A Proofs

A.1 Non-negative and monomial matrices

In this section, we show that if the inverse of a non-negative matrix A exists and is itself non-negative, then A has to be a monomial matrix. This is a known linear algebra fact; we provide a proof for completeness, adapted from (EuYu, 2012).

Definition 3. *A matrix A is called a* non-negative *matrix if all of its elements are ≥ 0, and a* positive *matrix if all of its elements are > 0.*

Definition 4. *A matrix A is called a* monomial *matrix if it has exactly one non-zero entry in each row and each column. In other words, it has the same sparsity pattern as a permutation matrix, though the non-zero elements are allowed to differ from one.*

Lemma 3. *If A is an invertible non-negative matrix and A^{-1} is also non-negative, then A must be a non-negative monomial matrix.*

Proof. Since A is invertible, every row of A must have at least one non-zero element. Consider the i-th row of A, and pick j such that $A_{ij} \neq 0$. Since $AA^{-1} = I$, we have that the dot product of the i-th row of A with the k-th column of A^{-1} must be 0 for all $i \neq k$. As A and A^{-1} are both non-negative, this dot product can only be 0 if every term in it is 0, including the product of A_{ij} with A^{-1}_{jk}. However, $A_{ij} \neq 0$ by construction, so A^{-1}_{jk} must be 0 for all $i \neq k$. In other words, the j-row of A^{-1} must be all 0 except for A^{-1}_{ji}.

Applying a symmetric argument, we conclude that the i-th row of A must be all 0 except for A_{ij}. Since this holds for all i, we have that A must have exactly one non-zero in each row. Moreover, these non-zeros must appear in distinct columns, else A would be singular. We thus conclude that A must be a monomial matrix. □

A.2 The Jacobians of monotone and order isomorphic functions

We recall the definition of monotone and order isomorphic functions from the main text:

Definition 1. *A function f is* monotone *if $u \preceq v \implies f(u) \preceq f(v)$ for all $u, v \in dom(f)$, where ordering is taken with respect to the positive orthant (i.e., $u \preceq v$ means $u_i \leq v_i$ for all i).*

Definition 2. *An injective function f is an* order isomorphism *if f and f^{-1} restricted to the image of f are both monotone, that is, $u \preceq v \iff f(u) \preceq f(v)$.*

In this section, we establish that monotonicity and order isomorphism impose strong constraints on the function Jacobians.

Lemma 4. *If a function $f\colon \mathbb{R}^{k_r} \to \mathbb{R}^{k_r}$ is twice differentiable and monotone, then the Jacobian of f evaluated at any $z \in \mathbb{R}^{k_r}$ is a non-negative matrix.*

Proof. Assume for contradiction that f is differentiable and monotone, but that there exists some $z \in \mathbb{R}^{k_r}$ such that the Jacobian $J_f(z)$ is not a non-negative matrix. By definition, this implies that we can find i and j such that the ij-th entry of $J_f(z)$ is negative.

Let e_j represent the j-th unit vector. By the remainder bound for Taylor approximations, twice differentiability implies that for any compact ball around z, we can find some constant M such that we can write $f(z + \delta e_j) \leq f(z) + \delta J_f(z) e_j + \frac{M}{2}\delta^2$. If we pick $\delta < 2|J_f(z)_{ij}|/M$, the first order term dominates. Since the ij-th entry is negative, this means that $f_i(z + \delta e_j) < f_i(z)$ even though $z + \delta e_j \succeq z$, contradicting the monotonicity of f. □

Lemma 5. *If $q\colon \mathbb{R}^{k_r} \to \mathbb{R}^{k_r}$ is twice continuously differentiable and an order isomorphism, then the Jacobian matrix $J_h(z)$ is a non-negative monomial matrix for all $z \in \mathbb{R}^{k_r}$.*

Proof. If q is an order isomorphism, then q and q^{-1} are both monotone. By Lemma 4, their respective Jacobian matrices are non-negative everywhere.

Now, for any $z \in \mathbb{R}^{k_r}$, the inverse function theorem tells us that $[J_q(z)]^{-1} = J_{q^{-1}}(q(z))$, so both $J_q(z)$ and its inverse $[J_q(z)]^{-1}$ are non-negative. Applying Lemma 3 gives us that $J_q(z)$ is a non-negative monomial matrix. □

A.3 Component-wise monotonicity of order isomorphisms

The conditions on the Jacobian of a twice differentiable order isomorphic function q imply a constrained form.

