JOURNAL OF
APPLIED
CRYSTALLOGRAPHY

Volume 53 (2020)

Supporting information for article:

Information gain from isotopic contrast variation in neutron reflectometry on protein–membrane complex structures

Frank Heinrich, Paul A. Kienzle, David P. Hoogerheide and Mathias Lösche

1. Implementation of multivariate normal distributions and Gaussian mixture models and basic performance testing

1.1. Gaussian approximations of PDFs

When using a Monte Carlo Markov Chain (MCMC) based global optimizer, the result of a model fit is a sample of an unnormalized joint posterior probability density function (PDF) $p(\theta|y) = p(\lambda, \varphi|y)$ of model parameters θ given the data y. In principle, the joint posterior PDF contains all information required to calculate full and marginal posterior entropies that are needed to determine the full and marginal information gain. However, due to the high dimensionality of the NR models, the sampling density is often insufficient to obtain those quantities from the sample of the posterior PDF, directly. Fortunately, assumptions about the shape of the posterior PDF can alleviate some of these problems. We found that distributions of parameter fit values in this work are sufficiently well described by Gaussian distributions, or linear combinations of thereof. We therefore use multivariate normal (MVN) and Gaussian mixture model (GMM) approximations in conjunction with the discrete MCMC sample of the posterior PDF to obtain posterior entropies.

A MVN approximation of the posterior PDF that uses a single multivariate Gaussian distribution takes the form:

$$p^{MVN}(\theta|y) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{\left(-\frac{1}{2}(\theta-\mu)^T\Sigma^{-1}(\theta-\mu)\right)} = \mathcal{N}(\theta, \mu, \Sigma)$$

μ is the mean of the sample of d-dimensional parameter vectors θ. $|\Sigma|$ denotes the determinant of the variance-covariance matrix Σ of θ. Both values can be defined in terms of an expectation value $\mathbb{E}$.

$$\mu = \mathbb{E}[\theta]$$

$$\Sigma = \mathbb{E}[(\theta-\mu)(\theta-\mu)^T] = \left[Cov\left[\theta_i, \theta_j\right], 0 < i, j < d\right]$$

The GMM approximation of the posterior PDF is a linear combination of k MVN distributions with weights ω_i:

$$p^{GMM}(\theta|y) = \sum_{i=1}^{k} \omega_i \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} e^{\left(-\frac{1}{2}(\theta-\mu_i)^T\Sigma_i^{-1}(\theta-\mu_i)\right)} = \sum_{i=1}^{n} \mu_i \mathcal{N}(\theta, \mu_i, \Sigma_i)$$

$$\sum_{i=1}^{k} \mu_i = 1$$

We typically use $k = 5\sqrt{d}$ weights, and a sub-sample size of $n = 10{,}000$ to compute a GMM approximation from a discrete sample of a PDF.

1.2. Entropy of the joint posterior PDF

The entropy of a continuous PDF $p(\theta|y)$ is given by:

$$H(\Theta|y) = -\int p(\theta|y) \log p(\theta|y)\, d\theta$$

In the case of the MCMC, an unnormalized, discrete sample of the PDF is obtained that requires a normalization factor of N to obtain absolute values of $p(\theta|y)$. An estimate of the entropy of such a PDF can be obtained via a MVN approximation by evaluating the determinant of the d-dimensional covariance matrix $|\Sigma|$ (Chen, Wang, Zhao, & Principe, 2016). Using this method, the normalization factor N does not need to be explicitly determined, as the MVN distribution is analytically normalized:

$$H^{MVN}(\Theta|y) = \frac{d}{2}\log 2\pi e + \frac{1}{2}\log|\Sigma|$$

In general, the entropy of a GMM distribution cannot be analytically obtained. A numerical approximation such as a Monte Carlo integration over a random sample of size n is required, instead:

$$H^{GMM}(\Theta|y) \approx \frac{1}{n}\sum_{i=1}^{n} \log p_i^{GMM}(\theta|y)$$

In previous work we used an entropy estimate proposed by Kramer et al. (KDN estimate) that retains the utilization of the unnormalized density values (log-likelihood values of the model fit) of every point in the MCMC sample for the entropy calculation (Kramer, Hasenauer, Allgöwer, & Radde, 2010). KDN uses a kernel density estimate (KDE) to globally obtain the normalization factor N. Entropy calculation is then realized as a Monte Carlo integration over n points from the MCMC posterior PDF, as the frequency of occurrence of a parameter vector θ in the MCMC sample is proportional to its probability density $p(\theta|y)$:

$$H^{KDN}(\Theta|y) = -\frac{1}{n}\sum_{i=1}^{n} \log\big(N\, p_i(\theta|y)\big) = -\,\mathbb{E}\big[\log\big(N\, p(\theta|y)\big)\big]$$

