nature machine intelligence

Article

https://doi.org/10.1038/s42256-022-00541-0

A context-aware deconfounding autoencoder for robust prediction of personalized clinical drug response from cell-line compound screening

Received: 24 October 2021

Accepted: 8 September 2022

Published online: 17 October 2022

Check for updates

Di He[1], Qiao Liu[2], You Wu[1] and Lei Xie[1,2,3] ✉

Accurate and robust prediction of patient-specific responses to a new compound is critical to personalized drug discovery and development. However, patient data are often too scarce to train a generalized machine learning model. Although many methods have been developed to utilize cell-line screens for predicting clinical responses, their performances are unreliable owing to data heterogeneity and distribution shift. Here we have developed a novel context-aware deconfounding autoencoder (CODE-AE) that can extract intrinsic biological signals masked by context-specific patterns and confounding factors. Extensive comparative studies demonstrated that CODE-AE effectively alleviated the out-of-distribution problem for the model generalization and significantly improved accuracy and robustness over state-of-the-art methods in predicting patient-specific clinical drug responses purely from cell-line compound screens. Using CODE-AE, we screened 59 drugs for 9,808 patients with cancer. Our results are consistent with existing clinical observations, suggesting the potential of CODE-AE in developing personalized therapies and drug response biomarkers.

Omics profiling, particularly transcriptomics, is a powerful technique to characterize cellular activity under various conditions, allowing the development of machine learning models for personalized phenotype compound screening[1–3]. However, the success of such predictive models largely relies on the availability of sufficient amounts of high-quality labelled data. In the early stage of drug discovery, cell-line and other in vitro models have been extensively applied to screen drug candidates. Unfortunately, the activity of a compound in vitro is poorly correlated with its efficacy in humans. This discrepancy is responsible for the high cost and low success rate of drug discovery. Even for drugs that have been tested in clinical, patient responses to the drug can remarkably vary. However, it is often difficult to collect a large number of coherent patient data with drug treatment and response history to reliably predict which patient will benefit from the drug. A robust predictive model that can utilize bioactivity data of a compound from a panel of in vitro screens to predict patients' clinical responses will no doubt fill in a critical knowledge gap between the in vitro activity and the clinical outcome of a drug candidate, thereby facilitating drug discovery and precision medicine. Nevertheless, it is a challenging task owing to the biological and environmental differences between in vitro models and humans as well as various confounding factors and overwhelming context-specific patterns that may mask intrinsic drug response signals.

The difficulty in predicting patient-specific clinical drug responses from in vitro screens using machine learning originates from a fundamental challenge of the out-of-distribution (OOD) problem. The underlying assumption of existing machine learning methods is that

[1]PhD program in Computer Science, Graduate Center, City University of New York, New York, NY, USA. [2]Department of Computer Science, Hunter College, City University of New York, New York, NY, USA. [3]Feil Family Brain and Mind Research Institute, Weill Cornell Medicine, Cornell University, New York, NY, USA. ✉e-mail: lxie@iscb.org

the data distribution of training data and unseen testing data are the same. When applying the machine learning model trained from in vitro data to patient samples, the performance could subtantially deteriorate owing to the data distribution shift. Current efforts in solving the OOD problem include domain adaptation and meta-learning. Many domain adaptation methods have been proposed in computer vision and natural language processing. However, their application to aligning in vitro with patient data could be sub-optimal owing to the noisy and heterogeneous nature of omics data. The data shift in omics data mainly comes from two sources: technical confounders such as batch effects and biological confounders (for example, sex and age). Early work used co-expression extrapolation (COXEN) to extract common drug response biomarkers from distinct datasets for translating drug activities in cell lines to clinical responses[4]. However, the performance of COXEN may be compromised by the high dimensionality of data and confounding factors. Several recent methods, notably Velodrome[5] and Celligner[6], have been developed to globally align transcriptomics profiles in a low-dimensional embedding space. These domain alignment methods are suitable for removing systemic biases resulting from the technical confounder but are incapable of disentangling intrinsic drug response biomarkers from the biological confounder. Adversarial deconfounding autoencoder (ADAE) is a method to facilitate the domain adaptation of gene expression profiles[7], but ADAE has not been tested for translating in vitro data to patient data. A meta-learning approach named Translation of Cellular Response Prediction (TCRP) has recently been proposed[8] to improve the transferability of predictive drug response models from in vitro screens to clinical settings. However, TCRP still requires a certain number of patient data for each drug tested to train the predictive model. It is often infeasible to obtain such data, especially for a new lead compound. Thus, the actual application of TCRP to drug discovery is limited. Another relevant work has applied variational autoencoder (VAE) pre-training followed by Elastic Net supervised training (VAEN) to learn cell-line models and applied them to impute clinical drug response[9]. However, VAEN is not optimized to reliably transfer cell-line data to patient samples and disentangle confounding factors[9] owing to the limitation of VAE.

The unsolved question is how we cannot only remove systematic biases between two data modalities but also extract and align their common drug response biomarkers from observed gene expressions that are entangled with context-specific signals so that we can robustly predict individual patient responses to a new drug that has never been tested in patients only using in vitro compound screens in the setting of zero-shot learning. To address this problem, we proposed a context-aware deconfounding autoencoder (CODE-AE). In CODE-AE, we devised a self-supervised (pre)training scheme to construct a feature encoding module that can be easily tuned to adapt to the different downstream tasks. We leverage both unlabelled cell lines and patient samples for the self-supervised (pre)training of the encoder. There are two unique features of CODE-AE. First, it can extract both common biological signals shared by incoherent samples and private representations unique to them, thus separating confounding factors between data modalities. Second, CODE-AE aligns drug response signals locally by separating them from confounders. In contrast, state-of-the-art domain adaptation methods align two data distributions globally. When drug response signals are entangled with other confounders, a global alignment will not guarantee that the drug response signals can be well aligned. Simply put, CODE-AE can be considered a unique feature selection procedure in an embedding space across incoherent data modalities using both labelled and unlabelled data. The biology-inspired design of CODE-AE allowed us to generalize existing cell-line omics data for the robust prediction of patient-specific clinical responses to new drugs in the setting of zero-shot learning, a critical component for patient-specific compound screening and personalized medicine.

To show the performance lift achieved by CODE-AE, we performed exhaustive comparative studies on CODE-AE variants and other competing methods on the breast cancer patient-derived tumour xenograft ex vivo patient-derived xenograft (PDX)-derived tumor cells (PDTC) dataset[10]. Moreover, to demonstrate the potential of CODE-AE in personalized medicine, we apply CODE-AE to predicting chemotherapy responses for patients, which is a critical obstacle to effective cancer therapy. Our extensive studies show that CODE-AE effectively alleviates the OOD problem when transferring the cell-line model to patient samples, significantly outperforms the state-of-the-art methods multi-layer perceptron neural network (MLP)[3], Elastic Net[11], Velodrome[5], Celligner[6], ADAE[7], TCRP[8], VAEN[9] and COXEN[4] that are specifically designed for transcriptomics data as well as other popular domain adaptation methods VAE[12], denoising autoencoder (DAE)[13], orrelation alignment for deep domain adaptation (Deep CORAL)[14] and domain separation network (DSN)[15] in terms of both accuracy and robustness. Using CODE-AE, we screened 59 drugs for 9,808 patients with cancer. The in vivo compound screening not only further validated CODE-AE but also discovered novel personalized anticancer therapies and drug response biomarkers. Thus CODE-AE provides a useful framework to take advantage of rich in vitro omics data for developing generalized clinical predictive models.

Results and discussion

Overview of CODE-AE

The goal of CODE-AE is to remove both biological and technical confounds and to extract common drug response biomarkers from distinct data domains (Fig. 1a). In practice, the drug response biomarker cannot be obtained directly, especially for the unlabelled target domain, and has to be inferred from observed data and represented in a high-dimensional or embedding space. Different from COXEN[4], which derives co-expressed genes as the common biomarker, CODE-AE infers common features in a nonlinearly mapped low-dimensional embedding space. Furthermore, CODE-AE explicitly separates the common biomarker from domain-specific features, and locally aligns the common biomarker to alleviate the data-shift problem, as illustrated in Fig. 1b.

Algorithmically, the training of CODE-AE follows a pre-training fine-tuning procedure. During the pre-training stage, CODE-AE uses unlabelled data from both the source domain and target domain to pre-train an autoencoder that minimizes a data reconstruction error (Methods). The architecture scheme of CODE-AE is shown in Fig. 1c. Different from conventional autoencoders such as VAE, CODE-AE has two unique features. First, it learns shared signals between the cell-line data (source domain) and the patient data (target domain) as well as private signals that are unique to the cell line and the patient. The rationale is to disentangle common biological signals between datasets from context-specific patterns that overwhelm drug response biomarkers[8]. Second, CODE-AE regularizes the embeddings of cell lines and patients to have their distributions be similar. In this way, the knowledge learned from the cell-line model can be transferred to patients. We test three regularization methods: simple concatenation of cell-line and patient embeddings (CODE-AE-BASE), minimization of their maximum mean discrepancy (MMD) loss (CODE-AE-MMD) and minimization of their adversarial loss (CODE-AE-ADV). After the unsupervised pre-training, a supervised drug response model is trained to fine-tune the aligned common embedding using labelled cell-line data for a specific compound. During the inference stage, patient-specific drug responses to the compound are predicted from the trained cell-line model based on the pre-trained fine-tuned common embedding of the patient. As shown in Supplementary Table 1 and Supplementary Fig. 1, the overall best performing CODE-AE variant is the CODE-AE-ADV. We will only compare CODE-AE-ADV with other baseline models and apply it to actual prediction tasks in the following sections.

