Supporting Information:

Seeding the Multi-dimensional Nonequilibrium Pulling for Hamiltonian Variation:

Indirect Nonequilibrium Free Energy Simulations at QM levels

Zhaoxi Sun[1*] and Qiaole He[2]

[1]Beijing National Laboratory for Molecular Sciences, College of Chemistry and Molecular Engineering, Institute of Theoretical and Computational Chemistry, Peking University, Beijing 100871, China

[2]AI Department of Enzymaster (Ningbo) Bio-Engineering Co., Ltd., North Century Avenue 333, 315100 Ningbo, China

*To whom correspondence should be addressed: z.sun@pku.edu.cn

Table S1. The timing information (in ns/day) of MM, SQM and QM calculations on a computing node with 384 GB memory and the CPU used is Intel(R) Xeon(R) Platinum 8160. A single core is used for this benchmark, which avoids the influence of parallelization-related issues. As the nonequilibrium trajectories are independent, the practical simulation speed is the same as these statistics. As the memory on each computing node is very large, no memory-related issue in QM calculations would influence the speed of calculation. The basis set of 6-31G* is employed in ab initio QM calculations. Different SQM simulations are of very similar computational costs, while the speed of ab initio QM calculations depends on the level of theory. Here, we only tested the HF Hamiltonian.

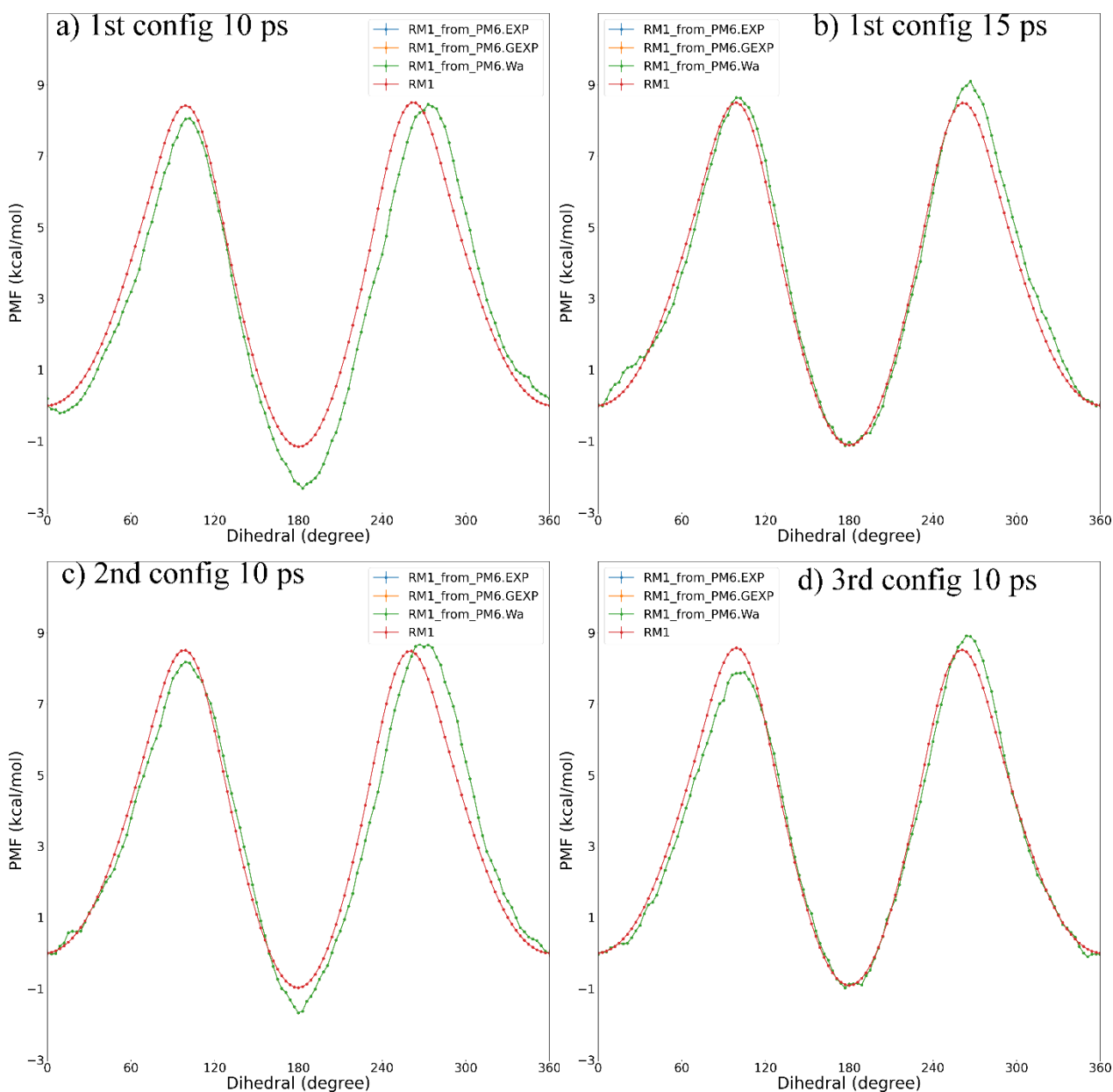| Hamiltonian Terms	SQM	HF	MM
speed(ns/day)	46.5	0.013907	1877.92
speedup(SQM/QM)	1.00	3343.64	0.02

Table S2. Mean absolute error of the indirect estimates at ab initio QM levels in kcal/mol. 7 different initial configurations are tested to initiate the multi-dimensional ASMD pulling. The pulling speed in the configurational space is 10 ps per 3° segment. The alchemical perturbation is finished in 10 time steps with a change of 0.1 per step in the first 7 trails, while the last trail 0.01 uses 100 steps.

