Supplementary Information

RPnet: A Reverse Projection Based Neural Network for Coarse-graining Metastable Conformational States for Protein Dynamics

Hanlin Gu,[a‡] Wei Wang,[b‡] Siqin Cao,[b‡#] Ilona Christy Unarta,[c#] Yuan Yao,[a] Fu Kit Sheong,[b,d*] Xuhui Huang[b,c#*]

[a]Department of Mathematics, Hong Kong University of Science and Technology, Kowloon, Hong Kong

[b]Department of Chemistry, Hong Kong University of Science and Technology, Kowloon, Hong Kong

[c]Department of Chemical and Biological Engineering, Hong Kong University of Science and Technology, Kowloon, Hong Kong

[d]Institute for Advanced Study, Hong Kong University of Science and Technology, Kowloon, Hong Kong

*To whom correspondence should be addressed. Email: fksheong@connect.ust.hk or xuhuihuang@ust.hk

‡ Hanlin Gu, Wei Wang and Siqin Cao contribute equally to this work.

#Current address: Department of Chemistry, University of Wisconsin-Madison, Madison, Wisconsin, 53706, U.S.A

Supplementary Text

1. Comparison of Y matrix obtained from different kinetic lumping methods

In this section, we present the underlying Y matrices that are used to compute the Y loss of the corresponding cases, namely the *Alanine Dipeptide* (lagtime of 5ps), 2D potential (lagtime of 3000 saving intervals), and RNAP (lagtime of 90ns) respectively (Fig. S1a-S1c). We used the deep blue colour to represent $y_{ij} = 1$ and light blue to represent $y_{ij} = 0$.

For the *Alanine Dipeptide* case (Fig. S1), all methods result in a Y matrix with close resemblance to the identity matrix, with the diagonal elements all close to 1 and off-diagonal elements all close to 0. In fact, RPnet, PCCA+ and MPP give exact same lumping results, and so their Y matrices are also the same.

For the 2D potential system (Fig. S2), the Y matrix corresponding to RPnet has larger diagonal elements and smaller off-diagonal elements when compared to that resulted from the PCCA+, consistent with the better state boundary partitioning of RPnet.

For the case of RNAP (Fig. S3), the Y matrix of RPnet is again closest to identity matrix. It can also be seen that the Y matrix corresponding to hierarchical clustering with Ward linkage actually has a mixing between the third and fourth eigenvectors, which correspond to the erroneous state partitioning shown in FIG. 5.

2. Comparison of the implied timescales of macrostate-MSMs generated by different kinetic lumping methods

Fig. S4 shows the implied timescale of the 2D potential dataset. Panel (a) presents the 9 slowest implied timescales of the microstate model. The implied timescales of the lumped macrostate models from RPnet and PCCA+ are shown in panel (b) and (c), respectively. As shown in Fig. S4, the implied timescales obtained from 4-macrostate MSMs generated by RPnet and PCCA+ are both consistent with the three slowest implied timescales predicted by the microstate-MSM.

3. Stability of RPnet in different lagtime

We have presented in the main text that our RPnet method performs better than other methods in several specific lag times. We will hereby show that our RPnet approach is also robust, where Y-loss is always low and stable. Fig. S5 displays the Y-loss values computed at different lag times in the three systems. In order to demonstrate the stability in the performance of RPnet, we compare our method to the PCCA+. The result demonstrates that Y-loss of RPnet is significantly less sensitive to the value of the lag time compared to PCCA+. Furthermore, we show that the Y-loss values of RPnet are always lower than those from PCCA+, even though PCCA+ can achieve comparable performance with RPnet in some specific lag times (see Fig. S5).

4. Performance of RPnet with number of macrostates

In this section, we examined the performance of our RPnet method by varying number of macrostates: i.e., $N = 2{,}6$ for 1D-potential, $N = 3$ for alanine dipeptide, $N = 2{,}5$ for RNAP.

