Electronic Supplementary Material (ESI) for CrystEngComm.

Kryptoracemic Compounds Hunting and Frequency in the Cambridge Structural Database

Simon Clevers and Gerard Coquerel

Supplementary material

SI-1 – Duplicates in CSD.

The following table gives the percentage of duplicates found in different subsets:

- NC-ND-NI : Non-centrosymmetric Not disordered Not ionic structures
- NC-ND-I : Non-centrosymmetric Not disordered ionic structures
- NC-D-NI : Non-centrosymmetric disordered Not ionic structures
- NC-D-I : Non-centrosymmetric disordered ionic structures
- C-ND-I : centrosymmetric Not disordered ionic structures
- C-ND-NI : centrosymmetric Not disordered Not ionic structures
- C-D-NI : centrosymmetric disordered Not ionic structures
- C-D-I : centrosymmetric disordered ionic structures

Subset	Export from conquest (hits)	frequency before filtering	Duplicates (hits)	% of Duplicates in the subset	Nafter	frequency without duplicates
NC-ND-NI	93417	22.5%	4838	4.9%	88579	22.5%
NC-ND-I	11541	2.8%	818	7.1%	10723	2.7%
NC-D-NI	14211	3.4%	510	3.4%	13701	3.5%
NC-D-I	3292	0.8%	165	5.7%	3127	0.8%
C-ND-NI	204476	49.3%	11644	5.5%	192832	49.1%
C-ND-I	31563	7.6%	2013	6.2%	29550	7.5%
C-D-NI	44115	10.6%	1644	3.7%	42471	10.8%
C-D-I	12552	3.0%	523	3.9%	12029	3.1%
Total	415167	100%	22155	5.2%	393012	100.0%

TableSI-1 repartition of duplicates structures over different subsets export from conquest and after removing of duplicates.

SI-2– Comparison between ChiPi and ChiralFinder.

The comparison of the filtering obtained with ChiralFinder (CF) and ChiPi (CP) for the achiral, chiral, racemic meso and not-treated crystal classes are summarized in following table:

Subset		Achiral	Chiral	Racemic	Meso	Not-treated
NC-ND-NI	CF	27.3%	65.3%	4.5%	0.9%	2.0%
	CP	**26.4%**	**67.0%**	**4.8%**	**0.8%**	**0.8%**
NC-ND-I	CF	42.0%	45.2%	2.1%	0.6%	10.1%
	CP	**44.6%**	**49.7%**	**2.9%**	**1.0%**	**1.5%**
NC-D-NI	CF	27.0%	62.0%	3.6%	0.8%	6.6%
	CP	**26.7%**	**63.0%**	**4.1%**	**1.9%**	**3.9%**
NC-D-I	CF	37.0%	37.2%	1.8%	0.8%	23.3%
	CP	**46.7%**	**41.6%**	**2.8%**	**2.3%**	**6.0%**
C-ND-NI	CF	67.1%	0%	26.0%	2.6%	4.3%
	CP	**68.6%**	**0.0%**	**27.5%**	**2.0%**	**1.9%**
C-ND-I	CF	77.3%	0.0%	7.7%	1.4%	13.6%
	CP	**87.2%**	**0.0%**	**9.1%**	**1.8%**	**1.9%**
C-D-NI	CF	62.9%	0.0%	20.5%	3.6%	13.1%
	CP	**66.6%**	**0.0%**	**21.8%**	**3.3%**	**8.3%**
C-D-I	CF	61.2%	0.0%	6.2%	2.0%	30.5%
	CP	**81.4%**	**0.0%**	**8.4%**	**2.6%**	**7.6%**

Table SI-2 Crystal class frequencies for the different subset from ChiralFinder (CF) program and ChiPi (CP) The different subsets are defined in SI-1 section.

SI – 3 Note about the determination of crystal chirality by ChiPi:

Two different notions must not be confused in the following: (i) the chirality of AU that represents the relation between the molecules in the AU and (ii) the chirality of the structure that represents the relation between molecules in the unit cell. For instance, a centrosymmetric crystal can be racemic with a chiral AU that contains two molecules of the same enantiomer (Z'=2).

