Removing Biases from Molecular Representations via Information Maximization

Joint work with

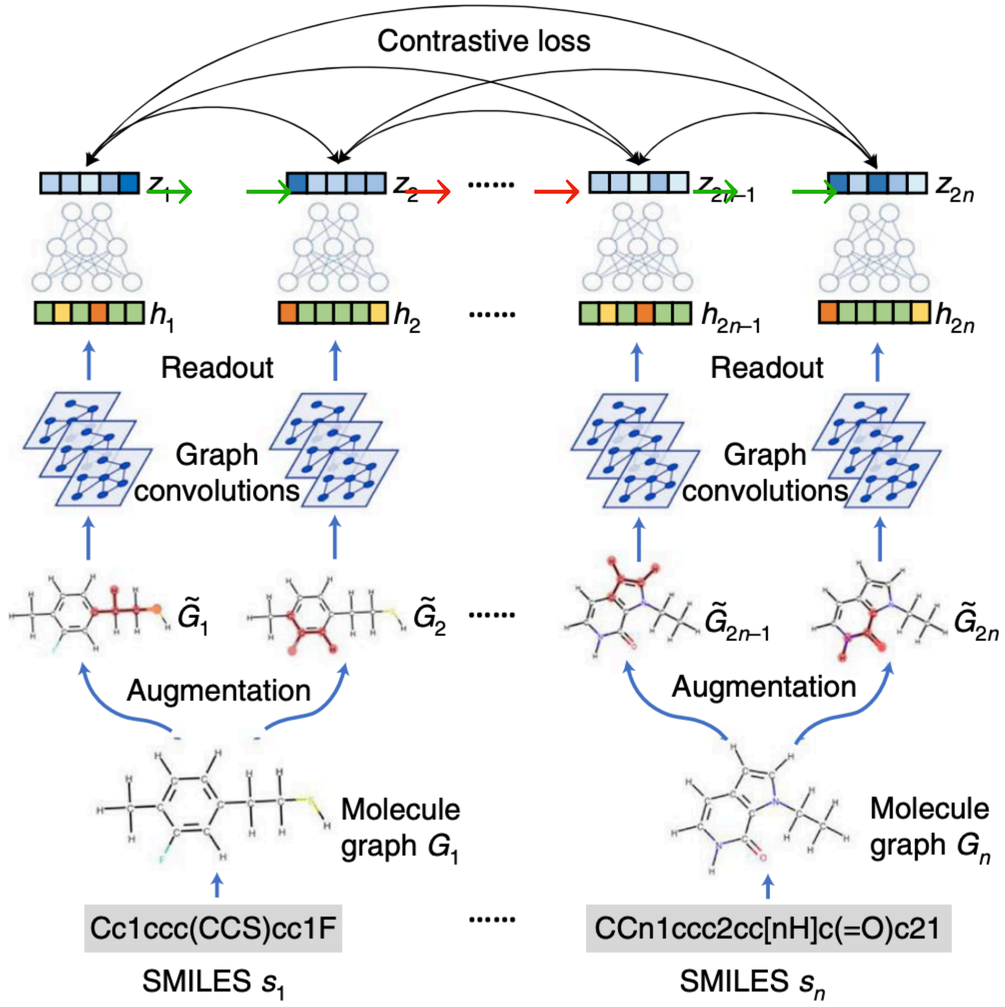
Chenyu Wang

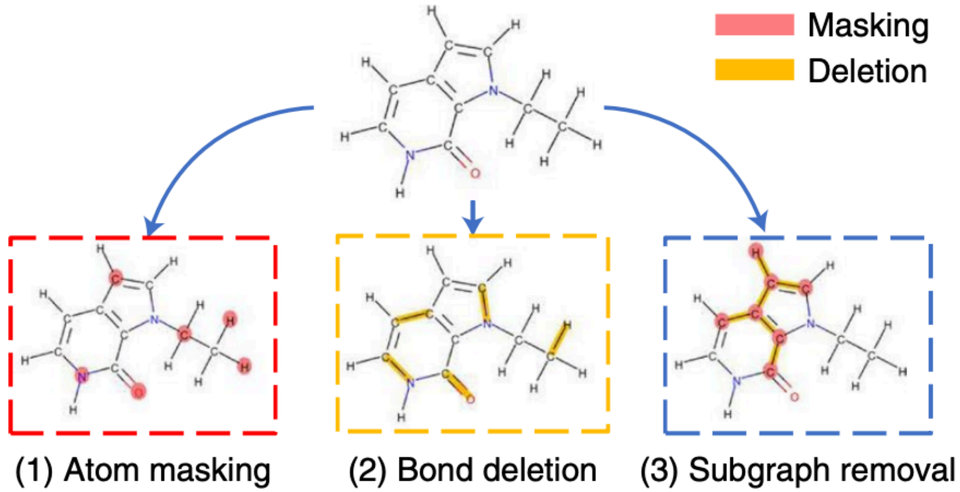
Sharut Gupta

Caroline Uhler

Tommi Jaakkola

Molecular Representation Learning

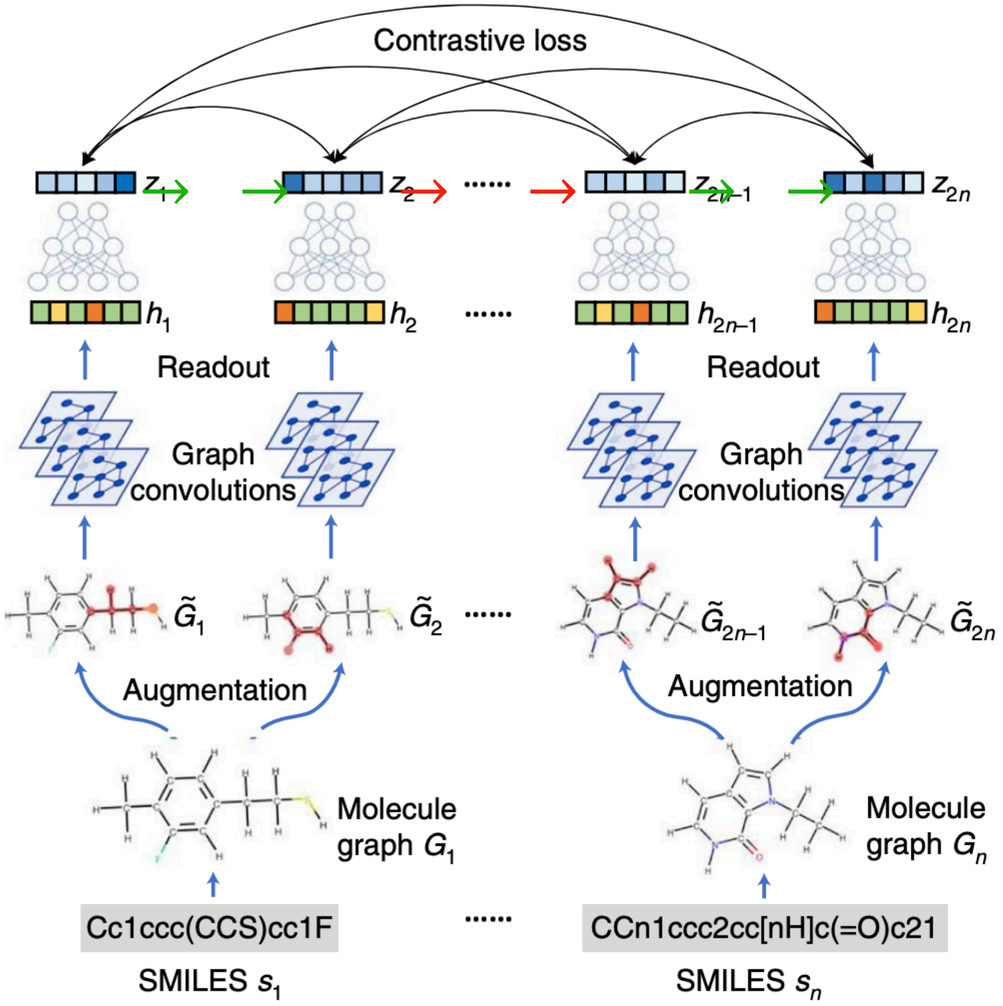




Wang, Yuyang, et al. "Molecular contrastive learning of representations via graph neural networks." Nature Machine Intelligence

