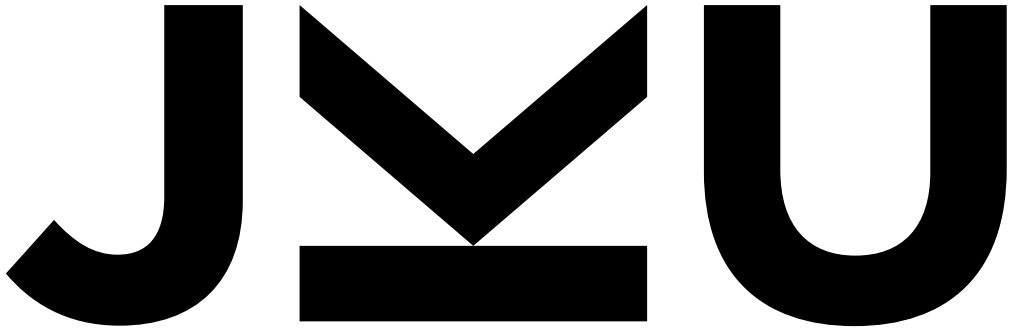


JOHANNES KEPLER
UNIVERSITY LINZ

Few- and zero-shot learning in drug discovery

Günter Klambauer
ELLIS Unit Linz & Institute for Machine Learning
https://ml-jku.github.io/
twitter: @gklambauer, slides available!

ellis unit LINZ

LIT AI LAB

AIDD: Advanced machine learning for Innovative Drug Discovery

D3Net

AI-SNN

madeSMART

Target Prediction

AI in Life Sciences Group @ LIT AI Lab:
Andreas Mayr, Philipp Renz, Theresa Roland, Elisabeth Rumetshofer, Ana Sanchez-Fernandez, Johannes Schimunek, Philipp Seidl, Florian Sestak, Emma Svensson, Andreu Vall

Overview

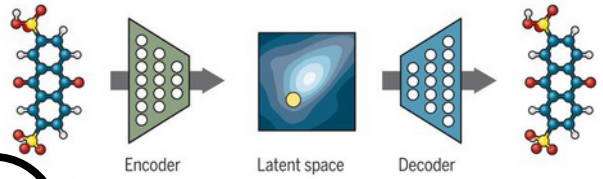
1. Introduction and motivation
2. Activity prediction and molecule encoders
   1. Drawbacks of current approaches
   2. Narrow AIs
   3. Multi-task deep networks
3. Zero- and few-shot learning
   1. Definition, problem setting
   2. Categories
4. Few-shot learning methods in drug discovery
   1. Data: FS-Mol
   2. Optimizer-based methods: fine-tuning, linear probing, MAML
   3. Embedding-based methods:
      1. Generalized framework
      2. Frequent hitters model
      3. Similarity search
      4. Neural similarity search
      5. IterRefLSTM
      6. ProtoNet
   4. Results
5. Zero-shot learning
   1. Proteo-chemometric models
   2. Text-based models
   3. Image-based models
6. Few- and zero-shot learning in other domains
   1. Chemical reactions
7. Summary

1. Main areas of Deep Learning in drug discovery

Activity and property prediction

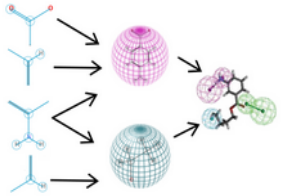
$$= f(\quad)$$

Molecule generation & optimization



$$\boldsymbol{x} = g_{\boldsymbol{\phi}}(\boldsymbol{y})$$

Molecular representations & molecular modeling

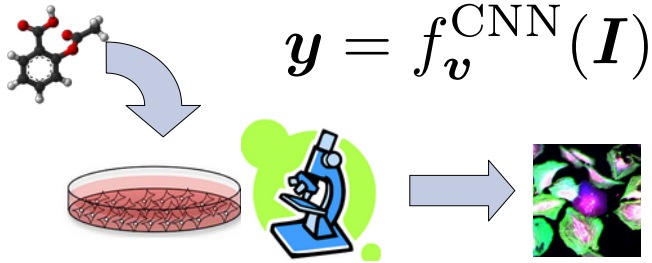


$$\boldsymbol{h_w}(\boldsymbol{x})$$

Reaction prediction & chemical synthesis planning

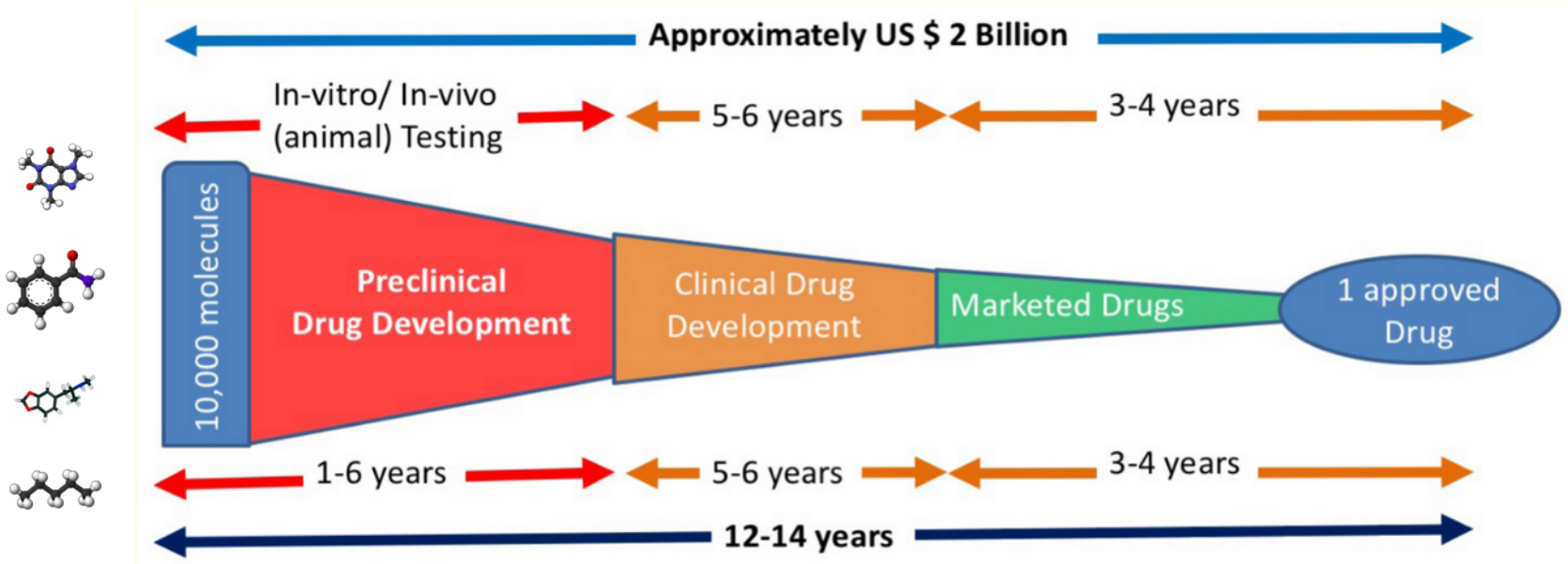
Image analysis

$$\boldsymbol{y} = f_{\boldsymbol{v}}^{\mathrm{CNN}}(\boldsymbol{I})$$



Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., & Blaschke, T. (2018). The rise of deep learning in drug discovery. Drug discovery today, 23(6), 1241-1250.

1. Computational methods to improve the time- and cost-intensive DD process

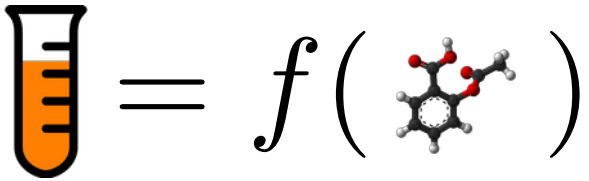


- Developing a new drug takes
  - **10-15 years**, up to **2 billion USD; many failures in late phases**
- Thus: more efficient ways are required; e.g. COVID-crisis
- All projects start with **few or zero data; late phases: few data**
  → **few-shot learning learning** is a promising technique to improve drug discovery

Mishra, M. (2018). An Opportunity in the Realm of Off-Target Effects. EC Pharmacology and Toxicology, 5, 134-140.

2. Activity/property prediction & QSAR

- Basic QSAR concept (Hansch, 1962): activity is a function of the structure of a molecule



$$y = f(m)$$

- ML methods model this with learned functions:

$$\hat{y} = g_{\boldsymbol{\theta}}(m)$$

Hansch, C., Maloney, P. P., Fujita, T., & Muir, R. M. (1962). Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. Nature, 194(4824), 178-180.

2. Classic activity prediction models

- Linear models or simple ML models

$$\hat{y} = g_{\boldsymbol{w}}(m) = \sigma(\boldsymbol{w}^T \boldsymbol{h}^{\mathrm{desc}}(m))$$

  - $\hat{y}$: predicted activity
  - m: representation of molecule
  - $\boldsymbol{h}^{\mathrm{desc}}(.)$: fixed *molecule encoder*; e.g. ECFP
  - $\boldsymbol{w}$: vector of adaptive parameters

- Other approaches: Random Forests, SVMs with molecule and graph kernels, naive Bayes, …

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. Journal of chemical information and computer sciences, 43(6), 1947-1958.
Ralaivola, L., Swamidass, S. J., Saigo, H., & Baldi, P. (2005). Graph kernels for chemical informatics. Neural networks, 18(8), 1093-1110.