Hamiltonians Trail	HF		
	AM1	MNDO	RM1
1	0.31	0.35	0.60
2	0.83	0.79	1.33
3	0.88	0.54	0.93
4	0.56	0.91	0.75
5	0.50	0.44	0.67
6	0.81	1.21	0.47
7	0.52	1.23	0.67
0.01	0.20	0.30	0.21

Table S3. Efficiency comparison of direct and indirect nonequilibrium free energy simulations with the current JI-based ASMD method and the previously proposed BAR-based method. The total simulation time in direct scheme is given by $N_{\text{segments}} * N_{\text{traj}} * (\phi_{\text{NEW}} + \phi_{\text{eq}})$, while the total simulation time in the indirect scheme is the sum of $N_{\text{segments,small}} * N_{\text{traj,small}} * (\phi_{\text{NEW,small}} + \phi_{\text{eq,small}})$ under the SQM Hamiltonian and $N_{\text{traj,small->large}} * (\phi_{\text{NEW,small->large}} + \phi_{\text{eq,small}}) + N_{\text{traj,large->small}} * (\phi_{\text{NEW,large->small}} + \phi_{\text{eq,large}})$ in the SQM-to-QM correction. N_{segments} is the number of segments and N_{traj} is the number of realizations per segment. The simulation time under the QM Hamiltonian is scaled by the ratio of computational cost QM/SQM in Table S1 to provide the effective simulation time at the SQM level, enabling the direct comparison between computational costs from methods. The statistics for the BAR-based method are borrowed from our previous work, Phys. Chem. Chem. Phys. 2019, 21, 21942-21959. As the BAR-based method is faster than ASMD in the construction of the free energy profiles in the configurational space, its computational cost of the direct free energy simulation (3.78 ns in the reference) is smaller than the ASMD one (24 ns). Therefore, the speedup of the multi-dimensional ASMD method is relatively modest compared with the ~1000-fold speedup shown in the previous table.

Terms / Simulation	ϕ_{eq} for each initial configuration (ps)	ϕ_{NEW} in each segment (ps)	Number of segments	Number of realizations per segment	Total simulation time (ps) scaled to SQM Hamiltonian	Relative efficiency
direct SQM	0	10	120	20	24000.00	526.62
SQM->HF	0	0.005	120	25	50154.60	-
HF->SQM	-	-	-	-	0.00	-
indirect HF	-	-	-	-	74154.60	170.44
direct HF	0.05	0.5x2=1	180	20	12638958.80	1.00

Fig. S1. Comparison between the direct and indirect estimates of the free energy profiles at the RM1 level initiated from different seeds (i.e., configurations) with different pulling speeds along the configurational CV or the alchemical CV. The configurational space is explored with the ASMD scheme with 20 samples in each stage at the PM6 level, and the RM1 result is obtained by the combination of the PM6 result and the PM6-to-RM1 unidirectional pulling. The pulling time 'x ps' denotes the pulling time for each segment along the configurational CV. The nonequilibrium transformation in the alchemical space is performed in 5 fs (i.e., 10 time steps with a change of 0.1 per step) for the first 5 subplots, while a smaller perturbation (0.01 per time step) and thus a longer pulling time is used for the last subplot. The exponential average EXP, the Gaussian approximation GEXP, and the ordinary average Wa are extremely similar, leading to overlaps of these curves in the plot. We can see that different initial configurations could lead to different systematic errors with a faster pulling speed along the alchemical CV, which could be eliminated when a slower pulling speed is employed.
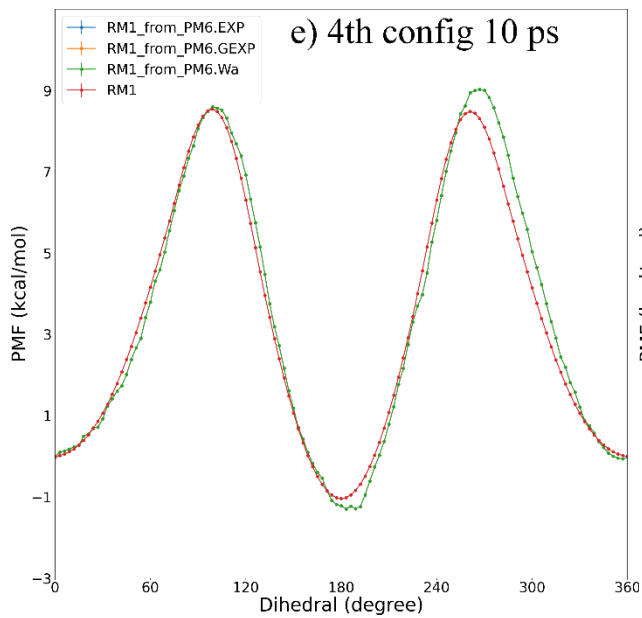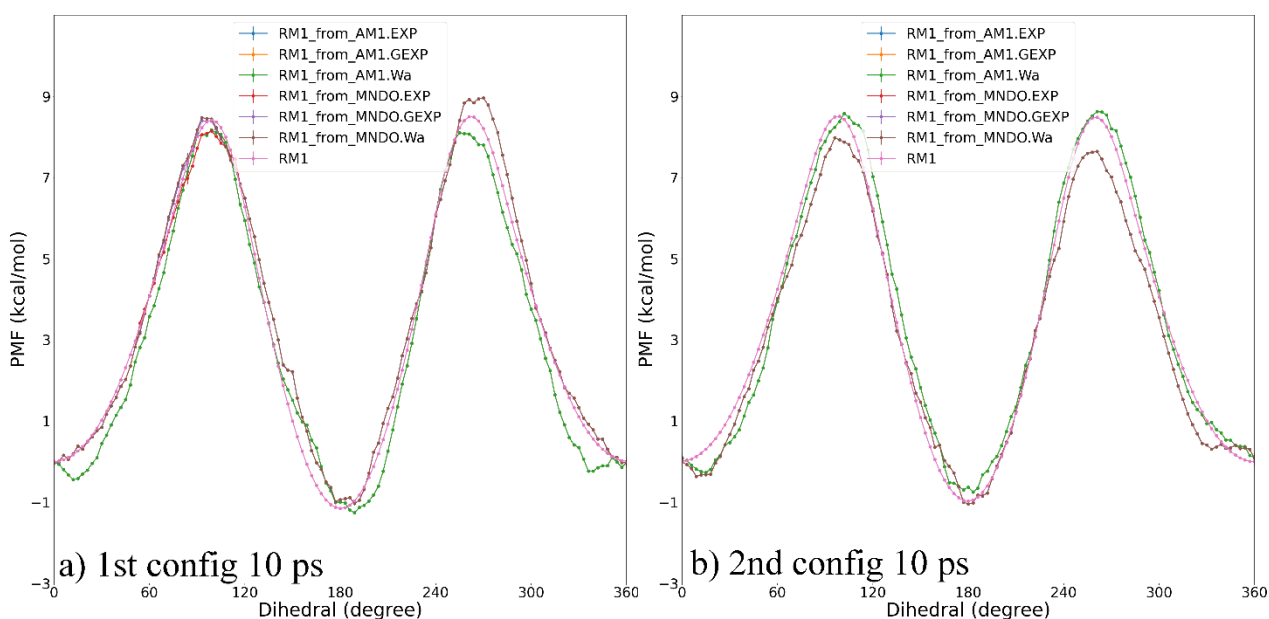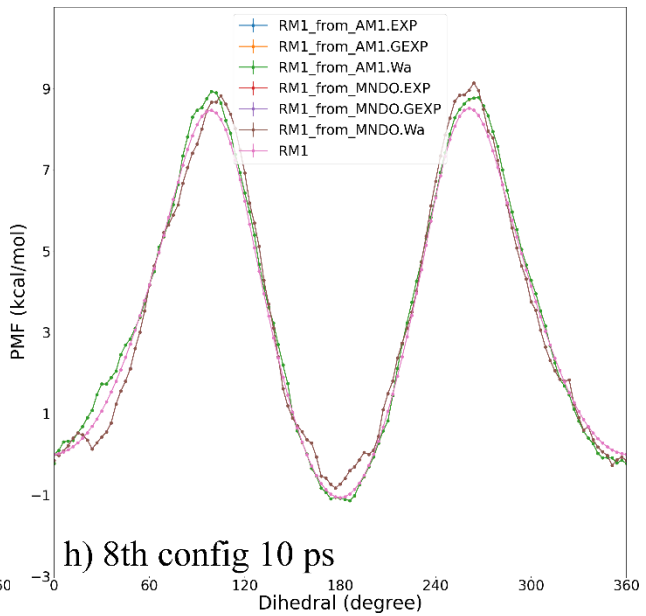
e) 4th config 10 ps