Molecular Representation Learning

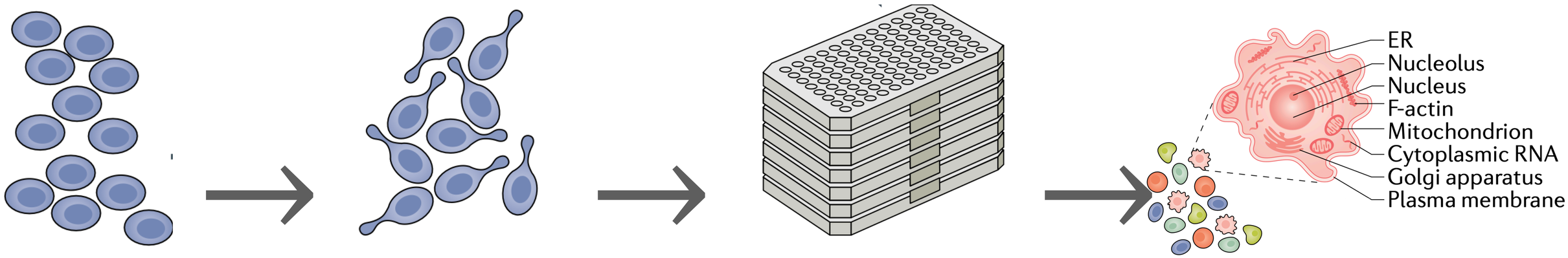
- Molecules with **similar** structures can have very **different** effects in the cellular context.



Wang, Yuyang, et al. "Molecular contrastive learning of representations via graph neural networks." Nature Machine Intelligence

High-content Drug Screens

- Output post-perturbation (i.e., after the application of a drug) cellular images and gene expression.



