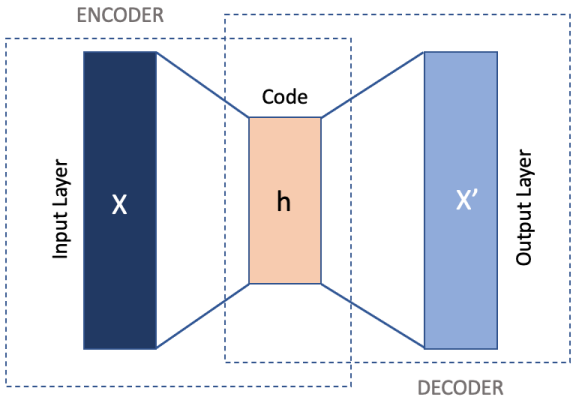
Unsupervised Deep Learning of Electrocardiograms Enables Scalable Human Disease Profiling

Supplementary Materials

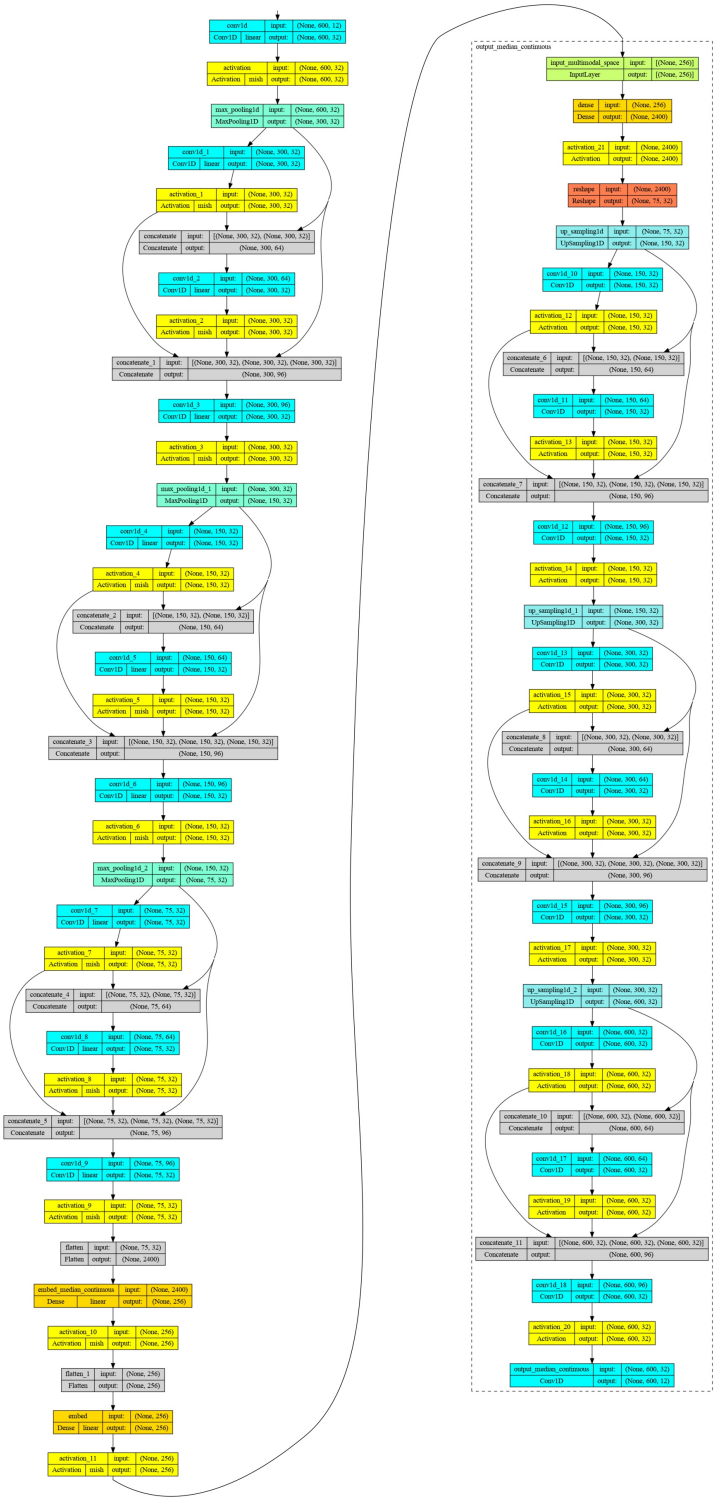
Table of Contents

Supplementary Figure 1. General autoencoder schematic.



