Electronic Supplementary Material (ESI) for Digital Discovery.

MACROCONF – DATASET & WORKFLOWS TO ASSESS CYCLIC PEPTIDE SOLUTION STRUCTURES

Daniel Crusius[1], Jason R. Schnell[1], Flaviu Cipcigan[2] and Philip C. Biggin[1*]

[1]*Department of Biochemistry, University of Oxford, South Parks Road, Oxford, OX1 3QU, UK*

[2]IBM Research Europe, The Hartree Centre STFC Laboratory, Sci-Tech Daresbury, Warrington WA4 4AD, U.K.;

*To whom correspondence should be addressed:

Philip.biggin@bioch.ox.ac.uk

ORCIDs: Daniel Crusius: 0000-0002-1305-9517, Jason Schnell: 0000-0002-8204-7608, Flaviu Cipcigan: 0000-0002-5015-1443, Philip Biggin: 0000-0001-5100-8836

Keywords: molecular dynamics, conformer generation, NMR, cheminformatics.

TABLE OF CONTENTS

Table of Contents *2*

The MacroConf dataset *3*

SI Text S1: Overview of the MacroConf dataset **3**

SI Text S2: Specification of the MacroConf dataset **4**

Molecular Dynamics Simulations *5*

SI Text S3: MacroConf Workflow: Setup of MD simulations **5**

SI Text S4: Clustering of MD Simulations **5**

SI Text S5: Convergence of MD Simulations **7**

SI Text S6: Computing NOE distance constraints **12**

SI Text S7: Outliers 24, 49 ..... **13**

Method Comparison *15*

SI Text S8: Significance tests for Method comparisons **15**

SI Text S9: Cheminformatics Conformer Generators: Bundle size effects & Distributions of NOE metrics **21**

SI Text S10: Effect of different Re-Weighting Methods of GaMD **22**

SI Text S11: Effect of a stepwise RMSD metric **23**

SI Text S12: NOE coverage for varying peptide sequence lengths **27**

SI Text S13: Comparison of Solvation properties by different methods **28**

The MacroConf Workflow *32*

SI Text S14: External Dependencies and installation **32**

Data Availability *33*

References *33*

THE MACROCONF DATASET

SI TEXT S1: OVERVIEW OF THE MACROCONF DATASET

The MacroConf dataset contains 68 compounds shown in Figure S1.

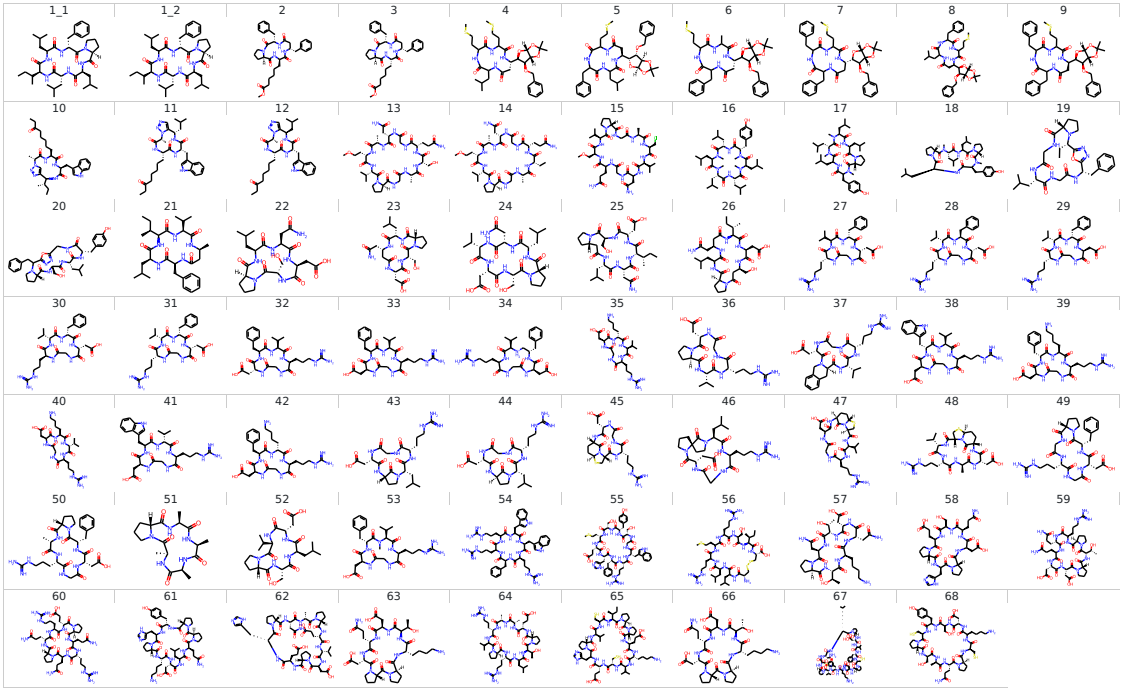

Figure S1. Overview of the MacroConf dataset showing the chemical structures of the 68 cyclic peptides and macrocycles that make up the dataset.

SI TEXT S2: SPECIFICATION OF THE MACROCONF DATASET

Every compound has the following properties, recorded in a *.json* file:

[ID].json

```
ID: unique compound identifier (int)
Compound name: name a compound is known as, (string, optional)
Compound name in original publication: Name of the compound in the original
publication (string, optional)
Publication: DOI of the publication the NMR experiment was published with (string)
Is natural cyclic peptide: Is the CP exclusively made up of natural L amino acids
only? (1: only natural L amino acids, 0: not only natural L amino acids) (0,1)
Is unnatural cyclic peptide: Is the CP exclusively made up of L/D amino acids only?
(1: only natural L/D amino acids, 0: not only natural L/D amino acids -> compound
is a macrocycle) (0,1)
Sequence: Sequence of the compound, as reported in the publication (string)
Sequence-1: 1 letter amino acid sequence (string)
Backbone size: No. of amino acids forming the CP (valid only if compound is an
amino acid like CP) (int)
Bonds: Specifier of ring bonds. [INSERT CODE HERE!]
Multi: Denotes whether there are distinct sets of experimental NOEs for different
conformers (e.g., cis/trans) [INSERT DETAILS HERE!]
Type of NOE available: Description of the type of NOEs available (e.g., upper
bounds only) (string, optional)
Solvent: Solvent that was used for the NMR experiment (string)
Type of NOE experiment: Type of reported NOE experiment (e.g., ROESY, NOESY, …)
(string)
NMR experiment details: Experimental parameters (mixing times, etc.) (string)
NOE computation details: Details of how the NOE values were derived (string)
Other NMR experiments conducted: Any other NMR experiments reported for the
compound (string)
SMILES: SMILES string of the compound (SMILES)
NOE quality: Quality assessment. 1:low quality (experiment was conducted only at 1
mixing time), 5: high quality (experiment was conducted with multiple mixing times)
(1,5)
No. of NOEs: No of NOEs available for the compound. Some NOEs might have been
removed in the compilation process, e.g., because there were errors or
inconsistencies (string)
NOEs removed for quality reasons: Reasons why some published NOEs were removed
(string)
raw_NOE columns: Contains information on what the column in the raw NOE file
contains
raw_NOE conversion factor: Conversion factor to convert the literature NOE values
to Angstrom
```

In addition, every compound has a *raw NOE file*. This is a *.csv* file containing the NOEs with the atom identifiers reported in the paper. However, this file does not include NOE values that were removed for quality reasons (see above).

raw_NOEs/[ID].csv:

```
columns 1, 2: atom identifiers, as reported in the publication.
Columns 3, 4, …: NOE values, as reported in the paper. The type of column
(distance, upper bound, lower bound, are described in the raw_NOE columns property.
A conversion factor to Angstroms is in raw_NOE conversion factor)
```

Further, every compound has a reference topology, given as .pdb, .mol2 and amber .prmtop file.

Topologies/[ID].xxx

In addition, every compound has a processed NOE .json file, which contains the converted NOEs that match the topology.

Processed_NOEs/[ID].json

MOLECULAR DYNAMICS SIMULATIONS

SI TEXT S3: MACROCONF WORKFLOW: SETUP OF MD SIMULATIONS

Each MD simulation is labelled with a hash that uniquely identifies the simulation based on the input parameters specified in the tabular input file.

Compound	Method	Solvent	Simtime	repeats
22	cMD, GaMD, aMD	H_2O	2000	1

hash = f({parameters})

hash	Compound	method	Solvent	Simtime	repeat
ea902b72	22	GaMD	H_2O	2000	1
734cdc50	22	cMD	H_2O	2000	1
bb375067	22	aMD	H_2O	2000	1

Figure S2: Top: example tabular input for the MacroConf workflow MD module. For convenience, multiple different parameters can be set in the same row. Bottom: as part of the workflow, a hash for every distinct parameter set / simulation run is computed.

MD topologies are created by the Snakemake rules 'md_build_leap' and 'md_make_topology'. The 'md_build_leap' rule creates a leap file that is then run by the 'md_make_topology' rule to produce AMBER topologies. Since tleap does not automatically recognize head-tail cyclic bonds, the head-tail cyclic bond was added for head-tail cyclic peptides during this process in tleap. After adding the ring-closing bond, the system was then solvated with the appropriate solvent to match the NMR solvent in octahedral solvent boxes.

For technical details about the implementation of the MD simulations in the MacroConf workflow, we refer the reader to the documentation, which can be found at https://github.com/bigginlab/Macroconf/blob/public/docs/MD.md.

SI TEXT S4: CLUSTERING OF MD SIMULATIONS

Many structures found during an MD simulation are qualitatively redundant with only minor coordinate changes. We clustered the simulation trajectories to identify representative conformations. As in Cipcigan, Smith [1], we used t-SNE (with a perplexity of 50, a learning rate of 400, and 2000 timesteps) as the dimensionality reduction method to better separate different structures. For this, we used the reduced dihedral angles as input features. We then clustered with the density based spatial clustering of applications with noise (DBSCAN) algorithm to identify different clusters in the t-SNE representation. The resulting clusters for compound 22 in a 2000 ns GaMD simulation in aqueous solvent are shown in Figure S3. Figure S4 shows which cluster is occupied over time. Frequent changes between different clusters and re-visiting of most clusters indicated convergence.

To get a representative structure for each cluster, we averaged the cartesian coordinates of all structures belonging to the same cluster. To avoid unphysical representative cluster structures, we selected the closest (lowest RMSD) observed structure in the simulation trajectory as the representative structure for each cluster. Finally, we compared the obtained representative cluster structures to the PCA representation to see if minima coincide with the representative cluster structures (see Figure 4A of the main manuscript).

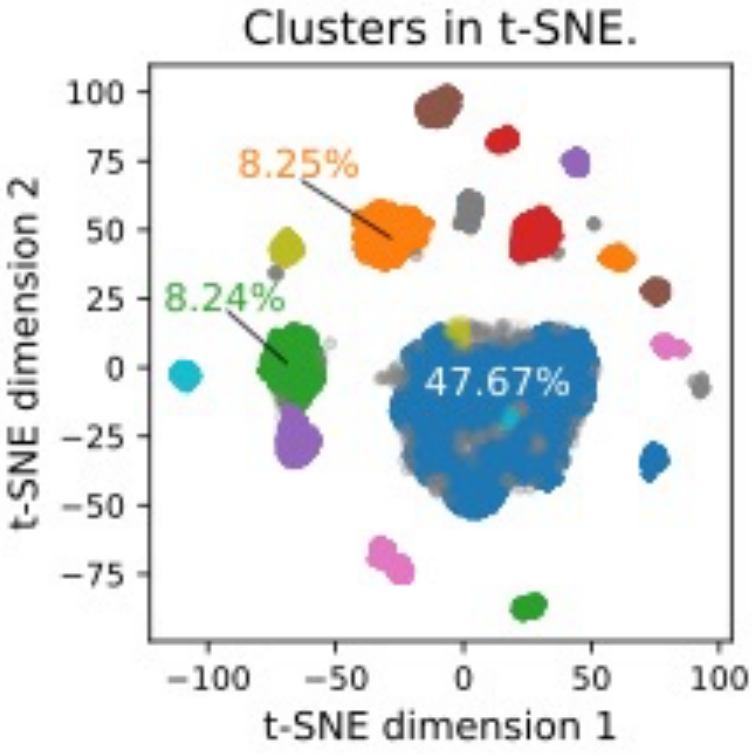

Figure S3 Structural clusters of the MD trajectory in t-SNE representation. The 3 most populated cluster are labelled with the % of snapshots associated with the cluster, relative to all snapshots from the MD trajectory.

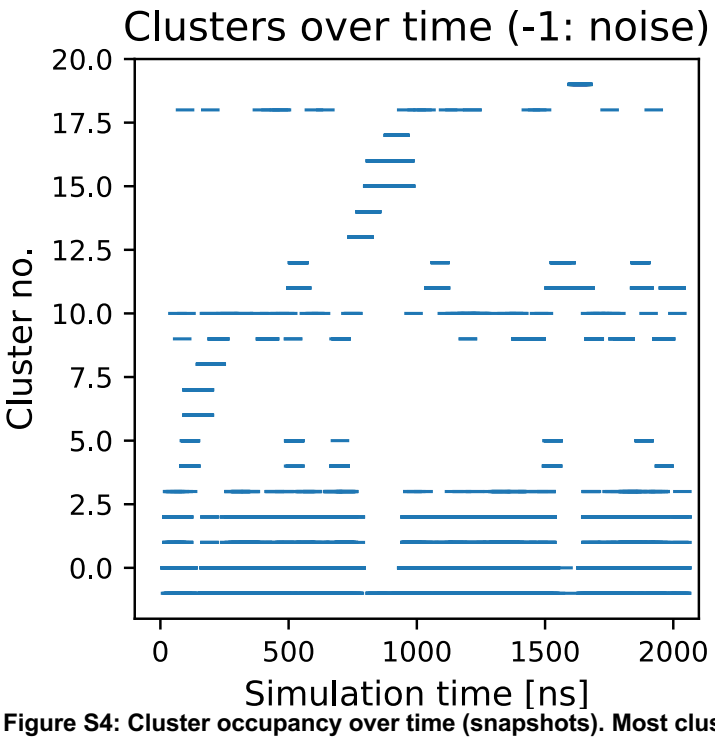

Figure S4: Cluster occupancy over time (snapshots). Most clusters are re-visited several times.

