ECEA 2019 5th International Electronic Conference on Entropy and Its Applications 18–30 November 2019

sciforum

*Conference Proceedings Paper*

# Systematic Coarse-Grained Models for Molecular Systems Using Entropy

**Evangelia Kalligiannaki [1]*, Vagelis Harmandaris [1,2]* and Markos Katsoulakis [3]**

[1] Institute of Applied and Computational Mathematics, Foundation for Research and Technology-Hellas, Greece; evangelia.kalligiannaki@iacm.forth.gr
[2] Department of Mathematics and Applied Mathematics, University of Crete & Institute of Applied and Computational Mathematics, Foundation for Research and Technology-Hellas, Greece; harman@uoc.gr
[3] Department of Mathematics and Statistics, University of Massachusetts, Amherst, USA; markos@math.umass.edu
* Correspondence: evangelia.kalligiannaki@iacm.forth.gr (E.K.); harman@uoc.gr (V.H.)

**Abstract:** The development of systematic coarse-grained mesoscopic models for complex molecular systems is an intense research area. Here we first give an overview of different methods for obtaining optimal parametrized coarse-grained models, starting from detailed atomistic representation for high dimensional molecular systems. We focus on methods based on information theory, such as relative entropy, showing that they provide parameterizations of coarse-grained models at equilibrium by minimizing a fitting functional over a parameter space. We also connect them with structural-based (inverse Boltzmann) and force matching methods. All the methods mentioned in principle are employed to approximate a many-body potential, the (n-body) potential of mean force, describing the equilibrium distribution of coarse-grained sites observed in simulations of atomically detailed models. We also present in a mathematically consistent way the entropy and force matching methods and their equivalence, which we derive for general nonlinear coarse-graining maps. We apply, and compare, the above-described methodologies in several molecular systems: a simple fluid (methane), water and a polymer (polyethylene) bulk system. Finally, for the latter we also provide reliable confidence intervals using a statistical analysis resampling technique, the bootstrap method.

**Keywords:** coarse-graining; data-driven; relative entropy; path-space; uncertainty quantification

## 1. Introduction

The enormous range of length and time scales involved in complex materials presents a challenging computational task, mainly, due to a wide range of relaxation times. A standard methodology to overcome problems of long relaxation times is to abandon the chemical detail and describe the molecular system by fewer degrees of freedom. Thus, systematic coarse-grained (CG) models are developed by averaging out the details at the molecular level, and by representing groups of atoms by a single CG particle. The challenge is to derive reliable coarse models both for reproducing the structural and the dynamical properties of systems. That is, to identify and effective approximate force field, approximating the potential of mean force (PMF), and then approximations to kinetic coefficients such as the friction.

Methods to approximate the PMF are well studied in the literature. Examples include: (a) The Boltzmann inversion methods, also known as structural-based, which rely on matching the radial distribution function. [1–6]. (b) The information theory based variational inference method relies on the minimization of the relative entropy (RE) between the configurational distributions of the system and the approximate one, [7–10]. (c) The Force Matching (FM) relies on minimizing the distance between the forces exerted on the CG particles and the approximate ones [11–13]. Recently, we have introduced

a path-space variational inference methods were introduced, capable of inferring dynamical models of coarse-grained systems, [9,14]. There the Relative Entropy Rate (RER) is defined as the appropriate quantity to infer the coarse dynamics for stationary system, while the path space force matching.

The purpose of the current work is to present a short review of the information theoretic methodologies ( relative entropy, and relative entropy rate) and their relation to the force matching and path-space force matching methodologies, through the application to different molecular systems.

