MolSets: Molecular graph deep sets learning for mixture property modeling

Hengrui Zhang, Jie Chen, James Rondinelli, Wei Chen

Northwestern University

March 5, 2024

arXiv:2312.16473

Introduction

(Generated by DALL·E)

Fascination of chemistry:
mix things up and see what happens

Molecular mixture: a broad search space
for materials discovery

coolant

fuel

battery
electrolyte

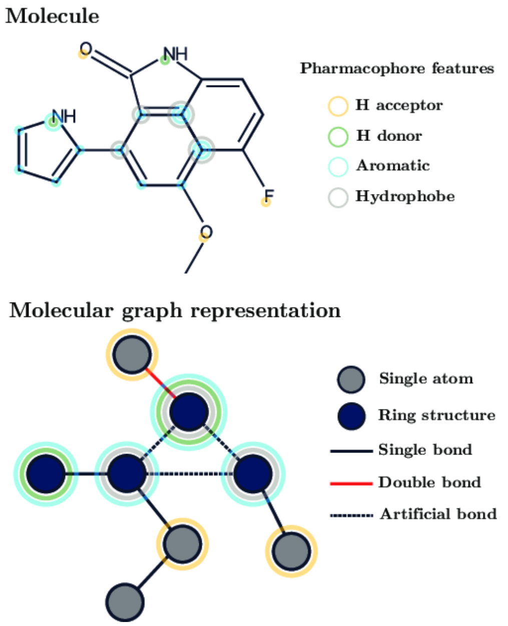
Large, combinatorial space
Machine learning (ML) expedites the search

Challenge: multi-level complexity

- Diverse constituents × configurations

Addressing local and global complexities

Locally, molecules have diverse chemistry and geometries.



