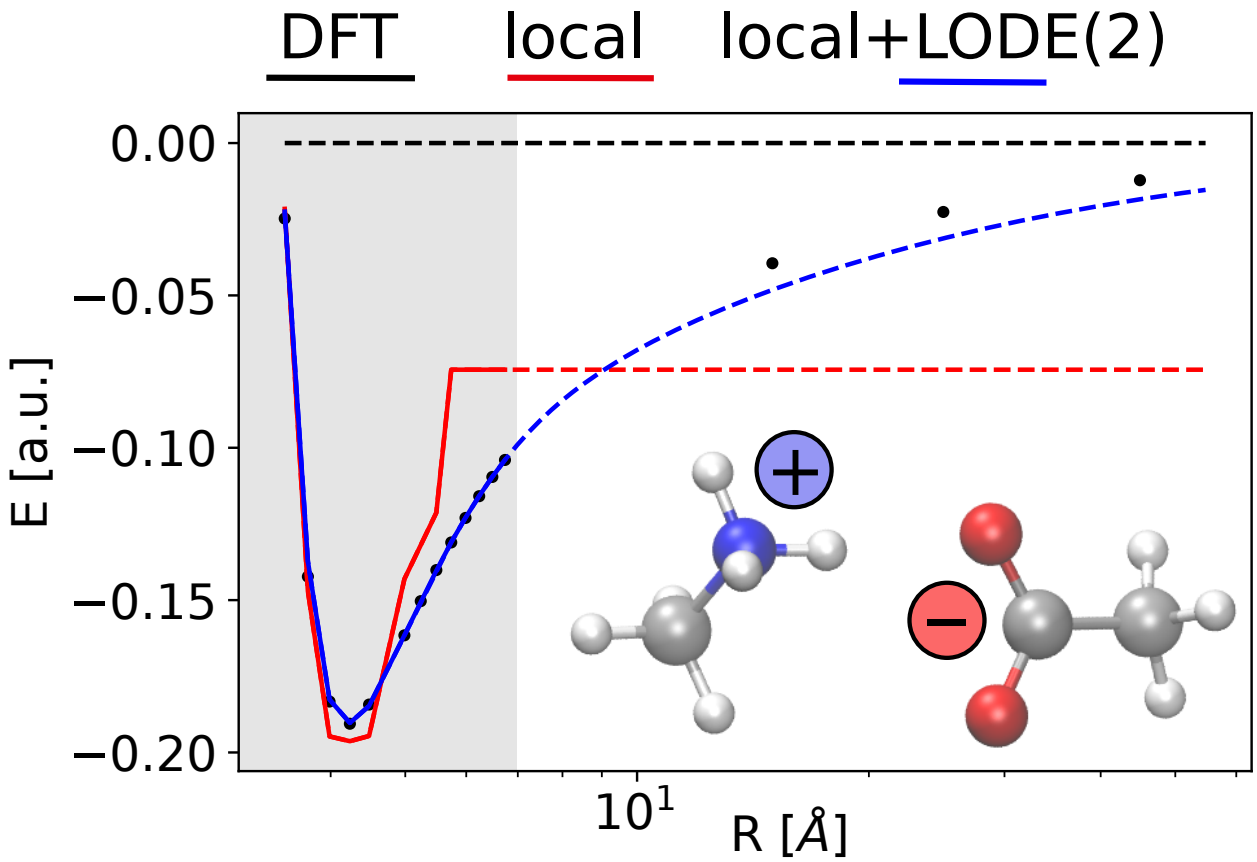
- A challenging test: rigid-molecule binding curves of charged dimers from the BioFragmentsDB
- Train on ~600 dimers, separations <8Å; test on ~60 dimers, up to $>$ 50Å
- Local ML alone fails, but SOAP+LODE combination extrapolates greatly for both monopole-monopole and monopole-dipole interactions




Grisafi, **MC**, arxiv:1909.04512

# Predicting binding curves for charged molecules

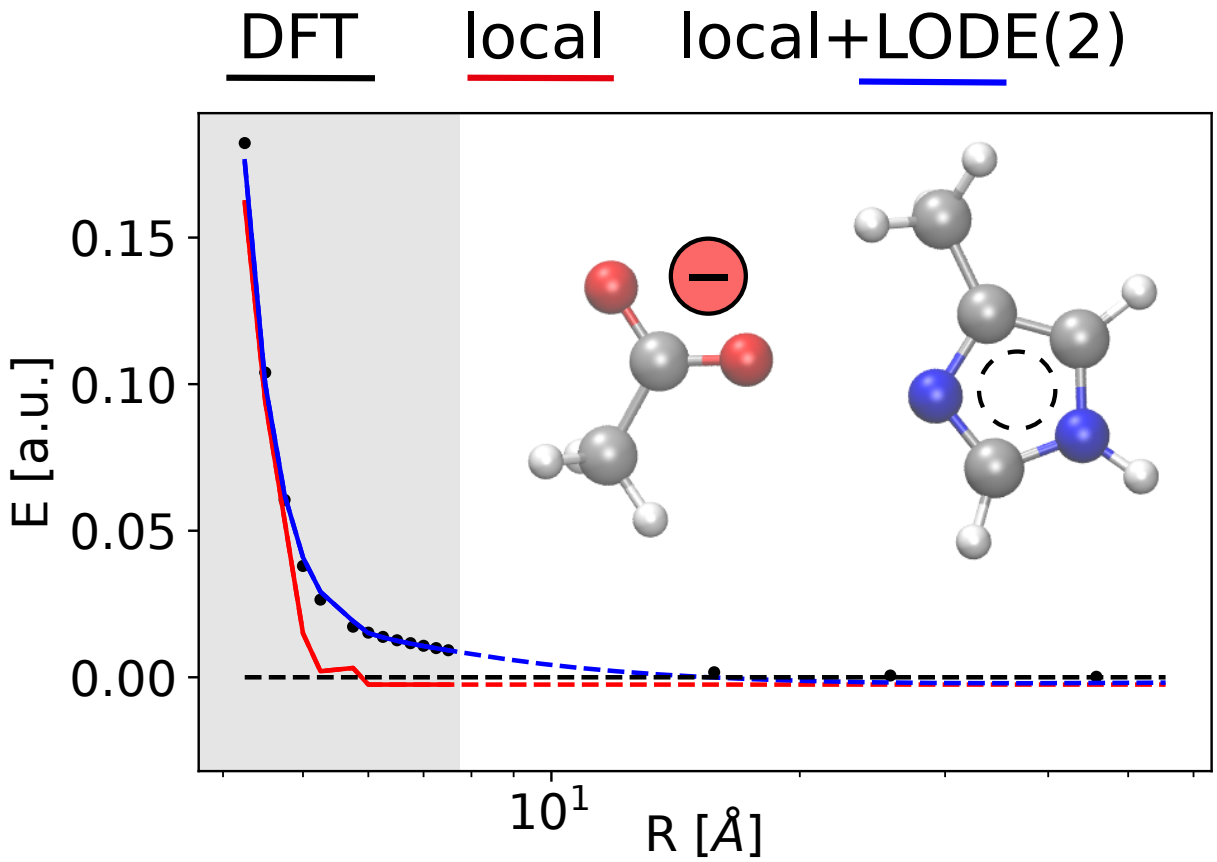
- A challenging test: rigid-molecule binding curves of charged dimers from the BioFragmentsDB
- Train on ~600 dimers, separations <8Å; test on ~60 dimers, up to $>$ 50Å
- Local ML alone fails, but SOAP+LODE combination extrapolates greatly for both monopole-monopole and monopole-dipole interactions




Grisafi, **MC**, arxiv:1909.04512

# Predicting binding curves for charged molecules

- A challenging test: rigid-molecule binding curves of charged dimers from the BioFragmentsDB
- Train on ~600 dimers, separations <8Å; test on ~60 dimers, up to $>$ 50Å
- Local ML alone fails, but SOAP+LODE combination extrapolates greatly for both monopole-monopole and monopole-dipole interactions

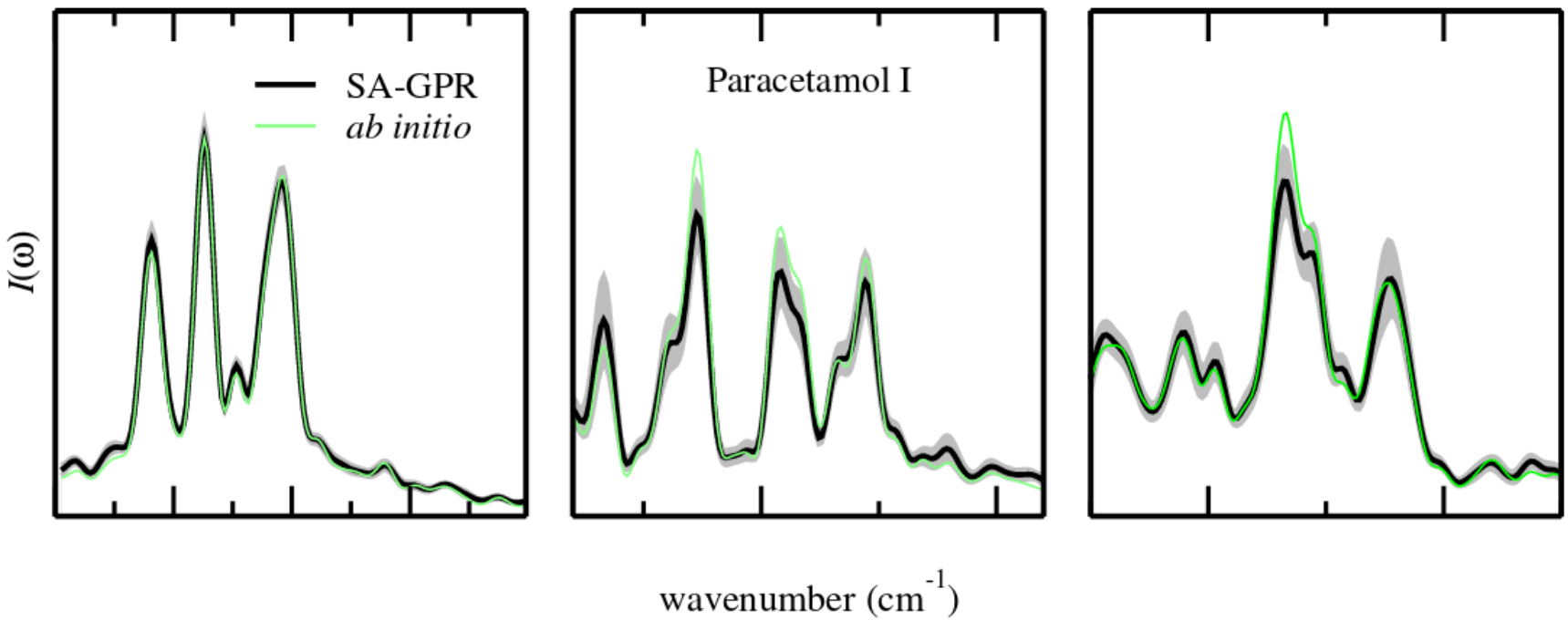



Grisafi, **MC**, arxiv:1909.04512

# Beyond single-point properties

# Raman spectra of molecular crystals

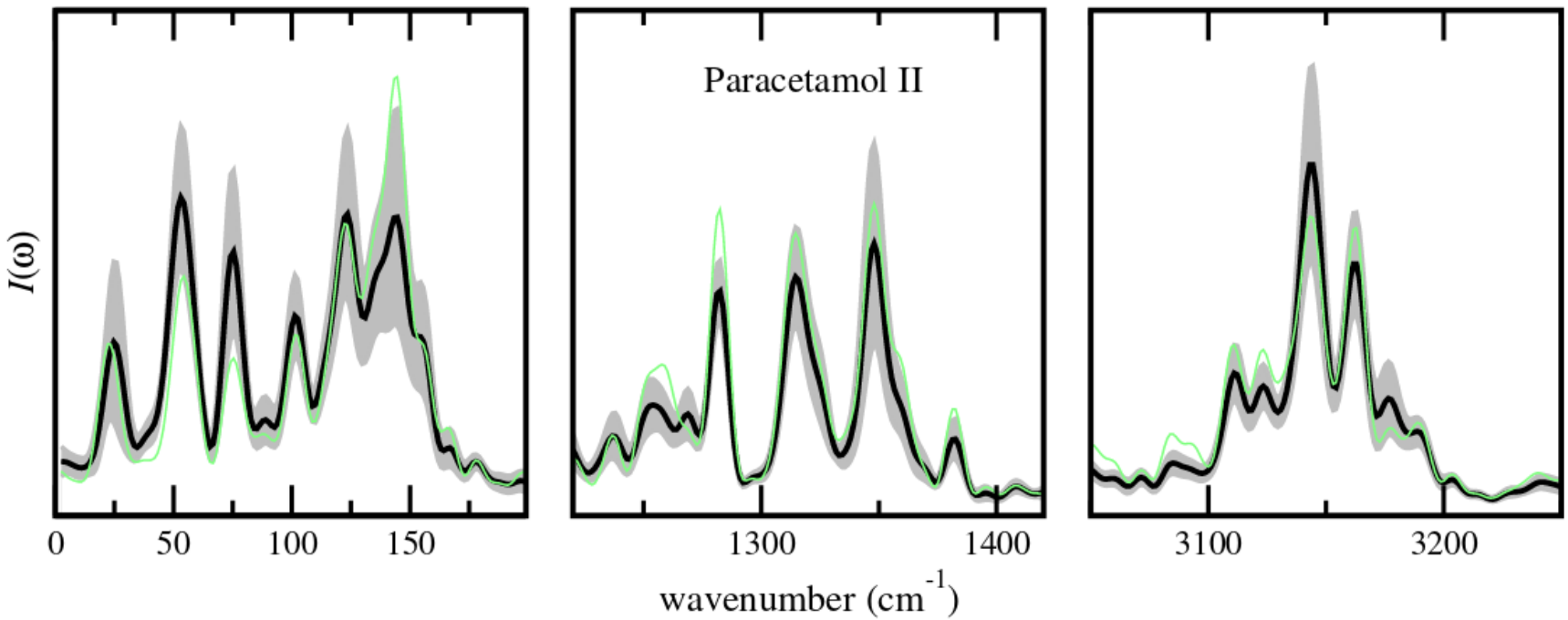
- Raman spectra are sensitive diagnostics for crystal polymorphism
- Symmetry-adapted ML predicts accurately the polarizability for paracetamol crystals, with small train set (~1000 structures)
- Model trained exclusively on paracetamol form I gives quantitative accuracy for form II




