Autoencoder Model for Translating Omics Signatures

Nikolaos Meimetis[1], Krista M. Pullen[1], Daniel Y. Zhu[1], Avlant Nilsson[1,2], Trong Nghia Hoang[3], Sara Magliacane[4,5] and Douglas A. Lauffenburger[1]*

1) Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

2) Department of Biology and Biological Engineering, Chalmers University of Technology, Gothenburg, SE 41296, Sweden

3) School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA, 99164-4236, USA

4) Institute of Informatics, University of Amsterdam, Amsterdam, The Netherlands

5) MIT-IBM Watson AI Lab, Cambridge, MA 02139, USA

* Corresponding author, lauffen@mit.edu

Abstract

The development of effective therapeutics and vaccines for human diseases requires a systematic understanding of human biology. While animal and in vitro culture models have successfully elucidated the molecular mechanisms of diseases in many studies, they yet fail to adequately recapitulate human biology as evidenced by the predominant likelihood of failure in clinical trials. To address this broadly important problem, we developed a neural network autoencoder framework to map omics profiles from designated species or cellular contexts into a global latent space, from which germane information can be mapped between different contexts. This approach performs as well or better than extant machine learning methods and can identify animal/culture-specific molecular features predictive of other contexts, without requiring homology matching. For an especially challenging test case, we successfully apply our framework to a set of inter-species vaccine serology studies, where no 1-1 mapping between human and non-human primate features exists.

Introduction

Animal and cellular models are essential tools for studying and gaining insights into the underlying biology of human diseases, but these insights are not always clinically translatable, resulting in the failure of numerous therapeutics in clinical trials[1,2]. A common approach is to choose orthologous biomolecules, including genes, proteins, and cellular pathways, to perform direct functional comparisons across species. However, functional divergence and the absence of orthologous biomarkers can hinder these direct comparisons between species [3–5]. Furthermore, within the same species, comparing different cellular models reveals that the transcriptional response to chemical stimuli is cell type-specific due to distinct genetic profiles, creating an additional barrier to understanding the mechanism of action of various therapeutics[6–9]. Consequently, computational systems-based approaches are needed to gain a better understanding of the biological insights needed for inter-system translation.

Advancements in omics and sequencing technologies have enabled the generation of large-scale datasets from both animal and human species, facilitating more powerful analyses and comparisons of molecular features between different biological systems[2,3,10–13]. Consequently, this has led to the development of numerous new statistical and machine learning models[3,13–17] for identifying similarities and differences between species and experimental models. Notably, most existing approaches focus on direct correlations between analogous biomarkers or processes across species despite known species and model system differences. To address this challenge, Brubaker et al. proposed a technique called "Translatable Components Regression"[18] (TransCompR), which can map human data into the principal component space of another species to identify translatable animal features that can predict human disease processes and phenotypes. Although this approach has been successfully applied to gain insights into some inflammatory pathologies[18,19], it requires the usage of homologs or at least comparable molecular features of the same dimensionality between species. Moreover, this linear approach is only capable of deciphering a part of inter-species relationships, which can be highly non-linear. We

50 emphasize, however, that TransCompR is not aimed centrally at aligning signatures between species but
51 rather at identifying what information inherent in a signature in one species is most germane for
52 understanding phenotype characteristics in the other – which is not the main goal of our new work here.

53 With the advent of large-scale datasets and deep learning approaches, autoencoders have the potential
54 to provide a deep learning framework that can approximate the non-linear relationships underlying
55 different biological systems and species. Autoencoders are artificial neural networks (ANNs) that can
56 embed raw input data into a lower dimensional space and then reconstruct the original data again[20].
57 Autoencoders have been successfully used in several tasks in biology including analyzing high
58 dimensional data[21,22], denoising single-cell RNA sequencing data[23–25], deciphering the hierarchical
59 structure of transcriptomic profiles[26,27], and predicting gene expression caused by various stimuli[28–30].

60 Regarding cellular models, DeepCellState[31], an autoencoder model, can predict the transcriptional
61 profile of a cell type after drug perturbations based on the behavior of another cell type. However, this
62 approach still requires a 1-1 mapping of molecular features, since it uses the same encoder for both cells
63 to capture a global cell representation. This suggests that an approach, similar to those used in language
64 translation autoencoder-based models, which create a global language representation[32,33], is more
65 appropriate and might improve biological inter-systems translation.

66 Relevant to the idea of building a global latent space, a recently proposed framework called
67 compositional perturbation autoencoder (CPA)[34] was used to construct a basal latent space devoid of
68 covariate and perturbational signals, capturing only the basal cell state in single-cell RNA sequence data.
69 CPA can be used to generate *in-silico* transcriptional profiles at the single-cell level for different
70 perturbations, cells, and species, although it still requires mapping to ortholog genes.

71 In this study, we build upon ideas presented in the CPA approach and ideas from language translation
72 models[32,33] to develop an ANN framework that utilizes separate autoencoders for each biological
73 system, enabling the mapping of samples into a global cell or species space, while providing feature

importance estimates for various predictive tasks. The basic model is trained to simultaneously minimize the reconstruction error of the input and the distance between samples coming from the same condition in the global latent space. Our framework is benchmarked, using the latest version of the L1000 dataset[12], against the established linear approaches of TransCompR[18] and FIT[15], as well as the ANN approach of DeepCellState[31], which all require 1-1 feature mapping. We demonstrate that our approach outperforms FIT and DeepCellState, while there is not a statistically significant difference when comparing with TransCompR in cellular models. Additionally, we present several variations of the model and we illustrate the adaptability of our framework by applying it to data of varying omics type and sample size. Furthermore, we demonstrate its biological interpretability, an aspect that deep learning models often struggle to attain, by using an integrated gradients approach[35]. To analyze the performance of the model in inter-species translation we used a mouse[36] and a human[37] lung fibrosis single-cell RNA sequencing datasets, as well as two small-scale serology datasets from humans and non-human primates vaccinated against HIV and SHIV, respectively[38,39]. The latter serves as a novel case study of computational inter-species translation where no 1-1 mapping between features exists. After building the model, we identified molecular features in non-human primates that are predictive of protection against HIV in humans, without analogous features necessarily being present in human data. These findings demonstrate that features derived from this approach can be predictive of the phenotypic profile of another cellular or animal model without requiring them to be homologs, allowing us to maximize the amount of information we can capture from different model systems to advance our understanding of complex human disease biology.

Results

A flexible Artificial Neural Network framework for omics translation

The framework can be modified depending on the task at hand, however, it contains some core elements and functionalities. The framework (Methods) consists of two separate ANN autoencoders for each biological system (e.g. cell line, species), while there are two main variations of a latent space

module which ensures that it removes the system-specific effect of perturbations and constructs a global latent space common for both autoencoders (Figure 1a). The separation of the autoencoders allows for the utilization of molecular features without 1-1 mapping between different systems. The first latent module consists of only one global latent space, created by maximizing the similarity and Mutual Information of embeddings derived from the same condition/perturbation in a different species or cell line. The second main latent module is based on the recently published compositional perturbation autoencoder (CPA)[34], where there are two separated latent spaces: i) a global/basal latent space and ii) a composed latent space. The global latent space is shaped the same way as in the first main module with the addition of an extra discriminator which attempts to remove the cell-line or species effect by penalizing the model when this classifier can distinguish between different cells or species[34]. In the composed latent space, a cell/species classifier is jointly trained to ensure there is a cell/species effect, which is added through a trainable covariate vector[34] or even completely non-linearly with two intermediate ANNs. Finally, we introduced a novel functionality where feature importance can be identified for various tasks based on integrated gradients[35] (Methods).

Benchmarking reconstruction and translation of gene expression profiles between two cell-lines

First, we compared our ANN framework with state-of-the-art techniques in the context of translating homologous genes between in-vitro models within the same species. The L1000 transcriptomics dataset[12] was used for benchmarking different approaches to translate between different human cell lines. The two main variations of our approach, as well as a variation where a classifier is jointly trained on the global latent space (Methods), are compared with three previously published approaches (DeepCellState[31], FIT[15], TransCompR[18]). Finally, the model is also compared to “direct translation”, where we directly use the gene expression profile in one cell line to predict an equivalent perturbation in the other cell line. The models are evaluated both in the task of translating the gene expression profile, as well as in the task of simply reconstructing the gene expression input. Performance is evaluated with: i) Pearson’s correlation between predicted gene expression and ground truth gene

124 expression, ii) the per sample Spearman's rank correlation, and iii) the accuracy in predicting correctly
125 the sign of regulation of a gene. When utilizing the 978 landmark genes measured in the L1000, all of
126 our framework's variations provide a statistically significant increase in performance compared to the
127 direct translation across all metrics (Figure 1b). The p-values for all these comparisons across all metrics
128 are provided in Supplementary Tables 1-4. Our main method outperforms FIT[15] and the basic
129 DeepCellState[31] (DCS) method when translating from the HT29 cell line to A375, while when translating
130 in the reverse direction our framework outperforms even the variations of DCS (Figure 1b left). An
131 interesting note is that the 2nd variation of DCS also enforces a similarity in the latent space when
132 training the model, and since it outperforms the basic DCS, this indicates the importance of enforcing
133 similarity in the global latent space via some distance metric. The basic DCS approach while taking a hit
134 in the task of translation greatly outperforms the other approaches in terms of reconstruction of the
135 input within the same cell line, while our approach still performs well and comparably with the other
136 methods (Figure 1b right). The other variations of our framework perform comparably, even in the case
137 of a global latent space where the cell line effect is not completely removed. When using 10,086 genes
138 all approaches take a significant hit in performance but still perform better than the baseline while there
139 is no statistically significant difference between our approach and other state-of-the-art approaches
140 (Figure 1c). Our approach only outperforms FIT in one direction of translation. Finally, our approach
141 performs better than direct translation also in the case of using different genes as input for each cell-line
142 (Figure 1f). The performance is comparable to that using the same genes for both cell lines, indicating
143 the potential to later extend the method in cases where no 1-1 mapping exists.

144 Performance in using predicted gene expression to infer transcription factor activity

145 In the case of using all 10,086 genes, there is a decrease in performance (Figure 1b-1c), but these
146 transcriptomic profiles may still be used to infer the activity of transcription factors (TFs). By inferring
147 transcription factor activity using the predicted gene expression values with the VIPER algorithm[40] and
148 the Dorothea regulon[41] (Methods), the model has increased performance compared with using the

149 10,086 genes and comparable to that in the case of the landmark genes (Figure 1d). For the case of
150 other gene sets by performing Gene Set Enrichment Analysis (Supplementary Method 1, Supplementary
151 Figure 3) this is true only for small well-defined gene sets like KEGG pathways. The increased
152 performance in inferring only TFs activity might be a result of the previously shown ability of
153 autoencoders to capture regulatory relationships between genes[26,31].

154 Creating cell-line-specific regions in the latent space enables robust cell classification

155 To establish the ability of the model to retrieve cell-line-specific information in different variations we
156 evaluated the performance in classifying the cell line used in a sample (Figure 1e) when using all the
157 10,086 genes of the L1000 (landmarks in Supplementary Table 11). The performance of ANN classifiers
158 trained on the original gene expression data serves as the baseline. First, classifiers trained with pre-
159 trained embeddings from DCS or our framework with one global latent space are expected to have
160 lower performance than the baseline, as these approaches generate embeddings aiming to filter the
161 cell-line effect as much as possible. Our framework seems to generate more global embeddings than
162 DCs. Interestingly when jointly training a classifier in the global latent space we can outperform the
163 baseline (Figure 1e) while the cell-line effect is still partially filtered in the higher dimensions
164 (Supplementary Figure 4). The CPA-based model in the composed latent space perfectly classifies cell
165 lines with 100% accuracy, even though the similarity of input gene expression data between training and
166 validation sets, as well as the latent space embedding similarity, is generally low (Supplementary Figure
167 5). The CPA-based framework can create very well-separated cell-line-specific regions (Supplementary
168 Figure 6) in the composed latent space, indicating the framework's ability to shape the latent spaces
169 with robust cell-line-specific regions and explaining the observed accuracy.

