Learning Chemical Sensitivity to Understand Disease Mechanisms

by
William Connell

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Pharmaceutical Sciences and Pharmacogenomics

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
Hani Goodarzi
FDD44359FCC6487...

Hani Goodarzi
Chair

DocuSigned by:
Michael J Keiser

Michael J Keiser

DocuSigned by:
LUKE GILBERT
71F73C69F83C48B...

LUKE GILBERT

Committee Members

Acknowledgments

My parents Sally and Brian are responsible for giving me the advantages of experience, education, and preparation. They have always encouraged me and assisted me in developing into the person I am today. They have done so much for me, and I sincerely appreciate it. Without their help, I could not have grown into leading a balanced life full of happiness. They taught me the work ethic required to accomplish difficult things when others depend on you. In doing so, I learned the joy of a challenge and the unexpected ways you learn and grow when committed to a long-term goal. Completing my graduate degree is one such example of that, but those lessons extend to the friendships and relationships I have made as well.

The knowledge my parents have shared with me comes from their own incredible parents, who established a pattern of virtue and kindness in all of their kin. Bill and Carol Wuesthoff inspired my desire to seek intellectual and personal autonomy, something that is directly tied to pursuing a scientific career. Allan and Nancy Connell taught me how to balance self-assuredness with humility to navigate life's everyday challenges. My sisters Hunter and Regan have shared their own lessons with me. Hunter has provided me with career wisdom and mentorship from her own experiences. Regan has shown me how to accept life gracefully in tough and easy times. Both share their unconditional love.

Along with family, I am fortunate to have many friends who genuinely care for me. They have broadened my outlook on life and introduced me to new ways of feeling and thinking. During my time working on the beginning of my Ph.D. I had incredible support and genuine interest from formative friends: Shawyan Yazdanmehr, Brandon Dang, Matt Wasilewski, Tori

McRae, Kyle Cuenin, Nyla Leonardi, Kacie Ring, Marina Kosturos, and Sam Sambado. Not only did these friends fill my life with adventure and joy, but Marina and Kyle introduced me to Jenny Hays. Jenny opened a world of new friendships for me, full of humor and intellect. Some of the important people to me here include Dani Jarvis, Ryan Montana, Carley Halsey, Chris Dunn, Fiona Dewar, Paul Joseph Corbae, Jenna Koopman, and Johannes Ferstad. Along the way, old friends have reappeared and reinforced the foundation of support my life is built on. I owe almost a lifetime of shared experience, mutual understanding, adventures, and true Bay Area sports enthusiasm to Tanner Pierce, Scott Webb, Roman Alcocer, and Justin Connolly.

I owe a large portion of these great friendships to my partner, Jenny Hays. Jenny has shared her friends, family, love, and food with me. Not only does her optimism make each day positive, but her groundedness puts all the useless things into perspective and highlights the most important things. Thank you for enduring my long-winded explanations and for supporting my career goals with genuine faith.

My friends and colleagues at UCSF have ushered me into scientific maturity. My advisor Michael Keiser trusted me to follow my own intellectual interests. Mike provided the correct balance of guidance and autonomy. On the backend Mike juggles many things to ensure his students can focus on scientific and personal growth, for which I am very grateful. Hani Goodarzi provided me with collaboration, mentorship, and confidence in my career aspirations. Luke Gilbert has been a trusted ear and completely dependable. Wilson Liao helped me to finish my first study in a personally meaningful field. Eugene Palovcak exposed me to my favorite part of science, collaboration. Many other friends and colleagues at UCSF and in the Keiser Lab have been there to see me along the way.

Contributions

Chapter 1 includes contributions from:

Connell, William T., Julie Hong, and Wilson Liao. 2021. “Genome-Wide Association Study of Ustekinumab Response in Psoriasis.” *Frontiers in Immunology* 12: 815121.

Chapter 2 includes contributions from:

Connell, William T., and Michael J. Keiser. 2020. “Predicting Cellular Drug Sensitivity Using Conditional Modulation of Gene Expression.” *Learning Meaningful Representations of Life Workshop, NeurIPS*, December. https://doi.org/10.1101/2021.03.15.435529.

Chapter 3 includes contributions from:

Connell, William T., Umair Khan, and Michael J. Keiser. 2022. “A Single-Cell Gene Expression Language Model.” *Learning Meaningful Representations of Life Workshop, NeurIPS*, October. https://doi.org/10.48550/arXiv.2210.14330.

Learning Chemical Sensitivity to Understand Disease Mechanisms

William T. Connell

Abstract

Patients differ in their response to clinical interventions. Both genetics and the environment contribute to the phenotype of a patient after a chemical intervention such as a drug[1,2]. Well-validated interventions have large population effect sizes, meaning they are valuable for an individual on average[3–5]. However, for complex conditions such as autoimmune disease and cancer, the precise outcomes of chemical interventions remain difficult to predict for an individual patient[6,7]. Here we present methods for predicting chemical response and apply these methods to understand biological mechanisms. We discover a new biomarker associated with psoriasis patient response to the drug ustekinumab and use this biomarker to stratify patient populations at various clinical endpoints. Moving beyond association analyses, we develop a machine learning model that integrates chemical structures and gene expression information to predict cellular responses to hundreds of chemicals, and apply methods to investigate how this model works. Finally, we improve model generalizability by pretraining a new model on massive amounts of gene expression data and then applying it to downstream prediction tasks. This work contributes models that predict biological responses to chemicals and methods to interpret how these models work. We anticipate this work will advance precision medicine by improving the ability to predict how patients respond to drugs. Simultaneously, this work provides an *in silico* platform for screening new biological models against a diverse set of chemical probes. Lastly, our pretrained model may find broad

applications in phenotype prediction tasks ranging from disease risk modeling to drug response prediction.

Table of Contents

INTRODUCTION **1**

CHAPTER 1: GENOME-WIDE ASSOCIATION STUDY OF USTEKINUMAB IN PSORIASIS **6**

1.1 SUMMARY 7

1.2 INTRODUCTION 8

1.3 RESULTS 9

I. Clinical characteristics do not correlate with ustekinumab response *9*

II. rs35569429 is associated with discrete levels of ustekinumab response *10*

III. rs35569429 is associated with ustekinumab response at multiple timepoints *12*

IV. rs35569429 association is replicated in an independent cohort *13*

*V. rs35569429 and HLA-C*06:02 combination stratifies responders and non-responders* *14*

1.4 DISCUSSION 15

1.5 MATERIALS AND METHODS 18

I. Study population *18*

II. Response variables *18*

III. Genome-wide association study *19*

IV. Statistical analysis *20*

V. Power analysis *20*

1.6 ACKNOWLEDGEMENTS 21

1.7 SUPPLEMENTARY MATERIAL 22

CHAPTER 2: LEARNING CHEMICAL SENSITIVITY REVEALS MECHANISMS OF CELLULAR RESPONSE **24**

2.1 SUMMARY 25

2.2 INTRODUCTION 25

2.3 RESULTS 27

I. Conditional modeling enhances cellular drug sensitivity prediction *27*

II. ChemProbe predicts breast cancer patient response *30*

III. ChemProbe predicts differential cellular sensitivity *34*

IV. Gene expression attribution vectors pass model interpretation sanity checks *38*

V. Model interpretation of transcriptomic features reflects compound pharmacology and network biology *39*

VI. Differential attribution analysis reflects ferroptosis biology and nominates novel therapeutic targets *43*

2.4 Discussion 47

2.5 Materials and methods 51

I. Pharmacogenomic dataset *51*

II. ChemProbe architecture, training, and evaluation *52*

III. Predictive modeling baselines *53*

IV. Dose-response modeling *54*

V. Retrospective I-SPY2 analysis *55*

VI. Prospective differential sensitivity predictions *56*

VII. Integrated gradients *56*

VIII. Attribution method sanity check *57*

IX. Attribution similarity analysis *58*

X. Attribution network analysis *58*

XI. Differential attribution analysis *60*

2.6 Supplementary Figures 61

CHAPTER 3: A SINGLE-CELL GENE EXPRESSION LANGUAGE MODEL 68

3.1 Summary 69

3.2 Introduction 69

3.3 Results 71

I. Exceiver accounts for discrete features and technical dropout in scRNA-seq self-supervised pretraining *71*

II. Exceiver learns representations that reflect known biology *73*

III. Pretrained Exceiver transfers to new datasets and predicts drug response *75*

3.4 DISCUSSION 78

3.5 MATERIALS AND METHODS 79

I. Model architecture *79*

II. Pretraining tasks *80*

III. Data and processing *82*

3.6 SUPPLEMENTARY MATERIALS 84

DISCUSSION 86

REFERENCES 90

List of Figures

Figure 1-1. Association analysis design and primary outcome. 10

Figure 1-2. Cohort 1 association analysis results. 12

Figure 1-3. Percent of psoriasis patients achieving PASI thresholds according to genotype. 13

Figure 1-4. Percent of psoriasis patients achieving PASI75 at week 12. 15

Figure 2-1. ChemProbe design and model interpretation. 28

Figure 2-2. I-SPY2 clinical trial retrospective analysis 32

Figure 2-3. Differential potency predictions. 35

Figure 2-4. Validation of differential sensitivity predictions 37

Figure 2-5. Feature attribution analysis of compound nominal targets. 41

Figure 2-6. Differential attribution analysis of ferroptosis-inducing compounds 45

Figure 3-1. Exceiver learns cell embeddings reflecting tissue and compartment. 72

Figure 3-2. Learned gene embedding similarity profiles reflect network biology. 75

Figure 3-3. Pretrained Exceiver encodes a new dataset. 76

List of Tables

TABLE 1-1. COHORT 1, 2 AND COMBINED ASSOCIATION ANALYSIS RESULTS. 11

TABLE 2-1. PREDICTIVE PERFORMANCE OF FEATURE MODULATION MODELS. 30

TABLE 3-1. LATENT SAMPLE REPRESENTATIONS REFLECT BIOLOGICAL ANNOTATIONS. 74

TABLE 3-2. PRETRAINED EXCEIVER PREDICTS DRUG RESPONSE. 78

Introduction

A phenotype is an observable trait of an organism. We largely rely on phenotypic change as an indicator of physical health: disease occurs upon deviation from a healthy baseline. Medical professionals quantify how well a treatment reverses disease by measuring a patient's phenotype before and after an intervention. The magnitude of this effect indicates whether an intervention is beneficial. Well-validated interventions have large population effect sizes, meaning they are valuable for an individual on average[3–5]. However, the precise outcomes of interventions remain difficult to predict on an individual patient basis[6,7]. This is especially evident for targeted chemical interventions, also known as drugs. Precision medicine is the promise of making treatment decisions for an individual in consideration of their genetics and environment[7,8]. To get there, we require computational methods that accurately predict the physical outcome of an intervention. In this work, I develop methods for predicting biological responses to chemicals and investigate the molecular patterns on which these methods rely.

There are several possible ways to model the relationship between molecular biology, chemical intervention, and patient response. Association analyses facilitate a straightforward interpretation of the relationship between molecular features and a phenotype[3]. However, due to the univariate specification of these models, association studies of complex diseases do not extrapolate well to predict outcomes for new patients. On the other hand, multivariate models predict phenotypes for new patients reasonably well but these models are challenging to interpret[9]. In the first chapter of my dissertation, we analyze the association between

genetic variants and psoriasis patient response to the drug ustekinumab. There are several drugs available for the treatment of psoriasis, a complex autoimmune disease[10]. Our analysis discovers genetic biomarkers that may inform clinical drug selection decisions for psoriasis patients. In the second chapter, we build on this line of work and propose a method that uses diverse data modalities to predict cellular response to a large panel of chemical probes. We show *in silico* chemical screening can be used to guide treatment decisions for cancer patients and discover mechanisms of chemical response. In the third chapter, we propose a method for learning general biological patterns that may be applied to various downstream tasks. This model provides a foundation to potentially predict diverse phenotypes, which we examine for drug response. In summary, this work presents methods for learning chemical sensitivity and applies these methods to understand biological mechanisms.

Patients respond differently to the same drug, which we observe as a response distribution. Knowing which genetic features are linked to drug response informs clinical treatment decisions[7]. The field of pharmacogenomics associates genetic features with drug response, supporting treatment decisions for complex conditions such as autoimmune disease and cancer[5,6]. Furthermore, genetic associations with drug response may not only reveal biomarkers but may also suggest disease mechanisms. We applied a genome-wide association study (GWAS) to examine the link between genetic variants and the response of psoriasis patients to the drug ustekinumab. Multiple monoclonal antibody treatments are available for psoriasis treatment, posing a need to support treatment decisions with molecular information[10]. We tested millions of genetic variants for association with response to ustekinumab[11]. Our analysis discovered biomarkers that may help guide treatment decisions

and map the molecular mechanisms underlying psoriasis. Association analyses apply a straightforward model which tests whether a single genetic variant predicts response, yielding an interpretable effect size. Such a univariate model is unlikely to either causally explain or predict a phenotype with high accuracy. These limitations beg methodology improvements that consider multiple features to extend predictive ability to many drugs.

Models that accurately predict the result of an intervention are not only valuable for making patient treatment decisions but also for determining molecular functions. A clinical study exploits natural biological variance to make response associations. A laboratory experiment, on the other hand, engineers biological changes in order to test for differences in response. In chemical biology, compounds with known mechanisms are used as tools to assess the functions of molecular components. These chemicals are known as probes because they intervene in defined biological processes and can therefore reveal dependence on a known mechanism[12]. Such an experiment measures the response of a normal and engineered biological model to a chemical probe. The altered biological component is involved in the chemical mechanism if there is a difference in response. Researchers often apply this type of experiment in the lab; however, scaling this approach to many biological models and chemical probes is prohibitively resource intensive. To provide access to large scale chemical probing, we built a model that accurately predicts the response of new biological models to a set of 545 diverse chemical probes and drugs[13]. This model, ChemProbe, combines transcriptome profiles and chemical structures to make accurate phenotype predictions. ChemProbe successfully predicted drug responses both retrospectively for patients in clinical trials and prospectively for new cell lines *in vitro*. To highlight the application of this model for discovery

research, we screened a genetically modified cell line against hundreds of chemical probes and validated a resistance function of the altered gene. Not only can our model support clinical treatment decisions, but ChemProbe can be applied to screen molecularly altered cell lines to discover mechanisms of chemical sensitivity and resistance.

It is relatively difficult to interpret which of many features a complex model, such as a neural network, relies on to make predictions. With ChemProbe we overcame this shortcoming by rigorously evaluating a model interpretation method. Our analysis revealed ChemProbe relies on similar gene expression patterns to make sensitivity predictions for compounds with the same target. Moreover, similarity analysis of model interpretations reflected network biology and molecular ontologies. ChemProbe's *in silico* chemical screening paradigm allows users to test the response of new biological models and patient tumors to many chemical probes, as well as understand how predictions are made. ChemProbe is still limited to predicting response to a finite set of chemical interventions. To advance precision medicine we require models that generalize to diverse phenotype prediction tasks.

Models that successfully predict many complex phenotypes will advance patient treatment. Current machine learning paradigms are relatively narrow in application: a model trained on a curated dataset will only apply to data from the same narrow distribution[14]. Supervised models inevitably fail when challenged with a distribution shift or a new task[15]. Some of this brittleness stems from the difficulty in obtaining enough labeled examples to train supervised models. Recent work in natural language processing (NLP) and computer vision (CV) leverages massive, unlabeled datasets to train models that learn general patterns and can be

“fine-tuned” on smaller, labeled datasets[16–18]. In fact, practitioners can reliably fine-tune these models on a diverse selection of tasks and increase predictive performance[19]. This approach, based on self-supervised learning, has found recent applications in biology, most notably in protein language modeling[20–23]. We hypothesized that training a language model on atlas-level single-cell RNA sequencing data would enable transfer learning on a range of phenotype prediction tasks. We built Exceiver, an attention-based neural network and formulated a novel self-supervised task for continuous values to train our model[24]. We found the latent representations of both individual cells and learned gene embeddings reflected biological annotations. Exceiver also learned expression patterns from a new dataset much more quickly than a model trained from scratch. We tested Exceiver in a transfer learning scenario and found robust generalization to several drug response prediction tasks. Exceiver helps to generalize our previous modeling efforts by capturing gene expression patterns useful for many different downstream tasks. As molecular measurement platforms continue to scale, computational methods that leverage unlabeled datasets will be essential for building general predictive systems for biology.

Chapter 1: Genome-wide association study of ustekinumab in psoriasis

1.1 Summary

Heterogeneous genetic and environmental factors contribute to the psoriasis phenotype, resulting in a wide range of patient response to targeted therapies. Here, we investigate genetic factors associated with response to the IL-12/23 inhibitor ustekinumab in psoriasis. To date, only HLA-C*06:02 has been consistently reported to associate with ustekinumab response in psoriasis. Genome-wide association testing was performed on the continuous outcome of percent change in Psoriasis Area Severity Index (PASI) at 12 weeks of ustekinumab therapy relative to baseline. A total of 439 European ancestry individuals with psoriasis were included (mean age, 46.6 years; 277 men [63.1%]). 310 (70.6%) of the participants comprised the discovery cohort and the remaining 129 (29.4%) individuals comprised the validation cohort. Chromosome 4 variant rs35569429 was significantly associated with ustekinumab response at 12 weeks at a genome-wide significant level in the discovery cohort and replicated in the validation cohort. Of psoriasis subjects with at least one copy of the deletion allele of rs35569429, 44% achieved PASI75 (75% improvement in PASI from baseline) at week 12 of ustekinumab treatment, while for subjects without the deletion allele, 75% achieved PASI75 at week 12. We found that differences in treatment response increased when rs35569429 was considered alongside HLA-C*06:02. Psoriasis patients with the deletion allele of rs35569429 who were HLA-C*06:02 negative had a PASI75 response rate of 35% at week 12, while those without the deletion allele who were HLA-C*06:02 positive had a PASI75 response rate of 82% at week 12. Through GWAS, we identified a novel SNP that is potentially associated with response to ustekinumab in psoriasis. Additional studies are needed to confirm these findings.

1.2 Introduction

Psoriasis is a common, chronic immune-mediated skin disease that affects at least 2% of the population worldwide[25]. Psoriasis is associated with psoriatic arthritis, cardiovascular disease, metabolic syndrome, and other comorbidities, which makes effective management of psoriasis critical. Moderate-to-severe psoriasis is treated with phototherapy and systemic agents, including targeted biologic inhibitors of TNF-α, IL-12/23, IL-17, and IL-23. Patient responses to biologic therapy can vary widely, from poor overall response to gradual loss of therapeutic sensitivity[26]. Response differences are largely influenced by patient weight and adherence, drug dose and bioavailability, and pharmacokinetic covariates, such as drug immunogenicity[27]. The molecular heterogeneity of psoriasis may also contribute to differential therapeutic responses. However, there are no molecular biomarkers routinely used in clinical practice to facilitate selection of the therapies tailored to individual patients.

Ustekinumab is a fully humanized immunoglobulin monoclonal antibody targeting the p40 subunit shared by IL-12 and IL-23. Phase 3 clinical trials showed that treatment with ustekinumab results in 75% improvement in the Psoriasis Area and Severity Index (PASI75) in ~66% of patients after 12 weeks of therapy[28–30]. Candidate gene studies have identified the HLA-C*06:02 allele as being associated with better ustekinumab responses in both European and Chinese patients with psoriasis[31–34]. A meta-analysis of eight studies including 1048 psoriasis patients showed that HLA-C*06:02 positive patients had a median PASI75 response rate of 92% after 6 months of ustekinumab therapy compared to a median PASI75 response rate of 67% in the HLA-C*06:02 negative patients[35].

Here, we performed an unbiased genome-wide association study (GWAS) to evaluate if additional genetic factors were associated with ustekinumab response. We evaluated our findings across multiple response timepoints and in conjunction with HLA-C*06:02. Our findings highlight a potentially novel variant associated with ustekinumab response in psoriasis, which may further facilitate the development of precision medicine approaches.

1.3 Results

I. Clinical characteristics do not correlate with ustekinumab response

In this study, we analyzed genetic data from two cohorts of psoriasis patients receiving ustekinumab. Following preprocessing and filtering for individuals of European genetic ancestry, the discovery cohort (cohort 1) totaled 310 individuals (181 males, 171 females) and the validation cohort (cohort 2) totaled 129 individuals (82 males, 47 females) (**Figure 1-1**). The average PASI score at baseline was 18.6 for cohort 1 and 18.8 for cohort 2 (**Supplementary Table 1-1**). Power analysis revealed the discovery cohort had 1-β>0.75 for MAF>0.05 and ES>7. The replication cohort had 1-β>0.75 for MAF>0.05 and ES>5 (**Supplementary Figure 1-2**). We used linear regression to perform genome-wide association testing on the percent improvement in PASI response at week 12 of ustekinumab therapy compared to baseline (**Figure 1-1**). There was no correlation between age, BMI, and duration of the disease with the primary outcome of percent PASI improvement, and so these clinical variables were not

included as covariates in the linear regression model (**Supplementary Figure 1-1**).

