Improving machine learning with polymer physics

Debra J. Audus
2024 Artificial Intelligence for Materials Science (AIMS) Workshop
July 18, 2024

National Institute of
Standards and Technology
U.S. Department of Commerce

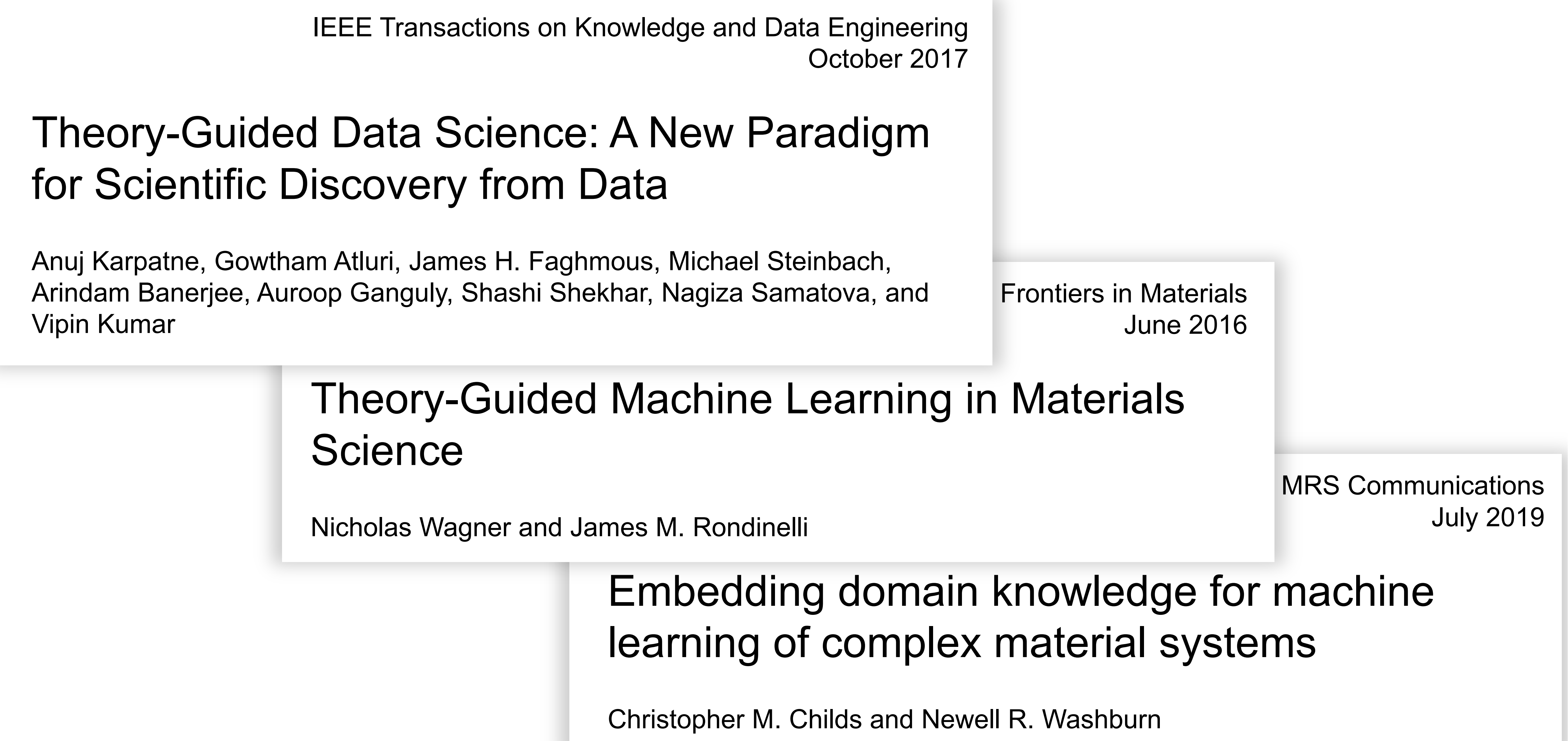
Outstanding challenges

DATA
(fuel for ML/AI)

EXTRAPOLATION
(go beyond the dataset)

EXPLAINABILITY
(answer scientific questions)

A path forward: use knowledge



IEEE Transactions on Knowledge and Data Engineering
October 2017

Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data

Anuj Karpatne, Gowtham Atluri, James H. Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar

Frontiers in Materials
June 2016

Theory-Guided Machine Learning in Materials Science

Nicholas Wagner and James M. Rondinelli

MRS Communications
July 2019

Embedding domain knowledge for machine learning of complex material systems

Christopher M. Childs and Newell R. Washburn

Impact of theory

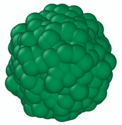
Need less
DATA
(fuel for ML/AI)

Improves
EXTRAPOLATION
(go beyond the dataset)

May provide
EXPLAINABILITY
(answer scientific questions)

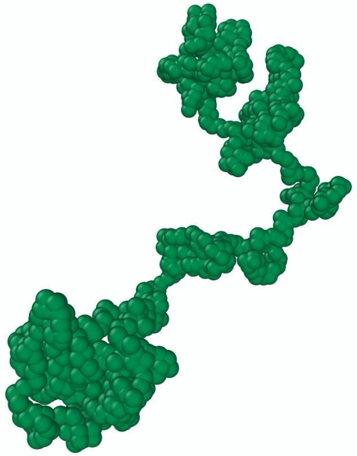
Case 1: a single polymer in solvent

Bad
Solvent



$R_g \sim N^{1/3}$

Theta
solvent



$R_g \sim N^{1/2}$

Good
solvent

$R_g \sim N^{0.588}$

$R_g \sim N^{\nu}$

ML Input:

- $\ln N$
- α (solvent quality)

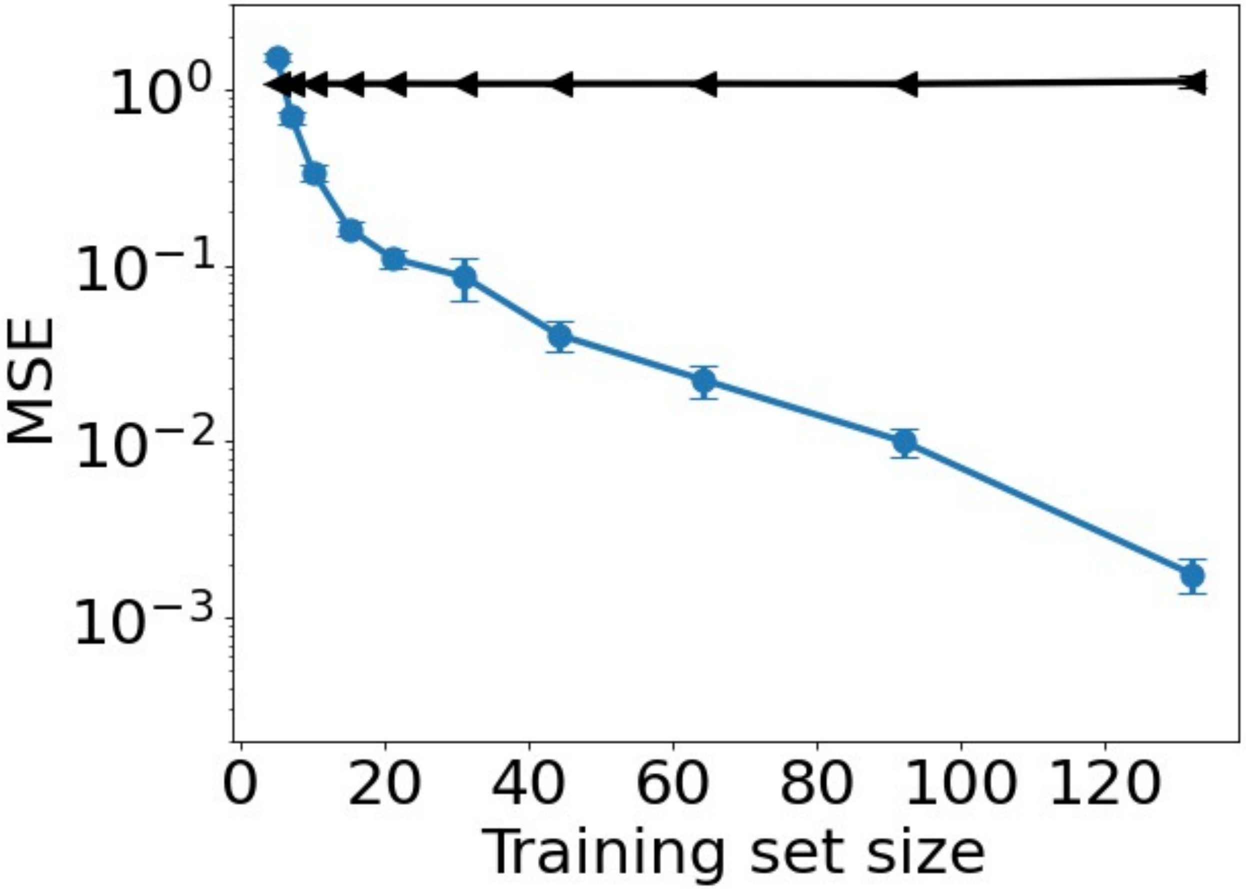
ML Output:

- $\ln R_g^2$

Audus *et. al.*, *ACS Macro Letters* **2023**, 11, 1117-1122; https://github.com/usnistgov/taml

Austin McDannald, Brian DeCost (NIST)

Benchmarks



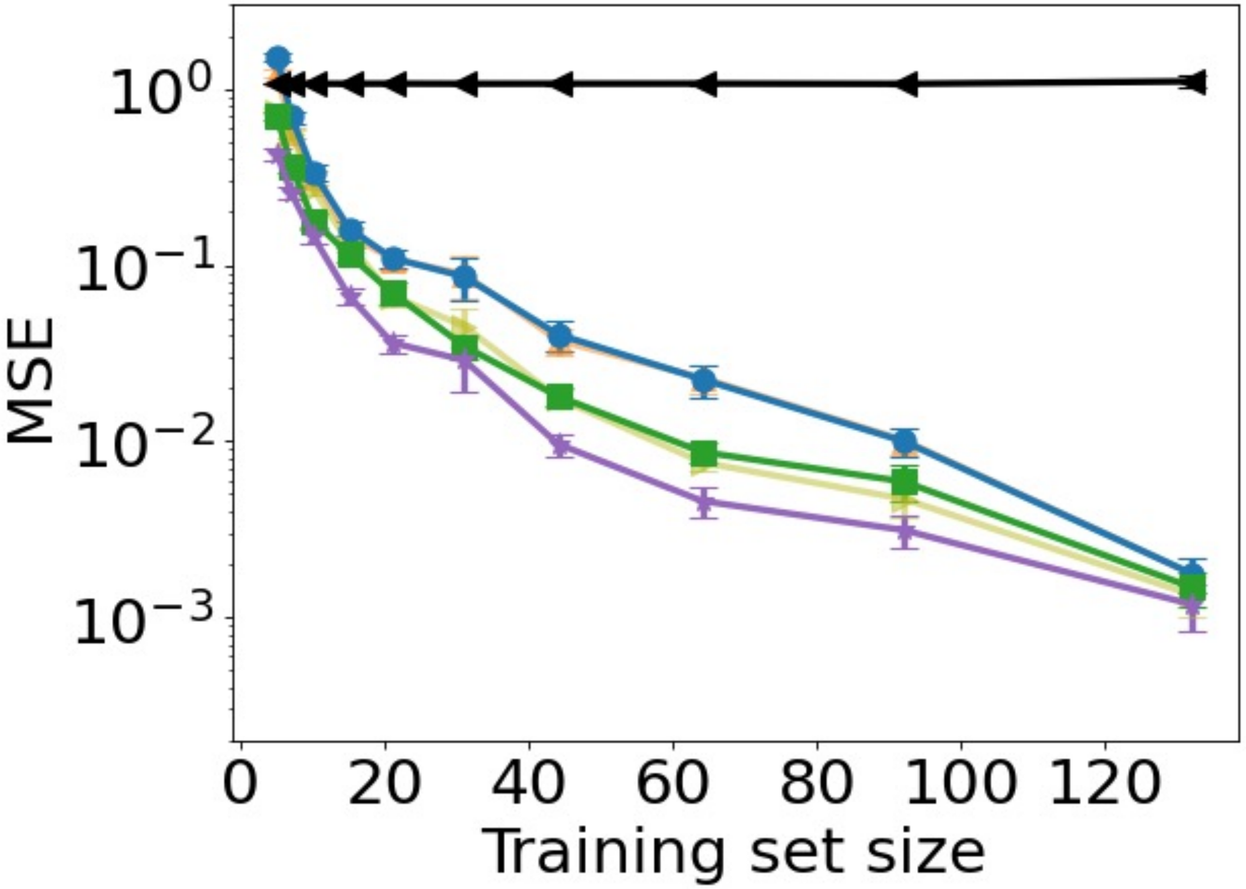
Direct $x \equiv [\ln N, \alpha] \rightarrow \boxed{\text{GPR}} \rightarrow \ln R_g^2$

Theory $t(x) = \ln N + \ln 1/6 = \ln R_g^2$

Purposely use an imperfect theory ($R_g \sim N^{1/2}$) because we often have theories that are only sometimes valid

Audus *et. al.*, *ACS Macro Letters* **2023**, 11, 1117-1122; https://github.com/usnistgov/taml

Parameterization: best!



Direct $x \equiv [\ln N, \alpha] \rightarrow \boxed{\text{GPR}} \rightarrow \ln R_g^2$

Theory $t(x) = \ln N + \ln 1/6 = \ln R_g^2$

Latent variable $x,\ t(x) \rightarrow \boxed{\text{GPR}} \rightarrow \ln R_g^2$

Difference $x \rightarrow \boxed{\text{GPR}} \rightarrow \ln R_g^2 - t(x) \rightarrow$; $t(x) \rightarrow$ $\boxed{\text{sum}} \rightarrow \ln R_g^2$

Linear prior $\ln R_g^2 = \mathcal{GP}(\kappa + 2\nu \ln N + \lambda\alpha, k(x, x'))$

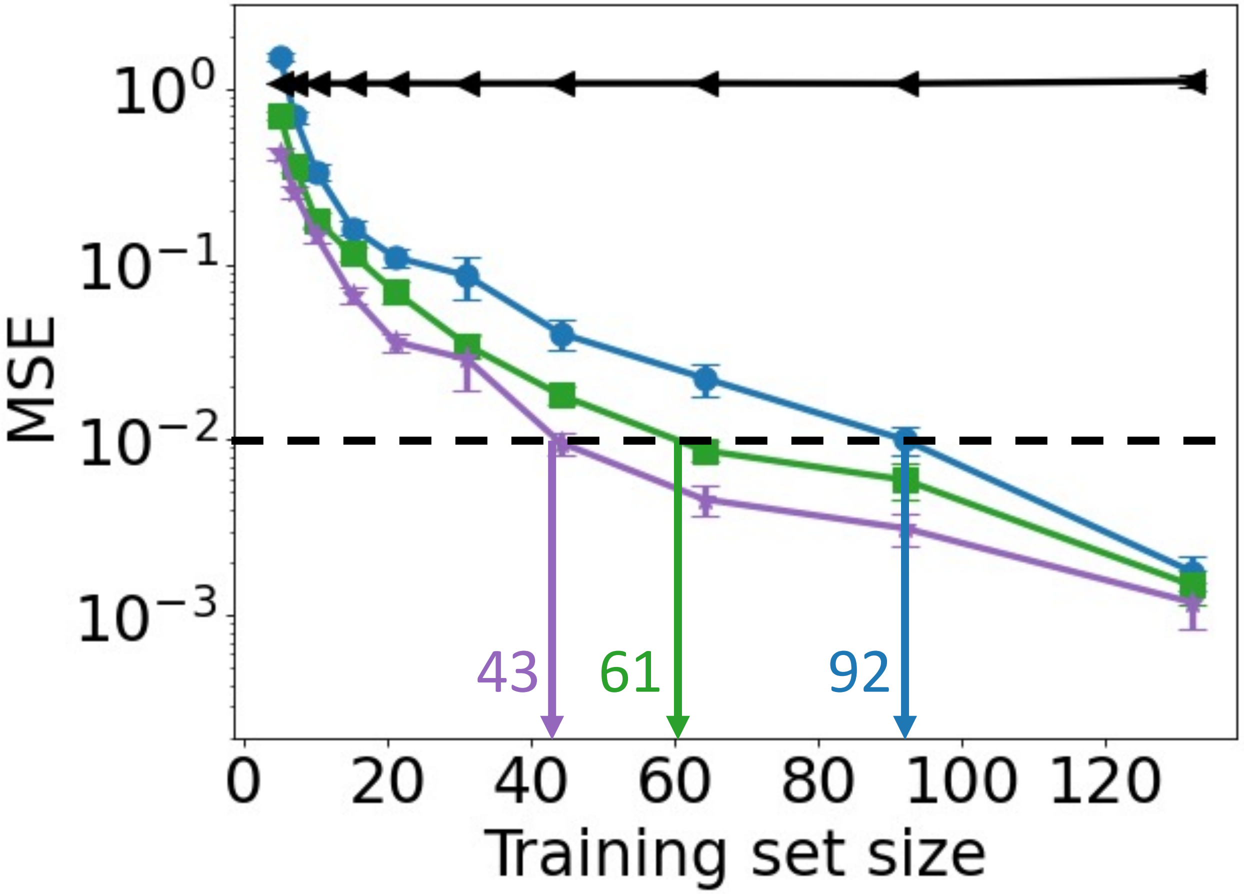
Parameterization $\ln R_g^2 = 2\nu\,[\alpha] \ln N + \kappa[\ln N, \alpha]$

