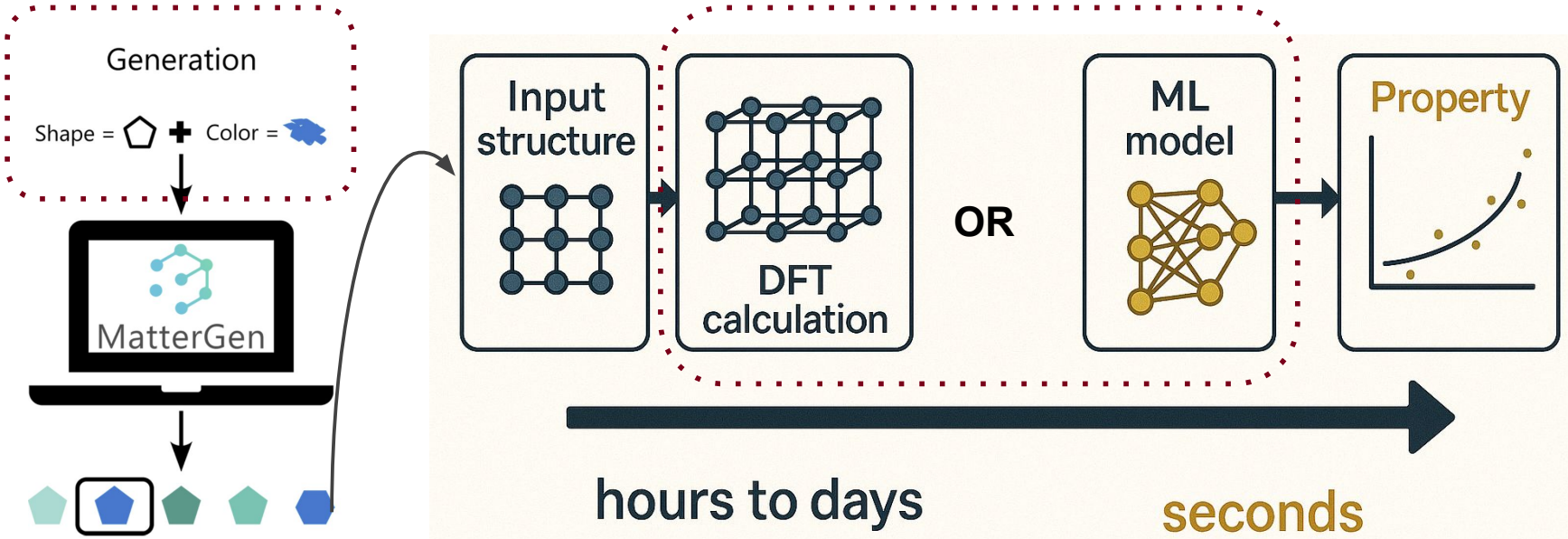
Accelerating Materials Discovery via Physics-Informed & Data-Driven Constraints