We found a good agreement between MVN and KDN with expected differences for non-normal PDFs but also observed that the KDE computation is less robust with respect to different samples from the posterior MCMC PDF. Occasional outliers were observed and eliminated during repeated entropy calculations.

1.3. Marginal entropies from the posterior PDF

The marginal entropy $H_\lambda(\Theta|y) = H(\Lambda|y)$ of a PDF with respect to the parameters of interest λ of the parameter vector $\theta = (\lambda, \delta)$ is defined as:

$$H_\lambda(\Theta|y) = H(\Lambda|y) \; = -\int p(\lambda|y) \log p(\lambda|y)\, d\lambda$$

The marginal PDF $p(\lambda|y)$ is generally computed by integrating the joint PDF $p(\lambda, \delta|y) = p(\theta|y)$ over the nuisance parameters δ:

$$p(\lambda|y) \; = \int p(\lambda, \delta|y)\, d\delta$$

This integral is easily obtained for the MVN estimate of the joint PDF. μ and Σ of the joint PDF are decomposed into contributions from the parameters of interest λ and nuisance parameters δ:

$$\mu = (\mu_\lambda, \mu_\delta)$$

$$\Sigma = \begin{pmatrix} \Sigma_{\lambda\lambda} & \Sigma_{\lambda\delta} \\ \Sigma_{\delta\lambda} & \Sigma_{\delta\delta} \end{pmatrix}$$

Then, the marginal PDF becomes:

$$p^{MVN}(\lambda|y) = \int p^{MVN}(\lambda, \delta|y)\, d\delta = \; \mathcal{N}(\lambda, \mu_\lambda, \Sigma_{\lambda\lambda})$$

The marginal entropy follows as:

$$H_\lambda^{MVN}(\Theta|y) = \frac{d_\lambda}{2} \log 2\pi e + \frac{1}{2} \log |\Sigma_{\lambda\lambda}|$$

d_λ constitutes the length of the sub-vector of parameters of interest. The marginal MVN entropy is computed from the MVN estimate of the joint PDF by dropping the rows and columns from Σ that are associated with nuisance parameters.

This approach is trivially extended to the PDF in the GMM approximation $p^{GMM}(\lambda|y)$ by dropping the rows and columns from the joint PDF $p^{GMM}(\theta|y)$ associated with nuisance parameters for all mean vectors and covariance matrices:

$$p^{GMM}(\lambda|y) = \sum_{i=1}^{n} \omega_i \, \mathcal{N}\left(\lambda, \mu_\lambda^i, \Sigma_{\lambda\lambda}^i\right)$$

$H_\lambda^{GMM}(\Theta|y)$ of is then calculated identically to the joint posterior PDF using a Monte Carlo sum over $p^{GMM}(\lambda|y)$.

If one wishes to make direct use of the probability densities obtained by the MCMC optimizer, a sampling scheme similar to the KDN estimate can be obtained starting with the conditional entropy:

$$\begin{aligned} H_\lambda(\Theta|y) &= H(\Lambda,\Phi|y) - H(\Phi|\Lambda,y) \\ &= H(\Theta|y) - H(\Phi|\Lambda,y) \end{aligned}$$

with

$$\begin{aligned} H(\Phi|\Lambda,y) &= -\int p(\phi|\lambda,y)\log p(\phi|\lambda,y)\,d\phi\; p(\lambda|y)\,d\lambda \\ &= -\int \frac{p(\lambda,\phi|y)}{p(\lambda|y)}\log p(\phi|\lambda,y)\;p(\lambda|y)\,d\lambda d\phi \\ &= -\int p(\lambda,\phi|y)[\log p(\lambda,\phi|y) - \log p(\lambda|y)]\;d\lambda d\phi \\ &= -\int p(\theta|y)[\log p(\theta|y) - \log p(\lambda|y)]\;d\theta \\ &= H(\Theta|y) + \int p(\theta|y)[\log p(\lambda|y)]\;d\theta \end{aligned}$$

