LING682 FINAL PROJECT

Controlled Molecular Generation

Using Sequential Monte Carlo Sampling with Small Language Models

Chih-Chao Hsu McGill University

November 2025

The Challenge

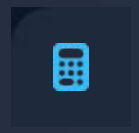
De Novo Discovery

Modern drug discovery aims to reveal novel, valid compounds satisfying specific physicochemical properties. Traditional methods are limited by predefined rules and fragment libraries that restrict accessible chemical space.

The Control Problem

A persistent challenge in AI-based generation is achieving reliable control over properties like **Molecular Weight (MW)**, **Lipophilicity (logP)**..etc without incurring massive computational costs.

Sequential Monte Carlo (SMC)

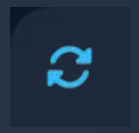


1. Importance Sampling

Estimates expectations under an intractable target distribution by drawing samples from a proposal distribution and reweighting them effectively.

2. Sequential Update

Weights are updated dynamically as tokens are appended. Potential functions score how well a partial sequence matches the desired constraints.



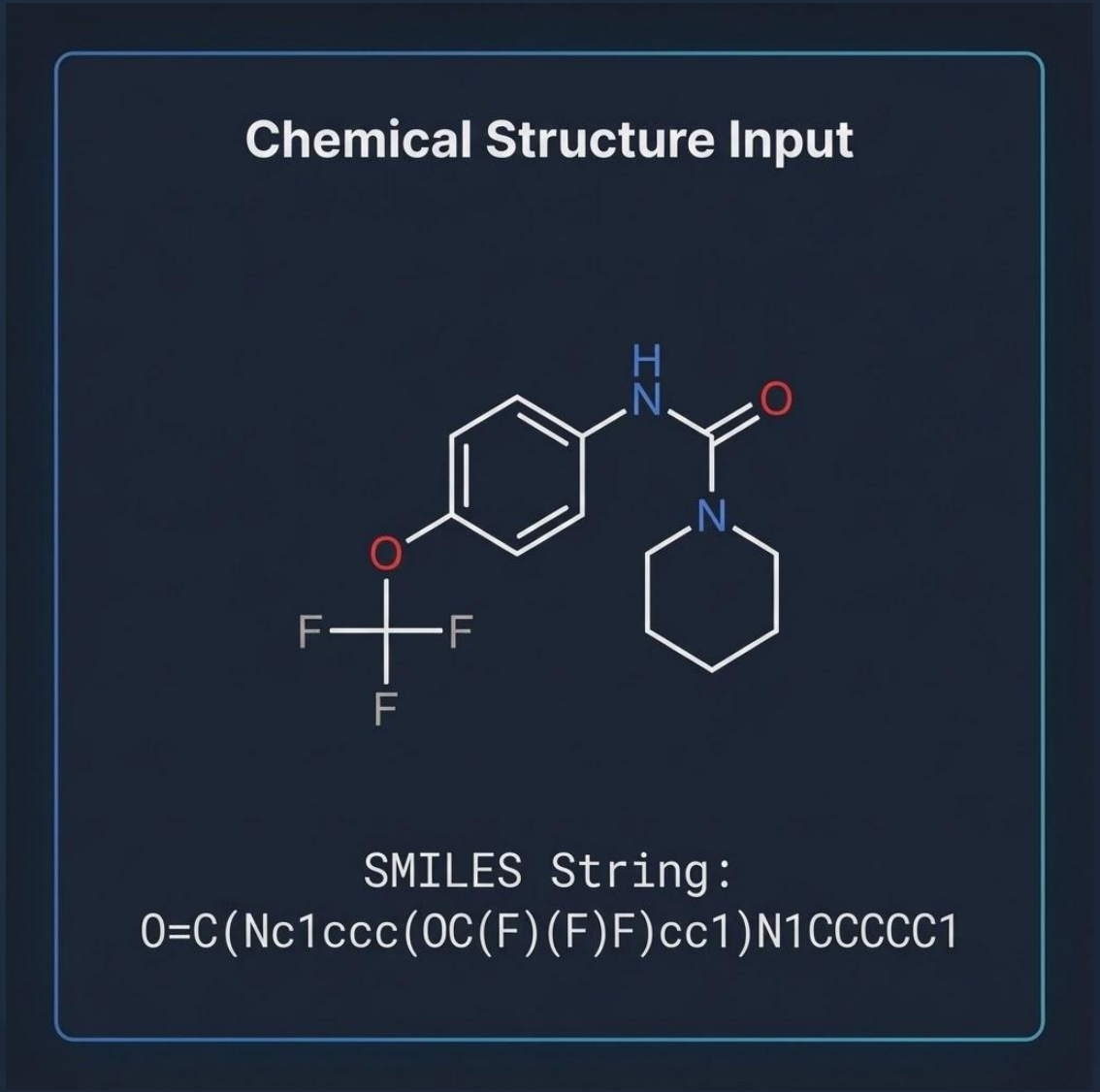
3. Adaptive Resampling

Prevents particle collapse by replicating high-weight particles and discarding low-weight ones.

