Digital Discovery

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2024, **3**, 2298

MolBar: a molecular identifier for inorganic and organic molecules with full support of stereoisomerism†

Nils van Staalduinen and Christoph Bannwarth *

Before a new molecular structure is registered to a chemical structure database, a duplicate check is essential to ensure the integrity of the database. The Simplified Molecular Input Line Entry Specification (SMILES) and the IUPAC International Chemical Identifier (InChI) stand out as widely used molecular identifiers for these checks. Notable limitations arise when dealing with molecules from inorganic chemistry or structures characterized by non-central stereochemistry. When the stereoinformation needs to be assigned to a group of atoms, widely used identifiers cannot describe axial and planar chirality due to the atom-centered description of a molecule. To address this limitation, we introduce a novel chemical identifier called the Molecular Barcode (MolBar). Motivated by the field of theoretical chemistry, a fragment-based approach is used in addition to the conventional atomistic description. In this approach, the 3D structure of fragments is normalized using a specialized force field and characterized by physically inspired matrices derived solely from atomic positions. The resulting permutation-invariant representation is constructed from the eigenvalue spectra, providing comprehensive information on both bonding and stereochemistry. The robustness of MolBar is demonstrated through duplication and permutation invariance tests on the Molecule3D dataset of 3.9 million molecules. A Python implementation is available as open source and can be installed *via pip install molbar*.

Received 30th June 2024
Accepted 24th September 2024

DOI: 10.1039/d4dd00208c

rsc.li/digitaldiscovery

1 Introduction

When enrolling a compound in a chemical database system or registry, it is customary to initially verify the novelty of its structure. Checking for duplicates, the molecular structure of the compound to be registered is compared with the structures already present in the database. To represent the molecular structure, connection table (CT) formats, which have been used since the early days of cheminformatics, are the most widely used in computer systems.[1–3] The molecular structure is represented as an undirected graph, with the atoms serving as nodes and the bonds as edges.[4] This approach goes back to the pioneering work of August Kekulé[5] and Gilbert N. Lewis,[6] who introduced the structural formula as a mathematical graph for chemical structures. The graphical representation of molecules has the advantage that duplicate identification is equivalent to determining whether the two graph representations are identical, a concept known as isomorphism in graph theory. Since most applications of chemoinformatics relate to organic molecules with well-defined covalent bonds, graph representation is a practical choice.[4] Classical cheminformatics representations rely on alphanumeric string line notations, such as SMILES[7] and InChI,[8] to only name two representations, which are derived from the CT representation. Line notations are widely used in cheminformatics and are supported by most chemical database systems. Several chemical line notation systems were proposed in the 1950s and 1960s, but the Wiswesser Line-formula Notation (WLN), introduced in 1949, became the most widely used.[9–11] The WLN remained popular until the early 1980s, when more flexible systems such as SMILES began to replace it. The WLN used simple digits 1–9 to represent unbranched alkyl chains and capital letters to denote individual atoms or common substructures, making it a compact and efficient system. However, the WLN eventually fell out of favor

Institute of Physical Chemistry, RWTH Aachen University, Melatener Str. 20, 52074 Aachen, Germany. E-mail: bannwarth@pc.rwth-aachen.de

† Electronic supplementary information (ESI) available: The ESI encompasses the first and second derivatives of the bond, angle, dihedral, and Coulomb terms of the unification force field. It also contains the specified ideal geometries for matching the actual structure to a reference structure (SI.pdf). Additional ESI material further contains the full and the filtered benchmark dataset as pickle files with 3D coordinates of the molecules (*molecule3D.pkl* and *filtered_molecule3D.pkl*). Further, each dataset entry discussed in this work is presented as XYZ files, identifiable by their ID of the Molecule3D dataset. Additionally, the data from permutation test is provided (*permutation_data.pkl*) and from the timings (*timings.pkl*). The ESI material contains the Python code for generating the identifiers (*generate_identifiers.py*) and running the tests (*run_duplication_test.py*, *run_permutation_test.py*, *run_timing_test.py*). Finally, the MolBar package in the version used in this work is provided as a ZIP file. All if this additional data can be obtained from **https://zenodo.org/records/13827787**. See DOI: **https://doi.org/10.1039/d4dd00208c**

due to its inability to represent complex features such as stereochemistry and its reliance on rigid, complicated rules.[12] The Simplified Molecular Input Line Entry System (SMILES) was created by David Weininger in 1986 and is based on a few basic rules.[7] Molecules are depicted by a chain of connected atoms. The atoms themselves are denoted by their atomic symbols. Hydrogen atoms are omitted, since free valences are assumed to be saturated with hydrogen atoms. Only double and triple bonds are portrayed by their symbols = and #. Atoms in aromatic substructures are expressed by lowercase atom symbols. Branches are illustrated by brackets around the branch, and digits are used to denote ring closures. In addition, stereochemistry is represented by the symbols @ and @@ for tetrahedral chiral centers and / and \ for double bonds. SYBYL Line Notation (SLN), originally inspired by SMILES, has since diverged significantly from it.[13,14] A key difference lies in the treatment of valence and aromaticity. SMILES allows implicit valences as part of the language structure, while SLN makes no assumptions about the valence of an atom. In addition, aromaticity is treated differently: SMILES treats aromaticity as an atomic property, while SLN explicitly assigns aromaticity to bonds. Another notation system, the Representation Of Structure Diagram Arranged Linearly (ROSDAL), was developed facilitating searches in the Beilstein database.[15] The IUPAC International Chemical Identifier (InChI) is a non-proprietary identifier for chemical substances, developed since 2000 and nowadays used in many chemical database systems.[8] The InChI is generated based on a workflow that includes normalization, followed by a canonicalization and finally a serialization of the structure. Normalization intends to unify different input representations of single compounds, such as different tautomeric or mesomeric structures by applying a consistent chemical model. Canonicalization ensures a unique numbering of the atoms to be independent of the order of the input atoms. Finally, the canonicalized structure is serialized to a string, which is the InChI. Serialization refers to the process of converting the labeled atoms to a string line notation. The InChI is a hierarchical identifier, which consists of several layers, each of which is separated by a forward slash delimiter. The main layer consists of the chemical formula of the compound, the connection layer **/c** and the hydrogen layer **/h**. Further layers include the charge layer (charge sublayer **/q** and proton sublayer **/p**), the stereochemical layer representing double bond stereoconfiguration (**/b**) and tetrahedral stereochemistry (**/t**). Non-standard InChI can be used to distinguish between tautomers by the fixed-H layer (**/f**), for example.

While connection tables like Molfiles or SDFiles are extensively employed in cheminformatics, their utilization is not as prevalent in the realm of quantum chemistry.[16] In quantum chemistry, the molecular structure is typically conveyed through a list of atoms along with their Cartesian coordinates without any explicit information about the connectivity of the atoms. Recent algorithmic advances have facilitated the automatic generation of molecular structures from quantum chemical calculations such as *ab initio* reaction and molecule discovery workflows,[17–19] conformer ensemble sampling tools[20] or *ab initio* electron ionization mass spectra,[21,22] pinpointing novel structures as local minima on the potential energy surface represented by their Cartesian coordinates. However, registering these structures in a quantum chemical database system would then require relying on additional pre-processing software to assess the connectivity of the atoms before generating an identifier. More critically, these black-box tools can generate structures that are drastically different from the input. These cases are not always related to changing stereoinformation, as the resulting structures may have chemically unreasonable geometries. In other cases, the VSEPR geometry of metal centers may change from square-planar to tetrahedral due to an insufficient level of theory to describe the underlying electronic structure. Automated information about molecular shape changes is crucial for evaluating the correctness of the vast number of generated structures. Furthermore, existing approaches reach their limits when it comes to supporting all possible inorganic stereochemistries and accounting for non-central chirality in (in)organic molecules. While SMILES supports an incomplete set of inorganic stereochemistry,[23] InChI separates all bonds to metals, resulting in a complete loss of such information.[8] Both SMILES and InChI take an atom-centric Lewis picture, making them unable to effectively represent non-central chirality, such as axial or planar chirality, where the stereochemistry information may relate to a group of atoms rather than a single atom.[24]

2 Design goals of the molecular barcode

This motivates the introduction of a new concept for a molecular identifier, inspired by the principles of electronic structure theory and optimized for quantum chemical structure databases and chemical space exploration with electronic structure theory methods. To fulfill the criteria for use in computational chemistry, the identifier must effectively differentiate between structures that are separated by a reasonably high potential energy surface barrier. Consequently, the identifier must not only distinguish between constitutional isomers and stereoisomers, but by default also between (prototropic) tautomers, explicitly considering hydrogen atoms. The identifier must retain information about the molecular shape, *e.g.*, including VSEPR geometries, but at the same time, different conformations should be mapped to a single identifier. To achieve this, the identifier should integrate an evaluation of molecular topology using a black-box approach. This ensures clarity and prevents divergent interpretations of atomic connectivity. This level of distinction is crucial for the individual registration of these structures in quantum chemical database systems. Such precision facilitates computational studies, as each minimum on the potential energy surface possesses unique and distinguishable properties from electronic structure theory. However, the identifier should follow a hierarchical approach with different levels to categorize the relationship between different stored structures, *e.g.* tautomerism, which is crucial for the evaluation of the correctness of the generated structures. For comprehensive utility in organic and inorganic molecular computational studies, the identifier must describe all forms of stereochemistry, including tetrahedral/center, axial, and planar chirality. In addition to an atom-centric approach, the identifier should adopt a fragment-centric perspective capable of characterizing the stereochemistry of a group of atoms.

3 Theory of the molecular barcode

3.1 Topology: describing atomic connectivity

The description of molecular topology is a crucial first step in the creation of a molecular identifier. The covalent bonds determine how atoms are connected to each other and how they share their electrons, and form the basis of a molecule's identity.

The specification of bonds already allows to differentiate between constitutional isomers, since their differences are based on different connectivity of atoms (Fig. 1). As discussed before, simple structure diagrams of molecules can be interpreted mathematically as graphs. A set of vertices $v_i \in V_{\text{atoms}}$ and a set of edges $e_{ij} \in E_{\text{atoms}}$ is called a graph $G_{\text{atoms}} = (V_{\text{atoms}}, E_{\text{atoms}})$. The adjacency matrix $\mathbf{A} \in \mathbb{R}^{n_{\text{atoms}} \times n_{\text{atoms}}}$ is a mathematical representation of a graph G, which describes, in analogy to the Hückel matrix, if two atoms are adjacent. If both atoms v_i and v_j are connected by a bond e_{ij}, the matrix element a_{ij} is one (eqn (1)).

$$a_{ij} := \begin{cases} 0 & \forall\, i = j \\ 1 & \forall\, i \neq j \wedge e_{ij} \in E(G) \\ 0 & \forall\, i \neq j \wedge e_{ij} \notin E(G) \end{cases} \tag{1}$$

The disadvantage of the matrix graph representation is that its size grows quadratically with the system size and the matrix itself is not permutation invariant with respect to the atomic order. Interchanging atoms is equivalent to interchanging rows and columns in the adjacency matrix, resulting in a different matrix representation. However, coming back to the analogy of the Hückel matrix, the sorted eigenvalue spectrum of such a matrix is independent of the atomic order. By calculating the spectrum of an adjacency matrix in general, the molecular topology can be represented in an *a priori* permutation invariant vector.[25,26] Furthermore, this representation does not grow quadratically with system size and corresponds to a desired line notation. Yet, a simple adjacency matrix contains no information about the nature of the involved atoms. If one element were replaced by another, the adjacency matrix would remain the same and so would the spectrum. Therefore, a modified adjacency matrix must include atomic information. For a topology representation, we introduce a modified extended adjacency matrix that is inspired by the Hückel matrix. In analogy, the diagonal elements of the modified extended adjacency matrix thus contain atomic information in form of the atomic number and the coordination number of the atom to represent its local environment. The coordination number CN_i is the number of bonds in which atom v_i is involved. If two atoms are bonded, the respective off-diagonal element is the arithmetic mean of the respective diagonal elements. The matrix elements of the topology matrix $\mathbf{H}_{\text{topology}}$ are given by:

$$h_{ij} := \begin{cases} (\text{CN}_i + 1)Z_i & \forall\, i = j \\ \frac{1}{2}\left(h_{ii} + h_{jj}\right) & \forall\, i \neq j \wedge e_{ij} \in E(G) \\ 0 & \forall\, i \neq j \wedge e_{ij} \notin E(G) \end{cases} \tag{2}$$

where Z_i represents the nuclear charge of the i-th atom, and h_{ii} and h_{jj} are the diagonal matrix elements corresponding to the i-th and j-th atoms, respectively. The generation of the molecular graph in the form of $\mathbf{H}_{\text{topology}}$ for a given molecule and its diagonalization to obtain its eigenvalue spectrum $\sigma(\mathbf{H}_{\text{topology}}) \in \mathbb{R}^{n_{\text{atoms}}}$ leads to a representation of the molecular topology that is invariant to the atomic order.[25,26] To obtain a spectrum composed only of integers, we multiply the eigenvalues by a factor of ten and round them to the nearest integer values.

3.2 Topography: describing spatial atomic arrangement

While the topology is sufficient for the distinction of constitutional isomers, stereoisomers require additional criteria, since they do not differ in their atomic connectivity. For this purpose, the atoms' spatial arrangement has to be taken into account. The adjacency matrix does not contain any information about the 3D position of atoms in space. Two molecules with the same topology that differ in their relative configurations, such as in Fig. 2, cannot be distinguished as their topological spectra are the same. In the field of molecular machine learning, 3D representations have been used to describe the shape of the molecule and correlate them with various molecular

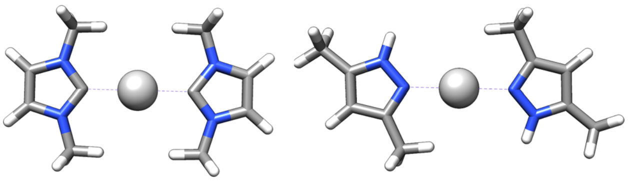


Fig. 1 Example of two constitutional isomers of a silver complex. Both structures differ by different connected atoms while sharing the same molecular formula.

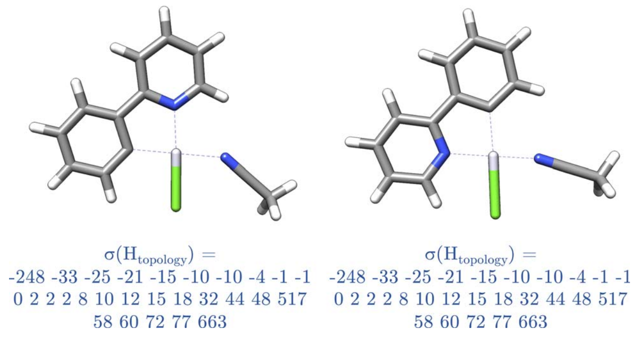


Fig. 2 Configurational isomers of two platinum complexes. Both complexes have the same atomic connectivity as indicated by the same eigenvalue spectrum $\sigma(\text{H}_{\text{topology}})$ of eqn (2), but differ generally in the relative position of the nitrogen atoms (blue-colored atoms). The eigenvalue spectrum was rounded to integer values for clarity reasons.

© 2024 The Author(s). Published by the Royal Society of Chemistry

properties.[29–31] 3D representations are derived from the molecular structure as an object in Cartesian space and include information about the relative positions of the atoms to each other. Those representations can be divided into local (nearer atomic vicinity) and global (considering the whole molecule) types of representations.[32] For a molecular identifier, in general, global representations are of interest as they capture the shape of molecules as a whole as they describe the long-range interactions.

A global 3D descriptor was introduced in the work of Rupp *et al.* in the context of machine learning for molecular atomization energies. The Coulomb matrix (CM) representation is defined by:[27]

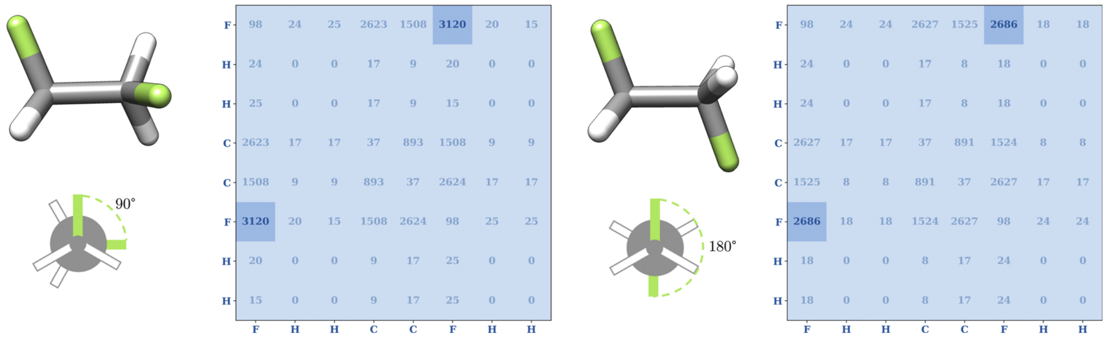
$$m_{ij} := \begin{cases} 0.5Z_i^{2.4} & \forall i = j \\ \dfrac{Z_i Z_j}{R_{ij}} & \forall i \neq j \end{cases} \tag{3}$$

The diagonal elements describe a polynomial fit of atomic energies to the nuclear charge Z_i. The off-diagonal elements of the matrix contain the Coulomb repulsion operator with the inverse of the Cartesian distance R_{ij} between pairs of atoms, allowing the characterization of the molecular shape. Since the relative positions, *e.g.*, of the nitrogen atoms differ for the complexes in Fig. 2, resulting in different interatomic distances and molecular shape, such a representation is able to distinguish between the two. In literature, it has been proven to be effective in describing constitutional isomers and diastereomers with the spectrum of the matrix as a permutation-invariant descriptor for the molecular shape.[33] However, it is important to note that this matrix is not invariant in conformational space and thus impractical to be used as molecular identifier on its own. To yield a conformer-independent representation, modifications to eqn (3) are required. Overall, conformers exhibit variations primarily due to rotations around single bonds, resulting in different Coulomb spectra with evolving off-diagonal matrix elements within eqn (3) as the interatomic distance changes. This is illustrated by the highlighted off-diagonal matrix elements in Fig. 3 for two conformers of 1,2-difluoroethane.

