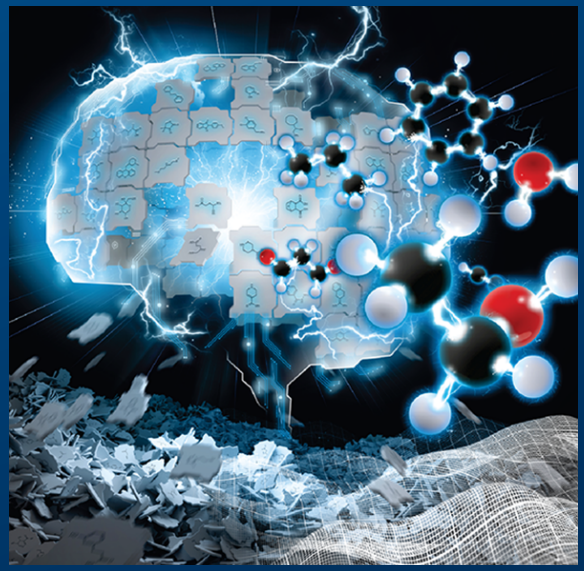
AIDD Spring School: Advanced Machine Learning for Drug Discovery

Constructing accurate machine learning force fields for flexible molecules

Leonardo Medrano Sandonas

Department of Physics and Materials Science,
University of Luxembourg, Luxembourg

May 10, 2022



O.T. Unke *et al.*, *Chem. Rev.* **121**, 10142-10186, (2021).

Email: leonardo.medrano@uni.lu

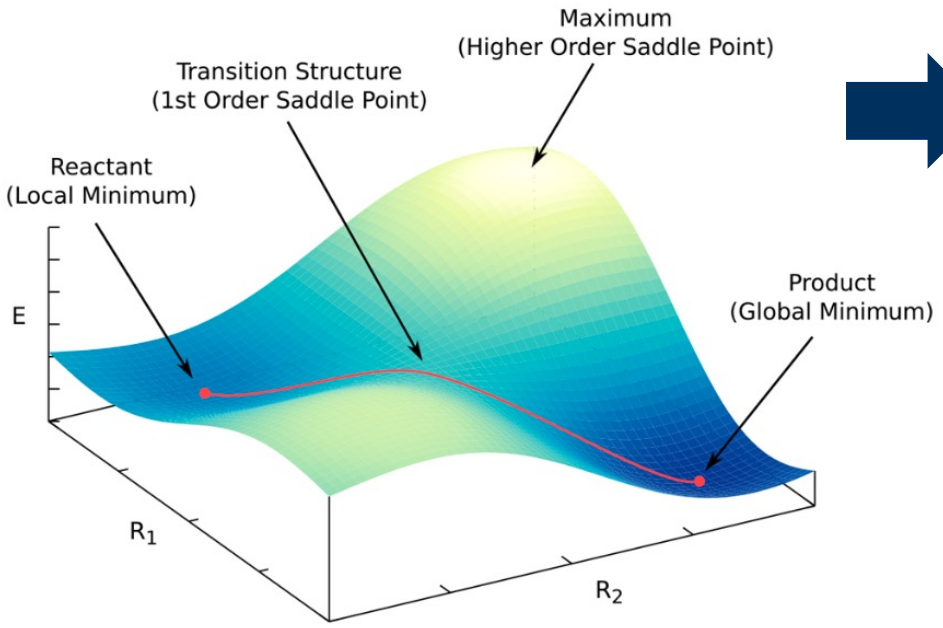
The role of system dimensionality

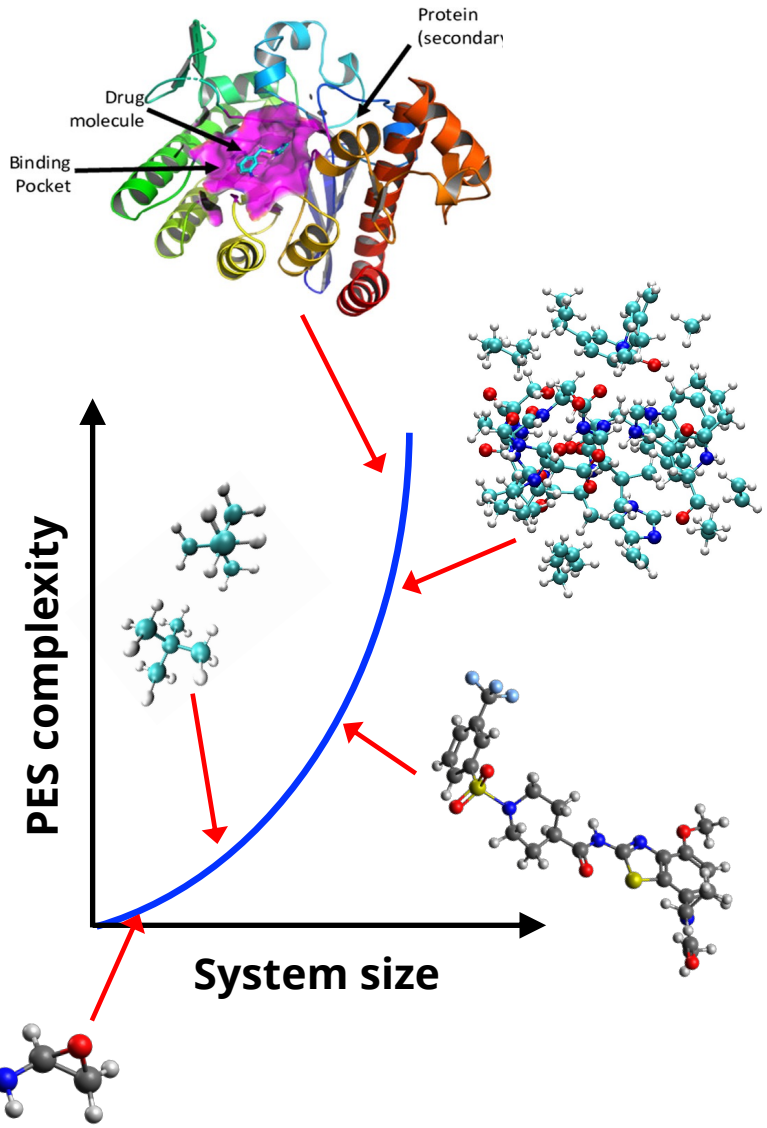
Potential energy surfaces (PES)

Describe the energy of a system (molecule) in terms of certain parameters (positions, bonds, angles, etc).

$$\text{PES: } U(R_1, R_2)$$

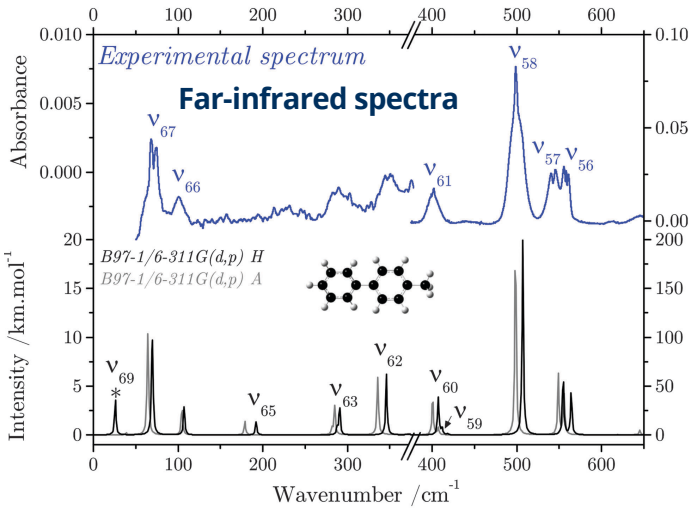
$$\vec{F} = -\nabla U$$





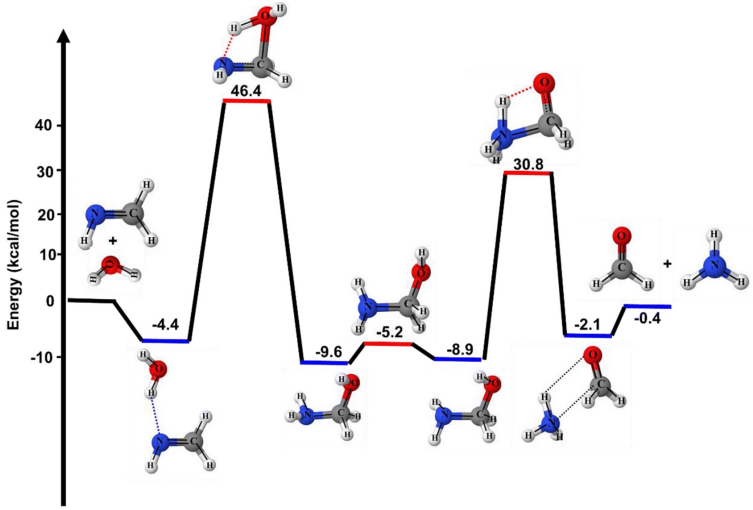
Potential energy surface (PES): what's good for?

Vibrational spectrum



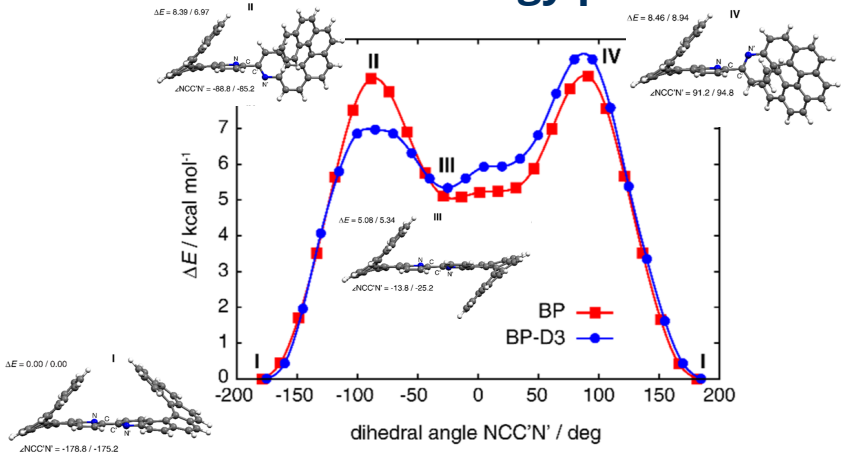
M.A. M.-Drumel *et al.*, *Phys. Chem. Chem. Phys.* **16**, 22062, (2014).

Chemical reactions



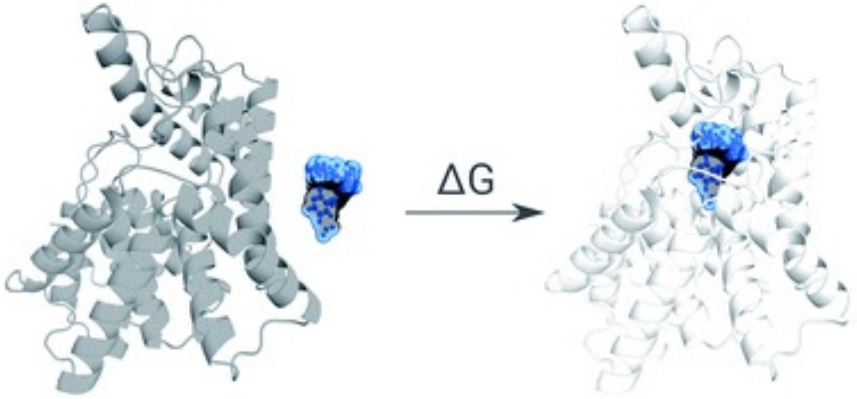
Y. Ali, *Sci. Rep.* **10**, 10995, (2020).