170 Framework’s performance analysis

171 We investigated the factors affecting the performance of the framework, using the CPA-based approach
172 due to its ability to shape the latent spaces. The framework has similar behavior and performance across

173 cell-line pairs (Figure 2a). Some cell-line pairs perform slightly worse, as the original correlation between
174 the same perturbations in the cell-line pair correlates with the model's (Figure 2b). Additionally, for all
175 cell lines, ~600 total training samples are enough to train a high-performance model (Figure 2a).
176 Interestingly, the amount of paired conditions required to train the framework can be low, since even
177 with ~10-15% of the samples being paired we can shape a global latent space to facilitate translation
178 successfully (Figure 2c). Finally, it seems the model is not affected by low to medium imbalance between
179 conditions coming from each cell line (Figure 2d). The same analysis for using all 10,086 genes shows
180 similar trends and results (Supplementary Figure 7).

181 Evaluation of latent space embeddings

182 The validity of the global latent embeddings produced from our framework is evaluated based on three
183 main tasks (Figure 3a-3c): i) embeddings of pairs coming from the same cell line should not be more
184 similar to each other than embeddings from random pairs, ii) embeddings of pairs coming from the
185 same condition regardless of cell-line should be similar, and iii) embeddings of pairs that consist
186 biological duplicates (same drug tested on the same cell-line, with the same dose and duration) should
187 be similar to each other. These tasks are evaluated using cosine distance to quantify how close are
188 embeddings in the latent space. No cell-line effect is observed in the global latent space, both for
189 training and validation embeddings (Figure 3a, Supplementary Figure 8), proving that the framework
190 succeeds in creating a global cell-line space. Embeddings coming from the same condition are closer to
191 each other than embeddings coming from random pairs (Figure 3b), while biological duplicates are
192 slightly even closer (Figure 3c), validating that indeed we have constructed a stimuli-specific global
193 latent space. Similar patterns can be observed in the global latent space also when using the CPA-based
194 approach (Figure 3d), while in the composed latent space where a cell-line effect should be visible,
195 embeddings of pairs coming from the same cell-line are closer in the latent space than random (Figure
196 3d). Cohen's d is used to quantify the difference between the distributions of cosine distances across all

197 folds (Figure 3e), proving that indeed there is a much higher cell-line effect in the latent space than the
198 effect in the global latent space.

199 Framework's trained parameters that capture biological information

200 To better interpret the framework, we investigate the biological information captured by some parts of
201 it. Since only the cell-line effect is filtered in the global latent space in the CPA-based framework, then
202 the trainable covariate vectors should only add a basal cell-specific effect. Intuitively this would mean
203 that in the case of gene expression coming from untreated cells (control perturbations), the global latent
204 embeddings should capture a "zero"/basal cell state, and thus the trained covariate which is added to
205 that global representation should be similar to the composed latent space vectors which now captures
206 the cell line effect. On this front, we used control samples from the L1000 dataset not seen by the model
207 during training, as well as samples coming from untreated cell lines from the Cancer Cell Line
208 Encyclopedia[42] (CCLE), pre-processed and using only genes present in the L1000 landmark genes.
209 Additionally, for this case study two models were trained completely separately: the original benchmark
210 model of A375/HT29 cell lines and another model using the PC3 prostate cancer cell line and the HA1E
211 normal epithelial cell line. The latter pair was chosen to be used in the subsequent analysis because of
212 both the comparably high performance of the model (Figure 2a) and a comparable amount of paired
213 conditions (330 in PC3/HA1E vs 406 in A375/HT29), while these two cell lines are significantly different
214 in terms of biology. Each trained covariate, even though the models were trained separately, is observed
215 to be closer to its respective cell-line control signatures, both when using PCA for dimensionality
216 reduction (Figure 3f), where clearly defined cell-line specific regions are observed, as well as when using
217 t-SNE (Figure 3g).

218 Identification of important features that capture task-specific characteristics

219 The framework can identify both latent variables and genes, important for tasks of interest, enabling the
220 investigation of the framework's ability to identify important features that can be biologically

221 interesting. PC3 and HA1E cell lines were chosen, as described previously, to train a model with one
222 global latent space and one classifier trained jointly to classify the cell lines from which the samples
223 were derived (contradictory learning tasks). To identify the importance of input variables of different
224 parts of the model w.r.t their output, an integrated gradient-based approach[35] was utilized (Methods),
225 attributing an importance score to each variable of interest. It is interesting since the same genes are
226 used for both cell lines, to identify what the model considers important to translate, in the case of
227 paired conditions, a gene expression profile from one cell line to the same set of genes in another cell
228 line. For approximately, 48% of the 10,086 genes, the model considers a different gene than itself as
229 important (top 20% genes based on their absolute score) to be translated in the other cell line for the
230 same condition (Figure 4a). For the case of landmarks genes, this percentage slightly drops to 41%
231 (Figure 4b). Since the cell lines chosen are one cancerous and one non-cancerous this is an interesting
232 result indicating that the model potentially avoids the fallacy of using the same gene as a proxy for its
233 gene expression across quite different biological systems. Additionally, Spearman's correlation between
234 the absolute importance scores and the absolute gene expression across different samples is ~0.04 on
235 average (Figure 4c), proving that the model doesn't consider important just genes that are highly
236 regulated. Subsequently, the jointly trained classifier can identify a subset of important latent variables
237 of the global latent space for classifying samples. Even though the cell line effect is partially filtered and
238 embeddings coming from the same condition are globally close to each other (Supplementary Figure 4),
239 the model can still identify 11 latent variables that are important for classifying a cell line (Figure 4d), by
240 using a k-means based approach (see Methods). These latent variables can greatly separate the latent
241 vectors based on the cell line (Figure 4e), even though globally the cell line-specific separation of
242 embeddings is minimized. Genes considered important by the encoders to control these latent variables
243 should be either cell line-specific genes or a subset of genes that can easily distinguish between cell
244 lines. The importance scores of the genes for each cell line-specific encoder do not correlate at all and
245 are different between the two cell lines (Figure 4f). It is possible to even train a very simple generalized

246 linear model to classify cell lines based on gene expression, only using a subset of these important
247 genes, achieving high performance with only ~25-30 genes from each cell line (Figure 4g).

248 Performance in inter-species translation for lung fibrosis

249 To evaluate the framework in inter-species translation, we utilize the raw gene counts coming from
250 single-cell RNA-sequencing of a mouse[36] and human[37] dataset with lung fibrosis. Based on the original
251 CPA manuscript [34], the decoders predict a mean and a variance for every gene, derived from a negative
252 binomial distribution. Furthermore, both a trainable species vector and another trainable cell type
253 vector are added to the global space, where it is attempted to filter both species and cell type effect. We
254 compare the CPA-based framework using all genes available in the dataset and using only the homologs,
255 as well as the best-performing variation of DCS and TransCompR. We evaluate the performance in 10-
256 fold cross-validation in terms of R^2 of the mean and variance in the reconstruction of gene expression
257 profiles, as well as in translating between mouse and human, where we would expect to observe a
258 similar distribution in a successful translation, and thus mean and variance. Our framework outperforms
259 the other approaches in terms of R^2 of the means both in reconstruction and translation (Figure 5a). In
260 predicting the variance there is not always a statistically significant improvement, while all approaches
261 have generally low performance (Figure 5b). Finally, there is not a significant difference between using
262 all genes or just homologs (Figure 5a-5b). Additionally, the CPA-based approaches and TransCompR are
263 compared in terms of classifying fibrosis, species, and cell type. Cell type and species are also classified
264 when a signature is translated into another species in the composed latent space. Using all genes in our
265 framework outperforms homolog genes approaches in the fibrosis and species-translation tasks, while
266 the performance of all approaches is high (Figure 5c). Similar to what was observed in the L1000
267 dataset, species and cell type are predicted perfectly in our framework. Additionally, both the species
268 and cell-type effects are filtered (Figure 5e-5f, Supplementary Figures 9-10) in the global latent space
269 compared to the composed latent space. Finally, it is worth noting that based only on the human lung
270 fibrosis dataset, out of the top 10 genes contributing to the top principal components, 3 of them are not

271 homologs (Supplementary Figure 11). Regardless of performance, a method considering only
272 homologues would exclude important genes for lung fibrosis.

273 Generalization in other disease datasets

274 It is useful to investigate the extent to which the model can make predictions about disease, species,
275 and cell types in other datasets, used as external test sets, as well as different tissue and disease
276 datasets. On this front, an external mouse lung fibrosis[43] and human liver cirrhosis[44] dataset were used.
277 In the mouse dataset, even though different genes were measured than those in our model, the
278 performance is still decent in disease classification, and better than random (Figure 5d). In the human
279 dataset, which is an extreme case of fibrosis in a different organ, the model has significantly lower
280 performance but is still barely higher than random (Figure 5d). Interestingly, in both cases the model can
281 still perfectly identify cell types and species (Figure 5d), proving once again the ability of the model to
282 capture a very general species and cell type effect.

283 Building an inter-species model from serology data for predicting protection against HIV

284 We aimed to develop a framework for species translation when no 1-1 mapping between molecular
285 features exists, a limitation of existing cross-species models. For this purpose, we utilized non-human
286 primate and human serology datasets[38,39] following vaccination against SHIV and HIV, respectively. In
287 this special case, the species effect was added completely non-linearly during training, with 2 small
288 intermediate neural networks (Figure 6a). This modification made no significant difference in
289 performance and the properties of the global space (Supplementary Figures 12-14). We reason that
290 adding more non-linear parts in the model is more appropriate for species translation in this setting,
291 even if performance is not increasing, due to few data points. The framework is trained so that
292 protected individuals are close in the global latent space regardless of species, while two separate
293 classifiers try to predict protection and vaccination outcomes (Figure 6a). In the composed latent space,
294 a classifier predicts species. For the human serology features the performance is very high (Figure 6b, r =

295 0.8932±0.0115). In non-human primates while some features are not predicted well and there is a big
296 variation in performance between folds, the overall performance is still good (Figure 6c, r =
297 0.7107±0.0420). Finally, the performance across all classification tasks is exceptionally high (Figure 6d),
298 with achieving once again a perfect 100% accuracy in species classification and translation.

299 Non-human primates' serological features predictive of human protection against HIV

300 With our framework, we aimed to identify features predictive of human protection, even non-human
301 ones, by combining the integrated gradient approach with likelihood ratio tests (LRT) performed on each
302 latent variable (see Methods for more details). The likelihood of an alternative model accounting for
303 protection to derive the value of a latent variable is compared against a null model not accounting for
304 protection. Latent variables with an FDR-adjusted p-value<0.05 are considered significant and the sign of
305 the statistic signifies the directionality of the association with protection (Supplementary Figure 16). For
306 the integrated gradient approach, the integrated gradient of each latent variable to predict protection
307 was calculated for each human sample, according to the classifier in the global latent space. Then the
308 average score is used to signify importance. The intersection of latent variables between the two
309 approaches is used to identify the final latent variables contributing to human protection
310 (Supplementary Figure 17). Latent variables are denoted as important only if there is an agreement
311 between the likelihood ratio test results and the integrated gradients (and gradient score≥10%).

312 To identify features that are important for predicting human protection, we utilize integrated gradient
313 scores derived by the respective encoder of each species to estimate the importance of a feature in
314 controlling a latent variable associated with protection. The human features identified indeed have a
315 statistically significant difference between protected and non-protected individuals (Figure 6e). Finally,
316 we identified non-human features that have a high gradient score when translating to human
317 signatures, meaning these non-human primate features are predictive of human features linked with
318 viral protection (Figure 6f). These features are not necessarily associated with non-human primate viral
319 protection (Supplementary Figure 18) but they could be predictive of human protection.

320 ## Discussion

321 Here we developed an ANN framework that facilitates the translation of different omics profiles
322 between different biological systems, by building upon ideas of the CPA approach[34] and other species
323 and cellular translation methods[13,15,18,31], together with ideas from language translation models[32]. The
324 explicit central goal here is to align omics signatures between species, rather than identifying what
325 information inherent in a signature in one species is most germane for understanding phenotype
326 characteristics in the other (which has been the objective in many previous studies[16–19]). The framework
327 performs as well as (or even better in some cases) other state-of-the-art translation techniques, when
328 using homolog features, while it does not take a significant hit in performance when it utilizes different
329 features to represent two species or cell lines. Notably the framework can construct a truly global latent
330 space and manipulate it successfully to create stimuli-specific regions, with high-performing classifiers
331 trained jointly to make predictions for various tasks such as disease diagnosis.

