Meta

Visit nima.sh/jmp and play with our interactive visualization

# From Molecules to Materials: Pre-training Large Generalizable Models for Atomic Property Prediction


**AUTHORS**

Nima Shoghi, Adeesh Kolluru, John R. Kitchin, Zachary W. Ulissi, C. Lawrence Zitnick, Brandon M. Wood

**AFFILIATIONS**

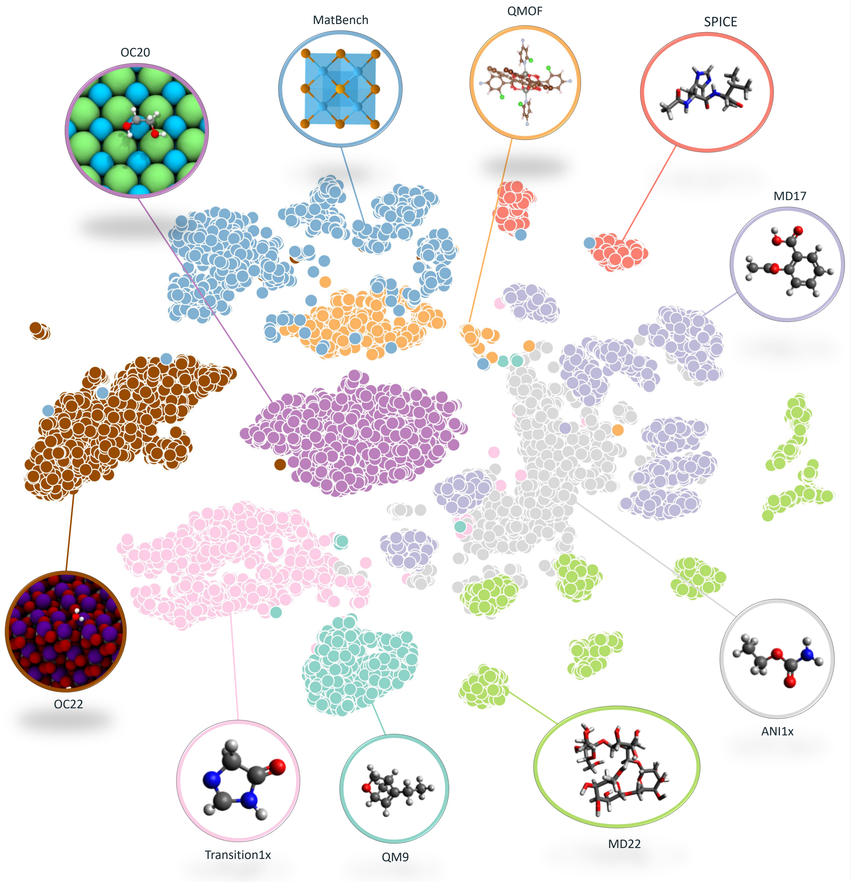
- Nima Shoghi, Zachary W. Ulissi, C. Lawrence Zitnick, Brandon M. Wood: Fundamental AI Research (FAIR) at Meta
- Nima Shoghi: Georgia Institute of Technology. Work done while at FAIR
- Adeesh Kolluru, John R. Kitchin: Carnegie Mellon University