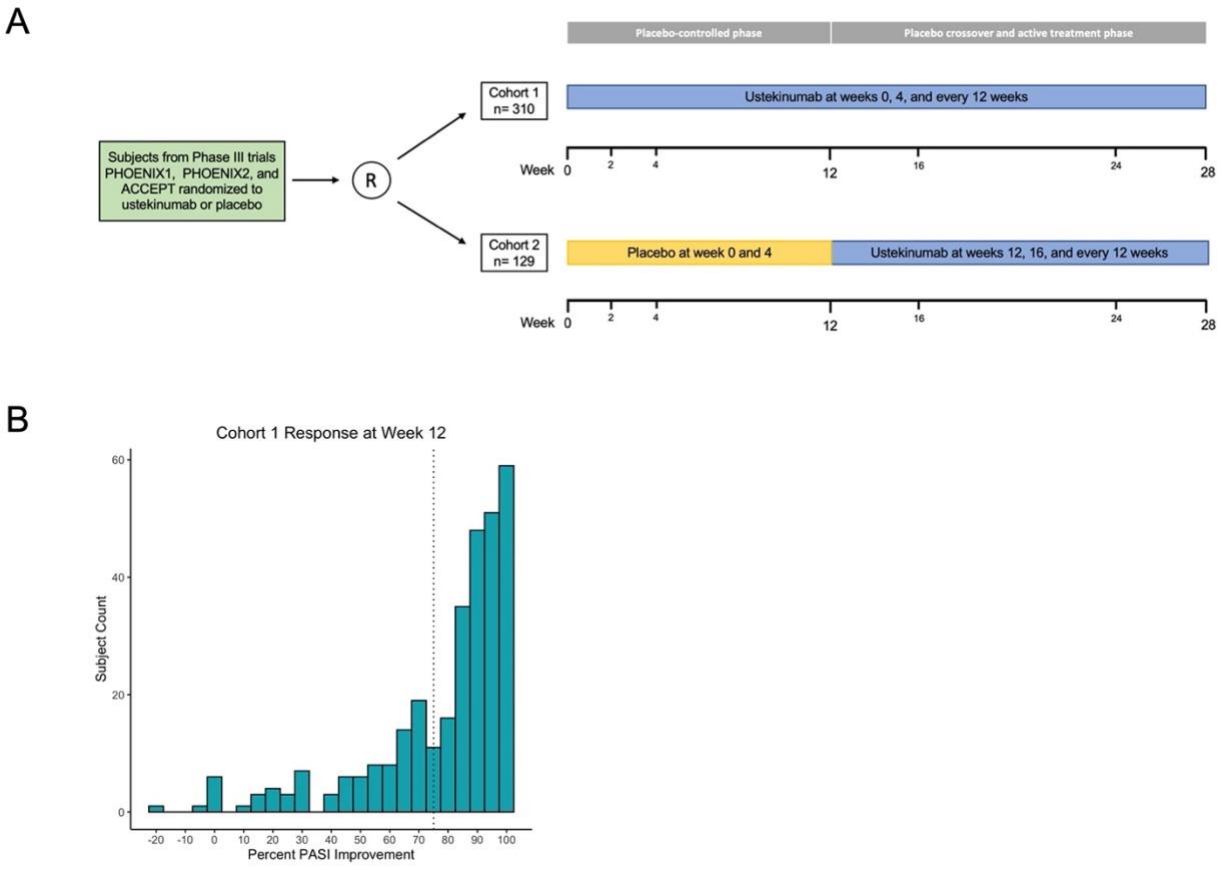


Figure 1-1. Association analysis design and primary outcome.

- **(a)** Association analysis design and primary outcome. Phase 3 clinical trial comprise discovery and validation cohorts
- **(b)** Histogram of cohort 1 percent PASI improvement at week 12; dashed line marks 75% improvement threshold

II. rs35569429 is associated with discrete levels of ustekinumab response

Genome-wide association testing of subjects in cohort 1 identified a single peak on chromosome 4 exceeding a genome-wide significance threshold of $p<5\times10^{-8}$ lead by rs35569429 (β, -19.84; 95% CI, -26.58 to -13.1; $p=1.98\times10^{-8}$) (**Figure 1-2, Table 1-1**). Directly genotyped SNP rs11722643 was in high linkage disequilibrium with imputed SNP rs35569429

and achieved a similar level of significance (R^2, 0.9; β, -19.31; 95% CI, -26.33 to -12.29; p=1.44×10^{-7}). To determine whether multiple SNPs contributed to the peak on chromosome 4, we performed conditional analysis on rs35569429. The conditional analysis completely attenuated the GWAS peak, indicating a single independent signal at this locus (**Figure 1-2**). The major allele of rs35569429 is "G" while the minor allele is a single nucleotide deletion of G, denoted as "Del". Subjects with at least one minor allele were labeled as the deletion positive group (Del+, N=55), and subjects with zero minor alleles were labeled the deletion negative group (Del-, N=255). Only one subject was homozygous for the minor allele. To understand the impact of this SNP at various discrete levels of PASI response, we examined the proportions of Del- and Del+ individuals who achieved PASI50, PASI75, PASI90, and PASI100 at Week 12. We found that in the Del- group, 235/255 (92.2%) achieved PASI50, 191/255 (74.9%) achieved PASI75, 121/255 (47.5%) achieved PASI90, and 48/255 (18.8%) achieved PASI100 at Week 12. In the Del+ group, 39/55 (80.9%) achieved PASI50, 24/55 (43.6%) achieved PASI75, 12/55 (21.8%) achieved PASI90, 5/55 (9.1%) achieved PASI100 at Week 12.

Table 1-1. Cohort 1, 2 and combined association analysis results.
Abbreviations: MAF, mean allele frequency.

	SNP	**MAF**	β	***P* value**
Cohort 1	rs35569429	0.090	-19.84	1.98E-08
Cohort 2	rs35569429	0.097	-6.71	0.042
Cohort 1+2 Analysis	rs35569429	0.092	-15.83	2.42E-09

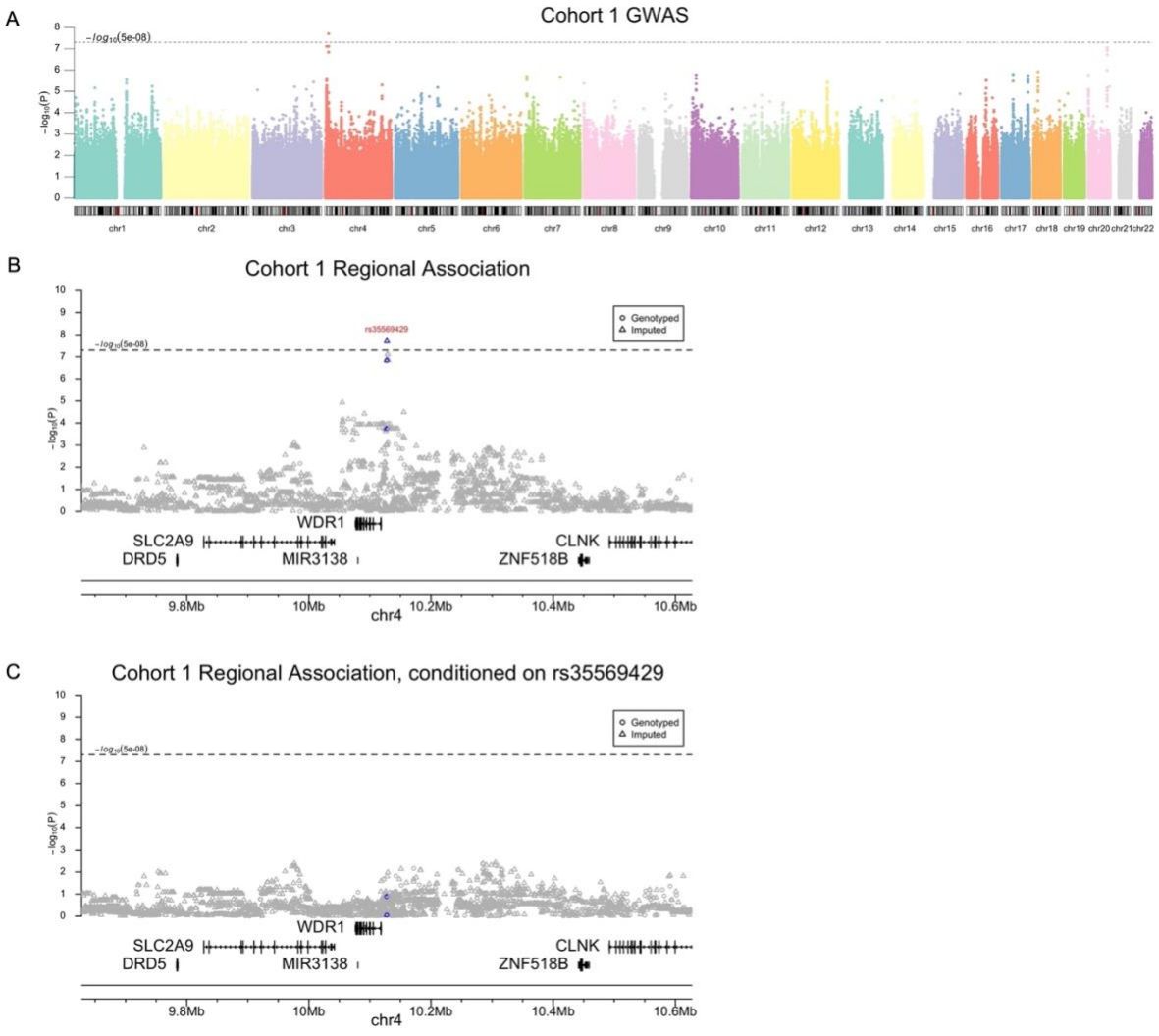


Figure 1-2. Cohort 1 association analysis results.

- **(a)** Genome-wide association Manhattan plot. Blue indicates variants in high linkage disequilibrium (R^2>0.95) with rs35569429.
- **(b)** Regional Manhattan plot.
- **(c)** Conditional association analysis Manhattan plot.

III. rs35569429 is associated with ustekinumab response at multiple timepoints

To further investigate the validity of rs35569429, we analyzed its association with PASI outcomes in cohort 1 at timepoints that were not part of the original GWAS analysis (i.e.

timepoints other than week 12). We found that a greater proportion of individuals in the Del- group achieved PASI75 compared to the Del+ group at Week 2 (1.57% vs 0%), Week 4 (17.6% vs 10.9%), Week 24 (76.5% vs 61.8%), and Week 28 (73.3% vs 52.7%) (**Figure 1-3**). Similarly, the Del- group had a higher proportion of individuals achieving PASI50, PASI90, and PASI100 than the Del+ group at weeks 2, 4, 24, and 28. The difference in PASI responses between Del- and Del+ groups were generally comparable if not greater than the difference in PASI responses between HLA-C*06:02 positive and HLA-C*06:02 negative individuals (**Figure 1-3**), where HLA-C*06:02 represents a previously well-validated locus associated with ustekinumab response[35].

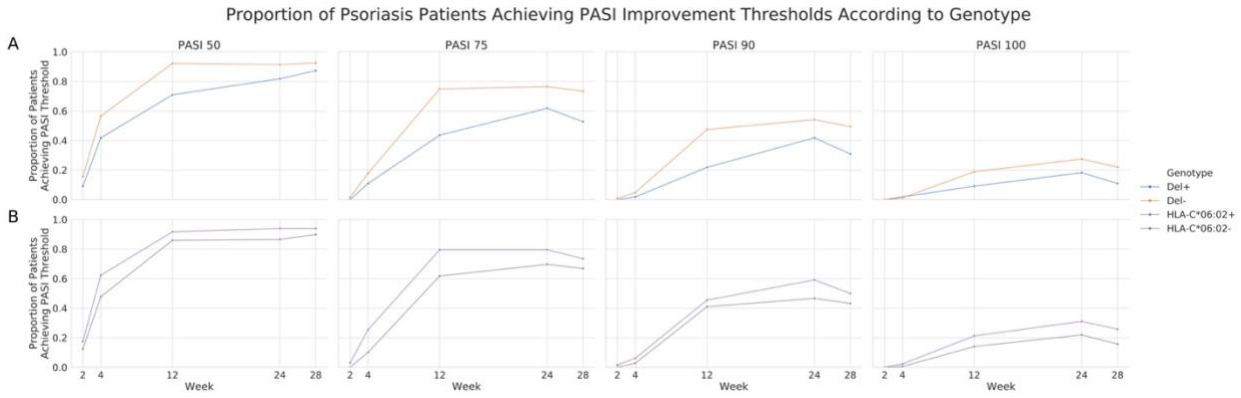


Figure 1-3. Percent of psoriasis patients achieving PASI thresholds according to genotype.

- **(a)** PASI 50, 75, 90 and 100 achievement across weeks 2, 4, 12, 24 and 28 for rs35569429 genotype.
- **(b)** PASI achievement across weeks for HLA-C*06:03 genotypes.

IV. rs35569429 association is replicated in an independent cohort

We next investigated the association of rs35569429 with response to ustekinumab in an independent cohort 2. We found the same direction of effect at week 12 for rs35569429 (β, -

6.71; 95% CI, -13.13 to -0.30; *P*=0.042) (**Table 1-1**). In the Del- group, 102/106 (94.5%) subjects achieved PASI50, 81/106 (76.4%) subjects achieved PASI75, 45/106 (42.5%) achieved PASI90, and 26/106 (24.5%) achieved PASI100 at Week 12. In the Del+ group, 20/23 (87.0%) subjects achieved PASI50, 13/23 (56.5%) achieved PASI75, 9/23 (39.1%) achieved PASI90, and 2/23 (8.7%) achieved PASI100 at Week 12. Association testing for rs35569429 in cohort 1 and cohort 2 combined at week 12 yielded a genome-wide significant result (β, -15.83; 95% CI, -20.72 to -10.74; $p=2.42\times10^{-9}$). We ran a sensitivity analysis on the full sample of cohorts 1 and 2 combined at week 12. We observed the expected genome-wide significant peak at rs35569429, with the most significant SNP being rs11722643, which is in high linkage disequilibrium with rs35569429 (R^2, 0.88; β, -16.64; 95% CI, -22.04 to -11.25; $p=3.25\times10^{-9}$). We also observed a single additional genome-wide significant loci on chromosome 14, which could not be further confirmed (rs994384156; β, -14.94; 95% CI, -20.02 to -9.86; $p=1.58\times10^{-8}$).

V. rs35569429 and HLA-C*06:02 combination stratifies responders and non-responders

Finally, we explored how the combination of rs35569429 and HLA-C*06:02 affects PASI75 response in cohort 1 and 2 at week 12, since HLA-C*06:02 is an allele previously established to be associated with a more favorable responses to ustekinumab in psoriasis (11). In cohort 1 at week 12, 82.4% Del-/HLA-C*06:02+ individuals achieved PASI75 compared to 68.8% in Del-/HLA-C*06:02-, 61.1% in Del+/HLA-C*06:02+, and 35.1% in Del+/HLA-C*06:02- (**Figure 1-4**). In cohort 2 at week 12, 88.6% Del-/HLA-C*06:02+ individuals achieved PASI75 compared to 79.2% in Del-/HLA-C*06:02-, 72.7% in Del+/HLA-C*06:02+, and 50.0% in Del+/HLA-C*06:02-. In cohort 1 and cohort 2 combined at week 12, 84.4% Del-/HLA-C*06:02+ individuals

achieved PASI75 compared to 71.6% in Del-/HLA-C*06:02-, 65.5% in Del+/HLA-C*06:02+, and 38.8% in Del+/HLA-C*06:02.

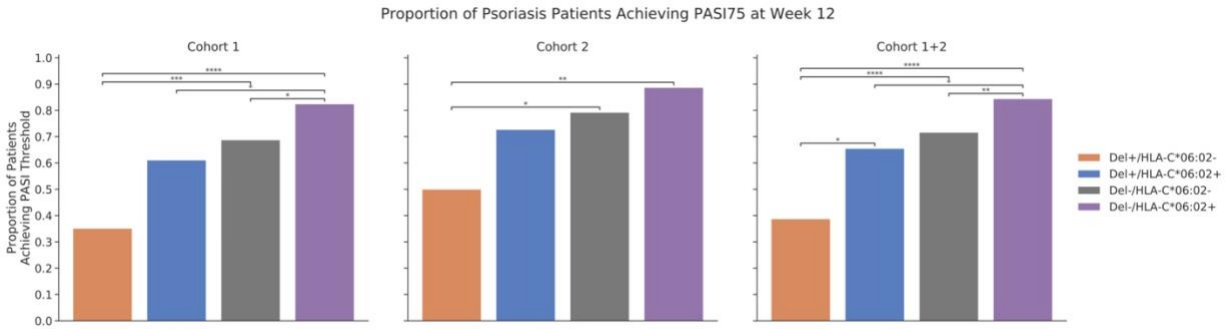


Figure 1-4. Percent of psoriasis patients achieving PASI75 at week 12.
*p<=5×10-2; **p<=1×10-2; ***p<=1×10-3; ****p<=1×10-4.

1.4 Discussion

This genetic association study found a genome-wide significant association between intergenic variant rs35569429 and response to ustekinumab for the treatment of moderate to severe psoriasis. In our primary association analysis, absence of the minor allele (Del-) was significantly associated with a larger PASI improvement at 12 weeks from baseline. More favorable PASI responses in Del- individuals compared to Del+ individuals were also observed at weeks 2, 4, 24, and 28. The association of rs35569429 with ustekinumab response was validated in an independent cohort of psoriasis patients. Conditional analysis revealed a single independent signal at the locus of interest.

rs35569429 is characterized by a G deletion minor allele. This variant is located in an intergenic region 9 kB upstream of *WDR1*. Functional analysis by GeneHancer Regulatory

Elements strongly associates a 10.6 kB region (GH04J010114) 1.2 kB downstream of this variant with promoter/enhancer activity influencing proximal protein coding genes *WDR1* and *SLC2A9*[36]. The WDR1 protein is involved in actin filament disassembly, a critical process of cytoskeleton dynamics, especially in highly motile and interacting immune cells[37]. Impaired actin dynamics as a result of WDR1 deficiency have been causally linked to primary immunodeficiencies and autoinflammatory phenotypes[38,39]. SLC2A9 is a transporter mainly expressed in the kidneys and primarily involved in urate reabsorption. Mutations of *SLC2A9* lead to poor reabsorption and Renal Hypouricemia type-2, as caused by increased urate excretion[40]. Future studies are needed to fine-map the causal and functional SNPs in linkage disequilibrium with rs35569429.

Stratification of ustekinumab responses was greatest when rs35569429 was considered in combination with HLA-C*06:02. Individuals who were Del-/HLA-C*06:02+ achieved PASI75 84.4% of the time, while those were Del+/HLA-C*06:02- achieved PASI75 38.8% of the time, a more than two-fold difference.

Pharmacogenomics continues to play an increasingly important role in precision medicine for dermatology. In 2018, five dermatologic drugs had clinically actionable pharmacogenomic tags that either require or advise testing of genomic biomarkers before treatment[41]. Single FDA-approved biomarkers currently dominate this list; however, multi-gene marker panels will continue to gain importance for informing clinical decisions. Understanding the role of multiple SNPs in disease pathogenesis is important in advancing precision medicine.

Conclusions from this study are limited due to the moderate sample size of the discovery and replication cohorts; our study was not powered for detection of small to moderate effects. Given the polygenicity of complex autoimmune diseases such as psoriasis, in the future, prospective design of large study cohorts are essential for thorough investigation of the biology contributing to therapeutic response. In general, validation in additional, independent cohorts will provide evidence with respect to the genomic signals discovered herein. Furthermore, the index SNP rs35569429 requires further investigation to identify the causal variant(s) associated with this locus and further characterization of functional effects on psoriatic response to ustekinumab.

1.5 Materials and methods

I. Study population

This study involved analysis of individuals with moderate to severe psoriasis who participated in at least one of three placebo-controlled randomized clinical trials: PHOENIX I, PHOENIX II, and ACCEPT[28,29,42]. Participants were originally approached for retrospective collection of DNA samples by investigators analyzing the association between the HLA-C*06:02 allele and response to IL-12/23 inhibition[31]. In total, 439 patients of European descent were used to assess genetic associations between ustekinumab treatment and response.

The GWAS discovery cohort consisted of 310 individuals who were treated with 45mg (n=146) or 90mg (n=164) of ustekinumab for 40 weeks, with the lower or higher dose given according to body weight less than or greater than 100 kg, respectively. The validation cohort consisted of 129 trial participants who crossed-over from placebo to ustekinumab treatment at week 12 and continued ustekinumab for 16 weeks, again dose-stratified by body weight (45 mg: n=64; 90 mg: n=65). In both cohorts, ustekinumab was given with two loading doses 4 weeks apart and every 12 weeks thereafter (**Figure 1-1**).

II. Response variables

In the ustekinumab phase 3 clinical trials, the primary endpoint was achievement of PASI75 at week 12. PASI75 is a binary outcome converted from percent PASI improvement from baseline. To maximize power for the GWAS, we focused on the continuous outcome measure of percent PASI improvement from baseline to 12 weeks after ustekinumab therapy. Phenotypic

response to ustekinumab was recorded at weeks 2, 4, 12, 28, and 40 for the majority of patients in the discovery cohort (cohort 1). In order to validate findings, the placebo to ustekinumab cross-over patients acted as a validation cohort (cohort 2). PASI responses for cohort 2 were measured after 12 weeks of ustekinumab therapy compared to trial start.

III. Genome-wide association study

Genotyping was performed using Illumina HumanOmni2.5-8 v1.2 BeadChips. Imputation was performed using the Michigan Imputation Server (https://imputationserver.sph.umich.edu/index.html)[43] The 1000 Genomes Phase 3 data was used as a reference panel for imputation[44]. Files were converted to PLINK (v1.9) format, which along with R (v3.5.1) and python (v3.7.4), was used for data manipulation, visualization, and association analysis. Quality control and population stratification was performed following methods outlined by Marees et al.[45]. Single nucleotide polymorphisms (SNPs) and individuals with missingness greater than 2% were removed. Duplicate, non-biallelic, and poor imputation quality (R^2<0.7) SNPs were filtered. Non-autosomal SNPs with a low minor allele frequency (MAF<0.05) and significant deviation from Hardy-Weinberg equilibrium (p<1×10^{-6}) were removed. In total 6,799,417 SNPs passed quality control, of which 1,696,820 were directly genotyped. Individuals with a heterozygosity rate +/-3 standard deviation from the mean were filtered, as well as the individual with the lowest call rate within a pair of cryptically related individuals ($\hat{\pi}$>0.2). In total, 310 individuals (181 males, 129 females) passed quality control. The previously described quality control steps were applied to the 1000 Genomes Phase 3 data prior to merging with cohort data for population stratification. Multidimensional scaling (MDS)

was applied to the merged genotype information. The presence of ethnic outliers was evaluated by qualitative alignment with the European superpopulation cluster along the top 2 MDS components. We included the top 10 MDS components as covariates in linear regression models for association testing.

