Supporting Information for the Paper

Automated Reaction Transition State Search for Bimolecular Liquid-Phase Reactions Using Internal Coordinates: A Test Case for Neutral Hydrolysis

Leen Fahoum[a], Alon Grinberg Dana[a,b,*]

[a]Wolfson Department of Chemical Engineering, Technion – Israel Institute of Technology, Haifa 3200003, Israel
[b]Grand Technion Energy Program (GTEP), Technion – Israel Institute of Technology, Haifa 3200003, Israel
* Corresponding author, alon@technion.ac.il

Section S1: Figures and Tables

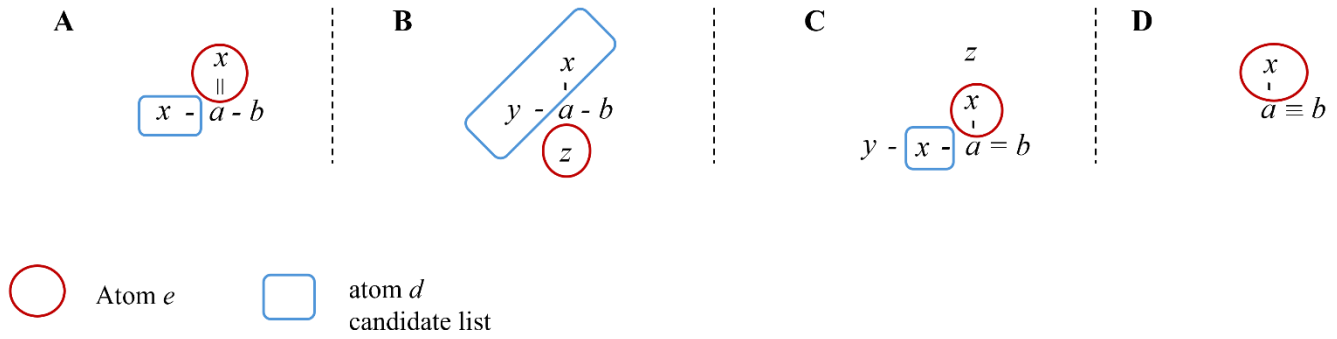


Figure S1. Illustration of e and d atom assignments in different neighboring atom scenarios. Atoms a and b are defined in the main text. Atoms x, y, and z are generic substituents attached to atom a, ranked according to their intrinsic electronegativity values (χ), on the Pauling scale, with z having the highest χ, followed by y, and then x with the lowest.

Fig. S1 illustrates four example scenarios for determining the atoms labeled e and d based on their ${\chi_i}^{eff}$:

- Case A: Atom a has two neighboring atoms (in addition to b), with identical intrinsic electronegativity values (χ). However, one of them forms a double bond with atom a, while the other is connected via a single bond. As a result, the atom forming the double bond has a higher effective electronegativity (${\chi_i}^{eff}$, based on Eq. 1) and is labeled e. The other is assigned as atom d.
- Case B: Atom a is bonded to three different neighboring atoms (x, y, and z), each via a single bond. In this case, the atom with the highest intrinsic electronegativity (atom z) is labeled as atom e, while the next in rank (atom y) is assigned as atom d, and atom x is retained as an alternative atom d candidate for generating additional TS guesses if initial attempts fail.
- Case C: Atom a is bonded to two atoms (besides atom b) with identical intrinsic electronegativity and bond order, making their ${\chi_i}^{eff}$ values equal. To break the tie, the total ${\chi_i}^{eff}$ of each atom's own neighbors is considered. One of the atoms is bonded to atom z (higher intrinsic electronegativity), while the other is bonded to atom y (lower intrinsic electronegativity). The atom connected to z is therefore labeled atom e and the other is labeled atom d.
- Case D: Atom a has only one neighboring atom in addition to atom b. The neighboring atom is labeled as e, and no d atom is defined.

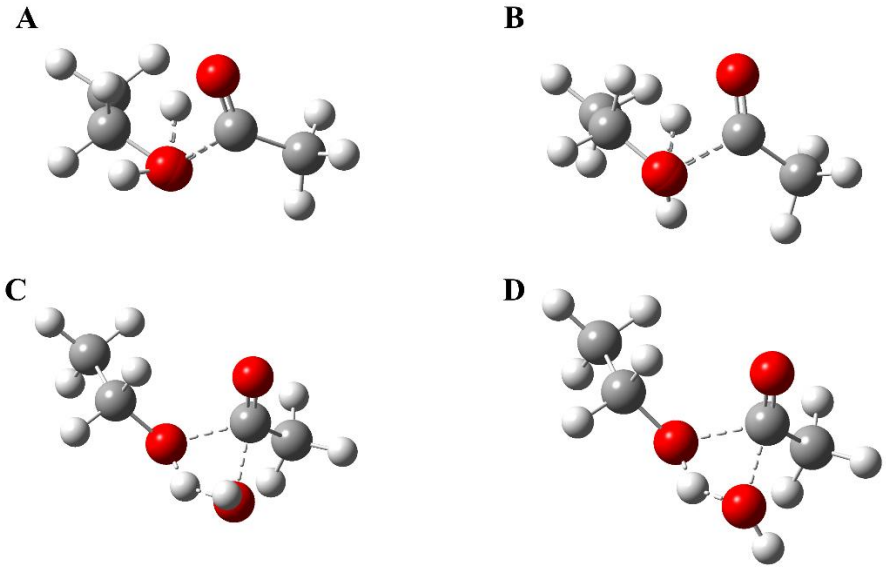


Figure S2. Example that illustrates the effect of dihedral sign combinations on water placement in TSG generation. This example shows an ester class hydrolysis, belonging to the carbonyl-based family, however this behavior is valid for all three different studied families. Four initial TS guesses were generated from the same substrate using different sign assignments for the two water-positioning dihedrals (φ_1 and φ_3): A. (+, +), B. (+, -), , C. (-, +), , and D. (-, -), respectively.

Fig. S2 shows that configurations A and B lead to severe steric clashes between the ester oxygen and the water oxygen. In contrast, C and D represent sterically feasible approaches. After full DFT refinement, configuration D resulted in the lowest energy structure, and therefore this sign combination was selected as the final TSG for this reaction. This example demonstrates that the sign of the dihedral angles cannot be treated as symmetrical equivalent. Even though the magnitudes φ_1 and $\varphi_3$may lie within similar ranges, only specific sign pairings produce sterically reasonable TSG. Therefore, both signs must be exhaustively sampled during TSG generation, since the chemically valid approach direction depends on the substrate's steric and electronic environment and the ultimately favorable TS is determined only after quantum refinement.

