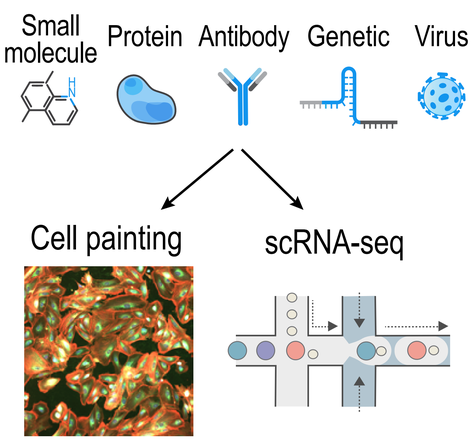
Biological Cartography: Building and Benchmarking Representations of Life

Safiye Celik, Jan-Christian Hütter†, Sandra Melo Carlos†, Nathan H. Lazar*, Rahul Mohan†, Conor Tillinghast*, Tommaso Biancalani†, Marta Fay*, Berton A. Earnshaw*, Imran S. Haque**

*Recursion
†Genentech
A Member of the Roche Group

The continued scaling of genetic perturbation technologies combined with high-dimensional assays (microscopy and RNA-seq) has enabled genome-scale reverse-genetics experiments that go beyond single-endpoint measurements of growth or lethality. Datasets emerging from these experiments can be combined to construct "maps of biology", in which perturbation readouts are placed in unified, relatable embedding spaces to capture known biological relationships and discover new ones.

Perturbations



Embed
Represent perturbation units with vector

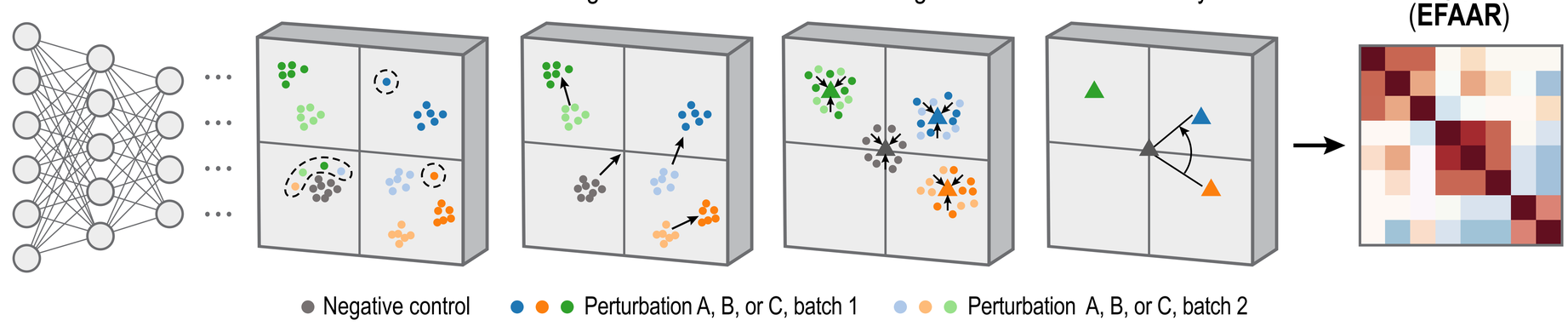
Filter
Remove outliers

Align
Align batches, center at negative controls

Aggregate
Aggregate perturbation units into single vector

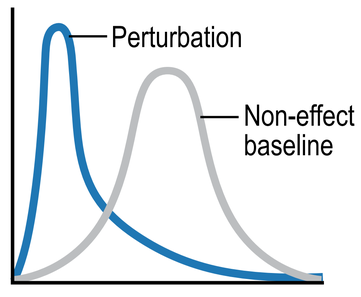
Relate
Quantify pairwise similarity

Results from map-building pipeline (**EFAAR**)



Univariate Benchmark – Does perturbation have an effect?

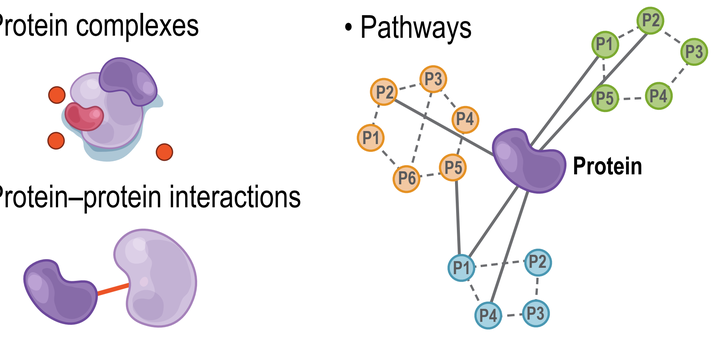
Compare perturbation to non-effect baseline, such as negative control or random perturbation units

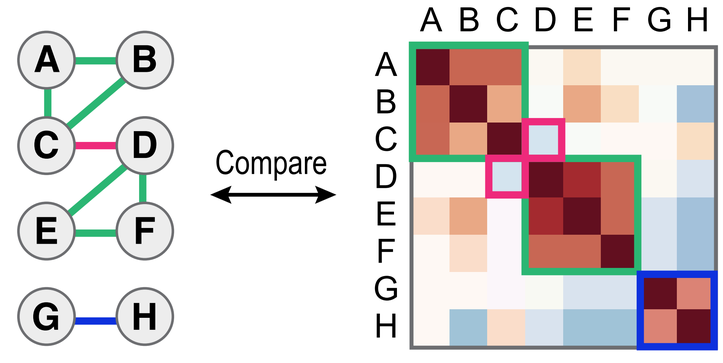


Multivariate Benchmark – Does map recapitulate known biology?

