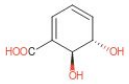
QuChemPedIA

Quantum Chemistry Collaborative EncyclopedIA

Thomas CAUCHY <thomas.cauchy@univ-angers.fr>
Benoit DA MOTA <benoit.damota@univ-angers.fr>

Intelligence artificielle pour la chimie des matériaux - 25 Sept. 2018

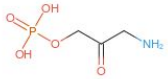
10 results for **1**.



- [x] OPT
- [ ] FREQ
- [ ] TD
- [ ] SP
- [ ] OPT_ES
- [ ] FREQ_ES

Formula : $C_7H_8O_4$
Charge : 0
Multiplicity: 1
Solvent : gas / **Solvatation**
method : Unknown

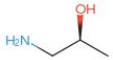
IUPAC : Unknown
GAMESS
B3LYPV1R / **Basis set :** 6-31G* (181 functions)
Ending energy : -572.4028395 a.u.



- [x] OPT
- [ ] FREQ
- [ ] TD
- [ ] SP
- [ ] OPT_ES
- [ ] FREQ_ES

Formula : $C_3H_8NO_5P$
Charge : 0
Multiplicity: 1
Solvent : gas / **Solvatation**
method : Unknown

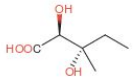
IUPAC : Unknown
GAMESS
B3LYPV1R / **Basis set :** 6-31G* (170 functions)
Ending energy : -891.4226034 a.u.



- [x] OPT
- [ ] FREQ
- [ ] TD
- [ ] SP
- [ ] OPT_ES
- [ ] FREQ_ES

Formula : C_3H_9NO
Charge : 0
Multiplicity: 1
Solvent : gas / **Solvatation**
method : Unknown

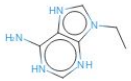
IUPAC : Unknown
GAMESS
B3LYPV1R / **Basis set :** 6-31G* (93 functions)
Ending energy : -249.6945983 a.u.



- [x] OPT
- [ ] FREQ
- [ ] TD
- [ ] SP
- [ ] OPT_ES
- [ ] FREQ_ES

Formula : $C_6H_{12}O_4$
Charge : 0
Multiplicity: 1
Solvent : gas / **Solvatation**
method : Unknown

IUPAC : Unknown
GAMESS
B3LYPV1R / **Basis set :** 6-31G* (174 functions)
Ending energy : -536.7636026 a.u.

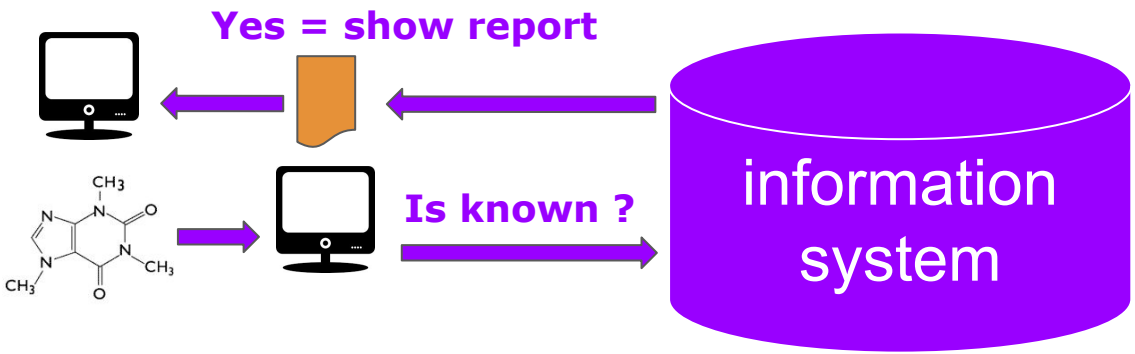


- [x] OPT
- [ ] FREQ
- [ ] TD
- [ ] SP
- [ ] OPT_ES
- [ ] FREQ_ES

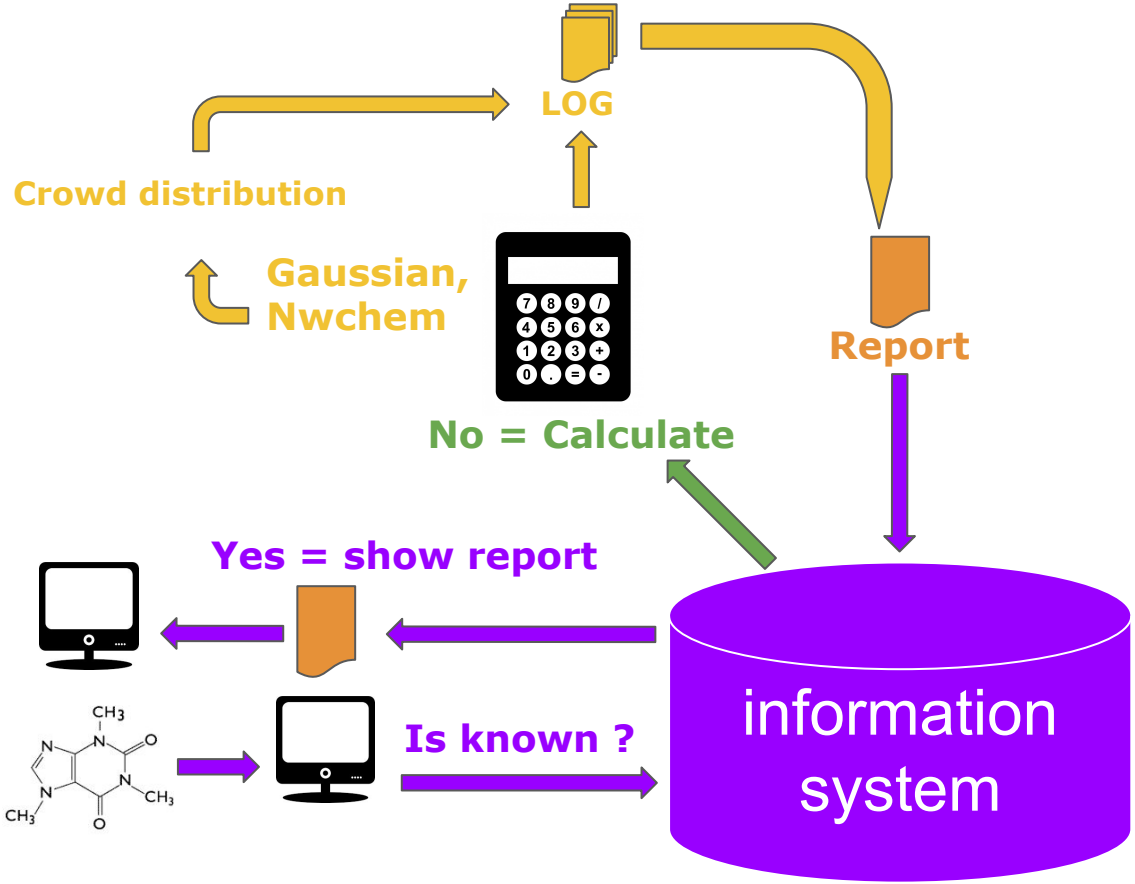
Formula : $C_7H_9N_5$
Charge : 0
Multiplicity: 1
Solvent : gas / **Solvatation**
method : Unknown

