LEARNING AT DIFFERENT SPEEDS: EQUILIBRIUM ROUTING FOR AI-DRIVEN SCIENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

AI for Science increasingly trains a single model to predict multiple observables of the same physical system. Coupled targets often share latent physical structure, allowing joint prediction to exploit common mechanisms and mutual constraints. However, shared physics does not imply synchronized optimization. In data-limited scientific regimes, one observable may saturate early and become prone to overfitting, while another still requires longer training to learn slower or higher-dimensional physical modes. We study this failure mode as asynchronous convergence. We introduce Task-Aware Equilibrium Routing Mixture-of-Experts (TAER-MoE), a convergence-aware MoE framework that routes by generalization. TAER-MoE does not treat large loss or high expert affinity as sufficient evidence for assigning more capacity. Instead, it estimates task-level learning potential from validation descent and route-level overfitting from expert-task generalization gaps. A spectral bias–variance perspective links asynchronous convergence to task-dependent physical spectra and different optimal stopping times. Guided by this perspective, TAER-MoE uses an entropy-regularized equilibrium router to suppress expert-task routes with large train–validation gaps and reallocate training pressure toward routes where additional optimization remains useful. Across tropical cyclone forecasting, QM9 molecular property prediction, and a transverse-field Ising model benchmark, TAER-MoE improves joint test prediction over single-task, shared, MoE-based, gradient-based, and ensemble baselines. These results provide empirical evidence that validation-aware routing can improve scientific multi-task learning when coupled observables converge asynchronously.

1 INTRODUCTION

Scientific prediction often requires modeling multiple coupled observables of the same physical system. A tropical cyclone is simultaneously a trajectory, a wind field, and a pressure system (Roy & Kovordányi, 2012; Chavas et al., 2017; Huang et al., 2025). A molecule is at once a geometry, a spectrum, and an electronic structure (Ramakrishnan et al., 2014; Wu et al., 2018; Gilmer et al., 2017). These targets are not independent labels, but different projections of the same underlying physics (Carleo et al., 2019; Karniadakis et al., 2021). Joint prediction is therefore a natural paradigm for AI for Science, because shared models can exploit common latent mechanisms and cross-target constraints that separate models may miss (Zhang & Yang, 2018; 2021).

Yet physical coupling does not imply synchronized optimization. Targets from the same system may learn at different speeds (Chen et al., 2018; Guo et al., 2018; Kendall et al., 2018). Some depend on local and readily identifiable patterns, whereas others require longer temporal memory, broader context, or slower evidence accumulation (Zamir et al., 2018; Liu et al., 2019). One target may saturate early and become vulnerable to overfitting, while another still benefits from continued optimization. Joint learning then becomes fragile, because updates that remain useful for the slower target can degrade the representation already sufficient for the faster one (Sener & Koltun, 2018; Yu et al., 2020; Liu et al., 2021; Javaloy & Valera, 2021; Navon et al., 2022).

This asynchronous convergence is especially problematic for Mixture-of-Experts routing. Raw task losses and instantaneous gradient magnitudes are often poor proxies for useful training need. A large loss may reflect noise, scale, or irreducible error, while a small loss may indicate that a target has already saturated and is vulnerable to overfitting. Similarly, high expert affinity only indicates that an

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

expert currently fits the routed samples; it does not imply that sending more samples through this route will improve generalization. For scientific MoE training, this issue is even more fine-grained. The relevant question is not only which task should receive more pressure, but which expert-task route still contains reducible and generalizable signal. In data-limited scientific regimes, capacity assigned beyond this reducible signal can amplify finite-sample artifacts and degrade already-saturated targets.

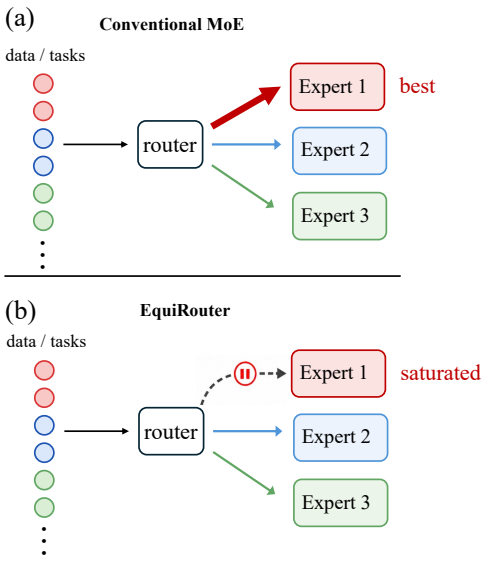

Figure 1: From winner-take-more to saturation-avoidance routing. Conventional MoE reinforces high-affinity experts, whereas TAER-MoE suppresses overfitting expert-task routes and redirects capacity to routes with remaining generalizable signal.