IV. Statistical analysis

A threshold of $p<5\times10^{-8}$ was established in the discovery cohort to determine the associated markers for further replication. We took linkage-disequilibrium into account when interpreting multiple significant association results from the same region. Clumping was employed to greedily assign groups around index variants with $p<5\times10^{-6}$. Variants with an $R^2>0.5$ and less than 1MB away were assigned representation by the index variant. We modeled the additive effect of allele dosage with the quantitative phenotype of interest using linear regression. When considering cohort 1 index variants in replication analyses, a 2-sided *t*-test with $p<0.05$ was considered statistically significant. A two-sided normal test for proportions (p <0.05) was applied to assess PASI threshold achievement differences based on genotype. The combined cohort association study followed the same procedures outlined for analysis of discovery cohort results.

V. Power analysis

We performed power calculations for the discovery and replication cohorts assuming an additive linear model for our quantitative trait of interest. Each power calculation was performed under consideration of the established type 1 error rates for the respective cohort

(cohort 1 α, 5×10^{-8}; cohort 2 α, 5×10^{-2}). We examined power across a range of MAF (0.05-0.25) and effect sizes (ES) (1-9). The genpwr (v1.0.4) R package was used for all calculations.

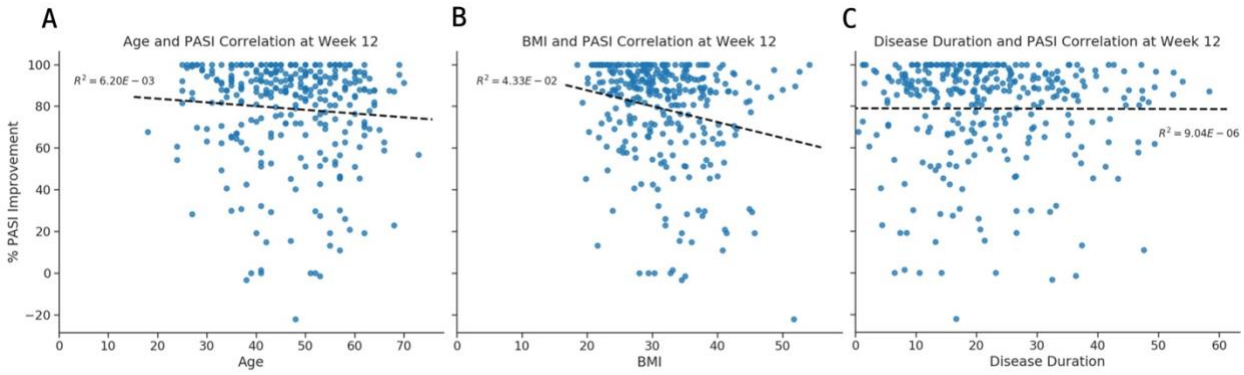
1.6 Acknowledgements

We would like to acknowledge Anne Fourie, Xuejun Liu, Amy Hart, and Dawn Waterworth at Janssen Research & Development, LLC for providing access to the data and for providing analysis feedback. We would also like to acknowledge Joanne Nititham for assistance with data handling. Janssen provided the genotyping and clinical response data for the study, but was not involved in the study design, data analysis, data interpretation, manuscript writing, or decision to submit the manuscript for publication. W.L has received research grant funding from Abbvie, Amgen, Janssen, Leo, Novartis, Pfizer, Regeneron, and TRex Bio. W.C and J.H. have no conflicts of interest to declare. W.L had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

1.7 Supplementary material

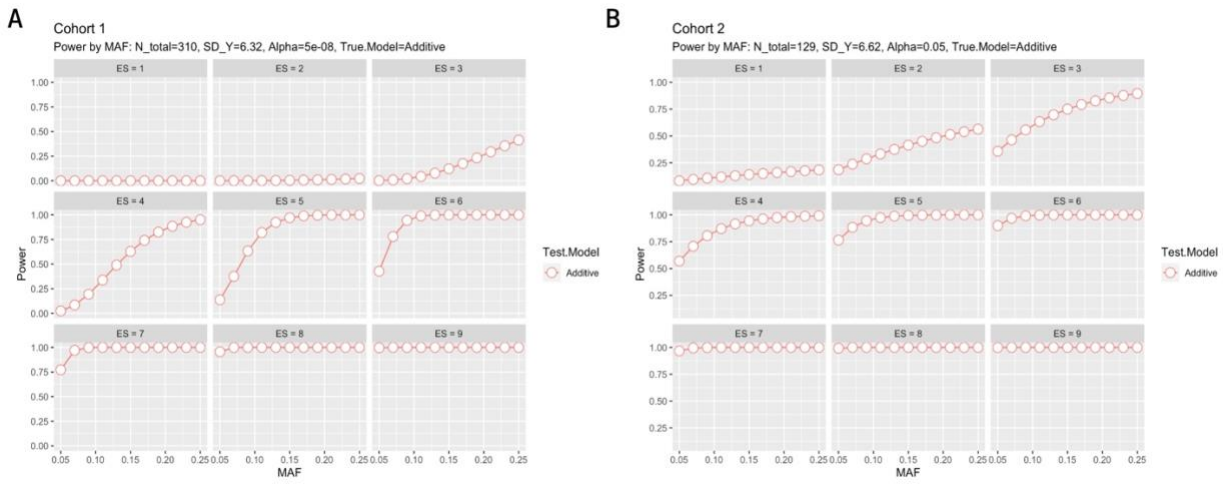
Supplementary Table 1-1. Baseline PASI scores.
Average and standard deviation of baseline PASI for cohort 1, 2 and combined.

	Baseline avg week 0	STD
Cohort 1	18.6	6.32
Cohort 2	18.8	6.62
Cohort 1+2	18.6	6.40



Supplementary Figure 1-1. Correlation between cohort 1 clinical covariates and PASI improvement at week 12.

- **(a)** Age and PASI correlation at week 12. Formula and dashed annotation indicate line of best fit.
- **(b)** BMI and PASI correlation at week 12.
- **(c)** Disease duration and PASI correlation at week 12.



Supplementary Figure 1-2. Power analysis.

(a) Cohort 1 power analysis across a range of effect sizes (1-9) and minimum allele frequencies (0.05-0.25).

(b) Cohort 2 power analysis.

Chapter 2: Learning chemical sensitivity reveals mechanisms of cellular response

2.1 Summary

Drugs elicit different responses in individuals with the same disease phenotype. While complicating drug selection for a tumor, this variability in response can be used to shed light on disease mechanisms. One way to measure response variance is through a diverse chemical screen against a variety of biological models, however, this method is unreasonably resource-intensive. To address this problem, we built a deep learning model that integrates cellular features and chemical structures to predict chemical sensitivity to a library of diverse molecular probes. Using transcriptomic profiles of 842 cell lines and their viability with 545 small-molecules, neural networks trained under a conditional formulation generalize to drug response in clinical tumor samples and laboratory disease models (R^2=0.7113 ± 0.0081). Using deep learning interpretation methods, we found highly attributed transcriptome features reflect known compound mechanisms of action and biological network modules. Finally, we integrate these findings into a new method and nominate genes driving the chemical induction of ferroptosis. This *in silico* cellular screening tool can inform molecular mechanisms of response variance and aid precision oncology treatment decisions.

2.2 Introduction

Chemical probes are small molecules with high potency and selectivity against a known mechanism of action[12]. These tools help define the functions of targeted proteins in biological processes and diseases. Target-specific chemical probes have helped us understand the cell cytoskeleton, immunosuppresion, mTOR signaling, molecular chaperones, and the function of many protein kinases[46,47]. These small molecules have also served as structural

seeds for drug development[12]. Drugs are similarly useful as chemical probes despite the fact that drugs are ultimately optimized for safety, efficacy, and desirable PK/PD properties. Pharmacogenomics leverages clinical observations of drug sensitivity to stratify patient populations with biomarkers and uncover disease mechanisms[3,48].

Complex diseases exhibit a similar phenotype due to a variety of mechanisms. In cancer, genetic alterations that confer a fitness advantage can accumulate into uncontrolled cellular growth over time. Under selection pressure, cancers explore a continuum of molecular heterogeneity throughout an individual's disease progression. Addressing this heterogeneity necessitates precision clinical treatment strategies and research into mechanisms that control disease resistance and sensitivity[49]. As we improve our understanding of gene expression patterns contributing to variance in drug response, we can develop better solutions for cancer patients that exploit specific tumor vulnerabilities.

To better understand complex disease mechanisms and treat patients precisely, it would be ideal to test both new disease models and patient samples against a wide range of chemical probes. In such an experiment, a change in cellular sensitivity in a sample relative to a baseline indicates reliance on the target of a chemical probe. However, screening biological samples against a large library of chemical probes is prohibitively resource intensive. To address this problem, we built a model that predicts the sensitivity of new cellular samples to a panel of chemical probes and drugs.

ChemProbe models chemical sensitivity as a function of basal cellular state, thereby learning gene abundance patterns that result in specific outcomes (**Figure 2-1A**). ChemProbe

generalizes to new biological samples, and the model permits the interpretation of learned gene features relevant to known compound mechanisms. ChemProbe accurately models chemical response without biological priors, enabling *in silico* chemical screening of biological models and mechanistic interpretation of learned gene dependencies.

2.3 Results

I. Conditional modeling enhances cellular drug sensitivity prediction

We hypothesized that a model could learn to combine gene expression with chemical structure features to predict cellular drug sensitivity (**Figure 2-1**). We leveraged publicly available datasets to match cancer model basal cellular transcriptomes to large-scale drug screens. The Cancer Therapeutics Response Portal (CTRP) reports the viability of 842 cancer cell lines in response to 545 compounds and compound pairs across a range of concentrations[50]. These compounds were originally selected based on their diversity of cell circuitry targets. The Cancer Cell Line Encyclopedia (CCLE) provides basal transcriptomic characterizations of all 842 CTRP cell lines[51]. We combined compound structures and concentrations from the CTRP with protein-coding gene transcriptomes from the CCLE to create a dataset of compound-cell line pairs consisting of approximately 5.8 million labeled examples (Methods).

We formulated the cellular drug sensitivity prediction task as a conditional model $y = f(x|n)$, where y is cellular viability, x is a matrix of standardized RNA abundance values, and n is a matrix of chemical features (Methods). Thus the model's prediction of cellular viability depends on a cell's transcriptomic profile in the context of a chemical structure and concentration. In practice, the models achieve this by learning to use chemical features to

modulate gene expression through linear transformations of internal gene expression representations (**Figure 2-1**).

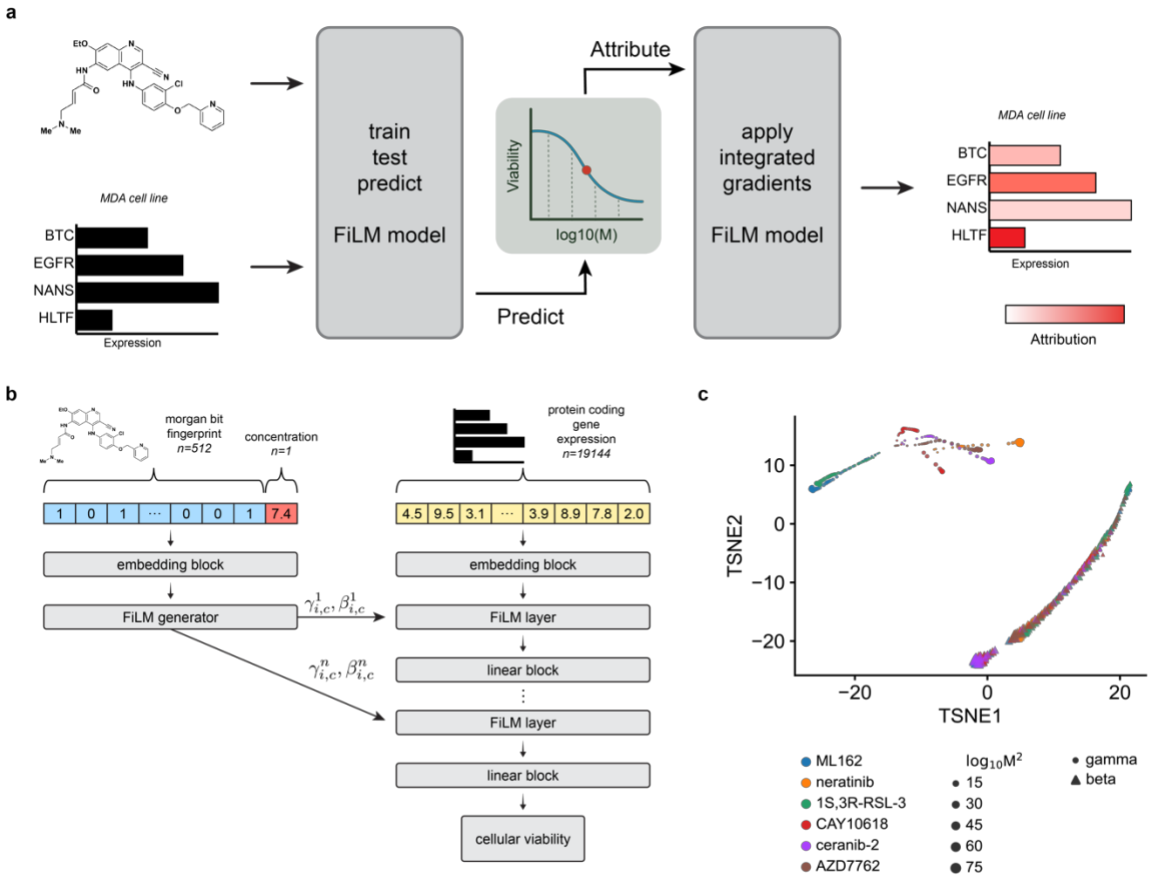


Figure 2-1. ChemProbe design and model interpretation.

(a) Workflow of model training, validation, prediction, dose-response modeling, and feature attribution. We trained a model to predict drug sensitivity at different compound concentrations and fit log-logistic models to predictions. We derived compound pharmacodynamics from dose response curves and applied integrated gradients to predicted IC50 to derive input feature attributions. IC50, inhibitory concentration of compound at 50% cellular viability.

(b) Architecture of the conditional neural network (ChemProbe) that was trained to predict cell line viability from molecular features and compound structure. ChemProbe learns an embedding of protein-coding gene expression features conditioned by parameters learned from an embedding of compound structure and concentration, across multiple neural network layers.

(c) t-SNE decomposition of learned conditioning parameters. Points represent compound-concentration samples; color indicates compound; size indicates concentration; and shape indicates parameter.

We tested several ways to combine cellular features and chemical information within a single model, as assessed by their average maximum coefficient of determination (R^2). Accounting for comparable model sizes, we trained, validated, and hyperparameter-optimized different model architectures on the same data folds (Methods). We compared three methods of learned feature conditioning against a baseline feature concatenation approach[52]. All conditioning approaches outperformed feature concatenation by a notable margin (**Table 2-1**). Among the conditioning models, scaling, shifting, and linearly modulating gene expression by chemical features performed similarly. A t-distributed stochastic neighbor embedding (t-SNE) decomposition of learned FiLM parameters demonstrated that scaling and shifting operations encoded distinct chemical features (**Figure 2-1**). Hierarchical clustering of scaling (γ) parameters grouped compounds by identity (**Supplementary Figure 2-1**), whereas compound concentration correlated with the first principal component of shifting (β) parameters (p=1.72e-55, **Supplementary Figure 2-1**). Thus the learned conditioning parameters interpretably reflected compound structure and concentration in the drug-response modeling task as an emergent property of the training.

Table 2-1. Predictive performance of feature modulation models.
Concatenation of compound features and gene expression features provided a baseline performance. Learned conditional transformations performed similarly. Compound structural ablation largely reduced predictive capacity.

Model	R2
concatenation	0.6066 ± 0.0165
shift	0.7060 ± 0.0304
scale	0.7113 ± 0.0081
FiLM	0.7089 ± 0.0040
Structural ablation	0.3016 ± 0.0304

Cellular response commonly follows a sigmoidal relationship to drug concentration. Thus to quantify whether compound dosage alone was driving drug sensitivity predictions, we performed a feature ablation control experiment, wherein we purposefully removed crucial data from the model's training and compared it to the actual model. For the "straw model," we replaced chemical fingerprints with unique but structurally uninformative and randomized numerical values. The "straw model" trained this way failed catastrophically, consistent with the importance of compound structural features in the modeling task (**Table 2-1**)[53]. Explicitly modeling chemical information using conditioning is a valuable inductive bias for chemical sensitivity prediction and gives insights into the predictive mechanisms of the model. We averaged predictions from an ensemble of 5 ChemProbe FiLM models trained on independent data folds in subsequent experiments.

II. ChemProbe predicts breast cancer patient response

We next asked whether learned transcriptional patterns would generalize to an *in vivo* cellular context. We measured how well ChemProbe, trained solely on the CTRP, could predict drug response in clinical samples. We used gene expression and patient drug response data

from the I-SPY2 adaptive, randomized, phase II clinical trial of neoadjuvant therapies for early-stage breast cancer. I-SPY2 assigned patients to treatment arms based on defined biomarkers including hormone receptor status, human epidermal growth factor receptor-2 expression, and MammaPrint status. Absence of invasive cancer in the breast and regional lymph nodes at the time of surgery defined the endpoint of pathological complete response (pCR) (nonresponse, pCR=0; response, pCR=1)[54–58].

The I-SPY2 dataset also introduced a major change in the input data's modality. CCLE quantifies gene expression by high-throughput RNA sequencing, but I-SPY2 collected pre-treatment patient gene expression by microarray. Microarrays measure gene expression with lower overall specificity and sensitivity and capture a smaller dynamic range of gene abundance[59]. In this case, the microarray library mapped to only 90% of the protein-coding genes in ChemProbe's training. To test how far out of the model's training distribution the I-SPY2 data might be, we mean-imputed the missing features and compared I-SPY2 expression profiles to the first two principal components of CCLE's expression profiles. A portion of I-SPY2 data fell outside the training data distribution, consistent with the different technology platforms used to measure expression profiles. This mismatch underscores the challenge of generalizing across different means to collect gene expression data (Methods, **Supplementary Fig 2-2**).

We assessed whether ChemProbe could retrospectively predict responders versus non-responders within I-SPY2 trial arms. Five drugs from I-SPY2 matched those in the training dataset. We found a significantly lower predicted scaled area under the curve (AUC) in the

responder group for 4 of the 5 drugs (**Figure 2-2**). Additionally, we computed receiver operating characteristic (ROC) curves for each drug response prediction relative to trial arm outcomes. The area under the ROC (auROC) curve for each drug spanned 0.60 (paclitaxel and neratinib) to 0.73 (veliparib), with a macro-average auROC of 0.65 across all considered drugs (**Figure 2-2**).

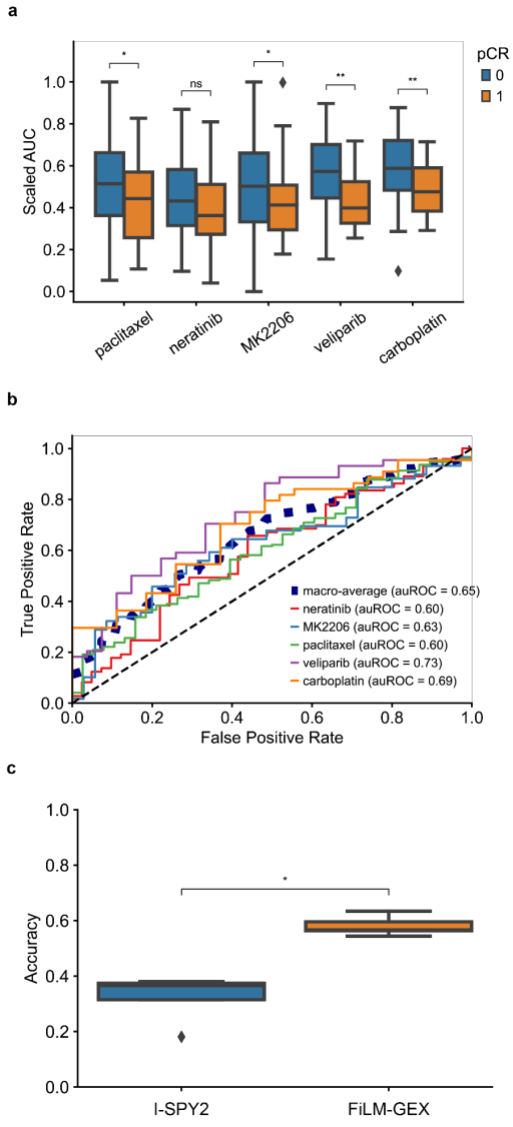


Figure 2-2. I-SPY2 clinical trial retrospective analysis.

(a) Predicted AUC of dose-response curve for I-SPY2 patients treated with each drug. AUC scaled between by the minimum and maximum predicted AUC of patients treated with

each drug. Blue = non-responder, orange = responder; Wilcoxon rank-sum; ns: *p*<=1e1, **p*<=5e-2, ***p*<=1e-2.

(b) Receiver operating characteristic curve of patients treated with each drug and corresponding auROC.

(c) Accuracy of I-SPY2 predictions and ChemProbe predictions for non-responders/responders. Wilcoxon rank-sum; **p*<=5e-2.

Finally, we evaluated ChemProbe predictions for clinical decision-making. We used ChemProbe to classify patients as responders (+) and non-responders (-) (ChemProbe+/-). As the model predicts cellular viability as a function of drug concentration, we defined a decision threshold for responders. We binarized drug response at each I-SPY2 treatment arm's median predicted AUC across participants. ChemProbe+/- classification accuracy significantly exceeded I-SPY2's (p<5e-2, **Figure 2-2**). While I-SPY2 predictions achieved a higher true positive rate for the considered drugs (0.30, I-SPY2; 0.21, ChemProbe), ChemProbe+/- classifications greatly reduced the false positive rate (0.70, I-SPY2; 0.37, ChemProbe) while adding relatively few false negative predictions (0.00, I-SPY2; 0.095, ChemProbe) (**Supplementary Figure 2-2**). ChemProbe+/- greatly increased the true negative rate of drug-response classification relative to I-SPY2, correctly predicting a portion of patients with a low likelihood of response to the proposed drug—essential information for clinical decision-making. Despite training solely on isogenic cell lines, the ChemProbe model generalized to heterogeneous tumors from clinical patient samples.

