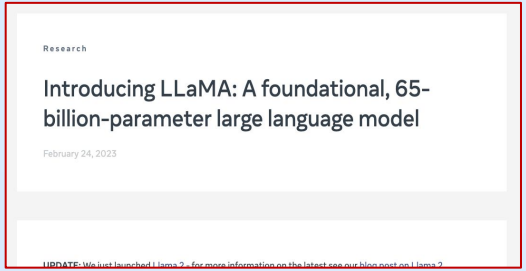
Beyond Experimental Structures: Advancing Materials Discovery with Generative AI

Anuroop Sriram
Research Engineer, FAIR Chemistry, Meta

Meta AI

FAIR & FAIR Chemistry

- Meta FAIR = Meta Fundamental AI Research
- FAIR is committed to open and reproducible research
- World-class compute resources for dataset and model development
- FAIR Chemistry: catalysis, direct air capture, and, more recently, display materials for AR/VR

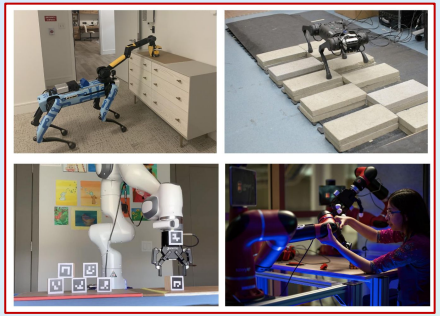


Language

LLaMA[1,2,3]:
ai.meta.com/llama

Vision

Segment Anything,
Make-a-Scene, Make-A-Video,
Masked Autoencoders, etc.

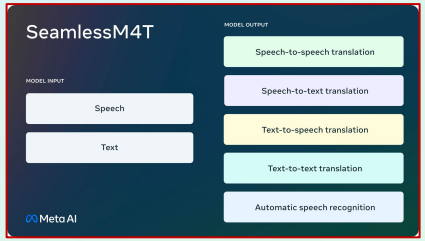


Embodied AI

Navigation, mobile
manipulation,
tactile sensing, etc.

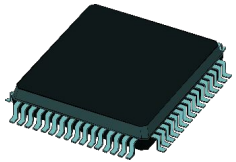
Speech

Seamless, No Language
Left Behind, etc.

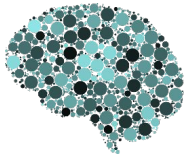


Material Discovery with AI

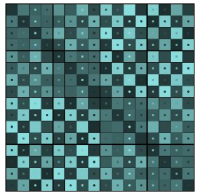
Simulate ~10^5 compounds



Train machine learning model

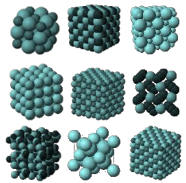


Explore all known compounds

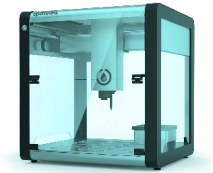


Rapid ML model screening

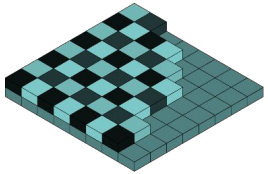
~10^3 – 10^4 compounds



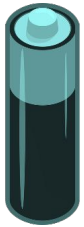
Robotic synthesis



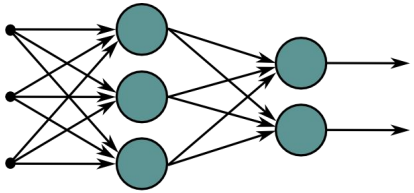
Rapid characterization



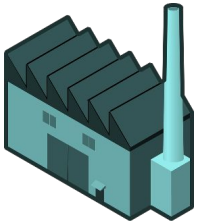
Device design



Learning with AI



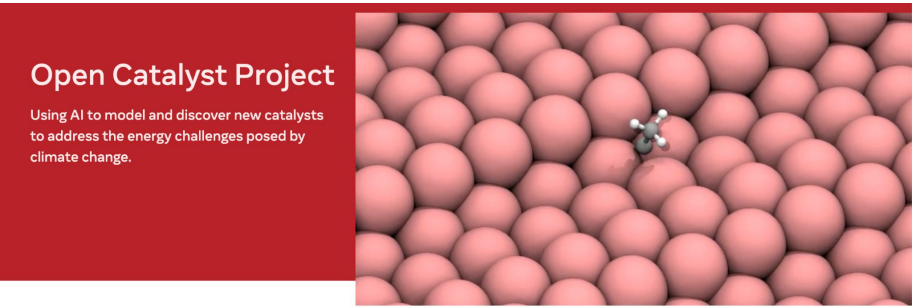
To market



AI informs experimental design

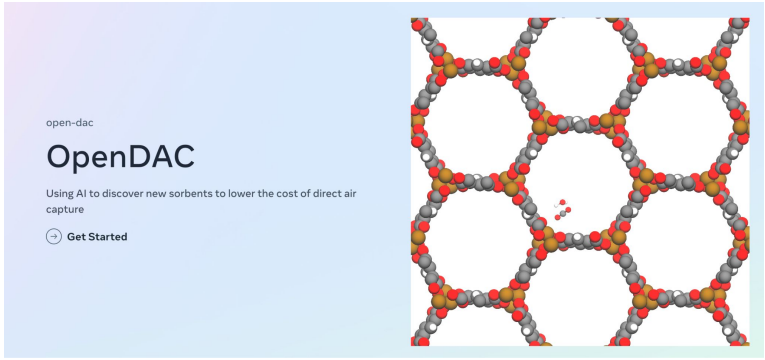
FAIR Chemistry Datasets

Discover novel catalysts for energy storage



OC20 & OC22 datasets
>1M DFT relaxations, 80 adsorbates
https://opencatalystproject.org/

