Beyond MD17: The xxMD Dataset as a Chemically Meaningful Benchmark for Neural Force Fields Development

Zihan Pengmei
Department of Chemistry
The University of Chicago

Junyu Liu[*]
Pritzker School of Molecular Engineering
The University of Chicago

Yinan Shu[†]
Department of Chemistry
University of Minnesota

Abstract

Neural force fields (NFFs) have gained prominence in computational chemistry as surrogate models, superseding quantum-chemistry calculations in ab initio molecular dynamics. The prevalent benchmark for NFFs has been the MD17 dataset and its subsequent extension. These datasets predominantly comprise geometries from the equilibrium region of the ground electronic state potential energy surface, sampling from direct adiabatic dynamics. However, many chemical reactions entail significant molecular deformations, notably bond breaking. We demonstrate the constrained distribution of internal coordinates and energies in the MD17 datasets, underscoring their inadequacy for representing systems undergoing chemical reactions. Addressing this sampling limitation, we introduce the xxMD (Extended Excited-state Molecular Dynamics) dataset, derived from non-adiabatic dynamics. This dataset encompasses energies and forces ascertained from both multireference wave function theory and density functional theory. Furthermore, its nuclear configuration spaces authentically depict chemical reactions, making xxMD a more chemically relevant dataset. Our re-assessment of equivariant models on the xxMD datasets reveals notably higher mean absolute errors than those reported for MD17 and its variants. This observation underscores the challenges faced in crafting a generalizable NFF model with extrapolation capability.

1 Introduction

The data-driven evolution of molecular force fields is predominantly benchmarked against the **MD17** dataset introduced by Chmiela et al. [5] and its derivative, the rMD17 [6]. These datasets encompass dynamic data of ten small to medium-sized gas-phase molecules. While it's a common practice in machine learning community to consider both training and testing data as in-domain, this assumption is precarious for molecular dynamics. Such data, inherently being time-series sequences, mandates careful sampling to prevent unintended leakage of future information. A closer scrutiny of MD17 and its variant reveals their sampling bias, majorly confined to an exceedingly narrow **potential energy surface (PES)** in proximity to the equilibrium structure. This limited exploration translates to a circumscribed conformation and energy space. To highlight these shortcomings, we employ internal

---

[*]Kadanoff Center for Theoretical Physics, The University of Chicago, Chicago, IL 60637, USA. qBraid Co., Chicago, IL 60615, USA. SeQure, Chicago, IL 60615, USA

[†]To whom the correspondence should be addressed: shuxx055@umn.edu

coordinate analysis, revealing that existing molecular dynamics datasets are suboptimal concerning their segmentation strategy and the scope of molecular conformation space they cover.

For our analysis, we label those preliminary molecular dynamic datasets as **in-distribution (ID)** datasets. Nevertheless, significant chemical processes typically fall under the **out-of-distribution (OOD)** category. To underscore this discrepancy, consider the simplest chemical reaction depicted in Figure 1. The nuclear configuration space in this instance encompasses three pivotal regions: the reactant, the product, and the transition state (or the dividing plane) bridging the reactant and product. Merely sampling from the reactant region is inadequate and does not authentically represent chemical reactions. Consequently, calibrating NFF models on these skewed datasets induces a bias towards the reactant region. This can not only skew the NFF models but might also lead them to offer qualitatively inaccurate insights.

In this study, we introduce the **extended excited-state molecular dynamics (xxMD)** dataset, a successor to the MD17 dataset. While it retains the foundational aspect of encapsulating trajectory data from small to medium-sized gas-phase molecules, the xxMD stands apart due to its incorporation of **non-adiabatic** trajectories, emphasizing the dynamics associated with excited electronic states.[8, 24, 23, 4] Comprising four photo-sensitive molecules, these entities start within the excited electronic states, endowing them with a notably higher initial energy in contrast to their MD17 or rMD17 counterparts. This distinction leads the xxMD dataset to span a broader nuclear configuration space, ensuring representation across the complete chemical reaction PES: reactants, transition states, and products. A salient feature of the xxMD is its capture of regions near the conical intersections—crossings of potential energies spanning different electronic states. By encompassing these critical regions, including reactants, transition states, products, and conical intersections, the xxMD dataset promises to set novel benchmarks and challenges for NFF models.

For our xxMD dataset, we diverge from the MD17's framework by employing the trajectory surface hopping dynamics algorithm in tandem with the **state-averaged complete active state self-consistent field (SA-CASSCF)** electronic theory. This approach stands in contrast to the adiabatic dynamics used in MD17. The reason for the pivot to SA-CASSCF lies in its adeptness at handling electronic correlation effects at strongly deformed geometries, where KS-DFT often falls short, especially in scenarios involving multiple electronic states and conical intersections. Nevertheless, to ensure compatibility with prevalent datasets like MD17, we also computed single-point **spin-polarized KS-DFT (or unrestricted KS-DFT)** values. These calculations leverage the **M06**[27] exchange-correlation functional—a notably superior **meta-GGA** functional relative to PBE. This dual approach culminates in two datasets: **xxMD-CASSCF** and **xxMD-DFT**. The former captures potential energies and forces across the first three electronic states for azobenzene, dithiophene, malonaldehyde, and stilbene. The latter provides recomputed ground-state energy and force values, anchored to the same trajectories. Both xxMD datasets are structured via a **temporal split** method, partitioning training and testing data based on trajectory timesteps.

We evaluated six prominent message-passing NFF models on the xxMD datasets: SchNet[19], DimeNet++ (DPP)[7], SphereNet (SPN)[12], NequIP[3], Allegro[16], and MACE[2]. Each model was mostly used with its default parameters, and in line with convention, we trained the NFFs emphasizing more on force losses. While hyperparameter optimization could potentially fine-tune performance (See Supplementary Information for an example), it remains outside the scope of this study. Therefore, the presented results might not showcase the absolute best performance for each model. Given our observations, we urge researchers aiming to apply NFFs in practical scenarios to conduct rigorous re-benchmarks tailored to their specific chemical systems and objectives.

2 Preliminaries

In the realm of quantum mechanics, the behavior of nuclei is ideally described by the time-dependent Schrödinger equation. Yet, practical computation limits restrict nuclear quantum dynamics simulations to small systems with just 5 or 6 atoms. Consequently, in many cases, the nuclei are treated as classical particles. This premise paves the way for classical Molecular Dynamics (MD) and adiabatic Ab Initio Molecular Dynamics (AIMD), wherein the dynamics are propagated based on a **single electronic state**.

2.1 Dynamics

At the heart of classical MD is the Newtonian equation of motion:

$$m_i \frac{d^2 \mathbf{r}_i}{dt^2} = \mathbf{F}_i \tag{1}$$

where m_i denotes the mass of atom i, $\mathbf{r}_i$ its position, and $\mathbf{F}_i$ the force exerted on it. This force can be described as the negative gradient of the potential energy V at the atom's location:

$$\mathbf{F}_i = -\nabla V(\mathbf{r}_i) \tag{2}$$

The ground state electronic potential energy, $V(\mathbf{r}_i)$, in the absence of an external field, forms the basis for the PES. Classical force fields offer an analytical approximation of this energy based on nuclear configuration:

$$V(\mathbf{r}) = V_{\text{bond}}(\mathbf{r}) + V_{\text{angle}}(\mathbf{r}) + V_{\text{dihedral}}(\mathbf{r}) + V_{\text{non-bonded}}(\mathbf{r}) \tag{3}$$

This classical approximation often falls short under quantum mechanical scenarios, particularly during bond breaks, necessitating improvements in force field formulations. Upon electronic excitation, as observed in solar cells or photochemical reactions, nuclei confront electronic potentials beyond the ground state. Herein, dynamics involving multiple electronic states emerge. Nonadiabatic dynamics, particularly pertinent when energy levels soar, may either adopt the trajectory surface hopping method or the semiclassical Ehrenfest dynamics, depending on the specific conditions.