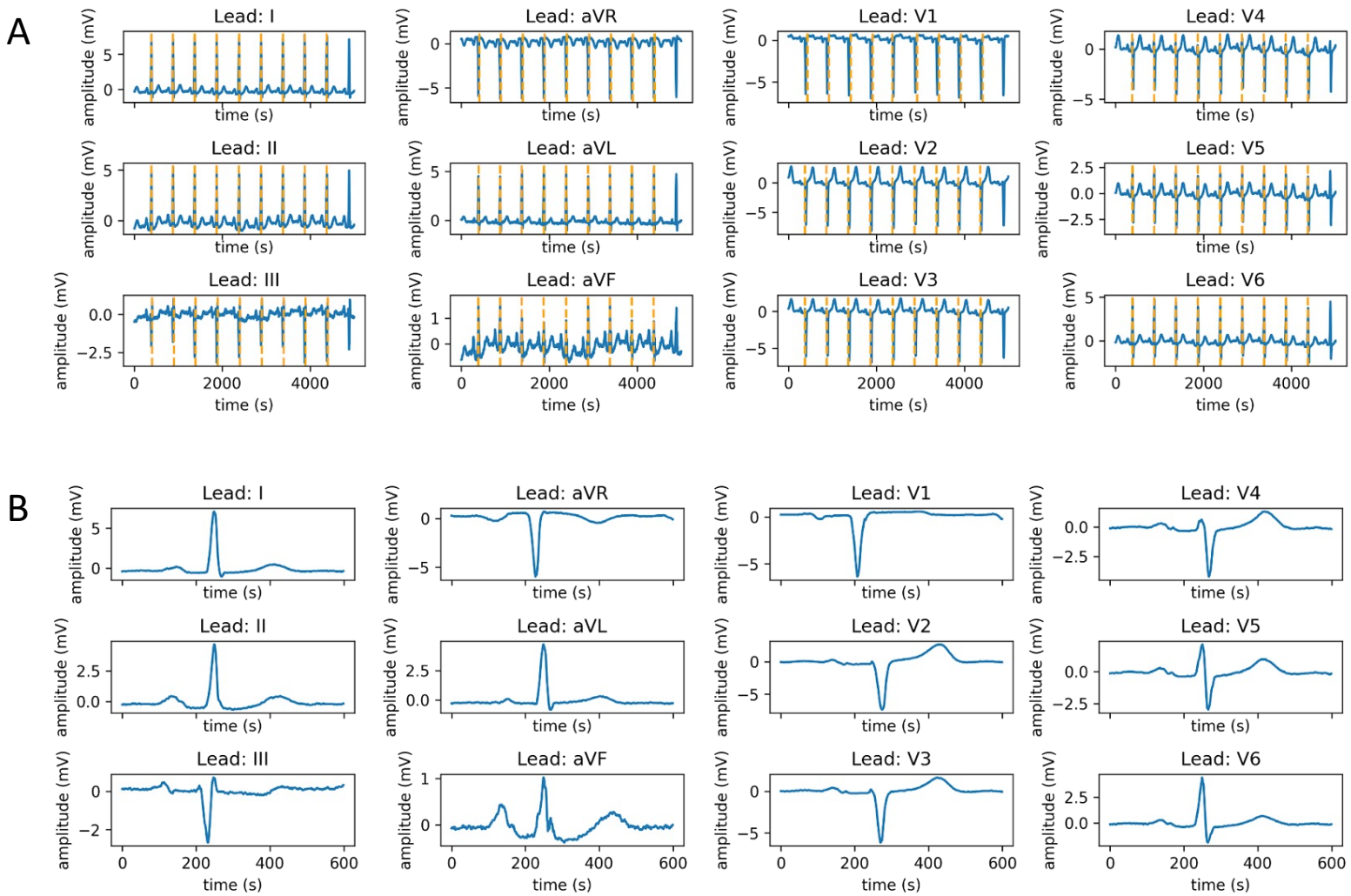
General autoencoder structure. The input, X, is encoded into a multi-dimensional latent space, h. This "code" is then used to reconstruct the input as accurately as possible at the output layer, X'.

Supplementary Figure 2. Autoencoder architecture.



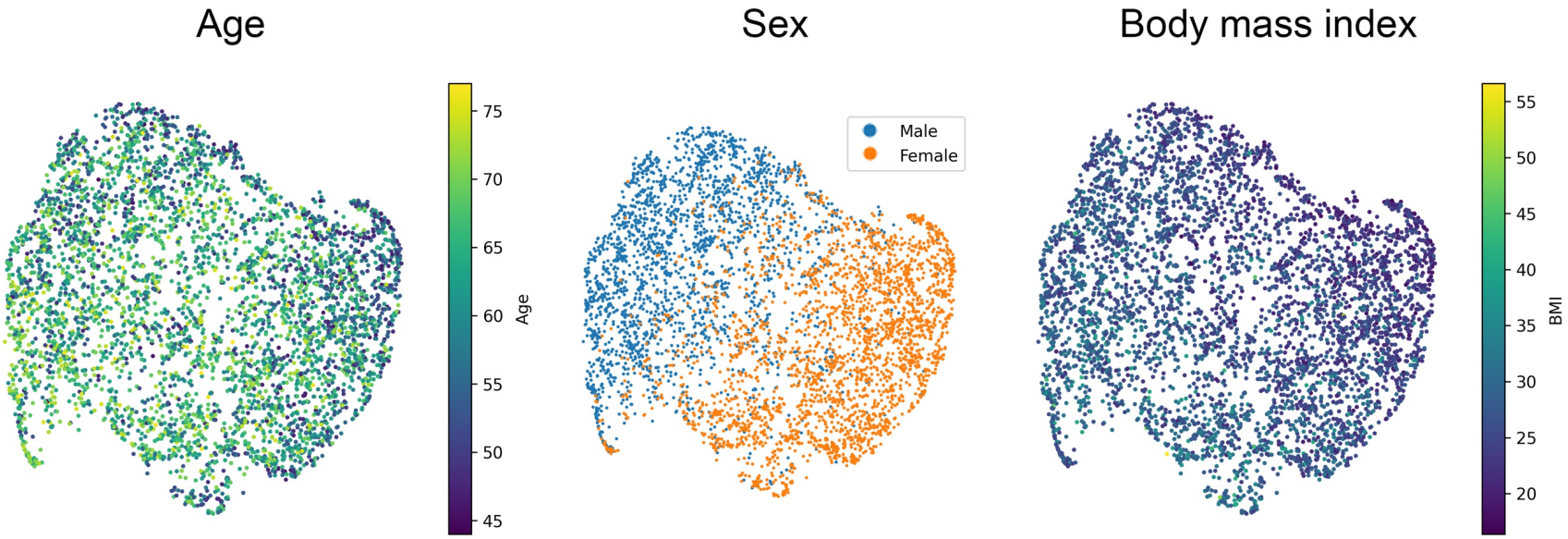
Depicted is a summary of the architecture of the convolutional autoencoder using in the current study. The encoder (left) takes 10 seconds of continuous 12-lead ECG waveform data as input to generate a lower dimensional latent space. The decoder (right) takes the latent space as input to reconstruct the 12-lead ECG. Box colors correspond to the following layer types: green = input, cyan = convolution, yellow = activation, teal = pooling, red = reshape, and orange = dense.

