AutoChemSchematic AI: Agentic Physics-Aware Automation for Chemical Manufacturing Scale-Up

S S Srinivas, Shivam Gupta, Venkataramana Runkana

Tata Research Development and Design Center

December 13, 2025

Limitations of Current AI Pipelines in Molecule and Material Design

- Most current AI pipelines for molecular and material design are not truly end-to-end.
  - **De novo molecule design for specialty chemicals** primarily focuses on:
    - **Generating and optimizing candidate molecules by predicting properties and selecting target-fit candidates**.
  - **Inverse material design for high-performance materials** primarily focuses on:
    - **Using target properties and constrained optimization to identify optimal material structures**.
- However, they often overlook industrial manufacturability — creating a **synthesis gap where discoveries remain theoretical**:
  - The main drawbacks are:
    - **Production feasibility** — can the designed molecule be synthesized at scale?
    - **Process engineering** — how to manufacture efficiently.
    - **Manufacturability schematics** — e.g., generate industrial diagrams (PFDs and PIDs) for production?
- Bridging lab-scale or simulation-based design with industrial-scale manufacturing remains a major bottleneck and limits end-to-end AI pipelines.
- **Auto-generating industrial production diagrams (PFDs and PIDs) is essential for:**
  - How the molecule can be synthesized at **full industrial scale**.
  - What the complete **industrial process and flowsheet** look like.
  - How the process should be **monitored, controlled, and stabilized**.
  - What **equipment, utilities, and control systems** are required for reliable operation.

Introduction

- **Process Flow Diagrams (PFDs)**
  - Show how **raw materials** are transformed into **intermediates** and **final products**.
  - Show major steps in the process: flow of **materials** and **energy**.
- **What PFDs depict**
  - **Major equipment**: reactors, pumps, heat exchangers, etc.
  - **Material streams**: flow rate, composition.
  - **Process conditions**: temperature, pressure.
  - **Utilities**: steam, cooling water.
- **Focus**
  - **What happens** in the process (material and energy transformations).
  - **Where it happens** (location and role of major equipment).
- **Used For**
  - Process **design**, **simulation**, and **optimization**.

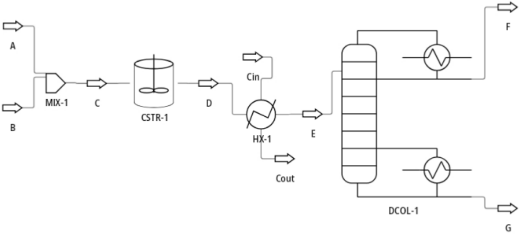


Figure: Schematic illustrating the process flow diagram (i.e., core process operations and transformations).

Introduction

- **Process and Instrumentation Diagram**
- **Purpose**
  - Depicts detailed instrumentation and control of the chemical process.
- **Enhances the PFD by adding**
  - **Instrumentation**: sensors (temperature, pressure, flow), control valves, indicators.
  - **Safety systems**: alarms, interlocks, relief systems.
- **Focus**
  - How the **process is monitored and controlled**.
    - **Product Quality** — to meet product specifications.
    - **Efficiency & Energy Optimization** — minimizes energy use and maximizes throughput.
    - **Reliable Operation** — ensures stable plant performance.
- **Used For**
  - Detailed **design, troubleshooting, process optimization**.

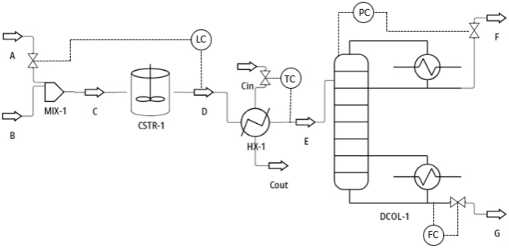


Figure: Schematic illustrates PID of a chemical process showing instrumentation and control systems.

High-level Goal

Given a chemical product, design its industrial production diagrams

- Represent industrial PFDs and PIDs as graph structures (nodes = equipment, edges = streams).
- Chemical flowsheets follow a multi-step process sequence, where each process step defines *what transformation must occur* (e.g., reaction, separation, purification, heat exchange).
- **For each process step (WHAT must happen):**
  - Which feasible unit operations can achieve it (HOW it is done)?
  - What design and operating decisions are required?
  - How the step integrates with upstream and downstream units.
  - **Key challenges:**
    - Identifying feasible alternatives.
    - Defining safe and operable conditions.
    - Ensuring valid mass and energy balance across the entire process plant.
    - Incorporating valid, reliable control systems.

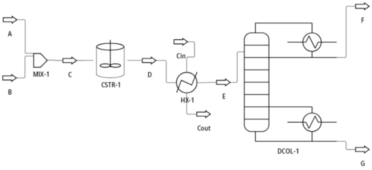


Figure 1: Process Flow Diagram (PFD).

