Unraveling the molecular mechanism of collagen flexibility during physiological warmup using molecular dynamics simulation and machine learning

Wei-Han Hui[1, ‡], Pei-Hsin Chiu[1, 2, ‡], Ian-Ian Ng[3], Shu-Wei Chang[1, 4, *], Chia-Ching Chou[2, *], Hsiang-Ho Chen[5, 6, *]

[1]Department of Civil Engineering, National Taiwan University, Taipei City, Taiwan
[2]Institute of Applied Mechanics, National Taiwan University, Taipei City, Taiwan
[3]School of Biomedical Engineering, Taipei Medical University, Taipei City, Taiwan
[4]Department of Biomedical Engineering, National Taiwan University, Taipei City, Taiwan
[5]Center for Biomedical Engineering, Department of Biomedical Engineering, Chang Gung University, Taoyuan City, Taiwan
[6]Bone and Joint Research Center, Department of Orthopedic Surgery, Chang Gung Memorial Hospital Linkou Branch, Taoyuan 33305, Taiwan

[‡]**Contributed equally to this work**

[*]**Corresponding authors**
Shu-Wei Chang, PhD
Department of Civil Engineering, National Taiwan University
No. 1, Sec. 4, Roosevelt Rd., Da'an Dist., Taipei City 1017, Taiwan
E-mail: changsw@ntu.edu.tw

Chia-Ching Chou, PhD
Institute of Applied Mechanics, National Taiwan University
No. 1, Sec. 4, Roosevelt Rd., Da'an Dist., Taipei City 1017, Taiwan
E-mail: ccchou@iam.ntu.edu.tw

Hsiang-Ho Chen, PhD
Graduate Institute of Biomedical Engineering, Chang Gung University
No. 259, Wenhua 1st Rd., Guishan Dist., Taoyuan City 33302, Taiwan
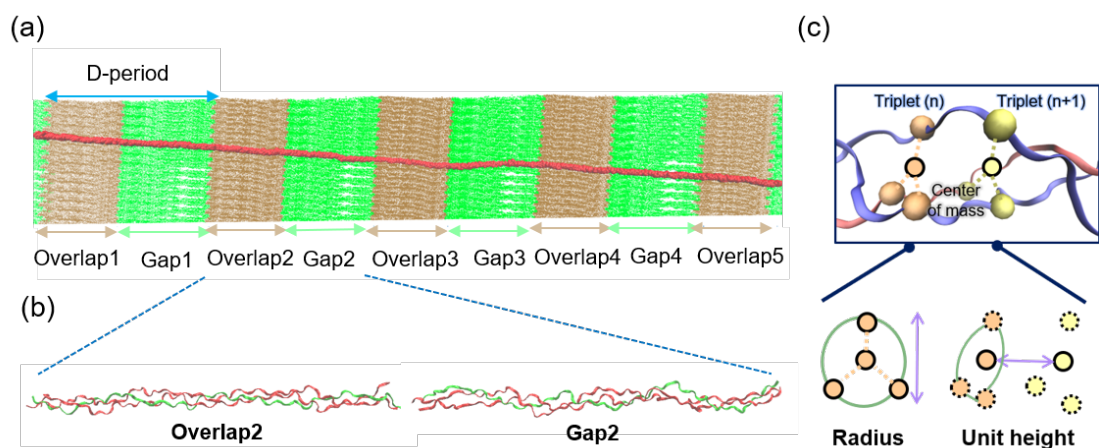E-mail: hchen@mail.cgu.edu.tw

Figure S1. Snapshot of the simulation models. (a) The preequilibrium fibril model. After the equilibrium, we select fragments from gap2 and overlap2 regions to build the molecular models. (b) The molecular models in the gap region and overlap region. (c) The illustration of unit height and radius. The unit height represents the longitudinal structure and the radius is related to the cross-section property.

Table S2. The values of the parameters in the VAE model.