## 2. Methodology

Let a prototypical problem of $N$ classical atoms in a box of volume $V$ at temperature $T$. We denote $\mathbf{q} = (\mathbf{q_1}, ..., \mathbf{q_N}) \in \mathbb{R}^{3N}$ the position vector and $\mathbf{p} = (\mathbf{p_1}, ..., \mathbf{p_N}) \in \mathbb{R}^{3N}$ the momentum vector of the $N$ atoms. The probability of an elementary configuration $\mathbf{q}$ is given by the Gibbs probability,

$$\mu(\mathbf{q}) = Z^{-1} \exp\{-\beta U(\mathbf{q})\}, \tag{1}$$

where $U(\mathbf{q})$ is potential energy of a state $\mathbf{q}$, $Z$is the normalization constant (partition function), and $\beta = \frac{1}{k_B T}$ with $k_B$ the Boltzmann constant and $T$ the temperature. In the above relation the kinetic part of the Hamiltonian has been integrated out. *Coarse-graining* (CG) is a standard methodology to overcomes the large range of length and time scales by averaging out the details of the atomistic level at the molecular level through representing groups of atoms by a single particle. The CG map $\Pi : \mathbb{R}^{3N} \to \mathbb{R}^{3M}$ determines the position vectors of $M$ CG particles (or beads) $\bar{\mathbf{q}} = (\bar{\mathbf{q}}_1, ..., \bar{\mathbf{q}}_M) \in \mathbb{R}^{3M}$. Note that $M < N$ but still $M >> 1$. From now on, we will use the bar "¯" notation for objects related to the CG model. The probability that the CG system has configuration $\bar{\mathbf{q}}$ is given by

$$\bar{\mu}(d\bar{\mathbf{q}}) = \int_{A(\bar{\mathbf{q}})} \mu(\mathbf{q}) d\mathbf{q} = Z^{-1} \int_{A(\bar{\mathbf{q}})} e^{-\beta U(\mathbf{q})} d\mathbf{q}, \;\; A(\bar{\mathbf{q}}) = \{\mathbf{q} \in \mathbb{R}^{3N} : \mathbf{\Pi}(\mathbf{q}) = \bar{\mathbf{q}}\}. \tag{2}$$

The quantity

$$\bar{U}^{PMF}(\bar{\mathbf{q}}) = -\frac{1}{\beta} \ln \int_{A(\bar{\mathbf{q}})} e^{-\beta U(\mathbf{q})} d\mathbf{q},$$

is the $M-$body potential of mean force (PMF). The corresponding conservative force is thus $\bar{\mathbf{F}}^{PMF}(\bar{\mathbf{q}}) = -\nabla \bar{U}^{PMF}(\bar{\mathbf{q}})$. Although the above formula is exact, the accurate calculation of the PMF, for a realistic model of a complex molecular system, is an extremely difficult task, due to the high dimensionality of the integral, and of the $M$ vector as well. For this reason, we develop methods in order to find an effective potential in a parameterized form, $\bar{U}_{eff}(\bar{\mathbf{q}}; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, which best approximates the PMF, i.e.:

$$\bar{U}_{eff}(\bar{\mathbf{q}}; \boldsymbol{\theta}) \approx \bar{U}^{PMF}(\bar{\mathbf{q}}).$$

Moreover, we assume that the evolution of the particles is described by a continuous time process $\{X_t\}_{t \geq 0} = (\mathbf{q}_t, \mathbf{p}_t)_{t \geq 0}$, with path space distribution $P_{[0,T]}$, and invariant measure the Gibbs probability (2). The approximate coarse space dynamics we adopt are described by a Markov process $\{\bar{X}_t\}_{t \geq 0}$ in $\mathbb{R}^m$ with a parametric path space distribution $\bar{Q}^{\boldsymbol{\theta}}_{[0,T]}$, $\boldsymbol{\theta} \in \tilde{\Theta}$.

