Supplementary Information: Exploring Active Learning Strategies for Excited State Dynamics: Application to Uracil

Juan Carlos San Vicente Veliz,[a] Mark DelloStritto,[b] and Spiridoula Matsika*[a]

* Corresponding
smatsika@temple.edu

This document provides additional information on the ML models. Section S1 shows the uracil structure. Section S2 describes the correlation of some of the test cases of our benchmark, such as the influence of hopping and weights on the test set of our training. Section S3 analyzes the influence of each of these benchmark cases on the MAE and KRC. Section S4 shows the final electronic state population for each of the cases. Section S5 shows the populations after systematic addition of CI S_2/S_1; S_1/S_0 data. Section S6 provides additional information on the performance of active learning in potential energy surfaces and conical intersections (Section 3.5 in manuscript). Section S7 shows a comparison of different threshold calculation criteria for active learning. Section S8 includes additional performance analysis based on hoping events. Section S9 includes the cartesian coordinates of the conical intersections and the g and h vectors used to construct the additional 882 QM/CASSCF points to enhance the model. Section S10 provides additional information about active learning cycles. Section S11, provides electronic state population comparison of including CI geometries in the training (AL-CI).

S1 Uracil structure

The minimum of the ground state of uracil used in this work is shown in Figure S1.

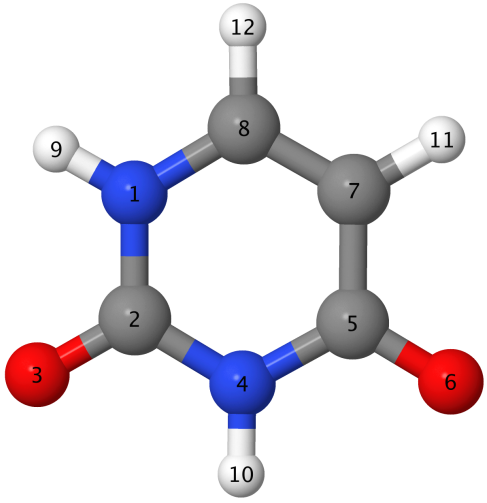

Figure S1: Optimized geometry of Uracil including labeling of the atoms.

S2 Benchmarking: hopping and weight influence

Correlation plots equivalent to Figure 1 in the main paper are given here (Figures S2,S3) for the additional cases in Table 1 of the main paper.

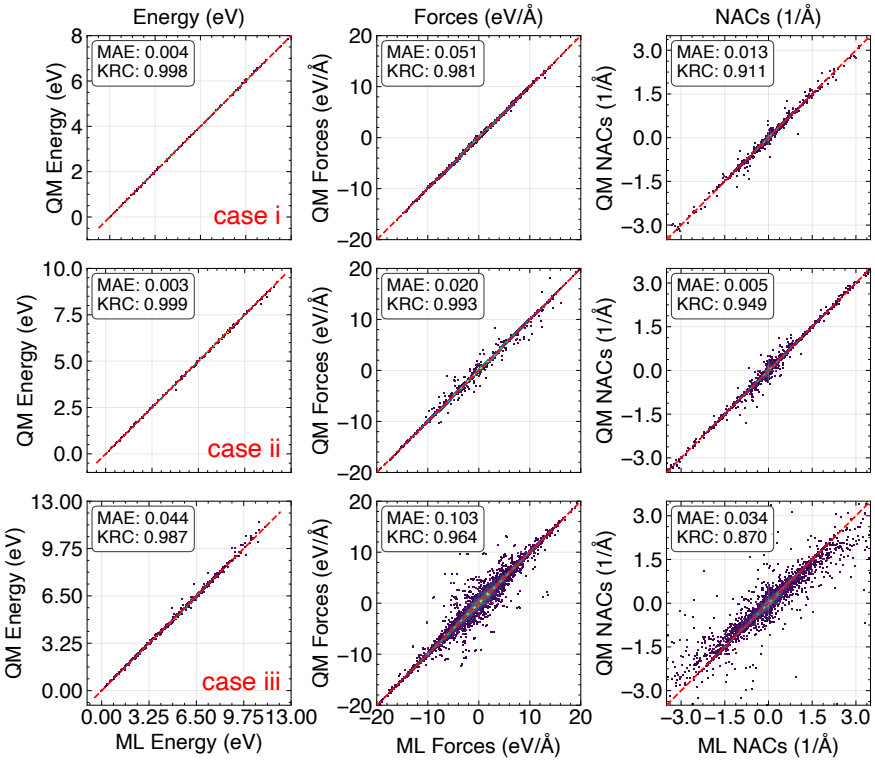

Figure S2: Correlation for Energy, Forces and Non-Adiabatic Couplings for test set. Case $i - iii$ for 1, 2 or 11 trajectories with hopping. Mean absolute error (MAE) are added for each property.

