Meta-matching as a simple framework to translate phenotypic predictive models from big to small data

Supplementary Materials

This supplemental material is divided into *Supplemental Methods, Supplemental Tables* and *Supplemental Figures*.

Supplementary Methods

This section provides additional implementation details of the meta-matching. Section S1 provides details about meta-matching with DNN. Section S2 provides details about meta-matching with DNN finetuning.

S1. Details about basic meta-matching (DNN)

In this section, we provide implementation details of DNN, which we utilized for basic meta-matching (DNN), as well as both advanced meta-matching algorithms.

- The DNN we considered is a generic feedforward neural network, which was implemented with default libraries (class "nn.Linear") in PyTorch[1].
- The loss function was MSE (mean squared error) loss. The output layer has 33 nodes, which is the number of training meta-set non-brain-imaging phenotypes (phenotypes).
- We used the HORD algorithm[2, 3, 4] to automatically tune the hyperparameters using the validation set (N = 5370) within the training meta-set. By setting a specific search range for multiple hyperparameters, the HORD algorithm was able to tune these hyperparameters within these ranges automatically. HORD does not perform well when there are too many hyperparameters to tune. Therefore, several hyperparameters were set based on our manual tuning using the training meta-set. These

[1] Paszke, A., Chanan, G., Lin, Z., Gross, S., Yang, E., Antiga, L., Devito, Z., 2017. Automatic differentiation in PyTorch. Adv. Neural Inf. Process. Syst. 30 1–4.
[2] Eriksson, D., Bindel, D., Shoemaker, C.A., 2019. pysot: Surrogate Optimization Toolbox [WWW Document]. GitHub. URL https://github.com/dme65/pySOT
[3] Ilievski, I., Akhtar, T., Feng, J., Shoemaker, C.A., 2017. Efficient hyperparameter optimization of deep learning algorithms using deterministic RBF surrogates, in: 31st AAAI Conference on Artificial Intelligence, AAAI 2017. pp. 822–829.
[4] Regis, R.G., Shoemaker, C.A., 2013. Combining radial basis function surrogates and dynamic coordinate search in high-dimensional expensive black-box optimization. Eng. Optim. 45, 529–555. https://doi.org/10.1080/0305215X.2012.687731

hyperparameters were stochastic gradient descent (SGD) with 0.9 momentum, 128 for batch size and Xavier uniform for weight initialization.

- Table S4 shows the search ranges of hyperparameters tuned by the HORD algorithm. We ran 200 HORD evaluation rounds. For each HORD evaluation round, 1000 epochs were run. DNN was trained on the training set (within the training meta-set) and evaluated on the validation set (within the training meta-set) for each epoch. The epoch with the best coefficient of determination (COD) on the validation set was chosen as the optimal epoch.

Hyperparameter tuned	Range
Number of layers	2 to 5
Number of nodes for each layer (separately)	2 to 512
Dropout rate	0 to 0.8
Starting learning rate	1e-2 to 1e-4
Epochs to decrease the learning rate	10 to 1000
Weight decay rate	1e-3 to 1e-7

Table S4. Search ranges of hyperparameters tuned by the HORD algorithm.

- Table S5 shows the final set of hyperparameters estimated by the HORD algorithm. The final DNN structure is a 4-layer DNN. The optimal epoch on the validation set is 118 epochs. After we obtained the best DNN on the training meta-set, we applied the trained DNN to the test meta-set.

Hyperparameter	Value
Number of layers	4
Number of nodes for each layer (separately)	87/386/313/33
Dropout rate	0.242
Starting learning rate	3.646e-03
Epochs to decrease learning rate	312
Weight decay rate	8.447e-04

Table S5. Final DNN hyperparameters estimated by the HORD algorithm.

S2. Lists of selected and removed UK Biobank non-brain-imaging phenotypes

We performed phenotype selection using kernel ridge regression (KRR) with 1000 randomly selected subjects. Here we include the full list of selected and removed UK Biobank phenotype (Data-Field) ID.

- 265 phenotypes have been selected: [age[5], 3, 31, 46, 47, 48, 49, 50, 77, 78, 93, 94, 95, 102, 129, 130, 135, 137, 398, 404, 709, 767, 777, 845, 864, 1070, 1090, 1160, 1578, 1588, 1845, 2946, 3062, 3063, 3064, 3085, 3143, 3144, 3147, 3148, 3659, 4079, 4080, 4100, 4101, 4104, 4105, 4106, 4119, 4120, 4123, 4124, 4125, 4138, 4143, 4144, 4145, 4146, 4194, 4230, 4250, 4253, 4255, 4256, 4286, 4288, 4289, 4429, 4440, 5089, 5100, 5101, 5106, 5109, 5114, 5115, 5157, 5162, 5257, 5262, 5263, 5306, 5983, 5984, 5986, 6032, 6033, 6333, 6348, 6373, 6374, 6382, 6772, 6773, 12143, 12144, 12336, 12340, 20007, 20008, 20009, 20015, 20016, 20023, 20075, 20127, 20133, 20149, 20150, 20151, 20153, 20155, 20156, 20157, 20159, 20161, 20162, 20195, 20200, 20229, 20230, 21001, 21002, 21003, 21004, 21621, 21631, 21651, 21663, 21664, 21671, 21811, 21821, 21822, 21825, 21831, 21834, 21842, 21851, 21861, 21862, 21863, 21864, 21865, 21866, 21871, 22003, 22009, 22022, 22023, 22670, 22671, 22672, 22673, 22674, 22675, 22676, 22677, 22678, 22679, 22680, 22681, 22702, 22704, 23098, 23099, 23100, 23101, 23102, 23104, 23105, 23106, 23107, 23108, 23109, 23110, 23111, 23112, 23113, 23114, 23115, 23116, 23117, 23118, 23119, 23120, 23121, 23122, 23123, 23124, 23125, 23126, 23127, 23128, 23129, 23130, 23323, 23324, 24508, 26410, 26414, 30002, 30010, 30012, 30020, 30022, 30030, 30032, 30040, 30042, 30050, 30052, 30062, 30072, 30080, 30082, 30090, 30102, 30122, 30132, 30142, 30152, 30162, 30180, 30182, 30192, 30202, 30212, 30222, 30240, 30242, 30250, 30252, 30262, 30270, 30272, 30280, 30282, 30290, 30292, 30300, 30302, 30502, 30512, 30522, 30532, 30620, 30630, 30650, 30670, 30700, 30720, 30730, 30740, 30750, 30760, 30770, 30790, 30800, 30830, 30840, 30850, 30870, 30880, 40008]
- 436 phenotypes have been removed: [4, 5, 6, 84, 87, 189, 399, 400, 403, 630, 699, 757, 796, 874, 884, 894, 904, 914, 1080, 1568, 1598, 1807, 1873, 1883, 2139, 2149, 2217, 2355, 2405, 2867, 2887, 2897, 2926, 2966, 3083, 3084, 3137, 3526, 3761, 3786, 3809, 4139, 4140, 4141, 4195, 4196, 4233, 4241, 4244, 4254, 4282, 4283, 4285, 4290, 4407, 4418, 4609, 4620, 4700, 5057, 5084, 5085, 5086, 5087, 5088, 5096, 5097, 5098, 5099, 5102, 5103, 5104, 5105, 5107, 5108, 5110, 5111, 5112, 5113, 5116, 5117, 5118, 5119, 5132, 5133, 5134, 5135, 5156, 5158, 5159, 5160, 5161, 5163, 5198, 5201, 5208, 5221, 5237, 5251, 5254, 5255, 5256, 5264, 5265,

[5] Age was computed by date of attending assessment centre (Data-Field 53) - birth year (Data-Field 34) and month (Data-Field 52), since date of birth (Data-Field 33) is restricted.

5276, 5292, 5375, 5386, 5993, 6022, 6038, 6039, 6349, 6350, 6351, 6383, 12338, 12654, 20006, 20019, 20021, 20022, 20074, 20128, 20132, 20134, 20135, 20136, 20137, 20138, 20154, 20191, 20240, 20247, 20248, 20400, 20420, 20433, 20434, 20442, 20455, 21021, 21611, 21622, 21625, 21634, 21642, 21661, 21662, 21665, 21666, 21836, 21838, 22004, 22005, 22024, 22025, 22026, 22033, 22034, 22037, 22038, 22039, 22040, 22507, 22700, 23321, 23322, 24003, 24004, 24005, 24006, 24007, 24008, 24010, 24011, 24012, 24016, 24017, 24018, 24019, 24020, 24021, 24022, 24023, 24024, 24500, 24501, 24502, 24503, 24504, 24505, 24506, 24507, 26411, 26412, 26413, 26415, 26416, 26417, 26427, 26428, 26429, 26430, 26431, 26432, 26433, 26434, 30000, 30060, 30070, 30092, 30100, 30110, 30112, 30120, 30130, 30140, 30150, 30160, 30172, 30190, 30200, 30210, 30220, 30232, 30260, 30600, 30601, 30610, 30611, 30621, 30631, 30640, 30641, 30651, 30660, 30661, 30671, 30680, 30681, 30690, 30691, 30701, 30710, 30711, 30721, 30731, 30741, 30751, 30761, 30771, 30780, 30781, 30791, 30801, 30810, 30811, 30820, 30821, 30831, 30841, 30851, 30860, 30861, 30871, 30881, 30890, 30891, 30897, 40005, 40009, 42014, 90010, 90011, 90012, 90013, 90019, 90020, 90021, 90022, 90023, 90024, 90025, 90027, 90028, 90029, 90030, 90031, 90032, 90033, 90034, 90035, 90036, 90037, 90038, 90039, 90040, 90041, 90042, 90043, 90044, 90045, 90046, 90047, 90048, 90049, 90050, 90051, 90052, 90053, 90054, 90055, 90056, 90057, 90058, 90059, 90060, 90061, 90062, 90063, 90064, 90065, 90066, 90067, 90068, 90069, 90070, 90071, 90072, 90073, 90074, 90075, 90076, 90077, 90078, 90079, 90080, 90081, 90082, 90083, 90086, 90087, 90088, 90089, 90091, 90092, 90093, 90094, 90095, 90096, 90097, 90098, 90099, 90100, 90101, 90102, 90103, 90104, 90105, 90106, 90107, 90108, 90109, 90110, 90111, 90112, 90113, 90114, 90115, 90116, 90117, 90118, 90119, 90120, 90121, 90122, 90123, 90124, 90125, 90126, 90127, 90128, 90129, 90130, 90131, 90132, 90133, 90134, 90135, 90136, 90137, 90138, 90139, 90140, 90141, 90142, 90143, 90144, 90145, 90146, 90159, 90160, 90161, 90162, 90163, 90164, 90165, 90166, 90167, 90168, 90169, 90170, 90171, 90172, 90173, 90174, 90175, 90176, 90177, 90179, 90182, 90183, 90184, 90185, 90186, 90187, 90188, 90189, 90190, 90191, 90192, 90193, 90194, 90195, 110006]