N. Raimbault, A. Grisafi, **MC**, M. Rossi, New J. Phys. (2019)

# Raman spectra of molecular crystals

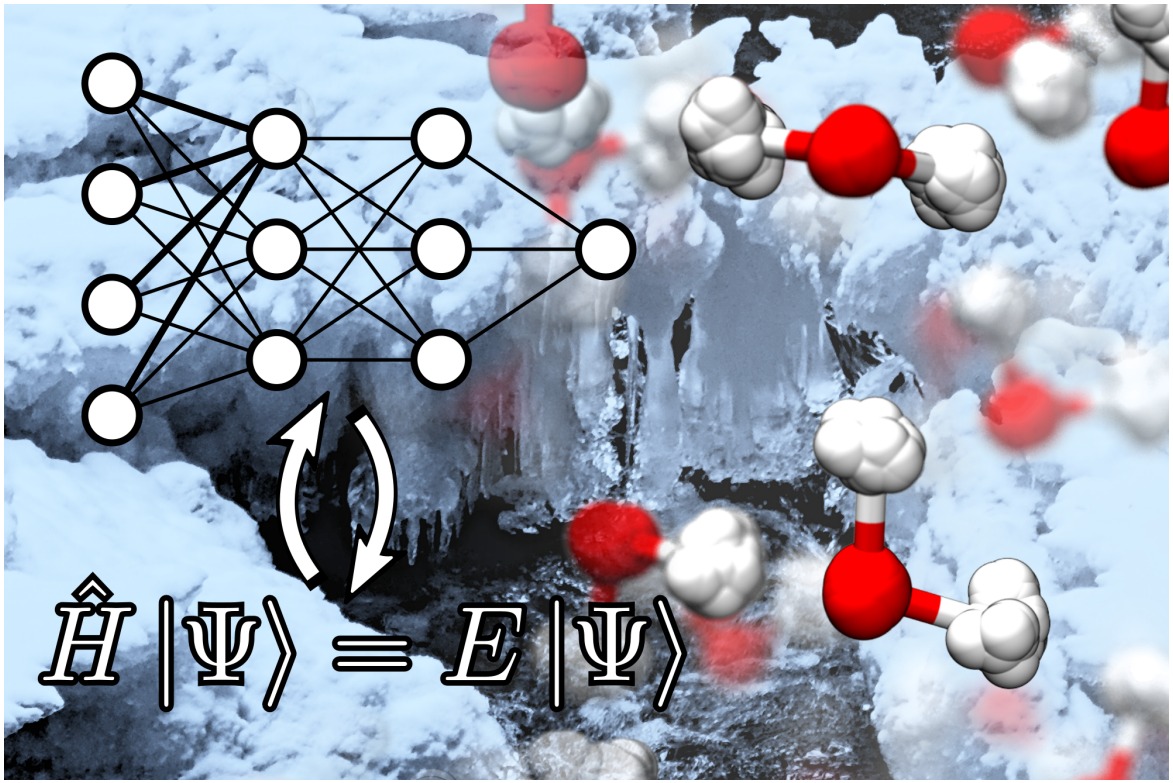
- Raman spectra are sensitive diagnostics for crystal polymorphism
- Symmetry-adapted ML predicts accurately the polarizability for paracetamol crystals, with small train set (~1000 structures)
- Model trained exclusively on paracetamol form I gives quantitative accuracy for form II




N. Raimbault, A. Grisafi, **MC**, M. Rossi, New J. Phys. (2019)

# Ab initio thermodynamics of water

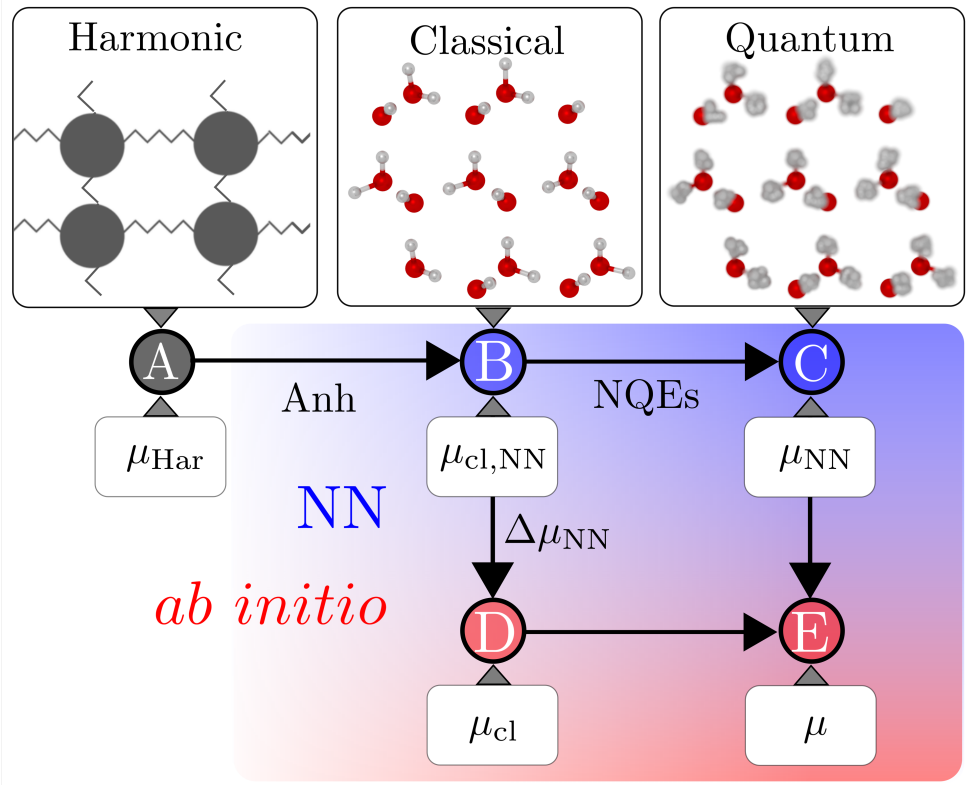
- Huge challenge: determining the thermodynamics of Ih/Ic polymorphism (and the liquid, as a bonus) at the hybrid DFT+D3 level
- Use a machine-learned potential (Behler-Parrinello NN) as a stepping stone, for all thermodynamic integration, and free-energy perturbation to promote to full DFT level




Cheng, Engel, Behler, Dellago, **MC**, PNAS (2019)

# Ab initio thermodynamics of water

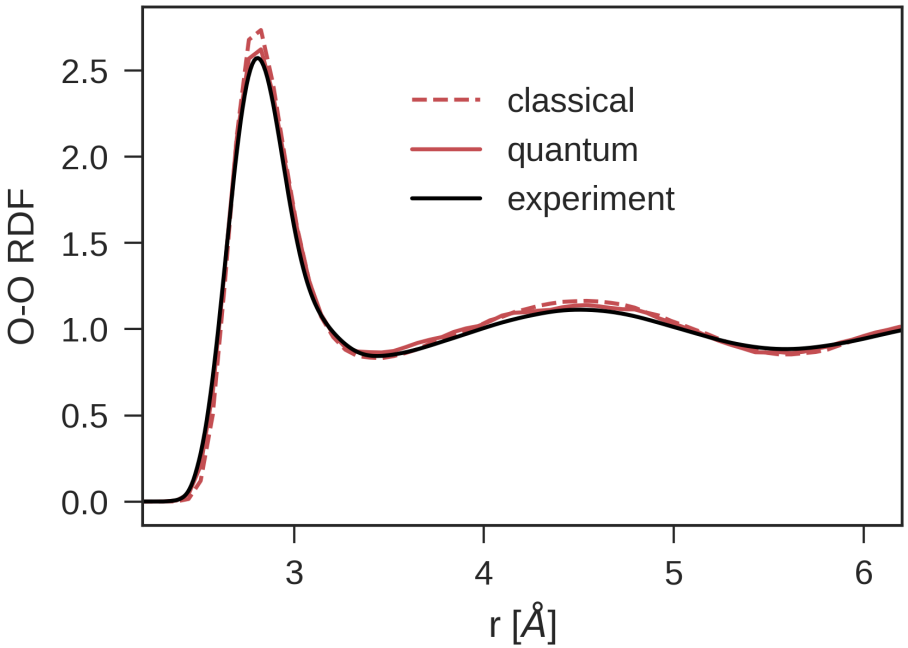
- Huge challenge: determining the thermodynamics of Ih/Ic polymorphism (and the liquid, as a bonus) at the hybrid DFT+D3 level
- Use a machine-learned potential (Behler-Parrinello NN) as a stepping stone, for all thermodynamic integration, and free-energy perturbation to promote to full DFT level

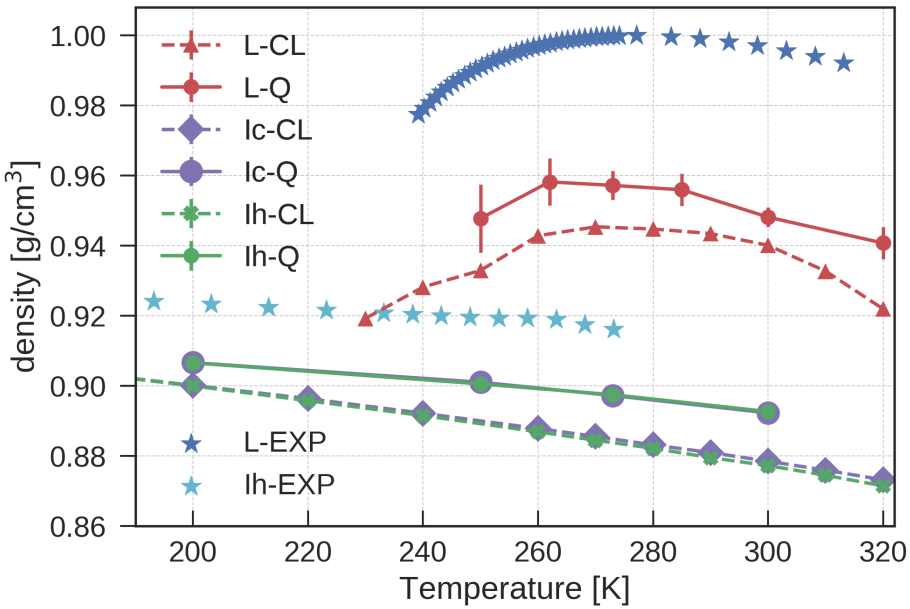



Cheng, Engel, Behler, Dellago, **MC**, PNAS (2019)

# The right results for the right reasons

- Excellent agreement with experiments, thanks to reference hybrid DFT (REVPBE0-D3) that works very well for water
- Getting the melting point of water within ~5K!
- Nuclear quantum effects contribute a small (crucial!) stabilization to the hexagonal phase

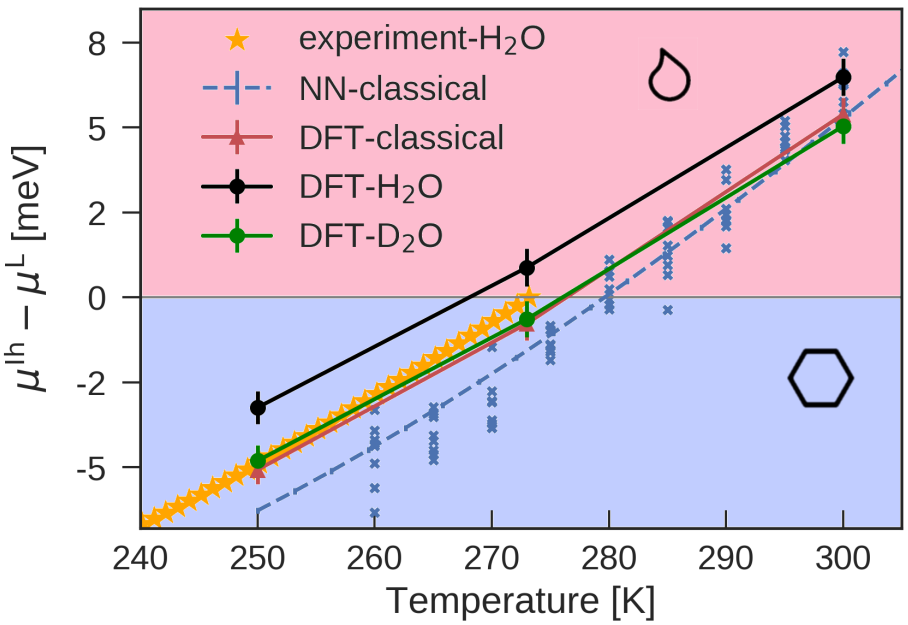


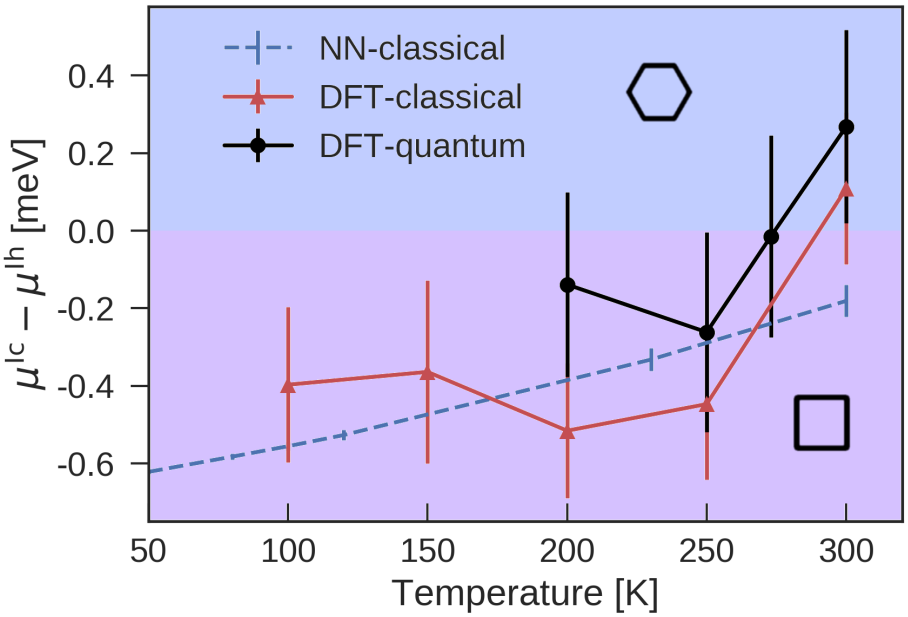




Cheng, Engel, Behler, Dellago, **MC**, PNAS (2019)