Graph representation

- Atoms → nodes
- Atomic properties
- Bonds → edges

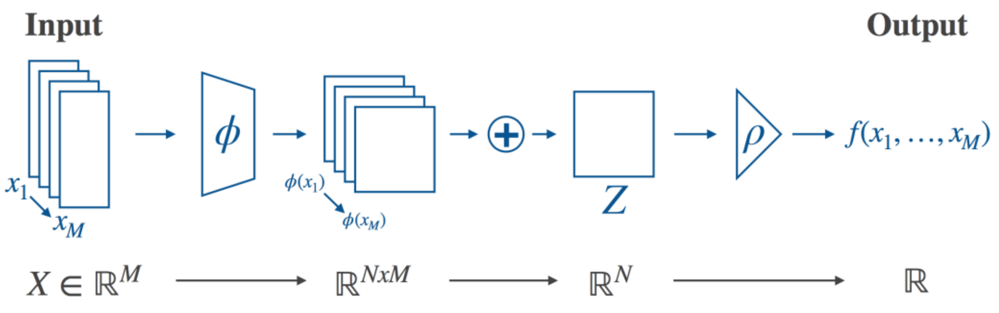
Graph neural network

- Message passing between nodes
- Readout

Globally, mixture should be permutation invariant

$$f(\{x_1, x_2, x_3\}) = f(\{x_2, x_1, x_3\})$$

The “Deep Sets” model architecture