SI TEXT S5: CONVERGENCE OF MD SIMULATIONS

To ensure convergence, several metrics that capture structural diversity were tracked over the course of the simulation. In Figure S5, we considered RMSD of atomic positions of different atom types in reference to the first simulation frame. In Figure S6, we followed the ω-dihedral angles, which indicates whether any cis/trans flips happen in the backbone. Structures of different RMSDs are revisited several times throughout the 2000 ns GaMD simulation of compound 22, which indicates convergence. Similar behaviour is observed for the ω-dihedral angles: several cis/trans flips of some dihedral angles are observed multiple times, also indicating convergence.

We repeated many MD simulations with increasing simulation times to ensure convergence. Viewing multiple simulations in the same dihedral PCA space of a reference simulation allowed to compare sampling of different simulations. Observing structures that are present in one but not in another identical repeat simulation implies the simulations are not converged. Figure S7 shows different GaMD simulations of compound 22 with different total simulation times. Increasing the simulation time from 1000 ns to 2000 ns only changed the PES landscape minimally, whereas shorter simulations, such as 100 ns, did not converge and cover only a fraction of the PES surface of the longer simulations.

The GaMD simulations for compound 22 converged at 1000 ns simulation time. To ensure convergence, we ran most simulations for 2000 ns. Other solvents might take different simulation times to converge. As an additional convergence check, we performed 2000 ns GaMD simulations with different starting structures. We used the structural clusters identified in our analysis of the 2000 ns GaMD run in H_2O as starting structures. Figure S8 shows comparable dPCA-PES for all replica runs, therefore we consider the simulations converged.

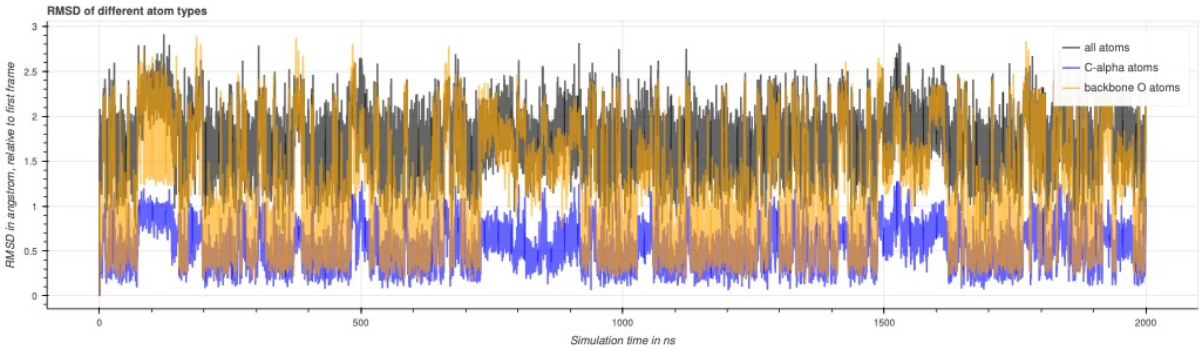

Figure S5: RMSD of different atom types to assess convergence.

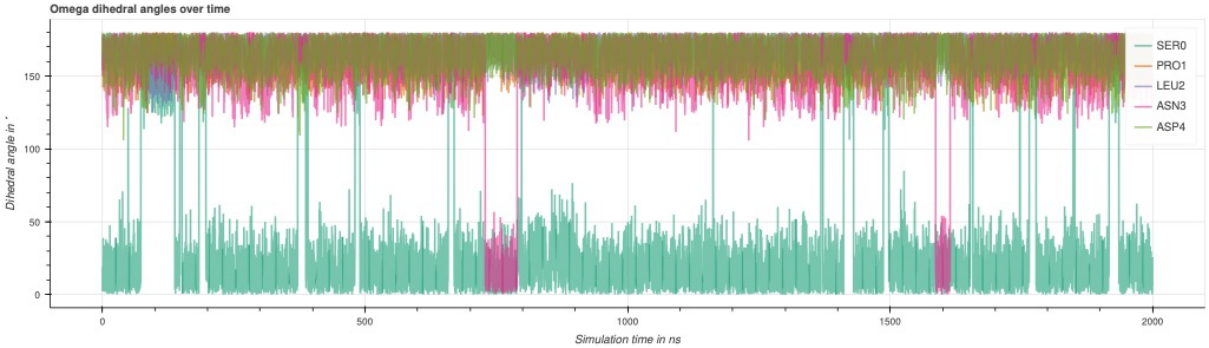

Figure S6: Omega dihedral angles of the cyclic peptide to assess convergence.

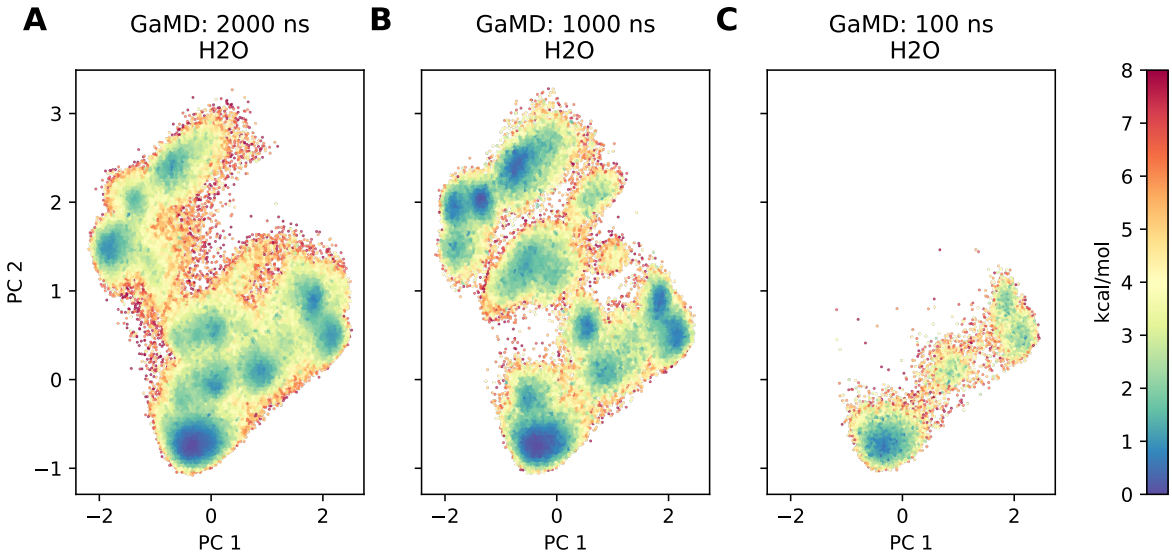

Figure S7 dPCA PES of 2000, 1000, 100 ns GaMD simulations in H_2O. The middle and right plots are within the PCA space of the left plot.

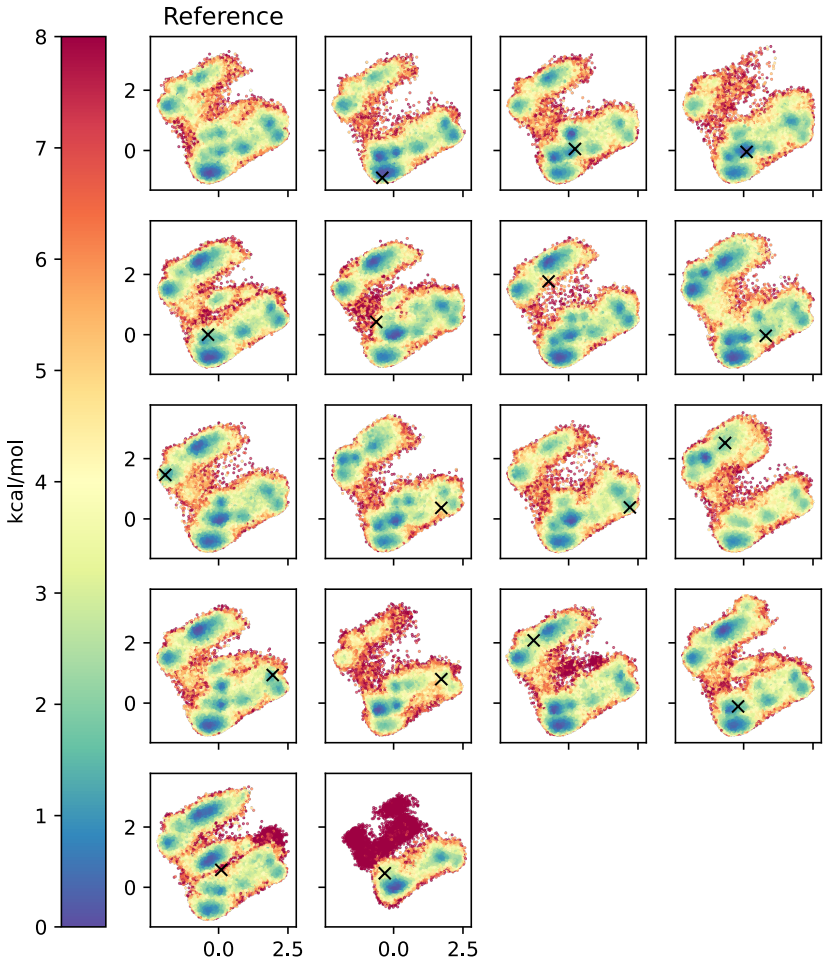

Figure S8: Extended convergence check, showing the reference simulation (top left), and replica simulations starting at every cluster that was identified in the reference simulation. The respective starting structures of the replica (clusters in the reference) are denoted with a black x. All replica simulations are in the same dihedral PCA space as the reference simulation for comparison. The axes show the first and second principal components.

To investigate the required sampling times in a more rigorous way, we introduce a metric termed *convergence time*. We define the convergence time of a simulation as the time at which 95% of the final d-PCA surface is reproduced. To compute the convergence time, we divide the final d-PCA surface into a (10 x 10) grid and track the fraction of occupied grid cells every 100 ns relative to the total simulation (see Figure S10). An exemplary convergence plot for compound 22 is shown in Figure S9. The convergence time for this simulation is ~ 500 ns.

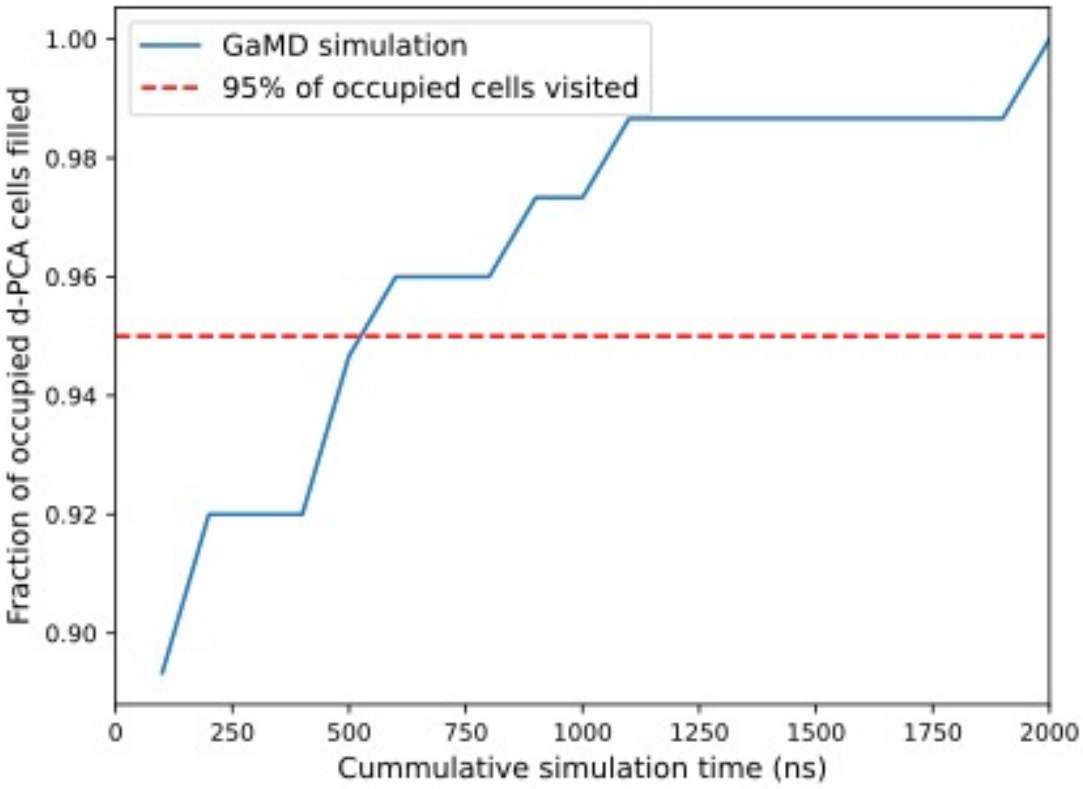

Figure S9: Convergence analysis of a 2000 ns GaMD simulation of compound 22.

