Unifying Genetic and Chemical Perturbagen Representation through a Hybrid Deep Learning Framework

Yiming Li[1,2], Jun Zhu[2,3], Linjing Liu[2,4], Fang Wang[2], Min Zeng[1], Longkai Huang[2], Fan Yang[2,*], Min Li[1,*] , Jianhua Yao[2,*]

[1] Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha, 410083, China

[2] AI for Life Sciences Lab, Tencent, Shenzhen, China

[3] Tsinghua-Peking Joint Center for Life Sciences, School of Life Sciences, Tsinghua University, Beijing, 100084, China

[4] Department of Computer Science, City University of Hong Kong, Hong Kong, China

* Correspondence to jianhuayao@tencent.com; limin@mail.csu.edu.cn; fionafyang@tencent.com

The integration of genetic and chemical perturbations has driven transformative advances in elucidating cellular mechanisms and accelerating drug discovery. However, the lack of a unified representation for diverse perturbagen types limits comprehensive analysis and joint modeling of multi-domain perturbation agents (molecular cause space) and their resulting phenotypes (phenotypic effect spaces). Here, we present UniPert, a hybrid deep learning framework that encodes genetic and chemical perturbagens into a shared semantic representation space. UniPert employs tailored encoders to address the inherent molecular-scale differences across perturbagen types and leverages contrastive learning with experiment-driven compound-target interactions to bridge these domains. Extensive experiments validate UniPert's versatility across diverse application. The generated representations effectively capture hierarchical pharmacological relationships of perturbagens, facilitating annotations of understudied targets and compounds. UniPert can be plugged in advanced frameworks to enhance performance in both genetic and chemical perturbation outcome prediction tasks. Notably, UniPert paves the way for cross-domain perturbation modeling, driving a novel *genetic-to-chemical* perturbation transfer learning paradigm, boosting context-specific *in silico* drug screening efficiency and accelerating the development of personalized therapies.

Introduction

Advances in sequencing and CRISPR (Clustered Regularly Interspaced Short Palindromic Repeat) technologies[1–3], coupled with the development of numerous integrative analytical and computational tools for biological data[4–7], have exponentially increased our ability to detect disease-relevant genetic variants and furthered progress in precision medicine[8–11]. With the growing prevalence of target-specific basket trials[12–16], the promise of individualized treatments is becoming increasingly tangible. Perturbation-based studies[17,18], leveraging both genetic and chemical interventions, play an indispensable role in this transition[19]. The former modifying biomolecule (DNA, RNA and protein) expressions to elucidate their functions and identify potential therapeutic targets[20–22], while the latter introducing small molecules into cells to reveal their action mechanisms and facilitate novel drug discovery[23,24].

Large-scale initiatives, such as the Connectivity Map[25,26], Cancer Dependency Map[27,28], and Cell Painting Gallery[29,30], have generated vast perturbation resources across genetic and chemical domains, contributing extensive efforts to uncover connections between diseases, genes, and small molecules. Joint analysis on such cross-data resources offers great potential in accelerating data-driven personalized therapeutic development by simultaneously advancing target identification and drug discovery [31–34].

However, existing large-scale perturbation data analysis primarily focus on representing and connecting phenotypic states (effects)[35–37], largely overlooking the underlying relationships and systematic representations of perturbagens (causes). Most studies remain confined to single perturbation domains—either genetic or chemical—and, in some studies, even joint analyses rely heavily on manually curated annotations of the perturbagens[26,38], such as protein families, drug mechanisms of action (MOA), and drug-target interactions. While valuable, this approach is constrained by the scope and completeness of the available annotations, which may introduce biases and fail to account for understudied perturbagens. Moreover, the intrinsic differences and representation gap between genetic and chemical perturbagens also constrain advanced deep learning-based perturbation effect simulation methods (e.g., GEARS[39] and chemCPA[40]) to modeling data from a single perturbagen type. These limitations hinder the analysis and modeling for complex cross-domain perturbation *cause-and-effect* data and restrict the potential for comprehensive insights into cellular mechanisms, underscoring the need for a shared, interpretable, and scalable representation space that unifies diverse perturbagen types.

To date no unified representation method has been developed to simultaneously encode genetic and chemical perturbagens, primarily due to several challenges. First, the inherent molecular-scale differences of distinct perturbagen types render the application of a one-fits-all encoder architecture impractical. For instance, advanced encoders such as large language models for

proteins and small molecules[36,37], typically accept input in the form of amino acid-level sequences or atomic-level SMILES, respectively. Second, encoding these molecule sequences with significant different scales and lengths into global feature vectors of uniform dimensions can easily lose crucial local information (e.g., protein conserved regions and compound functional groups). Third, establishing biological semantic relationships between and within perturbagen types requires rich reliable empirical knowledge to minimize gaps in the joint representation space.

To tackle these challenges, we propose UniPert, a hybrid deep learning framework that encodes genetic and chemical perturbagens into a unified semantic representation space. A hybrid representation architecture is adopted to enable UniPert tailors its encoders to accommodate the inherent differences between large and small molecules. By integrating contrastive learning with an extensive collection of experimental compound-target interactions, UniPert bridges genetic and chemical perturbagens, allowing agents involved in the same cellular pathway form closer semantic relationships in the representation space.

The proposed framework, along with its ability to generate a unified representation of distinct perturbagens, demonstrates significant utility in the realms of mechanism elucidation and drug discovery. The generated representations exhibit an exceptional ability to distinguish hierarchical pharmacological classes of well-annotated genetic and chemical perturbagens (e.g., enzymes, GPCRs, protein families, and drug MOAs), while facilitates the class annotations of understudied targets and compounds[43]. UniPert, as a pre-trained perturbagen encoder, could be plugged in state-of-the-art perturbation prediction frameworks to enhance performance in both genetic and chemical perturbation outcome prediction tasks, particularly in out-of-distribution scenarios.

Most importantly, UniPert overcomes critical bottlenecks in cross-domain perturbation effect modeling that were inaccessible previously. With the innovative unified representation ability, UniPert drives a novel *genetic-to-chemical* perturbation outcome transfer learning paradigm. By pre-training on contextualized genetic perturbation data and fine-tuning on limited chemical perturbation data, this paradigm fills the gaps between pooled genetic screens and arrayed chemical screens in terms of screen library size, experimental technology development, and data production efficiency. This paradigm was validated through proof-of-concept experiments on transcriptome-based dual-domain perturbation profiles, reducing chemical perturbation costs by over 60% and boosting *in silico* context-specific drug screening efficiency. The advancement offers new avenues for cost-effective, data-driven personalized target identification and drug screening, thereby accelerating progress in precision medicine.

Results

Overview of UniPert model

The goal of UniPert is to vectorize diverse perturbagens from their original sequences and unify them into an interpretable low-dimensional embeddings space, aggregating i) genetic perturbagens with similar (or close) functional roles, ii) chemical perturbagens with the same MOA, and iii) genetic and chemical perturbagens involved in the same biological pathway (Fig.1a). To achieve this goal, a unique hybrid design is implemented in the UniPert framework across multiple ways: multiscale molecular inputs, adapted model architectures, and combined training strategies (Fig.1b).

For chemical perturbagens, UniPert accepts SMILES as inputs, enabling the encoding of any small molecule within the vast chemical space, thereby ensuring the model's scalability. The inputted compound is initially embedded by binary topological extended-connectivity fingerprints (ECFPs)[44] for substructure characterization. Subsequently, the sparse embedding is converted into a dense embedding for refining the essential functional groups and enabling alignment with genetic perturbagen embeddings for downstream applications.

For genetic perturbagens, considering that proteins are the fundamental units of cellular function and the common key targets impacted by multiple perturbagen modalities, UniPert accepts amino acid sequences of proteins coded by the genes of interest as inputs. Unlike small molecules, mapping protein sequences into low-dimensional embeddings that preserve functional similarities is challenging due to their longer length and intrinsic evolutionary conservation. Large protein language models (PLMs), such as ESM[41], enable the generation of informative representations from primary sequences. However, directly compressing these high-dimensional representations—numerical matrices with sequence length multiplied by residue-level feature dimension—into global low-dimensional vectors risks losing critical functional characteristics in local sequence regions. To address this challenge, UniPert combines a traditional sequence alignment method with an advanced pre-trained PLM, leveraging graph neural networks (GNNs)[45] to capture both conserved sequence motifs and broader contextual features. Specifically, for a given protein, UniPert first compares it against a genome-wide protein reference database by using a Smith-Waterman-based local alignment algorithm[46], resulting in multiple paired sequence with varying similarity scores. Then, UniPert generates global sequence embeddings using ESM[41]. These similarity scores and embeddings form the edges and initial node embeddings of a weighted graph, respectively. Last, the node embeddings are updated and projected through a GNN-based message-passing mechanism that integrates functional information from similar sequences. In this way, UniPert better identifies conserved sequence regions and enhances the capacity to capture protein functional similarities.

Given these inputs from multiple perturbation modalities, UniPert leverages a dual optimization strategy to establish relationships in the shared latent embedding space. On one hand, to complement the sequence alignment algorithm and the PLM-based representations in learning robust protein embeddings, UniPert employs a graph self-supervised learning approach[47]. This method encourages consistency in protein node embeddings across different graph variants generated by randomly dropping edges and node features (see more details in Methods). On the other hand, UniPert maps compounds and proteins in the embedding space by integrating a contrastive learning-based alignment module[48] with a large collection of compound-target interactions, including 81,397 pairs involving 4,250 proteins and 8,533 compounds (see more details in Methods). This alignment allows UniPert to calibrate the distances between paired compounds and targets, bringing them closer to reflect their functional roles in shared biological pathways.

Together, these components establish UniPert as a versatile representation framework for multi-scale molecules, bridging genetic and chemical perturbagens within a unified, interpretable, and scalable embedding space. The framework facilitates various downstream applications, including pharmacology-based molecule classification, genetic and chemical perturbation outcome prediction, as well as large-scale cross-domain perturbation analysis and modeling (Fig.1c).

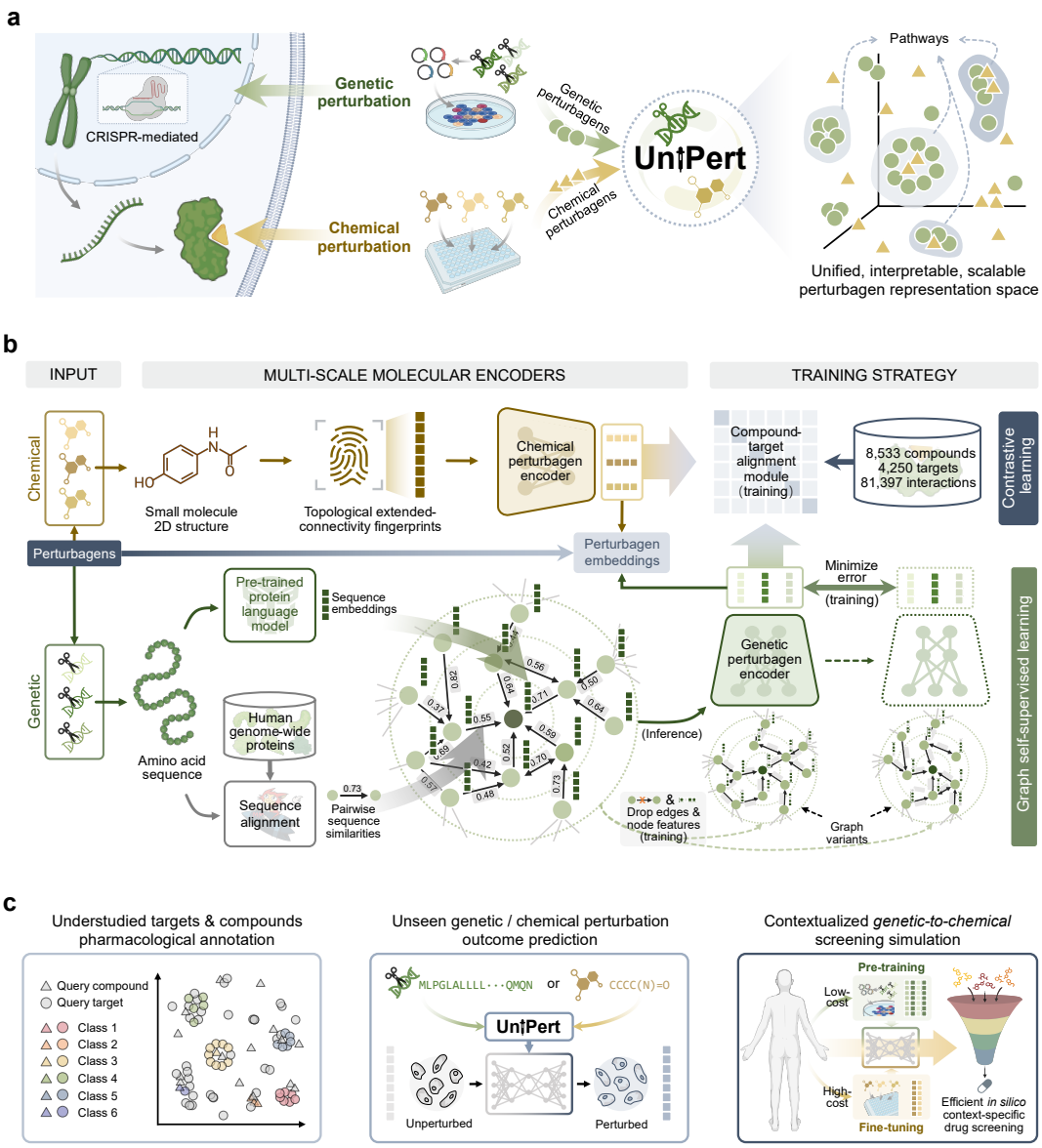


Fig. 1 | Overview of UniPert.

a, Problem formulation of UniPert framework: given genetic perturbagens (green) and chemical perturbagens (yellow), UniPert represents these large/small-scale molecules within a unified, biologically interpretable, and scalable representation space (right).

b, Hybrid design of UniPert framework: multi-scale molecular inputs (left), including amino acid sequences for genetic perturbagens and SMILES strings for chemical perturbagens, are processed through tailored model architectures (middle) trained by combined training strategies (right). For genetic perturbagens, the framework combines sequence alignment algorithm, large PLM, and graph neural networks to preserve sequence features. For chemical perturbagens, binary ECFP fingerprints are adopted for capturing functional groups. Training involves graph self-supervised learning (bottom right) to refine genetic perturbagen embeddings and contrastive learning (top right) to align perturbagen modalities using experimentally validated compound-target interaction data.

c, Downstream applications of UniPert. The representation interpretability supports pharmacological annotation for understudied targets and compounds, including families, interactions, regulatory relationships, and MOA (left). Its scalability positions UniPert as an efficient tool for enhancing genetic and chemical perturbation outcome predictions, especially for unseen perturbagens (middle). By unifying genetic and chemical perturbagens, UniPert advances large-scale cross-domain perturbation analysis and modeling, driving a *genetic-to-chemical* perturbation outcome transfer learning paradigm and accelerating *in silico* context-specific drug screening (right).

UniPert Distinguishes Genetic and Chemical Perturbagen Pharmacological Classes

To evaluate the biological semantic interpretability of the perturbagen embeddings generated by UniPert, we performed visual clustering analysis and semantic similarity analysis on 4,417 human protein targets with expert-curated pharmacological classification information and 2,812 small molecule drugs with known MOAs, respectively.