IUPAC : Unknown
GAMESS
B3LYPV1R / **Basis set :** 6-31G* (198 functions)
Ending energy : -545.9492504 a.u.

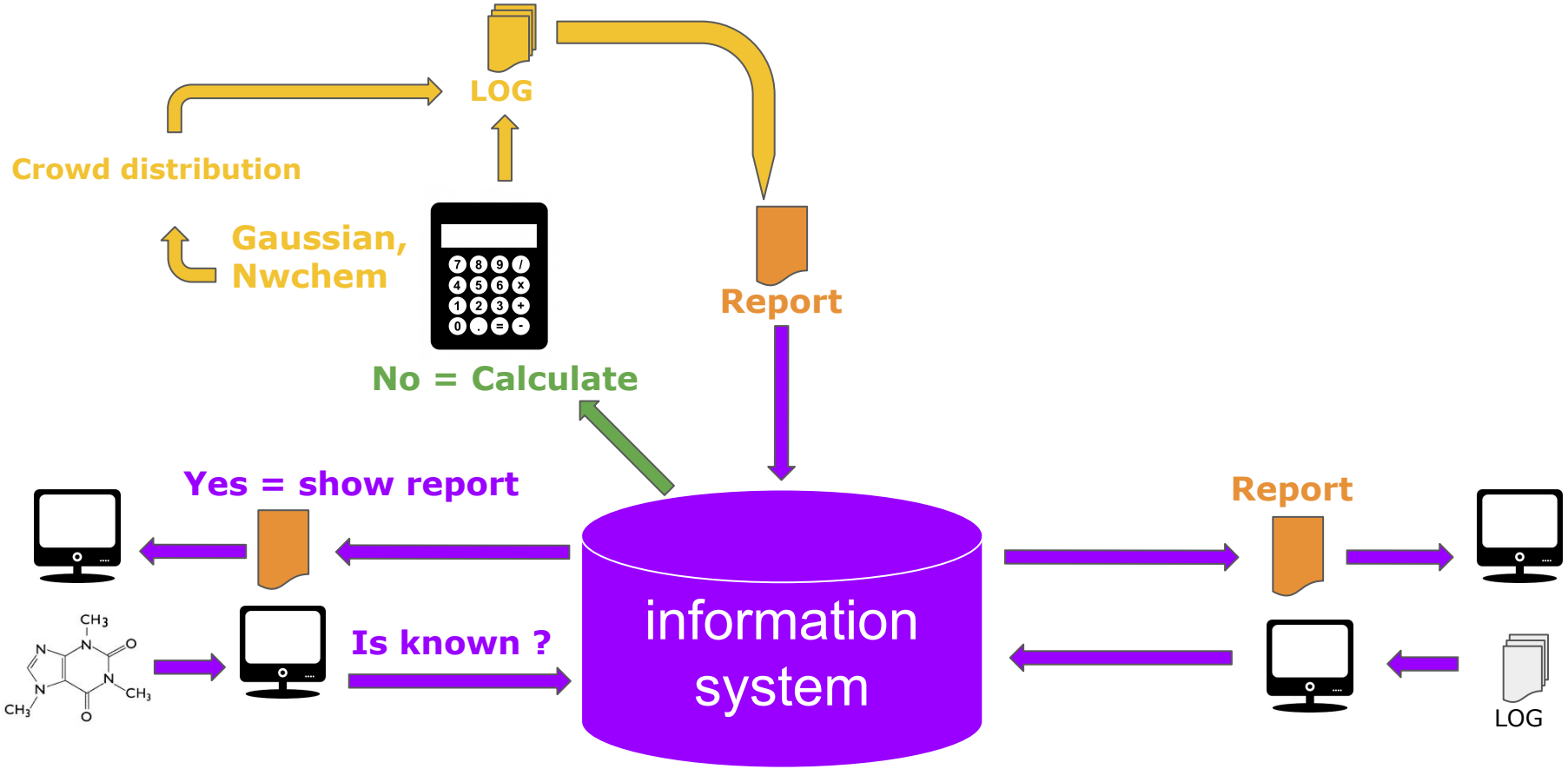
The QuChemPedIA project



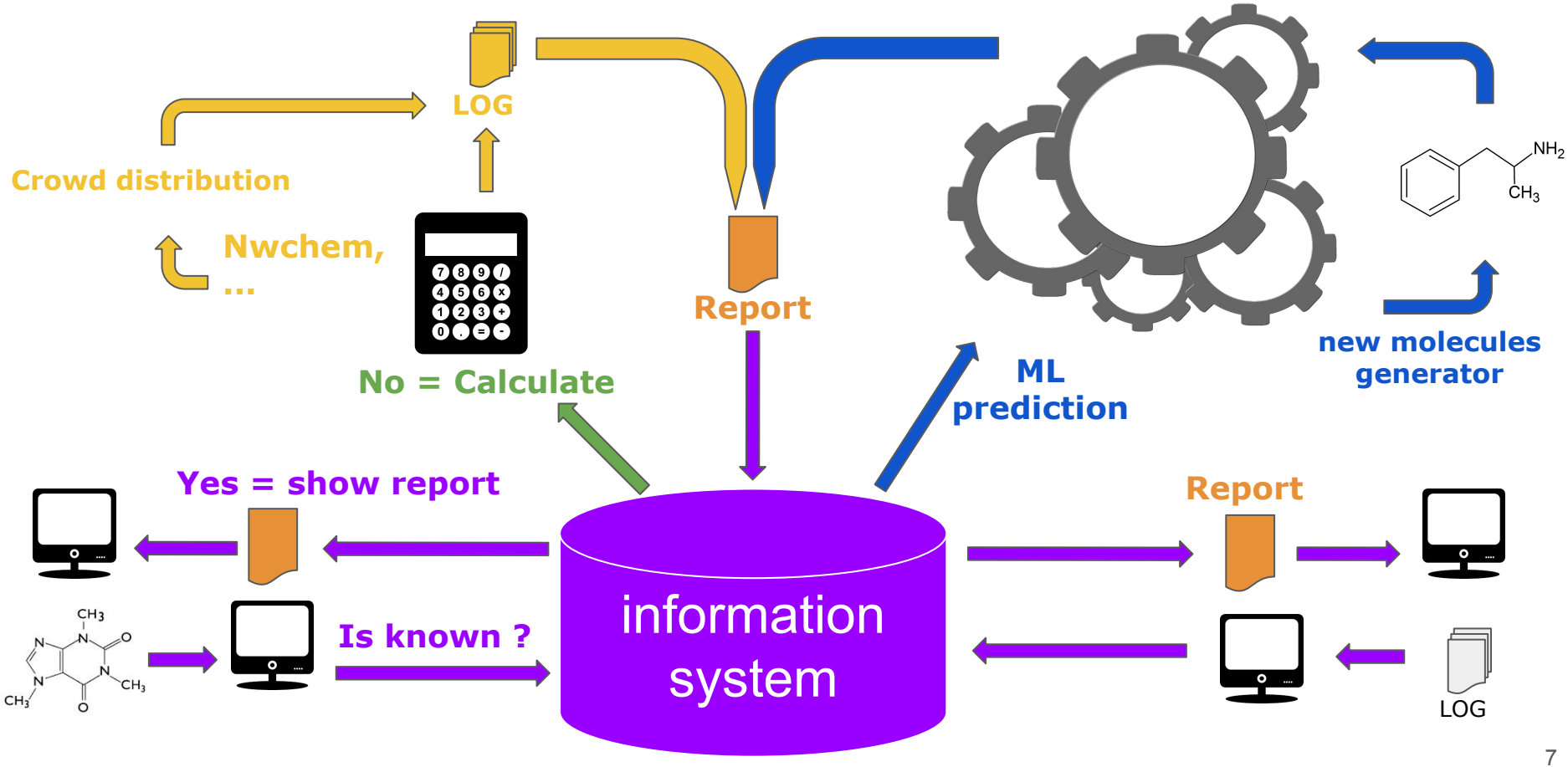
The QuChemPedIA project



The QuChemPedIA project

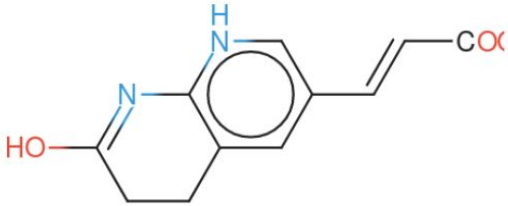


The QuChemPedIA project



Molecule

InChI	1S/C11H10N2O3 /c14-9-3-2-8-5-7(1-4-10(15)16)6-12-11(8)13-9 /h1,4-6H,2-3H2,(H,15,16)(H,12,13,14)/b4-1+
