Supporting Information: Enhancing chemical databases for atomistic machine learning by sampling conformational space

Luis Itza Vazquez-Salazar[a)] and Markus Meuwly[b)]

Department of Chemistry, University of Basel, Basel, Switzerland

(Dated: 26 February 2024)

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

[b)]Electronic mail: m.meuwly@unibas.ch

I. SUPPLEMENTARY TABLES

TABLE S1. Statisical summary of the performance of the initially generated databases on its test set used for training.

Subset	MAE(kcal/mol)	RMSE(kcal/mol)
1a	0.3918	0.6908
1b	0.4264	0.8564
2a	0.4636	0.7792
2b	0.5044	0.8868
2c	0.517	0.8725
3a	0.4379	0.6599
3b	0.4138	0.7649
4	0.5181	0.9275

TABLE S2. Number of samples added to the database as a percentage of the total number of samples used for training the different databases. Note: For the 25% of set 3b, only the number of converged molecules was used.

Dataset	1 %	5 %	10 %	25 %
1 and 2	250	1250	2500	6250
3a	87	435	870	2175
3b	206	1080	2060	5138*
4	125	625	1250	3125

II. SUPPLEMENTARY FIGURES

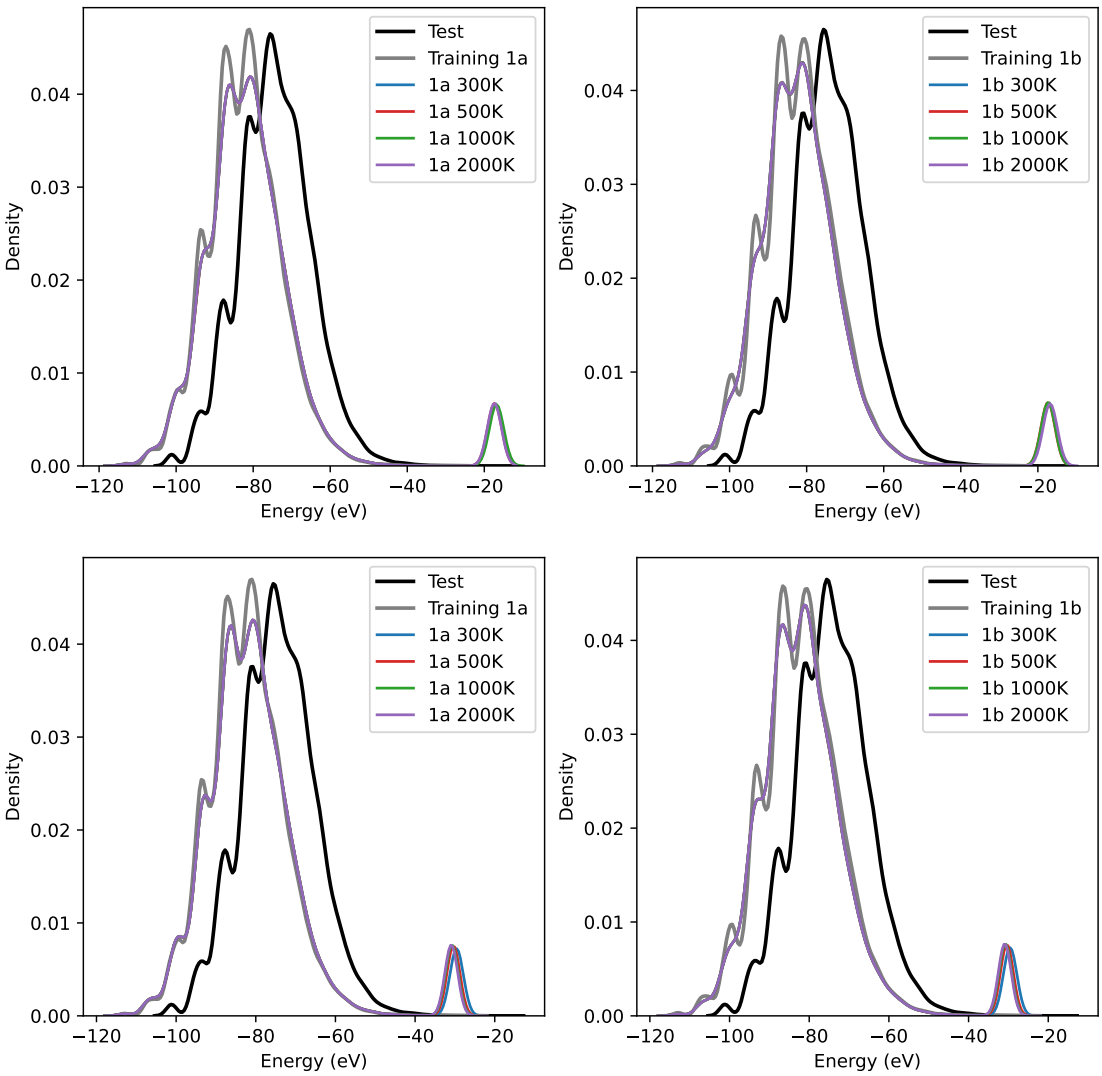

FIG. S1. Energy distribution for the testing, initial training dataset and the enhanced datasets by temperature for set 1.

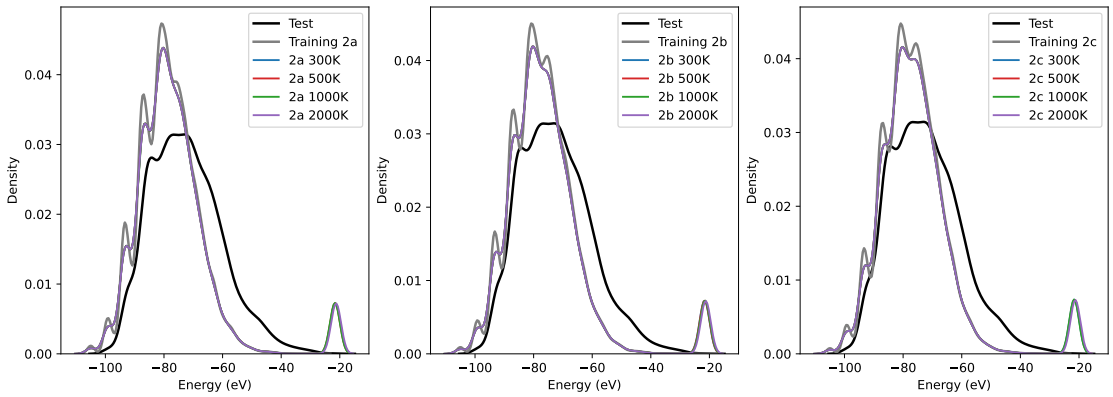

FIG. S2. Energy distribution for the testing, initial training dataset and the enhanced datasets by temperature for set 2.

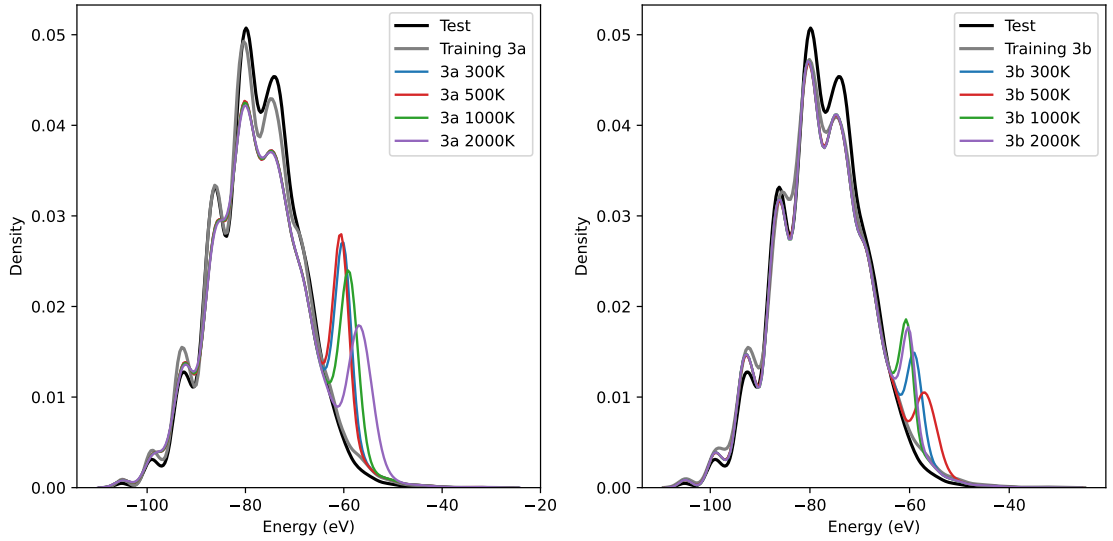

FIG. S3. Energy distribution for the testing, initial training dataset and the enhanced datasets by temperature for set 3.

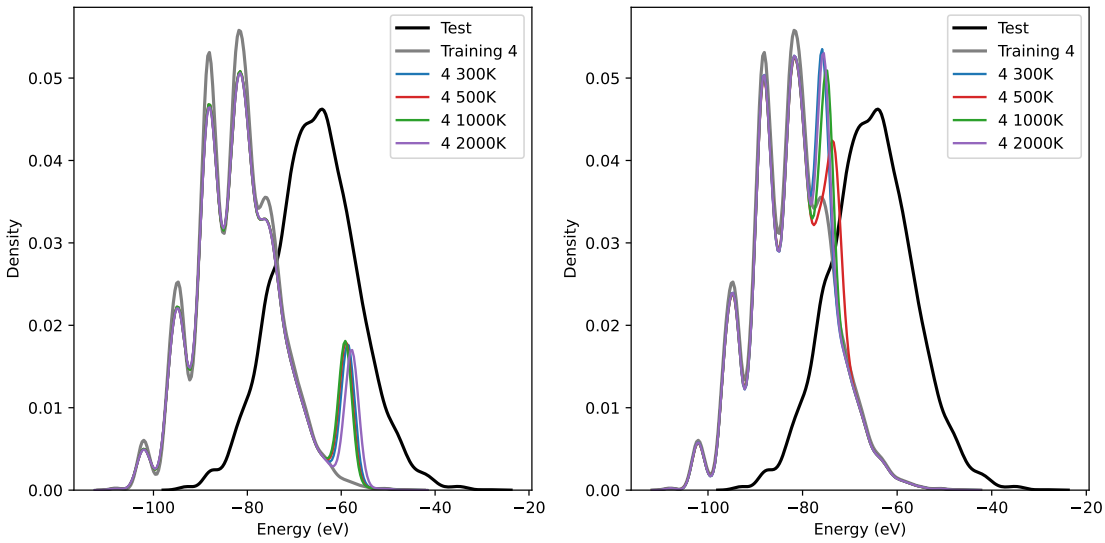

FIG. S4. Energy distribution for the testing, initial training dataset and the enhanced datasets by temperature for set 4.

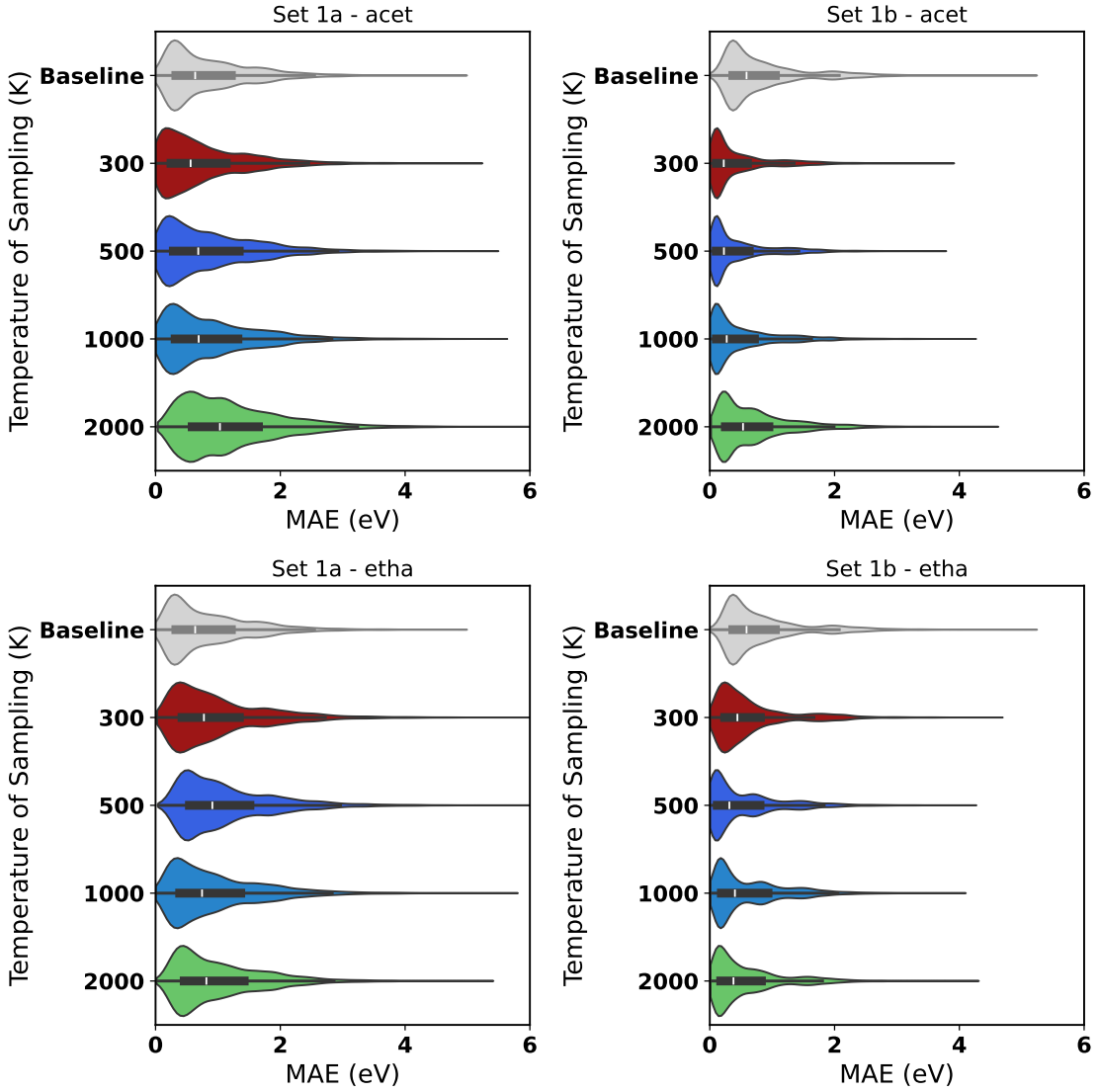

FIG. S5. Violin plot of the MAE for the datasets of set1 at different temperatures.

