Supporting Information: Uncertainty quantification for predictions of atomistic neural networks

Luis Itza Vazquez-Salazar,[1, a)] Eric D. Boittier,[1] and Markus Meuwly[1, b)]

Department of Chemistry, University of Basel, Basel, Switzerland

(Dated: 23 February 2024)

[a)] Electronic mail: luisitza.vazquezsalazar@unibas.ch

[b)] Department of Chemistry, Brown University, USA; Electronic mail: m.meuwly@unibas.ch

I. CALCULATION OF THE MEAN DISTANCE BETWEEN TRAINING AND TEST MOLECULES IN FEATURE SPACE.

The mean distance between molecules from the tautobase and the molecules in QM9 was calculated as follows. First, for each of the molecules in the test set, molecules with the same number of atoms in the tautobase were filtered. This is necessary because the size of the matrices for the Radial Base Functions (RBFs) and the Atomic Embeddings (AtEs) need to be equal. In the second step, the pairwise distance between the test molecule and the filtered molecules was calculated using the Euclidean norm between points as

$$||X||_i = \frac{1}{n}\sqrt{\sum_{j=1}^{n}\sum_{k=1}^{n}(x_j - y_k^i)^2} \tag{1}$$

where x_j is the element in the RBF/AtE matrix of the test molecule, y_k^i is the element in the RBF/AtE matrix for each of the retained molecules, and n is the element on the RBF matrix/embedding matrix. Therefore the value obtained from equation 1 is the mean distance between the RBF/AtE matrix of the test molecule and the ith-molecule on the training dataset.

In the next step, the distance between each of the molecules and the test molecule is averaged to obtain the average distance in feature space (RBF and AtE) between the target molecule and the molecules with the same number of atoms in embedding space as:

$$\langle \text{RBF/AtE} \rangle = \frac{1}{N}\sum_{i=1}^{N}||X||_i \tag{2}$$

A. Construction of polar plots

The polar plots in Figures 7B, C, and S14 to S16 represent a projection of the distance between a test molecule (at the center) and the closest molecules in embedding space. The value for ||AtE|| and ||RBF|| between the test molecule and the molecule in the training dataset obtained from 1 are transformed to polar values according to the following expressions:

$$r = \sqrt{||AtE||^2 + ||RBF||^2} \tag{3}$$

$$\theta = \arctan \frac{||AtE||}{||RBF||} \tag{4}$$

II. TABLES

λ	Calibration	MSE	MV	Sharpness	Dispersion	ENCE	C_v	MA	Sensitivity	Precision	Accuracy
0.01	Poor	0.0012	3.3895	0.0078	0.5500	0.1930	1.3041	0.1765	0.0678	0.8000	0.9283
0.1	Poor	0.0009	0.0094	0.0133	0.4400	0.1465	0.3196	0.0184	0.0657	0.7917	0.9120
0.2	Regular	0.0013	0.0099	0.0137	0.5700	0.1416	0.3080	0.0229	0.0928	0.8800	0.9302
0.4	Good	0.0017	0.0058	0.0113	0.4500	0.1422	0.6542	0.0549	0.1289	0.9615	0.9456
0.5	Poor	0.0010	0.4827	0.0189	0.5800	0.1383	0.1775	0.0435	0.0370	0.8000	0.8992
0.75	Good	0.0014	0.0011	0.0179	0.5100	0.1545	0.2554	0.0242	0.1042	0.6818	0.9130
1	Poor	0.0009	0.0016	0.0196	0.4700	0.1649	0.2020	0.0735	0.0988	0.6538	0.8950
1.5	Regular	0.0012	0.0009	0.0165	0.3800	0.1551	0.2357	0.0418	0.1474	0.6491	0.9251
2	Poor	0.0020	0.0008	0.0220	0.1500	0.1684	0.2542	0.0952	0.2836	0.1932	0.8778

TABLE S1. Summary of the properties tested for calibration of the evaluated models. Units when necessary are in eV. MSE: Mean Square Error, MV: Mean Variance, ENCE: Expected Normalized Calibration Error, MA: Miscalibration Area.

III. FIGURES

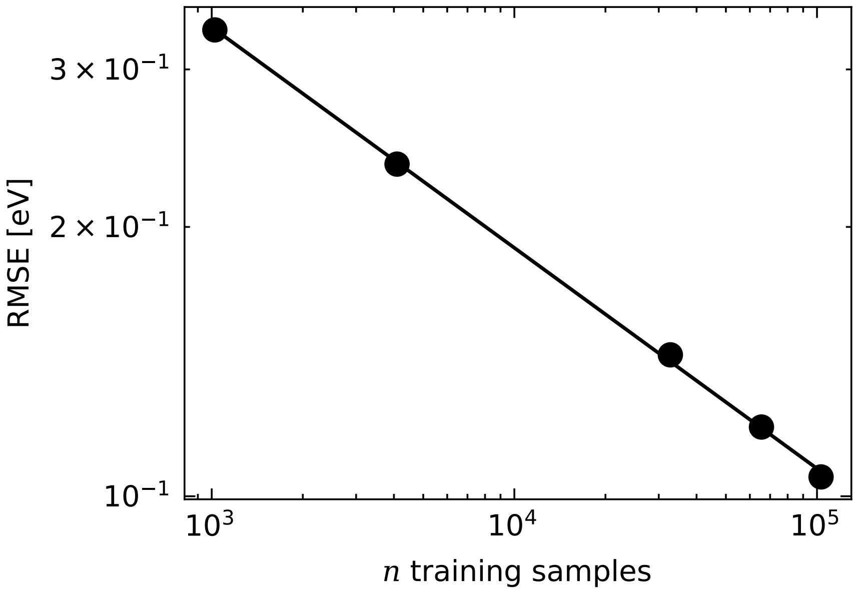

FIG. S1. Learning curve showing the model improvement with respect to training set size and are consistent with expectations.

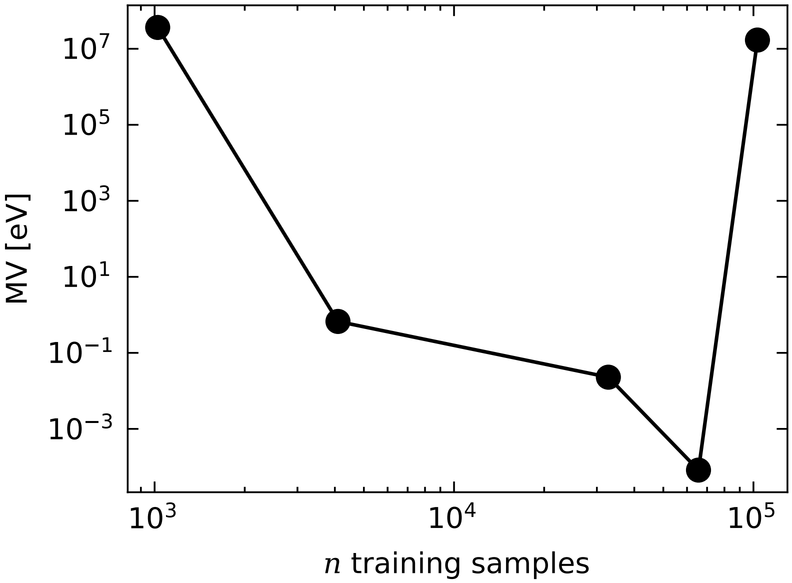

FIG. S2. Impact of the training set size on the mean variance of predicted energies. Values outside the 99th percentile were removed. Variance is observed to decrease with increasing training set size until a certain point where models trained on the most extensive training corpus predicted higher variance.