Experimental Setup

We compare a small model against a massive fine-tuned model on constrained molecules generation.

Feature	GPT2-ZINC	Smiley Llama
Size	87 Million	8 Billion
Training	ZINC (~480M)	ChEMBL (~2M)
Input	SMILES Prefix	Natural Language
Method	SMC-Guided	Fine-tuning

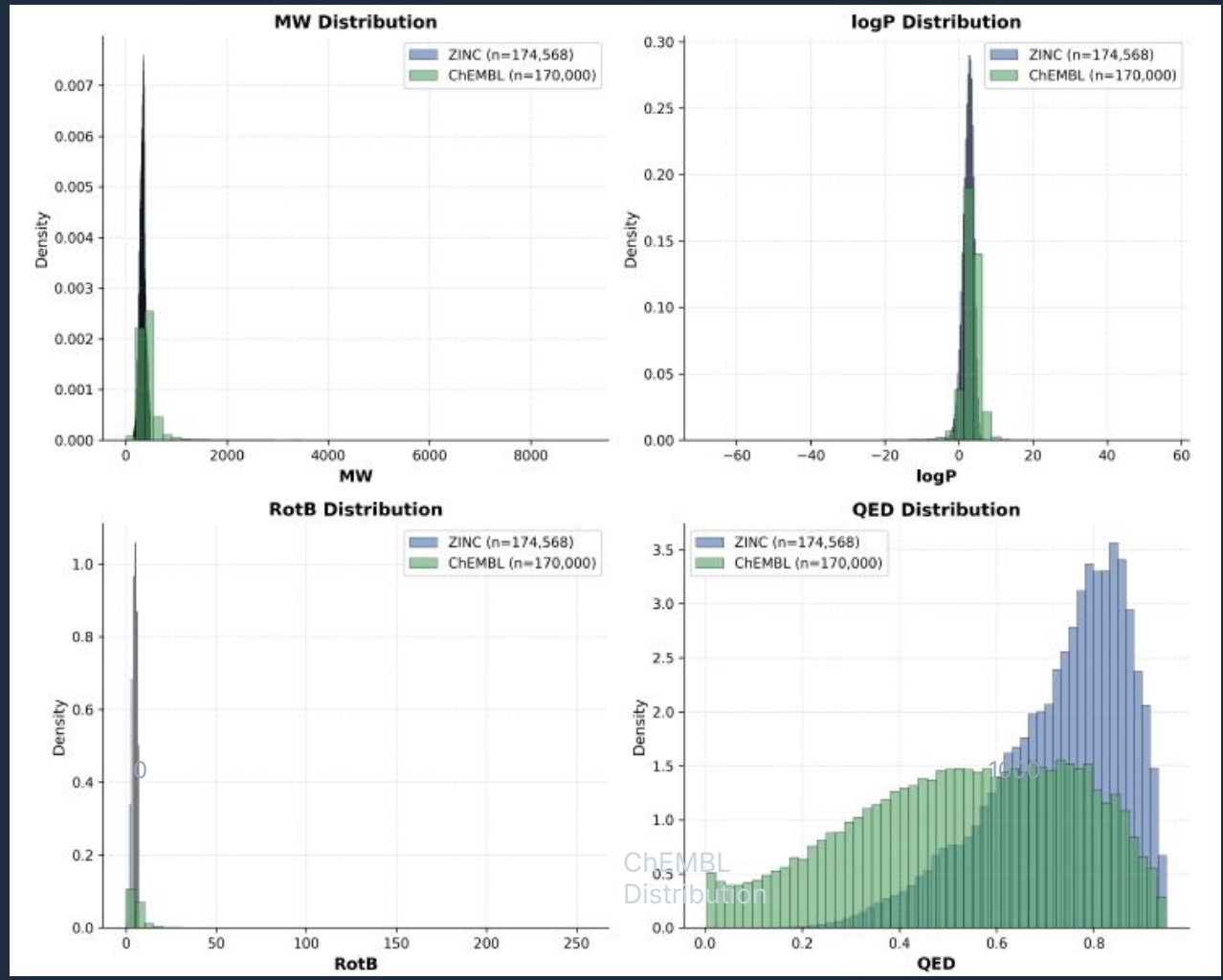


Data Distribution Analysis

⚖ Property Overlap

Despite the immense difference in dataset size (ZINC: ~480M vs ChEMBL: ~2M), the distributions of key properties are remarkably similar.