S3. Details about advanced meta-matching (finetune)

In this section, we provide implementation details of advanced meta-matching (finetune). The trained DNN (previous section) was applied to the K participants in the test meta-set. For a given test meta-set phenotype,

- The best DNN output that gave the best prediction for the test phenotype (based on the K participants) was selected
- We took the trained DNN and removed all output nodes except the best DNN output node (selected in the previous step). We then performed finetuning on this DNN using the K participants. The loss function was MSE (mean squared error) loss. The evaluation metric was COD.
- Finetuning was only performed on the weights of the last two layers. The weights of the earlier layers were frozen. We split the K subjects into training and validation sets (4:1 ratio). We ran the finetuning for 100 epochs using the training set and checked the performance in the validation set every 10 epochs. The DNN from the epoch with the best performance in the validation set was used for predicting the phenotype in the remaining 10,000 – K participants. If the performance in the validation set was worse than the original DNN (without finetuning), then we simply applied the original DNN to the remaining 10,000 – K participants. We did not perform cross-validation like the classical (KRR) baseline, because the runtime would be increased multiple folds.
- Furthermore, because of the small number of participants K, we decided not to optimize the hyperparameters of the finetuning procedure for fear of overfitting. Optimizing the hyperparameters would also be computationally too expensive. More specifically, it took 6 days (on one GPU) to run 34 meta-set phenotypes for 100 repetition of K-shots across different values of K. Optimizing the hyperparameters using HORD would dramatically increase the runtime to 6 x 200 = 1200 days (since we utilized 200 HORD rounds).
- Therefore, we simply set the hyperparameters to the following generic values: stochastic gradient descent (SGD) with 0.9 momentum. The learning rate was set to be 1e-3. The batch size was set to be the minimum of K and 32. So if K was less than 32, the batch size was set to be K. Otherwise, the batch size was set to be 32.

Supplementary Tables

Label	Description
ECG C1	ECG measures principal component 1
Sex	sex
Sex G C2	genotype sex inference principal component 2
Body C2	anthropometry principal component 2
Grip C1	hand grip strength principal component 1
Body C1	anthropometry principal component 1
Bone C3	bone-densitometry of heel principal component 3
BP eye C4	blood pressure & eye measures principal component 4
Matrix C1	matrix pattern completion principal component 1
#Mem C1	numeric memory principal component 1
Matrix C2	matrix pattern completion principal component 2
Fluid Int.	fluid intelligence
Hearing	hearing signal-to-noise-ratio (snr) of triplet (left)
Illness C1	non-cancer illness principal component 1
#household	number of people in household
Time TV	time spent watching television (tv) per day
BP eye C2	blood pressure & eye measures component 2
Body C3	anthropometry principal component 3
ECG C6	ECG measures principal component 6
ECG C2	ECG measures principal component 2
Illness C4	non-cancer illness principal component 4
Smoke C1	smoke principal component 1
BP eye C3	blood pressure & eye measures principal component 3
BP eye C6	blood pressure & eye measures principal component 6
Urine C1	urine assays principal component 1
Sex G C1	genotype sex inference principal component 1
Bone C1	bone-densitometry of heel principal component 1
Matrix C3	matrix pattern completion principal component 3
Time walk	number of days walked 10+ minutes per week
BP eye C5	blood pressure & eye measures principal component 5
ECG C3	ecg measures principal component 3
Genetic C1	genetic principal components and heterozygosity principal component 1
Sleep	sleep duration per day

Table S1. Dictionary of 33 training meta-set non-brain-imaging phenotypes. For UK Biobank IDs, please see GITHUB_LINK.

Label	Description
Alcohol 3	average weekly beer plus cider intake
Blood C2	blood assays principal component 2
Breath C1	spirometry principal component 1
Age	age
Cancer C1	cancer principal component 1
Carotid C1	carotid ultrasound principal component 1
Match-o	pairs matching online
Trail C1	trail making principal component 1
Digit-o C1	symbol digit substitution online principal component 1
Digit 1	symbol digit substitution principal component 1
Match	pairs matching
ProMem C1	prospective memory principal component 1
RT C1	reaction time principal component 1
Trail-o C1	trail making online principal component 1
Tower C1	tower rearranging principal component 1
Family C1	family history (parent's age) principal component 1
Blood C5	blood assays principal component 5
Dur C4	process durations principal component 4
Dur C2	process durations principal component 2
Loc C1	location principal component 1
Dur C1	process durations principal component 1
Digit-o C6	symbol digit substitution online principal component 6
Trail-o C4	trail making online principal component 4
Blood C4	blood assays principal component 4
Alcohol 2	average weekly champagne plus white wine intake
Carotid C5	carotid ultrasound principal component 5
Time drive	time spent driving per day
Travel	frequency of travelling from home to job workplace per week
Work	weekly length of working hour for main job
Age edu	age completed full time education
Deprive C1	multiple deprivation principal component 1
Blood C3	blood assays principal component 3
Alcohol 1	average monthly spirits intake
Neuro	neuroticism score

Table S2. Dictionary of 34 test meta-set non-brain-imaging phenotypes. For UK Biobank IDs, please see GITHUB_LINK.

Description	HCP field
Visual Episodic Memory	PicSeq_Unadj
Cognitive Flexibility (DCCS)	CardSort_Unadj
Inhibition (Flanker Task)	Flanker_Unadj
Fluid Intelligence (PMAT)	PMAT24_A_CR
Vocabulary (Pronunciation)	ReadEng_Unadj
Vocabulary (Picture Matching)	PicVocab_Unadj
Processing Speed	ProcSpeed_Unadj
Delay Discounting	DDic_AUC_40K
Spatial Orientation	VSPLOT_TC
Sustained Attention – Spec.	SCPT_SPEC
Working Memory (List Sorting)	ListSort_Unadj
Cognitive Status (MMSE)	MMSE_Score
Sleep Quality (PSQI)	PSQI_Score
Walking Endurance	Endurance_Unadj
Walking Speed	GaitSpeed_Unadj
Manual Dexterity	Dexterity_Unadj
Grip Strength	Strength_Unadj
Taste Intensity	Taste_Unadj
Emotional Face Matching	Emotion_Task_Face_Acc
Arithmetic	Language_Task_Math_Avg_Difficulty_Level
Story Comprehension	Language_Task_Story_Avg_Difficulty_Level
Relational Processing	Relational_Task_Acc
Working Memory (N-back)	WM_Task_Acc
Agreeableness (NEO)	NEOFAC_A
Openness (NEO)	NEOFAC_O
Conscientiousness (NEO)	NEOFAC_C
Extraversion (NEO)	NEOFAC_E
Anger – Aggression	AngAggr_Unadj
Fear – Affect	FearAffect_Unadj
Sadness	Sadness_Unadj
Life Satisfaction	LifeSatisf_Unadj
Meaning & Purpose	MeanPurp_Unadj
Loneliness	Loneliness_Unadj
Perceived Stress	PercStress_Unadj
Self-Efficacy	SelfEff_Unadj

Table S3. Dictionary of 35 HCP non-brain-imaging phenotypes and corresponding descriptive labels used in the manuscript.