# The right results for the right reasons

- Excellent agreement with experiments, thanks to reference hybrid DFT (REVPBE0-D3) that works very well for water
- Getting the melting point of water within ~5K!
- Nuclear quantum effects contribute a small (crucial!) stabilization to the hexagonal phase

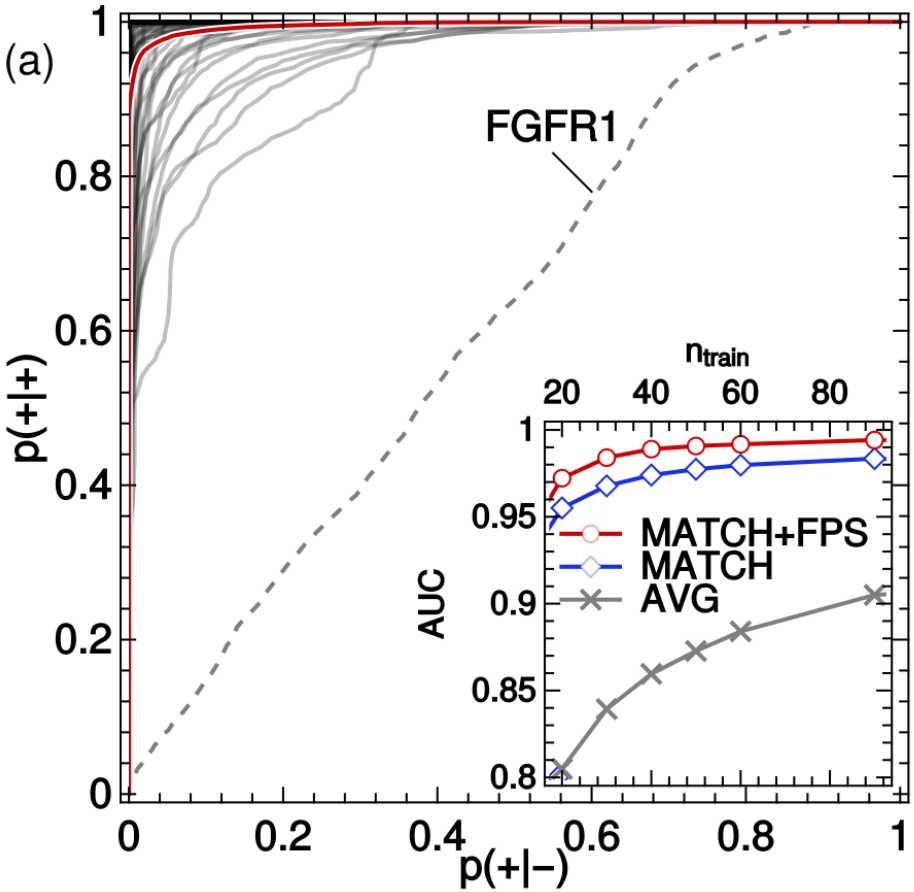


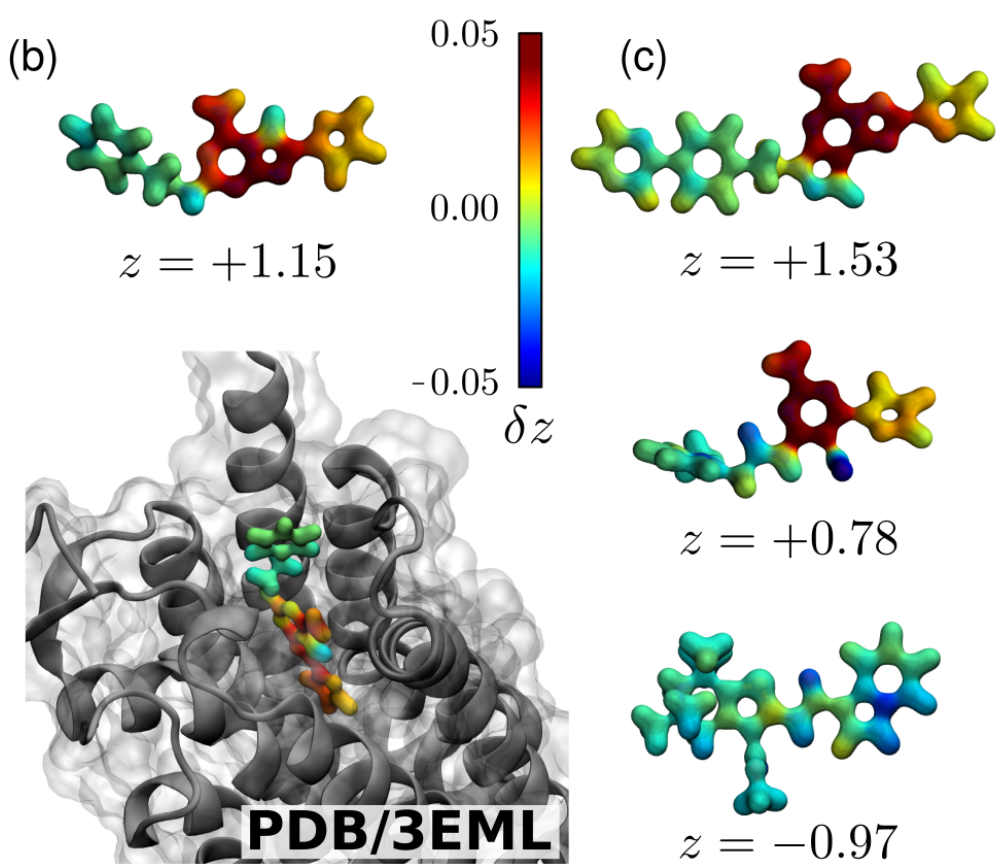




Cheng, Engel, Behler, Dellago, **MC**, PNAS (2019)

# Recognizing active protein ligands

- A SOAP-REMatch-based KSVM classifies active and inactive ligands with 99% accuracy; non-additive model is crucial!
- Sensitivity analysis help identify the active “warhead” and could guide drug design and optimization

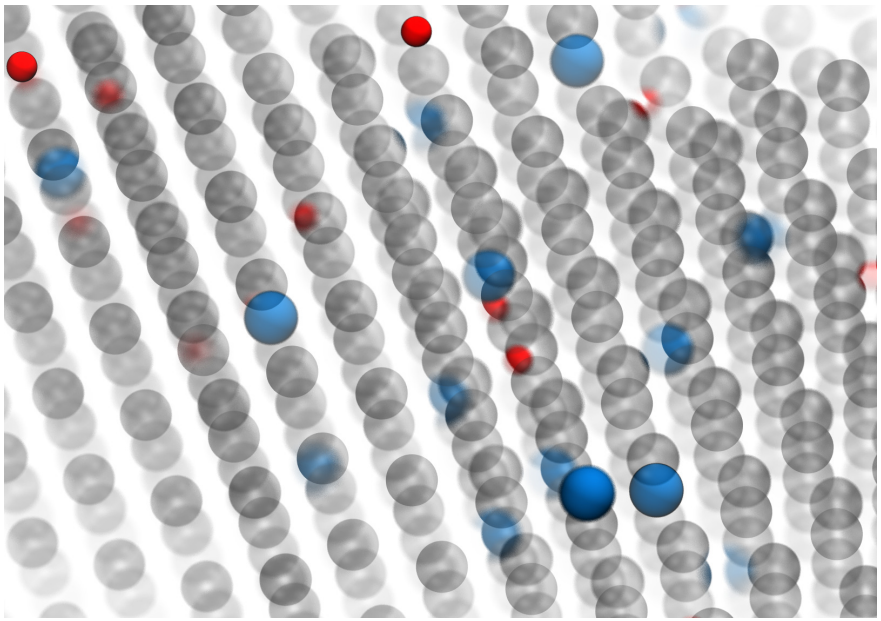


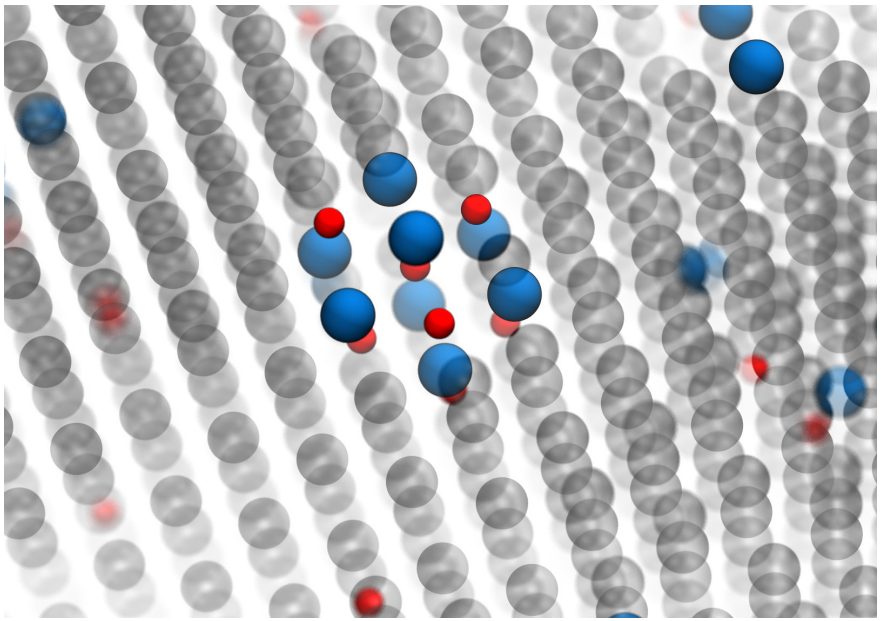




Bartok, De, Poelking, Kermode, Bernstein, Csanyi, **MC**, Science Advances (2017) *[data: DUD-E, Shoichet]*

# Machine learning between physics and data

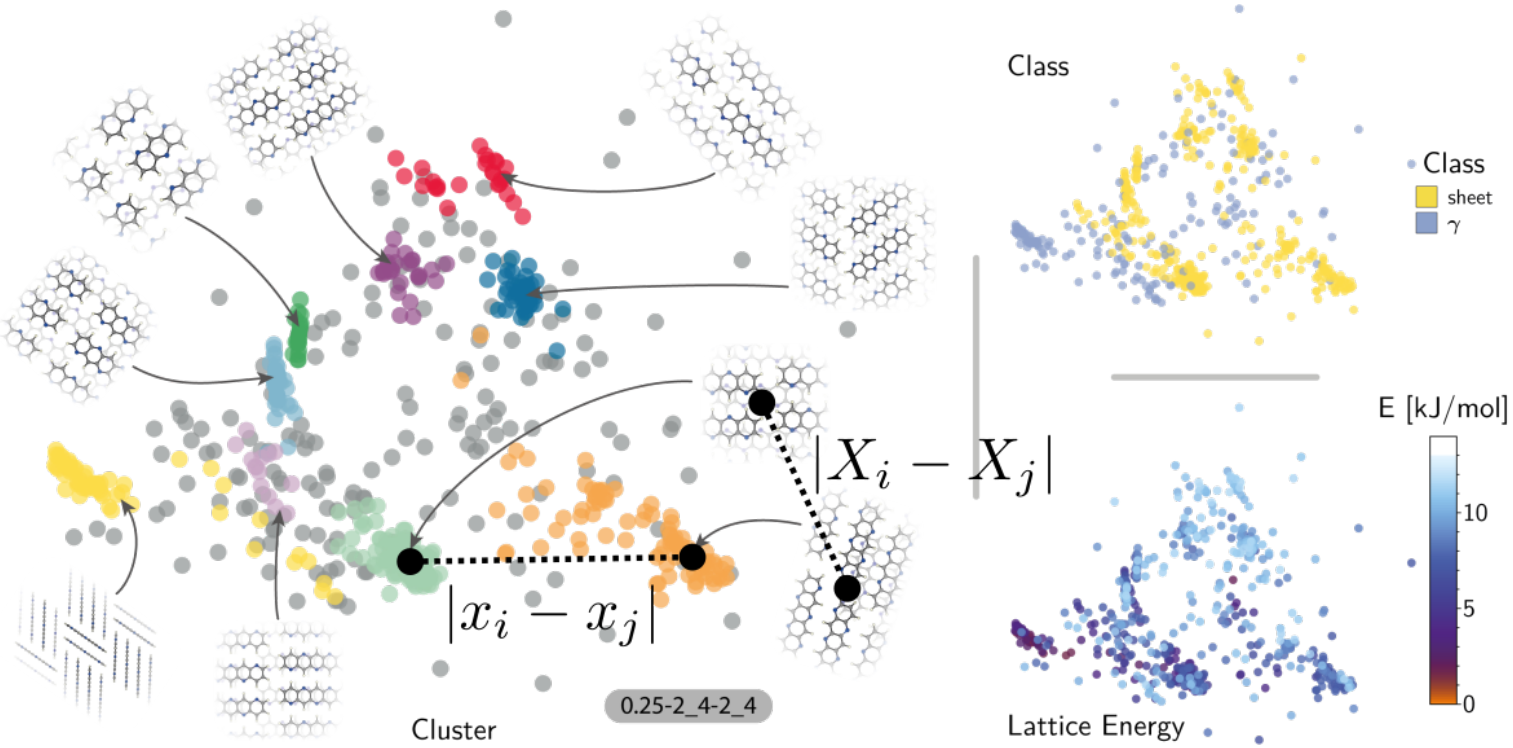
- Machine learning is an enabling technology to do serious stat mech and kinetics with first-principles energetics
- Machine learning benefits greatly from incorporating physical principles
  - Symmetries of representations and target quantities
  - Locality, additivity, smoothness, conservation laws. . .
  - Long-range interactions
- Physical-chemical insight can also emerge from a ML exercise