Databases of known relations

- Protein complexes
- Protein–protein interactions
- Pathways





EFAAR choices

	Genome-Wide Perturb-Seq (GWPS)[1]	
	PCA pipeline	**scVI[2] pipeline**
Align	Compute z-score according to the distribution of non-targeting controls (NTC) in the same batch	Compute scVI[2] embedding w/ 128 latent dimensions
Embed	Compute top 100 PCs	
Align	Center at mean of NTCs	
Aggregate	Compute mean embedding of each perturbation	
Filter	Exclude perturbations without transcriptoprint	
Relate	Pairwise cosine similarity	

	Phenomics (Recursion)	
	Centerscale pipeline	**TVN[3] pipeline**
Embed	Apply, to the well images, a convolutional neural network (CNN) model trained to be partially resilient to batch effects (CNN-BC)	
Align		
Filter	Filter outlier image embeddings	
Align	Center and scale (z-scoring)	Typical Variation Normalization (TVN)[3]
Aggregate	Compute mean embeddings	
Filter	Exclude perturbations without phenoprint	
Relate	Pairwise cosine similarity	

Results

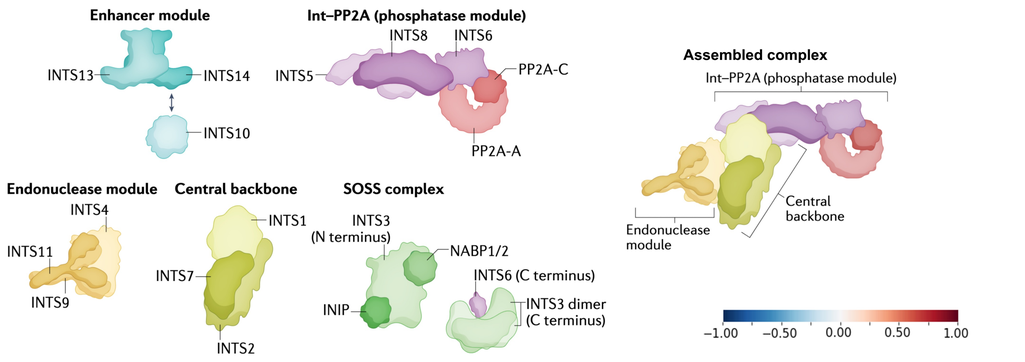
		GWPS	
		PCA	**scVI**
Uni-variate	Consistency	52%	**61%**
	Distance	51%	**53%**
Multi-variate	Reactome pairwise recall	18.6%	**19.8%**
	SIGNOR pairwise recall	12.4%	**12.9%**
	CORUM clusters recall	30.2%	**33.0%**
	Reactome clusters recall	16.1%	**16.9%**
	SIGNOR clusters recall	11.4%	**13.3%**

		Phenomics	
		Centerscale	**TVN**
Uni-variate	Consistency	-1.9%	**+100%**
	Distance	+5.8%	**+109%**
Multi-variate	Reactome pairwise recall	+14.5%	**+82%**
	SIGNOR pairwise recall	-4.9%	**+55%**
	CORUM clusters recall	+18.2%	**+107%**
	Reactome clusters recall	+6.5%	**+63%**
	SIGNOR clusters recall	-11.5%	**+53%**

* Relative to CNN-BC baseline

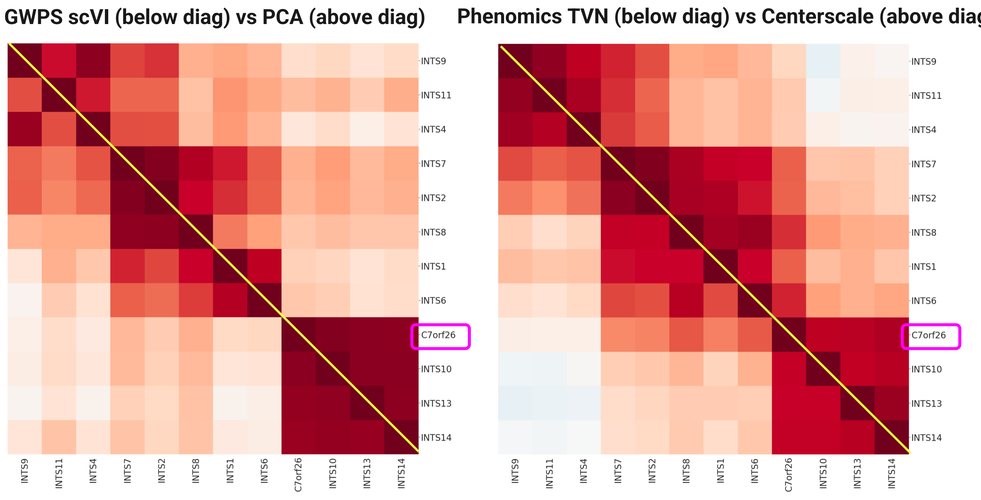
Example: Integrator complex known and novel biology[3]

C7orf26 was officially renamed INTS15 in Jan 2022.