Supplementary Figures

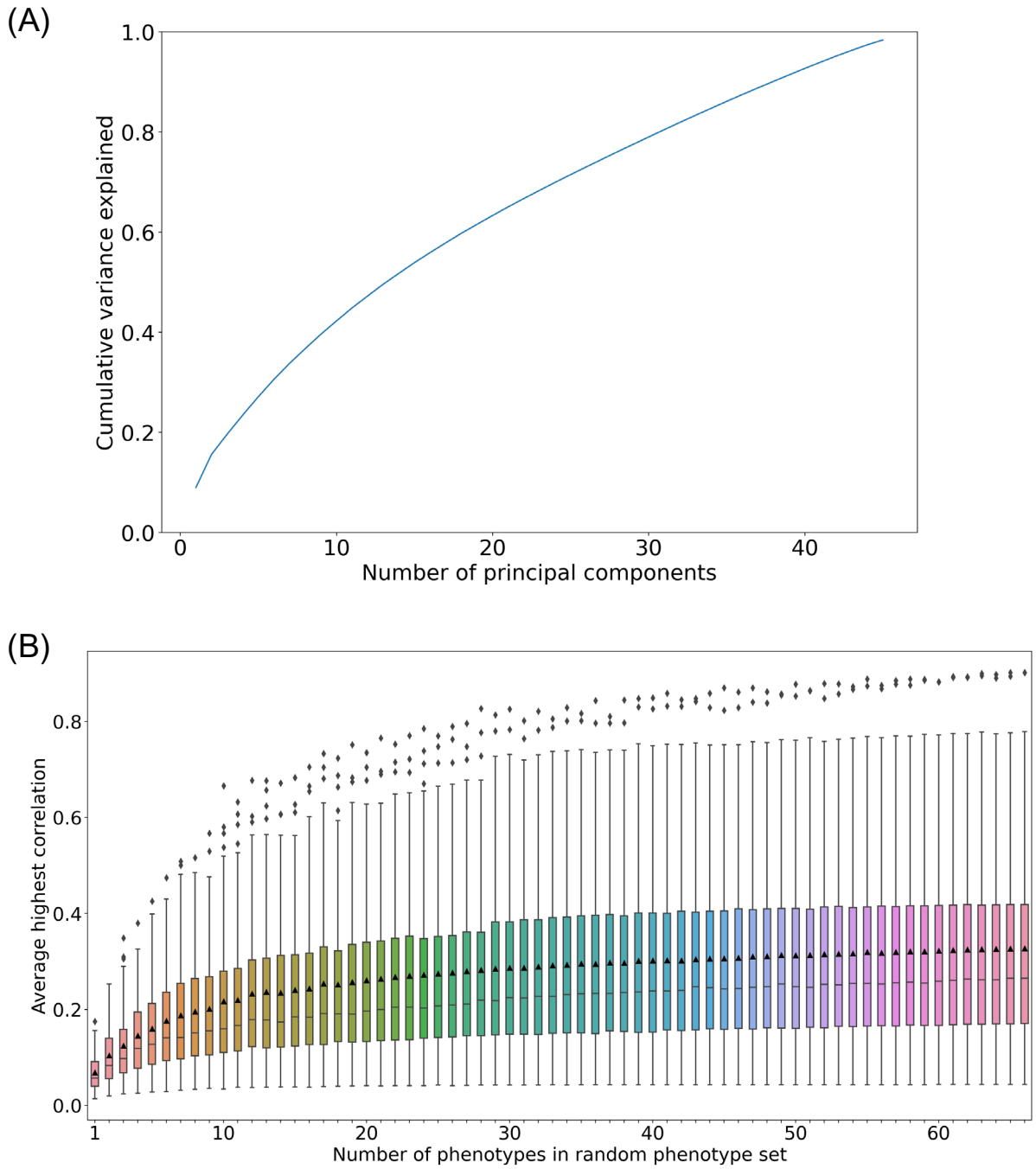


Figure S1. Quantifying low dimensionality of 67 UK Biobank non-brain-imaging phenotypes. (A) Top 14 principal components were sufficient to explain 50% of the variance among 67 UK Biobank non-brain-imaging phenotypes (N = 36,848). We applied principal component analysis to 67 non-brain-imaging phenotypes. Horizontal axis is the number of principal components. Vertical axis is the cumulative variance explained. (B) Maximum absolute correlation between UK Biobank phenotypes and randomly selected sets of phenotypes (N = 36,848). For each of 67 UK Biobank phenotypes, we randomly selected N phenotypes from remaining 66 phenotypes. Maximum absolute correlation with the N random phenotypes were computed. This procedure was repeated 100 times. Y axis is the

maximum absolute correlation averaged across the 100 repetitions. For each boxplot, the horizontal line indicates the median and the black triangle indicates the mean. The bottom and top edges of the box indicate the 25th and 75th percentiles respectively. Whiskers correspond to 1.5 times the interquartile range. Outliers are defined as data points beyond 1.5 times the interquartile range. The maximum correlation increased with more phenotypes in the random phenotype set, but the improvement tapers off at around 20 phenotypes. The very wide quantiles in the boxplots suggest that certain phenotypes were much strong correlated with other phenotypes.

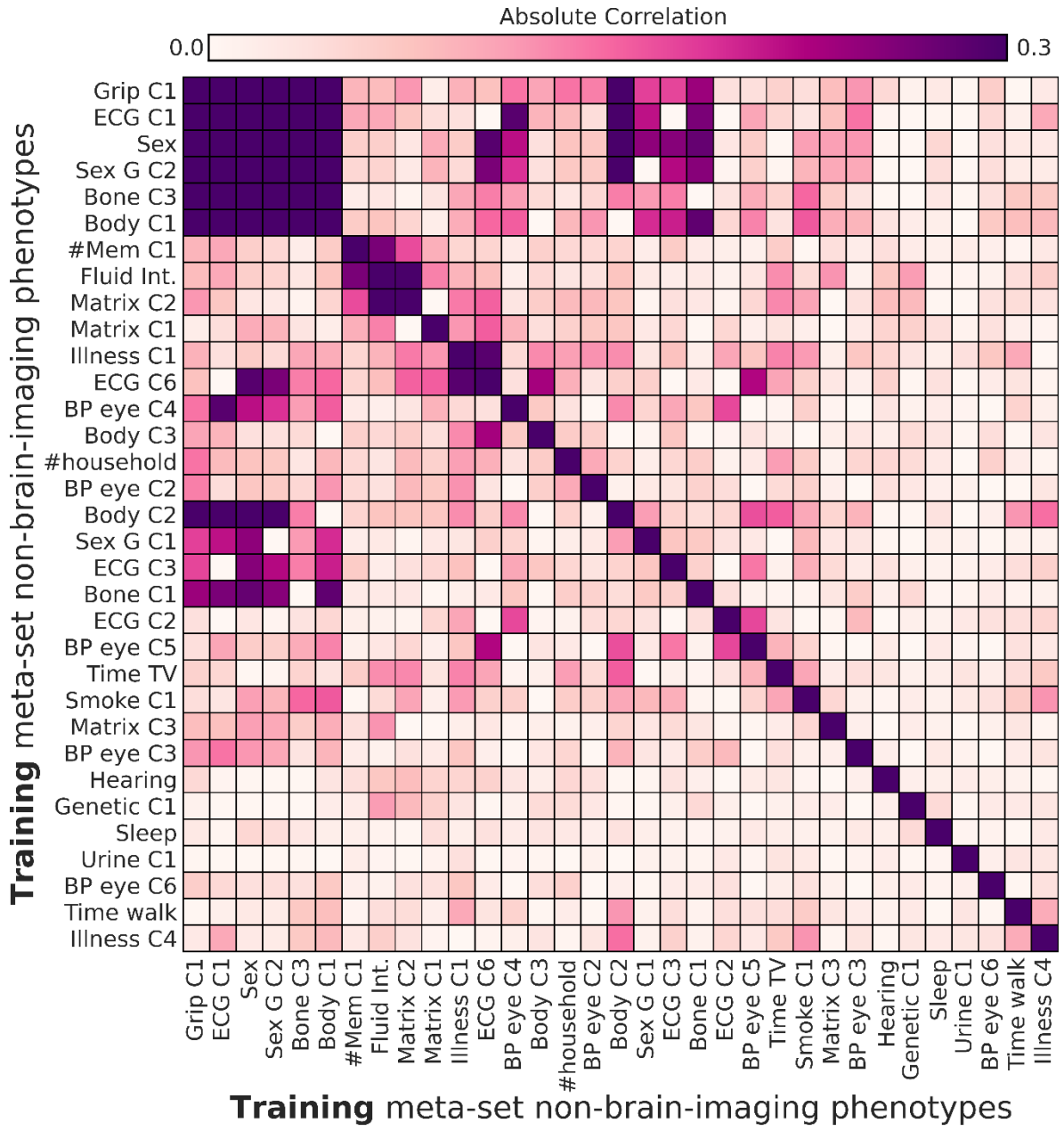


Figure S2. Absolute Pearson's correlation among 33 non-brain-imaging phenotypes in the training meta-set in the UK Biobank.

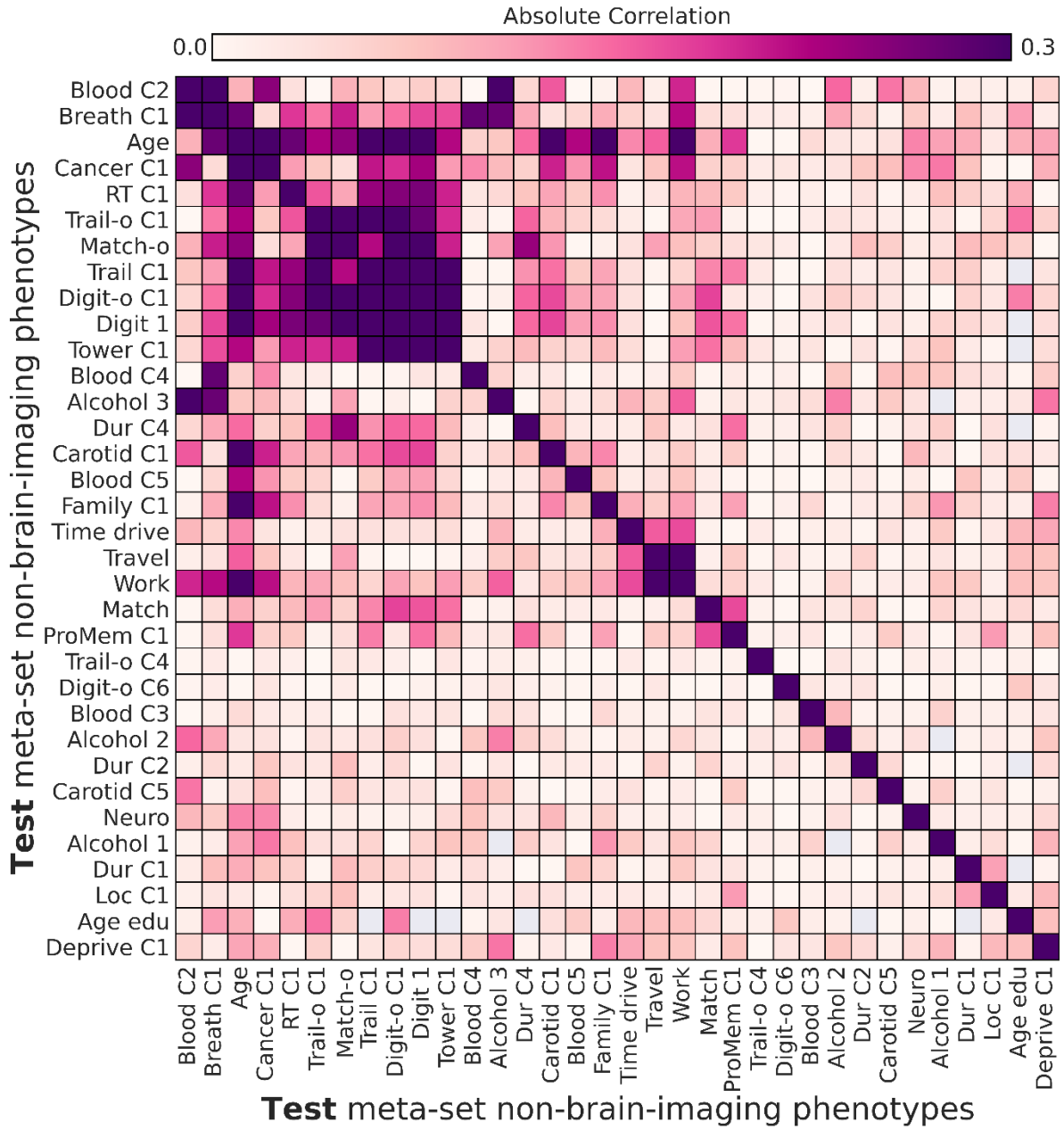
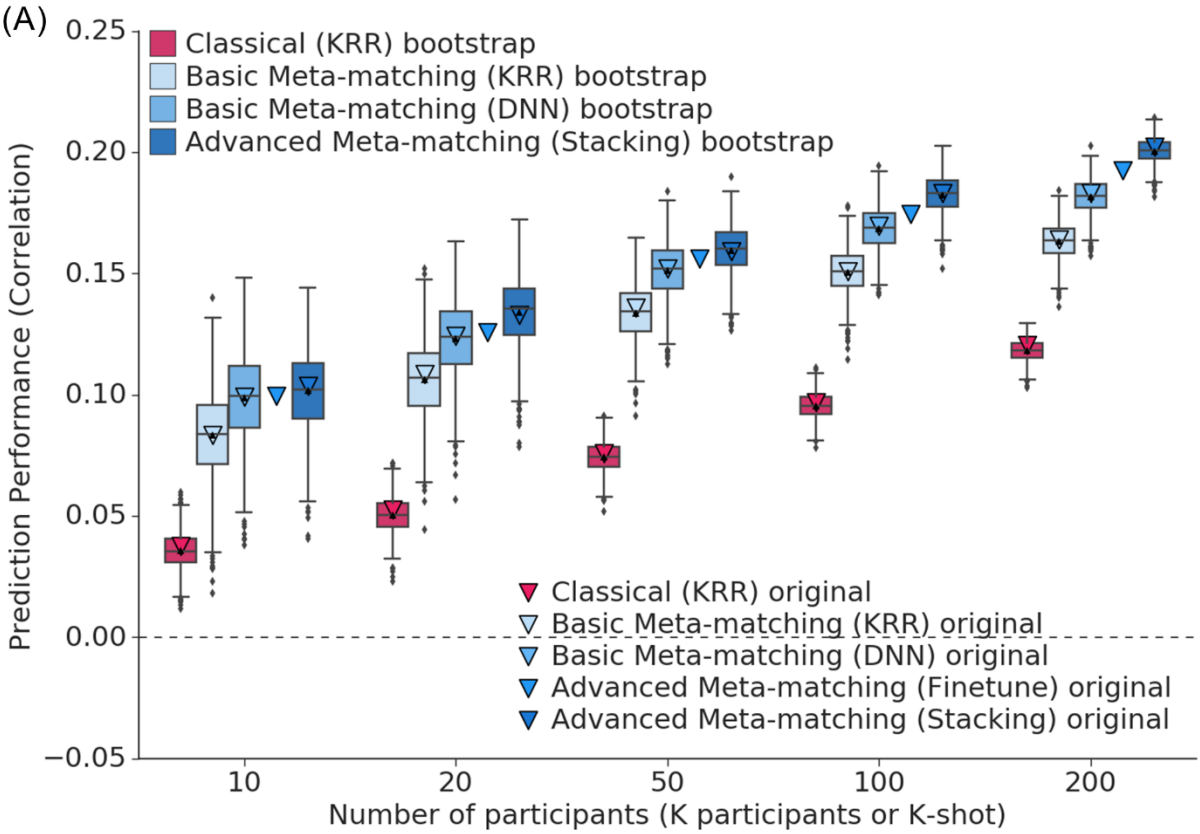