Canonical SMILES	OC(=O)/C=C /c1cnc2c(c1)CCC(=N2)O
Monoisotopic mass	218.06914219
Formula	$C_{11}H_{10}N_2O_3$
Charge	0
Spin multiplicity	1



Associated calculations

Job type	Author	Description
OPT	Brice Harismendy	N/A

Authorship

Original log file	Download
Primary author	Brice Harismendy
Affiliation	None

Computational details

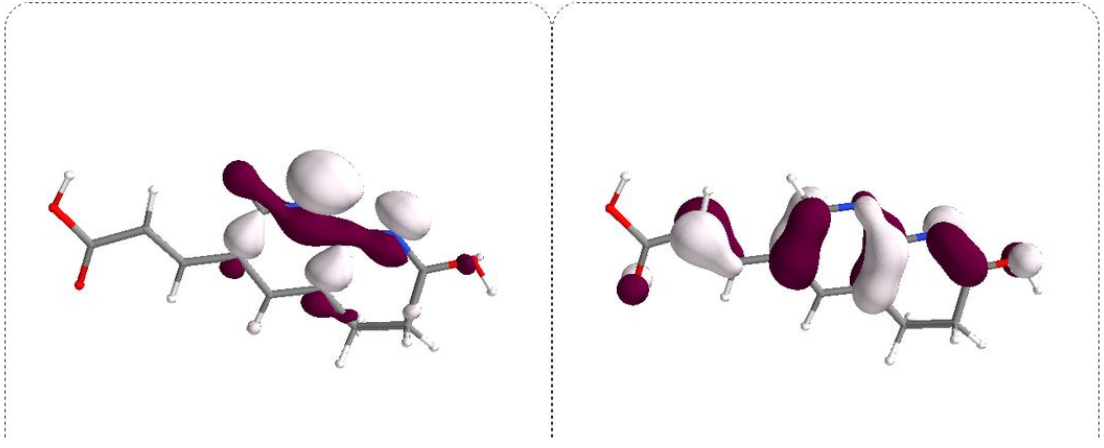
Software	GAMESS (1MAY2013)
Computational method	DFT
Functional	B3LYPV1R
Basis set name	6-31G*