embed aggregate transform

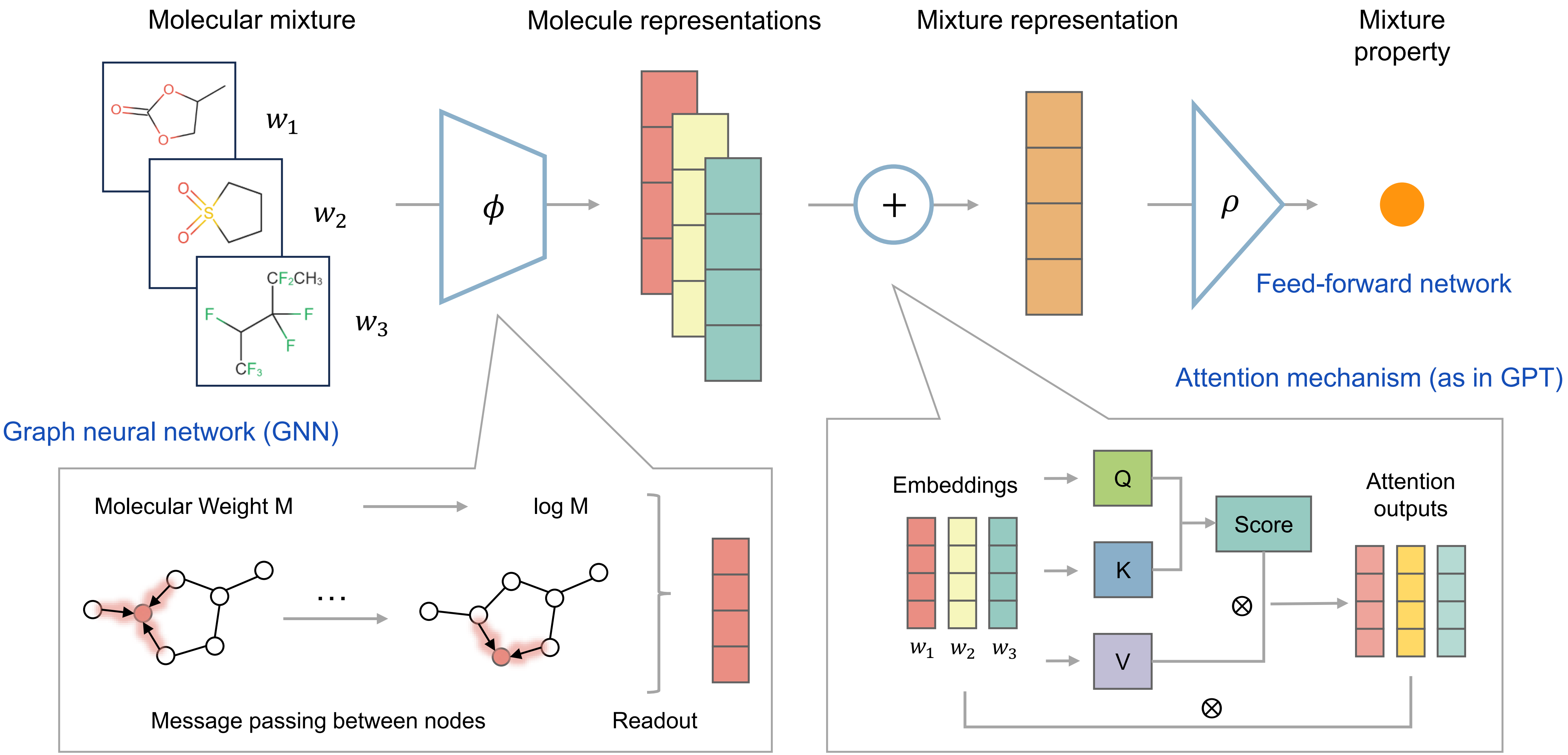
Molecular mixtures as sets:

$$\{(x_1, w_1), (x_2, w_2), \ldots\}$$

* w: weight fraction, ≠ “influence” fraction

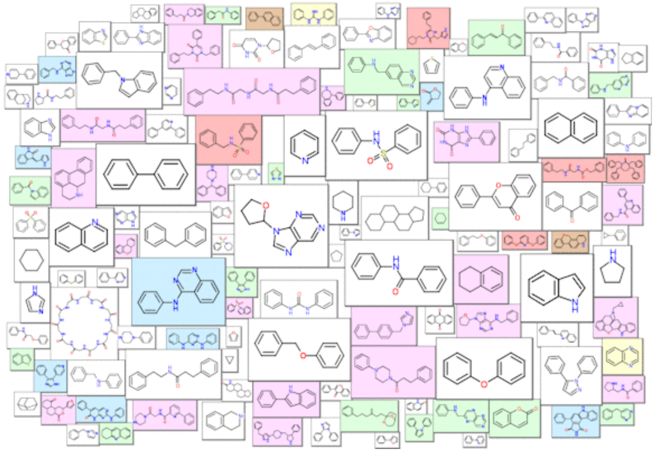
Hernandez, M. et al. *J. Chem. Inf. Model.* 59, 10, 4475–85 (2019); Zaheer, M. et al. *NeurIPS* (2017).