Figure S3. Absolute Pearson's correlation among 34 non-brain-imaging phenotypes in the test meta-set in the UK Biobank.

(A)



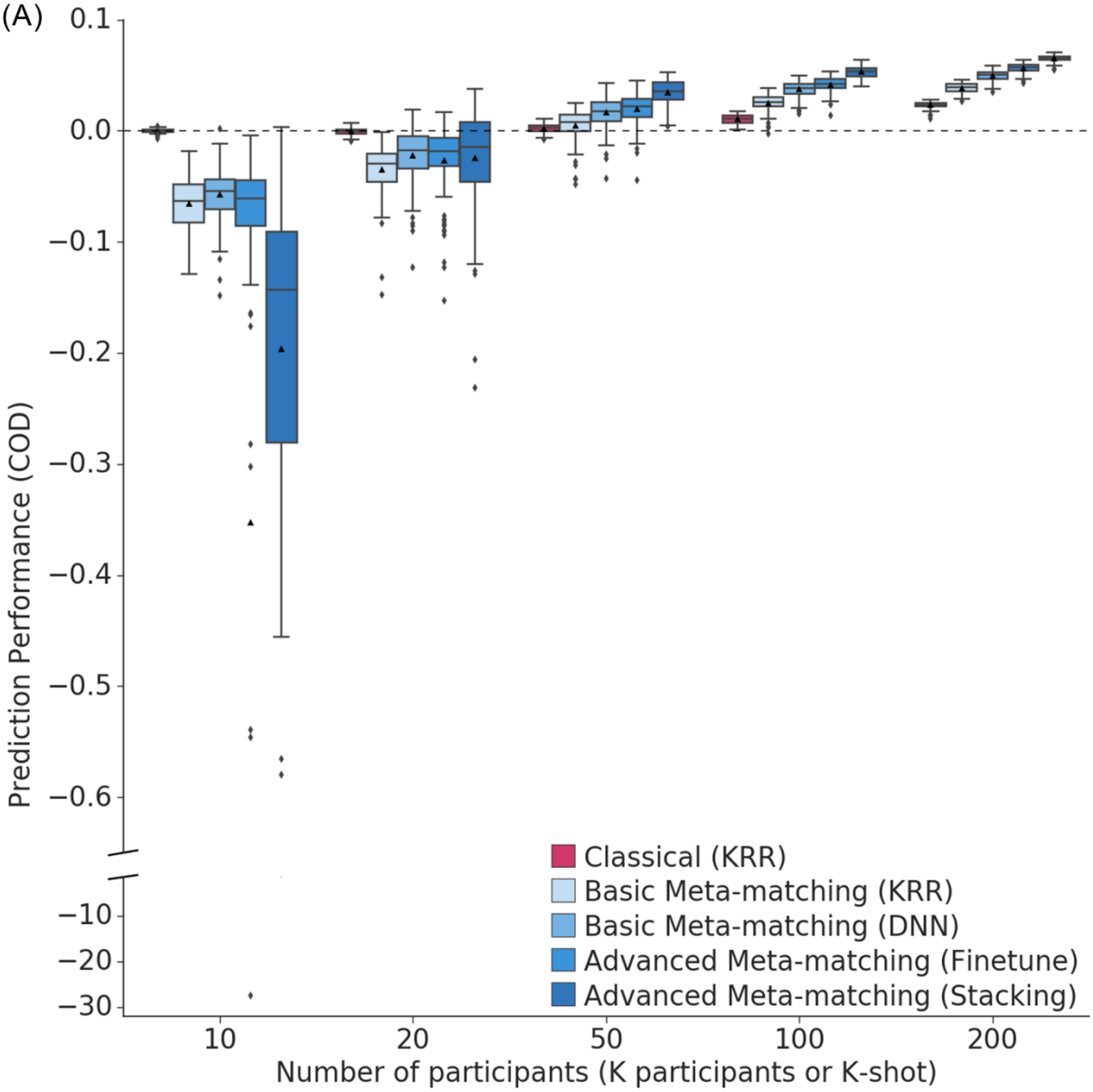
(B)

K Shot	10	20	50	100	200
Classical (KRR) vs Basic Meta-matching (KRR)	**0.0135**	**0.0005**	**3.8E-07**	**2.5E-09**	**3.5E-09**
Classical (KRR) vs Basic Meta-matching (DNN)	**0.0014**	**1.3E-05**	**3.6E-10**	**1.5E-15**	**8.0E-18**
Classical (KRR) vs Advanced Meta-matching (Finetune)	**≈0**	**≈0**	**≈0**	**≈0**	**≈0**
Classical (KRR) vs Advanced Meta-matching (Stacking)	**0.0002**	**2.1E-08**	**5.3E-17**	**4.9E-30**	**7.8E-61**
Basic Meta-matching (KRR) vs Basic Meta-matching (DNN)	0.4212	0.3651	0.2003	0.0512	**0.0143**
Basic Meta-matching (KRR) vs Advanced Meta-matching (Finetune)	0.3891	0.1979	0.0495	**0.0084**	**5.6E-05**
Basic Meta-matching (KRR) vs Advanced Meta-matching (Stacking)	0.2947	0.0832	0.0269	**0.0003**	**8.7E-08**
Basic Meta-matching (DNN) vs Advanced Meta-matching (Finetune)	0.9712	0.8671	0.6743	0.5136	0.1245
Basic Meta-matching (DNN) vs Advanced Meta-matching (Stacking)	0.8873	0.5552	0.5110	0.1067	**0.0039**
Advanced Meta-matching (Finetune) vs Advanced Meta-matching (Stacking)	0.8989	0.5740	0.7273	0.2850	0.0997

Figure S4. Meta-matching outperformed classical kernel ridge regression (KRR) baseline in the UK Biobank (N = 10,000 – K). (A) Prediction performance (Pearson's correlation) with different number of participants. This plot is the same as Figure 4A, but the boxplots now show the bootstrap distribution of each approach based on 1000 bootstrapped samples. The triangles show the average performance (Pearson's correlation) of 34 non-brain-imaging phenotypes using the original 100 random repeats (Figure 4A). We observe that the mean of the bootstrap distributions matches the mean of the original experiments (Figure 4A) quite well. Bootstrapping could not be performed for advanced meta-matching (finetune) because 1000 bootstrap samples would have required 60 days of compute time. For each boxplot, the horizontal line indicates the median and the black triangle indicates the mean. The bottom and top edges of the box indicate the 25th and 75th percentiles respectively. Whiskers correspond to 1.5 times the interquartile range. Outliers are defined as data points beyond 1.5 times the interquartile range. (B) Statistical differences among the different algorithms. P values were calculated based on a two-sided bootstrapping procedure (see Methods). For rows comparing advanced meta-matching (finetune) and another

algorithm X, p values were derived by comparing the mean of advanced meta-matching (finetune) with algorithm X's bootstrap distribution (assuming Gaussanity). For other rows comparing algorithms X and Y, bootstrap distributions were available for both X and Y. Therefore, one p value was obtained by comparing the original mean of X with Y's bootstrap distribution and another p value was obtained by comparing the original mean of Y with X's bootstrap distribution. The larger of the two p values were reported. Bold indicates statistical significance after FDR correction ($q < 0.05$).

