# Self-supervised pre-training of atomic and molecular representations with SE(3) invariant graph neural networks

Aryan Pedawi[1], Kate Stafford[1], Andreana Rosnik[1], Saulo de Oliveira[1], Brandon M. Anderson[1], Jon Sorenson[1]

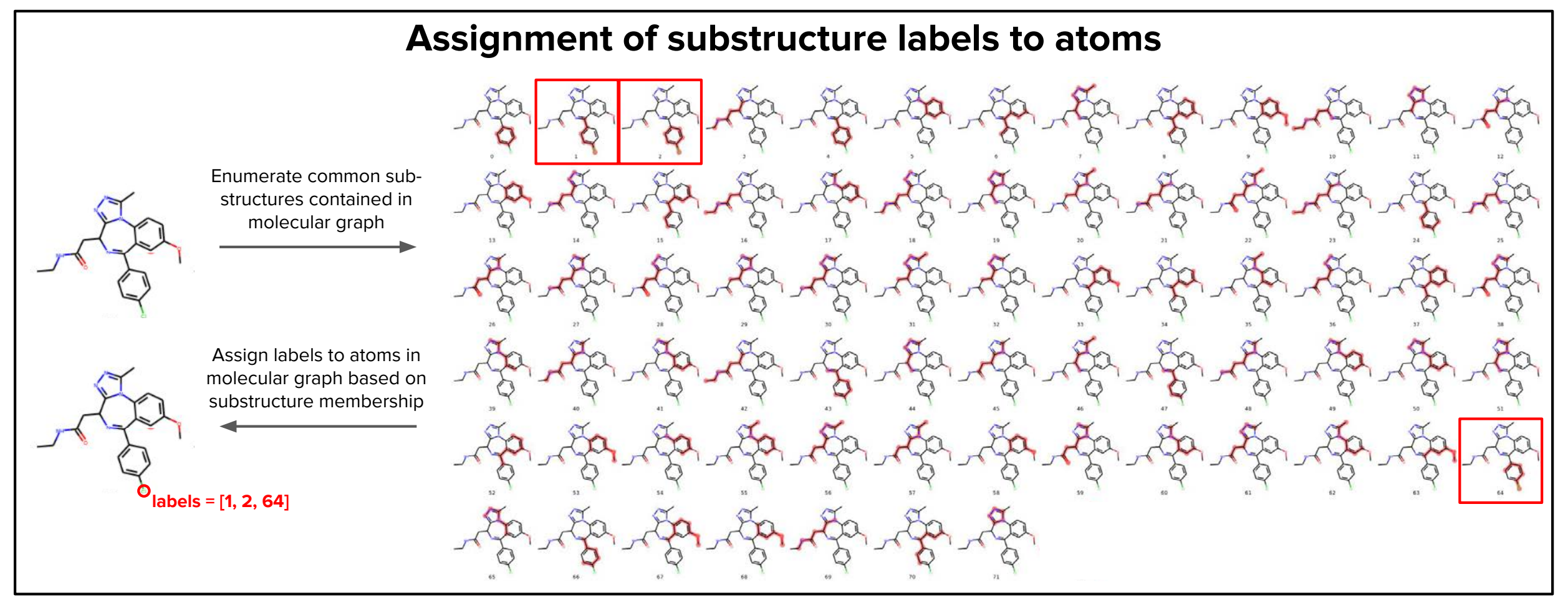
[1] Atomwise Inc., San Francisco, CA 94103 USA

