Thousands of exact duplicates in high-profile structural databases

Olga Anosova, Materials Innovation Factory (Liverpool, UK) and the National Institute for Theory and Mathematics in Biology (Chicago)

THE ROYAL SOCIETY

Identical numerical data?

We consider **exact duplicates** whose entries in structural databases have identical numerical data, after rounding to a few decimal places:

Google's GNoME (Nov 2023, 385K entries),

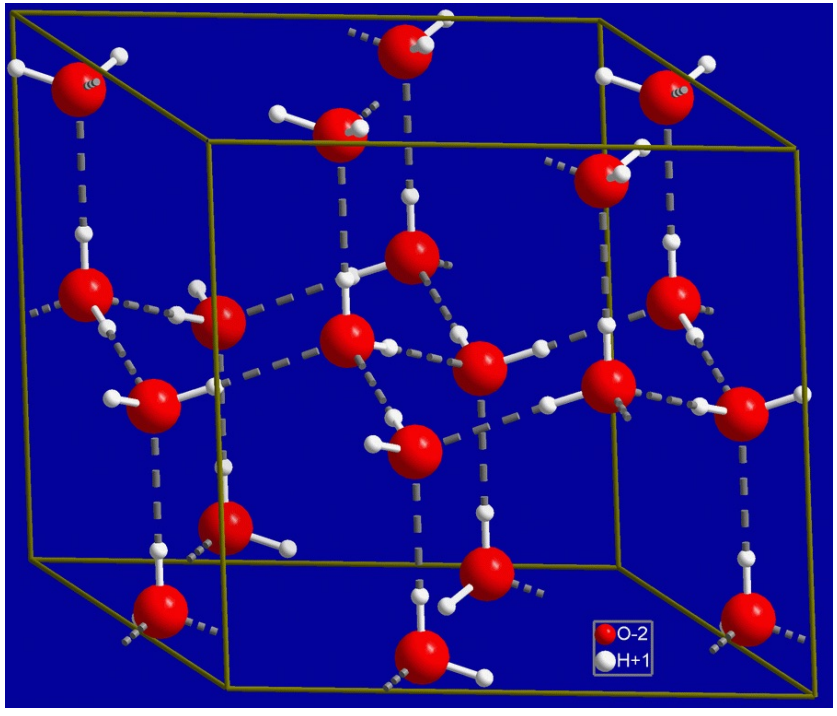
Protein Data Bank (May 2024, 220K entries),

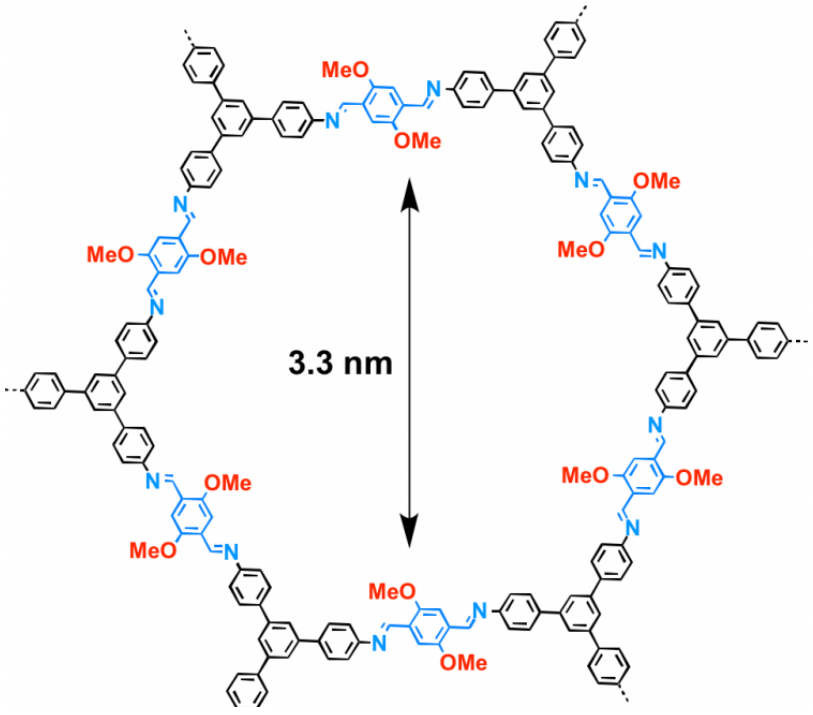
Quick discovery of near-duplicates under equivalence relations, such as rigid motion, is now possible with the theory of Geometric Invariants developed in Data Science Theory and Applications group.

All types of periodic crystals

We study solid crystalline materials at the atomic level.

What is a crystal on the left?





Left: Hexagonal ice. Right: a Metal Organic Framework.

How are these crystals represented in a digital form?

A periodic point set (crystal)

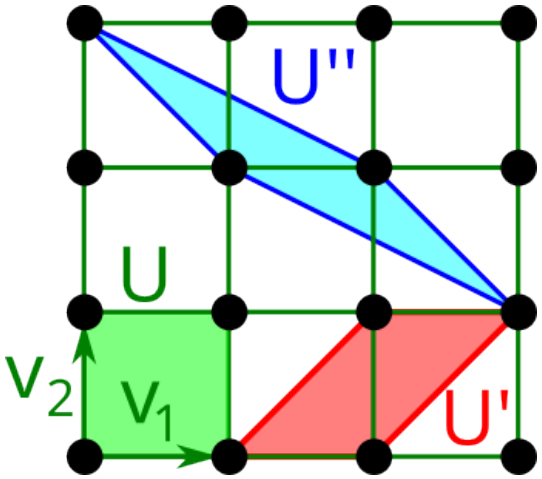
Any linear basis $v_1, \ldots, v_n$ of $\mathbb{R}^n$ defines

