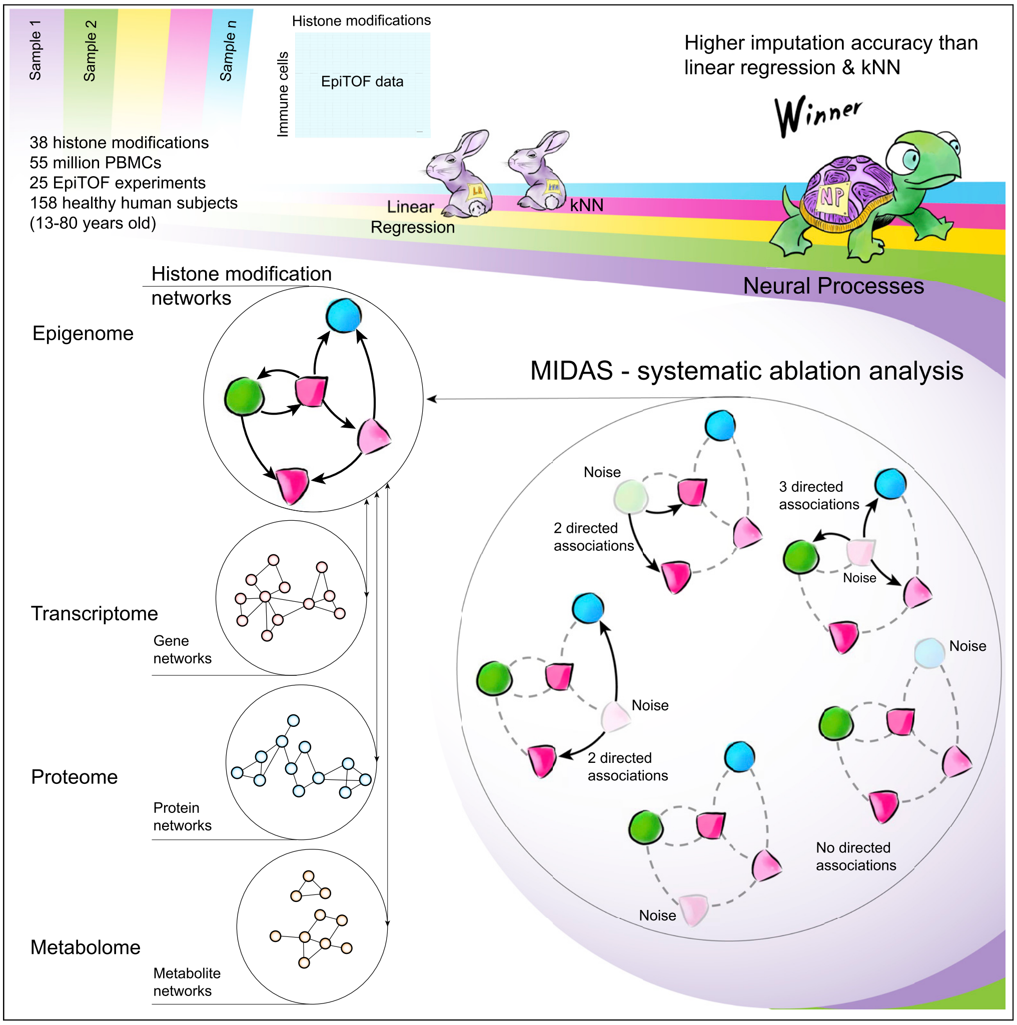
iScience

CellPress
OPEN ACCESS

Article

Inferring direction of associations between histone modifications using a neural processes-based framework



Ananthakrishnan Ganesan, Denis Dermadi, Laurynas Kalesinskas, Michele Donato, Rosalie Sowers, Paul J. Utz, Purvesh Khatri

pkhatri@stanford.edu

Highlights

Predicting the direction of association between HPTMs using current techniques is difficult

Neural Processes allow joint modeling of subject- and cell-level variability

Neural Processes have higher imputation accuracy than several machine learning methods

Ablation-based analysis enables inference of directed associations between HPTM pairs

Ganesan et al., iScience *26*, 105756
January 20, 2023 © 2022 The Authors.
https://doi.org/10.1016/j.isci.2022.105756

iScience

CellPress OPEN ACCESS

Article

Inferring direction of associations between histone modifications using a neural processes-based framework

Ananthakrishnan Ganesan,[1,2,3,6] Denis Dermadi,[1,3,6] Laurynas Kalesinskas,[1,3,6] Michele Donato,[1,3,6] Rosalie Sowers,[1,3] Paul J. Utz,[1,4] and Purvesh Khatri[1,3,5,7,*]

SUMMARY

Current technologies do not allow predicting interactions between histone post-translational modifications (HPTMs) at a system-level. We describe a computational framework, imputation-followed-by-inference, to predict directed association between two HPTMs using EpiTOF, a mass cytometry-based platform that allows profiling multiple HPTMs at a single-cell resolution. Using EpiTOF profiles of >55,000,000 peripheral mononuclear blood cells from 158 healthy human subjects, we show that neural processes (NP) have significantly higher accuracy than linear regression and *k*-nearest neighbors models to impute the abundance of an HPTM. Next, we infer the direction of association to show we recapitulate known HPTM associations and identify several previously unidentified ones in healthy individuals. Using this framework in an influenza vaccine cohort, we identify changes in associations between 6 pairs of HPTMs 30 days following vaccination, of which several have been shown to be involved in innate memory. These results demonstrate the utility of our framework in identifying directed interactions between HPTMs.

INTRODUCTION

Histone post-translational modifications (HPTMs) play a vital role in the regulation of gene expression, cell differentiation, and different processes centered around DNA.[1] Dysregulation of HPTMs has been implicated in human diseases such as cancer,[2,3] infectious diseases,[4] mental illnesses,[5] and autoimmune disorders.[6,7] HPTMs are also known to play an important role in the immune response following vaccination.[8,9] Interaction of multiple HPTMs has been recognized as the language of histone crosstalk.[10,11] Despite the advances in understanding the chromatin organization and transcriptional regulation through HPTMs, histone crosstalk has been understudied due to technological limitations of ChIP-seq[12] to investigate more than 2-3 HPTMs at once.

Recently, we described a mass cytometry-based technology, EpiTOF, for high-throughput profiling of HPTMs at a single-cell resolution.[13] EpiTOF can profile up to 38 HPTMs and histone variants across 2 non-overlapping panels. However, both panels include 11 cell phenotypic marks (CPMs) for cell type identification. Prior studies using EpiTOF profiling of immune cells have quantified increased epigenetic noise with aging in the human immune system,[13] identified a novel epigenetic regulation of monocyte-to-macrophage differentiation,[14] and epigenetic mechanism causally linked with innate memory after vaccination.[8] Further, EpiTOF data have also elucidated comprehensive correlation networks of HPTMs that consist of modules either conserved across immune cell types or are specific to a cell lineage.[15]

Despite its demonstrated advantage in identifying novel epigenetic mechanisms, EpiTOF is limited in the number of HPTMs measured in a panel due to the chemical and physical properties of heavy metal isotopes used in mass cytometry. However, the large number of cells per sample profiled using EpiTOF and recent advances in machine learning provide an opportunity to use *in silico* models to impute abundances for a subset of HPTMs, which in turn could be replaced by other HPTMs to increase the number of HPTMs profiled per cell. On the other hand, an advantage of EpiTOF is that the large number of cells profiled using it provides an unprecedented opportunity to create interpretable imputation models to infer the underlying associations between HPTMs, further elucidating the histone language.[10] Several machine learning (ML)

[1] Institute for Immunity, Transplantation and Infection, School of Medicine, Stanford University, Stanford, CA 94305, USA

[2] Institute for Computational and Mathematical Engineering, School of Engineering, Stanford University, Stanford, CA 94305, USA

[3] Center for Biomedical Informatics Research, Department of Medicine, School of Medicine, Stanford University, Stanford, CA 94305, USA

[4] Division of Immunology and Rheumatology, Department of Medicine, School of Medicine, Stanford University, Stanford, CA 94305, USA