III. ChemProbe predicts differential cellular sensitivity

We prospectively evaluated ChemProbe's ability to differentiate drug sensitivity *in vitro* between a pair of primary breast cancer cell lines (HCC1806 and MDAMB231) against 545 CTRP compounds. To check for batch effects, we assessed where the cell lines fell within the training data by the top two gene-expression principal components. Although CTRP included both cell lines, the gene expression profiles we collected experimentally varied substantially from their CTRP counterparts (**Supplementary Figure 2-3**). The variability of cell culture protocols, reagents, and genetic drift may account for transcriptional differences across isogenic cell lines.

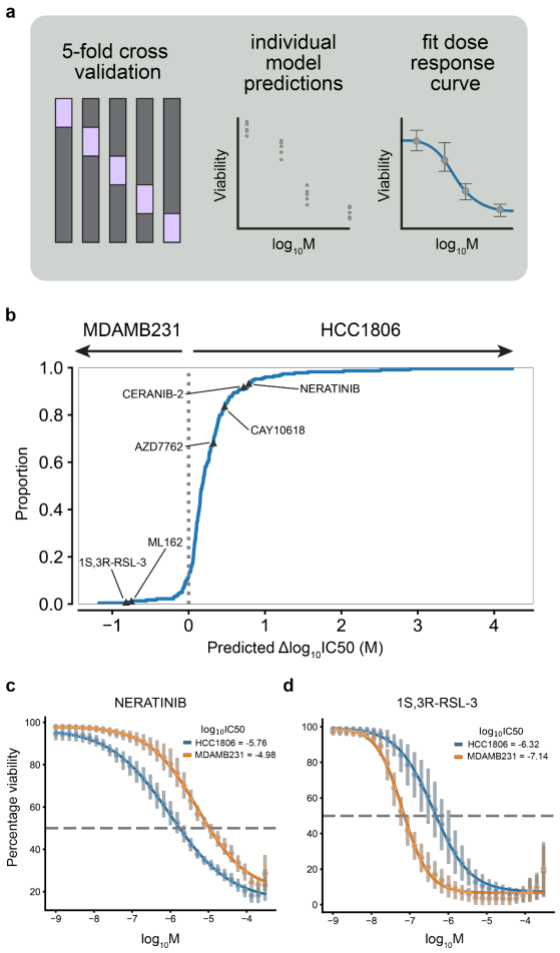


Figure 2-3. Differential potency predictions.

(a) Detailed approach to model training and dose-response modeling. Individual models were trained on held-out cell line dataset splits by 5-fold cross validation. Log-logistic models were fit to cross validated model predictions and pharmacodynamic features were derived.

(b) Expected cumulative distribution plot of predicted compound IC50 differences between HCC1806 and MDAMB231 cell lines. Compounds selected for *in vitro* dose-response testing highlighted.

(c) Predicted dose-response relationships of HCC1806 and MDAMB231 response to neratinib.

(d) Predicted dose-response relationships of HCC1806 and MDAMB231 response to 1S,3R-RSL-3.

We predicted sensitivity across 32 drug concentrations (1e-3–300 uM), fit log-logistic models, and determined 50% inhibitory concentration (IC50) values of each *in silico* dose-

response curve. We focused on compounds with the largest differences in IC50 between the cell lines (**Figure 2-3**). ChemProbe predicted that HCC1806 would be more sensitive than MDAMB231 for 88.16% (201/228) of the compounds with calculated dose-response curves (**Figure 2-3**). We selected four compounds whose predicted IC50s were strongest against HCC1806 (neratinib, ceranib-2, CAY10618, and AZD7762) and two compounds with IC50s favoring MDAMB231 (ML162 and 1S,3R-RSL-3) (**Figure 2-3**). *In vitro* testing confirmed ChemProbe's predictions for all six compounds (**Figure 2-4**). Predicted differences in IC50s between the two cell lines significantly correlated with true differences (p=0.043, **Figure 2-4**). Likewise, predicted IC50s correlated highly with true IC50s for individual cell lines after correcting for an outlier (HCC1806 treated with neratinib) (p=0.055, **Figure 2-4**, **Supplementary Figure 2-4**). These results were consistent with the initial concentration range-finding experiments, where 5/6 compounds had shown predicted differences in IC50s and relevant IC50s specific to individual lines (**Supplementary Figure 2-4**). In prospective testing, ChemProbe predicted the drug sensitivities of independently collected and characterized cell line samples, despite their meaningfully different transcriptomic profiles from the cell, tissue, and cancer subtypes in the model's training dataset.

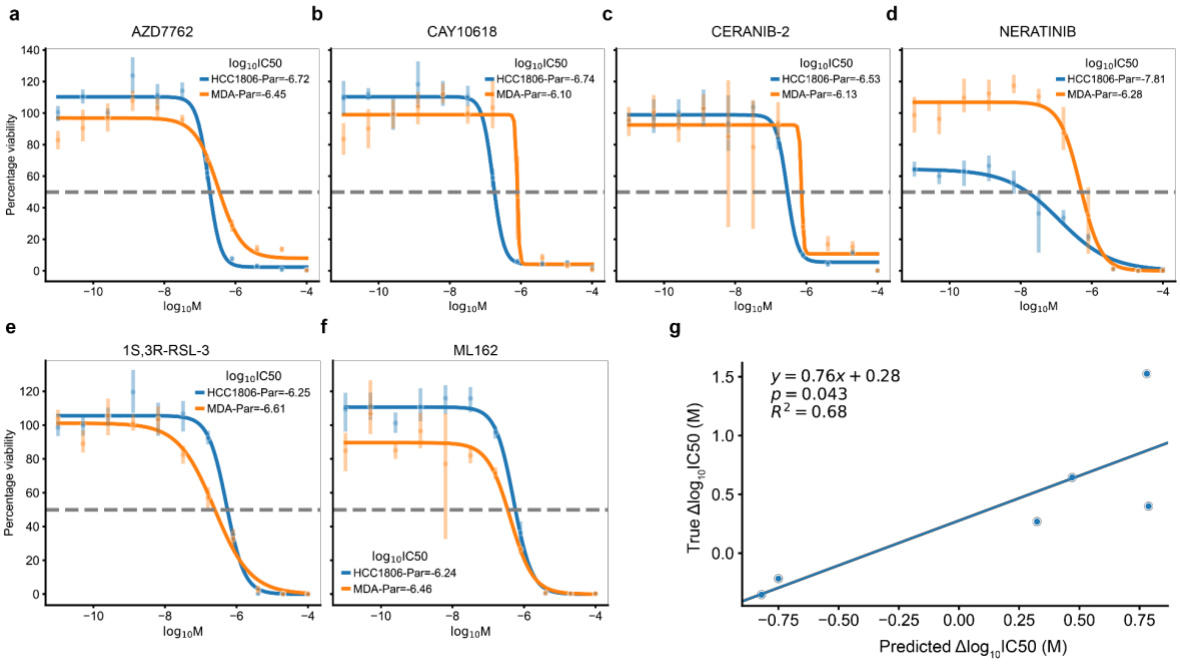


Figure 2-4. Validation of differential sensitivity predictions.

(a) *In vitro* dose-response relationships of HCC1806 differentially potent compound AZD7762.

(b) *In vitro* dose-response relationships of HCC1806 differentially potent compound CAY10618.

(c) *In vitro* dose-response relationships of HCC1806 differentially potent compound ceranib-2.

(d) *In vitro* dose-response relationships of HCC1806 differentially potent compound neratinib.

(e) *In vitro* dose-response relationships of MDAMB231 differentially potent compound 1S,3R-RSL-3.

(f) *In vitro* dose-response relationships of MDAMB231 differentially potent compound ML162.

(g) Relationship between predicted difference in IC50 and true difference in IC50 between HCC1806 and MDAMB213 across tested compounds.

IV. Gene expression attribution vectors pass model interpretation sanity checks

We next asked whether ChemProbe had learned biologically relevant patterns by deep learning feature attribution. Controlling for cell line-specific effects, we tested whether the model's learned gene expression saliency reflected known compound pharmacology and network biology relationships. We prospectively screened seven new cell lines, including three from primary tissue (SW480, colon; MDAMB231, breast; and HCC1806, breast) and four metastatic derivatives (SW480-LvM2, liver; MDAMB231-LM2, lung; HCC1806-LM2b and HCC1806-LM2c, lung). We used integrated gradients on predicted IC50s to calculate transcriptome attribution vectors for each cell line-compound pair. Given the low inter-model prediction variance of ChemProbe, we tested if attribution vectors were dually consistent across models. Transcriptomic attribution vectors correlated across all independently trained models, indicating their consistency. Integrated gradient attribution vectors can correlate closely with input feature magnitudes, confounding their use in quantifying feature importance. Accordingly, the first two principal components of attribution vectors were distinguished by cell line, and attribution vectors strongly correlated with transcriptome vectors (**Supplementary Figure 2-5**). Gene expression magnitudes vary by cell line; therefore we normalized attribution vectors by line to keep them from dominating the feature-importance calculations. This normalization decoupled the correlation between attribution and transcriptome vectors and decreased cell-line-specific effects in the principal component analysis (**Supplementary Figure 2-5**).

We applied sanity checks to test whether the calculated feature attributions depended on (1) learned model parameters or (2) the relationship between data features and labels[60]. For the first test, we randomly initialized parameters of architecturally identical models, applied integrated gradients, and compared true-model and random-model attribution vectors. For the second test, we trained a model on permuted labels and compared true-model and permuted-model attribution vectors. We tested uncorrected (raw) attribution vectors as well as cell line-effect corrected (adjusted) attribution vectors. Raw attribution vectors correlated closely with transcriptome profiles, random-model, and permuted-model attribution vectors, failing the sanity check. By contrast, the adjusted attribution vectors did not correlate with transcriptome profiles, random-model attributions, or permuted-model attributions, indicating adjusted attribution vectors do not simply reflect artifacts of the model architecture or the raw data.

V. Model interpretation of transcriptomic features reflects compound pharmacology and network biology

We suspected that integrated gradients could help interpret drug sensitivity IC50 predictions. Neural networks are notoriously difficult to interpret, and highly attributed transcriptome features may reflect causative mechanisms or correlative biomarkers. We hypothesized that ChemProbe should rely on similar gene features to make predictions for compounds with the same known targets. We defined a control compound set (CCS) based on nominal target classes containing at least two compounds. To test whether transcriptomic attribution vector similarity corresponded to known compound mechanisms of action (MOA), we applied K-means clustering to the CCS attribution vectors and computed the adjusted

mutual information (AMI) between cluster labels and nominal target classes. As structural similarity alone begets target profile similarity, we also examined the AMI between structural clusters produced from compound fingerprints and nominal target classes. The AMI of attribution vector clusters was significantly greater than that of structural clusters, a randomly initialized model, and a model trained on permuted labels (**Figure 2-5**). We also evaluated whether our model made predictions based on nominal targets and found that compounds belonging to the same target class frequently had high nominal target attributions relative to other compounds (**Figure 2-5**, **Supplementary Figure 2-5**).

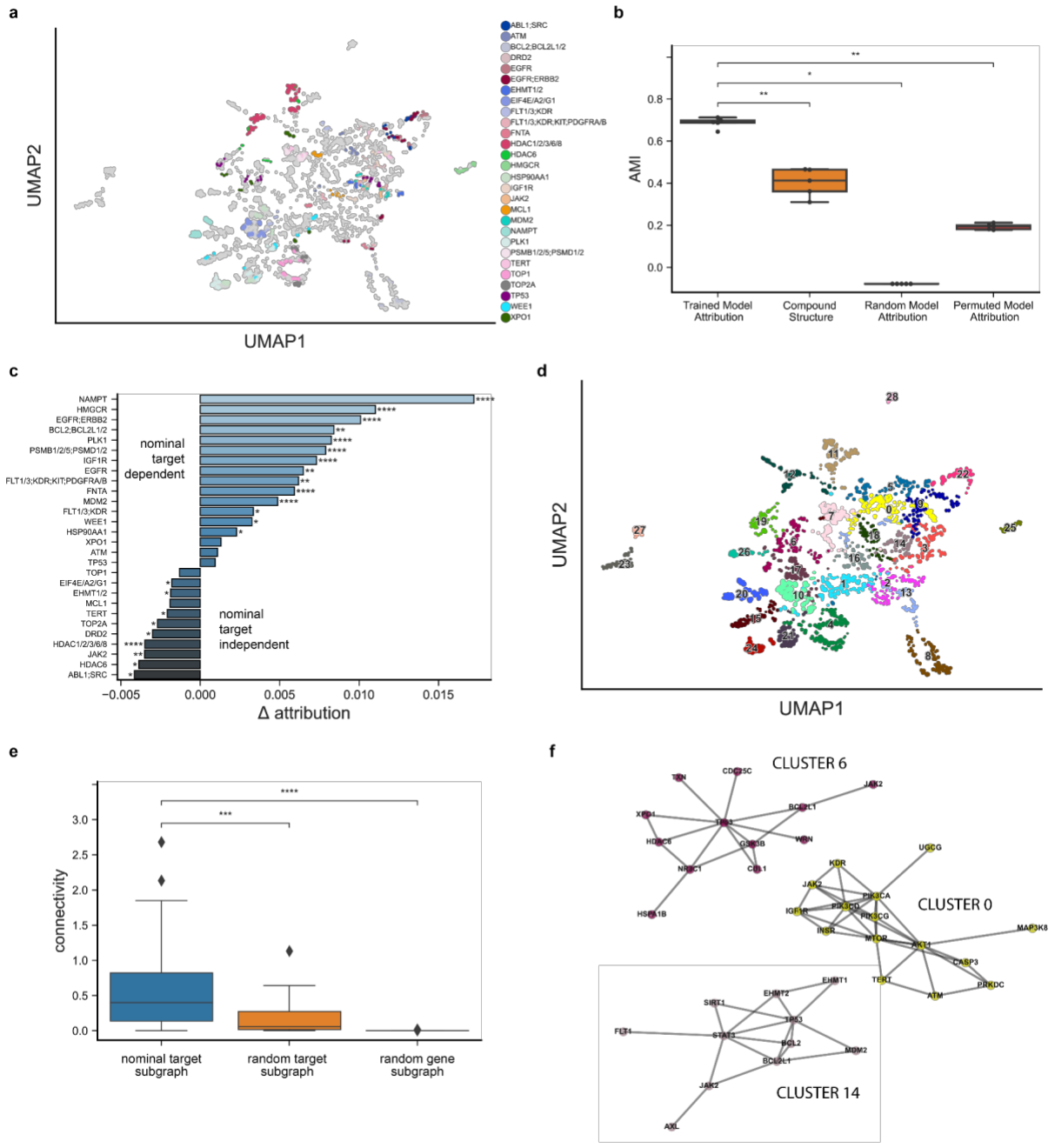


Figure 2-5. Feature attribution analysis of compound nominal targets.

(a) UMAP decomposition of adjusted attribution vectors derived at compound IC50s for predicted and fitted dose-response relationships in MDAMB231, MDAMB231-LM2, HCC1806, HCC1806-LM2b/c, SW480, and SW480-LvM2 cell lines. Control compound set (CCS) attribution vectors colored by nominal target class.

(b) Comparison of adjusted mutual information (AMI) derived from CCS nominal target labels and K-means clustering of trained model adjusted attribution vectors, compound fingerprints, random model adjusted attribution vectors and permuted model adjusted attribution vectors. Wilcoxon rank-sum; *p<=5e-2, **p<=1e-2.

(c) Average attribution difference between highest significance target of nominal target class and all other target classes. Wilcoxon rank-sum; *p<=5e-2, **p<=1e-2, ***p<=1e-3, ****p<=1e-4.

(d) Leiden clustering of all attribution vectors.

(e) Comparison of PPI subgraph connectivity derived from clustered target profiles, random target profiles and random protein coding genes. Wilcoxon rank-sum; ***p<=1e-3, ****p<=1e-4.

(f) Network representation of select clustered target profile subgraphs.

We reasoned that because feature encoding in neural networks is hierarchical, attribution vectors may reflect the structure of higher-order biological systems. We expanded our analysis to all attribution vectors to test if the nominal targets of similar attribution vectors enriched for protein modules (**Figure 2-5**). We looked into the network topology of nominal target classes using the STRING database of high-confidence protein-protein interactions[61]. We clustered attribution vectors, sampled nominal target annotations, and extracted the subgraph of protein targets (Methods). We found that the connectivity of nominal target subgraphs was significantly greater than the connectivity of subgraphs generated from random target protein sets or random protein sets of equal size (**Figure 2-5**). Finally, we tested if attribution-defined nominal target modules of action (ModOA) were not only highly connected but reflected protein interaction enrichment (Methods). 50% of ModOA reflected significant

network interaction enrichment as well as a variety of functional enrichments from gene ontologies, KEGG pathways, and Reactome pathways (**Figure 2-5**).

Given that STRING likely does not reflect all protein-protein associations, we reasoned that unenriched ModOA might still reflect true, albeit untested, biological relationships. We produced target modules from our clustering analyses in order to infer potential ModOA for compounds with high promiscuity or uknown targets. These model investigations support the hypothesis that ChemProbe learns patterns of compound pharmacology and systems biology.

VI. Differential attribution analysis reflects ferroptosis biology and nominates novel therapeutic targets

We then wanted to know if the most highly attributed gene features were related to compound MOA. We used linear regression to test for differences in gene attribution between groups, which we call differential attribution analysis (DAA) (Methods). DAA generates ranked gene lists, which we used as marker genes to arrange attribution clusters hierarchically (**Figure 2-6**).

We noticed clusters 26 and 28 reflected divergent phenotypes to ferroptosis-inducing compounds (**Figure 2-6**). Ferroptosis is a recently recognized type of cell death that has been implicated in a variety of biological contexts and has the potential to be a targeted pathway in cancer, immunity, development, and aging[62,63]. These attribution clusters included compounds ML162 and 1S,3R-RSL-3, reflecting the differential cellular sensitivity we confirmed in our *in vitro* experiments. Additional compounds with ferroptosis-inducing mechanisms of action in this cluster included ML210, erastin, CIL56, and CIL70.

Next, we investigated the attribution vectors of cell lines sensitive (cluster 26) and resistant (cluster 28) to ferroptosis-inducing compounds to understand if model interpretations reflected known ferroptosis biology. First, we collapsed the leaf nodes clusters 26 and 28 into a single ferroptosis-inducing compound cluster and applied DAA. Since a variety of mechanisms induce ferroptosis, we queried differential attributions of several ferroptosis-associated genes including GPX4, SCD, SLC7A11, FSP1, and LRP8. All ferroptosis-associated genes were in the most highly attributed genes of the ferroptosis-inducing compound cluster (**Figure 2-6**). To verify that these signals did not merely arise due to transcriptome inputs or relative expression differences, we performed differential expression analysis (DEA) between MDAMB231 and HCC1806 and queried the ferroptosis gene set. Besides GPX4, a key ferroptosis regulator, no ferroptosis-associated genes rose to significance (**Supplementary Figure 2-6**).

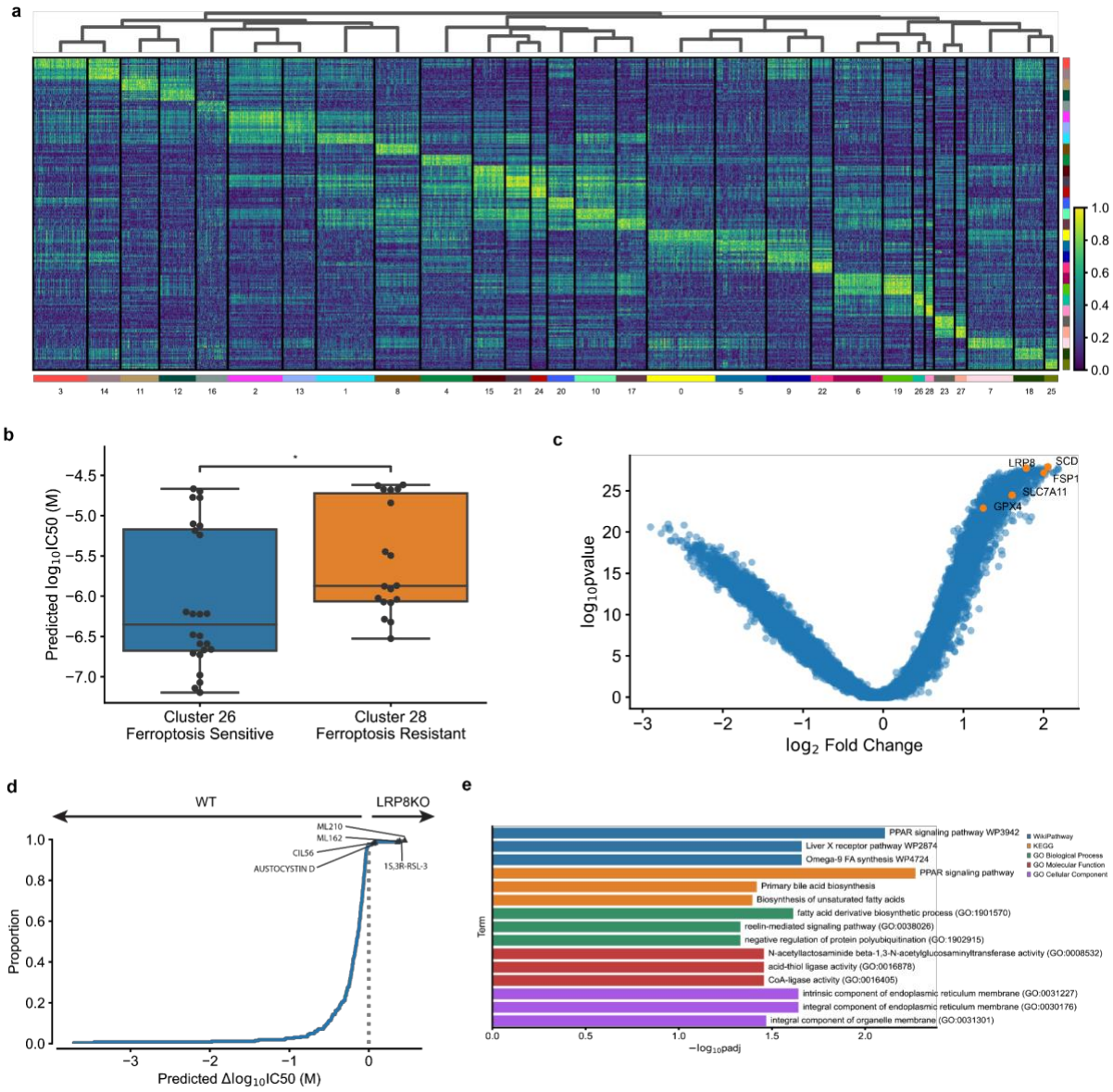


Figure 2-6. Differential attribution analysis of ferroptosis-inducing compounds.

