nature research

https://doi.org/10.1038/s41557-020-0527-z

Supplementary information

Quantum machine learning using atom-in-molecule-based fragments selected on the fly

In the format provided by the authors and unedited

Supplementary Information for "Quantum machine learning using atom-in-molecule-based fragments selected on-the-fly"

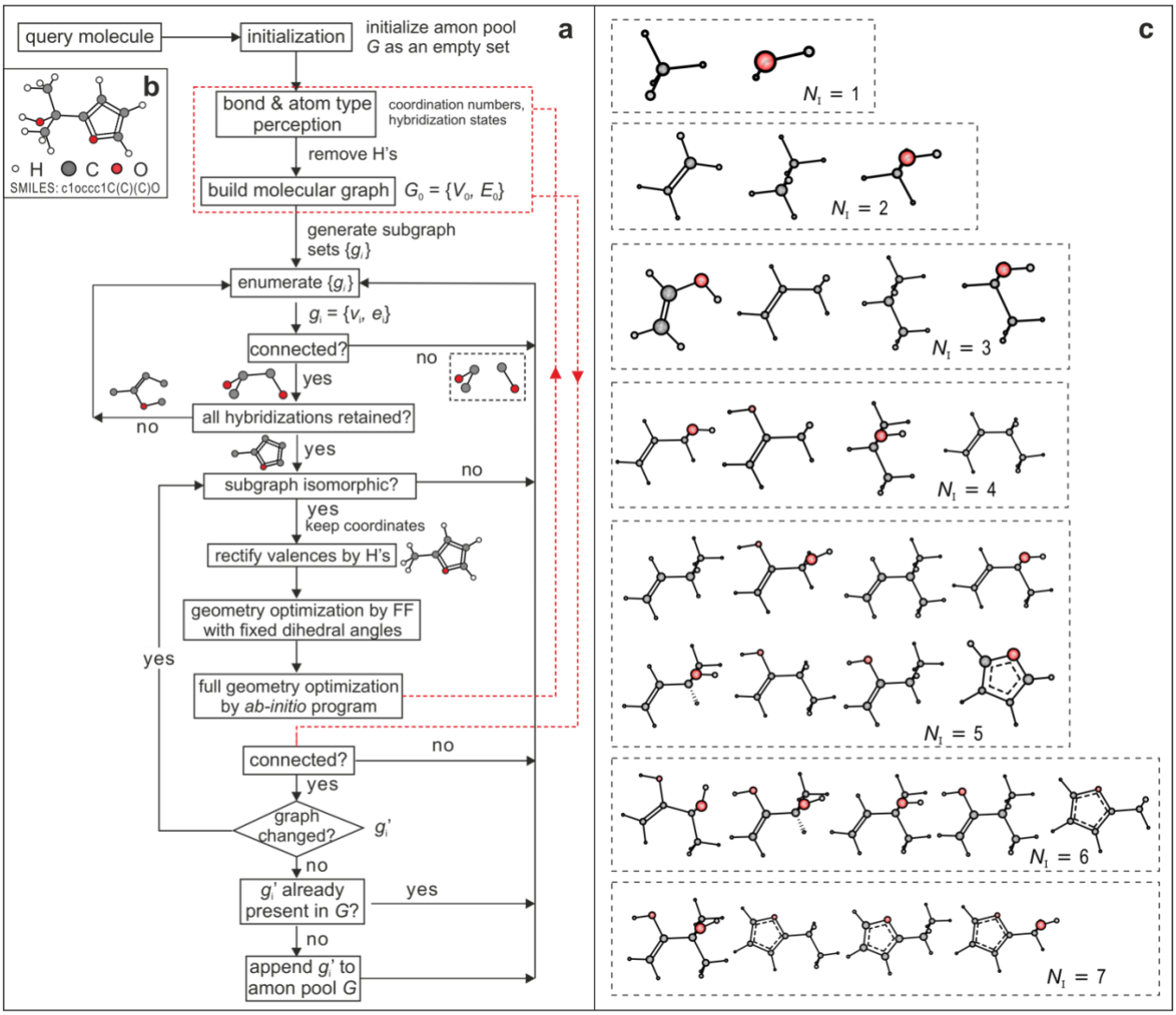
Bing Huang and O. Anatole von Lilienfeld*

Institute of Physical Chemistry and National Center for Computational Design and Discovery of Novel Materials (MARVEL),
Department of Chemistry, University of Basel,
Klingelbergstrasse 80, 4056 Basel, Switzerland

To whom correspondence should be addressed; E-mail: anatole.vonlilienfeld@unibas.ch.