We examined hierarchical clustering for UniPert protein embeddings spanning multiple levels of pharmacological classes, which ranges from broad categories such as target class to more specific subclasses such as functional families and binding domains, reflecting the increasing granularity in biological and pharmacological relationships. Compared to ESM, UniPert effectively clusters targets across multi-level pharmacological classes and subclasses, achieving a 78.1% and 48.9% increase in Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI), respectively (Fig. 2a). By conducting in-depth embedding visualization and clustering comparisons across 5 major target classes (enzyme, membrane receptor, epigenetic regulator, transporter, and ion channel) and their subclasses (Fig. 2b-f and Extended Data Fig. 1), we demonstrate that UniPert learns interpretable embedding space that spatially enriches proteins with pharmacological relationships at a fine-grained scale. Further investigation into the embedding space of minority and unclassified targets reveals UniPert's ability in annotating meaningful protein relationships and protein family groups (Extended Data Fig. 2), which is significant for annotations of understudied proteins. For example, UniPert captures the regulatory relationship between FOMX1 and KIF proteins and interactions between subunits of the asymmetric PD-1/PD-L1 complex, and successfully clusters various protein families, including zinc finger proteins, the kinesin superfamily, tubulin proteins, etc. Additionally, UniPert presents significant advantages in organizing stratified protein family groups (superfamily, family, and subfamily) over other advanced PLMs such as ESM[41] and OntoProtein[49] (Extended Data Fig. 3).

In the semantic similarity analysis of drug MOAs, we quantitatively evaluated the differences in drug embedding similarities between intra-MOA and overall groups. UniPert demonstrates superior performance compared to the molecular ECFP approach[44], exhibiting improved intra-class semantic similarity in 13 out of 18 MOA categories (Fig. 2g). UniPert effectively clusters compounds with shared MOAs and preserves clear semantic boundaries between distinct MOA classes, highlighting its capacity to capture subtle mechanistic differences among pharmacologically similar compounds. Notably, we observed that drugs within two MOA classes—DNA inhibitor and $GABA_A$ receptor positive allosteric modulator—exhibited low similarity when measured with traditional molecular fingerprints, underscoring the limitations of traditional similarity-based virtual screening method for these MOAs. In contrast, UniPert closes their semantic distance in the representation space, thereby enhancing the potential for

discovering novel drugs within these challenging MOA categories.

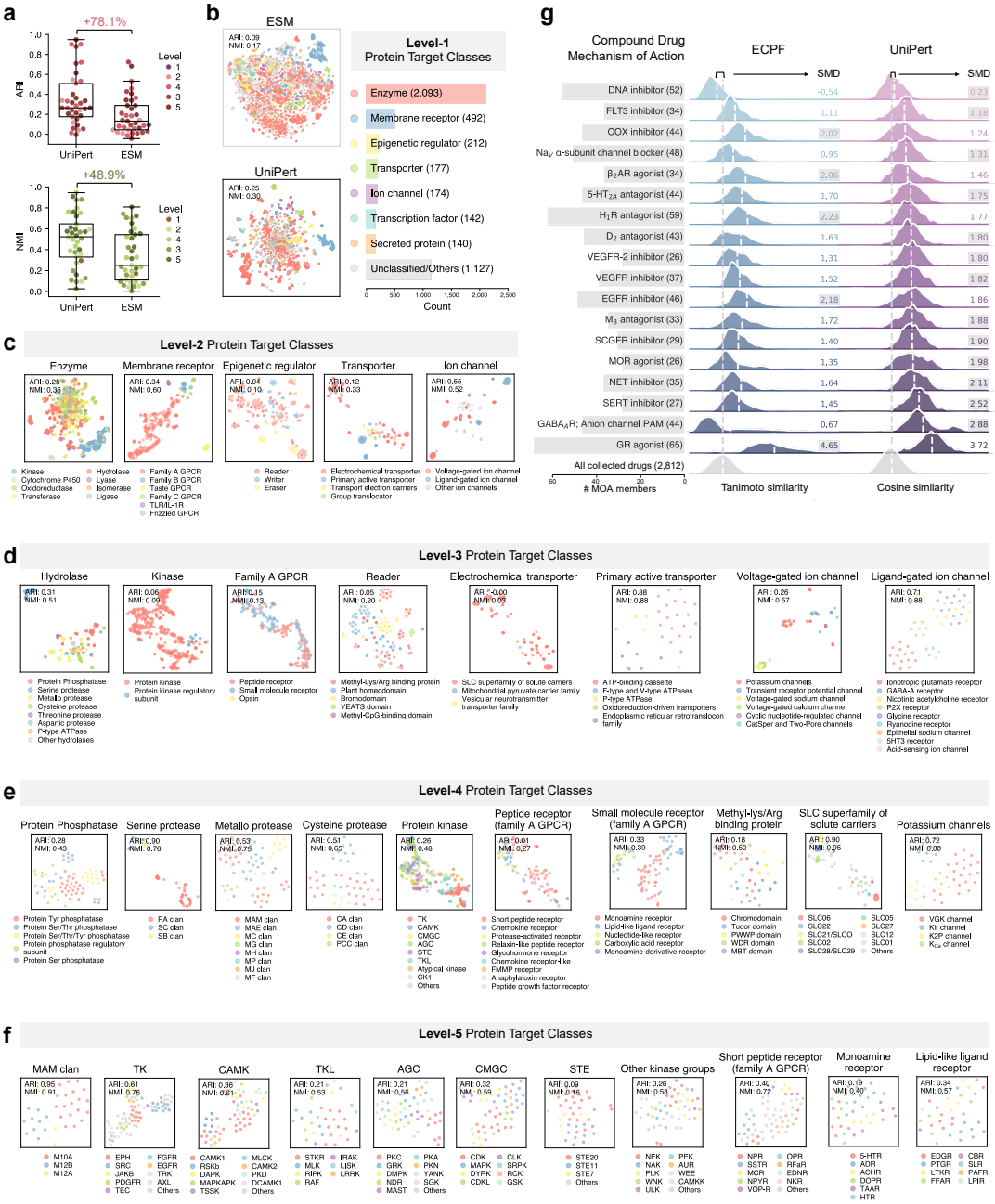


Fig. 2 | UniPert distinguishes hierarchical pharmacological target classes and converges compound members of MOA classes.

a, Overall performance comparison of UniPert and ESM in the visual cluster analysis across hierarchical pharmacological target classes, spanning five levels from the top to the bottom of the hierarchy. Each data point represents the adjusted Rand index (ARI) or normalized mutual information (NMI) score for a specific category across the 5 levels to measure clustering performance of UniPert and ESM embeddings against true protein classes using the K-means method.

b, T-SNE visualization comparison (left) and category distribution (right) of protein embeddings (n=4,417) at level 1 pharmacological class by UniPert (black-framed panel) and ESM (gray-framed panel). The top 7 categories are highlighted in color, and visualizations of the remaining minority and unclassified proteins are detailed in Extended Data Fig. 2.

c-f, T-SNE visualizations of UniPert protein embedding space for multiple pharmacological classes from level 2 to 5. Each panel represents to a specific target class at the respective pharmacological class level. Colors of scatters (representing proteins) correspond to distinct subclasses (bottom) within each parent class (top). ARI and NMI scores are calculated and displayed in the top-left corner to measure clustering performance. Detailed t-SNE plots for individual categories and the specific performance metrics comparisons with ESM are provided in Extended Data Fig. 1.

g, Performance comparison of UniPert and ECFP in the semantic similarity analysis across compound drug MOA classes. Each row represents the pairwise drug embedding similarity distribution (cosine similarity for UniPert and Tanimoto similarity for ECFP, as describe in the Methods) for all drugs within a MOA class with over 25 members. The standardized mean difference metric, SMD, was employed to quantify the magnitude of separation between MOA class members and all collected drugs (n=2,812). Higher scores indicate better semantic separation performance of diverse MOA drugs and are marked with a grey background.

UniPert Enhances Outcome Prediction for Unseen Genetic Perturbations

Building on UniPert's ability to distinguish pharmacological classes for both genetic and chemical perturbagens, we next explore its potential in enhancing prediction performance for perturbation outcomes. We examined the predictive ability of UniPert for unseen single and combination genetic perturbation outcomes within the GEARS-derived framework[39]. GEARS is a state-of-the-art genetic perturbation simulator capable of predicting outcomes of genetic perturbagens that have been not experimentally investigated. It learns functional perturbagen embeddings by leveraging the Gene Ontology graph and GNN architecture (Fig. 3a). However, reliance on prior knowledge limits its scalability and robustness, as the biological significance of learned embeddings depend heavily on the quality of the underlying annotations. Considering these limitations, we extended GEARS framework by adopting various pre-trained/pre-defined representation methods, including PseAAC[50], ESM[41], OntoProtein[49] and UniPert, to evaluate their impact on perturbation outcome prediction (Fig. 3a).

We first conducted experiments in the single-gene perturbation scenario using the *Dixit et al.* dataset. Compared to alternative methods, UniPert presents enhanced performance on the independent test perturbagens, as evidenced across multiple evaluation metrics (Fig. 3b). In the case of predicting the outcome of individual perturbagen *TOR1AIP1*, UniPert not only accurately predicted the direction of post-perturbation expression for the top 20 differentially expressed genes but also achieved higher accuracy than GEARS in estimating the magnitude of these changes (Fig. 3c).

Next, we expanded the comparison to the more complex compositional perturbation scenario using the *Norman et al.* dataset, which includes three test conditions: i) both genes in the combination occurred during training, ii) one gene occurred, and iii) neither gene was encountered during training. Across these conditions, UniPert outperformed other models in the latter two, especially when both genes were unseen during training (Fig. 3d). Cases of gene-level post-perturbation expression analysis also highlight UniPert's superior performance in combination genetic perturbation. For example, when simulating the response after perturbing of a combination of two unseen genes (*CBL* with *CNN1*), UniPert's predictions are more closely aligned with the observed ground truth (Fig. 3e). Moreover, UniPert consistently outperformed in the case of their individual perturbagens, with another example (*FOSB* with *CEBPB*) showing similar improvements (Extended Data Fig. 4). These improvements underscore UniPert's ability to generalize effectively to out-of-domain data, making it an effective tool for simulating perturbational screens involving unseen single and combination genetic perturbagens.

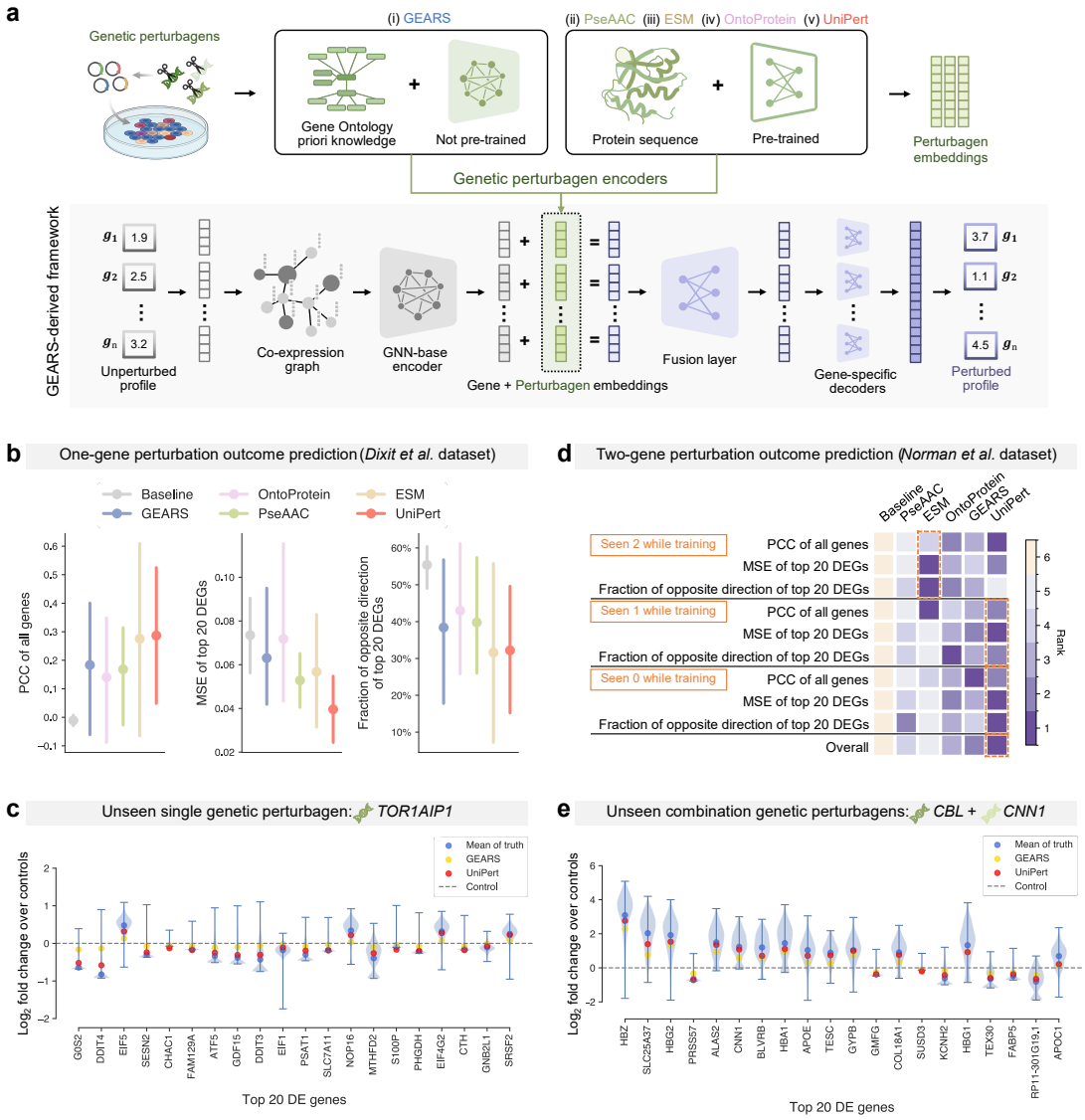


Fig. 3 | UniPert enhances unseen genetic perturbations outcome prediction in the GEARS-derived framework.

a, Schematic of the GEARS-derived genetic perturbations outcome prediction framework. Given unperturbed profiles (bottom left), i.e., the control group, each gene in the profile is encoded through graph neural networks (GNNs) with a gene co-expression graph. Embeddings of the genetic perturbagens (top left) can be learned by incorporating Gene Ontology prior knowledge and GNNs, as in i) the original GEARS model, or extracted from extensible pre-trained/pre-defined protein sequence representation methods, such as ii) PseAAC, iii) ESM, iv) OntoProtein, and v) our proposed UniPert. Single/multiple perturbagen embeddings (green) are then added to each gene embedding (grey) to obtain post-perturbation gene embeddings (purple) that will be finally converted into predicted post-perturbation gene expression values after sequentially passing through the fusion and gene-specific decoders.

b, Performance comparison for the one-gene perturbation outcome prediction. UniPert and alternative methods are evaluated across three metrics, with error bars showing 95% confidence

intervals.

c, Mean predictions for the top 20 DEGs after perturbations of *TOR1AIP1* (unseen single) with the original GEARS and UniPert-enhanced prediction frameworks, compared to ground truths.

d, Performance comparison across test scenarios in two-gene perturbation predictions. UniPert and alternative methods are tested in three scenarios where 0, 1, or 2 genes were observed during training in two-gene pairs.

e, Mean predictions for the top 20 DEGs after perturbations of *CBL* + *CNN1* (unseen pair) with the original GEARS and UniPert-enhanced prediction frameworks, compared to ground truths.

In b) and d), the baseline represents metric calculation of real unperturbed profiles versus real perturbated profiles, demonstrating the actual phenotypic differences between the control and treatment groups of the test perturbagens.

In c) and e), blue areas show ground truth distributions with blue dots as their mean, while red and yellow marks indicate mean predictions from UniPert and GEARS. The grey dotted line represents the mean unperturbed control expression.

UniPert Improves Performance in Chemical Perturbation Outcome Prediction

We further evaluated the ability of UniPert in predicting chemical perturbation outcomes within a CPA-derived framework[40]. CPA is an autoencoder-based model with chemical perturbation outcome prediction capabilities for combinatorial drugs, as well as multiple dosages, cell types, and time points (Fig. 4a). Its extension, chemCPA, implements the prediction for unseen drugs by incorporating an external molecular fingerprint encoder. This encoder can be replaced with alternative advanced molecule representation models such as Uni-Mol[51], KPGT[42], and our proposed UniPert (Fig. 4a), which is expected to improve the performance of perturbation response predictions.

