Electronic Supplementary Information

Simple coordination geometry descriptors allow to accurately predict metal binding sites in proteins

Giuseppe Sciortino,[§,#] Eugenio Garribba,[#] Jaime Rodríguez-Guerra Pedregal,[*,§] and Jean-Didier Maréchal.[*,§]

[§] Departament de Química, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallés, Barcelona, Spain

[#] Dipartimento di Chimica e Farmacia, Università di Sassari, Via Vienna 2, I-07100 Sassari, Italy

* Corresponding authors. E-mail: jaime.rodriguezguerra@uab.cat (J. R-G. P.), jeandidier.marechal@uab.cat (J-D.M.)

Table of Contents

Table S1. Specification of the validation dataset.

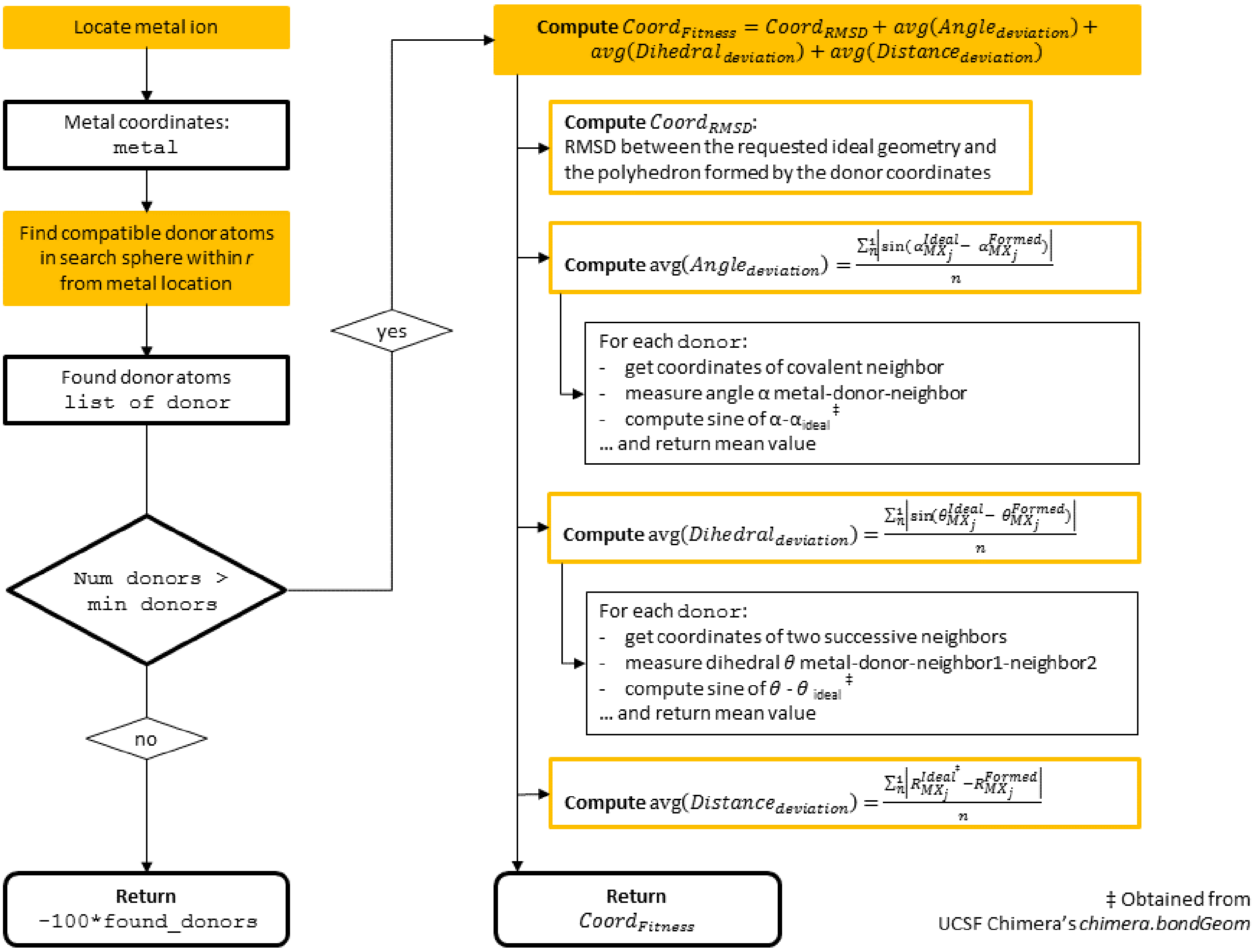
PDB	Res.(Å)	Geom.[d]	Donors[e]
Mg			
1AOX	2.1	*oct*	$(^{-}OOC_{Asp});2(^{-}O_{Ser})$
1AUS	2.2	*oct*	$(^{-}OOC_{Asp});(^{-}OOC_{Glu});(^{-}OOC_{FMT})$
1B8C	2.0	*oct*	$4(^{-}OOC_{Asp});(OC_{Amid})$
1BGM	2.5	*oct*	$3(OC_{Amid});(NCO_{Gln})$
1CHN	1.76	*oct*	$2(^{-}OOC_{Asp});(OC_{Amid})$
1DIE	2.5	*oct*	$(^{-}OOC_{Asp})^{\S};(^{-}OOC_{Glu});(N_{His})$
1DP0^{A}	1.7	*oct*	$(OCN_{Gln});(^{-}OOC_{Asp});3(OC_{Amid})$
1DP0^{B}	1.7	*oct*	$2(^{-}OOC_{Glu});(N_{His})$
1EC7	1.9	*oct*	$(OCN_{Asn});(^{-}OOC_{Asp});(^{-}OOC_{Glu})$
1H1L	1.9	*oct*	$2(^{-}OOC_{Asp});(^{-}OOC_{Glu});(OC_{Amid})$
1HUJ	2.1	*oct*	$3(^{-}OOC_{Asp})$
1F49^{A}	2.5		
1F49^{B}	2.5		
Fe			
1AHJ	2.65	*spy*	$3(^{-}S_{Cys});2(^{-}N_{Amid})$
1DFX	1.9	*spy*	$4(N_{His});(^{-}S_{Cys})$
1VRB	2.60	*spy*	$(^{-}OOC_{Asp});2(N_{His})$
1YGE	1.4	*spy*	$(^{-}OOC_{Ile});2(N_{His})$
5AFS	2.22	*spy*	$(^{-}OOC_{Asp})^{\S};(^{-}OOC_{Glu});2(N_{His})$
1AUI	2.1	*oct*	$2(^{-}OOC_{Asp});(N_{His})$
1B0L^{A}	2.2	*oct*	$2(^{-}O_{Tyr});(^{-}OOC_{Asp});(N_{His})$