### 2.1. Information Theoretic Variational Inference: The Relative Entropy

Here we adopt the *information theoretic variational inference* approach as the methodology to derive optimal approximate coarse models bot at equilibrium and dynamical regimes. This variational approach encompasses the minimization of the Relative Entropy (RE) between probability measures. The relative entropy (Kullback-Leibler divergence), [15], of two probability measures $P(d\omega)$ and $Q(d\omega)$ on a common measurable space $(\Omega, \mathcal{B})$ is given by

$$\mathcal{R}(P|Q) = \int_\Omega \log \frac{dP(\omega)}{dQ(\omega)} P(d\omega), \tag{3}$$

provided $P \ll Q$, i.e., $P$ is absolutely continuous with respect to $Q$, and $\mathcal{R}(P|Q) = +\infty$ otherwise. The functional $\mathcal{R}(P|Q)$ defines a pseudo-distance between two measures as $\mathcal{R}(P|Q) \geq 0$ and $\mathcal{R}(P|Q) = 0$ if and only if $P = Q$, $P$-a.s. In the case these probability measures have corresponding probability densities $p(\omega)$ and $q(\omega)$ relation (3) becomes $\mathcal{R}(P|Q) = \int_\Omega \log \frac{p(\omega)}{q(\omega)}\, p(\omega) d\omega$. The optimization problem in path-space is,

$$\min_{\boldsymbol{\theta}\in\Theta} \mathcal{R}\left(\Pi_* P_{[0,T]} | \bar{Q}^{\boldsymbol{\theta}}_{[0,T]}\right), \tag{4}$$

where $\Pi_*\mu$ denotes the push-forward of the microscopic measure $\mu$. When the system is at equilibrium the optimization principle is

$$\min_{\boldsymbol{\theta}\in\Theta} \mathcal{R}\left(\Pi_*\mu | \bar{\mu}^{\boldsymbol{\theta}}\right).$$

When considering continuous time observations, in work [14] we prove that the path-space minimization principle (4) reduces to the path-space force matching (PSFM). In stationary dynamics the Relative Entropy Rate (RER) is the

$$\mathcal{H}(P \,|\, Q) = \lim_{T\to\infty} \frac{1}{T} \mathcal{R}\left(\Pi_* P | Q^{\boldsymbol{\theta}}\right), \tag{5}$$

where $P$ and $Q$ denote the corresponding stationary processes.

For discrete time observations **(a)** from the microscopic Gibbs density at $\mathcal{D}_{n_t} = \{X_1, \ldots, X_{n_t}\}$, and **(b)** the path-space distribution $P_{[0,T]}$ at dynamical regimes, $\mathcal{D}_{n_s,n_t} = \{X_1^k, \ldots, X_{n_t}^k\}_{k=1}^{n_s}$, encountering the estimator for the relative entropy the optimal parameter estimate,[14], is given by

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \sum_{k=1}^{n_s} \sum_{i=1}^{n_t} \log \frac{\bar{p}(\Pi X_i^k, \Pi X_{i+1}^k)}{\bar{q}^{\boldsymbol{\theta}}(\Pi X_i^k, \Pi X_{i+1}^k)}. \tag{6}$$

$\bar{p}$ and $\bar{q}^{\boldsymbol{\theta}}$ are microscopic and coarse space transition probability densities of the Markov processes $X_t$ and $\bar{X}_t$ respectively. Note that if time series are stationary, the RER optimization is

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{n_t-1} \log q^{\boldsymbol{\theta}}(\Pi X_i, \Pi X_{i+1}) \tag{7}$$

### *2.2. Relative Entropy and Force-Matching*

The Force-Matching (FM) method estimates an effective CG potential that reproduces best the potential at the reference all-atom system, by solving the optimization problem

$$\min_{\boldsymbol{\theta}} \mathbb{E}_\mu \left[ \|\mathbf{F}(\mathbf{q}) - \bar{\mathbf{F}}(\Pi(\mathbf{q}); \boldsymbol{\theta})\|^2 \right], \tag{8}$$

