Autoencoder Model for Translating Omics Signatures

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

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

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

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

4) 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, thus 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 comparison 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 a number of 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

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

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

59 Regarding cellular models Umarov, Li, and Arner developed DeepCellState[31], an autoencoder model
60 based on DeepFake technology[32], that can predict the transcriptional profile of a cell type after drug
61 perturbations based on the behavior of another cell type. However, this approach still requires a 1-1
62 mapping of molecular features, as perturbations from both models or species would need to pass
63 through the same part of the model. Also of note, the model captures a global cell representation based
64 on unsupervised learning of disentangled stimuli representations, which is an idea that has been
65 recently challenged[33]. Both of these limitations suggest that an approach, similar to those used in
66 language translation autoencoder-based models, which create a global language representation[34,35],
67 might improve biological inter-systems translation.

68 Relevant to the idea of building a global latent space for biological tasks, recently it was demonstrated in
69 an interpretable framework called compositional perturbation autoencoder (CPA)[36], how autoencoders
70 can be used to construct a basal latent space devoid of covariate and perturbational signals, capturing
71 only the basal cell state in single-cell RNA sequence data. CPA can be used to generate *in-silico*

72 transcriptional profiles at the single-cell level for different drug doses, cell types, time points, and
73 species when mapped to respective ortholog genes.

74 In this study we build upon ideas presented in the CPA approach, together with ideas from language
75 translation models[34,35] to develop an ANN framework that utilizes separate autoencoders for each
76 biological system, enabling the mapping of samples into a global cell or species space, while providing
77 feature importance estimates for various predictive tasks. The model in its basic form is trained to
78 simultaneously minimize the reconstruction error of the input omic profile and the distance between
79 samples coming from the same condition or perturbation in the global latent space. Our framework is
80 benchmarked, using the latest version of the L1000 dataset[12], against the established linear approaches
81 of TransCompR and FIT[15] (a linear regression model), as well as the neural network approach of
82 DeepCellState, which all require 1-1 feature mapping or input data of similar dimensions. We
83 demonstrate that our approach outperforms FIT and DeepCellState, while there is not a statistically
84 significant difference when comparing with TransCompR in cellular models where 1-1 mapping exists.
85 Additionally, we present several variations of the model and we illustrate the adaptability of our
86 framework by applying it to data of varying omics type and sample size. Furthermore, we demonstrate
87 its biological interpretability, an aspect that deep learning models often struggle to attain, by using an
88 integrated gradients approach[37]. To analyze the performance of the model in inter-species translation
89 we used a mouse[38] and a human[39] lung fibrosis single-cell RNA sequencing datasets, as well as two
90 small-scale serology datasets from humans and non-human primates vaccinated against HIV and SHIV,
91 respectively[40,41]. The latter serves as a novel case study of computational inter-species translation where
92 no 1-1 mapping between features exists. After building the model, we identified molecular features in
93 non-human primates that are predictive of protection against HIV in humans, without analogous
94 features necessarily being present in human data. These findings demonstrate that features derived
95 from this approach can be predictive of the phenotypic profile of another cellular or animal model
96 without requiring them to be homologs, allowing us to maximize the amount of information we can

97 capture from different model systems to advance both our understanding of complex human disease
98 biology and potentially develop more effective therapeutics and vaccines.

99 Results

100 A flexible Artificial Neural Network framework for omics translation

101 The framework is adaptive can be modified depending on the task at hand, however, it contains some
102 core elements and functionalities. The framework (see Methods for details) consists of two separate
103 ANN autoencoders for each biological system (e.g. cell line, species), while there are two main variations
104 of a latent space module which ensures that it removes the system-specific effect of perturbations and
105 constructs stimuli specific global latent space common for both autoencoders (Figure 1a). The
106 separation of the autoencoders allows for the utilization of molecular features without 1-1 mapping
107 between different systems. The first latent module consists of only one global latent space, created by
108 maximizing the similarity and Mutual Information (estimated via an ANN discriminator) of embeddings
109 coming from signatures that represent the same condition/perturbation in a different species or cell
110 line, while another prior discriminator forces the latent variables' distribution to be normal. The second
111 main latent module is based on the recently published compositional perturbation autoencoder (CPA)[36],
112 where there are two separated latent spaces: i) a global/basal latent space and ii) a composed latent
113 space. The global latent space is shaped the same way as in the first main module with the addition of
114 an extra discriminator/classifier which attempts to remove the cell-line or species effect by penalizing
115 the model when this classifier can distinguish between different cells or species, as in the original CPA.
116 In the composed latent space, a cell/species classifier is jointly trained to ensure there is a cell/species
117 effect, which is added through a trainable covariate vector or even completely non-linearly with two
118 intermediate ANNs. Finally, we introduced a novel functionality where feature importance can be
119 identified for various tasks based on integrated gradients[37] (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 dataset[12] provides an ideal transcriptomics dataset for benchmarking different approaches for translating gene expression profiles between different established 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 to hide cell-line specific information in a few latent variables (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, similar to any autoencoder framework. The metrics used to evaluate the performance are: i) Pearson's correlation between predicted gene expression and ground truth gene expression, ii) the per sample Spearman's rank correlation, and iii) the accuracy in predicting correctly the sign of regulation of a gene. For visualization purposes only statistically significant comparisons between our first main variation and other techniques, in terms of Pearson's r, are shown (the p-values for all these comparisons across all metrics are provided in Supplementary Tables 1-4) When utilizing the 978 landmark genes measured in the L1000, all of our framework's variations provide a statistically significant increase in performance compared to the direct translation across all metrics (Figure 1b). Our main method outperforms FIT[15] and the basic DeepCellState[31] (DCS) method when translating from the HT29 cell line to A375, while when translating in the reverse direction our framework outperforms even the variations of DCS (Figure 1b left). An interesting note is that the 2nd variation of DCS also enforces a similarity in the latent space when training the model, and since it outperforms the basic DCS, this indicates the importance of enforcing similarity in the global latent space via some distance metric. The basic DCS approach while taking a hit in the task of translation greatly outperforms the other approaches in terms of simple

145 reconstruction of the input within the same cell line. Nevertheless, our approach performs comparably
146 with other established methods in reconstruction (Figure 1b right). The other variations of our
147 framework perform comparably, even in the case of a global latent space where the cell line effect is not
148 completely removed (when training jointly a classifier). When using 10,086 genes (measured landmarks
149 and computationally well-inferred genes) all approaches take a significant hit in performance but still
150 perform better than the baseline while there is no statistically significant difference between our
151 approach and other state-of-the-art approaches (Figure 1c). Our approach only outperforms FIT in one
152 direction of translation. Finally, our approach performs better than direct translation also in the case of
153 using different genes as input for each cell-line (Figure 1f). The performance is comparable to that using
154 the same genes for both cell lines, indicating the potential to later extend the method in cases where no
155 1-1 mapping exists.

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

157 In the case of using all ~10k genes, there is a decrease in performance (Figure 1b-1c), but it is possible
158 that the predicted values of these genes can still be summarized to infer the activity of transcription
159 factors (TFs) or other gene sets. Indeed by inferring transcription factor activity using the predicted gene
160 expression values with the VIPER algorithm[42] and the Dorothea regulon[43] (Methods) the model has
161 increased performance compared with using the ~10k genes and comparable to that in the case of the
162 landmark genes, while it still performs better than the baseline (Figure 1d). For the case of other gene
163 sets by performing Gene Set Enrichment Analysis (Supplementary Method 1, Supplementary Figure 3)
164 this is true only for small well-defined gene sets like KEGG pathways. The increased performance in
165 inferring only TFs activity might be a result of the previously shown ability of autoencoders to capture
166 regulatory relationships between genes[26,31].

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

168 To establish the ability of the model to retrieve cell-line-specific information in different variations we
169 evaluated the accuracy and F1 score in identifying the cell-line from which a sample is derived (Figure
170 1e). The performance of ANN classifiers trained on the original gene expression data and the PCA-
171 transformed data serves as the baseline. All classifiers are trained using the 10,086 genes of the L1000,
172 but the results in 978 landmark genes are similar (Supplementary Table 11). First, classifiers trained with
173 pre-trained embeddings from DCS or our framework with one global latent space are expected to have
174 lower performance, as embeddings coming from these approaches generate embeddings with the cell-
175 line effect being filtered as much as possible. Indeed, the performance is lower than the baseline and
176 our framework seems to generate more global embeddings than DCs making it more difficult to later
177 retrieve cell-line information. Interestingly when jointly training a classifier in the global latent space we
178 can outperform the baseline (Figure 1e) while the cell-line effect is still partially filtered in the higher
179 dimensions and embeddings coming from the same condition are close to each other regardless of
180 coming from different cells (Supplementary Figure 4). The CPA-based model in the composed latent
181 space perfectly classifies cell lines with, an even suspicious, 100% accuracy. The similarity of input gene
182 expression data in the validation sets, as well as the latent space embeddings similarity, is generally low
183 (Supplementary Figure 5) and no information leakage leads to this kind of performance. The CPA-based
184 framework can create very well-separated cell-line-specific regions (Supplementary Figure 6) in the
185 composed latent space, indicating the framework's ability to shape the latent spaces with robust cell-
186 line-specific regions and explaining the observed accuracy.

187 ## Framework’s performance analysis

188 Due to the CPA-based approach's ability to shape the latent space and its classification performance, we
189 investigated the factors affecting the performance of this method, despite its slightly lower performance
190 in translation compared to that of a single global latent space. The framework has similar behavior and
191 performance for other cell-line pairs, as the original A375/HT29 pair used to initially benchmark the

192 approach against other models (Figure 2a). Some cell-line pairs perform slightly worse, as the original
193 correlation between the same perturbations in the cell-line pair correlates with the model's
194 performance and sets an upper bound in the model's performance (Figure 2b). Additionally, for all cell
195 lines, ~600 total training samples are enough to train a high-performance model (Figure 2a).
196 Interestingly, the amount of paired conditions required to train the framework can be low, as even with
197 ~10-15% of the samples being paired is enough to shape a global latent space to facilitate translation
198 (Figure 2c). Finally, it seems the model is not affected by low to medium imbalance between conditions
199 coming from each cell line (Figure 2d). The same analysis for using all 10,086 genes shows similar trends
200 and results (Supplementary Figure 7).