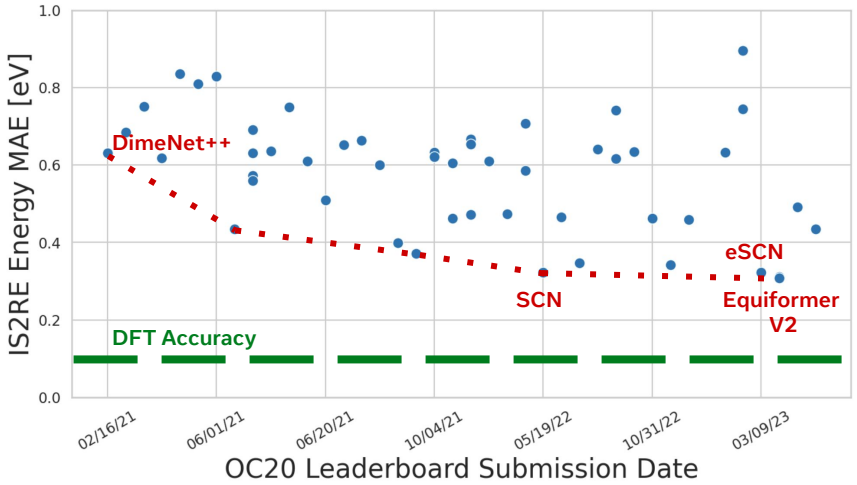
Discover new sorbents for carbon capture



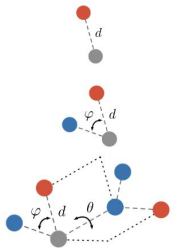
OpenDAC23 dataset
175K DFT relaxations with ~9K MOFs
https://open-dac.github.io/

