# Efficient exploration of high-dimensional configuration spaces for the generation of chemical datasets

Xuewen Xiao,[†] Abdul-Rahman Allouche,[¶] Emmanuel Dartois,[†] Frédéric Magoulès,[‡] and Daniel Peláez[*,†]

[†]*Université Paris-Saclay, CNRS, Institut des Sciences Moléculaires d'Orsay, 91405, Orsay, France.*

[‡]*MICS, CentraleSupélec, Paris-Saclay University, Gif-sur-Yvette, France.*

[¶]*Institut Lumière Matiére, CNRS, Université Claude Bernard, Lyon, France.*

[§]*Faculty of Engineering and Information Technology, University of Pécs, Pécs, Hungary.*

E-mail: daniel.pelaez-ruiz@universite-paris-saclay.fr

**Abstract**

In this work, we introduce an automated methodology for the efficient and relatively inexpensive exploration of large high-dimensional chemical spaces, with particular focus on number-of-atoms conserving processes, such as in mechanochemical reactions. Our approach combines: (1) a physically-motivated stochastic global-landscape exploration phase (mechanochemical distortion) which efficiently overcomes entropic barriers and (2) a local exploration phase in which the previously determined local basins are sampled with molecular dynamics and graph-theory. Specifically, this last phase makes use of the (vdW-) Transition State Search Using Chemical Dynamical Simulations algorithm. Our methodology requires a minimum input from the user. As a case study, we have

explored the conformational landscape, including transition states and minimum energy paths, of the $C_{60}H_{10}$ hydrogen-carbon clusters owing to their astrochemical relevance as potential carriers of the Aromatic Infrared Bands. From a single initial seed (geometry), we have obtained a series of 212 mechanochemically-relevant conformers and from just 3 of them, we have obtained a set of >13000 minima spanning the domain of our interest. The underlying chemical network has been fully characterized and rationalized using statistical analysis tools. Our case study perfectly illustrates the potential of our approach in the automatic generation of chemical databases, in other words, annotated data for the training of data-hungry Deep Learning models in Chemistry.

# 1 Introduction

Artificial Intelligence (AI), at least in its major flavor, supervised learning, has permeated all layers of society, technology, and science. Its widespread use is concomitant to a boost in investment,[1] which, in turn, reflects two intertwined factors: on the one hand, the ubiquitousness current and foreseen AI-impacted applications, (e.g. internet-of-things, industry 4.0); and the availability of vast amounts of data as well as the plethora of tools, methods, and associated software to handle them, collectively referred to as Big Data[2]. In turn, all these realizations have been synergistically enabled by the different AI paradigms and associated architectures together with major technological developments both in software and hardware that have permitted their efficient implementations.[3] This reality has been globally acknowledged by the award of the last Nobel Prizes in Physics and Chemistry on AI related topics.[4,5]

In the context of Chemistry, the supervised training of accurate AI agents relies on the possibility of obtaining a sufficiently large set of *clean* (and reliable) reference data[6]. It is widely acknowledged that homogeneous molecular datasets for AI duties are scarce, and its gathering is a computationally demanding task which, oftentimes, is not automated.[7] Hence for most of the AI applications, training/validation data are extracted from pre-existing

databases (DB).[8] This circumstance reflects the existing imbalance between the availability of tools for constructing AI models and the counterparts in DB generation.

Methodologies for the generation of *clean* reference datasets should fulfill several conditions. They should be able to produce efficiently large amounts of data in a consistent manner, being reproducible and meaningful in a statistical sense. The distribution of data should be representative of the target distribution of feature values. Second, the data should remain as close as possible to the *ground truth*. In the case of (theoretical) applications in Chemistry, ground truth refers to the highest- possible level of electronic structure theory. In this way, one minimizes *garbage-in, garbage-out* issues.[9] Third, our (tedious) selection process should be as automatized as possible, no-human-in-the-loop, in order to reduce human bias while at the same time retaining the target physics/chemistry of the problem in hand.

In this work, we present a quasi-automated methodology (i.e. grey-box) for the extraction of statistically meaningful sets of data-points from *huge*-dimensional chemical spaces. Note that, in spite of its high degree of automatization, we have consciously avoided the term black-box, since it is hardly attainable in real life applications. Our approach overcomes all of the major issues discussed above except the high-level of electronic structure theory which is, so far, an independent problem. Promising synergistic mixed AI/quantum mechanical approaches exist[10] and it is just a matter of time before the efficient combination of such powerful solutions with automated methods such as ours arises. Finally, the present study relies on the so-called adiabatic approximation, that is, the total neglect of all non-adiabatic terms.[11] One should be aware of the fact that such an approximation, common to many automated search methods, will imply that a fraction of the raw data generated might lack of physico-chemical meaning and will most likely require a multiconfigurational description.[12,13] In spite of this limitation, automated global search method such as the one used in this work constitute an excellent means of assessing the quality and reliability of newly developed DFT

functionals[14] or reparametrized semiempirical Hamiltonians[15,16] by simple comparison of the associated reaction network(RXN) with that of a reference method.[17]

One of the novelties of our approach is its focus on mechanochemical reactions.[18,19] It should be clear from the start that our goal is not simulating time-dependent processes. Conversely, our algorithm samples the configuration space available to a target system, under mechanochemical conditions. This, in turn, is a key aspect in circumventing the, arguably, major issue in automated molecular dynamics (MD) based exploration: the entropic barriers. Thanks to our mechanochemical step, we overcome this in a physically-sound manner.

From an algorithmic perspective, following the classification proposed by Reiher and coworkers,[20,21] we would ascribe our method (denoted by Mech-AMK, this name will be explained below), to the (unimolecular) forward-exploration-with-open-end class. This means Mech-AMK will try to characterize *all* possible chemical transformations stemming from a *single* initial geometry. In other words, for a given a molecular system of empirical formula $A_l B_m C_n \ldots$ the total number of atoms $N = l + m + n + \ldots$ will remain constant. Hence, Mech-AMK provides a statistical distribution of chemically-meaningful rearrangements of the target system. From a graph-theoretical perspective, where nodes are minima and transition states edges, this would result in a ever growing sparsely connected network of structures. The same group has also proposed three classes of exploration algorithms:[20] (i) those aiming at a complete exploration of a chemical space (automated, total breadth, total depth); (ii) those based on some hopping plus chemical heuristics (automated, trade-off of breadth and depth) and (iii) those mainly controlled by human/chemical intuition (partially automated, selective breadth and depth). Mech-AMK has a large overlap with the second category, since it combines a physically-oriented stochastic move (mechanochemical reaction) with an MD based exploration of a basin. However, for this last step, we will make use of our previous (human) experience to favour the in-depth exploration of certain regions of interest.[22]

Arguably, the most relevant and representative information contained on any chemical space is carried by the stationary points, minimum energy paths connecting them, and dissociation channels, associated to a chemical system.[15] In this respect, it is estimated that the number of minima grows exponentially with the number of atoms $N_{\text{min}} \propto N \ \exp(\xi N)$ with $\xi \approx 1$, as empirically shown for the case of small atomic clusters.[23] Consequently, the potential energy landscapes associated to the chemical spaces of interest are characterized by large numbers of basins and associated to families of structurally-related local minima (see Figure TOC Graphic). These basins correspond to major distinct regions within a given descriptor space, connected among them through entropy controlled paths. Conversely, the passage among local minima is mainly controlled by energy barriers. Logically, as early recognized by Wales and Doye,[23,24] to tackle this problem one has to rely on a combination of global (stochastical) and local (possibly deterministic) searches. In our specific case, we combine a global stochastic, physically-oriented, mechanochemical sampling strategy together with an efficient and (more) local sampling technique relying on the powerful Transition State Search using Chemical Dynamics Simulations (TSSCDS)[25,26] in any of its flavours: covalently bound systems or non-covalently bound ones.[17,27] Consequently, hereafter, we will refer to our approach as Mech-AMK, indicating the combination of the stochastic mechanochemical (global) and AutoMekin (AMK) local search. One of the major achievements of Mech-AMK is an almost complete automatization of all these tasks. This is essential for the generation of annotated data in the context of supervised learning.

To illustrate the capabilities of our approach, we have considered as case-study the transformation network associated to a medium-sized carbon-hydrogen (CH) cluster $C_{60}H_{10}$. These family of systems, either pure carbon clusters or their hydrogenated counterparts, are object of intense theoretical scrutiny[28–32] owing to their crucial role in the formation of complex organic molecules[33,34]. Moreover, they have been revealed as potential candidates to carriers of the Aromatic Infrared Bands through the mechanochemical experiments of Dartois *et al.*.[22]

The rest of the article is structured as follows. In Section 2, we introduce the theoretical and methodological aspects involved in our global/local sampling scheme Mech-AMK. More specifically, we describe the different uses we make of graph theory (Subsection 2.1); then Subsection 2.2 is devoted to a comprehensive presentation of the methodology and its main features. In Section 3, we present a case study, the $C_{60}H_{10}$ system together with a detailed analysis of the resulting transformation networks. Section 4 concludes the paper.

# 2 Description of the Mech-AMK methodology

As succinctly discussed in the Introduction section, we have devised a workflow process that combines a physically inspired global stochastic sampling with an efficient local sampling method. The former corresponds to a series of optimizations under mechanochemical conditions for which a force is exerted along a random direction (see Section 2.2.2). The latter relies on the (vdW-)TSSCDS family of algorithms[17,26] as implemented in the AutoMekin software package[27] (see Section 2.2.4). All of the aforementioned computational methods have been integrated using the Python-based Atomic Simulation Environment (ASE) framework[35]. Regarding the analysis of our extensive results, we have utilized the graph data management platform `Neo4j` that provides seamless integration with Python.[36]

## 2.1 Graph theory as a means of representing chemical networks

Graph theory, or more specifically network analysis, is a powerful tool for the rationalization of dependencies in intricate dynamical systems.[27,37,38] To the best of our knowledge, its use in chemical networks dates back to the 1970s.[39] In this work, we use graph theory in two different contexts: (i) to represent chemical networks, as indicated above; and (ii) to represent the topology (connectivity) within a given target molecular system. This last application, common in Machine Learning applications[40–44], will be discussed in Section 3.2.