"learning_rate"	0.38
"n_estimators"	60
"max_depth"	6
"subsample"	0.53
"colsample_bytree"	0.7
"gamma"	0
"reg_alpha"	0.18
"reg_lambda"	0.14

Table S3. The parameters of the XGBoost regression model in the overlap case.

"learning_rate"	0.26
"n_estimators"	85
"max_depth"	6
"subsample"	0.3
"colsample_bytree"	0.7
"gamma"	0
"reg_alpha"	0.22
"reg_lambda"	0.84
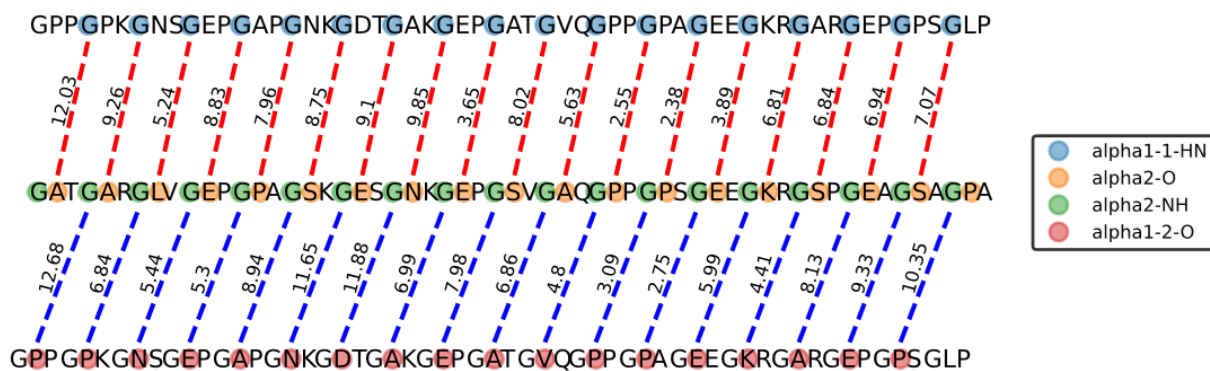
Figure S4. Interchain distances between the pairs of donors and acceptors along the sequence of the overlap region at 310 K.
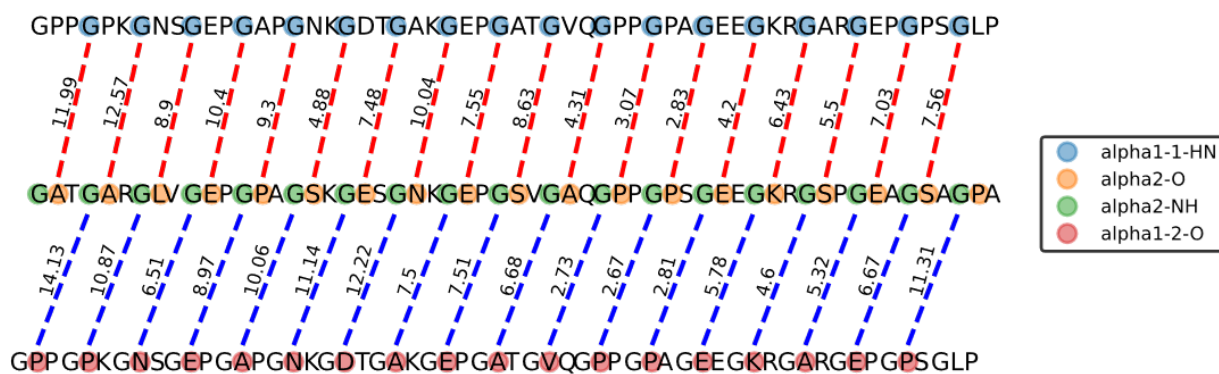
Figure S5. Interchain distances between the pairs of donors and acceptors along the sequence of the overlap region at 313 K.
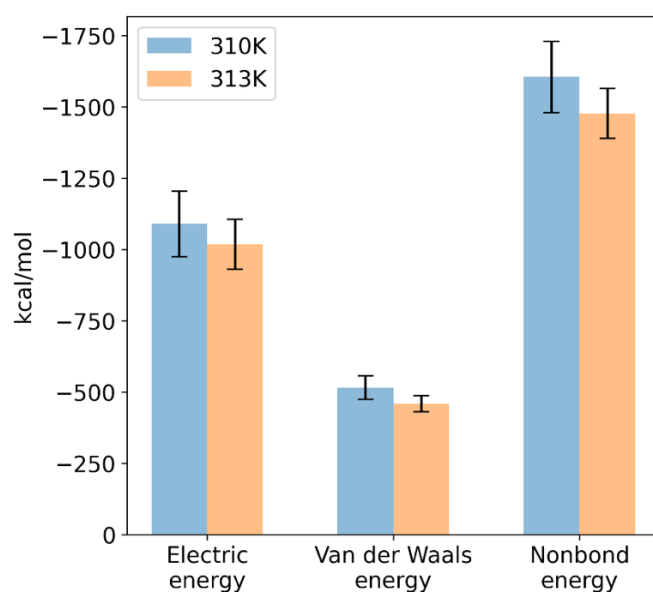
Figure S6. The nonbonded potential of the overlap region at 310 K and 313 K.

