Supporting Information

Comparisons of Molecular Structures Generation Methods Based on Fragments Assemblies and Genetic Graphs.

Philippe Gantzer, Benoit Creton, Carlos Nieto-Draghi*

IFP Energies nouvelles, 1 et 4 avenue de Bois-Préau, 92852 Rueil-Malmaison, France.

* benoit.creton@ifpen.fr

Representation of the chemical space by $\mathbb{C}$

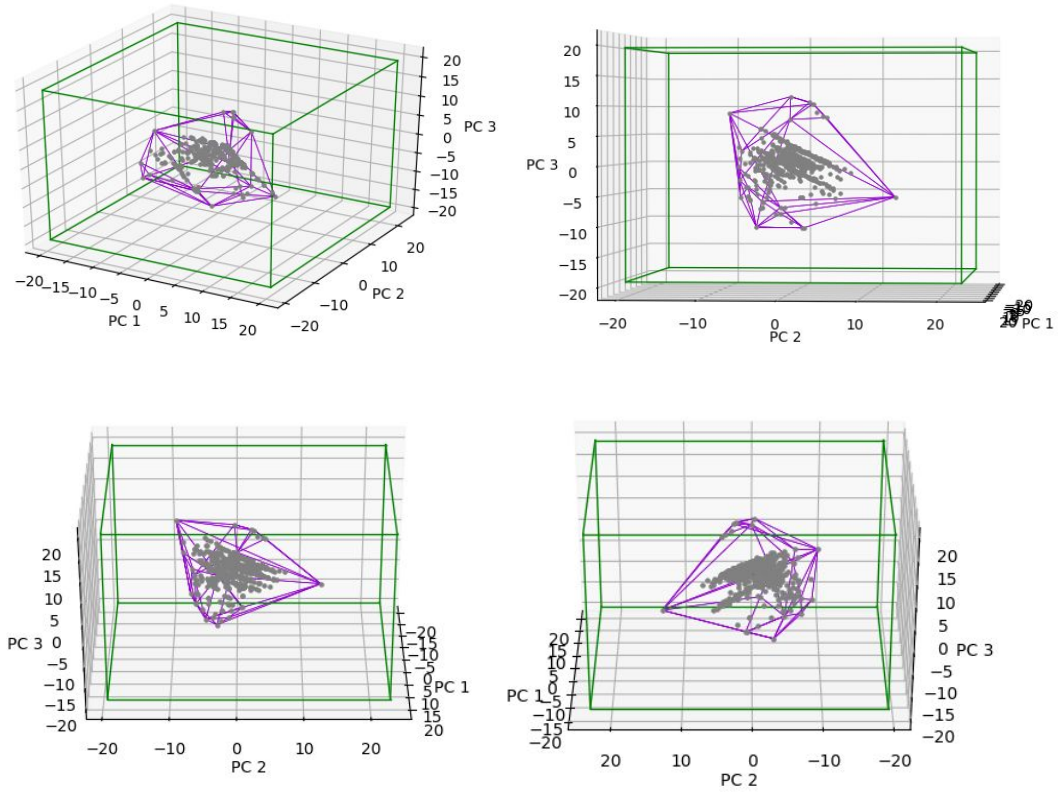

Figure S1: Additional representations of $\mathbb{C}$, with projected initial data points (grey dots), the initial chemical space (purple convex hull) and the extended chemical space (green parallelepiped rectangle).

Table S1: Limits of the initial and extended chemical spaces.