Ju Sun
McKnight Land-Grant Professor
Computer Science & Engineering
May 15, 2026

University of Minnesota
Driven to Discover℠

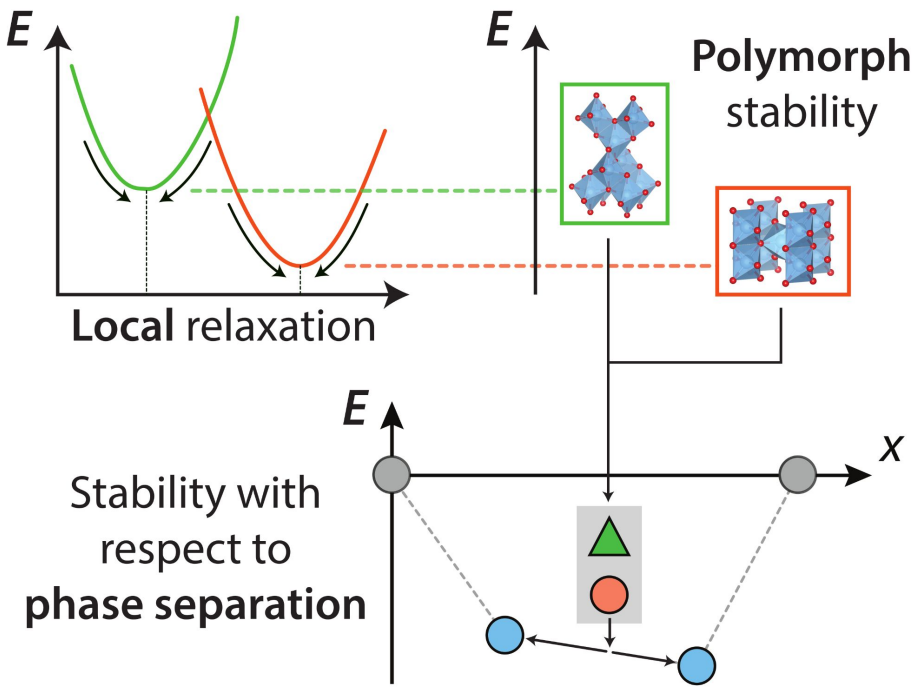
AI-accelerated materials discovery

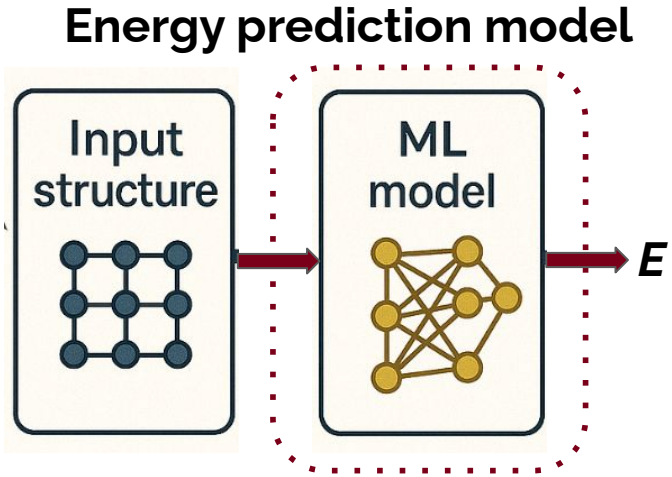


Material generation

Property prediction

Property prediction: thermodynamic stability



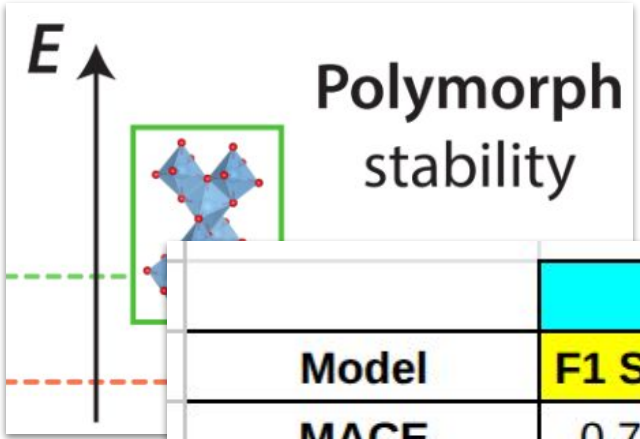


Training objective

$$\min_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \ell(\boldsymbol{y}_i, f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)) + \Omega(\boldsymbol{\theta})$$

… with physics-informed constraints

$$\min_{\boldsymbol{\theta}} \frac{1}{N}\sum_{i=1}^{N} \ell(\boldsymbol{y}_i, f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)) + \Omega(\boldsymbol{\theta})$$

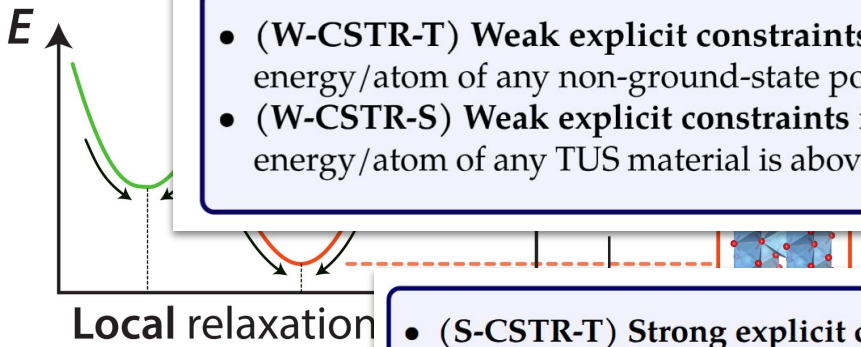


$$\min_{\boldsymbol{\theta}} \frac{1}{N}\sum_{i=1}^{N} \ell(\boldsymbol{y}_i, f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)) + \Omega(\boldsymbol{\theta})$$

$$\text{s.t.} \quad f_{\boldsymbol{\theta}}(\boldsymbol{x}_p) \geq f_{\boldsymbol{\theta}}(\boldsymbol{x}_q)\ \forall\ (p,q) \in \mathcal{O}_{\text{W-CSTR-T}},$$

	Test			
Model	F1 Score	L1 Score	Ground State Score	Kendall Tau
MACE	0.7873	0.0317	0.006	0.5966
MACE + CDL	0.82	0.0327	0.005	0.6474
Improvement	0.0327	0.001	-0.001	0.0508