2.2 Chemical Reactions and PES

Chemical reactions within the classical MD or adiabatic AIMD framework involve a system transition between different PES minima. This movement is influenced by differences in electronic potential energy across nuclear configurations. However, systems generally adopt paths that demand minimal effort, known as reaction pathways. Accurate NFFs demand two crucial ingredients: a quantum chemical dataset spanning reactants, transition states, and products, and a proficient machine learning model capable of extrapolating across the PES. Notably, nonadiabatic dynamics can help curate datasets that offer richer nuclear configuration insights compared to adiabatic AIMD.
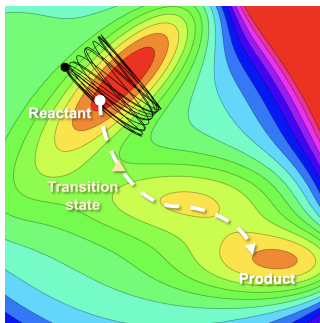

Figure 1: Illustration of trajectories on a sample potential energy surface[9]. The contour plot showcases the energy landscape, with varied hues depicting distinct energy tiers. Predominantly, trajectories navigate regions surrounding the minima, emphasizing systems' inclination towards states with diminished potential energy in or near equilibrium.

3 Dataset

3.1 MD17 and Revised MD17 Datasets (rMD17)

The use of the PBE density functional in conjunction with single-valence basis sets often falls short in simulating chemically meaningful reactions. Chmiela et al.[5] conducted adiabatic AIMD simulations on small gas-phase molecules at room temperature, purporting the electronic potential energies to be computed at the KS-DFT level. However, detailed specifics about the density functional, basis set, spin-polarization, and the software employed were notably absent from both the main

manuscript and supporting information. Later, Christensen et al.[6] revisited the potential energies and forces of the MD17 dataset, recalculating them at the PBE[17]/def2SVP[26] level of theory with an enhanced grid precision, leading to the inception of the rMD17 dataset. These datasets have been embraced widely as benchmarks in NFF studies[3, 2]. It's pivotal to underscore the inadequacy of employing the PBE density functional and def2SVP basis set for accurate chemical reaction simulations, as they can sometimes yield both quantitatively and qualitatively erroneous outcomes. Despite this, they might offer a continuous PES dependent on nuclear configuration, suggesting that the MD17 and rMD17 could be considered well-behaved datasets.

Adiabatic molecular dynamics datasets at low energy can not benefit from uniform sampling and cross-validation. It's important to highlight that the initial low-energy conditions in these adiabatic AIMD simulations constrain the nuclear configuration space considerably. As illustrated in Figure 1, adiabatic AIMD simulations at ambient conditions tend to produce trajectories that linger primarily within the reactant zone of the PES. To scrutinize the nuclear configuration space covered in the MD17 and rMD17 datasets, we analyzed the internal coordinate distributions of two specific molecules: the C-N=N-C dihedral angle and N=N bond length for azobenzene, as well as the C-C-C=O dihedral angle and C=O bond length for malonaldehyde. These distributions are visualized in Figure 2. Additionally, Figure 2 portrays the relative electronic potential energies, benchmarked against the ground state minimum geometry, and the force norm (essentially the electronic potential energy derivative) corresponding to these geometries. It's evident from the findings that the internal coordinate distribution is notably constricted. Consequently, this suggests that the training and testing samples in the MD17 and rMD17 datasets overlap significantly. Such overlaps flag concerns related to potential data leakage, which could inadvertently lead to unduly optimistic benchmark outcomes as cited in previous literature [1, 2, 3, 20, 12].
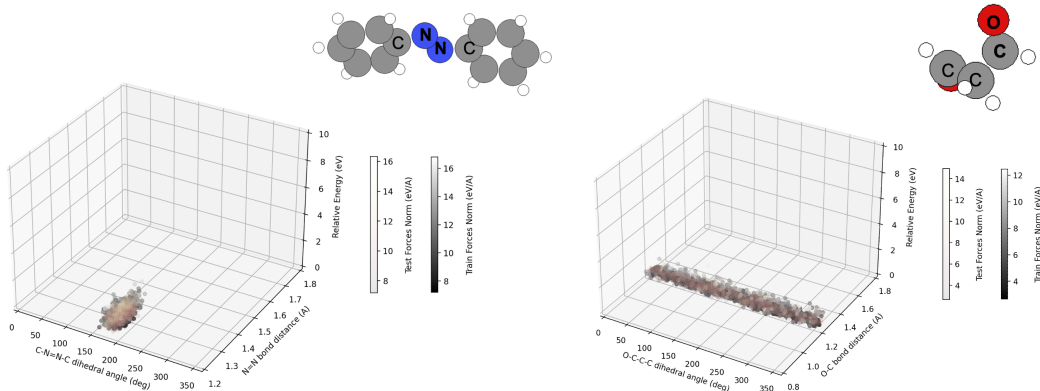

Figure 2: Illustration of training and testing sets using the reference split indices for azobenzene and malonaldehyde datasets in rMD17. The X-axis depicts dihedral angles (marked by 'C', 'N', and 'O'), the Y-axis denotes bond distances (highlighted by bold letters), and the Z-axis shows relative energy. Training and testing samples are differentiated by color, correlating to force norms. Note that training samples overlap with testing ones.

3.2 xxMD dataset

An updated alternative to the MD17 dataset is imperative for the effective development and benchmarking of NFFs. To meet this need, we introduce the **xxMD** dataset, curated specifically for benchmarking NFFs in the context of chemical reactions. This dataset encompasses four molecular entities: azobenzene, malonaldehyde, stilbene, and dithiophene. Key geometries along their respective reaction trajectories are depicted in Figure 6. It's noteworthy that both azobenzene and malonaldehyde are constituents of the MD17 and rMD17 datasets. This inclusion facilitates a comparative analysis of the internal coordinate distribution between the xxMD and the previously mentioned datasets. We refer readers to the supplementary information for the comparison.

The cornerstone of the xxMD datasets is the non-adiabatic dynamic sampling. We employed the trajectory surface hopping (TSH) nonadiabatic dynamics, interfaced with the SA-CASSCF [18] electronic structure method, for all included molecules. In stark contrast to KS-DFT, SA-CASSCF

ensures qualitative precision in representing the global PES—encompassing both bond-breaking geometries and conical intersections[11, 21]. Our data selection prioritized samples from energy-conserving trajectories exclusively, with the criteria for total energy conservation. The sample size is listed in Table 9. Current NFF methodologies have demonstrated the capability to attain chemical precision vis-à-vis the foundational electronic structure calculations, even with a dataset of fewer than 1,000 training samples [19, 12, 3, 2, 20]. Comprehensive computational specifics can be found in the Supplementary Information.

The ensemble-averaged radial distribution function (RDF) and mean square displacement (MSD) from the xxMD datasets demonstrate a more expansive and meaningful sampling of the PES than that of MD17. Both the RDF and MSD of nuclear configurations as functions of time are depicted in the left and right columns of Figure 3, respectively. While the RDF quantifies the average likelihood of finding a particle at a particular radial distance from a reference configuration, the MSD gauges the average squared movement of molecules over time. The nuclear configuration shifts evident in nonadiabatic dynamics—and by extension, in the xxMD dataset—are substantial. This dynamic sampling breadth sets the xxMD dataset distinctly apart from the MD17 datasets and its variants. With the xxMD dataset capturing diverse regions on the PESs, we anticipate that mastering the PESs for molecules within this dataset will pose a significantly greater challenge.

