Emergent Symbolic Structure in Health Foundation Models: Extraction, Alignment, and Cross-Modal Transfer

Gajendra Katuwal [1] **Advait Koparkar** [1] **Salar Abbaspourazad** [1] **Anshuman Mishra** [1] **Sarvesh Kirthivasan** [1]

Abstract

We show that information can be transferred post-hoc across independently trained health foundation models (FMs), each pretrained on ∼20M minutes of wearable sensor data from ∼172K participants, by aligning their data-dependent coordinate systems. From frozen embeddings we extract candidate symbol-like components using linear decomposition methods, and align them across models with simple linear maps. Aligned symbols associate selectively with health conditions and physiological attributes, with associations similar across modalities and architectures. A classifier trained on one model's symbols and applied to another retains more than 95% of its in-domain performance, with similar retention in both directions. Overall, our results indicate that independently trained health FMs converge toward a common representation of the same underlying physiology.

1. Introduction

Health FMs trained on wearable biosignals (Spathis et al., 2020; Abbaspourazad et al., 2023; 2024; Narayanswamy et al., 2024; Pillai et al., 2025; Gu et al., 2025) have demonstrated that health-relevant structure can be learned from large amounts of unlabeled sensor data. However, like other deep learning systems, they lack explicit symbolic structure for transparent, structured reasoning (Kahneman, 2011). Bridging this gap is a central goal of neurosymbolic AI (Garcez & Lamb, 2020), and is especially pressing in healthcare, where interpretability and structured decision-making are critical. Even partial progress, surfacing symbol-like components from learned representations, would aid interpretability and unlock new capabilities such as cross-modal transfer. A natural question is whether such symbols must be imposed externally, or whether they already emerge within the learned representations and can be surfaced post-hoc.

Recent work provides theoretical motivation to expect that interpretable, shared structure exists within neural embeddings. The Linear Representation Hypothesis (LRH) (Elhage et al., 2024; Park et al., 2024), supported by mechanistic interpretability work (Olah et al., 2020; Elhage et al., 2022a), posits that neural networks encode individual concepts as linear directions in representation space. The Universality Hypothesis (Li et al., 2016; Olah et al., 2020; Chughtai et al., 2023) and the Platonic Representation Hypothesis (Huh et al., 2024) go further, suggesting that networks trained under differing architectures, modalities, and objectives converge toward similar latent structures, implying that symbol-like components may be partially shared across modalities despite very different sensing mechanisms.

Motivated by these hypotheses, we ask:

> *Do health FM embeddings admit a symbol-like decomposition that is consistent across modalities? If so, do the resulting components carry interpretable health meaning, and can they support cross-modal transfer without joint training?*

Such shared symbolic structure is not directly visible in raw embeddings, where latent factors are entangled across coordinates (Hinton et al., 1986; Bengio et al., 2013) and feature superposition compounds the problem (Olah et al., 2020; Elhage et al., 2022b). We address this by re-expressing *frozen embeddings* in data-dependent coordinate systems whose components are more interpretable and comparable across models (§3). We use *symbol* as an operational term for a component with selective health associations that can be extracted, aligned, and transferred across modalities. Like classical symbols (Newell & Simon, 1976), these are manipulable units of knowledge, but grounded in continuous embedding geometry rather than logical predicates. We evaluate them empirically as candidate symbol-like factors, not as symbols in the strong cognitive sense.

Authors ordered by contribution. [1]Apple. Correspondence to: Gajendra Katuwal <gkatuwal@apple.com>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

Contributions We introduce a unified framework for post-hoc symbol extraction, alignment, and transfer from frozen embeddings without joint training (§3). To our knowledge, this is the first work to study emergent symbolic structure in health FMs and at this scale (∼20M minutes, ∼172K participants). Using this framework we show:

1. Extracted components exhibit selective associations with health conditions and physiological targets, and these associations are partially shared across sensing modalities and architectures (§4.1, §4.2).
2. Embeddings from different modalities share a subspace (as defined in §3.2.2) rich in physiological information that simple linear maps can recover, enabling cross-modal transfer without joint training from only a fraction of the paired cohort (§4.3).

2. Related Work

Our framework connects to two lines of research: decomposing neural representations into interpretable units, and aligning representations across domains.

Interpretability through decomposition. Sparse autoencoders recover monosemantic features from polysemantic neurons (Bricken et al., 2023; Cunningham et al., 2024), scaling to frontier models (Templeton et al., 2024; Gao et al., 2024), extending to causal circuits (Marks et al., 2024) and multi-dimensional features (Engels et al., 2025). (Yamagiwa et al., 2023) showed that ICA applied to word and image embeddings reveals shared geometric structure across modalities. Concept-based methods explain models through predefined directions (Kim et al., 2018; Koh et al., 2020; Oikarinen et al., 2023) or population-level representation engineering (Zou et al., 2023; Marks & Tegmark, 2024). Our work differs in that we study health FM embeddings, require no concept supervision, and discover symbols from frozen embeddings using classical linear decompositions evaluated post-hoc against clinical attributes.

Cross-modal alignment and transfer. Most cross-modal alignment methods require joint training (Radford et al., 2021). Post-hoc alternatives include model stitching, which learns a linear layer between intermediate representations (Bansal et al., 2021), and relative representations, which re-express embeddings as similarities to shared anchor points (Moschella et al., 2023). Comparison measures such as Centered Kernel Alignment (CKA) (Kornblith et al., 2019) quantify representational similarity but do not produce a mapping for transfer; for a broader taxonomy see (Sucholutsky et al., 2023). Our approach operates on frozen embeddings through explicit symbol alignment (one-to-one matching or linear map), enabling both transfer and interpretability within the same framework.

3. Method

3.1. Dataset and Model Pretraining

We use three separately trained FMs, each with its own architecture, modality, and training objective rather than a single multi-modal model. Each is trained on approximately 20 million minutes of unlabeled sensor data collected under informed consent from the same ∼172 thousand participants in the Apple Heart and Movement Study (AHMS) (MacRae, 2021):

- Vision Transformer trained on PPG (PPG ViT)
- EfficientNet trained on PPG (PPG EfficientNet)
- Vision Transformer trained on accelerometer (Accel ViT)

See (Abbaspourazad et al., 2023) and (Abbaspourazad et al., 2024) for architecture and pretraining details. Note that the Accel ViT used here is the uni-modal encoder trained solely on accelerometer data, not the knowledge-distilled variant from (Abbaspourazad et al., 2024).

Each model produces a d=256-dimensional embedding per 1 min input window. We aggregate these into subject-level embeddings by averaging across all windows for a given participant, yielding one vector $h \in \mathbb{R}^d$ per subject per model. This averaging discards within-subject temporal structure in favor of a stable per-subject representation suitable for population-level analysis. Symbol extraction, alignment, meaning analysis, and transfer experiments all use a held-out cohort of 30 thousand subjects not seen during pretraining. We split this cohort 80/20 into an operator-fit set and an evaluation set. Symbol operators and downstream classifiers are fitted on the operator-fit set. All reported metrics use the evaluation set with frozen operators.

3.2. Symbolic Framework

The framework consists of two steps: *symbol extraction*, which re-expresses each model's embeddings as candidate symbols, and *symbol alignment*, which maps symbols across models.

3.2.1. Symbol Extraction

We view symbol extraction as applying an *extraction operator* $\mathcal{E}$ to frozen embeddings, re-expressing each model in a new component representation whose factors are candidate symbols. Each operator acts as a *representation lens*: a different way of viewing the same embedding space rather than a competing method, with the choice of extractor depending on the downstream use case. In this paper, we focus on linear decomposition methods for symbol extraction.

Let $H_m \in \mathbb{R}^{n \times d}$ denote the matrix of subject-level embeddings from model m, where n is the number of subjects

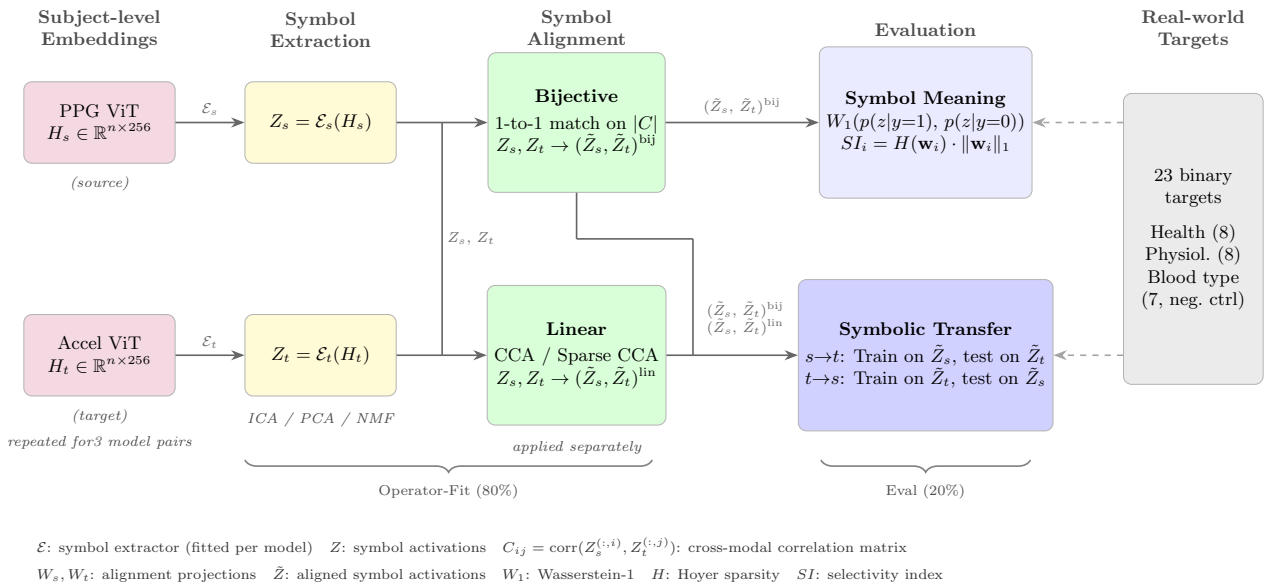

Figure 1. Overview of the symbolic framework. Frozen embeddings are re-expressed as symbol activations via linear decompositions (ICA, PCA, or NMF), then aligned across models using bijective matching (Hungarian algorithm on $|C|$) or CCA. Aligned symbols are evaluated for selective meaning (W1 distance against 23 targets) and cross-modal transfer (classifier trained on source, tested on aligned target). All operators are fitted on an 80% operator-fit split of a 30K held-out cohort; metrics use the remaining 20%.