CODE-AE alleviates OOD problem on gene expression profiles

We used the shared encoder from pre-trained CODE-AE-ADV to generate the new representations for clinical The Cancer Genome Atlas (TCGA)

https://doi.org/10.1038/s42256-022-00541-0

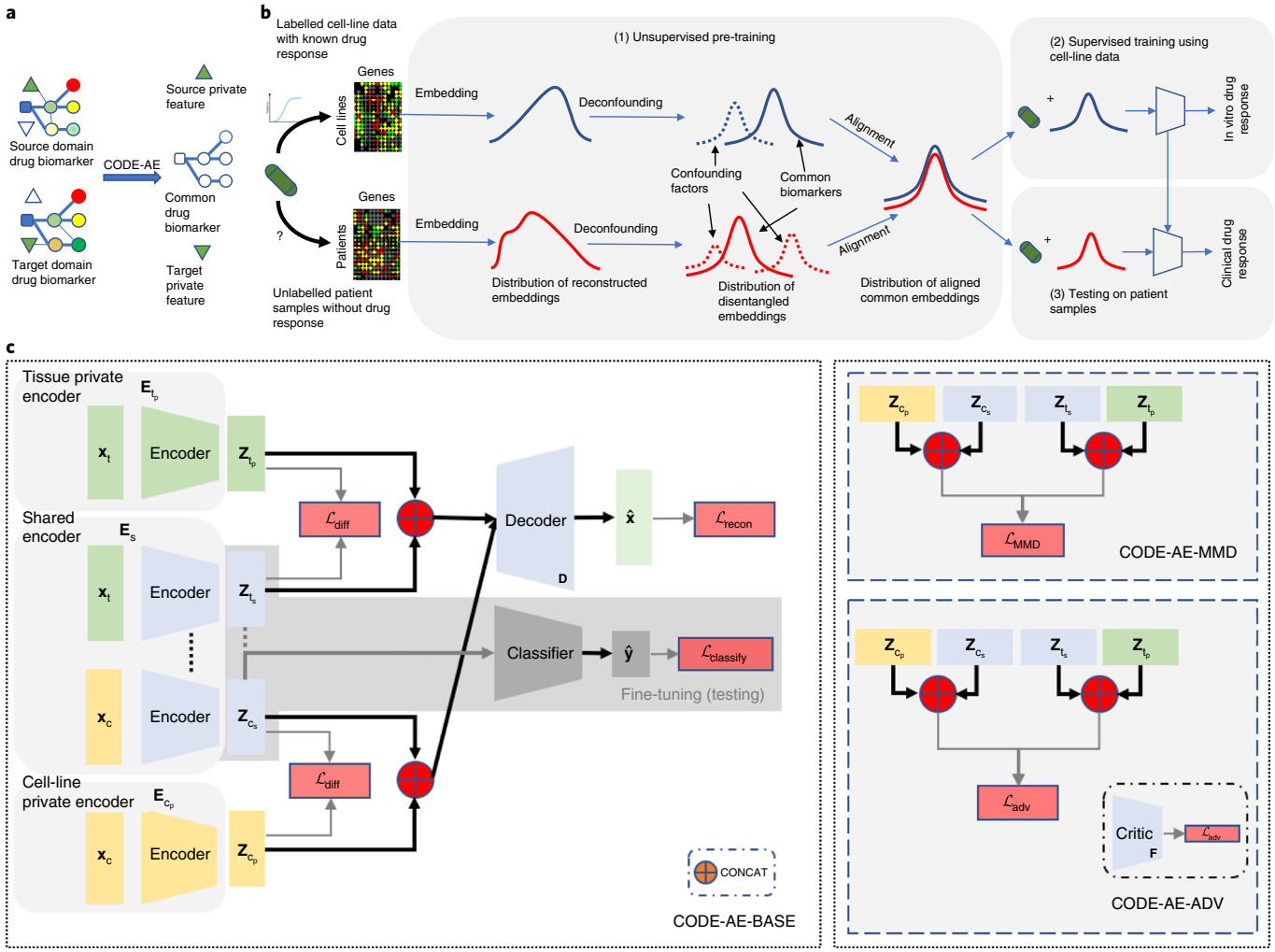


Fig. 1 | Overview of CODE-AE. a, Rationale of CODE-AE. Mechanistically, drug response biomarkers are a drug target and the downstream and upstream genes regulated by them can be characterized by observed gene expressions. The common biomarkers between the two data domains are their shared drug regulatory network. Private features for each domain are unique features that may contribute to the drug response for each domain but are not shared between domains. **b**, Illustration of CODE-AE. Given labelled cell-line drug response data, the aim of CODE-AE is to predict individual patient clinical responses to drugs that have been tested in the cell-line model but have never been tested in the patient. Conceptually, CODE-AE consists of three steps: pre-training, fine-tuning and inference. (1) During the pre-training stage, unlabelled gene expression profiles of both cell lines and patients are mapped into an embedding space using unsupervised learning. Biological confounding factors are disentangled from intrinsic biomarkers in the embedding. The distribution of embeddings of patients is aligned with that of cell lines to remove systems biases (for example, batch effect). (2) During the fine-tuning stage, a supervised model is appended to the pre-trained CODE-AE and trained based on the deconfounded and aligned embedding of cell lines using labelled cell-line drug response data. (3) During the inference stage, the deconfounded and aligned embedding of a patient is first obtained from the pre-trained CODE-AE, then used to predict the patient's response to a drug by the model trained from the fine-tuning in stage 2. **c**, Architecture of CODE-AE. Left: CODE-AE base architecture. A layer-tying shared encoder $\mathbf{E}_s$ learns to map both cell-line and tissue samples to extract common intrinsic biological signals. Private encoders $\mathbf{E}_p$ learn to represent cell-line/tissue context-specific information as private embeddings. A shared decoder **D** reconstructs the input samples through the concatenation of private and shared embeddings and the reconstruction quality is measured with $\mathcal{L}_{recon}$. CONCAT: Concatenation. The private and shared embeddings are pushed apart through soft subspace orthogonality loss $\mathcal{L}_{diff}$. The shared encoder $\mathbf{E}_s$ appended with an additional classifier network will be trained during fine-tuning and perform inference during the testing phase. Top right: CODE-AE-MMD. A variation of CODE-AE-BASE where the concatenation of private and shared embeddings are kept similar via optimizing $\mathcal{L}_{MMD}$. Bottom right: CODE-AE-ADV. A variation of CODE-AE-BASE where the concatenation of private and shared embeddings are kept similar via optimizing $\mathcal{L}_{adv}$. $\mathcal{L}_{adv}$ is in the form of min–max optimization between a critic network **F** and encoder components.

patient samples and in vitro Cancer Cell Line Encyclopedia (CCLE) cell-line samples. To inspect how well the embeddings of cell-line data and patient samples are aligned, we generated t-distributed stochastic neighbor embedding (tSNE) plots to visualize their embeddings, as shown in Fig. 2a–c. The embeddings of TCGA and CCLE samples from CODE-AE-ADV largely overlap in tSNE manifolds. This indicates that CODE-AE-ADV is effective in aligning cell lines and patients' representations. As a comparison, the low-dimensional representations of CCLE and TCGA data are clearly separated when using original gene expression profiles or a vanilla autoencoder. Thus, CODE-AE-ADV is more effective in addressing the OOD problems than the embedding algorithms that are used by state-of-the-art methods such as VAEN[9].

CODE-AE is successful in deconfounding biological variables

To evaluate whether CODE-AE-ADV can generate transferable embedding through deconfounding uninteresting confounders while preserving true biological signals present in expression data, we selected the gene expression datasets used in ADAE[7] that represent the state of the art for deconfounding biological variables and performed the same evaluation process. The question to be answered here is whether we can

https://doi.org/10.1038/s42256-022-00541-0

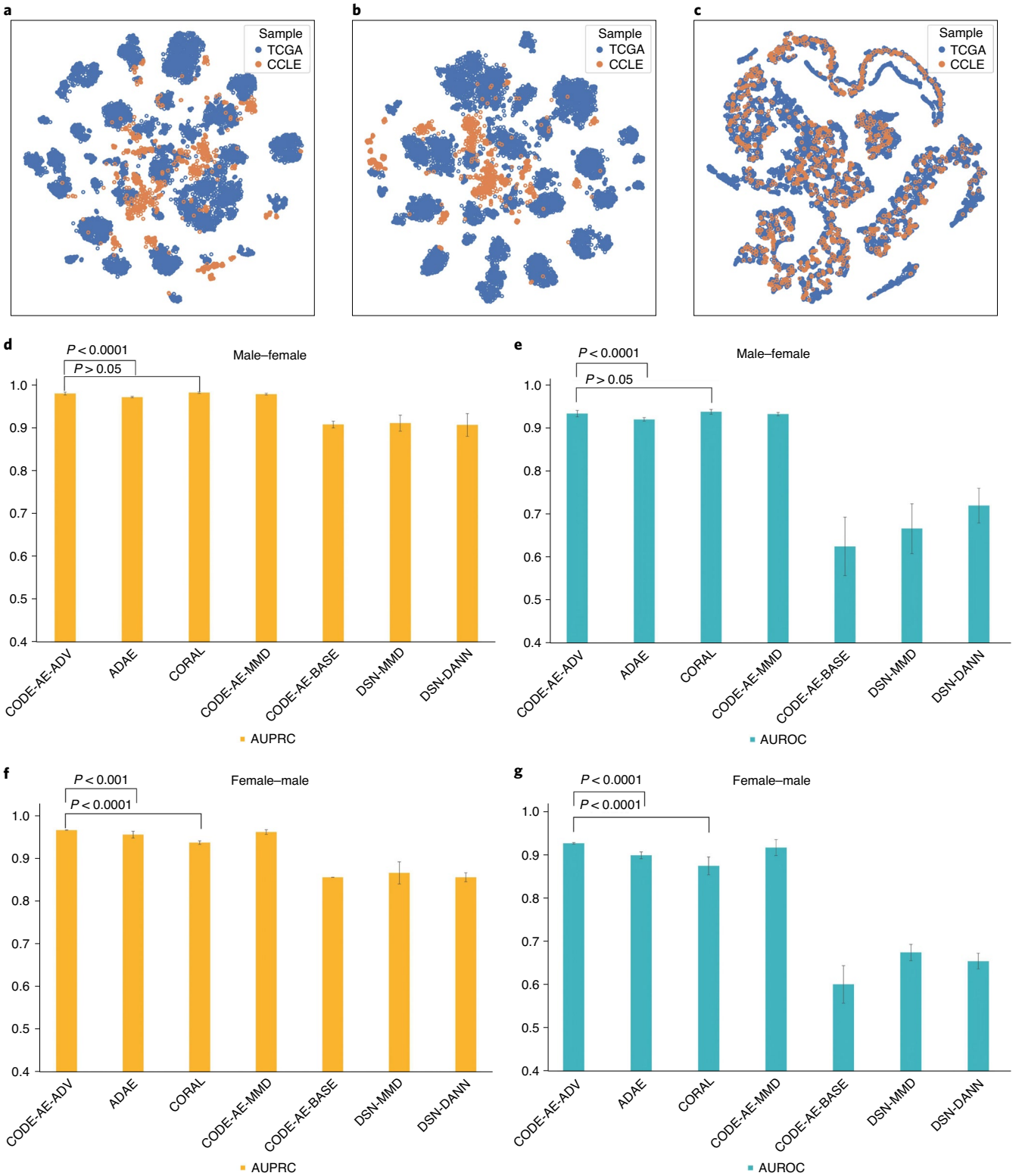


Fig. 2 | Evaluation of CODE-AE-ADV. a–**c**, tSNE plots of embeddings: original expression (**a**), embeddings generated by standard autoencoder (**b**) and embeddings generated by CODE-AE-ADV (**c**). **d**–**g**, Performance comparison on cancer subtype prediction with sex as a confounding factor. Error bars represent the standard deviation of cross-validations, and P values present the statistical significance of difference between the two evaluated models. Male–female (**d**,**e**): models were trained using only male samples and evaluated using female samples. Female–male (**f**,**g**): models were trained using only female samples and evaluated using male samples. The performance was evaluated using both AUPRC (**d**,**f**) and AUROC (**e**,**g**).