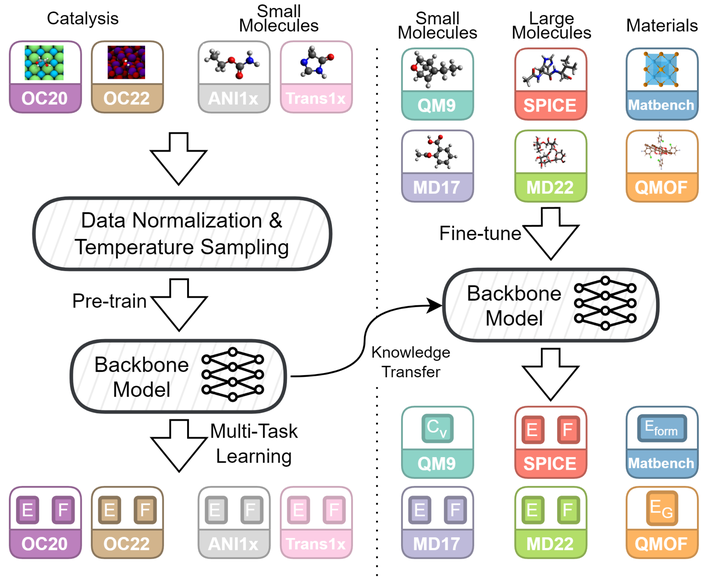

See our interactive visualization!

## Unlocking the Potential of Diverse Chemical Data: A Multi-Task Approach to Atomic Pre-training

- Inspired by the success of foundation models (FMs) in NLP & CV, we aim to train large generalizable models that are **widely useful across chemistry.**
- In chemistry, it is common for datasets to be created for specific chemical domains (e.g., small molecules, proteins, materials). **Unifying this fragmented data** is a core challenge to building more general models.
- We propose **Joint Multi-domain Pre-training (JMP):** Simultaneously pre-training on multiple datasets utilizing a multi-task learning framework.
- Using JMP, we demonstrate a **59% average improvement** over training from scratch and set **SOTA on 34/40** evaluated tasks.

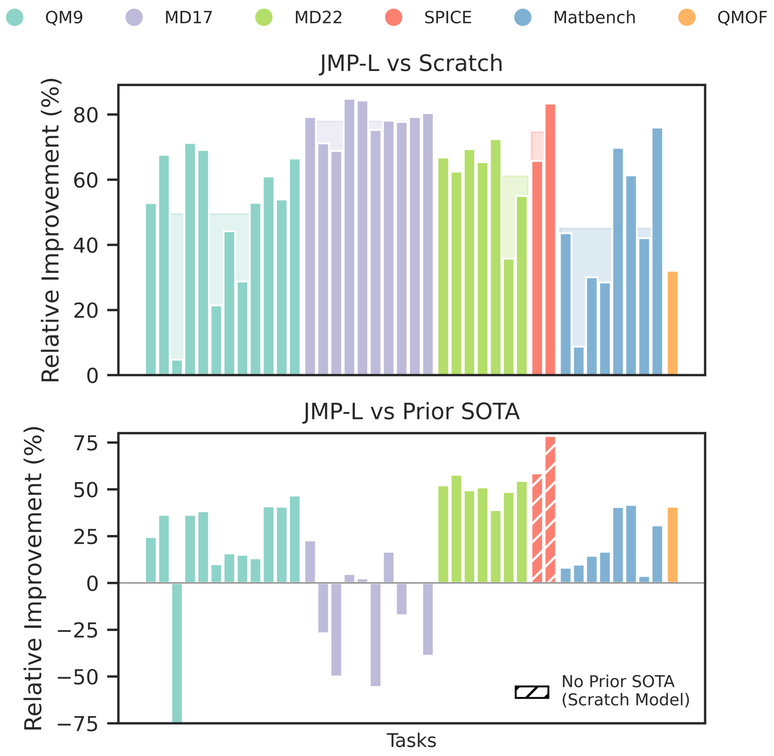
| Dataset | Domain | Labels | Elements | Avg size | Train Set | Description |
|---|---|---|---|---|---|---|
| **Pretraining Datasets** | | | | | | |
| OC20 | Catalyst | E, F | 55 | ~ 73 (7-225) | 100M | Catalyst relaxations |
| OC22 | Catalyst | E, F | 51 | ~ 80 (17-228) | 8M | Oxide catalyst relaxations |
| ANI-1x | Small Molecule | E, F | H, C, N, O | ~ 15 (4-63) | 2M | MD simulations |
| Transition-1x | Small Molecule | E, F | H, C, N, O | ~ 14 (4-23) | 10M | Reactions database |
| **Finetuning Datasets** | | | | | | |
| Matbench | Materials (OOD) | ID / OOD | 84 | ~30 (4-444) | ~600–130k | Material properties |
| QMOF | Materials (OOD) | OOD | 77 | ~109 (17, 500) | 10k | MOF properties |
| MD17 | Small Mols. (ID) | ID | H, C, N, O | ~13 (9-21) | 1k | MD simulation |
| QM9 | Small Mols. (ID) | ID / OOD | H, C, N, O | ~18 (3-29) | ~130k | QM properties |
| SPICE | Large Mols. (OOD) | ID | H, C, N, O, S | ~ 46 (26-96) | 1300, ~34k | MD simulations |
| MD22 | Large Mols. (OOD) | ID | H, C, N, O | ~67 (42-370) | ~600–8k | MD simulations |