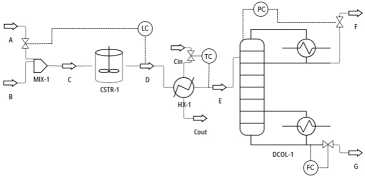


Figure 2: Process Instrumentation Diagram (PID).

What We Propose

Closed-Loop, End-to-End AI Framework Enabling Lab-to-Plant Scale-Up

- Introduces a **self-driving lab framework** for **auto-generating high-fidelity PFDs and PIDs** for chemical processes.
- Streamlines the transition from **simulation** → **lab** → **pilot** → **plant**.
- Ensures industrial viability by advancing only **sustainable, efficient, and scalable** process routes.

Closed-Loop Optimization

- Functions as an **end-to-end process design and flowsheeting tool** with minimal human intervention.
- Automates the **design, simulation, and optimization** of chemical processes.
- Integrates **first-principles (or physics)-informed modeling and process simulation** with **adaptive learning** into a continuous feedback loop for robust process design.
- Continuously **self-improves** through iterative simulation feedback, enhancing reliability and performance.

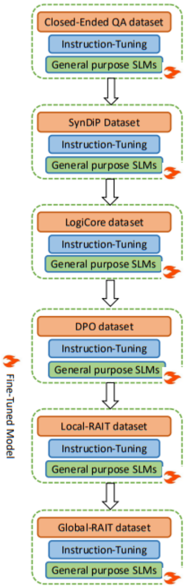
Methodology

- General-purpose, lightweight, white-box SLMs are trained primarily on English literature.
- They lack high-fidelity knowledge specific to chemical manufacturing.
- They require customization to achieve expert-level accuracy, reliability, and detail.

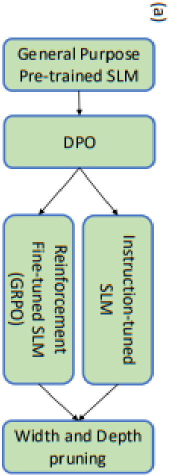
- **Purpose:** Domain specialization
  - Train small models (e.g., Llama-3.2-1B, SmolLM2-135M) on chemical manufacturability tasks such as PFD and PID generation, interpretation and analysis tasks.
- **Method:** Multi-stage instruction tuning
  - Use synthetic instruction–response datasets from teacher LLMs (knowledge distillation + transfer learning)
  - Validate Q&A pairs and filter with reward models (e.g., Nemotron-4-340B)
- **Outcome:**
  - Instruction-tuned SLMs interpret PFDs, PIDS accurately
  - Analyze and reason over PIDs, PFDs
  - Generate chemical manufacturability descriptions reliably for chemicals.

- Curated a custom database of 1,120+ chemicals across pharmaceuticals, FMCG, petrochemicals, and so on.
- Data extracted from major industrial manufacturers (e.g., BASF, Dow, DuPont, Solvay, Mitsubishi, Bayer, Evonik, SABIC, so on).
- Dataset contains two components:
  - **ChemAtlas**: Core set of 1,020 chemicals.
  - **ChemEval**: Evaluation subset for benchmarking.

Methodology



- **Factual QA**
  - Builds foundational industrial manufacturing knowledge and factual recall.
- **SynDIP Dataset**
  - Contains instruction–response pairs describing PFD and PID details for industrial chemicals across sectors.
- **LogiCore**
  - Focuses on multi-step reasoning: justifying process design choices and validating continuous-flow sequencing.
- **DPO (Direct Preference Optimization)**
  - Uses preference-labeled instruction–response pairs to align SLM outputs with expert-like reasoning.
- **Local RAIT**
  - Grounds responses in localized retrieved chunks for high-precision, context-aware generation.
- **Global RAIT**
  - Supports multi-hop, cross-document reasoning via retrieval over semantically clustered sources.


(a)
General Purpose Pre-trained SLM
DPO
Reinforcement Fine-tuned SLM (GRPO)
Instruction-tuned SLM
Width and Depth pruning

Inference

Overall Architecture

SLM Domain Adaptation

- SLM is aligned with DPO.
- Further refined via instruction tuning or policy-gradient RL.
- Width/depth pruning optionally improves efficiency.

Meta-Agent orchestrates the operational RAG workflow

- The SLM retrieves context from memory and graph databases.

Inference Acceleration

- **Flash Attention**: Fast attention.
- **Paged KV Caching**: Memory-efficient key–value reuse for long-context decoding.
- **Lookahead Decoding**: Multi-step token prediction for faster decoding.
- **Test-Time Scaling**: Optimizes decoding hyperparameters for reliable generation.

Critique-Agent for Response Refinement

- Nemotron-4-340B reward model, LLM-as-a-judge (GPT-4o), and Human evaluation.

