Supporting Information

Knowledge distillation of neural network potential for molecular crystals

Takuya Taniguchi*[1]

[1] Center for Data Science, Waseda University, 1-6-1 Nishiwaseda, Shinjuku-ku, Tokyo 169-8050, Japan

* Correspondence to takuya.taniguchi@aoni.waseda.jp

Contents

Supporting Figures

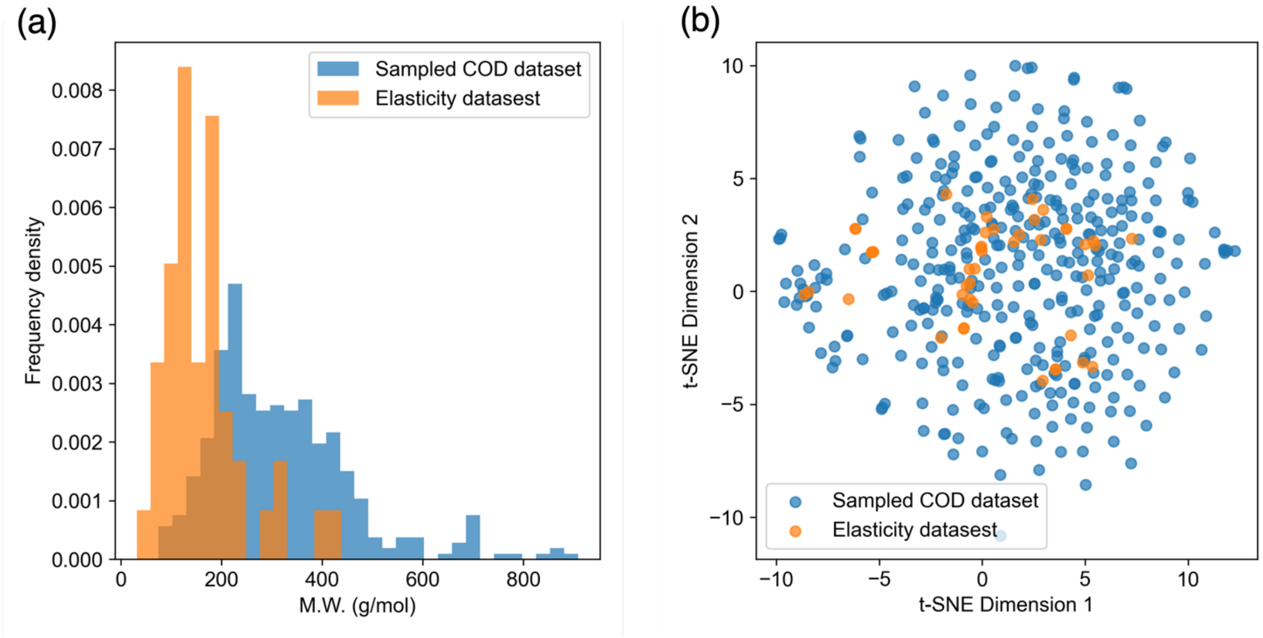

Figure S1. Comparison of the sampled COD dataset and the elasticity dataset. (a) Frequency density of molecular weight of two datasets. (b) Two-dimensional visualization of two datasets using t-SNE. The molecular structures are vectorized by extended-connectivity fingerprint (ECFP).