Nevertheless, the particular spatial orientation of these fluorine atoms with respect to each other has no significance for the identification of the molecule as 1,2-difluoroethane. This is because the transition between conformers is rapid and primarily reflects conformational changes rather than changes in the fundamental identity of the molecule. A simple strategy to achieve conformational independence is to allow interatomic interactions in the CM matrix only if the relative distances between the atoms remain unchanged under rotations around the respective single bonds. Therefore, we introduce a matrix similar to the eqn (2) by including the inverse atomic distances (eqn (4)). However, this CM matrix contains only off-diagonal nonzero elements between atoms that form rigid substructures within the molecule and will be denoted by the indicator function $I(v_i, v_j)$ in the following. If two atoms belong to the same rigid fragment, the indicator function yields $I(v_i, v_j) = 1$. Conversely, if the spatial distance between two atoms in the conformational space changes, the indicator function is $I(v_i, v_j) = 0$. These rigid fragments are thus defined as substructures within the molecule if the relative positions of the atoms can be consistently unified into identical relative three-dimensional positions regardless of the current conformation of the whole molecule. The expression for the indicator function and the structure unification workflow will be discussed in Section 4. The matrix elements of $\mathbf{B}_{\text{topography}}$ are given by the following expression

$$b_{ij} := \begin{cases} (\mathrm{CN}_i + 1)Z_i & \forall i = j \\ \dfrac{1}{2}\left(b_{ii} + b_{jj}\right) \Big/ R_{ij} & \forall i \neq j \wedge I\left(v_i, v_j\right) = 1 \\ 0 & \forall i \neq j \wedge I\left(v_i, v_j\right) = 0 \end{cases} \tag{4}$$

where b_{ii} and b_{jj} are the diagonal matrix elements. Unlike to the global 3D representation in eqn (3), eqn (4) ranges from being a global to local representation, depending on the size of the rigid fragments. The topography of a molecule is then described by diagonalizing $\mathbf{B}_{\text{topography}} \in \mathbb{R}^{n_{\text{atoms}} \times n_{\text{atoms}}}$ to obtain its spectrum $\sigma(\mathbf{B}_{\text{topography}}) \in \mathbb{R}^{n_{\text{atoms}}}$, which leads to a vector representation of the molecular topography that is invariant to the atomic order and invariant to rotation and translation in Cartesian space.[26,27,33]



	F	H	H	C	C	F	H	H
F	98	24	25	2623	1508	3120	20	15
H	24	0	0	17	9	20	0	0
H	25	0	0	17	9	15	0	0
C	2623	17	17	37	893	1508	9	9
C	1508	9	9	893	37	2624	17	17
F	3120	20	15	1508	2624	98	25	25
H	20	0	0	9	17	25	0	0
H	15	0	0	9	17	25	0	0

	F	H	H	C	C	F	H	H
F	98	24	24	2627	1525	2686	18	18
H	24	0	0	17	8	18	0	0
H	24	0	0	17	8	18	0	0
C	2627	17	17	37	891	1524	8	8
C	1525	8	8	891	37	2627	17	17
F	2686	18	18	1524	2627	98	24	24
H	18	0	0	8	17	24	0	0
H	18	0	0	8	17	24	0	0

Fig. 3 Conformational dependence of the Coulomb matrix in eqn (3) (ref. 27) for two conformers of 1,2-difluoroethane optimized at the GFN2-xTB level of theory.[28] Rotation around the C–C single bond changes the distance between the two fluorine atoms, leading to different off-diagonal matrix elements (highlighted in darker blue).

3.3 Chirality: describing absolute configuration

Chirality is of extraordinary importance in various fields of chemistry and biochemistry, as it strongly influences molecular properties.[34] Enantiomers possess the ability to have different pharmacological activities, variations in chemical reactivity, and exhibit different circular dichroism spectra.[34] Chirality occurs in a variety of forms, with chiral carbon atoms being the most common (*c.f.* Fig. 4a), although the center of chirality does not to be located on a single atom, as illustrated by the example of tetrasubstituted adamantane (*c.f.* Fig. 4b).[34]

In addition, chirality can originate either from an axis, where atoms arrange sequentially clockwise or counterclockwise around an axis (*c.f.* Fig. 4c) or from a plane with atoms located on one side or the other with respect to a chiral plane (*c.f.* Fig. 4d).[34] Chiral molecules exhibit rigidity within the specific molecular segment that defines the chirality – whether it is a chiral center, axis, or plane. This property can be exploited by the introduced concept of rigid fragments in eqn (4), which only allows interactions between atoms that are within the boundaries of the same rigid fragment. To characterize the absolute configuration for the whole molecule, the inherent rigidity of these fragments can be effectively utilized by solely characterizing the absolute configuration of these fragments. In general, two enantiomers have the same topology and topography (after structure unification), but cannot be superimposed. As a result, the spectra of the topology matrix of eqn (2) (blue spectrum in Fig. 5) and of the topography matrix in eqn (4) (red spectrum in Fig. 5) are identical.

Therefore, an additional matrix is needed to fully describe chirality. To design an identifier that also covers axial and planar chirality, a matrix of size $n_{\text{atoms}} \times n_{\text{atoms}}$ describing only atomic properties is not sufficient, since in these cases the chiral properties cannot be assigned to individual atoms. Therefore, an alternative to the atomic matrix based on rigid fragments is obligatory. This matrix has the dimensions of $n_{\text{fragments}} \times n_{\text{fragments}}$, where $n_{\text{fragments}}$ indicates the number of rigid fragments in the molecule.

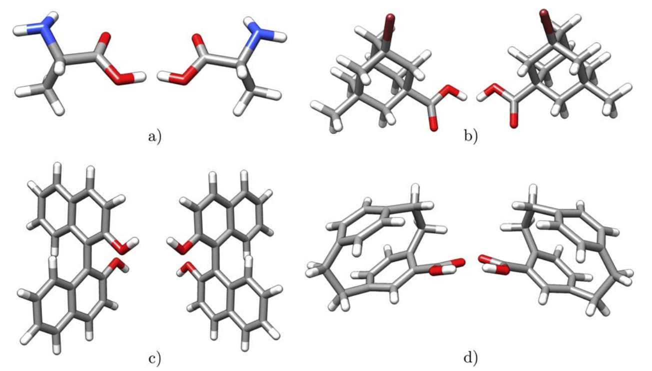


Fig. 4 Examples of chiral molecules with different types of stereogenic units: (a) alanine (central chirality) (b) 3-bromo-5-methyl-adamantane-1-carboxylic acid (central chirality) (c) [1,1′-binaphthalene]-2,2′-diol (axial chirality) (d) 4-carboxy[2.2]para-cyclophane (planar chirality).

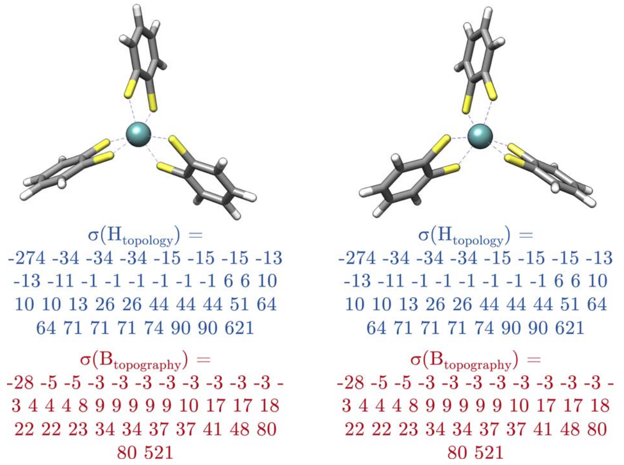


Fig. 5 Two enantiomers of a technetium complex exhibiting axial chirality. Both complexes have identical topology and topography, so they cannot be distinguished based on the spectra of the matrices of the eqn (2) (blue spectrum) and (4) (red spectrum). The eigenvalue spectra were rounded to integer values for clarity reasons.

In this matrix, the diagonal elements describe the absolute configuration of the individual fragments, while the off-diagonal elements represent the relationships between the (a) chiral fragments. In this context, the variable $G_a \in \{-1, 0, 1\}$ serves as a chirality index and determines whether fragment a is achiral (0) or one of its enantiomers (−1 or 1). An expression for G_a is discussed in Section 4. The parameter p_a denotes the priority of the fragment a, analogous to the Cahn–Ingold–Prelog rules but extended to the entire fragment in the molecule. This parameter characterizes each unique fragment in the molecule. More details on the concept behind the fragment priority can be found in Section 4. Since the relative arrangements of rigid fragments can change due to conformational changes, d_{ab} denotes the graph distance between two fragments, which is the number of edges along the shortest path between them.

Diagonalizing $\boldsymbol{G}_{\text{chirality}}$, we obtain the spectrum $\sigma(\boldsymbol{G}_{\text{chirality}}) \in \mathbb{R}^{n_{\text{fragments}}}$, which decomposes the absolute configuration of the molecule into contributions of different fragments and is not restricted to describing a chiral center located on an atom. Following the eqn (4), we hereby introduce the absolute configuration matrix tailored to these fragments, denoted $\boldsymbol{G}_{\text{chirality}}$, where g_{aa} and g_{bb} are the diagonal matrix elements.

$$g_{ab} = \begin{cases} G_a p_a & \forall a = b \\ \frac{1}{2}(g_{aa} + g_{bb}) \Big/ d_{ab} & \forall a \neq b \end{cases} \tag{5}$$

4 Implementation of the molecular barcode

In a nutshell, MolBar uses a set of matrices, more precisely their rounded eigenvalue spectra, to represent the topology, topography of the molecule as well as the absolute configuration of fragments. The latter two matrices rely on a partitioning procedure that decomposes the molecule into rigid substructures. This facilitates the representation of the molecule as a 3D object consistent within the conformational space. In the following sections, the comprehensive MolBar workflow (Fig. 6) and the algorithms used to obtain the 3D molecular information required for matrix construction are discussed in more detail. The *molbar* implementation is provided *via* an open-source *Python* module that is partially written in *Fortran* to handle resource-intensive tasks. The program package and its documentation are available at **https://git.rwth-aachen.de/bannwarthlab/molbar**. The *molbar* package is installable *via* the Python package index (PyPI) using the command *pip install molbar*. The package takes 3D coordinates of the molecule with explicit hydrogen atoms as input, supporting various file formats as defined in the documentation.

4.1 Step 1: topology

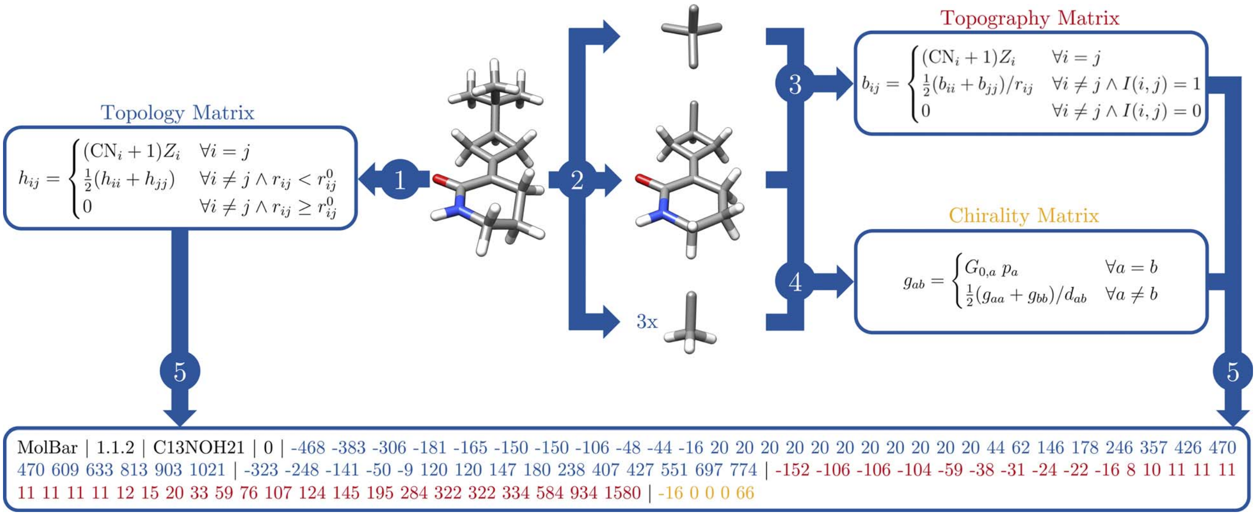
To construct the topology matrix according to eqn (2), it is important to determine which atoms are bonded. Starting from 3D coordinates, the connectivity evaluation can be carried out by comparing the Cartesian distance r_{ij} between the two atoms with a certain reference value r^0_{ij}.[35] For a given molecular geometry represented by Cartesian coordinates $\mathbf{X}$ in 3D space, a molecular graph G can therefore be constructed:

$$E = \{(v_i,v_j)|v_i,v_j \in V_{\text{atoms}} \wedge r_{ij} < r^0_{ij}\} \tag{6}$$

This reference value r^0_{ij} can be a simple scaled summation of the two tabulated covalent radii, as used in the DFT-D3 dispersion,[37] or a more elaborate expression that takes into account the chemical environment of the individual atoms, as used in the force field GFN-FF by the Grimme group.[35] While the former approach is found to be appropriate for organic chemistry molecules, the latter approach was chosen for the *molbar* implementation. This choice allows for a more satisfactory treatment of metal complex systems. Here, the precalculated reference value r^0_{ij} describes the usual bonding distance between these two atoms in their actual environment. Those values had been fitted to reproduce equilibrium bond lengths at the PBEh-3c[38]/B97-3c[39] level of theory. More information can be found in the ESI† of ref. 35. This force field has been successfully applied in computational studies across a wide range of chemical fields, including metal–organic frameworks,[40] transition metal complexes,[41] and supramolecular systems and biological macromolecules.[42] These applications demonstrate the force field's reliability in accurately defining molecular topologies, which is why it has been adapted as a standalone implementation within MolBar. By implementation of the same bond identification routine, MolBar does not rely on external packages to determine the topology in molecules. Furthermore, only atomic information such as the coordination number CN_i of atom v_i and its nuclear charge Z_i are required. The nuclear charges are tabulated for each element and are assigned by the element information specified in the input file.

4.2 Step 2.1: fragmentation

The introduction of the concept of rigid fragments in eqn (4) allows a consistent representation of the atomic 3D positions in conformational space. Consequently, a rigidity analysis must be conducted prior to constructing the topography and absolute configuration matrix in eqn (4) and (5). For that, in the following, the indicator function $I(v_i, v_j)$ is defined that states if two atoms are part of the same fragment.


Topography Matrix
$b_{ij} = \begin{cases} (\text{CN}_i+1)Z_i & \forall i=j \\ \frac{1}{2}(b_{ii}+b_{jj})/r_{ij} & \forall i \neq j \wedge I(i,j)=1 \\ 0 & \forall i \neq j \wedge I(i,j)=0 \end{cases}$
Topology Matrix
$h_{ij} = \begin{cases} (\text{CN}_i+1)Z_i & \forall i=j \\ \frac{1}{2}(h_{ii}+h_{jj}) & \forall i \neq j \wedge r_{ij} < r^0_{ij} \\ 0 & \forall i \neq j \wedge r_{ij} \geq r^0_{ij} \end{cases}$
Chirality Matrix
$g_{ab} = \begin{cases} G_{0,a}\, p_a & \forall a=b \\ \frac{1}{2}(g_{aa}+g_{bb})/d_{ab} & \forall a \neq b \end{cases}$
1
2
3
4
5
5
3x

```
MolBar | 1.1.2 | C13NOH21 | 0 | -468 -383 -306 -181 -165 -150 -150 -106 -48 -44 -16 20 20 20 20 20 20 20 20 20 20 20 44 62 146 178 246 357 426 470
470 609 633 813 903 1021 | -323 -248 -141 -50 -9 120 120 147 180 238 407 427 551 697 774 | -152 -106 -106 -104 -59 -38 -31 -24 -22 -16 8 10 11 11 11
11 11 11 11 12 15 20 33 59 76 107 124 145 195 284 322 322 334 584 934 1580 | -16 0 0 0 66
```

Fig. 6 Flowchart for MolBar generation: (1) use of an extended adjacency matrix (eqn (2)) to represent the molecular topology. Identification of bonds by comparing Cartesian distances r_{ij} with reference values r^0_{ij}.[35] (2) Partitioning of the molecule into rigid fragments using bond orders and ring structures, then 3D structure unification of fragments using a specialized force field. (3) Use of a modified Coulomb matrix (eqn (4)) to describe the 3D atomic arrangement in the unified fragments. (4) Describing absolute configuration based on Osipov–Pickup–Dunmur indices $\{G_{0,a}\}$[36] for each fragment, incorporated into an additional matrix. (5) Concatenation of the eigenvalue spectra of the matrices to yield MolBar. All eigenvalue spectra are first multiplied by ten and then rounded to the nearest integer. These spectra are preceded by the identifier name, version, the chemical formula, and its (user-specified) charge.