To evaluate the effectiveness of UniPert in the CPA-derived framework, we compared its performance against chemCPA using the massive drug screening dataset sciPlex3 (Fig. 4b). The dataset was processed using a pathway-based hierarchical splitting, with 20% of the drugs randomly selected from pathways containing over 10 members as unseen drugs for independent testing, while the rest for training. The results show that UniPert outperformed chemCPA in terms of Pearson correlation coefficient (PCC) and mean squared error (MSE) at multiple top DEG scales across different cellular contexts (Fig. 4c and Extended Data Fig. 5).

Beyond the overall performance comparison on large-scale data, we also explored UniPert's ability in leave-one-drug-out experiments using the HDAC inhibitor dataset. Drawing from existing evidences[52,53] suggesting that drugs targeting epigenetic regulators results in very similar perturbation responses, we extracted post-perturbation transcriptional profiles of 17 histone deacetylase (HDAC) inhibitors from the sciPlex3 dataset, and projected them into a lower-dimensional space for visualization. We observed that, under the condition of same concentration (1μM) and same time point (24h), cells treated with these HDAC inhibitors exhibited similar response trajectories across three cell lines, with Panobinostat eliciting the strongest response (Fig. 4d). Furthermore, HDAC inhibitors with significant responses tend to be non-selective that can inhibit multiple classes of HDAC, such as Panobinostat, Quisinostat, and Abexinostat. In contrast, less potent inhibitors displayed higher selectivity for specific HDAC targets, such as Tacedinaline.

Subsequent analysis showed that UniPert embeddings effectively captured the pharmacological distinctions and pairwise similarities of HDAC inhibitors, outperforming extended-connectivity fingerprints (Fig. 4e, f). In leave-one-drug-out experiments of the top 9 most responsive HDAC inhibitors, UniPert demonstrated superior performance in handling unseen compounds compared to other molecular representation methods (Fig. 4g). Structural comparison of these inhibitors with Panobinostat highlighted key similarities in the zinc-binding group (ZBG) and cap regions, which aligns with known pharmacological

characteristics[54], further validating UniPert's ability to capture meaningful molecular features.

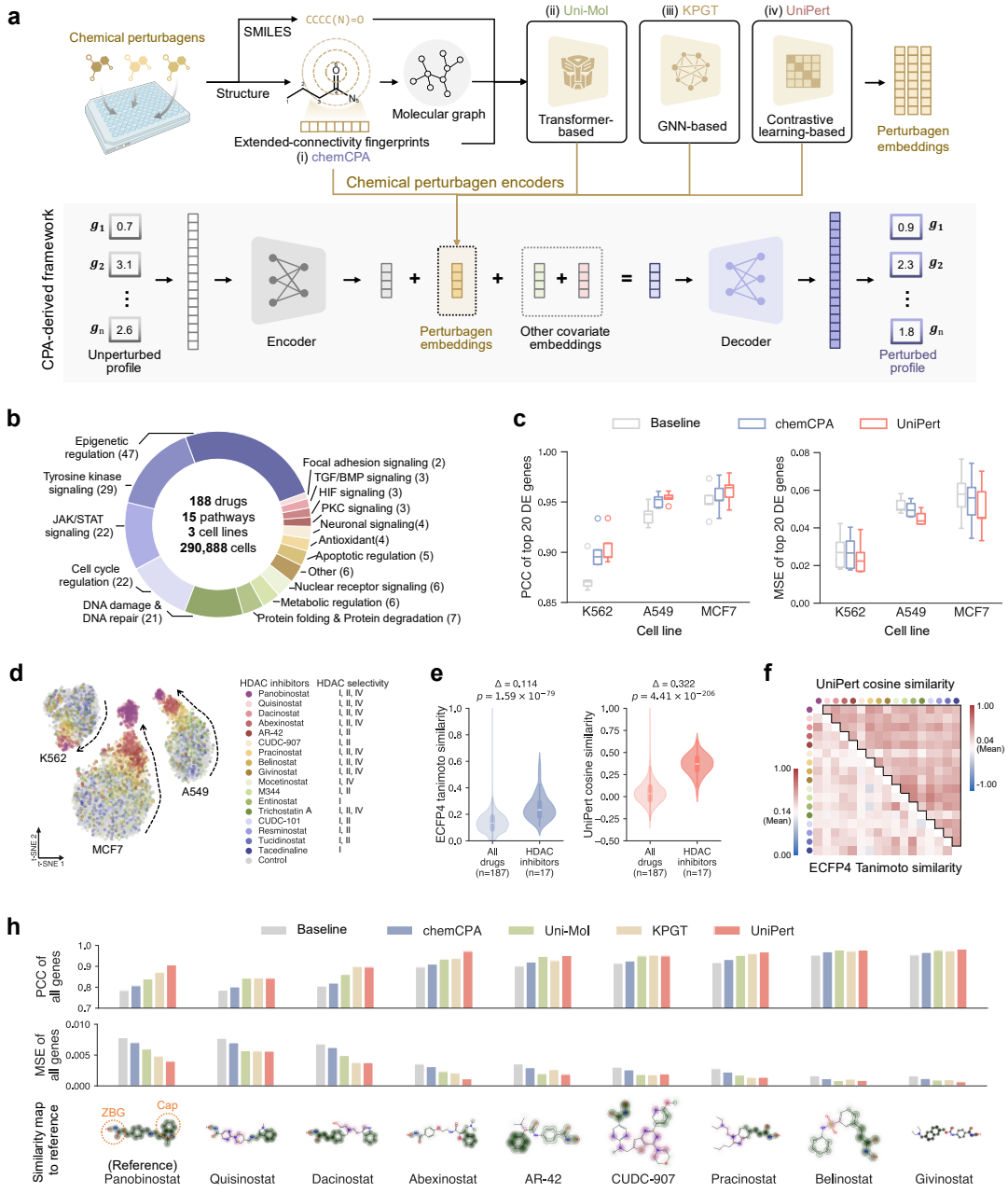


Fig. 4 | UniPert improves chemical perturbation outcome prediction performance in the CPA-derived framework.

a, Schematic of the CPA-derived chemical perturbations outcome prediction framework. Unperturbed profile (bottom left) vectors are projected into a low-dimensional latent space, while chemical perturbagens (top left), i.e., small molecules, are numericized using traditional molecular fingerprint features like i) chemCPA model, or, be encoded with advanced representation methods such as ii) Uni-Mol, iii) KPGT, and our proposed iv) UniPert. Perturbagen embeddings (yellow) along with other covariate embeddings are then added to the

latent control embedding (grey), producing predicted post-perturbation gene profile vectors (purple) after decoding.

b, Donut plot showing the distribution of drug pathways in the extensive drug screen dataset (sciPlex3), with statistics displayed in the center.

c, Performance comparison of UniPert and chemCPA in predicting drug perturbation outcomes. Box plots display the PCC and MSE for the top 20 differentially expressed genes (DEGs) across the sciPlex3 dataset, spanning three cell lines. Evaluation results are shown for different experimental trials from five data splits, with error bars representing 95% confidence intervals. Results for multiple DEG scales can be found in Extended Data Fig. 5.

d, T-SNE visualization highlighting the response trajectory over HDAC inhibitor and vehicle-treated cells. Colors correspond to the 17 HDAC inhibitors. The black-dotted trajectory starts from the control group (vehicle-treated) and points towards cells exhibiting stronger responses under 1μM drug treatment.

e, Comparative distribution of embedding similarity scores between all drugs and HDAC inhibitors. Violin plots show the distribution of ECFP4 Tanimoto similarity (left) and UniPert cosine similarity (right) among 'All drugs' (n=187) and 'HDAC inhibitors' (n=17) groups. Mean differences (Δ) and statistical significance (p-values) annotated above each plot were calculated using the independent samples *t*-test.

f, Heatmap showing pairwise HDAC inhibitor embedding similarity by UniPert and ECFP4.

g, Performance comparison in leave-one-drug-out perturbation outcome prediction. Top and middle panels show the PCC and MSE metrics results for UniPert and competitive methods. The bottom panel exhibits 2D molecular structures and structural similarities between Panobinostat (reference) and other eight HDAC inhibitors, which highlights the ZBG (zinc-binding group) and Cap regions.

In c) and h), the baseline refers to the metric calculation between control profiles and real perturbed profiles, highlighting the actual phenotypic differences observed between control and treatment groups for the test perturbagens.

Establish the *Genetic-to-chemical* Perturbation Outcome Transfer Learning Paradigm with UniPert

Extensive comparative experiments not only demonstrate the versatility and efficiency of UniPert in improving outcome predictions for both genetic and chemical perturbations but also underscore its potential to support reliable cross-domain perturbation modeling in advancing data-driven precision medicine. The conceptual process of personalized drug discovery typically involves identifying patient-specific targets followed by iterative, target-oriented drug screening[55,56] (Fig. 5a). However, the vast chemical space (>10^{63}) renders drug screening libraries unaffordable, even with the assistance of virtual screening, which narrows it to a gigascale yet still comprise countless compounds, making these processes become time-consuming and costly[57]. Additionally, unlike genetic screening technologies such as CRISPR-mediated approaches[58], which have achieved significant scalability through pooled experiments that targets thousands of genes within a single assay[59–61], chemical screening still requires array-based libraries, where each compound must be treated individually. This technological gap further results in slower generation and less availability of high-resolution chemical perturbation data compared to genetic data. For example, the latest scPerturb database[62] compiles 58 single-cell level perturbation datasets, of which only 10 pertain to chemical perturbation.

Considering these limitations and leveraging UniPert's capability to unify perturbagen representations, we introduce a UniPert-driven *genetic-to-chemical* perturbation outcome transfer learning paradigm (Fig. 5b). This paradigm begins a perturbation outcome prediction model with pre-training on low-cost genetic perturbation data, enabling it to systematically learn the genetic characteristics of the cellular environment and response patterns to perturbagens targeting distinct pathways. Then, fine-tuning with high-cost chemical perturbation data further adapts the model to the specific dynamics of chemical perturbation responses. The genetic-guided perturbation outcome prediction framework is expected to accelerate the efficient and cost-effective simulations for massively contextual chemical perturbations, thereby addressing key challenges in scalability and resource constraints in chemical screening efforts.

To validate the efficacy of the paradigm, we conducted a comprehensive investigation, collecting and integrating up-to-date available large-scale dual-domain perturbation datasets, including image-based CPJIMP1[38] and transcriptome-based LINCS[26] datasets. Specifically, the processed CPJIMP1 datasets contains well-level morphological profiles (838-dimensional features) extracted using the image processing software, involving 160 CRISPR-based genetic and 302 chemical perturbagens across two cell lines. The processed LINCS datasets includes gene signature profiles (978 landmark transcripts) from thousands of CRISPR-knockout-based

genetic and chemical screenings across five different cell lines (Fig. 5c). Based on these multi-omics and multi-domain perturbation data, we conducted extensive experiments to comprehensively evaluate the proposed paradigm's predictive power by holding out various proportions (20%, 40%, 60%, and 80%) of chemical data for testing (Fig. 5d).

For image-based perturbation datasets, experimental results show that both pre-trained and non-pre-trained settings yield less satisfactory performances in two cell lines (Fig. 6f). Although genetic pre-training enhanced prediction accuracy and robustness certain trials, these improvements are not substantial. This limited performance can be attributed to several factors: i) the limited number of perturbagens in both domains results in insufficient training data for capturing useful patterns, ii) the suboptimal morphological profile extraction tool fails to adequately represent post-perturbation states, and iii) insufficient quality control may introduce significant noise into the data (Extended Data Fig. 6). Similar concerns regarding these limitations have been raised in recent studies[29,63].

In contrast, for transcriptome-based perturbation datasets with a significantly larger number of perturbagens and more refined data preprocessing workflows compared to image-based profiles, the proposed paradigm yielded more insightful results (Fig. 5f). Specifically, when modeling chemical perturbation outcome prediction without genetic pre-training, the PCC improved progressively as the proportion of training data increased. However, after pre-training on genetic data, the model consistently delivered significantly more robust and accurate predictions across multiple cell lines. Remarkably, even with only 20% of the data used for fine-tuning, the pre-trained models outperformed the non-pre-trained models trained on 80% of chemical data, which directly reduced the cost required for chemical perturbation simulation by up to 60%. Moreover, we noticed that cell line datasets with fewer chemical perturbagens (<5000), such as A375 and HT29, struggled to achieve desirable performance even with 80% data training, while genetic pre-training compensated for this deficiency. Additionally, we observed that unlike PCC, which exhibited variability across experimental settings, the MSEs between the real post-perturbation outcomes of the same perturbagen were consistently higher than those between predicted perturbation outcomes of the same perturbagen or between its real post-perturbation outcomes versus the control. This observation suggests that, in profiles generated from real perturbation experiments, the gene expression changes induced by the same perturbagen exhibit greater consistency and more predictability in their directional trends than in the magnitude of the changes.

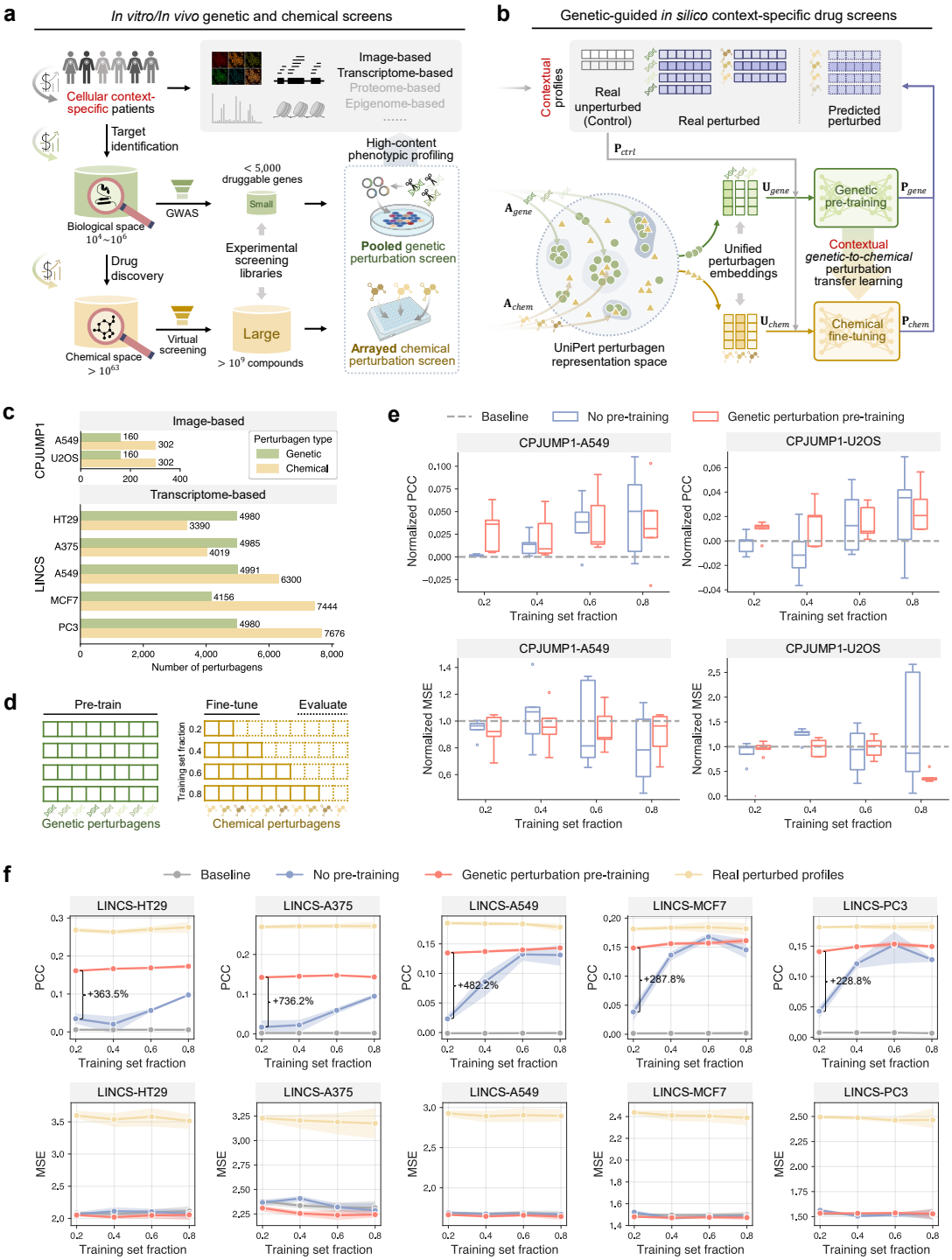


Fig. 5 | Establish the *genetic-to-chemical* perturbation outcome paradigm with UniPert and its application to morphology and transcriptomic profiles.

a, Schematic of the envisioned personalized drug discovery process involving both *in vitro/in vivo* genetic and chemical perturbations. This process is resource-intensive due to the heterogeneity of cellular contexts, the vast size of chemical libraries, and the cumbersomeness and costliness of large-scale arrayed chemical screens.

b, Overview of UniPert-driven *genetic-to-chemical* perturbation outcome transfer learning

paradigm.

c, Summary of integrated multi-domain perturbation datasets employed in the paradigm. Bar plots indicate the number of genetic (green) and chemical (yellow) perturbagens across diverse cellular contexts in CPJUMP1 (image-based profiles) and LINCS (transcriptome-based profiles) datasets.

d, Illustration of data splitting for the genetic-guided chemical perturbation outcome prediction task. All genetic perturbation data is used for pre-training, followed by fine-tuning and evaluation on chemical perturbation data at varying proportions of total chemical perturbagen numbers.

e, Performance comparison on image-based datasets. Box plots show normalized metrics calculated between predicted profiles and ground truth under genetic-pre-training and non-pre-training settings. The baseline (grey dotted lines) reflects normalized control group performance.

f, Performance comparison on transcriptome-based datasets. Point plots illustrate trends in performance as the chemical training set size increases for models with (red) and without (blue) genetic pre-training. The baseline (grey) represents comparisons between control and real perturbed profiles without a prediction model, serving as a reference for evaluating the phenotypic differences between the control and predicted perturbed profiles. Points represent means and error bars represent 95% confidence intervals of five experimental trials with different data split seeds.