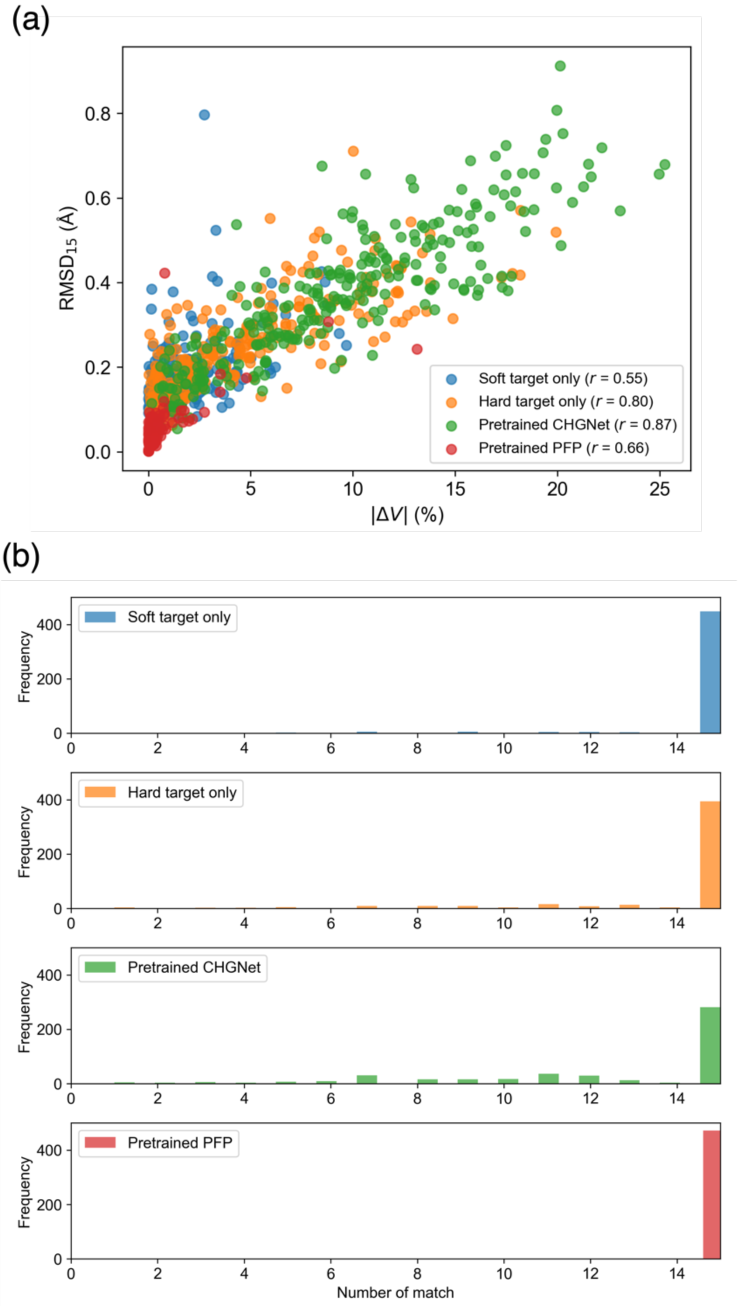

Figure S2. Distribution of RMSD$_{15}$ in the COD dataset. (a) The correlation between the absolute value of cell volume reproducibility and RMSD$_{15}$. Pearson's correlation coefficients (r) show medium to strong positive correlation. (b) Number of matched molecules when calculating RMSD$_{15}$.