| RM1_from_PM6.EXP |
| RM1_from_PM6.GEXP |
| RM1_from_PM6.Wa |
| RM1 |

f) 1st config 15 ps 0.01

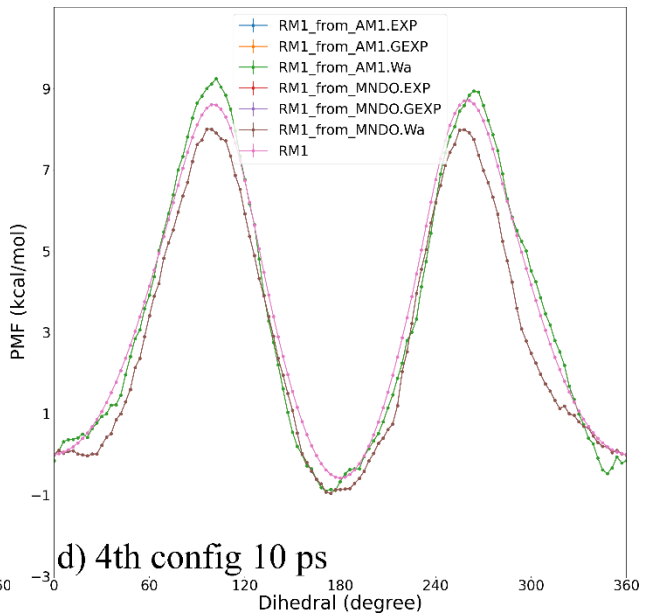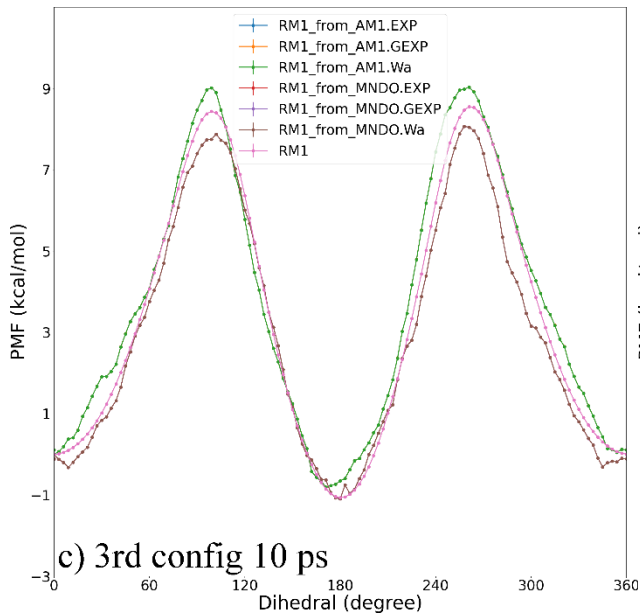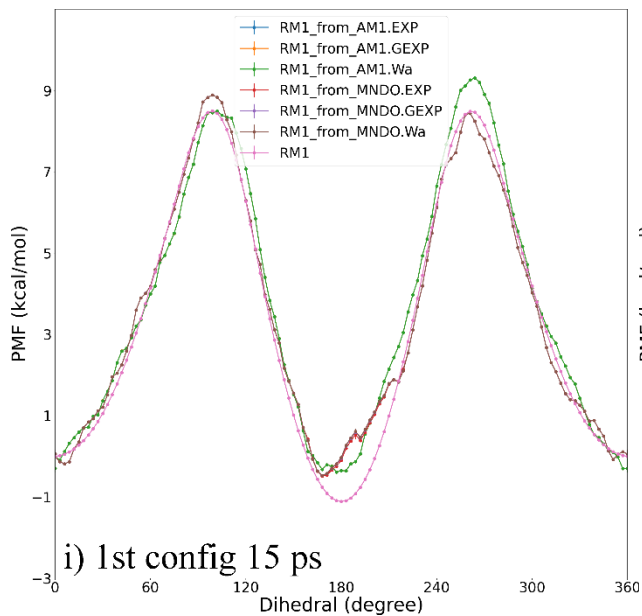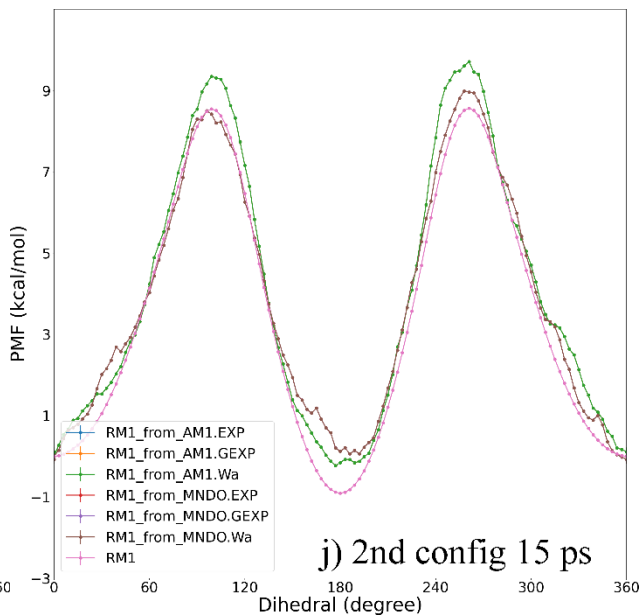| RM1_from_PM6.EXP |
| RM1_from_PM6.GEXP |
| RM1_from_PM6.Wa |
| RM1 |
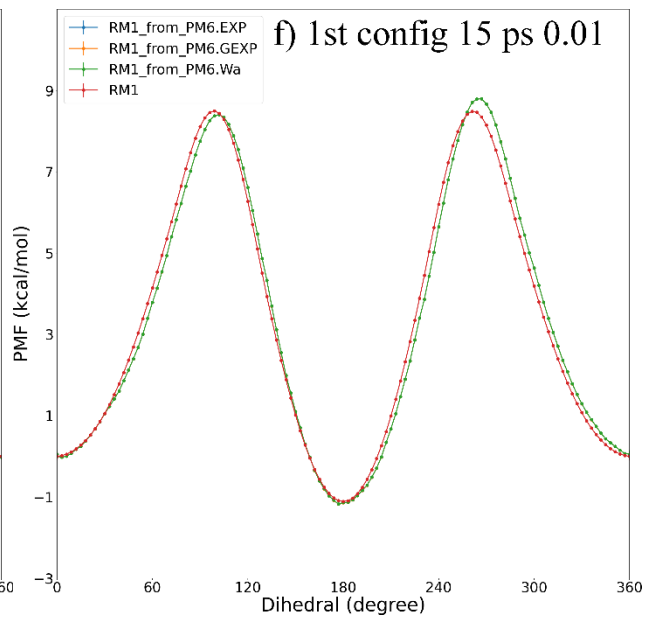
Fig. S2. Comparison between the direct and indirect estimates of the free energy profiles at the RM1 level initiated from different seeds (i.e., configurations) with different pulling speeds along the configurational CV. The nonequilibrium transformation in the alchemical space is performed in 5 fs (i.e., 10 time steps with a change of 0.1 per step). The configurational space is explored with the ASMD scheme with 20 samples in each stage under the AM1 or MNDO Hamiltonians, and the RM1 result is obtained by the combination of the AM1 or MNDO result and the AM1-to-RM1 or MNDO-to-RM1 unidirectional pulling. The pulling time 'x ps' denotes