Joint analysis of UniPert-derived simulated perturbation *cause-and-effect* spaces

After validating the proposed paradigm on the LINCS datasets as a proof-of-concept, we further extended its application to the joint analysis of perturbagen representation space and simulated cell-line-specific perturbation phenotype spaces generated by UniPert. This extension is particularly significant in the context of precision medicine, as it addresses cellular heterogeneity by modeling perturbation responses across diverse cell lines.

To facilitate cross-evaluation and comparison of perturbation prediction models across multiple cell lines, we redefined the perturbagen sets from five datasets and retrained the models accordingly. Specifically, we referenced the original Connectivity Map (CMap) workflow[26], which defines well-annotated perturbagens as *Touchstone* (CMap-*Touchstone*) and unannotated compounds as *Discovery* (CMap-*Discovery*) to establish their query methodology. In alignment with our genetic-guided phenotype space simulation paradigm, we defined all genetic perturbagens (~5,000) in each dataset as *Guide* for pre-training context-specific perturbation outcome prediction models. We then defined a new set of 1,012 chemical perturbagens, termed *Touchstone*, which is overlapped with the CMap-*Touchstone*. The *Touchstone* was used for fine-tuning the models, with 452 involved in distinct pathways for training and the rest for early stopping. This process yielded five cell-line-specific perturbation prediction models under similar training conditions (Fig. 6a right).

We performed cross-cell-line validation with five trained models on compounds excluding *Touchstone* (term as *Discovery* with 2,378-6,664 compounds) in each cell line dataset (Fig. 6b). The PCC results demonstrated the expected specificity of the perturbation models across different cell lines. While for MSE, although each cell line's test set performed optimally in its corresponding model, the overall MSE differences across datasets were more pronounced, reinforcing our earlier conclusion that perturbation direction is more robust and predictable than perturbation magnitude.

Simultaneously, we constructed a non-context-specific, cross-domain perturbagen space based on UniPert representations for perturbagens in the *Guide*, *Touchstone*, and *Discovery* (Fig. 6a left). By integrating PharmGKB pathway evidence[64], including manually curated relationships between drugs and genes involved in pharmacokinetics (PK) and pharmacodynamics (PD), we annotated this space. T-SNE plots of perturbagens and contour density plots of pathway labels revealed that UniPert effectively clusters genes and drugs involved in the same pathway (Fig. 6c). Notably, certain pathways exhibited similar patterns, such as the selective serotonin reuptake inhibitor pathway and the sympathetic nerve pathway, as well as the Vemurafenib and Sorafenib pathways. Moreover, the pathway patterns of the same drug involved in PK and PD showed distinct contours, exemplified by Ibuprofen.

To evaluate the similarity between perturbagens in the UniPert space, we computed pairwise cosine similarity scores across all pairs, resulting in a perturbagen association matrix. Among over 82 million pairwise similarity scores, we observed distinct distribution patterns for genetic-genetic (gene-gene) and chemical-chemical (chem-chem) perturbagen pairs (Fig. 6d). The chem-chem pairs exhibited a broader and more normally distributed range of similarity scores, reflecting inherent structural diversity and molecular flexibility of chemical compounds. In contrast, the gene-gene pairs displayed a more restricted and non-normal distribution, which can be attributed to the evolutionary conservation and functional constraints of biological molecules. The gene-chem pairs closely mirrored that of the chem-chem pairs but with a slightly lower mean similarity, suggesting that UniPert effectively integrates and coherently represents different perturbagen types within a unified space. In the original CMap analysis[26], this perturbagen association matrix was manually curated through literature-based annotations of shared descriptors, such as MOA, Gene Ontology, and pathway membership. However, this process i) is cumbersome and time-consuming, ii) is binary and does not reflect the strength of relationships, and iii) is overly reliant on prior knowledge, failing to address understudied targets and unannotated compounds. UniPert's unified sequence-based representation framework overcomes these limitations effectively.

To validate whether the UniPert-based perturbagen similarity matrix can recover known compound-target relationships, we compared approximately 40 million gene-chem perturbagen pairs with 81,397 compound-target interactions used in the UniPert model training. We referred to these interaction pairs as true relationship pairs and the remaining pairs as null pairs. The receiver operating characteristic (ROC) analysis yielded an area under the curve (AUC) > 0.98, with an optimal cutoff of 0.127, indicating the efficiency and promising of using UniPert embeddings for similarity-based compound-target interaction inferences (Fig. 6e).

To assess the effectiveness of our simulated *cause-and-effect* perturbation spaces and gain deeper insights with them, we conducted joint analyses of two genes with their agonists (Fig. 6f), which were previously reported in the original CMap study[26]. For *ESR1* (Estrogen Receptor 1), a key mediator of estrogen signaling in tissues such as the breast, we observed that the average similarity of *ESR1* and its agonists in the cause space representation was 0.23, exceeding the optimal cutoff. This indicates that UniPert effectively captures the relationships between *ESR1* and its agonists within the unified perturbagen space. Furthermore, the connectivity of *ESR1* and its agonists varied across different cell-line-specific cause spaces, with the strongest connectivity observed in HT29 and MCF7 cell lines. Notably, in MCF7 (a breast cancer cell line with high *ESR1* expression), the connectivity pattern was most consistent with the cause similarity, indicating a highly coherent perturbation effect. Another case of *PPARG* (Estrogen Receptor 1) and its agonists, which exhibited the highest phenotype

connectivity in HT29 and PC3 cell lines, was also consistent with the findings reported in the original CMap study.

By leveraging these joint analyses, we can reveal how perturbation effects vary across different cell lines within the simulated *cause-and-effect* perturbation spaces constructed with UniPert, providing a critical insight into the complex nature of biological responses. This context-dependent and cross-domain connectivity highlights the utility of UniPert in precision medicine, where the specific cellular environment can greatly influence the response to genetic or chemical perturbations. Incorporating such dynamic, context-aware perturbation representations into therapeutic strategies allows for the identification of more personalized and accurate interventions.

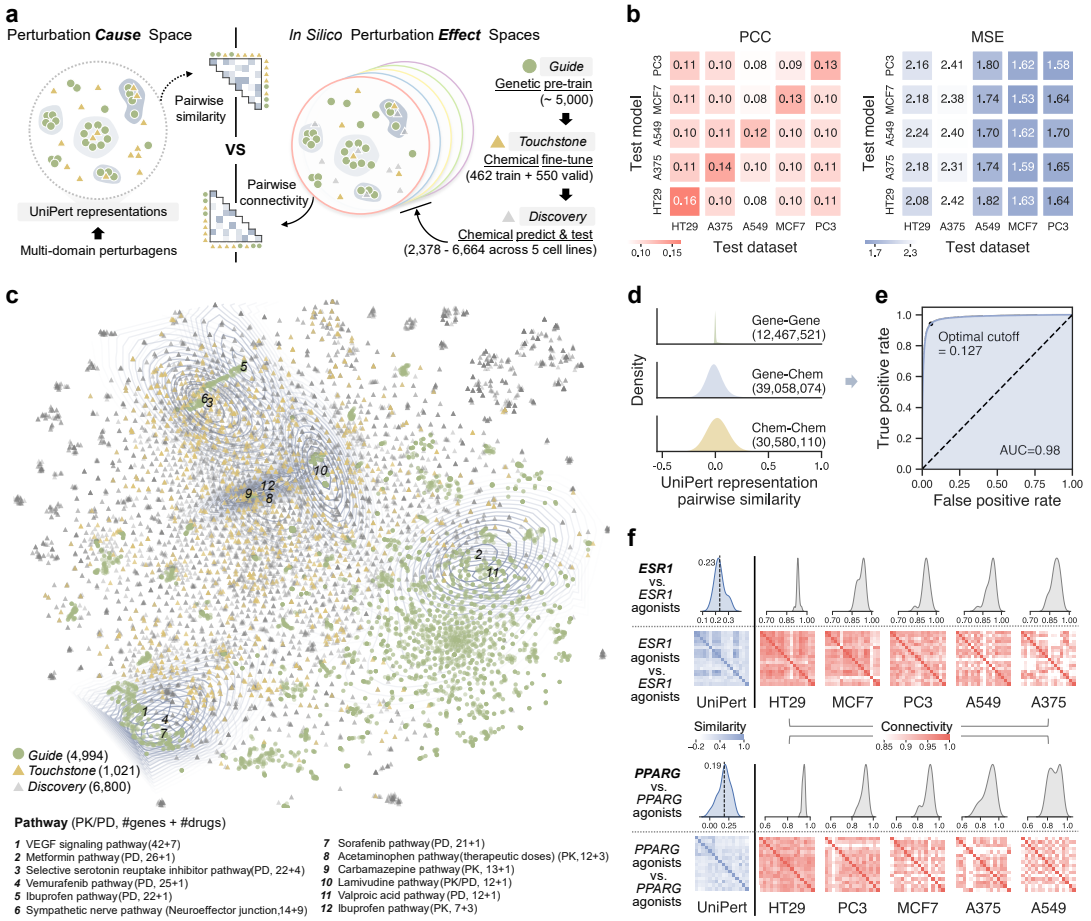


Fig. 6 | Joint analysis of cause space (perturbagens) and *in silico* context-specific effect spaces (perturbation phenotypic states) derived using UniPert.

a, Perturbation cause space (left) and *in silico* context-specific effect spaces (right) generated using UniPert and UniPert-driven *genetic-to-chemical* perturbation outcome transfer learning paradigm. The perturbation cause space is composed of multi-scale perturbagen entities and formed through UniPert's unified encoding framework. The effect spaces are simulated by

trained perturbation outcome prediction models within the proposed paradigm, capturing context-specific phenotypic responses.

b, Cross-cell-line evaluation of chemical perturbation outcome prediction models trained following the *genetic-to-chemical* perturbation outcome transfer learning paradigm. Each cell line model was pre-trained with all available genetic perturbation data and fine-tuned with equal amount of chemical perturbation data. The resulting five cell line-specific models were evaluated on the remaining chemical perturbagens of each cell line dataset.

c, T-SNE plot of UniPert representations with contour density plots of pathway labels collected from PharmGKB resource for the *Guide* (4,994 genetic perturbagens), *Touchstone* (1,021 chemical perturbagens), and *Discovery* (6,800 chemical perturbagens). PK refers to pharmacokinetics, and PD refers to pharmacodynamics.

d, Distribution of UniPert representation similarities of approximately 82 million perturbagen pairs grouping into gene-gene, gene-chem, chem-chem categories across 4,994 genetic and 7,821 chemical perturbagens.

e, ROC curve comparing UniPert representation pairwise cosine similarities with 81,397 experimentally validated compound-target interactions as positive references.

f, Perturbagen similarity and context-specific perturbation phenotype connectivity of *ESR1* and *PPARG* with their agonists. Distribution plots show the cross-domain gene-chem perturbagen similarity (blue) or phenotype connectivity across 5 cell lines (grey) for the indicated gene and its agonists. Thumbnail heatmaps show inter-domain chem-chem perturbagen similarity (blue) and phenotype connectivity (red) of agonists targeting the indicated gene. Top: *ESR1* / *ESR1* agonists (n=44). Bottom: *PPARG* / *PPARG* agonists (n=16).

Discussion

Unifying genetic and chemical perturbation representations is becoming increasingly crucial as perturbation analysis plays a central role in disease research and therapeutic development. Previous efforts primarily focused on representing and connecting perturbation phenotypes, overlooking the need for systematic representations of perturbation entities.

UniPert addresses this need by constructing a unified, interpretable, and scalable representation space that bridges genetic and chemical perturbagens. UniPert presents three key advantages: i) hybrid encoding architectures enable multi-scale perturbagen representation from their primary sequences. ii) functional and pharmacologically relevant perturbagen embeddings improve the prediction of genetic and chemical perturbation outcomes. iii) the novel contextualized *genetic-to-chemical* perturbation transfer learning paradigm accelerates large-scale *in silico* drug screening. Together, these advantages lay the foundation for further joint analysis and modelling with growing genetic and chemical perturbation data across both perturbagen (cause) and phenotype (effect) spaces.

To our knowledge, UniPert is the first method capable of representing diverse perturbagens from both genetic and chemical domains within a unified framework while generalizing effectively to unseen queries. This flexibility allows UniPert to be applied to pharmacological functional annotation and clustering for understudied proteins and poorly annotated drug-like compounds. In addition, when performing joint analysis with perturbation phenotype data across both domains, UniPert can eliminate the need for cumbersome manual annotation, streamlining the processing and broadens the scope for the investigations. Extensive experimental validation across various model frameworks, perturbation complexities, and data scales demonstrates UniPert's representation ability to improve perturbation outcome predictions. These strengths further contribute to the cross-domain perturbation modelling, establishing a cost-effective paradigm for large-scale, context-specific drug screening simulations by transferring learned perturbation effects from genetic to chemical domains. This paradigm highlights UniPert potential to advance personalized medicine, enabling the development of personalized treatment strategies by predicting individual responses to small molecular interventions based on the patient's unique genetic profiling[19]. While this paradigm has only been efficiently validated on transcriptome-based but not image-based perturbation profiles, it is important to note that image-based phenotypic data holds considerable potential for future model development as its more accessible, less expensive, higher content and throughput cellular profiling compared to transcriptome data[65–68]. Furthermore, the paradigm is poised to expand into other omics fields with the relevant perturbation data continues to grow[69], such as proteomics[70–72] and epigenomics[73,74].

Despite its strengths, one limitation of UniPert is that it does not fully incorporate genetic perturbagens at a granular level. While UniPert cleverly uses protein sequences as inputs for genetic perturbagens and associates them with chemical perturbagens based on protein-compound interaction knowledge, it does not account for other molecular layers underling genetic changes, such as DNA and RNA. Additionally, UniPert cannot encode non-coding sequences, which are important for gene regulation and alternative splicing. Future iterations of UniPert could integrate multimodal biological sequence language models, such as Evo[75], to provide a more comprehensive representation of genetic perturbagens.