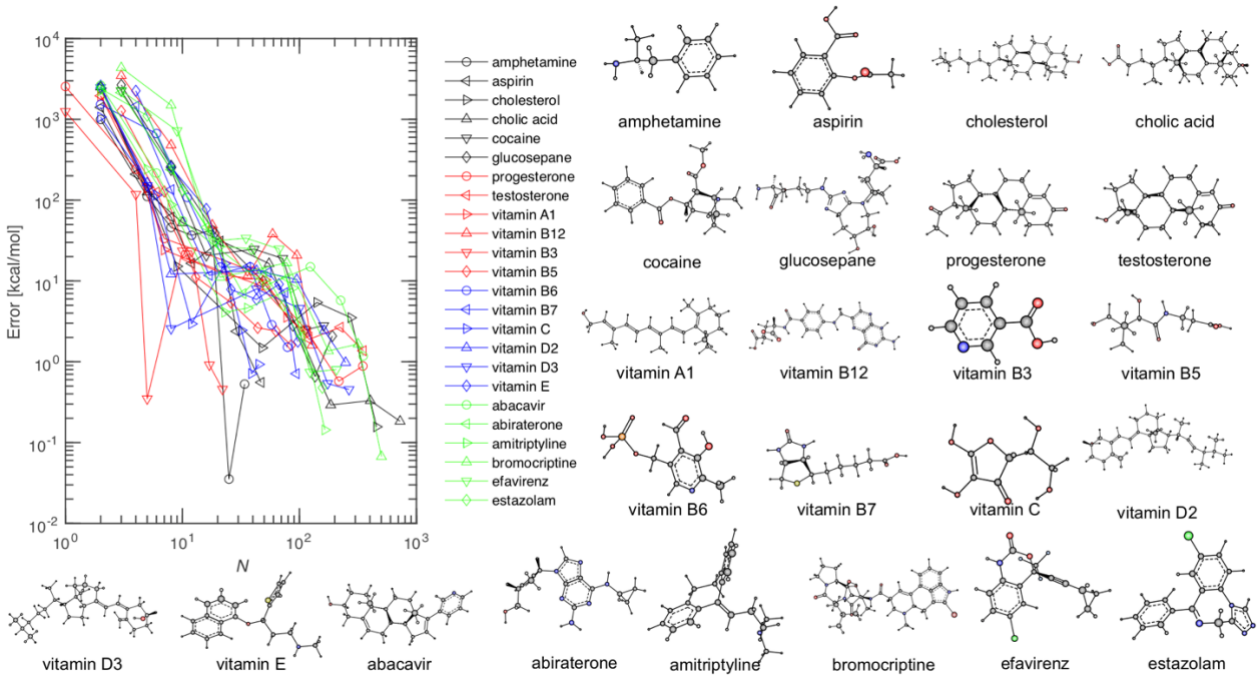
Table of Contents

Supplementary Figure 1 - Detailed flowchart for amon selection



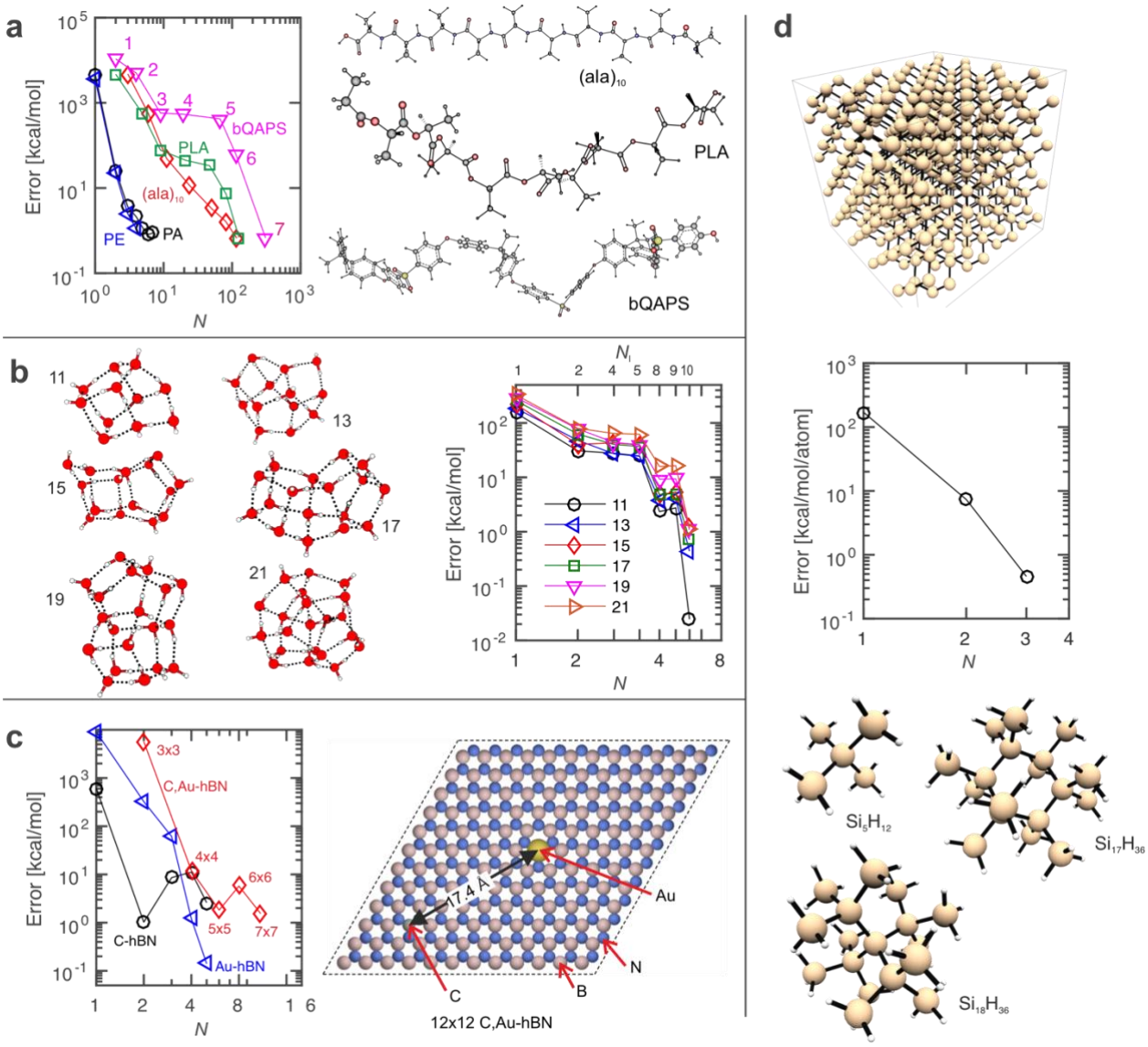
The flowchart in (**a**) is demonstrated for the query molecule 2-(furan-2-yl)propan-2-ol (**b**). The resulting amons selected are displayed in (**c**).

Supplementary Figure 2 – Scalability of AML demonstrated for 24 biomolecules.