>1B CPU Hours of Compute

Progress in ML Potentials



GemNet-OC

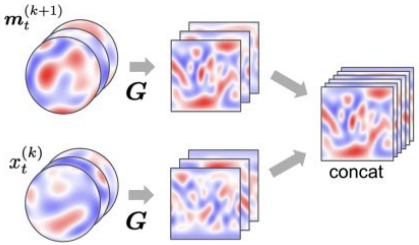


$O(n^2)$ 30 neighbors

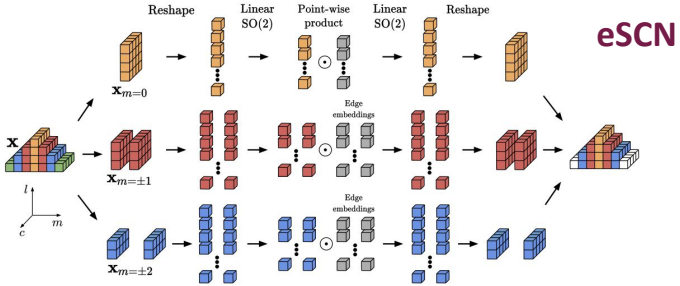
$O(n^3)$ 30 neighbors

$O(n^4)$ 8 neighbors

SCN



eSCN

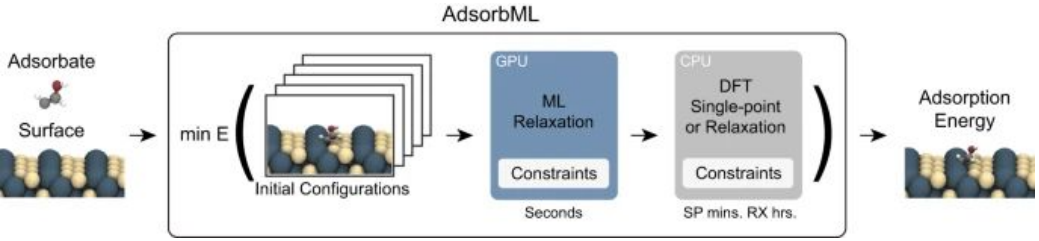


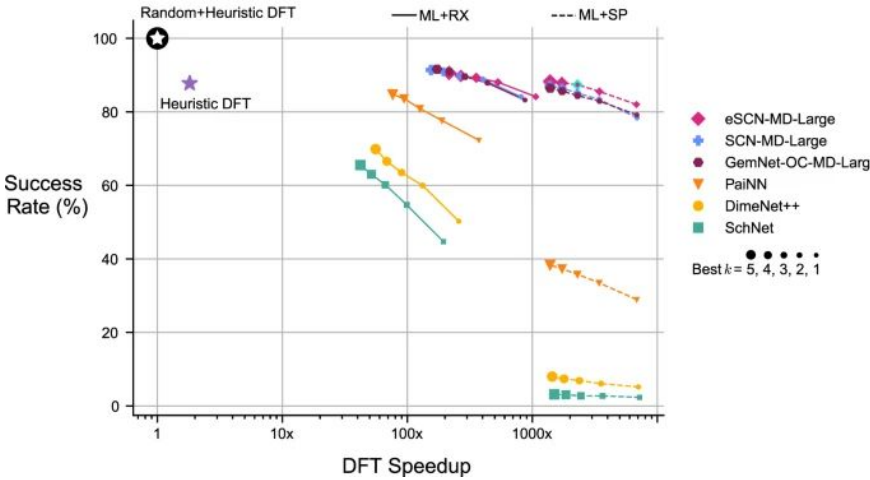
J. Gasteiger, et al. TMLR (2022)

C.L. Zitnick, et al. NeurIPS (2022)

S Passaro, C.L. Zitnick, ICML (2023)

AdsorbML: 1000x faster Catalyst Screening

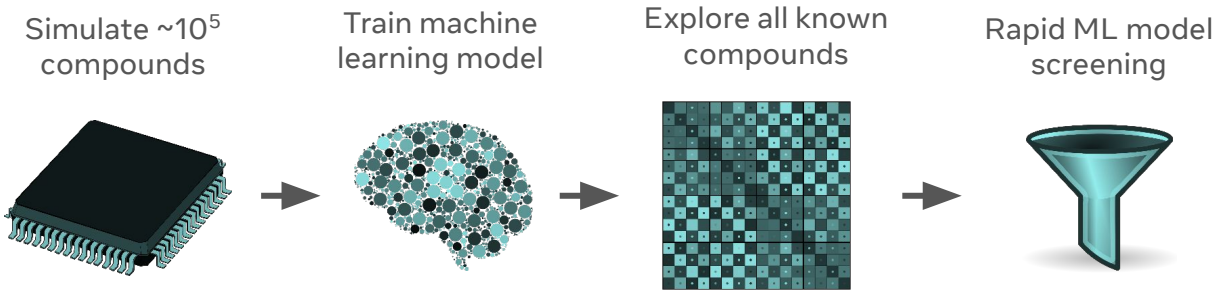




Demo & API: https://open-catalyst.metademolab.com/

Lan, et al. "*AdsorbML: a leap in efficiency for adsorption energy calculations using generalizable machine learning potentials*". Nature Computational Materials, 2023.

What's Next? Materials discovery beyond known materials



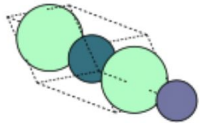
This pipeline is limited to known materials. How do we expand the search space beyond these known materials?

Use generative models to discover new materials!

Crystal Generation

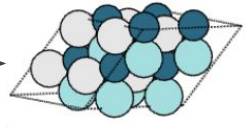
Crystal Structure Prediction

Pm_2ZnRh → Generative Model →



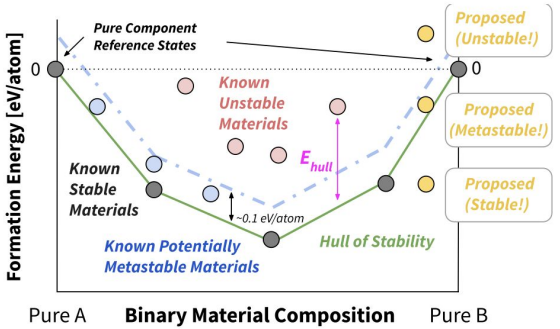
De Novo Generation

Generative Model →

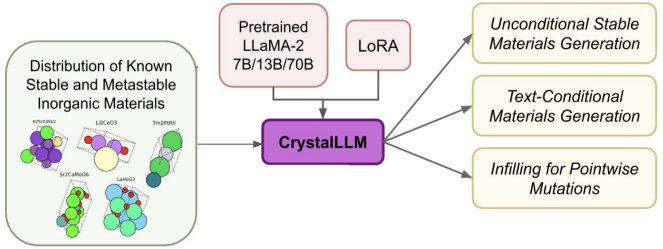


Match Rate = Percentage of generated materials that match the ground truth structure

Stability Rate = Percentage of generated materials that are stable (E-hull < 0)

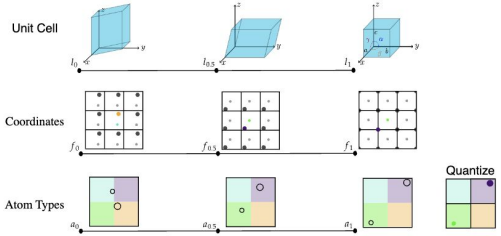


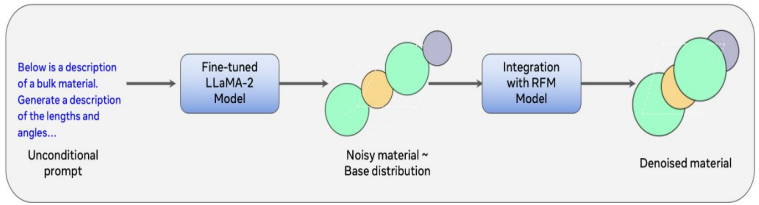
Generative Models



CrystalLLM
Fine-tuned LLaMA2 models for materials
(Gruver et al., ICLR 2024)