UniPert holds significant potential for advancing perturbation studies in multiple dimensions. First, by integrating interpretable methodology[76] and performing more targeted in-depth analyses, we can further refine the context-specific *genetic-to-chemical* perturbation outcome transfer learning paradigm and uncover the relationships between individualized genetic factors and potential therapeutic drugs. This approach will enable the development of more robust, data-driven frameworks for personalized treatment prediction, facilitating the identification of patient-specific therapeutic strategies and accelerating the transition to personalized medicine. Second, with the successful application of single-cell language models[36,77] and the rapid growth of perturbation data driven by the Perturbation Cell Atlas initiative[18], UniPert can help to build a computational foundation model for multi-domain perturbation data. Third, UniPert holds promise as a foundational tool in the creation of AI-powered virtual cells[78], aiding in the identification of drug targets, prediction of cellular outcomes, and hypothesis testing at scale. We expect UniPert to play an important role in advancing the field of perturbation biology and contribute substantially to drug discovery and precision medicine.

Methods

The Methods describe in turn i) the model architecture, training strategy, and implementation details of UniPert, ii) the implementation for genetic and chemical perturbation outcome prediction tasks, iii) the implementation for *genetic-to-chemical* screening transfer learning, iv) all datasets for UniPert training and downstream applications, and v) all statistical analyses and evaluation metrics used.

UniPert

Model architecture. A schematic overview of UniPert is depicted in Fig. 1b. UniPert integrates chemical and genetic perturbagen encoder modules into a unified neural network framework for end-to-end representation learning and generalization to new perturbagens.

Chemical perturbagen encoder module. UniPert first characterizes genetic perturbagens using the topology-based ECFP algorithm[44]. Specifically, for a given small molecule $\mathbf{c}$, UniPert employs the RDKit tool[79] to calculate the molecular fingerprint, generating an initial K-dimensional binary vector $\mathbf{h_c} \coloneqq \mathcal{F}_{ECFP}(\mathbf{c}) \in \{0,1\}^K$. The vector is then passed through a fully connected layer and transformed into a D-dimensional dense vector: $\mathbf{u_c} = \mathcal{E}_{\theta_{chem}}(\mathbf{h_c})$, where θ_{chem} represents the learnable parameters of the chemical perturbagen encoder $\mathcal{E}_{\theta_{chem}}$, and $\mathbf{u_c} \in \mathbb{R}^D$ denotes the generated UniPert representation for the query $\mathbf{c}$.

Genetic perturbagen encoder module. UniPert represents genetic perturbagens through a combination of sequence alignment algorithm, pre-trained PLM, and GNN. Specifically, UniPert first constructs a reference graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$, where $\mathbf{V}$ denotes the protein node set of the human genome-wide proteins and their corresponding sequences (N = 19,187) downloaded from the HGNC[80] and UniProt[81] resources, and $\mathbf{E}$ is the weighted edge set representing pairwise Smith-Waterman-based[82] sequence alignment similarity scores between proteins, calculated using MMseqs2[83]. Given an unseen amino acid sequence $\mathbf{t}$, UniPert adds it as an new node to $\mathbf{G}$, resulting in a update graph $\mathbf{G_t} \coloneqq (\mathbf{V_t}, \mathbf{E_t})$, where $\mathbf{V_t} = \mathbf{V} \cup \{\mathbf{t}\}$ and $\mathbf{E_t} = \mathbf{E} \cup \mathbf{e_t}$, with $\mathbf{e_t}$ representing the new edges between $\mathbf{t}$ and the nodes in $\mathbf{V}$ based on their sequence similarities. To obtain and update the embedding for $\mathbf{t}$, UniPert employs the pre-trained ESM model (esm2_t33_650M_UR50D version)[41] to initialize the node embeddings for $\mathbf{V_t}$ with P-dimensional (P=1280) vectors $\mathbf{X_{V_t}} \in \mathbb{R}^{(N+1)\times P}$. Subsequently, a two-layer GNN is applied to perform message passing on $\mathbf{G_t}$, updating the node embeddings to $\mathbf{d}$-dimensional vectors, $\mathbf{U_{V_t}} \coloneqq \mathcal{E}_{\theta_{gene}}(\mathbf{X_{V_t}}, \mathbf{G_t}) \in \mathbb{R}^{(N+1)\times D}$,where θ_{gene} represents the learnable parameters of the genetic perturbagen encoder $\mathcal{E}_{\theta_{gene}}$, and $\mathbf{u_t} \in \mathbb{R}^{\mathrm{D}}$ extracted from $\mathbf{U_{V_t}}$ represents the generated UniPert representation for the query $\mathbf{t}$.

Training strategy. UniPert employs a dual optimization strategy to effectively train the model, including graph self-supervised learning for genetical perturbagen embedding enhancement and contrastive learning for genetical-chemical perturbagen alignment, performing the overall loss $\mathcal{L} = \mathcal{L}_{enhance} + \mathcal{L}_{align}$.

Graph self-supervised learning loss. Inspired by the successful applications of bootstrap mechanisms in representation learning within the fields of computational vision[84] and large-scale graphs[47], UniPert adopts a Bootstrapped Graph Latents approach[47] for efficient self-supervised representation learning of genetic perturbagens. Specifically, for the constructed reference graph $\mathbf{G} = (\mathbf{X}, \mathbf{A})$ with initialized ESM node features $\mathbf{X} \in \mathbb{R}^{N \times P}$ and similarity-based adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$, UniPert produces two augmented graph variants: $\mathbf{G_1} = (\widetilde{\mathbf{X}}_\mathbf{1}, \widetilde{\mathbf{A}}_\mathbf{1})$ and $\mathbf{G_2} = (\widetilde{\mathbf{X}}_\mathbf{2}, \widetilde{\mathbf{A}}_\mathbf{2})$, by applying stochastic masking to edges and node features. These variants are processed separately through an online encoder $\mathcal{E}_{\theta_{gene}}$ and an offline encoder $\mathcal{E}_{\varphi_{gene}}$, producing corresponding node representations $\widetilde{\mathbf{H}}_\mathbf{1} = \mathcal{E}_{\theta_{gene}}(\widetilde{\mathbf{X}}_\mathbf{1}, \widetilde{\mathbf{A}}_\mathbf{1})$ and $\widetilde{\mathbf{H}}_\mathbf{2} = \mathcal{E}_{\varphi_{gene}}(\widetilde{\mathbf{X}}_\mathbf{2}, \widetilde{\mathbf{A}}_\mathbf{2})$. The online representations are then fed into a node-level predictor $p_{\theta_{pred}}$, yielding $\widetilde{\mathbf{Z}}_\mathbf{1} = p_{\theta_{pred}}(\widetilde{\mathbf{H}}_\mathbf{1})$ aimed at predicting the offline representations $\widetilde{\mathrm{H}}_2$, with a cosine similarity-based loss $\mathcal{L}_{\mathrm{enhance}}(\theta, \varphi)$ defined as

$$\mathcal{L}_{enhance}(\theta, \varphi) = -\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\widetilde{\mathbf{Z}}_{1,i}{}^{T}\widetilde{\mathbf{H}}_{2,i}}{\|\widetilde{\mathbf{Z}}_{1,i}\| \cdot \|\widetilde{\mathbf{H}}_{2,i}\|}\right), \tag{1}$$

where N is the number of graph nodes, and $\widetilde{\mathbf{Z}}_{1,i} \in \mathbb{R}^D$ and $\widetilde{\mathbf{H}}_{2,i} \in \mathbb{R}^D$ represent the embeddings of the i-th node from the online predictor $p_{\theta_{pred}}$ and offline encoder $\mathcal{E}_{\varphi_{gene}}$, respectively. During training, the online parameters θ are updated through gradient descent, while the offline encoder parameters φ are stop-gradient and updated as an exponential moving average of θ_{gene} at each step. This approach alleviates the need for numerous negative samples and ensures stable and efficient learning on the large-scale protein graph.

Contrastive learning loss. To establish functional and mechanistic relationships between genetic and chemical perturbagens, UniPert adopts a contrastive learning strategy[85] to align representations of paired perturbagens (e.g., a compound and its protein target). Specifically, given a minibatch of M experimentally validated compound-target interactions, the representations of the i-th pair ($\mathbf{u}_{c,i}$ and $\mathbf{u}_{t,i}$) are encoded as $\mathbf{u}_{c,i} = \mathcal{E}_{\theta_{chem}}(\mathcal{F}_{ECFP}(\mathbf{c})) \in \mathbb{R}^{\mathrm{D}}$ and $\mathbf{u}_{t,i} = \mathcal{E}_{\theta_{gene}}(\mathbf{X}_{\mathbf{V}_{t,i}}, \mathbf{G}_{t,i}) \in \mathbb{R}^{\mathrm{D}}$, respectively. To promotes a dense clustering of positive

pairs in the embedding space, UniPert optimizes the symmetric contrastive loss $\mathcal{L}_{align}$

$$\mathcal{L}_{align} = -\frac{1}{2M}\sum_{i=1}^{M}\left(\log\frac{\exp\left(\mathbf{u}_{c,i}{}^{T}\mathbf{u}_{t,i}/\tau\right)}{\sum_{j=1}^{M}\exp\left(\mathbf{u}_{c,j}\mathbf{u}_{t,j}/\tau\right)} + \log\frac{\exp\left(\mathbf{u}_{t,i}{}^{T}\mathbf{u}_{c,i}/\tau\right)}{\sum_{j=1}^{M}\exp\left(\mathbf{u}_{t,i}{}^{T}\mathbf{u}_{c,i}/\tau\right)}\right), \quad (2)$$

where $\tau > 0$ is the temperature parameter that regulates the similarity distribution. This alignment ensures embedding coherence across molecules at different scales, enhancing the biological interpretability and robustness of the unified representation space.

Implementation details.

We implemented UniPert model training using PyTorch (Version 1.12.1)[86] and PyTorch Geometric (Version 2.5.2)[87] on a single NVIDIA Tesla V100-SXM2-32GB GPU. Optimal hyperparameters were determined through grid search, with selection based on the comprehensive performance ranking across $\mathcal{L}_{enhance}$ and $\mathcal{L}_{align}$ losses. TensorBoard[88] was employed for real-time tracking and visualization of training performance.

To facilitate direct inference of representations from new query gene or compound names using our trained UniPert model, we also integrated the getSequence[89] and PubChemPy[90] packages to encapsulate UniPert. This integration allows UniPert to retrieve corresponding amino acid sequences or SMILES strings via online API interfaces and then generate perturbagen embeddings.

Implementation for genetic perturbation outcome prediction task

To evaluate the performance of different genetic perturbagen encoders on the task, we employed the GEARS-derived framework (illustrated in Fig. 3a). We compared and evaluated the impact of various encoders, including the original GEARS[39] encoder, PseAAC[50], ESM[41], OntoProtein[49] and UniPert. Apart from the original GEARS re-learns perturbagen representations for each dataset training based on prior GO knowledge, other methods rely on pre-trained/pre-defined models that capable of directly obtaining sequence-derived features and generalize to out-of-domain amino acid sequences. These encoding methods are described below:

PseAAC[50]. A widely used pseudo amino acid composition method that considers the sequential effects of amino acids in a protein. We utilized the propy3 package[91] to calculate 68-dimensional PseAAC embeddings with the parameter λ set to 24 to accommodate the shortest sequence length of 24 in the reference protein library.

ESM[41]. A transformer-based large-scale pre-trained protein language model that generates 1280-dimensional informative embedding (esm2_t33_650M_UR50D version) with evolutionary patterns for each given sequence.

OntoProtein[49]. A knowledge-enhanced protein pre-training model incorporating GO knowledge graph information that generate 1024-dimensional embeddings for each sequence.

These methods generate embeddings of varying dimensionalities, which are subsequently mapped to a unified dimension through fully connected layers for a fair comparison. The training hyperparameter settings used for genetic perturbation outcome prediction tasks are consistent with the original GEARS study[39].

Implementation for chemical perturbation outcome prediction task

To evaluate the performance of different chemical perturbagen encoders in predicting chemical perturbation outcomes, we employed a CPA-derived framework (illustrated in Fig. 4a). The comparative encoders including ECFP[44] (used in chemCPA[40]), Uni-Mol[51], KPGT[42], and UniPert, , all of which are capable of representing unseen compounds. The details of each encoder are described as follows:

ChemCPA[40]. Following the tutorial of extended CPA to predict unseen compound perturbation response, we utilized the RDKit tool[39] to compute 2048-dimensional ECFP features with a radius of 2 for each given small molecule.

Uni-Mol[51]. A transformer-based 3D molecular representation model that produces 512-dimensional embedding for each small molecule.

KPGT[42]. A knowledge-guided pre-trained graph transformer model that incorporates molecular graph structures and external knowledge to generate 2304-dimensional embeddings for each small molecule.

The training hyperparameter settings for CPA-derived chemical perturbation outcome prediction training with all encoders followed the CPA tutorial[40].

Implementation for genetic-to-chemical perturbation transfer learning

A two-phase *genetic-to-chemical* perturbation outcome transfer learning strategy (illustrated in Fig. 5b) is designed to train a contextualized perturbation response prediction model $\mathcal{C}_{\xi}^{\mathcal{D}}$ with learnable parameters ξ for accelerating large-scale *in silico* drug screening by effectively transferring response from perturbagen genetic to chemical perturbagens with UniPert encoders $\mathcal{E}_{\theta_{gene}}$ and $\mathcal{E}_{\theta_{chem}}$. Specifically, given a context-specific dual-domain perturbation dataset $\mathcal{D} = \{\mathbf{P}_{ctrl}, \mathbf{P}_{gene}, \mathbf{A}_{gene}, \mathbf{P}_{chem}, \mathbf{A}_{chem}\}$, containing L-dimensional unperturbed (control group)

phenotype profiles $\mathbf{P}_{\text{ctrl}} \in \mathbb{R}^{N_{\text{ctrl}} \times L}$, post-genetic-perturbation phenotype profiles $\mathbf{P}_{gene} \in \mathbb{R}^{N_{gene} \times L}$, post-chemical-perturbation $\mathbf{P}_{chem} \in \mathbb{R}^{N_{chem} \times L}$, along with their corresponding genetic and chemical perturbagen annotations $\mathbf{A}_{gene} = \{ \boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_{N_{\text{gene}}} \}$ and $\mathbf{A}_{chem} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_{N_{chem}}\}$.

Pre-training phrase. $\mathcal{C}_{\xi}^{\mathcal{D}}$ is initially trained using genetic perturbation data, $\mathcal{D}_{gene} = \{\mathbf{P}_{ctrl}, \mathbf{P}_{gene}, \mathbf{A}_{gene}\}$, which, due to advancements in current screening technologies, usually owns a more variety of perturbagen types compared to lab-produced chemical perturbation data. The model predicts post-genetic-perturbation phenotype profiles as $\mathbf{P}_{gene} = \mathcal{C}_{\xi}^{\mathcal{D}}(\mathbf{P}_{ctrl}, \mathbf{U}_{gene})$, where $\mathbf{U}_{gene} = \mathcal{E}_{\theta_{gene}}(\mathbf{A}_{gene})$ represent generated genetic perturbagen embeddings by UniPert. Following pre-training, the model $\mathcal{C}_{\xi}$ systematically learns the genetic characteristics of the cellular environment and the response patterns to perturbagens targeting distinct molecular pathways. This enables it to effectively predict phenotypic outcomes of genetic perturbations and establish a robust foundation for subsequent chemical perturbation simulation.