Spin-unpolarized DFT is ill-suited for simulating chemical reactions involving bond-breaking or bond-forming events. From the geometries sampled through TSH with SA-CASSCF dynamics, we recalculated the ground-state potential energy and atomic forces employing spin-polarized KS-DFT at the M06[27]/6-31g level. It's worth noting that M06 is a hybrid meta-GGA exchange-correlation functional, which delivers notably superior accuracy compared to PBE. Crucially, spin-unpolarized (or unrestricted) KS-DFT falls short when simulating chemical reactions, particularly during bond-breaking events due to the reactions' open-shell nature. For clarity, we've named the dataset derived from these recalculations as the **xxMD-DFT** dataset, whereas our principal dataset is referred to as the **xxMD-CASSCF** dataset.

Temporal splitting presents a formidable yet insightful test for NFFs in predicting unexplored sections of the PES. In curating the xxMD datasets, we favored temporal over random splitting. While the latter uniformly divides data by shuffling all sampled geometries and randomly segregating them into training, validation, and testing sets, temporal splitting allocates time-series data based on timesteps. A designated range of timesteps in the dynamic sequence is reserved as the hold-out test set, adhering to a 50:25:25 splitm, granting a unique vantage point to scrutinize the predictive prowess of the model over unseen portions of PESs. This distinction becomes evident in the right columns of Figure 3 where trajectories rapidly deviate over time. This behavior underscores that the temporal-split xxMD datasets adeptly challenge the extrapolation and predictive power of NFFs.

4 Experiment

We picked six representative equivariant NFFs to benchmark. The hyperparameters and training details of models in the supplementary information. We used a weighted loss of 1:1000 on energy and forces. We stress that our purpose is not to perform an extensive comparison of models over multiple choices of hyperparameters. Rather, we limit ourselves to showing the performance of the models in the default configurations.

4.1 Complete xxMD-CASSCF and xxMD-DFT

We first evaluate the regression precision of all models on the first three electronic states, which are labeled as S_0, S_1, and S_2 respectively (Label S denotes the singlet spin state which is a widely used notation in photochemistry) by using the temporal splitting approach. The MAE of the predictive energy and forces for test sets are shown in Table 1. Similarly, we present such results of using xxMD-DFT datasets in Table 2. Additional results on the validation sets are available in the supporting information. Note that validation sets depict the nuclear configurations that are closer to the training sets. Therefore, the MAE shown in validation sets are in general lower than that for test sets.
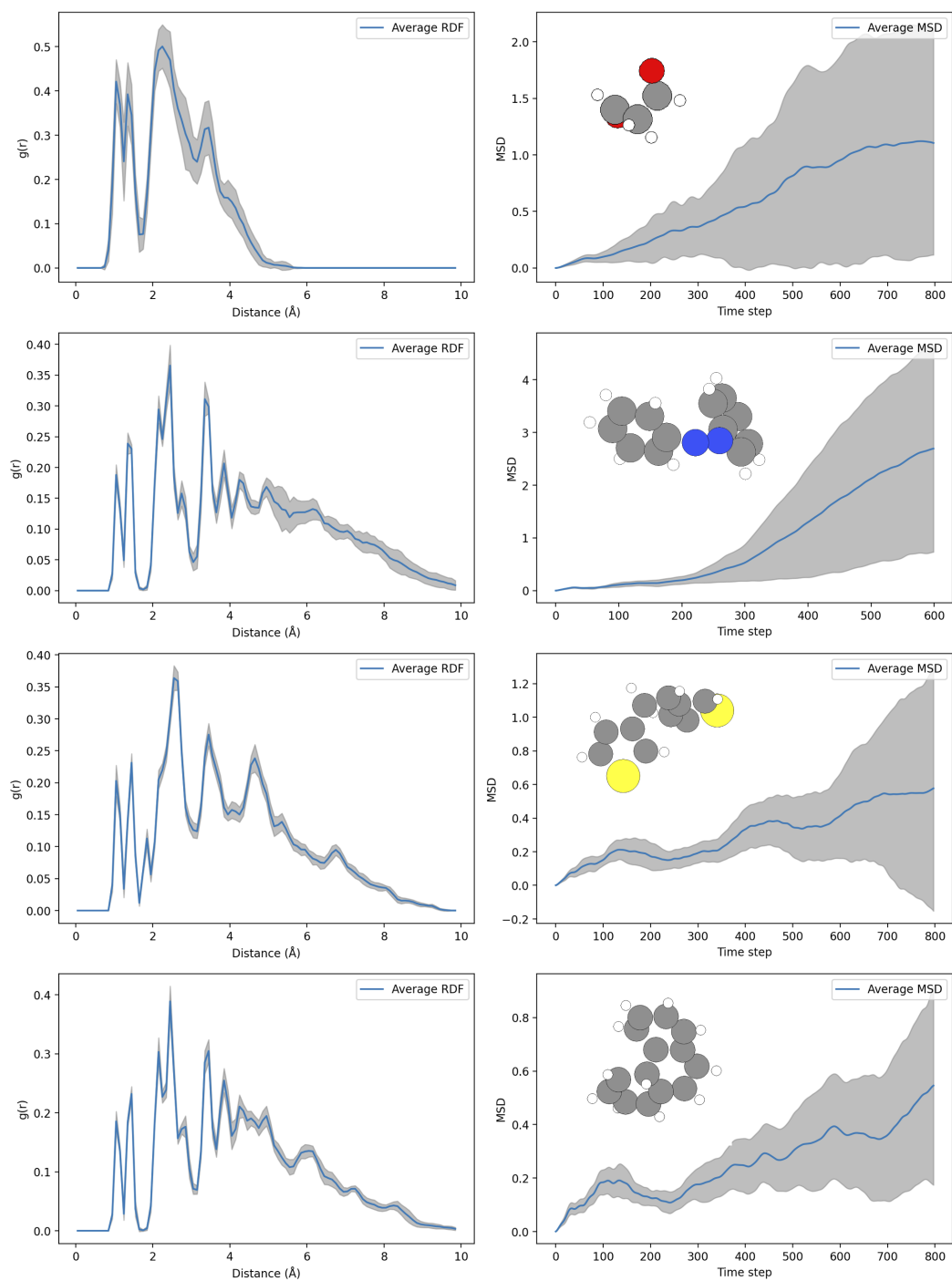
Figure 3: Comparison of Average RDFs and MSDs Across Multiple Trajectories. Each row corresponds to a group of trajectories, with RDF on the left (indicating particle density as a function of distance) and MSD on the right (showing particle displacement over time). Shaded regions represent standard deviations.

Table 1: Comparison of predictive MAE of energy(E, meV) and forces(F, meV/A) on hold-out testing set for different models on temporally split xxMD-CASSCF datasets and tasks.

Dataset	State	Task	MACE	Allegro	NequIP	SchNet	DPP	SPN
Azobenzene	S_0	E	527	**437**	870	648	528	493
		F	**63**	82	76	156	102	96
	S_1	E	599	524	1160	619	497	**494**
		F	**78**	98	85	157	91	88
	S_2	E	881	783	1957	894	837	**831**
		F	**191**	216	215	284	224	231
Dithiophene	S_0	E	293	296	295	306	295	**290**
		F	**14**	31	21	94	30	31
	S_1	E	205	211	224	217	**204**	205
		F	**37**	81	49	103	41	44
	S_2	E	246	255	259	262	**244**	246
		F	52	10	70	121	**51**	54
Malonaldehyde	S_0	E	530	443	770	515	452	**442**
		F	**105**	142	166	220	138	137
	S_1	E	528	**458**	1227	482	482	462
		F	164	189	189	260	165	**161**
	S_2	E	679	**528**	159	653	610	615
		F	276	307	309	353	251	**238**
Stilbene	S_0	E	538	544	529	604	**519**	544
		F	**72**	87	112	191	91	114
	S_1	E	391	353	370	424	**313**	352
		F	**58**	66	85	142	88	93
	S_2	E	604	669	674	678	550	**529**
		F	**117**	142	178	259	148	159

Table 2: Comparison of predictive MAE of energy(E, meV) and forces(F, meV/A) on hold-out testing set for different models xxMD-DFT datasets and tasks with temporal split.