For the 1D potential, we have performed two different manual partitioning and examine the Y-loss in each case. we show that when N=2 (with the state boundary correctly located at the

highest energy barrier, see the left panel of Fig. S6(b)), the reverse projected modes and the original microstate eigenmodes are still in good agreement (see Fig. S6(b)). However, when N=6, several state boundaries split energy basins (indicating poor metastability of the lumped macrostate model). In this case, we observer a large discrepancy between the original and reverse projected ones, especially for fast modes (Fig. S7).

Also, for the alanine dipeptide with N=3, we show the three macrostates obtained from our RPnet in Fig. S8(b). Our RPnet indeed yields 3 macrostates that can still reasonably separate the metastable regions. In particular, two macrostates (in green and orange, see Fig. S8(d)) in the 4-state model are merged (in green, see Fig. S8(b)).

For the RNAP with N=2,5, as shown in Fig. S9, our RPnet method yields similar state boundaries with PCCA+ when N=2 (Fig. S9(a) v.s. Fig.S9(d)) and N=5 (Fig. S9(c) v.s. Fig.S9(f)). When comparing the Y-loss values of various macro-state models obtained from RPnet with different N, we show that the Y-loss value for N=4 (*Y-loss=0.015*) is lower than N=5 (*Y-loss=0.029*). Interestingly, we also found that the model with N=2 generates a very small Y-loss value (*Y-loss=0.002*). We anticipate that the reason behind is that the binomial assignment when N=2 only captures one dynamic mode (the slowest mode), while other dynamic modes: e.g., the 2^{nd}, 3^{rd} (when N=4 and 5) or even the 4^{th} (when N=5) slowest dynamic mode, may contribute to the increased Y-loss values when N=4 or 5.

REFERENCES

1 I. C. Unarta, S. Cao, S. Kubo, W. Wang, P. P.-H. Cheung, X. Gao, S. Takada and X. Huang, *Proc. Natl. Acad. Sci.*, , DOI:10.1073/pnas.2024324118.

Supplementary Figures:

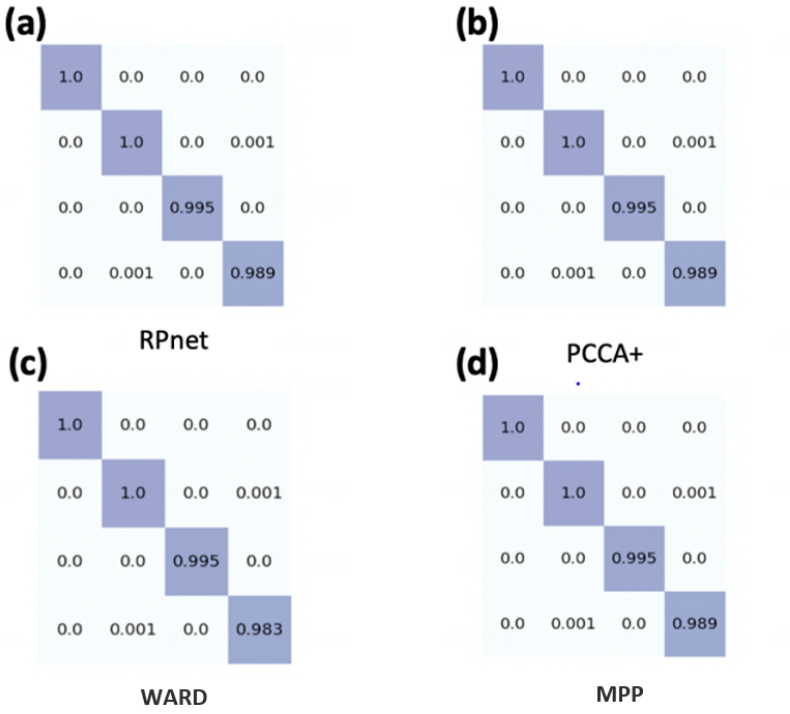

Fig. S1: The Y matrix of *Alanine Dipeptide* built with the lagtime of 5ps. The microstate model has 100 states, and the macrostate models have 4 states. In (a-d), the macrostate models are generated by RPnet, PCCA+, hierarchical clustering with Ward linkage, and MPP, respectively.