(a) Heatmap of top-10 differentially attributed genes within Leiden clusters from Figure 2-5d. Clusters ordered by hierarchical clustering of DAA profiles.

(b) Comparison of predicted IC50s between cluster 26 (ferroptosis-sensitive) and cluster 28 (ferroptosis-resistant). Wilcoxon rank-sum; *p<=5e-2.

(c) Volcano plot of DAA results derived by comparing ferroptosis-inducing compound attributions to all other compound attributions. Known ferroptosis-mediating genes highlighted in orange.

(d) Expected cumulative distribution plot of predicted compound IC50 differences between HCC1143 WT and LRP8 KO cell lines. Ferroptosis-inducing compounds predicted differentially potent in LRP8 KO highlighted.

(e) Enrichment analysis of top-10 differentially attributed genes of ferroptosis-inducing compound samples.

Changes in compound sensitivity following gene knockout (KO) or overexpression inform gene-dependent protection or resistance. We assessed ChemProbe's utility for screening gene-dependent ferroptosis resistance. Lipoprotein receptor LRP8 has recently been shown to act as a ferroptosis resistance factor by maintaining cellular selenium levels and appropriate translation of GPX4. In LRP8 KO models, selenium uptake is reduced, leading to ribosome stalling and early translation termination of GPX4—ultimately sensitizing cells to ferroptosis[21]. We tested if ChemProbe correctly predicted reduced sensitivity to ferroptosis-inducing compounds in an LRP8KO cell line relative to wild-type. Concordant with previous research, ChemProbe predicted LRP8KO cells were sensitive to known ferroptosis-inducing compounds ML210, 1S,3R-RSL-3, ML162, and CIL56 (**Figure 2-6**)[64].

We noticed several highly attributed genes of ferroptosis-inducing compounds reflected expression-response correlations in the CTRP (**Supplementary Figure 2-6**). Under this consideration, we sought to understand if highly attributed genes not only reflected simple correlations, but played functional roles related to ferroptosis. We extracted the 10 highest differentially attributed genes and applied a functional enrichment analysis (**Supplementary Figure 2-6**). We observed enrichment of terms related to lipid transport and fatty acid metabolic processes, pathways adjacent to lipid peroxidation and ferroptosis (**Figure 2-6**).

With evidence that highly attributed genes not only reflect expression-response relationships but are functionally involved in related cellular mechanisms, we applied DAA to nominate genes potentially associated with ferroptosis sensitivity and resistance. (**Supplementary Figure 2-6**). Overall, our findings show that transcriptomic attributions in different phenotype clusters reflect ferroptosis biology, allowing us to screen genetic dependencies for new biological mechanisms.

2.4 Discussion

Here we investigated the gene dependencies of a model that learned the effect of compound features on basal cellular transcriptomic state to predict cellular sensitivity. This study takes an unbiased approach to modeling cellular sensitivity to various chemical probes and drugs. We demonstrate that an accurate model of cellular sensitivity learns higher-order aspects of compound pharmacology, reflects systems biology, and nominates genes functionally related to compound MOA.

ChemProbe takes a principled approach to modeling how compound features influence cellular transcriptomic state. We find that learning an affine transformation of transcriptome representations from compound representations is a valuable bias in the drug sensitivity prediction task (**Table 2-1**). Analysis of learned parameters reflects the influence of compound structure and concentration on gene expression representations. This observation led us to apply model interpretability methods to explore the relationship between learned gene dependencies and compound mechanisms.

Prior to model investigation, we verified the utility of our approach in two challenging scenarios. First, we retrospectively evaluated ChemProbe on patient drug-response results from I-SPY2, an adaptive clinical trial designed to evaluate neoadjuvant therapies for breast cancer treatment. In contrast to the isogenic cell line transcriptomes used to train ChemProbe, I-SPY2 clinical tumor samples are biologically heterogenous, quantified by microarray, and missing approximately 10% of gene expression features. ChemProbe correctly predicted a significant difference in patient response for 4/5 drugs and was significantly more accurate in classifying non-responders/responders compared to original trial arm designations (**Figure 2-2**). Second, we performed a prospective analysis that evaluated the ability of ChemProbe to predict differential drug sensitivity. We observed a significant correlation between predicted and true differential IC50s, as well as a high correlation between predicted and true IC50s (**Figure 2-4**). Together these results support the generalization ability of our model in both clinical and research settings.

As molecular information becomes increasingly important in the clinical setting, ChemProbe contributes a robust computational tool that may guide future clinical treatment selection in precision oncology. Clinical adoption of new computational tools, however, relies on method interpretability. Although ChemProbe is structured as a traditional “black-box” neural network, our analyses support the conclusion that ChemProbe has learned patterns of biology causally related to phenotype response. Our proposed DAA method is a tool to explore gene patterns driving individual disease and discover new disease-gene relationships. Research investments into model interpretability will advance precision medicine by providing insight into systems pharmacodynamics driving patient responses.

In the research setting, ChemProbe is an *in silico* chemical perturbation tool. New models of biology may be screened against this set of diverse chemical probes to evaluate sensitivity to specific MOA classes. In one example, we screened an LRP8KO cellular model and, within minutes, correctly predicted differential sensitivity to ferroptosis-inducing compounds (**Figure 2-6**). These results support the role of LRP8 as a ferroptosis resistance factor and potential therapeutic target in cancer. Computational access to the CTRP screening library via ChemProbe should reduce resources spent searching for specific disease vulnerabilities and the outcomes of perturbing specific biological mechanisms.

Extending this paradigm, we leveraged *in silico* perturbation results to validate interpretation of ChemProbe and propose new ModOA for common target proteins. Initially, we found that adjusted attribution vectors of transcriptome profiles reflected compound MOA similarity better than compound structures (**Figure 2-5**). We followed this up by analyzing the connectivity of nominal target PPI subgraphs. Clustered nominal target attributions reflected high PPI connectivity, indicating model feature dependence aligned with broader systems biology structure (**Figure 2-5**). We used this insight to propose target protein sets that were likely functionally related (but not necessarily annotated) for interaction. An avenue of future work will be to test if novel compound pairs associated by ModOA promote synergistic, antagonistic, or additive effects through perturbation of the shared protein module. On the other hand, diseases of diverse mechanisms, such as cancer, may be more vulnerable to network redundancy disruption by targeting orthogonal ModOA.

An outstanding question remained with respect to the value of interpreting individual gene attributions. We applied a nonparametric test to compare the distributions of individual gene attributions between groups, termed Differential Attribution Analysis. We used DAA to identify which genes ChemProbe relies on to make drug sensitivity predictions conditional on treatment with ferroptosis-inducing compounds. Ferroptosis is driven by iron-dependent peroxidation of membrane lipids[63]. Four independent mechanisms that induce this form of cell death have been recognized; however, many outstanding research directions remain, including identification of key gene participants in ferroptosis[62,65]. We showed that DAA of ferroptosis-inducing compounds enriches several genes critically implicated in ferroptosis (**Figure 2-6**). Furthermore, ChemProbe predicted increased sensitivity of an LRP8KO model to ferroptosis-inducing compounds, providing additional evidence that implicates LRP8 as a ferroptosis resistance factor (**Figure 2-6**).

When we performed enrichment analyses on the top gene set from DAA, we found biological processes related to lipid synthesis, endoplasmic reticulum (ER) membrane components, and functional enrichment of terms related to acyl-coenzyme A (CoA) ligation (**Figure 2-6**). The highest enriched pathways were related to peroxisome proliferator-activated receptor (PPAR) signaling. SCD and SLC27A5 contributed to enrichment of PPAR signaling: SCD is known to drive resistance to ferroptosis; however, SLC27A5 has not been mechanistically implicated in ferroptosis[66]. SLC27A5 is a homolog of very long-chain acyl-CoA synthetase (VLCS) associated with the ER and plays a primary role in lipid synthesis. Like SCD, SLC27A5 expression is highly correlated with ferroptosis resistance in the CTRP (**Supplementary Figure 2-6**). Additional research is necessary to test the role of SLC27A5 and other DAA nominated

genes in ferroptosis. Ultimately, measuring the responses of many diverse biological contexts to mechanistic probes will support learning and interpreting gene dependencies across tissues, cell types, and models of disease.

2.5 Materials and methods

I. Pharmacogenomic dataset

Drug sensitivity data was obtained from the Cancer Therapeutic Response Portal v1 and v2 (CTRP v1/2). These datasets comprise 864 cell line responses to 481 individual compounds and 64 compound pairs across a range of concentrations. Response phenotypes were quantified by cellular viability, a normalized measure characterizing complete cell killing to cell stasis (0-1) and cell growth (>1). We utilized predicted cellular viability derived from fitted dose-response curves of each experimental set, in which replicate cell line-compound experiments were fit with a log-logistic function and predicted cellular viability was derived at the original experimental concentrations. Compound structure was represented as 512-bit Morgan fingerprints (radius=2) converted by RDKit from SMILES provided by the CTRP. Experimental micromolar compound concentrations were concatenated with Morgan fingerprints, resulting in 513-length compound feature vectors. We matched CTRP cell lines with the Cancer Cell Line Encyclopedia (CCLE) molecular characterizations and extracted protein-coding gene expression measurements, resulting in 19144-length cell line feature vectors. In total, 545 total compounds or compound pairs and 860 cell lines compromised 366,710 unique pairs and 5,849,340 individual examples of cellular response at various concentrations.

II. ChemProbe architecture, training, and evaluation

In the context of pharmacological intervention, phenotypic response results from a change in basal cellular state dependent on a chemical presence. As such, we hypothesized that better integration of cell state features with compound features would result in more accurate predictions of drug sensitivity. Simple concatenation of features is a common means of conditional biasing. Alternatively, hierarchical representations of gene expression features can be combined with learned small molecule feature representations in multiple downstream layers such that a network learns various conditional transformation parameters. Under Perez et al.'s general formulation of this approach, termed feature-wise linear modulation (FiLM), a neural network learns an affine transformation of input features by conditional information. Model parameters are learned by functions, g and h, dependent on compound representations, n: $\gamma = g(n)$, $\beta = h(n)$ (**Figure 2-1**). Learned parameters modulate intermediate gene expression representations by element-wise transformation: $f(x|n) = F * \gamma + \beta$, where F denotes the activations at a given layer.

The ChemProbe model defines a conditional encoder to embed compound features into a vector of length c and an inputs encoder that embeds gene expression features into a vector of length g. A FiLM generator ingests compound embeddings and predicts gamma and beta parameters of length g. The FiLM layer applies an affine transformation of gene expression embeddings by parameters γ, β. The modulated gene expression embeddings are then passed through a linear block consisting of a linear layer, ReLU activation, batch normalization and dropout. This sequence is repeated across 2 FiLM layers. The final linear block compresses

feature maps to a vector of length 1 and the mean-squared error is calculated between predicted cellular viability and true cellular viability (**Figure 2-1**). Neural network parameters are optimized by standard back propagation and gradient descent.

Models were trained and performance was evaluated using cross validation. Cell line-compound pairs were split into 5 groups of approximately equal size by cell line to ensure no cell line was present in the same training/validation split. As our goal was to assess generalization to unseen gene expression programs, this cell line-based dataset split avoided data leakage and performance inflation. Five individual models were trained in a leave-one-out cross validation scenario in which four data folds were combined into a training dataset and the model was evaluated on the left-out fold. We applied 20 rounds of hyperparameter optimization to all five individually trained models. Average R^2 of the best performing models was 0.7173±0.0052. ChemProbe is implemented in PyTorch and hyperparameter optimization was applied with Optuna on a GPU.

III. Predictive modeling baselines

In our experiments, we compared models that conditionally modulate gene expression features by compound structure and concentration through several different transformations. For our "concatenation" architecture baseline, we simply concatenated gene expression features and compound features and fed the single vector into a multi-layer perceptron (MLP). The "scale" and "shift" variants of the ChemProbe model evaluated the isolated effects of learning transformation types independently. Shift parameters were held constant ($\beta = 0$) for the "scale" model and scale parameters were held constant ($\gamma = 1$) for the "shift" model:

$f_{scale}(x|n) = F * \gamma + 0$; $f_{shift}(x|n) = F * 1 + \beta$. We assessed ChemProbe conditional dependence on compound concentration with our "permuted" model. We defined random binary fingerprints for each compound and trained a new ChemProbe model on this dataset lacking structural information. All models were trained and evaluated over 3 rounds of hyperparameter optimization via 5-fold cross validation on the originally defined dataset splits.

IV. Dose-response modeling

We generated predicted dose-response curves by fitting log-logistic functions to each set of cell line-compound predictions from the five individually trained ChemProbe models. We defined a sequence of conditions for quality control of each dose-response relationship. First, cellular viability at any of the four largest compound concentrations was checked for increases of 20% or more from the fifth largest compound concentration. If so, the viability prediction at the largest concentration was dropped. This condition was replied recursively and had to be met by a minimum of 16 data points else no dose-response curve was fit. If the minimum predicted cellular viability was greater than 0.4, no dose-response curve was fit. For cell line-compound pairs that passed quality control, a 4-parameter log-logistic function was fit: $y = c + (d - c) / (1 + e^{(b * \log(x) - \log(e))})$. If this optimization failed, a 3-parameter log-logistic function was fit with d = 1. If this optimization failed, a 2-parameter log-logistic function was fit with c = 0 and d = 1. In our analyses of predicted dose-response curves, we performed additional quality control by filtering out log-logistic functions with undetermined parameters and with predicted EC50 < 1e-3 or EC50 > 300. Scipy was used to fit parameters of log-logistic functions to dose-response relationships.

V. Retrospective I-SPY2 analysis

We downloaded I-SPY2 clinical trial metadata and microarray characterizations of 988 patient transcriptomes from GEO (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE194040). We matched 90% of recorded genes to our training dataset, mean-imputed the remaining 10% of genes, standardized the data with a Z-score transformation, and evaluated I-SPY2 patient data alignment with CCLE cell line training data across the first two principal components. We predicted drug sensitivity across a range of 32 concentrations (1e-3–300 uM) for each patient in response to all 545 compounds and compound pairs in the CTRP. We averaged patient-drug response predictions across independent models and computed the AUC of each predicted dose-response assay. The AUC of each patient-drug prediction was scaled between 0-1 based on the minimum and maximum predicted AUC of the drug across all I-SPY2 patients.

I-SPY2 participants were placed in treatment arms based on analyses of clinical and molecular information including clinical characteristics, gene expression patterns measured via microarray, and protein abundance as measured by RPPA. The I-SPY2 trial assessed the success of various combination therapies relative to paclitaxel treatment, a clinical standard of care. Drugs matched between I-SPY2 treatment arms and the CTRP included: paclitaxel, neratinib, MK2206, veliparib and carboplatin. In the I-SPY2 experimental arms, patients were treated with the combination of paclitaxel and an additional drug(s) to assess response relative paclitaxel treatment only. Given that the predictive ability of ChemProbe was only evaluated with respect to the available compounds and compound pairs in the CTRP, ChemProbe predictions for I-SPY2

patients reflected predicted patient response to a single compound, rather than a combination therapy.

VI. Prospective differential sensitivity predictions

Differential sensitivity predictions were made by taking the difference between HCC1806 and MDAMB231 predicted IC50s between compounds that passed dose-response modeling. We visually inspected fitted dose-response curves of the predicted largest 50 differentially sensitive compounds to select compounds for *in vitro* tests. Selection criteria was based upon dose-response curve completeness in each cell line, including sufficient E_{max} and E_{min} boundaries within the predicted concentration range.

To determine appropriate concentration points for the prospective dose-response experiment, we performed a range-finding dose-response experiment across a wider range of concentrations than our predictions (**Supplementary Figure 2-4**). In the following experiment, we reduced the concentration range to capture response granularity (**Figure 2-4**).

VII. Integrated gradients

Integrated gradients is a path-based model attribution method that quantifies how much feature gradients change relative to a baseline feature vector. First, n feature vectors are linearly interpolated between a specified baseline and the query feature vector. We used a zero-vector baseline for both compounds features and gene expression features and a step size of $n = 50$. At each interpolated feature vector step, gradients of the interpolated feature inputs are computed with respect to the corresponding prediction. Finally, the integral of each

feature is computed along the path of feature gradients between the baseline vector and the query vector.

We controlled attribution vector differences resulting from different cellular responses by computing integrated gradients relative to the predicted compound IC50 for each cell line-compound pair. This resulted in an attribution vector for each cell line-compound pair at the predicted IC50. We extracted the cell line feature attribution vector of each pair to analyze how conditional compound information influenced gradient changes of the input gene expression features. To account for cell line-specific effects we standardized transcriptome attribution vectors of each cell line independently with a Z-score transformation, resulting in adjusted attribution vectors (**Supplementary Figure 2-5**).

VIII. Attribution method sanity check

The sanity check reveals attribution method insensitivity to (1) the learned parameters and (2) the learned dependence between data features and labels if attribution vectors between trained models and alternative models are highly correlated. To test for (1) model-dependent attribution method invariance, we randomly initialized parameters of architecturally identical models, applied integrated gradients, and compared true-model and random-model attribution vectors. To test for (2) data-dependent attribution method invariance, we permuted data labels, trained architecturally identical models, applied integrated gradients, and compared true-model and permuted-model attribution vectors (**Supplementary Figure 2-5**).

IX. Attribution similarity analysis

We analyzed the relationship between learned gene expression feature dependence conditional on compounds by comparing attribution vector similarity to compound MOAs. Compound MOA classes were extracted from CTRP annotations. We filtered attribution vectors by considering compound MOA classes with at least two compounds successfully attributed in all 7 cell lines to garner MOA classes with enough samples for our analysis (control compound set). We used nominal target classes as a true label baseline to assess concordance with unsupervised labels generated by K-means clustering of attribution vectors (**Figure 2-5**).

Compound MOA classes consisted of several nominal targets. To assess model dependence on individual nominal targets, we analyzed gene target attributions within each MOA class relative to all other classes. We applied a Wilcoxon-rank sum test between groups for each nominal target in the MOA class of interest and adjusted for false discovery rate (FDR). We visualized the nominal target with the largest average attribution difference between groups for each MOA class (**Figure 2-5**).

X. Attribution network analysis

We expanded our analysis to include all attribution vectors generated from the 7-cell line test set. We randomly selected a single nominal target from each compound MOA class to avoid bias towards closely associated targets. We reasoned that the nominal targets of a single compound likely fall in close network proximity and downstream network analysis of target sets would reflect artificial over-connectivity. For example, the MOA class of neratinib includes nominal targets EGFR and HER2, which are involved in the same pathway. Given that these

targets are closely associated by virtue of a shared MOA, we randomly choose one target from this set.

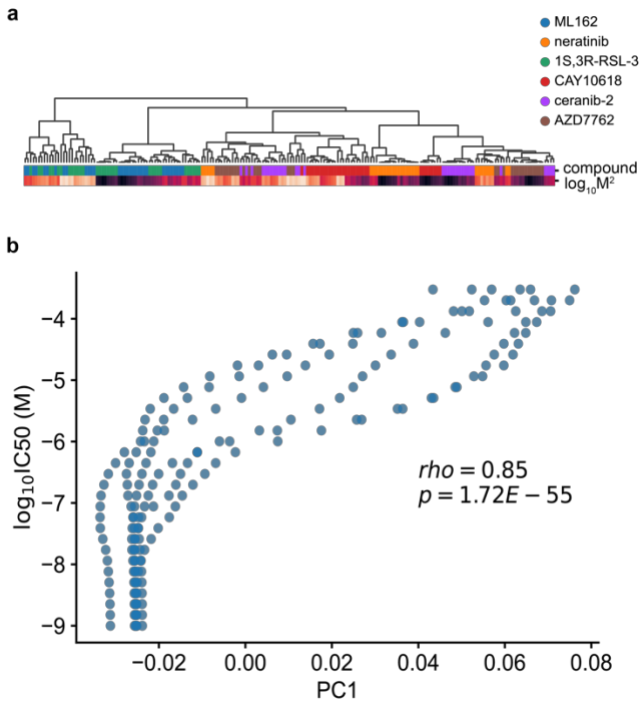
First, we applied Leiden clustering for unsupervised discovery of attribution clusters. We defined attribution cluster MOA classes by random target selection from each compound MOA class as described above and filtered the STRING database to all high confidence protein-protein associations (combined score > 0.7). We queried STRING for attribution cluster nominal targets and computed the connectivity of the resulting subgraph. To account for random subgraph connectivity due to target biases in STRING, we randomly sampled from available targets, queried the filtered STRING database, and computed connectivity. We repeated this procedure with randomly sampled protein coding genes to account for random protein associations (**Figure 2-5**). Networkx was used for network analysis.

To test for protein interaction enrichment, we again defined attribution cluster nominal targets by random target selection from each compound MOA class as described above. Next, we queried the STRING API for protein-protein interaction enrichment in the network of high-confidence protein-protein associations. Statistical enrichment was computed by the hypergeometric test. This tests if a query set of proteins has more interactions than would be expected relative to the background proteome-wide interaction distribution. The hypergeometric test was also applied to test for functional enrichment of GO terms, KEGG pathways, and Reactome pathways. We utilized the stringdb python package to access the STRING API. To infer potential modules of action (ModOA) for compounds, we returned the unique set of all nominal targets present in attribution cluster MOA classes.

XI. Differential attribution analysis

We extended our nominal target attribution analysis to perform an unbiased assessment of model dependence on individual genes within attribution clusters. We applied a Wilcoxon-rank sum test to each gene to analyze gene attributions within cluster samples relative all other samples and adjusted FDR. Scanpy was used to apply Wilcoxon-rank sum tests across genes in each cluster relative to the background.