Healthy and diseased patient cell lines

Drugs or genetic perturbations

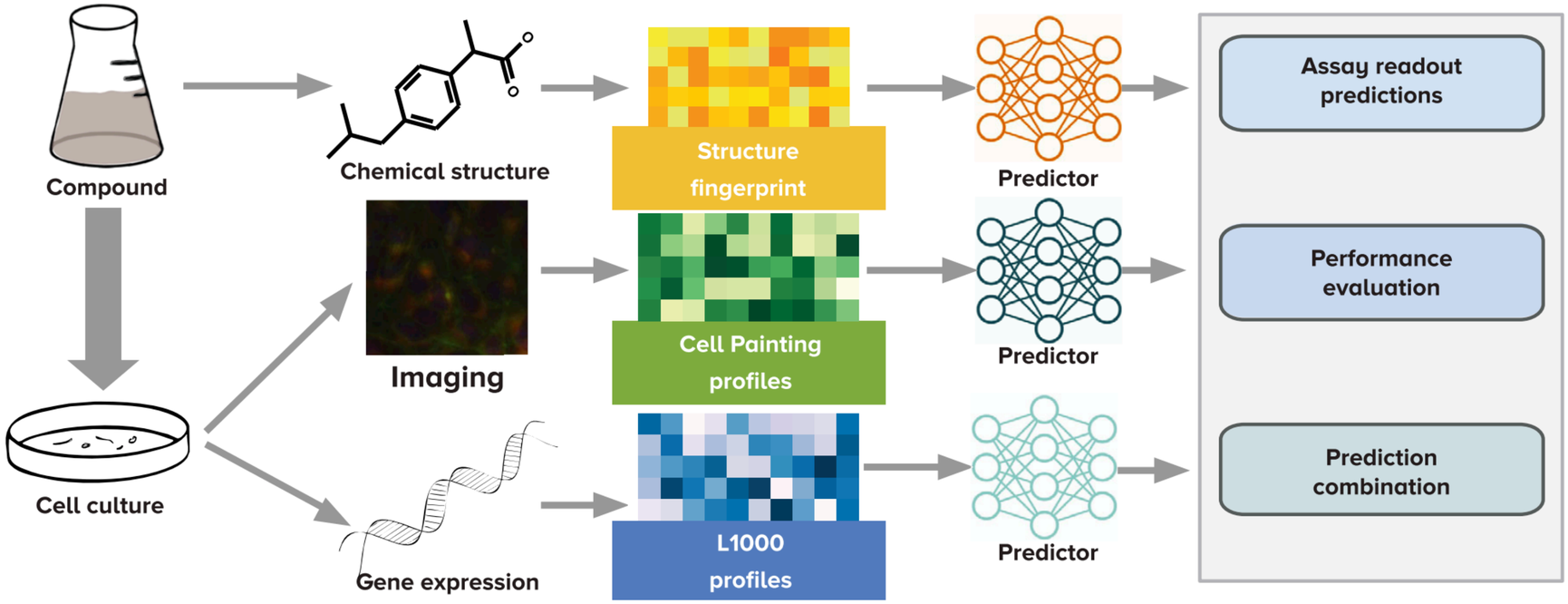
High-throughput staining and imaging: e.g. Cell Painting assay

Drug screens

Chandrasekaran, Srinivas Niranj, et al. "Image-based profiling for drug discovery: due for a machine-learning upgrade?." Nature Reviews Drug Discovery

Chemical Structure and High-content Drug Screens

- High-content drug screens improve our understanding of the biological effect of a compound. However, due to experimental constraint, we can't screen each molecule in wet-lab, and *multimodal molecular representations* are necessary.



Moshkov, Nikita, et al. "Predicting compound activity from phenotypic profiles and chemical structures." Nature Communications

Batch effect!

- In molecular biology, a batch effect is a change in data that is caused by **non-biological factors** in an experiment. Batch effects can lead to inaccurate conclusions if their causes are correlated with outcomes of interest in an experiment.

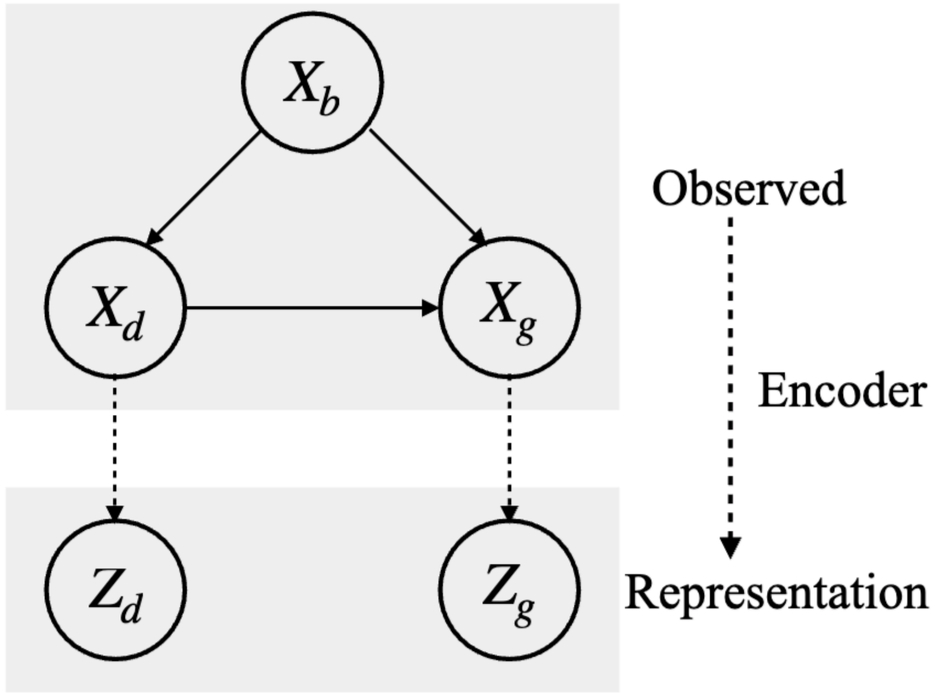
Moshkov, Nikita, et al. "Predicting compound activity from phenotypic profiles and chemical structures." Nature Communications

Batch effect!