201 Evaluation of latent space embeddings

202 The validity of the global latent embeddings (by using the 978 landmark genes) produced from our
203 framework is evaluated based on three main tasks (Figure 3a-3c): i) embeddings of pairs coming from
204 the same cell line should not be more similar to each other compared to embeddings from random
205 pairs, ii) embeddings of pairs coming from the same condition regardless of cell-line should be similar,
206 compared to embeddings from random pairs, and iii) embeddings of pairs that consist biological
207 duplicate signatures (meaning same drug tested on the same cell-line, with the same dose and duration)
208 should similar to each other compared to embeddings from random pairs. These tasks are evaluated
209 using cosine distance to quantify how close are embeddings coming from two gene expression
210 signatures. No cell-line effect is observed in the global latent space, both for training and validation
211 embeddings (Figure 3a, Supplementary Figure 8), proving that the framework succeeds in creating a
212 global cell-line space. Embeddings coming from the same condition are closer to each other than
213 embeddings coming from random pairs (Figure 3b), while biological duplicates are slightly even closer
214 (Figure 3c), validating that indeed we have constructed a stimuli-specific global latent space. Similar
215 patterns can be observed in the global latent space also when using the CPA-based approach (Figure
216 3d), while in the composed latent space where a cell-line effect should be visible, embeddings of pairs

217 coming from the same cell-line are closer in the latent space than random (Figure 3d). Cohen's d is used
218 to quantify the difference between the distributions of cosine distances across all folds (Figure 3e),
219 proving that indeed there is a much higher cell-line effect in the latent space than the effect in the global
220 latent space (not zero though).

221 Framework's trained parameters that capture biological information

222 To better interpret the framework, we investigate the biological information captured by some parts of
223 it. Since only the cell-line effect is filtered in the global latent space in the CPA-based framework then
224 the trainable covariate vectors should only add a basal cell-specific effect. Intuitively this would mean
225 that in the case of gene expression coming from untreated cells (control perturbations), the global latent
226 embeddings should capture a "zero"/basal cell state, and thus the trained covariate which is added to
227 that global representation should be similar to the composed latent space vectors which now capture
228 both the cell line effect. On this front, we used high-quality control samples from the L1000 dataset not
229 seen by the model during training, as well as samples coming from sequenced untreated cell lines from
230 the Cancer Cell Line Encyclopedia[44] (CCLE), pre-processed to use only genes present in the L1000
231 landmark genes and z-scored like the gene expression profiles used in the L1000 dataset. Additionally,
232 for this case study two models trained completely separately were trained: the original benchmark
233 model of A375/HT29 cell lines and another model using the PC3 prostate cancer cell line and the HA1E
234 normal epithelial cell line. The latter pair was chosen to be used in the subsequent analysis because of
235 both the comparably high performance of the model (Figure 2a) and a comparable amount of paired
236 conditions (330 in PC3/HA1E vs 406 in A375/HT29), while these two cell lines are significantly different
237 in terms of biology. Each trained covariate, even though the models were trained separately, is observed
238 to be closer to its respective cell-line control signatures, even those coming from an external dataset,
239 both when using PCA for dimensionality reduction (Figure 3f), where clearly defined cell-line specific
240 regions are observed, as well as when using t-SNE (Figure 3g).

241 ## Identification of important features that capture task-specific characteristics

242 The framework can identify both latent variables and genes, important for tasks of interest, enabling the
243 investigation of the framework's ability to identify important features that can be biologically
244 interesting. PC3 and HA1E cell lines were chosen, as described previously, to train a model with one
245 global latent space and one classifier trained jointly to classify the cell lines from which the samples
246 were derived (contradictory learning tasks). To identify the importance of input variables of different
247 parts of the model w.r.t their output, an integrated gradient-based approach[37] was utilized (see
248 Methods), attributing an importance score to each variable of interest. It is interesting since the same
249 genes are used for both cell lines, to identify what the model considers important to translate, in the
250 case of paired conditions, a gene expression profile from one cell line to the same set of genes in
251 another cell line. For approximately, 48% of the 10,086 genes, the model considers a different gene than
252 itself as important (important means in the top 20% genes based on their absolute importance score) to
253 be translated in the other cell line for the same condition (Figure 4a). For the case of landmarks genes,
254 this percentage slightly drops to 41% (Figure 4b). Since the cell lines chosen are one cancerous and one
255 non-cancerous this is an interesting result indicating that the model potentially avoids the fallacy of
256 using the same gene as a proxy for its gene expression across quite different biological systems.
257 Additionally, Spearman's correlation between the absolute importance scores and the absolute gene
258 expression across different samples is ~0.04 on average (Figure 4c), proving that the model doesn't
259 consider important just genes that are highly regulated. Subsequently, the jointly trained classifier can
260 identify a subset of important latent variables of the global latent space for classifying samples. Even
261 though the cell line effect is generally filtered and embeddings coming from the same condition are
262 close to each other globally (Supplementary Figure 4) the model can still identify 11 latent variables
263 (overlapping with the top-ranked variables based on absolute gradient score) that are important for
264 classifying a cell line (Figure 4d), by using a k-means based approach together with the calculated
265 importance scores (see Methods). These latent variables can greatly separate the latent vectors based

266 on the cell line (Figure 4e), even though globally the cell line-specific separation of embeddings is
267 minimized. Genes considered important by the encoders to control these latent variables should be
268 either cell line-specific genes or a subset of genes that can easily distinguish between cell lines. The
269 importance scores of the genes for each cell line-specific encoder do not correlate at all and are
270 different between the two cell lines (Figure 4f). It is possible to even train a very simple generalized
271 linear model to classify cell lines based on gene expression, only using a subset of these important
272 genes, achieving high performance and observing a saturation with only ~25-30 genes from each cell
273 line (Figure 4g).

274 Performance in inter-species translation for lung fibrosis

275 To evaluate the framework in inter-species translation, we utilize the raw gene counts coming from
276 single-cell RNA-sequencing of a mouse[38] and human[39] dataset with lung fibrosis. Based on the original
277 CPA manuscript [36], the decoders predict a mean and a variance for every gene, derived from a negative
278 binomial distribution. Furthermore, both a trainable species vector and another trainable cell type
279 vector are added to the global space, where it is attempted to filter both species and cell type effect. We
280 compare the CPA-based framework using all genes available in the dataset and using only the homologs,
281 as well as the best performing (in the L1000) variation of DCS and TransCompR. We evaluate the
282 performance in 10-fold cross-validation in terms of R^2 of the mean and variance in the reconstruction of
283 gene expression profiles, as well as in translating between mouse and human, where we would expect
284 to observe a similar distribution in a successful translation, and thus mean and variance. Our framework
285 outperforms the other approaches in terms of R^2 of the means both in reconstruction and translation
286 (Figure 5a). In predicting the variance there is not always a statistically significant improvement, while all
287 approaches have generally low performance (Figure 5b). Finally, there is not a significant difference
288 between using all genes or just homologs (Figure 5a-5b). Additionally, the CPA-based approaches and
289 TransCompR, when a classifier is trained, are compared in terms of classifying fibrosis, species, and cell
290 type. Cell type and species are also classified when a signature is translated into another species in the

291 composed latent space. Using all genes in our framework outperforms homolog genes approaches in the
292 fibrosis and species-translation tasks, while the performance of all approaches is high (Figure 5c). Similar
293 to what was observed in the L1000 dataset, species and cell type are predicted perfectly in our
294 framework. Additionally, both the species and cell-type effects are filtered (Figure 5e-5f, Supplementary
295 Figures 9-10) in the global latent space compared to the composed latent space. Finally, it is worth
296 noting that based only on the human lung fibrosis dataset, out of the top 10 genes contributing to the
297 top principal components, 3 of them are not homologs (Supplementary Figure 11). Regardless of
298 performance, a method considering only homologues would exclude important genes for lung fibrosis.

299 Generalization in other disease datasets

300 It is useful to investigate the extent to which the model can make predictions about disease, species,
301 and cell types in other datasets, used as external test sets, as well as different tissue and disease
302 datasets. On this front, an external mouse lung fibrosis[45] and human liver cirrhosis[46] dataset were used.
303 In the mouse dataset, even though different genes were measured than those in our model, the
304 performance is still decent in disease classification, and better than the random 50% (Figure 5d). In the
305 human dataset, which is an extreme case of fibrosis in a different organ, the model has significantly
306 lower performance but is still barely higher than random (Figure 5d). Interestingly, in both cases the
307 model can still perfectly identify cell types and species (Figure 5d), proving once again the ability of the
308 model to capture a very general species and cell type effect.

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

310 We aimed to develop a framework for species translation that is still viable in the case where no 1-1
311 mapping between molecular features exist, a limitation of existing cross-species models. For this
312 purpose, we utilized non-human primate and human serology datasets[40,41] following vaccination against
313 SHIV and HIV, respectively. In this special case, the species effect was added completely non-linearly
314 during training, with 2 small intermediate neural networks (Figure 6a). This modification made no

315 significant difference in performance and in the properties of the global space, both in the case of the
316 L1000 and these serological datasets (Supplementary Figures 12-14). We reason that adding more non-
317 linear parts in the model is more appropriate for species translation in this setting, even if performance
318 is not increasing, due to few data points. The framework is trained so that protected individuals are
319 close in the global latent space regardless of species, while two separate classifiers try to predict
320 protection and vaccination outcomes (Figure 6a). In the composed latent space, a classifier predicts
321 species. The overall performance of the framework in 10-fold cross-validation in reconstructing the
322 input features is high (Figure 6b-6c). For the human serology features the performance is very high (r =
323 0.8932±0.0115), while in non-human primates some features are not predicted well and there is a big
324 variation in performance between folds, the overall performance is still good (r = 0.7107±0.0420).
325 Finally, the performance across all classification tasks is exceptionally high (Figure 6d), with achieving
326 once again a perfect 100% accuracy in species classification and translation.