Next: more powerful constraints



- **(W-CSTR-T) Weak explicit constraints informed by phase transition**: For each composition, the energy/atom of any non-ground-state polymorph is greater than that of the ground-state polymorph
- **(W-CSTR-S) Weak explicit constraints informed by phase separation**: For each chemical space, the energy/atom of any TUS material is above the lower convex envelope in the composition-energy space

- **(S-CSTR-T) Strong explicit constraints informed by phase transition**: For each composition, t ordering of the energy/atom of all polymorphs is preserved
- **(S-CSTR-S) Strong explicit constraints informed by phase separation**: For each chemical space, t ordering of energy/atom of all materials is preserved, and the energy/atom of any TUS material above the lower convex envelope in the composition-energy space

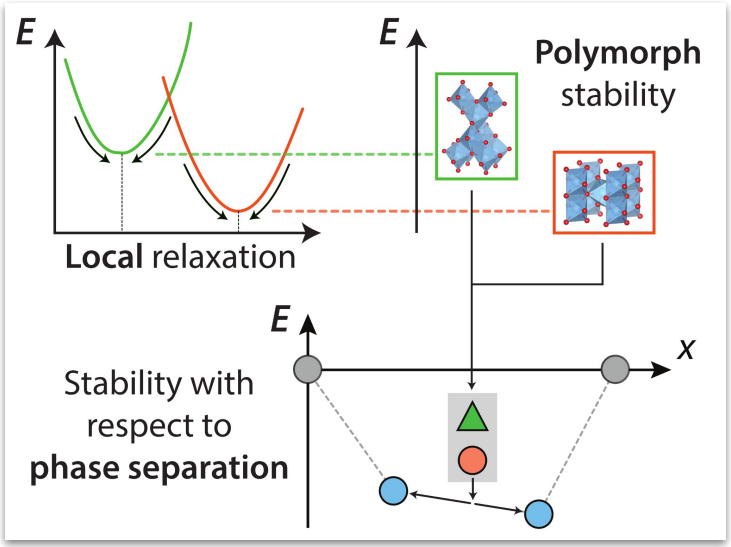
E

Stability with respect to **phase separation**

NSF

ACED: Accelerating Computing-Ena Scientific Discovery (ACED)

Deep learning with nontrivial "**hard**" constraints



$$\min_{\boldsymbol{\theta}} \frac{1}{N}\sum_{i=1}^{N} \ell(\boldsymbol{y}_i, f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)) + \Omega(\boldsymbol{\theta})$$

$$\text{s.t.} \quad f_{\boldsymbol{\theta}}(\boldsymbol{x}_p) \geq f_{\boldsymbol{\theta}}(\boldsymbol{x}_q) \;\forall\; (p,q) \in \mathcal{O}_{\text{W-CSTR-T}},$$

Robustness evaluation of DL models

$$\max_{\boldsymbol{x}'} \ell(\boldsymbol{y}, f_{\boldsymbol{\theta}}(\boldsymbol{x}'))$$

$$\text{s.t. } \boldsymbol{x}' \in \Delta(\boldsymbol{x}) = \{\boldsymbol{x}' \in [0,1]^n : d(\boldsymbol{x}, \boldsymbol{x}') \leq \varepsilon\}$$

$$\min_{\boldsymbol{x}' \in [0,1]^n} d(\boldsymbol{x}, \boldsymbol{x}') \quad \text{s.t. } \max_{\ell \neq y} f_{\boldsymbol{\theta}}^{\ell}(\boldsymbol{x}') \geq f_{\boldsymbol{\theta}}^{y}(\boldsymbol{x}')$$

Imbalance Learning

$$\max_{\boldsymbol{\theta}, t} \frac{\sum_{i=1}^{N} \mathbb{1}\{y_i = +1\}\, \mathbb{1}\{f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) > t\}}{\sum_{i=1}^{N} \mathbb{1}\{f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) > t\}}$$

$$\text{s.t. } \frac{\sum_{i=1}^{N} \mathbb{1}\{y_i = +1\}\, \mathbb{1}\{f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) > t\}}{\sum_{i=1}^{N} \mathbb{1}\{y_i = +1\}} \geq \alpha$$