Respects a scaling exponent dependent on solvent quality

Best performance when full functional form of theory is used

Audus *et. al.*, *ACS Macro Letters* **2023**, 11, 1117-1122; https://github.com/usnistgov/taml

Need substantially less data with theory



Direct $x \equiv [\ln N, \alpha] \rightarrow \boxed{\text{GPR}} \rightarrow \ln R_g^2$

Theory $t(x) = \ln N + \ln 1/6 = \ln R_g^2$

Latent variable $x \rightarrow$, $t(x) \rightarrow \boxed{\text{GPR}} \rightarrow \ln R_g^2$

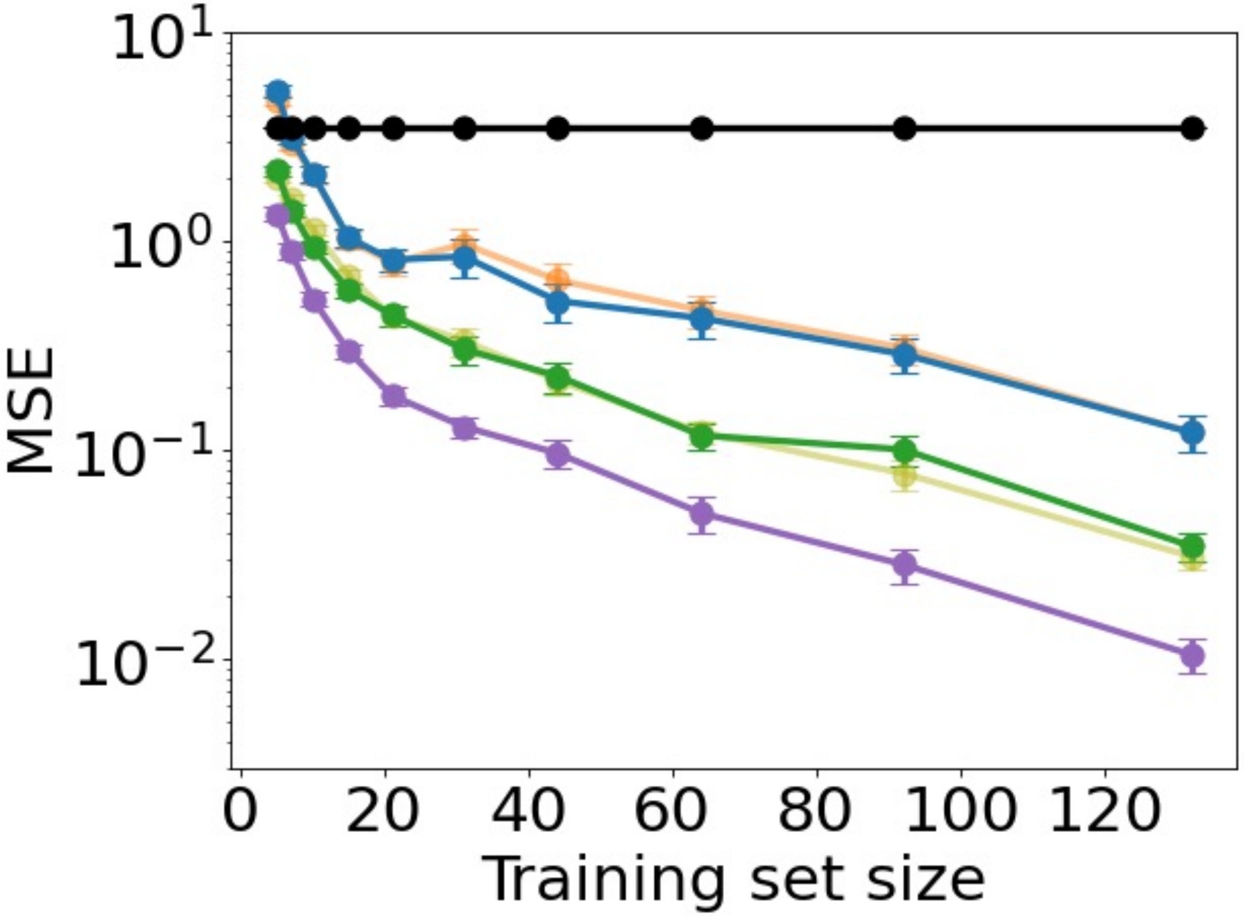
Difference $x \rightarrow \boxed{\text{GPR}} \rightarrow \ln R_g^2 - t(x) \rightarrow$, $t(x) \rightarrow \boxed{\text{sum}} \rightarrow \ln R_g^2$

Linear prior $\ln R_g^2 = \mathcal{GP}(\kappa + 2\nu \ln N + \lambda\alpha, k(x, x'))$

Parameterization $\ln R_g^2 = 2\nu\,[\alpha] \ln N + \kappa[\ln N, \alpha]$

Audus *et. al.*, *ACS Macro Letters* **2023**, 11, 1117-1122; https://github.com/usnistgov/taml

Testing extrapolation (larger N)



Direct $x \equiv [\ln N, \alpha] \rightarrow \boxed{\text{GPR}} \rightarrow \ln R_g^2$

Theory $t(x) = \ln N + \ln 1/6 = \ln R_g^2$

Latent variable

$$\begin{tikzcd} x \arrow[r] & \boxed{\text{GPR}} \arrow[r] & \ln R_g^2 \\ t(x) \arrow[ru] & & \end{tikzcd}$$

Difference

$$\begin{tikzcd} x \arrow[r] & \boxed{\text{GPR}} \arrow[r] & \ln R_g^2 - t(x) \arrow[r] & \boxed{\text{sum}} \arrow[r] & \ln R_g^2 \\ & & t(x) \arrow[ru] & & \end{tikzcd}$$

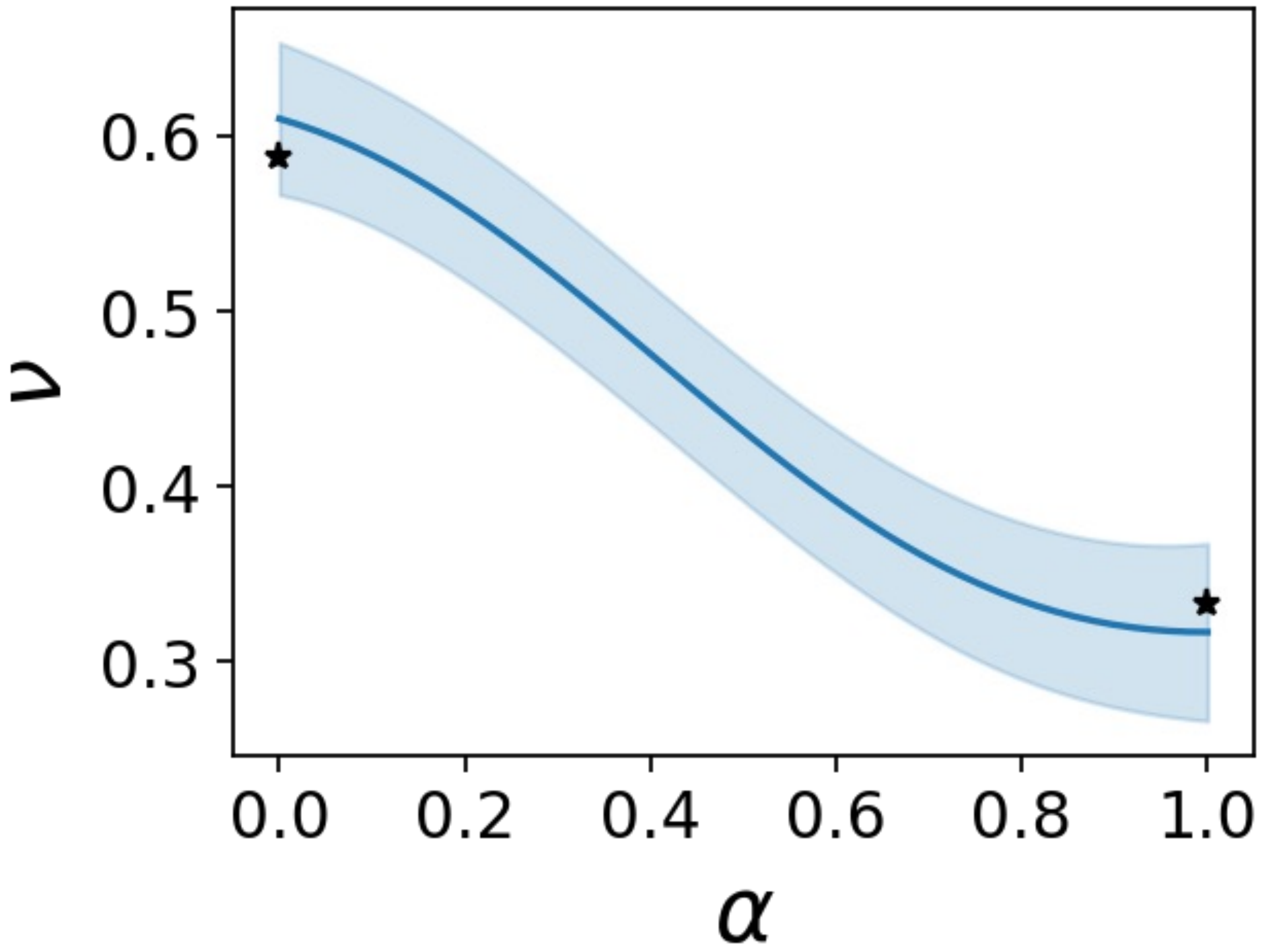
Linear prior $\ln R_g^2 = \mathcal{GP}(\kappa + 2\nu \ln N + \lambda\alpha, k(x, x'))$