the pulling time for each segment along the configurational CV. 8 different initial configurations are used for the seeding SMD simulations with the pulling speed of 10 ps per segment in the first 8 subplots a-h), while the first 3 configurations are used to initiate the seeding SMD simulations with a slower pulling speed of 15 ps per segment for the last 3 subplots i-k). The exponential average EXP, the Gaussian approximation GEXP, and the ordinary average Wa are extremely similar, leading to overlaps of these curves in the plot. Different initial configurations introduce different systematic errors, and averaging over these configurations could eliminate this systematic error, as shown in the main article. The pulling speeds of 10 ps/segment and 15 ps/segment along the configurational CV have little influence on the systematic error of the indirect results, as the latter is mainly introduced in the alchemical perturbation term.
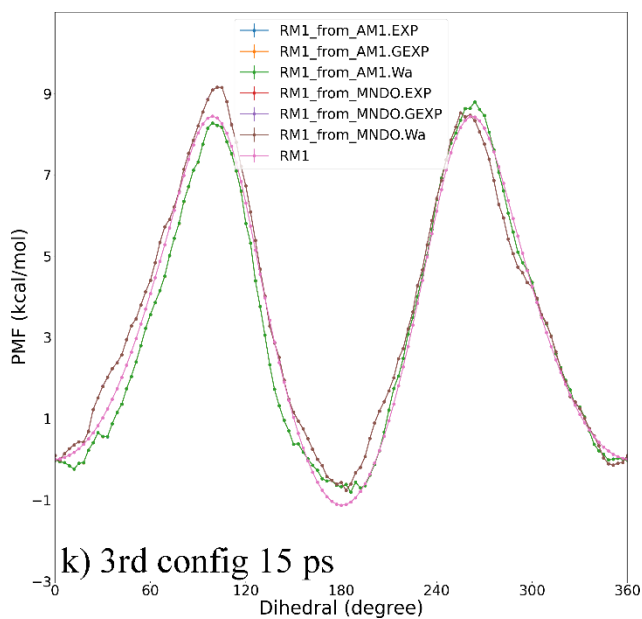
c) 3rd config 10 ps

d) 4th config 10 ps

e) 5th config 10 ps

f) 6th config 10 ps

g) 7th config 10 ps

h) 8th config 10 ps
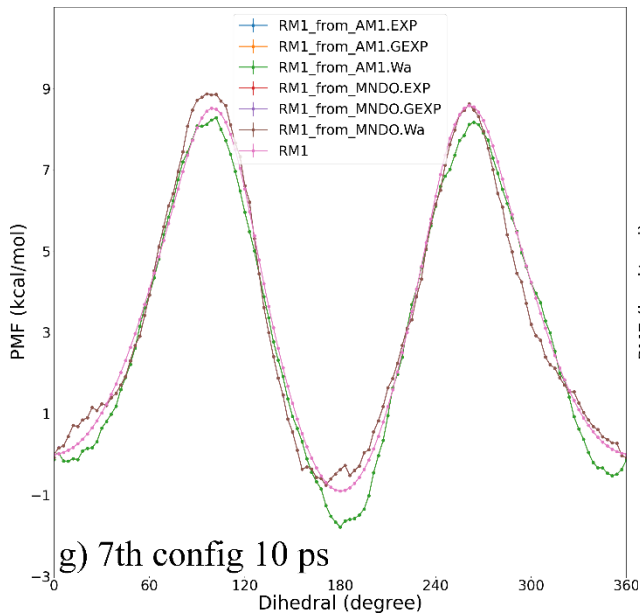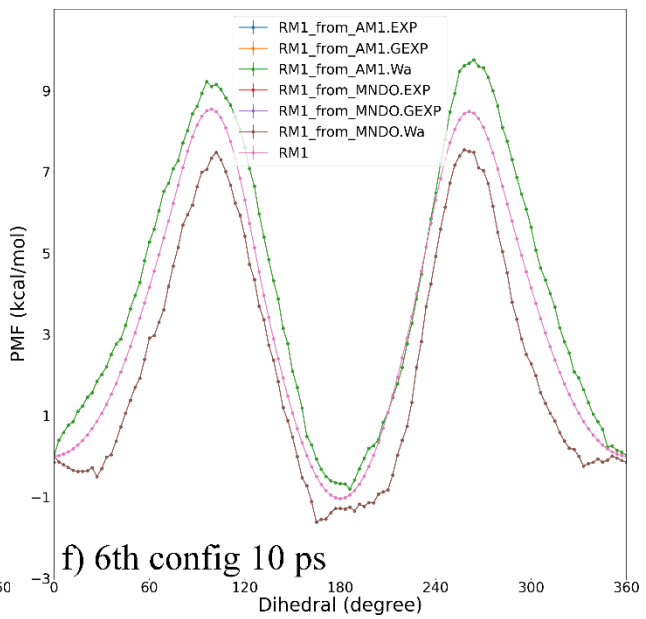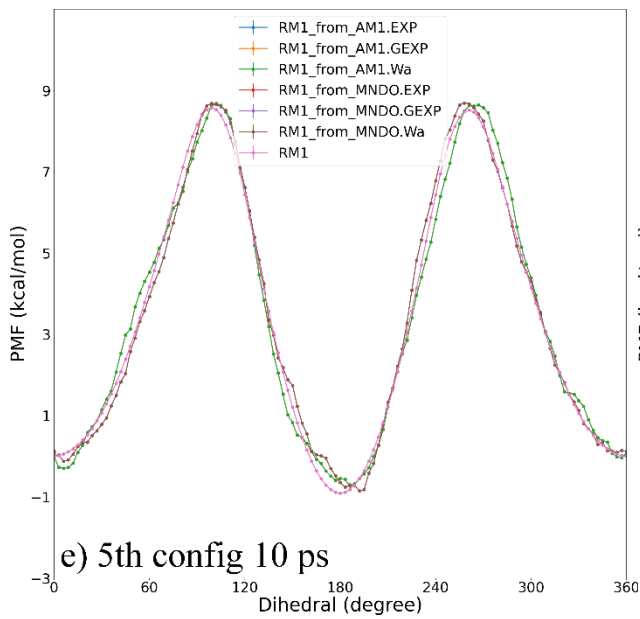
i) 1st config 15 ps