332 Current approaches[13–18] require homolog features and they utilize linear transformations to facilitate
333 translation, thus being able to better approximate only the linear inter-species relationships. Even non-
334 linear ANN approaches like DeepCellState[31] require homology of the molecular features used to describe
335 biological systems. Our framework could fill these gaps by non-linearly modeling relationships between
336 different biological systems, without requiring any kind of homology, while it achieves high performance
337 with only a small percentage of paired conditions. This enables for the first time the training of a
338 framework with serology datasets where no 1-1 mapping of the features between the two biological
339 models exists. This way the identification of relationships between very different molecular profiles that
340 correlate with specific phenotypes, e.g. protection against infection, is possible, through the
341 interpretation of the framework.

342 Interpretability of deep learning models in biology remains a challenge and such models are often
343 criticized due to poor understanding of the biological relationships captured by the models[45,46]. On this
344 front, we demonstrate in our framework how integrated gradient approaches[35] can be used to estimate

345 the importance of features used by different parts of the framework for various tasks, enabling to some
346 extent the interpretation of the model. We were able to propose serological features predictive of
347 human protection against HIV, regardless of species. Finally, different parts of the framework can be
348 used to interpret and retrieve successfully the effect of species or cell types, filtered from the global
349 latent space. This can explain the ability of the framework to predict with very high performance,
350 general cell types and species, even in external disease datasets, derived from different organs/tissues.
351 However, there are still limitations in the generalization of the approach in external datasets.

352 The performance in such datasets drops significantly as we move to more different pathologies and
353 tissues. Even for the same disease, different features measured can lead to reduced performance
354 regarding predictions made about disease diagnosis. Despite our framework being trained successfully
355 even with fewer samples, still it contains a lot of parameters, especially when using a larger number of
356 features, leading to overfitting and scalability issues.

357 Some of these shortcomings can be alleviated by combining our framework with larger datasets, such as
358 ARCHS4[10], which contains hundreds of thousands of publicly available RNA-sequencing data from
359 humans and mice. Training with more data and more diverse unique conditions may enable higher
360 generalization and higher granularity in modeling different biological covariates. Additionally, with the
361 advent of Natural Language Processing (NLP) models[47] and attention-based models[48], our encoder could
362 be modified with NLP-based models. Recently, Geneformer[49], an attention-based model, was pre-
363 trained on a corpus of 30 million single-cell transcriptomic profiles. This approach was proven to be
364 context-aware of the system it encodes and although it still requires some level of homology and a
365 large-scale dataset to be trained, it paves the way to utilize NLP approaches for transfer learning in
366 biology, and ultimately translation.

367 The flexibility of our framework allows the modeling of many different biological systems, even with a
368 small sample size. This could lead to prior computational optimization of biological systems and assays
369 aiming to model human pathology. Features that can explain or predict human biology, without

370 requiring homology, can be identified and optimized, while it is possible to use the framework to
371 explore potential transcriptional modifications to design better disease models.

372 Methods

373 Pre-processing *in-vitro* transcriptomics benchmark dataset

374 The L1000 CMap resource[12] contains bulk gene expression data from drug perturbations across different
375 cell lines and provides a benchmark dataset with diverse conditions and a large sample size (a total of
376 720,216 samples of drug perturbations of varying quality). Additionally, a sufficient number of
377 equivalent perturbations across different biological systems is available (a maximum of 406 *paired*
378 *conditions* for the case of A375 and HT29 cell lines after filtering and pre-processing) to evaluate the
379 performance in translating omics profiles. We selected high-quality drug perturbations from the latest
380 version of the L1000 dataset (accessed via clue.io). The level 5 z-score transformed and pre-processed
381 differential gene expression data of 978 landmark genes, measured with the L1000 assay, and
382 additionally, 9,196 computationally well-inferred genes in the CMap resource, were considered in the
383 subsequent analysis. High-quality perturbations consist of signatures with more than three replicates,
384 where at least half of them passed standard quality control protocols in the assay, and were not
385 identified as statistical outliers. Additionally, for multiple-signature perturbagens (technical duplicates),
386 only the signature with the highest transcriptional activity score (TAS) across these technical duplicates
387 was kept in the dataset. These signatures are called exemplars and are specifically designated for further
388 analysis in the CLUE platform[50]. TAS is a metric inherent to the L1000 dataset that quantifies signal
389 strength and reproducibility and is calculated as the geometric mean of the number of differentially
390 expressed transcripts and the 75th quantile of pairwise replicate correlations. Finally, the ability to
391 distinguish between random signatures and true biological duplicates, meaning the same perturbagen
392 tested on the same cell line for the same duration and dosage, was evaluated for different parts of the
393 dataset, split using varying TAS thresholds (Supplementary Figure 19) and we decided to keep samples
394 with TAS≥0.3. Only 13,699 samples remain in total after filtering and pre-processing, while 1107

395 experiments are available in total in the HT29 cell line and 1213 in the A375 cell line. For the case of
396 control signatures, we followed the same procedure without filtering based on the TAS number.

397 Pre-processing single-cell RNA sequencing interspecies datasets

398 For the human and mouse single-cell RNA-sequencing datasets, we first summarized each annotated cell
399 into one of the four many cell types: i) immune cells, ii) mesenchymal cells, iii) epithelial cells, iv)
400 endothelial cells, and iv) stem cells. These general cell-type labels were used later to filter cell effect
401 from the global latent space and they were also used in the subsequent cell-type classification. Finally,
402 while the raw gene counts are used for reconstruction from the decoder in the loss function, the
403 encoders take as input the log-transformed counts ($x_{input} = \log_{10}(count + 1)$).

404 Pre-processing serology dataset preprocessing

405 For all serology data, we aimed to build a model with just antibody and receptor measurements. From
406 the human serology data, avidity molecular features were first dropped and then the data were scaled
407 by z-scoring per feature. For the non-human primates' samples taken in week 28 were used, while
408 antibody-dependent cellular function features and mass spectrometry data were dropped. This was
409 done to utilize a similar modality of antibody and receptor measurements, as we did in humans. Then
410 the data are log-transformed ($x = \log_{10}(MFI + 1)$), since they consist of Median Fluorescence Intensity
411 (MFI) values and the median per feature from controls is subtracted from each feature to standardize
412 the data and create a differential expression effect. Finally, the data are again scaled by z-scoring per
413 feature.

414 The general framework and the training procedure

415 Generally, the framework always consists of two independent encoders and two independent decoders
416 for each of the species or cell lines in the pair attempting translation, while inside a latent module, the
417 global latent space is shaped (Figure 1a). Both the encoders and the decoders are multi-layer fully
418 connected neural networks, with each layer consisting sequentially of a fully-connected layer, a batch

419 normalization layer[51], an ELU activation function[52], and a dropout layer[53]. Only their final output layer
420 consists of only one fully-connected layer without a trainable bias term.

421 For the construction of the global latent space, in every variation of the framework, the distance
422 ($\boldsymbol{L_{distance}}$) between embeddings of profiles coming from the same perturbation is minimized, their
423 cosine similarity ($\boldsymbol{L_{cosine}}$) and mutual information ($\boldsymbol{L_{MI}}$, see details below) is maximized, and finally the
424 divergence of the distribution of the latent variables from a random uniform distribution is also
425 minimized ($\boldsymbol{L_{prior}}$). Both cosine similarity and Euclidean distance losses were added to enforce the
426 strongest possible filtering of species and cell type effect, while the cosine similarity also enforces
427 normalization of the latent embeddings. The last two tasks mentioned above are achieved via two
428 different ANN discriminators, as proposed in the MINE[54], Deep InfoMax[55] and InfoGraph[56] approaches,
429 where the Jensen-Shannon Mutual Information between embeddings coming from the same
430 perturbation is estimated and the extra prior loss is also calculated and added in the final loss, according
431 to the following equations (with a more detailed description in the paragraph after the equations). The
432 actual implementation is taken from the deepSNEM model[57]
433 (https://github.com/BioSysLab/deepSNEM).

434 - $\boldsymbol{L_{prior}} = \frac{\mathbf{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{N}}\left[\mathbf{log}\left(\boldsymbol{Discr_2}(\boldsymbol{v_i})\right) + \mathbf{log}\left(\mathbf{1} - \boldsymbol{Discr_2}\left(\boldsymbol{z_{g_i}}\right)\right)\right]$, where v_i is a randomly
435 sampled embedding from a prior random uniform distribution ranging from 0 to 1 and z_{g_i} is a
436 global latent space embedding. N is the number of samples in a batch during training.
437 - $\boldsymbol{L_{MI}} = -(\boldsymbol{MI^+} - \boldsymbol{MI^-}),\ \boldsymbol{MI} = \frac{\mathbf{1}}{\boldsymbol{N1}+\boldsymbol{N2}}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{N1}}\sum_{\boldsymbol{j}=\mathbf{1}}^{\boldsymbol{N2}}\boldsymbol{Discr_1}\left(\boldsymbol{z_{g_i}}, \boldsymbol{z_{g_j}}\right)$, where MI^+ and MI^- are
438 respectively the mutual information between pairs derived from the same conditions and pairs
439 coming from different conditions, averaged for every possible pair in a batch during training.
440 z_{g_i}, z_{g_j} are global latent space embeddings, whose mutual information is estimated by the
441 $Discr_1$ discriminator.

442 $Discr_1$ is the discriminator estimating the mutual information between two embeddings from the global
443 latent space. It takes as input two global latent space embeddings and passes them through the same 3
444 fully-connected layers, each of them followed by a ReLU activation function[52] and one fully-connected
445 skip connection. Then the product of the result of this non-linear transformation of the two embeddings
446 is used to approximate a lower bound of their Mutual Information, as proposed originally in MINE[54] and
447 Deep InfoMax[55]. $Discr_2$ is the second discriminator which takes as input an embedding vector and
448 calculates the probability a point in this embedding space is sampled from a specific distribution. This
449 way L_{prior} forces each feature of the learned embeddings to be sampled from a distribution close to the
450 random uniform distribution ranging from 0 to 1. It has three similar fully-connected layers and the final
451 scalar output is passed through a sigmoid activation function[52]. These regularization loss terms
452 ($\boldsymbol{L_{distance}}, \boldsymbol{L_{cosine}}, \boldsymbol{L_{MI}}$) are calculated and averaged across every pair of global embeddings (z_{g_i}, z_{g_j})
453 that are coming from the same condition. The $\boldsymbol{L_{prior}}$ is calculated for every sample in the dataset,
454 meaning every global latent embedding and averaged across samples.

455 For the case of the L1000 dataset, we consider similar perturbations those that are coming from
456 experiments of the same drug, tested on the same cell line, with the same dose and time duration. For
457 the lung fibrosis dataset, similar profiles are considered those coming from samples that have the same
458 diagnosis (fibrosis or not). For the serology datasets, we train the framework so that embeddings
459 coming from protected individuals against HIV are close to each other regardless of species (and even
460 vaccination status)

461 The basic task of this autoencoder framework is reconstruction, which is achieved by minimizing some
462 kind of reconstruction loss ($\boldsymbol{L_{recon}}$). In the case of simple z-scored profiles coming from bulk data, this is
463 done by minimizing the mean sum of squared errors between the input of the encoders and the output
464 of the decoders. The sum of squares error is averaged across samples. For only the case of single-cell
465 RNA-sequencing data, based on the implementation proposed in the original CPA manuscript[34] (found
466 here https://github.com/facebookresearch/CPA), the negative binomial negative log-likelihood is used

467 to optimize the reconstruction, by assuming that the data are derived from a negative-binomial
468 distribution. In that case, the decoders are not just predicting one value per molecular feature, but a
469 mean and variance. The negative binomial negative log-likelihood loss is calculated for every sample and
470 the average across all samples in the batch is taken to be minimized.

471 Finally, classifiers, which also consist of multiple fully connected layers and a final SoftMax activation
472 function for the output, are used for different classification tasks. The average entropy loss across
473 samples for every classification task in the latent space (global or not) is minimized: $entropy_i =$
474 $\frac{1}{N}\sum_{j=1}^{N} CrossEntropy\left(Classifier_i\left(z_j\right), label_j\right)$, where $entropy_i$ is the average cross entropy between
475 every j[th] prediction of a classifier taking a latent vector as input and the true label for that sample.

476 Finally, L2-regularization of the weights and bias of the encoders ($\boldsymbol{L2_{encoder,i}}$), decoders ($\boldsymbol{L2_{decoder,i}}$),
477 and classifiers ($\boldsymbol{L2_{classifier,i}}$) is also enforced. This is implemented by calculating and minimizing the
478 sum of squares for the aforementioned trainable parameters.