- ✓ **ZINC :** High density in drug-like space.
- ✓ **ChEMBL :** Broad pharmacological coverage.
- ✓ **Conclusion:** The chosen properties (MW, logP, RotB) follow similar manifolds in both datasets, ensuring a fair comparison of generative capabilities.



Constraint Paradigms

Range-Based

Testing strict dual-bound control derived from data percentiles.

- **Loose:** 5th-95th percentile
- **Tight:** 25th-75th percentile
- **Ultra-tight:** 40th-60th percentile

Gradual (Upper-Bound)

Matching instruction-following formats with increasing difficulty.

- **Loosen:** MW ≤ 500
- **Tight:** MW ≤ 400, logP ≤ 4
- **Ultra-tight:** MW ≤ 350, logP ≤ 3.5, RotB ≤ 8

Potential & Reward Design

Core Philosophy: Binary filtering creates optimization cliffs. We make it smooth..

Dual-Profile Shaping

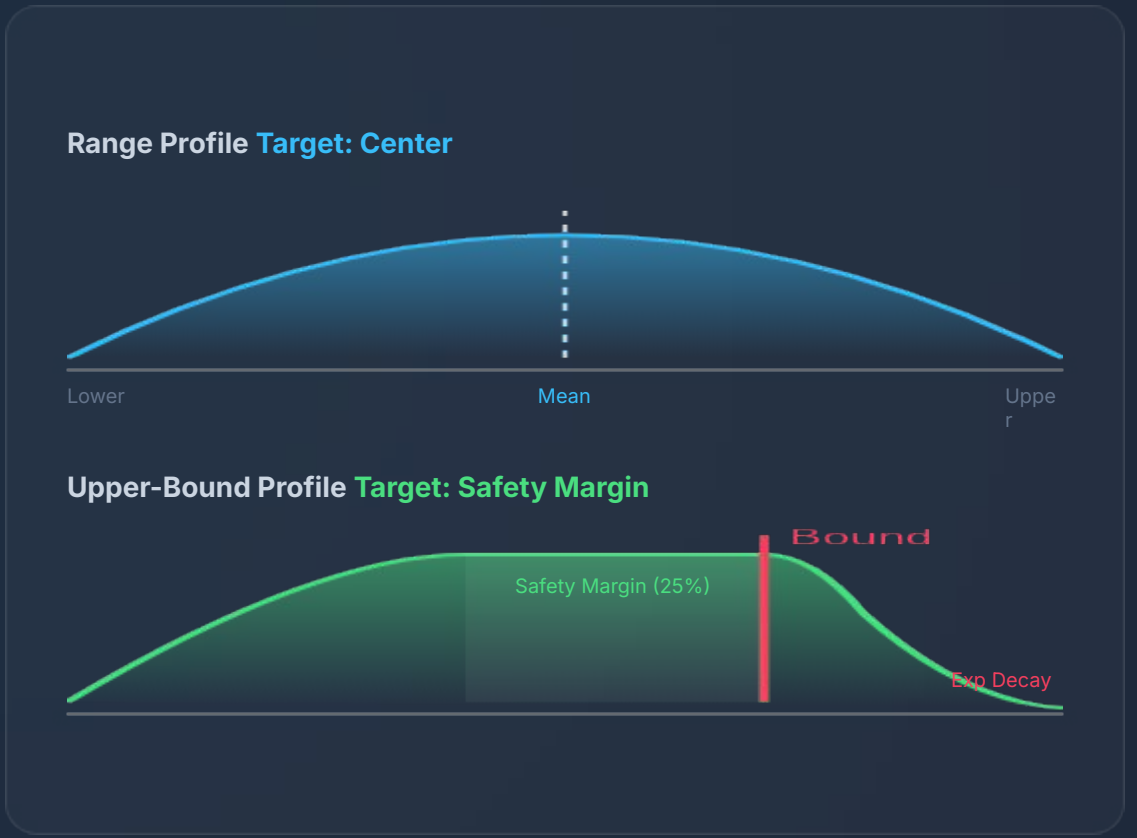
We shape the reward surface differently based on the constraint paradigms to guide the SMC sampler.

1. Range Constraints (Symmetric)

Standard Gaussian centering. The optimizer is pushed toward the exact middle of the range to maximize the chance to be within the range.

2. Upper Bound (Asymmetric)

Right: Exponential penalty for violations.
Sweet Spot: Maintains peak reward within a 25% "Safety Margin" below the bound.
Left: Decays as range profile.



Initialization Prefixes

To prevent mode collapse and ensure structural diversity, we initialize SMC particles with 20 distinct chemical prefixes ranging from aromatic rings to aliphatic chains.