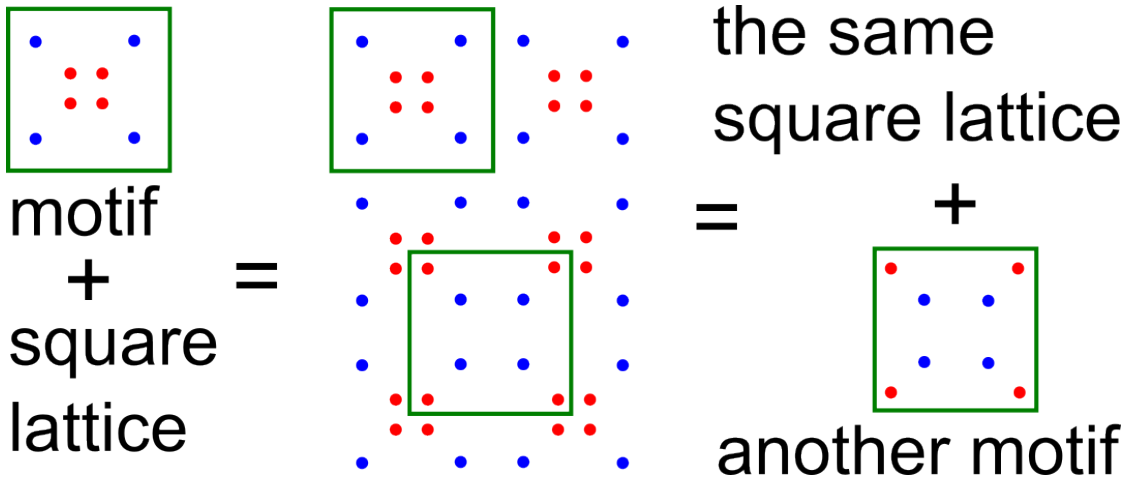
the **unit cell** $U = \{\sum_{i=1}^{n} t_i v_i \ : \ 0 \le t_i < 1\}$ and

the **lattice** $\Lambda = \{\sum_{i=1}^{n} c_i v_i \ : \ c_i \in \mathbb{Z}\}$.

For any finite **motif** $M \subset U$ of atoms, the **periodic crystal**

is the infinite set $S = \Lambda + M = \{v + p \mid v \in \Lambda, p \in M\}$.





A Crystallographic Information File (CIF) has unit cell parameters (3 lengths and 3 angles) and fractional coordinates of atoms in the cell.

```
_symmetry_space_group_name_H-M    'P1'
_symmetry_Int_Tables_number       1
_symmetry_cell_setting            triclinic
loop_
_symmetry_equiv_pos_as_xyz
  x,y,z
_cell_length_a                    12.7573
_cell_length_b                    7.4980
_cell_length_c                    12.7591
_cell_angle_alpha                 90.0000
_cell_angle_beta                  59.9998
_cell_angle_gamma                 90.0000
loop_
_atom_site_label
_atom_site_type_symbol
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
_atom_site_U_iso_or_equiv
_atom_site_adp_type
_atom_site_occupancy
C1     C     0.33327   0.57807   0.83334
```

Nature papers in November 2023

Google's GNoME: 2.2M "new" crystals, "800 years worth of human knowledge".
Berkeley's A-lab: claimed to have synthesised 43 of 58.

The review by R.Palgrave et al. (2024): "none of the materials produced by A-lab were new: the large majority were misclassified, and a smaller number were correctly identified but already known".
D.Widdowson, V.Kurlin: review in Scientific Reports (2025): this "small number"=0, all crystals had near-duplicates in the ICSD, missed by a manual search.

Google's GNoME database

GNoME paper made public 384K+ '*stable*' crystals (close to the boundary of the convex hull): any such '*stable*' crystal can be perturbed to get many more 'new (?) stable' crystals.

Review "Artificial Intelligence Driving Materials Discovery?" by A.Cheetham and R.Seshadri (2024) found "scant evidence for compounds that fulfill the trifecta of novelty, credibility, and utility".

Our reviews: Anosova et al, IUCrJ 11(4), 2024,
CSD and GNoMe: Pattern Recognition, 2025.

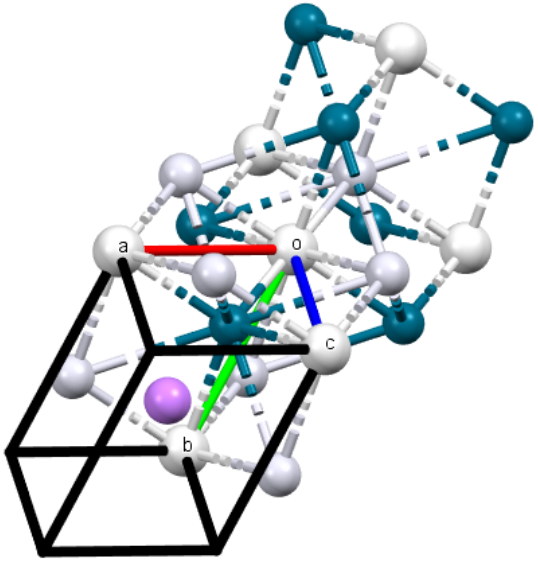
Thousands of (near)-duplicates

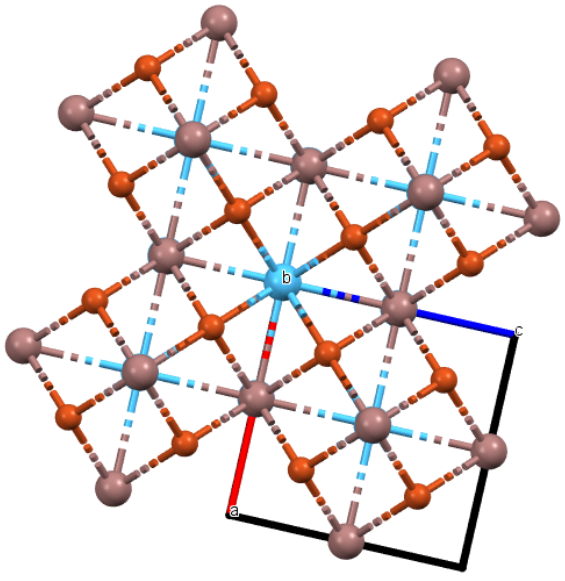
Many GNoME's crystals have geometric near- duplicates in the ICSD and Materials Project, measured by EMD on PDDs with $k = 100$.