and d is the embedding dimensionality (d=256). A linear decomposition represents these embeddings as coefficients over a learned symbol dictionary:

$$H_m \approx Z_m S_m^\top, \tag{1}$$

where $S_m \in \mathbb{R}^{d\times k}$ contains the candidate symbol components and $Z_m \in \mathbb{R}^{n\times k}$ contains their activations across subjects. In this work we use $k = d = 256$. More generally, a symbol extractor $\mathcal{E}_m$ maps embeddings to activations,

$$Z_m = \mathcal{E}_m(H_m). \tag{2}$$

We instantiate $\mathcal{E}_m$ using three symbol extractors: Principal Component Analysis (PCA), orthogonal directions capturing maximal variance; Independent Component Analysis (ICA), components capturing statistically independent sources; and Non-negative Matrix Factorization (NMF), yielding parts-based, additive components. These methods impose distinct structural constraints on the components and activations (Table 3), offering complementary lenses on the symbolic structure of representation spaces.

In all cases, the extracted symbols are fixed linear components of the embedding space, and Z_m denotes the corresponding activation or coefficient matrix. Thus these methods surface linearly recoverable structure in the frozen embeddings. The framework also accommodates nonlinear extractors such as sparse autoencoders; extending the analysis to these methods is left to future work (Supplementary 7.8).

3.2.2. Symbol Alignment

Given symbol activations $Z_{\text{source}}, Z_{\text{target}} \in \mathbb{R}^{n\times d}$ after symbol extraction from two domains, alignment finds operators W_s and W_t that map both into a shared coordinate system where corresponding components are maximally correlated. We use two alignment techniques.

(1) Bijective alignment. This technique finds a one-to-one matching that maximizes pairwise similarity, preserving individual symbol identity for interpretability and meaning analysis. We compute a similarity matrix

$$C_{ij} = \text{corr}\big(Z_{\text{source}}^{(:,i)},\ Z_{\text{target}}^{(:,j)}\big), \tag{3}$$

where $Z^{(:,i)}$ denotes the activation of the ith symbol across all subjects. The optimal one-to-one assignment is found via the Hungarian algorithm (Kuhn, 1955) on $|C|$, yielding W_s and W_t that establish bijective correspondence between symbols of the source and target domains, with polarity corrections ensuring all matched pairs are positively correlated. After Hungarian matching on $|C|$, we flip the sign of any target component with $C_{ij} < 0$ so each matched pair is positively correlated. We keep every matched pair and use $|C_{ij}|$ as its correlation strength.

(2) Linear alignment. This technique relaxes the one-to-one constraint, allowing shared structure to span multiple dimensions. It is better suited for cross-modal transfer where maximizing overall alignment matters more than preserving

Table 1. Target condition groups for symbol meaning analysis.

Group	Labels	n	Role
Health conditions	cardiac, arrhythmia, vascular, cv risk, mental, sleep, hepatorenal, musculoskeletal	8	Positive
Physiological markers	sex (male), age (18–40, 40–60, 60+), BMI (<18.5, 18.5–25, 25–30, 30+)	8	Positive
Blood type	A+, A−, B+, B−, AB+, O+, O−	7	Negative

Positive: health conditions and physiological markers affect PPG waveform morphology and motion patterns through known physiological pathways. *Negative*: ABO/Rh antigens have no proximal mechanism to affect PPG or motion signals. Total: 23 conditions across 3 groups.

individual symbol identity. Canonical Correlation Analysis (CCA) (Hotelling, 1936) finds linear projections W_s, W_t of each domain's symbols into a shared canonical space by maximizing cross-domain correlation:

$$(W_s^*, W_t^*) = \arg\max_{W_s, W_t} \operatorname{tr}\big(W_s^\top \Sigma_{st} W_t\big) - \lambda\big(\|W_s\|_1 + \|W_t\|_1\big), \tag{4}$$

subject to $W_s^\top \Sigma_{ss} W_s = I$ and $W_t^\top \Sigma_{tt} W_t = I$, where Σ_{st} is the cross-covariance between symbol activations and Σ_{ss}, Σ_{tt} are their respective covariance matrices. Standard CCA sets $\lambda = 0$ and increasing λ enforces sparser projections known as sparse CCA (Witten et al., 2009), constraining each canonical direction to depend on only a few source symbols. Concretely, CCA returns linear maps W_s, W_t such that the aligned coordinates $Z_s W_s$ and $Z_t W_t$ are maximally correlated coordinate-by-coordinate. When we refer to a *shared subspace* in what follows, we mean the span of these correlated canonical directions. This is a property of the aligned coordinates, not a claim that the original embedding spaces share a canonical, model-independent subspace.

3.3. Evaluation

We evaluate the symbolic framework against a set of target labels derived from participant metadata in the AHMS, organized into three groups comprising 23 binary targets (Table 1). Health condition labels are constructed from participants' self-reported survey responses following the protocol in (Abbaspourazad et al., 2023); we aggregate related items into eight condition categories (Supplementary Table 2). Physiological markers (sex, age, BMI) are from enrollment data. Blood type is user-reported via the Apple Health app profile. Health conditions and physiological markers serve as *positive controls* with expected physiological signal; blood type serves as a *negative control* with no expected signal in either modality.

3.3.1. Quantifying Symbol Meaning

We quantify the association of a symbol to a target by measuring the groupwise difference in activation along that symbol direction using the Wasserstein-1 (W1) distance (Villani, 2009), with Cohen's d as a robustness check (see Supplementary 7.2 for details). Computing W1 for each symbol-target pair yields a condition profile $\mathbf{w}_i \in \mathbb{R}^T$ per symbol i.

A high-quality symbol should associate strongly with a small number of targets rather than responding diffusely to many. We define the *Selectivity Index* to capture both the shape and magnitude of this profile:

$$\mathrm{SI}_i = H(\mathbf{w}_i) \cdot \|\mathbf{w}_i\|_1 \tag{5}$$

where H is the Hoyer sparsity (Hoyer, 2004) (shape) and $\|\cdot\|_1$ is the L1 norm (magnitude). See Supplementary 7.3 for derivation details.

3.3.2. Symbolic transfer learning

Having established that aligned symbols carry physiological meaning, we quantify how well this structure transfers across domains. We train a logistic regression classifier f_θ on source-domain symbols to predict target labels y, then apply it to aligned target-domain symbols (subjects are paired, so the same y serves as ground truth):

$$\theta^* = \arg\min_\theta \ \mathcal{L}\Big(f_\theta\big(\underbrace{\mathcal{E}_{\text{source}}(H_{\text{source}})\, W_s}_{\tilde{Z}_{\text{source}}}\big),\, y\Big), \tag{6}$$

$$\hat{y}_{\text{target}} = f_{\theta^*}\Big(\underbrace{\mathcal{E}_{\text{target}}(H_{\text{target}})\, W_t}_{\tilde{Z}_{\text{target}}}\Big). \tag{7}$$

For transfer, both the alignment operators and source-domain classifiers are fit only on the operator-fit split; all AUCs are computed on held-out evaluation subjects. We use AUC (area under the receiver operating characteristic curve) as the primary transfer metric (see Supplementary 7.4 for justification). We define *transfer retention* as

$$\text{Transfer Retention} = \frac{\mathrm{AUC}_{\text{transfer}}}{\mathrm{AUC}_{\text{in-domain}}}, \tag{8}$$

where $\mathrm{AUC}_{\text{transfer}}$ is the AUC of the source classifier applied to aligned target symbols and $\mathrm{AUC}_{\text{in-domain}}$ is the AUC when both training and evaluation use the same domain. A value near 100% indicates nearly lossless transfer. We evaluate transfer separately for positive-control targets (health conditions, physiological markers) and the negative-control group (blood type), expecting the former to transfer and the latter to remain near chance.

3.3.3. Transfer efficiency

A practical question for post-hoc alignment is how many paired subjects are needed to learn a useful mapping be-

tween independently trained models. If the shared structure is simple, few paired subjects may suffice. Whereas if it is complex, more pairs will be needed. We evaluate this by sweeping the training fraction $\rho \in \{0.1, 0.2, \ldots, 1.0\}$ of participant overlap between domain pairs and measuring transfer retention as a function of ρ.

4. Results

4.1. Do Symbols Align Across Modalities?

If independently trained models on different sensor modalities converge on a shared representational structure, their learned features should align across modalities. Figure 2 shows results for PPG EfficientNet vs Accel ViT, the hardest pair (different modality and architecture); the other two pairs show similar patterns.

Cross-modal correspondence is present in embeddings and symbol extraction reshapes it. Comparing raw embedding dimensions without alignment yields near-zero cross-modal correlation. Bijective alignment recovers substantial correspondence even on raw embeddings (Figure 2a), and symbol extraction further reshapes how this correspondence is distributed across components. Overall, cross-modal correspondence is already present in the raw embeddings; symbol extraction reshapes how it is distributed, with ICA concentrating it into a few top symbols and NMF spreading it more broadly (see Supplementary 7.1 for details).

CCA reveals shared cross-modal linear structure. Differences between extractors largely disappeared under CCA alignment (Figure 2c). All representations converged to near-identical performance ($r \approx 0.99$ at top ranks). This convergence is expected because CCA (Hotelling, 1936) solves for the linear projections that maximize cross-modal correlation, effectively bypassing whatever basis the extractor produced. These results suggest that PPG and accelerometer embeddings share a linearly accessible subspace, and that CCA recovers a maximally correlated coordinate system for it. Whether this subspace encodes meaningful physiology is tested in Section 4.2.