Fine-tuning phrase. $\mathcal{C}_{\xi}^{\mathcal{D}}$ continues to be trained using chemical perturbation data $\mathcal{D}_{chem} = \{\mathbf{P}_{ctrl}, \mathbf{P}_{chem}, \mathbf{A}_{chem}\}$, following $\mathbf{P}_{chem} = \mathcal{C}_{\xi}^{\mathcal{D}}(\mathbf{P}_{ctrl}, \mathbf{U}_{chem})$, where the chemical perturbagen embeddings $\mathbf{U}_{chem} = \mathcal{E}_{\theta_{chem}}(\mathbf{A}_{chem})$ are generated with UniPert. Leveraging the pre-trained knowledge from genetic perturbations, $\mathcal{C}_{\xi}^{\mathcal{D}}$ is further fully fine-tuned to accurately model chemical perturbation responses, thereby effectively transferring the learned perturbation dynamics across domains to enhance prediction performance.

All experiments involving *genetic-to-chemical* perturbation transfer learning, conducted on both image-based and transcriptome-based datasets in this study, were implemented within the GEARS-derived perturbation outcome prediction framework[39].

Datasets

Compound-target interaction dataset. We curated a comprehensive compound-target interaction dataset according to the involved compounds of the CMAP project[26]. This dataset was gathered from the PubChem resource that integrates 8 related databases, which consists of 81,397 pairs across 8,533 compounds and 4,250 proteins. The corresponding SMILES sequences and amino acid sequences were collected for UniPert model training[81].

Target and drug pharmacological class annotations. Pharmacological class information for

protein targets and small-molecule drugs is obtained from the ChEMBL database[92], including 4,417 protein targets with multi-level hierarchical pharmacological classes, as well as 2,812 drugs across 928 MOA classes. The top 18 MOA categories used for drug embedding semantic similarity analysis cover 726 drugs. Detailed information on pharmacological class full names and abbreviations is provided in Supplementary Tables 1 and 2.

***Dixit et al.* dataset.** The Dixit et al. dataset is a single-gene perturbation dataset downloaded from the processed data provided by the GEARS study[39,60], including 44,375 profiles of 20 genes perturbing the K562 cell line.

***Norman et al.* dataset.** The Norman et al. dataset is a combination genetic perturbation dataset, also downloaded from the processed data available through the GEARS study[93,39], comprising 91,205 profiles from 283 single and combination genes perturbing the K562 cell line.

sciPlex3 dataset. The sciPlex3 dataset is a large-scale chemical perturbation dataset involving 187 drugs and 3 cell lines, downloaded from the processed data from the CPA study[53,40]. It was used to evaluate the performance in predicting perturbation outcomes for drugs held out in pathway-based hierarchical splits.

HDAC inhibitor dataset. The HDAC inhibitor dataset is a small-scale chemical perturbation dataset extracted from the sciPlex3 dataset for leave-one-drug-out evaluation. It includes phenotype data from both the untreated control group and cells treated with 17 HDAC inhibitors at 1 μM for 24 hours across 3 cell lines.

CPJUMP1 datasets. The CPJUMP1 datasets[38] include dual-domain perturbation data for two cell lines, i.e., A549 and U2OS. We downloaded pre-processed well-level morphological profiles, with each profile containing 838-dimensional features generated by the image feature extraction software CellProfiler[94] and the pre-processing tool Pycytominer[95], resulting the CPJUMP1-A549 and CPJUMP1-U2OS datasets with detailed information in Supplementary Table 3.

LINCS datasets. We extracted transcriptome-based perturbation data from the CMap LINCS resource[26] (level 5 signatures of LINCS2020 Release) for cell lines that underwent both CRISPR knockout and compound screening. These data were subjected to several quality control and pre-processing steps, including batch correction, differential expression normalization, and replicate consensus processing, which ensured greater robustness and more significant perturbation effects compared to the original readouts. We selected cell lines with the top 5 number of perturbagens under specific conditions (CRISPR sgRNA infection for 96h and compound treatment at 10uM for 24h), forming five cell-line-specific dual-domain

perturbation datasets, i.e., LINCS-A375, LINCS-A549, LINCS-HT29, LINCS-MCF7, LINCS-PC3, with detailed information in Supplementary Table 3.

Evaluation metrics

Target embedding clustering analysis. To evaluate and compare the biological interpretability of target embeddings generated by UniPert and other sequence-based encoders, we apply two clustering metrics: normalized mutual information (NMI) and adjusted Rand index (ARI). These metrics measure the consistency between the predicted clusters and the true target categories, with higher values suggesting better clustering performance.

Compound embedding semantic similarity analysis. To quantify and compare the semantic distance between MOA-specific drug embeddings and all collected drug embeddings from UniPert and ECFP, we use squared mean difference (SMD). SMD measures the effect size between different embedding groups, with larger values indicating greater differences between the compared embedding groups.

Perturbation outcome prediction. For all perturbation outcome prediction task, we use two evaluation metrics: Pearson correlation coefficient (PCC) and mean squared error (MSE). Both metrics provide complementary insights into the prediction performance, with higher PCC values indicating better directional alignment and lower MSE values indicating more accurate magnitude predictions.

Perturbation effect connectivity analysis. To quantify the phenotypic similarity between two post-perturbation phenotype profiles $\mathbf{P}_i$ and $\mathbf{P}_j$ of perturbagens i and j, we calculate their connectivity based on the weighted connectivity score employed in the CMap analyses[26], which is a composite, bi-directional version of the weighted Kolmogorov-Smirnov enrichment statistic (ES)[96]. The connectivity between $\mathbf{P}_i$ and $\mathbf{P}_j$ is calculated as follows:

$$\text{connectivity}(\mathbf{P}_i, \mathbf{P}_j) = \begin{cases} (\text{ES}_{up} - \text{ES}_{down})/2, & if\ \text{sgn}(\text{ES}_{up}) \neq \text{sgn}(\text{ES}_{up}) \\ 0, & otherwise \end{cases}, \quad (3)$$

where ES_{up} is the enrichment of $\mathbf{P}_i{}^{up}$ in $\mathbf{P}_j$ and ES_{down} is the enrichment of $\mathbf{P}_i{}^{down}$ in $\mathbf{P}_j$. Genes with positive scores are considered upregulated and those with negative scores are considered downregulated in our inference perturbated profiles of the LINCS datasets. This approach allows us to assess the similarity in perturbation-induced phenotypic responses across various perturbagen conditions, effectively capturing the connectivity of perturbagen effects in a given cellular context.

Data availability

All datasets used in this work were obtained from public repositories. The human genome-wide proteins, as well as their corresponding sequences and family annotations were obtained from https://www.genenames.org/ and https://www.uniprot.org/. The compound-target interaction dataset was obtained from https://pubchem.ncbi.nlm.nih.gov/. The target and drug pharmacological class annotations were obtained from https://www.ebi.ac.uk/chembl/. The *Dixit et al.* and *Norman et al.* datasets were obtained from https://github.com/snap-stanford/GEARS. The sciPlex3 dataset was obtained from https://github.com/theislab/cpa. The CPJUMP1 datasets were obtained from https://github.com/jump-cellpainting. The LINCS datasets were obtained from https://clue.io/data. The pathway annotations were obtained from https://www.pharmgkb.org/.

Author contributions

Y.L., F.Y., M.L. and J.Y. conceived the study. Y.L. developed the algorithm and conducted experiments under the supervision of F.Y. and J.Y.. L.L. and J.Z. provided valuable advice on UniPert framework and downstream task design. L.H. gave suggestions for the cross-domain transfer learning implications. Y.L. wrote the manuscript. Y.L. finished the figures under the guidance of F.Y. and J.Y.. F.Y., J.Y., L.L, F.W., M.Z. and M.L. revised the manuscript. J.Z. and F.Y. conducted the code review. All authors reviewed and approved the manuscript.

Acknowledgements

The authors appreciate informative conversations with Yu Rong and Zetian Zheng, helpful visualization supports and suggestions from Shaoning Li, and Yin Fang. This work was partially supported by the National Natural Science Foundation of China (Grant 62225209), the Young Elite Scientists Sponsorship Program by CAST (2023QNRC001 to F.Y.), and the Fundamental Research Funds for the Central Universities of Central South University (No. 2023ZZTS0627). Figures 1a was created with BioRender.com.

References

1. Jinek, M. *et al.* A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **337**, 816–821 (2012).

2. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).

3. Meyers, S., Demeyer, S. & Cools, J. CRISPR screening in hematology research: from bulk to single-cell level. *J Hematol Oncol* **16**, 1–22 (2023).

4. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

5. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**, 15 (2018).

6. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol* **36**, 983–987 (2018).

7. Hao, Y. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* **42**, 293–304 (2024).

8. Finan, C. *et al.* The druggable genome and support for target identification and validation in drug development. *Science Translational Medicine* **9**, eaag1166 (2017).

9. Das, S., Bano, S., Kapse, P. & Kundu, G. C. CRISPR based therapeutics: a new paradigm in cancer precision medicine. *Mol Cancer* **21**, 85 (2022).

10. Pun, F. W., Ozerov, I. V. & Zhavoronkov, A. AI-powered therapeutic target discovery. *Trends in Pharmacological Sciences* **44**, 561–572 (2023).

11. Moffat, J., Komor, A. C. & Lum, L. Impact of CRISPR in cancer drug discovery. *Science* **386**, 378–379 (2024).

12. Hnatiuk, A. P., Briganti, F., Staudt, D. W. & Mercola, M. Human iPSC modeling of heart disease for drug development. *Cell Chemical Biology* **28**, 271–282 (2021).

13. Kim, G. *et al.* Genome-wide CRISPR screen reveals v-ATPase as a drug target to lower levels of ALS protein ataxin-2. *Cell Reports* **41**, (2022).

14. Hendriks, D. *et al.* Engineered human hepatocyte organoids enable CRISPR-based target discovery and drug screening for steatosis. *Nat Biotechnol* **41**, 1567–1581 (2023).

15. Awad, S. A. *et al.* Integrated drug profiling and CRISPR screening identify BCR::ABL1-independent vulnerabilities in chronic myeloid leukemia. *CR Med* **5**, (2024).

16. Yang, F. *et al.* Genome-wide CRISPR screens identify CLC-2 as a drug target for anti-herpesvirus therapy: tackling herpesvirus drug resistance. *Sci. China Life Sci.* **68**, 515–526 (2025).

17. Abraham, V. C., Taylor, D. L. & Haskins, J. R. High content screening applied to large-

scale cell biology. *Trends in Biotechnology* **22**, 15–22 (2004).

18. Rood, J. E., Hupalowska, A. & Regev, A. Toward a foundation model of causal cell and tissue biology with a Perturbation Cell and Tissue Atlas. *Cell* **187**, 4520–4545 (2024).

19. Wang, B., Xu, Y., Wan, A. H., Wan, G. & Wang, Q.-P. Integrating genome-wide CRISPR screens and in silico drug profiling for targeted antidote development. *Nat Protoc* **19**, 2739–2770 (2024).

20. Pacini, C. *et al.* A comprehensive clinically informed map of dependencies in cancer cells and framework for target prioritization. *Cancer Cell* **42**, 301-316.e9 (2024).

21. Chavez, M., Chen, X., Finn, P. B. & Qi, L. S. Advances in CRISPR therapeutics. *Nat Rev Nephrol* (2022) doi:10.1038/s41581-022-00636-2.

22. Vazquez, F. & Sellers, W. R. Are CRISPR Screens Providing the Next Generation of Therapeutic Targets? *Cancer Research* **81**, 5806–5809 (2021).

23. Woo, J. H. *et al.* Elucidating Compound Mechanism of Action by Network Perturbation Analysis. *Cell* **162**, 441–451 (2015).

24. Jones, L. H. & Bunnage, M. E. Applications of chemogenomic library screening in drug discovery. *Nat Rev Drug Discov* **16**, 285–296 (2017).

25. Lamb, J. *et al.* The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science* **313**, 1929–1935 (2006).

26. Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **171**, 1437-1452.e17 (2017).

27. Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170**, 564-576.e16 (2017).

28. Arafeh, R., Shibue, T., Dempster, J. M., Hahn, W. C. & Vazquez, F. The present and future of the Cancer Dependency Map. *Nat Rev Cancer* 1–15 (2024) doi:10.1038/s41568-024-00763-x.

29. Seal, S. *et al.* Cell Painting: a decade of discovery and innovation in cellular imaging. *Nat Methods* 1–15 (2024) doi:10.1038/s41592-024-02528-8.

30. Weisbart, E. *et al.* Cell Painting Gallery: an open resource for image-based profiling. *Nat Methods* **21**, 1775–1777 (2024).

31. Vazquez, F. & Boehm, J. S. The Cancer Dependency Map enables drug mechanism-of-action investigations. *Molecular Systems Biology* **16**, e9757 (2020).

32. McDonagh, E. M. *et al.* Human Genetics and Genomics for Drug Target Identification and Prioritization: Open Targets' Perspective. *Annual Review of Biomedical Data Science* **7**, 59–81 (2024).

33. Rad, A. N. *et al.* Connectivity Map and perturbation-based sensitivity analysis identifies MEK inhibitors as senolytics in human lung fibroblasts. 2024.11.16.623933 Preprint at https://doi.org/10.1101/2024.11.16.623933 (2024).

34. Gonçalves, E. *et al.* Drug mechanism-of-action discovery through the integration of pharmacological and CRISPR screens. *Molecular Systems Biology* **16**, e9405 (2020).

35. Heumos, L. *et al.* Pertpy: an end-to-end framework for perturbation analysis. *bioRxiv* 2024.08.04.606516 (2024) doi:10.1101/2024.08.04.606516.

36. Heimberg, G. *et al.* A cell atlas foundation model for scalable search of similar human cells. *Nature* 1–3 (2024) doi:10.1038/s41586-024-08411-y.

37. Lotfollahi, M., Yuhan Hao, Theis, F. J. & Satija, R. The future of rapid and automated single-cell data analysis using reference mapping. *Cell* **187**, 2343–2358 (2024).

38. Chandrasekaran, S. N. *et al.* Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. *Nat Methods* **21**, 1114–1121 (2024).

39. Roohani, Y., Huang, K. & Leskovec, J. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nat Biotechnol* 1–9 (2023) doi:10.1038/s41587-023-01905-6.

40. Lotfollahi, M. *et al.* Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology* **19**, e11517 (2023).

41. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

42. Li, H. *et al.* A knowledge-guided pre-training framework for improving molecular representation learning. *Nat Commun* **14**, 7568 (2023).

43. Kustatscher, G. *et al.* Understudied proteins: opportunities and challenges for functional proteomics. *Nat Methods* **19**, 774–779 (2022).

44. Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).

45. Kipf, T. N. & Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. (2016) doi:10.48550/ARXIV.1609.02907.

46. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *Journal of Molecular Biology* **147**, 195–197 (1981).

47. Thakoor, S. *et al.* Large-Scale Representation Learning on Graphs via Bootstrapping. (2021) doi:10.48550/ARXIV.2102.06514.

48. Hadsell, R., Chopra, S. & LeCun, Y. Dimensionality Reduction by Learning an Invariant Mapping. in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* vol. 2 1735–1742 (2006).

49. Zhang, N. *et al.* OntoProtein: Protein Pretraining With Gene Ontology Embedding. (2022) doi:10.48550/ARXIV.2201.11147.

50. Chou, K.-C. Prediction of protein cellular attributes using pseudo-amino acid composition.

Proteins: Structure, Function, and Bioinformatics **43**, 246–255 (2001).

51. Zhou, G. *et al.* Uni-Mol: A Universal 3D Molecular Representation Learning Framework. in (2022).

52. Dini, A. *et al.* A multiplex single-cell RNA-Seq pharmacotranscriptomics pipeline for drug discovery. *Nat Chem Biol* 1–11 (2024) doi:10.1038/s41589-024-01761-8.

53. Srivatsan, S. R. *et al.* Massively multiplex chemical transcriptomics at single-cell resolution. *Science* (2020) doi:10.1126/science.aax6234.

54. Yue, K. *et al.* Comparison of three zinc binding groups for HDAC inhibitors – A potency, selectivity and enzymatic kinetics study. *Bioorganic & Medicinal Chemistry Letters* **70**, 128797 (2022).

55. Bredel, M. & Jacoby, E. Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nat Rev Genet* **5**, 262–275 (2004).

56. Gray, J. W. & Mills, G. B. Large-Scale Drug Screens Support Precision Medicine. *Cancer Discovery* **5**, 1130–1132 (2015).