2. Deep Learning based activity prediction models

- Deep neural networks

$$\hat{y} = g_{\boldsymbol{w},\boldsymbol{v}}(m) = \sigma(\boldsymbol{w}^T \boldsymbol{h}_{\boldsymbol{v}}(m))$$

  - $\boldsymbol{h}_{\boldsymbol{v}}(.)$: *adaptive* (learned) *molecule encoder* "molecular representation learning"
  - $\boldsymbol{v}$: adaptive parameters of molecule encoder

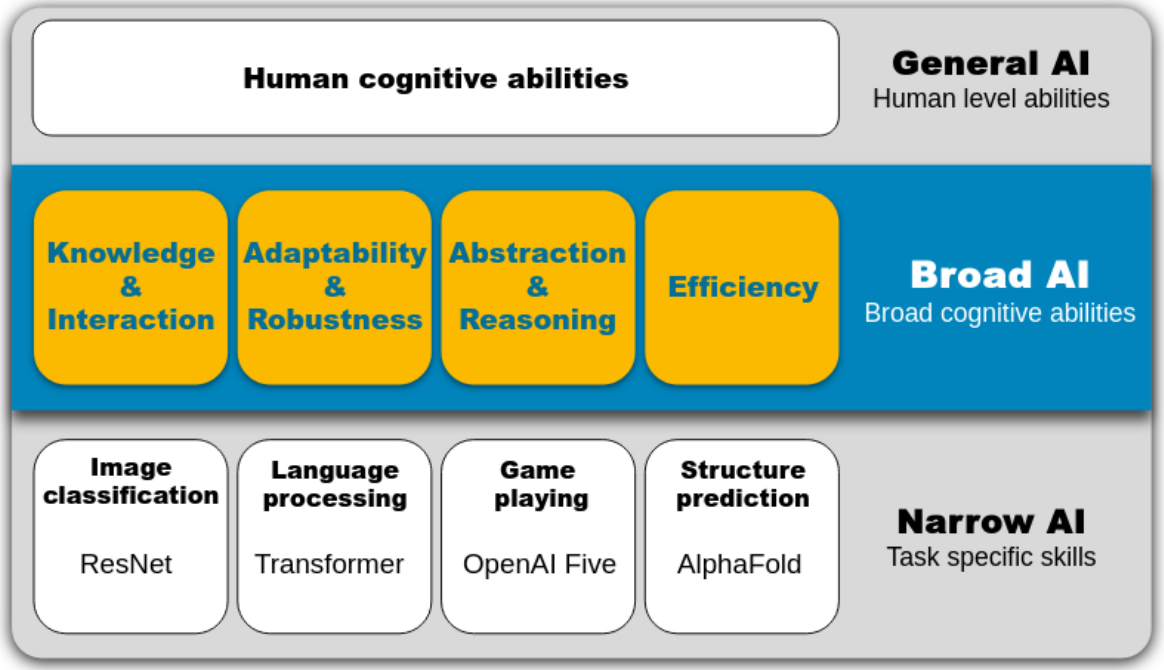
- Research focuses on structure of molecule encoders:
  DAG-RNN (Lusci, 2013); MT-DNN (Unterthiner, 2014; Dahl, 2014; Mayr, 2016), M-GConv (Kearnes, 2016), DGCNN (Zhang, 2018), GraphSage (Hamilton, 2017), ECC (Simonovsky, 2017), MPNN (Gilmer, 2017), SmilesLSTM (Mayr, 2018), GIN (Xu, 2018), ChemNet (Preuer, 2018), DiffPool (Ying, 2018), chemprop (Yang, 2019), MAT (Maziarka, 2020), CMPNN (Song, 2020), ChemBERTA (Chithrananda, 2020), Trans-CNN (van Deursen, 2020), **etc...**

2.1 Strengths and weaknesses of DL-based activity prediction

- Strengths:
  - Good predictive quality
- Weaknesses:
  - Lots of training data required
  - Re-training or fine-tuning required for new task
  - Not robust to domain shifts
  - Hardly uses prior knowledge
  - Difficult to interact with humans or other systems
  - …

Marcus, G. (2018). Deep learning: A critical appraisal. arXiv preprint arXiv:1801.00631.

2.2 Narrow AIs in drug discovery



- We are currently making progress on **adaptability** of our methods; a small step towards **Broad AI**

Chollet, F. (2019). On the measure of intelligence. arXiv preprint arXiv:1911.01547.
Hochreiter, S. (2022). Toward a broad AI. Communications of the ACM, 65(4), 56-57.

2.3 Multi-task deep networks

- Multi-task deep networks (MT-DNN) for activity/property prediction:

$$\hat{y} = g(m, \boldsymbol{t}) = \sigma(\boldsymbol{t}^T \boldsymbol{W} \boldsymbol{h_v}(m))$$

  one-hot description of task: $\boldsymbol{t}^T = (0, \ldots, 0, 1, 0, \ldots, 0)$

- Fundamental change: information can be carried over from one task to another
- Advantage: molecule encoder can be shared across tasks → "multi-task learning effect"
- Note: no similarities of tasks!

Mayr, A., Klambauer, G., Unterthiner, T., & Hochreiter, S. (2016). DeepTox: toxicity prediction using deep learning. Frontiers in Environmental Science, 3, 80.
Ramsundar, B., Kearnes, S., Riley, P., Webster, D., Konerding, D., & Pande, V. (2015). Massively multitask networks for drug discovery. arXiv preprint arXiv:1502.02072.

3. Zero- and few-shot learning: intuitive definitions

- 3.1 **Few-shot learning**: a new task represented by a small set $Z = \{X, y\}$ given; produce a predictive model for that task
  - Usually supervised
  - Similar tasks are available for training
- 3.2 Support set Z can be used in arbitrary way
  - **Data-based FSL**: Use prior information to improve data
    - augment data and train on Z
  - **Optimizer-based FSL**: prior information to constrain optimizer
    - Train a model from scratch on Z
    - Fine-tune model on Z
  - **Model-based FSL**: prior information to constrain model
    - Use Z as input for another model
- **Zero-shot learning** (ZSL): a description of the task $z \in \mathcal{Z}$ instead of a support set is given
  - E.g. textual description of task (“dog”, “cat”)

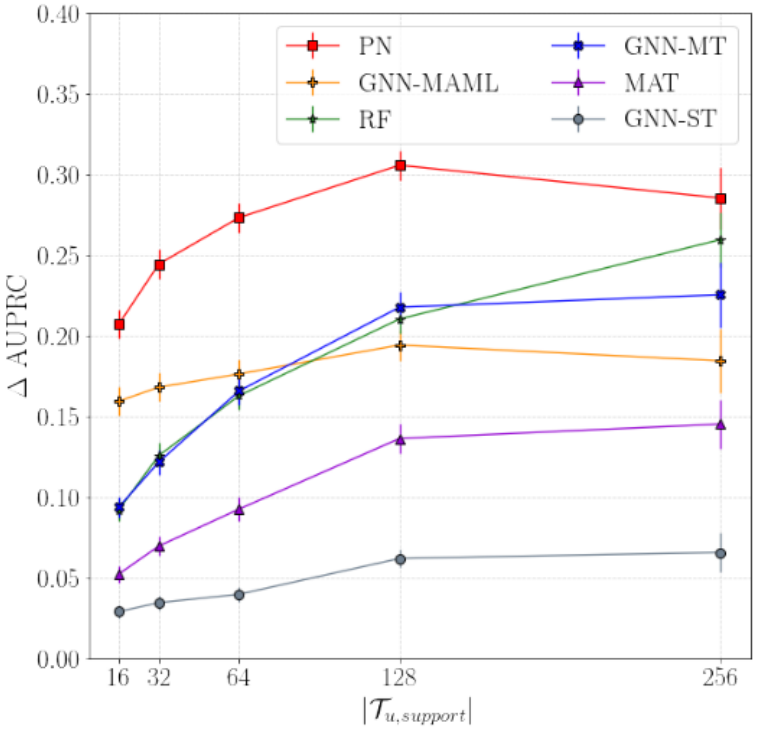
4.1 FS-Mol

	Datasets			
	ExCAPE-ML	PCBA	LSC	FS-Mol
# measurements	49,316,517	34,017,170	5,100,411	489,133
# compounds	955,386	437,929	449,391	233,786
# tasks	526	128	1310	5120
Mean # compounds / task	93,758	265,759	3872	94
Median # compounds / task	1820	309,562	320	46
Mean inactive:active / task	268:1	46:1	7:1	1:1
Raw values available?	Yes	No	No	Yes
Source	PubChem/ChEMBL	PubChem	ChEMBL18	ChEMBL27

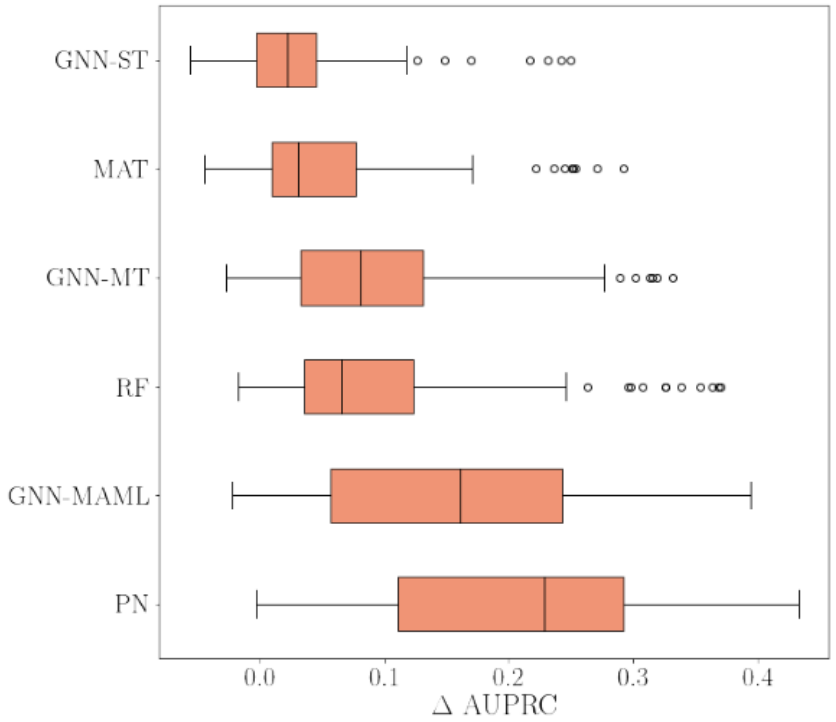
- 4938 training tasks, 40 validation task, 157 test tasks

Stanley, M., Bronskill, J. F., Maziarz, K., Misztela, H., Lanini, J., Segler, M., ... & Brockschmidt, M. (2021, August). Fs-mol: A few-shot learning dataset of molecules. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).

4.1 FS-Mol



(a) Mean performance on unseen tasks $\mathcal{T}_u$ as the support set size available for adaptation is increased. We include errors in the means for each point.



(b) Performance across all independent unseen tasks from $\mathcal{D}_{test}$, at support set size 16. The boxes represent the interquartile range across tasks, the extended lines are the (5, 95) percentiles and additional points represent outliers.

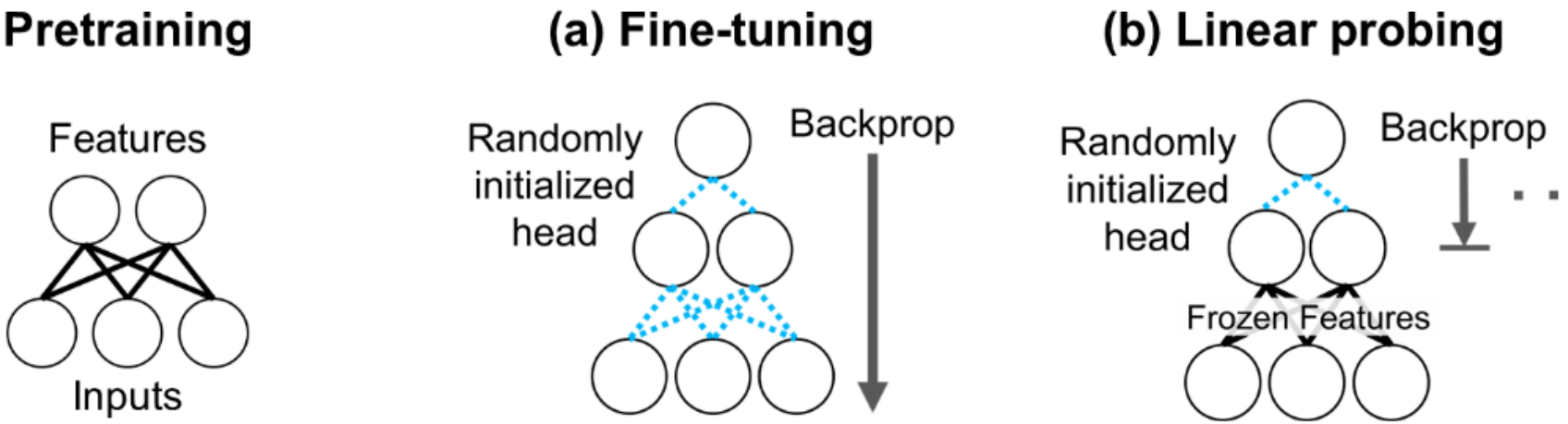
Stanley, M., Bronskill, J. F., Maziarz, K., Misztela, H., Lanini, J., Segler, M., ... & Brockschmidt, M. (2021, August). Fs-mol: A few-shot learning dataset of molecules. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).