FlowMM
Riemannian Flow Matching adopted to materials
(Miller et al., ICML 2024)

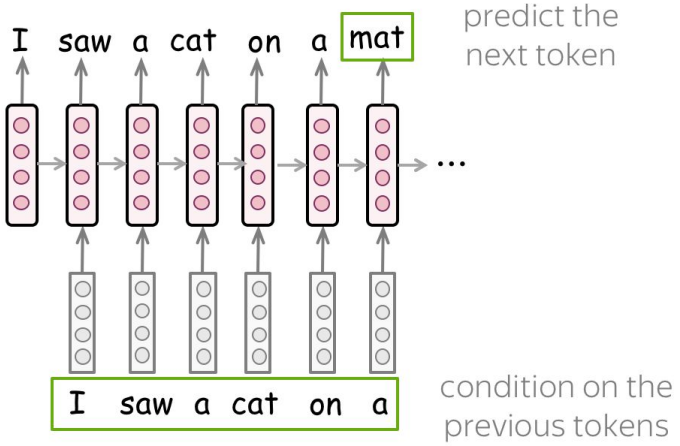


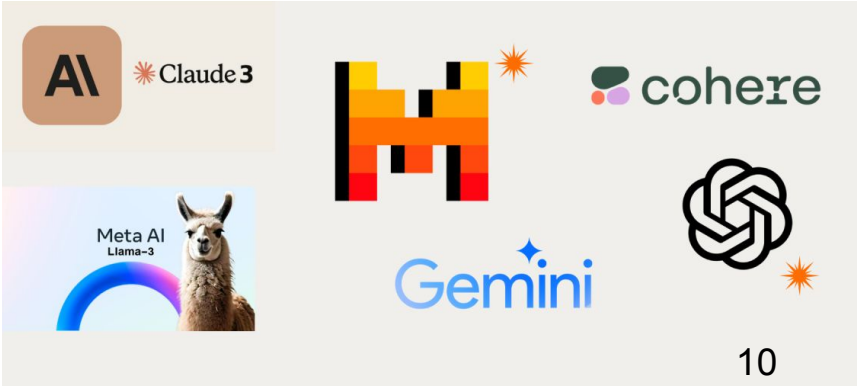


FlowLLM
Combined LLaMA2 fine-tuning + Flow Matching
(Sriram et al., Under review at NeurIPS 2024)

Large Language Models

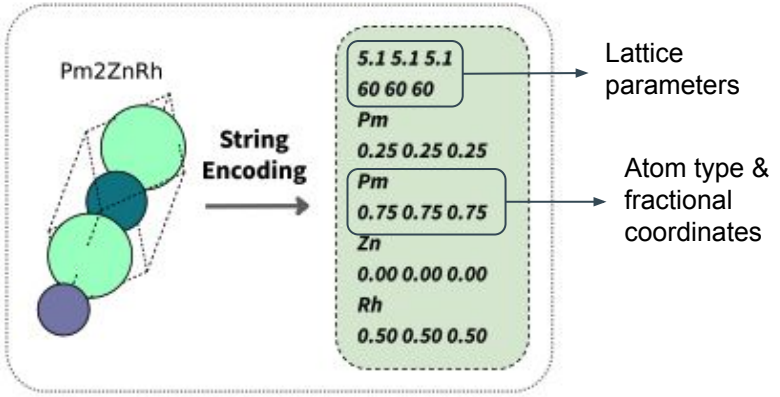
- Language models model **P(text)** by predicting each token given the previous tokens.
- They can model pretty much anything that can be represented as text (sequence of discrete tokens).
- Fine-tuning base LLMs can adapt these LLMs to new domains.





CrystalLLM: LLMs for Materials Generation

- Represent crystals as strings, then fine-tune LLAMA-2 models to generate stable materials directly.
- Very simple approach, yet obtained state-of-the-art performance.



"*Fine-Tuned Language Models Generate Stable Inorganic Materials as Text*", Gruver et al. ICLR 2024

CrystalLLM: LLMs for Materials Generation

- Represent crystals as strings, then fine-tune LLAMA-2 models to generate stable materials directly.
- Very simple approach, yet obtained state-of-the-art performance.
- Can support complex prompting.
  - Supports conditional generation, de novo generation, infilling etc.
  - Can condition on formula, properties, molecule etc.

Generation Prompt

<s>Below is a description of a bulk material. [The chemical formula is Pm2ZnRh]. Generate a description of the lengths and angles of the lattice vectors and then the element type and coordinates for each atom within the lattice:

[Crystal string]</s>

Below is a description ...