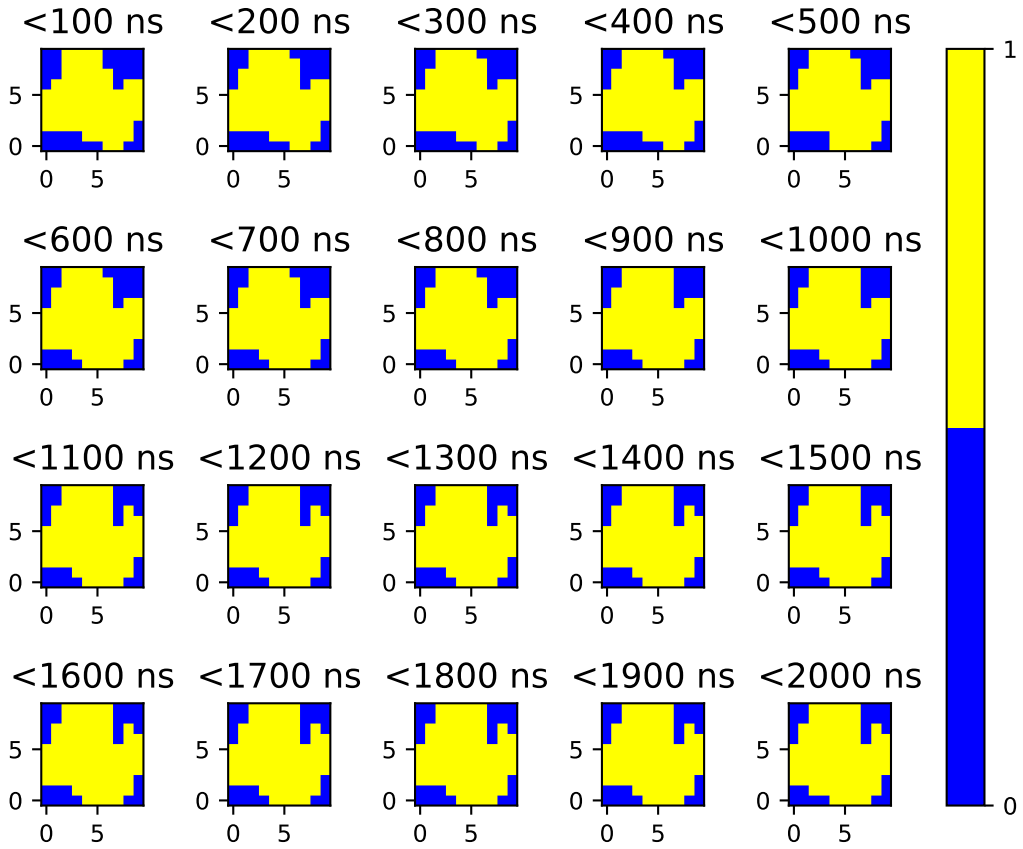

Figure S10: Shown is the dPCA grid (10x10) that was used for the convergence analysis. Values of 0 indicate that a grid cell was not visited, values of 1 indicate that the simulation visited a grid cell.

We determined the convergence time for all GaMD simulations to investigate if the total simulation time of 2000 ns could have been reduced. Figure S11 shows a histogram of the convergence time for all compounds simulated. We observe a wide-spread in convergence times. The mean convergence time is 830 ± 530 ns. This indicates that the simulation time of 2000 ns could have been reduced for some compounds. However, other compounds required almost the full simulation times to reach convergence in their respective d-PCA space. Compounds 27 (penta-peptide), 30 (penta-peptide), 49 (hexa-peptide), 58 (hepta-peptide), 61 (nona-peptide) required the longest convergence times.

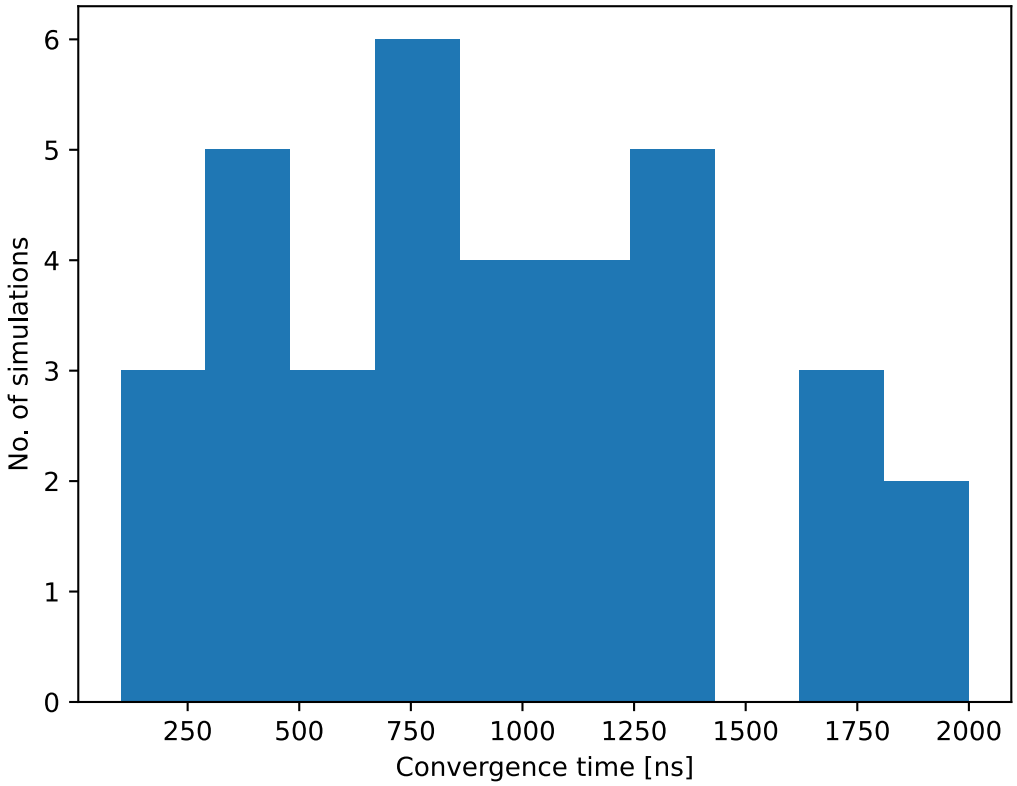

Figure S11: Shown is a histogram of the convergence time of all GaMD simulations performed.

SI TEXT S6: COMPUTING NOE DISTANCE CONSTRAINTS

Figure S12 shows the reweighted NOE distances - energy profiles. The 1d reweighting was performed separately for each NOE pair. Due to the reweighting and r^{-6} averaging, simulated NOE distances are not symmetrically distributed around the computed average. Mirroring Kamenik, Lessel [2] we computed population weighted RMSDs from the mean NOE distance separately for values above and below the mean. These two values give an indication of the variance of the mean of the simulated NOE distances. These values are not indicative of the full fluctuations observed in the MD simulations, which can be inferred from Figure S12.

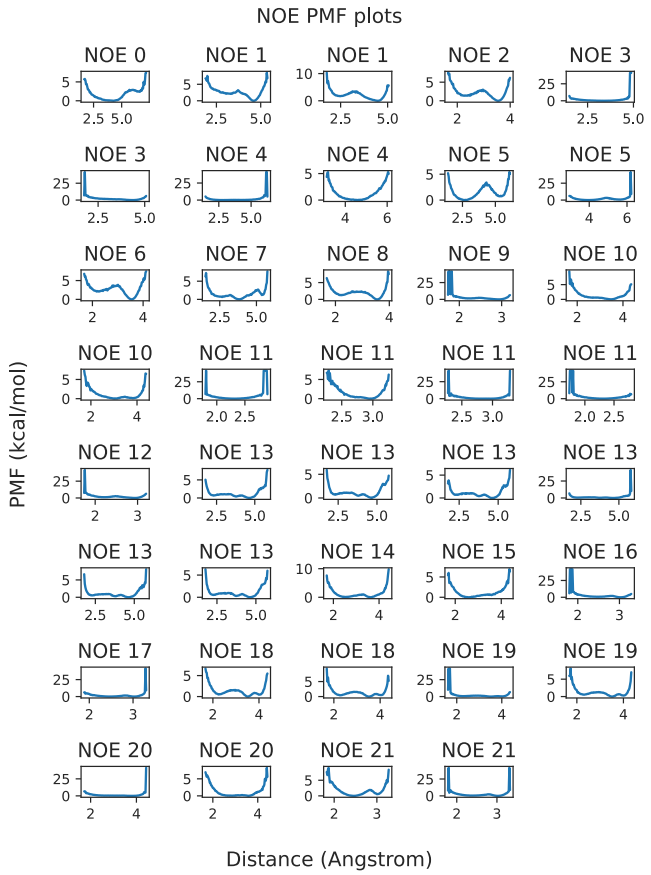

Figure S12: Potential of mean force (PMF) plotted against distances (in Angstroms). PMF are derived via reweighing of corresponding NOE distances.

SI TEXT S7: OUTLIERS 24, 49

Shown below are two compounds (24, 49) for which cMD only sampled the trans isomers. GaMD, however, was able to sample both cis and trans structures.

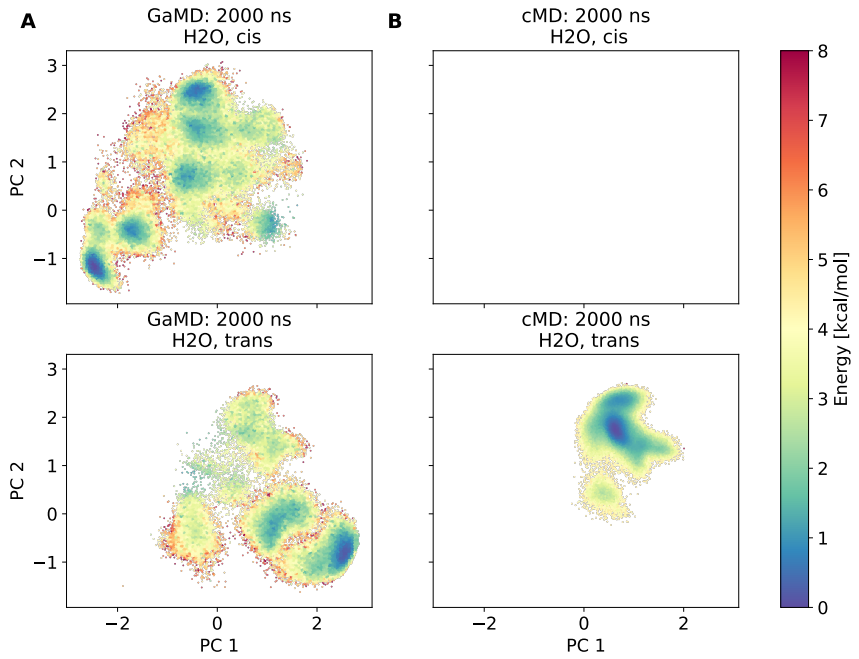

Figure S13: Dihedral PCA plots of compound 24. A: GaMD simulations sample both cis and trans isomers. B: cMD only samples the trans isomers.

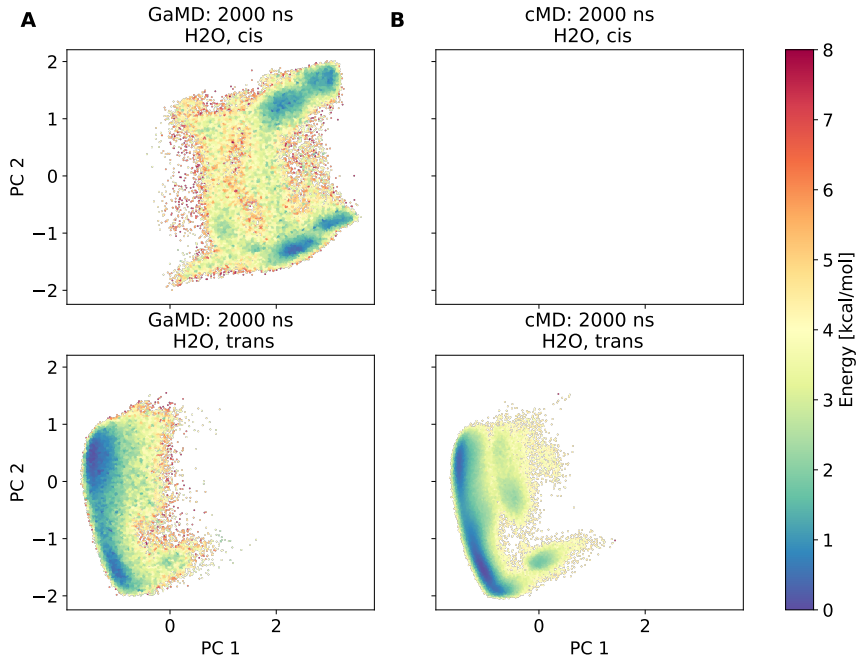

Figure S14: Dihedral PCA plots of compound 49. A: GaMD simulations sample both cis and trans isomers. B: cMD only samples the trans isomers.

METHOD COMPARISON

SI TEXT S8: SIGNIFICANCE TESTS FOR METHOD COMPARISONS

To test for statistically significant performance differences between simulation methods, we used the paired Student's t-test and Wilcoxon signed-rank tests to consider differences in mean and mean signed rank, respectively. P-Values of these tests are shown in the following Figures. For all statistical tests, NaN values (i.e., when no conformers existed) were omitted.

As described in the main manuscript, we applied the method of Holm to control the family-wise error rate, which results in scaled p-values. The method of Holm was applied separately for each metric considered (% of NOE distance restraints satisfied and RMSD) but combined for the two statistical tests applied. In total, we compared 8 different methods, which led us to perform 56 statistical tests per metric. In the below figures, we show the p-values of statistical tests performed as 2d heatmaps. The upper triangular part of each heatmap shows the via the method of Holm corrected p-values, the lower triangular part shows the uncorrected p-values.