479 All the above can be combined in the following loss function, which consists of the basic loss used to
480 train the basic format of the framework:

481 $Loss_{basic} = \lambda_{recon} * L_{recon} + \lambda_{distance} * L_{distance} + \lambda_{MI} * L_{MI} + \lambda_{prior} * L_{prior} + \sum_{i=1}^{2}\left(\lambda_{enc,i} *$
482 $L2_{encoder,i}\right) + \sum_{i=1}^{2}\left(\lambda_{dec,i} * L2_{decoder,i}\right) + \sum_{i=1}^{M}\left(\lambda_{L2class,i} * L2_{classifier,i}\right) + \sum_{i=1}^{M}\left(\lambda_{class,i} *$
483 $entropy_i\right) - \lambda_{cosine} * L_{cosine}$, where M is the number of classifiers and thus individual classification
484 tasks, and the rest of the terms, together with how they are calculated, have already been described in
485 the previous paragraphs of this section.

486 For values for each of the λ used in the loss function see Supplementary Table 5.

487 Translation of omics profiles

488 For translating an omics profile, we first encode it in the global latent space, using the encoder
489 representing the respective biological system (meaning species or cell line). Then the decoder of the

490 other biological system in the pair of the currently trained framework is using directly the embedding of
491 the global latent space to reconstruct the profile in the other system and thus translate the signature. In
492 the case of the CPA-based framework first, the species and cell type effect are added, via trained vectors
493 (Figure 1a) or intermediate small neural networks (Figure 6a), and then this composed latent signature is
494 passed through the decoder. During the evaluation of the model we utilize paired conditions, where the
495 same perturbation is present in both species/cell lines, and we know the ground truth expression profile
496 in both cases (more details in the Evaluation procedure and metrics section).

497 Global latent space with a jointly and competitively trained classifier

498 In this case, we aim to construct a generally global latent space, but also hide some species or cell line
499 information in some latent variables. We jointly train a simple classifier to correctly predict, in our case
500 study in the L1000, the cell line label. So just an entropy loss is added in the original description of the
501 framework. The construction of a global latent space and the training of the classifier are competing
502 tasks, where the framework is trained to achieve a stable trade-off.

503 CPA-based framework

504 For the CPA-based framework, we utilize the same methodology for building the global latent space and
505 the same basic loss function with some additions.

506 First of all, as described in the original CPA manuscript[34], during training we are iterating between
507 training a classifier (updating only its parameters) in the global latent space, to predict cell type labels
508 and species labels, and training the rest of the framework with the extra addition of adding a penalty
509 ($\boldsymbol{entropy_{adverse}}$) if the classifier can correctly classify species and cell types. This way we are training
510 to update the parameters of the encoders, decoders, trainable vectors, and other classifiers so that the
511 entropy derived from this adverse classifier is maximized. This adverse classifier is initially pre-trained
512 only with encoders and discriminators so that it can already distinguish cell types and species in the
513 global space, making it more difficult to trick it, and thus trying harder to remove the species/cell effect.

514 Then the species and cell type effects are added via trainable vectors, to which is also applied L2-
515 regularization ($\boldsymbol{L2_{trained_{effect}}}$). This vector is trainable for the model to learn the unknown species and
516 cell type effects filtered from the global latent space. In the newly composed latent space, from which
517 the decoders are sampling embeddings, classifiers are jointly trained to correctly classify cell types, and
518 species (or even disease diagnosis).

519 All the above can be summarized in this new loss function:

520 $$Loss = Loss_{basic} - \lambda_{adverse} * entropy_{adverse} + \lambda_{trained_{effect}} * L2_{trained_{effect}}$$

521 ## Framework for the serology datasets

522 In the serology dataset we utilize the CPA-based framework but instead of adding the species effect with
523 trainable vectors, small artificial neural networks are utilized. These do not change the dimensionality of
524 the latent space and consist only of two fully-connected layers. This way they act as signal translators
525 adding species effect in a completely non-linear way. L2-regularization terms for these small ANNs are
526 added to the training loss function.

527 Additionally, it is aimed to later identify features predictive of protection or vaccination status regardless
528 of species. For this reason, we train two classifiers predicting vaccination and protection status in the
529 global latent space, which in other models would be trained in the composed latent space. We care
530 more about protection and thus, as described previously, we aim to create similar embeddings and
531 minimize their distance in the global latent space just by looking at protection status.

532 ## Framework's basic hyperparameters

533 Here we present the basic parameters used to train the model. No thorough hyperparameter tuning was
534 performed, and values were selected based on empirical values and tuned so that there is convergence
535 in the training loss and the training reconstruction performance (Pearson's r and/or R^2). Additionally,
536 these values were also tuned so that the performance in training (not validation) is sufficiently high,

537 meaning that the model is at least able to fit the given data. This empirical tuning was done only based
538 on the 1st training set in 10-fold cross-validation.

539 The latent space dimension was chosen to be as small as possible until the model's performance
540 dropped in both training and validation of only the 1st fold. Based on this latent dimension and the
541 original input dimension of the data the sizes of hidden layers of the encoders were chosen to be in-
542 between, gradually reducing the input dimension to that of the latent space. The actual size and number
543 were constrained by practical memory limits. The decoders had the same number and sizes of hidden
544 layers as those of the encoders, but now they increase the size of the embeddings from the latent
545 dimension to the original input dimension.

546 For using the L1000 978 landmark genes:

547 - Latent dimension = 292
548 - hidden encoder layers dimensions = [640,384]
549 - hidden decoder layers dimensions = [384,640]
550 - cell line classifier (also adverse for CPA) hidden layer dimensions = [256,128,64]
551 - total batch size = 512
552 - number of epochs = 1000
553 - learning rate = 0.001

554 For using the L1000 10,086 genes:

555 - Latent dimension = 1024
556 - hidden encoder layers dimensions = [4096,2048,1024,512]
557 - hidden decoder layers dimensions = [512,1024,2048, 4096]
558 - cell line classifier (also adverse for CPA) hidden layer dimensions = [512,256,128]
559 - total batch size = 512
560 - number of epochs = 1000

- learning rate = 0.001

For using the lung fibrosis datasets:

- Latent dimension = 512
- hidden encoder layers dimensions = [4096,2048]
- hidden decoder layers dimensions = [2048, 4096]
- cell type classifier hidden layer dimensions = [256,128,64,32]
- species classifier hidden layer dimensions = [256,128,64,32]
- fibrosis classifier hidden layer dimensions = [256,128,64,32]
- adverse classifiers hidden layers dimensions = [512,256,128,64]
- total batch size = 1024
- number of epochs = 200
- learning rate = 0.001

For using the serology datasets:

- Latent dimension = 32
- hidden encoder layers dimensions = [64]
- hidden human decoder layers dimensions = [64]
- hidden primate decoder layers dimensions = [128]
- all classifiers hidden layer dimensions = [32,16,8]
- intermediate ANN translators hidden layers dimensions= [32]
- total batch size = 50
- number of epochs = 2000
- learning rate = 0.001

583 ## Evaluation procedure and metrics

584 A part of the data was hidden during training and used to evaluate performance in unseen data, by using
585 10-fold cross-validation splits, where 90% of the data from each system (species or cell line in the case
586 of L1000) are used for training a model and 10% of the data are used for validation. A different 10% of
587 the data is used for validation each time, resulting in all data points being used for validation, where
588 each validation set is not the same in the end. In the L1000 dataset, for evaluating the translation of the
589 whole omics profile, we made sure that for the case of paired conditions, the perturbation in both cell
590 lines was hidden during training.

591 For the classification tasks we simply used the total accuracy and F1-score (micro F1 for multiple
592 categories):

593 $$\mathrm{Accuracy} = \frac{\sum_{i=1}^{K} \mathrm{TP_i} + \sum_{i=1}^{K} \mathrm{TN_i}}{\sum_{i=1}^{K} \mathrm{TP_i} + \sum_{i=1}^{K} \mathrm{TN_i} + \sum_{i=1}^{K} \mathrm{FP_i} + \sum_{i=1}^{K} \mathrm{FN_i}},$$

594 $$\mathrm{F1-micro} = \frac{\sum_{i=1}^{K} \mathrm{TP_i}}{\sum_{i=1}^{K} \mathrm{TP_i} + \frac{1}{2} * \left(\sum_{i=1}^{K} \mathrm{FP_i} + \sum_{i=1}^{K} \mathrm{FN_i}\right)},$$

595 where K is the total number of classes in multi-class classification, TP and FP symbolize true and false
596 positives, and TN and FN symbolize true and false negatives. For the case of multiple classes, we define
597 as positives the samples belonging to that specific class while everything else is a negative sample. Using
598 this definition of positives and negatives we further calculate the TP, FP, TN and FN per class. In the case
599 of cell-type classification in lung fibrosis K=5.
600 For the cell line classification in L1000, species classification both in lung fibrosis and the serology
601 datasets, and vaccination and protection status in the serology dataset we simply use the standard F1
602 score and accuracy for binary classification ($\mathrm{Accuracy} = \frac{\mathrm{TP+T}N}{TP+TN+FP+FN}$, $F1 = \frac{TP}{TP+\frac{1}{2}(FP+FN)}$).

603 To evaluate the validity of the predictions ($\hat{y}$) of whole signatures in translation and reconstruction,
604 compared to the ground truth (y), we utilized:

605 i. the global Pearson's correlation $r(\hat{y}, y) = \frac{\sum(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})}{\sqrt{\sum(y_i-\bar{y})^2 \sum(\hat{y}_i-\bar{\hat{y}})^2}}$, where $\hat{y}$ and y are flattened and the

606 i^{th} element is the ith point in these flattened vectors.

607 ii. the average per sample Spearman's correlation $r_s = \frac{\sum_{i=1}^{N} r(\mathrm{Rank}(\hat{y}),\mathrm{Rank}(y))_i}{N}$, where N is the

608 number of samples and Rank() means ranking the gene based on their differential gene

609 expression and using these ranks to calculate Spearman's correlation.

610 iii. the average per sample sign accuracy $= \frac{\mathrm{TP+TN+TrueZeros}}{\text{total predictions}}$, where TP signifies the genes that

611 have a positive sign regulation both in the actual data and predictions, TN signifies the genes in

612 the sample that have a negative sign regulation both in the actual data and predictions, and

613 TrueZeros are the genes that have an absolute expression $\leq 10^{-6}$ both in the actual data and

614 predictions (we did not choose exactly zero for numerical reasons).

615 For the single-cell RNA-sequencing data where we predict the per gene mean and variance we calculate

616 the R^2 of per gene mean and variance, similar to the original CPA manuscript[34]. In general, R^2 is

617 calculated as: $R^2 = 1 - \frac{RSS}{TSS}, where\ RSS = \sum(\hat{y}_i - y_i)^2\ \ and\ TSS = \sum(y_i - \bar{y})^2$

618 ## Separation of latent space embeddings

619 To evaluate how similar are embeddings of different signatures, and whether there is a separation

620 based on cell, species, or conditions in the latent space we utilize cosine distance, ranging from 0 (the

621 same) to 2 (completely) different: $cosine\ distance = 1 - cosine\ similarity = 1 - \frac{\sum_{i=1}^{d} z_{1,i} z_{2,i}}{\sqrt{\sum_{i=1}^{d} z_{1,i}^2}\sqrt{\sum_{i=1}^{d} z_{2,i}^2}}$,

622 where z_1 a and z_2 are two latent space vectors to be compared and d is the total number of elements in

623 the vector, meaning the latent dimension.

624 To estimate if there is a cell, species, or condition effect, and compare it between the composed and

625 global latent space we utilize Cohen's d to estimate the effect size between the distributions of cosine

626 distances, derived from random pairs of embeddings and pairs coming from the same cell, species, or

condition. The effect size is thus calculated using the mean and standard deviations of two cosine distance (cos) distributions as: $d = \frac{\overline{cos}_1 - \overline{cos}_2}{\sqrt{\frac{\left((n_1-1)s_1^2+(n_2-1)s_2^2\right)}{n_1+n_2-2}}}$, where n_1, n_2 is the number of samples of each of the two distance distributions, $\overline{cos}_1, \overline{cos}_2$ are the means of the cosine distance distributions and s_1,s_2 are the standard deviations of the cosine distance distributions.