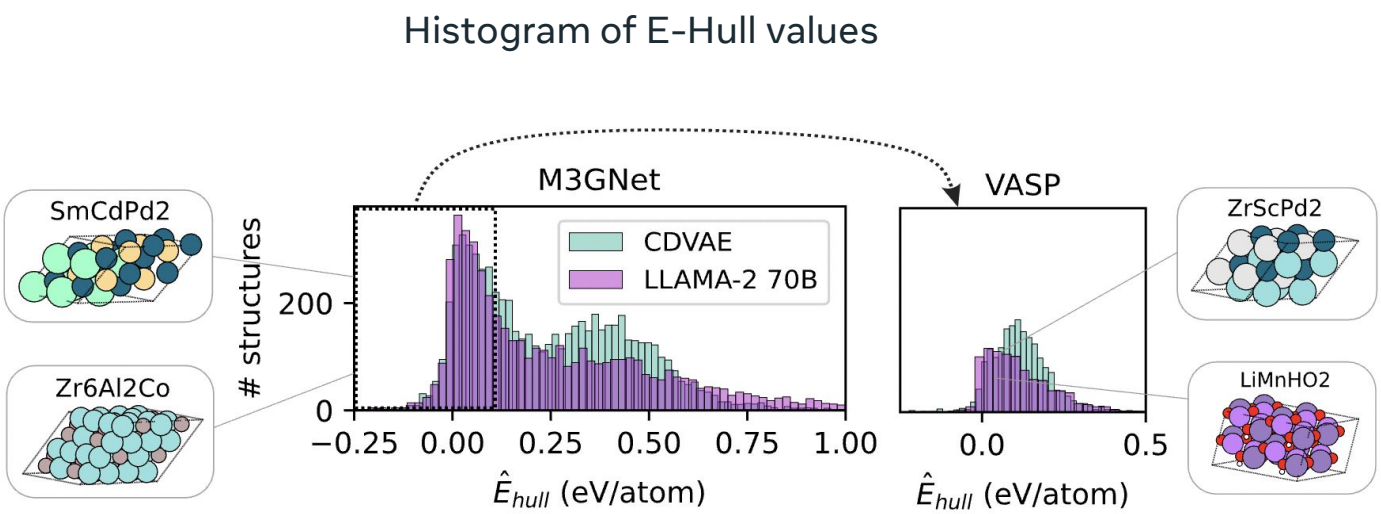
formula is PrAlO3

space group is 221

E above hull is 0.011

Generate ...

CrystalLLM: De Novo Generation



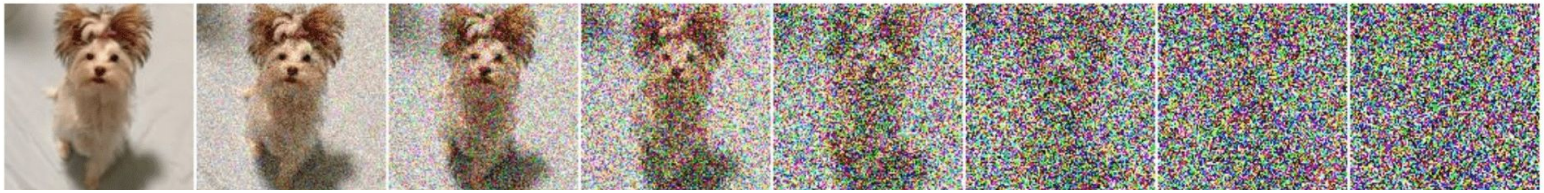
FlowMM: Riemannian Flow Matching for Materials

- **Conditional Flow Matching:**
  - Maps samples from any base distribution to any target distribution
  - Generalizes diffusion models
  - Much faster to sample than diffusion models
- **Reimannain Flow Matching** (RFM; Chen et al., 2023) generalizes flow matching to general geometries.
- **FlowMM** adapts RFM to material generation.

"*FlowMM: Generating Materials with Riemannian Flow Matching*", Miller et al. ICML 2024

Flow Matching

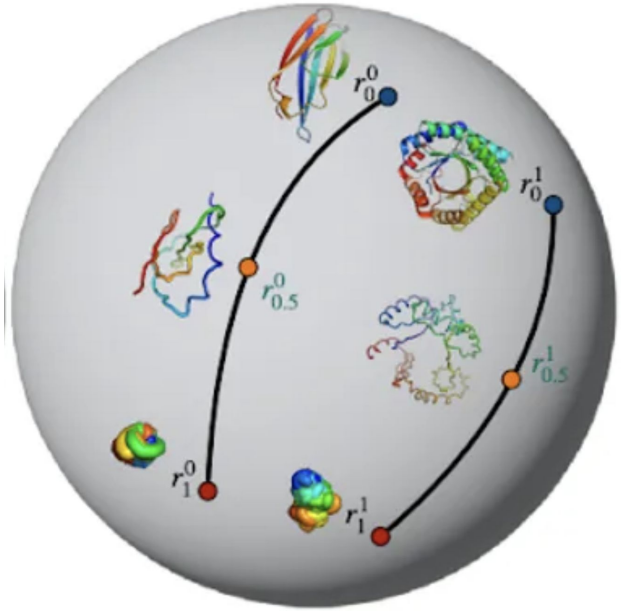
- Define a time-dependent vector field called the **velocity vector field**
  - Guides samples from base to target distribution
  - Modeled using a Neural Network
  - Choose velocity to move from noise to target in a straight line ⇒ much faster sampling than diffusion
- Generating samples: Start from a base distribution sample and integrate the velocity.