Lemma 1 (restated)**.** *If $q\colon \mathbb{R}^{k_r} \to \mathbb{R}^{k_r}$ is an order isomorphism and twice continuously differentiable, q must be expressible as a permutation followed by a component-wise strictly monotone transform.*

Proof. Since q is bijective, q^{-1} exists everywhere, which implies that $J_q(r)$ must have full rank everywhere. Since $J_q(r)$ is a monomial matrix by Lemma 5, this means that the sparsity pattern of $J_q(r)$ cannot vary with r; otherwise, by the intermediate value theorem, there will be some r where $J_q(r)$ where a row has greater than one nonzero or no nonzeros and thus is not monomial. By definition, a monomial matrix can be decomposed into a positive diagonal matrix and a permutation. Applying the fundamental theorem of calculus to each diagonal entry recovers the strictly monotone tranform, and the permutation matrix defines the

permutation. The existence of the antiderivative is guaranteed by construction of J_q as the derivative of q. □

A.4 Identifiability in the noiseless setting

We start by establishing two helpful lemmas:

Lemma 6. *If functions f_1 and f_2 are both monotone, then $f_1 \circ f_2$ is also monotone.*

If f_1 and f_2 are both bijective order isomorphisms, then $q \stackrel{\text{def}}{=} f_2^{-1} \circ f_1$ is also a bijective order isomorphism.

Proof. The first part of the lemma follows from the transitivity of partial orders: $x \prec y \implies f_1(x) \prec f_1(y) \implies f_2(f_1(x)) \prec f_2(f_1(x))$.

For the second part, note that q is bijective because it is the composition of two bijective functions. Now, since f_1 and f_2 are both order isomorphisms, we know that f_1, f_1^{-1}, f_2, and f_2^{-1} are all monotone. By the first part of the lemma, we conclude that $q = f_2^{-1} \circ f_1$ and $q^{-1} = f_1^{-1} \circ f_2$ are both monotone, implying that q is an order isomorphism. □

Lemma 7. *If a continuous, univariate, strictly monotone function q_i is measure preserving for a random variable x, q_i must be the identity map (on the support of x).*

Proof. By strict monotonicity, $c_1 < c_2$ implies $q(c_1) < q(c_2)$ and thus the CDF is preserved implying that $P(x < c) = P(q(x) < q(c)) = P(x < q(c))$. The last step follows from measure preservation of q.

Now assume for contradiction that q_i is not the identity map. We can then pick some c such that $q(c) \neq c$ and $P(c) > 0$. This implies that $P(x < q(c)) \neq P(x < c)$ which is a contradiction. □

We can now state and prove identifiability in the noiseless setting:

Proposition 1 (restated). *If f_1 and f_2 and their inverses are twice continuously differentiable and order-isomorphic functions such that $f_1(tr) \stackrel{d}{=} f_2(tr) \stackrel{d}{=} x_t$ for some $t > 0$, then f_1 and f_2 are identical up to a permutation.*

Proof. We consider the difference map $q \stackrel{\text{def}}{=} f_2^{-1} \circ f_1$, which maps latent rates of aging implied by f_1 to that of f_2. Our aim is to show that q must be a permutation, which will give the desired result.

From Lemma 6, we know that q is itself an order isomorphism. Thus, by Lemma 1, it must be expressible as the composition of a component-wise strictly monotone map and a permutation.

We can further show that this component-wise strictly monotone transformation has to be the identity transformation. Since both f_1 and f_2 map $rt \mapsto x_t$, q is measure preserving on rt. In other words, it maps the probability distribution of rt to itself. We can therefore apply Lemma 7 to conclude that q can only be a permutation.

Applying f_2 to both sides of $q = f_2^{-1} \circ f_1$, we get that f_1 and f_2 have to be permutations of each other, as desired. □

A.5 Checking order isomorphisms

Lemma 2 (restated). *Let $a(x) = Ax$, where $A \in \mathbb{R}^{d \times k}$. If we can write $A = P \begin{bmatrix} B \\ C \end{bmatrix}$ where P is a permutation matrix, B is a non-negative monomial matrix, and C is a non-negative matrix, then a is an order isomorphism.*

Proof. a is monotone since A is non-negative. To verify that the inverse of a over its image is monotone, let $I_k = [I; 0] \in \mathbb{R}^{k \times d}$ be the matrix selecting the first k coordinates. If $Ax \prec Ay$, every coordinate of Ax is smaller than the corresponding coordinate of Ay, so we can jointly permute the rows (i.e., left-multiplying by a permutation matrix) or select a subset of coordinates while preserving ordering. Thus, $Ax \prec Ay \implies I_k P^{-1} Ax \prec I_k P^{-1} Ay$. By construction, $I_k P^{-1} A = B$ is a non-negative monomial matrix. Applying a similar permutation argument, we have that $I_k P^{-1} Ax \prec I_k P^{-1} Ay \implies x \prec y$. □