use a model trained purely from female data to classify cancer sub-types for males or vice versa by removing the sex confounder. Specifically, we chose the TCGA brain cancer expression dataset with sex information as confounding factors and brain cancer subtype classification as targeted tasks (see Methods for the details of training procedure). Using the model built from female data to predict male cancer sub-types,

https://doi.org/10.1038/s42256-022-00541-0

Table 1 | Average ranks of different methods on PDTC test dataset and patient chemotherapy response prediction

	Method	PDTC rank (average)		Chemotherapy prediction rank (average)	
		AUROC	AUPRC	AUROC	AUPRC
With pre-training	**CODE-AE-ADV**	**2.20±1.48**	**2.34±1.64**	**1.29±0.49**	**1.57±0.98**
	DSN-MMD	6.14±3.57	9.42±3.33	5.29±3.86	6.71±3.99
	DSN-DANN	8.04±2.82	9.28±2.77	5.71±1.80	6.29±2.21
	AE	9.94±3.45	10.70±2.78	6.43±2.64	8.86±3.02
	DAE	9.94±2.91	10.40±2.73	7.29±3.59	8.00±3.27
	VAE	4.52±2.77	2.64±1.54	9.57±2.15	7.00±3.79
	CORAL	9.10±3.05	10.16±3.05	7.57±3.21	7.86±3.02
	ADAE	4.30±2.06	4.08±1.83	5.00±2.94	2.43±1.57
	VAEN[1]	6.32±3.07	2.52±1.52	8.00±4.49	5.86±3.58
	Velodrome[1]	7.56±3.95	8.44±3.47	7.42±2.23	10.71±5.96
	Celligner+Elastic Net	12.38±4.07	11.14±4.37	13.00±4.58	12.00±4.36
	COXEN[2]+Elastic Net	8.08±5.50	8.72±4.79	11.57±7.09	10.71±5.96
	COXEN[2]+Random Forest	7.92±5.26	8.50±4.26	13.71±2.75	12.57±3.74
Without pre-training	TCRP	12.66±3.13	13.06±3.10	11.14±3.18	12.14±2.54
	MLP	14.54±2.20	14.98±1.65	11.00±3.37	11.14±2.54
	Elastic Net	13.86±2.96	11.10±4.02	14.29±4.64	14.14±4.26
	Random Forest	14.66±2.14	15.12±2.22	14.29±0.76	13.43±2.07

The best performer and the second-best performer are highlighted in bold and underlined, respectively. Three methods, VAEN, Velodrome and COXEN, are specially designed for the transfer learning of drug response predictions using gene expression data. Superscript 1 and 2 indicate that the method is to only align two datasets and only extract common features between two datasets, respectively

CODE-AE-ADV significantly outperforms ADAE, the second-best performer measured by both the area under the receiver operating curve (AUROC) and the area under the precision–recall curve (AUPRC). When applying the model trained from male data to predict female cancer sub-types, the performance of CODE-AE-ADV is slightly worse than CORAL, but the difference is not statistically significant. Both CODE-AE-ADV and CORAL significantly outperform the state-of-the-art deconfounding method ADAE ($P \leq 0.05$). In addition, the adversarial loss is more effective than MMD loss. Overall, CODE-AE-ADV performs the best for removing biological confounders.

CODE-AE improves ex vivo drug response predictions

Given that CODE-AE-ADV could disentangle confounding factors and extract meaningful biological signals in embedding space, we next evaluated whether CODE-AE-ADV could predict patient-specific responses to a new compound using the model trained from cell-line screens. We first compared CODE-AE-ADV with baseline models using ex vivo drug response data from PDTC[10]. As shown in Table 1, CODE-AE-ADV is overall the best performer for the PDTC test dataset when evaluated by the average of predicted ranks in terms of both AUROC and AUPRC. When evaluated by AUROC, CODE-AE significantly outperformed the second-best performer ADAE that was specially designed to remove confounders[7] and VAEN[9] that represented the state of the art for predicting clinical drug responses from in vitro screens. When evaluated by AUPRC, CODE-AE-ADV is still significantly better than ADAE but only slightly better than VAEN. Among the domain adaptation methods DSN, CORAL and Domain-Adversarial training of Neural Networks (DANN), DSN performs the best. DSN uses the same idea as CODE-AE to separate common and unique features between two domains. This observation suggests the importance of disentangling shared and private information between cell lines and patient samples. It is noted that all models used the exact same data and training procedure for the pre-training (detailed in Methods).

Figure 3 shows the drug-wise performance of each algorithm, as measured by the AUROC and AUPRC of predicted drug responses for each drug across all mice. CODE-AE-ADV performed the best for three drugs BX795, obatoclax mesylate and axitinib with the AUROC above 0.85. The AUPRC of CODE-AE-ADV was quite stable. It was above 0.75 for most drugs, demonstrating the robustness of CODE-AE-ADV. Although the average AUPRC rank of the second-best performer VAEN is comparable to that of CODE-AE-ADV, the best-ranked drug by VAEN was only around half as much as that by CODE-AE-ADV (12 versus 22).

CODE-AE improves clinical drug response predictions

We further evaluated the performance of CODE-AE-ADV for predicting clinical responses to single chemotherapy in two aspects: either a lack of reduction in tumour size following chemotherapy (marked as a diagnosis in Fig. 4) or the occurrence of clinical relapse after an initial 'positive response to treatment'[16] as detailed in Methods. Again, we used the AUROC and AUPRC of predicted drug responses for each drug across patients to evaluate the performance (Fig. 4). Consistent with the results from the PDTC dataset, CODE-AE-ADV consistently outperformed baseline models in most cases when evaluated by AUROC. CODE-AE-ADV significantly outperforms the second-ranked ADAE in most cases. This observation further supports that CODE-AE-ADV can enhance the signal-to-noise ratio in biomarker identification because the major difference between CODE-AE-ADV and ADAE is to disentangle shared and private embeddings between cell lines and patient tissues. When the performance is evaluated by AUPRC, the overall performance ranking is similar to that based on AUROC (Table 1). Side-by-side comparisons of AUPRCs in Fig. 4 show that the AURPCs

https://doi.org/10.1038/s42256-022-00541-0

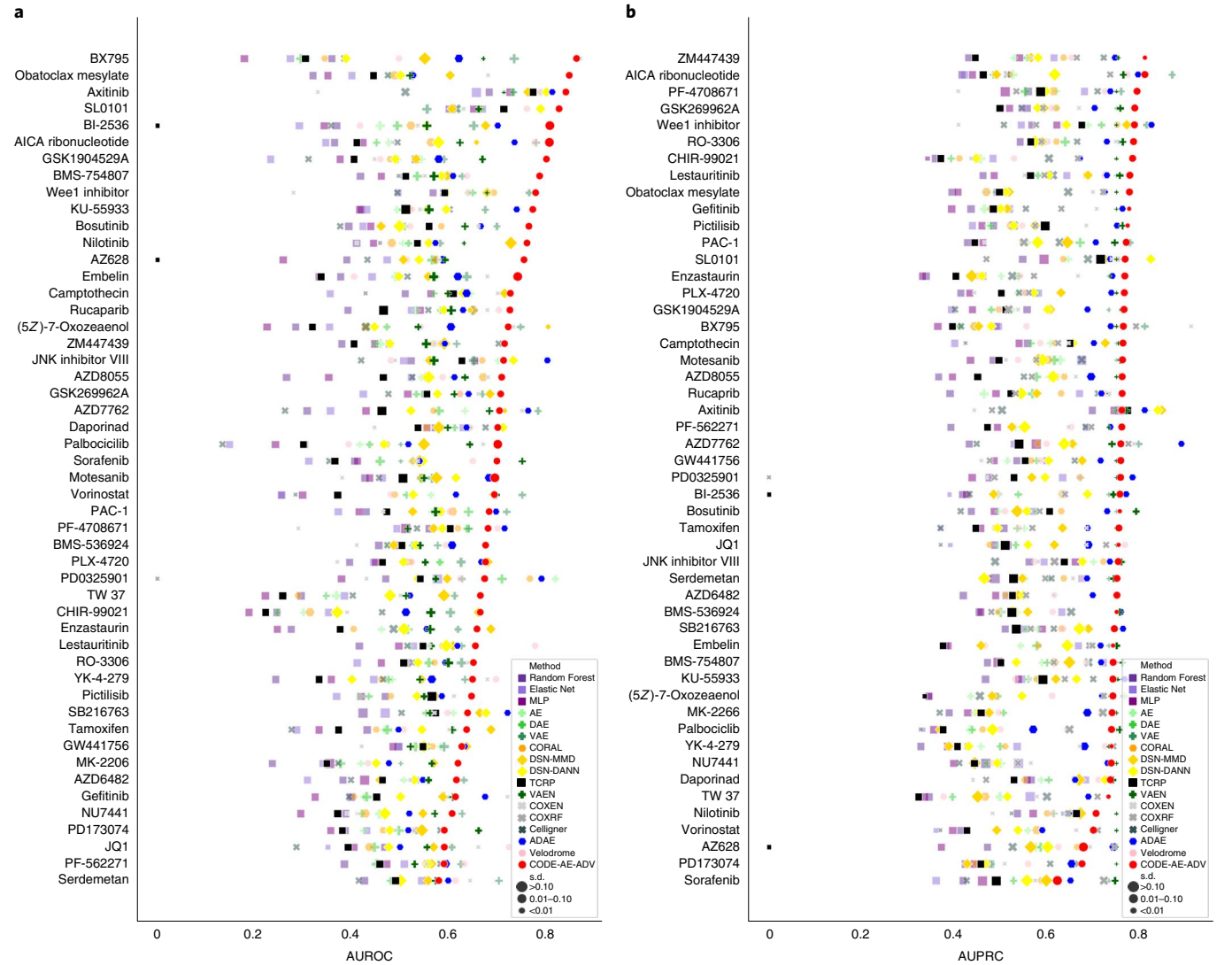


Fig. 3 | Drug-wise performance of algorithms. a,**b**, Performance comparison of PDTC drug response classification as measured by AUROC (**a**) and AUPRC (**b**). COXRF: COXEN + Random Forest.

of ADAE are slightly higher than that of CODE-AE-ADV in two cases ($P \geq 0.05$), respectively. However, the average ranks of ADAE are significantly worse than that of CODE-AE-ADV.