4.1 FS-Mol

- Methods compared in FS-Mol
  - **RF**: standard training (optimizer-based)
  - **GNN-ST**: fine-tuning (optimizer-based)
  - **GNN-MT**: linear probing (optimizer-based)
  - **GNN-MAML**: model-agnostic meta-learning (optimizer-based)
  - **MAT**: self-supervised pre-training plus fine-tuning (optimizer-based)
  - **ProtoNet**: learned embeddings yield prototypes for each class (embedding-based)

Stanley, M., Bronskill, J. F., Maziarz, K., Misztela, H., Lanini, J., Segler, M., ... & Brockschmidt, M. (2021, August). Fs-mol: A few-shot learning dataset of molecules. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).

4.2 Pre-training, fine-tuning, and linear probing

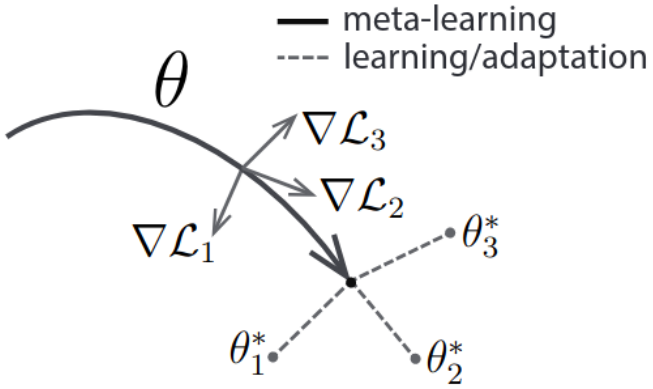


Alain, G., & Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. arXiv preprint arXiv:1610.01644.

Kumar, A., Raghunathan, A., Jones, R., Ma, T., & Liang, P. (2022). Fine-tuning can distort pretrained features and underperform out-of-distribution. arXiv preprint ICLR2022

4.2 Model-agnostic meta-learning (MAML)

- Learns a good starting point for fine-tuning
  - Learns an initialization
- Intuitively: look ahead
  - Which starting parameters would have led to low loss at sampled few-shot tasks



Algorithm 1 Model-Agnostic Meta-Learning

Require: $p(\mathcal{T})$: distribution over tasks
Require: α, β: step size hyperparameters
1: randomly initialize θ
2: **while** not done **do**
3: Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$
4: **for all** $\mathcal{T}_i$ **do**
5: Evaluate $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$ with respect to K examples
6: Compute adapted parameters with gradient descent: $\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$
7: **end for**
8: Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'})$
9: **end while**

Finn, C., Abbeel, P., & Levine, S. (2017, July). Model-agnostic meta-learning for fast adaptation of deep networks. In International conference on machine learning (pp. 1126-1135). PMLR.

4.3 Embedding-based few-shot learning methods

- Embedding-based few-shot learning methods:

$$\hat{y} = g_{\boldsymbol{\theta}}(m, \boldsymbol{Z})$$

  - $\boldsymbol{Z} = \{\boldsymbol{X}, \boldsymbol{y}\}$: support set of molecules, a "description" of the task

Altae-Tran, H., Ramsundar, B., Pappu, A. S., & Pande, V. (2017). Low data drug discovery with one-shot learning. ACS central science, 3(4), 283-293.
Stanley, M., Bronskill, J. F., Maziarz, K., Misztela, H., Lanini, J., Segler, M., ... & Brockschmidt, M. (2021, August). FS-Mol: A Few-Shot Learning Dataset of Molecules. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).

4.3.1 Generalized framework for embedding-based FSL methods

- Drug-target association networks: generalized framework

$$\hat{y} = g_{\boldsymbol{w}}(m, \{\boldsymbol{X}, \boldsymbol{y}\})$$

$$\hat{y} = \boldsymbol{h}^{\text{assoc}}\left(\boldsymbol{h}^{\text{mol}}(\boldsymbol{m}), \boldsymbol{h}^{\text{set}}(\{\boldsymbol{X}, \boldsymbol{y}\})\right)$$

  - molecule, memory and association encoder
  - Any “DeepSets”-like methods as encoders
  - Simple forms implemented and tested

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., & Smola, A. J. (2017). Deep sets. Advances in neural information processing systems, 30.

Schimunek, J., Friedrich, L., Kuhn, D., Rippmann, F., Hochreiter, S., & Klambauer, G. (2021). A generalized framework for embedding-based few-shot learning methods in drug discovery. ELLIS Machine Learning for Molecules workshop.

4.3.2 A naive baseline: The frequent-hitters (FH) model

- A method that does not consider the support set:

$$\hat{y} = g_{\boldsymbol{w}}(m, \{\boldsymbol{X}, \boldsymbol{y}\}) = g_{\boldsymbol{w}}^{\mathrm{FH}}(m)$$

- Learns general activity
  - **Frequent hitters**: molecules that are active in many tasks or for many targets
  - **Dark matter**: molecules that are almost always inactive or do not interact with targets
- During training: has to predict both “active” and “inactive” for the same molecule → average activity is learned

Schimunek, J., Friedrich, L., Kuhn, D., Rippmann, F., Hochreiter, S., & Klambauer, G. (2021). A generalized framework for embedding-based few-shot learning methods in drug discovery. ELLIS Machine Learning for Molecules workshop.

4.3.3 Classic similarity search as used in cheminformatics

- Traditional similarity search

$$\hat{y} = g(m, \boldsymbol{A}) = \frac{1}{N} \sum_{n=1}^{N} k(\boldsymbol{h}^{\mathrm{fp}}(m), \boldsymbol{h}^{\mathrm{fp}}(a_n)),$$

  - $\boldsymbol{h}^{\mathrm{fp}}$: some fixed molecule encoder (fingerprints, descriptors)
  - $k(.,.)$: some similarity measure (Tanimoto, MinMax similarity)
  - $\boldsymbol{A}$: active molecules; support set

Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Garcia-Vallvé, S., & Pujadas, G. (2015). Molecular fingerprint similarity search in virtual screening. Methods, 71, 58-63.

4.3.4 Neural variant of similarity search

- Neural similarity search

$$\hat{y} = g(m, \boldsymbol{A}) = \frac{1}{N} \sum_{n=1}^{N} k(\boldsymbol{h_w}(m), \boldsymbol{h_v^a}(a_n))$$

  - $\boldsymbol{h}$: some adaptive (learned) molecule encoders
  - $k(.,.)$: some similarity measure (Tanimoto, MinMax similarity); must be differentiable
  - $\boldsymbol{A}$: active molecules; support set; Can be extended to using also negatives!

4.3.4 Neural variant of similarity search

- A variant of eural similarity search

$$\hat{y} = \sigma\left(\tau^{-1}\frac{1}{N}\sum_{n=1}^{N} y_n \; \boldsymbol{h_w}(\boldsymbol{m})^T \; \boldsymbol{h_w}(\boldsymbol{x}_n)\right),$$

  - $\boldsymbol{h}$: some adaptive (learned) molecule encoders; returns normalized embeddings
  - τ^{-1} : scaling hyperparameter

Koch, G., Zemel, R., & Salakhutdinov, R. (2015, July). Siamese neural networks for one-shot image recognition. In ICML deep learning workshop (Vol. 2, p. 0).