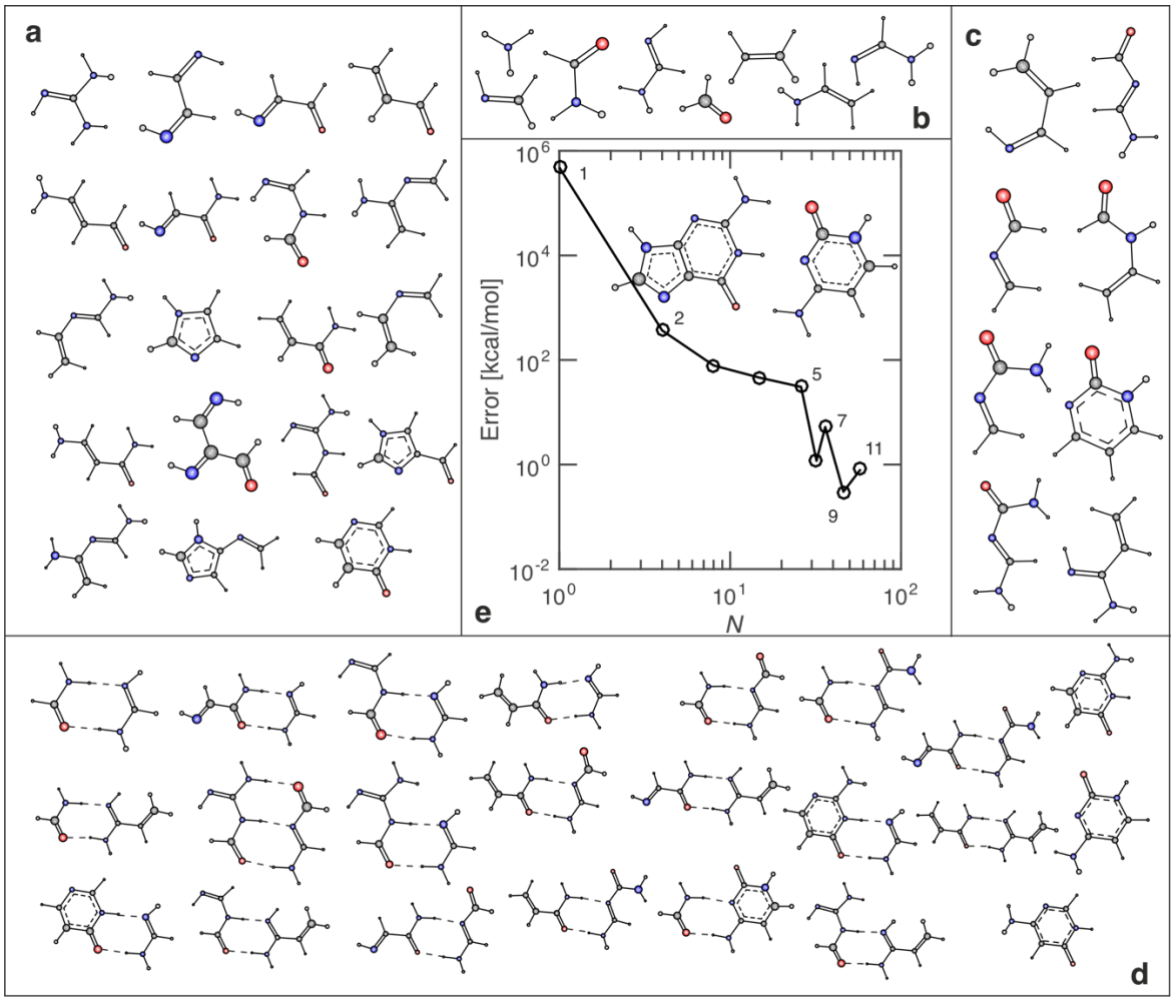
Displayed above are learning curves (absolute prediction error as a function of number of amons used for training, left) for 24 query biomolecules (right, also shown in Fig. 3 of the main text).

Supplementary Figure 3 – Scalability of the AML demonstrated by polymers, 2D materials and Si bulk



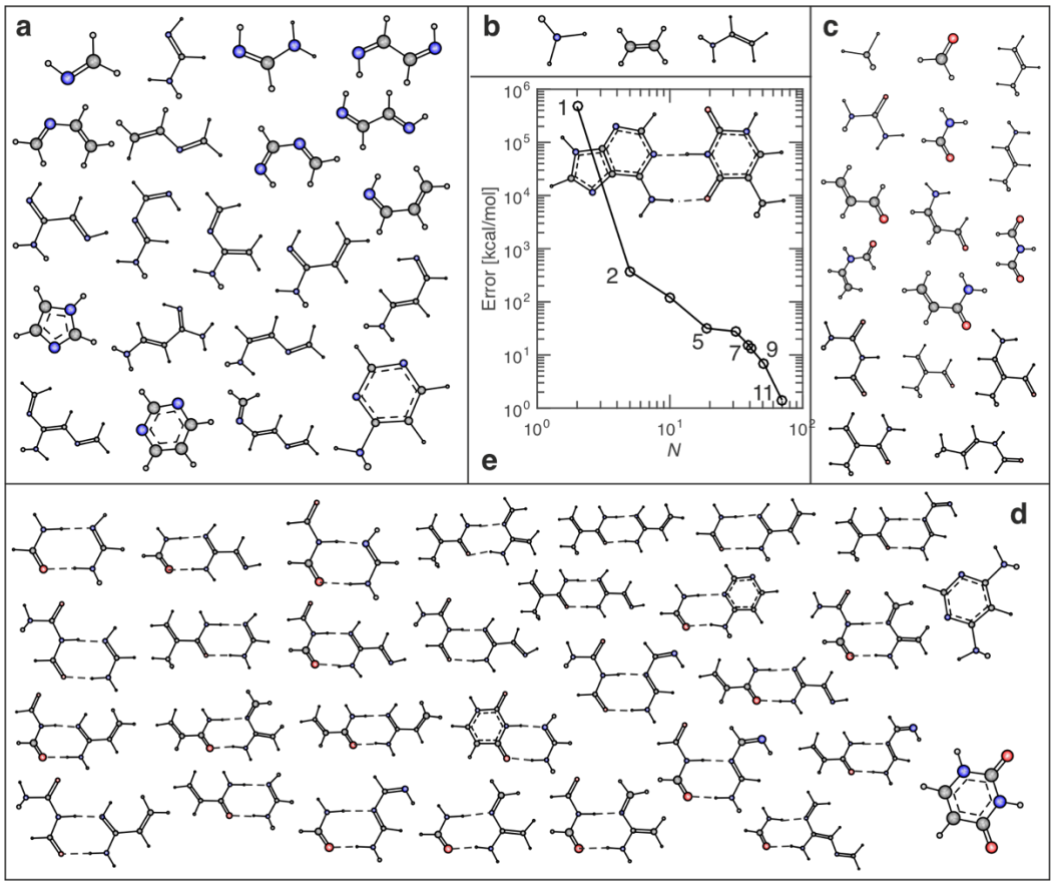
Learning curves are displayed for (**a**) six common polymers, including polyethylene (PE, with 28 monomers), polyacetylene (PA, with 15 monomers), alanine peptide ($(ala)_{10}$), polylactic acid (PLA, with 10 monomers) and the backbone of quaternary ammonium polysulphone (bQAPS, with 3 monomers). See Supplementary Fig. 8 for corresponding amons.; (**b**) 6 water clusters with number of water molecules being 11, 13, 15, 17, 19, and 21; (**c**) 2-D hexagonal BN sheets with one B atom replaced by gold (Au-hBN), or carbon (C-hBN), or two B atoms replaced by one gold and carbon (Au,C-hBN). Absolute errors are shown for per unit cell (each unit cell is of size 12x12). See Supplementary Fig. 10 for corresponding amons; (**d**) Bulk silicon. Amons are shown in the bottom panel, with chemical formula attached to each amon. Inset integers indicate the size of amons (i.e., number of heavy atoms).

