Materials Property Axiom: Scaling Foundation Models to Experimental Property Generalists via Multi-phase Training

Deep Principle Team

Large-scale experimental property prediction is becoming a central bottleneck in discovering novel materials, where the models are required to learn from heterogeneous experimental data and be rapidly adapted to new assays, measurements, and product-relevant endpoints. However, current materials foundation models focus primarily on computational properties, especially thermodynamics and stability. Inspired by the training paradigm of large language models, in which broad pre-training of all available corpus is followed by mid-training alignment of multiple high-quality sources and finally post-training with task-specific supervised fine tuning, we ask whether the same philosophy can improve experimental materials property prediction. Here we introduce Materials Property Axiom (MPA), a three-phase framework comprising general pre-training, domain-aligned mid-training, and downstream post-training. Across a vast validation of 40 experimental properties, MPA consistently claims state-of-the-art performance relative to direct fine-tuning from a pretrained model, improving mean absolute error by 15% on average and by up to 55% on select individual properties. This gap widens with a mid-training strategy where high-quality subsets from various experimental sources and increasing first-principle computational data are aligned, suggesting that mid-training captures a common structure across heterogeneous materials data that scales. Together, these findings establish multi-phase training as a general strategy for transforming materials foundation models into reliable surrogates of experimental properties for accelerating real-world materials discovery. MPA is ready to use in https://sciclaw.ai.

Date: June 1, 2026
Correspondence: Deep Principle Team
Use: https://sciclaw.ai

Deep Principle

1 Introduction

Deep learning has reshaped materials property prediction models and its use in drug discovery, molecular design, and materials discovery.[1–4] Materials foundation models extend this success by learning transferable representations through large-scale pre-training, which can then be adapted to downstream prediction tasks through fine-tuning.[5–7] As reusable backbones, these models improve data efficiency and generalization across a wide range of property prediction problems. Most effort in this area has been devoted to predicting computational properties such as thermal stability and absorption energy, with numerous foundation models developed via Materials Project, Open Catalyst Project, and Matbench Discovery.[8–11] These models, however, are still "one step away" from direct downstream applications in fulfilling the actual industry need of discovering new materials, where rapid assessment of a diverse set of experimental properties is essential.

Current materials foundation models differ mainly in how supervision enters representation learning. Some rely on structure-only objectives, such as masked-token prediction, autoregressive chemical language modeling, graph self-supervision, or three-dimensional reconstruction, as in MoLFormer, ChemFM, GROVER, and UniMol.[12–16] Others inject readily available chemical knowledge, including RDKit or Mordred descriptors, fingerprints, and generic properties, as in ChemBERTa, KPGT,

and CheMeleon.[17–19] A third group uses large-scale property supervision, including biological, phenotypic, or computational labels, to make representations more task-aware, as in MolE, MolGPS, and MoleVers.[20–22] Together, these studies show that property-related supervision can complement structural pre-training. Yet existing auxiliary tasks are typically selected for scale, availability, or broad coverage, rather than for physical alignment with a target experimental endpoint. Consequently, they offer limited guidance on which intermediate task should improve which measured property, or why. Because experimental endpoints often reflect specific thermodynamic, electronic, interfacial, or biological mechanisms, the value of physically aligned intermediate representation alignment remains poorly understood.[23] More importantly, representation learning alone does not determine property prediction: the readout and training objective also define how structure is mapped to a scalar endpoint. Because endpoints differ physically—some extensive, others intensive or local—even strong 3D pre-training can underperform when the adaptation data or readout bias is mismatched to the target property.

This motivates our central hypothesis: transfer across materials property tasks is governed by shared physics, not shared data statistics alone. This view parallels mid-training in large language models, where broad pre-training is followed by targeted intermediate training that aligns general representations with the capabilities needed for downstream use.[24, 25] For materials foundation models, the corresponding mid-training signal should be physical rather than merely generic: an auxiliary task should help a downstream endpoint when both depend on related thermodynamic, electronic, interfacial, or biological quantities, and a readout should help when its inductive bias matches the structure of the target property. If so, intermediate supervision and readout design can be chosen prospectively from physical principles, rather than discovered by trial and error over available labels. First-principle computation can then serve as physically aligned mid-training, bridging general materials representations to experimentally measured properties of practical interest.

We validate this principle through the largest systematic training of materials foundation models on experimental properties to date, spanning 40 experimentally measured datasets. Using a three-phase training pipeline—pre-training, physics-aligned mid-training, and downstream post-training—we show that transfer is not a generic consequence of more supervision, but a consequence of physical alignment: models improve when mid-training labels encode the same thermodynamic, electronic, interfacial, or solvation physics that controls the downstream endpoint, and when the readout matches the physical underlying embedding of the property. This converts first-principle computation from a source of generic auxiliary labels into targeted proxy supervision for scarce experimental measurements. The resulting Materials Property Axiom (MPA) model improves with data scale, reducing MAE by approximately 15% on average and by up to 55% under out-of-distribution evaluation relative to direct fine-tuning, and achieves state-of-the-art performance on 35/40 experimental properties. Together, these results establish physical alignment as a scaling principle for adapting materials foundation models to real-world experimental prediction and thus is anticipated to accelerate the discovery of new materials.

2 Three-stage training framework

The MPA framework consists of three stages (Fig. 1). First, a pretrained foundation model provides a transferable molecular backbone, initialized through geometry-based 3D pre-training (Fig. 1a). Second, a supervised intermediate stage, termed mid-training, adapts the backbone using inexpensive property-related supervision chosen to reflect the physics of the target endpoint (Fig. 1b). Third, the adapted model is fine-tuned on downstream datasets of experimentally measured properties, a final target-specific stage that we refer to as post-training (Fig. 1c).

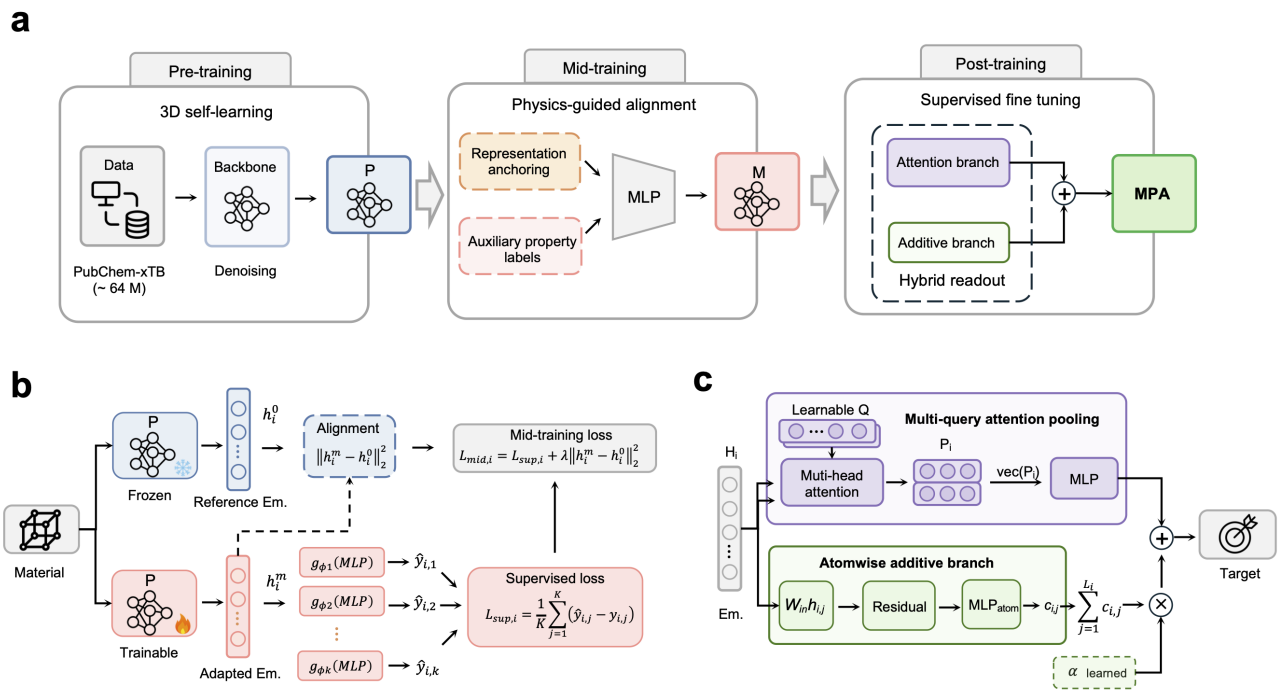

Fig. 1 | Overview of the Materials Property Axiom (MPA) framework. (a) MPA is organized as a three-stage adaptation workflow. Geometry-based 3D self-supervised pre-training initializes a molecular backbone from the PubChem-xTB dataset; mid-training aligns the representation with physics-relevant auxiliary property labels while anchoring it to the pretrained representation; and post-training adapts the model to experimentally measured downstream targets through supervised fine-tuning. The final readout combines an attention branch with an atom-wise additive branch. (b) Mid-training uses a frozen pretrained model to compute reference embeddings $\mathbf{h}_i^0$ and a trainable copy to compute adapted embeddings $\mathbf{h}_i^m$. Property-specific heads optimize auxiliary-label prediction, while the representation-anchoring term penalizes $\|\mathbf{h}_i^m - \mathbf{h}_i^0\|_2^2$ to preserve the pretrained chemical representation space. (c) The post-training readout combines multi-query attention pooling with an atom-wise additive branch weighted by a learnable scalar α. This hybrid design allows the model to combine non-additive molecular summaries with an explicit atom-wise summation prior for additive-compatible properties.