4.3.5 IterRefLSTM

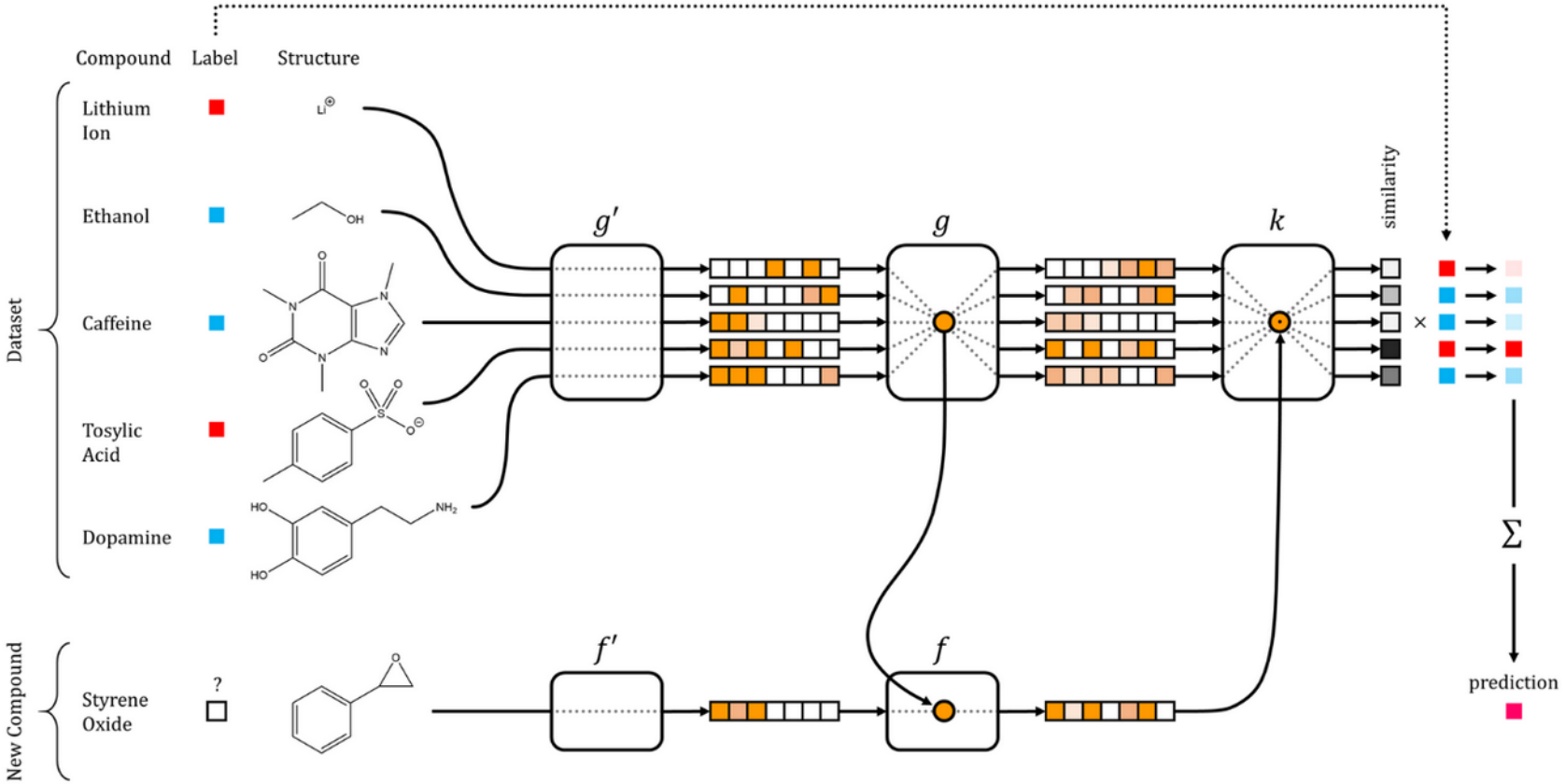
- Altae-Tran's few-shot method

$$\hat{y} = g_{\boldsymbol{w}}(m, \{\boldsymbol{X}, \boldsymbol{y}\})$$
$$(\boldsymbol{m}', \boldsymbol{H}) = \mathrm{peLSTM}(\{\boldsymbol{m}, \boldsymbol{X}\})$$
$$\hat{y} = \boldsymbol{y}^T \ \mathrm{softmax}(\boldsymbol{H}\boldsymbol{m}')$$

  - $(\boldsymbol{X}, \boldsymbol{y})$: Support set
  - $\mathrm{peLSTM}(.)$: permutation-equivariant LSTM

Altae-Tran, H., Ramsundar, B., Pappu, A. S., & Pande, V. (2017). Low data drug discovery with one-shot learning. ACS central science, 3(4), 283-293.

4.3.5 IterRefLSTM



4.3.5 IterRefLSTM

Table 1. ROC-AUC Scores of Models on Median Held-out Task for Each Model on Tox21[a]

Tox21	RF (100 trees)	Graph Conv	Siamese	AttnLSTM	IterRefLSTM
10+/10−	0.586 ± 0.056	0.648 ± 0.029	0.820 ± 0.003	0.801 ± 0.001	**0.823 ± 0.002**
5+/10−	0.573 ± 0.060	0.637 ± 0.061	0.823 ± 0.004	0.753 ± 0.173	**0.830 ± 0.001**
1+/10−	0.551 ± 0.067	0.541 ± 0.093	**0.726 ± 0.173**	0.549 ± 0.088	0.724 ± 0.008
1+/5−	0.559 ± 0.063	0.595 ± 0.086	0.687 ± 0.210	0.593 ± 0.153	**0.795 ± 0.005**
1+/1−	0.535 ± 0.056	0.589 ± 0.068	0.657 ± 0.222	0.507 ± 0.079	**0.827 ± 0.001**

[a]Numbers reported are means and standard deviations. Randomness is over the choice of support set; experiment is repeated with 20 support sets. The Appendix contains results for all held-out Tox21 tasks. The result with highest mean in each row is highlighted. The notation 10+/10− indicates supports with 10 positive examples and 10 negative examples.

- Good performance on held-out tasks on Tox21
- **However:** No transfer to new domains

Table 4. ROC-AUC Scores of Models Trained on Tox21 on Median SIDER Task for Each Model on SIDER[a]

SIDER from Tox21	Siamese	AttnLSTM	IterRefLSTM
10+/10−	0.511 ± 0.031	0.509 ± 0.014	0.509 ± 0.012

[a]Note that models are evaluated on all SIDER tasks and not just the held-out SIDER tasks from previous section. Numbers reported are means and standard deviations. Randomness is over the choice of support set; experiment is repeated with 20 support sets. The result with highest mean in each row is highlighted. The notation 10+/10- indicates supports with 10 positive examples and 10 negative examples.

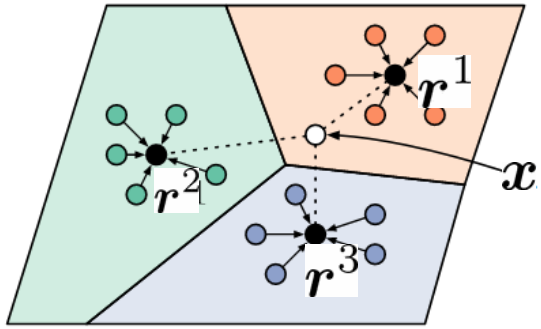
4.3.6 Prototypical networks

- Calculate embedding of prototype for each class

$$h^{\text{set}} : Z \mapsto (r^+, r^-)$$

$$r^+ = \frac{1}{|Z^+|} \cdot \sum_{(x,y)\in Z^+} h_w(x)$$

$$r^- = \frac{1}{|Z^-|} \cdot \sum_{(x,y)\in Z^-} h_w(x),$$



(a) Few-shot

- Prediction via associating query molecule with each prototype

$$h^{\text{assoc}} : \left(h_w(m), r^+, r^-\right) \mapsto \hat{y} \in \mathbb{R}$$

$$\hat{y} = \frac{\exp(-d(h_w(m), r^+))}{\exp(-d(h_w(m), r^+)) + \exp(-d(h_w(m), r^-))},$$

Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. Advances in neural information processing systems, 30.

4.4 Methods compared

Table 1: Results on FS-MOL [ΔAUC-PR]. The best method is marked bold. Error bars represent standard errors across tasks according to Stanley et al. (2021).

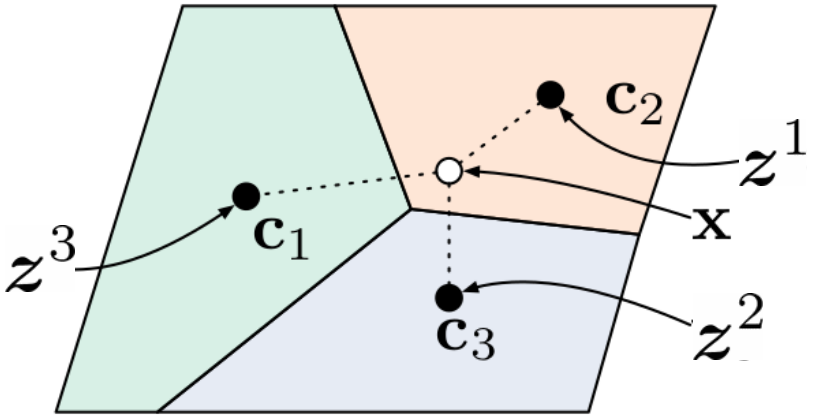
Method	All	Kinases	Hydrolases	Oxidored
GNN-ST[a] (Stanley et al., 2021)	.029 ± .004	.027 ± .004	.040 ± .018	.020 ± .016
MAT[a] (Maziarka et al., 2020)	.052 ± .005	.043 ± .005	.095 ± .019	.062 ± .024
Random Forest[a] (Breiman, 2001)	.092 ± .007	.081 ± .009	.158 ± .028	.080 ± .029
GNN-MT[a] (Stanley et al., 2021)	.093 ± .006	.093 ± .006	.108 ± .025	.053 ± .018
Similarity Search[b]	.118 ± .011	.113 ± .008	.117 ± .009	.157 ± .012
GNN-MAML[a] (Finn et al., 2017)	.159 ± .009	.177 ± .009	.105 ± .024	.054 ± .028
Frequent hitters (this work)	.198 ± .010	.220 ± .009	.136 ± .011	.064 ± .003
ProtoNet[a] (Snell et al., 2017)	.207 ± .008	.215 ± .009	.209 ± .030	.095 ± .029
Neural Sim Search (Schimunek et al., 2021)	**.226** ± .010	**.222** ± .010	**.230** ± .010	**.213** ± .013

[a] metrics from Stanley et al. (2021).
[b] metrics from Schimunek et al. (2021)

- Frequent hitters model outperforms almost all other methods
- Only embedding-based methods reach better performance than this baseline

Schimunek, J., Friedrich, L., Kuhn, D., Rippmann, F., Hochreiter, S., & Klambauer, G. (2021). A generalized framework for embedding-based few-shot learning methods in drug discovery. ELLIS Machine Learning for Molecules workshop.

5. Zero-shot learning



(b) Zero-shot

- Zero-shot learning:
  - Producing a model without any training data
  - Only a description of the task (or class) is available
  - Drug discovery, e.g.:
    - Description of drug target (protein)
    - Description of activity (bioassay)

5.1 Zero-shot learning via proteo-chemometrics

- Proteo-chemometric methods

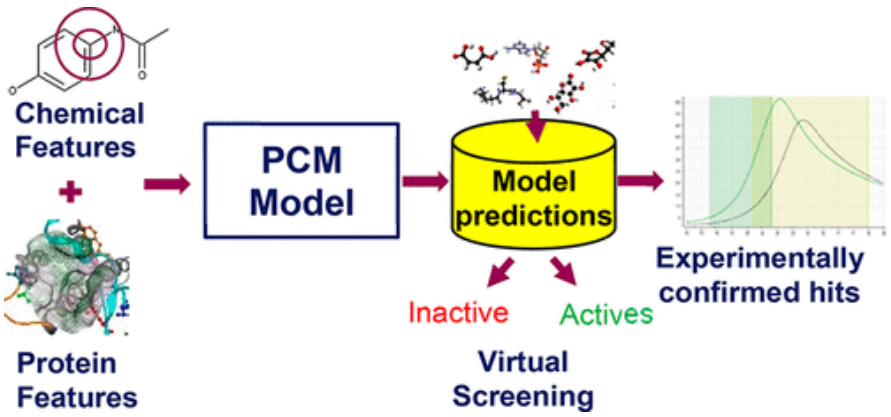
$$\hat{y} = g_{\boldsymbol{\theta}}(m, s)$$