j) 2nd config 15 ps

k) 3rd config 15 ps

Fig. S3. Comparison between the direct and indirect estimates of the free energy profiles at the RM1 level initiated from different seeds (i.e., initial configurations) with different pulling speeds along the configurational CV. The nonequilibrium transformation in the alchemical space is performed in 50 fs (i.e., 100 time steps with a change of 0.01 per step). The configurational space is explored with the ASMD scheme with 20 samples in each stage under the AM1 or MNDO Hamiltonians, and the RM1 result is obtained by combining the AM1 or MNDO result and the AM1-to-RM1 or MNDO-to-RM1 unidirectional pulling. The pulling time 'x ps' denotes the pulling time for each segment along the configurational CV. Different initial configurations are used for the seeding SMD simulations with the pulling speed of 10 ps per segment for the first 4 subplots a-d), while the first 2 configurations are used to initiate the seeding SMD simulations with a slower pulling speed of 15 ps per segment for the last 2 subplots e-f). The exponential average EXP, the Gaussian approximation GEXP, and the ordinary average Wa are extremely similar, leading to overlaps of these curves in the plot. Compared with the indirect estimates obtained with a faster pulling speed along the alchemical CV shown in the previous figure, the current results are much closer to the direct free energy estimates, which indicates that the sampling in the nonequilibrium ensemble successfully eliminates the systematic bias introduced in the alchemical perturbation term. The pulling speed in the configurational space have little impact on the outcome.
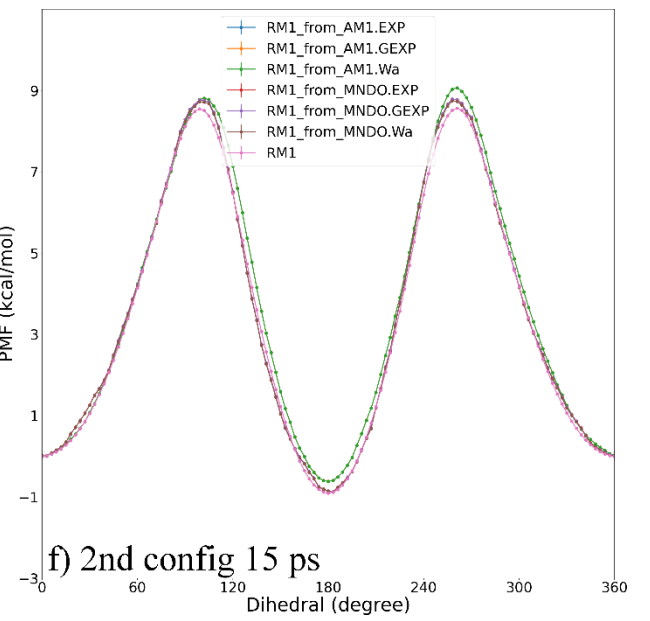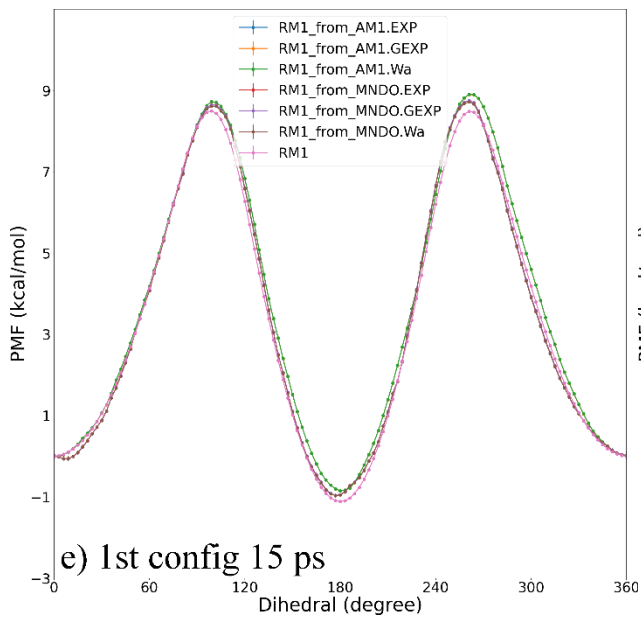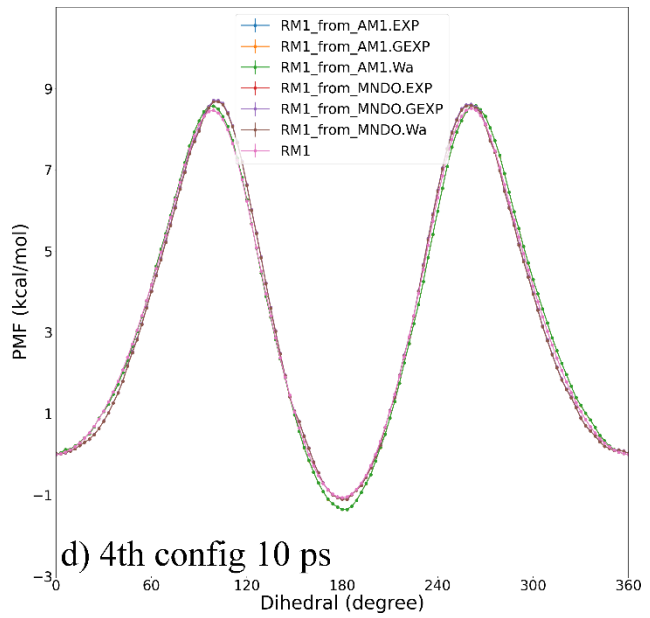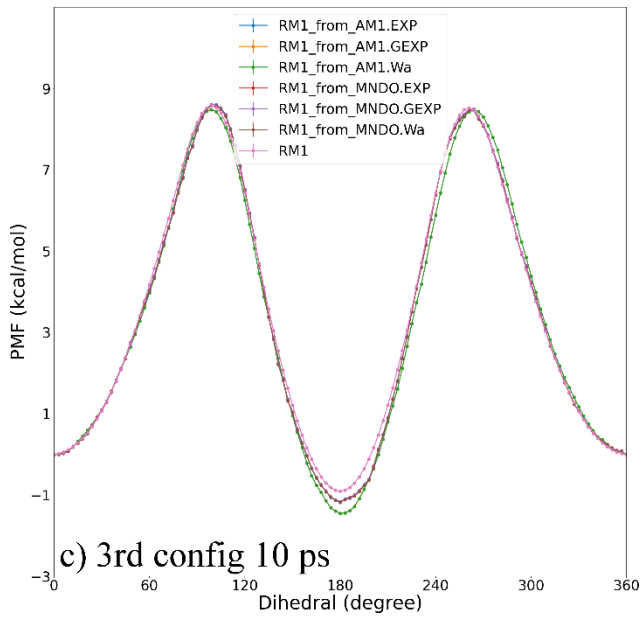
c) 3rd config 10 ps