Supplementary Figure 3. Example of median waveforms generated from an original 12-lead electrocardiogram.



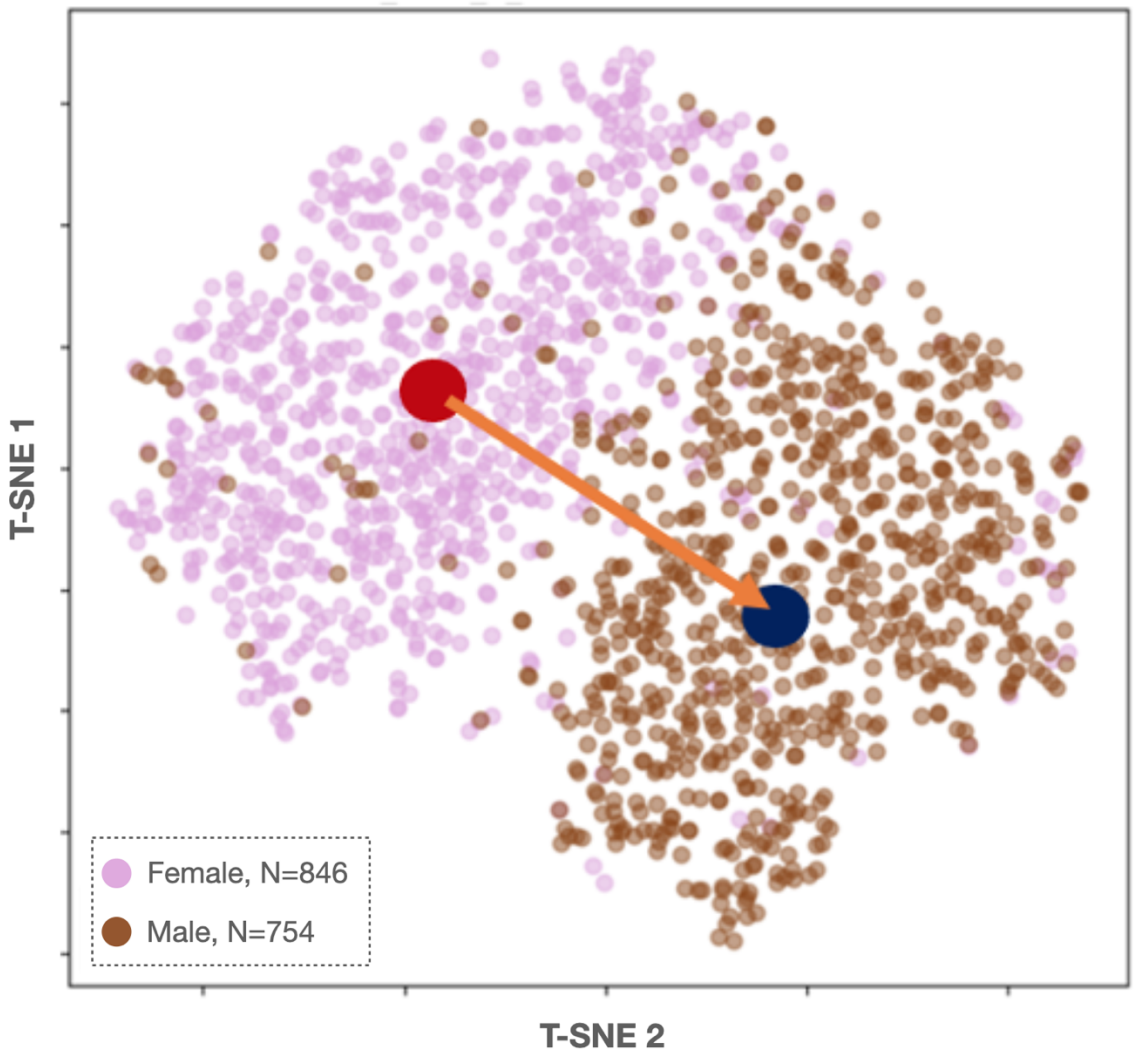
Panel A represents an exemplar original 12-lead tracing, and panel B represents the corresponding median waveform for each lead.