ChiPi program can accept refcode in gcd format. After removing duplicates, each subset was exported from Conquest in gcd format and analyzed by ChiPi. The program proceeds as follow:

1) For each refcode, ChiPi searches for crystallographic data in CSD.

2) Then, ChiPi analyzes each component in the asymmetric unit (AU). If the structure has an unusual Z' in combination with high symmetry (*e.g.* MUSKAM) where the AU only contains a fraction of the component, ChiPi regenerates the whole molecule for each fraction of component in the AU. ChiPi starts to assign (i) unknown bond types, (ii) missing hydrogen atoms, (iii) canonically reorder atoms of the component, (iv) check and assign a unique label for each atom of the component. If errors occur during this procedure, the structure is not treated.
3) For each component; the script determines if the molecule has a stereocenter atom by using the implemented algorithm of the Python CCDC API. The function can determine using Cahn-Ingold-Prelog (CIP) priority rules (including special CIP rules) the R or S label. In the case of stereogenic centers with more than four substituents, an M label (for 'mixed') may be assigned if the order of CIP priorities does not uniformly increase or decrease. If the assignment fails, an error is raised. The structures with error or M atom labels are not treated.
4) Once the assignment of chiral centers is achieved for each molecule, the program will determine the chemical nature of each component and will compare all similar chiral chemicals together (if more than one chiral molecule is in the AU).

   ChiPi compares each chiral center of the molecule and will assign the relationships between the same chemical molecules in the AU:

   - "racemic" if two components exhibit opposite chirality (e.g pair of enantiomers),
   - "scalemic" if the proportion of opposite enantiomers are not equimolar
   - "diastereomeric" (e.g if two components exhibit different stereochemistry for some sterocenters),

- "meso" (eg if the chiral molecule has a molecular point group incompatible with chirality)
- "chiral" if there is only one enantiomer (e.g enantiopure compound, same enantiomer with Z'>1 in the AU)
- "achiral" if the structure does not contain any chiral centers.

ChiPi cannot determined relationship between molecules exhibiting axial chirality (atropisomerism)

If the crystal structure contains more than one chiral chemical (*e.g.* a co-crystal where the two components are chiral), the program will assign the chirality of the AU for each chemical family. For instance, the racemic structure SITBOM has a [chiral, chiral] AU unit (it contains one molecule of S aspartate and one molecule of S-Arginine).

5) Based on the relationship(s) determined in step (4) and on the space group (SG) of the crystal, the program assigns the crystal structure in different classes:
    a. RACEMIC: structure containing enantiomers in racemic proportion (*e.g*. SITBOM)
    b. SCALEMIC: structures containing enantiomers in scalemic proportion (*e.g MUSKAM*.) One can note that this class is reserved to chiral space group.
    c. MESO: Structure containing MESO compound
    d. CHIRAL: Structure containing chiral molecules in enantiopure proportion.
    e. ACHIRAL: Structures with no chiral molecule
    f. ERROR- NOT TREATED: structures containing Mixed atoms or if an error occurred during previous steps.

In addition, previous crystal classes can also contain subclasses:

g. DIAST: for structures containing at least a couple of diastereoisomers (of the same chemical)
h. KRC: structure crystallizing in Sohncke SG and containing racemic proportion of enantiomers. An extend definition also include unbalanced compound (ie SCALEMIC composition) crystallizing in Sohncke SG.

SI- 4 Note about conformation comparisons

For each chiral chemical family in the AU, the program compares the conformational difference between each molecule of this family (for structures with Z'>1) and returns the root mean square deviation (rmsd) in angstrom. It will also return the Tanimoto index (Tanimoto coefficient is based on the two molecular fingerprints in the form of strings of binary variables. This calculation results in a single coefficient which effectively gives a measure of the similarity between the molecules based on their fingerprints. The similarity value will be in the range of 0 to 1, with 0 being completely dissimilar and 1 being identical, always in terms of fingerprints This coefficient was kept in the software for people that want to use it.)