1B0L^{B}	2.2	*oct*	$2(^{-}O_{Tyr});(^{-}OOC_{Asp});(N_{His})$
1B71	1.9	*oct*	$3(^{-}OOC_{Glu});(N_{His})$
1BSM	1.35	*oct*	$(^{-}OOC_{Asp});3(N_{His})$
3V83^{A}	2.1	*oct*	$2(^{-}O_{Tyr});(^{-}OOC_{Asp});(N_{His})$
3V83^{B}	2.1	*oct*	$2(^{-}O_{Tyr});(^{-}OOC_{Asp});(N_{His})$
1D9Y	2.2	*oct*	$2(^{-}O_{Tyr});(^{-}OOC_{Asp});(N_{His})$
1DMW	2.0	*oct*	$(^{-}OOC_{Glu});2(N_{His})$
1D06^{B}	1.4	*oct*	$(^{-}S_{Cys});4(N_{His})$
1DQI	1.7	*oct*	$(^{-}OOC_{Asp});(^{-}S_{Cys});4(N_{His})$
1D3K	1.8	*oct*	
Ni			
1FRV	2.85	*spy*	$4(^{-}S_{Cys})$
1OID	2.1	*spy*	$(^{-}OOC_{Asp});(OCN_{Asn});(N_{His})$
1T6U	1.30	*spy*	$2(^{-}S_{Cys});(N_{His});(^{-}N_{Amid});(N_{Amid})$
1XCV	2.1	*spy*	$(^{-}OOC_{Asp});(^{-}OOC_{Glu});(N_{His});(S_{Met});(OC_{Amid})$
2C21	2.0	*spy*	$2(^{-}OOC_{Glu});2(N_{His})$
2QJV	1.90	*spy*	$(^{-}OOC_{Glu});3(N_{His})$
2QQH	2.5	*spy*	$(^{-}OOC_{Glu});2(N_{His})$
3C7J	2.1	*spy*	$(OCN_{Asn});3(N_{His})$
3CGM	2.41	*spy*	$5(N_{His})$
3DKQ	2.26	*spy*	$(^{-}OOC_{Asp});3(N_{His})$
3NY0	3.09	*spy*	$5(N_{His})$
4I4A	1.35	*spy*	$(^{-}OOC_{Glu});3(N_{His})$
1F9Z	1.5	*oct*	$2(^{-}OOC_{Glu});2(N_{His})$
1IAE	1.83	*oct*	$(^{-}O_{Tyr});3(N_{His})$
1P1M	1.5	*oct*	$(^{-}OOC_{Asp});3(N_{His})$
1QXJ	1.8	*oct*	$(^{-}OOC_{Glu});3(N_{His})$
1J5Y	2.3	*oct*	$(^{-}OOC_{Glu})^{\S};3(N_{His})$
1J6P	1.9	*oct*	$(^{-}OOC_{Asp});3(N_{His})$
1B9M	1.75	*oct*	$(^{-}OOC_{Asp})^{\S};(^{-}OOC_{Asp});2(N_{His})$
1DDN	3.0	*oct*	$(^{-}OOC_{Asp});(^{-}OOC_{Glu});(N_{His});(S_{Met});(OC_{Amid})$
Cu			
1AG0	2.4	*spy*	$(^{-}OOC_{Asp})^{\S};2(N_{His});(OC_{Amid})$
1GOG	1.7	*spy*	$2(^{-}O_{Tyr});2(N_{His})$
3AWS	1.24	*spy*	$(^{-}OOC_{Asp});(N_{His});(OC_{Amid})$
4EIS	1.37	*spy*	$(^{-}O_{Tyr});2(N_{His});(N_{Amid})$

