Learning Coarse-Grained Particle Latent Space with Auto-Encoders

Wujie Wang and Rafael Gómez-Bombarelli *
Department of Materials Science and Engineering
Massachusets Institute of Technology
77 Massachusetts Avenue, Cambridge, MA 02319
{wwj, rafagb}@mit.edu

Abstract

Molecular dynamics simulations provide theoretical insight into the microscopic behavior of materials in condensed phase and, as a predictive tool, enable the computational design of new compounds. However, because of the large temporal and spatial scales of physical processes in materials, atomistic simulations are often computationally infeasible to predict phenomena at long time-scale. Coarse-graining methods allow simulating larger systems, by reducing the dimensionality of the simulation, propagating longer timesteps, and averaging out fast motions. We propose a generative modeling framework based on auto-encoders to unify the tasks of learning discrete coarse-grained variables and decoding back to atomistic details.

1 Introduction

Coarse-Grained (CG) molecular modeling has been used extensively to simulate complex molecular processes at a lower cost than all-atom simulations [1, 2]. By compressing the full atomistic model into a reduced number of pseudo atoms, CG methods focus on the slow collective atomic motions and average out fast local motions. The use of structure-based coarse-grained strategy enabled important theoretical insights to probe length scales that are otherwise inaccessible in studying polymer dynamics [3–5] and lipid membranes [6].Fitting such structure-based coarse-grained potentials have been studied extensively[7] and recently attempted by numerous machine learning efforts[8–12]. Beyond parameterizing accurate CG potentials given a pre-defined mapping, the choice of all-atom to CG mapping plays an important role in recovering consistent CG dynamics, structural correlation and thermodynamics[13, 7] due to the loss of atomistic sub-ensemble and how they coupled to the coarse-grained variables. Thus, there has been a gap to reversibly bridge information hierarchy between simulations at different scales.

We propose to use unsupervised learning to optimize CG representations from atomistic simulations. As a powerful unsupervised learning technique, variational auto-encoders (VAEs) compress data through an information bottleneck [14] that continuously maps an otherwise complex data set into a low dimensional space. auto-encoder-based models have been used to learn latent representation of molecular configurational space [15, 16]. Compared to continuously parameterized latent space given simple statistical priors, coarse-grained coordinates, as discrete latent variables encoded in 3D space, need specially designed parameterization to maintain its Hamiltonian structure for discrete particle dynamics. Inspired generative modeling perspective, we propose an unsupervised Auto-Encoder based model to learn discrete coarse-grained latent variables for molecular configurational space.

Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019), Vancouver, Canada.

2 Theory

The essential idea of Coarse-Grained Auto-Encoders is to treat coarse-grained coordinates as latent variables that are the most predictive of atomistic distributions while having a smooth underlying free energy landscape. We show that this is achieved by minimizing the reconstruction loss and the instantaneous mean force regularizer.

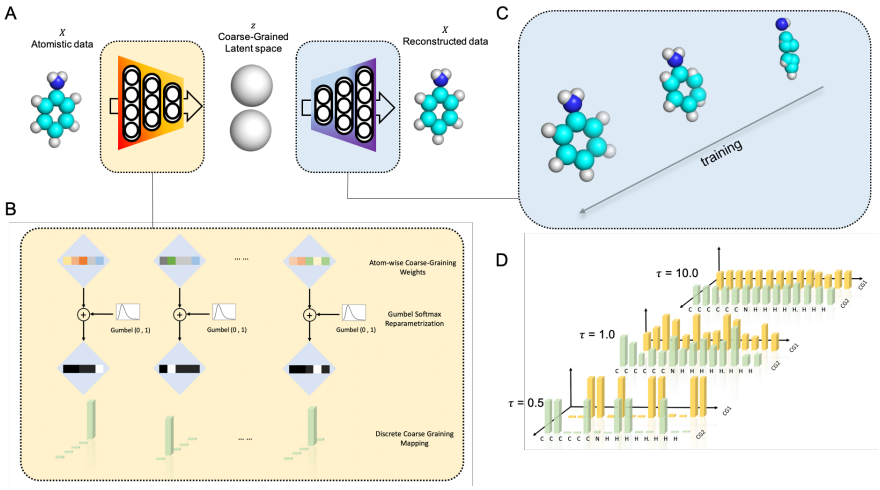

Figure 1: Variational coarse-graining auto-encoding framework. A-B The latent space of The discrete optimization is done using the Gumbel-softmax reparametrization [17, 18]. C. The learning task of reconstruction molecules conditioned on the CG variables in training time. D. Demonstration of continuously relaxation of CG mapping as in equation 2.