Supplementary Figure 4. Latent space representations of age, sex, and body mass index



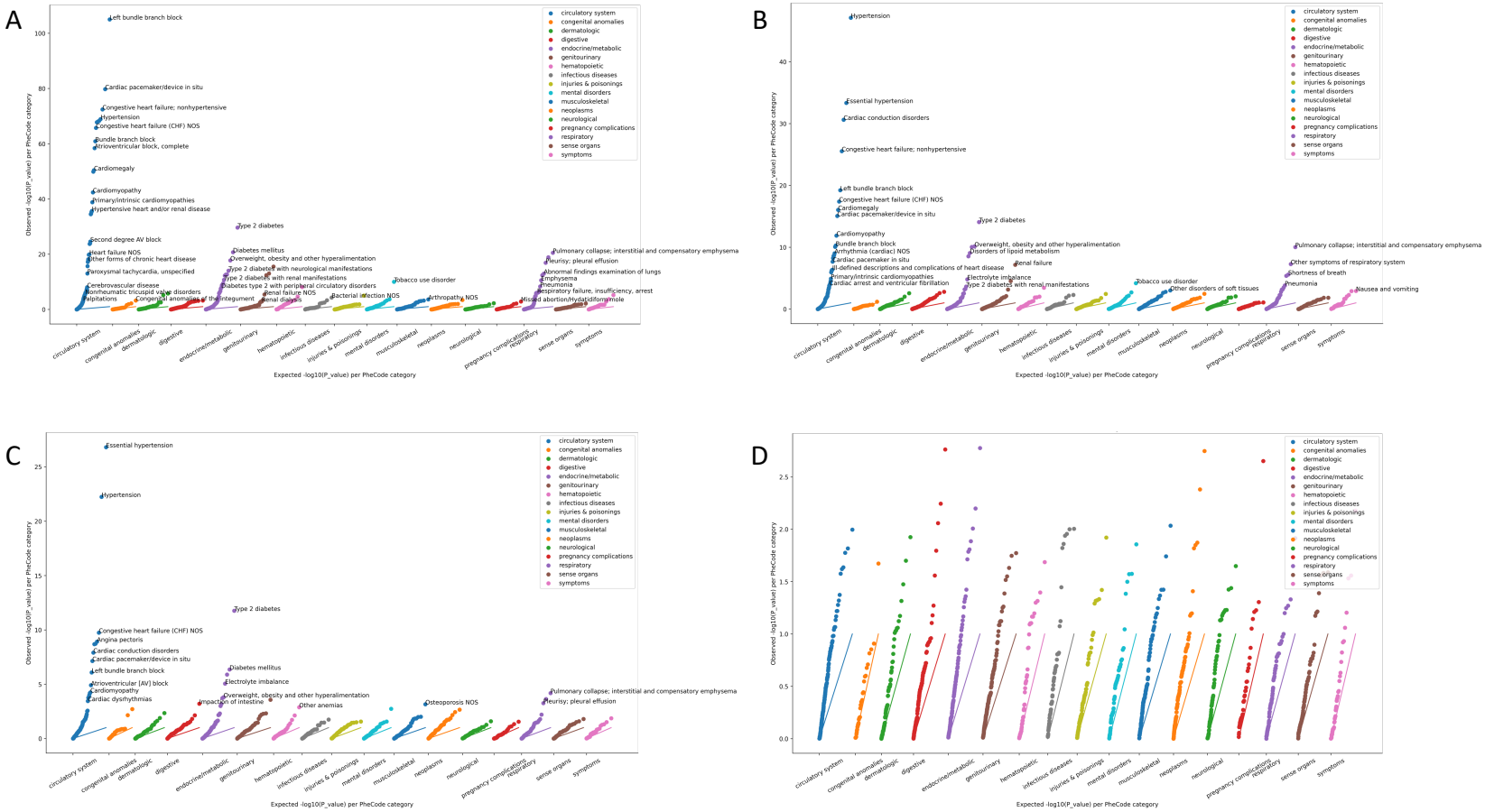
Figures show uniform manifold approximation and projections (UMAPs) depicting depictions of the ECG latent space stratified by age (left), sex (middle), and body mass index (right). Maps were derived using ECG encodings in latent space of XX.

Supplementary Figure 5. Cluster centroids and phenotype vector derivation.



