Supporting Information: CGCompiler: Automated coarse-grained molecule parameterization via noise-resistant mixed-variable optimization

Kai Steffen Stroh,[†,‡] Paulo C. T. Souza,[¶,§] Luca Monticelli,[¶] and Herre Jelger Risselada[*,†,‡,‖]

†*Department of Physics, Technische Universität Dortmund, 44227 Dortmund, Germany*

‡*Institute for Theoretical Physics, Georg-August University Göttingen, 37077 Göttingen, Germany*

¶*Molecular Microbiology and Structural Biochemistry (MMSB, UMR 5086), CNRS & University of Lyon, 69367 Lyon, France*

§*Laboratory of Biology and Modeling of the Cell, École Normale Supérieure de Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5239 and Inserm U1293, 69007 Lyon, France*

‖*Leiden Institute of Chemistry, Leiden University, 2333 CC Leiden, The Netherlands*

E-mail: jelger.risselada@tu-dortmund.de

1 Melting temperature

Due to the slow kinetics of the gel-liquid phase transition, estimating the melting temperature can be difficult to estimate in simulations.[S1] For use in automated parameterization methods, where the melting temperature of many candidate solutions has to be estimated, the trade-off between accuracy and computational cost is of particular high importance. Preparing the system in stripes, i.e., half gel and half fluid (cf. Figure S1A), bypasses the slowest step in the transition, the nucleation.[S1] By simulating the biphasic system at a range of temperatures, observing the direction and rate of domain growth, and fitting rates to an Arrhenius-like equation, Coppock and Kindt have estimated T_m for atomistic DPPC and DPSM.[S1] In tests performed by us with Martini DPPC and DPSM, this procedure did not provide results reliable enough for application in a high-throughput manner. A similar biphasic system setup was used by Carpenter et al.[S2] in Martini lipid refinement. Instead of fitting domain growth rates, they used the area per lipid as a proxy for which phase prevails at a certain temperature, utilizing the fact that the highly ordered tails in the gel phase result in a much smaller area per lipid compared to the liquid phase. Phase identification with this procedure is very fast and reliable at temperatures more than a few Kelvin away from the transition temperature. The melting temperature is then given as a range between the highest temperature where the system converges to a gel phase and the lowest temperature the system ends up in the liquid phase. The accuracy of this method is strongly influenced by the employed temperature-spacing. Close to the transition temperature, longer and repeated simulations are necessary, due to the stochasticity of the melting/freezing process (cf Figure S1B).

The area per lipid has a positive linear relationship with temperature and a sharp increase at the melting temperature, as shown in Figure S2a. Therefore, to estimate the melting temperature from the temperature-dependent APL data, we fit the sigmoidal function

$$\mathrm{APL}(T) = \mathrm{APL}_0 + c \cdot T + \frac{\Delta\mathrm{APL}}{(1 + \exp(-k \cdot (T - T_\mathrm{m})))} \tag{1}$$