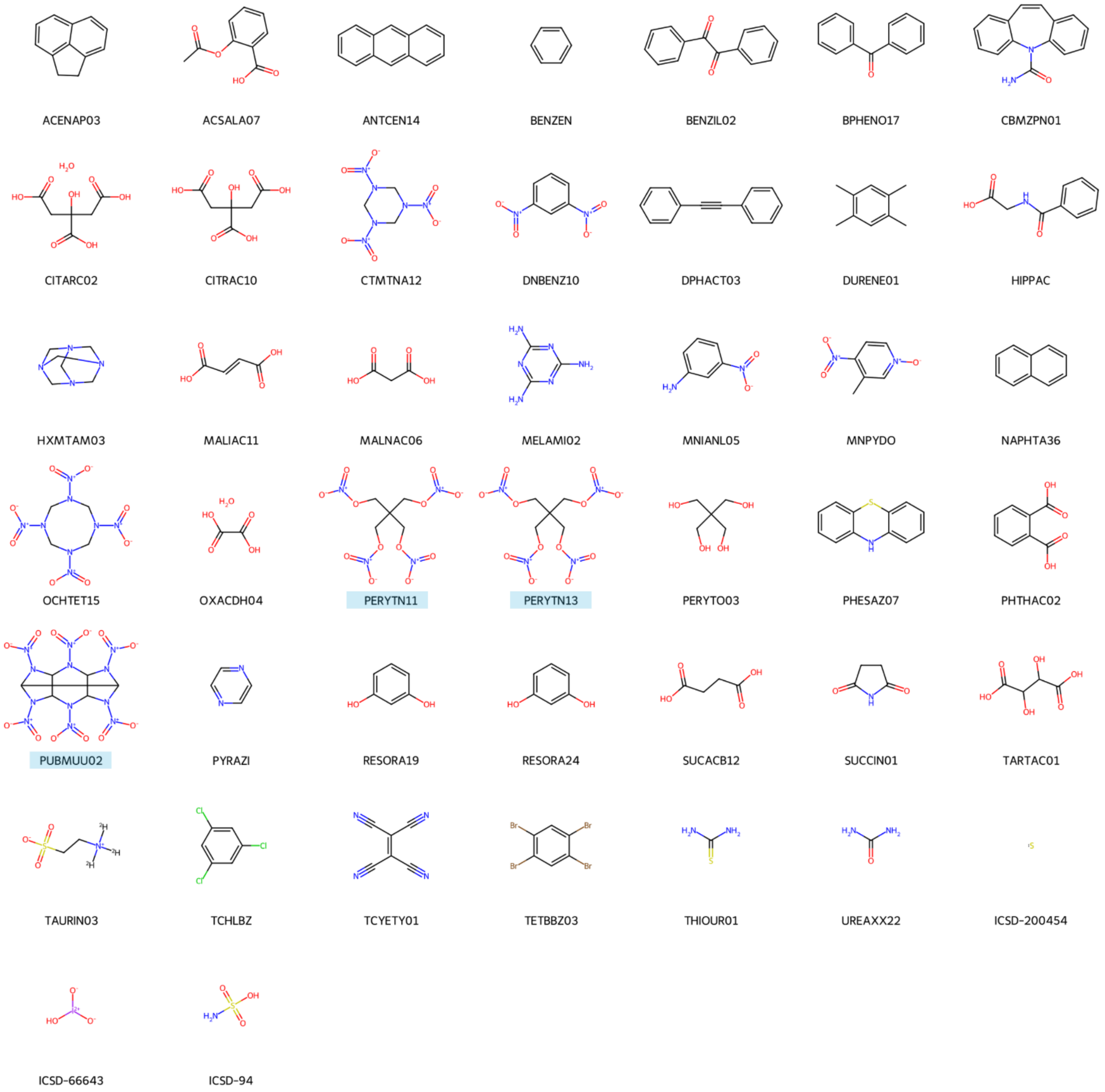

Figure S3. Molecular structure formula of the elasticity dataset. Identifiers of CSD or ICSD are written below each molecule. Most molecules satisfy the Lipinski's rule of five, which is an empirical rule for determining whether a chemical compound has a high probability of being an orally active drug in humans. Some molecules highlighted in blue do not satisfy this rule.