In general, bonds with a bond order higher than one are considered as rigid substructures. The restriction of rotational motion around a multiple bond arises from the intrinsic nature of the π-bond.[43] This rigidity leads to important structural properties such as *E*–*Z* isomerism, which is distinguishable in the topography matrix, where interactions are allowed exclusively between atoms with more or less fixed relative positions. Therefore in $I(v_i, v_j)$, information about bond orders must be included. For a molecular graph G, we define the function w, which gives the integer bond order for the bonds in E_{atoms}.

$$w: E_{\text{atoms}} \rightarrow \mathbb{Z} \tag{7}$$

Following, we define a set of bonds with a bond order greater than one to only include bonds where the rotation around the bond is restricted.

$$M = \{w(e_{ij}) > 1 | e_{ij} \in E_{\text{atoms}}\} \tag{8}$$

Furthermore, ring structures can be considered as rigid substructures in the molecule. At first glance, this is particularly evident in aromatic rings, but as well as in aliphatic ring structures, where the alteration of dihedral angles within the system is limited, ensuring the integrity of the rings. To start, we define a set C as the set of all possible cycles c_a in the graph G:

$$C = \{c_a, c_b, \ldots, c_z\} \tag{9}$$

, where

$$c_a = (v_i, v_j, v_k, \ldots, v_i) \in V_{\text{atoms}} \times V_{\text{atoms}} \times \ldots \times V_{\text{atoms}} \tag{10}$$

and $e_{ij} \in E_{\text{atoms}}$ as well as $e_{jk} \in E_{\text{atoms}}$ and so on. Further, $|c_a| \geq 3$, where $|c_a|$ is defined as the number of vertexes in this cycle. Now, we denote a function that gives the smallest cycle c_a for a given vertex v_i to obtain a chemically meaningful cycle definition.

$$f(v_i) = \text{argmin}\{|c_a| \forall c_a \in C, v_i \in c_a\} \tag{11}$$

With that given function, we define the subset of chemical meaningful cycles C^{s}.

$$C^{\text{s}} = \{f(v_i) | v_i \in V_{\text{atoms}}\} \tag{12}$$

With that we also can define a subset of E_{atoms} termed $E^{\text{s}}_{\text{atoms}}$, containing edges that make up the chemical cycles in C^{s}.

$$E^{\text{s}}_{\text{atoms}} = \{e_{ij} | v_i \in c_a \wedge v_j \in c_a \wedge c_a \in C^{\text{s}}\} \tag{13}$$

Now, two vertices belong to the same fragment if no edge in the shortest path connecting them is a non-cycle edge with a bond order of 1. A path $\hat{p}_a \in P(v_i, v_l)$ is in general defined as a sequence of vertices connecting two vertices v_i and v_l, where $P(v_i, v_l)$ is a set of all possible paths between the two vertices.

$$\hat{p}_a = (v_i, v_j, v_k, \ldots, v_l) \in V_{\text{atoms}} \times V_{\text{atoms}} \times \ldots \times V_{\text{atoms}} \tag{14}$$

with $e_{ik} \in E_{\text{atoms}}$ as well as $e_{jk} \in E_{\text{atoms}}$ and so on. The shortest $p^s(v_i, v_l)$ between two vertices is the path that minimizes the number of edges in that path.

$$\hat{p}^{\text{s}}(v_i, v_l) = \text{argmin}\{|p_a| | \forall \hat{p}_a \in P(v_i, v_l)\} \tag{15}$$

With this, we can now define the indicator function $I(v_i, v_j)$ needed for the eqn (4). So that two vertices are part of the same rigid fragment, the function returns the value one if all edges on the shortest path between both vertices either have a bond order greater than one or are part of a cycle. Also, the function returns the value one if both vertices are adjacent (number of vertices in $\hat{p}^{\text{s}}$ is equal to two), so that there is an overlap between the vertices of the fragment. This is done so that there is an interaction between the fragments and information about the relation between two fragments is still included in eqn (4) (the matrix does not consist entirely of block matrices). On the other hand, if there is at least one non-cyclic single bond between the two edges or if the two vertices are not adjacent, the value zero is returned.

$$I(v_i, v_j) = \begin{cases} 1 & \left(\forall \ e_{kl} \in \hat{p}^s(v_i, v_j) : e_{kl} \in (M \cup C^s) \vee \left|\hat{p}^s(v_i, v_j)\right| = 2\right) \\ 0 & \left(\exists \quad e_{kl} \in \hat{p}^s(v_i, v_j) : e_{kl} \notin (M \cup C^s)\right) \end{cases} \tag{16}$$

Before the indicator function can be applied, the bond orders must first be assigned to the bonds and cycles determined according to their definition in eqn (12). Bond orders can be calculated by quantum mechanical calculations such as the Wiberg (or Mayer) bond orders (WBO).[44,45] However, because of computational cost, such a quantum mechanical calculation is not preferable for a molecular identifier. In the past, an algorithm based on graph theory was proposed by Y. Kim and W. Y. Kim, requiring only information about the edges present and element-specific parameters such as the number of possible valences for a given element and the number of valence electrons.[46] This algorithm has already been implemented in *xyz2mol*[47] and *RDKit*[48] and is also implemented in a modified version in *molbar*. To find the minimum size cycles (eqn (12)) for a graph, several algorithms have been developed in the past. In the implementation, the algorithm proposed by Kavitha *et al.*[49] as implemented in the *Python* package *networkx*[50] is used. In the context of solving the shortest path problem, the standard Dijkstra algorithm[51] is implemented using the *networkx* library as well.

Since the fragmentation into rigid substructures is based solely on bond orders and ring structures, atropisomerism cannot be automatically captured with this rigidity function in its current form in MolBar. Atropisomerism refers to restricted rotation around a single bond.[34] For example, this fragmentation procedure applied to [1,1′-binaphthalene]-2,2′-diol in Fig. 4c would result in two separate 2-naphthol units, losing the chirality information. While atropisomerism cannot yet be handled automatically, the Python implementation allows for manual input to specify rigidity and thus accurately describe atropisomerism. An additional consequence of this current fragmentation algorithm is the separation of mechanically interlocked molecular architectures, such as catenanes,[52] rotaxanes,[53] or molecular knots,[54] into their covalent substructures. As a result, MolBar in its current form cannot capture the spatial arrangements of these architectures.

To illustrate the concept of fragmentation, consider the second step in Fig. 6, in which the molecule 3-(3-*tert*-butylcyclobutylidene)-piperidin-2-one is fragmented. This results in four tetrahedral fragments and one larger fragment consisting of two rings linked by a double bond. The definition of the indicator function in eqn (16) leads to overlapping fragments where the atoms are shared by several fragments, due to the rule that adjacent atoms always belong to the same fragment.

4.3 Step 2.2: structure unification

After fragmentation to eliminate the conformational dependence of the eqn (4), multiple fragments are obtained with their respective 3D coordinates. However, structural differences between identical fragments of the same molecule may remain due to variations attributable to different sources of the coordinates provided. For example, the application of different DFT functionals in a geometry optimization may result in different bond lengths, angles, and dihedral angles for the same molecule.[38,55,56] To obtain a uniform identifier for all input structures of the same molecule, structural unification of bond lengths, angles, and dihedral angles is essential. An effective method to achieve this unification is to use force field optimization characterized by harmonic potentials.[57] By using such harmonic terms of the form of $k(x_0 - x)^2$ with sufficiently large force constants, the system can be forced to a reference value x_0:

$$\begin{aligned} E_{\text{FF}} &= E_{\text{bond}} + E_{\text{angle}} + E_{\text{dihedral}} + E_{\text{Coulomb}} \\ &= \sum k_{\text{bond}}\left(r_{ij}^{\text{cov}} - r_{ij}\right)^2 + \sum k_{\text{angle}}\left(\alpha_{ijk}^0 - \alpha_{ijk}\right)^2 \\ &\quad + \sum k_{\text{dihedral}}\left(\left(\sin\theta_{ijkl}^0 - \sin\theta_{ijkl}\right)^2 + \left(\cos\theta_{ijkl}^0 - \cos\theta_{ijkl}\right)^2\right) \\ &\quad + \sum \frac{Q}{r_{ij}} \end{aligned} \tag{17}$$

E_{bond} sums harmonic terms for each edge $e_{ij} \in E_{\text{atoms}}$ to constrain the actual bond length to an element-pair specific reference value r_{ij}^{cov} as the sum of the covalent radii proposed by Pyykkö and Atsumi.[58] E_{angle} sums harmonic terms for all angles that can be defined for each fragment atom ν_i as the central atom. The reference angle α^0 is derived, *e.g.*, from the basic VSEPR theory but also adapted from more distorted geometries, with the values described in the ESI.† [59] Future developments will also include reference geometries other than VSEPR, such as those required for metal complexes. Currently, such systems with unknown VSEPR reference, will be treated exclusively *via* the distance constraints and the repulsive term.

Consequently, a method is needed to represent the structural arrangement of neighboring atoms around an atom ν_i and assign a reference geometry classification to it. For example, the Smooth-Overlap of Atomic Positions (SOAP) descriptor can be used to evaluate and compare different chemical environments around an atom.[60] This is done by creating a neighborhood density around the atom by summing Gaussian functions centered on neighboring atoms by spherical harmonics and radial basis functions. The resulting expansion coefficients c_{nlm} then find application in a similarity kernel equation for comparing different chemical environments:

$$k\left(\rho, \rho'\right) = \sum_{n,n',l,m,m'} c_{nlm}\left(c'_{nlm'}\right)^*(c_{nlm})^* c'_{n'lm'} \tag{18}$$

$$k\left(\rho, \rho'\right) = \sum_{n,n',l} p_{nn'l} p'_{nn'l} \tag{19}$$

Here,

$$p_{nn'l} = \sum_m c_{nlm}(c_{n'lm})^* \tag{20}$$

represents the power spectrum of the neighborhood density. Thus, the similarity kernel equation corresponds to the dot product between the power spectra of two neighborhood densities. To assign reference geometry classes to an atom ν_i, we construct the neighborhood density by considering only the set of adjacent atoms. The SOAP kernel is normalized as

$$K\left(\rho, \rho'\right) = \left(\frac{k(\rho, \rho')}{\sqrt{k(\rho, \rho)k(\rho', \rho')}}\right)^2. \tag{21}$$

When calculating the power spectrum in eqn (20) for each ν_i, the real geometry is compared with ideal reference structures by the similarity measure in eqn (21), which ranges from 0 (not similar) to 1 (very similar). The overall workflow for finding the best set of reference angles $\{\alpha_{ijk}^0\}$ is shown in Fig. 8. Since the assignment of the reference angles is not straightforward for distorted geometries or such ones, where not all angles are identical, several steps are necessary. First, the workflow cuts out the immediate neighborhood of the target atom from the overall structure, taking into account the coordinates of both the central atom and the adjacent atoms (as shown in the first step of Fig. 8). In preparation for the power spectrum calculations, all bond lengths are set uniformly to 1 Å, and placeholder elements are introduced that align with the ideal reference structure, as shown in the second step of Fig. 8. This alignment is crucial because the calculation of the power spectrum depends on the atomic elements and bond lengths.

In the next step, the power spectrum given in eqn (20) is calculated for the rescaled local structure in question. Then, using the SOAP kernel described in eqn (21), a comparison of this power spectrum with those of the ideal reference structures is performed. This process is illustrated in the third step of Fig. 8. Once the best match to an ideal geometry class is identified, the Kabsch algorithm is used to calculate the optimal rotation matrix to rotate the real structure into the ideal geometry.[61] This matrix facilitates the mapping of the atoms of the real structure to the corresponding atoms in the ideal geometry. In this alignment process, both the real local geometry and the reference geometry are shifted so that their center of mass coincides with the origin of the coordinate system. Then, a covariance matrix labeled $\mathbf{H}$ is computed in eqn (22), where $\mathbf{P}_{\text{real}}$ and $\mathbf{Q}_0$ are the centered Cartesian coordinates of the two structures as matrices of dimensions $n_{\text{atoms}} \times 3$ (fourth step of the Fig. 8).

$$\mathbf{H} = \mathbf{P}_{\text{real}}^{\text{T}}\mathbf{Q}_0 \tag{22}$$

The optimal rotation matrix $\hat{\mathbf{R}}$ to rotate $\mathbf{P}_{\text{real}}$ into $\mathbf{Q}_0$ can be determined with a Singular Value Decomposition (SVD) of the covariance matrix $\mathbf{H}$.

$$\mathbf{H} = \mathbf{U}\Sigma\mathbf{V}^{\text{T}} \tag{23}$$

The rotation matrix then follows by

$$\hat{\mathbf{R}} = V\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & d \end{pmatrix}\mathbf{U}^T \tag{24}$$

where d ensures a right-handed coordinate system:

$$d = \text{sign}(\det(\mathbf{V}\mathbf{U}^{\text{T}})) \tag{25}$$

Since $\mathbf{P}_{\text{real}}$ and $\mathbf{Q}_0$ must be in the same atomic order, all atomic permutations of $\mathbf{P}_{\text{real}}$ must be considered to find the best match. In the final step, the best reference angles are determined by using the angles between the mapped atoms in the ideal geometry after aligning them with the atoms in the real structure. Nitrogen is an exception to the above procedure as its usual trigonal pyramidal geometry can easily be inverted. Therefore, nitrogen always obtains angular constraints for a trigonal planar structure. Stereocenters involving nitrogen can occur when the configuration is fixed by a ring structure or other constraints.

E_{dihedral} sums the harmonic terms for all dihedral angles used to constrain different structures within the molecule. For example, the dihedral constraint ensures that C═C double bonds and aromatic rings are constrained to be planar. Furthermore, these terms are also used to planarize aliphatic ring structures so that different conformations of, for example, cyclohexane are unified into a single structure. The resulting unified structure does not need to be physically meaningful, since the crucial information is retained even after structure unification, namely whether the atoms are above or below the ring plane. Thus, the topography matrix in eqn (4) is conformer-independent even for non-aromatic ring systems and still contains the information necessary to distinguish *cis*/*trans* isomers for example. The sine and cosine basis was chosen to prevent unreasonably high energy contributions upon geometry optimization, due to a possible sign change of a dihedral angle close to 180°. When using the sine and cosine basis, as shown in eqn (17), the sign reversal problem is circumvented and a more stable representation of the dihedral angles is obtained as $\sin(180°) = \sin(-180°) = 0$ and $\cos(180°) = \cos(-180°) = -1$.

E_{Coulomb} is introduced to enforce repulsive interatomic interactions, which results in maximizing the interatomic distance under constraints. Thus, E_{Coulomb} sums over all atom–atom combinations with system-independent scaling constant Q. Table 1 shows the default values for the parameters used in the force field to unify the input structures. The geometry optimization is performed with a Newton-CG algorithm[62] as implemented in the *scipy* package.[63] The gradients and the Hessian matrix are calculated analytically, and the derivatives can be found in the ESI.†

Table 1 Default values for the scaling constants used in the force field to unify the input structures

Constant	Default value
k_{bond}	1×10^6 Å^{-2}
k_{angle}	2×10^3 rad^{-2}
k_{dihedral}	2×10^3
Q	1×10^2 Å

4.4 Step 3: topography matrix

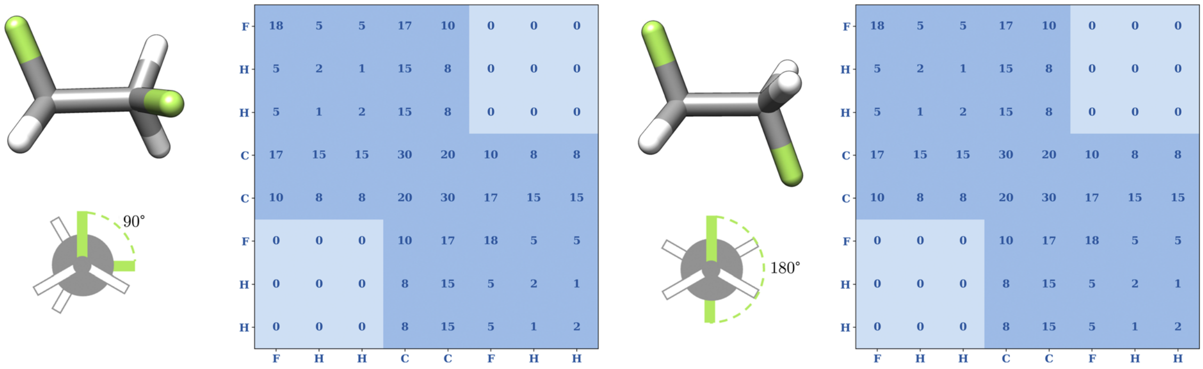
After the fragmentation and structure unification of all fragments, the matrix of eqn (4) can be set up by calculating the interatomic distances for the atoms within a fragment. The coordination number and the nuclear charges are already known from the first step. In contrast to Fig. 3 with the CM matrix in eqn (3) by Rupp *et al.*,[27] the topography matrix is conformer independent due to fragmentation and structure unification (*c.f.* Fig. 7). The rotation around a C–C single bond changes the interatomic distances between the substituents of the two carbon atoms. The fact that the substituents of the carbon atoms in 1,2-difluoroethane do not belong to the same molecular fragment, *e.g.*, for the two fluorine atoms $I(1, 6) = 0$, does not lead to any change in the matrix despite the inherent changes in the spatial arrangement. Consequently, for the substituents of the two carbon atoms represented by zero entries in the matrix, no 3D information of those two atoms is kept in the MolBar topography matrix, leading almost to a set of block matrices. The complete formation of block matrices is prevented only by the overlapping of the fragments, since the adjacent carbon atom belongs in each case to the fragment of the other carbon atom.