d) 4th config 10 ps

e) 1st config 15 ps

f) 2nd config 15 ps
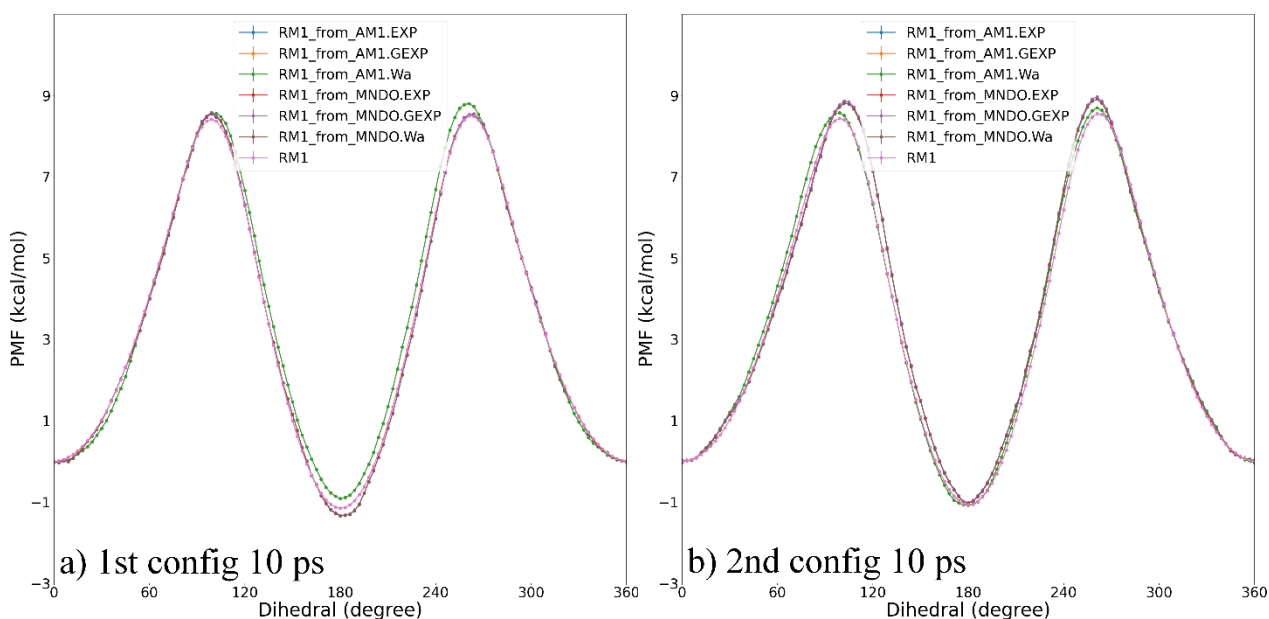
Fig. S4. The deviation of the indirect EXP, GEXP and Wa estimates from the direct result at the RM1 level for the initial-seed-averaged simulations with 100 steps in the alchemical transformation.
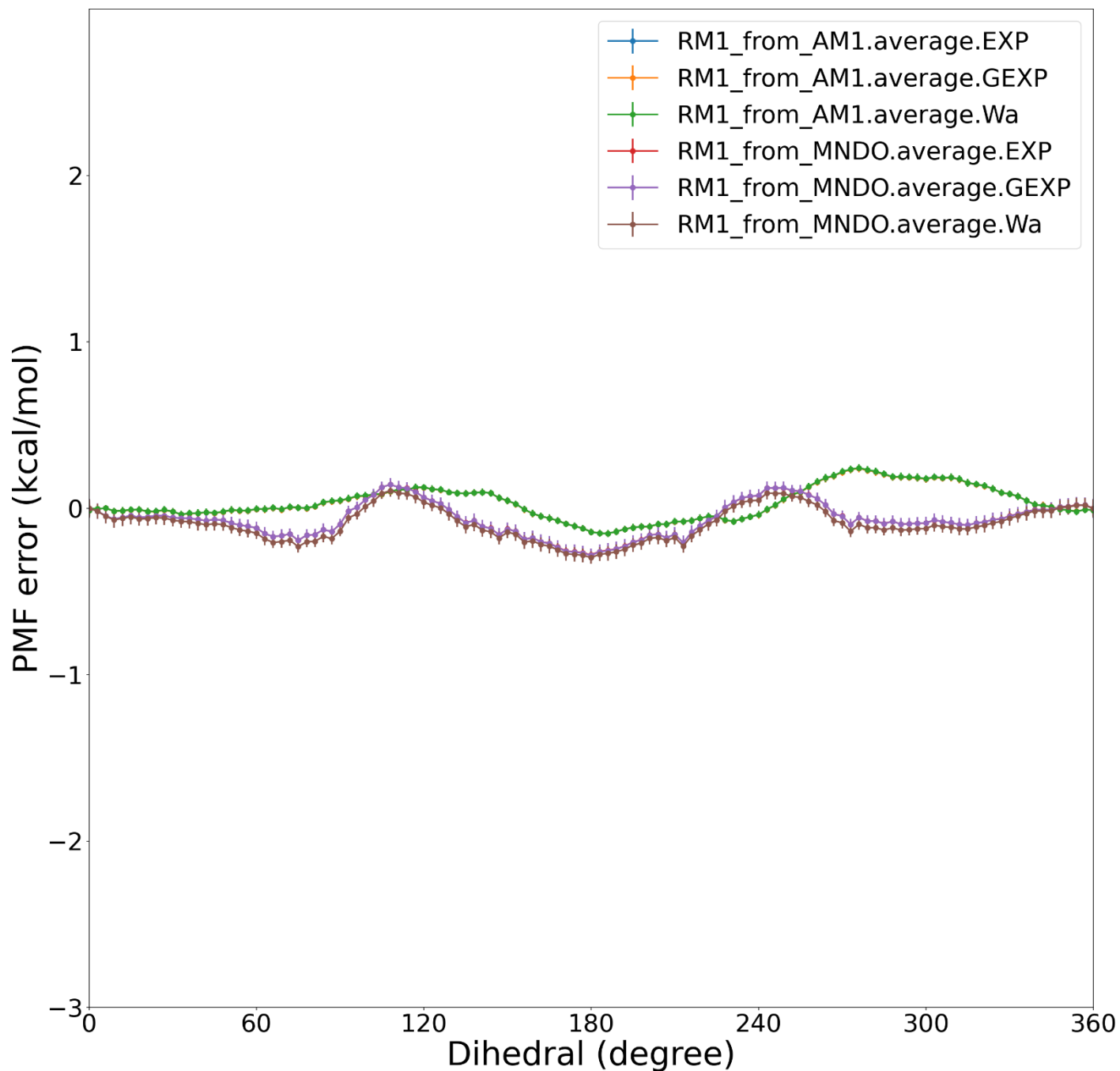
Fig. S5. The correlations between the MM and SQM (RM1) energetics calculated from 10 ns trajectories generated at 1300 K and 300 K. The sampling interval is 2 ps and there are 5000 independent configurations in total. The RMSE and MAE of the original parameter set and the newly fitted force-matching set are also presented. The RMSE of atomic forces is also improved from 23.9 kcal/(mol·Å·atom) to 13.1 kcal/(mol·Å·atom) at 300 K and from 37.6 kcal/(mol·Å·atom) to 25.0 kcal/(mol·Å·atom) at 1300 K.