[5] Present address: Inflammatix, Inc., Sunnyvale, CA 94085, USA

[6] These authors contributed equally

[7] Lead contact

*Correspondence: pkhatri@stanford.edu

https://doi.org/10.1016/j.isci.2022.105756

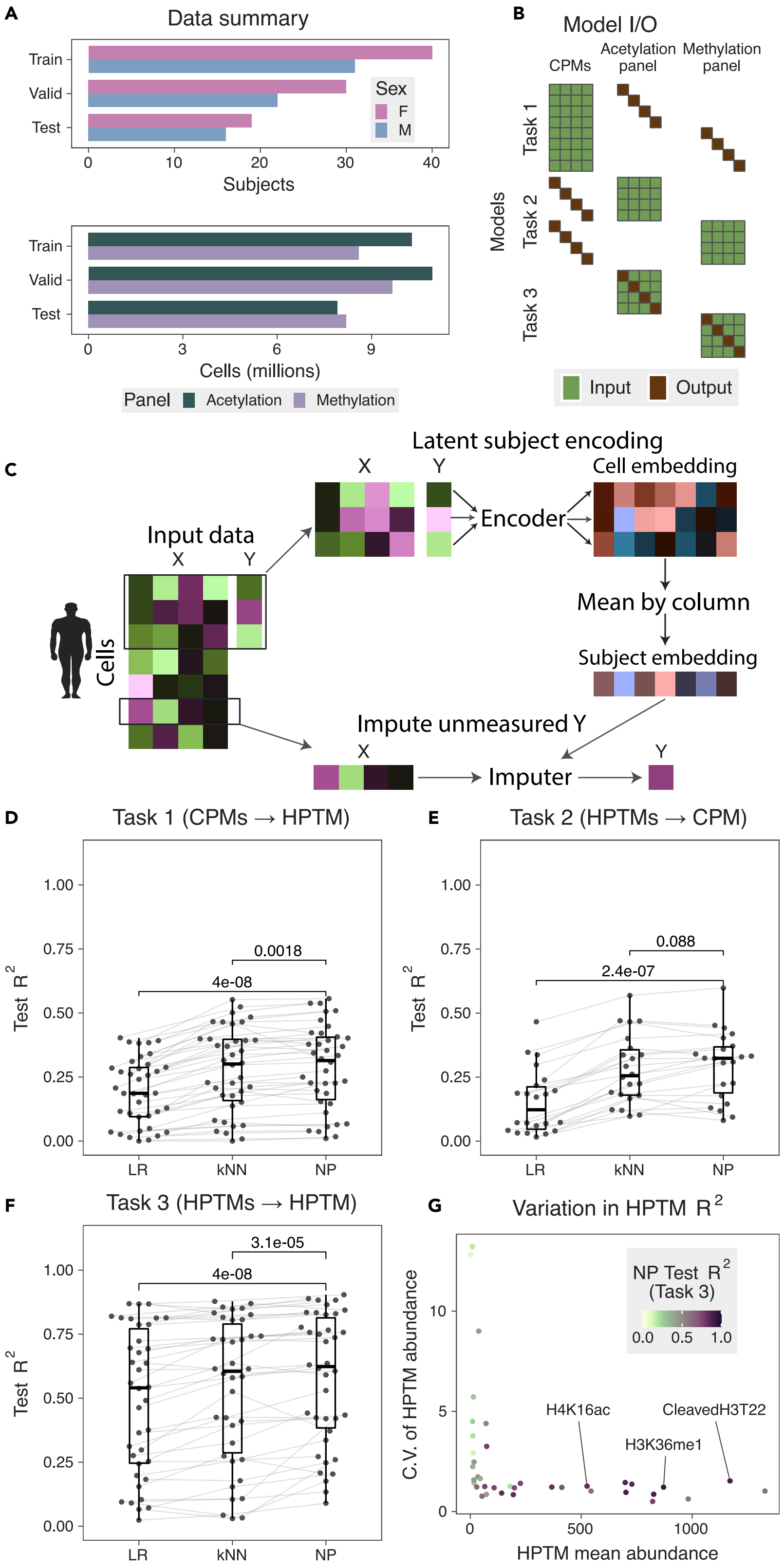

A
Data summary
Train
Valid
Test
Sex
F
M
0
10
20
30
40
Subjects
Train
Valid
Test
0
3
6
9
Cells (millions)
Panel
Acetylation
Methylation
B
Model I/O
CPMs
Acetylation panel
Methylation panel
Models
Task 1
Task 2
Task 3
Input
Output
C
Latent subject encoding
X
Y
Encoder
Cell embedding
Input data
X
Y
Cells
Mean by column
Subject embedding
Impute unmeasured Y
X
Y
Imputer
D
Task 1 (CPMs → HPTM)
Test R^2
1.00
0.75
0.50
0.25
0.00
0.0018
4e-08
LR
kNN
NP
E
Task 2 (HPTMs → CPM)
Test R^2
1.00
0.75
0.50
0.25
0.00
0.088
2.4e-07
LR
kNN
NP
F
Task 3 (HPTMs → HPTM)
Test R^2
1.00
0.75
0.50
0.25
0.00
3.1e-05
4e-08
LR
kNN
NP
G
Variation in HPTM R^2
C.V. of HPTM abundance
10
5
0
NP Test R^2 (Task 3)
0.0
0.5
1.0
H4K16ac
CleavedH3T22
H3K36me1
0
500
1000
HPTM mean abundance

Figure 1. Summary of EpiTOF data and results from imputation models

(A) Number of subjects (top) and PBMCs (bottom) used in train, validation, and test sets for acetylation and methylation panels.

(B) Illustration of the three imputation tasks with corresponding inputs and outputs. Each row represents a unique model, and each column represents a variable.

(C) Flow chart for NP models. X represents the input columns and Y represents the column to be imputed. *Encoder* and *Imputer* are fully connected neural networks.

(D–F) Summary of LR, kNN, and NP models for (D) *task 1*, (E) *task 2*, and (F) *task 3*. Each dot represents a model. Lines connect models predicting the same variable (HPTM or CPM) across algorithms. p-values were computed using one-sided, paired Wilcoxon test was used for the improvement in R^2 due to the NP model (n = 38 HPTMs or 22 CPMs per algorithm).

(G) Comparison of R^2 from *task 3* NP models with mean and coefficient of the variation of HPTM abundances. Each dot represents a unique HPTM. Color of each dot corresponds to the imputation R^2.

See also Figures S1 and S2; Tables S1–S3.

approaches have been used to impute multi-panel cytometry data. For instance, CyTOFmerge uses a *k*-nearest neighbors (*k*NN)-based approach to create an imputation function for each subject in the dataset.[16] Such subject-specific models are advantageous since individuals are exposed to different environments that are known to have different effects on epigenetics and HPTMs.[13,17] However, it is difficult to interpret the importance of input variables in *k*NN because it is non-parametric. Other approaches use parametric non-linear models such as support vector machines, boosting trees, and canonical neural networks.[18] Although these parametric models can be interpreted to quantify the importance of each input variable, they use the same imputation function for all subjects and do not distinguish between them, which may not capture the effects of difference in environmental exposure between individuals.

Neural process (NP), a recently described neural network-based model, combines the advantages of both *k*NN and parametric non-linear models.[19,20] Similar to *k*NN-based models, NP-based models use subject-specific functions for imputation. Because NP-based models are parametric, the importance of each input variable can be interpreted. Importantly, NP models are scalable to complex functions and large datasets.[19] Because the biological processes regulating the interactions between HPTMs are complex, we hypothesized that NPs would more accurately impute HPTMs and CPMs, and reveal biologically meaningful relationships between them.

Here, we adapted the neural network architecture from NP to impute multi-panel mass cytometry-based HPTM abundances using subject-specific parametric models. Using more than 55 million human peripheral mononuclear blood cells (PBMCs) from 158 healthy subjects across 25 independent experiments, we evaluated linear regression (LR)-, kNN- and NP-based methods for accuracy of imputation. We also developed an interpretation framework to infer the direction of association between two HPTMs through systematic perturbations. We reproduced several known HPTM associations and identified several previously unidentified associations between pairs of HPTMs. Most of the associations were conserved across all healthy individuals. By combining the inferred pair-wise associations, we created a system-wide directed network of HPTM associations from the NP models. Finally, we demonstrated the utility of the NP models and the interpretation framework by identifying HPTMs whose associations were modified in response to the trivalent inactivated seasonal influenza vaccine (TIV) in healthy adults.

RESULTS

Data and model types

We profiled 55.6 million PBMCs from 158 healthy subjects (13-80 years old) across 25 EpiTOF experiments and measured 38 HPTMs and histone variants, 2 core histones (H3 and H4), and 11 cell phenotypic markers (CPMs) across two panels referred to as the methylation panel and the acetylation panel (Figures 1A and S1; Table S1, STAR Methods). We have previously described extensive validation of antibodies for HPTMs using Western blot, flow cytometry, and different cell lines with *in vitro* manipulation.[13] We divided 158 subjects into three sets: training set (15 experiments, 71 subjects, 18.9 million cells), validation set (5 experiments, 52 subjects, 20.6 million cells), and test set (5 experiments, 35 subjects, 16.1 million cells) (Table S2, STAR Methods). These 25 experiments were performed over 4 years (2016-2019), collectively representing a broad range of technical heterogeneity due to different CyTOF machines, batches of reagents, and operators. We ensured that all subjects from each experiment were assigned to exactly one set for unbiased evaluation of imputation models.

CPMs are expressed on the cell surface and are used to determine the immune cell sub-type, whereas HPTMs are inside the cell nucleus. Given these biological differences between CPMs and HPTMs, we compared the accuracy of LR, *k*NN, and NP models for imputing a CPM or an HPTM using three imputation tasks: impute an HPTM using CPMs (*task 1*), impute a CPM using HPTMs (*task 2*), and impute an HPTM using other HPTMs on the same panel (*task 3*) (Figure 1B). We used the training and validation sets for creating imputation models. Next, we locked each model and evaluated its accuracy using the test set. We evaluated 98 combinations of inputs and outputs for 294 models across the three tasks for each of the ML methods (LR, *k*NN, and NP; Figure 1C and Table S3). We assessed the accuracy of a model using the coefficient of determination (R^2), where an R^2 of 1 indicates an accurate imputation whereas an R^2 of 0 indicates a failure to impute accurately. Importantly, we only report results from the test set experiments because the imputation models are expected to have high accuracy in training and validation set experiments.