4.5 Step 4: absolute configuration matrix

Moreover, the absolute configuration matrix in eqn (5) can be constructed once the fragments are defined and their structures are unified. Unlike eqn (4), the matrix contains the fragmentary information rather than the atomic information. For this purpose, a fragment graph is constructed where each vertex $v_a \in V_{\text{fragments}}$ represents a fragment and each edge $e_{ab} \in E_{\text{fragment}}$ describes whether two fragments are adjacent, *i.e.*, there is an overlap between the atoms $v_i \in V_{\text{atoms}}$ in both fragments (*cf.* Fig. 9). The interaction between two fragments v_a and v_b is described by the graph distance d_{ab}. Thus, d_{ab} is the number of edges in the shortest path between two fragments v_a and v_b, calculated by the standard Dijkstra algorithm.[51]

4.6 Step 4.1: chirality index

Each fragment v_a, must be characterized in terms of its absolute configuration with an index, denoted G_a. Requirements



	F	H	H	C	C	F	H	H
F	18	5	5	17	10	0	0	0
H	5	2	1	15	8	0	0	0
H	5	1	2	15	8	0	0	0
C	17	15	15	30	20	10	8	8
C	10	8	8	20	30	17	15	15
F	0	0	0	10	17	18	5	5
H	0	0	0	8	15	5	2	1
H	0	0	0	8	15	5	1	2

	F	H	H	C	C	F	H	H
F	18	5	5	17	10	0	0	0
H	5	2	1	15	8	0	0	0
H	5	1	2	15	8	0	0	0
C	17	15	15	30	20	10	8	8
C	10	8	8	20	30	17	15	15
F	0	0	0	10	17	18	5	5
H	0	0	0	8	15	5	2	1
H	0	0	0	8	15	5	1	2

Fig. 7 Conformational independence of the topography matrix in eqn (4) for two conformers of 1,2-difluoroethane optimized at the GFN2-xTB level of theory. In general, rotation around the C–C single bond changes the distance between the two fluorine atoms, but since the two fluorine atoms are not part of the same fragment, *i.e.*, $I(1, 6) = 0$. In MolBar, we enforce no 3D information between atoms of two different fragments by zeroing out the corresponding Coulomb matrix elements. This mimics the effect of conformational averaging and, thus, changing the F–F distance does no longer affect the topography matrix.

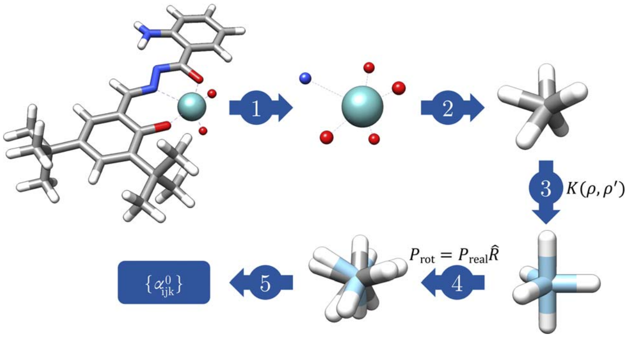

1
2
3
$K(\rho,\rho')$
$P_{\mathrm{rot}} = P_{\mathrm{real}}\hat{R}$
4
5
$\{\alpha^0_{\mathrm{ijk}}\}$

Fig. 8 Flowchart to determine the best set of reference angles $\{\alpha^0_{ijk}\}$: (1) isolation of local structure. (2) Insertion of dummy elements and rescaling all bond lengths to 1 Å. (3) Calculate eqn (20) and select the most resembling reference geometry class based on eqn (21). (4) Apply Kabsch algorithm for optimal rotation matrix.[61] (5) Obtain best reference angles by comparing mapped real and ideal structure.

for such an index are that it must be zero for nonchiral and non-zero for chiral fragments. Moreover, it should change sign upon spatial inversion of a chiral fragment, which allows the determination of an absolute configuration.[64] Various chirality indices have been proposed in the past such as the Hausdorff measure,[65,66] the helicity tensor of Ferrarini and Nordio[67] or a measure based on mean torsion in a molecule by Luzanov and Babich.[68]

Another approach is to derive a chirality index based on optical activity theory, since chiral molecules have different specific angle of rotation for example. Therefore, it is logical to start with an optical activity tensor to derive a geometric chirality measure. Osipov, Pickup, and Dunmur stated that the pseudoscalar behaviour of a chiral molecule described by its Cartesian coordinates **X** can be described by the trace of the gyration tensor **G** (eqn (26)),[36]

$$G_0(\mathbf{X}) = \frac{1}{3}\mathrm{Tr}\mathbf{G}(\mathbf{X}) \tag{26}$$

$$G_0(\mathbf{X}) = \int \rho(\mathbf{r}_1)\rho(\mathbf{r}_2)\rho(\mathbf{r}_3)\rho(\mathbf{r}_4) \times \frac{[(\mathbf{r}_{12}\times\mathbf{r}_{34})\cdot\mathbf{r}_{14}](\mathbf{r}_{12}\cdot\mathbf{r}_{23})(\mathbf{r}_{23}\cdot\mathbf{r}_{34})}{(r_{12}r_{23}r_{34})^n r_{14}^m} \mathrm{d}\mathbf{r}_1\mathrm{d}\mathbf{r}_2\mathrm{d}\mathbf{r}_3\mathrm{d}\mathbf{r}_4, \tag{27}$$

where $\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j$ and r_{ij} is the magnitude of the vector $\mathbf{r}_{ij}$. $\rho(\mathbf{r}_i)$ is an arbitrary density that describes the molecule of interest. n and m are arbitrary integer values. If $n = 2$ and $m = 1$ are used the index is dimensionless which is used in this work. G_0 is defined as the isotropic chirality index, which is invariant under rotation and translation. As a pseudoscalar, G_0 changes its sign for space inversion for chiral objects, while being zero only for achiral molecules. This means, the index is theoretically

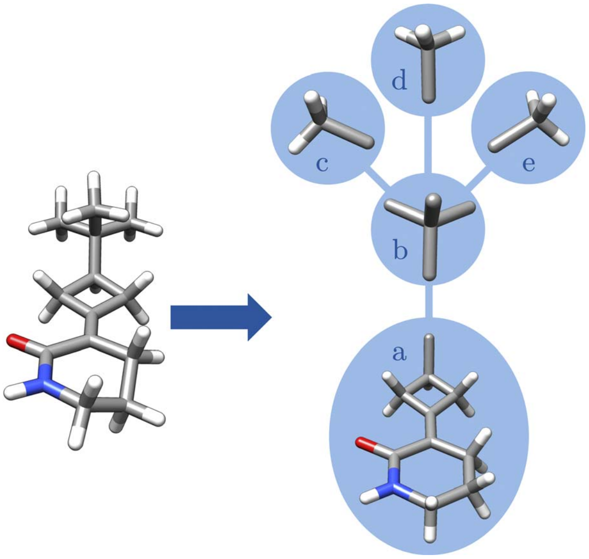


Fig. 9 Visualization of the fragment graph for 3-(3-*tert*-butylcyclobutylidene)piperidin-2-one: each vertex $v_a \in V_{\mathrm{fragments}}$ (blue circle) represents an unified fragment and each edge $e_{ab} \in E_{\mathrm{fragments}}$ describes whether two fragments are adjacent, *i.e.*, there is an overlap between the atoms $v_i \in V_{\mathrm{atoms}}$ in both fragments.

capable of describing the chirality of a molecule in general terms, *i.e.*, regardless of whether central, axial, or planar chirality is present. If $\rho(\mathbf{r})$ is replaced by point masses p_i through the delta functions $\delta(|\mathbf{r} - \mathbf{r}_i|)$, G_0 reduces to a discrete form in eqn (28):

$$G_0(\mathbf{X}) = \sum_{ijkl} p_i p_j p_k p_l \frac{(\mathbf{r}_{ij}\mathbf{r}_{kl}\mathbf{r}_{il})(\mathbf{r}_{ij}\mathbf{r}_{jk})(\mathbf{r}_{jk}\mathbf{r}_{kl})}{(r_{ij}r_{jk}r_{kl})^n r_{il}^m} \tag{28}$$

For the purpose of the identifier, only the sign is of interest, so the used expression for the chirality index G_a of a fragment, described with the Cartesian coordinates after the unification process $\mathbf{X}_a^{\text{uni}}$, writes as:

$$G_a = \begin{cases} 1 & \forall\, G_0(\mathbf{X}_a^{\text{uni}}) > 0 \\ 0 & \forall\, G_0(\mathbf{X}_a^{\text{uni}}) = 0 \\ -1 & \forall\, G_0(\mathbf{X}_a^{\text{uni}}) < 0 \end{cases} \tag{29}$$

However, Millar, Weinberg, and Mislow have pointed out that the utilization of pseudoscalar functions as measures of chirality theoretically leads to situations where a chirality index becomes zero, even when the object is inherently chiral.[64] The argument is that for each enantiomer $\mathbf{X}$, there exists a path of geometric distortion transforming it into its enantiomer $\tilde{\mathbf{X}}$, where only chiral objects exist along that path. As there must be change of sign for G_0, there must also be chiral structure with $G_0 = 0$. Future research needs to investigate whether this occurrence is frequent, as current tests show its usefulness to describe the absolute configuration of unified fragments (*c.f.* Section 5). The unification process, which precludes any geometric distortion path between $\mathbf{X}$ and $\tilde{\mathbf{X}}$, as every input fragment structure is unified into either $\mathbf{X}_\text{a}^\text{uni}$ or $\tilde{\mathbf{X}}_\text{a}^\text{uni}$, may potentially reduce the incidence of chiral zeros.

4.7 Step 4.2: priorities

The process requires the use of pseudo masses p_i for an atom ν_i in eqn (28). If the atomic masses m_i were employed, identical atoms of the same element but in different environments would be treated as equivalent. In the case of fragment *b* in Fig. 9, the four adjacent carbon atoms are not equivalent due to their substituents, since three carbon atoms belong to a methyl group and one to a ring system. Therefore, we introduce a pseudo mass p_i termed as the atomic priority for each atom ν_i in the molecule, which takes the atom itself and its environment into account. In theory, this priority can be anything as long as it groups equivalent atoms together. Coming from theoretical chemistry and quantum mechanics, equivalent atoms share the same atom-partitioned electron density. The Mulliken population analysis is often used in quantum chemistry to get the number of electrons associated to individual atoms in a molecule.[57] The total electron density is partitioned into atomic contributions based on the coefficients of the atomic orbitals that contribute to each molecular orbital. The analysis provides insights into the distribution of electrons across the atoms within a molecule, aiding in understanding bonding characteristics and reactivity. The partitioning reads as follows:

$$\sum_k^{n_\text{orb}} n_k \int |\phi_k(\mathbf{r})|^2 \text{d}^3\mathbf{r} = \sum_{\mu\nu}^{n_\text{basis}} \left(\sum_k^{n_\text{orb}} n_k c_{\mu k} c_{\nu k} \right) S_{\mu\nu} \tag{30}$$

$$\sum_k^{n_\text{orb}} n_k \int |\phi_k(\mathbf{r})|^2 \text{d}^3\mathbf{r} = \sum_{\mu\nu}^{n_\text{basis}} P_{\mu\nu} S_{\mu\nu} \tag{31}$$

$$\sum_k^{n_\text{orb}} n_k \int |\phi_k(\mathbf{r})|^2 \text{d}^3\mathbf{r} = \sum_{\mu}^{n_\text{basis}} (\mathbf{PS})_{\mu\mu} = \text{Tr}(\mathbf{PS}) \tag{32}$$

$$\sum_k^{n_\text{orb}} n_k \int |\phi_k(\mathbf{r})|^2 \text{d}^3\mathbf{r} = n_\text{elec} \tag{33}$$

with the coefficients of the atomic orbitals $c_{\mu k}$, the occupation number n_k, the density matrix $\mathbf{P}$ and the overlap matrix $\mathbf{S}$. The matrix element $(\mathbf{PS})_{\mu\mu}$ can be interpreted as the number of electrons associated with the basis function ϕ_ν. Summing over all basis functions on an atom ν_i yields the electron density at that atom. Usually, this analysis is performed for electronic structure theory calculations. However, no such calculations are conducted for a molecular identifier like MolBar, as those calculations are computationally too expensive. In simple Hückel theory, the eigenvectors of the Hückel matrix represent the molecular orbitals (MOs) of the molecule. Similarly, the eigenvectors of the MolBar topography matrix can be interpreted as non-physical orbitals, mathematical constructs with information about the 3D molecular shape including (a-)symmetry. Diagonalizing the MolBar topography matrix (eqn (34)) yields the MolBar orbitals and their corresponding eigenvalues:

$$\mathbf{H}_\text{topography}\mathbf{C} = \mathbf{EC} \tag{34}$$

These orbitals can then be used to calculate an artificial electron density at each atom. Determining how to populate the MolBar orbitals is challenging since the topography matrix lacks physical significance. In electronic structure theory, the lowest eigenvalues are populated according to the Aufbau principle, as this corresponds to the most stable occupation. In MolBar, however, the orbitals with highest eigenvalues show fewer nodal planes. Hence, all orbitals with positive eigenvalues are occupied, with each orbital having an arbitrarily chosen occupation number n_k of two. This ensures a defined rule, with degenerate orbitals (capturing symmetry information) always occupied in the same manner. This process results in the formulation of the MolBar density matrix, denoted as $\mathbf{P}^\text{MolBar}$:

$$\mathbf{P}^\text{MolBar} = \mathbf{C}\mathbf{N}_\text{occ}\mathbf{C}^\top \tag{35}$$

with the occupation matrix $\mathbf{N}_\text{occ}$, indicating whether an orbital is occupied or not. Further, assuming zero overlap between the orbitals (as in Hückel theory), the overlap matrix can be simplified to the identity matrix.

$$\sum_\mu^{n_\text{basis}} P_{\mu\mu}^\text{MolBar} = \text{Tr}(\mathbf{P}^\text{MolBar}) = n_\text{elec} \tag{36}$$

Further, as the topography matrix has the dimensions $n_{\text{atoms}} \times n_{\text{atoms}}$, equivalent to a minimal basis set, the artificial atomic density for each atom just simply the density matrix element:

$$P_{\mu\mu}^{\text{MolBar}} = \rho_i \tag{37}$$

Based on that procedure, each atom $\{v_i\}$ gets assigned an artificial electron density ρ_i. These densities are ranked yielding a set of atomic priorities $\{p_i\}$. The priorities obtained naturally include information about connectivity, but even if two branches differ only in the configuration of the double bond, using the topography matrix as a source for the orbitals. Equivalent atoms have the same density so they get assigned the same priority, while non-equivalent atoms receive different priorities. For the absolute configuration matrix in eqn (5), not atomic priorities but fragment priorities $\{p_a\}$ are needed. Those fragment priorities are obtained by comparing atomic priorities of fragments (*c.f.* Fig. 10). The highest atomic priority within a fragment is compared to the highest atomic priority of another fragment. The fragment with the highest atomic priority is then assigned the highest fragment priority. In case of a tie between several fragments, the comparison continues with the evaluation of the second highest priority until all fragments receive different priorities or are considered equal.

4.8 Step 5: concatenation of modified eigenvalue spectra and MolBar generation

In the final step, the eigenvalue spectra of the matrices from eqn (2), (4), and (5) are combined to generate MolBar. The eigenvalues are first multiplied by ten and then rounded to the nearest integer in order to minimize the size of the MolBar. Previous results have shown that it is not necessary to use more digits. We term this the "modified eigenvalue spectra". A MolBar example can be found in Table 2 using 3-(3-*tert*-butylcyclobutylidene)piperidin-2-one as an example (*c.f.* Fig. 6).

The first segment contains the version number, the molecular formula and the total molecular charge. The purpose of specifying the version is to ensure that only MolBar identifiers with the same version are compared, taking into account that future changes or bug fixes may result in different spectra. The concatenation of eigenvalue spectra follows a specific order: $\lambda_{\text{topology}}$, $\lambda_{\text{topology},Z_i>Z_{\text{H}}}$, $\lambda_{\text{topography}}$, and $\lambda_{\text{chirality}}$. First, the eigenvalue spectra $\lambda_{\text{topology}}$ of the topology matrix (eqn (2)) are shown, followed by the spectra of the topology matrix without hydrogen atoms. By explicitly including hydrogen atoms in one matrix and excluding them from the other, it is possible to classify two different molecules as prototropic tautomers. By default, therefore, tautomeric structures are given a different MolBar identifier. This approach is advantageous for quantum chemical structure databases, as it guarantees that each structure receives a unique database entry, since it is characterized by its unique properties derived from electronic structure theory. Prototropic tautomers exist when, under the same molecular formula, the total topology spectrum is different but the heavy atom topology spectrum is identical. It must be kept in mind that the tautomeric picture is rather simplistic, as it does not take into account large energy barriers for hydrogen shifts,[69] but it is sufficient for quantum chemical structure databases. Then the topography spectrum $\lambda_{\text{topography}}$ is given (eqn (4)), followed by the absolute configuration spectrum (eqn (5)).

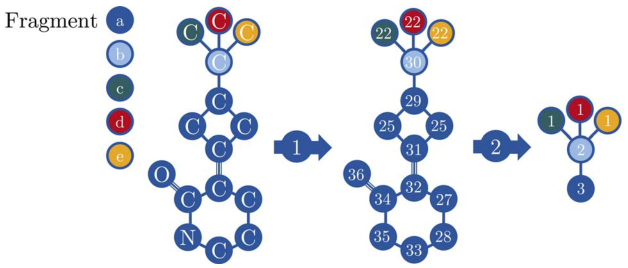


Fig. 10 (1) Each fragment is assigned a priority based on the atomic priorities of its atoms. (2) The highest atomic priority within a fragment is compared to the highest atomic priority of another fragment. The fragment with the highest atomic priority is then assigned the highest fragment priority.