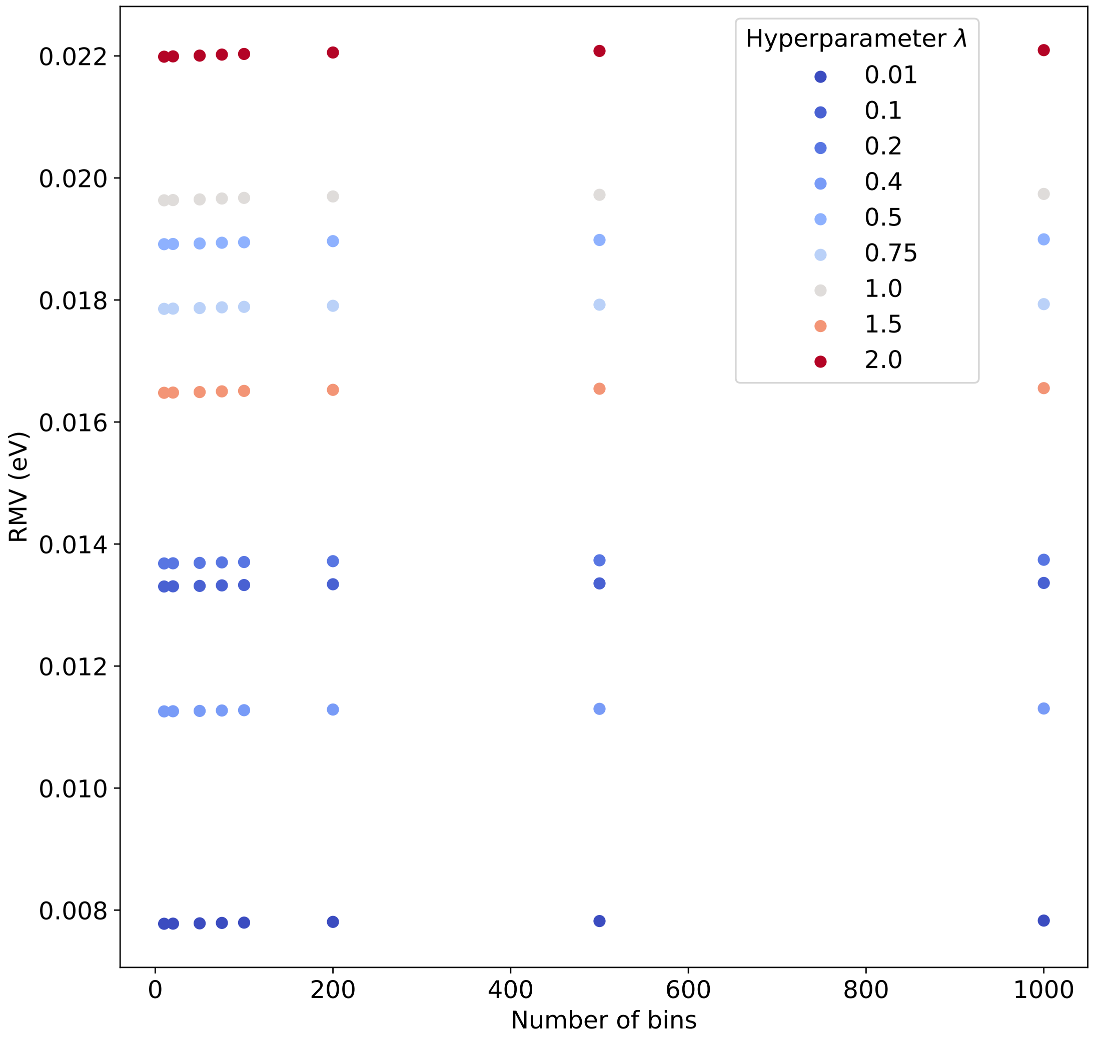

FIG. S3. **Root Mean Variance depending on the number N of bins for different values of the hyperparameter λ used for calculation. In the main manuscript $N = 100$ was used.**

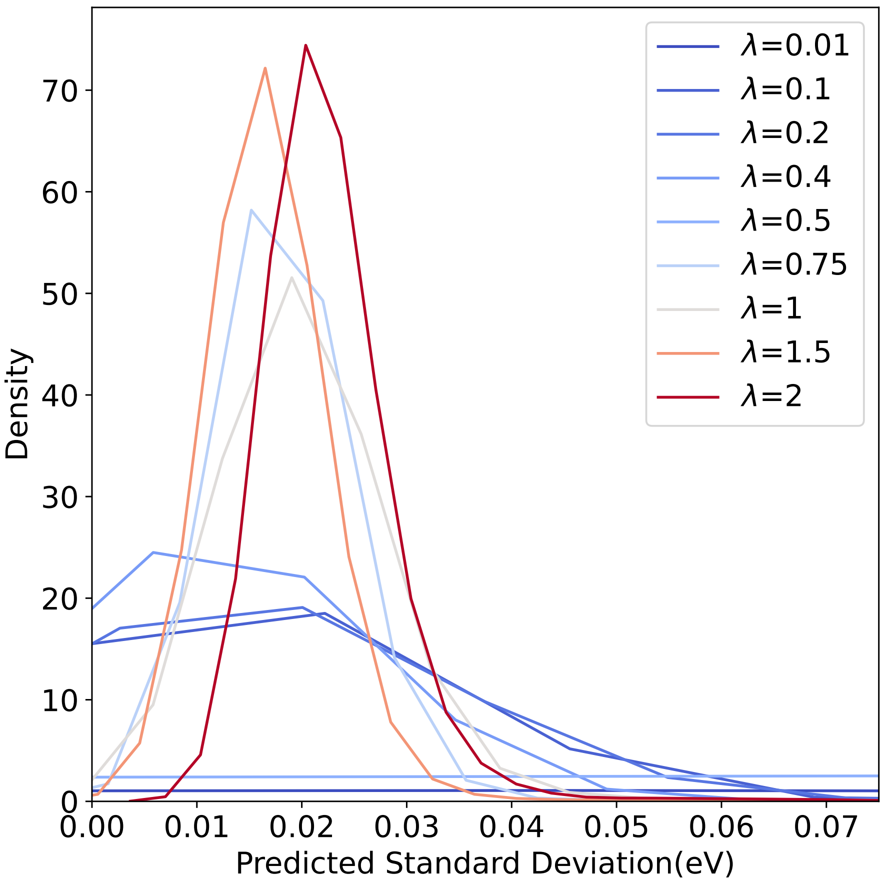

FIG. S4. Distribution of standard deviation($\sigma = \sqrt{Var}$) for different values of hyperparameter λ.