Generation

Reimannian Flow Matching

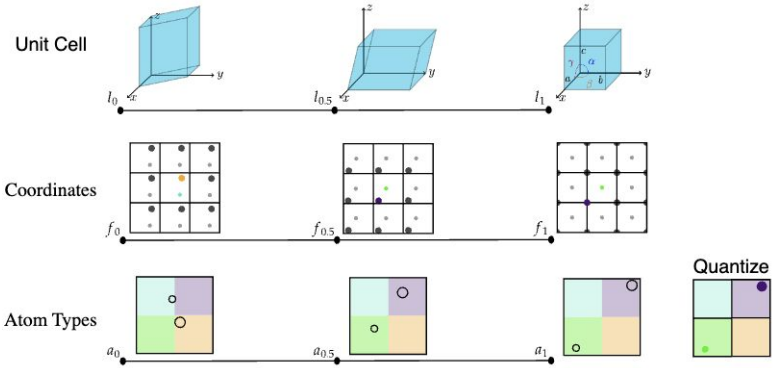
- Extends CFM to generate objects on any Reimannian manifold
  - Define base and target distributions on the manifold
  - Integrate the flow on the manifold along the shortest path on the manifold (“geodesic”)



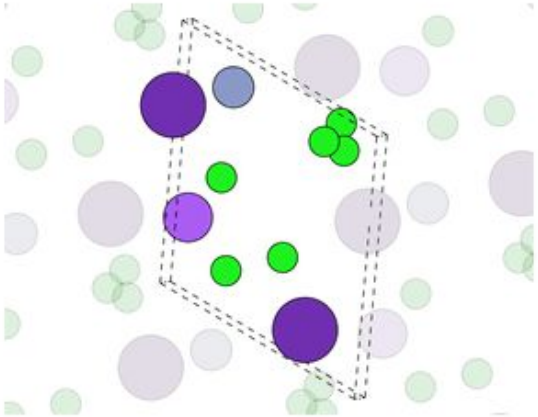
Source: Bose et al, ICLR 2024

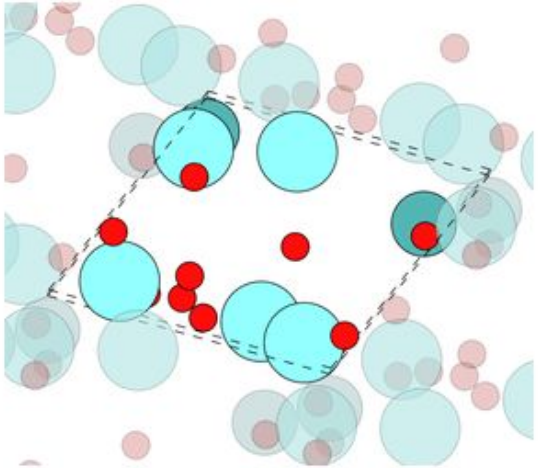
FlowMM: Reimannian Flow Matching for Materials

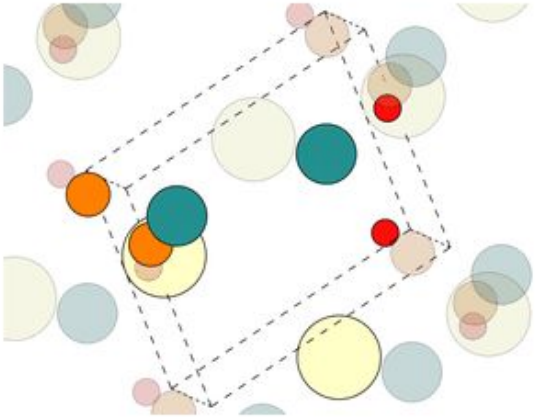
- RFM adapted to material generation: Jointly model all material attributes as a product manifold
  - Unit Cell (angles & lengths): Euclidean (*with suitable transformation of angles)
  - Coordinates: Flat 3D torus (for periodic boundary conditions)
  - Atom types: Binary coding
- Allows handling all attributes uniformly
- Use an equivariant GNN to model the velocity