2.1 Pre-Train stage

The first stage is geometry-based 3D pre-training, which provides a general molecular representation before physics-guided adaptation. We adopt a conventional self-supervised pre-training procedure in which the model learns to recover corrupted molecular geometry and atom-level features from three-dimensional conformations. This objective encourages the backbone to capture both local chemical identity and global spatial organization. The pre-training corpus is PubChem-xTB, a self-constructed dataset of approximately 100 million molecular structures. Importantly, MPA does not rely on a specific external pretrained model or on a particular implementation of the pre-training architecture. Instead, the pre-training stage serves as a generic 3D molecular initialization, after which mid-training and post-tuning introduce the physics-guided components of the framework.

2.2 Supervised Mid-Training for Physics-Guided Adaptation

Mid-training adapts a pretrained molecular backbone to auxiliary property supervision before target-specific post-tuning. For molecule i, let $\mathbf{h}_i^0$ denote the fixed representation produced by the initial pretrained backbone, and let $\mathbf{h}_i^m$ denote the representation produced by the backbone after mid-training. Given a scalar prediction head $g_\phi(\cdot)$, the auxiliary-property prediction is

$$\hat{y}_i = g_\phi\left(\mathbf{h}_i^m\right), \tag{1}$$

and the single-property mid-training objective is

$$\mathcal{L}_{\text{mid}} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i - y_i\right)^2 + \lambda\frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{h}_i^m - \mathbf{h}_i^0\right\|_2^2. \tag{2}$$

Here y_i is the auxiliary-property label, N is the number of mid-training samples, and λ controls the strength of the representation-anchoring regularizer. The representation $\mathbf{h}_i^0$ is computed from the initial pretrained backbone and kept fixed during optimization, while $\mathbf{h}_i^m$ is updated through the mid-trained backbone. The head $g_\phi(\cdot)$ is implemented as a multilayer perceptron.

The first term in Eq. (2) is the supervised regression loss on the auxiliary property. The second term anchors the mid-trained representation to the original pretrained representation, discouraging unconstrained drift from the chemical and geometric representation space learned during pre-training. This regularization preserves general transferability while still allowing the backbone to adapt along directions supported by the mid-training signal.

Some mid-training sources contain a single auxiliary label, whereas others provide multiple property labels for the same molecule. We therefore consider two head configurations. In the single-head setting, each auxiliary property is used in isolation with one scalar prediction head, allowing us to measure the transfer effect of that property separately. In the multi-head setting, a shared backbone is optimized jointly with multiple property-specific heads, so that several auxiliary signals can shape the same representation. For K auxiliary properties, the multi-head prediction for task k is

$$\hat{y}_{ik} = g_{\phi_k}\left(\mathbf{h}_i^m\right), \quad k = 1, \ldots, K, \tag{3}$$

where $g_{\phi_k}(\cdot)$ is the prediction head for property k. When labels are missing for some molecule–property pairs, we use a binary mask M_{ik} and define the supervised multi-head loss as

$$\mathcal{L}_{\text{sup}}^{\text{multi}} = \frac{1}{\sum_{i=1}^{N}\sum_{k=1}^{K} M_{ik}}\sum_{i=1}^{N}\sum_{k=1}^{K} M_{ik}\left(\hat{y}_{ik} - y_{ik}\right)^2. \tag{4}$$

The full multi-head mid-training objective is obtained by replacing the first term in Eq. (2) with Eq. (4), while retaining the same representation-anchoring regularizer.

2.3 Post-Tuning for Downstream Property Prediction

In the post-tuning stage, the mid-trained backbone is adapted to each downstream experimentally measured property. A standard protocol for pretrained molecular models is to attach an MLP readout and optimize a mean-squared-error objective. For a downstream dataset with targets y_i, this baseline objective can be written as

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i - y_i\right)^2, \tag{5}$$

where $\hat{y}_i$ is the predicted value for molecule i and N is the number of training samples.

We modify this post-tuning stage in two aspects. First, we replace MSE with the Huber loss,[26] which behaves quadratically for small residuals and linearly for large residuals. This provides a smooth approximation to MAE while reducing the sensitivity of training to outliers. To make the loss scale comparable across endpoints, all downstream targets are standardized within each training fold:

$$\tilde{y}_i = \frac{y_i - \mu_{\text{train}}}{\sigma_{\text{train}}}, \tag{6}$$

where μ_{train} and σ_{train} are computed only from the training split. The model is trained to predict $\hat{\tilde{y}}_i$, and the Huber loss is applied in standardized units:

$$\mathcal{L}_{\text{Huber}} = \frac{1}{N}\sum_{i=1}^{N} \ell_\delta\left(\hat{\tilde{y}}_i - \tilde{y}_i\right), \tag{7}$$

with

$$\ell_\delta(r) = \begin{cases} \frac{1}{2}r^2, & |r| \le \delta, \\ \delta\left(|r| - \frac{1}{2}\delta\right), & |r| > \delta. \end{cases} \tag{8}$$

In our experiments, we set $\delta = 1$, so the transition from quadratic to linear penalty occurs at one standard deviation of the training targets. Reported errors are computed after transforming predictions back to the original physical units.

Second, we replace the standard MLP readout with a hybrid readout that combines a normalized attention-pooling branch with an atomwise additive branch. Given the atom-level backbone features of molecule i,

$$\mathbf{H}_i = \{\mathbf{h}_{i,j}\}_{j=1}^{L_i},$$

where L_i is the number of valid atoms, the hybrid readout predicts the standardized target as

$$\hat{\tilde{y}}_i = g_{\text{attn}}(\mathbf{H}_i) + \alpha\, g_{\text{add}}(\mathbf{H}_i), \tag{9}$$

where $g_{\text{attn}}(\cdot)$ is a multi-query attention-pooling branch, $g_{\text{add}}(\cdot)$ is an atomwise additive branch, and α is a learnable scalar initialized to $\alpha_{\text{init}} = 0.1$. Setting $\alpha = 0$ recovers the attention-only readout, which we use as the reference configuration in the readout ablation.

The attention branch uses Q learnable query vectors to pool information from the atom tokens through multi-head attention:[27]

$$\mathbf{P}_i = \text{MHA}\left(\mathbf{Q}, \mathbf{H}_i, \mathbf{H}_i\right), \qquad \mathbf{P}_i \in \mathbb{R}^{Q\times d}, \tag{10}$$

where $\mathbf{Q} \in \mathbb{R}^{Q\times d}$ denotes the learnable queries and $\text{MHA}(\cdot)$ denotes multi-head attention. The resulting query-specific pooled representations are concatenated and passed through an MLP:

$$g_{\text{attn}}(\mathbf{H}_i) = \text{MLP}_{\text{attn}}\left(\text{vec}(\mathbf{P}_i)\right), \tag{11}$$

where $\text{vec}(\mathbf{P}_i)$ concatenates the Q pooled query representations into a single vector. Because attention pooling uses normalized weights over atom tokens, this branch provides a flexible non-additive summary of the molecule, but does not impose size extensivity.

The additive branch instead imposes an explicit atomwise decomposition. Each atom representation is first projected into a hidden space and updated through residual MLP blocks:

$$\mathbf{z}_{i,j}^{(0)} = \mathbf{W}_{\text{in}}\mathbf{h}_{i,j}, \qquad \mathbf{z}_{i,j}^{(b+1)} = \mathbf{z}_{i,j}^{(b)} + F_b\left(\mathbf{z}_{i,j}^{(b)}\right), \quad b = 0, \ldots, B-1, \tag{12}$$

where $F_b(\cdot)$ denotes the b-th residual MLP block. The atomwise contributions are then summed over valid atoms:

$$g_{\text{add}}(\mathbf{H}_i) = \sum_{j=1}^{L_i} \text{MLP}_{\text{atom}}\left(\mathbf{z}_{i,j}^{(B)}\right). \tag{13}$$

The two branches encode complementary physical priors. The attention branch provides a normalized, non-additive molecular summary that is suitable for intensive or emergent properties. The additive branch enforces a size-extensive form, making it appropriate for properties that scale approximately additively with molecular composition. The learnable coefficient α allows the model to control how strongly the extensive prior contributes for each endpoint. Thus, the post-tuning readout is not only a prediction head, but also a mechanism for matching the inductive bias of the model to the physical structure of the target property.

3 Data curation

3.1 Pre-Training data generation

The pretraining dataset was constructed from the PubChem database.[28] Compared with the full dataset, we restricted the molecular space to compounds whose SMILES strings contained only the elements C, H, O, N, P, S, F, Cl, Br, I, Si, and B. All 3D structures were regenerated from the corresponding SMILES representations. Initial conformers were generated using the ETKDGv3 algorithm implemented in RDKit[29, 30], followed by conformer–rotamer ensemble sampling with CREST[31] at the GFN-FF level[32]. The lowest-energy conformer from each ensemble was subsequently optimized at the GFN2-xTB level[33]. Overall, this workflow yielded optimized geometries for approximately 64 million molecules.

