†©Electronic Supplementary Material (ESI) for Chemical Science.
This journal is ©The Royal Society of Chemistry 2022

Electronic Supplementary Information†

Molecular Dynamics Energetics of Polymers in Solution from Supervised Machine Learning

JamesAndrews[1,2], Olga Gkountouna[2] and Estela Blaisten-Barojas[1,2,*]
[1]*Center for Simulation and Modeling, George Mason University, Fairfax, Virginia, USA*
[2]*Department of Computational and Data Sciences, George Mason University, Fairfax, Virginia, USA*
[*]Corresponding author: blaisten@gmu.edu

May 9, 2022

Table of Contents

S1 DSPE-PEG and ethyl acetate atomic interactions.

The DSPE-PEG(2000) polymer-lipid macromolecule and the ethyl acetate solvent molecules were modeled with all-atom force fields. A succinct enumeration of the analytical terms composing both the GAFF and Amber-Lipid17 [1, 2, 3] includes *bonded* and *non-bonded* terms. The intra-molecule potential energy is a sum of four bonded terms and two non-bonded terms. The bonded terms are: the Morse potential for the stretching between two contiguous atoms, the harmonic angle bending between three contiguous bonded atoms, the harmonic out-of plane bending for moving an atom bonded to three others out of the plane, and the torsion energy due to the twist of four contiguous bonded atoms forming a dihedral angle. The non-bonded interactions are Coulomb and Lennard-Jones potentials. Non-bonded terms in the intra-molecule potential energy includes a sum of pairs between 4th neighboring atoms and beyond. In addition, each atom is assigned a *type* according to its local chemical environment. Depending on the atom types, the parameters entering in duplets, triplets, or quartet of atoms are assigned different values. The force field parameters are included in a curated database embedded in the Amber package [1, 4].

Once the force field is defined for a specific system, its multitude of parameters are transferable to the molecular dynamics of choice that supports such force field. Hence, the GAFF-Lipid17 were transferred to the GROMACS 2018-2020 [4] package used for our MD simulations. The GROMACS topology files for the ethyl acetate and the DSPE-PEG(2000) are provided open access in the Zenodo archive [5].

Therein, *PE* terms the sum of the four DSPE-PEG(2000) intra-macromolecule potential energies and the sum of their interaction energy with the solvent is termed *IE*.

S2 The RNN hyperparameters

Figure S1 summarizes the behavior of the RNN hyperparameters depicting the validation errors for LSTM and GRU when one parameter was varied while the others were kept constant.

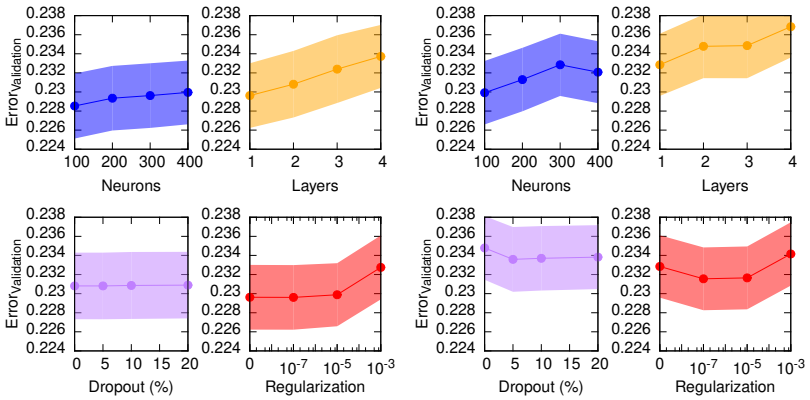

Figure S1: Validation error comparison between various hyperparameters of the GRU(left) and LSTM (right). Data corresponds to Ensemble$_{100}$ Set 3. Errors are shown after 40 epochs, or at the minimum validation error if sooner as in the case of 4 layers. Dots are mean values over the ensemble of series and shaded area is $\pm$ one standard deviation