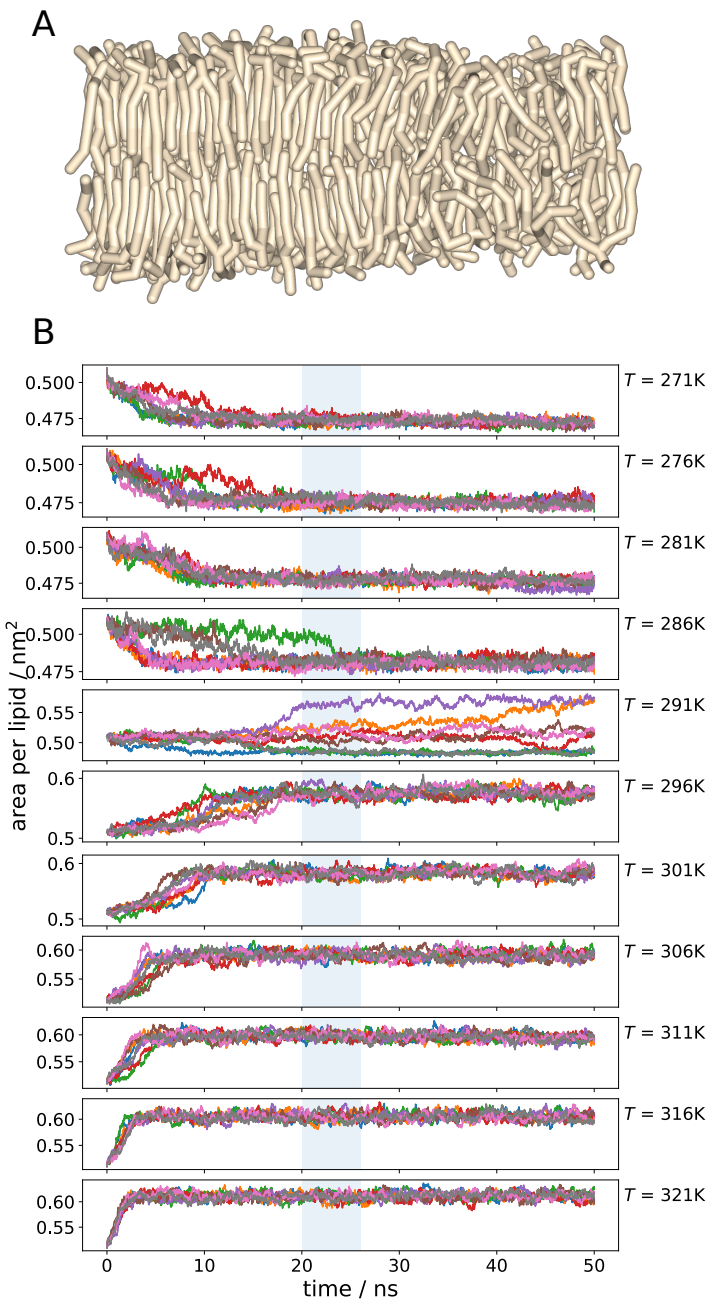
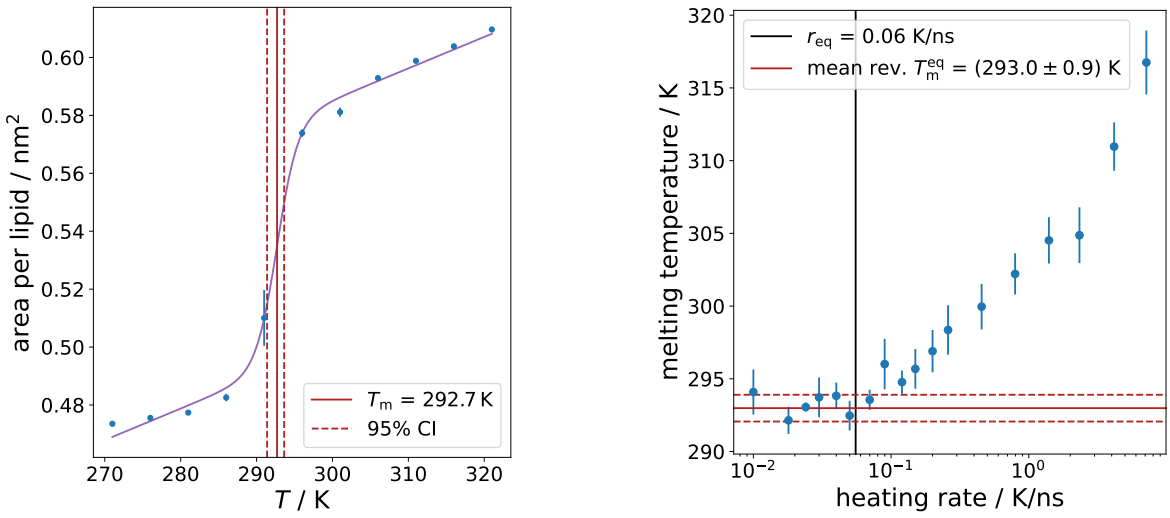


Figure S1: Melting temperature estimation. **A**: Snapshot of initial configuration of a DPSM bilayer, where half of the lipids is in the gel phase, while the other half is in the fluid phase. **B**: Area per lipid vs. time at different temperatures. The shaded area represents the time window over which the area per lipid is averaged for use in the fit of Eq. 1. Production simulations during PSO are typically 25 ns long.



(a) Tm APL fit, $T_m = 292.7\,K$ with a 95% CI of $[291.4\,K, 293.6\,K]$

(b) melting temperature.[S3]

Figure S2: Estimation of melting temperatures with two independent methods for the old DPSM model.

where T is the temperature, APL_0 is the theoretical area per lipid at $T = 0\,K$, c is the slope in the linear regime, ΔAPL describes the height of the APL jump at the melting temperature T_m, and k determines the broadness of the transition.

