



Atlantisc
2 · 0 · 2 · 0

Recherche,
Formation
& Innovation
en PAYS de la LOIRE

AI-based assistant for molecular QUantum chemistry

Challenge RFI Atlantisc - Nantes 20/04/2018

Le consortium



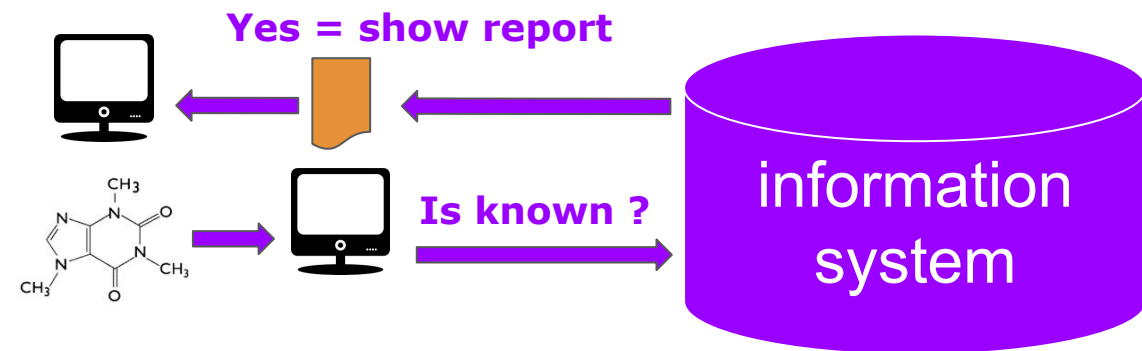
- Benoit DA MOTA
- Gilles HUNAULT
- Béatrice DUVAL
- David LESAINT

Optimisation combinatoire
Big Data
Intelligence artificielle
Calcul haute performance
Apprentissage artificiel
Science des données
Programmation par contraintes

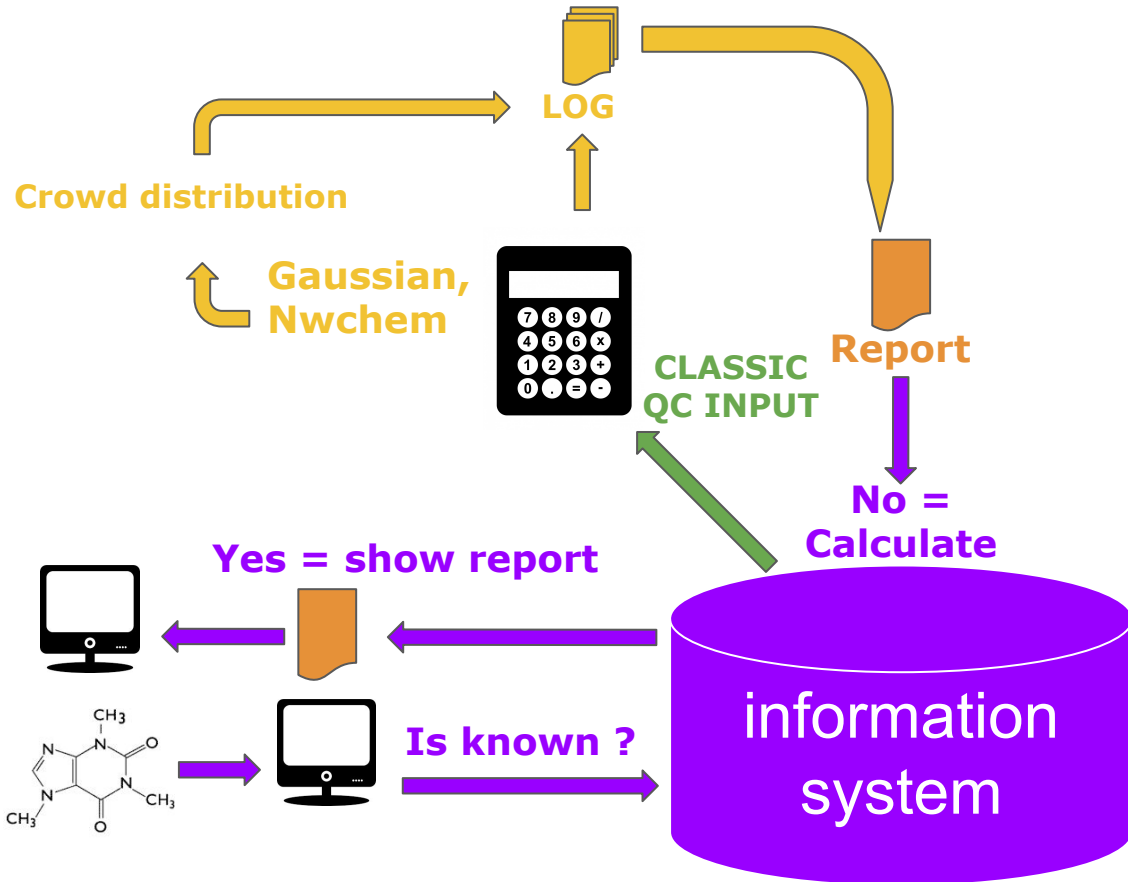


- Thomas CAUCHY = Chimie théorique
- Yohann MORILLE = Calcul scientifique
- Collègues expérimentateurs