Thereby, one obtains

$$H_\lambda(\Theta|y) = -\int p(\theta|y)[\log p(\lambda|y)]\;d\theta$$

This means that $H_\lambda(\Theta|y)$ can be computed using Monte Carlo sampling of $\log p(\lambda|y)$ over the sample of the posterior PDF obtained by the MCMC optimizer. The marginal PDF $p(\lambda|y)$ at the sampling points has to be computed from the MVN or the GMM estimates as shown above, or alternatively, $p(\lambda|y)$ can be obtained as the ratio of $p(\theta|y)$ and $p(\phi|\lambda,y)$ using the formula for conditional probability:

$$p(\lambda|y) = \frac{p(\theta|y)}{p(\phi|\lambda,y)}$$

Following our previous work (Treece et al., 2019), we initially used a KDN estimate for $p(\theta|y)$ and computed a KDE estimate of the conditional PDF $p(\phi|\lambda,y)$ using the *statsmodels* Python package (Seabold & Perktold, 2010). We found that most NR models in this study were too complex (high-dimensional) to support robust computations of KDE estimates, while GMM estimates provided reliable and for practical purposes identical estimates. We therefore used a GMM estimate of $\log p(\lambda|y)$ in the Monte Carlo integration to calculate $H_\lambda(\Theta|y)$.

1.4. Comparison of different methods to calculate the marginal entropy

Fig. S1 shows marginal entropies calculated using the MVN and GMM methods for the model parameters describing the thickness, volume fraction, and nSLD of the porous layer in the test structure discussed in section 3.1 and shown in Fig. 1. The information gain from a single virtual NR measurement was determined under systematic variation of the values of ρ_n of the porous layer and the bulk solvent. Both methods yield quantitatively similar results, showing lowest information gain when the solvent nSLD matches that of the porous layer and was close to that of the Si substrate ($\rho_n = 2 \times 10^{-6}$ Å^{-2}). There from we concluded to report only GMM estimates throughout the paper.

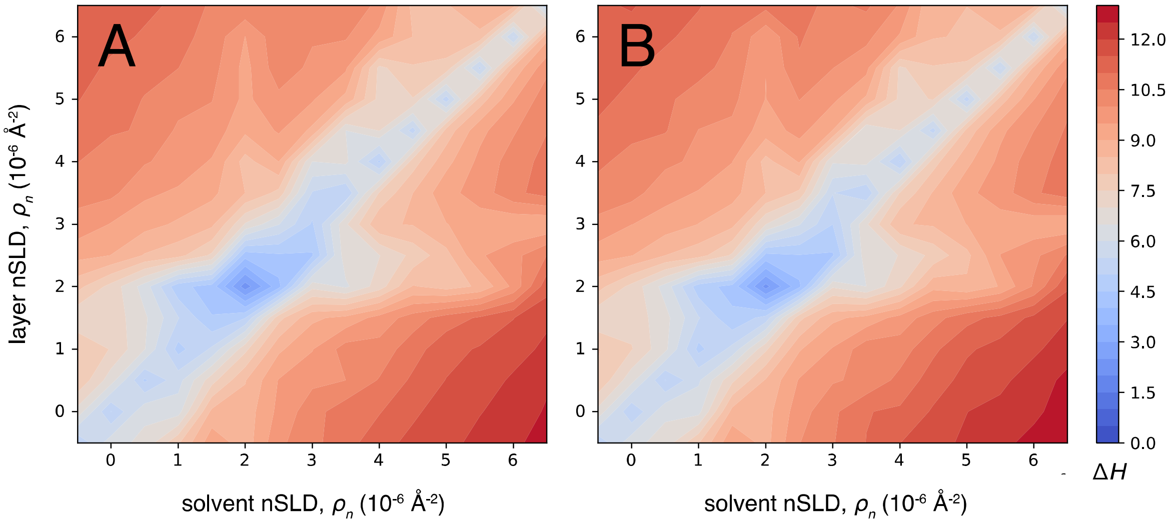

Figure S1 Information gain ΔH from a single NR measurement on the model structure in Figure 1 as a function of the nSLDs of the porous layer material and the aqueous solvent. nSLDs were varied in steps of 0.5×10^{-6} Å^{-2}. Entropies of the posterior were calculated using the marginal MVN (panel A) and marginal GMM (panel B) approximations.