Axis	Initial chemical space		Extended chemical space	
	Minimum	Maximum	Minimum	Maximum
PC 1	-11.8	12.4	-19.8	20.4
PC 2	-9.2	13.0	-19.2	23.0
PC 3	-11.2	11.7	-19.2	19.7

Molecular diversity additional indices definition:

I_{2b} is defined as one minus the total variation distance, *i.e.* the maximal difference between P_x^m and P_x^T values over all x unit cubes (Equation (S1)). I_{2b} informs about the biggest difference of occupancy rates within a cube, between molecules issued from a specific generation method and molecules issued from all generation methods.

$$I_{2b} = 1 - \max_x \left| P_x^m - P_x^T \right| \quad \text{(S1)}$$

I_{2c} is defined as one minus the squared Euclidean distance. The Euclidean distance is the sum of the squared P_x^m and P_x^T values differences over all x unit cubes (Equation (S2)). I_{2c} considers all occupancy rates differences although I_{2b} only informs about the biggest occupancy rate difference.

$$I_{2c} = 1 - \sum_{x=0}^{n} (P_x^m - P_x^T)^2 \quad \text{(S2)}$$

I_{2d} is defined as one minus the M^{th} root of the ratio between the square of the Euclidean distance and the sum of the square of the global occupancy rates of each unit cube, weighted by the inverse of the number of occupied cubes by all generated molecules C, where M is the number of compared methods (Equation (S3)). The $1/C$ coefficient normalize I_5 values between 0 and 1. The use of the M^{th} root avoids a fast convergence to values close to zero.

$$I_{2d} = 1 - \sqrt[M]{\frac{\sum_x^n \left(P_x^m - P_x^T\right)^2}{\sum_x^n \left(P_x^T\right)^2} * \frac{1}{C}} \tag{S3}$$

Molecular diversity additional indices' use to compare generations of diverse molecules:

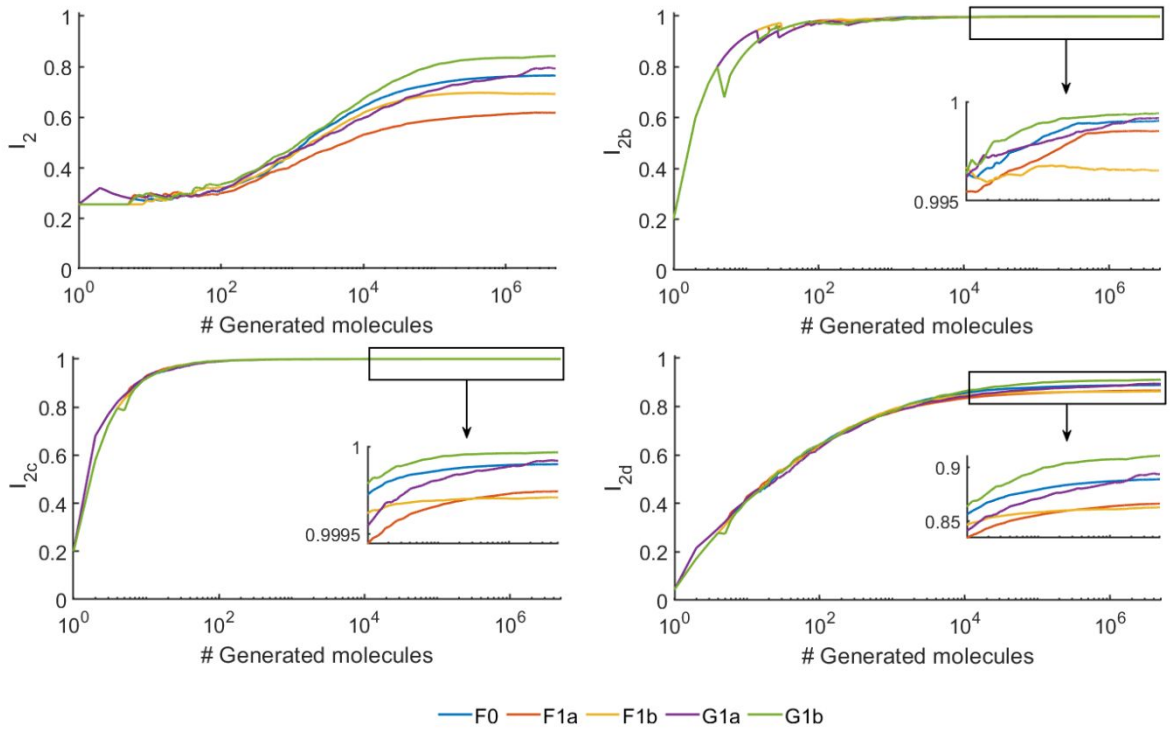

Figure S2: Evolution of AD representativeness's indices with the number of generated molecules, for each generation method.