A Cohen's d around 0.8 is a large effect size (around 2 is considered a huge effect size) while around 0.5 is a medium effect size, and around 0.2 and below is considered small or very small[58,59].

Feature importance using integrated gradients

To estimate the importance of features we utilize integrated gradients[35] from the Captum library[60].

$$InterGrad_i(x) = (x_i - x_i') \int_{a=0}^{1} \frac{dF\left(x' + a(x - x')\right)}{dx_i} da \, , x' = baseline = 0$$

The importance scores to make a prediction are calculated based on the gradient with respect to the input of the model, and thus the higher the absolute integrated gradient the higher the importance of that input feature to control the output. A negative score means the variable has a negative effect pushing the prediction to the other class, while a positive score has a positive effect.

For example, if we want to identify important latent variables to classify a sample as one coming from a particular cell line we calculate the integrated gradient of every latent variable to make the classification and take the average across all samples. Similarly, if for example, we are aiming to calculate the importance of genes to control latent variables in the global latent space we can calculate the integrated gradient score of every gene for every variable in every sample, and then take the average across samples.

K-means-based separation of important latent variables

Latent variables can be separated into important and unimportant ones using k-means, inspired by an approach that was used to identify important connections between latent components and genes in

649 microbial organisms by using the weights derived from independent component analysis [61,62]. We
650 assume that only 2 main clusters of latent variables exist, one containing important variables and one
651 containing unimportant ones. On this front, the latent variables are clustered based on their absolute
652 gradient scores into 3 clusters, where 3rd cluster is assumed to be a very small cluster of outliers. The
653 midpoint between the variable with the highest score in the unimportant cluster and the variable with
654 the lowest score in the important cluster is used as a threshold to distinguish between significantly
655 important and unimportant latent variables. As a sanity check the important variables are also
656 compared with the top-ranked variables based on their score.

657 ## Likelihood ratio tests for the identification of important latent variables

658 To identify which latent embeddings correlate with viral protection after accounting for vaccination
659 status and species, a likelihood ratio test (LRT) was performed on each individual latent variable. Here,
660 the likelihood (L) of the alternative model (H_A): $latent\ variable_i\ embeddings \sim protection +$
661 $vaccination + species$ was compared to the likelihood of the nested model, or null hypothesis,
662 (H_0): $latent\ variable_i\ embeddings \sim vaccination + species$ in $LRT = -2\ln\left(\frac{L(H_0)}{L(H_A)}\right)$. We rejected
663 H_0 for $latent\ variable_i$ when the FDR-adjusted p-value of the chi-square test was less than 0.05,
664 concluding that the model including protection has a statistically significant better fit than the model
665 without protection. In the volcano plots, the $-\log(pvalue)$ is plotted against the t-value for the
666 protection term in the alternative model. This method assumes that the relationship between the latent
667 variable embeddings and protection is linear. R package *lmtest*[63] (version 0.9.40) was used to perform
668 these statistical tests. Finally, the intersection of these latent variables with significant latent variables
669 (average percentage importance score across folds ≥10%), based on their gradient score from the
670 trained protection classifier, is used for the final identification of robust latent variables associated with
671 viral protection. We keep latent variables that the sign of correlation with protection agrees in both
672 approaches.

673 ## Identification of protection-associated serological features

674 The importance of the serological features is calculated as previously described with the integrated
675 gradient score of every feature for every latent variable that was identified to be statistically significant
676 for predicting viral protection of humans, averaged across samples coming from the respective species.
677 Serological features with high scores (and at least ≥20%) can control latent variables in the global latent
678 space associated with human viral protection, and thus they are predictive of human protection. For
679 human features, we also validate that the univariate differences between protected and unprotected
680 individuals are indeed significant, by using a non-parametric Wilcoxon test, with Bonferroni correction
681 for multiple hypothesis testing.

682 Finally, we calculate the integrated gradient score for translating each non-human primate serological
683 profile to a human profile. The non-human features with high scores the human ones associated with
684 protection, can be considered serological non-human features predictive of human viral protection.

685 ## Transcription factor activity inference

686 To infer the transcription factor activity, we utilized the VIPER algorithm[40] together with the Dorothea
687 Regulon[41]. The VIPER algorithm is testing for regulon enrichment on gene expression signatures, based
688 on a prior knowledge transcription regulatory network. This way based on the expression of
689 downstream genes known to be regulated by a specific transcription factor, the activity of that TF is
690 inferred. The Dorothea regulon contains known regulatory interactions, annotated based on the
691 confidence that this interaction exists. We utilized only interactions of confidence levels A and B.

692 ## Hardware and software specifications

693 All deep learning models were trained using the PyTorch framework[64](version 1.12) in Python version
694 3.8.8. When using the 978 landmark genes and for the serology case study, the models were trained in
695 an NVIDIA GeForce RTX 3060 Laptop GPU with 6 GB of memory. The larger models (using 10,086 genes
696 and the single-cell lung fibrosis data) were trained on the MIT Satori GPU cluster using NVIDIA V100

697 32GB memory GPU cards. Pre-processing and statistical analysis of the results were done in the R
698 programming language version 4.1.2. Visualization of results was done mainly using *ggplot2*[65]. More
699 information about the versions of each library used can be found in the GitHub provided in the Date and
700 code availability section.

701 Data and code availability

702 The study did not produce new experimental data. All analyzed data that were used to train our models
703 and produce all tables and figures, as well as all the code to generate those data, figures and tables are
704 available at https://github.com/NickMeim/OmicTranslationBenchmark .

705 Acknowledgments

706 The authors would like to thank Brian Joughin for his valuable input on this work. We acknowledge
707 funding from MIT-IBM Watson AI Lab, National Institutes of Health (NIH) IMPAcTB contract
708 #75N93019C00071, NIH grant U01-AI67892, US Army Research Office Cooperative Agreement W911NF-
709 19-2-0026.

710 References

711 1. Mak, I. W., Evaniew, N. & Ghert, M. Lost in translation: animal models and clinical trials in cancer
712 treatment. *Am J Transl Res* **6**, 114–118 (2014).
713 2. Brubaker, D. K. & Lauffenburger, D. A. Translating preclinical models to humans. *Science* **367**, 742–
714 743 (2020).
715 3. Rhrissorrakrai, K. *et al.* Understanding the limits of animal models as predictors of human biology:
716 lessons learned from the sbv IMPROVER Species Translation Challenge. *Bioinformatics* **31**, 471–483
717 (2015).
718 4. Shay, T. *et al.* Conservation and divergence in the transcriptional programs of the human and mouse
719 immune systems. *Proceedings of the National Academy of Sciences* **110**, 2946–2951 (2013).

720 5. Gharib, W. H. & Robinson-Rechavi, M. When orthologs diverge between human and mouse. *Briefings*
721 *in Bioinformatics* **12**, 436–441 (2011).
722 6. Niepel, M. *et al.* Common and cell-type specific responses to anti-cancer drugs revealed by high
723 throughput transcript profiling. *Nat Commun* **8**, 1186 (2017).
724 7. Iorio, F. *et al.* Discovery of drug mode of action and drug repositioning from transcriptional
725 responses. *PNAS* **107**, 14621–14626 (2010).
726 8. Iwata, M., Sawada, R., Iwata, H., Kotera, M. & Yamanishi, Y. Elucidating the modes of action for
727 bioactive compounds in a cell-specific manner by large-scale chemically-induced transcriptomics. *Sci*
728 *Rep* **7**, 40164 (2017).
729 9. Fotis, C., Meimetis, N., Sardis, A. & G. Alexopoulos, L. DeepSIBA: chemical structure-based inference
730 of biological alterations using deep learning. *Molecular Omics* **17**, 108–120 (2021).
731 10. Lachmann, A. *et al.* Massive mining of publicly available RNA-seq data from human and mouse.
732 *Nat Commun* **9**, 1366 (2018).
733 11. Wilks, C. *et al.* recount3: summaries and queries for large-scale RNA-seq expression and splicing.
734 *Genome Biology* **22**, 323 (2021).
735 12. Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 Platform and the First
736 1,000,000 Profiles. *Cell* **171**, 1437-1452.e17 (2017).
737 13. Poussin, C. *et al.* The species translation challenge—A systems biology perspective on human
738 and rat bronchial epithelial cells. *Sci Data* **1**, 140009 (2014).
739 14. Seok, J. Evidence-Based Translation for the Genomic Responses of Murine Models for the Study
740 of Human Immunity. *PLOS ONE* **10**, e0118017 (2015).
741 15. Normand, R. *et al.* Found In Translation: a machine learning model for mouse-to-human
742 inference. *Nat Methods* **15**, 1067–1073 (2018).
743 16. Brubaker, D. K., Proctor, E. A., Haigis, K. M. & Lauffenburger, D. A. Computational translation of
744 genomic responses from experimental model systems to humans. *PLOS Computational Biology* **15**,
745 e1006286 (2019).

746 17. Brubaker, D. K. *et al.* Proteogenomic Network Analysis of Context-Specific KRAS Signaling in
747 Mouse-to-Human Cross-Species Translation. *Cell Systems* **9**, 258-270.e6 (2019).
748 18. Brubaker, D. K. *et al.* An interspecies translation model implicates integrin signaling in infliximab-
749 resistant inflammatory bowel disease. *Science Signaling* **13**, eaay3258 (2020).
750 19. Lee, M. J. *et al.* Computational Interspecies Translation Between Alzheimer's Disease Mouse
751 Models and Human Subjects Identifies Innate Immune Complement, TYROBP, and TAM Receptor
752 Agonist Signatures, Distinct From Influences of Aging. *Frontiers in Neuroscience* **15**, (2021).
753 20. Schmidhuber, J. Deep Learning in Neural Networks: An Overview. *Neural Networks* **61**, 85–117
754 (2015).
755 21. Tan, J., Hammond, J. H., Hogan, D. A. & Greene, C. S. ADAGE-Based Integration of Publicly
756 Available Pseudomonas aeruginosa Gene Expression Data with Denoising Autoencoders Illuminates
757 Microbe-Host Interactions. *mSystems* **1**, e00025-15 (2016).
758 22. Wang, D. & Gu, J. VASC: Dimension Reduction and Visualization of Single-cell RNA-seq Data by
759 Deep Variational Autoencoder. *Genomics, Proteomics & Bioinformatics* **16**, 320–331 (2018).
760 23. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-
761 cell transcriptomics. *Nat Methods* **15**, 1053–1058 (2018).
762 24. Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. Single-cell RNA-seq denoising
763 using a deep count autoencoder. *Nat Commun* **10**, 390 (2019).
764 25. Lotfollahi, M., Wolf, F. A. & Theis, F. J. scGen predicts single-cell perturbation responses. *Nat*
765 *Methods* **16**, 715–721 (2019).
766 26. Chen, L., Cai, C., Chen, V. & Lu, X. Learning a hierarchical representation of the yeast
767 transcriptomic machinery using an autoencoder model. *BMC Bioinformatics* **17**, S9 (2016).
768 27. Lotfollahi, M. *et al.* Biologically informed deep learning to query gene programs in single-cell
769 atlases. *Nat Cell Biol* **25**, 337–350 (2023).
770 28. Rampášek, L., Hidru, D., Smirnov, P., Haibe-Kains, B. & Goldenberg, A. Dr.VAE: improving drug
771 response prediction via modeling of drug perturbation effects. *Bioinformatics* **35**, 3743–3751 (2019).