Sparse CCA preserves alignment with interpretability. As shown above, CCA alignment is insensitive to the choice of extractor, but its canonical projections are dense and individual symbols difficult to interpret. We examined whether interpretability could be improved by imposing sparsity on the CCA projection weights without sacrificing alignment (Figure 2d). Low-to-moderate L1 regularization ($\lambda \leq 2 \times 10^{-2}$) maintained $> 95\%$ of standard CCA's alignment capacity while yielding sparser projections. Only at high sparsity ($\lambda > 4 \times 10^{-2}$) did alignment degrade

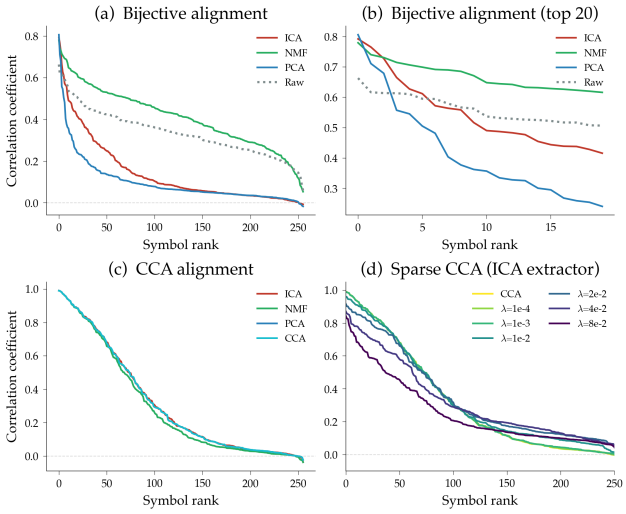

Figure 2. Cross-modal symbol alignment analysis [PPG EfficientNet vs Accel ViT]. **(a)** Bijective matching across symbol extractors: cross-modal correspondence is already present in raw embeddings (black dotted). Symbol extraction reshapes how it is distributed — ICA concentrates it into a few top-ranked symbols, while NMF spreads it more broadly across ranks. **(b)** Zoomed view of top-20 symbols showing ICA's stronger peak versus NMF's consistently high alignment. **(c)** CCA alignment: all extractors converge to near-identical curves, indicating strongly shared cross-modal linear structure that is less sensitive to extraction basis. **(d)** Sparse CCA: L1 penalty on projection weights $\leq 2 \times 10^{-2}$ preserves alignment while yielding sparser projections suggesting that much of the shared linear structure can be described by relatively few dimensions.

substantially. This suggests that much of the shared linear structure can be captured by a relatively small number of contributing dimensions.

4.2. Do Symbols Carry Interpretable Health Meaning?

Having established that symbols align across modalities, we next ask whether they carry selective, interpretable meaning. We focus on PPG ViT vs Accel ViT (Figure 3), which share the same architecture so that differences in symbol meaning reflect modality rather than model design; other model pairs show similar patterns (Supplementary Figure 6).

Symbols show semantic grounding and selective meaning. We focus on ICA because its independence constraint produces the sparsest and most interpretable target profiles. Figure 3a shows W1 heatmaps for PPG ViT and Accel ViT, where each cell shows the association strength between the corresponding symbol (column) and target (row). Individual symbols activate for specific targets while most symbol-target pairs are near zero, quantified per symbol by the Selectivity Index (Eq. 5). The same selective structure appears in both modalities, confirming that the underlying

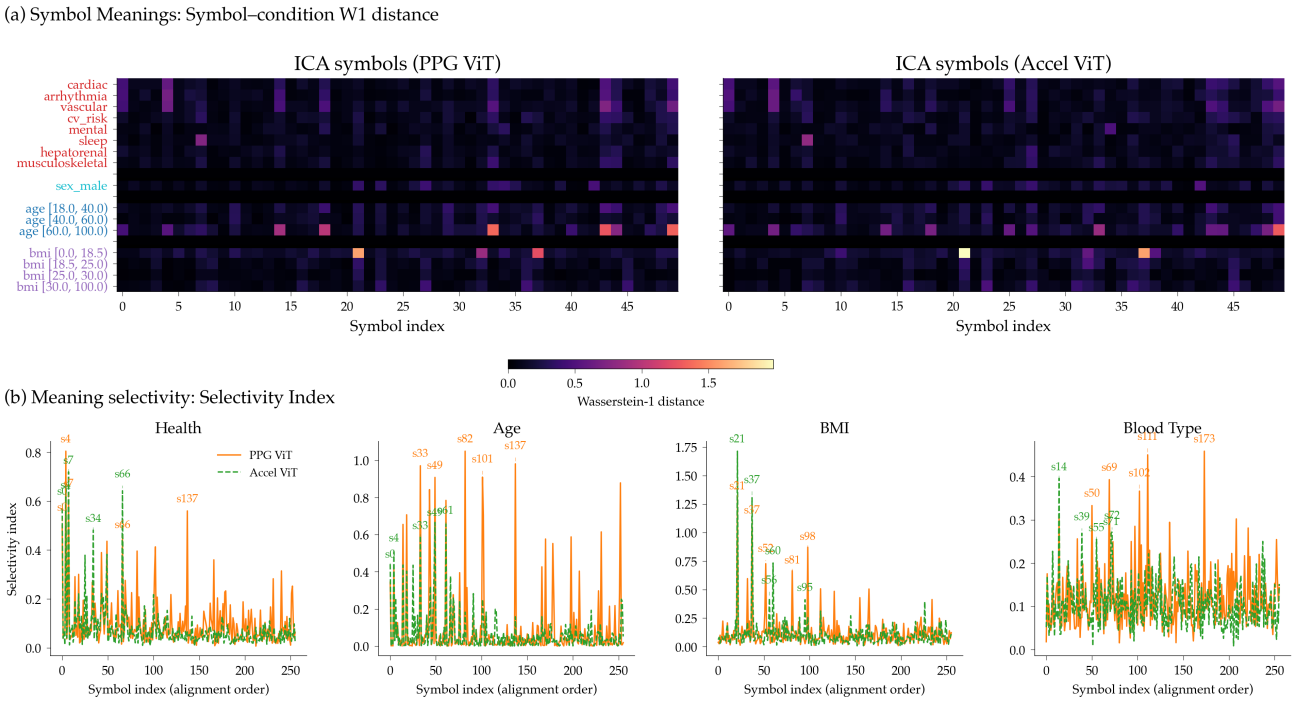

Figure 3. Symbol meaning and selectivity (ICA, PPG ViT vs Accel ViT). **(a)** W1 heatmaps for the 50 most cross-modally aligned ICA symbols: each column is a symbol (an extracted direction in embedding space), each row a health or physiological target, and each cell the W1 distance between the symbol's activation distributions for target-positive vs target-negative subjects (higher values indicate a stronger association). PPG ViT (left) and Accel ViT (right) exhibit the same selective structure — most cells are near zero while a few symbols activate for specific targets. The full set of 256 symbols, the raw-embedding comparison, and the blood-type control are shown in Supplementary Figure 5. **(b)** Selectivity Index ($\mathrm{SI} = H \cdot \|\mathbf{w}\|_1$) per symbol in alignment order for each target group; the five highest-SI symbols per curve are labeled. Health conditions, age, and BMI reach high SI, whereas blood type (negative control) remains uniformly low in both modalities.

models encode shared health-relevant structure. The corresponding raw-embedding heatmaps, which are far more diffuse, and the blood-type negative control, which produces little or no signal in any representation, are shown in Supplementary Figure 5.

While detailed interpretation of individual symbols is beyond the scope of this paper, a few patterns are worth noting. Symbol S7 peaks at sleep and is the brightest accelerometer symbol, consistent with sleep patterns directly affecting motion. In the PPG heatmap, S7 does not overlap with other health targets, whereas heart-related conditions (cardiac, vascular, arrhythmia) share clustered symbols, reflecting common cardiovascular physiology. Hepatorenal and mental health targets show weaker cross-modal correspondence, consistent with more indirect physiological signatures in wearable data.

Figure 3b quantifies this selectivity using the Selectivity Index (SI, Eq. 5) per symbol in alignment order. Accelerometer SI peaks concentrate among the most cross-modally aligned symbols, whereas PPG distributes SI more broadly, consistent with the richer physiological encoding of the pulse waveform. Blood type SI remains uniformly low in both modalities and, unlike health conditions, shows no cross-modal consistency, confirming that the negative control carries neither selective signal nor shared structure.

Quantifying cross-modal meaning consistency. Having seen this consistency qualitatively in the heatmaps and SI concentration (Figure 3a,b), we measure it directly as the cosine similarity between each symbol's W1 target profile in PPG and Accel. Even raw embeddings reach a median cosine of 0.84, so the cross-modal structure is already present, and ICA sharpens it further, with 91% of symbols above cosine 0.8 under bijective alignment (median 0.89). Supplementary Figure 6 extends this across three model pairs and four extraction methods with broadly consistent results.

The selectivity results above are robust to replacing W1 with $|\text{Cohen's } d|$ as the group-difference metric (Supplementary Figure 7).

4.3. Can Symbols Support Cross-Domain Transfer?

We evaluated cross-modal symbolic transfer using all three foundation models and six directed domain pairs using the pipeline described in §3.3.2.