Supplementary Figure 4 – Amons of Watson-Crick base-pair GC and energy prediction by AML



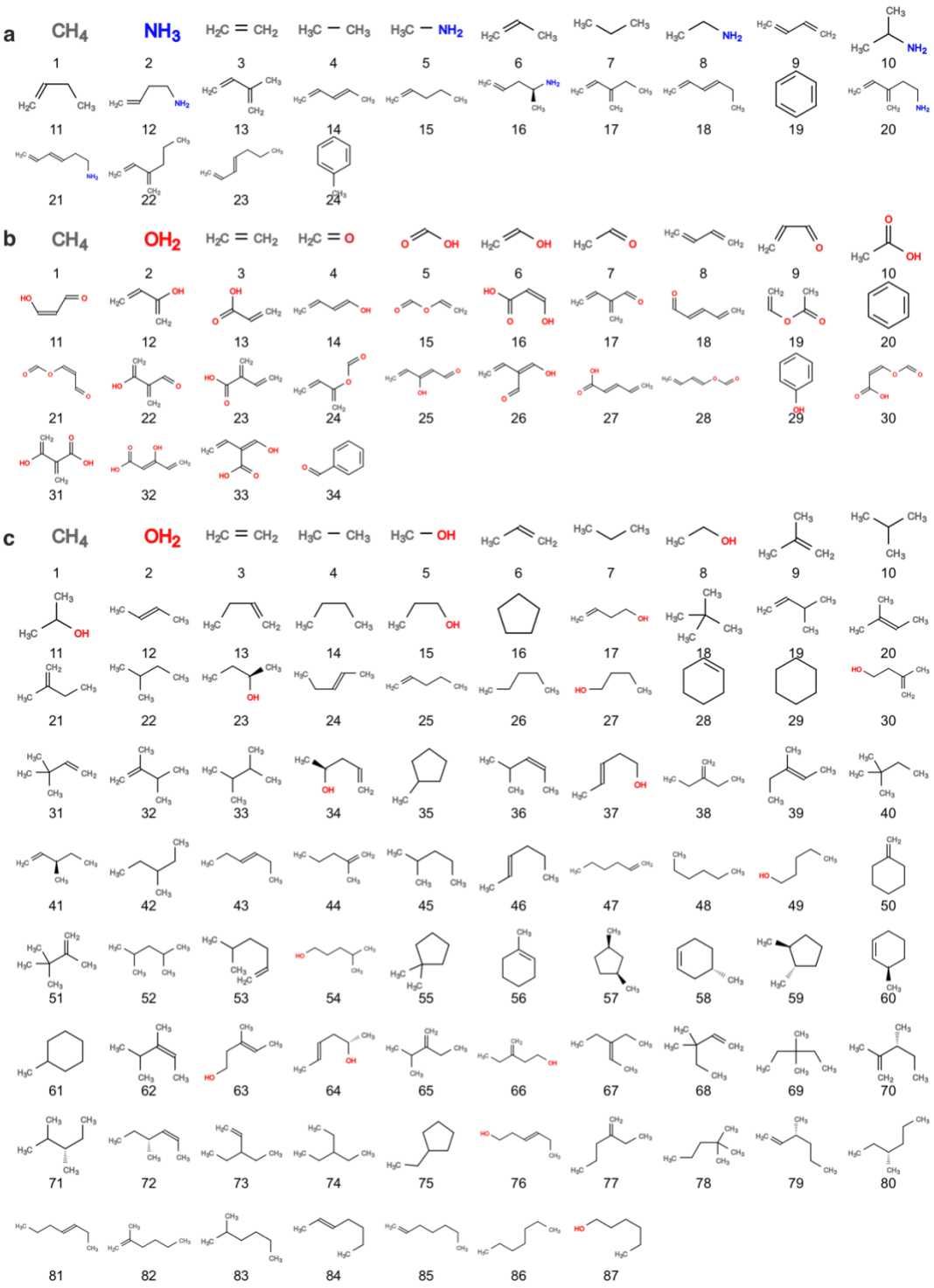
a. Amons of DNA base Guanine; **b.** Amons shared by Cytosine and Guanine; **c.** Amons of Cytosine; **d.** Amons of Watson-Crick bonding pattern, all complexes containing at least two hydrogen-bonds; **e.** Learning curve for total energy prediction of Cytosine- Guanine (CG, inset) base pair. Trained on all amons in A-D, AML underestimates the DFT energy of GC by 0.81 kcal/mol.

Supplementary Figure 5 – Amons of Watson-Crick base-pair AT and energy prediction by AML