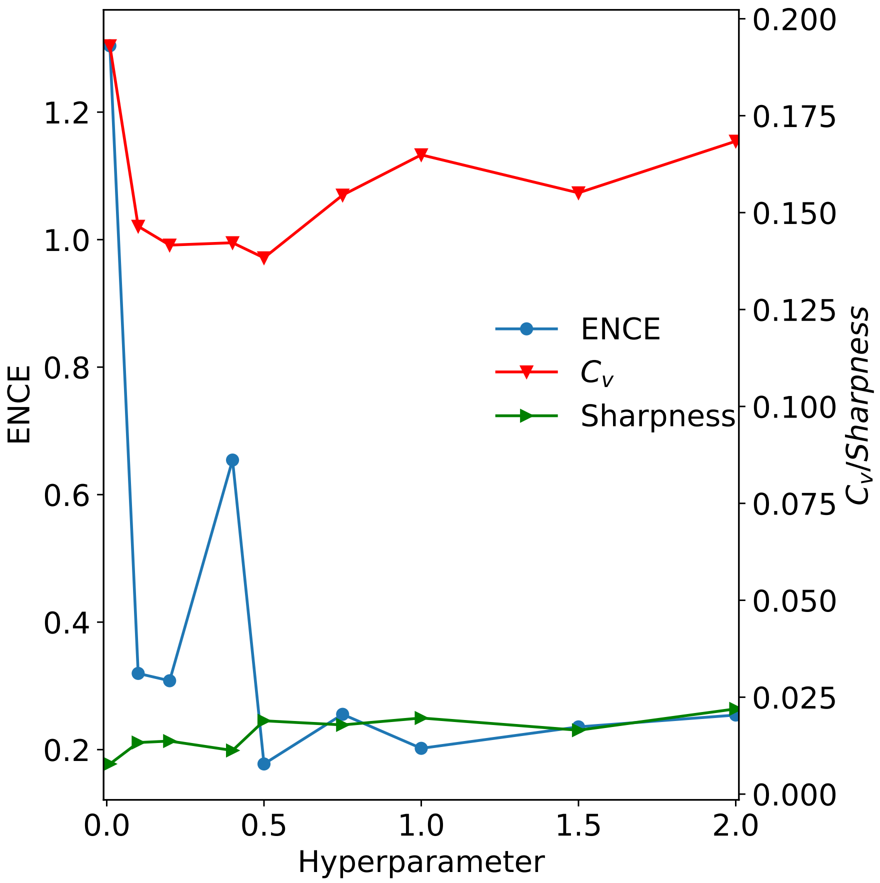

FIG. S5. Evolution of the Expected Normalized Calibration Error (ENCE), sharpness, and the Coefficient of Variation (C_v) depending on λ for 95% of the variance distribution.

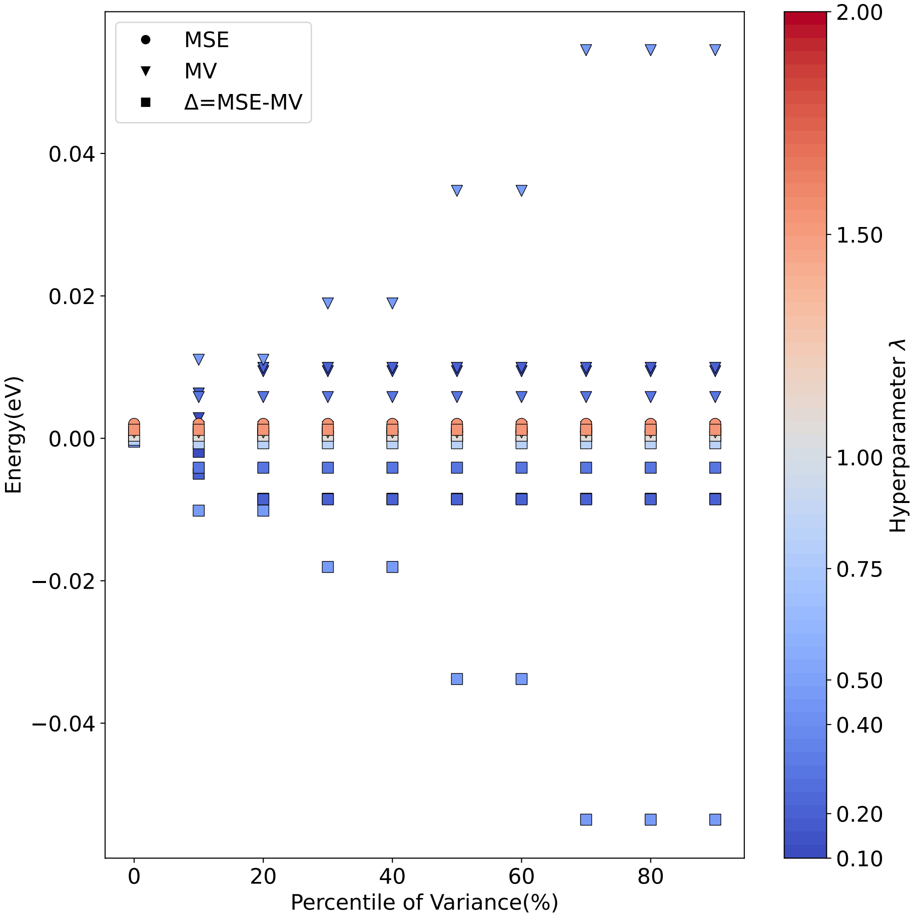

FIG. S6. Difference between Mean Squared Error (MSE) and Mean Variance (MV) for different percentiles of the predicted variance. Values for MSE, MV and its difference (Δ) at a given quantile are shown with different labels. The color bar indicates the values of the hyperparameter λ. The different dots are colored accordingly to its λ value. Value for 0.01 was excluded for clarity.

