



HAL
open science

Accuracy of Computational Chemistry Methods to Calculate Organic Contaminant Molecular Properties

Kevin Bonnot, Pierre Benoit, Sophie Hoyau, Laure Mamy, Dominique Patureau, Rémi Servien, Mathias Rapacioli, Fabienne Bessac

► **To cite this version:**

Kevin Bonnot, Pierre Benoit, Sophie Hoyau, Laure Mamy, Dominique Patureau, et al.. Accuracy of Computational Chemistry Methods to Calculate Organic Contaminant Molecular Properties. *ChemistrySelect*, 2022, 7 (48), pp.e202203586. 10.1002/slct.202203586 . hal-03920181

HAL Id: hal-03920181

<https://hal.science/hal-03920181v1>

Submitted on 3 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **Accuracy of Computational Chemistry Methods to Calculate Organic Contaminant Molecular**
2 **Properties**

3 Dr. Kevin Bonnot^{[a],[b]}, Prof. Dr. Pierre Benoit^[b], Dr. Sophie Hoyau^[c], Prof. Dr. Laure Mamy^[b], Prof. Dr.
4 Dominique Patureau^[a], Dr. Rémi Servien^[a], Dr. Mathias Rapacioli^[c], Dr. Fabienne Bessac^{[c],[d]*}

5 [a] INRAE, Univ Montpellier, LBE, 102 avenue des Etangs, 11100, Narbonne, France.

6 [b] Université Paris-Saclay, INRAE, AgroParisTech, UMR ECOSYS 78850 Thiverval-Grignon, France.

7 [c] Université de Toulouse; Laboratoire de Chimie et Physique Quantiques (UMR 5626), UPS, CNRS;
8 118, route de Narbonne; F-31062 Toulouse, France.

9 [d] Université de Toulouse; INPT; Ecole d'Ingénieurs de Purpan; 75, voie du TOEC, BP 57611, F-31076
10 Toulouse Cedex 03, France.

11 *fabienne.bessac@irsamc.ups-tlse.fr

12 <https://www.lcpq.ups-tlse.fr/spip.php?article1234>

13

14 **Abstract**

15 The quantitative structure activity relationship (QSAR) methodology has been developed and
16 extensively used to predict unknown environmental data for compounds that have not been
17 experimentally studied yet. QSAR is based on a large series of descriptors: such as the number of
18 atoms, the number of bonds... (descriptive), or based on the 2D structure of the molecule (connectivity
19 indices...) or on its 3D structure (dipole moment, polarizability...). Among them, quantum-based 3D
20 descriptors appear as promising tools to predict macroscopic environmental properties. For a set of
21 104 pharmaceuticals and personal care products, four quantum-based 3D descriptors (electric dipole
22 moment, polarizability, HOMO energy and ionization potential) were calculated using different
23 computational chemistry strategies involving a conformational search followed by local quenches
24 within three different frameworks: density functional theory (DFT), semi-empirical Austin Model 1
25 (AM1) approach, and density functional based tight binding (DFTB). Comparing the results obtained
26 using each framework highlights the necessity of a comprehensive conformational search and the use
27 of an accurate potential for the local quenches. Using the combination of a global exploration through
28 molecular dynamics with local quenches at B3LYP/6-31G* (DFT) allows the calculation of accurate and
29 trackable quantum-based 3D descriptors.

30

31 Introduction

32 Pharmaceuticals and personal care products (PPCP) have been increasingly studied over the past
33 twenty years. This emerging contaminants family contains diverse groups of organic compounds,
34 named according to their use or biological response, such as antibiotics, hormones, anti-inflammatory
35 drugs, antiepileptic drugs, blood lipid regulators, β -blockers, contrast media and cytostatic drugs for
36 pharmaceuticals; antimicrobial agents, synthetic musk, insect repellents, preservatives, fragrance
37 components and sunscreen UV filters for personal care products^[1,2]. Understanding the fate and
38 effects of PPCP in the environment is therefore essential from human and environmental health
39 perspectives. However, due to the high number of substances and of their transformation products,
40 their fate cannot be studied experimentally on a case-by-case basis^[3]. Consequently, numerous tools
41 using computational and statistical methods have been developed to predict unknown environmental
42 data for compounds that have not been experimentally studied yet. Among them, Quantitative
43 Structure Activity Relationships (QSAR) are based on the assumptions that the structure of the lowest-
44 energy conformer of a molecule in the gas phase contains the features responsible for its physical and
45 chemical properties despite the environmental and/or biological interactions, which inevitably
46 modifies the conformation of the compound in use. For instance, QSAR has been used to predict the
47 fate of various organic compounds^[4-6]. Mamy *et al.*^[6,7] have compiled many of these QSAR used to
48 predict volatilization, transformation, etc. and have identified the common and generic structural
49 parameters used in these equations. Forty common structural parameters, combining constitutional,
50 topological, geometric, and electronic properties, were thus calculated for 500 organic compounds
51 and transformation products, and implemented in the database of a recently developed clustering
52 tool: TyPol (Typology of Pollutants). TyPol allows the classification of organic compounds according to
53 both their overall behavior in the environment (characterized through mobility, persistence,
54 volatility...) and ecotoxicological effects (bioconcentration factor, no observed effect concentration...),
55 and to the set of the 40 structural parameters^[8]. These latter gather molecular descriptors such as the
56 number of atoms, the number of bonds... (descriptive) or based on the 2D structure of the molecule
57 (connectivity indices...) or on its 3D structure (dipole moment, polarizability...). Most molecular
58 descriptors used in TyPol are provided by the Dragon software^[9] but the calculation of the 3D quantum-
59 based descriptors that appear especially powerful to predict macroscopic environmental
60 properties^[6,10], needs to use other softwares and methods. Some properties have already been
61 computed for organic molecules either on few 3D quantum based descriptors or on relatively small
62 molecules^[11-13]. Moreover, the level of theory used to compute 3D quantum-based descriptors
63 influences the obtained values by two ways: (i) the structural conformation of the lowest-energy
64 isomer; (ii) the accuracy of the methods to compute the set of quantum descriptors. Semi-empirical
65 approaches such as Austin Model 1 (AM1) or PM3 (Parametric Method 3) are often used to predict the
66 physicochemical properties of a family of molecules^[14,15] or their behavioral parameters^[6]. The main
67 advantage of those approaches relies on their low computational costs (time and memory). For
68 example, up to now, AM1 is used in TyPol after a conformational search made "by hand", to evaluate
69 the 3D quantum descriptors listed in the database^[8].

70 Therefore, the objective of this work was to compare several computational chemistry strategies to
71 calculate the quantum descriptors of organic compounds from the case study of 104 PPCP
72 implemented in TyPol. The update of the database will be performed by non-computational chemists.
73 Thus, an easy-to-use program chain will be implemented. In this paper, the selected potentials will be
74 first described, followed by the impact of the conformational search method on the determination of

75 the lowest-energy isomer. Then, the effect of the potential used into the quantum descriptor
76 calculation is discussed. Afterwards, a comparison between the quantum descriptor values presently
77 in TyPol and the ones computed at the recommended level is done. Then, the effect of the temperature
78 on the set of quantum descriptors is studied using a mean value over the five lowest-energy isomers
79 with weights based on the Boltzmann distribution. Finally, connections between the quantum and the
80 non-quantum descriptors are attempted.

81

82 **Materials and Methods**

83 **Computational Chemistry Methods**

84 Three different frameworks were used for the quantum calculation of the molecular electronic
85 structures of PPCP in the gas phase:

- 86 • the density functional theory (DFT);
- 87 • an approximated scheme of DFT, the Density Functional based Tight Binding (DFTB);
- 88 • a semi-empirical method (Austin Model 1, AM1).

89 *DFT*. Within the DFT framework,^{[16] [17]} the calculations were performed using the B3LYP hybrid
90 functional^[17–19]. Two basis sets have been used in this work. On the one hand, the 6-31G* basis set^[20]
91 ^{[21] [22]} was called “basis1” (see SI for more details). On the other hand, the 6-311+G(2d,2p) basis set^[23]
92 ^{[24,25] [22]} was denoted “basis2” (see SI for more details). In previous work on similar molecules
93 (pesticides), the authors showed that the B3LYP/basis1 level is a nice compromise between accuracy
94 and tractability to compute geometries with respect to correlated methods such as MP2 and CCSD(T).
95 However, using basis1 or basis2 could slightly change the order of the isomers close in energy.^[26,27] The
96 calculations were run using Gaussian09 package^[28].

97 *DFTB*. The mathematical expression for the DFTB energy is derived from DFT first principle theorems
98 and is parameterized from DFT reference calculations^{[29–31] [32–34]}. In this work, we have used the third
99 order version of DFTB^[32–34] (simply labelled DFTB in the following) with the 3-ob parameters set^[34]
100 (downloaded from the www.dftb.org website), a correction for the hydrogen bond interactions (see
101 Gaus et al.^[33]) and an empirical long-range dispersion correction^[35]. We used a Fermi electronic
102 distribution (temperature of 500 K) to avoid convergency issues. It was however removed (Fermi
103 temperature set to 0 K) for the final local optimization steps (see below). The polarizabilities were
104 obtained from the scheme proposed by Witek *et al.*^[36]. All DFTB calculations were performed with the
105 deMonNano code^[37] (see SI for more details).

106 *AM1*. AM1 is a parametric quantum mechanical molecular model^[38,39]. The parameterization of the
107 model was carried out with a particular attention on dipole moments, ionization potentials, and
108 geometries of molecules^[38,39]. Up to now, AM1 is used in TyPol to compute 3D quantum descriptors.