A particle based coarse-grained latent space needs to preserve the structure of classical mechanical phase space[19]. To ensure that, we make our encoding function a linear projection in Cartesian space $E(x) : \mathbb{R}^{3n} \rightarrow \mathbb{R}^{3N}$ where n is the number of atoms and N is the desired number of coarse-grained particles. Let x be atomistic coordinates and z be the coarse-grained coordinates. The encoding function should satisfy the following requirements [20, 19]:

1. $z_{ik} = E(x) = \sum_{j=1}^{n} E_{ij} x_{jk} \in \mathbb{R}^3, i = 1 \dots N, j = 1 \dots n$ and k represents the Cartesian coordinates.
2. $\sum_j E_{ij} = 1$ and $E_{ij} \geq 0$
3. Each atom contributes to at most one coarse-grained variable z

where E_{ij} defines the encoding weight toward coarse-grained variables, j is atom index , i is the coarse-grained particle index. Requirement (2) defines the coarse-grained variables to be the center of geometries of contributing atoms. In order to maintain the momentum space consistency based on the coarse-grained mapping, the coarse-grained masses are redefined as $M_i = (\sum_j \frac{E_{ij}^2}{m_j})^{-1}$ [20, 19] ($m_j$is the mass of atom j) and this definition of mass is a corollary of requirement (3).

The encoder function parameters are initialized randomly as atom-wise vectors ϕ which are continuously parameterized as one-hot assignment C_{ij} with Gumbel-softmax reparametrization and the coarse-graining encoding weights are obtained by normalizing over the total number of contributing atoms per coarse-grained atoms to satisfy requirement (2):

$$E_{ij} = \frac{C_{ij}}{\sum_j^n C_{ij}} \qquad (1) \qquad C_{ij} = \frac{e^{(log\phi_{ij} + g_i)/\tau}}{\sum_j e^{(log\phi_{ij} + g_i)/\tau}} \qquad (2)$$

where g_i is sampled from Gumbel distribution via inverse transformation $g_i = -log(-log(u_i))$ and u_i is sampled from uniform distribution from 0 to 1. During training τ is gradually decreased and the

one-hot categorical encoding C_{ij} is the atom-wise discrete Coarse-Graining one-hot vector in the limit of small τ, so that requirement (3) is satisfied.

For the decoding of atomistic coordinates conditioned on coarse-grained coordinates, we opt for a simple decoding approach via geometrical projection using a matrix $\mathbf{D}$ of dimension n by N that maps coarse-grained variables back to the original space so that $\hat{x} = D(z) = \sum_{i=1}^{i=N} \mathbf{D}_{ji} z_{ik}$ where $\hat{x}$ is the reconstructed atomistic frame. Hence, both the encoding and decoding mappings are deterministic. Although deterministic reconstruction via a low dimensional space leads to irreversible information loss that is analogous to mapping entropy introduced in Shell *et al* [21], the decoder and encoder functions are sufficient to decode to the mean underlying atomistic configurations based on maximum likelihood and hence minimize the information loss due to coarse-graining.

$$L_{AE} = \frac{1}{N} \mathbb{E}_{x \sim P(x)} [\| D((E(x)) - x \|_2^2] \tag{3}$$

