Supporting Information: Impact of the characteristics of quantum chemical databases on machine learning predictions of tautomerization energies.

Luis Itza Vazquez-Salazar,[†] Eric Boittier,[†] Oliver T. Unke,[‡,¶] and Markus Meuwly*[,§,‖]

†*Department of Chemistry, University of Basel, Klingelbergstrasse 80 , CH-4056 Basel, Switzerland*

‡*Machine Learning Group, Technische Universität Berlin, 10587 Berlin, Germany*

¶*DFG Cluster of Excellence "Unifying Systems in Catalysis" (UniSysCat), Technische Universität Berlin, 10623 Berlin, Germany*

§*Department of Chemistry, University of Basel, Klingelbergstrasse 80 , CH-4056 Basel, Switzerland.*

‖*Department of Chemistry, Brown University, Providence RI, USA*

E-mail: m.meuwly@unibas.ch

June 2, 2021

Supplementary Methods

Generation of the geometries for the molecules of the SAMPL2 challenge

The molecules of the SAMPL2 challenge[1] were used to generate a third database for the evaluation of the energies by the NN models containing only equilibrium molecules. The initial geometries generation was done using OpenBabel[2] for the force fields GAFF,[3] UFF[4] and Ghemical[5] for 5000 steps from the SMILES representations.

For the Gromos[6] force field, the parametrization of the molecules was done using the ATB server.[7] Meanwhile, the parameters for the CHARMM[8] force field were generated using SwissParam.[9] Finally, the parameters of the OPLS[10] force field were generated using the LigParGen[11] code. Once the parameters from those force fields were generated, the molecules were optimized for 5000 steps using the GROMACS code.

Table S1: Values of the Kullback-Leibler (KL) divergence of the different training sets and the tautobase at specific intervals for different types of bonds for molecules with $n_{\text{atoms}} \leq 9$. The values of the bond lengths were taken from Allen, F.H., *et al.*, 2006[12]

Type of Bond	QM9	PC9	ANI-1E	Interval (Å)
C(sp)-C(sp)(triple)	0.003	0.000	0.005	1.167-1.197
$C(sp^2)$-$C(sp^2)$(double)	-0.136	-0.083	-0.076	1.280-1.405
$C(sp^2)$-$C(sp^2)$(single)	0.254	0.236	0.167	1.400-1.568
$C(sp^3)$-$C(sp^3)$(single)	0.372	0.312	0.224	1.458-1.610
C(sp)-C(sp)(single)	-0.114	-0.125	-0.099	1.374-1.474
C(ar)-C(ar)	-0.131	-0.132	-0.121	1.350-1.440
$C(sp^3)$-$C(sp^2)$	0.136	0.135	0.174	1.470-1.538
$C(sp^3)$-C(ar)	0.151	0.161	0.182	1.479-1.539
$C(sp^3)$-C(sp)	-0.038	-0.048	-0.027	1.436-1.481
$C(sp^2)$-C(ar)	-0.022	-0.065	0.031	1.441-1.512
$C(sp^2)$-C(sp)	-0.015	-0.016	-0.013	1.425-1.441
C(ar)-C(sp)	-0.017	-0.019	-0.015	1.430-1.448
Carbon-Nitrogen bonds				
$C(sp^3)$-N(4)	0.1541	0.0765	0.0205	1.482-1.510
$C(sp^3)$-N(3)	0.5254	0.6454	0.4214	1.446-1.572
$C(sp^3)$-N(2)	0.3340	0.4596	0.2156	1.461-1.506
$C(sp^2)$-N(3)	-0.2290	-0.2421	-0.2186	1.314-1.419
$C(sp^2)$-N(2) (Imidazole)	-0.0432	-0.0488	-0.0402	1.369-1.384
C(ar)-N(4)	0.0790	0.2000	0.1227	1.461-1.470
C(ar)-N(3)	0.0290	0.2841	0.2339	1.340-1.476
C(ar)-N(2)	-0.0080	-0.0087	0.0113	1.422-1.442
$C(sp^2)$-N(3) (furoxan)	-0.0202	-0.0213	-0.0216	1.311-1.324

Continued on next page..

Table S1: Values of the Kullback-Leibler (KL) divergence of the different training sets and the tautobase at specific intervals for different types of bonds for molecules with $n_{\text{atoms}} \leq 9$. The values of the bond lengths were taken from Allen, F.H., *et al.*, 2006[12](cont.)

Type of Bond	QM9	PC9	ANI-1E	Interval (Å)
C(sp^2)-N(2)	-0.1046	-0.0753	-0.0409	1.273-1.339
C(ar)-N(3)	-0.0897	-0.0923	-0.0962	1.325-1.369
C(ar)-N(2)	-0.0807	-0.0801	-0.0853	1.300-1.348
C(sp)-N(2)	0.0131	0.0001	0.0024	1.140-1.148
C(sp)-N(1)	-0.1771	-0.2281	-0.0925	1.131-1.449
Carbon-Oxygen bonds				
C(sp^3)-O(2) (Alcohols)	0.7924	0.6993	0.7431	1.395-1.449
C(sp^3)-O(2)(Dialkyl ethers)	0.7733	0.6979	0.6591	1.405-1.458
C(sp^3)-O(2)(aryl alkyl ethers)	0.4180	0.4313	0.3602	1.417-1.438
C(sp^2)-O(2)[1]	0.2570	0.1608	0.0713	1.435-1.501
C(sp^2)-O(2)(Ring systems)	0.3480	0.2537	0.1312	1.430-1.501
C(sp^2)-O(2)(Enols)	-0.0370	-0.0262	-0.0359	1.324-1.342
C(sp^2)-O(2)(enol esters)	-0.0516	-0.0156	-0.0380	1.341-1.363
C(sp^2)-O(2)(acids)	-0.0245	-0.0172	-0.0248	1.279-1.320
C(sp^2)-O(2)(esters)	0.2340	0.1834	0.3607	1.328-1.420
C(sp^2)-O(2)(anhydrides)	0.0068	-0.0116	0.0138	1.379-1.393
C(sp^2)-O(2)(ring systems)	-0.0932	-0.0388	-0.0697	1.332-1.377
C(ar)-O(2)(Phenols)	-0.0390	-0.0100	-0.0245	1.353-1.373
C(ar)-O(2)(aryl alkyl ethers)	-0.0215	-0.0096	-0.0127	1.363-1.377
C(ar)-O(2)(diaryl ethers)	-0.0010	-0.0151	0.0059	1.375-1.391
C(ar)-O(2)(esteres)	0.1042	0.0453	0.1445	1.394-1.408

Continued on next page..

[1]Aryl alkyl ethers, alkyl esters of carboxilic acids, alkyl esters of alpha, beta unsaturated acids, alkyl esterets of benzoic acid

Table S1: Values of the Kullback-Leibler (KL) divergence of the different training sets and the tautobase at specific intervals for different types of bonds for molecules with $n_{\text{atoms}} \leq 9$. The values of the bond lengths were taken from Allen, F.H., *et al.*, 2006[12](cont.)