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

328 With our framework, we aimed to identify features predictive of human protection, even non-human
329 ones, by combining the integrated gradient approach (used in the L1000) with likelihood ratio tests (LRT)
330 performed on each latent variable (see Methods for more details). The likelihood of an alternative
331 model accounting for protection to derive the value of a latent variable is compared against a null model
332 not accounting for protection. Latent variables with an FDR-adjusted p-value<0.05 are considered
333 significant and the sign of the statistic signifies the directionality of the association with protection
334 (Supplementary Figure 16). For the integrated gradient approach, the integrated gradient of each latent
335 variable to predict protection was calculated for each human sample, according to the classifier in the
336 global latent space. Then the average score is used to signify importance. The intersection of latent
337 variables between the two approaches is used to identify the final latent variables contributing to
338 human protection (Supplementary Figure 17). Latent variables are denoted as important only if there is

339 an agreement between the likelihood ratio test results and the integrated gradients (and gradient
340 score≥10%).

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

349 Discussion

350 Here we developed an ANN framework that facilitates the translation of different omics profiles
351 between different biological systems, by building upon ideas introduced in the CPA approach[36] and
352 other species and cellular translation methods[13,15,18,31], together with ideas from language translation
353 models[34]. The explicit central goal here is to align omics signatures between species, rather than
354 identifying what information inherent in a signature in one species is most germane for understanding
355 phenotype characteristics in the other (which has been the objective in much of our recent work[16–19]).

356 The framework performs as well as (or even better in some cases) other state-of-the-art translation
357 techniques, when using homolog features, while it does not take a significant hit in performance when it
358 utilizes different features or non-homologs to represent two species or cell lines. Notably the framework
359 can construct a truly global latent space and manipulate it successfully to create stimuli-specific regions,
360 with high-performing classifiers trained jointly to make predictions for various tasks such as disease
361 diagnosis.

362 Current approaches[13–18] require homolog features and they utilize linear transformations to respectively
363 build their common latent space or translate signatures, thus being able to better approximate only the
364 linear inter-species relationships. Even the non-linear neural network approaches like DeepCellState[31]
365 require homology of the molecular features used to describe biological systems, without shaping the
366 latent space in some semi-supervised way. Additionally, approaches like FIT require a large number of
367 paired conditions to achieve high performance. Our framework could fill these gaps by non-linearly
368 modeling relationships between different biological systems, without requiring any kind of homology,
369 while it achieves high performance with only a small percentage of paired conditions used to construct
370 and tune the global latent space. This enables for the first time the training of a framework with
371 serology datasets where no 1-1 mapping of the features between the two biological models exists. This
372 way the identification of relationships between very different molecular profiles that correlate with
373 specific phenotypes, e.g. protection against infection, is possible, through the interpretation of the
374 framework.

375 Interpretability of deep learning models in biology remains a challenge and such models are often
376 criticized due to poor understanding of the biological relationships captured by the models[47,48]. On this
377 front, we demonstrate in our framework how integrated gradient approaches[37] can be used to estimate
378 the importance of features used by different parts of the framework for various tasks and decisions,
379 enabling to some extent the interpretation of the model. We were able to propose serological features
380 predictive of human protection against HIV, regardless of species. Furthermore, we explored the
381 biological information captured in the framework.

382 Different parts of the framework can capture and retrieve successfully the effect of species or cell types,
383 filtered from the global latent space, composing that way a new latent space that perfectly separates
384 species and cell types. This can explain the ability of the framework to predict, with very high
385 performance, general cell types (unfortunately not with high granularity regarding cell subtypes) and
386 species, even in external disease datasets, derived even from different organs/tissues, enabling this way

387 successful translation. However, there are still limitations in the generalization of the approach in
388 external datasets.

389 The performance in external disease datasets drops significantly as we move to more different
390 pathologies and tissues. Even for the same disease, different features measured can lead to reduced
391 performance regarding predictions made about disease diagnosis. The more features contained in the
392 molecular profile of a sample the more difficult is for the model to reconstruct this profile, while as
393 observed in the serology dataset not all features' predictions can be trusted. Even though our
394 framework can be trained even with fewer samples, still it contains a lot of parameters, especially when
395 using a larger number of molecular features, leading to overfitting, while training such large models is
396 not computationally trivial.

397 Some of these shortcomings can be alleviated by combining our framework with larger datasets, such as
398 ARCHS4[10], which contains the majority of publicly available RNA-sequencing data from humans and
399 mice. The Gene Expression Omnibus (GEO)[49] contains thousands of datasets, derived from different
400 types of molecular features, opening even the potential to utilize our framework for combining different
401 modalities. Training the framework with more data, more diverse unique conditions, and even training
402 to translate simultaneously between multiple systems and not just pairs, may enable higher
403 generalization and higher granularity in modeling different biological covariates, e.g. more specific cell
404 types.

405 Finally, the flexibility and adaptability of the framework allows it to be combined with other approaches,
406 to model many different biological systems. This could potentially lead to prior computational
407 optimization of biological systems and assays aiming to model human pathology. Features that are
408 informative and can explain or predict human biology, without requiring homology, can be identified
409 and optimized, while it is possible to use the framework to explore potential transcriptional
410 modifications to design better disease models, capturing the human phenotype in the clinic.

411 # Methods

412 ## Pre-processing *in-vitro* transcriptomics benchmark dataset

413 The L1000 CMap resource[12] contains bulk gene expression data from drug perturbations across different
414 cell lines and provides a benchmark dataset with diverse conditions and a large sample size.
415 Additionally, a sufficient number of equivalent perturbations across different biological systems is
416 available (termed *paired conditions* for the rest of the manuscript) to evaluate the performance in
417 translating omics profiles. We selected high-quality drug perturbations from the latest version of the
418 L1000 dataset (accessed via clue.io). The level 5 z-score transformed and pre-processed differential gene
419 expression data of 978 landmark genes, measured with the L1000 assay, and additionally, 9,196
420 computationally well-inferred genes in the CMap resource, were considered in the subsequent analysis.
421 High-quality perturbations consist of signatures with more than three replicates, where at least half of
422 them passed standard quality control protocols in the assay, and were not identified as statistical
423 outliers. Additionally, for multiple-signature perturbagens (technical duplicates), only the signature with
424 the highest transcriptional activity score (TAS) across these technical duplicates was kept in the dataset.
425 These signatures are called exemplars and are specifically designated for further analysis in the CLUE
426 platform[50]. TAS is a metric inherent to the L1000 dataset that quantifies signal strength and
427 reproducibility and is calculated as the geometric mean of the number of differentially expressed
428 transcripts and the 75th quantile of pairwise replicate correlations. Finally, the ability to distinguish
429 between random signatures and true biological duplicates, meaning the same perturbagen tested on
430 the same cell line for the same duration and dosage, was evaluated for different parts of the dataset,
431 split using varying TAS thresholds (Supplementary Figure 19) and it was decided to keep samples with
432 TAS≥0.3. For the case of control signatures, we followed the same procedure without filtering based on
433 the TAS number.

434 ## Pre-processing single-cell RNA sequencing interspecies datasets

435 For the human and mouse single-cell RNA-sequencing datasets, first, each cell annotated was
436 summarized into one of the four many cell types: i) immune cells, ii) mesenchymal cells, iii) epithelial
437 cells, iv) endothelial cells, and iv) stem cells. These were used later to filter cell effect from the global
438 latent space and these are the categories used in subsequent cell-type classification. Finally, while the
439 raw gene counts are used for reconstruction from the decoder in the loss function, the encoders take as
440 input the log-transformed counts ($x_{input} = \log_{10}(count + 1)$).

441 ## Pre-processing serology dataset preprocessing

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

450 ## The general framework and the training procedure

451 Generally, the framework always consists of two independent encoders and two independent decoders
452 for each of the species or cell lines in the pair attempting translation, while inside some latent module, a
453 global latent space is shaped (Figure 1a). Both the encoders and the decoders are multi-layer fully
454 connected neural networks, with each layer consisting sequentially of a fully-connected layer, a batch
455 normalization layer, an ELU activation function, and a dropout layer. Only their final output layer
456 consists of only one fully-connected layer without a bias term.

457 For the construction of the global latent space, in every variation of the framework, the distance
458 ($L_{distance}$) between embeddings of profiles coming from the same perturbation is minimized, their
459 cosine similarity (L_{cosine}) and mutual information (L_{MI}) is maximized, and finally the divergence of the
460 distribution of the latent variables from a normal distribution is also minimized (L_{prior}). Both cosine
461 similarity and Euclidean distance losses were added to enforce the strongest possible filtering of species
462 and cell type effect, while the cosine similarity also enforces normalization of the latent embeddings.
463 The last two tasks mentioned above are achieved via two different estimated ANN discriminators, based
464 on the Deep InfoMax and InfoGraph approaches[51,52], where the Jensen-Shannon Mutual Information
465 between embeddings coming from the same perturbation is estimated and the extra prior loss is added,
466 according to the following equation:

467 $$L = (L_{MI}) + (L_{prior}) = \left(-\frac{1}{N1 + N2} \sum_{i=1}^{N1} \sum_{j=1}^{N2} D_1(z_i, z_j) \right) + (D_2(Z||N))$$

468 Where Z denotes the global latent space embeddings, D_1 is the discriminator estimating the mutual
469 information between two embeddings and D_2 is the second discriminator which forces the distribution
470 of the features of each learned embedding to be close to the normal distribution N. D_1 consists of 3
471 fully-connected layers, each of them followed by a ReLU function and one fully-connected skip
472 connection. Likewise, D_2 has three similar fully-connected layers and the final scalar output is passed
473 through a sigmoid function. These regularization loss terms ($L_{distance}$, L_{cosine}, L_{MI}) are calculated and
474 averaged across every pair of global embeddings (z_i, z_j) that are coming from the same condition. The
475 L_{prior} is calculated for every sample in the dataset, meaning every global latent embedding and
476 averaged across samples.