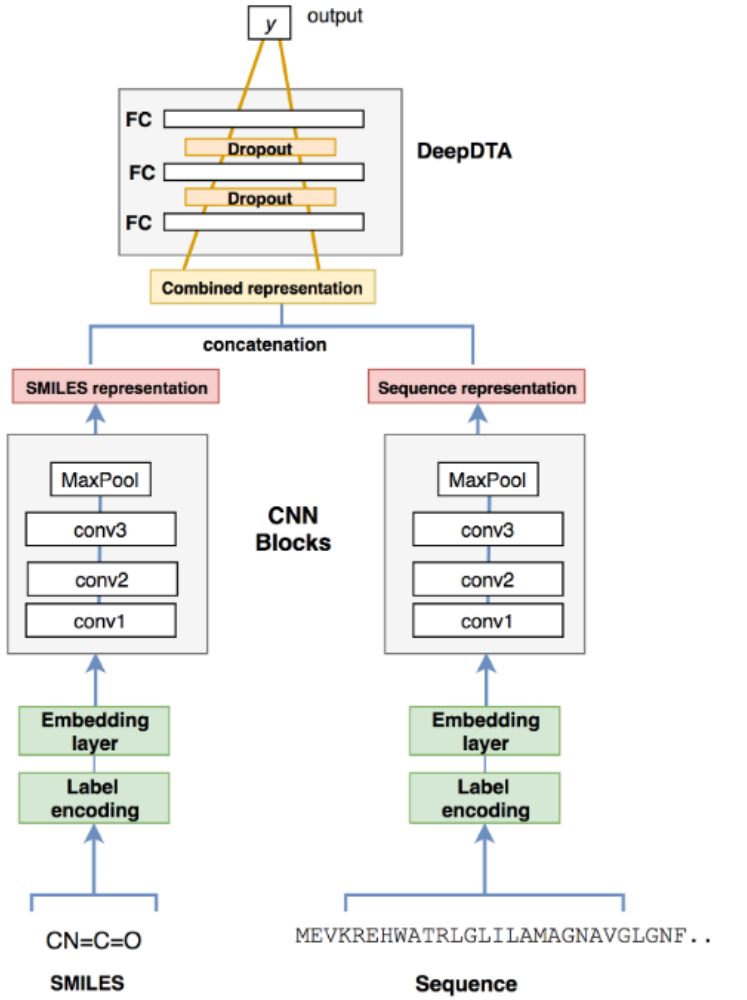
  - s : representation of the protein target
  - $g_{\boldsymbol{\theta}}(.,.)$: neural network
- Allows for making zero-shot predictions, i.e. for new proteins

Lapinsh, M., Prusis, P., Gutcaits, A., Lundstedt, T., & Wikberg, J. E. (2001). Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions. Biochimica et Biophysica Acta (BBA)-General Subjects, 1525(1-2), 180-190.

Lenselink, E. B., Ten Dijke, N., Bongers, B., Papadatos, G., Van Vlijmen, H. W., Kowalczyk, W., ... & Van Westen, G. J. (2017). Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. Journal of cheminformatics, 9(1), 1-14.

5.1 Proteo-chemometric



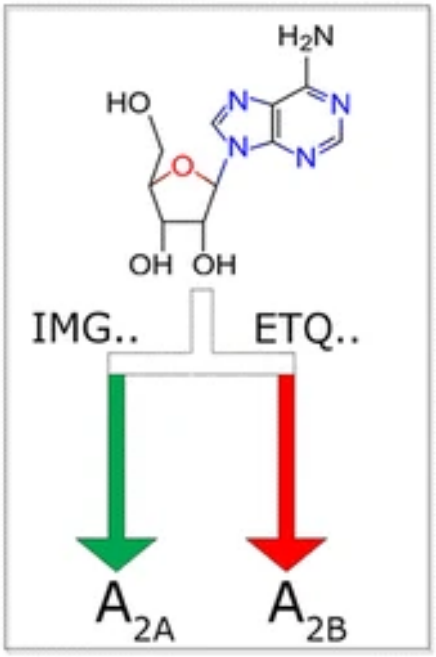


Giblin, K. A., Hughes, S. J., Boyd, H., Hansson, P., & Bender, A. (2018). Prospectively validated proteochemometric models for the prediction of small-molecule binding to bromodomain proteins. Journal of Chemical Information and Modeling, 58(9), 1870-1888.

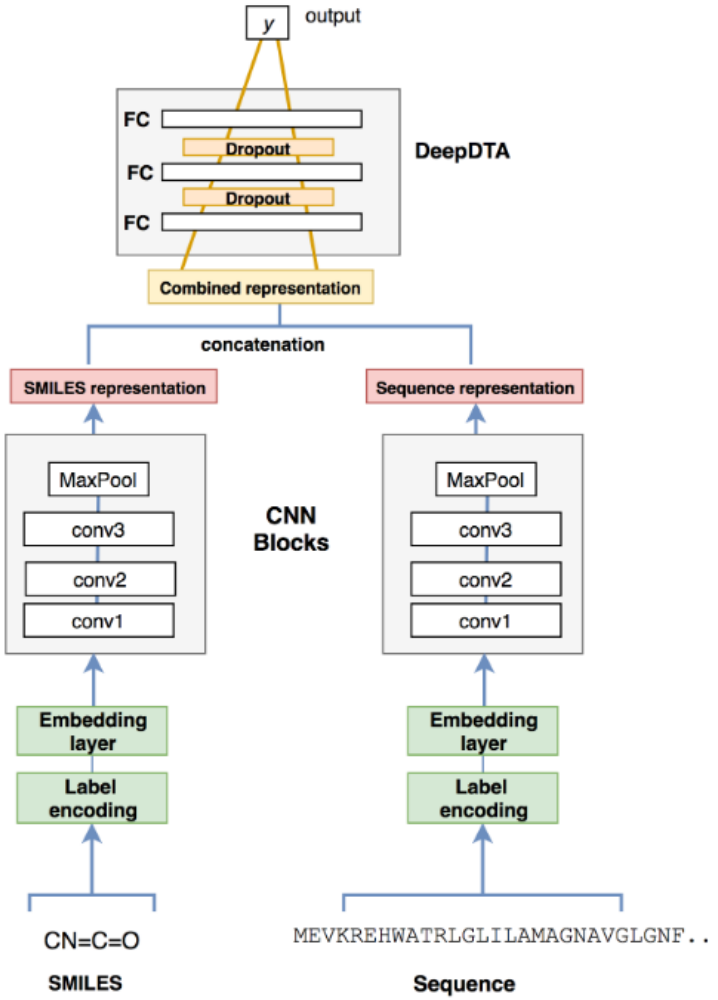
Lenselink, E. B., Ten Dijke, N., Bongers, B., Papadatos, G., Van Vlijmen, H. W., Kowalczyk, W., ... & Van Westen, G. J. (2017). Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. Journal of cheminformatics, 9(1), 1-14.

Öztürk, H., Özgür, A., & Ozkirimli, E. (2018). DeepDTA: deep drug–target binding affinity prediction. Bioinformatics, 34(17), i821-i829.

5.1 Proteo-chemometric



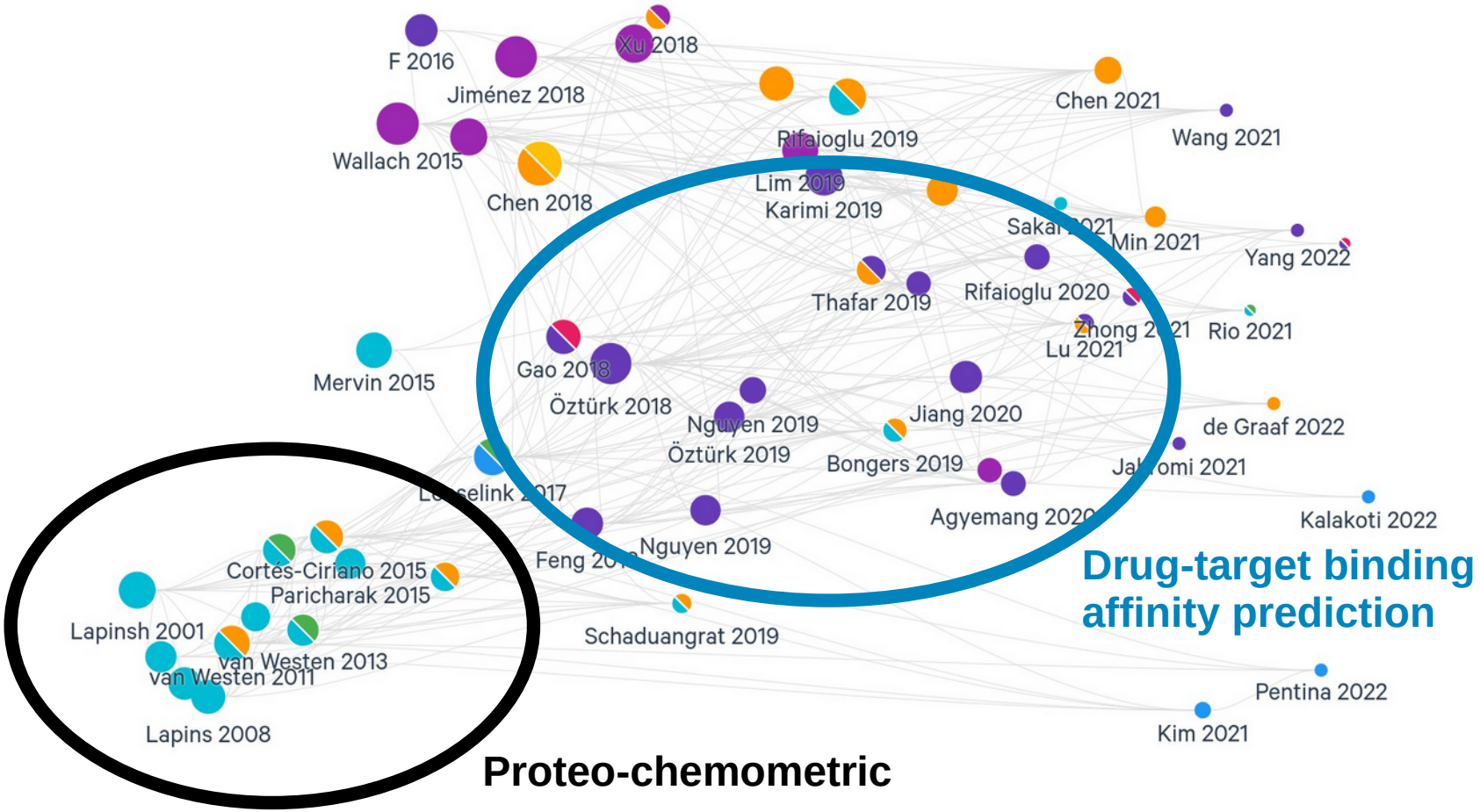
"For the DNN_PCM, we found that for targets with few data points in the training set, the PCM models were able to extrapolate predictions"



Lenselink, E. B., Ten Dijke, N., Bongers, B., Papadatos, G., Van Vlijmen, H. W., Kowalczyk, W., ... & Van Westen, G. J. (2017). Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. Journal of cheminformatics, 9(1), 1-14.

Öztürk, H., Özgür, A., & Ozkirimli, E. (2018). DeepDTA: deep drug–target binding affinity prediction. Bioinformatics, 34(17), i821-i829.

5.1 Proteo-chemometric



https://app.litmaps.co/shared/workspace/DF7EF612-299D-4ADD-8CE3-034CE99B772E
by Emma Svensson, 2022.