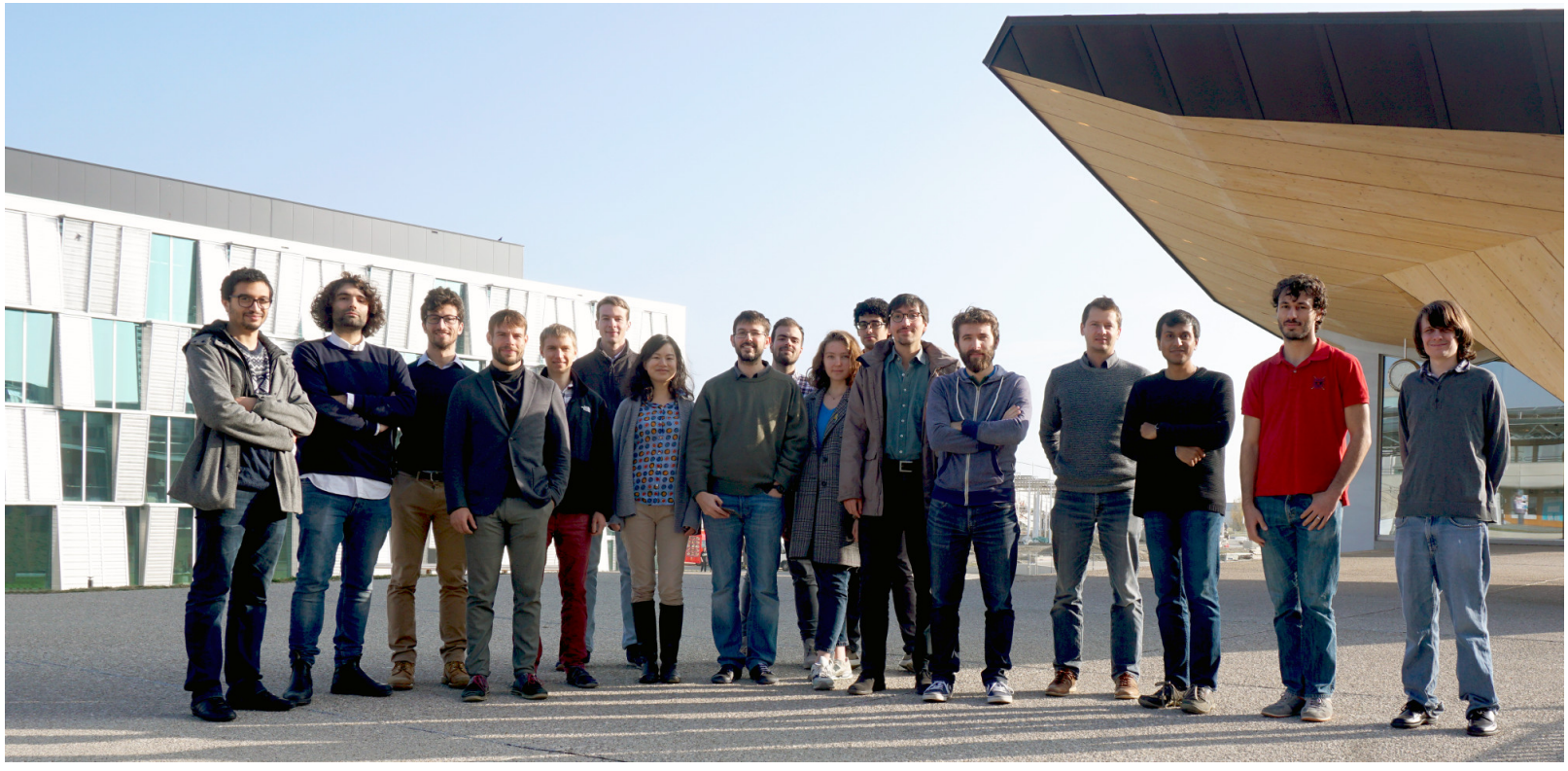
# Machine learning between physics and data

- Machine learning is an enabling technology to do serious stat mech and kinetics with first-principles energetics
- Machine learning benefits greatly from incorporating physical principles
  - Symmetries of representations and target quantities
  - Locality, additivity, smoothness, conservation laws. . .
  - Long-range interactions
- Physical-chemical insight can also emerge from a ML exercise



http://cosmo.epfl.ch

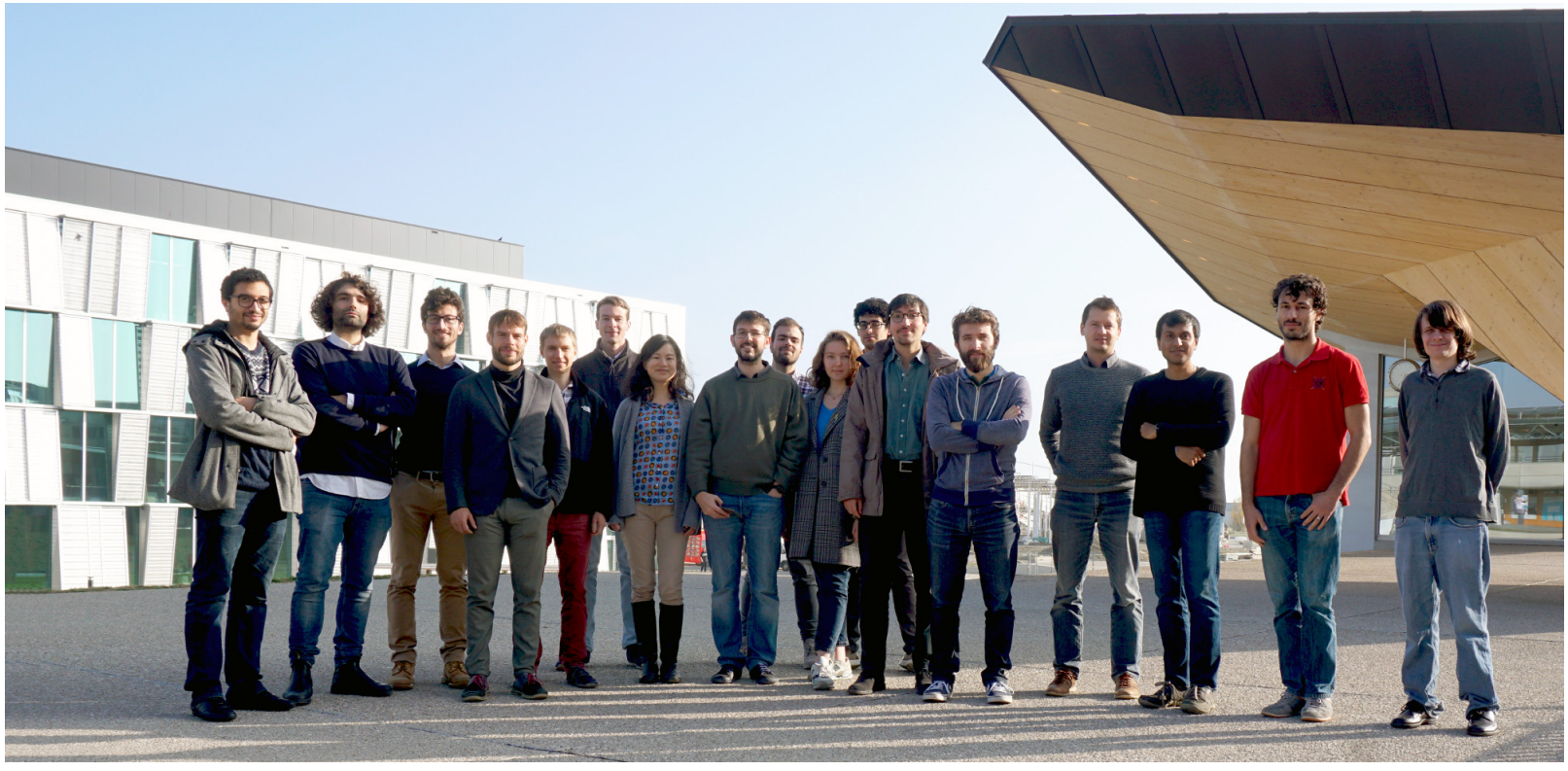
Follow @COSMO_EPFL



C.Cominboeuf, W.Curtin, L.Emsley, A.Fontcuberta, R.Logé, N.Marzari, S.Roke, F.Sorin, F.Stellacci
G.Csányi, S.Althorpe, C.Pickard (Cambridge) J.Kermode (Warwick) D.Manolopoulos (Oxford)
J.Behler (Göttingen) T.Paxton (KCL) G.Day (SHampton) R.DiStasio (Cornell) S.Auerbach (UMass)
M.Rossi (MP Hamburg) G.Tribello (Belfast) T.Markland (Stanford) A.Cooper (Liverpool)

http://cosmo.epfl.ch

Follow @COSMO_EPFL



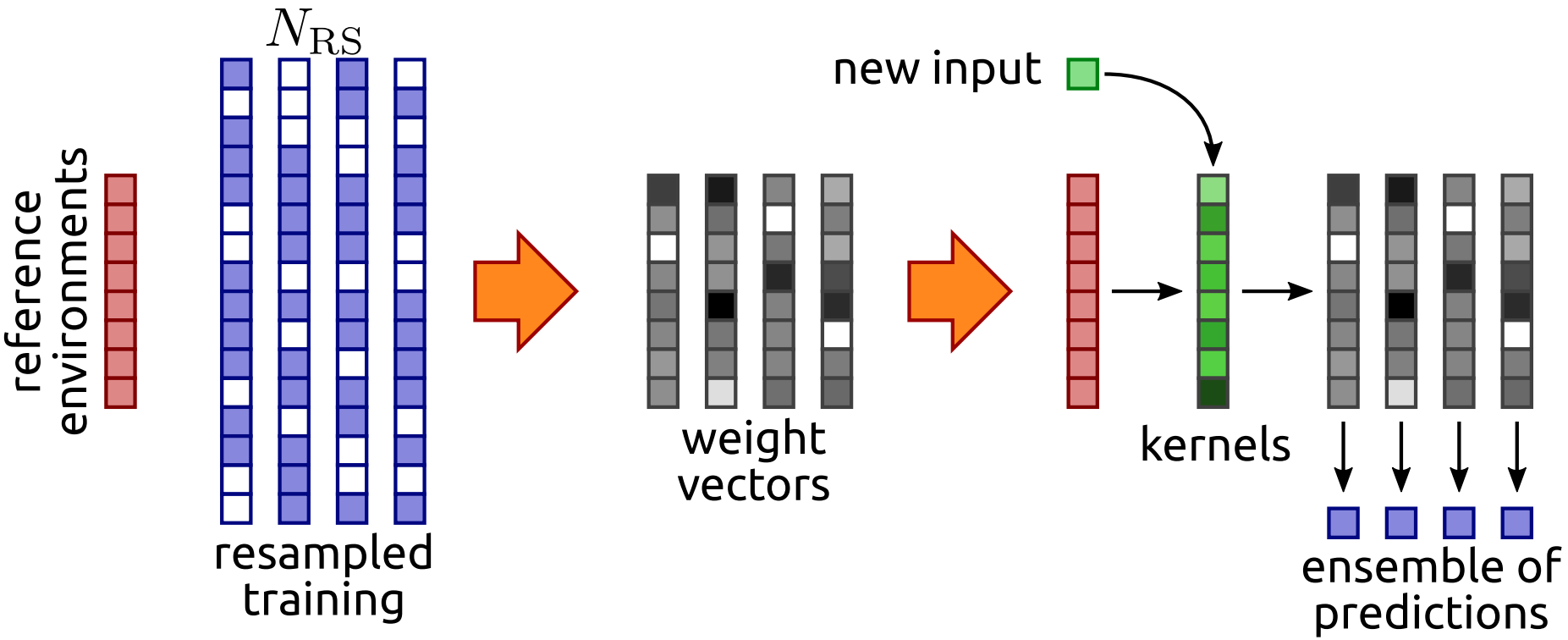
Deep connections between structure representations ..... Willatt et al. JCP (2019)
Strategies to reduce the computational cost ..... Imbalzano et al. J. Chem. Phys. (2018)
Feature optimization: efficiency and insight ..... Willatt et al. PCCP (2018)
Fast and accurate error estimation ..... Musil et al. JCTC (2019)
Symmetry-adapted regression for tensors: ..... Grisafi et al., Phys. Rev. Lett. (2018)
Molecular polarizability ..... Wilkins et al. PNAS (2019)
Electron density ..... Grisafi et al., ACS Central Science (2019)
Applications from water to biomolecules .....
Bartók et al. Science Adv. (2017); Musil et al., Chem. Sci. (2018);
Raimbault et al. New J. Phys. (2019); Paruzzo et al. Nat.Comm. (2018); Cheng et al., PNAS (2019);

# An accurate & inexpensive error estimation

- Generate an ensemble of GPR models, and use distribution of predictions

$$y(\mathcal{X}) = \frac{1}{N_{RS}} \sum_i y^{(i)}(\mathcal{X}), \qquad \sigma^2(\mathcal{X}) = \frac{1}{N_{RS}-1} \sum_i \left( y^{(i)}(\mathcal{X}) - y(\mathcal{X}) \right)^2$$

