

Of graphs, sets, and molecules

ML4Molecules - ELLIS UnConference

2025-12-02

Daniel Probst, Wageningen University & Research

It's always graphs

- SMILES (or SELFIES, DeepSMILES, ...) are a way to represent (and store) graphs as strings
- ... but then again, “Transformers are Graph Neural Networks” [1]
- The baseline for GNNs applied to molecular property prediction is quite strong (when also adding RDKit descriptors)

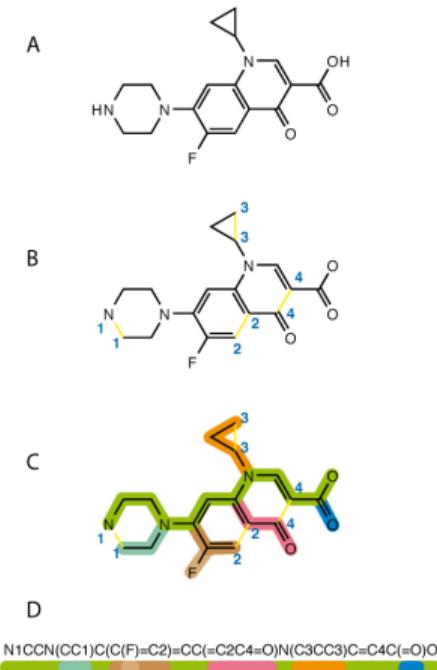


Figure 1: SMILES construction
(Source: Wikipedia)

What about sets?

- ... so we got (almost) rid of the bonds and thought of molecules as bags of atoms [2]
- This worked very well on the MoleculeNet benchmark... but everything does perform well on this

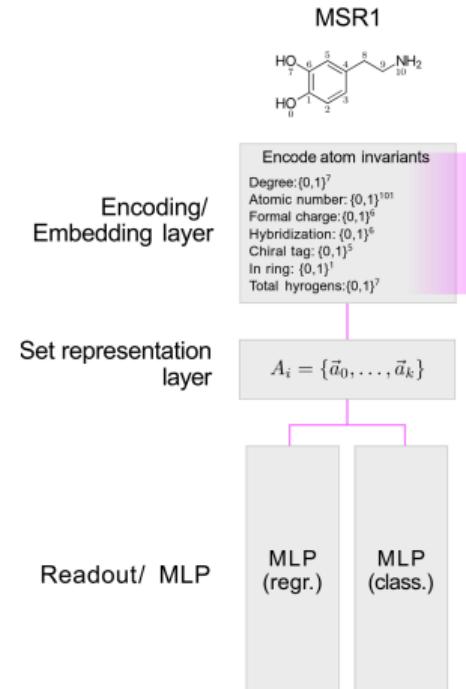


Figure 2: Molecular set representation architecture

What about sets?

4 / 21

Data Set	Metric	MGCN	SchNet	GCN	GIN	D-MPNN	MSR1
HIV	AUROC	73.8 ± 1.6	70.2 ± 3.4	71.6 ± 4.0	75.3 ± 1.9	<u>75.0 ± 2.1</u>	72.3 ± 2.2
BACE	AUROC	73.4 ± 3.0	76.6 ± 1.1	71.6 ± 2.0	70.1 ± 5.4	<u>85.3 ± 5.3</u>	75.5 ± 1.9
BBBP	AUROC	85.0 ± 6.4	84.8 ± 2.2	71.8 ± 0.9	65.6 ± 0.5	71.2 ± 3.8	<u>71.4 ± 0.0</u>
Tox21	AUROC	70.7 ± 1.6	77.2 ± 2.3	70.9 ± 2.6	74.0 ± 0.8	68.9 ± 1.3	<u>72.1 ± 5.0</u>
SIDER	AUROC	55.2 ± 1.8	53.9 ± 3.7	53.6 ± 3.2	57.3 ± 1.6	<u>63.2 ± 2.3</u>	61.4 ± 7.3
ClinTox	AUROC	63.4 ± 4.2	71.5 ± 3.7	62.5 ± 2.8	58.0 ± 4.4	<u>90.5 ± 5.3</u>	86.6 ± 1.2
ESOL	RMSE	1.27 ± 0.15	1.05 ± 0.06	1.43 ± 0.05	1.45 ± 0.02	0.98 ± 0.26	<u>0.59 ± 0.03</u>
FreeSolv	RMSE	3.35 ± 0.01	3.22 ± 0.76	2.87 ± 0.14	2.76 ± 0.18	2.18 ± 0.91	<u>1.94 ± 0.24</u>
Lipo	RMSE	1.11 ± 0.04	0.91 ± 0.10	0.85 ± 0.08	0.85 ± 0.07	<u>0.65 ± 0.05</u>	0.85 ± 0.03
QM7	MAE	77.6 ± 4.7	74.2 ± 6.0	122.9 ± 2.2	124.8 ± 0.7	105.8 ± 13.2	<u>85.9 ± 13.1</u>
QM8	MAE	0.022 ± 0.002	0.020 ± 0.002	0.037 ± 0.001	0.037 ± 0.001	<u>0.014 ± 0.002</u>	0.023 ± 0.010