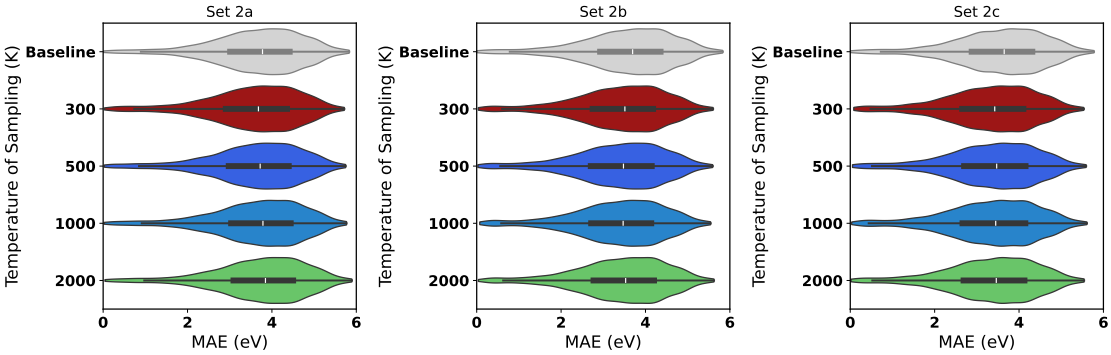

FIG. S6. Violin plot of the MAE for the datasets of set2 at different temperatures.

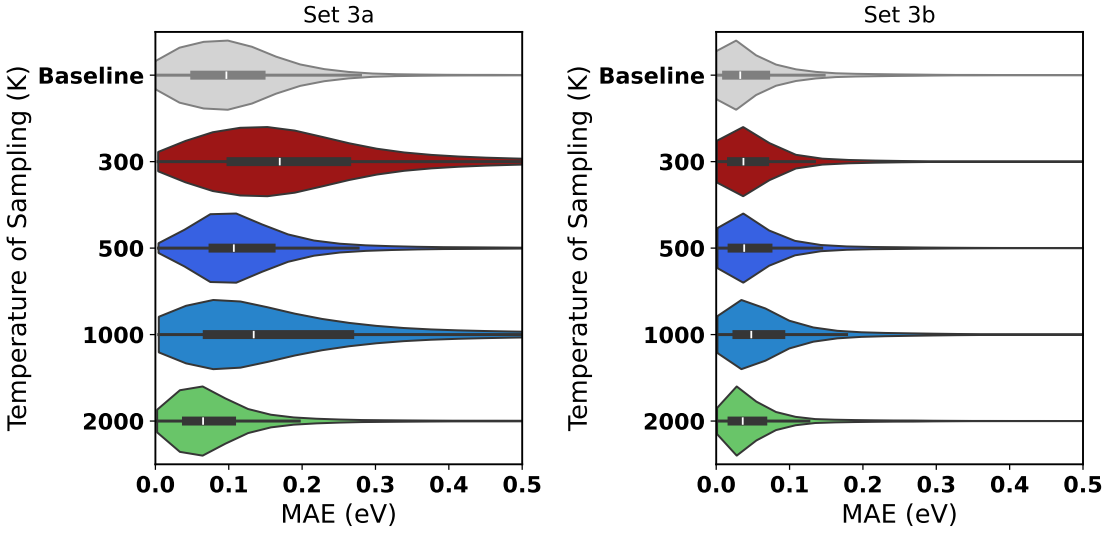

FIG. S7. Violin plot of the MAE for the datasets of set3 at different temperatures.

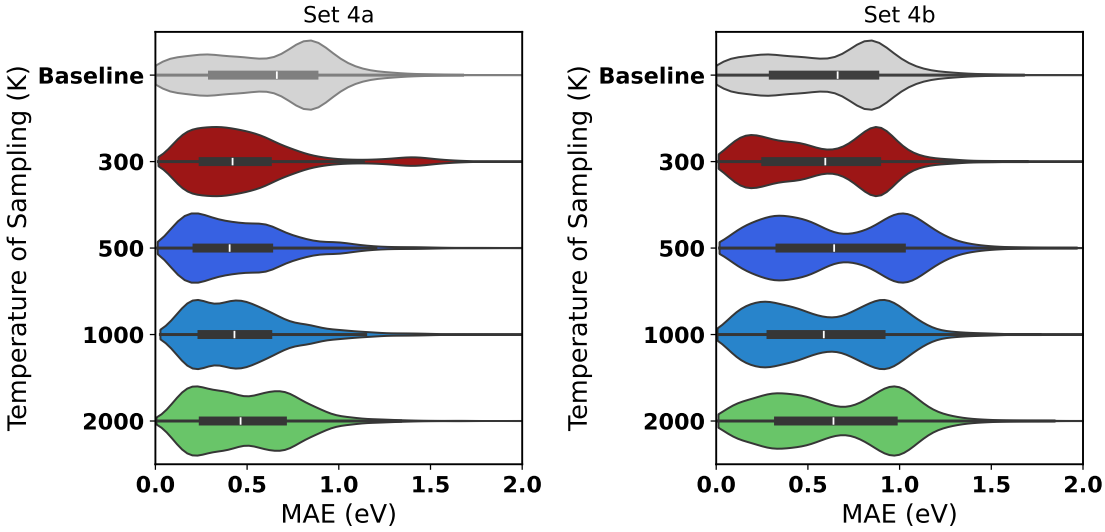

FIG. S8. Violin plot of the MAE for the datasets of set4 at different temperatures.

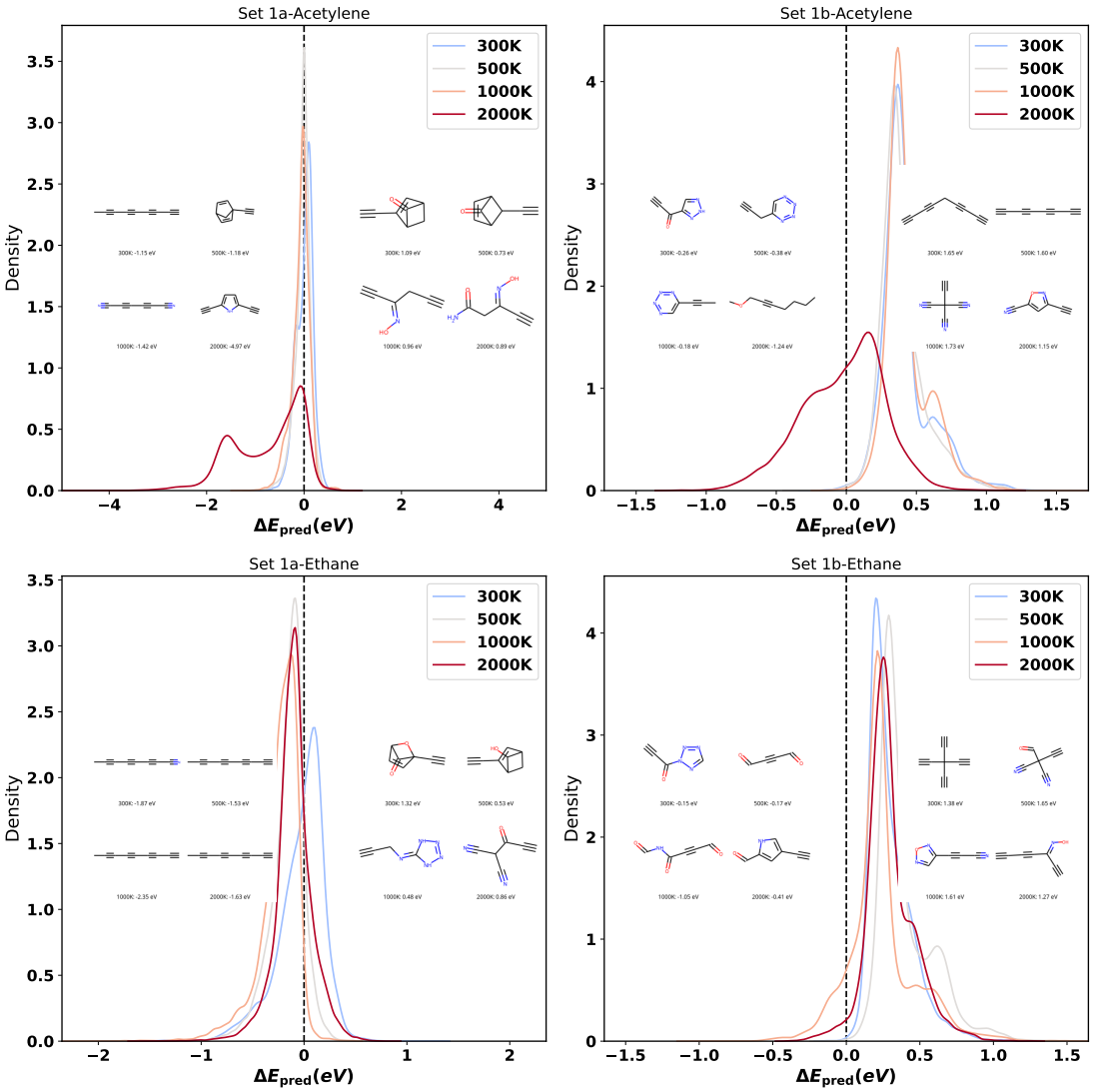

FIG. S9. Distribution of change in predicted energy to the temperature ($\Delta E = E_0 - E_T$, here $T \in \{300, 500, 1000, 2000\}$K) for the datasets of set1. Each panel shows the molecule with the largest decrease or increase in ΔE for the different temperatures

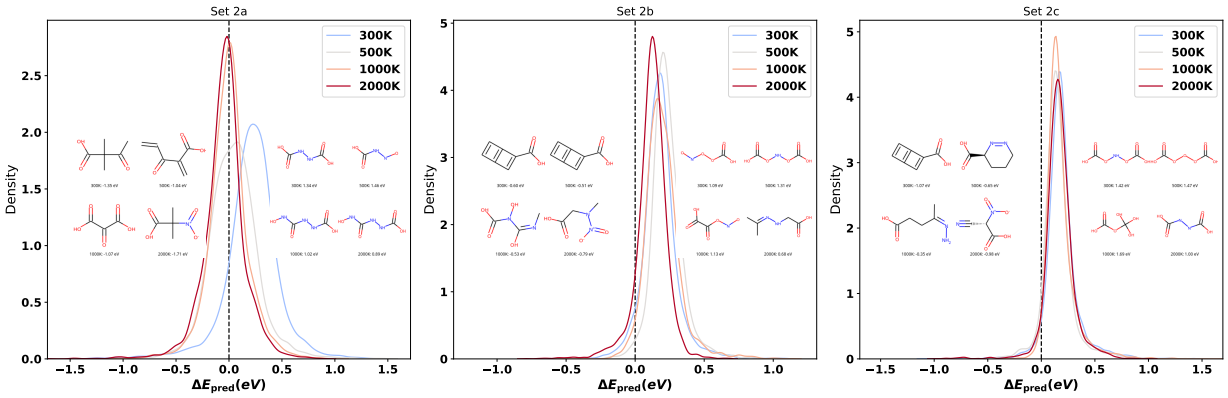

FIG. S10. Distribution of change in predicted energy to the temperature ($\Delta E = E_0 - E_T$, here $T \in \{300, 500, 1000, 2000\}$K) for the datasets of set2. Each panel shows the molecule with the largest decrease or increase in ΔE for the different temperatures.

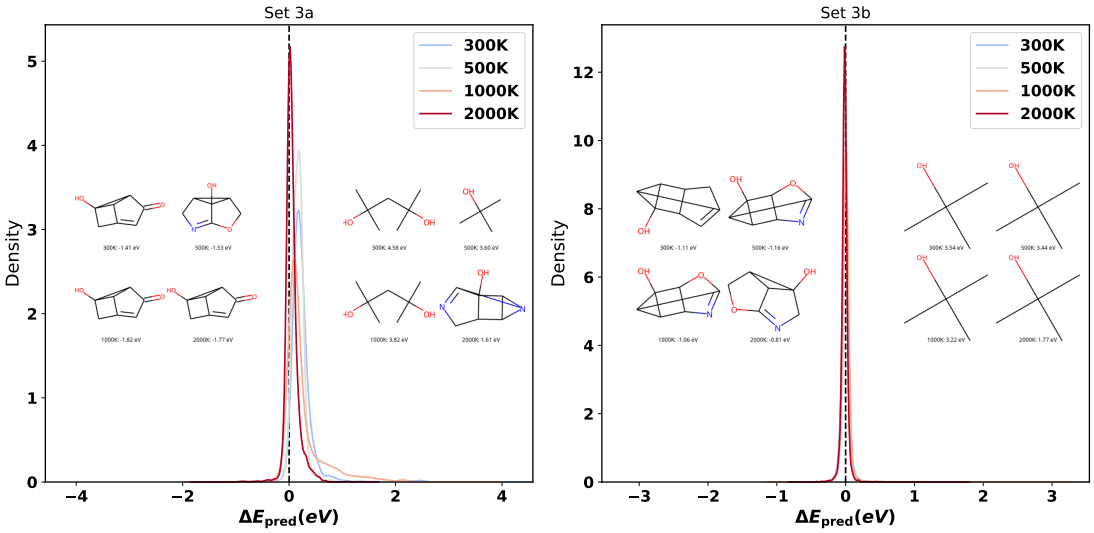

FIG. S11. Distribution of change in predicted energy to the temperature ($\Delta E = E_0 - E_T$, here $T \in \{300, 500, 1000, 2000\}$K) for the datasets of set3. Each panel shows the molecule with the largest decrease or increase in ΔE for the different temperatures.