5 Results and discussion

5.1 Topology test

Many approaches for describing molecular topologies are based on graph relaxation algorithms to create a canonical atomic order.[70] The classical Morgan algorithm[1] faces challenges with molecules with high topological symmetry.[71] Since then many improved algorithms have been published that benchmark their approaches on such highly symmetrical molecules.[70,72,73] Unlike these methods, MolBar does not rely on any canonicalization process. MolBar starts with the topology matrix defining the molecular graph. The corresponding topology matrix is then diagonalized to obtain the sorted eigenvalues as a permutation invariant vector. In this approach, each eigenvalue represents a topological substructure composed of multiple atoms, rather than focusing on individual atoms. To test MolBar on systems with difficult topology and high symmetry, the topology barcode was tested on the set of 1812 unique isomers of C_{60} fullerene.[74] In fact, all 1812 topology barcodes generated by MolBar were unique. Thus, MolBar is able to identify each isomer even though the topology is very similar. Moreover, considering the full barcode, within these isomers, 1722 are identified as chiral, exhibiting non-central chirality. Barcodes and structures for all these isomers are included in the ESI.†

As a side note, in general, two distinct graphs can be isospectral, meaning they possess the same eigenvalue spectrum. This situation is particularly common when using only the adjacency matrix to represent the graph. However, the probability of finding isospectral graphs is highly dependent on the employed type of matrix. The number of such isospectral graphs can be significantly reduced by using a Laplacian matrix or by combining the eigenvalue spectra from several different matrices.[75] However, MolBar addresses this issue by using three matrices to describe molecular connectivity and is not purely graph-based: the topology matrix, the heavy atom topology matrix, and the non-graph-based topography matrix, the latter

Table 2 Structure of MolBar as in the example for chiral 3-(3-*tert*-butylcyclobutylidene)piperidin-2-one

MolBar	Example
Version	1.1.2
Molecular formula	$C_{13}NOH_{21}$
Total molecular charge	0
Topology	−468 −383 −306 −181 −165 −150 −150 −106 −48 −44 −16 20 20 20 20 20 20 20 20 20 20 20 44 62 146 178 246 357 426 470 470 609 633 813 903 1021
Heavy atom topology	−323 −248 −141 −50 −9 120 120 147 180 238 407 427 551 697 774
Topography	−152 −106 −106 −104 −59 −38 −31 −24 −22 −16 8 10 11 11 11 11 11 11 11 12 15 20 33 59 76 107 124 145 195 284 322 322 334
Chirality	−16 0 0 0 66

of which encodes information indirectly through bonding force field constraints. For two distinct molecules to be isospectral within MolBar, they would need to have identical eigenvalues across all three matrices, thereby minimizing the likelihood of such occurrences. So far, in all tested cases, including in the dataset of the duplication test discussed below and other cases beyond it, no instances of such isospectral molecules have been identified. However, continued research is necessary to further explore the robustness of this approach.

5.2 Duplication test

Molecular identifiers play a critical role in merging of and eliminating duplicate entries in molecular databases. A benchmark set is critical for evaluating different molecular identifiers in database deduplication tasks because it provides a standardized test bed that allows for fair and consistent comparisons. To our knowledge, there is no existing benchmark set for deduplication tasks. However, several datasets containing millions of molecules have been proposed, particularly in the field of molecular machine learning, where molecular structures are represented by 3D coordinates. Such a dataset has been proposed by Xu *et al.* with the benchmark set Molecule3D for predicting 3D geometries from molecular graphs.[76] It consists of 3 899 647 molecules with ground state 3D molecular geometries with an average of 29.11 atoms per molecule, an average of 14.08 heavy atoms and an average of 29.53 bonds. The source of the molecular data is the PubChemQC database,[77] with optimized molecular structures at the B3LYP/6-31G* level of theory.[78–90]

A pickle file *molecule3D.pkl* with all molecular 3D coordinates of that dataset is provided in the ESI.† Throughout the benchmark tests, the molecules are named by the ID specified in the Molecule3D dataset. As a side note, this ID initially corresponds to the CID in the PubChem database, but due to structural modifications, the structures may differ from the database in some cases due to factors such as hydrogen shifts or inversion of the absolute configuration of stereocenters. This discrepancy is not investigated further in this work, as the correspondence of the structures to the PubChem database is not relevant to the following discussion.

The Molecule3D dataset contains organic molecules with typical 2c–2e bonding patterns. The stereoisomerism in this dataset is present through stereocenters as well as *E*/*Z* and *cis*/*trans* isomerism, aspects that are captured by the InChI identifier in a highly robust manner.[8] This robustness and reliability in representing chemical structures makes InChI an ideal reference for evaluating MolBar on standard organic molecules with typical stereochemistry. To provide a reliable reference, the InChI identifier is generated using *RDKit* 2023.09.02 and *OpenBabel* 3.1.0 with the FixedH option to differentiate between tautomers. Typically, the input for InChI is a Molfile, for example, but since MolBar is designed to be generated from 3D coordinates, which is the starting point for computational chemists, Cartesian coordinates are used as the input for both MolBar and InChI. Using these two different tools with *RDKit* and *OpenBabel* helps to eliminate potential misinterpretations of the 3D geometry prior to InChI generation from the discussion. In 8.4% of 3 899 647 molecules, the InChIs generated by *RDKit* and *OpenBabel* are different. These differences could stem from a variety of reasons, likely unrelated to InChI itself. For instance, they may arise from the process of converting Cartesian coordinates into connectivity, bonding, and stereochemistry information before inputting them into InChI. To streamline the discussion, these molecules are excluded from the duplication test, resulting in a final count of 3 570 657 molecules (referred to as *filtered_molecule3D.pkl* in the ESI†).

First, we examine unique dataset entries identified by both MolBar and InChI. The Molecule3D dataset initially includes a variety of molecules, including constitutional isomers and stereoisomers. To explore the basic properties of MolBar and demonstrate its ability to discriminate between stereoisomers, several examples of stereoisomerism are presented in Fig. 11. Typically, these stereoisomers have the same topology barcode but differ in their topography and chirality barcodes.

The first two examples show classical cases with carbon stereocenters. In Fig. 11a, the topography barcodes are identical, which means that the interatomic distances within the rigid fragments are the same. Here, the fragments consist mainly of the benzene ring, while the remaining atoms form monoatomic fragments, as the atoms are all connected by rotatable bonds. However, the chirality barcodes differ only in sign, indicating that the molecules are enantiomers, as the absolute configurations of the two chiral fragments are interchanged. In Fig. 11a, with two chiral fragments, this results in four non-zero eigenvalues, while in Fig. 11b, with one chiral fragment, only two eigenvalues are non-zero.

The examples in Fig. 11c and d show cases where the topography barcodes differ, indicating that the interatomic distances within the rigid fragments are not the same. In both cases, this corresponds to a different configuration of a double bond: the C═N double bond in Fig. 11c and the C═C double bond in Fig. 11d. In addition, the chirality barcodes in both examples consist only of zeros, which indicates that all fragments are achiral.

In Fig. 11e and f the topography barcodes differ again, but here due to a different relative configuration of the ring substituents. These two examples also make it clear that the chirality barcodes can only be compared if the topography barcodes are the same. Although the absolute configuration of a chiral carbon center is different in both structures in Fig. 11f, the chirality barcode remains unchanged. In Fig. 11e, the chirality barcodes differ in sign, but both structures are not enantiomers. In Fig. 11e and f, the stereocenters are located within larger fragments (here the ring structures), which changes the interatomic distances. This changes the 3D shape of the fragments, which is then captured by the topography barcode. In general, it can be said that the sign determined by the Osipov–Pickup–Dunmur index strongly depends on the 3D shape of the fragment, and small changes can lead to different signs. Since the Osipov–Pickup–Dunmur index is derived from the theory of optical activity, this is comparable to how ECD spectra differ in orientation for two conformers.

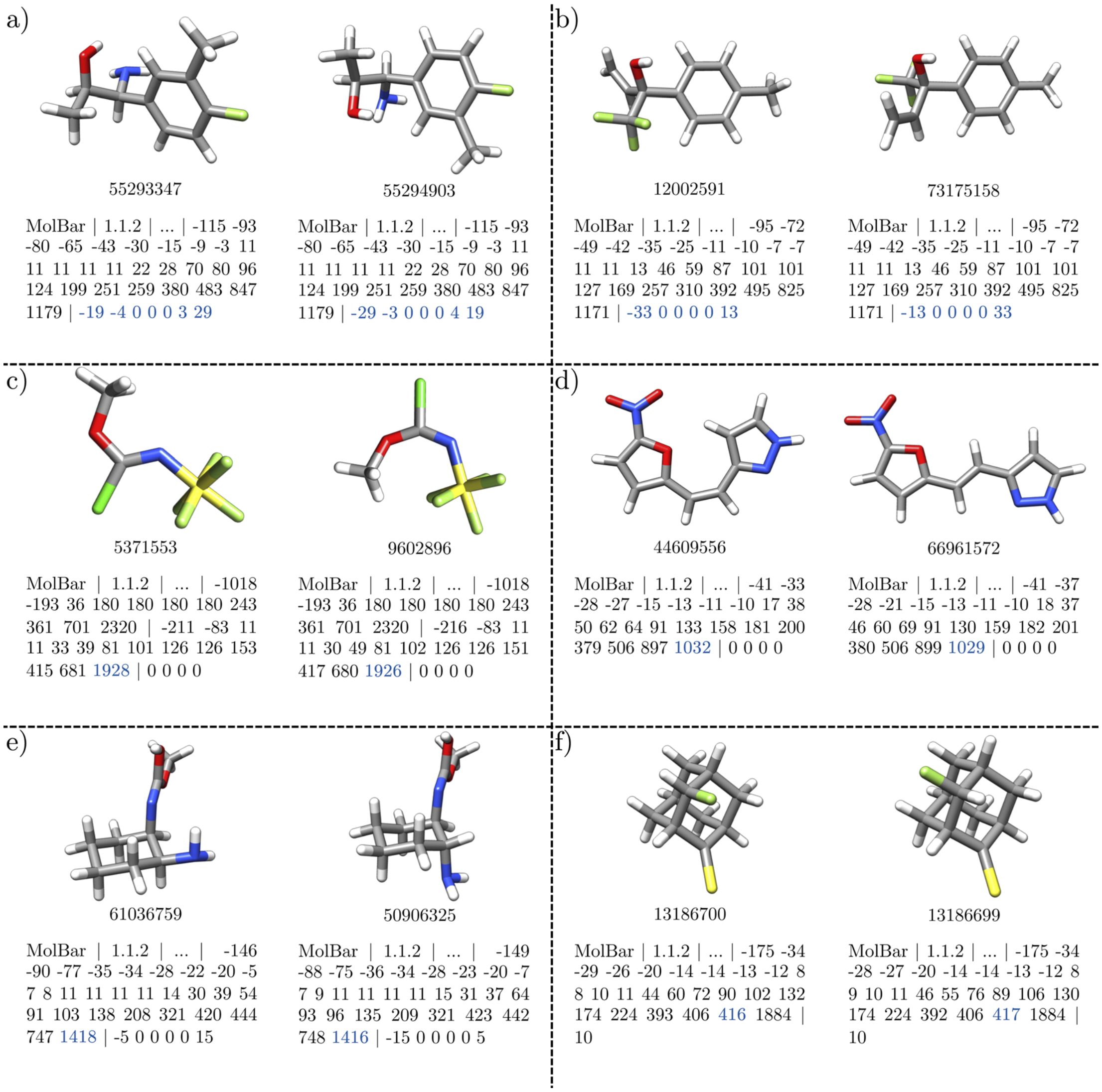

a)
55293347
MolBar | 1.1.2 | ... | -115 -93 -80 -65 -43 -30 -15 -9 -3 11 11 11 11 11 22 28 70 80 96 124 199 251 259 380 483 847 1179 | -19 -4 0 0 0 3 29
55294903
MolBar | 1.1.2 | ... | -115 -93 -80 -65 -43 -30 -15 -9 -3 11 11 11 11 11 22 28 70 80 96 124 199 251 259 380 483 847 1179 | -29 -3 0 0 0 4 19
b)
12002591
MolBar | 1.1.2 | ... | -95 -72 -49 -42 -35 -25 -11 -10 -7 -7 11 11 13 46 59 87 101 101 127 169 257 310 392 495 825 1171 | -33 0 0 0 0 13
73175158
MolBar | 1.1.2 | ... | -95 -72 -49 -42 -35 -25 -11 -10 -7 -7 11 11 13 46 59 87 101 101 127 169 257 310 392 495 825 1171 | -13 0 0 0 0 33
c)
5371553
MolBar | 1.1.2 | ... | -1018 -193 36 180 180 180 180 243 361 701 2320 | -211 -83 11 11 33 39 81 101 126 126 153 415 681 1928 | 0 0 0 0
9602896
MolBar | 1.1.2 | ... | -1018 -193 36 180 180 180 180 243 361 701 2320 | -216 -83 11 11 30 49 81 102 126 126 151 417 680 1926 | 0 0 0 0
d)
44609556
MolBar | 1.1.2 | ... | -41 -33 -28 -27 -15 -13 -11 -10 17 38 50 62 64 91 133 158 181 200 379 506 897 1032 | 0 0 0 0
66961572
MolBar | 1.1.2 | ... | -41 -37 -28 -21 -15 -13 -11 -10 18 37 46 60 69 91 130 159 182 201 380 506 899 1029 | 0 0 0 0
e)
61036759
MolBar | 1.1.2 | ... | -146 -90 -77 -35 -34 -28 -22 -20 -5 7 8 11 11 11 11 14 30 39 54 91 103 138 208 321 420 444 747 1418 | -5 0 0 0 0 15
50906325
MolBar | 1.1.2 | ... | -149 -88 -75 -36 -34 -28 -23 -20 -7 7 9 11 11 11 11 15 31 37 64 93 96 135 209 321 423 442 748 1416 | -15 0 0 0 0 5
f)
13186700
MolBar | 1.1.2 | ... | -175 -34 -29 -26 -20 -14 -14 -13 -12 8 8 10 11 44 60 72 90 102 132 174 224 393 406 416 1884 | 10
13186699
MolBar | 1.1.2 | ... | -175 -34 -28 -27 -20 -14 -14 -13 -12 8 9 10 11 46 55 76 89 106 130 174 224 392 406 417 1884 | 10

Fig. 11 Examples of stereoisomerism in the Molecule3D dataset. Only the topography and chirality barcodes are shown. In MolBar, the stereoisomers have the same topology barcode, but differ in the topography and chirality barcodes. Enantiomers are identified by chirality barcodes with opposite signs. The sign of the chirality indices depends strongly on the 3D geometry of the molecule, so that the chirality barcode can only be compared if two molecules have the same topography barcode.

For the duplication test, a molecule is considered a duplicate if it shares the same respective identifier with another molecule. From 3 570 657 molecules in the filtered Molecule3D dataset, around 162 300 duplicates are found by both identifiers, which corresponds to a duplication rate of 4.5% (see Fig. 12). Therefore, the MolBar identifier is able to identify duplicates with a similar performance to InChI. However, in 70 instances, MolBar and InChI identify different duplicates. This means that one identifier recognizes a molecule as a duplicate, while the other does not. Only one molecule is identified as a MolBar-exclusive duplicate. This case involves a different bonding motif in the InChI for the first and duplicate dataset entry, despite nearly identical 3D geometries. This discrepancy is likely due to the conversion tool from coordinates to bonding information rather than an issue with InChI. Therefore, it should not be discussed here but can be found in the ESI.† In contrast, 69 molecules are identified as InChI-exclusive duplicates. The 69 differences can be divided into eight different classes, the examples of which are shown in Fig. 13. All cases are categorized in their class and provided as XYZ files in the ESI.†

Fig. 13a shows a case in which both structures are enantiomers and exhibit axial chirality through the ring scaffold (1 case). Since the entire structure is a single fragment, the chirality barcode consists of only one eigenvalue. Therefore, the entire fragment is characterized by a chirality index, and this type of decentralized chirality can be detected. The two chirality barcodes differ only in the sign between the enantiomers. Fig. 13b, on the other hand, shows a case of centrochirality with phosphorus or sulfur (18 cases). In this example with phosphorus, the atom has four different substituents, indicating that it is a stereocenter. However, this case can also be considered as two tautomers, where the hydrogen atom can be exchanged between the oxygen groups. Consequently, the hydrogen position can change rapidly, inverting the absolute configuration. For InChI, the structures are therefore considered identical. For MolBar, however, the tautomers are always treated explicitly. The distinction between 11506511 and 16072129 is desirable, as both structures have different properties for the calculation of *e.g.* ECD spectra. Fig. 13c shows a case of chiral structures with two nitrogen stereocenters (1 case). Both structures are diastereomers. Normally, the nitrogen in a trigonal pyramidal geometry is not a stereocenter because it is rapidly transformed by pyramidal inversion. In this case, however, the configuration of the stereocenters is fixed by the ring structure. Fig. 13d and e highlight the importance of preserving 3D shape information for identifiers in computational chemistry and why MolBar's motivation deviates from InChI in some cases. Modern tools like CREST[20] and Nanoreactor[17] produce new structures as 3D point clouds, either during conformational sampling or reaction discovery. Given the vast number of generated geometries, manually verifying their chemical correctness is impractical. Fig. 13d demonstrates a scenario where the sulfur atom's geometry is trigonal planar in one structure and trigonal pyramidal in another (15 instances). MolBar identifies these differences using eqn (21) and applies distinct force field constraints to eqn (17), resulting in varied unified structures, topography matrices, and final barcodes. Fig. 13e shows a chemically unreasonable geometry, far from the global minimum (21 cases). Here, hydrogen atoms on one carbon point towards the center of the ring, unlike in the other structure. The examples given by Fig. 13d and e can also occur during black-box structure generation. For example, CREST allows users to define a starting structure with a specific geometry, but during its black-box conformational search, the geometry might change depending on the theoretical level used. Users need to be notified if the geometry alters during optimization or sampling processes. Identifying such discrepancies is crucial for ensuring the chemical reasonableness of geometries generated during black-box chemical space exploration. Further evaluation of this use case is discussed in Section 5.5.