- The batch identifier is predictable from both the phenotypic screens (batch effect) and the molecular structure (batch confounder).
- For example, in *Bray 2017 dataset*:
    - CellProfiler features: accuracy > 90% (versus 1% with a random predictor)
    - Molecular structure + mol2vec featurizer: accuracy ~50%

Relation to Conditional Mutual Information

- X_d : Drug structure
- X_b : Experimental batch number
- X_g : Phenotypic change induced by drug perturbation

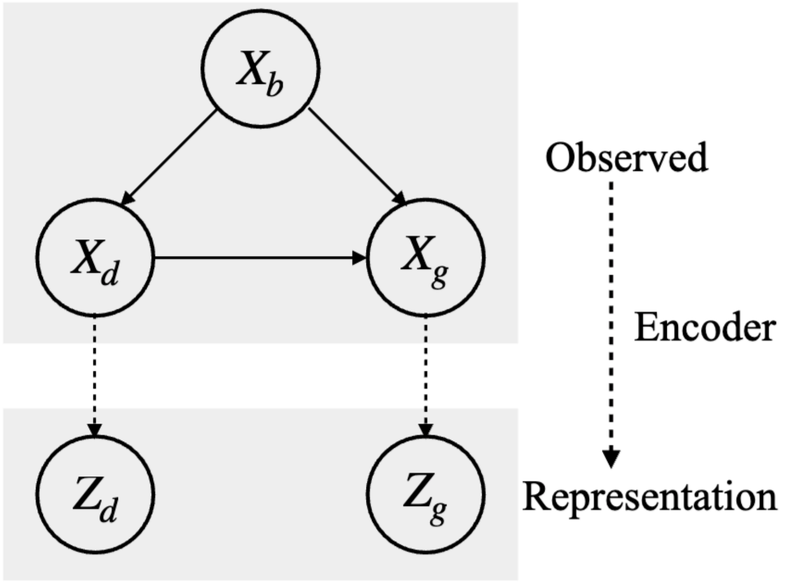


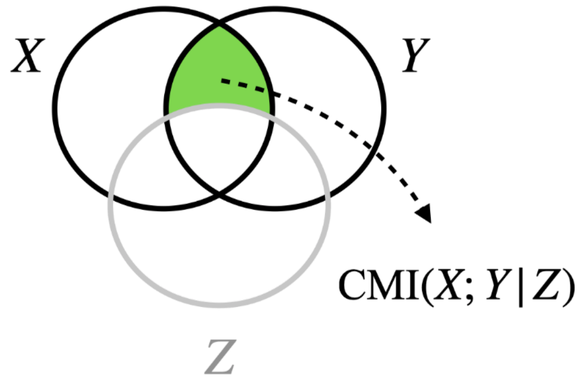
Relation to Conditional Mutual Information

- X_d : Drug structure
- X_b : Experimental batch number
- X_g : Phenotypic change induced by drug perturbation

$$\max_{\theta_d,\theta_g} \frac{1}{2}\left(I(Z_d; X_g|X_b; \theta_d) + I(Z_g; X_d|X_b; \theta_g)\right)$$

Further justified in Robinson et al., 2021 and Ma et al., 2021



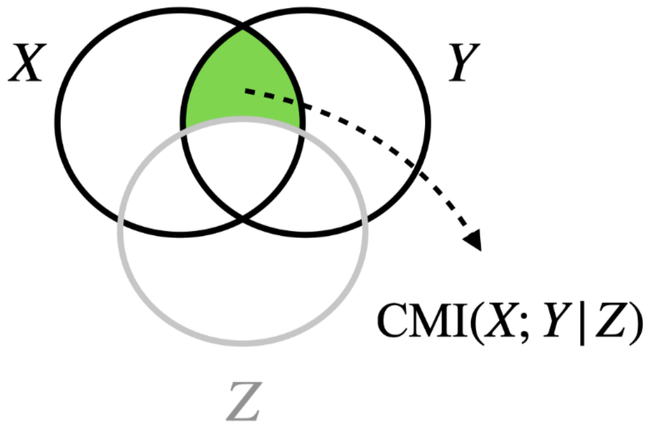


Relation to Conditional Mutual Information

$$\max_{\theta_d,\theta_g} \frac{1}{2}\left(I(Z_d; X_g|X_b;\theta_d) + I(Z_g; X_d|X_b;\theta_g)\right) \geq \max_{\theta_d,\theta_g} I(Z_d; Z_g|X_b;\theta_d,\theta_g)$$

This objective function emphasizes drug's bioactivity by focusing on shared features of the two modalities that are **unrelated** to batch

- CLIP → Conditional CLIP with negatives drawn from $p(z_d|x_b)$ and $p(z_g|x_b)$ [*Ma et al., 2021*]



InfoNCE as a lower bound to Mutual Information

- Success of InfoNCE is based on maximizing mutual information $I(X;Z) \geq I(Z;Z^1)$

Positive

$$L_{\text{InfoNCE}} = \mathbb{E}_{p(z,z^1)p(z^{2:K})}\left[-\log \frac{e^{f(z,z^1)}}{\frac{1}{K}\sum_{i=1}^{K} e^{f(z,z^i)}}\right]$$

