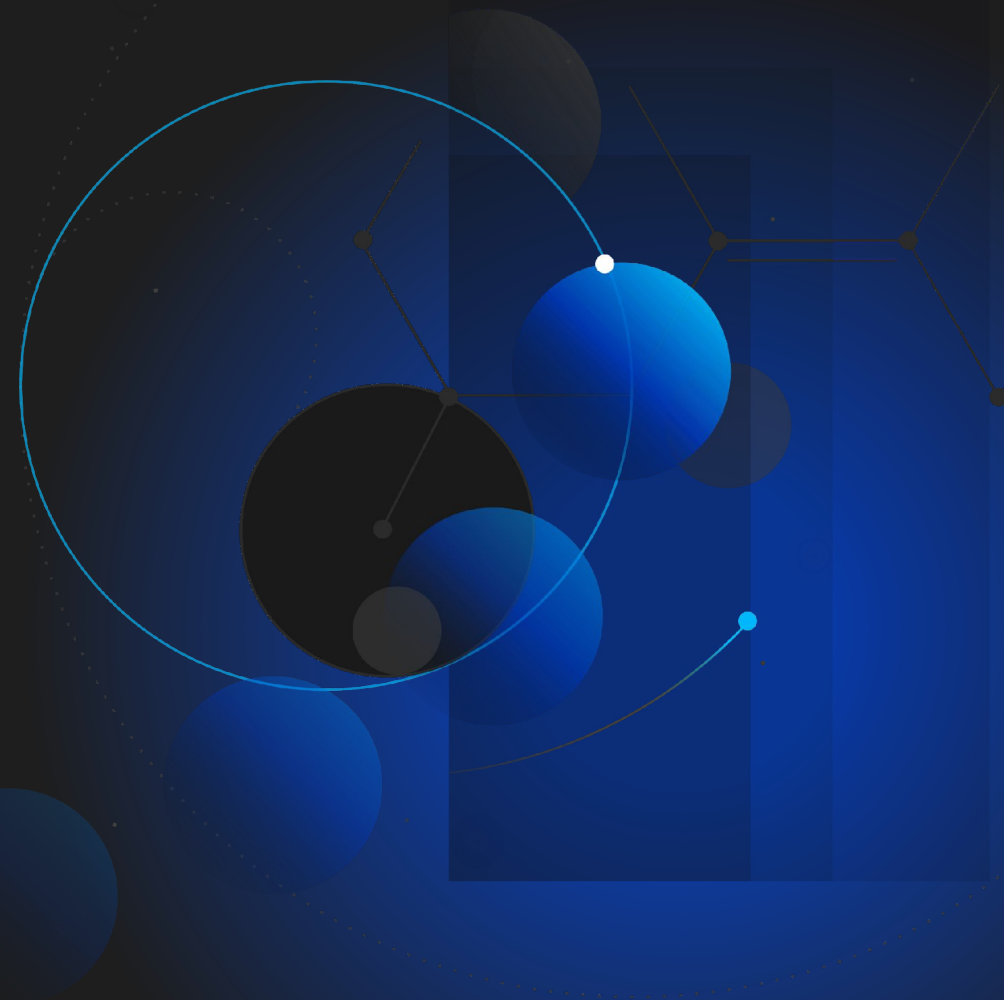


Valence

Bag of Features: A Multi-Instance
Learning Perspective on QSAR



Building tooling to support discovery programs from start to finish

Screening & Scoring

Structure-based

- Large-scale docking
- ML-enabled scoring
- Interaction profiling
- 3D-aware representations

Ligand-based

- Representation learning
- Uncertainty estimation & active learning
- Out-of-distribution generalization
- Few-shot, meta, & transfer learning

Multiparameter Optimization

Hit expansion & LO

- Large-scale molecular search and Bayesian optimization
- Ligand-based generative models
- Evolutionary algorithms for pareto optimization

Scaffold hopping

- Scaffold-invariant property optimization

Generative Design

Structure-based

- Structure-constrained 3D design
- Target-conditioned generation
- Fragment-based linkage
- Docking optimization

Ligand-based

- Synthetically-accessible reinforcement learning
- Retrosynthesis optimized fragment-based design
- Adversarial design
- Graph-based design

Library generation

- Rule-based molecular enumeration
- ML-augmented matched molecular pairs analysis

Integration & Collaboration

ReactR

- Molecule review platform
- Chemist in the loop active learning

Patentor

- Automated SAR extraction from patent data

Kernel

- File and data management
- Data visualization
- Design cycle management

Circus

- Rapid, large scale similarity search
- Retrosynthesis



How can we adapt AI models to localized chemistry in the absence of large datasets?