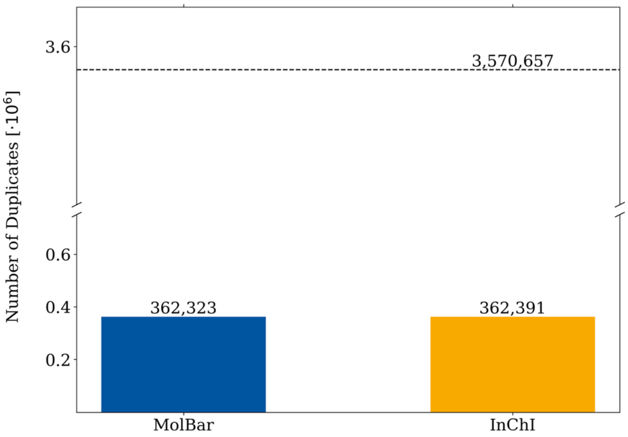


Fig. 12 Duplication test of the Molecule3D dataset. The filtered dataset comprises 3 570 657 molecules, each molecule represented by MolBar and InChI. A database entry is considered to be duplicated if the respective identifier already existed in previous entries.

Fig. 13f shows chemically unreasonable, broken structures, either missing hydrogen atoms or with open ring structures (5 instances). Fig. 13g and h highlight cases where MolBar needs improvement. In Fig. 13g, both structures are identical, but MolBar identifies one case where phosphorus atoms are closer than the threshold defined by connectivity evaluation. As only 6 out of 162 324 duplicate structures are effected by this, this indicates a rare issue. In general, it was found that the covalent radii of phosphorus or sulfur are too large for the connectivity evaluation and need to be adjusted.

Fig. 13h shows a case where, despite applying identical force field constraints, the exact same energetic minimum in the force field unification is not achieved (2 out of 162 324 duplicate structures). This issue is attributed to convergence issues with the optimizer used for the force field optimization. Overall the duplication test demonstrates that MolBar is capable of differentiating between stereoisomers and constitutional isomers and identifying duplicates with a similar performance to InChI. Differences between MolBar and InChI are mainly due to the different motivation from which the identifiers were developed, as MolBar treats all 3D structures explicitly.

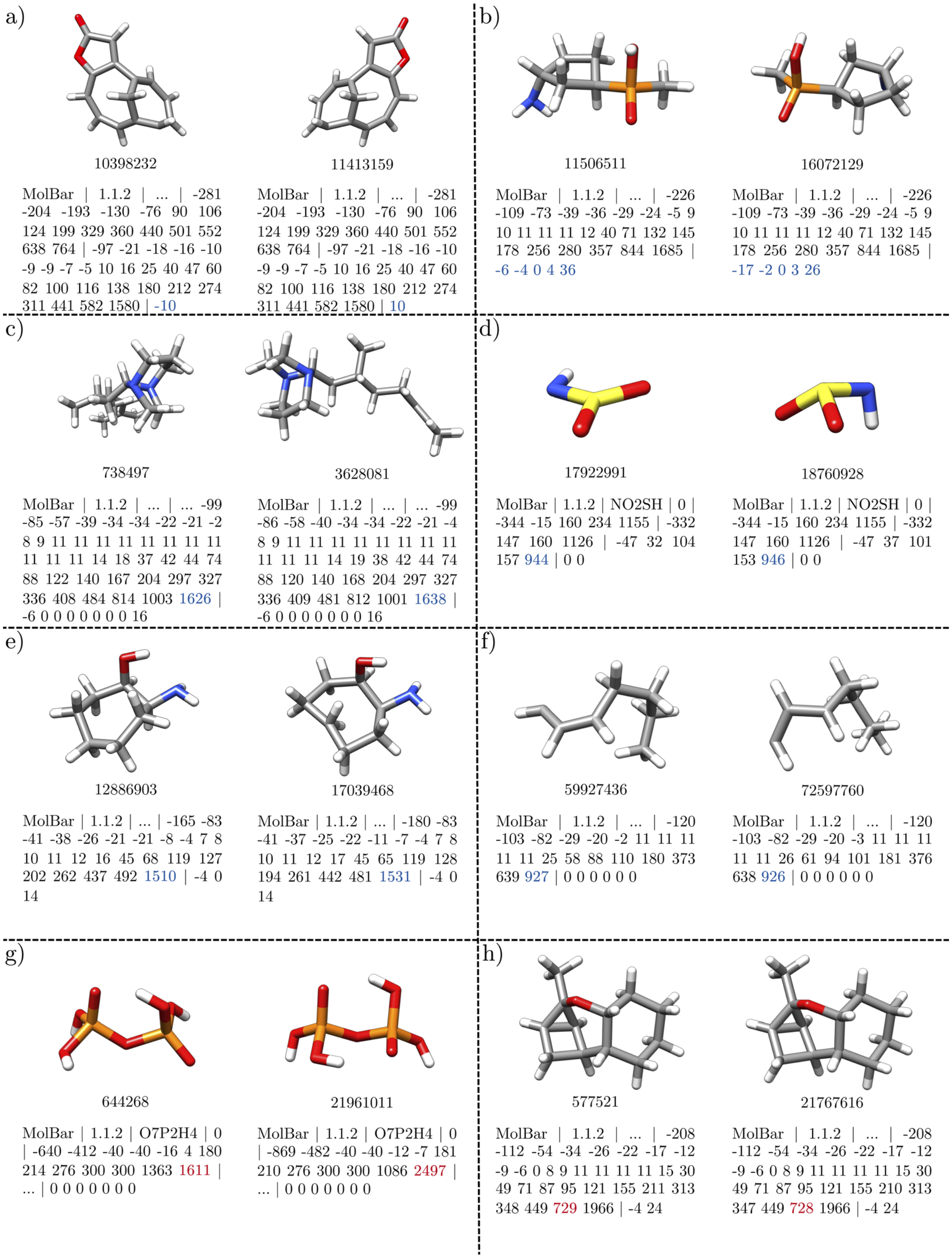

a)
10398232
MolBar | 1.1.2 | ... | -281 -204 -193 -130 -76 90 106 124 199 329 360 440 501 552 638 764 | -97 -21 -18 -16 -10 -9 -9 -7 -5 10 16 25 40 47 60 82 100 116 138 180 212 274 311 441 582 1580 | -10
11413159
MolBar | 1.1.2 | ... | -281 -204 -193 -130 -76 90 106 124 199 329 360 440 501 552 638 764 | -97 -21 -18 -16 -10 -9 -9 -7 -5 10 16 25 40 47 60 82 100 116 138 180 212 274 311 441 582 1580 | 10
b)
11506511
MolBar | 1.1.2 | ... | -226 -109 -73 -39 -36 -29 -24 -5 9 10 11 11 11 12 40 71 132 145 178 256 280 357 844 1685 | -6 -4 0 4 36
16072129
MolBar | 1.1.2 | ... | -226 -109 -73 -39 -36 -29 -24 -5 9 10 11 11 11 12 40 71 132 145 178 256 280 357 844 1685 | -17 -2 0 3 26
c)
738497
MolBar | 1.1.2 | ... | ... -99 -85 -57 -39 -34 -34 -22 -21 -2 8 9 11 11 11 11 11 11 11 11 11 11 11 14 18 37 42 44 74 88 122 140 167 204 297 327 336 408 484 814 1003 1626 | -6 0 0 0 0 0 0 0 16
3628081
MolBar | 1.1.2 | ... | ... -99 -86 -58 -40 -34 -34 -22 -21 -4 8 9 11 11 11 11 11 11 11 11 11 11 11 14 19 38 42 44 74 88 120 140 168 204 297 327 336 409 481 812 1001 1638 | -6 0 0 0 0 0 0 0 16
d)
17922991
MolBar | 1.1.2 | NO2SH | 0 | -344 -15 160 234 1155 | -332 147 160 1126 | -47 32 104 157 944 | 0 0
18760928
MolBar | 1.1.2 | NO2SH | 0 | -344 -15 160 234 1155 | -332 147 160 1126 | -47 37 101 153 946 | 0 0
e)
12886903
MolBar | 1.1.2 | ... | -165 -83 -41 -38 -26 -21 -21 -8 -4 7 8 10 11 12 16 45 68 119 127 202 262 437 492 1510 | -4 0 14
17039468
MolBar | 1.1.2 | ... | -180 -83 -41 -37 -25 -22 -11 -7 -4 7 8 10 11 12 17 45 65 119 128 194 261 442 481 1531 | -4 0 14
f)
59927436
MolBar | 1.1.2 | ... | -120 -103 -82 -29 -20 -2 11 11 11 11 11 25 58 88 110 180 373 639 927 | 0 0 0 0 0 0
72597760
MolBar | 1.1.2 | ... | -120 -103 -82 -29 -20 -3 11 11 11 11 11 26 61 94 101 181 376 638 926 | 0 0 0 0 0 0
g)
644268
MolBar | 1.1.2 | O7P2H4 | 0 | -640 -412 -40 -40 -16 4 180 214 276 300 300 1363 1611 | ... | 0 0 0 0 0 0 0
21961011
MolBar | 1.1.2 | O7P2H4 | 0 | -869 -482 -40 -40 -12 -7 181 210 276 300 300 1086 2497 | ... | 0 0 0 0 0 0 0
h)
577521
MolBar | 1.1.2 | ... | -208 -112 -54 -34 -26 -22 -17 -12 -9 -6 0 8 9 11 11 11 11 15 30 49 71 87 95 121 155 211 313 348 449 729 1966 | -4 24
21767616
MolBar | 1.1.2 | ... | -208 -112 -54 -34 -26 -22 -17 -12 -9 -6 0 8 9 11 11 11 11 15 30 49 71 87 95 121 155 210 313 347 449 728 1966 | -4 24

Fig. 13 Examples of differences between MolBar and InChI in the Molecule3D dataset. Out of approximately 162 300 duplicates, there are 69 cases where there is a difference between MolBar and InChI. The examples fall into eight different categories: (a) non-central chirality (1 case) (b) stereocenters with phosphorus or sulfur (18 cases), (c) stereocenters with nitrogen (1 case), (d) different VSEPR geometries (15 cases), (e) chemically unreasonable geometries, far from the global energetic minimum (21 cases), (f) broken structures (5 cases), (g) incorrect connectivity assessment by MolBar (6 cases), (h) convergence problems in force field unification (2 cases). For all cases, except (g) only the topography and chirality barcode is shown. Example (g) shows the topology barcode and the chirality barcode.

5.3 Permutation invariance test

Registry and storage systems require a unique and unambiguous identifier, where uniqueness requires a one-to-one correspondence between the identifier and a distinct molecular structure. In the context of a molecule consisting of *n* atoms, there are *n*! possible permutations of these atoms. With

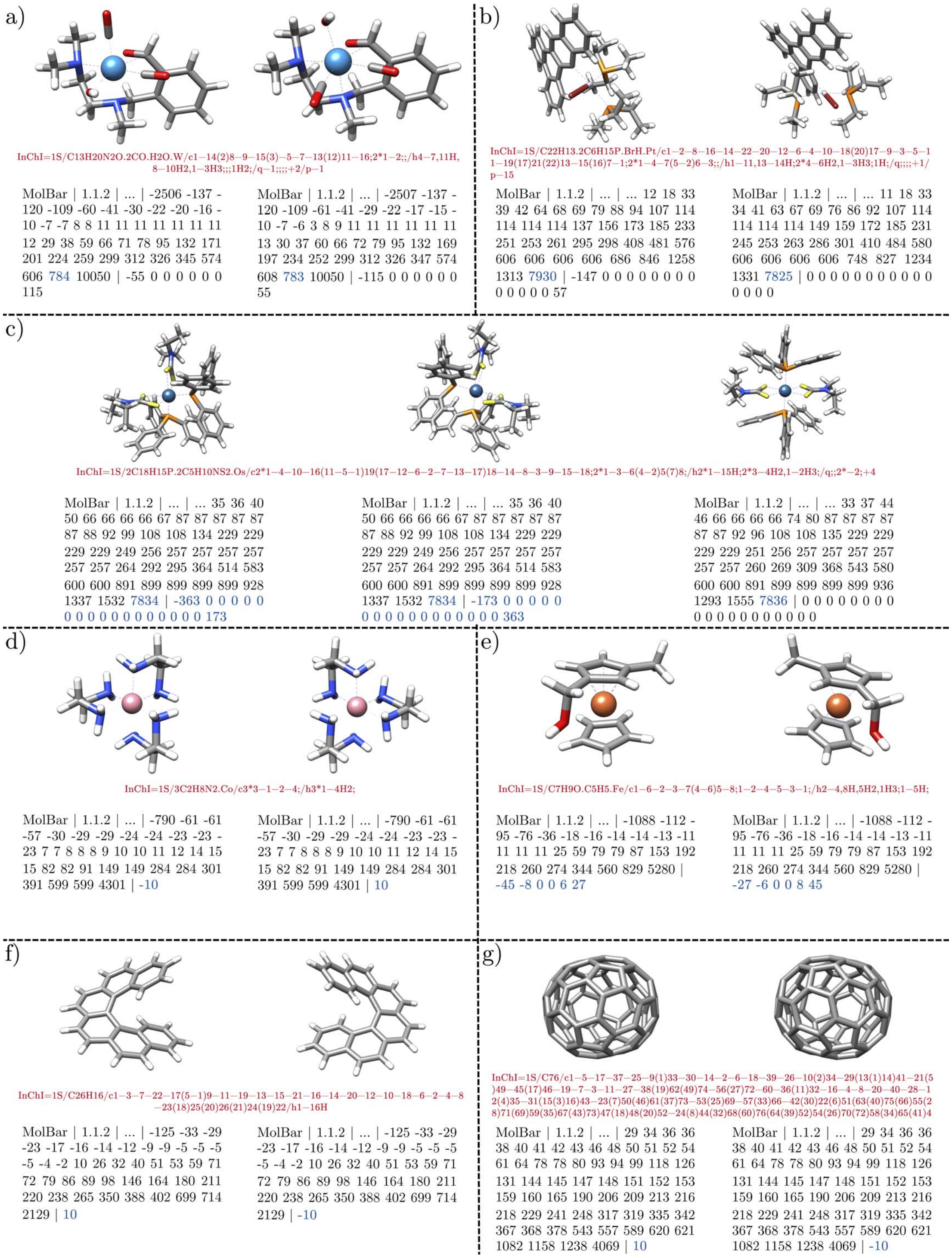
InChI, a canonicalization procedure is applied to obtain a unique numbering of the atoms. As demonstrated in previous studies, InChI is permutation-invariant.[91] To evaluate the robustness of MolBar against changes in atomic order, a permutation test was conducted on 100 000 randomly selected molecules from *Molecule3D*. The atomic numbering of these molecules was randomly permuted five times. For each permutation, the MolBar identifiers were generated and compared to the original identifiers. In the sample set, no permutation errors were found for MolBar. Dataframes containing the original and permuted geometries, as well as the calculated MolBar identifiers, are available in the ESI.† At first glance, this might seem straightforward, as eigenvalue spectra in MolBar are inherently permutation invariant. However, it is essential that the entries in the different MolBar matrices also maintain permutation invariance. For example, an identical set of flexible, freely rotatable bonds is crucial for fragmenting the molecule consistently into matching sets of fragments, regardless of atomic order. Additionally, force field constraints must be defined in a permutation-invariant manner to ensure a consistently unified structure. While the permutation invariance test shows that MolBar is generally robust against shuffling atomic order, some exceptional cases may still occur.

5.4 Examples of metal complex isomerism and non-central chirality

MolBar departs from the conventional atomic description and instead attempts to characterize the relative and absolute configuration of entire fragments. Therefore, MolBar is theoretically able to represent a wide range of molecules, including both organic and inorganic variants, and going beyond the limits of centrochirality. The data set considered so far, Molecule3D, contains mainly molecules that do not require a decentralized description, such as classical organic molecules with stereocenters as well as *E*/*Z*-configured double bonds. In the following, some cases are examined where the atomistic description reaches its limits, taken from the tmQM dataset.[92] For example, Fig. 14 shows stereoisomers involving both inorganic and organic molecules, as well as examples of planar and axial chirality. The first example (Fig. 14a) shows two stereoisomers of a tungsten complex with octahedral coordination geometry. For MolBar, the central fragment corresponds to the octahedron itself. The different ligand arrangements within the octahedron distinguish the stereoisomers; in particular, the position of a carbon monoxide ligand is exchanged with that of a hydroxide ligand. As a result, the interatomic distances in the topography matrix vary, allowing MolBar to discriminate between the two stereoisomers. A parallel scenario unfolds in Fig. 14b, where a square-planar platinum complex exhibits *cis*/*trans* isomerism. The stereoisomers in Fig. 14c show an octahedral osmium complex with two bidentate dithiocarbamate ligands and two phosphine ligands. Depending on the arrangement of the ligands, the complex can be chiral or achiral. Axial chirality occurs when the two bidentate ligands are not in the same plane. In such cases, the complex consists of a single chiral fragment, the octahedron with the metal atom at the center. The absolute configuration spectra of the two enantiomers, therefore, differ only in sign. If both bidentate ligands are in the same plane, the complex is considered achiral. The achiral variant has different relative ligand positions within the octahedron compared to both chiral complexes, which can be seen in the topography spectrum. Achirality is captured by the absolute chirality spectrum, which consists exclusively of zeros. A similar example is shown in Fig. 14d, which shows the classic example of Δ/Λ isomerism with tris(ethylenediamine)cobalt(III). In contrast to the previous case, the entire molecule corresponds to a singular fragment characterized by a single eigenvalue in the absolute configuration spectrum. Another stereogenic unit can occur in metallocenes, as shown in Fig. 14e, where complexation of the iron atom to one side of a suitably substituted aromatic ring results in planar chirality. The chirality of the hexahelicene in Fig. 14f results from the tendency to relieve steric tension. This leads to a right- or left-handed helical form within the polyaromatic ring system. Another notable example is the chiral fullerene C_{76} (Fig. 14g), which exists in two enantiomeric forms. The absolute configuration of hexahelicenes and fullerenes is correctly described by the Osipov–Pickup–Dunmur index, which characterizes a single fragment and leads to a differentiation of the respective enantiomers with MolBar.