Most intense Mulliken atomic charges
mean = 0.000 e, std = 0.342

Atom	number	Mulliken partial charges
O	15	-0.585
O	13	-0.561
O	14	-0.463
N	12	-0.458
N	11	-0.457
C	2	-0.374
C	1	-0.345
C	10	0.364
H	24	0.414
C	8	0.502
C	9	0.610
H	25	0.407

Geometry

Nuclear repulsion energy in atomic units 988.60699 a.u.



Molecular databases

PubChem

PubChemQC Project: A Large-Scale First-Principles Electronic Structure Database for Data-Driven Chemistry

Maho Nakata*† and Tomomi Shimazaki‡
† Advanced Center for Computing and Communication, RIKEN, 2-1 Hirosawa, Wako, Saitama 351-0198 Japan
‡ Advanced Institute for Computational Science, RIKEN, 7-1-26 Minatojima-minami-machi, Chuo-ku, Kobe, Hyogo 650-0047 Japan

J. Chem. Inf. Model., **2017**, *57* (6), pp 1300–1308
DOI: 10.1021/acs.jcim.7b00083
Publication Date (Web): May 8, 2017

*E-mail: maho@riken.jp.

ioChem-BD Browse - Barcelona Supercomputing Center / BSC host

PubChemDFT Collection home page

This is a **live** project. It uses spare computer time to compute, process, store, and publish open-access DFT results of molecules contained in the PubChem database. For each entry, we provide its optimized geometry, energies, charges, vibrational frequencies, cube files for electron density and electrostatic potential, etc ... Average rate is close to 1000 molecs/day. Started in April 2017, in July 2017 70000 molecules have been completed.

Our Twitter bot @MolecuBot, born July 2017, tweets each time a new molecule is completed and published.

Discover

Author
Central, ioChem-BD 182959

Program name
Gaussian 182959

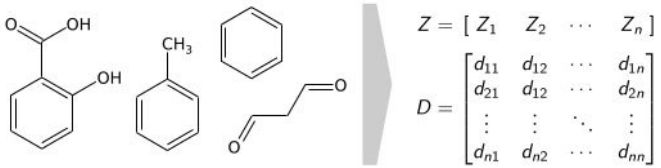
Calculation type
Geometry optimization Minimum 182959

Machine learning

First step :
Predicting interatomic distances

Related works

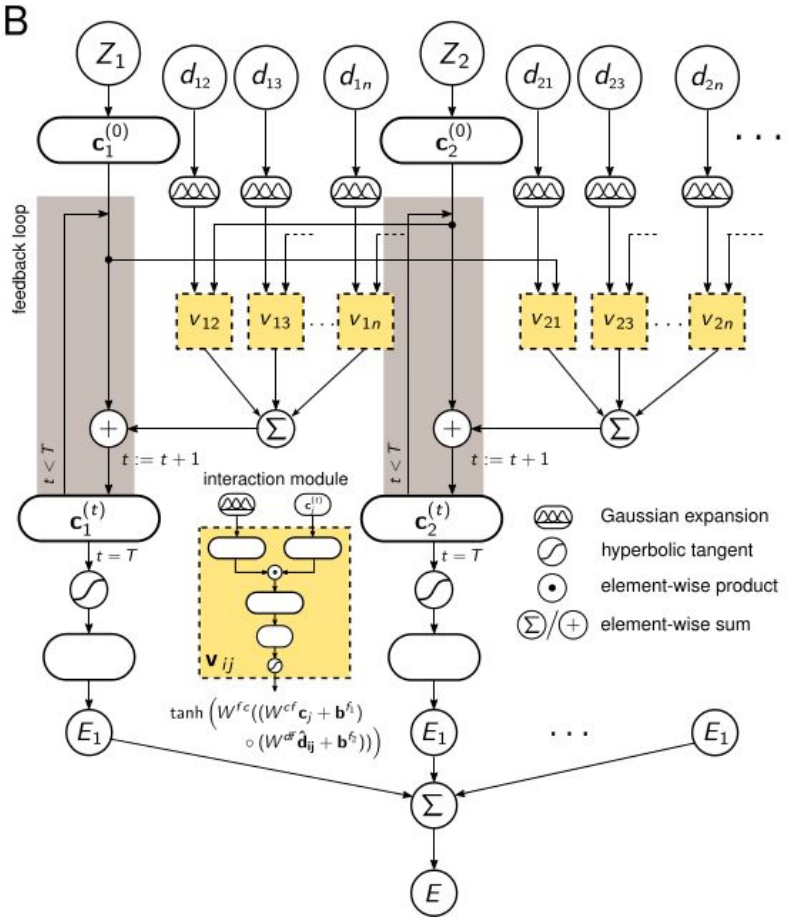
[Schütt et al.] Quantum-Chemical Insights from Deep Tensor Neural Networks. *Nature Communications 2017*.



$$Z = [\, Z_1 \quad Z_2 \quad \cdots \quad Z_n \,]$$

$$D = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{12} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix}$$