- Better uncertainty calibration !
- Few shot learning algorithm with better OOD generalization !
- **Better representation !**



Molecular representation for machine learning: **key to success**



“A method cannot save an unsuitable representation which cannot remedy irrelevant data for an ill thought-through question”

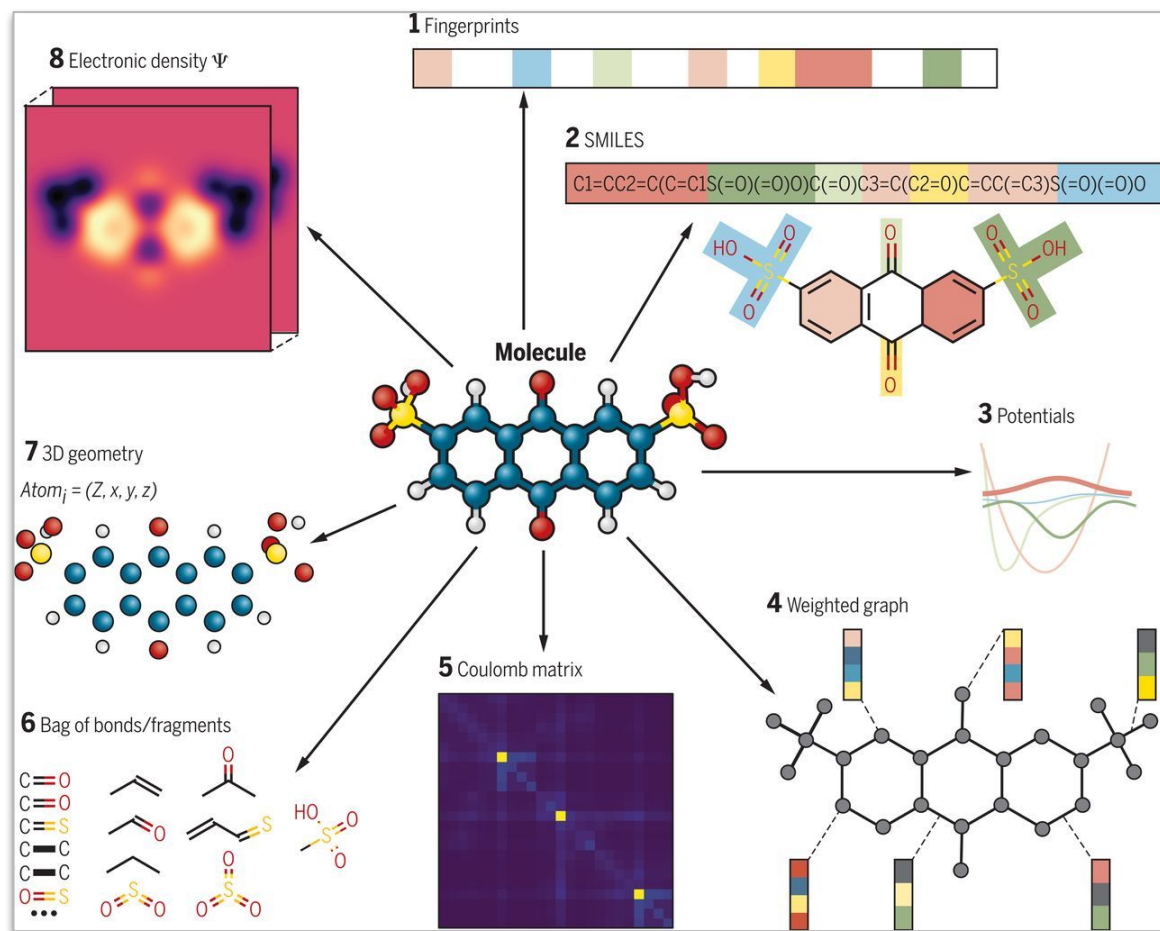
Bender & Cortes-Ciriano, 2021: doi/10.1016/j.drudis.2020.11.037

But what constitutes a “good” molecular representation (for Machine Learning) ?

- Information preservation
- Accuracy and robustness
- Compatibility with SOTA ML algorithms
- Ease of conversion
- **Relevance to the task**