Cell phenotypic mark and histone post-translational modification abundances are mostly independent of each other

When imputing an HPTM using CPMs in *task 1*, each of the three ML methods had low to moderate accuracy. Specifically, the mean R^2 for LR, *k*NN, and NP models were 0.19 (range: 0-0.4), 0.28 (range: 0-0.55), and 0.29 (range: 0.01-0.56), respectively (Figure 1C). Although NP had significantly higher mean R^2 compared to LR and *k*NN models ($p < 0.002$), the low to moderate imputation accuracy suggests that CPMs alone are not sufficient to impute HPTMs across panels accurately.

Similarly, when imputing a CPM using HPTMs in *task 2*, each of the three methods had low to moderate accuracy. Specifically, the mean R^2 for LR, *k*NN, and NP were 0.15 (range: 0.02-0.47), 0.28 (range: 0.10-0.57), and 0.30 (range: 0.08-0.60), respectively (Figure 1D). NP had a significantly higher mean of R^2 compared to LR and *k*NN models ($p < 0.09$). Taken together, our results indicate that the abundances of CPMs and HPTMs are largely independent of each other. These results are in line with single-cell RNA-seq datasets repeatedly showing that despite the expression of the same CPM, PBMCs are highly heterogeneous.[21,22]

Histone post-translational modifications with high abundance impute each other accurately

We recently observed that HPTMs form conserved modules across immune cell types.[15] Therefore, we hypothesized that an HPTM could be more accurately imputed using other HPTMs than using CPMs. Indeed, each of the three ML models demonstrated higher accuracy for imputing an HPTM using other HPTMs (*task 3*) instead of CPMs (*task 1*) (Figures 1E and S2). The mean R^2 for LR, *k*NN, and NP were 0.49 (range: 0.02-0.87), 0.54 (range: 0.03-0.88), and 0.58 (range: 0.09-0.90), respectively. NP had significantly higher R^2 compared to LR and *k*NN ($p < 4e-05$). Interestingly, HPTMs imputed with high R^2 were those with higher abundance and low variability across cells (Figure 1F). It is important to note that the NP models are unaware of the mean abundance or variability of these HPTMs because the abundances of each HPTM are independently scaled by the subject before being used in the models. Taken together, our results demonstrate HPTMs with high abundances are highly predictive of other HPTMs, and strongly associated with each other.

Interpretation of neural processes models reveals histone post-translational modification association networks

The higher accuracy of *k*NN and NP in imputing an HPTM from other HPTMs (*task 3*) compared to LR suggested non-linear associations between HPTMs, which are modeled better with *k*NN and NP than LR. To infer pairwise HPTM associations and their directionality, we developed an interpretation algorithm using noise perturbations (STAR Methods). Briefly, without re-training or modifying a *k*NN or NP model, we imputed an HPTM Y by systematically replacing the abundance of each input HPTM X_i, one at a time, with noise. We quantified the strength of the directed association from X_i to Y as the percent decrease in R^2, with 100% denoting the strongest association and 0% denoting the weakest association. The pairs of HPTMs with strong associations formed directed networks for each panel.

Interpretation of NP models identified several known directed interactions between HPTMs in both panels. In the acetylation panel, the strongest directional interaction was between histone H3 cleaved at threonine 22 (cleaved H3T22) and H3.3S31ph (Figures 2A and S3). Replacing cleaved H3T22 with noise led to a 90% reduction in R^2 of H3.3S31ph. Importantly, perturbing H3.3S31ph had a lower effect on cleaved H3T22, with

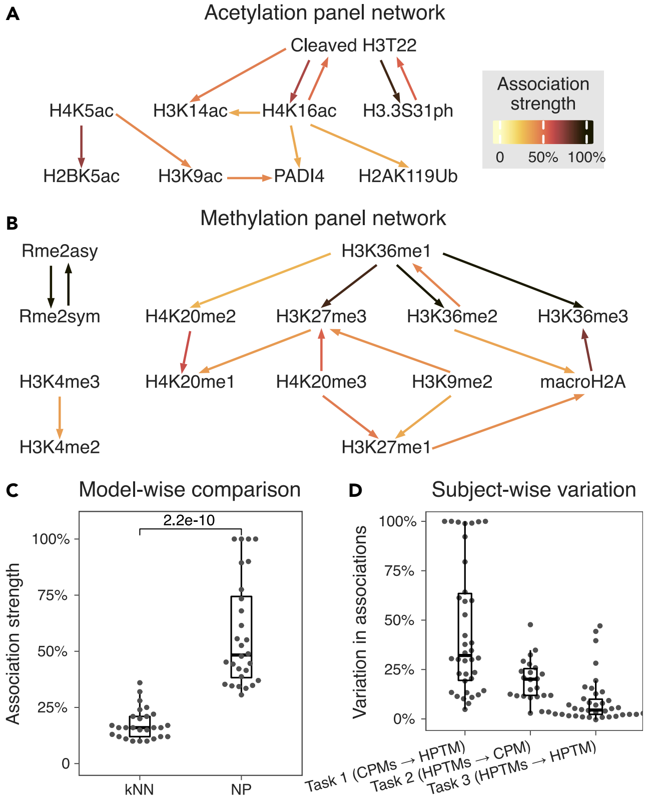


Figure 2. Inferred HPTM associations and subject-wise variability in healthy subjects

(A and B) Acetylation and B) methylation panel HPTM networks inferred from *task 3* (HPTMs → HPTM) NP models. Arrows go from input to output HPTM.

(C) Comparison of inferred association strengths from NP and *k*NN models. Each dot represents the association strength of a pair of HPTMs in the association networks from the corresponding models. Wilcoxon test was used to compute the significance in the difference of means (n = 27, 28 HPTM associations for *k*NN, NP respectively).

(D) Summary of variation in model association strength across subjects. Each dot represents a model. Association strength of a model is the strength of all associations containing input and output of the model.

See also Figures S3–S6.

R^2 reduced by 55%. The directionality of the identified association suggests biologically relevant causality, which is in line with our recent results showing that the abundance of H3.3S31ph is impacted by cleaved H3T22 in a cell line genetically modified to prevent cleavage of H3T22.[14] We also found that perturbing H3K9ac abundance reduced the accuracy of imputing PADI4 abundance, suggesting the functional interaction between the two, which is in line with results from Kolodziej et al. demonstrating a locus-specific interaction between H3K9ac and PADI4.[23]

For the methylation panel, perturbation analysis identified three distinct network components (Figures 2B and S3). Asymmetric and symmetric arginine dimethylation (Rme2asy/sym) strongly influenced each other and were disconnected from the lysine methylation network. This is consistent with the fact that the arginine and lysine methylations are catalyzed by different sets of enzymes. The second component consisted of H3K4me2/3, which is known to promote gene expression and occur near promoters and enhancers. The third component, on the other hand, contained known markers of gene repression and heterochromatin, including H3K27me3, H3K9me2, H4K20me3, and macroH2A. In this network component, we found that perturbing H3K9me2 abundance reduced the accuracy of imputing H3K27me3 abundance, which is in line with H3K9me2 and H3K27me3 cooperating to maintain heterochromatin.[24]

In addition to these known interactions, we identified directional *trans* histone interactions between cleaved H3T22 and H4K16ac, and between H3K9ac, H4K5ac and H2BK5ac. We also found directional *cis* interactions between modifications of the same amino acid such as H3K36me1/2/3 and H4K20me2/3