Symbols enable cross-modal transfer without joint training. Figure 4a summarizes transfer performance; best results are achieved with the CCA family of alignments. Phys-

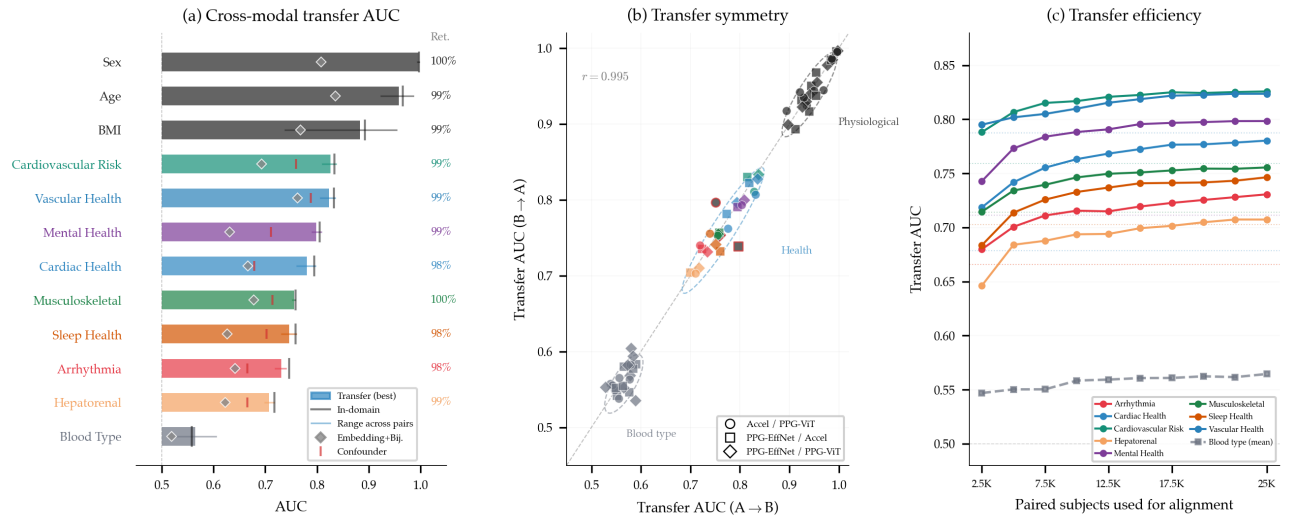

Figure 4. Cross-modal symbolic transfer. **(a)** Classifiers trained in one symbol space retain >98% of in-domain AUC after cross-modal alignment; blood type (negative control) remains near chance. Best transfer is achieved with the CCA family of alignments. **(b)** Transfer retention is similar in both directions ($r = 0.995$), consistent with a low-complexity alignment recoverable by linear maps; we report this empirically and do not claim a formal isomorphism between symbol spaces. **(c)** Transfer saturates with a fraction of the paired cohort, suggesting a low-complexity alignment. Overall, independently trained health models share physiologically rich structure that can be leveraged for cross-modal transfer without joint training.

iological markers transfer most readily (AUC 0.88–1.00, ≈100% retention). Health conditions follow with AUC 0.71–0.83 (98.7% ± 1.6% retention), while blood type remains near chance (AUC ≈ 0.56) as expected for a negative control (full per-target AUC in Supplementary Tables 4–6). Without symbol extraction and alignment, raw embedding transfer yields near-chance AUC (∼0.50). Bijective matching on raw embeddings (grey diamonds in Figure 4a) recovers partial transfer (health mean AUC = 0.67) but remains well below CCA, because raw dimensions are entangled and a single bijective mapping cannot disentangle them.

A natural concern is that cross-modal transfer merely propagates demographic information shared between modalities. To rule this out, we include a per-condition confounder baseline that predicts each health target from age, sex, and BMI alone (red ticks in Figure 4a). Every health condition exceeds its confounder baseline, confirming that symbolic transfer carries condition-specific information that demographics alone do not explain.

Transfer retention is similar in both directions. Figure 4b shows that forward and reverse transfer AUC are tightly correlated across all targets and domain pairs ($r = 0.995$, mean $|\Delta\text{AUC}| = 0.009$). The high retention above, together with its similarity across both transfer directions, indicates that the two representations carry much of the same predictive information with no direction-specific loss. How much paired data is needed to estimate the cross-modal alignment map (the linear map that places the two symbol spaces in correspondence) is a separate question that we examine next.

Transfer retention plateaus before the full paired cohort. We evaluate how much paired data is needed for effective cross-modal transfer by sweeping the number of paired subjects used for alignment (Figure 4c). All eight health conditions plateau above 95% retention by ∼7.5K paired subjects, roughly 30% of the ∼24K operator-fit cohort, with only marginal gains from additional data. We discuss the implications of this saturation in §5.

5. Discussion

A shared, physiologically meaningful structure across modalities. Regardless of which extractor or modality pair we use, CCA recovers near-identical cross-modal correlation (Figure 2c). This suggests that the coordinates CCA recovers, the shared subspace as defined in §3.2.2, are largely invariant to modality pairing and extraction basis (Lyu et al., 2022). Beyond this geometric alignment, aligned symbols carry consistent health associations across modalities, in line with the high transfer retention (>98%). This shared structure persists even for raw embeddings, suggesting that extraction sharpens rather than creates it, and that different extractors simply offer complementary lenses on the same structure. We read this as each modality offering a view of the same underlying physiological system, consistent with the Platonic Representation Hypothesis (Huh et al., 2024): independently trained models can converge toward compati-

ble latent structure when trained on different observations of the same process.

The cross-modal alignment is simple, with few effective degrees of freedom. Sparse CCA shows that much of the alignment is preserved even when many projection weights are driven to zero (Figure 2d), and transfer retention saturates with only a fraction of the paired cohort (§4.3). These results suggest that the alignment recovered in our data has few effective degrees of freedom. The similar retention in both transfer directions further supports this interpretation. Both modality-specific maps, W_s and W_t, recover the downstream-predictive coordinates with comparable fidelity. Much of the transferable signal therefore appears to lie in a shared predictive component, so alignment mainly brings corresponding coordinates into register rather than learning a full cross-modal translation.

A parallel to cross-lingual word embeddings. Independently trained word embedding spaces can often be brought into correspondence by simple linear maps, often approximately orthogonal and structure preserving (Xing et al., 2015; Smith et al., 2017; Artetxe et al., 2016). This was first shown to enable word translation from only a small seed dictionary (Mikolov et al., 2013), and later with no parallel data at all (Conneau et al., 2018). In our setting, the analogy is not a single source-to-target map, but the fact that independently trained representations can expose compatible coordinates when their inputs are different views of the same underlying structure. Pulse and motion are different views of the same physiology, so the practical consequence is similar: the cross-modal alignment can be estimated from a fraction of the paired cohort rather than learned from scratch.

Symbols as an interpretable interface: directions and open problems. Our results suggest that health FM embeddings already contain symbol-like structure that can be surfaced post-hoc. These symbols are not only interpretable within a single model but partially shared across modalities, enabling a common vocabulary through which separately trained models can be compared and composed. Connecting symbols to model predictions through causal circuits (Marks et al., 2024) would deepen mechanistic understanding of what drives health predictions. Exploring nonlinear extraction methods such as sparse autoencoders may surface structure missed by linear extractors. Individual symbol meaning analysis with comprehensive stratified analysis across subpopulations remains an important open direction. The symbolic layer also enables targeted modal unlearning, where specific symbols are suppressed to remove learned biases without retraining the underlying model. Symbolic reasoning over extracted components, composing symbols to derive higher-order health inferences, is a natural next step toward neurosymbolic health AI. Finally, projecting symbols from human-uninterpretable modalities such as accelerometry into more legible spaces such as PPG or ECG offers a path toward interpretability transfer, where the meaning of opaque sensor signals is explained in terms of physiologically familiar ones.

Limitations. Our extraction uses linear decompositions, which may miss nonlinear structure (Engels et al., 2025). We also emphasize that the alignment quality we report is empirical rather than theoretically guaranteed. Our claims about shared structure describe what these linear operators recover in the data-dependent coordinates they produce on this dataset, not an extrinsic or canonical correspondence between the embedding spaces. All three models share the same $\sim$172K training population, so alignment may partly reflect shared covariance; extending this analysis to models trained on entirely disjoint populations would further corroborate whether alignment reflects shared physiology rather than shared demographics. Three observations argue against a purely demographic explanation: the blood-type negative control produces no signal despite sharing population statistics; health conditions transfer with varying strength (cardiovascular strong, hepatorenal weak); and every condition exceeds a confounder baseline trained on age, sex, and BMI.

6. Conclusion

We showed that frozen health FM embeddings can be re-expressed via linear decompositions into symbol spaces that are selectively grounded in health conditions, partially shared across modalities, and useful for cross-modal transfer with limited paired data.

Acknowledgments

We would like to thank participants in the Apple Heart and Movement Study and study staff at The Brigham and Women's Hospital, without whom this work would not have been possible. We are grateful to Vincent Man, Nandita Bhaskhar, and Lin Yang for valuable internal reviews, and to Calum MacRae and Angella Spillane for helpful feedback on the manuscript. Finally, we thank Ian Shapiro, Lindsay Hislop, Laura Rhodes, and Jonathan Varbel for publication coordination.

References

Abbaspourazad, S., Elachqar, O., Miller, A. C., Emrani, S., Nallasamy, U., and Shapiro, I. Large-scale training of foundation models for wearable biosignals. *arXiv preprint arXiv:2312.05409*, 2023. URL https://arxiv.org/abs/2312.05409.

Abbaspourazad, S., Mishra, A., Futoma, J., Miller, A. C., and Shapiro, I. Wearable accelerometer foundation models for health via knowledge distillation. *arXiv preprint arXiv:2412.11276*, 2024. URL https://arxiv.org/abs/2412.11276.

Artetxe, M., Labaka, G., and Agirre, E. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2289–2294, 2016.

Bansal, Y., Nakkiran, P., and Barak, B. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. URL https://arxiv.org/abs/2106.07682.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features.

Chughtai, B., Nanda, N., Wattenberg, M., and Carter, S. A toy model of universality: Reverse engineering how networks learn group operations. *arXiv preprint arXiv:2302.03025*, 2023. URL https://arxiv.org/abs/2302.03025.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. Word translation without parallel data. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://arxiv.org/abs/1710.04087.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Elhage, N., Nanda, N., Olah, C., et al. Linear representations: The linear representation hypothesis. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/july-update/index.html#linear-representations.

Elhage, N. et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2022a. URL https://transformer-circuits.pub/2021/framework/index.html.