5.2 Zero-shot learning via rich textual descriptions

- The task description is a text representation of the prediction task

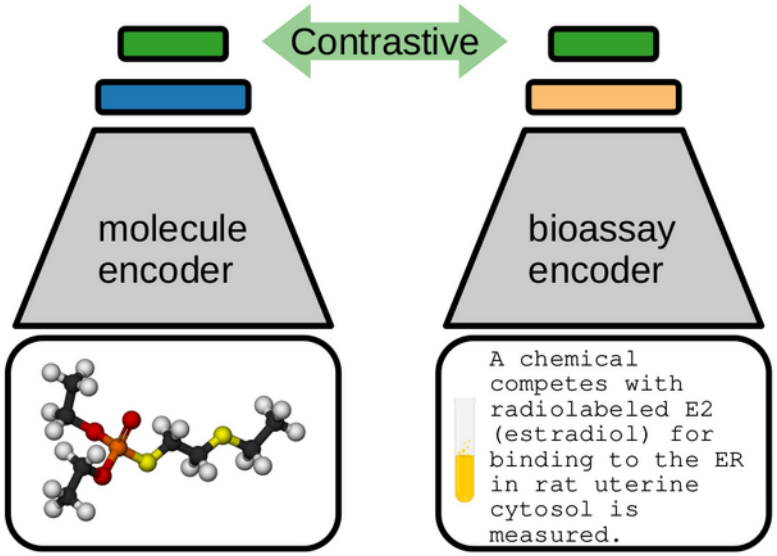
$$\hat{y} = g_{\boldsymbol{\theta}}(m, a)$$

  - a : text description of the activity or task
  - $g_{\boldsymbol{\theta}}(.,.)$: neural network
- Allows for zero-shot transfer learning

Vall, A., Hochreiter, S., & Klambauer, G. (2021) BioassayCLR: Prediction of biological activity for novel bioassays based on rich textual descriptions. ELLIS Machine Learning for Molecules workshop.

5.2 Using a text description of the biological effect

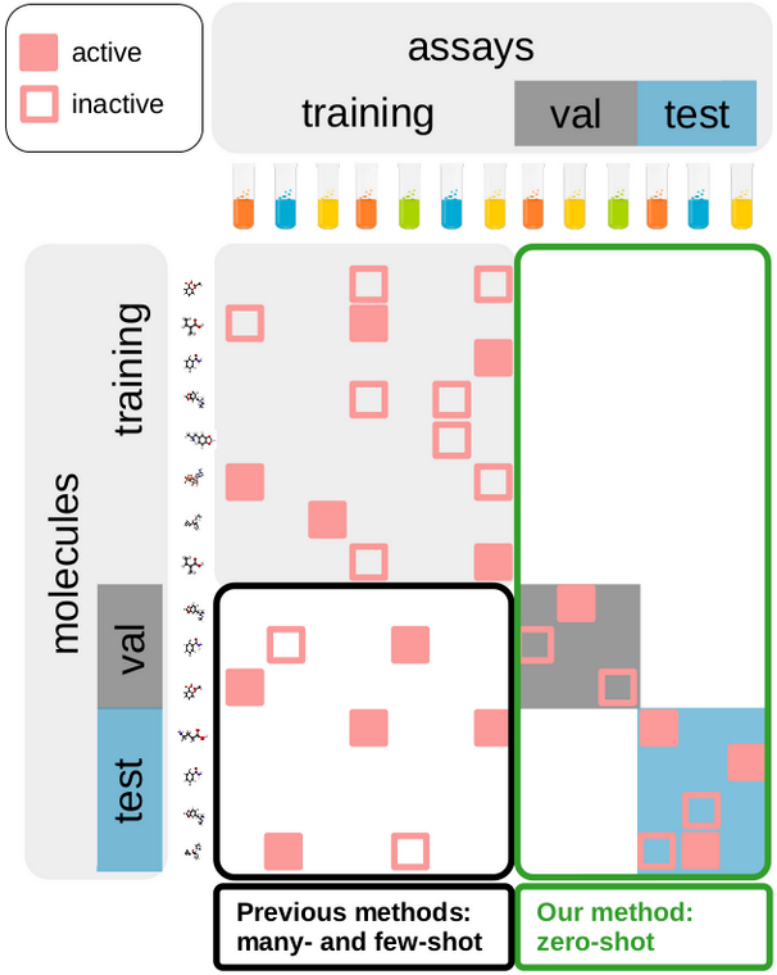
- Available for non-target related tasks
- Large amounts of data in PubChem
  - Text description of wet-lab procedures
- Biomedical texts
  - However: few molecules and activities



Vall, A., Hochreiter, S., & Klambauer, G. (2021) BioassayCLR: Prediction of biological activity for novel bioassays based on rich textual descriptions. ELLIS Machine Learning for Molecules workshop.

5.2 Using a text description of the biological effect

- Available for non-target related tasks
- Large amounts of data in PubChem
  - Text description of wet-lab procedures
- Biomedical texts
  - However: few molecules and activities



Vall, A., Hochreiter, S., & Klambauer, G. (2021) BioassayCLR: Prediction of biological activity for novel bioassays based on rich textual descriptions. ELLIS Machine Learning for Molecules workshop.

5.2 Using a text description of the biological effect

Table 1: Mean of AUROC, average precision (AVGP) and negative-class average precision (Neg-AVGP) over 615 test bioassays for zero-shot transfer learning. The table shows the mean and one standard deviation of this mean value over five runs initialized with different random seeds.

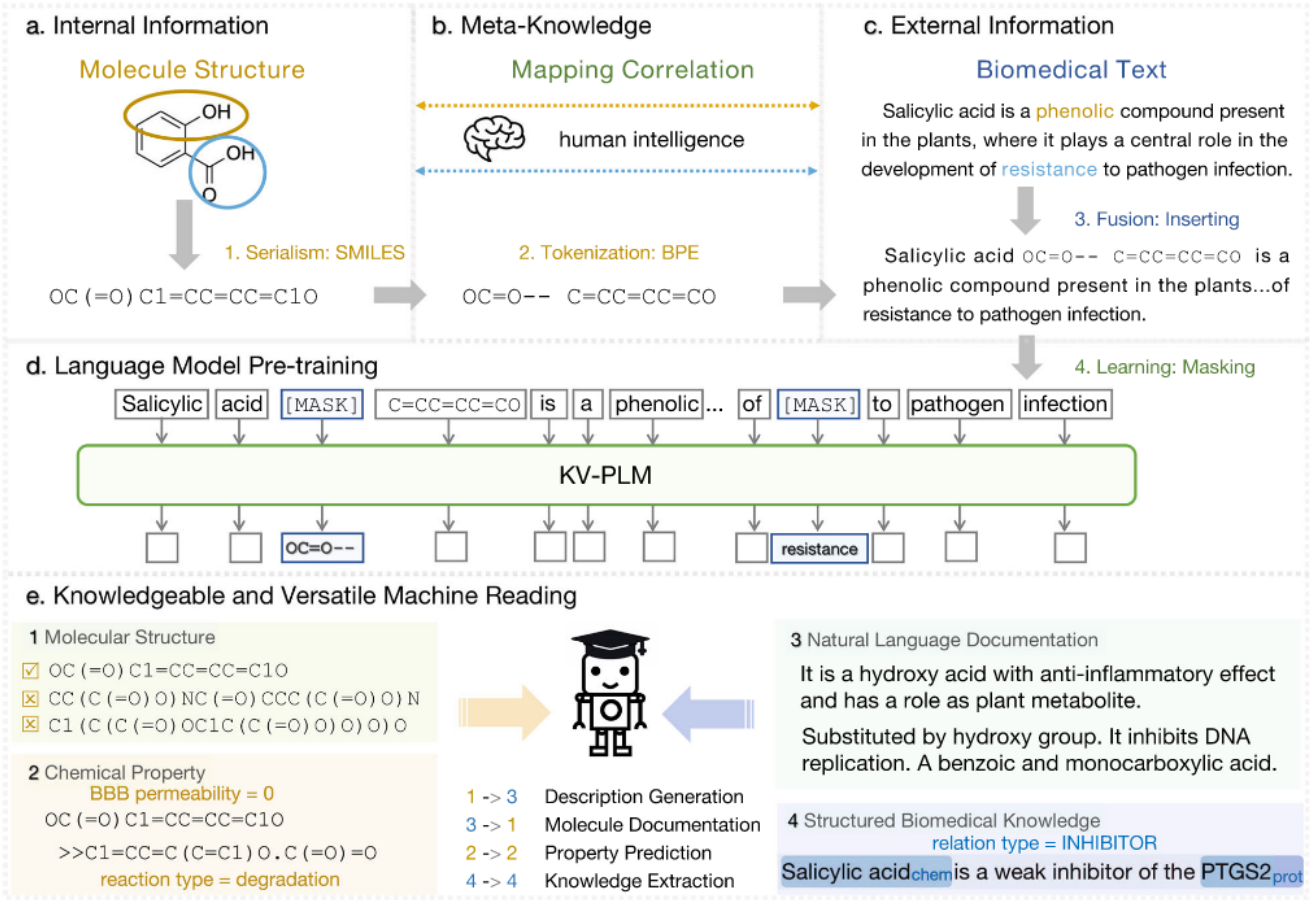
Method	Bioassay encoder	AUROC [%]	AVGP [%]	NegAVGP [%]
BioassayCLR (ours)	LSA	**63.97 ± 0.47**	**46.34 ± 0.64**	75.14 ± 0.48
soft-NN (baseline)	LSA	61.99 ± 0.32	43.31 ± 0.81	75.69 ± 0.53
1-NN (baseline)	LSA	57.16 ± 0.92	41.25 ± 1.09	72.43 ± 0.52
BioassayCLR (ours)	BioBERT	62.52 ± 0.93	44.93 ± 0.72	75.28 ± 0.45
soft-NN (baseline)	BioBERT	61.71 ± 0.77	42.58 ± 0.70	75.31 ± 0.49
1-NN (baseline)	BioBERT	55.15 ± 0.76	40.89 ± 0.64	72.23 ± 0.56
MT-DNN[1] [9, 49, 31, 41]	–	49.68 ± 0.49	38.48 ± 0.45	69.35 ± 0.17

[1] equivalent to a random classifier, in this case

- Zero-shot results:
  - Predictive quality without any training data (actives/inactives)
  - Only text description of assay is available

Vall, A., Hochreiter, S., & Klambauer, G. (2021) BioassayCLR: Prediction of biological activity for novel bioassays based on rich textual descriptions. ELLIS Machine Learning for Molecules workshop.

5.2 Biomedical texts



- Transformer training on human language and structural tokens
- Some zero-shot capabilities; however: limited data in biomedical texts

Zeng, Z., Yao, Y., Liu, Z., & Sun, M. (2022). A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. Nature communications, 13(1), 1-11.