What about sets of node embeddings?

5 / 21

- What about combining the set representation approach with GNNs?
- Embedding nodes, followed by a set representation layer instead of global pooling
- Turns out, this works quite well on more representative benchmarks

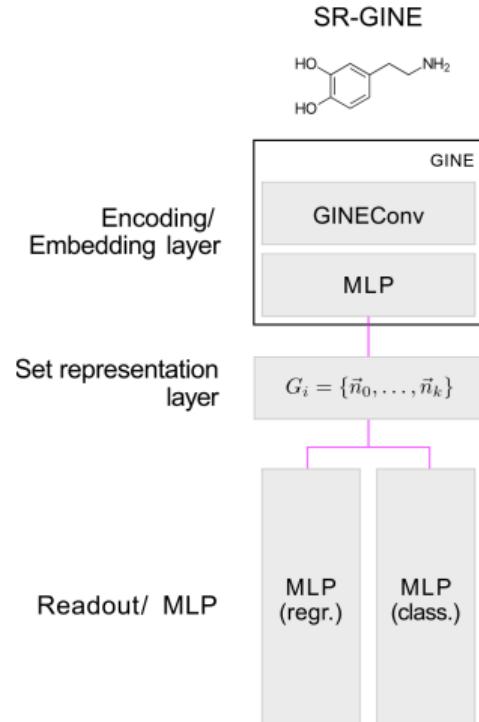


Figure 3: SR-GINE architecture

What about sets of node embeddings?

6 / 21

Data from Biogen ADME in vitro assays. The metric used is the Pearson correlation. Similar results on OCELOT chromophore data containing chemically diverse π -conjugated molecules [3], [4].

Endpoint	D-MPNN	D-MPNN+	MSR1	GINE	SR-GINE
HLM	<u>0.65</u>	0.68	0.58 ± 0.00	0.59 ± 0.01	0.68 ± 0.01
MDR1-MDCK ER	<u>0.72</u>	0.78	0.68 ± 0.02	0.67 ± 0.00	0.80 ± 0.01
Solubility	0.64	0.59	0.47 ± 0.04	0.55 ± 0.02	<u>0.63 ± 0.02</u>
RLM	<u>0.72</u>	0.74	0.57 ± 0.01	0.61 ± 0.02	<u>0.70 ± 0.01</u>
hPPB	<u>0.74</u>	0.77	0.68 ± 0.01	0.68 ± 0.02	0.77 ± 0.01
rPPB	<u>0.70</u>	0.70	0.55 ± 0.02	0.63 ± 0.02	0.73 ± 0.02
Average	0.695 ± 0.037	<u>0.710 ± 0.064</u>	0.588 ± 0.08	0.622 ± 0.05	0.718 ± 0.06

Sets of molecules (reactions)

- There are other sets, such as ...
- ... sets of atoms of different types
- ... sets of molecules