Application of CODE-AE-ADV to personalized medicine

To further validate CODE-AE-ADV with patient data and demonstrate its utility in personalized medicine as well as to comprehend whether the common embedding learns biological meaningful signals, we applied CODE-AE-ADV (per drug) trained with CCLE data to screen 59 drugs for 9,808 patients with cancer from TCGA. It is noted that the target variable used for the model training was the cell viability in cell lines, which does not directly correspond to clinical drug responses in humans. Thus, the absolute value of the predicted score for a patient is less meaningful. Instead, we ranked predicted scores for each drug or normalized them as *z*-scores. If the ranking was used, the patients with the top 5% and bottom 5% ranked scores are classified as drug-responsive or drug-resistant, respectively. In the case of *z*-score, a threshold was selected to separate drug-responsive and drug-resistant patients. Our major findings are summarized below.

We first verify our predictions by inspecting the association of our predicted drug response with the gene expression values of drug targets. If the predicted patient response to the targeted therapy is correlated with the drug target, it provides the validation of our prediction. We found that 47 out of 50 targeted therapies are statistically significant (false discovery rate ≤0.05) associated with the differential target gene expression between drug-sensitive and drug-resistant patients (Supplementary Table 2). This indicates that CODE-AE-ADV could capture the drug mode of action.

We applied spectral biclustering[17] to divide 9,808 patients into 100 clusters and 59 drugs into 30 clusters from the predicted drug response matrix. In this way, patients with similar drug response profiles were grouped together. The clustering result for lung squamous cell carcinoma (LSCC), a type of non-small cell lung cancer (NSCLC), is shown in Fig. 5a; 498 LSCC patients were clustered into 45 groups (Fig. 5b). The number of patients in each group ranged from 1 (0.2%) to 60 (12.0%).

Among 59 drugs tested, the top 3 most responsive drugs to the LSCC are gefitinib, 5-aminoimidazole-4-carboxamide ribonucleotide (AICAR) and gemcitabine. Gefitinib is an epidermal growth factor receptor (EGFR) tyrosine kinase inhibitor for the first-line treatment of NSCLC[18]. AICAR is an AMP-activated protein kinase (AMPK) agonist that can block the growth of cancer cells harbouring the activated EGFR mutant[19,20]. Gemcitabine, a chemotherapy, has long been used as one of the most effective treatments for NSCLC that may not harbour the EGFR mutation[21,22]. Consistent with their drug mode of action, the CODE-AE

https://doi.org/10.1038/s42256-022-00541-0

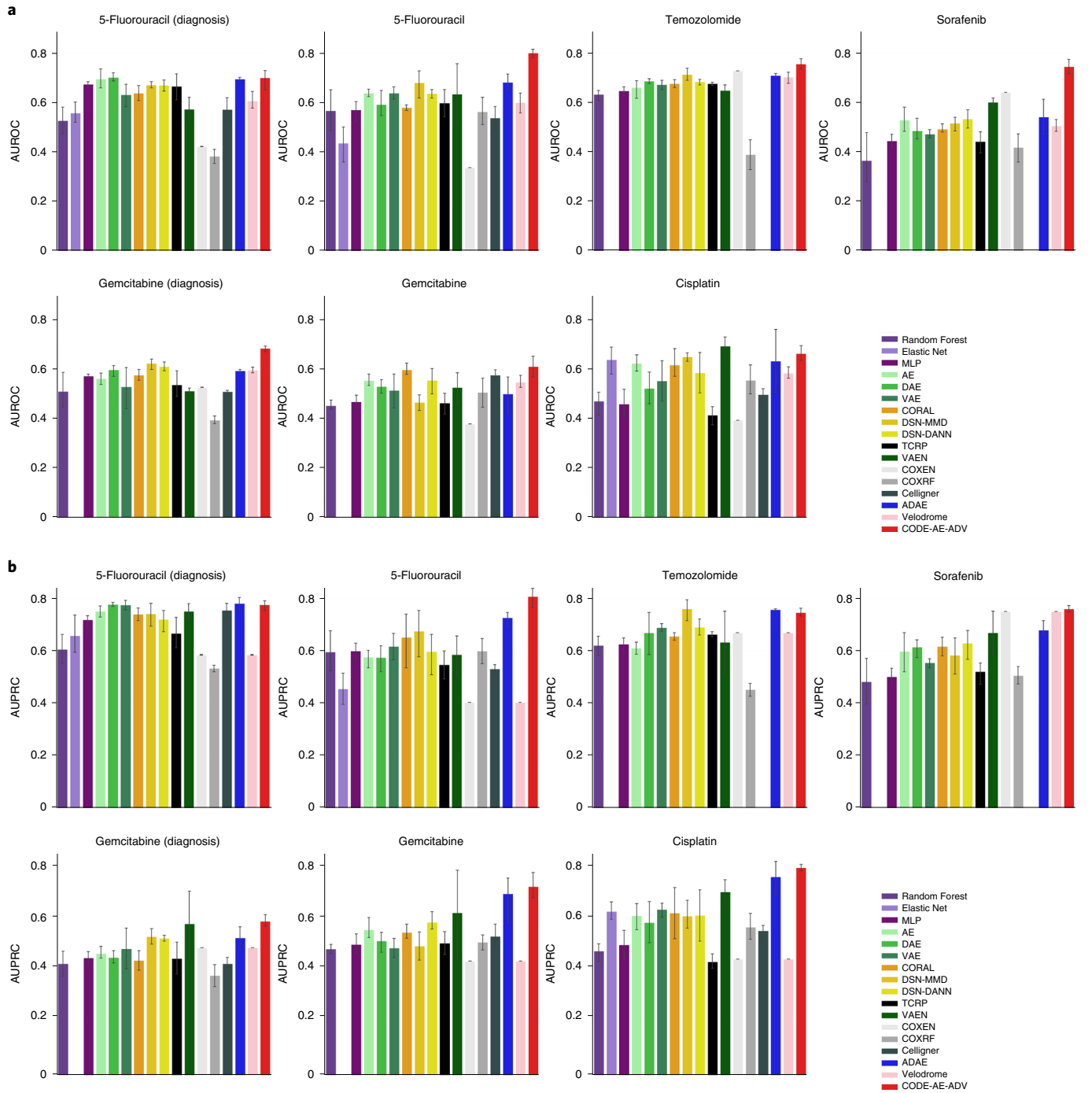
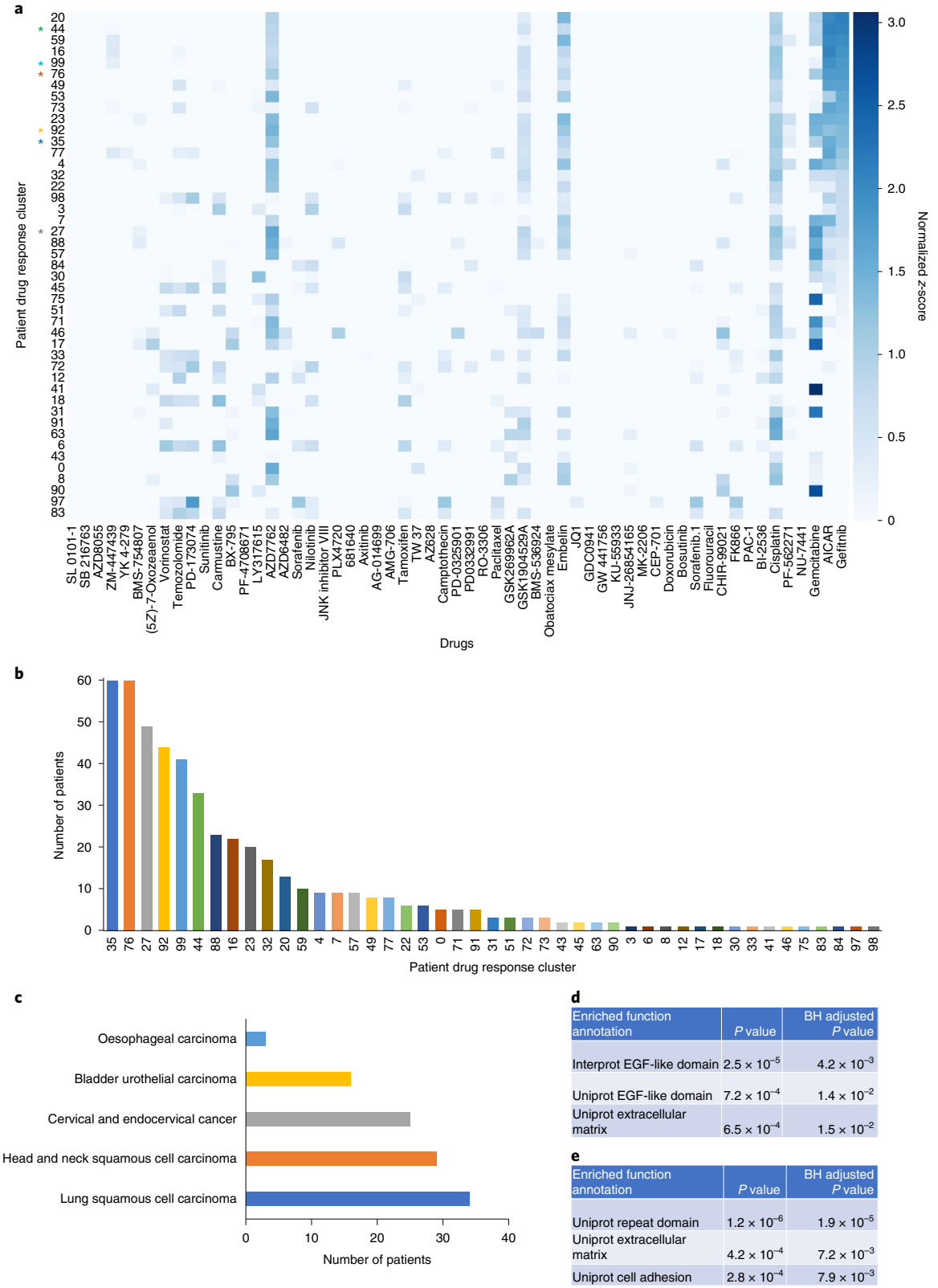


Fig. 4 | Performance of CODE-AE-ADV for predicting clinical responses to single chemotherapy. a,**b**, Performance comparison of patient chemotherapy response prediction as measured by AUROC (**a**) and AUPRC (**b**). Error bars represent the standard deviation of cross-validations. Two clinical endpoints are evaluated: reduction in tumour size following chemotherapy (charts marked as diagnosis) and occurrence of clinical relapse (remaining charts). The model trained using features from Celligner alignment could not make meaningful predictions for two drugs (temozolomide and sorafenib). They were not included in the chart.

predicted patient's drug response profiles of gefitinib and AICAR are similar, but different from that of chemotherapy gemcitabine for cancer cells not harbouring the EGFR mutations (Fig. 5a).