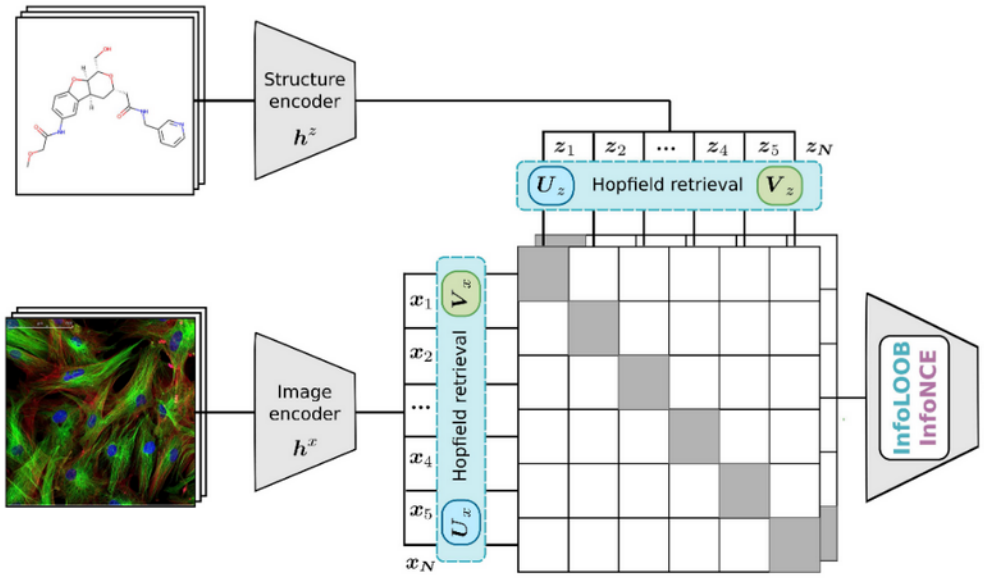
5.3 Zero-shot learning via biological images

- The task description is a text representation of the prediction task

$$\hat{y} = g_{\boldsymbol{\theta}}(m, \mathrm{x})$$

  - x : image describing the task
  - $g_{\boldsymbol{\theta}}(.,.)$: neural network
- Allows for zero-shot transfer learning

5.3 Co-learning of image- and structure-based molecule representations

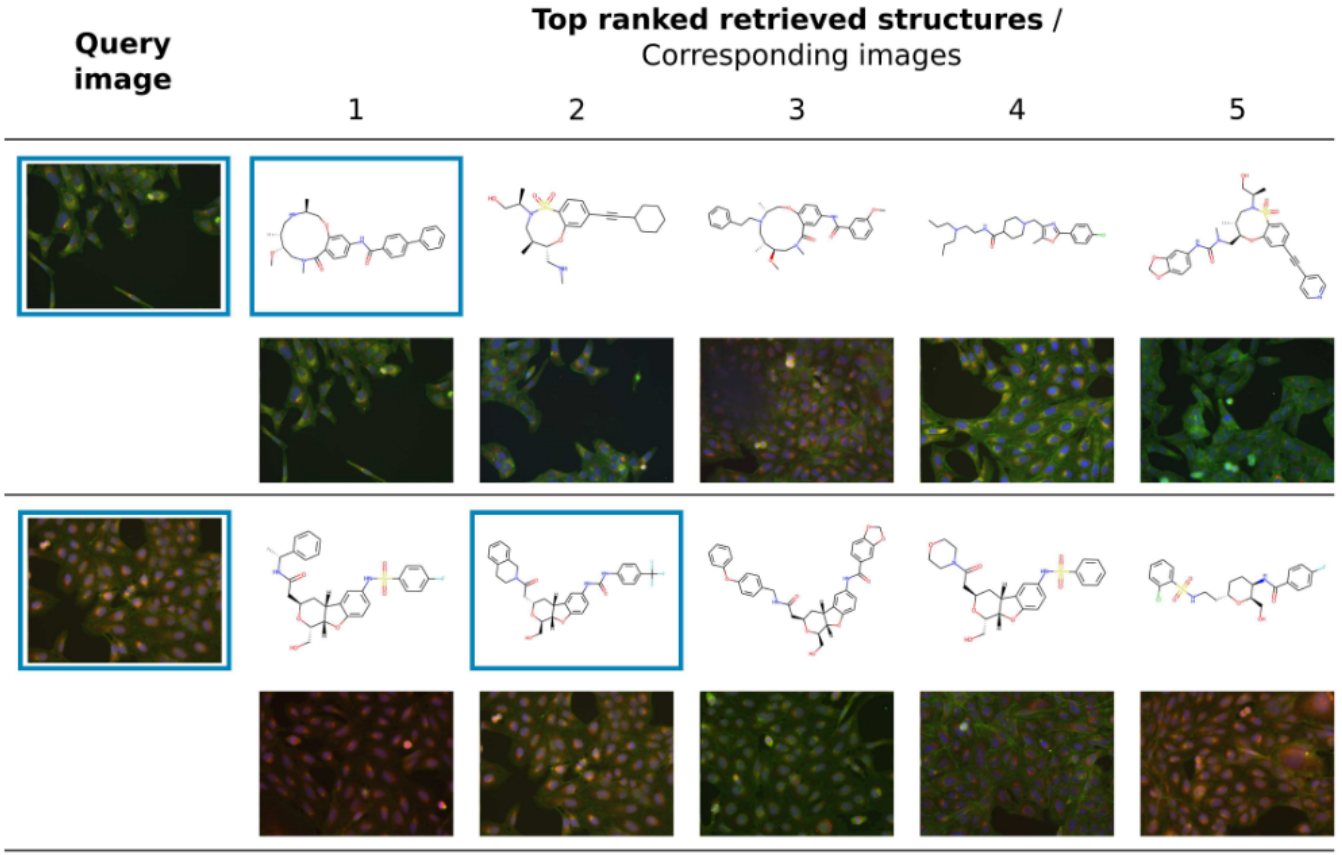


- Molecule-perturbed images allow to train both **image-based** and **structure-based encoders** together
- Similar to CLIP/DALL-E2: those are currently considered some of the strongest recent advances of AI; spectacular zero-shot transfer learning capabilities!

Sanchez-Fernandez, A., Rumetshofer, E., Hochreiter, S. & Klambauer, G. (2022, March). Contrastive learning of image-and structure-based representations in drug discovery. In ICLR2022 Machine Learning for Drug Discovery.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (pp. 8748-8763). PMLR.

5.3 Co-learning of image- and structure-based molecule representations



- Correctly retrieves the correct structure based on image in 3[2.5-4.0]% of cases
  - Considered impossible by human experts

Sanchez-Fernandez, A., Rumetshofer, E., Hochreiter, S. & Klambauer, G. (2022, March). Contrastive learning of image-and structure-based representations in drug discovery. In ICLR2022 Machine Learning for Drug Discovery.

5.3 Co-learning of image- and structure-based molecule representations

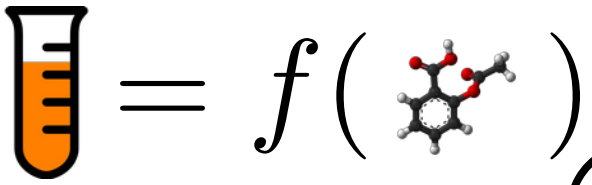
- Linear probing on activity prediction tasks

Type	Method	AUC	F1	AUC >0.9	AUC >0.8	AUC >0.7
Linear probing on self-supervised	CLOOME	**0.714**±0.20	0.395±0.32	57	84	109
Supervised	ResNet	**0.731**±0.19	0.508±0.30	68	94	119
	DenseNet	0.730±0.19	0.530±0.30	61	98	121
	GapNet	0.725±0.19	0.510±0.29	63	94	117
	MIL-Net	0.711±0.18	0.445±0.32	61	81	105
	M-CNN	0.705±0.19	0.482±0.31	57	78	105
	SC-CNN	0.705±0.20	0.362±0.29	61	83	109
	FNN	0.675±0.20	0.361±0.31	55	71	90

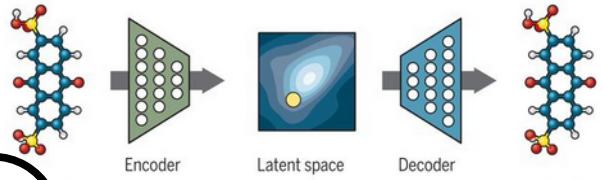
Sanchez-Fernandez, A., Rumetshofer, E., Hochreiter, S. & Klambauer, G. (2022, March). Contrastive learning of image-and structure-based representations in drug discovery. In ICLR2022 Machine Learning for Drug Discovery.

6. Few and zero-shot learning in other domains

Activity and property prediction

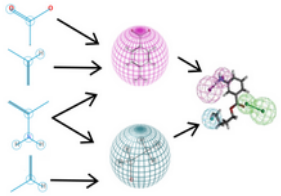


Molecule generation & optimization



$$\boldsymbol{x} = g_{\boldsymbol{\phi}}(\boldsymbol{y})$$

Molecular representations & molecular modeling

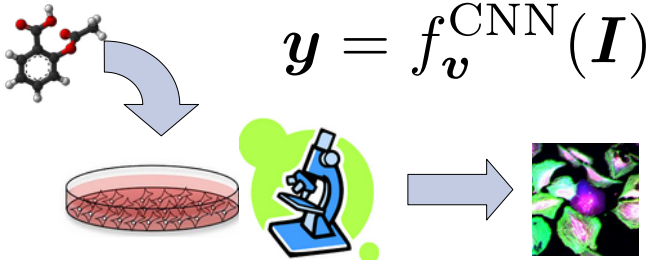


$$\boldsymbol{h_w}(\boldsymbol{x})$$

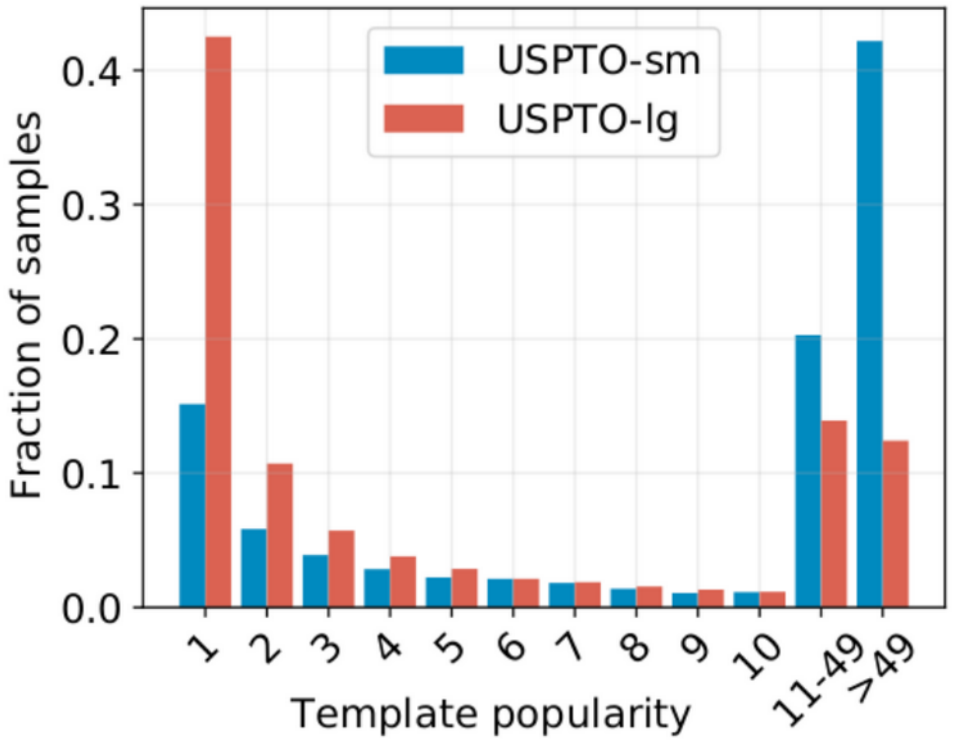
Reaction prediction & chemical synthesis planning