For that purpose, ChiPi makes the distinction between comparison of the same enantiomer and comparison with the opposite enantiomer. In the former, ChiPi will directly perform an overlay comparison between both molecules. For the latter, the molecule is firstly inverted and it is compared to the opposite enantiomer. The rmsd comparison is a trivial task for structures where chiral molecules of the chemical equals two. Nevertheless, it starts to become complicated for higher number of molecules because the program will compare each molecule together leading to rmsd values for each couple (eg, 6 rmsd comparisons for Z'=4). In order to clarify and quickly compare the rmsd values for these structures the program

returns the mean rmsd in angstrom together with the standard deviation (std), the minimum and the maximum rmsd values. Of course, this method suffers from a loss information especially when great conformational disparities exist in the structure but should be highlighted by a high standard deviation value and greater difference between maximum and minimum values. In addition, it is possible to modify the program to keep all the rmsd values if wanted.

SI – 5 ChiPi result file and examples

ChiPi returns results a csv file format. The results are organized in different columns (see for example the organic teaching subset file). The significances of some headers are obvious, other are described in the following table:

ChiPi Header	Notice
Refcode	The CCDC refcode
Class	It is the final result: • ACHIRAL: the structure only contains achiral molecule(s) • CHIRAL: the structure contains chiral molecules(s) with enantiopure ratio, e.g 8 molecules 'R'. • MESO: the structure contains meso molecule • RACEMIC: the structure contains enantiomers with racemic composition in the AU (e.g. 1 'R'+1'S') • SCALEMIC: the structure contains enantiomers with scalemic ratio (e.g. 2'R'+1'S'). • DIAST: the structure contains diastereomeric molecules (e.g if two molecules exhibit different stereochemistry for some stereocenters), • ERROR: The structure was not treated.
Note	Information about the chemical composition of the AU. (Asymmetric Unit) (NB: 'NA' has no signification)
Kryptoracemic?	Boolean: true if the structure corresponds to a Kryptoracemic Compound
Diast?	Boolean: true if the structure contains diasterosomer(s)
Is Sohncke?	Boolean: true if the structure belongs to Sohncke Space Group set
Has Disorder?	Boolean: true if the structure exhibits disorder
Number of chiral resd	Number of chiral molecules in the AU
Number of chiral families	Number of different chiral chemical(s) in the AU. Each stereoisomer of a chemical belongs to the same chiral family. (NB: Achiral molecules are excluded)
Number of chiral center(s)	Total number of stereocenters in the AU
Chiral Atom Symb	Atom symbol(s) of chiral center (sorted)