Heuristic handling of hard constraints is problematic

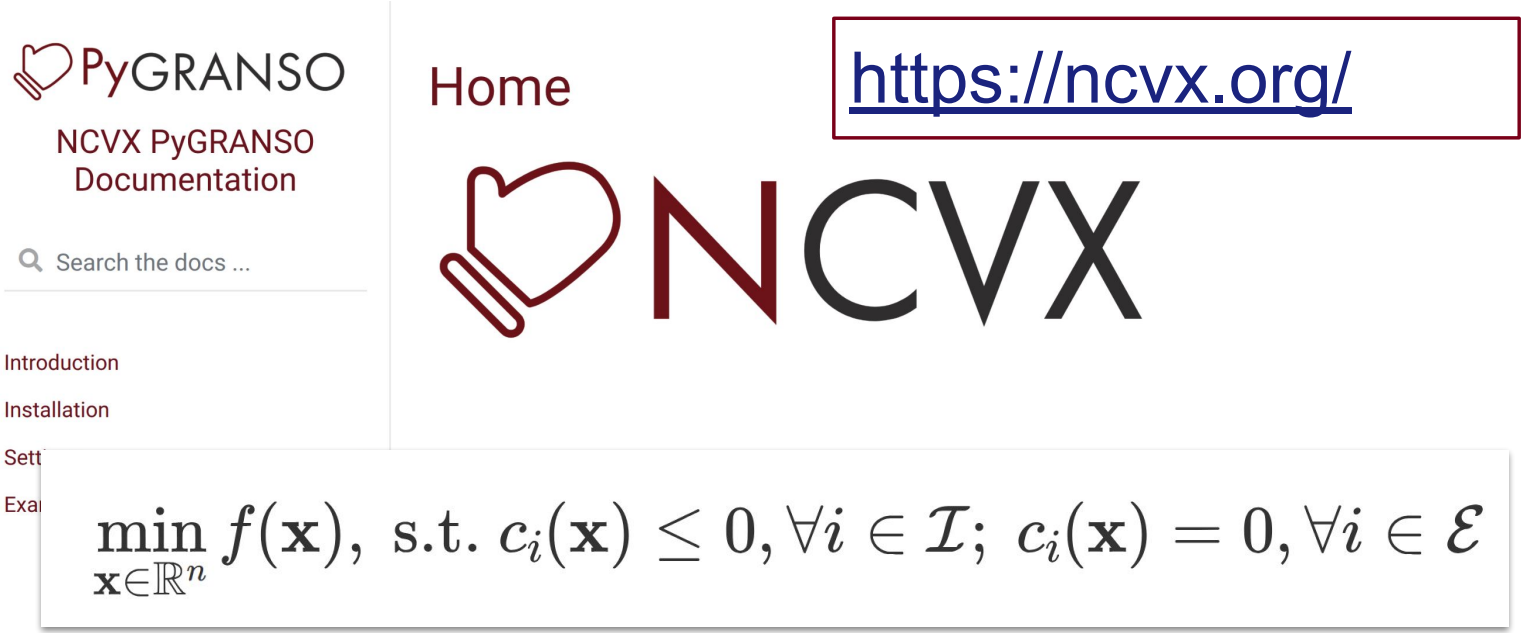
	cross-entropy loss		margin loss	
Method	**Viol. (%) ↓**	**Att. Succ. (%) ↑**	**Viol. (%) ↓**	**Att. Succ. (%) ↑**
Fast-LPA	73.8	3.54	41.6	56.8
LPA	**0.00**	80.5	**0.00**	97.0
PPGD	5.44	25.5	**0.00**	38.5
PWCF (ours)	0.62	93.6	**0.00**	100

LPA, Fast-LPA: penalty methods **PPGD**: Projected gradient descent

Penalty methods tend to encounter
large constraint violation (i.e., infeasible so
theory) or **suboptimal solution**