In brief, a graph ($G = (\mathcal{V}, \mathcal{E})$), a collection of nodes ($\mathcal{V} = \{v\}_{i=1}^{N}$) and edges ($\mathcal{E} = \{e\}_{i=1}^{M}$) can be seen as an abstract structured representation of a chemical network, reactive[17] or not[37]. In our specific context, the graph analysis of a chemical network will be restricted for a given basin, that is, a restricted entropically-confined region of the full global energy landscape. Landscapes in full might more suitably be rationalized by means of disconnectivity graphs[45].

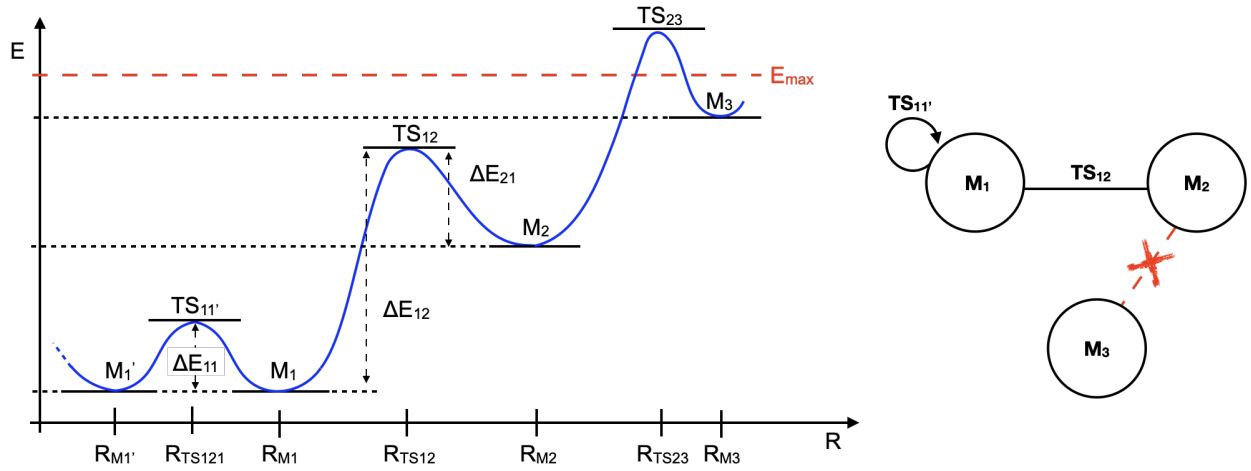



Figure 1: Graphical depiction of the transformation of a chemical reaction path (left) to a undirected graph representation. Minima are given by nodes and connecting transition states (TS) as edges. For simplicity, we have used the notation transition state (TS) instead of the more general term first-order saddle point which could also apply to non-covalently linked networks (see Ref. 37). It should be noted that all TS above the a given energy threshold ($E_{max}$) are not included in the RXN graph. This has been illustrated by a crossed red dashed line. Conversely, other stationary points, minima for instance, are shown as isolated units, possibly multiply connected.

In a chemical reaction network graph, minima are represented by nodes and minimum energy paths (comprising a first-order saddle point) are given by edges (see Figure 1). Consequently, the number of edges stemming from a node (its *degree*) represents all the possible chemical processes departing from a given minimum. Furthermore, in the special case of valley inflection ridges, double-ended edges are necessary. Indeed, these are topographical features occurring when a saddle point of symmetry $\Gamma_A$ connects one minimum (also $\Gamma_A$) with two equivalent minima whose point group symmetry fulfills $\Gamma_B \subset \Gamma_A$ (see for instance the case

of water-benzene[17] or pyrene-$NO_2$[37]). A last possibility are self-referencing nodes (loops) which indicate a minimum connecting with itself. In chemical terms, this corresponds to an isomerization connecting two equivalent minima. Note that in such case, a node would represent a basin and not a unique minimum.[37] Nodes and edges can be further characterized by colors (e.g. indicating their nature, dissociation fragments, minimum) and by weights (e.g. activation energy), respectively. This greatly enhances the amount of information contained in the graph, beyond its mere topology. In Figure 1, we provide an illustrative example including the most common features occurring in energy landscapes such as ours.

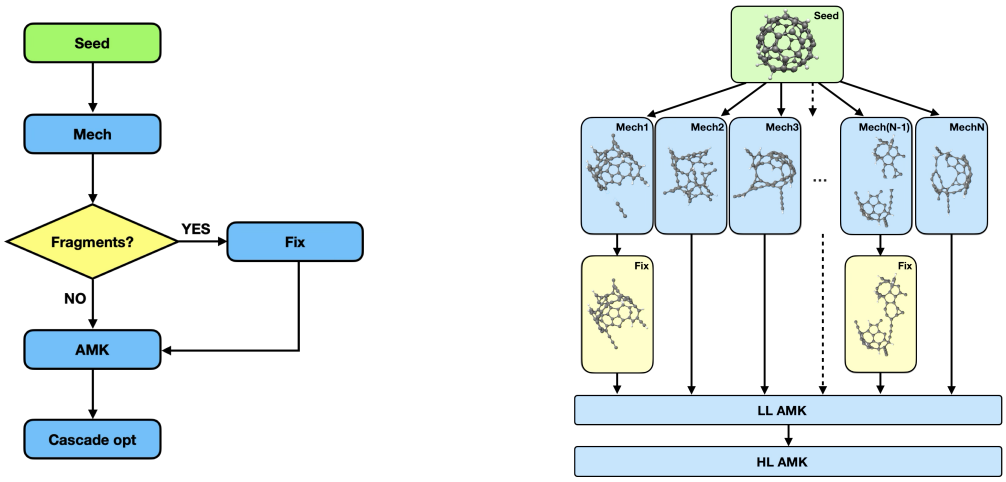



Figure 2: Workflow for automated generation methodology. The left figure presents the simplified workflow and the the right one shows the corresponding example structures for each procedures in the methodology.

## 2.2 Mech-AMK workflow

In this subsection, we outline the main steps involved in Mech-AMK which are schematically presented in Figure 2. In a nutshell, our methodology aims at characterizing the stationary points (minima, transition states) as well as the minimum energy paths connecting them, associated to the chemical (aka transformation) network resulting from a mechanochemical process. This solvent-free and scalable technique has been shown to be particularly suited for the type of system in hand.[18] As discussed below, from a microscopic molecular perspective, the collision with the milling balls can be represented by the interaction with a sphere of infinite radius, hence a plane.[46,47]

### 2.2.1 Initialization

As it is the case in the (vdW-)TSSCDS approaches[27] from which our methodology is derived, Mech-AMK requires an initial seed in the form of a single input geometry which is optimized at a low-level of theory ($\vec{R}_0 \rightarrow \vec{R}_0^{\text{opt, LL}}$). Obviously, N judiciously chosen seeds may lead to an increase of the efficiency of the process by a certain factor proportional to N. This proportionality factor is difficult to estimate and depends on the system under study.

### 2.2.2 Global exploration

The global (in-breadth) exploration phase starts with the optimized seed ($\vec{R}_0^{\text{opt, LL}}$) being further subjected to a series of geometrical optimizations under the influence of a mechanochemical force given by a randomly-oriented external potential which is added to the one resulting from electronic structure one[46]. The original PES is deformed so that bond breaking is favoured or hindered (see Figure 3). In other words, a molecular structure might become more prone to collapse under such a mechanical strain, thus facilitating the overcoming of energy or entropy barriers.[48] Owing to the random nature of the external force, the mechanochemical optimization can be viewed as a Monte Carlo step enabling over-

coming entropic barriers. However, in contrast to such methods[49–51], possibly epitomized by the Basin-Hopping algorithm[24], our stochastic step is directly linked to such a physical process.

We have employed the the `WallPotential` object from the `OpenMechanochem` Python package[47] that is seamlessly integrated with ASE. This module provides an imaginary constraining plane that interacts with the chemical system via a specified potential function accounting for the high-energy, localized impacts induced by the milling balls. At a molecular scale, the effect of the impinging ball can be seen as the pressure exerted by a plane or wall. Thence, we will use the adjective *wall-optimized* to refer to all the geometries obtained in the presence of such a force. Owing to the time-scale ratio between the mechanochemical process and the molecular dynamics, one can consider a *huge* number of collisions on a series of random impact sites. To generate the latter for the most general case of a three-dimensional molecular system, we generate randomly a 3-tuple of distinct atoms-labels ($[n_1, n_2, n_3]$) in a non-linear arrangement. Our mechanochemical wall is given by the vector associated to the plane defined by these three points. Of course, one can also consider replacing one (or two) coordinates of an atom by one (or two) point(s) outside the molecular system or, even, generate a distribution of wall angles with respect to a single atom. This would account for shear stress[52].

In our specific system ($C_{60}H_{10}$), we restrict the definition of walls to planes solely defined by a tuple of atoms. The maximum number of possible combinations (i.e. planes) equals $\binom{70}{3} = 54740$. This value constitutes a rather crude estimate of the maximum number of *potentially* accessible *wall-optimized* structures and does not correspond to the *actual* number of stable structures (aka minima). Indeed, the chemical space accessible is determined by the physicochemical properties of the system in hand.