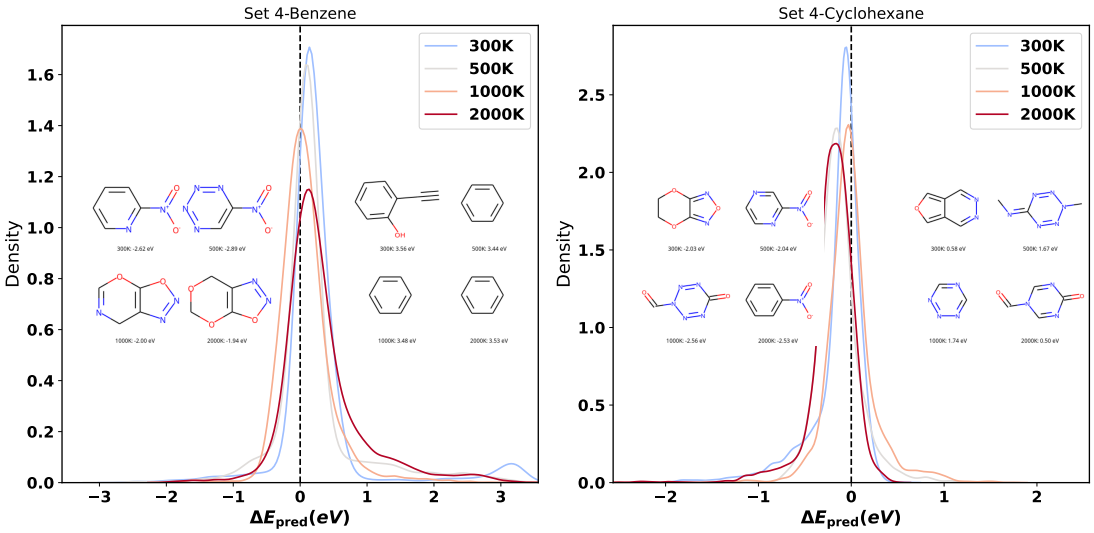

FIG. S12. Distribution of change in predicted energy to the temperature ($\Delta E = E_0 - E_T$, here $T \in \{300, 500, 1000, 2000\}$K) for the datasets of set4. Each panel shows the molecule with the largest decrease or increase in ΔE for the different temperatures.

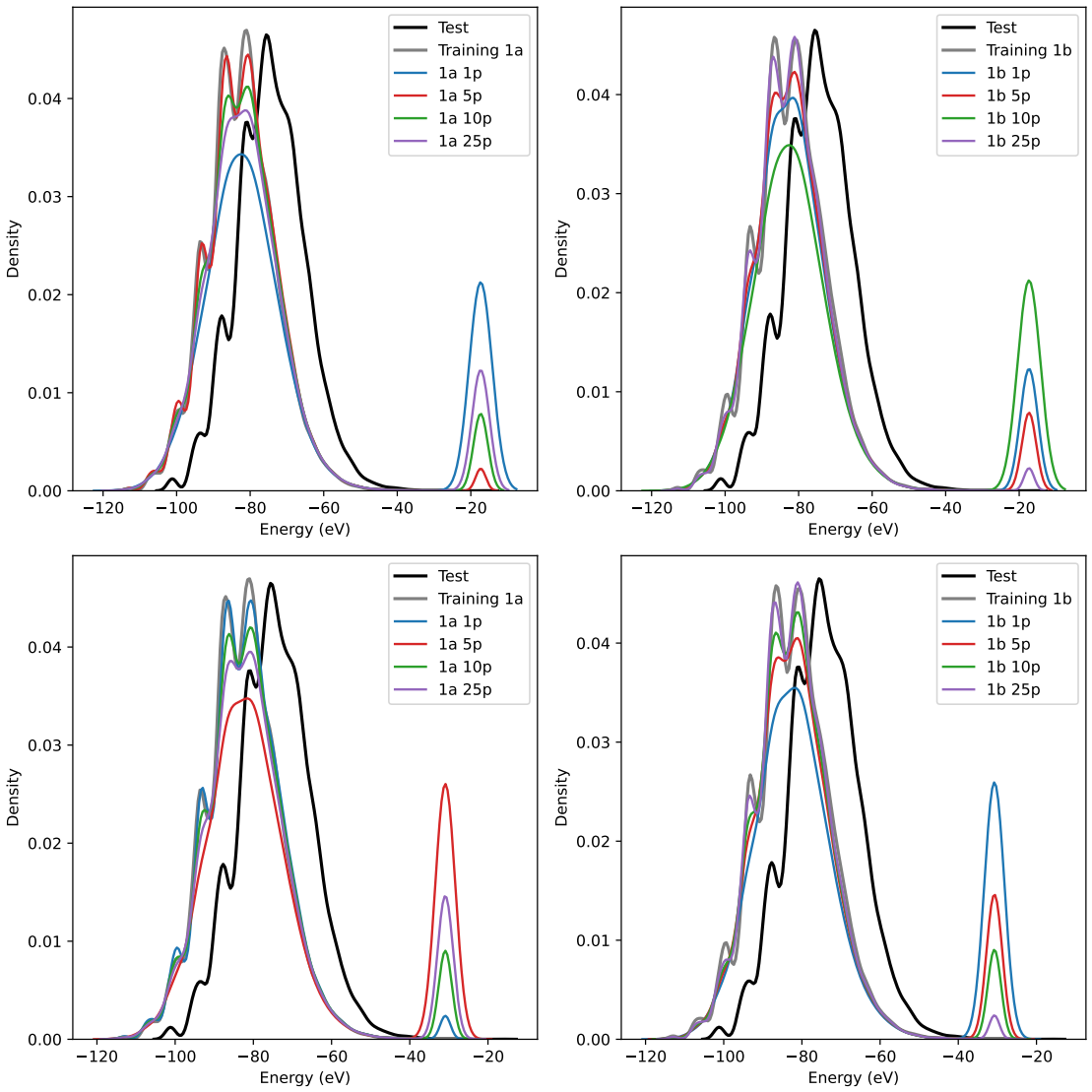

FIG. S13. Energy distribution for the testing, initial training dataset and the enhanced datasets by different percentages of added molecules for set 1.

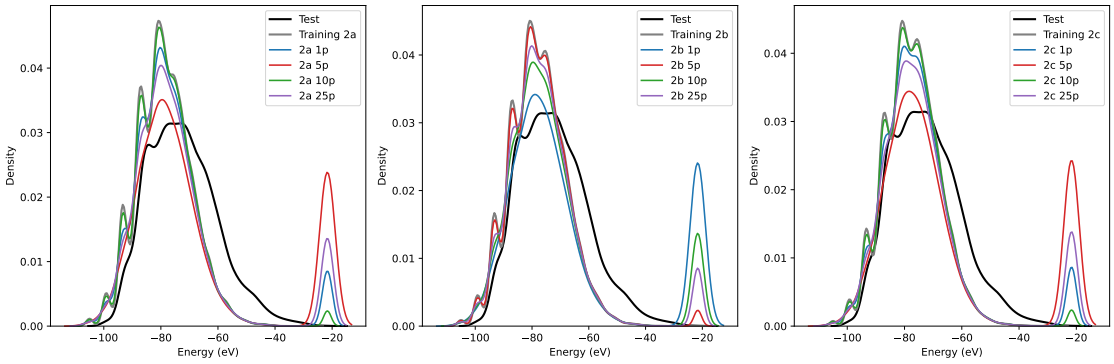

FIG. S14. Energy distribution for the testing, initial training dataset and the enhanced datasets by different percentages of added molecules for set 2.

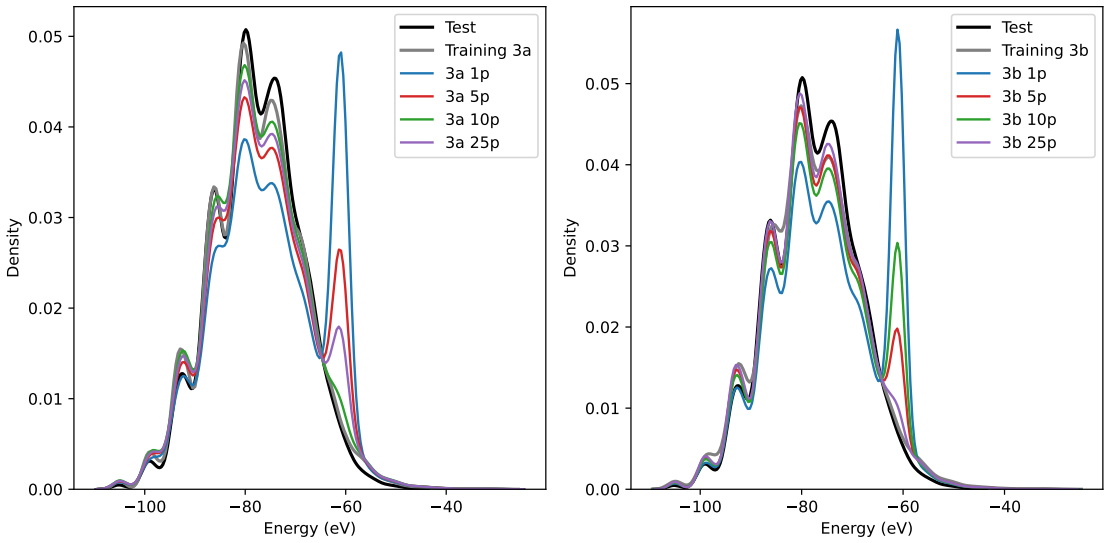

FIG. S15. Energy distribution for the testing, initial training dataset and the enhanced datasets by different percentages of added molecules for set 3.

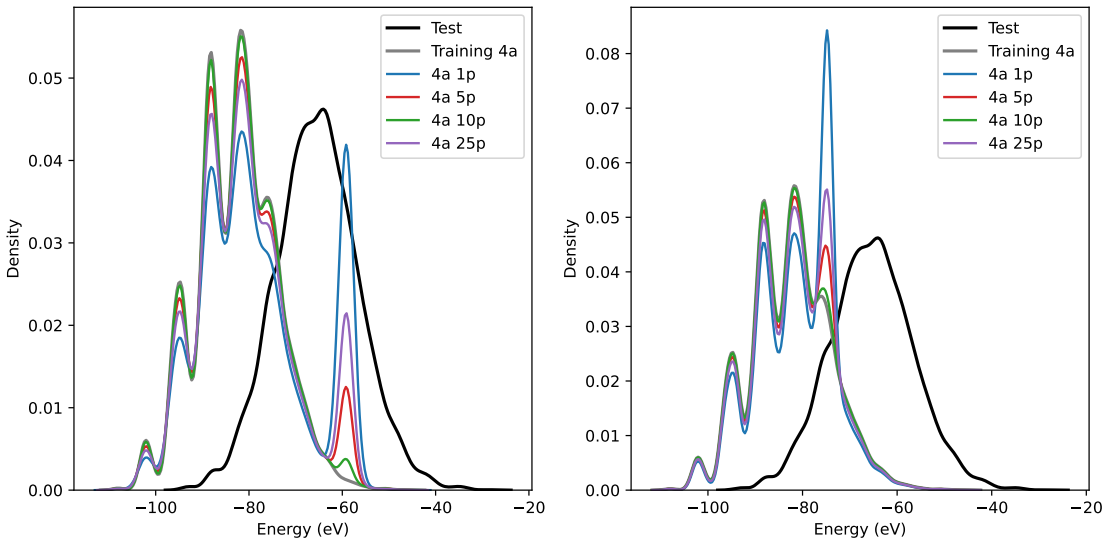

FIG. S16. Energy distribution for the testing, initial training dataset and the enhanced datasets by different percentages of added molecules for set 4.

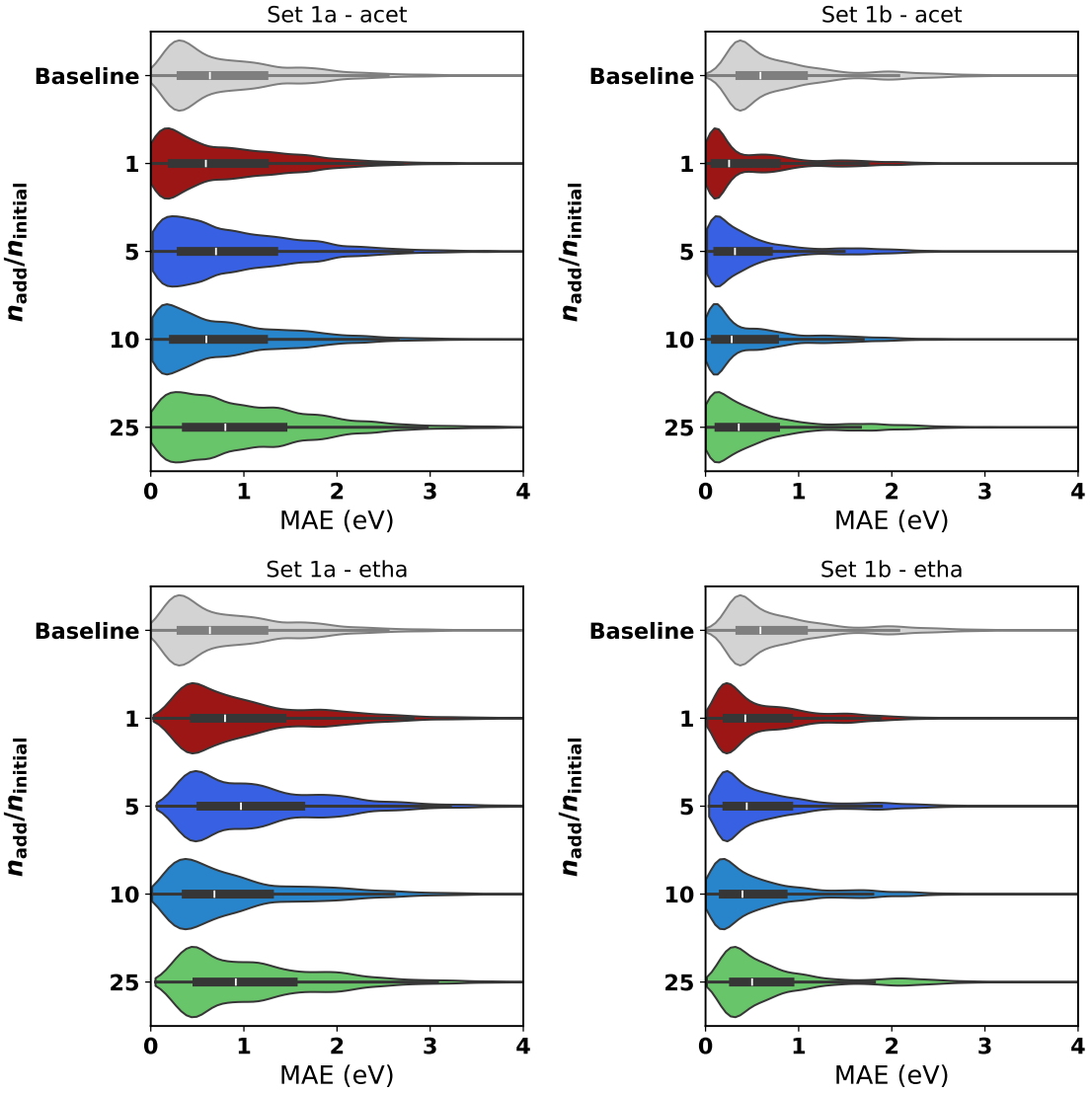

FIG. S17. Violin plot of the MAE for the datasets of set1 for different fractions of added molecules.

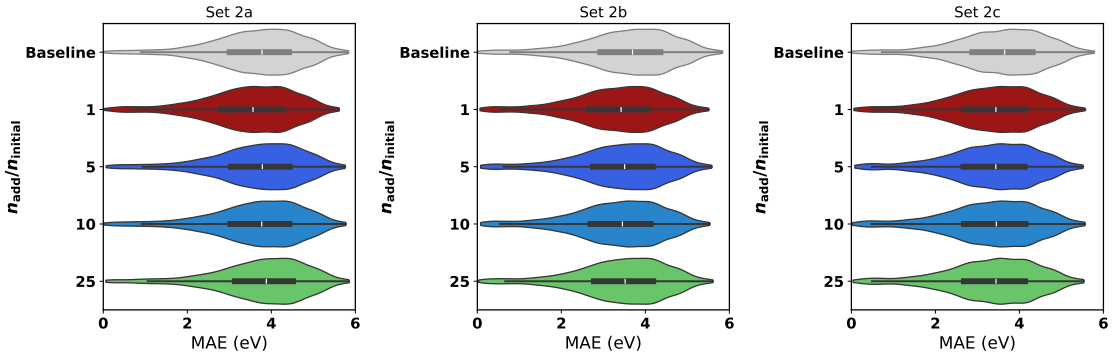

FIG. S18. Violin plot of the MAE for the datasets of set2 for different fractions of added molecules.

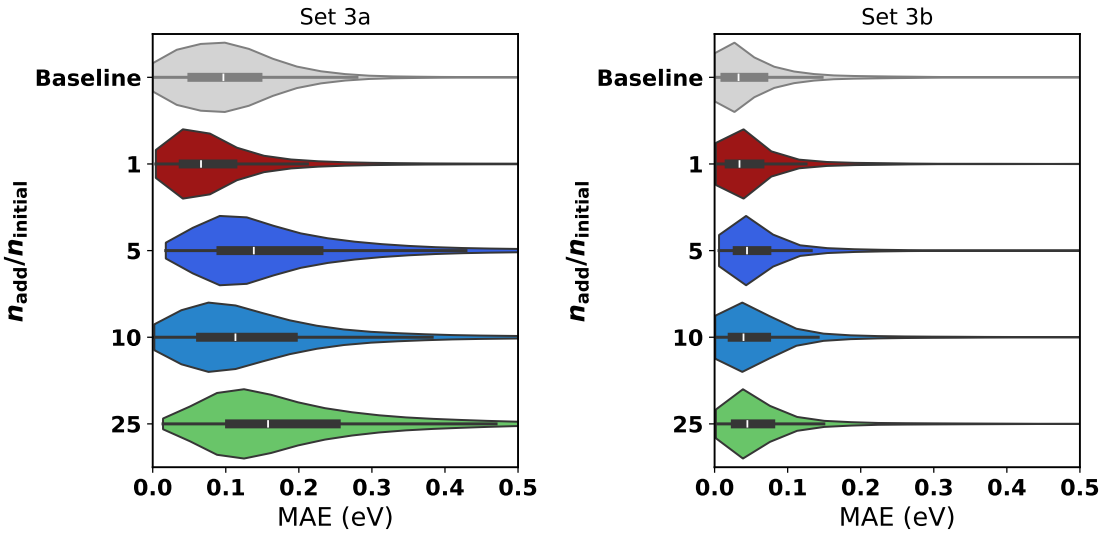

FIG. S19. Violin plot of the MAE for the datasets of set3 for different fractions of added molecules.

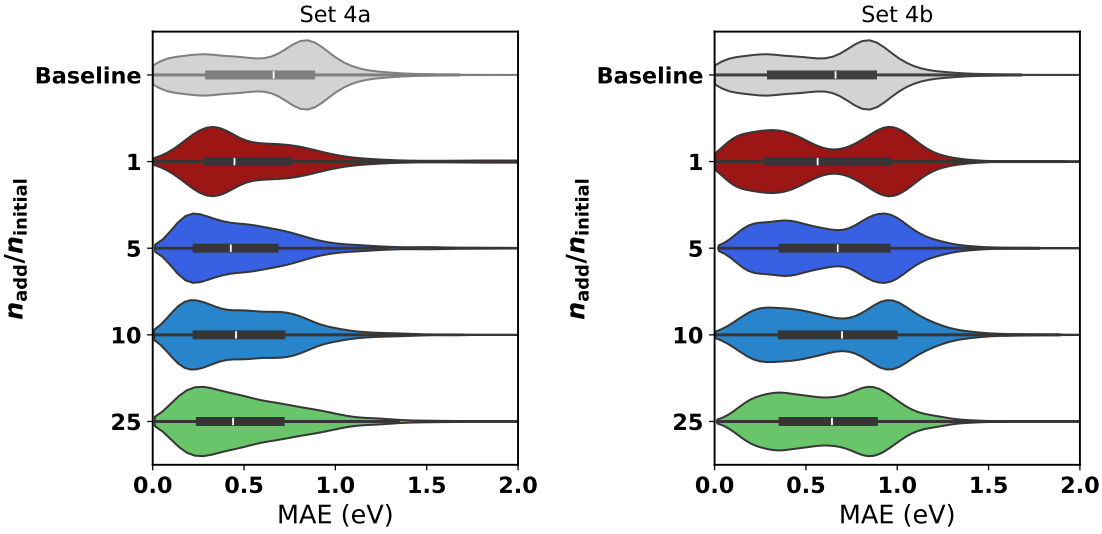

FIG. S20. Violin plot of the MAE for the datasets of set4 for different fractions of added molecules.

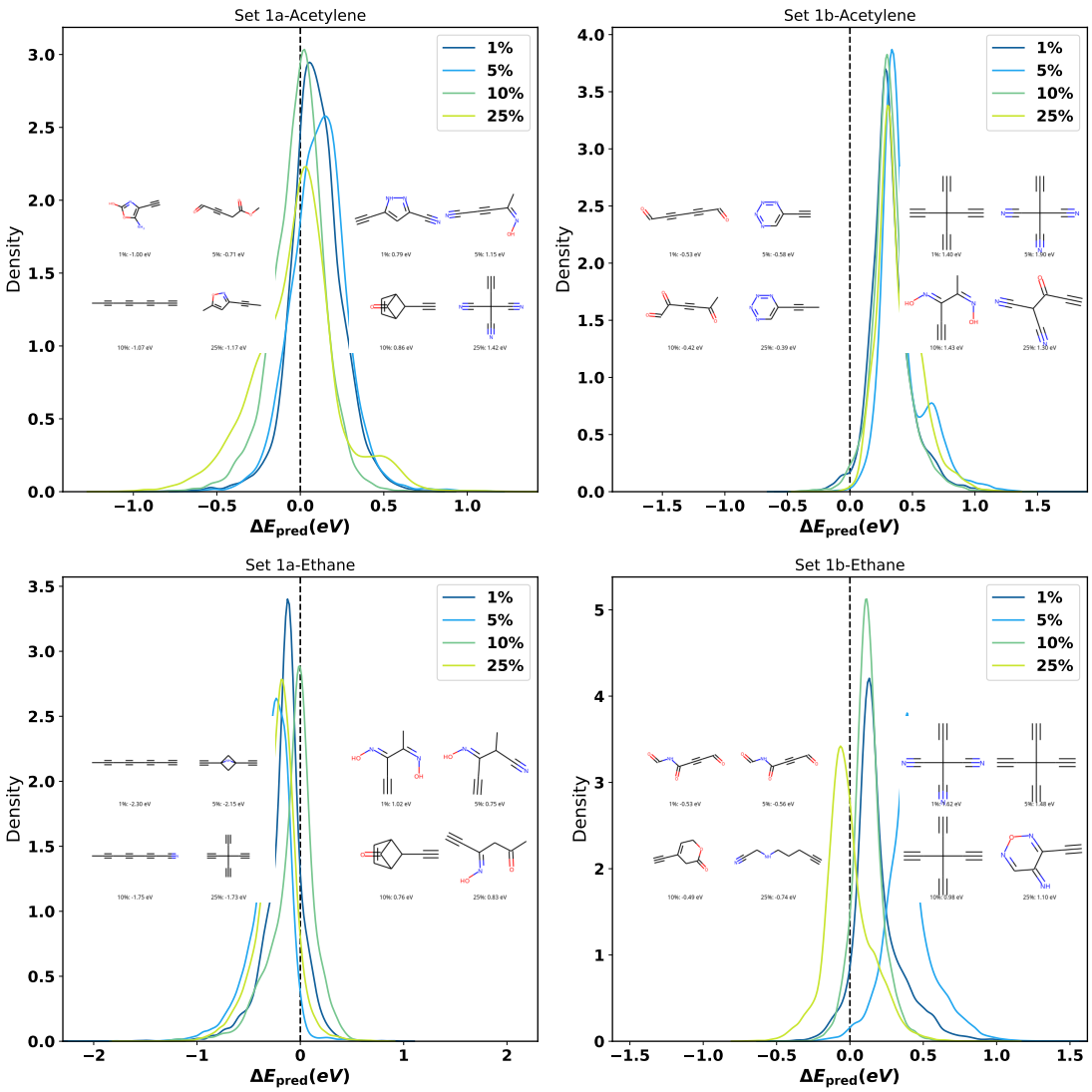

FIG. S21. Distribution of change in predicted energy to the percentages of samples added ($\Delta E = E_0 - E_i$, here $i \in \{1, 5, 10, 25\}\%$) for the datasets of set1. Each panel shows the molecule with the largest decrease or increase in ΔE for the different percentages.

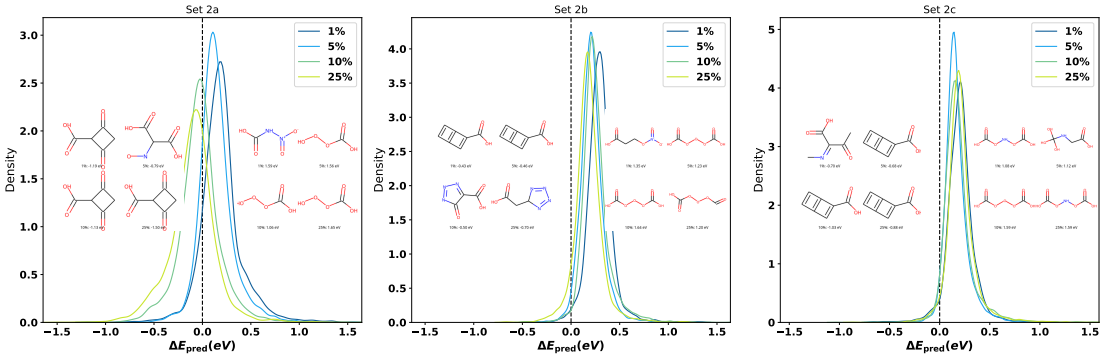

FIG. S22. Distribution of change in predicted energy to the percentages of samples added ($\Delta E = E_0 - E_i$, here $i \in \{1, 5, 10, 25\}\%$) for the datasets of set2. Each panel shows the molecule with the largest decrease or increase in ΔE for the different percentages.

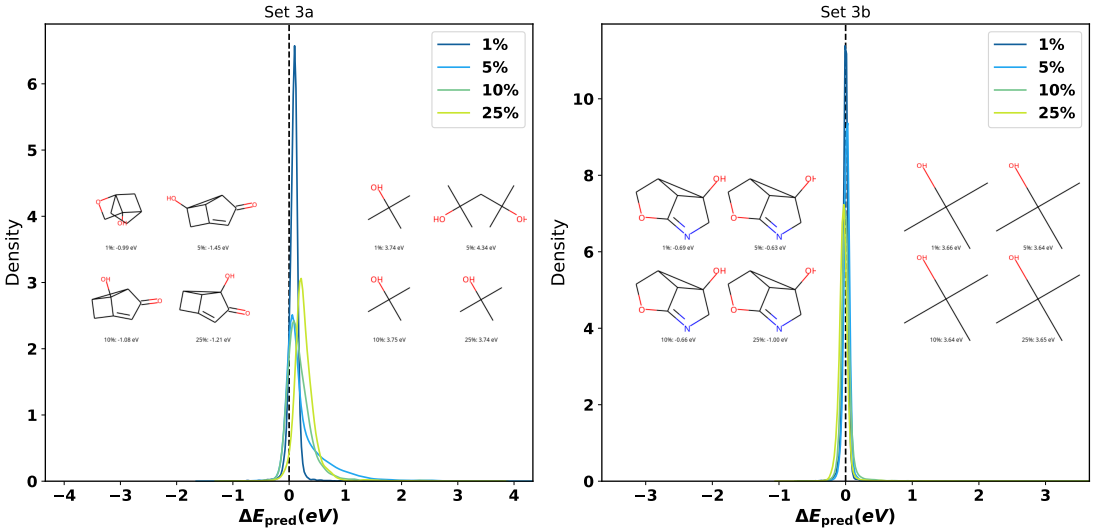

FIG. S23. Distribution of change in predicted energy to the percentages of samples ($\Delta E = E_0 - E_i$, here $i \in \{1, 5, 10, 25\}\%$) for the datasets of set3. Each panel shows the molecule with the largest decrease or increase in ΔE for the different percentages.

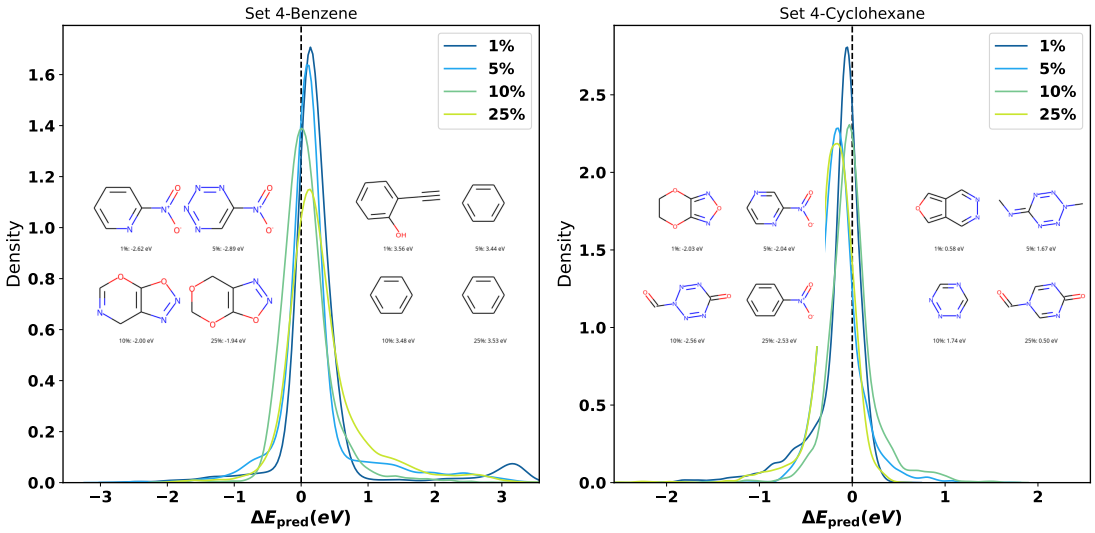

FIG. S24. Distribution of change in predicted energy to the percentages of samples ($\Delta E = E_0 - E_i$, here $i \in \{1, 5, 10, 25\}\%$) for the datasets of set4. Each panel shows the molecule with the largest decrease or increase in ΔE for the different percentages.

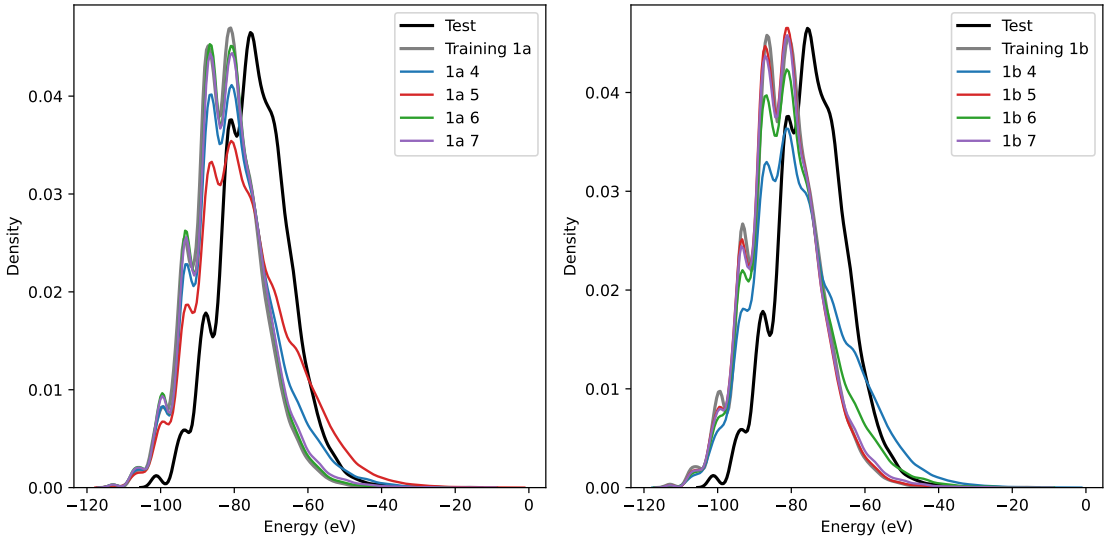

FIG. S25. Energy distribution for the testing, initial training dataset and the enhanced datasets by different amon size for set 1.

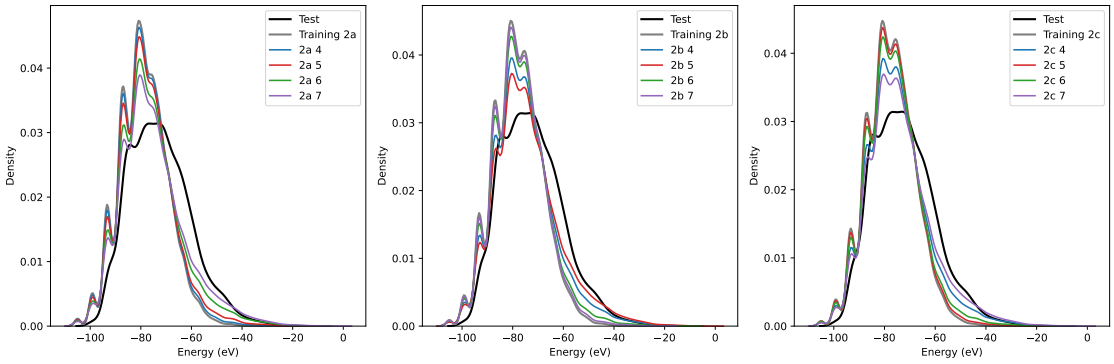

FIG. S26. Energy distribution for the testing, initial training dataset and the enhanced datasets by different amon size for set 2.

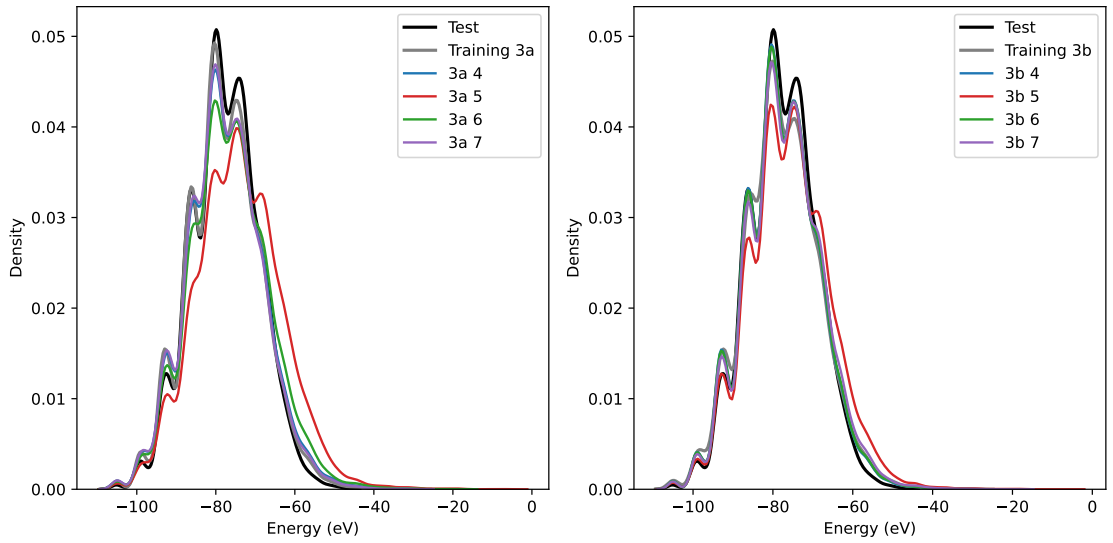

FIG. S27. Energy distribution for the testing, initial training dataset and the enhanced datasets by different amon size for set 3.

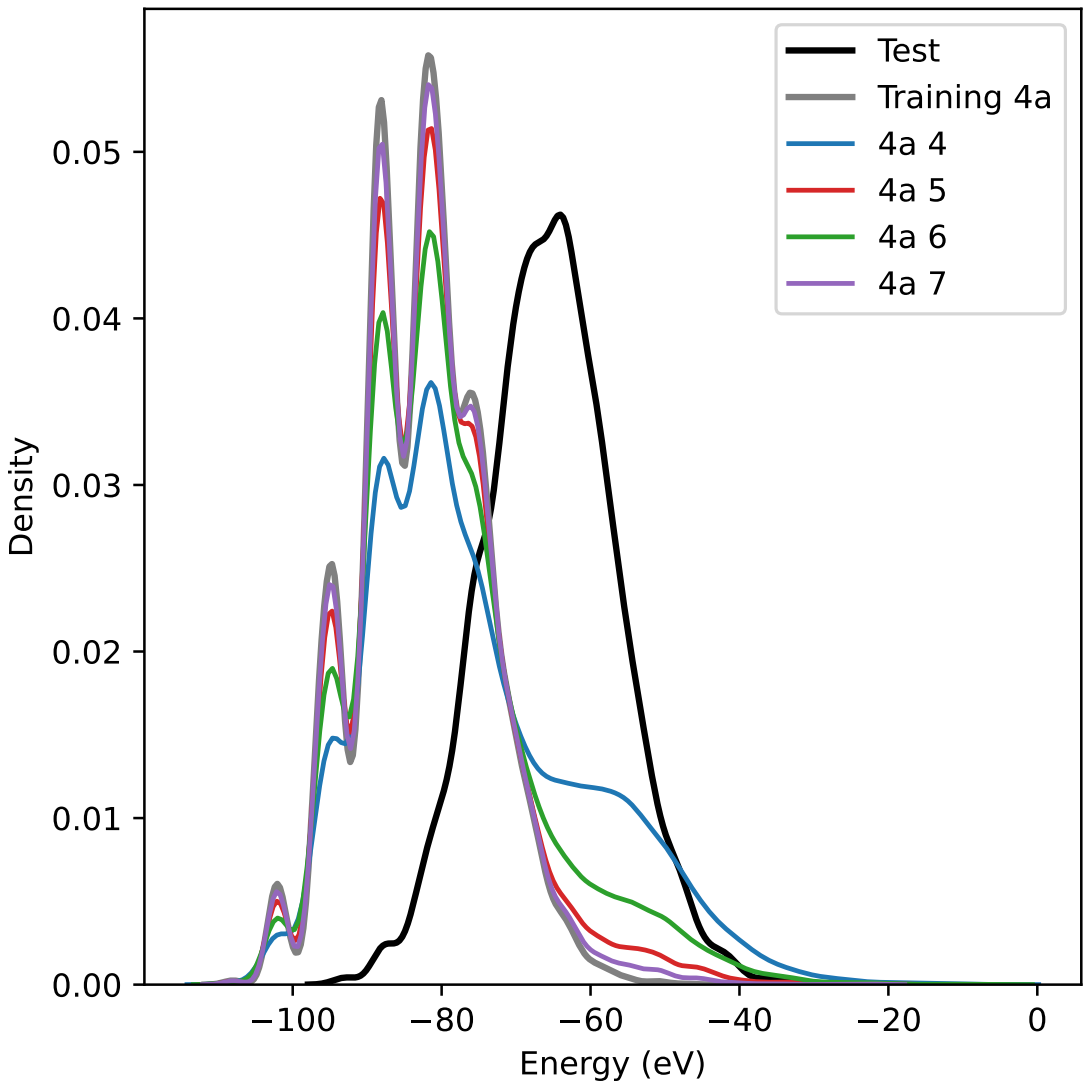

FIG. S28. Energy distribution for the testing, initial training dataset and the enhanced datasets by different amon size for set 4.

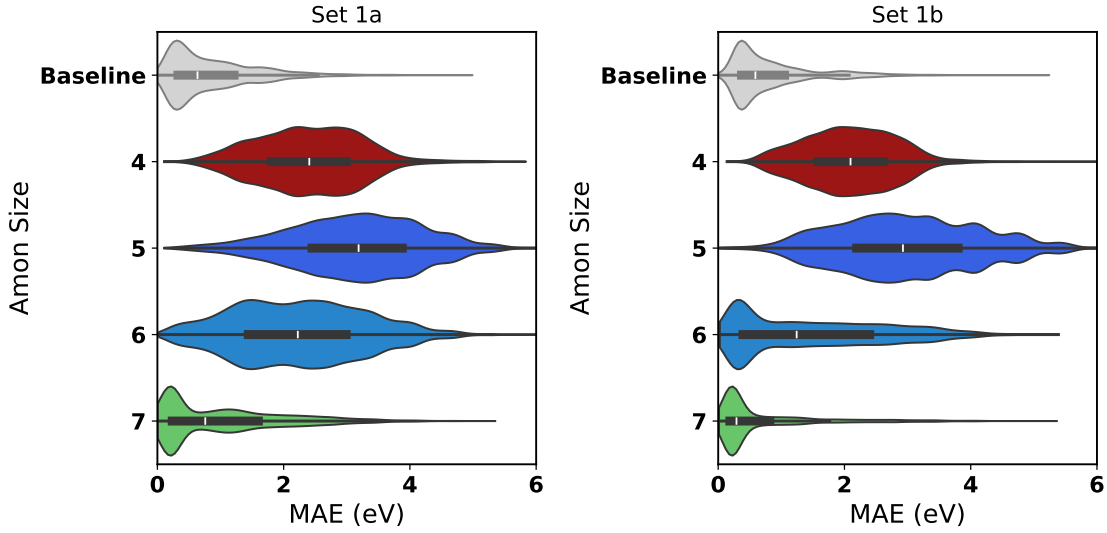

FIG. S29. Violin plot of the MAE for the datasets of set1 for different amon size.

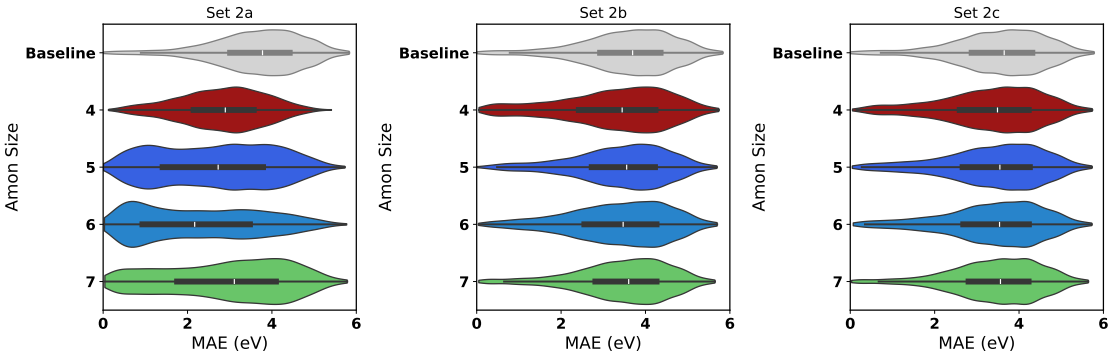

FIG. S30. Violin plot of the MAE for the datasets of set2 for different amon size.

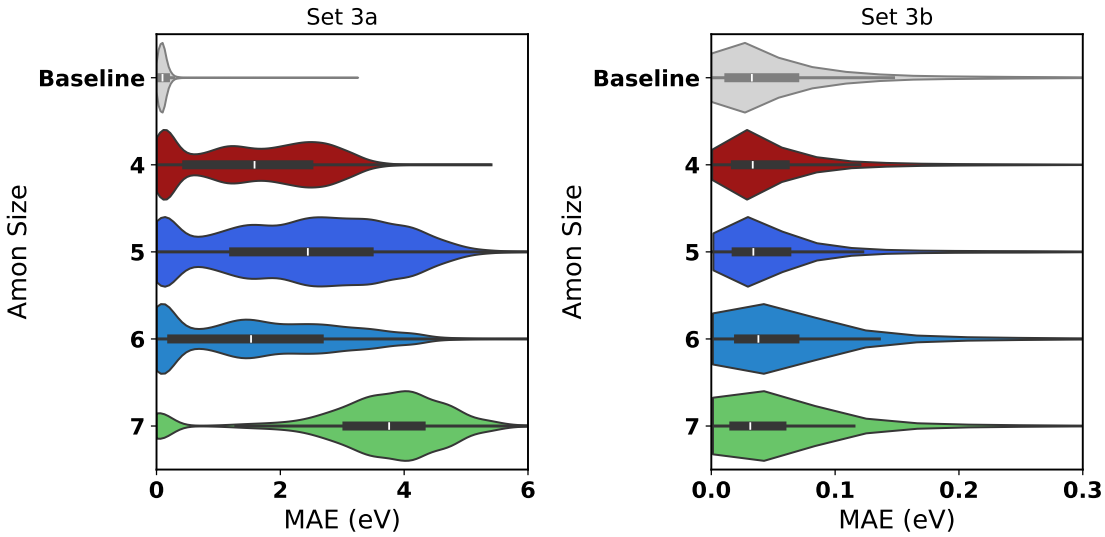

FIG. S31. Violin plot of the MAE for the datasets of set3 for different amon size.