5.5 Conformational invariance and application to chemical space exploration

The fragmentation of molecules into rigid substructures was performed to ensure the MolBar identifier remains conformationally invariant. An additional step of force field unification was undertaken to map structural variances of the input to a consistent structure. To demonstrate the practical application of this concept, molecular dynamics (MD) simulations were conducted on two examples. The first example, labeled 3 628 081 in Fig. 13, is an organic molecule featuring freely rotatable bonds and two connected rings. The second example, an achiral osmium complex representative of inorganic chemistry, was already discussed in Fig. 14c. The MD simulations were conducted using the GFN2-xTB level of theory with *xtb* 6.6.1 and standard settings.[28] The results are illustrated in Fig. 15, where the energies of different MD snapshots are presented. The MolBar was calculated for each snapshot and compared to the original MolBar of the input structure. For both examples, the MolBar values remained constant throughout the simulation, demonstrating the conformational invariance of MolBar for those examples. This demonstrates that the chosen reference values in the force field, such as VSEPR geometries, are robust for the two examples. Despite fluctuations in geometry during the simulations, the same force field parameters, including the angles derived from VSEPR theory, were consistently applied. This resulted in the exact same unified structure after force field optimization, ensuring identical MolBar identifiers throughout the molecular dynamics simulation. This robustness of the references is also proven by the duplication test as otherwise MolBar would not be able to find the same duplicates as InChI. However, the example in Fig. 13h may also be considered as



a)

```
InChI=1S/C13H20N2O.2CO.H2O.W/c1–14(2)8–9–15(3)–5–7–13(12)11–16;2*1–2;;/h4–7,11H,
8–10H2,1–3H3;;;1H2;/q–1;;;;+2/p–1
```

```
MolBar | 1.1.2 | ... | -2506 -137 -
120 -109 -60 -41 -30 -22 -20 -16 -
10 -7 -7 8 8 11 11 11 11 11 11 11
12 29 38 59 66 71 78 95 132 171
201 224 259 299 312 326 345 574
606 784 10050 | -55 0 0 0 0 0 0
115
```

```
MolBar | 1.1.2 | ... | -2507 -137 -
120 -109 -61 -41 -29 -22 -17 -15 -
10 -7 -6 3 8 9 11 11 11 11 11 11
13 30 37 60 66 72 79 95 132 169
197 234 252 299 312 326 347 574
608 783 10050 | -115 0 0 0 0 0 0
55
```

b)

```
InChI=1S/C22H13.2C6H15P.BrH.Pt/c1–2–8–16–14–22–20–12–6–4–10–18(20)17–9–3–5–1
1–19(17)21(22)13–15(16)7–1;2*1–4–7(5–2)6–3;;/h1–11,13–14H;2*4–6H2,1–3H3;1H;/q;;;;+1/
p–15
```

```
MolBar | 1.1.2 | ... | ... 12 18 33
39 42 64 68 69 79 88 94 107 114
114 114 114 137 156 173 185 233
251 253 261 295 298 408 481 576
606 606 606 606 686 846 1258
1313 7930 | -147 0 0 0 0 0 0 0
0 0 0 0 0 57
```

```
MolBar | 1.1.2 | ... | ... 11 18 33
34 41 63 67 69 76 86 92 107 114
114 114 114 149 159 172 185 231
245 253 263 286 301 410 484 580
606 606 606 606 748 827 1234
1331 7825 | 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0
```

c)

```
InChI=1S/2C18H15P.2C5H10NS2.Os/c2*1–4–10–16(11–5–1)19(17–12–6–2–7–13–17)18–14–8–3–9–15–18;2*1–3–6(4–2)5(7)8;/h2*1–15H;2*3–4H2,1–2H3;/q;;2*–2;+4
```

```
MolBar | 1.1.2 | ... | ... 35 36 40
50 66 66 66 66 67 87 87 87 87 87
87 88 92 99 108 108 134 229 229
229 229 249 256 257 257 257 257
257 257 264 292 295 364 514 583
600 600 891 899 899 899 899 928
1337 1532 7834 | -363 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 173
```

```
MolBar | 1.1.2 | ... | ... 35 36 40
50 66 66 66 66 67 87 87 87 87 87
87 88 92 99 108 108 134 229 229
229 229 249 256 257 257 257 257
257 257 264 292 295 364 514 583
600 600 891 899 899 899 899 928
1337 1532 7834 | -173 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 363
```

```
MolBar | 1.1.2 | ... | ... 33 37 44
46 66 66 66 66 74 80 87 87 87 87
87 87 92 96 108 108 135 229 229
229 229 251 256 257 257 257 257
257 257 260 269 309 368 543 580
600 600 891 899 899 899 899 936
1293 1555 7836 | 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0
```

d)

```
InChI=1S/3C2H8N2.Co/c3*3–1–2–4;/h3*1–4H2;
```

```
MolBar | 1.1.2 | ... | -790 -61 -61
-57 -30 -29 -29 -24 -24 -23 -23 -
23 7 7 8 8 8 9 10 10 11 12 14 15
15 82 82 91 149 149 284 284 301
391 599 599 4301 | -10
```

```
MolBar | 1.1.2 | ... | -790 -61 -61
-57 -30 -29 -29 -24 -24 -23 -23 -
23 7 7 8 8 8 9 10 10 11 12 14 15
15 82 82 91 149 149 284 284 301
391 599 599 4301 | 10
```

e)

```
InChI=1S/C7H9O.C5H5.Fe/c1–6–2–3–7(4–6)5–8;1–2–4–5–3–1;/h2–4,8H,5H2,1H3;1–5H;
```

```
MolBar | 1.1.2 | ... | -1088 -112 -
95 -76 -36 -18 -16 -14 -14 -13 -11
11 11 11 25 59 79 79 87 153 192
218 260 274 344 560 829 5280 |
-45 -8 0 0 6 27
```

```
MolBar | 1.1.2 | ... | -1088 -112 -
95 -76 -36 -18 -16 -14 -14 -13 -11
11 11 11 25 59 79 79 87 153 192
218 260 274 344 560 829 5280 |
-27 -6 0 0 8 45
```

f)

```
InChI=1S/C26H16/c1–3–7–22–17(5–1)9–11–19–13–15–21–16–14–20–12–10–18–6–2–4–8
–23(18)25(20)26(21)24(19)22/h1–16H
```

```
MolBar | 1.1.2 | ... | -125 -33 -29
-23 -17 -16 -14 -12 -9 -9 -5 -5 -5
-5 -4 -2 10 26 32 40 51 53 59 71
72 79 86 89 98 146 164 180 211
220 238 265 350 388 402 699 714
2129 | 10
```

```
MolBar | 1.1.2 | ... | -125 -33 -29
-23 -17 -16 -14 -12 -9 -9 -5 -5 -5
-5 -4 -2 10 26 32 40 51 53 59 71
72 79 86 89 98 146 164 180 211
220 238 265 350 388 402 699 714
2129 | -10
```

g)

```
InChI=1S/C76/c1–5–17–37–25–9(1)33–30–14–2–6–18–39–26–10(2)34–29(13(1)14)41–21(5
)49–45(17)46–19–7–3–11–27–38(19)62(49)74–56(27)72–60–36(11)32–16–4–8–20–40–28–1
2(4)35–31(15(3)16)43–23(7)50(46)61(37)73–53(25)69–57(33)66–42(30)22(6)51(63(40)75(66)55(2
8)71(69)59(35)67(43)73)47(18)48(20)52–24(8)44(32)68(60)76(64(39)52)54(26)70(72)58(34)65(41)4
```

```
MolBar | 1.1.2 | ... | 29 34 36 36
38 40 41 42 43 46 48 50 51 52 54
61 64 78 78 80 93 94 99 118 126
131 144 145 147 148 151 152 153
159 160 165 190 206 209 213 216
218 229 241 248 317 319 335 342
367 368 378 543 557 589 620 621
1082 1158 1238 4069 | 10
```

```
MolBar | 1.1.2 | ... | 29 34 36 36
38 40 41 42 43 46 48 50 51 52 54
61 64 78 78 80 93 94 99 118 126
131 144 145 147 148 151 152 153
159 160 165 190 206 209 213 216
218 229 241 248 317 319 335 342
367 368 378 543 557 589 620 621
1082 1158 1238 4069 | -10
```

Fig. 14 Examples illustrating MolBar's ability to discriminate stereoisomers beyond atom-centered stereochemistry: (a) two stereoisomers of a tungsten complex that differ in ligand positions within the octahedron, (b) *cis*/*trans* isomerism in a square-planar platinum complex. (c) Chiral and achiral configurations in an octahedral osmium complex, (d) Δ/Λ isomerism in tris(ethylenediamine)cobalt(III), (e) planar chirality in metallocenes, and (f) axial chirality in hexahelicene, (g) axial chirality in fullerene C_{76}.

showing conformational variance due to convergence issues during geometry optimization. Such cases are rare but can occur.

The duplication test highlighted the importance of retaining 3D shape information for molecular identifiers used in computational chemistry. This is particularly important for identifiers used in conformational sampling or reaction discovery, where novel structures are generated as 3D coordinates in a black-box fashion. In some cases, the generated structures may not be chemically reasonable, as shown in

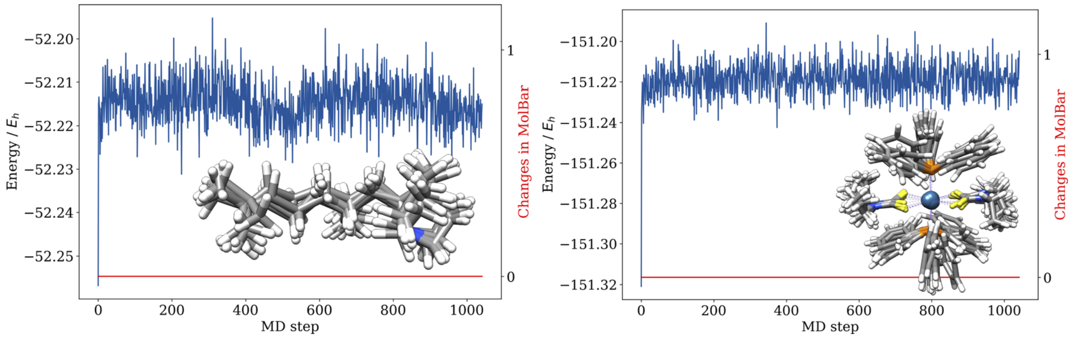


Fig. 15 Conformational invariance of MolBar: (a) MD simulation of an organic molecule with freely rotatable bonds, taken from Fig. 13c, (b) MD simulation of an achiral osmium complex, see Fig. 14c. No changes in the MolBar identifier are observed. The MDs were all carried out on the GFN2-xTB level of theory.

Fig. 13e. In other cases, the VSEPR geometry may change due to the level of theory used during the structure optimization, as shown in Fig. 13d. Given the large number of generated geometries, manual verification of their chemical correctness is impractical. Fig. 16 shows another example. The input structure is a Pt complex with a square planar geometry. The conformer ensemble was generated at the GFN-FF level of theory using the *crest* software 3.0.1,[20] with a charge of 0 and the ALPB solvation model for water.[35,93] The MolBar identifier was calculated for each conformer and compared to the MolBar of the input structure. None of the structures in the ensemble were identical to the input structure. Further evaluation revealed that the square planar geometry was not preserved in the conformer ensemble. However, all structures in the ensemble shared the same topography and topology. Unlike the input structure, the conformer ensemble consisted of structures with tetrahedral geometry. In addition, since the metal has four different ligands, the metal center became a stereocenter in the generated conformer ensemble, resulting in two enantiomers. Consequently, the entire ensemble can be separated into two enantiomeric ensembles. This example highlights the importance of identifying discrepancies in the geometry of generated structures during black-box conformational sampling.

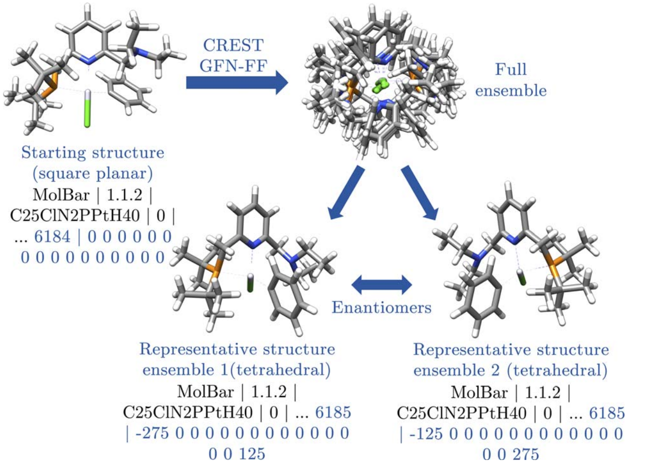


Fig. 16 Conformational sampling of a Pt complex with square planar geometry was performed using the *crest*[20] software. The input structure is displayed in the top left corner. The resulting conformer ensemble consists of structures with tetrahedral geometry, which can be divided into two enantiomeric sub-ensembles.

5.6 Computational cost

To evaluate the computational cost of MolBar, the time required to calculate the identifier was measured using a random sample of 10 000 molecules from the Molecule3D dataset. The calculations were conducted on Apple Silicon M1 hardware using the *molbar* Python package version 1.1.2 with the function *get_molbar_from_coordinates*. The timing measurements were performed three times for each molecule, and the results were averaged. Additionally, the average time was calculated for molecules with the same number of atoms. It is important to note that the time required to calculate the MolBar identifier depends on several factors, including the number of rings, the rigidity of the molecule, and the number of freely rotatable bonds, which can lead to varying calculation times for molecules with the same number of atoms. In result, those timings are only indicative, should be interpreted with caution and highly depend on the specific molecule. The results are presented in Fig. 17, where the number of atoms includes the heavy and hydrogen atoms.

The overall computational cost of MolBar scales to the power of 0.93 for molecules. The computational cost is higher compared to the generation of SMILES and InChI. However, for its primary application in chemical space exploration with quantum chemical or reactive force field methods, MolBar is practical and never poses a rate-limiting step in these workflows. Nonetheless, the algorithm is continuously being developed to reduce computational overhead, making it more suitable for applications outside of quantum chemistry requiring lower computational costs. The most computationally demanding aspect of MolBar is the force field geometry optimization step during structure unification, which ensures the input structure is mapped to a consistent form. The cost of the geometry optimization process largely depends on the optimizer and the convergence criteria. Currently, the Newton-CG optimizer from the *scipy* package is used.[63] More specialized convergence criteria, such as convergence based on the stability

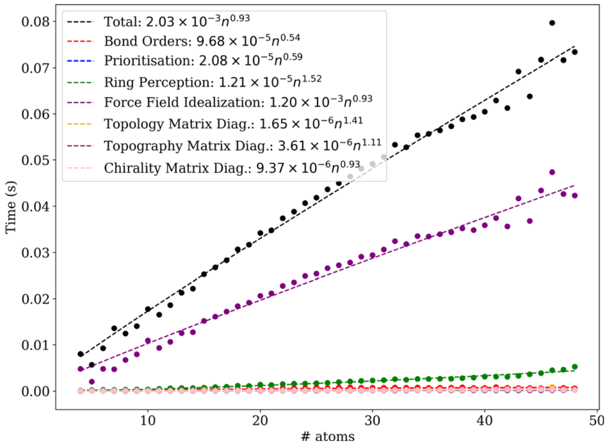


Fig. 17 Computational cost of MolBar for the Molecule3D dataset. The time required to calculate the MolBar identifier is shown for molecules with the same number of atoms.

of the rounded eigenvalue spectrum, could significantly reduce the computational burden. Additionally, while the analytical gradient and Hessian of the MolBar force field are implemented in Fortran, they are currently accessed through Python's *scipy* optimizer, which introduces overhead. Future improvements will focus on using a dedicated optimizer directly in Fortran without using external packages with MolBar-specific convergence criteria. The other steps, such as the diagonalization of the topography and chirality matrices, are less computationally demanding and negligible in comparison.