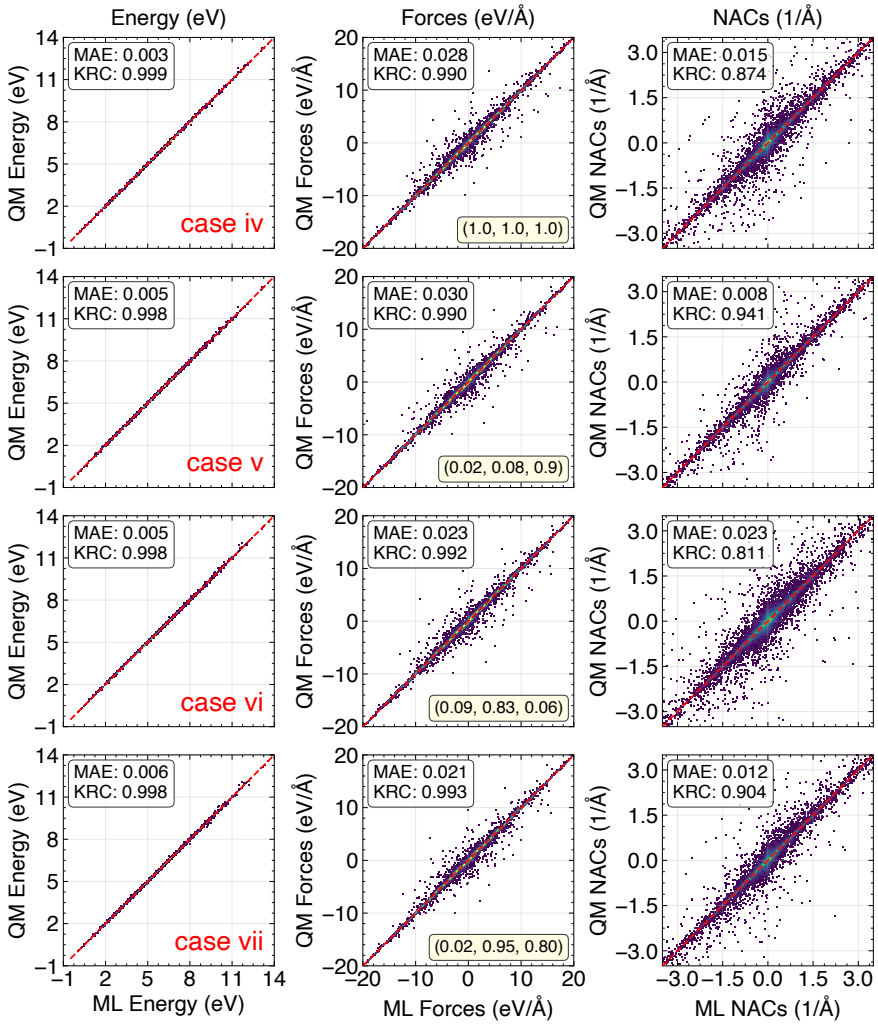

Figure S3: Correlation for Energy, Forces and Non-Adiabatic Couplings for test set. Case $iv - vii$ with different weights on the individual properties of the total loss function: (iv) (1.0, 1.0, 1.0), (v) (0.02, 0.08, 0.9), (vi) (0.09, 0.83, 0.06) and (vii) (0.02, 0.95, 0.80). Weights (w_1, w_2, w_3) for energies, forces and NACs. Mean absolute error (MAE) are added for each property.

S3 Hop, Weight and Loss function influence

Figure S4 is a summary figure of the influence of hop(open or closed circle), weight (closed triangle), and loss function (closed squared). This figure is a complementary figure to Table 1 in the MS. It highlights the impact hopping trajectories have on learning, since hopping events in a trajectory increase MAE and KRC. It also highlights that NAC is the most difficult to train accurately. The influence of changes to weights and loss function is more prominent for energies and forces than NACs.

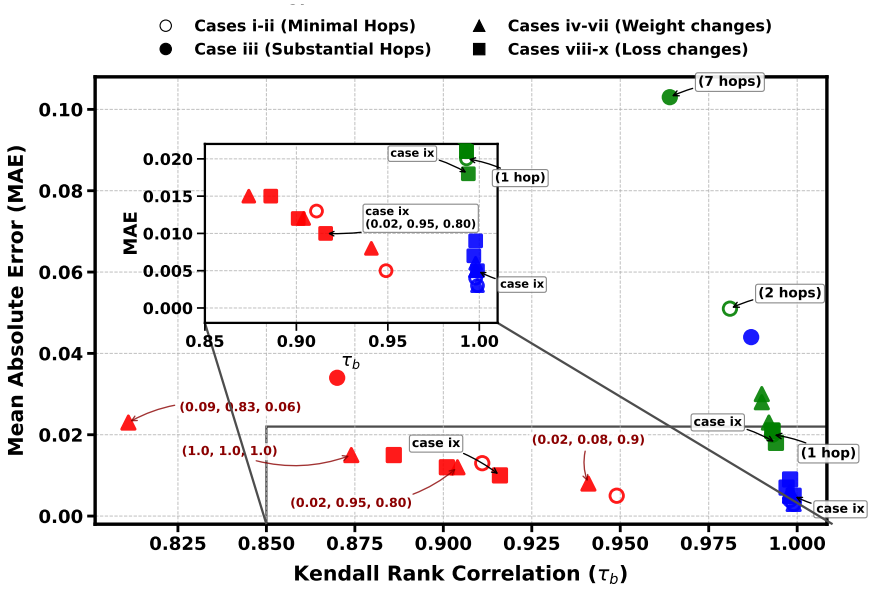

Figure S4: Mean absolute error (MAE) vs Kendall Rank Correlation (KRC) for all the cases in Table 1 of MS: Hop (open; closed circle), Weight (closed triangle) and Loss function (closed square).

S4 Benchmarking: Population: cases $iv - x$

The populations obtained from the various cases in Table 1 are shown in Figure S4, while Table S1 provides the final populations at 500 fs. There are no important differences between the cases.

Table S1: Comparison of electronic state populations at 500 fs for the different models compared to the QM reference.