MolSets model architecture

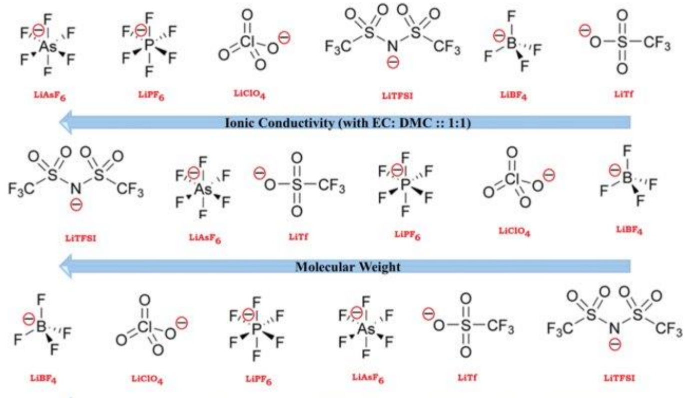


Zhang, H. et al. arXiv:2312.16473; under review (2024).

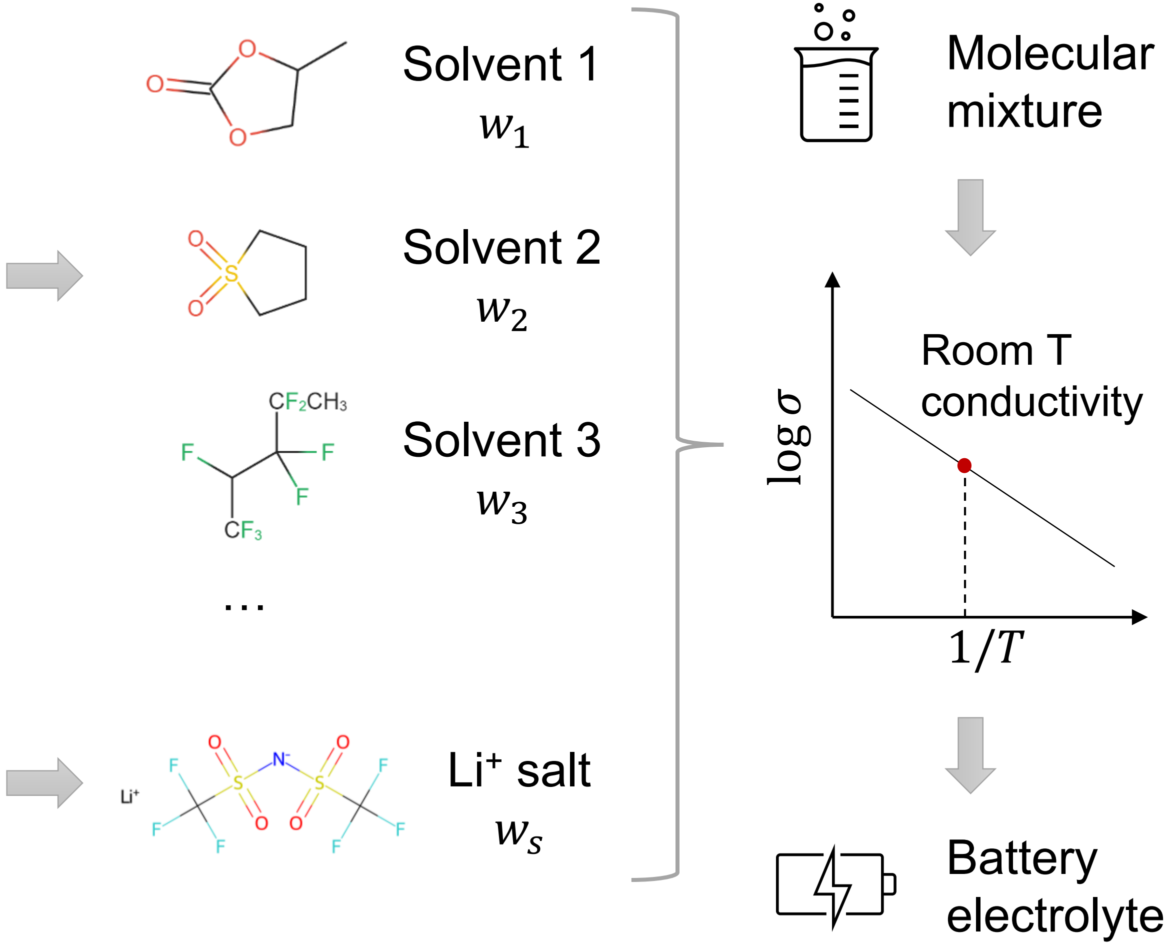
Molecular mixture electrolytes



~200 molecules (incl. polymers)



~50 Li$^+$ salts



Solvent 1 w_1

Solvent 2 w_2

Solvent 3 w_3

…

Li$^+$ salt w_s

Molecular mixture

Room T conductivity

Battery electrolyte

Ertl, P. & Rohde, B., *J. Cheminform.* 4, 12 (2012); Arya, A. & Sharma, A., *J. Mater. Sci.* 55, 6242–6304 (2020).