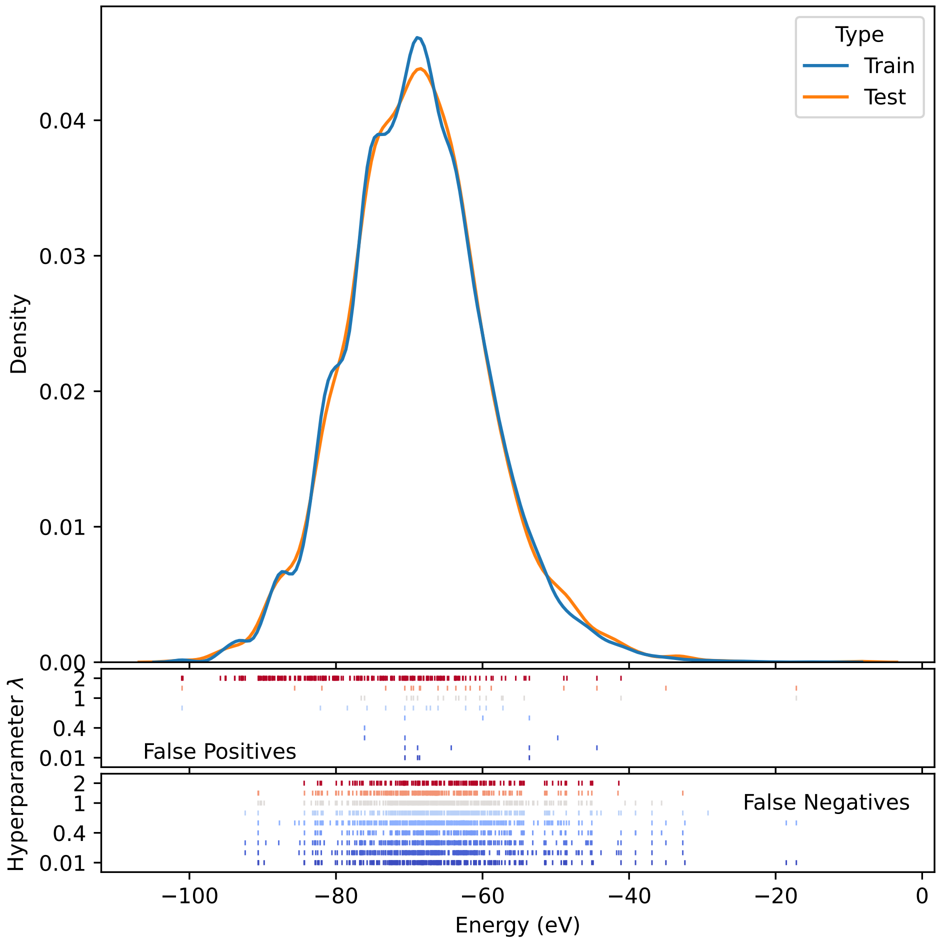

FIG. S7. Distribution of the energy of the molecules in the QM9 database used for training (blue line) and testing (orange line). The energies in the two sets are very similarly distributed. In the bottom half of the figure, rug plots, which are one dimensional scatter plots to show the location of the real values in the distribution, for all values of the hyperparameter λ. The top rug plot shows the molecules classified as False Positives, ($\varepsilon_i < \varepsilon^*$ and $\sigma_i > \sigma^*$). The bottom rug plot displays the molecules classified as False Negatives ($\varepsilon_i > \varepsilon^*$ and $\sigma_i < \sigma^*$). Here ε^* is the Mean Squared Error (MSE) and σ^* is the Mean Variance (MV). The color code for the rug plots correspond to the different values of the hyperparameter λ. It is noted that the number of FPs decreases considerably for decreasing λ whereas for FNs this number is rather insensitive to the value of the hyperparameter, see Figure 5A.

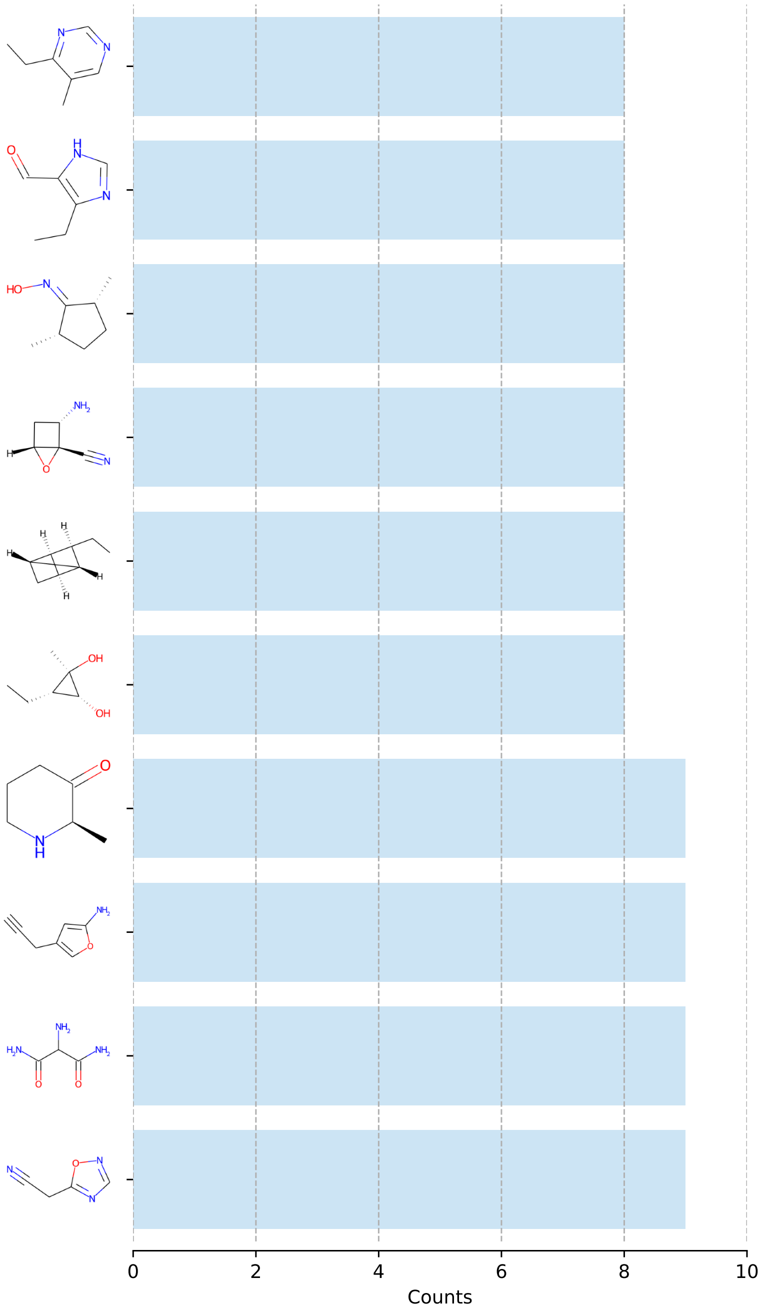

FIG. S8. Top 10 Common True Positive (TP) molecules ($\varepsilon_i > \varepsilon_{max}$ and $\sigma_i > \sigma_{max}$). The $x-$axis shows how often a molecule appear as TP for the different values of hyperparameter λ, the $y-$axis show the characteristic chemical structure. There are only show molecules that appear for at least four different values of λ