Proposition 2 (restated). *Let $f \colon \mathbb{R}^k \to \mathbb{R}^d = s_2 \circ a \circ s_1$, where $s_1 \colon \mathbb{R}^k \to \mathbb{R}^k$ and $s_2 \colon \mathbb{R}^d \to \mathbb{R}^d$ are continuous, component-wise monotone transformations, and $a \colon \mathbb{R}^k \to \mathbb{R}^d$ is a linear transform. If a satisfies Lemma 2, then f is an order isomorphism.*

Proof. The proof follows from the fact that order preservation is transitive. $a \circ s_1$ is an order isomorphism onto its image, since s_1 is an order isomorphism on the entire $\mathbb{R}^k$ and a is order isomorphic onto its image by Lemma 2. Thus for any $x \prec y \iff a(s_1(x)) \prec a(s_1(y))$. Since s_2 is an order isomorphism on $\mathbb{R}^d$, we have $x \prec y \iff a(s_1(x)) \prec a(s_1(y)) \iff s_2(a(s_1(x))) \prec s_2(a(s_1(y)))$. □

B UK Biobank dataset and processing

Phenotype filtering. We selected Biobank phenotypes that were measured for a large proportion of the dataset and that captured diverse and important dimensions of aging and general health. After removing phenotypes which were missing data for

many people, redundant (e.g., there are multiple measurements of BMI), or discrete (e.g., categorical responses from a survey question), we were left with 52 phenotypes (Table 2) across the following categories: spirometry (a measure of lung function), bone density, body type/anthropometry, cognitive function, vital signs (blood pressure and heart rate), physical activity, hand grip strength, and blood test results. By visual inspection, we categorized the 52 phenotypes into monotone features (45/52) and non-monotone features (7/52) for the cross-sectional model. In the combined longitudinal/cross-sectional model, we modeled an additional 8 features as non-monotone because they increased in the longitudinal data but not in the cross-sectional data, or vice versa.

Sample filtering. We removed individuals with non-European ancestry, as identified from their genetic principal components, as is commonly done in studies of the UK Biobank to minimize spurious correlations with ancestry particularly in genetic analysis (Lane et al., 2016; Wain et al., 2015). (The vast majority of individuals in UK Biobank are of European ancestry.) We also removed individuals who were missing data in any of our selected phenotypes.

After filtering, we were left with a train/development set of 213,510 individuals; we report all results on a test set of 53,174 individuals not used in model development or selection. While these samples are cross-sectional (with a measurement at only a single timepoint), we have a single longitudinal followup visit for an additional 8,470 individuals, on which we assess longitudinal progression. UK Biobank data contains two followup visits; we use only longitudinal data from the first followup visit (2-6 years after the initial visit), not the second, because some of the phenotypes we use in model fitting were not measured at the second followup.

Phenotype processing. We normalized each phenotype to have mean 0 and variance 1. In fitting the model, we first transformed all phenotypes so they were positively correlated with age, by multiplying all phenotypes which were not by negative one, so we could assume that monotone features were monotone increasing. However, all results in the paper are shown with the original phenotype signs.

Diseases, mortality, and risk factors. We examined associations with 91 diseases which were reported by at least 5,000 individuals in the entire UKBB dataset. Diseases were retrospectively assessed via interview (i.e., subjects developed the disease prior to the measurement of x_{t_0}). Second, we examined associations between rates of aging and mortality. In contrast to disease status, mortality was measured after x_{t_0} (all subjects were obviously alive when x_{t_0} was measured); thus, examining associations with mortality serves as an indication that rates of aging predict future outcomes. Finally, we examined 5 binary risk factors: whether the individual currently smokes, if they are a heavy drinker, if they are above the 90th percentile in Townsend deprivation index (a measure of low socioeconomic status), if they have type 2 diabetes, and if they report no days of moderate or vigorous exercise in a typical week.

We examined associations between rates of aging and mortality using a Cox proportional hazards model which controlled for age, sex, and the first five genetic principal components. We report the hazard ratios for a one standard-deviation increase in rate of aging. For the 5 binary risk factors and the 91 diseases, we examined associations using a linear regression model, where the dependent variable was the rate of aging and the independent variable was the risk factor or disease. We controlled for age, sex, and the first five genetic principal components. We filtered for associations which passed a statistical significance threshold of $p = 0.05$, with Bonferroni correction for the number of tests performed. Figure 3 reports the diseases/risk factors with the largest positive associations and an effect size of a greater than 1% increase in the rate of aging; if more than five diseases or risk factors pass this threshold, we report the top five.