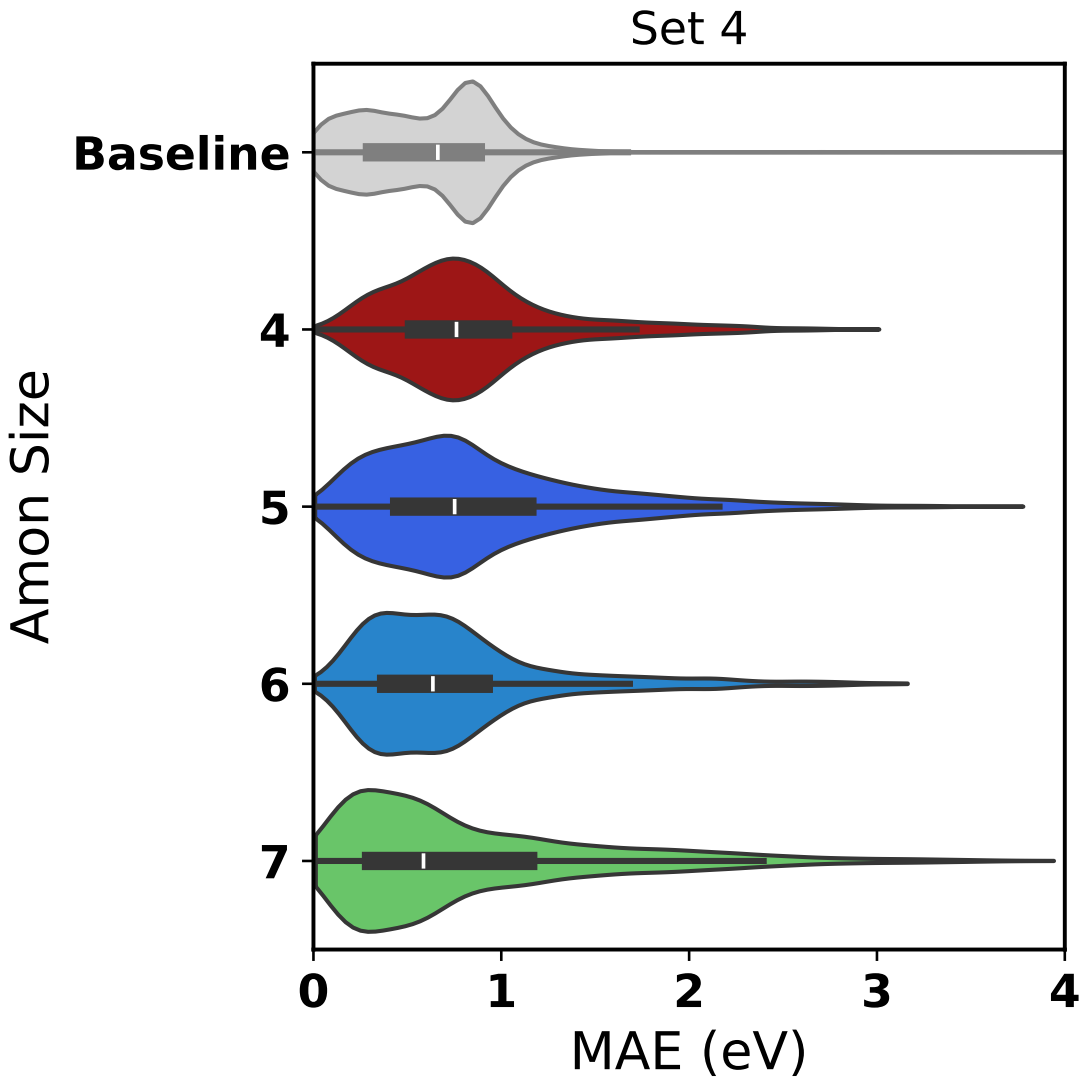

FIG. S32. Violin plot of the MAE for the datasets of set4 for different amon size.

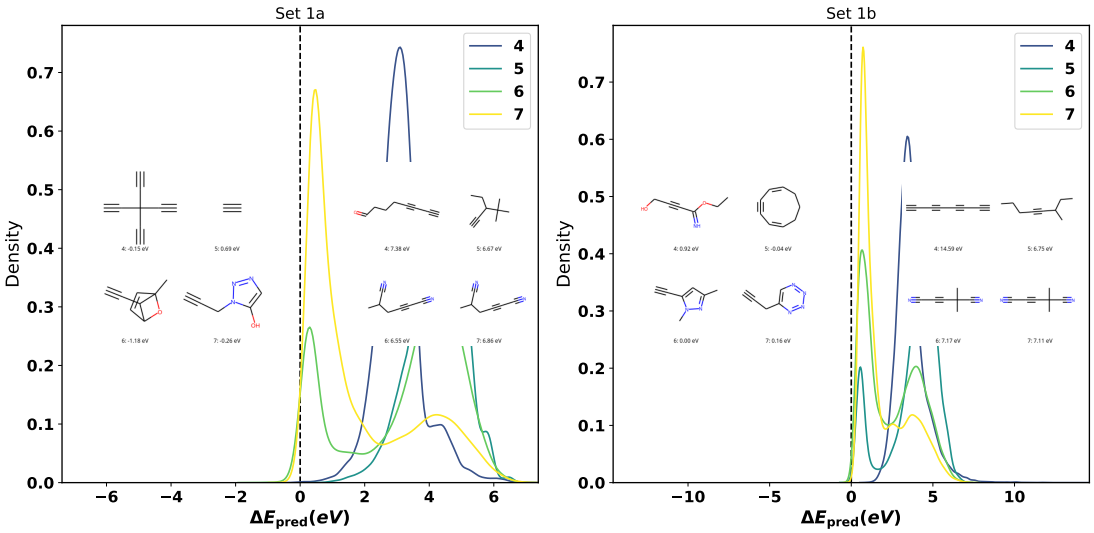

FIG. S33. Distribution of change in predicted energy to the size of the amon size added ($\Delta E = E_0 - E_i$, here $i \in \{4, 5, 6, 7\}$) for the datasets of set1. Each panel shows the molecule with the largest decrease or increase in ΔE for the different percentages.

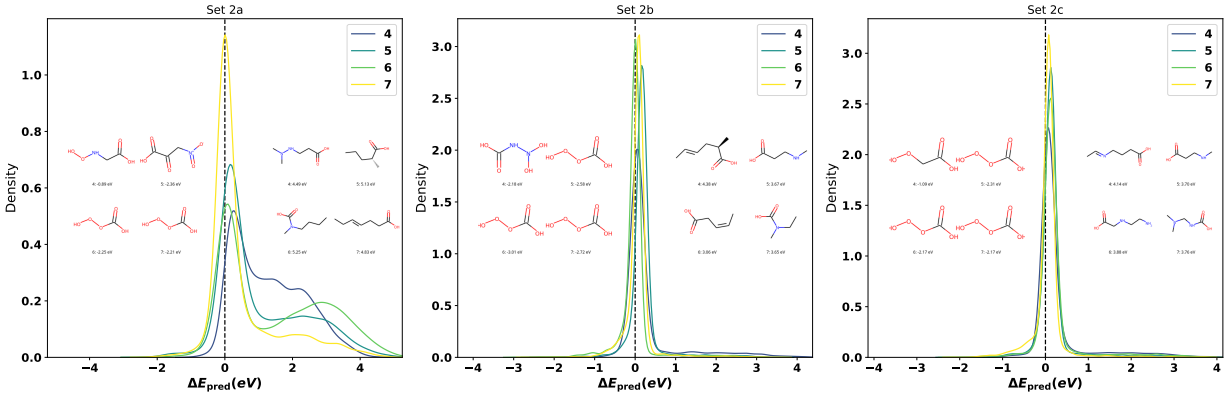

FIG. S34. Distribution of change in predicted energy to the size of the amon size added ($\Delta E = E_0 - E_i$, here $i \in \{4, 5, 6, 7\}$) for the datasets of set2. Each panel shows the molecule with the largest decrease or increase in ΔE for the different percentages.

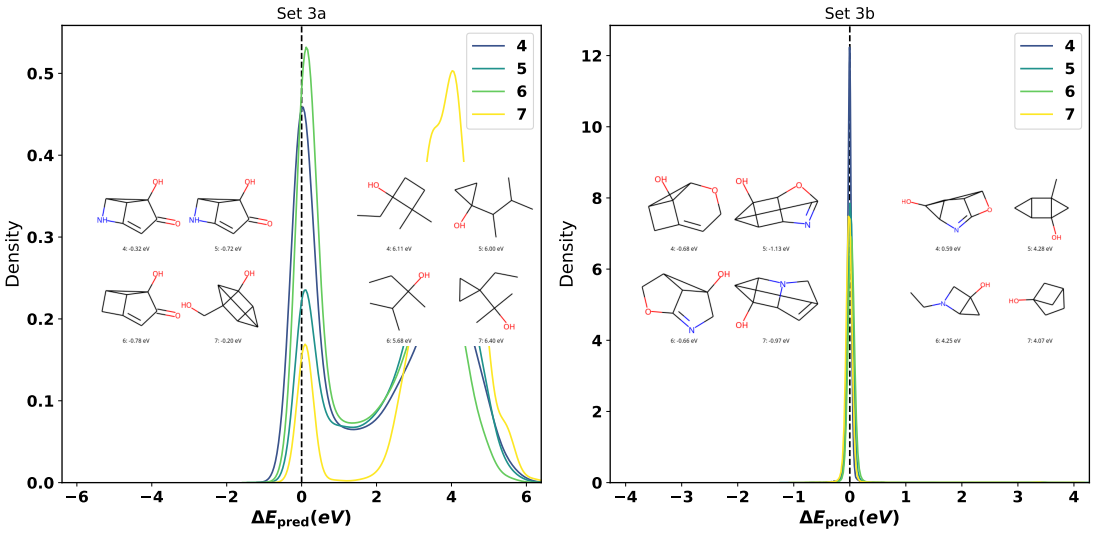

FIG. S35. Distribution of change in predicted energy to the size of the amon size added ($\Delta E = E_0 - E_i$, here $i \in \{4, 5, 6, 7\}$) for the datasets of set3. Each panel shows the molecule with the largest decrease or increase in ΔE for the different percentages.

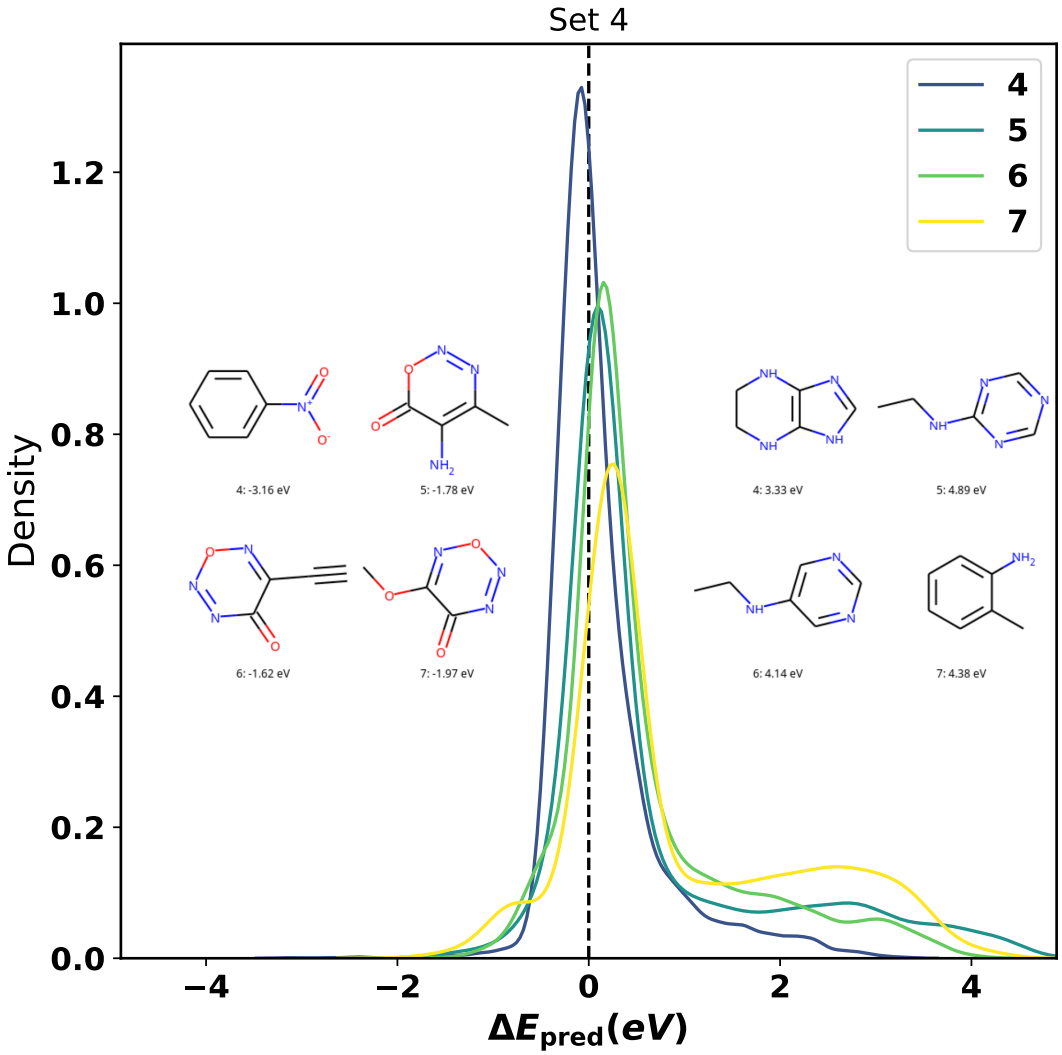

FIG. S36. Distribution of change in predicted energy to the size of the amon size added ($\Delta E = E_0 - E_i$, here $i \in \{4, 5, 6, 7\}$) for the datasets of set4. Each panel shows the molecule with the largest decrease or increase in ΔE for the different percentages.

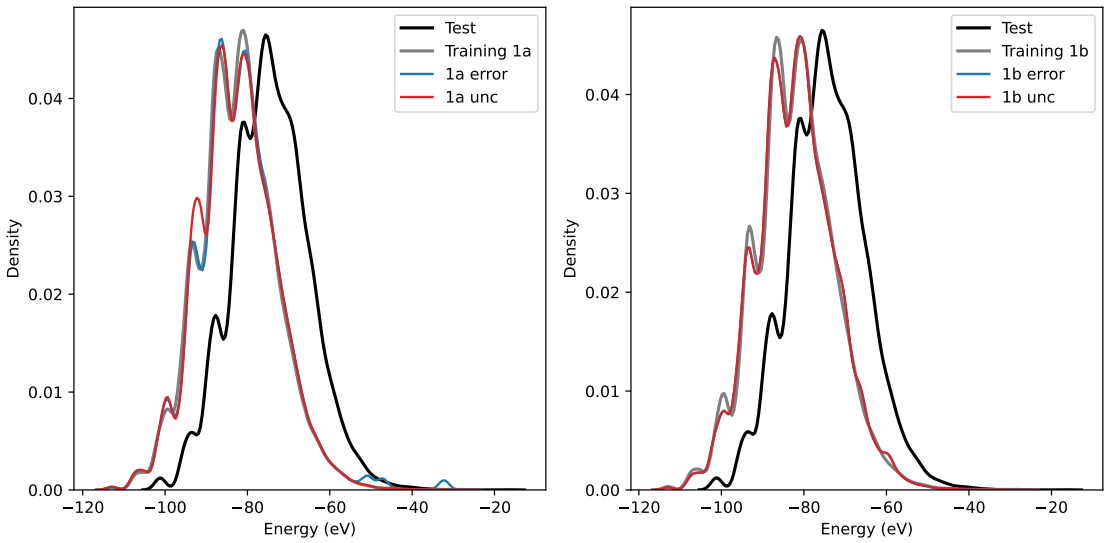

FIG. S37. Energy distribution for the testing, initial training dataset and the enhanced datasets by different method of addition for set 1.

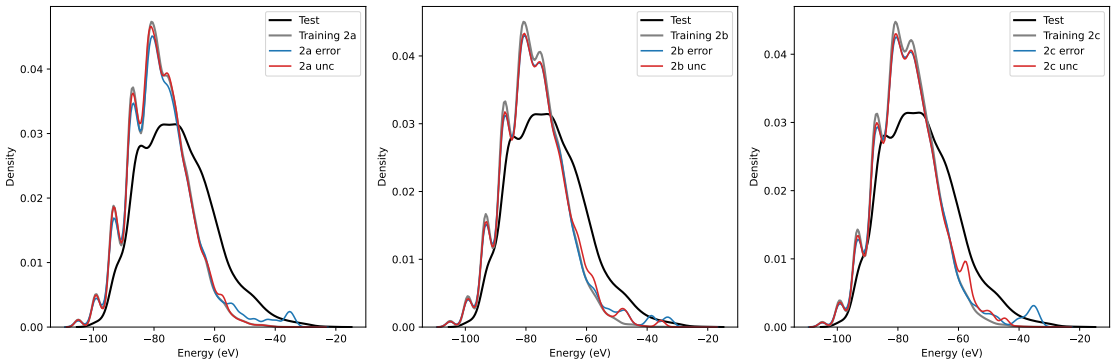

FIG. S38. Energy distribution for the testing, initial training dataset and the enhanced datasets by different method of addition for set 2.

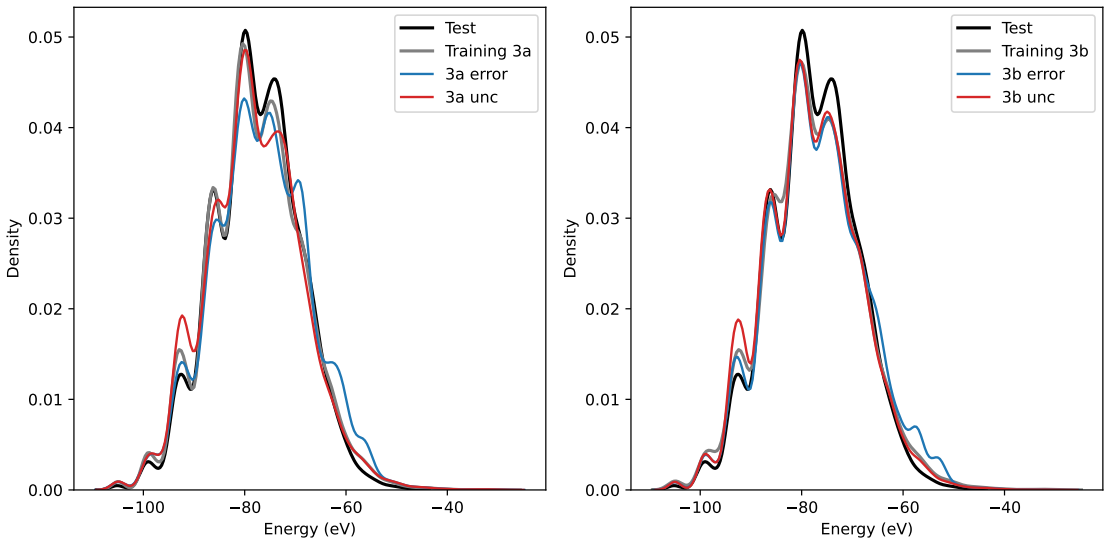

FIG. S39. Energy distribution for the testing, initial training dataset and the enhanced datasets by different method of addition for set 3.

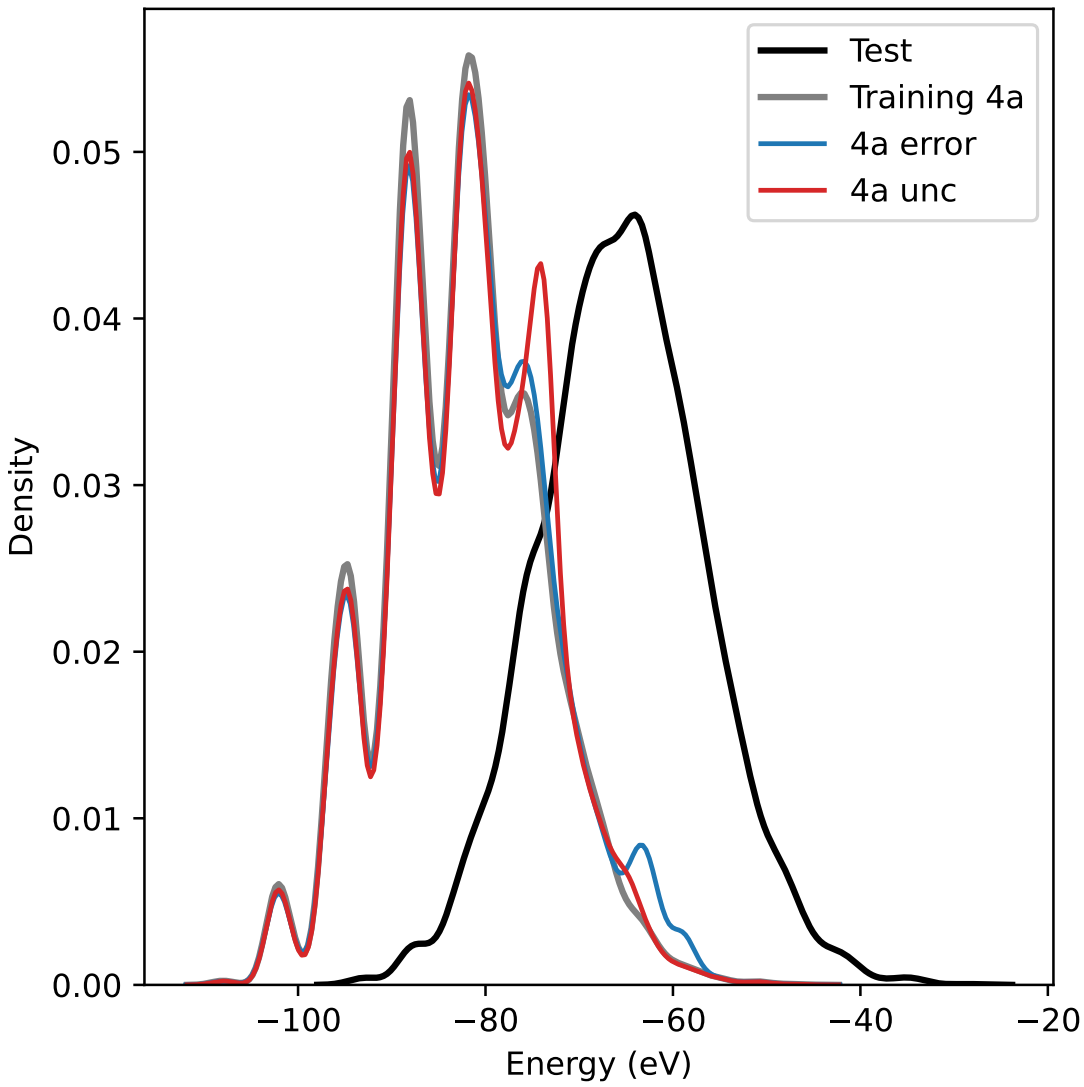

FIG. S40. Energy distribution for the testing, initial training dataset and the enhanced datasets by different method of addition for set 4.

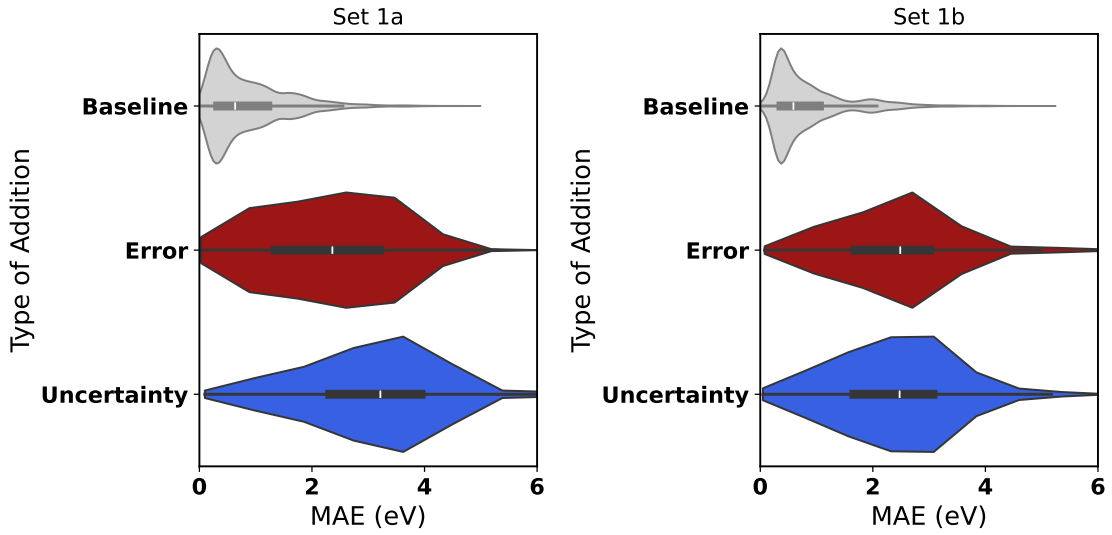

FIG. S41. Violin plot of the MAE for the datasets of set1 for the addition based on uncertainty or error.

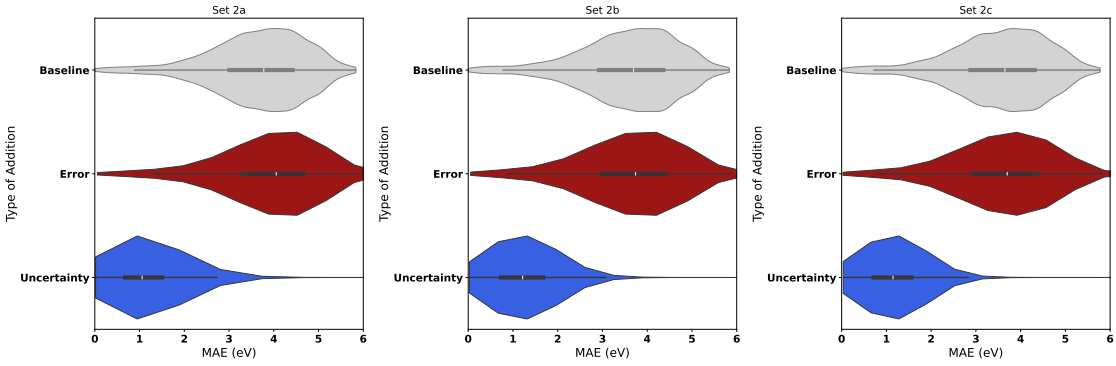

FIG. S42. Violin plot of the MAE for the datasets of set2 for the addition based on uncertainty or error.

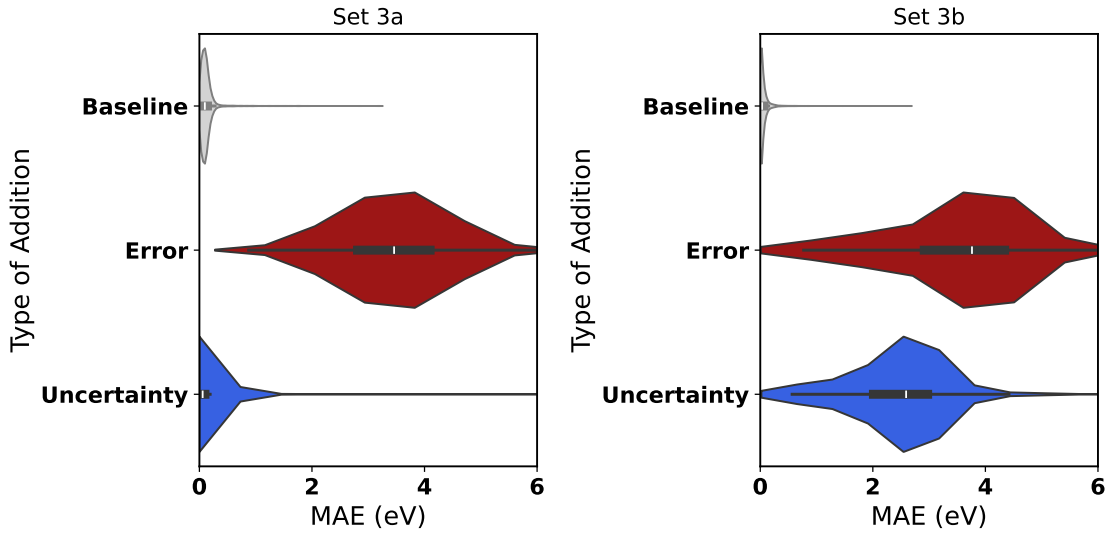

FIG. S43. Violin plot of the MAE for the datasets of set3 for the addition based on uncertainty or error.