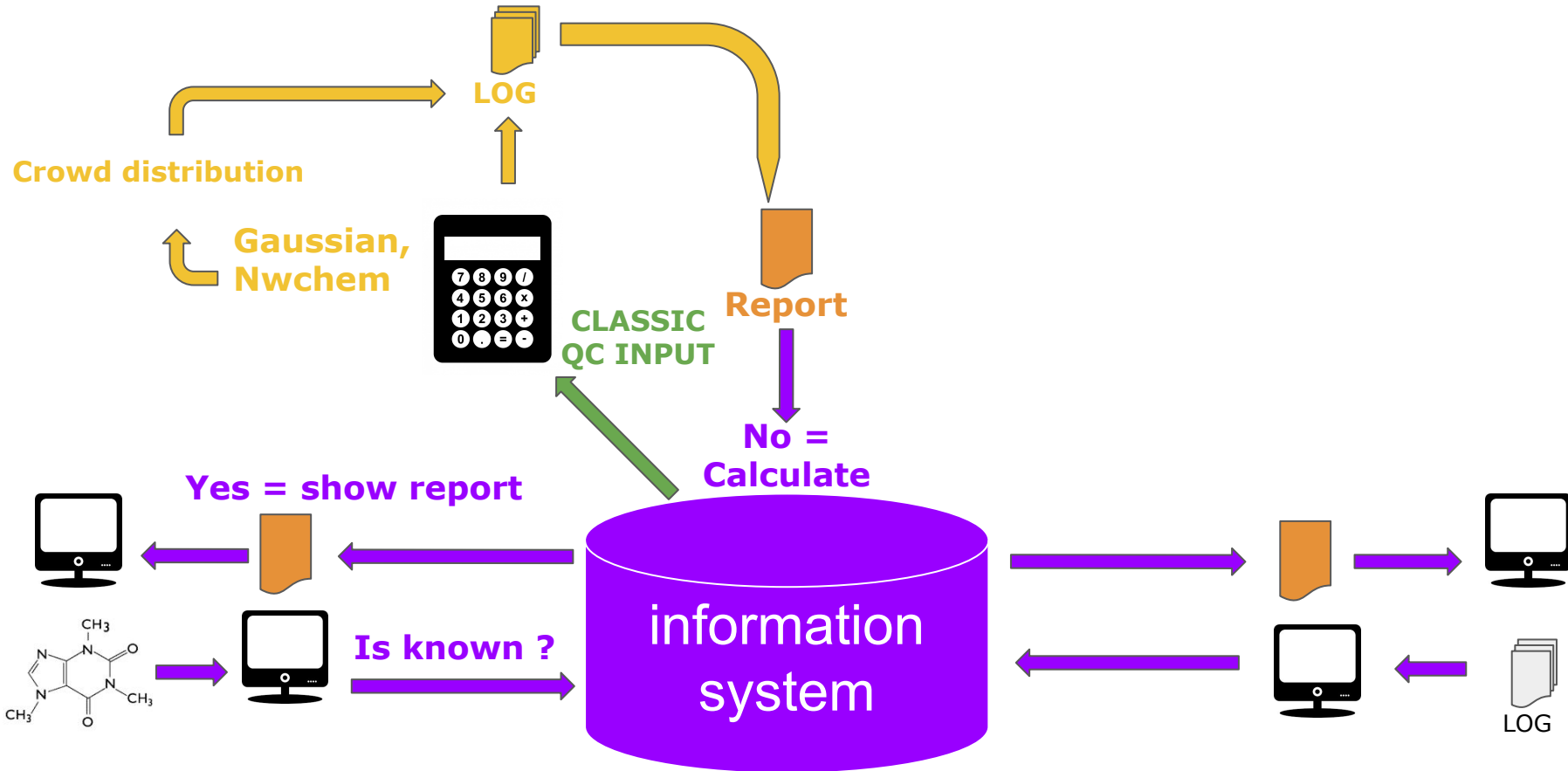
The QuChemPedIA project



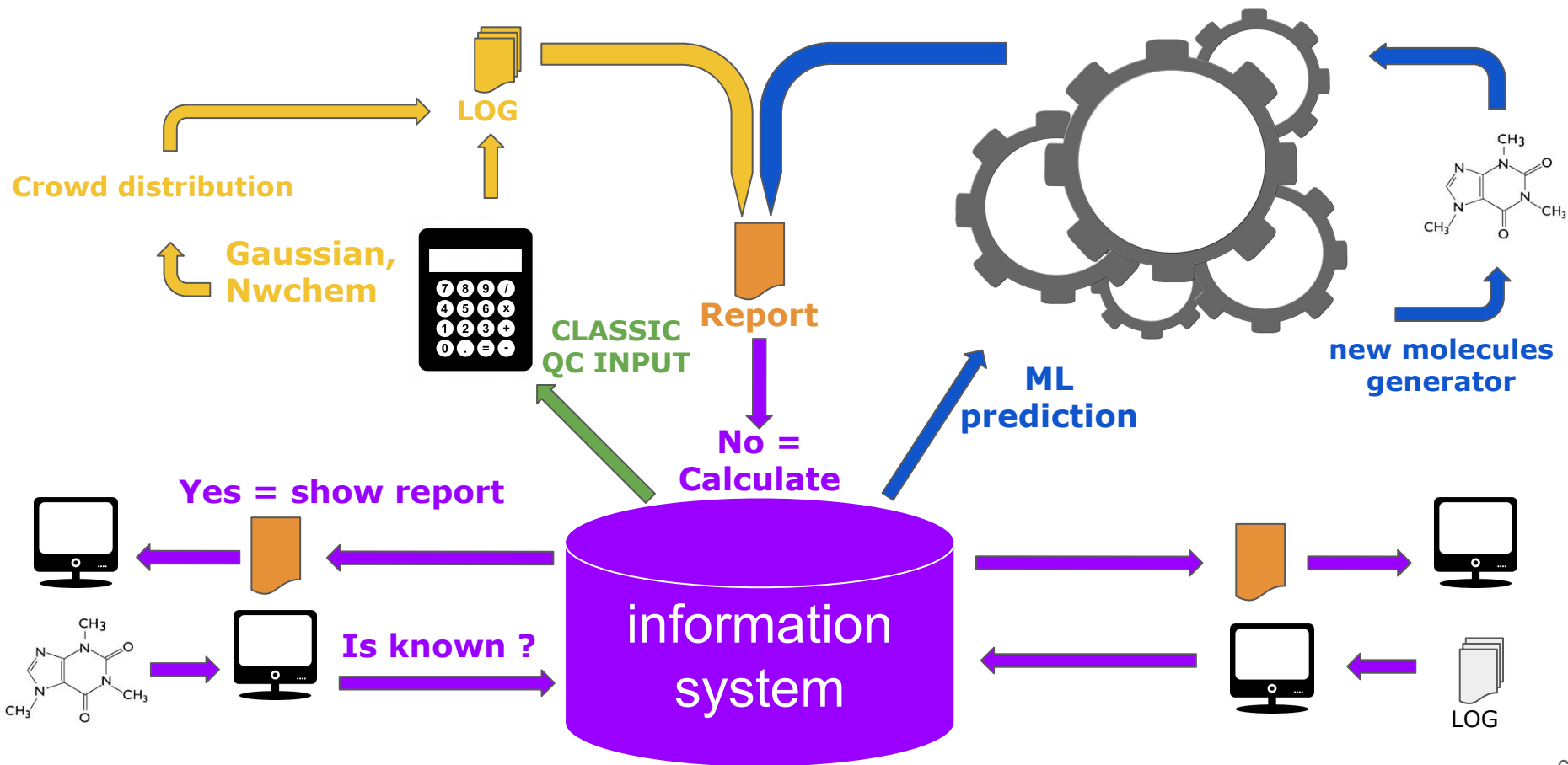
The QuChemPedIA project



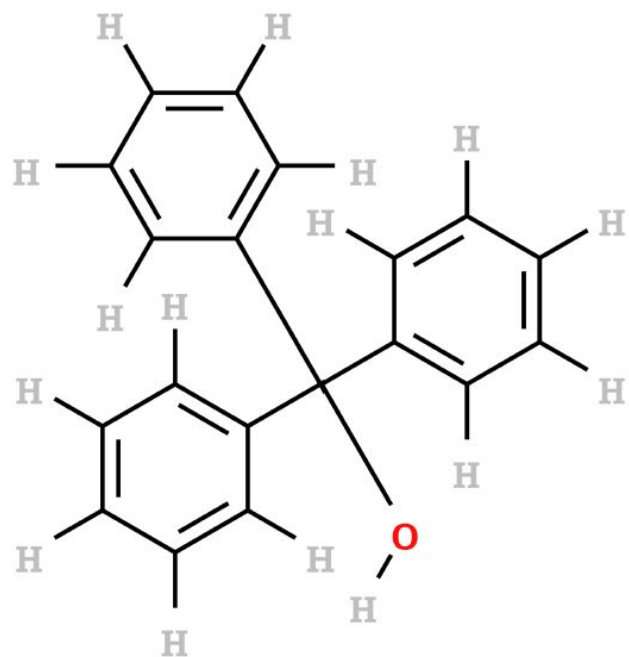
The QuChemPedIA project



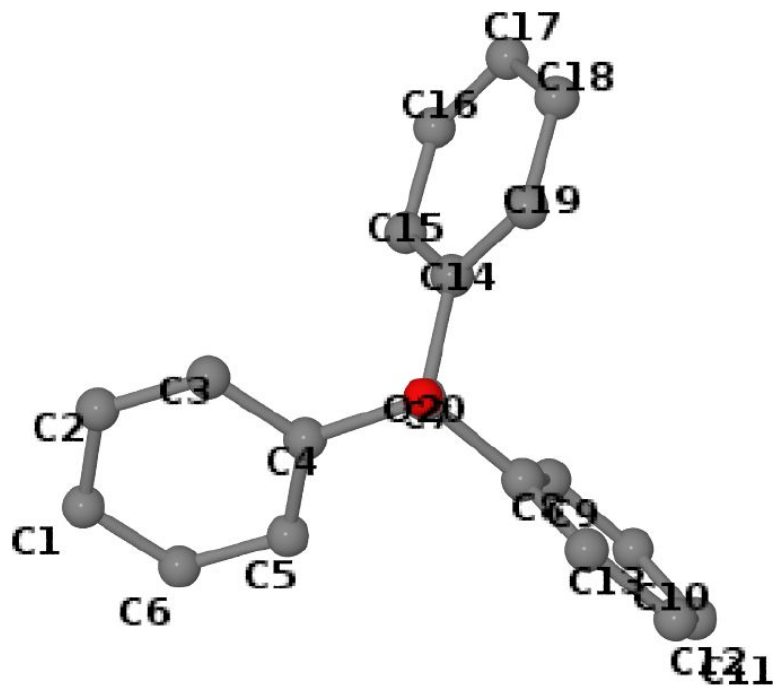
The QuChemPedIA project



L'objet : molécule

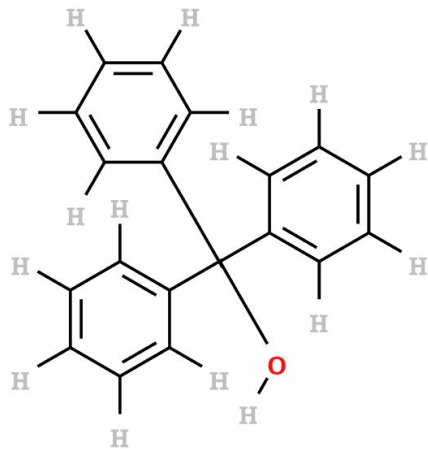


Chemical structure diagram



Géométrie 3D

L'objet : molécule

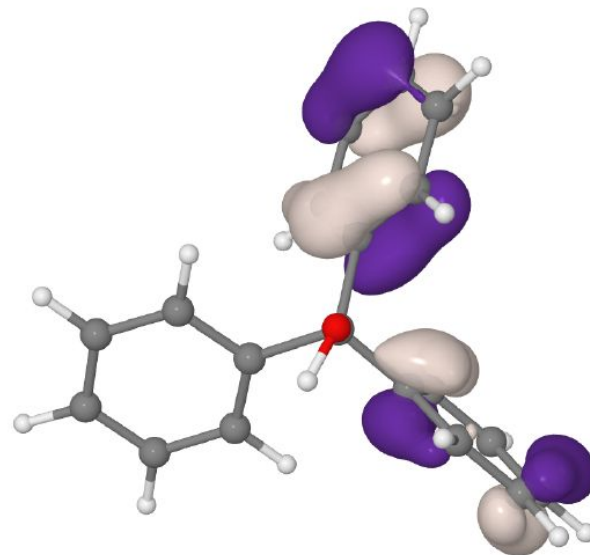


Chemical structure diagram

Most intense Mulliken atomic charges

mean = -0.000 e, std = 0.186

Atom	number	Mulliken charges
O	20	-0.647
C	5	-0.192
H	36	0.394



Représentation d'une isosurface
décrivant 1 électron
(cubes de voxels)

Les bases de données disponibles



NCBI Resources ▾ How To ▾

[Sign in to NCBI](#)

PubChem
Compound

PubChem Compound ▾

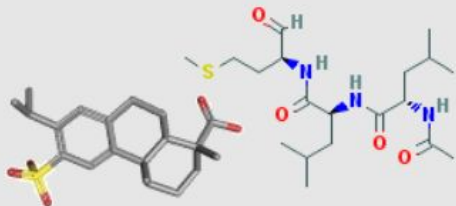
[Limits](#) [Advanced](#)

Search

[Help](#)



Your search request did not contain a term.



PubChem Compound

The PubChem Compound Database contains validated chemical depiction information provided to describe substances in PubChem Substance. Structures stored within PubChem Compounds are pre-clustered and cross-referenced by identity and similarity groups.

Base de données expérimentale avec > 100 millions de composés.

Les bases de données disponibles

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

Copyright © 2017 American Chemical Society

*E-mail: maho@riken.jp.

 **Cite this:** *J. Chem. Inf. Model.* 57, 6, 1300-1308

 RIS Citation 

Base de données théorique avec 3.5 millions de composés.
Un seul niveau de théorie. Pas d'apprentissage automatique !