C Model architecture and hyperparameters

Model architecture. Figure G.4 illustrates our model architecture. The monotone function $f = s \circ a$ is parametrized as the composition of a monotone elementwise transformation $s\colon \mathbb{R}^{d'} \to \mathbb{R}^{d'}$ with a monotone linear transform $a\colon \mathbb{R}^{k_r} \to \mathbb{R}^{d'}$. We parametrize the linear transformation a using a matrix A constrained to have non-negative entries, and implement each component $s_i(v)\colon \mathbb{R}_+ \to \mathbb{R}_+$ of s as the sum of positive powers of $v \in \mathbb{R}_+$ with non-negative coefficients $s_i(v) = \sum_{p_j \in S} w_j v^{p_{ij}}$, where w_{ij} are learned non-negative weights, and S is a hyperparameter. (For example, $S = [\frac{1}{2}, 1, 2]$ yields the function class $s(v) = w_0 v^{\frac{1}{2}} + w_1 v + w_2 v^2$. We illustrate some of the learned S in Appendix Figure G.5). We verified that the learned model's A matrix (part of the monotone function f) can be row-permuted into a combination of an approximately monomial matrix and positive matrix, indicating that we learned an f that was order-isomorphic.

We use neural networks to parametrize the non-monotone functions $\tilde{f}$ and g as well as the encoder (which approximates the posterior over the latent

Table 2: UK Biobank features used in model fitting. * denotes features which are modeled as non-monotone in age when fitting the cross-sectional model. ** denotes additional features which are modeled as non-monotone in age when fitting the model which uses both longitudinal and cross-sectional data. All features which are modeled as non-monotone in the cross-sectional analysis are also modeled as non-monotone in the combined longitudinal/cross-sectional model.

Feature
Spirometry: forced vital capacity
Spirometry: peak expiratory flow
Spirometry: forced expiratory volume in 1 second (FEV1)
Bone density: heel bone mineral density
Bone density: heel broadband ultrasound attenuation
Bone density: heel quantitative ultrasound index
Body type: body fat percentage
Body type: body mass index
Body type: hip circumference
Body type: impedance of whole body**
Body type: sitting height
Body type: standing height
Body type: waist circumference
Body type: whole body fat free mass
Body type: whole body fat mass
Body type: whole body water mass
Cognitive function: duration to first press of snap button
Cognitive function: mean time to correctly identify matches
Vital signs: diastolic blood pressure*
Vital signs: pulse rate**
Vital signs: systolic blood pressure
Physical activity: days/week of moderate activity (10+ min)
Physical activity: days/week of vigorous activity (10+ min)
Physical activity: days/week walked (10+ min)
Physical activity: time spent driving
Physical activity: time spent using computer**
Physical activity: time spent watching television
Hand grip strength: hand grip strength left
Hand grip strength: hand grip strength right
Blood: basophil percentage
Blood: eosinophil percentage
Blood: haematocrit percentage*
Blood: haemoglobin concentration*
Blood: high light scatter reticulocyte percentage
Blood: immature reticulocyte fraction
Blood: lymphocyte percentage*
Blood: mean corpuscular haemoglobin
Blood: mean corpuscular haemoglobin concentration**
Blood: mean corpuscular volume**
Blood: mean platelet thrombocyte volume
Blood: mean reticulocyte volume
Blood: mean sphered cell volume
Blood: monocyte percentage**
Blood: neutrophil percentage*
Blood: platelet count
Blood: platelet crit
Blood: platelet distribution width
Blood: red blood cell erythrocyte count*
Blood: red blood cell erythrocyte distribution width
Blood: reticulocyte count**
Blood: reticulocyte percentage**
Blood: white blood cell leukocyte count*

variables r and b). We adopt the following priors: $r \sim \text{lognormal}(0, \sigma_r^2 I)$; $b \sim \mathcal{N}(0, I)$; and $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 I)$. We use a lognormal distribution for r to ensure positivity; set $\sigma_r = 0.1$ to reflect a realistic distribution of the rates of biological aging (Belsky et al., 2015); and optimize over σ_ϵ. Finally, we simply take t to be an individual's age, although we could also have optimized over some constant t_0 and taken $t = \text{age} - t_0$.

