Supplemental Material

Synthesizability score distribution for all iterations

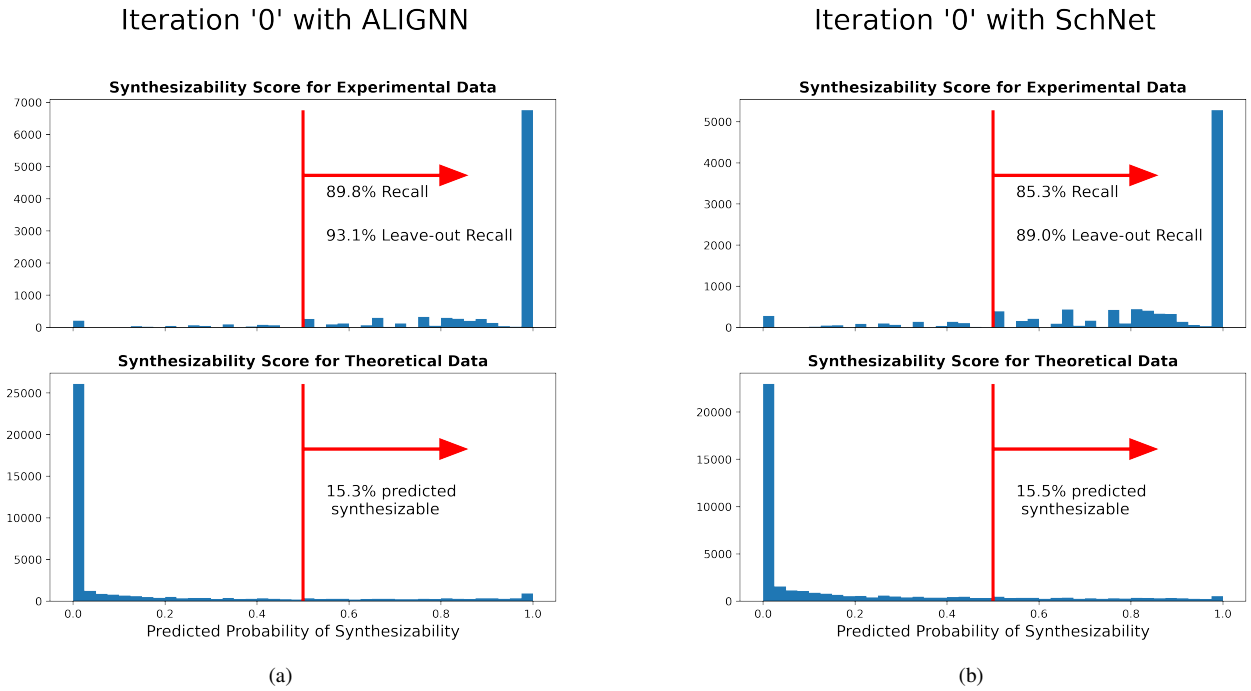


Fig. 1: *Synthesizability score distribution for Iteration '0' for the a) ALIGNN0 series and b) SchNet0 series.*

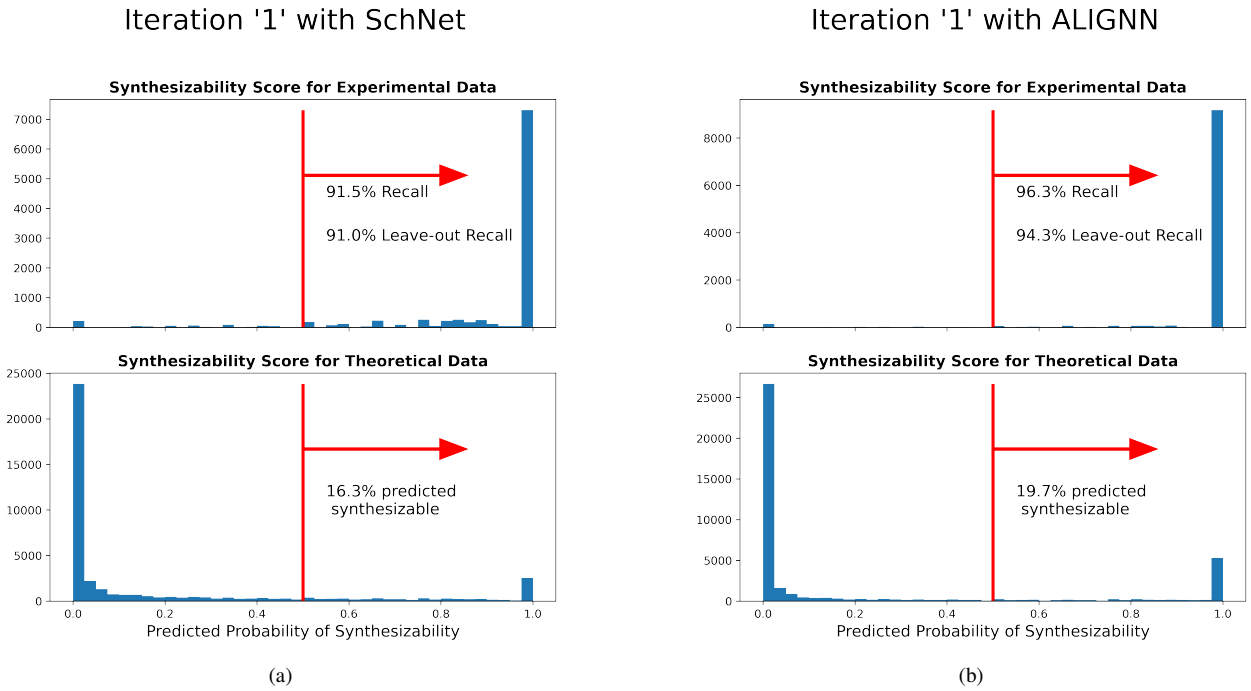
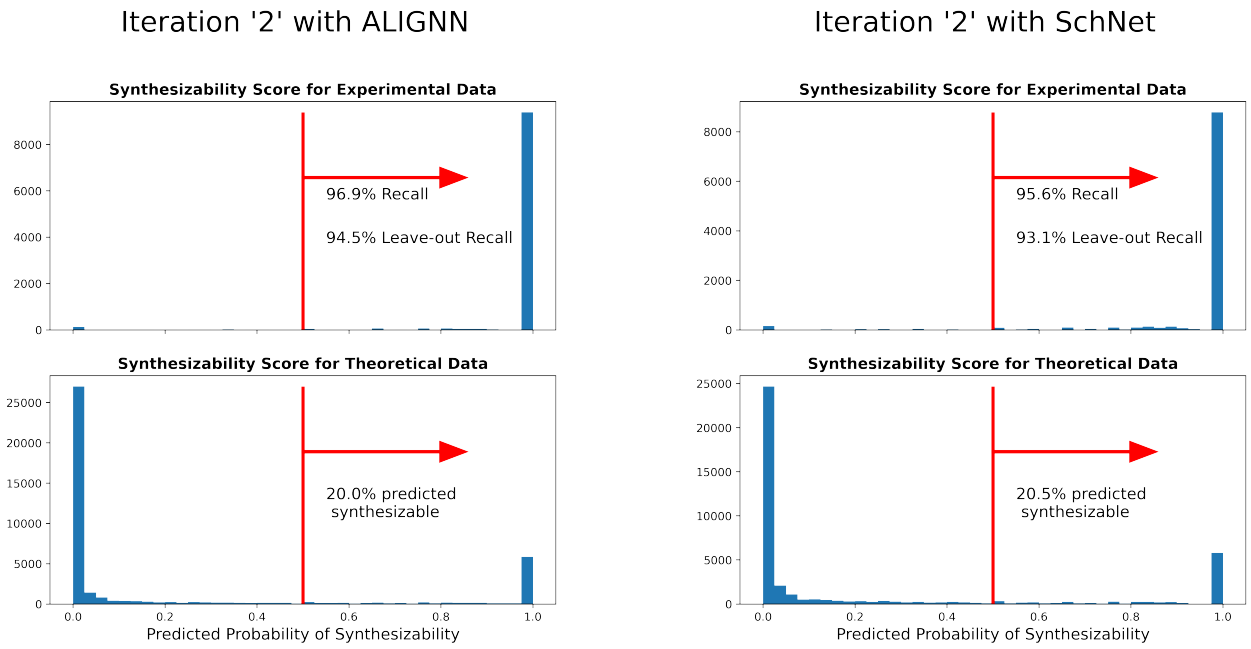
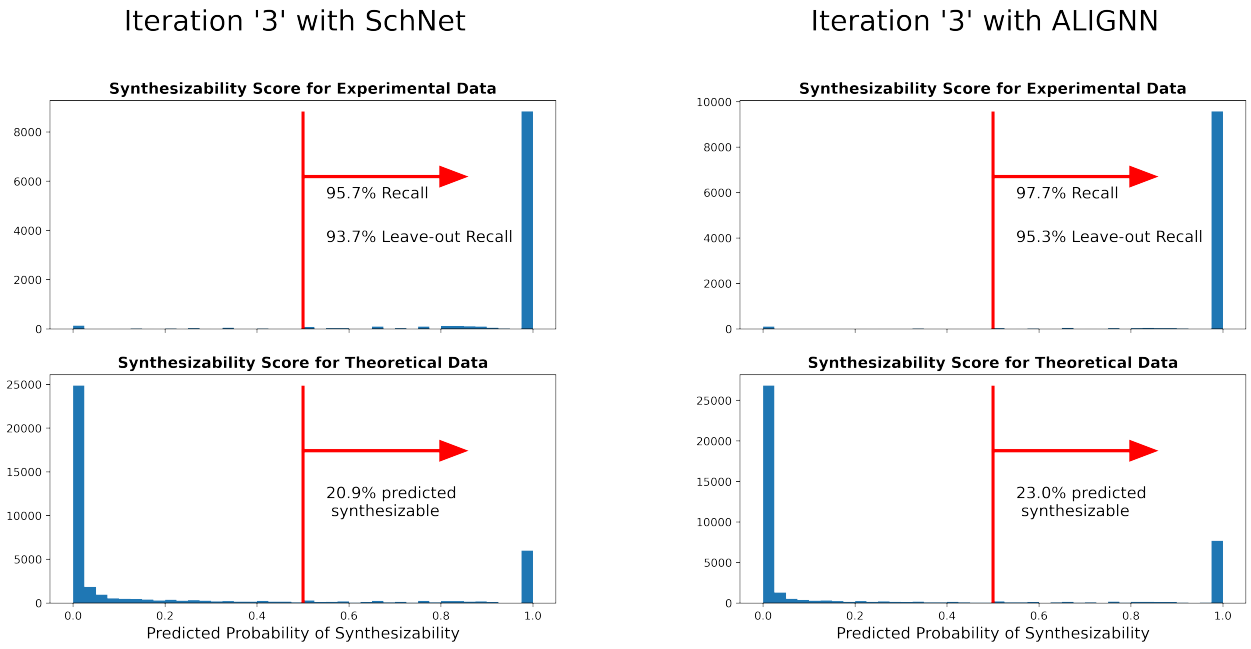


Fig. 2: *Synthesizability score distribution for Iteration '1' for the a) ALIGNN0 series and b) SchNet0 series.*



(a) (b)

Fig. 3: *Synthesizability score distribution for Iteration '2' for the a) ALIGNN0 series and b) SchNet0 series.*



(a) (b)

Fig. 4: *Synthesizability score distribution for Iteration '3' for the a) ALIGNN0 series and b) SchNet0 series.*

Ground truth stability set-up

This auxiliary experiment does not affect the synthesizability prediction results but serves as a demonstration of the reliability of the Recall values. To set up this experiment, we needed a threshold for the energy above the convex hull, which would allow us to label materials as stable (1) or unstable (0). While the obvious choice would have been 0.1 eV, a commonly used threshold for stability, it would not work well for our demonstration.

The performance of the bagging PU Learning algorithm depends on the proportion of unlabeled data belonging to the positive class—what Mordelet and Vert referred to as contamination.[1] The algorithm assumes that part of the unlabeled data is from the negative class, and the smaller the contamination, the more accurate this assumption becomes, resulting in better performance. However, about 72% of our data have an energy above the hull of 0.1 eV or less. If we used 0.1 eV as the threshold, the Recall values, while consistent with the ground truth, would be quite low—making it a poor demonstration of the algorithm's abilities. Instead, we chose 0.015 eV as the threshold for stability, which labels approximately a quarter (26%) of our data as stable. This proportion aligns better with what we expect for synthesizability and provides a more suitable demonstration of the model's capabilities.

There is a valid concern that with such a low threshold, we may be approaching the precision limits of DFT, potentially producing unreliable labels. While this would be a legitimate issue if the goal were to predict stability accurately, we argue that this threshold remains useful for our purposes. Neural networks are highly capable models, able to learn any target, including the imperfections of computational models—much like what is achieved in Δ-Machine Learning.[2] Therefore, it is worth evaluating our results to see if the model has, in fact, learned stability as a property.

As with synthesizability, stability labels were generated from iteration '2'. The label accuracy was 87%, with 6602 crystals misclassified compared to the ground truth label produced by the 0.015 eV threshold. Of these 6602, 97% were cases where the prediction was 1 (stable) but the actual label was 0 (unstable). Upon closer inspection, 85% of these misclassified data points had an energy above the hull below 0.1 eV, compared to 72% for the full dataset. This indicates that the model has indeed learned to associate lower energy above the hull with stability, even when it makes mistakes. To verify this further, a Z-test was conducted to compare whether data points misclassified as label 1 had a significantly higher proportion of materials with energy above the hull below 0.1 eV compared to the rest of the dataset. The p-value was highly significant ($p < 0.0001$).

This result strongly suggests that the model has learned stability as a property, even if our chosen threshold differs from the conventional one.

References

[1] Mordelet F, Vert JP. A bagging SVM to learn from positive and unlabeled examples. Pattern Recognition Letters. 2014 Feb;37:201-9. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0167865513002432.

[2] Ramakrishnan R, Dral PO, Rupp M, von Lilienfeld OA. Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach. Journal of Chemical Theory and Com-

putation. 2015 May;11(5):2087-96. Publisher: American Chemical Society. Available from: https://doi.org/10.1021/acs.jctc.5b00099.