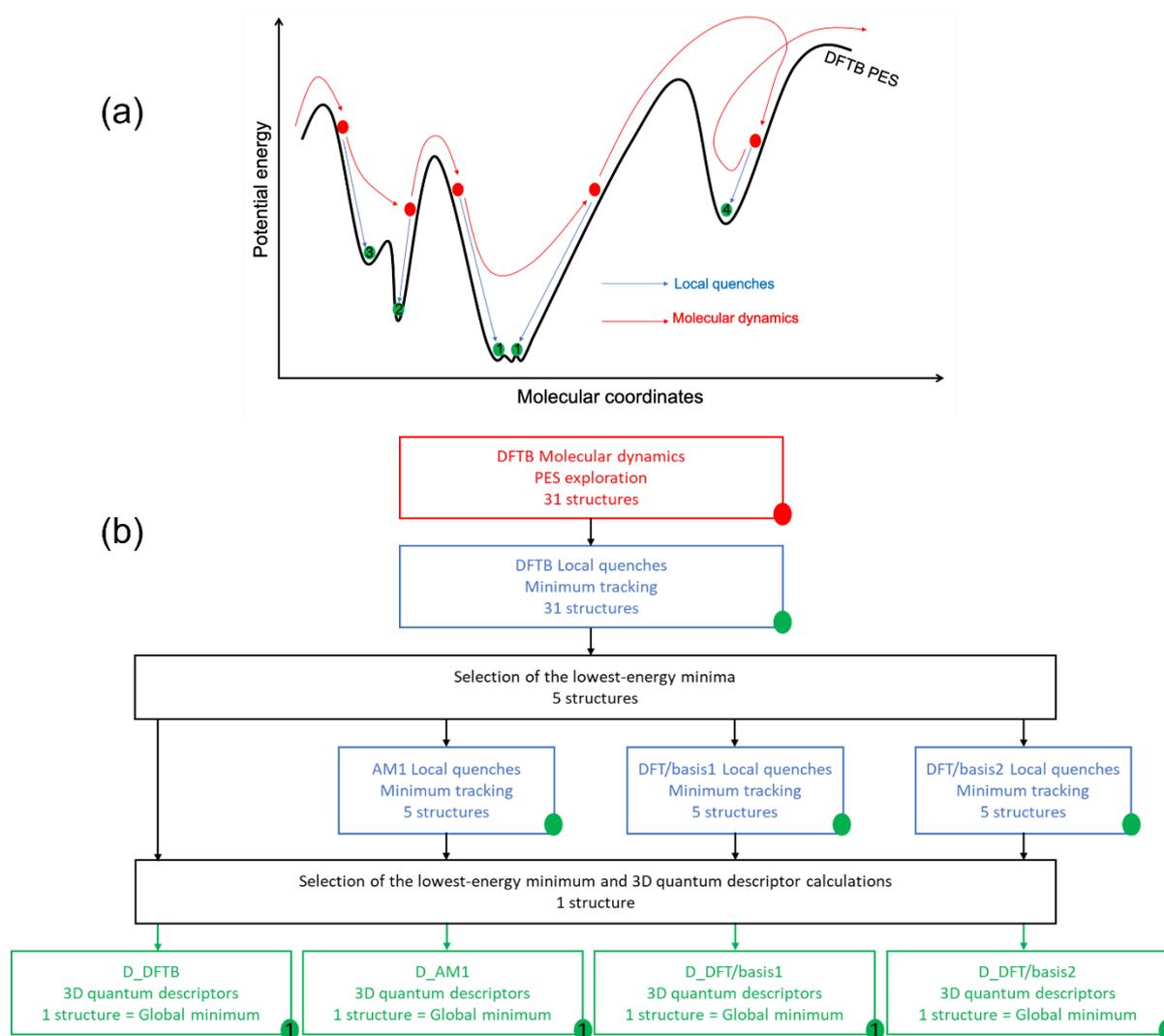
109 **Strategy**

110 Two strategies were compared to obtain the lowest-energy isomer structures of the 104 PPCP. The
111 first one, up to now used in TyPol, relies on angle and dihedral rotations made “by hand” to generate,
112 for each molecule, a guessed geometry, which is further optimized locally, making use of the
113 eigenvalue-following algorithm^[40–42] combined with the AM1 potential. These structures will be

114 labelled H_AM1 (for “by hand” designed + AM1 optimization) in the following. The second one relies
115 on a combination of global exploration through molecular dynamics (MD) with periodic local quenches
116 to improve the conformational search.

117 Molecular Dynamics

118 The MD exploration is performed at the DFTB level at high temperature to ensure that barriers
119 separating basins associated to various conformers can easily be overcome (Figure 1a).



120

121 Figure 1. (a) Schematic representation of the potential energy surface (PES) explored by DFTB
122 molecular dynamics simulations for a molecular system. (1) is the global minimum; (2), (3) and (4)
123 are local minima. In red, PES structures reached by the DFTB/MD exploration; in blue, path followed by
124 the quench (geometry optimization) leading to the closest local minimum. (b) Graphical representation
125 of the steps followed for the four molecular dynamic (MD) based strategies to calculate the quantum
126 descriptors of organic compounds from the case-study of 104 PPCP. Red and green dots have the same
127 meaning as in (a).

128

129 The temperature was maintained through a chain of 5 Nose Hoover thermostats at 800 K associated
130 to an energy exchange frequency of 300 cm⁻¹. Ten ps were simulated for each molecule starting from
131 a guessed structure. The timestep was set to 1 fs, which would probably be too large to extract
132 properties from such simulations but sufficient for the exploration purpose of the present work.

133 Local Quenches

134 Thirty-one structures were then extracted corresponding to snapshots equally spaced along the MD
135 run. The latter were further optimized at the DFTB level (conjugated gradient) and the five lowest-
136 energy ones were selected for further local optimizations at the AM1, B3LYP/basis1 and B3LYP/basis2
137 levels of theory using the Bery algorithm^[43]. These structures will be labelled D_DFTB, D_AM1,
138 D_DFT/basis1 and D_DFT/basis2 in the following, D_ accounting for the (DFTB/MD + quenches)
139 exploration scheme followed by the level of theory used for the final local optimization procedure. The
140 various strategies are summarized in Table I and Figure 1b.

141 Table I. Computational chemistry strategies used in this work with the associated labels. The first part
142 of the label refers to the potential energy surface (PES) exploration: H_ by hands or D_ using DFTB3
143 molecular dynamics simulations (MD). The second part of the label refers to the local optimization
144 level.

LABEL	PES exploration		Local optimization
	By hand	MD	
H_AM1	×		AM1
D_AM1		×	AM1
D_DFTB		×	DFTB3
D_DFT/basis1		×	B3LYP/6-31G*
D_DFT/basis2		×	B3LYP/6-311+G(2d,2p)

145

146 Properties

147 This work focuses on four 3D quantum-based chemical descriptors frequently used in QSAR, three of
148 them already present in TyPol^[6,8]: the dipole moment (Debye), the polarizability (Bohr³), the HOMO
149 energy (eV); and the ionization potential (eV).

150 The ionization potential IP (electron-Volt) is the minimal energy necessary to extract one electron from
151 a gas phase molecule in its neutral fundamental state. IP is the energy difference between the cation
152 (formed after the electron extraction) and the neutral molecule: $IP = E_{\text{cation}} - E_{\text{neutral}}$. In this work, we
153 only consider the vertical IP, *i.e.*, E_{cation} is the energy of the cation at the optimized neutral geometry.
154 In practice, IP can be computed by direct or indirect methods. In the indirect method, E_{cation} and E_{neutral}
155 are computed from two independent calculations. The direct method relies on the Koopman's
156 Theorem (KT)^[44], which stipulates that the first IP equals the opposite value of the HOMO (Highest
157 Occupied Molecular Orbital) energy of the neutral molecule: $IP_K = -E_{\text{HOMO}}$ (electron-Volt). Differences
158 exists between the two approaches. The KT considers that the electron is extracted from the HOMO.

159 Then, it only considers the HOMO energy and neglects the electronic reorganization after extracting
160 the electron (see SI for more details^{[45],[46–50]}).

161 **Dataset**

162 The dataset consisted of 104 PPCP. The most represented use families are antibiotics (34% of the data
163 set), antidepressants (8%), fragrance components (8%), hormones and NSAIDs (Non-Steroidal Anti-
164 Inflammatory Drugs) (7%). At the H_AM1 level, dipole moments, polarizabilities, and HOMO energies
165 were calculated for the 104 PPCP (the complete molecule set). At the D_AM1, D_DFTB, and
166 D_DFT/basis1 levels, dipole moments, polarizabilities, HOMO energies, and ionization potentials were
167 obtained for the 104 PPCP. At the D_DFT/basis2 level, dipole moments, polarizabilities, HOMO
168 energies, and ionization potentials were determined for 100 PPCP. For four molecules: azithromycin,
169 erythromycin, oleandomycin and roxithromycin; there were missing values because local geometry
170 optimizations did not converge.

171 **Results and Discussion**

172 **Conformational Search**

173 For extended molecules presenting very soft modes, it is worth noticing that the basins of the PES
174 often present multiple isomers very close in energy and differing by small geometrical atomic
175 displacements (Figure 1a). The fine details of these wells are sensitive to the level of theory used to
176 describe the PES. Therefore, it is not always an easy task to state if two structures optimized with two
177 different levels of theory correspond to the same isomer: it requires some degree of arbitrariness in
178 the selected procedures, which will be detailed in the following.

179 First, the results from the strategies involving a DFTB/MD exploration, namely D_AM1, D_DFTB,
180 D_DFT/basis1 and D_DFT/basis2, are discussed. The five lowest-energy isomers obtained with the
181 D_DFTB strategy were labelled according to their energetic position (1 to 5) before further
182 optimization at the AM1 or DFT levels of theory. If two strategies found the same isomer name (same
183 number) for their lowest-energy isomers, we considered that they agree on the structure of the most
184 stable isomer. This approach was called the identification procedure 1 (IdP1). The 104 molecules were
185 classified within five groups labelled by **0**, **1**, **2**, **3** and **4** indices as follows (these groups are **exclusive**,
186 see Table SI):

- 187 • If the lowest-energy isomer has the same name with D_AM1, D_DFTB, D_DFT/basis1 and
188 D_DFT/basis2 strategies, the molecule is indexed **1**;
- 189 • If the lowest-energy isomer has the same name only for D_DFTB, D_DFT/basis1 and
190 D_DFT/basis2 strategies, the molecule is indexed **2**;
- 191 • If the lowest-energy isomer has the same name for D_DFT/basis1 and D_DFT/basis2 strategies
192 but not with the D_DFTB strategy, the molecule is indexed **3**;
- 193 • If the lowest-energy isomers at D_DFT/basis2 and D_DFT/basis1 strategies have different
194 names, the molecule is indexed **4**;
- 195 • for four molecules, the D_DFT/basis2 local geometry optimizations did not converge and they
196 were indexed **0**.

197

198 Indices **0**, **1**, **2**, **3** and **4** gathered 3.8 %, 27.9%, 15.4%, 33.7% and 19.2% of the molecule set,
 199 respectively. In other words, the four levels of theory led to the same most stable structure after local
 200 optimization of the same five isomers for only 27.9% of the molecule set (index **1**). For 15.4% of the
 201 104 compounds, only the DFT like levels of theory, namely D_DFTB, D_DFT/basis1 and D_DFT/basis2
 202 gave the same lowest-energy isomer name (index **2**). For 19.2 % of the set, changing the basis for the
 203 DFT/B3LYP calculations, led to different most stable isomers (index **4**). These results point out the
 204 strong dependence of the identified most stable structures on the choice of the potential. The largest
 205 correspondence between the various computational methods was obtained for D_DFT/basis1 and
 206 D_DFT/basis2, which agreed for 80% of the molecule set (Table II). The agreement with the most
 207 accurate strategy, D_DFT/basis2, drops to less than 50% when the local optimization was performed
 208 at the DFTB or AM1 levels of theory (Table II), showing that working within the density functional
 209 theory framework brings a decisive and unquestionable improvement. D_DFTB and D_AM1 present
 210 similar abilities (between 44 and 50%) at recovering the D_DFT/basis2 or DFT/basis1 lowest-energy
 211 minima (Table II). It must be underlined that both parametrized strategies, D_DFTB and D_AM1, only
 212 agree for 49% of the molecule set (Table II).

213

214 Table II. Percentages of isomer correspondences among the studied computational chemistry
 215 strategies. The percentages written in black correspond to when both methods found the same isomer
 216 number for their lowest-energy isomers. The percentages in red correspond to when both methods
 217 found the same total energy (total energy difference $< 10^{-3}$ au) when the structure of the lowest-energy
 218 isomer of the most accurate method is reoptimized at the least accurate level of theory.

	<i>H_AM1</i>	<i>D_AM1</i>	<i>D_DFTB</i>	<i>D_DFT/basis1</i>	<i>D_DFT/basis2</i>
<i>H_AM1</i>		36	35	28	32
<i>D_AM1</i>			49/59	46/54	44/57
<i>D_DFTB</i>				50/50	48/52
<i>D_DFT/basis1</i>					80/90
<i>D_DFT/basis2</i>					

219

220 The identification procedure IdP1 only applies to strategies involving the DFTB/MD exploration as a
 221 first step. Consequently, as the isomer structures obtained at the H_AM1 level are not products of the
 222 MD exploration, the isomer numbering is not comparable to the other methods (D_) arising from the
 223 MD exploration. To allow comparison with the H_AM1 strategy, we defined a second procedure to
 224 compare the most stable isomers found by the two strategies: the structure of the lowest-energy
 225 isomer of the most accurate method (M1) was reoptimized at the least accurate level of theory (M2)
 226 giving E_{totM1} . Then, we compared this energy to E_{totM2} , the total energy of the lowest-energy isomer
 227 at M2 level by calculating the total energy difference $\Delta E_{tot} = E_{totM1} - E_{totM2}$. If $|\Delta E_{tot}| < 10^{-3}$ a.u., it
 228 was considered that both strategies found the same lowest-energy isomer. This new identification
 229 strategy, IdP2, was used to provide the red number in Table II.