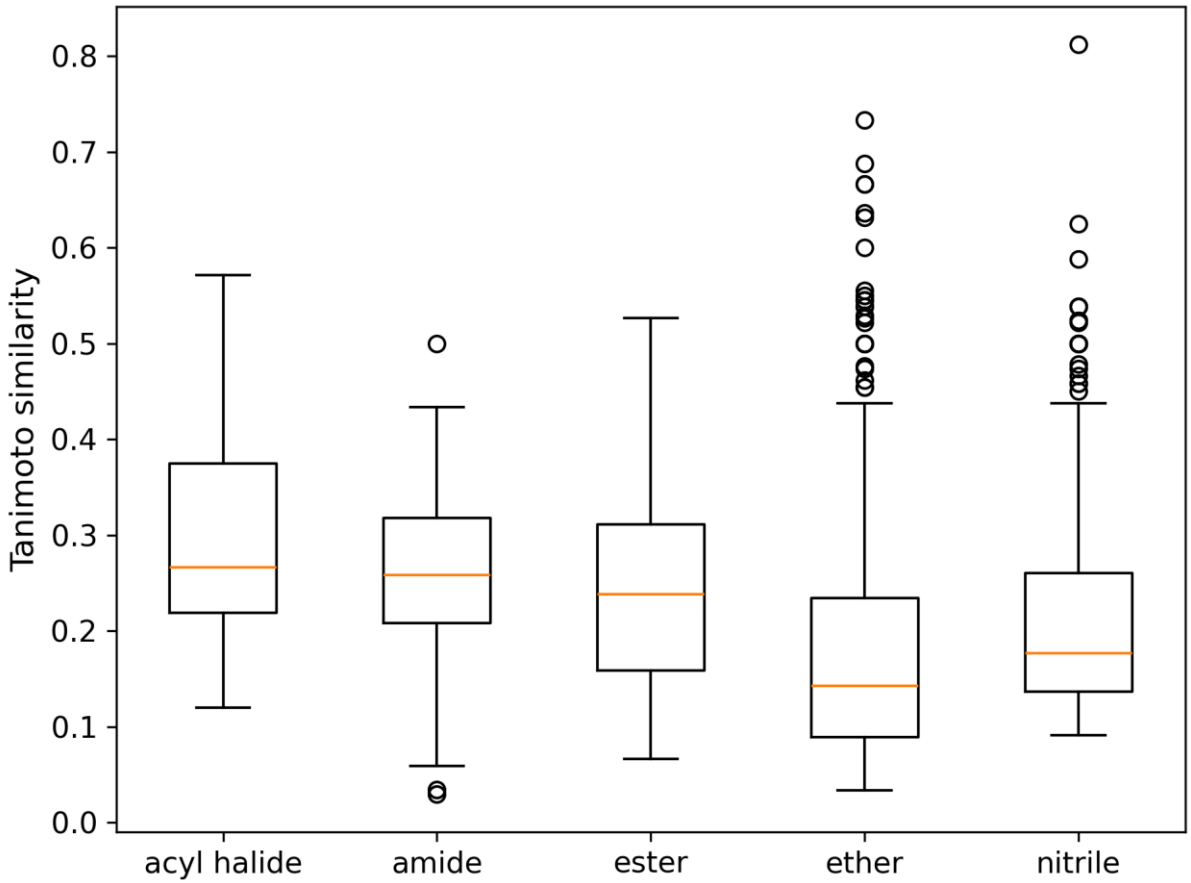


Figure S3. Within-class similarity distributions of pairwise Tanimoto similarity values (Morgan fingerprints, radius = 2) for the validation cases jobs. Circles represent outlier similarity values.

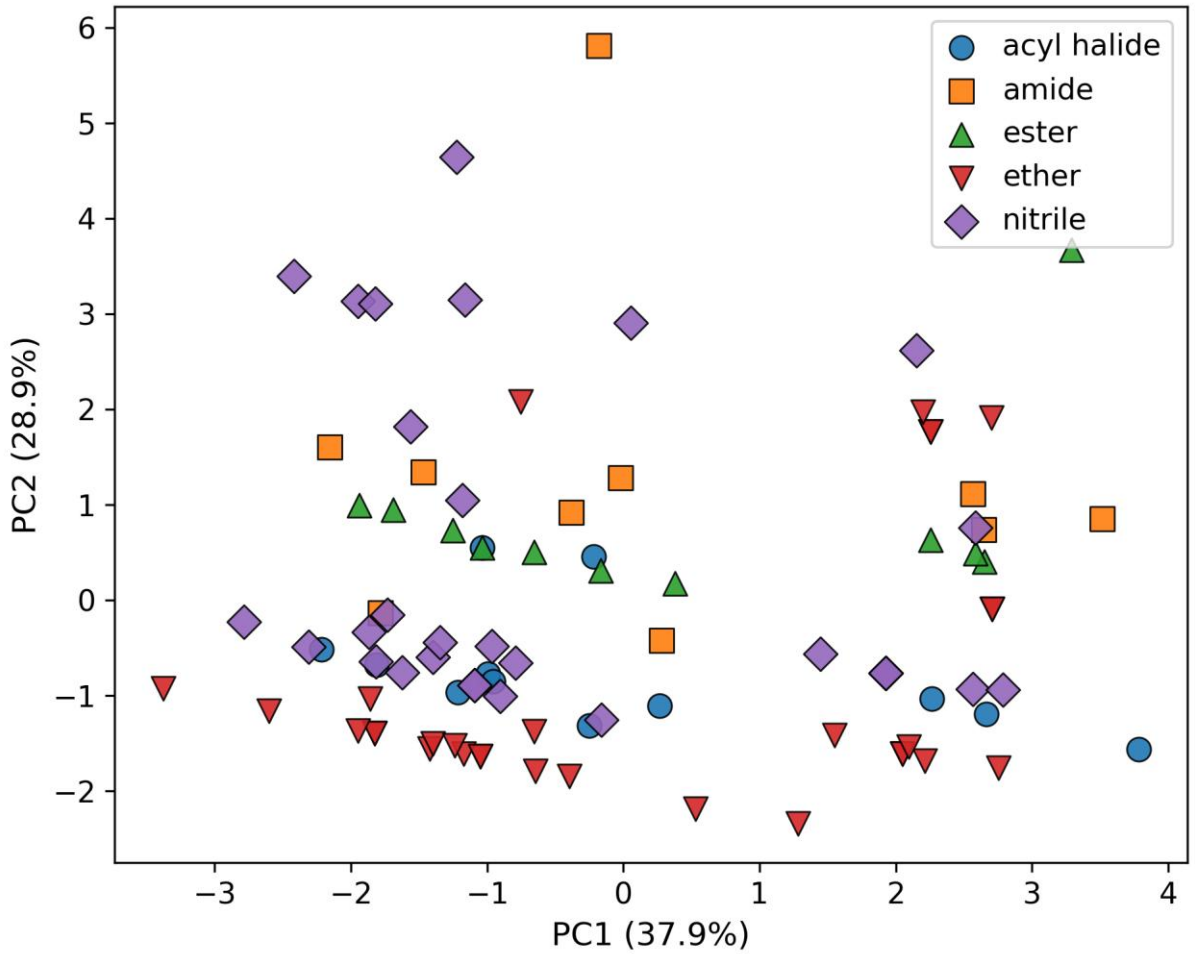


Figure S4. PCA for the validation cases jobs of nine standardized RDKit descriptors. PC1 (28.9%) reflects molecular size and hydrophobicity; PC2 (37.9%) contrasts polarity and H-bonding with aromaticity for all studied classes.

Table S1: Ester class hydrolysis transition state (TS) parameters manually extracted from the optimized geometries of the study cases. All TSs were optimized at the CBS-QB3 level of theory using Gaussian 09 and each structure was verified to contain a single imaginary frequency (provided below) corresponding to the reaction coordinate. ' r ' denotes a bond length, ' α ' represents a bond angle, and ' φ ' refers to a dihedral angle. The labeling method of each parameter is explained in the main text. The averages of $\varphi_1$and $\varphi_3$were computed using their absolute values. In addition to mean $\pm$SD, each parameter is also reported with its median and interquartile range (IQR = Q1–Q3).

Molecule SMILES	r_1(Å)	α_1(°)	φ_1(°)	r_2(Å)	α_2(°)	φ_2(°)	r_3(Å)	α_3(°)	φ_3(°)	**Original** r_{a-b}(Å)	**TS** r_{a-b}(Å)	**a-b bond stretch factor**	**Imag. Freq.** (cm^{-1})
O=C(OCC)C	1.84	76.65	-140.22	1.22	70.86	-0.61	0.96	111.86	-106.93	1.35	1.82	1.34	-1191.29
CC(=O)OC1=CC=CC=C1	1.85	73.36	147.25	1.13	75.45	2.36	0.97	109.50	110.33	1.37	2.01	1.46	-877.58
C1CC(=O)O1	1.86	75.90	150.27	1.15	71.95	4.66	0.97	109.79	-99.44	1.38	1.87	1.36	-1142.81
CSC(=O)C	1.70	79.89	-143.30	1.17	84.01	-3.10	0.97	108.50	102.95	1.80	2.45	1.36	-1117.18
COP(=O)(OC)O	2.04	68.02	-147.80	1.32	74.17	7.12	0.97	106.08	109.59	1.59	1.93	1.22	-804.24
COC=O	1.76	79.46	138.53	1.21	72.14	4.88	0.96	111.49	-98.83	1.34	1.74	1.30	-1272.43
CCCOC(=O)	1.78	78.79	-139.92	1.22	71.51	-0.57	0.97	112.41	-105.00	1.34	1.75	1.30	-1239.14