When using this method in an automated parameterization setting, the case that T_m might be outside of the predefined temperature range has to be handled properly. To this end we additionally fit a line to the $APL(T)$ data and use the Akaike information criterion (AIC)[S4,S5] with the modification for small sample sizes (AIC_c)[S6] to determine which model (sigmoidal or linear) is a better description of the data. If the linear model is better, i.e., it has a lower AIC_c, and all APL values are above or below a threshold (the average initial APLs), T_m is considered to be out of range and T_m is set to a particular low or high value, respectively. Hereby, candidate solutions with a melting temperature far off the target value receive a high cost value in the PSO. If the sigmoidal model is selected, or the linear model is a better fit but the APL values cross the threshold, T_m is taken from the sigmoidal fit. A few typical examples of this procedure are shown in Figure S3.

All of the above biphasic methods are sensitive to the construction of the stripe structure.

In particular, an improperly equilibrated gel phase can lead to an underestimation of T_m.[S1] An alternative, independent approach to estimate T_m is based on a two-state kinetic rate model from Kowalik et al.[S3] In this approach a system in the gel phase is heated with different heating rates. According to the two-state kinetic rate model, the melting process can be divided into regimes of reversible and irreversible melting. In the reversible melting regime, for slow heating rates, the system is assumed to be close to thermal equilibrium and melting and freezing can both occur. In this regime the apparent melting temperature is independent of the heating rate r, i.e., $T_\mathrm{m}^\mathrm{app}(r) \approx T_\mathrm{m}^\mathrm{eq}$. For fast heating rates, melting is assumed to be irreversible. In the irreversible melting regime, the two-state model predicts a dependency of the apparent melting temperature on the heating rate which can be approximated by $T_\mathrm{m}^\mathrm{app}(r) \propto \ln r$. Both regimes are divided by a characteristic heating rate r_eq.

Kowalik et al. used a series of melting simulations with fast heating rates, i.e., in the irreversible regime, to obtain several $T_\mathrm{m}^\mathrm{app}(r)$ values, determine the characteristic melting rate r_eq, and finally extrapolate the equilibrium melting temperature T_m^eq. Based on the two-state kinetic rate model from Kowalik et al., Sun and Böckmann[S7] simply used a broad range of heating rates, including the reversible regime. The equilibrium melting temperature was calculated by averaging over the $T_\mathrm{m}(r)$ in the reversible regime, i.e., $r < r_\mathrm{eq}$.

Due to the slow rates and concomitant long simulation times, we use this approach only for validation. To minimize bias caused by the quenched starting conformations we typically generate eight independent conformations for each validated candidate solution.

The rate dependent melting temperatures in this approach are obtained by fitting

$$H(T) = H_0 + c_p \cdot T + \frac{\Delta H}{(1 + \exp(-k \cdot (T - T_\mathrm{m})))} \tag{2}$$

where H is the enthalpy, c_p the heat capacity, H_0 is the enthalpy at $T = 0\,\mathrm{K}$, ΔAPL describes the height of the APL jump at the melting temperature T_m, and k determines the broadness of the transition. The functional form is the same as in Eq. 1. Figure S2b shows

results of this approach for the old DPSM model. The heating rate dependency matches the prediction of the two-state model. Comparision of Figures S2a and S2b show that the melting temperatures for the old DPSM model, obtained with both approaches, are in good agreement.