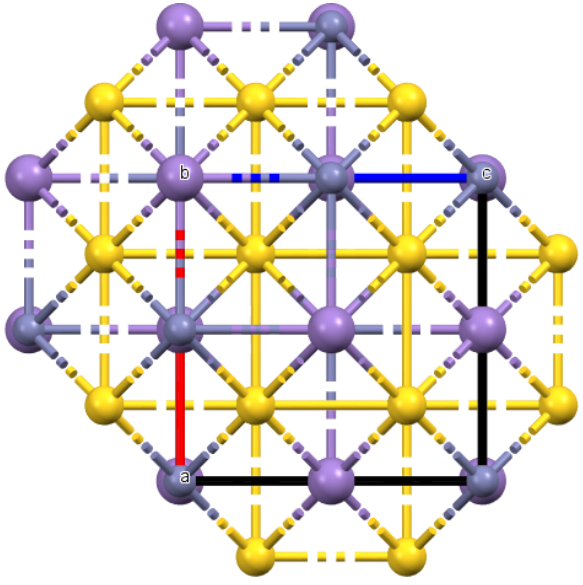
$\mathrm{EMD} \leq$	10^{-5}	10^{-4}	10^{-3}Å	0.01	0.02	0.03
ICSD	38	303	757	2454	6002	13165
Mat. Proj.	83	452	848	3457	10725	24416

Since the smallest inter-atomic distance is about 1$\mathring{A}$, any perturbations of atoms up to a small fraction of 1$\mathring{A}$ look the same when visualised.

Some near-duplicates in GNoME







These crystals are perturbations up to $10^{-4}\mathring{A}$.

crystal	database	ID	composition
1st	GNoME	01cd76eb18	LiScPdPt
2nd	ICSD	54594	$HfInCu_2$
3rd	Mat. Project	1186003	$MnZnAu_2$

Nearly identical CIFs in the GNoME

Filtering by unit cells and fractional coordinates detected numerous near-duplicates.

group size = #CIFs	CIFs are identical texts	all numbers coincide	rounding to 4 digits	rounding to 2 digits
10	0	0	0	1
9	0	1	1	0
7	0	1	1	2
6	0	2	2	4
5	0	2	3	18
4	1	8	12	92
3	43	72	96	670
2	1,089	1,481	1,932	7,856
all CIFs	2,311	3,248	4,243	18,228

Euclidean coordinates

This table shows groups of identical crystals after rounding all 6 cell parameters and Euclidean coordinates of all atoms (in Å).

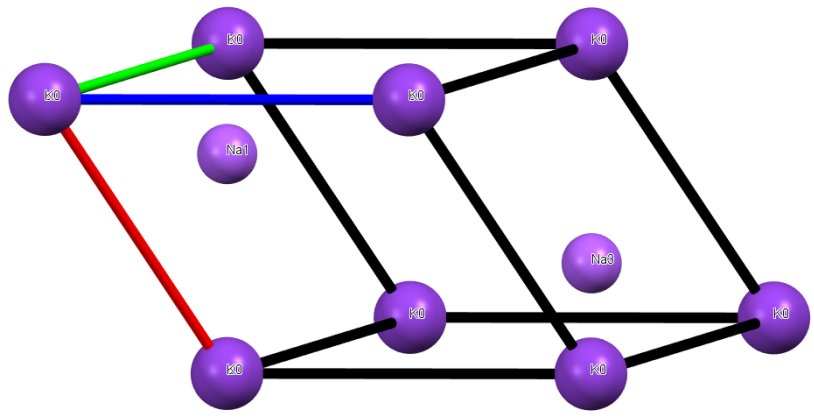
group size = # files	**rounding to 4 digits**	**rounding to 3 digits**	**rounding to 2 digits**	**rounding to 1 digit**
10 or more	0	0	0	1747
9	1	1	1	51
8	0	0	0	51
7	1	1	1	113
6	2	2	3	165
5	3	3	10	350
4	12	13	23	745
3	99	113	192	2,361
2	1,983	2,144	2,668	13,690
total	4,354	4,722	6,088	43,624

The largest group of 9+1 duplicates

GNoME id	chemical formula	all digits are equal
082738d51d	$Dy_1Y_6Ho_{13}Cd_6Ru_2$	in a group of 9
1fba8c028f	$Dy_2Y_4Ho_{14}Cd_6Ru_2$	9
39fe92e2ee	$Tb_2Y_4Ho_{14}Cd_6Ru_2$	9
6d47ae3d9f	$Tb_3Y_3Ho_{14}Cd_6Ru_2$	9
703ed1d823	$Tb_6Ho_{14}Cd_6Ru_2$	9
78fcd9d814	$Tb_1Y_5Ho_{14}Cd_6Ru_2$	9
976f8cb279	$Y_6Ho_{14}Cd_6Ru_2$	9
a30e9d8c9b	$Tb_5Y_1Ho_{14}Cd_6Ru_2$	9
b8c0e953e2	$Tb_4Y_2Ho_{14}Cd_6Ru_2$	9
a18d30a9fc	$Tb_6Ho_{14}Cd_6Re_2$	in a group of 1

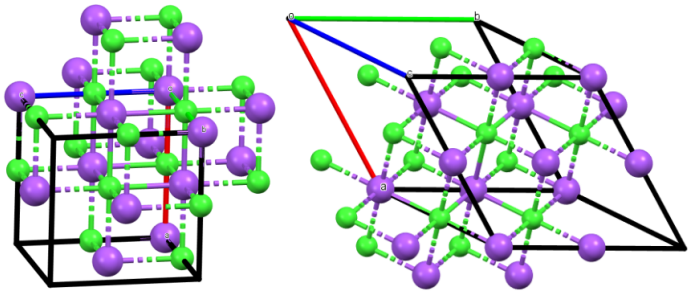
The most striking duplicates in GNoME

```
20  _atom_site_type_symbol
21  _atom_site_label
22  _atom_site_symmetry_multiplicity
23  _atom_site_fract_x
24  _atom_site_fract_y
25  _atom_site_fract_z
26  _atom_site_occupancy
27   K   K0   1   0.000000   0.000000   0.000000   1
28   Na   Na1   1   0.250000   0.250000   0.250000   1
29   Na   Na2   1   0.250000   0.250000   0.250000   1
30   Na   Na3   1   0.749999   0.749999   0.749999   1
31   Na   Na4   1   0.749999   0.749999   0.749999   1
```



Different entries cdc06a1a2a and 0e2d8f26d6 have identical CIFs and two pairs of atoms (Na1=Na2, Na3=Na4) at the same positions.

More crystals with different CIFs



Software by Tatiana K (UCL): 2.2M realistic crystals in 16 hours on a small laptop.