Elhage, N. et al. Toy models of superposition. *Transformer Circuits*, 2022b.

Engels, J., Michaud, E. J., Liao, I., Gurnee, W., and Tegmark, M. Not all language model features are linear. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2405.14860.

Gao, L., Dupré la Tour, T., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., and Wu, J. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024. URL https://arxiv.org/abs/2406.04093.

Garcez, A. S. d. and Lamb, L. C. Neurosymbolic artificial intelligence: The state of the art. *arXiv preprint arXiv:2012.05876*, 2020. URL https://arxiv.org/abs/2012.05876.

Gu, H., Tang, M., Han, C., Sangha, V., et al. Sensing cardiac health across scenarios and devices: A multimodal foundation model pretrained on 1.7 million individuals. *arXiv preprint arXiv:2507.01045*, 2025. URL https://arxiv.org/abs/2507.01045.

Hinton, G. E., McClelland, J. L., and Rumelhart, D. E. Distributed representations. *Parallel Distributed Processing*, 1986.

Hotelling, H. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

Hoyer, P. O. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004.

Huh, M., Cheung, B., Wang, T., and Isola, P. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024. URL https://arxiv.org/abs/2405.07987.

Hyvärinen, A. and Oja, E. Independent component analysis: Algorithms and applications. *Neural Networks*, 13:411–430, 2000.

Kahneman, D. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., and Sayres, R. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine*

Learning (ICML), 2018. URL https://arxiv.org/abs/1711.11279.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. Concept bottleneck models. In *International Conference on Machine Learning (ICML)*, 2020. URL https://arxiv.org/abs/2007.04612.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, 2019. URL https://arxiv.org/abs/1905.00414.

Kuhn, H. W. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2): 83–97, 1955.

Lee, D. and Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

Li, Y., Yosinski, J., Clune, J., and Lipson, H. Convergent learning: Do different neural networks learn the same representations? In *International Conference on Learning Representations (ICLR)*, 2016. URL https://arxiv.org/abs/1511.07543.

Lyu, Q., Fu, X., Wang, W., and Lu, S. Understanding latent correlation-based multiview learning and self-supervision: An identifiability perspective. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2106.07115.

MacRae, C. A. Apple heart & movement study, 2021. URL https://clinicaltrials.gov/study/NCT04198194. Clinical trial registration NCT04198194, clinicaltrials.gov.

Marks, S. and Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2024. URL https://arxiv.org/abs/2310.06824.

Marks, S., Rager, C., Michaud, E. J., Belinkov, Y., Bau, D., and Mueller, A. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *arXiv preprint arXiv:2403.19647*, 2024. URL https://arxiv.org/abs/2403.19647.

Mikolov, T., Le, Q. V., and Sutskever, I. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013. URL https://arxiv.org/abs/1309.4168.

Moschella, L., Maiorca, V., Fumero, M., Norelli, A., Locatello, F., and Rodolà, E. Relative representations enable zero-shot latent space communication. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2209.01605.

Narayanswamy, G., Liu, X., Ayush, K., Yang, Y., Xu, X., Liao, S., Garrison, J., Tailor, S., Sunshine, J., Liu, Y., Althoff, T., Narayanan, S., Kohli, P., Zhan, J., Malhotra, M., Patel, S., Abdel-Ghaffar, S., and McDuff, D. Scaling wearable foundation models. *arXiv preprint arXiv:2410.13638*, 2024. URL https://arxiv.org/abs/2410.13638.

Newell, A. and Simon, H. A. Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, 19(3):113–126, 1976.

Oikarinen, T., Das, S., Nguyen, L. M., and Weng, T.-W. Label-free concept bottleneck models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2304.06129.

Olah, C. et al. Zoom in: An introduction to circuits. *Distill*, 2020.

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2024. URL https://arxiv.org/abs/2311.03658.

Pillai, A. S., Spathis, D., Kawsar, F., and Malekzadeh, M. PaPaGei: Open foundation models for optical physiological signals. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2410.20542.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. URL https://arxiv.org/abs/2103.00020.

Smith, S. L., Turban, D. H. P., Hamblin, S., and Hammerla, N. Y. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/abs/1702.03859.

Spathis, D., Perez-Pozuelo, I., Brage, S., Wareham, N. J., and Mascolo, C. Self-supervised transfer learning of physiological representations from free-living wearable data. In *NeurIPS ML for Mobile Health Workshop*, 2020. URL https://arxiv.org/abs/2011.12121.

Sucholutsky, I., Muttenthaler, L., Weller, A., Peng, A., Bobu, A., Kim, B., Love, B. C., et al. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*, 2023. URL https://arxiv.org/abs/2310.13018.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Villani, C. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009.

Witten, D. M., Tibshirani, R., and Hastie, T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009. doi: 10.1093/biostatistics/kxp008.

Xing, C., Wang, D., Liu, C., and Lin, Y. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 1006–1011, 2015.

Yamagiwa, H., Oyama, M., and Shimodaira, H. Discovering universal geometry in embeddings with ICA. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. URL https://arxiv.org/abs/2305.13175.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, J. Z., and Hendrycks, D. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023. URL https://arxiv.org/abs/2310.01405.

7. Supplementary

7.1. Symbol Alignment Details

Applying bijective matching to raw embedding dimensions recovers substantial cross-modal correspondence (Head-10 ≈ 0.56, area under the correlation curve ≈ 0.32), establishing a strong baseline. ICA yields the strongest top-ranked matches but decays rapidly, suggesting it isolates a small set of highly shared modes. In contrast, NMF maintains strong alignment across a much broader range of symbols (area under the correlation curve: 0.39 vs. 0.21 for ICA; Figure 2b), indicating a more distributed shared vocabulary. This behavior is consistent with their inductive biases: ICA favors statistically independent sources (Hyvärinen & Oja, 2000), whereas NMF's non-negativity promotes parts-based factors that align more consistently across many components (Lee & Seung, 1999).

7.2. Symbol Meaning Quantification

Letting z denote a symbol activation and $y \in \{0, 1\}$ indicate condition presence, we compute the Wasserstein-1 (W1) distance between $p(z \mid y{=}1)$ and $p(z \mid y{=}0)$. Embeddings are z-scored to unit variance per symbol before computing W1, so that values are in units of standard deviations.

We use two complementary metrics. W1 captures full distributional differences (shifts, spreads, and shape changes), making it sensitive to any systematic difference between groups. However, W1 is influenced by sample size imbalance, producing a nonzero floor for minority groups even under the null. Cohen's d measures the standardized mean difference and has $\mathbb{E}[d] = 0$ under the null regardless of group size, eliminating this floor. However, it only captures mean shifts and misses higher-order distributional differences. Together, they complement each other: W1 provides a complete distributional view while Cohen's d confirms that observed selectivity reflects genuine effect sizes rather than sample-size artifacts (Figure 7).

7.3. Selectivity Index Details

Both the *shape* (how peaked the profile is) and the *magnitude* (how strong the associations are) matter for assessing symbol quality: a symbol with profile $[0.002, 0.001, 0.001]$ is peaked but indistinguishable from noise. Standard sparsity measures such as the Hoyer index (Hoyer, 2004) capture only shape, not magnitude. The Selectivity Index (Eq. 5) combines both. L1 is preferred over L2 because L2 upweights large values via squaring, which is redundant with Hoyer's own sensitivity to concentrated peaks. The product $\text{Hoyer} \times \|\mathbf{w}\|_1$ gives a clean factorization: Hoyer captures shape (how concentrated the profile is) and L1 captures magnitude (how large the associations are).

7.4. Choice of AUC as Transfer Metric

Because symbol alignment optimizes correlation rather than scale, the predicted distribution may shift relative to the source domain. AUC is invariant to monotonic shifts in the predicted scores. It depends only on the ranking of predictions, not their absolute values. This makes AUC a suitable metric for evaluating whether alignment preserves the discriminative structure of the source classifier, independent of any scale or offset mismatch introduced by the alignment operators.

7.5. Cross-Modal Meaning Consistency

Figure 6 shows per-symbol W1 cosine similarity across three model pairs and four extraction methods.

7.6. Robustness Check: Cohen's d Selectivity Index

Figure 7 replicates the SI analysis using |Cohen's d| instead of W1, confirming that the main findings are robust to the choice of distributional distance.

7.7. Target Label Definitions

Table 2 defines the health condition groups used as targets throughout the paper.

Table 2. Health condition group definitions. A subject is labeled positive if *any* constituent condition is self-reported as present; negative if *all* are reported absent; otherwise missing.

Group	**Positive if any of**
Cardiac	heart attack, heart disease, heart failure, beta-blocker use
Arrhythmia	atrial fibrillation, heart rhythm disorder, pacemaker
Vascular	stroke/TIA, artery disease, anticoagulant use, antiplatelet use, calcium-channel blocker use
CV risk	diabetes, hypertension
Mental health	depression, anxiety, antianxiety use, antidepressant use, antipsychotic use
Sleep	sleep apnea, sleep medication use
Hepatorenal	kidney disease, liver disease, urinary disorder, diuretic use
Musculoskeletal	arthritis, hip/knee disorder, osteoporosis, low-back disorder, neck disorder
Physiological markers	
Sex	binary (male/female) from enrollment
Age	binned: [18,40), [40,60), [60,100)
BMI	binned: $<$18.5, 18.5–25, 25–30, 30+
Negative control	
Blood type	ABO/Rh from Apple Health profile; one-hot encoded into 7 types