Not all molecular representations are equal

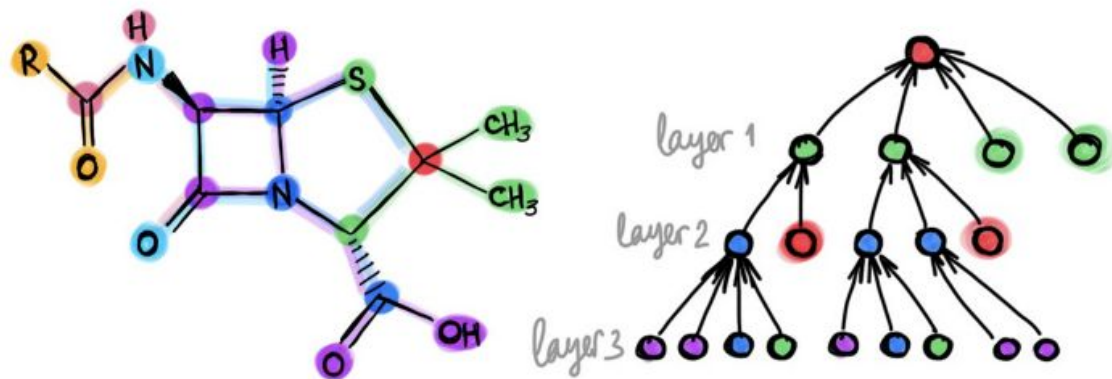


Sanchez-Lengeling and Aspuru-Guzik, 2018

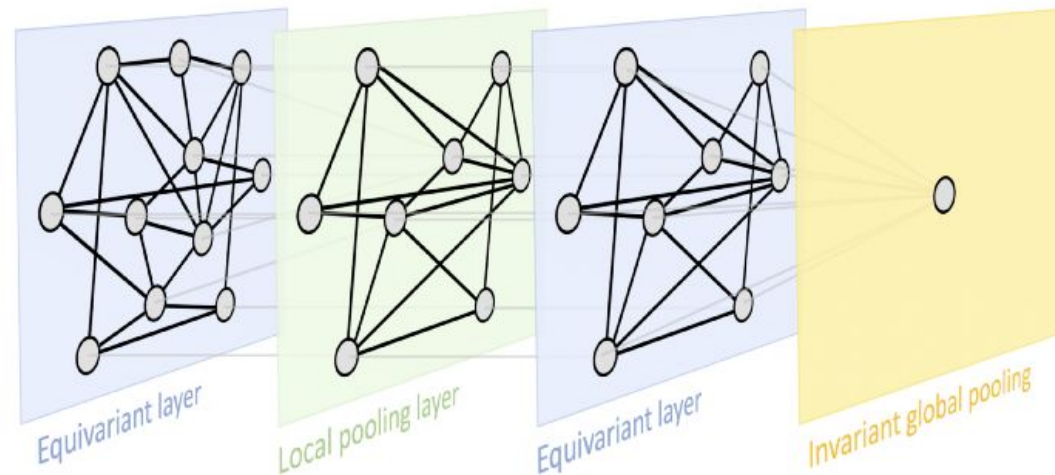
- Multiple ways to represent molecules for generative/predictive modelling
- Each with its own strength and limitation and learning algorithm that matches best
- Hand-crafting molecular features vs learning features for task specific/agnostic objectives



Geometric deep learning: promise vs reality



Original image: M. Bronstein Do we need deep graph neural networks?



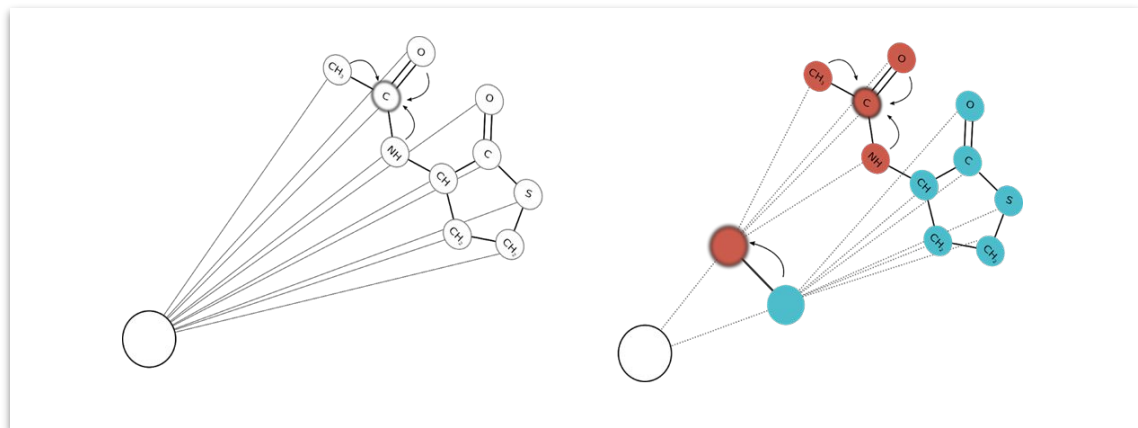
Original image: Geometric Deep Learning blueprint, M. Bronstein

- Extraction of relevant predictive features (from ligands and ligand-target complexes)
- Ability to learn more abstract features (larger receptive field from deeper architecture)

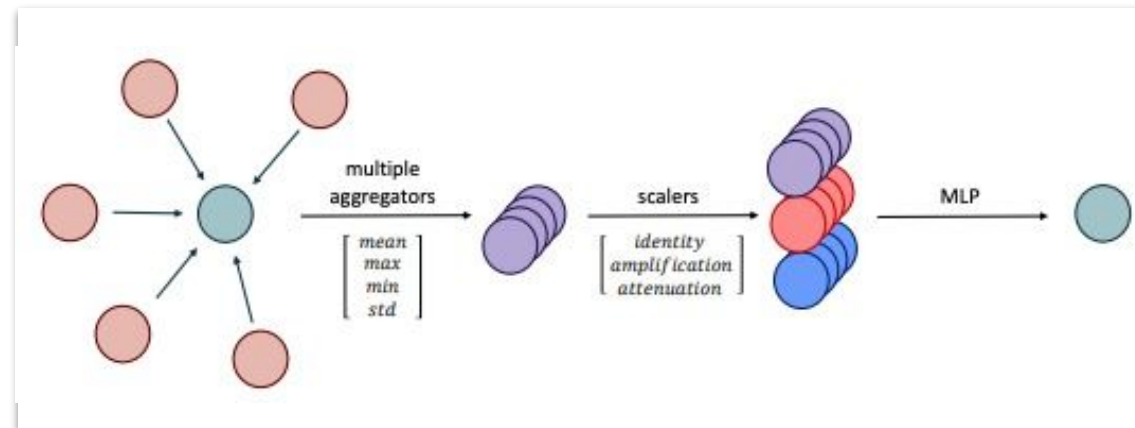
- Limited expressive power
- Oversmoothing
- Bottleneck (over squashing)
- No formal gain in low data regime