Parameterization $\ln R_g^2 = 2\nu\,[\alpha] \ln N + \kappa[\ln N, \alpha]$

Incorporating theory improves extrapolation

Audus *et. al.*, *ACS Macro Letters* **2023**, 11, 1117-1122; https://github.com/usnistgov/taml

Interpretability for parameterization



$$\ln R_g^2 = 2\nu\,[\alpha]\ln N + \kappa[\ln N, \alpha]$$

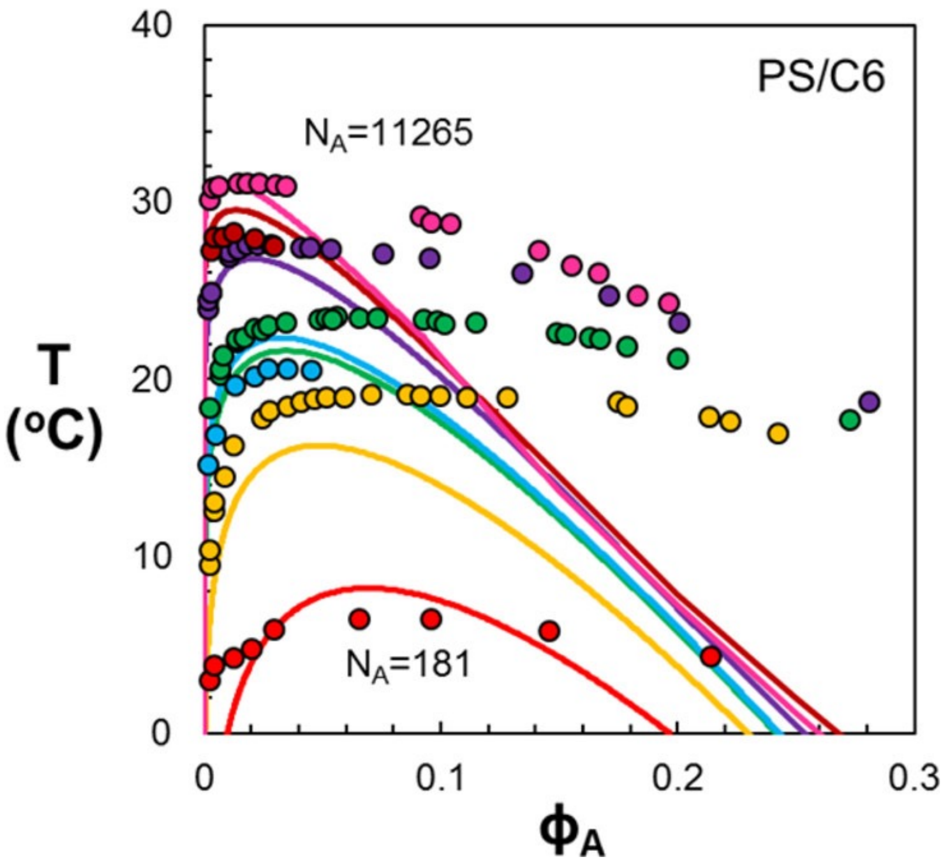
Get a prediction for the scaling exponent that includes known limits

Solid lines = prediction; Shaded regions = 95% confidence intervals

Audus *et. al.*, *ACS Macro Letters* **2023**, 11, 1117-1122;
https://github.com/usnistgov/taml

Case 2: Polymer Solution Phase Diagrams

$$\frac{\Delta G}{n_T k_B T} = \underbrace{\frac{\phi_A}{N_A}\log\phi_A + \frac{\phi_B}{N_B}\log\phi_B}_{\text{Configurational Entropy}} + \underbrace{\chi_{AB}\phi_A\phi_B}_{\text{Pairwise Interactions (Enthalpy)}}$$

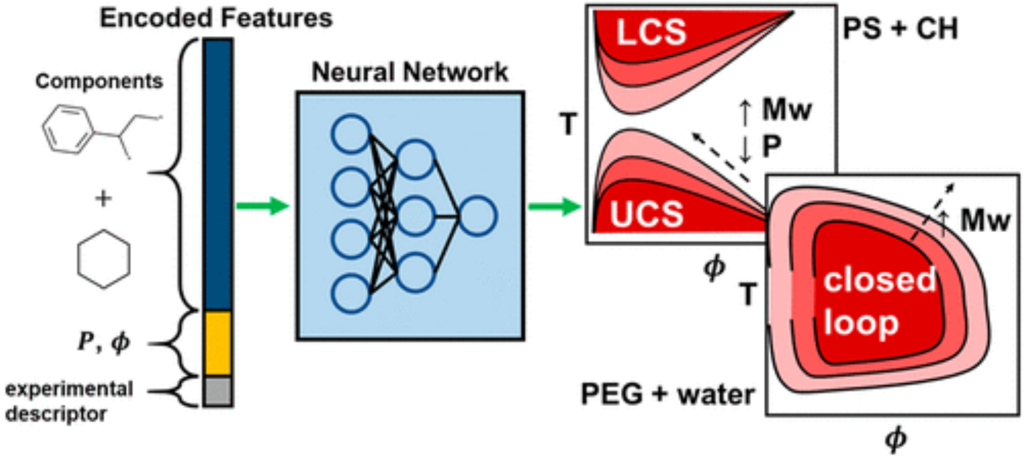


Current predictive models – poor global agreement, requires empirical parameters (e.g. χ(T, p))

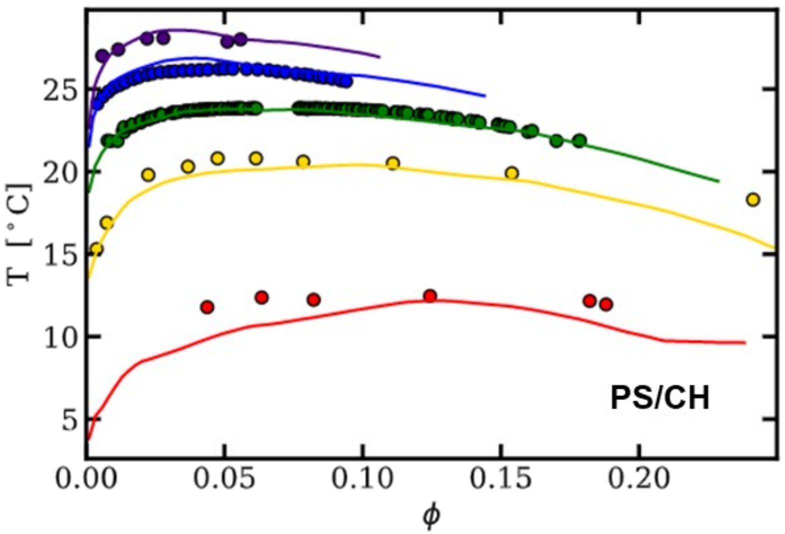
Jeffrey Ethier, Devin Ryan, Richard Vaia (AFRL)