S3 Expectation maximization clustering of PE and IE time series

Table S1: Energetic properties of EM clustered patterns from the time series in Ensemble$_{100}$. EM groups are numbered in ascending order of PE. The radius of gyration R_g of the DSPE-PEG aggregate is mass weighted

	Cluster No.	0	1	2	3	4	5
	% of samples	11.7	24.4	16.0	16.9	20.4	10.6
	PE (MJ/mol)	8.83 ± 0.12	8.88 ± 0.11	8.95 ± 0.11	8.99 ± 0.11	9.10 ± 0.11	9.14 ± 0.11
Set 1	IE (MJ/mol)	-5.99 ± 0.08	-6.215 ± 0.06	-6.385 ± 0.07	-6.10 ± 0.06	-6.25 ± 0.07	-6.45 ± 0.09
	E_{coh} (MJ/mol)	-0.53 ± 0.08	-0.508 ± 0.08	-0.49 ± 0.07	-0.553 ± 0.09	-0.547 ± 0.08	-0.516 ± 0.08
	Rg (nm)	2.05 ± 0.08	2.08 ± 0.08	2.14 ± 0.08	2.06 ± 0.08	2.09 ± 0.09	2.14 ± 0.10
	% of samples	16.9	17.8	19.8	21.8	12.6	11.1
	PE (MJ/mol)	8.64 ± 0.10	8.65 ± 0.14	8.77 ± 0.11	8.80 ± 0.11	8.94 ± 0.12	8.96 ± 0.12
Set 2	IE (MJ/mol)	-6.07 ± 0.07	-5.882 ± 0.07	-6.241 ± 0.07	-6.042 ± 0.07	-6.357 ± 0.08	-6.17 ± 0.07
	E_{coh} (MJ/mol)	-0.45 ± 0.05	-0.47 ± 0.05	-0.45 ± 0.05	-0.46 ± 0.05	-0.49 ± 0.06	-0.49 ± 0.07
	Rg (nm)	2.28 ± 0.10	2.29 ± 0.12	2.28 ± 0.10	2.28 ± 0.10	2.26 ± 0.08	2.29 ± 0.10
	% of samples	13.1	16.4	19.7	21.2	14.0	15.6
	PE (MJ/mol)	8.93 ± 0.12	8.95 ± 0.11	9.06 ± 0.10	9.12 ± 0.11	9.24 ± 0.11	9.25 ± 0.13
Set 3	IE (MJ/mol)	-5.83 ± 0.09	-6.06 ± 0.07	-6.18 ± 0.09	-5.95 ± 0.09	-6.14 ± 0.09	-6.43 ± 0.10
	E_{coh} (MJ/mol)	-0.76 ± 0.06	-0.68 ± 0.07	-0.67 ± 0.07	-0.75 ± 0.06	-0.73 ± 0.07	-0.66 ± 0.08
	Rg (nm)	2.11 ± 0.07	2.08 ± 0.09	2.05 ± 0.09	2.08 ± 0.08	2.06 ± 0.07	1.99 ± 0.08