i.e., we minimize the average difference between the atomistic $\mathbf{F}(\mathbf{q})$ forces and the corresponding CG forces $\bar{\mathbf{F}}(\Pi(\mathbf{q}); \boldsymbol{\theta})$, where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^{3M}$ and $\mathbb{E}_\mu[\cdot]$ averages with respect to the probability Gibbs measure $\mu(d\mathbf{q})$. The minimization problem for the discrete observations, and a linear parametric representation of the force $\bar{F}(\cdot; \boldsymbol{\theta}) = G(\cdot)\boldsymbol{\theta}$,

$$\boldsymbol{\theta}^\star = \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta} \frac{1}{3M} \frac{1}{n_t} \sum_{l=1}^{n_t} \sum_{I=1}^{M} \left\| \mathbf{F}_I(\mathbf{q}_l) - \sum_{d=1}^{N_d} \theta_d \mathbf{G}_{I;d}(\Pi(\mathbf{q}_l)) \right\|^2. \tag{9}$$

The path-space Force matching optimization problem is, [14],

$$\boldsymbol{\theta}^*(T) = \operatorname*{argmin}_{\boldsymbol{\theta}} \mathbb{E}_{P_{[0,T]}} \left[ \frac{1}{2\sigma^2} \int_0^T \left\| \Pi_{\mathbf{p}} \mathbf{f}(\mathbf{q}_s) - \bar{\mathbf{F}}(\Pi_{\mathbf{q}} \mathbf{q}_s; \boldsymbol{\theta}) \right\|^2 ds \right].$$

for which the discrete optimization problem becomes

$$\hat{\boldsymbol{\theta}}^*(T) = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{3M}\frac{1}{n_p}\frac{1}{n_t}\sum_{l=1}^{n_t}\sum_{I=1}^{M}\sum_{n=1}^{n_p}\left\|\mathbf{F}_I(\mathbf{q}_{l,n}) - \sum_{d=1}^{N_d}\theta_d \mathbf{G}_{I;d}(\blacksquare_{\mathbf{q}}(\mathbf{q}_{l,n}))\right\|^2 .$$

*2.3. Relative Entropy and Structural-based Methods*

The structural-based methods, (Direct inverse Boltzmann, DBI, Iterative Boltzmann Inversion, and Inverse Monte Carlo (IMC)) methods use the pair correlation function $g^{(2)}(\bar{q})$ and the assumption that the interactions depend only on the distance $R$ between particles, that is $g^{(2)}(\bar{q}) = \bar{g}(R)$. $\bar{g}(R)$ is called the radial distribution function. Thus the CG effective interaction is given by

$$\bar{U}_{\text{eff}}(R) = -\frac{1}{\beta}\log \bar{g}(R)\,, \tag{10}$$

where

$$\bar{g}(R) = \frac{(M-1)M}{\rho^2}\int_{\{x:\Pi(x)=Q\}} \mathbf{1}_{B(Q_2,r)}(Q_1)\mu(x)dx\,,$$

that is the average density of finding the CG particle 1 at a distance $R$ from the particle 2.

The structural methods are thus based on the pair correlation function between CG particle, in contrast to the RE which is considering the total joint probability distribution of the CG particle. In case the PMF can be exactly described by pair functions the the RE and structural methods coincide.

## 3. Results and Discussion

In the current section, we present the application of the variational inference methods, RE and FM, for a few representative molecular systems: a simple fluid (bulk methane), a system of water molecules and a polyethylene melt, at equilibrium conditions. We moreover study thebulk methane system out-of equilibrium, specifically we apply the PSFM at a transient time regime.

*3.1. Bulk Methane*

The molecular system consists of 666 methane molecules at temperature $T = 100\,K$. We employed molecu;lar dynamics simulations to generate the microscopic space data for which we applied the inference methods. Details on the atomistic simulations are given in [16]. For the coarse-grained representation of methane we have used a one-site representation with a pair potential. The pair potentials we have tested are (a) expansions with linear and cubic B-splines, with 48 parameters, and (b) Lennard-Jones parametric form, with two parameters.