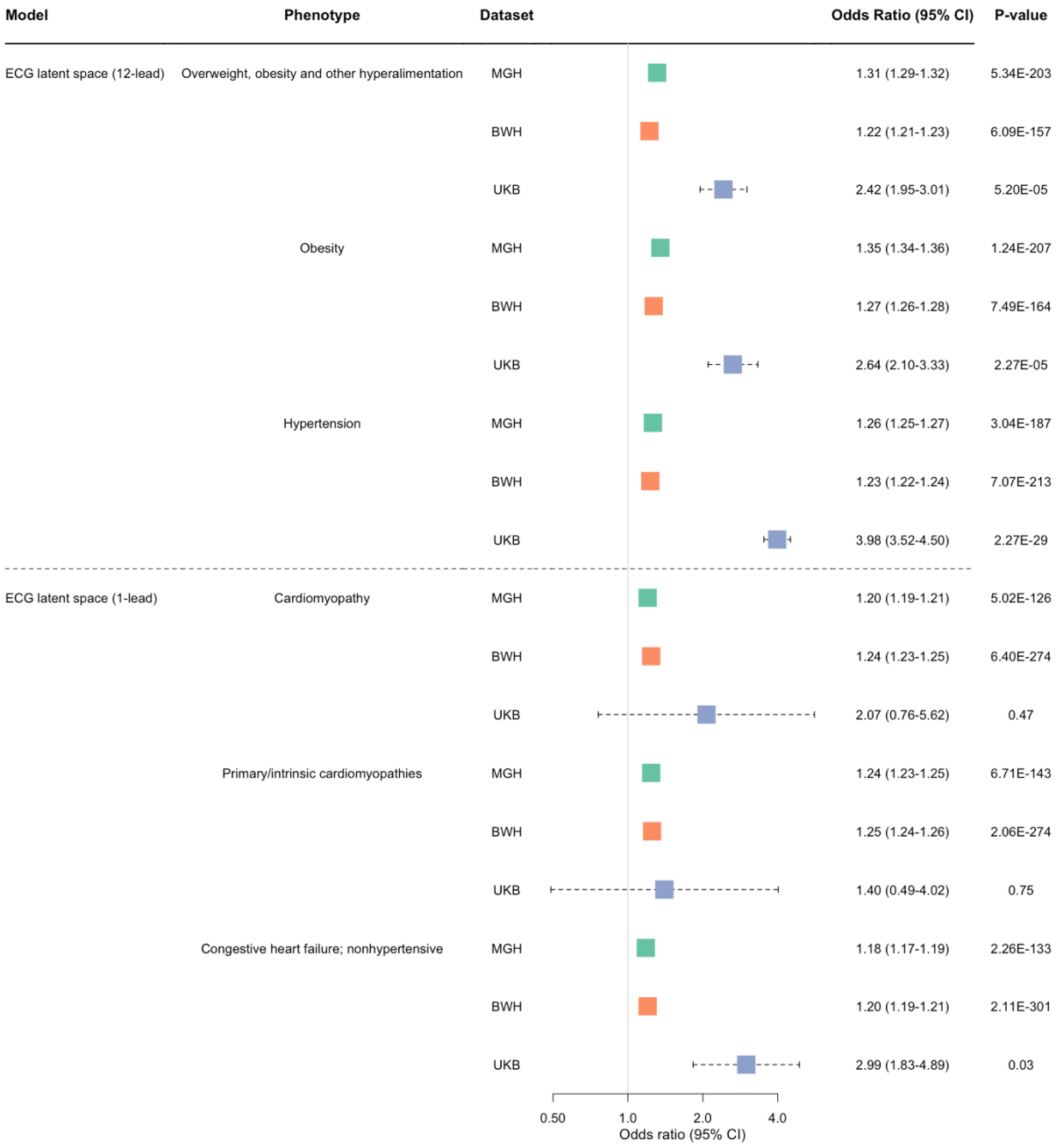
Illustrative example of phenotype vector derivation. The figure shows a two-dimensional t-Distributed Stochastic Neighbor Embedding (t-SNE) scatter plot and demonstrates how sex affects the distribution of electrocardiogram (ECG) encodings in latent space of 1600 UKB participants. ECG encodings from female participants (small pink dots) and male participants (small brown dots) form distinct phenotypic clusters. The larger red and blue dots mark the centroids of these clusters, respectively. The orange arrow represents the phenotype vector for sex.

Supplementary Figure 6. Latent space 12-lead electrocardiogram phenome-wide association study perturbation tests.



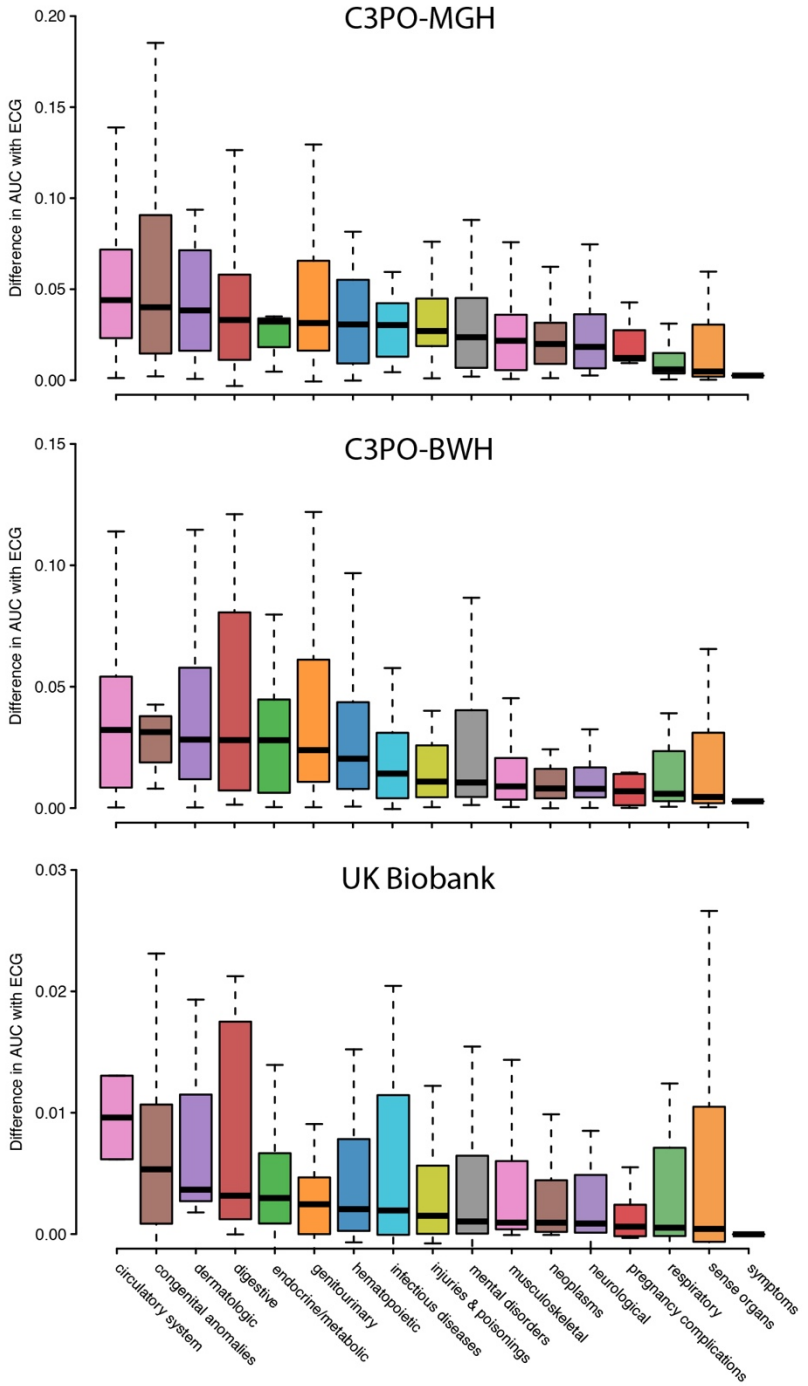
Panel A represents the results of the phenome-wide association study results in the Massachusetts General Hospital Community Care Cohort Project sample at baseline. Each phecode tested for association is represented as a single point on the plot. The x-axis represents the phenotype category and the y-axis represents the -log10(p-value) for the association test. The remaining panels represent the results in which the phecode labels are randomly reclassified with (B) 10%, (C) 20%, and D (100%) reclassification.