Les bases de données disponibles

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 moles/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

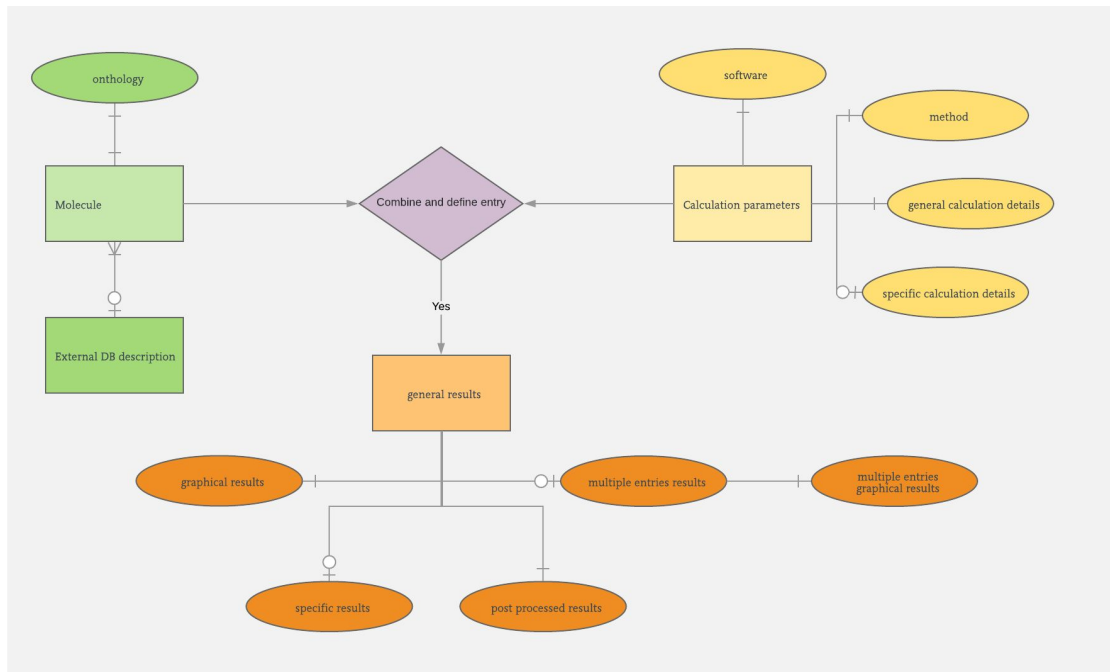
182959

Base de données théorique avec 200 k composés.

Un seul niveau de théorie différent de PubChemQC. Pas d'apprentissage automatique !

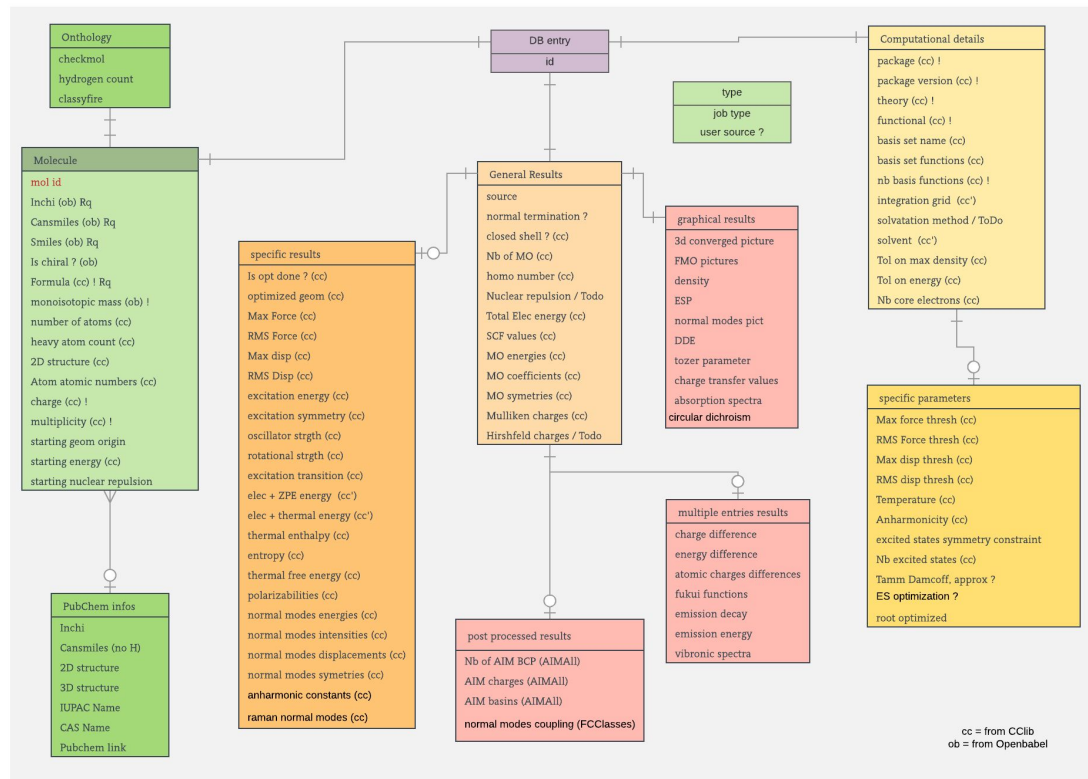
L'objet dans notre système d'information

- objectif 100 millions de molécules
- Plusieurs paramètres de calculs
- choix des informations à sauver



L'objet dans notre système d'information

- objectif 100 millions de molécules
 - Plusieurs paramètres de calculs
 - choix des informations à sauver
-
- Calculs Gaussian
 - 500 Mo / molécule
 - de 10 à 1000 h / molécule
 - Objectifs :
 - 10 Mo / molécule
 - soit 1 Po pour les 100 millions

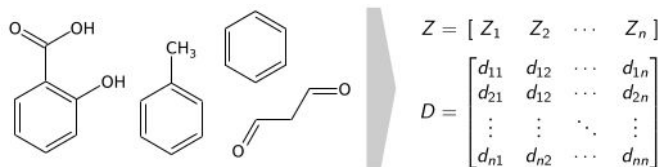


Projet AIQU

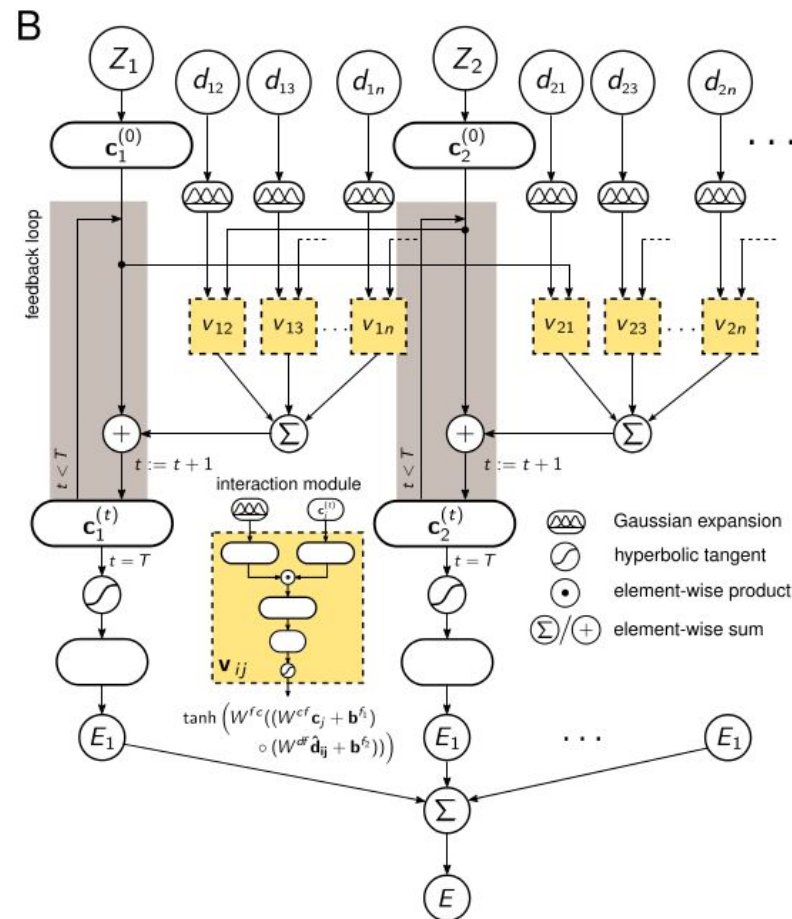
Objectif 1 :
Prédire les propriétés chimique

Travaux existants

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



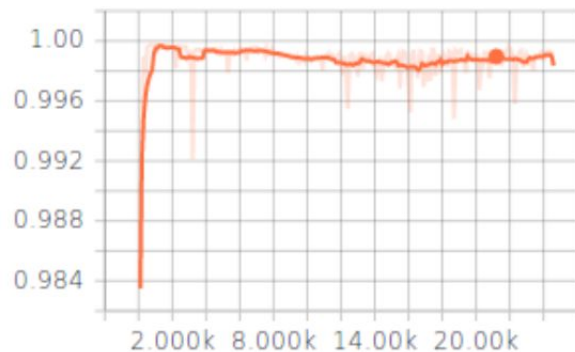
- GDB-9 dataset :
 - combinatoire de l'espace moléculaire pour 9 atomes lourds parmi C, N, O et F
 - ~134k petites molécules "théoriques"
- Données en entrée pas très homogènes
- RNN très structuré avec de nombreux a priori
- Matrice des distances (D) insensible aux translations et rotations.
 - Problématique de passage à l'échelle
- Prédiction uniquement de l'énergie



Travaux préliminaires : Nicolas Roux (M1)

- PubChemQC dataset
 - > 3 millions de molécules “réelles”
 - échantillon de l'espace moléculaire général
- Données homogènes (DFT, B3LYP, 6-31G*)
- NN simple (3 couches entièrement connectées)
- Prévu pour le passage à l'échelle
 - Matrice partielle des distances: insensible aux translations et rotations.
- Prédit les distances pour l'état fondamental à partir de distances “bruitées”.

