LLM-Orchestrated Mixture-of-Experts for Solubility Prediction

Samantha Alonso[1] Surya Appana[1] Ameya Kiwalkar[1] Hilary Wang[1] Alan Cheng[2] Claire Suen[2] Alec Glisman[2] David He[2] Xiadong Zhu[2]
[1]University of California Berkeley [2]Merck & Co.

Background

Computational models for predicting molecular properties are valuable tools in early-stage drug discovery, enabling researchers to prioritize promising candidates and reduce experimental costs. Since no single architecture can fully capture the vast complexity of chemical space, researchers typically select one model per endpoint, discarding valuable information from competing architectures. A mixture-of-experts (MoE) framework offers a superior alternative: combining predictions from multiple models using adaptive weighting schemes driven by a gating neural network. However, these neural networks assign numerical weights without scientific justification, which poses a transparency problem for researchers. To overcome this, our project explores a three-agent LLM-based framework that synthesizes model outputs dynamically and integrates molecular context to produce a final solubility prediction with an interpretable reasoning chain.

Data

"CC(=O)Oc1ccccc1C(O)=O"

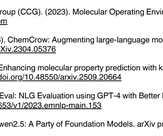


Figure 1: Simplified Molecular Input Line Entry System (SMILES)
Compact, ASCII-based text representations of molecular structures.
Our dataset[1] contains ~7,000 drug-like compounds represented as canonicalized SMILES strings with corresponding log solubility values, standardized using RDKit[2].

- Organized around molecular scaffolds, enabling model performance evaluation across distinct chemical families. This allows for stratified analysis of how structural variations influence solubility, ensuring a representative validation across the training and evaluation phases.
- Additional molecular descriptors calculated using Molecular Operating Environment[3] are provided for each compound. These serve as the primary features for individual experts and provide the critical molecular context used by the agentic framework to inform its weighting logic.

Overview

We evaluate two MoE approaches to multi-model prediction of molecular solubility. The first is a traditional mixture-of-experts baseline, consisting of five machine learning models. The second is our custom three-agent pipeline, where weighting is performed by an LLM rather than a neural network, which uses molecular context and past predictions to inform its decisions. The feasibility of this approach is grounded in work like ChemCrow[4], which showed LLMs can reason effectively over chemical problems when given the right contextual tools. By comparing these two approaches, we assess whether LLM-driven orchestration provides measurable improvement or advantages over conventional mixture-of-experts methods.

References

[1] Chen, J., et al. (2025). Data Scaling and Generalization Insights for Medicinal Chemistry Deep Learning Models. J. Chem. Inf. Model. https://doi.org/10.1021/acs.jcim.5c00538

[2] Landrum, G., et al. (2023). RDKit: Open-source cheminformatics. Zenodo. https://doi.org/10.5281/zenodo.8053810

[3] Chemical Computing Group (CCG). (2023). Molecular Operating Environment (MOE), version 2022.02. Montreal, QC. https://www.chemcomp.com

[4] Bran, A. M., et al. (2023). ChemCrow: Augmenting large-language models with chemistry tools. arXiv preprint arXiv:2304.05376. https://doi.org/10.48550/arXiv.2304.05376

[5] Zhou, P., et al. (2025). Enhancing molecular property prediction with knowledge from large language models. arXiv preprint arXiv:2509.20664. https://doi.org/10.48550/arxiv.2509.20664

[6] Liu, Y., et al. (2023). G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. Proc. EMNLP 2023 https://doi.org/10.18653/v1/2023.emnlp-main.153

[7] Qwen Team. (2024). Qwen2.5: A Party of Foundation Models. arXiv preprint arXiv:2412.15115.https://arxiv.org/abs/2412.15115

[8] ChromaDB Team. (2023). Chroma: The AI-native open-source embedding database. https://www.trychroma.com

[9] Heid, E., et al. (2024). Chemprop: A Deep Learning Package for Molecular Property Prediction. J. Chem. Inf. Model.https://doi.org/10.1021/acs.jcim.3c01250

[10] Walters, P. (2026). Practical cheminformatics tutorials. GitHub. https://github.com/PatWalters/practical_cheminformatics_tutorials

Methods

Traditional Mixture-of-Experts

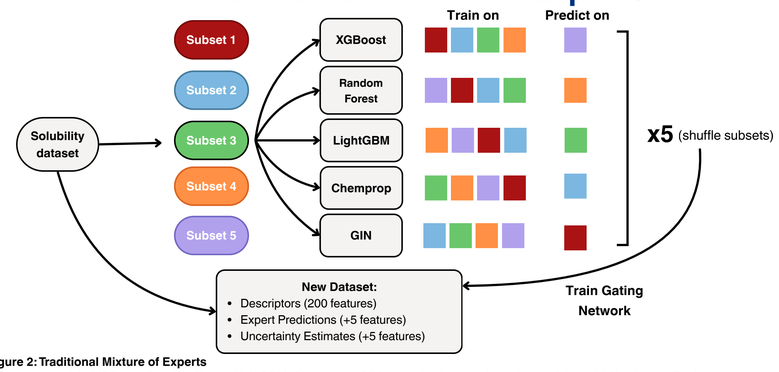


Figure 2: Traditional Mixture of Experts
Five expert models (XGBoost, Random Forest, LightGBM, Chemprop[9], GIN) are trained on rotating subsets of the solubility dataset. Their predictions and uncertainty estimates form a new dataset used to train a neural gating network to assign expert weights. Base experts pre-fit using Out-of-Fold (OOF) cross-validation to establish unbiased out-of-sample error baselines, ensuring gating network (and later, Agent 2), prioritize true generalization over training-set memorization.

Agentic Mixture-of-Experts

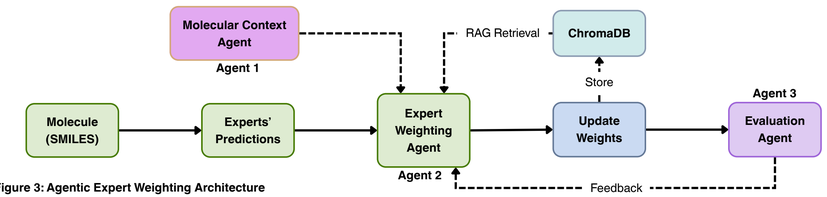


Figure 3: Agentic Expert Weighting Architecture
Input molecules are processed by an ensemble of five expert models (same as Traditional MoE), whose predictions are passed to the Agent 2 (Expert Weighting). Utilizing the Qwen2.5-14B[7] model, this agent performs chemical reasoning, augmented by Agent 1's molecular context and prior RAG-retrieved decisions, to dynamically re-weight the experts. Agent 3 (Evaluation) scores and stores each weighting decision in a ChromaDB[8] for future retrieval.

Architecture Modes:

1. **TRAIN_COLD (Sequential Learning)**: Establishes initial memory baseline. Agent 2 adapts weights molecule-by-molecule using Agent 1 molecular context (no RAG) to correlate structural features with expert error, populating the ChromaDB[8] with context-aware initial decisions. EMA momentum (m=0.20) is applied to dampen noisy per-molecule updates. Consolidated Agent 3 judge feedback is provided at end of training loop.
2. **TRAIN_WARM (Parallel Learning)**: Molecules processed in batches (size=5), with RAG and Context enabled. Agent 2 retrieves high-scoring prior decisions and relevant molecular context to refine weighting. Lower momentum (m=0.15) allows faster adaptation to judge feedback provided at end of each batch, while preventing "reasoning drift" in Agent 2.
3. **EVAL (Workflow Generalization):** Final performance is validated on remaining molecules using frozen weights from TRAIN_WARM phase. Agent 2 employs score-weighted RAG for zero-shot inference of molecules in batches (size=5) without updates. Agent 3 scores without providing Agent 2 feedback in order to test LLM's learnt chemical reasoning.

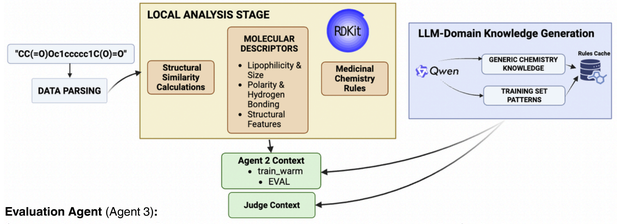
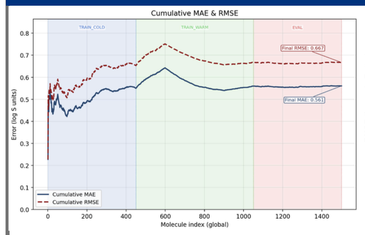


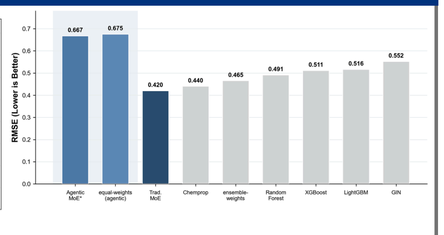
Figure 4: Molecular Context Architecture (Agent 1)
SMILES strings are standardized and featurized into RDKit[2] descriptors and medicinal chemistry alerts[10] (e.g., Lipinski, PAINS). This multi-dimensional profile is processed by a Qwen2.5-14B[7] domain-knowledge engine[5] to generate natural language insights, providing qualitative structural context for Agent 2's reweighting logic and a ground-truth reference for Agent 3's G-Eval evaluation.

Evaluation Agent (Agent 3):
Scores and evaluates Agent 2's weight updates and reasoning to provide dynamic, live-feedback.
1. Weighting (Deterministic): Validates Agent 2 assigned higher weights to models with smaller prediction errors, incentivizing Agent 2 to have higher confidence in low-error models.
2. Calibration (Deterministic): Measures mathematical proportionality of weight changes relative to the variance in expert error, preventing erratic oscillations in weight updates.
3. Reasoning (LLM-as-a-judge): Utilizes DeepEval's G-Eval[6] framework and Qwen2.5-14B[7] model to score Agent 2's reasoning. Decisions are graded based on structure, use of numeric evidence, and consistency with molecular context & prior high-scoring decisions.

Score-Weighted RAG Retrieval ("Memory"):
Agent 2 decisions (weight updates, reasoning, Agent 3 scores) are embedded into a ChromaDB[8] using SMILES-based structural vectors. During TRAIN_WARM and EVAL, past "lessons" are retrieved based on structural similarity and Agent 3's quality scores per decision, ensuring Agent 2 draws on high-scoring and successful past experience when reweighting experts.

Results





MoE Benchmark

- Best configuration of traditional MoE (n=70k, top-2 routing, $\lambda = 10^{-4}$) achieved RMSE≈0.420, outperforming all individual experts (e.g., Chemprop[9] ≈0.440) and average-weighted ensemble (≈0.465)
- Expert collapse was successfully avoided; weights converged to a globally optimal, near-uniform distribution across scaffolds (Chemprop: ~30%, others: ~15-20%).

Agentic MoE

- On a data-constrained stratified split (n=3k), the Agentic MoE (RMSE≈0.667) outperformed the equal-weight baseline (≈0.675). The consistent, yet gradual error decay observed in the TRAIN_WARM phase confirms Agent 2's ability to calibrate expert weights dynamically
- Agentic MoE maintained performance during the EVAL phase using frozen weights (RMSE Reduction ≈ -0.0078), providing evidence that LLM-driven chemical reasoning can generalize to unseen molecules without overfitting.
- G-Eval[6] logic scores improved from 0.62 (Cold) to 0.76 (Warm) as RAG provided historical context. Prompt step-compliance rose from 46% to 60%, showing that Agent 2's "chemical intuition" became more structured and reliable over time.

Conclusion and Future Work

We presented an agentic MoE framework that replaces "black-box" gating with interpretable chemical reasoning. Our large-scale traditional MoE benchmark (n=70,000) achieved a strong performance of RMSE≈0.420, validating the mixture-of-experts architecture as a robust baseline. For our Agentic MoE, while absolute performance gains were incremental (RMSE Net Gain ≈ 0.008), consistent improvement across phases provides evidence that LLM-guided orchestration is viable even with a fraction of the training data. Though a direct comparison between Agentic MoE (n=3,000, stratified) and the traditional benchmark (n=70,000, random) is confounded by differing sampling distributions and sizes, the agentic workflow's ability to maintain a performance lift despite training on a fraction of the data suggests high efficiency and generalization. A core benefit of our research is architectural transparency. By grounding expert weights in RAG-retrieved reasoning and molecular descriptors, this framework provides an auditable "glass-box" lens for high-stakes drug discovery, where understanding chemical logic is as critical as the prediction itself.

Implementation Details

- **Computational Setup**: LLM reasoning is powered by Qwen2.5-14B[7] across all 3 agents. To optimize latency, models were hosted on Modal GPU clusters using vLLM for high-throughput inference and paged attention.
- **Constraint Note**: Due to high compute requirements, agentic MoE was validated on 3,000 stratified sampled molecules representative of the dataset (1500 fit, 1500 LLM-based)

Future Directions

- **RLHF / RLAI**: Explore RL techniques to fine-tune Agent 2's underlying LLM weights based on accumulated Agent 2 judge feedback, enabling true agent learning across molecules
- **Property & Model Scaling**: Expand agentic framework to molecular properties beyond solubility; evaluate higher-parameter models over more time for deeper reasoning insights
- **Inference Optimization**: Consolidate the Molecular Context Agent architecture into a single-call inference pipeline to reduce computational overhead and latency