Analysis of Atom-level pretraining with QM data for Graph Neural Networks Molecular property models

Abstract

Despite the rapid and significant advancements in deep learning for Quantitative Structure-Activity Relationship (QSAR) models, the challenge of learning robust molecular representations that effectively generalize in real-world scenarios to novel compounds remains an elusive and unresolved task. This study examines how atom-level pretraining with quantum mechanics (QM) data can mitigate violations of assumptions regarding the distributional similarity between training and test data and therefore improve performance and generalization in downstream tasks. In the public dataset Therapeutics Data Commons (TDC), we show how pretraining on atom-level QM improves performance overall and makes the activation of the features distributes more Gaussian-like which results in a representation that is more robust to distribution shifts. To the best of our knowledge, this is the first time that hidden state molecular representations are analyzed to compare the effects of molecule-level and atom-level pretraining on QM data.

Preprint. Under review.

A Supplementary Materials

A.1 Distribution of activations of features for scratch, HOMO-LUMO pretrained and Atom-level pretrained networks in TDC Lipophilicity Dataset

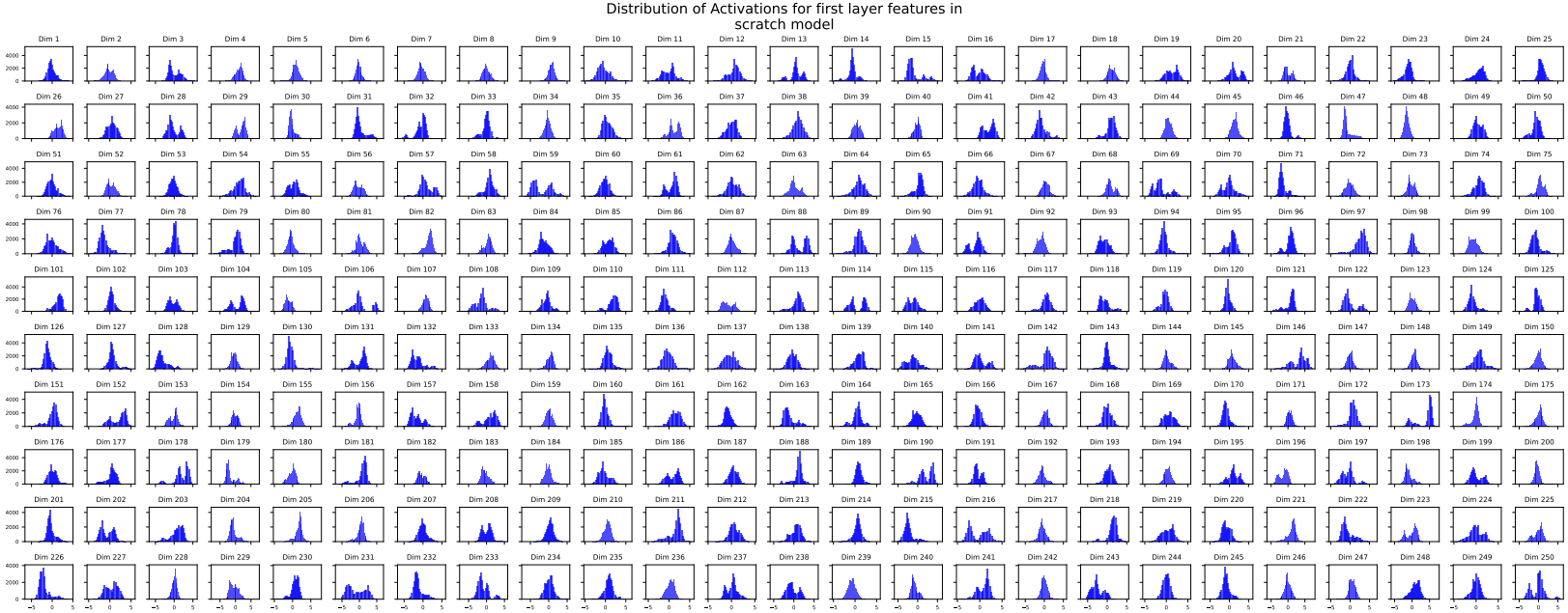

Figure 1: Distribution plots Scratch for Lipophilicity test set

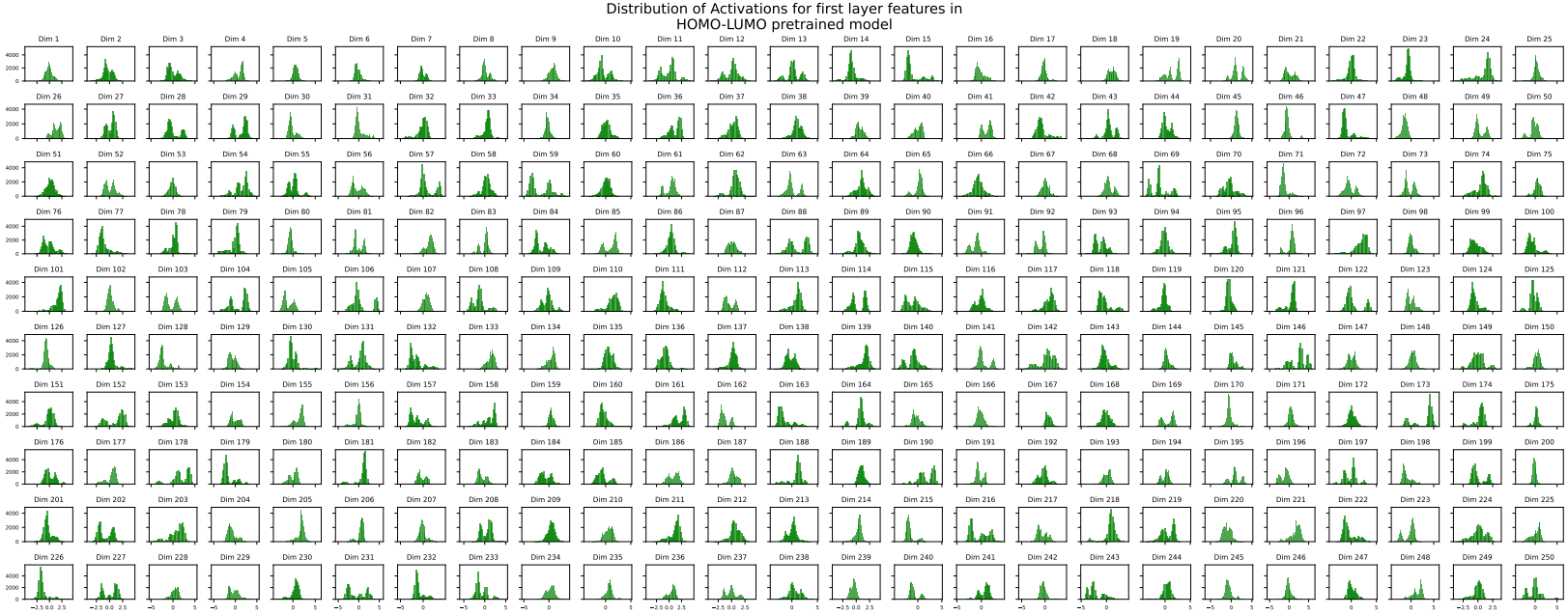

Figure 2: Distribution plots HOMO-LUMO pretrained for Lipophilicity test set

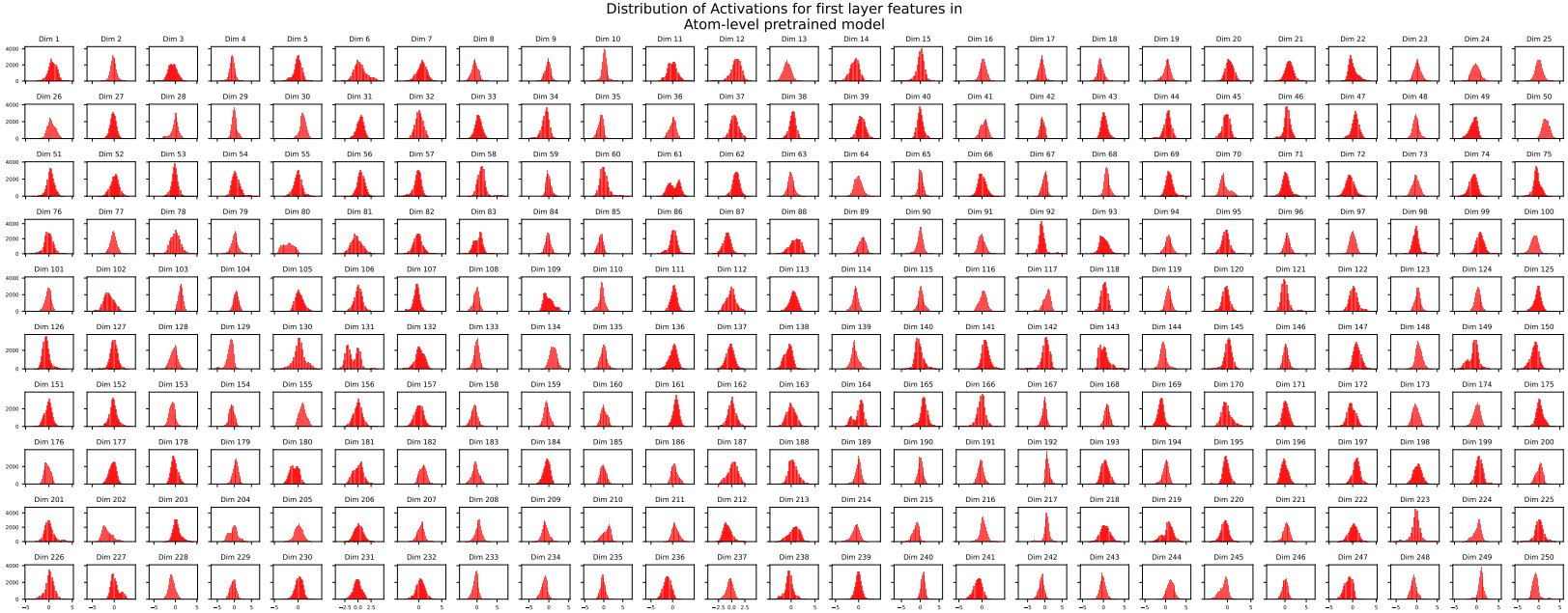

Figure 3: Distribution plots Atom-level pretrained for Lipophilicity test set

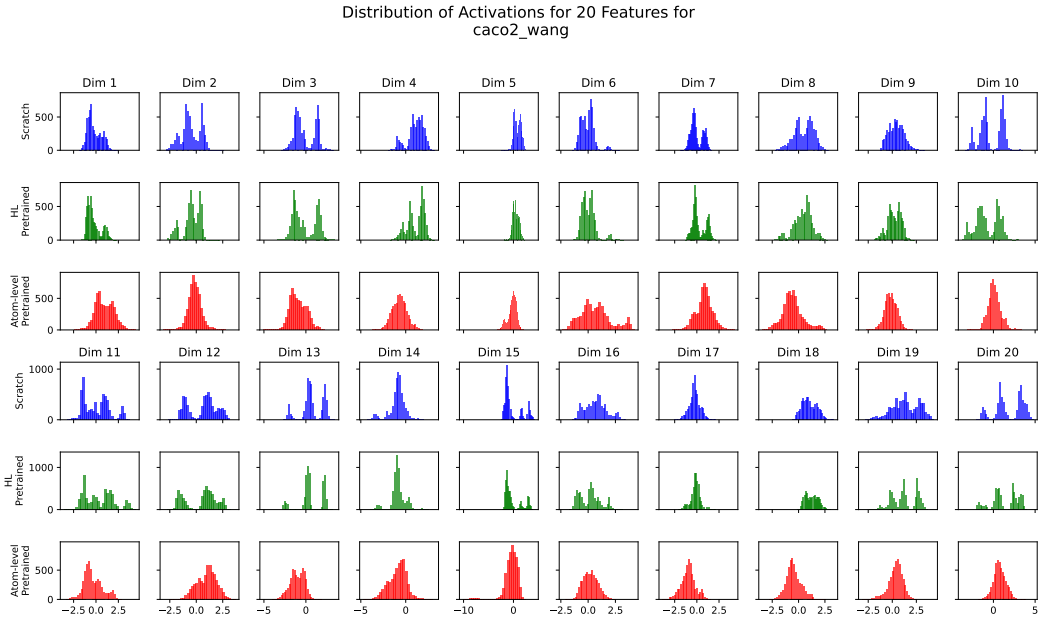

Figure 4: Distribution of first 20 features from the first layer of the Graphormer network for three different training approaches —scratch, HOMO-LUMO pretrained and atom-level pretrained— across test split of caco2 wang dataset.

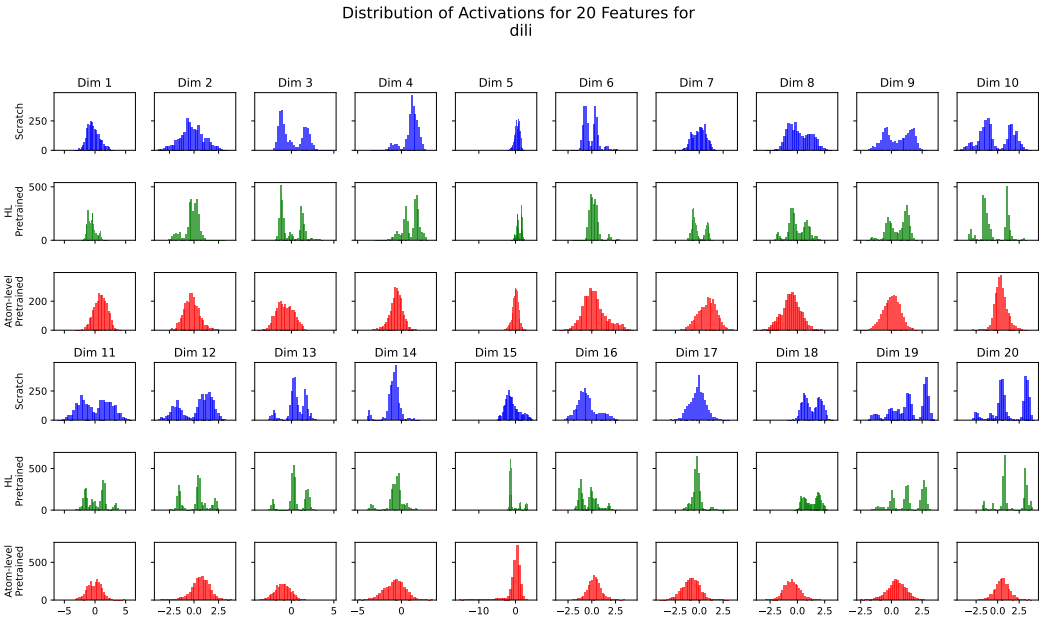

Figure 5: Distribution of first 20 features from the first layer of the Graphormer network for three different training approaches —scratch, HOMO-LUMO pretrained and atom-level pretrained— across test split of dili dataset.