Case	S_0	S_1	S_2
REF. (QM)	7	17	21
CASE *iv*	3	22	20
CASE *v*	4	24	17
CASE *vi*	4	20	21
CASE *vii*	2	27	16
CASE *viii*	1	26	18
CASE *ix*	2	25	18
CASE *x*	3	22	20

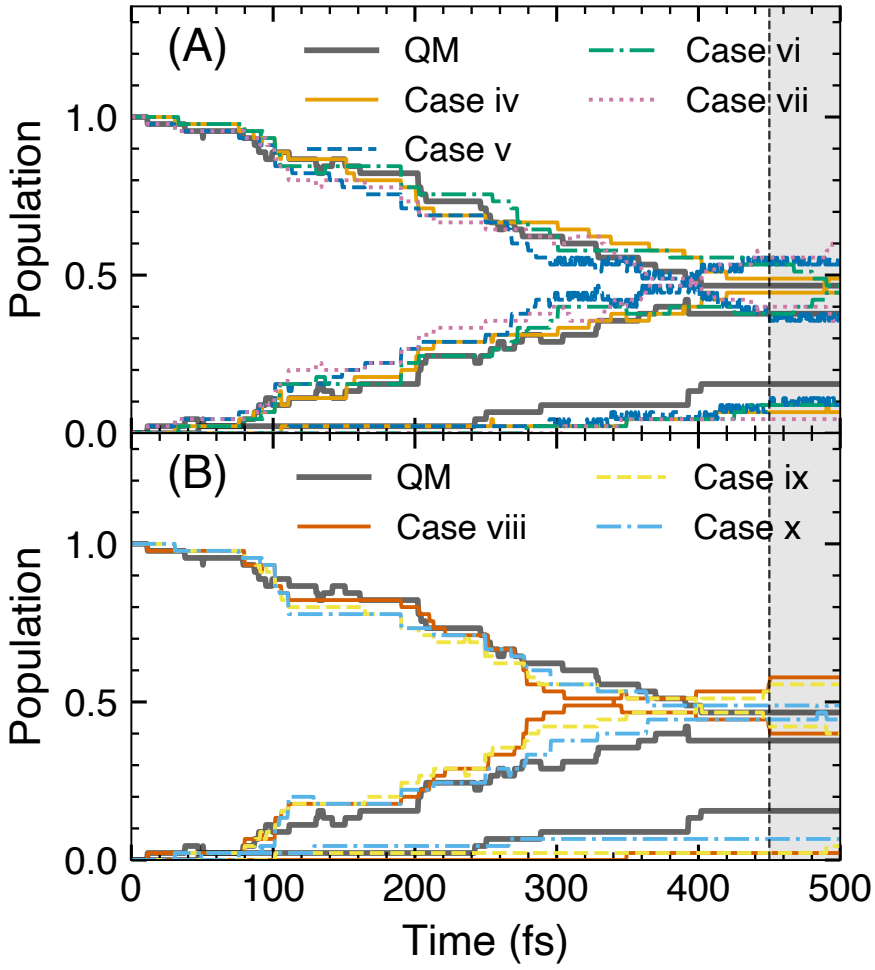

Figure S5: Population plot comparing, ML and 45 reference (QM) trajectories that completed successfully. ML model predictions $iv - vii$ (A) and $vii - x$ (B) are added for comparison.

S5 Influence of systematic addition of CI S_2/S_1; S_1/S_0 data

Figure S5 provides a direct comparison between the populations obtained when adding points along the branching plane and the populations obtained after active learning.

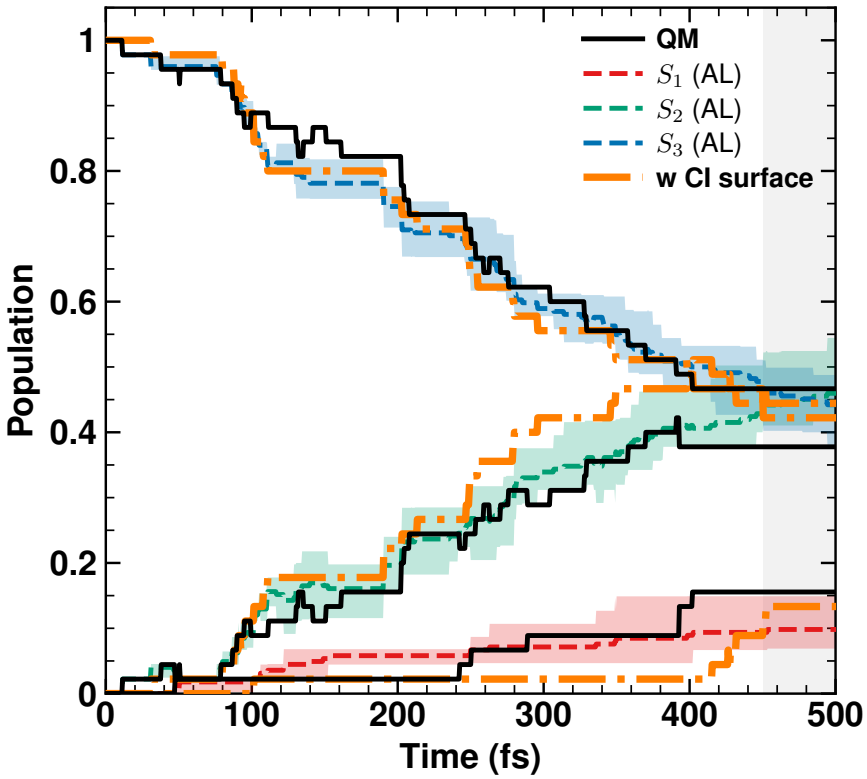

Figure S6: Electronic state populations as a function of time. QM population from 45 trajectories (black solid line) is compared with 95 % confidence of the 5 ML models after active learning: S_2 (shaded blue), S_1 (shaded green) and S_0 (shaded red). ML enhanced model with 882 CI structures from g and h vectors (dash orange line) is added for comparison.