772 29. Way, G. P. & Greene, C. S. Extracting a biologically relevant latent space from cancer
773 transcriptomes with variational autoencoders. in *Biocomputing 2018* 80–91 (WORLD SCIENTIFIC,
774 2017). doi:10.1142/9789813235533_0008.
775 30. Xie, R., Wen, J., Quitadamo, A., Cheng, J. & Shi, X. A deep auto-encoder model for gene
776 expression prediction. *BMC Genomics* **18**, 845 (2017).
777 31. Umarov, R., Li, Y. & Arner, E. DeepCellState: An autoencoder-based framework for predicting
778 cell type specific transcriptional states induced by drug treatment. *PLOS Computational Biology* **17**,
779 e1009465 (2021).
780 32. Escolano, C., Costa-jussà, M. R. & Fonollosa, J. A. R. (Self-Attentive) Autoencoder-based
781 Universal Language Representation for Machine Translation. Preprint at
782 https://doi.org/10.48550/arXiv.1810.06351 (2018).
783 33. Mohiuddin, T. & Joty, S. Unsupervised Word Translation with Adversarial Autoencoder.
784 *Computational Linguistics* **46**, 257–288 (2020).
785 34. Lotfollahi, M. *et al.* Predicting cellular responses to complex perturbations in high-throughput
786 screens. *Molecular Systems Biology* **n/a**, e11517 (2023).
787 35. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic Attribution for Deep Networks. in *Proceedings of*
788 *the 34th International Conference on Machine Learning* 3319–3328 (PMLR, 2017).
789 36. Strunz, M. *et al.* Alveolar regeneration through a Krt8+ transitional stem cell state that persists
790 in human lung fibrosis. *Nat Commun* **11**, 3559 (2020).
791 37. Habermann, A. C. *et al.* Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial
792 and mesenchymal lineages in pulmonary fibrosis. *Science Advances* **6**, eaba1972 (2020).
793 38. Chung, A. W. *et al.* Dissecting Polyclonal Vaccine-Induced Humoral Immunity against HIV Using
794 Systems Serology. *Cell* **163**, 988–998 (2015).
795 39. Barouch, D. H. *et al.* Evaluation of a mosaic HIV-1 vaccine in a multicentre, randomised, double-
796 blind, placebo-controlled, phase 1/2a clinical trial (APPROACH) and in rhesus monkeys (NHP 13-19).
797 *The Lancet* **392**, 232–243 (2018).

798 40. Alvarez, M. J. *et al.* Functional characterization of somatic mutations in cancer using network-
799 based inference of protein activity. *Nat Genet* **48**, 838–847 (2016).
800 41. Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D. & Saez-Rodriguez, J. Benchmark and
801 integration of resources for the estimation of human transcription factor activities. *Genome Res.* **29**,
802 1363–1375 (2019).
803 42. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer
804 drug sensitivity. *Nature* **483**, 603–607 (2012).
805 43. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional
806 profibrotic macrophage. *Nat Immunol* **20**, 163–172 (2019).
807 44. Ramachandran, P. *et al.* Resolving the fibrotic niche of human liver cirrhosis at single-cell level.
808 *Nature* **575**, 512–518 (2019).
809 45. Ching, T. *et al.* Opportunities and obstacles for deep learning in biology and medicine. *Journal of*
810 *The Royal Society Interface* **15**, 20170387 (2018).
811 46. Wysocka, M., Wysocki, O., Zufferey, M., Landers, D. & Freitas, A. A systematic review of
812 biologically-informed deep learning models for cancer: fundamental trends for encoding and
813 interpreting oncology data. *BMC Bioinformatics* **24**, 198 (2023).
814 47. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional
815 Transformers for Language Understanding. *arXiv.org* https://arxiv.org/abs/1810.04805v2 (2018).
816 48. Vaswani, A. *et al.* Attention is all you need. in *Advances in neural information processing systems*
817 5998–6008 (2017).
818 49. Theodoris, C. V. *et al.* Transfer learning enables predictions in network biology. *Nature* 1–9
819 (2023) doi:10.1038/s41586-023-06139-9.
820 50. [clue.io]. https://clue.io/.
821 51. Ioffe, S. & Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing
822 Internal Covariate Shift. in *Proceedings of the 32nd International Conference on Machine Learning*
823 448–456 (PMLR, 2015).

52. Rasamoelina, A. D., Adjailia, F. & Sinčák, P. A Review of Activation Function for Artificial Neural Network. in *2020 IEEE 18th World Symposium on Applied Machine Intelligence and Informatics (SAMI)* 281–286 (2020). doi:10.1109/SAMI48414.2020.9108717.

53. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* **15**, 1929–1958 (2014).

54. Belghazi, M. I. *et al.* MINE: Mutual Information Neural Estimation. Preprint at https://doi.org/10.48550/arXiv.1801.04062 (2021).

55. Hjelm, R. D. *et al.* Learning deep representations by mutual information estimation and maximization. Preprint at https://doi.org/10.48550/arXiv.1808.06670 (2019).

56. Sun, F.-Y., Hoffmann, J., Verma, V. & Tang, J. InfoGraph: Unsupervised and Semi-supervised Graph-Level Representation Learning via Mutual Information Maximization. Preprint at https://doi.org/10.48550/arXiv.1908.01000 (2020).

57. Fotis, C. *et al. DeepSNEM: Deep Signaling Network Embeddings for compound mechanism of action identification*. 2021.11.29.470365 https://www.biorxiv.org/content/10.1101/2021.11.29.470365v1 (2021) doi:10.1101/2021.11.29.470365.

58. Sawilowsky, S. New Effect Size Rules of Thumb. *Journal of Modern Applied Statistical Methods* **8**, (2009).

59. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. (Routledge, 2013).

60. Kokhlikyan, N. *et al.* Captum: A unified and generic model interpretability library for PyTorch. Preprint at https://doi.org/10.48550/arXiv.2009.07896 (2020).

61. McConn, J. L., Lamoureux, C. R., Poudel, S., Palsson, B. O. & Sastry, A. V. Optimal dimensionality selection for independent component analysis of transcriptomic data. *BMC Bioinformatics* **22**, 584 (2021).

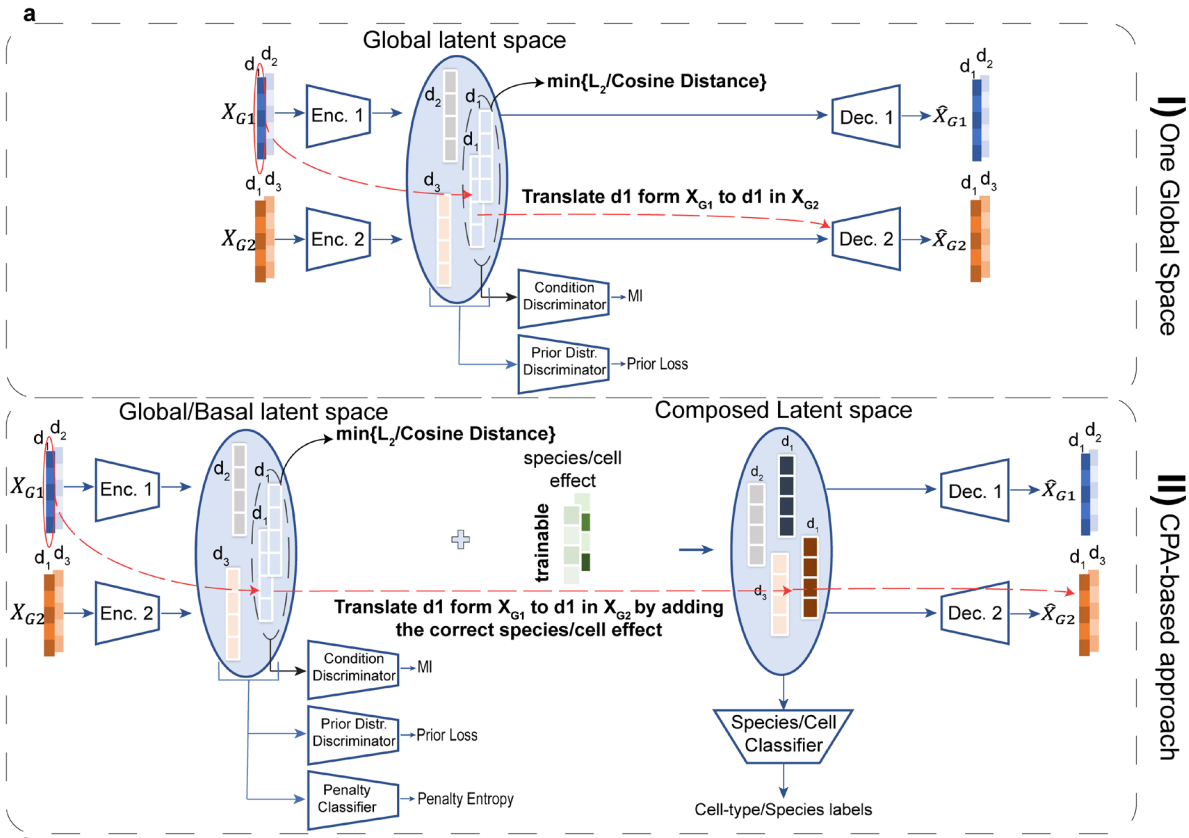
62. Sastry, A. V. *et al.* Mining all publicly available expression data to compute dynamic microbial transcriptional regulatory networks. 2021.07.01.450581 Preprint at https://doi.org/10.1101/2021.07.01.450581 (2021).

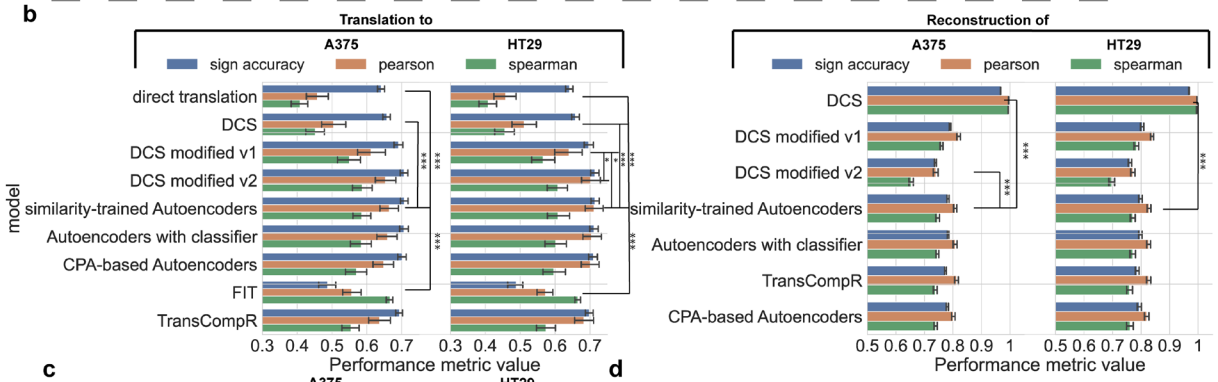
63. Torsten Hothorn, A. Z. Diagnostic Checking in Regression Relationships. *R News* **2**, 7–10 (2002).

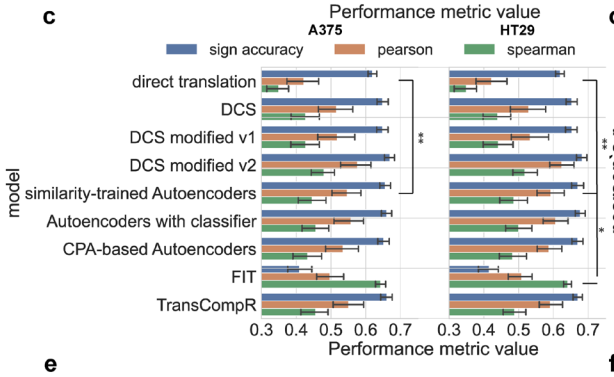
64. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. in *Advances in Neural Information Processing Systems* vol. 32 (Curran Associates, Inc., 2019).

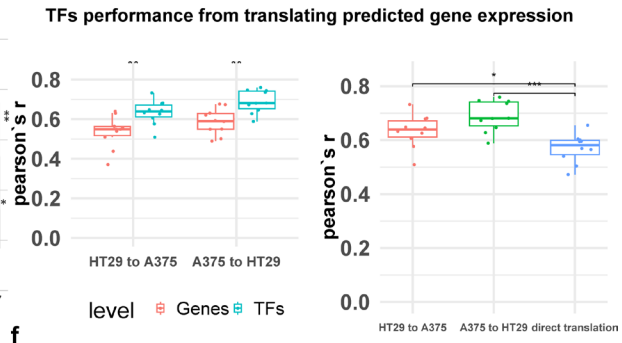
65. Villanueva, R. A. M. & Chen, Z. J. ggplot2: Elegant Graphics for Data Analysis (2nd ed.). *Measurement: Interdisciplinary Research and Perspectives* **17**, 160–167 (2019).