CC(C)OC(=O)C	1.84	76.62	-139.77	1.22	70.60	-0.34	0.96	112.15	-107.01	1.35	1.81	1.34	-1173.01
CCCC(=O)OC	1.85	76.51	-140.34	1.23	70.35	-0.99	0.96	111.08	-106.97	1.35	1.82	1.35	-1146.60
COC(=O)C1=CC=CC=C1	1.82	76.56	138.46	1.22	71.87	2.26	0.97	111.37	104.83	1.35	1.82	1.35	-1184.19
AVG	1.83	76.18	142.59	1.21	73.29	1.57	0.97	110.42	105.19			1.34	
SD	0.09	3.44	4.32	0.05	4.09	3.23	0.00	1.98	3.86			0.06	
Median	1.84	76.59	140.28	1.22	71.91	0.96	0.97	111.23	105.97			1.35	
[Q1, Q3]	[1.79, 1.85]	[76.06, 78.25]	[139.81, 146.26]	[1.18, 1.22]	[71.02, 73.95]	[-0.60, 4.08]	[0.96, 0,97]	[109.57, 111.46]	[103.42, 107.00]			[1.31, 1,36]	
IQR	0.06	2.19	6.46	0.04	2.93	4.68	0.00	1.89	3.58			0.04	

Table S2: Amide class hydrolysis transition state (TS) parameters manually extracted from the optimized geometries of the study cases. All TSs were optimized at the CBS-QB3 level of theory using Gaussian 09 and each structure was verified to contain a single imaginary frequency (provided below) corresponding to the reaction coordinate. ' r ' denotes a bond length, ' α ' represents a bond angle, and ' φ ' refers to a dihedral angle. The labeling method of each parameter is explained in the main text. The averages of $\varphi_1$and $\varphi_3$were computed using their absolute values. In addition to mean $\pm$SD, each parameter is also reported with its median and interquartile range (IQR = Q1–Q3).

Molecule SMILES	r_1(Å)	α_1(°)	φ_1(°)	r_2(Å)	α_2(°)	φ_2(°)	r_3(Å)	α_3(°)	φ_3(°)	**Original r_{a-b}(Å)**	**TS r_{a-b}(Å)**	**a-b bond stretch factor**	**Imag. freq. (cm^{-1})**
O=C(N)C	1.89	80.90	122.79	1.38	67.22	-4.73	0.96	115.47	94.60	1.37	1.63	1.19	-1024.59
CC(=O)NC	1.97	78.87	128.21	1.43	65.58	-8.02	0.96	127.67	102.83	1.37	1.61	1.18	-902.72
CC(=O)N(C)C	1.96	79.72	-125.44	1.38	65.01	-4.45	0.96	115.27	94.64	1.38	1.64	1.19	-947.46
C1CCCC(=O)NC1	1.95	79.86	-127.66	1.42	65.69	3.71	0.96	124.75	-102.69	1.37	1.60	1.17	-785.37
CC(=S)N	1.74	85.42	127.27	1.27	70.71	4.77	0.97	113.26	-95.87	1.35	1.53	1.14	-1460.01
C(=O)N	1.85	81.94	128.25	1.36	68.06	5.11	0.97	116.46	-96.02	1.36	1.60	1.18	-1140.45
C1=CC=CC=C1C(=O)N	1.91	80.12	129.97	1.38	66.87	6.38	0.96	118.23	-92.69	1.37	1.61	1.17	-1014.19
CCCC(=O)N	1.91	80.83	128.10	1.41	66.65	4.54	0.96	115.58	-94.35	1.37	1.62	1.19	-947.45

C(=O)NC	1.92	80.06	-128.02	1.39	66.12	7.80	0.96	124.56	-100.59	1.36	1.59	1.17	-942.85
CC(=O)N(CC)CC	2.04	77.72	-124.82	1.44	63.47	8.93	0.96	129.09	-101.66	1.37	1.61	1.17	-620.78
AVG	1.91	80.54	127.05	1.39	66.54	2.40	0.96	120.03	97.59			1.17	
SD	0.08	2.06	2.10	0.05	1.94	5.90	0.00	5.85	3.90			0.02	
Median	1.92	80.09	127.84	1.39	66.38	4.65	0.96	117.34	95.94			1.18	
[Q1, Q3]	[1.89, 1.96]	[79.75, 80.76]	[125.90, 128.18]	[1.38, 1.42]	[65.61, 66.82]	[-2.41, 6.06]	[0.96, 0.96]	[115.50, 124.71]	[94.61, 101.39]			[1.17, 1,18]	
IQR	0.06	1.01	2.28	0.03	1.21	8.47	0.00	9.21	6.79			0.01	

Table S3: Acyl halides class hydrolysis transition state (TS) parameters manually extracted from the optimized geometries of the study cases. All TSs were optimized at the CBS-QB3 level of theory using Gaussian 09 and each structure was verified to contain a single imaginary frequency (provided below) corresponding to the reaction coordinate. ' r ' denotes a bond length, ' α ' represents a bond angle, and ' φ ' refers to a dihedral angle. The labeling method of each parameter is explained in the main text. The averages of $\varphi_1$and $\varphi_3$were computed using their absolute values. In addition to mean $\pm$SD, each parameter is also reported with its median and interquartile range (IQR = Q1–Q3).

Molecule	r_1(Å)	α_1(°)	φ_1(°)	r_2(Å)	α_2(°)	φ_2(°)	r_3(Å)	α_3(°)	φ_3(°)	**Original** r_{a-b}(Å)	**TS** r_{a-b}(Å)	**a-b bond stretch factor**	**Imag. freq.** (cm^{-1})
CC(=O)Cl	1.80	75.84	153.17	1.03	87.45	5.01	0.96	106.28	113.05	1.80	2.56	1.42	-358.05
C1=CC=CC=C1C(=O)Cl	1.87	75.16	151.14	1.03	87.76	2.10	0.97	106.07	-102.58	1.84	2.58	1.40	-275.55
CCC(=O)Cl	1.80	75.22	153.84	1.04	87.06	5.14	0.97	105.12	112.64	1.85	2.69	1.46	-299.31
CCCC(=O)Cl	1.81	74.36	152.55	1.04	88.48	-0.91	0.97	105.62	-105.34	1.85	2.63	1.42	-270.57
CCCCC(=O)Cl	1.81	74.31	152.59	1.04	88.45	-0.90	0.97	105.61	-105.31	1.85	2.63	1.43	-269.38
ClC(C(=O)Cl)(Cl)Cl	1.64	78.93	147.05	1.07	90.42	0.75	0.97	108.30	-105.03	1.78	2.45	1.38	-457.90
C1CCCCC1C(=O)Cl	1.84	73.36	153.96	1.04	88.26	-1.05	0.97	105.35	-105.12	1.86	2.68	1.44	-257.05
CS(=O)Cl	1.81	76.30	150.06	1.03	87.56	-1.00	0.96	106.08	-104.81	2.19	2.84	1.30	-318.75

ClCC(=O)Cl	1.72	76.79	155.17	1.04	90.45	1.33	0.97	107.54	-106.75	1.83	2.55	1.39	-309.39
CC(C)C(=O)Cl	1.82	74.06	152.91	1.04	88.37	-0.93	0.97	105.55	-105.23	1.85	2.65	1.43	-267.16
AVG	1.79	75.43	152.24	1.04	88.43	0.95	0.97	106.15	106.59			1.41	
SD	0.07	1.62	2.32	0.01	1.16	2.45	0.00	1.01	3.45			0.04	
Median	1.81	75.19	152.75	1.04	88.31	-0.08	0.97	105.85	105.27			1.42	
[Q1, Q3]	[1.80, 1,82]	[74.32, 76.08]	[151.49, 153.61]	[1.04, 1.04]	[87.61, 88.47]	[-0.93, 1.24]	[0.97, 0.97]	[105.57, 106.13]	[105.05, 106.27]			[1.40, 1.43]	
IQR	0.02	1.76	2.12	0.01	0.86	2.16	0.00	0.57	1.22			0.03	

Table S4: Ether hydrolysis transition state (TS) parameters manually extracted from the optimized geometries of the study cases. All TSs were optimized at the CBS-QB3 level of theory using Gaussian 09 and each structure was verified to contain a single imaginary frequency (provided below) corresponding to the reaction coordinate. ' r ' denotes a bond length, ' α ' represents a bond angle, and ' φ ' refers to a dihedral angle. The labeling method of each parameter is explained in the main text. The averages of $\varphi_1$and $\varphi_3$were computed using their absolute values. In addition to mean $\pm$SD, each parameter is also reported with its median and interquartile range (IQR = Q1–Q3).