3.2 Mid-Training Data Curation

To test whether physically aligned auxiliary supervision can guide transfer, we constructed mid-training datasets that share a common molecular basis but differ in the physical information they provide. We first curated an experimental $\log P$ dataset from OPERA,[34] removed molecules that overlapped with the downstream benchmark, and retained approximately 10,000 molecules. This dataset provides a lipophilicity-oriented mid-training signal.

For the same molecules, we generated thermochemical labels using Taffi Component Increment Theory (TCIT),[35, 36] including the enthalpy of formation at 0K and 298K, the Gibbs free energy of formation at 298K, the heat capacity at 298K, and the standard entropy at 298K. These labels provide a thermochemistry-oriented source of auxiliary supervision. We further computed dipole moments and HOMO-LUMO gaps for the same molecular set using g-xTB,[37] providing an electronic-structure-oriented signal.

In addition to these matched-label datasets, we assembled a larger TCIT thermochemistry set containing approximately 4 million molecules. This larger set allows us to test whether scale improves transfer when the auxiliary labels are physically relevant to the downstream endpoint. Each mid-training source is used both individually, yielding single-source mid-training, and in combination, yielding multi-source mid-training. This design lets us distinguish whether downstream gains arise from generic additional supervision or from physically aligned auxiliary information.

3.3 Benchmark Curation

Data collection and task organization. We constructed a downstream benchmark designed to evaluate molecular property prediction across a broad range of physical regimes, rather than around a single endpoint family. The benchmark aggregates experimentally measured or experimentally grounded molecular property records from public databases, handbook compilations, and literature-derived collections.[34, 38–47] We then applied a unified curation pipeline to canonicalize molecular identifiers, remove duplicates and ambiguous records, reduce label noise, and harmonize property units where necessary.

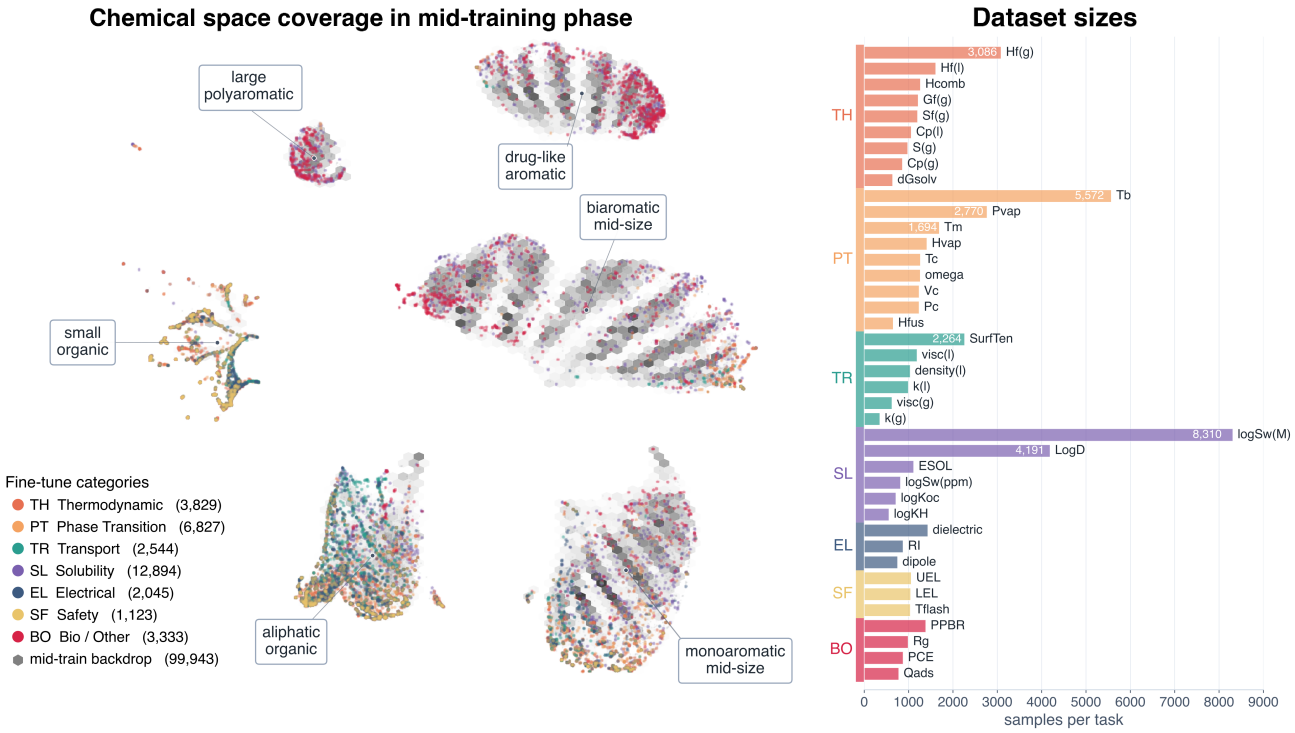

Fig. 2 | Chemical-space coverage and downstream benchmark composition. (Left) UMAP projection[48] of molecular fingerprints[49] for the mid-training data (sampled 10k) and downstream fine-tuning molecules. The gray background represents the mid-training chemical universe, while colored points indicate molecules from downstream property categories. The downstream benchmark spans several chemically distinct regions, including small organics, aliphatic molecules, monoaromatic and biaromatic mid-size compounds, drug-like aromatics, and larger polyaromatic structures, while remaining embedded within the broader mid-training chemical space. (Right) Sample counts for downstream tasks grouped by physical property category. Dataset sizes vary substantially across endpoints, from small safety, electrical, and transport datasets to larger boiling-point and aqueous-solubility datasets. This heterogeneity motivates a unified transfer-aware evaluation protocol across both random and scaffold splits.

The final benchmark contains 40 scalar property prediction tasks. We organize these tasks into seven physically motivated property families: thermodynamic properties (TH), phase-transition properties (PT), density and transport properties (TR), solubility and partitioning properties (SL), electrical and polarization properties (EL), safety and flammability properties (SF), and bio/other molecular-property proxies (BO). This grouping is used throughout the analysis to test whether transfer follows physical relatedness between the mid-training signal and the downstream endpoint. The categories are not intended as a strict ontology of molecular properties; rather, they provide a consistent physical

organization for comparing transfer behavior across endpoints with different governing mechanisms.

Table 1. Summary of the 40 downstream molecular property prediction tasks included in the curated benchmark.

Group	Property	Abbrev.	Additive	N	Units
TH	Gas-phase formation enthalpy	$H_f(g)$	Yes	3,085	kJ/mol
TH	Liquid-phase formation enthalpy	$H_f(l)$	Yes	1,613	kJ/mol
TH	Gas-phase formation Gibbs free energy	$G_f(g)$	Yes	1,216	kJ/mol
TH	Gas-phase entropy	S(g)	Yes	980	J/(mol·K)
TH	Gas-phase entropy of formation	$S_f(g)$	Yes	1,200	J/(mol·K)
TH	Heat of combustion	H_{combus}	Yes	1,266	kJ/mol
TH	Gas-phase heat capacity at 298 K	$C_p(g)$	Yes	861	J/(mol·K)
TH	Liquid-phase heat capacity at 298 K	$C_p(l)$	Yes	1,060	J/(mol·K)
TH	Solvation free energy	dG_{solv}	Yes	641	kcal/mol
PT	Boiling point	T_b	Yes	5,570	K
PT	Critical temperature	T_c	No	1,266	K
PT	Critical pressure	P_c	No	1,234	bar
PT	Critical molar volume	V_c	Yes	1,235	cm^3/mol
PT	Acentric factor	omega	No	1,263	—
PT	Log vapor pressure	P_{vap}	No	2,763	$\log_{10}$(mmHg)
PT	Enthalpy of vaporization at boiling point	H_{vap}	Yes	1,412	kJ/mol
PT	Melting point	T_m	No	1,693	K
PT	Enthalpy of fusion at melting point	H_{fuse}	Yes	653	kJ/mol
TR	Liquid density at 298 K	density(l)	No	1,033	g/cm^3
TR	Gas viscosity at 298 K	visc(g)	Yes	626	μPa·s
TR	Liquid viscosity at 298 K	visc(l)	Yes	1,188	cP
TR	Surface tension at 298 K	SurTen	Yes	2,262	mN/m
TR	Gas thermal conductivity at 298 K	k(g)	Yes	353	W/(m·K)
TR	Liquid thermal conductivity at 298 K	k(l)	No	992	W/(m·K)
SL	ESOL log solubility	ESOL	Yes	1,115	log mol/L
SL	Log Henry constant	logKH	Yes	558	log atm· m^3/mol
SL	Aqueous solubility	logSw(M)	Yes	8,305	log mol/L
SL	Aqueous solubility	logSw(ppm)	Yes	815	log ppm
SL	Lipophilicity	logD	No	4,191	log D
SL	Soil organic carbon partition coefficient	logKoc	No	710	log Koc
EL	Refractive index at 298 K	RI	No	876	—
EL	Dielectric constant at 298 K	dielectric	No	1,432	—
EL	Dipole moment	dipole	Yes	753	Debye
SF	Flash point	T_{flash}	Yes	1,039	K
SF	Lower explosive limit	LEL	No	1,046	vol %
SF	Upper explosive limit	UEL	No	1,059	vol %
BO	Plasma protein binding rate	PPBR	No	1,386	%
BO	Cell entry permeability proxy	PCE	No	875	nm/s
BO	Radius of gyration	Rg	No	989	Å
BO	Adsorption capacity at 10 ppmv	Q_{ads}	Yes	778	mg/g

Figure 2 summarizes both the chemical-space coverage and the task composition of the benchmark. The downstream molecules occupy multiple chemically distinct regions, including small organics, aliphatic compounds, monoaromatic and biaromatic mid-size molecules, drug-like aromatic compounds, and larger polyaromatic structures. At the same time, they remain embedded within the broader mid-training chemical universe, allowing us to study transfer from physically motivated auxiliary supervision without restricting evaluation to a narrow chemical domain. Dataset sizes vary

substantially across tasks, from hundreds to several thousand molecules per endpoint, making a unified per-task evaluation protocol essential for comparing performance across property families without allowing the largest datasets to dominate the conclusions.