The counterpart of the regularization term in VAE is the Relative Entropy framework[21] in the coarse-graining theory. However, computing the normalization constant is intractable for Boltzmann distribution, main methods in developing coarse-grained model is matching the gradient in the free energy surface to fit the potential of mean force [20]. The mean force (negative gradient of free energy) is:

$$F(z) = \langle F_{inst} \rangle_{E(x)=z} = \langle F(z) + \epsilon(z) \rangle \rangle_{E(x)=z} = \langle -\mathbf{b} \nabla V(x) \rangle_{E(x)=z} \tag{4}$$

In the case of coarse-graining encoding being atom-wise one-hot vectors, $\mathbf{b} = \mathbf{C}$. We propose a gradient domain regularization by estimating the local mean forces from atomistic dynamics data to smooth-en the coarse-grained free energy surface by minimizing the mean force and fluctuations $\epsilon(E(x))$. We regularize the learning by minimize the mean squared instantaneous forces $\|F_{inst}\|_2^2 \approx \|F(E(x)) + \epsilon(E(x))\|_2^2$ per mini-batch for a smoother coarse-grained free energy surface. By including the instantaneous force loss, we present a regularized loss function that is optimized using Algorithm 1:

$$L_{AE} = \frac{1}{N} \mathbb{E}_{x \sim P(x)} [\| D((E(x)) - x \|_2^2 + \rho \| F_{inst}(E(x)) \|_2^2] \tag{5}$$

Algorithm 1 Variational Coarse-graining Auto-Encoding

$\phi_{ij}, D_{ji}, \tau, \Delta\tau \leftarrow$ initialize parameters
repeat
  $x \leftarrow$ random mini-batch molecular dynamics frames $x \sim P(x)$
  $g_{ij} \leftarrow$ Gumbel(0, 1)
  $C_{ij} \leftarrow \frac{e^{(log \phi_{ij} + g_{ij})/\tau}}{\sum_j e^{(log \phi_{ij} + g_{ij})/\tau}}$
  $E_{ij} \leftarrow \frac{C_{ij}}{\sum_i^n C_{ij}}$
  $g \leftarrow \nabla_{\phi_{ij}, D_{ji}} L_{AE}(\phi_{ij}, D_{ji}; \tau, g_{ij})$
  $\phi_{ij}, D_{ji} \leftarrow$ update parameters using gradients g
  $\tau \leftarrow \tau - \Delta\tau$
until convergence of L_{AE}

3 Experiments

We present the unsupervised auto-encoding process for gas-phase ortho-terphenyl (OTP) and aniline (C6H7N) in Figure 2 trained on atomistic trajectories of 3000 frames sampled by Langevin dynamics at 300K. The results show that the optimized reconstruction loss decreases with the increasing of coarse-graining resolutions: a few coarse-grained atoms have the potential to capture collective motions of the underlying atomistic process. The reconstruction loss represents the lower bound of the information capacity of coarse-grained particles to represent collective atomistic motions through a deterministic projection and atomistic structures can be represented quite well with a few

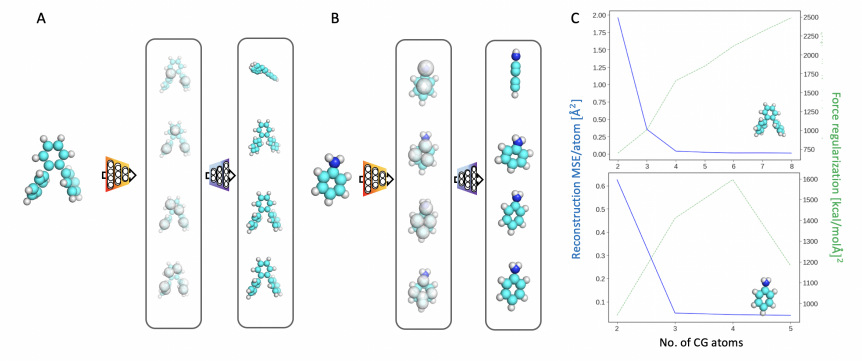

Figure 2: Coarse-Graining encoding and decoding for OTP (A) and aniline (B) with different resolutions. C. average instantaneous force residue and reconstruction loss of trained model. The average mean forces increases with the coarse-graining resolutions increasingly rough underlying free-energy landscape that involves fast motions like bond stretching in the fine-grained descriptions.