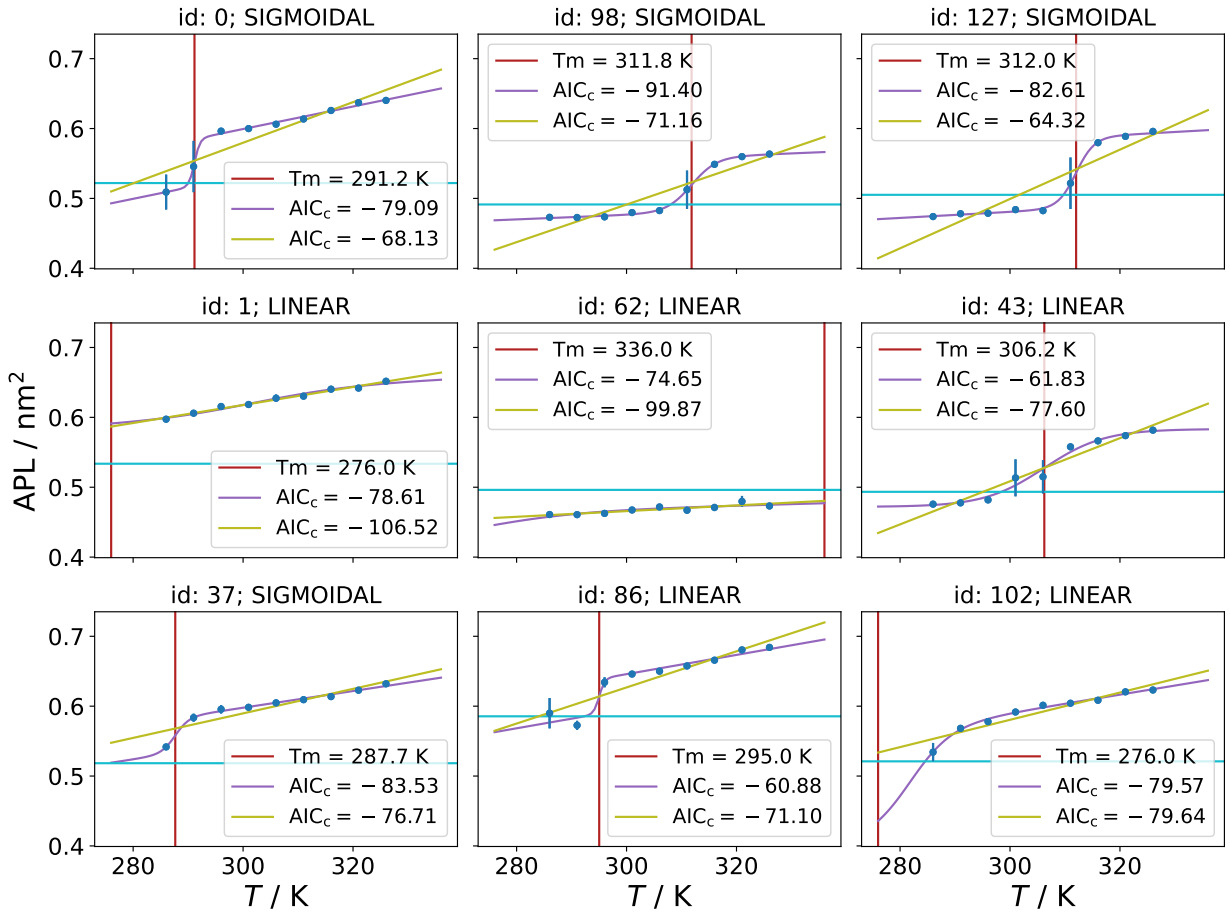