Figure S7. (a) Original heatmap of combinations of sequences in the overlap region. (b) Heatmap of different combinations of sequences in the overlap region after ML prediction.

Figure S8. (a) Original heatmap of combinations of sequences in the gap region. (b) Heatmap of different combinations of sequences in the gap region after ML prediction.

The sequence of the GAR-GSK model:

α1-
GPOGPOGPOGARGARGARGARGARGARGARGARGARGARGARGARGARGA RGARGARGARGARGARGPOGPOGPO and

α2-
GPOGPOGPOGSKGSKGSKGSKGSKGSKGSKGSKGSKGSKGSKGSKGSKGS KGSKGSKGSKGSKGPOGPOGPO.

The sequence of the GEP-GSK model:

α1-
GPOGPOGPOGEPGEPGEPGEPGEPGEPGEPGEPGEPGEPGEPGEPGEPGEPG
EPGEPGEPGEPGEPGPOGPOGPO and
α2-
GPOGPOGPOGSKGSKGSKGSKGSKGSKGSKGSKGSKGSKGSKGSKGSKGSKGS
KGSKGSKGSKGSKGPOGPOGPO.

(a) (b)

(c)

Figure S9. Validation of the ML model by molecular dynamics simulation. (a) The end-to-end distance for combinations of GAR-GSK and GEP-GSK. (b) A snapshot of the combinations of GAR-GSK and (c) a snapshot of the combinations of GEP-GSK. The end-to-end distance of GEP-GSK model is shorter than GAR-GSK under 310 K. These results are consistent with the prediction of the machine learning model.

Table S10. The ground truth data and prediction data from ML model. Black: training data. Red: testing data.

Sequence	Ground truth	Prediction
GPP,GAT	0.86	0.85
GPK,GAR	0.37	0.39
GNS,GLV	0.47	0.48
GNK,GSK	0.62	0.61
GAK,GNK	1.09	0.94
GEP,GEP	0.5	0.5

GAT,GSV	0.44	0.4
GVQ,GAQ	0.25	0.25
GPP,GPP	0.23	0.26
GEE,GEE	0.24	0.25
GKR,GKR	0.62	0.65
GAR,GSP	0.45	0.48
GEP,GEA	0.59	0.58
GLP,GPA	0.05	0.1
GAP,GPA	0.28	0.2
GDT,GES	0.77	0.74
GPA,GPS	0.19	0.22
GPS,GSA	0.52	0.49
GEP,GEP	0.44	0.5