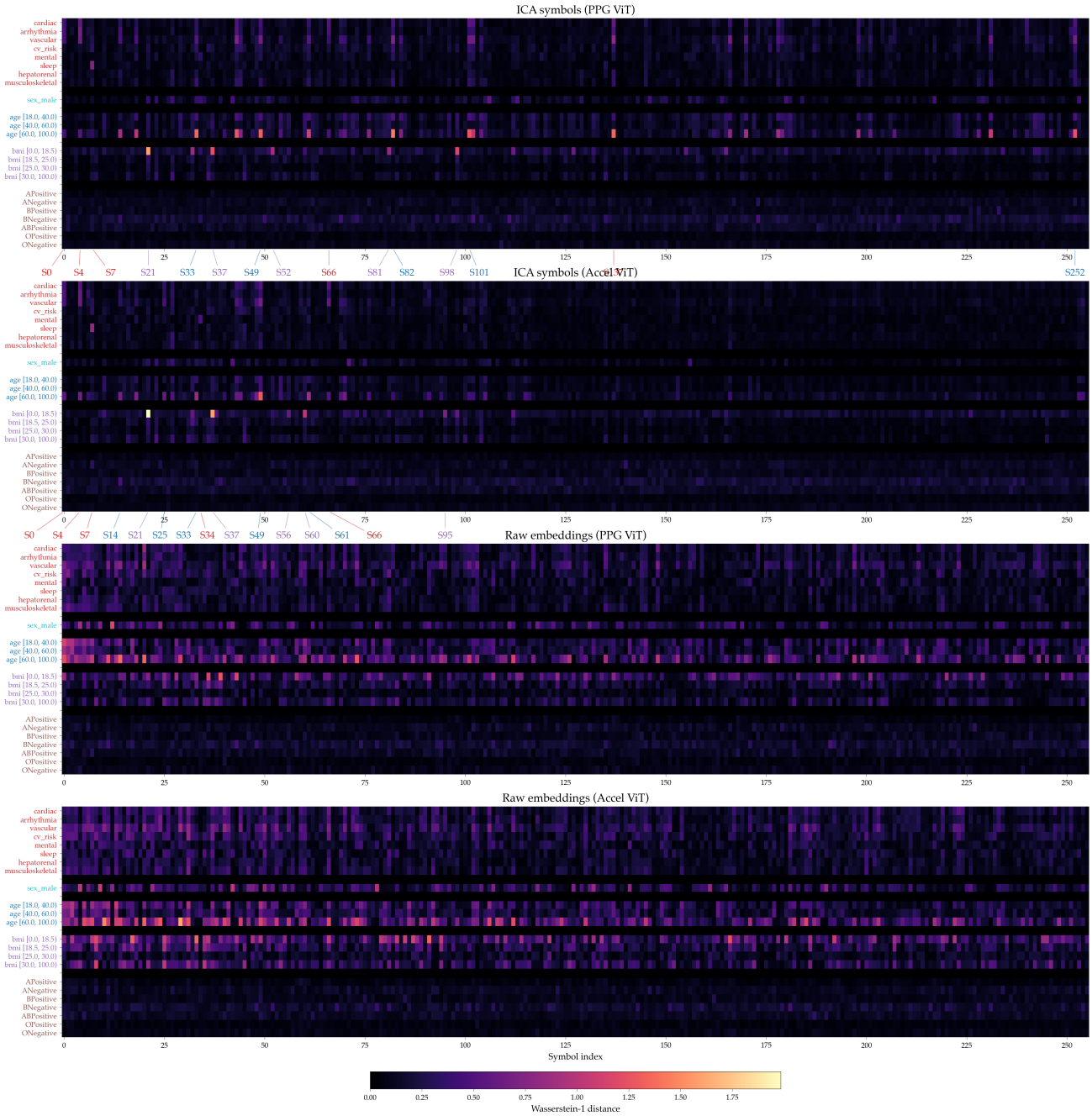

Figure 5. Full symbol-meaning heatmaps (companion to Figure 3a). All 256 symbols are shown for ICA (top two rows) and raw embeddings (bottom two rows), for PPG ViT and Accel ViT, including the blood-type negative control. Each column is a symbol, each row a target, and each cell the W1 distance between target-positive and target-negative activation distributions (shared global color scale). ICA concentrates target associations into a sparse set of symbols, whereas raw embedding dimensions are diffuse; blood type produces little signal in any representation. Colored markers indicate the five highest-SI symbols per target group.

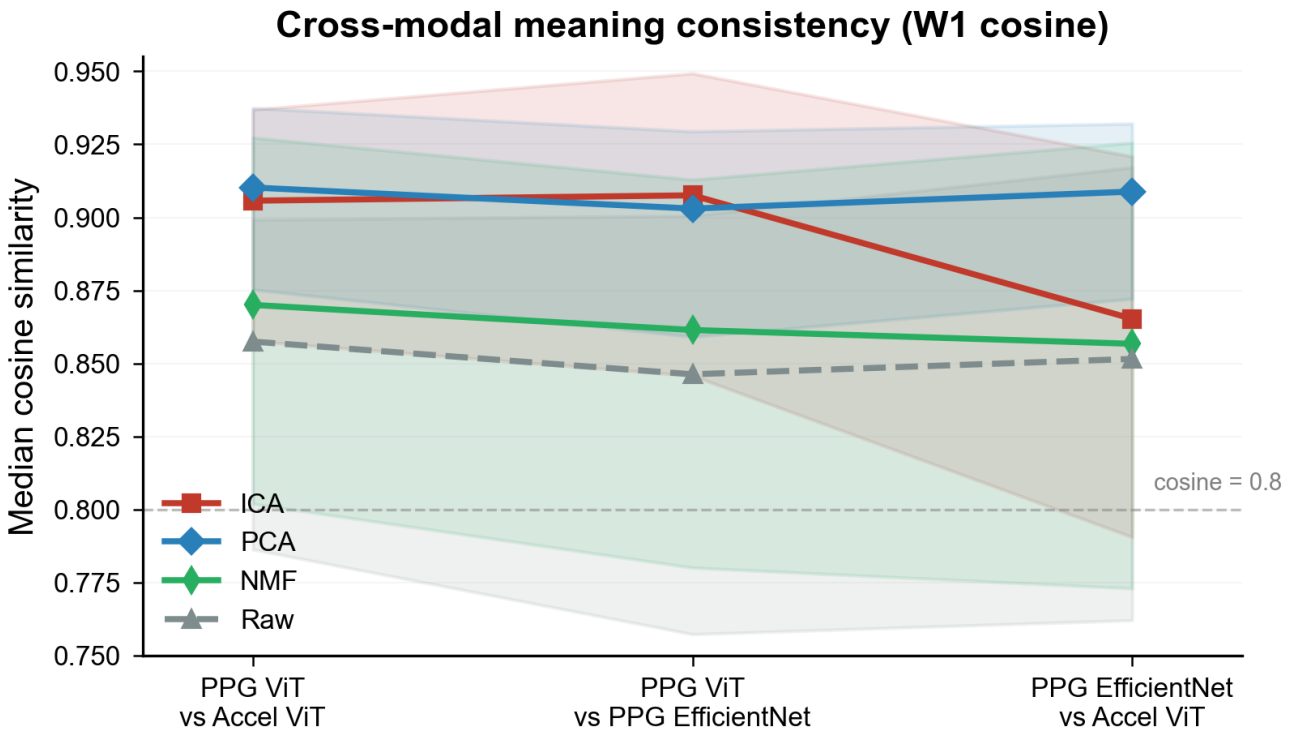

Figure 6. Cross-modal meaning consistency across model pairs and symbol extraction methods. Lines show median per-symbol W1 cosine similarity; shaded bands show interquartile range (Q25–Q75). PCA maintains 0.90–0.91 across all pairs (architecture-invariant), while ICA drops from 0.91 to 0.87 when both modality and architecture differ—higher-order statistics exploited by ICA are sensitive to architectural nonlinearities. Even raw embeddings achieve median 0.85, confirming that cross-modal physiological structure is genuine; symbol extraction adds disentanglement, sharpening the alignment.

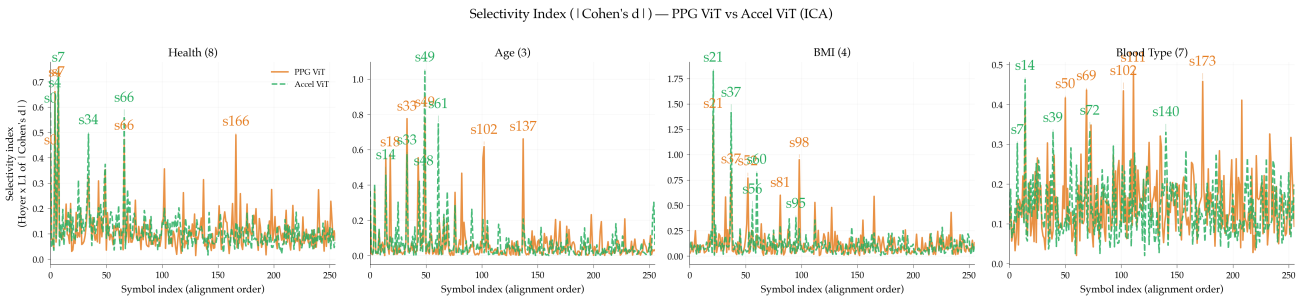

Figure 7. Selectivity Index computed on |Cohen's d| effect size (PPG ViT vs Accel ViT, ICA+bijective). Cohen's d has $\mathbb{E}[d] = 0$ under the null regardless of group size, eliminating the minority-group inflation floor present in W1. Age 60+ and BMI underweight remain dominant, confirming genuine physiological effect sizes.

7.8. Symbol Extraction Methods

Table 3 summarizes the symbol extraction methods used throughout the paper.

Table 3. Unified view of the symbol extraction methods used to re-express frozen embeddings in alternative coordinate systems. Each method is an extraction operator $\mathcal{E}$ mapping embeddings X to symbol activations Z, with a corresponding reconstruction back to the embedding space. For PCA and ICA the activations are an explicit linear readout $Z = XA$ (A the readout/encoder matrix); NMF has no such readout and obtains Z by non-negative least squares against the dictionary S. The dictionary S (columns are the candidate symbol directions) reconstructs embeddings via $\hat{X} = ZS^\top$; for orthonormal PCA the readout and dictionary coincide ($A = S$). In nonlinear cases, reconstruction is a learned inverse without assuming a linear symbol basis. We use the activations Z directly; the reconstruction column is included only to characterize each method.