Image analysis

$$\boldsymbol{y} = f_{\boldsymbol{v}}^{\mathrm{CNN}}(\boldsymbol{I})$$



6.1 Few-shot learning for template-based reaction prediction

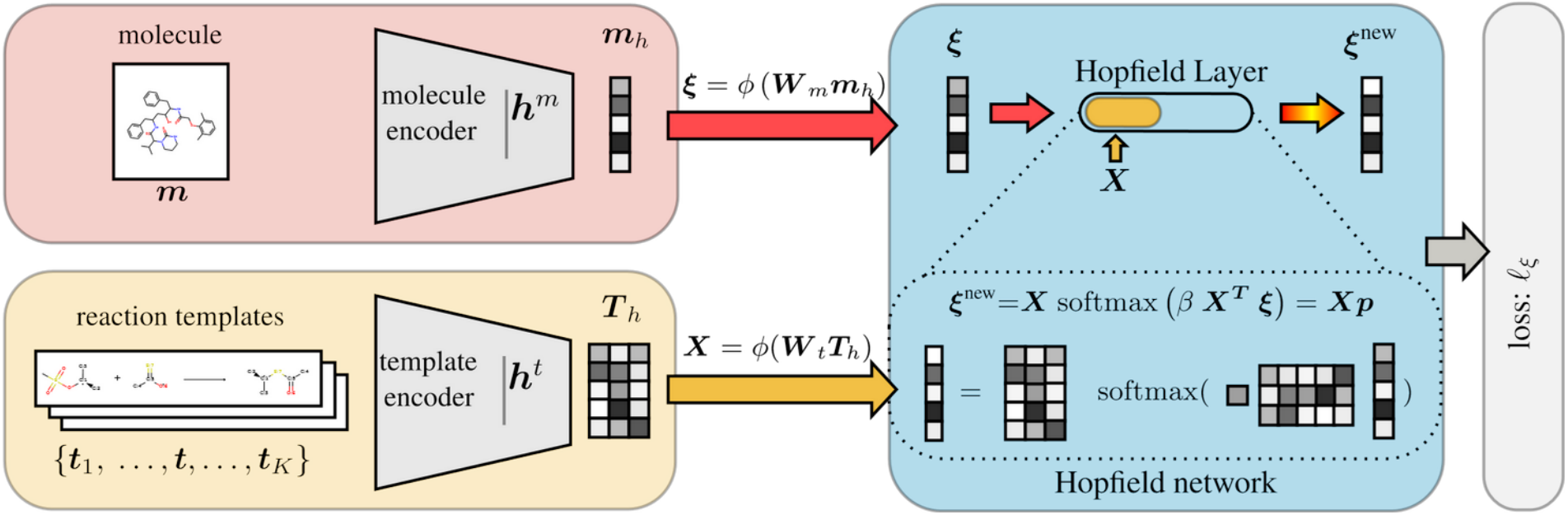


- Template-based methods rely on transformation rules
- Transformation rules are often rare

JOHANNES KEPLER UNIVERSITY LINZ

Seidl, P., Renz, P., Dyubankova, N., Neves, P., Verhoeven, J., Wegner, J. K., ... & Klambauer, G. (2022). Improving Few-and Zero-Shot Reaction Template Prediction Using Modern Hopfield Networks. Journal of Chemical Information and Modeling.

6.1 Few-shot learning for template-based reaction prediction



Modern Hopfield network (MHN):
Reaction templates stored and retrieved from an associative memory

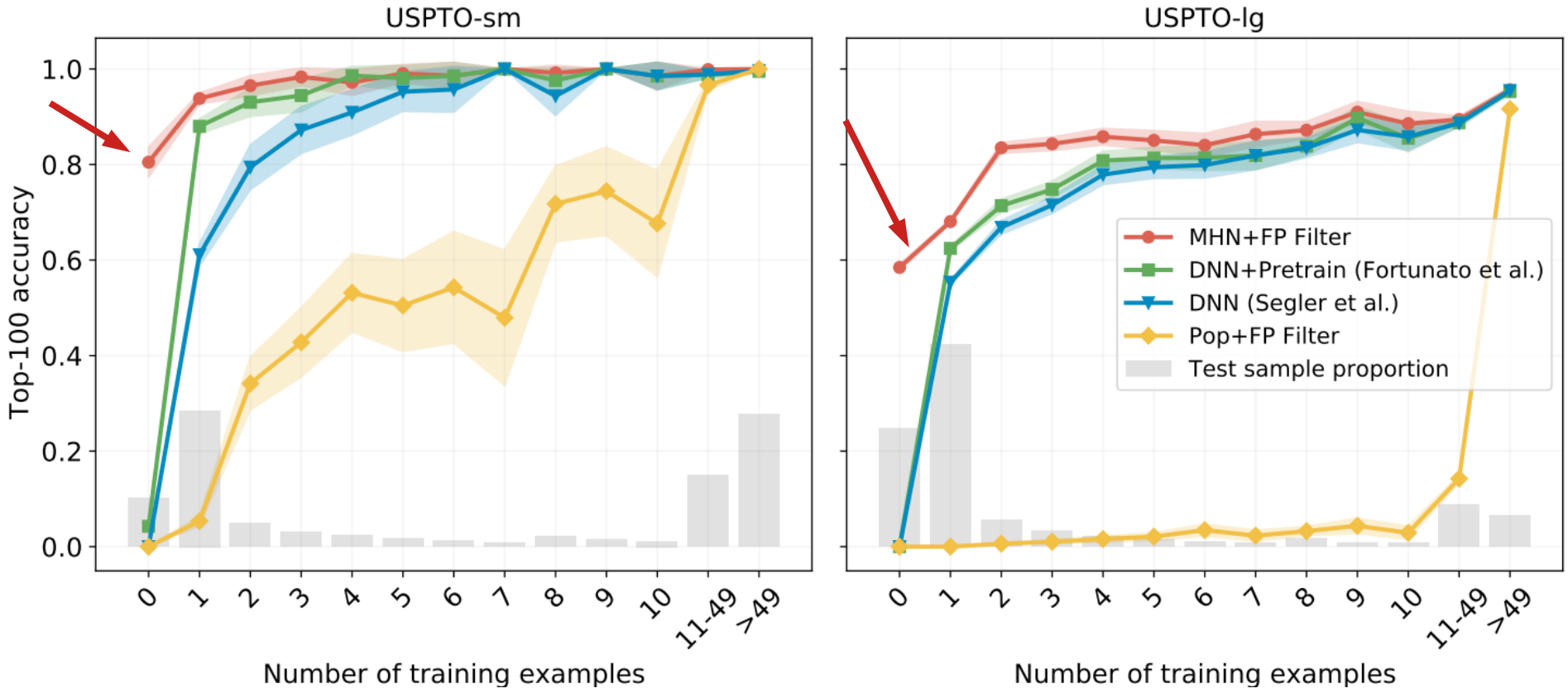
Previous approaches

$$\hat{y} = \text{softmax}(W \; h^m(m))$$

Our approach (simplified):

$$\hat{y} = \text{softmax}(h^t(T) \; h^m(m))$$

6.1 Few-shot learning for template-based reaction prediction



6.1 Few-shot learning for template-based reaction prediction

Abbr.	Ref.	Cat.	Top-1	Top-3	Top-5	Top-10	Top-20	Top-50
MHNreact	ours	tb	$51.8^{\pm.2}$	$\mathbf{74.6}^{\pm.3}$	$\mathbf{81.2}^{\pm.2}$	$\mathbf{88.1}^{\pm.2}$	$\mathbf{92.0}^{\pm.1}$	$\mathbf{94.0}^{\pm.0}$
Neuralsym	[11]	tb	$45.2^{\pm.2}$	$67.9^{\pm.5}$	$75.8^{\pm.2}$	$83.5^{\pm.2}$	$89.1^{\pm.1}$	$93.5^{\pm.1}$
Pop		tb	18.4	38.7	48.6	63.0	75.8	89.8
Transformer		tf	43.7	59.7	65.1	70.1	73.5	75.0
Dual-TB	[66]	tb	**55.2**	**74.6**	80.5	86.9		
Dual-TF	[66]	tf	53.3	69.7	73.0	75.0		
ATx100	[20]	tf	53.5		81.0	85.7		
GLN	[29]	tb	52.5	69.0	75.6	83.7	89.0	92.4
RetroPrime	[67]	tf	51.4	70.8	74.0	76.1		
G2G	[24]	tf	48.9	67.6	72.5	75.5		
MEGAN	[68]	tf	48.1	70.7	78.4	86.1	90.3	93.2
GET-LT1	[69]	tf	44.9	58.8	62.4	65.9		
Neuralsym	[11] [29]	tb	44.4	65.3	72.4	78.9	82.2	83.1
GOPRO	[70]	tf	43.8	57.2	61.4	66.6		
SCROP	[71]	tf	43.7	60.0	65.2	68.7		
LV-Trans	[72]	tf	40.5	65.1	72.8	79.4		
Trans	[19]	tf	37.9	57.3	62.7			
Retrosim	[31]	tb	37.3	54.7	63.3	74.1	82.0	85.3

Overview

1. Introduction and motivation
2. Activity prediction and molecule encoders
   1. Drawbacks of current approaches
   2. Narrow AIs
   3. Multi-task deep networks
3. Zero- and few-shot learning
   1. Definition, problem setting
   2. Categories
4. Few-shot learning methods in drug discovery
   1. Data: FS-Mol
   2. Optimizer-based methods: MAML
   3. Embedding-based methods:
      1. Generalized framework
      2. Frequent hitters model
      3. Similarity search
      4. Neural similarity search
      5. IterRefLSTM
      6. ProtoNet
   4. Results
5. Zero-shot learning
   1. Proteo-chemometric models
   2. Text-based models
   3. Image-based models
6. Few- and zero-shot learning in other domains
   1. Chemical reactions
7. Summary

7. Summary

- Current methods in drug discovery suffer from the usual problems of Deep Learning methods: narrow AIs, task-specific, data-hungry
- A step towards broad AIs: **adaptability** via
  - few-shot learning methods
  - zero-shot learning methods
- Introduced and presented several FSL and ZSL methods
- Applications for activity prediction and chemical reactions shown
- Advance machine learning, save the world!