Supplementary Information

ArcaNN: automated enhanced sampling generation of training sets for chemically reactive machine learning interatomic potentials

Rolf David,* Miguel de la Puente, Axel Gomez, Olaia Anton, Guillaume Stirnemann,* and Damien Laage*

PASTEUR, Département de Chimie, École Normale Supérieure, PSL University, Sorbonne Université, CNRS, 75005 Paris

E-mail: rolf.david@ens.psl.eu; guillaume.stirnemann@ens.psl.eu; damien.laage@ens.psl.eu

Contents

Details on the initial *ab initio* MD simulations for the S_N2 reaction

The first step consisted on the preparation of the initial datasets by generating reactant structures, by *ab initio* MD simulations. In order to perform the MD simulations, we have to construct an initial structure of the system. A 15 $\mathring{A}^3$ cubic box was constructed with packmol,[1] containing one bromide ion, one chloromethane molecule and 38 acetonitrile molecules. An energy minimization of the system was then performed using the Amber22 software. The AMBER's built in General Forcefield (GAFF)[2] parameters were used for the chloromethane, the acetonitrile and the bromide ion, with the AM1-BCC charge model was used to generate atomic charges Next, the system was heated to a temperature of 300 K in the NVT ensemble for 20 ps, and then equilibrated in the NPT ensemble for 200 ps and at a pressure of 1 bar and a temperature of 300 K, both with a timestep of 2 fs. From this equilibration, 20 snapshots were extracted, with ten of them having their bromine and chlorine atoms swapped (to generate the product structures). With these initial structures, *ab initio* MD simulations were performed with the CP2K software[3] at the BLYP[4,5] level of theory and with the D3 dispersion correction.[6] The DZVP-MOLOPT-SR[7,8] basis set was used in conjunction with the GTH pseudopotentials.[9–11] Each run was performed within the NVT ensemble at 300 K with a timestep of 0.5 fs for 2 ps. The temperature control was enabled by the use of a CSVR thermostat[12] with a time constant of 0.1 ps^{-1}.

Timings of the ArcaNN training

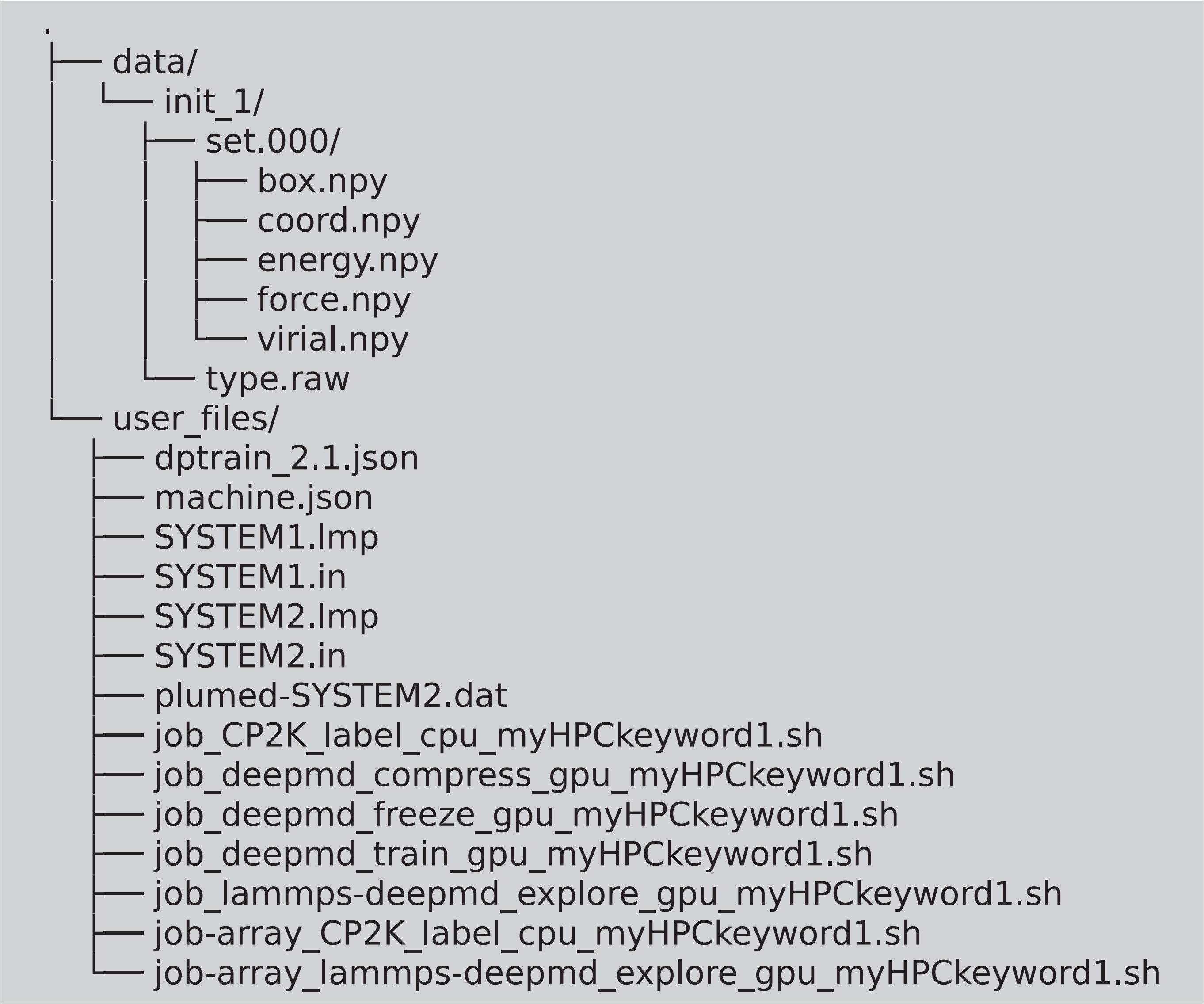
Table S1: Summary of Timings for the initial aiMD, the training, exploration, and labeling Phases for the S_N2 reaction

Phase	**Hardware Used**	**Average Time per Cycle**	**Total Time**
Initial aiMD	AMD EPYC 7H12	-	26897.4 core.hours
Training	Nvidia V100 SXM2	14.67 gpu.hours	190.76 gpu.hours
Exploration	Nvidia V100 SXM2	59.04 gpu.hours	767.57 gpu.hours
Labeling	Intel Cascade Lake 6248	1836.79 core.hours	23878.21 core.hours

Table S2: Summary of Timings for the initial aiMD, the training, exploration, and labeling Phases for the Diels-Alder reaction

Phase	**Hardware Used**	**Average Time per Cycle**	**Total Time**
Initial aiMD	Cascade Lake 6248	-	1928.72 core.hours
Training	Nvidia A100 SXM4	3.58 gpu.hours	32.24 gpu.hours
Exploration	Nvidia V100 SXM2	27.81 gpu.hours	250.33 gpu.hours
Labeling	Intel Cascade Lake 6248	297.56 core.hours	2380.45 core.hours

User-provided tree folder structure



```
.
├── data/
│   └── init_1/
│       ├── set.000/
│       │   ├── box.npy
│       │   ├── coord.npy
│       │   ├── energy.npy
│       │   ├── force.npy
│       │   └── virial.npy
│       └── type.raw
└── user_files/
    ├── dptrain_2.1.json
    ├── machine.json
    ├── SYSTEM1.lmp
    ├── SYSTEM1.in
    ├── SYSTEM2.lmp
    ├── SYSTEM2.in
    ├── plumed-SYSTEM2.dat
    ├── job_CP2K_label_cpu_myHPCkeyword1.sh
    ├── job_deepmd_compress_gpu_myHPCkeyword1.sh
    ├── job_deepmd_freeze_gpu_myHPCkeyword1.sh
    ├── job_deepmd_train_gpu_myHPCkeyword1.sh
    ├── job_lammps-deepmd_explore_gpu_myHPCkeyword1.sh
    ├── job-array_CP2K_label_cpu_myHPCkeyword1.sh
    └── job-array_lammps-deepmd_explore_gpu_myHPCkeyword1.sh
```

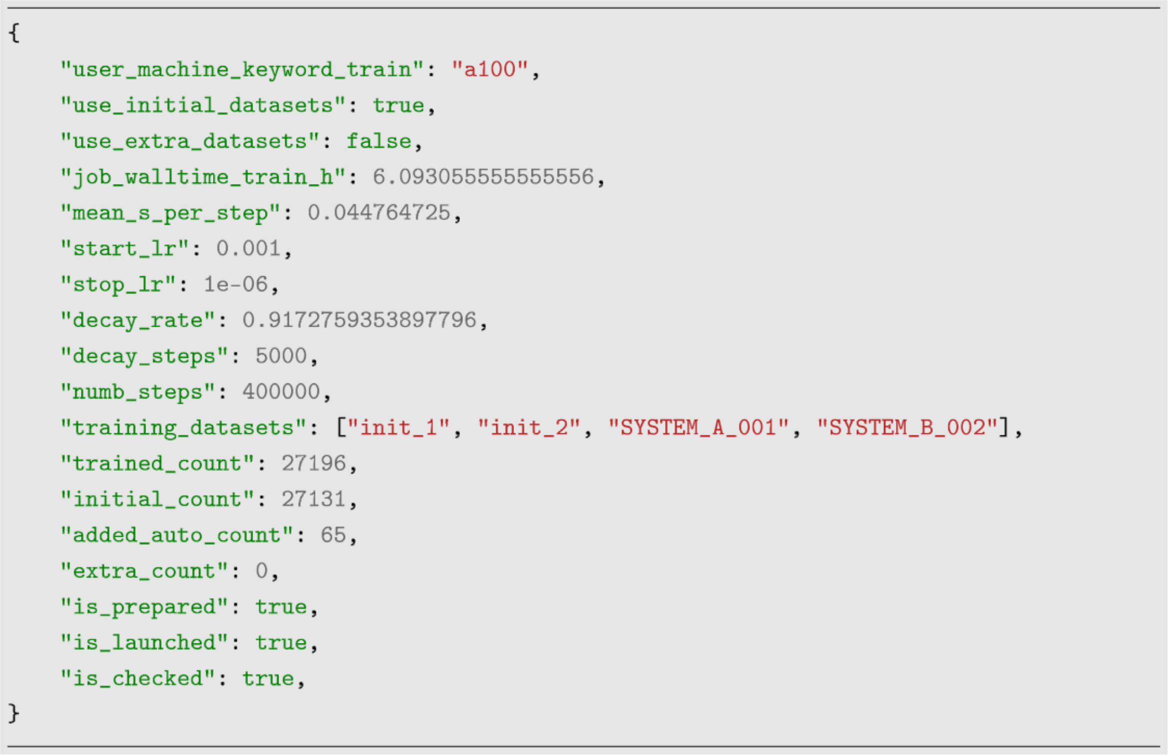
Figure S1: Example of the tree folder structure used by ArcaNN.

Machine JSON user file used by ArcaNN

```json
{
    "myHPCkeyword1":
    {
        "hostname": "myHPC1",
        "walltime_format": "hours",
        "job_scheduler": "slurm",
        "launch_command": "sbatch",
        "max_jobs" : 200,
        "max_array_size" : 500,
        "mykeyword1": {
            "project_name": "myproject",
            "allocation_name": "myallocationcpu",
            "arch_name": "cpu",
            "arch_type": "cpu",
            "partition": "mypartitioncpu",
            "subpartition": null,
            "qos": {"mypartitioncpu1": 72000, "mypartitioncpu2": 360000},
            "valid_for":
            ↪  ["labeling","freezing","compressing","exploration","test","training"],
            "default":
            ↪  ["labeling","freezing","compressing","exploration","test","training"]
        }
    }
}
```

Figure S2: Example of a *machine.json* file for configuring HPC resources in ArcaNN.

JSON control files written by ArcaNN



```
{
    "user_machine_keyword_train": "a100",
    "use_initial_datasets": true,
    "use_extra_datasets": false,
    "job_walltime_train_h": 6.093055555555556,
    "mean_s_per_step": 0.044764725,
    "start_lr": 0.001,
    "stop_lr": 1e-06,
    "decay_rate": 0.9172759353897796,
    "decay_steps": 5000,
    "numb_steps": 400000,
    "training_datasets": ["init_1", "init_2", "SYSTEM_A_001", "SYSTEM_B_002"],
    "trained_count": 27196,
    "initial_count": 27131,
    "added_auto_count": 65,
    "extra_count": 0,
    "is_prepared": true,
    "is_launched": true,
    "is_checked": true,
}
```

Figure S3: A pruned control training JSON file.

```json
{
    "user_machine_keyword_exp": "v100",
    "deepmd_model_version": 2.1,
    "nnp_count": 3,
    "systems_auto": {
        "SYSTEM_A": {
            "nb_steps": 320000,
            "print_every_x_steps": 3200,
            "nb_atm": 790,
            "exploration_type": "lammps",
            "traj_count": 2,
            "temperature_K": 298.15,
            "timestep_ps": 0.0005,
            "previous_start": true,
            "print_interval_mult": 0.01,
            "max_exp_time_ps": 400,
            "completed_count": 6,
            "mean_s_per_step": 0.0074279635416666665,
            "max_candidates": 50,
            "sigma_low": 0.2,
            "sigma_high": 0.7,
            "sigma_high_limit": 1.0,
            "ignore_first_x_ps": 0.5,
            "mean_deviation_max_f": 0.12157408506666667,
            "total_count": 600,
            "candidates_count": 13,
            "rejected_count": 0,
            "selected_count": 13,
            "discarded_count": 0
        }
    },
    "is_locked": true,
    "is_launched": true,
    "is_checked": true,
    "is_deviated": true,
    "is_extracted": true,
    "nb_sim": 156
}
```

Figure S4: A pruned control exploration JSON file.

```json
{
    "labeling_program": "cp2k",
    "user_machine_keyword_label": "cpu",
    "systems_auto": {
        "SYSTEM_A": {
            "walltime_first_job_h": 0.5,
            "walltime_second_job_h": 0.5,
            "nb_nodes": 1,
            "nb_mpi_per_node": 10,
            "nb_threads_per_mpi": 1,
            "candidates_count": 5,
            "disturbed_candidates_count": 0,
            "timings_s": [382.4943999999999, 598.353],
            "candidates_skipped_count": 0,
        },
        "total_to_label": 126,
        "launch_all_jobs": true,
        "is_locked": true,
        "is_launched": true,
        "is_checked": true,
        "is_extracted": true
    }
}
```

Figure S5: A pruned control labeling JSON file.

```
{
    "user_machine_keyword_test": "v100",
    "job_email": "",
    "job_walltime_h": 2.0,
    "is_compressed": true,
    "deepmd_model_version": 2.1,
    "graph_1_002_compressed": {
        "SYSTEM_A": {
            "energy_rmse": 0.3445179,
            "energy_rmse_per_atom": 0.0004360987,
            "force_rmse": 0.0486603,
            "virial_rmse": 11.81919,
            "virial_rmse_per_atom": 0.014961,
            "number_of_test_data": 2.0,
            "trained": true
        },
    }
}
```

Figure S6: A pruned control testing JSON file.

Evolution of the candidate and rejected structures with exploration time per reactive exploration iteration for the S_N2 reaction

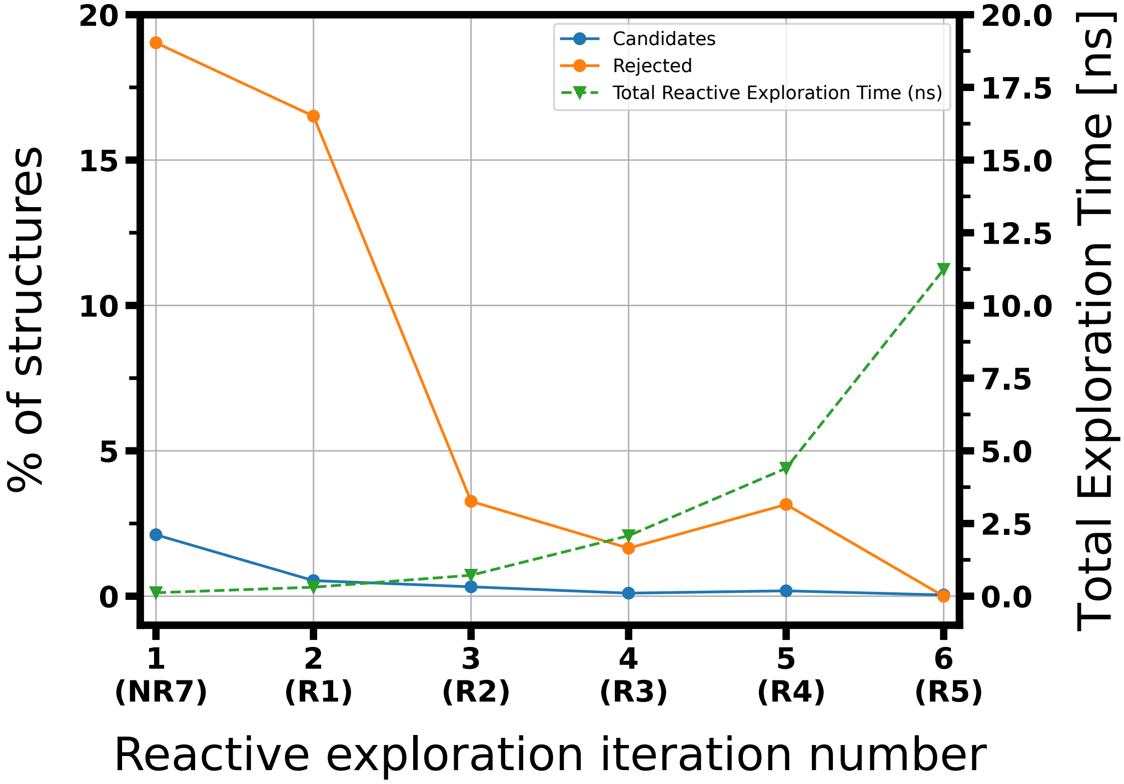


Figure S7: Percentage of candidate structures (solid blue line), rejected structures (solid orange line), and total exploration time (dashed green line) for each reactive exploration step, with the associated training dataset name in parentheses.

Validation of the $R5$ (production) NNP for the S_N2 reaction

In Figure S8, we report the component-wise force RMSE and the maximum component-wise force error along δd on the test dataset.

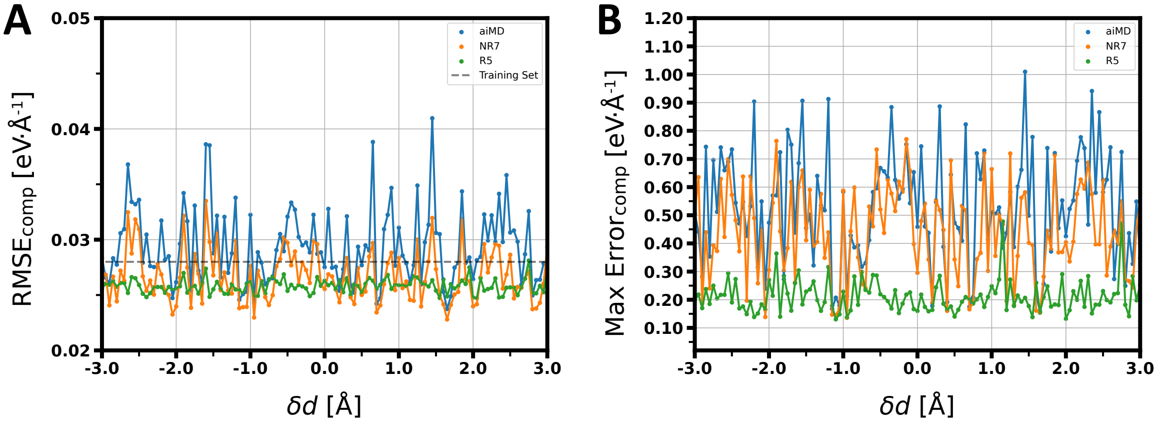


Figure S8: For three generation of NNPs, $aiMD$ (blue), $NR7$ (orange) and $R5$ (green): (A) the component-wise force RMSE along δd on the test dataset. (B) the maximum component-wise force error along δd on the test dataset.

In Figure S9, we report the probability density of the component-wise force errors and the probability density of the magnitude per atom force errors on the training and test datasets for the last ($R5$) reactive cycle. The RMSE on component-wise forces for training dataset is equal to $0.028\,\text{eV}\cdot\text{Å}^{-1}$ and for test dataset is equal to $0.026\,\text{eV}\cdot\text{Å}^{-1}$, while the RMSE on the magnitude per atom force errors for training dataset is equal to $0.048\,\text{eV}\cdot\text{Å}^{-1}$ and for test dataset is equal to $0.045\,\text{eV}\cdot\text{Å}^{-1}$.

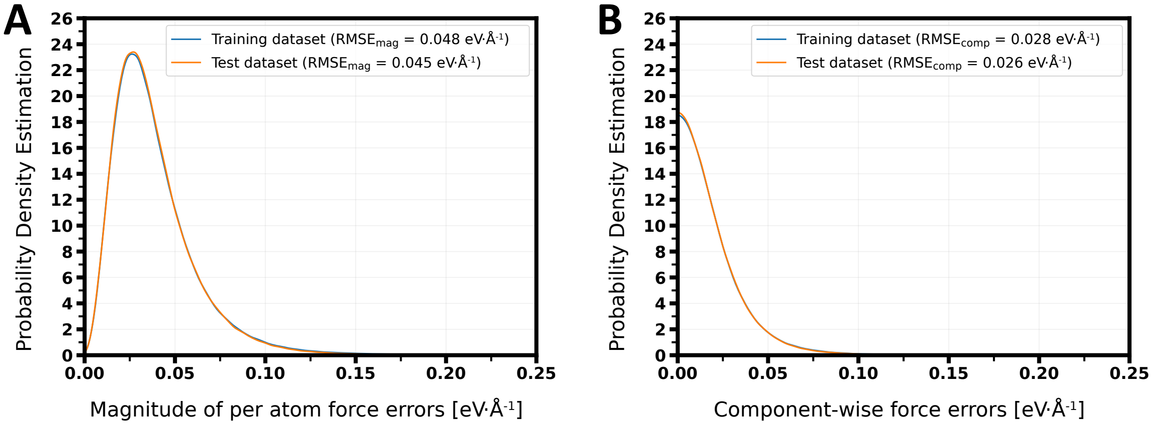


Figure S9: (A) Probability density of the magnitude of per-atom force errors on the training dataset and the test dataset. (B) Probability density of the component-wise force errors on the training dataset and the test dataset.

Free energy profiles and CV for the $NR2$ and $NR3$ NNPs for the S_N2 reaction

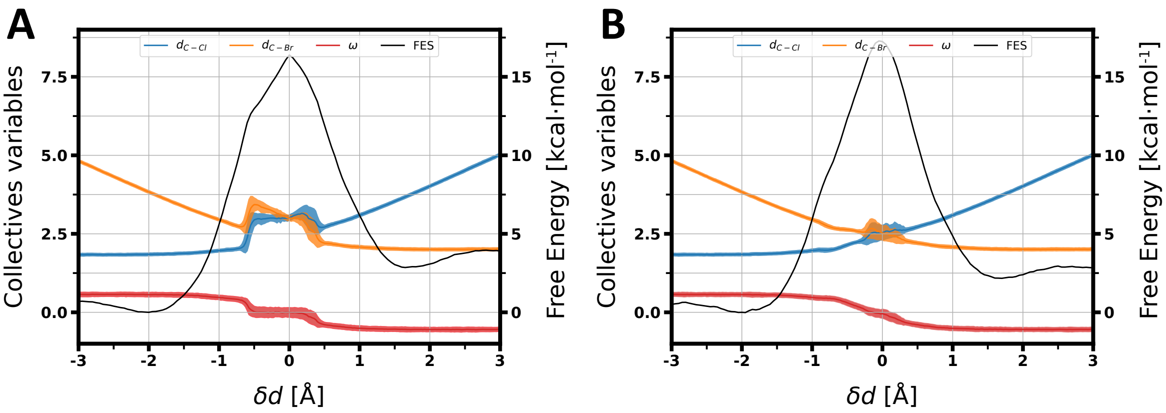


Figure S10: Free energy surface obtained from the Umbrella Sampling simulations (black) and the average value of the collective variables (as well as the 95% confidence interval in shaded color); (A) with the NNP trained on the $NR2$ dataset. (B) with the NNP trained on the $NR3$ dataset.

Validation of the $R8$ (production) NNP for the Diels-Alder reaction

In Figure S11, we report the probability density of the magnitude per atom force errors on the training and test datasets for the last ($R8$) reactive cycle. The RMSE on component-wise forces for training dataset is equal to $0.070\,\mathrm{eV}\cdot\text{Å}^{-1}$ and for test dataset is equal to $0.071\,\mathrm{eV}\cdot\text{Å}^{-1}$ (see Figure 7C), while the RMSE on the magnitude per atom force errors for training dataset is equal to $0.122\,\mathrm{eV}\cdot\text{Å}^{-1}$ and for test dataset is equal to $0.123\,\mathrm{eV}\cdot\text{Å}^{-1}$.

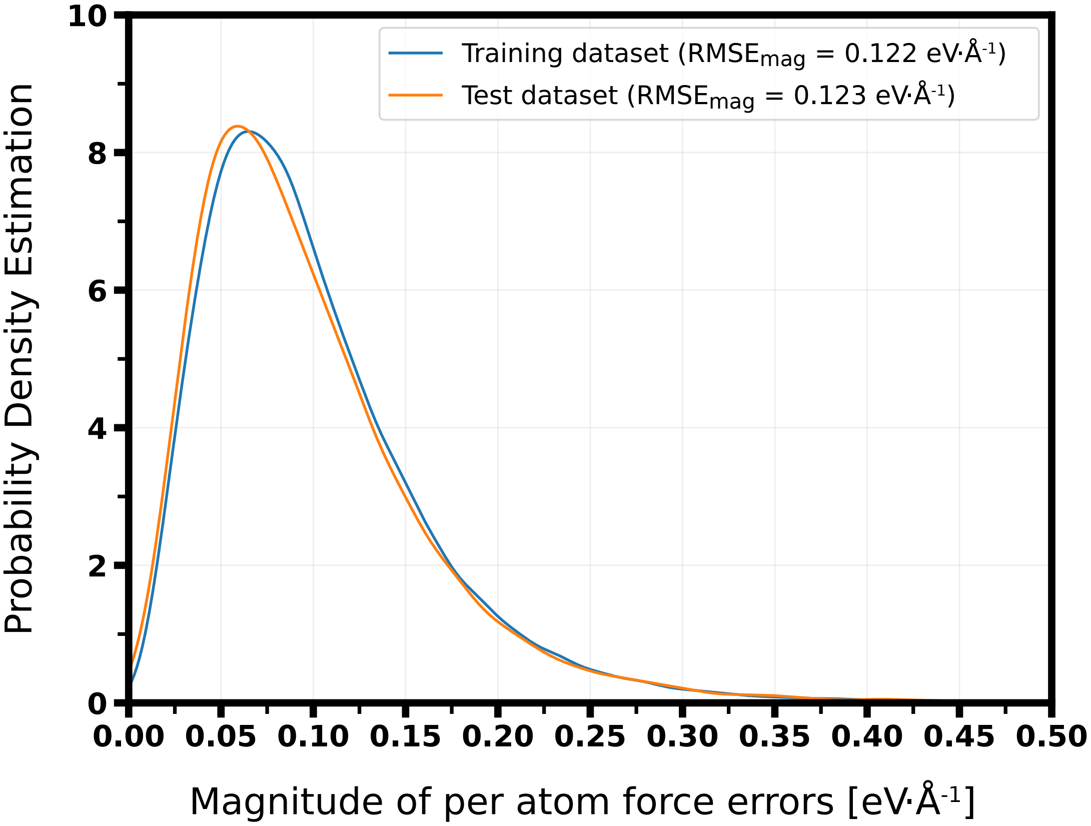


Figure S11: Probability density of the magnitude of per-atom force errors on the training dataset and the test dataset.

In Figure S12, we report the component-wise force RMSE and the maximum component-wise force error along $\bar{d}$ on the test dataset.

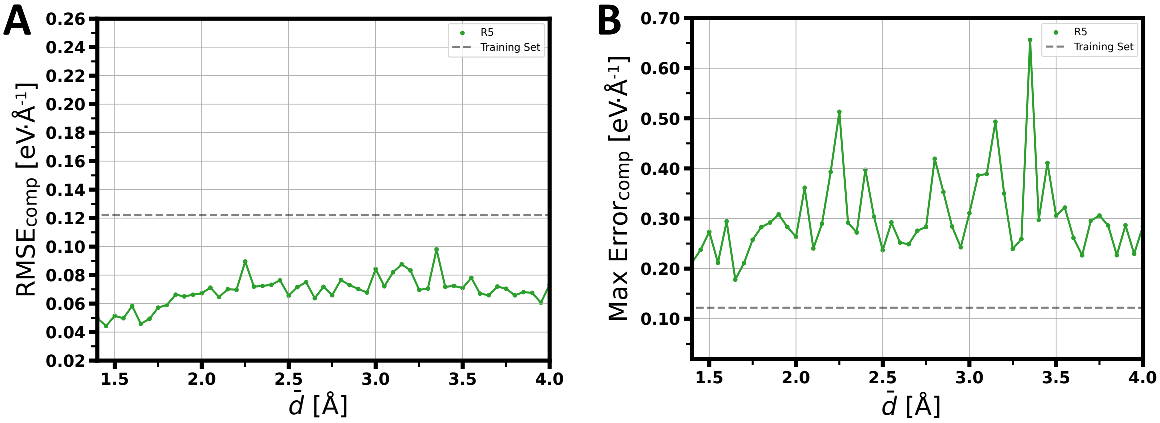


Figure S12: (A) the component-wise force RMSE along δd on the test dataset for the $R8$ NNP. (B) the maximum component-wise force error along δd on the test dataset for the $R8$ NNP.

Joint density distribution of the two main distances for the Diels-Alder reaction

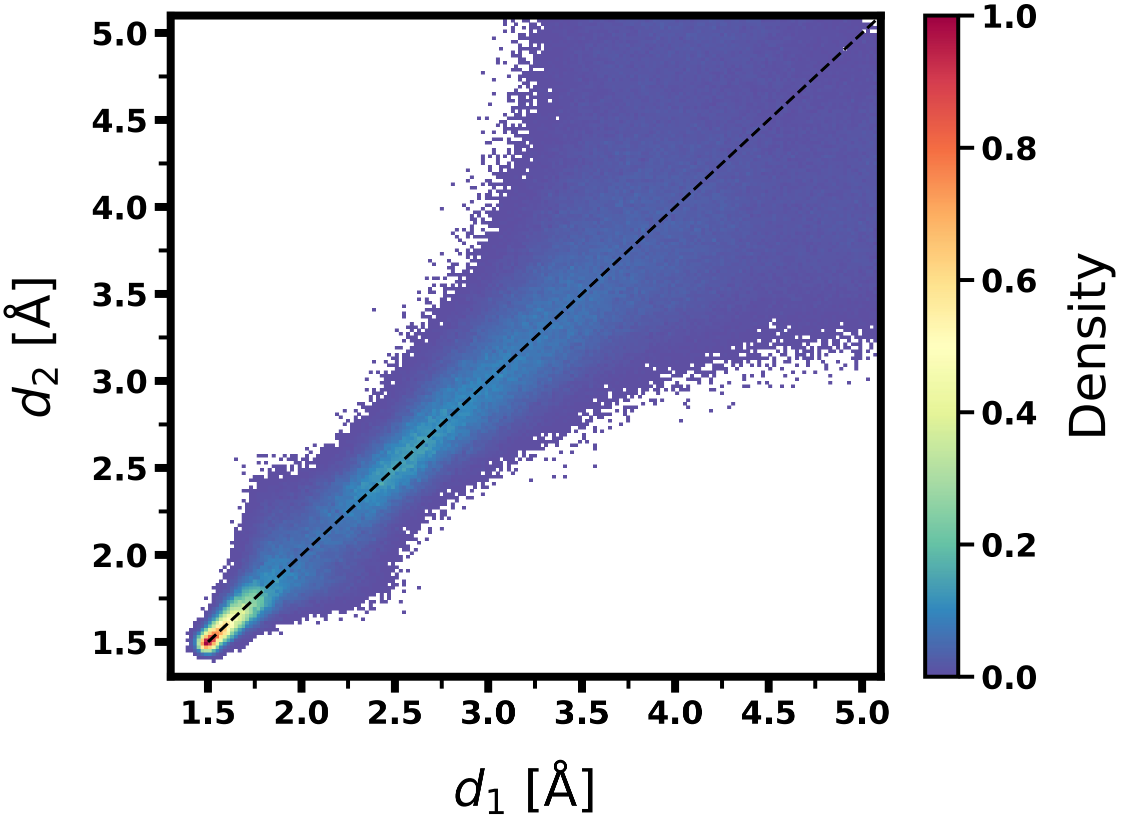


Figure S13: Joint density distribution of the distance d_1 and the distance d_2, with the dotted line representing $d_1 = d_2$, in the Umbrella Sampling simulations.

Joint density distribution of the two main distances in each important training dataset for the S_N2 reaction

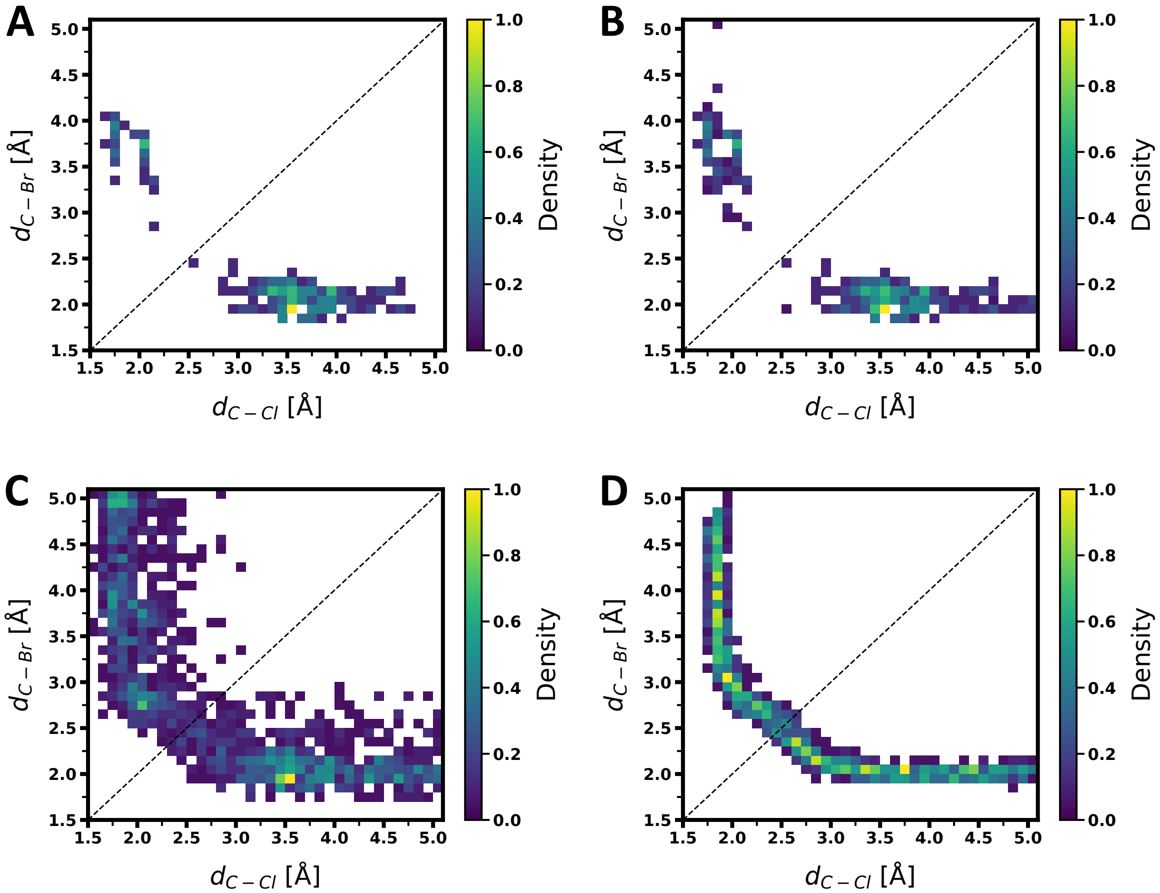


Figure S14: Joint density distribution of the distance d_{C-Cl} and the distance d_{C-Br}, with the dotted line representing $\delta d = 0\,\mathring{A}$ of the structures; (A) in the $aiMD$ dataset. (B) in the $NR7$ dataset. (C) in the $R5$ dataset. (D) in the test dataset.

OPES 1D free energy profile and CV from the $R5$ NNP for the S_N2 reaction

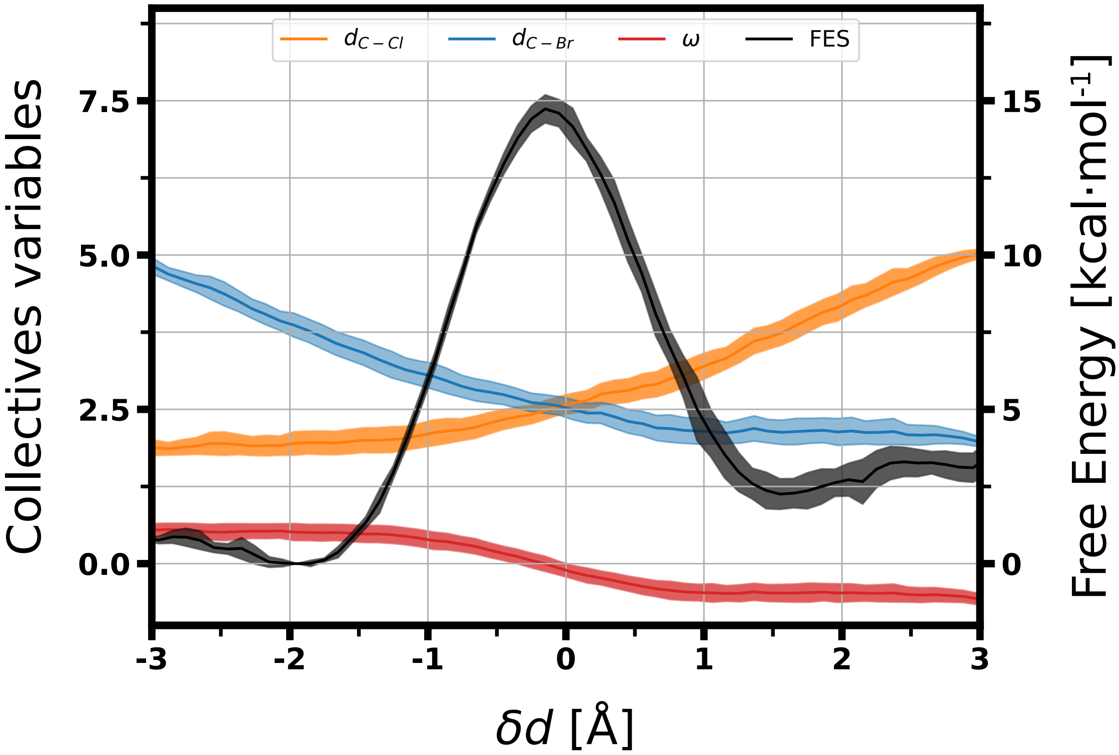


Figure S15: Free energy surface obtained from the OPES simulation (with the NNP trainined on the $R5$ dataset) (black) and the average value of the collective variables (as well as the 95% confidence interval in shaded color)

References

(1) Martínez, L.; Andrade, R.; Birgin, E. G.; Martínez, J. M. PACKMOL: A Package for Building Initial Configurations for Molecular Dynamics Simulations. *J. Comput. Chem.* **2009**, *30*, 2157–2164.

(2) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(3) Kühne, T. D. et al. CP2K: An Electronic Structure and Molecular Dynamics Software Package - Quickstep: Efficient and Accurate Electronic Structure Calculations. *J. Chem. Phys.* **2020**, *152*, 194103.

(4) Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic Behavior. *Phys. Rev. A* **1988**, *38*, 3098–3100.

(5) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B* **1988**, *37*, 785–789.

(6) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate *Ab Initio* Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.

(7) Weigend, F.; Ahlrichs, R. Balanced Basis Sets of Split Valence, Triple Zeta Valence and Quadruple Zeta Valence Quality for H to Rn: Design and Assessment of Accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297.

(8) VandeVondele, J.; Hutter, J. Gaussian Basis Sets for Accurate Calculations on Molecular Systems in Gas and Condensed Phases. *J. Chem. Phys.* **2007**, *127*, 114105.

(9) Goedecker, S.; Teter, M.; Hutter, J. Separable Dual-Space Gaussian Pseudopotentials. *Phys. Rev. B* **1996**, *54*, 1703–1710.

(10) Hartwigsen, C.; Goedecker, S.; Hutter, J. Relativistic Separable Dual-Space Gaussian Pseudopotentials from H to Rn. *Phys. Rev. B* **1998**, *58*, 3641–3662.

(11) Krack, M. Pseudopotentials for H to Kr Optimized for Gradient-Corrected Exchange-Correlation Functionals. *Theor. Chem. Acc.* **2005**, *114*, 145–152.

(12) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity Rescaling. *J. Chem. Phys.* **2007**, *126*, 14101.