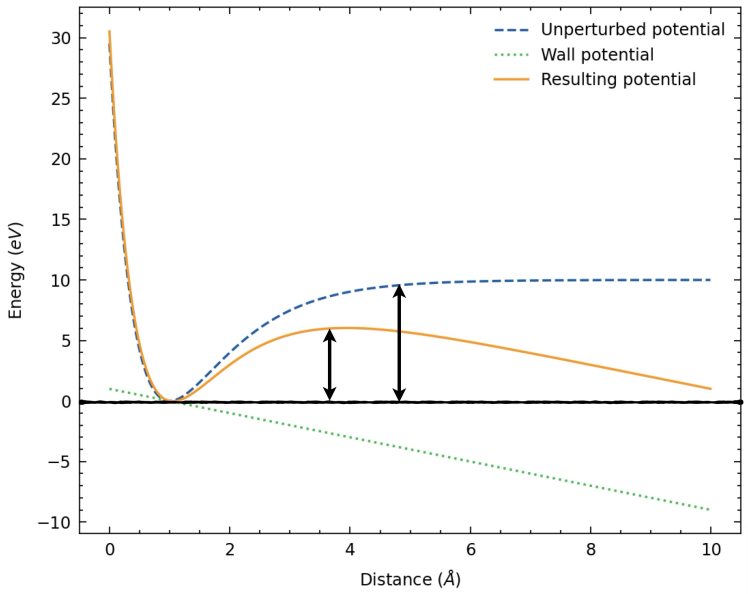



Figure 3: Illustration for mechanism of 'Mech' process. The blue dashed line represents the unperturbed potential, while the green dotted line illustrates the potential energy associated with the wall, where its slope corresponds to the force value. The resulting potential energy is given by the solid orange line. The double-arrows illustrate the difference in dissociation energy.

### 2.2.3 Fragmentation

As a result of a wall-optimisation, a given distortion ($\Delta\vec{R}_{n_1,n_2,n_3}$) will be induced $\vec{R}_{n_1,n_2,n_3} = \vec{R}_0 + \Delta\vec{R}_{n_1,n_2,n_3}$. Here the label $n_1, n_2, n_3$ refers to the defintion of the plane of mechanochemical distorsion. It should be obvious that this will most likely induce the breaking of some covalent chemical bonds and hence to the generation of fragments $\vec{R}_0 \rightarrow \vec{R'}_{n_1,n_2,n_3} + \vec{R''}_{n_1,n_2,n_3}$. Even though, this is a perfectly valid stochastic move, we shall restrict our study to a constant number of atoms, a necessary condition for the exploration of transformation (conformational) landscapes. Consequently, we merge these fragments and restore a single chemical entity. We do this by identifying fragments, analyzing the interatomic distances among them, $\vec{R'}_{n_1,n_2,n_3}$ and $\vec{R''}_{n_1,n_2,n_3}$, and sticking them together through an (unconstrained) optimization. The identification of the fragments is carried out through graph theory. To this end, we transform the set of 3D Cartesian coordinates into a 2D graph and analyze the different groups of connected components. In the following, for the sake of generality, we will consider that we

carry out this fragment fixing, although this step is optional. We have used the `NetworkX`[53] Python package for this. A given structure ($\vec{R} = \{R_i\}_{i=1}^{n}$) can be seen as a collection of $n$ atoms. To every $i$-th atom, we associate the set:

$$\{R_i\} = \{T_i, \vec{C}_i\},\ i \in [1, n] \tag{1}$$

where $T_i$ is the atom type, a string, and $\vec{C}_i$, the set of Cartesian coordinates of the $i$-th atom. Turning a molecular structure into a graph additionally requires the list of interatomic distances:

$$D : \{d_{i,j}\}, i, j \in [1, n]\ and\ i \neq j \tag{2}$$

and, since we are just interested in the topology of our molecular structure and not in its actual physical dimensions, we will define *threshold* distances ($\delta$), which depend only by the type of atoms. For example for $C_{60}H_{10}$, we have only 2 threshold values: $\delta_{C-C}$ and $\delta_{C-H}$ (see Section 3.1).

In the context of our `Fragmentation` module, the graph derived from a single structure is taken as undirected, thereby indicating that each edge within this framework is unidirectional. The algorithm we employed for constructing a graph from a given structure $\{R_i\}$ is presented in Algorithm 1.

**Algorithm 1** Conversion of structure $\vec{R}$ into a graph

**Require:** $\{R_i\}, \{d_{i,j}\}, \delta_{T_i - T_j}$
**Ensure:** $i, j \in [1, n]\ and\ i \neq j,\ n > 2\ and\ n \in \mathbb{Z}$
  $G$ add nodes from $\{R_i\}$ ▷ add nodes function from `NetworkX`
  **for all** combinations of $(i, j)$ **do**
    **if** $d_{i,j} \leq \delta_{T_i - T_j}$ **then**
      $G$ add edge $e_{i,j}$
    **end if**
  **end for**
    **return** $\{v_i, e_{i,j}\}$

Once the fragmented molecular graph has been generated, we employ the `NetworkX` Python

library to identify, in an automated fashion, all isolated subgraphs, hereafter referred to as *islands*, within the full graph along with their associated node indices. These subgraphs correspond to dissociated atoms or molecular fragments. Let us assume that after a wall-optimization, we obtain a fragmented geometry ($\vec{R}_i$). This is turned into a molecular graph ($G$). Since it is fragmented, it presents a series of $N$ *islands* collectively denoted as $\{\mathcal{S}_k\}_{k=1}^{N}$. If the wall-optimized structure has not been fragmented then the count of connected components equals 1 ($N = 1$). The number of nodes within $S_k$ is $m_k$. We denote $S_1$ as the *central component*, that is, the *island* harboring the highest number of nodes ($m_1$). The remaining *islands* ($m_{k>1} \leq m_1$) are denoted *fragments* and will be re-integrated (re-connected) to the central component ($S_1$). Algorithm 2 below delineates the process for relocating fragments. The reference interval for $T_i - T_j$ bond lengths are defined as ranges of distance between the central component and fragments after relocation. The updated distances are randomly selected from these intervals.

**Algorithm 2** The function to relocate fragments

**Require:** $\{\mathcal{S}\}_{k=1}^{N}, \{d_{i,j}\}, \{R_i\}, \{\delta_{min}^{T_i-T_j}, \delta_{max}^{T_i-T_j}\}$
**Ensure:** $k \in [1, N]$ *and* $N \in \mathbb{Z}^+,\ i, j \in [1, n]$ *and* $i \neq j,\ n > 2$ *and* $n \in \mathbb{Z}^+$

$p \leftarrow N$
$S_1^p \leftarrow S_1^N$ ▷ The most connected component in $\{\mathcal{S}_k\}^N$
$S_2^p \leftarrow S_2^N$. ▷ The second most connected component in $\{\mathcal{S}_k\}^N$
**while** $p > 1$ **do**
  $d_0 = min\{d_{i,j}\}_{i \in S_1^p, j \in S_2^p}$
  $i_0, j_0 \leftarrow i, j\ for\ d_0$
  $d_{i_0,j_0}^{p-1} = rand[\delta_{min}^{T_i-T_j}, \delta_{max}^{T_i-T_j}]$
  $\lambda = d_{i_0,j_0}^{p-1}/d_0$
  $\vec{C}_{j_0}^{p-1} = (1-\lambda)\vec{C}_{i_0}^{p} + \lambda\vec{C}_{j_0}^{p}$ ▷ Relocate the nearest node to central component
  **for all** $j \in S_2^p$ **do**
    $\vec{C}_{j}^{p-1} = \vec{C}_{j_0}^{p-1} + \vec{C}_{j}^{p} - \vec{C}_{j_0}^{p}$ ▷ Move the remaining fragment nodes concurrently
    **for all** $i \notin S_2^p$ **do**
      $d_{i,j}^{p-1} = \|C_i^p - C_j^{p-1}\|$ ▷ Update distances between nodes
    **end for**
  **end for**
  $p = p - 1$
**end while**
  **return** $\{R_i^{\text{new}}\} = \{T_i, C_i^1\}$

### 2.2.4 Local exploration and geometry refinement

The mechanochemical *wall-optimisation* process generates a set of $\{\vec{R}^{(i)}_{n_1,n_2,n_3}\}_{i=1}^{N}$ structures which, in general, should constitute a good set of representatives of the different basins constituting the energy landscape of the system (see Figure 5 and related discussion). In turn, in Mech-AMK these geometries are employed in a subsequent basin (local) exploration. In our terminology, local refers to an energy-barrier controlled region surrounded by large entropic barriers. To this end, we use the (vdW-)TSSCDS family of methods. These algorithms (and others) are implemented in the AutoMekin (AMK), a freely available software package, designed for the automatic characterization of chemical (or transformation) networks, that is the set of minima, saddle-points, dissociation products and minimum energy paths connecting them underlying a given basin.[27] Hereafter, for short, we will refer to both algorithm and software collectively as AMK. AMK aims at locating transition states (or more generally first-order saddle points[37]) which, arguably, constitute the most-difficult-to-locate stationary points. AMK combines classical Molecular Dynamics with a graph-theory based algorithm that detects structural changes in the system, potentially related to the occurrence of a transition state. In short, it analyzes, at a given sampling rate, changes in the adjacency (connectivity) matrix of the MD generated structures. The AMK workflow relies on two distinct levels of theory. An initial sampling is carried our at a low-Level (LL) of theory. It should be highlighted that, recently, even DFT has been used as LL.[54] The so-obtained transition state TS-candidate structures (LL) are subsequently optimized at a user-defined high-Level (HL) of theory. This workflow is particularly efficient for the local exploration of basins and hence, in this work, we will use it mainly for this purpose. The numerical illustration of this statement is discussed around Table 1 (see Section 3.2).

In general, one can employ canonical or microcanonical trajectories for sampling configuration space. In our specific case, we employed microcanonical MD. The number of trajectories per batch is set to 10, with a total of 50 batches, and the remaining parameters have been left

at their default values. Specifically, we set low convergence criteria for the LL geometrical optimizations in order to obtain clouds of candidate structures. In this way, we minimize the probability of missing TS once optimized at higher levels of theory.[17] From the TSs (LL or HL), Intrinsic Reaction Coordinate (IRC) calculations are launched to determine the associated minima. In terms of output, AMK provides a comprehensive array of data, encompassing molecular geometries of the stationary points located (Cartesian coordinates), the delineation of the reaction network (RXN), and a database file categorizing the different stationary points: minima, transition states, and dissociation products.