Hyperparameter selection. We conducted a random search over the encoder architecture, decoder architecture, learning rate, elementwise nonlinearity, and whether there was an elementwise nonlinearity prior to the linear transformation matrix. We selected a configuration which performed well (as measured by low reconstruction error/high out-of-sample evidence lower bound (ELBO)) across a range of latent state sizes. Our final architecture uses a learning rate of 0.0005, encoder layer sizes of [50, 20] prior to the latent state, and decoder layer sizes of [20, 50]. Our elementwise nonlinearity is parametrized by $s(y) = \sum_{p_i \in S} w_i y^{p_i}$, where $S = [\frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 1, 2, 3, 4, 5]$. We found that using an elementwise nonlinearity prior to the linear transformation was not necessary in our dataset, so we only used a nonlinearity after the linearity transformation for interpretability and ease in training. We used Adam for optimization (Kingma and Ba, 2014) and ReLUs as the nonlinearity.

D Baselines

Linear baselines. We compare to three linear baselines (PCA, contrastive PCA, and mixed criterion PCA), using the same number of dimensions as in the original model ($k_\text{r} + k_\text{b} = 15$). We compare to PCA because it is commonly used in biological studies (Relethford et al., 1978) and serves as a good baseline for reconstruction performance. (We evaluate PCA reconstruction loss both when PCA is provided age as an input, and when it is not; its reconstruction loss is virtually identical regardless). However, because PCA does not naturally isolate age-related variation, a key goal of our analysis, we also compare to two linear baselines which naturally incorporate age information: contrastive PCA (Abid et al., 2018) and mixed-criterion PCA (Bair et al., 2006).

Contrastive PCA takes as input a *foreground dataset* and a *background dataset*, and finds a set of latent components which maximize variance in the foreground space while minimizing variance in the background space (trading off between the two objectives using a weighting α). The latent components v optimize

$$\max_{||v||=1} v^T C_\text{foreground} v - \alpha v^T C_\text{background} v \tag{3}$$

where $C_\text{foreground}$ and $C_\text{background}$ are the empirical covariance matrices of the foreground and background datasets, respectively. This corresponds to taking the eigenvectors of the matrix $C_\text{foreground} - \alpha C_\text{background}$. Because we seek to isolate age-related variation, we use as our foreground dataset the entire dataset of Biobank participants (aged 40-69), and as the background set participants aged 40-49. Contrastive PCA will thus identify components which explain variation in the population as a whole but not within participants of similar ages (40-49). Following the original authors, we experiment with a set of weightings α logarithmically spaced between 0.1 and 1,000. We report results with $\alpha = 10$ because this weighting reconstructs the data almost as well as PCA but does not learn identical latent dimensions, indicating that the weighting is having an effect; however, the patterns we report in the main text hold with other α as well.

Mixed-criterion PCA, like contrastive PCA, uses a two-term objective: the PCA objective (weighted by $1-\alpha$), and a second term (weighted by α) which encourages the learned components to correlate with age:

$$\max_{||v||=1} (1-\alpha)\text{Var}(Xv) + \alpha\text{Cov}(Xv, t) \tag{4}$$

where X is the matrix of observed features and t is age. When $\alpha = 0$, mixed-criterion PCA reduces to standard PCA; when $\alpha = 1$, it learns a single component which is the linear combination of observed features which correlates most strongly with age. We experiment with a range of α and report results with $\alpha = 0.99$, because this yields several top principal components which correlate with age; using a significantly smaller α produces results very similar to PCA, and using a significantly larger α produces only a single meaningful component which is strongly correlated with age, severely harming reconstruction performance.

Non-linear baseline: non-monotone model. We use the same hyperparameter settings as for the monotone model but remove the constraint that the age decoder must be linear. Thus, all observed features are represented as an arbitrary function of the age latent state rt plus an arbitrary function of the bias latent state b.

E Learning from both cross-sectional and longitudinal data

Our model can naturally incorporate any available longitudinal data by optimizing the joint likelihood of the cross-sectional and longitudinal data. As cross-sectional and longitudinal data can display different biases (Fry et al., 2017; Louis et al., 1986; Kraemer et al., 2000), this can produce models that are less affected by the biases in a particular dataset.