Rotational energy profiles



H. Isla *et al.*, *J. Org. Chem.* **84**, 5383, (2019).

Thermodynamic binding constants



Y. Khalak *et al.*, *Chem. Sci.* **12**, 13958, (2021).

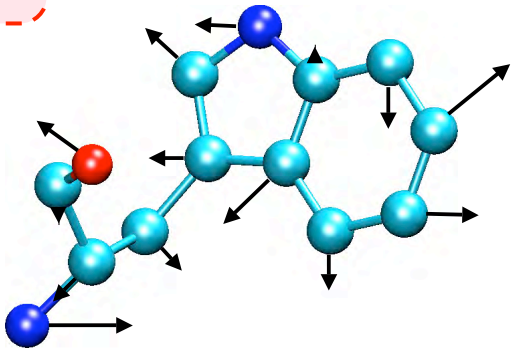
Potential energy surface (PES)

"Original" energy function for molecules: Molecular mechanics (MM)

$$U = \underbrace{\sum_{bonds} k_i^{bond} (r_i - r_0)^2}_{U_{\text{bond}}} + \underbrace{\sum_{angles} k_i^{angle} (\theta_i - \theta_0)^2}_{U_{\text{angle}}} +$$

$$\underbrace{\sum_{dihedrals} k_i^{dihe} [1 + \cos(n_i \phi_i + \delta_i)]}_{U_{\text{dihedral}}} +$$

$$\underbrace{\sum_i \sum_{j \neq i} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \sum_i \sum_{j \neq i} \frac{q_i q_j}{\epsilon r_{ij}}}_{U_{\text{non-bond}}}$$



U_{bond} = Oscillations about the equilibrium bond length

U_{angle} = Oscillations of 3 atoms about an equilibrium bond angle

U_{dihedral} = Torsional rotation of 4 atoms about a central bond

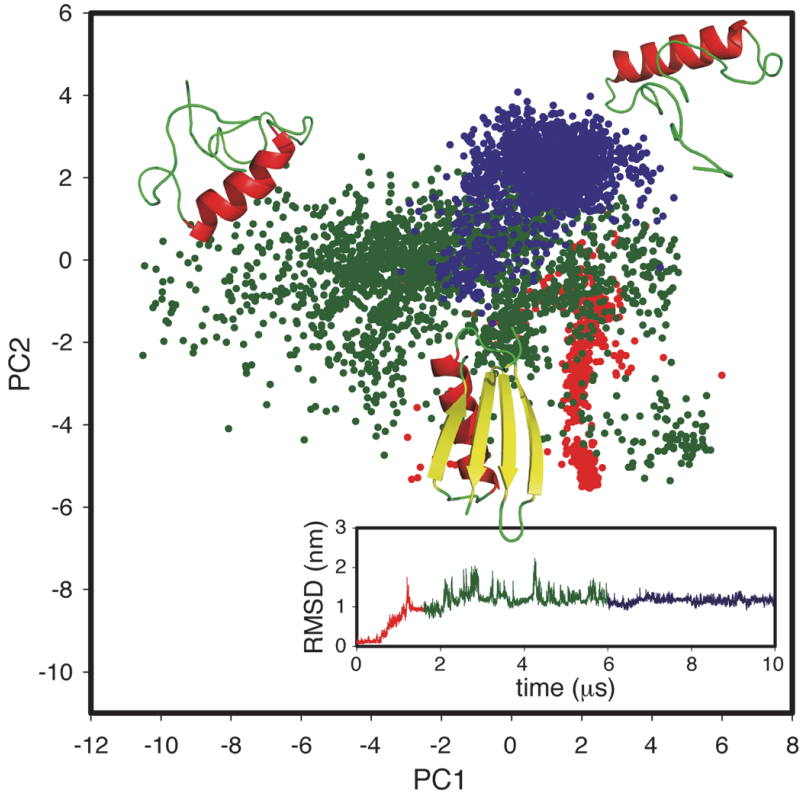
$U_{\text{non-bond}}$ = Non-bonded energy terms (Lennard-Jones and electrostatics)

Potential energy surface (PES)

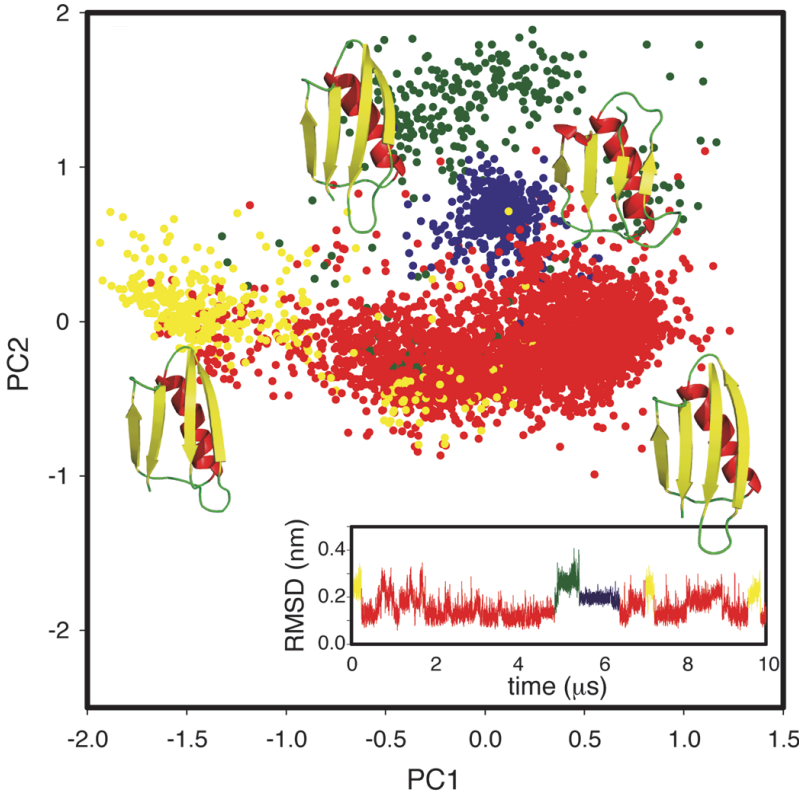
“Original” energy function for molecules: Molecular mechanics (MM)

Analysis of the conformational sampling of B3 domain of Protein G (GB3)

CHARMM22



Amber-ff03*

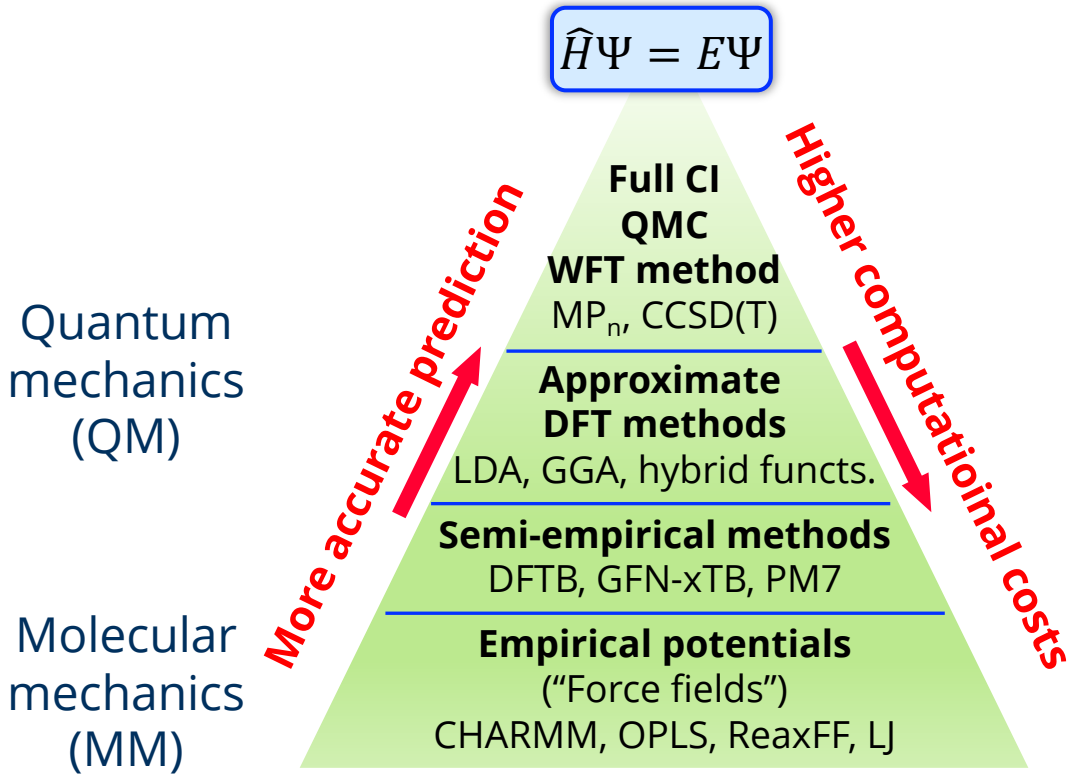


Conformations and energies highly depend on the chosen force field.

F. Martin-Garcia *et al.*, *PLoS One* **10**, 3, (2015).

Potential energy surface (PES)

Hierarchies in atomistic modeling methods



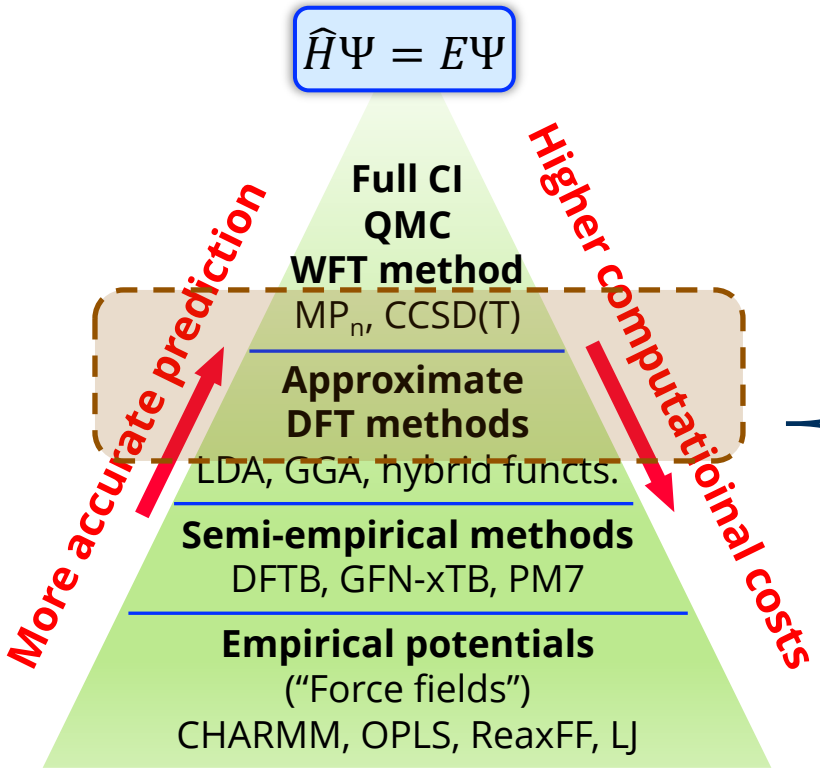
http://quantum-machine.org
J. Hoja *et al.*, *Sci. Data* **8**, 43, (2021).
J.S. Smith *et al.*, *Sci. Data* **4**, 170193, (2017).
J. Rezac *et al.*, *J. Chem. Theory Comput.* **7**, 8, (2011).
Z.M. Sparrow *et al.*, *J. Chem. Phys.* **155**, 184303, (2021).