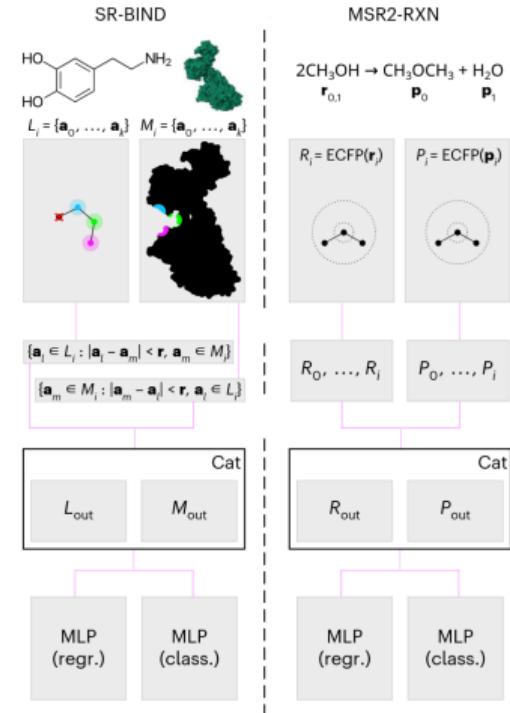


Figure 4: MSR: Complexes and reactions

Sets of molecules (reactions)

8 / 21

Yield prediction on Buchwald-Hartwig reactions from high-throughput experiments (HTE) and extracted from an electronic laboratory notebook (ELN) [5], [6]. The metric used is R^2 .

Data Set	Subset/Split	Yield-BERT	DRFP	YieldGNN	MSR2-RXN
HTE	Rand 70/30	0.95 ± 0.005	0.95 ± 0.005	0.96 ± 0.005	0.94 ± 0.005
	Rand 50/50	0.92 ± 0.01	0.93 ± 0.01	-	0.93 ± 0.01
	Rand 30/70	0.88 ± 0.01	0.89 ± 0.01	-	0.90 ± 0.01
	Rand 20/80	0.86 ± 0.01	0.87 ± 0.01	-	0.87 ± 0.01
	Rand 10/90	0.79 ± 0.02	0.81 ± 0.01	-	0.80 ± 0.02
	Rand 5/95	0.61 ± 0.04	0.73 ± 0.02	-	0.69 ± 0.03
	Rand 2.5/97.5	0.45 ± 0.05	0.62 ± 0.04	-	0.57 ± 0.05
	Avg. 1-4	0.73	0.71 ± 0.16	-	0.72 ± 0.15
ELN	-	-0.006 ± 0.11	0.20 ± 0.05	0.05 ± 0.07	0.13 ± 0.08

Speaking of sets and reactions ...

- We recently published an update on the differential reaction fingerprint (DRFP) where we set SOTA on the ELN data set
- See results on previous slide
- More on sets, molecules, and reactions coming soon ...

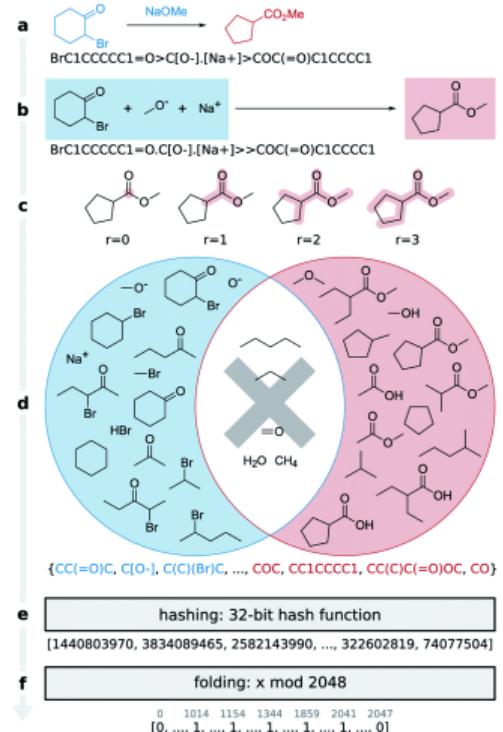


Figure 5: DRFP Scheme

Speaking of sets and reactions ...