- GDB-9 dataset :
  - combinatorial molecular space with 9 heavy atoms: C, N, O and F.
  - ~134k small theoretical molecules
- full distances matrix (D): scaling issue
- “only” energy prediction



Preliminary results

- PubChemQC dataset
  - > 3 millions of real molecules
  - general sampling of the real molecular space (organic chemistry)
- homogeneous data (DFT, B3LYP, 6-31G*)
- simple neural networks (from 3 up to 9 fully connected layers)
- designed for strong scaling

Model	Objective	Given data
1	All interatomic distances	(partial) distance matrix
2	1 distance (CC, CH, OH)	all distances with other atoms
3	1 distance (CC, CH, OH)	distances with neighbors

Partial distances matrix with trilateration

Distances better than coordinates: invariant by rotation and translation

But full distances matrix for n atoms $\rightarrow n^2$ distances bad scaling

How to reconstruct converged geometry? 3D trilateration

Solution : distances from fixed points $\rightarrow 4n$ distances good scaling

d_{a_0,p_0}	d_{a_0,p_1}	d_{a_0,p_2}	d_{a_0,p_3}
d_{a_1,p_0}	d_{a_1,p_1}	d_{a_1,p_2}	d_{a_1,p_3}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
d_{a_n,p_0}	d_{a_n,p_1}	d_{a_n,p_2}	d_{a_n,p_3}

+ Atomic masses $[m_{a0}, \ldots, m_{an}]$

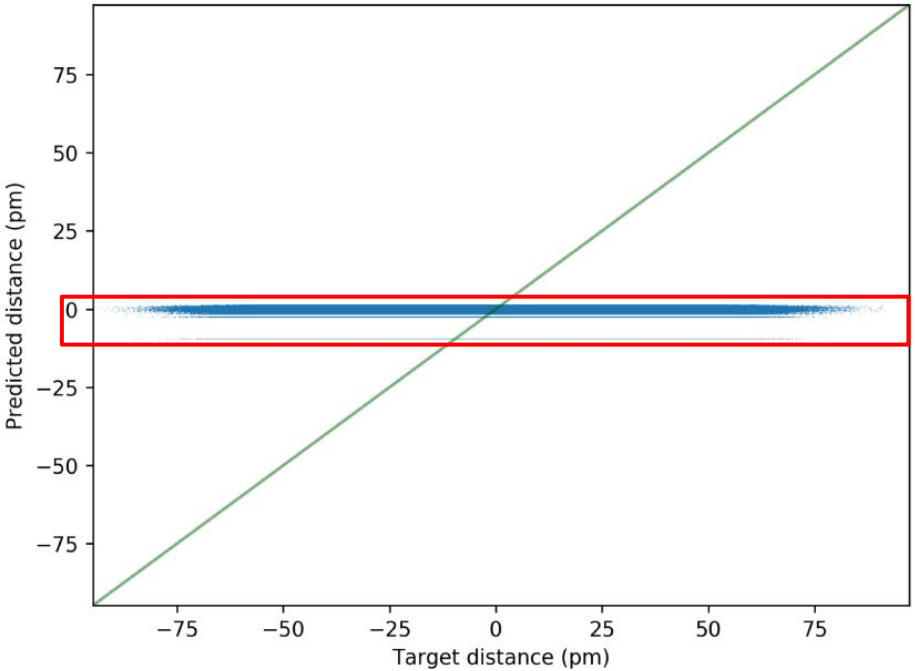
Many tested models !

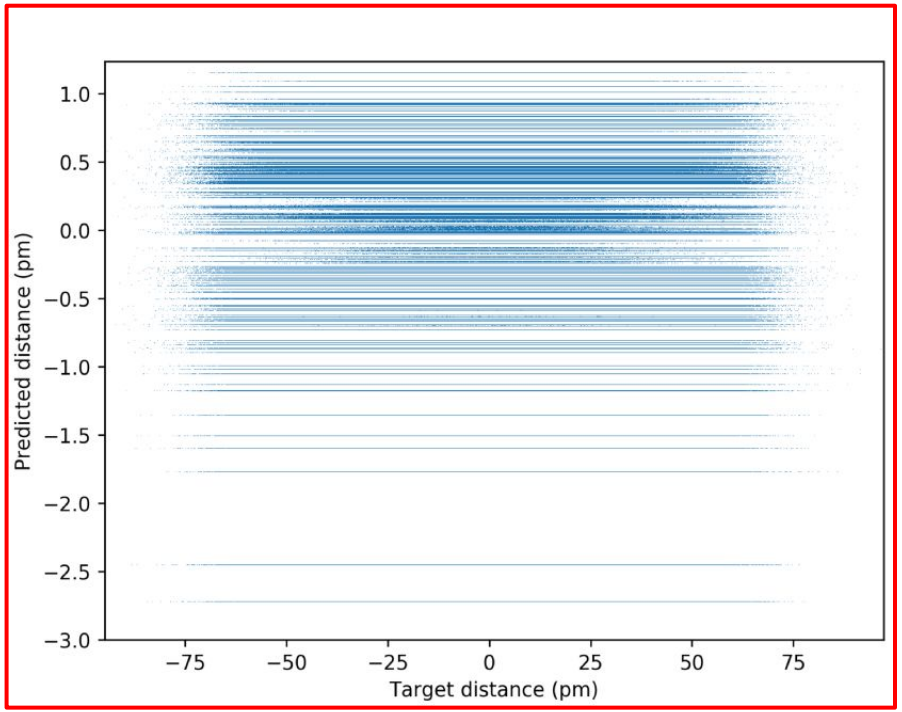
Neural networks with fully connected layers:

- Loss function: RMSE
- Learning rate: {0.1, 0.0001, 0.00001}
- Adam Optimizer ε: {1000, 0.0001}
- Weights initialization : {0.2, 0.002}
- Hidden layers activation function: {elu, crelu}
- Exit layer activation function: linear
- Weight decay: {0.1, 0.01, 0.001}
- Layers width: 500 (up to 100 atoms per molecule)
- Networks depth: {3, 7}
- Batch size : {500, 200}
- Number of epochs : 3

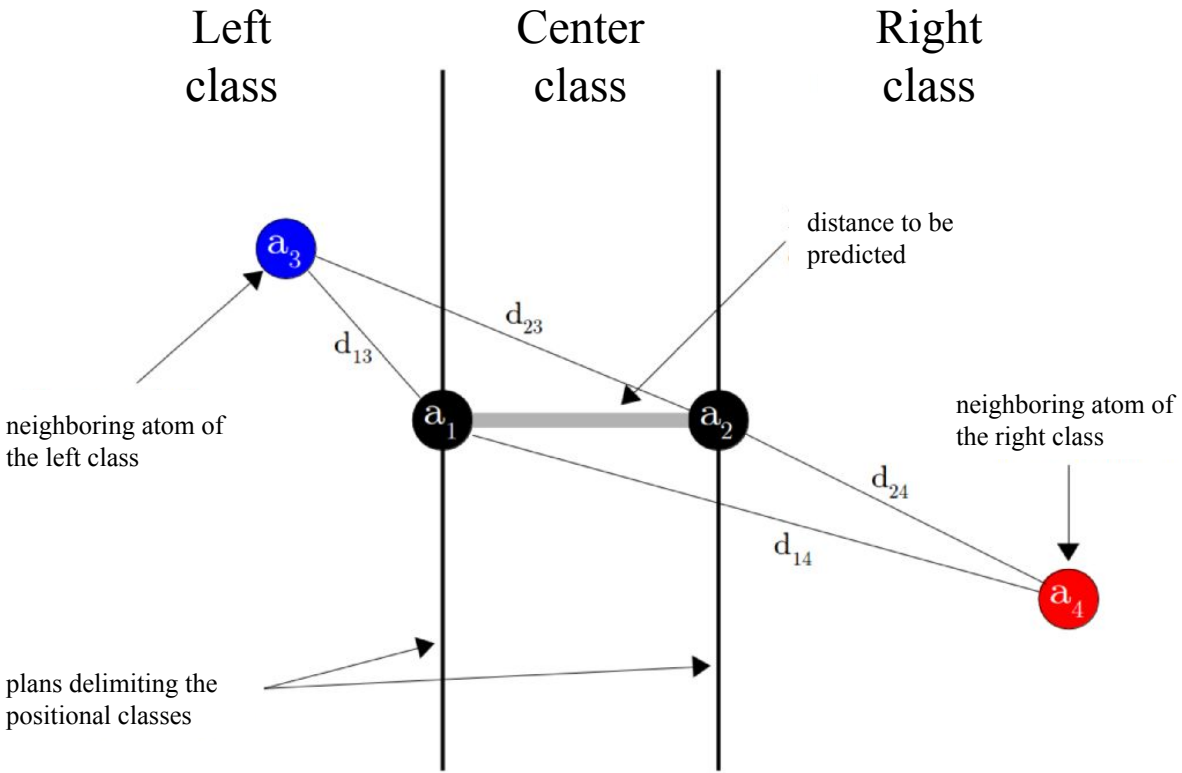
→Networks trained to predicted Δd from starting distances to converged distances

Results

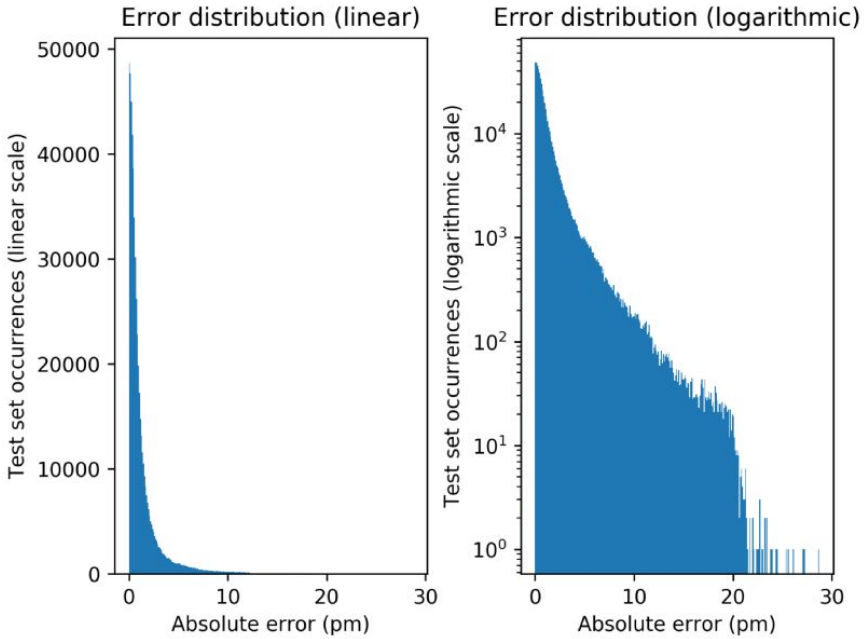

Predicted distance (pm)
75
50
25
0
−25
−50
−75
−75
−50
−25
0
25
50
75
Target distance (pm)