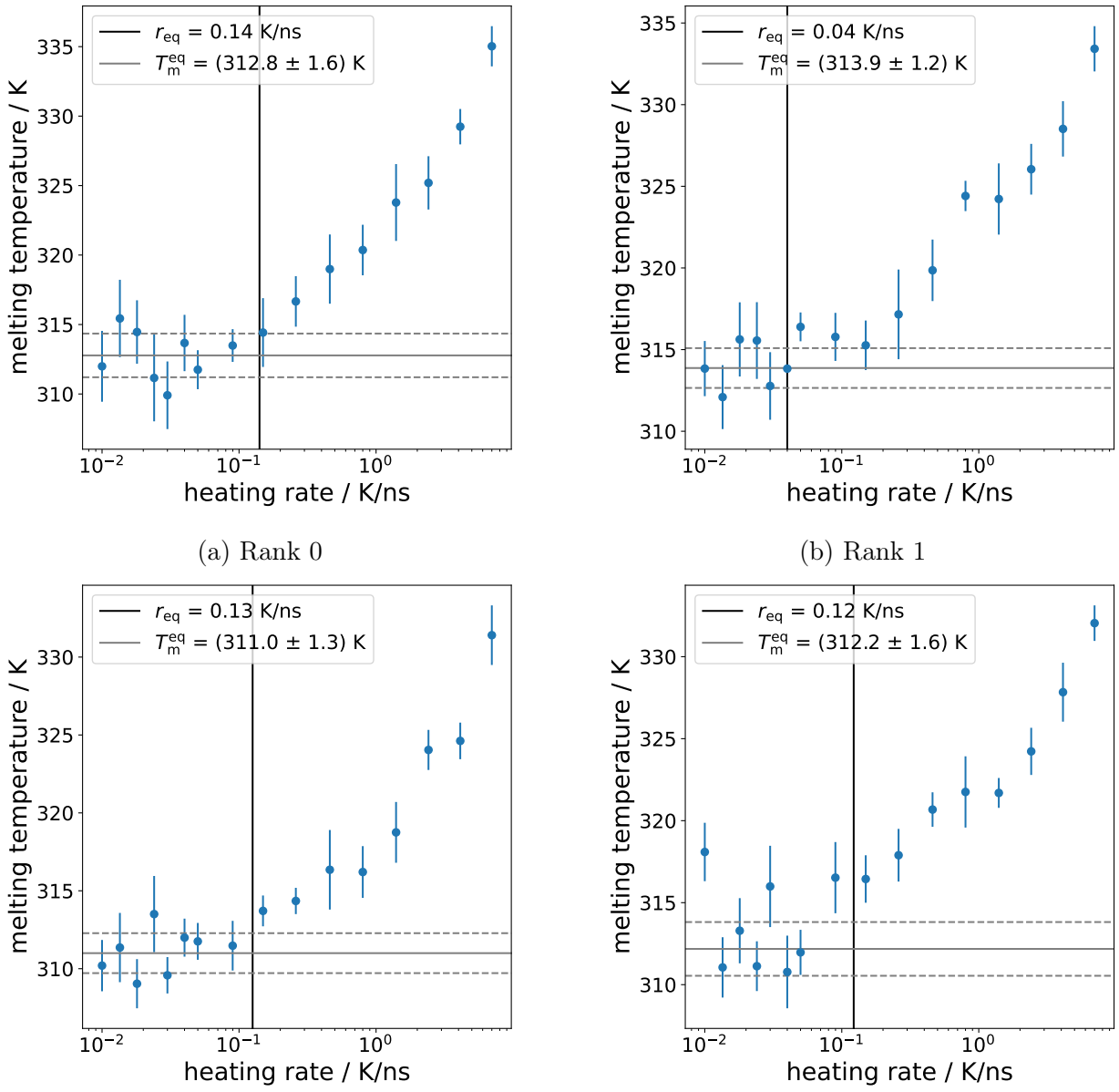