Potential energy surface (PES)

Hierarchies in atomistic modeling methods

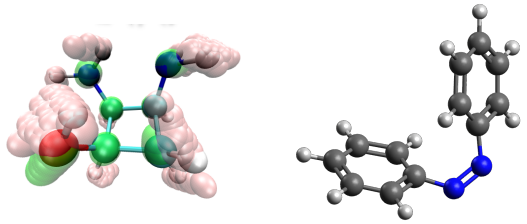
Quantum mechanics (QM)

Molecular mechanics (MM)

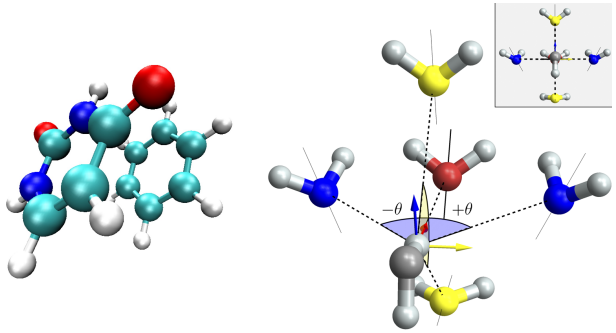


QM datasets of small molecules for PES prediction

- Single molecules
  - QM7-X (PBE0+MBD)
  - ANI-1 (ω97x/6-31G(d))
  - MD17 (PBE+TS)



- Molecular dimers
  - S66x8 (CCSD(T)/CBS)
  - NENCI (CCSD(T)/CBS)

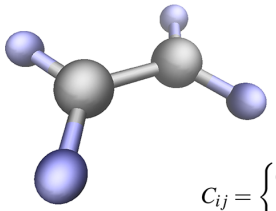


http://quantum-machine.org
J. Hoja *et al.*, *Sci. Data* **8**, 43, (2021).
J.S. Smith *et al.*, *Sci. Data* **4**, 170193, (2017).
J. Rezac *et al.*, *J. Chem. Theory Comput.* **7**, 8, (2011).
Z.M. Sparrow *et al.*, *J. Chem. Phys.* **155**, 184303, (2021).

QM dataset of small molecules

Molecular representations (3D geometric descriptors)

❑ Coulomb matrix



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

$$\mathbf{C} =$$

	H	H	C	C	H	H
H	0.5	0.3	2.9	1.5	0.2	0.2
H	0.3	0.5	2.9	1.5	0.2	0.2
C	2.9	2.9	36.9	14.3	1.5	1.5
C	1.5	1.5	14.3	36.9	2.9	2.9
H	0.2	0.2	1.5	2.9	0.5	0.3
H	0.2	0.2	1.5	2.9	0.3	0.5

❑ FCHL19 (previous FCHL18)

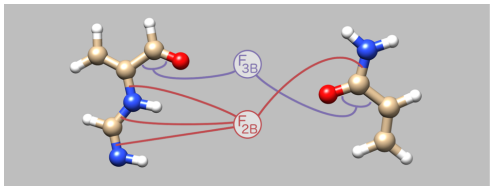
Two-body term:

$$G^{\text{2-body}} = \xi_2(r_{IJ}) f_{\text{cut}}(r_{IJ}) \frac{1}{R_s \sigma(r_{ij})\sqrt{2\pi}} \exp\left(-\frac{(\ln R_s - \mu(r_{ij}))^2}{2\sigma(r_{ij})^2}\right),$$

Three-body term:

$$G^{\text{3-body}} = \xi_3 G^{\text{3-body}}_{\text{Radial}} G^{\text{3-body}}_{\text{Angular}} f_{\text{cut}}(r_{IJ}) f_{\text{cut}}(r_{JK}) f_{\text{cut}}(r_{KI}).$$

❑ Spectrum of London and Axilrod-Teller-Muto potential (SLATM)



Two-body term:

$$\frac{1}{2} \sum_{J \neq I} Z_J \delta(\mathbf{r} - \mathbf{R}_{IJ}) g(\mathbf{r})$$

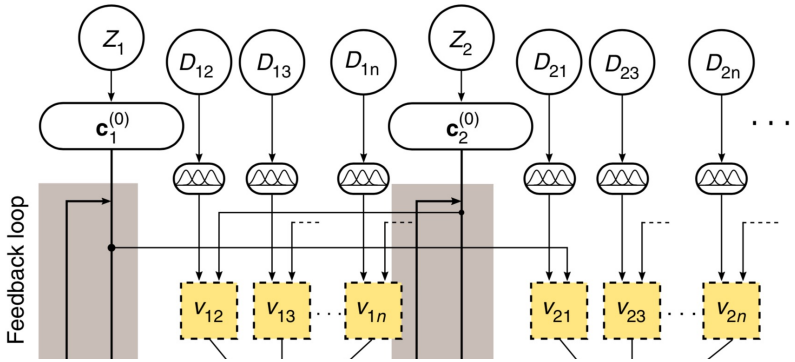
Three-body term:

$$\frac{1}{3} \sum_{J \neq K \neq I} Z_J Z_K \delta(\theta - \theta_{IJK}) h(\theta, \mathbf{R}_{IJ}, \mathbf{R}_{IK})$$

❑ Neural Network representation (SchNet)

➢ Distances are expanded with radial basis functions,

$$e_k(\mathbf{r}_i - \mathbf{r}_j) = \exp(-\gamma \| d_{ij} - \mu_k \|^2)$$



➢ Many-body atomic interactions.

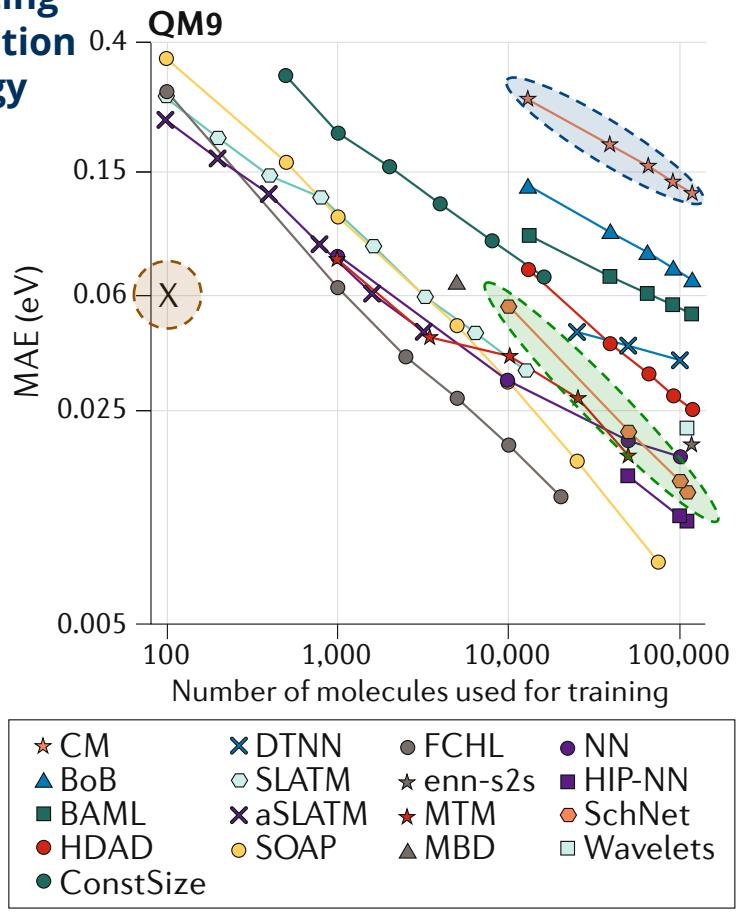
K. Hansen *et al.*, *J. Chem. Theory Comput.* **9**, (2013).
W. Pronobis *et al.*, *J. Chem. Theory Comput.* **14**, (2018).

QM dataset of small molecules

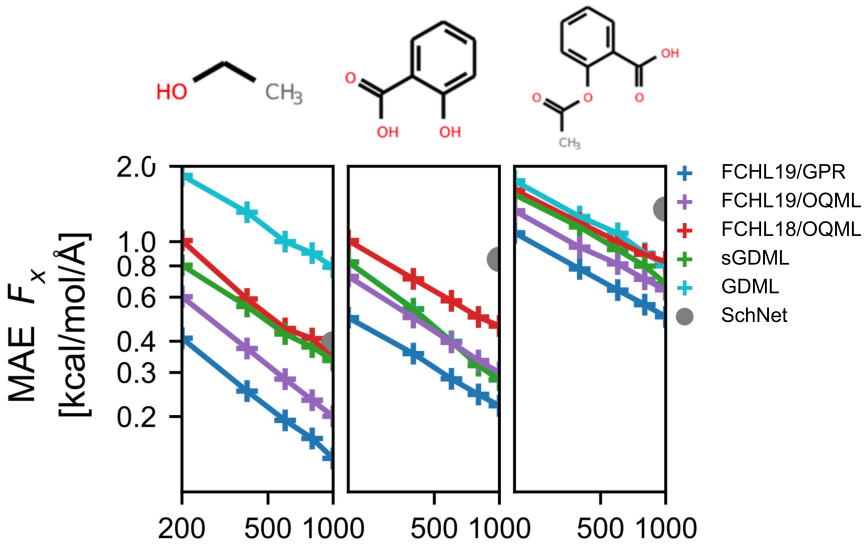
Molecular representations (3D geometric descriptors)

❑ Physicochemical properties

Predicting atomization energy



❑ Atomic forces prediction (MD17 dataset)



Shortcomings for application in **large molecules:**

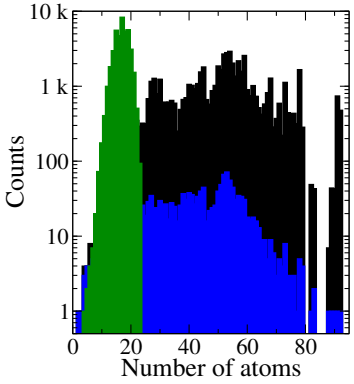
- ➢ High-dimensional geometric descriptors.
- ➢ Large degrees of freedom.
- ➢ Strong non-covalent interaction.

O.A. von Lilenfeld *et al.*, *Nat. Rev. Chem.* **4**, 347, (2020).
F. Faber *et al.*, *J. Chem. Phys.* **148**, 241717, (2018).