# 3 A first Mech-AMK case-study: $C_{60}H_{10}$

In this section, we present and discuss in detail the application of our Mech-AMK methodology to a system of astrophysical interest, a $C_{60}H_{10}$ C-H cluster[22]. Indeed, the conformational distribution in this type of clusters is expected to provide valuable insights for the disentanglement of the structural features of the (still) unknown products obtained via mechanochemical milling of fullerene in $H_2$ atmosphere.[22]

$C_{60}H_{10}$ constitutes a medium size cluster ($N$=70) beyond the reach of non-automated methods and similar recent works[29]. We address all practical aspects from the preparation of an initial seed down to the (collective/statistical) analysis of the individual structures, together with the tools we have employed and/or developed.

## 3.1 Initialization, progress and conclusion of the simulations

Any sampling stochastic/deterministic sampling strategy requires one or multiple initial structures (seeds). In our case, in order to illustrate the power of our approach, we have opted for utilizing a single structure (Figure 4). This seed has been generated through a mechanochemical optimization of a fullerene molecule ($C_{60}$) with 10 randomly oriented dihydrogen molecules, including treatment of the fragments (see Section 2.2.3). This system

constitutes a preliminary model for recent mechanochemical studies in CH-clusters[22]. It should be noted that at this benchmark stage, we have realized some methodological choices. First, for illustrative purposes, we have limited our study to the so-called low level of theory, specifically the semiempirical potential PM7[55]. The latter constitutes the low-level of choice for preliminary explorations[17,27] and has been shown to work even in the case of weak non-covalent interactions[17,37]. Hereafter, unless otherwise indicated, all optimizations will be carried out for the neutral species and will employ this level of electronic structure theory. Concerning the mechanochemical operatory conditions, we have chosen a linear force of 5 eV/Å.

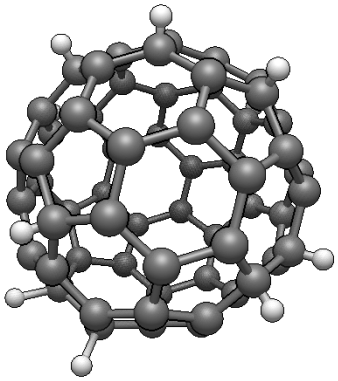


Figure 4: Our initial geometry (seed) of $C_{60}H_{10}$

An initial $C_{60}H_{10}$ seed was obtained using our `fragmentation` module followed by a geometry optimization (see Figure 4). As thresholds distances, we have taken somewhat elongated bond-length values: $\delta_{C-C} = 1.8$ Å[56] and $\delta_{C-H} = 1.2$ Å. This loose criteria responds to the large variability of geometries obtained and enables a clear identification of broken bonds and fragments beyond a mere lack of connectivity between neighboring atoms. The ranges of distances for the relocation of the fragments to a given *central* component are defined as $[\delta^{C-C}_{min}, \delta^{C-C}_{max}] = [1.5\ \mathring{A}, 1.6\ \mathring{A}]$ and $[\delta^{C-H}_{min}, \delta^{C-H}_{max}] = [1.0\ \mathring{A}, 1.1\ \mathring{A}]$, respectively. These values were empirically determined for the $C_{60}H_{10}$ system and were found to minimize the number of steps in the process.

Subsequently, a seed undergoes a series of 500 stochastical mechanochemical optimizations yielding a total of 212 distinct structures, referred to as *wall-optimized* structures ($\mathrm{AMKi}_{i=1}^{212}$). It should be emphasized that several optimizations might lead to the same minima hence reducing the ratio between the (*potential*) initial seeds and the resulting number of *actual* ones. The wall-optimized structures can be considered as representatives of the different basins. As such, they are subsequently taken as seeds for the local AMK sampling of the different basins. To illustrate this, we have randomly selected three *wall-optimized* seeds (AMKx, x=1,2,3) from the pretzel/cage region (*vide infra*). The corresponding AMK searches have resulted in a total of 13644 geometries, including initial wall-optimized seeds and the subsequent AMK basin-searches. For instance, for AMK1, we carried out a total of 9 AMK basin-search iterations. The total number of trajectories was set to 5000. The average time per trajectory was 4.94 minutes(296 seconds). For the sake of comparison, we have additionally re-optimized a series of 35 wall-optimized structures using the B3LYP/cc-pVDZ level of theory. The average CPU time for these calculations was 9398.59 minutes ($>5.6 \times 10^5$ seconds), and the average *elapsed* time was 2586.60 minutes ($1.5 \times 10^5$ seconds). These figures make clear that a straightforward DFT exploration is prohibitively expensive above a certain molecular size. The solution should hence involve the use of some type of artificial-intelligence-enhanced quantum-chemical approaches such as the AIQM family of methods[57].

In Figure 5, we present the distribution of structures in terms of their asphericity (see Supplementary Information for a definition) and a fraction of $sp^2$ atoms. Based on these two parameters, the geometries can be classified into four categories: (i) cages, (ii) pretzels, (iii) flakes, and (iv) branched structures, as suggested in previous works[29,58]. In short, the cages category includes molecules with an $sp^2$ atom fraction exceeding 75%. Pretzels are defined by an $sp^2$ fraction below 75% and an asphericity ( $\Omega_A$ ) less than 0.3. Flakes have intermediate $sp^2$ fractions (45%–75%) and $\Omega_A$ values exceeding 0.3. The remaining geometries, characterized by $sp^2$ fractions below 45% and $\Omega_A > 0.3$, are classified as branched structures[30,58]. The

number of geometries per group and category is presented in Table 1. Pretzels constitute by far the largest proportion, accounting for 96.45% of the entire dataset. Cages represent a significantly smaller fraction (3.26%), yet they remain more prevalent than flakes (0.08%) or branched structures (0.21%). As observed, our results span the whole range of structures from cages to branched, passing through pretzels. This is common to pure C-clusters as well as to smaller CH-clusters[58,59]. Moreover, we observe a much denser sampling in our specific local sampling region, the low-asphericity one, which is more interesting to us.[22] For a more detailed discussion, see Section 3.2. The distributions with respect to other features and aphericity are represented in the Figure S3.

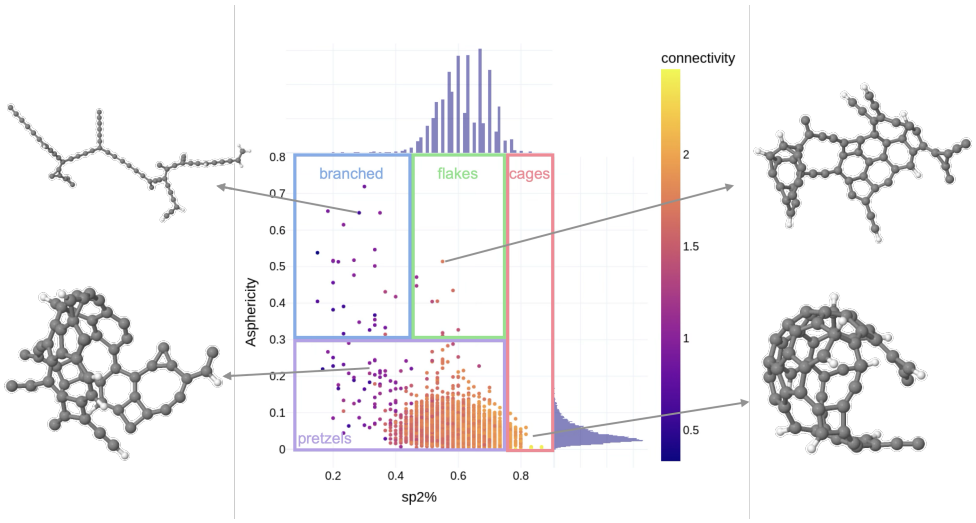



Figure 5: Distribution of 13644 $C_{60}H_{10}$ optimized structures. Both global and local results are presented. The horizontal axis corresponds to the fraction of $sp^2$ C atoms. The vertical axis indicates the asphericity values, that is the deviation from a spherical shape. A detailed description is provided in the Supporting Information. The average node connectivity is indicated using a continuous color scale shown on the right side of the graph. Illustrative geometries for each category are also displayed.

Table 1: Number of molecules per group and category

| Categories \ Groups | All | AMK1 | AMK2 | AMK3 | Wall-optimised |
|---|---|---|---|---|---|
| cages | 445 (3.26%) | - | 127 (1.91%) | 287 (13.39%) | 31 (14.62%) |
| pretzels | 13160 (96.45%) | 4658 (99.99%) | 6499 (98.06%) | 1857 (86.61%) | 146 (68.86%) |
| flakes | 10 | 3 | 1 | - | 6 |
| branched | 29 (0.21%) | - | - | - | 29 (13.68%) |
| All | 13644 | 4661 | 6627 | 2144 | 212 |

## 3.2 Graph-theoretical and statistical analysis of the data

We have shown that our Mech-AMK approach is able to efficiently sample the vast configurational space of the $C_{60}H_{10}$ target system. In this subsection, we analyze these results with a twofold purpose: (i) to assess the degree of coverage of our chemical space and (ii) to unveil the global chemical nature of the determined structures. For the first goal, we will use a metric based on some global shape descriptors which have been shown to be suitable for chemically-related systems. Specifically, we utilize a series of nine descriptors, namely: (1) total number of bonds; (2) average node connectivity; (3) number of cycles in the molecular graph; fraction of atoms that (4) singly (N1); (5) doubly, (N2); (6) triply (N3); and (7) quadruply (N4) connected to other atoms; (8) radius of gyration; and (9) asphericity. These features have been adapted from Refs. 29,60. It should be noted that all descriptors, except the last two ones, are related to the topology of the molecular graph, where atoms correspond to (atom labelled) nodes and bonds to edges. This results in an undirected, unweighted graph for each molecular structure. The non-topological features can be directly computed using the `RDKit`[61] package. The detailed description of these features together with the corresponding Principal Component Analysis(PCA) are provided in the Supporting Information. Finally, concerning the chemical nature of our compounds, as illustrative examples, we will analyze their bond-length and type of cycle distributions.