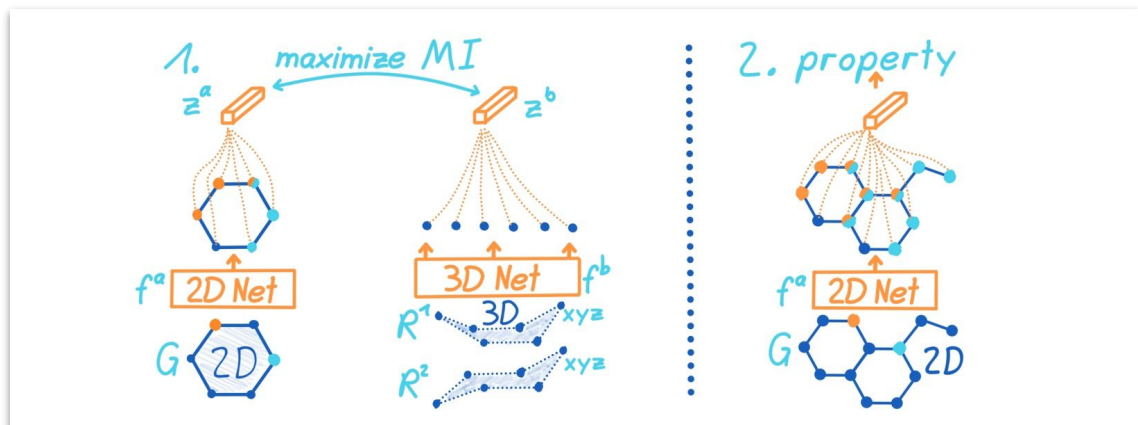
Recent work in GNN space to offset known limitations



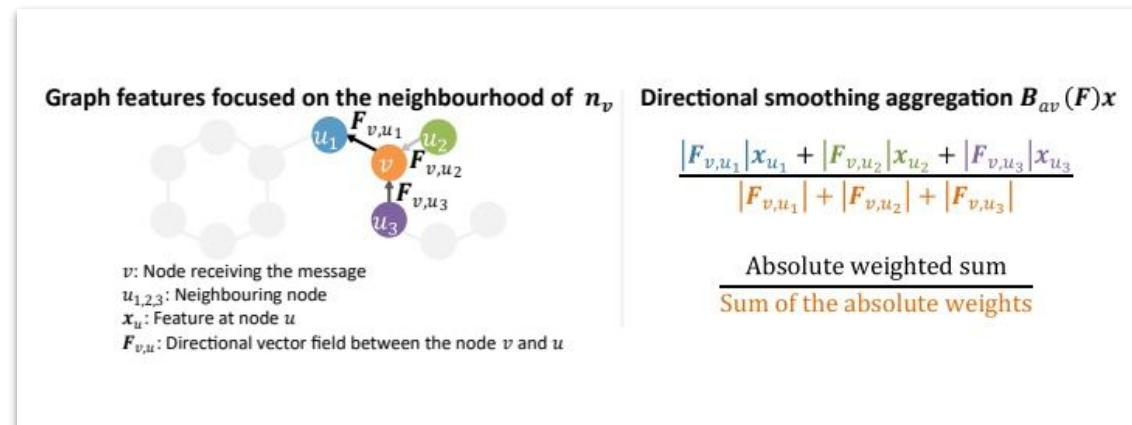
LaPool: Noutahi et al. 2019, arXiv



PNA: Corso et al., 2020, NeurIPS



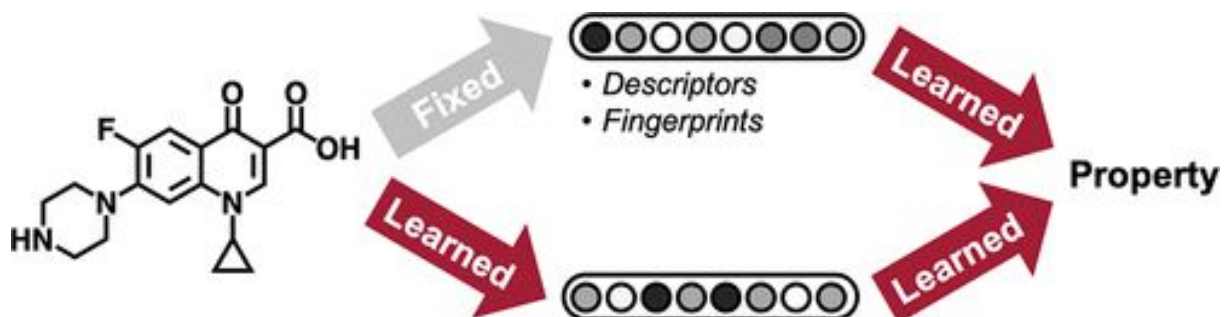
3D Pretraining, Stärk et al. 2021, under review