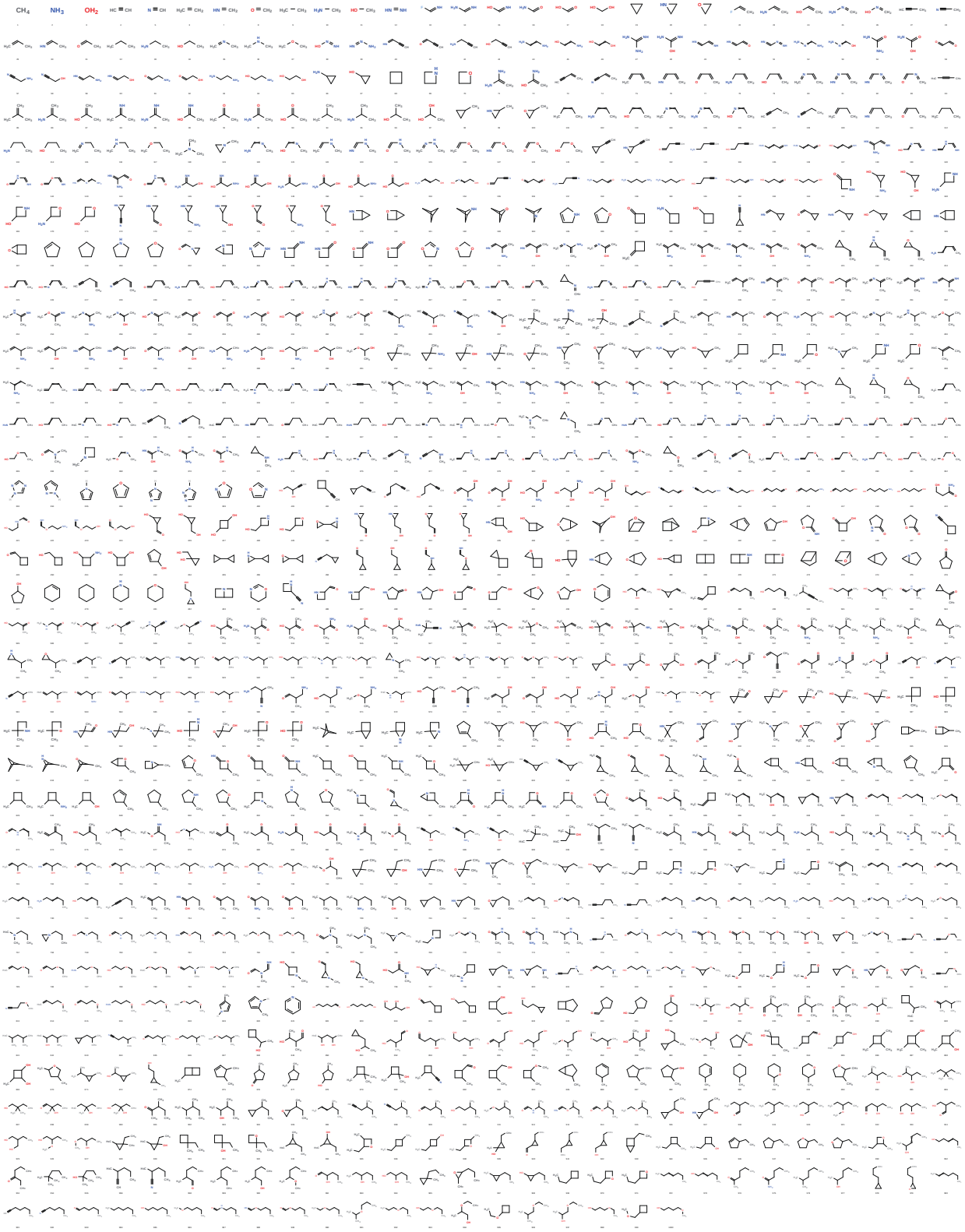
a. Amons of DNA base Adenine; **b.** Amons shared by Adenine and Thymine; **c.** Amons of Thymine; **d.** Amons of Watson-Crick bonding pattern, all complexes containing at least two hydrogen-bonds. **e.** Learning curve for the total energy prediction of Adenine-Thymine (AT, inset) base pair. Trained on all amons in A-D, AML underestimates the DFT energy of AT by 1.41 kcal/mol.

Supplementary Figure 6 – Amons of bio-moelcules

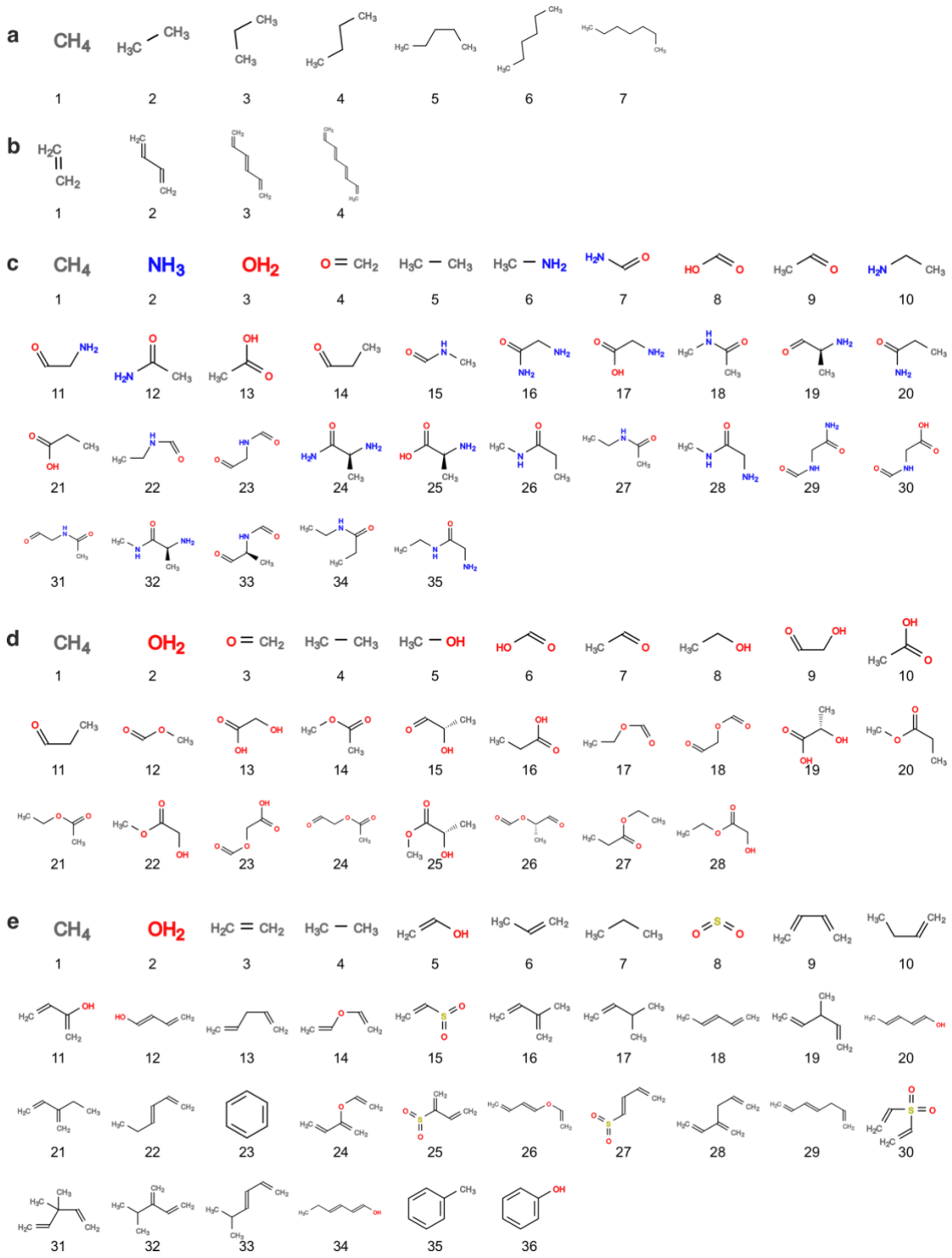


Displayed are amons of amphetamine (**a**), aspirin (**b**) and cholesterol (**c**). For better visualization, only amon graphs are shown.