Dataset	Task	MACE	Allegro	NequIP	SchNet	DPP	SPN
Azobenzene	E	292	**174**	1754	722	300	260
	F	**85**	110	129	283	173	168
Stilbene	E	**315**	332	647	397	439	477
	F	**149**	189	156	291	162	168
Malonaldehyde	E	190	**151**	244	360	179	185
	F	**166**	210	227	394	257	255
Dithiophene	E	100	103	243	323	**61**	76
	F	**51**	75	101	177	74	90

4.2 Model behavior comparison between xxMD and (r)MD17 datasets

In this section, we analyze model behavior for azobenzene and malonaldehyde from xxMD and (r)MD17 datasets. Benchmarks for (r)MD17 reveal that the accuracy of MACE, NequIP, and SPN exceeds that of traditional electronic structure methods[2, 3, 12, 14]. It's essential to note that typical errors for KS-DFT in predicting relative transition state energy can be several kcal/mol. For instance, the MAE of HTBH38 and NHTBH38 is about 9.1 kcal/mol for PBE and 2.4 kcal/mol for M06. Thus, an NFF fitting error below 50 meV would surpass the accuracy of modern density functionals. However, such claims are pertinent mainly to ground state potential energies, given that excited state calculations are often less precise. Therefore, given the reported MAEs, these NFF models perform admirably on (r)MD17 datasets.

However, this conclusion might be deceiving. Previous discussions highlight the constrained nuclear configuration space in MD17 and rMD17. A comparative analysis of MAEs for the six NFF models on azobenzene and malonaldehyde from xxMD-DFT and (r)MD17 is presented in Table 3. Literature-derived MD17/rMD17 results indicate that all models used 1,000 training samples[2, 3, 12]. Predictably, the predictive prowess of NFF models diminishes when applied to the xxMD dataset.

Table 3: Comparison of predictive MAE on hold-out testing sets of NFF models on azobenzene and malonaldehyde in (r)MD17 and xxMD-DFT datasets. (r)MD17 benchmarks with 1,000 samples are taken from [2, 12, 19].

Molecule	Dataset	Task	MACE	Allegro	NequIP	SchNet	DPP	SPN
Azobenzene	rMD17	E	1.2	1.2	0.7	N/A	N/A	N/A
		F	3.0	2.6	2.9	N/A	N/A	N/A
	xxMD	E	292	174	1754	722	300	260
		F	85	110	129	283	173	168
Malonaldehyde	(r)MD17	E	0.8	0.6	0.8	5.6	4.5	N/A
		F	4.1	3.6	5.1	28.6	16.6	7.5
	xxMD	E	190	151	244	360	179	185
		F	166	210	227	394	257	255

5   Conclusion

The effectiveness of NFF models largely depends on the datasets they are benchmarked against. Historically, the (r)MD17 datasets have been the gold standard for this purpose. However, our study highlights the potential shortcomings of relying solely on (r)MD17 datasets. Given that they primarily capture a narrow nuclear configuration space from ground state ab initio molecular dynamics, they fall short of encompassing the holistic nuclear configuration pertinent to chemical reactions. Training NFF models on such datasets can be somewhat trivial and could result in misleading conclusions about their true capabilities.

Addressing this gap, we introduced the xxMD dataset, derived from nonadiabatic dynamics trajectories. The xxMD dataset offers a comprehensive representation of the nuclear configuration space, encapsulating the reactant, transition state, product, and conical intersection regions of potential energy surfaces. Its inclusion of several low-lying excited state potential energy surfaces underscores its importance and the challenges it presents for NFF model development. Our benchmarks of prevailing NFF models on the xxMD dataset have revealed pronounced difficulties. Utilizing default hyperparameters, the chosen NFF models struggled to offer quantitatively or even qualitatively accurate force field models for specific systems. We anticipate that our findings will galvanize the community towards pioneering more advanced NFF models better equipped to study intricate chemical reactions.

References

[1] Ilyes Batatia, Simon Batzner, Dávid Péter Kovács, Albert Musaelian, Gregor NC Simm, Ralf Drautz, Christoph Ortner, Boris Kozinsky, and Gábor Csányi. The design space of e (3)-equivariant atom-centered interatomic potentials. *arXiv preprint arXiv:2205.06643*, 2022.

[2] Ilyes Batatia, David P Kovacs, Gregor Simm, Christoph Ortner, and Gábor Csányi. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. *Advances in Neural Information Processing Systems*, 35:11423–11436, 2022.

[3] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1): 2453, 2022.

[4] N. C. Blais and D. G. Truhlar. Trajectory-surface-hopping study of na(3p 2p) + h2 -> na(3s 2s) + h2(v', j', theta). *The Journal of chemical physics*, 79:1334–1342, 1983.

[5] Stefan Chmiela, Alexandre Tkatchenko, Huziel E Sauceda, Igor Poltavsky, Kristof T Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science advances*, 3(5):e1603015, 2017.

[6] Anders S Christensen and O Anatole Von Lilienfeld. On the role of gradients for machine learning of molecular energies and forces. *Machine Learning: Science and Technology*, 1(4): 045018, 2020.

[7] Johannes Gasteiger, Shankari Giri, Johannes T Margraf, and Stephan Günnemann. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. *arXiv preprint arXiv:2011.14115*, 2020.

[8] Michael F Herman. Nonadiabatic semiclassical scattering. i. analysis of generalized surface hopping procedures. *The Journal of chemical physics*, 81(2):754–763, 1984.

[9] Wolfram Research, Inc. Mathematica, Version 13.3. URL `https://www.wolfram.com/mathematica`. Champaign, IL, 2023.

[10] Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N Groves, Bjørk Hammer, Cory Hargus, et al. The atomic simulation environment—a python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002, 2017.

[11] B. G. Levine, C. Ko, J. Quenneville, and T. J. Martinez. Conical intersections and double excitations in time-dependent density functional theory. *Molecular Physics*, 104:1039–1051, 2006.

[12] Yi Liu, Limei Wang, Meng Liu, Xuan Zhang, Bora Oztekin, and Shuiwang Ji. Spherical message passing for 3d graph networks. *arXiv preprint arXiv:2102.05013*, 2021.

[13] Sebastian Mai, Philipp Marquetand, and Leticia González. Nonadiabatic dynamics: The sharc approach. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 8(6):e1370, 2018.

[14] Narbe Mardirossian and Martin Head-Gordon. Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. *Molecular Physics*, 115(19):2315–2372, 2017.

[15] Donald Allan McQuarrie and John Douglas Simon. Molecular thermodynamics. *(No Title)*, 1999.

[16] Albert Musaelian, Simon Batzner, Anders Johansson, Lixin Sun, Cameron J Owen, Mordechai Kornbluth, and Boris Kozinsky. Learning local equivariant representations for large-scale atomistic dynamics. *Nature Communications*, 14(1):579, 2023.

[17] John P Perdew, Kieron Burke, and Yue Wang. Generalized gradient approximation for the exchange-correlation hole of a many-electron system. *Physical review B*, 54(23):16533, 1996.

[18] Björn O Roos, Peter R Taylor, and Per EM Sigbahn. A complete active space scf method (casscf) using a density matrix formulated super-ci approach. *Chemical Physics*, 48(2):157–173, 1980.