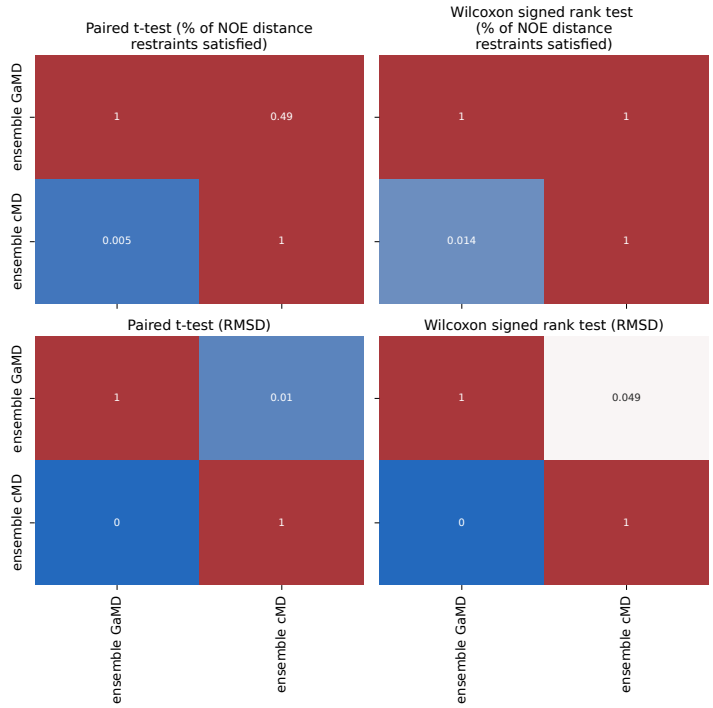

Figure S15: Significance tests for comparison of Gaussian accelerated MD vs conventional MD (Figure 8 in the main manuscript). The upper triangular part shows the corrected p-values (method of Holm), the lower triangular part shows the uncorrected p-values.

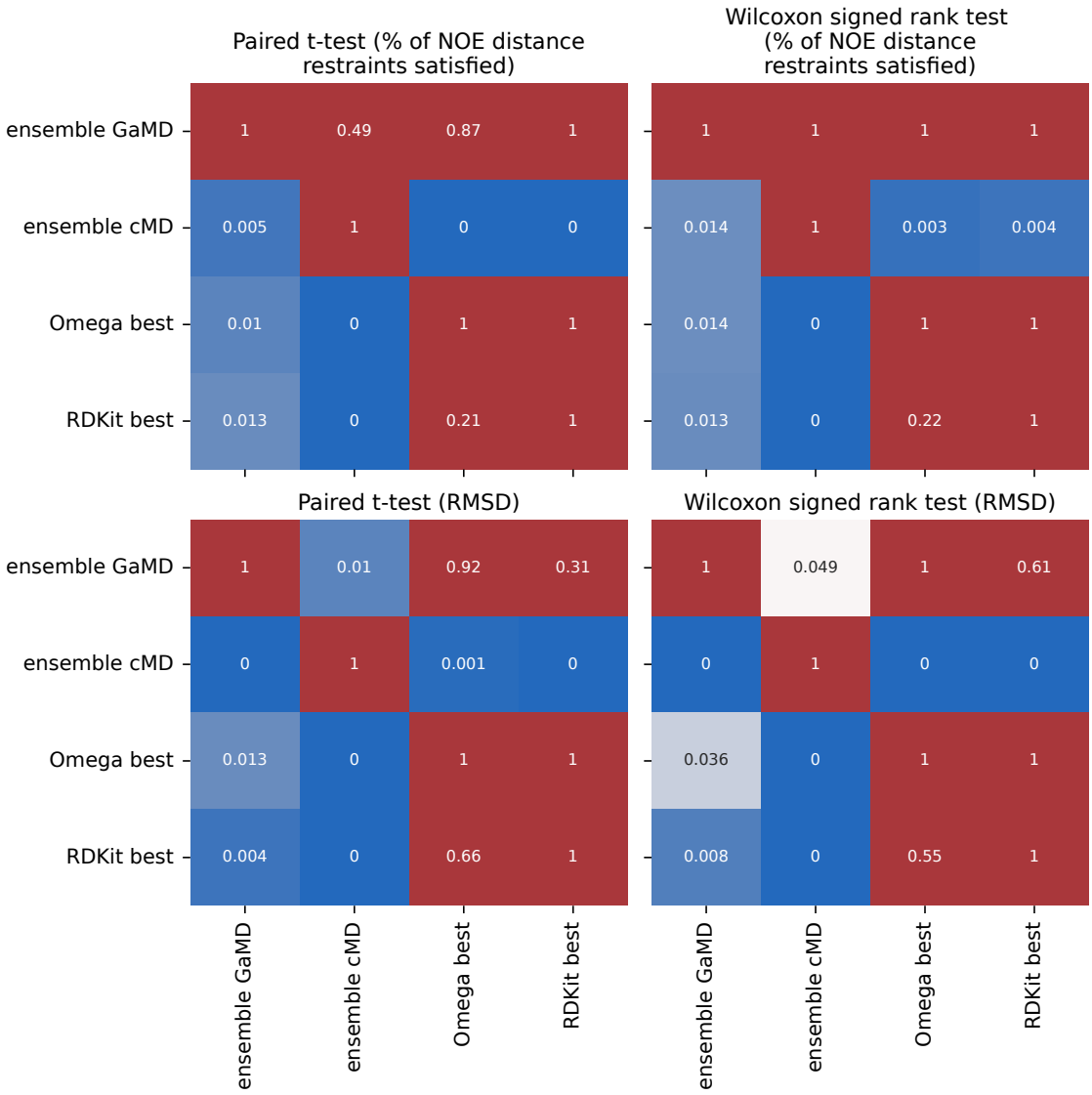

Figure S16: Significance tests for comparison of best cheminformatics structures vs MD ensembles (Figure 9 in the main manuscript). The upper triangular part shows the corrected p-values (method of Holm), the lower triangular part shows the uncorrected p-values.

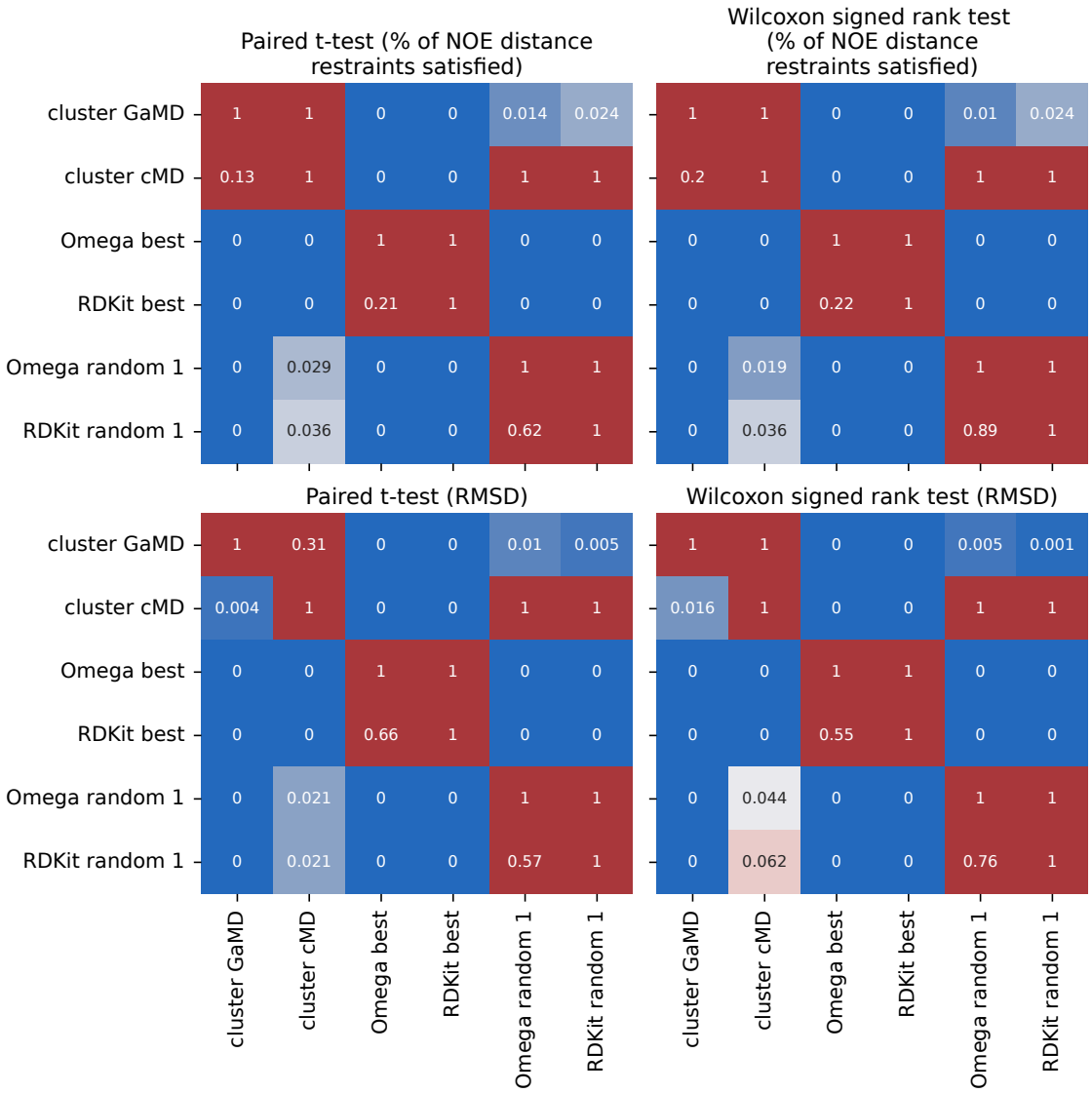

Figure S17: Significance tests for comparison of best cheminformatics structures vs most populated cluster from MD (Figure 10 in the main manuscript). The upper triangular part shows the corrected p-values (method of Holm), the lower triangular part shows the uncorrected p-values.

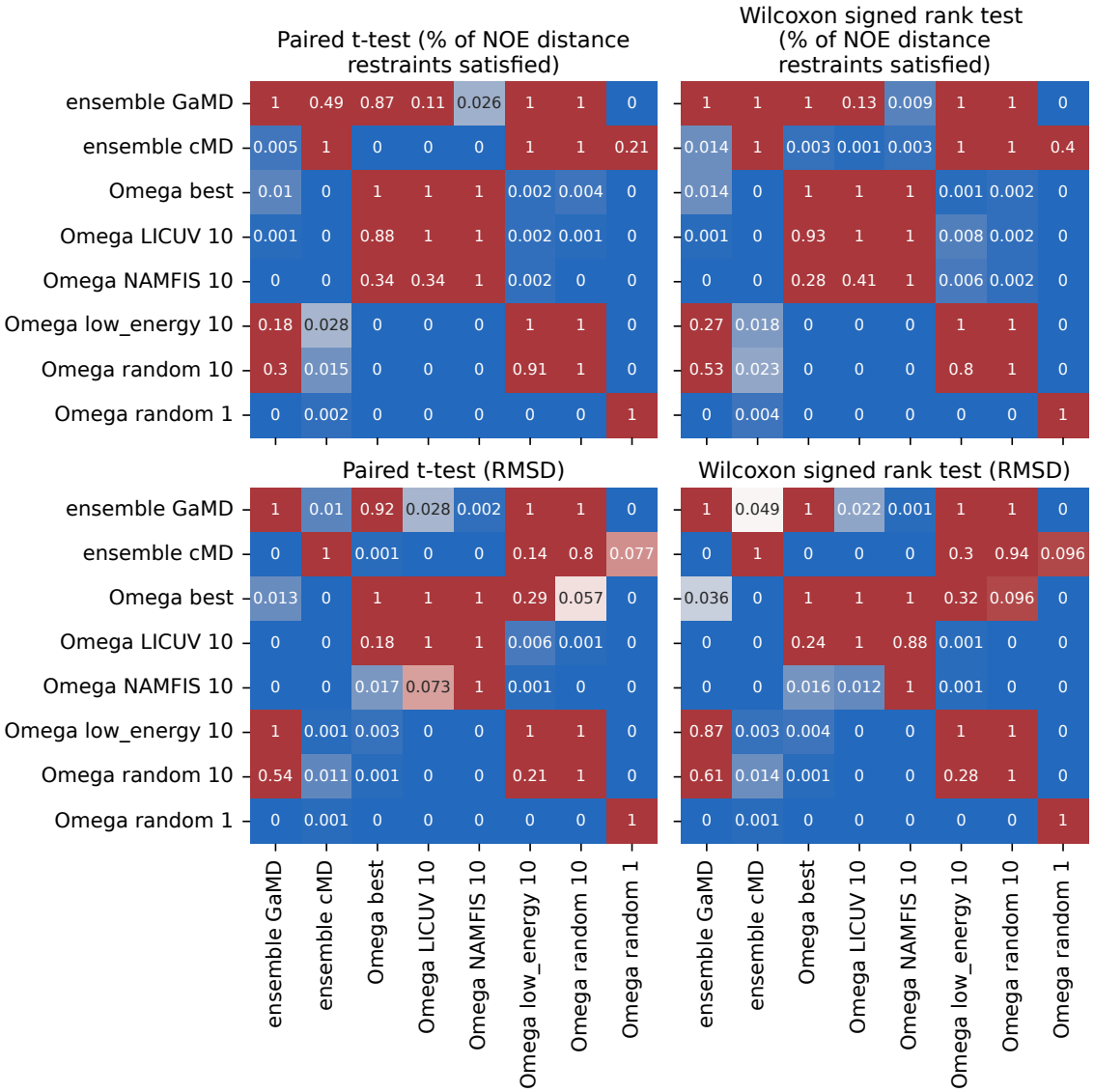

Figure S18: Significance tests for comparison of bundles of cheminformatics structures (Omega Macrocycle) vs MD ensembles (Figure 11A in the main manuscript). The upper triangular part shows the corrected p-values (method of Holm), the lower triangular part shows the uncorrected p-values.