## Introduction

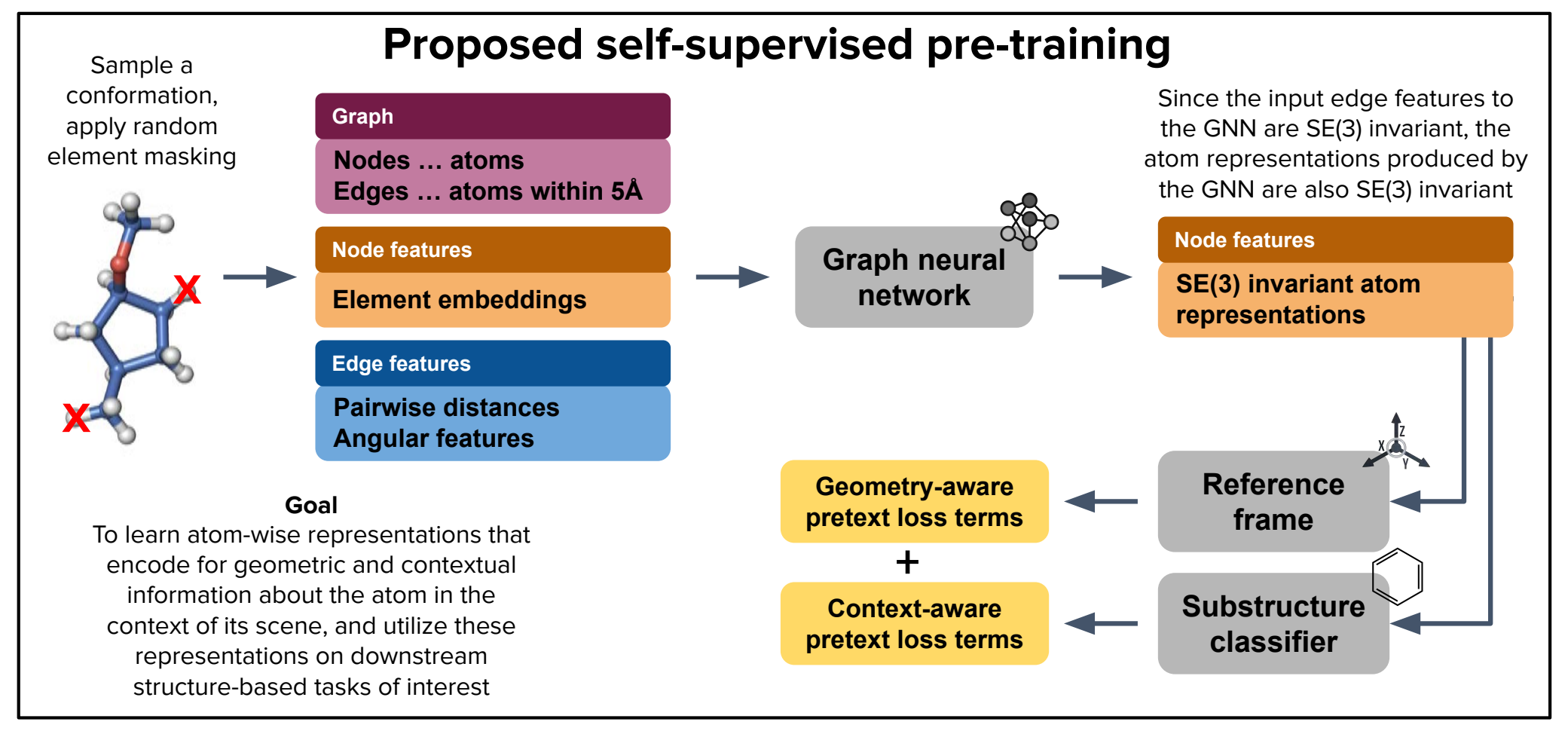
- Many problems in structural biology, chemistry, and physics seek to make predictions from an input structure, represented as a graph with 3D coordinate information.
  - Predicting quantum chemical properties given a 3D conformation of a small molecule.
  - Predicting binding affinity given a docked ligand-protein pose.
- SE(3) invariant/equivariant graph neural networks (GNNs) have emerged as a popular modeling choice in recent years for such problems, achieving state-of-the-art results on various benchmarks of interest.
- These networks are typically trained from random initialization on labeled datasets and are therefore required to learn all pertinent information from scratch.
- This can present challenges in settings where the amount of labeled data for tasks of interest is low, as sufficiently expressive models may not learn the relevant contextual or geometric inductive biases required to generalize appropriately.
- Self-supervised learning has received significant attention in computer vision and natural language settings as a method for pre-training, seeking to learn representations that are generally useful for a variety of relevant downstream tasks.
- The quality of the such downstream models therefore depends on high quality pretext tasks aimed at capturing important inductive biases
- Further, the pretext tasks should be non-trivial and generation of labels should be inexpensive to collect in large quantities.
- In this work, we develop and investigate self-supervised pretext tasks for problems involving 3D molecular conformations of small molecules. In particular, we learn SE(3) invariant representations for molecular conformations that encode for useful chemical and geometrical priors, which can then be fine-tuned on downstream tasks of interest.