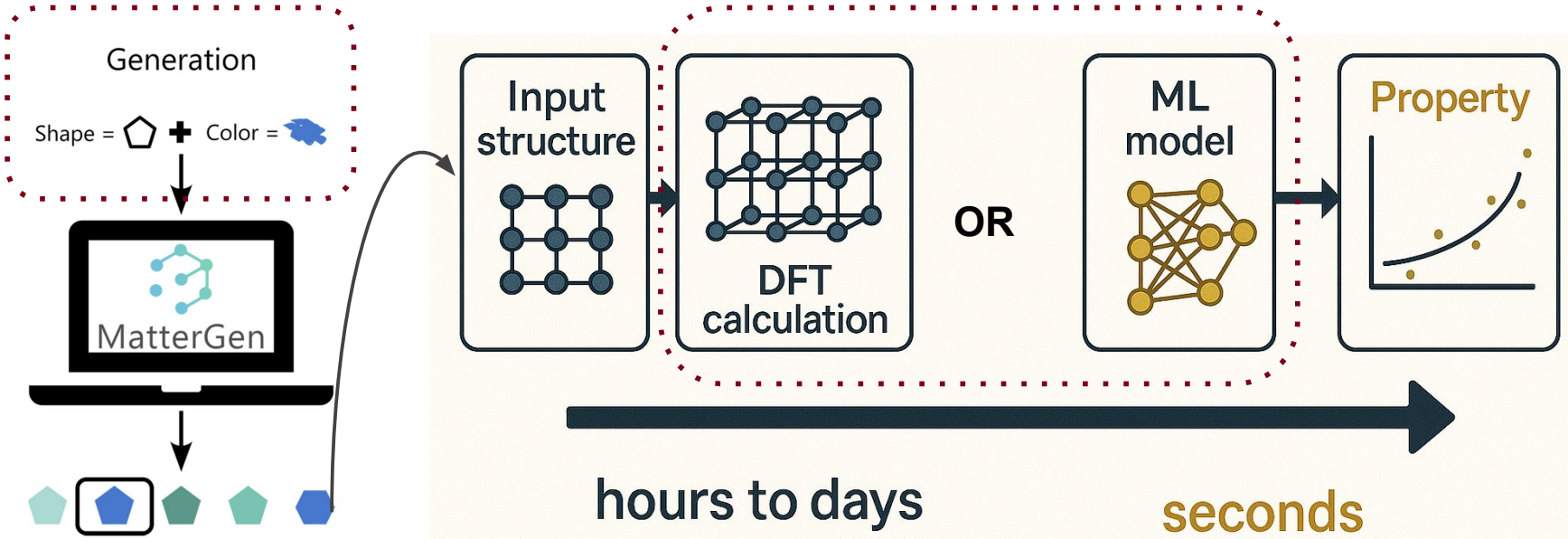
Unreliable computation leads to unreliable reliability evaluation

NCVX: A principled solver for constrained DL



- **First general-purpose solver for hard-constrained deep learning problems**
- **Recently updated to be compatible with PyTorch 2.8**