Data splits. For each downstream property, we evaluate all models using the same five-fold cross-validation protocol under two partitioning schemes: a random split and a scaffold split. The random split measures in-distribution interpolation, where training and test molecules are sampled from the same property dataset without enforcing structural separation. The scaffold split provides a harder out-of-distribution evaluation by testing whether a model can generalize to molecular scaffolds that were not observed during training. This paired protocol is important because random splits can place structurally similar analogues in both training and test folds, often producing optimistic estimates of practical predictive performance.

The scaffold split enforces structural separation using Bemis–Murcko frameworks.[50] Molecules sharing the same scaffold are assigned to the same fold, so that no scaffold appears in both the training and test sets within a given cross-validation split. To improve fold balance while preserving structural separation, scaffolds are grouped by fingerprint similarity and then allocated across folds with a preference for comparable fold sizes. Thus, the two split types provide complementary views of model behavior: random splits assess conventional in-distribution accuracy, whereas scaffold splits assess robustness under scaffold-level distribution shift.

Applying the same split construction to all 40 tasks enables direct comparison across property families, mid-training sources, readout architectures, and external models. This uniform evaluation is central to the benchmark: it allows performance differences to be interpreted as differences in transfer behavior, rather than as artifacts of inconsistent dataset splitting or endpoint-specific evaluation practice.

4 Results and Discussion

4.1 MPA converts geometry pre-training into broad downstream gains

We first ask what the Materials Property Axiom (MPA) adds beyond geometry-based molecular pre-training alone. The comparison baseline is a pretrained-only model in which the same MPA initialization is fine-tuned directly on each downstream endpoint, without physics-aligned mid-training and without the MPA post-training modifications. This matched comparison isolates the two components that define MPA after pre-training, namely physically selected intermediate supervision and a physically structured readout.

MPA substantially improves over this pretrained-only adaptation baseline under both evaluation regimes (Fig. 3). Under the random split, MPA improves 38 of 40 comparable endpoints, reduces mean MAE by 14.0%, and achieves a median MAE reduction of 10.4%. Under the scaffold split, MPA again improves 38 of 40 endpoints, with a mean MAE reduction of 14.6% and a median reduction of 13.5%. The stronger median improvement under scaffold evaluation is the more consequential result. Random splits primarily measure interpolation among structurally related molecules, whereas scaffold splits test generalization to new molecular families. MPA therefore gains most in the regime where transferable physical information is most needed.

The improvements are broad rather than concentrated in a few favorable endpoints. As summarized in Table 2, MPA improves every thermodynamic, solubility/partitioning, electrical, and bio/other endpoint under the scaffold split, and improves nearly all phase-transition and transport endpoints. The largest gains occur in properties whose physical structure is most compatible with the MPA design, especially additive-compatible thermodynamic quantities and size- or composition-correlated

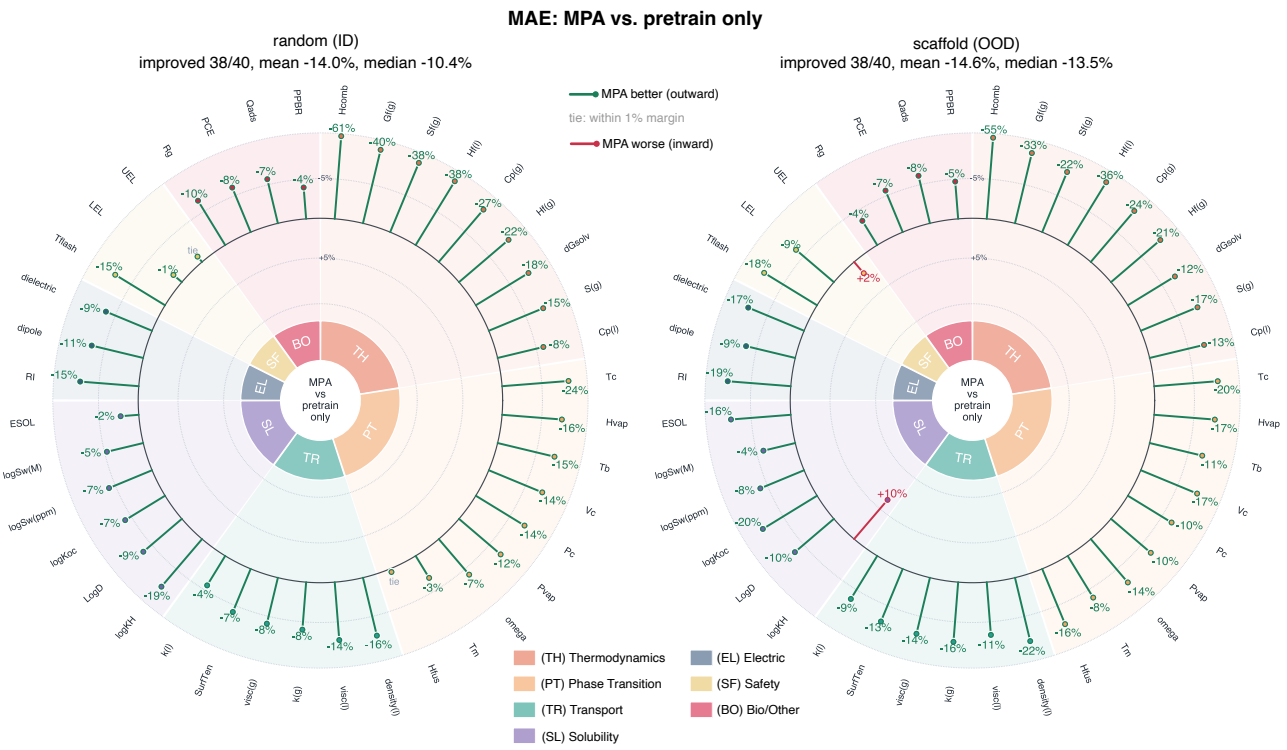

Fig. 3 | MPA improves broadly over the pretrained-only adaptation baseline. Each radial spoke reports the relative MAE change of MPA compared with direct target fine-tuning from the geometry-pretrained MPA initialization, without physics-aligned mid-training or the MPA post-training modifications. Properties are grouped by physical category. Green outward segments indicate lower MAE for MPA, red inward segments indicate higher MAE, and pale markers denote ties with $|\Delta\text{MAE}| < 1\%$. The radial axis uses a symmetric-log scale, with a linear region near the baseline and a shared cap across panels. MPA improves 38 of 40 comparable endpoints under both random and scaffold evaluation. Under the random split, MPA reduces mean MAE by 14.0% with a median reduction of 10.4%. Under the scaffold split, it reduces mean MAE by 14.6% with a median reduction of 13.5%, indicating that physics-guided adaptation is especially valuable under scaffold-level distribution shift.

Table 2. Category-level coverage of MPA improvements relative to the pretrained-only adaptation baseline. Improved counts are reported within each physical property family. Ties are defined as $|\Delta\text{MAE}| < 1\%$. The overall row reports the mean and median relative MAE reduction across all 40 comparable endpoints.

Category	N	Random	Scaffold
Thermodynamic (TH)	9	9/9 improved	9/9 improved
Phase transition (PT)	9	8/9 improved, 1 tie	9/9 improved
Transport (TR)	6	6/6 improved	5/6 improved
Solubility/partitioning (SL)	6	6/6 improved	6/6 improved
Electrical/polarization (EL)	3	3/3 improved	3/3 improved
Safety/flammability (SF)	3	2/3 improved, 1 tie	2/3 improved
Bio/other (BO)	4	4/4 improved	4/4 improved
Overall	40	38/40 improved; mean -14.0%	38/40 improved; mean -14.6%

phase-transition endpoints. At the same time, the improvement extends across transport, solubility, electrical, safety, and bio/pharma-related properties, showing that MPA is not merely a specialist model for thermochemistry. It is a general adaptation strategy for converting a geometry-pretrained molecular representation into a stronger downstream predictor.