1.5. Degenerate model parameters

To test the robustness of the implementation of the marginal entropy against degenerate parameters, we introduced a pair of highly correlated parameters by splitting the porous layer of the model structure shown in Fig. 1 into two sublayers, each having a thickness of one half of that of the original layer, which is 15 Å. The volume fraction and nSLD values of the two sublayers are shared and their nominal values are identical to those of the original single layer. Fit boundaries for the layer thicknesses of both sublayers are 15.0 Å ± 2.2 Å, ensuring that the prior entropy of the two layers is identical to the one of the single layer of the non-degenerate model with fit boundaries of 30 Å ± 10 Å. Per design, the thicknesses parameters of the porous layers in the degenerate model cannot be individually resolved, but are highly correlated while the total thickness of both layers and the uncertainty is the same as those obtained for the single porous layer in the non-degenerate model (see Figure S2).

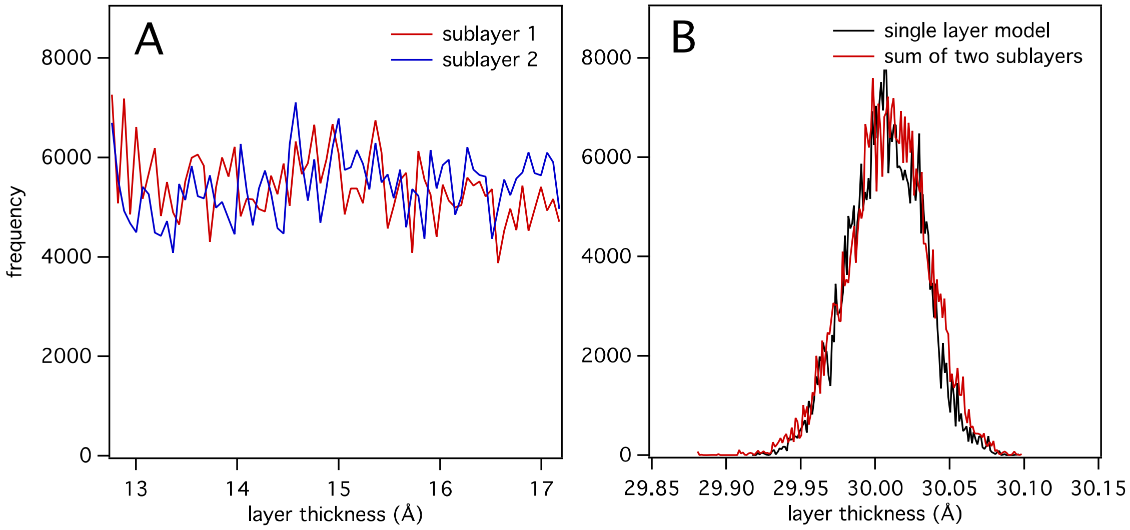

Figure S2 Fit results of a single NR measurement using one bulk solvent contrast of the original test structure (Fig. 1) and a degenerate model. (A) Thicknesses of the two layers with otherwise identical values for nSLD and volume fraction. (B) PDFs of the thickness of the single layer and the sum of the thicknesses of the two sublayers shown in panel A.

To illustrate the contribution of the layer thickness parameters to the information gain (one parameter for the original model, two parameters for the degenerate model), ΔH values were calculated for different combinations of parameters that describe the porous layer: (1) layer thickness(es), volume fraction and nSLD; (2) layer thickness(es) only; and (3) the volume fraction and nSLD only (see Fig. S3). As expected, the information gain from the non-degenerate and degenerate models trace each other very closely. A difference in ΔH of 2 bit is observed for combinations that include layer thickness parameters. The origin of this difference is related to inherent difficulties in computing the entropy of highly correlated parameters with broad individual PDFs.