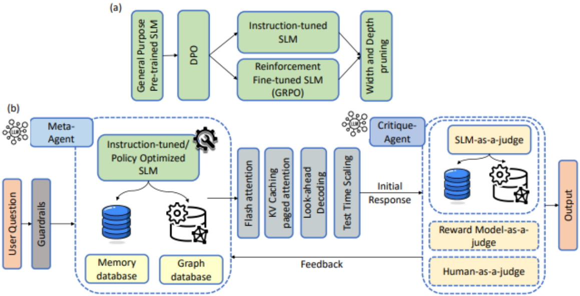


Figure: Overview of the integrated framework.

Experimental Results

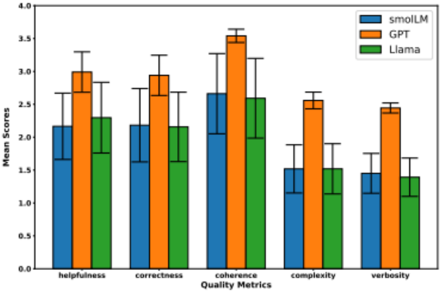


Figure: Framework performance evaluated on the ChemEval benchmark using Nemotron-4-340B (0–4 reward scale).

- GPT-4o: highest performance.
- Llama-3.2-1B (fine-tuned): second; lower verbosity, higher variance.
- SmolLM2-135M(fine-tuned): lowest, but matches Llama-3.2-1B in complexity and verbosity.

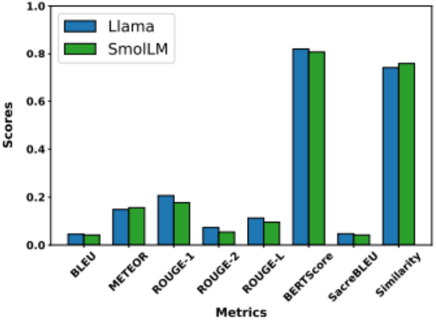


Figure: Compared Llama-3.2-1B and SmolLM2-135M on ChemEval benchmark dataset using BLEU, METEOR, ROUGE, SacreBLEU, BERTScore, and cosine similarity..

- Llama-3.2-1B: higher overlap-based scores across BLEU/METEOR/ROUGE.
- Both models show similar semantic similarity (BERTScore, cosine).

Future Work

Not Fully-Closed?

Inference Acceleration

- AI framework uses an offline chemical simulator-in-the-loop to verify feasibility of generated chemical process diagrams.
- However, this validation is post-generation and does not feed back into the model during training or inference.
- It is not a fully closed-loop system, as it lacks real-time or iterative feedback from the simulator during training or inference.

Extend the Chemical Database

- Current database (1,120 chemicals) is impressive but still small compared to real industrial diversity.

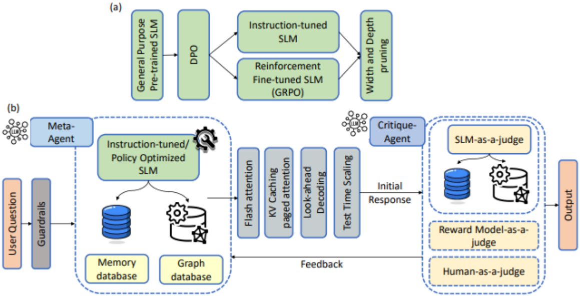
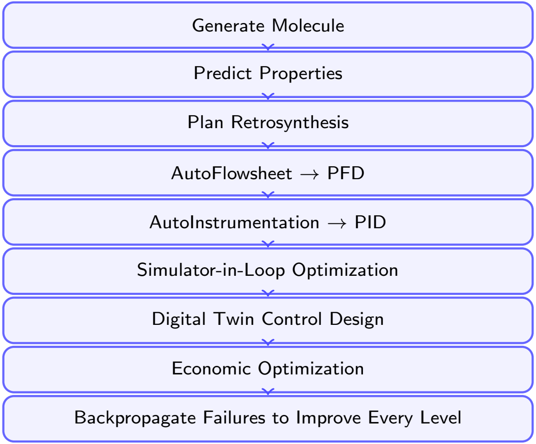


Figure: Overview of the integrated framework.

End-to-End Autonomous Chemical Manufacturing Pipeline



Generate Molecule

Predict Properties

Plan Retrosynthesis

AutoFlowsheet → PFD

AutoInstrumentation → PID

Simulator-in-Loop Optimization

Digital Twin Control Design

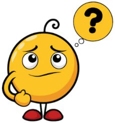
Economic Optimization

Backpropagate Failures to Improve Every Level

So future AI systems will propose families of manufacturable molecules optimized for downstream synthesis and production.

Thank You

Questions?



For More Details, Scan Me