Drag and Drop / Select Files

Or use an example crystal instead

File Name: NaCl_1.cif

CIF contents:

```
data_NaCl
_cell_length_a 5.58812644
_cell_length_b 5.58812644
_cell_length_c 5.58812644
_cell_angle_alpha 90.00000000
_cell_angle_beta 90.00000000
_cell_angle_gamma 90.00000000
_cell_volume 174.50130186

loop_
_atom_site_type_symbol
_atom_site_label
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
Na+ Na0 0.00000000 0.00000000 0.00000000
Na+ Na1 0.00000000 0.50000000 0.50000000
Na+ Na2 0.50000000 0.00000000 0.50000000
Na+ Na3 0.50000000 0.50000000 0.00000000
Cl- Cl4 0.00000000 0.00000000 0.50000000
Cl- Cl5 0.00000000 0.50000000 0.00000000
Cl- Cl6 0.50000000 0.00000000 0.00000000
Cl- Cl7 0.50000000 0.50000000 0.50000000
```

Disguise a crystal

Input any real crystal, and this program will generate a new-looking CIF based on it!

Instructions:

1) First, either input your own CIF file or use the toggle switch to load an example crystal on the left.
2) Next, select which of the 3 transformations you would like to apply using the toggle switches below.
3) Finally, generate the new file using the buttons on the right.

Transform the basis | Extend the unit cell | Shift atoms slightly

Matrix Transformation info:

0	-1	-1
0	2	1
1	1	1

The 3x3 matrix on the left is used to change the unit cell.
The transformation preserves all geometry and volume of unit cell

*Max integer generated may not be the max number that appears in the matrix on the left because it is a combination of 2 randomly generated matrices.

Max integer generated*: 1

Unit Cell Extension info:

Input max scale factor: 2

Name of crystal within CIF | Submit

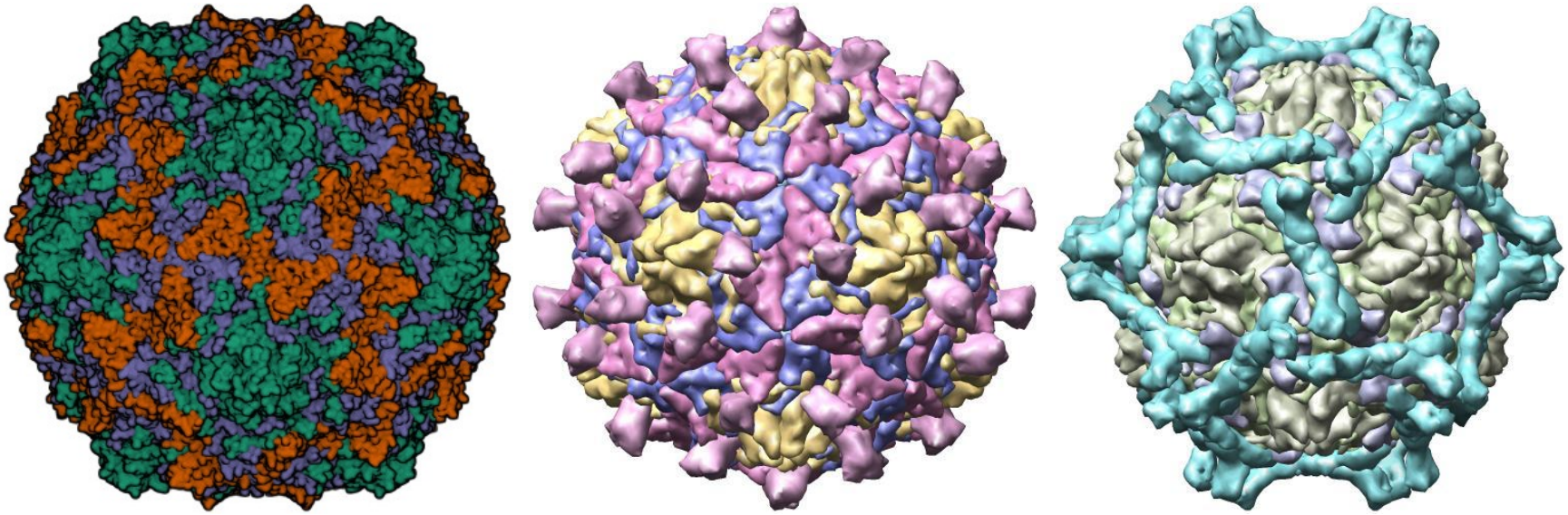
Generate new CIF | Download CIF

```
data_disguised
_cell_length_a 11.17625288
_cell_length_b 13.6880584
_cell_length_c 9.67891891
_cell_angle_alpha 19.47122063
_cell_angle_beta 54.73561032
_cell_angle_gamma 65.90515745
_cell_volume 349.00260458

loop_
_atom_site_label
_atom_site_type_symbol
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
Na+1 Na+ 0.99965316 0.00034167 0.99893752
Na+2 Na+ 0.24977722 0.50093238 0.4980313
Na+3 Na+ 0.99923756 0.49934318 0.00185047
Na+4 Na+ 0.25014822 0.99738209 0.50386005
Cl-1 Cl- 0.24966755 0.00138265 0.99773412
Cl-2 Cl- 0.9996186 0.50025784 0.50063548
Cl-3 Cl- 0.24928238 0.50000091 0.00012204
Cl-4 Cl- 0.00042397 0.99941199 0.49998203
Na+5 Na+ 0.50031141 0.000409 0.99920739
Na+6 Na+ 0.74916051 0.50053112 0.50012889
Na+7 Na+ 0.50025432 0.49917051 0.00115204
Na+8 Na+ 0.75014703 0.00129511 0.49773894
Cl-5 Cl- 0.75033765 0.00160787 0.99747864
Cl-6 Cl- 0.49993941 0.5006407 0.49900401
Cl-7 Cl- 0.75028128 0.50203556 0.99755844
Cl-8 Cl- 0.5000433 0.99887213 0.50079855
```

The Protein Data Bank (PDB)

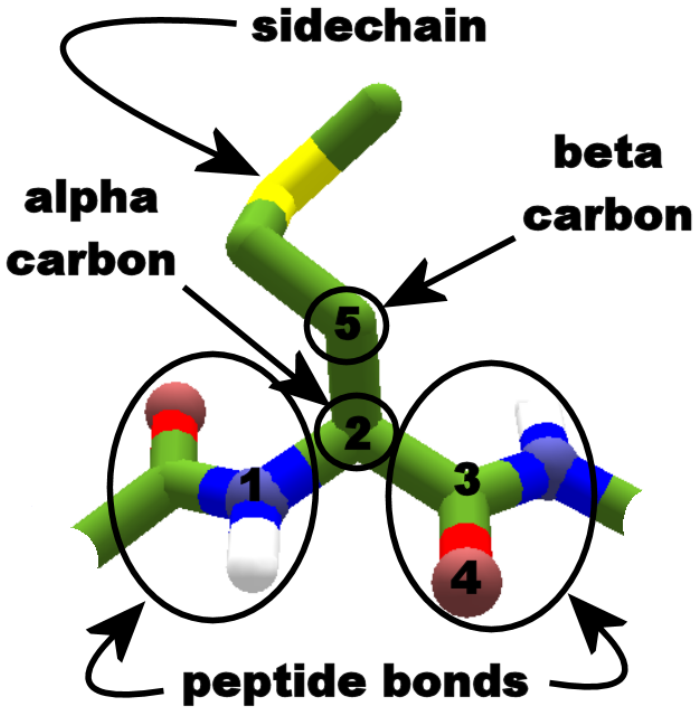
The PDB (www.rcsb.org) is a ‘gold standard’ collection of 220K+ experimental proteins given by 4-symbol codes: 1cov, 1 jew, 1m11.