### 3.2.1 Basin analysis through graph theory

As discussed in Section 2.1, our RXNs (one per AMKx basin) are stored as directed graphs (see Figure 1). For the detailed analysis of each of the AMKx RXN graphs, we investigate the connectivity relationships among nodes embedded within these complex architectures. To this end, we first identify the most-connected islands (subgraphs), meaning those subgraphs that comprise the greatest number of nodes (or vertices) within a given graph. To determine them, we start from the most connected node and perform an in-depth traversal along all possible paths/nodes in a recursive manner. From these subgraphs, one can directly access the

encoded chemical information. For instance, in Figure 6, we present the graph corresponding to the RXN associated to AMK1. It should be noted that the circular form is merely for representation purposes, and no meaning should be ascribed to it. Each node (blue-colored dot) graph represents a minimum, and the line connecting them (edges) corresponds to the transition states. By progressively zooming in, we can arrive at a *traditional* chemical reaction scheme. In this specific case, we have selected two highly-connected islands and, in each of them, we have illustrated a given chemical path: a series of hydrogen transpositions (bottom-left) and a C-C skeleton rearrangement (top-right). We have highlighted the molecular region of interest for each reaction. As it can be observed, a large fraction of subgraphs are isolated from lowly-connected units. In graph theory, a connected component is a subgraph in which every pair of nodes is connected by a path. This can be explained by two facts: (i) the actual TS lies above our energy threshold (our AMK simulation is microcanonical), and (ii) the AMK algorithm did not manage to find the corresponding TS probably due to not fully converged results owing to the stochastic nature in the choice of the initial conditions in the AMK simulations.[17,25,26,37] In our case, RXN convergence is not relevant since we are mainly interested in minima and their statistical distribution. Of course, one can deal with these two issues. The first one requires simply an increase in the total MD energy. In other words, a *completely* connected graph could be obtained only for small systems. Hence, for most of the cases, a partially disconnected graph is obtained. The second issue is a bit trickier since it would require a full convergence of our simulation. This requires again a *large* number of trajectories and possibly a fine-tuning of several AMK parameters, such as the sampling rate or the redefinition of the normalized distances.[17]

### 3.2.2 Analysis of the C-C and C-H bond-length distributions

As a final structural analysis, we aim at rationalizing the most relevant geometrical features of the whole set of structures we have obtained, either globally or particular to a basin. Specifically, we will analyze the different, C-X (X=C, H) bond-length distributions.[62,63] Indeed,

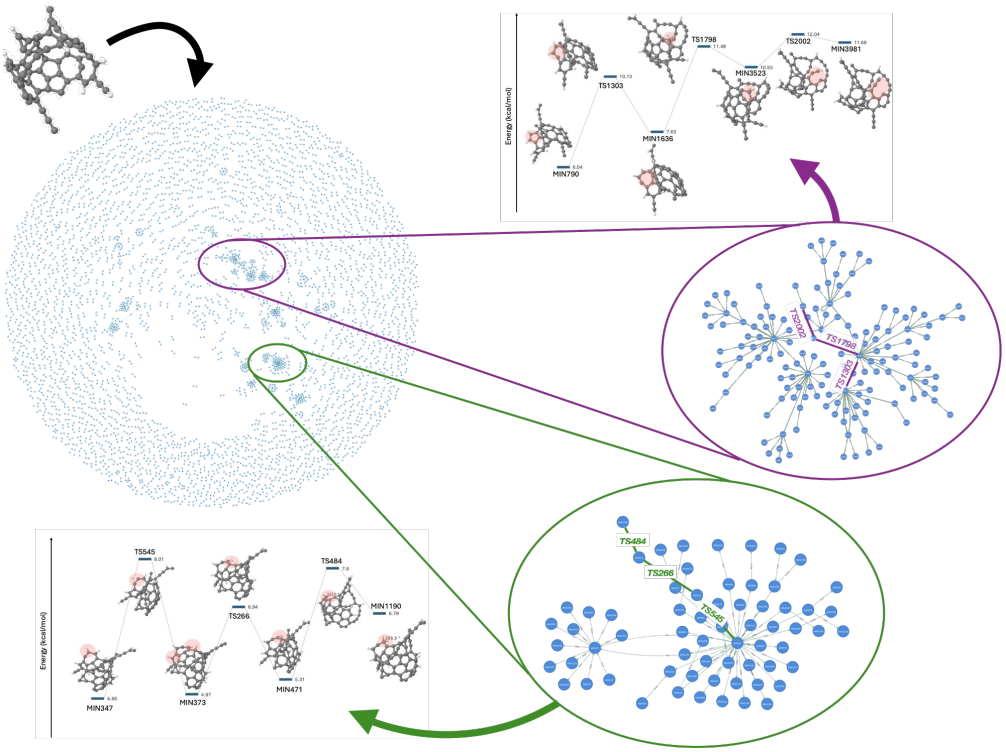
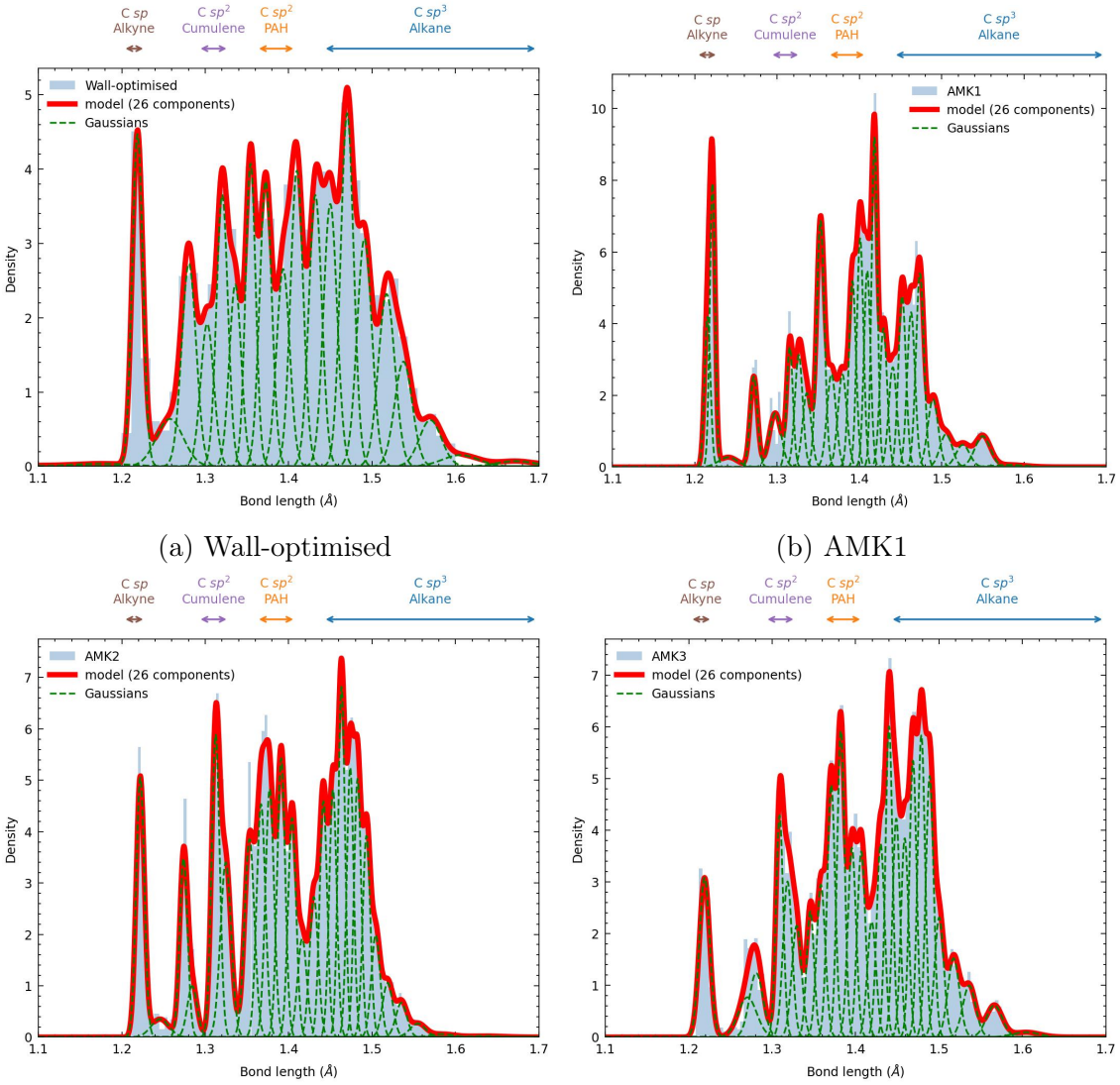



Figure 6: Reaction Network associated to the AMK1 basin presenting all stationary points, minima and TS. The basin exploration corresponds to an AMK simulation from a single wall-optimized $C_{60}H_{10}$ conformer. In the reaction paths, we have colored the molecular region of interest. See main text for further details.

the distribution of bond-length values constitutes a signature of the different Chemical Local Environments (CLE) present in a given system. For example, in the case of a simple alkane, such as pentane (Figure S4), in its equilibrium geometry, there are 4 C-C $sp^3$ bonds distributed in two different bond lengths. This seems to suggest that different bond-lengths may, in principle, provide a good hint on the various CLEs within a molecule. For this analysis, we have computed the C-X distances for all of the clusters and represented them in the form of a histogram. We have made a histogram with *Freedman Diaconis Estimator* as bin. This method is robust and accounts for outliers, data variability, and data size.[64] The resulting histograms have been interpolated into continuous distributions by means of a fit in terms of a linear combination of Gaussian functions. Centers, variances, weights, and the number of Gaussian functions have been optimized to maximize the agreement with the underlying histogram. The quality of these fits has been measured by means of the Integrated Square

Error (ISE). In Figure S5, we present the ISE for (i) the whole set of $C_{60}H_{10}$ structures, (ii) the wall-optimized molecules only, and (iii) the different AMKx clusters. As can be observed, the ISE achieves convergence after 26 components for all groups, which indicates the stability of the solution.




(a) Wall-optimised

(b) AMK1

(c) AMK2

(d) AMK3

Figure 7: The distribution of C-C bond lengths for four groups of molecules(wall-optimized molecules and three clusters) and its Gaussian fitting.