Molecule	r_1(Å)	α_1(°)	φ_1(°)	r_2(Å)	α_2(°)	φ_2(°)	r_3(Å)	α_3(°)	φ_3(°)	**Original** r_{a-b}(Å)	**TS** r_{a-b}(Å)	**a-b bond stretch factor**	**Imag. freq.** (cm^{-1})
CC(C)(C)OC	2.12	66.25	95.05	1.10	73.57	-0.51	0.96	105.91	100.58	1.41	2.18	1.54	-855.08
C1CO1	2.21	69.42	-91.54	1.03	71.80	-1.26	0.96	104.74	-101.89	1.43	2.10	1.47	-532.90
C1CCOC1	2.21	63.39	108.05	1.11	72.22	-1.31	0.96	106.84	101.53	1.43	2.30	1.60	-845.07
C1=CC=C(C=C1)COCC2=CC=CC=C2	2.33	61.89	99.75	1.04	72.89	0.16	0.96	105.09	-104.74	1.43	2.46	1.73	-392.01
COCCOC	2.18	64.41	96.69	1.12	71.70	-1.37	0.96	107.78	103.15	1.42	2.23	1.57	-680.10
CCOCC	2.17	64.29	-98.43	1.11	72.41	-0.04	0.96	107.08	104.08	1.42	2.25	1.59	-919.56
CCOC	2.16	64.78	96.92	1.11	72.41	-0.06	0.96	106.63	103.92	1.42	2.23	1.57	-939.07
COCCC	2.15	66.78	100.20	1.10	73.21	-0.55	0.97	106.69	100.94	1.41	2.15	1.52	-889.82

COc1ccccc1	1.79	75.51	-137.53	1.19	74.72	7.10	0.96	115.27	-99.32	1.36	1.79	1.31	-1263.04
NCCOC	2.13	66.14	93.62	1.09	73.04	0.85	0.96	106.00	101.72	1.41	2.18	1.55	-820.85
AVG	2.14	66.29	101.78	1.10	72.80	0.30	0.96	107.20	102.19			1.55	
SD	0.14	3.84	13.34	0.04	0.90	2.49	0.00	2.98	1.74			0.10	
Median	2.16	65.46	97.67	1.10	72.65	-0.28	0.96	106.66	101.80			1.56	
[Q1, Q3]	[2.13, 2.20]	[64.32, 66.65]	[95.46, 100.09]	[1.09, 1.11]	[72.26, 73,17]	[-1.09, 0.11]	[0.96, 0.96]	[105.94, 107.02]	[101.09, 103.72]			[1.53, 1.58]	
IQR	0.07	2.33	4.63	0.02	0.90	1.20	0.00	1.08	2.64			0.06	

Table S5: Nitrile hydrolysis transition state (TS) parameters manually extracted from the optimized geometries of the study cases. All TSs were optimized at the CBS-QB3 level of theory using Gaussian 09 and each structure was verified to contain a single imaginary frequency (provided below) corresponding to the reaction coordinate. ' r ' denotes a bond length, ' α ' represents a bond angle, and ' φ ' refers to a dihedral angle. The labeling method of each parameter is explained in the main text. The averages of φ_1 and φ_3 were computed using their absolute values. In addition to mean $\pm$SD, each parameter is also reported with its median and interquartile range (IQR = Q1–Q3).

Molecule	r_1(Å)	α_1(°)	φ_1(°)	r_2(Å)	α_2(°)	φ_2(°)	r_3(Å)	α_3(°)	φ_3(°)	**Original** r_{a-b}(Å)	**TS** r_{a-b}(Å)	**a-b bond stretch factor**	**Imag. freq.** (cm^{-1})
CC#N	1.83	97.31	-174.36	1.32	58.09	-1.81	0.97	114.42	104.05	1.15	1.20	1.04	-1940.22
CCC#N	1.83	97.43	174.31	1.32	58.07	1.83	0.97	114.24	-103.82	1.15	1.20	1.04	-1940.01
CCCC#N	1.86	96.23	-173.53	1.34	57.26	-1.92	0.97	114.85	103.72	1.15	1.20	1.04	-1900.90
C1=CC=CC=C1C#N	1.83	97.01	177.75	1.32	58.10	1.21	0.97	112.63	-102.34	1.16	1.21	1.04	-1956.87
CC(C)C#N	1.85	96.39	174.37	1.34	57.33	1.83	0.97	114.56	-103.43	1.15	1.20	1.04	-1907.48
ClCC#N	1.81	98.63	-172.52	1.32	58.06	-1.69	0.97	114.10	104.67	1.15	1.20	1.04	-1961.84
C1CCCCC1C#N	1.86	95.90	175.35	1.34	57.23	2.10	0.97	115.27	-103.57	1.15	1.20	1.04	-1893.65

CCC(C)C#N	1.85	96.22	173.74	1.34	57.37	1.80	0.97	114.68	-103.44	1.15	1.20	1.04	-1904.68
C1=CC=CC=C1CCC#N	1.82	97.48	-174.36	1.32	58.16	-1.89	0.97	114.39	103.80	1.15	1.20	1.04	-1943.12
N#CC(C#N)	1.78	99.65	-172.56	1.30	59.26	-1.62	0.97	113.67	104.87	1.15	1.20	1.04	-1984.96
AVG	1.83	97.23	174.29	1.32	57.89	-0.02	0.97	114.28	103.77			1.04	
SD	0.03	1.18	1.50	0.01	0.62	1.88	0.00	0.72	0.70			0.00	
Median	1.83	97.16	174.33	1.32	58.06	-0.20	0.97	114.41	0.69			1.04	
[Q1, Q3]	[1.82, 1.85]	[96.27, 97.47]	[173.58, 174.37]	[1.32, 1.34]	[57.34, 58.10]	[-1.78, 1.82]	[0.97, 0.97]	[114.14, 114.65]	[103.47, 104.00]			[1.04, 1.04]	
IQR	0.03	1.19	0.79	0.02	0.75	3.60	0.00	0.51	0.52			0.00	

Table S6: Comparison between the initial TS guess and the optimized TS geometry for methyl acetate hydrolysis. The job was run in ARC using the heuristic-based TS generation adapter with all calculations performed at the ωB97X-D/jul-cc-pVTZ level of theory and the SMD solvation model for water.
' r ' denotes a bond length, ' α ' represents a bond angle, and ' φ ' refers to a dihedral angle. The labeling method of each parameter is explained in the main text. The reported similarity values correspond to those defined in Section 3 of the article, representing the normalized geometric agreement between the guessed and optimized internal coordinates.

	Bond length				**Angels**			**Dihedral Angles**		
	$\boldsymbol{r_1}$(Å)	$\boldsymbol{r_2}$(Å)	$\boldsymbol{r_3}$(Å)	$\boldsymbol{r_{a-b}}$(Å)	$\boldsymbol{\alpha_1}$(°)	$\boldsymbol{\alpha_2}$(°)	$\boldsymbol{\alpha_3}$(°)	$\boldsymbol{\varphi_1}$(°)	$\boldsymbol{\varphi_2}$(°)	$\boldsymbol{\varphi_3}$(°)
Guess	1.85	1.21	0.97	1.72	77.00	76.00	112.00	-140.00	1.64	-103.00
Optimized	1.70	1.23	0.96	1.63	82.51	71.45	109.84	-134.48	0.48	-105.25
Error	0.09	0.02	0.01	0.06	0.03	0.03	0.01	0.03	0.01	0.01
Similarity(%)	91.20	98.03	99.33	94.43	96.94	97.47	98.80	96.93	99.35	98.75
Avg similarity(%)	95.75				97.74			98.35		

For each parameter, a normalized deviation was calculated and converted into a similarity value. Bond deviations were computed as $e_r = \frac{|r_g - r_o|}{r_o}$, where r_g and r_o are the guessed and optimized bond lengths, respectively.
For bond angles, the deviation was normalized to the maximum possible difference, 180°, ensuring a consistent scale for all angular values: $e_\alpha = \frac{|\alpha_g - \alpha_o|}{180°}$, where α_g and α_o are the guessed and optimized bond angles, respectively.
Because dihedral angles are periodic, the difference, $\Delta\varphi = \varphi_g - \varphi_o$, was first wrapped to the range $[-180°, 180°]$, after which the deviation was computed as $e_\varphi = \frac{|\Delta\varphi|}{180°}$.
Each normalized deviation $e_i$was converted to a per-parameter similarity $s_i = 1 - e_i$ and the overall geometric similarity was obtained as the means of all $s_i$values, expressed as a percentage.
To better understand the algorithm's performance, the results were also analyzed separately for bonds, bond angles, and dihedral angles.