We handle longitudinal data similarly to cross-sectional data, but with an additional term in the model objective that captures the expected log-likelihood of observing the longitudinal follow-up x_{t_1} given our posterior of r and b. We control the relative weighting between cross-sectional and longitudinal data with a single parameter λ_{lon}; when $\lambda_{\text{lon}} = 1$, the longitudinal and cross-sectional losses per sample are equally weighted; when $\lambda_{\text{lon}} = 0$, the model tries to fit only the cross-sectional data, and when $\lambda_{\text{lon}} \gg 1$, the model tries to fit only the longitudinal data. We fit the longitudinal model using the same model architecture and hyperparameters as the cross-sectional experiments (Appendix C), varying only the longitudinal loss weighting λ_{lon}. The loss for cross-sectional samples is the negative evidence lower bound (ELBO), as before. The loss for longitudinal samples has an additional term that captures the expected log-likelihood of observing the longitudinal follow-up x_{t_1} given our posterior of r and b. We use the same model architecture as for the cross-sectional model. In particular, to avoid overfitting on the small number of longitudinal samples, we share the same encoder; this means that the approximate posterior over r and b for a longitudinal sample is calculated only using x_{t_0}. Because we have far more cross-sectional samples than longitudinal samples, we train the model by sampling longitudinal batches with replacement, with one longitudinal batch for every cross-sectional batch. In addition to the 7 non-monotonic features used in the cross-sectional experiments, we add an additional 8 features to the non-monotonic list because they increase in longitudinal data and not in cross-sectional data, or vice versa (Table 2).

We search over a range of values of λ_{lon} and find that test longitudinal loss (i.e., the negative evidence lower bound on the likelihood of x_{t_0} and x_{t_1}) is minimized when $\lambda_{\text{lon}} = 1$. This indicates that the model achieves the best longitudinal generalization performance by using cross-sectional data and the small amount of available longitudinal data. With higher λ_{lon}, the model overfits to the small longitudinal dataset. Repeating our longitudinal extrapolation task (Section 7.1) on a held-out test set of 1687 participants with longitudinal data and comparing to the same three benchmarks, we found that the model with $\lambda_{\text{lon}} = 1$ outperforms just predicting x_{t_0} on 83% of people with followups > 5 years (compared to 66% with purely cross-sectional data, as in Section 7.1); pure reconstruction on 79% (vs 61%); and the average-cross-sectional-change baseline on 80% (vs. 60%). The longitudinal model also outperforms benchmarks on the full longitudinal dataset (as opposed to just individuals with followups > 5 years) by similarly large margins. These results illustrate the benefits of methods which exploit both cross-sectional and longitudinal data.

F Model stability

We evaluated the stability of the learned rates of aging in response to various model and data perturbations. To compare the rates of aging learned by two different models, we defined ρ_r as the correlation between the $r^{(i)}$'s learned by the 2 models, averaged over each component of r, and maximized over permutations of components We found that learned rates of aging were stable over random seeds and changes to:

1. The number of non-monotone features. ρ_r with the original model remained high even as we tripled the number of non-monotone features from the original 7, to 25 (for which $\rho_r = 0.84$). (We did this by removing monotone constraints on randomly chosen features.)

2. Random subsets of training data. Models trained on different subsets, each containing 70% of the overall data, learned similar rates r (average ρ_r of 0.82 between models).

3. The dimensions k_r and k_b of the time-dependent and bias latent variables. When we altered k_r, the model learned many of the same rates of aging: e.g., for $k_r = 4$, ρ_r with the original model ($k_r = 5$) was 0.89, and for $k_r = 6$ it was 0.92. Results were also stable when we altered k_b and compared to the original $k_b = 10$: $\rho_r > 0.8$ for $8 \leq k_b \leq 12$.