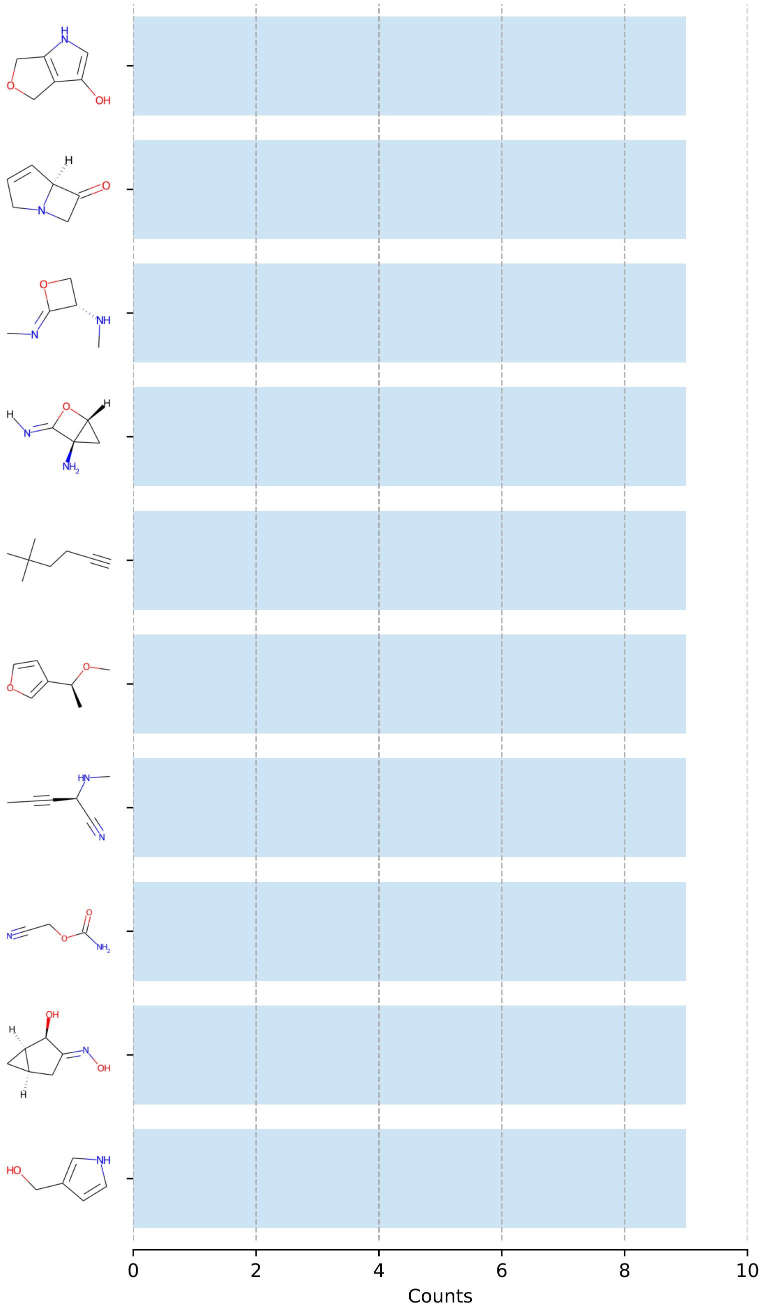

FIG. S9. Top 10 Common True Negative (TN) molecules ($\varepsilon_i < \varepsilon_{max}$ and $\sigma_i < \sigma_{max}$). The $x-$axis shows how often a molecule appear as TN for the different values of hyperparameter λ, the $y-$axis show the characteristic chemical structure.

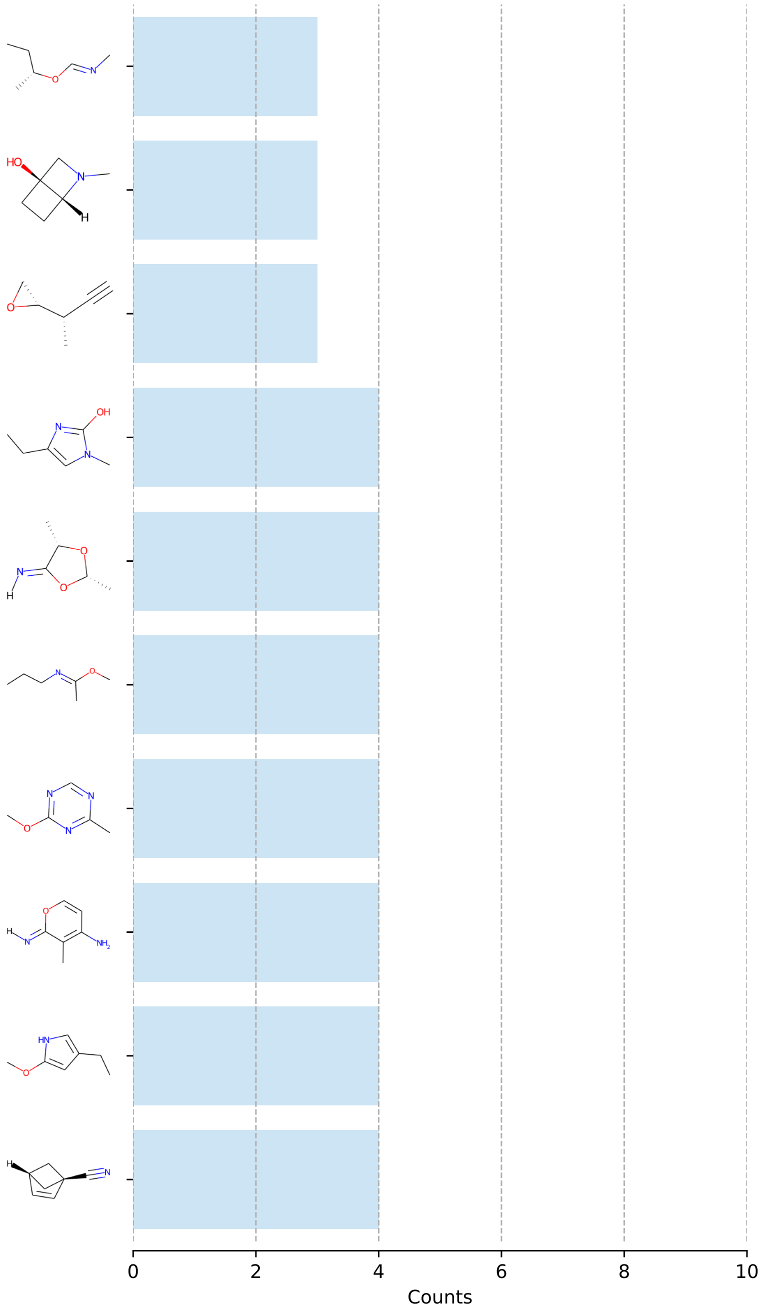

FIG. S10. Top 10 Common False Positive (FP) molecules ($\varepsilon_i < \varepsilon_{max}$ and $\sigma_i > \sigma_{max}$). The $x-$axis shows how often a molecule appear as FP for the different values of hyperparameter λ, the $y-$axis show the characteristic chemical structure.

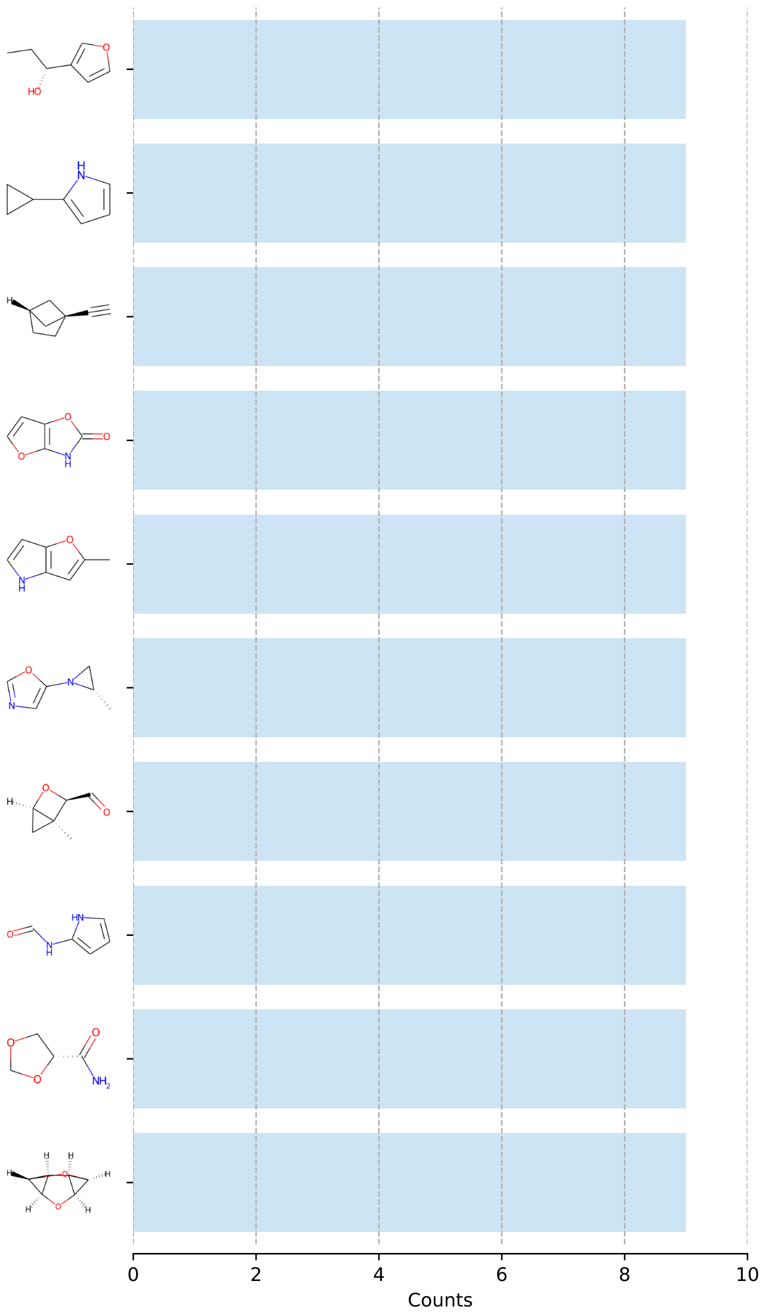

FIG. S11. Top 10 Common False Negative (FN) molecules ($\varepsilon_i > \varepsilon_{max}$ and $\sigma_i < \sigma_{max}$). The $x-$axis shows how often a molecule appear as FN for the different values of hyperparameter λ, the $y-$axis show the characteristic chemical structure.

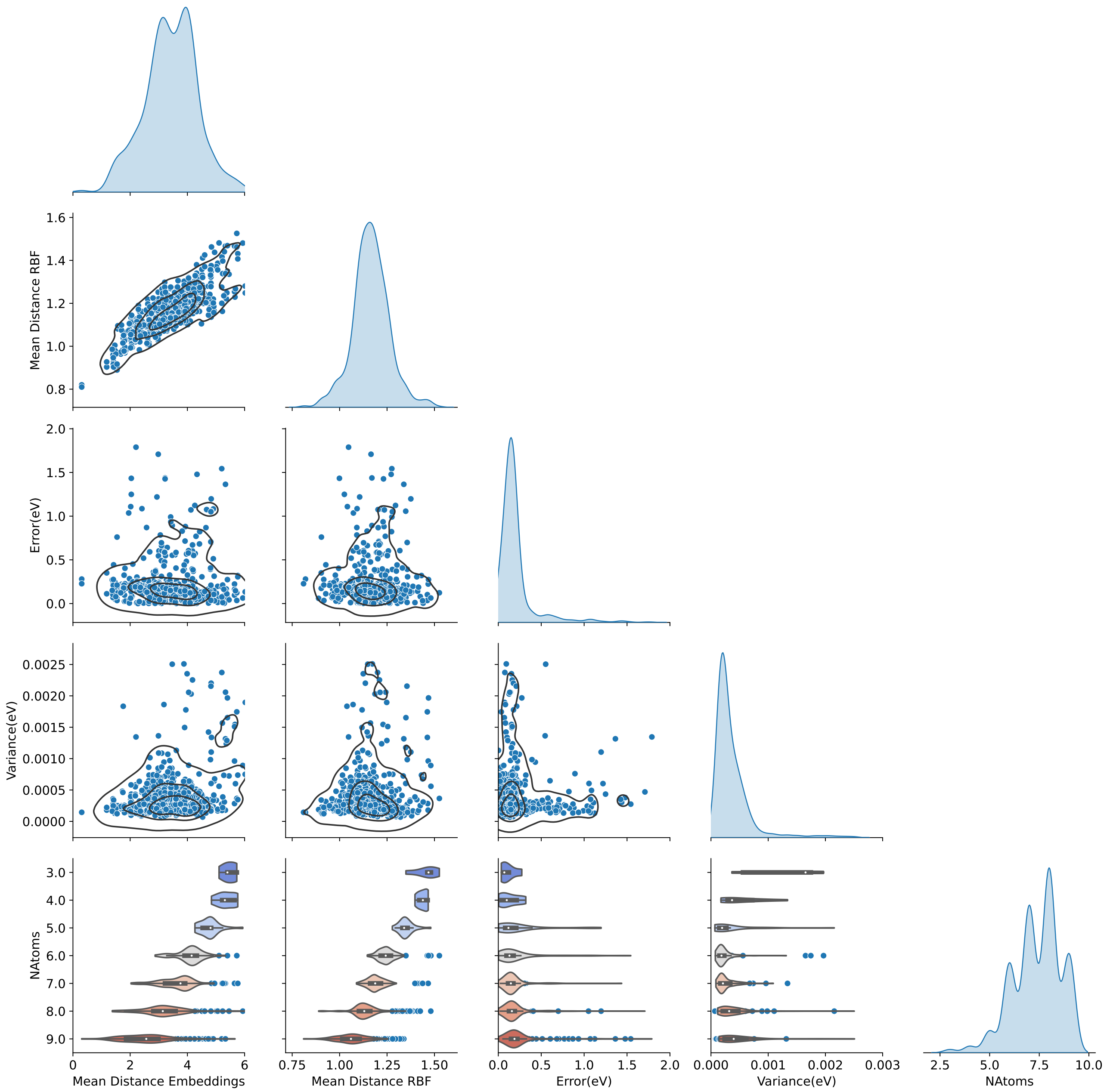

FIG. S12. Overview of the results for the evaluation of molecules on the tautobase for $\lambda = 0.2$. The diagonal of the figure shows the kernel density estimate of the considered properties (Mean Distance Embeddings, Mean Distance RBF, Error (eV),Variance (eV) and Number of Atoms). For each of the panels a correlation plot between the variable and a 2D kernel density estimate is shown. In the last row, violin plots for the different considered properties with respect to the number of atoms is shown.