Each entry can have a few entities (molecules), models (versions), and chains given by IDs.

The backbone of a protein chain

Proteins are large biomolecules consisting of one or more chains. Any **protein chain** has a sequence (**primary structure**) of residues (made of 20 standard amino acids), which are sequentially joined by peptide bonds.



A **protein backbone** is a sequence of ordered triplets of
(1) nitrogen N_i,
(2) alpha-carbon A_i,
(3) carbonyl carbon C_i,
where $i = 1, \ldots, m$ ($\#$ residues).

Anosova et al. MATCH (2025).

Exact geometric duplicates

9366 pairs turned out to have x, y, z coordinates of the main chain atoms N, C_α, C in all residues *identical to the last digit* without rigid motion.

763 such pairs are in *different PDB entries*.

In 9 pairs, geometric duplicates surprisingly differ by primary sequences of amino acids,

which seems physically impossible because replacing one amino acid with a different one should affect main atoms at least slightly.

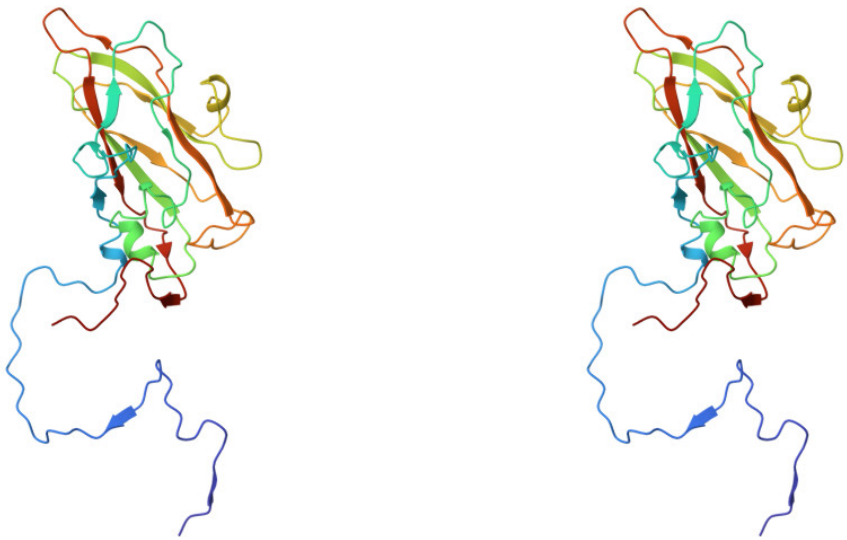
Coincidences and differences

chain 1	chain 2	# identical C_α	# different acids
1a0t-B(Q)	1oh2-B(Q)	413	9
2hqe-A	2o4x-A	217	1, GLN$\neq$GLU
1m11-D(3)	1cov-C(3)	238	72
1m11-D(3)	1jew-D(3)	238	72

Kay Diederichs accepted the duplication of 1a0t, 1oh2:
"*PDB entries can multiply on their own! ... your geometric comparison method identified an error in the PDB*."

Viruses: same or different?

1cov-C(3) and 1jew-D(3) are Coxsackieviruses CVB3, 1m11-D(3) is Echovirus E7, those are known to be different in cellular entry mechanisms and interactions with receptors, CVB3 are associated with myocarditis.



After our discussions, the PDB validation team updated other geometric duplicates 1cov, 1gli, 1ruj, 3hhb, 4rhv.

8fdz.cif vs 8fe0.cif: GLN $\leftrightarrow$ ALA

When a new protein is deposited, it is validated individually, no comparison with all structures is done. Since 2022, more CIF's were deposited, which differ as text files but all atom coordinates are identical.

```
1934 ATOM   731  C CG2  . VAL A 1 98  ? 25.693 -9.510  -28.294 1.00 33.04 ? 904  VAL A CG2
1935 ATOM   732  N N    . GLN A 1 99  ? 25.276 -8.797  -32.806 1.00 37.89 ? 905  GLN A N
1936 ATOM   733  C CA   . GLN A 1 99  ? 24.861 -9.009  -34.190 1.00 39.75 ? 905  GLN A CA
1937 ATOM   734  C C    . GLN A 1 99  ? 26.064 -9.292  -35.086 1.00 35.10 ? 905  GLN A C
1938 ATOM   735  O O    . GLN A 1 99  ? 26.027 -10.217 -35.909 1.00 33.37 ? 905  GLN A O
1939 ATOM   736  C CB   . GLN A 1 99  ? 24.064 -7.802  -34.702 1.00 37.88 ? 905  GLN A CB
1940 ATOM   737  N N    . LYS A 1 100 ? 27.152 -8.528  -34.921 1.00 32.26 ? 906  LYS A N

1916 ATOM   731  C CG2  . VAL A 1 98  ? 25.693 -9.510  -28.294 1.00 33.04 ? 904  VAL A CG2
1917 ATOM   732  N N    . ALA A 1 99  ? 25.276 -8.797  -32.806 1.00 37.89 ? 905  ALA A N
1918 ATOM   733  C CA   . ALA A 1 99  ? 24.861 -9.009  -34.190 1.00 39.75 ? 905  ALA A CA
1919 ATOM   734  C C    . ALA A 1 99  ? 26.064 -9.292  -35.086 1.00 35.10 ? 905  ALA A C
1920 ATOM   735  O O    . ALA A 1 99  ? 26.027 -10.217 -35.909 1.00 33.37 ? 905  ALA A O
1921 ATOM   736  C CB   . ALA A 1 99  ? 24.064 -7.802  -34.702 1.00 37.88 ? 905  ALA A CB
1922 ATOM   737  N N    . LYS A 1 100 ? 27.152 -8.528  -34.921 1.00 32.26 ? 906  LYS A N
```

To avoid wasting time, we should first compare chains in the PDB as lists of atomic coordinates.