accuracy_ns/Validation



Loss/Validation

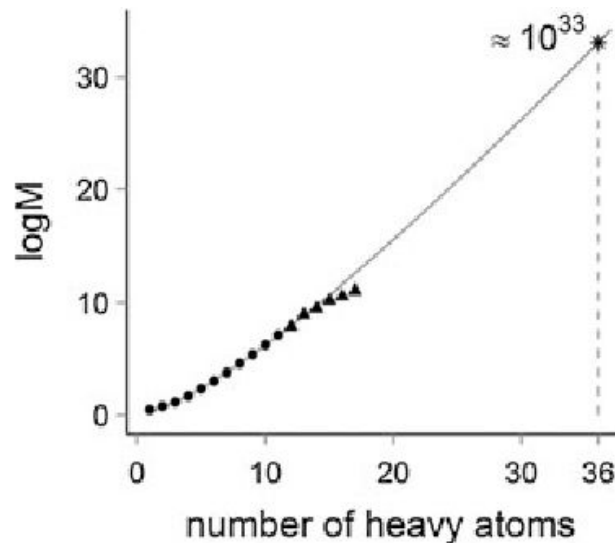


Projet AIQU

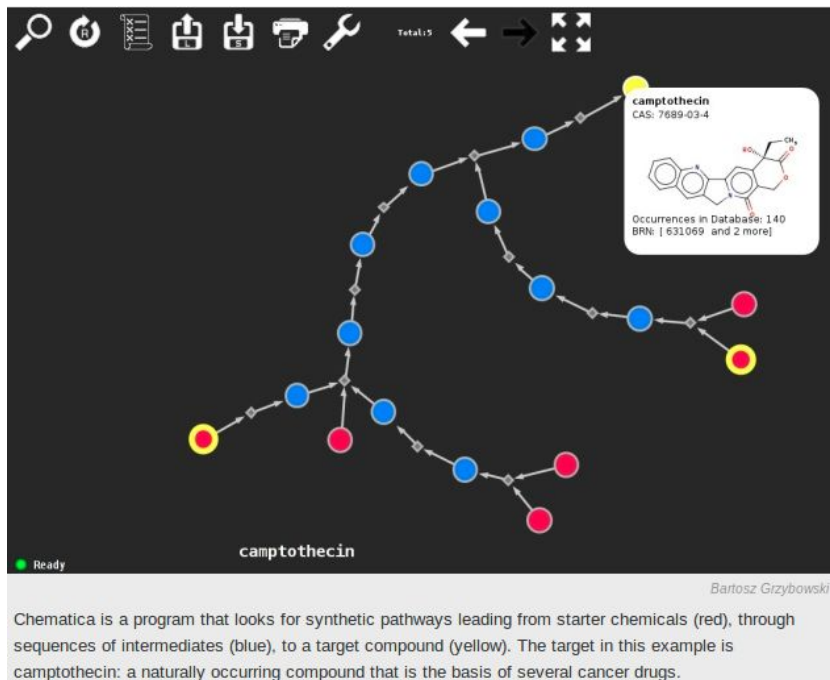
Objectif 2 :
Parcourir l'espace moléculaire

Fonctions d'évaluation + relations de voisinage = algorithmes de recherche locale

- Fonctions d'évaluation
 - Prédire les propriétés chimiques
 - Prédire le coût de synthèse
- Relations de voisinage
 - Produire des molécules proches, pouvant exister et stables
- Comparaison avec les techniques génératives issues de l'apprentissage profond



Evaluer le coût de synthèse ?



- La **rétrosynthèse** (1) planifie l'ensemble des réactions permettant la synthèse d'une cible à partir d'un catalogue de composés
 - Informatif pour les chimistes
 - Chiffrage précis du nombre de réactions
 - Difficile d'associer un coût
 - Coûteux en temps de calcul
- Solutions envisagées :
 - **Apprentissage** (2) : Prédiction rapide mais imparfaite du coût de synthèse
Bowen L. et al. Retrosynthetic reaction prediction using neural sequence-to-sequence models. CoRR 2017.
 - **Programmation par contraintes** (3)
 - (1) + (3) pour entraîner (2)

Parcourir le voisinage des molécules stables ?

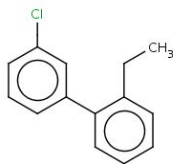
Search done!

Retrieved **1000** neighbors of c2ccc(Cc1ccccc1)cc2 from GDB-17 using **1.096** seconds server time.

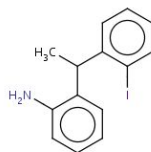
Store complete result to SMILES file

(Click on image below to get smiles for each molecule)

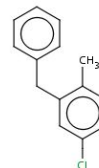
Molecules



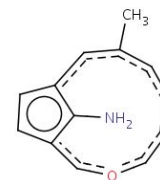
1, Dist:- 3



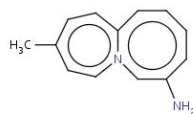
2, Dist:- 3



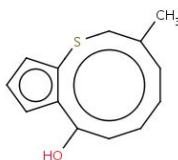
3, Dist:- 4



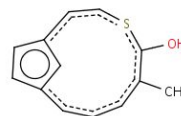
4, Dist:- 4



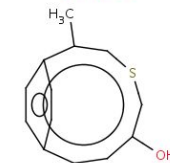
5, Dist:- 4



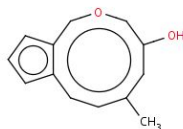
6, Dist:- 4



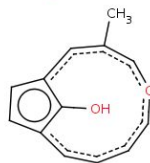
7, Dist:- 4



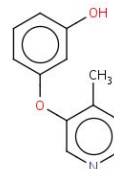
8, Dist:- 4



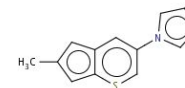
9, Dist:- 4



10, Dist:- 4



11, Dist:- 4

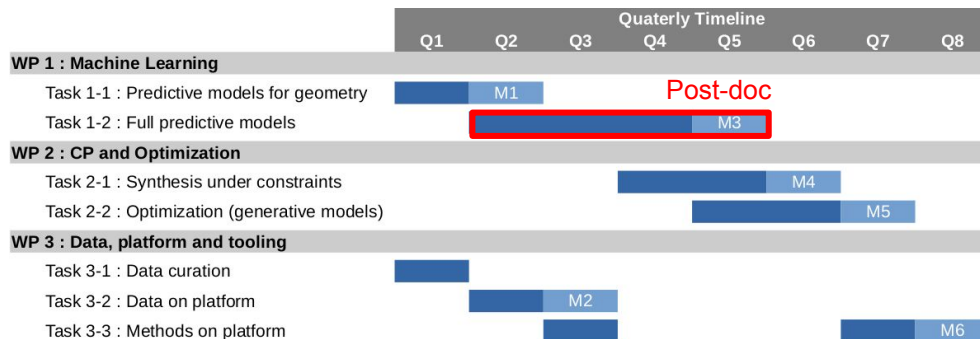


12, Dist:- 4

Conclusion

Jalons du projet, post-doc et valorisation

- **M1** : Modèles prédictifs de la géométrie
- **M2** : Données publiques sur la plateforme
- **M3** : Modèles prédictifs d'un calcul en chimie quantique
- **M4** : Modélisation des contraintes et du coût de synthèse
- **M5** : Modèles génératifs
- **M6** : Outils méthodologiques et techniques sur plateforme



Stratégie de valorisation :

1. Une plateforme publique internationale
2. Des publications en chimie et en informatique
3. Des modèles prédictifs libres et utilisables
4. Un projet voué à s'ouvrir à des partenaires extérieurs

Quantum chemistry

POSTULATE I. For any possible state of a system, there is a function, Ψ , of the coordinates of the parts of the system and time that completely describes the system.

$$\Psi = \Psi(x, y, z, t).$$

The quantity $\Psi^* \Psi d\tau$ is proportional to the probability of finding the particles of the system in the volume element, $d\tau = dx dy dz$. We require that the total probability be unity (1) so that the particle must be *somewhere*. That is,

$$\int_{\text{all space}} \Psi^* \Psi d\tau = 1. \quad (2.6)$$

If this condition is met, then Ψ is *normalized*. In addition, Ψ must be *finite*, *single valued*, and *continuous*. These conditions describe a “well-behaved” wave function. The reasons for these requirements are as follows:

Quantum chemistry

POSTULATE IV. The state function, Ψ , is given as a solution of

$$\hat{H}\Psi = E\Psi, \quad (2.38)$$

where \hat{H} is the operator for total energy, the *Hamiltonian operator*.

$$\hat{H} = \hat{T} + \hat{V}, \quad (2.39)$$

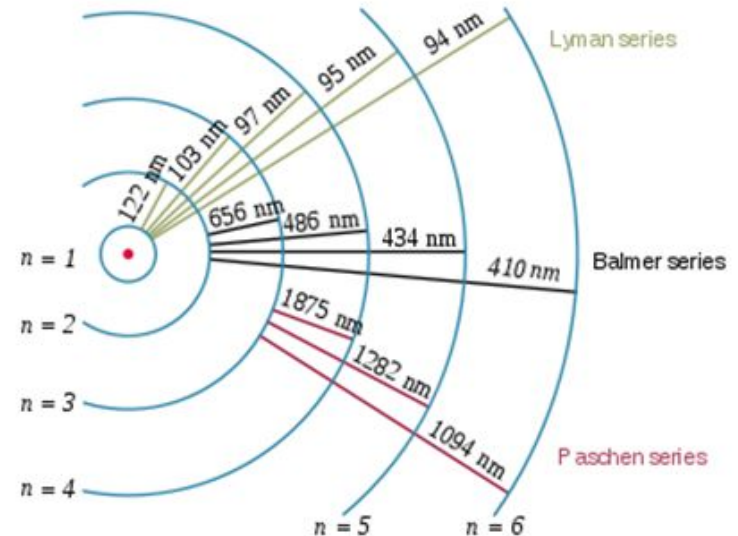
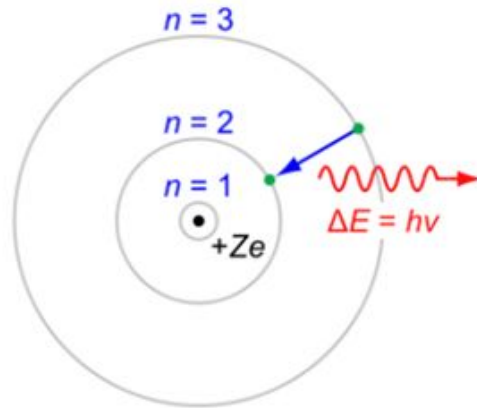
where \hat{T} is the operator for kinetic energy and \hat{V} is the operator for potential energy.