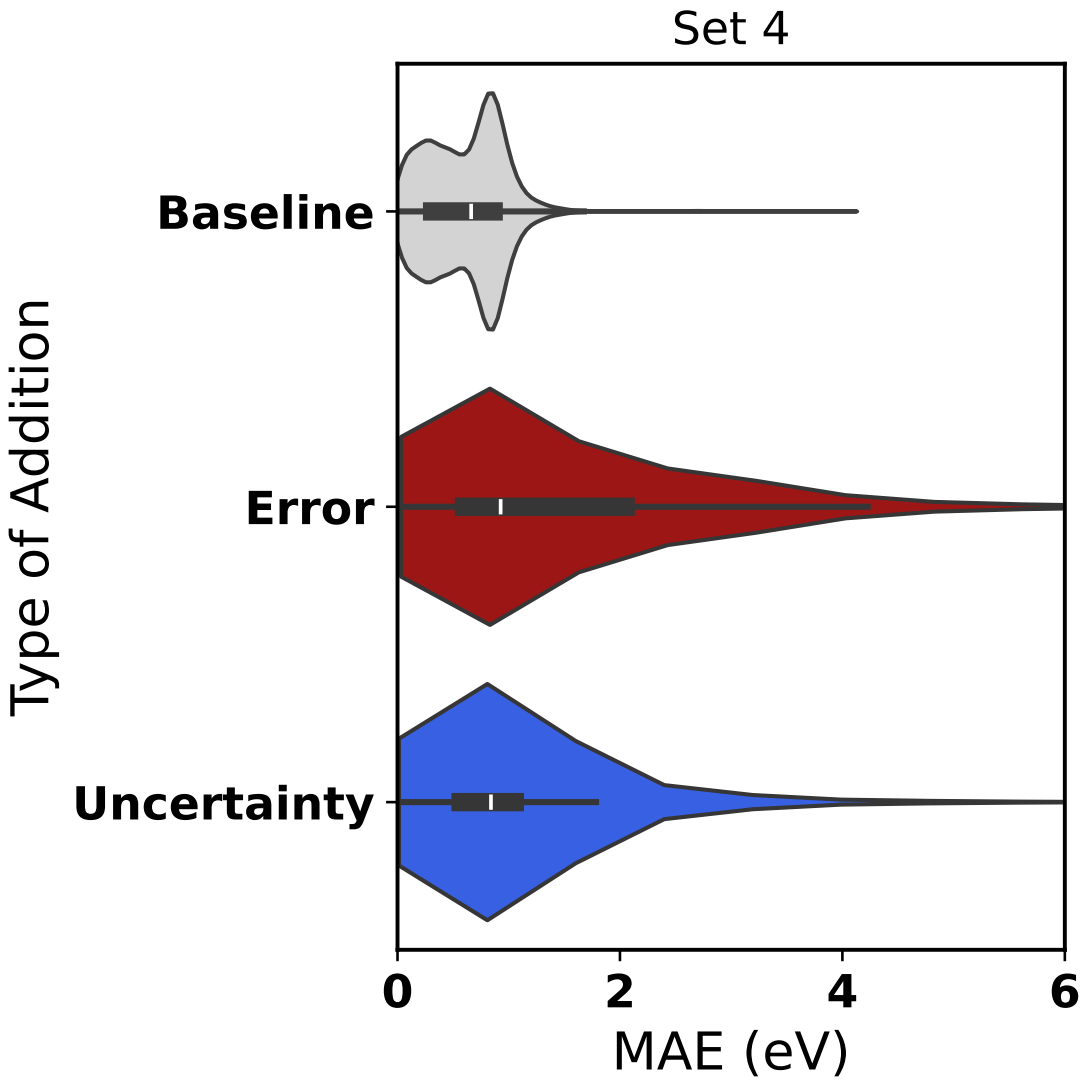

FIG. S44. Violin plot of the MAE for the datasets of set4 for the addition based on uncertainty or error.

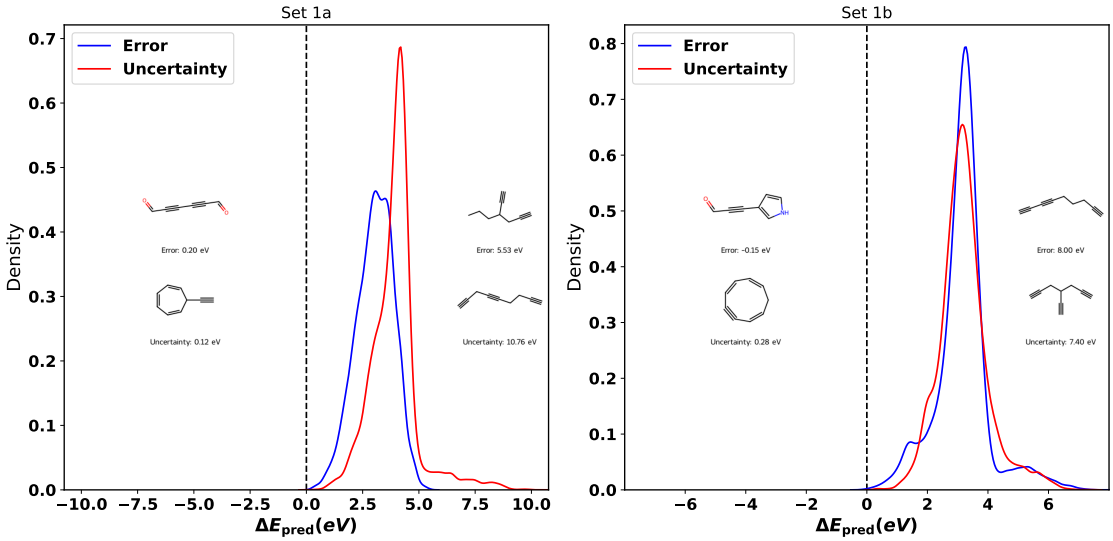

FIG. S45. Distribution of change in predicted energy to the method of data addition ($\Delta E = E_0 - E_i$, here $i \in \{Uncertainty, Error\}$) for the datasets of set1. Each panel shows the molecule with the largest decrease or increase in ΔE for the different percentages.

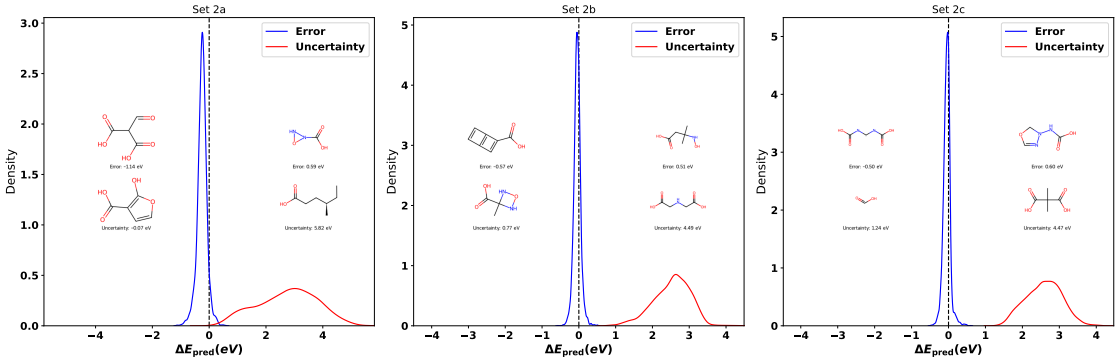

FIG. S46. Distribution of change in predicted energy to the method of data addition ($\Delta E = E_0 - E_i$, here $i \in \{Uncertainty, Error\}$) for the datasets of set2. Each panel shows the molecule with the largest decrease or increase in ΔE for the different percentages.

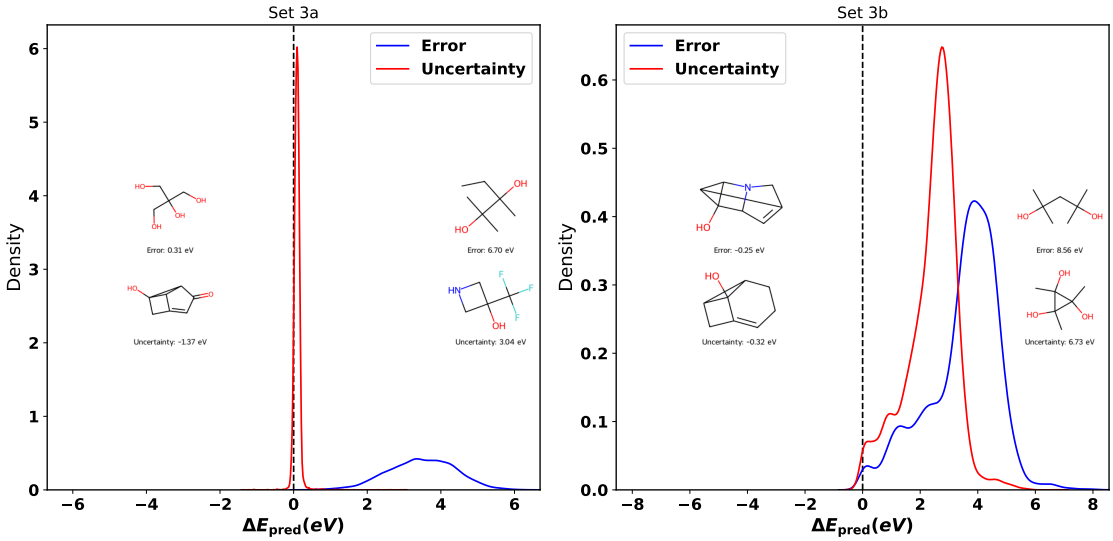

FIG. S47. Distribution of change in predicted energy to the method of data addition ($\Delta E = E_0 - E_i$, here $i \in \{Uncertainty, Error\}$) for the datasets of set3. Each panel shows the molecule with the largest decrease or increase in ΔE for the different percentages.

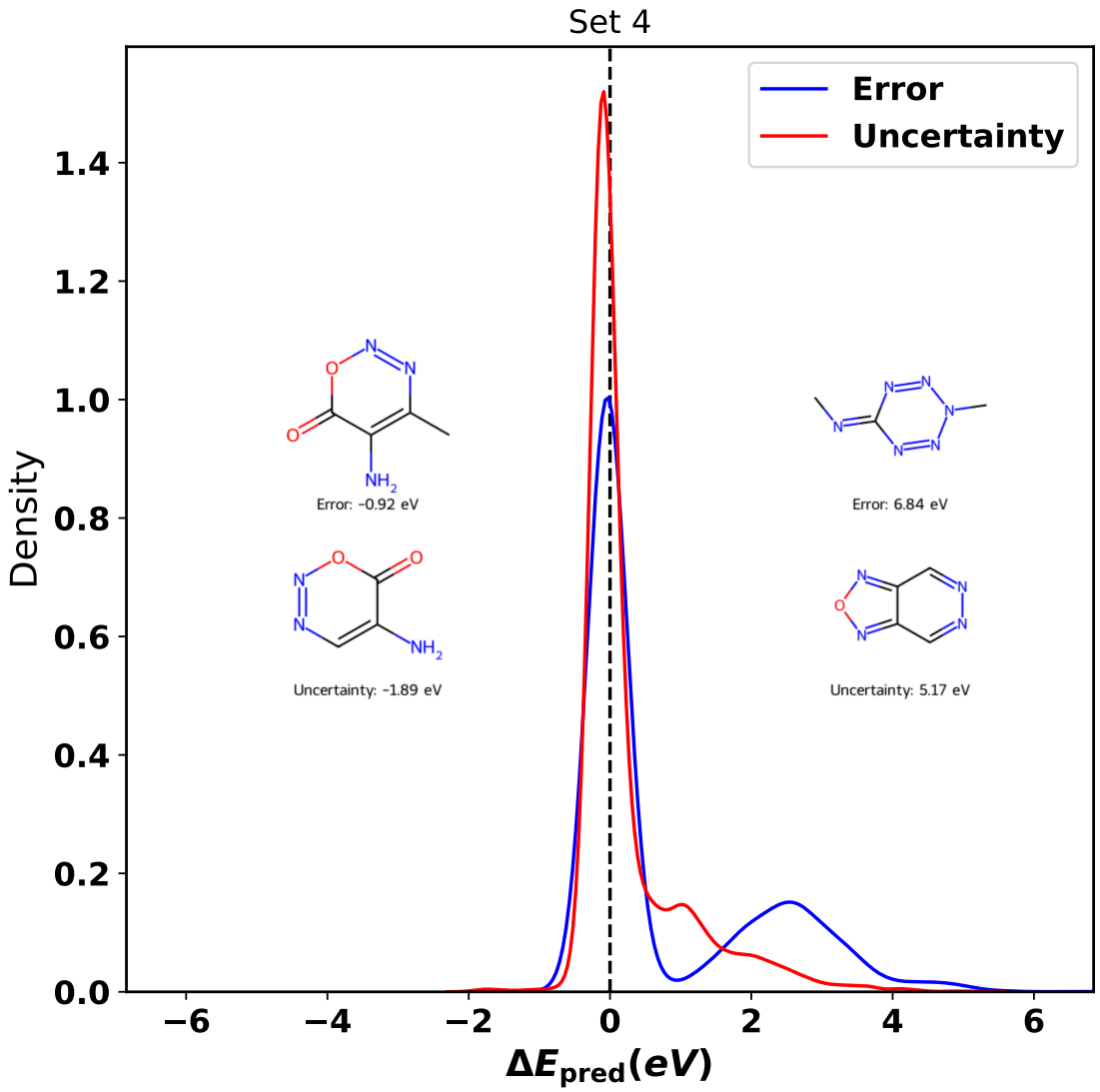

FIG. S48. Distribution of change in predicted energy to the method of data addition ($\Delta E = E_0 - E_i$, here $i \in \{Uncertainty, Error\}$) for the datasets of set4. Each panel shows the molecule with the largest decrease or increase in ΔE for the different percentages.