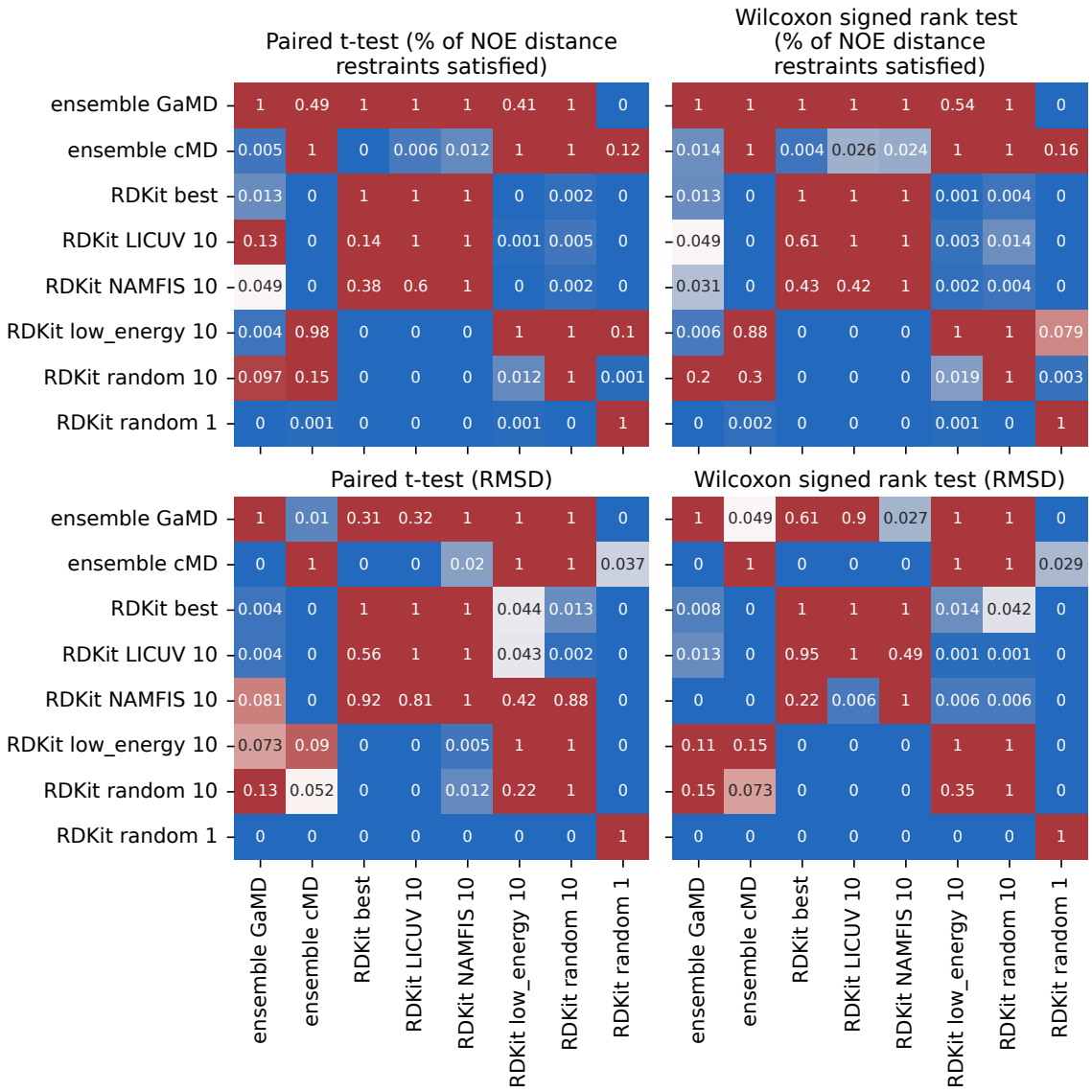

Figure S19: Significance tests for comparison of bundles of cheminformatics structures (RDKit ETKDG) vs MD ensembles (Figure 11B in the main manuscript). The upper triangular part shows the corrected p-values (method of Holm), the lower triangular part shows the uncorrected p-values.

Table S1: Percentage of NOE distance restraints satisfied values [%] for different MD and cheminformatics methods.

	mean	std	min	50%	max
ensemble GaMD	74.41	18.73	20.83	75.76	100.00
cluster GaMD	66.76	22.50	20.83	68.49	94.19
ensemble cMD	67.10	20.77	31.25	65.52	100.00
cluster cMD	63.75	23.51	20.83	59.09	95.83
Omega best	78.26	15.35	47.83	80.00	100.00
Omega LICUV 10	78.41	17.48	39.13	80.85	100.00
Omega low_energy 10	70.64	19.09	29.17	72.50	100.00
Omega random 1	59.14	19.05	28.33	62.00	85.42
Omega random 10	70.89	18.82	33.04	75.81	100.00
Omega NAMFIS 10	79.76	16.33	42.86	86.96	100.00
RDKit best	77.22	16.96	43.48	80.00	100.00
RDKit LICUV 10	74.97	20.12	21.74	80.65	100.00
RDKit low_energy 10	66.15	22.52	20.83	74.24	100.00
RDKit random 1	58.86	19.96	25.65	59.30	92.00
RDKit random 10	68.65	20.36	16.52	72.94	100.00
RDKit NAMFIS 10	75.76	20.60	26.09	83.33	100.00

Table S2: RMSD values [Å] for different MD and cheminformatics methods.

	mean	std	min	50%	max
ensemble GaMD	0.62	0.35	0.19	0.46	1.34
cluster GaMD	0.73	0.27	0.23	0.74	1.26
ensemble cMD	0.75	0.29	0.26	0.70	1.29
cluster cMD	0.84	0.26	0.42	0.76	1.46
Omega best	0.54	0.30	0.21	0.38	1.17
Omega LICUV 10	0.51	0.28	0.22	0.37	1.27
Omega low_energy 10	0.60	0.32	0.20	0.44	1.37
Omega random 1	0.97	0.44	0.42	0.81	1.94
Omega random 10	0.62	0.34	0.20	0.49	1.37
Omega NAMFIS 10	0.48	0.30	0.16	0.33	1.23
RDKit best	0.53	0.29	0.24	0.39	1.14
RDKit LICUV 10	0.55	0.33	0.23	0.36	1.37
RDKit low_energy 10	0.68	0.41	0.22	0.49	1.64
RDKit random 1	1.01	0.47	0.52	0.75	2.16
RDKit random 10	0.64	0.38	0.25	0.45	1.40
RDKit NAMFIS 10	0.54	0.37	0.16	0.33	1.44

SI TEXT S9: CHEMINFORMATICS CONFORMER GENERATORS: BUNDLE SIZE EFFECTS & DISTRIBUTIONS OF NOE METRICS

When comparing different bundling methods in [3], a bundle size of 10 was chosen. However, different bundle sizes can also impact the performance. Figure S20 shows the effect of different bundle sizes and methods for compound 22 and Omega Macrocycle. For a bundle size of 1, LICUV and NAMFIS coincide, since they both select the conformer, which best agrees with the experimental NOE values. As can be seen from the distribution of the % of NOE distance restraints satisfied metric for all produced conformers in Figure S21, the best conformer selected in Figure S20 makes up the very tail of the distribution, which is centred around 0.45 % of NOE distance restraint satisfied values.

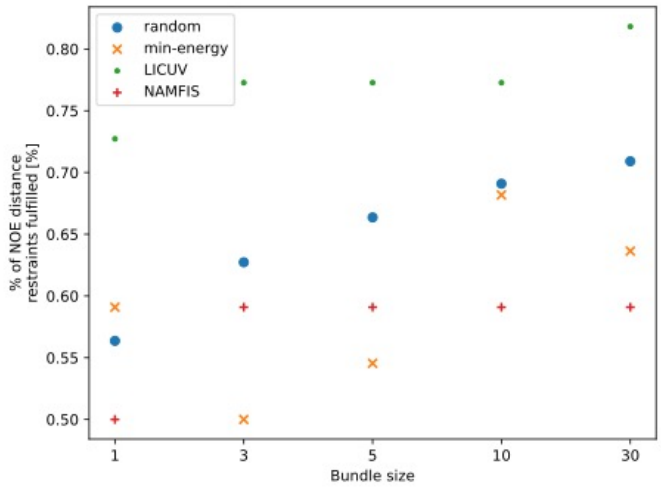

Figure S20: Compound 22, Omega Macrocycle. Comparison of different bundling methods and bundle sizes.

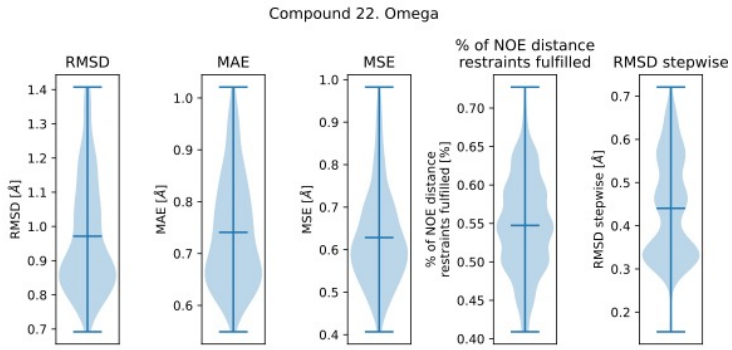

Figure S21: Compound 22, Omega Macrocycle. NOE metrics were computed for each produced conformer separately. Shown are the resulting distributions over the different NOE metrics.

SI TEXT S10: EFFECT OF DIFFERENT RE-WEIGHTING METHODS OF GAMD

To produce physical averages from biased GaMD trajectories, reweighting is required. For the main analysis, GaMD simulations were reweighted via Maclaurin series reweighting. However, Boltzmann reweighting is also a viable reweighting method, although it tends to be noisier. In Figure S22, we compare the two reweighting methods for all compounds simulated. Maclaurin reweighting of the NOEs leads to better agreement with the experimentally observed NOEs, compared to Boltzmann reweighting. As can be seen from Fig. S22. these differences are statistically significant.

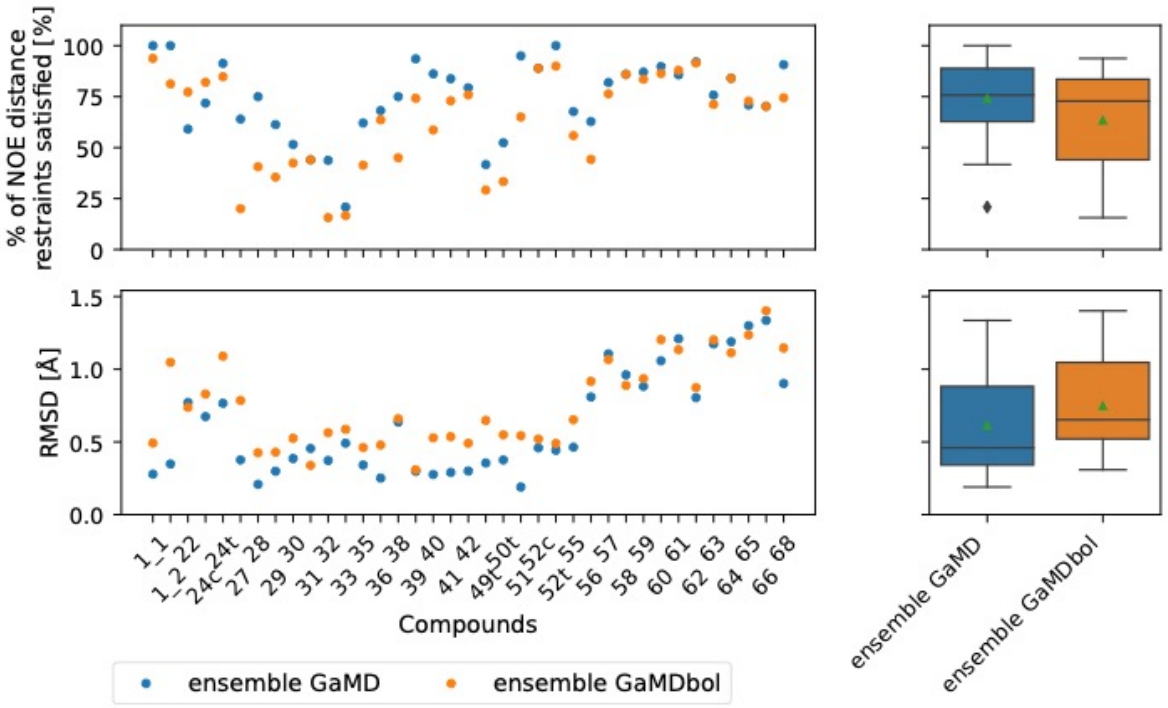

Figure S22: Comparison of GaMD ensembles, obtained via Maclaurin series reweighting (blue), and via Boltzmann reweighting (ensemble GaMDbol, orange). Maclaurin series reweighting leads to better agreement with the experimentally observed NOEs.