Figure S3: Examples of fit model selection in the biphasic approach to estimate T_m.

1.1 Melting temperature validation of 4 best candidate solutions



(a) Rank 0

(b) Rank 1

(c) Rank 2

(d) Rank 3

Figure S4: T_m of the 4 best candidate solutions with the reversible melting approach.

2 Noise

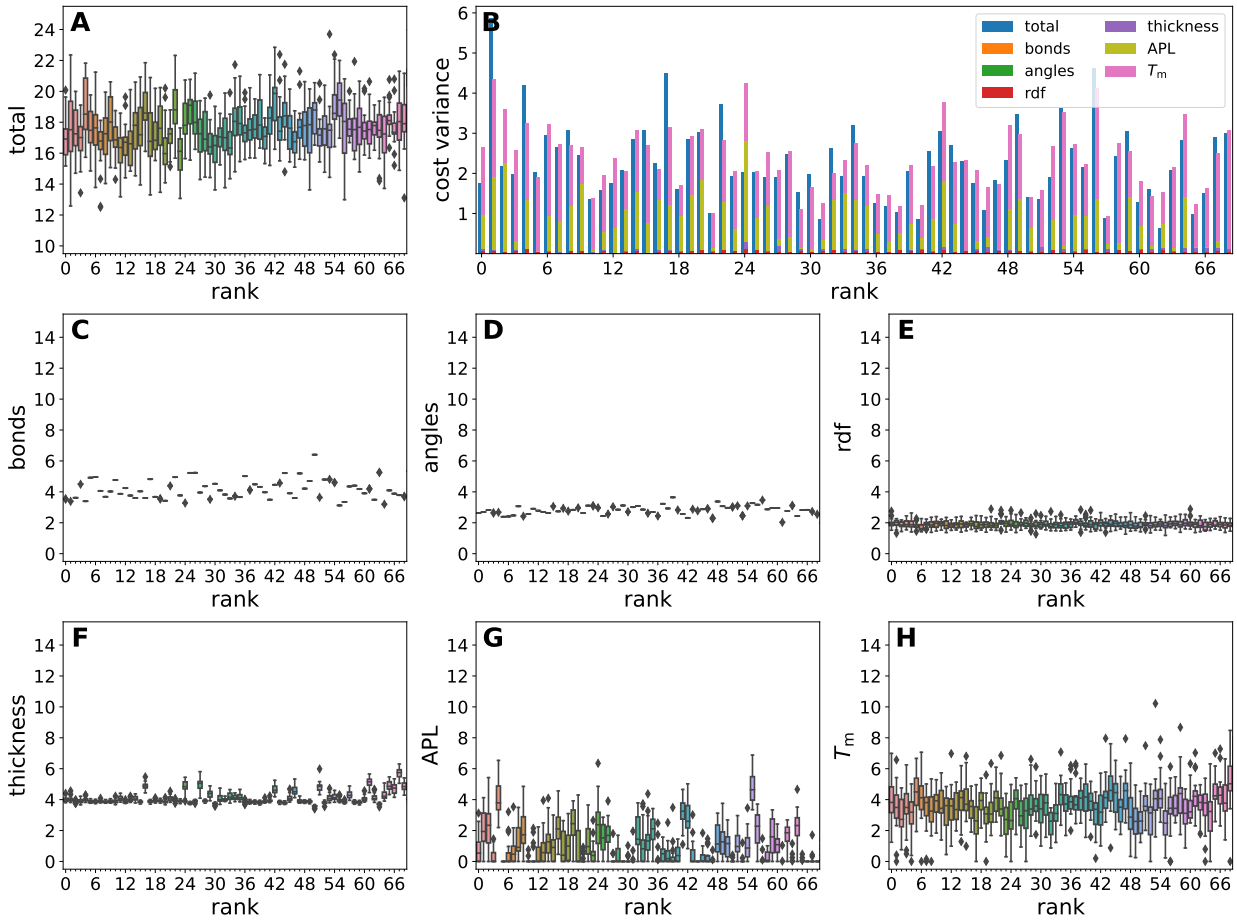
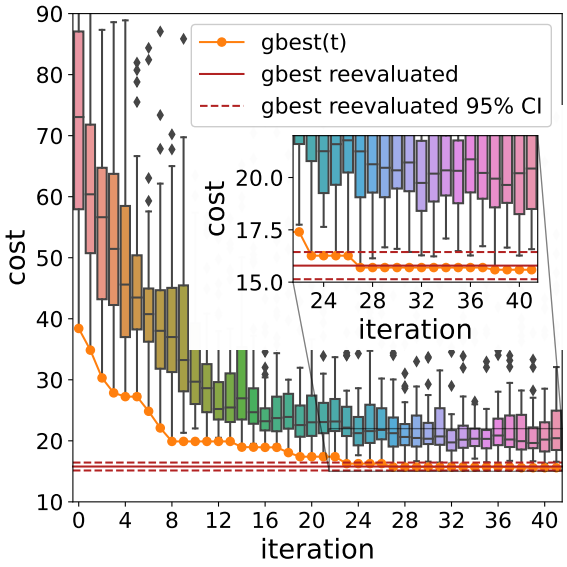
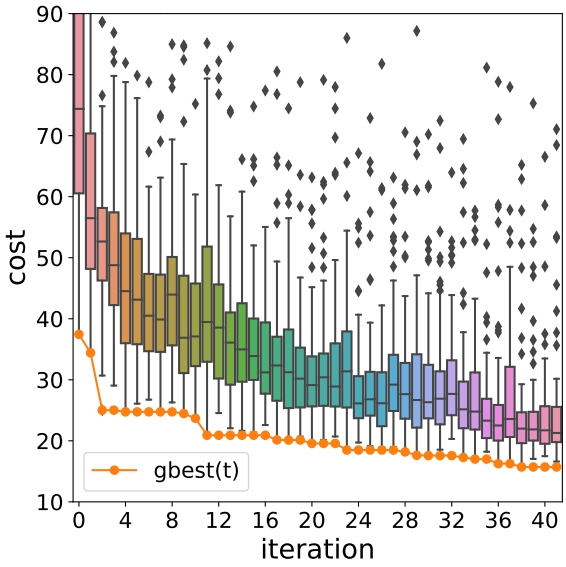


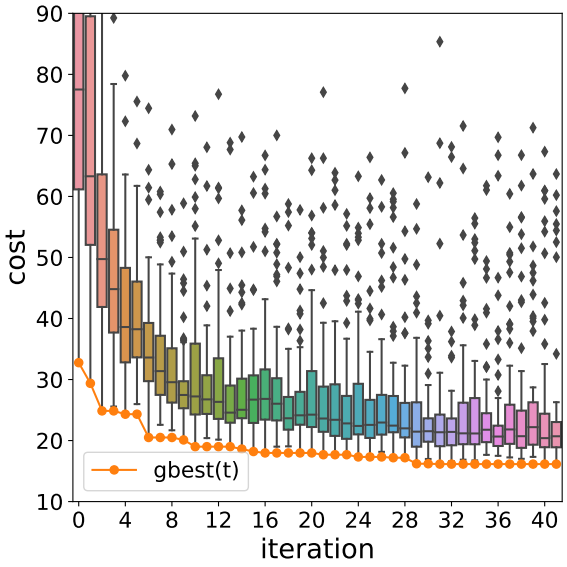
Figure S5: Noise levels of individual observables.