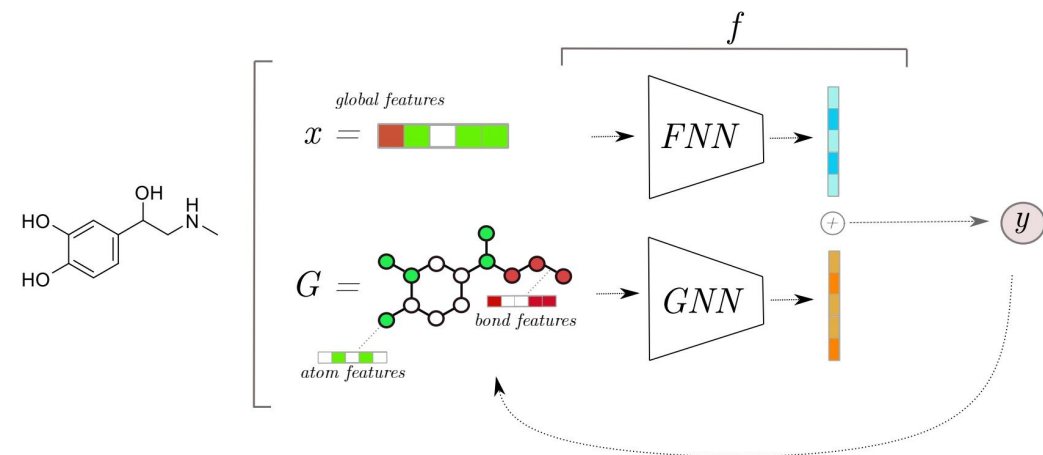
DGN: Beaini et al. 2021, ICML



GNN for QSPR in practice: supplementing learned features with fixed descriptors



ChemProp: Yang and al. 2019 doi/10.1021/acs.jcim.9b00237



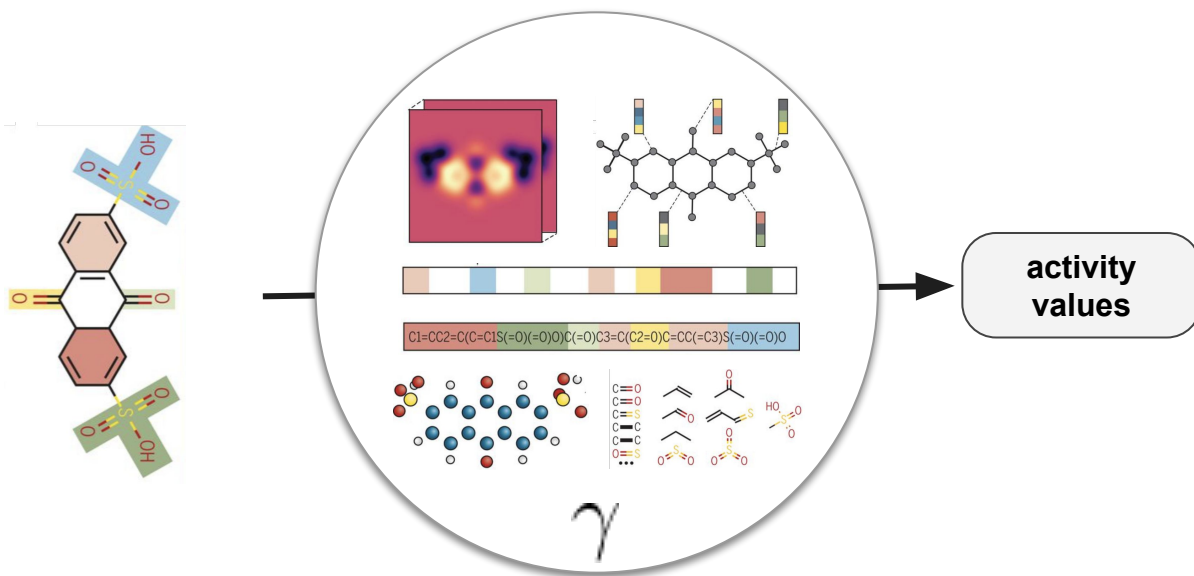
Jiménez-Luna et al. 2020 doi/10.26434/chemrxiv.13252286.v1

Incorporation of external information from computed features/descriptors to GNN can be very beneficial

Hypothesis: Using multiple molecular perspective would yield richer representation which in turn would enable improvement on QSPR tasks



Multi-instance learning to enable multiple molecular perspectives



There exists a perspective γ_i on the molecules of maximum information, w.r.t to the predictive task:

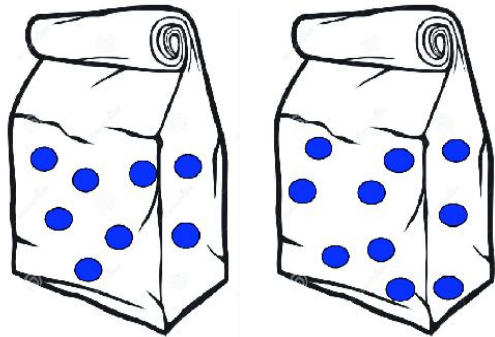
$$\left\{ \inf_{\gamma \in \Gamma^*} \Delta[f(\gamma(x), y)] \right\}$$

Should we search over all perspectives ?

How do we deal with increased feature space and avoid overfitting ?

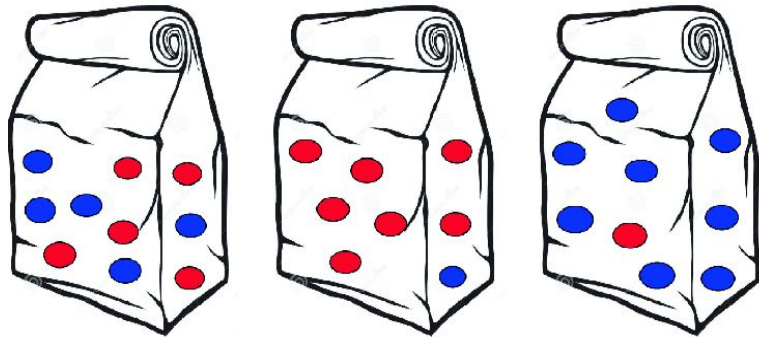


A quick introduction to multi-instance learning



Negative Bags

Label = 0



Positive Bags

Label = +1

- **Objective:** label a **bag of samples** instead of a **sample**
- Involves the notion of **set encoding** or **prediction aggregation** due to permutation invariance in sets

$$S(X) = g\left(\sum_{\mathbf{x} \in X} f(\mathbf{x})\right) \quad |S(X) - g(\max_{\mathbf{x} \in X} f(\mathbf{x}))| < \varepsilon$$