We further inspected the patient cluster that was predicted to be the most responsive to gefitinib with a number of patients with LSCC larger than 30. Besides LSCC, four other cancer types—head and neck squamous cell carcinoma (HNSCC), cervical and endocervical cancer, bladder urothelial carcinoma and oesophageal carcinoma—were included in this cluster (Fig. 5c). Several clinical studies have been carried out and are undergoing for the use of gefitinib in the treatment of HNSCC (for example, https://clinicaltrials.gov/ct2/show/NCT00024089) owing to the fact that EGFR is over-expressed in over 90% of patients with HNSCC patients[23–25]. Similarly, clinical trials for gefitinib to treat cervical cancer (https://clinicaltrials.gov/ct2/show/NCT00049556), bladder urothelial carcinoma (https://clinicaltrials.gov/ct2/show/NCT00246974) and oesophageal carcinoma (https://clinicaltrials.gov/ct2/show/NCT01243398) are undergoing, as certain patients diagnosed with these cancers were observed to respond to

https://doi.org/10.1038/s42256-022-00541-0



d

Enriched function annotation	*P* value	BH adjusted *P* value
Interprot EGF-like domain	2.5×10^{-5}	4.2×10^{-3}
Uniprot EGF-like domain	7.2×10^{-4}	1.4×10^{-2}
Uniprot extracellular matrix	6.5×10^{-4}	1.5×10^{-2}

e

Enriched function annotation	*P* value	BH adjusted *P* value
Uniprot repeat domain	1.2×10^{-6}	1.9×10^{-5}
Uniprot extracellular matrix	4.2×10^{-4}	7.2×10^{-3}
Uniprot cell adhesion	2.8×10^{-4}	7.9×10^{-3}

Fig. 5 | Predicted 498 LSCC patient responses to 59 drugs. Patients are clustered into 45 patient clusters. A drug response score for each cluster is calculated by averaging the individual patient response score. Then the averaged cluster scores are normalized across all clusters and drugs, and a normalized *z*-score is calculated. The higher the *z*-score is, the more sensitive patients are to the drug. **a**, Heat map of normalized *z*-score of drug responses across 59 drugs and 45 patient clusters. For clarity, all clusters with a *z*-score less than 0 are marked as 0. The six largest patient clusters are indicated with coloured stars. **b**, The distribution of LSCC patients in 45 patient clusters. **c**, The distribution of patients with different primary tumours in the patient cluster 44. **d**, Enriched functional annotations of top 50 genes with the highest differential somatic mutation rates between patients resistant and sensitive to chemotherapy AICAR. BH: Benjamini-Hochberg. **e**, Enriched functional annotations of top 50 genes with the highest differential somatic mutation rates between patients resistant and sensitive to chemotherapy gemcitabine.

EGFR inhibitors[26–28]. Thus, the predictions from CODE-AE are largely consistent with clinical observations.

The patient response to gefitinib can be explained by the differential gene expression of EGFR between drug-responsive and drug-resistant patients (Supplementary Table 2). It is interesting to know whether the mutation of certain genes is responsible for the patients' resistance to AICAR and gemcitabine chemotherapies. We ranked genes based on their mutation rate differences between patients that are sensitive and resistant to the drugs (Supplementary Data 1 and 2). *TP53* and *FBN1* are the highest-ranked genes for AICAR and gemcitabine, respectively. It is well known that *TP53* contributes to anticancer drug sensitivity[29]. In particular, *TP53* plays a role in the drug resistance of EGFR inhibitors in NSCLC[30]. A recent study provided evidence that *FBN1* prompts chemotherapy resistance[31]. Furthermore, Fig. 5d,e lists the statistically significantly over-represented function annotations of top 50 ranked genes for AICAR and gemcitabine, respectively. The mutated genes with the EGF-like domain are accountable for the patient's resistance to AICAR. For gemcitabine, the drug resistance is mainly due to the mutations in the genes with a repeat domain or involved in the cell adhesion. Several studies have shown that the dysregulation of repeat domains is responsible for drug resistance in lung cancers[32,33]. It is well known that cell adhesion is a key determinant in cancer drug resistance[34]. The mutation in the extracellular matrix has an impact on the patient's resistance to both drugs and affects the efficacy of chemotherapies[35]. In summary, the mutation biomarkers for drug sensitivity based on the predicted patient drug responses are supported by existing experimental evidences.

Conclusion

In this Article, we have introduced a new transfer learning framework CODE-AE to predict individual patient drug responses from a neural network model trained on cell-line data. Extensive benchmark studies demonstrate the advantage of CODE-AE over the state of the arts in terms of both accuracy and robustness. When CODE-AE is applied to predict drug responses for patients in TCGA, the predictions are largely consistent with existing clinical observations. CODE-AE could be further improved in several directions. In principle, integrating multiple omics data may benefit drug response predictions[36,37]. We performed a preliminary study on encoding propagated somatic mutations on a protein–protein interaction network using CODE-AE (Supplementary Figs. 2 and 3). Simple integration of CODE-AE embeddings of gene expressions and mutations did not improve the performance, as shown in Supplementary Table 3. More sophisticated methods such as the framework of cross-level information transmission[38] may be needed. The interpretability of CODE-AE can be improved, for example, by incorporating Gene Ontology or biological pathway information[39]. Finally, estimating the prediction uncertainty for each new case, especially those that are remote to the labelled data in the embedding space, may further improve the performance of CODE-AE and is critical for clinical applications. Although CODE-AE is applied to only precision oncology here, it can be a general framework for other transfer learning tasks where two data modalities have shared and unique features.

Methods

CODE-AE

We proposed CODE-AE to generate biologically informative gene expression embeddings to transfer knowledge from in vitro data into patient samples. CODE-AE employed the standard autoencoder as the backbone to leverage the unlabelled gene expression datasets. Inspired by the work on factorized latent space[40] and DSN[15], we encoded the samples (from cell lines or tumour tissues) into two orthogonal embeddings, namely private embeddings and shared embeddings. The first one is designed to separate the context-specific signals that overwhelm the common biomarkers. The latter contains the deconfounded common intrinsic biological signals used to transfer knowledge across cell lines and tissues.

CODE-AE-BASE. As shown in Fig. 1, the CODE-AE takes expression vectors from in vitro cell lines and patient tumour tissue samples as input. Let $X_t = \{x_t^{(i)}\}_{i=1}^{N_t}$ and $X_c = \{x_c^{(i)}\}_{i=1}^{N_c}$ represent the unlabelled dataset of N_t patient tumour tissue samples and N_c in vitro cancer cell-line samples, respectively. Each sample x will be encoded into two separate embeddings through its corresponding cell-line or tissue private encoder $\mathbf{E}_{\cdot_p}$ and also the weight-sharing encoder $\mathbf{E}_s$. The concatenation of these two embeddings of each sample is expected to be able to reconstruct the original gene expression vector x through a shared decoder $\mathbf{D}$, and the reconstruction is done as

$$\widehat{\mathbf{x}_{\cdot}^{(i)}} = D(\mathbf{E}_s(\mathbf{x}_{\cdot}^{(i)}) \bigoplus \mathbf{E}_{\cdot_p}(\mathbf{x}_{\cdot}^{(i)})) \tag{1}$$

where $\mathbf{x}_{\cdot}^{(i)}$ represents the input gene expression profile and $\widehat{\mathbf{x}_{\cdot}^{(i)}}$ is the corresponding reconstructed input sample through the autoencoder component. $\bigoplus$ stands for the vector concatenation operation. We measure the quality of autoencoder reconstruction through the mean squared error between the original samples and the reconstruction output as below

$$\mathcal{L}_{\text{recon}} = \frac{1}{N_c}\sum_{i=1}^{N_c}\left\|\mathbf{x}_c^{(i)} - \widehat{\mathbf{x}_c^{(i)}}\right\|_2^2 + \frac{1}{N_t}\sum_{i=1}^{N_t}\left\|\mathbf{x}_t^{(i)} - \widehat{\mathbf{x}_t^{(i)}}\right\|_2^2 \tag{2}$$

In our formulation, we factorized each sample's latent space into two different subspaces to capture both domain-specific and common information separately. To minimize the redundancy between the factorized latent spaces, we included an additional penalty term, $\mathcal{L}_{\text{diff}}$ in the form of orthogonality constraint. The difference loss $\mathcal{L}_{\text{diff}}$ is applied to both cell-line and tissue samples and encourages the shared and private encoder to encode different aspects of the inputs. We define the loss via soft subspace orthogonality constraint as below

$$\mathcal{L}_{\text{diff}} = \left\|\mathbf{Z}_{c_s}^T\mathbf{Z}_{c_p}\right\|_F^2 + \left\|\mathbf{Z}_{t_s}^T\mathbf{Z}_{t_p}\right\|_F^2 \tag{3}$$

where $\mathbf{Z}_{c_s}$ and $\mathbf{Z}_{t_s}$ are embedding matrices whose rows are the shared embeddings from the corresponding common encoders for cell-line and tissue samples, respectively, while $\mathbf{Z}_{c_p}$ and $\mathbf{Z}_{t_p}$ are embedding matrices whose rows are the private embedding from corresponding private encoders for cell-line and tissue samples, respectively. The superscript T stands for the transpose of matrix. Each column in the matrix corresponds to a sample. It is obvious that $\mathcal{L}_{\text{diff}}$ tends to push the embeddings to meaningless all-zero-valued vectors. To avoid such a scenario, we append an additional instance normalization layer after the output layer of each encoder to avoid embeddings with minimal norm. Lastly, the loss for CODE-AE-BASE is defined with the weighted combination between $\mathcal{L}_{\text{recon}}$ and $\mathcal{L}_{\text{diff}}$ as below

$$\mathcal{L}_{\text{code-ae-base}} = \mathcal{L}_{\text{recon}} + \alpha\mathcal{L}_{\text{diff}} \tag{4}$$

where α is the embedding difference loss coefficient with a default value of 1.0 and not optimized.