PDB	Res.(Å)	Geom.[f]	Donors‡
Ca			
1AVW	1.75	*oct*	$3(^{-}OOC_{Asp});2(OC_{Amid})$
Mn			
1AX1	1.95	*spy*	$2(^{-}OOC_{Asp});(^{-}OOC_{Glu});\ (N_{His})$
1D5N	1.55	*oct*	$3(^{-}OOC_{Asp});(N_{His})$
1DO8	2.2	*oct*	$2(^{-}OOC_{Asp});(^{-}OOC_{Glu})$
1DCK	2.0	*oct*	$2(^{-}OOC_{Asp});(OC_{Amid})$
1DID^{A}	2.5	*oct*	$(^{-}OOC_{Asp})^{\S};(^{-}OOC_{Asp});(^{-}OOC_{Glu});(N_{His})$
1DID^{B}	2.5	*oct*	$2(^{-}OOC_{Asp});2(^{-}OOC_{Glu})$
1EJJ	1.9	*oct*	$(^{-}OOC_{Asp})^{\S};(^{-}OOC_{Asp})\ (^{-}O_{Ser});(N_{His})$
1F1R	1.8	*oct*	$(^{-}OOC_{Glu});2(N_{His})$
1F9C	2.5	*oct*	$2(^{-}OOC_{Asp});(^{-}OOC_{Glu})$
1DO8	2.2		
1FFS	2.4	*oct*	$2(^{-}OOC_{Asp});(OC_{Amid})$
1FI2	1.6	*oct*	$(^{-}OOC_{Glu});3(N_{His})$
1FQW	1.37	*oct*	$2(^{-}OOC_{Asp});(OC_{Amid})$
1FX5	2.2	*oct*	$2(^{-}OOC_{Asp});(^{-}OOC_{Glu});(N_{His})$
1I9A	2.5	*oct*	$(^{-}OOC_{Asp})^{\S};(^{-}OOC_{Asp});3(N_{His})$
Co			
1C0W^{A}	3.2	*spy*	$(^{-}S_{Cys});(S_{Met});(N_{His});(OC_{Amid})$
1C0W^{B}	3.2	*spy*	$2(^{-}OOC_{Glu});(OCN_{Gln});2(N_{His})$
1IAB	1.79	*spy*	$(^{-}O_{Tyr});3(N_{His})$
1UGS	2.0	*spy*	$3(^{-}S_{Cys});2(^{-}N_{Amid})$
2DXB	2.25	*spy*	$(^{-}S_{Cys});(S_{Cso});(S_{Csd});2(^{-}N_{Amid})$
4ERA	2.4	*spy*	$(^{-}OOC_{Asp});(^{-}O_{Tyr});3(N_{His})$
1CAH	1.88	*oct*	$3(N_{His})$
1DZI	2.1	*oct*	$2(^{-}O_{Ser});(^{-}OOC_{Glu});(OC_{Amid})$
1FA6	1.9	*oct*	$2(^{-}OOC_{Glu});2(N_{His})$
1FOF	2.0	*oct*	$2(^{-}OOC_{Glu});(N_{His})$
1FX7	2.0	*oct*	$2(^{-}OOC_{Glu});2(N_{His});(OCN_{Gln})$
1LNA	1.9	*oct*	$2(^{-}OOC_{Asp});(OC_{Amid})$
1QT1	1.85	*oct*	$2(^{-}OOC_{Asp});(^{-}OOC_{Glu});(N_{His})$
1R8K	2.1	*oct*	$3(N_{His})$
1RR2	2.0	*oct*	$(^{-}OOC_{Kcx})^{\S};(^{-}OOC_{Asp});2(N_{His})$
1RV8	2.3	*oct*	$(^{-}OOC_{Glu})^{\S};3(N_{His})$
1V29	2.6	*oct*	$3(^{-}S_{Cys});2(^{-}N_{Amid})$
Zn			
1XJS	NMR	*spy*	$3(^{-}S_{Cys});(^{-}OOC_{Asp})^{\S}$
1Y7P	1.9	*spy*	$5(OC_{Amid})$
2DI3	1.8	*spy*	$(^{-}OOC_{Asp});3(N_{His})$
2JNE^{A}	NMR	*spy*	$(^{-}OOC_{Asp});3(^{-}S_{Cys});(N_{His})$
2KO0	NMR	*spy*	$(^{-}OOC_{Asp});3(^{-}S_{Cys});(N_{His})$
2LQ6^{A}	NMR	*spy*	$(^{-}O_{Thr});3(^{-}S_{Cys});(N_{His})$
2LQ6^{B}	NMR	*spy*	$3(^{-}S_{Cys});(N_{His})$
2LZE^{A}	NMR	*spy*	$(^{-}OOC_{Asp});(^{-}OOC_{Glu});2(^{-}S_{Cys});(N_{His})$
2OBA	2.33	*spy*	$(^{-}OOC_{Glu});3(N_{His})$
2Y3G^{A}	1.91	*spy*	$(^{-}OOC_{Asp})^{\S};3(N_{His})$
2Y3G^{B}	1.91	*spy*	$(^{-}OOC_{Glu});2(N_{His})$
2Y3G^{C}	1.91	*spy*	$(^{-}OOC_{Glu});3(N_{His})$
2Y12	3.1	*spy*	$4(^{-}S_{Cys});(OC_{Amid})$
3IJ6	2.0	*spy*	$(^{-}OOC_{Asp});3(N_{His})$
1LR5	1.9	*oct*	$(^{-}OOC_{Glu});3(N_{His})$
2OI0	2.0	*oct*	$3(N_{His})$
1GE6	2.2	*oct*	$(^{-}OOC_{Asp})^{\S};2(N_{His})$
1IM5	1.65	*oct*	$(^{-}OOC_{Asp});2(N_{His})$
1OI0		*oct*	$(^{-}OOC_{Asp})^{\S};2(N_{His})$