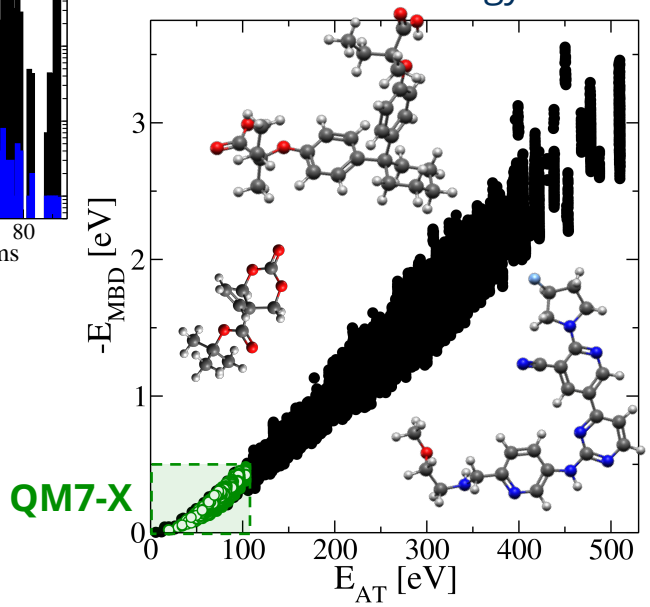
Problems to consider when studying large systems

Long range interactions

- **UniLu-Janssen dataset**
- **60,082** molecular conformations (**1673** unique compositions). Elements: H, C, N, O, S, Cl, P, F.
- Structures containnig up to **92** atoms (**54** non-hydrogen/heavy atoms).
- **~43 QM properties:** PBE0(tight)+many body dispersion (MBD) with MPB implicit solvent

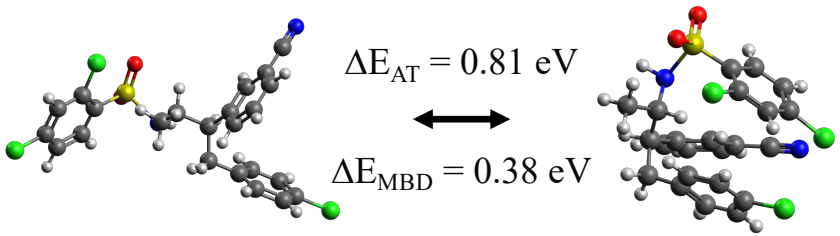


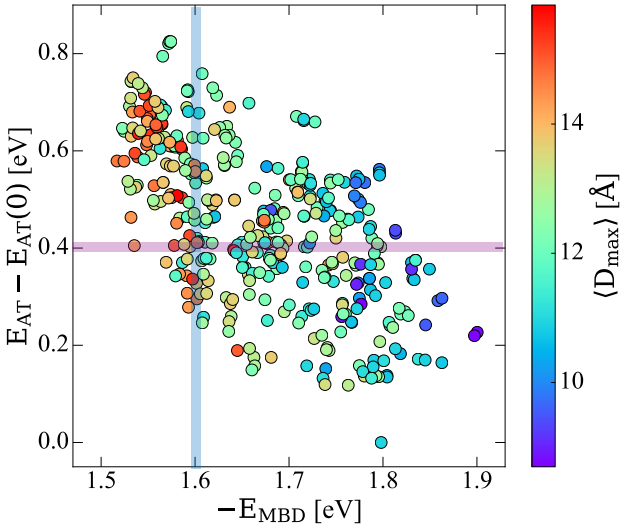
Atomization energy vs MBD energy



Conformational sampling

$C_{19}H_{23}N_2O_2SCl_3$ (50 atoms, 400 conformers)

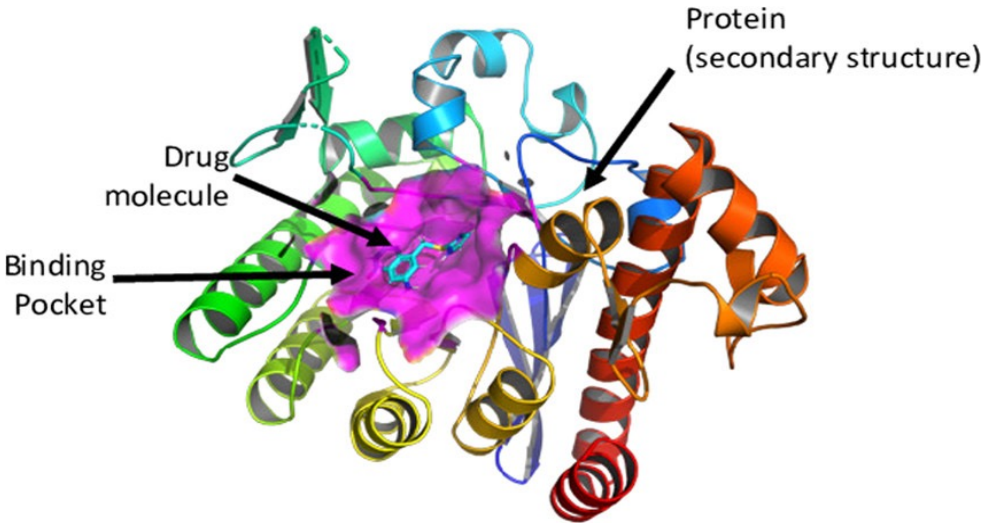




- Representations must consider long range terms (vdW, electrostatics).
- Conformer identification is more challenging by only considering energies.

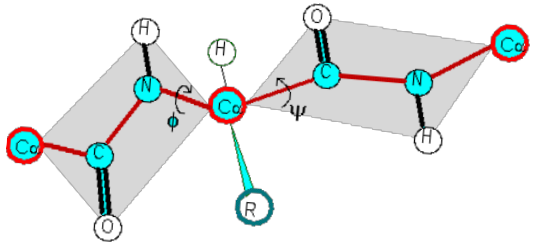
Problems to consider when studying large systems

Much larger and complex molecular systems



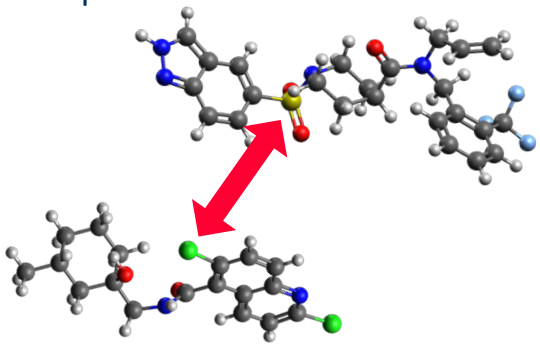
Intramolecular interaction

Electrostatic, chemical bonds



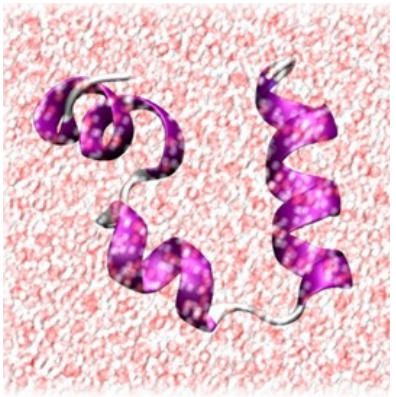
Intermolecular interaction

Electrostatic, dispersion, polarization effects



Interaction with chemical environment

Solvation

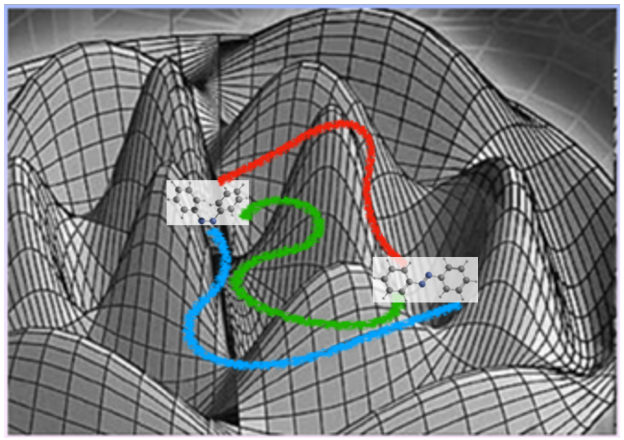


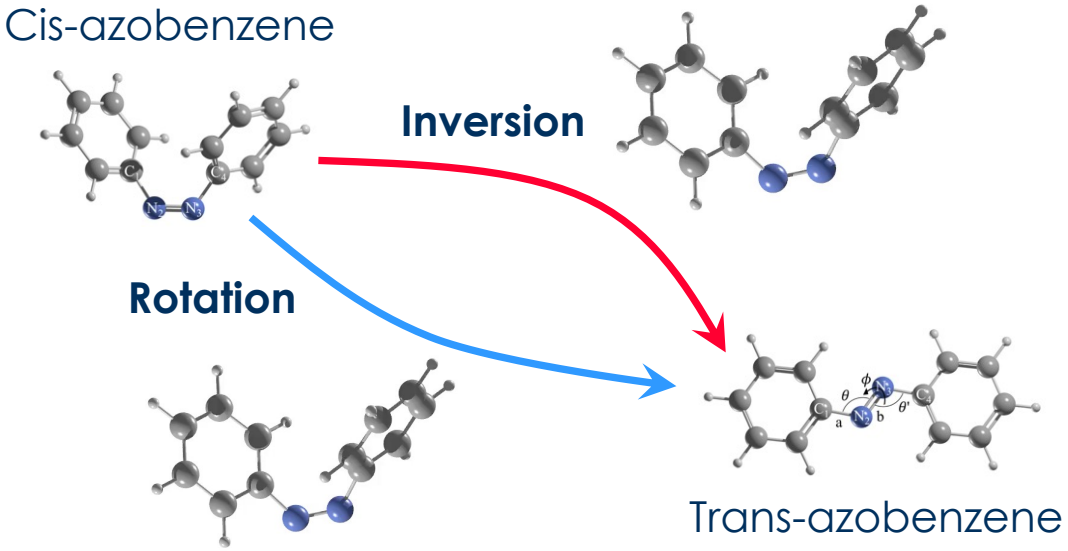
M. Stöhr and A. Tkatchenko, *Sci. Adv.* **5**, 12, (2019).

Methods for developing machine learning force fields

Method 1: Kernel ridge regression

Isomerization: Azobenzene reaction paths



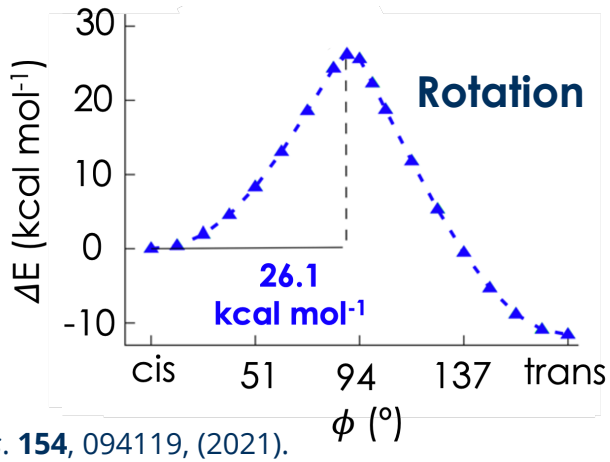


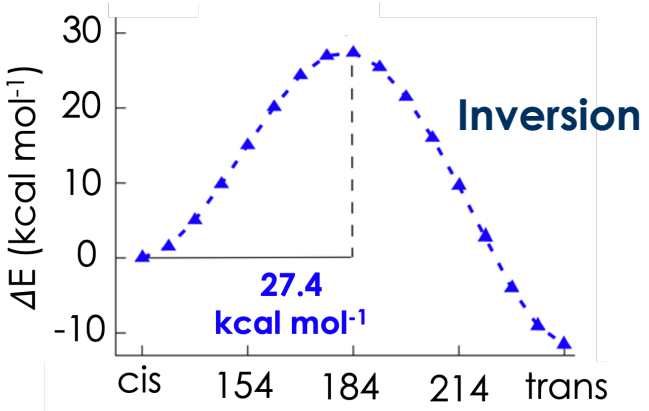
Transition paths: reference data

PBE+TS / Tight

FHI-aims
The ab initio materials simulation package

i-PI





V. Vassilev-Galindo *et al.*, *J. Chem. Phys.* **154**, 094119, (2021).

Method 1: Kernel ridge regression

Isomerization: Azobenzene reaction paths

Training/Testing scheme

- Training sets from 100 up to 1000 points.
- Subsets of size equal to five times the number of training points.
- 5-fold cross-validation on each subset.
- Used one fold for training and the rest for testing.