## Self-supervision tasks for 3D molecular conformations

- We consider two categories of pretext tasks, which we call **context-aware** and **geometry-aware** tasks.
- Context-aware pretext tasks.
  - Given a representation of an atom in a 3D conformation, we would like to make predictions about the molecular substructures the atom participates in.
  - We enumerate a large set of SMARTS substructures present in training library and assign multi-labels to each atom based on the substructures they belong to, as shown below (left).
  - As such, the model learns atom-wise representations that are predictive of the substructures they participate in.
  - We apply random masking of element tokens during training, drawing inspiration from BERT-style pre-training, masked autoencoders/language models, in-painting.


**Assignment of substructure labels to atoms**

Enumerate common sub-structures contained in molecular graph
Assign labels to atoms in molecular graph based on substructure membership
labels = [1, 2, 64]


**Proposed self-supervised pre-training**

Sample a conformation, apply random element masking

Graph: Nodes … atoms; Edges … atoms within 5Å

Node features: Element embeddings

Edge features: Pairwise distances; Angular features

Graph neural network

Since the input edge features to the GNN are SE(3) invariant, the atom representations produced by the GNN are also SE(3) invariant

Node features: SE(3) invariant atom representations

Reference frame

Substructure classifier

Geometry-aware pretext loss terms + Context-aware pretext loss terms

**Goal**
To learn atom-wise representations that encode for geometric and contextual information about the atom in the context of its scene, and utilize these representations on downstream structure-based tasks of interest

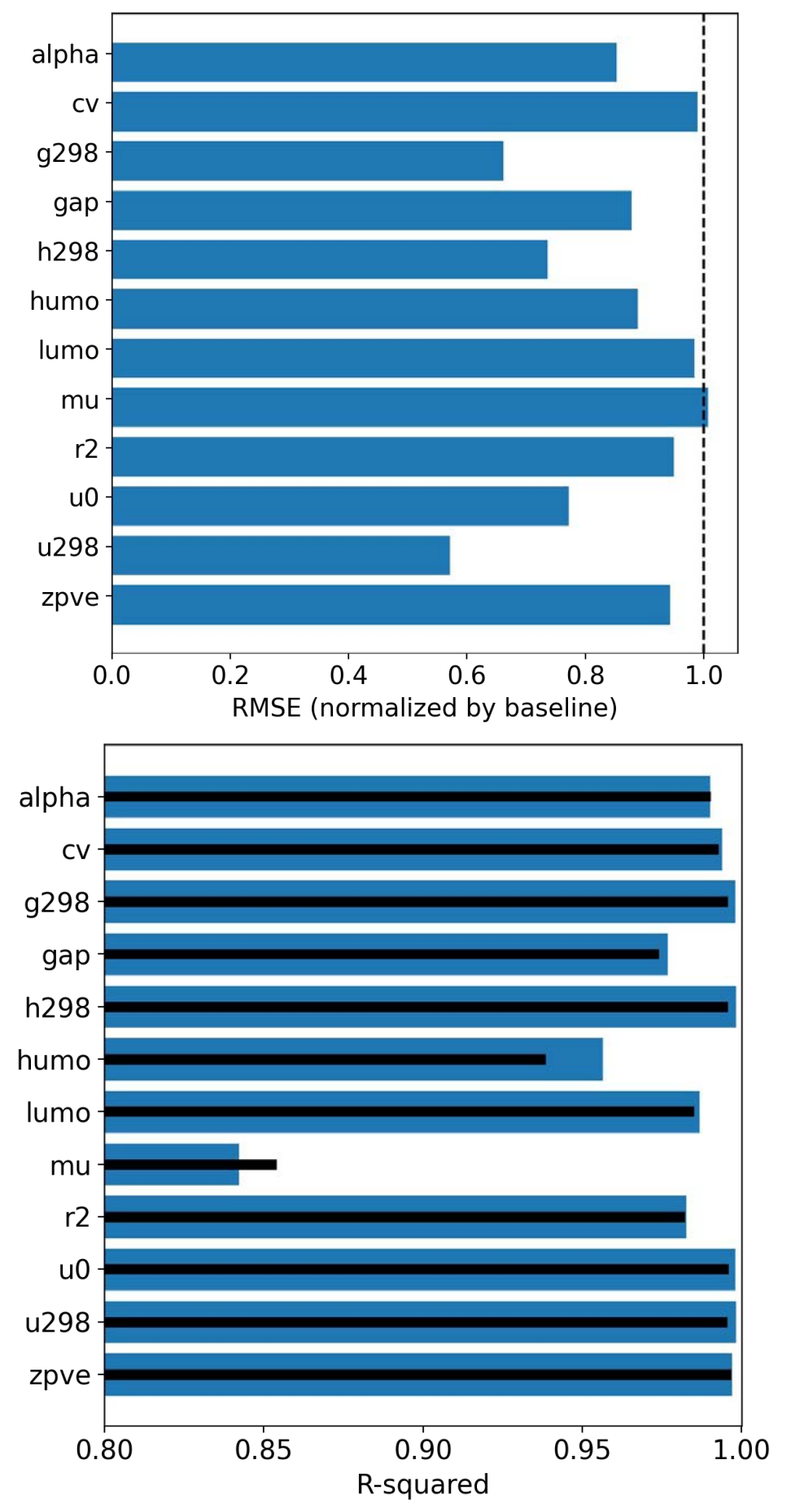
- Geometry-aware pretext tasks.
  - Given a representation of an atom in a 3D conformation, we would like to make predictions about its geometric relationships with other atoms in the scene.
  - We project the atom representations onto three learned axes which serves as a learned global reference frame.
  - We require that two-, three-, and four-body terms in the form of pairwise distances, angles, and dihedrals are reconstructed accurately in the global reference frame.
  - Since we use SE(3) invariant GNNs, these pretext tasks encourage the learned SE(3) invariant atom representations to be informative about relevant geometric quantities pertaining to the atom from the original (non-invariant) coordinate set.
- During self-supervised pre-training, we take a minibatch of molecular graphs, use RDKit to sample a random conformation for each, generate the self-labels (atom membership in substructure set, geometric targets)
- The figure above (right) illustrates the self-supervised pre-training scheme.