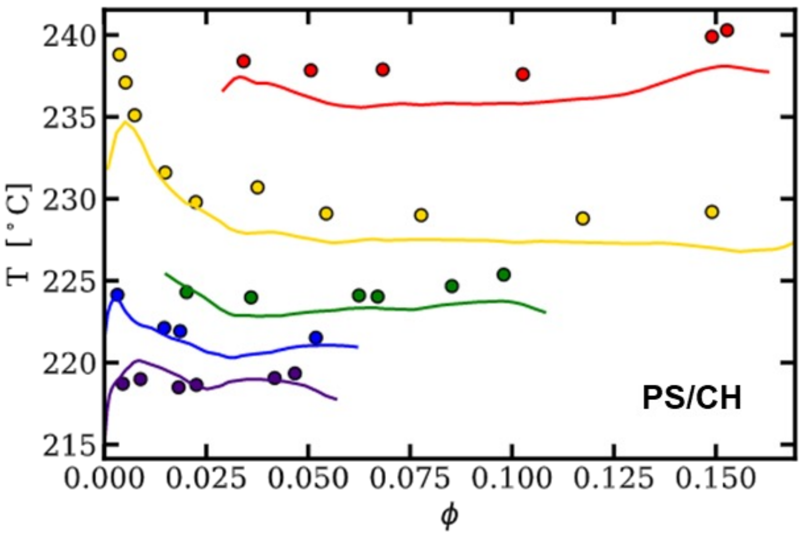
Knychała *et al. Macromolecules* **2017**, *50* (8), 3051–3065

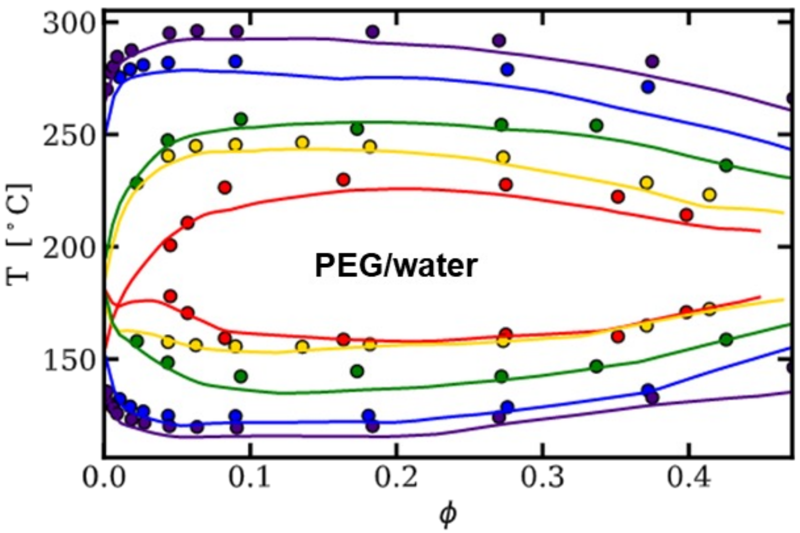
Data driven predictions



Can we improve predictions with theory?





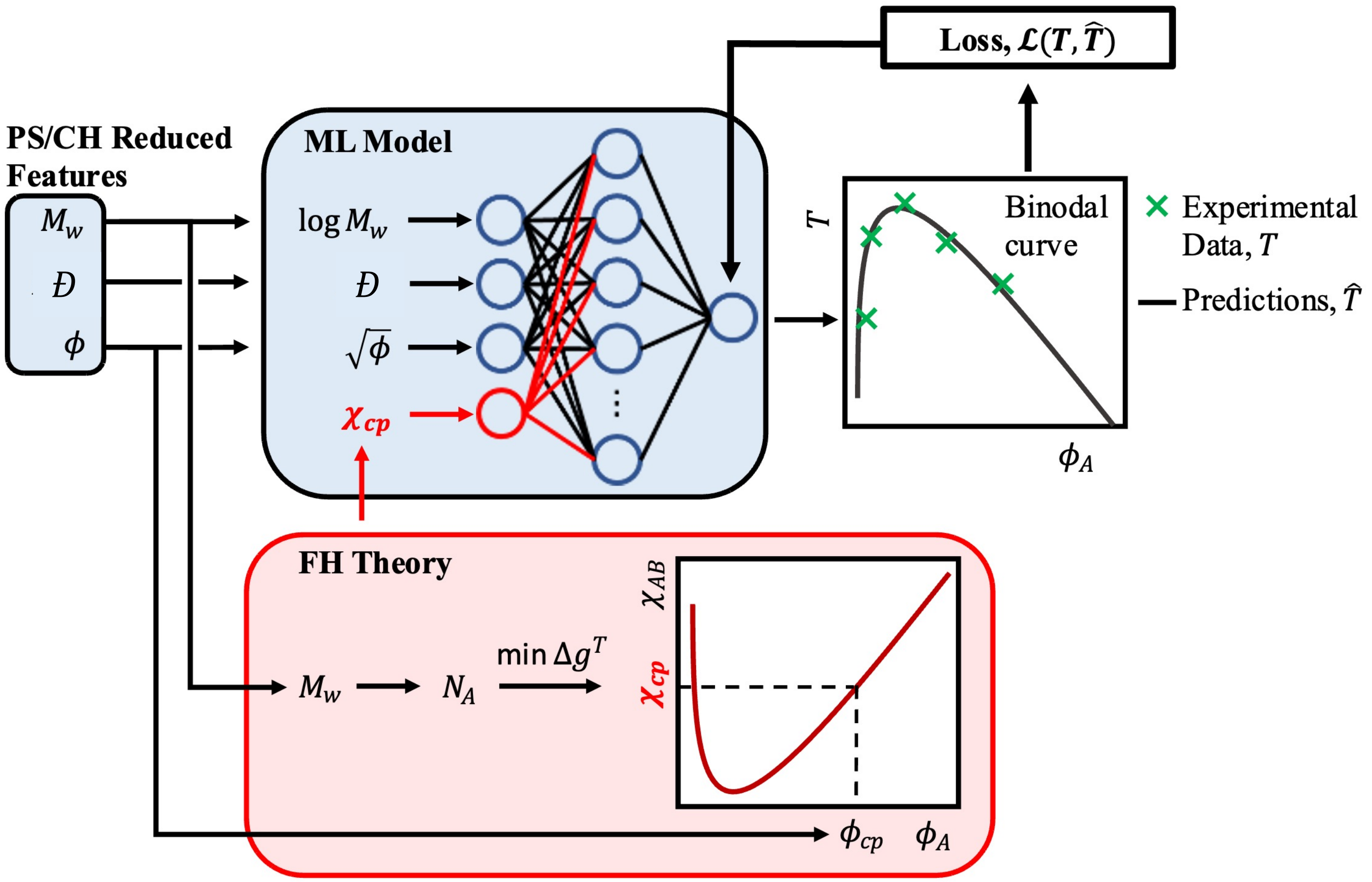


Dataset and predictions at pppdb.uchicago.edu

Ethier et. al., ACS Macro Letters **2021**, 10, 749-754

Ethier et. al., Macromolecules **2022**, 55, 2691-2702

Theory informed models



Baseline Model (Prior Knowledge)