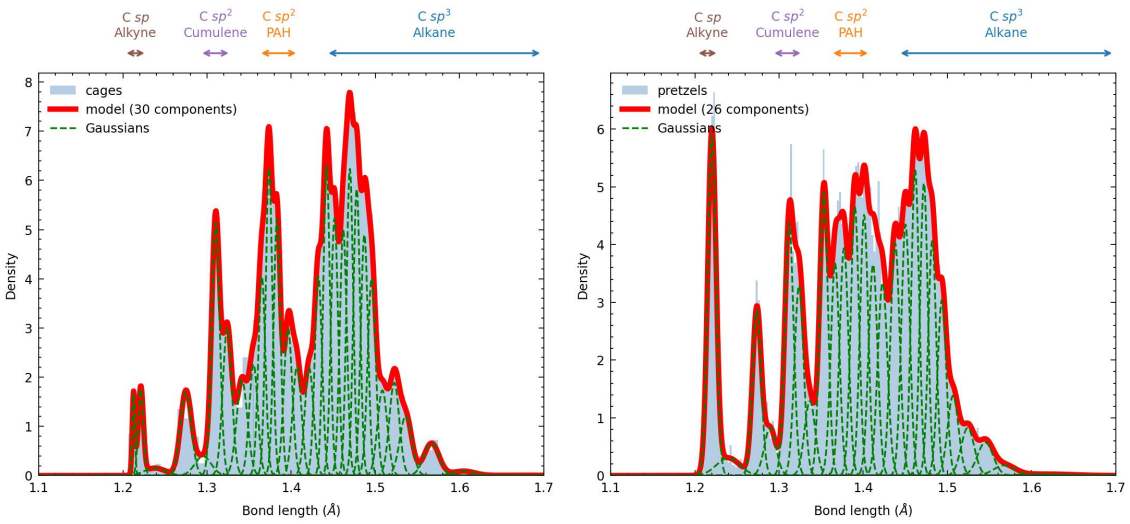
In the following, the features of the 26 Gaussian components will be used to rationalize the associated C-C bond-length distribution for the full set of structures (shown in Figure 7 and S6). In the Supporting Information, in Table S3, we display the mean value for the Gaussian

functions in the mixture. The Gaussian means for C–C bond lengths are filtered using two criteria: the mean values are restricted to the range 1.1 Å–1.7 Å and the Gaussian variance must be below $5 \times 10^{-3}$ Å$^2$. The first condition ensures consistency with the bond length restrictions mentioned earlier, while the second excludes Gaussian functions that represent flat distributions rather than distinct peaks. These means correspond to the most frequent C-C bond lengths in the molecule which, in turn, points towards a certain type of recurrent molecular sub-structure within them. To compare the distribution of C-C bond lengths among different sets of data, the mean values have been listed in ascending order, and the domain of definition has been restricted to chemically meaningful values between 1.1 Å and 1.7 Å in order to remove possible outliers.

However, in the case of branched structures, the ISE diverges (see Figure S7), most likely due to the limited sample size for this category. This somehow shows that the ISE criterion alone might not be able to provide a proper and consistent number of components for various categories. Given the potential variation in the distribution of C–C bond lengths across different categories, selecting the number of components separately for each category seems to be a more sensitive approach. On top of that, we have adopted two additional quality criteria: (i) the *Akaike Information Criterion* (AIC) and (ii) the *Bayesian Information Criterion* (BIC). Both are widely used in model selection in statistics and machine learning.[65] AIC evaluates the relative quality of the statistical model for a given dataset by balancing goodness of fit with model complexity. Similarly, BIC penalizes models with more parameters but applies a stricter penalty, favoring simpler models more strongly. Both lower AIC and BIC scores indicate a better-fitting model. While the definitions of AIC and BIC may vary across the literature, we adopted the implementation provided in the `scikit-learn` Python package.[66]

With these considerations, our best strategy for determining the optimal number of Gaussian

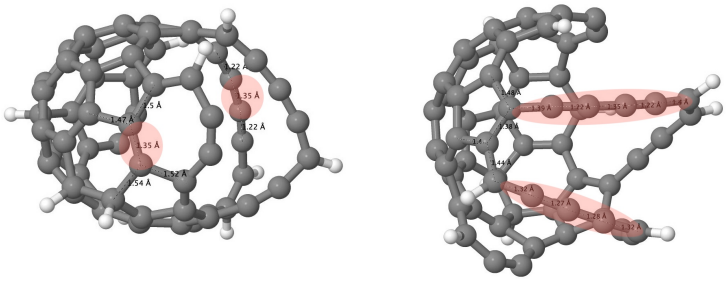
components would be: (i) selection of the number of components leading to the lowest AIC/BIC scores; (ii) identification of the *elbow* point[65] which refers to the point on the curve where the rate of decrease in AIC/BIC values undergoes a noticeable change, if the scores continue to decrease as the number of components increases; and (iii) prioritization of the BIC score when it differs from the AIC result, as BIC is stricter and tends to favor smaller models.[65] Using these more stringent criteria, we obtained a new set of Gaussian components: 30, 26, 6, and 6 for cages, pretzels, flakes, and branched structures, respectively (see Figure S8). It is important to note that the number of Gaussian components used here should be considered as a reference for fitting the distributions. While this number may fluctuate, the overall shape of the distribution remains the most consistent and reliable feature to consider. Figure 8 illustrates the distribution of C–C bond lengths for cages and pretzels, along with the corresponding Gaussian fitting. The fitting function for distributions of flakes and branched structures are less reliable due to their limited sample sizes and are provided in the appendix (see Figure S9).




(a) cages (b) pretzels

Figure 8: The distribution of C-C bond lengths for two categories of molecules(cages and pretzels) and its Gaussian fitting.The bond length ranges for typical substructures have been highlighted on top of each figure (see detailed values in Table S2).

The meaning of these distributions is straightforward. For instance, in the case of pretzels, it shows the occurrence of a set of 26 C-C bond-lengths sub-distributions. It is tempting to ascribe the latter as the direct mapping of the different CLEs, thus reflecting the chemical nature of the various C-C bonds in a pretzel cluster. Obviously, this would be very interesting from a structure/property relationship perspective, particularly in the context of the prediction of, for instance, spectroscopical properties. However, the situation is a bit more complex than this. Indeed, it can be easily understood that totally different CLEs carry associated C-C bonds that lie under the umbrella of the same peak. Or, less evident, C-C bonds within seemingly similar CLEs may exhibit different bond lengths and lie under two different peaks. The former case can be understood as a quasi-random factor due to the specific conformation of the system and the associated non-covalent intramolecular interactions. The latter case, however, is less intuitive and does not follow this *pretzel logic*, that is, different CLE but *same* distance. To illustrate this, we present these two cases in Figure 9. Specifically, one can observe how this is true due to the bond length alternation of cumulenes.[67]



(a) CLE1 (b) CLE2

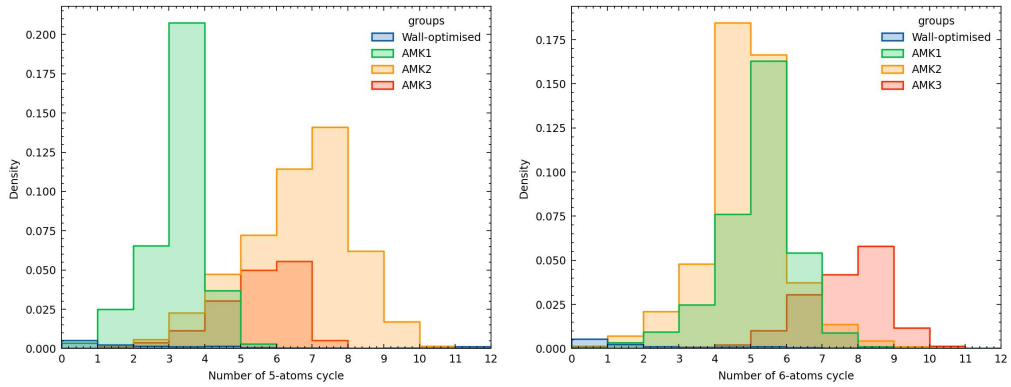
Figure 9: Example geometries from with different CLE but similar C-C bond lengths(left) and similar CLE but different C-C bond lengths(right).

As illustrated in Figure 5 and Table 1, pretzels constitute a significant proportion of both the entire set of structures and each AMKx cluster. Specifically, pretzels account for 96.45% of the complete dataset, 99.99% of AMK1, and 98.06% of AMK2. The high proportion of pretzels explains why the number of Gaussian components for pretzels matches that of

the entire dataset. Similarly, the large proportion of pretzels in AMK2 explains the close similarity between the distributions and fitting models of the complete dataset (Figure S6) and AMK2 (Figure 7c) with those of the pretzels. In spite of their common *pretzel* character (with a slightly higher proportion of pretzels in AMK1), AMK1 (Figure 7b) and AMK2 distributions (Figure 7b) show a clear difference in their CLE distribution: AMK1 shows less similarity with pretzels than AMK2.

This discrepancy can be attributed to two factors. The more straightforward reason is that the number of pretzels present in AMK2 (6,649) largely exceeds that of AMK1 (4,658). And second, despite both being classified as pretzels, AMK1 (Figure S10a) presents a much fewer number of rings than AMK2 (Figure S10b). This is illustrated in Figure 10 for all AMKx clusters, in fact these histograms clearly reveal the different nature of the AMKx (x=1,2,3) basins and the need for different representations to assess the occurrence of different basins. These structural differences underlie the aforementioned variations in the C–C bond-length distributions. The peaks around 1.22 Å and 1.27–1.29 Å in the cages distribution exhibit a lower frequency when compared to pretzels. These C–C bond lengths are primarily associated with cumulene sub-structures, reflecting the lower proportion of cumulenes in cages. The C–C bond lengths of 5-atom and 6-atom cycles ($R_5, R_6$) show two concentrated ranges around 1.38 Å and 1.45 Å (see Figure S11), which is consistent with the cages distribution in clear contrast to the pretzels. This reflects the prevalence of 5-atom and 6-atom cycles in cages when compared to pretzels.