477 For the case of the L1000 dataset, similar perturbations are considered the same drug, tested on the
478 same dose and time duration just on similar cell lines. For the lung fibrosis dataset, similar profiles are
479 considered those coming from samples that have the same diagnosis (fibrosis or not). For the serology

480 datasets, we train the framework so that embeddings coming from protected individuals against HIV are
481 close to each other regardless of species (and even vaccination status)

482 The basic task of this autoencoder framework is reconstruction, which is achieved by minimizing some
483 kind of reconstruction loss (L_{recon}). In the case of simple z-scored profiles coming from bulk data, this is
484 done by minimizing the mean sum of squared errors between the input of the encoders and the output
485 of the decoders. The sum or squares error is averaged across samples. For single-cell RNA-sequencing
486 data, based on the implementation proposed in the original CPA manuscript[36] (found here
487 https://github.com/facebookresearch/CPA), the negative binomial negative log-likelihood is used to
488 optimize the reconstruction, by assuming that the data are derived from a negative-binomial
489 distribution. In that case, the decoders are not just predicting one value per molecular feature, but a
490 mean and variance. The negative binomial negative log-likelihood loss is calculated for every sample and
491 the average across all samples in the batch is taken to be minimized.

492 Finally, classifiers, which also consist of multiple fully connected layers and a final SoftMax activation
493 function for the output, are used for different classification tasks. The average entropy loss across
494 samples for every classification task in the latent space (global or not) is minimized.

495 L2-regularization of the weights and bias of the encoders, decoders, and classifiers is also enforced. This
496 is implemented by calculating and minimizing the sum of squares for the aforementioned trainable
497 parameters.

498 All the above can be summarized in the following loss function, which consists of the basic loss used to
499 train the basic format of the framework (for the CPA-based approach see the corresponding section in
500 Methods, and for all other variations see Supplementary Method 2):

501 $$Loss_{basic} = \lambda_{recon} * L_{recon} + \lambda_{distance} * L_{distance} + \lambda_{MI} * L_{MI} + \lambda_{prior} * L_{prior} + \sum_{i=1}^{2} \lambda_{enc,i} *$$
502 $$L2_{encoder,i} + \sum_{i=1}^{2} \lambda_{dec,i} * L2_{decoder,i} + \sum_{i=1}^{M} \lambda_{class,i} * entropy_i - \lambda_{cosine} * L_{cosine}$$

503 Where M is the number of classifiers and thus individual classification tasks.

504 For values for each of the λ used in the loss function see Supplementary Table 12.

505 Translation of omics profiles

506 For translating an omics profile, we first encode it in the global latent space, using the encoder
507 representing the respective biological system (meaning species or cell line). Then the decoder of the
508 other biological system in the pair of the currently trained framework is using directly the embedding of
509 the global latent space to reconstruct the profile in the other system and thus translate the signature. In
510 the case of the CPA-based framework first, the species and cell type effect are added, via trained vectors
511 (Figure 1a) or intermediate small neural networks (Figure 6a), and then this composed latent signature is
512 passed through the decoder. During the evaluation of the model we utilize paired conditions, where the
513 same perturbation is present in both species/cell lines, and we know the ground truth expression profile
514 in both cases (more details in the Evaluation procedure and metrics section).

515 Global latent space with a jointly and competitively trained classifier

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

521 CPA-based framework

522 For the CPA-based framework, we utilize the same tactics for building the global latent space and the
523 same basic loss function with some additions.

524 First of all, as described in the original CPA manuscript[36], during training we are iterating between
525 training classifier (updating only its parameters) in the global latent space, to predict cell type labels and
526 species labels, and training the rest of the framework with the extra addition of adding a penalty

527 ($entropy_{adverse}$) if the classifier can correctly classify species and cell types. This way we are training to
528 update the parameters of the encoders, decoders, trainable vectors, and other classifiers so that the
529 entropy derived from this adverse classifier is maximized. This adverse classifier is initially pre-trained
530 only with encoders and discriminators so that it can already distinguish cell types and species in the
531 global space, making it more difficult to trick it, and thus trying harder to remove the species/cell effect.

532 Then the species and cell type effects are added via trainable vectors, to which is also applied L2-
533 regularization. This vector is trainable for the model to learn the unknown species and cell type effects
534 filtered from the global latent space. In the newly composed latent space, from which the decoders are
535 sampling embeddings, classifiers are jointly trained to correctly classify cell types, and species (or even
536 disease diagnosis).

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

538 $$Loss = Loss_{basic} - \lambda_{adverse} * entropy_{adverse} + \lambda_{trained_{effect}} * L2_{trained_{effect}}$$

539 ## Framework for the serology datasets

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

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

550 ### Framework's basic hyperparameters

551 Here we present the basic parameters used to train the model. For a more detailed overview of all
552 hyperparameters see Supplementary Table 13. No thorough hyperparameter tuning was performed,
553 and values were selected based on empirical values and tuned so that there is convergence in the
554 training loss and the training reconstruction performance (Pearson's r and/or R^2). Additionally, these
555 values were also tuned so that the performance in training (not validation) is sufficiently high, meaning
556 that the model is at least able to fit the given data. This empirical tuning was done only based on the 1st
557 training set in 10-fold cross-validation.

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

565 For using the L1000 978 landmark genes:

566 - Latent dimension = 292
567 - hidden encoder layers dimensions = [640,384]
568 - hidden decoder layers dimensions = [384,640]
569 - cell line classifier (also adverse for CPA) hidden layer dimensions = [256,128,64]
570 - total batch size = 512
571 - number of epochs = 1000
572 - learning rate = 0.001

573 For using the L1000 10,086 genes:

- Latent dimension = 1024
- hidden encoder layers dimensions = [4096,2048,1024,512]
- hidden decoder layers dimensions = [512,1024,2048, 4096]
- cell line classifier (also adverse for CPA) hidden layer dimensions = [512,256,128]
- total batch size = 512
- 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

Evaluation procedure and metrics

A part of the data was hidden during training and used to evaluate performance in unseen data, by using 10-fold cross-validation splits. For evaluating the translation of the whole omics profile, we made sure that for the case of paired conditions, the perturbation in both species or cell lines was hidden during training.

For the classification tasks simply, total accuracy and F1-score (micro F1 for multiple categories) were used:

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

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

Where K is the total number of classes.

To evaluate the validity of the predictions ($\hat{y}$) of whole signatures in translation and reconstruction, compared to the ground truth (y), we utilized: i) the global Pearson's r, ii) per sample Spearman's correlation (r_s), and iii) sign accuracy

$r(\hat{y}, y) = \frac{\sum(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum(y_i - \bar{y})^2 \sum(\hat{y}_i - \bar{\hat{y}})^2}}$, where $\hat{y}$ and y are flattened and the ith element is the ith point in these flattened vectors

$r_s = \frac{\sum_{i=1}^{N} r(\text{Rank}(\hat{y}), \text{Rank}(y))_i}{N}$, where N is the number of samples

$$\text{sign accuracy} = \frac{\text{TP+TN+TrueZeros}}{\text{total predictions}}$$

For the single-cell RNA-sequencing data where we predict the per gene mean and variance we calculate the R^2 of per gene mean and variance, similar to the original CPA manuscript[36]. In general, R^2 is calculated as: $R^2 = 1 - \frac{RSS}{TSS}, where\ RSS = \sum(\hat{y}_i - y_i)^2\ \ and\ TSS = \sum(y_i - \bar{y})^2$

Separation of latent space embeddings

To evaluate how similar are embeddings of different signatures, and whether there is a separation based on cell, species, or conditions in the latent space we utilize cosine distance, ranging from 0 (the same) to 2 (completely) different: $cosine\ distance = 1 - cosine\ similarity = 1 - \frac{\sum_{i=1}^{d} z_{1,i} z_{2,i}}{\sqrt{\sum_{i=1}^{d} z_{1,i}^2}\sqrt{\sum_{i=1}^{d} z_{2,i}^2}}$,

where z_1 a and z_2 are two latent space vectors to be compared and d is the total number of elements in the vector, meaning the latent dimension.

To estimate if there is a cell, species, or condition effect, and compare it between the composed and global latent space we utilize Cohen's d to estimate the effect size between the distributions of cosine 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: $d = \frac{\overline{cos}_1 - \overline{cos}_2}{\sqrt{\frac{\left((n_1-1)s_1^2 + (n_2-1)s_2^2\right)}{n_1+n_2-2}}}$.

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[53,54].

Feature importance using integrated gradients

To estimate the importance of features we utilize integrated gradients[37] from the Captum library[55].

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

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

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

648 K-means-based separation of important latent variables

649 Latent variables can be separated into important and unimportant ones using k-means, inspired by an
650 approach that was used to identify important connections between latent components and genes in
651 microbial organisms by using the weights derived from independent component analysis [56,57]. We
652 assume that only 2 main clusters of latent variables exist, one containing important variables and one
653 containing unimportant ones. On this front, the latent variables are clustered based on their absolute
654 gradient scores into 3 clusters, where 3rd cluster is assumed to be a very small cluster of outliers. The
655 midpoint between the variable with the highest score in the unimportant cluster and the variable with
656 the lowest score in the important cluster is used as a threshold to distinguish between significantly
657 important and unimportant latent variables. As a sanity check the important variables are also
658 compared with the top-ranked variables based on their score.

659 Likelihood ratio tests for the identification of important latent variables

660 To identify which latent embeddings correlate with viral protection after accounting for vaccination
661 status and species, a likelihood ratio test (LRT) was performed on each individual latent variable. Here,

662 the likelihood (L) of the alternative model (H_A): $latent\ variable_i\ embeddings \sim protection +$

663 $vaccination + species$ was compared to the likelihood of the nested model, or null hypothesis,

664 (H_0): $latent\ variable_i\ embeddings \sim vaccination + species$ in $LRT = -2\ln\left(\frac{L(H_0)}{L(H_A)}\right)$. We rejected

665 H_0 for $latent\ variable_i$ when the FDR-adjusted p-value of the chi-square test was less than 0.05,

666 concluding that the model including protection has a statistically significant better fit than the model

667 without protection. In the volcano plots, the $-\log(pvalue)$ is plotted against the t-value for the

668 protection term in the alternative model. This method assumes that the relationship between the latent

669 variable embeddings and protection is linear. R package *lmtest*[58] (version 0.9.40) was used to perform

670 these statistical tests. Finally, the intersection of these latent variables with significant latent variables

671 (average percentage importance score across folds ≥10%), based on their gradient score from the

672 trained protection classifier, is used for the final identification of robust latent variables associated with

673 viral protection. We keep latent variables that the sign of correlation with protection agrees in both

674 approaches.