## Fine-tuning on QM9 property prediction tasks

- We investigate whether the proposed self-supervised pre-training strategy improves prediction of quantum chemical properties in QM9.
- For this experiment, we carry out an 80-20 random split of the molecules in QM9 into train-holdout sets.
- We apply self-supervised pre-training to the QM9 training set, and subsequently fine-tune on the 12 quantum chemical properties as a supervised multi-task problem.
- We compare against a baseline that skips pre-training and starts the multi-task learner from random initialization.
- We report reductions in prediction error for 11 of the 12 tasks.
  - For this experiment, however, we note gaps across the 12 tasks from SOTA (2-3 order of magnitude difference), which is likely due to choice of a simplified SE(3) invariant GNN and to multi-task training (possibility of negative transfer).
  - Advantages of pre-training on this task may diminish when using more performant GNNs.

### Comparison of results with and without pre-training

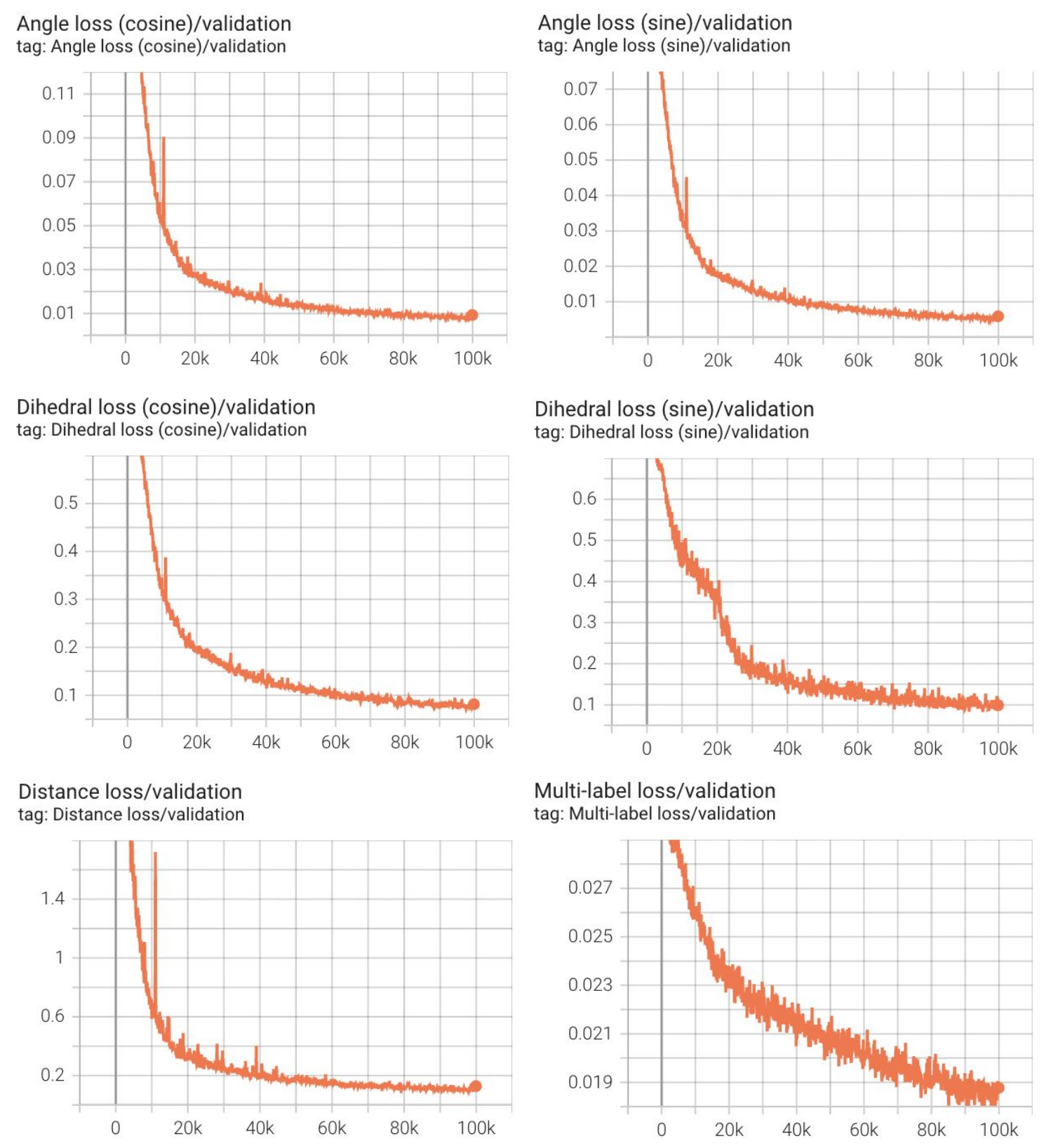
| Property | Units | MAE (w/ pre-training) | MAE (w/o pre-training) |
|---|---|---|---|
| alpha | $Bohr^3$ | **0.498** | 0.532 |
| cv | cal/molK | **0.119** | 0.158 |
| g298 | eV | **0.045** | 0.076 |
| gap | eV | **0.082** | 0.110 |
| h298 | eV | **0.043** | 0.072 |
| homo | eV | **0.081** | 0.085 |
| lumo | eV | **0.072** | 0.078 |
| mu | Debye | 0.204 | **0.168** |
| r2 | $Bohr^2$ | **0.153** | 0.210 |
| u0 | eV | **0.054** | 0.099 |
| u298 | eV | **0.059** | 0.097 |
| zpve | eV | **0.009** | 0.014 |