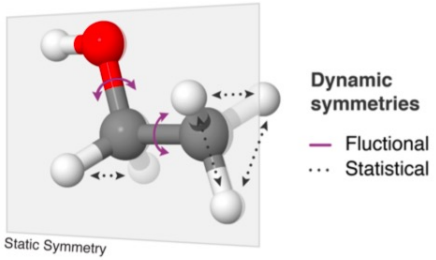
KRR procedures

sGDML

Global kernel-based

Descriptor

$$D_{ij} = ||\vec{R}_i - \vec{R}_j||^{-1}$$



Kernel

$$\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}') = \nabla_{\boldsymbol{x}} K_E(\boldsymbol{x}, \boldsymbol{x}') \nabla_{\boldsymbol{x}'}^T$$

S. Chmiela *et al.*, *Sci. Adv.* **3**, 1603015, (2017).

GAP/SOAP

Kernel-based with cut-off

Descriptor

$$\rho_{X_i}(\boldsymbol{r}) = \sum_{i \in X} exp\left[-\frac{|\boldsymbol{r} - \boldsymbol{r}_i|^2}{2\sigma^2}\right] f_{cut}(|\boldsymbol{r}|)$$

Kernel

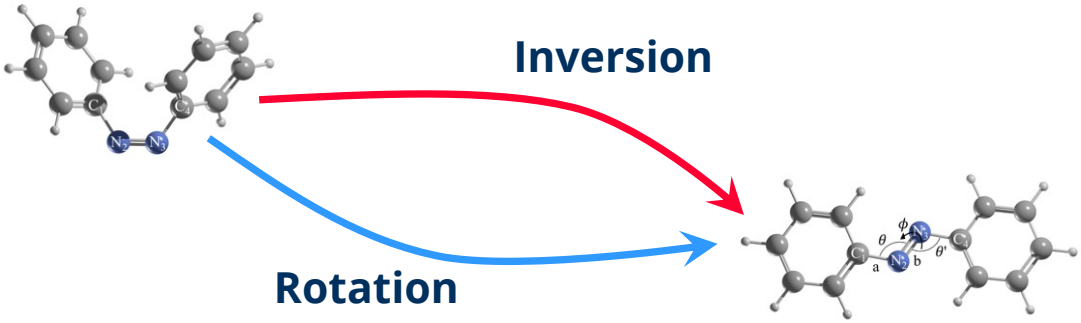
$$K(\rho, \rho') = \int d\mathcal{R} \left| \int \rho(\boldsymbol{r}) \rho'(\mathcal{R}\boldsymbol{r}) d\boldsymbol{r} \right|^n$$

A.P. Bartók *et al.*, *Phys. Rev. Lett.* **104**, 136403, (2010).

V. Vassilev-Galindo *et al.*, *J. Chem. Phys*. **154**, 094119, (2021).

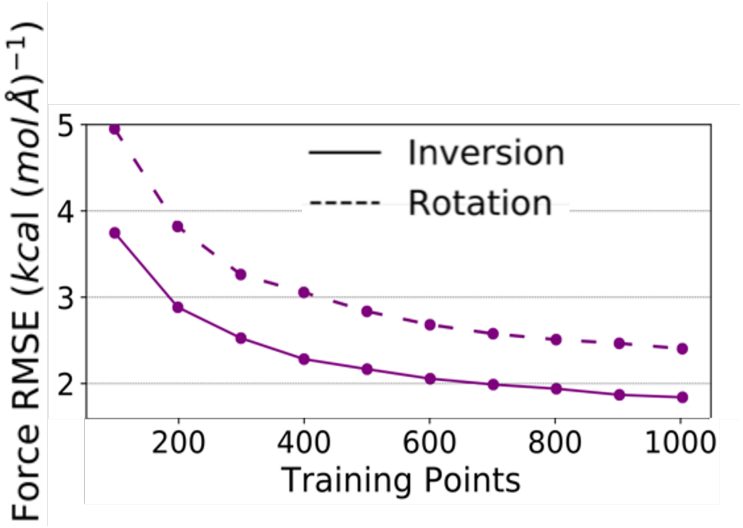
Method 1: Kernel ridge regression

Isomerization: Azobenzene reaction paths

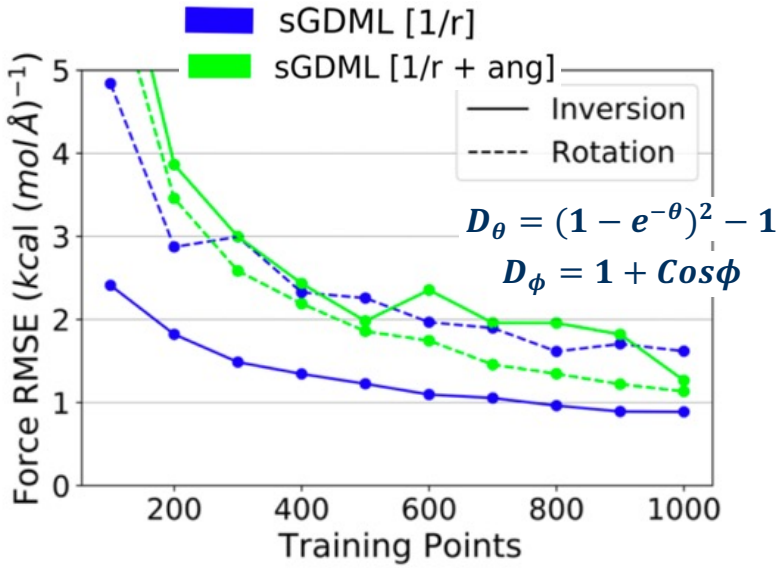


Predicting atomic forces

GAP/SOAP



sGDML



A single descriptor is not able to optimally resolve all different states on a PES.

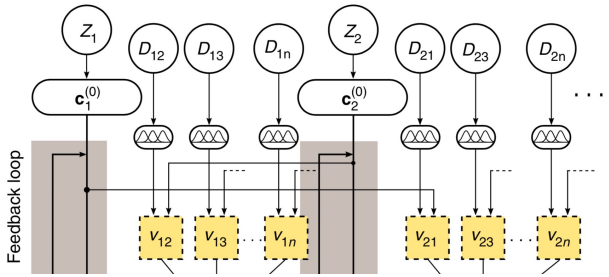
V. Vassilev-Galindo *et al.*, *J. Chem. Phys*. **154**, 094119, (2021).

Method 2: First neural network architectures

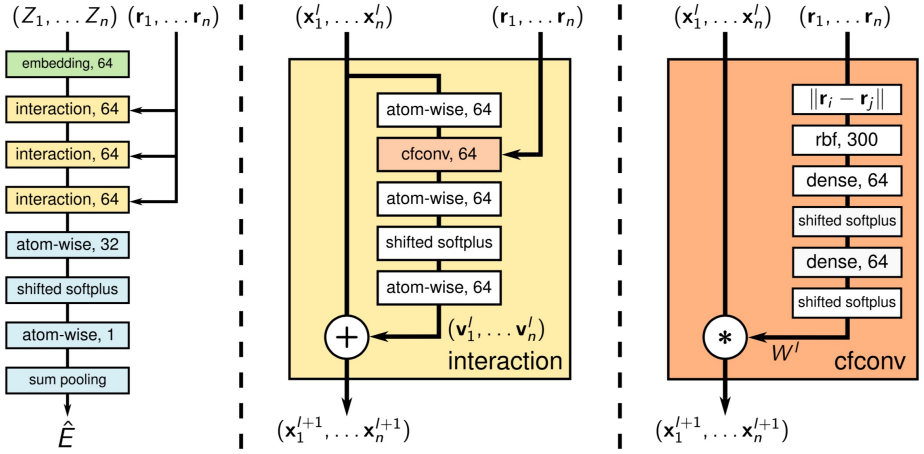
SchNetPack: end-to-end NN with cut-off

- Distances are expanded with radial basis functions,

$$e_k(\mathbf{r}_i - \mathbf{r}_j) = \exp(-\gamma \| d_{ij} - \mu_k \|^2)$$



- Many-body atomic interactions.

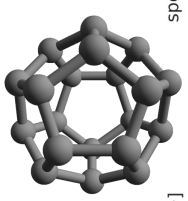


Custom loss function: energies and forces

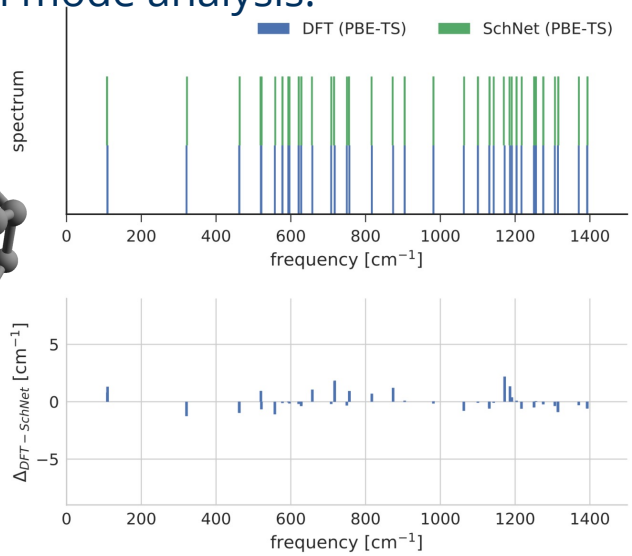
$$\ell(\hat{E}, (E, \mathbf{F}_1, \dots, \mathbf{F}_n)) = \rho \| E - \hat{E} \|^2 + \frac{1}{n} \sum_{i=0}^{n} \left\| \mathbf{F}_i - \left(-\frac{\partial \hat{E}}{\partial \mathbf{R}_i} \right) \right\|^2$$

T	F	ρ	Energy	Forces
3	64	0.010	0.228	0.401
6	64	0.010	0.202	0.217
3	128	0.010	0.188	0.197
6	128	0.010	**0.1002**	**0.120**
6	128	0.100	**0.027**	0.171
6	128	0.010	0.100	0.120
6	128	0.001	0.238	0.061
6	128	0.000	0.260	**0.058**

- Normal mode analysis:



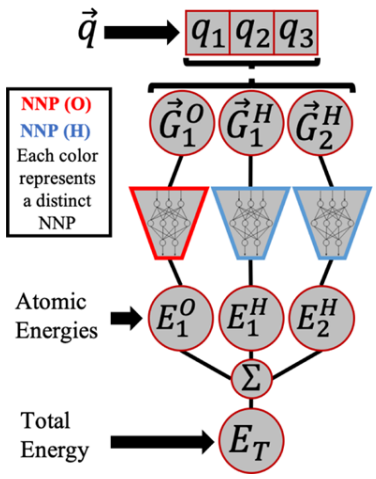
C_{20}



K.T. Schütt *et al.*, *J. Chem. Phys.* **148**, 241722, (2018).

Method 2: First neural network architectures

ANI potentials



- Heavily modified Behler-Parrinello symmetry functions.

$$G_m^{\mathrm{R}} = \sum_{j\neq i}^{\mathrm{all\ atoms}} \mathrm{e}^{-\eta\left(R_{ij}-R_{\mathrm{s}}\right)^2} f_{\mathrm{C}}\left(R_{ij}\right)$$

$$G_m^{\mathrm{A_{mod}}} = 2^{1-\zeta} \sum_{j,k\neq i}^{\mathrm{all\ atoms}} \left(1+\cos\left(\theta_{ijk}-\theta_{\mathrm{s}}\right)\right)^{\zeta} \times \exp\left[-\eta\left(\frac{R_{ij}+R_{ik}}{2}-R_{\mathrm{s}}\right)^2\right] f_{\mathrm{C}}\left(R_{ij}\right) f_{\mathrm{C}}\left(R_{ik}\right)$$

- Each different chemical symbol has a distinct NN.