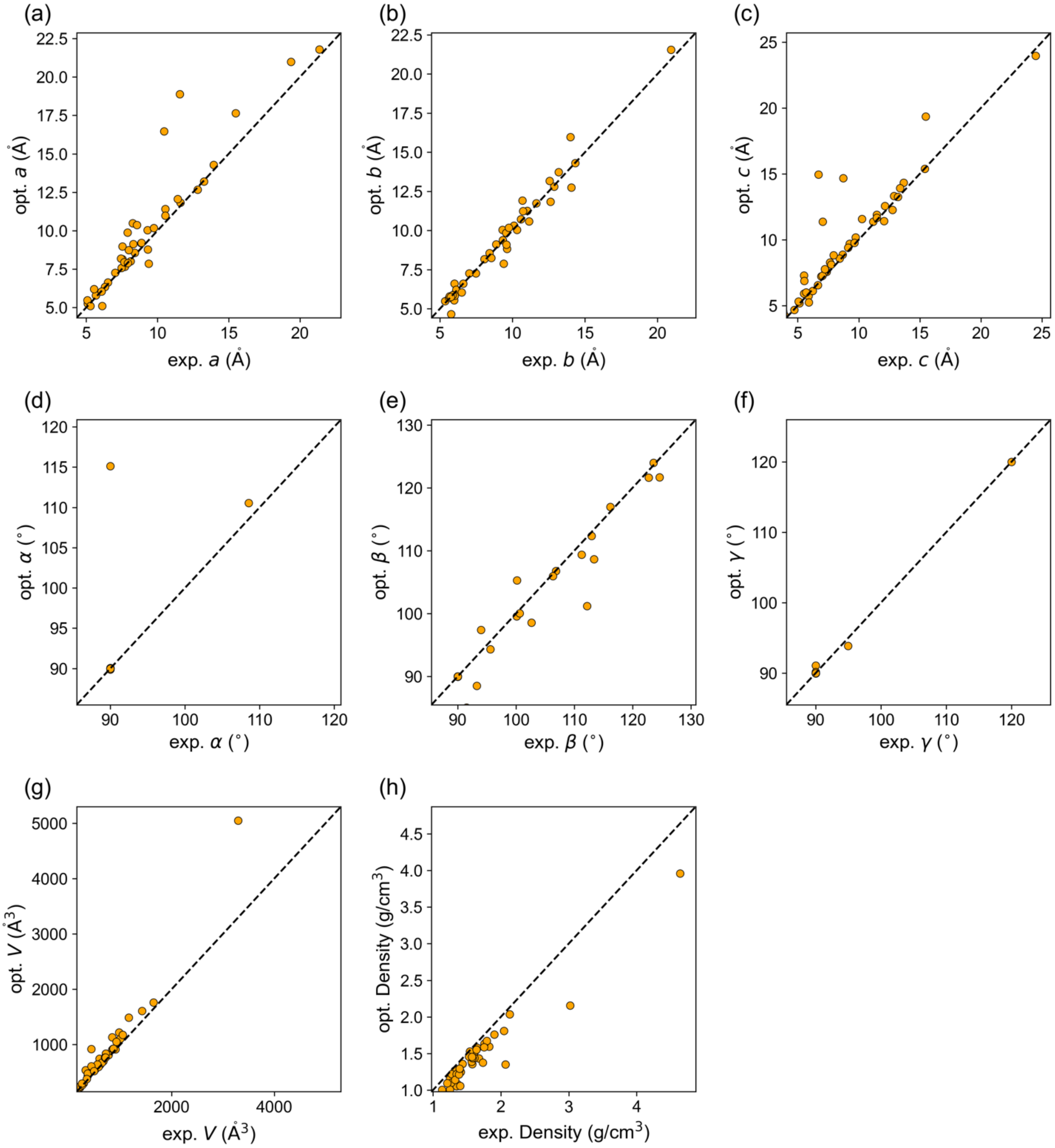

Figure S4. The lattice changes before and after the cell optimization using pre-trained CHGNet. Elasticity dataset was used for evaluation.