Thereafter, we turn our attention to the C-H bond-length distribution. The situation for the C-H bond is somewhat similar to the C-C bonds mentioned above. Besides, the C-H bond contains one carbon atom and one hydrogen atom, thus greatly simplifying the number of possible CLEs. In other words, from a topological perspective, hydrogen can only connect with one carbon atom. Figure S13 illustrates this: they show a much smaller number of peaks than the distribution of C-C bond lengths. This is reflected in the ISE convergence




(a) $R_5$ (b) $R_6$

Figure 10: Distribution of number of cycles

(Figure S12), which stabilizes after 8 Gaussian components. The mean values of 8 Gaussian functions that fitted the C-H bond length distribution for each group are given in Table S3. For the above reasons, the mean values are listed in ascending order, and the scale of C-H bond lengths is restricted between 1.00 Å and 1.25 Å. Moreover, the corresponding variance is below $5 \times 10^{-4}$.

In addition, we note that the use of Gaussians as fitting functions implies that the variability of a given bond-length is given by a random perturbation, in our case, a random arrangement of the environment of the targeted bond. For our systems, with a relatively low number of atoms and associated conformations, this assumption might not perfectly hold. However, this should be true for larger systems such as the ones in the mechanochemical experiments.[22] Finally, to further illustrate the power of our approach in generating specific basin-based distributions, we present in the Supplementary Information our preliminary results on a much larger system consisting on 130 atoms, namely the $C_{120}H_{10}$ (294D).

# 4 Conclusions

In this work we have presented our Mech-AMK methodology for the automated global characterization of transformation and/or reaction networks associated to high-dimensional chemical spaces. Its goal is to provide a simple means of accessing statistically significant data set which is essential to ensure meaningful predictions in any Deep Learning model. Our methodology efficiently combines: (i) a global (in-breadth) exploration of the conformational landscape by means of a physically-oriented stochastic mechanochemical hopping strategy and (ii) an efficient local (in-depth) sampling MD technique which relies on the Transition State Search using Chemical Dynamics Simulations(TSSCDS). This will greatly enhance the efficiency and statistical reliability of our approach while keeping relatively low computational costs.

# 5 Data availability

The set of structures $C_{60}H_{10}$ can be retrieved from: https://doi.org/10.5281/zenodo.15262282

# References


(1) *Venture Capital Investments in Artificial Intelligence: Analysing Trends in VC in AI Companies from 2012 through 2020*; OECD Digital Economy Papers 319, 2021; Vol. 319.

(2) Philip Chen, C. L.; Zhang, C.-Y. Data-Intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data. *Inf. Sci.* **2014**, *275*, 314–347.

(3) Keith, J. A.; Vassilev-Galindo, V.; Cheng, B.; Chmiela, S.; Gastegger, M.; Müller, K.-R.; Tkatchenko, A. Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems. *Chem. Rev.* **2021**, *121*, 9816–9872.

(4) Nobel Prize Outreach AB The Nobel Prize in Physics 2024. 2024; https://www.nobelprize.org/prizes/physics/2024/summary/, Accessed: 2024-12-20.

(5) Nobel Prize Outreach AB The Nobel Prize in Chemistry 2024. 2024; https://www.nobelprize.org/prizes/chemistry/2024/summary/, Accessed: 2024-12-20.

(6) Karagiorgi, G.; Kasieczka, G.; Kravitz, S.; Nachman, B.; Shih, D. Machine Learning in the Search for New Fundamental Physics. *Nat. Rev. Phys.* **2022**, *4*, 399–412.

(7) Wu, Z.; Ramsundar, B.; Feinberg, E.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **2018**, *9*, 513–530.

(8) Ullah, A.; Chen, Y.; Dral, P. O. Molecular Quantum Chemical Data Sets and Databases for Machine Learning Potentials. *Mach. Learn.: Sci. Technol.* **2024**, *5*, 041001.

(9) Shumailov, I.; Shumaylov, Z.; Zhao, Y.; Papernot, N.; Anderson, R.; Gal, Y. AI Models Collapse When Trained on Recursively Generated Data. *Nature* **2024**, *631*, 755–759.

(10) Chen, Y.; Dral, P. O. AIQM2: Better Reaction Simulations with the 2nd Generation of General-Purpose AI-enhanced Quantum Mechanical Method. 2024.

(11) Worth, G. A.; Cederbaum, L. S. Beyond Born-Oppenheimer: Conical intersections and their impact on molecular dynamics. *Ann. Rev. Phys. Chem.* **2004**, *55*, 127–158.

(12) Arenas, J. F.; Centeno, S. P.; López-Tocón, I.; Peláez, D.; Soto, J. DFT and CASPT2 Study of Two Thermal Reactions of Nitromethane: C–N Bond Cleavage and Nitro-to-Nitrite Isomerization. An Example of the Inverse Symmetry Breaking Deficiency in Density Functional Calculations of an Homolytic Dissociation. *J. Mol. Struct. THEOCHEM* **2003**, *630*, 17–23.

(13) Soto, J.; Arenas, J. F.; Otero, J. C.; Peláez, D. Effect of an S1/S0 Conical Intersection on the Chemistry of Nitramide in Its Ground State. A Comparative CASPT2 Study of the Nitro-Nitrite Isomerization Reactions in Nitramide and Nitromethane. *J. Phys. Chem. A* **2006**, *110*, 8221–8226.

(14) Mardirossian, N.; Head-Gordon, M. Thirty Years of Density Functional Theory in Computational Chemistry: An Overview and Extensive Assessment of 200 Density Functionals. *Mol. Phys.* **2017**, *115*, 2315–2372.

(15) Panadés-Barrueta, R. L.; Martínez-Núñez, E.; Peláez, D. Specific Reaction Parameter Multigrid POTFIT (SRP-MGPF): Automatic Generation of Sum-of-Products Form Potential Energy Surfaces for Quantum Dynamical Calculations. *Front. Chem.* **2019**, *7*, 576.

(16) C. Chapleski, R.; Zhang, Y.; Troya, D.; R. Morris, J. Heterogeneous Chemistry and Reaction Dynamics of the Atmospheric Oxidants, O 3 , NO 3 , and OH, on Organic Surfaces. *Chem. Soc. Rev.* **2016**, *45*, 3731–3746.

(17) Kopec, S.; Martínez-Núñez, E.; Soto, J.; Peláez, D. vdW-TSSCDS—An Automated and Global Procedure for the Computation of Stationary Points on Intermolecular Potential Energy Surfaces. *Int. J. Quant. Chem.* **2019**, *119*.

(18) Báti, G.; Csókás, D.; Yong, T.; Tam, S. M.; Shi, R. R. S.; Webster, R. D.; Pápai, I.; García, F.; Stuparu, M. C. Mechanochemical Synthesis of Corannulene-Based Curved Nanographenes. *Angew. Chem. Int. Ed.* **2020**, *59*, 21620–21626.

(19) Horst, M.; Meisner, J.; Yang, J.; Kouznetsova, T. B.; Craig, S. L.; Martínez, T. J.; Xia, Y. Mechanochemistry of Pterodactylane. *J. Am. Chem. Soc.* **2024**, *146*, 884–891.

(20) Simm, G. N.; Vaucher, A. C.; Reiher, M. Exploration of Reaction Pathways and Chemical Transformation Networks. *J. Phys. Chem. A* **2019**, *123*, 385–399.

(21) Unsleber, J. P.; Reiher, M. The Exploration of Chemical Reaction Networks. *Ann. Rev. Phys. Chem.* **2020**, *71*, 121–142.

(22) Dartois, E.; Charon, E.; Engrand, C.; Pino, T.; Sandt, C. Mechanochemical Synthesis of Aromatic Infrared Band Carriers. *A&A* **2020**, *637*, A82.

(23) Wales, D. J. Exploring Energy Landscapes. *Ann. Rev. Phys. Chem.* **2018**, *69*, 401–425.

(24) Wales, D. J.; Doye, J. P. K. Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms. *J. Phys. Chem. A* **1997**, *101*, 5111–5116.

(25) Martínez-Núñez, E. An Automated Transition State Search Using Classical Trajectories Initialized at Multiple Minima. *Phys. Chem. Chem. Phys.* **2015**, *17*, 14912–14921.

(26) Martínez-Núñez, E. An Automated Method to Find Transition States Using Chemical Dynamics Simulations. *J. Comp. Chem.* **2015**, *36*, 222–234.

(27) Martínez-Núñez, E.; Barnes, G. L.; Glowacki, D. R.; Kopec, S.; Peláez, D.; Rodríguez, A.; Rodríguez-Fernández, R.; Shannon, R. J.; Stewart, J. J. P.; Tahoces, P. G.; Vazquez, S. A. AutoMeKin2021 : An Open-source Program for Automated Reaction Discovery. *J. Comp. Chem.* **2021**, *42*, 2036–2048.

(28) Furman, D.; Naumkin, F.; Wales, D. J. Energy Landscapes of Carbon Clusters from Tight-Binding Quantum Potentials. *J. Phys. Chem. A* **2022**, *126*, 2342–2352.

(29) Calvo, F.; Simon, A.; Parneix, P.; Falvo, C.; Dubosq, C. Infrared Spectroscopy of Chemically Diverse Carbon Clusters: A Data-Driven Approach. *J. Phys. Chem. A* **2021**, *125*, 5509–5518.

(30) Dubosq, C.; Pla, P.; Dartois, E.; Simon, A. Spectroscopic Investigation of Interstellar Hydrogenated Carbon Clusters: Infrared and Optical Spectra of C $_{24}$ H $_{n}$ Populations. *A&A* **2023**, *670*, A175.

(31) Laranjeira, J.; Strutyński, K.; Marques, L.; Martínez-Núñez, E.; Melle-Franco, M. C60+C60 Molecular Bonding Revisited and Expanded. *Carbon* **2023**, *213*, 118209.