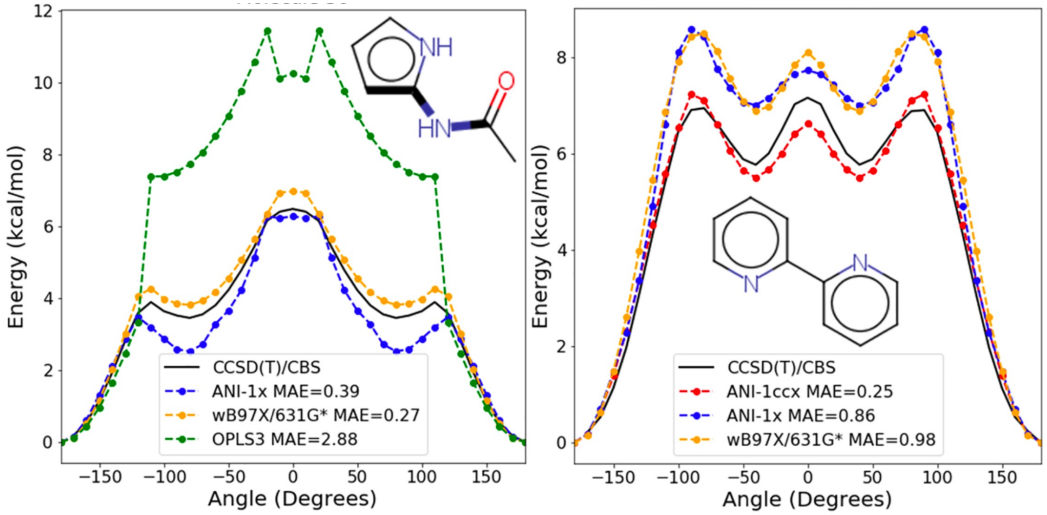
Total loss function:

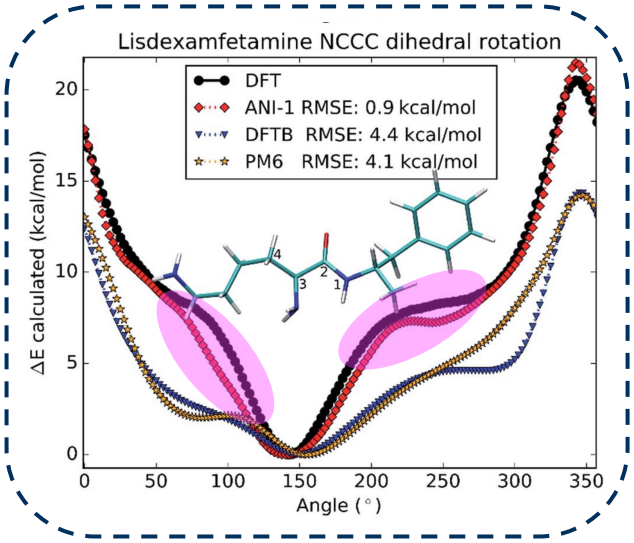
target property

$$L^{\mathrm{tot}} = \frac{1}{N}\sum_{t}^{T}\sum_{p}\sum_{i}^{N} w_t w_p \left(y_{tpi} - \hat{y}_{tpi}\right)^2$$

Atomic forces of the ANI potentials are calculated by using the automatic differentiation feature of PyTorch library.

Rotational energy profiles



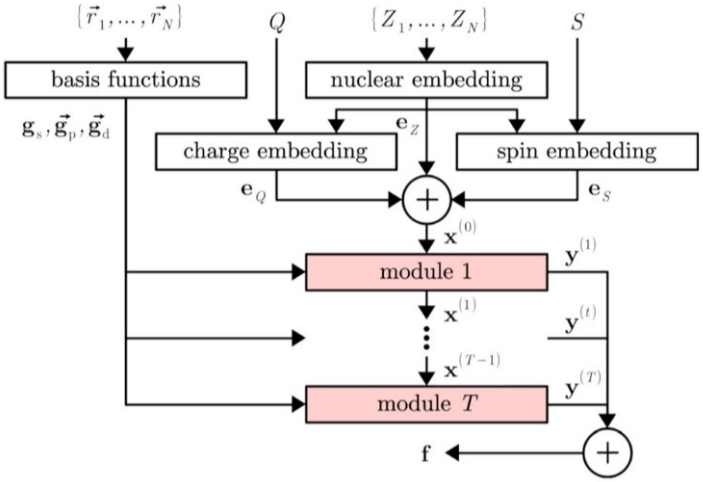


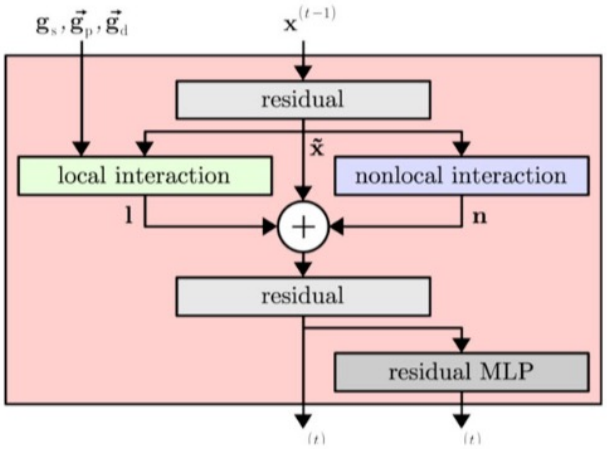
J.S. Smith *et al.*, *Chem. Sci.* **8**, 3192, (2017).

Method 3: Recent Physics-inspired NN potentials

Scalable and accurate ML force field

SpookyNet: It models electronic degrees of freedom and non-local interactions using attention in a transformer architecture.





Physics-based components

$$E_{\text{pot}} = \sum_{i=1}^{N} E_i + E_{\text{rep}} + E_{\text{ele}} + E_{\text{vdw}}$$

$$E_{\text{rep}} = k_e \sum_i \sum_{j>i} \sum_{\in \mathcal{N}(i)} \frac{Z_i Z_j}{r_{ij}} f_{\text{cut}}(r_{ij}) \cdot \left(\sum_{k=1}^{4} c_k e^{-a_k r_{ij} (Z_i^p + Z_j^p)/d} \right) .$$

Repulsive energy

$$E_{\text{ele}} = k_e \sum_i \sum_{j>i} q_i q_j \left(\frac{f_{\text{switch}}(r_{ij})}{\sqrt{r_{ij}^2 + 1}} + \frac{1 - f_{\text{switch}}(r_{ij})}{r} \right)$$

Electrostatic energy

$$E_{\text{vdw}} = -\sum_i \sum_{j>i} \sum_{n=6,8} s_n \frac{C_{(n)}^{ij}}{r_{ij}^n} f_{\text{damp}}^{(n)}(r_{ij})$$

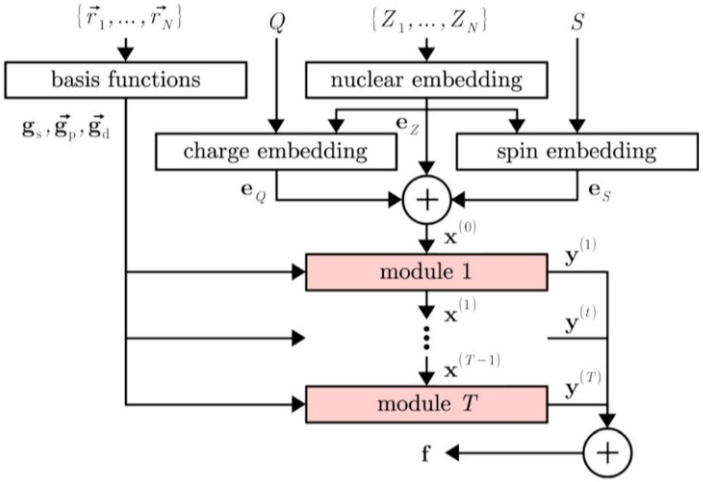
van der Waals energy

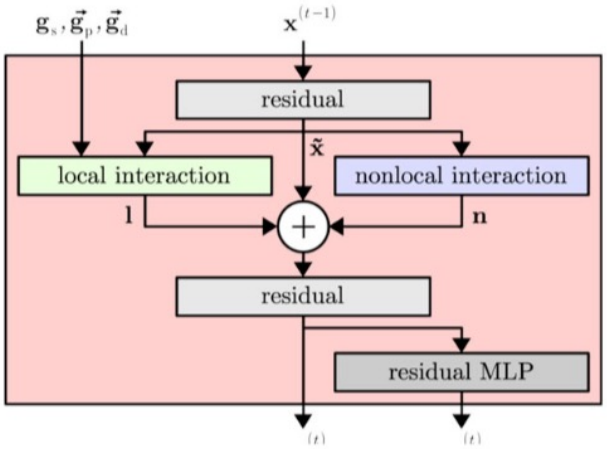
O. T. Unke *et al.*, *Nat. Commun.* **12**, 7273, (2021).

Method 3: Recent Physics-inspired NN potentials

Scalable and accurate ML force field

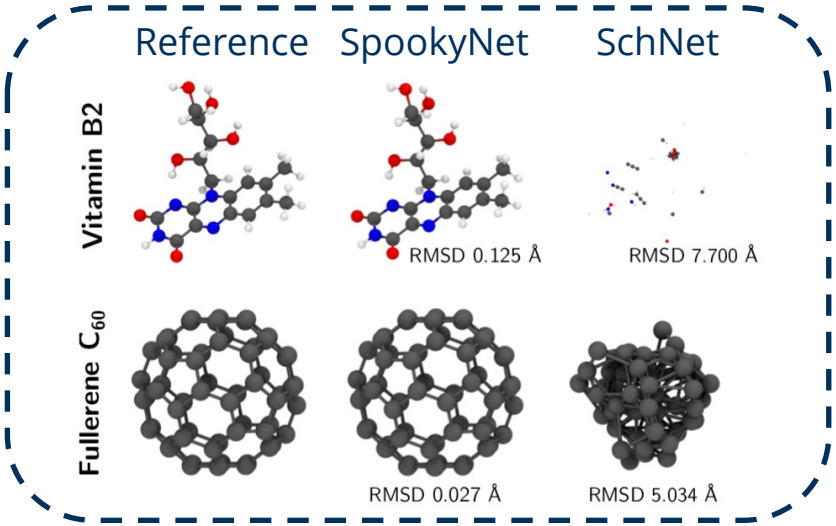
SpookyNet: It models electronic degrees of freedom and non-local interactions using attention in a transformer architecture.





SpookyNet model: QM7-X dataset

- Training set: ~4 M of equilibrium and non-equilibrium molecules up to 23 atoms.
- Level of theory: PBE0+MBD.
- Features: 128.
- Cut-off = 5.29 Å.
- Model parameters: 3 630 142.

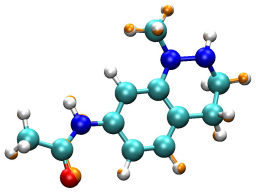
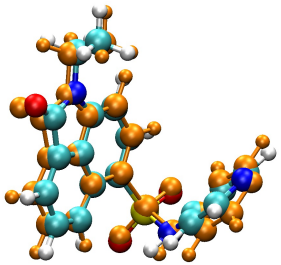
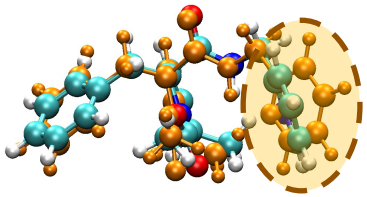
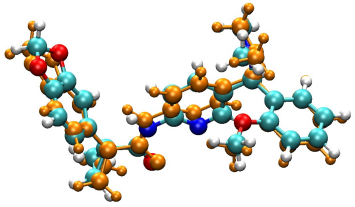
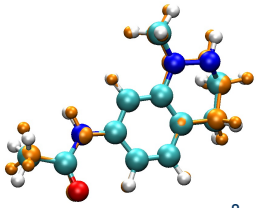
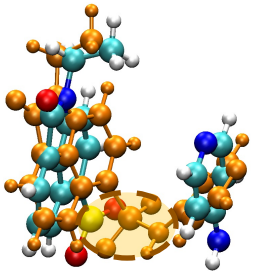


O. T. Unke *et al.*, *Nat. Commun.* **12**, 7273, (2021).

Method 3: Recent Physics-inspired NN potentials

Scalable and accurate ML force field

- Four molecules: 30, 40, 50, and 60 atoms.
- References geometries optimized at PBE0(tight)+MBD.