S6 Performance of active learning in potential energy surfaces and conical intersections

Figure S6 shows a linear interpolation plot connecting the ground state minimum to the conical intersection. Geometries are taken from previous work.[1] Energy points from ML after active learning are compared to the QM energies. Figure S7 shows cuts along $g = 0$ and $h = 0$ for the conical intersection surfaces shown in Figure 9 of the main text.

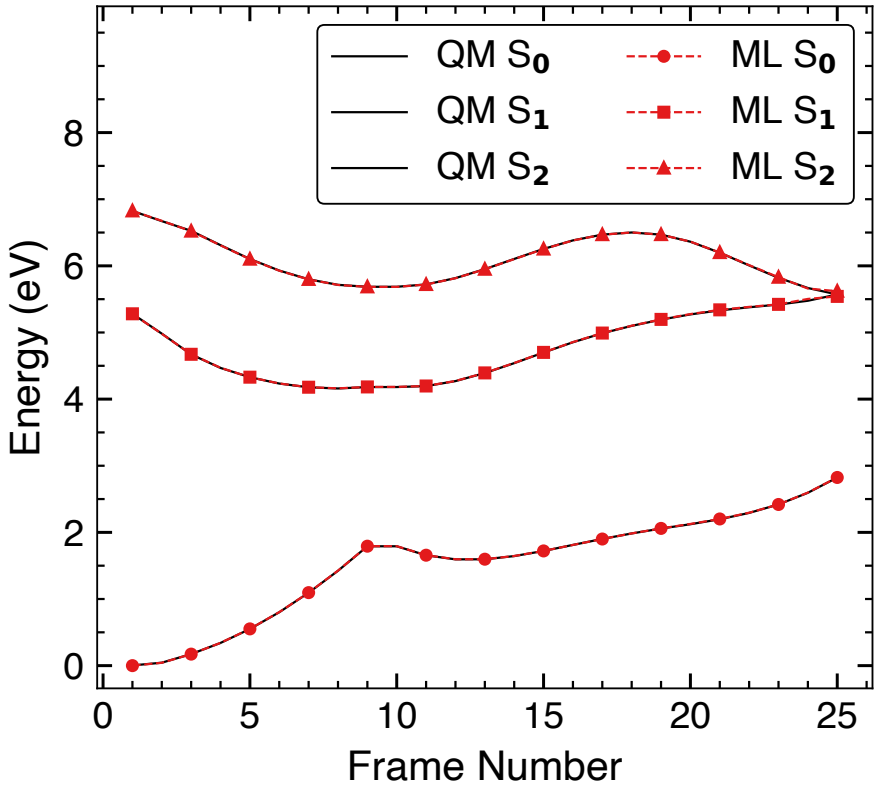

Figure S7: Linear interpolation from the ground state minimum geometry to the conical intersection S2/S1. We compared the CASSCF(12,9)/cc-pVDZ single point calculations(solid black line) with evaluating the same geometries with our ML model (red dash with symbol lines).

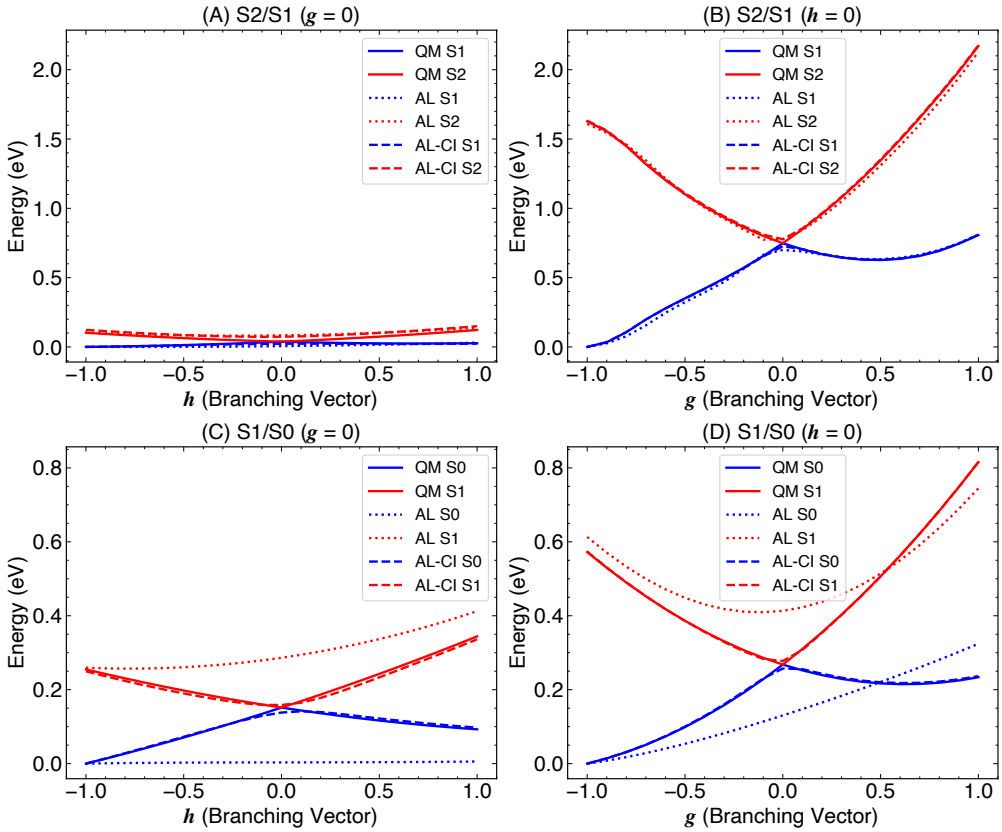

Figure S8: $h = 0$ and $g = 0$ cuts for QM CASSCF(12,9)/cc-pVDZ calculations (QM) and ML evaluation after 4 cycles of active learning (AL) and enhanced by the added CI data (AL-CI) along the g and h vectors. Both conical intersection S_2/S_1 and S_1/S_0 are explored. These are obtained at the CASSCF level.