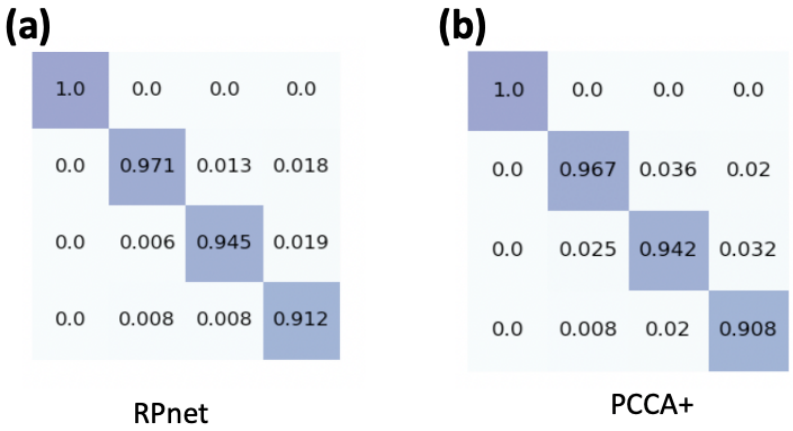

Fig. S2: The Y matrix of 2D-potential with the lag time of 3000 steps. The microstate model has 961 states while the macrostate models have 4 states. The macrostate models are generated by: (a) RPnet and (b) PCCA+, respectively.

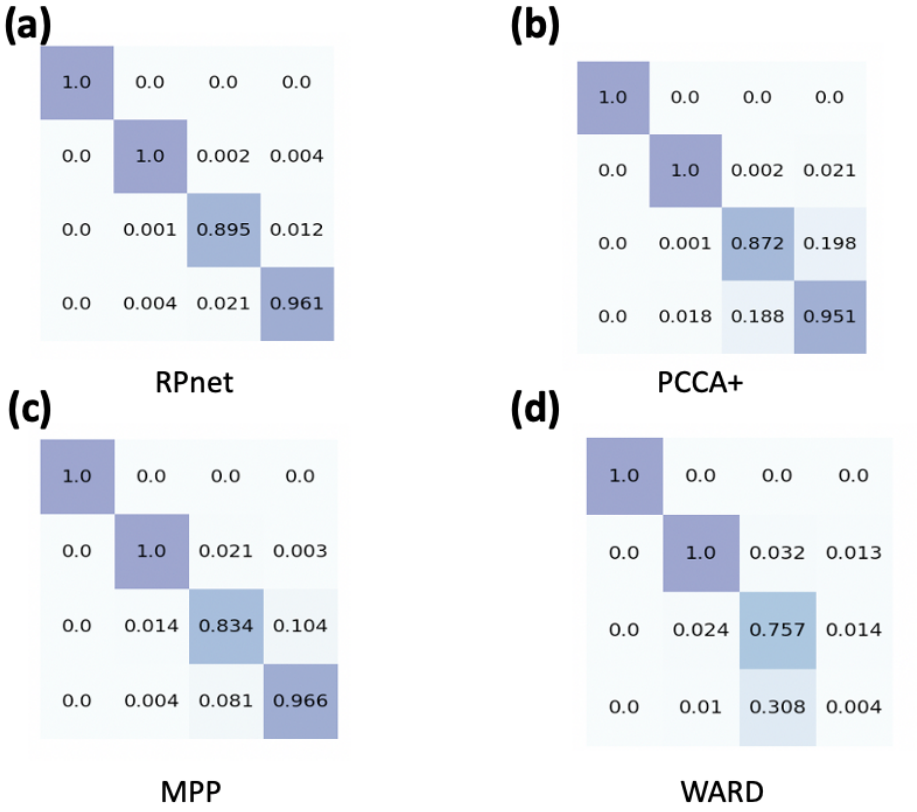

Fig. S3: The Y matrix of RNAP with 90 ns lag time. The macrostate model has 100 states, while the macrostate models have 4 states. In (a-d), the macrostate models are generated by RPnet, PCCA+, MPP and hierarchical clustering with Ward linkage, respectively.