Supplementary Figure 7. Forest plots demonstrating associations for the top 3 phecodes in each of the datasets for both 12-lead and single-lead ECG latent space PheWAS analyses.



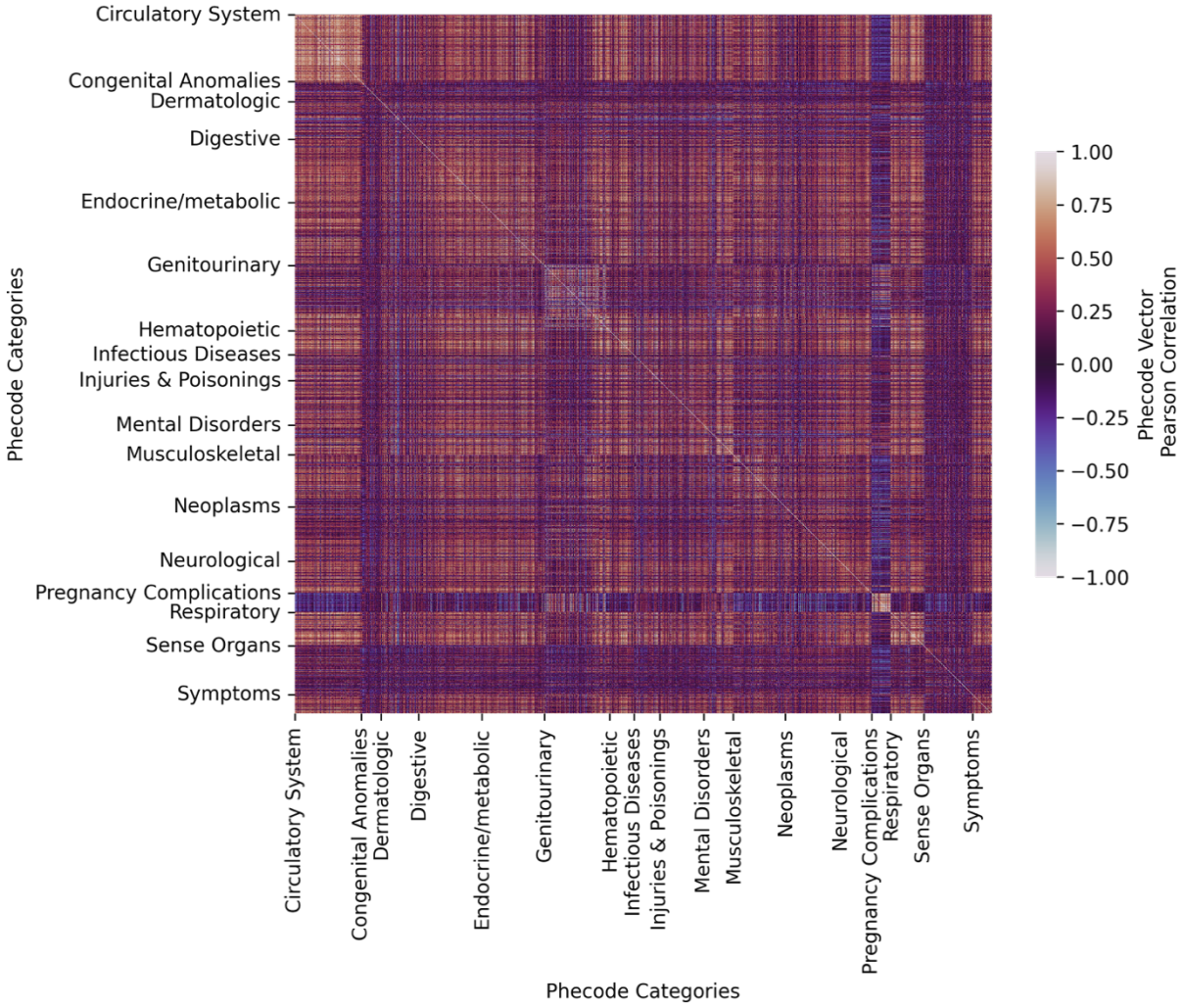
Depicted is a forest plot summarizing the top three strongest Phecode associations with the autoencoder model using 12-lead ECG (top) and single-lead ECG (bottom) in the primary meta-analysis. Points depict study-specific odds ratios (per 1-point increase in vector component score) and error bars 95% confidence intervals.