Table S7: Within-class similarity statistics for the study jobs cases: the table summarizes the molecular similarity metrics for class, including the mean and median Tanimoto similarity values, the 10th (p10) and 90th (p90) percentiles, and the calculated diversity index. The Tanimoto similarity and diversity index are defined in Eqs. S1 and S2, respectively. The percentile values provide information on the distribution of similarities within each class, where p10 indicates the lower bound of similarity (10th percentile) and p90 represents the upper bound (90^{th} percentile).

Class	n	Average T	Median T	p10_T	p90_T	Diversity index
acyl halide	10	0.29	0.24	0.15	0.46	0.71
amide	10	0.16	0.13	0.04	0.36	0.84
ester	10	0.17	0.17	0.05	0.27	0.83
ether	10	0.16	0.13	0.00	0.45	0.84
nitrile	10	0.24	0.22	0.12	0.38	0.76

$$\boldsymbol{Eq\ S1}.\quad T(A,B) = \frac{c}{a+b-c}$$

$$\boldsymbol{Eq\ S2}.\quad Diversity\ index = 1 - \prec T \succ$$

Table S8: Between-classes similarity statistics for the study jobs cases: the table summarizes the pairwise similarity distributions between classes. The table reports the mean, median, 10th (p10), and 90th (p90) percentiles of the Tanimoto similarity values for each pair of classes.

Class_A	Class_B	**Average T**	**Median T**	**p10_T**	**p90_T**
acyl halide	amide	0.15	0.13	0.04	0.27
acyl halide	ester	0.15	0.13	0.05	0.27
acyl halide	ether	0.05	0.04	0.00	0.14
acyl halide	nitrile	0.05	0.04	0.00	0.14
amide	ester	0.15	0.13	0.04	0.27
amide	ether	0.05	0.05	0.00	0.11
amide	nitrile	0.04	0.03	0.00	0.12
ester	ether	0.11	0.08	0.00	0.23
ester	nitrile	0.04	0.04	0.00	0.13
ether	nitrile	0.06	0.04	0.00	0.16

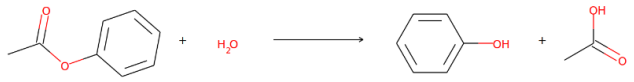
Table S9: Within-class similarity statistics for the validation jobs cases: the table summarizes the molecular similarity metrics for each class, including the mean and median Tanimoto similarity values, the 10th (p10) and 90th (p90) percentiles, and the calculated diversity index. The Tanimoto similarity and diversity index are defined in Eqs. S1 and S2, respectively. The percentile values provide information on the distribution of similarities within each class, where p10 indicates the lower bound of similarity (10th percentile) and p90 represents the upper bound (90th percentile).

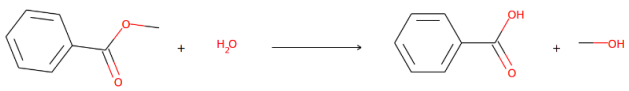
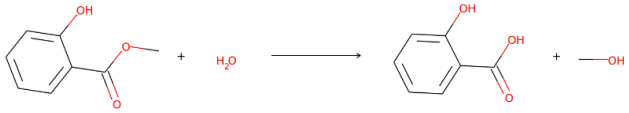
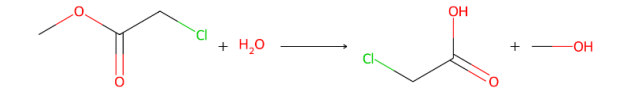
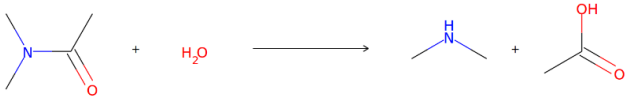
Class	**n**	**Average T**	**Median T**	**p10_T**	**p90_T**	**Diversity index**
acyl halide	12	0.30	0.27	0.14	0.46	0.70
amide	10	0.25	0.26	0.08	0.37	0.75
ester	11	0.24	0.24	0.09	0.41	0.76
ether	30	0.19	0.14	0.07	0.35	0.81
nitrile	30	0.21	0.18	0.12	0.36	0.79

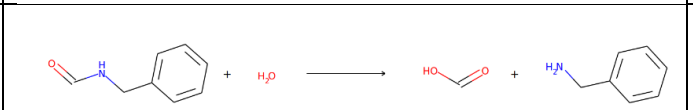
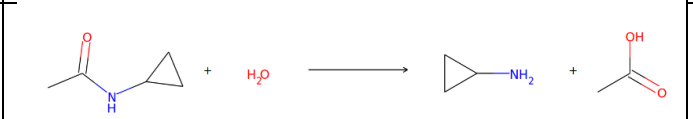
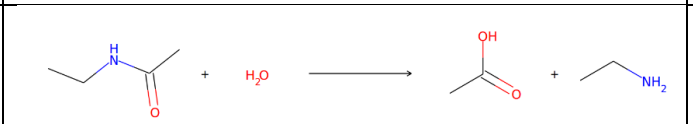
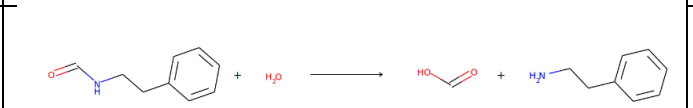
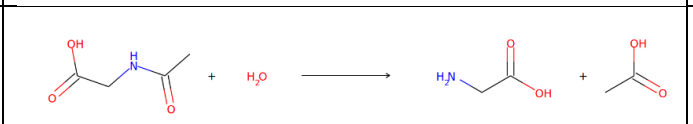
Table S10: Between-classes similarity statistics for the validation jobs cases: the table summarizes the pairwise similarity distributions between classes. The table reports the mean, median, 10th (p10), and 90th (p90) percentiles of the Tanimoto similarity values for each pair of classes.

Class_A	Class_B	**Average T**	**Median T**	**p10_T**	**p90_T**
acyl halide	amide	0.20	0.19	0.10	0.30
acyl halide	ester	0.21	0.19	0.09	0.36
acyl halide	ether	0.07	0.05	0.00	0.17
acyl halide	nitrile	0.05	0.04	0.00	0.13
amide	ester	0.19	0.20	0.07	0.32
amide	ether	0.07	0.05	0.03	0.16
amide	nitrile	0.06	0.04	0.00	0.13
ester	ether	0.14	0.11	0.06	0.27
ester	nitrile	0.07	0.05	0.00	0.15
ether	nitrile	0.07	0.05	0.00	0.18

Table S11: Carbonyl based family validation jobs summary: all calculations were performed at the ωB97X-D/jul-cc-pVTZ level of theory using the SMD (water) solvation model for geometry optimization, frequency analysis, constrained scans, and IRC calculations, with single-point energies computed at the same level.
Each entry reports the reaction scheme, TS convergence status, E_0 validation, confirmation of a single imaginary frequency, normal-mode displacement (NMD) inspection, IRC connectivity, generated TSGs number, and overall success.
The overall success indicates whether the validation job was ultimately considered successful. Jobs marked "X" under IRC but still classified as successful were accepted because NMD inspection clearly confirmed the correct bond-forming and bond-breaking motion for the targeted hydrolysis step.