Number of Carbon Chiral Atom	Total number of chiral carbon atoms among “Chiral Atom Symb”
Number of Chiral Center having H	Number of Chiral Center having Hydrogen atoms as one of the four substituents
R	Number of atom(s) having ‘R’ chiral configuration
S	Number of atom(s) having ‘S’ chiral configuration
M	Number of atom(s) having ‘M’ chiral configuration.
Number and type of chiral configuration	Gives the number and the chirality for enantiomers in the AU; e.g. (3, ‘RSS’), (2, ‘SRR’) means that there are 3 molecules RSS and 2 molecules SRR in the AU.
Number of chiral configurations:	Number of chiral configuration(s) for each chemical chiral family, e.g. [1, 2] means that there are two different chiral chemical family. The first family being in enantiopure ratio, the second having 2 different chiral configurations.
Number of molecule(s) in Family:	Number of molecules in chiral family.
AU_chirality:	Chirality of the AU: gives the relationship between molecules for each chiral family (e.g. [‘chiral’, ‘racemic’] means that there is a chemical in enantiopure ratio and other chemical family in racemic ratio in the AU)
point_group_of_chiral_mol:	Molecular point group of the chiral molecule(s) (sorted). NB: the determination of the point group is done for molecule having atom numbers comprise between 3 and 125 (excluded hydrogen atoms)
Unique Chemical Units	Number of different chemicals in the AU.
Note1	Note 1 to Note 4 gives supplementary information about the structure but principally serve to know which part of the code is called.
Note2	
Note 3:	
Note 4:	
Time (in seconds)	Calculation time
NB: for “rmsd”, “Tanimoto”, “Rmsd_same_chirality”, “Tanimoto_same_chirality “, “Rmsd_not_same_chirality” and “Tanimoto_not_same_chirality" headers, the values: -1" ,"ND", or "blank" should not be considered (arbitrary value): It means that no molecular overlay was performed. For each chiral family, ChiPi returns 4 values (mean, std, min and max values of overlay comparisons). Detailed examples are given at the end of the SI-5.	
rmsd (mean, std, min max value)/chiral family	Gives the result of molecular overlay(s) for all the molecules of a chiral family, i.e. ChiPi compares each enantiomer together as well as its antipode together. (e.g. for a structure containing 2’R’ and 2’S’ enantiomers, the comparisons are performed between R^1/R^2, R^1/S^1, R^1/S^2, R^2/S^1, R^2/S^2 and S^1/S^2)
Tanimoto (mean, std, min, and max value)/chiral family	Similar to “rmsd” header but return the Tanimoto index.
Rmsd_same_chirality (mean, std, min, and max value)/chiral family	"same chirality" only considers comparisons between molecules of the same chirality (e.g. for a structure containing 2’R’ and 2’S’ enantiomers, the comparison is performed between R^1/R^2 and S^1/S^2) NB: if the AU contains only 2 molecules having same chirality, the std, min and max values are not calculated (because only 1 overlay is performed)
Tanimoto_same_chirality (mean, std, min, and max value)/chiral family	Similar to “rmsd_same_chirality” header but ChiPi returns the Tanimoto index instead of rmsd value.
Rmsd_not_same_chirality (mean, std, min, and max value)/chiral family	“not same chirality" only considers comparisons between molecules of the opposite chirality (e.g. for a structure containing 2’R’ and 2’S’ enantiomers, the comparison is performed between R^1/S^1, R^1/S^2, R^2/S^1, R^2/S^2)
Tanimoto_not_same_chirality (mean, std, min, and max value)/chiral family	Similar to “rmsd_not_same_chirality” header but ChiPi returns the Tanimoto index instead of rmsd value.

Example 1: BAVLOZ

The following table sums up information obtained by ChiPi for BAVLOZ structure:

Refcode	Class	Note	N of chiral residue	N of chiral family	N and type of chiral configuration	Tanimoto (mean, sdt, min, max)/N of chiral family
BAVLOZ	CHIRAL	Enantiopure	3	1	(3, 'SRR')	0.9998, 0.0001, 0.9997, 0.9999

ChiPi detects that BAVLOZ (“refcode”) is a CHIRAL structure (“class”). This structure is enantiopure (“Note”) because the AU only contains chiral molecules. In the AU, there are 3 chiral residues (“N of chiral residue” = 3), all belonging to the same chiral family (“N of chiral family” = 1). ChiPi also determines the following chiral configuration ‘SRR’ for the three molecules in the AU (“N and type of chiral configuration” = 3, ‘SRR’). In this case, all the molecules in the AU correspond to the same enantiomer. Thus, because the antipode (‘RSS’) is not in the AU, the column “rmsd-not same chirality”, or “Tanimoto not-same chirality“ corresponding to overlay between antipodes are empty (ND or -1 values are arbitrary values and were added for consistency reasons in the code – it means that no calculation is performed). The column ‘rmsd’ (or ‘Tanimoto’) gives an average of all the molecular overlays performed by ChiPi (for BAVLOZ, it also corresponds to the rmsd same_chirality and Tanimoto_same chirality, respectively).

Therefore, there are three molecular overlays. For instance, in Tanimoto column, ChiPi calculates a mean value of 0.9998 with a standard deviation of 0.0001 and a minimum value of 0.9997 and a maximum value of 0.9999. The mean value is very closed to 1 and the standard deviation is almost 0 meaning that the conformations of the three molecules are almost identical.