  - Verify accuracy by the distribution of errors $P\left(|y(\mathcal{X}) - y_{\text{ref}}(\mathcal{X})| \;\middle|\; \sigma(\mathcal{X})\right)$
  - Use maximum-likelihood to calibrate the uncertainty $\sigma(\mathcal{X}) \to \alpha\sigma(\mathcal{X})^{\gamma-1}$



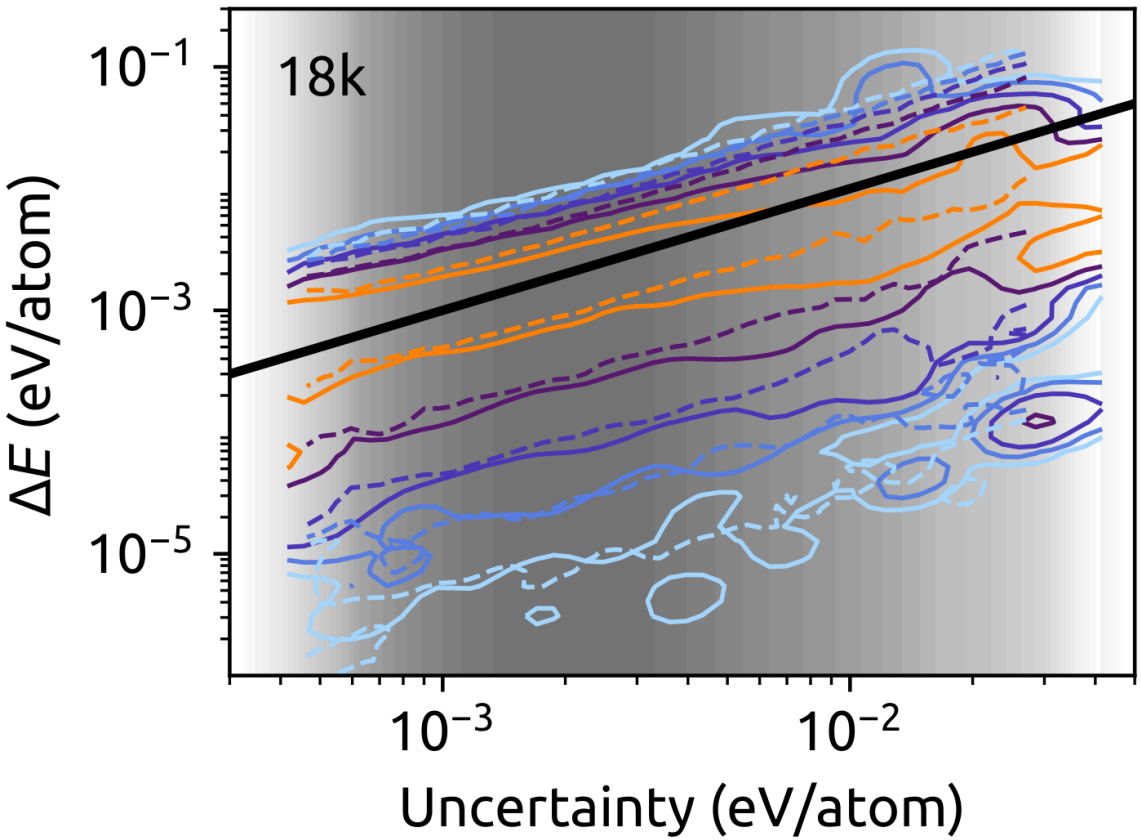

Musil, Willatt, **MC** JCTC (2019)

# An accurate & inexpensive error estimation

- Generate an ensemble of GPR models, and use distribution of predictions

$$y(\mathcal{X}) = \frac{1}{N_{RS}} \sum_i y^{(i)}(\mathcal{X}), \qquad \sigma^2(\mathcal{X}) = \frac{1}{N_{RS}-1} \sum_i \left( y^{(i)}(\mathcal{X}) - y(\mathcal{X}) \right)^2$$

  - Verify accuracy by the distribution of errors $P\left(|y(\mathcal{X}) - y_{\text{ref}}(\mathcal{X})| \;\middle|\; \sigma(\mathcal{X})\right)$
  - Use maximum-likelihood to calibrate the uncertainty $\sigma(\mathcal{X}) \to \alpha\sigma(\mathcal{X})^{\gamma-1}$



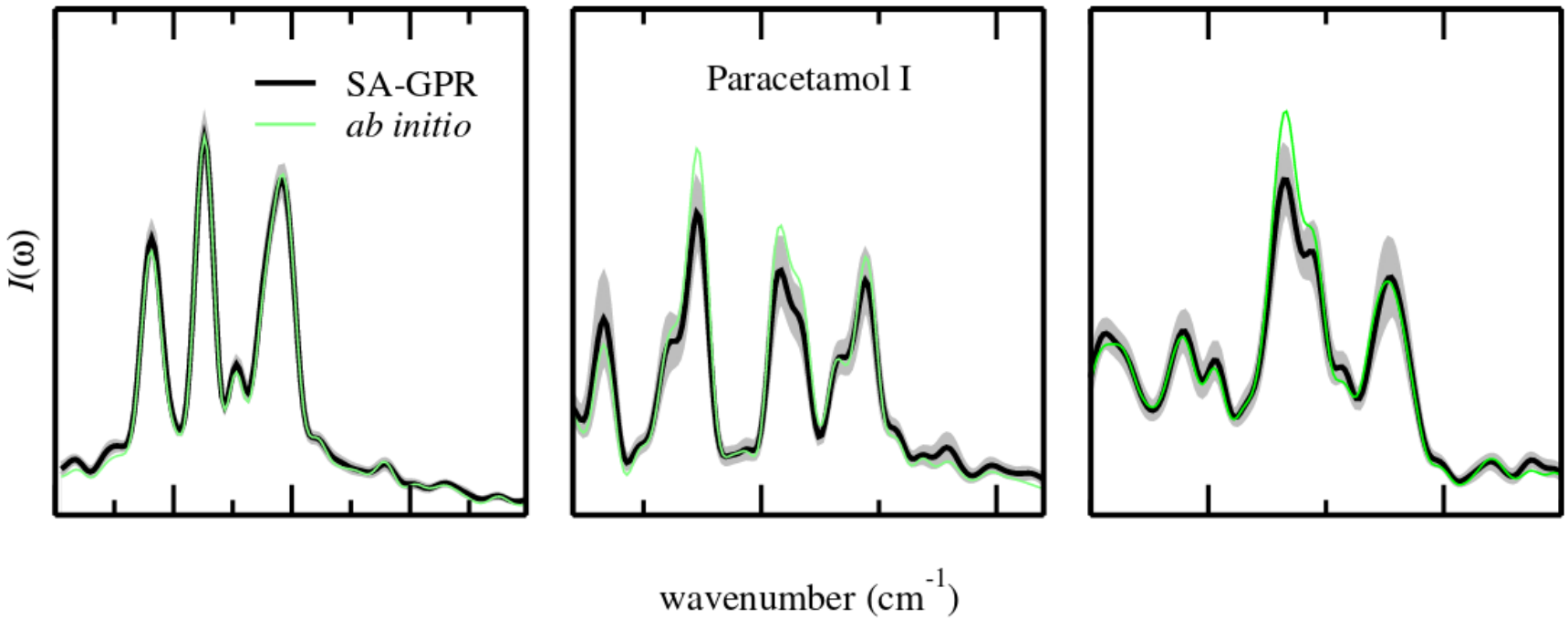

Musil, Willatt, **MC** JCTC (2019)

# An accurate & inexpensive error estimation

- Generate an ensemble of GPR models, and use distribution of predictions

$$y(\mathcal{X}) = \frac{1}{N_{RS}} \sum_i y^{(i)}(\mathcal{X}), \qquad \sigma^2(\mathcal{X}) = \frac{1}{N_{RS}-1} \sum_i \left( y^{(i)}(\mathcal{X}) - y(\mathcal{X}) \right)^2$$

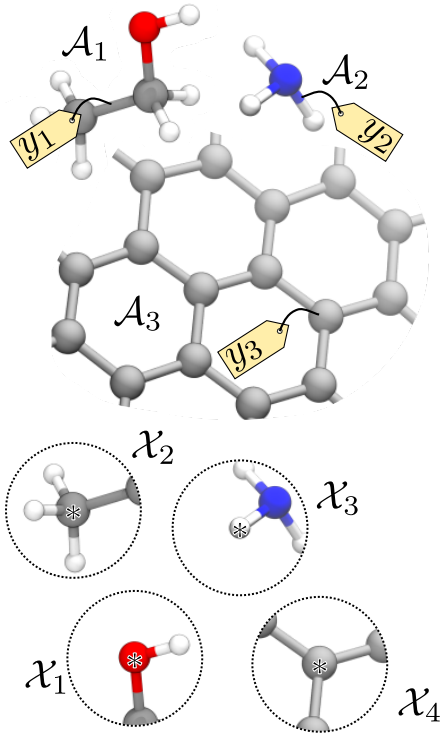
  - Verify accuracy by the distribution of errors $P\left(|y(\mathcal{X}) - y_{\text{ref}}(\mathcal{X})| \;\middle|\; \sigma(\mathcal{X})\right)$
  - Use maximum-likelihood to calibrate the uncertainty $\sigma(\mathcal{X}) \to \alpha\sigma(\mathcal{X})^{\gamma-1}$

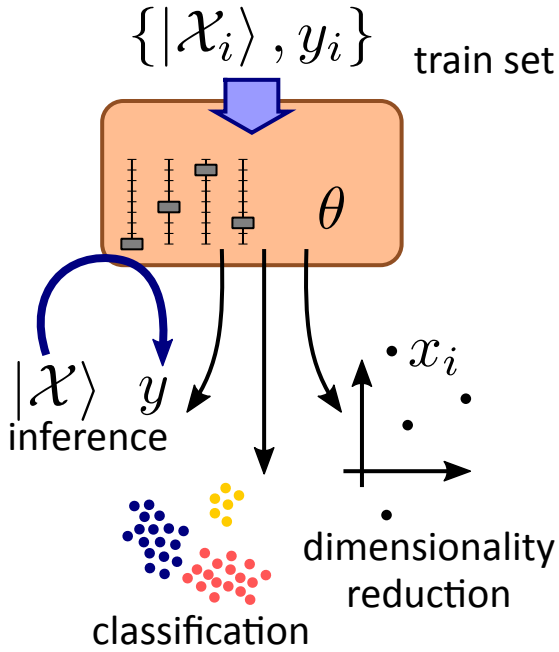



N. Raimbault, A. Grisafi, **MC**, M. Rossi, New J. Phys. (2019)

# Machine learning at the atomic scale

- Chemical structures (inputs) and their properties (labels) are fed to a learning scheme, tuned by hyperparameters $\theta$, that can then be used to perform different tasks on new data
- Structures must be cast to feature vectors $|\mathcal{A}\rangle$ /distances $d(\mathcal{A},\mathcal{B})$/kernels $k(\mathcal{A},\mathcal{B})$ to obtain a representation that must be concise, complete, and incorporating physical principles

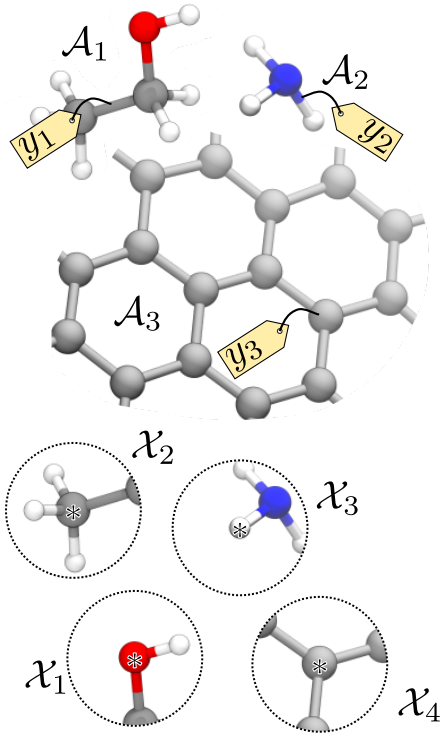


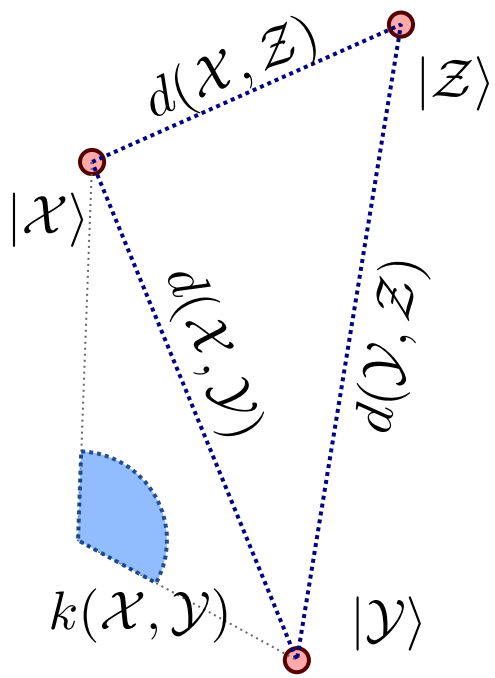



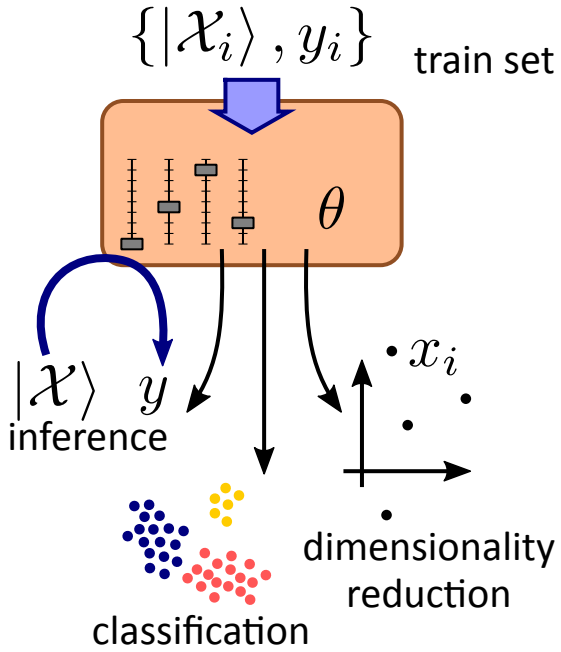

JCP Perspective, **MC** (2019)