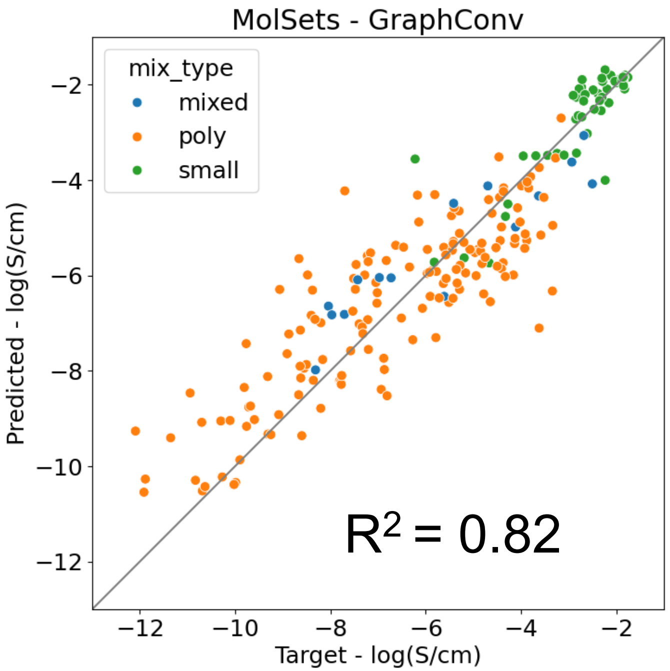
Benchmark

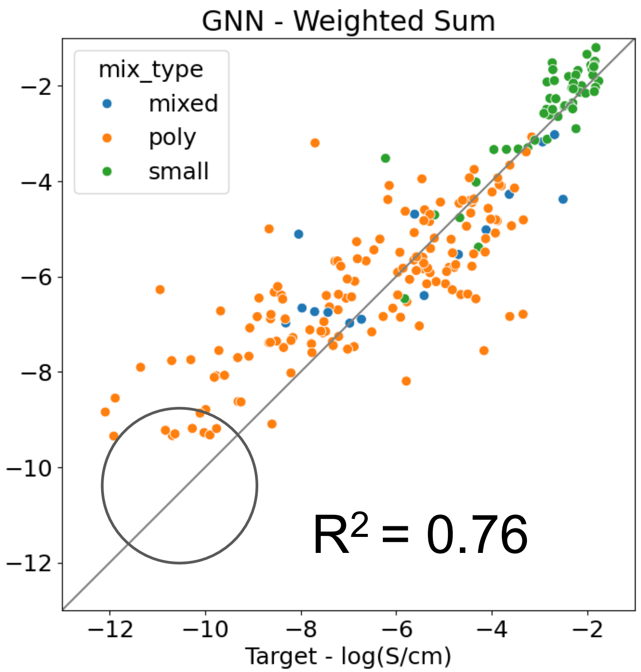
MolSets

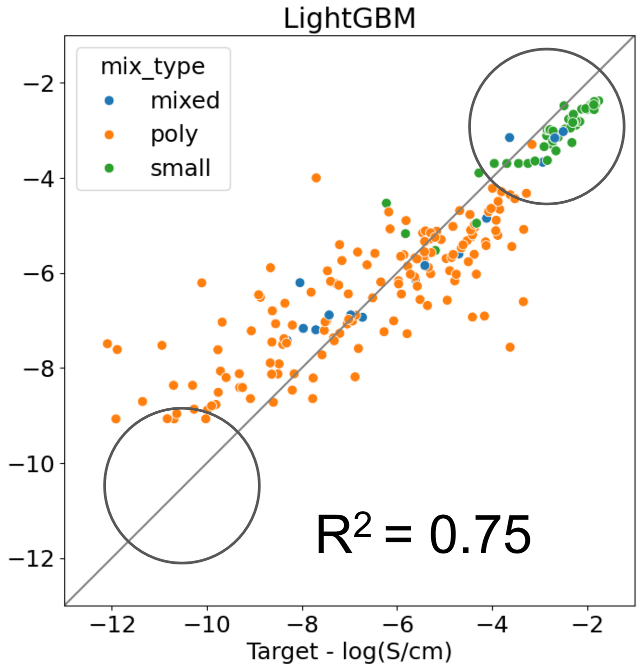
Ablation test:

replace ⊕ with weighted sum

Gradient boosting







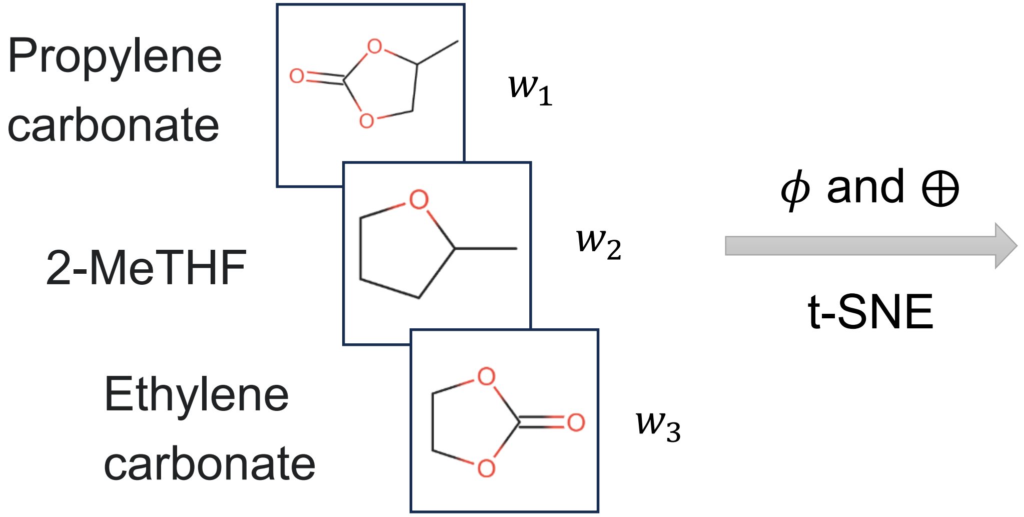
Tests performed on an electrolyte dataset collected in [1] from experimental literature

~1100 unique mixtures; train : validation : test = 3 : 1 : 1

[1] Bradford, G. et al. *ACS Cent. Sci.* 9, 2, 206–216 (2023).

Interpretation

- How is mixture representation different from linear combination of constituents?