Type of Bond	QM9	PC9	ANI-1E	Interval (Å)
C(sp^2)-O(1)(double) (Aldehydes and ketones)	-0.1119	-0.1178	-0.1090	1.188-1.238
C(sp^2)-O(1)(double)[2]	-0.0370	-0.0513	-0.0399	1.232-1.262
C(sp^2)-O(1)(double)(carboxylic acids)	-0.0887	-0.0917	-0.0811	1.203-1.241
C(sp^2)-O(1)(double)(esters)	-0.0477	-0.0509	-0.0523	1.181-1.207
C(sp^2)-O(1)(anhydrides)	-0.0198	-0.0164	-0.0199	1.184-1.193
C(sp^2)-O(1)(lactones)	0.0000	0.0000	0.0000	1.187-1.209
C(sp^2)-O(1)(double)(amides)	-0.1086	-0.1185	-0.1058	1.193-1.243
Nitrogen-nitrogen bonds				
N(4)-N(3)	-0.0109	0.0094	0.0150	1.412-1.418
N(3)-N(3)	-0.1170	0.1075	0.2943	1.384-1.457
N(3)-N(2)	0.0426	-0.1165	-0.0721	1.345-1.375
N(2)-N(2)(aromatic)	0.2894	-0.1551	-0.1649	1.287-1.375
N(2)-N(2)	0.0477	0.0710	-0.0100	1.202-1.262
N(2)-N(1) (azides)	0.0000	0.0203	0.0003	1.114-1.137
Nitrogen-Oxygen bonds				
N(3)-O(2)	0.5630	0.1865	0.5170	1.388-1.468
N(3)-O(1)	-0.1038	-0.0670	-0.0907	1.228-1.316
N(2)-O(2)	0.6880	0.0412	0.8593	1.365-1.420
N(3)-O(1)	-0.0948	0.0117	-0.0875	1.203-1.251

[2]Delocalized double bonds in carboylate anions

Table S2: Values of the Kullback-Leibler (KL) divergence of the different training sets and the tautobase at specific intervals for different types of bonds for molecules with $n_{\text{atoms}} > 9$. The values of the bond lengths were taken from Allen, F.H., *et al.*, 2006[12]

Type of Bond	QM9	PC9	ANI-1E	Interval (Å)
C(sp)-C(sp)(triple)	0.003	0.000	0.005	1.167-1.197
$C(sp^2)$-$C(sp^2)$(double)	-0.121	-0.022	0.000	1.280-1.405
$C(sp^2)$-$C(sp^2)$(single)	0.767	0.719	0.657	1.400-1.568
$C(sp^3)$-$C(sp^3)$(single)	0.937	0.853	0.746	1.458-1.610
C(sp)-C(sp)(single)	-0.183	-0.217	-0.158	1.374-1.474
C(ar)-C(ar)	-0.202	-0.208	-0.191	1.350-1.440
$C(sp^3)$-$C(sp^2)$	0.499	0.482	0.582	1.470-1.538
$C(sp^3)$-C(ar)	0.513	0.513	0.584	1.479-1.539
$C(sp^3)$-C(sp)	-0.014	-0.031	0.003	1.436-1.481
$C(sp^2)$-C(ar)	0.146	0.069	0.249	1.441-1.512
$C(sp^2)$-C(sp)	-0.021	-0.024	-0.018	1.425-1.441
C(ar)-C(sp)	-0.019	-0.022	-0.016	1.430-1.448
Carbon-Nitrogen bonds				
$C(sp^3)$-N(4)	0.1634	0.0808	0.0236	1.482-1.510
$C(sp^3)$-N(3)	0.5185	0.6255	0.4057	1.446-1.572
$C(sp^3)$-N(2)	0.3322	0.4451	0.2122	1.461-1.506
$C(sp^2)$-N(3)	-0.2625	-0.2728	-0.2479	1.314-1.419
$C(sp^2)$-N(2) (Imidazole)	-0.0560	-0.0593	-0.0518	1.369-1.384
C(ar)-N(4)	0.0733	0.1890	0.1165	1.461-1.470
C(ar)-N(3)	-0.0104	0.2425	0.1849	1.340-1.476
C(ar)-N(2)	-0.0027	-0.0028	0.0118	1.422-1.442
$C(sp^2)$-N(3) (furoxan)	-0.0226	-0.0240	-0.0221	1.311-1.324

Continued on next page..

Table S2: Values of the Kullback-Leibler (KL) divergence of the different training sets and the tautobase at specific intervals for different types of bonds for molecules with $n_{\text{atoms}} > 9$. The values of the bond lengths were taken from Allen, F.H., *et al.*, 2006[12](cont.)

Type of Bond	QM9	PC9	ANI-1E	Interval (Å)
C(sp^2)-N(2)	-0.0917	-0.0424	-0.0127	1.273-1.339
C(ar)-N(3)	-0.0984	-0.0977	-0.1007	1.325-1.369
C(ar)-N(2)	-0.0855	-0.0850	-0.0863	1.300-1.348
C(sp)-N(2)	0.0136	0.0005	0.0028	1.140-1.148
C(sp)-N(1)	-0.1404	-0.1741	-0.0339	1.131-1.449
Carbon-Oxygen bonds				
C(sp^3)-O(2) (Alcohols)	0.6515	0.5870	0.5996	1.395-1.449
C(sp^3)-O(2)(Dialkyl ethers)	0.6108	0.5655	0.5093	1.405-1.458
C(sp^3)-O(2)(aryl alkyl ethers)	0.3302	0.3556	0.2713	1.417-1.438
C(sp^2)-O(2)[3]	0.1719	0.0874	0.0270	1.435-1.501
C(sp^2)-O(2)(Ring systems)	0.2398	0.1593	0.0673	1.430-1.501
C(sp^2)-O(2)(Enols)	-0.0349	-0.0245	-0.0335	1.324-1.342
C(sp^2)-O(2)(enol esters)	-0.0418	-0.0013	-0.0211	1.341-1.363
C(sp^2)-O(2)(acids)	-0.0261	-0.0196	-0.0271	1.279-1.320
C(sp^2)-O(2)(esters)	0.2461	0.2170	0.3624	1.328-1.420
C(sp^2)-O(2)(anhydrides)	0.0175	-0.0017	0.0245	1.379-1.393
C(sp^2)-O(2)(ring systems)	-0.0718	-0.0090	-0.0364	1.332-1.377
C(ar)-O(2)(Phenols)	-0.0261	0.0091	-0.0040	1.353-1.373
C(ar)-O(2)(aryl alkyl ethers)	-0.0112	0.0045	0.0018	1.363-1.377
C(ar)-O(2)(diaryl ethers)	0.0114	-0.0029	0.0192	1.375-1.391
C(ar)-O(2)(esteres)	0.1064	0.0513	0.1394	1.394-1.408

Continued on next page..

[3]Aryl alkyl ethers, alkyl esters of carboxilic acids, alkyl esters of alpha, beta unsaturated acids, alkyl esterets of benzoic acid

Table S2: Values of the Kullback-Leibler (KL) divergence of the different training sets and the tautobase at specific intervals for different types of bonds for molecules with $n_{\text{atoms}} > 9$. The values of the bond lengths were taken from Allen, F.H., *et al.*, 2006[12](cont.)

Type of Bond	QM9	PC9	ANI-1E	Interval (Å)
C(sp^2)-O(1)(double) Aldehydes and ketones)	-0.1140	-0.1154	-0.1149	1.188-1.238
C(sp^2)-O(1)(double)[4]	-0.0404	-0.0553	-0.0426	1.232-1.262
C(sp^2)-O(1)(double)(Carboxylic acids)	-0.0978	-0.0981	-0.0938	1.203-1.241
C(sp^2)-O(1)(double)(esters)	-0.0373	-0.0375	-0.0425	1.181-1.207
C(sp^2)-O(1)(anhydrides)	-0.0143	-0.0118	-0.0150	1.184-1.193
C(sp^2)-O(1)(lactones)	0.0000	0.0000	0.0000	1.187-1.209
C(sp^2)-O(1)(double)(amides)	-0.1143	-0.1195	-0.1149	1.193-1.243
Nitrogen-Nitrogen bonds				
N(4)-N(3)	-0.0088	0.0221	0.0297	1.412-1.418
N(3)-N(3)	-0.0963	0.2138	0.4141	1.384-1.457
N(3)-N(2)	0.0599	-0.1094	-0.0745	1.345-1.375
N(2)-N(2)(aromatic)	0.2938	-0.1645	-0.1646	1.287-1.375
N(2)-N(2)	0.0053	0.0423	-0.0506	1.202-1.262
N(2)-N(1) (azides)	0.0000	0.0203	0.0003	1.114-1.137
Nitrogen-Oxygen bonds				
N(3)-O(2)	0.8954	0.4577	0.8386	1.388-1.468
N(3)-O(1)	-0.1448	-0.1724	-0.1143	1.228-1.316
N(2)-O(2)	1.0017	0.1725	1.2526	1.365-1.420
N(3)-O(1)	-0.1417	-0.1208	-0.1158	1.203-1.251

[4]Delocalized double bonds in carboylate anions

Table S3: Mean Absolute (MAE) and Root-Mean-Squared Error (RMSE) for the prediction of tautomerization energy ΔE_{Tauto}, and the single isomer energies, E_{SI}, for the entire Tautobase (1257 tautomeric pairs) for each of the datasets evaluated for the MMFF94 force field geometries.