S7 Changes in threshold: MAD vs MAE

Changes in threshold criteria are explored. Figure S9 shows the active learning performance for the first 3 cycles or iterations in the same test set. We compare the approach used in this work based on the MAE (solid lines) with a published approach based on MAD[2] (dash lines).

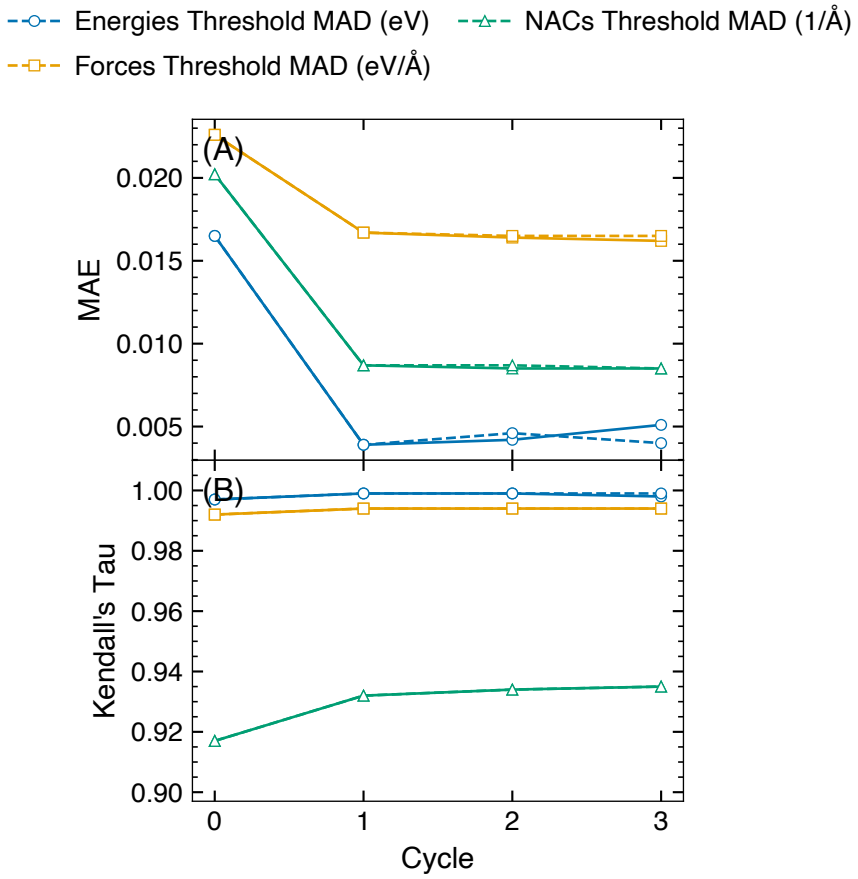

Figure S9: Test performance as a function of active learning cycles for threshold calculation based on MAE (solid lines) and Median Absolute Deviation MAD (dash lines). The average MAE (top) and Kendall Rank Coefficient (bottom) for the 5 models are plotted as a function of active learning cycle.

S8 Analysis of the hops

We explore the ML hop performance for 45 trajectories in a simulation window of 500 fs. Figure S10 shows the cumulative hops ($S_2 \rightarrow S_1$; $S_1 \rightarrow S_0$) and back-hops ($S_1 \rightarrow S_2$; $S_0 \rightarrow S_1$) between ML and QM. As discussed in the main text the cumulative hops show that forward hops are described well up to 200 fs and they are underestimated after that. Backward hops are always underestimated.

Figure S11 shows the comparison of the hop event times between QM and ML after active learning (AL). This figure indicates comparable times between the QM and ML times.

Lastly, Figure S12, shows the absolute mean error (MAE) performance for each of the 45 trajectories, along the bond length (C5-C6). Markers are added for QM hops from $S_2 \rightarrow S_1$ (black X), $S_1 \rightarrow S_0$ (red X) and the back hops $S_1 \rightarrow S_2$ (orange triangle) and $S_0 \rightarrow S_2$ (green triangle) products of a hop $S_2 \rightarrow S_0$ at the previous time step. No $S_0 \rightarrow S_1$ hops are found for this dataset. Figure S13 shows a 2d analysis of two degrees of freedom (C5-C6 bond) and (H12-C5-C6-H11) dihedral angle and measures the average combine MAE of the two properties $\Gamma = \sqrt{\left(\frac{|\Delta \text{Bond}|}{P_{98}}\right)^2 + \left(\frac{|\Delta \text{Dihedral}|}{P_{98}}\right)^2}\sqrt{2}$, where P_{98} is the 98 % of the data (for normalization purposes); symbol size represents the magnitude of this combine MAE.