675 ## Identification of protection-associated serological features

676 The importance of the serological features is calculated as previously described with the integrated

677 gradient score of every feature for every latent variable that was identified to be statistically significant

678 for predicting viral protection of humans, averaged across samples coming from the respective species.

679 Serological features with high scores (and at least ≥20%) can control latent variables in the global latent

680 space associated with human viral protection, and thus they are predictive of human protection. For

681 human features, we also validate that the univariate differences between protected and unprotected

682 individuals are indeed significant, by using a non-parametric Wilcoxon test, with Bonferroni correction

683 for multiple hypothesis testing.

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

687 Transcription factor activity inference

688 To infer the transcription factor activity, we utilized the VIPER algorithm[42] together with the Dorothea
689 Regulon[43]. The VIPER algorithm is testing for regulon enrichment on gene expression signatures, based
690 on a prior knowledge transcription regulatory network. This way based on the expression of
691 downstream genes known to be regulated by a specific transcription factor, the activity of that TF is
692 inferred. The Dorothea regulon contains known regulatory interactions, annotated based on the
693 confidence that this interaction exists. We utilized only interactions of confidence levels A and B.

694 Hardware and software specifications

695 All deep learning models were trained using the PyTorch framework[59](version 1.12) in python version
696 3.8.8. When using the 978 landmark genes and for the serology case study, the models were trained in
697 an NVIDIA GeForce RTX 3060 Laptop GPU with 6 GB of memory. The larger models (using 10,086 genes
698 and the single-cell lung fibrosis data) were trained on the MIT Satori GPU cluster using NVIDIA V100
699 32GB memory GPU cards. Pre-processing and statistical analysis of the results were done in the R
700 programming language version 4.1.2. Visualization of results was done mainly using *ggplot2*[60]. More
701 information about the versions of each library used can be found in the GitHub provided in the Date and
702 code availability section.

703 Data and code availability

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

Acknowledgements

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

References

1. Mak, I. W., Evaniew, N. & Ghert, M. Lost in translation: animal models and clinical trials in cancer treatment. *Am J Transl Res* **6**, 114–118 (2014).
2. Brubaker, D. K. & Lauffenburger, D. A. Translating preclinical models to humans. *Science* **367**, 742–743 (2020).
3. Rhrissorrakrai, K. *et al.* Understanding the limits of animal models as predictors of human biology: lessons learned from the sbv IMPROVER Species Translation Challenge. *Bioinformatics* **31**, 471–483 (2015).
4. Shay, T. *et al.* Conservation and divergence in the transcriptional programs of the human and mouse immune systems. *Proceedings of the National Academy of Sciences* **110**, 2946–2951 (2013).
5. Gharib, W. H. & Robinson-Rechavi, M. When orthologs diverge between human and mouse. *Briefings in Bioinformatics* **12**, 436–441 (2011).
6. Niepel, M. *et al.* Common and cell-type specific responses to anti-cancer drugs revealed by high throughput transcript profiling. *Nat Commun* **8**, 1186 (2017).
7. Iorio, F. *et al.* Discovery of drug mode of action and drug repositioning from transcriptional responses. *PNAS* **107**, 14621–14626 (2010).
8. Iwata, M., Sawada, R., Iwata, H., Kotera, M. & Yamanishi, Y. Elucidating the modes of action for bioactive compounds in a cell-specific manner by large-scale chemically-induced transcriptomics. *Sci Rep* **7**, 40164 (2017).

731 9. Fotis, C., Meimetis, N., Sardis, A. & G. Alexopoulos, L. DeepSIBA: chemical structure-based inference
732 of biological alterations using deep learning. *Molecular Omics* **17**, 108–120 (2021).
733 10. Lachmann, A. *et al.* Massive mining of publicly available RNA-seq data from human and mouse.
734 *Nat Commun* **9**, 1366 (2018).
735 11. Wilks, C. *et al.* recount3: summaries and queries for large-scale RNA-seq expression and splicing.
736 *Genome Biology* **22**, 323 (2021).
737 12. Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 Platform and the First
738 1,000,000 Profiles. *Cell* **171**, 1437-1452.e17 (2017).
739 13. Poussin, C. *et al.* The species translation challenge—A systems biology perspective on human
740 and rat bronchial epithelial cells. *Sci Data* **1**, 140009 (2014).
741 14. Seok, J. Evidence-Based Translation for the Genomic Responses of Murine Models for the Study
742 of Human Immunity. *PLOS ONE* **10**, e0118017 (2015).
743 15. Normand, R. *et al.* Found In Translation: a machine learning model for mouse-to-human
744 inference. *Nat Methods* **15**, 1067–1073 (2018).
745 16. Brubaker, D. K., Proctor, E. A., Haigis, K. M. & Lauffenburger, D. A. Computational translation of
746 genomic responses from experimental model systems to humans. *PLOS Computational Biology* **15**,
747 e1006286 (2019).
748 17. Brubaker, D. K. *et al.* Proteogenomic Network Analysis of Context-Specific KRAS Signaling in
749 Mouse-to-Human Cross-Species Translation. *Cell Systems* **9**, 258-270.e6 (2019).
750 18. Brubaker, D. K. *et al.* An interspecies translation model implicates integrin signaling in infliximab-
751 resistant inflammatory bowel disease. *Science Signaling* **13**, eaay3258 (2020).
752 19. Lee, M. J. *et al.* Computational Interspecies Translation Between Alzheimer's Disease Mouse
753 Models and Human Subjects Identifies Innate Immune Complement, TYROBP, and TAM Receptor
754 Agonist Signatures, Distinct From Influences of Aging. *Frontiers in Neuroscience* **15**, (2021).
755 20. Schmidhuber, J. Deep Learning in Neural Networks: An Overview. *Neural Networks* **61**, 85–117
756 (2015).

757 21. Tan, J., Hammond, J. H., Hogan, D. A. & Greene, C. S. ADAGE-Based Integration of Publicly
758 Available Pseudomonas aeruginosa Gene Expression Data with Denoising Autoencoders Illuminates
759 Microbe-Host Interactions. *mSystems* **1**, e00025-15 (2016).
760 22. Wang, D. & Gu, J. VASC: Dimension Reduction and Visualization of Single-cell RNA-seq Data by
761 Deep Variational Autoencoder. *Genomics, Proteomics & Bioinformatics* **16**, 320–331 (2018).
762 23. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-
763 cell transcriptomics. *Nat Methods* **15**, 1053–1058 (2018).
764 24. Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. Single-cell RNA-seq denoising
765 using a deep count autoencoder. *Nat Commun* **10**, 390 (2019).
766 25. Lotfollahi, M., Wolf, F. A. & Theis, F. J. scGen predicts single-cell perturbation responses. *Nat*
767 *Methods* **16**, 715–721 (2019).
768 26. Chen, L., Cai, C., Chen, V. & Lu, X. Learning a hierarchical representation of the yeast
769 transcriptomic machinery using an autoencoder model. *BMC Bioinformatics* **17**, S9 (2016).
770 27. Lotfollahi, M. *et al.* Biologically informed deep learning to query gene programs in single-cell
771 atlases. *Nat Cell Biol* **25**, 337–350 (2023).
772 28. Rampášek, L., Hidru, D., Smirnov, P., Haibe-Kains, B. & Goldenberg, A. Dr.VAE: improving drug
773 response prediction via modeling of drug perturbation effects. *Bioinformatics* **35**, 3743–3751 (2019).
774 29. Way, G. P. & Greene, C. S. Extracting a biologically relevant latent space from cancer
775 transcriptomes with variational autoencoders. in *Biocomputing 2018* 80–91 (WORLD SCIENTIFIC,
776 2017). doi:10.1142/9789813235533_0008.
777 30. Xie, R., Wen, J., Quitadamo, A., Cheng, J. & Shi, X. A deep auto-encoder model for gene
778 expression prediction. *BMC Genomics* **18**, 845 (2017).
779 31. Umarov, R., Li, Y. & Arner, E. DeepCellState: An autoencoder-based framework for predicting
780 cell type specific transcriptional states induced by drug treatment. *PLOS Computational Biology* **17**,
781 e1009465 (2021).

782 32. Nguyen, T. T. *et al.* Deep learning for deepfakes creation and detection: A survey. *Computer*
783 *Vision and Image Understanding* **223**, 103525 (2022).
784 33. Locatello, F. *et al.* Challenging Common Assumptions in the Unsupervised Learning of
785 Disentangled Representations. in *Proceedings of the 36th International Conference on Machine*
786 *Learning* 4114–4124 (PMLR, 2019).
787 34. Escolano, C., Costa-jussà, M. R. & Fonollosa, J. A. R. (Self-Attentive) Autoencoder-based
788 Universal Language Representation for Machine Translation. Preprint at
789 https://doi.org/10.48550/arXiv.1810.06351 (2018).
790 35. Mohiuddin, T. & Joty, S. Unsupervised Word Translation with Adversarial Autoencoder.
791 *Computational Linguistics* **46**, 257–288 (2020).
792 36. Lotfollahi, M. *et al.* Predicting cellular responses to complex perturbations in high-throughput
793 screens. *Molecular Systems Biology* **n/a**, e11517 (2023).
794 37. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic Attribution for Deep Networks. in *Proceedings of*
795 *the 34th International Conference on Machine Learning* 3319–3328 (PMLR, 2017).
796 38. Strunz, M. *et al.* Alveolar regeneration through a Krt8+ transitional stem cell state that persists
797 in human lung fibrosis. *Nat Commun* **11**, 3559 (2020).
798 39. Habermann, A. C. *et al.* Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial
799 and mesenchymal lineages in pulmonary fibrosis. *Science Advances* **6**, eaba1972 (2020).
800 40. Chung, A. W. *et al.* Dissecting Polyclonal Vaccine-Induced Humoral Immunity against HIV Using
801 Systems Serology. *Cell* **163**, 988–998 (2015).
802 41. Barouch, D. H. *et al.* Evaluation of a mosaic HIV-1 vaccine in a multicentre, randomised, double-
803 blind, placebo-controlled, phase 1/2a clinical trial (APPROACH) and in rhesus monkeys (NHP 13-19).
804 *The Lancet* **392**, 232–243 (2018).
805 42. Alvarez, M. J. *et al.* Functional characterization of somatic mutations in cancer using network-
806 based inference of protein activity. *Nat Genet* **48**, 838–847 (2016).