	ΔE_{Tauto}		E_{SI}	
Database	MAE	RMSE	MAE	RMSE
QM9	7.40	10.60	8.36	11.76
PC9	6.59	9.72	9.65	14.95
ANI-1E	6.11	9.31	16.70	16.70
ANI-1	5.57	8.79	20.17	30.35
ANI-1x	3.42	5.79	5.97	8.38

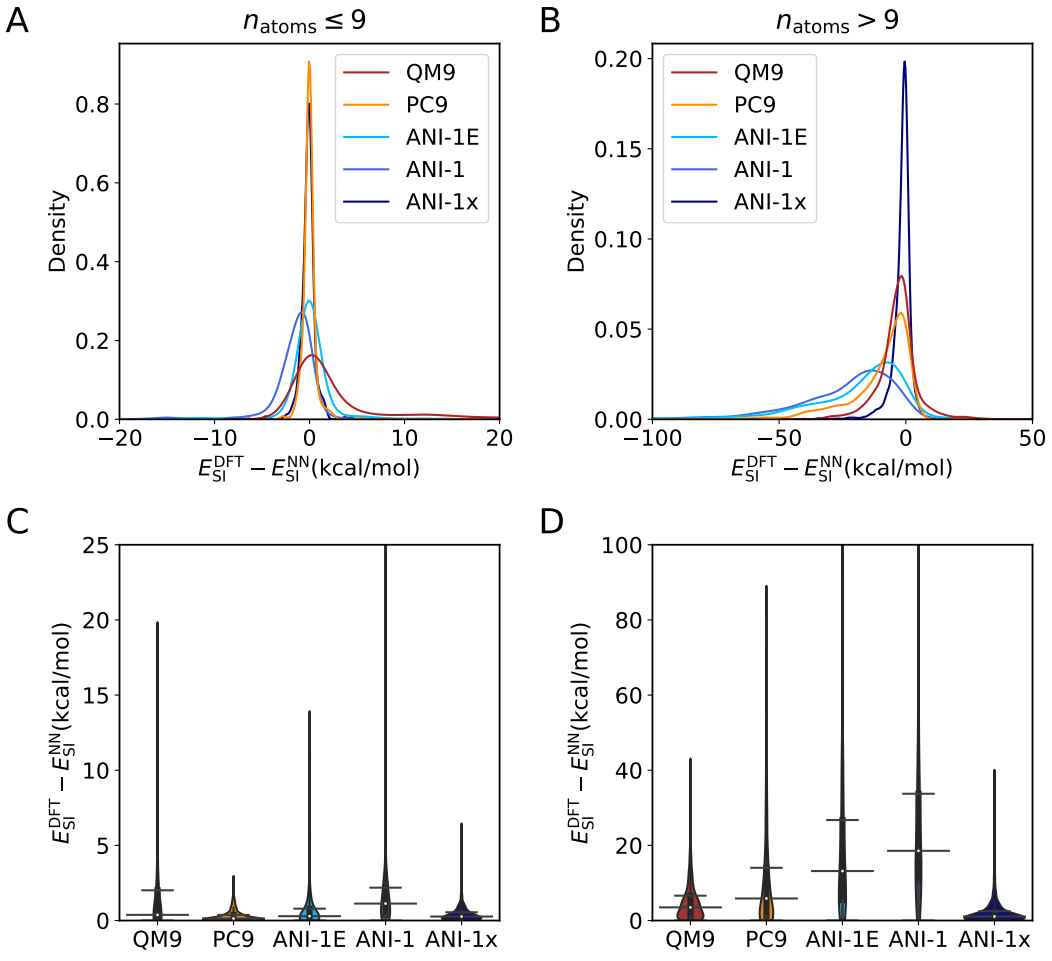


Figure S1: Error analysis for the prediction of E_{SI}. Panels A and B: Kernel density estimate for prediction of E_{SI} for the different databases. Panels C and D: Normalized error distribution up to the 95% quantile of the different datasets. The blackbox inside spans between the 25% and 75% quantiles with a white dot indicating the mean of the distribution. The whisker marks indicate the 5% and 95 % quantiles. The left and right columns are for SetLE9 and SetG9, respectively.

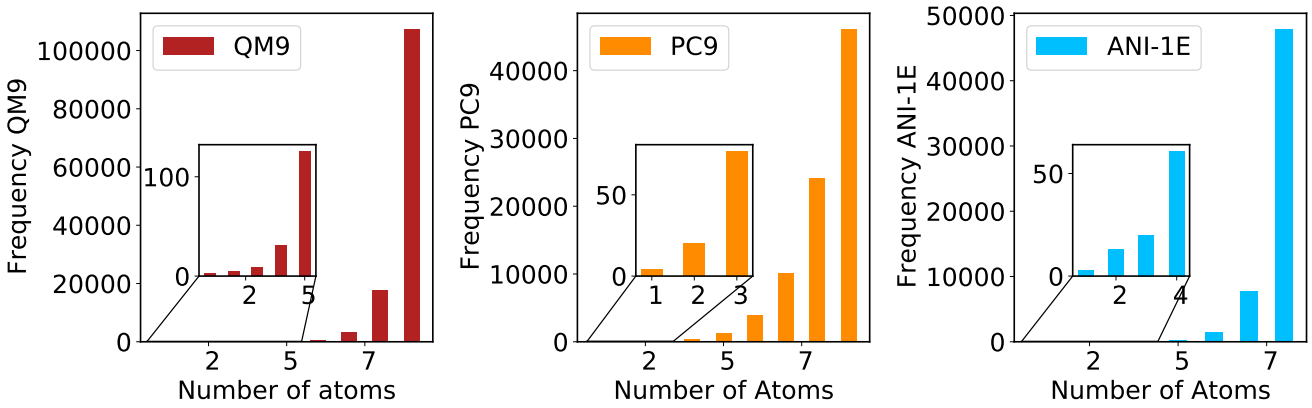


Figure S2: Number of heavy atoms (C,N,O) in the QM9, PC9, and ANI-1E databases (from left to right) used in the present work. The insets show enlargements for cases with few representatives.

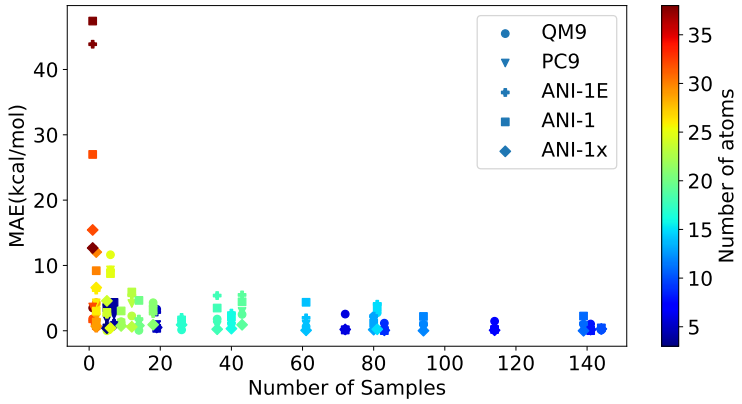


Figure S3: Mean Absolute Error (MAE) by number of samples for the tautomerization energy. The color bar provides the color code for the number of heavy atoms (C, O, N).