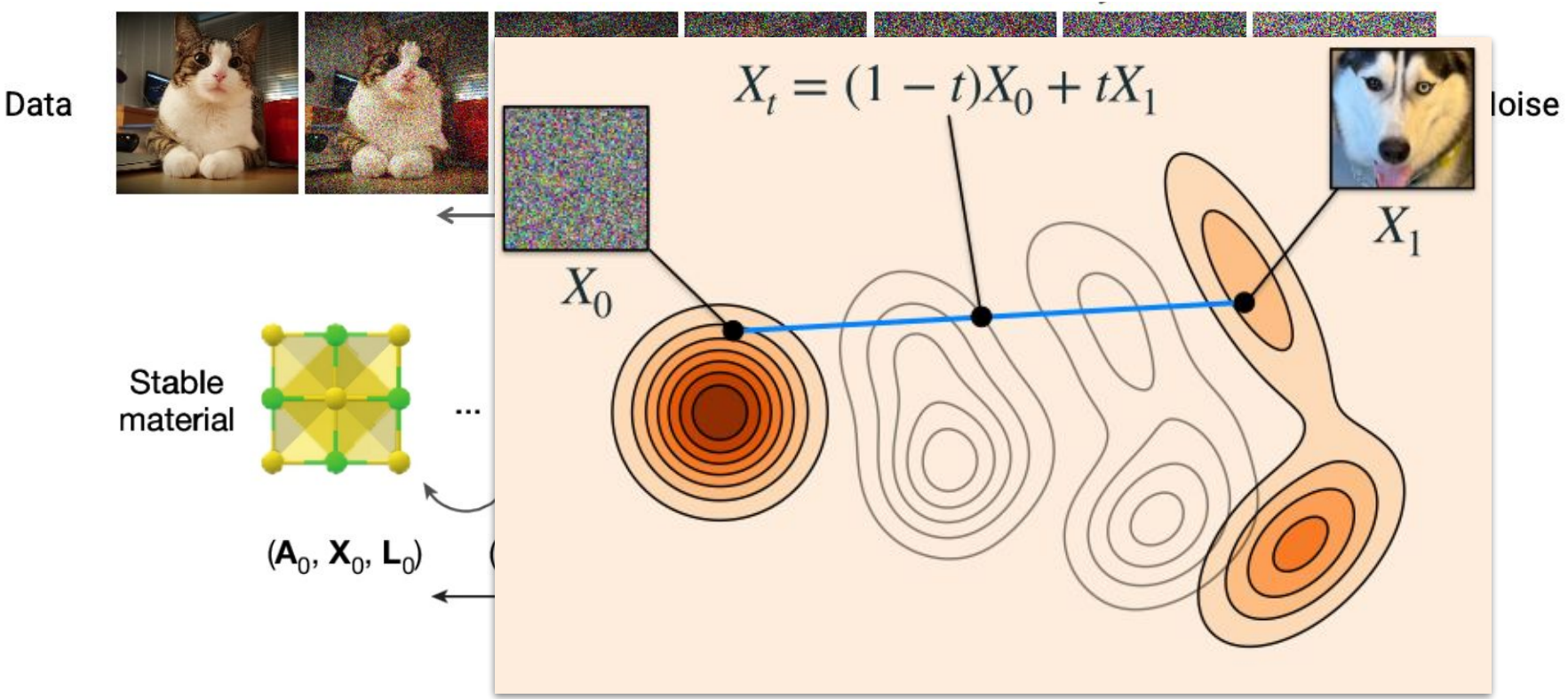
AI-accelerated materials discovery



Material generation

Property prediction

Generative models



Data bottlenecks for numerous domains

IMAGES

LAION-

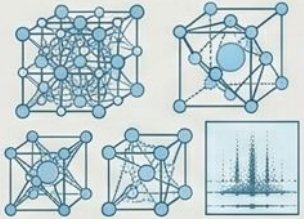
Size: 5.8

Key Deta
for SOTA
models;
English

VIDEOS

3D OBJECTS

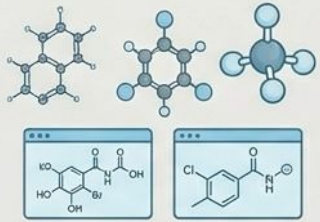
CRYSTAL STRUCTURES



MATERIALS PROJECT / OQMD

Size: **>150k unique compounds**
Key Detail: Density functional theory (DFT) calculated properties; foundation for new crystal discovery.

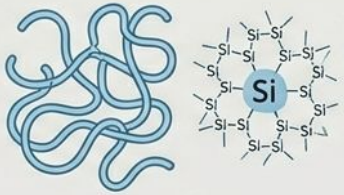
ORGANIC MOLECULES



ZINC20 / PubChem

Size: **>1.5 Billion distinct conformers**
Key Detail: Synthesizable chemical space; pre-training for generative molecular graph and SMILES models.

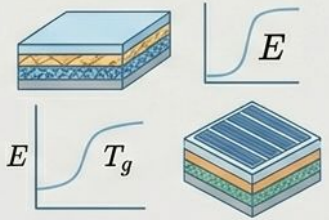
POLYMERS & AMORPHOUS



KHASHKHOV / Specialized Sets

Size: **>10k annotated entries**
Key Detail: Processed from literature & experiments; focuses on complex chain architectures and