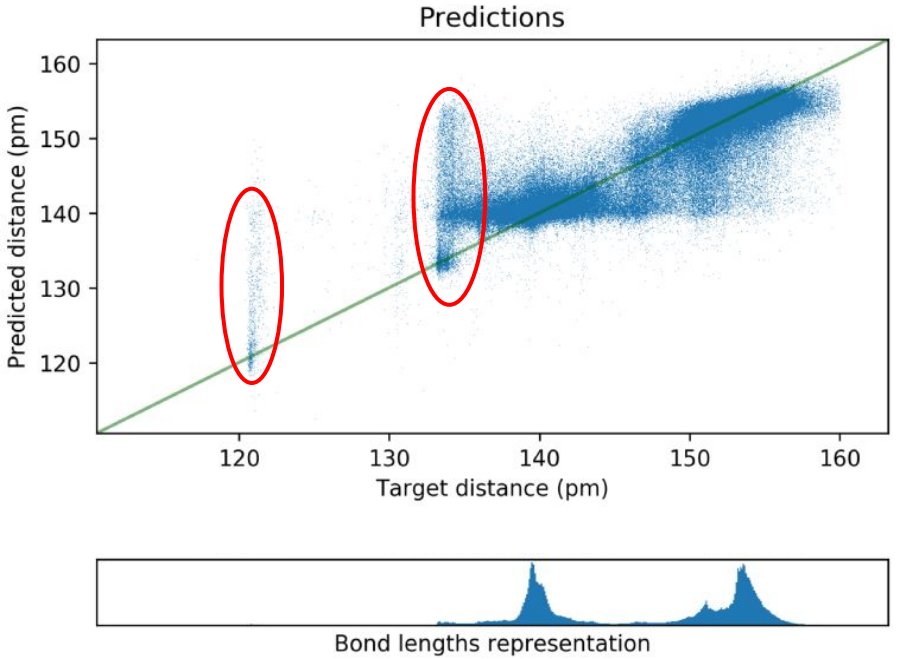

Predicted distance (pm)
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
−75
−50
−25
0
25
50
75
Target distance (pm)

One interatomic distance (CC, OH and CH)



Results for Carbon-Carbon





Example of folded molecule (Pubchem CID 328310)

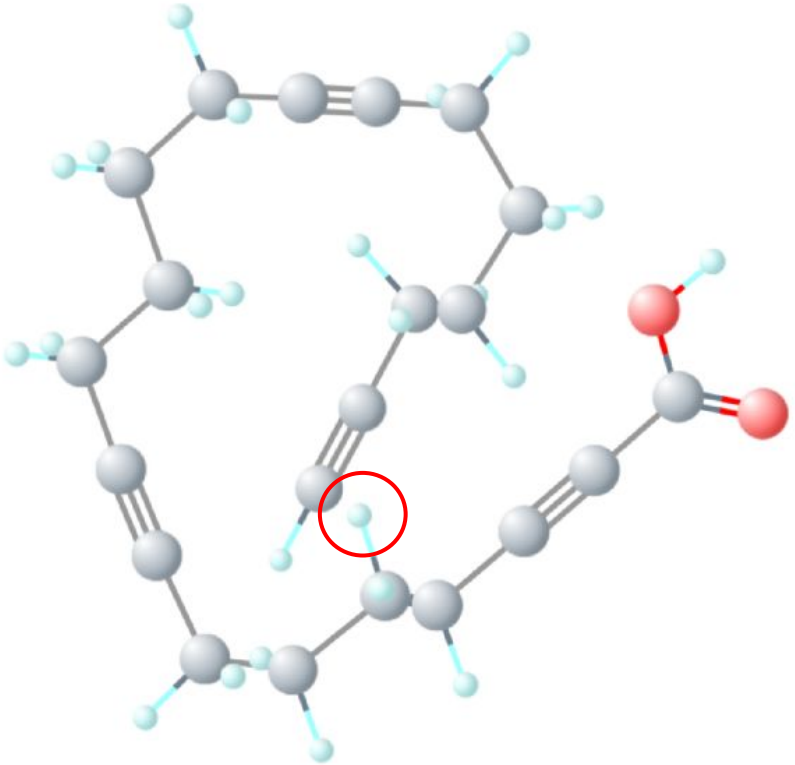
Unusually close atoms = “false” bonding.

Bad surrounding predicted distances.