10 / 21

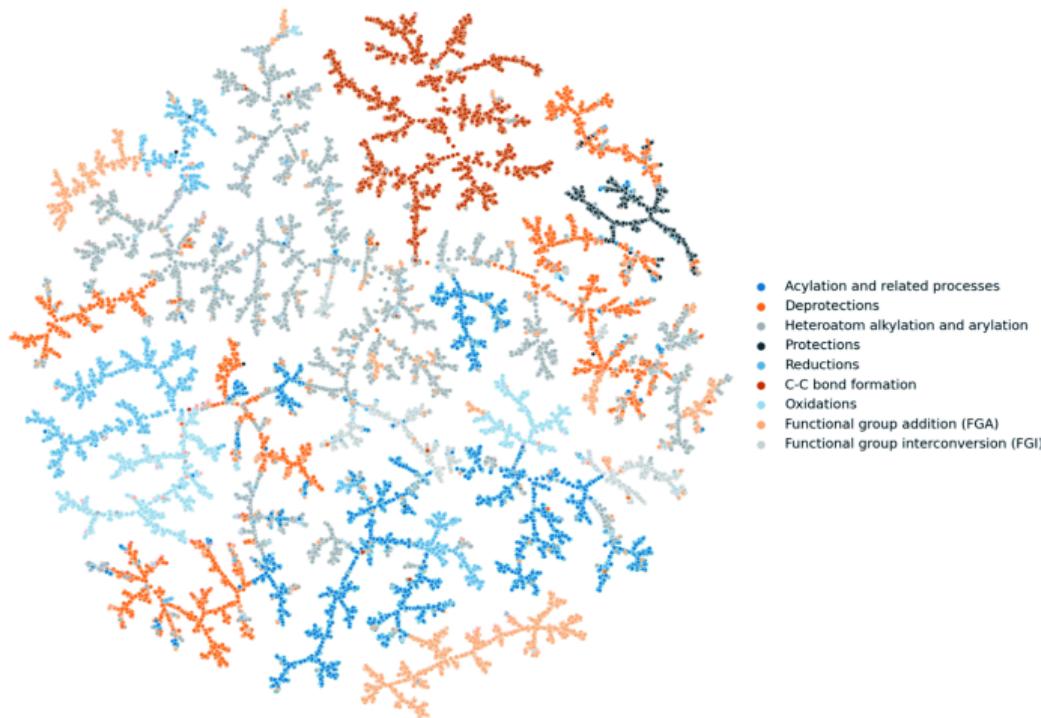


Figure 6: DRFP-encoded USPTO subset visualised using TMAP.

- Mass spectra are commonly binned to account for different numbers of peaks between samples
- Why not represent them as graphs and predict the quantitative estimate of drug-likeness (QED)?
- Intensities as node attributes and m/z differences as edge weights [7]

Model	Params	MAE (↓)	RMSE (↓)	Pearson's r (↑)	R^2 (↑)
MLP	11.0 M	0.145 ± 0.008	0.200 ± 0.008	0.736 ± 0.011	0.437 ± 0.043
SetTransformer	0.4 M	0.134 ± 0.002	0.174 ± 0.001	0.758 ± 0.004	0.572 ± 0.006
GNN (GAT)	12.6 M	0.110 ± 0.006	0.144 ± 0.006	0.843 ± 0.015	0.709 ± 0.025

Molecular dynamics as graphs

12 / 21

- Idea: Combine structure and dynamics in a heterogeneous graph [8]
- 8 ns MD per protein (or complex) from the MISATO data set [9]
- Significant performance improvements in binding site detection, binding affinity prediction, and atomic adaptability prediction

