Dr. Clara Kirkvold | 09.01.2025

Foundational Models: Can they do it all?

Laying the foundation

Foundational models (FM) that can make accurate predictions for a diverse set of tasks (*multimodal predictions*).

FMs are pre-trained on **HUGE** datasets. After this pretraining, the model can be finely tuned to perform more specific downstream tasks.

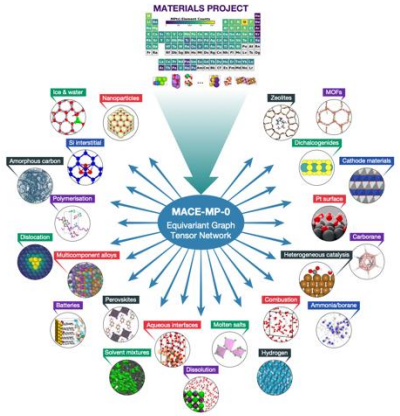

Table 1. Examples of foundational models for chemistry and materials science.

Model	Pretraining Dataset	Representation
SMI-TED289M[1]	91 million molecules from PubChem	SMILES-based
MOLFORMER[2]	1.1 billion molecules from PubChem and ZINC	SMILES-based
MACE-MP-0[3]	1.5 million configurations from Materials Project	Graph-based
CHGNet[4]	1.5 million configurations from Materials Project	Graph-based

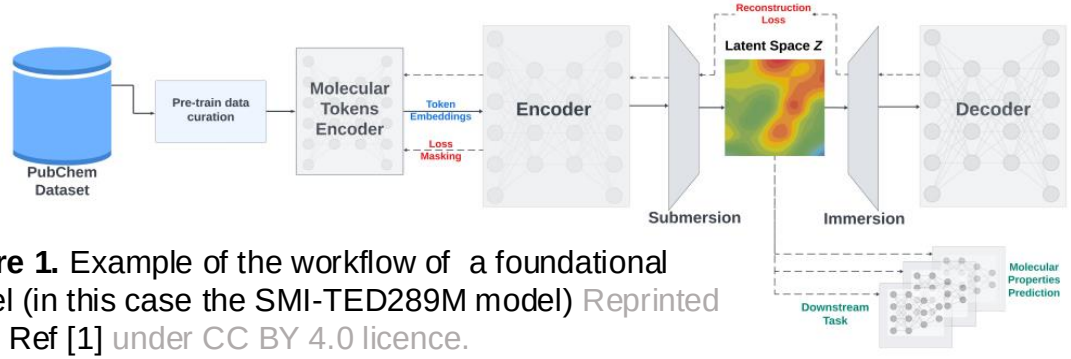

Figure 1. Example of the workflow of a foundational model (in this case the SMI-TED289M model) Reprinted from Ref [1] under CC BY 4.0 licence.

Performance on downstream tasks

FMs can be **fine-tuned** for specific tasks by training them on smaller, task-specific datasets.

This approach can be more data efficient and result in better performance (since the model leverages the broad knowledge it gained during pretraining) than a training new model from scratch.

Table 2. Performance SMI-TED289M model for regression task of benchmark datasets compared to other methods. Reprinted from Ref [1] under CC BY 4.0 licence.

Method	Dataset				
	QM9	QM8	ESOL	FreeSolv	Lipophilicity
D-MPNN [36]	3.241 ± 0.119	0.0143 ± 0.0022	0.98 ± 0.26	2.18 ± 0.91	0.65 ± 0.05
N-Gram [37]	2.51 ± 0.19	0.0320 ± 0.003	1.074 ± 0.107	2.688 ± 0.085	0.812 ± 0.028
PretrainGNN [38]	-	-	1.100 ± 0.006	2.764 ± 0.002	0.739 ± 0.003
$GROVER_{Large}$ [30]	-	-	0.895 ± 0.017	2.272 ± 0.051	0.823 ± 0.010
ChemBERTa-2 [32]	-	-	0.89	-	0.80
SPMM [35]	-	-	0.818 ± 0.008	1.907 ± 0.058	0.692 ± 0.008
$MolCLR_{GIN}$ [39]	2.357 ± 0.118	0.0174 ± 0.0013	1.11 ± 0.01	2.20 ± 0.20	0.65 ± 0.08
Hu et al. [40]	4.349 ± 0.061	0.0191 ± 0.0003	1.22 ± 0.02	2.83 ± 0.12	0.74 ± 0.00
MoLFormer [35]	1.5894 ± 0.0567	0.0102	0.880 ± 0.028	2.342 ± 0.052	0.700 ± 0.012
SMI-TED289M (Frozen Weights)	7.4883 ± 0.0659	0.0179 ± 0.0004	0.7045 ± 0.0344	1.668 ± 0.0616	0.6499 ± 0.012
SMI-TED289M (Fine-tuned)	**1.3246 ± 0.0157**	**0.0095 ± 0.0001**	**0.6112 ± 0.0096**	**1.2233 ± 0.0029**	**0.5522 ± 0.0194**

Prospects and limitations

Chemistry-focused FMs demonstrate versatility across chemical domains but face challenges. They rely on large datasets that are often limited, incomplete, or at low levels of theory. Their accuracy is constrained by their underlying architecture (e.g., neural network potentials often lack long-range electrostatics), and it is unclear if, when fine-tuning, they require less data than models trained from scratch, especially for complex chemistries beyond standard benchmarks.

[1] Soares *et al.* arXiv:2407.20267 (2022).
[2] Ross *et al*. *Nat Mach Intell* **4**, 1256–1264 (2022).
[3] Batatia *et al.* arXiv:2401.00096 (2024).
[4] Deng *et al. Nat Mach Intell* **5**, 1031–1041 (2023).