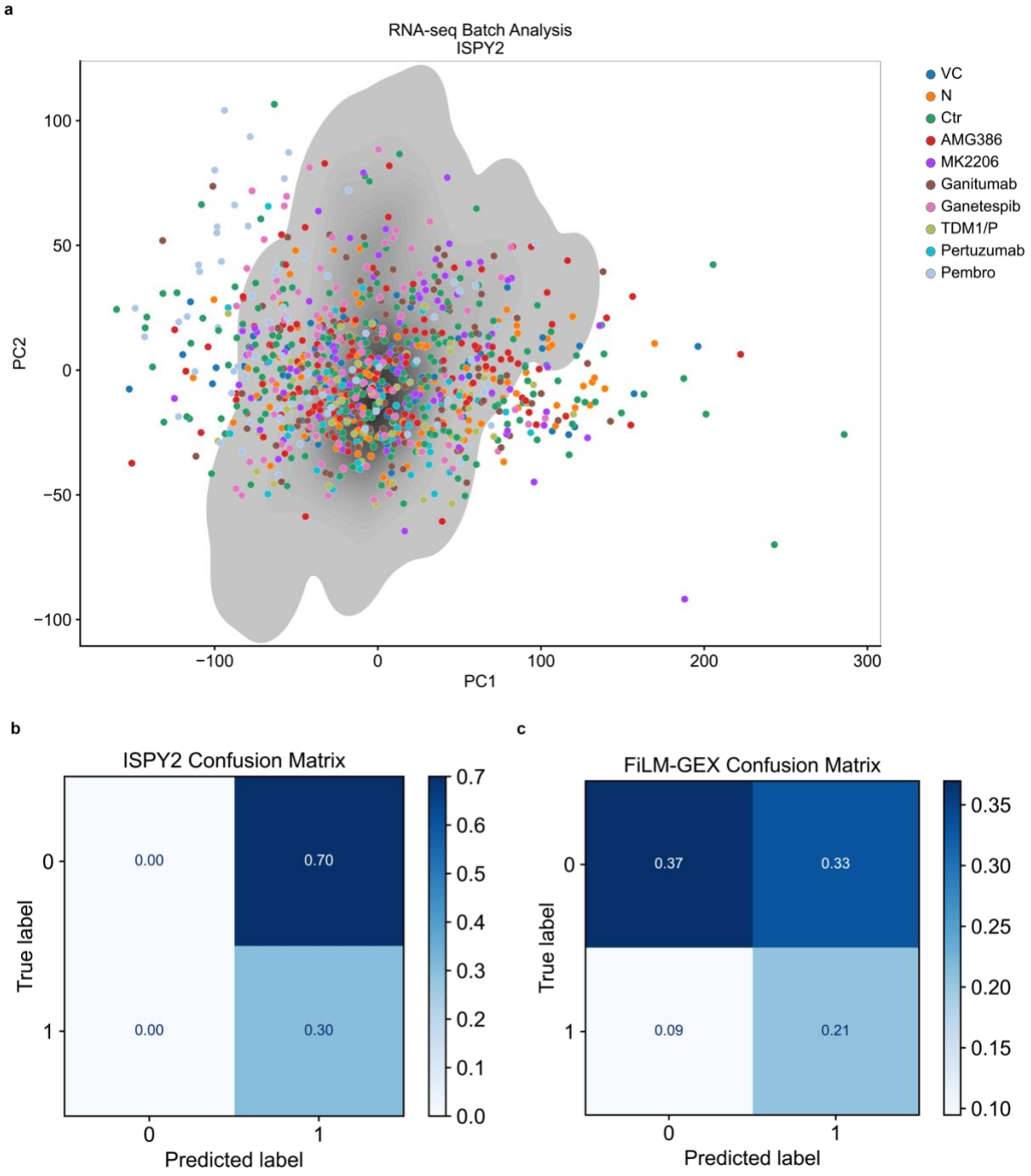
2.6 Supplementary Figures



Supplementary Figure 2-1. Conditioning parameter analysis.

(a) Hierarchical clustering of learned gamma parameters. Color bars indicated compound identity and squared compound concentration.

(b) Relationship between principal component 1 (PC1) of learned beta parameters and input concentration.



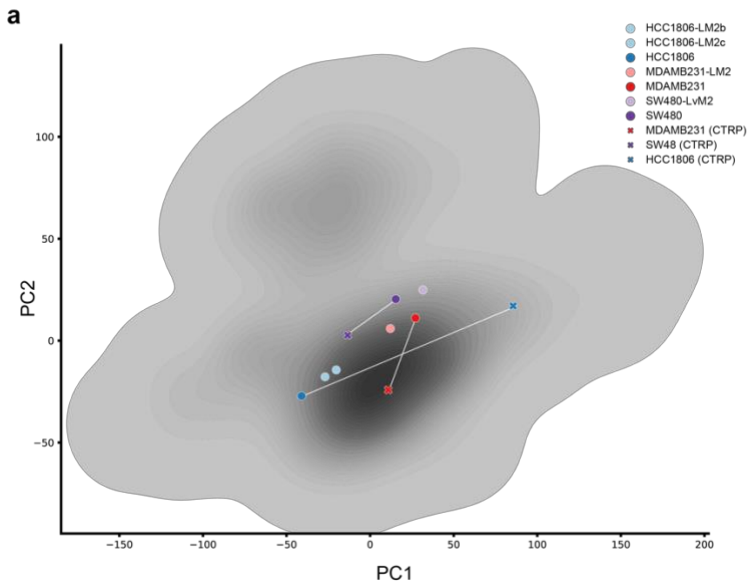
Supplementary Figure 2-2. I-SPY2 analysis.

(a) Batch analysis of I-SPY2 microarray data and HT-RNAseq training data. PCA decomposition of gene expression values of training dataset cell lines and I-SPY2 patient samples. Training distribution in grey, patients are points and colors are clinical trial

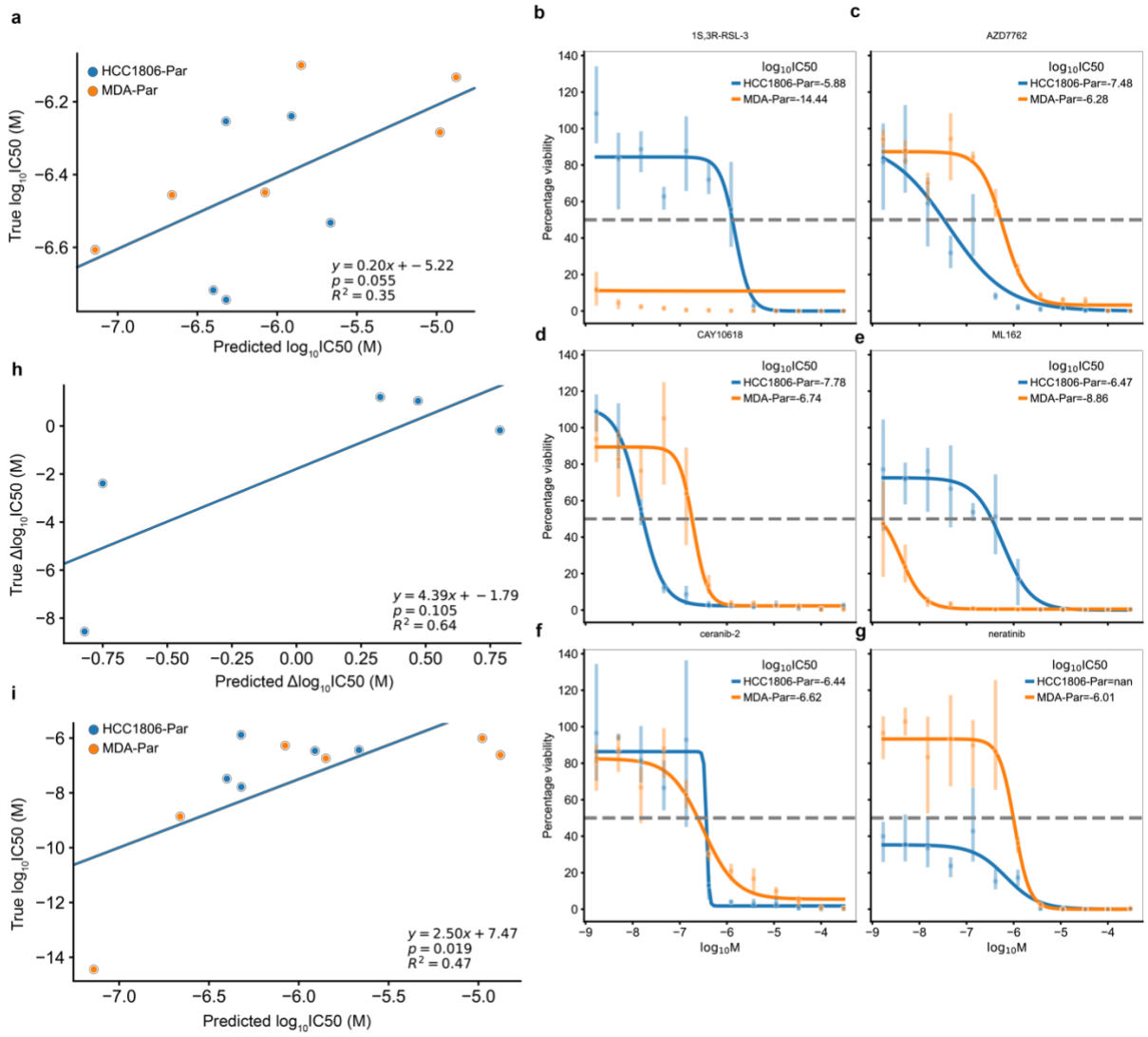
arms. VC = vincristine/carboplatin, N = neratinib, Ctr = paclitaxel, TDM1/P = TDM-1/pertuzimab, Pembro = pembrolizumab.

(b) Confusion matrix of predicted non-responders/responders (0/1) and true non-responders/responders (0/1) from I-SPY2 for the following trial arms: Ctr, N, VC, MK2206.

(c) ChemProbe confusion matrix.



Supplementary Figure 2-3. Cell line transcriptome similarity analysis.
PCA decomposition of gene expression values of training dataset cell line distribution (grey) and test dataset cell lines. Shape of cell line points represents samples from training dataset (CTRP) and test dataset.



Supplementary Figure 2-4. Concentration range-finding for dose-response experiment.

(a) Relationship between predicted IC50 and true IC50 for test compounds in HCC1806 and MDAMB231 cell lines.

(b) *In vitro* dose-response relationships of HCC1806 and MDAMB231 to 1S,3R-RSL-3.

(c) *In vitro* dose-response relationships of HCC1806 and MDAMB231 to AZD7762.

(d) *In vitro* dose-response relationships of HCC1806 and MDAMB231 to CAY10618.

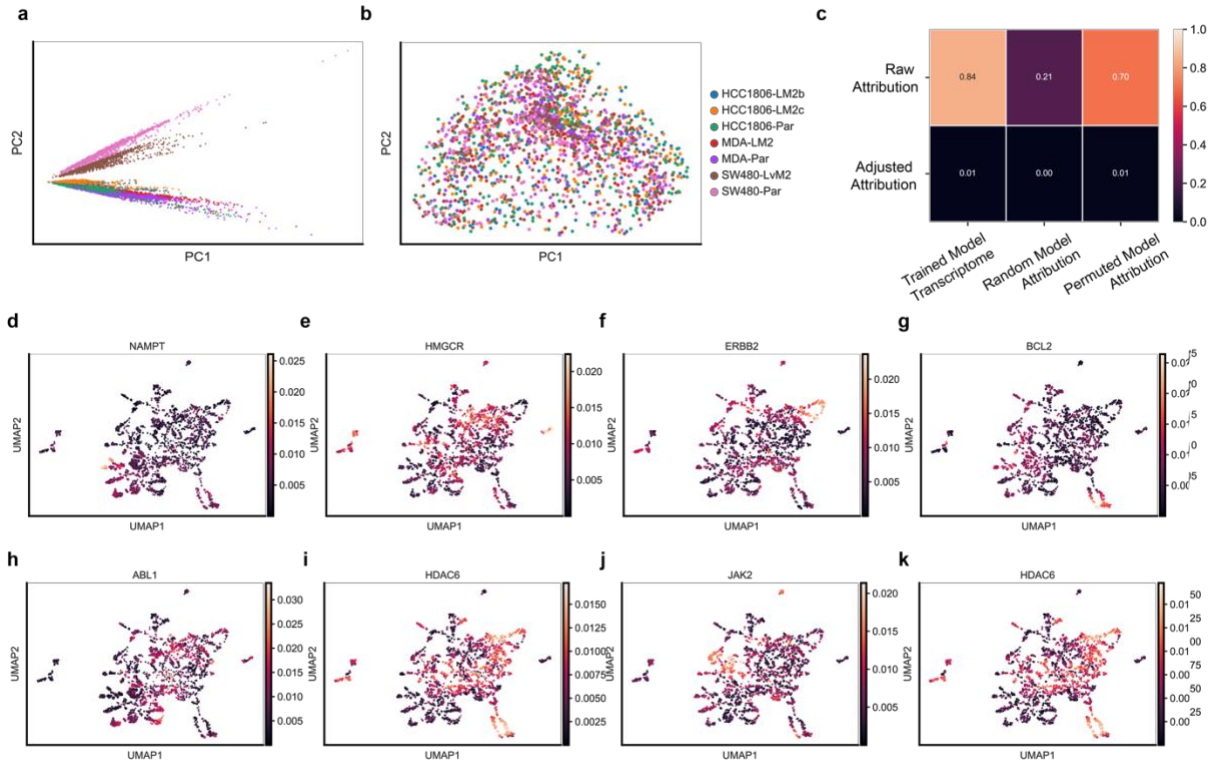
(e) *In vitro* dose-response relationships of HCC1806 and MDAMB231 to ML162.

(f) *In vitro* dose-response relationships of HCC1806 and MDAMB231 to ceranib-2.

(g) *In vitro* dose-response relationships of HCC1806 and MDAMB231 to neratinib.

(h) Relationship between predicted differential IC50 and true differential IC50.

(i) Relationship between predicted IC50 and true IC50.



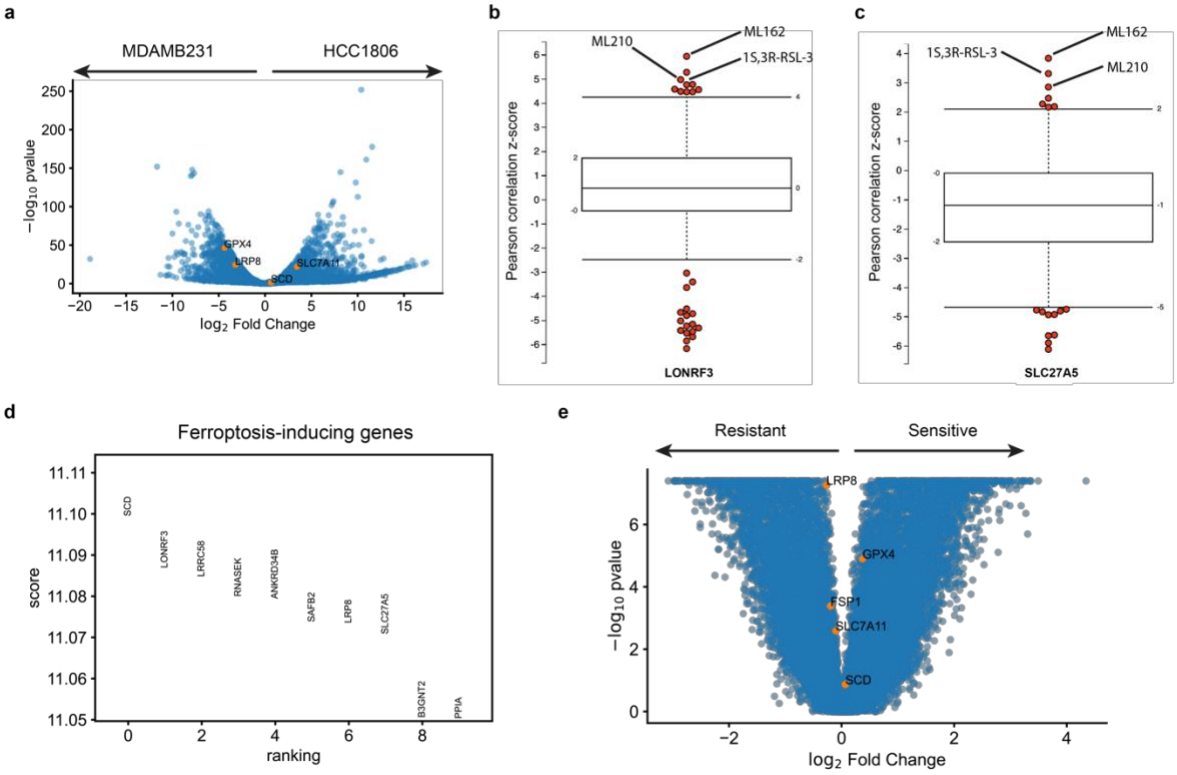
Supplementary Figure 2-5. Attribution vector analysis.

- **(a)** PCA decomposition of raw attribution vectors colored by cell line.
- **(b)** PCA decomposition of adjusted attribution vectors colored by cell line.
- **(c)** Heatmap of correlations between raw attributions (top row) and adjusted attributions (bottom row) and input transcriptomes (column 1), random model attributions (columns 2), and permuted model attributions (column 3).
- **(d)** UMAP decomposition of adjusted attributions colored by nominal target dependent attribution for NAMPT.
- **(e)** UMAP decomposition of adjusted attributions colored by nominal target dependent attribution for HMGCR.
- **(f)** UMAP decomposition of adjusted attributions colored by nominal target dependent attribution for ERBB2.
- **(g)** UMAP decomposition of adjusted attributions colored by nominal target dependent attribution for BCL2.
- **(h)** UMAP decomposition of adjusted attributions colored by nominal target independent attribution for ABL1.

(i) UMAP decomposition of adjusted attributions colored by nominal target independent attribution for HDAC6.

(j) UMAP decomposition of adjusted attributions colored by nominal target independent attribution for JAK2.

(k) UMAP decomposition of adjusted attributions colored by nominal target independent attribution for HDAC6.



Supplementary Figure 2-6.

(a) Volcano plot of DEA results between MDAMB231 and HCC1806 cell lines. Known ferroptosis-mediating genes highlighted in orange.

(b) Boxplot of Z-score Pearson correlation between ferroptosis-inducing compounds and LONRF3 expression.

(c) Boxplot of Z-score Pearson correlation between ferroptosis-inducing compounds and SLC27A5 expression.

(d) Top-10 DAA derived ferroptosis-inducing genes, ranked by Wilcoxon rank-sum Z-score.

(e) Volcano plot of DAA results between ferroptosis-resistant cluster and ferroptosis-sensitive cluster attribution vectors. Known ferroptosis-mediating genes highlighted in orange.

Chapter 3: A single-cell gene expression language model

3.1 Summary

Gene regulation is a dynamic process that connects genotype and phenotype. Given the difficulty of physically mapping mammalian gene circuitry, we require new computational methods to learn regulatory rules. Natural language is a valuable analogy to the communication of regulatory control. Machine learning systems model natural language by explicitly learning context dependencies between words. We propose a similar system applied to single-cell RNA expression profiles to learn context dependencies between genes. Our model, Exceiver, is trained across a diversity of cell types using a self-supervised task formulated for discrete count data, accounting for feature sparsity. We found agreement between the similarity profiles of latent sample representations and learned gene embeddings with respect to biological annotations. We evaluated Exceiver on a new dataset and a downstream prediction task and found that pretraining supports transfer learning. Our work provides a framework to model gene regulation on a single-cell level and transfer knowledge to downstream tasks.

3.2 Introduction

Many biological processes regulate the relationship between genotype and phenotype. On one hand, classical genetics defines simple hereditary rules. On the other hand, complex regulatory networks mediate response to the environment. Eventually, we may hope to model molecular circuitry comprehensively to accurately predict phenotypes. In this direction, learned generalizations of biological processes may support medical interventions such as individual disease risk prediction and patient therapy selection.

Large-scale cellular assays capture snapshots of complex and dynamic biological processes such as gene regulation, with RNA abundances being one measurable outcome. Single-cell RNA sequencing (scRNA-seq) observations can relate cellular states and mRNA expression relationships, revealing gene programs corresponding to disease processes, genetic perturbations, and therapeutic interventions[67]. Given the difficulty of physically mapping regulatory circuitry explicitly, we hypothesized a model trained on a large volume of transcriptomic profiles would instead implicitly learn RNA expression dependencies that reflect regulatory logic.

Pretrained models in natural language processing, computer vision, and protein modeling motivate a similar approach in systems biology[17,68,69]. Pretrained models that transfer to downstream tasks share three components leveraging domain-specific inductive biases. First, sufficient unlabeled data volumes provide enough information for highly parameterized models to learn complex relationships between features. Second, models learn from unlabeled data in a self-supervised manner, often by feature masking, in which unmasked features are used to predict a fraction of values that are masked. Third, an attention mechanism learns the dependencies between features. Traditionally, a transformer applies self-attention to learn context-dependent feature representations. Given the success of this recipe across various domains, we propose to model gene regulation similarly.

Building on sequence modeling, Exceiver (Expression-Perceiver) is a single-cell gene expression language model pretrained on an atlas of transcriptomic data. We leveraged the Perceiver IO framework to train a long-context sequence model on all protein-coding genes in a

self-supervised manner[70,71]. We evaluated latent sample representations with respect to metadata labels including cell compartment, tissue, and cell ontology. We analyzed the similarity of learned gene embeddings relative to known molecular interactions. Finally, we assessed pretrained Exceiver models on new datasets and a downstream task. Exceiver provides a framework to learn gene regulatory logic from unlabeled single-cell transcriptomes and transfer knowledge to new domains.

3.3 Results

I. Exceiver accounts for discrete features and technical dropout in scRNA-seq self-supervised pretraining

Exceiver builds on Perceiver IO to encode single-cell transcriptomic profiles. Perceiver IO scales linearly with the size of inputs and outputs, allowing tractable Transformer-based encoding and decoding of long-context sequences. Exceiver retains the core Perceiver IO architectural components: a cross-attention encoder, a self-attention latent process module, and a cross-attention decoder (**Figure 3-1**; Methods).