G Supplementary Figures

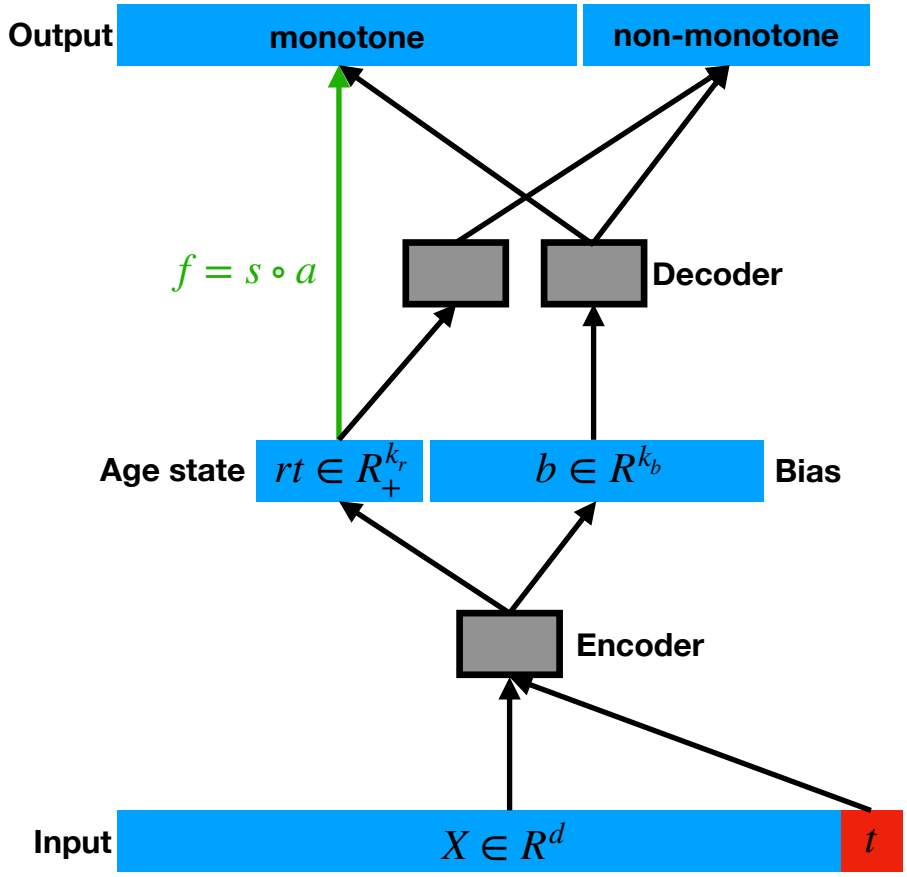

Figure G.4: The model structure. The features X and age t are fed into the encoder to approximate the posterior over the rates of aging r and bias b. The grey boxes indicate functions parametrized by neural networks. While both the monotone and non-monotone outputs are a function of both the age state rt and the bias b, only the relationship between rt and the monotone outputs (green arrow) is constrained to be monotone and parametrized by $f = s \circ a$.

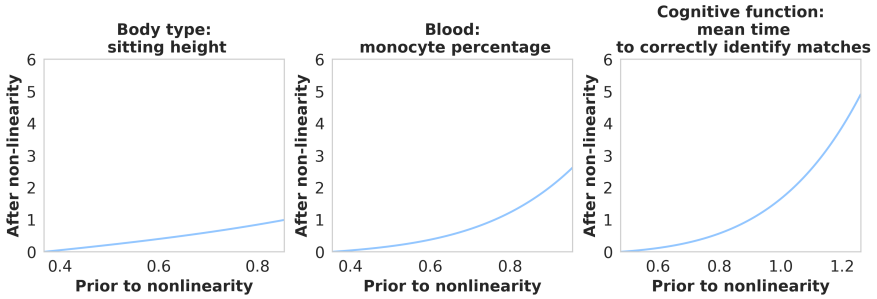

Figure G.5: Representative elementwise transformations s. Most elementwise transformations are close to linear, like the left plot, but some are not (right two plots). To determine the relevant domain for each elementwise transformation, we sample latent rt from the fitted cross-sectional model (for $t = 40$-69), feed it through the linear transformation a, and compute the 0.1th and 99.9th percentiles of the resulting distribution for each monotonic feature. This yields the relevant domain over which each elementwise transformation operates.