Method	**Primary Assumption / Inductive Bias**	**Extraction**	**Reconstruction**
PCA	Dominant structure captured by maximal variance; orthogonality	$Z = XA$	$\hat{X} = ZS^\top \approx X$
ICA	Latent factors are statistically independent	$Z = XA$	$\hat{X} = ZS^\top \approx X$
NMF	Representations are additive and non-negative	$Z = \arg\min_{Z\geq 0}\lVert X - ZS^\top\rVert^2,\ S \geq 0$	$\hat{X} = ZS^\top$
CCA	Shared structure exists between paired embedding spaces	$Z_1 = X_1A_1,\ Z_2 = X_2A_2$	$\hat{X}_1 \approx Z_1S_1^\top,\ \hat{X}_2 \approx Z_2S_2^\top$
Sparse Autoencoder	Meaningful structure expressed via sparse, reusable activations	$Z = \phi_\theta(X)$	$\hat{X} = \psi_\theta(Z) \approx X$

The methods differ in how the coefficient matrix Z_m is obtained. For PCA and ICA, once the decomposition is fitted, activations are computed by an explicit linear readout,

$$Z_m = H_m A_m, \tag{9}$$

where $A_m \in \mathbb{R}^{d\times k}$ is the encoder/readout matrix. For orthonormal PCA, $A_m = S_m$. For ICA, the readout matrix and reconstruction dictionary are generally distinct, corresponding to the unmixing and mixing matrices. For NMF, activations are constrained to be non-negative: $H_m \approx Z_m S_m^\top$ with $Z_m, S_m \geq 0$. They are therefore estimated by non-negative least-squares fitting against the learned dictionary rather than by an unconstrained linear readout.

The framework also supports *nonlinear* extractors. A sparse autoencoder, for example, computes symbol activations through a nonlinear function,

$$z = \phi_\theta(h), \qquad \hat{h} = \psi_\theta(z). \tag{10}$$

Because ϕ_θ and ψ_θ may include nonlinearities and learned encoder–decoder structure, the resulting symbols are not restricted to fixed linear directions in embedding space. Such methods can in principle capture curved or conditional structure that a single linear direction would miss, and can also support overcomplete symbol dictionaries. This is why nonlinear extractors are a natural extension of the present framework, while the current paper focuses on the linearly recoverable structure exposed by PCA, ICA, NMF, and CCA-based alignment.

7.9. Transfer AUC by Extraction and Alignment Method

Tables 4–6 report transfer AUC and retention for each symbol extraction method (ICA, PCA, NMF) across the full range of alignment strategies, from bijective matching through sparse CCA at six regularization strengths to dense CCA. Embed+Bijective is included as a baseline that uses raw embeddings without symbol extraction. Several patterns emerge. First, alignment is necessary: the Embed+Bijective baseline already recovers substantial transfer, but CCA-based alignment consistently improves over bijective matching for all three extractors. Second, the choice of extraction method interacts with the alignment method: PCA with sparse CCA ($\lambda \leq 10^{-3}$) achieves the highest retention across most targets, while ICA

paired with dense CCA is competitive but less robust across domain pairs. NMF generally underperforms ICA and PCA, likely because its non-negativity constraint is a poor match for the centered embedding space. Third, blood type targets (negative control) remain near chance regardless of SE or SA, confirming that alignment does not create spurious transfer.

Table 4. Cross-modal transfer AUC and retention (mean across six domain pairs) for ICA symbol extraction across alignment methods. Each cell shows transfer AUC with retention in parentheses. Sparse CCA columns show decreasing λ.

		Embed	ICA							
Target	In-d.	Bij.	Bij.	8×10^{-2}	4×10^{-2}	2×10^{-2}	10^{-2}	10^{-3}	10^{-4}	CCA
Cardiac	0.80	0.67 (84%)	0.68 (86%)	0.57 (74%)	0.62 (79%)	0.68 (85%)	0.64 (81%)	0.68 (86%)	0.69 (87%)	0.69 (87%)
Arrhythmia	0.75	0.64 (86%)	0.67 (92%)	0.60 (82%)	0.64 (87%)	0.64 (88%)	0.65 (88%)	0.65 (89%)	0.65 (88%)	0.65 (89%)
Vascular	0.83	0.76 (92%)	0.74 (89%)	0.59 (72%)	0.67 (81%)	0.68 (82%)	0.71 (86%)	0.71 (85%)	0.73 (87%)	0.73 (87%)
CV Risk	0.83	0.69 (83%)	0.70 (84%)	0.56 (69%)	0.62 (74%)	0.68 (82%)	0.72 (86%)	0.72 (86%)	0.72 (86%)	0.72 (87%)
Mental	0.81	0.63 (78%)	0.69 (85%)	0.57 (72%)	0.60 (74%)	0.65 (81%)	0.68 (85%)	0.70 (87%)	0.71 (89%)	0.71 (88%)
Sleep	0.76	0.63 (83%)	0.68 (90%)	0.60 (80%)	0.61 (81%)	0.65 (86%)	0.63 (84%)	0.67 (88%)	0.66 (88%)	0.66 (88%)
Hepatorenal	0.72	0.62 (87%)	0.63 (89%)	0.54 (77%)	0.60 (85%)	0.63 (88%)	0.63 (89%)	0.61 (86%)	0.64 (90%)	0.64 (90%)
Musculoskeletal	0.76	0.68 (89%)	0.67 (88%)	0.53 (71%)	0.61 (81%)	0.64 (85%)	0.66 (87%)	0.67 (89%)	0.67 (89%)	0.67 (89%)
Sex	1.00	0.81 (81%)	0.83 (84%)	0.66 (66%)	0.75 (75%)	0.80 (80%)	0.82 (82%)	0.82 (82%)	0.82 (82%)	0.81 (82%)
Age 18–39	0.96	0.86 (90%)	0.77 (81%)	0.56 (59%)	0.64 (67%)	0.74 (77%)	0.77 (81%)	0.81 (85%)	0.82 (86%)	0.81 (85%)
Age 40–59	0.94	0.78 (83%)	0.74 (78%)	0.58 (63%)	0.60 (64%)	0.75 (79%)	0.75 (80%)	0.80 (85%)	0.80 (85%)	0.80 (85%)
Age 60+	0.99	0.89 (90%)	0.88 (89%)	0.70 (72%)	0.75 (76%)	0.80 (81%)	0.82 (84%)	0.85 (87%)	0.85 (86%)	0.85 (86%)
BMI $<$18.5	0.95	0.85 (90%)	0.82 (87%)	0.78 (83%)	0.76 (81%)	0.80 (85%)	0.79 (84%)	0.80 (85%)	0.81 (86%)	0.81 (86%)
BMI 18.5–25	0.91	0.81 (89%)	0.72 (79%)	0.59 (66%)	0.63 (70%)	0.71 (79%)	0.76 (84%)	0.77 (84%)	0.77 (85%)	0.78 (86%)
BMI 25–30	0.78	0.57 (73%)	0.65 (84%)	0.54 (71%)	0.55 (71%)	0.61 (79%)	0.65 (83%)	0.67 (86%)	0.68 (88%)	0.68 (88%)
BMI 30+	0.93	0.84 (90%)	0.76 (81%)	0.60 (65%)	0.66 (70%)	0.76 (81%)	0.78 (84%)	0.78 (84%)	0.79 (84%)	0.79 (85%)
A+	0.56	0.53 (94%)	0.52 (94%)	0.52 (93%)	0.52 (93%)	0.53 (94%)	0.52 (93%)	0.52 (94%)	0.53 (96%)	0.53 (96%)
A−	0.59	0.54 (95%)	0.54 (97%)	0.53 (96%)	0.54 (97%)	0.54 (96%)	0.53 (96%)	0.53 (96%)	0.53 (96%)	0.53 (95%)
B+	0.58	0.55 (95%)	0.53 (93%)	0.51 (90%)	0.53 (94%)	0.54 (94%)	0.53 (94%)	0.55 (96%)	0.55 (96%)	0.55 (96%)
B−	0.51	0.51 (107%)	0.50 (105%)	0.49 (102%)	0.48 (101%)	0.50 (107%)	0.49 (105%)	0.48 (102%)	0.47 (99%)	0.47 (100%)
AB+	0.54	0.49 (94%)	0.51 (97%)	0.53 (101%)	0.52 (100%)	0.52 (99%)	0.53 (101%)	0.53 (101%)	0.52 (98%)	0.52 (98%)
O+	0.58	0.52 (92%)	0.54 (96%)	0.53 (94%)	0.52 (93%)	0.54 (95%)	0.53 (94%)	0.54 (96%)	0.54 (96%)	0.54 (96%)
O−	0.54	0.50 (94%)	0.51 (96%)	0.50 (95%)	0.52 (98%)	0.52 (98%)	0.52 (97%)	0.52 (98%)	0.52 (97%)	0.52 (97%)

Table 5. Cross-modal transfer AUC and retention (mean across six domain pairs) for PCA symbol extraction across alignment methods. Each cell shows transfer AUC with retention in parentheses. Sparse CCA columns show decreasing λ.