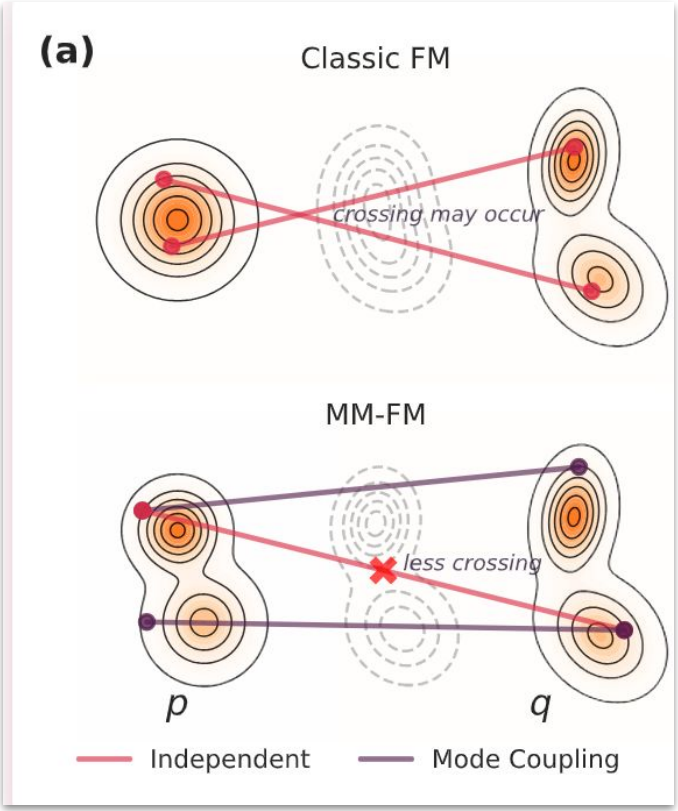
POLYMERS & AMORPHOUS

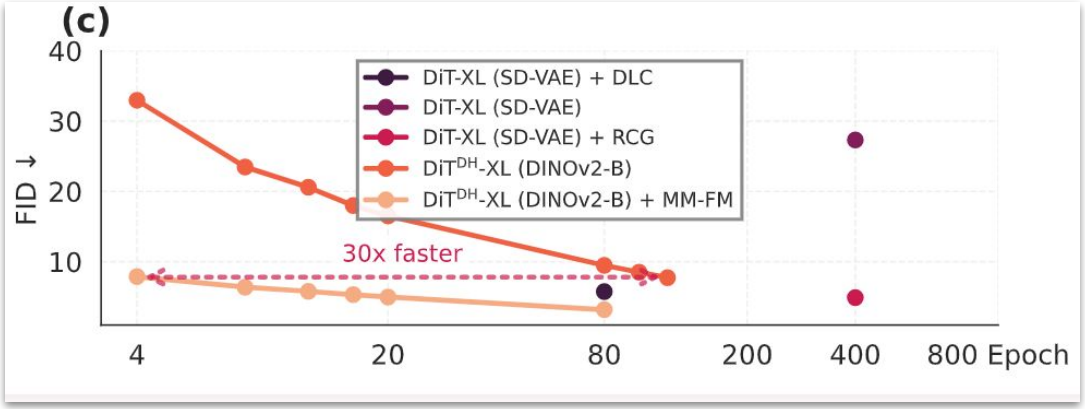


KHASHKHOV / Specialized Sets

Size: **>10k annotated entries**
Key Detail: Processed from literature & experiments; focuses on chain architectures and composite design.

Maximize data-efficient for generative models





Better
data/training/inference efficiencies

For more details

Flow Matching for Multimodal Distributions

Gaoxiang Luo[1,*] , Frank Cole[2,*] , Sihang Zhang[3] , Yuxiang Wan[1] , Yulong Lu[2] , Ju Sun[1]

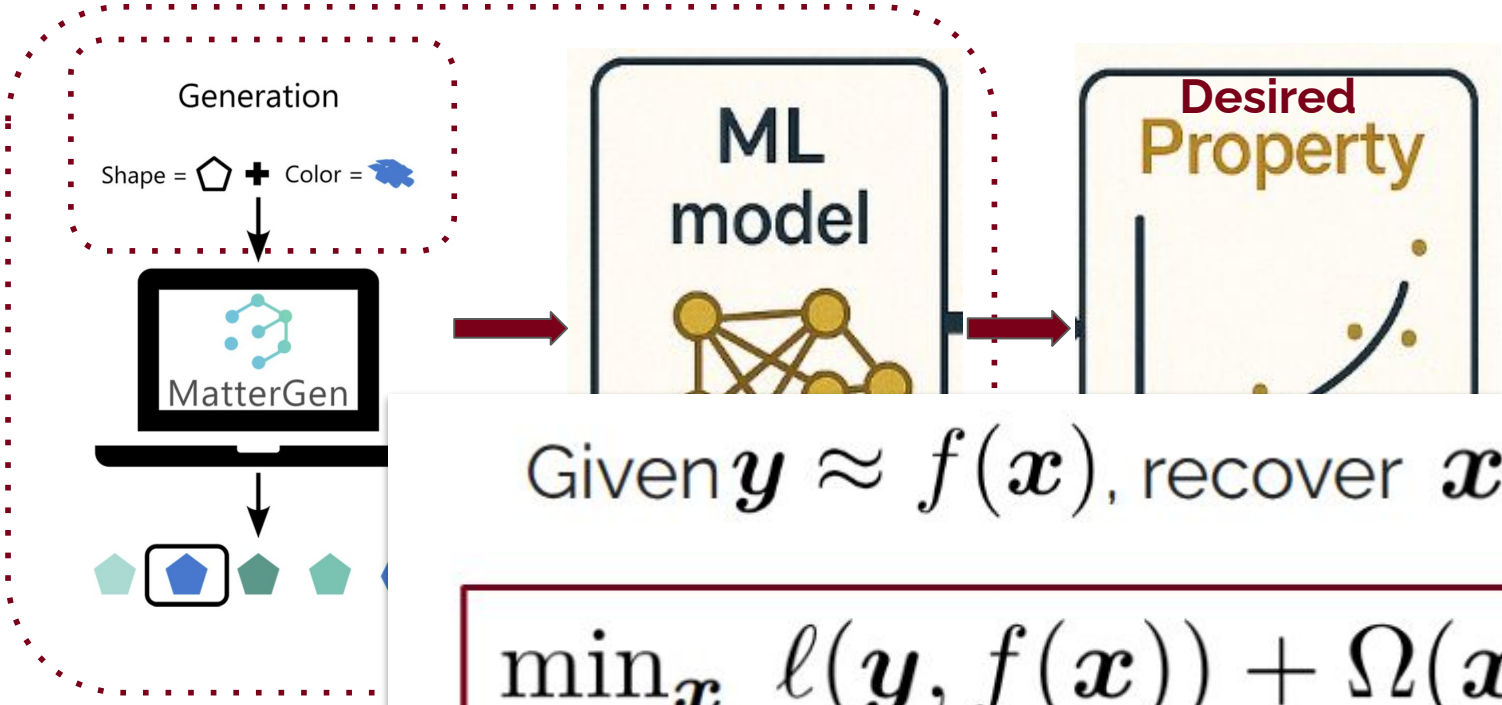
[*]Equal Contribution [1]Department of Computer Science and Engineering
[2]School of Mathematics [3]School of Statistics
University of Minnesota Twin Cities