57. Sadybekov, A. V. & Katritch, V. Computational approaches streamlining drug discovery. *Nature* **616**, 673–685 (2023).

58. Bock, C. *et al.* High-content CRISPR screening. *Nat Rev Methods Primers* **2**, 8 (2022).

59. Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat Methods* **11**, 783–784 (2014).

60. Dixit, A. *et al.* Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853-1866.e17 (2016).

61. Morris, J. A., Sun, J. S. & Sanjana, N. E. Next-generation forward genetic screens: uniting high-throughput perturbations with single-cell analysis. *Trends in Genetics* **40**, 118–133 (2024).

62. Peidli, S. *et al.* scPerturb: harmonized single-cell perturbation data. *Nat Methods* 1–10 (2024) doi:10.1038/s41592-023-02144-y.

63. Chandrasekaran, S. N. *et al.* JUMP Cell Painting dataset: morphological impact of 136,000 chemical and genetic perturbations. *bioRxiv* 2023.03.23.534023 (2023) doi:10.1101/2023.03.23.534023.

64. Hewett, M. *et al.* PharmGKB: the Pharmacogenetics Knowledge Base. *Nucleic Acids Research* **30**, 163–165 (2002).

65. Moshkov, N. *et al.* Learning representations for image-based profiling of perturbations. *Nat Commun* **15**, 1–17 (2024).

66. Gu, J. *et al.* Mapping multimodal phenotypes to perturbations in cells and tissue with CRISPRmap. *Nat Biotechnol* 1–15 (2024) doi:10.1038/s41587-024-02386-x.

67. Ramezani, M. *et al.* A genome-wide atlas of human cell morphology. 2023.08.06.552164

Preprint at https://doi.org/10.1101/2023.08.06.552164 (2023).

68. Fay, M. M. *et al.* RxRx3: Phenomics Map of Biology. *bioRxiv* 2023.02.07.527350 (2023) doi:10.1101/2023.02.07.527350.

69. Gavriilidis, G. I., Vasileiou, V., Orfanou, A., Ishaque, N. & Psomopoulos, F. A mini-review on perturbation modelling across single-cell omic modalities. *Computational and Structural Biotechnology Journal* **23**, 1886–1896 (2024).

70. Mitchell, D. C. *et al.* A proteome-wide atlas of drug mechanism of action. *Nat Biotechnol* **41**, 845–857 (2023).

71. Qian, L. *et al.* AI-empowered perturbation proteomics for complex biological systems. *Cell Genomics* **0**, (2024).

72. Messner, C. B. *et al.* The proteomic landscape of genome-wide genetic perturbations. *Cell* **186**, 2018-2034.e21 (2023).

73. Metzner, E., Southard, K. M. & Norman, T. M. Multiome Perturb-seq unlocks scalable discovery of integrated perturbation effects on the transcriptome and epigenome. 2024.07.26.605307 Preprint at https://doi.org/10.1101/2024.07.26.605307 (2024).

74. Liscovitch-Brauer, N. *et al.* Profiling the genetic determinants of chromatin accessibility with scalable single-cell CRISPR screens. *Nat Biotechnol* **39**, 1270–1277 (2021).

75. Nguyen, E. *et al.* Sequence modeling and design from molecular to genome scale with Evo. *Science* **386**, eado9336 (2024).

76. Chen, V. *et al.* Applying interpretable machine learning in computational biology—pitfalls, recommendations and opportunities for new developments. *Nat Methods* **21**, 1454–1461 (2024).

77. Theodoris, C. V. *et al.* Transfer learning enables predictions in network biology. *Nature* (2023) doi:10.1038/s41586-023-06139-9.

78. Bunne, C. *et al.* How to Build the Virtual Cell with Artificial Intelligence: Priorities and Opportunities. *arXiv.org* https://arxiv.org/abs/2409.11654v1 (2024).

79. Landrum, G. *et al.* rdkit/rdkit: 2024_09_3 (Q3 2024) Release. (2024).

80. Seal, R. L. *et al.* Genenames.org: the HGNC resources in 2023. *Nucleic Acids Research* **51**, D1003–D1009 (2023).

81. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research* **51**, D523–D531 (2023).

82. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *Journal of Molecular Biology* **147**, 195–197 (1981).

83. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**, 1026–1028 (2017).

84. Grill, J.-B. *et al.* Bootstrap your own latent: A new approach to self-supervised Learning.

Preprint at https://doi.org/10.48550/arXiv.2006.07733 (2020).

85. Radford, A. *et al.* Learning Transferable Visual Models From Natural Language Supervision. Preprint at https://doi.org/10.48550/arXiv.2103.00020 (2021).

86. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. in *Advances in Neural Information Processing Systems* vol. 32 (Curran Associates, Inc., 2019).

87. Fey, M. & Lenssen, J. E. Fast Graph Representation Learning with PyTorch Geometric. Preprint at https://doi.org/10.48550/arXiv.1903.02428 (2019).

88. Abadi, M. *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems.

89. getSequence: A CLI to get a uniprot sequence returned to terminal.

90. PubChemPy: A simple Python wrapper around the PubChem PUG REST API.

91. propy3: Compute protein descriptors.

92. Zdrazil, B. *et al.* The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research* **52**, D1180–D1192 (2024).

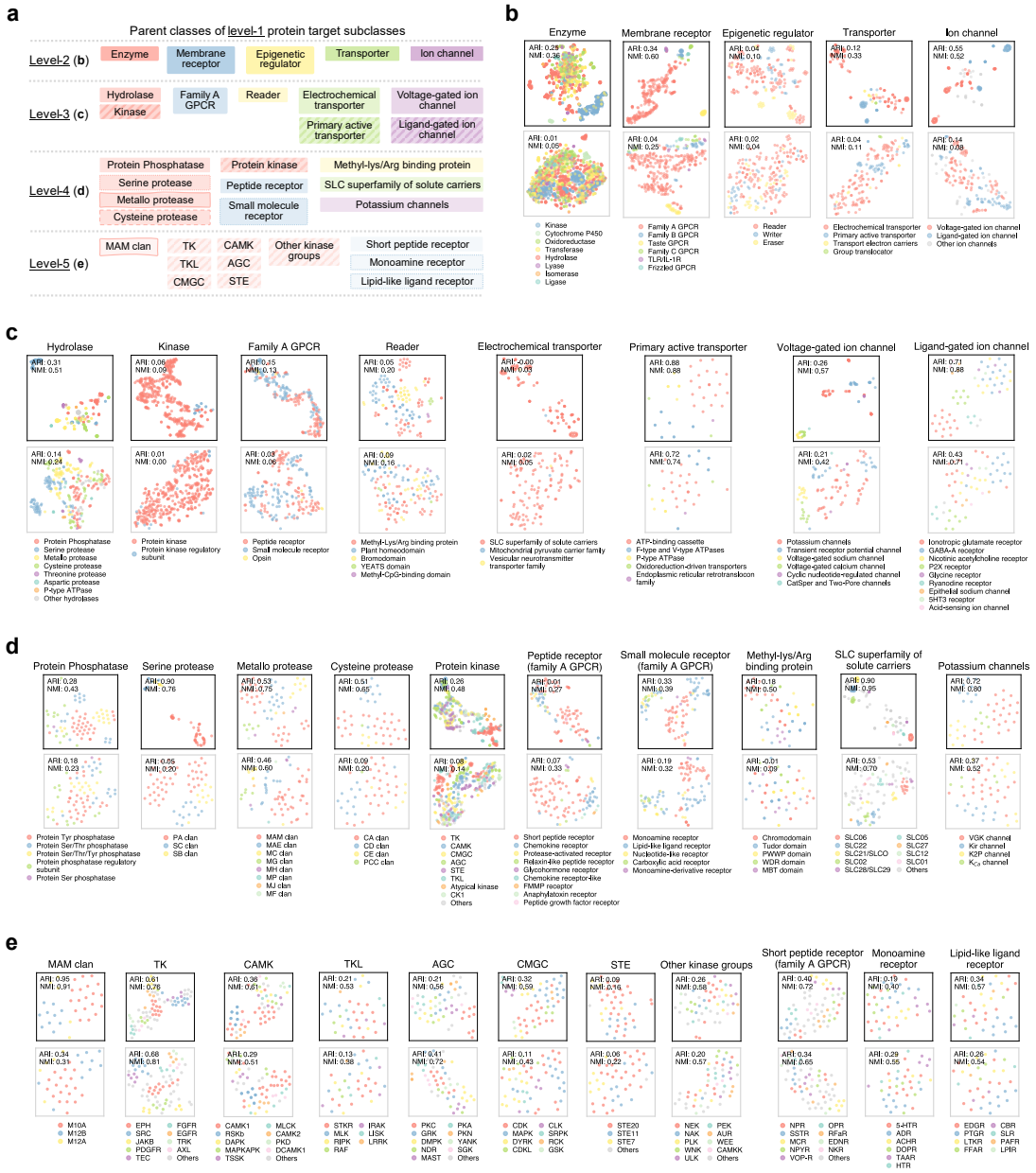
93. Norman, T. M. *et al.* Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science* (2019) doi:10.1126/science.aax4438.

94. Stirling, D. R. *et al.* CellProfiler 4: improvements in speed, utility and usability. *BMC Bioinformatics* **22**, 433 (2021).

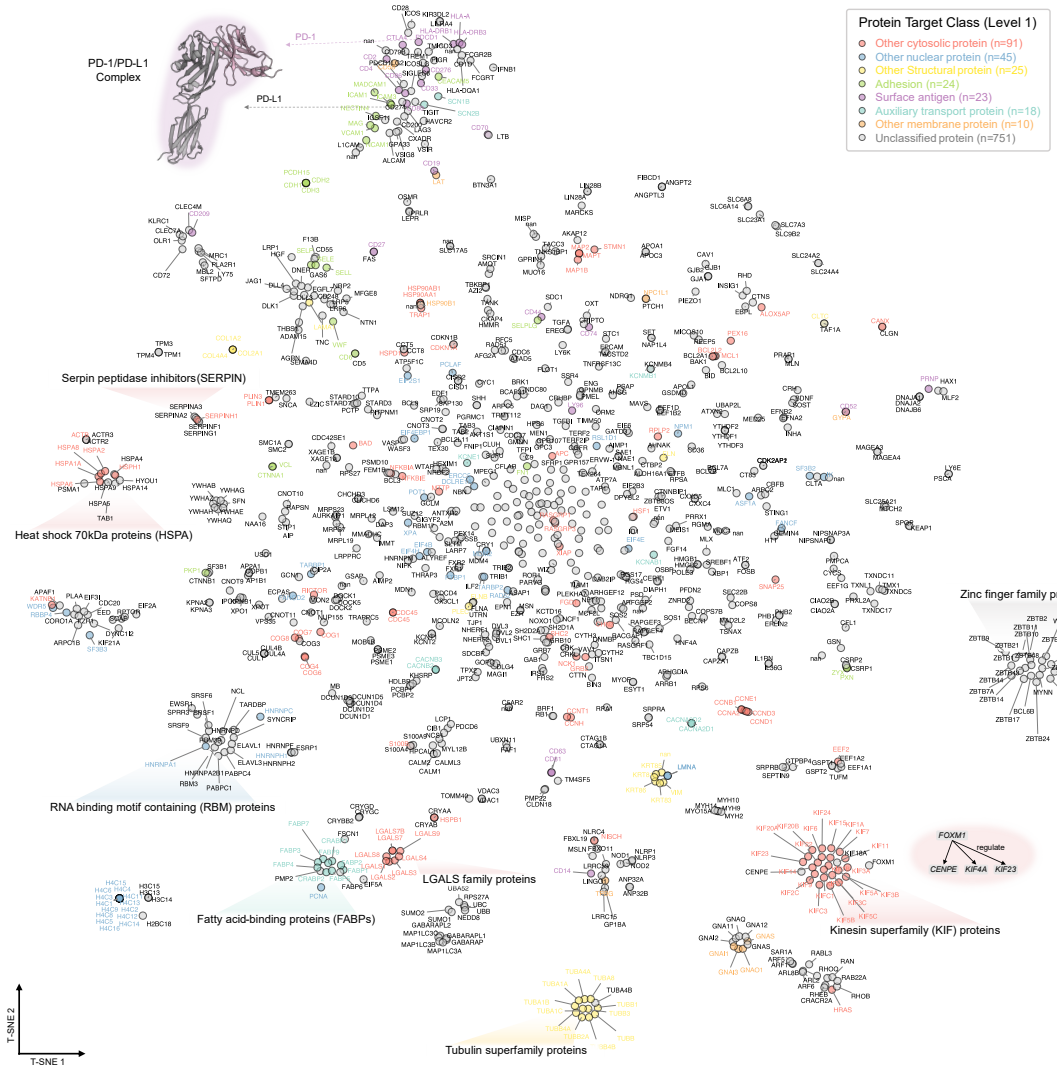
95. Serrano, E. *et al.* Reproducible image-based profiling with Pycytominer. Preprint at https://doi.org/10.48550/arXiv.2311.13417 (2024).

96. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 15545–15550 (2005).