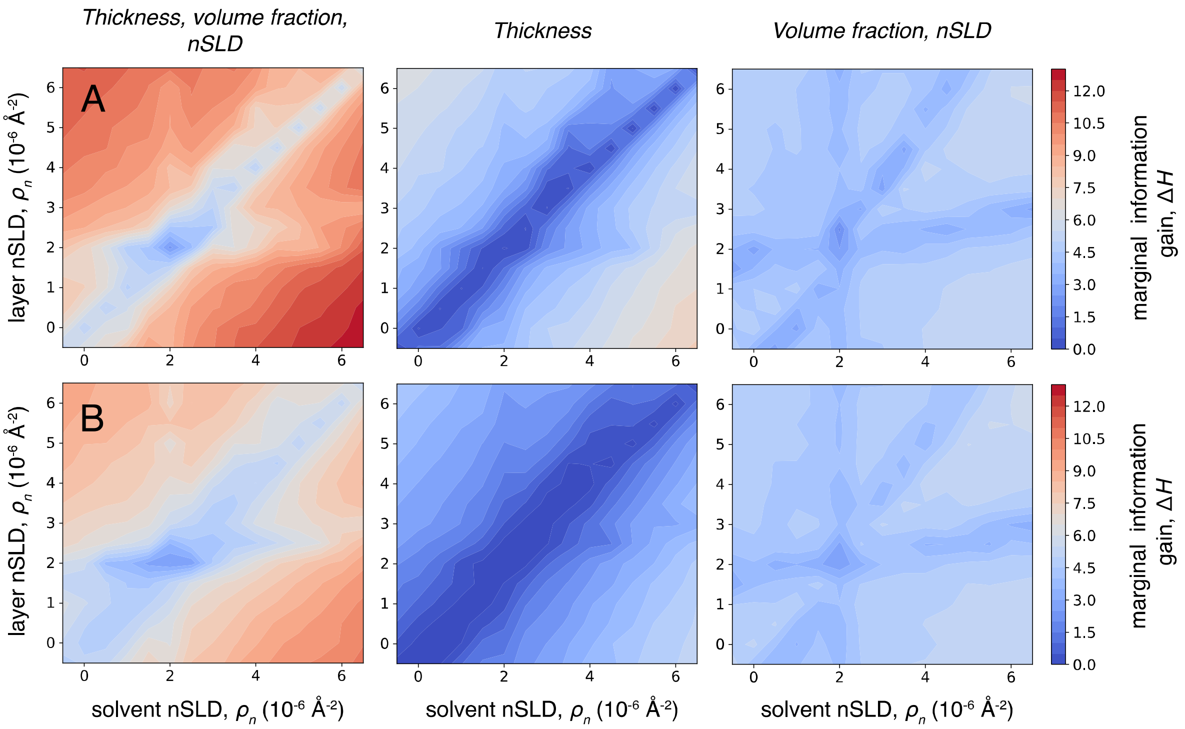

Figure S3 Marginal information gain ΔH from a virtual experiment using one bulk solvent contrast for the non-degenerate system (A) and the degenerate system (B). The columns refer to different combinations of parameters describing the porous layer for which ΔH was calculated.