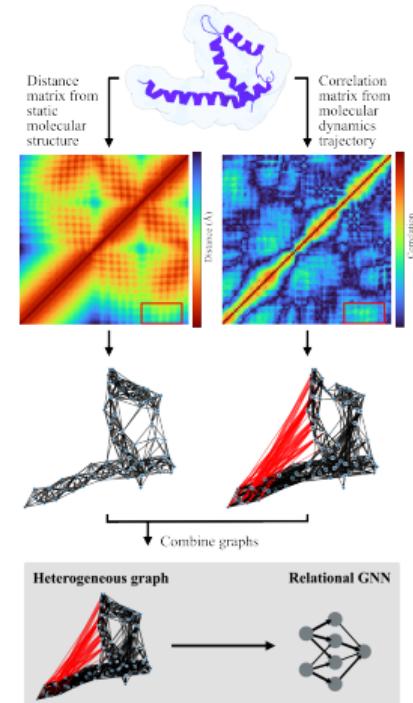


Figure 7: Static-dynamic RGNN

Molecular dynamics as graphs

- Hypothesis: In addition to the dynamics information, dynamics edges reduce over-squashing and improve long-range information propagation
- More recent research showed that random edges do not have a similar effect on performance

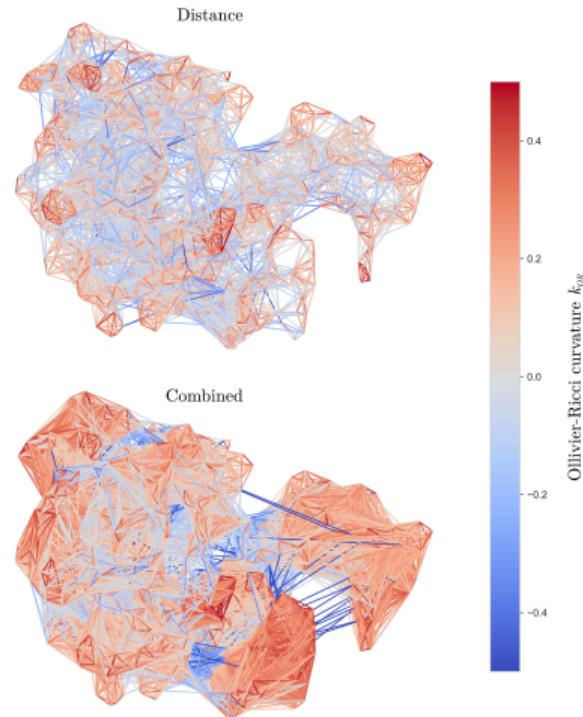


Figure 8: Ollivier-Ricci curvature

Modelling protein conformations

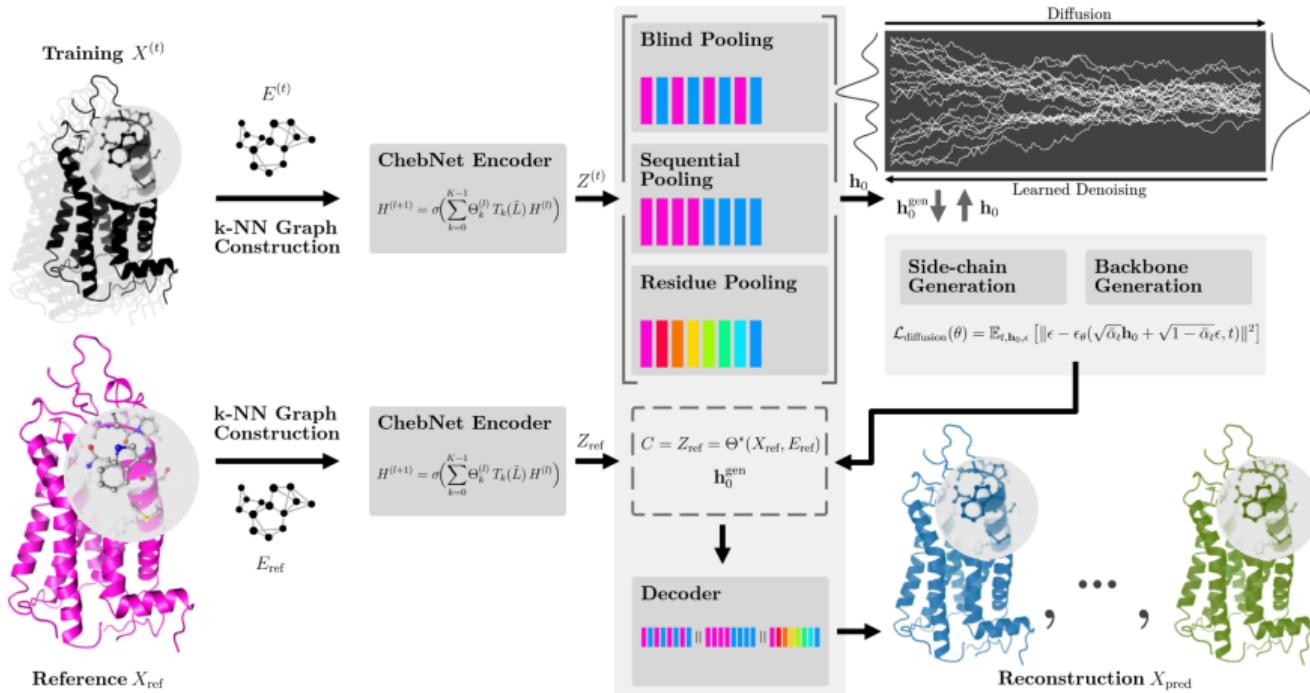


Figure 9: Schematic of the LD-FPG framework.

Modelling protein conformations

15 / 21

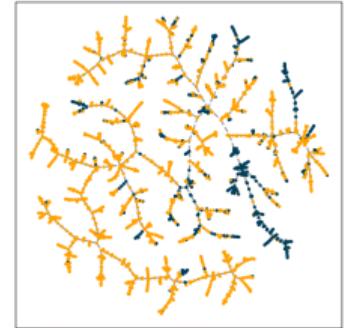
- Autoencoder architecture utilising **ChebNet** combined with residue-based pooling [10]
- As it turns out, ChebNet performs well compared to other GNN architectures on long-range tasks