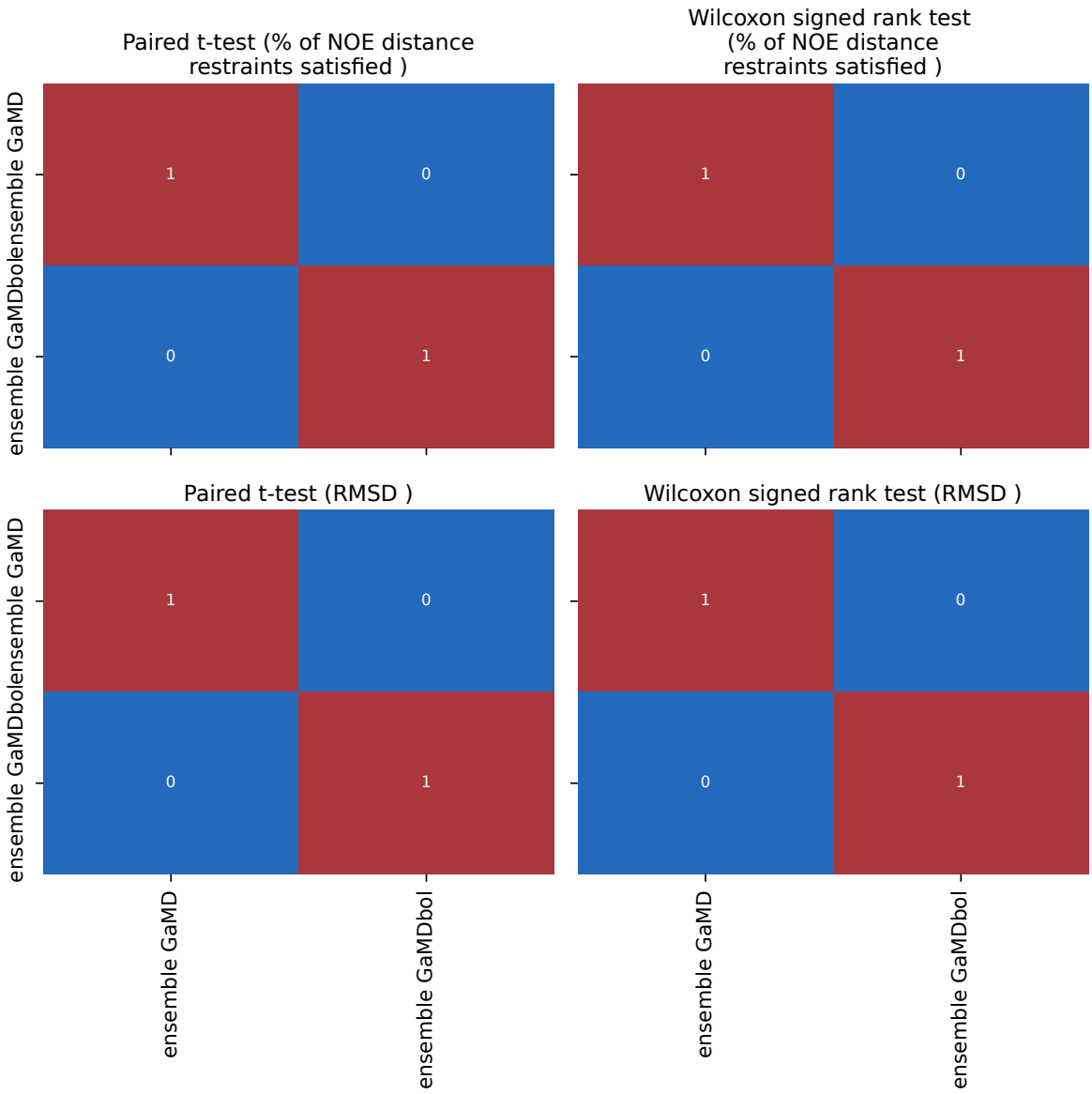

Figure S23: Significance tests for Maclaurin and Boltzmann reweighting (Figure S22 above).

SI TEXT S11: EFFECT OF A STEPWISE RMSD METRIC

The two main metrics we use to assess how well the conformer generators agree with the experimental NOE data are % of NOE distance restraints satisfied and RMSD. The former metric takes experimentally reported bounds into account, however not in a quantitative way. The latter metric measures more quantitatively how far the simulated NOEs deviate from the reference values, but it does not consider the experimental bounds. To combine the two approaches, we adopt a stepwise RMSD definition:

RMSD = {0 if within bounds, > 0 if outside bounds.

This metric should provide a more quantitative measure of NOE violations. We observe that the stepwise RMSD metric is similar to the % of NOE distance restraints satisfied metric: The cMD ensembles perform better than the GaMD ensembles in this metric, as for % of NOE

distance restraints satisfied. Alike to the % of NOE distance restraints satisfied metric; this difference is not statistically significant (see Fig. 9 and S17; and S24, S25 below).
As for comparing cheminformatics conformer generators to the MD methods (Fig. 10, S16), the stepwise RMSD again performs equivalently to the % of NOE distance restraints satisfied metric (Fig. S26, S27).

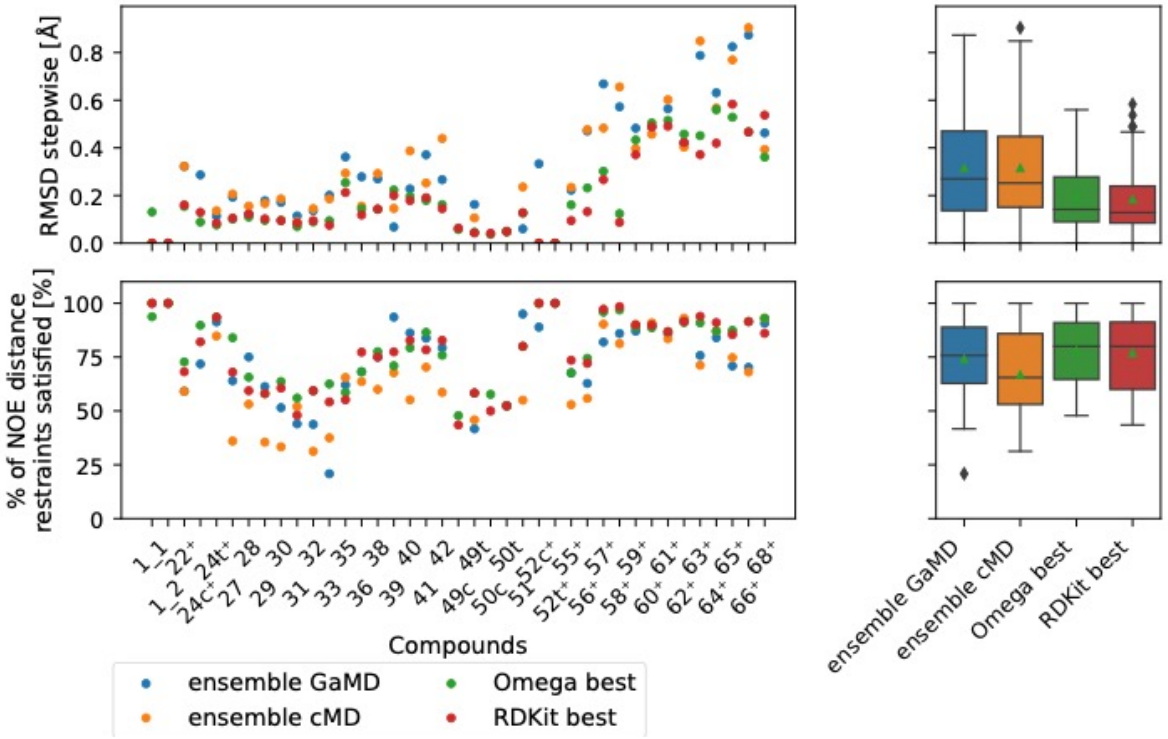

Figure S24: Comparison of the stepwise RMSD metric to the % of NOE distance restraints satisfied metric.

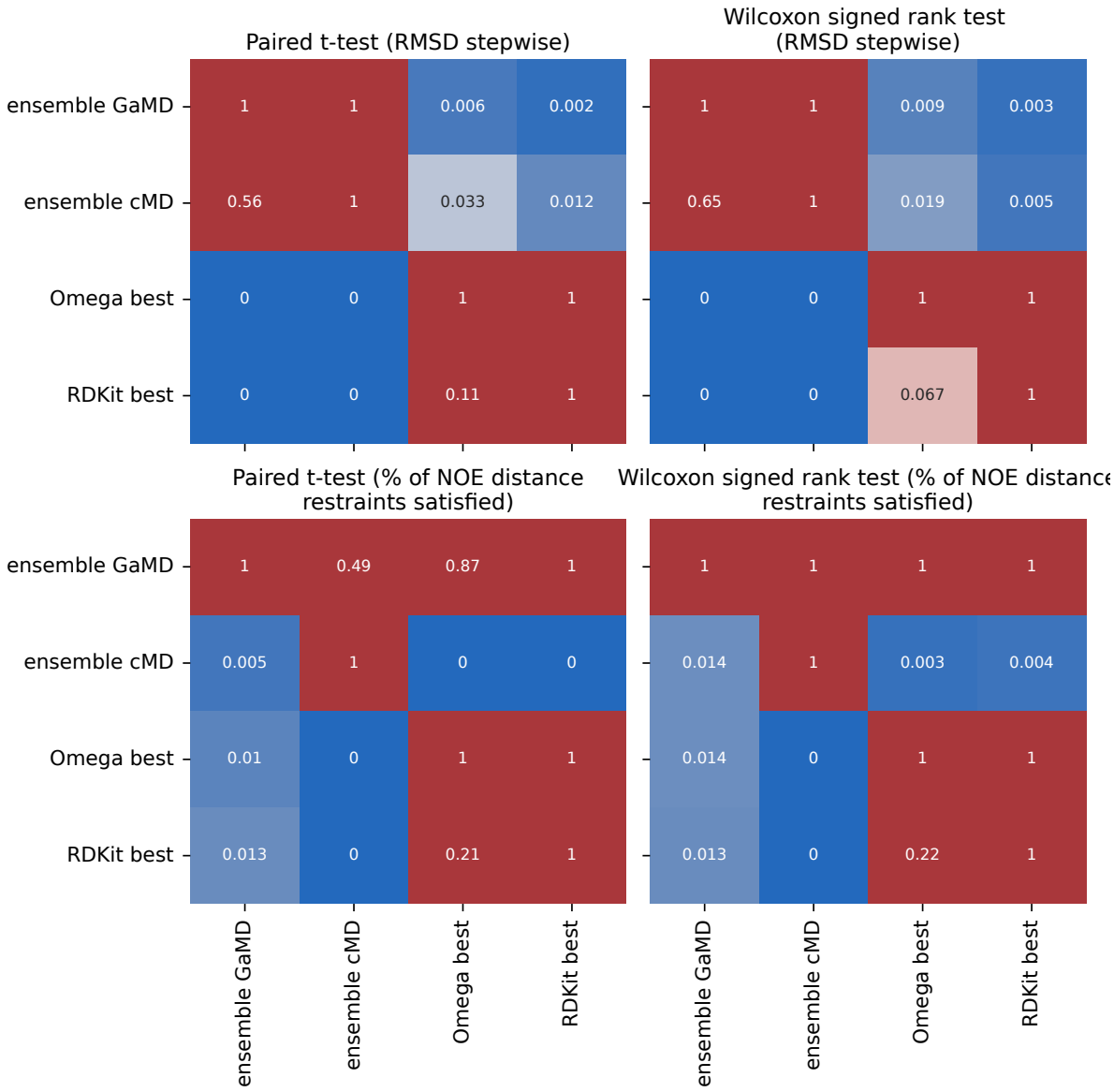

Figure S25: Significance tests for comparison of the stepwise RMSD metric to the % of NOE distance restraints satisfied metric. The upper triangular part shows the corrected p-values (method of Holm), the lower triangular part shows the uncorrected p-values. Both metrics perform similar, the stepwise RMSD metric shows significant differences between the GaMD ensembles and Omega/RDKit, which are not significant with the % of NOE distance restraints satisfied metric.

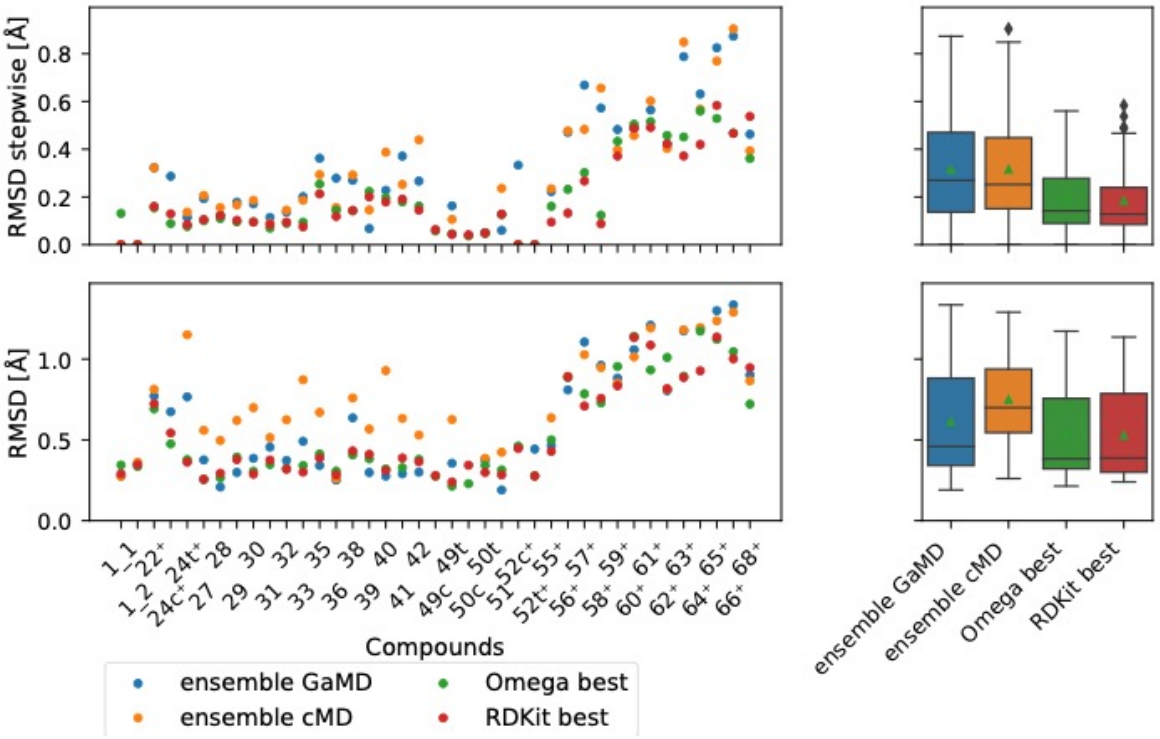

Figure S26: Comparison of the stepwise RMSD metric to the RMSD metric.