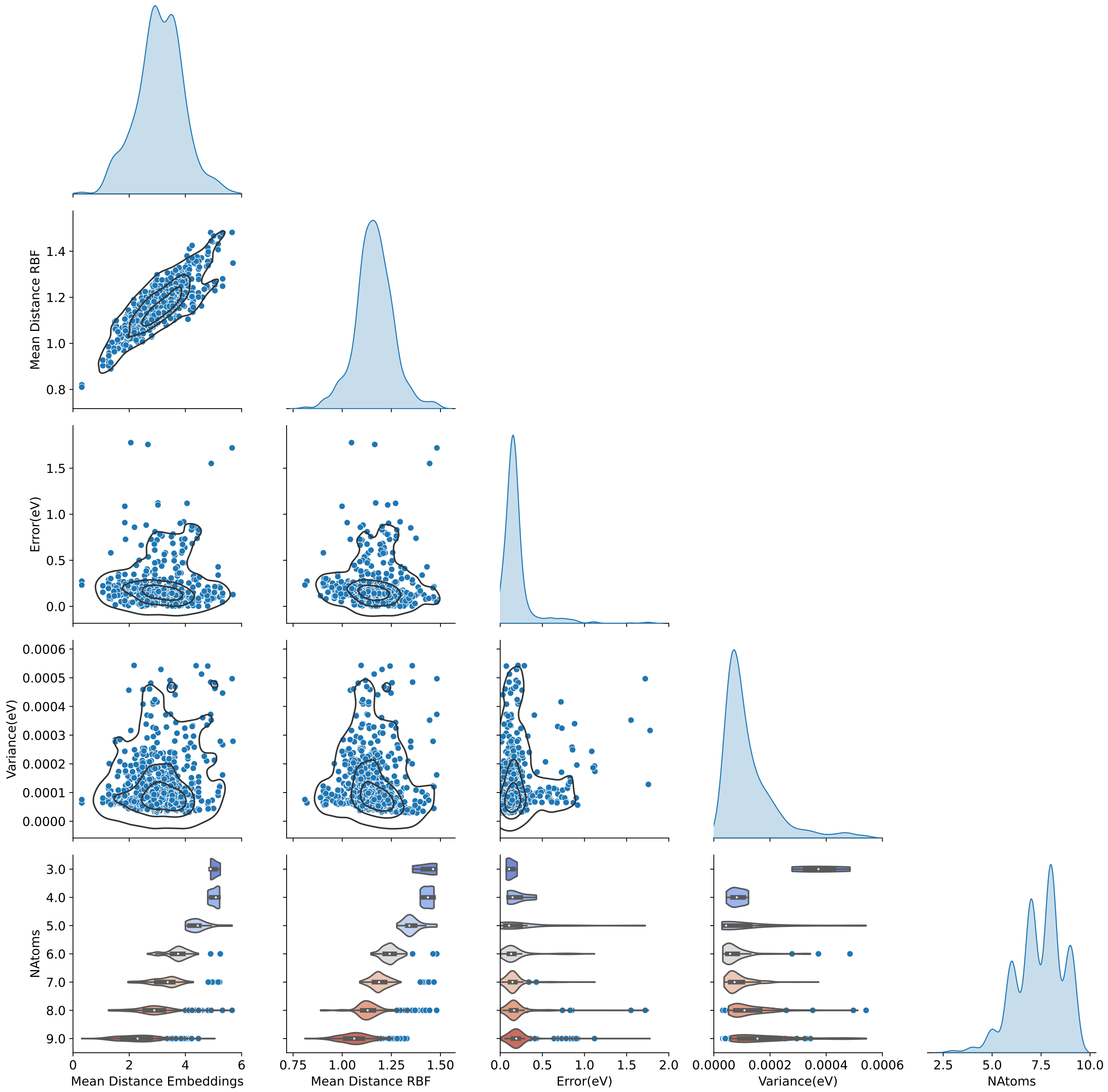

FIG. S13. Overview of the results for the evaluation of molecules on the tautobase for $\lambda = 0.4$. The diagonal of the figure shows the kernel density estimate of the considered properties (Mean Distance Embeddings, Mean Distance RBF, Error (eV), Variance (eV) and Number of Atoms). For each of the panels a correlation plot between the variable and a 2D kernel density estimate is shown. In the last row, violin plots for the different considered properties with respect to the number of atoms is shown.

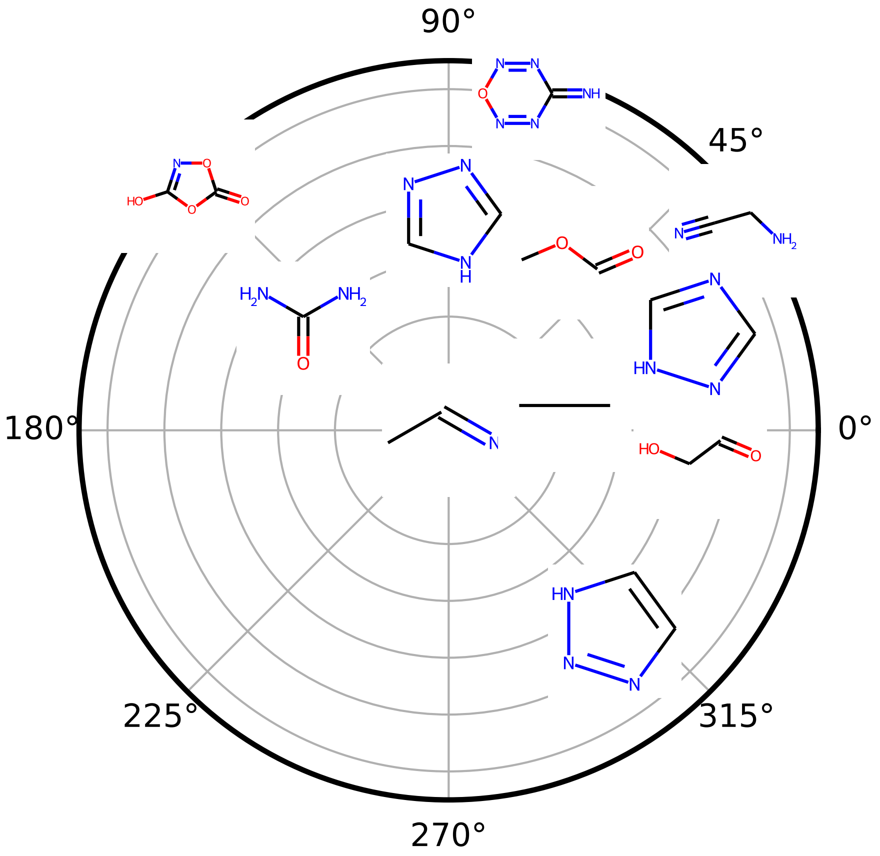

FIG. S14. Polar plot of the mean distance between a molecule in the test set (Ethanimine) and molecules on the training set. The molecule in the center corresponds to the sample with the largest mean distance in RBF and embedding space.