[f] *spy* = square pyramidal; *oct* = octahedral. § Symmetric bidentate coordination. [e] The donor atom is write in any case in the first position. [a,b] the superscript a or b, following the formalism of the webserver *MetalPDB,* indicate the specific region in case of multiple binding site structures.

Coordination scoring algorithm

Figure S1. Coordination objective flowchart.


Locate metal ion
Metal coordinates: metal
Find compatible donor atoms in search sphere within *r* from metal location
Found donor atoms list of donor
Num donors > min donors
yes
no
Return -100*found_donors
Compute $Coord_{Fitness} = Coord_{RMSD} + avg(Angle_{deviation}) + avg(Dihedral_{deviation}) + avg(Distance_{deviation})$
Compute $Coord_{RMSD}$: RMSD between the requested ideal geometry and the polyhedron formed by the donor coordinates
Compute avg$(Angle_{deviation}) = \frac{\sum_n^1 \left|\sin(\alpha_{MX_j}^{Ideal} - \alpha_{MX_j}^{Formed})\right|}{n}$
For each donor:
- get coordinates of covalent neighbor
- measure angle α metal-donor-neighbor
- compute sine of α-α_{ideal} ‡
... and return mean value
Compute avg$(Dihedral_{deviation}) = \frac{\sum_n^1 \left|\sin(\theta_{MX_j}^{Ideal} - \theta_{MX_j}^{Formed})\right|}{n}$
For each donor:
- get coordinates of two successive neighbors
- measure dihedral θ metal-donor-neighbor1-neighbor2
- compute sine of θ - θ_{ideal} ‡
... and return mean value
Compute avg$(Distance_{deviation}) = \frac{\sum_n^1 \left|R_{MX_j}^{Ideal‡} - R_{MX_j}^{Formed}\right|}{n}$
Return $Coord_{Fitness}$

‡ Obtained from UCSF Chimera's *chimera.bondGeom*

1. Locate the metal center (*probe*) and search for compatible coordinating atoms (*donor*) which match the following criteria:

- The atom type must be contained in the user-supplied atom type list.
- The position must be within the user-specified search radius.