S4 The statistical distributions of the PE and IE energies time series

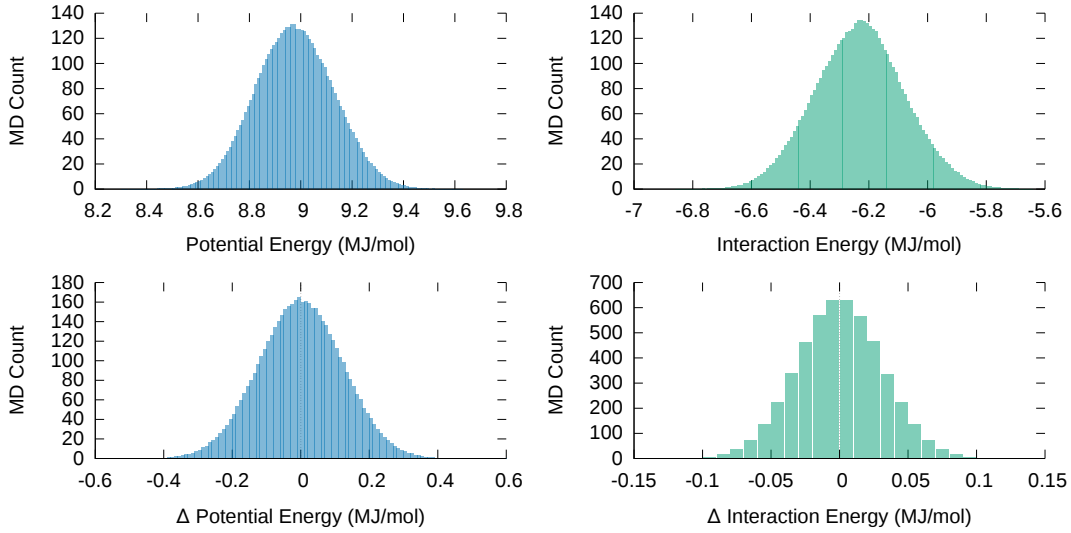

Figure S2: Statistical distributions of the PE and IE time points in Ensemble$_{10}$ for Set 1. Top histograms correspond to the MD time points. Bottom histograms correspond to series of $\Delta PE = (PE_{t_{n+1}} - PE_{t_n})$ and $\Delta IE = (IE_{t_{n+1}} - IE_{t_n})$, where t_n identifies each of the time values in the series composing the ensemble that were spaced by 10 fs.

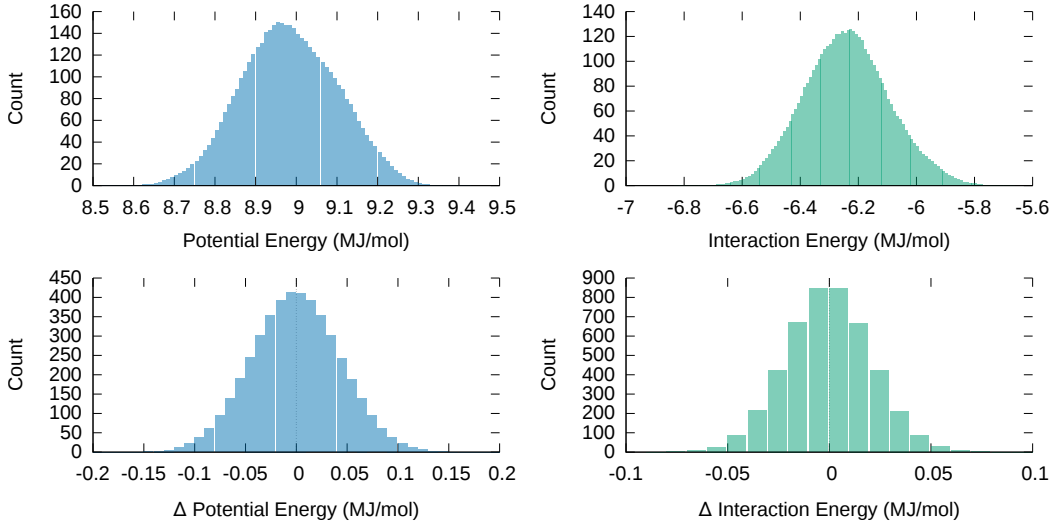

Figure S3: Approximated statistical distributions of the PE and IE time points reproduced by the GRU models for Ensemble$_{10}$ for Set 1. Top histograms reproduce the MD time points. Bottom histograms are the corresponding $\Delta PE = (PE_{t_{n+1}} - PE_{t_n})$ and $\Delta IE = (IE_{t_{n+1}} - IE_{t_n})$, where n identifies each of the time values in the GRU generated series that are spaced by 10 fs.