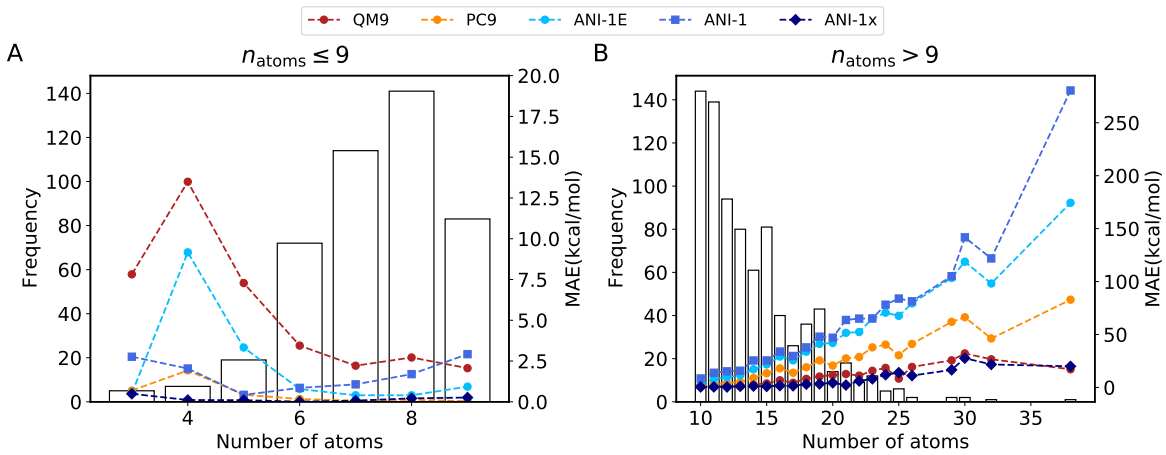


Figure S4: Mean Absolute Error (MAE) by number of heavy atoms (C, O, N) for E_{SI}. A histogram of the number of molecules for different numbers of heavy atoms is shown in the background. Panel A for SetLE9 and panel B for SetG9.

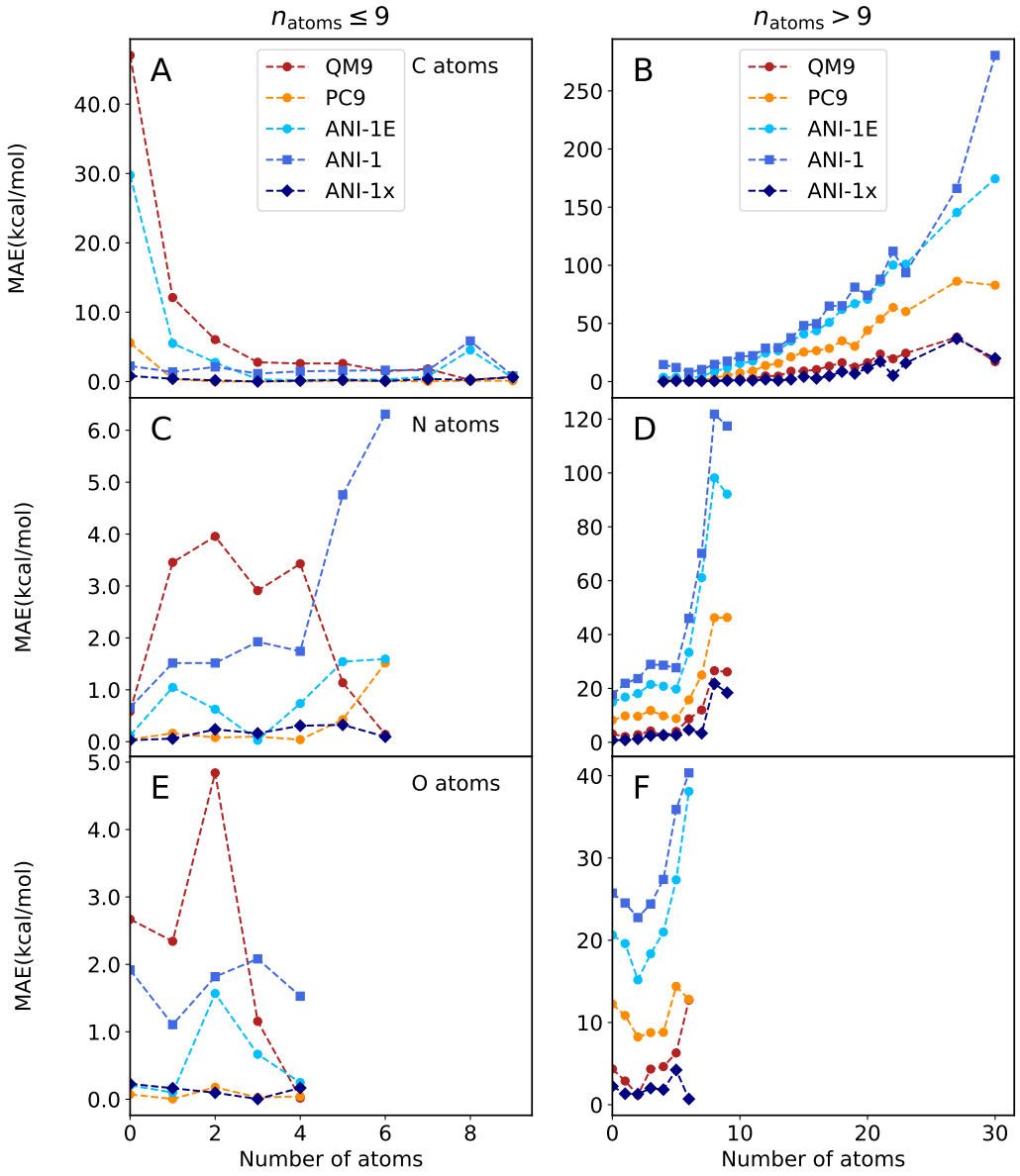


Figure S5: Mean Absolute Error (MAE) by number of atoms of a given element (carbon (top), nitrogen (middle), oxygen (bottom)) for the energy of a single isomer E_{SI}. Panels A and B shows the results by number of C atoms. Panels C and D shows the results by number of N atoms. Finally, panels E and F show the results by number of O atoms. Left and right columns for SetLE9 and SetG9, respectively.