2. If the number of donors meets the minimum number of ligand atoms defined in the objective configuration:

A) Compute the RMSD (*Coord_RMSD*) value as reported by a rigid Coherent Point Drift registration between the ideal convex polyhedron and the best *formed polyhedron* built using *donors* as possible vertices. The ideal polyhedral positions are extracted from a database of normalized vector sets centered at the cartesian origin.[1]

B) Additionally, for every atom that matches the criteria:

- Test if the *donor* can act as a bidentate ligand (i.e. Asp or Glu) or not. If that's the case, use a virtual position (as defined by the mean coordinates of both donor atoms of the bidentate) to be reported as *donor*.
- Get the coordinates for *donor*, *1st_neighbor*, and *2nd_neighbor* (if available).
- Compute the *formed distance* between *probe* and *donor*.
- Compute the *formed angle* between *probe, donor,* and *1st_neighbor*.
- Compute the *formed dihedral* between *probe*, *donor*, *1st_neighbor* and *2nd_neighbor* (if available).
- An *Distance_Dev* is obtained by computing the ideal distance deviation as the absolute difference between the ideal element-element distance reported by the *chimera.bondGeom* routines and the *formed distance* measured.
- A *Coord_Directionality* value is obtained as a sum of averages values of:
    - Absolute difference of sines of the ideal angle and the *Formed Angle* $\left(\alpha_{MX_j}^{Ideal}\right)$.
    - Absolute sine of the ideal dihedrals and *Formed Dihedral* $\left(\theta_{MX_j}^{Ideal}\right)$.

    The ideal angles are obtained from UCSF Chimera routines ('chimera.bondGeom' module) that consider the van der Waals radius of each element and its hybridation

C) The final *Coord_Fitness* is obtained as a linear sum of the latter two values as reported in eq. S1

$$Coord_{Fitness} = \left(\frac{\sum_n^1 \left|\sin(\alpha_{MX_j}^{Ideal} - \alpha_{MX_j}^{Formed})\right|}{n}\right) + \left(\frac{\sum_n^1 \left|\sin(\theta_{MX_j}^{Ideal} - \theta_{MX_j}^{Formed})\right|}{n}\right) + \left(\frac{\sum_n^1 \left|R_{MX_j}^{Ideal} - R_{MX_j}^{Formed}\right|}{n}\right) + Coord\_RMSD \quad \textbf{eq. S1}$$

3. Otherwise, the score is proportional to the number of missing ligand atoms. Since the score is to be minimized by the algorithm, this acts as a dynamic penalty that tells the genetic algorithm how far this candidate is from obtaining a valid geometry:

$$Coord\_Fitness = N_{missing} * 100 \quad \textbf{eq. S2}$$

Discarded strategies for increasing performance

During the development of this update, we strived to make the most of this algorithm and devised complementary strategies that could bring the reported results to an even higher standard. Here we report two additional approaches that while superior to the original implementation featured in GaudiMM,[2] could not compete with the strategy presented in the main manuscript: a center of mass correction, and a local optimization step.

The original implementation

The original implementation present in GaudiMM did not feature distance deviation correction but, when submitted to this protocol, already generated exciting results. The mean of the *Coord_Fitness* was close to 3.0 units (smaller due to the absent sum term, but not necessarily more accurate), displaying a general agreement between the experimental and the simulated geometry with the developed protocol. Concerning the RMSD value, the mean is very small (0.564 Å ± 0.541). It should be highlighted that, for all the proposed solutions, the reported RMSD falls under the X-ray spectra resolution. A detailed analysis of the GaudiMM solutions for the entire dataset is summarized in **Table S2**, which shows that the crystallographic binding site is reproduced with a success rate of 92.4% with a RMSD ≤ 1.0.

Table S2. Summary of RMSD deviations between the experimental and the predicted binding sites (first column) and *Coord_Fitness* distribution of the dataset.