A discussion in A.Wlodawer et al. Duplicate entries in the PDB. Acta Cryst D (2025).

Resolution (Resol.) and maximum deviation between two corresponding atoms (Max. dev.) are given in Å. R_{free} is shown if reported in the PDB deposition. The number of residues that are identified as different in the two depositions is indicated as No. diff. res.

PDB ID 1	PDB ID 2	No. of residues	No. diff. res.	Resol. 1 (Å)	R_{free} 1	Resol. 2 (Å)	R_{free} 2	Max. dev. (Å)
1ac4†	1aen	291	0	2.1	—	2.1	—	0
1aeb†	1aef	291	0	2.1	—	2.1	—	0
1buv	1bqq	184	0	2.75	0.248	2.75	0.248	0
1c77	1c78	130	0	2.3	0.267	2.3	0.267	0
1c79	1c78	130	0	2.3	0.267	2.3	0.267	0
1c79	1c77	130	0	2.3	0.267	2.3	0.267	0
1ffy	1qu2	917	0	2.2	0.281	2.2	0.281	0
1hdu	1hee	307	0	1.75	0.229	1.75	0.229	0

PDB ID 1	PDB ID 2	No. of residues	No. diff. res.	Resol. 1 (Å)	R_{free} 1	Resol. 2 (Å)	R_{free} 2	Max. dev. (Å)
1ac4†	1aen	291	0	2.1	—	2.1	—	0
1aeb†	1aef	291	0	2.1	—	2.1	—	0
1buv	1bqq	184	0	2.75	0.248	2.75	0.248	0
1c77	1c78	130	0	2.3	0.267	2.3	0.267	0
1c79	1c78	130	0	2.3	0.267	2.3	0.267	0
1c79	1c77	130	0	2.3	0.267	2.3	0.267	0
1ffy	1qu2	917	0	2.2	0.281	2.2	0.281	0
1hdu	1hee	307	0	1.75	0.229	1.75	0.229	0
1n6j	1n6j	93	0	2.2	0.268	2.2	0.268	0
1npw	1npa	99	0	2.0	—	2.0	—	0
1nv5	1nv1	331	0	1.9	0.245	1.9	0.255	0
1o3g	1o3f	223	0	1.55	0.205	1.55	0.213	0
1oh2_*B*	1a0t_*B*	413	9	2.4	0.246	2.4	0.246	0
1oh2_*A*	1a0t_*A*	413	0	2.4	0.246	2.4	0.246	0
1rai	1rah	310	0	2.5	—	2.5	—	0
1tc8	1po8	118	0	2.7	0.233	2.71	0.2387	0
1w21	1w20	389	0	2.08	0.195	2.08	0.195	0
1wte	1wte	272	0	1.9	0.2232	1.9	0.2232	0
1yrq	1yrq	545	0	2.1	0.22	2.1	0.22	0
1zbl	1zbl	133	0	2.2	0.253	2.2	0.253	0
1zzz	1yyy	223	0	1.9	—	2.1	—	0
2bsq	2h1o	143	0	3.0	0.271	3.0	0.269	0
2c0a	1wbh	214	0	1.55	0.217	1.55	0.217	0
2clk	1wbj	390	0	1.5	0.202	1.5	0.202	0
2fap	1nsg	107	0	2.2	0.266	2.2	0.265	0
2gzj	2gyk	130	0	1.6	0.232	1.6	0.232	0
2mll_*A*	1ce7_*A*	241	1	2.7	0.319	2.7	0.319	0
2mll_*B*	1ce7_*B*	255	0	2.7	0.319	2.7	0.319	0
2o4x	2hqe	217	1	2.0	0.2492	2.0	0.2492	0
2pgi	1b0z	442	0	2.3	0.278	2.3	0.258	0
2pr4	2f5a	213	0	2.05	0.27	2.05	0.27	0
2pt2	3f11	316	0	2.0	0.212	2.0	0.212	0
2qr1	2qrc	91	0	2.7	0.2895	2.7	0.2895	0
3g9o	3g9p	75	0	1.65	0.2029	1.65	0.2029	0
3lt9	3lt8	69	0	2.55	0.239	2.55	0.239	0
4f2u	4i9m	304	0	2.19	0.2358	2.2	0.2358	0
4f9n	4i47	246	0	2.65	0.196	2.65	0.1993	0.001
4g94	4g6d	62	0	2.0	0.2426	2.0	0.2426	0
4lf8	4lf7	235	0	3.15	0.2054	3.15	0.2054	0
4ov7	4oor	73	0	2.7	0.2312	2.7	0.2326	0
4rrw	4rrz	552	0	2.57	0.2162	2.57	0.2162	0
4y72	5hq0	264	0	2.3	0.2519	2.3	0.2518	0
5cry	5hbc	348	0	2.79	0.27287	2.79	0.2727	0.003
5hrf	8ild	178	0	2.25	0.2632	2.25	0.2632	0
5mlb	5o2s	165	0	3.22	0.2354	3.22	0.2354	0
6eda	6ebe	257	0	1.88	0.2277	1.88	0.2277	0
6jy9	6jy7	264	0	1.9	0.1895	1.8	0.1895	0
6kft	6a3b	210	0	2.51	0.2544	2.51	0.2544	0
6viu	6vb4	99	0	2.33	0.2433	2.33	0.2433	0
6xk7	6xmc	362	0	1.85	0.2423	1.85	0.2423	0
7cjn	7d9m	115	0	2.66	0.4123	2.66	0.4117	0
7kc9	7kdm	258	0	2.3	0.2377	2.3	0.2377	0
8fdz	8fe0	200	1	2.48	0.2516	2.22	0.2516	0
8fr1	6oti	257	0	2.0	0.2543	2.0	0.2543	0
8gv6	7wvd	318	0	3.4	0.2614	3.39	0.2614	0
8p1o	8r6q	126	0	2.17	0.2894	2.17	0.2894	0