Exceiver extends this general architecture to accommodate various biological and experimental priors. To account for discrete RNA abundances, Exceiver represents individual genes as learnable embeddings. Global gene embedding vectors are scaled by expression values to incorporate observed RNA counts. Exceiver can also augment the learning processes with prior knowledge through the integration of experimental metadata. An auxiliary classification option encodes sample class labels, such as tissue type, as an additional embedding. In addition to biological priors, the Exceiver framework acknowledges that technical dropout influences

scRNA-seq experimental measurements. Exceiver masks attention computation at these values to mitigate feature sparsity biases in the learning process (**Figure 3-1**; Methods).

To leverage unlabeled single-cell atlases for pretraining, we propose discrete noise masking (DNM), a novel self-supervised task for count-based RNA abundances. DNM randomly chooses a fraction of genes to mask each time a cell is sampled for training. A mask embedding replaces sampled gene embeddings and a noised expression value replaces the true expression value. In our experiments DNM simply samples the mean of the feature distribution for expression noising; however, DNM may extend to other distributions. (**Figure 3-1**; Methods).

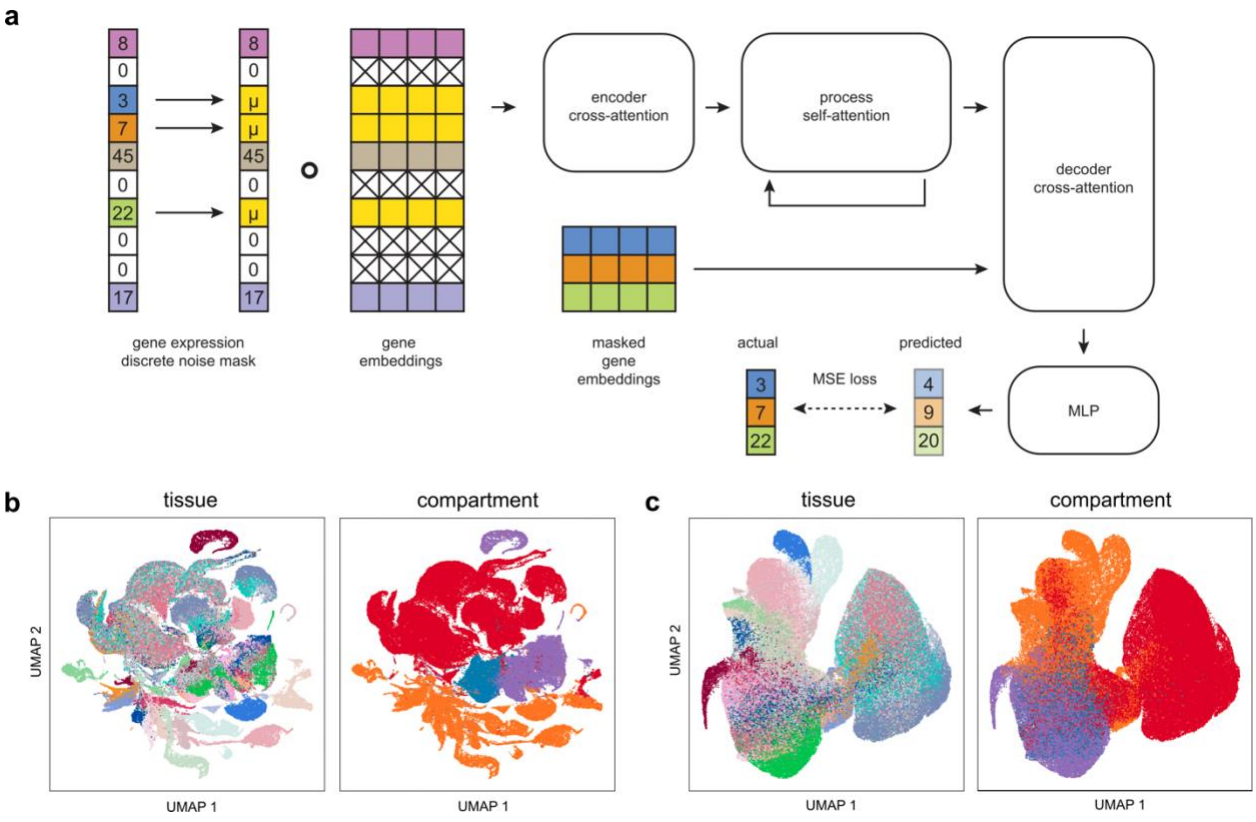


Figure 3-1. Exceiver learns cell embeddings reflecting tissue and compartment.

(a) Architectural overview and pretraining strategy.

(b) UMAP of original data PCA embeddings colored by tissue type and compartment.

(c) UMAP of Exceiver sample embeddings colored by tissue type and compartment.

II. Exceiver learns representations that reflect known biology

We trained Exceiver on the Tabula Sapiens, a healthy human atlas of approximately 500,000 single cells from 24 organs of 15 individuals[72]. We randomly split the data into 70% training and 30% validation sets and pretrained Exceiver with DNM and tissue identity as an auxiliary task. The model converged to an explained variance (EV) of approximately 0.73 and an accuracy of 0.61 on the validation dataset.

To assess whether Exceiver learned sample representations that reflect biological relationships, we evaluated latent representation similarity relative to that of metadata annotations. Qualitatively, both the original data and Exceiver latent representations reflected tissue of origin and cell compartment (evolutionary lineage) (**Figure 3-1**). We applied k-means clustering to samples and computed the adjusted mutual information (AMI) between derived cluster labels and true labels. Exceiver latent representations achieved a considerably higher AMI than original samples by tissue, donor, and compartment labels (**Table 3-1**). Given its role as the auxiliary classification task, we expected latent representations to cluster by tissue. However, structure also increased for donor and compartment labels. Cell ontology, a fine-grained label, saw decreased clustering relative to original samples. Exceiver's learned sample representations reflected known relationships and led us to interrogate similar structure in learned gene embeddings.

Table 3-1. Latent sample representations reflect biological annotations.
AMI was computed between k-means clustering derived labels and true labels of original sample embeddings and Exceiver latent sample representations. (AMI, adjusted mutual information).

sample label	original samples (AMI)	latent representations (AMI)
tissue	0.25	**0.42**
method	0.00	0.00
donor	0.08	**0.16**
cell ontology	**0.56**	0.43
compartment	0.07	**0.44**
gender	0.01	0.00

We hypothesized that similarity profiles of learned gene embeddings may reflect known gene associations. To analyze learned gene relationships, we extracted the vocabulary of global gene embeddings and applied unsupervised Leiden clustering (**Figure 3-2**). Then, we queried the STRING database with the gene list from each cluster and calculated network enrichment[61]. 66% (39/59) of Exceiver gene clusters had more interactions than expected by chance (**Figure 3-2**). Additionally, we investigated functional gene associations through Gene Ontology (GO). For example, cluster 17 reflected significant enrichment of GO process terms associated with muscle function (**Figure 3-2**, **Supplementary Figure 3-1**). In another case, an extremely enriched cluster contained a large portion of ribosomal genes (**Supplementary Figure 3-1**). Overall, Exceiver gene embeddings reflected network biology associations and captured functional gene relationships.

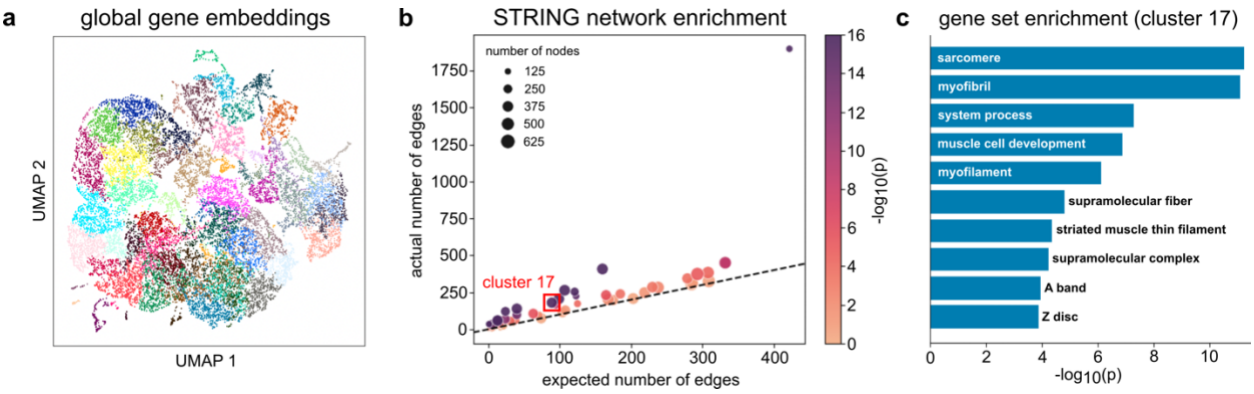


Figure 3-2. Learned gene embedding similarity profiles reflect network biology.

(a) UMAP of global gene embeddings colored by Leiden cluster.

(b) STRING network enrichment plot of gene clusters.

(c) GSEA functional enrichment analysis of cluster 17.

III. Pretrained Exceiver transfers to new datasets and predicts drug response

Next, we evaluated the transferability of a pretrained Exceiver model to a new dataset. We expected that, compared to an untrained Exceiver model, a pretrained model would more rapidly converge to higher performance on a new dataset. We tested this on a scRNA-seq dataset generated at a different time, in a different lab, of a novel disease physiology. Bi et al. investigated tumor and immune cell reprogramming of patients treated with immune checkpoint blockade for metastatic renal cell carcinoma[67]. Biopsies from 8 patients were characterized by scRNA-seq for 35,000 cells. We trained five randomly-initialized Exceiver models on this dataset with DNM and no auxiliary task. We likewise fine-tuned five pretrained Exceiver models. Prior to fine-tuning, the pretrained Exceiver models predicted masked expression values with an average EV of 0.52. The models then converged to an average EV of 0.94 in under 10 epochs. This is in contrast to the baseline models, which converged to an

average EV of 0.73 over a 30x longer training interval (~350 epochs) (**Figure 3-3**). An Exceiver model pretrained on the self-supervised DNM task learns information that is sufficiently general to apply across new datasets and biological contexts.

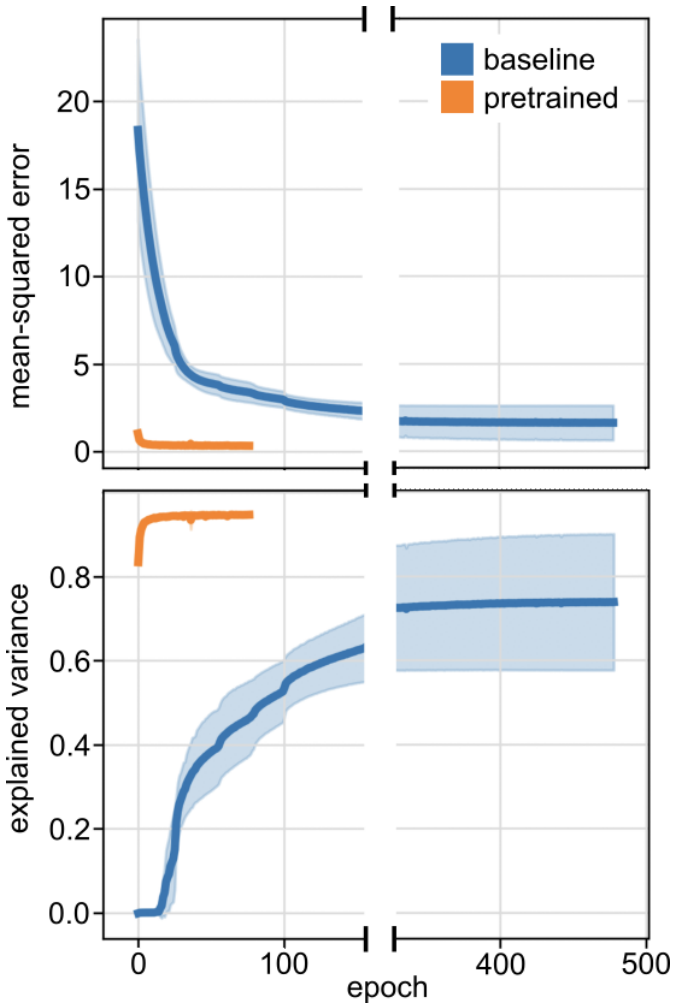


Figure 3-3. Pretrained Exceiver encodes a new dataset.
Validation loss and explained variance curves of baseline and pretrained Exceiver models. Models were trained and evaluated with DNM on Bi et al. (DNM, discrete noise masking).

Additionally, we hypothesized that a pretrained Exceiver model would support transfer learning to a new downstream task. We turned to a MIX-seq dataset of pooled cell lines that

were transcriptionally characterized after drug treatments. As in the original study, we matched drug response (quantified by area under the dose-response curve) from the Genomics of Drug Sensitivity in Cancer Project (GDSC) screen to MIX-seq transcriptional profiles[73,74]. We trained five baseline and pretrained Exceiver models on each drug-treated cell line pool (Methods). Consistent with the original study, neither approach learned a relationship between post-perturbation transcriptional profiles and drug responses for navitoclax, taselisib, bortezomib, or gemcitabine, likely due to matching drug responses from an entirely different dataset. By contrast, trametinib and dabrafenib treated cell line pools matched the most samples and here the pretrained Exceiver succeeded whereas the baseline Exceiver did not (p<5e-2, Mann-Whitney U; **Table 3-2**). Noting the high training variances of the baseline models, further optimization may be warranted, but it is clear that the pretrained Exceiver model was the more robust learner in this challenging scenario.

Table 3-2. Pretrained Exceiver predicts drug response.
Explained variance (EV) and standard error of baseline and pretrained drug response prediction Exceiver models. MIX-Seq post-perturbation transcriptomic profiles of pooled single-cell drug screens were trained to predict AUC from GDSC screens. (AUC, area under the curve; GDSC, Genomics of Drug Sensitivity in Cancer Project).

Drug	n samples	baseline (EV)	pretrained (EV)	*p*-value
dabrafenib	6744	-0.4566 ± 0.08904	0.4869 ± 0.1780	0.007937
trametinib	6696	-0.2071 ± 0.1461	0.1394 ± 0.008390	0.007937
navitoclax	5910	-1.028 ± 0.3961	-0.5686 ± 0.1491	0.3100
taselisib	1327	0.004383 ± 0.004002	0.02918 ± 0.05300	0.8413
bortezomib	913	0.01568 ± 0.01057	0.05706 ± 0.06736	0.2222
gemcitabine	736	0.01773 ± 0.01340	0.01530 ± 0.04873	0.4206

3.4 Discussion

We present Exceiver, a single-cell gene expression language model, whose attention-based transformer backbone encodes long-context transcriptomic profiles. We introduce discrete noise masking, a procedure that masks expression values and enables self-supervised learning on unlabeled, continuously-valued datasets. We show that an Exceiver model trained on the Tabula Sapiens with a self-supervised task learns low dimensional representations that reflect sample annotations. Moreover, learned gene embedding similarity reflects molecular network interactions and functional associations. Finally, we find that a pretrained Exceiver

model transfers to new datasets and a drug response prediction task. Exceiver provides a framework to leverage publicly available scRNA-seq datasets and learn robust gene regulatory logic across diverse biological contexts. Exceiver may provide utility in transferring systems knowledge to downstream tasks, from the interrogation of molecular functions to the prediction of comprehensive phenotypes.

3.5 Materials and methods

I. Model architecture

Perceiver IO is a general purpose model architecture that adapts to any task with structured input and output. Since self-attention complexity scales quadratically with input size, it cannot be directly applied to high-dimensional data, such as scRNA-seq readouts. Perceiver IO addresses this issue by introducing cross-attentional encoder and decoder mechanisms that project to and from a lower-dimensional latent space where full self-attention can be applied. Each of the blocks in the Perceiver IO architecture are transformer-style modules characterized by query-key-value attention followed by a multilayer perceptron (MLP) and residual connections. For more details please reference the original Perceiver IO manuscript.

Exceiver implicitly models gene expression as a discrete variable sampled from a count-based distribution. Global embeddings are used to represent gene identities, which are scaled by expression values upon model input. Exceiver also accounts for dropout, a sequencing bias characterized by experimental failure to capture the comprehensive set of genes expressed in an individual cell. This limitation implies that unobserved genes may be either truly unexpressed or merely unmeasured. Exceiver masks the attention computation at zero-valued

expression positions to prevent learning based on this artifact. Upon decoding, Exceiver reuses unscaled gene embeddings in the output query matrix as a means to residually connect discrete gene identities within the learning process. Exceiver also provides an option to bias learning with an auxiliary classification task. An additional classification embedding token is concatenated to each input sequence and used to query the decoder. An auxiliary MLP classifies the output of the decoder from the classification token, and an additional cross-entropy loss is added to the mean squared error from the DNM task. Code is available at https://github.com/keiserlab/exceiver.

II. Pretraining tasks

Exceiver employs a self-supervised pretraining task as well as an optional auxiliary supervised classification task. The self-supervised task draws inspiration from masking tasks applied for NLP pretraining. However, there are several crucial differences between gene expression profiles and natural language. First, scRNA-seq features are a set rather than a sequence. Exceiver does not positionally encode gene embeddings, though features that describe spatial gene dependencies (such as the linear position of a gene or its relative position in the three-dimensional chromosome) may prove valuable in future work. Second, as previously referenced, expression features are counts of discrete features (genes), rather than identities of discrete features (words). As such, pretraining is a regression task rather than a classification task. Exceiver employs a new self-supervised task to account for the discrete distributions of features in the gene set.

The primary pretraining task we apply is discrete noise masking (DNM). First, we heuristically determine the number of genes to mask for each sample. We mask 15% of the median number of recorded genes across the training dataset. We avoid masking a majority fraction of genes captured in poorly sampled cells by removing cells below a minimum number of recorded genes. In our experiments, we set this threshold at twice the number of masked genes. The training data therefore reflects a distribution of masked genes not exceeding 50% of the observed genes in a given cell (**Supplementary Figure 3-2**). A "mask" embedding replaces the randomly selected gene embeddings, and corresponding expression values are noised. Masked gene embeddings scaled by noised expression values pass through encoder and process modules. Finally, the embeddings of masked genes query the decoder. The resulting "contextual gene embeddings" pass through the final MLP, outputting expression predictions for masked genes (**Figure 3-1**).

Additionally, we propose an auxiliary classification task based on metadata labels. The Perceiver IO architecture is flexible enough to accommodate multimodal tasks. Exceiver implements an auxiliary classification task as described in the original paper. A classification token is initialized as a global embedding and passed through Exceiver with each gene set. This procedure allows for attention computation between all genes and the classification token. The original classification embedding queries the decoder along with masked genes, and the resulting "contextual classification embedding" is passed through an auxiliary MLP head, outputting a vector of class logits.

III. Data and processing

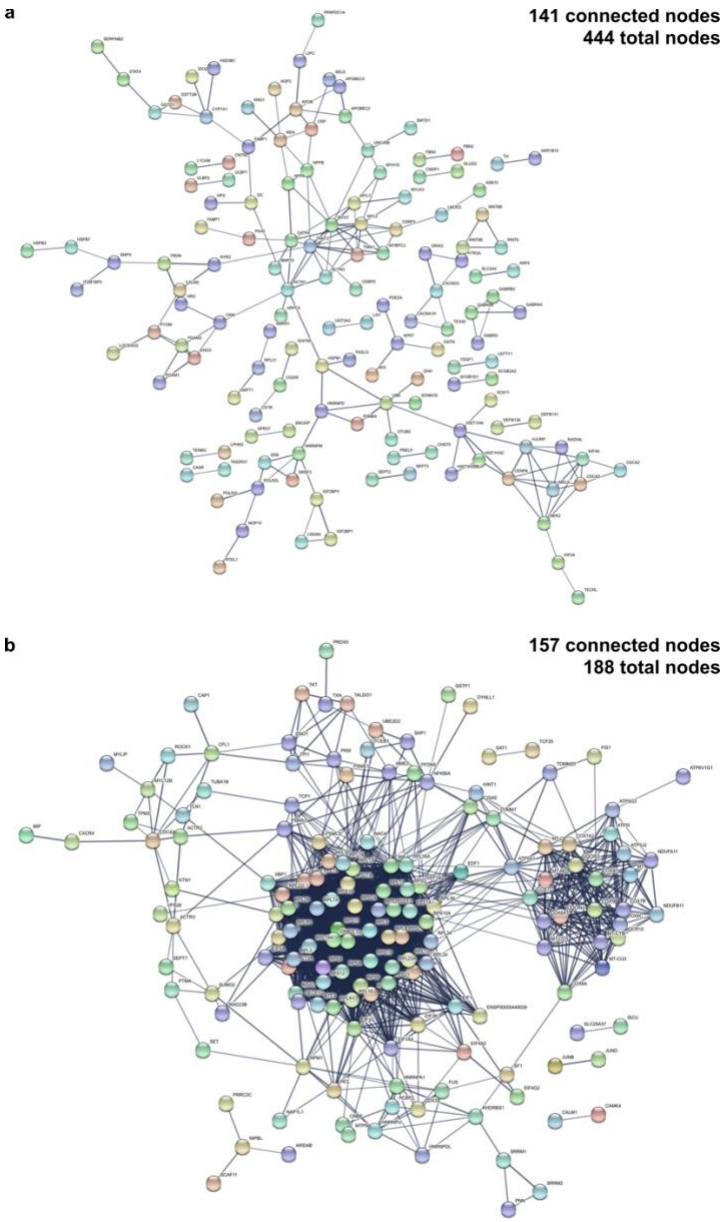
The Tabula Sapiens (TS) data was retrieved from figshare (https://figshare.com/ndownloader/files/34702114). Python packages anndata and scanpy were used for data storage and statistics calculations. Features were subsetted to all protein-coding genes as defined in the Cancer Cell Line Encyclopedia (CCLE) (s=19067). DecontX-corrected count matrices were used for training to account for ambient RNA. TS was shuffled and split into 70% training and 30% validation datasets. Counts were normalized to a maximum of 1e4 per cell. Normalized counts were log-transformed and scikit-learn was used to apply a Z-score transformation to each dataset individually. Expression values were shifted to a mean-center of 1, implying the learned identity of global gene embeddings corresponds to average expression.