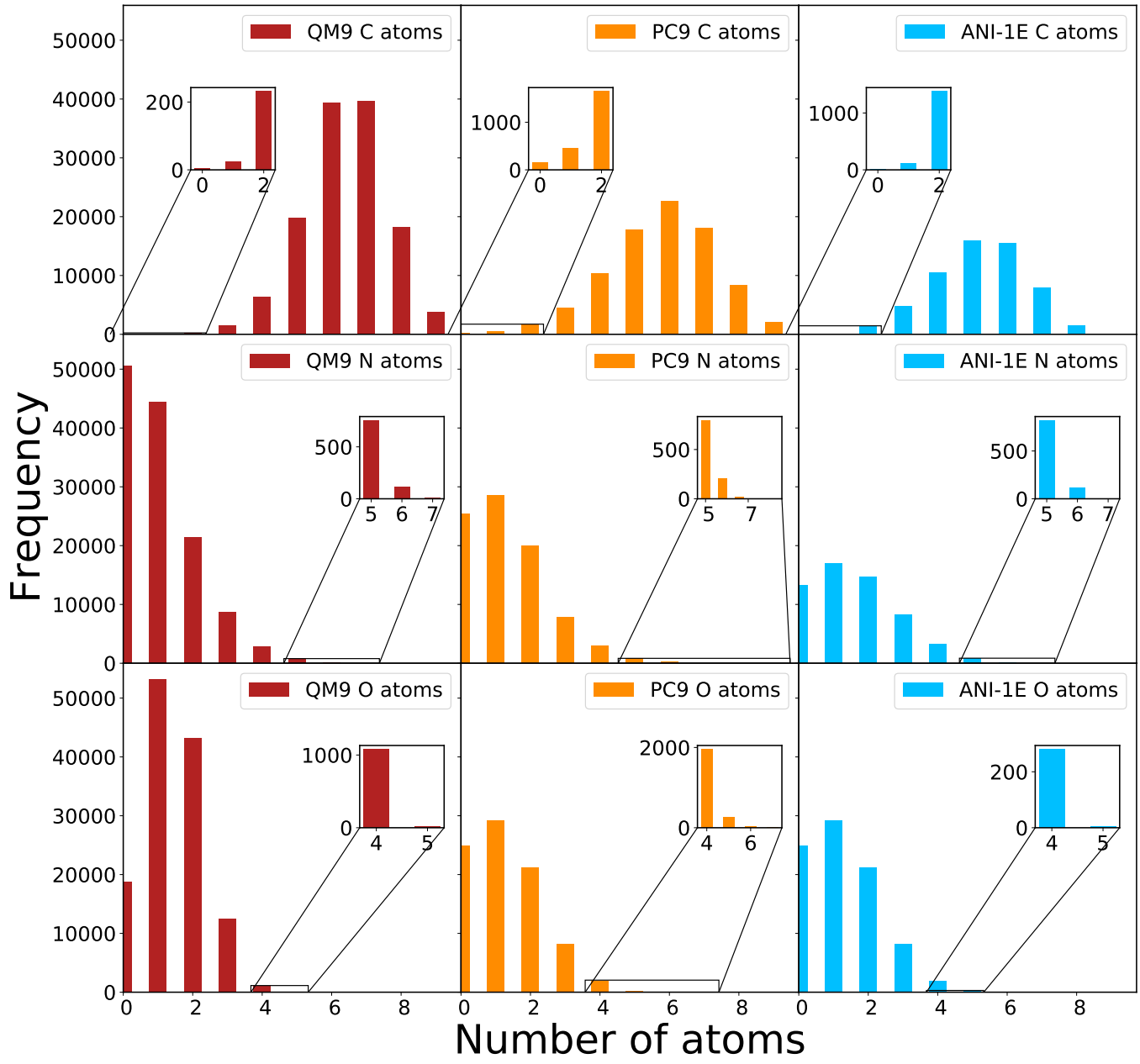


Figure S6: Number of C-, N-, and O-atoms (from top to bottom) in the QM9, PC9, and ANI-1E databases (from left to right) used in the present work. The insets show enlargements for cases with few representatives.

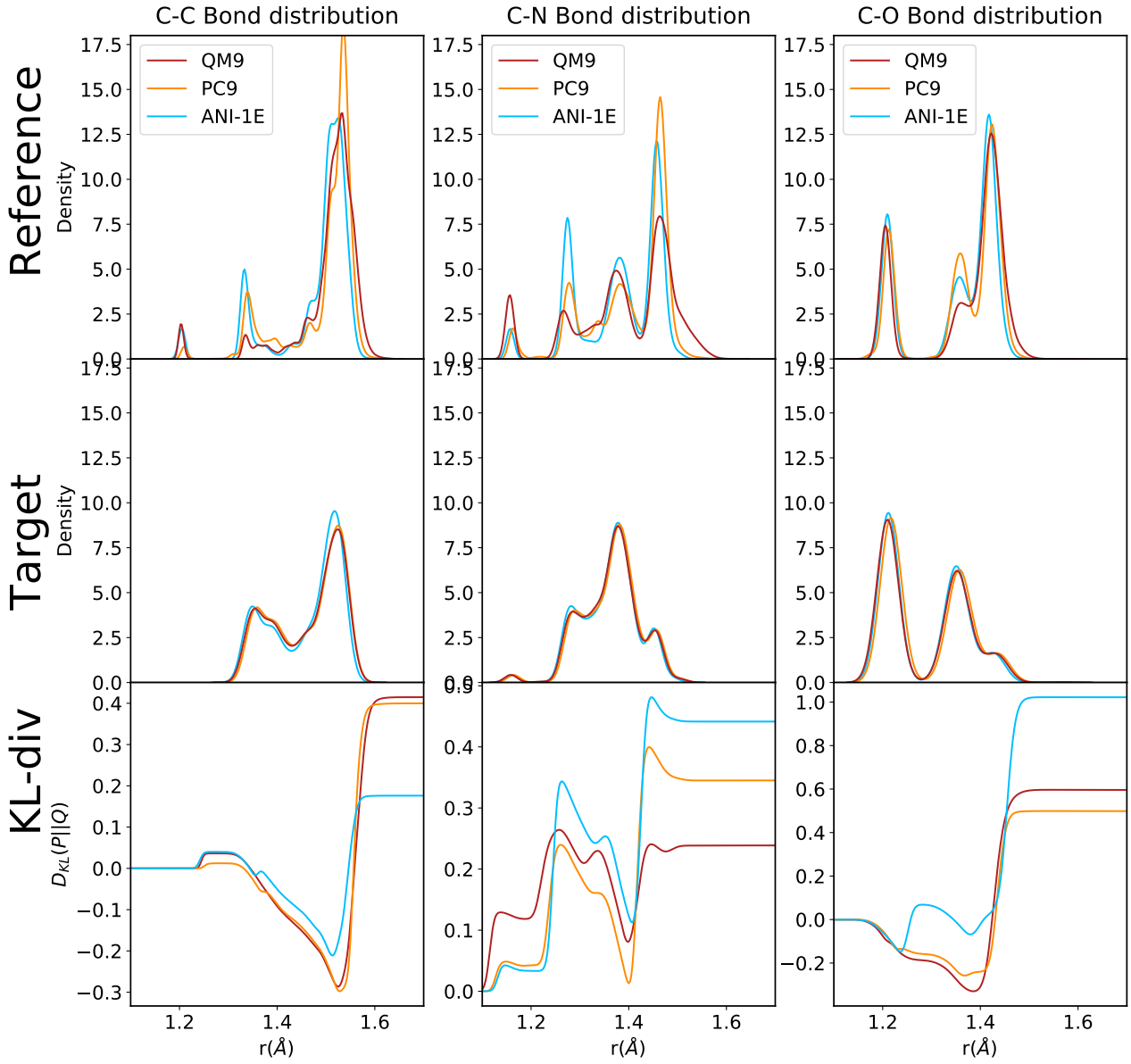


Figure S7: Kernel distributions of different types of bond lengths involving C atoms. Top row are the results for the reference sets (PC9, QM9 and ANI-1E). Middle row shows the results of the geometries of tautobase optimized at level of theory of the different databases for SetLE9. The KL-divergence between reference and target data set distributions is reported in the bottom row.