As mentioned earlier, the Molecule3D dataset consists of structures with an average of 29.11 atoms per molecule, with no molecule exceeding 45 atoms in this analysis. To account for larger systems in the computational cost analysis, we also examined a supramolecule and a protein. The Cucurbit[6]urils host with *n*-butylammonium as a guest ($BuNH_4^+$@CB6), containing 125 atoms, took an average of 2.4 seconds to compute.[94,95] Meanwhile, the photoactive yellow protein, consisting of 1931 atoms, required approximately 141 seconds on average to calculate the MolBar identifier.[96]

6 Conclusions

We introduced a unique molecular identifier named the Molecular Barcode (MolBar), which is designed to comprehensively describe both organic and inorganic molecules. MolBar includes full support for stereochemistry beyond the constraints of 2c–2e bonds and atom-centered stereochemistry. This identifier takes a unique approach to describe molecules with eigenvalue spectra from a Hückel-like adjacency matrix and Coulomb-inspired matrix. Using the Coulomb-inspired matrix, MolBar retains information about molecular shape, including local atom geometries, while mapping different conformations to a single identifier. This is achieved by fragmenting the molecule into rigid substructures and applying geometry optimization with a specialized force field for additional structural uniformity. The 3D structure of these unified fragments is then described by the Coulomb-inspired matrix. Thus, MolBar relies on a fragmentary picture rather than solely on an atomistic description, unlike standard molecular identifiers. This fragmentary approach allows MolBar to describe a wide range of molecules, both organic and inorganic, surpassing the limits of centrochirality. MolBar excels in distinguishing (in)organic molecules characterized by non-local chirality, such as axial or planar chirality in hexahelicene or chiral substituted ferrocene derivatives. MolBar distinguishes between tautomers by default, which is essential for quantum chemical structural databases since these structures have different electronic properties. The robustness of MolBar was demonstrated in deduplication tasks and permutation invariance tests on a dataset of approximately 3.9 million molecules. Differences in identifying duplicates between MolBar and InChI were observed, which are explainable by the different design strategies behind the identifiers and considerations of non-central chirality.

In future updates, we plan to address existing limitations, particularly those arising from significant deviations in coordination geometry from the VSEPR reference. Challenges also arise with η bonds, as small changes in the metal–ligand distance can easily affect the coordination sphere of the metal center. Caution is advised when dealing with metal complexes with η bonds, and additional analysis, such as examining geometry optimization trajectories for each fragment provided by the *molbar* package, should be performed if necessary.

In addition, the current rigidity analysis is based on bond orders and ring structures. Consequently, in cases of atropisomerism, such as in the case of 1,1′-bi-2-naphthol, the analysis does not account for rotational hindrance around covalent single bonds. This omission results in the loss of information about the configuration of the atropisomer, as both 2-naphthol substructures are split into two fragments, and axial chirality is not detected. Currently, the *molbar* implementation allows the user to set an additional constraint to account for this barrier. However, an upcoming update will provide automatic detection of atropisomerism and appropriate constraints.

While the procedure for defining molecular topology is generally robust, challenges in determining whether two atoms are bonded may still arise, as discussed by rare the cases. This is particularly true in situations where the bonding is ambiguous, especially when working with systems beyond organic chemistry. However, users can verify the topology through the debug options provided in the Python package.

All in all, a novel molecular identifier able to distinguish different stereoisomers based only on their 3D geometry is presented here. To our knowledge, it is more capable than any other identifier for computational chemistry currently in use. It can be used as a gatekeeper to avoid data duplicates in quantum chemical structure databases and to evaluate geometries in electronic structure theory-based exploration of chemical space.

Looking ahead, our future developments aim to introduce a MolBar version that supports Molfiles without 3D coordinates, along with the ability to convert MolBar strings back to the input structure. Additionally, our next focus will be on streamlining the MolBar string, as topology information is redundantly encoded in the topography spectrum by force field bond constraints. Nevertheless, the topology spectrum will be retained for the time being, as we are considering the potential benefits of back-converting from this spectrum.

Author contributions

C. B. conceived the initial concepts of the MolBar identifier and functioned as academic supervisor. Both authors contributed equally to the further conceptual development of the MolBar algorithm in its current form. The code development including the derivation of the unifying force field, implementation and benchmark of the MolBar package was carried out by N. v. S. He also prepared the manuscript and all figures with support from C. B. All authors read and approved the final manuscript.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

C. B. gratefully acknowledges the support of the Ministry of Culture and Science of the State of North Rhine-Westphalia (MKW) through the NRW Rückkehrprogram. The authors thank Sonja Herres-Pawlis and Ulrich Schatzschneider for helpful discussions. Furthermore, we thank Ying Chen, Alessandro Currà, Felix Bröcker, Hannah Schulz, and Janis Sprenger from the Bannwarth group at RWTH Aachen University as well as Johannes Gorges (University of Bonn) for testing.

References

1 H. L. Morgan, *J. Chem. Doc.*, 1965, **5**, 107–113.

2 M. F. Lynch, *J. Chem. Doc.*, 1968, **8**, 130–133.

3 A. E. Petrarca, M. F. Lynch and J. E. Rush, *J. Chem. Doc.*, 1967, **7**, 154–165.

4 T. Engel, in *Chemoinformatics: Basic Concepts and Methods*, ed. T. Engel and J. Gasteiger, Wiley-VCH, Weinheim, 1st edn, 2018, ch. 2, p. 15.

5 A. Kekulé, *Liebigs Ann. Chem.*, 1872, **162**, 77–124.

6 G. N. Lewis, *J. Am. Chem. Soc.*, 1916, **38**, 762–785.

7 D. Weininger, *J. Chem. Inf. Comput. Sci.*, 1988, **28**, 31–36.

8 S. R. Heller, A. McNaught, I. Pletnev, S. Stein and D. Tchekhovskoi, *J. Cheminf.*, 2015, **7**, 1–34.

9 W. J. Wiswesser, *Chem. Eng. News*, 1952, **30**, 3523–3526.

10 W. J. Wiswesser, *J. Chem. Doc.*, 1968, **8**, 146–150.

11 W. J. Wiswesser, *J. Chem. Inf. Comput. Sci.*, 1982, **22**, 88–93.

12 W. A. Warr, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2011, **1**, 557–579.

13 S. Ash, M. A. Cline, R. W. Homer, T. Hurst and G. B. Smith, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 71–79.

14 R. W. Homer, J. Swanson, R. J. Jilek, T. Hurst and R. D. Clark, *J. Chem. Inf. Model.*, 2008, **48**, 2294–2307.

15 J. M. Barnard, C. J. Jochum and S. M. Welford, in *Chemical Structure Information Systems*, ed. W. A. Warr, ACS Publications, 1989, ch. 8, pp. 76–81.

16 T. Verstraelen, W. Adams, L. Pujal, A. Tehrani, B. D. Kelly, L. Macaya, F. Meng, M. Richer, R. Hernández-Esparza, X. D. Yang, et al., *J. Comput. Chem.*, 2021, **42**, 458–464.

17 L.-P. Wang, A. Titov, R. McGibbon, F. Liu, V. S. Pande and T. J. Martínez, *Nat. Chem.*, 2014, **6**, 1044–1048.

18 R. Xu, J. Meisner, A. M. Chang, K. C. Thompson and T. J. Martínez, *Chem. Sci.*, 2023, **14**, 7447–7464.

19 G. N. Simm, A. C. Vaucher and M. Reiher, *J. Phys. Chem. A*, 2018, **123**, 385–399.

20 P. Pracht, F. Bohle and S. Grimme, *Phys. Chem. Chem. Phys.*, 2020, **22**, 7169–7192.

21 S. Grimme, *Angew. Chem., Int. Ed.*, 2013, **52**, 6306–6312.

22 C. A. Bauer and S. Grimme, *J. Phys. Chem. A*, 2016, **120**, 3755–3766.

23 M. Quirós, S. Gražulis, S. Girdzijauskaitė, A. Merkys and A. Vaitkus, *J. Cheminf.*, 2018, **10**, 1–17.

24 A. H. Cheng, A. Cai, S. Miret, G. Malkomes, M. Phielipp and A. Aspuru-Guzik, *Digital Discovery*, 2023, **2**, 748–758.

25 R. Merris, *Linear Algebra Appl.*, 1994, **197**, 143–176.

26 G. Strang, *Linear Algebra and Its Applications*, Harcourt Brace Jovanovich, San Diego, 1988.

27 M. Rupp, A. Tkatchenko, K.-R. Müller and O. A. Von Lilienfeld, *Phys. Rev. Lett.*, 2012, **108**, 058301.

28 C. Bannwarth, S. Ehlert and S. Grimme, *J. Chem. Theory Comput.*, 2019, **15**, 1652–1671.

29 K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko and K.-R. Müller, *J. Chem. Theory Comput.*, 2013, **9**, 3404–3419.

30 G. Ferré, T. Haut and K. Barros, *J. Chem. Phys.*, 2017, **146**, 114107.

31 F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley and O. A. von Lilienfeld, *J. Chem. Theory Comput.*, 2017, **13**, 5255–5264.

32 M. F. Langer, A. Goeßmann and M. Rupp, *npj Comput. Mater.*, 2022, **8**, 41.

33 J. Schrier, *J. Chem. Inf. Model.*, 2020, **60**, 3804–3811.

34 L. Poppe and J. Nagy and G. Hornyánszky and Z. Boros, in *Stereochemistry and Stereoselective Synthesis: An Introduction*, ed. L. Poppe and J. Nagy, Wiley-VCH, Weinheim, 2018, ch. 2, pp. 44–46.

35 S. Spicher and S. Grimme, *Angew. Chem., Int. Ed.*, 2020, **59**, 15665–15673.

36 M. Osipov, B. Pickup and D. Dunmur, *Mol. Phys.*, 1995, **84**, 1193–1206.

37 S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *J. Chem. Phys.*, 2010, **132**, 154104.

38 S. Grimme, J. G. Brandenburg, C. Bannwarth and A. Hansen, *J. Chem. Phys.*, 2015, **143**, 054107.

39 J. G. Brandenburg, C. Bannwarth, A. Hansen and S. Grimme, *J. Chem. Phys.*, 2018, **148**, 064104.

40 S. Spicher, M. Bursch and S. Grimme, *J. Phys. Chem. C*, 2020, **124**, 27529–27541.

41 M. Bursch, A. Hansen, P. Pracht, J. T. Kohn and S. Grimme, *Phys. Chem. Chem. Phys.*, 2021, **23**, 287–299.

42 S. Spicher and S. Grimme, *J. Phys. Chem. Lett.*, 2020, **11**, 6606–6611.

43 S. S. Shaik and P. C. Hiberty, *A Chemist's Guide to Valence Bond Theory*, John Wiley & Sons, Hoboken, NJ, 2007.

44 K. B. Wiberg, *Tetrahedron*, 1968, **24**, 1083–1096.

45 I. Mayer, in *Simple Theorems, Proofs, and Derivations in Quantum Chemistry*, Springer US, Boston, MA, 2003, pp. 227–249.

46 Y. Kim and W. Y. Kim, *Bull. Korean Chem. Soc.*, 2015, **36**, 1769–1777.

47 Jensen Group, *xyz2mol*, https://github.com/jensengroup/xyz2mol, accessed: 25 September 2024.

48 RDKit Community, *RDKit: Open-source cheminformatics*, https://www.rdkit.org, accessed: 25 September 2024.

49 T. Kavitha, K. Mehlhorn, D. Michail and K. E. Paluch, *Algorithmica*, 2008, **52**, 333–349.

50 A. A. Hagberg, D. A. Schult and P. J. Swart, Exploring Network Structure, Dynamics, and Function using NetworkX, *Python in Science Conference*, Pasadena, CA, USA, 2008, pp. 11–15.

51 E. W. Dijkstra, *Numer. Math.*, 1959, **1**, 269–271.

52 P. R. Ashton, C. L. Brown, E. J. T. Chrystal, T. T. Goodnow, A. E. Kaifer, K. P. Parry, D. Philp, A. M. Z. Slawin, N. Spencer, J. F. Stoddart and D. J. Williams, *J. Chem. Soc., Chem. Commun.*, 1991, 634–639.

53 I. T. Harrison and S. Harrison, *J. Am. Chem. Soc.*, 1967, **89**, 5723–5724.

54 O. Lukin and F. Vögtle, *Angew. Chem., Int. Ed.*, 2005, **44**, 1456–1477.

55 N. Mardirossian and M. Head-Gordon, *J. Chem. Phys.*, 2016, **144**, 214110.

56 C. Zhao, R. Wu, S. Zhang and X. Hong, *J. Phys. Chem. A*, 2023, **127**, 6791–6803.

57 F. Jensen, *Introduction to Computational Chemistry*, John Wiley & Sons, Hoboken, NJ, 2nd edn, 2017.

58 P. Pyykkö and M. Atsumi, *Chem. - Eur. J.*, 2009, **15**, 186–197.

59 R. J. Gillespie, *Chem. Soc. Rev.*, 1992, **21**, 59–69.

60 A. P. Bartók, R. Kondor and G. Csányi, *Phys. Rev. B*, 2013, **87**, 184115.

61 W. Kabsch, *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.*, 1976, **32**, 922–923.

62 J. Nocedal and S. J. Wright, in *Numerical Optimization*, Springer New York, New York, NY, 2006, pp. 101–134.

63 P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt and SciPy 1.0 Contributors, *Nat. Methods*, 2020, **17**, 261–272.

64 G. Millar, N. Weinberg and K. Mislow, *Mol. Phys.*, 2005, **103**, 2769–2772.

65 A. B. Buda and K. Mislow, *J. Am. Chem. Soc.*, 1992, **114**, 6006–6012.

66 A. B. Buda, T. A. der Heyde and K. Mislow, *Angew. Chem., Int. Ed.*, 1992, **31**, 989–1007.

67 A. Ferrarini and P. L. Nordio, *J. Chem. Soc., Perkin Trans. 2*, 1998, (2), 455–460.

68 A. V. Luzanov and E. N. Babich, *J. Mol. Struct.: THEOCHEM*, 1995, **333**, 279–290.

69 R. A. Sayle, *J. Comput.-Aided Mol. Des.*, 2010, **24**, 485–496.

70 N. Schneider, R. A. Sayle and G. A. Landrum, *J. Chem. Inf. Model.*, 2015, **55**, 2111–2120.

71 M. Randić, *J. Chem. Inf. Comput. Sci.*, 1975, **15**, 105–108.

72 M. Randic, G. M. Brissey and C. L. Wilkins, *J. Chem. Inf. Comput. Sci.*, 1981, **21**, 52–59.

73 Z. Ouyang, S. Yuan, J. Brandt and C. Zheng, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 299–303.

74 R. Sure, A. Hansen, P. Schwerdtfeger and S. Grimme, *Phys. Chem. Chem. Phys.*, 2017, **19**, 14296–14305.

75 R. C. Wilson and P. Zhu, *Pattern Recognit.*, 2008, **41**, 2833–2841.

76 Z. Xu, Y. Luo, X. Zhang, X. Xu, Y. Xie, M. Liu, K. Dickerson, C. Deng, M. Nakata and S. Ji, *arXiv*, preprint, 2021, DOI: 10.48550/arXiv.2110.01717.v1.

77 M. Nakata and T. Shimazaki, *J. Chem. Inf. Model.*, 2017, **57**, 1300–1308.

78 A. D. Becke, *J. Chem. Phys.*, 1992, **96**, 2155–2160.

79 C. Lee, W. Yang and R. G. Parr, *Phys. Rev. B*, 1988, **37**, 785.

80 S. H. Vosko, L. Wilk and M. Nusair, *Can. J. Phys.*, 1980, **58**, 1200–1211.

81 P. J. Stephens, F. J. Devlin, C. F. Chabalowski and M. J. Frisch, *J. Phys. Chem.*, 1994, **98**, 11623–11627.

82 J. S. Binkley and J. A. Pople, *J. Chem. Phys.*, 1977, **66**, 879–880.

83 J. D. Dill and J. A. Pople, *J. Chem. Phys.*, 1975, **62**, 2921–2923.

84 R. Ditchfield, W. J. Hehre and J. A. Pople, *J. Chem. Phys.*, 1971, **54**, 724–728.

85 M. M. Francl, W. J. Pietro, W. J. Hehre, J. S. Binkley, M. S. Gordon, D. J. DeFrees and J. A. Pople, *J. Chem. Phys.*, 1982, **77**, 3654–3665.

86 M. S. Gordon, J. S. Binkley, J. A. Pople, W. J. Pietro and W. J. Hehre, *J. Am. Chem. Soc.*, 1982, **104**, 2797–2803.

87 P. C. Hariharan and J. A. Pople, *Theor. Chem. Acc.*, 1973, **28**, 213–222.

88 W. J. Hehre, R. Ditchfield and J. A. Pople, *J. Chem. Phys.*, 1972, **56**, 2257–2261.

89 V. A. Rassolov, J. A. Pople, M. A. Ratner and T. L. Windus, *J. Chem. Phys.*, 1998, **109**, 1223–1229.

90 V. A. Rassolov, M. A. Ratner, J. A. Pople, P. C. Redfern and L. A. Curtiss, *J. Comput. Chem.*, 2001, **22**, 976–984.

91 J. M. Goodman, I. Pletnev, P. Thiessen, E. Bolton and S. R. Heller, *J. Cheminf.*, 2021, **13**, 40.

92 D. Balcells and B. B. Skjelstad, *J. Chem. Inf. Model.*, 2020, **60**, 6135–6146.

93 S. Ehlert, M. Stahn, S. Spicher and S. Grimme, *J. Chem. Theory Comput.*, 2021, **17**, 4250–4261.

94 R. Sure and S. Grimme, *J. Chem. Theory Comput.*, 2015, **11**, 3785–3801.

95 W. L. Mock and N. Y. Shih, *J. Am. Chem. Soc.*, 1989, **111**, 2697–2699.

96 S. Rajagopal and K. Moffat, *Proc. Natl. Acad. Sci. U.S.A.*, 2003, **100**, 1649–1654.