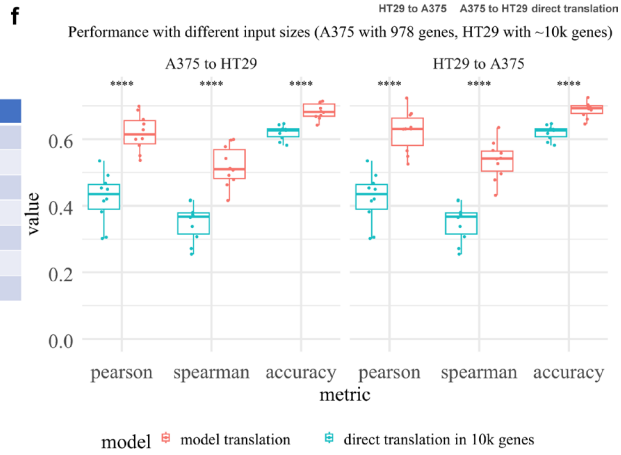
Figures





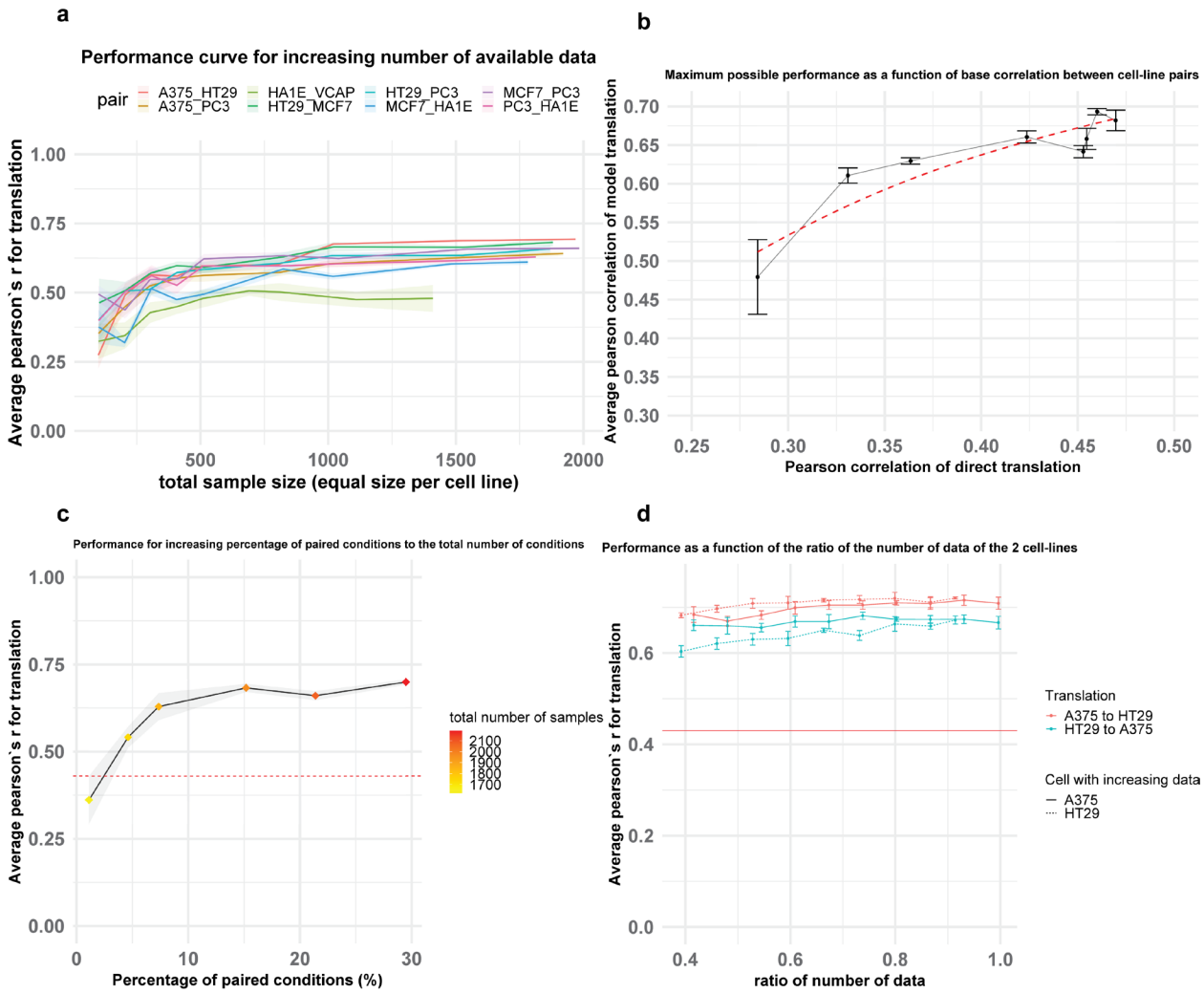






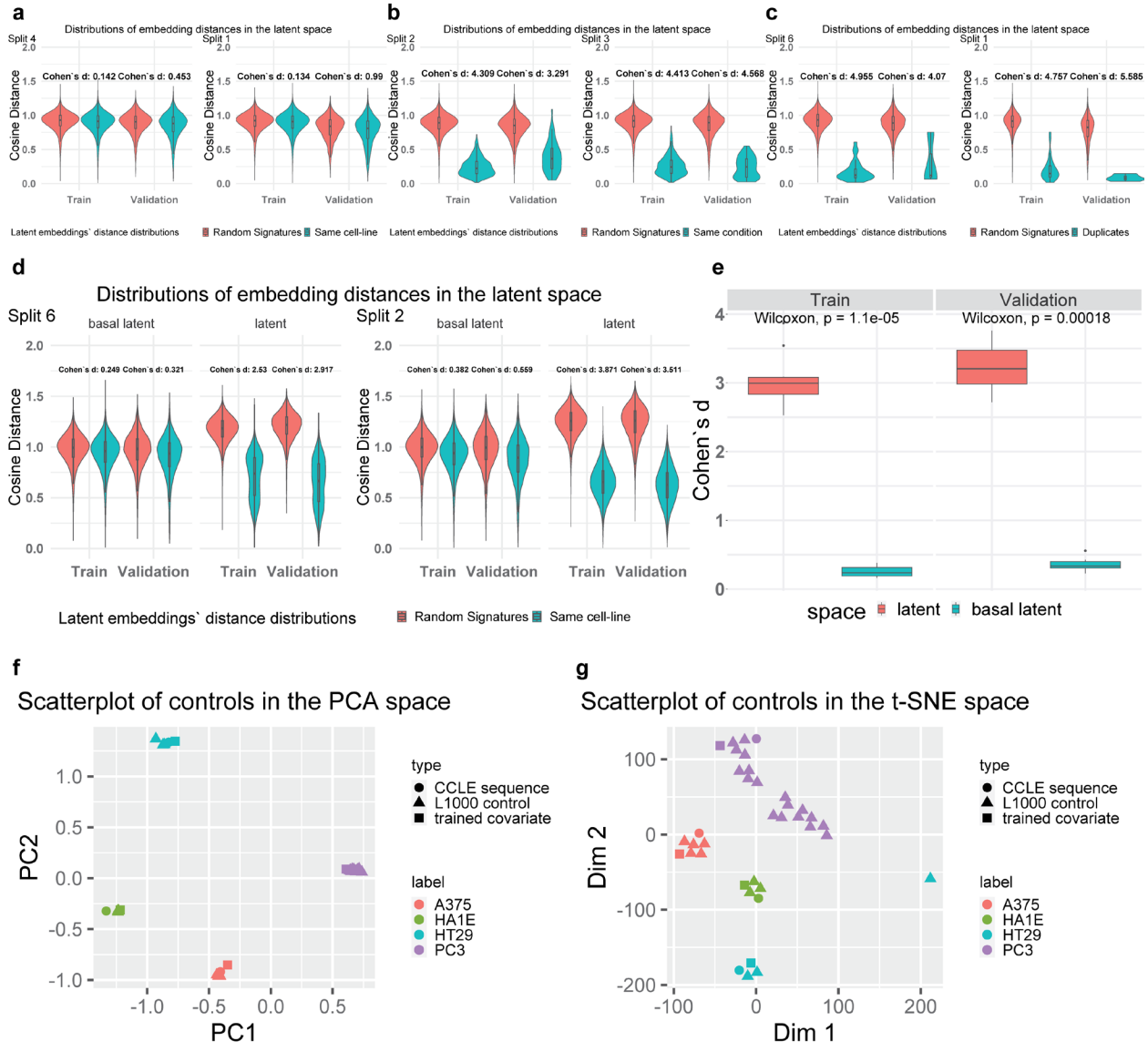
Approach	Accuracy (%)	F1 score (%)
Gene expression classifier	95.00±0.96	95.15±0.96
PCA space classifier	96.83±0.60	96.85±0.62
TransCompR transformed classifier	99.83±0.09	99.82±0.09
DCS global embeddings classifier	92.22±0.41	92.40±0.45
Global Latent space classifier	84.70±0.79	85.05±0.77
Autoencoders joint with classifier	**99.62±0.16**	**99.61±0.17**
CPA-based Autoencoders	**100±0**	**100±0**

861 **Figure 1: a)** Framework architecture main variations: I) One global space is constructed by mapping omic
862 profiles in a space where the distance between embeddings coming from the same perturbation is
863 minimized. II) CPA-based architecture where the latent space is separated into two, one global devoid of
864 species/cell effect and a composed latent space. **b)** Model performance in reconstructing and translating
865 gene expression profiles between the two cell lines with the most common perturbations in the L1000
866 dataset, A375 and HT29, by using only the 978 measured landmark genes. For DCS modified v1-v2 see
867 Supplementary Methods 2.2-2.3. It is worth noting that DCS modified v2 has a distance term and a
868 direct translation term in its training loss. **c)** Model performance in reconstructing and translating gene
869 expression profiles between A375 and HT29 by using all 10,086 genes that are either measured or
870 belong to those that are well-inferred computationally. **d)** Performance in inferring transcription factors
871 activity by using the translated/predicted gene expression. **e)** Performance in correctly classifying cell
872 lines in different cases. **f)** Performance by using different inputs in the L1000.



873

874 **Figure 2: a)** Performance in the translation task of the CPA-based approach across different cell-line
875 pairs and different sizes of training data. **b)** Model performance in translation as a function of the initial
876 similarity of 2 cell lines. **c)** Model performance in translation for different percentages of paired
877 conditions. **d)** Model performance in translation for low-to-medium cell-line imbalance in the
878 conditions of the training samples.



<u>Figure 3:</u> The two splits in 10-fold cross-validation that are shown each time here are the ones where the maximum and minimum difference between the two distributions is observed. For every other split, the difference is between these two extreme cases. Additionally, panels a-c come from the first variation of the model with one global latent space while the rest come from the CPA-based approach **a)** Cosine distance between embeddings coming from random pairs of signatures and pairs coming from the same cell line. **b)** Cosine distance between embeddings coming from random pairs of signatures and pairs coming from the same condition tested on a different cell line **c)** Cosine distance between embeddings coming from random pairs of signatures and pairs being biological duplicates **d)** Cosine distance between embeddings coming from random pairs of signatures and pairs coming from the same cell-line in the global and then the composed latent space in the CPA-based approach. **e)** Cohen's d between distributions of cosine distances between random pairs of embeddings and embeddings coming from the same cell distribution. **f-g)** 2D-Visualization of L1000 control conditions, untreated cell lines from the CCLE dataset, and the trainable vectors of the CPA-based framework containing the cell line basal effect added to perturbations.