[19] Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems*, 30, 2017.

[20] Kristof Schütt, Oliver Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International Conference on Machine Learning*, pages 9377–9388. PMLR, 2021.

[21] Y. Shu, K. A. Parker, and D. G. Truhlar. Dual-functional tamm-dancoff approximation: A convenient density functional method that correctly describes s1/s0 conical intersections. *Journal of Physical Chemistry Letters*, 8:2107–2112, 2017.

[22] Sina Stocker, Johannes Gasteiger, Florian Becker, Stephan Günnemann, and Johannes T Margraf. How robust are modern graph neural network potentials in long and hot molecular dynamics simulations? *Machine Learning: Science and Technology*, 3(4):045010, 2022.

[23] J. C. Tully and R. K. Preston. Trajectory surface hopping approach to nonadiabatic molecular collisions: The reaction of h+ with d2. *The Journal of chemical physics*, 55:562–572, 1971.

[24] John C Tully. Molecular dynamics with electronic transitions. *The Journal of Chemical Physics*, 93(2):1061–1071, 1990.

[25] Justin M Turney, Andrew C Simmonett, Robert M Parrish, Edward G Hohenstein, Francesco A Evangelista, Justin T Fermann, Benjamin J Mintz, Lori A Burns, Jeremiah J Wilke, Micah L Abrams, et al. Psi4: an open-source ab initio electronic structure program. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2(4):556–565, 2012.

[26] Florian Weigend. Accurate coulomb-fitting basis sets for h to rn. *Physical chemistry chemical physics*, 8(9):1057–1065, 2006.

[27] Yan Zhao and Donald G Truhlar. The m06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four m06-class functionals and 12 other functionals. *Theoretical chemistry accounts*, 120:215–241, 2008.

A Brief introduction of chosen neural force fields

In this study, we picked six representative neural network architectures for NFF applications, namely, SchNet[19], DPP[7], SPN[12], NequIP[3], Allegro[16] and MACE[2]. In general, those approaches can be devided into two catagories based on the representation of the feature space. SchNet, DPP and SPN are the so-called scaler-based NFFs, while NequIP, MACE and Allegro are vector-based NFFs, as we summarized in Table 4.

The key concept in SchNet is the continuous-filter convolution, which involves two steps: interaction and update. In the interaction step, the model calculates pairwise interaction features between all atoms based on their distances, using a set of radial bessel basis. The update step then uses these interaction features to update the atom-centered descriptors. In DPP, a higher-order feature, bond angle has been introduced to enhance the expressiveness of the neural network. DPP uses a concept called spherical functions to account for the directionality of the interactions between atoms. The DPP architecture uses 'interaction blocks' to propagate information through the molecular graph. Each interaction block consists of a radial and a spherical part. The radial part captures the distance-based interactions, similar to SchNet. The spherical part captures the angular interactions among any three atoms in the molecule, which is unique to DPP. As a continuation of the DPP, SPN further introduces another higher-order feature called dihedral angles among any four atoms in the molecule. These improvements are chemically-intuitive since bond lengths, angles, and dihedral angles are very common descriptors in classical force fields [15].

Table 4: Summary of models, their features, and the corresponding years of introduction.

Model	Feature	Year
SchNet	Bond length	2017
DPP	Bond length, Bond angle	2020
SPN	Bond length, Bond angle, Dihedral angle	2021
NequIP	SO(3) vector	2021
Allegro	SO(3) vector	2022
MACE	SO(3) vector	2022

On the other hand, NequIP, Allegro, and MACE are examples of group equivariant NFFs that based on SO(3) relative displacement vectors between any two atoms in the molecule. These networks use the representation theory of the three-dimensional orthogonal group to construct neurons that obey equivariance with respect rotations and reflections of a molecular system's pose. We visualize this concept in Figure 4. Atomic types are embedded as node features, and relative displacement vectors that contains the positional information are converted into activations that transform according to irreducible representations (irreps) of the orthogonal group. Nonlinearities for these activations are constructed using the tensor product, followed by applying the Clebsch-Gordon decomposition to convert the product back into irreducible components.

SchNet, DimeNet++ and SphereNet are available as implemented in the Dive Into Graphs package (`https://github.com/divelab/DIG.git`). NequIP package is available at `https://github.com/mir-group/nequip.git`. Allegro package is available at `https://github.com/mir-group/allegro`. MACE package is available at `https://github.com/ACEsuit/mace.git`. All packages are up-to-date at the data of the publication. All the trainings are done with single precision float format. SchNet, DPP and SPN models are initialized using the default hyperparameters shipped with the packages. Allegro hyperameters can be found at `https://github.com/mir-group/allegro/blob/main/configs/example.yaml`, NequIP hyperparameters are mainly `https://github.com/mir-group/nequip/blob/main/configs/example.yaml`, MACE hyperparameters are mainly `https://github.com/ACEsuit/mace`. Since Dive Into Graphs package doesn't implement the scale and shift of the energy, we manually rescaled the energy by substracting the energy of the configuration with the lowest potential energy.
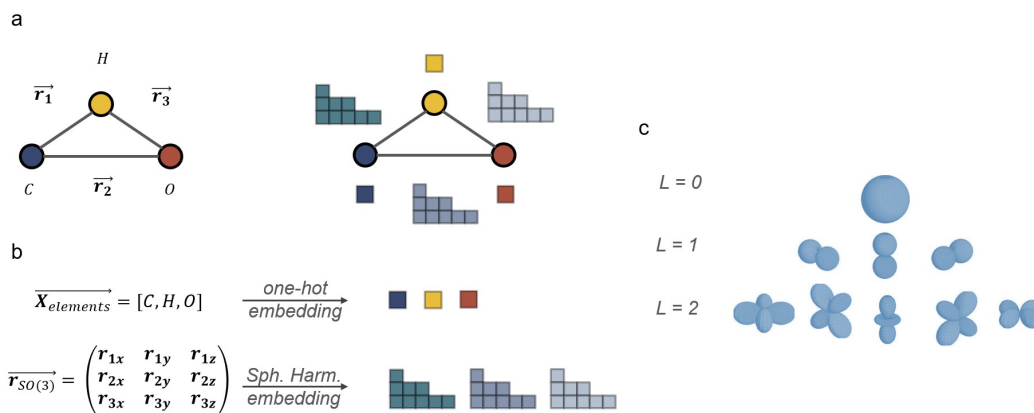
Figure 4: (a) depicts three atoms with their relative displacement vectors. (b) illustrates the details of atomic embedding with E(3) equivariant activations based on spherical harmonics and one-hot encoding for chemical elements. (c) gives an illustration of spherical harmonics with quantum numbers $L = 0, 1, 2$. The Clebsch-Gordan coefficients are used during the aggregation step to ensure rotational equivariance when combining activations with different irreps.
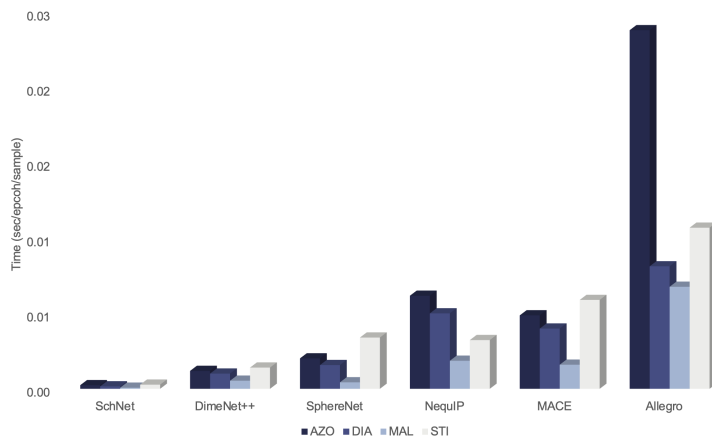

Figure 5: Comparison of average computational time for NFFs. The timing is specific to the chosen hyperparameters. All NFFs, except MACE, operate with single precision. Generally, group-equivariant NFFs are significantly more computationally expensive.