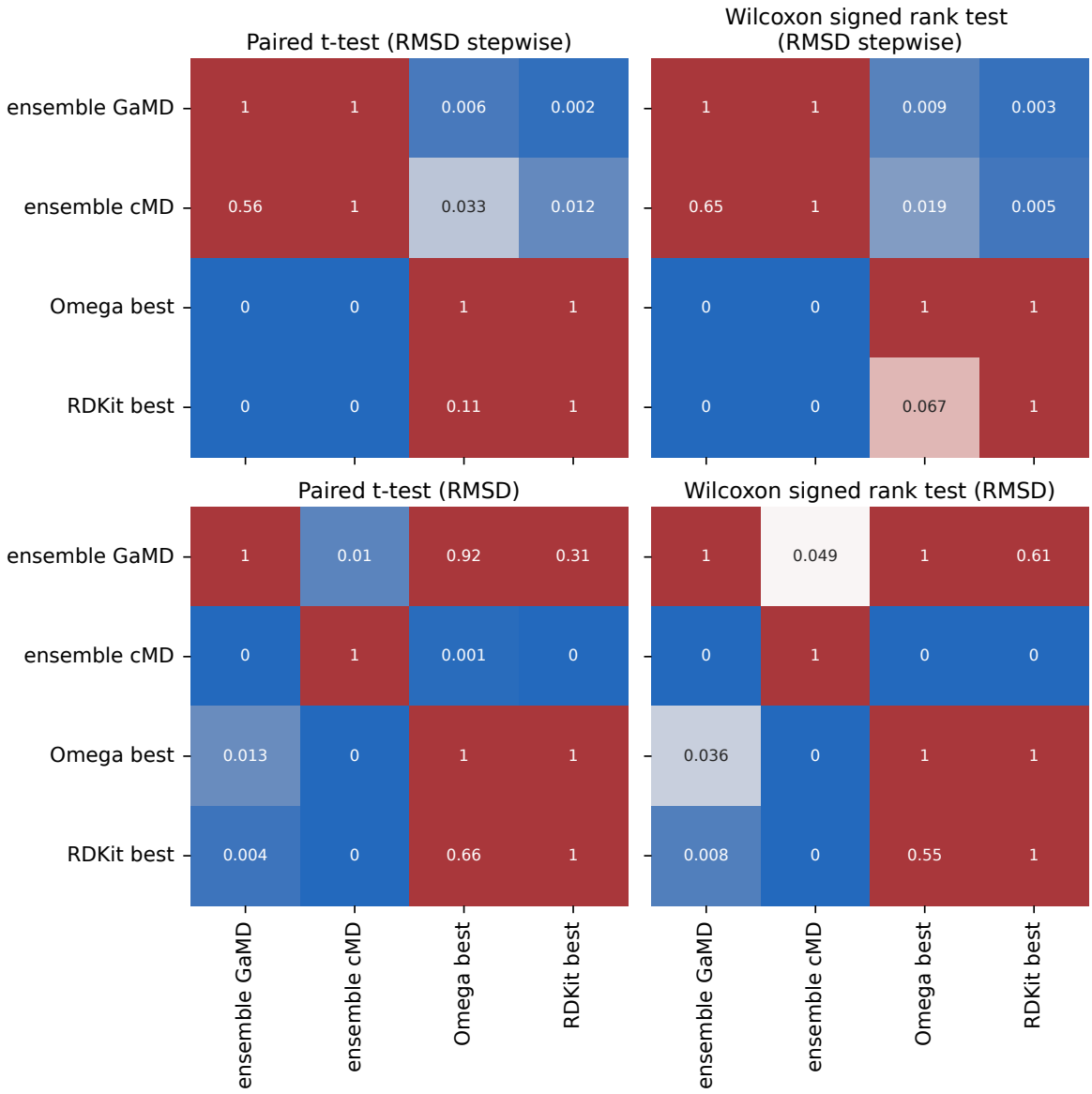

Figure S27: Significance tests for comparison of the stepwise RMSD metric to the RMSD metric. The upper triangular part shows the corrected p-values (method of Holm), the lower triangular part shows the uncorrected p-values. Both metrics perform similar, the stepwise RMSD metric shows significant differences between the GaMD ensembles and Omega/RDKit, which are not significant with the RMSD metric.

SI TEXT S12: NOE COVERAGE FOR VARYING PEPTIDE SEQUENCE LENGTHS

Performance of MD and cheminformatics conformer generators varies between different compounds. To investigate a possible dependence on the sequence length, we show how the % of NOE distance restraints satisfied and RMSD metrics differ for subsets of the MacroConf datasets in Figure S28. However, we would like to highlight that comparisons between different subsets are problematic for several reasons: 1. The MacroConf dataset is imbalanced, there are many more compounds with shorter sequence lengths (see Figure 2B), 2. Different compounds were measured in different experimental conditions and to varying quality levels. This makes NOE violations not comparable across different compounds. 3. Different NOEs can have varying restraining power, which makes a direct comparison

between subsets of compounds problematic. We can however consider relative performance differences within subsets for different methods. As can be seen in Figure S28, there are no performance variations between methods for different subsets of compounds (sequence lengths = 5, 6, > 6).

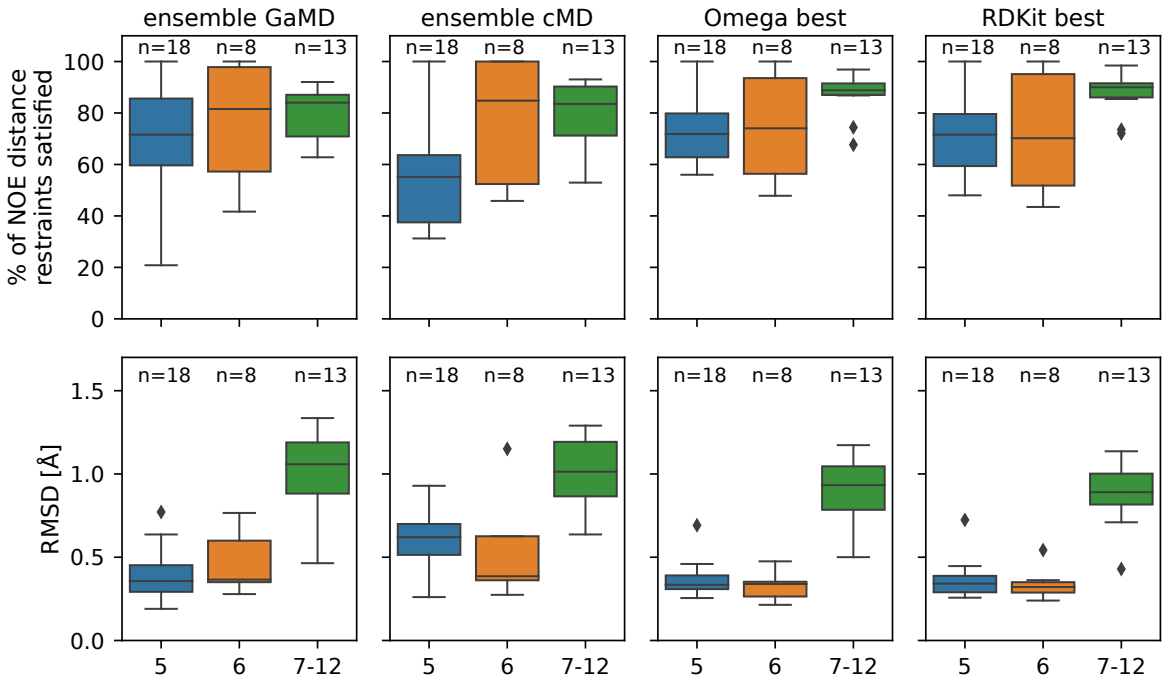

Figure S28: Shown are 3 subsets of the simulated compounds with sequence lengths of 5, 6, and 7-12. Between different methods, there are no clear performance differences.

SI TEXT S13: COMPARISON OF SOLVATION PROPERTIES BY DIFFERENT METHODS

We compare solvent accessible and polar surface area (SASA, PSA) of MD with the cheminformatics conformer generators. For GaMD, we report the most likely value of SASA and PSA after reweighting. For cMD, we report the mean value of SASA and PSA. The cheminformatics methods produce conformers that generally agree with the values from GaMD and cMD simulations.

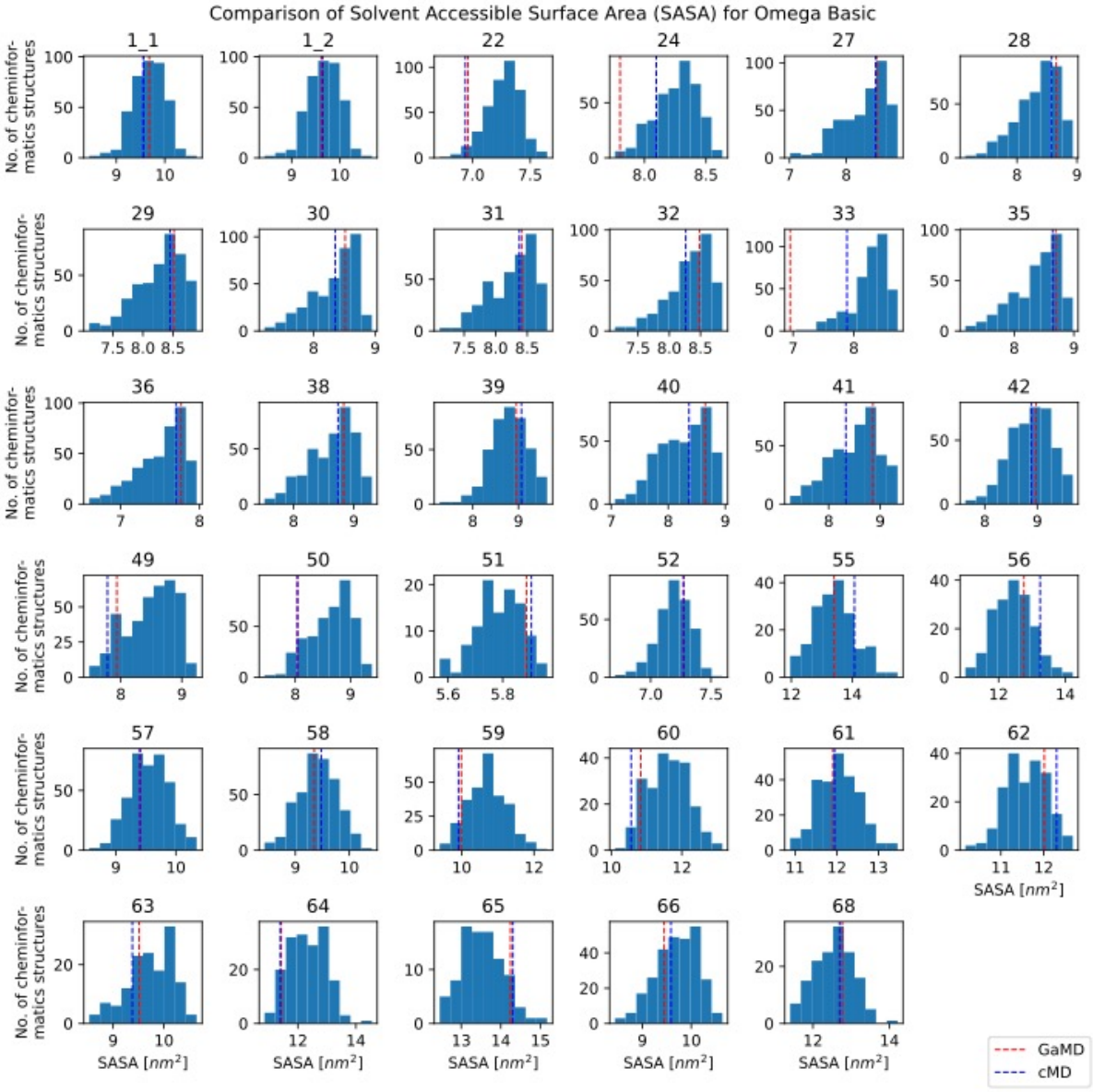

Figure S29: Shown are histograms of the Solvent Accessible Surface Area (SASA) for the Omega Macrocycle conformer generator. The dashed lines show the most likely SASA value for GaMD after reweighting, and the mean SASA value for cMD. For most compounds, the values from the MD simulations are reproduced by the Omega conformer generator.