- But you have probably heard enough about protein ensemble generation already ...

Back to SMILES and ... GZIP

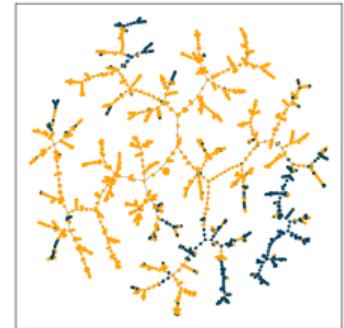
16 / 21

Model		RMSE	MAE	R
GraphDTA	GCN	1.735 ± 0.034	1.343 ± 0.037	0.613 ± 0.016
	GAT	1.765 ± 0.026	1.354 ± 0.033	0.601 ± 0.016
	GIN	1.640 ± 0.044	1.261 ± 0.044	0.667 ± 0.018
	GAT-GCN	1.562 ± 0.022	1.191 ± 0.016	0.697 ± 0.008
GNN-based	GNN-DTI	1.492 ± 0.025	1.192 ± 0.032	0.736 ± 0.021
	D-MPNN	1.493 ± 0.016	1.188 ± 0.009	0.729 ± 0.006
	MAT	1.457 ± 0.037	1.154 ± 0.037	0.747 ± 0.013
	DimeNet	1.453 ± 0.027	1.138 ± 0.026	0.752 ± 0.010
	CMPNN	1.408 ± 0.028	1.117 ± 0.031	0.765 ± 0.009
Compression-based	MolZip	1.508 ± 0.000	1.190 ± 0.000	0.720 ± 0.000
	MolZip Aug	<u>1.422 ± 0.017</u>	<u>1.131 ± 0.014</u>	<u>0.757 ± 0.007</u>