Supplementary Figure 8. Improvements in Phecode discrimination using ECG vector component scores



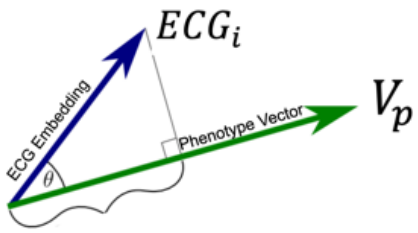
Plots depict the difference in the area under the receiver operator characteristic curve (AUC) between models with and without the ECG vector component scores, stratified by Phecode category and dataset. Only conditions meeting Bonferroni-corrected phenome-wide significance in the primary meta-analysis ($p<3.1x10^{-5}$) are included. Multivariable logistic regression models were fit with adjustment for age, sex, race, and vector component score. In the MGH-C3PO and BWH-C3PO datasets, models were further adjusted for ECG acquisition date and zero padding as described in the methods.

Supplementary Figure 9. Phenotype vector correlation matrix.



Correlation matrix for phenotype vectors based on the 12-lead electrocardiogram model, organized by disease category. The color scale represents Pearson correlation coefficients.

Supplementary Figure 10. Vector component score illustration



Depicted is an illustration of the calculation of vector component scores. Each ECG encoding ("ECG embedding", ECG_i) projects onto each phenotype vector, V_p. The projected component is calculated from the angle between the ECG encoding and the phenotype vector, scaled by the length of the ECG. Thus, the projected component signifies the latent space position of a single individual along a single phenotype vector and therefore represents a disease-specific "vector component score".

Supplementary Tables 1–10.
See separate data file.

Supplementary Table 11. Top associations by effect size across disease category for incident disease

Disease grouping*	N events	Odds ratio (95% CI)	p
Circulatory system			
Cardiac defibrillator in situ	166	1.80 (1.65-1.95)	3.70×10^{-43}
Right bundle branch block	468	1.77 (1.70-1.84)	7.47×10^{-183}
Bundle branch block	835	1.69 (1.63-1.75)	4.22×10^{-162}
Left bundle branch block	409	1.64 (1.57-1.72)	5.06×10^{-112}
Paroxysmal ventricular tachycardia	495	1.61 (1.53-1.70)	2.58×10^{-65}
Congenital anomalies			
Cardiac congenital anomalies	1051	1.33 (1.28-1.39)	1.46×10^{-40}
Cardiac and circulatory congenital anomalies	1307	1.32 (1.27-1.38)	2.04×10^{-41}
Valvular heart disease/ heart chambers	285	1.27 (1.18-1.38)	1.74×10^{-09}
Cardiac shunt/ heart septal defect	494	1.20 (1.13-1.28)	2.36×10^{-08}
Dermatologic			
Chronic ulcer of leg or foot	1147	1.24 (1.20-1.29)	5.41×10^{-37}
Chronic ulcer of unspecified site	532	1.24 (1.17-1.31)	1.13×10^{-13}
Decubitus ulcer	650	1.24 (1.17-1.30)	9.49×10^{-15}
Chronic ulcer of skin	1856	1.24 (1.20-1.27)	1.95×10^{-49}
Cellulitis and abscess of trunk	804	1.17 (1.12-1.22)	3.22×10^{-12}
Digestive			
Hemorrhage from gastrointestinal ulcer	319	1.34 (1.22-1.47)	7.84×10^{-10}
Portal hypertension	123	1.31 (1.18-1.46)	8.84×10^{-07}
Gastric ulcer	470	1.29 (1.20-1.38)	1.50×10^{-11}
Liver abscess and sequelae of chronic liver disease	267	1.28 (1.19-1.38)	9.40×10^{-11}
Splenomegaly	398	1.25 (1.18-1.31)	4.99×10^{-16}
Endocrine/Metabolic			
Cachexia	253	1.40 (1.27-1.54)	1.43×10^{-11}
Dysmetabolic syndrome X	367	1.37 (1.27-1.47)	7.44×10^{-18}
Other disorders of pancreatic internal secretion	341	1.35 (1.25-1.46)	2.42×10^{-13}
Disorders of magnesium metabolism	747	1.34 (1.27-1.41)	2.26×10^{-28}
Acid-base balance disorder	1320	1.33 (1.29-1.38)	1.26×10^{-60}
Genitourinary			
Renal dialysis	613	1.33 (1.27-1.39)	1.04×10^{-35}
End stage renal disease	363	1.31 (1.23-1.40)	8.49×10^{-18}
Nephritis and nephropathy in diseases classified elsewhere	184	1.31 (1.21-1.42)	4.25×10^{-11}
Acute renal failure	2959	1.29 (1.26-1.31)	2.71×10^{-119}
Chronic Kidney Disease, Stage III	1621	1.26 (1.22-1.30)	1.00×10^{-40}
Hematopoietic			
Anemia in chronic kidney disease	453	1.26 (1.19-1.33)	1.07×10^{-15}
Acute posthemorrhagic anemia	1306	1.25 (1.21-1.30)	4.48×10^{-34}
Hemorrhagic disorder due to intrinsic circulating anticoagulants	170	1.25 (1.14-1.37)	3.74×10^{-06}
Secondary thrombocytopenia	547	1.24 (1.17-1.31)	1.40×10^{-14}
Encounter for long-term (current) use of anticoagulants	2661	1.24 (1.21-1.27)	7.91×10^{-63}
Infectious diseases			
Septicemia	1848	1.31 (1.28-1.35)	5.20×10^{-78}
Bacteremia	905	1.27 (1.22-1.33)	3.17×10^{-28}
Infection/inflammation of internal prosthetic device; implant; and graft	601	1.26 (1.20-1.32)	1.34×10^{-20}
Gram positive septicemia	173	1.26 (1.14-1.39)	5.80×10^{-06}
Methicillin resistant Staphylococcus aureus	340	1.24 (1.14-1.35)	3.60×10^{-07}
Injuries and poisonings			