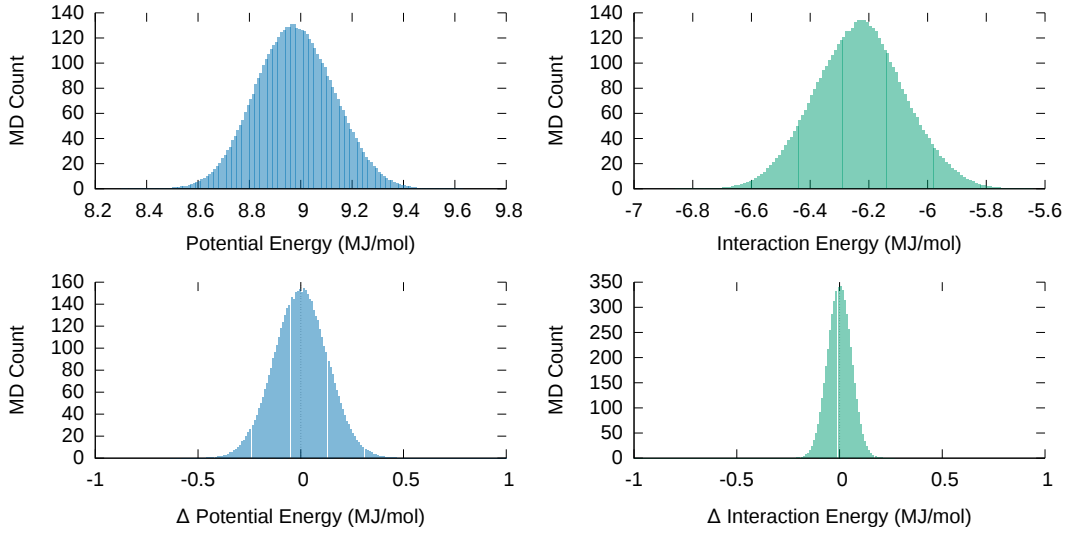

Figure S4: Statistical distributions of the PE and IE time points in Ensemble$_{100}$ for Set 1. Top histograms correspond to the MD time points. Bottom histograms correspond to series of $\Delta PE = (PE_{t_{n+1}} - PE_{t_n})$ and $\Delta IE = (IE_{t_{n+1}} - IE_{t_n})$, where n identifies each of the time values in the series composing the ensemble that were spaced by 100 fs.