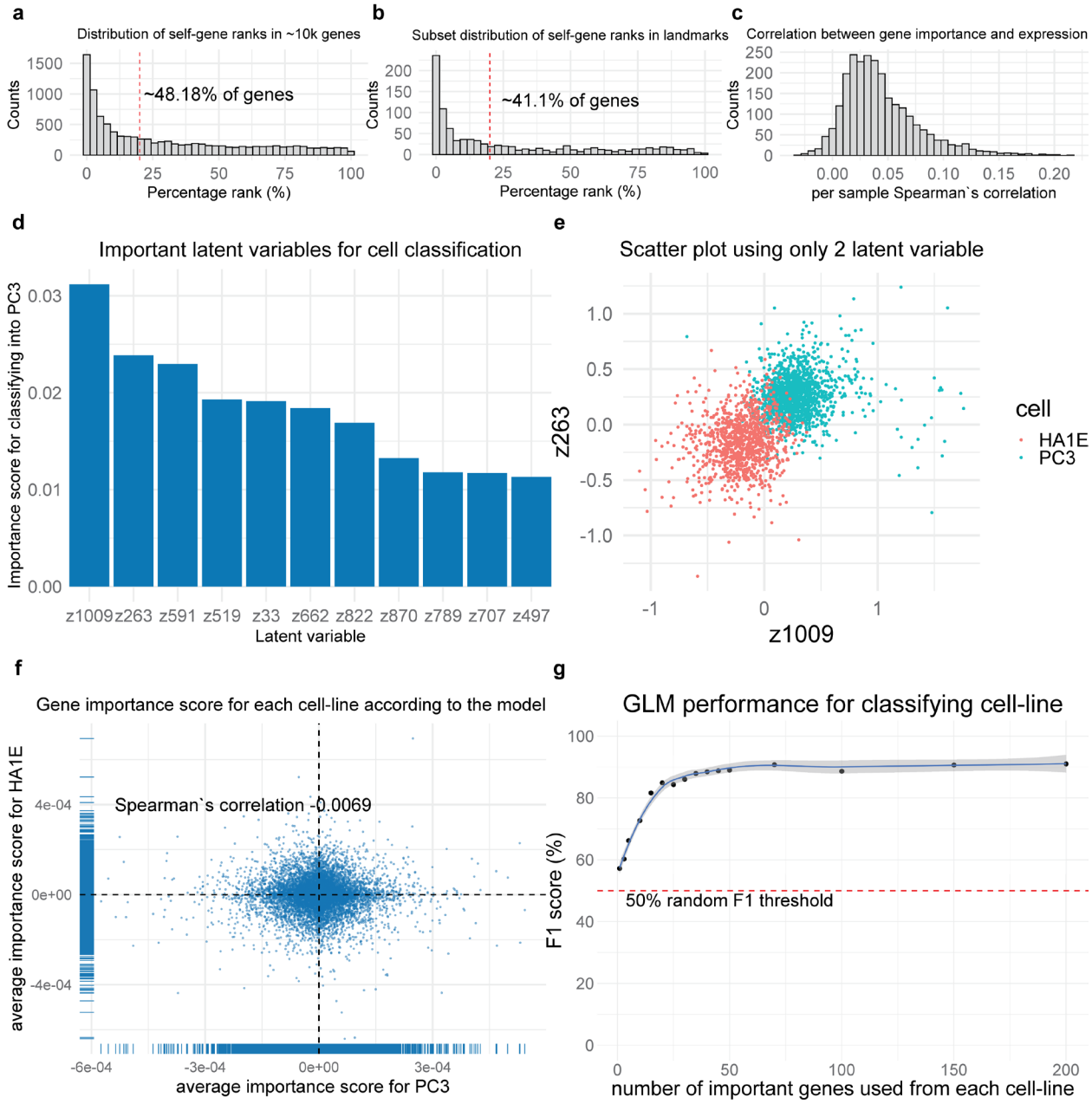
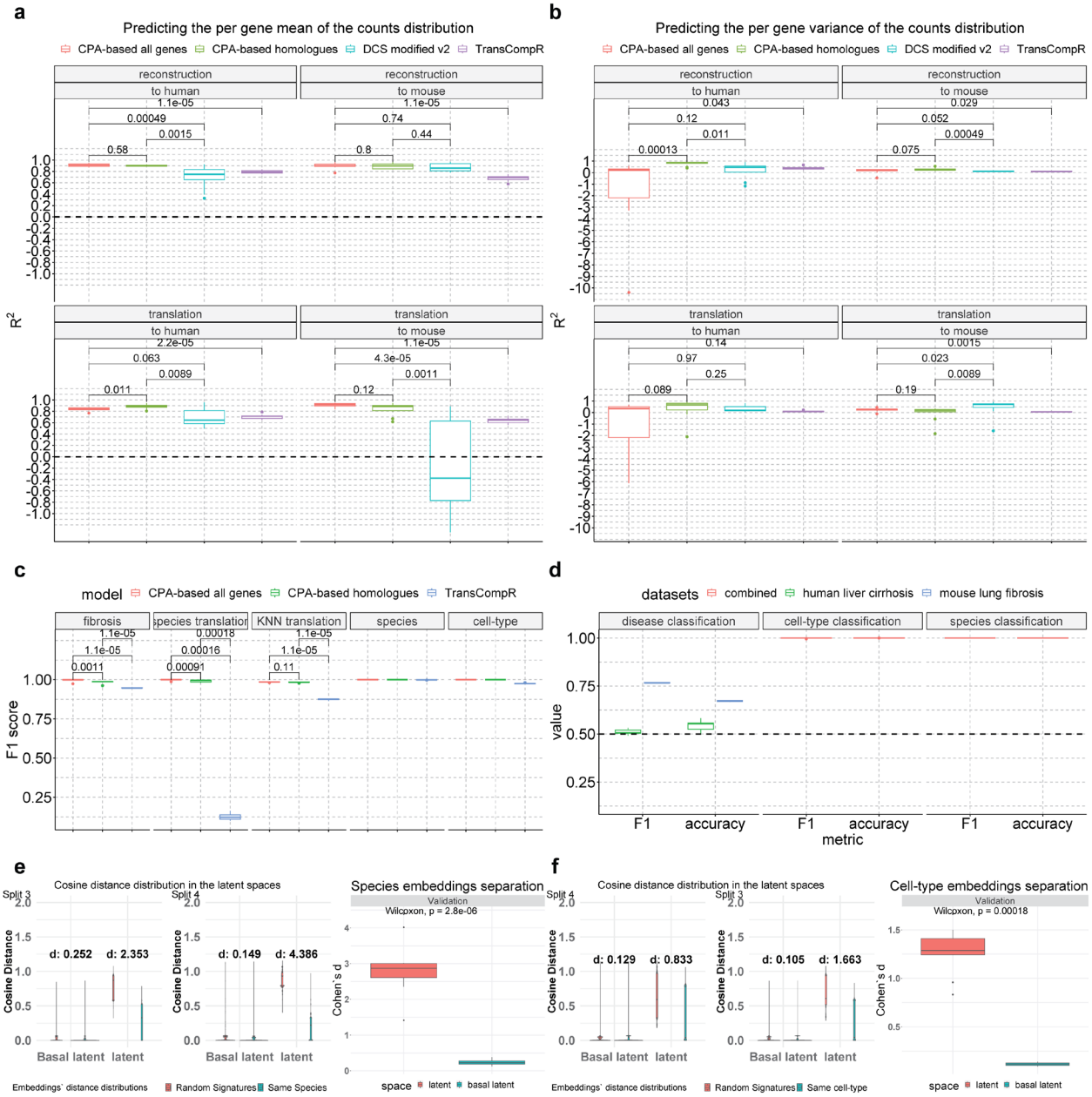


Figure 4: **a-b)** Distribution of percentage rank in terms of the importance of a gene to translate its expression to itself, using the 10,086 genes and the 978 landmark genes respectively in the L1000 dataset. **c)** Per sample Spearman's correlation between absolute gene importance scores and absolute gene expression. **d)** Important latent variable to classify a sample as PC3 or HA1E in the global latent space, when a classifier is jointly trained. **e)** Separation of cell lines based on the top 2 most important latent variables according to the classifier. **f)** Average importance scores of genes from PC3 to control cell-specific latent variables versus the importance scores from HA1E, according to the individual encoders. **g)** Generalized Linear model classification performance by using increasingly more important genes.



<u>Figure 5:</u> a-b) Performance (R^2) in predicting the per gene mean and variance of single-cell RNA-sequencing data for the tasks of reconstruction and species translation in the human-mouse lung fibrosis datasets. The comparison is done between the CPA-based framework using all genes or homologs and TransCompR **c)** Classification performance comparison in different tasks. **d)** Classification performance of the CPA-based framework using all genes in external disease datasets. **e)** Embeddings separation based on species in the global latent space versus the composed latent space. The effect size d is calculated as Cohen's d. **f)** Embeddings separation based on the cell type in the global latent space versus the composed latent space. The effect size d is calculated as Cohen's d.

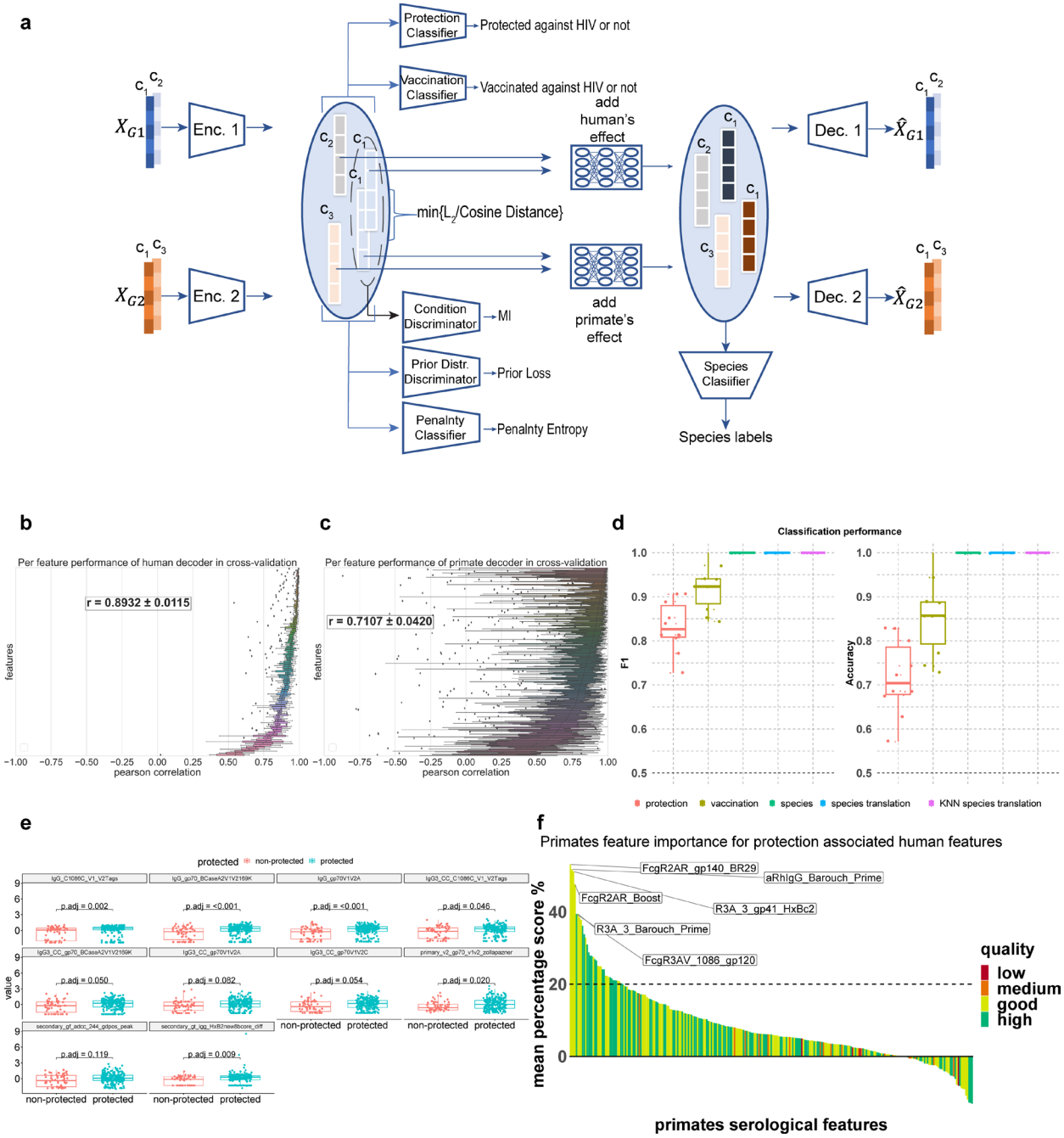


Figure 6: a) Framework architecture for inter-species translation in the serology dataset. Instead of adding the species effect with a trained vector, it is adding completely non-linear with 2 small intermediate neural networks. **b)** Per feature Pearson correlation across in 10-fold cross-validation for human's features. **c)** Per feature Pearson correlation across in 10-fold cross-validation for non-human primates' features. **d)** Classifiers' performance in various tasks. **e)** Univariate differences between protected and non-protected individuals, for human serological features related to viral protection. **f)** Non-human primate features predictive of human protection, by using importance score for translating them into human features. Their quality is assigned based on the average reconstruction Pearson's r of these features in 10-fold cross-validation: i) low quality = 0<r<0.25, ii) medium quality = 0.25≤r<0.5, iii) good quality = 0.5≤r<0.75, iv) high quality = r≥0.75