B   Timing

In this practical view, we present a comprehensive analysis of the operational time of multiple NFFs examined in our study as illustrated in Figure 5. It is important to note that the specific runtime of each NFF model is contingent upon the chosen setup and hyperparameter selection. For example, the radius cutoff utilized for generating locally fully-connected graphs can yield varying numbers of edges and nodes in each mini-batch. Within our findings, we have diligently reported the time required for processing each sample in a mini-batch using the designated hyperparameters. Consequently, we emphasize that while we employed mostly default hyperparameters as a practical reference.
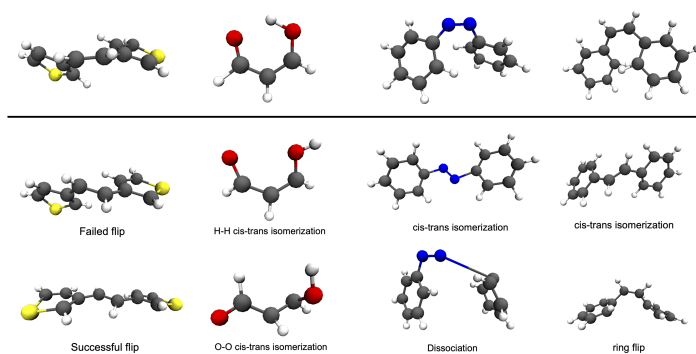
C   Additional illustration of xxMD datasets

Figure 6: Schematic representation of the photodynamic processes featured in the xxMD dataset.

Here we provide addition illustration (Figure 7) of the xxMD-CASSCF datasets with the ground-state energy and forces as the internal coordinate analysis of MD17. For azobenzene, the primary reaction path involves the cis-trans isomerization of the two phenyl groups along the N=N bond. For malonaldehyde, the reaction path involves either a H-H cis-trans isomerization occurs along the O=C bond or a O-O cis-trans isomerization occurs along the carbon skeleton. The reaction path of stilbene involves the cis-trans isomerization of the two phenyl rings along the C=C double bond and the flip of the phenyl rings to opposite directions. The reaction path of dithiophene is also the cis-trans isomerization of two five-member rings along the C=C double bond.

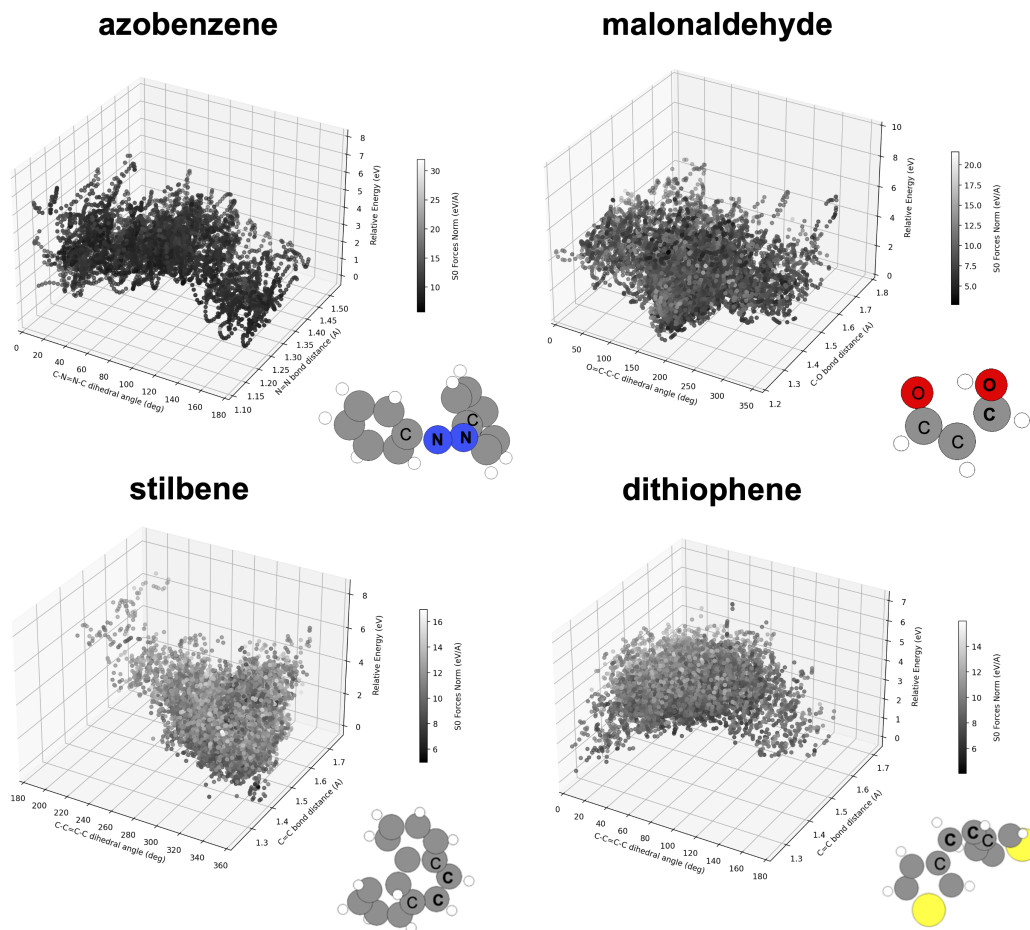Figure 7: Illustration of xxMD datasets using similar internal coordinates as MD17 analysis. Only the ground-state (in black/white color scheme) energies are visualized for clarity. This figure clearly indicates the breadth of the conformation space explored by using direct non-adiabatic dynamics as compared to MD simulation with room temperature.

D Benchmark results on the validation sets of xxMD-CASSCF and xxMD-DFT

Table 5: Comparison of predictive MAE on validation set for different models on temporally split xxMD-CASSCF datasets and tasks. Energy(E) has the unit of meV, while forces(F) have the unit of meV/A.

Dataset	State	Task	MACE	Allegro	NequIP	SchNet	DPP	SPN
Azobenzene	S_0	E	527	**367**	530	682	526	494
		F	**50**	70	69	141	83	79
	S_1	E	474	**308**	869	483	478	428
		F	**67**	83	74	134	79	74
	S_2	E	864	**742**	1590	897	804	801
		F	**163**	185	180	257	191	185
Dithiophene	S_0	E	300	295	304	302	**286**	287
		F	**10**	21	17	76	22	24
	S_1	E	259	208	226	219	**206**	208
		F	65	78	46	101	**33**	36
	S_2	E	246	258	256	259	**244**	249
		F	50	104	69	119	**49**	51
Malonaldehyde	S_0	E	488	**386**	583	470	419	415
		F	**84**	147	109	179	108	109
	S_1	E	507	**406**	828	469	446	451
		F	**144**	184	168	233	147	145
	S_2	E	556	**457**	858	526	512	512
		F	221	255	281	301	197	**188**
Stilbene	S_0	E	517	514	**359**	505	467	461
		F	54	71	**12**	145	71	75
	S_1	E	322	293	**262**	351	294	316
		F	38	45	**20**	97	62	61
	S_2	E	494	505	**377**	596	486	473
		F	80	98	**31**	176	104	106

Table 6: Comparison of predictive MAE on validation set for different models xxMD-DFT datasets and tasks with temporal split. Energy(E) has the unit of meV, while forces(F) have the unit of meV/A.