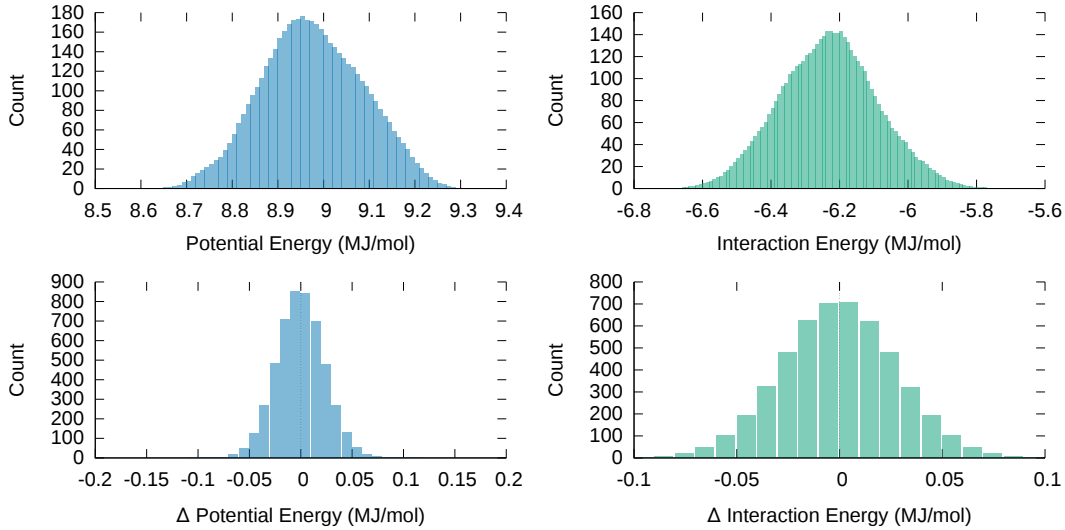

Figure S5: Approximated statistical distributions of the PE and IE time points reproduced by the GRU models for Ensemble$_{100}$ for Set 1. Top histograms reproduce the MD time points. Bottom histograms are the corresponding $(\Delta PE = (PE_{t_{n+1}} - PE_{t_n})$ and $\Delta IE = (IE_{t_{n+1}} - IE_{t_n})$, where n identifies each of the time values in the GRU generated series that are spaced by 100 fs.

S5 List of RNN training/testing scenarios attempted

The main paper describes in detail the best scenario concerning the RNN forecast model. However, along our investigation several other possibilities were evaluated for the construction of the RNN data models, as listed below:

- Input of 8 time series corresponding to 4 intra-potential energy of each DSPE-PEG macromolecule, and 4 single macromolecule interaction energy with the solvent. The RNN data models cpu time for their creation increases by one order of magnitude and the forecast model has no improvement.
- Input of 3 time series corresponding to PE, IE and macromolecular aggregate cohesive energy yields inadequate forecasts.
- Input of 4 time series corresponding to PE, IE, macromolecular aggregate cohesive energy and R_g yields incorrect means for the desired energetic forecast
- Input of 7 time series, PE and the 6 series with PE time-patterns from the EM clustering is cpu time demanding and yields almost identical forecast that not using them.
- Input of 7 time series, IE and the 6 series with IE patterns obtained from the EM clustering is cpu time demanding and yields almost identical forecast that not using them.

S6 Time dependent autocorrelation function of PE and IE time series

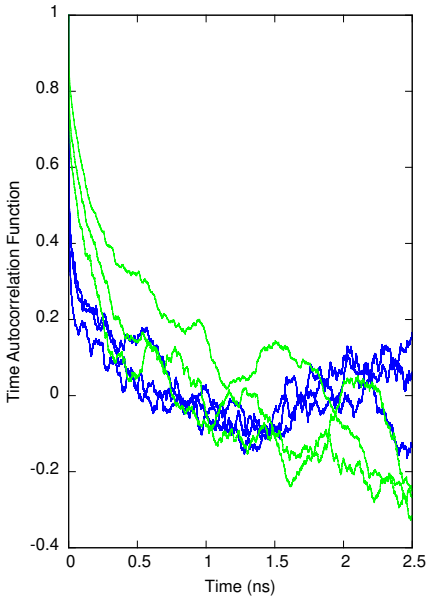

Figure S6: Time autocorrelation function of the PE (blue) and IE (green) as a function of time for the three Sets 1, 2, 3.

References

[1] The 2018 reference manual for Amber18 and AmberTools18, https://ambermd.org/doc12/Amber18.pdf. Online: accessed February 2, 2021.

[2] J. Wang, R. Wolf, J. Caldwell, P. Kollman and D. Case, J. Comput. Chem., 2004, 25, 1157-1174.

[3] C. J. Dickson, B. D. Madej, A. A. Skjevik, R. M. Betz, K. Teigen, I. R. Gould and R. C. Walker, J. Chem. Theory Comput., 2014, 10, 865-879.

[4] GROMACS 2018-20 Manual, https://manual.gromacs.org/documentation/2020/index.html. Online: accessed February 22, 2021.

[5] J. Andrews, O. Gkountouna and E. Blaisten-Barojas, 2022, https://doi.org/10.5281/zenodo.6503359. Online: accessed May 1, 2022.