Quantum model of atoms

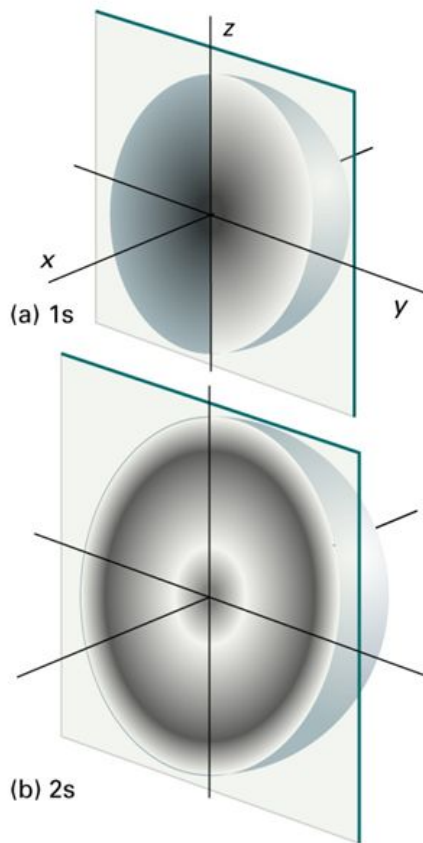
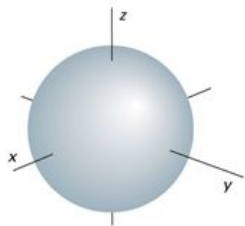
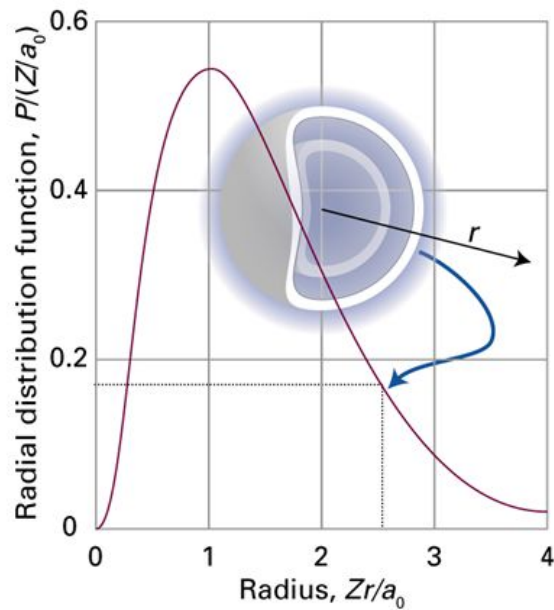


© The Nobel Foundation

Niels Bohr
1885-1962
Université
de Copenhague,
Danemark



Atomic orbitals



Molecular orbitals

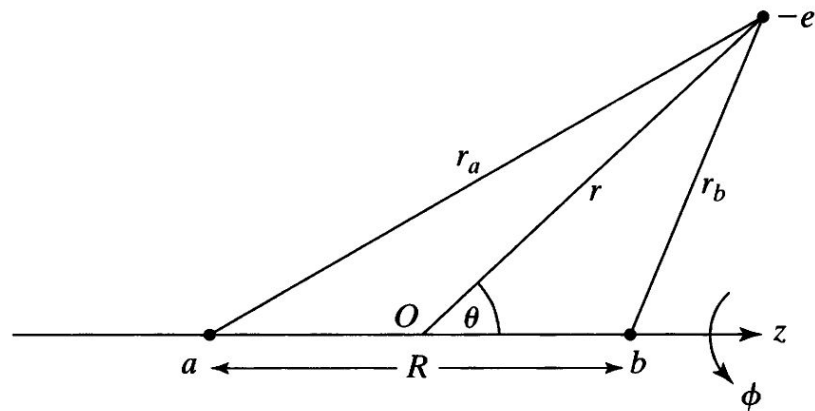


FIGURE 13.3 Interparticle distances in H_2^+ .

$$\hat{T} = - \sum_i^N \frac{\hbar^2}{2m_e} \nabla_i^2 - \sum_k^M \frac{\hbar^2}{2m_k} \nabla_k^2$$

$$V = - \sum_i^N \sum_k^M \frac{Z_k e^2}{4\pi\epsilon_0 r_{ik}} + \sum_i^N \sum_{j>i}^N \frac{e^2}{4\pi\epsilon_0 r_{ij}} + \sum_k^M \sum_{l>k}^M \frac{Z_k Z_l e^2}{4\pi\epsilon_0 R_{kl}}$$

Molecular orbitals

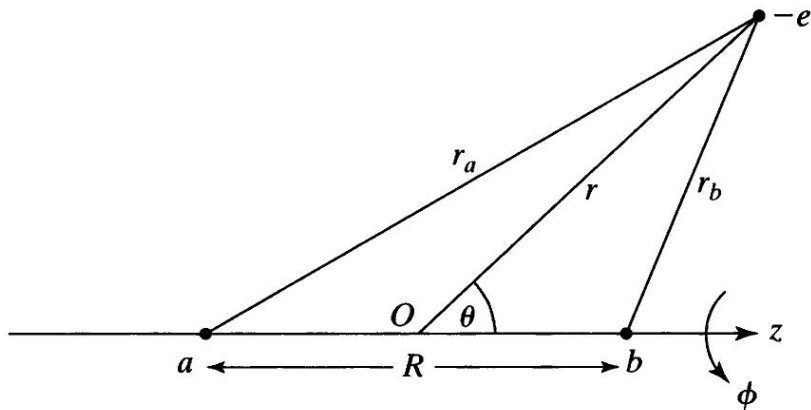
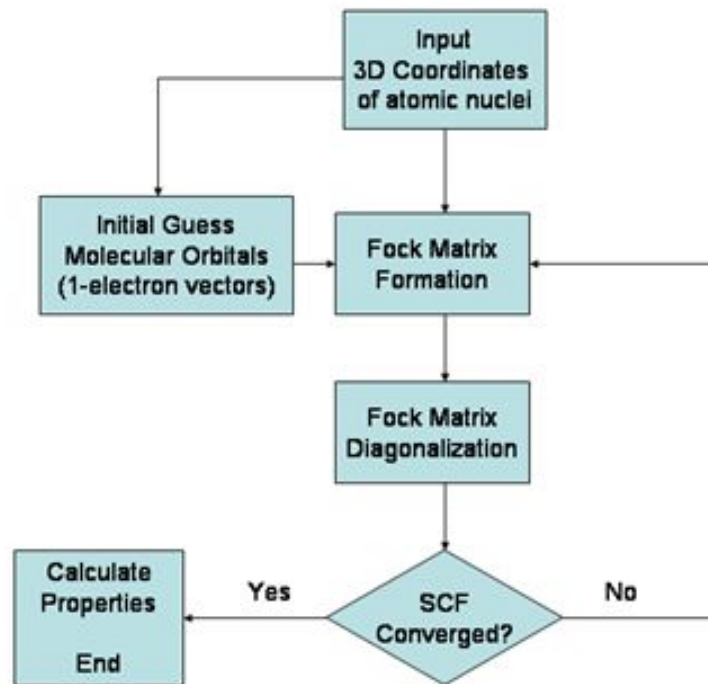


FIGURE 13.3 Interparticle distances in H_2^+ .

$$\hat{T} = - \sum_i^N \frac{\hbar^2}{2m_e} \nabla_i^2 - \sum_k^M \frac{\hbar^2}{2m_k} \nabla_k^2$$

$$V = - \sum_i^N \sum_k^M \frac{Z_k e^2}{4\pi\epsilon_0 r_{ik}} + \sum_i^N \sum_{j>i}^N \frac{e^2}{4\pi\epsilon_0 r_{ij}} + \sum_k^M \sum_{l>k}^M \frac{Z_k Z_l e^2}{4\pi\epsilon_0 R_{kl}}$$