Example 2: QAMZAG

More complicated cases can be encountered. For instance, the structure QAMZAG (not in the teaching subset):

Refcode	Class	Note	N of Chiral residue	N of chiral family	N and type of chiral configuration	Tanimoto (mean, std, min, max)/ N of chiral family
QAMZAG	RACEMIC	Contains also at least an other enantiomer in enatiopure ratio + Achiral molecule	12	2	(8, 'R'); (2'SS'), (2,'RR')	0.9948, 0.0053, 0.9848, 1 0.9989, 0.0007, 0.998, 0.9999

In this structure classified as RACEMIC there are 12 chiral molecules in the AU belonging to 2 different "chiral families" (moreover, the "Note" entry informs that the structure also contains achiral molecule(s)). There are 8 molecules with the 'R' chiral configuration (first family) and 2 'SS' + 2 'RR" molecules (in the second family). As noticed in the note information, it means that the structure has one family of enantiomer with enantiopure ratio (8:0). and the other chemical family with racemic ratio (2:2).

The molecular overlays are numerous: 28 for the first family and 6 for the second.

In the case of the first family only one of the two antipodes is in the structure, thus only the headers "rmsd (or Tanimoto)" and "rmsd (or Tanimoto)_same chirality" are filled.

For the second family "rsmd", "rsmd same chirality "and "rmsd_not same chirality" are likewise filled because both antipodes are in the AU. In detail, for the second family, ChiPi will performed 6 molecular overlays comparing all molecules together, 2 comparisons for the molecules of the same chirality (RR(1)/RR(2) and SS(1)/SS2) and 4 comparisons for molecules of different chiralities (RR(1)/ SS(1), RR(1)/SS2, RR2/ SS(1), RR2/SS2).

All these molecular overlays are, for the sake of clarity, summarized in four values (mean, std, min and max). For instance, the first set of values in "Tanimoto" (0.9948, 0.0053, 0.9848, 1) corresponds to the mean, std, min and max values of the 28 molecular overlays of the first

family and the second set (0.9989,0.0007,0.998,0.9999) corresponds to the 6 comparisons performed for the second chiral family.

Example 3: MUSKAM

The structure MUSKAM is a good example to explain the behavior of ChiPi for structures with unusual Z' with high symmetry and C2 molecular point group. The molecule in the crystal structure has two stereocenters. MUSKAM is a scalemic structure with 3x1/2 parts of a molecule in the asymmetric unit (AU). To determine the chirality of the AU, ChiPi will consider the complete molecules and not only the ½ parts (see Figure SI-5.1)

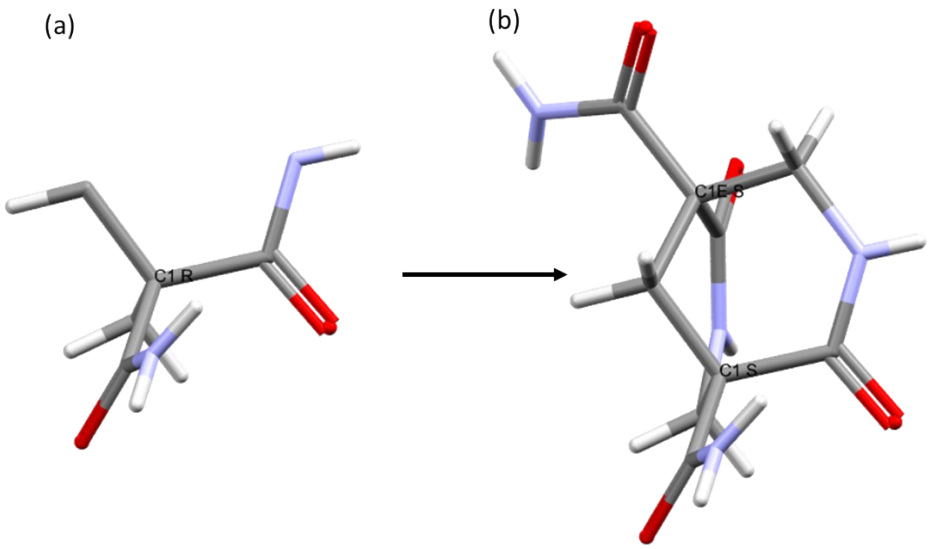


Figure SI-5.1 One of the three molecular fragment in the AU of MUSKAM (a), regeneration of the whole molecule by CCDC API (b).