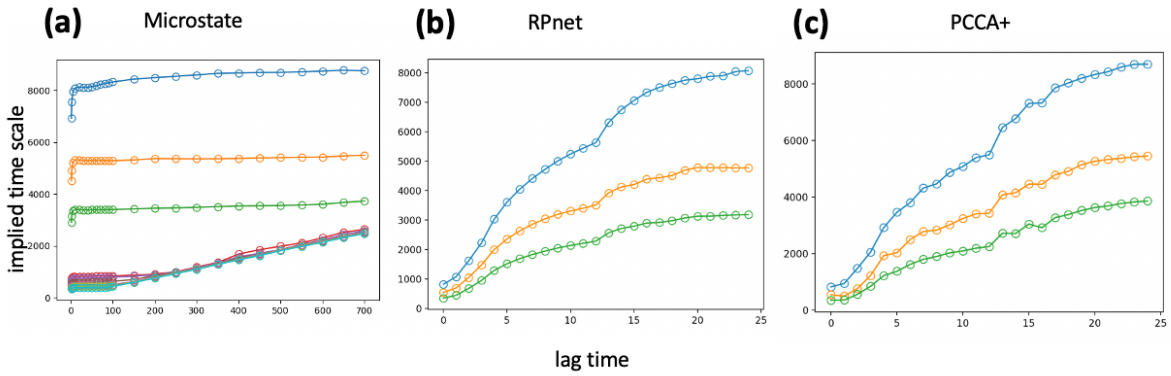

Fig. S4: The implied time scales of 2D potential system with different lag time. (a) The implied time scale of Microstates. (b) The implied timescale plots of the macrostate model generated by RPnet, (c) The implied time scale plots of the macrostate model generated by PCCA+.