		Embed	PCA							
Target	In-d.	Bij.	Bij.	8×10^{-2}	4×10^{-2}	2×10^{-2}	10^{-2}	10^{-3}	10^{-4}	CCA
Cardiac	0.80	0.67 (84%)	0.55 (69%)	0.67 (85%)	0.76 (96%)	0.74 (94%)	0.76 (96%)	0.75 (95%)	0.78 (98%)	0.68 (86%)
Arrhythmia	0.75	0.64 (86%)	0.58 (79%)	0.69 (93%)	0.70 (94%)	0.71 (96%)	0.72 (97%)	0.72 (98%)	0.73 (99%)	0.64 (87%)
Vascular	0.83	0.76 (92%)	0.65 (78%)	0.80 (96%)	0.80 (96%)	0.79 (95%)	0.79 (95%)	0.79 (95%)	0.78 (94%)	0.71 (85%)
CV Risk	0.83	0.69 (83%)	0.63 (76%)	0.71 (86%)	0.74 (90%)	0.78 (93%)	0.79 (95%)	0.82 (98%)	0.82 (99%)	0.71 (85%)
Mental	0.81	0.63 (78%)	0.58 (73%)	0.72 (89%)	0.69 (86%)	0.75 (94%)	0.76 (94%)	0.79 (98%)	0.80 (99%)	0.70 (87%)
Sleep	0.76	0.63 (83%)	0.54 (72%)	0.67 (88%)	0.68 (90%)	0.69 (92%)	0.70 (93%)	0.74 (98%)	0.75 (99%)	0.67 (89%)
Hepatorenal	0.72	0.62 (87%)	0.53 (74%)	0.64 (89%)	0.66 (93%)	0.68 (95%)	0.69 (97%)	0.69 (97%)	0.71 (99%)	0.63 (88%)
Musculoskeletal	0.76	0.68 (89%)	0.60 (79%)	0.66 (87%)	0.69 (91%)	0.69 (91%)	0.73 (96%)	0.74 (98%)	0.75 (99%)	0.67 (88%)
Sex	1.00	0.81 (81%)	0.57 (57%)	0.96 (97%)	0.98 (99%)	0.98 (99%)	0.99 (99%)	1.00 (100%)	1.00 (100%)	0.83 (84%)
Age 18–39	0.96	0.86 (90%)	0.60 (63%)	0.87 (92%)	0.87 (91%)	0.89 (94%)	0.94 (98%)	0.95 (99%)	0.95 (99%)	0.80 (84%)
Age 40–59	0.94	0.78 (83%)	0.57 (60%)	0.82 (87%)	0.79 (84%)	0.87 (93%)	0.89 (94%)	0.90 (96%)	0.93 (99%)	0.79 (84%)
Age 60+	0.99	0.89 (90%)	0.61 (62%)	0.95 (96%)	0.92 (93%)	0.96 (97%)	0.97 (99%)	0.98 (99%)	0.98 (100%)	0.82 (83%)
BMI <18.5	0.95	0.85 (90%)	0.57 (61%)	0.91 (97%)	0.88 (94%)	0.93 (98%)	0.93 (99%)	0.93 (99%)	0.93 (99%)	0.80 (85%)
BMI 18.5–25	0.91	0.81 (89%)	0.55 (60%)	0.84 (93%)	0.78 (86%)	0.85 (93%)	0.86 (95%)	0.89 (98%)	0.90 (99%)	0.77 (85%)
BMI 25–30	0.78	0.57 (73%)	0.52 (67%)	0.61 (79%)	0.63 (80%)	0.68 (88%)	0.69 (89%)	0.75 (96%)	0.76 (98%)	0.68 (88%)
BMI 30+	0.93	0.84 (90%)	0.56 (60%)	0.85 (91%)	0.88 (94%)	0.90 (96%)	0.90 (97%)	0.92 (99%)	0.92 (99%)	0.79 (84%)
A+	0.56	0.53 (94%)	0.51 (92%)	0.54 (98%)	0.52 (94%)	0.52 (94%)	0.53 (95%)	0.54 (96%)	0.55 (98%)	0.53 (95%)
A−	0.59	0.54 (95%)	0.50 (91%)	0.55 (96%)	0.53 (94%)	0.53 (95%)	0.54 (98%)	0.53 (96%)	0.54 (97%)	0.51 (92%)
B+	0.58	0.55 (95%)	0.50 (87%)	0.55 (96%)	0.55 (96%)	0.57 (100%)	0.56 (97%)	0.58 (101%)	0.58 (101%)	0.55 (97%)
B−	0.51	0.51 (107%)	0.47 (99%)	0.50 (104%)	0.52 (110%)	0.48 (101%)	0.47 (102%)	0.46 (99%)	0.46 (98%)	0.48 (101%)
AB+	0.54	0.49 (94%)	0.50 (96%)	0.51 (98%)	0.52 (99%)	0.53 (100%)	0.53 (102%)	0.53 (101%)	0.53 (101%)	0.53 (101%)
O+	0.58	0.52 (92%)	0.50 (89%)	0.54 (94%)	0.54 (96%)	0.54 (95%)	0.54 (96%)	0.55 (98%)	0.56 (99%)	0.54 (96%)
O−	0.54	0.50 (94%)	0.52 (98%)	0.51 (95%)	0.51 (95%)	0.51 (97%)	0.51 (97%)	0.53 (99%)	0.53 (100%)	0.52 (98%)

Table 6. Cross-modal transfer AUC and retention (mean across six domain pairs) for NMF symbol extraction across alignment methods. Each cell shows transfer AUC with retention in parentheses. Sparse CCA columns show decreasing λ.

		Embed	NMF							
Target	In-d.	Bij.	Bij.	8×10^{-2}	4×10^{-2}	2×10^{-2}	10^{-2}	10^{-3}	10^{-4}	CCA
Cardiac	0.80	0.67 (84%)	0.50 (63%)	0.54 (69%)	0.63 (80%)	0.64 (81%)	0.65 (82%)	0.68 (86%)	0.71 (89%)	0.59 (75%)
Arrhythmia	0.75	0.64 (86%)	0.53 (71%)	0.56 (77%)	0.62 (84%)	0.63 (86%)	0.63 (86%)	0.64 (86%)	0.65 (89%)	0.59 (80%)
Vascular	0.83	0.76 (92%)	0.56 (68%)	0.62 (75%)	0.68 (82%)	0.70 (84%)	0.69 (83%)	0.70 (84%)	0.74 (89%)	0.63 (76%)
CV Risk	0.83	0.69 (83%)	0.50 (60%)	0.56 (67%)	0.65 (78%)	0.70 (84%)	0.69 (83%)	0.68 (82%)	0.67 (80%)	0.57 (69%)
Mental	0.81	0.63 (78%)	0.50 (62%)	0.55 (69%)	0.59 (74%)	0.61 (77%)	0.65 (81%)	0.63 (79%)	0.68 (85%)	0.60 (75%)
Sleep	0.76	0.63 (83%)	0.50 (66%)	0.52 (70%)	0.62 (82%)	0.66 (87%)	0.66 (88%)	0.65 (86%)	0.64 (85%)	0.57 (76%)
Hepatorenal	0.72	0.62 (87%)	0.53 (74%)	0.56 (79%)	0.60 (84%)	0.61 (86%)	0.62 (87%)	0.62 (87%)	0.61 (86%)	0.58 (81%)
Musculoskeletal	0.76	0.68 (89%)	0.50 (66%)	0.53 (71%)	0.63 (83%)	0.65 (85%)	0.64 (84%)	0.61 (81%)	0.61 (81%)	0.56 (74%)
Sex	1.00	0.81 (81%)	0.50 (50%)	0.61 (62%)	0.77 (77%)	0.82 (83%)	0.82 (83%)	0.83 (83%)	0.82 (83%)	0.66 (67%)
Age 18–39	0.96	0.86 (90%)	0.52 (55%)	0.57 (60%)	0.72 (75%)	0.75 (79%)	0.78 (82%)	0.79 (83%)	0.79 (83%)	0.65 (68%)
Age 40–59	0.94	0.78 (83%)	0.50 (53%)	0.59 (63%)	0.67 (71%)	0.72 (77%)	0.74 (79%)	0.73 (78%)	0.74 (78%)	0.63 (67%)
Age 60+	0.99	0.89 (90%)	0.50 (51%)	0.69 (70%)	0.80 (81%)	0.79 (80%)	0.81 (82%)	0.82 (83%)	0.82 (83%)	0.62 (63%)
BMI <18.5	0.95	0.85 (90%)	0.50 (53%)	0.73 (78%)	0.78 (83%)	0.78 (83%)	0.75 (79%)	0.78 (83%)	0.78 (83%)	0.66 (70%)
BMI 18.5–25	0.91	0.81 (89%)	0.54 (60%)	0.56 (62%)	0.66 (73%)	0.67 (74%)	0.74 (81%)	0.70 (77%)	0.70 (78%)	0.63 (70%)
BMI 25–30	0.78	0.57 (73%)	0.50 (64%)	0.52 (68%)	0.56 (72%)	0.58 (74%)	0.58 (75%)	0.59 (77%)	0.60 (77%)	0.58 (75%)
BMI 30+	0.93	0.84 (90%)	0.53 (56%)	0.58 (63%)	0.72 (77%)	0.76 (82%)	0.77 (83%)	0.77 (83%)	0.77 (83%)	0.64 (69%)
A+	0.56	0.53 (94%)	0.50 (90%)	0.50 (91%)	0.51 (92%)	0.52 (93%)	0.52 (93%)	0.52 (94%)	0.52 (93%)	0.52 (92%)
A−	0.59	0.54 (95%)	0.50 (86%)	0.53 (93%)	0.53 (93%)	0.54 (94%)	0.53 (93%)	0.53 (94%)	0.52 (91%)	0.51 (91%)
B+	0.58	0.55 (95%)	0.51 (89%)	0.51 (90%)	0.53 (93%)	0.54 (93%)	0.54 (94%)	0.55 (96%)	0.55 (95%)	0.52 (91%)
B−	0.51	0.51 (107%)	0.50 (100%)	0.48 (101%)	0.51 (106%)	0.49 (101%)	0.49 (103%)	0.48 (100%)	0.49 (103%)	0.50 (104%)
AB+	0.54	0.49 (94%)	0.50 (98%)	0.51 (99%)	0.52 (100%)	0.51 (96%)	0.52 (100%)	0.52 (99%)	0.52 (99%)	0.53 (101%)
O+	0.58	0.52 (92%)	0.51 (89%)	0.51 (91%)	0.53 (93%)	0.53 (94%)	0.53 (94%)	0.53 (94%)	0.53 (95%)	0.52 (92%)
O−	0.54	0.50 (94%)	0.50 (93%)	0.52 (96%)	0.51 (96%)	0.51 (95%)	0.52 (97%)	0.51 (96%)	0.51 (96%)	0.50 (94%)