RMSD[a,b]	**Total**	***Coord_Fitness***[d] ***≤ 3.5***	***3.5 < Coord_Fitness***[d] ***≤ 5***
0.5 ≤	70	45	22
> 0.5 < 1.0	27	21	3
> 1.0 ≤ 1.5	4	3	-
bad	4	-	-

[a] Value reported in Å. [b] RMSD computed via UCSF Chimera v1.11. [d] Value reported by GaudiMM using the eq. S3.

$$Coord_{Fitness} = \left(\frac{\sum_{n}^{1} \left| \sin(\alpha_{MX_j}^{Ideal} - \alpha_{MX_j}^{Formed}) \right|}{n} \right) + \left(\frac{\sum_{n}^{1} \left| \sin(\theta_{MX_j}^{Ideal} - \theta_{MX_j}^{Formed}) \right|}{n} \right) + Coord\_RMSD \qquad \textbf{eq. S3}.$$

Furthermore, a ranking analysis shows that the solutions with lowest RMSD, in the 74.8% of the simulations, are situated in the first cluster in terms of *Coord_Fitness.* In all cases, they are placed in the most populated one. The complete set of the data is reported in **Table S1**.

Center of mass correction

A first attempt to increase the accuracy of the original *coordination* objective was the implementation of a *center of mass correction*. This objective considers an alternative sum term in the score that, instead of computing the ideal distances, evaluates the absolute distance between the metal *probe* and the center of mass of the coordinated donors as reported in eq. 2, supported by the intuition that fully-coordinated metal centers are approximately positioned in the center of mass of the coordinating atoms. An analysis of the result of this second benchmark is reported in **Table S3** and represented in **Figure S2b**.

$$Fitness = Coord_{Fitness} = \left(\frac{\sum_{n}^{1} \left| \sin(\alpha_{MX_j}^{Ideal} - \alpha_{MX_j}^{Formed}) \right|}{n} \right) + \left(\frac{\sum_{n}^{1} \left| \sin(\theta_{MX_j}^{Ideal} - \theta_{MX_j}^{Formed}) \right|}{n} \right)$$

$$+ \left| r_{metal} - \frac{1}{M} \sum_{i}^{1} m_i r_i \right| + Coord\_RMSD \qquad \textbf{eq. S4}$$

where M is the sum of the donors’ mass m_i and r the atom coordinate

Table S3. Summary of RMSD deviations between the experimental and the predicted binding sites (first column) and *Coord_Fitness* distribution of the dataset using the Center of Mass correction.

RMSD[a,b]	**Total**	***Coord_Fitness***[d] **≤ *3.5***	***3.5 < Coord_Fitness***[d] ***≤ 5***
0.5 ≤	70	45	22
> 0.5 < 1.0	27	21	3
> 1.0 ≤ 1.5	4	3	3
bad	4	-	-

[a] Value reported in Å. [b] RMSD computed via UCSF Chimera. [d] Value extrapolated by GaudiMM using eq. **S4**.

The mean of *Coord_Fitness* is close to 4.0 units. The mean RMSD is lower than the one obtained with the original method (0.507 Å) with an associated standard deviation value of 0.235, which highlights an improvement of the error distribution. The crystallographic binding site is reproduced with a success rate of 93.3% with RMSD ≤ 1.0 (Table 3). A ranking analysis shows that the solutions with lowest RMSD are situated in the first cluster (in terms of *Coord_Fitness* score) increases up to 85.3% proportion; all are placed in the most populated one. However, it must be highlighted that this correction would only improve the accuracy if the metal ion under evaluation has no coordination vacancies; otherwise the center of mass would be skewed towards the present coordinating atoms. The complete set of the data is reported in Table S2 of the Supporting Information.

A comparison between the results of the three benchmarks is summarized in **Table S4** and shown in **Figure S2**, highlighting the improvement in terms of a) success rate from 92.4% (original method), to 93.3% (center of mass correction) and 100.0% (ideal distances deviation); b) RMSD mean from 0.564, to 0.507 and 0.519 Å, respectively, c) standard deviation from 0.541, to 0.235 and 0.175, respectively, and d) percentage of solutions in the first position of the scoring ranking: 74.8%, 85.3%, and 86.7% respectively.

Table S4. Percentage of solutions in the first position of the scoring ranking, RMSD mean, *Coordination_Fitness* mean and associated standard deviation obtained in the three benchmarks.