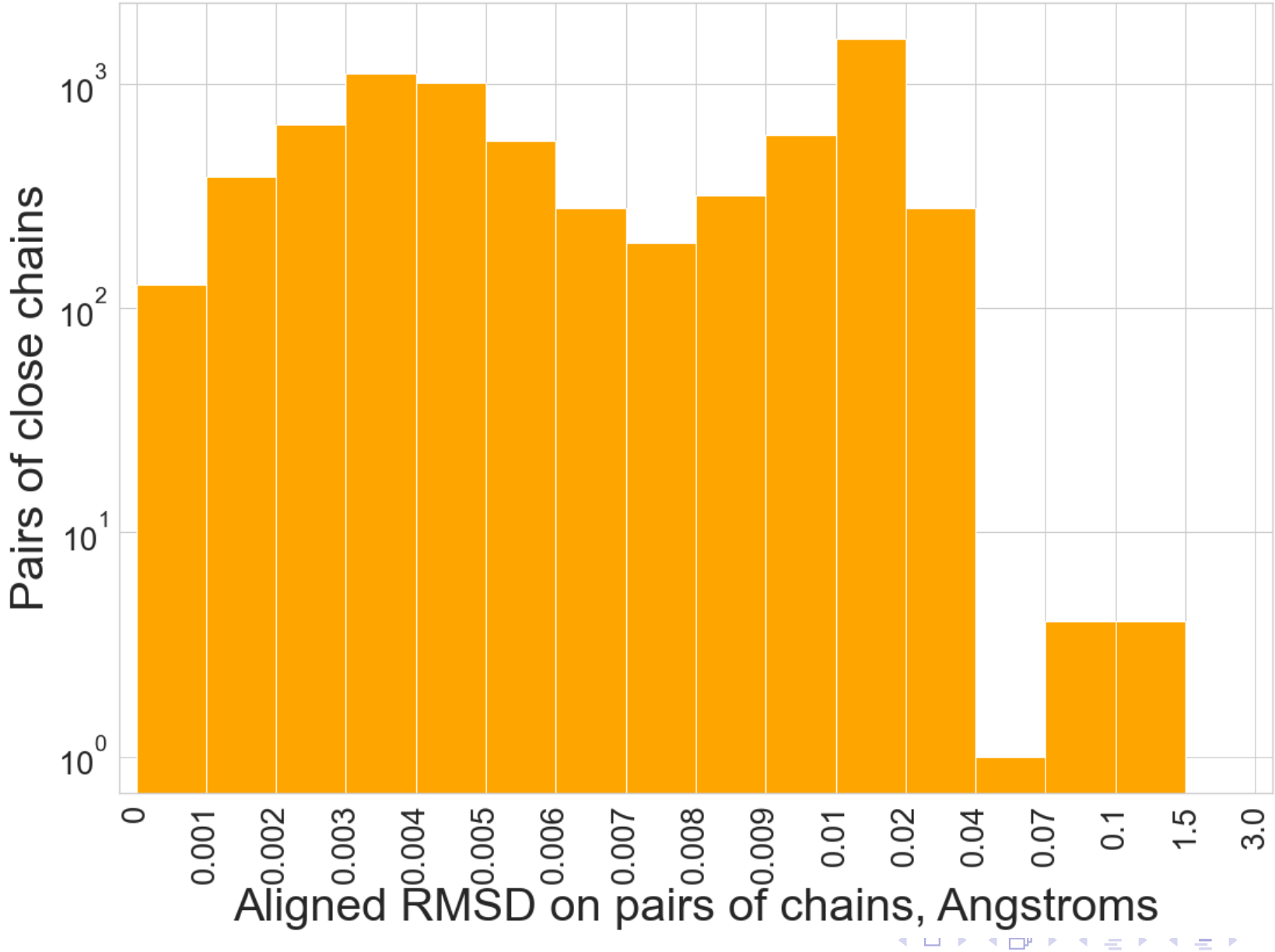
Full Table 1 in the paper: representative pairs from among multiple depositions with identical unit-cell parameters within each group of PDB entries 1aeu, 1aen and 1ac4 and PDB entries 1aeb, 1aed, 1aee, 1aef, 1aeg, 1aeh, 1aej, 1aek, 1aem, 1aeo and 1aeq. The supplementary file has 616 pairs.

The maximum deviation of coordinates is almost always 0.

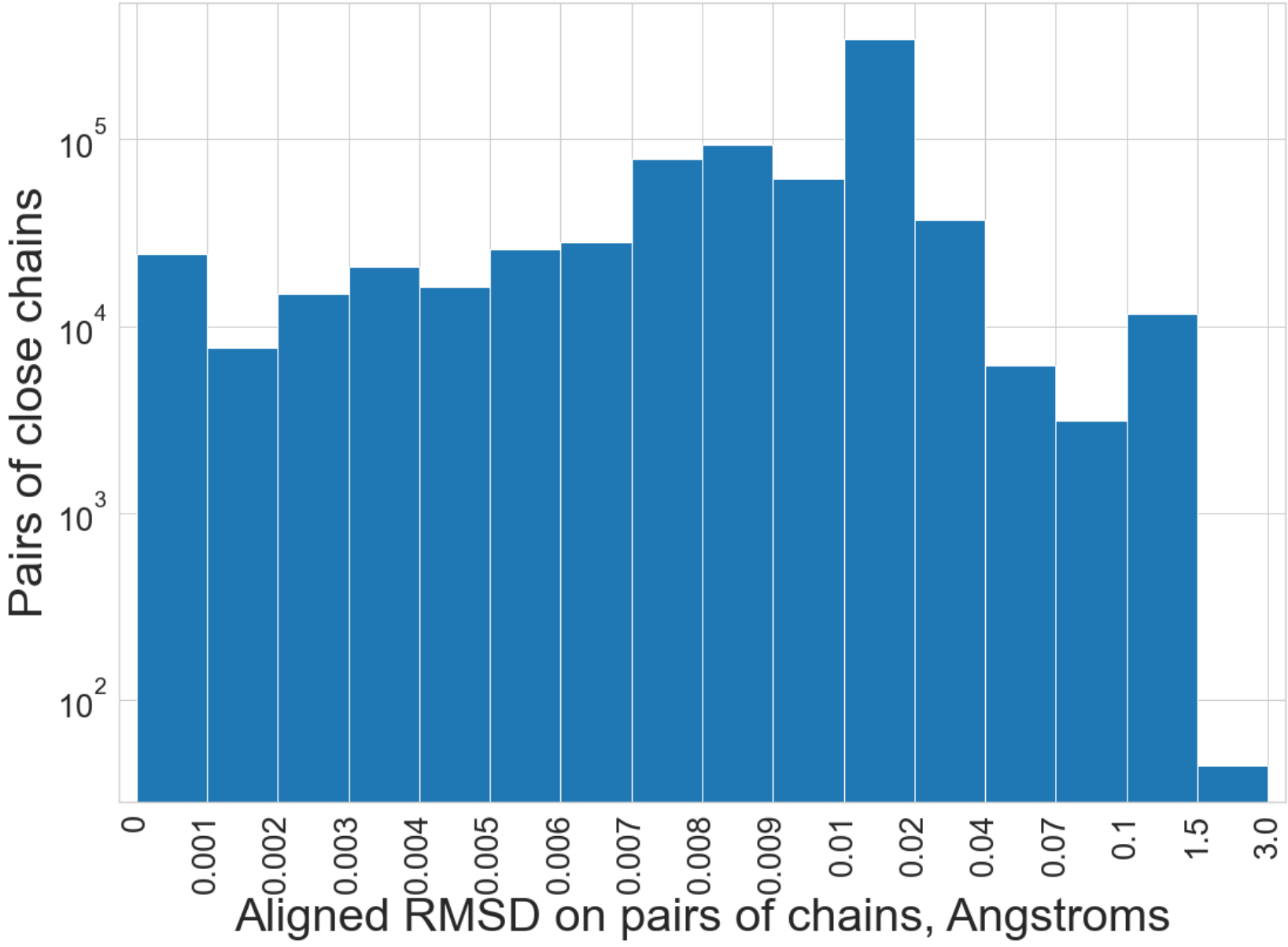
Identified cases & causes

- Subsequent redeposition in the PDB: 1a0t/1oh2 (still in the PDB).
- Deposited close together: 3lt8/3lt9.
- Redepositions should be different by description, but are identical in structure: 1hdu/1hee (potentially non-experimental modelling) or 2f5a/2pr4 (“refinement” by removal of some atoms).
- “Other peculiarities”: 1npw/1npa (same authors, 1997 and 2003) have different unit-cell parameters with exactly the same atomic coordinates and other factors.

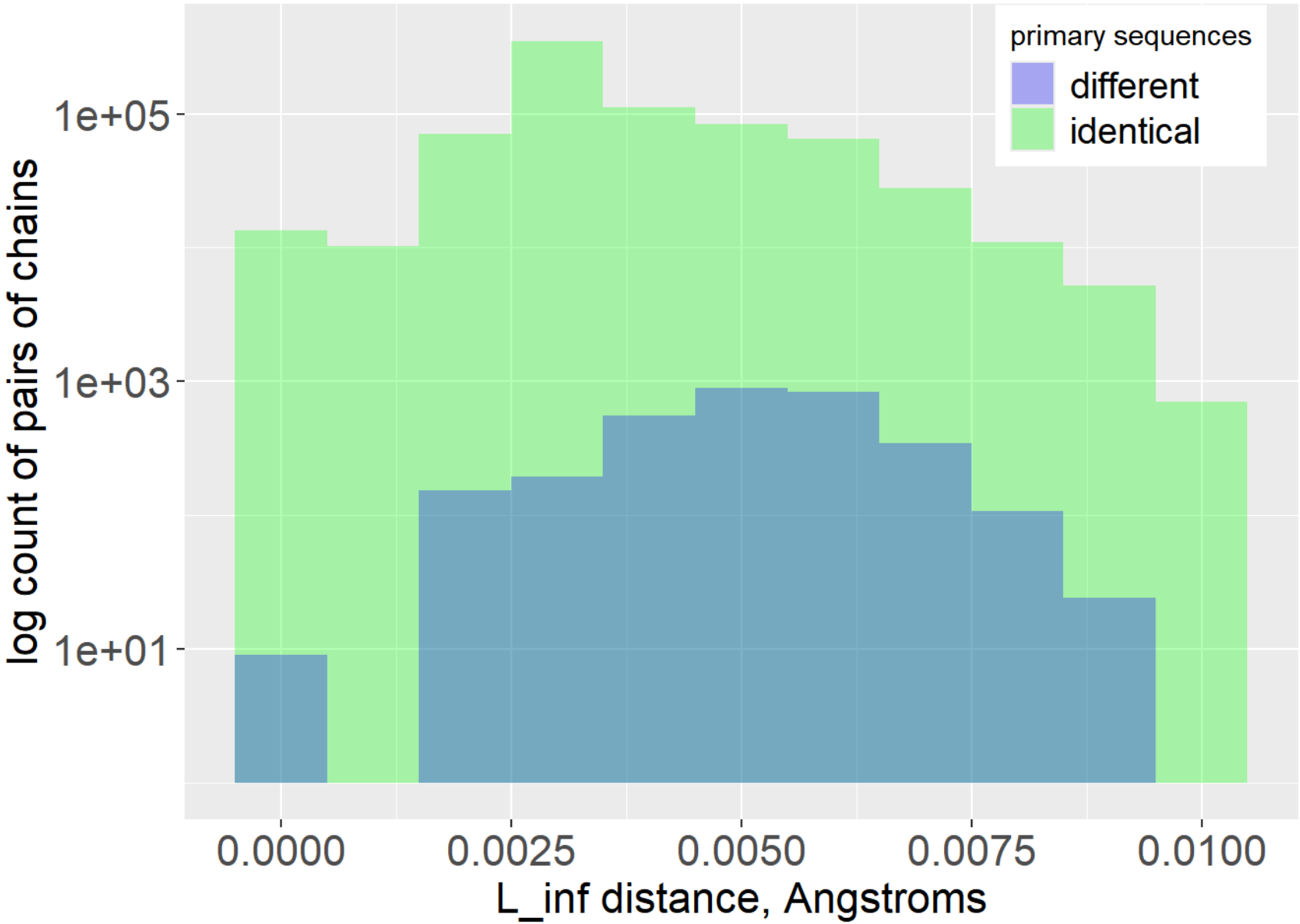
Different sequences, near-duplicates by RMSD



Identical sequences, near-duplicates by RMSD



Many more near-duplicates



Summary & references

D.Widdowson, V.Kurlin. Geographic-style maps with a local novelty distance help navigate in the materials space. Scientific Reports, 2025

O.Anosova et al. Recognition of near-duplicate periodic patterns by continuous metrics with approximation guarantees. Pattern Recognition, 2025

O.Anosova, V.Kurlin, M.Senechal. The importance of definitions in crystallography. IUCrJ, 2024

Anosova et al. A complete and bi-continuous invariant of protein backbones under rigid motion. MATCH, 2025

Wlodawer et al. A detailed discussion of duplicates. Acta Cryst D, 2025.