The few exceptions are also informative. Under the random split, the only non-improvements are ties within the $|\Delta\text{MAE}| < 1\%$ threshold. Under the scaffold split, two endpoints degrade, `k(l)` and `UEL`. These are both cases where the target mechanism is not naturally captured by a simple atom-wise-additive or local group-contribution structure. Their behavior reinforces the central claim of MPA. Physical alignment governs not only when the model improves, but also where a given inductive bias is inappropriate.

4.2 MPA reaches state-of-the-art performance

We next compare MPA with external molecular prediction models under the same random and scaffold protocols. The comparison spans SMILES-based language modeling,[18] 2D graph message passing,[51] descriptor-augmented foundation modeling,[19] Uni-Mol-style 3D molecular pre-training,[16] and the most recent Suiren molecular foundation model.[52] Across the 40-endpoint benchmark, MPA achieves the strongest aggregate performance under both ID random and OOD scaffold splits (Fig. 4). Under the random split, MPA is best on 21 of 40 endpoints, followed by Suiren on 16 and Uni-Mol2 on 3, while ChemBERTa, Chemprop, and CheMeleon do not rank first on any endpoint. Under the scaffold split, the ranking becomes far more decisive, with MPA gets best results on 35 of 40 endpoints, Suiren on 4, and Uni-Mol2 on 1. The random split thus reflects a competitive foundation-model regime, whereas the scaffold split consolidates the benchmark around MPA.

Table 3. Best-model win counts in the cross-model benchmark. A win means that the model achieves the lowest MAE among all evaluated models for a given endpoint and split. Counts are computed over the same 40 endpoints shown in Fig. 4.

Model	Random / ID wins	Scaffold / OOD wins
MPA	**21/40**	**35/40**
Suiren	16/40	4/40
Uni-Mol2	3/40	1/40
ChemBERTa	0/40	0/40
Chemprop	0/40	0/40
CheMeleon	0/40	0/40

The category-level pattern shows that MPA is not merely accumulating marginal wins on easy endpoints. Its strongest scaffold-split behavior appears in families where physical structure matters most, namely thermodynamic, transport, solubility/partitioning, electrical, and bio/other endpoints. Against Suiren, the competition is closer under the random split, but the scaffold split shifts the advantage toward MPA, especially where geometry, polarity, intermolecular cohesion, or additive-compatible structure are central. This ID-to-OOD reversal is consistent with the central mechanism, since physically aligned mid-training and a physically structured readout become more valuable as structural novelty increases.

The comparison also highlights a natural limitation of lower-dimensional molecular representations. SMILES and 2D graph models can be highly effective, but their inputs do not explicitly encode conformational geometry, spatial charge distribution, or orientation-dependent interactions. For geometry-sensitive endpoints such as dipole moment, thermodynamic quantities, phase-transition properties, and transport-related measurements, these physical variables must be inferred indirectly

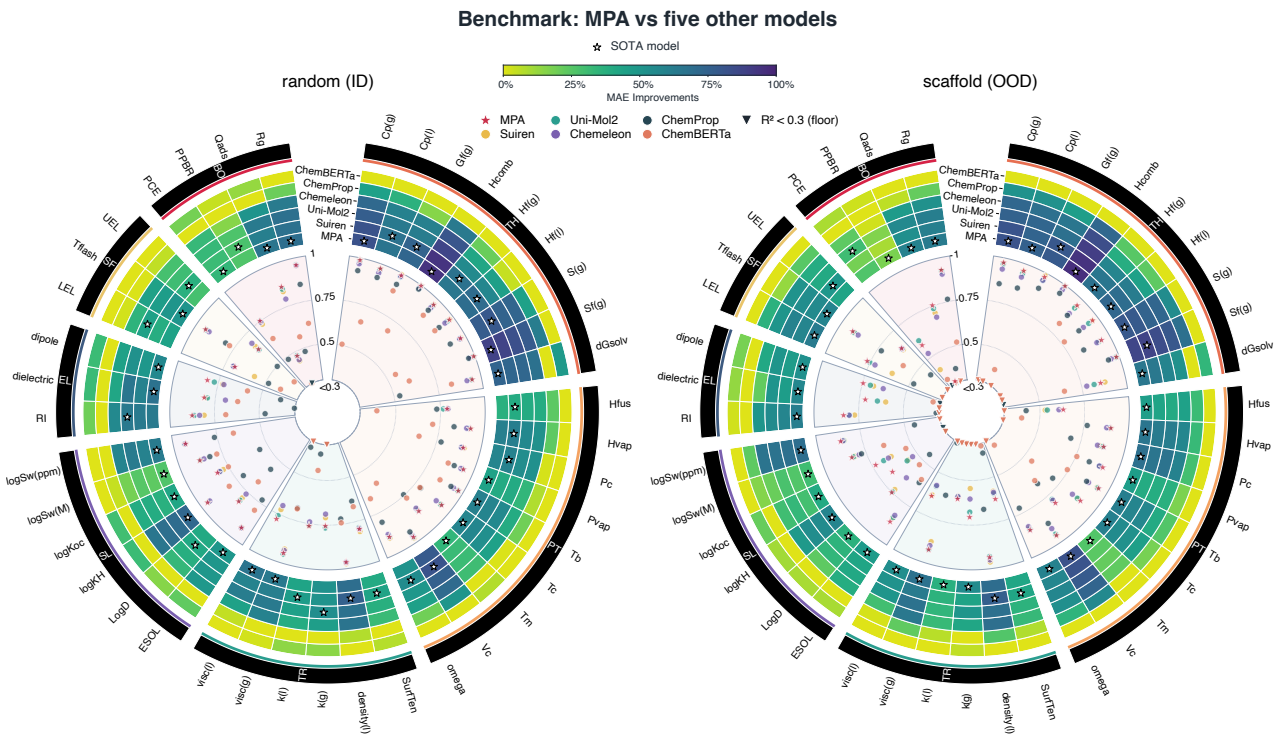

Fig. 4 | MPA reaches state-of-the-art performance under both in-distribution (ID) and out-of-domain (OOD) molecular endpoint evaluation. The cross-model benchmark compares MPA with five molecular prediction models under random and scaffold splits. Each sector corresponds to one downstream endpoint group. The outer ring identifies the endpoint category, the middle heatmap rings summarize normalized MAE improvement relative to the worst model for each endpoint, and the inner dots report R^2 values, with low-R^2 cases pinned to the plotting floor. Stars mark the best-performing model for each endpoint. Across the 40-endpoint benchmark, MPA is the most frequent best model under both random and scaffold evaluation, with its advantage becoming dominant under scaffold-level distribution shift.

from sequence or graph statistics. MPA reduces this information bottleneck by starting from geometry-based 3D pre-training and then aligning both intermediate supervision and readout structure with the physical form of the target endpoint.

MPA also shows the most robustness to scaffold-level shift. Among the foundation-model-class methods, it shows the smallest average MAE degradation when moving from random to scaffold evaluation: 25.7%, compared with 29.5% for the pretrained-only adaptation baseline, 29.6% for CheMeleon, and 31.8% for Suiren. MPA therefore combines high absolute accuracy with strong scaffold-level robustness, the combination needed for prospective molecular discovery.

4.3 Ablation studies reveal the two mechanisms of MPA

Having established the full-system performance of MPA, we isolate the two components that distinguish it from pretrained-only adaptation, physics-guided mid-training and physically structured post-training. Mid-training controls what physical information enters the representation, whereas post-training controls how that representation is converted into a scalar prediction. The ablation tests both whether each component improves performance and whether the improvements follow the physical-alignment principle. The gains in Fig. 5 are broad but uneven. They vary by endpoint family, indicating that each component helps most when its physical content or inductive bias matches the