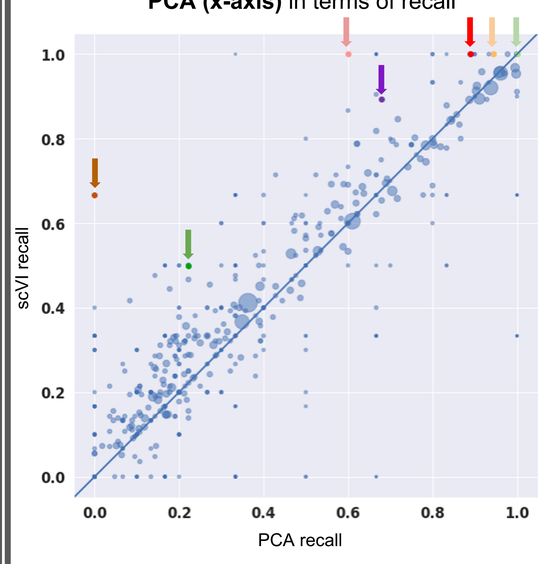
GWPS scVI (below diag) vs PCA (above diag)

Phenomics TVN (below diag) vs Centerscale (above diag)



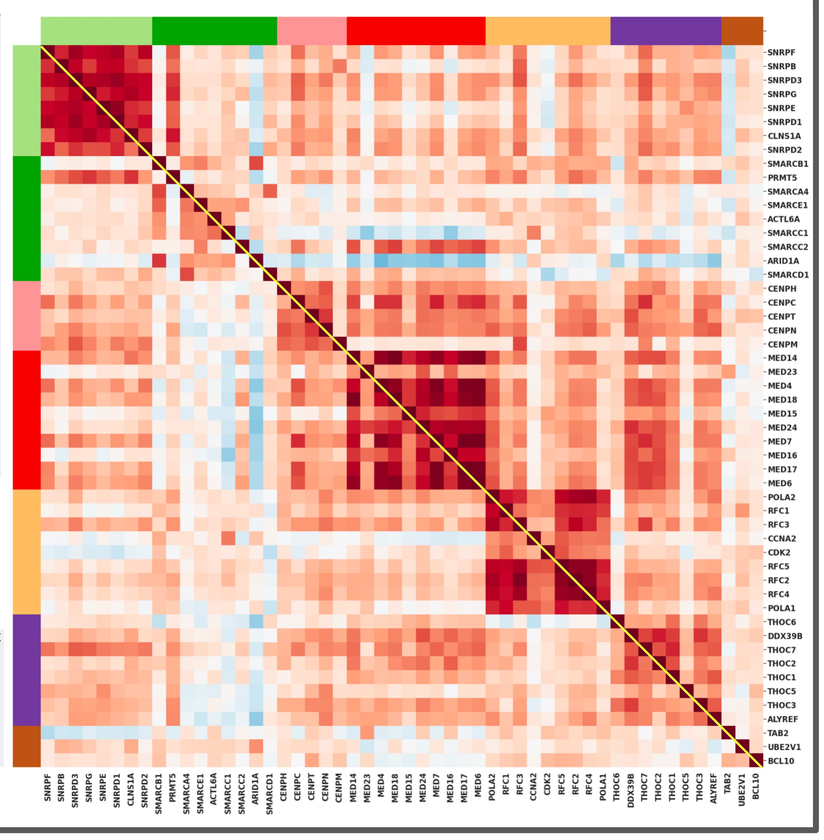
CORUM cluster recapitulation for two EFAAR pipelines in GWPS data[1]

Cluster-level comparison of **scVI (y-axis)** vs **PCA (x-axis)** in terms of recall



Each dot represents a complex, and marker sizes represent cluster sizes. scVI performs better for more clusters, consistent with the summary metrics.

Cosine similarity of complexes that are identified with highest recall by **scVI (below diag)** than the alternative pipeline **PCA (above diag)**.



References

1. Replogle, et al. "Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq". Cell, 2022.
2. Lopez et al. "Deep Generative Modeling for Single-cell Transcriptomics". Nature Methods 2018.
3. Ando et al. "Improving Phenotypic Measurements in High-Content Imaging Screens". BioRxiv 2017.
4. Welsh & Gardini. "Genomic regulation of transcription and RNA processing by the multitasking Integrator complex". Nature Reviews Molecular Cell Biology, 2022.

Recursion references: rxrx.ai, rxrx.ai/lmrl21