`C1=CC=` Aromatic Core	`CCN` Amino Aliphatic	`COC` Polar Ether	`CC(=O)` Carbonyl
`C1CN` Heterocycle	`CCS` Sulfur Motif	`CCOCON` Extended Polar	`CCC` Generic Aliphatic
`C1CCCCC1` Aliphatic Ring	`c1ccccc1` Benzene	`n1ccccc1` Pyridine	`N1CCCCC1` Piperidine
`O1CCNCC1` Morpholine	`NC(=O)` Amide	`NS(=O)(=O)` Sulfonamide	`NC(=O)N` Urea
`C#N` Nitrile	`Clc1ccccc1` Aryl Chloride	`P(=O)(O)O` Phosphonate	`COC` Ether Builder

Results

100%

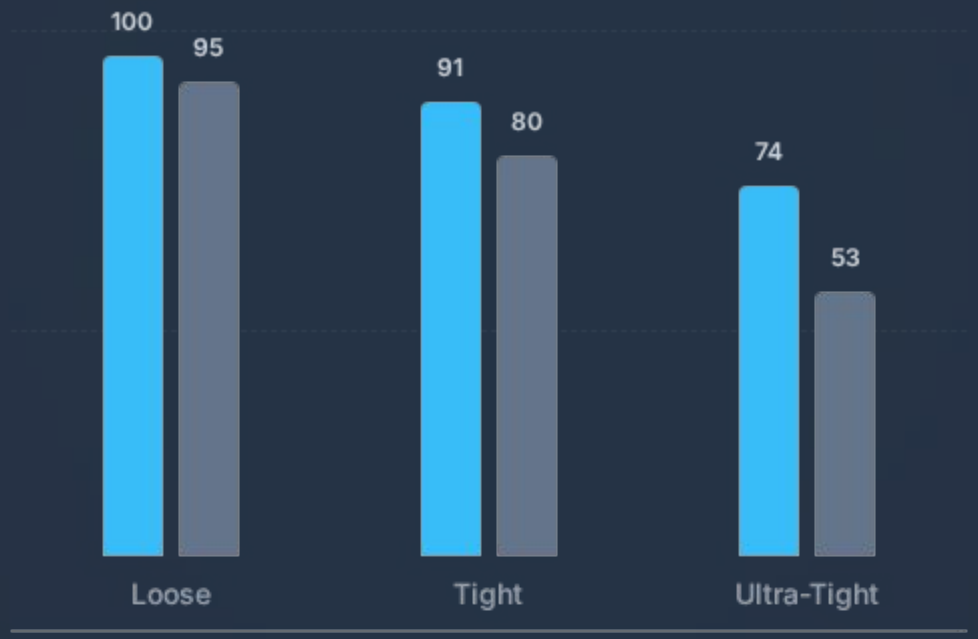
Validity

Perfect structural validity achieved across all models and constraint settings.

Adherence %

GPT2+SMC SmileyLlama

Comparison under "Gradual" constraints.



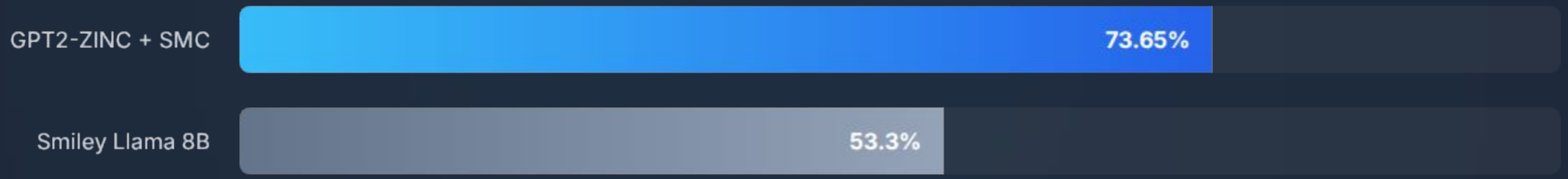
0.91

Diversity

High chemical diversity maintained via distinct prefix initialization.

Highlight: Ultra-Tight Constraints

Comparing the percentage of generated molecules that strictly adhere to the most difficult "Gradual" constraints.



Insight: Despite being ~100x smaller, the SMC-guided model outperforms the 8B parameter model by over 20 percentage points.

Key Takeaways

Small Models Could Win

A specialized small model (87M) with SMC can rival or surpass a larger fine-tuned model (8B).

SMC Effectiveness

Potential-based rewards steer generation towards desired properties without fine-tuning.

Computational Efficiency (Not 100% confirmed)

This approach operates at a fraction of the cost of large-scale inference.

Current Limitations

Extremely narrow, dual-bound ranges remain a difficult open problem.

Future Work

Non-Linear Restrictions

Refining SMC potentials to handle more complex chemical constraints beyond simple ranges.

Multi-Property Balancing

Developing sophisticated weighting mechanisms to balance conflicting objectives.

Core-Driven Synthesis

Exploring scaffold-based generation where the model builds around a fixed core.