- Why not multi-view? Accommodates better to changing and missing “**instances**”
- **Should there be confidence in number?**

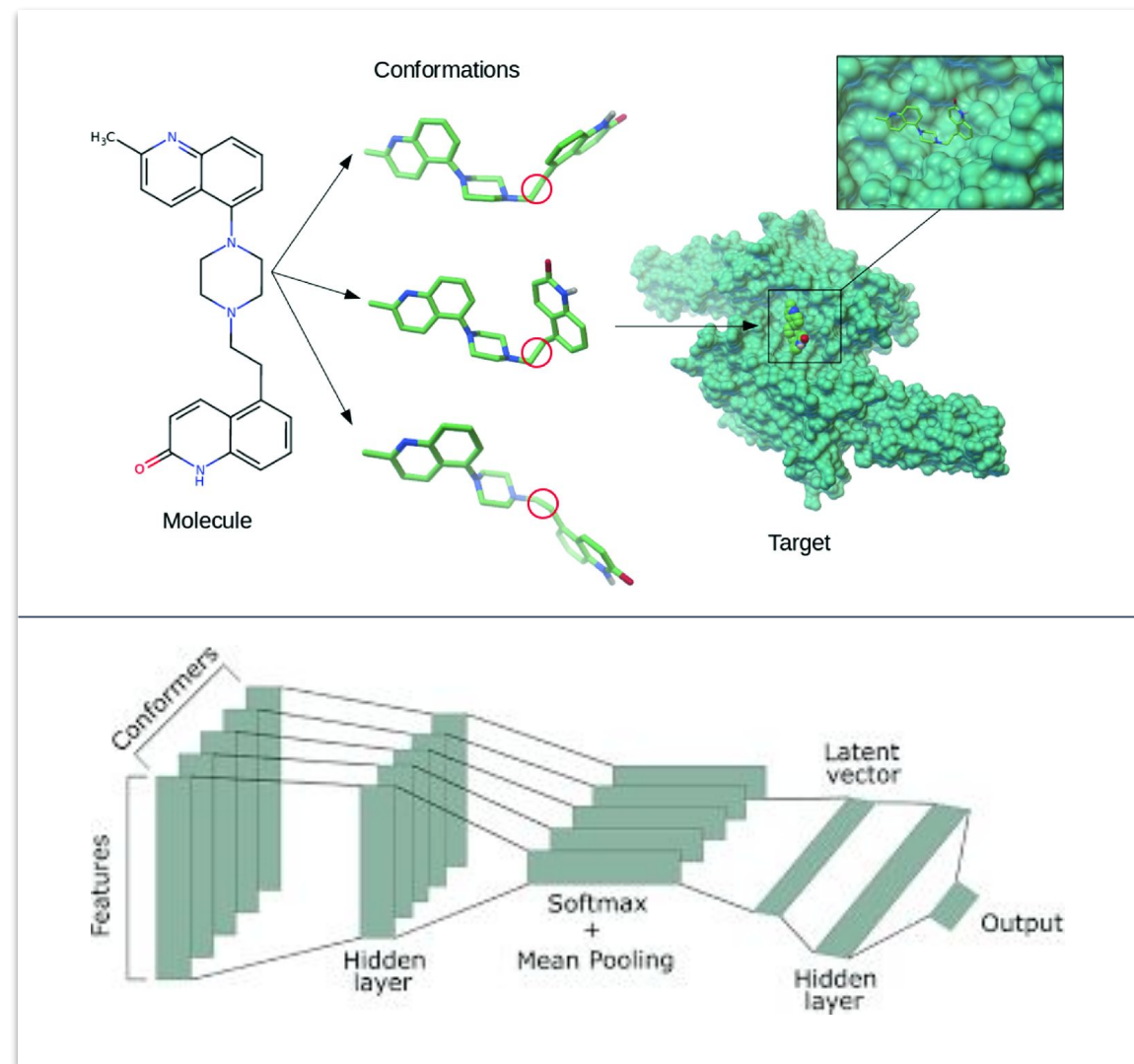


Sound familiar?

- **Objective:** label a **bag of molecular representation/state** instead of a **single state**
- Involves the notion of **set encoding** or **prediction aggregation** due to permutation invariance in sets

$$S(X) = g\left(\sum_{\mathbf{x} \in X} f(\mathbf{x})\right) \quad |S(X) - g(\max_{\mathbf{x} \in X} f(\mathbf{x}))| < \varepsilon$$

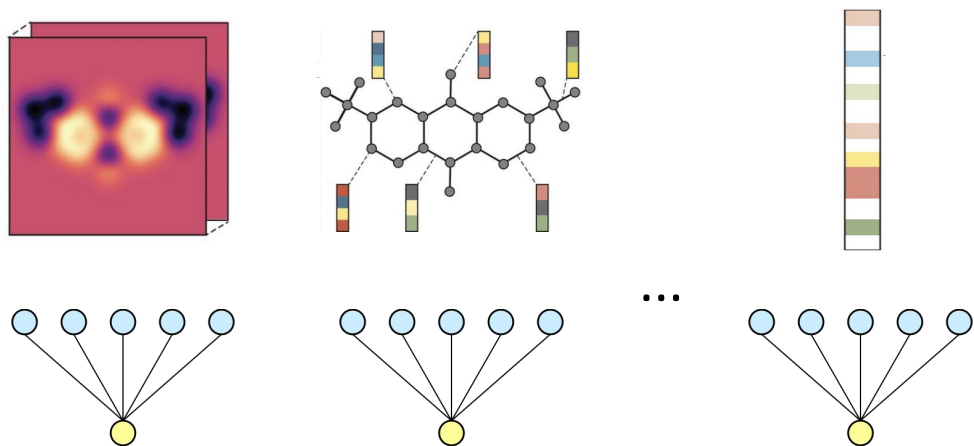
- Why not multi-view? Accommodates better to changing and missing “**molecular perspectives**”
- **Should there be confidence in number?**