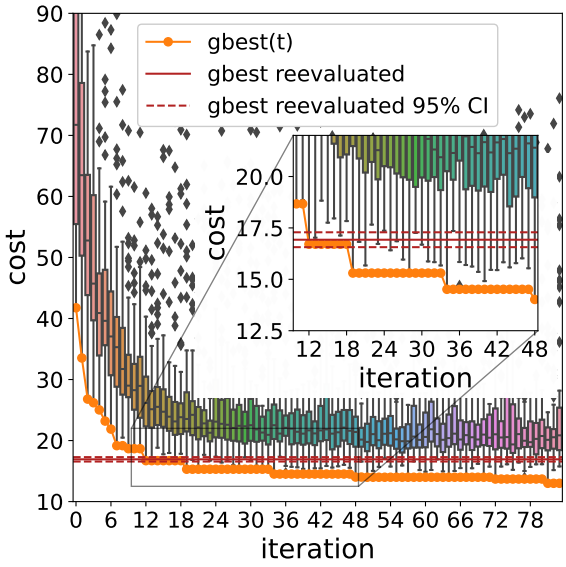
(a) 16 particles resampled 1+12 Tm samples.



(b) 32 particles resampled 1+6 Tm samples.



(c) 64 particles resampled 1+3 Tm samples.



(d) No resampling during optimization. But twice as many iterations

Figure S6: Comparison of cost evolution during optimization with and without noise-mitigation through resampling.

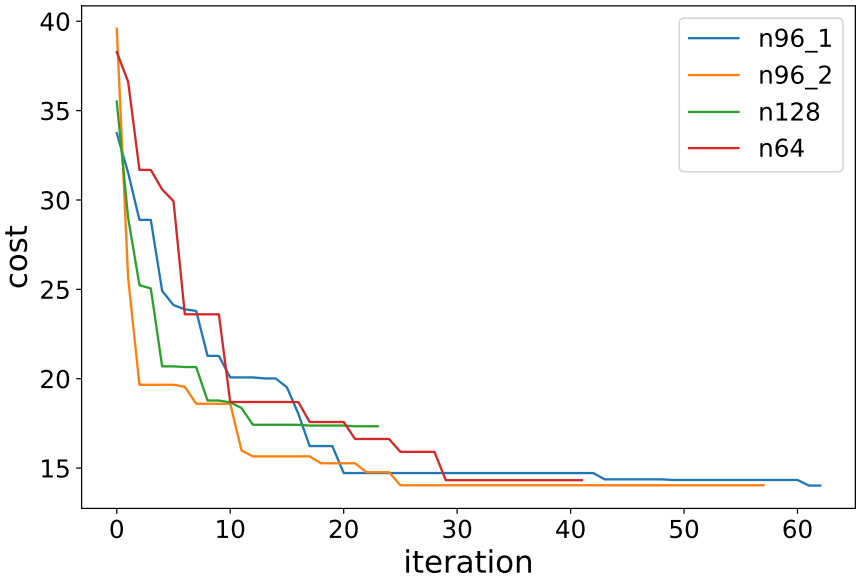


Figure S7: Evolution of the cost of gbest(t) with different swarm sizes.