# Machine learning at the atomic scale

- Chemical structures (inputs) and their properties (labels) are fed to a learning scheme, tuned by hyperparameters $\theta$, that can then be used to perform different tasks on new data
- Structures must be cast to feature vectors $|\mathcal{A}\rangle$ /distances $d(\mathcal{A},\mathcal{B})$/kernels $k(\mathcal{A},\mathcal{B})$ to obtain a representation that must be concise, complete, and incorporating physical principles



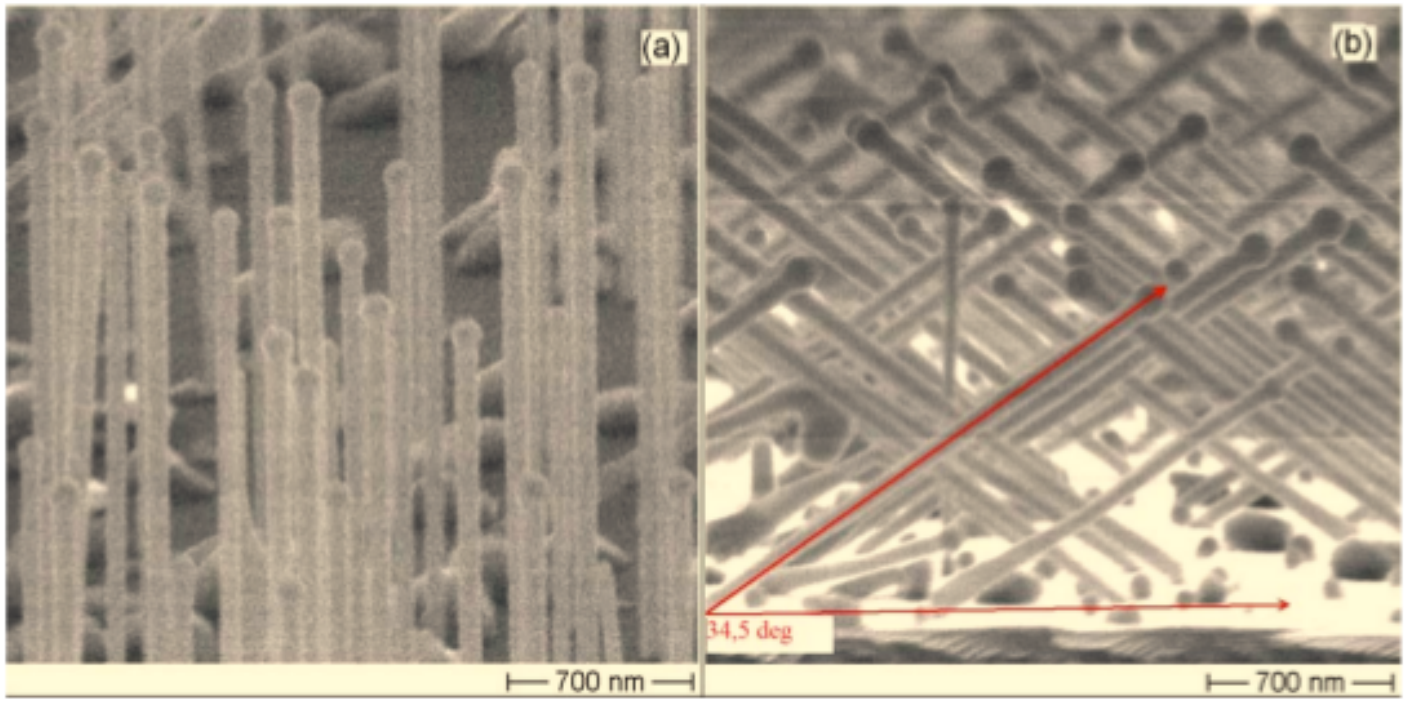




JCP Perspective, **MC** (2019)

# Synthesis of GaAs nanowires

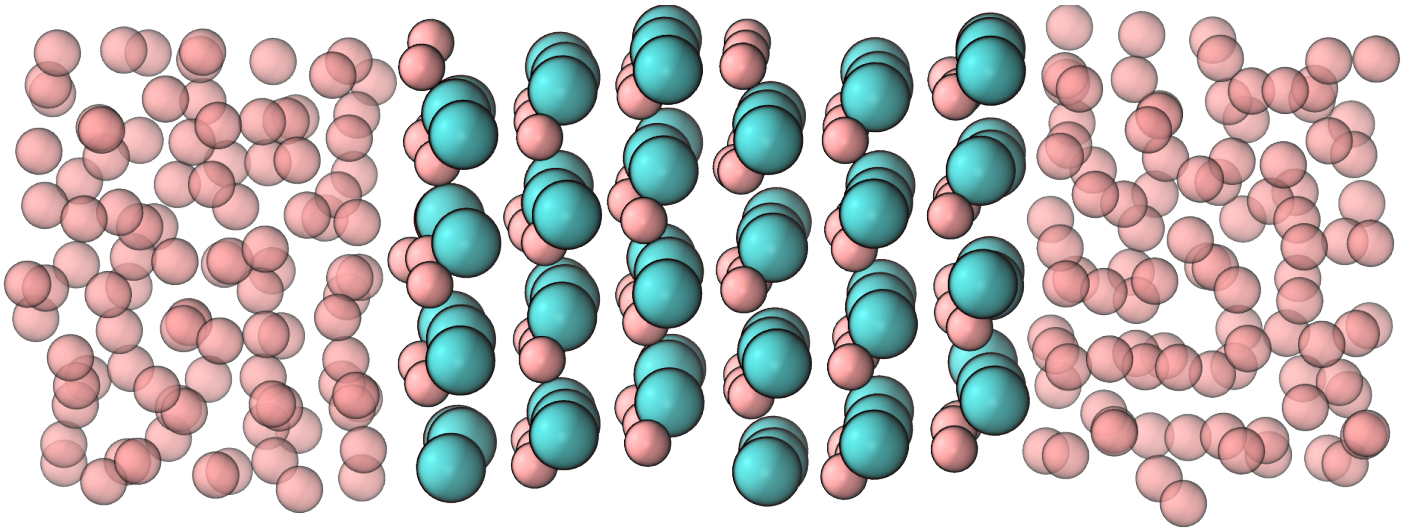
- GaAs nanowires can be grown with solid-liquid-gas process based on droplets of molten Ga
- Problem: controlling the zinc-blende/wurtzite polymorphism.
- Important role played by surface polarity
- MD simulation with DFT/NN to study liquid ordering at the interfaces




w/ Giulio Imbalzano, Mahdi Zamani, Anna Fontcuberta, CIME

# Synthesis of GaAs nanowires

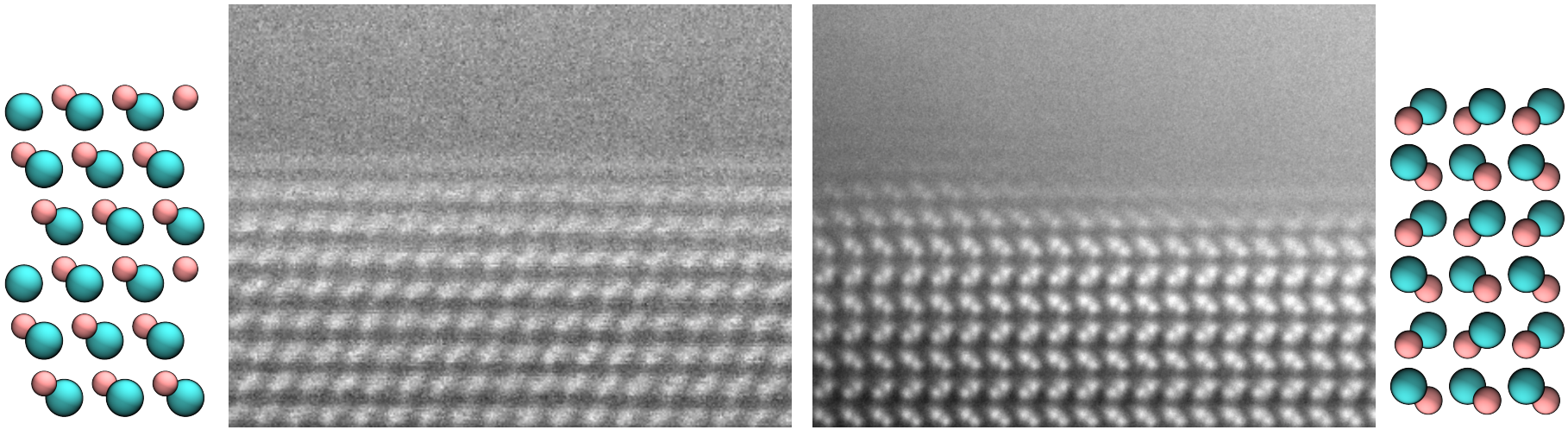
- GaAs nanowires can be grown with solid-liquid-gas process based on droplets of molten Ga
- Problem: controlling the zinc-blende/wurtzite polymorphism.
- Important role played by surface polarity
- MD simulation with DFT/NN to study liquid ordering at the interfaces



w/ Giulio Imbalzano, Mahdi Zamani, Anna Fontcuberta, CIME

# Synthesis of GaAs nanowires

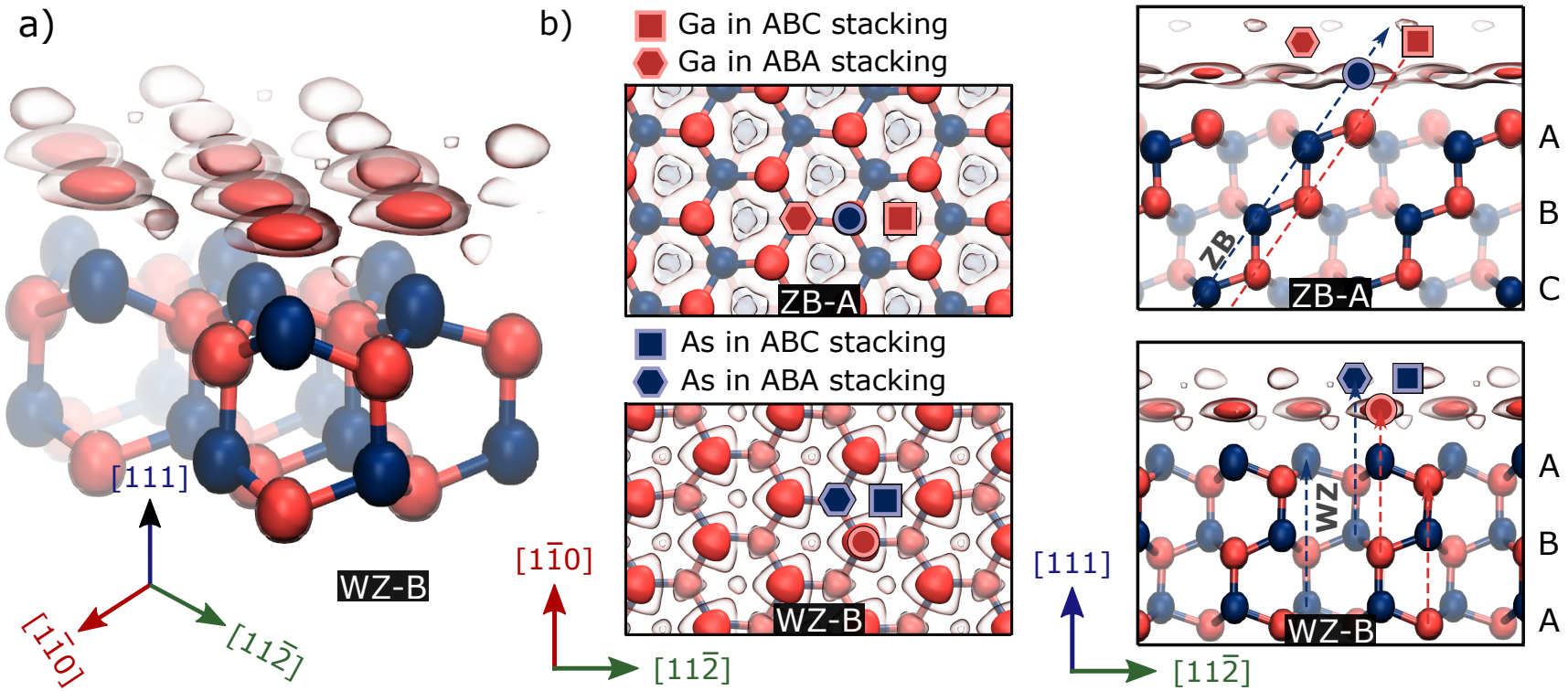
- GaAs nanowires can be grown with solid-liquid-gas process based on droplets of molten Ga
- Problem: controlling the zinc-blende/wurtzite polymorphism.
- Important role played by surface polarity
- MD simulation with DFT/NN to study liquid ordering at the interfaces