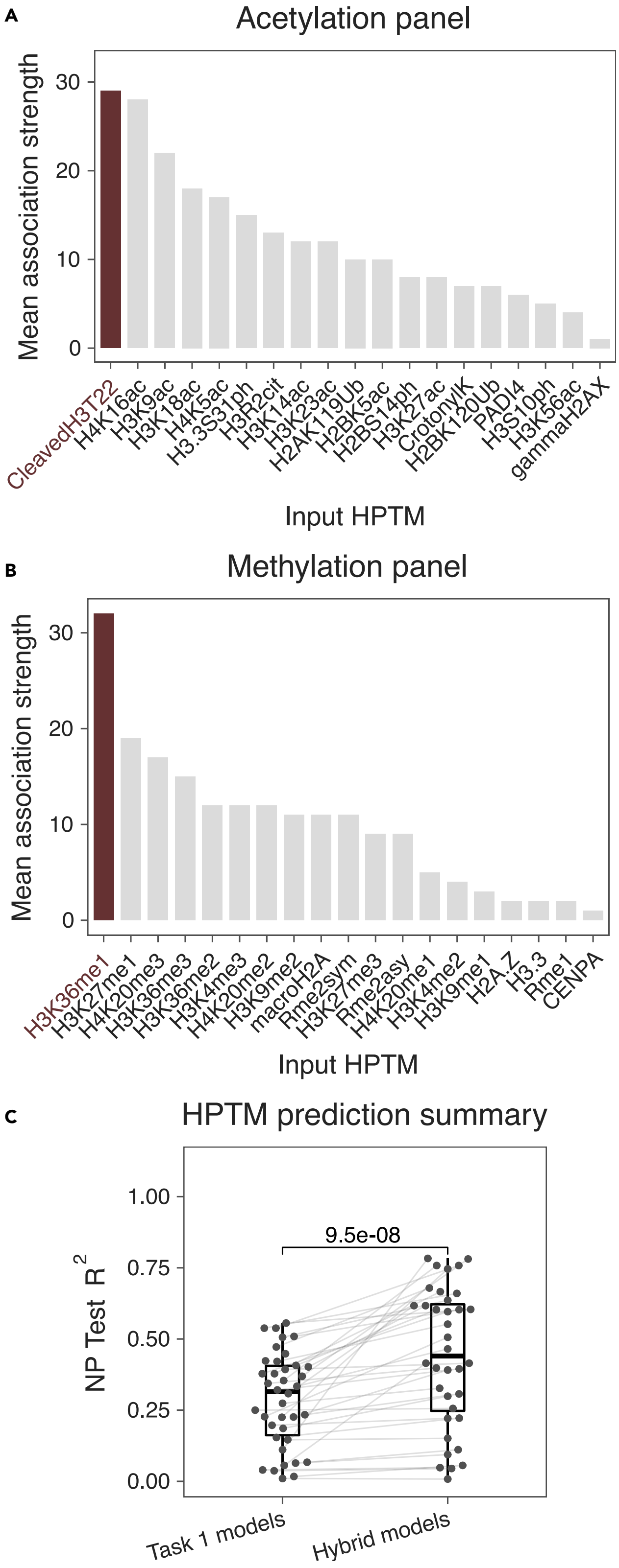

A
Acetylation panel
Mean association strength
0
10
20
30
CleavedH3T22
H4K16ac
H3K9ac
H3K18ac
H4K5ac
H3.3S31ph
H3R2cit
H3K14ac
H3K23ac
H2AK119Ub
H2BK5ac
H2BS14ph
H3K27ac
CrotonylK
H2BK120Ub
PADI4
H3S10ph
H3K56ac
gammaH2AX
Input HPTM
B
Methylation panel
Mean association strength
0
10
20
30
H3K36me1
H3K27me1
H4K20me3
H3K36me3
H3K36me2
H3K4me3
H4K20me2
H3K9me2
macroH2A
Rme2sym
H3K27me3
Rme2asy
H4K20me1
H3K4me2
H3K9me1
H2A.Z
H3.3
Rme1
CENPA
Input HPTM
C
HPTM prediction summary
NP Test R^2
0.00
0.25
0.50
0.75
1.00
9.5e-08
Task 1 models
Hybrid models

Figure 3. Hybrid NP models using CPMs and highly predictive HPTMs have higher imputation accuracy than using only CPMs

(A and B) Average association strength for each HPTM in the (A) acetylation and (B) methylation panels.

(C) Summary of hybrid NP models that use CPMs and cleaved H3T22 (acetylation panel) or H3K36me1 (methylation panel) as inputs to impute an HPTM and comparison with models in *task 1* (CPMs → HPTM). Each dot represents a unique model. Lines connect models predicting the same HPTM. One-sided, paired Wilcoxon test was used to compute the significance of the improvement in R^2 due to the hybrid models (n = 36 HPTMs per task).

See also Figure S7.

(discussion). For several HPTMs imputed with $R^2 > 0.5$ (e.g., H3K36me2/3, Rme2asy/sym, cleaved H3T22, H3.3S31ph, H4K16ac), there was at least one input HPTM leading to >50% decrease in R^2 when its abundance was replaced with the noise, suggesting highly conserved interactions between these HPTMs (Figure S3). Conversely, several HPTMs imputed with $R^2 > 0.5$ (e.g., H3K36me1, H4K20me2/3, H3K9ac, H4K5ac, PADI4, macroH2A) had no such inputs. Since the latter HPTMs were accurately imputed, detected associations suggest interactions between their inputs (discussion). Taken together, our analyses identified several known associations between pairs of HPTMs together with their directionality. We also identified several associations that lead to new hypotheses to further explore the crosstalk of HPTMs.

In contrast, when we applied our perturbation algorithm to the *k*NN models, the strength of associations was significantly lower (p = 2.2e-10; Figures 2C, S4, and S5). This is expected because *k*NN models are non-parametric and perturbation to any input HPTM leads to the recalibration of the kNN model, which in turn can result in lower strength of associations. In contrast, NP models do not recalibrate during the perturbation analysis. Furthermore, *k*NN models only identified a subset of known HPTM associations (e.g., H3K4me2-H3K4me3 and Rme2sym-Rme2asy), but failed to identify the directionality of the other known associations. For instance, *k*NN found an equally strong bidirectional association between cleaved H3T22 and H3.3S31ph (Figure S5) and did not identify the association between PADI4 and H3K9ac. Overall, NP more accurately imputed HPTM abundances and better identified direction of interaction between HPTMs than *k*NN.

Histone post-translational modification associations are conserved across subjects

Each NP model is subject-specific and learns a unique embedding for each subject from the input data. However, we averaged the inferred HPTM associations across all subjects in the test set. Thus, it is possible that our overall associations are driven by a subset of subjects. Therefore, we modified our perturbation algorithm to quantify the between-subject variability in the associations (STAR Methods). Briefly, we replaced the embedding for a subject with that of a different subject and estimated the between-subject variability as the percentage change in R^2. Variability of 0% implies the model is independent of the subject embedding and the pair-wise associations for a given NP model are conserved, whereas variability of 100% implies the associations are distinct for each subject.

The NP models predicting an HPTM from CPMs (*task 1*) showed the highest between-subject variability (median = 32.1%; Figure 2D). This is in line with single-cell RNA-seq studies repeatedly showing a large amount of heterogeneity in immune cells identified using CPMs.[21,22] The NP models predicting a CPM from HPTMs (*task 2*) had moderate between-subject variability (median = 20.1%), which is also expected since we have observed immune cell type-specific HPTM profiles that are affected by aging.[13,14] The NP models predicting an HPTM from other HPTMs (*task 3*) had the lowest between-subject variability (median = 4.5%; Figure S6). Overall, our results show that the associations between CPMs and HPTMs vary between subjects, presumably affected by their different environmental exposures. On the other hand, low between-subject variability in the interactions between HPTMs suggests that most of the interactions between HPTMs are highly conserved such that changes in one HPTM lead to predictable changes in the other HPTMs.

A subset of histone post-translational modifications accurately imputed several histone post-translational modifications

Based on low between-subject variability in predicting an HPTM using other HPTMs, combined with our previous observation of highly conserved HPTM correlation modules,[15] we hypothesized that a subset of HPTMs may be more informative than others. To test this hypothesis, we ranked HPTMs using their average association strength, defined as the mean strength of all the associations containing the given HPTM as the input (Figures 3A, 3B, and S3). Cleaved H3T22 and H4K16ac in the acetylation panel, and H3K36me1 in the methylation panel were the strongest predictors of 12 other HPTMs (mean association strength >25%), suggesting an important role for them in HPTM networks. While cleaved H3T22 and H3K36me1 were among the top four most abundant HPTMs, H4K16ac was not among the top 10 most abundant HPTMs (Figure 1F). Hence, our results suggest that the abundance of HPTM is not confounding the average strength of HPTMs associations.

Next, we investigated whether the integration of these HPTMs with CPMs would improve the imputation accuracy of other HPTMs compared to using only CPMs as in *task 1*. The new hybrid NP models using 11 CPMs and either cleaved H3T22 or H3K36me1 to impute HPTMs had significantly higher R^2 (mean = 0.44, range: 0.01-0.78) than the models using only CPMs (p = 9.5e-08; Figure 3C). Importantly, the hybrid NP models had 17 HPTMs with $R^2 \geq 0.5$; more than threefold increase in accurately imputed HPTMs compared to HPTMs imputed from only CPMs (*task 1*). Although including H4K16ac with cleaved H3T22 and 11 CPMs (mean = 0.453, range: 0-0.78) significantly increased the R^2 (p = 0.011) compared to including only cleaved H3T22, it did not increase the number of HPTMs imputed with $R^2 \geq 0.5$ (Figure S7). These results show that a modified panel with H3K36me1 and cleaved H3T22 measured in both panels would enable imputing multiple HPTMs across panels with higher accuracy, further increasing the power of EpiTOF. This also highlights the power of our interpretation algorithm in accurately identifying associations and ranking the predictive power of each HPTM.

Neural processes models and interpretation algorithms identify influenza vaccine-associated epigenetic changes

We demonstrated that the NP models can accurately impute HPTM abundances and infer known and previously unidentified associations for healthy subjects. Next, we investigated if the same models, without any modification or re-training, could be used to impute HPTM abundances from healthy subjects affected by external stimuli such as vaccinations. We recently showed that for healthy subjects receiving the trivalent inactivated seasonal influenza vaccine (TIV), HPTM abundances are most changed 30 days after vaccination in myeloid cells, which are associated with innate memory.[8] Therefore, we evaluated the NP models imputing an HPTM from other HPTMs (*task 3*) and inferred pair-wise HPTM associations in a cohort of 21 healthy subjects sampled before (*Day 0*) and 30 days after (*Day 30*) receiving TIV (Figure 4A, STAR Methods). The imputation R^2 at both these time points were strongly correlated with those from the healthy subjects in the test set (Figure 4B, R = 0.9 and 0.87, and p < 1e-12 for *Day 0* and *Day 30* samples, respectively), indicating that the same models can be used to impute HPTMs without significant change in imputation performance (Figure S8). Although the inferred association strengths at both these time points were also highly correlated with those from the healthy subjects in the test split (Figure 4C, R = 0.94, 0.91 and p < 2.2e-16 for *Day 0* and *Day 30* samples, respectively), some association strengths showed large deviations from the test set (Figure S8), suggesting that most HPTM associations are conserved but a few are significantly modified due to TIV.