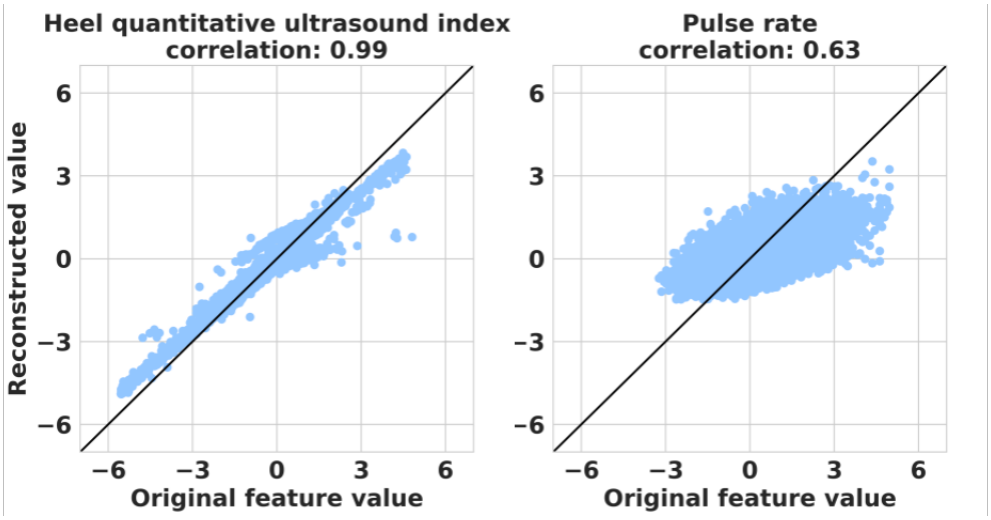

Figure G.6: Reconstructed vs. actual features. The figure plots the reconstructed $f(rt) + g(b)$ against the actual x_t for the 2 features with the highest ($\rho = 0.99$, left) and lowest correlation ($\rho = 0.63$, right). Overall, the model fits the data well: reconstructed features are highly correlated with actual features (mean $\rho = 0.88$), with most resembling the left plot.

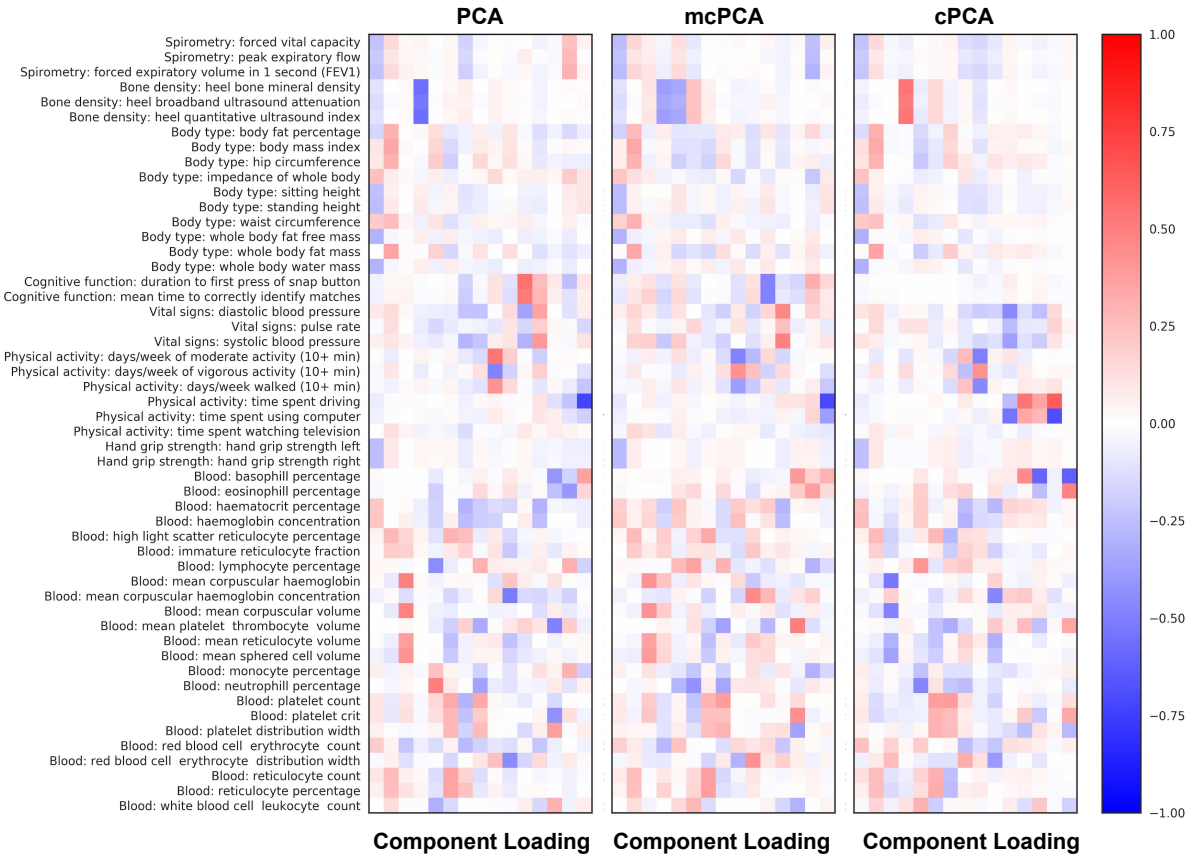

Figure G.7: Loadings for the three linear baselines (with 15 latent dimensions) reveal non-sparse latent dimensions which are difficult to interpret and do not clearly differentiate between age and non-age variation. Each cell shows the loading for one component (horizontal axis) and observed feature (vertical axis).