We introduce **T**ask-**A**ware **E**quilibrium **R**outing **M**ixture-**o**f-**E**xperts (TAER-MoE), a convergence-aware MoE framework for scientific multi-task learning. TAER-MoE allocates expert capacity according to remaining generalizable signal rather than raw loss, task difficulty, or current expert affinity. It combines task-level learning potential with route-conditioned generalization gaps. The former estimates whether a target is still improving, while the latter identifies expert-task paths that fit training samples without improving validation generalization. These signals define an entropy-regularized equilibrium router that suppresses overfitting expert-task routes while preserving input-dependent expert affinity. As illustrated in Fig. 1, TAER-MoE changes routing from reinforcing what already fits to reallocating capacity where additional optimization remains useful.

We evaluate TAER-MoE in three scientific settings with different scales and physics, including tropical cyclone forecasting, QM9 molecular property prediction, and a transverse-field Ising model benchmark. Across macroscopic geophysical dynamics, microscopic quantum chemistry, and static quantum many-body prediction, TAER-MoE improves joint test prediction over single-task, shared, MoE-based, gradient-based, and ensemble baselines. Ablations and routing analyses show that the gains arise from training-time generalization-aware capacity allocation.

In summary, this work makes three contributions.

- We characterize asynchronous convergence as a learning-stage mismatch in scientific multi-task learning, where physically coupled targets may reach their best generalization stages at different times.
- We introduce TAER-MoE, a validation-aware MoE training controller that combines task-level validation descent with expert-task generalization gaps.
- We evaluate TAER-MoE on tropical cyclone forecasting, QM9 molecular property prediction, and transverse-field Ising prediction, and provide ablation and routing analyses that are consistent with the proposed stage-aware capacity-allocation mechanism.

2 Related Work

Multi-task learning methods such as Gradient Normalization (GradNorm) (Chen et al., 2018), Projected Conflicting Gradient (PCGrad) (Yu et al., 2020), and Conflict-Averse Gradient Descent (CAGrad) (Liu et al., 2021) reduce inter-task interference by adjusting loss weights or gradients. Progress-aware methods such as Achievement-based Training Progress Balancing (Yun & Cho, 2023), Improvable Gap Balancing (Dai et al., 2023) and ExcessMTL (He et al., 2024) further show that raw losses or instantaneous gradients are unreliable proxies for useful training need. These methods mainly regulate task-level pressure. TAER-MoE targets a finer-grained setting, where capacity must be assigned not only across tasks but also across expert-task routes that differ in remaining generalizable signal.

Mixture-of-Experts architectures allocate capacity through learned routers (Yuksel et al., 2012; Eigen et al., 2013; Fedus et al., 2022; Dai et al., 2022; Rajbhandari et al., 2022; Lin et al., 2026). Recent work improves routing stability (Dai et al., 2022), sparsity (Puigcerver et al., 2023), specialization (Chen et al., 2023; Shi et al., 2024; Liu et al., 2024), and adaptive expert usage, as in AdaMV-MoE. However, high expert affinity or larger expert capacity does not guarantee better generalization. TAER-MoE augments routing with route-conditioned train–validation gaps, so expert allocation reflects both input affinity and expert-task generalization state.

Scientific learning often combines shared physical structure with limited or imbalanced data (Carleo et al., 2019; Karniadakis et al., 2021; Bodnar et al., 2025). Over-parameterized models can exploit common mechanisms, but they can also overfit targets whose useful modes mature early. Our spectral view links this behavior to task-dependent physical spectra and different optimal training stages, motivating routing toward targets and expert paths that retain reducible and generalizable signal.

3 METHOD

In this section, we introduce Task-Aware Equilibrium Routing Mixture-of-Experts (TAER-MoE), a convergence-aware routing method for scientific multi-task learning. The method addresses the following question.

> Can a multi-task scientific model allocate expert capacity according to validation generalization when coupled targets reach their best learning stages at different times?

TAER-MoE is based on two ideas. First, task-level validation descent estimates which targets still contain learnable signal. Second, expert-task generalization gaps estimate which routes have begun to fit training samples without improving validation generalization. The first signal controls task pressure. The second signal corrects expert routing. Detailed derivations and implementation details are provided in Supplementary Section B.

3.1 STAGE SEPARATION IN LOCAL FEATURE SPACE

We first formalize the learning-stage mismatch that motivates TAER-MoE. The analysis below should be interpreted as a local spectral approximation to nonlinear MoE training around a validation checkpoint, rather than as a global convergence guarantee. At checkpoint c, the encoder, experts, and routers induce checkpoint-dependent and route-dependent feature spaces. For small subsequent updates, optimization can be viewed through the spectrum of the corresponding local feature or tangent space. Scientific targets from the same system may share the same nonlinear backbone but depend on different spectral components of these local route features.

For task i and expert route e, let

$$z_{i,e,c}(x) = E_e(f_{\theta_c}(x))$$

denote the local feature induced by expert e at checkpoint c. Let the covariance operator of $z_{i,e,c}(x)$ have eigenpairs (λ_m, v_m). For task i, consider the local target expansion

$$y_i = \langle h_i^\star, z_{i,e,c}(x)\rangle + \varepsilon_i, \qquad h_i^\star = \sum_{m\geq 1} c_m^{(i,e,c)} v_m, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2).$$

Here $c_m^{(i,e,c)}$ measures the dependence of task i on the m-th local spectral mode of route e at checkpoint c. Let r_i be the slowest relevant local mode for task i.

Theorem 1. Assume the local route-conditioned spectrum satisfies $\lambda_m \asymp m^{-\beta}$ with $\beta > 0$. Under gradient-flow learning over this local spectral basis, the optimal training stage of task i satisfies

$$T_i^\star = \Theta\left(\frac{\log(n c_{r_i}^{(i,e,c)}/\sigma_i)}{\lambda_{r_i}}\right), \qquad T_2^\star/T_1^\star = \Theta\big((r_2/r_1)^\beta\big). \tag{1}$$

For a fixed joint training horizon T, a stylized excess-risk comparison gives

$$\max_i \mathcal{R}_i(\theta_T^{\text{joint}}) \geq \min_i \mathcal{R}_i(\theta_{T_i^\star}^{\text{single}}) \left[1 + \Omega\left(\frac{(T_2^\star - T_1^\star)^2}{(T_2^\star + T_1^\star)^2}\right)\right]. \tag{2}$$

This local result is an explanatory model rather than a global guarantee for nonlinear MoE training. It motivates the observable signals used by TAER-MoE. Since the task-specific optimum $T_i^\star$ is unavailable during training, we use validation descent to estimate whether additional optimization remains useful. Since the route-conditioned spectra and coefficients $c_m^{(i,e,c)}$ are also unobserved, we use expert-task train–validation gaps to detect routes that fit finite-sample artifacts rather than transferable signal. Thus, task-level descent controls the pressure $\alpha_i(c)$, while route-level gaps correct the expert logits in Sec. 3.2–3.3.

This local view suggests that tasks relying on slower local modes may require later optimal stages, and that a fixed joint schedule can become suboptimal when these stages separate. TAER-MoE therefore uses validation feedback to allocate optimization toward routes where continued training remains useful. Importantly, longer training is not harmful by itself; the issue is target-dependent. A slow target may still gain reusable structure, while a saturated target may convert additional updates into finite-sample fit.

3.2 Validation Signals

TAER-MoE uses validation-stage statistics as observable proxies for the unobserved optimal stages in Theorem 1. For task-level progress, we measure validation descent over a window of length W:

$$S_i(c) = 1 - \frac{\mathcal{L}_{\mathrm{val}}^{(i)}(c)}{W^{-1}\sum_{j=1}^{W}\mathcal{L}_{\mathrm{val}}^{(i)}(c-j)+\varepsilon}, \qquad u_i(c) = \mathrm{EMA}([S_i(c)]_+). \tag{3}$$

The resulting task pressure is

$$\alpha_i(c) = \mathrm{clip}\left[N_{\mathrm{task}}\frac{\exp(u_i(c)/\tau_u)}{\sum_{m=1}^{N_{\mathrm{task}}}\exp(u_m(c)/\tau_u)}, \alpha_{\min}, \alpha_{\max}\right]. \tag{4}$$

For route-level generalization, let $\hat{y}_k^{(i)}(x) = H_i(E_k(f_\theta(x)))$ and $\bar{w}_k^{(i)} = \mathrm{stopgrad}(w_k^{(i)})$. For $s \in \{\mathrm{train}, \mathrm{val}\}$, define

$$\mathcal{E}_{s,k}^{(i)}(c) = \frac{\sum_{\mathcal{D}_s}\bar{w}_k^{(i)}(x;c)\ell_i(\hat{y}_k^{(i)}(x), y^{(i)})}{\sum_{\mathcal{D}_s}\bar{w}_k^{(i)}(x;c)+\varepsilon}, \qquad G_k^{(i)}(c) = \mathcal{E}_{\mathrm{val},k}^{(i)}(c) - \mathcal{E}_{\mathrm{train},k}^{(i)}(c). \tag{5}$$

We normalize gaps within each task:

$$\widetilde{G}_k^{(i)}(c) = \left[\frac{G_k^{(i)}(c) - \mathrm{median}_j\, G_j^{(i)}(c)}{\mathrm{MAD}_j\, G_j^{(i)}(c)+\varepsilon}\right]_+. \tag{6}$$

The descent signal controls how much each task is trained, and the gap signal controls which expert routes receive that task's samples.

3.3 Practical Update

The backbone is a standard multi-task MoE. A shared encoder produces $\mathbf{h} = f_\theta(x)$. For task i, router R_i produces logits $\mathbf{g}^{(i)}(x) = R_i(\mathbf{h})$ over K experts. TAER-MoE modifies the router by adding a validation-stage correction to the expert logits:

$$w_k^{(i)}(x;c) = \mathrm{softmax}_k\left(\frac{g_k^{(i)}(x) - \gamma\,\mathrm{stopgrad}(\widetilde{G}_k^{(i)}(c))}{\tau_r}\right). \tag{7}$$

This update is the closed-form solution of an entropy-regularized free-energy objective. The affinity term selects input-relevant experts. The gap term lowers the probability of routes with weak validation generalization.

The routed prediction is

$$\hat{y}^{(i)}(x) = H_i\left(\sum_{k=1}^{K} w_k^{(i)}(x;c)E_k(\mathbf{h})\right). \tag{8}$$

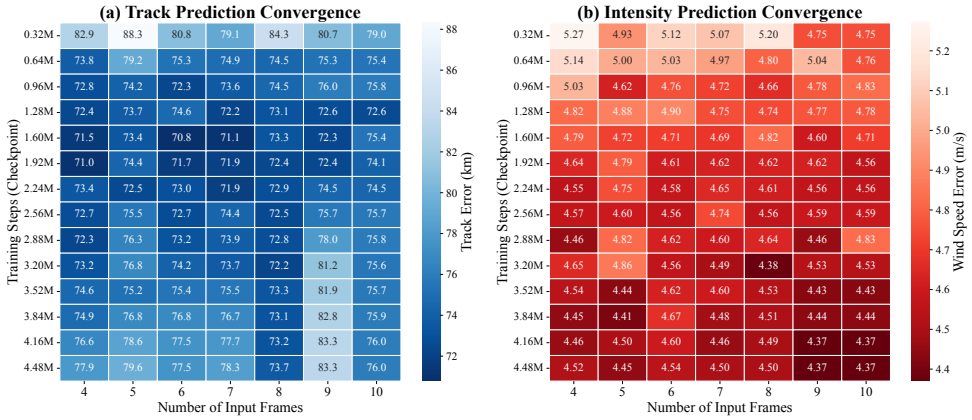

Figure 2: Asynchronous convergence in tropical cyclone forecasting. Track prediction reaches its best validation stage earlier and favors shorter input windows, whereas intensity prediction benefits from longer optimization and longer temporal context. Darker colors indicate lower error.

The training objective is

$$\mathcal{L}_{\text{TAER-MoE}} = \sum_{i=1}^{N_{\text{task}}} \alpha_i(c)\mathbb{E}_{\mathcal{D}_{\text{train}}}\left[\ell_i(\hat{y}^{(i)}(x), y^{(i)})\right] + \beta_{\text{load}}\mathcal{L}_{\text{load}} + \beta_{\text{reg}}\mathcal{L}_{\text{reg}}. \tag{9}$$

At each validation checkpoint, TAER-MoE updates $S_i(c)$, $u_i(c)$, $\alpha_i(c)$, and $\widetilde{G}_k^{(i)}(c)$. Between checkpoints, these quantities remain fixed. At inference time, the final gap correction is a fixed expert-task bias and requires no validation pass.

4 Case Study I: Tropical Cyclone Forecasting

Tropical cyclone (TC) forecasting provides a macroscopic testbed for asynchronous convergence. A storm jointly evolves through trajectory, wind, and pressure, making multi-target prediction physically meaningful. However, these coupled targets need not optimize synchronously: track is governed mainly by large-scale steering flows, whereas intensity depends more on thermodynamic accumulation, ventilation, and inner-core organization. They can therefore reach their best generalization stages at different times.

We evaluate global TC forecasting across six ocean basins using storm-centered environmental fields and historical storm states sampled every 6 hours. Training and validation use storms from 1950–2016 with a TC-level split, and 2017–2023 is held out for testing. All methods use the same inputs, preprocessing, prediction horizon, checkpoint rule, and training protocol, and all results are averaged over five random seeds. Trajectory is evaluated by average displacement error (ADE), while intensity is evaluated by mean absolute error for maximum wind speed and minimum central pressure.

Table 1: Global TC forecasting on held-out 2017–2023. Shared storm-centered protocol; 1950–2016 train/validation storms are TC-level split; results average five seeds.

Method	ADE (km)↓	Press (hPa)↓	Wind (m/s)↓
GraphCast	133.72 +135.1%	10.94 +107.6%	8.65 +135.6%
FuXi	138.95 +144.3%	10.56 +100.4%	9.45 +157.4%
Pangu-Weather	125.37 +120.4%	11.64 +120.9%	10.96 +198.5%
TCN	92.18 +62.1%	5.42 +2.8%	3.76 +2.4%
Env-MLP	94.41 +66.0%	5.60 +6.3%	3.92 +6.8%
MMoE	95.86 +68.5%	5.345 +1.4%	3.794 +3.3%
PLE	94.37 +65.9%	5.401 +2.5%	3.842 +4.6%
PCGrad	110.38 +94.1%	5.308 +0.7%	3.803 +3.6%
TAER-MoE	**56.88±1.03**	**5.27±0.03**	**3.67±0.03**

Table 1 compares TAER-MoE with foundation weather models, TC-oriented predictors, and representative multi-task baselines. Under the same storm-centered protocol and held-out 2017–2023 test period, TAER-MoE achieves the best overall joint performance. The gain is especially large for track

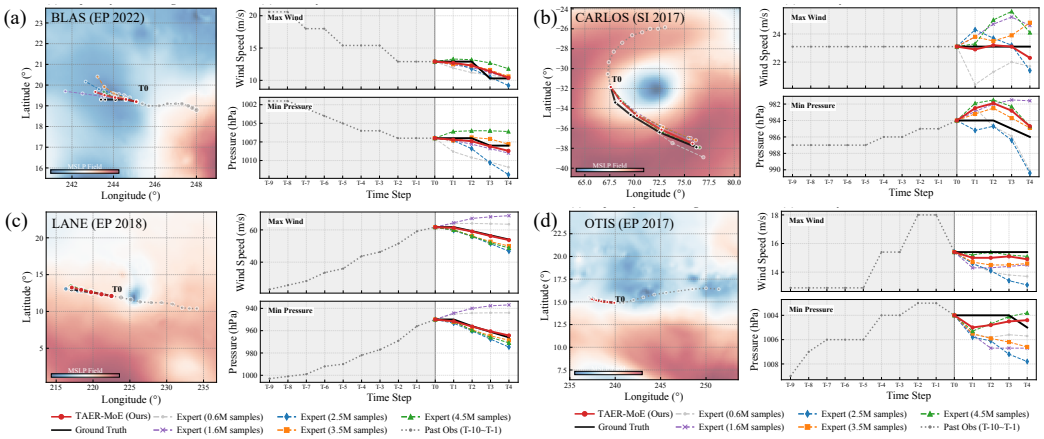

Figure 3: Representative TC forecasting cases. TAER-MoE better balances trajectory and intensity prediction than single-scale experts across storms from different basins.

prediction, while the improvements on pressure and wind are smaller but consistent, indicating a better joint track–intensity trade-off.

Figure 2 shows why this trade-off arises. Using identical architectures, we vary the training scale from 320k to 4.48M samples and the input length from 4 to 10 frames. Track reaches its best validation stage relatively early and then degrades with further optimization, whereas intensity continues to benefit from longer context and longer training. A single shared training schedule can therefore over-train the faster-saturating target or under-train the slower one.

TAER-MoE addresses this mismatch through task-level validation descent and route-level generalization gaps. When a target stops improving on validation data, its training pressure is reduced; when an expert-task route fits the training set but generalizes poorly, its routing probability is suppressed. In TC forecasting, this prevents repeated reinforcement of saturated track routes while preserving capacity for slower intensity-related modes.

Figure 3 shows representative storms from different basins. TAER-MoE preserves trajectory accuracy while maintaining coherent wind-speed and pressure evolution. Together, the quantitative results and case studies support the need for stage-aware capacity allocation in coupled TC prediction.

5 Case Study II: Transverse-Field Ising Model

We further evaluate TAER-MoE on a controlled quantum many-body benchmark based on the one-dimensional transverse-field Ising model (TFIM). In contrast to TC forecasting, TFIM removes temporal storm dynamics, basin dependence, and observational uncertainty by using labels generated from exactly solvable Hamiltonians. This provides a clean setting where two observables are derived from the same ground state but exhibit different learning stages.

We consider a spin-$1/2$ chain of length $L = 200$ with site-dependent Ising couplings and transverse fields. For each sample, the Hamiltonian and input are

$$H(\mathbf{J}, \mathbf{h}) = -\sum_{i=1}^{L-1} J_i \sigma_i^z \sigma_{i+1}^z - \sum_{i=1}^{L} h_i \sigma_i^x, \qquad x = \{(J_i, h_i)\}_{i=1}^{L} \in \mathbb{R}^{L\times 2}. \tag{10}$$

Here $\sigma_i^{x,z}$ are Pauli operators, J_i is the nearest-neighbor Ising coupling, and h_i is the local transverse field. We generate 20,000 Hamiltonian samples and compute ground-state observables exactly using the free-fermion/Majorana representation, which reduces the original 2^L Hilbert space to a quadratic $2L \times 2L$ representation. All baselines use the same generated data, split, and evaluation protocol.

For each Hamiltonian, the supervised targets are the ground-state energy density and half-chain entanglement entropy. Let $|\psi_0\rangle$ be the ground state with energy E_0. For the bipartition $A =$

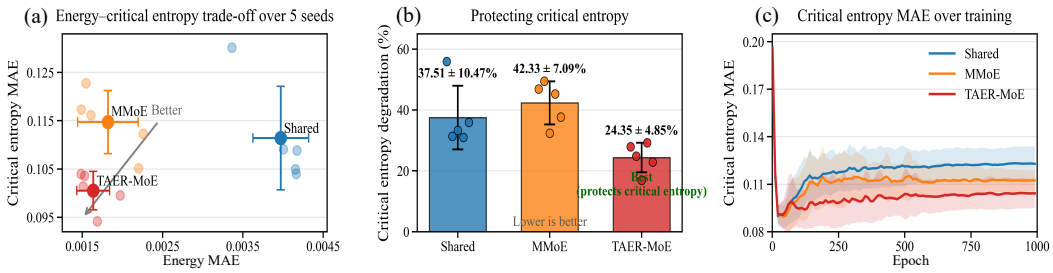

Figure 4: **TFIM benchmark. (a)** Test energy–critical entropy trade-off over five seeds, with large markers showing mean ± std. **(b)** Critical entropy degradation at the energy-optimal checkpoint. **(c)** Critical-region entropy MAE during training. Lower is better in all panels.

Table 2: TFIM test performance over five seeds. All methods use the same generated Hamiltonians, data split, checkpoint rule, and evaluation protocol.

Method	Energy MAE ↓	Critical entropy MAE ↓	Critical entropy degradation ↓
Shared	0.003971 ± 0.000347 +142.9%	0.111411 ± 0.010720 +10.8%	$37.51 \pm 10.47\%$ +54.0%
MMoE	0.001819 ± 0.000378 +11.3%	0.114729 ± 0.006529 +14.1%	$42.33 \pm 7.09\%$ +73.8%
TAER-MoE	$\mathbf{0.001635 \pm 0.000203}$	$\mathbf{0.100529 \pm 0.003983}$	$\mathbf{24.35 \pm 4.85\%}$

$\{1, \ldots, L/2\}$ and $B = \{L/2+1, \ldots, L\}$,

$$e_0 = \frac{E_0}{L} = \frac{1}{L}\langle\psi_0|H|\psi_0\rangle, \qquad S_{L/2} = -\mathrm{Tr}(\rho_A \log \rho_A)\,, \quad \rho_A = \mathrm{Tr}_B[|\psi_0\rangle\langle\psi_0|]\,. \tag{11}$$

Thus, the learning problem is

$$\{(J_i, h_i)\}_{i=1}^{L} \longmapsto \big(e_0, S_{L/2}\big)\,. \tag{12}$$

This static benchmark shows that asynchronous convergence is not limited to temporal forecasting.

To characterize the critical region, we use

$$r_{\text{eff}} = \exp\left(\frac{1}{L}\sum_{i=1}^{L}\log h_i - \frac{1}{L}\sum_{i=1}^{L}\log J_i\right), \qquad r_{\text{eff}} \approx 1 \text{ marks the quantum critical region.} \tag{13}$$

We report both overall entropy error and critical-region entropy MAE, since half-chain entanglement is most sensitive near criticality.

TFIM exposes a clear stage mismatch. The entropy target saturates early, especially near the critical region, while energy density continues to benefit from longer optimization. Energy-based checkpoint selection can therefore degrade the already-saturated entropy observable. This reproduces the same failure mode observed in TC within a static quantum system.

Figure 4 summarizes the TFIM results. Panel (a) shows the test energy–critical entropy trade-off over five seeds. TAER-MoE achieves the best lower-left trade-off by improving energy prediction while preserving critical entropy accuracy. Panel (b) measures critical entropy degradation under energy-based checkpoint selection. Panel (c) shows that TAER-MoE maintains lower critical-region entropy error during late training.

Table 2 reports test performance over five random seeds. Shared training underfits energy, while MMoE improves energy but leaves high critical entropy degradation. TAER-MoE achieves the best energy–critical entropy trade-off and reduces critical entropy degradation by approximately 42.5% relative to MMoE. These results show that generalization-aware routing protects a fast-saturated quantum observable while continuing to optimize a slower one.

The TFIM case complements TC by removing temporal forecasting, basin dependence, and observational uncertainty. TC tests macroscopic geophysical dynamics, while TFIM provides a controlled quantum many-body setting with exact labels. Across both regimes, physically coupled targets benefit from shared learning but require stage-aware capacity allocation.

We next examine QM9 molecular property prediction, which provides a microscopic quantum-chemical setting complementary to both TC and TFIM.

6 Case Study III: QM9 Molecular Property Prediction

We further evaluate TAER-MoE on QM9 molecular property prediction, a microscopic quantum-chemical setting where multiple molecular observables arise from the same structure but depend on different physical components. We first screen QM9 learning curves to identify heterogeneous convergence behavior. Geometry- and inertia-related targets, such as rotational constants, tend to stabilize earlier, whereas electronic and energetic targets, such as HOMO and internal energy, show longer convergence tails. Based on this screening, we predefine three representative two-task pairs before final evaluation: HOMO + internal energy U, HOMO + rotational constant C, and HOMO + zero-point vibrational energy (ZPVE). The full screening curves and pair-selection rationale are provided in the supplementary material.

Table 3: **QM9 two-task prediction.** Test MAE for each target pair; red percentages denote baseline error relative to TAER-MoE.

Method	HOMO + U	HOMO + C	HOMO + ZPVE
Shared	0.04645 / 29.31 +10.4 / 44.9%	0.05489 / 0.03041 +9.3 / 26.0%	0.04742 / 0.03290 +11.9 / 78.5%
MMoE	0.04382 / 27.84 +4.1 / 37.6%	0.05420 / 0.02725 +7.9 / 12.9%	0.04259 / 0.02099 +0.5 / 13.9%
TAER-MoE	**0.04208 / 20.23**	**0.05021 / 0.02413**	**0.04237 / 0.01843**

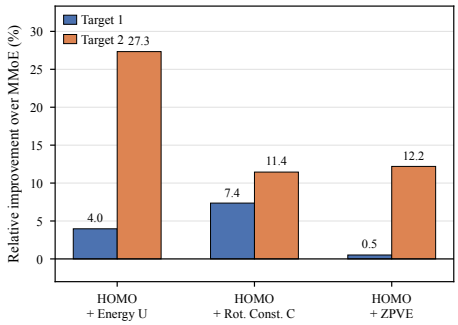

Figure 5: QM9 relative improvement over MMoE. TAER-MoE improves joint test prediction across electronic–thermodynamic, electronic–geometric, and electronic–vibrational task pairs.

Table 3 compares TAER-MoE with a shared multi-task model and a conventional MMoE baseline under the same training protocol. TAER-MoE improves joint test performance across all selected molecular pairs, with the largest gain on HOMO + Energy U, where task scale and convergence profile differ most strongly.

Figure 5 summarizes the relative improvement over MMoE. Across electronic–thermodynamic, electronic–geometric, and electronic–vibrational couplings, TAER-MoE consistently improves the paired targets. These results show that convergence-stage routing is not specific to TC forecasting or TFIM: it also helps molecular prediction when coupled observables share an underlying structure but reach different learning stages.

7 Mechanism and Ablation

We analyze whether the gain of TAER-MoE comes from convergence-stage control rather than from a fixed task preference, post-hoc ensembling, or simply using more experts. Figure 6 shows the controller dynamics in a representative TC run. Early in training, track and intensity exhibit similar validation progress, and the controller remains close to ordinary joint training. When track begins to saturate while intensity continues to improve, the stage-separation signal increases the controller activation λ_c. The allocation then shifts training pressure away from the saturated track route and toward the still-improving intensity route. As the validation trends realign, the activation weakens and the allocation returns toward balance.

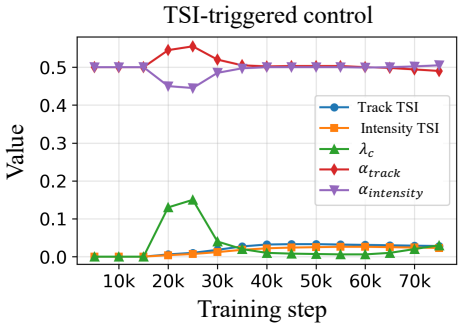

Figure 6: TC controller dynamics. TSI tracks task-wise validation progress, while λ_c activates under stage separation and shifts allocation between track and intensity.

This behavior suggests that TAER-MoE acts as a transient training controller. Its intervention appears when task learning stages separate and relaxes once the mismatch decreases, so the improvement is unlikely to be explained by a permanent preference for the harder task. It also differs from post-hoc ensembling, because the controller changes the training trajectory before the final model is obtained.

Table 4 further separates routing feedback from expert count and static averaging. The post-hoc ensemble remains substantially weaker, indicating that the gain is not obtained by averaging inde-

pendently trained experts. Adding convergence feedback improves the same expert pool, supporting the role of validation-stage information. The sparse-expert variant shows that raw capacity is not the main driver: better-timed allocation can regularize the saturated trajectory target while preserving intensity performance.

Table 4: **Ablation and efficiency analysis.** Controlled ablations separate routing feedback from expert count and post-hoc averaging.

Configuration	ADE ↓	Wind ↓	Purpose
Post-hoc ensemble	144.55±1.62	4.89±0.12	Static avg.
14 experts + TSI	61.59	3.56	TSI feedback
5 experts + TSI	58.60	3.55	Sparse routing
TAER-MoE	**56.88±1.30**	**3.67±0.03**	Full framework

The additional cost mainly comes from periodic validation used to update task pressure and expert-task gap statistics. The controller is lightweight and introduces no validation feedback at inference time; the final gap correction is absorbed as a fixed expert-task bias. In our setting, this adds about 15% wall-clock training overhead while leaving inference cost essentially unchanged.

Overall, these results support the central mechanism of TAER-MoE: its benefit comes from training-time capacity reallocation under asynchronous convergence, rather than from static ensembling, larger MoE capacity, or a fixed task-reweighting heuristic.

8 Future Work and Limitations

TAER-MoE targets coupled scientific observables with asynchronous convergence. Its benefits may diminish when tasks improve and saturate together.

Estimating task progress and route-conditioned generalization gaps requires periodic training-set and validation-set evaluation, adding training overhead without increasing inference cost. For large models or expensive scientific tasks, cheaper proxies based on gradients, uncertainty, or curvature may reduce this overhead.

Although our experiments span TC forecasting, QM9, and TFIM, they cover a limited number of target groups. Scaling to larger sets of coupled observables requires controllers that accommodate different learning stages without unstable competition. Broader evaluation across scientific domains is also needed to establish generality. In high-stakes applications, TAER-MoE should support expert judgment, physical forecasting systems, and uncertainty-aware risk assessment.

9 Conclusion

Scientific targets can share physical structure while reaching their best learning stages at different times. This paper studies asynchronous convergence, where one target saturates and becomes prone to overfitting while another still benefits from optimization. This is a learning-stage mismatch beyond gradient conflict or loss-scale imbalance.

We introduced Task-Aware Equilibrium Routing Mixture-of-Experts (TAER-MoE) for scientific multi-task learning. A spectral bias–variance view links task-dependent physical dimensions to different optimal training stages. TAER-MoE combines task-level validation descent with per-expert per-task generalization gaps, and derives routing from an entropy-regularized equilibrium objective.

Across TC forecasting, QM9 molecular property prediction, and TFIM ground-state observable prediction, TAER-MoE improves joint prediction over shared, MoE-based, gradient-based, and ensemble baselines. These results support a general principle for AI-driven science. Coupled targets benefit from shared representations, but routing should account for when each task and expert path no longer improves validation generalization.

AI Use Statement

Generative AI tools were used only for language editing and improving the clarity of the manuscript. All scientific content, experiments, analyses, and conclusions were developed and verified by the authors.

Reproducibility Statement

Implementation details, data processing, experimental protocols, and additional derivations are provided in the main text and supplementary material.

References

Cristian Bodnar, Wessel P Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A Weyn, Haiyu Dong, et al. A foundation model for the earth system. *Nature*, 641(8065):1180–1187, 2025.

Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová. Machine learning and the physical sciences. *Reviews of Modern Physics*, 91(4):045002, 2019.

Daniel R Chavas, Kevin A Reed, and John A Knaff. Physical understanding of the tropical cyclone wind-pressure relationship. *Nature communications*, 8(1):1360, 2017.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pp. 794–803. PMLR, 2018.

Zitian Chen, Yikang Shen, Mingyu Ding, Zhenfang Chen, Hengshuang Zhao, Erik G Learned-Miller, and Chuang Gan. Mod-squad: Designing mixtures of experts as modular multi-task learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11828–11837, 2023.

Damai Dai, Li Dong, Shuming Ma, Bo Zheng, Zhifang Sui, Baobao Chang, and Furu Wei. Stablemoe: Stable routing strategy for mixture of experts. *arXiv preprint arXiv:2204.08396*, 2022.

Yanqi Dai, Nanyi Fei, and Zhiwu Lu. Improvable gap balancing for multi-task learning. In *Uncertainty in artificial intelligence*, pp. 496–506. PMLR, 2023.

David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*, 2013.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. Pmlr, 2017.

Michelle Guo, Albert Haque, De-An Huang, Serena Yeung, and Li Fei-Fei. Dynamic task prioritization for multitask learning. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 270–287, 2018.

Yifei He, Shiji Zhou, Guojun Zhang, Hyokun Yun, Yi Xu, Belinda Zeng, Trishul Chilimbi, and Han Zhao. Robust multi-task learning with excess risks. *arXiv preprint arXiv:2402.02009*, 2024.

Cheng Huang, Pan Mu, Jinglin Zhang, Sixian Chan, Shiqi Zhang, Hanting Yan, Shengyong Chen, and Cong Bai. Benchmark dataset and deep learning method for global tropical cyclone forecasting. *Nature Communications*, 16(1):5923, 2025.

Adrián Javaloy and Isabel Valera. Rotograd: Gradient homogenization in multitask learning. *arXiv preprint arXiv:2103.02631*, 2021.

George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7482–7491, 2018.

Bin Lin, Zhenyu Tang, Yang Ye, Jinfa Huang, Junwu Zhang, Yatian Pang, Peng Jin, Munan Ning, Jiebo Luo, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *IEEE Transactions on Multimedia*, 2026.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. *Advances in Neural Information Processing Systems*, 34:18878–18890, 2021.

Shengchao Liu, Yingyu Liang, and Anthony Gitter. Loss-balanced task weighting to reduce negative transfer in multi-task learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 9977–9978, 2019.

Xu Liu, Juncheng Liu, Gerald Woo, Taha Aksu, Yuxuan Liang, Roger Zimmermann, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Moirai-moe: Empowering time series foundation models with sparse mixture of experts. *arXiv preprint arXiv:2410.10469*, 2024.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. *arXiv preprint arXiv:2202.01017*, 2022.

Joan Puigcerver, Carlos Riquelme, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. *arXiv preprint arXiv:2308.00951*, 2023.

Samyam Rajbhandari, Conglong Li, Zhewei Yao, Minjia Zhang, Reza Yazdani Aminabadi, Ammar Ahmad Awan, Jeff Rasley, and Yuxiong He. Deepspeed-moe: Advancing mixture-of-experts inference and training to power next-generation ai scale. In *International conference on machine learning*, pp. 18332–18346. PMLR, 2022.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):1–7, 2014.

Chandan Roy and Rita Kovordányi. Tropical cyclone track forecasting techniques—a review. *Atmospheric research*, 104:40–69, 2012.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-moe: Billion-scale time series foundation models with mixture of experts. *arXiv preprint arXiv:2409.16040*, 2024.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836, 2020.

Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193, 2012.

Hayoung Yun and Hanjoo Cho. Achievement-based training progress balancing for multi-task learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16935–16944, 2023.

Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3712–3722, 2018.

Yu Zhang and Qiang Yang. An overview of multi-task learning. *National Science Review*, 5(1): 30–43, 2018.

Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE transactions on knowledge and data engineering*, 34(12):5586–5609, 2021.