2. Supplemental for optimizations

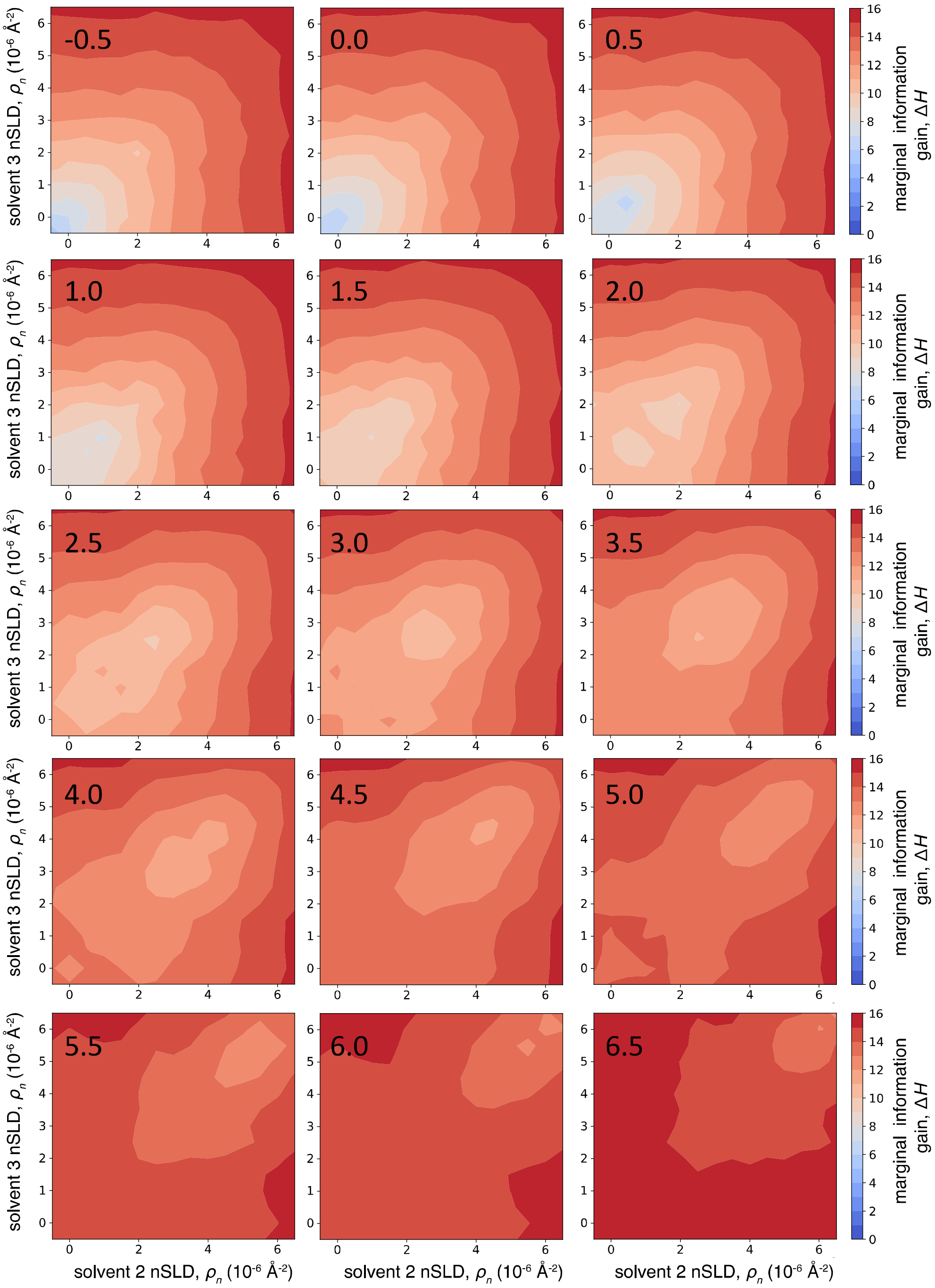

Figure S4 Marginal information gain on the thickness, volume fraction, and nSLD of a porous layer (Fig. 1) with an nSLD close to that of a lipid bilayer (ρ_n = -0.5 × 10^{-6} Å^{-2}) from an NR experiment consisting of three virtual measurements with systematically varied, independent bulk solvent nSLDs. The nSLD of one bulk solvent is given in the panels.