The same applies for the overlay analysis. Here, the consideration or not of the molecular fragment or the whole molecule will not change the result, but it can be noticed that the chirality of atoms is not the same (see table SI-5.1). It is worth mentioning that if only one part of the molecule in the AU is considered for calculation the classification could be drastically impacted if one stereocenter coincides with an internal symmetry axis of the molecule and or

crystallographic symmetry element. For instance, YURMIH (1/2 part of molecule in the AU) belongs to RACEMIC subset but is classified in ACHIRAL subset if only one part of molecule in the AU is considered.

Atom label	½ part	Complete molecule (ChiPi)
C1	R	S
C6	R	S
C11	S	R
C1E	-	S
C6E	-	S
C11E	-	R
Crystal Chirality and detected chiral configuration by ChiPi	Scalemic (R, R,S)	Scalemic (SS, SS,RR)

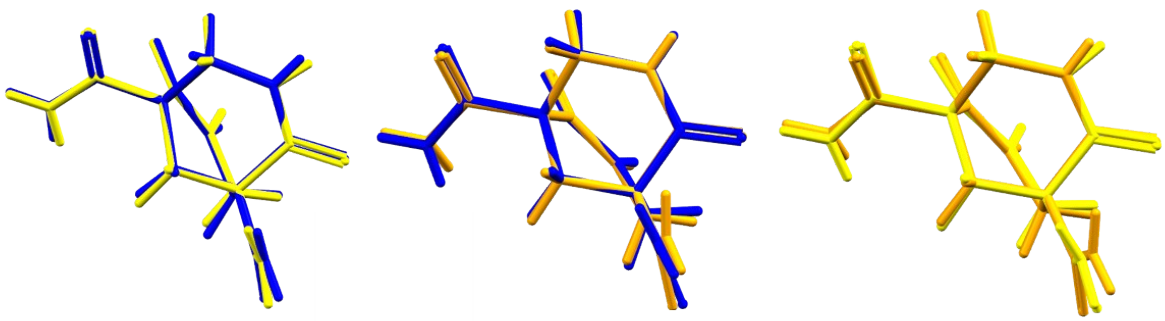
Table SI-5.1 Atom Chirality in MUSKAM considering ½ part of molecules or complete molecules in the AU.

For molecular overlay, the matching of molecules is also performed on the complete molecules and not on the fragments. For MUSKAM, ChiPi has detected three molecules with two different chiral configurations: two molecules having "SS" and one having "RR" chiral configurations. ChiPi will thus performed one comparison between SS molecules (rmsd- "same chirality") and between RR and the two SS molecules (after inversion, rmsd –"not same chirality"). ChiPi generates the following values (in angstrom):

Rmsd "same chirality"				Rmsd "not same chirality"			
Mean	Std	Min	Max	Mean	Std	Min	Max
0.1479	ND	ND	ND	0.0753	0.0318	0.0435	0.1071

Table SI-5.2 Rmsd values in angstrom for molecular overlays in MUSKAM obtained with ChiPi.

The different molecular overlays are presented in Figure SI-5.2.



RR/SS(1)
Rmsd= 0.0435 Å

RR/SS(2)
Rmsd=0.1071 Å

SS(1)/SS(2)
Rmsd=0.1479 Å

Figure SI-5.2 – Molecular overlays performed by ChiPi for MUSKAM

Therefore, the rmsd value for the overlay of the two molecules “SS” is 0.1479 Å. For overlay of molecules having different chiral configurations, Chipi performed two comparisons. As stated in the supplementary (SI-4), if the number of overlay is higher than one, the rmsd values of each comparison is not saved and, in order to clarify and quickly compare the rmsd values for structures having large number of molecules in the AU, the program returns the mean rmsd in angstrom together with the standard deviation (std), the minimum and the maximum rmsd value. Nevertheless, in MUSKAM case (Z’=1.5), three molecules are considered in the AU and thus the minimum and maximum values for the comparison between molecules having different chiralities correspond to each overlay. Thus, for one couple of antipodes the rmsd is 0.0435 Å (both are almost identical) and for other couple the rmsd is 0.1071 Å highlighting greatest molecular differences.

One can also notice that the rsmd value is higher for the overlay of the molecules having the same chirality than for overlays involving antipodes.