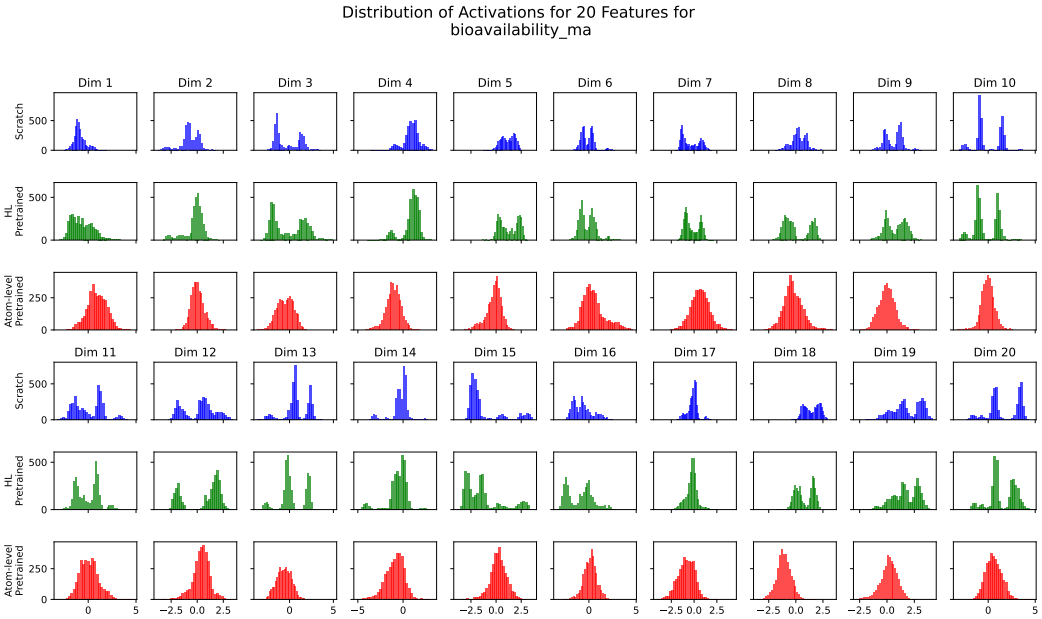

Figure 6: Distribution of first 20 features from the first layer of the Graphormer network for three different training approaches —scratch, HOMO-LUMO pretrained and atom-level pretrained— across test split of bioavailability ma dataset.

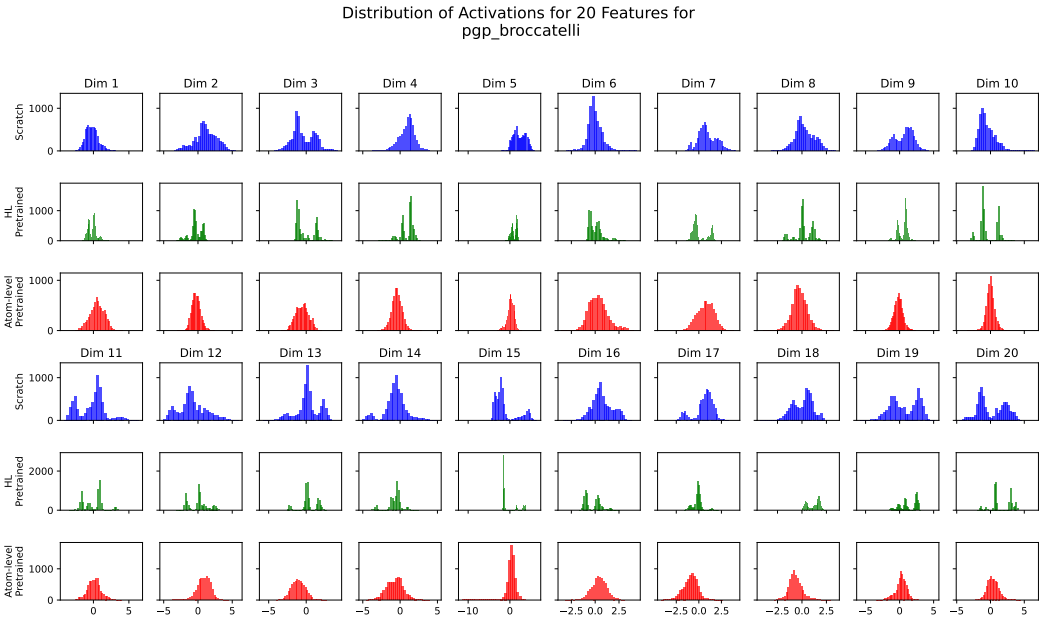

Figure 7: Distribution of first 20 features from the first layer of the Graphormer network for three different training approaches —scratch, HOMO-LUMO pretrained and atom-level pretrained— across test split of pgp broccatelli dataset.

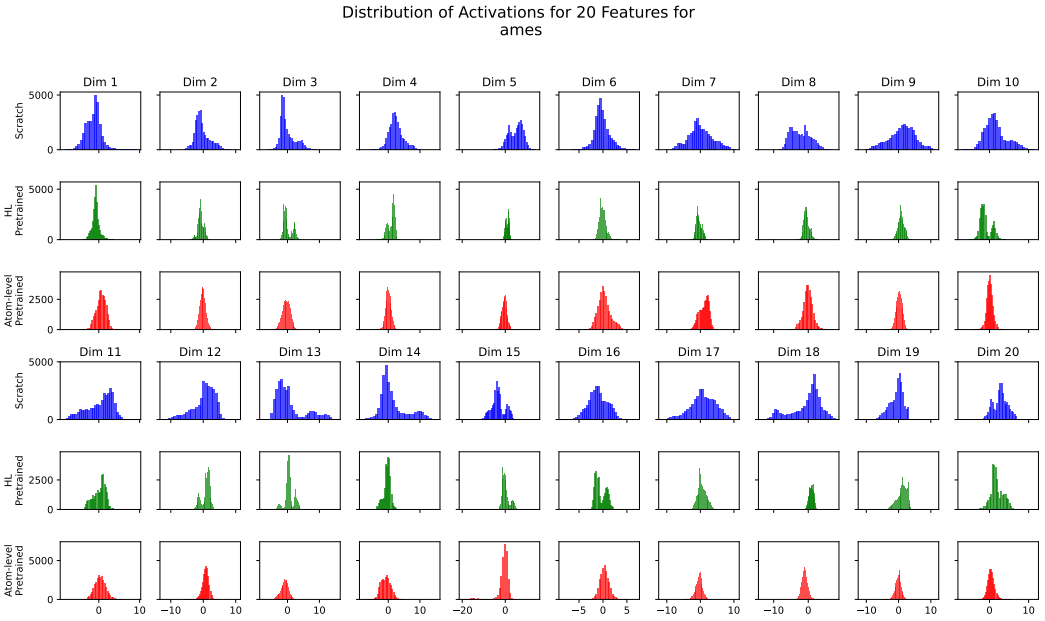

Figure 8: Distribution of first 20 features from the first layer of the Graphormer network for three different training approaches —scratch, HOMO-LUMO pretrained and atom-level pretrained— across test split of ames dataset.

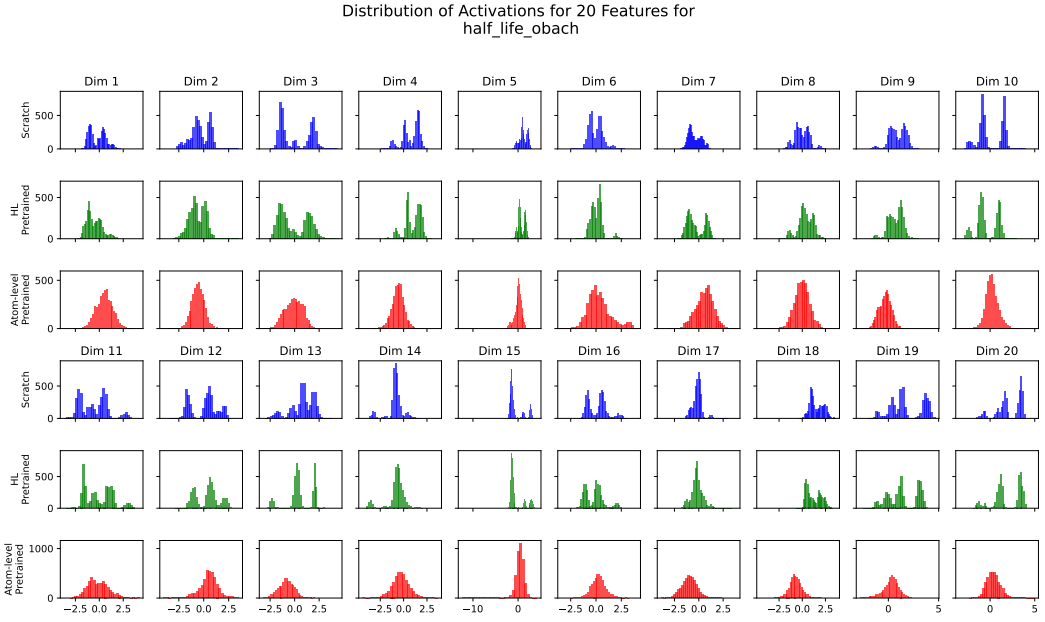

Figure 9: Distribution of first 20 features from the first layer of the Graphormer network for three different training approaches —scratch, HOMO-LUMO pretrained and atom-level pretrained— across test split of half life obach dataset.