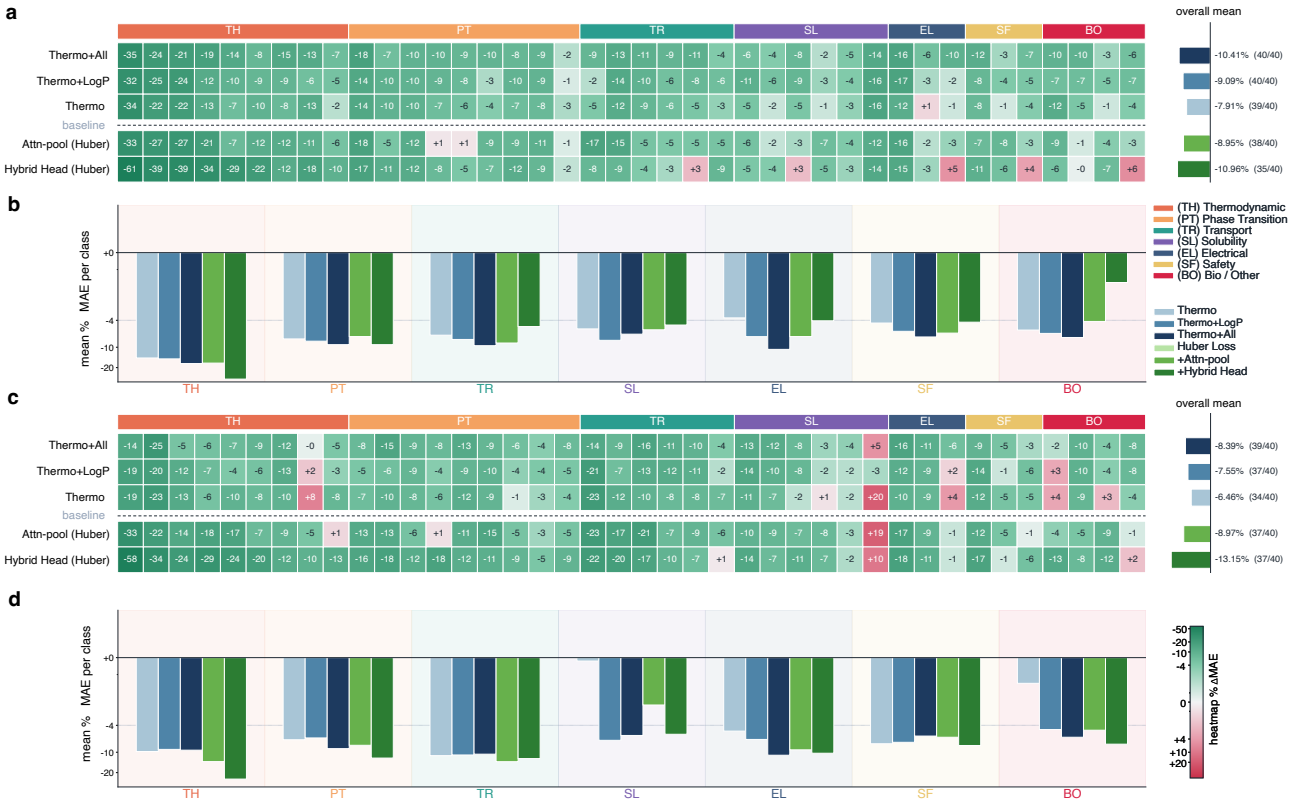

Fig. 5 | Ablation analysis of the MPA components under random and scaffold evaluation. (a,c) Per-endpoint heatmaps of relative MAE change with respect to the corresponding ablation baseline under random and scaffold splits, respectively. Rows above the dashed baseline compare mid-training variants: thermochemical supervision alone, thermochemistry plus logP, and thermochemistry plus all auxiliary property signals. Rows below the dashed baseline compare post-training variants: attention pooling with Huber loss and the hybrid readout with Huber loss. Green cells indicate lower MAE and red cells indicate higher MAE. The bars on the right report the overall mean relative MAE change and the number of improved endpoints for each setting. Endpoints share the same left-to-right order in both panels, grouped by family: thermodynamic (Hcomb, Hf(l), Gf(g), Sf(g), Cp(g), Hf(g), S(g), dGsolv, Cp(l)); phase transition (Tc, Hvap, Pc, Vc, ω, Hfus, Tb, Pvap, Tm); transport (density(l), k(g), visc(g), SurfTen, k(l), visc(l)); solubility (logKoc, ESOL, logSw(ppm), LogD, logSw(M), logKH); electrical (RI, dielectric, dipole); safety (Tflash, LEL, UEL); and bio/other (Rg, PCE, Qads, PPBR). (b,d) Endpoint-family averages under random and scaffold splits. The structured pattern of gains shows that MPA is not improved by generic extra training alone. Mid-training improves endpoints whose governing physics overlaps with the auxiliary labels, while post-training improves endpoints whose physical form matches the readout and loss bias.

target.

Mid-training selects transferable physical supervision. The mid-training ablations improve nearly every endpoint under both evaluation regimes, and the gain grows as the auxiliary supervision broadens. With thermochemical supervision alone, the mean MAE reduction is 7.91% under the random split, improving 39 of 40 endpoints, and 6.46% under the scaffold split, improving 34 of 40 endpoints. Adding logP widens the transfer profile, reducing MAE by 9.09% under the random split and by 7.55% under the scaffold split, with 40 of 40 and 37 of 40 endpoints improved. The strongest mid-training setting is Thermo+All, which combines thermochemical supervision with the full set of

auxiliary property signals. It reduces MAE by 10.41% under the random split and by 8.39% under the scaffold split, improving 40 of 40 and 39 of 40 endpoints, respectively.

The benefit persists under scaffold-level distribution shift, where 34 to 39 of 40 endpoints remain improved. This is the central point. The injected physical information transfers to molecules drawn from new scaffolds rather than reflecting local interpolation among structurally similar training examples. A generic regularizer would not be expected to retain such broad coverage once the test molecules leave the training neighbourhood. The monotonic improvement from thermochemistry to logP to the full auxiliary set, present under both splits, instead points to transferable physical content that accumulates as the supervision spans more of the relevant physics.

The endpoint-family pattern supports this interpretation. Thermochemical supervision is most aligned with thermodynamic and phase-transition endpoints, where formation energies, entropies, heat capacities, vaporization behavior, and cohesive interactions are directly relevant. Adding logP introduces a solvation and partitioning axis, which broadens the transfer profile toward solubility, partitioning, and bio/other endpoints. Thermo+All provides the widest coverage because it combines multiple physical factors, including thermochemistry, polarity and electronic structure, and lipophilicity-related behavior. The best mid-training signal is therefore not simply the largest auxiliary dataset, but the one whose labels span the physical factors that govern the downstream endpoint.

The comparison also reveals a breadth–specialization trade-off. A single physically matched source can sharpen performance within its natural endpoint family, whereas Thermo+All gives the strongest average performance across the heterogeneous benchmark. Targeted applications may therefore benefit from a carefully chosen single auxiliary source, while broad deployment across many endpoint types benefits from multi-property mid-training.

Post-training aligns the readout with the target structure. The post-training ablations isolate the readout, comparing two physically structured prediction heads against the baseline, attention pooling and the hybrid head, both trained with the Huber objective. The final head determines how the learned molecular representation is expressed as a scalar endpoint, and its effect grows as the readout acquires more physically appropriate structure.

Attention pooling reduces MAE by 8.95% under the random split and by 8.97% under the scaffold split, improving 38 and 37 of 40 endpoints, respectively. It provides a flexible molecular summary and is useful for endpoints governed by nonlocal or emergent molecular features. The pooled representation remains largely non-additive, however, and therefore does not fully encode the physical structure of endpoints that are well described by local atomic, fragment, or group contributions.

The hybrid head gives the strongest post-training effect, and its advantage is largest under scaffold evaluation. It reduces MAE by 10.96% under the random split and by 13.15% under the scaffold split, improving 35 and 37 of 40 endpoints, respectively. Its strongest gains occur in thermodynamic and additive-compatible endpoint families, where atomwise or group-contribution structure is a natural approximation. The attention branch preserves flexibility for non-additive molecular summaries, while the additive branch supplies an explicit sum-like prior (Eq. 13) for endpoints whose values depend on local contributions or size-correlated effects.

To test whether this additive prior helps where it should, we classify the 40 endpoints by an *a priori* additive-compatibility criterion, using physical reasoning rather than observed performance so that the assignment cannot be tuned to the result it explains. A property is assigned to Class A, the additive-compatible class, when it satisfies at least one of three conditions. It is size-extensive or approximately extensive, so that its magnitude scales with molecular size or composition and admits a local atomic or group-contribution interpretation,[53, 54] as for formation, combustion, and vaporization enthalpies, heat capacities, and entropies. It is sum-structured, meaning its physical definition involves a sum over

local contributions even when the reported scalar is not itself extensive, as for the molecular dipole moment $\boldsymbol{\mu} = \sum_j q_j \mathbf{r}_j$, which aggregates atom-wise charge-position terms. Or it is formally intensive but size- or composition-correlated across molecular series, as for boiling point, critical temperature, and melting point, which are governed by fragment-level interactions and cohesive forces that an additive decomposition can capture. The remaining endpoints form Class B, the additive-mismatched class, whose values are primarily normalized, ratio-based, interfacial, recognition-driven, or otherwise governed by collective mechanisms not naturally represented by atom-wise summation, such as specific protein binding or fractional binding measurements. Applying this criterion before any model is trained yields 24 Class A and 16 Class B targets. Because the assignment is fixed in advance, it provides a falsifiable prediction. The additive branch should improve Class A targets more consistently than Class B targets.

The ablation confirms this prediction. For Class A properties, the hybrid head reduced MAE by 14.92% and 15.27% under the random and scaffold splits, respectively. For Class B properties, the corresponding reductions were smaller, at 5.01% and 9.96%. Attention pooling showed a less class-dependent pattern, reducing MAE by 10.06% and 9.55% for Class A properties under the random and scaffold splits, and by 7.29% and 8.08% for Class B properties, respectively. The hybrid head therefore offers a stronger physical inductive bias when the target admits an additive interpretation, while attention pooling behaves more uniformly across property classes. The additive branch is not universally optimal. Its value depends on whether its inductive bias is aligned with the physical nature of the endpoint, helping most when atom-wise summation is physically appropriate and less when the target is governed by molecular recognition, device-level effects, or other nonlocal mechanisms.

Together, the two MPA components are complementary and each is aligned with a distinct aspect of the target physics. Auxiliary supervision supplies transferable physical content, while the readout supplies a compatible prediction structure, so the full workflow improves most where both forms of alignment hold.