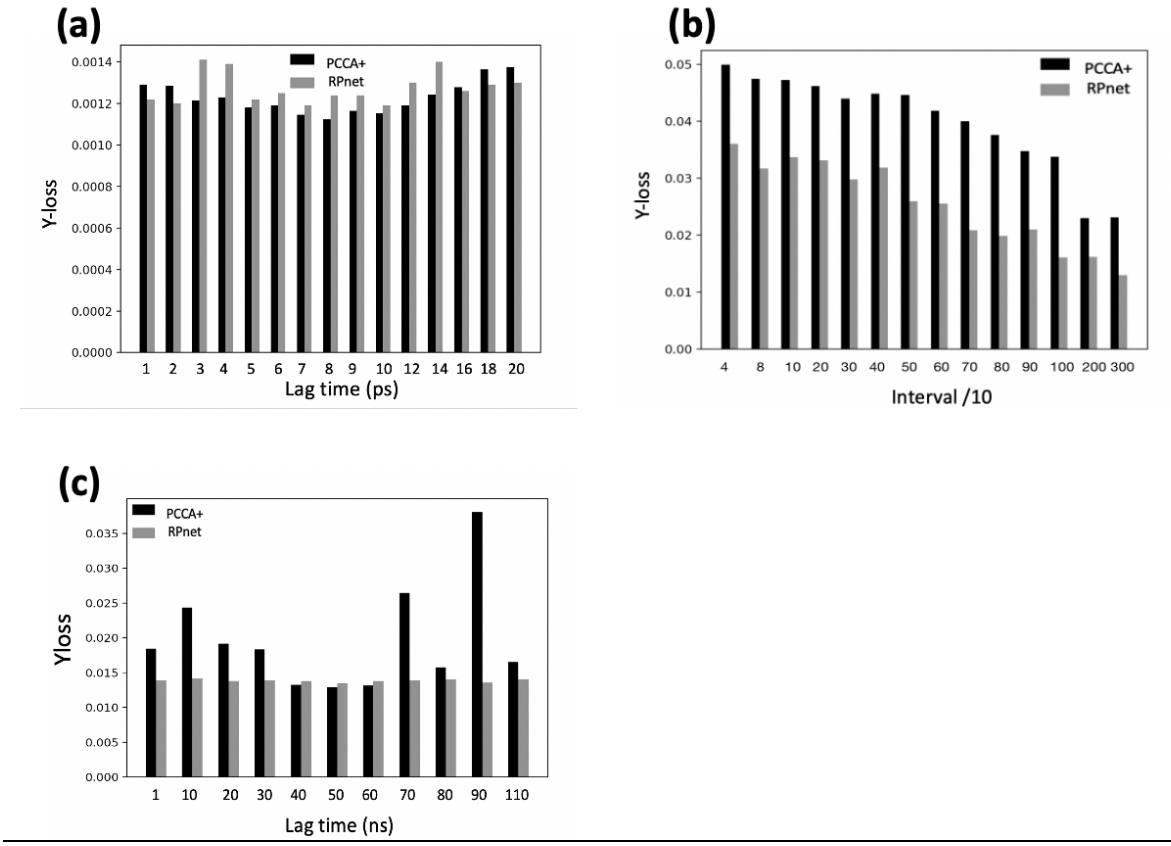

Fig. S5: The Y-loss result with different lag time. (a) The Y-loss change in the *Alanine Dipeptide*. (b) The Y-loss change in the 2D potential. (c) The Y-loss change in the RNAP.

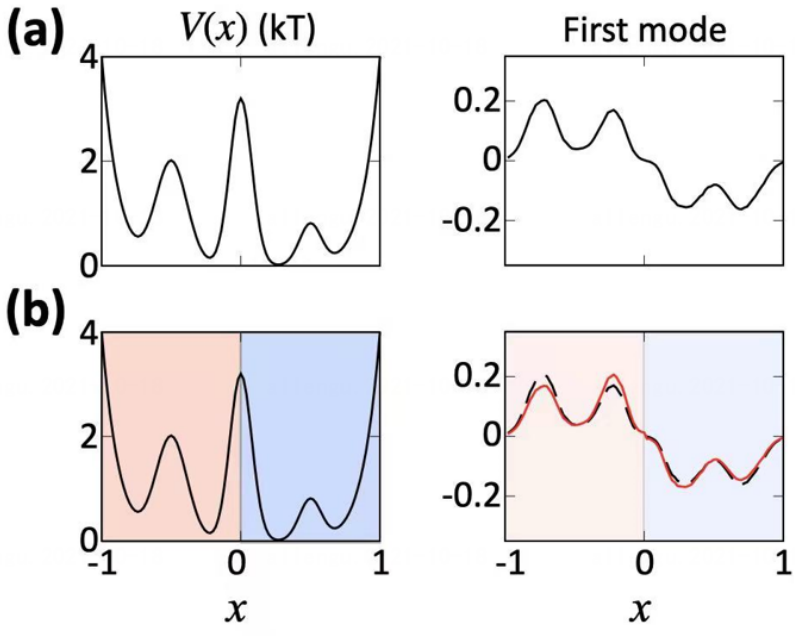

Fig. S6: Illustration of reverse projected modes in 1D potential with N=2. (a) Energy landscape $V(x)$ and the corresponding microstate transition mode. (b) Reverse projected mode of the low-resolution lumping. It is clear from the Fig. that the reverse projected mode is smooth within each macrostate region, but at the boundaries between two macrostates, discontinuities could be present.