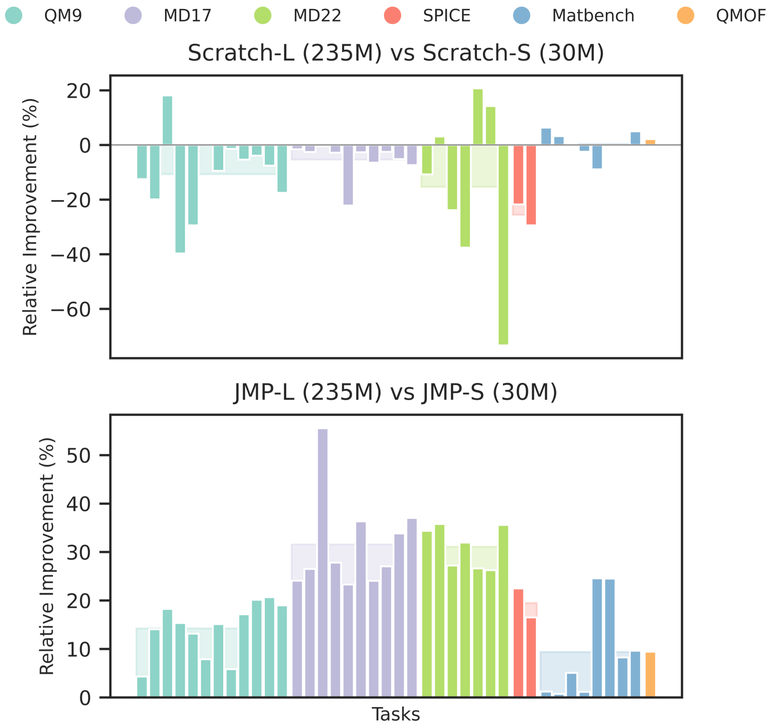




JMP concurrently trains on over 120 million diverse equilibrium and non-equilibrium atomic structures by framing each chemical domain as a separate pre-training task in a **multi-task framework.**

## JMP Improves Downstream Performance



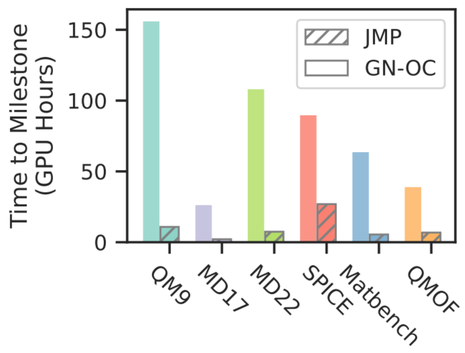

## JMP Enables Larger Models




We compare the relative improvement of JMP-L (235M) over JMP-S (30M) to the relative improvement of the scratch variants of the same models. On average, JMP shows a **21% improvement** in performance while the scratch model shows an **8% decrease** in performance.

## JMP Speeds Up Downstream Training

While JMP's pre-training is computationally expensive, this upfront cost is recovered by enabling over **12x faster fine-tuning** compared to training from scratch.




See our interactive visualization!

Georgia Institute of Technology

Carnegie Mellon University