Table 4. Summary of MPA ablation settings under random and scaffold evaluation. Relative MAE changes are computed with respect to the corresponding ablation baseline; negative values indicate lower MAE. The improved columns report the number of endpoints with lower MAE than the baseline.

		Random		Scaffold	
Stage	Setting	Mean ΔMAE (%)	Improved	Mean ΔMAE (%)	Improved
Mid-training	Thermo	−7.91	39/40	−6.46	34/40
	Thermo+LogP	−9.09	40/40	−7.55	37/40
	Thermo+All	**−10.41**	**40/40**	**−8.39**	**39/40**
Post-training	Attention pooling + Huber	−8.95	**38/40**	−8.97	**37/40**
	Hybrid head + Huber	**−10.96**	35/40	**−13.15**	**37/40**

5 Conclusion

This work establishes multi-phase training as an effective strategy for pushing materials foundation models to diverse experimental property prediction. Validating MPA model across 40 experimental properties, we show that foundation models can be transformed from general materials representations into accurate predictors of sparse, noisy, and application-targeted experimental endpoints. The key insight is that transferability depends on physical alignment. Mid-training improves downstream performance when auxiliary supervision reflects the physics of the target property, while tailored readout architectures further improve prediction. This mirrors the training paradigm of large language

models: broad pre-training learns general representations, mid-training aligns the model with domain-relevant signals, and post-training adapts it to specific tasks. In materials science, the alignment signal comes from thermodynamic, electronic, interfacial, solvation, or biological structure rather than human preference data.

Consistent with this view, MPA benefits from scale and physically aligned supervision, reducing MAE by 10% on average and by up to 51% under out-of-distribution evaluation while achieving state-of-the-art performance across the 40 experimental benchmarks. These results suggest that materials foundation models will scale most effectively through physically meaningful multi-phase training, providing a practical guide of how computational and experimental data should be scaled in the future to best boost the accuracy and transferability on diverse downstream applications of chemistry and materials industry.

6 Acknowledgement

We thank the entire Deep Principle team for discussion and support. We thank MIRA for conducting initial research, adapting and updating backbone foundation models, automating the training and evaluation loop, analyzing the results, and writing the initial version of a report.

References

[1] Walters, W. P. & Barzilay, R. Applications of Deep Learning in Molecule Generation and Molecular Property Prediction. In: *Accounts of Chemical Research* 54.2 (2021), pp. 263–270. URL: https://doi.org/10.1021/acs.accounts.0c00699.

[2] Wieder, O., Kohlbacher, S., Kuenemann, M., Garon, A., Ducrot, P., Seidel, T. & Langer, T. A compact review of molecular property prediction with graph neural networks. In: *Drug Discovery Today: Technologies* (2020). URL: https://doi.org/10.1016/j.ddtec.2020.11.009.

[3] Li, Z., Jiang, M., Wang, S. & Zhang, S. Deep learning methods for molecular representation and property prediction. In: *Drug Discovery Today* 27.12 (2022), p. 103373. URL: https://doi.org/10.1016/j.drudis.2022.103373.

[4] Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. In: *Nature* 559.7715 (2018), pp. 547–555. URL: https://doi.org/10.1038/s41586-018-0337-2.

[5] Merchant, A., Batzner, S., Schoenholz, S. S., Aykol, M., Cheon, G. & Cubuk, E. D. Scaling deep learning for materials discovery. In: *Nature* 624.7990 (2023), pp. 80–85. URL: https://doi.org/10.1038/s41586-023-06735-9.

[6] Szymanski, N. J., Rendy, B., Fei, Y., Kumar, R. E., He, T., Milsted, D., McDermott, M. J., Gallant, M., Cubuk, E. D., Merchant, A., Kim, H., Jain, A., Bartel, C. J., Persson, K., Zeng, Y. & Ceder, G. An autonomous laboratory for the accelerated synthesis of inorganic materials. In: *Nature* 624.7990 (2023), pp. 86–91. URL: https://doi.org/10.1038/s41586-023-06734-w.

[7] Zeni, C., Pinsler, R., Zügner, D., Fowler, A., Horton, M., Fu, X., Wang, Z., Shysheya, A., Crabbé, J., Ueda, S., Sordillo, R., Sun, L., Smith, J., Nguyen, B., Schulz, H., Lewis, S., Huang, C.-W., Lu, Z., Zhou, Y., Yang, H., Hao, H., Li, J., Yang, C., Li, W., Tomioka, R. & Xie, T. A generative model for inorganic materials design. In: *Nature* 639.8055 (2025), pp. 624–632. URL: https://doi.org/10.1038/s41586-025-08628-5.

[8] Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G. & Persson, K. A. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. In: *APL Materials* 1.1 (2013), p. 011002. URL: https://doi.org/10.1063/1.4812323.

[9] Chanussot, L., Das, A., Goyal, S., Lavril, T., Shuaibi, M., Riviere, M., Tran, K., Heras-Domingo, J., Ho, C., Hu, W., Palizhati, A., Sriram, A., Wood, B., Yoon, J., Parikh, D., Zitnick, C. L. & Ulissi, Z. Open Catalyst 2020 (OC20) Dataset and Community Challenges. In: *ACS Catalysis* 11.10 (2021), pp. 6059–6072. URL: https://doi.org/10.1021/acscatal.0c04525.

[10] Dunn, A., Wang, Q., Ganose, A., Dopp, D. & Jain, A. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. In: *npj Computational Materials* 6.1 (2020), p. 138. URL: https://doi.org/10.1038/s41524-020-00406-3.

[11] Riebesell, J., Goodall, R. E. A., Benner, P., Chiang, Y., Deng, B., Ceder, G., Asta, M., Lee, A. A., Jain, A. & Persson, K. A. A framework to evaluate machine learning crystal stability predictions. In: *Nature Machine Intelligence* 7 (2025), pp. 836–847. URL: https://doi.org/10.1038/s42256-025-01055-1.

[12] Ross, J., Belgodere, B., Chenthamarakshan, V., Padhi, I., Mroueh, Y. & Das, P. Large-scale chemical language representations capture molecular structure and properties. In: *Nature Machine Intelligence* 4.12 (2022), pp. 1256–1264. URL: https://doi.org/10.1038/s42256-022-00580-7.

[13] Cai, F., Zacour, K., Zhu, T., Tzeng, T.-R., Duan, Y., Liu, L., Pilla, S., Li, G. & Luo, F. ChemFM as a scaling law guided foundation model pre-trained on informative chemicals. In: *Communications Chemistry* 9.1 (2025), p. 3. URL: https://doi.org/10.1038/s42004-025-01793-8.

[14] Rong, Y., Bian, Y., Xu, T., Xie, W., Wei, Y., Huang, W. & Huang, J. "Self-Supervised Graph Transformer on Large-Scale Molecular Data". In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 12559–12571. URL: https://proceedings.neurips.cc/paper/2020/hash/94aef38441efa3380a3bed3faf1f9d5d-Abstract.html.

[15] Zhou, G., Gao, Z., Ding, Q., Zheng, H., Xu, H., Wei, Z., Zhang, L. & Ke, G. "Uni-Mol: A Universal 3D Molecular Representation Learning Framework". In: *International Conference on Learning Representations*. 2023. URL: https://openreview.net/forum?id=IfFZr1gl0b.

[16] Ji, X., Wang, Z., Gao, Z., Zheng, H., Zhang, L., Ke, G. & E, W. Uni-Mol2: Exploring Molecular Pretraining Model at Scale. In: *arXiv preprint arXiv:2406.14969* (2024). URL: https://arxiv.org/abs/2406.14969.

[17] Ahmad, W., Simon, E., Chithrananda, S., Grand, G. & Ramsundar, B. ChemBERTa-2: Towards Chemical Foundation Models. In: *arXiv preprint arXiv:2209.01712* (2022). URL: https://arxiv.org/abs/2209.01712.

[18] Singh, R., Barsainyan, A. A., Irfan, R., Amorin, C. J., He, S., Davis, T., Thiagarajan, A., Sankaran, S., Chithrananda, S., Ahmad, W., Jones, D., McLoughlin, K., Kim, H., Bhutani, A., Sathyanarayana, S. V., Viswanathan, V., Allen, J. E. & Ramsundar, B. ChemBERTa-3: an open source training framework for chemical foundation models. In: *Digital Discovery* 5.2 (2026), pp. 662–685. URL: https://doi.org/10.1039/D5DD00348B.

[19] Burns, J. W., Zalte, A. S., Abreu, C. R. A., Sieg, J., Feldmann, C., Mathea, M. & Green, W. H. Deep Learning Foundation Models from Classical Molecular Descriptors. In: *arXiv preprint arXiv:2506.15792* (2025). URL: https://arxiv.org/abs/2506.15792.

[20] Méndez-Lucio, O., Nicolaou, C. A. & Earnshaw, B. MolE: a foundation model for molecular graphs using disentangled attention. In: *Nature Communications* 15.1 (2024), p. 9431. URL: https://doi.org/10.1038/s41467-024-53751-y.

[21] Sypetkowski, M., Wenkel, F., Poursafaei, F., Dickson, N., Suri, K., Fradkin, P. & Beaini, D. "On the Scalability of GNNs for Molecular Graphs". In: *Advances in Neural Information Processing Systems*. 2024. URL: https://proceedings.neurips.cc/paper_files/paper/2024/hash/2345275663a15ee92a06bc957be54a2c-Abstract-Conference.html.