Zankov et al. 2020 https://link.springer.com/chapter/10.1007/978-3-030-39575-9_7



Two paradigms for multi-instance learning

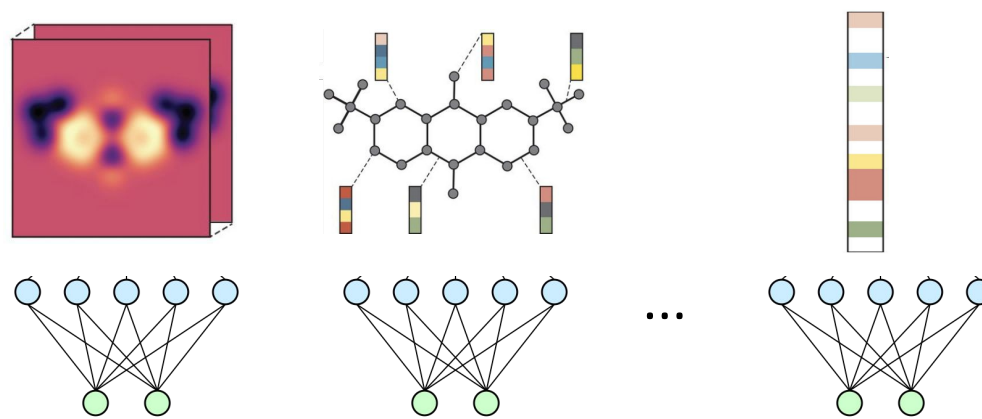


g

prediction aggregation

activity values

Instance-level approach (f is an instance level predictor)



Features 1

Features 2

Features N

Feature aggregation

g

activity values

Feature-level approach (g is a predictor on aggregation of instance embeddings)



Attention-based mechanism for aggregation

Mean	Max	LSE (LogSumExp)	Additive Attention	Gated-Attention	Set2Set
$\mathbf{z} = \frac{1}{K} \sum_{k=1}^K \mathbf{h}_k$	$z_m = \max_{k=1, \dots, K} \{\mathbf{h}_{km}\}$	$z = \log \left(\sum_k \exp(\mathbf{h}_k) \right)$	$\mathbf{z} = \sum_{k=1}^K a_k \mathbf{h}_k$ $a_k = \frac{\exp\{\mathbf{w}^\top \tanh(\mathbf{V}\mathbf{h}_k^\top)\}}{\sum_{j=1}^K \exp\{\mathbf{w}^\top \tanh(\mathbf{V}\mathbf{h}_j^\top)\}}$	$\mathbf{z} = \sum_{k=1}^K a_k \mathbf{h}_k$ $a_k = \frac{\exp\{\mathbf{w}^\top (\tanh(\mathbf{V}\mathbf{h}_k^\top) \odot \text{sigm}(\mathbf{U}\mathbf{h}_k^\top))\}}{\sum_{j=1}^K \exp\{\mathbf{w}^\top (\tanh(\mathbf{V}\mathbf{h}_j^\top) \odot \text{sigm}(\mathbf{U}\mathbf{h}_j^\top))\}}$	$\mathbf{z}_t = \text{LSTM}(\mathbf{z}_{t-1}^*)$ $\alpha_{i,t} = \text{softmax}(\mathbf{x}_i \cdot \mathbf{z}_t)$ $\mathbf{z}_t = \sum_{i=1}^N \alpha_{i,t} \mathbf{h}_i$ $\mathbf{z}_t^* = \mathbf{z}_t \parallel \mathbf{r}_t$


- Attention is trainable, flexible and adaptive
- A good choice for both instance and feature level aggregation
- Interpretable: which features are the most important for the task on a given molecule ?



Does this framework improve prediction on *relevant* QSAR endpoints ?

Setup

- 5 datasets from **TDC** (Therapeutics Data Commons)
- 5 Molecular Perspectives:
 - **MACCS keys, FCFP** radius 2, **2D Descriptors** (RDKit), **GCN, MPNN**
- Fixed HP search budget of 50 (including aggregator framework) using **optuna**
- Separate test set, single network, 5 splits train/valid