Job name	Reaction	Class	TS conversion	E_0 validation	Imaginary frequency	NMD	IRC	Generated TSGs number	Overall success
H164		ester	V	V	V	V	V	4	V
H165		ester	V	V	V	V	V	6	V
H166		ester	V	V	V	V	V	4	V
H167		ester	V	V	V	V	V	4	V
H168		ester	V	V	V	V	V	6	V

H169		ester	V	V	V	V	V	6	V
H170		ester	V	V	V	V	X	6	V
H272		ester	V	V	V	V	V	4	V
H187		ester	V	V	V	V	V	4	V
H188		amide	V	V	V	V	V	10	V
H189		amide	V	V	V	V	V	6	V
H171		amide	V	V	V	V	V	8	V
H172		amide	V	V	V	V	V	4	V

H173		amide	V	V	V	V	V	8	V
H174		amide	V	V	V	V	X	6	V
H175		amide	V	V	V	V	X	4	V
H176		amide	V	V	V	V	V	4	V
H177		amide	V	V	V	V	V	2	V
H190		amide	V	V	V	V	V	6	V
H191		amide	V	V	V	V	X	6	V
H178		amide	V	V	V	V	V	6	V

H179		acyl halide	V	V	V	V	V	7	V
H180		acyl halide	V	V	V	V	V	6	V
H181		acyl halide	V	V	V	V	V	6	V
H182		acyl halide	V	V	V	V	V	8	V
H183		acyl halide	V	V	V	V	X	6	V
H184		acyl halide	X	-	-	-	-	0	X
H185		acyl halide	V	V	V	V	V	8	V
H186		acyl halide	V	V	V	V	V	6	V

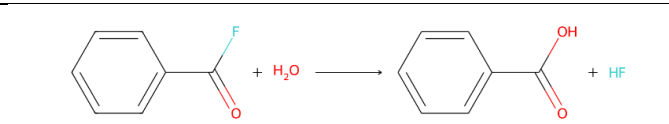
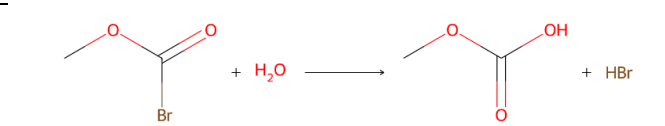
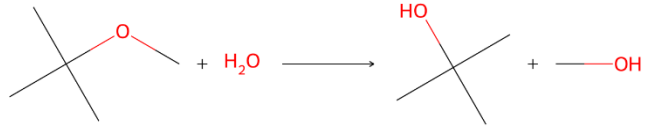
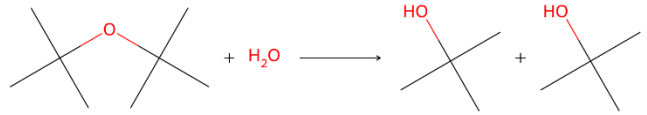
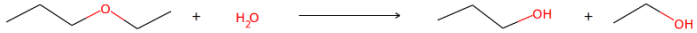
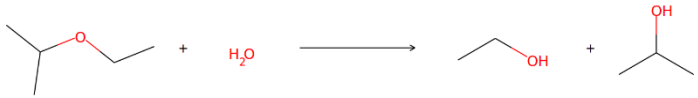
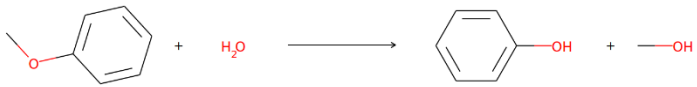
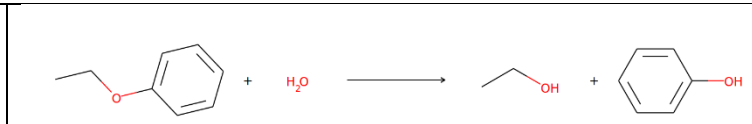
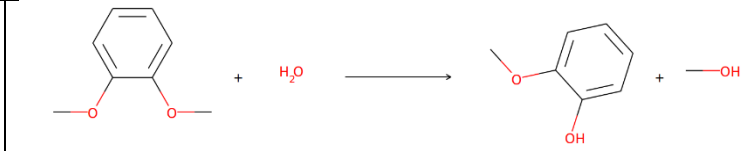
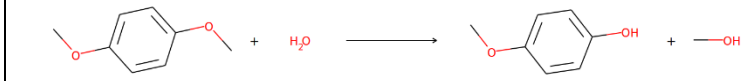
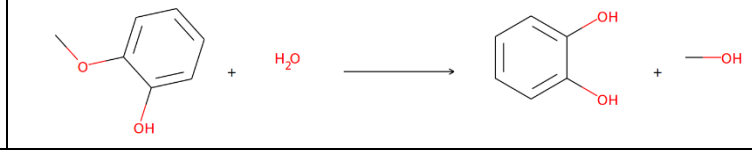
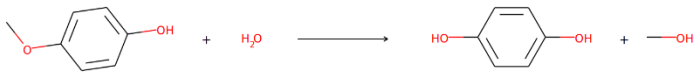
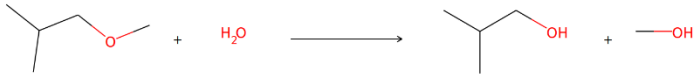
H192		acyl halide	V	V	V	V	V	8	V
H273		acyl halide	V	V	V	V	V	6	V
H274		acyl halide	V	V	V	V	V	6	V
H275		acyl halide	V	V	V	V	V	6	V

Table S12: Ether hydrolysis family validation jobs summary : all calculations were performed at the ωB97X-D/jul-cc-pVTZ level of theory using the SMD (water) solvation model for geometry optimization, frequency analysis, constrained scans, and IRC calculations, with single-point energies computed at the same level.
Each entry reports the reaction scheme, TS convergence status, E_0 validation, confirmation of a single imaginary frequency, normal-mode displacement (NMD) inspection, IRC connectivity, generated TSGs number, and overall success.
The overall success indicates whether the validation job was ultimately considered successful. Jobs marked "X" under IRC but still classified as successful were accepted because NMD inspection clearly confirmed the correct bond-forming and bond-breaking motion for the targeted hydrolysis step.

Job name	Reaction	TS conversion	E_0 validation	Imaginary frequency	NMD	IRC	Generated TSGs number	Overall success
H193		V	V	V	V	V	4	V
H194		V	V	V	V	V	10	V
H195		V	V	V	V	V	2	V
H196		V	V	V	V	V	24	V

H197		X	-	-	-	-	0	X
H198		X	-	-	-	-	0	X
H199		X	-	-	-	-	12	X
H200		X	-	-	-	-	0	X
H201		V	V	V	V	V	2	V
H202		V	V	V	V	V	1	V
H203		V	V	V	V	V	32	V

H204		V	V	V	V	V	28	V
H205		V	V	V	V	V	16	V
H206		V	V	V	V	V	26	V
H207		V	V	V	V	V	3	V
H208		V	V	V	V	V	32	V
H209		X	-	-	-	-	32	X
H210		V	V	V	V	V	13	V

H211		V	V	V	V	V	23	V
H212		V	V	V	V	V	23	V
H276		V	V	V	V	V	5	V
H277		X	-	-	-	-	0	X
H278		X	-	-	-	-	0	X
H279		V	V	V	V	V	4	V
H280		V	V	V	V	V	2	V

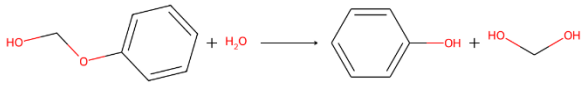
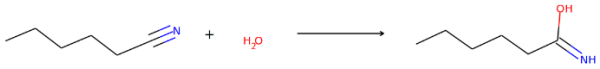
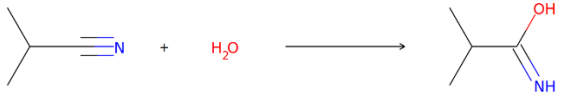
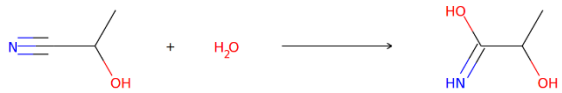
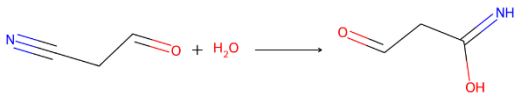
H281		V	V	V	V	V	4	V
H282		V	V	V	V	V	22	V
H283		X	-	-	-	-	0	X
H284		V	V	V	V	V	16	V
H285		V	V	V	V	V	6	V