807 43. Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D. & Saez-Rodriguez, J. Benchmark and
808 integration of resources for the estimation of human transcription factor activities. *Genome Res.* **29**,
809 1363–1375 (2019).
810 44. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer
811 drug sensitivity. *Nature* **483**, 603–607 (2012).
812 45. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional
813 profibrotic macrophage. *Nat Immunol* **20**, 163–172 (2019).
814 46. Ramachandran, P. *et al.* Resolving the fibrotic niche of human liver cirrhosis at single-cell level.
815 *Nature* **575**, 512–518 (2019).
816 47. Ching, T. *et al.* Opportunities and obstacles for deep learning in biology and medicine. *Journal of*
817 *The Royal Society Interface* **15**, 20170387 (2018).
818 48. Wysocka, M., Wysocki, O., Zufferey, M., Landers, D. & Freitas, A. A systematic review of
819 biologically-informed deep learning models for cancer: fundamental trends for encoding and
820 interpreting oncology data. *BMC Bioinformatics* **24**, 198 (2023).
821 49. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and
822 hybridization array data repository. *Nucleic Acids Research* **30**, 207–210 (2002).
823 50. [clue.io]. https://clue.io/.
824 51. Hjelm, R. D. *et al.* Learning deep representations by mutual information estimation and
825 maximization. Preprint at https://doi.org/10.48550/arXiv.1808.06670 (2019).
826 52. Sun, F.-Y., Hoffmann, J., Verma, V. & Tang, J. InfoGraph: Unsupervised and Semi-supervised
827 Graph-Level Representation Learning via Mutual Information Maximization. Preprint at
828 https://doi.org/10.48550/arXiv.1908.01000 (2020).
829 53. Sawilowsky, S. New Effect Size Rules of Thumb. *Journal of Modern Applied Statistical Methods* **8**,
830 (2009).
831 54. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. (Routledge, 2013).

55. 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).

56. 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).

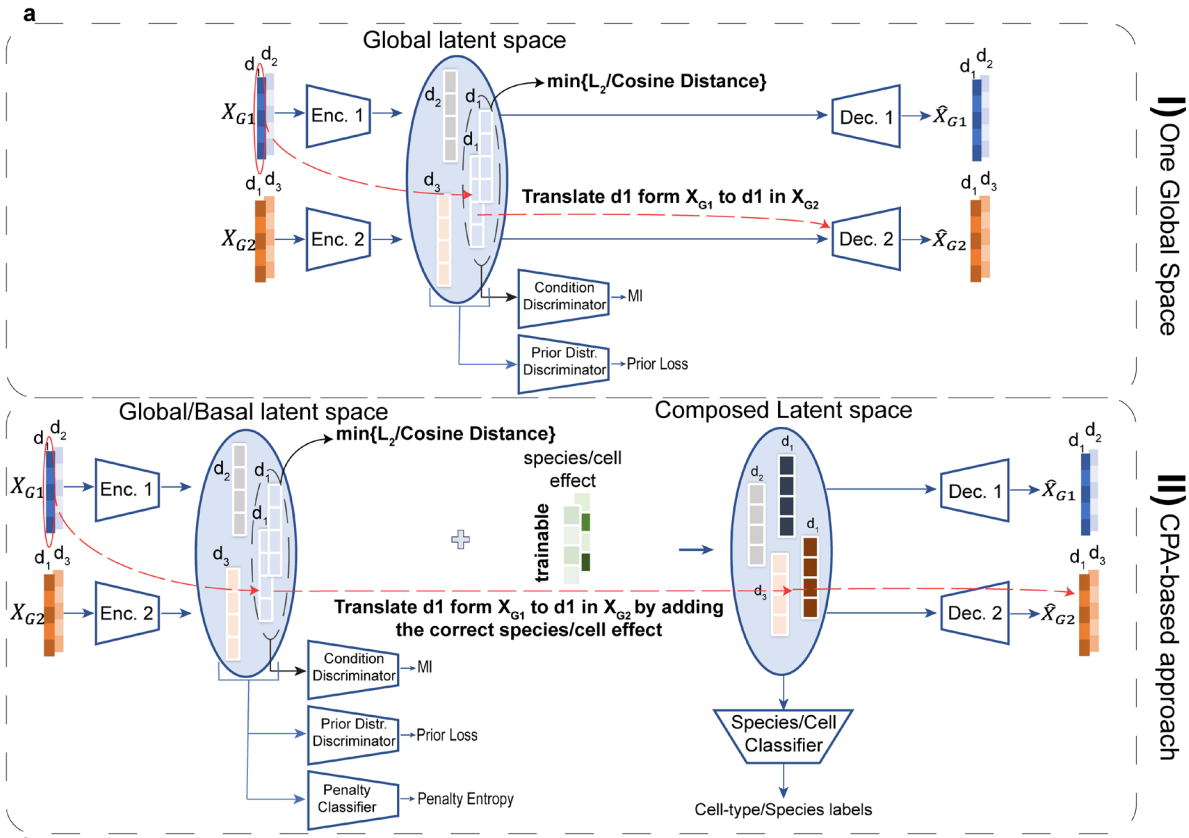
57. Sastry, A. V. *et al.* Mining all publicly available expression data to compute dynamic microbial transcriptional regulatory networks. 2021.07.01.450581 Preprint at https://doi.org/10.1101/2021.07.01.450581 (2021).

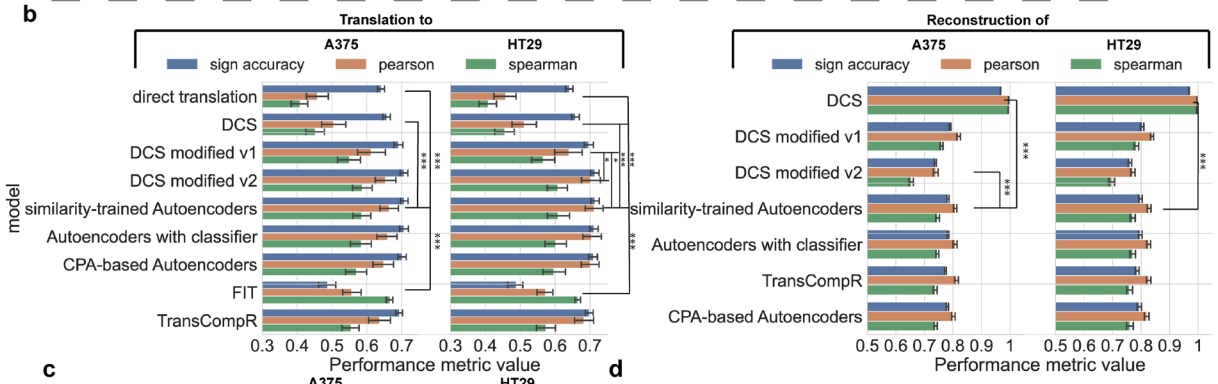
58. Torsten Hothorn, A. Z. Diagnostic Checking in Regression Relationships. *R News* **2**, 7–10 (2002).

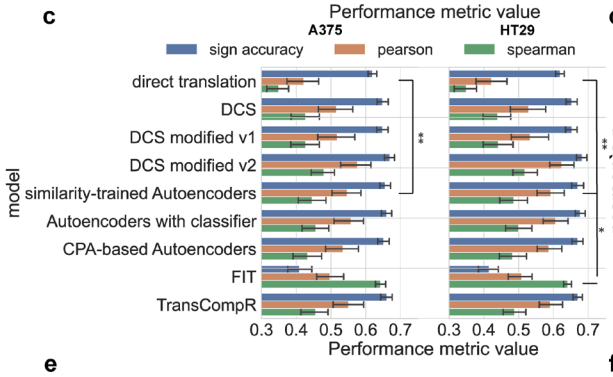
59. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. in *Advances in Neural Information Processing Systems* vol. 32 (Curran Associates, Inc., 2019).

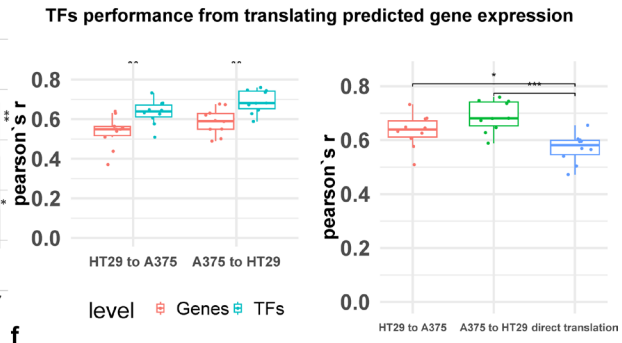
60. Villanueva, R. A. M. & Chen, Z. J. ggplot2: Elegant Graphics for Data Analysis (2nd ed.). *Measurement: Interdisciplinary Research and Perspectives* **17**, 160–167 (2019).