	$C_{11}H_{15}N_3O$	$C_{18}H_{15}N_3O_3S$	$C_{21}H_{23}N_3O_3$	$C_{25}H_{28}N_3O_4$
SpookyNet model (~4 M, int 6, 128 features, cut-off = 5.29 Å)	 RMSD = 0.045 Å	 RMSD = 0.322 Å	 RMSD = 0.587 Å	 RMSD = 0.558 Å
TorchMD-Net model: Trained on only energies and forces (~3.2 M, cut-off = 5.29 Å)	 RMSD = 0.131 Å			

Other alternatives:

- NequIP (S. Batzner *et al.*, *Nat. Commun.* **13**, 2453, (2022)).
- Allegro (A. Musaelian *et al.*, (2022). arXiv:2204.05249).

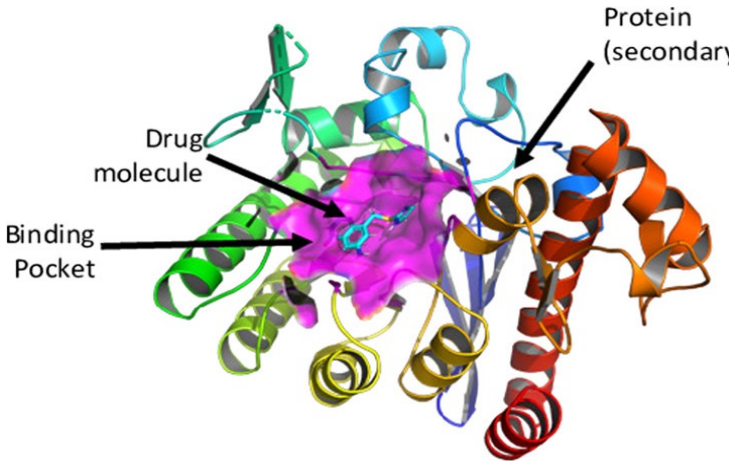
P. Thölke and G. De Fabritiis, (2022). arXiv:2202.02541
O. T. Unke *et al.*, *Nat. Commun.* **12**, 7273, (2021).

Method 4: Hybrid ML/Molecular-Mechanics potentials

Binding free energy calculations

$$U_{\mathrm{ML/MM}}\left(X_{\mathrm{P}}, X_{\mathrm{L}}\right) =$$

$$U_{\mathrm{MM}}\left(X_{\mathrm{P}}, X_{\mathrm{L}}\right) \text{ - } U_{\mathrm{MM}}^{vacuum}\left(X_{\mathrm{L}}\right) \text{ + } U_{\mathrm{ML}}^{vacuum}\left(X_{\mathrm{L}}\right)$$



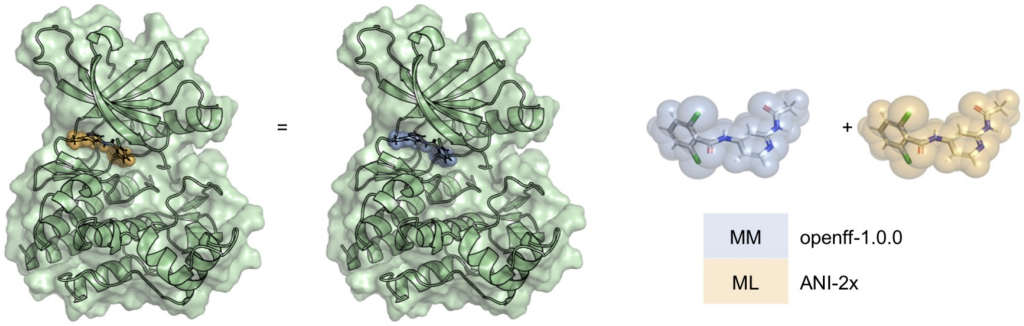
D.M. Rufa *et al.*, (2020). bioRxiv: 2020.07.29.227959.

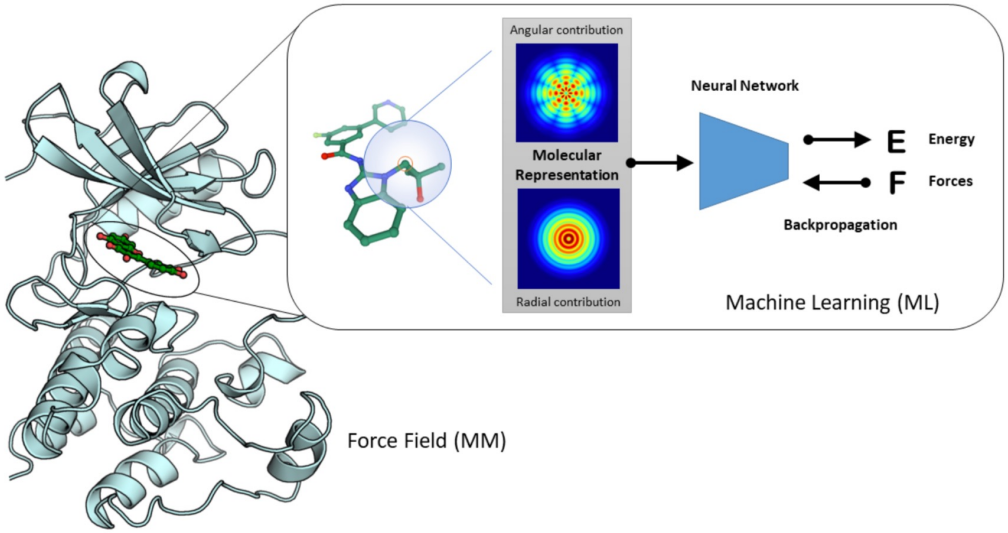
Method 4: Hybrid ML/Molecular-Mechanics potentials

Binding free energy calculations

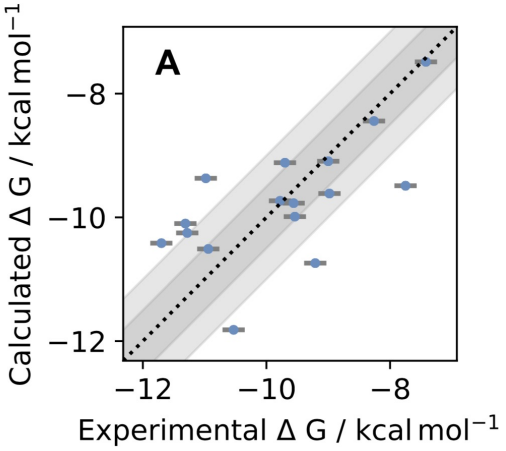
$$U_{\mathrm{ML/MM}}(X_{\mathrm{P}}, X_{\mathrm{L}}) =$$

$$U_{\mathrm{MM}}(X_{\mathrm{P}}, X_{\mathrm{L}}) \; - \; U_{\mathrm{MM}}^{vacuum}(X_{\mathrm{L}}) \; + \; U_{\mathrm{ML}}^{vacuum}(X_{\mathrm{L}})$$

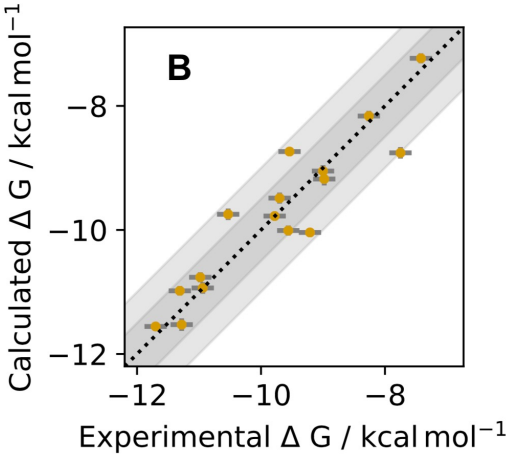




MM: openFF (+solvent)



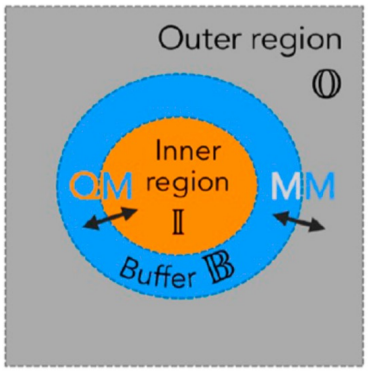
ML/MM: openFF (+solvent) **& ANI2x**



D.M. Rufa *et al.*, (2020). bioRxiv: 2020.07.29.227959.

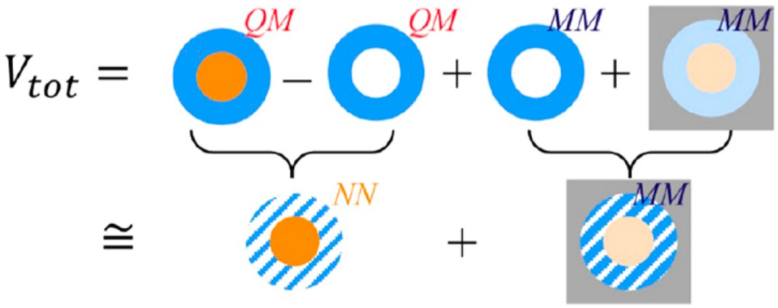
Method 4: Hybrid ML/Molecular-Mechanics potentials

BuRNN: Buffer region NN for polarizable embedding



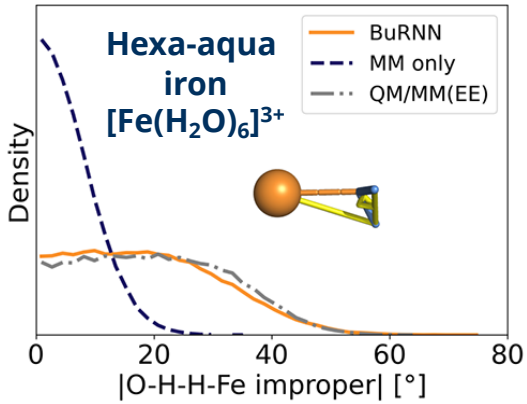
$$V_{tot} = V^{QM}_{\mathbb{I}+\mathbb{B}} - V^{QM}_{\mathbb{B}} + V^{MM}_{\mathbb{B}} + V^{MM}_{\mathbb{O}(\mathbb{I}+\mathbb{B}+\mathbb{O})}$$

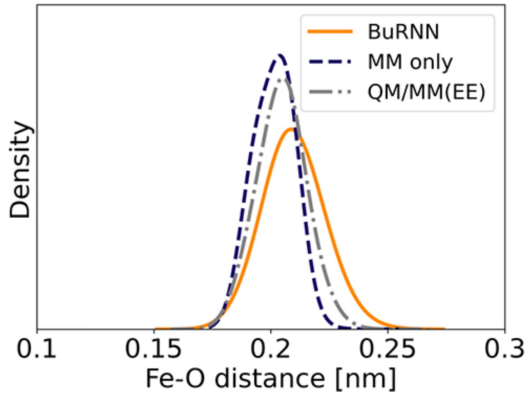
$$\cong \quad V^{NN}_{\mathbb{I}+\Delta\mathbb{B}} \quad + \quad V^{MM}_{\mathbb{B}+\mathbb{O}}$$



Features:

- Predict the difference between two QM regions.
- Polarizable embedding of buffer region at full QM level.