Molecular orbitals

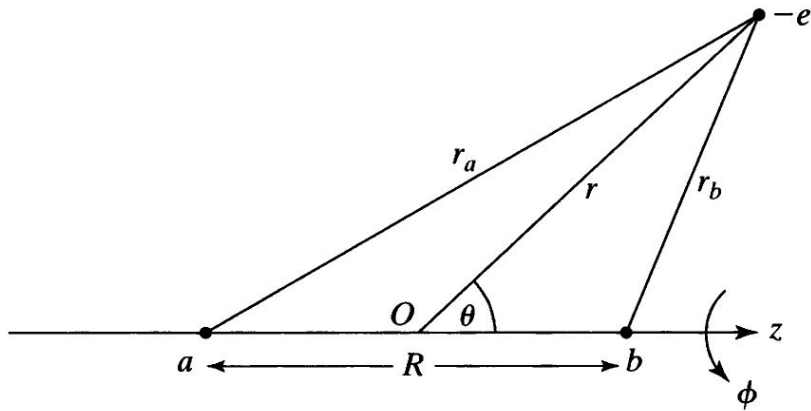
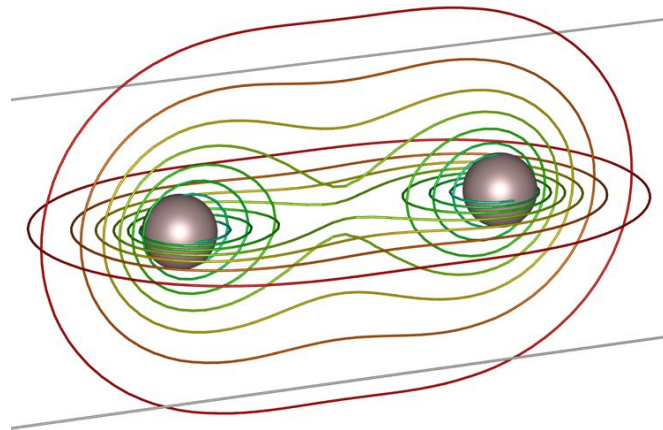
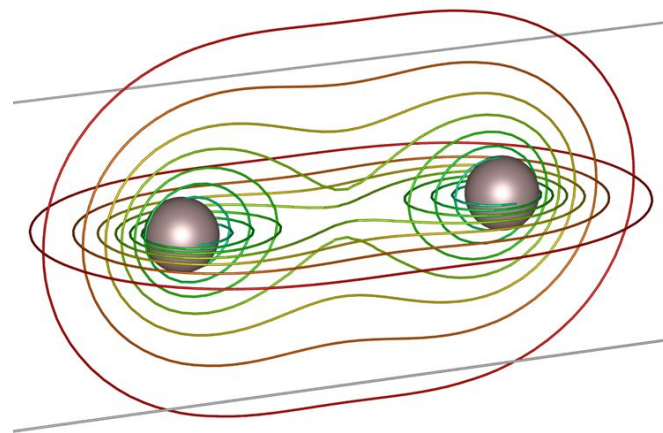
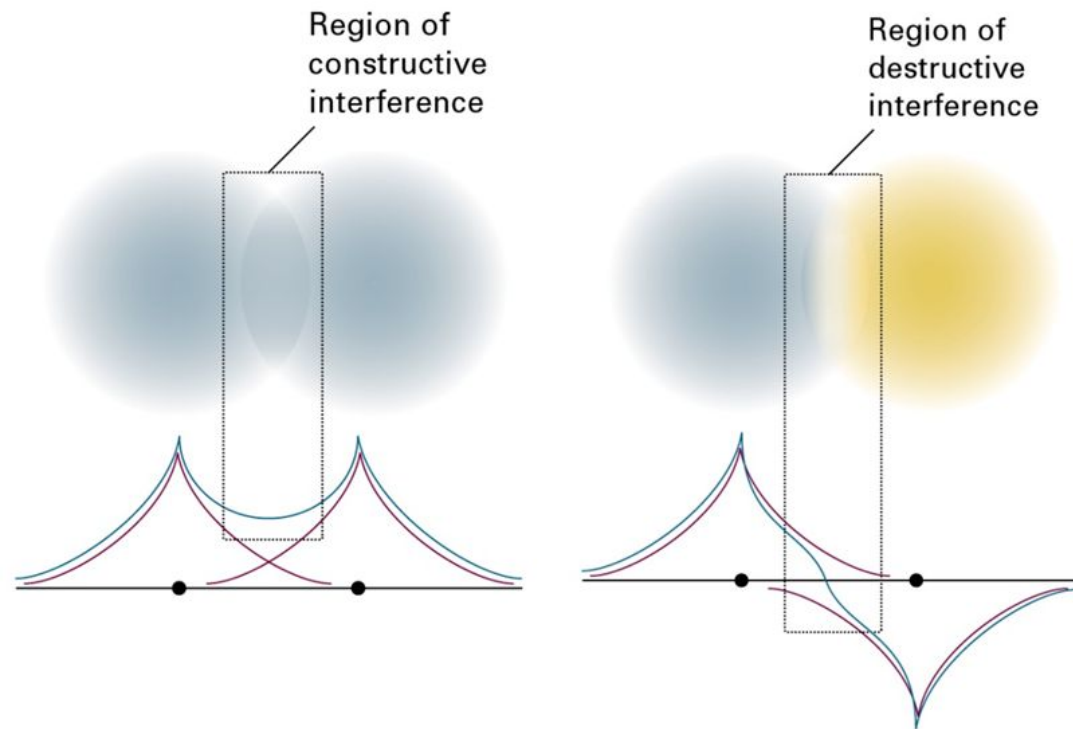


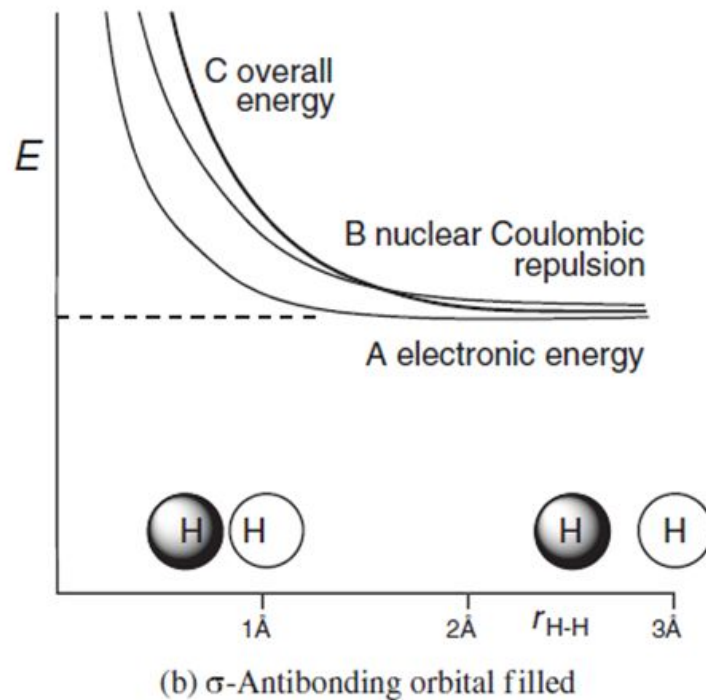
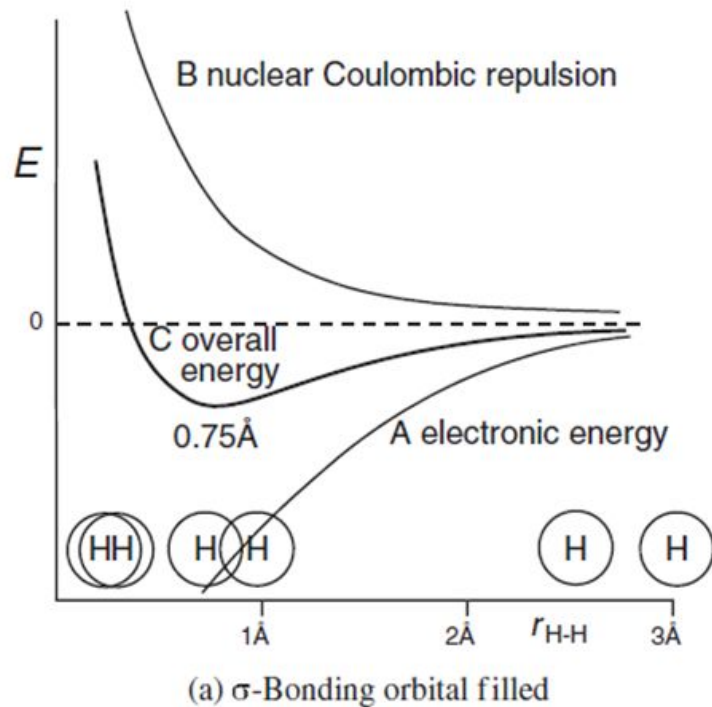
FIGURE 13.3 Interparticle distances in H_2^+ .



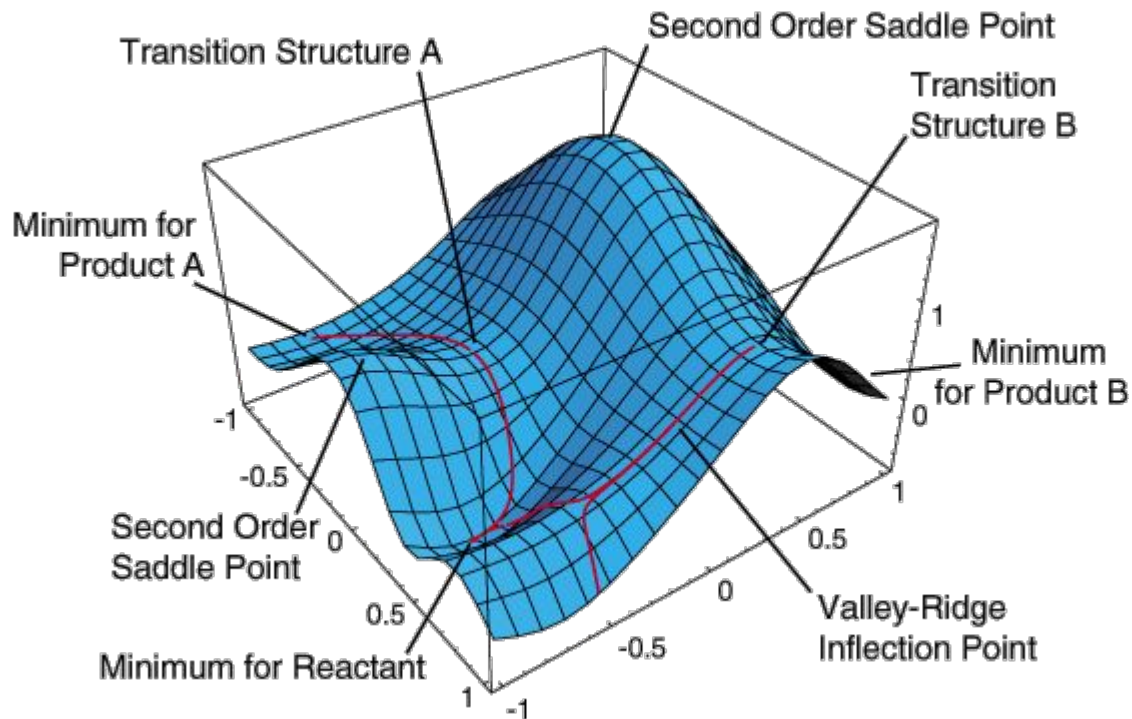
Molecular orbitals



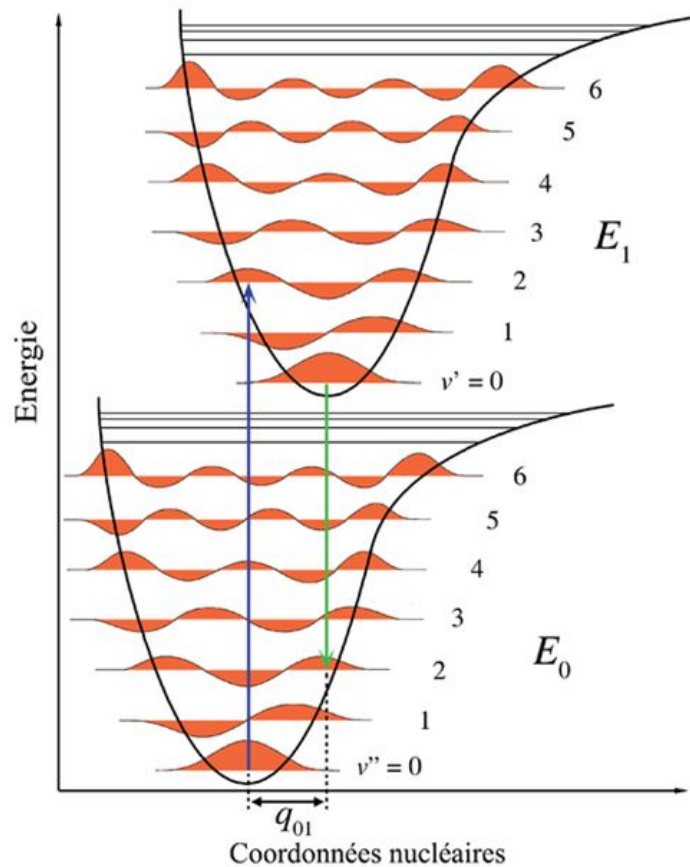
Molecular orbitals



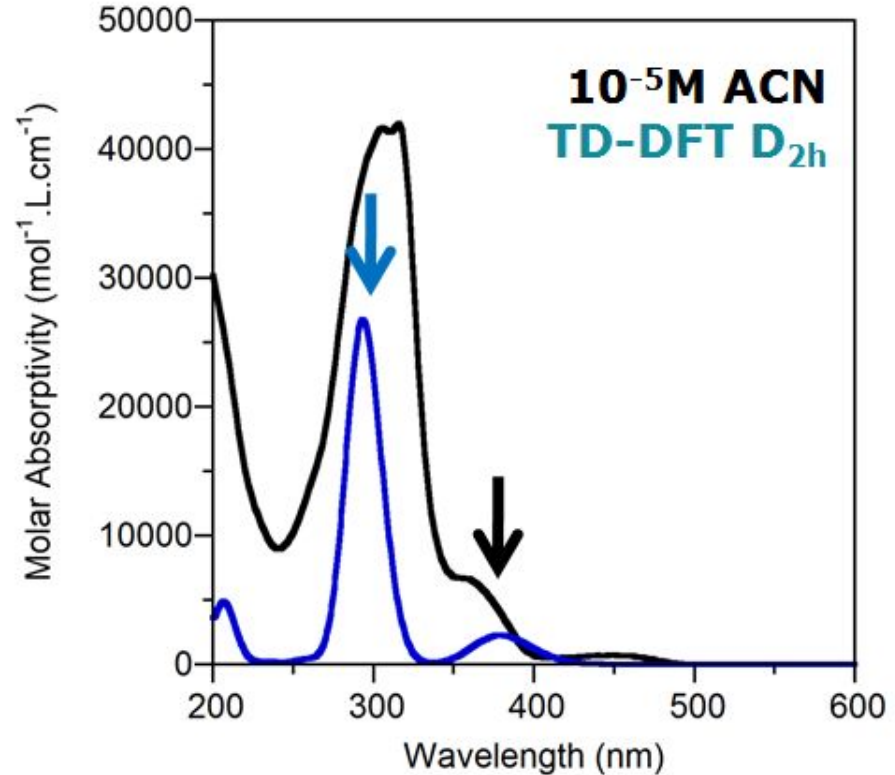
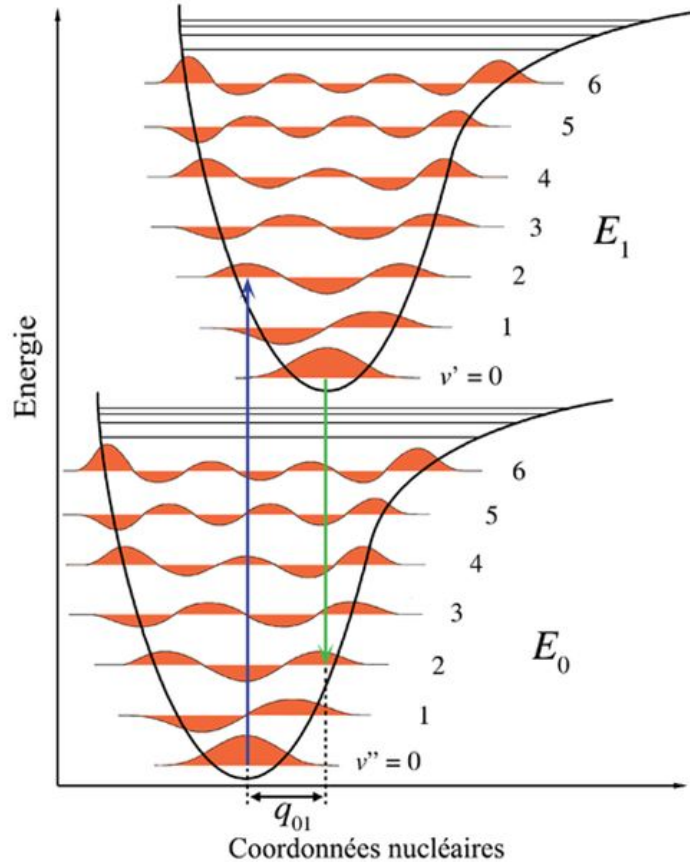
Molecular calculations: after SCF, OPT



Molecular calculations: after OPT, freq puis TD

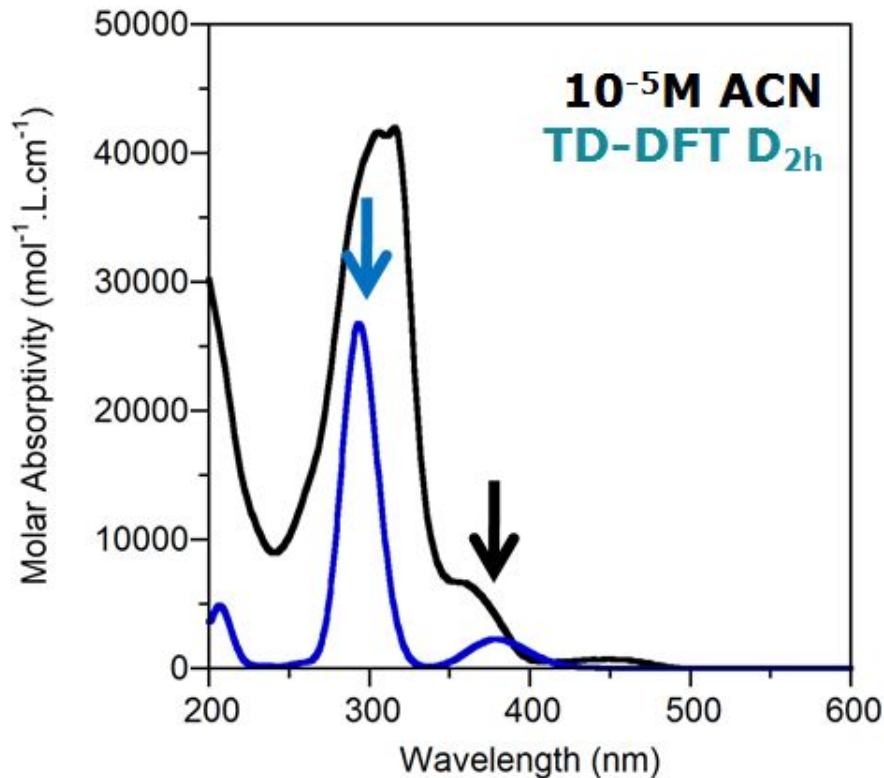
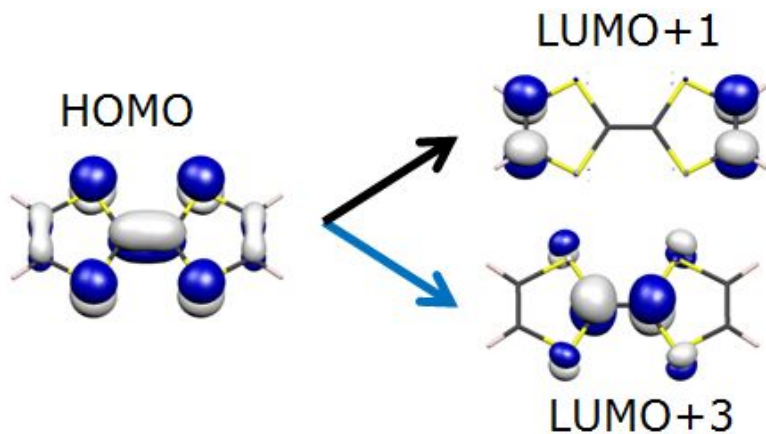


Molecular calculations: TD = états excités



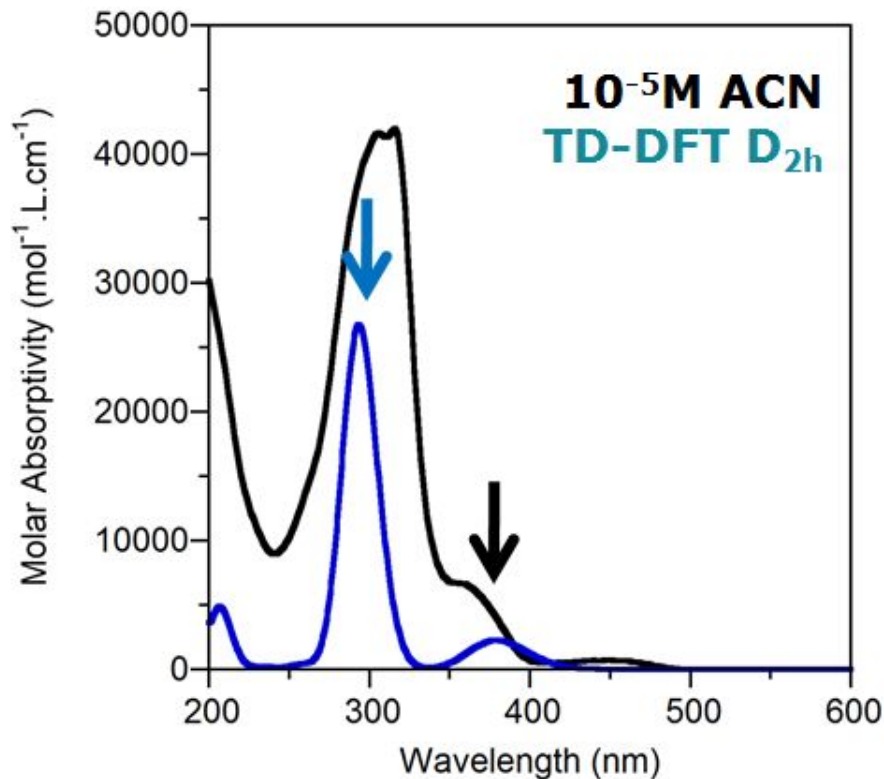
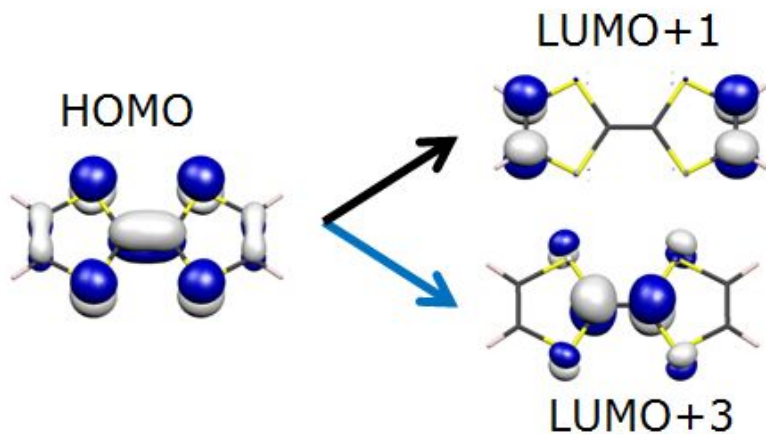
Molecular calculations: TD = états excités

λ (nm)	$f.$	transition
379	0.03	HO \rightarrow LU+1 (99%)
293	0.37	HO \rightarrow LU+3 (99%)

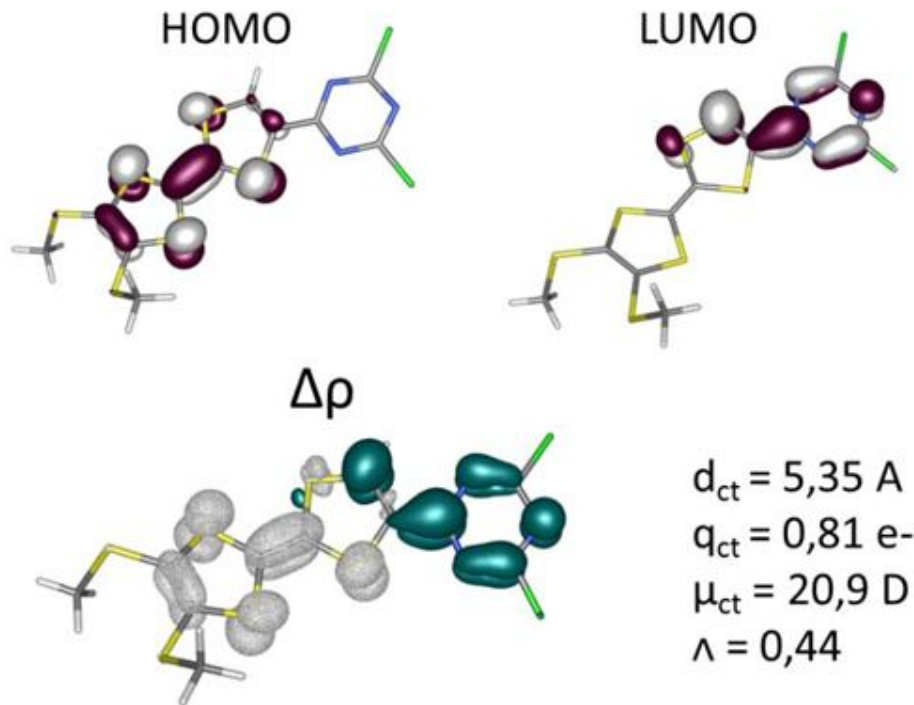
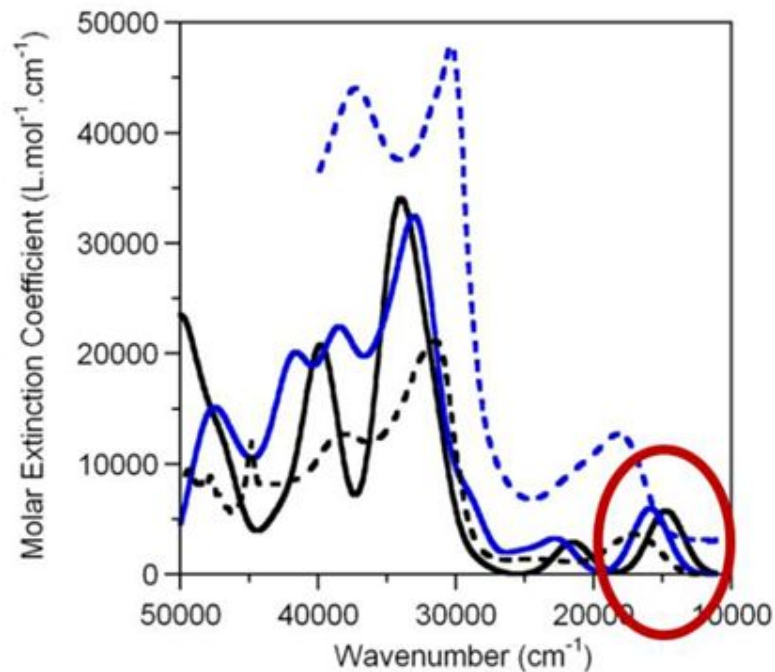


Molecular calculations: TD = états excités

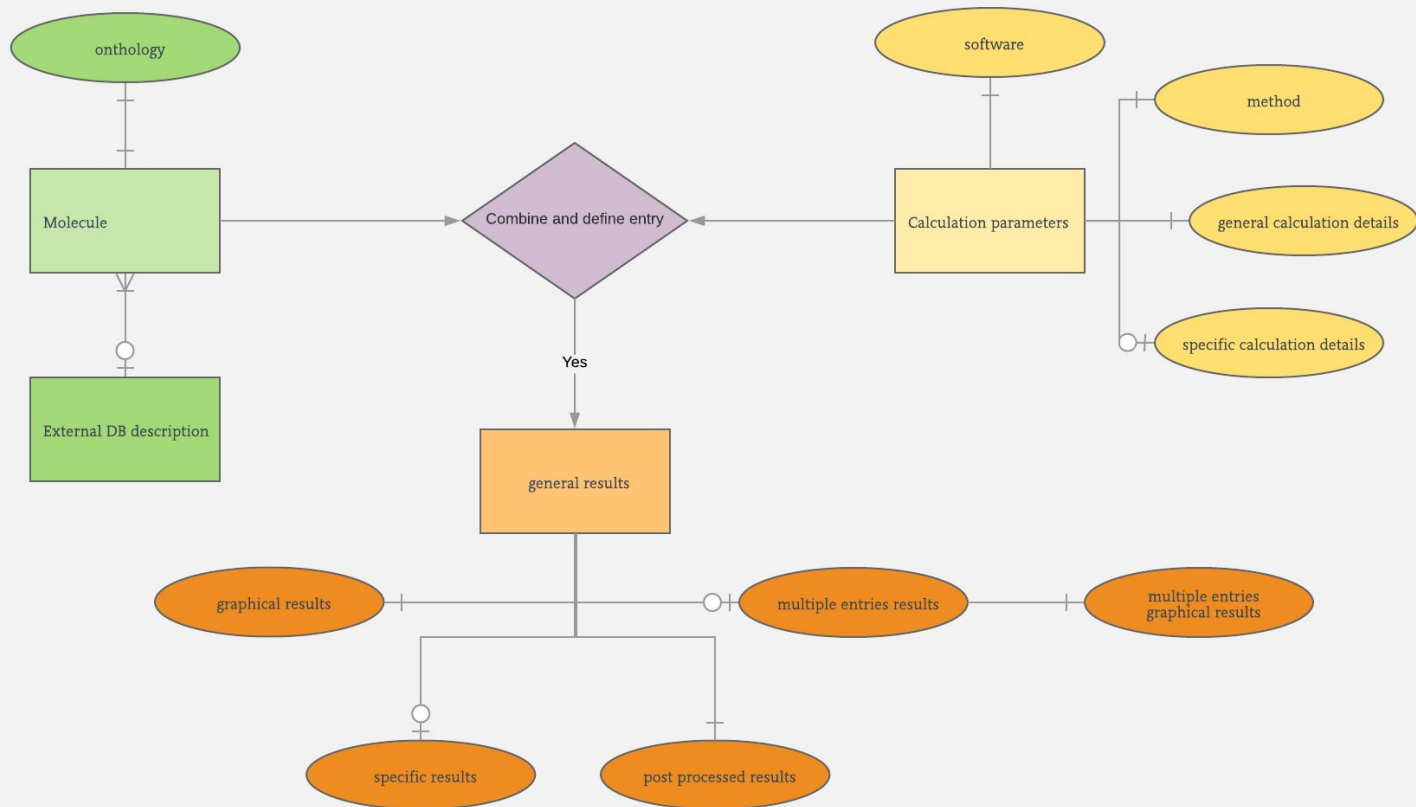
λ (nm)	$f.$	transition
379	0.03	HO \rightarrow LU+1 (99%)
293	0.37	HO \rightarrow LU+3 (99%)



Molecular calculations: TD = états excités



Results



Results