coarse-grained particles. In the case of OTP, an intuitive 3-bead mapping that partitioned each of the phenyl rings is learned. When the coarse-grained degrees of freedoms increase up to 4, the additional beads are able to encode more configurational information than three-bead models and therefore can decode back into atomistic coordinates with high accuracy. However, such encoding loses the configuration information of the relative rotation of the two side rings, so the decoded structures yields higher error. Learning to perform stochastic generation of atomistic coordinates conditioned on the coarse-grained data is our future research focus. We further apply the auto-encoding framework to a small peptide molecule trajectories of 5000 frames to test for its representation power of the critical collective variables conditioned on coarse-grained representation. The coarse-grained latent variables can faithfully represent different states in the Ramanchadran map as the coarse-grained resolution is increased (Figure 3). However, the fast degrees of freedom like hydrogen atom fluctuations are lost during the process and cannot be decoded with full resolutions.

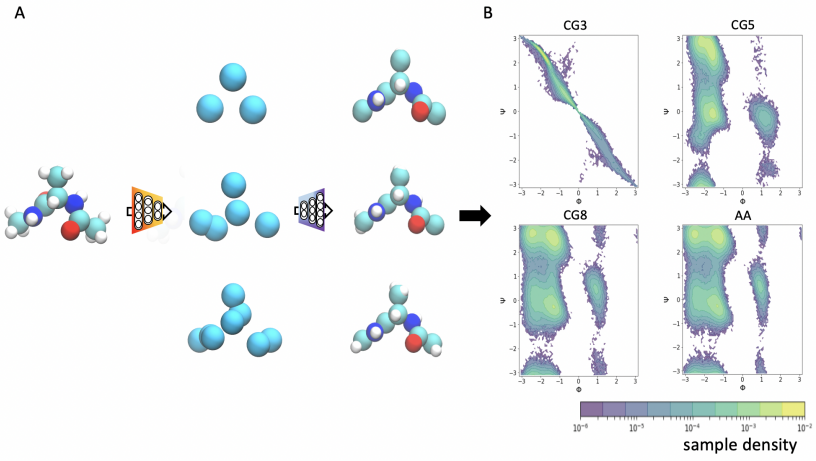

Figure 3: Coarse-Graining encoding and decoding for alanine dipeptide. (A) coarse-graining alanine dipeptide molecules at three different resolutions (3, 5, 8). (B) comparison of dihedral correlation (Ramachandran map) between decoded atomistic distributions and atomistic data. The critical back-bone structures can be inferred with high accuracy with above 5 CG atom resolution and the whole molecule has 32 atoms

4 Conclusion

In summary, we propose an Auto-Encoding framework by treating coarse-grained coordinates as latent variables which can be sampled with coarse-grained molecular dynamics. By regularizing the

latent space instantaneous force minimisation, we train the encoding mapping and a deterministic decoding that can be used to map larger systems to a reduced representation and back to infer atomistic configurations represented by coarse-grained variables. Our work opens up possibilities to use statistical learning as a basis to bridge across multi-scale coarse-grained simulations.

Acknowledgments

WW thanks Toyota Research Institute for financial support. RGB thanks MIT DMSE and Toyota Faculty Chair for support. WW and RGB thank Prof. Adam P. Willard (Massachusetts Institute of Technology) and Prof. Salvador Leon Cabanillas (Universidad Politécnica de Madrid) for helpful discussions.

References

[1] Massimo D. Agostino, Herre Jelger Risselada, Anna Lürick, Christian Ungermann, and Andreas Mayer. A tethering complex drives the terminal stage of SNARE-dependent membrane fusion. *Nature.*, 551(7682):634, nov 2017.

[2] David M Huang, Roland Faller, Khanh Do, Adam J Moulé, Adam J Moul, and Adam J Moule. Coarse-grained computer simulations of polymer / fullerene bulk heterojunctions for organic photovoltaic applications. *J. Chem. Theory Comput.*, 6(2):1–11, feb 2010.