To identify associations that are modified due to TIV, we compared the HPTM association networks independently inferred from the *Day 0* and *Day 30* samples. 12 out of 16 *Day 0* associations in the acetylation panel network and 17 out of 19 *Day 0* associations in the methylation panel network were conserved with identical association strengths in *Day 30* samples (Figures 4D and 4E). However, in the acetylation panel, the association between cleaved H3T22 and H3K14ac, H4K16ac and PADI4, H3K18ac and H3R2cit, and H2K9ac and H4K5ac were present only on *Day 0* but not on *Day 30* (Figure 4D). In the methylation panel, the association between H4K20me3 and H3K27me3, and H3K9me2 and H3K9me1 was present only on *Day 0* but not on *Day 30* (Figure 4E). Across both panels, all the HPTM pairs associated on *Day 30* were also associated on *Day 0*. Taken together, the networks revealed that the associations between 6 pairs of 12 HPTMs were significantly different on *Day 30* post-vaccination.

Furthermore, across both EpiTOF panels, PBMCs from the two-time points were distributed differently in the UMAP[25] space defined by these HPTMs (Figures 4F and 4H). To further quantify these distinct patterns, we clustered[26] the PBMCs (Figures 4G and 4I), computed the median abundance of the HPTMs for each cluster in each panel, and compared the sample-wise proportion of cells in a cluster across the two-time points (Figures 4J and 4K, STAR Methods). In the acetylation panel, the cell proportions were significantly different in clusters *3* (FDR = 7.2%), *7* (FDR = 4.1%), and *12* (FDR = 5.1%) (Figures 4J and S9). Similarly, in the methylation panel, the cell proportions were significantly different in cluster *3* (FDR = 8.4%) (Figures 4K and S9). These suggest that the changes in these HPTM associations are driven by a change in cell proportions in these clusters.

We have previously shown that the overall proportions of immune cell sub-types are not significantly different between *Day 30* and *Day 0* samples.[8] However, based on the clustering analysis, we hypothesized that the proportions of immune cell sub-types could be significantly different in the clusters where the overall cell

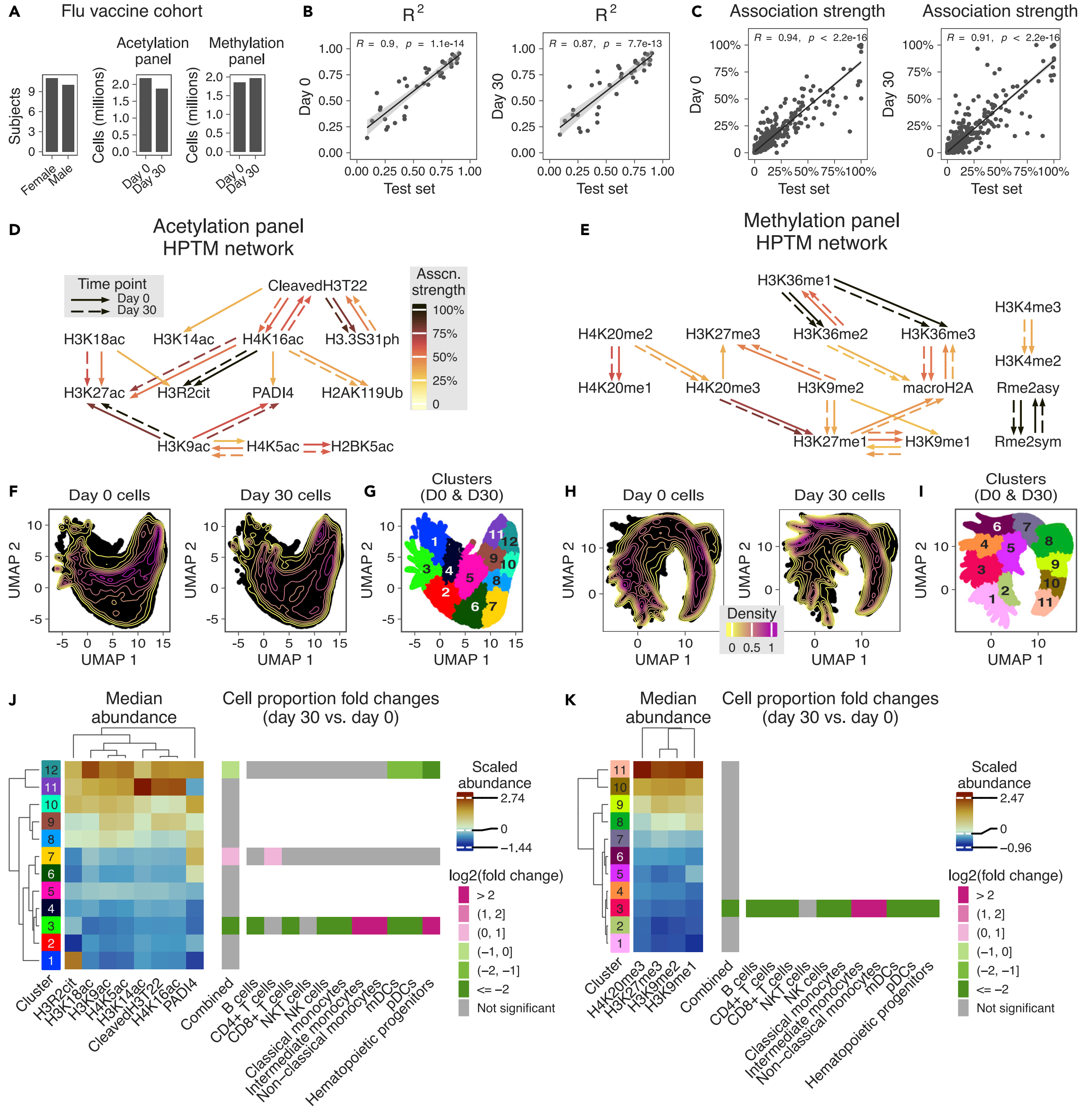


Figure 4. Inferred HPTM networks identify changes in HPTM associations following influenza vaccination

(A) Summary of Flu vaccine cohort.

(B) Comparison of NP *task 3* R^2 between Flu vaccine cohort samples and test set. Each dot represents a unique HPTM (n = 38 per time point).

(C) Comparison of HPTM association strength between Flu vaccine cohort samples and test split. Each dot represents a unique pair of HPTMs (n = 684 per time point). T-test was used to compute the significance of the Pearson correlation coefficient.

(D and E) (D) Acetylation and (E) methylation panel HPTM networks from *Day 0* and *Day 30* time points. Arrows go from input to output HPTM.

(F) UMAP visualization of cells in acetylation panel based on HPTMs that change in the association network (cleaved H3T22, H3K14ac, H4K16ac, PADI4, H3K18ac, H3R2cit, H2K9ac, and H4K5ac). Each time point contains 250,000 cells sampled randomly.

(G) PhenoGraph clustering of cells in acetylation panel. Clustering was performed in the UMAP space and used 25,000 cells per time point randomly sampled from those used for UMAP.

(H) UMAP visualization of cells in methylation panel based on HPTMs that change in the association network (H4K20me3, H3K27me3, H3K9me2, and H3K9me1). Each time point contains 250,000 cells sampled randomly.

Figure 4. *Continued*

(I) PhenoGraph clustering of cells in methylation panel. Clustering was performed in the UMAP space and used 25,000 cells per time point randomly sampled from those used for UMAP.

(J and K) Median abundance of HPTMs and fold changes in sample-wise cell proportion between *Day 30* and *Day 0* in the (J) acetylation and (K) methylation panels. All cells from all samples were used to calculate the medians and proportions. FDR-adjusted Wilcoxon test was used to compute the significance of differences in proportion means across groups (n = 21 subjects per time point).

See also Figures S8 and S9.

proportions are significantly different. Therefore, we compared the cell proportions within clusters *3*, *7*, and *12* in the acetylation panel and cluster *3* in the methylation panel (Figures 4J, 4K, and S9). Cluster *3* in the acetylation panel was defined by the low abundance of the HPTMs and was dominated by cells from *Day 0* samples. Importantly, in this cluster, the proportions of B cells, CD8$^+$ T cells, NK cells, classical monocytes, mDCs, and pDCs significantly decreased, whereas those of intermediate and non-classical monocytes, and hematopoietic progenitors significantly increased in *Day 30* samples compared to *Day 0*. The proportions of intermediate and non-classical monocytes are known to increase in response to infections and inflammatory conditions.[27] In contrast, cluster *7*, which was also defined by low abundances of several HPTMs and moderate abundance of PADI4, had a significantly higher proportion of cells from *Day 30* samples, which was solely due to a significant increase in CD4$^+$ T cell proportions. Cluster *12*, defined by moderate to high abundances of the HPTMs, had a significantly lower proportion of cells from *Day 30* samples compared to *Day 0*, which was due to significantly reduced proportions of mDCs, pDCs, and hematopoietic progenitors on *Day 30* post-vaccination. In the methylation panel, the proportion of cells from *Day 30* samples was significantly decreased in cluster 3 compared to *Day 0*. In this cluster, defined by the low abundance of the HPTMs, the proportions of B cells, CD4$^+$ and CD8$^+$ T cells, NK cells, classical monocytes, mDCs, pDCs, and hematopoietic progenitors decreased significantly in *Day 30* samples compared to *Day 0*. However, the proportions of intermediate and non-classical monocytes were significantly higher, similar to cluster *3* in the acetylation panel. Thus, NP models and perturbation analysis led to the identification of HPTMs whose associations changed due to TIV, and clusters defined by these HPTMs showed significant differences in cell proportions. The observation of cluster-specific cell proportion changes, but not at the overall level, suggests epigenetic reprogramming of the immune cells in these clusters, which is in line with recent studies demonstrating that memory in the innate immune system is driven by epigenetic reprogramming.[9,28]

In summary, we have shown that the NP models trained using EpiTOF profiles of healthy subjects can be used to impute HPTM abundances in post-vaccinated subjects without any change in imputation accuracy. The associations between 6 pairs of HPTMs are modified due to TIV, and the clusters of cells that possibly drive these differences in associations show differences in proportions of immune cell sub-types.