A comparison of the RE, FM, and IBI methods, is depicted in Figure 1, [17]. The result depicts slight difference of the FM method to the RE and IBI.Figure 2 presents the performance of the FM and PSFM methods at equilibrium verifying the validity of the PSFM and its reduction to the FM method. A study at transient time regimes is presented in work [16].

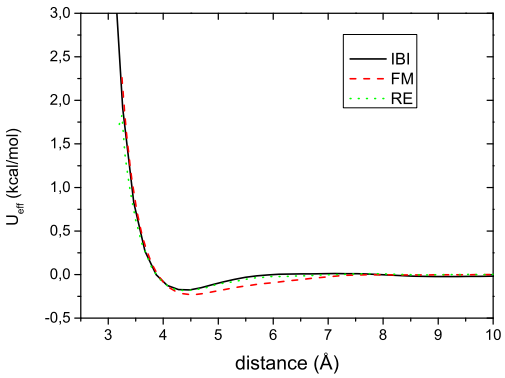


**Figure 1.** Methane: The effective pair potential for a one-site methane melt, derived with the RER, FM, and IBI methods.

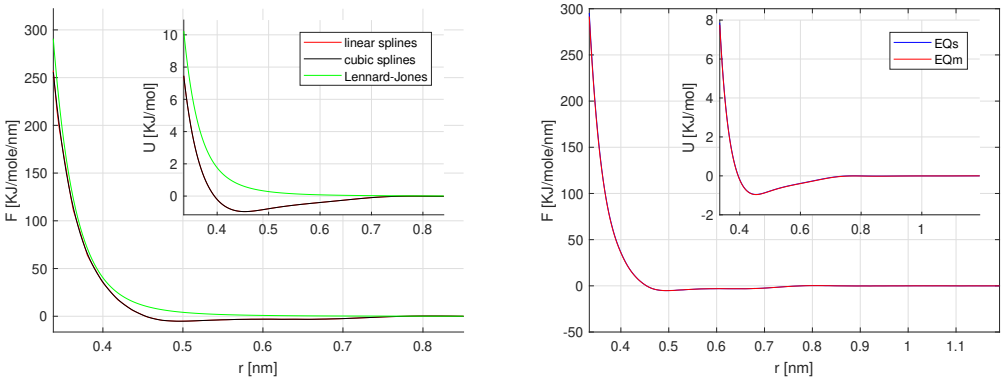


**Figure 2.** Methane: The FM and PSFM methods at equilibrium. (a)The FM pair force with linear and cubic B-splines, and Lennard-Jones parametrizations. (b)The PSFM reproduces the FM method.

### *3.2. Water*

The model system consists of 1192 molecules at ambient conditions (T $=$ 300 K, $P$ $=$ 1 atm). Details on the atomistic simulations are given in [17]. For the coarse-grained representation of $H_2O$, we have also used a one-site representation with a pair potential. Figure 3 a depicts the resulting pair potential obtained with the RE and FM methods. The RE and FM potentials have a very similar structure with two minima, though the actual values of the potential are different. In Figure 3b shows that the pair correlation function derived by CG simulations with the RE potential and the target one (from atomistic simulations) are very close, that the RE potential can reproduce with sufficient accuracy the pair correlation.