Variations of AD representativeness's indices with the number of generated structures can be observed on Figure S2. At the beginning of the generations, indices' values are at their minimum, meaning an important deviation between distributions of generated molecules by each method and the distribution of molecules generated by all methods. Then, indices' values increase with the number of generated structures, showing that each method started to generate structures projected in a more similar way in $\mathbb{C}$. Plateau values are observed from 10^5 generated

structures. Values of AD representativeness's indices at five million generated structures are reported on **Table S2**.

Table S2: Calculated indices after generation of five million molecules with each method.

Method	I_2	I_{2b}	I_{2c}	I_{2d}
F0	0.77	0.9990	0.99990	0.89
F1a	0.62	0.9985	0.99974	0.87
F1b	0.69	0.9965	0.99971	0.86
G1a	0.79	0.9992	0.99992	0.89
G1b	0.84	0.9994	0.99997	0.91

I_{2b}, I_{2c} and I_{2d} rank the generation methods according to AD representativeness as follows: G1b>G1a>F0>F1a>F1b. F1a and F1b are ranked differently by I_2 than by I_{2b}, I_{2c} and I_{2d} since I_2 uses a difference of square rooted values whilst other indices use a squared difference of

value; I_2 gives then more importance to small differences of values between occupancy and global occupancy rates than other indices.

Among the developed AD representativeness indices, we selected I_2 to be computed when targeting a property range, due to its abilities to output more diffuse values between methods in our study and to give more importance to small differences of values between occupancy and global occupancy rates than other indices.

QSPR model accuracy inside the space $\mathbb{C}$:

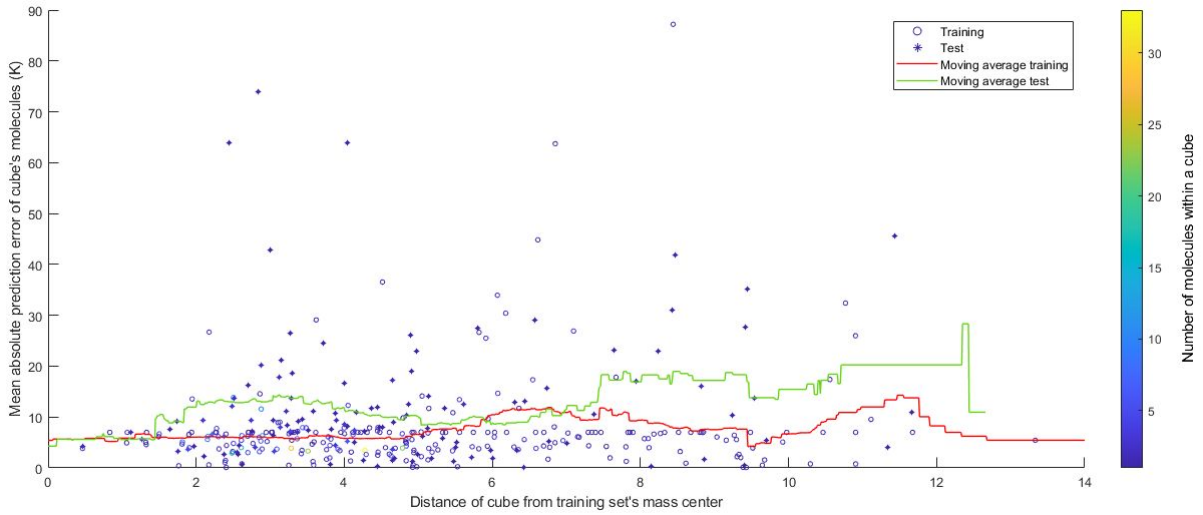

Figure S3: Evolution against the distance to the dataset mass centre of the absolute difference between experimental and predicted FP value using the QSPR model, for training and test sets' molecules projected in $\mathbb{C}$.