DISCUSSION

We profiled 38 HPTMs and histone variants with 11 CPMs across 2 panels in more than 55 million PBMCs from 158 healthy controls across 25 experiments using EpiTOF. Using this unique dataset, we developed a computational framework for inferring directional associations between two HPTMs. As part of the development, we evaluated three machine learning methods for imputation to conclusively show that NP models impute HPTMs and CPMs with significantly higher accuracy, measured as R^2, compared to LR and *k*NN models.

Increased accuracy of NP models required larger amounts of training data, computational resources, and computational time to train than LR and *k*NN models. However, once trained, NP models offered several advantages over LR and *k*NN models. First, NP models better captured associations from the underlying data than *k*NN models and provided new biological insights. Second, NP models were faster than *k*NN models when making predictions during testing. Third, the learnings from the trained NP models can be transferred to other settings using transfer learning approaches. For instance, when a new HPTM is added to the EpiTOF panel, the data and computational time required for training can be significantly reduced by using transfer learning approaches. Importantly, the NP approach can be applied to impute other multi-panel data sources such as CyTOF. Taken together, the advantages of the NP model outweigh its limitations for increased resources.

We found that the CPMs and HPTMs impute each other with low to moderate R^2. The best imputed CPM (using HPTMs as input) was CD11c (R^2 = 0.6), which is abundant only in cells from the myeloid lineage. This suggests a distinct pattern of HPTM abundances in the myeloid and lymphoid lineages. Indeed, we recently found that myeloid and lymphoid cells follow epigenetically distinct trajectories during their differentiation from hematopoietic progenitors.[15] However, other CPMs were imputed with $R^2 < 0.5$. This is not surprising, since these CPMs delineate PBMCs into broad immune cell subtypes (e.g., CD4+/CD8+ T cells, classical/non-classical monocytes) whereas chromatin states, and hence HPTM abundances, which lead to mRNA expression, show substantial variability within the same cellular population.[29]

We found that the HPTMs accurately imputed other HPTMs, and the perturbation-based analysis allowed inferring directionality between a pair of HPTMs. Our analysis correctly identified several known HPTM interactions. Importantly, our analyses also identified several previously unidentified associations and HPTM networks that should be further investigated in the future studies to decode the crosstalk between HPTMs. For example, we identified directed interactions between mono-, di-, and tri-methylation of lysine 36 on histone H3. Although H3K36me2 and H3K36me3 have been well studied, not much is known about H3K36me1.[30] *NSD1* is known to catalyze H3K36me1/2 and *SETD2* catalyzes H3K36me3 *in vivo*. However, it is unclear whether the substrate for the *SETD2*-dependent catalysis of H3K36me3 is H3K36me1, H3K36me2, or both.[31] Our perturbation analysis strongly suggests that H3K36me1, not H3K36me2, is a stronger predictor of H3K36me3, which in turn suggests that H3K36me1 is the substrate for *SETD2*-dependent catalysis of H3K36me3.

We identified cleaved H3T22 and H3K36me1 as strong predictors for HPTMs in the acetylation and methylation panels respectively and showed that a hybrid model using cleaved H3T22 and H3K36me1 with the CPMs significantly improves imputation accuracy for HPTMs. We note that both cleaved H3T22 and H3K36me1 are not measured in ENCODE[32] and are understudied.[30,33] Our results indicate a pair of highly abundant yet understudied HPTMs to play a central role in HPTM associations and crosstalk.

The inferred HPTM associations also revealed several accurately imputed HPTMs (e.g., H3K36me1, H4K20me2/3, H3K9ac, H4K5ac, PADI4, and macroH2A) that had no single strong predictor. These likely suggest that several predictors interact with each other such that the removal of a single predictor doesn't significantly affect the imputation of these HPTMs. Inferring these interactions would require performing perturbation analyses on several combinations of inputs for a single output. In this work, we only focus on pair-wise associations between an input and an output.

Limitations of the study

Our study has a few limitations. First, although the networks inferred from the interpretation framework provide valuable insights about associations between HPTMs, our analysis does not identify the immune cell sub-type where these associations are identified in this proof-of-concept study. As more data become available, future studies should train NP models per cell type to potentially further increase the accuracy of imputation and derive cell type-specific HPTM associations. Second, we focus only on the pair-wise associations from an input to an output. Future studies should perturb multiple combinations of inputs to quantify the effect of their interactions in the prediction of an output.

In summary, we described a machine learning-based framework to impute abundances for a subset of HPTMs with high accuracy. We found that NP-based models, which had the highest accuracy, have several advantages compared to other imputation methods. First, NP-based models capture variability at the cellular and subject levels using subject-specific functions. Second, because the NP models are based on deep learning, current trained models can be adapted using transfer learning to impute HPTMs added to the EpiTOF panel in the future. We also demonstrated an interpretation algorithm that infers valuable insights into the underlying biological processes from the data. Our methods have the potential to be applied to other single-cell frameworks such as CyTOF and scRNA-seq to infer directed associations.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ EpiTOF data

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.105756.

ACKNOWLEDGMENTS

PK is funded in part by the Bill and Melinda Gates Foundation (OPP1113682); the National Institute of Allergy and Infectious Diseases (NIAID) grants 1U19AI167903-01, 5U01AI165527-02, 1U19AI109662, U19AI057229, and 5R01AI125197; National Institutes of Health contract 75N93022C00052; Department of Defense contracts W81XWH-18-1-0253 and W81XWH1910235; and the Ralph & Marian Falk Medical Research Trust. PJU is supported in part by the Donald E. and Delia B. Baxter Foundation, Elizabeth F. Adler, the Henry Gustav Floren Trust, the Bill & Melinda Gates Foundation (OPP1113682), and the NIH grants U19 AI110491 (Autoimmunity Center of Excellence), R01 AI125197.

AUTHOR CONTRIBUTIONS

PK conceived the study. PK and PJU obtained funding. AG designed and performed computational analyses, and wrote code with help from RS. AG, DD, LK, and MD interpreted data. PJU supervised EpiTOF profiling. AG and PK wrote the article with contributions from all coauthors. PK and PJU supervised the study. All authors approved the article.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 26, 2022
Revised: November 19, 2022
Accepted: December 6, 2022
Published: January 20, 2023

REFERENCES

1. Paksa, A., and Rajagopal, J. (2017). The epigenetic basis of cellular plasticity. Curr. Opin. Cell Biol. *49*, 116–122. https://doi.org/10.1016/j.ceb.2018.01.003.

2. Flavahan, W.A., Gaskell, E., and Bernstein, B.E. (2017). Epigenetic plasticity and the hallmarks of cancer. Science *357*, eaal2380. https://doi.org/10.1126/science.aal2380.

3. Zhao, Z., and Shilatifard, A. (2019). Epigenetic modifications of histones in cancer. Genome Biol. *20*, 245. https://doi.org/10.1186/s13059-019-1870-5.

4. Bannister, S., Messina, N.L., Novakovic, B., and Curtis, N. (2020). The emerging role of epigenetics in the immune response to vaccination and infection: a systematic review. Epigenetics *15*, 555–593. https://doi.org/10.1080/15592294.2020.1712814.

5. Wang, J., Hodes, G.E., Zhang, H., Zhang, S., Zhao, W., Golden, S.A., Bi, W., Menard, C., Kana, V., Leboeuf, M., et al. (2018). Epigenetic modulation of inflammation and synaptic plasticity promotes resilience against stress in mice. Nat. Commun. *9*, 477. https://doi.org/10.1038/s41467-017-02794-5.

6. Klein, K., and Gay, S. (2013). Epigenetic modifications in rheumatoid arthritis, a review. Curr. Opin. Pharmacol. *13*, 420–425. https://doi.org/10.1016/j.coph.2013.01.007.

7. Nemtsova, M.V., Zaletaev, D.V., Bure, I.V., Mikhaylenko, D.S., Kuznetsova, E.B., Alekseeva, E.A., Beloukhova, M.I.,

Deviatkin, A.A., Lukashev, A.N., and Zamyatnin, A.A. (2019). Epigenetic changes in the pathogenesis of rheumatoid arthritis. Front. Genet. *10*, 570. https://doi.org/10.3389/fgene.2019.00570.

8. Wimmers, F., Donato, M., Kuo, A., Ashuach, T., Gupta, S., Li, C., Dvorak, M., Foecke, M.H., Chang, S.E., Hagan, T., et al. (2021). The single-cell epigenomic and transcriptional landscape of immunity to influenza vaccination. Cell *184*, 3915–3935.e21. https://doi.org/10.1016/j.cell.2021.05.039.