CVPR 2026

Paper (Coming soon) Code (Coming soon) Models (Coming soon)

https://mm-flow.github.io/

Materials discovery as inverse design



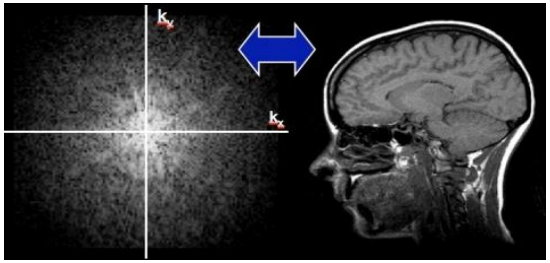
Given $\boldsymbol{y} \approx f(\boldsymbol{x})$, recover $\boldsymbol{x}$

$$\min_{\boldsymbol{x}} \ \ell(\boldsymbol{y}, f(\boldsymbol{x})) + \Omega(\boldsymbol{x})$$

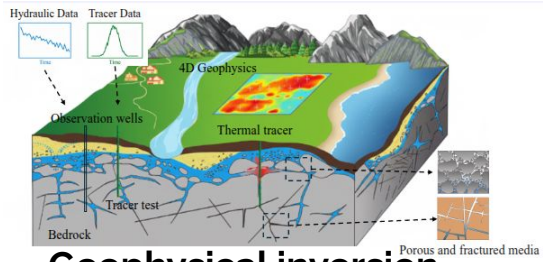
media

Inverse modeling

(inverse problems, reconstruction, data assimilation, **inverse design/control,** **conditional generation**, …)



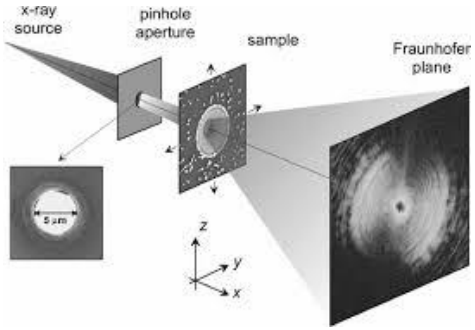
MRI reconstruction



Geophysical inversion

Given $\boldsymbol{y} \approx f(\boldsymbol{x})$, recover $\boldsymbol{x}$

$$\min_{\boldsymbol{x}} \ \ell(\boldsymbol{y}, f(\boldsymbol{x})) + \Omega(\boldsymbol{x})$$



Coherent diffraction imaging

Inverse shape control

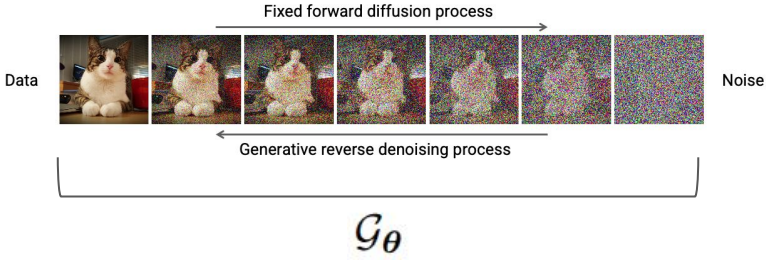
Challenges:

- Linear vs. **nonlinear** f
- Unconstrained vs. **constrained** (e.g., PDE (x, y) = 0)
- Explicit vs. **implicit** f

Inverse modeling with generative priors

object-only datasets $\{\boldsymbol{x}_i\}_{i=1,\dots,N}$

Distribution learning via deep generative models



Plug in pretrained deep generative priors

$$\min_{\boldsymbol{z}} \ \mathcal{L}(\boldsymbol{z}) \doteq \ell(\boldsymbol{y}, \mathcal{A} \circ \mathcal{G}_{\boldsymbol{\theta}}(\boldsymbol{z})) + \Omega \circ \mathcal{G}_{\boldsymbol{\theta}}(\boldsymbol{z})$$

DMPlug: A Plug-in Method for Solving Inverse Problems with Diffusion Models

Hengkang Wang, Xu Zhang, Taihui Li, Yuxiang Wan, Tiancong Chen, Ju Sun

NeurIPS'24; with **domain-specific** priors

Saving Foundation Flow-Matching Priors for Inverse Problems

Yuxiang Wan, Ryan Devera, Wenjie Zhang, Ju Sun

ICML'26; with **foundation** priors (e.g., stable diffusion models)