The Bi et al. data was retrieved from the Human Cell Atlas collection (https://singlecell.broadinstitute.org/single_cell/study/SCP1288). Raw count matrices were split into 70% training and 30% validation, and the same normalization and transformation procedure was applied. Missing genes were imputed with a value of zero.

The MIX-seq data from McFarland et al. was retrieved from figshare (https://figshare.com/s/139f64b495dea9d88c70). Sanger GDSC2 AUC data was downloaded from the Dependency Map (https://depmap.org/portal/download/custom/). Drugs from 24-hr treatment experiments with sufficient matching GDSC2 experiments were selected for analysis. Post-perturbation transcriptome datasets were preprocessed individually in the manner as

described previously with zero-value missing gene imputation. Cell line-drug transcriptome pairs were matched to dose-response curve AUC values of GDSC2 drug screening results.

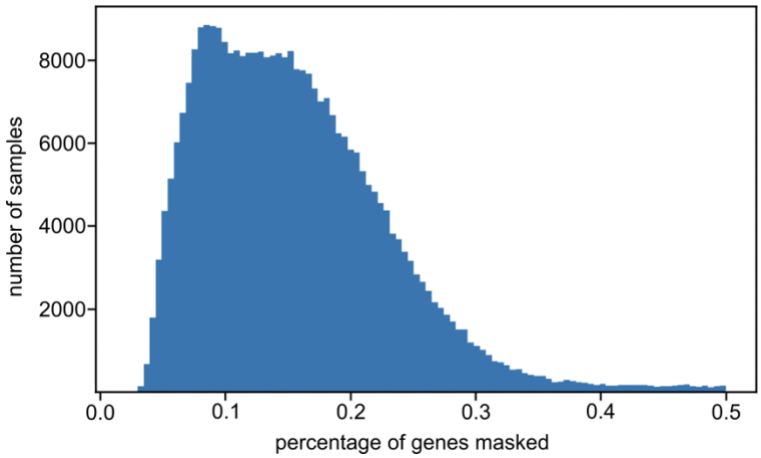
3.6 Supplementary materials



Supplementary Figure 3-1. Example clusters visualized using STRING.

(a) Cluster 17 connected nodes from the STRING network of high confidence interactions (interaction score > 0.7).

(b) Cluster 45 connected nodes from the STRING network of high confidence interactions (interaction score > 0.7)



Supplementary Figure 3-2. Distribution of masked genes.
Histogram of the percentage of genes masked across the Tabula Sapiens training dataset. Feature sparsity results in a distribution of masked genes per sample.

Discussion

Our analysis of psoriasis patients treated with ustekinumab discovered a genetic variant on chromosome 4 associated with drug response. This signal was validated in an independent cohort at multiple clinical endpoints, however, the causality of this variant remains unclear. rs35569429 appears to be a complementary signal to the major psoriasis risk allele HLA-C*06:06 – ustekinumab responders and nonresponders were stratified most when both variants were considered. Association studies such as this will remain a critical tool for discovery research in pharmacogenomics. Most notably, they afford computational efficiency and are easily interpretable. Discovery of strong associations between genetics and phenotype requires large sample sizes. Even then, large studies have found that genetic associations still only explain a portion of phenotypic variance for complex diseases[75–77]. Even when interpretable univariate signals are combined into Polygenic Risk Scores, genetics alone fails to predict complex phenotypes. These limitations motivate the need for efficient multivariate models that learn to combine complementary data types.

ChemProbe advances this direction by explicitly learning to incorporate chemical structure and concentration with gene expression information. ChemProbe successfully predicted the sensitivity of both clinical samples and engineered cell lines to a large set of chemical probes. In one example, ChemProbe predicted that an LRP8 gene knockout cell line would be more susceptible to chemical induction of ferroptosis than a wild-type cell line. This experiment demonstrates the use of ChemProbe to unravel the function of particular genes or proteins in chemically induced mechanisms. CRISPR technology affords the opportunity to

perform whole-genome knockouts of disease models followed by RNA sequencing. ChemProbe can screen each knockout transcriptome for changes in sensitivity to the set of chemical probes. Such an *in silico* chemical screen may reveal disease-specific mechanisms of gene sensitivity or resistance to protein targets. This approach may be extended to reduce the experimental burden of any perturbation paired with the set of chemical probes. For example, ChemProbe can screen the transcriptome of a drug-treated cell line to test for chemical combinations that result in sensitivity changes. These use cases are not without limitations: given that ChemProbe was trained on bulk RNA-seq of cancer cell lines, biological models with sufficiently different gene expression patterns (due to unique biology or experimental effects) may cause model failure. Although we found evidence that ChemProbe generalizes to microarray technology, an immediate question is whether ChemProbe will apply to single-cell RNA sequencing data and be able to take advantage of highly scalable assays such as Perturb-seq.

The failure of machine learning models to generalize beyond a distribution or to a new task is a major challenge in precision medicine. Most labeled datasets are relatively small and highly specific to a task, question, or disease. Models only trained within a supervised learning paradigm will remain brittle and clinically unusable[78]. The path to generalizable models that gain FDA approval and adoption in the clinic will likely leverage large, unlabeled datasets for learning. We applied a self-supervised learning approach to single-cell RNA sequencing data and found evidence that our model, Exceiver, learned patterns that reflect known biology and that this knowledge transferred to new datasets and modeling tasks. The generalizability of Exceiver requires further investigation to understand if pretraining meaningfully increased

performance on challenging downstream tasks such as disease risk modeling. It remains unclear how much information a model requires to sufficiently generalize to new tasks through either transfer learning or zero-shot prediction. If domain-specific challenges are addressed, such "foundation" models may prove as valuable for precision medicine as they have for natural language processing[79]. For one, we must rigorously evaluate how the variance of experimental measurements may bias large models. High experimental variance is a problem unique to molecular biology assays. Since building large datasets requires integrating experimental data from diverse assays, protocols, and model systems, the community must determine if large models learn to regularize batch effects or if we require more principled methods to do so. Finally, the computational resources and technical expertise required to collect, integrate, store data, and train large models will be a challenge for individual operators. The community will depend on large collaborations, such as the OpenFold Consortium, to ensure the most powerful predictive models of biology are publicly accessible and usable for research.

A well-specified model that reflects true biology is more robust than a model that relies on correlative signals. To build trust in predictive models and garner biological insights, we must understand if well-performing models rely on causal relationships to make predictions[9,80]. In our association study, we hypothesized that rs35569429 may be a *cis*-acting regulator of WDR1 or SLC2A9. Although the literature reflects some support for these hypotheses, each requires experimental testing. This story is true for any model we wish to interpret and use to understand biology. Neural networks, such as ChemProbe, pose a more daunting challenge to interpret than linear models. We applied a model interpretation method

and evaluated if feature attributions depended on the learned model weights or the relationship between features and labels. We found this to be the case and applied similarity analyses to attribution vectors, finding that they reflected compound mechanisms of action. Such experiments that evaluate model interpretation results with respect to the model, data, and known biology, are critical for determining if models reflect true biological processes. If so, we may consider interpreting such models and posing new hypotheses. The machine learning community requires novel model interpretation methods, and experimental approaches that eliminate alternative hypotheses, to successfully do so. Research into causal models that reflect data generation processes and biological priors remains a promising direction for model interpretation.

The promise of precision medicine remains, and our ability to accurately predict phenotypes will change medical practice and ultimately health outcomes. In parallel, understanding how such models predict complex phenotypes poses opportunities to discover new therapeutic interventions and engineer biology.

References

1. Wilson JF, Weale ME, Smith AC, Gratrix F, Fletcher B, Thomas MG, et al. Population genetic structure of variable drug response. Nat Genet. 2001;29: 265–269.

2. Bachtiar M, Lee CGL. Genetics of Population Differences in Drug Response. Curr Genet Med Rep. 2013;1: 162–170.

3. McInnes G, Yee SW, Pershad Y, Altman RB. Genomewide Association Studies in Pharmacogenomics. Clin Pharmacol Ther. 2021;110: 637–648.

4. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. Nat Rev Genet. 2019;20: 467–484.

5. Daly AK. Genome-wide association studies in pharmacogenomics. Nat Rev Genet. 2010;11: 241–246.

6. Feng F, Shen B, Mou X, Li Y, Li H. Large-scale pharmacogenomic studies and drug response prediction for personalized cancer medicine. J Genet Genomics. 2021;48: 540–551.

7. Ashley EA. Towards precision medicine. Nat Rev Genet. 2016;17: 507–522.

8. Kosorok MR, Laber EB. Precision Medicine. Annu Rev Stat Appl. 2019;6: 263–286.

9. Zhang Y, Tiňo P, Leonardis A, Tang K. A Survey on Neural Network Interpretability. IEEE Transactions on Emerging Topics in Computational Intelligence. 2021;5: 726–742.

10. D'Adamio S, Silvaggio D, Lombardo P, Bianchi L, Talamonti M, Galluzzo M. The safety of

anti-interleukins monoclonal antibodies for the treatment of psoriasis. Expert Opin Drug Saf. 2019;18: 1031–1041.

11. Connell WT, Hong J, Liao W. Genome-Wide Association Study of Ustekinumab Response in Psoriasis. Front Immunol. 2021;12: 815121.

12. Arrowsmith CH, Audia JE, Austin C, Baell J, Bennett J, Blagg J, et al. The promise and peril of chemical probes. Nat Chem Biol. 2015;11: 536–541.

13. Connell WT, Keiser MJ. Predicting Cellular Drug Sensitivity using Conditional Modulation of Gene Expression. Learning Meaningful Representations of Life Workshop, NeurIPS. 2020 [cited 7 Dec 2020]. doi:10.1101/2021.03.15.435529

14. Barbiero P, Squillero G, Tonda A. Modeling Generalization in Machine Learning: A Methodological and Computational Study. arXiv [cs.LG]. 2020. Available: http://arxiv.org/abs/2006.15680

15. Quinonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND. Dataset Shift in Machine Learning. MIT Press; 2008.

16. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv [cs.CL]. 2018. Available: http://arxiv.org/abs/1810.04805

17. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language Models are Few-Shot Learners. Adv Neural Inf Process Syst. 2020. pp. 1877–1901.

18. Du Y, Liu Z, Li J, Zhao WX. A Survey of Vision-Language Pre-Trained Models. arXiv [cs.CV]. 2022. Available: http://arxiv.org/abs/2202.10936

19. Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, et al. A Comprehensive Survey on Transfer Learning. Proc IEEE. 2021;109: 43–76.

20. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. PNAS. 2021. doi:10.1101/622803

21. Rao R, Meier J, Sercu T, Ovchinnikov S, Rives A. Transformer protein language models are unsupervised structure learners. bioRxiv. 2020. p. 2020.12.15.422761. doi:10.1101/2020.12.15.422761

22. Hsu C, Verkuil R, Liu J, Lin Z, Hie B, Sercu T, et al. Learning inverse folding from millions of predicted structures. bioRxiv. 2022. Available: https://www.biorxiv.org/content/10.1101/2022.04.10.487779.abstract

23. Nijkamp E, Ruffolo J, Weinstein EN, Naik N, Madani A. ProGen2: Exploring the Boundaries of Protein Language Models. arXiv [cs.LG]. 2022. Available: http://arxiv.org/abs/2206.13517

24. Connell W, Khan U, Keiser MJ. A single-cell gene expression language model. Learning Meaningful Representations of Life Workshop, NeurIPS. 2022. doi:10.48550/arXiv.2210.14330

25. Parisi R, Symmons DPM, Griffiths CEM, Ashcroft DM, Identification and Management of Psoriasis and Associated ComorbidiTy (IMPACT) project team. Global epidemiology of psoriasis: a systematic review of incidence and prevalence. J Invest Dermatol. 2013;133: 377–385.

26. Warren RB, Smith CH, Yiu ZZN, Ashcroft DM, Barker JNWN, Burden AD, et al. Differential Drug Survival of Biologic Therapies for the Treatment of Psoriasis: A Prospective Observational Cohort Study from the British Association of Dermatologists Biologic Interventions Register (BADBIR). J Invest Dermatol. 2015;135: 2632–2640.

27. Mahil SK, Capon F, Barker JN. Update on psoriasis immunopathogenesis and targeted immunotherapy. Semin Immunopathol. 2016;38: 11–27.

28. Leonardi CL, Kimball AB, Papp KA, Yeilding N, Guzzo C, Wang Y, et al. Efficacy and safety of ustekinumab, a human interleukin-12/23 monoclonal antibody, in patients with psoriasis: 76-week results from a randomised, double-blind, placebo-controlled trial (PHOENIX 1). Lancet. 2008;371: 1665–1674.

29. Papp KA, Langley RG, Lebwohl M, Krueger GG, Szapary P, Yeilding N, et al. Efficacy and safety of ustekinumab, a human interleukin-12/23 monoclonal antibody, in patients with psoriasis: 52-week results from a randomised, double-blind, placebo-controlled trial (PHOENIX 2). Lancet. 2008;371: 1675–1684.

30. Young MS, Horn EJ, Cather JC. The ACCEPT study: ustekinumab versus etanercept in moderate-to-severe psoriasis patients. Expert Rev Clin Immunol. 2011;7: 9–13.

31. Li K, Huang CC, Randazzo B, Li S, Szapary P, Curran M, et al. HLA-C*06:02 Allele and Response to IL-12/23 Inhibition: Results from the Ustekinumab Phase 3 Psoriasis Program. J Invest Dermatol. 2016;136: 2364–2371.

32. Talamonti M, Galluzzo M, van den Reek JM, de Jong EM, Lambert JLW, Malagoli P, et al. Role of the HLA-C*06 allele in clinical response to ustekinumab: evidence from real life in a large cohort of European patients. Br J Dermatol. 2017;177: 489–496.

33. Talamonti M, Galluzzo M, Chimenti S, Costanzo A. HLA-C*06 and response to ustekinumab in Caucasian patients with psoriasis: Outcome and long-term follow-up. J Am Acad Dermatol. 2016;74: 374–375.

34. Chiu H-Y, Wang T-S, Chan C-C, Cheng Y-P, Lin S-J, Tsai T-F. Human leucocyte antigen-Cw6 as a predictor for clinical response to ustekinumab, an interleukin-12/23 blocker, in Chinese patients with psoriasis: a retrospective analysis. Br J Dermatol. 2014;171: 1181–1188.

35. van Vugt LJ, van den Reek JMPA, Hannink G, Coenen MJH, de Jong EMGJ. Association of HLA-C*06:02 Status With Differential Response to Ustekinumab in Patients With Psoriasis: A Systematic Review and Meta-analysis. JAMA Dermatol. 2019;155: 708–715.

36. Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database . 2017;2017. doi:10.1093/database/bax028

37. Pfajfer L, Mair NK, Jiménez-Heredia R, Genel F, Gulez N, Ardeniz Ö, et al. Mutations affecting the actin regulator WD repeat–containing protein 1 lead to aberrant lymphoid

immunity. J Allergy Clin Immunol. 2018;142: 1589-1604.e11.

38. Standing ASI, Malinova D, Hong Y, Record J, Moulding D, Blundell MP, et al. Autoinflammatory periodic fever, immunodeficiency, and thrombocytopenia (PFIT) caused by mutation in actin-regulatory gene WDR1. J Exp Med. 2017;214: 59–71.

39. Kuhns DB, Fink DL, Choi U, Sweeney C, Lau K, Priel DL, et al. Cytoskeletal abnormalities and neutrophil dysfunction in WDR1 deficiency. Blood. 2016;128: 2135–2143.

40. Ruiz A, Gautschi I, Schild L, Bonny O. Human Mutations in SLC2A9 (Glut9) Affect Transport Capacity for Urate. Front Physiol. 2018;9: 476.

41. Do LHD, Maibach H. Pharmacogenomics/updated for precision medicine in dermatology. J Dermatolog Treat. 2019;30: 410–413.

42. Griffiths CEM, Strober BE, van de Kerkhof P, Ho V, Fidelus-Gort R, Yeilding N, et al. Comparison of ustekinumab and etanercept for moderate-to-severe psoriasis. N Engl J Med. 2010;362: 118–128.

43. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48: 1284–1287.

44. Siva N. 1000 Genomes project. Nat Biotechnol. 2008;26: 256.

45. Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. Int J Methods Psychiatr Res. 06 2018;27: e1608.

46. Workman P, Collins I. Probing the probes: fitness factors for small molecule tools. Chem Biol. 2010;17: 561–577.

47. Blagg J, Workman P. Choose and Use Your Chemical Probe Wisely to Explore Cancer Biology. Cancer Cell. 2017;32: 9–25.

48. Roden DM, McLeod HL, Relling MV, Williams MS, Mensah GA, Peterson JF, et al. Pharmacogenomics. Lancet. 2019;394: 521–532.

49. Tyner JW, Haderk F, Kumaraswamy A, Baughn LB, Van Ness B, Liu S, et al. Understanding Drug Sensitivity and Tackling Resistance in Cancer. Cancer Res. 2022;82: 1448–1460.

50. Rees MG, Seashore-Ludlow B, Cheah JH, Adams DJ, Price EV, Gill S, et al. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. Nat Chem Biol. 2016;12: 109–116.

51. Ghandi M, Huang FW, Jané-Valbuena J, Kryukov GV, Lo CC, McDonald ER 3rd, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature. 2019;569: 503–508.

52. Perez E, Strub F, De Vries H, Dumoulin V. Film: Visual reasoning with a general conditioning layer. arXiv preprint arXiv. 2017. Available: https://arxiv.org/abs/1709.07871

53. Chuang KV, Keiser MJ. Adversarial Controls for Scientific Machine Learning. ACS Chem Biol. 2018;13: 2819–2821.

54. Wolf DM, Yau C, Wulfkuhle J, Brown-Swigart L, Gallagher RI, Lee PRE, et al. Redefining

breast cancer subtypes to guide treatment prioritization and maximize response: Predictive biomarkers across 10 cancer therapies. Cancer Cell. 2022;40: 609-623.e6.

55. Chien AJ, Tripathy D, Albain KS, Symmans WF, Rugo HS, Melisko ME, et al. MK-2206 and Standard Neoadjuvant Chemotherapy Improves Response in Patients With Human Epidermal Growth Factor Receptor 2-Positive and/or Hormone Receptor-Negative Breast Cancers in the I-SPY 2 Trial. J Clin Oncol. 2020;38: 1059–1069.

56. Nanda R, Liu MC, Yau C, Shatsky R, Pusztai L, Wallace A, et al. Effect of Pembrolizumab Plus Neoadjuvant Chemotherapy on Pathologic Complete Response in Women With Early-Stage Breast Cancer: An Analysis of the Ongoing Phase 2 Adaptively Randomized I-SPY2 Trial. JAMA Oncol. 2020;6: 676–684.

57. Park JW, Liu MC, Yee D, Yau C, van ’t Veer LJ, Symmans WF, et al. Adaptive Randomization of Neratinib in Early Breast Cancer. N Engl J Med. 2016;375: 11–22.

58. Rugo HS, Olopade OI, DeMichele A, Yau C, van ’t Veer LJ, Buxton MB, et al. Adaptive Randomization of Veliparib-Carboplatin Treatment in Breast Cancer. N Engl J Med. 2016;375: 23–34.

59. Mantione KJ, Kream RM, Kuzelova H, Ptacek R, Raboch J, Samuel JM, et al. Comparing bioinformatic gene expression profiling methods: microarray and RNA-Seq. Med Sci Monit Basic Res. 2014;20: 138–142.

60. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity Checks for Saliency Maps. arXiv [cs.CV]. 2018. Available: http://arxiv.org/abs/1810.03292

61. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43: D447-52.

62. Stockwell BR. Ferroptosis turns 10: Emerging mechanisms, physiological functions, and therapeutic applications. Cell. 2022;185: 2401–2421.

63. Li J, Cao F, Yin H-L, Huang Z-J, Lin Z-T, Mao N, et al. Ferroptosis: past, present and future. Cell Death Dis. 2020;11: 88.

64. Li Z, Ferguson L, Deol KK, Roberts MA, Magtanong L, Hendricks JM, et al. Ribosome stalling during selenoprotein translation exposes a ferroptosis vulnerability. Nat Chem Biol. 2022; 1–11.

65. Hadian K, Stockwell BR. A roadmap to creating ferroptosis-based medicines. Nat Chem Biol. 2021;17: 1113–1116.

66. Ahmadian M, Suh JM, Hah N, Liddle C, Atkins AR, Downes M, et al. PPARγ signaling and metabolism: the good, the bad and the future. Nat Med. 2013;19: 557–566.

67. Bi K, He MX, Bakouny Z, Kanodia A, Napolitano S, Wu J, et al. Tumor and immune reprogramming during immunotherapy in advanced renal cell carcinoma. Cancer Cell. 2021;39: 649-661.e5.

68. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. Available:

https://openaccess.thecvf.com/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf

69. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med. 2018;50: 1–14.

70. Jaegle A, Gimeno F, Brock A, Zisserman A, Vinyals O, Carreira J. Perceiver: General perception with iterative attention. Proceedings of Machine Learning Research. 2021. Available: http://proceedings.mlr.press/v139/jaegle21a/jaegle21a.pdf

71. Jaegle A, Borgeaud S, Alayrac J-B, Doersch C, Ionescu C, Ding D, et al. Perceiver IO: A General Architecture for Structured Inputs & Outputs. arXiv [cs.LG]. 2021. Available: http://arxiv.org/abs/2107.14795

72. Tabula Sapiens Consortium*, Jones RC, Karkanias J, Krasnow MA, Pisco AO, Quake SR, et al. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. Science. 2022;376: eabl4896.

73. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. 2013;41: D955-61.

74. McFarland JM, Paolella BR, Warren A, Geiger-Schuller K, Shibue T, Rothberg M, et al. Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. Nat Commun. 2020;11: 1–15.

75. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461: 747–753.

76. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010;11: 446–450.

77. Young AI. Solving the missing heritability problem. PLoS Genet. 2019;15: e1008222.

78. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med. 2019;380: 1347–1358.

79. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. On the Opportunities and Risks of Foundation Models. arXiv [cs.LG]. 2021. Available: http://arxiv.org/abs/2108.07258

80. Carvalho DV, Pereira EM, Cardoso JS. Machine Learning Interpretability: A Survey on Methods and Metrics. Electronics. 2019;8: 832.

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:
William Connell
2F4CC450BDB5491...
Author Signature

12/5/2022
Date