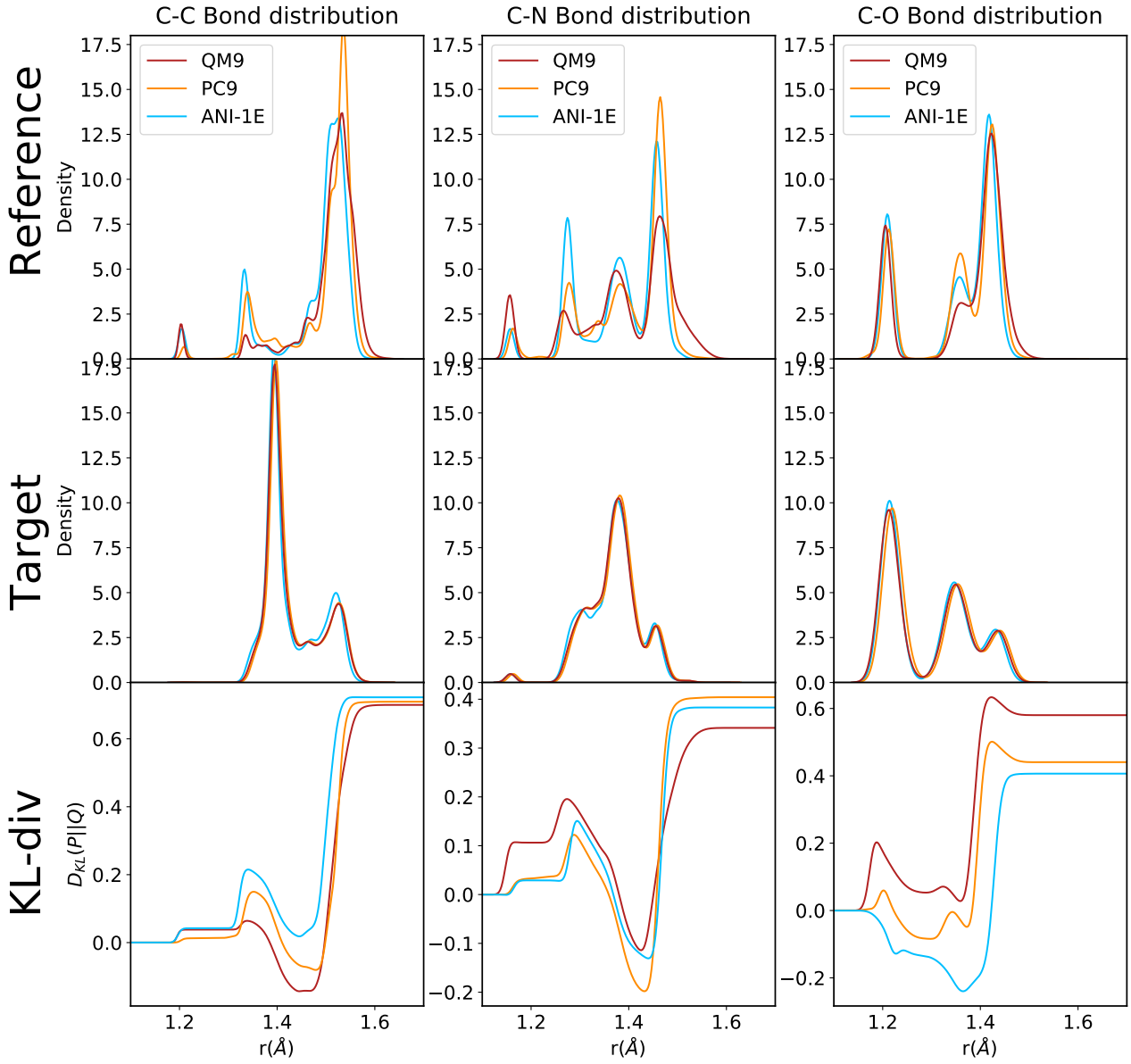


Figure S8: Kernel distributions of different types of bond lengths involving C atoms. Top row are the results for the reference sets (PC9, QM9 and ANI-1E). Middle row shows the results of the geometries of tautobase optimized at level of theory of the different databases for SetG9. The KL-divergence between reference and target data set distributions is reported in the bottom row.

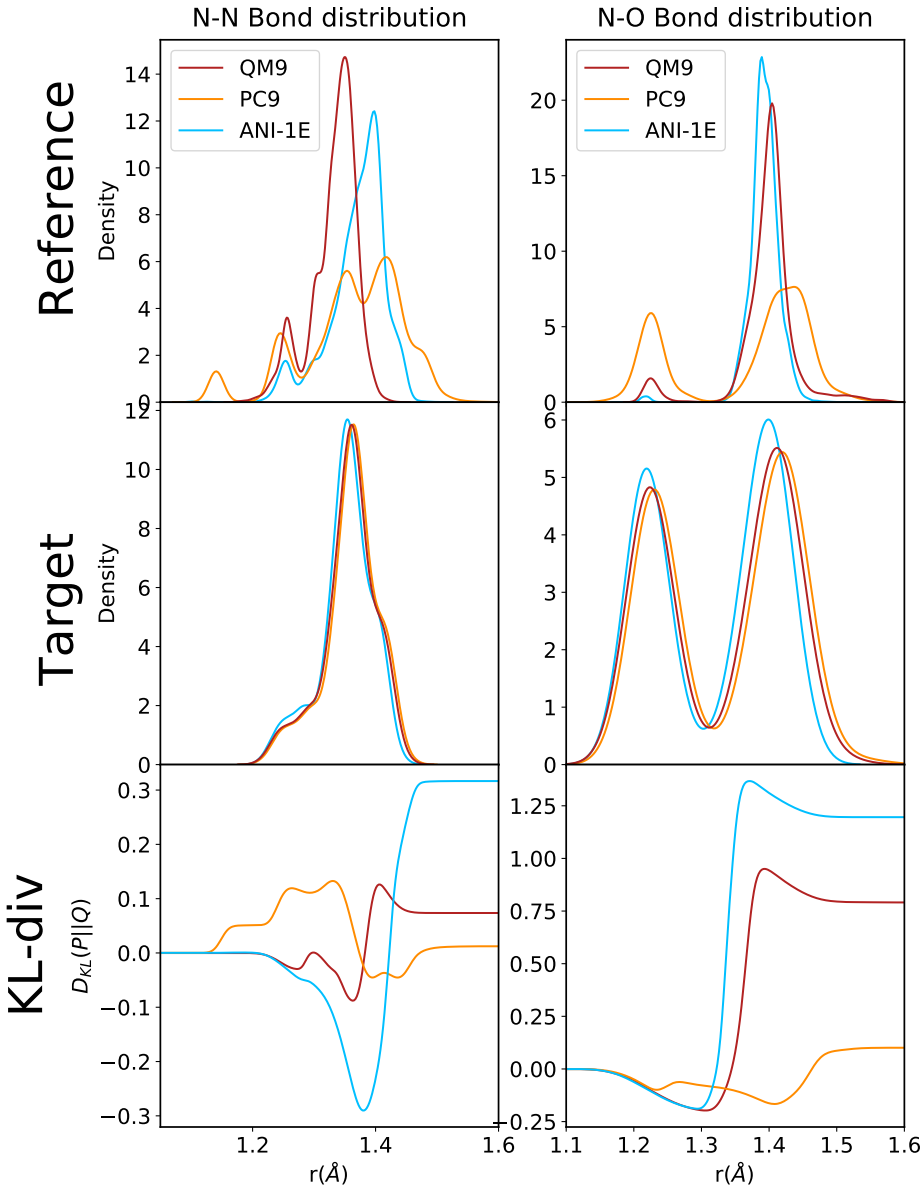


Figure S9: Kernel distributions of different types of bond lengths involving N atoms. Top row are the results for the reference sets (PC9, QM9 and ANI-1E). Middle row shows the results of the geometries of tautobase optimized at level of theory of the different databases for SetLE9. The KL-divergence between reference and target data set distributions is reported in the bottom row..