Negatives

- Bi-modal contrastive learning, the InfoNCE objective → CLIP

Positive Image pair

Positive Text pair

$$\frac{1}{2}\left[\mathbb{E}_{p(z_T^1,z_I^1)p(z_I^{2:K})}\left[-\log \frac{e^{f(z_T^1,z_I^1)}}{\frac{1}{K}\sum_{i=1}^{K} e^{f(z_T^1,z_I^i)}}\right] + \mathbb{E}_{p(z_T^1,z_I^1)p(z_T^{2:K})}\left[-\log \frac{e^{f(z_T^1,z_I^1)}}{\frac{1}{K}\sum_{i=1}^{K} e^{f(z_T^i,z_I^1)}}\right]\right]$$

Negative Image pair

Negative Text pair

InfoNCE as a lower bound to Mutual Information

- Success of InfoNCE is based on maximizing mutual information $I(X;Z) \geq I(Z;Z^1)$

$$I(X_1;Y) \geq 1 + \mathbb{E}_{p(x_{1:K})p(y|x_1)}\left[\log \frac{e^{f(x_1,y)}}{a(y;x_{1:K})}\right] - \mathbb{E}_{p(x_{1:K})p(y)}\left[\frac{e^{f(x_1,y)}}{a(y;x_{1:K})}\right]$$

$$I(X_1;Y) \geq \mathbb{E}_{p(x_{1:K})p(y|x_1)}\left[\log \frac{e^{f(x_1,y)}}{a(y;x_{1:K})}\right]$$

$$a(y;x_{1:K}) = m(y;x_{1:K}) = \frac{1}{K}\sum_{i=1}^{K} e^{f(x_i,y)}$$

InfoCORE

Variables for anchor-positive pair

Proposition 2. *Given samples* (z_d^1, z_g^1, x_b^1) *drawn from the joint distribution* $(Z_d^1, Z_g^1, X_b^1) \sim p(z_d, z_g, x_b)$ *and* $z_d^{2:K}$ *drawn i.i.d. from the marginal distribution* $Z_d^i \sim p(z_d)$ *for* $i = 2, ..., K$*, then the conditional mutual information* $I(Z_d^1; Z_g^1 | X_b^1)$ *has the following lower bound:*

Negatives

$$I(Z_d^1; Z_g^1 | X_b^1) \geq -L_{CLIP} - L_{CLF} + C - H(X_b^1), \quad \text{(3)}$$

constant

$$\text{where } L_{CLIP} = -\frac{1}{2}\left[\mathbb{E}_{p(z_d^1, z_g^1, x_b^1)p(z_d^{2:K})}\left[\log \frac{e^{f(z_g^1, z_d^1)}}{\frac{1}{K}\sum_{i=1}^{K} e^{f(z_g^1, z_d^i)} \cdot \hat{p}_g(x_b^1 | z_g^1, z_d^i)}\right]\right.$$

$$\left. + \mathbb{E}_{p(z_d^1, z_g^1, x_b^1)p(z_g^{2:K})}\left[\log \frac{e^{f(z_g^1, z_d^1)}}{\frac{1}{K}\sum_{i=1}^{K} e^{f(z_g^i, z_d^1)} \cdot \hat{p}_d(x_b^1 | z_g^i, z_d^1)}\right]\right],$$

Posterior reweighing

$$L_{CLF} = \frac{1}{2}\left[\mathbb{E}_{p(z_d^1)}\left[D_{\mathrm{KL}}\left(p\left(x_b^1 | z_d^1\right) \| \hat{p}\left(x_b^1 | z_d^1\right)\right)\right] + \mathbb{E}_{p(z_g^1)}\left[D_{\mathrm{KL}}\left(p\left(x_b^1 | z_g^1\right) \| \hat{p}\left(x_b^1 | z_g^1\right)\right)\right]\right],$$

Classification Loss

$$C = \frac{1}{2}\,\mathbb{E}_{p(z_d^1, z_g^1, x_b^1)}\left[\log \frac{\hat{p}_g(x_b^1 | z_g^1, z_d^1) \cdot \hat{p}_d(x_b^1 | z_g^1, z_d^1)}{\hat{p}(x_b^1 | z_g^1) \cdot \hat{p}(x_b^1 | z_d^1)}\right].$$

InfoCORE — Reweighting Factor Estimation

- Estimating posterior distribution of batch given both modalities is challenging
  - The corresponding empirical observations are absent (especially when $i \neq 1$)
  - Poor OOD generalization
  - Computationally intensive for each pair

Tradeoff between CLIP and bias removal

$$\hat{p}_g(x_b^1|z_g^1, z_d^i) = \alpha \cdot \hat{p}(x_b^1|z_g^1) + (1-\alpha) \cdot \hat{p}(x_b^1|z_d^i),\ \hat{p}_d(x_b^1|z_g^i, z_d^1) = \alpha \cdot \hat{p}(x_b^1|z_d^1) + (1-\alpha) \cdot \hat{p}(x_b^1|z_g^i)$$