(A)



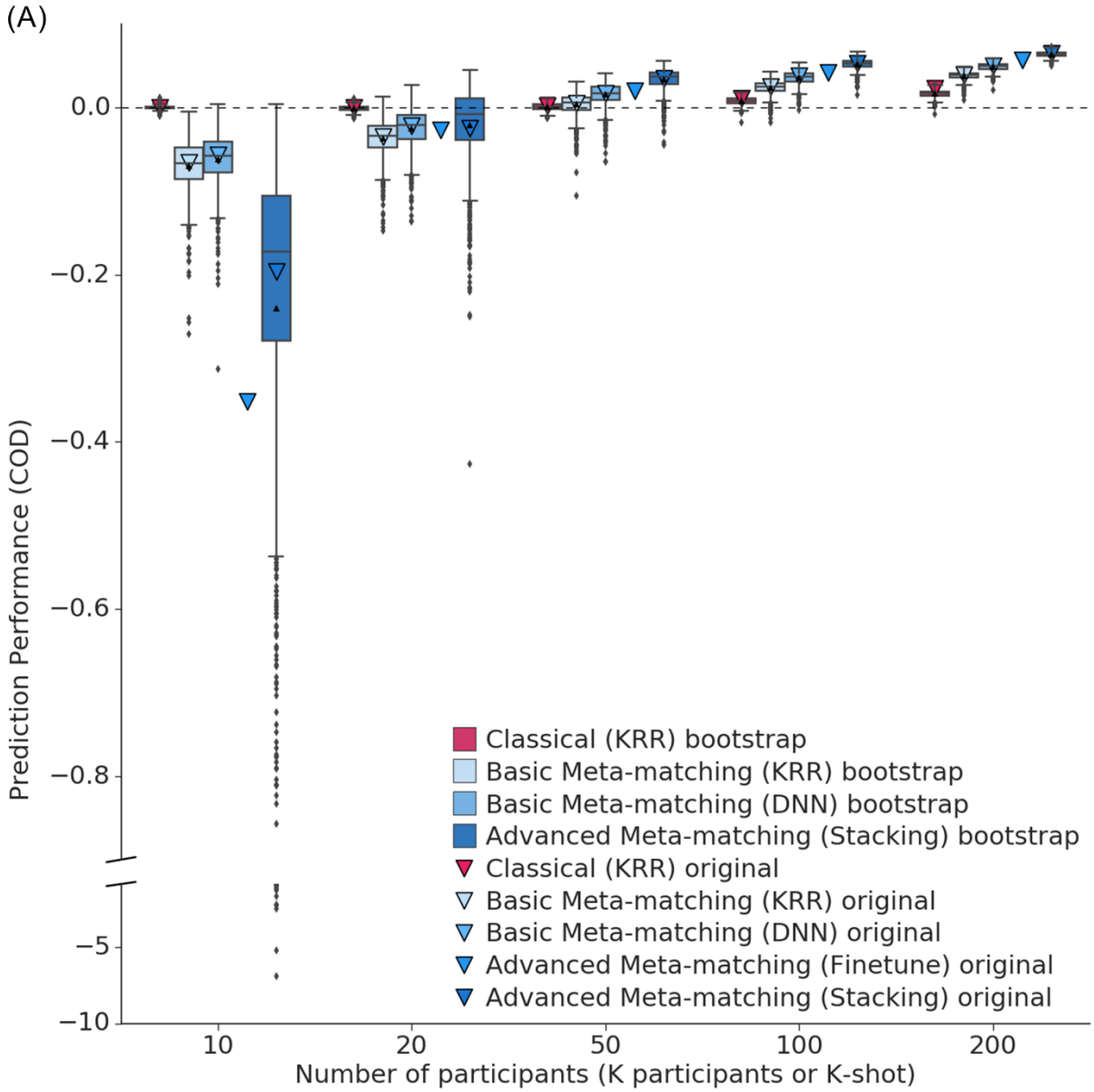
(B)

K Shot	10	20	50	100	200
Classical (KRR) vs Basic Meta-matching (KRR)	*	n.s.	n.s.	n.s.	*
Classical (KRR) vs Basic Meta-matching (DNN)	n.s.	n.s.	n.s.	*	***
Classical (KRR) vs Advanced Meta-matching (Finetune)	***	***	**	***	***
Classical (KRR) vs Advanced Meta-matching (Stacking)	n.s.	n.s.	*	***	***

Figure S5. Meta-matching outperformed classical kernel ridge regression (KRR) baseline in the UK Biobank. (A) Prediction performance (coefficient of determination; COD) averaged across 34 non-brain-imaging phenotypes in the test meta-set (N = 10,000 –

K). The K participants were used to train and tune the models (Figure 3). Boxplots represent variability across 100 random repeats of K participants (Figure 2A). For each boxplot, the horizontal line indicates the median and the black triangle indicates the mean. The bottom and top edges of the box indicate the 25th and 75th percentiles respectively. Whiskers correspond to 1.5 times the interquartile range. Outliers are defined as data points beyond 1.5 times the interquartile range. (B) Statistical difference between the prediction performance (COD) of classical (KRR) baseline and meta-matching algorithms. P values were calculated based on a two-sided bootstrapping procedure (see Methods). “n.s.” indicates that difference was not statistically significant after multiple comparisons correction (FDR $q < 0.05$). "*" indicates $p < 0.05$ and statistical significance after multiple comparisons correction (FDR $q < 0.05$). "**" indicates $p < 0.001$ and statistical significance after multiple comparisons correction (FDR $q < 0.05$). "***" indicates $p < 0.00001$ and statistical significance after multiple comparisons correction (FDR $q < 0.05$). Green indicates that meta-matching outperforms classical (KRR) baseline. Red indicates that classical (KRR) baseline outperforms meta-matching. Observe that all algorithms performed poorly ($COD \leq 0$) when there were less than 50 participants ($K < 50$), suggesting chance or worse than chance prediction for all algorithms. The actual p values and statistical comparisons among all algorithms are found in Figure S5.

(A)



(B)

K Shot	10	20	50	100	200
Classical (KRR) vs Basic Meta-matching (KRR)	**0.0209**	0.1016	0.9113	0.1064	**0.0026**
Classical (KRR) vs Basic Meta-matching (DNN)	0.0481	0.2986	0.3349	**0.0014**	**4.8E-08**
Classical (KRR) vs Advanced Meta-matching (Finetune)	**≈0**	**7.3E-17**	**1.3E-05**	**8.4E-15**	**≈0**
Classical (KRR) vs Advanced Meta-matching (Stacking)	0.4811	0.6593	**0.0116**	**2.6E-12**	**4.1E-41**
Basic Meta-matching (KRR) vs Basic Meta-matching (DNN)	0.8895	0.6694	0.4278	0.1881	0.0345
Basic Meta-matching (KRR) vs Advanced Meta-matching (Finetune)	**3.7E-21**	0.6371	0.2350	0.0330	**0.0001**
Basic Meta-matching (KRR) vs Advanced Meta-matching (Stacking)	0.6083	0.7706	**0.0201**	**0.0004**	**3.3E-08**
Basic Meta-matching (DNN) vs Advanced Meta-matching (Finetune)	**5.0E-21**	0.9451	0.7705	0.4471	0.0881
Basic Meta-matching (DNN) vs Advanced Meta-matching (Stacking)	0.5910	0.9913	0.1692	0.0244	**0.0005**
Advanced Meta-matching (Finetune) vs Advanced Meta-matching (Stacking)	0.7426	0.9102	0.2605	0.0740	**0.0198**

Figure S6. Meta-matching outperformed classical kernel ridge regression (KRR) baseline in the UK Biobank (N = 10,000 – K). (A) Prediction performance (coefficient of determination; COD) with different number of participants. This plot is the same as Figure S4A, but the boxplots now show the bootstrap distribution of each approach based on 1000 bootstrapped samples. The triangles show the average performance (COD) of 34 non-brain-imaging phenotypes using the original 100 random repeats (Figure S4A). We observe that the

mean of the bootstrap distributions matches the mean of the original experiments (Figure S4A) quite well. Bootstrapping could not be performed for advanced meta-matching (finetune) because 1000 bootstrap samples would have required 60 days of compute time. For each boxplot, the horizontal line indicates the median and the black triangle indicates the mean. The bottom and top edges of the box indicate the 25th and 75th percentiles respectively. Whiskers correspond to 1.5 times the interquartile range. Outliers are defined as data points beyond 1.5 times the interquartile range. (B) Statistical differences among the different algorithms. P values were calculated based on a two-sided bootstrapping procedure (see Methods). For rows comparing advanced meta-matching (finetune) and another algorithm X, p values were derived by comparing the mean of advanced meta-matching (finetune) with algorithm X's bootstrap distribution (assuming Gaussanity). For other rows comparing algorithms X and Y, bootstrap distributions were available for both X and Y. Therefore, one p value was obtained by comparing the original mean of X with Y's bootstrap distribution and another p value was obtained by comparing the original mean of Y with X's bootstrap distribution. The larger of the two p values were reported. Bold indicates statistical significance after FDR correction ($q < 0.05$).