[22] Wijaya, K. T., Guo, M., Sun, M., Seidel, H.-P., Matusik, W. & Babaei, V. Two-Stage Pretraining for Molecular Property Prediction in the Wild. In: *arXiv preprint arXiv:2411.03537* (2024). URL: https://arxiv.org/abs/2411.03537.

[23] Deng, J., Yang, Z., Wang, H., Ojima, I., Samaras, D. & Wang, F. A systematic study of key elements underlying molecular property prediction. In: *Nature Communications* 14 (2023), p. 6395. URL: https://doi.org/10.1038/s41467-023-41948-6.

[24] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. & Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. In: *Journal of Machine Learning Research* 21.140 (2020), pp. 1–67. URL: https://jmlr.org/papers/v21/20-074.html.

[25] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J. & Lowe, R. "Training language models to follow instructions with human feedback". In: *Advances in Neural Information Processing Systems*. Vol. 35. 2022, pp. 27730–27744. URL: https://arxiv.org/abs/2203.02155.

[26] Huber, P. J. Robust Estimation of a Location Parameter. In: *The Annals of Mathematical Statistics* 35.1 (1964), pp. 73–101. URL: https://doi.org/10.1214/aoms/1177703732.

[27] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. "Attention is all you need". In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017. URL: https://arxiv.org/abs/1706.03762.

[28] Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J. & Bolton, E. E. PubChem 2023 update. In: *Nucleic Acids Research* 51.D1 (2023), pp. D1373–D1380. URL: https://doi.org/10.1093/nar/gkac956.

[29] *RDKit: Open-source cheminformatics*. https://www.rdkit.org. 2016. URL: https://www.rdkit.org.

[30] Riniker, S. & Landrum, G. A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. In: *Journal of chemical information and modeling* 55.12 (2015), pp. 2562–2574. URL: https://doi.org/10.1021/acs.jcim.5b00654.

[31] Pracht, P., Grimme, S., Bannwarth, C., Bohle, F., Ehlert, S., Feldmann, G., Gorges, J., Müller, M., Neudecker, T., Plett, C., Spicher, S., Steinbach, P., Wesołowski, P. A. & Zeller, F. CREST—A program for the exploration of low-energy molecular chemical space. In: *The Journal of Chemical Physics* 160.11 (2024). URL: https://doi.org/10.1063/5.0197592.

[32] Spicher, S. & Grimme, S. Robust Atomistic Modeling of Materials, Organometallic, and Biochemical Systems. In: *Angewandte Chemie International Edition* 59.36 (2020), pp. 15665–15673. URL: https://doi.org/10.1002/anie.202004239.

[33] Bannwarth, C., Ehlert, S. & Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. In: *Journal of chemical theory and computation* 15.3 (2019), pp. 1652–1671. URL: https://doi.org/10.1021/acs.jctc.8b01176.

[34] Mansouri, K., Grulke, C. M., Judson, R. S. & Williams, A. J. OPERA models for predicting physicochemical properties and environmental fate endpoints. In: *Journal of cheminformatics* 10.1 (2018), p. 10. URL: https://doi.org/10.1186/s13321-018-0263-1.

[35] Zhao, Q. & Savoie, B. M. Self-Consistent Component Increment Theory for Predicting Enthalpy of Formation. In: *Journal of Chemical Information and Modeling* 60.4 (2020), pp. 2199–2207. URL: https://doi.org/10.1021/acs.jcim.0c00092.

[36] Zhao, Q., Iovanac, N. C. & Savoie, B. M. Transferable Ring Corrections for Predicting Enthalpy of Formation of Cyclic Compounds. In: *Journal of Chemical Information and Modeling* 61.6 (2021), pp. 2798–2805. URL: https://doi.org/10.1021/acs.jcim.1c00367.

[37] Froitzheim, T., Müller, M., Hansen, A. & Grimme, S. g-xTB: A General-Purpose Extended Tight-Binding Electronic Structure Method for the Elements H to Lr (Z=1–103). In: (2025). URL: https://doi.org/10.26434/chemrxiv-2025-bjxvt.

[38] Yaws Carl, L. *Yaws' Handbook of Thermodynamic and Physical Properties of Chemical Compounds*. Knovel, 2003. URL: https://search.worldcat.org/title/51621359.

[39] Haynes, W. M. *CRC handbook of chemistry and physics*. CRC press, 2016. URL: https://books.google.com/books?id=VVezDAAAQBAJ.

[40] Huang, K., Fu, T., Gao, W., Zhao, Y., Roohani, Y., Leskovec, J., Coley, C. W., Xiao, C., Sun, J. & Zitnik, M. Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Development. In: *arXiv preprint arXiv:2102.09548* (2021). URL: https://arxiv.org/abs/2102.09548.

[41] Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K. & Pande, V. MoleculeNet: a benchmark for molecular machine learning. In: *Chemical science* 9.2 (2018), pp. 513–530. URL: https://doi.org/10.1039/C7SC02664A.

[42] Biswas, S., Chung, Y., Ramirez, J., Wu, H. & Green, W. H. Predicting Critical Properties and Acentric Factors of Fluids Using Multitask Machine Learning. In: *Journal of Chemical Information and Modeling* 63.15 (2023), pp. 4574–4588. URL: https://doi.org/10.1021/acs.jcim.3c00546.

[43] Goussard, V., Duprat, F., Ploix, J.-L., Dreyfus, G., Nardello-Rataj, V. & Aubry, J.-M. A New Machine-Learning Tool for Fast Estimation of Liquid Viscosity. Application to Cosmetic Oils. In: *Journal of Chemical Information and Modeling* 60.4 (2020), pp. 2012–2023. URL: https://doi.org/10.1021/acs.jcim.0c00083.

[44] Chew, A. K., Sender, M., Kaplan, Z., Chandrasekaran, A., Chief Elk, J., Browning, A. R., Kwak, H. S., Halls, M. D. & Afzal, M. A. F. Advancing material property prediction: using physics-informed machine learning models for viscosity. In: *Journal of Cheminformatics* 16.1 (2024), p. 31. URL: https://doi.org/10.1186/s13321-024-00820-5.

[45] Krasnov, L., Malikov, D., Kiseleva, M., Tatarin, S., Sosnin, S. & Bezzubov, S. BigSolDB 2.0, dataset of solubility values for organic compounds in different solvents at various temperatures. In: *Scientific Data* 12.1 (2025), p. 1236. URL: https://doi.org/10.1038/s41597-025-05559-8.

[46] Sorkun, M. C., Khetan, A. & Er, S. AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. In: *Scientific data* 6.1 (2019), p. 143. URL: https://doi.org/10.1038/s41597-019-0151-1.

[47] Gao, P., Andersen, A., Sepulveda, J., Panapitiya, G. U., Hollas, A., Saldanha, E. G., Murugesan, V. & Wang, W. SOMAS: a platform for data-driven material discovery in redox flow battery development. In: *Scientific Data* 9.1 (2022), p. 740. URL: https://doi.org/10.1038/s41597-022-01814-4.

[48] McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. In: *Journal of Open Source Software* 3.29 (2018), p. 861. URL: https://doi.org/10.21105/joss.00861.

[49] Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. In: *Journal of Chemical Information and Modeling* 50.5 (2010), pp. 742–754. URL: https://doi.org/10.1021/ci100050t.

[50] Bemis, G. W. & Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. In: *Journal of medicinal chemistry* 39.15 (1996), pp. 2887–2893. URL: https://doi.org/10.1021/jm9602928.

[51] Graff, D. E., Morgan, N. K., Burns, J. W., Doner, A. C., Li, B., Li, S.-C., Manu, J., Menon, A., Pang, H.-W., Wu, H., Zalte, A. S., Zheng, J. W., Coley, C. W., Green, W. H. & Greenman, K. P. Chemprop v2: An Efficient, Modular Machine Learning Package for Chemical Property Prediction. In: *Journal of Chemical Information and Modeling* 66.1 (2026), pp. 28–33. URL: https://doi.org/10.1021/acs.jcim.5c02332.

[52] An, J., Lu, X., Shi, Y.-F., Xu, L.-C., Zhang, N., Qu, C., Qi, Y. & Cao, F. Suiren-1.0 Technical Report: A Family of Molecular Foundation Models. In: *arXiv preprint arXiv:2603.21942* (2026). URL: https://arxiv.org/abs/2603.21942.

[53] Benson, S. W., Cruickshank, F. R., Golden, D. M., Haugen, G. R., O'Neal, H. E., Rodgers, A. S., Shaw, R. & Walsh, R. Additivity rules for the estimation of thermochemical properties. In: *Chemical Reviews* 69.3 (1969), pp. 279–324. URL: https://doi.org/10.1021/cr60259a002.

[54] Joback, K. G. & Reid, R. C. ESTIMATION OF PURE-COMPONENT PROPERTIES FROM GROUP-CONTRIBUTIONS. In: *Chemical Engineering Communications* 57.1-6 (1987), pp. 233–243. URL: https://doi.org/10.1080/00986448708960487.