9. Netea, M.G., Schlitzer, A., Placek, K., Joosten, L.A.B., and Schultze, J.L. (2019). Innate and adaptive immune memory: an evolutionary continuum in the host's response to pathogens. Cell Host Microbe *25*, 13–26. https://doi.org/10.1016/j.chom.2018.12.006.

10. Lee, J.-S., Smith, E., and Shilatifard, A. (2010). The Language of histone crosstalk. Cell *142*, 682–685. https://doi.org/10.1016/j.cell.2010.08.011.

11. Strahl, B.D., and Briggs, S.D. (2021). The SAGA continues: the rise of cis- and trans-histone crosstalk pathways. Biochim. Biophys. Acta. Gene Regul. Mech. *1864*, 194600. https://doi.org/10.1016/j.bbagrm.2020.194600.

12. Grosselin, K., Durand, A., Marsolier, J., Poitou, A., Marangoni, E., Nemati, F., Dahmani, A., Lameiras, S., Reyal, F., Frenoy, O., et al. (2019). High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. Nat. Genet. *51*, 1060–1066. https://doi.org/10.1038/s41588-019-0424-9.

13. Cheung, P., Vallania, F., Warsinske, H.C., Donato, M., Schaffert, S., Chang, S.E., Dvorak, M., Dekker, C.L., Davis, M.M., Utz, P.J., et al. (2018). Single-cell chromatin modification profiling reveals increased epigenetic variations with aging. Cell *173*, 1385–1397.e14. https://doi.org/10.1016/j.cell.2018.03.079.

14. Cheung, P., Schaffert, S., Chang, S.E., Dvorak, M., Donato, M., Macaubas, C., Foecke, M.H., Li, T.-M., Zhang, L., Coan, J.P., et al. (2021). Repression of CTSG, ELANE and PRTN3-mediated histone H3 proteolytic cleavage promotes monocyte-to-macrophage differentiation. Nat. Immunol. *22*, 711–722. https://doi.org/10.1038/s41590-021-00928-y.

15. Dermadi, D., Kalesinskas, L., Ganesan, A., Kuo, A., Cheung, P., Cheng, S., Dvorak, M., Scriba, T.J., Habtezion, A., Donato, M., et al. (2022). Crosstalk of histone modifications in the healthy human immune system. Preprint at BioRxiv. https://doi.org/10.1101/2022.01.21.477300.

16. Abdelaal, T., Höllt, T., van Unen, V., Lelieveldt, B.P.F., Koning, F., Reinders, M.J.T., and Mahfouz, A. (2019). CyTOFmerge: integrating mass cytometry data across multiple panels. Bioinformatics *35*, 4063–4071. https://doi.org/10.1093/bioinformatics/btz180.

17. Brodin, P., and Davis, M.M. (2017). Human immune system variation. Nat. Rev. Immunol. *17*, 21–29. https://doi.org/10.1038/nri.2016.125.

18. Becht, E., Tolstrup, D., Dutertre, C.-A., Morawski, P.A., Campbell, D.J., Ginhoux, F., Newell, E.W., Gottardo, R., and Headley, M.B. (2021). High-throughput single-cell quantification of hundreds of proteins using conventional flow cytometry and machine learning. Sci. Adv. *7*, eabg0505. https://doi.org/10.1126/sciadv.abg0505.

19. Garnelo, M., Rosenbaum, D., Maddison, C.J., Ramalho, T., Saxton, D., Shanahan, M., Teh, Y.W., Rezende, D.J., and Eslami, S.M.A. (2018). Conditional neural processes. Preprint at ArXiv. https://doi.org/10.48550/arXiv.1807.01613.

20. Garnelo, M., Schwarz, J., Rosenbaum, D., Viola, F., Rezende, D.J., Eslami, S.M.A., and Teh, Y.W. (2018). Neural processes. Preprint at ArXiv. https://doi.org/10.48550/arXiv.1807.01622.

21. Papalexi, E., and Satija, R. (2018). Single-cell RNA sequencing to explore immune cell heterogeneity. Nat. Rev. Immunol. *18*, 35–45. https://doi.org/10.1038/nri.2017.76.

22. Chen, H., Ye, F., and Guo, G. (2019). Revolutionizing immunology with single-cell RNA sequencing. Cell. Mol. Immunol. *16*, 242–249. https://doi.org/10.1038/s41423-019-0214-4.

23. Kolodziej, S., Kuvardina, O.N., Oellerich, T., Herglotz, J., Backert, I., Kohrs, N., Buscató, E.I., Wittmann, S.K., Salinas-Riester, G., Bonig, H., et al. (2014). PADI4 acts as a coactivator of Tal1 by counteracting repressive histone arginine methylation. Nat. Commun. *5*, 3995. https://doi.org/10.1038/ncomms4995.

24. Boros, J., Arnoult, N., Stroobant, V., Collet, J.-F., and Decottignies, A. (2014). Polycomb repressive complex 2 and H3K27me3 cooperate with H3K9 methylation to maintain heterochromatin protein 1 at chromatin. Mol. Cell Biol. *34*, 3662–3674. https://doi.org/10.1128/MCB.00205-14.

25. McInnes, L., Healy, J., and Melville, J. (2020). UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at ArXiv. https://doi.org/10.48550/arXiv.1802.03426.

26. Levine, J.H., Simonds, E.F., Bendall, S.C., Davis, K.L., Amir, E.a.D., Tadmor, M.D., Litvin, O., Fienberg, H.G., Jager, A., Zunder, E.R., et al. (2015). Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. Cell *162*, 184–197. https://doi.org/10.1016/j.cell.2015.05.047.

27. Wong, K.L., Yeap, W.H., Tai, J.J.Y., Ong, S.M., Dang, T.M., and Wong, S.C. (2012). The three human monocyte subsets: implications for health and disease. Immunol. Res. *53*, 41–57. https://doi.org/10.1007/s12026-012-8297-3.

28. Rodriguez, R.M., Suarez-Alvarez, B., and Lopez-Larrea, C. (2019). Therapeutic epigenetic reprogramming of trained immunity in myeloid cells. Trends Immunol. *40*, 66–80. https://doi.org/10.1016/j.it.2018.11.006.

29. Carter, B., and Zhao, K. (2021). The epigenetic basis of cellular heterogeneity. Nat. Rev. Genet. *22*, 235–250. https://doi.org/10.1038/s41576-020-00300-0.

30. Li, J., Ahn, J.H., and Wang, G.G. (2019). Understanding histone H3 lysine 36 methylation and its deregulation in disease. Cell. Mol. Life Sci. *76*, 2899–2916. https://doi.org/10.1007/s00018-019-03144-y.

31. Hyun, K., Jeon, J., Park, K., and Kim, J. (2017). Writing, erasing and reading histone lysine methylations. Exp. Mol. Med. *49*, e324. https://doi.org/10.1038/emm.2017.11.

32. Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., Kaul, R., et al.; ENCODE Project Consortium (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature *583*, 699–710. https://doi.org/10.1038/s41586-020-2493-4.

33. Kragesteen, B.K., and Amit, I. (2021). Heads or tails: histone tail clipping regulates macrophage activity. Nat. Immunol. *22*, 678–680. https://doi.org/10.1038/s41590-021-00941-1.

34. Slight-Webb, S., Thomas, K., Smith, M., Macwana, S., Bylinska, A., Kheir, J., Donato, M., Dvorak, M., Chang, S., Kuo, A., et al. (2021). Ancestry-based differences in the immune system are associated with lupus severity. Preprint at Res. Sq. https://doi.org/10.21203/rs.3.rs-1149629/v1.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
EpiTOF data (train, validation, and test set, and flu vaccine cohort)	Cheung et al., Dermadi et al., Slight-Webb et al., and Wimmers et al.[8,13,15,34]	https://khatrilab.stanford.edu/EpiTOF
Software and algorithms		
Neural processes	Garnelo et al.[20]	https://github.com/deepmind/neural-processes
MIDAS (Perturbation-based interpretation algorithm)	This paper	https://doi.org/10.5281/zenodo.7388350

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Purvesh Khatri (pkhatri@stanford.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- All datasets, software, and algorithms used in this study are publicly available and listed in the key resources table.
- This paper did not generate any unique datasets.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

EpiTOF data

The 158 healthy subjects were enlisted through 6 cohorts and their EpiTOF profiles were measured across 25 experiments (Table S2). We have previously described all 6 cohorts- the Atlanta cohort[8] (50 subjects, 13 experiments), BR (24 subjects, 2 experiments) and Twins cohorts[13] (40 subjects, 2 experiments), Oklahoma cohort[34] (18 subjects, 4 experiments), Stanford (16 subjects, 3 experiments) and South Africa cohorts[15] (10 subjects, 1 experiment). The effects of TIV were studied on 21 subjects from the Atlanta cohort. The *Day 30* time point samples were not present in any of training, validation, or test splits.

METHOD DETAILS

Methods for sample randomization and blinding (including their inclusion and exclusion criteria), imputation tasks and their corresponding inputs and outputs, neural processes and the corresponding neural network architecture, and inferrence of direction based on perturbation analysis are described below.