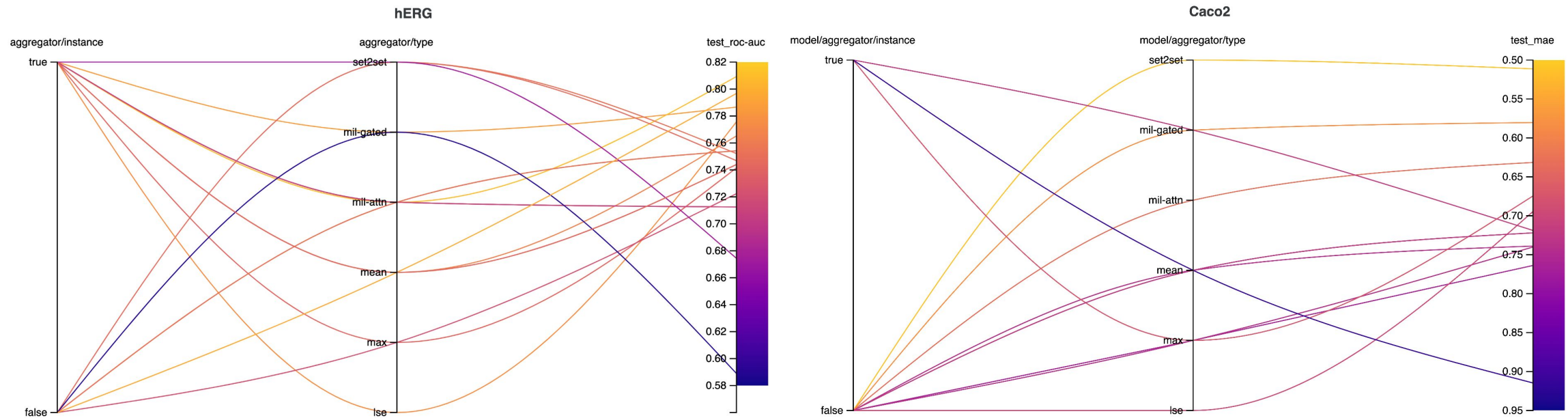
	Task	Metric	Size	TDC Baseline	Chembag
Classification	hERG (hERG blockers)	AUROC ↑	648	0.841 ± 0.020	0.853 ± 0.010
	BBB_Martins (Blood-Brain Barrier)	AUROC ↑	1,975	0.889 ± 0.016	0.902 ± 0.004
	AMES (Mutagenicity)	AUROC ↑	7,255	0.823 ± 0.011	0.859 ± 0.003
Regression	CACO2 (Permeability)	MAE ↓	906	0.393 ± 0.024	0.371 ± 0.029
	Solubility_AqSolDB (Solubility)	MAE ↓	9,982	0.827 ± 0.047	0.835 ± 0.021



Importance of aggregation framework and type

Setup

- Fix all (hyper) parameters except for aggregation function and type
- Compare performance on classification and regression tasks



- **Aggregation function is evidently important in measured performance.**
- **Feature wise aggregation performs better on average, especially for regression tasks**
- **Some combination clearly are hits or misses , but ATTENTION/SET2SET pooling consistent**



Can we transfer learned features on downstream tasks ?

Setup

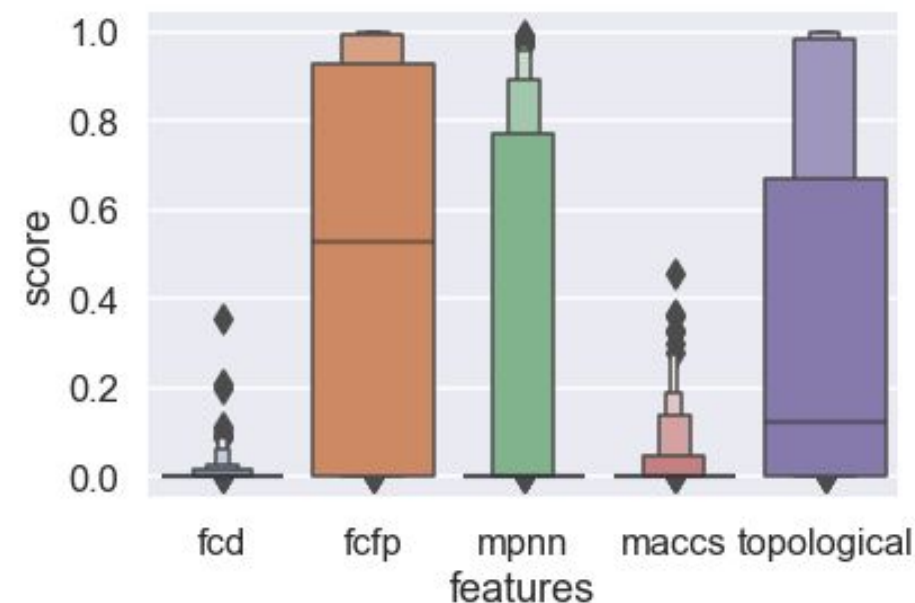
- Use a pretrained model (feature aggregator) on a larger dataset: PCBA (400K molecules, 128 tasks) ~ 40 epochs
- Can features extracted from that model be predictive of a new task? Using a simple machine learning model?

Task	Size	TDC Baseline	Chembag	Chembag Pretrained (PCBA)
hERG (hERG blockers)	648	0.841 ± 0.020	0.853 ± 0.010	0.835 ± 0.069
BBB_Martins (Blood-Brain Barrier)	1,975	0.889 ± 0.016	0.902 ± 0.004	0.892 ± 0.010
AMES (Mutagenicity)	7,255	0.823 ± 0.011	0.859 ± 0.003	0.823 ± 0.009



Conclusion

- Multi-instance framework has interesting applications in QSAR
- Very competitive across a wide range of tasks, with pretraining on large dataset an avenue to be explored further
- Attention weights can be used to rank features' contributions to predictive accuracy on a given task
- Formal characterization and further exploration of the framework still needed



“Model Interpretability on BBB prediction”



Valence

To learn more, please visit:
www.valencediscovery.com