	Success rate	**RMSD mean**	**Std. dev**	**1st rank**	**Coord_Fitness mean**
Original Method	92.4%	0.564	0.541	74.8	3.019 [d]
Center of Mass correction	93.3%	0.507	0.235	85.3	3.853 [e]
Ideal Distances (main)	100.0%	0.519	0.175	86.7	4.003 [e]

[a] Value reported in Å. [b] RMSD computed with UCSF Chimera. [c] Standard deviation (SD). [d] Value reported by GaudiMM using **eq. S3**. [e] Coord_Fitness increment corresponding to the addition of an additional term in the original *Coordination* objective (equations **S4** and **S1**, respectively).

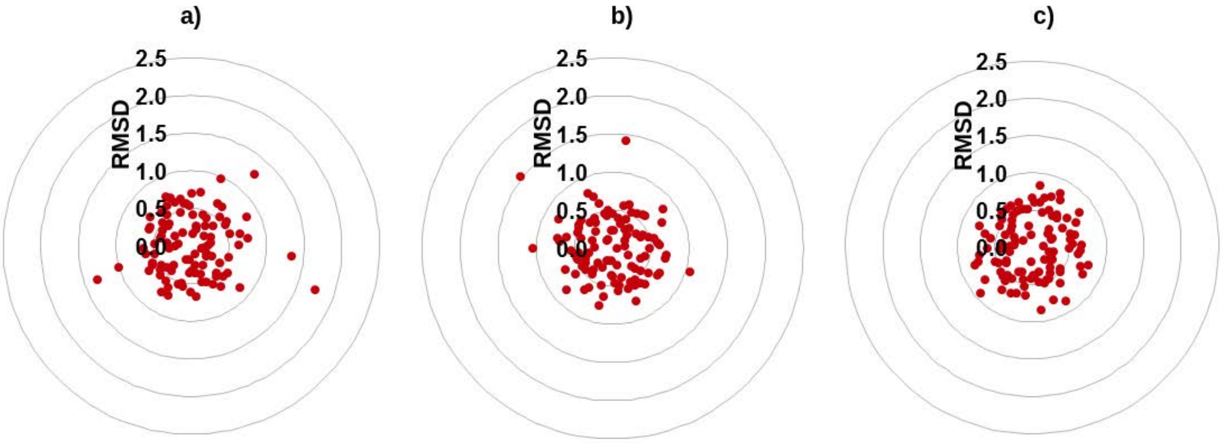


Figure S2. Graphical representation of the RMSD distribution of the GaudiMM benchmarks: a) standard method, b) center of mass correction, and c) ideal bonds deviations correction. Method c) is recommended as it shows improved RMSD mean, standard deviation.

Local search

The algorithm implemented in GaudiMM (NSGA-II[3]) performs a *global search* procedure that can quickly localize high *fitness* regions of vast search spaces, but it may result less suited for fine-tuned predictions. Fortunately, GA performance can be generally improved by introducing several local search evaluations during the calculation.[4] Literature reports that some *hybrid GAs* have been specifically designed to provide an alternation between global and local searches in a specific portion of the population.[5-7]

In this section, we perform a test on the accuracy improvement by adding a local search procedure in GaudiMM. We extracted six predicted solutions of the original implementation benchmark in which the RMSD ≥ 1.0 Å (PDB code: 1F49, 1BGM, 1D80, 1DZI, 2LQ6^a and 2LQ6^b) and redocked them along 20 generations with a population of 100 individuals using a local evaluation sphere of 5 Å centered on the metal site proposed by the global search. The results, summarized in **Table S5**, show that a local search significantly improves the accuracy of the docking prediction showing a mean RMSD decrease of about 66.8%.

Table S5. Summary of RMSD and *Coord_Fitness* values obtained after implementing a local search procedure in GaudiMM.

Structure (PDB)	**RMSD[a]**		
	Global	Local	Improvement (%)
1F49^a [b]	1.743	0.813	53.54
1BGM	1.266	0.894	29.4
1D80	5.226	0.533	89.8
1DZI	1.344	0.234	82.6
2LQ6^a [b]	1.324	0.473	64.3
2LQ6^b [b]	1.007	0.187	81.4
		mean	**66.8**

[a] RMSD value computed via UCSF Chimera, reported in Å. [b] The superscript ^{a} or ^{b}, following the formalism of the webserver MetalPDB, indicate the specific region in case of multiple binding site structures.

Thus, the local search implementation allowed to rescue six structures by increasing the success rate up to the 98.1% with a RMSD ≤ 1.0 Å. However, still worse than the ideal distances deviation strategy featured in the main text. In **Figure S3** a comparison between six simulated structures after and before the local search implementation is reported.