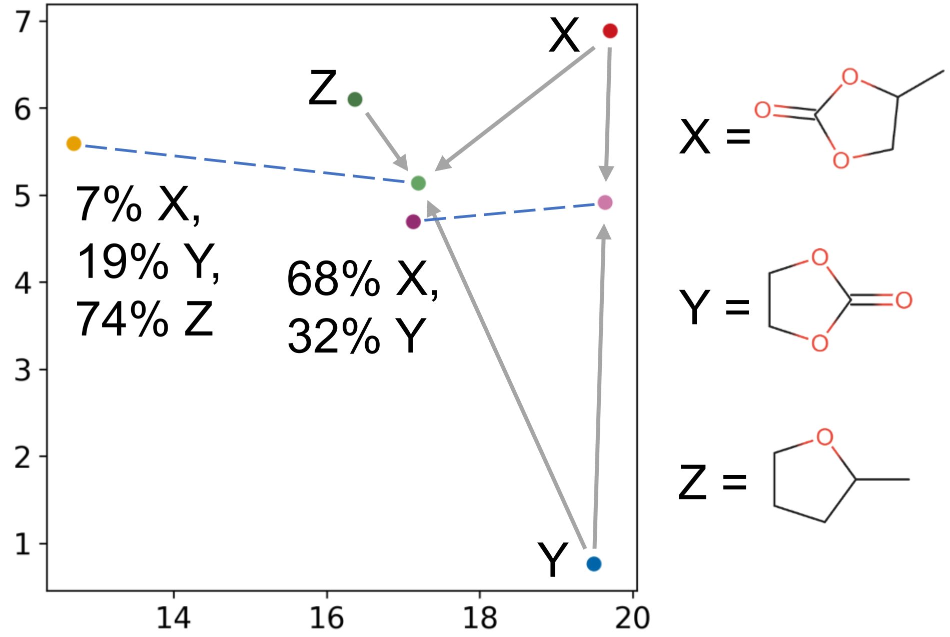
Propylene carbonate

2-MeTHF

Ethylene carbonate

ϕ and $\oplus$

t-SNE



7% X, 19% Y, 74% Z

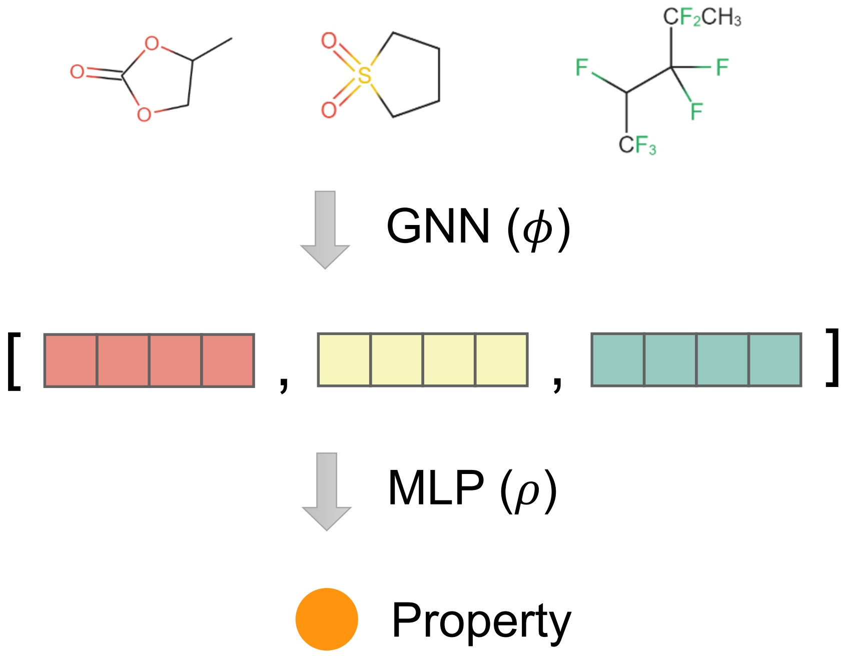
68% X, 32% Y

ϕ and $\oplus$ learns a representation space of mixtures
t-SNE projects the space into 2 dimensions