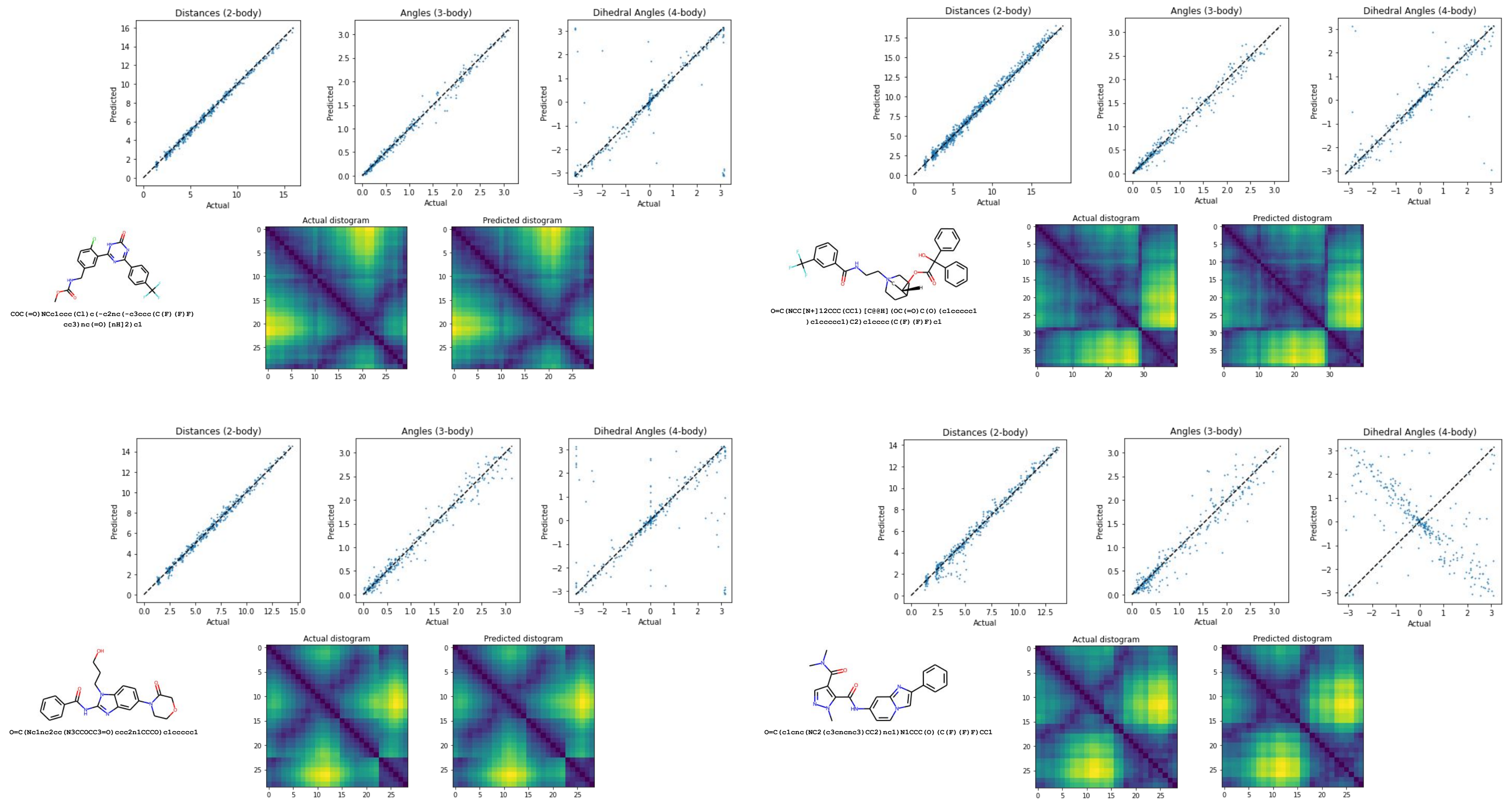
## Self-supervision results on unseen molecules and conformations from ChEMBL

- We apply the self-supervised pre-training strategy to a subset of molecules from ChEMBL.
- We present results on held-out ChEMBL molecules, showing that the self-supervised representations generalize to unseen molecules and their conformations, reconstructing the relevant geometric quantities and atom-substructure membership accurately.
  - Average post-alignment RMSD between raw and reconstructed coordinates (via the global ref. frame) is 0.53Å.
  - Lowest atom-wise AUROC across the 304 SMARTS substructures is 0.88, suggesting atom representations are highly descriptive of the substructures the atom participates in.
- Fine-tuning experiments on downstream structure-based tasks forthcoming.

### Loss curves on held-out molecules



### Example reconstructions on held-out molecules



## Conclusion

We consider new small molecule self-supervised tasks for SE(3) invariant GNNs. The strategies we propose are shown to encode for relevant geometrical and chemical context about an atom in its scene. Experiments on QM9 quantum chemical property prediction are encouraging, but further experimentation is warranted and evaluation is needed on other benchmarks.

#ACSSpring2022

info.atomwise.com/acs_spring2022