230 When both ways to evaluate isomer correspondences (IdP1 and IdP2) were possible, IdP1 always gave
231 percentages of correspondence lower than those of IdP2 although presenting the same trend. The
232 largest difference is of 13 points (D_AM1 versus D_DFT/basis2), the smallest of 0 point (D_DFTB versus
233 D_DFT/basis1). Going vertically down through a column the percentages increase a lot (variations up
234 to 58%), whereas going horizontally left to right through a line, they remain mostly unchanged
235 ($|\text{variations}| \leq 8\%$).

236 H_AM1 had the lowest correspondence rates with any of the other methods, below 36%, meaning that
237 the MD PES exploration has provided a tremendous improvement in the conformational search and in
238 particular, the identification of the lowest-energy isomer (Table II). It was particularly striking to obtain
239 the poor agreement of 36%, when comparing H_AM1 and D_AM1 because the two methods differed
240 only by the MD exploration prescreening, and the final local optimization was performed at the same
241 level (AM1). In addition, when going from H_AM1 to D_AM1, the agreement with D_DFT/basis1
242 increases from 28 to 54 % (from 32 to 57 % for the agreement with D_DFT/basis2) (Table II).

243 As a conclusion, the MD PES exploration is mandatory due to the multiple minima exhibited by the PES
244 on such extended molecules. Moreover, facing the lack of experimental data for the molecules of the
245 set under study, comparisons will be made with respect to the highest level of theory of this study,
246 D_DFT/basis2. However, comparing to experimental results, Hait and Head-Gordon showed that B3LYP
247 with aug-pc-4, a polarization consistent basis set, gave 6.98% and 6.24% root-mean-square relative
248 errors in dipole moments and polarizabilities, respectively when CCSD(T) gave 3.95% and 1.62%^[12,13].
249 Their databases contained 132 (polarizability) and 200 (dipole moment) relatively small molecules (6
250 atoms max), while we deal with 104 compounds having between 16 and 134 atoms. Thus, our
251 calculations at the DFT/basis2 level have already been very heavy given the size of the molecules. A
252 reference level as CCSD(T) associated with a large basis set is simply not feasible for most of our
253 molecules, again given their size.

254

255 **Effects of the Computational Chemistry Methods on the Values of 3D Quantum Descriptors**

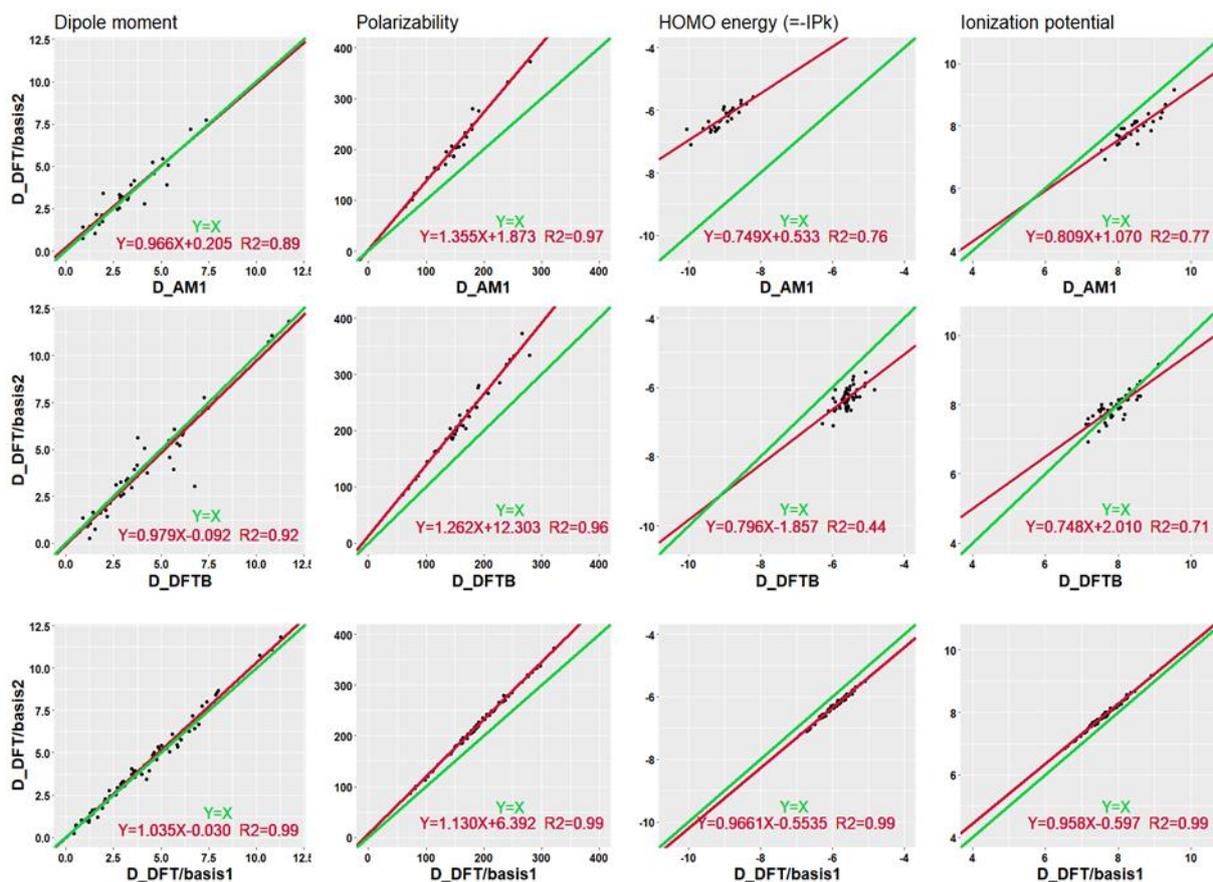
256 As the influence of the computational method on the value of the 3D quantum descriptors was
257 determined considering the D_DFT/basis2 strategy, only the matching molecules at both levels were
258 considered (IdP1): 29 molecules for D_AM1 (index **1**), 45 molecules for D_DFTB (indices **1** and **2**) and
259 80 molecules for D_DFT/basis1 (indices **1**, **2** and **3**). Each quantum descriptor computed at
260 D_DFT/basis2 level was compared to the descriptor computed at either D_AM1 or D_DFTB or
261 D_DFT/basis1 (Figure 2 and Table III).

262

263

264

265



266

267 Figure 2. For the following quantum descriptors: dipole moment (Debye), polarizability (Bohr³), HOMO
 268 energy (= -IP_K) (eV) and ionization potential (eV); straight lines in red are confronting D_DFT/basis2
 269 results on the vertical axis to D_AM1, D_DFTB and D_DFT/basis1 results. On each graph, the equation,
 270 $y=ax+b$, of the linear regression line is written in red with the corresponding R^2 regression coefficient,
 271 while the green straight line represents the first bisector: $y=x$. The set of molecules used to produce
 272 the graphs were: (i) index 1, for D_AM1 versus D_DFT/basis2 (29 molecules); (ii) indices 1 and 2 for
 273 D_DFTB versus D_DFT/basis2 (45 molecules); (iii) indices 1, 2 and 3 for D_DFT/basis1 versus
 274 D_DFT/basis2 (80 molecules). The indices are defined in **Results and Discussion**.

275

276

277

278

279

280

281

282

283 Table III. Relative absolute errors (%) (mean, median, minimum, and maximum) for four 3D quantum
 284 descriptors (dipole moment, polarizability, HOMO energy (= $-IP_K$) and IP, ionization potential)
 285 calculated at D_AM1, D_DFTB and D_DFT/basis1 in comparison to the highest level of theory of this
 286 study, D_DFT/basis2. The set of molecules used to produce the numbers were: (i) index **1**, for D_AM1
 287 versus D_DFT/basis2 (29 molecules); (ii) indices **1** and **2** for D_DFTB versus D_DFT/basis2 (45
 288 molecules); (iii) indices **1**, **2** and **3** for D_DFT/basis1 versus D_DFT/basis2 (80 molecules). The indices
 289 are defined in **Results and Discussion**.

		Dipole Moment	Polarizability	HOMO energy = $-IP_K$	Ionization Potential IP
Mean	D_AM1	15	27	45	7
	D_DFTB	25	26	11	2
	D_DFT/basis1	9	14	6	4
Median	D_AM1	11	27	45	7
	D_DFTB	12	26	11	2
	D_DFT/basis1	7	14	6	4
Min	D_AM1	1	21	39	2
	D_DFTB	0	16	3	0
	D_DFT/basis1	0	12	4	2
Max	D_AM1	48	36	53	15
	D_DFTB	393	32	20	7
	D_DFT/basis1	65	18	8	5

290

291 For the dipole moment, D_DFT/basis2 vs. D_DFT/basis1 led to the best correlation coefficient ($R^2 =$
 292 0.99). D_DFTB gives more dispersion ($R^2=0.92$) than D_DFT/basis1, and D_AM1 led to a correlation
 293 coefficient ($R^2=0.89$) lower than those of D_DFTB and D_DFT/basis1. At the same time, the slope the
 294 closest to one ($a=0.979$) was obtained with D_DFTB but the y-intercept the closest to zero was found
 295 for D_DFT/basis1 ($b=0.030$). Table III shows that, on average, the absolute values obtained for the
 296 dipole moment are better described at the D_DFT/basis1 level than at D_AM1 or D_DFTB levels. The
 297 average relative error is of 9% in D_DFT/basis1 compared to D_DFT/basis2 with a maximum of 65%
 298 and a minimum of 1%.

299 Concerning polarizability, the three methods underestimated the values with respect to D_DFT/basis2
 300 as shown by the deviation from the first bisector. D_DFT/basis1 was the best method to reproduce
 301 polarizability ($R^2=0.99$, $a=1.130$). D_DFTB and D_AM1 also performed well ($R^2=0.96$ and 0.97
 302 respectively) but led to a greater underestimation than D_DFT/basis1. The closest y-intercept to zero
 303 ($b=1.873$) was obtained with D_AM1. Again, according to Table III, D_DFT/basis 1 is the best method
 304 to reproduce the absolute values of polarizability. With respect to D_DFT/basis2, the relative errors on
 305 this descriptor are only 14% in average with a maximum of 18% and a minimum of 12%.

306 D_DFT/basis1 allowed the best determination of $IP_K (= -E_{HOMO})$ ($R^2=0.99$) although slightly
 307 underestimating IP_K comparing to D_DFT/basis2 ($a=0.9661$ and $b=-0.5535<0$). D_DFTB showed an

308 important dispersion ($R^2=0.44$) while it slightly underestimated the IP_k values. D_AM1 strongly
309 overestimated IP_k for all the molecules of this set with an intermediate dispersion ($R^2=0.76$) (index 1,
310 29 molecules). The average relative error made on IP_k in D_DFT/basis1 compared to D_DFT/basis2 is
311 only 6% (Table III), a value is clearly lower than those obtained for D_AM1 (45%) and D_DFTB (11%).