B. Lier *et al.*, *J. Phys. Chem. Lett.* **13**, 3812, (2022).

Method 4: Hybrid ML/Molecular-Mechanics potentials

System specific MD simulations with HDNNP

- ❑ Challenges
  - Long range interactions (electrostatic, vdW)
  - Large phase space
  - Long time scales
- ❑ Approach
  - Symmetry functions as in ANI-x models
  - Δ-learning scheme with DFTB as baseline

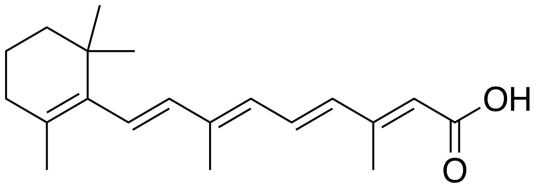
Loss functions:

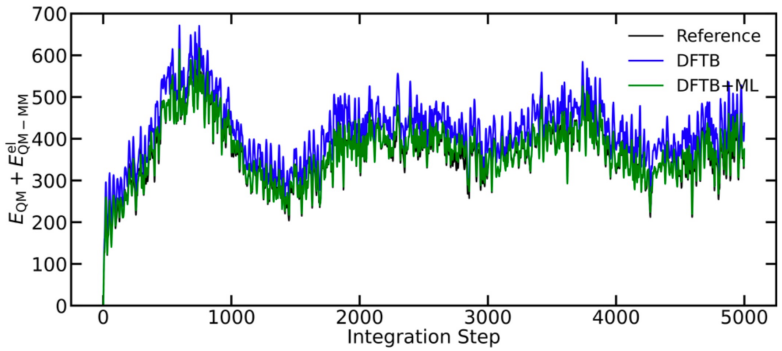
$$L = \frac{1}{N} \cdot \sum_{i}^{N} (E_i - \tilde{E}_i)^2$$

$$L' = L + \frac{\omega_0}{3N_{\mathrm{QM}}} \cdot \sum_{i}^{N_{\mathrm{QM}}} \sum_{\alpha}^{3} (-F_{i\alpha} + \tilde{F}_{i\alpha})^2 + \frac{\omega_1}{3N_{\mathrm{MM}}} \sum_{i}^{N_{\mathrm{MM}}} \sum_{\alpha}^{3} (-F_{i\alpha} + \tilde{F}_{i\alpha})^2$$

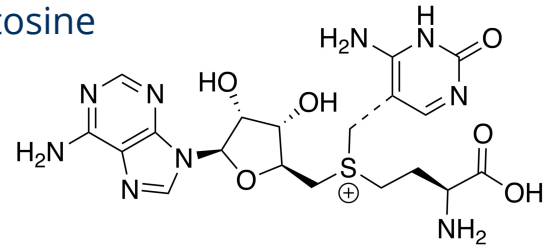
Molecules in water

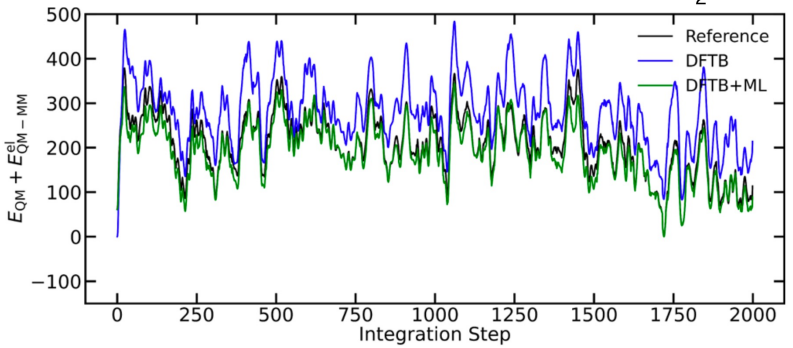
Retionic acid





Transition state of S-adenosylmethionate (SAM) with cytosine





L. Böselt *et al.*, *J. Chem. Theory Comput.* **17**, 2641, (2021).

Method 5: Hybrid ML/Quantum-Mechanics potentials

Density functional tight-binding (DFTB) method

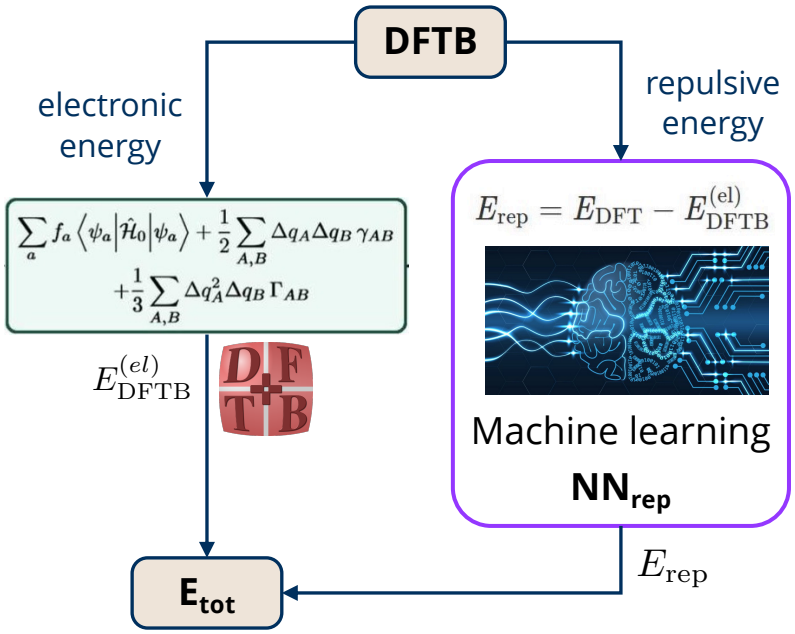
$$E_{\mathrm{DFTB}} = \sum_a f_a \langle \psi_a | \hat{\mathcal{H}}_0 | \psi_a \rangle + \frac{1}{2} \sum_{A,B} \Delta q_A \Delta q_B \gamma_{AB} + \frac{1}{3} \sum_{A,B} \Delta q_A^2 \Delta q_B \Gamma_{AB} + E_{\mathrm{rep}}$$

Electronic energy **Repulsive energy**

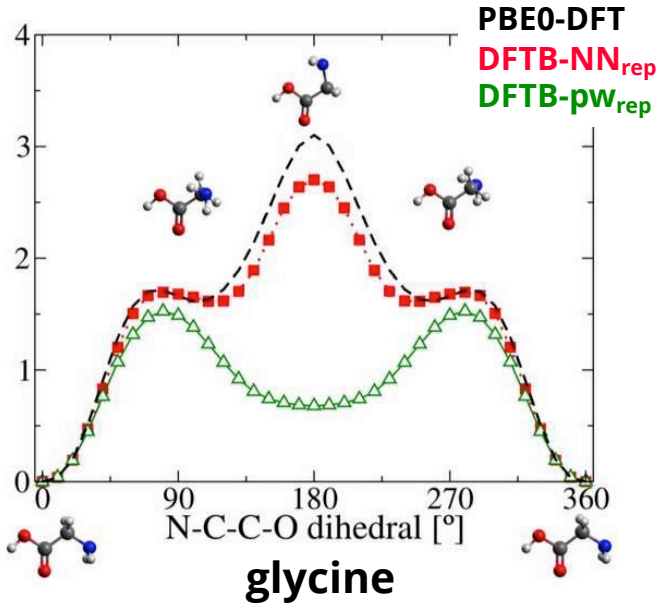
$$E_{\mathrm{rep}} = E_{\mathrm{DFT}} - E_{\mathrm{DFTB}}^{(\mathrm{el})} \approx \sum_{A<B} \mathcal{V}_{\mathrm{rep}}^{(AB)}(R_{AB})$$

- Problems of accuracy and general validity.

Many-body NN repulsive potentials



Rotational energy profile:

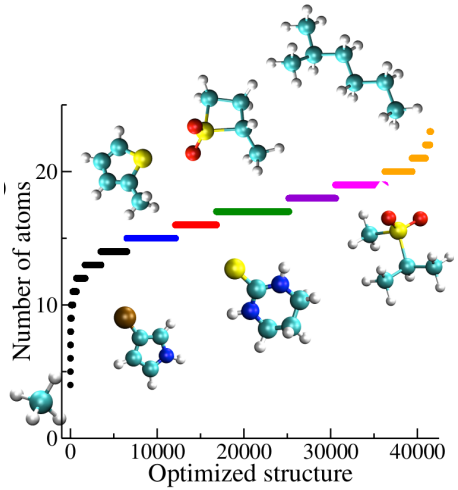


M. Stöhr *et al.*, *J. Phys. Chem. Lett.* **11**, 6835, (2020).

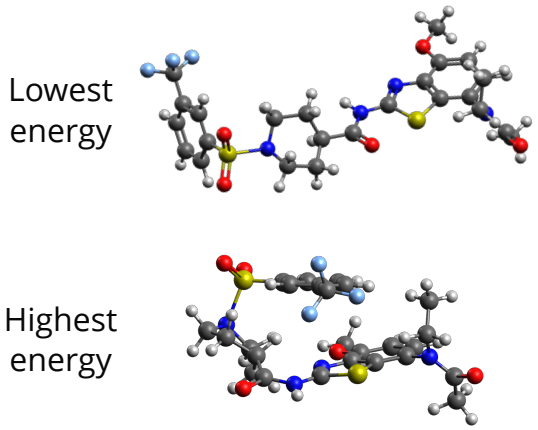
Summary

Increasing the system dimensionality

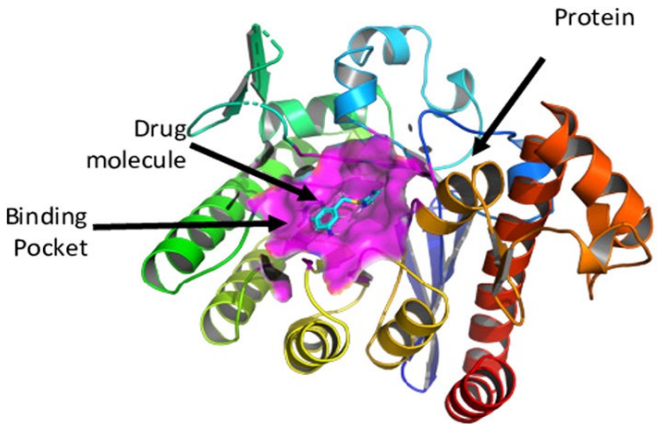
Small molecules



Large molecules



Drug-protein binding



Increasing the system complexity

- ✓ 2-and 3-body geometric descriptors
- ✓ Few degrees of freedom
- ✓ Weak non-covalent interaction

- ➢ High-dimensional geometric descriptors
- ➢ Large degrees of freedom
- ➢ Strong non-covalent interaction
- ➢ Strong intermolecular interaction
- ➢ Dependence on the chemical environment
- ➢ Hybrid QM/ML models

Acknowledgments

University of Luxembourg

Funding and computational resources

European Research Council

Fonds National de la Recherche Luxembourg

External collaborators

- Prof. Dr. Klaus-Robert Müller (TU Berlin)
- Prof. Dr. Robert DiStasio Jr. (Cornell univ.)
- Dr. Hugo Ceulemans (Janssen Phar.)
- Dr. Dries van Rompaey (Janssen Phar.)

Many thanks for your attention.

Email: leonardo.medrano@uni.lu

@Lmedranos88