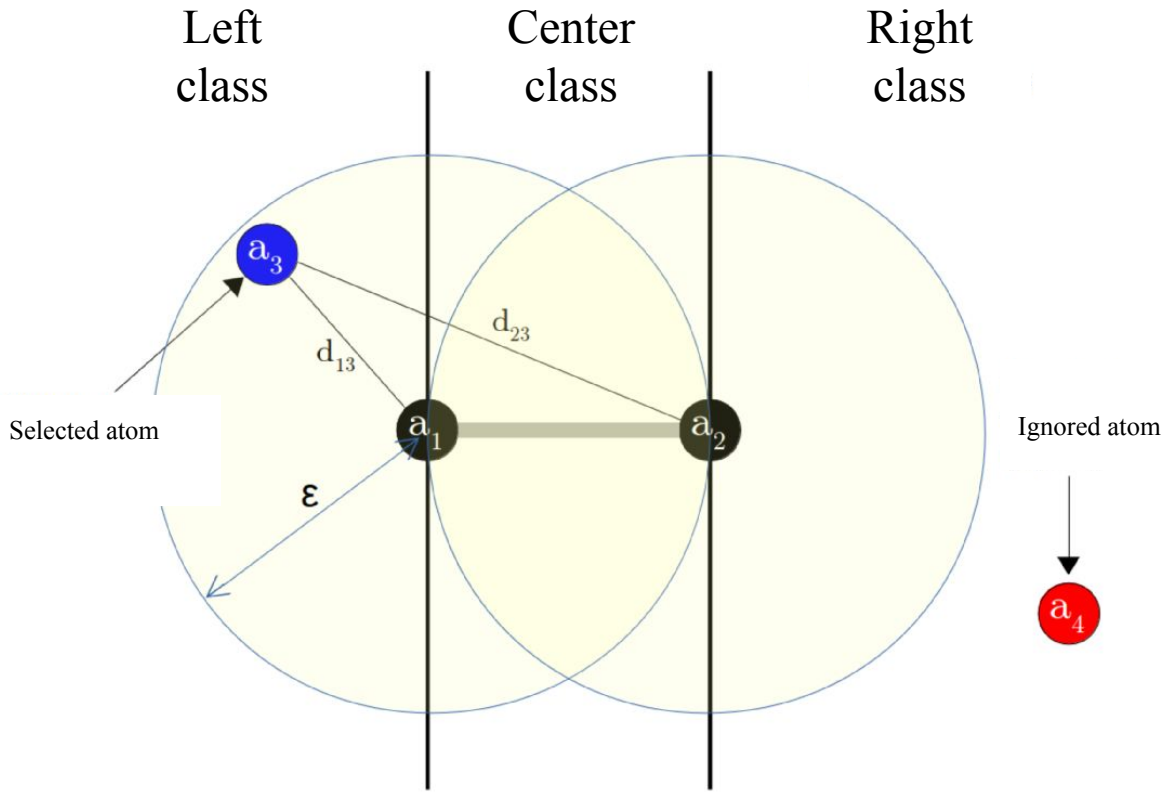
Neural network needs

feature engineering
domain specific knowledge

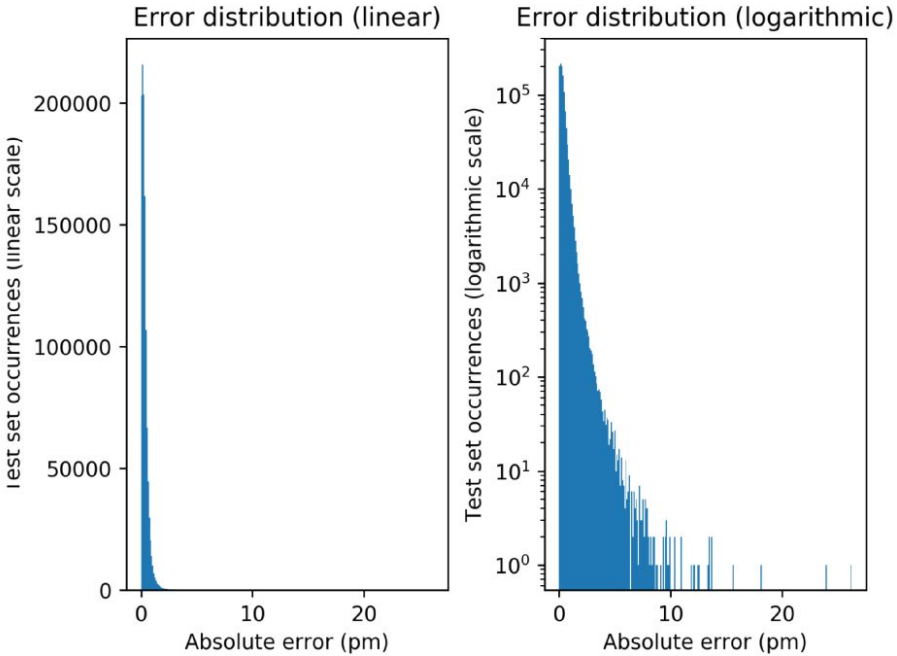
Limitation to the covalent neighboring region! (200 pm)

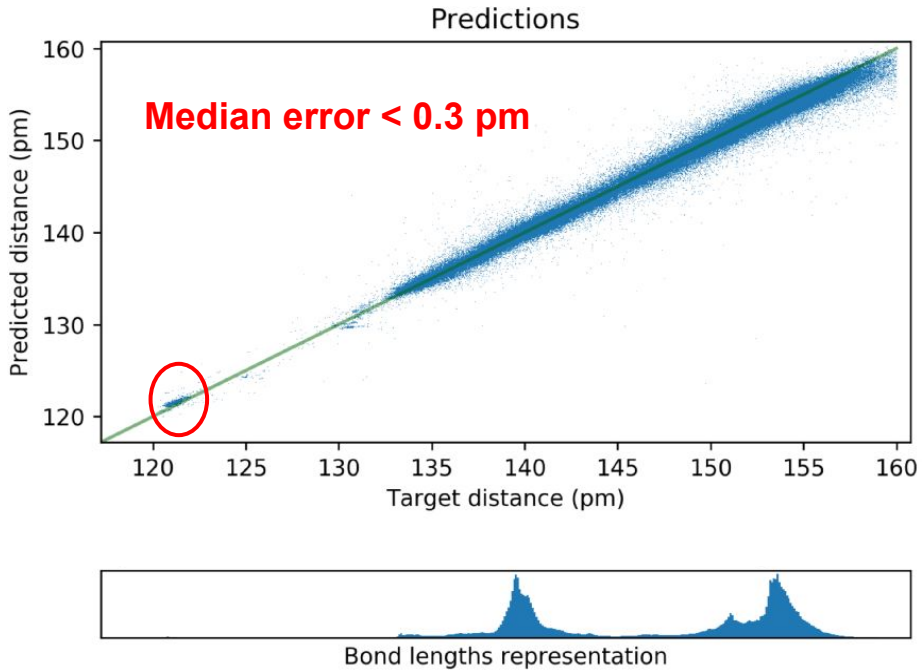


One interatomic distance with neighborhood



Results for Carbon-Carbon





20 problematic CC bonds on non curated data!

Conclusion and perspectives

- Median error for CC < 0.3 pm; for CH < 0.2 pm and for OH < 0.1 pm
  - > 3 millions of real molecules (previously non curated)
  - general sampling of the real molecular space (organic chemistry)
- Post-doc position available (now) on:
  - iterative geometry optimization coupling different models (NN, KRR,...)
  - prediction of the wavefunction or UV-visible absorption/emission
  - generative models (GAN, autoencoders) or combinatorial optimization algorithms (objective and neighboring function) to efficiently explore the molecular space