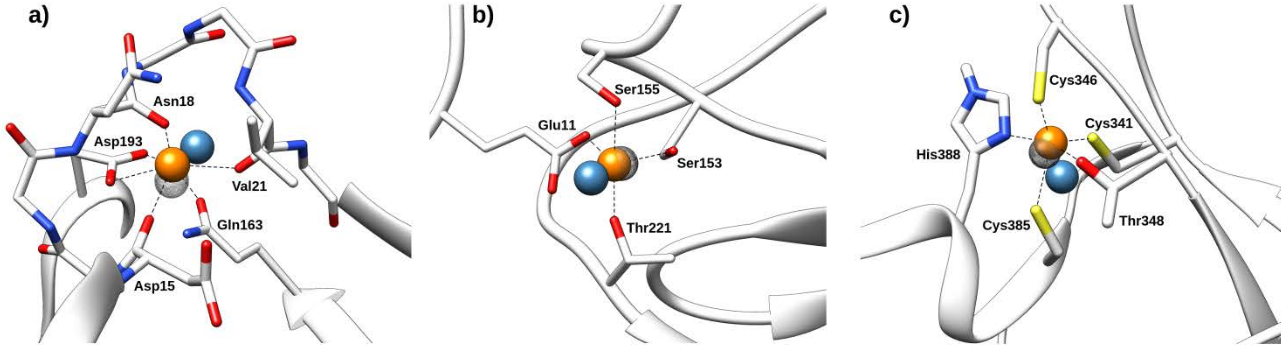


Figure S3. Comparison between the three simulated structures a) 1F49, b) 1DZI, d) 2LQ6[a] after (in orange) and before (in blue) the local search implementation. The original XRD structure is also shown (in dots surface).

Benchmark input details

The benchmark was performed with GaudiMM v0.0.3+7.g77615c9.

The *precision* parameter was set to 5 decimal places, a value that guarantees an adequate consideration of the *"diversity"* of the solutions proposed by the *search* gene.

The μ and λ genetic algorithm parameters were set to 1.0 and 4.0, respectively, to reach a final number of individuals equal to the initial population size while temporally expanding the population in the variation stage. The probability associated to mutation *mut* and crossover *cx* were both set to 0.50. The results were collected after running three calculations for each dataset entry.

The full input file (*benchmark.yaml*) can be found attached as part of the accompanying ZIP file.

Running the benchmark

A Python script called *‘benchmark_all.py’* is attached in the accompanying ZIP file. It expects a directory containing the metal-containing protein PDB files corresponding to the dataset entries, and a YAML input file for GaudiMM (*benchmark.yaml* is attached as an example). This script requires GaudiMM to be installed, along with its dependencies. Documentation on how to install it is available at http://gaudi.readthedocs.io/en/latest/. Once installed, run ‘*pychimera benchmark_all.py -h*` for help.

Multisite.py script

The Python script to precompute potential binding sites in biological scaffold is available at https://github.com/insilichem/scripts/blob/master/multisite.py. Documentation is included within the code itself. A downloaded copy is also attached in the accompanying ZIP file.

References

1. E. W. Weisstein, Mathworld--a Wolfram Web Resource, http://mathworld.wolfram.com/ConvexPolyhedron.html, 2018.
2. J. Rodríguez-Guerra Pedregal, G. Sciortino, J. Guasp, M. Municoy and J.-D. Maréchal, Gaudimm: A Modular Multi-Objective Platform for Molecular Modeling, *J. Comput. Chem.*, DOI: 10.1002/jcc.24847.
3. K. Deb, A. Pratap, S. Agarwal and T. Meyarivan, A Fast and Elitist Multiobjective Genetic Algorithm: Nsga-Ii, *IEEE Transactions on Evolutionary Computation*, 2002, **6**, 182-197.
4. C. García-Martínez and M. Lozano, in *Advances in Metaheuristics for Hard Optimization*, eds. P. Siarry and Z. Michalewicz, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, DOI: 10.1007/978-3-540-72960-0_10, pp. 199-221.
5. G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew and A. J. Olson, Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function, *J. Comput. Chem.*, 1998, **19**, 1639-1662.
6. J. Fuhrmann, A. Rurainski, H. P. Lenhof and D. Neumann, A New Lamarckian Genetic Algorithm for Flexible Ligand-Receptor Docking, *J. Comput. Chem.*, 2010, **31**, 1911-1918.
7. G. Boxin, Z. Changsheng and N. Jiaxu, Edga: A Population Evolution Direction-Guided Genetic Algorithm for Protein–Ligand Docking, *J. Comput. Biol.*, 2016, **23**, 585-596.