3 DPSM-CHOL 2d center-of-mass radial distribution function

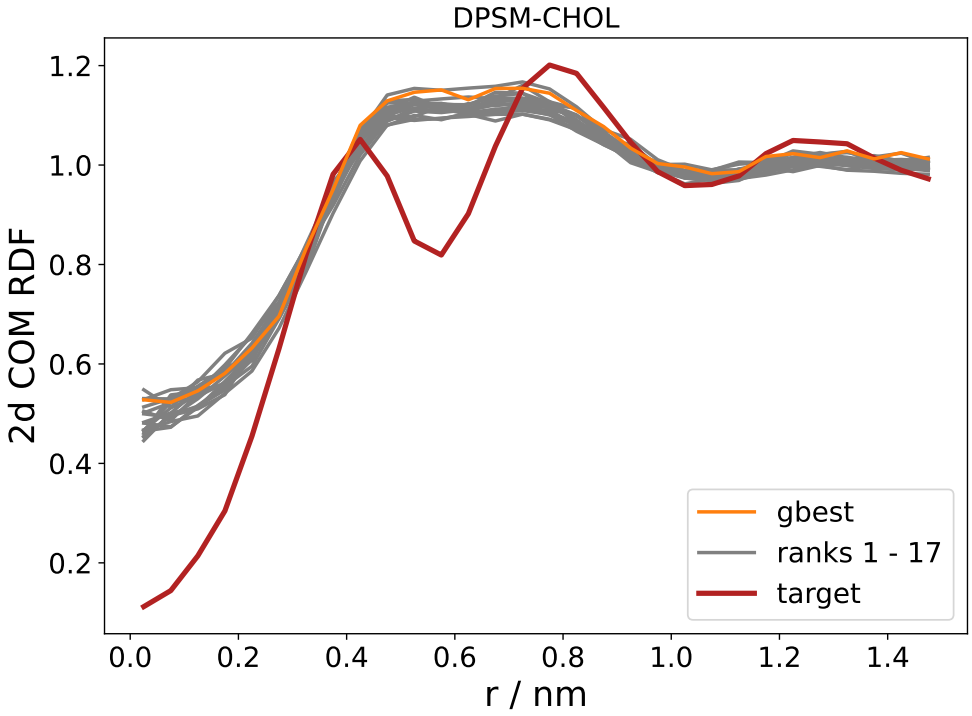


Figure S8: 2d COM radial distribution function (RDF). The distance r is measured in the x-y plane, i.e., parallel to the membrane. The 2d RDF is calculated per leaflet and averaged.

4 DPSM topology for Gromacs

```
[ moleculetype ]
; molname    nrexcl
DPSM      1

[ atoms ]
; id     type   resnr  resname atomname cgnr charge
   1       Q1       1   DPSM    NC3      1   1.0
   2       Q5       1   DPSM    PO4      2  -1.0
   3      SP2       1   DPSM    AM1      3   0.0
   4       P1       1   DPSM    AM2      4   0.0
   5       C1       1   DPSM    T1A      5   0.0
   6       C1       1   DPSM    C2A      6   0.0
   7       C1       1   DPSM    C3A      7   0.0
   8       C1       1   DPSM    C1B      8   0.0
   9       C1       1   DPSM    C2B      9   0.0
  10       C1       1   DPSM    C3B     10   0.0
  11       C1       1   DPSM    C4B     11   0.0

[ bonds ]
;   i    j    funct   r0   fc
    1     2   1   0.40000   7000
    2     3   1   0.33632   8207
    3     4   1   0.29241   6909
    3     5   1   0.50190   5239
    5     6   1   0.47000   3800
    6     7   1   0.47000   3800
    4     8   1   0.34964   4483
    8     9   1   0.47000   3800
    9    10   1   0.47000   3800
   10    11   1   0.47000   3800

[ angles ]
; i   j   k   funct   theta0  fc
    2     3     4   2   147.805  81.78
    2     3     5   2   173.718  88.64
    3     5     6   2   180.000  45.16
    5     6     7   2   180.000     35
    4     8     9   2   180.000  96.23
    8     9    10   2   180.000     35
    9    10    11   2   180.000     35
```

References

(S1) Coppock, P. S.; Kindt, J. T. Determination of Phase Transition Temperatures for Atomistic Models of Lipids from Temperature-Dependent Stripe Domain Growth Kinetics. *J. Phys. Chem. B* **2010**, *114*, 11468–11473.

(S2) Carpenter, T. S.; López, C. A.; Neale, C.; Montour, C.; Ingólfsson, H. I.; Natale, F. D.; Lightstone, F. C.; Gnanakaran, S. Capturing Phase Behavior of Ternary Lipid Mixtures with a Refined Martini Coarse-Grained Force Field. *J. Chem. Theory Comput.* **2018**, *14*, 6050–6062.

(S3) Kowalik, B.; Schubert, T.; Wada, H.; Tanaka, M.; Netz, R. R.; Schneck, E. Combination of MD Simulations with Two-State Kinetic Rate Modeling Elucidates the Chain Melting Transition of Phospholipid Bilayers for Different Hydration Levels. *J. Phys. Chem. B* **2015**, *119*, 14157–14167.

(S4) Akaike, H. Information Theory and an Extension of the Maximum Likelihood Principle. *Proc. of the Second Internat. Symp. on Information Theory, edited by B. N. Petrov and S. Caski.* **1973**, 267–281.

(S5) Akaike, H. A new look at the statistical model identification. *IEEE Trans. Automat. Control* **1974**, *19*, 716–723.

(S6) Hurvich, C. M.; Tsai, C.-L. Regression and time series model selection in small samples. *Biometrika* **1989**, *76*, 297–307.

(S7) Sun, L.; Böckmann, R. A. Membrane phase transition during heating and cooling: molecular insight into reversible melting. *Eur. Biophys. J.* **2017**, *47*, 151–164.