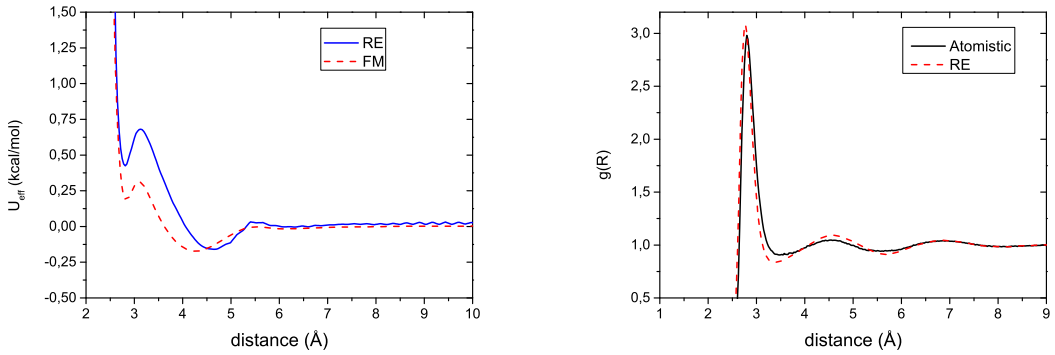


**Figure 3.** Water: (a) The effective pair potential with RE and FM (b) RE derived potential reproduces well the target pair correlation

### *3.3. Polyethylene Melt*

The model system consists of 96 polyethylene chains of 99 monomer units ($-CH_2-$), i.e., $N = 9504$. The simulations were performed under NVT conditions at temperature $T = 450$ K. For the coarse-grained representation we consider a $3:1$ mapping representation, i.e. three monomer units form one CG particle. With this application we study the effect of the size of the available observations (system configurations), and quantify uncertainties due to the small number of observations. Figure 4a depicts the derived FM potential for a large set of observations. In addition, Figure 4b shows the 95% confidence set obtained with a statistical analysis resampling technique (bootstrap method) of a small observations set, which captures the large-set outcome.

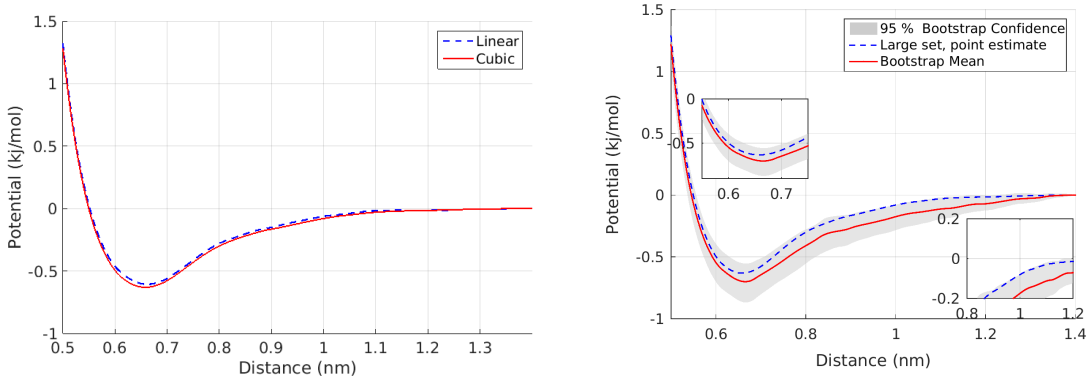


**Figure 4.** Polyethylene: (a)The FM potential for linear and cubic B-splines, for a set of 2000 observations (b) 95% Bootstrap Confidence interval for the FM potential, with a set of 200 observations and cubic B-splines.

## 4. Conclusions

In the current work we presented a short review of the information theoretic variational inference method for coarse-graining molecular systems, for systems at- and out-of- equilibrium. Moreover, we presented the connection to the Force Matching method and its relation to the structural based methods. The application of all methods to the methane system shows that the RE and IBI methods give similar results while the FM differs slightly. While for the water model the RE and FM resulting potentials differ substantially, which is not surprising as we know that the two methods are equivalent only asymptotically. We verify the validity of the PSFM, i.e. deriving the piar potential using time-series data, as it gives the same results to the FM, i.e., with identically distributed data. Finally, with the application to the polyethylene system, we show that when the availability of observations is limited the bootstrapping method can provide reliable confidence intervals to the pair potential.

**Funding:** E.K. acknowledges support by the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat for Research and Technology (GSRT), under grant agreement No [52].