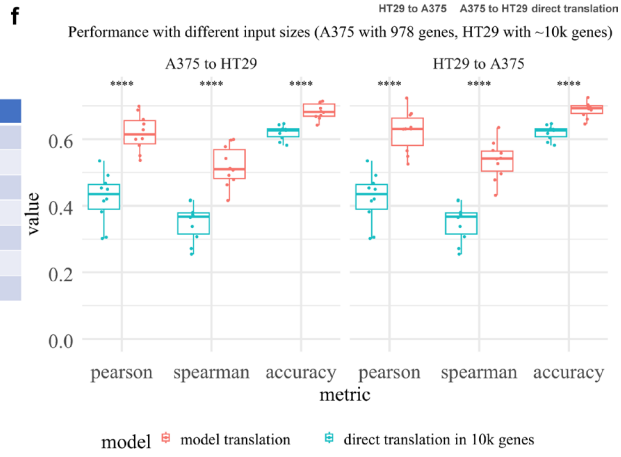
Figures





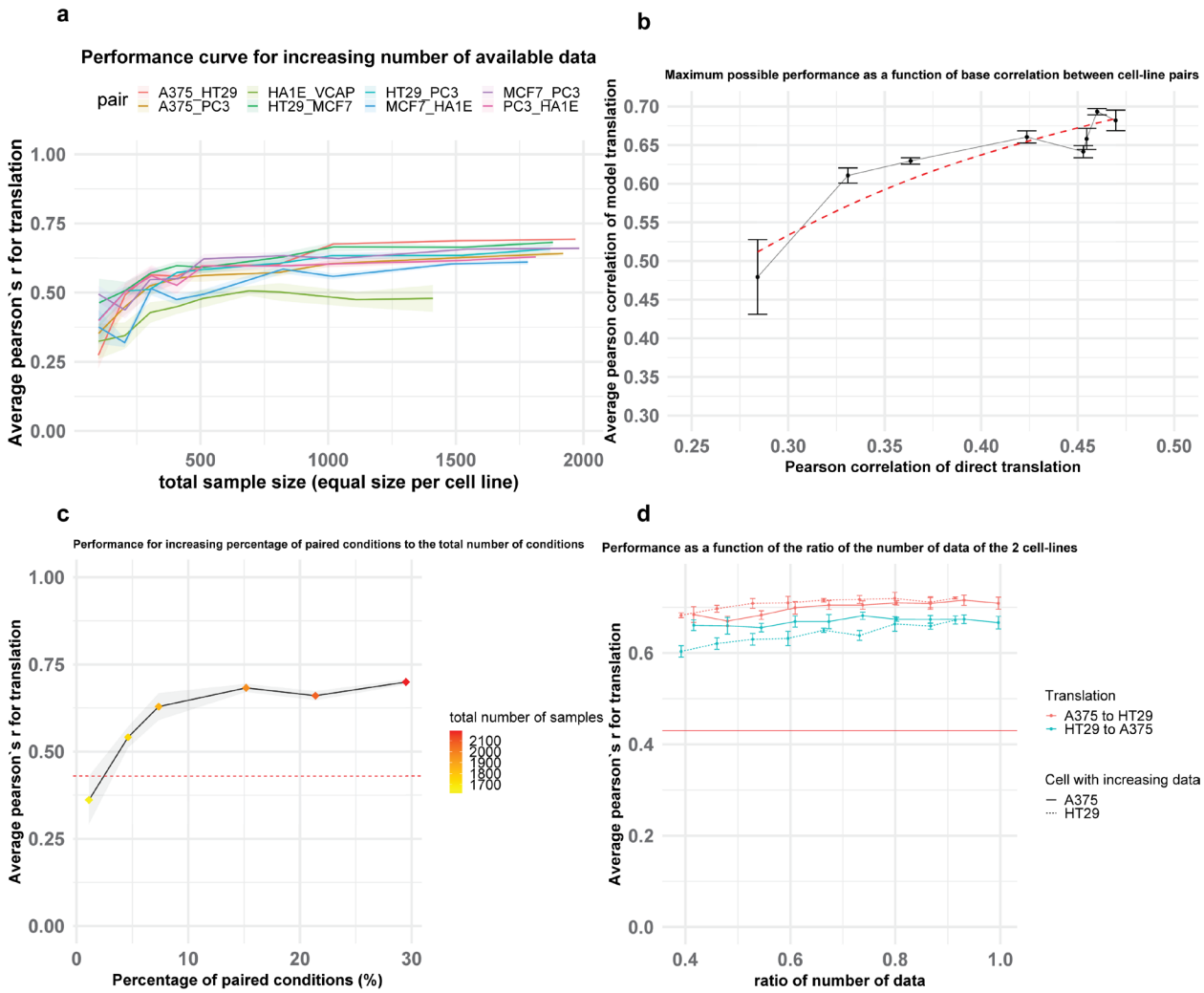






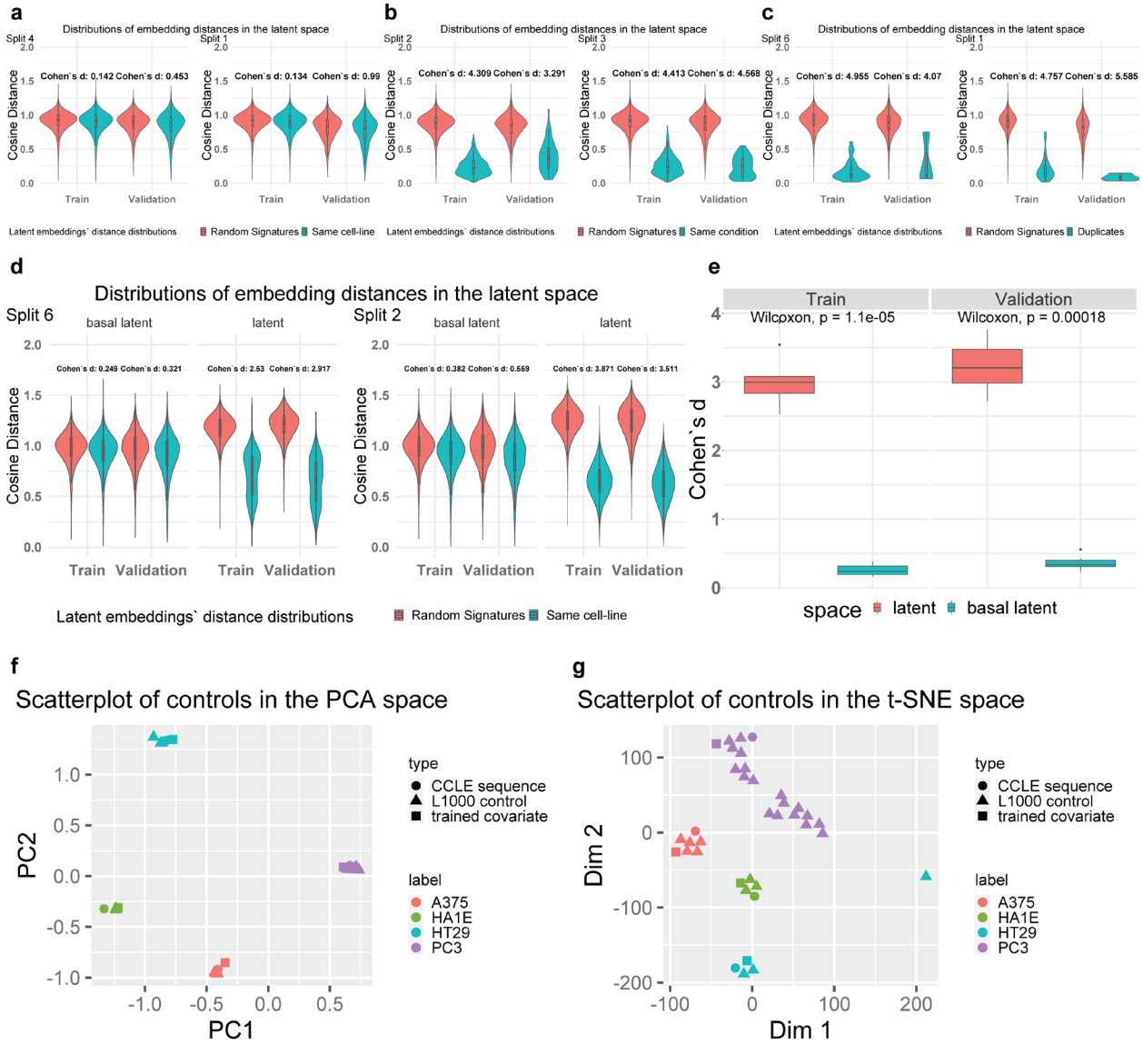
Approach	Accuracy (%)	F1 score (%)
Gene expression classifier	95.00±0.96	95.15±0.96
PCA space classifier	96.83±0.60	96.85±0.62
TransCompR transformed classifier	99.83±0.09	99.82±0.09
DCS global embeddings classifier	92.22±0.41	92.40±0.45
Global Latent space classifier	84.70±0.79	85.05±0.77
Autoencoders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

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



861

862 **<u>Figure 2:</u> a)** Performance in the translation task of the CPA-based approach across different cell-line
863 pairs and different sizes of training data. **b)** Model performance in translation as a function of the initial
864 similarity of 2 cell lines. **c)** Model performance in translation for different percentages of paired
865 conditions. **d)** Model performance in translation for low-to-medium cell-line imbalance in the
866 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.