312 For IP (computed with the indirect method), D_DFT/basis1 was the most efficient method ($R^2=0.99$)
313 leading to a small underestimation of the values. D_DFTB showed a larger dispersion ($R^2=0.71$) with
314 small over- or under- estimations while D_AM1 led to a comparable one ($R^2=0.77$) and overestimated
315 IP values. The lowest average relative error regarding the IP evaluation with respect to D_DFT/basis2
316 (Table III) is obtained at the D_DFTB level (2%) but the performance of D_DFT/basis1 is almost
317 equivalent (4%). However, with D_DFTB, the IP is either underestimated or overestimated, depending
318 on the molecule, whereas with D_DFT/basis1, IP is systematically overestimated (Figure 2).
319 Consequently, D_DFTB can hardly be used to predict IP trends on the opposite to D_DFT/basis1.

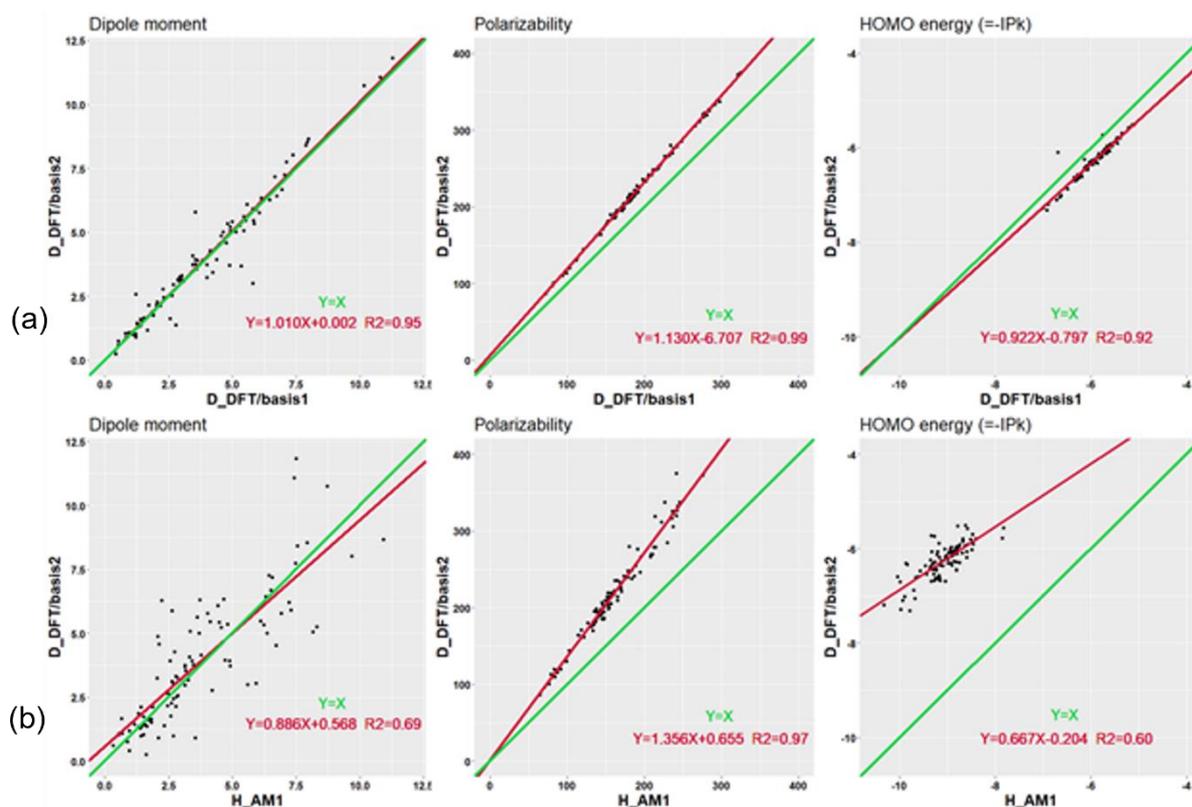
320 Finally, D_DFT/basis1 was the preferred method to compute the quantum descriptors as it gave the
321 best regression coefficients (Figure 2) for all quantum descriptors and the smallest average relative
322 errors (Table III) for all quantum descriptors except IP, with respect to D_DFT/basis2. Moreover, the
323 computations are more tractable with D_DFT/basis1 than with the highest computational level (for
324 local quench plus frequency calculations between 2.5 and 13 times more CPU time depending on the
325 molecules). Consequently, the D_DFT/basis1 strategy can be recommended to calculate 3D quantum
326 descriptors.

327
328 Moreover, for the molecules indexed 1, for which the four levels of theory led to the same most stable
329 structure (27.9% of the molecule set), computing the lowest-energy isomer (local quench plus
330 frequency calculation) at AM1 is between 91 and 1436 times faster than at DFT/basis1, and between
331 1204 and 34 425 times faster than at DFT/basis2. Local quench plus frequency calculations for the
332 lowest-energy isomer lasted between 2 and 73 seconds (~ 1 minute) at AM1, between 441 and 70 738
333 seconds (~ 20 hours) at DFT/basis1, and between 5 178 and 1 401 534 seconds (~ 16 days) at DFT/basis2.
334 Thus, a conformational search through molecular dynamics (D_) followed by local quenches at
335 DFT/basis1 was the best compromise to reach accuracy and a reasonable computational cost.

336

337 **H_AM1 versus D_DFT/basis1**

338 After discussing separately, the effect of a global exploration and the choice of a potential to compute
339 the quantum descriptors, we shine a light on the comparison between the previously implemented
340 strategy in TyPol, namely H_AM1, and the new recommended strategy D_DFT/basis1. The quantum
341 descriptor values determined by the H_AM1 strategy (presently in the TyPol database) and those
342 calculated using D_DFT/basis1 strategy were compared (Figure S1) for the set of molecules with indices
343 1, 2, 3 and 4 (100 molecules). D_DFT/basis1 and H_AM1 quantum descriptors are discussed with
344 respect to those computed at the D_DFT/basis2 level (see Figure 3a and 3b, respectively).



345

346 Figure 3. For the following quantum descriptors: dipole moment (Debye), polarizability (Bohr³) and
 347 HOMO energy (= -IP_K) (eV); straight lines in red are confronting D_DFT/basis2 results on the vertical
 348 axis to: (a) D_DFT/basis1 results; (b) H_AM1 results. On each graph, the equation, $y=ax+b$, of the linear
 349 regression line is written in red with the corresponding R² regression coefficient, while the green
 350 straight line represents the first bisector: $y=x$. Only the set of molecules with indices **1, 2, 3** and **4** (100
 351 molecules) was used to produce the graphs. The indices are defined in **Results and Discussion**.

352

353 For the dipole moment calculation, D_DFT/basis1 best matched D_DFT/basis2 ($a=1.01$, $b=0.002$) while
 354 less satisfactory results were obtained with H_AM1 ($a=0.886$, $b=0.568$) (Figure 3). In addition, the
 355 dispersion was significantly lower for D_DFT/basis2 vs. D_DFT/basis1 ($R^2 = 0.95 > R^2 = 0.69$ for
 356 D_DFT/basis2 vs. H_AM1) (Figure 3). Therefore, dipole moment values as listed in TyPol will be
 357 substantially improved using the D_DFT/basis1 strategy. As a set of 100 PPCP molecules was
 358 considered, the dispersion was more important than those observed on Figure 2 because only similar
 359 isomers according to IdP1 were selected (29 molecules D_DFT/basis2 vs. D_AM1; see **Effects of the**
 360 **Computational Chemistry Methods on the Values of 3D Quantum Descriptors**): $R^2 = 0.89$ for
 361 D_DFT/basis2 vs. D_AM1, and $R^2 = 0.99$ for D_DFT/basis2 vs. D_DFT/basis1. The differences originate
 362 from the molecular conformation, which was not the lowest in energy for many molecules using
 363 H_AM1 compared to D_AM1 (69 molecules D_DFT/basis2 vs. H_AM1 and 57 molecules D_DFT/basis2
 364 vs. D_AM1). It should be noted that the dipole moment strongly depends on the molecular
 365 conformation. Thus, a global exploration (D_) is needed to properly compute dipole moment values
 366 (see also **Temperature effects**).

367 For the polarizability, the dispersions obtained with D_DFT/basis2 vs. H_AM1 and D_DFT/basis2 vs.
 368 D_DFT/basis1 were acceptable ($R^2=0.97$ and 0.99 , respectively) (Figure 3). Indeed, the polarization,

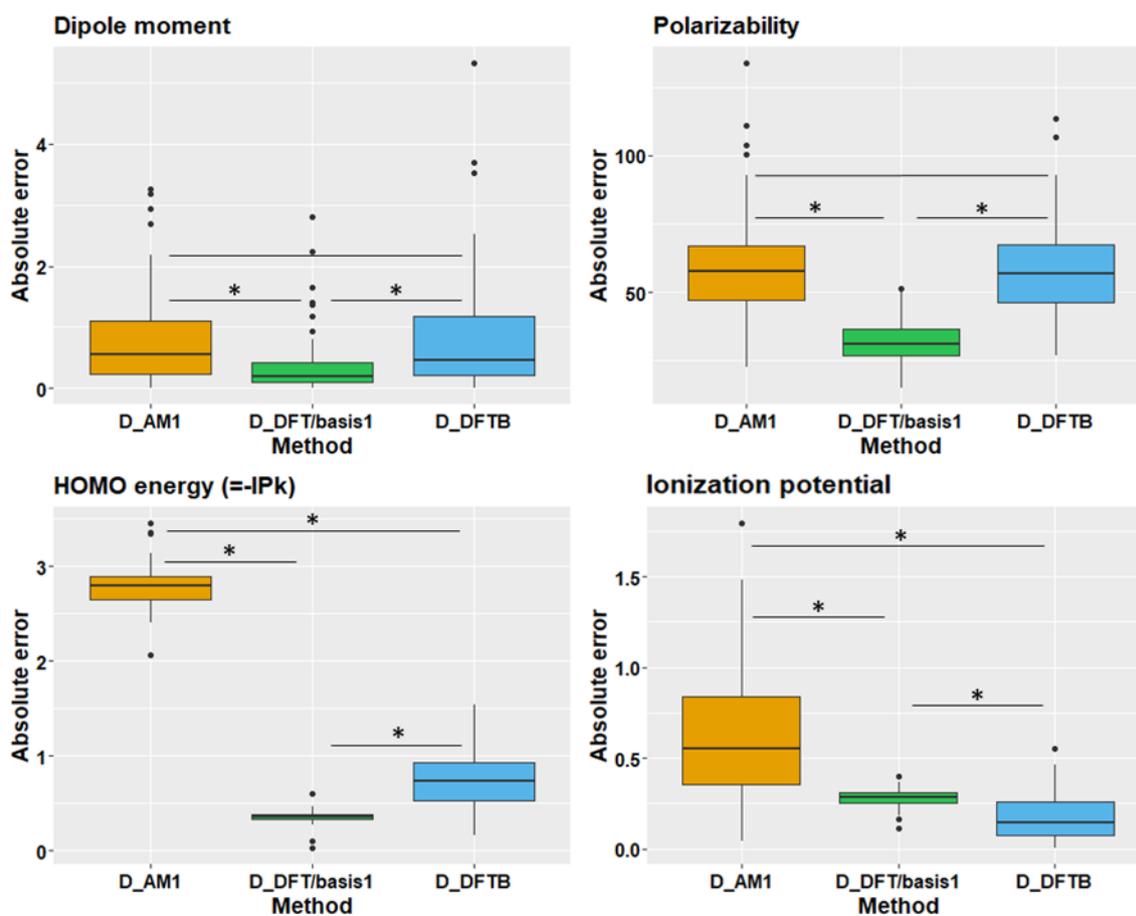
369 depending on the size of the electronic cloud, is quite unchanged when the conformation of the
370 molecule varies. This explains the small dispersion observed with D_DFT/basis2 vs. H_AM1. However,
371 as mentioned in the previous section, all methods significantly underestimate polarizabilities with
372 respect to the D_DFT/basis2 level (Figure 3). There was a clear underestimation of the polarizabilities
373 calculated at H_AM1 ($a=1.356$ and $b=0.655$) and to a lesser extent at D_DFT/basis1 ($a=1.130$ and $b=-$
374 6.707), compared to D_DFT/basis2 (Figure 3), which could be partly corrected by including atomic
375 polarizabilities as implemented in the MOPAC package^[8] but not in the Gaussian one. Up to now, TyPol
376 quantum descriptors were calculated at the H_AM1 level of theory with the MOPAC package.
377 Polarizabilities listed in TyPol are corrected and then accurate^a. To compare polarizabilities computed
378 with H_AM1 presently listed in TyPol to the other strategies, we calculated polarizabilities on H_AM1
379 structures via Gaussian package.

380 As explained in **Materials and Methods**, the ionization potential can be computed using the HOMO
381 energy ($= -IP_K$, Koopman's Theorem). The AM1 potential led to a large overestimation of IP_K (D_AM1
382 on Figure 2). A huge overestimation of IP_K was observed for H_AM1 ($a=0.667$ and $b=-0.204$, $R^2=0.60$,
383 Figure 3b) but an improved agreement was found for D_DFT/basis2 vs. D_DFT/basis1 ($a=0.922$ and $b=-$
384 0.797 , Figure 3a) leading also to a correlation coefficient close to one: $R^2=0.916$ (Figure 3). Indeed, the
385 HOMO energy depends on the conformation of the molecule, so the dispersion is larger with H_AM1
386 considering the whole set of molecules ($R^2=0.60$ D_DFT/basis2 vs. H_AM1) than considering only 29
387 molecules ($R^2=0.76$ D_DFT/basis2 vs. D_AM1, Figure 2). Finally, the HOMO energy computed with AM1
388 could not be used to compute properly the ionization potential. We suggest replacing IP_K computed at
389 the AM1 level by D_DFT/basis1 IP_K . Moreover, for a better accuracy of IP values, the indirect method
390 using the cation and the neutral species energies is recommended (see discussion below, **Results and**
391 **Discussion**).

392 As a conclusion, the H_AM1 strategy could lead to a reasonable calculation of polarizabilities if atomic
393 corrections were added. However, the determination of accurate dipole moments and IP_K implies to
394 make a more comprehensive conformational search through molecular dynamics (D_), and to use a
395 higher level of theory for the local optimizations such as DFT/basis1 or DFT/basis2.

396 In addition, to evidence the potential to be chosen for 3D quantum molecular descriptor calculation,
397 the absolute errors of the three different levels D_AM1, D_DFTB and D_DFT/basis1 were compared
398 with respect to the D_DFT/basis2 values for each quantum descriptors under investigation in this work
399 (Figure 4).

^aMOPAC^[51] manual specifies that "Polarizability volumes calculated using NDDO methods are too low by about 30%." "In 2004, the polarizability volume reported was modified by the use of additive corrections." "For normal organic compounds, the average error in polarizability should be less than 2%."
<http://openmopac.net/index.html>



400

401 Figure 4. For the following quantum descriptors: dipole moment (Debye), polarizability (Bohr³), HOMO
 402 energy (eV) and IP, ionization potential (eV), absolute error between D_AM1, D_DFTB and
 403 D_DFT/basis1 results and D_DFT/basis2 ones are plotted using boxplots. Only the set of molecules with
 404 indices **1, 2, 3** and **4** (100 molecules) was used to produce the graphs. Kruskal-Wallis and Dunnett test
 405 were performed to evaluate the difference between the methods (*: p-value < 0.001, hypothesis: the
 406 methods are significantly different). The indices are defined in **Results and Discussion**.

407

408 For the dipole moment, polarizability and HOMO energy, D_DFT/basis1 gave lower errors than D_AM1
 409 and D_DFTB. Moreover, those errors were statistically different (* in Figure 4). For the ionization
 410 potential, all levels gave errors statistically different from each other but D_DFTB gives the lowest
 411 mean absolute errors compared to D_DFT/basis2. However, one must keep in mind that, as seen in
 412 the previous section, D_DFT/basis1 has the advantage to systematically overestimate the ionization
 413 potential values compared to D_DFT/basis2, while respecting the IP ordering of the various molecules,
 414 and could be used in a QSAR scheme where it is mandatory to respect the trends. On the opposite,
 415 D_DFTB over- or underestimates those values depending on the molecule, preventing its use to
 416 correlate IP to macroscopic properties.

417 Moreover, for the molecules indexed **1**, for which the four levels of theory led to the same most stable
 418 structure (27.9% of the molecule set), computing the lowest-energy isomer (local quench plus
 419 frequency calculation) at AM1 is between 91 and 1436 times faster than at DFT/basis1, and between
 420 1204 and 34 425 times faster than at DFT/basis2. Local quench plus frequency calculations for the

421 lowest-energy isomer lasted between 2 and 73 seconds (~1 minute) at AM1, between 441 and 70 738
422 seconds (~20 hours) at DFT/basis1, and between 5 178 and 1 401 534 seconds (~16 days) at
423 DFT/basis2. Thus, a conformational search through molecular dynamics (D_) followed by local
424 quenches at DFT/basis1 was the best compromise to reach accuracy and a reasonable computational
425 cost considering the large size of the database.

426

427 **Temperature Effects**

428

429 As said in **Materials and Methods – Global Exploration**, for each compound of the dataset, the five
430 lowest-energy conformers were selected for local optimizations at each level of theory of this study,
431 in particular at D_DFT/basis1. In this section, we highlight the temperature effects at D_DFT/basis1
432 level on the three following quantum descriptors: dipole moment, polarizability, and HOMO energy.

433 First, for 17 compounds, only one isomer was found. For 19 compounds only two isomers were found;
434 for 16 compounds, 3 isomers were localized and for 25, 4. Finally, 5 isomers of low energies were found
435 for 27 compounds. Among these 27 compounds, the fifth lowest-energy isomer abundance using a
436 Boltzmann distribution at room temperature is at most 6.3% (diphenhydramine) and for 24 of these
437 compounds, this abundance is less than 2%.

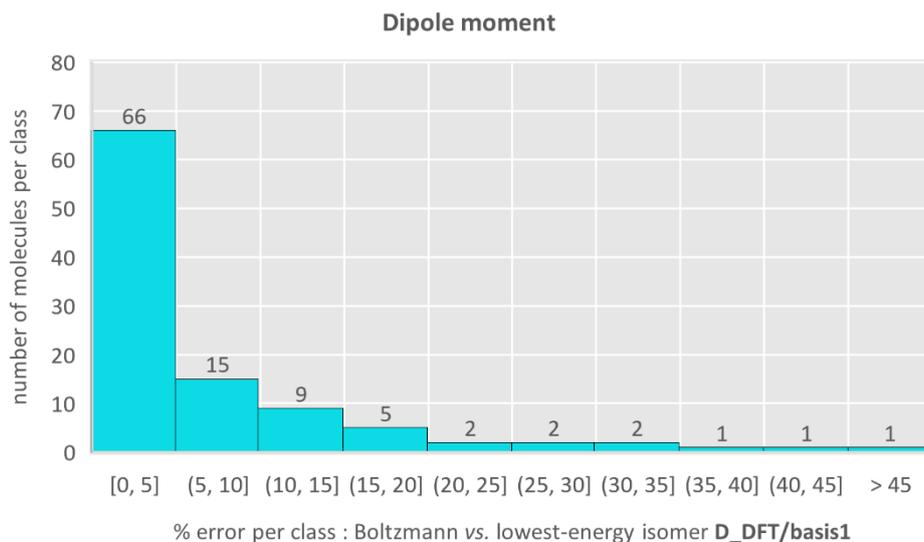
438 For the 104 dataset compounds, the polarizability and HOMO energy computed from the lowest-
439 energy isomer differ by less than 2% from those resulting from Boltzmann weighted values obtained
440 with the five lowest-energy structures. It can therefore be concluded that taking the Boltzmann
441 distribution into account is not essential in the calculation of these two quantum descriptors.

442 However, if we now look at the values of the dipole moment, the effects of temperature are quite
443 different (see Figure 5).

444

445

446



447

448 Figure 5. Percentage of errors by evaluating the dipole moment values using the lowest-energy isomer
 449 value instead of the Boltzmann averaged value (5 lowest-energy isomers) at room temperature at the
 450 D_DFT/basis1 level.

451

452 The difference between the lowest-energy value and the Boltzmann averaged value is less than 5 %
 453 for 66 compounds, between 5 and 10% for 15 compounds and between 10 and 20% for 14 compounds.
 454 Finally, for 9 compounds, the error is greater than 20%: tetrabromobisphenolA (69%), omeprazole
 455 (44%), estriol (36%), gemfibrozil (34%), 17-ethinylestradiol (30%), risperidone (29%), metoprolol (26%),
 456 oxytetracycline (24%), mestranol (23%), and olaquinox (20%). TetrabromobisphenolA has two
 457 isomers with Boltzmann equiprobabilities (54% and 46%) but as already mentioned (**Quantum**
 458 **Descriptor Analysis**), the lowest-energy isomer of tetrabromobisphenolA has a small $\mu = 0.533$ D (at
 459 D_DFT/basis1), whereas the second lowest-energy isomer exhibits a larger μ value 3.108 D (at
 460 D_DFT/basis1). Consequently, the Boltzmann weighted dipole moment value of 1.719 D is far from
 461 both isomer values. For the 8 other compounds, the observation is to a lesser extent the same: a wide
 462 range of variation of the dipole moment for all the conformers combined with non-negligible
 463 Boltzmann weights.

464 To correctly evaluate the dipole moment, Boltzmann distribution must be considered but to get correct
 465 polarizabilities and HOMO energies, the values obtained for the lowest-energy conformer is sufficient.
 466 As quantum descriptor calculations were performed for the five lowest-energy isomers, we suggest
 467 adding the Boltzmann average dipole moment in the Typol database.

468

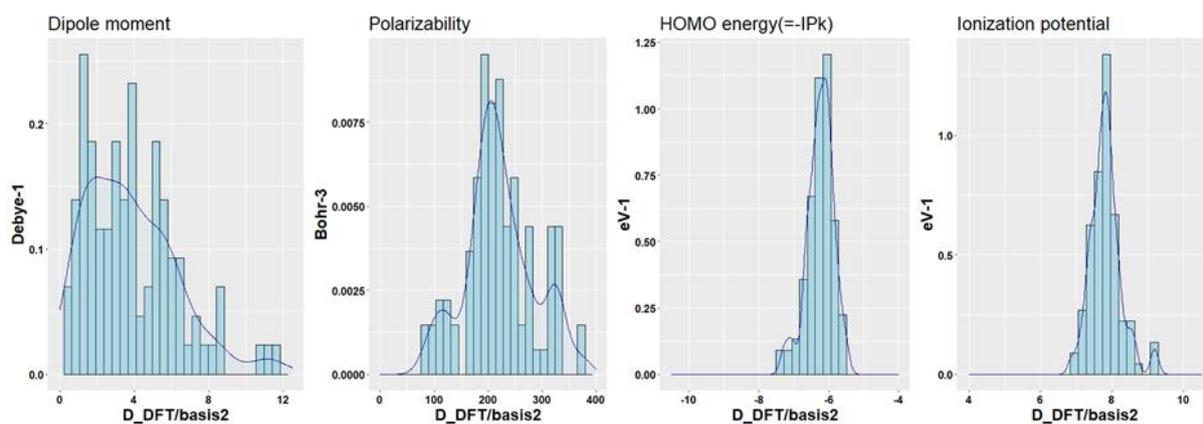
469 **Quantum Descriptor Analysis**

470

471 In this section, we seek to show that the information carried by the 3D quantum descriptors is not
 472 already contained in descriptors simpler to calculate, which do not require the determination of the
 473 lowest-energy conformer followed by quantum chemistry calculations. Thus, the distribution functions

474 of the 3D quantum descriptors computed with the D_DFT/basis2 strategy, are discussed regarding
475 non-quantum descriptors, *i.e.*, descriptive or based on 2D or 3D structural formula (Table SII). For
476 example, we will look for a link: i) between the dipole moment and the number of heteroatoms (nHet)
477 of the molecule; ii) between the polarizability and the molecular weight (MW); iii) between the
478 ionization potential and the number of heteroatoms (nHet) or the number of aromatic bonds (nAB).

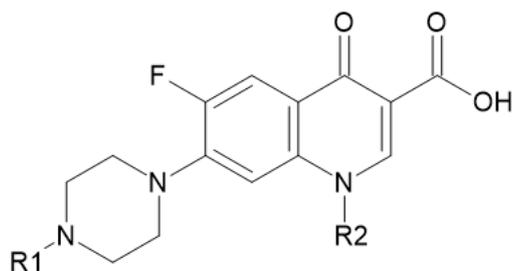
479 The dipole moment (μ) results from the vectorial linear combination of the bond and lone pair μ
480 present in the molecule. The greater the difference in the electronegativity between two atoms, the
481 more the bond is polarized and the greater the bond dipole moment (bond μ). Consequently, the
482 molecules containing heteroatoms, N, O, S, F, Cl, Br and I, which are more electronegative than C or H,
483 will have more polarized bonds and thus, non-zero bond μ . However, dipole moment values are not
484 completely determined by the number of heteroatoms in a molecule because of the tremendous
485 influence of the molecule conformation in the result of the vectorial linear combination of bond μ to
486 obtain the μ of the molecule. In the molecule set, μ goes from 0.253 D for estriol to 11.825 D for
487 norfloxacin (Figure 6).



489 Figure 6. Distribution functions of the following quantum descriptors calculated at the D_DFT/basis2
490 level of theory: dipole moment (Debye), polarizability (Bohr³), HOMO energy (= -IP_k) (eV) and IP,
491 ionization potential (eV). Only the set of molecules with indices **1, 2, 3** and **4** (100 molecules) was used
492 to produce the graphs. The indices are defined in **Results and Discussion**.

493

494 Fifty-seven molecules have a μ between 0.253 D and 3.8 D, the mean value for the dipole moment.
495 The μ values are not normally distributed (Figure 6). Consequently, in Table III, mean and median do
496 not correspond. Only seven molecules have a μ > 8.0 D and are part of the fluoroquinolones having
497 the 2D formula shown on Figure 7 and differing by R1 and R2 groups. Members of the fluoroquinolones
498 contain at least one fluorine atom, which is the most electronegative element of the whole periodic
499 classification.



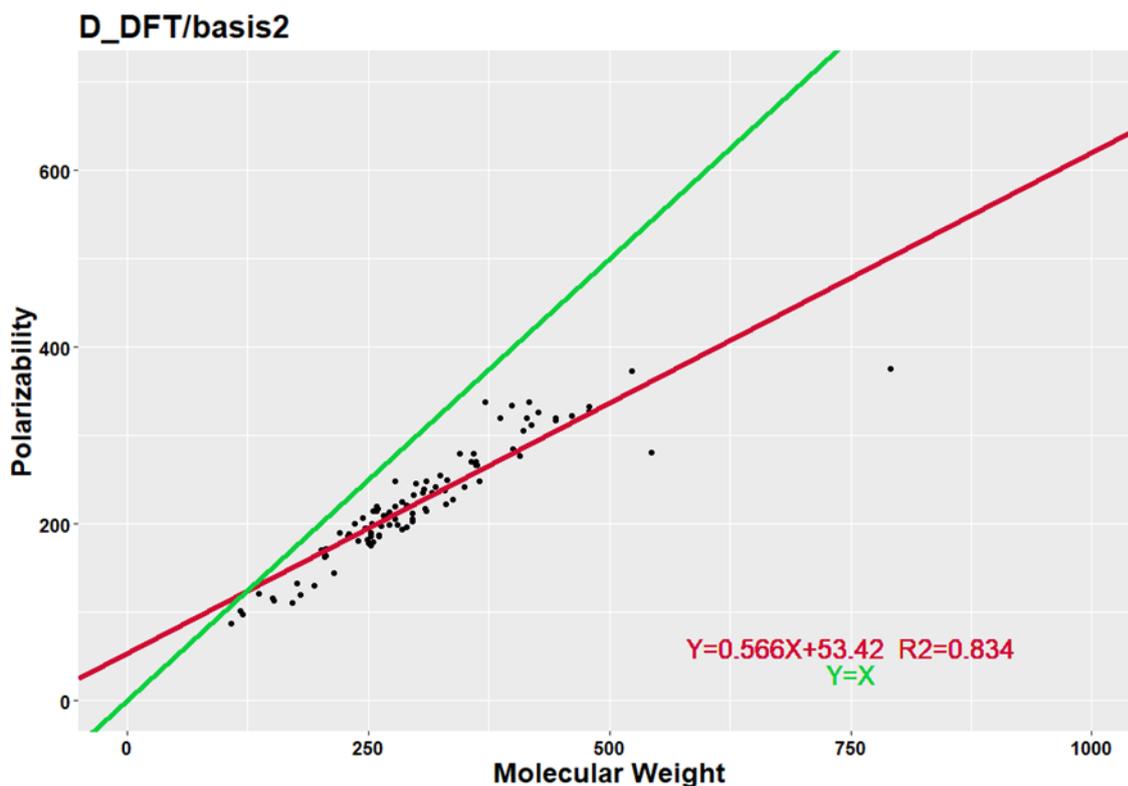
500

501 Figure 7. Fluoroquinolones 2D formula: the members of this family differ by R1 and R2 groups.

502

503 The great polarity of the C—F bond can partly explain the large values of μ in the fluoroquinolones. As
 504 expected, μ values are not completely determined by the number of heteroatoms (nHet), for instance,
 505 the lowest-energy isomer of tetrabromobisphenolA has a small μ of 0.757 D although it has four
 506 bromine and two oxygen atoms whereas the second lowest-energy isomer of this compound exhibits
 507 a larger μ value of 3.236 D as mentioned previously in **Temperature Effects**.

508 The polarizability (α) is the aptitude of the electronic cloud of a molecule to deform when an electric
 509 field is applied. It depends on the size of the electronic cloud and thus, on the number of electrons
 510 within the molecule. For an atom $\frac{A}{Z}X$, the number of electrons Z and the atomic weight A are clearly
 511 related. Consequently, for a molecule, α increases with the molecular weight (MW). Looking at the
 512 distribution function for α on Figure 6, ten molecules belong to the eleven lowest α values ($< 135 \text{ Bohr}^3$)
 513 and the eleven lowest MW values ($< 185 \text{ g/mol}$) and have less than twenty-seven atoms. At the same
 514 time, the twelve highest MW values ($> 414 \text{ g/mol}$) share nine molecules with the twelve highest α
 515 values ($> 319 \text{ Bohr}^3$). Both extreme values are found for the same molecules for α and MW: the minima
 516 are for ortho-cresol (108.15 g/mol, 86.44 Bohr³); the maxima for iopromide (791.15 g/mol, 375.11
 517 Bohr³). Moreover, looking at the molecules around the average values (av) of MW and α ($av \pm \frac{1}{3}\sigma$, σ
 518 the standard deviation), twenty-three are in common over thirty-two molecules for α and over thirty-
 519 one for MW. In fact, MW and α distribution functions are *quasi* stackable. Moreover, as polarizabilities
 520 are little influenced by the conformation of the molecules, polarizabilities were represented versus
 521 MW for the whole set of molecules (except index 0) (Figure 8).



522

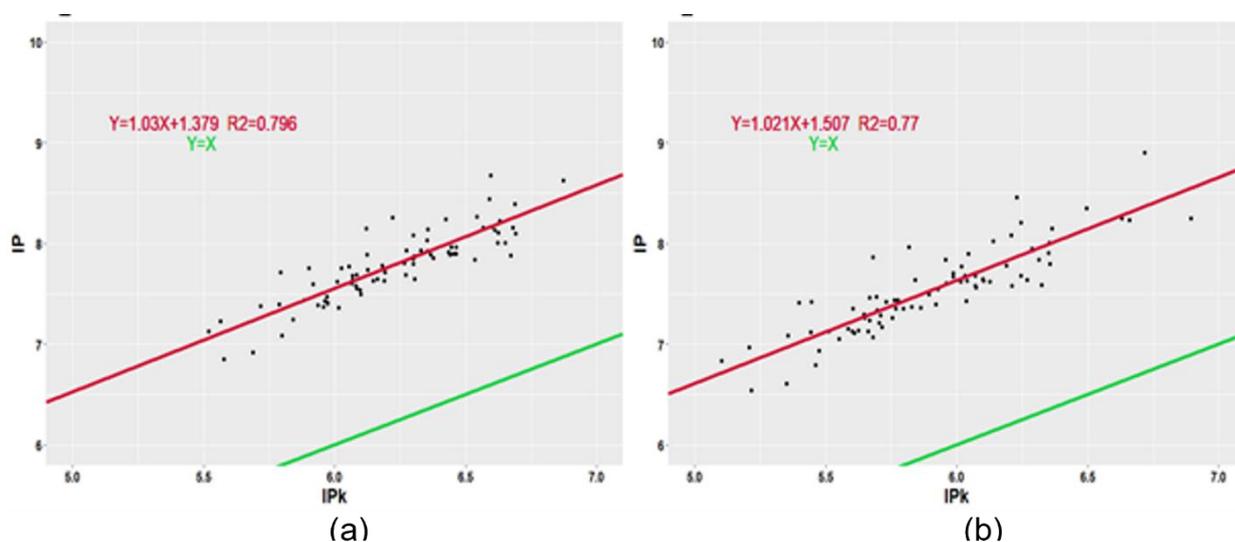
523 Figure 8. Polarizabilities (Bohr³) at D_DFT/basis2 level versus molecular weight (MW) values (g/mol).
 524 The equation, $y=ax+b$, of the linear regression line is written in red with the corresponding R²
 525 regression coefficient, while the green straight line represents the first bisector: $y=x$. Only the set of
 526 molecules with indices **1, 2, 3** and **4** (100 molecules) was used to produce the graphs. The indices are
 527 defined in **Results and Discussion**.

528

529 Linear regression leads to the following equation: α (Bohr³) = 0.566 MW (g/mol) + 53.42; with an
 530 acceptable dispersion ($R^2 = 0.834$). Two points were clearly more distant to the straight line: iopromide
 531 (791.15 g/mol, 375.11 Bohr³) and tetrabromobisphenolA (543.87 g/mol, 280.17 Bohr³). Iopromide is
 532 the only molecule of the set containing iodine and tetrabromobisphenolA, the only molecule
 533 containing bromine (Figure 8). Iodine and bromine are the heaviest elements present in the set of
 534 molecules under study. The greatest dispersion for both molecules could be explained by the quality
 535 of the basis set used for these two elements: only basis1 was used as basis2 was not available (α are
 536 then underestimated). Moreover, we did not consider scalar relativistic effects with core
 537 pseudopotentials, which could be harmful for these two elements. Anyway, knowing MW for a PPCP
 538 molecule, we can predict its polarizability value with an acceptable accuracy ($R^2 = 0.834$) *via* the
 539 equation: α (Bohr³) = 0.566 MW (g/mol) + 53.42.

540 For the ionization potential, we have two objectives here. The first one is to relate this potential to the
 541 descriptors nAB and nHet and to highlight that the information carried by the ionization potential is
 542 not already contained in these two descriptors. The second objective is to show that the calculation of
 543 the ionization potential is improved if the electronic reorganization is considered: IP *versus* IP_k (see
 544 **Materials and Methods**). The ionization potential is the minimal energy to extract one electron from
 545 a molecule. If IP differs from IP_k, IP gives a more reliable value as it takes electronic reorganization into

546 account. Figure 9 presents IP as a function of IP_K both computed at D_DFT/basis2 (a) or D_DFT/basis1
547 (b). For both linear regression lines, the slopes are close to one: $a=1.03$ (D_DFT/basis2) and $a=1.021$
548 (D_DFT/basis1).



550 Figure 9. At both D_DFT/basis2 (a) and D_DFT/basis1 (b) levels, straight lines in red are confronting IP
551 (eV) values on the vertical axis to IP_K (= - HOMO energy) (eV) values. On both graphs, the equation,
552 $y=ax+b$, of the linear regression line is written in red with the corresponding R^2 regression coefficient,
553 while the green straight line represents the first bisector: $y=x$. (a) Only the set of molecules with indices
554 **1, 2, 3** and **4** (100 molecules) was used to produce the graphs. (b) All the set of molecules was used to
555 produce the graphs: indices **0, 1, 2, 3** and **4** (104 molecules). The indices are defined in **Results and**
556 **Discussion**.

557

558 Thus, the lines are almost parallel to the first bisector. IP_K underestimates IP and shifting IP_K by $b=1.379$
559 eV (D_DFT/basis2) and by $b=1.507$ eV (D_DFT/basis1) gives almost the value of the corresponding IP.
560 Moreover, the dispersion observed ($R^2=0.796, 0.77$ respectively) indicates that the electronic
561 reorganization depends on the molecule under interest. The indirect method is then recommended to
562 obtain more accurate IP values. On Figure 6, both IP and HOMO energy distribution functions are
563 represented. Looking at the molecules around the average values (av) of IP and HOMO energy ($av \pm$
564 $\frac{1}{3}\sigma$, σ the standard deviation), 15 are in common over 36 molecules for IP, and over 25 for HOMO
565 energy. For the extreme values: (i) eight molecules belong to both the ten lowest HOMO energies and
566 the ten highest IP values; (ii) seven molecules belong to both the ten highest HOMO energies and the
567 ten lowest IP values. Overall, as similar patterns are observable for IP and IP_K (= -HOMO energy)
568 distribution functions, we will thus focus on the description of IP. For the 100 molecules under study,
569 IP values are going from 6.855 eV (tamoxifen) to 9.237 eV (metrodinazole) with an average value of
570 7.832 eV (median = 7.810 eV). The 36 molecules with IP values around the average, have IP values
571 between 7.693 (pyrimethamine) and 7.979 eV (iopromide). When the electron is extracted from the
572 HOMO, this orbital is either a heteroatom lone-pair or a π orbital. For instance, among the largest IP,
573 the electron is extracted from the π HOMO for aspirin, and from the carbonyl oxygen lone pair for
574 acetophenone. For metrodinazole, the HOMO of the neutral molecule is the hydroxyl oxygen lone pair
575 whereas, in the cation, the depopulated orbital is a π orbital delocalized over the imidazole cycle. The

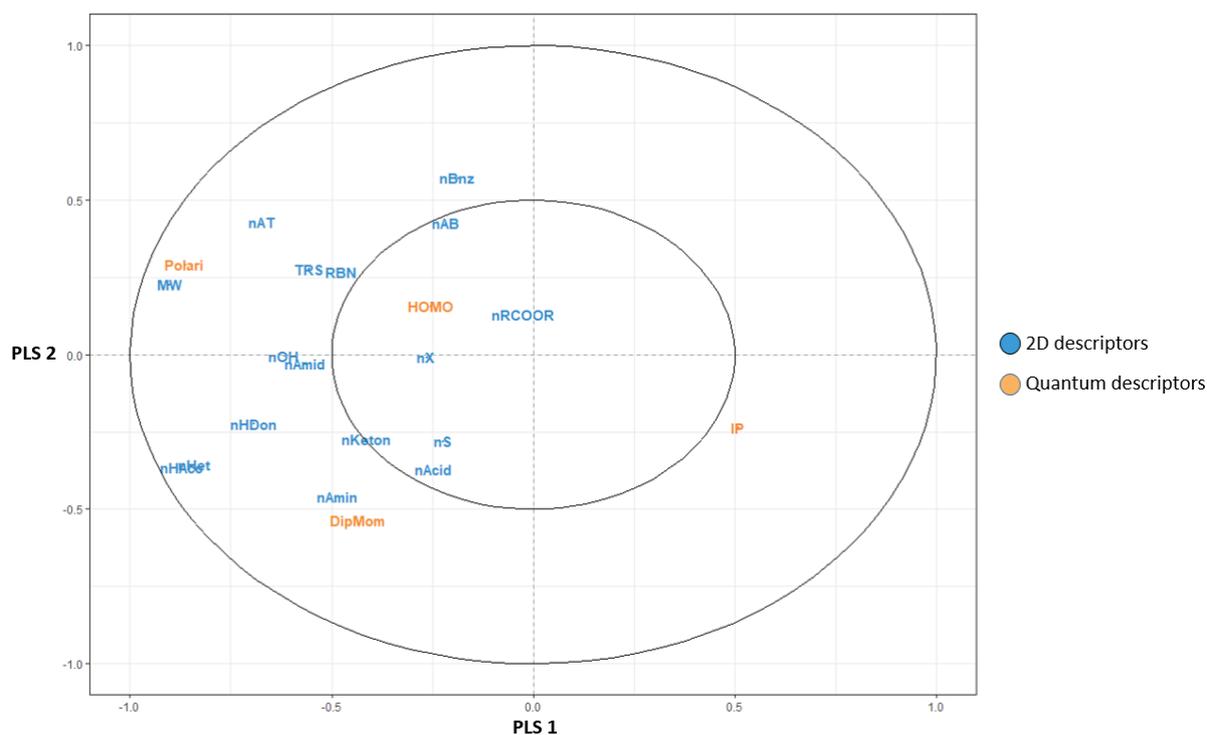
576 electronic reorganization is evidently important between the neutral and cationic forms, in this case,
577 Koopman's Theorem fails. Finally, we tried to relate IP and descriptive descriptors as nHet (number of
578 heteroatoms) or nAB (number of aromatic bonds) but no clear correlation was found.

579 In the following, we intend to connect quantum descriptors with non-quantum descriptors using a
580 partial least square regression (PLS) to evidence the possible correlations between both types of
581 descriptors. A PLS was performed considering the non-quantum descriptors not depending on the
582 conformation of the molecule as the predictive variables (X) (see Table SII) and the quantum
583 descriptors as predicted variables (Y). The non-quantum descriptors were obtained from the Dragon
584 software while the quantum descriptors used were calculated with the D_DFT/basis2 method. The
585 number of PLS components was four. Components 1 and 2 were the ones collecting the highest
586 explained variance. For the predictive variables (X), the explained variance on component 1 was 26.5%
587 and 10.4% for component 2. For the predicted variables (Y), the explained variance was 11.1% and
588 10.9% on components 1 and 2, respectively. The explained variance is low on both components for the
589 predictive and the predicted variables. Thus, the information provided by quantum descriptors is
590 poorly described by non-quantum descriptors. This fact constitutes a first proof of the importance of
591 considering quantum descriptors in the TyPol database to best describe the molecules and hope to
592 predict their environmental behavior.

593 On Figure 10, the correlation circle showed that the weight of the component 1 came from the
594 molecular weight (MW), the number of hydrogen acceptor (nHAcc), the number of heteroatoms
595 (nHet), the number of hydrogen donor (nHDon) and the number of hydroxyl groups (nOH). The weight
596 of component 2 is carried by the number of aromatic bonds (nAB) and the number of benzene type
597 rings (nBnz) (Figure 10). The Cluster Image Map (CIM) was also constructed from the PLS regression
598 data. The CIM allows to obtain the correlation level between the predictive and the predicted variables
599 according to the set of variables.

600

601



602

603 Figure 10. Circles of correlations of the non-quantum molecular descriptors (in blue) and '3D quantum
 604 descriptors' (in orange) variables on the two main components of the PLS (PLS1 and PLS2). HOMO
 605 stands for HOMO energy, IP for ionization potential, Polari for polarization and DipMom for dipole
 606 moment. See Table SII for the description of the non-quantum molecular descriptors.

607

608 The HOMO energy was the least explained quantum descriptor (Figure 10). The IP seems to be
 609 negatively correlated with both descriptors nAB and nBnz according to the CIM of component 2. IP is
 610 also negatively correlated to nHAcc, nHet and MW on the component 1 (Figure 10). The dipole
 611 moment, which tends to the lower left part of the correlation circle, is positively correlated with the
 612 amine number (nAmin) and ketone number (nKeton), and negatively correlated with nAB and nBnz.
 613 Finally, polarizability is the quantum descriptor that correlates best with the descriptors (Figure 10).
 614 This quantum descriptor is related to MW, nHAcc, nHet, Total Ring Size (TRS) and the number of
 615 rotatable bonds (RBN).

616 The strongest correlations established by the PLS between quantum and non-quantum descriptors had
 617 all been anticipated but remain very weak given the explained variances. Therefore, the information
 618 carried by the quantum descriptors (dipole moment, polarizability, HOMO energy and ionization
 619 potential) is not redundant with that carried by the non-quantum descriptive, 2D and 3D descriptors.
 620 The intrinsic properties of the molecule are better described when the quantum descriptors are
 621 included.

622

623 Conclusion

624 As a conclusion, the importance of doing a comprehensive conformational search through molecular
 625 dynamics was characterized. Investigating the quality and efficiency of various levels of theory (AM1,

626 DFTB and DFT) for computing quantum descriptors, namely dipole moment, polarizability, ionization
627 potential and HOMO energy, we propose a strategy, the so-called D_DFT/basis1, as a good
628 compromise between accuracy and computational cost. It consists in an exhaustive MD
629 conformational search followed by a local optimization at B3LYP/6-31G*. Moreover, we showed that
630 dipole moment values are not completely determined by the number of heteroatoms in a molecule,
631 indeed, the conformation of the molecule has a tremendous influence on the dipole moment value.
632 On the contrary, knowing the molecular weight of a molecule allows to predict with an acceptable
633 accuracy its polarizability. At the same time, ionization potential values are not correlated with the
634 number of heteroatoms or the number of aromatic bonds in the compound. Finally, information bared
635 by quantum descriptors is not redundant with the one bared only by the non-quantum descriptive, 2D
636 and 3D descriptors. Therefore, quantum descriptors (dipole moment, polarizability, HOMO energy,
637 ionization potential) should be considered into TyPol database. At present, the quantum descriptors
638 calculated at the D_DFT/basis1 level are being calculated for all 500 organic contaminants in the TyPol
639 database. We showed that averaging the quantum descriptors values obtained for the lowest-energy
640 isomers with a Boltzmann distribution provides values very close to those of the most stable isomer
641 for the polarizabilities and the HOMO energies. On the opposite, the dipole moments can be strongly
642 affected by the introduction of higher energy isomers contributions. Once the database has been
643 updated, we will be able to test the consequences of these improvements on the classifications already
644 published,^[52] but above all to consider new and more ambitious projects. However, the improvement
645 of the base is continuous. New descriptors can be integrated describing, for example, solvent effects:
646 a first step towards quantum descriptors directly linked to the behavior of a contaminant in the
647 environment.