Non-healing surgical wound	181	1.34 (1.22-1.48)	5.08×10^{-09}
Traumatic amputation	124	1.33 (1.17-1.52)	2.19×10^{-05}
Certain early complications of trauma or procedure	463	1.31 (1.22-1.40)	2.84×10^{-15}
Septic shock	561	1.30 (1.23-1.37)	2.82×10^{-23}
Sepsis	1452	1.30 (1.25-1.34)	1.37×10^{-56}
Mental disorders			
Delirium due to conditions classified elsewhere	1193	1.26 (1.22-1.31)	7.90×10^{-37}
Altered mental status	2465	1.24 (1.21-1.27)	5.92×10^{-53}
Alcoholic liver damage	176	1.23 (1.13-1.34)	3.81×10^{-06}
Psychosis	540	1.22 (1.15-1.29)	1.70×10^{-12}
Alteration of consciousness	1357	1.18 (1.14-1.23)	4.87×10^{-19}
Musculoskeletal			
Acute osteomyelitis	209	1.30 (1.20-1.41)	6.41×10^{-11}
Unspecified osteomyelitis	349	1.28 (1.21-1.35)	9.71×10^{-18}
Osteomyelitis	446	1.26 (1.20-1.33)	2.50×10^{-18}
Chronic osteomyelitis	177	1.24 (1.13-1.37)	5.86×10^{-06}
Osteomyelitis, periostitis, and other infections involving bone	474	1.24 (1.18-1.3)	6.46×10^{-17}
Neoplasms			
Cancer of connective tissue	240	1.23 (1.12-1.34)	8.73×10^{-06}
Cancer of bone and connective tissue	388	1.18 (1.11-1.26)	1.36×10^{-07}
Cancer within the respiratory system	1045	1.17 (1.12-1.23)	4.84×10^{-13}
Cancer of bronchus; lung	961	1.17 (1.12-1.22)	4.32×10^{-13}
Secondary malignant neoplasm of liver	374	1.17 (1.10-1.24)	6.45×10^{-07}
Neurological			
Encephalopathy, not elsewhere classified	754	1.24 (1.18-1.31)	8.10×10^{-16}
Chronic pain syndrome	662	1.20 (1.13-1.27)	1.25×10^{-09}
Convulsions	1054	1.16 (1.12-1.21)	1.04×10^{-13}
Epilepsy, recurrent seizures, convulsions	1207	1.15 (1.11-1.20)	2.37×10^{-14}
Other conditions of brain	1712	1.15 (1.11-1.19)	3.71×10^{-16}
Respiratory			
Pseudomonal pneumonia	103	1.62 (1.43-1.84)	2.79×10^{-14}
Pulmonary congestion and hypostasis	1958	1.35 (1.31-1.39)	5.26×10^{-107}
Respiratory failure	1134	1.35 (1.30-1.40)	2.16×10^{-52}
Emphysema	759	1.32 (1.25-1.39)	8.90×10^{-26}
Pneumonitis due to inhalation of food or vomitus	692	1.31 (1.25-1.37)	2.57×10^{-28}
Sense organs			
Other nondiabetic retinopathy	282	1.18 (1.10-1.27)	3.52×10^{-06}
Blindness and low vision	769	1.11 (1.06-1.17)	3.52×10^{-05}
Presbyopia	1118	1.07 (1.04-1.10)	1.41×10^{-06}
Dry eyes	1579	1.06 (1.03-1.09)	3.18×10^{-05}
Senile cataract	3649	1.05 (1.03-1.07)	5.34×10^{-10}
Symptoms			
Cardiogenic shock	313	1.45 (1.34-1.56)	1.31×10^{-22}
Shock	747	1.38 (1.32-1.44)	2.71×10^{-44}
Gangrene	209	1.37 (1.25-1.50)	5.46×10^{-12}
Debility unspecified	223	1.28 (1.15-1.43)	4.77×10^{-06}
Nonspecific abnormal findings on radiological and other examination of other intrathoracic organs (echocardiogram, etc)	1632	1.14 (1.10-1.18)	1.41×10^{-13}

*Displayed are significant associations with the top 5 largest effect sizes within each disease category. In cases where there are fewer than 5 significant associations, all significant associations are shown.

Supplementary Tables 12–13.
See separate data file.