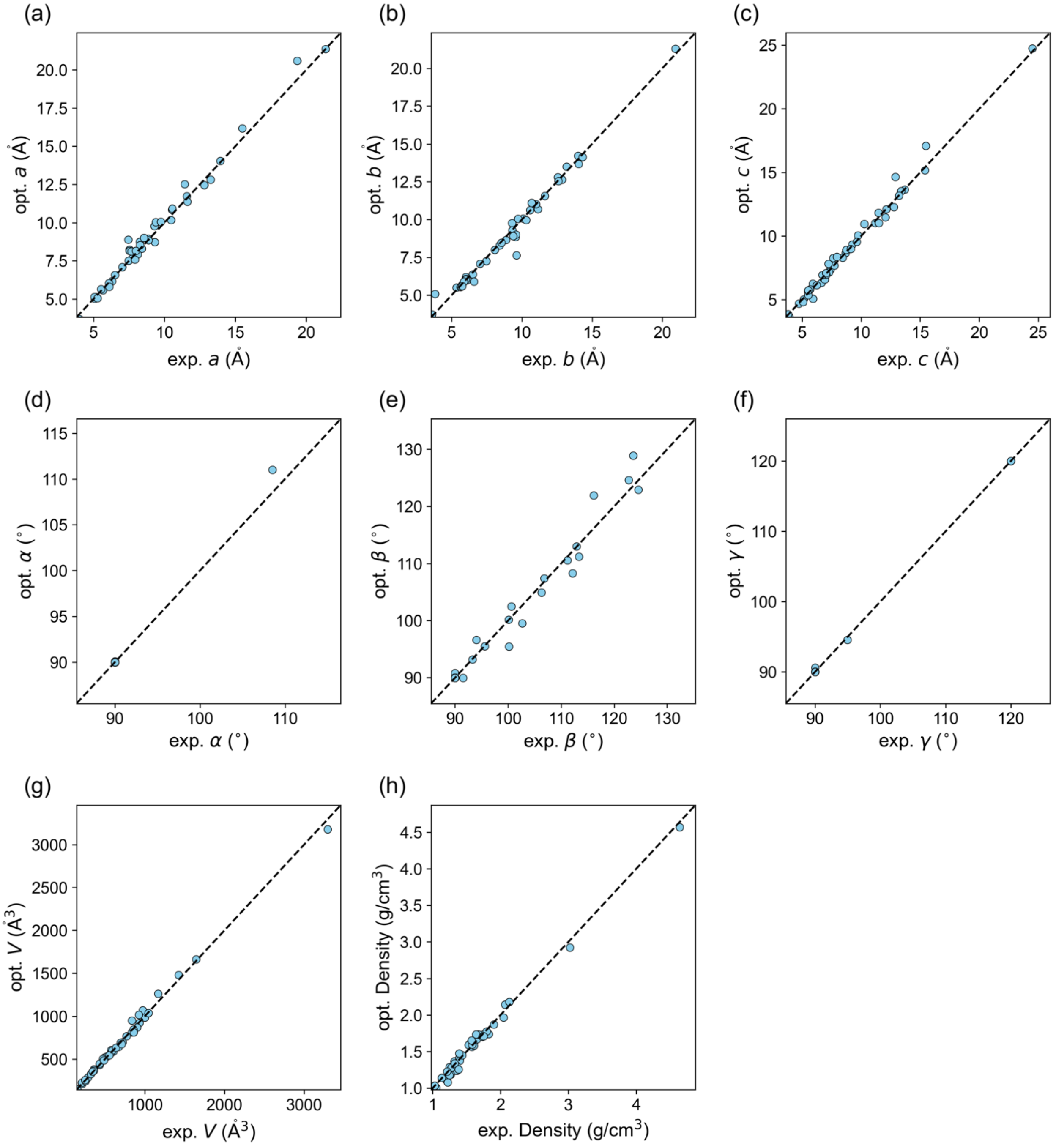

Figure S5. The lattice changes before and after the cell optimization using CHGNet tuned by knowledge distillation. Elasticity dataset was used for evaluation.