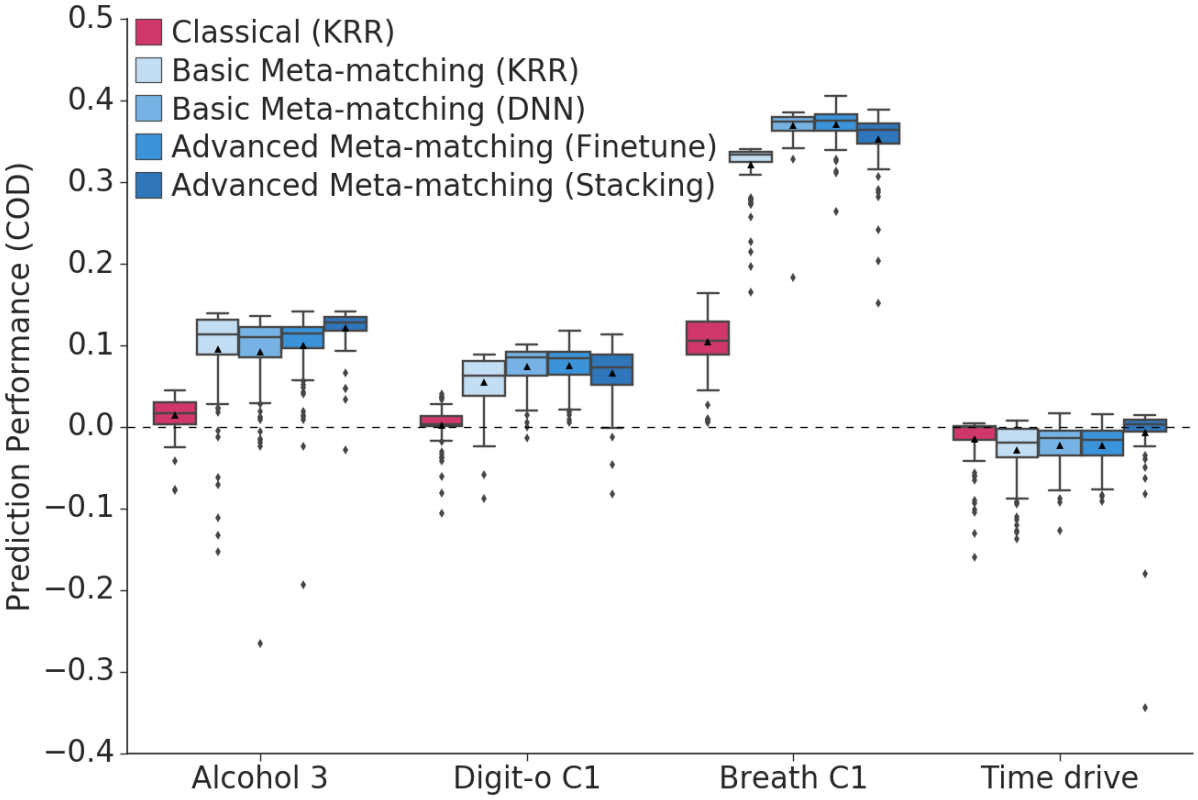


Figure S7. Examples of non-brain-imaging phenotypic prediction performance in the test meta-set in the case of 100-shot learning in the UK Biobank (N = 9,900). Here, prediction performance was measured using coefficient of determination (COD). "Alcohol 3" (average weekly beer plus cider intake) was most frequently matched to "Bone C3" (bone-densitometry of heel principal component 3). "Digit-o C1" (symbol digit substitution online principal component 1) was most frequently matched to "Matrix C1" (matrix pattern completion principal component 1). "Breath C1" (spirometry principal component 1) was most frequently matched to "Grip C1" (hand grip strength principal component 1). "Time drive" (Time spent driving per day) was most frequently matched to "BP eye C3" (blood pressure & eye measures principal component 3). For each boxplot, the horizontal line indicates the median and the black triangle indicates the mean. The bottom and top edges of the box indicate the 25th and 75th percentiles respectively. Whiskers correspond to 1.5 times the interquartile range. Outliers are defined as data points beyond 1.5 times the interquartile range.

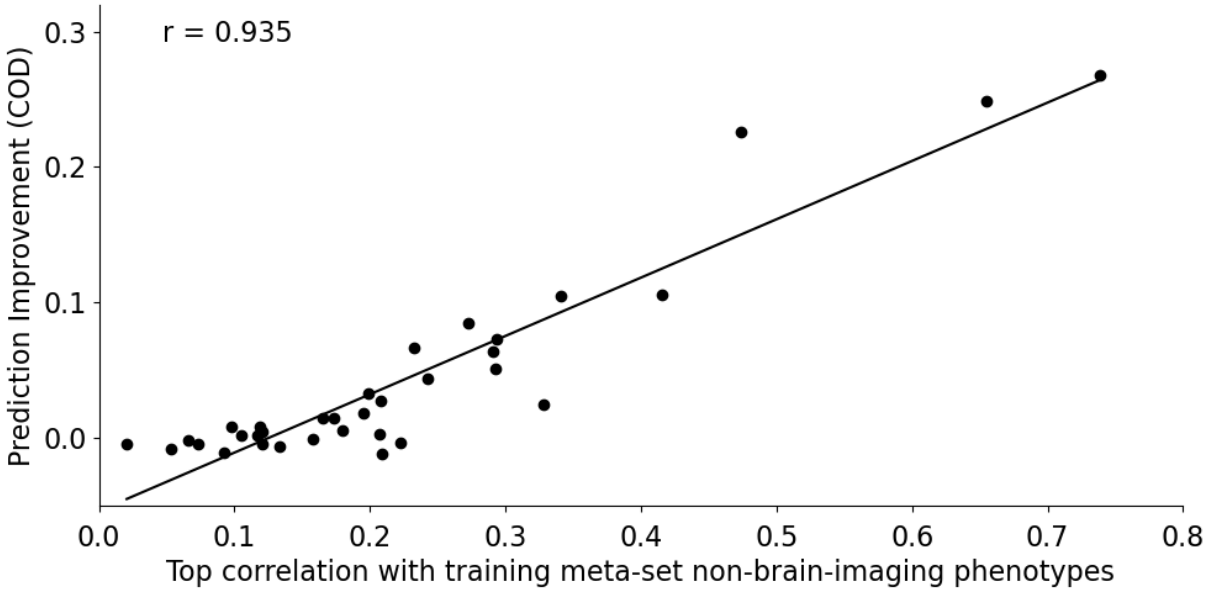


Figure S8. Prediction improvements were driven by correlations between training and test meta-set phenotypes in the UK Biobank. Vertical axis shows the prediction improvement of advanced meta-matching (stacking) with respect to classical (KRR) baseline under the 100-shot scenario. Prediction performance was measured using coefficient of determination (COD). Each dot represents a test meta-set phenotype. Horizontal axis shows each test phenotype's top absolute Pearson's correlation with training phentoypes computed using participants from the test meta-set. Test phenotypes with stronger correlations with at least one training phenotype led to greater prediction improvement with meta-matching.

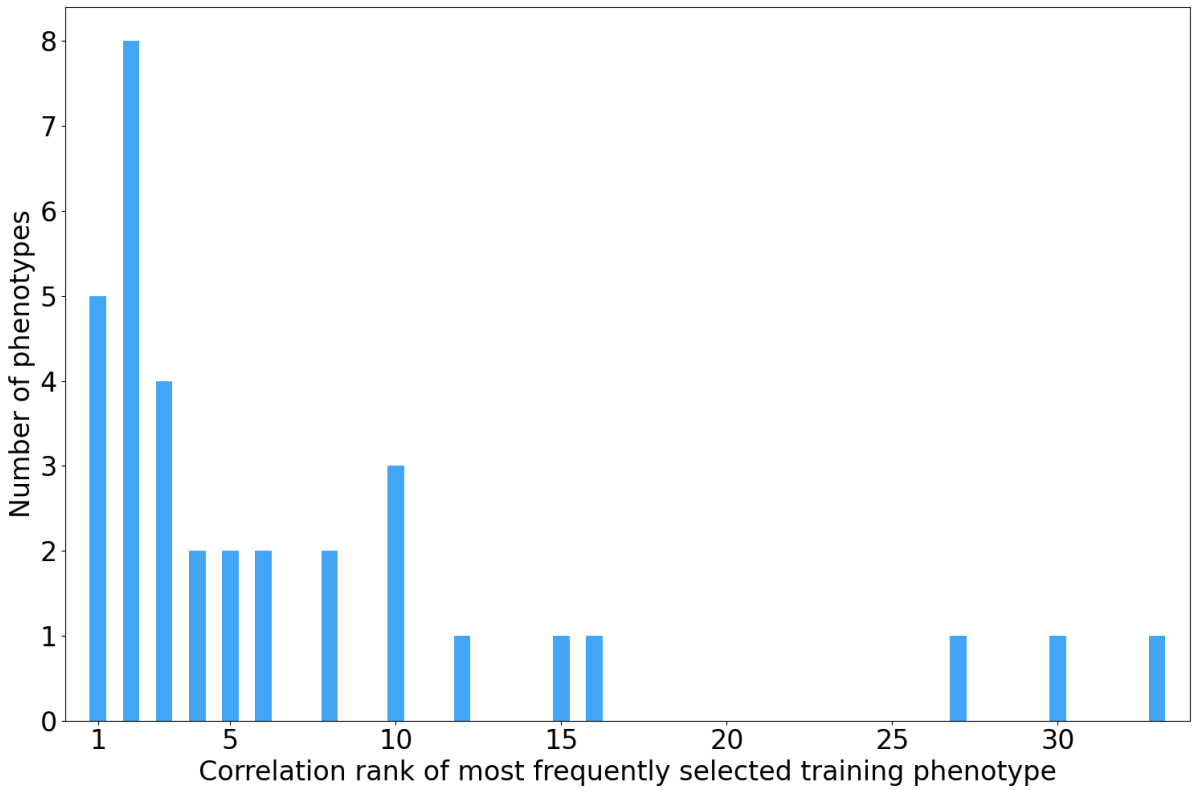


Figure S9. For most test meta-set phenotypes, basic meta-matching (DNN) was able to select training phenotypes most strongly correlated with the test phenotypes. For each

test phenotype, we considered the training phenotype most frequently selected by basic meta-matching (DNN) in the 100-shot scenario. Horizontal axis is the rank of correlation between the test phenotype and most frequently selected training phenotype out of all the correlations between the test phenotype and all training phenotypes. Here, correlations were computed using participants from the test meta-set. Vertical axis shows the number of test phenotypes. For example, the figure shows that for 8 test phenotypes, the most frequently selected training phenotype (out of 100 repetitions in the 100-shot scenario) was the 2nd most correlated training phenotype.