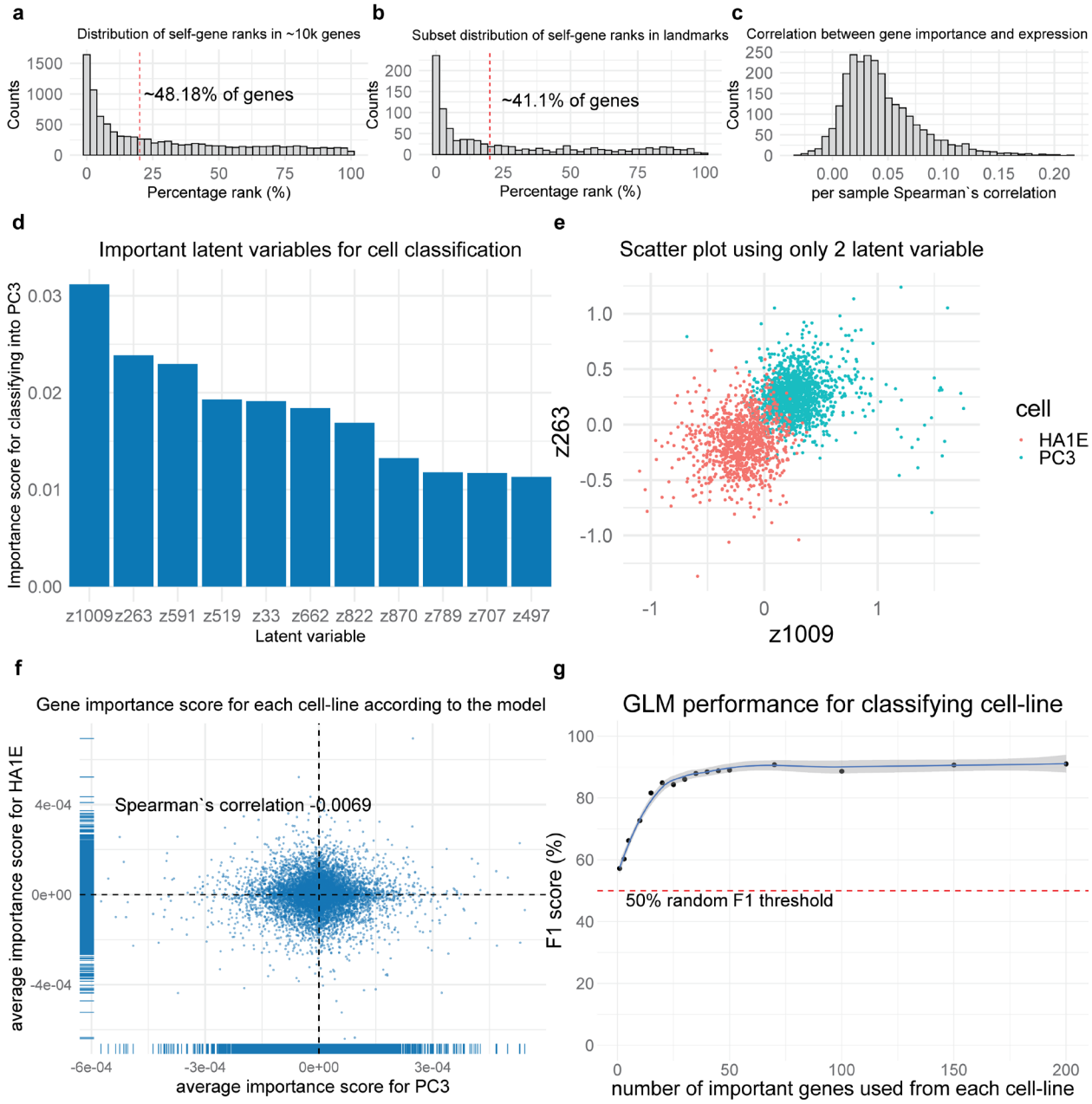


Figure 4: **a-b)** Distribution of percentage rank in terms of the importance of a gene to translate its expression to itself, using the 10,086 genes and the 978 landmark genes respectively in the L1000 dataset. **c)** Per sample Spearman's correlation between absolute gene importance scores and absolute 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 the number of important genes used is increased.

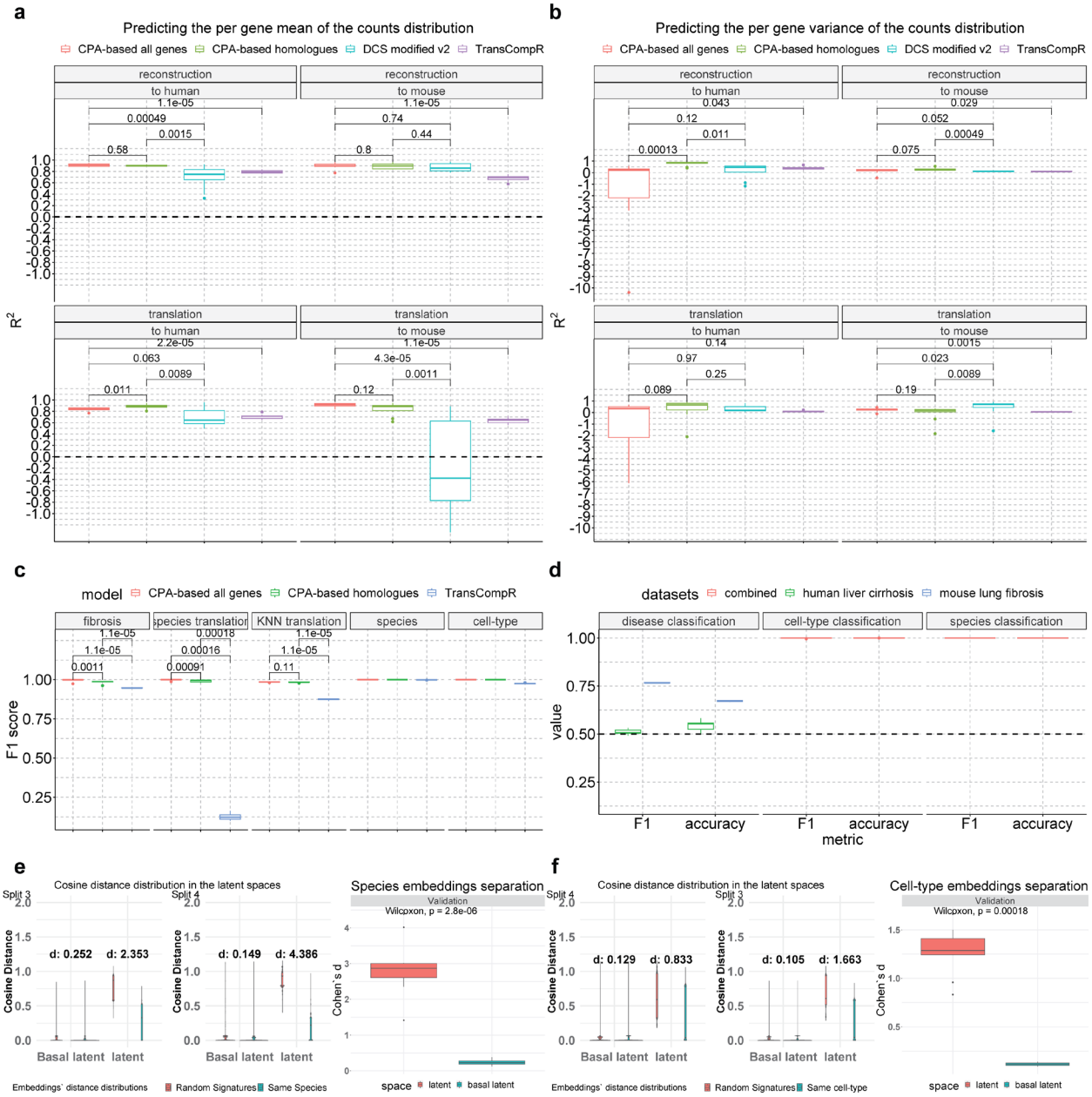


Figure 5: a-b) Performance (R^2) in predicting the per gene mean and variance of single-cell RNA-sequencing data for the tasks of reconstruction and species translation in the human-mouse lung fibrosis datasets. The comparison is done between the CPA-based framework using all genes or homologs and TransCompR **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.

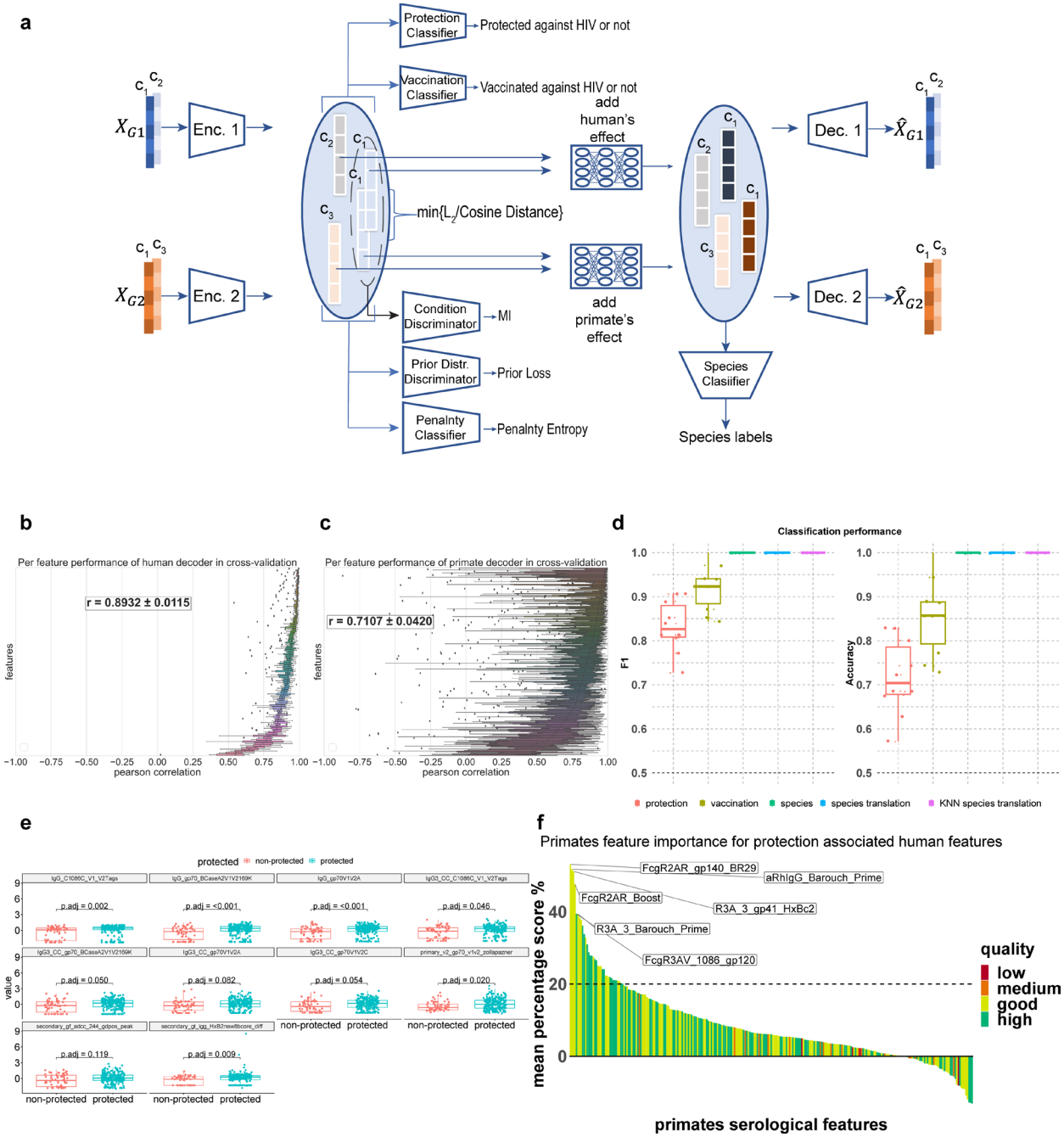


Figure 6: a) Framework architecture for inter-species translation in the serology dataset. Instead of adding the species effect with a trained vector, it is adding completely non-linear with 2 small intermediate neural networks. **b)** Per feature Pearson correlation across in 10-fold cross-validation 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