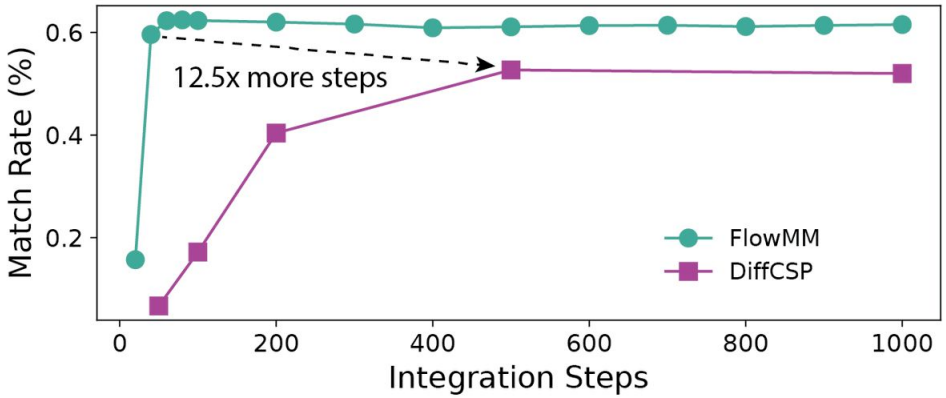
FlowMM: Reimannian Flow Matching for Materials



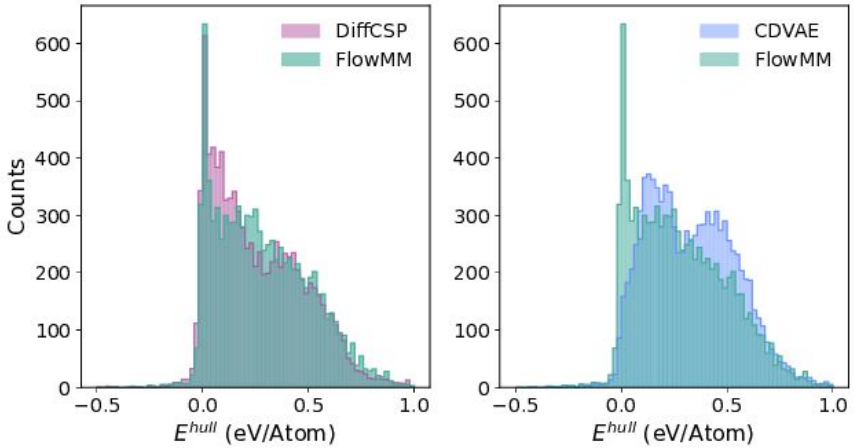




FlowMM: Reimannian Flow Matching for Materials



Crystal Structure Prediction
Match Rate vs Generation time for FlowMM vs SOTA Diffusion model



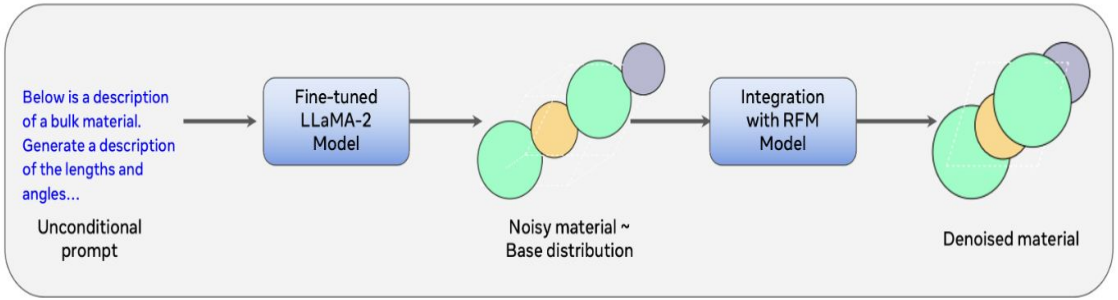
De Novo Generation
Histogram of E-Hull values for FlowMM vs SOTA Diffusion models

Combining LLMs & FlowMM ⇒ FlowLLM

- LLMs are good at generating discrete information (atom types)
- Flow matching & diffusion are good at generating continuous values (atom positions, lattice coordinates)

Can we get the best of both worlds?

FlowLLM

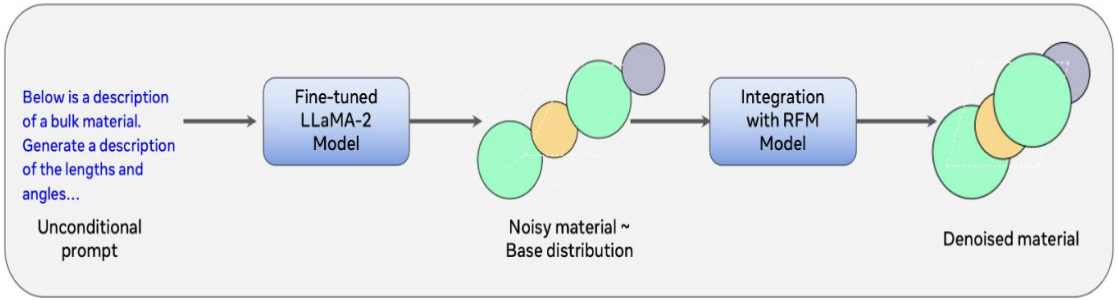


"*FlowLLM: Flow Matching for Material Generation with Learned Base Distributions*", **Sriram** & Miller. Under review at NeurIPS 2024