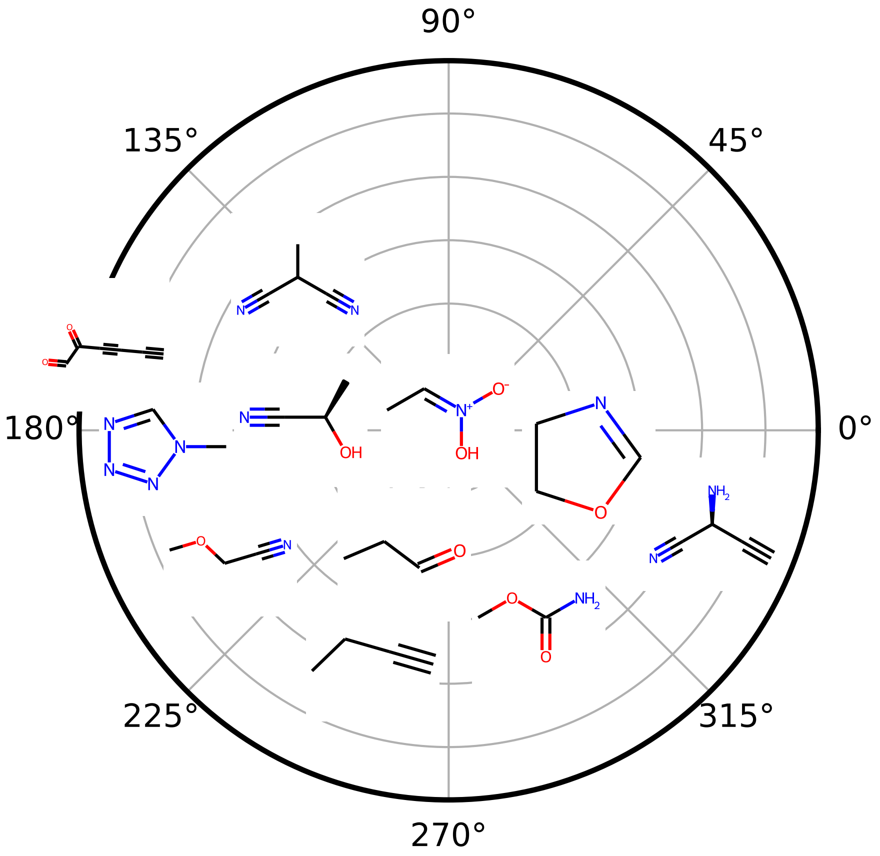

FIG. S15. Polar plot of the mean distance between a molecule in the test set ((1E)-Ethylideneazinic acid) and molecules in the training set. The molecule in the center corresponds to the molecule in the test set with the largest prediction error at 95th percentile.

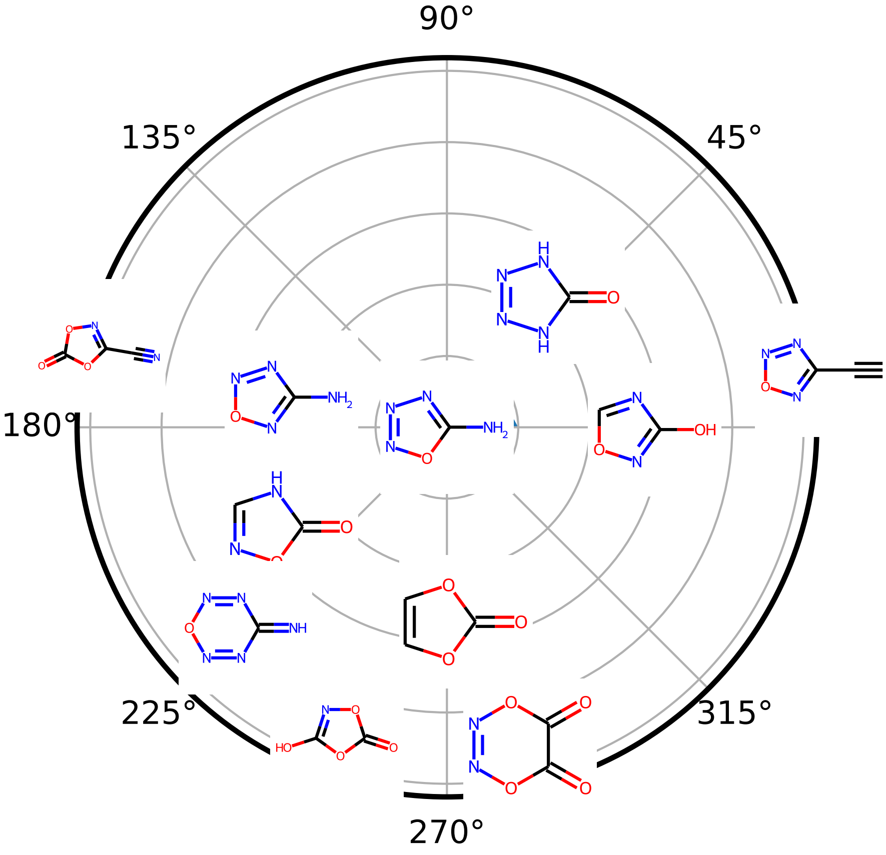

FIG. S16. Polar plot of the mean distance between a molecule in the test set (1,2,3,4-Oxatriazol-5-amine) and molecules in the training set. The molecule in the center corresponds to the molecule in the test set with the largest predicted variance at 95th percentile.

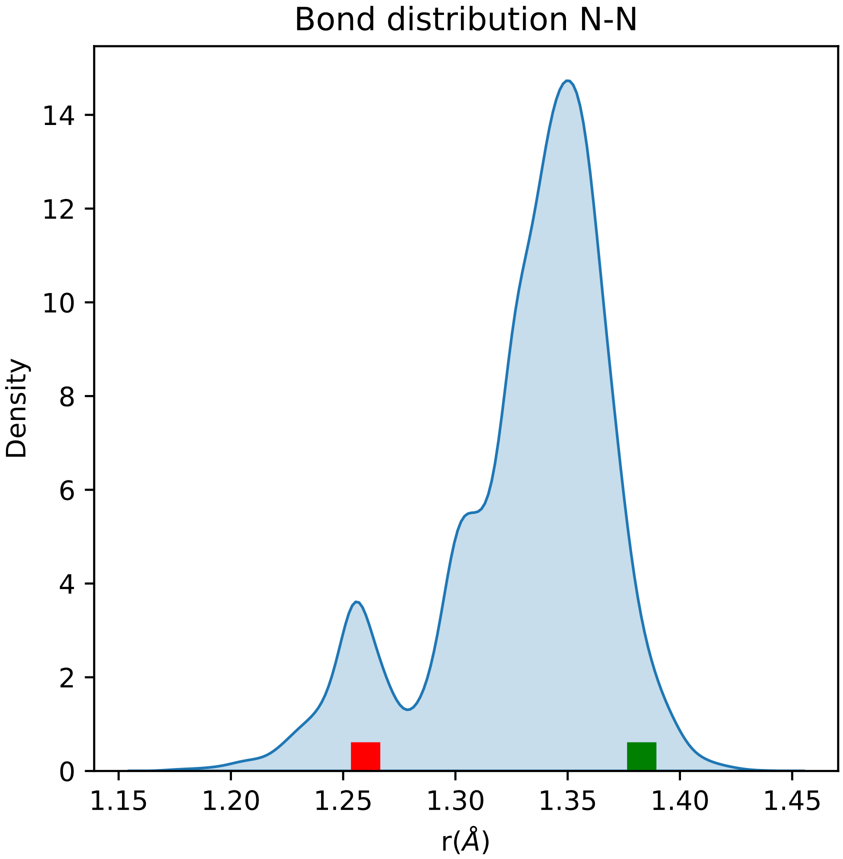

FIG. S17. N-N bond distribution from all molecules in the QM9 database[1]. The red square indicates the N-N bond distance for molecule A1 in Figure 8B and the green square indicates that for molecule B1.

REFERENCES

[1]R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. Von Lilienfeld, Sci. Data **1**, 140022 (2014).