Proposition 3. *When estimating $\hat{p}_g(x_b^1|z_g^1, z_d^i)$ as the weighted average of $\hat{p}(x_b^1|z_g^1)$ and $\hat{p}(x_b^1|z_d^i)$, and analogously for $\hat{p}_d(x_b^1|z_g^i, z_d^1)$, the term C defined in Proposition 2 is lower bounded by zero.*

$$L_{\text{InfoCORE}} = L_{\text{CLIP}} + L_{\text{CLF}}$$

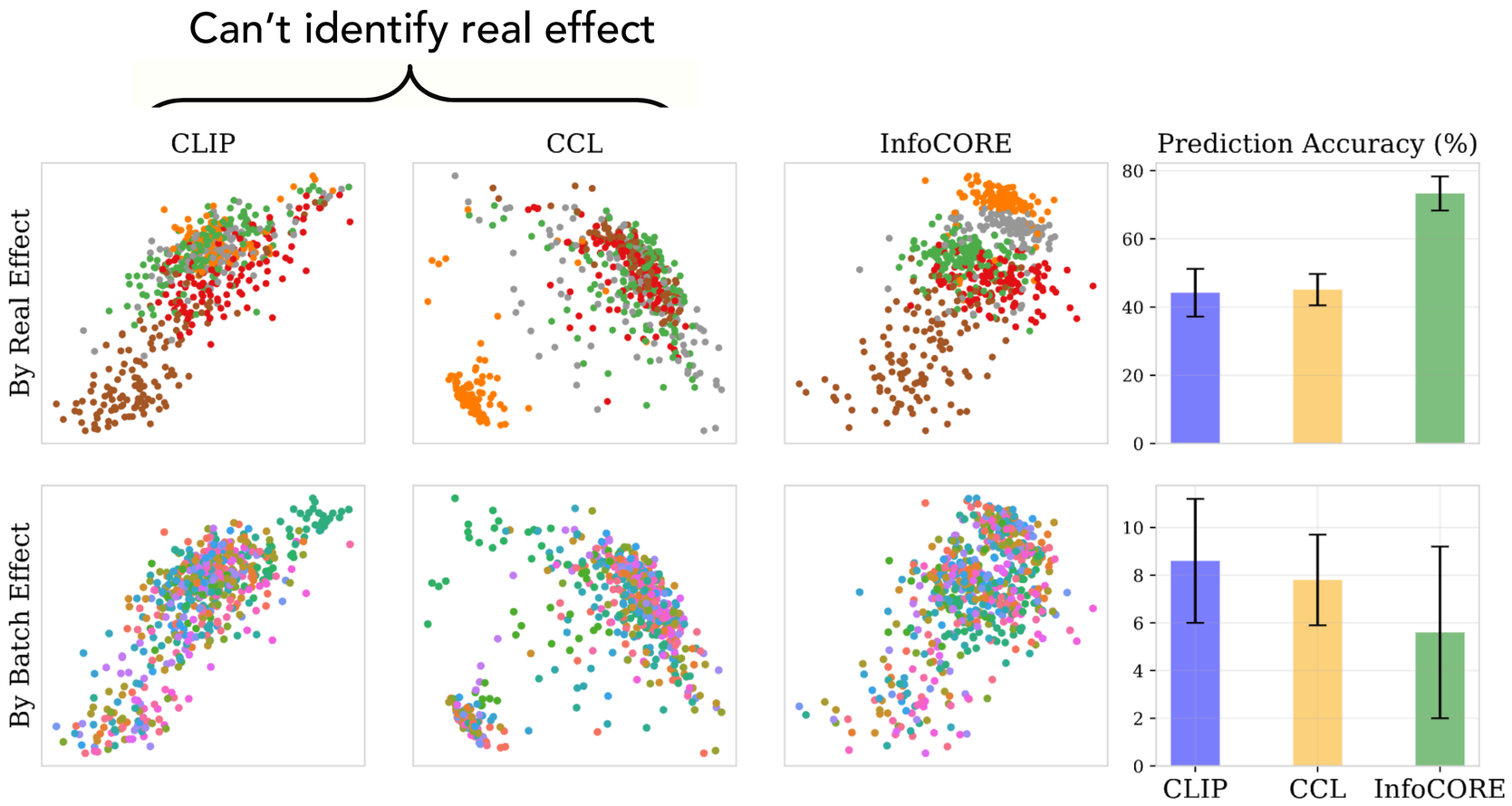
Advantages of estimating batch distribution

- Confounding varies across batches — some batches might still have random assignment
- Using latents implicitly adjusts training to stop reweighing when debiasing is complete!