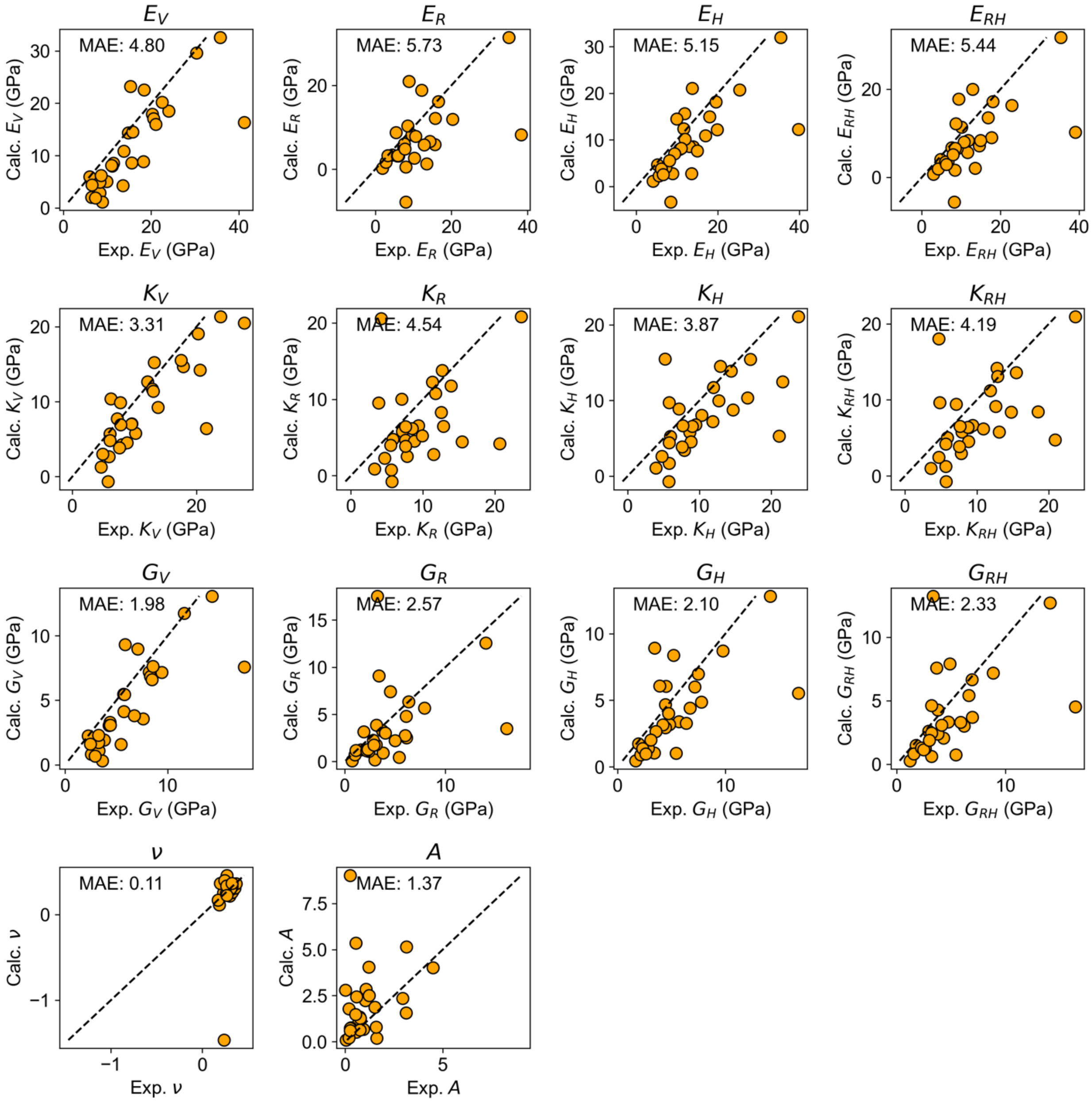

Figure S6. The observed-predicted plot of elastic properties calculated by pre-trained CHGNet.