648

649 **Supporting Information Summary**

650 More details on the computational chemistry methods are given for DFT and DFTB (1.). Similarly,
651 further information about ionization potential calculations is presented (2.c.). The molecule set of 104
652 PPCP is separated into 5 groups further details about those groups: indices, exclusivity, percentages;
653 are given in Table SI. The list of the 20 non-quantum descriptors from Dragon software used in this
654 work are gathered in Table SII. A comparison of 3 quantum-based 3D descriptors calculated at
655 D_DFT/basis1 (chosen strategy) and at H_AM1 (presently in the TyPol database) is represented on
656 Figure S1. An archive file is provided including the most stable isomer structures obtained with the
657 chosen strategy, D_DFT/basis1.

658

659 **Acknowledgement**

660 This work was granted access to the HPC resources of CALMIP supercomputing center under the
661 allocation 2016-[P1222].

662 **Funding Information**

663 We would like to thank ADEME (French Agency for Ecological Transition) and INRAE (French National
664 Research Institute for agriculture, food, and environment) for the PhD grant of Kevin Bonnot and OFB
665 (French Office for Biodiversity) for its financial and scientific supports.

666 **Keywords**
667 Database
668 DFT
669 DFTB
670 Molecular descriptors
671 QSAR
672

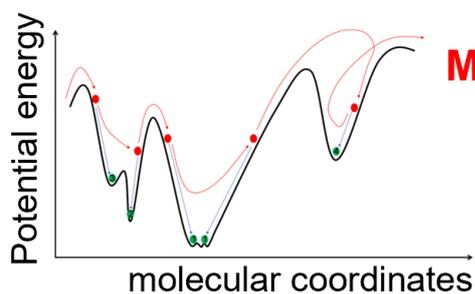
673 **References**

- 674 [1] C. G. Daughton, T. A. Ternes, *Environmental Health Perspectives* **1999**, *107*, 907–938.
- 675 [2] J. Kagle, A. W. Porter, R. W. Murdoch, G. Rivera-Cancel, A. G. Hay, in *Advances in Applied*
676 *Microbiology*, Elsevier, **2009**, pp. 65–108.
- 677 [3] D. C. G. Muir, P. H. Howard, *Environ. Sci. Technol.* **2006**, *40*, 7157–7166.
- 678 [4] J. C. Dearden, P. Rotureau, G. Fayet, *SAR and QSAR in Environmental Research* **2013**, *24*, 279–
679 318.
- 680 [5] G. Fayet, P. Rotureau, L. Joubert, C. Adamo, *Journal of Molecular Modeling* **2010**, *16*, 805–812.
- 681 [6] L. Mamy, D. Patureau, E. Barriuso, C. Bedos, F. Bessac, X. Louchart, F. Martin-Laurent, C. Miede,
682 P. Benoit, *Critical Reviews in Environmental Science and Technology* **2015**, *45*, 1277–1377.
- 683 [7] L. Mamy, E. Barriuso, B. Gabrielle, *Pest Management Science* **2005**, *61*, 905–916.
- 684 [8] R. Servien, L. Mamy, Z. Li, V. Rossard, E. Latrille, F. Bessac, D. Patureau, P. Benoit, *Chemosphere*
685 **2014**, *111*, 613–622.
- 686 [9] *Dragon 7.0, Software for Molecular Descriptor Calculation.*, Talete S.R.L., **2017**.
- 687 [10] M. Karelson, V. S. Lobanov, A. R. Katritzky, *Chem. Rev.* **1996**, *96*, 1027–1044.
- 688 [11] A. Soyemi, T. Szilvási, *J. Phys. Chem. A* **2022**, *126*, 1905–1921.
- 689 [12] D. Hait, M. Head-Gordon, *Phys. Chem. Chem. Phys.* **2018**, *20*, 19800–19810.
- 690 [13] D. Hait, M. Head-Gordon, *J. Chem. Theory Comput.* **2018**, *14*, 1969–1981.
- 691 [14] A. R. Katritzky, S. Perumal, R. Petrukhin, E. Kleinpeter, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 569–
692 574.
- 693 [15] A. R. Katritzky, L. Pacureanu, D. Dobchev, M. Karelson, *J Mol Model* **2007**, *13*, 951–963.
- 694 [16] P. Hohenberg, W. Kohn, *Phys. Rev.* **1964**, *136*, B864–B871.
- 695 [17] W. Kohn, L. J. Sham, *Phys. Rev.* **1965**, *140*, A1133–A1138.
- 696 [18] C. Lee, W. Yang, R. G. Parr, *Phys. Rev. B* **1988**, *37*, 785–789.
- 697 [19] A. D. Becke, *The Journal of Chemical Physics* **1993**, *98*, 5648–5652.
- 698 [20] M. M. Francl, W. J. Pietro, W. J. Hehre, J. S. Binkley, M. S. Gordon, D. J. DeFrees, J. A. Pople, *The*
699 *Journal of Chemical Physics* **1982**, *77*, 3654–3665.
- 700 [21] V. A. Rassolov, M. A. Ratner, J. A. Pople, P. C. Redfern, L. A. Curtiss, *J. Comput. Chem.* **2001**, *22*,
701 976–984.
- 702 [22] M. N. Glukhovtsev, A. Pross, M. P. McGrath, L. Radom, *The Journal of Chemical Physics* **1995**, *103*,
703 1878–1885.
- 704 [23] A. D. McLean, G. S. Chandler, *The Journal of Chemical Physics* **1980**, *72*, 5639–5648.
- 705 [24] M. P. McGrath, L. Radom, *The Journal of Chemical Physics* **1991**, *94*, 511–516.
- 706 [25] L. A. Curtiss, M. P. McGrath, J. Blaudeau, N. E. Davis, R. C. Binning, L. Radom, *The Journal of*
707 *Chemical Physics* **1995**, *103*, 6104–6113.
- 708 [26] F. Bessac, S. Hoyau, *Computational and Theoretical Chemistry* **2011**, *966*, 284–298.
- 709 [27] F. Bessac, S. Hoyau, *Computational and Theoretical Chemistry* **2013**, *1022*, 6–13.
- 710 [28] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani,
711 V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. Marenich, J. Bloino, B. G. Janesko,
712 R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-
713 Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe,
714 V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J.
715 Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A.
716 Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N.
717 Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S.
718 Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L.
719 Martin, K. Morokuma, O. Farkas, J. B. Foresman, D. J. Fox, **n.d.**
- 720 [29] D. Porezag, T. Frauenheim, T. Köhler, G. Seifert, R. Kaschner, *Phys. Rev. B* **1995**, *51*, 12947–12957.
- 721 [30] G. Seifert, D. Porezag, T. Frauenheim, *Int. J. Quantum Chem.* **1996**, *58*, 185–192.
- 722 [31] M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai, G. Seifert, *Phys.*
723 *Rev. B* **1998**, *58*, 7260–7268.

- 724 [32] Y. Yang, H. Yu, D. Uork, Q. Cui, M. Elstner, *J. Phys. Chem. A* **2007**, *111*, 10861–10873.
725 [33] M. Gaus, Q. Cui, M. Elstner, *Journal of Chemical Theory and Computation* **2011**, *7*, 931–948.
726 [34] M. Gaus, A. Goetz, M. Elstner, *Journal of Chemical Theory and Computation* **2013**, *9*, 338–354.
727 [35] L. Zhechkov, T. Heine, S. Patchkovskii, G. Seifert, H. A. Duarte, *J. Chem. Theory Comput.* **2005**, *1*,
728 841–847.
729 [36] H. A. Witek, K. Morokuma, A. Stradomska, *The Journal of Chemical Physics* **2021**, *121*, 5171–5178.
730 [37] T. Heine, M. Rapacioli, S. Patchkovskii, J. Frenzel, A. Koster, P. Calaminici, H. A. Duarte, S.
731 Escalante, R. Flores-Moreno, A. Goursot, J. Reveles, D. Salahub, A. Vela, *deMonNano*,
732 <http://demon-nano.ups-tlse.fr/> **2009**.
733 [38] M. J. S. Dewar, W. Thiel, *J. Am. Chem. Soc.* **1977**, *99*, 4899–4907.
734 [39] M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, J. J. P. Stewart, *J. Am. Chem. Soc.* **1985**, *107*, 3902–
735 3909.
736 [40] A. Banerjee, N. Adams, J. Simons, R. Shepard, *J. Phys. Chem.* **1985**, *89*, 52–57.
737 [41] J. Baker, *J. Comput. Chem.* **1986**, *7*, 385–395.
738 [42] P. Culot, G. Dive, V. H. Nguyen, J. M. Ghuysen, *Theoret. Chim. Acta* **1992**, *82*, 189–205.
739 [43] X. Li, M. J. Frisch, *J. Chem. Theory Comput.* **2006**, *2*, 835–839.
740 [44] T. Koopmans, *Physica* **1934**, *1*, 104–113.
741 [45] D. Danovich, *Journal of Molecular Structure: THEOCHEM* **1997**, *401*, 235–252.
742 [46] P. Politzer, F. Abu-Awwad, *Theoretical Chemistry Accounts: Theory, Computation, and Modeling*
743 *(Theoretica Chimica Acta)* **1998**, *99*, 83–87.
744 [47] S. Hamel, P. Duffy, M. E. Casida, D. R. Salahub, *Journal of Electron Spectroscopy and Related*
745 *Phenomena* **2002**, *123*, 345–363.
746 [48] R. Vargas, J. Garza, A. Cedillo, *J. Phys. Chem. A* **2005**, *109*, 8880–8892.
747 [49] J. Luo, Z. Q. Xue, W. M. Liu, J. L. Wu, Z. Q. Yang, *J. Phys. Chem. A* **2006**, *110*, 12005–12009.
748 [50] T. Tsuneda, J.-W. Song, S. Suzuki, K. Hirao, *The Journal of Chemical Physics* **2010**, *133*, 174101.
749 [51] James J. P. Stewart, Stewart, *MOPAC2009 Computational Chemistry*, Colorado Springs, CO, USA,
750 **2009**.
751 [52] P. Benoit, L. Mamy, R. Servien, Z. Li, E. Latrille, V. Rossard, F. Bessac, D. Patureau, F. Martin-
752 Laurent, *Science of The Total Environment* **2017**, *574*, 781–795.
753

754

755 ToC



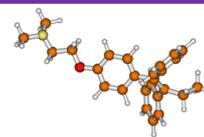
Molecular Dynamics

Local Quenches

Properties

μ , α , E_{HOMO} , IP

104 PPCP



ex. Tamoxifen

756

757 For 104 pharmaceuticals and personal care products (PPCP), four quantum-based 3D descriptors
758 (electric dipole moment, polarizability, HOMO energy and ionization potential) were calculated using
759 different computational chemistry strategies combining a molecular dynamics global exploration with
760 local quenches within different frameworks (semi-empirical, DFTB, DFT).