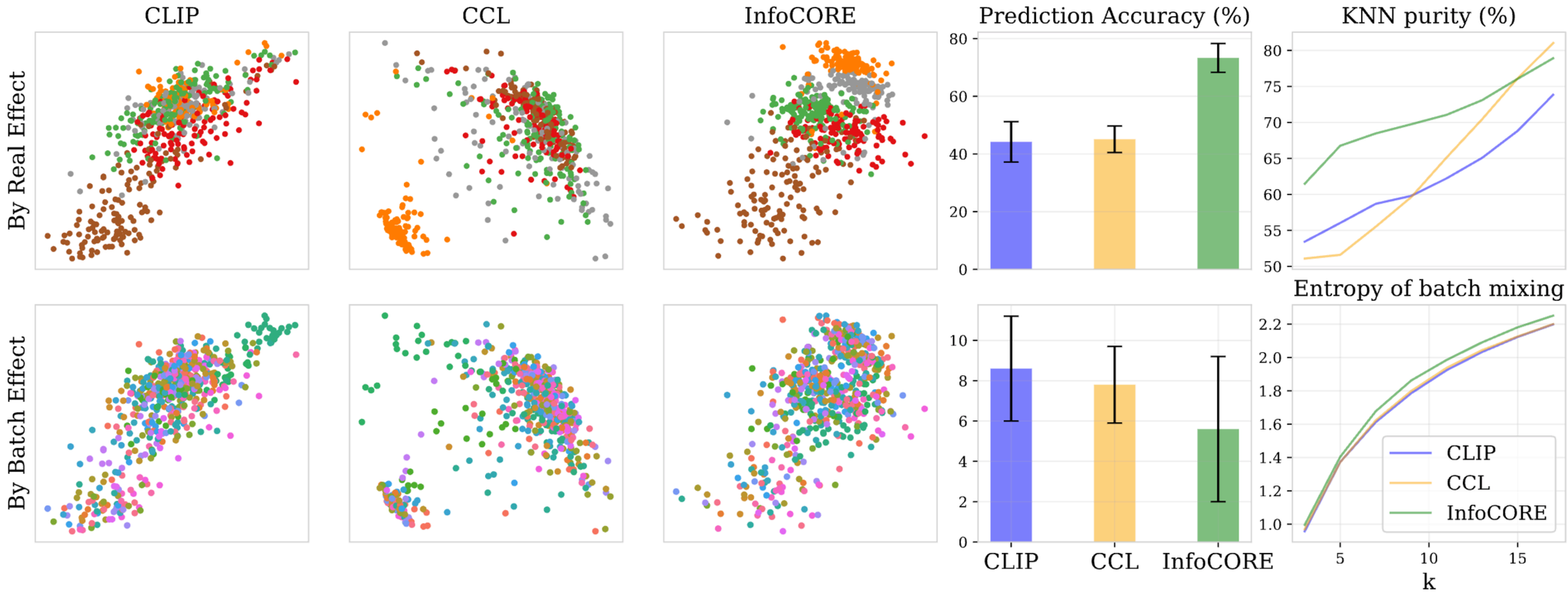
Synthetic Simulation — Data generation

- Randomly assign a real effect identifier (1-5) and a batch effect identifier (1-25) to each of the 1250 samples.
- Each category ≡ 10-D random Gaussian vector
- Each sample (x) ≡ 30-D vector of real effect, batch effect, noise
- $m_1 = MLP_1(x)$ and $m_2 = MLP_2(x)$ represent paired modality data

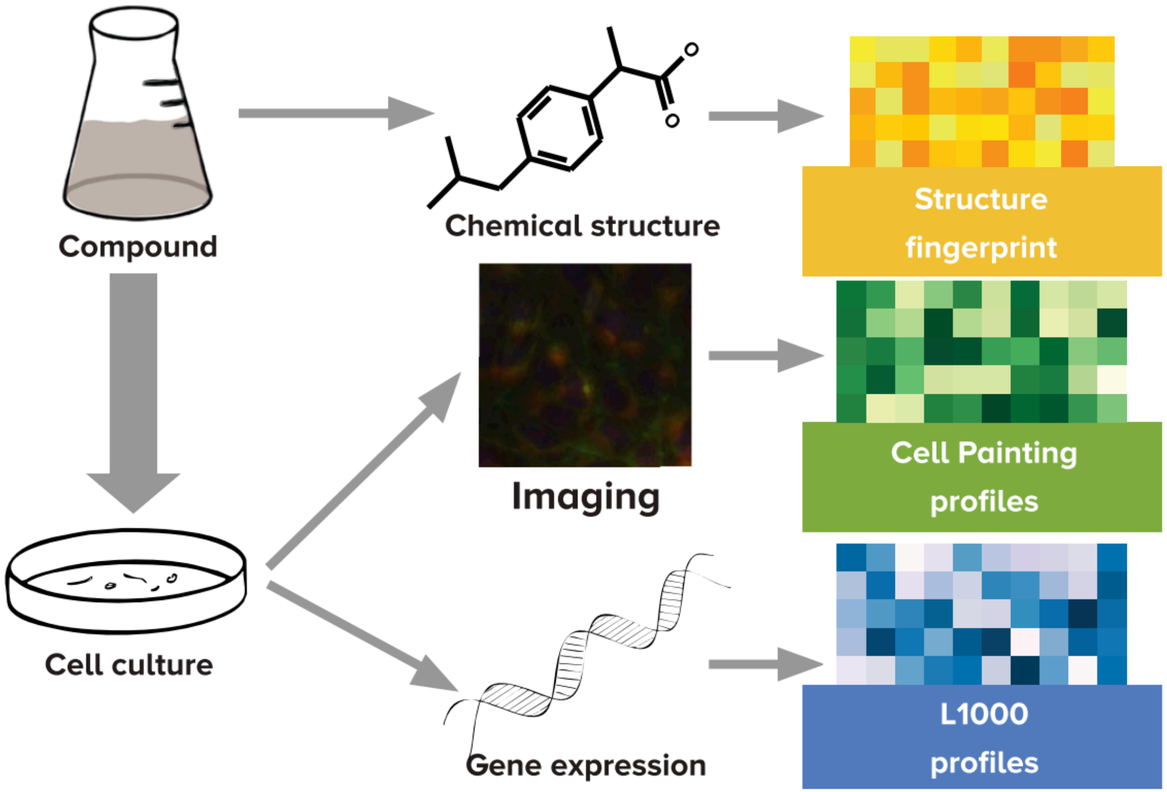
Experimental Results — Synthetic Simulation



Experimental Results — Synthetic Simulation



Experimental Results — Dataset



Drug Chemical Structure

- Mol2vec to get features

Cell imaging profiles obtained from the Cell Painting assay (Bray et al., 2017)