Dataset	Task	MACE	Allegro	NequIP	SchNet	DPP	SPN
Azobenzene	E	257	106	393	539	184	**168**
	F	**71**	98	119	248	150	140
Stilbene	E	190	200	161	**156**	224	248
	F	**104**	116	117	196	114	125
Malonaldehyde	E	156	**91**	134	257	116	127
	F	**135**	162	173	326	208	204
Dithiophene	E	89	54	86	198	**49**	69
	F	**47**	59	81	158	61	78

E   Addition experiment of hyperparameter tuning

We would like to stress again, our purpose is to give a initial view of the datasets using common hyperparameters without tuning, and we don't aim to strictly test models listed. We left most hyperparameters unchanged as default, and uses a loss weight heavily focused on the forces following the literatures [22, 2, 3]. However, users should carefully use the hyperparameters before applying to specific chemical problems.

We used a default MACE model and one subset of xxMD-CASSCF dataset and varied the weights on the energy and forces, and we found that by simply tuning this hyperparameter, MACE would perform noticeably differently. For instance, the regression accuracy on force is not improved and

Table 7: Predictive MAE of energy (meV) and forces (meV/A) on the ground-state azobenzene in xxMD-CASSCF dataset using various loss weights and default MACE model.

Loss E:F ratio	Testing		Validation	
	E	F	E	F
1000:1	325	210	291	186
100:1	311	104	266	87
10:1	338	72	327	58
1:1	446	66	458	53
1:10	516	64	524	50
1:100	541	65	544	50
1:1000	527	63	527	50

accuracy on energy deteriorates quickly when the weight on the force gradually increase from 1 to 1000. On the contrary, putting slightly more weights on the energy greatly improve the overall performance, as we laid out in Table 7. Thus, we would like to leave a note to future users that exploring the hyperparameter spaces is important.

F Computational details

The active space and basis set used for SA-CASSCF for all four molecules are shwon in Table 9. The total number of trajectories simulated are vary, but finally selected number of points for each molecule in xxMD dataset is summarized in Table 9 as well. These points are selected from energy conserving trajectories only, and we used the criteria for the total energy conservation as listed in Table 9. Therefore, all trajectories fail to conserve the total energy below the threshold are discarded. We show the total energy conservation in Figure 8

Table 8: Summary of the computational methods, number of samples used in direct non-adiabatic dynamics sampling for four molecules, and number of data points for all studied molecules. The number in the parenthesis indicates the number of active electrons and orbitals. The total energy conservation (Total E. Con.) criteria has a unit of eV.

Molecule	Method	Total E. Con.	Num. of Samples
Dithiophene	SA-CASSCF(10e,10o)/6-31g	0.2	24769
Azobenzene	SA-CASSCF(6e,6o)/6-31g	0.6	8414
Malonaldehyde	SA-CASSCF(8e,6o)/6-31g	0.3	25568
Stilbene	SA-CASSCF(2e,2o)/6-31g*	0.2	27965

Table 9: Summary of the number of samples used in direct non-adiabatic dynamics sampling for four molecules, and number of data points for all studied molecules.

Molecule	Num. of Samples	Train	Valid	Test
Dithiophene	24769	12400	6169	6200
Azobenzene	8414	4200	2114	2100
Malonaldehyde	25568	14000	6965	7000
Stilbene	27965	12800	6368	6400

F.1 Dynamics

We used SHARC-MD[13] 2.1.2 package to run non-adiabatic dynamic simulations, the software is available at `https://github.com/sharc-md/sharc`. Initial conformations are generated by Wigner-Sampling of the optimized ground-state structure with the same level of electronic structure method. For each conformation, a single-point calculation is performed to acquire the energy of states without spin-orbit calculations. To select initial excited-states, the MCH representation of the Hamiltonian is used to simulate delta-pulse excitation based on excitation energies and oscillators strengths with an excitation window of 0.0 to 10.0 eV.

For azobenzene, we conducted 300 fs SHARC dynamics simulations with a time step of 0.5 fs. For dithiophene, we conducted 500 fs SHARC dynamics simulations with a time step of 0.5 fs. For malonaldehyde, we conducted 300 fs SHARC dynamics with a timestep of 0.25 fs. For stilbene, we performed 500 fs SHARC dynamics with a time step of 0.5 fs. Local diabatizatrion scheme was used to calculate the non-adiabatic coupling vectors by calculating the overlap matrix of wavefunctions between steps. Non-adiabatic coupling vectors are included in the gradient transformation. kinetic energy are adjusted by rescaling the velocity vectors during a surface hop. When the surface hop is refused due to insufficient energy, the velocity doesn't reflect at a frustrated hop. Default energy-based decoherence scheme was used for decoherence correction. The standard SHARC surface hopping probabilities was used as the surface hopping scheme. All gradients and non-adiabatic couplings of active states were calculated at each time step. For azobenzene, dithiophene, malonaldehyde, and stilbene the threshold of total energy was set to 0.6 eV, 0.2 eV, 0.3 eV and 0.2 eV.
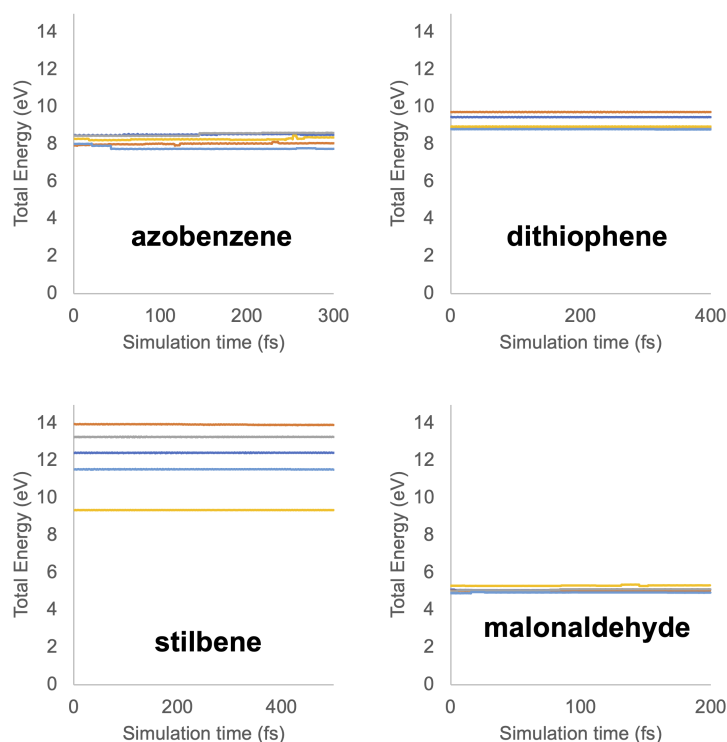
Figure 8: Illustration of total energy conservation over the simulation time of trajectories in xxMD-CASSCF datasets. All trajectories follow the total energy conservation threshold.

F.2 Complete Active Space Self-Consistent Field (CASSCF)

In quantum chemistry, accurately capturing electron correlation—the interaction of electrons relative to one another—is pivotal for an in-depth understanding of a molecule's electronic structure. While standard methods like Hartree-Fock (HF) have their strengths, they can falter in specific scenarios. This is where the CASSCF method becomes instrumental.

Central to CASSCF is the categorization of molecular orbitals into three distinct groups:

1. **Inactive (core) orbitals:** These are fully occupied orbitals, exempted from the correlation treatment.

2. **Active orbitals:** A defined number of electrons within these orbitals undergo correlation across a predetermined set of orbitals. The flexibility in electron configuration within the active space encapsulates static electron correlation.

3. **Virtual (secondary) orbitals:** Remaining unoccupied, these orbitals are sidelined from the primary correlation procedure.