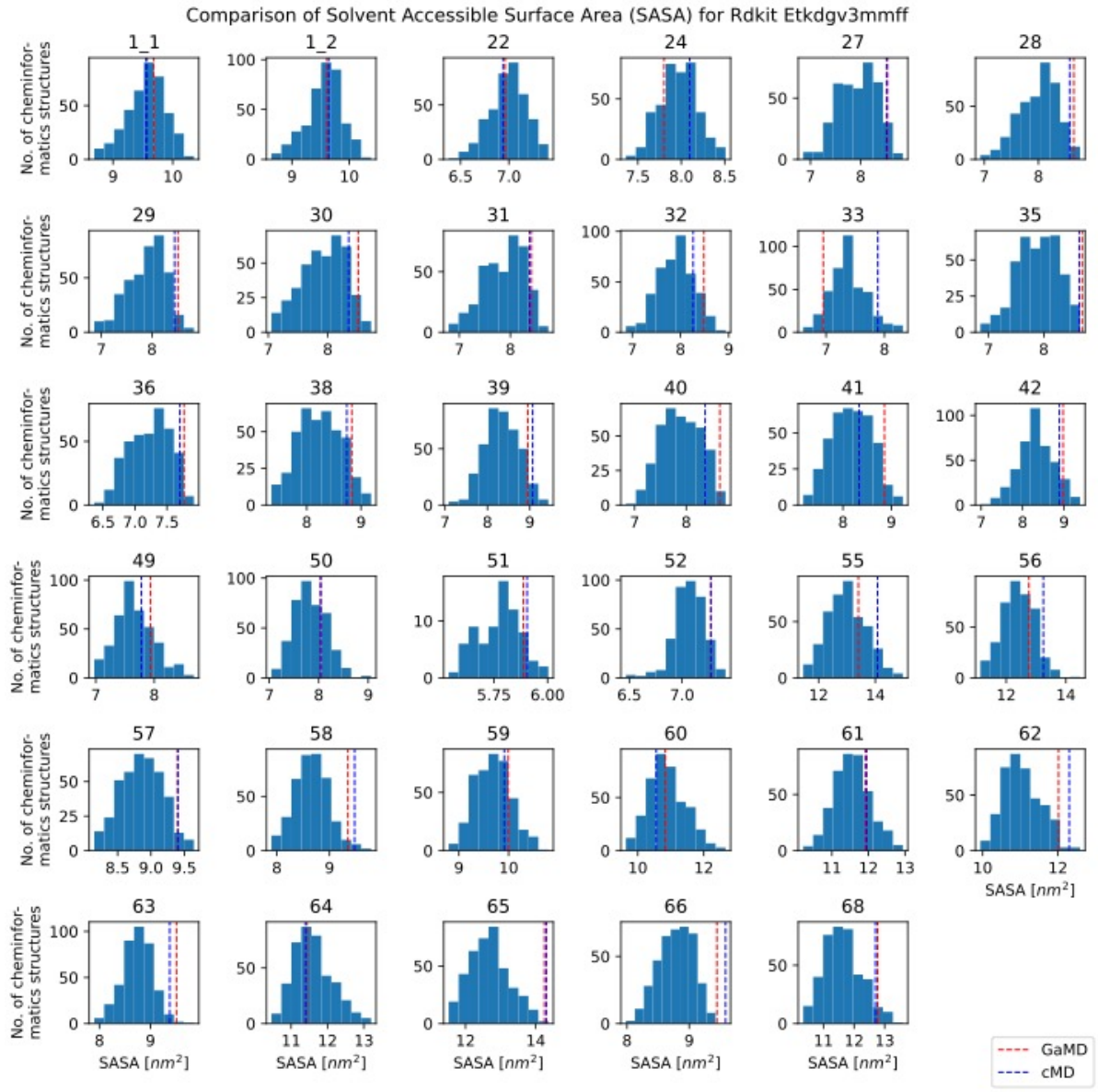

Figure S30: Shown are histograms of the Solvent Accessible Surface Area (SASA) for the RDKit ETKDGv3 conformer generator. The dashed lines show the most likely SASA value for GaMD after reweighting, and the mean SASA value for cMD. For most compounds, the values from the MD simulations are reproduced by the RDKit conformer generator.

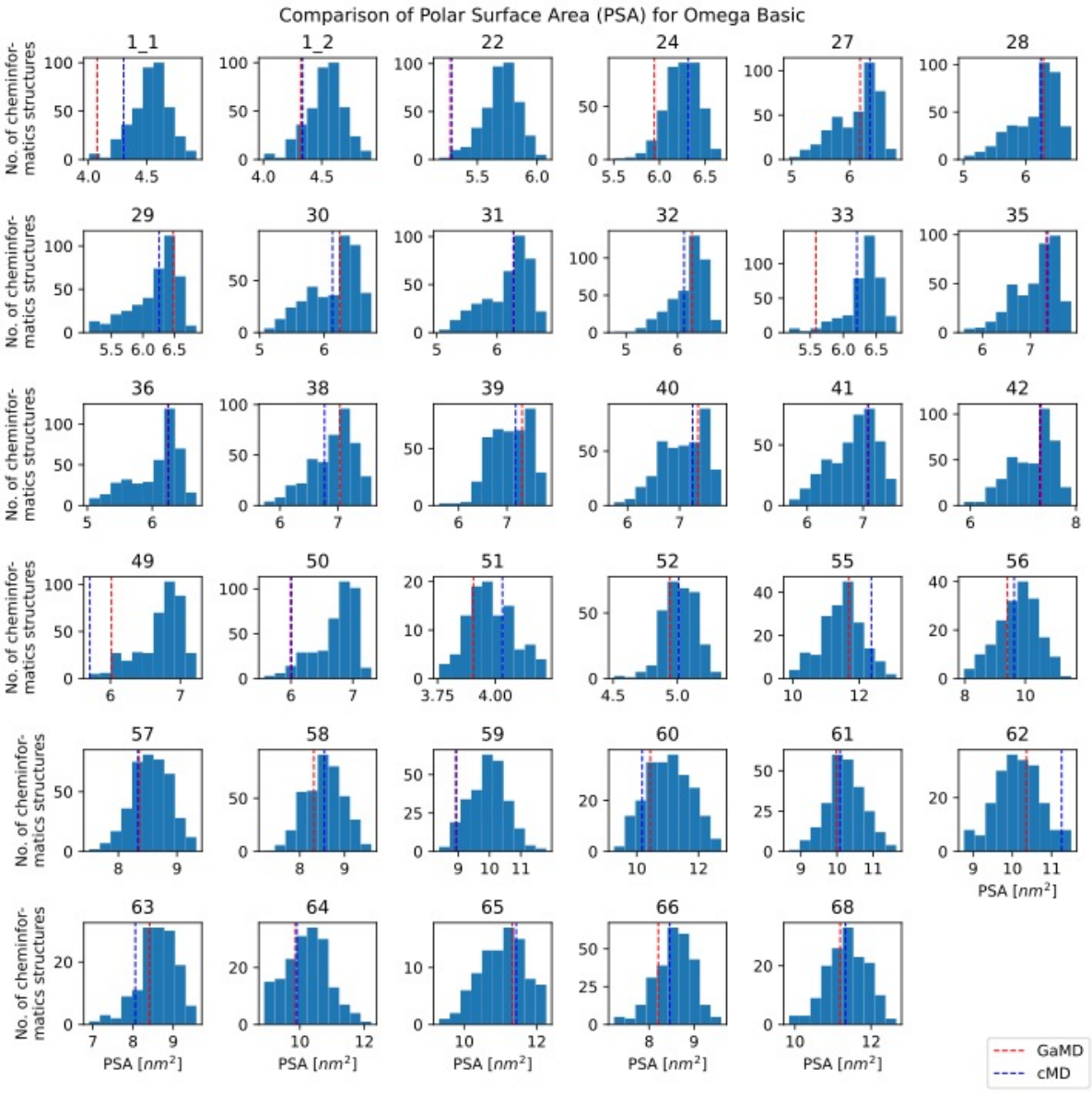

Figure S31: Shown are histograms of the Polar Surface Area (PSA) for the Omega Macrocycle conformer generator. The dashed lines show the most likely PSA value for GaMD after reweighting, and the mean SASA value for cMD. For most compounds, the values from the MD simulations are reproduced by the Omega conformer generator.

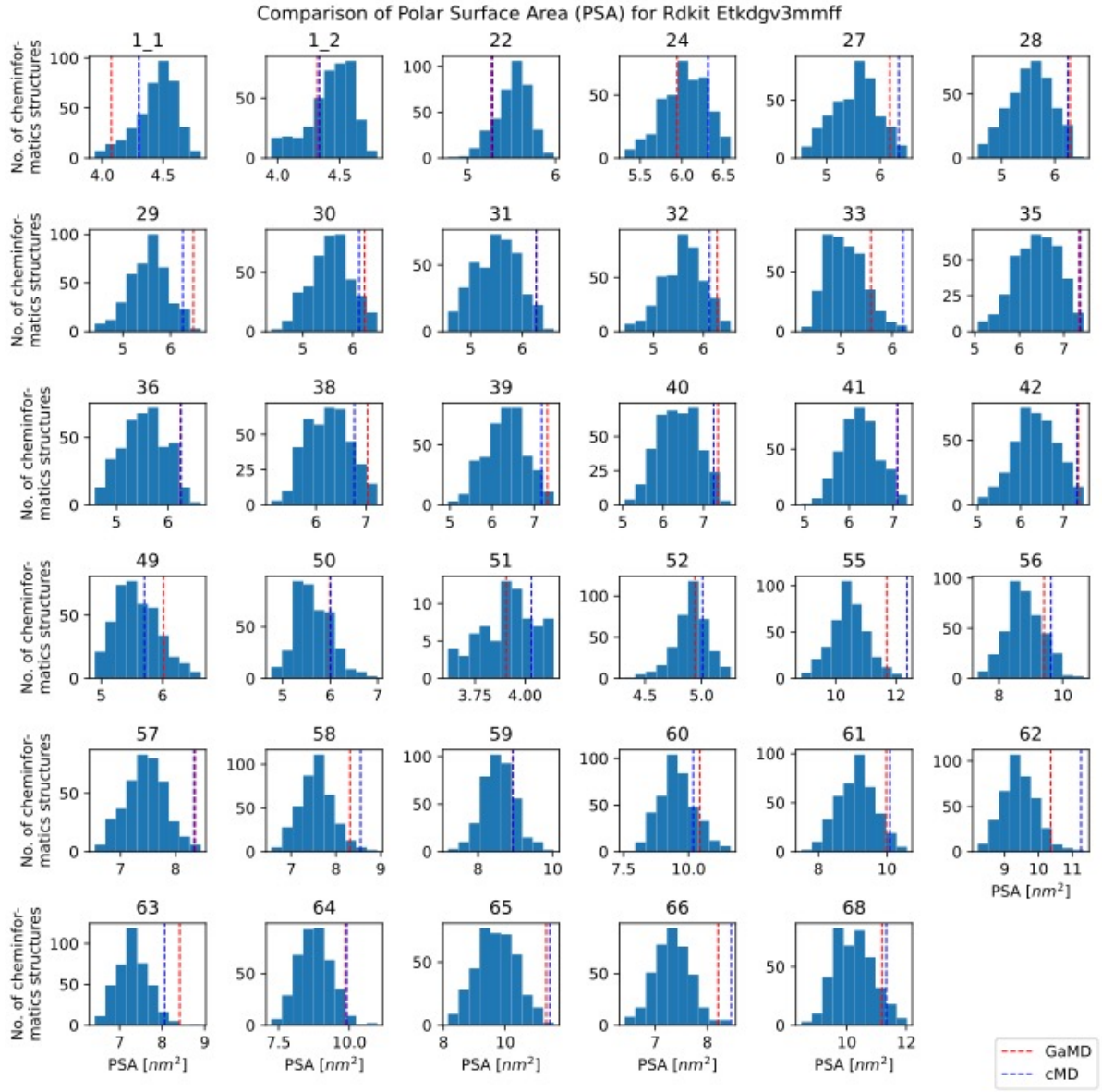

Figure S32: Shown are histograms of the Polar Surface Area (PSA) for the RDKit ETKDGv3 conformer generator. The dashed lines show the most likely PSA value for GaMD after reweighting, and the mean SASA value for cMD. For most compounds, the values from the MD simulations are reproduced by the RDKit conformer generator.

THE MACROCONF WORKFLOW

SI TEXT S14: EXTERNAL DEPENDENCIES AND INSTALLATION

The MacroConf package requires lots of external dependencies. Wherever possible, these are automatically installed via Snakemake through the conda package manager. Definitions of the conda environments are provided in the workflow/envs folder. The following additional programs are required:

- Conda: https://docs.conda.io/projects/conda/en/stable/user-guide/install/index.html

- Amber: https://ambermd.org/Installation.php
- OpenEye toolkits and OpenEye Omega: https://www.eyesopen.com/omega, https://docs.eyesopen.com/applications/gettingstarted.html

DATA AVAILABILITY

The MacroConf dataset is publicly available. All simulation data can be found at: https://github.com/D-Cru/Macroconf

CODE AVAILABILITY

An open-source software implementation of the computational workflow used to simulate and analyse the MacroConf dataset is available at https://github.com/D-Cru/Macroconf.

REFERENCES

1. Cipcigan, F., et al., *Membrane Permeability in Cyclic Peptides is Modulated by Core Conformations.* J Chem Inf Model, 2021. **61**(1): p. 263-269.
2. Kamenik, A.S., et al., *Peptidic Macrocycles - Conformational Sampling and Thermodynamic Characterization.* J Chem Inf Model, 2018. **58**(5): p. 982-992.
3. Wang, S., et al., *Incorporating NOE-Derived Distances in Conformer Generation of Cyclic Peptides with Distance Geometry.* J Chem Inf Model, 2022.