w/ Giulio Imbalzano, Mahdi Zamani, Anna Fontcuberta, CIME

# Synthesis of GaAs nanowires

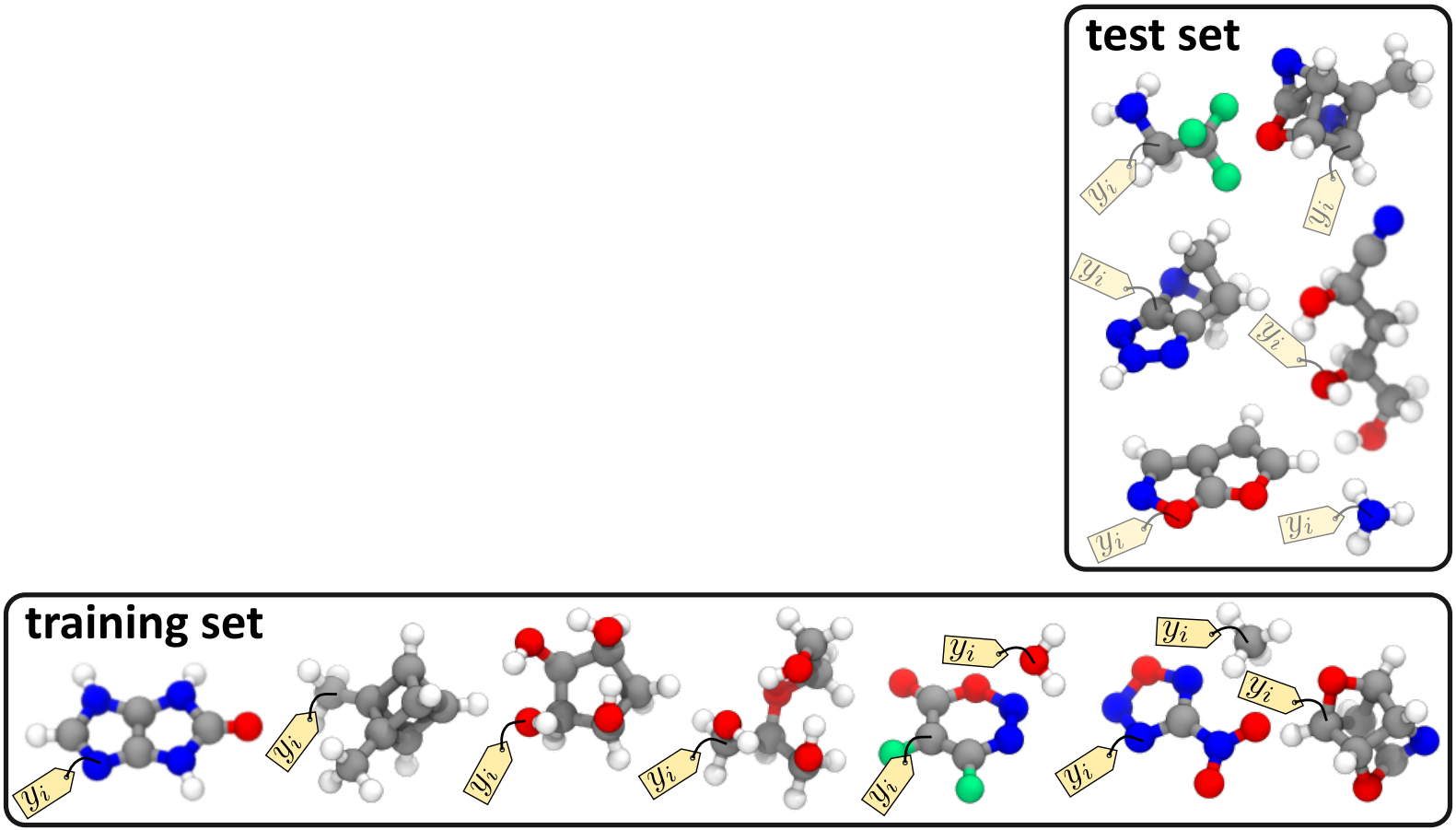
- GaAs nanowires can be grown with solid-liquid-gas process based on droplets of molten Ga
- Problem: controlling the zinc-blende/wurtzite polymorphism.
- Important role played by surface polarity
- MD simulation with DFT/NN to study liquid ordering at the interfaces




w/ Giulio Imbalzano, Mahdi Zamani, Anna Fontcuberta, CIME

# Learning to predict

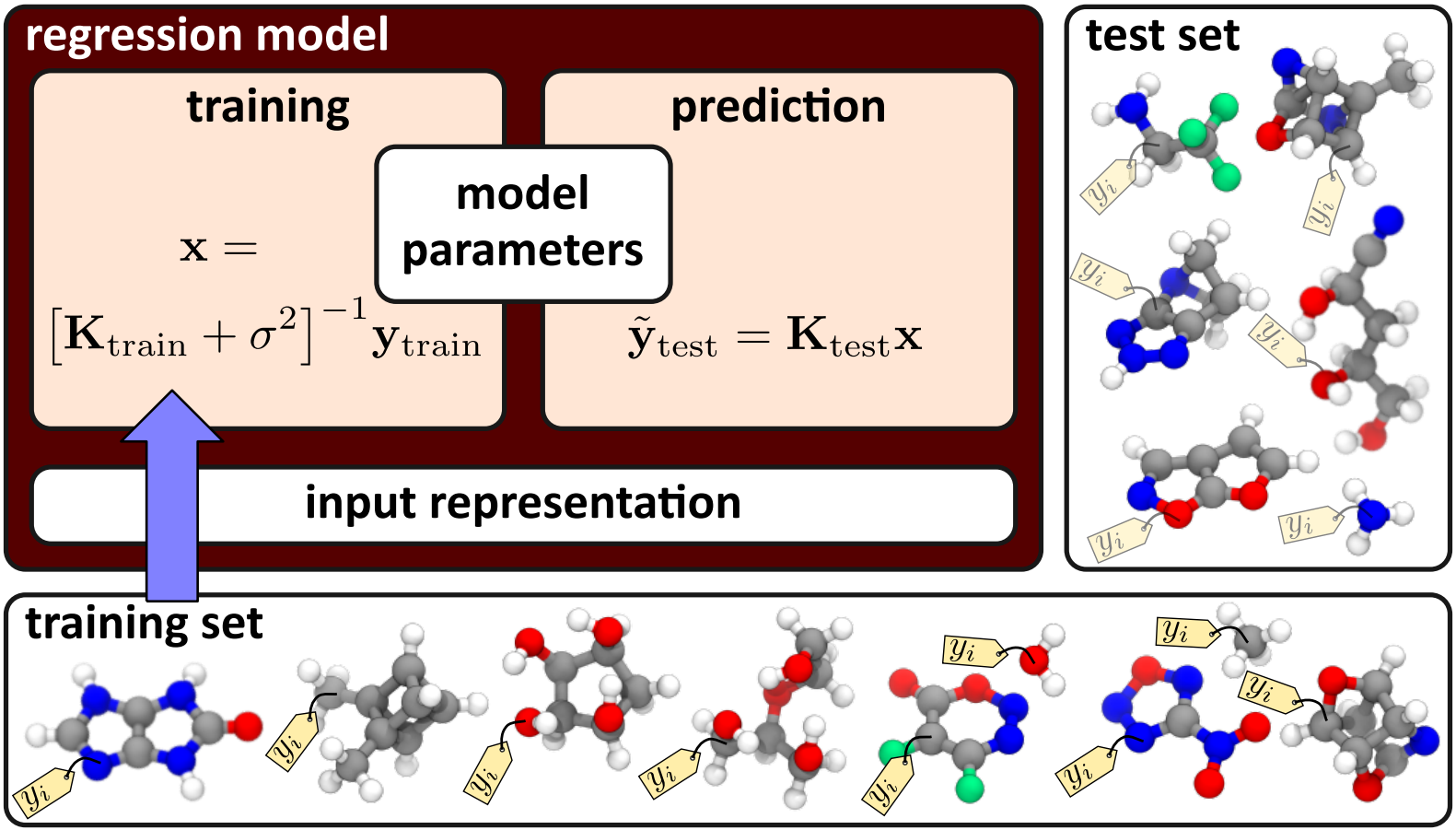
- Data is split between a training set - used to determine the parameters of the model - and a test/validation set used to verify accuracy of predictions
- Learning curves provide diagnostics to understand data and model




**MC**, Willatt, Csányi, Handbook of Materials Modeling, Springer (2018)

# Learning to predict

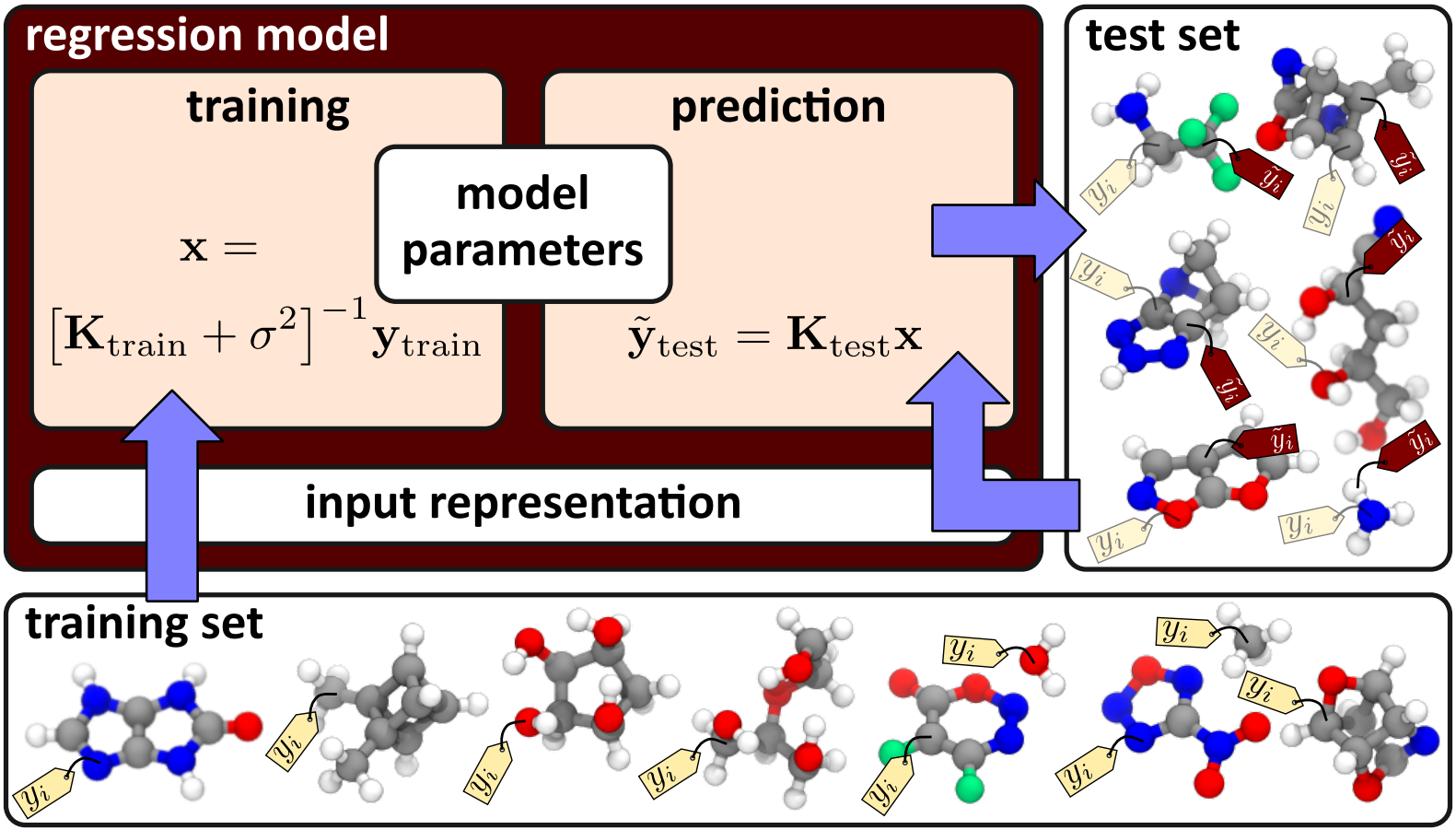
- Data is split between a training set - used to determine the parameters of the model - and a test/validation set used to verify accuracy of predictions
- Learning curves provide diagnostics to understand data and model




**MC**, Willatt, Csányi, Handbook of Materials Modeling, Springer (2018)

# Learning to predict

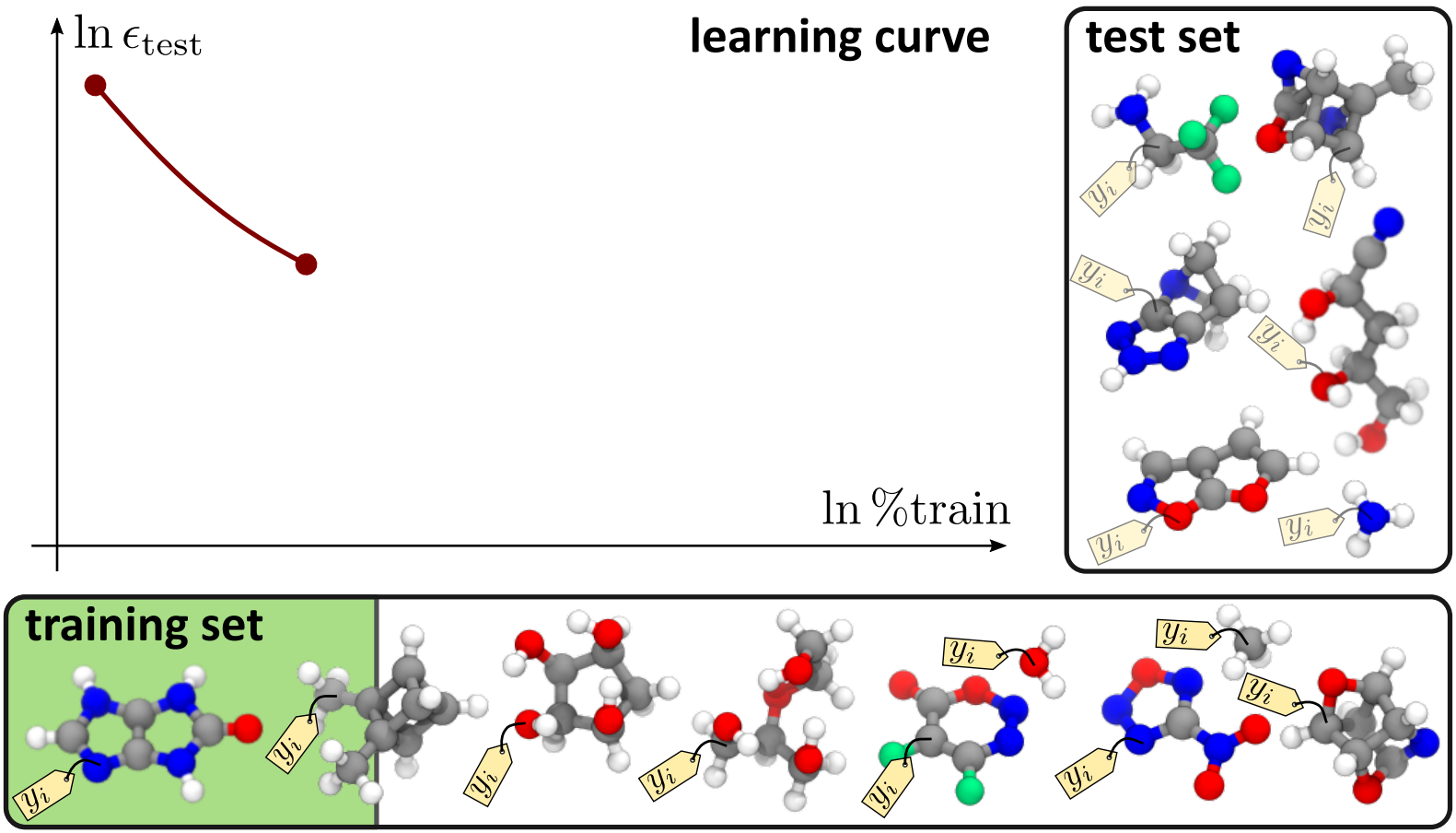
- Data is split between a training set - used to determine the parameters of the model - and a test/validation set used to verify accuracy of predictions
- Learning curves provide diagnostics to understand data and model