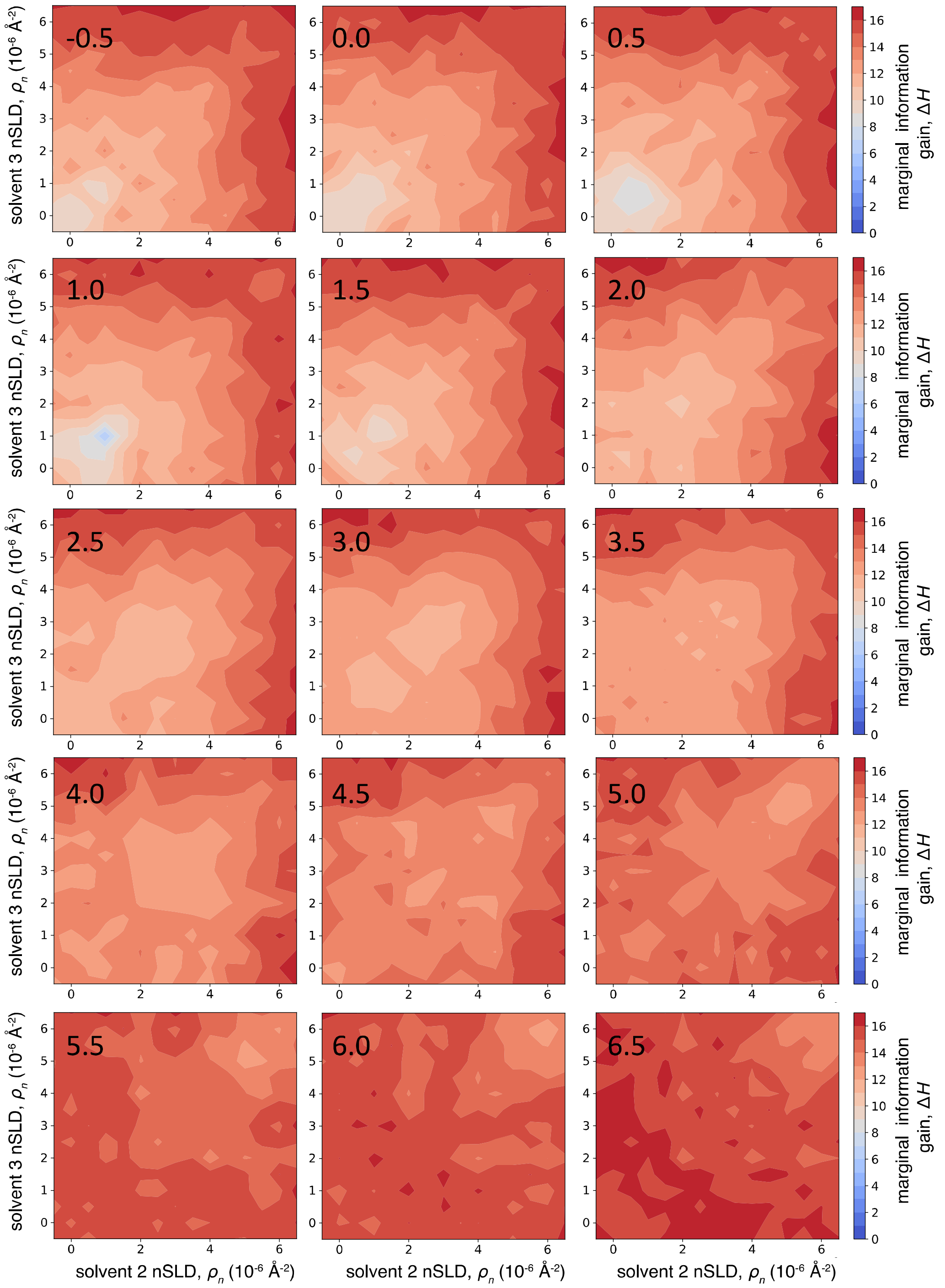

Figure S5 Marginal information gain on the thickness, membrane volume fraction, and fraction of DPPC-d$_{62}$ of the solid supported lipid bilayer shown in Fig. 3 for a particular nSLD value of the lipid hydrocarbon chains of ρ_n = -0.5 × 10^{-6} Å^{-2}. Three bulk solvent nSLDs in a virtual NR experiment were systematically varied.

References

Chen, B., Wang, J., Zhao, H., & Principe, J. C. (2016). Insights into Entropy as a Measure of Multivariate Variability. *Entropy*, *18*(5). http://doi.org/10.3390/e18050196

Kramer, A., Hasenauer, J., Allgöwer, F., & Radde, N. (2010). Computation of the posterior entropy in a Bayesian framework for parameter estimation in biological networks (pp. 493–498). Presented at the Control (MSC), IEEE. http://doi.org/10.1109/CCA.2010.5611198

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. *Proceedings of the 9th Python in Science Conference*, 57–61.

Treece, B. W., Kienzle, P. A., Hoogerheide, D. P., Majkrzak, C. F., Lösche, M., & Heinrich, F. (2019). Optimization of reflectometry experiments using information theory. *Journal of Applied Crystallography*, *52*(Pt 1), 47–59. http://doi.org/10.1107/S1600576718017016