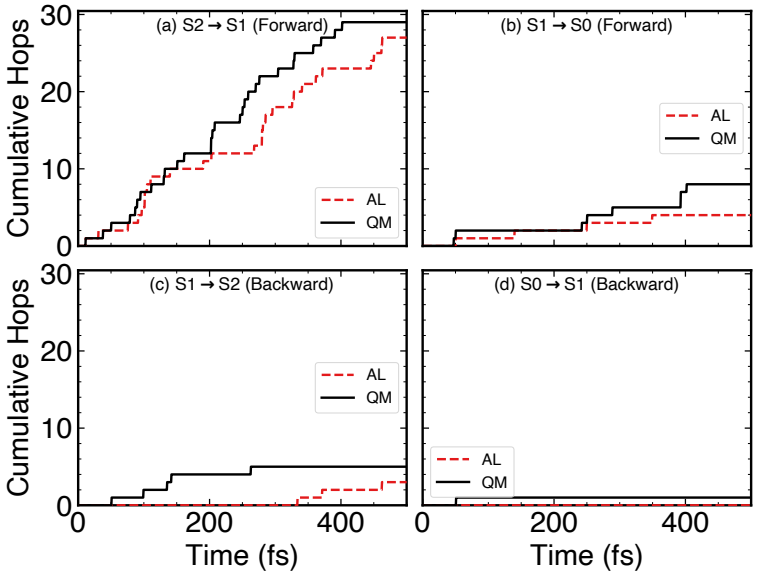

Figure S10: Cumulative hops as a function of time for QM (solid black line) and ML after active learning (AL) (dash red line) simulations. Simulation window is 0-500 fs for 45 trajectories evaluated.

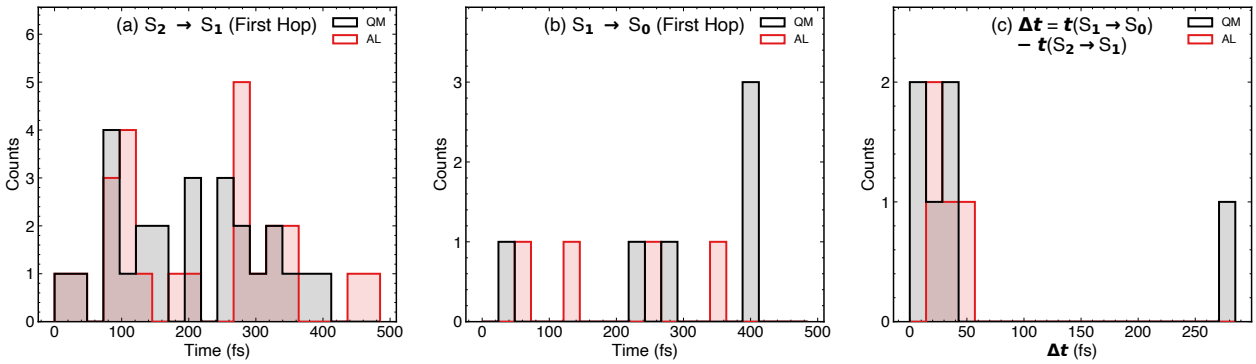

Figure S11: First hop histogram for $S_2 \rightarrow S_1$ (a); $S_1 \rightarrow S_0$ (b), and the time difference between the first two hops (c). We compare the QM reference with the predicted ML results after active learning (AL).

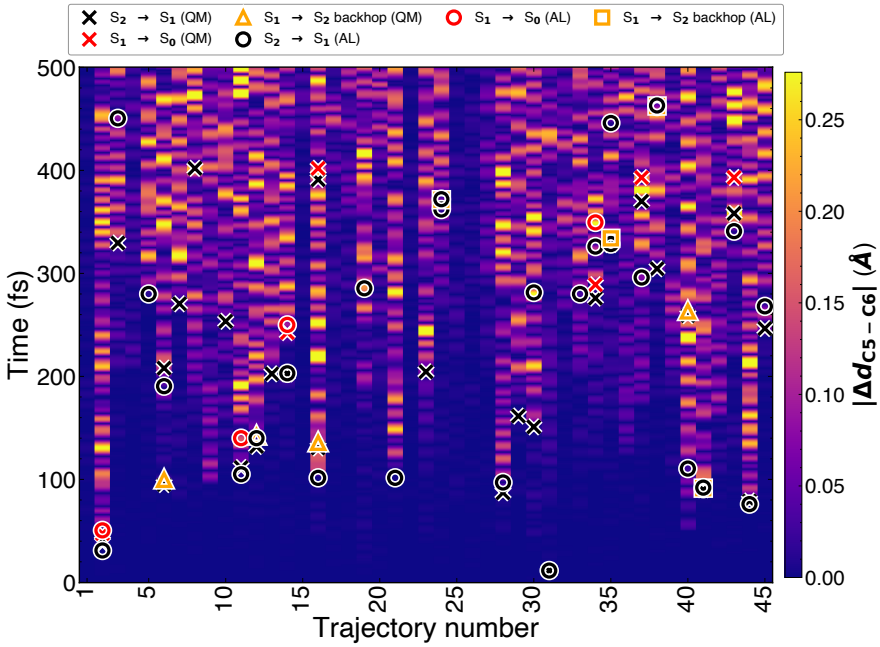

Figure S12: Mean absolute error (MAE) for the ML prediction after active learning (AL). MAE= $|X^{QM} - X^{ML}|$, where X is the bond length (C5-C6). We evaluated 45 trajectories from the training set for a simulation window of 500 fs. Markers are added for the QM hops from $S_2 \rightarrow S_1$ (black X), $S_1 \rightarrow S_0$ (red X), and back hops $S_1 \rightarrow S_2$ (orange triangle) and $S_0 \rightarrow S_2$ (green triangle) product of a hop $S_2 \rightarrow S_0$ at the previous time step. No $S_0 \rightarrow S_1$ hops are found for this dataset.

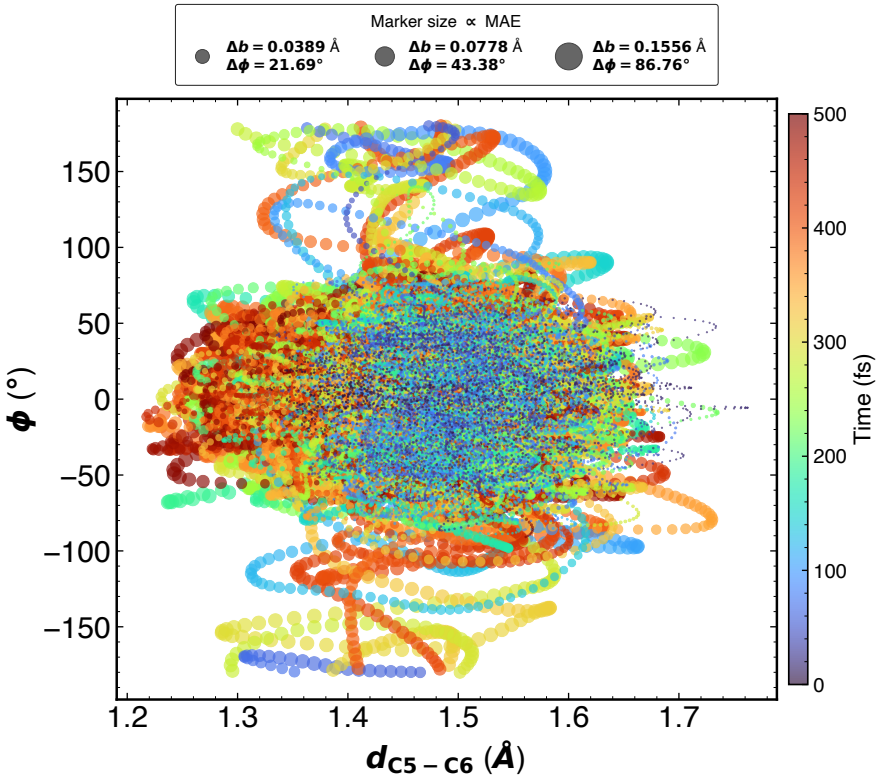

Figure S13: 2D map of the dihedral angle (H12-C5-C6-H11) and bond length (C5-C6) with the marker size representing the average combine MAE of the two properties $\Gamma = \sqrt{\left(\frac{|\Delta\text{Bond}|}{P_{98}}\right)^2 + \left(\frac{|\Delta\text{Dihedral}|}{P_{98}}\right)^2}\sqrt{2}$, where P_{98} is the 98 % of the data (for normalization purposes). The time component is added with the colour palette.

S9 Cartesian coordinates of conical intersections and the g and h vectors

Conical intersection structures and branching vectors g, h were obtained from our previous work at the CASSCF level of theory.[1]

Table S2: Optimized Cartesian coordinates (Å) for the S_1/S_0 conical intersection of uracil.

Atom	X (Å)	Y (Å)	Z (Å)
N	-0.68437806	-0.22664609	-0.02156979
N	0.70380627	1.65358206	0.13752205
C	-0.59540325	1.10866126	0.25749077
C	1.74812117	0.86266105	-0.12277018
C	1.70394810	-0.50608977	0.29576934
C	0.39793388	-1.13606153	-0.14459186
O	0.18308792	-2.25828666	-0.46711587
O	-1.52430935	1.82410308	0.48225986
H	-1.61360008	-0.58569444	-0.12577273
H	0.72382399	2.64779075	0.00090489
H	2.56692718	1.32664895	-0.66541764
H	1.56041723	-0.49046687	1.39044338

Table S3: Components of the g vector (Å) for the S_1/S_0 conical intersection of uracil.

Atom	X (Å)	Y (Å)	Z (Å)
N	-0.00425988	0.00188916	-0.00028046
N	0.02295571	-0.00235484	-0.01071055
C	-0.00859384	-0.01265792	0.00438159
C	-0.02771830	-0.00772070	0.01831482
C	0.02281812	-0.00127532	-0.01236158
C	-0.00086256	0.00349786	0.00397412
O	-0.00075143	0.00406937	-0.00018521
O	0.00286285	0.00224900	-0.00067206
H	-0.00104777	0.00122769	-0.00041805
H	-0.00089960	0.00222784	-0.00224900
H	0.00139703	0.00374657	0.00107423
H	-0.00590562	0.00509598	-0.00086256

Table S4: Components of the h vector (Å) for the S_1/S_0 conical intersection of uracil.

Atom	X (Å)	Y (Å)	Z (Å)
N	0.00245009	0.00030163	0.00104777
N	0.01085343	0.00076731	0.01314476
C	0.00016404	-0.00077789	0.00043922
C	-0.00245009	0.00398470	-0.00721269
C	0.00356665	0.00343965	0.00857796
C	0.00772070	-0.01485930	-0.00869967
O	-0.00396883	0.00067206	0.00037572
O	0.00092606	-0.00045509	-0.00073026
H	0.00071968	-0.00040217	-0.00041276
H	0.00022225	0.00107952	0.00370424
H	-0.00459326	-0.00243951	-0.01012845
H	-0.01560544	0.00867851	-0.00011642

Table S5: Optimized Cartesian coordinates (Å) for the S_2/S_1 conical intersection of uracil. .