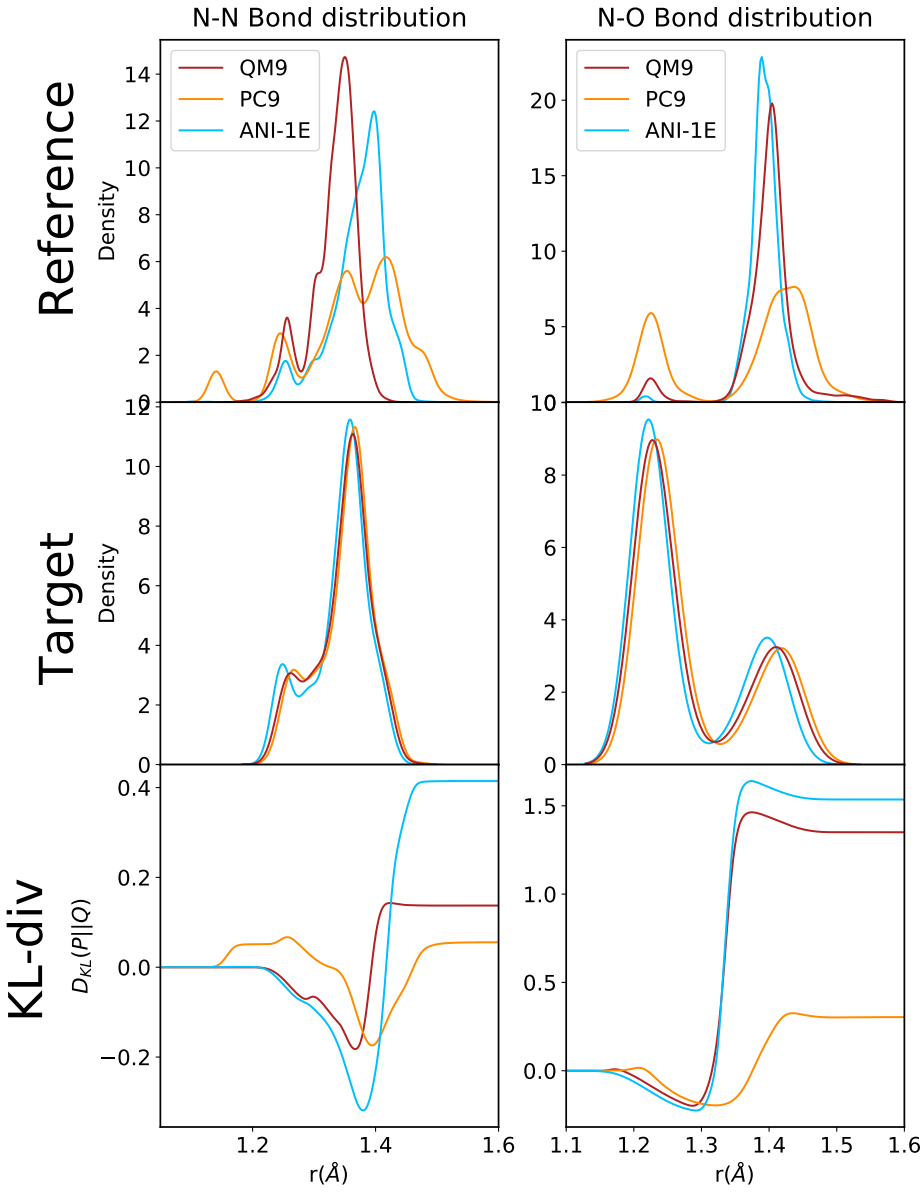


Figure S10: Kernel distributions of different types of bond lengths involving C atoms. Top row are the results for the reference sets (PC9, QM9 and ANI-1E). Middle row shows the results of the geometries of tautobase optimized at level of theory of the different databases for SetG9. The KL-divergence between reference and target data set distributions is reported in the bottom row.

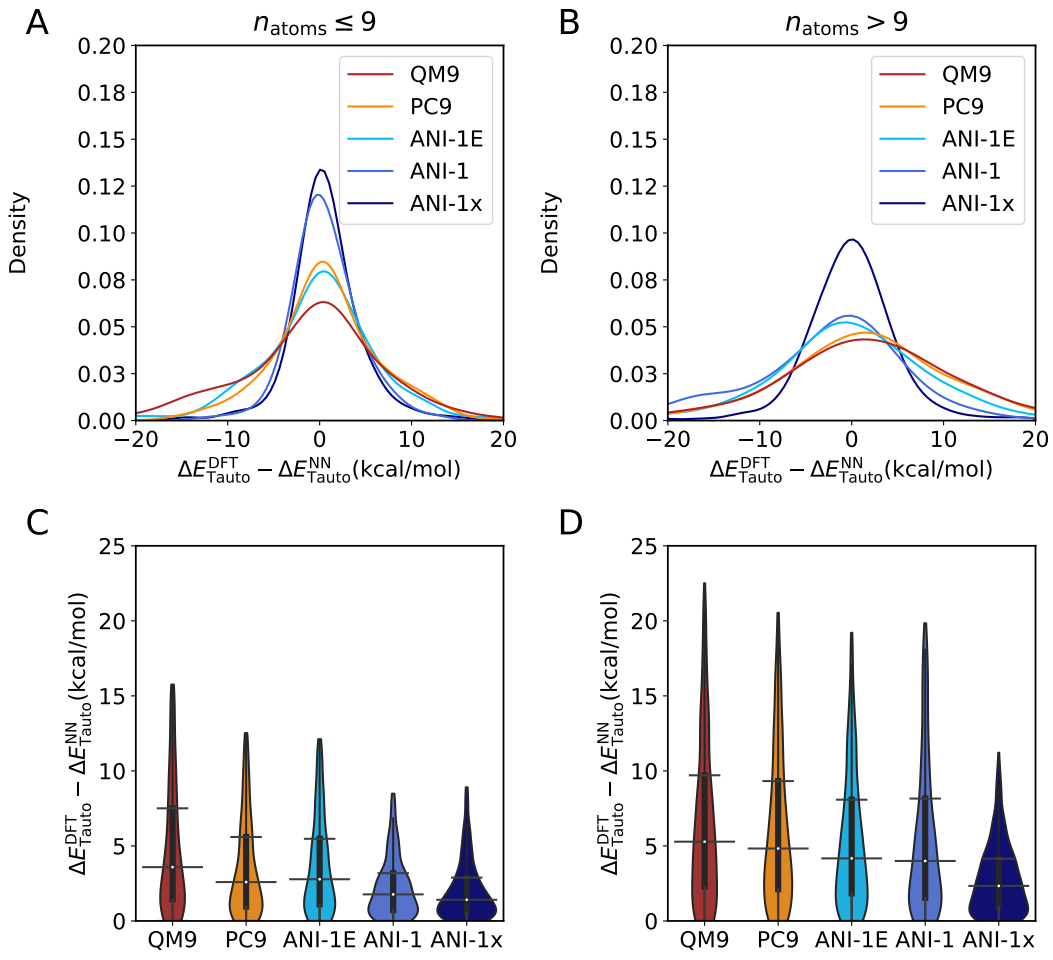


Figure S11: Error analysis on the prediction of the tautomerization energies using an optimized geometry with MMFF94 force field. Panels A and B: Kernel density estimate for prediction of the tautomerization energies for the different databases. Panels C and D: Normalized error distribution up to the 95% quantile of the different datasets evaluated on this work for the tautomerization energy. The blackbox inside spans between the 25% and 75% quantiles with a white dot indicating the mean of the distribution. The whisker marks indicate the 5% and 95 % quantiles. The left and right columns are for SetLE9 and SetG9, respectively.