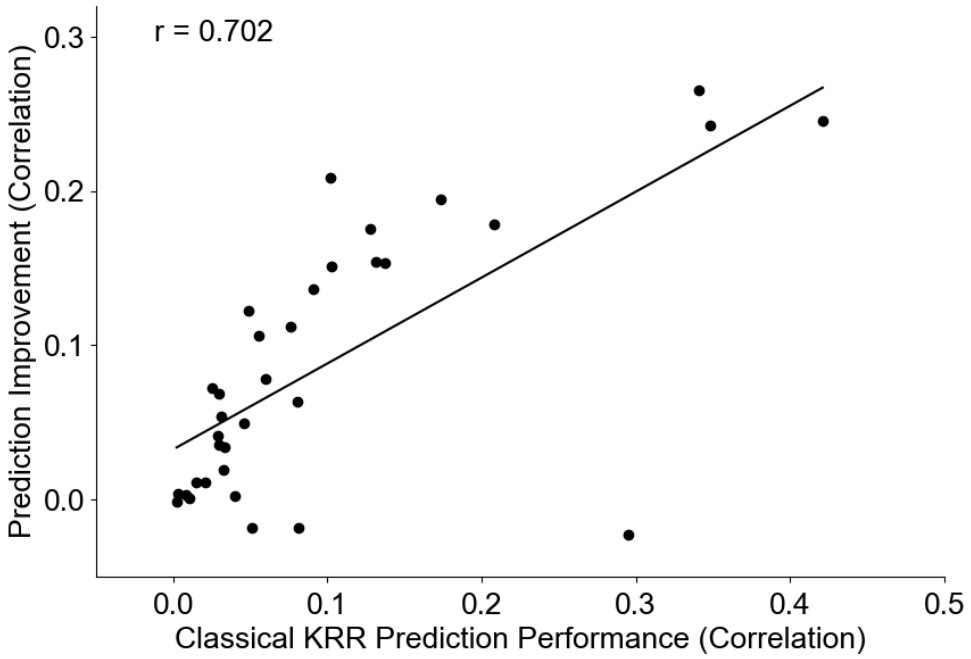


Figure S10. Phenotypes better predicted by classical kernel ridge regression benefited more from meta-matching in the UK Biobank. Vertical axis shows the prediction improvement of advanced meta-matching (stacking) with respect to classical (KRR) baseline under the 100-shot scenario. Prediction performance was measured using Pearson's correlation. Each dot represents a test meta-set phenotype. Horizontal axis shows the prediction performance with the classical (KRR) baseline under the 100-shot scenario. Similar conclusions were obtained with coefficient of determination (Figure S9).

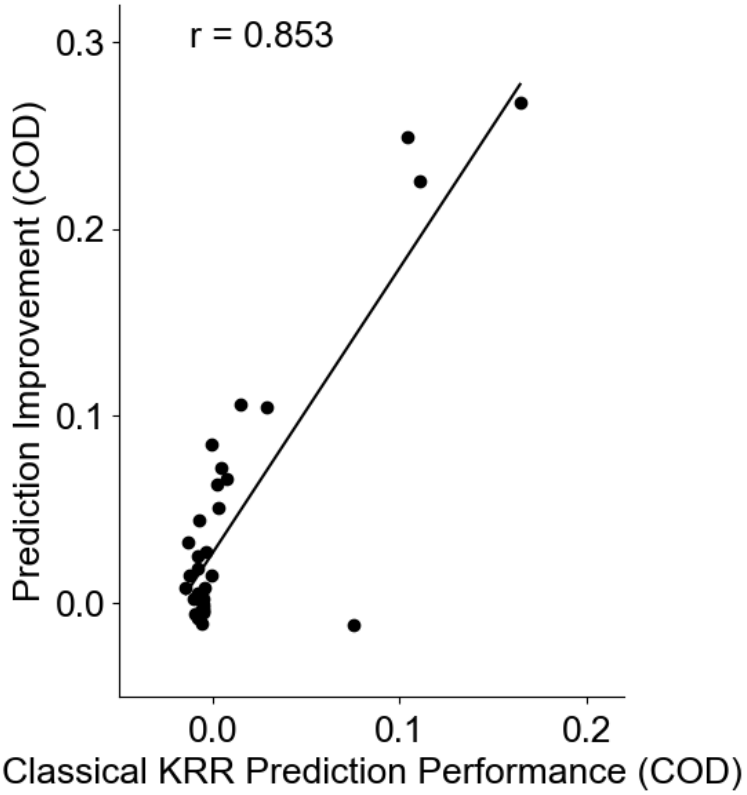
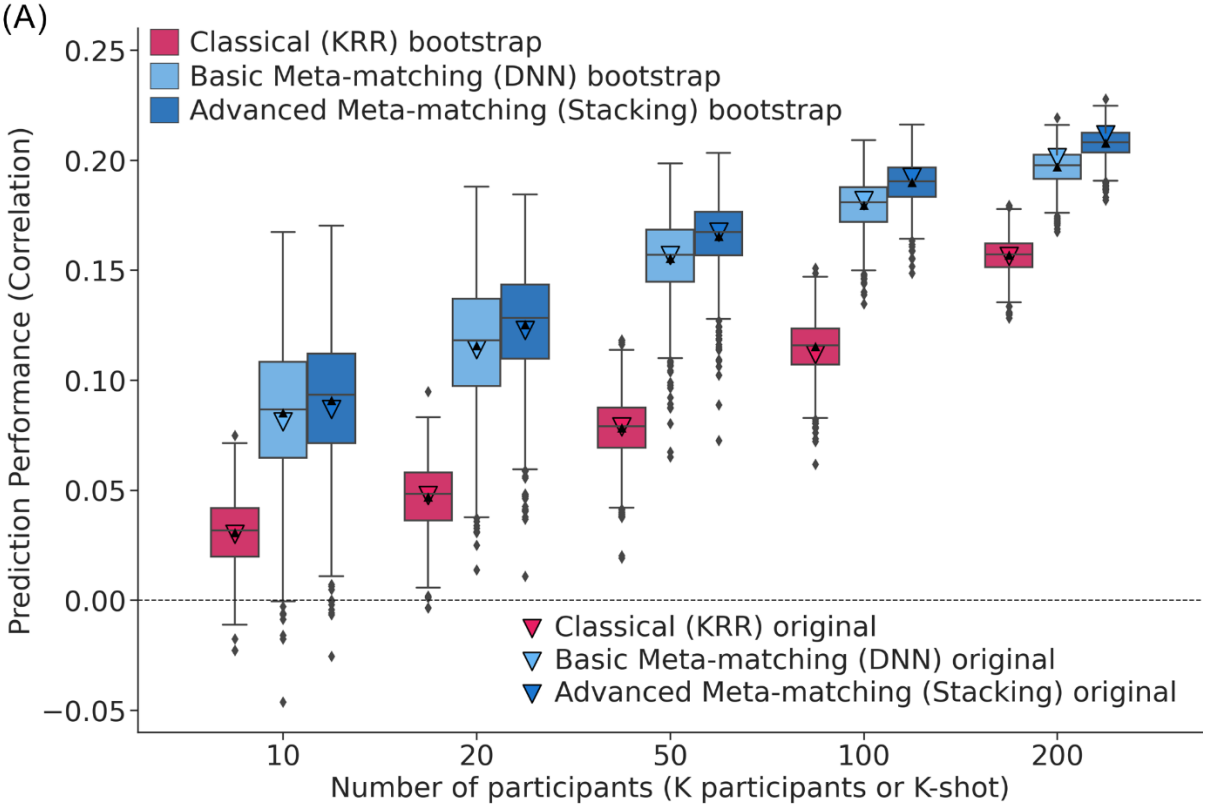


Figure S11. Phenotypes better predicted by classical kernel ridge regression benefited more from meta-matching in the UK Biobank. Vertical axis shows the prediction improvement of advanced meta-matching (stacking) with respect to classical (KRR) baseline under the 100-shot scenario. Prediction performance was measured using coefficient of determination (COD). Each dot represents a test meta-set phenotype. Horizontal axis shows the prediction performance with the classical (KRR) baseline under the 100-shot scenario.

(A)