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Soper, A. Empirical potential Monte Carlo simulation of fluid structure. *Chemical Physics* **1996**, *202*, 295 – 306.
2. Lyubartsev, A.P.; Laaksonen, A. On the Reduction of Molecular Degrees of Freedom in Computer Simulations. Novel Methods in Soft Matter Simulations; Karttunen, M.; Lukkarinen, A.; Vattulainen, I., Eds., 2004, Vol. 640, *Lecture Notes in Physics, Berlin Springer Verlag*, pp. 219–244. doi:10.1007/b95265.
3. Tschöp, W.; Kremer, K.; Hahn, O.; Batoulis, J.; Bürger, T. Simulation of polymer melts. I. Coarse-graining procedure for polycarbonates. *Acta Polym.* **1998**, *49*, 61.
4. Müller-Plathe, F. Coarse-Graining in Polymer Simulation: From the Atomistic to the Mesoscopic Scale and Back. *ChemPhysChem* **2002**, *3*, 754–769.
5. Harmandaris, V.A.; Adhikari, N.P.; van der Vegt, N.F.A.; Kremer, K. Hierarchical Modeling of Polystyrene: From Atomistic to Coarse-Grained Simulations. *Macromolecules* **2006**, *39*, 6708.
6. Briels, W.J.; Akkermans, R.L.C. Coarse-grained interactions in polymer melts: a variational approach. *J. Chem. Phys.* **2001**, *115*, 6210.
7. Shell, M. The relative entropy is fundamental to multiscale and inverse thermodynamic problems. *The Journal of Chemical Physics* **2008**, *129*, –.
8. Chaimovich, A.; Shell, M.S. Anomalous waterlike behavior in spherically-symmetric water models optimized with the relative entropy. *Phys. Chem. Chem. Phys.* **2009**, *11*, 1901–1915. doi:10.1039/B818512C.
9. Katsoulakis, M.A.; Plechac, P. Information-theoretic tools for parametrized coarse-graining of non-equilibrium extended systems. *J. Chem. Phys.* **2013**, *139*, 4852–4863.
10. Kalligiannaki, E.; Harmandaris, V.; Katsoulakis, M.; Plecháč, P. The geometry of generalized force matching and related information metrics in coarse-graining of molecular systems. *The Journal of Chemical Physics* **2015**, *143*. doi:http://dx.doi.org/10.1063/1.4928857.
11. Izvekov, S.; Voth, G. Effective force field for liquid hydrogen fluoride from ab initio molecular dynamics simulation using the force-matching method. *The Journal of Physical Chemistry. B* **2005**, *109*, 6573–6586.
12. Noid, W.G.; Liu, P.; Wang, Y.; Chu, J.; G.S. Ayton S. Izvekov, H.A.; Voth, G. The multiscale coarse-graining method. II. Numerical implementation for coarse-grained molecular models. *The Journal of Chemical Physics* **2008**, *128*, 244115.
13. Rudzinski, J.; Noid, W. Coarse-graining entropy, forces, and structures. *The Journal of Chemical Physics* **2011**, *135*, 214101.
14. Harmandaris, V.; Kalligiannaki, E.; Katsoulakis, M.; Plecháč, P. Path-space variational inference for non-equilibrium coarse-grained systems. *Journal of Computational Physics* **2016**, *314*, 355–383.
15. Cover, T.; Thomas, J. *Elements of Information Theory*; John Wiley & Sons, 1991.
16. Baxevani, G.; Kalligiannaki, E.; Harmandaris, V. Study of the transient dynamics of coarse-grained molecular systems with the path-space force-matching method. *Procedia Computer Science* **2019**, *156*, 59 – 68. doi:https://doi.org/10.1016/j.procs.2019.08.180.
17. Kalligiannaki, E.; Chazirakis, A.; Tsourtis, A.; Katsoulakis, M.; Plecháč, P.; Harmandaris, V. Parametrizing coarse grained models for molecular systems at equilibrium. *The European Physical Journal Special Topics* **2016**, *225*, 1347–1372. doi:10.1140/epjst/e2016-60145-x.