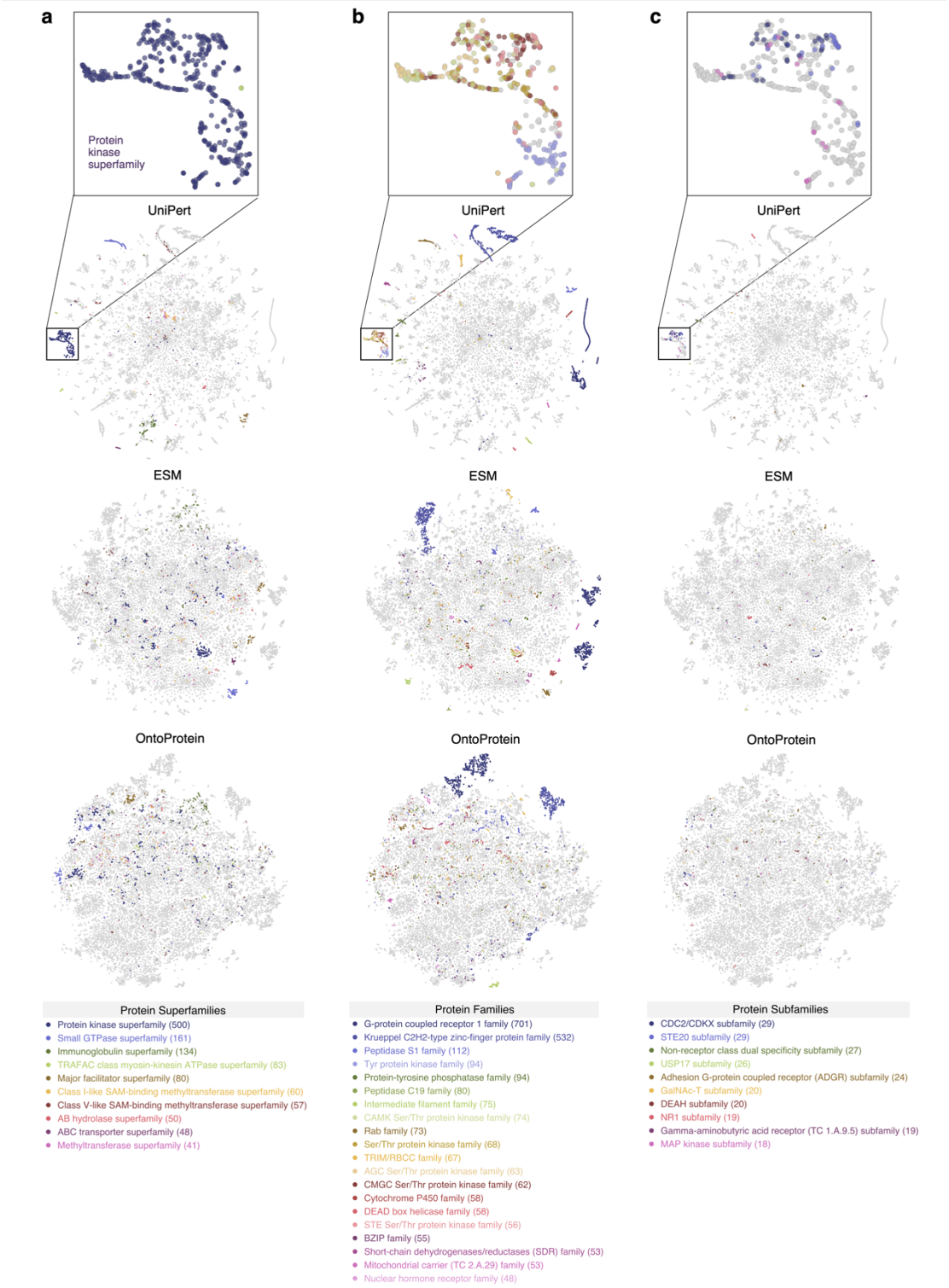
Extended Data



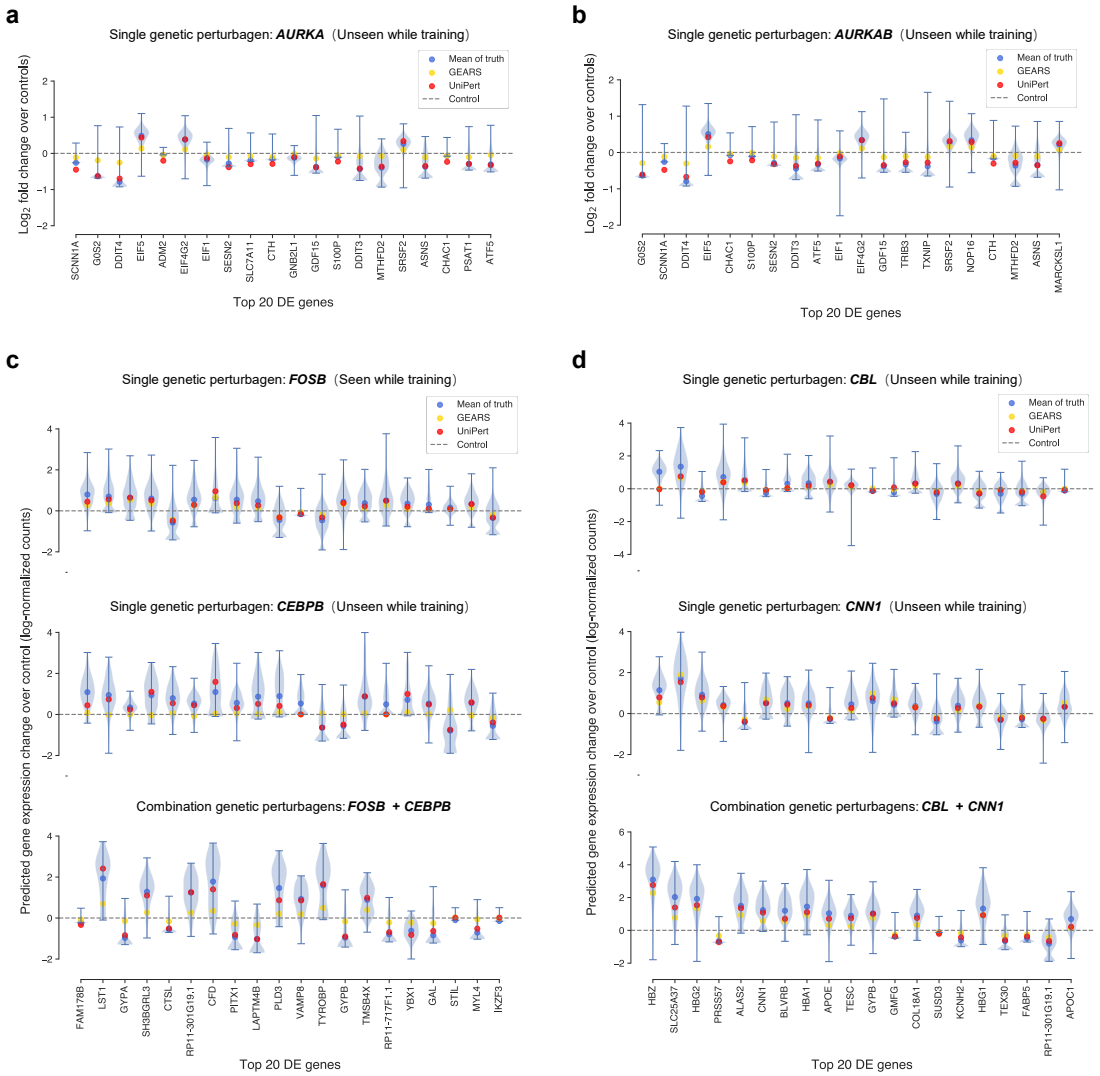
Extended Data Fig. 1 | Performance comparison in identifying hierarchical pharmacological protein target classes between UniPert and ESM embeddings. **a)** Hierarchical relationships of target pharmacological classes with more than 30 members. **b-e)** T-SNE visualizations of protein embeddings by UniPert (top panels, black-framed) and ESM (bottom panels, gray-framed) across hierarchical pharmacological classes at **b)** level 2, **c)** level 3, **d)** level 4, and **e)** level 5. Each panel represents to a specific target class at the corresponding pharmacological class level. Colors of scatters (represent proteins) correspond to distinct subclasses (title) within each parent class (legend). Adjusted Rand index (ARI) and normalized mutual information (NMI) scores were calculated for each panel (upper left) to compare clustering accuracy of UniPert and ESM embeddings against true protein subclasses using the K-means method.



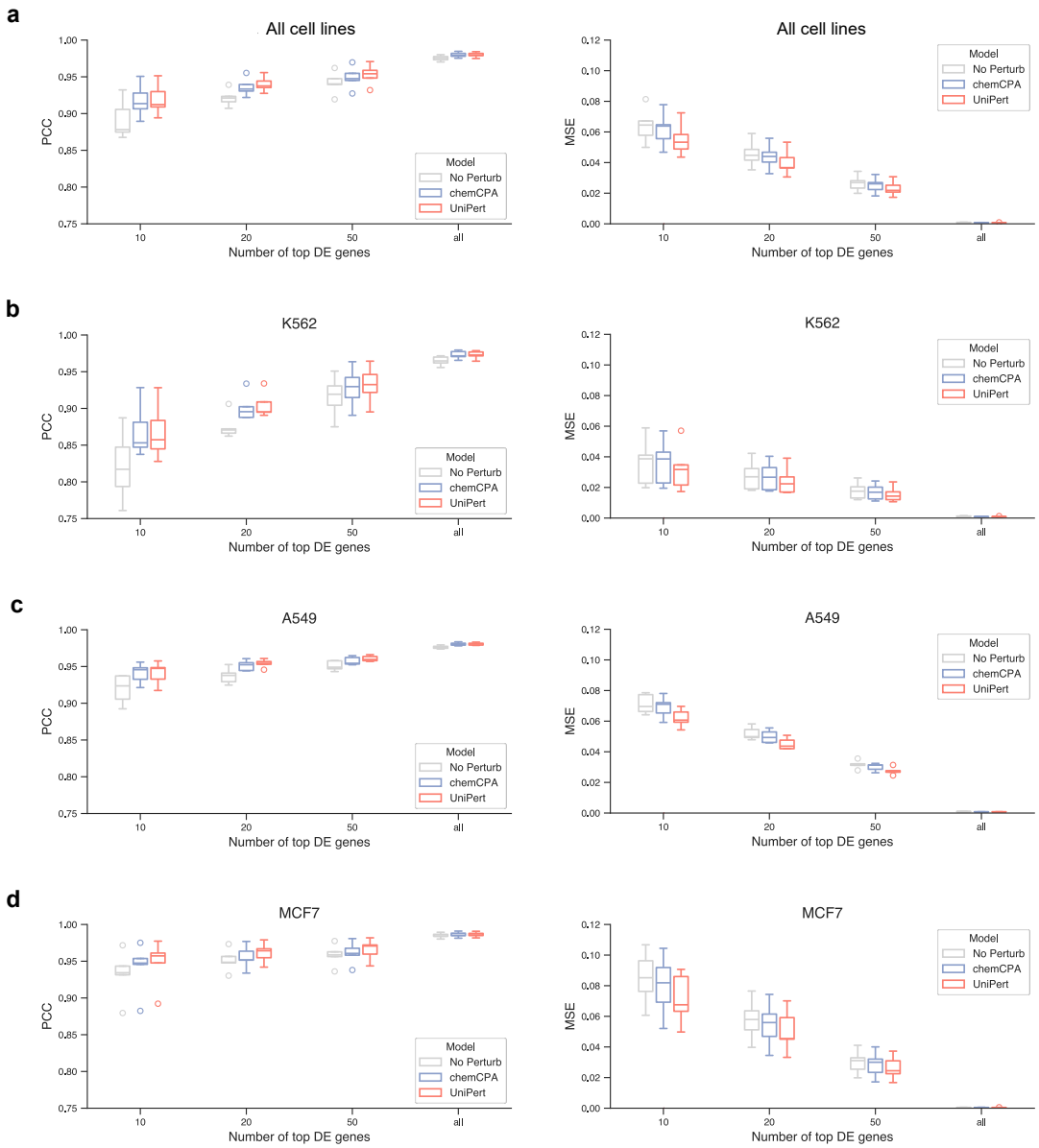
Extended Data Fig. 2 | UniPert annotates protein families and key relationships in the embedding space. T-SNE visualizations of UniPert target embeddings of minority and unclassified entries of level 1 pharmacological protein target classes. Protein families are clustered notably in the embedding space, such as zinc finger proteins (right), kinesin superfamily (KIF) proteins (right bottom), tubulin proteins (bottom), heat shock 70kDa proteins (left), fatty acid-binding proteins (left bottom), LGALS proteins (bottom), RNA binding motif (RBM) proteins (left bottom), and serpin peptidase inhibitors (SERPIN) (left top). The close position of FOMX1 to the KIF proteins (right bottoms) corresponds to the regulatory association between them. Additionally, the subunits of PD-1/PD-L1 complex (top) appear in proximity underscoring key interaction patterns.



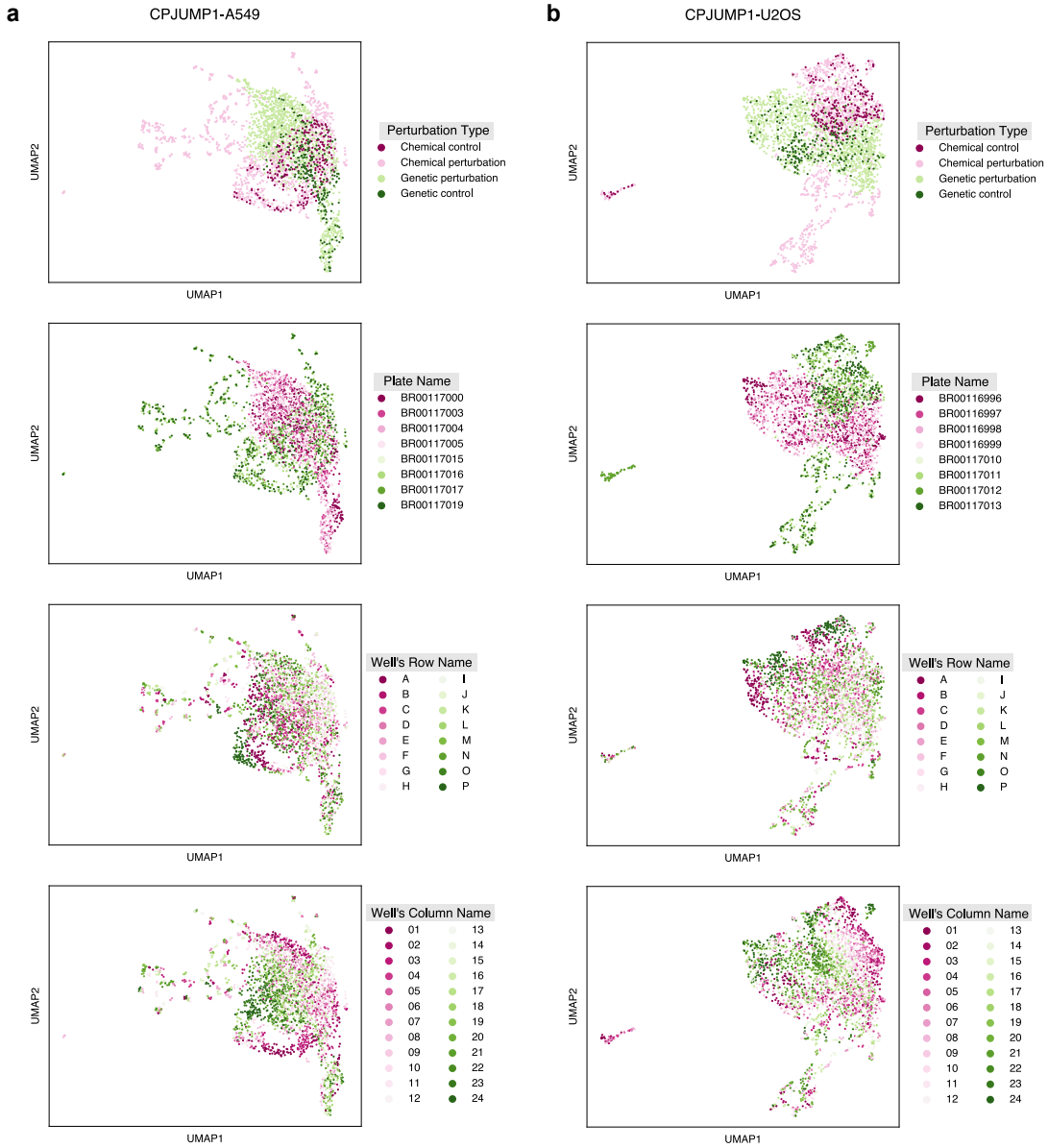
Extended Data Fig. 3 | UniPert distinguishes hierarchical protein family groups in genome-wide protein embedding space. T-SNE visualizations of genome-wide protein (n=19,443) are presented using UniPert, along with state-of-the-art protein language models, ESM and OntoProtein. Colors denote **a)** the top 10 superfamilies, **b)** the top 20 families, and **c)** the top 10 subfamilies, with corresponding annotations provided on the right. The black frames (top) highlight the superior clustering capability of UniPert for the kinase superfamily.



Extended Data Fig. 4 | Perturbation outcome prediction of top 20 DEGs after single gene and gene combination perturbations using UniPert and GEARS. Mean predicted values of the top 20 DEGs after **a)** unseen single gene *AURKA* and **b)** unseen single gene *AURKAB* perturbing in the *Dixit et al.* dataset. **c)** Mean predicted values of top 20 DEGs after seen single gene *FOSB*, unseen single gene *CEBPB*, and gene combination *FOSB+CEBPB* perturbing, as well as **d)** unseen single gene *FOSB*, *CEBPB*, and unseen gene combination *CBL+CNN1* perturbing in the *Norman et al.* dataset.



Extended Data Fig. 5 | Performance comparison at multiple DEG scales and cell lines of UniPert and chemCPA in predicting drug perturbation outcomes. **a)** Overall performance on the sciPlex3 dataset across three cell lines, assessed using PCC (left) and MSE (right). **b)-d)** Performance results shown separately for each cell line, highlighting the comparative performance of UniPert.



Extended Data Fig. 6 | Batch effects in image-based profiled CPJUMP1 datasets. **a)** CPJUMP1-A549 dataset. **b)** CPJUMP1-U2OS dataset. UMAP plots display projections of profiles under different conditions (i.e., perturbation types, plate identities, and plate well layouts) arranged sequentially from top to bottom. Variations caused by differences in experimental conditions, such as imaging batches, plate layouts, or technical processing, have the potential to introduce non-biological artifacts into the data. Notably, the CPJUMP1 dataset demonstrates a pronounced influence of perturbation types and well column positions on the profiles.