Combining LLMs & FlowMM ⇒ FlowLLM

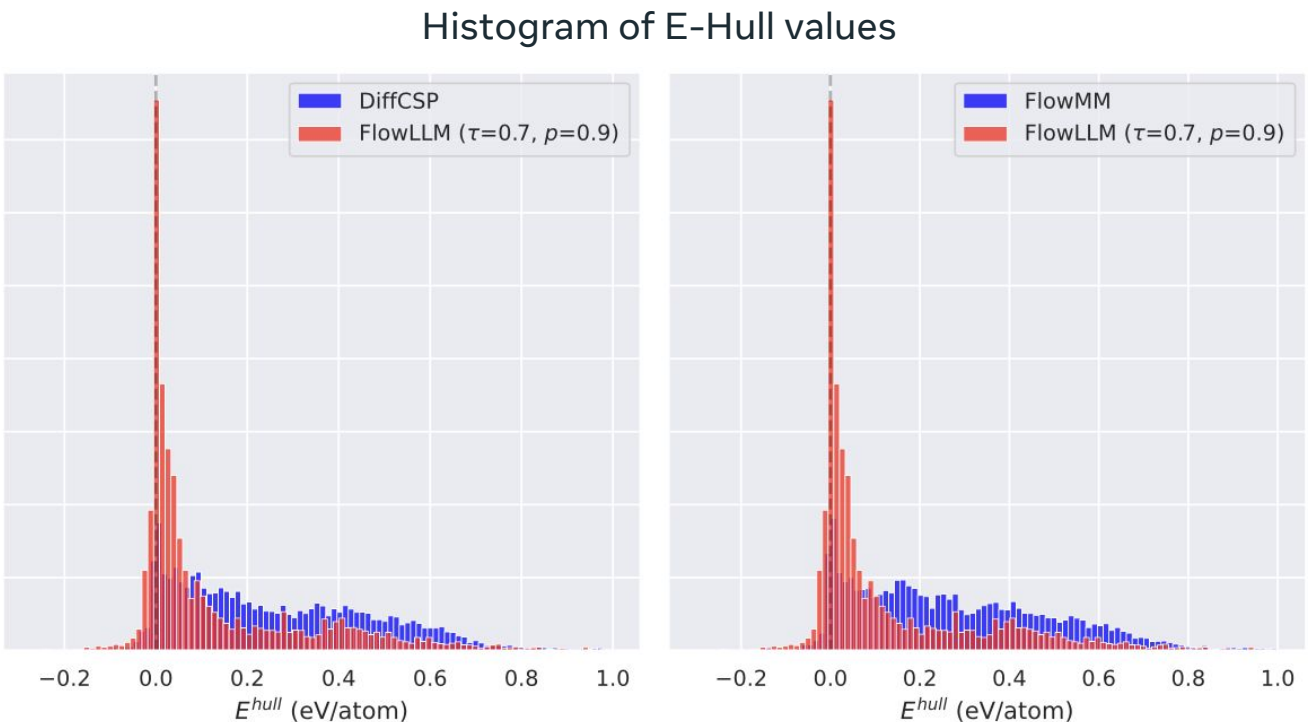
- FlowLLM significantly outperforms both LLMs & RFM. Why?
- Learning a good base distribution
  - Flow matching learns to map samples from base distribution to target distribution. A good base distribution makes this much easier.
  - In FlowLLM, the LLM learns a good base distribution close to the target distribution.
- Allows natural language prompting.

FlowLLM



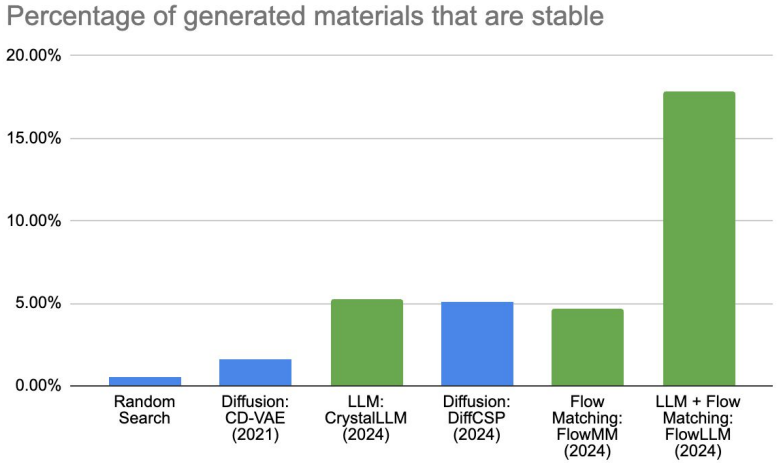
"*FlowLLM: Flow Matching for Material Generation with Learned Base Distributions*", **Sriram** & Miller. Under review at NeurIPS 2024

FlowLLM: De Novo Generation



Conclusion

- Material generation models are progressing quickly: GAN, LLM, Flow Matching, …
- What's next?
  - Extensions to new domains: molecular crystals, MOFs, …
  - Inverse design: guide material generation to optimize for properties
  - Synthesizability: Generate material and synthesis procedure jointly?
  - Integration with automated labs?



Thank you!