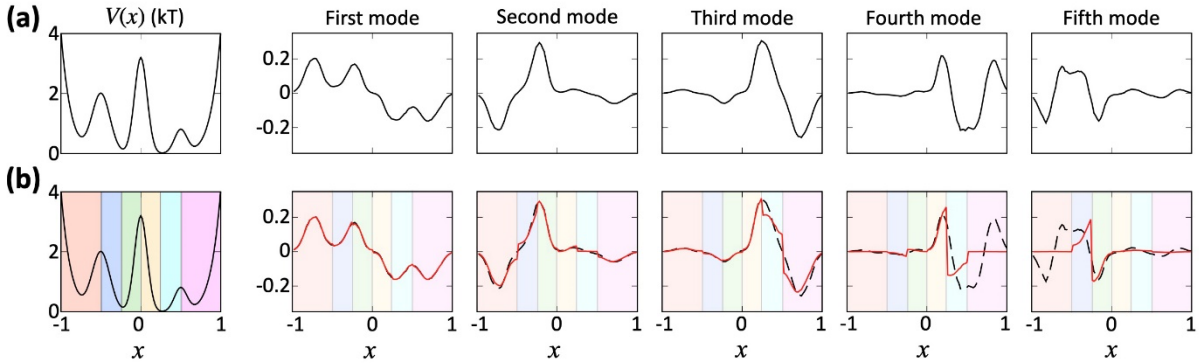

Fig. S7: Illustration of reverse projected modes in 1D potential with N=6. (a) Energy landscape $V(x)$ and the corresponding microstate transition modes. (b) Reverse projected modes of the lumping. It is clear from the Fig. that the reverse projected modes are smooth within each macrostate region, but at the boundaries between two macrostates, clear discontinuities could be present. It is also obvious that the fourth and fifth reverse projected modes failed to reproduce the original ones, indicating that the lumping is suboptimal.

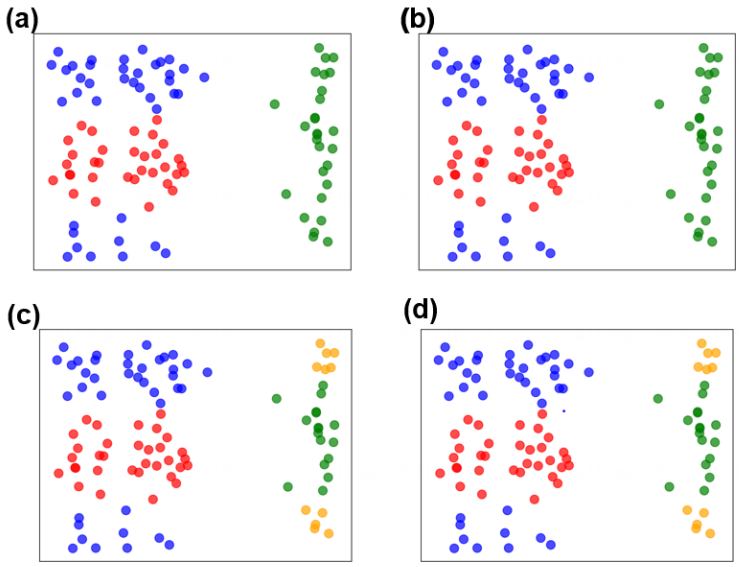

Fig. S8: Lumping assignment in alanine dipeptide with N=3 and 4. Fig. (a) and (b) are the lumping assignments of PCCA+ and RPnet with N=3; Fig. (c) and (d) are the lumping assignments of PCCA+ and RPnet with N=4 (same as the two in the main text, shown here only for comparison). It is clear that when N changes from 4 to 3, RPnet gives a "low resolution" lumping that merges the green and the orange macrostates into one, and the state boundary still respects barriers of the underlying energy landscape.