CODE-AE variants. With CODE-AE-BASE, we could split cell lines or tissue samples' inherent information into private and shared streams. However, in our baseline experiments, we often found that it was sub-optimal or demonstrated varied performance. Thus, we proposed two variants that showed better and generally more stable performance. Under the CODE-AE framework, for each input sample, CODE-AE factorized it into two orthogonal embeddings. The concatenation of these two embeddings is considered as the new

representation of the original input. Given that all samples in our consideration are gene expression profiles regardless of cell line or patient, we assumed that the new representation of original input in the factorized latent space is close to each in terms of distributional differences. Hence, we incorporated an additional feature alignment component into the CODE-AE-BASE framework. Specifically, the distributional difference of the concatenated representation of private and shared embeddings from both cell-line and tumour tissue samples is minimized via the following two approaches.

CODE-AE-MMD. The first variant, named CODE-AE-MMD, utilized the well-known MMD[41] as the distance measurement between the latent representation of cell-line and tissue samples. MMD loss[41] is a kernel-based distance function between samples from two distributions. In particular, we used an approximate version of the exact MMD loss in CODE-AE-MMD as below

$$\mathcal{L}_{\mathrm{MMD}}(\mathbf{Z}_c, \mathbf{Z}_t) = \frac{1}{N^2}\sum_{i,j=0}^{N} \kappa(\mathbf{z}_c^{(i)}, \mathbf{z}_c^{(j)}) + \frac{1}{N^2}\sum_{i,j=0}^{N} \kappa(\mathbf{z}_t^{(i)}, \mathbf{z}_t^{(j)}) - \frac{2}{N^2}\sum_{i,j=0}^{N} \kappa(\mathbf{z}_c^{(i)}, \mathbf{z}_t^{(j)}) \tag{5}$$

where $\mathbf{Z}_c$ and $\mathbf{Z}_t$ are embedding matrices for cell line and tissue samples respectively, whose rows are the concatenations of each sample's private and shared embedding. $\mathbf{z}_{\cdot}^{(i)}$ and $\mathbf{z}_{\cdot}^{(j)}$ are the ith or jth samples' corresponding embedding vectors. In practice, N will be the batch size. Accordingly, the loss of CODE-AE-MMD is given as below

$$\mathcal{L}_{\mathrm{code-ae-mmd}} = \mathcal{L}_{\mathrm{code-ae-base}} + \beta\mathcal{L}_{\mathrm{MMD}} \tag{6}$$

where β is the MMD loss coefficient with a default value of 1.0 and not optimized.

CODE-AE-ADV. The second variant, CODE-AE-ADV, employed adversarial training to push the representations of cell-line and tissue samples to be similar to each other. Specifically, we appended a critic network F that scores representations with the objective that consistently gives higher scores for representations of cancer cell-line samples. The encoders for tissue samples are given an additional objective to generate the embedding that could fool the critic network to produce high scores. In this manner, critic network and tissue sample encoders will play a min–max game in the form of an alternative training schedule, which is adopted by Wasserstein generative adversarial networks (WGANs)[42]. To avoid unstable training commonly existing in alternative training schedules, instead of a standard WGAN[42], we used a WGAN with a gradient penalty[43]. Its affiliated loss terms are defined as below

$$\mathcal{L}_{\mathrm{adv}}: \begin{cases} \mathcal{L}_{\mathrm{critic}} = \frac{1}{N_t}\sum_{i=1}^{N_t} F(\mathbf{z}_t^{(i)}) - \frac{1}{N_c}\sum_{i=1}^{N_c} F(\mathbf{z}_c^{(i)}) + \lambda\left(\left\|\nabla_{\tilde{\mathbf{z}}} F(\tilde{\mathbf{z}})\right\|_2 - 1\right)^2 \\ \mathcal{L}_{\mathrm{gen}} = -\frac{1}{N_t}\sum_{i=1}^{N_t} F(\mathbf{z}_t^{(i)}) \end{cases} \tag{7}$$

where $\mathbf{z}_{\cdot} = \mathbf{z}_{\cdot_s} \bigoplus \mathbf{z}_{\cdot_p}$ stands for new representation of input and $\tilde{\mathbf{z}} = \epsilon\mathbf{z}_c + (1-\epsilon)\mathbf{z}_t$ and $\epsilon \approx \mathbf{U}(0, 1)$, the standard uniform distribution exclusive between 0 and 1. A detailed CODE-AE-ADV learning procedure can be found in Procedure 1.

After the encoder training with unlabelled data as mentioned above, the shared encoder $\mathbf{E}_s$ could be used to directly generate the deconfounded biological meaningful embedding vectors or append a neural network module for specific downstream tasks. In the latter case, it is widely known as the pre-training fine-tuning scheme that has gained massive popularity in recent natural language processing (NLP) applications[44,45]. In our drug response prediction benchmark experiments mentioned above, we have followed this approach and adopted strategies such as gradual unfreezing and decayed learning rate scheduler to improve task-specific performance. Specifically, during the pre-training phase, we trained CODE-AE with unlabelled gene expressions from two groups of samples: cancer cell lines and patient tissues (TCGA) to extract common signals and disentangle confounders between them. Then in the fine-tuning phase, we kept only the pre-trained shared encoder and appended it to a multi-layer neural network and fine-tuned the model with labelled cell-line samples to predict their response to a drug, specifically, binary labels based on AUC as the target label. Finally, during inference, the complete trained model from the fine-tuning step was used to directly generate drug response prediction for in vivo (patients) or ex vivo (PDTC) samples given a specific drug. It is noted that the task labels during the inference stage are different from the ones used in the fine-tuning stage, more details can be found below.

Procedure 1 CODE-AE-ADV training

Input: $\{\mathbf{x}_c^{(i)}\}_{i=1}^{N_c}, \{\mathbf{x}_t^{(i)}\}_{i=1}^{N_t}$
Require:
N, the batch size
λ, generator loss coefficient
n_w, number of warm-up epochs
n_t, number of training epochs
n_{critic}, number of steps per encoders update
for epoch = 1 to n_w **do**
  for t = 1 to $\frac{\min(N_c, N_t)}{N}$ **do**
    sample $\{\mathbf{x}_c\}$ of size N from $\{\mathbf{x}_c^{(i)}\}_{i=1}^{N_c}$ (without rep)
    sample $\{\mathbf{x}_t\}$ of size N from $\{\mathbf{x}_t^{(i)}\}_{i=1}^{N_t}$ (without rep)
    Update E_{t_p}, E_{c_p}, E_s, D with $\mathcal{L}_{\mathrm{code-ae-base}}$
  end for
end for
for epoch = 1 to n_t **do**
  for t = 1 to $\frac{\min(N_c, N_t)}{N}$ **do**
    sample $\{\mathbf{x}_c\}$ of size N from $\{\mathbf{x}_c^{(i)}\}_{i=1}^{N_c}$ (without rep)
    sample $\{\mathbf{x}_t\}$ of size N from $\{\mathbf{x}_t^{(i)}\}_{i=1}^{N_t}$ (without rep)
    Update F with $\mathcal{L}_{\mathrm{critic}}$
    if $t \,\%\, n_{\mathrm{critic}} == 0$ **then**
      Update E_{t_p}, E_{c_p}, E_s, D with $\mathcal{L}_{\mathrm{code-ae-base}} + \lambda\mathcal{L}_{\mathrm{gen}}$
    end if
  end for
end for

Experiments set-up

Baseline models. We compared CODE-AE with the following baseline models that include unlabelled pre-training: VAEN[9], standard AE[46], DAE[13] and VAE[12] as well as representative domain adaptation methods including Velodrome[5], Celligner[6], Deep CORAL[14] and DSN[15] of both MMD (DSN-MMD) and adversarial (DSN-DANN) training variants. Furthermore, we included a more recent ADAE[7] given its similar formation to DANN[47] and state-of-the-art performance in transcriptomics datasets. We also included COXEN[4], an advanced gene selection method to predict patient responses from cell-line screens. In addition, for CODE-AE variants, we also explored different configurations, such as with/without hidden layer normalization, performing a downstream task with concatenated representation, or shared representation in an ablation study with the PDTC test dataset.

For fair comparisons, all the encoders and decoders trained in the experiments share the same architecture. Specifically, the hidden representation is of dimension 128. The encoders and decoder are 2-layer neural network modules of dimensions (512, 256) and (256, 512), respectively, with the rectified linear activation function. Appended modules such as the critic network in CODE-AE-ADV and classifier network used for fine-tuning are 2-layer neural networks of dimension (64, 32) with rectified linear activation, have one output node with linear activation in the critic network, and sigmoid activation in

classifier networks. Further, the loss weight terms in CODE-AE-MMD and CODE-AE-ADV are all specified as 1.0.

Moreover, for models that do not include unlabelled pre-training, we compared CODE-AE with TCRP[8] as well as deep neural network (denoted as MLP), the Elastic Net classifier and Random Forest classifier. TCRP incorporates a model agnostic meta-learning technique and is one of the most successful methods for predicting individual patient drug responses from the cell-line data so far.