- 30,204 small molecules screened in one cell line i.e. U2OS (a human bone cancer cell line)
- Hand crafted features through *CellProfiler*

L1000 gene expression profiles (Subramanian et al., 2017)

- Nine core cell lines tested for 17,753 drugs
- In total — 82,914 drug-cell line pairs

e.g. A549 (Lung Cancer), MCF-7 (Breast Cancer)

Experimental Results — Drug Representations

- The drug discovery task ≡ identify molecules (e.g., from a **drug repurposing library**) that are most likely to induce a given desired phenotypic change (i.e., gene expression change from diseased towards normal).
- **Molecule-Phenotype Retrieval for Drug Repurposing**. Identifying molecules from a retrieval library (whole / batch) that are most likely to induce a given phenotypic change.

All molecules in held-out set

held-out set molecules that are in the same experimental batch as the retrieving target

Experimental Results — Drug Representations

- **Molecule-Phenotype Retrieval for Drug Repurposing**. Identifying molecules from a retrieval library (whole / batch) that are most likely to induce a given phenotypic change.

Dataset	Gene Expression (GE)						Cell Painting (CP)					
Retrieval Library	*whole*			*batch*			*whole*			*batch*		
Top N Acc (%)	N=1	N=5	N=10	N=1	N=5	N=10	N=1	N=5	N=10	N=1	N=5	N=10
Random	0.03	0.13	0.27	1.58	7.90	15.81	0.02	0.08	0.17	1.59	7.97	15.94
CLIP	5.96	18.59	27.17	12.23	30.29	42.63	**7.23**	**20.95**	**28.89**	13.20	37.78	52.72
CCL	1.93	5.85	8.37	12.76	32.39	45.77	1.31	4.93	7.38	13.20	37.99	53.13
InfoCORE	**6.39**	**18.99**	**27.18**	**14.03**	**33.63**	**46.78**	6.93	20.65	28.22	**13.26**	**38.50**	**53.13**

CCL — Few negatives in the same batch

CLIP — Biased

Experimental Results — Drug Representations

- Molecular property prediction (bioactivity) downstream task

Blood-Brain Barrier Penetration

Clinical Toxicity

Side Effect Resource

Human Immunodeficiency Virus

	Datasets	BBBP	BACE	ClinTox	Tox21	ToxCast	SIDER	HIV	Avg.	PRISM
		Classification (ROC-AUC %) ↑								Reg (R^2 %) ↑
	# Molecules	2039	1513	1478	7831	8575	1427	41127	-	3172
	# Tasks	1	1	2	12	617	27	1	-	5
	Mol2vec	70.7(0.4)	82.9(0.7)	84.9(0.3)	76.0(0.1)	74.4(0.5)	64.9(0.3)	77.7(0.1)	75.9	8.5(0.7)
GE	CLIP	73.5(0.4)	86.1(0.4)	89.6(2.1)	77.3(0.0)	75.7(0.6)	63.7(0.6)	77.7(0.6)	77.6	13.9(0.4)
GE	CCL	73.0(0.8)	85.9(0.6)	90.5(1.0)	77.0(0.2)	**75.8(0.2)**	63.4(0.5)	77.5(0.9)	77.6	**16.0(0.5)**
GE	InfoCORE	**73.5(0.3)**	**86.6(0.3)**	**91.9(1.9)**	**77.4(0.4)**	75.7(0.2)	**64.8(0.6)**	**78.5(0.2)**	**78.3**	14.8(0.1)
CP	CLIP	73.4(0.8)	**85.2(0.4)**	87.3(0.1)	76.4(0.1)	76.7(0.1)	64.8(0.6)	78.2(0.4)	77.4	16.2(0.2)
CP	CCL	73.7(0.5)	84.9(0.9)	87.7(1.8)	75.9(0.3)	75.7(0.4)	65.2(0.4)	**79.3(0.3)**	77.5	14.7(0.3)
CP	InfoCORE	**74.0(0.8)**	85.0(0.2)	**89.3(0.5)**	**76.6(0.1)**	**76.9(0.1)**	**65.2(0.1)**	78.7(0.1)	**78.0**	**16.2(0.3)**

Post-Perturbation Cell Viability

Experimental Results — Representation Fairness

Method	UCI Adult			Law School			Compas		
	Acc↑	EO↓	EOPP↓	Acc↑	EO↓	EOPP↓	Acc↑	EO↓	EOPP↓
CLIP	85.1(0.1)	20.7(1.8)	15.2(1.7)	**83.1(0.2)**	30.9(1.4)	7.9(0.8)	**60.8(2.3)**	18.4(2.4)	11.7(1.9)
CCL	85.1(0.2)	19.0(3.3)	13.3(2.8)	83.0(0.3)	27.8(1.5)	6.7(0.9)	59.5(2.3)	17.1(3.4)	10.1(3.0)
InfoCORE	**85.2(0.1)**	**14.9(1.1)**	**9.7(0.8)**	82.7(0.4)	**25.4(3.6)**	**6.0(1.4)**	60.1(2.1)	**15.3(2.5)**	**9.3(0.8)**

Equalized Odds (EO): sum of the absolute difference of TPR and FPR of the model predictions between two groups

Equality of opportunity (EOPP): absolute difference of TPR of the model predictions between two groups

Thank You

Questions?

arXiv > cs > arXiv:2312.00718