Supplementary Figure 7 – The 1k most frequent amons of QM9 molecules

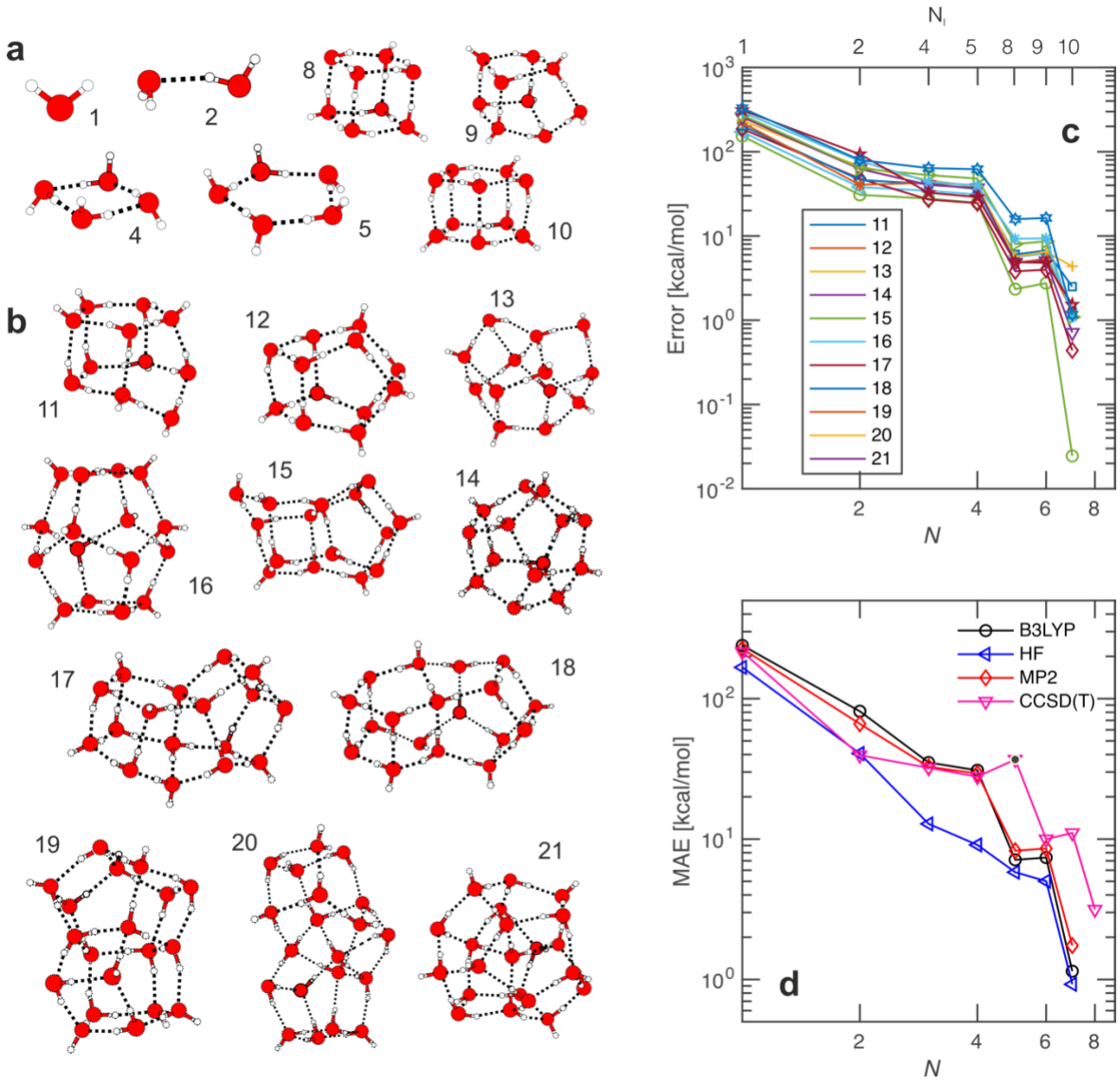


Supplementary Figure 8 – Amons of polymers



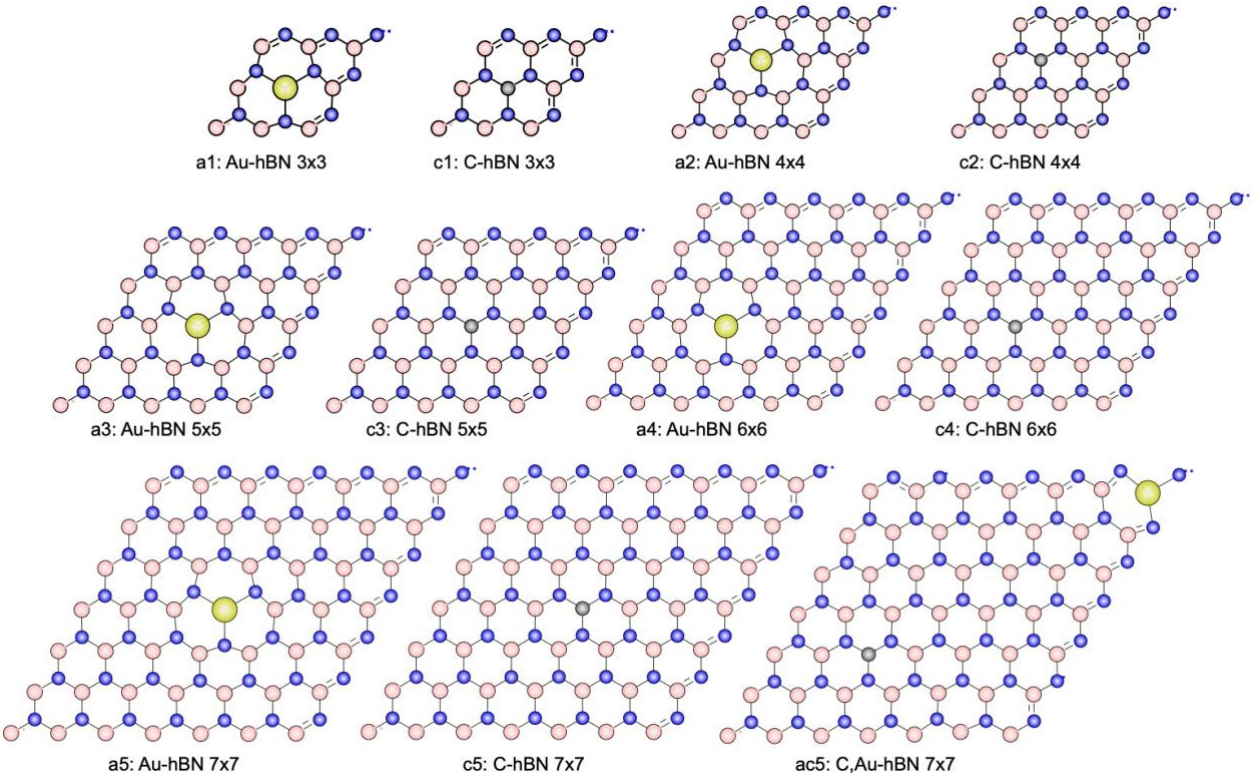
Amons of polyethylene (**a**), polyacetylene (**b**), alanine peptide (**c**), polylactic acid (**d**) and the backbone of quaternary ammonium poly- sulphone (bQAPS, **e**). Only amon graphs are shown for better visual effect.

Supplementary Figure 9 – AML predictions for water clusters



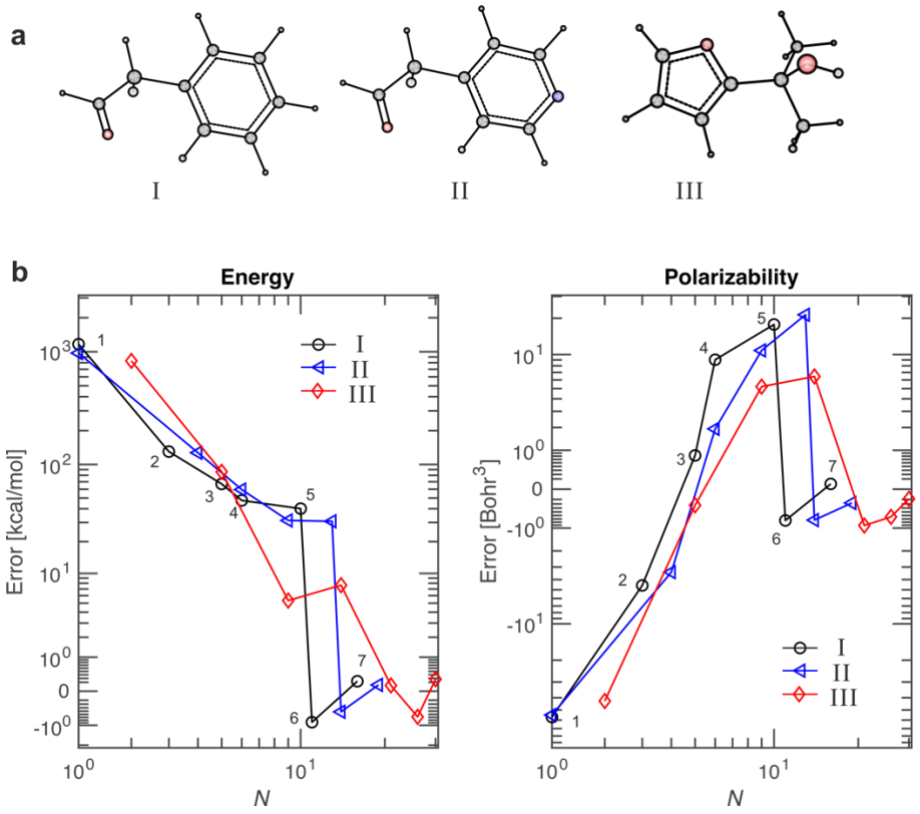
a. Small water clusters with $N_I \leq 10$, used as amons. Inset integer indicates the number of water molecules in each cluster; **b.** query water clusters, ranging from $(H_2O)_{11}$ to $(H_2O)_{21}$; **c.** Absolute error of predicted energies by AML as a function of amon number/size for each of the query water clus- ters in **b**. Geometries and energies were obtained at the level of Gaussian 09/B3LYP/6-31G*; **d.** Learning curve, reported as the mean absolute error (MAE) of total energy predictions by AML for all query water clusters as a function of number of amons. Reference data calculated at 4 levels of theory were used for training and test: HF (with HF coordinates, left-pointing blue triangle), MP2 (with MP2 coordinates, red diamond), B3LYP (with B3LYP coordinates, black circle) and CCSD(T) (with MP2 coordinates, downward-pointing pink triangle) (all with the same Pople basis 6-31G*).

Supplementary Figure 10 – Amons for 2D *h*BN sheets (w/wo dopant)



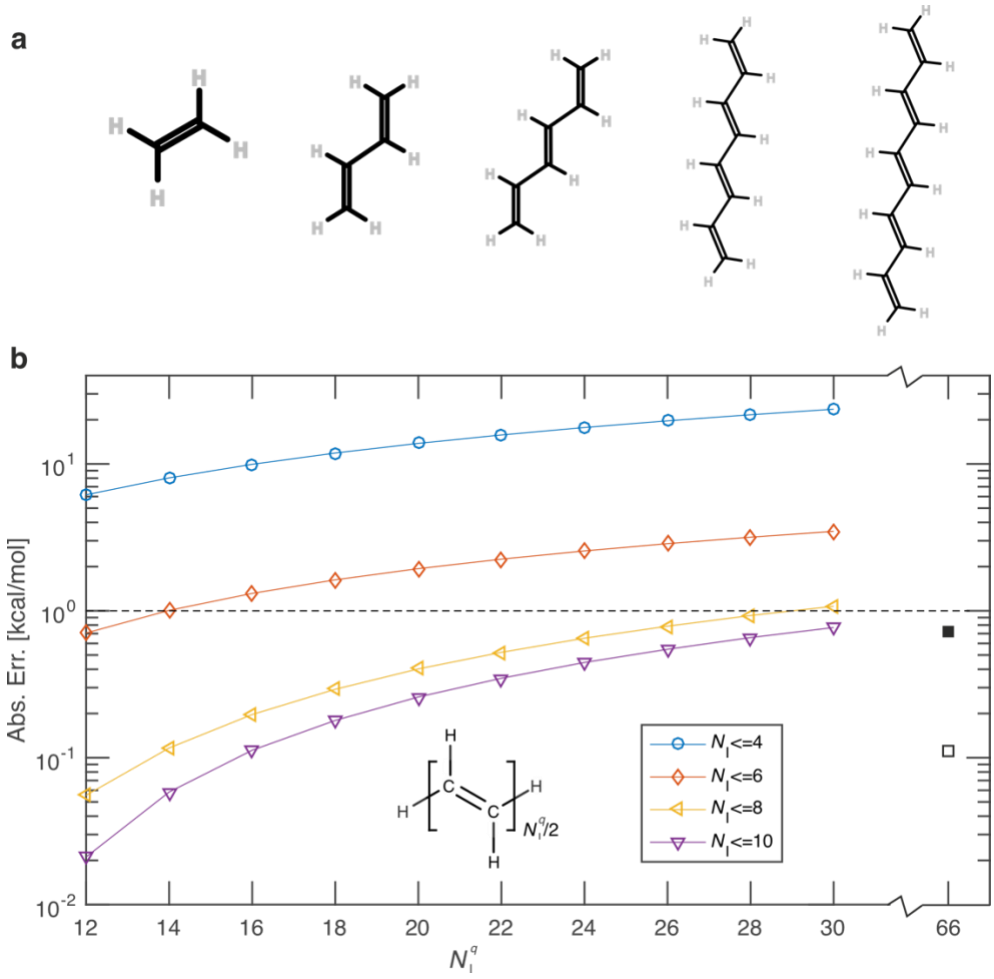
a1-a5: amons of Au-hBN 12x12, c1-c5: amons of C-hBN 12x12 (c1-c5). For C,Au-hBN 12x12, its amons include all structures displayed above.

Supplementary Figure 11 – AML predictions of energies and polarizabilities for three QM9 molecules



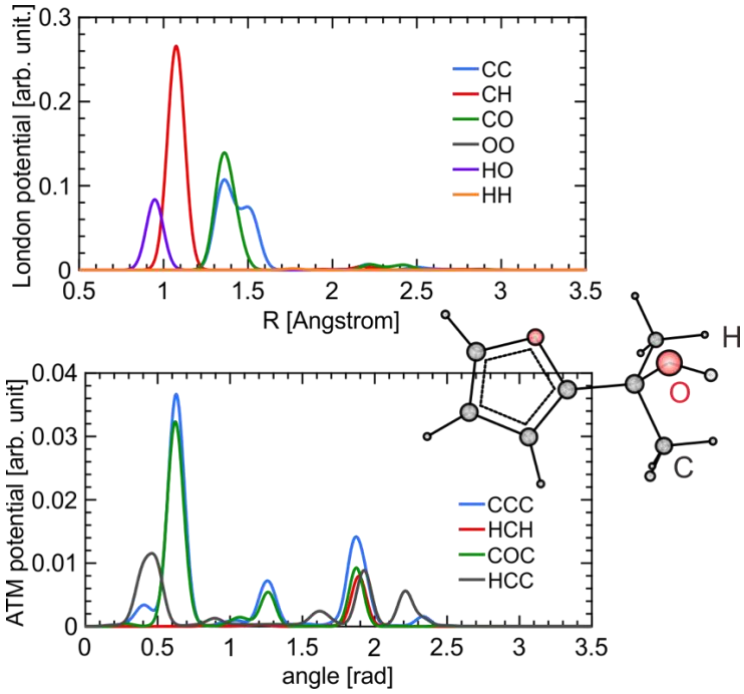
Learning curves (**b**) are shown as the signed difference beweetn AML-predicted values and corresponding DFT values (total energy: left panel of **b**, polarizability: right panel of **b**), plotted with respect to increasing number as well as size of amons, for the three QM9 molecules (**a**) in Fig. 2c.

Supplementary Figure 12 – Scalability of AML model demonstrated by (trans) poly-acetylene



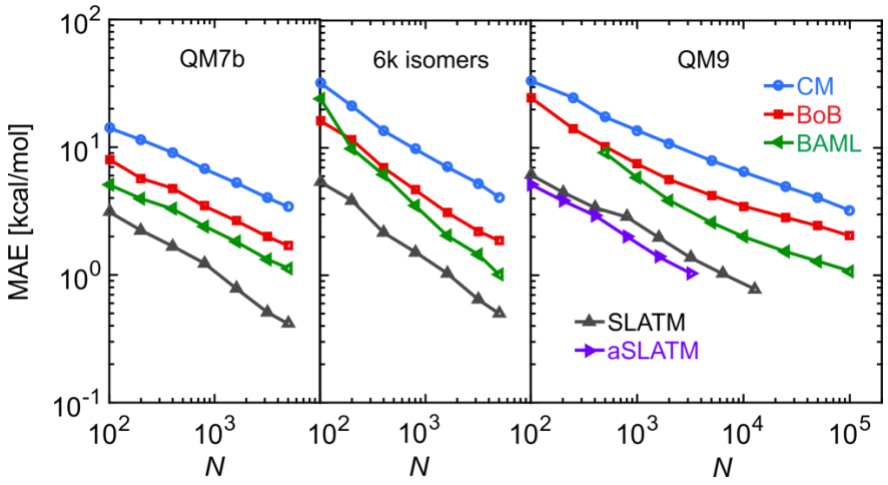
a. Amons employed with up to $N_I = 10$ heavy atoms. **b.** AML prediction error of atomization energy as a function of size of the query (for PA strands made from up to 15 acetylene monomers). Coloured symbols correspond to AML models trained on amons including increasingly larger instances with up to 4, 6, 8, or 10 heavy atoms, respectively. Squares show, for comparison, the error of a linear-scaling ab initio method for PA with $N_I = 56$ taken from reference 36.

Supplementary Figure 13 – Visualization of (a)SLATM representation



Distribution of 2-body London potential (upper panel) and 3-body ATM potential (lower pannel) for the inset query molecule 2-(furan-2-yl)propan-2-ol.

Supplementary Figure 14 –Comparison between (a)SLATM and other representation-based KRR models



Overall five representations are considered: Coulomb matrix (CM) [4], Bag of Bond (BoB) [5], Bond, Angle based ML (BAML) [6], SLATM and aSLATM for the prediction of total molecular potential energy for three datasets: QM7b (left panel), QM9 (right panel) and 6k constitutional isomers from QM9 (middle panel). For details of the KRR models, see the subsection "The SLATM representation" in the Methods section.