Datasets. Training dataset. The unlabelled pre-training (in vitro and in vivo) datasets used for encoder pre-training include cancer cell-line and patient tumour tissue gene expression profiles. Specifically, we collected 1,305 cancer cell-line samples with corresponding gene expression profiles from the DepMap portal[48] and 9,808 patient tumour tissue samples from TCGA[49]. All gene expression data are metricized by the standard transcripts per million bases for each gene, with additional log transformation. In addition, we used the gene selection method in ref. [50] to select the top 1,000 varied genes measured by the percentage of unique values in gene expression samples for cancer cell lines and tumour tissue samples separately. Then we combined the two sets of top 1,000 varied genes as the input features. There are a total of 1,426 genes in the feature set. We also explored other feature selection approaches including variance and mean absolute difference. We reported only the results based on the genes selected by the percentage of unique values because baseline methods in consideration showed overall better performance. We trained all baseline models only using gene expression data.

Besides gene expression features, we also evaluated CODE-AE when using somatic mutations as features from the samples mentioned above. For the somatic mutation data, we kept only non-silent genes and assembled as a binary-valued sparse vector. Furthermore, we applied pyNBS[51], a random walk with restart algorithm, to transform the binary-valued mutation profile into continuous-valued features by performing mutation score propagation on the search tool for the retrieval of interacting genes/proteins (STRING) gene–gene interaction network. The network-regularized mutation profile will not only reduce the sparsity of features but also significantly boost its prediction power[51]. Lastly, we only kept the genes selected in gene expression sets for the benchmark experiments.

The labelled fine-tuning (in vitro) dataset used for the fine-tuning phase was collected from Genomics of Drug Sensitivity in Cancer (GDSC)[52,53]. GDSC recorded the cellular growth responses of cancer cell lines against a panel of drugs as the area under the drug response curve (AUC), which is defined as the fraction of the total area under the drug response curve between the highest and lowest screening concentration in GDSC. For each drug of interest, we first identified all cell lines with corresponding drug sensitivity measured in the area under the drug response curve (AUC) and then split the drug sensitivity of these cancer cell lines into binary labels, namely responsive or non-responsive (resistant). The categorization threshold is selected as the average AUC value of all available cell-line drug sensitivity for drugs tested. The model was trained in a fashion of drug-wise. The rationale is that different drugs may have different drug modes of action.

Test dataset. We evaluated the performance of CODE-AE in the setting of zero-shot learning, that is, the unseen OOD data have never been used in training. It is a more difficult but more realistic scenario than the state-of-the-art method TCRP[8] in which a small set of OOD data was used during the training. Specifically, the predictive model for each drug of interest was learned with only the aforementioned in vitro dataset. While in testing time, we evaluated the model performance with the following ex vivo and in vivo labelled datasets that were not used in the training phase on the prediction task of drug response classification in pre-clinical and clinical scenarios, respectively.

For the pre-clinical (ex vivo) dataset, we used data from breast cancer PDTC[10] to evaluate the performance of drug response classification in a pre-clinical context. The previous study collected 83 human breast tumour biopsies and established human cell culture from these tumours with mice as intermediaries. Each of these human cell cultures was exposed to a list of drugs. From the list of drugs available in PDTC, we further selected 50 drugs with known protein targets for which cell-line responses had also been recorded in GDSC as drugs of interest. The drug sensitivity classification of each drug was considered as a separate learning task. Similar to the labelled GDSC dataset used during training, the PDTC responses were categorized into binary labels using PDTC AUCs, where the classification threshold is specified as the median AUC value of all available PDTC AUCs of each drug of interest.

For the clinical (in vivo) dataset, to evaluate the performance of drug response classification in a clinical context, we primarily consider a practical problem: predict chemotherapy resistance given gene expression profiles of patients while training the predictive model only using the gene expression profile of cancer cell lines.

Clinical chemotherapy resistance can be defined as either a lack of reduction in tumour size following chemotherapy or the occurrence of clinical relapse after an initial 'positive response to treatment'[16]. Hence, we extracted datasets to assess these two aspects. The patient clinical drug response was acquired from a recent work[50], where patients' clinical response records of two chemotherapy agents gemcitabine and fluorouracil from TCGA[49] were extracted. The patients were split into two groups: responders who had a partial or complete response and non-responders who had the progressive clinical disease or stable disease diagnosis. Only patients on single-drug therapy through the entire duration of treatment were retained in the study. Patients treated by drug combinations were excluded. It is noted that the gene expression profile of these TCGA patients could be used in the unsupervised pre-training of CODE-AE and other baseline models, but the drug response data were not used in the supervised fine-tuning.

In addition to using clinical diagnosis to indicate patients' drug responses towards a particular drug, we extracted patients' 'new tumour events days after treatment' from TCGA[49] as the standard to divide patients into responders and non-responders. The median number of days of new tumour events was used as the threshold. Similar to the above dataset from ref. [50], we only included patients on single-drug therapy through the entire treatment duration in this test dataset. For the list of drugs included in this test dataset, the drugs with more than 20 labelled samples are kept.

Training procedure. For models that include an unlabelled pre-training phase, we first pre-train them for N epochs using the same unlabelled samples from both cancer cell lines and tumour tissues. With parameter grid search, N is selected based on the downstream task performance (over validation set). The pre-trained encoders will then be appended with a classification module to perform the downstream drug sensitivity classification task in the following fine-tuning step. We adopted the early stopping with validation performance in the fine-tuning phase (training phase for the model without unlabelled pre-training). Specifically, the labelled cell-line samples were split into five stratified folds (according to drug sensitivity categorization) to ensure balanced class distribution among different folds for unbiased performance evaluation. We used the Scikit-learn package (https://scikit-learn.org/) to perform stratified cross-validations. In one evaluation iteration, four out of five folds of the samples were used as the training set. The remaining one fold of samples was used as the validation dataset for early stopping. At last, the test performance of the classifier in each evaluation iteration was recorded.

All of the baseline models that followed the pre-training and fine-tuning scheme used the exact same training data. For models that were designed for domain alignment (ADAE, DSN-NMD, DSN-DANN, CORAL, Celligner and Velodrome), we tuned them using the exact same data split for source and target domains as CODE-AE. We used the same training procedure as CODE-AE to train these models except Velodrome, which was trained using the code provided by the paper[5].

For models that were not specifically designed for domain alignment (VAEN, AE, DAE and VAE), we included all the unlabelled data from both source and target domains in their pre-training and only optimized the reconstruction loss. For models that were specifically designed for anticancer drug sensitivity prediction (COXEN and TCRP), we used the same training procedures in the papers[4,8]. For baseline models that did not include a pre-training process using unlabelled data (TCRP, MLP, Elastic Net and Random Forest), we only used the labelled training sets that were used for fine-tuning of CODE-AE. For all models tested, we used the exact same parameters grid for tuning the common hyperparameters and the same testing data for the performance evaluation. Specifically, the following common hyperparameters are kept the same across models: the number of warm-up (pre-training) epochs and training epochs as well as the architecture of deep learning models, namely hidden layers' dimension, activation function, drop-out probability, training optimizer (except certain model specific customization according to respective original publication). In addition, we made the test data (five-fold train test splits) the same to all models.

Performance evaluation. We choose AUROC as the measurement metric owing to its insensitivity to changes in the test dataset's class distribution[54]. The model performance was measured in AUROC over the patient tissue expression data and corresponding drug response records. The performance of different methods was compared by the average of AUROCs of five iterations. It is noted that only cell-line data were used for the model training and hyperparameter selections, and all ex vivo tissues and patient data were purely used for the testing. In addition to AUROC, AUPRC is used as an additional metric.

Tumour classification with sex as a confounding factor. During the pre-training phase, we split the unlabelled gene expression samples of different sexes into two groups: female and male, and aligned embeddings of different sexes through a shared encoder in CODE-AE. Then following the approach adopted in ADAE[7], in the downstream brain cancer subtype classification task, we trained Elastic Net models with shared embeddings from CODE-AE of labelled samples from one of the sexes only and evaluated the classification prediction performance with samples of the opposite sex. As the evaluation procedure described in ref. [7], the classification performance is measured by the AUPRC and the AUROC of ten-fold cross-validations. Besides, we performed a two-sample *t*-test on the average performance between CODE-AE-ADV and the best non-CODE-AE methods ADAD and CORAL in each setting.

Gene set over-representation analysis
For a set of genes, their over-represented functional annotations were determined by DAVID[55].

Clustering analysis
We grouped 9,808 patients and 59 drugs into 100 patient clusters and 30 drug clusters with spectral biclustering methods based on the predicted drug responses profiles[17]. For each cluster, we averaged the predicted drug response scores and then calculated its normalized *z*-score across all drugs for each primary tumour type. The higher *z*-score indicates that this cluster of patients is more sensitive to the drug.

Mutation profile analysis
For patients with LSCC treated with gemcitabine and AICAR, we selected responsive/resistant clusters based on the *z*-scores (≥2.0 responsive, ≤0.5 resistant) and acquired their somatic mutation profiles. We compared the absolute difference in the average mutation rate of all genes between responsive and resistant patient samples.

Reporting summary
Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability
The original CCLE, GDSC, PDTC and TCGA data are publicly available datasets. CCLE data were downloaded from the DepMap portal (https://depmap.org/portal/download/). GDSC data were downloaded from the GDSC website (https://www.cancerrxgene.org/). PDTC datasets were obtained from Breast Cancer PDTX Encyclopedia (https://caldaslab.cruk.cam.ac.uk/bcape/). TCGA data were downloaded from UCSC Cancer Genome Browser Xena[56]. Other intermediate files and TCGA tissue sample predictions can be found at https://doi.org/10.5281/zenodo.7027757[57].

Code availability
The source code is available at https://doi.org/10.5281/zenodo.7027757[57] and on CodeOcean at https://doi.org/10.24433/CO.4762159.v1[58].