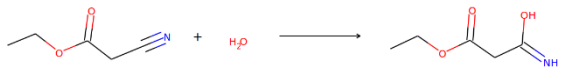
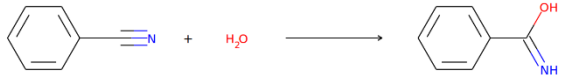
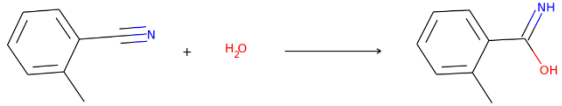
Table S13: Nitrile hydrolysis family validation jobs summary : all calculations were performed at the ωB97X-D/jul-cc-pVTZ level of theory using the SMD (water) solvation model for geometry optimization, frequency analysis, constrained scans, and IRC calculations, with single-point energies computed at the same level.
Each entry reports the reaction scheme, TS convergence status, E_0 validation, confirmation of a single imaginary frequency, normal-mode displacement (NMD) inspection, IRC connectivity, generated TSGs number, and overall success. The overall success indicates whether the validation job was ultimately considered successful.
Jobs marked "X" under IRC but still classified as successful were accepted because NMD inspection clearly confirmed the correct bond-forming and bond-breaking motion for the targeted hydrolysis step.

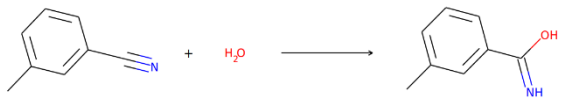
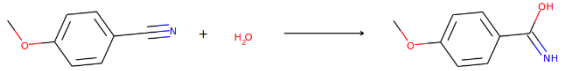
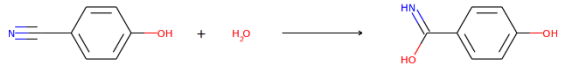
Job name	Reaction	TS conversion	E_0 validation	Imaginary frequency	NMD	IRC	Generated TSGs number	Overall success
H242		V	V	V	V	V	4	V
H243		V	V	V	V	V	4	V

H244		X	-	-	-	-	1	X
H245		V	V	V	V	V	4	V
H246		V	V	V	V	V	4	V
H247		V	V	V	V	V	2	V
H248		X	-	-	-	-	0	X

H249		X	-	-	-	-	0	X
H250		V	V	V	V	V	4	V
H251		V	V	V	V	V	4	V
H252		V	V	V	V	V	4	V
H253		X	-	-	-	-	4	X

H254		V	V	V	V	V	20	V
H255		V	V	V	V	V	4	V
H256		V	V	V	V	V	4	V
H257		V	V	V	V	V	1	V
H258		V	V	V	V	V	4	V

H259		V	V	V	V	V	4	V
H260		V	V	V	V	V	20	V
H261		V	V	V	V	V	2	V
H262		V	V	V	V	V	4	V
H263		V	V	V	V	V	2	V

H264		V	V	V	V	X	2	V
H265		V	V	V	V	V	2	V
H266		V	V	V	V	V	4	V
H267		V	V	V	V	X	4	V
H268		V	V	V	V	V	4	V

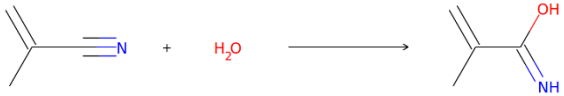
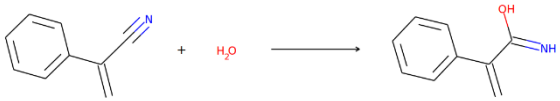
H269		V	V	V	V	V	10	V
H270		V	V	V	V	V	8	V
H271		V	V	V	V	V	4	V

Table S14: Detailed per-reactant results of the ablation study. Each row lists the tested molecule, the outcome of TSG generation and convergence after removing each heuristic component (dihedral modification, d/e atoms electronegativity ranking, and $\pm\varphi_1, \pm\varphi_3$ sign sampling). "Success" indicates that at least one TSG generated by the tool converged during DFT optimization and was verified using frequency and IRC validations. "Failed" indicates either that no TSG was generated by the tool or that none of the generated TSGs converge to a valid TS during optimization.

Class and type of molecule	**Molecule SMILES**	**Dihedral modifications**	**d/e atoms electronegativity ranking**	**$\pm\varphi$ sign sampling**
Ester - simple aliphatic	CC(=O)OC	Failed	Success	Failed

Ester - bulky substituted	CC(=O)OC(C)(C)C	Failed	Success	Failed
Ester - aromatic	O=C(OC)c1ccccc1	Failed	Success	Failed
Amide – polar and bulky(diamide)	CC(=O)NCC(=O)O	Failed	Failed	Failed
Amide - aromatic	CC(=O)Nc1ccccc1	Failed	Success	Failed
Amide - simple	CC(=O)NC	Failed	Success	Failed
Acyl halide - aromatic fluoride	O=C(F)c1ccccc1	Failed	Success	Success
Acyl halide - aliphatic bromide	CC(=O)Br	Failed	Failed	Failed
Acyl halide - aromatic chloride	O=C(Cl)c1ccc(Cl)cc1	Failed	Success	Failed
Ether - branched aliphatic	CC(C)OCC	Success	Success	Failed
Ether - aryl-alkyl	CCOc1ccccc1	Failed	Failed	Failed
Ether - simple dialkyl	CCOCC	Failed	Success	Success
Nitrile - aliphatic	OCCC#N	Success	Success	Failed
Nitrile - conjugated	C=CC#N	Failed	Success	Failed
Nitrile - polar substituted	NC(=O)CC#N	Failed	Success	Failed

Section S2: Structural Merging and Redundancy Screening

To ensure computational efficiency and avoid redundant DFT optimizations, the workflow employs a coordinate-based clustering algorithm. This logic is applied immediately after the heuristic Transition State Guesses (TSGs) are generated but before they are submitted for electronic structure optimization.

1. Pre-optimization Coordinate Comparison

The algorithm iterates through each newly generated heuristic geometry and compares it against the existing pool of TSGs stored within the ARC Species object. Two geometries are considered redundant if they satisfy the following criteria:

1. **Symbol Matching:** The ordered list of atomic symbols must be identical (trivial but still checked).
2. **Coordinate Proximity:** The Cartesian coordinates of all atoms must be nearly identical, determined using a numerical tolerance check. Specifically, the algorithm evaluates each coordinate component (x, y, z) of respective atoms using the following condition:
    - $|\mathbf{coord1 - coord2}| \leq (\mathbf{atol} + \mathbf{rtol} \times |\mathbf{coord2}|)$
    - Where **atol** (absolute tolerance) is **1e-04 Angstrom** and **rtol** (relative tolerance) is **1e-03 Angstrom**.

2. Handling Overlapping Methods

If a new heuristic guess is found to be redundant with an existing guess generated by a different method (e.g., a previous template-based guess), the algorithm does not discard the information. Instead, its "merges" the metadata by appending, e.g., "and Heuristics", to the method description of the existing guess. This ensures that if the subsequent DFT optimization is successful, the contribution of the heuristic approach to finding that specific saddle point is properly recorded along with any other method that yielded this successful TSG.

3. Selection of Unique Guesses

If a generated geometry does not meet the "almost equal" criteria for any existing guess, it is marked as unique. The unique structure is then:

- Initialized as a new TSGuess object in ARC (stored within the ARC Species object for the TS).
- Assigned a unique index within the reaction's transition state pool.
- Saved to a local directory as an .xyz file for auditing and as an initialization point for the downstream DFT solver.

This screening process significantly reduces the number of DFT optimization tasks by filtering out geometries that would likely relax into the same saddle point, especially when multiple dihedral perturbations may lead to nearly identical spatial configurations.

Section S3: Detailed PCA Descriptor Definitions and Methodological Specifics

The physicochemical diversity was quantified using nine standardized RDKit descriptors: molecular weight (MW), topological polar surface area (TPSA), predicted octanol/water partition coefficient (logP), number of hydrogen bond donors and acceptors, number of rotatable bonds, ring count, heavy atom count, and fraction of hybridized carbons. Prior to analysis, all descriptors were standardized to zero mean and unit variance.

The loadings reveal that PC1 is dominated by MW, atom count, and logP, capturing size and hydrophobicity, while PC2 is driven by TPSA and hydrogen-bonding capacity versus fraction of sp^3 (planarity/aromaticity), consistent with prior chemoinformatics PCA analyses[1].

For the study cases, PC1 and PC2 together, explain 61.4% of the variance, providing a reliable 2D visualization of the physicochemical diversity of the dataset[2].

Section S4: Ablation Study of Geometric Heuristics

To evaluate the role of individual geometric heuristics in the transition-state guess (TSG) generation process, three controlled modifications were introduced to the algorithm and tested on three representative reactions from each hydrolysis class. These reactions covered simple aliphatic, substituted, and aromatic derivatives to capture the structural and electronic variability within each class. Although esters, amides, and acyl halides belong to the same carbonyl-based hydrolysis family, they were analyzed separately to determine whether the same geometric perturbations would have distinct effects across these classes.

In the first knock-out modification, the neighboring atom selection step was simplified by removing the effective electronegativity (χ_i^{eff}) ranking used to define atoms d and e around the electrophilic center, a. In the original algorithm, e corresponds to the most electronegative substituent and d to the second most electronegative neighbor, which together determine the orientation of the local reactive plane. Without this ranking, the reactive plane orientation was determined by the (random) order of atom connectivity. This modification was expected to affect asymmetric ethers more strongly than linear nitriles, with a moderate impact on carbonyl-based systems where multiple substituents compete for labeling.

In the second knock-out modification, the $\pm\varphi$ sign-sampling procedure was disabled. In the full algorithm, both positive and negative values of the dihedrals φ_1 and φ_3 are tested to allow water attack from either side of the reactive plane. In the ablation variant, this explicit exploration was removed, and only a single, family-averaged orientation was retained. The circular mean, μ, for each dihedral was computed as:

$$\mu = atan2\left(\frac{1}{n}\sum_{i=1}^{n}\sin\theta_i, \frac{1}{n}\sum_{i=1}^{n}\cos\theta_i\right)$$

where θ_i are the parameterized dihedral values.
The resulting sign corresponded to the dominant attack direction observed in each family, ensuring a physically meaningful orientation while removing explicit sampling of the opposite side. This simplification was expected to sharply reduce success unless the retained orientation coincided with the optimal direction by chance.

In the third knock-out modification, all dihedral adjustments applied before placing the water molecule were removed. In the full algorithm, near-linear torsions around the reaction center are perturbed to prevent unfavourable alignments and open the reactive site. This correction is especially important for carbonyl-based reactions, which are otherwise planar, and for ethers, where substituents can restrict access. In contrast, nitriles are nearly linear and typically do not require this correction. The loss of dihedral modification was therefore expected to have the strongest effect on carbonyl-based systems, a smaller effect on ethers, and little to no impact on nitriles.

A full version of the results and per-reaction outcomes of the ablation study are reported in Table S14 (Section S1). The discussion below synthesizes these results to assess the functional role of each geometric heuristic.

Consistent with expectations, removal of the dihedral-modification step led to complete failure (0/3 success) across all carbonyl-based systems (esters, amides, and acyl halides), with none of the generated structures converging to valid TSs. Only two non-carbonyl cases, a branched ether (1/3) and an aliphatic nitrile (1/3), succeeded without dihedral modification, indicating that flexible or linear geometries can occasionally compensate for the lack of torsional adjustment.
Removal of the $\pm\varphi$ sign-sampling procedure caused widespread failure across all hydrolysis families, with only two successes out of the fifteen total cases tested. All ester, amide, and nitrile reactions failed to converge, while only one acyl halide and one ether succeeded. This near-total collapse demonstrates that exploring both attack directions is critical for correctly positioning the water molecule relative to the reactive plane.
In contrast, removal of the d/e electronegativity ranking produced a smaller and more variable effect. All ester reactions (3/3), two amides (2/3), and two acyl halides (2/3) still converged successfully. Although these results are not statistically significant, they do not contradict the expected sensitivity and instead indicate that, for these specific molecules, random plane assignment did not strongly distort water placement. Nitriles were unaffected (3/3), while ethers exhibited intermediate sensitivity (1/3), consistent with their more complex local connectivity.

Overall, the ablation study confirms that the geometric heuristics act cooperatively rather than independently. The strongest deterioration occurred when either the dihedral-modification step or the $\pm\varphi$ sign-sampling procedure was removed, as both directly control the spatial alignment of the attacking water molecule. The d/e ranking plays a secondary role, primarily fine-tuning local orientation and improving robustness across asymmetric environments.

Section S5: Datasets

- **Datasets S1-S4** containing molecular fingerprints, descriptor matrices, and PCA scores are available on Zenodo under DOI: 10.5281/zenodo.17524146
  The archive includes a README file describing all columns, units, and dataset contents. [3-6].
- **Dataset S5** containing quantum chemical validation output files (TS optimization, frequency, and IRC calculations, where available) for the neutral hydrolysis reactions is available on Zenodo under DOI: 10.5281/zenodo.18458225. The archive includes a README file describing the dataset organization and file contents.

Dataset S6. Example ARC file input for validation job (carbonyl-based hydrolysis family, specifically ester class):

This example input file is designed to be run directly with the Automated Rate Calculator (ARC) tool. To reproduce the results presented in this work, install **ARC** Version: 1.1.0 or later (available at https://github.com/ReactionMechanismGenerator/ARC).
The relevant commit that introduced the hydrolysis TS search features is *43d59cfc513e80d07af5c7cc5d49fe007b16776d* from January 5 2026.

```
project: H272

opt_level:
  method: wb97xd
  basis: jul-cc-pVTZ
  solvation_method: smd
  solvent: water

freq_level:
  method: wb97xd
  basis: jul-cc-pVTZ
  solvation_method: smd
  solvent: water

scan_level:
  method: wb97xd
  basis: jul-cc-pVTZ
  solvation_method: smd
  solvent: water

irc_level:
  method: wb97xd
  basis: jul-cc-pVTZ
  solvation_method: smd
  solvent: water
```

```
sp_level:
  method: wb97xd
  basis: jul-cc-pVTZ
  solvation_method: smd
  solvent: water


ts_adapters: ['heuristics']

compute_thermo: false

species:
- label: reactant
  smiles: ClCC(=O)OC

- label: H2O
  smiles: O

- label: ClCC(=O)O
  smiles: ClCC(=O)O

- label: CO
  smiles: CO

job_types:
  rotors: false
  sp: true
  opt: true
  fine: true
  freq: true
  irc: true

reactions:
- label: reactant + H2O <=> ClCC(=O)O + CO
```

Section S6: References

[1] Todeschini, R.; Consonni, V.; Xiang, H.; Holliday, J.; Buscema, M.; Willett, P. Similarity coefficients for binary chemoinformatics data: overview and extended comparison using simulated and real data sets. Journal of Chemical Information and Modeling 2012, 52, 2884–2901.

[2] Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? Drug Discovery Today 2002, 7, 903–911

[3] Muniba Faiza. (2023). *Tanimoto Similarities* [Computer software]. GitHub. https://github.com/MunibaFaiza/tanimoto_similarities

[4] Bender, Andreas, and Robert C. Glen. "Molecular similarity: a key technique in molecular informatics." *Organic & biomolecular chemistry* 2.22 (2004): 3204-3218..

[5] Gewers, Felipe L., et al. "Principal component analysis: A natural approach to data exploration." *ACM Computing Surveys (CSUR)* 54.4 (2021): 1-34.

[6] Exploring the chemical space by Principal Component Analysis (PCA) and clustering.” *Chem-Workflows.com*, accessed 2025. https://chem-workflows.com/content/PCA_compounds.html