**MC**, Willatt, Csányi, Handbook of Materials Modeling, Springer (2018)

# Learning to predict

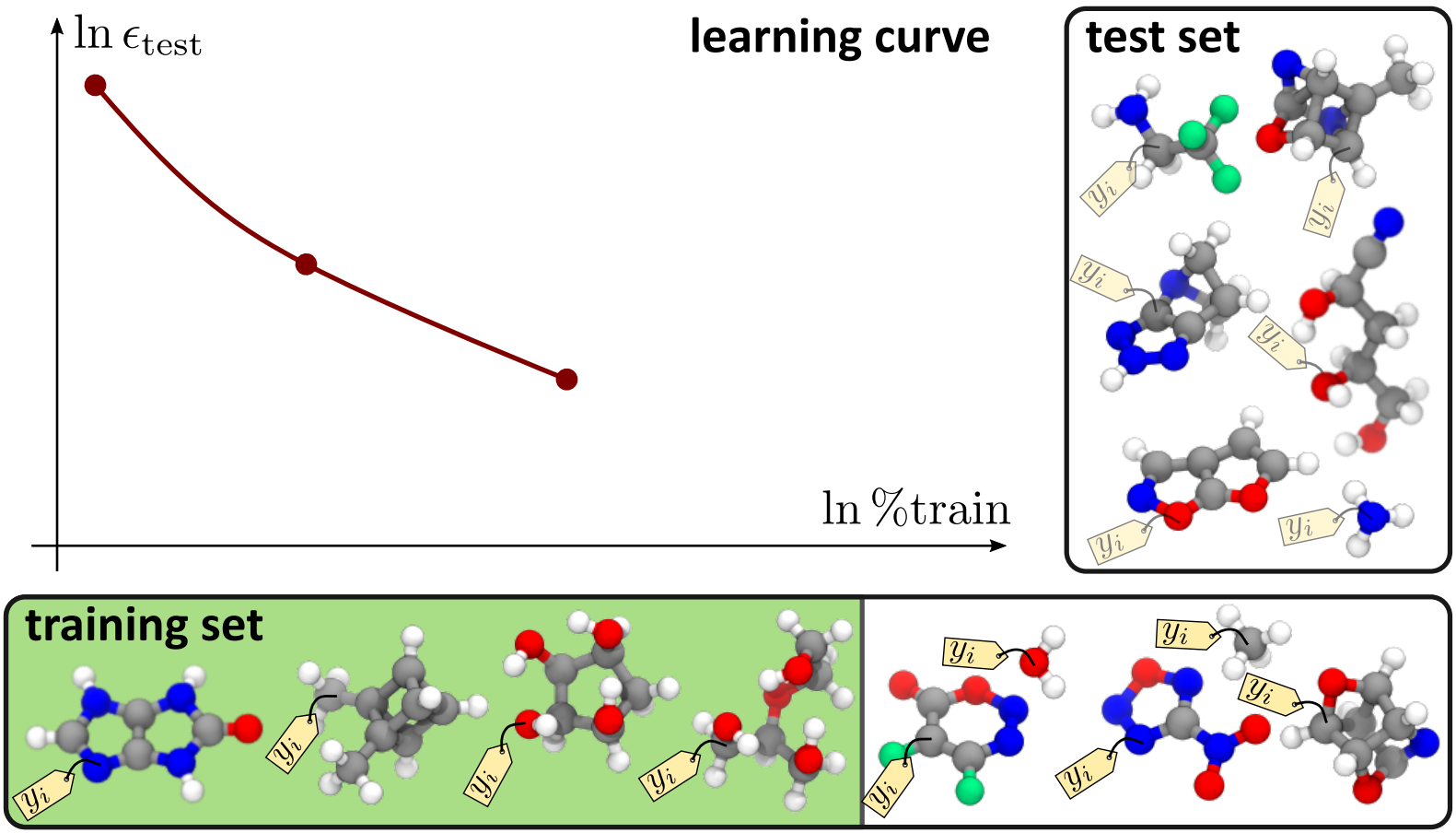
- Data is split between a training set - used to determine the parameters of the model - and a test/validation set used to verify accuracy of predictions
- Learning curves provide diagnostics to understand data and model




Huang, von Lilienfeld, JCP (2016)

# Learning to predict

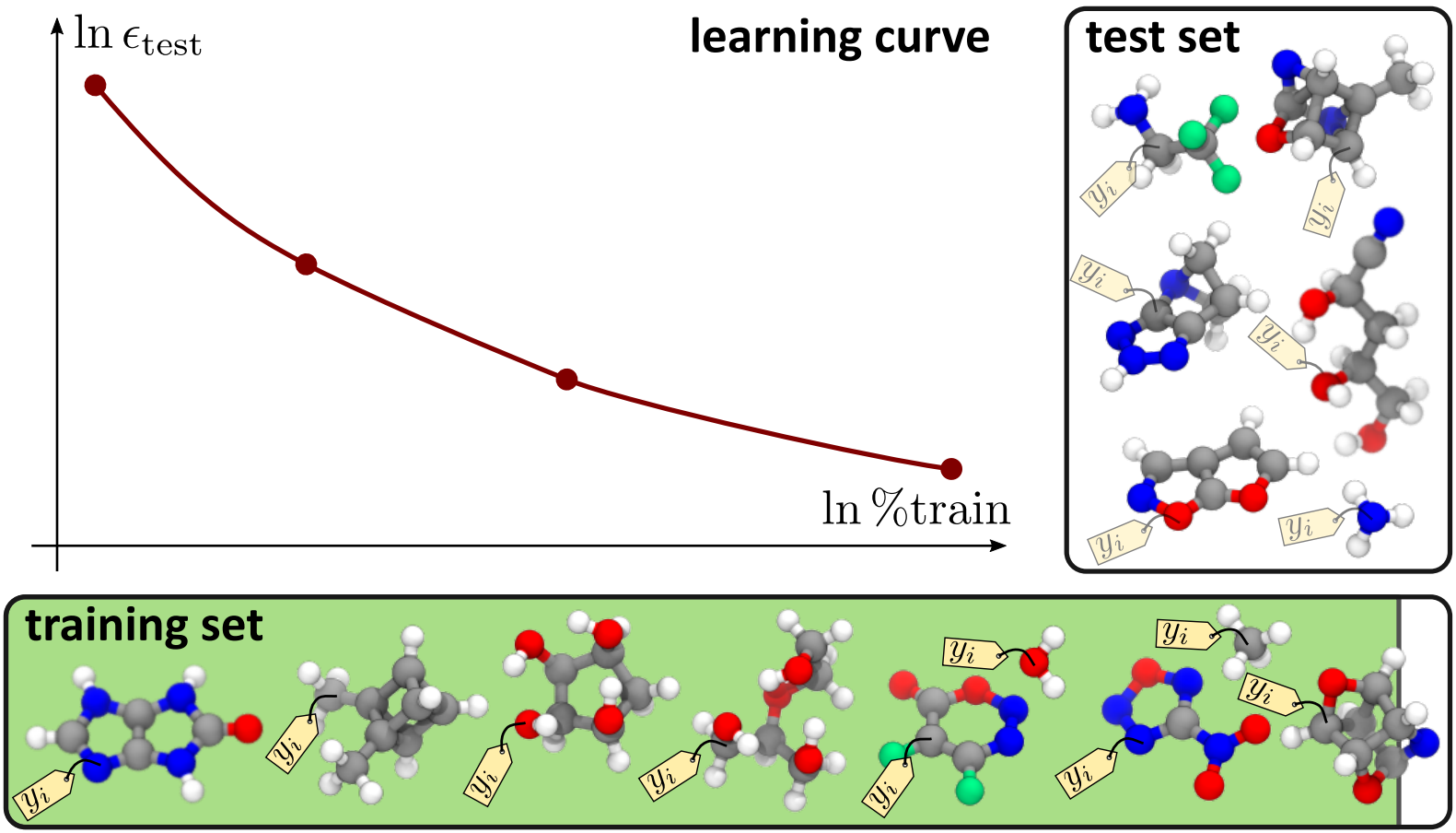
- Data is split between a training set - used to determine the parameters of the model - and a test/validation set used to verify accuracy of predictions
- Learning curves provide diagnostics to understand data and model




Huang, von Lilienfeld, JCP (2016)

# Learning to predict

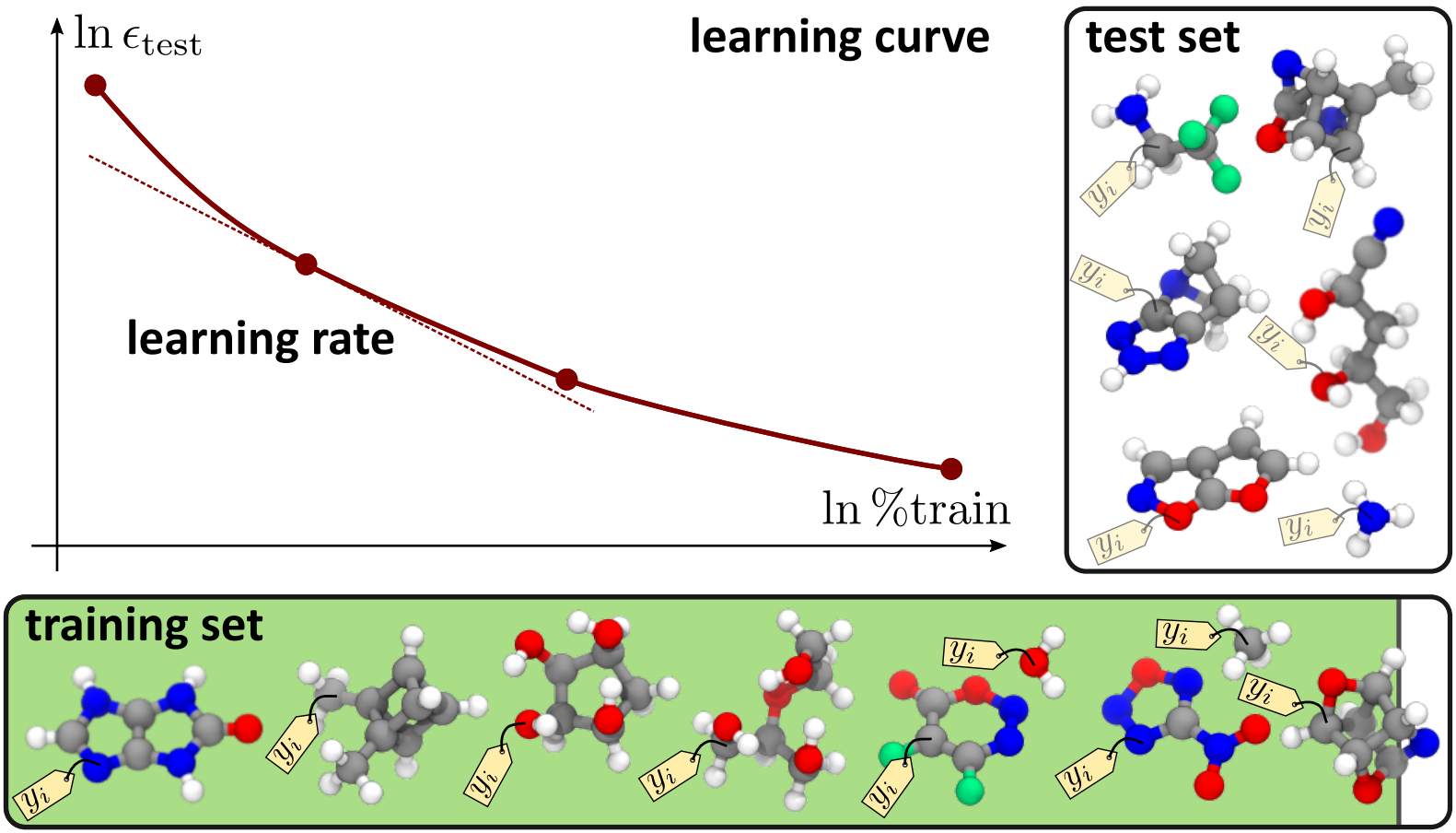
- Data is split between a training set - used to determine the parameters of the model - and a test/validation set used to verify accuracy of predictions
- Learning curves provide diagnostics to understand data and model




Huang, von Lilienfeld, JCP (2016)

# Learning to predict

- Data is split between a training set - used to determine the parameters of the model - and a test/validation set used to verify accuracy of predictions
- Learning curves provide diagnostics to understand data and model




Huang, von Lilienfeld, JCP (2016)