Atom	X (Å)	Y (Å)	Z (Å)
N	-0.93056892	-0.06037916	0.31994213
N	1.00329823	1.15780720	-0.11266002
C	-0.40773559	1.13782242	-0.09050753
C	1.68367895	0.10207898	0.35631385
C	1.10951562	-1.21669367	-0.06985805
C	-0.27423428	-1.27404071	-0.04918900
O	-1.03139852	-2.19308332	-0.43129811
O	-1.02981066	2.12928752	-0.34018708
H	-1.93189742	-0.09049985	0.29683545
H	1.39914061	2.07712111	-0.20090709
H	2.67872767	0.25659744	0.74273497
H	1.60820264	-1.71206138	-0.89649891

Table S6: Components of the g vector (Å) for the S_2/S_1 conical intersection of uracil.

Atom	X (Å)	Y (Å)	Z (Å)
N	0.00851446	0.01291722	-0.00033867
N	0.02211432	-0.01608170	0.00459326
C	-0.01946843	-0.01140377	0.00126473
C	-0.00563574	0.02032570	-0.00378362
C	0.00846684	-0.00386829	0.00333911
C	-0.04219130	-0.03994759	-0.01315005
O	0.02383414	0.02590323	0.01293309
O	0.00539232	0.00236542	-0.00016404
H	0.00021167	-0.00150816	-0.00103190
H	0.00112715	0.00228605	-0.00069851
H	-0.00028046	0.00186800	0.00226488
H	-0.00208496	0.00714918	-0.00522827

Table S7: Components of the h vector (Å) for the S_2/S_1 conical intersection of uracil.

Atom	X (Å)	Y (Å)	Z (Å)
N	0.00076731	0.00211142	-0.00010584
N	0.00388945	-0.00670997	0.00373070
C	-0.00475730	-0.00063501	0.00036513
C	-0.00091018	0.01020783	0.00447155
C	-0.01049888	-0.00865734	-0.00231250
C	0.00893780	0.00242363	-0.00111127
O	0.00024871	-0.00101073	0.00003175
O	0.00031751	0.00116948	-0.00114302
H	-0.00047097	0.00016404	0.00029105
H	0.00033867	0.00050801	-0.00062972
H	0.00121711	0.00052389	-0.00285756
H	0.00092077	-0.00009525	-0.00073026

S10 Active learning

Table S8 provides information on the number of structures that are added after each cycle of active learning, along with the MAE and Kendall rank correlations.

Table S8: Number of added geometries and MAE of test set as a function of active learning cycle. MAE units are eV (Energy), eV/Å (Forces), and $Å^{-1}$ (NACs). Kendall rank correlations (τ_b) are shown in parentheses

Cycle	# QM geometries	Energy	Forces	NACs
0	–	0.0165 (0.997)	0.0226 (0.992)	0.0202 (0.917)
1	184	0.0039 (0.999)	0.0167 (0.994)	0.0087 (0.932)
2	140	0.0042 (0.999)	0.0164 (0.994)	0.0085 (0.934)
3	138	0.0051 (0.998)	0.0162 (0.994)	0.0085 (0.935)
4	125	0.0049 (0.999)	0.0161 (0.994)	0.0085 (0.936)
5	123	0.0051 (0.999)	0.0163 (0.994)	0.0085 (0.936)
6	111	0.0057 (0.999)	0.0164 (0.994)	0.0087 (0.936)
7	96	0.0049 (0.999)	0.0164 (0.994)	0.0083 (0.936)

S11 Population after active learning and introducing g and h vectors

Figure S14 shows the electronic state population as a function of time (0-1 ps). We compare the addition of CI geometries to the model after active learning (AL-CI).

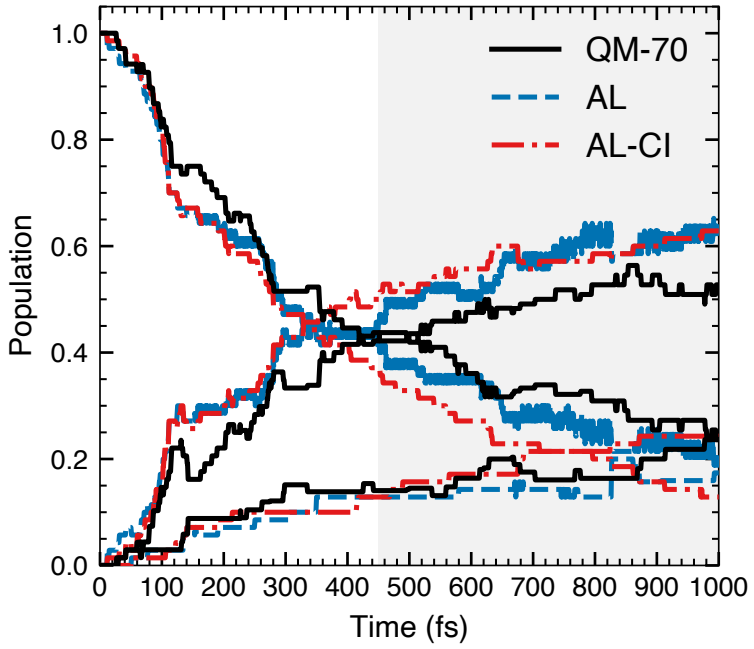

Figure S14: Electronic state population as a function of time (0-1000 fs). We compare QM simulations for 70 trajectories with ML predicted values after active learning (AL) and AL plus including CI geometries (AL-CI).

References

[1] P. Chakraborty, Y. Liu, T. Weinacht and S. Matsika, *Faraday Discuss.*, 2021, **228**, 266–285.

[2] M. d. O. Bispo, R. Souza Mattos, M. J. Pinheiro, B. C. Garain, P. O. Dral and M. Barbatti, *J. Chem. Theor. Comp.*, 2025, **21**, 11390–11400.