(32) Sure, R.; Hansen, A.; Schwerdtfeger, P.; Grimme, S. Comprehensive Theoretical Study of All 1812 C60 Isomers. *Phys. Chem. Chem. Phys.* **2017**, *19*, 14296–14305.

(33) Li, W.; Zhao, L.; Kaiser, R. I. A Unified Reaction Network on the Formation of Five-Membered Ringed Polycyclic Aromatic Hydrocarbons (PAHs) and Their Role in Ring Expansion Processes through Radical–Radical Reactions. *Phys. Chem. Chem. Phys.* **2023**, *25*, 4141–4150.

(34) Zhao, L.; Kaiser, R. I.; Lu, W.; Xu, B.; Ahmed, M.; Morozov, A. N.; Mebel, A. M.; Howlader, A. H.; Wnuk, S. F. Molecular Mass Growth through Ring Expansion in Polycyclic Aromatic Hydrocarbons via Radical–Radical Reactions. *Nat. Commun* **2019**, *10*, 3689.

(35) Hjorth Larsen, A. et al. The Atomic Simulation Environment—a Python Library for Working with Atoms. *J. Phys. Condens. Matter* **2017**, *29*, 273002.

(36) Neo4j. 2024; https://neo4j.com/, Graph database management system, available at https://neo4j.com/.

(37) Panadés-Barrueta, R. L.; Duflot, D.; Soto, J.; Martínez-Núñez, E.; Peláez, D. Automatic Determination of the Non-Covalent Stable Conformations of the NO2-Pyrene Cluster in Full Dimensionality (81D) Using the vdW-TSSCDS Approach. *ChemPhysChem* **2024**, *25*, e202301001.

(38) Ismail, I.; Chantreau Majerus, R.; Habershon, S. Graph-Driven Reaction Discovery: Progress, Challenges, and Future Opportunities. *J. Phys. Chem. A* **2022**, *126*, 7051–7069.

(39) Beineke, L. W. Chemical Applications of Graph Theory (A. T. Balaban, ed.). *SIAM REV.* **1978**, *20*, 869–870.

(40) Kotobi, A.; Singh, K.; Höche, D.; Bari, S.; Meißner, R. H.; Bande, A. Integrating Explainability into Graph Neural Network Models for the Prediction of X-ray Absorption Spectra. *J. Am. Chem. Soc.* **2023**, *145*, 22584–22598.

(41) Zhang, B.; Zhang, X.; Du, W.; Song, Z.; Zhang, G.; Zhang, G.; Wang, Y.; Chen, X.; Jiang, J.; Luo, Y. Chemistry-Informed Molecular Graph as Reaction Descriptor for Machine-Learned Retrosynthesis Planning. *Proc. Natl. Acad. Sci. U.S.A.* **2022**, *119*, e2212711119.

(42) Fang, Y.; Zhang, Q.; Zhang, N.; Chen, Z.; Zhuang, X.; Shao, X.; Fan, X.; Chen, H. Knowledge Graph-Enhanced Molecular Contrastive Learning with Functional Prompt. *Nat. Mach. Intell.* **2023**, *5*, 542–553.

(43) Matter, H.; Buning, C.; Stefanescu, D. D.; Ruf, S.; Hessler, G. Using Graph Databases to Investigate Trends in Structure–Activity Relationship Networks. *J. Chem. Inf. Model.* **2020**, *60*, 6120–6134.

(44) Rangarajan, S. Towards a Chemistry-Informed Paradigm for Designing Molecules. *Curr. Opin. Chem. Eng.* **2022**, *35*, 100717.

(45) Walters, W. P.; Barzilay, R. Applications of Deep Learning in Molecule Generation and Molecular Property Prediction. *Acc. Chem. Res.* **2021**, *54*, 263–270.

(46) Stauch, T.; Dreuw, A. Advances in Quantum Mechanochemistry: Electronic Structure Methods and Force Analysis. *Chem. Rev.* **2016**, *116*, 14137–14180, PMID: 27767298.

(47) De Chavez, D.; Hasegawa, J.-y. OpenMechanochem: A Python Module for Mechanochemical Simulations. *SoftwareX* **2021**, *16*, 100879.

(48) Weiss, P. S. A push for mechanochemistry. *Science* **2023**, *380*, 1013–1013.

(49) Brenner, D. W.; Shenderova, O. A.; Harrison, J. A.; Stuart, S. J.; Ni, B.; Sinnott, S. B. A Second-Generation Reactive Empirical Bond Order (REBO) Potential Energy Expression for Hydrocarbons. *J. Phys. Condens. Matter* **2002**, *14*, 783–802.

(50) Pla, P.; Dubosq, C.; Rapacioli, M.; Posenitskiy, E.; Alcamí, M.; Simon, A. Hydrogenation of $C_{24}$ Carbon Clusters: Structural Diversity and Energetic Properties. *J. Phys. Chem. A* **2021**, *125*, 5273–5288.

(51) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864–2875.

(52) Bhuiyan, F. H.; Li, Y.-S.; Kim, S. H.; Martini, A. Shear-Activation of Mechanochemical Reactions through Molecular Deformation. *Sci. Rep.* **2024**, *14*, 2992.

(53) Hagberg, A. A.; Schult, D. A.; Swart, P. J. Exploring Network Structure, Dynamics, and Function Using NetworkX. Python in Science Conference. Pasadena, California, 2008; pp 11–15.

(54) Vázquez, S. A.; Martínez-Núñez, E.; Preston, T. J. Exploring Unimolecular Reactions in Disilanol and Ethanol: Insights and Challenges. *Theor. Chem. Acc.* **2023**, *142*, 124.

(55) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods VI: More Modifications to the NDDO Approximations and Re-Optimization of Parameters. *J. Mol. Model.* **2013**, *19*, 1–32.

(56) Ishigaki, Y.; Shimajiri, T.; Takeda, T.; Katoono, R.; Suzuki, T. Longest C–C Single Bond among Neutral Hydrocarbons with a Bond Length beyond 1.8 Å. *Chem* **2018**, *4*, 795–806.

(57) Hou, Y.-F.; Wang, C.; Dral, P. O. Accurate and Affordable Simulation of Molecular Infrared Spectra with AIQM Models. 2024.

(58) Dubosq, C.; Falvo, C.; Calvo, F.; Rapacioli, M.; Parneix, P.; Pino, T.; Simon, A. Mapping the Structural Diversity of $C_{60}$ Carbon Clusters and Their Infrared Spectra. *A&A* **2019**, *625*, L11.

(59) Bonnin, M. A.; Falvo, C.; Calvo, F.; Pino, T.; Parneix, P. Simulating the Structural Diversity of Carbon Clusters across the Planar-to-Fullerene Transition. *Phys. Rev. A* **2019**, *99*, 042504.

(60) Todeschini, R.; Consonni, V. In *Handbook of Chemoinformatics*, 1st ed.; Gasteiger, J., Ed.; Wiley, 2003; pp 1004–1033.

(61) RDKit Development Team RDKit: Open-Source Cheminformatics. https://www.rdkit.org, 2024; Accessed: 2024-03-05.

(62) Siegel, J. S. Brevity for Bonds. *Nature* **2006**, *439*, 801–802.

(63) Vishnevskiy, Y. V.; Otlyotov, A. A.; Lamm, J.-H.; Stammler, H.-G.; Girichev, G. V.; Mitzel, N. W. Accurate Single Crystal and Gas-Phase Molecular Structures of Acenaphthene: A Starting Point in the Search for the Longest C–C Bond. *Phys. Chem. Chem. Phys.* **2023**, *25*, 11464–11476.

(64) Freedman, D.; Diaconis, P. On the Histogram as a Density Estimator:L2 Theory. *Z. Wahrscheinlichkeitstheorie verw Gebiete* **1981**, *57*, 453–476.

(65) Burnham, K. P.; Anderson, D. R.; Anderson, D. R. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed.; Springer, 2010.

(66) Machine learning in python - scikit-learn 0.16.1 documentation. https://scikit-learn.org/, available at https://scikit-learn.org/.

(67) Garner, M. H.; Bro-Jø rgensen, W.; Pedersen, P. D.; Solomon, G. C. Reverse Bond-Length Alternation in Cumulenes: Candidates for Increasing Electronic Transmission with Length. *J. Phys. Chem. C* **2018**, *122*, 26777–26789.

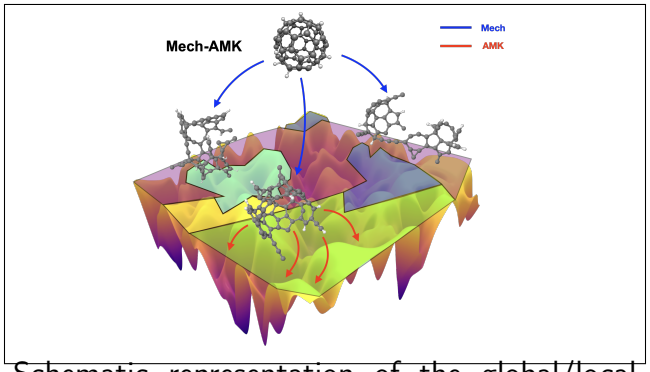
## Acknowledgement

The first author gratefully acknowledges the China Scholarship Council (CSC) for his Ph.D. funding as well as Institut des Sciences Moléculaires d'Orsay (ISMO), Université Paris-Saclay and the CNRS for financial support. This work was performed using computational resources from the "Mésocentre" computing center of Université Paris-Saclay, CentraleSupélec and École Normale Supérieure Paris-Saclay supported by CNRS and Région Île-de-France (https://mesocentre.universite-paris-saclay.fr/). The authors would like to thank kindly P. Çarçabal, M. Leoni, L. Coudert, J.-Y. Bazzara and A. Borissov for their kind computational support with the cluster.

## Supporting Information Available

An detailed description of the molecular descriptors employed for principal component analysis (PCA), accompanied by the resulting PCA outcomes and supplementary tables and figures depicting the dataset's distribution.

# TOC Graphic




Schematic representation of the global/local search using Mech-AMK.