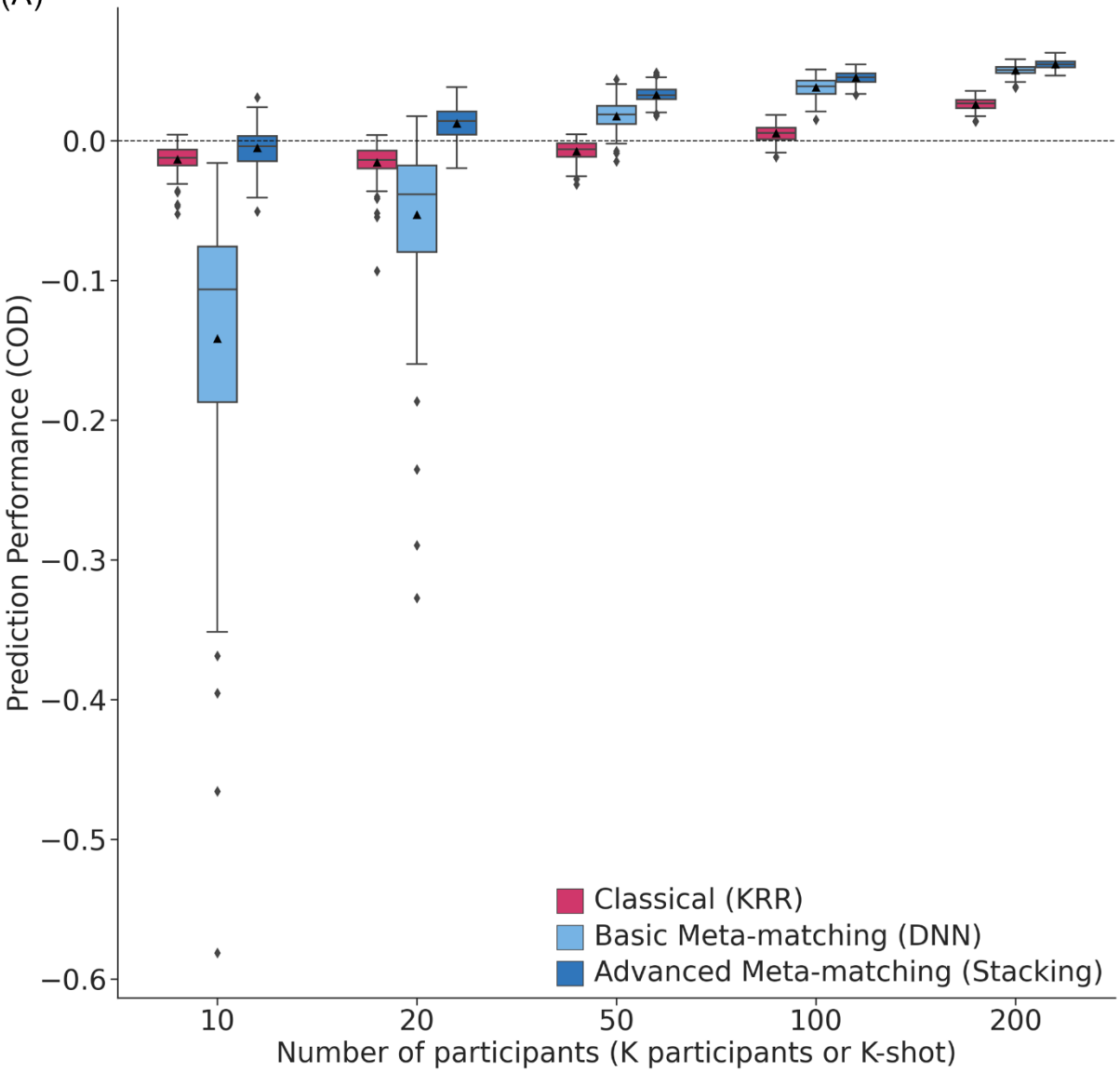


(B)

K Shot	10	20	50	100	200
Classical (KRR) vs Basic Meta-matching (DNN)	0.0916	**0.0179**	**3.2E-05**	**1.7E-07**	**8.3E-07**
Classical (KRR) vs Advanced Meta-matching (Stacking)	0.0513	**0.0028**	**7.1E-08**	**1.3E-09**	**1.0E-10**
Basic Meta-matching (DNN) vs Advanced Meta-matching (Stacking)	0.9626	0.8053	0.5948	0.4035	0.3590

Figure S12. Meta-matching methods outperforms classical kernel ridge regression (KRR) in the HCP dataset (N = 1,019 – K). (A) Prediction performance (Pearson's correlation) with different number of participants. This plot is the same as Figure 7A, but the boxplots now show the bootstrap distribution of each approach based on 1000 bootstrapped samples. The triangles show the average performance (Pearson's correlation) of 35 non-brain-imaging phenotypes using the original 100 random repeats (Figure 7A). We observe that the mean of the bootstrap distributions matches the mean of the original experiments (Figure 7A) quite well. For each boxplot, the horizontal line indicates the median and the black triangle indicates the mean. The bottom and top edges of the box indicate the 25th and 75th percentiles respectively. Whiskers correspond to 1.5 times the interquartile range. Outliers are defined as data points beyond 1.5 times the interquartile range. (B) Statistical differences among the different algorithms. P values were calculated based on a two-sided bootstrapping procedure (see Methods). Bold indicates statistical significance after FDR correction ($q < 0.05$).

(A)



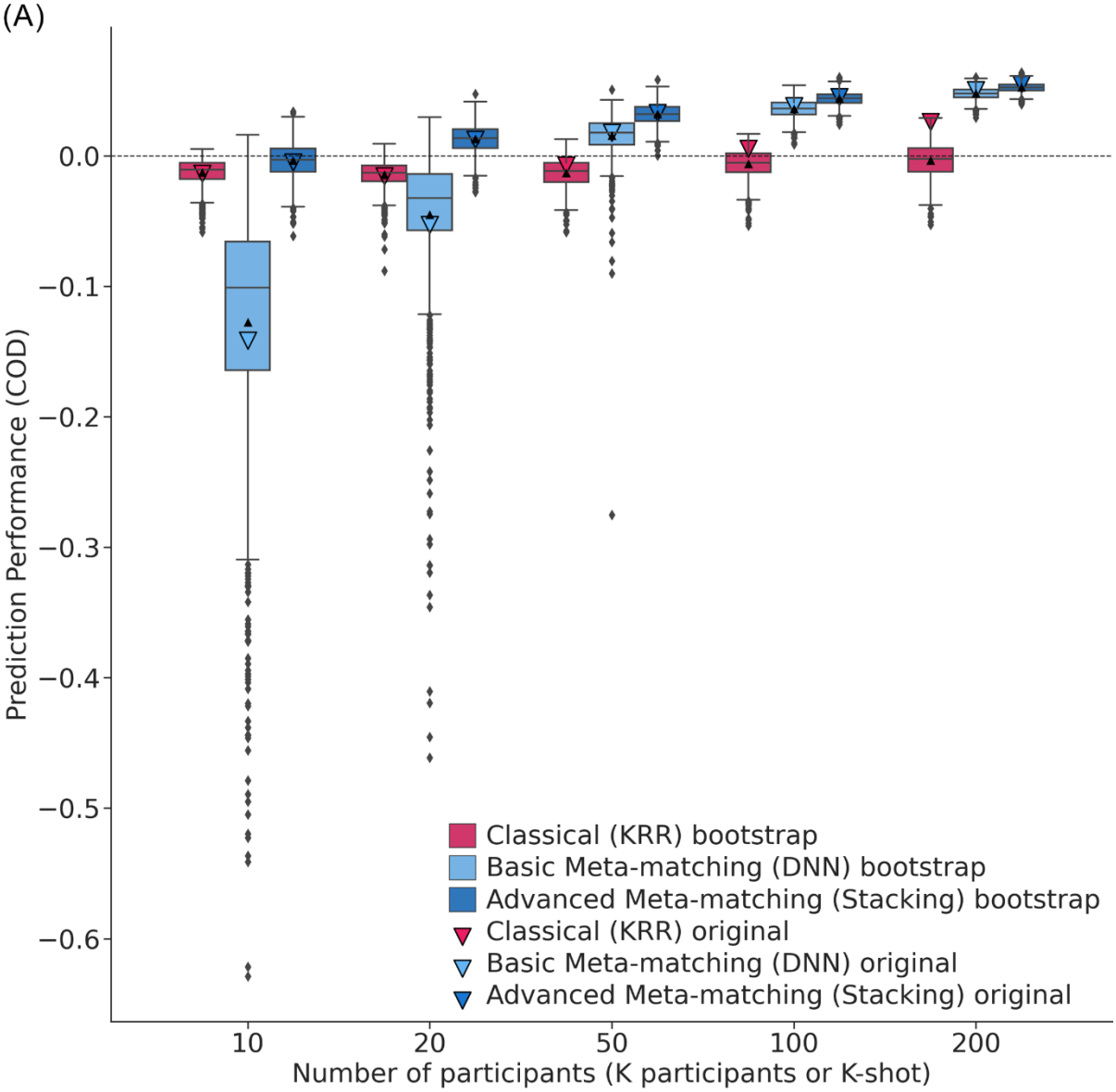
(B)

K Shot	10	20	50	100	200
Classical (KRR) vs Basic Meta-matching (DNN)	n.s.	n.s.	n.s.	**	**
Classical (KRR) vs Advanced Meta-matching (Stacking)	n.s.	*	**	***	**

Figure S13. Meta-matching outperformed classical kernel ridge regression (KRR) baseline in the HCP dataset. (A) Prediction performance (coefficient of determination; COD) averaged across 35 non-brain-imaging phenotypes in the test meta-set (N = 1,019 – K). The K participants were used to train and tune the models (Figure 6B). Boxplots represent variability across 100 random repeats of K participants (Figure 6A). For each boxplot, the horizontal line indicates the median and the black triangle indicates the mean. The bottom and top edges of the box indicate the 25th and 75th percentiles respectively. Whiskers correspond to 1.5 times the interquartile range. Outliers are defined as data points beyond 1.5 times the interquartile range. (B) Statistical difference between the prediction performance (COD) of classical (KRR) baseline and meta-matching algorithms. P values were calculated based on a two-sided bootstrapping procedure (see Methods). “n.s.” indicates that difference was not statistically significant after multiple comparisons correction (FDR $q < 0.05$). "*"

indicates $p < 0.05$ and statistical significance after multiple comparisons correction (FDR $q < 0.05$). "**" indicates $p < 0.001$ and statistical significance after multiple comparisons correction (FDR $q < 0.05$). "***" indicates $p < 0.00001$ and statistical significance after multiple comparisons correction (FDR $q < 0.05$). Green indicates that meta-matching outperforms classical (KRR) baseline. Red indicates that classical (KRR) baseline outperforms meta-matching. The actual p values and statistical comparisons among all algorithms are found in Figure S14.

(A)



(B)

K Shot	10	20	50	100	200
Classical (KRR) vs Basic Meta-matching (DNN)	0.2182	0.5849	0.1753	**6.2E-05**	**5.0E-05**
Classical (KRR) vs Advanced Meta-matching (Stacking)	0.4752	**0.0140**	**2.4E-05**	**3.5E-06**	**1.1E-05**
Basic Meta-matching (DNN) vs Advanced Meta-matching (Stacking)	0.1866	0.2900	0.2930	0.2791	0.5804

Figure S14. Meta-matching outperformed classical kernel ridge regression (KRR) baseline in the HCP dataset (N = 1,019 – K). (A) Prediction performance (coefficient of determination; COD) with different number of participants. This plot is the same as Figure S13A, but the boxplots now show the bootstrap distribution of each approach based on 1000 bootstrapped samples. The triangles show the average performance (COD) of 34 non-brain-imaging phenotypes using the original 100 random repeats (Figure S4A). We observe that the mean of the bootstrap distributions matches the mean of the original experiments (Figure S13A) quite well. For each boxplot, the horizontal line indicates the median and the black triangle indicates the mean. The bottom and top edges of the box indicate the 25th and 75th percentiles respectively. Whiskers correspond to 1.5 times the interquartile range. Outliers are defined as data points beyond 1.5 times the interquartile range. (B) Statistical differences among the different algorithms. P values were calculated based on a two-sided bootstrapping

procedure (see Methods). Bold indicates statistical significance after FDR correction ($q < 0.05$).