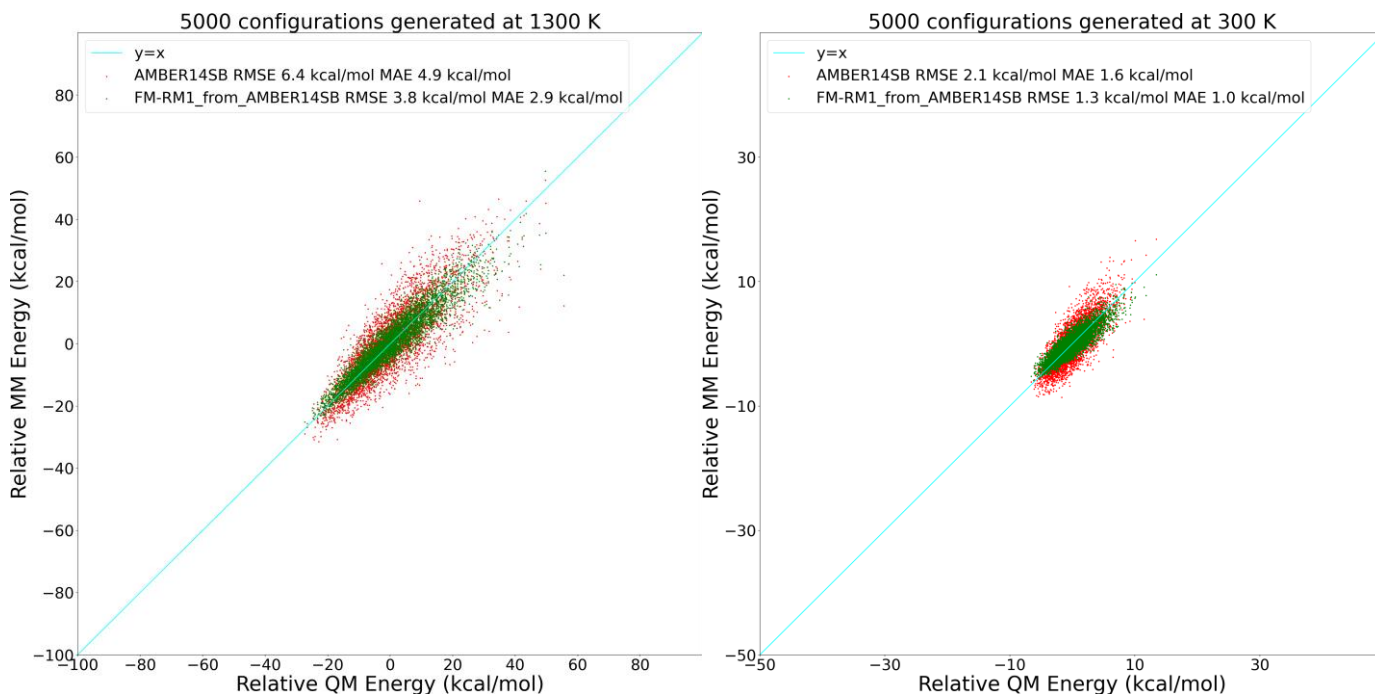
Fig. S6. Free energy profiles of ammonia inversion from direct simulations under the AM1, MNDO and PM6 Hamiltonians with different pulling times for each 3º segment. The number in the legend represents the number of time steps for each nonequilibrium trajectory in each segment. As we are using 0.5 fs time steps, the tested pulling times for each segment include 0.25 ps, 1 ps, 2 ps, 5 ps, and 10 ps. Similar to the PM3 case, the pulling speed of 5 ps per segment is sufficiently slow for converged estimates of the free energy profiles.
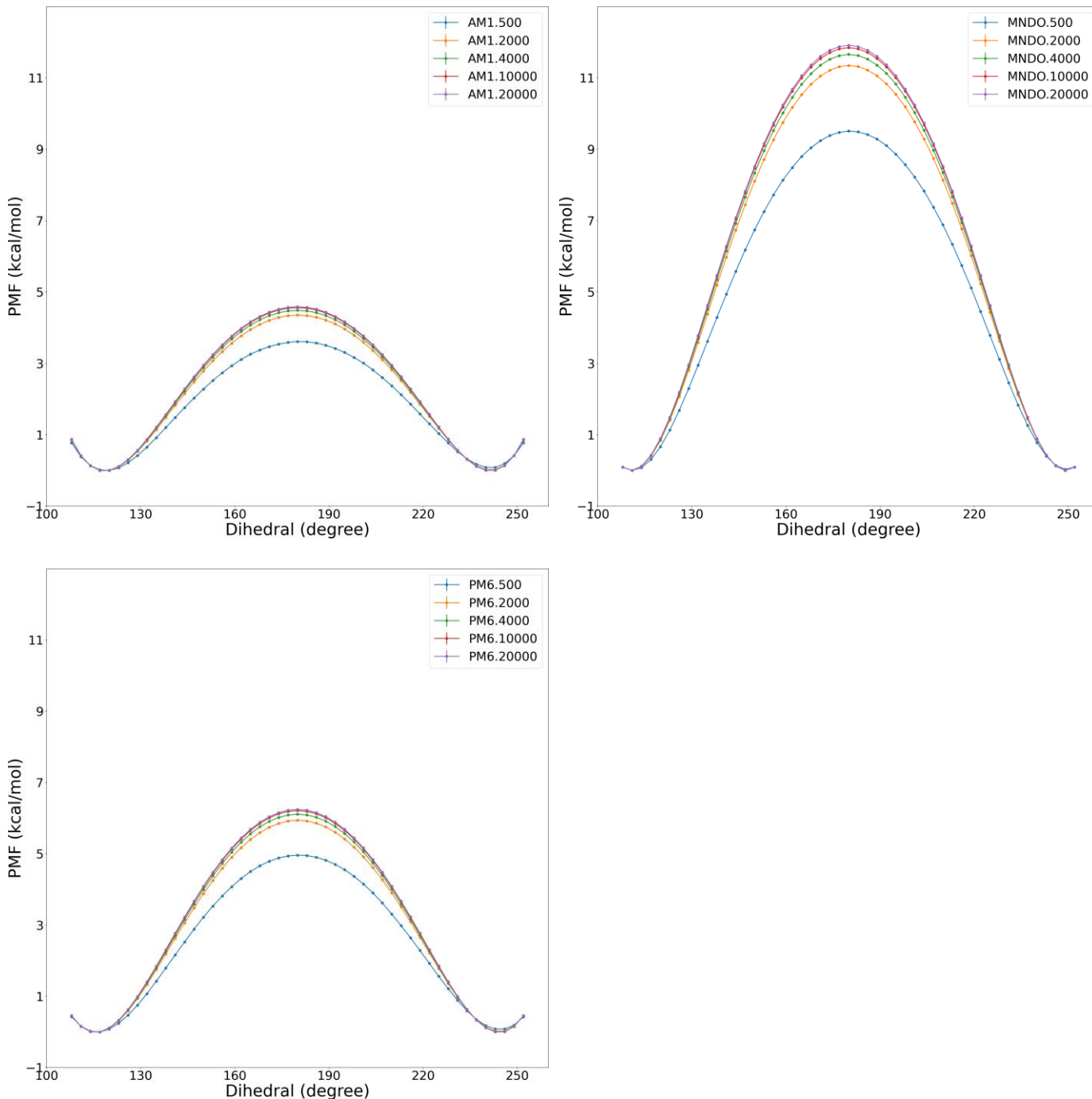
Fig. S7. Comparison between direct and indirect results at the PM3 level. Four new initial seeds are used to spawn the whole multi-dimensional ASMD pulling.