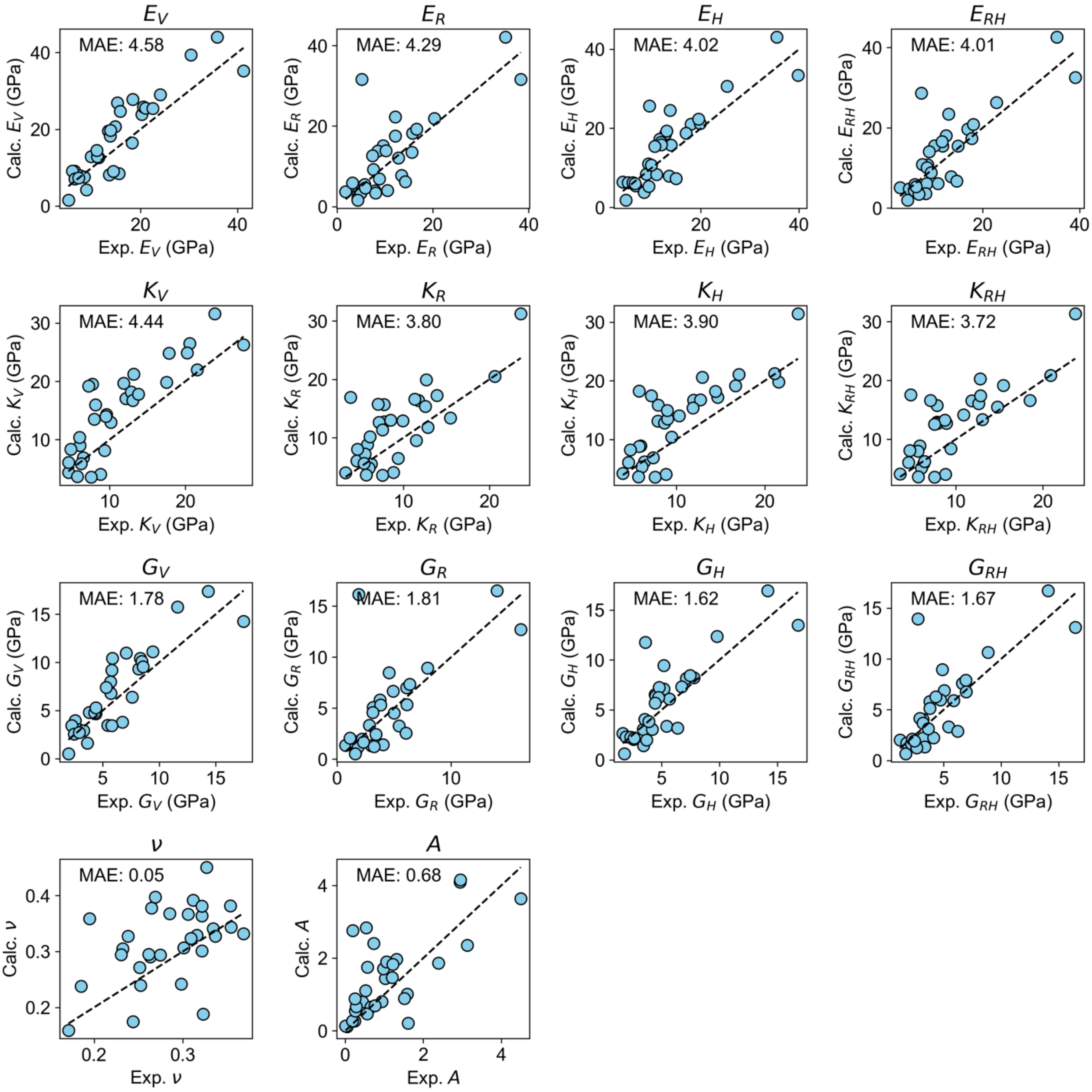

Figure S7. The observed-predicted plot of elastic properties calculated by CHGNet tuned by knowledge distillation.