QUANTIFICATION AND STATISTICAL ANALYSIS

To ensure our models were robust, generalizable, and accounted for heterogeneity in healthy subjects, our experimental design and analysis accounted for the following:

Replication

We used 25 independent experiments that profiled blood samples from 158 healthy subjects. Each subject was profiled in exactly 1 experiment.

Strategy for randomization and/or stratification

We stratified datasets by experiment into training, validation, and test sets to ensure robustness and generalizability of computational models. Specifically, we randomly assigned 15 experiments (71 subjects) for training, 5 experiments (52 subjects) for validation, and 5 experiments (35 subjects) for test of our computational models. In stratifying our datasets by experiments, we aimed to ensure that there was no information leakage between training, validation, and test sets.

Blinding at any stage of the study

5 experiments (35 subjects) used as test set were blinded until the computational models (LR, kNN, and NP) were optimized and locked using the training and validation samples.

Sample-size estimation and statistical method of computation

Sample size estimation was not performed because our models were at single-cell resolution, where we had >55 million cells.

Inclusion and exclusion criteria of any data or subjects

In order to account for heterogeneity in healthy subjects and demonstrate that our models are generalizable, we did not have any inclusion or exclusion criteria.

All statistical analyses were performed using R and corresponding details can be found on the figure legends. The statistical significance was computed using the non-parametric Wilcoxon test, which does not have any underlying assumptions about the data distributions.

Imputation tasks, model inputs and outputs

We evaluated models imputing a CPM or an HPTM across 4 tasks: an HPTM imputed using CPMs (*task 1*), a CPM imputed using HPTMs (*task 2*), an HPTM imputed using other HPTMs on the same panel (*task 3*), and an HPTM imputed using CPMs and best predicting HPTM (hybrid models). Each model also uses the abundances of H3 and H4, age, and sex of the subject as inputs.

Data scaling and splitting

We scaled each CPM and HPTM by subject to have mean 0 and standard deviation 1. Age was divided by 100 and sex was one-hot encoded. To prevent overfitting and ensure generalizability, we split the data by experiments into train, validation, and test, while keeping at least 5 experiments in each split (Table S2).

NP models: Design overview

We first describe how a trained and tested model will be used for imputing an HPTM across panel and use the example of imputing H3K36me1 (*Y*) from CPMs, H3, H4, age and sex (*X*) (*task 1*) for a subject *s* (Figure 1C). The NP model consists of two neural networks- the *encoder* and the *imputer*. The *encoder* processes cells from the panel where *Y* is measured and known. For H3K36me1, it is the methylation panel. We denote the cells processed by the *encoder* as the *context cells*. First, the *encoder* encodes each *context cell*:

$$cell\ encoding_{context\ cell^i} = encoder(X_i, y_i), \qquad \text{(Equation 1)}$$

where $X_i = [x_{CPM\ 1^i}, x_{CPM\ 2^i}, \ldots x_{CPM\ 11^i}, x_{H3^i}, x_{H4^i}, x_{age^s}, x_{male^s}, x_{female^s}] \in \mathbb{R}^{16}, y_i \in \mathbb{R},$ and $cell\ encoding_{context\ cell^i} \in R^{512}$. $x_{CPM\ 1^i}$ and y_i represent the measured abundance of CPM 1 and H3K36me1 respectively in the i^{th} *context cell* for subject *s*.

Next, the subject embedding for *s* is calculated as

$$subject\ embedding_s = mean_i\left(cell\ encoding_{context\ cell^i}\right), \qquad \text{(Equation 2)}$$

where $subject\ embedding_s \in \mathbb{R}^{512}$.

The *imputer* processes cells from the panel where *Y* needs to be imputed. For H3K36me1, it is the acetylation panel. We denote the cells processed by the *imputer* as the *target cells*. The *imputer* imputes H3K36me1 abundance for each *target cell*:

$$\widehat{y}_j = imputer\left(X_j, subject\ embedding_s\right), \qquad \text{(Equation 3)}$$

where $\widehat{y}_j$ represents the imputed abundance of H3K36me1 in the j^{th} *target cell*.

NP models: Training and testing

For training and testing NP models, we only used cells from the panel where Y is measured. We randomly divided the cells from this panel into two equal parts and used the first as the *context cells* and the second as the *target cells*. Due to GPU memory limitations, we split cells from each subject into 20 parts randomly and processed each as a separate subject.

We computed the training loss as the sum of two terms

$$loss = MSE\left(subject\ embedding_{s,context\ cells}, subject\ embedding_{s,target\ cells}\right) + MSE\left(\widehat{Y}_{target\ cells}, Y_{target\ cells}\right), \qquad \text{(Equation 4)}$$

where $MSE()$ is the mean squared error , $subject\ embedding_{s,context\ cells} \in \mathbb{R}^{512}$ and $subject\ embedding_{s,target\ cells} \in \mathbb{R}^{512}$ are computed by the *encoder* using the *context* and *target* cells respectively, and $Y_{target\ cells} \in \mathbb{R}^{nt}$ and $\widehat{Y}_{target\ cells} \in \mathbb{R}^{nt}$ are the measured and the *imputer* imputed values respectively for the output corresponding to the *nt target cells*. The first term corresponds to the *encoder* network loss and the second the *imputer* network loss.

We used the coefficient of determination on the test split to assess the model performance:

$$R^2 = coef.determination\left(\widehat{Y}_{target\ cells}, Y_{target\ cells}\right). \qquad \text{(Equation 5)}$$

NP models: Architecture

The *encoder* and *imputer* networks are fully connected with 2 hidden layers of dimension 256 (Table S3). The input dimensions for each network varies by task. The encoder output has dimension 512 and the imputer output is a scalar.

kNN and LR models

*k*NN models impute Y for a *target cell* based on its nearest neighbors in the *context cells*. To be comparable with the NP models, the division of cells into *context* and *target* groups are the same for both NP and *k*NN. We evaluated $k \in \{1, 5, 10, 20, 50, 100, 200\}$ on the validation split and used the best k for each model on the test split.

LR models impute Y for a *target cell* based on the learned linear combination of the inputs X. The LR models do not use the *context cells*.

Perturbation-based algorithm to infer HPTM associations

To infer the association strength between an input variable X_i and output Y, we designed a perturbation-based interpretation algorithm. First, we replaced the measured values of X_i in the *target cells* for subject s with random values drawn from $\mathcal{N}(0,1)$ and imputed the output variable:

$$\widehat{Y}_{s,X_i\ perturbed} = imputer\left(X_{s,X_i\ perturbed}, subject\ embedding_s\right). \qquad \text{(Equation 6)}$$

We then evaluated R^2 after input perturbation:

$$R^2_{Y,X_i\ perturbed} = coef.determination\left(\widehat{Y}_{X_i\ perturbed}, Y\right) \qquad \text{(Equation 7)}$$

and then obtained the strength of the directed association from X_i to Y:

$$association\ strength_{X_i\ to\ Y} = \frac{R^2_{Y,unperturbed\ data} - R^2_{Y,X_i\ perturbed}}{R^2_{Y,unperturbed\ data}} \times 100. \qquad \text{(Equation 8)}$$

Constructing directed HPTM association network from inferred pair-wise association strengths

To obtain the association network for HPTMs in a panel from the pair-wise association strengths, we first removed HPTMs imputed with $R^2 \leq 0.5$. Among the associations between the remaining HPTMs, we calculated the threshold strength as

$$threshold\ strength = mean(asscn.\ strengths) + std.dev.(asscn.\ strengths). \qquad \text{(Equation 9)}$$

We then removed all associations with strength lesser than the *threshold strength* and combined the remaining associations into a single directed network.

Perturbation-based algorithm to infer subject-wise variation in NP models

To infer the subject-wise variation in an NP model, we replaced the computed *subject embedding*$_s$ for subject s with those from a different random subject $\hat{s}$ and imputed the output variable:

$$\hat{Y}_{s,subject\ perturbed} = imputer\left(X_s, subject\ embedding_{\hat{s}}\right). \qquad \text{(Equation 10)}$$

We then evaluated R^2 after input perturbation:

$$R^2_{Y,subject\ perturbed} = coef.determination\left(\hat{Y}_{subject\ perturbed,target\ cells}, Y_{target\ cells}\right) \qquad \text{(Equation 11)}$$

and then obtained the subject-wise variation for the NP model:

$$subject\ wise\ variation_Y = \frac{R^2_{Y,unperturbed\ data} - R^2_{Y,subject\ perturbed}}{R^2_{Y,unperturbed\ data}} \times 100. \qquad \text{(Equation 12)}$$

subject wise variation$_Y$ = 0% implies the model, and the underlying associations, are independent of the subject. The greater this value, the more the variation in the underlying associations across subjects.

Dimensionality reduction and clustering for studying the effects of TIV

We computed UMAP[25] projections on a subset of 1,000,000 cells (250,000 randomly sampled cells from each panel and time point) using *n neighbors* = 15 and min dist = 0.1. We performed phenograph clustering[26] on the UMAP space using a subset of 100,000 cells (25,000 randomly sampled cells from those used for UMAP from each panel and time point) using k = 1000. We then transformed every cell from both the time points to the UMAP space without modifying the UMAP projection function and assigned clusters based on the 5 nearest neighbors in the subset where the clusters were computed. We computed the cluster-wise median HPTM abundances and sample-wise cell proportions using all the cells.