The CASSCF methodology initially optimizes the active space orbitals employing a comprehensive configuration interaction (CI) calculation. This act of considering all plausible electron configurations within the active ambit captures static correlation. To address dynamic correlation, supplementary methods, like multi-reference perturbation theory (MRPT), are often invoked.

Advantages of CASSCF:

- Offers a harmonized treatment of electron correlation.

- Particularly apt for systems with closely-spaced electronic states, encompassing transition states, metal complexes, and excited states.

18

However, one should note the substantial computational demands, especially with enlarging active spaces, which can potentially restrict its application or mandate approximate solutions.

For SA-CASSCF calculations, OpenMolcas 22.06 was used, which is available at `https://gitlab.com/Molcas/OpenMolcas`. The active space orbitals of the starting configurations are listed as following (Figure 9).
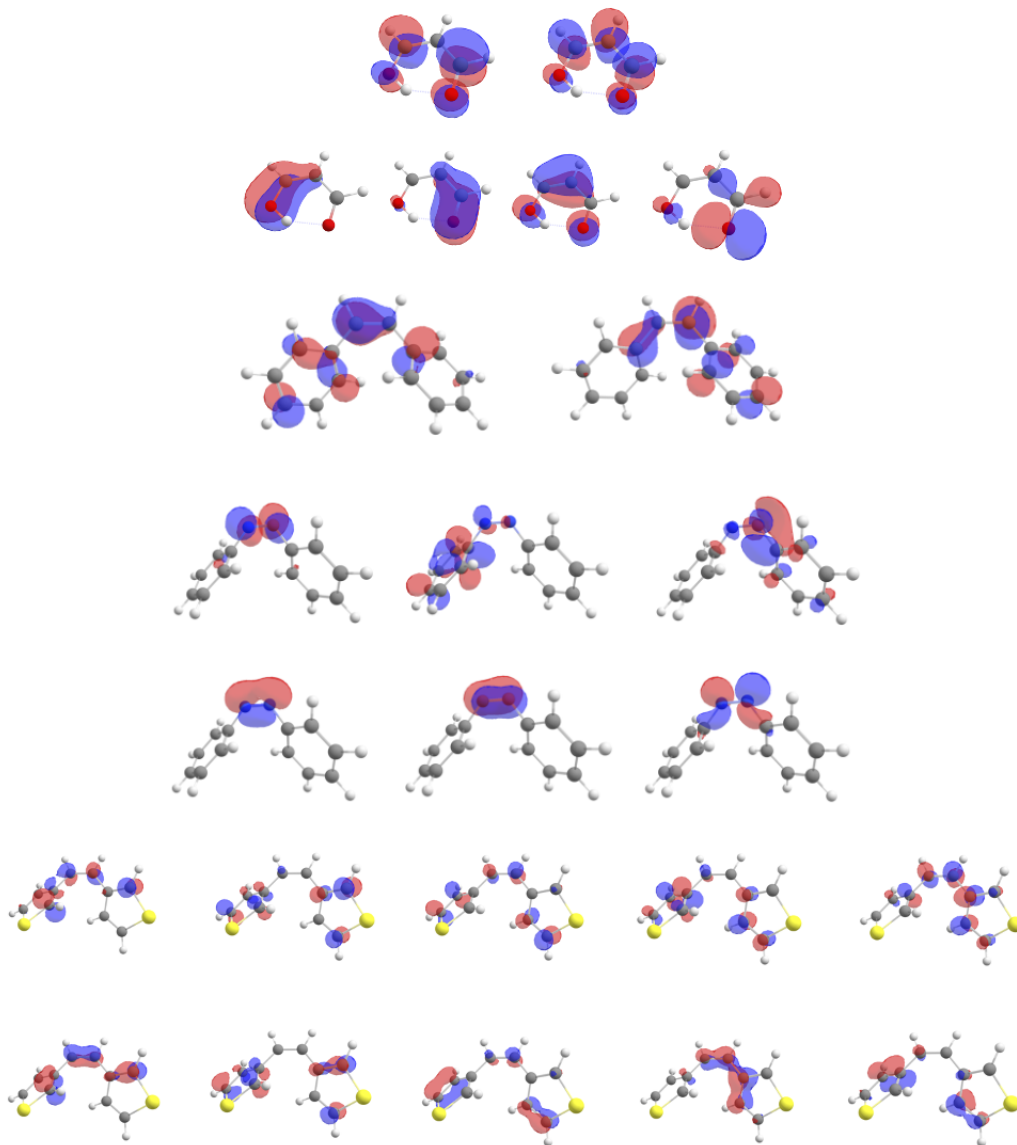

Figure 9: Active space orbitals used SA-CASSCF calculations for malonaldehyde, stilbene, azobenzene and dithiophine.

F.3 Unrestricted KS-DFT

In molecular modeling, the precise representation of electronic configurations during chemical reactions is paramount. The popular restricted KS-DFT inherently pairs electrons, enforcing identical spatial orbitals for both spin-up and spin-down states.

Consider the paradigmatic dissociation of hydrogen (H_2) into atomic hydrogen:

$$H_2 \rightarrow 2H \tag{4}$$

Within the confines of restricted KS-DFT, as H_2 dissociates, the emerging electrons—now localized on individual atoms—are still bound to identical spatial distributions. This treatment may distort the real physical scenario.

Unrestricted KS-DFT, on the other hand, permits differentiation between spin-up and spin-down spatial orbitals, enabling a nuanced portrayal of the process. In the H_2 example, would independently model the electron on each hydrogen atom, providing a truer representation of the physical system.

For all unrestricted KS-DFT calculations, we used M06[27] meta-GGA hybrid functional with 6-31g basis set. All calculations are done with the Psi4[25] package (available at `https://github.com/psi4/psi4`) interfaced the ASE[10] package (available at `https://github.com/rosswhitfield/ase`).

F.4 Dihydrogen dissociation: a comparative case of RKS, UKS and CASSCF

The limitation of using DFT, espeicialy restricted DFT becomes evident when examining the H-H bond-breaking process, as illustrated in Figure 10. Here, spin-unpolarized DFT yields an inaccurate yet smooth curve when juxtaposed against its spin-polarized counterpart, with CASSCF serving as the reference.
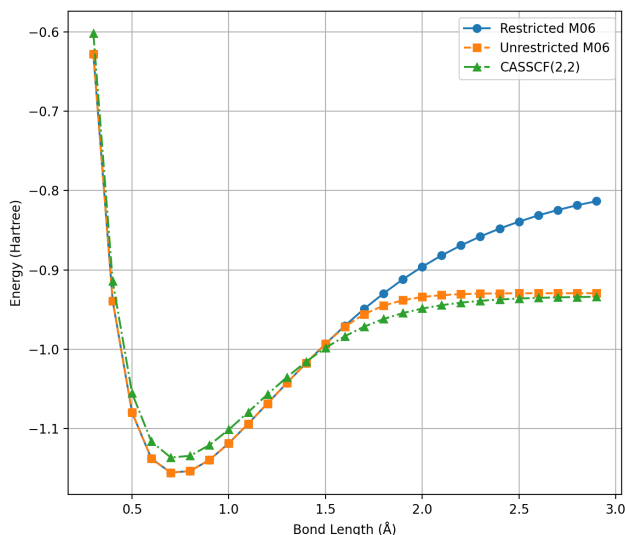

Figure 10: Dissociation Curve of dihydrogen molecule using RKS, UKS, and CASSCF(2,2) methods. RKS is inherently inadequate for capturing the true electronic structure nuances of bond-breaking events, as seen in the deviation from the CASSCF. In principle, multi-reference methods are essential for accurate modeling of such chemical reactions, ensuring a more holistic representation of the electronic correlation effects.