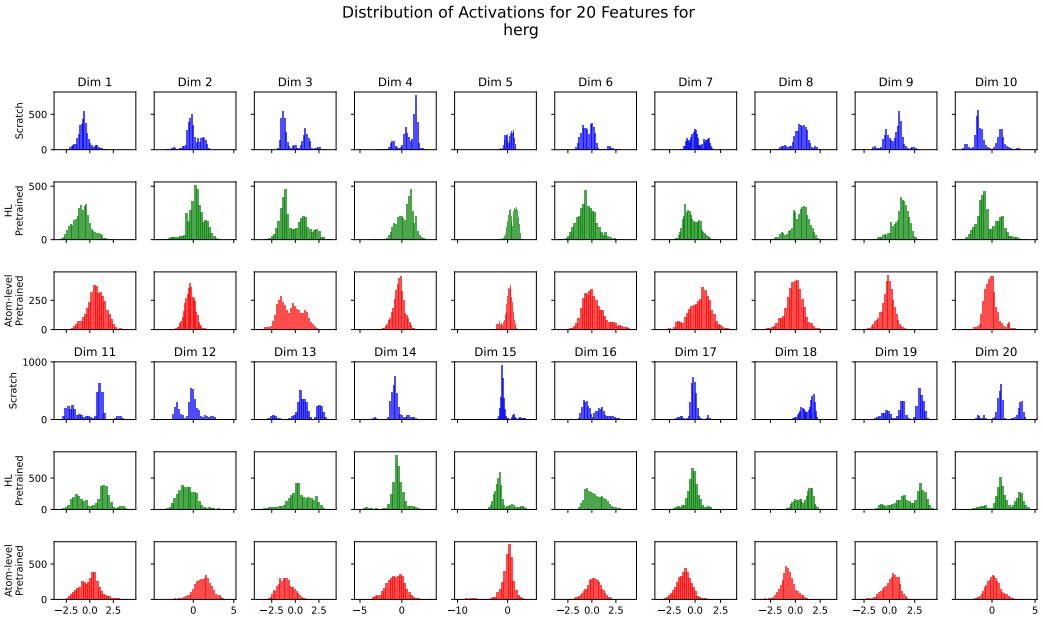

Figure 10: Distribution of first 20 features from the first layer of the Graphormer network for three different training approaches —scratch, HOMO-LUMO pretrained and atom-level pretrained— across test split of herg dataset.

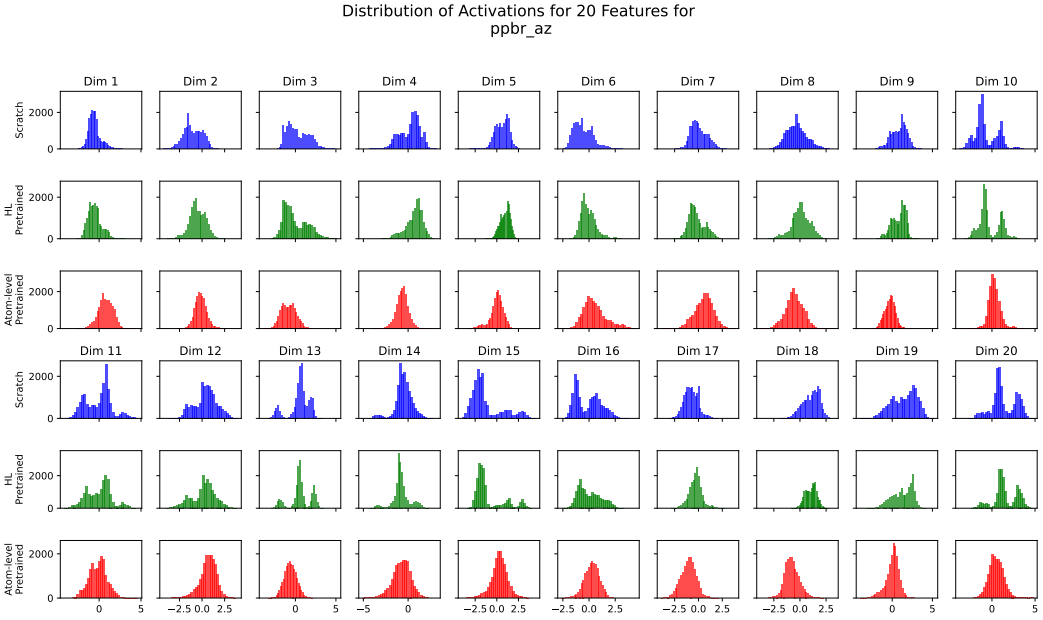

Figure 11: Distribution of first 20 features from the first layer of the Graphormer network for three different training approaches —scratch, HOMO-LUMO pretrained and atom-level pretrained— across test split of ppbr az dataset.

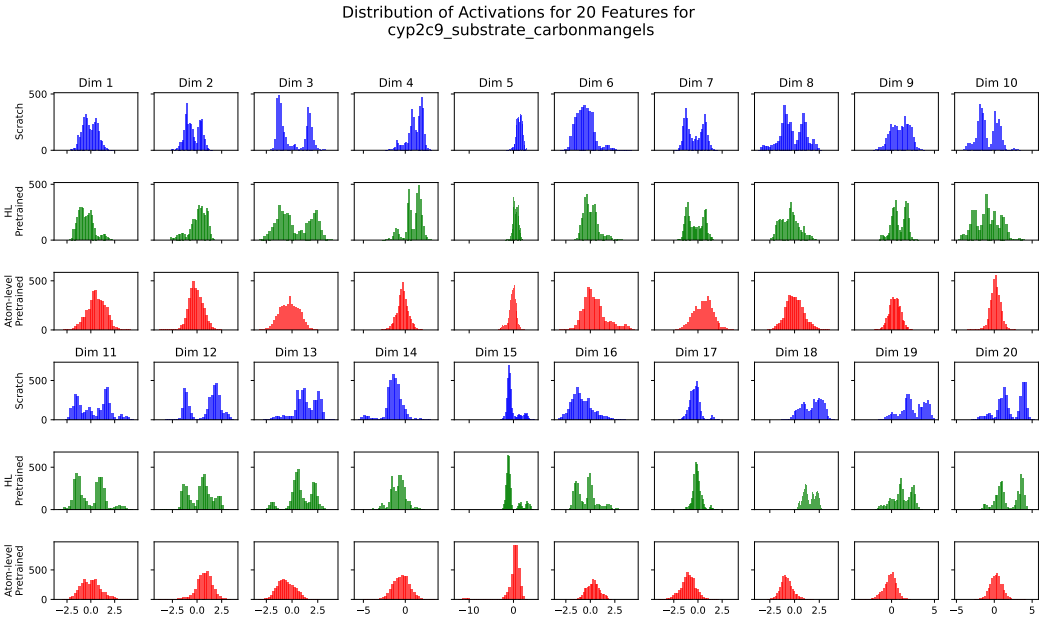

Figure 12: Distribution of first 20 features from the first layer of the Graphormer network for three different training approaches —scratch, HOMO-LUMO pretrained and atom-level pretrained— across test split of cyp2c9 substrate carbonmangels dataset.

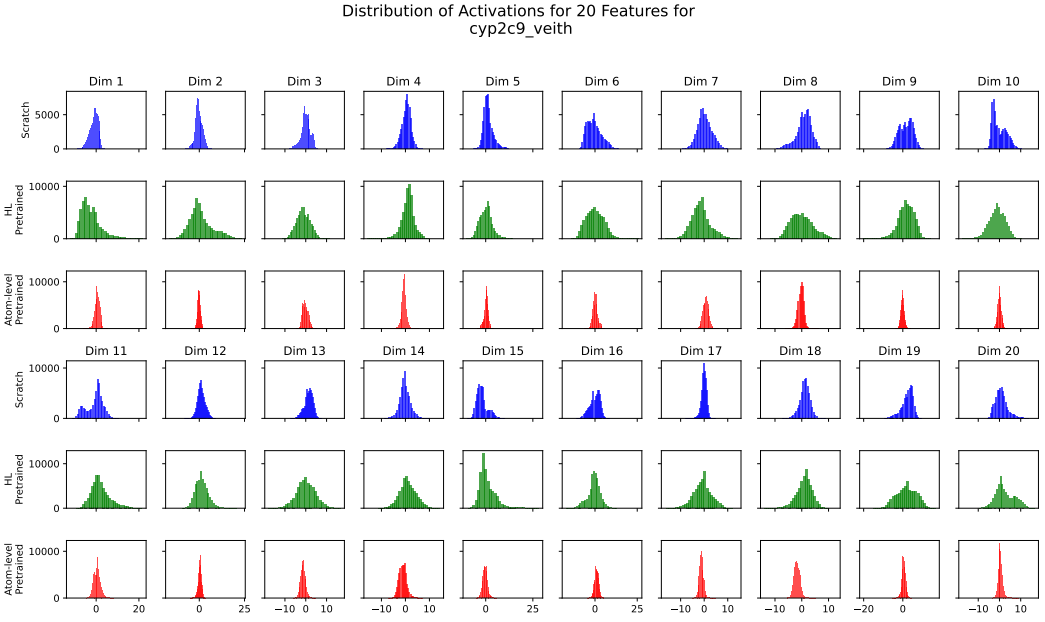

Figure 13: Distribution of first 20 features from the first layer of the Graphormer network for three different training approaches —scratch, HOMO-LUMO pretrained and atom-level pretrained— across test split of cyp2c9 veith dataset.

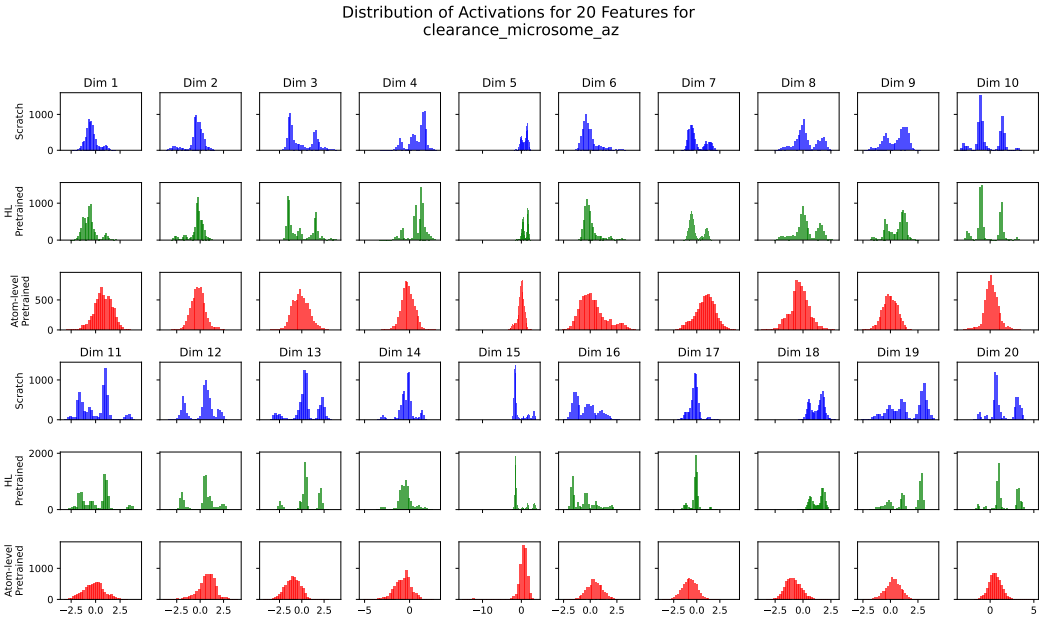

Figure 14: Distribution of first 20 features from the first layer of the Graphormer network for three different training approaches —scratch, HOMO-LUMO pretrained and atom-level pretrained— across test split of clearance microsome az dataset.

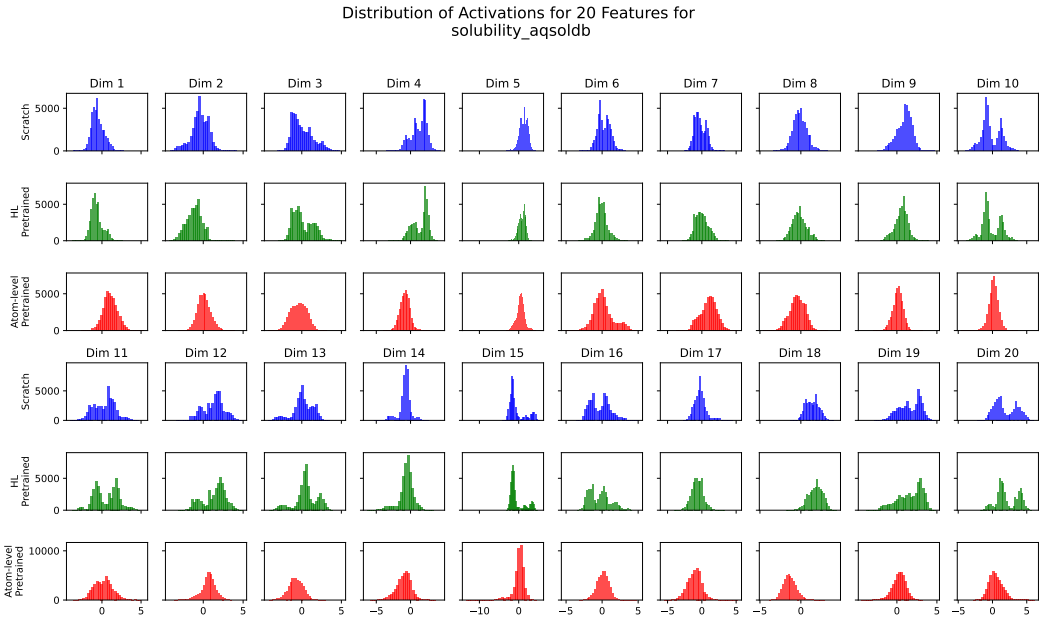

Figure 15: Distribution of first 20 features from the first layer of the Graphormer network for three different training approaches —scratch, HOMO-LUMO pretrained and atom-level pretrained— across test split of solubility aqsoldb dataset.

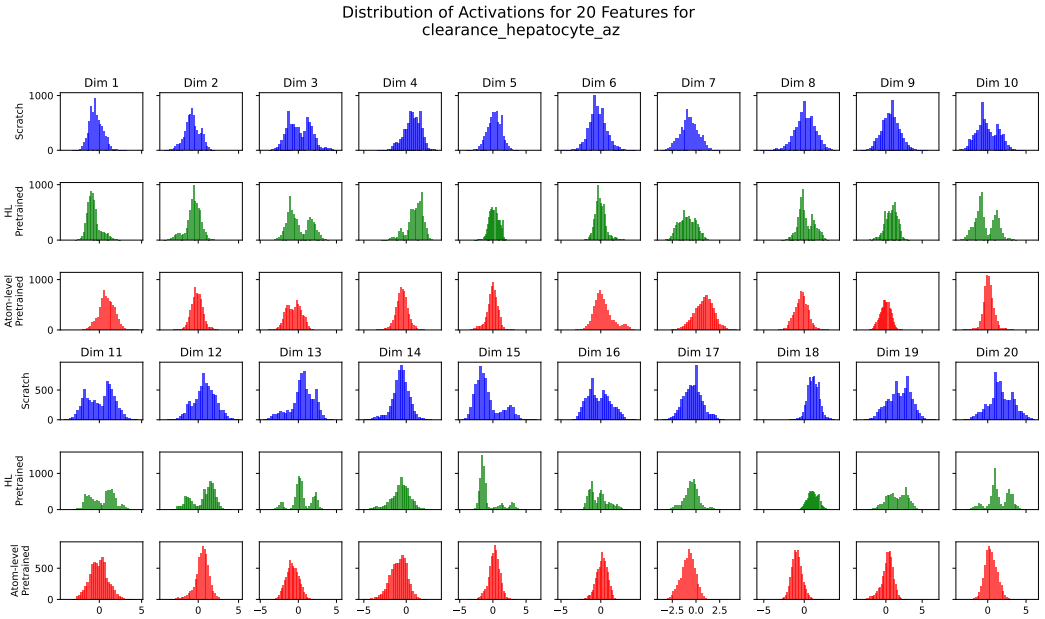

Figure 16: Distribution of first 20 features from the first layer of the Graphormer network for three different training approaches —scratch, HOMO-LUMO pretrained and atom-level pretrained— across test split of clearance hepatocyte az dataset.

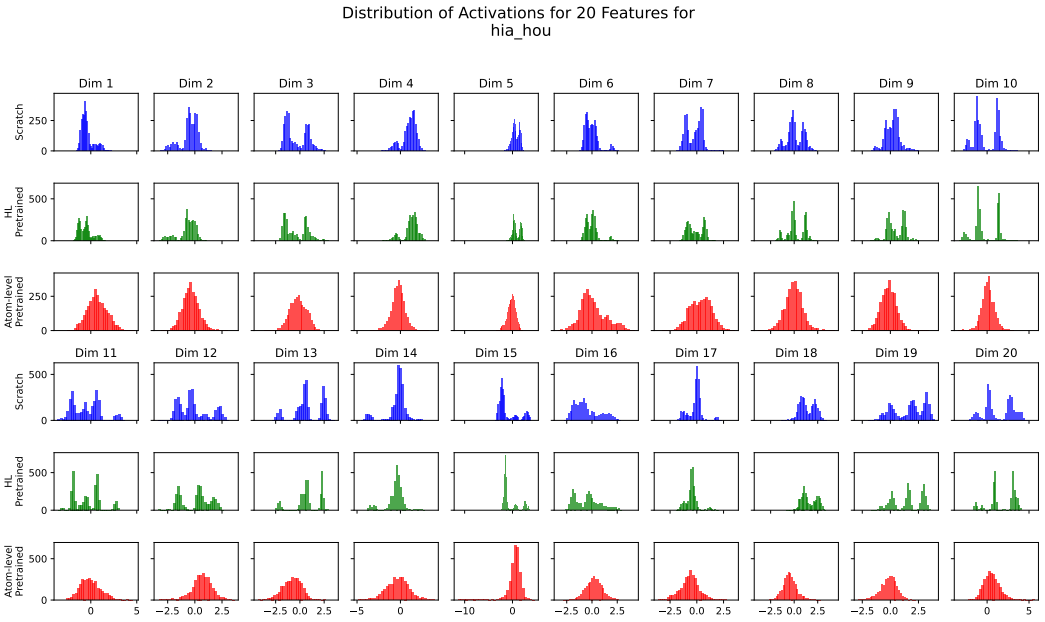

Figure 17: Distribution of first 20 features from the first layer of the Graphormer network for three different training approaches —scratch, HOMO-LUMO pretrained and atom-level pretrained— across test split of hia hou dataset.

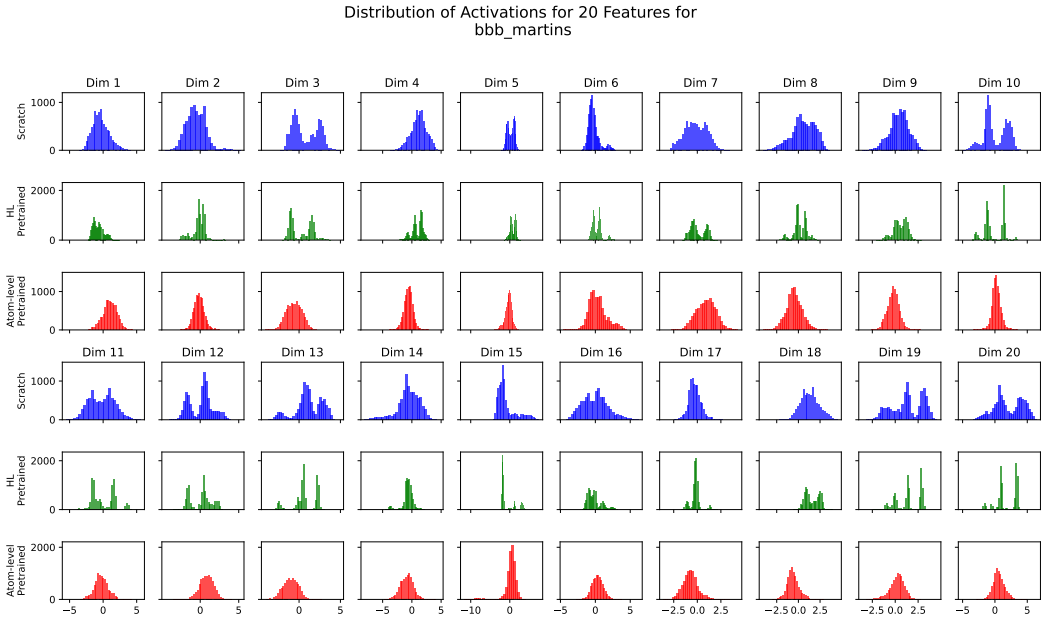

Figure 18: Distribution of first 20 features from the first layer of the Graphormer network for three different training approaches —scratch, HOMO-LUMO pretrained and atom-level pretrained— across test split of bbb martins dataset.

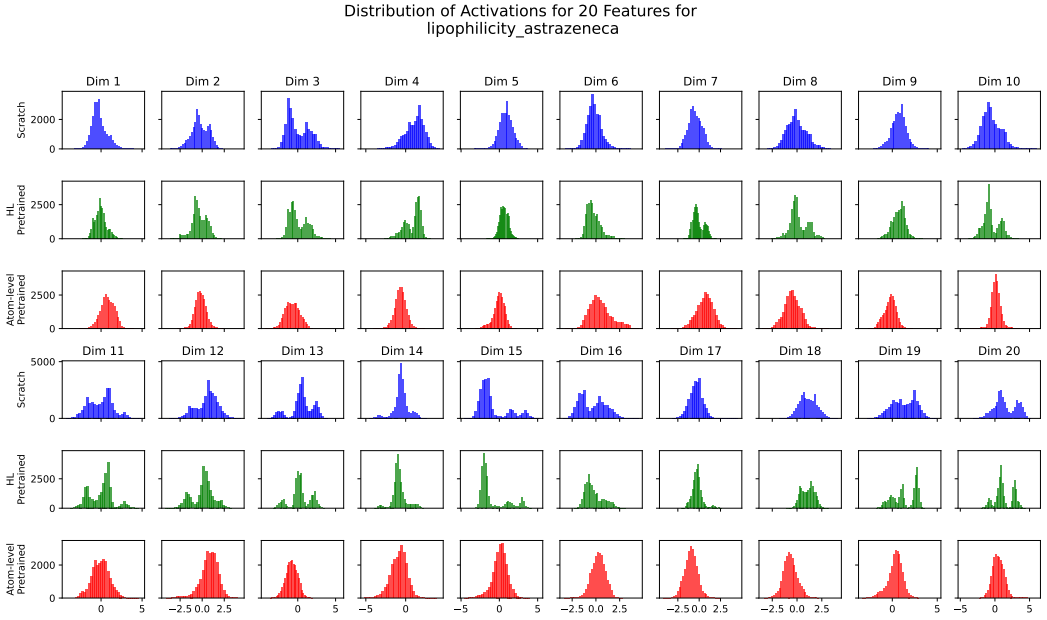

Figure 19: Distribution of first 20 features from the first layer of the Graphormer network for three different training approaches —scratch, HOMO-LUMO pretrained and atom-level pretrained— across test split of lipophilicity astrazeneca dataset.

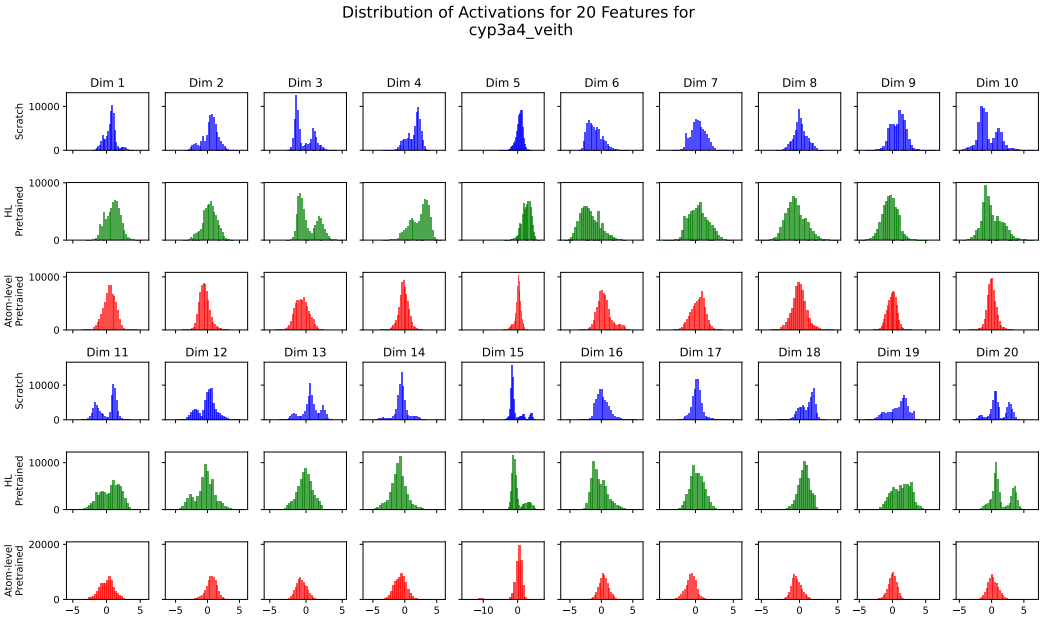

Figure 20: Distribution of first 20 features from the first layer of the Graphormer network for three different training approaches —scratch, HOMO-LUMO pretrained and atom-level pretrained— across test split of cyp3a4 veith dataset.

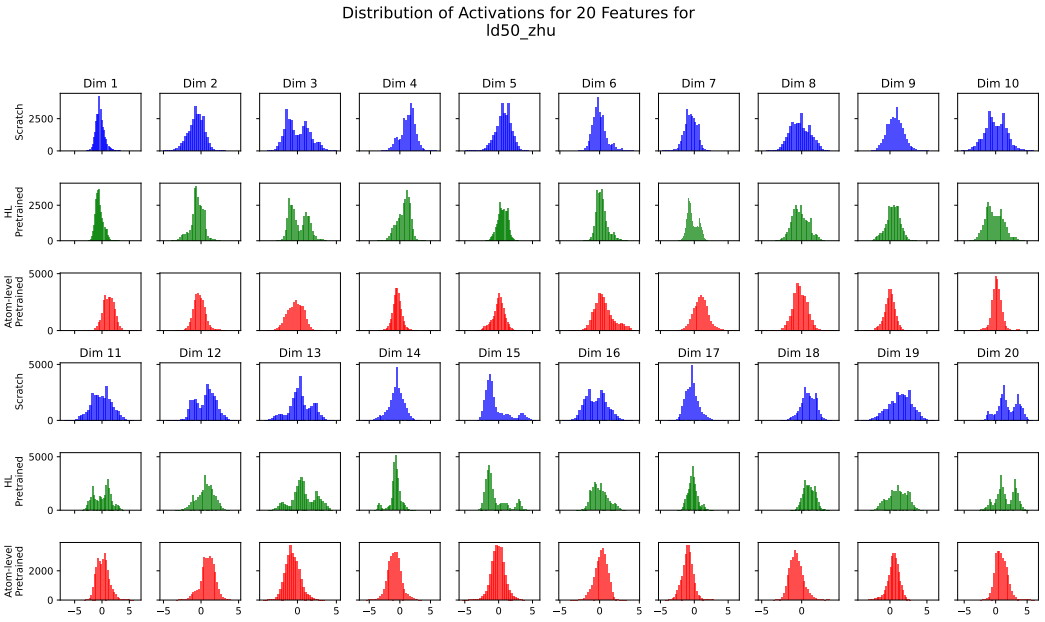

Figure 21: Distribution of first 20 features from the first layer of the Graphormer network for three different training approaches —scratch, HOMO-LUMO pretrained and atom-level pretrained— across test split of ld50 zhu dataset.

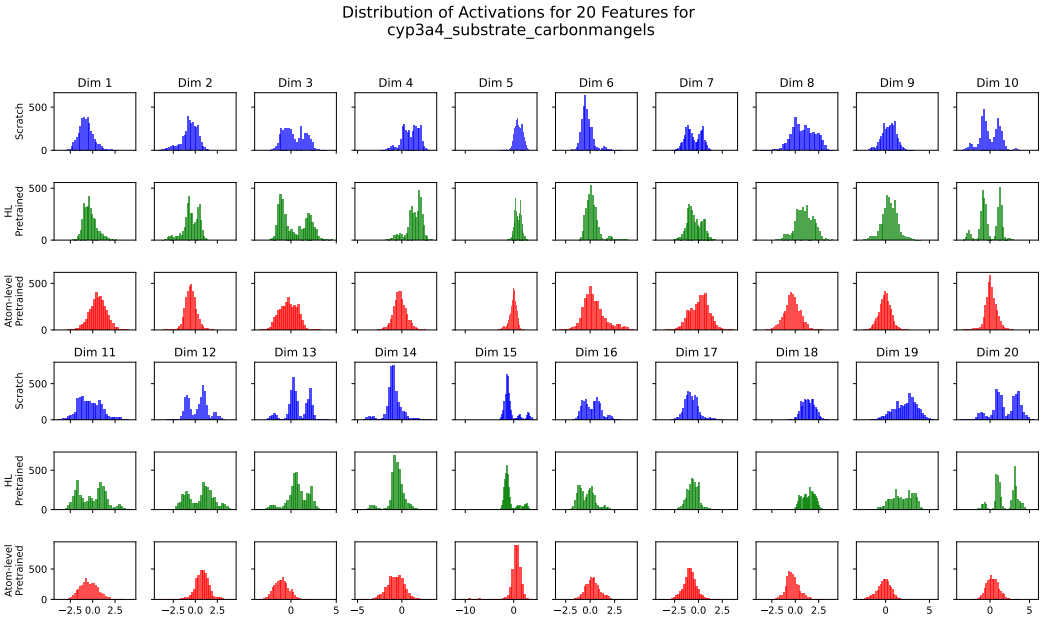

Figure 22: Distribution of first 20 features from the first layer of the Graphormer network for three different training approaches —scratch, HOMO-LUMO pretrained and atom-level pretrained— across test split of cyp3a4 substrate carbonmangels dataset.

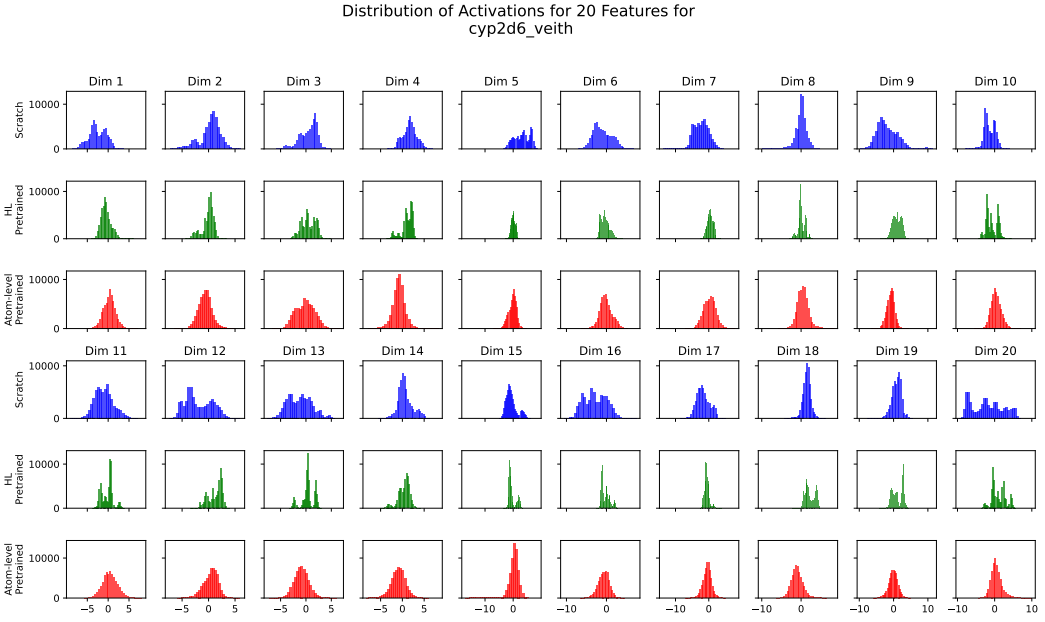

Figure 23: Distribution of first 20 features from the first layer of the Graphormer network for three different training approaches —scratch, HOMO-LUMO pretrained and atom-level pretrained— across test split of cyp2d6 veith dataset.

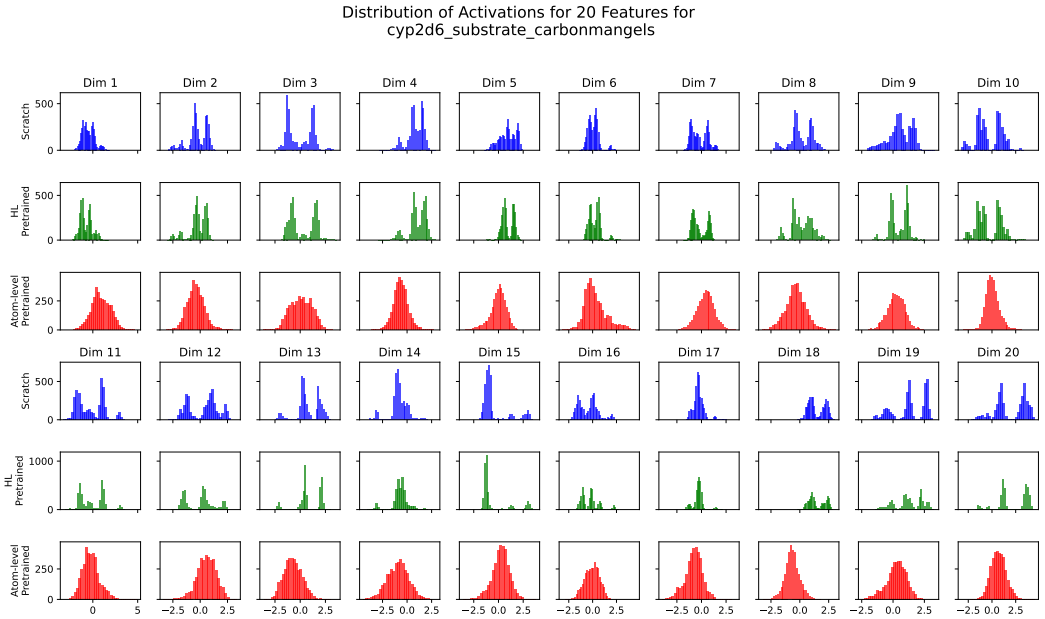

Figure 24: Distribution of first 20 features from the first layer of the Graphormer network for three different training approaches —scratch, HOMO-LUMO pretrained and atom-level pretrained— across test split of cyp2d6 substrate carbonmangels dataset.

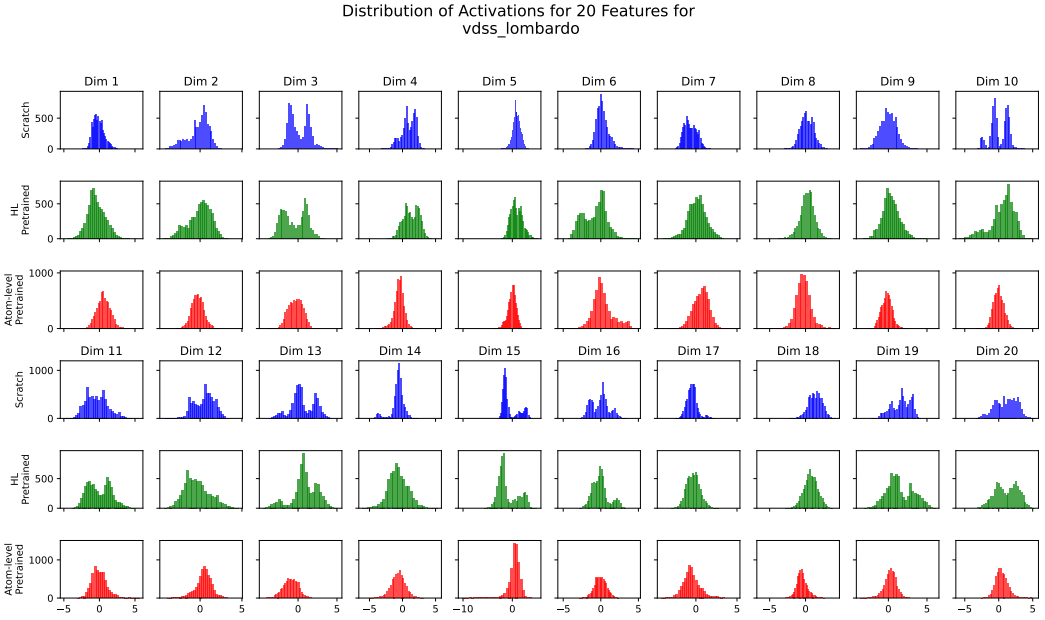

Figure 25: Distribution of first 20 features from the first layer of the Graphormer network for three different training approaches —scratch, HOMO-LUMO pretrained and atom-level pretrained— across test split of vdss lombardo dataset.