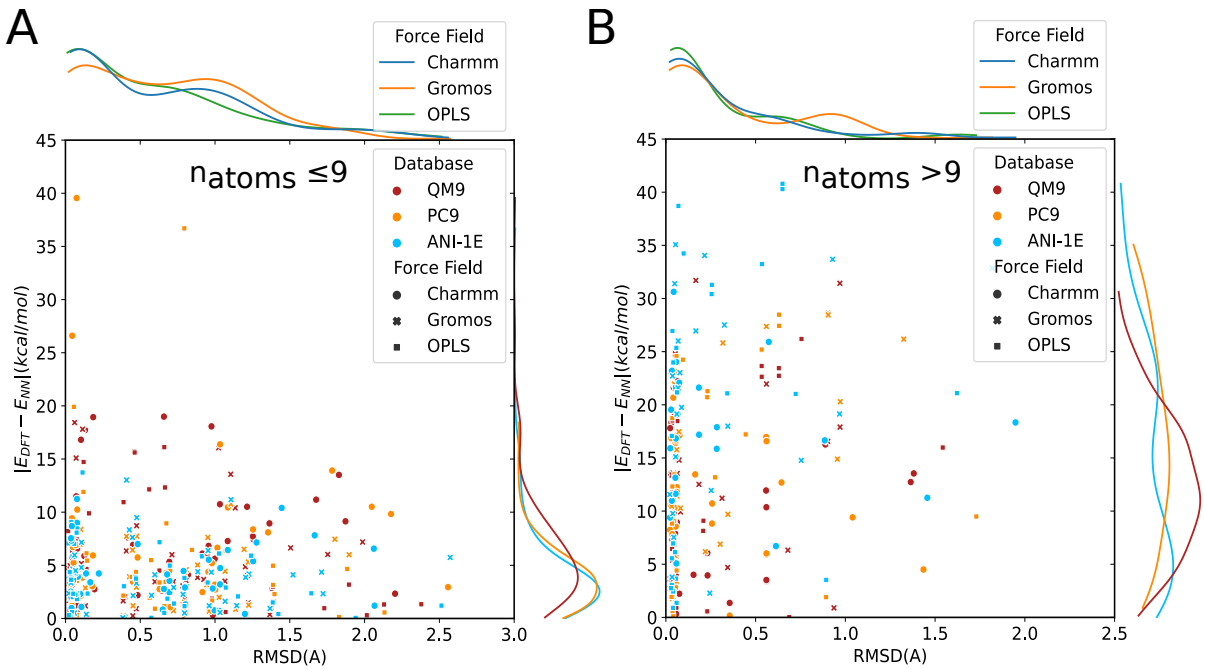
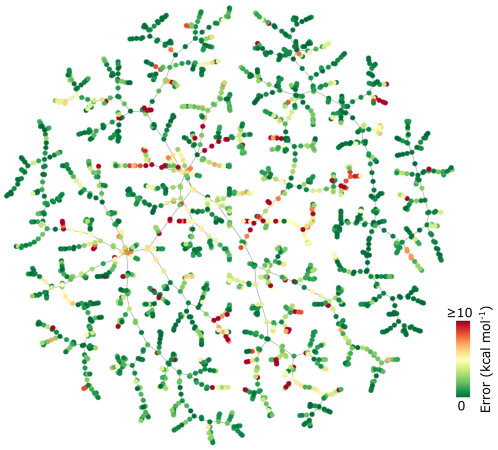
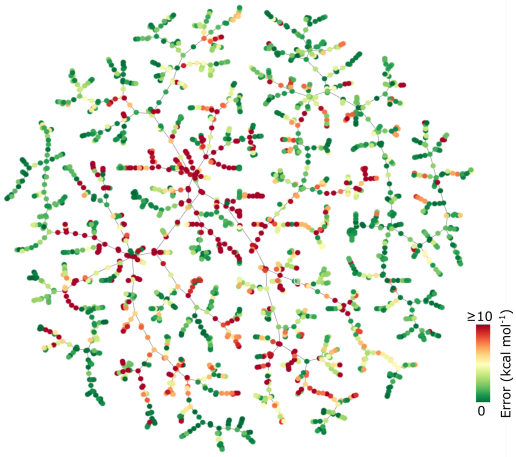


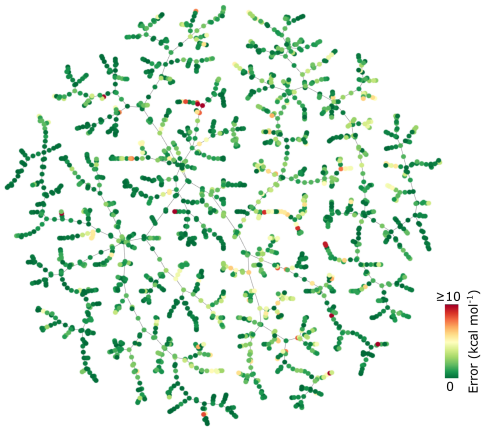
Figure S12: Effect of the initial geometry on the prediction of the energy by the NN models for the molecules from the SAMPL2 challenge.[1] Panel A: Root Mean Square Difference with respect to the optimized geometry vs. the absolute error between the energy obtained from DFT calculations and the energy from the NN model for $n_{\text{atoms}} \leq 9$. Panel B: Results for $n_{\text{atoms}} > 9$. On top of the $X-$axis the kernel distribution for the RMSD for the different force fields is given. Along the $Y-$axis the kernel distributions for the MAE of the energies is given.Results for GAFF, Ghemical and UFF are not show for clarity.



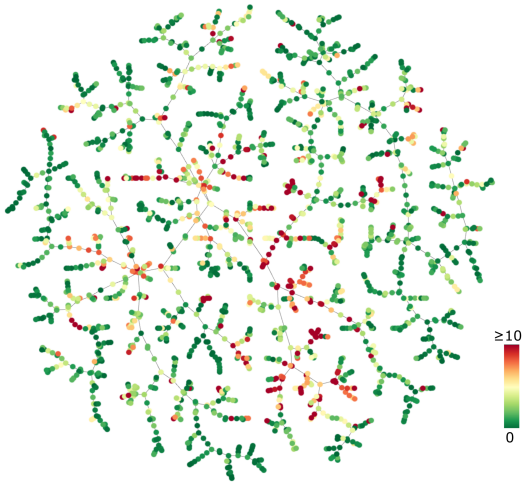
(a) PC9



(b) ANI-1



(c) ANI-1x



(d) ANI-1E

Figure S13: TMAP projection of chemical space for all molecules in the Tautobase, coloured by error in tautomerization energy calculated using the ML potentials trained on (a) PC9, (b) ANI-1, (c) ANI-1x and (d) ANI-1E.

References

(1) Geballe, M. T.; Skillman, A. G.; Nicholls, A.; Guthrie, J. P.; Taylor, P. J. The SAMPL2 blind prediction challenge: introduction and overview. *J. Comput. Aided Mol. Des.* **2010**, *24*, 259–279.

(2) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, 33.

(3) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(4) Rappé, A. K.; Casewit, C. J.; Colwell, K.; Goddard III, W. A.; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024–10035.

(5) Hassinen, T.; Peräkylä, M. New energy terms for reduced protein models implemented in an off-lattice force field. *J. Comput. Chem.* **2001**, *22*, 1229–1242.

(6) Schmid, N.; Eichenberger, A. P.; Choutko, A.; Riniker, S.; Winger, M.; Mark, A. E.; van Gunsteren, W. F. Definition and testing of the GROMOS force-field versions 54A7 and 54B7. *Eur. Biophys. J.* **2011**, *40*, 843–856.

(7) Koziara, K. B.; Stroet, M.; Malde, A. K.; Mark, A. E. Testing and validation of the Automated Topology Builder (ATB) version 2.0: prediction of hydration free enthalpies. *J. Comput. Aided Mol. Des.* **2014**, *28*, 221–233.

(8) Foloppe, N.; MacKerell, A. D., Jr All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data. *J. Comput. Chem.* **2000**, *21*, 86–104.

(9) Zoete, V.; Cuendet, M. A.; Grosdidier, A.; Michielin, O. SwissParam: a fast force field generation tool for small organic molecules. *J. Comput. Chem.* **2011**, *32*, 2359–2368.

(10) Jorgensen, W. L.; Tirado-Rives, J. The OPLS force field for proteins. Energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* **1988**, *110*, 1657–1666.

(11) Dodda, L. S.; Cabeza de Vaca, I.; Tirado-Rives, J.; Jorgensen, W. L. LigParGen web server: an automatic OPLS-AA parameter generator for organic ligands. *Nucleic Acids Res* **2017**, *45*, W331–W336.

(12) Allen, F.; Watson, D.; Brammer, L.; Orpen, A.; Taylor, R. Typical interatomic distances: organic compounds. *International tables for crystallography* **2006**,