[3] Sidath Wijesinghe, Dvora Perahia, and Gary S. Grest. Polymer Topology Effects on Dynamics of Comb Polymer Melts. *Macromolecules*, 51(19):7621–7628, oct 2018.

[4] K. Michael Salerno, Anupriya Agrawal, Brandon L. Peters, Dvora Perahia, and Gary S. Grest. Dynamics in entangled polyethylene melts. *The European Physical Journal Special Topics*, 225(8-9):1707–1722, oct 2016.

[5] K. Michael Salerno, Anupriya Agrawal, Dvora Perahia, and Gary S. Grest. Resolving Dynamic Properties of Polymers through Coarse-Grained Computational Studies. *Physical Review Letters*, 116(5):058302, feb 2016.

[6] Martin Vögele, Jürgen Köfinger, and Gerhard Hummer. Hydrodynamics of Diffusion in Lipid Membrane Simulations. *Physical Review Letters*, 120, 2018.

[7] W. G. Noid. Perspective: Coarse-grained models for biomolecular systems. *J. Chem. Phys.*, 139(9):90901, sep 2013.

[8] Linfeng Zhang, Jiequn Han, Han Wang, Roberto Car, and Weinan E. DeePCG: Constructing coarse-grained models via deep neural networks. *J. Chem. Phys.*, 149(3):034101, jul 2018.

[9] Karteek K. Bejagam, Samrendra Singh, Yaxin An, and Sanket A. Deshmukh. Machine-Learned Coarse-Grained Models. *J. Phys. Chem. Lett.*, 9(16):4667–4672, aug 2018.

[10] Tobias Lemke and Christine Peter. Neural Network Based Prediction of Conformational Free Energies - A New Route toward Coarse-Grained Simulation Models. *J. Chem. Theory Comput.*, 13(12):6213–6221, dec 2017.

[11] Jiang Wang, Simon Olsson, Christoph Wehmeyer, Adrià Pérez, Nicholas E. Charron, Gianni de Fabritiis, Frank Noé, and Cecilia Clementi. Machine Learning of Coarse-Grained Molecular Dynamics Force Fields. *ACS Central Science*, page acscentsci.8b00913, apr 2019.

[12] Lorenzo Boninsegna, Gianpaolo Gobbo, Frank Noé, and Cecilia Clementi. Investigating Molecular Kinetics by Variationally Optimized Diffusion Maps. *Journal of Chemical Theory and Computation*, 11(12):5947–5960, dec 2015.

[13] Joseph F. Rudzinski and William G. Noid. Investigation of Coarse-Grained Mappings via an Iterative Generalized Yvon–Born–Green Method. *J. Phys. Chem. B*, 118(28):8295–8312, jul 2014.

[14] Naftali Tishby and Noga Zaslavsky. Deep Learning and the Information Bottleneck Principle. *arXiv:1503.02406*, mar 2015.

[15] Christoph Wehmeyer and Frank Noé. Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. *J. Chem. Phys.*, 148(24), oct 2018.

[16] Andreas Mardt, Luca Pasquali, Hao Wu, and Frank Noé. VAMPnets for deep learning of molecular kinetics. *Nat. Commun.*, 9(1):5, dec 2018.

[17] Eric Jang, Shixiang Gu, and Ben Poole. Categorical Reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations*, nov 2017.

[18] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. In *International Conference on Learning Representations*, nov 2016.

[19] W G Noid, Jhih Wei Chu, Gary S Ayton, Vinod Krishna, Sergei Izvekov, Gregory A Voth, Avisek Das, and Hans C Andersen. The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *J. Chem. Phys.*, 128(24):243116, 2008.

[20] Eric Darve. Numerical Methods for Calculating the Potential of Mean Force. In *New Algorithms for Macromolecular Simulation*, pages 213–249. Springer-Verlag, Berlin/Heidelberg, 2006.

[21] M Scott Shell. Coarse-Graining With The Relative Entropy. In *Advances in Chemical Physics*, volume 161, pages 395–441. Wiley-Blackwell, sep 2016.