- Based on the normalised compression distance
- $\text{NCD}(x, y) = \frac{0.5(C(xy)+C(yx)) - \min\{C(xx), C(yy)\}}{\max\{C(xx), C(yy)\}}$
- approximates normalised information distance, specifically the Kolmogorov complexity, which is uncomputable (\rightarrow halting problem) in
- $\text{NID}(x, y) = \frac{\max\{K(x|y), K(y|x)\}}{\max\{K(x), K(y)\}}$
- Based on this, we build kNN classifiers and regressors [11]



(a) ECFP-based TMAP



(b) MolZip-based TMAP

Figure 10: TMAPs of BBBP

- daniel.probst@wur.nl
- probstlab.science
- github.com/daenuprobst/talks

- [1] C. K. Joshi, “Transformers are Graph Neural Networks.” [Online]. Available: <https://arxiv.org/abs/2506.22084>
- [2] M. Boulogouri, P. Vandergheynst, and D. Probst, “Molecular set representation learning,” *Nature Machine Intelligence*, vol. 6, no. 7, pp. 754–763, July 2024, doi: [10.1038/s42256-024-00856-0](https://doi.org/10.1038/s42256-024-00856-0).
- [3] C. Fang *et al.*, “Prospective Validation of Machine Learning Algorithms for Absorption, Distribution, Metabolism, and Excretion Prediction: An Industrial Perspective,” *Journal of Chemical Information and Modeling*, vol. 63, no. 11, pp. 3263–3274, June 2023, doi: [10.1021/acs.jcim.3c00160](https://doi.org/10.1021/acs.jcim.3c00160).
- [4] V. Bhat, P. Sornberger, B. S. S. Pokuri, R. Duke, B. Ganapathysubramanian, and C. Risko, “Electronic, redox, and optical property prediction of organic *pi*-conjugated

- molecules through a hierarchy of machine learning approaches”, *Chemical Science*, vol. 14, no. 1, pp. 203–213, Dec. 2022, doi: [10.1039/D2SC04676H](https://doi.org/10.1039/D2SC04676H).
- [5] D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, and A. G. Doyle, “Predicting reaction performance in C–N cross-coupling using machine learning,” *Science*, vol. 360, no. 6385, pp. 186–190, Apr. 2018, doi: [10.1126/science.aar5169](https://doi.org/10.1126/science.aar5169).
 - [6] M. Saebi *et al.*, “On the use of real-world datasets for reaction yield prediction,” *Chemical Science*, vol. 14, no. 19, pp. 4997–5005, 2023, doi: [10.1039/D2SC06041H](https://doi.org/10.1039/D2SC06041H).
 - [7] N. de Jonge, J. J. J. van der Hooft, and D. Probst, “To Bin or not to Bin: Alternative Representations of Mass Spectra.” [Online]. Available: <https://arxiv.org/abs/2502.10851>
 - [8] P. Guo, B. Correia, P. Vandergheynst, and D. Probst, “Boosting Protein Graph Representations through Static-Dynamic Fusion,” in *Proceedings of the 42nd International Conference on Machine Learning*, A. Singh, M. Fazel, D. Hsu, S. Lacoste-

References (iii)

21 / 21

- Julien, F. Berkenkamp, T. Maharaj, K. Wagstaff, and J. Zhu, Eds., in Proceedings of Machine Learning Research, vol. 267. PMLR, 2025, pp. 20777–20792.
- [9] T. Siebenmorgen *et al.*, “MISATO: machine learning dataset of protein–ligand complexes for structure-based drug discovery,” *Nature Computational Science*, vol. 4, no. 5, pp. 367–378, May 2024, doi: [10.1038/s43588-024-00627-2](https://doi.org/10.1038/s43588-024-00627-2).
- [10] A. Sengar, A. Hariri, D. Probst, P. Barth, and P. Vandergheynst, “Generative Modeling of Full-Atom Protein Conformations using Latent Diffusion on Graph Embeddings.” [Online]. Available: <https://arxiv.org/abs/2506.17064>
- [11] J. Weinreich and D. Probst, “Learning on compressed molecular representations,” *Digital Discovery*, vol. 4, no. 1, pp. 84–92, 2025, doi: [10.1039/d4dd00162a](https://doi.org/10.1039/d4dd00162a).