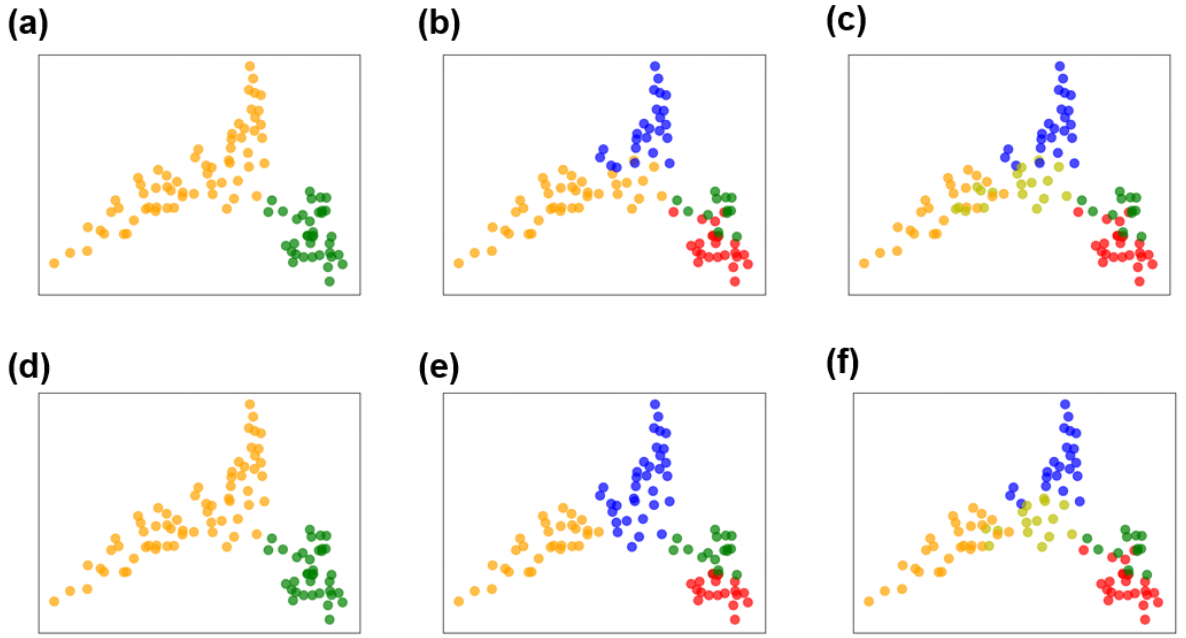

Fig. S9: Lumping assignments for the RNAP system with N=2, 4 and 5. Fig. (a) and (d) are the lumping assignments of PCCA+ and RPnet (Y-loss: 0.002) with N=2; Fig. (b) and (e) are the lumping assignments of PCCA+ and RPnet (Y-loss: 0.015) with N=4 (also shown in main text); Fig. (c) and (f) are the lumping assignments of PCCA+ and RPnet (Y-loss: 0.029) with N=5.

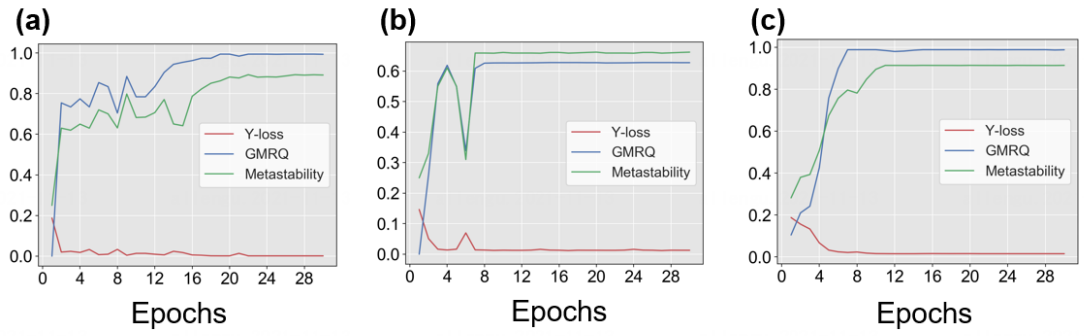

Fig. S10: The change of Y-loss, GMRQ and metastability upon optimization. The change of Y-loss, GMRQ and metastability as a function of Epoch number throughout the optimization process for the three systems under study are shown. (a) Alanine dipeptide with N=4 and the lag time of 5ps. (b) 2D potential with N=4 and the lagtime of 3,000 saving intervals. (c) RNAP with N=4 and the lag time of 90ns.