→ Linear combination of ingredients
--- Mixture representation

Interpretation

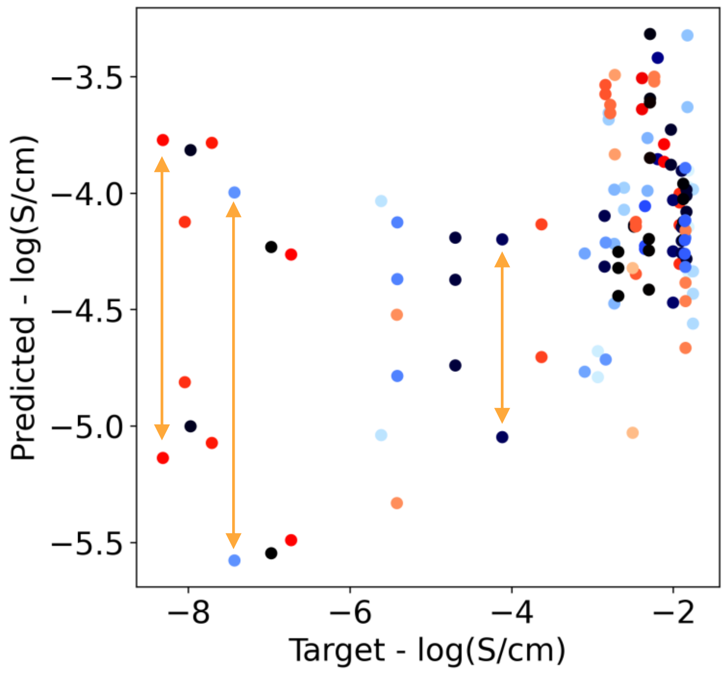
- What happens if a model doesn't preserve permutation invariance?



GNN (ϕ)

MLP (ρ)

Property



Concatenate molecular representations instead of aggregating

Same mixture, change sequence → predicted differently

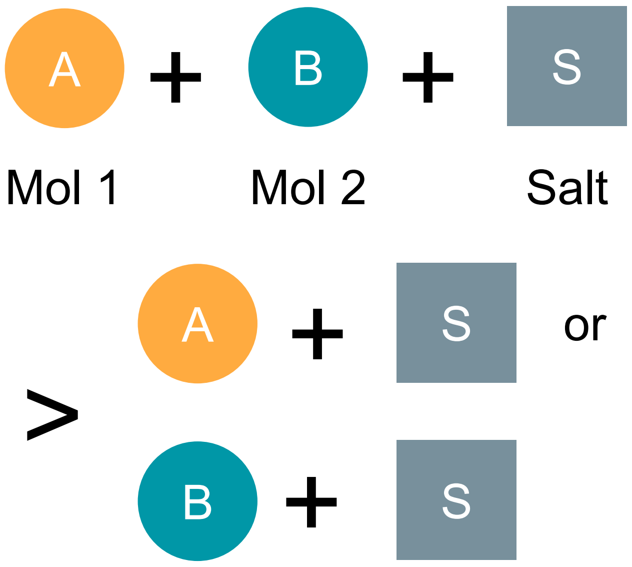
Virtual screening

- Train on all data from [1], predict 298 K conductivity for 11,340 new candidates
  - Equal-weight binary mixture among 28 molecules × 30 salts (1 mol/kg)
- Data available on Dryad

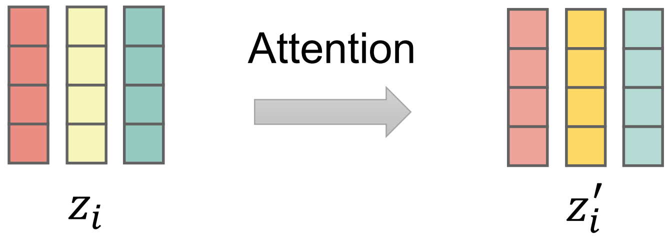
Dryad dataset

“Bowing effect”

(nonlinear mixing behavior)



Relative importance



Magnitude change $||z'||/||z||$ indicates importance

Molecules found important include:

- Cyclic carbonate esters (ion solvation ability)
- Benzene, toluene (incorrect)

Discussion

A limitation: current model didn’t learn solubility

Reason: limited data availability

- In experimental datasets, salt never exceeds solubility

⚠ Rationally choose salt molality in application

Future work directions

- Synergy with high-throughput experiments: “robots” make mixtures and measure properties
  - e.g., Coulomb efficiency, order parameters
- Build a data platform: access MolSets’ predictions

🏳 “AlphaFold” or “matterverse” for mixtures

arXiv:2312.16473

Thank you!

Acknowledgments:

Jie Chen

James Rondinelli

Wei Chen

NSF DMREF MATERIALS FOR OUR FUTURE

DMR-2324173
DMR-2219489

Northwestern INTERNATIONAL INSTITUTE FOR NANOTECHNOLOGY

Ryan Graduate
Fellowship