Inputs: $\log M_w, Đ, \sqrt{\phi}$

Outputs: $\hat{T}$

χ-Informed Model

Inputs: χ_{cp}, $\log M_w$, $Đ$, $\sqrt{\phi}$

Outputs: $\hat{T}$

Chi2T Model

Inputs: χ_{cp}

Outputs: $\hat{T}$

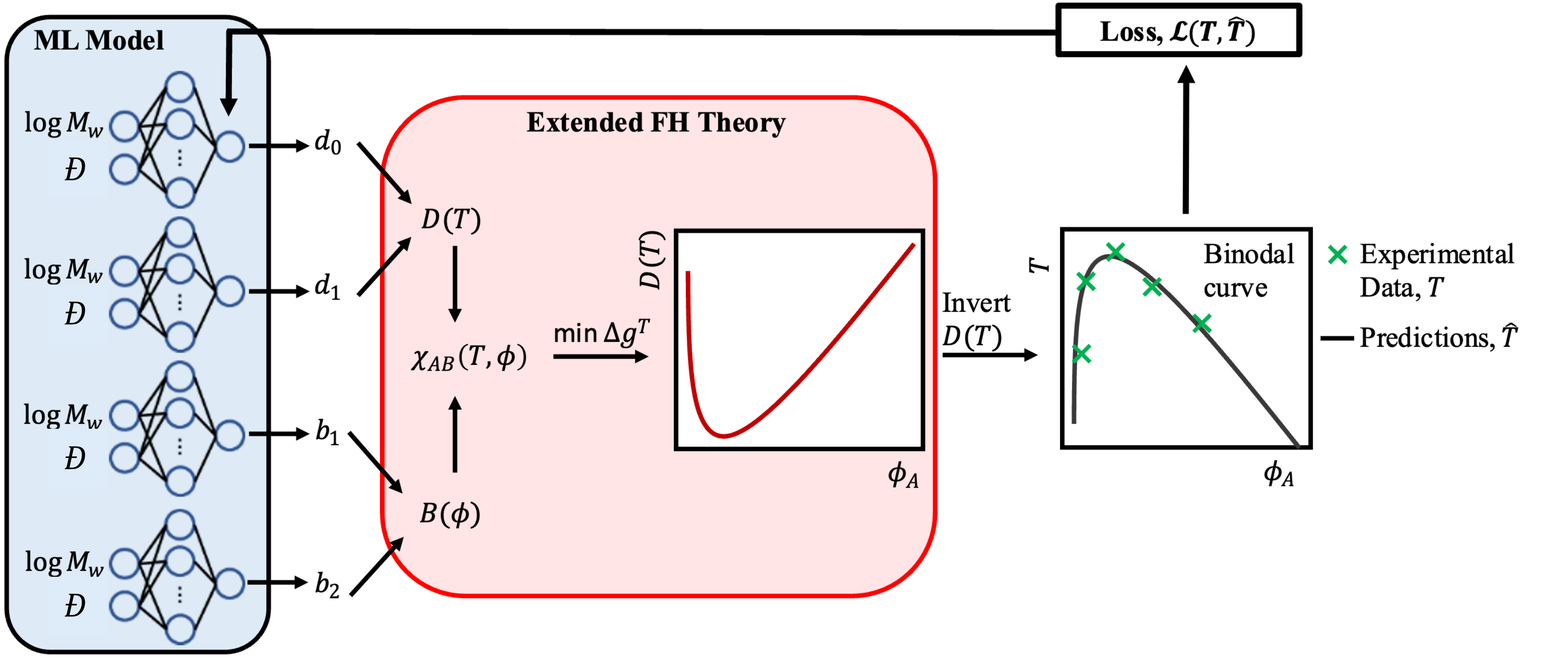
Ethier, Audus, et al. *Giant* **2023**, 15, 100171

Theory constrained model

$$\frac{\Delta G}{n_T k_B T} = \frac{\phi_A}{N_A}\ln \phi_A + \frac{(1-\phi_A)}{N_B}\ln(1-\phi_A) + \phi_A \int_{\phi_A}^{1} \chi_{AB}(T,\phi)d\phi$$

$$D(T)B(\phi_A)$$

$$d_0 + \frac{d_1}{T} \qquad 1 + b_1\phi_A + b_2\phi_A^2$$



Extended FH Model

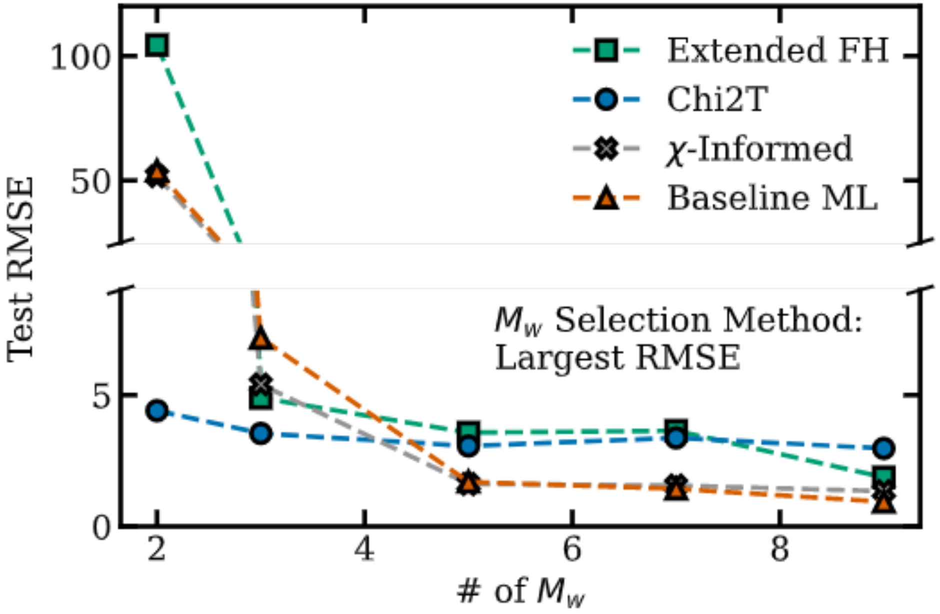
Inputs: $\log M_w$, $Đ$

Output: $\hat{T}$

Ethier, Audus, et al. *Giant* **2023**, 15, 100171

Interpolation

🟩 Extended FH (Theory-Constrained)
🟦 Chi2T (Theory-Informed)
⬜ χ-Informed (Theory-Informed)
🟧 Baseline Model

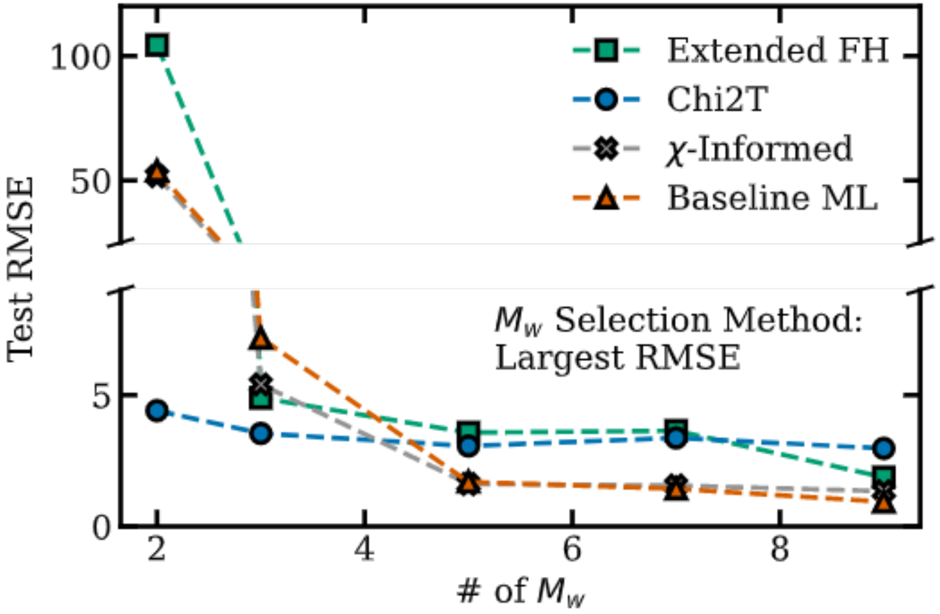


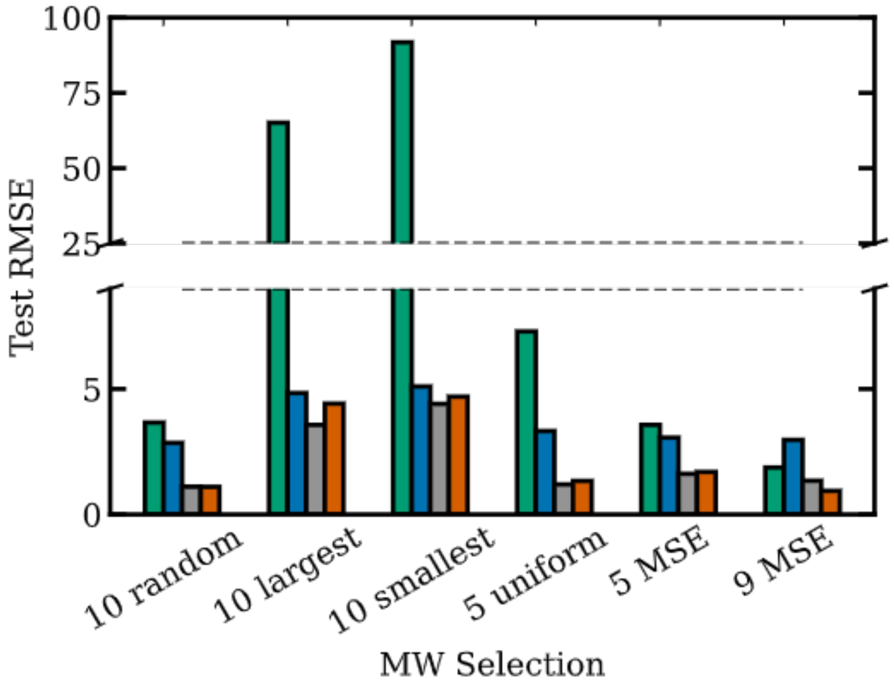
Chi2T works best for extreme data scarcity
χ-informed model works for intermediate and then saturates

Ethier, Audus, et al. *Giant* **2023**, 15, 100171

Interpolation & Extrapolation

🟩 Extended FH (Theory-Constrained)
🟦 Chi2T (Theory-Informed)
⬜ χ-Informed (Theory-Informed)
🟧 Baseline Model

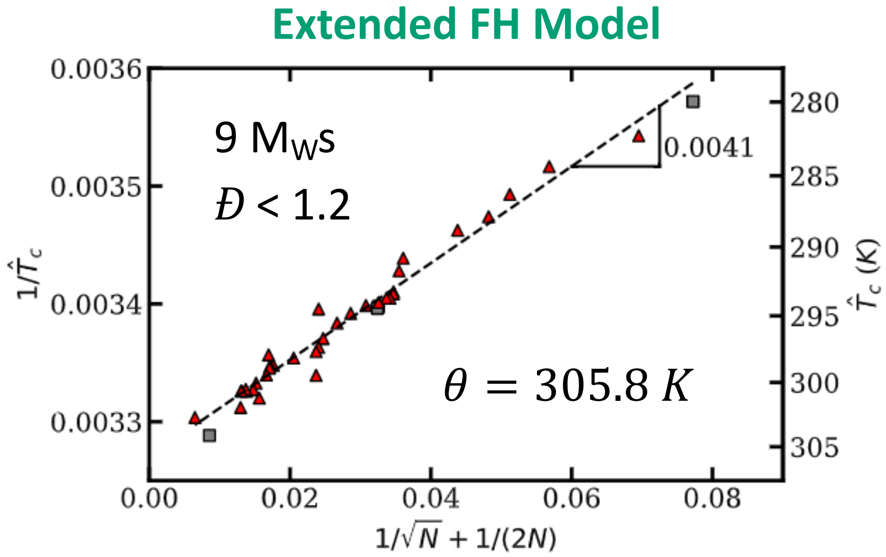


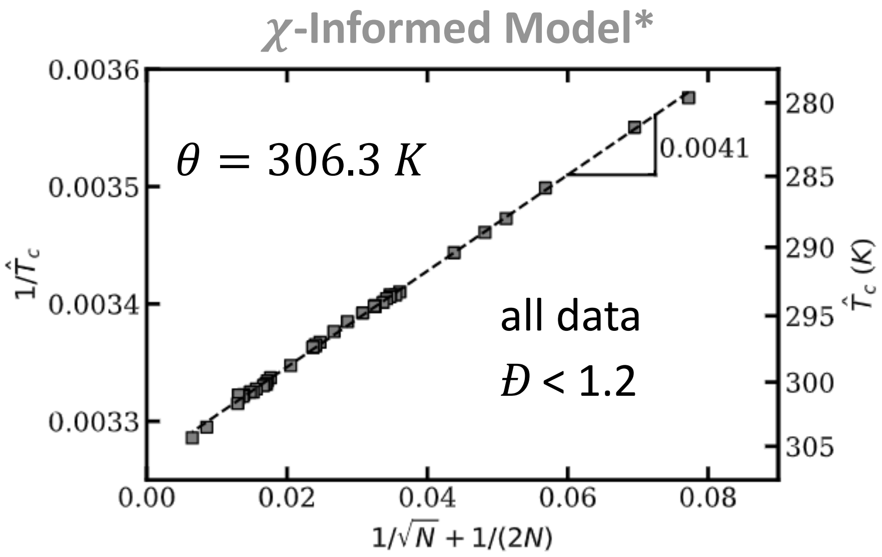


χ-informed model is best for extrapolation

Ethier, Audus, et al. *Giant* **2023**, 15, 100171

Interpretability: Critical Points





$$T_c \sim \frac{1}{\sqrt{N_A}} + \frac{1}{2N_A}$$

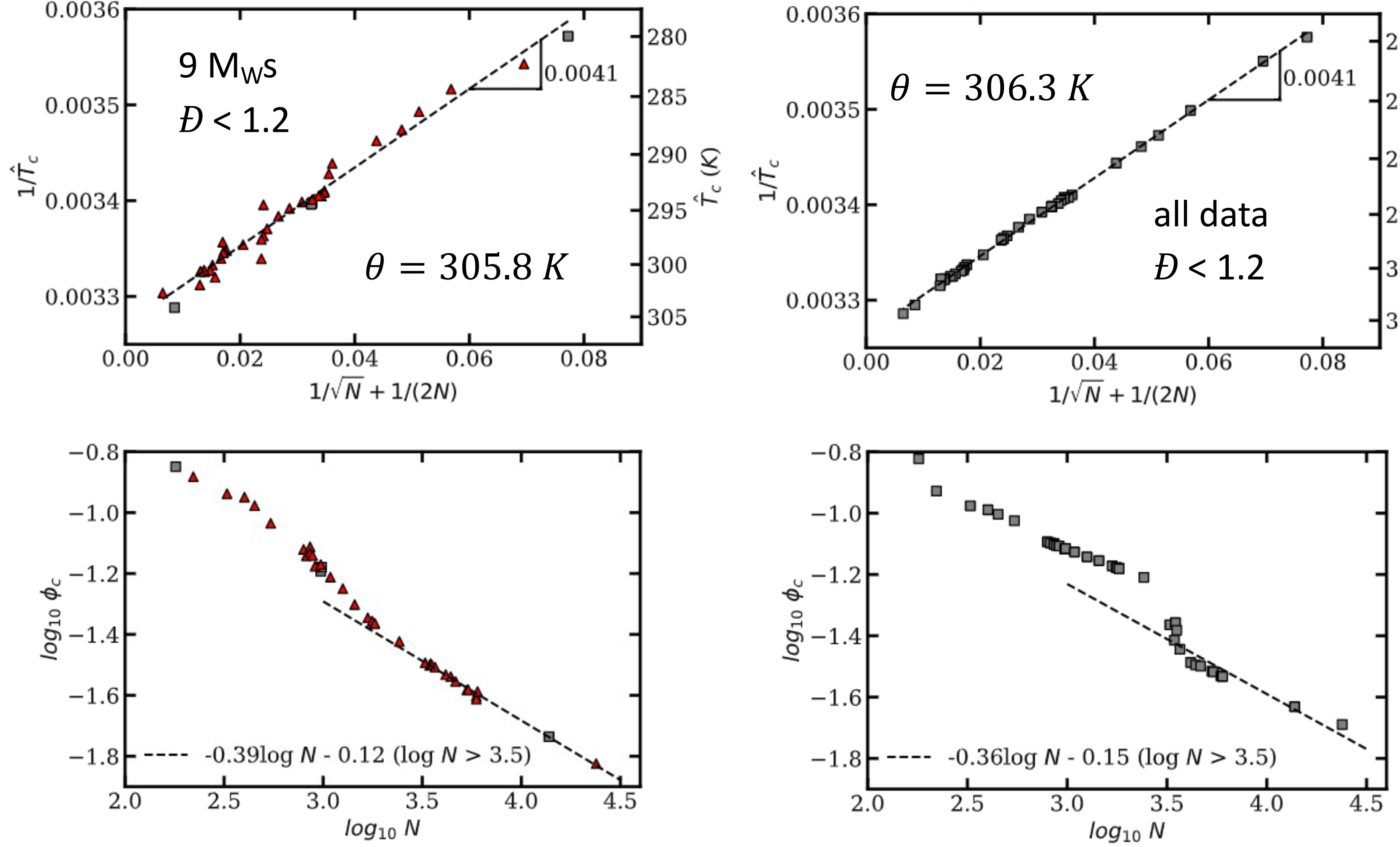
$\theta_{exp} = 307.2$ (Shultz & Flory)

Able to predict T_c

Ethier, Audus, et al. *Giant* **2023**, 15, 100171

*estimated at largest *T* in predicted cloud point curves

Interpretability: Critical Points


Extended FH Model

χ-Informed Model*

$$T_c \sim \frac{1}{\sqrt{N_A}} + \frac{1}{2N_A}$$

$\theta_{exp} = 307.2$ (Shultz & Flory)

Able to predict T_c

$\phi_c \sim N^{-0.38}$ for large N

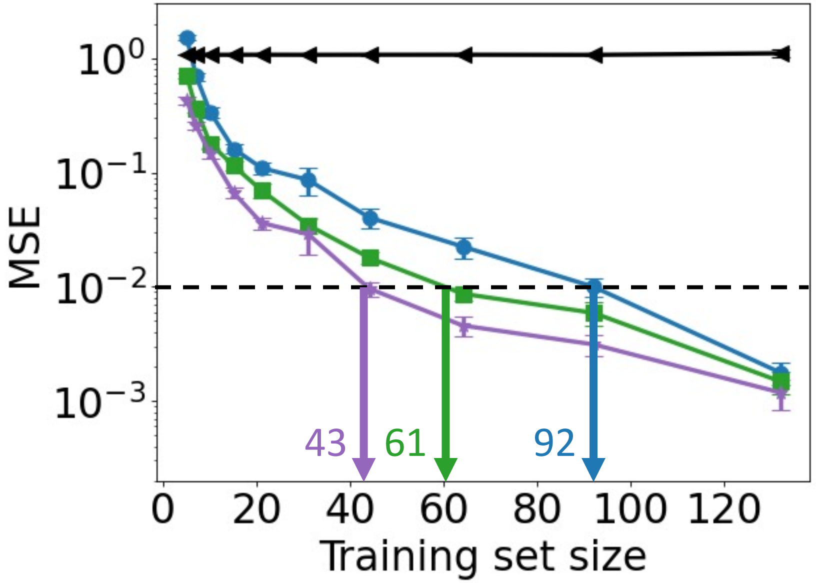
Able to capture scaling

Ethier, Audus, et al. *Giant* **2023**, 15, 100171

*estimated at largest T in predicted cloud point curves

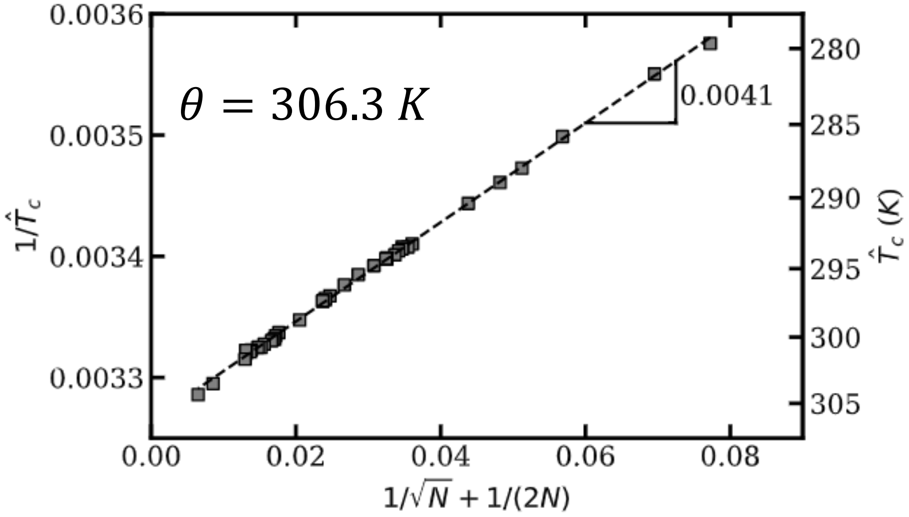
Thoughts on embedding knowledge in ML

- Theory can reduce the data burden, improve extrapolation and provide explainability
- Many methods exist for embedding theory
- Need to be careful not to make theory overly complex



Next steps:

- Explore extrapolation with other chemistries and polymer topologies



Audus *et. al.*, *ACS Macro Letters* **2023**, 11, 1117-1122;
https://github.com/usnistgov/taml

Ethier, Audus, et al. *Giant* **2023**, 15, 100171
https://pppdb.uchicago.edu

Outstanding challenges

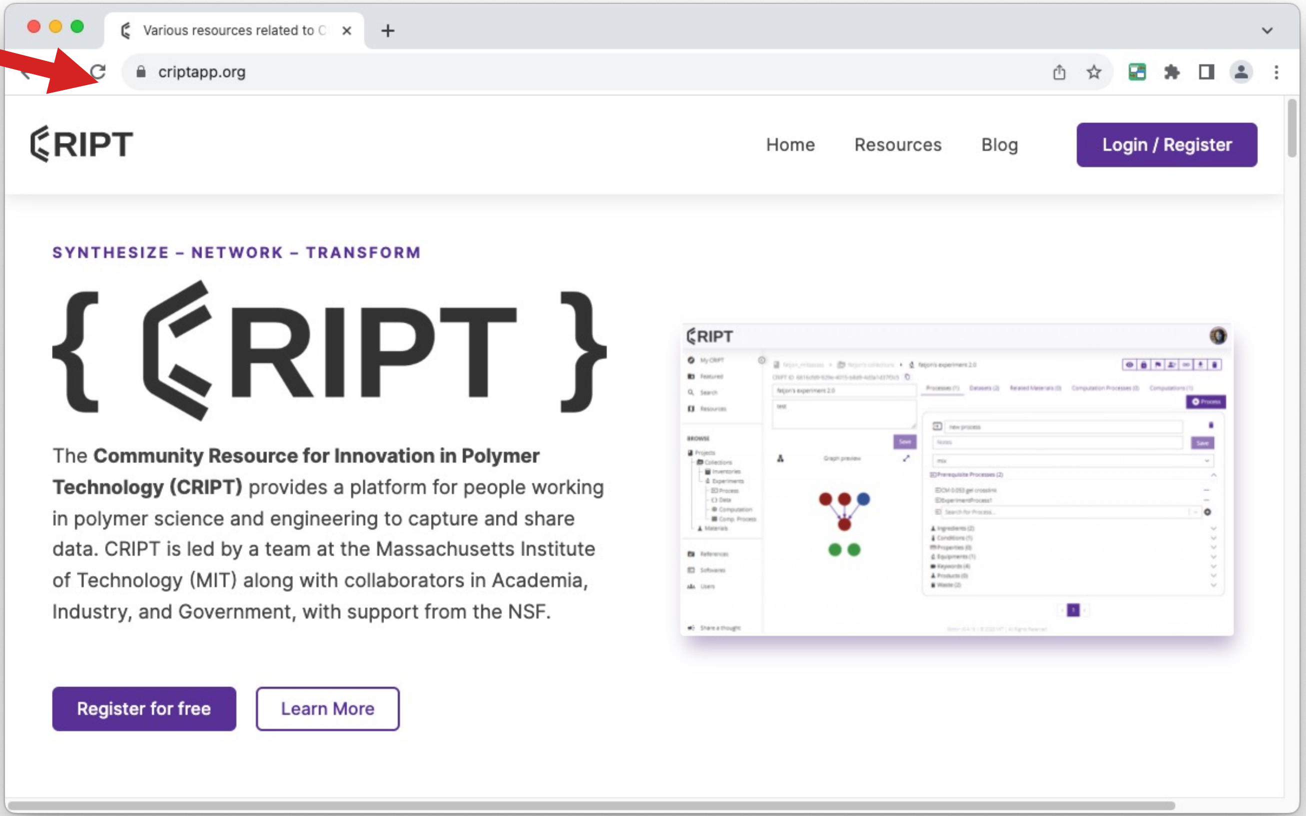
DATA
(fuel for ML/AI)

EXTRAPOLATION
(go beyond the dataset)

EXPLAINABILITY
(answer scientific questions)

Community Resource for Innovation in Polymer Technology (CRIPT)

https://criptapp.org



The **Community Resource for Innovation in Polymer Technology (CRIPT)** provides a platform for people working in polymer science and engineering to capture and share data. CRIPT is led by a team at the Massachusetts Institute of Technology (MIT) along with collaborators in Academia, Industry, and Government, with support from the NSF.

Goal: enable accelerated discovery

Bradley Olsen,* Dylan Walsh, Weizhong Zou, Nathan Rebello, Jiale Shi, Michael Deagan, Bruno Leao, Tzyy-Shyang Lin (MIT), Kaoru Aou (Dow), Ken Kroenlein (Citrine Informatics), Juan de Pablo, Ludwig Schneider, Joshua Mysona (U. Chicago) Debra Audus (NIST), Ardiana Osmani (CRIPT project leader), and the CRIPT Development Team

CRIPT publications

Data model: Walsh,…, Audus, *et al. ACS Cent Sci* **2023**, 9, 3, 330-338

CRIPT overview: Deagen,…, Audus *et al. Cell Reports Physical Science* **2022**, 3, 101126

BigSMILES: Lin *et al. ACS Cent Sci* **2019** 5, 9, 1523-1531

Scheme generation: Deagen *et al. Macromolecules* **2023** 57, 1, 42-53

Search: Rebello et al. *J. Chem. Inf. Model.* 2023 63, 6555-6568

Similarity: Shi, …, Audus, Olsen, *Macromolecules* **2023** 56, 18, 7344-7357