References

1. Pham, T.-H., Qiu, Y., Zeng, J., Xie, L. & Zhang, P. A deep learning framework for high-throughput mechanism-driven phenotype compound screening and its application to COVID-19 drug repurposing. *Nat. Mach. Intell.* **3**, 247–257 (2021).
2. Barretina, J. et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
3. Sakellaropoulos, T. et al. A deep learning framework for predicting response to therapy in cancer. *Cell Rep.* **29**, 3367–3373 (2019).
4. Zhu, Y. et al. Enhanced co-expression extrapolation (COXEN) gene selection method for building anti-cancer drug response prediction models. *Genes* **11**, 1070 (2020).
5. Sharifi-Noghabi, H., Alamzadeh Harjandi, P., Zolotareva, O., Collins, C. C. & Ester, M. Out-of-distribution generalization from labelled and unlabelled gene expression data for drug response prediction. *Nat. Mach. Intell.* **3**, 962–972 (2021).
6. Warren, A. et al. Global computational alignment of tumor and cell line transcriptional profiles. *Nat. Commun.* **12**, 22 (2021).
7. Dincer, A. B., Janizek, J. D. & Lee, S.-I. Adversarial deconfounding autoencoder for learning robust gene expression embeddings. *Bioinformatics* 36, i573–i582 https://doi.org/10.1093/bioinformatics/btaa796 (2020).
8. Ma, J. et al. Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients. *Nat. Cancer* **2**, 233–244 (2021).
9. Jia, P. et al. Deep generative neural network for accurate drug response imputation. *Nat. Commun.* **12**, 1740 (2021).
10. Bruna, A. et al. A biobank of breast cancer explants with preserved intra-tumor heterogeneity to screen anticancer compounds. *Cell* **167**, 260–274 (2016).
11. Kuenzi, B. M. et al. Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell* **38**, 672–684 (2020).
12. Kingma, D. P. & Welling, M. Auto-encoding variational Bayes. Preprint at https://arxiv.org/abs/1312.6114 (2013).
13. Vincent, P., Larochelle, H., Bengio, Y. & Manzagol, P.-A. Extracting and composing robust features with denoising autoencoders. In *Proc. 25th International Conference on Machine Learning* 1096–1103 (2008).
14. Sun, B. & Saenko, K. Deep CORAL: correlation alignment for deep domain adaptation. In *European Conference on Computer Vision* 443–450 (Springer, 2016).
15. Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D. & Erhan, D. Domain separation networks. In *Proc. 30th International Conference on Neural Information Processing Systems* 343–351 (2016).

16. Ben-Hamo, R. et al. Resistance to paclitaxel is associated with a variant of the gene *BCL2* in multiple tumor types. *npj Precis Oncol.* **3**, 1–11 2019).
17. Kluger, Y., Basri, R., Chang, J. T. & Gerstein, M. Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Res.* **13**, 703–716 (2003).
18. Rawluk, J. & Waller, C. F. in *Small Molecules in Oncology* 235–246. Ed. Martens, U. M. (Springer, 2018).
19. Guo, D. et al. The AMPK agonist AICAR inhibits the growth of EGFRvIII-expressing glioblastomas by inhibiting lipogenesis. *Proc. Natl Acad. Sci. USA* **106**, 12932–12937 (2009).
20. Chen, X. et al. Novel direct AMPK activator suppresses non-small cell lung cancer through inhibition of lipid metabolism. *Oncotarget* **8**, 96089 (2017).
21. Manegold, C. Gemcitabine (Gemzar®) in non-small cell lung cancer. *Expert Rev. Anticancer Ther.* **4**, 345–360 (2004).
22. Hayashi, H., Kurata, T. & Nakagawa, K. Gemcitabine: efficacy in the treatment of advanced stage nonsquamous non-small cell lung cancer. *Clin. Med. Insights Oncol.* **5**, 177–184 (2011).
23. Rehmani, H. S. & Issaeva, N. EGDR in head and neck squamous cell carcinoma: exploring possibilities of novel drug combinations. *Ann. Transl. Med* **8**, 13 (2020).
24. Wang, C.-J. et al. Shock wave therapy induces neovascularization at the tendon–bone junction. A study in rabbits. *J. Orthop. Res.* **21**, 984–989 (2003).
25. Tang, X. et al. Efficacy and safety of gefitinib in patients with advanced head and neck squamous cell carcinoma: a meta-analysis of randomized controlled trials. *J. Oncol.* **2019**, 6273438 (2019).
26. Chen, Q. et al. An EGFR-amplified cervical squamous cell carcinoma patient with pulmonary metastasis benefits from afatinib: a case report. *Onco Targets Ther.* **13**, 1845 (2020).
27. Hale, G. M. & Querry, M. R. Bladder cancers respond to EGFR inhibitors. *Cancer Discov.* **4**, 980–981 (2014).
28. Dragovich, T., & Campen, C. Anti-EGFR-targeted therapy for esophageal and gastric cancers: an evolving concept. *J. Oncol.* **2009**, 804108 (2009).
29. Hientz, K., Mohr, André, Bhakta-Guha, D. & Efferth, T. The role of p53 in cancer drug resistance and targeted chemotherapy. *Oncotarget* **8**, 8921 (2017).
30. Jung, S. et al. Contribution of p53 in sensitivity to egfr tyrosine kinase inhibitors in non-small cell lung cancer. *Sci. Rep.* **11**, 19667 (2021).
31. Bai, Y., Li, Y., Bai, J. & Zhang, Y. Hsa_circ_0004674 promotes osteosarcoma doxorubicin resistance by regulating the miR-342-3p/FBN1 axis. *J. Orthop. Surg. Res.* **16**, 510 (2021).
32. Takahashi, A. et al. Ankyrin repeat domain 1 overexpression is associated with common resistance to afatinib and osimertinib in EGFR-mutant lung cancer. *Sci. Rep.* **8**, 14896 (2018).
33. Sosa Iglesias, V., Giuranno, L., Dubois, L. J., Theys, J. & Vooijs, M. Drug resistance in non-small cell lung cancer: a potential for notch targeting? *Fron. Oncol.* **8**, 267 (2018).
34. Shain, K. H. & Dalton, W. S. Cell adhesion is a key determinant in de novo multidrug resistance (MDR): new targets for the prevention of acquired MDR. *Mol. Cancer Ther.* **1**, 69–78 (2001).
35. Henke, E., Nandigama, R. & Ergün, S. Extracellular matrix in the tumor microenvironment and its impact on cancer therapy. *Front. Mol. Biosci.* **6**, 160 (2020).
36. Manica, M. et al. Toward explainable anticancer compound sensitivity prediction via multimodal attention-based convolutional encoders. *Mol. Pharm.* **16**, 4797–4806 (2019).
37. Costello, J. C. et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.* **32**, 1202–1212 (2014).
38. He, D. & Xie, L. A cross-level information transmission network for hierarchical omics data integration and phenotype prediction from a new genotype. *Bioinformatics* **38**, 204–210 (2022).
39. Ma, J. et al. Using deep learning to model the hierarchical structure and function of a cell. *Nat. Methods* **15**, 290–298 (2018).
40. Salzmann, M., Ek, C. H., Urtasun, R. & Darrell, T. Factorized orthogonal latent spaces. In *Proc. Thirteenth International Conference on Artificial Intelligence and Statistics* 701–708 (2010).
41. Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. & Smola, A. A kernel two-sample test. *J. Mach. Learn. Res.* **13**, 723–773 (2012).
42. Arjovsky, M., Chintala, S. & Bottou, L. Wasserstein generative adversarial networks. In *International Conference on Machine Learning* 214–223 (PMLR, 2017).
43. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V. & Courville, A. C. Improved training of Wasserstein GANs. In *Proc. 31st International Conference on Neural Information Processing Systems* 5767–5777 (2017).
44. Howard, J. & Ruder, S. Universal language model fine-tuning for text classification. https://aclanthology.org/P18-1031 (ACL, 2018).
45. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. https://aclanthology.org/N19-1423 (NAACL, 2019).
46. Hinton, G. E. & Zemel, R. S. Autoencoders, minimum description length, and Helmholtz free energy. *Adv. Neural Inf. Process. Syst.* **6**, 3–10 (1994).
47. Ganin, Y. et al. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **17**, 2096–2030 (2016).
48. Ghandi, M. et al. Next-generation characterization of the cancer cell line encyclopedia. *Nature* **569**, 503–508 (2019).
49. Hutter, C. & Zenklusen, J. C. The Cancer Genome Atlas: creating lasting value beyond its data. *Cell* **173**, 283–285 (2018).
50. Clayton, E. A., Pujol, T. A., McDonald, J. F. & Qiu, P. Leveraging TCGA gene expression data to build predictive models for cancer drug response. *BMC Bioinformatics* **21**, 364 (2020).
51. Huang, J. K., Jia, T., Carlin, D. E. & Ideker, T. pyNBS: a python implementation for network-based stratification of tumor mutations. *Bioinformatics* **34**, 2859–2861 (2018).
52. Yang, W. et al. Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* **41**, D955–D961 (2012).
53. Iorio, F. et al. A landscape of pharmacogenomic interactions in cancer. *Cell* **166**, 740–754 (2016).
54. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **27**, 861–874 (2006).
55. Huang, Da. Wei, Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009).
56. Goldman, M., Craft, B., Brooks, A., Zhu, J. & Haussler, D. The UCSC Xena Platform for cancer genomics data visualization and interpretation. Preprint at *BioRxiv* (2018). https://doi.org/10.1101/326470
57. He, D., Liu, Q., Wu, Y. & Xie, L. A context-aware deconfounding autoencoder for robust prediction of personalized clinical drug response from cell line compound screening. *Zenodo* https://doi.org/10.5281/zenodo.7027757 (2022).
58. He, D., Liu, Q., Wu, Y. & Xie, L. Context-aware deconfounding autoencoder. *CodeOcean* https://doi.org/10.24433/CO.4762159.v1 (2022).

Acknowledgements

This work has been supported by the National Institute of General Medical Sciences of the National Institute of Health (R01GM122845) (L.X.) and the National Institute on Aging of the National Institute of Health (R01AD057555) (L.X.).

Author contributions

D.H. conceived the concept, prepared data, implemented the algorithms, performed the experiments, analysed data and wrote the manuscript. Q.L. performed the experiments, analysed data and wrote the manuscript. Y.W. performed the experiments and analysed data. L.X. conceived and planned the experiments, and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-022-00541-0.

Correspondence and requests for materials should be addressed to Lei Xie.

Peer review information *Nature Machine Intelligence* thanks Vassilis Gorgoulis, Yitan Zhu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022

nature research

Corresponding author(s): Lei Xie, Di He

Last updated by author(s): May 31, 2022

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data preprocessing scripts are available at https://github.com/XieResearchGroup/CODE-AE.
Data analysis	Data analysis scripts are available at https://github.com/XieResearchGroup/CODE-AE.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Benchmark datasets used in experiments are available at https://zenodo.org/record/4776448

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Not applicable.
Data exclusions	Not applicable.
Replication	Scripts to run the experiments are provided in https://github.com/XieResearchGroup/CODE-AE.
Randomization	Training data are stratified (according to drug response labels) sampled into training/validation parts (80% vs 20%) during fine tuning (with validation). Results reported are averaged with 5 repetitions.
Blinding	Not applicable.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging