Supplementary Information for "Expanding drug targets for 112 chronic diseases using a machine learning-assisted genetic priority score"

Supplemental Tables

Supplemental Figures

Supplementary Table 1: Multivariable linear regression for the proportion of genes identified by P that were also identified by C in genome-wide association testing of common variants.

Variable	Beta	SE	95% CI	p-value
Number of genes identified by P	-0.002	0.002	-0.006 to 0.002	0.293
Number of genes identified by C	0.001	0	0.001 to 0.002	0.003
Proportion of cases	1.342	1.749	-2.16 to 4.843	0.446
AUROC (holdout)	3.792	1.036	1.719 to 5.865	0.001
AUPRC (holdout)	-0.605	0.418	-1.442 to 0.232	0.153

Abbreviations: AUROC (area under the receiver operating characteristic curve); AUPRC (area under the precision-recall curve); SE (standard error); CI (confidence interval); P (observed case/control); C (continuous model probabilities).

Supplementary Table 2: Comparison of effect sizes for common, rare, and ultra-rare variants.

Variable	Rare versus common	Ultra-rare versus common	Ultra-rare versus rare
Two-sided rank-sum tests (unpaired)			
P	1.23, p = 1.21E-26	1.37, p = 2.97E-46	0.14, p = 0.123
B	0.71, p = 7.28E-51	1.34, p = 7.1E-98	0.63, p = 1.38E-5
C	0.23, p = 1.94E-211	0.23, p < 1E-323	0.003, p = 0.018
Two-sided Wilcoxon signed-rank tests (paired by gene)			
P	1.08, p = 9.77E-4, N = 11	1.21, p = 9.77E-4, N = 11	-0.19, p = 0.25, N = 8
B	0.99, p = 1.19E-7, N = 24	0.89, p = 1.19E-7, N = 24	0.12, p = 1.00, N = 17
C	0.20, p = 1.26E-14, N = 86	0.27, p = 1.87E-19, N = 108	-0.16, p = 2.23E-4, N = 68

For common and rare variant analyses, we calculated the median absolute effect size (|Beta|) of all significant variants for each gene, whereas for ultra-rare variant analyses, we used the absolute effect size from the most significant test. The first number in each cell refers to the difference in medians. Abbreviations: P (observed case/control); B (binarized model probabilities/predicted case control); C (continuous model probabilities).

Supplemental Table 3: Odds ratios for drug indication in Open Targets and SIDER of individual variables.

Variable	Analysis	Open Targets	SIDER
ML-GPS features			
EVA-ClinVar	Existing	5.54 (3.86-7.96)	8.28 (5.29-12.96)
HGMD		5.02 (4.26-5.91)	7.24 (5.83-8.99)
OMIM		13.58 (7.86-23.44)	11.71 (5.19-26.42)
L2G		6.62 (5.25-8.35)	6.26 (4.43-8.83)
P	Common variant	7.56 (5.08-11.26)	10.16 (6.11-16.88)
	Rare variant	16.46 (5.95-45.59)	13.69 (3.22-58.22)
	Ultra-rare variant	6.87 (1.95-24.21)	34.31 (10.78-109.22)
B	Common variant	6.28 (4.55-8.68)	7.15 (4.50-11.35)
	Rare variant	15.62 (7.16-34.06)	11.47 (2.75-47.88)
	Ultra-rare variant	8.66 (4.03-18.59)	12.61 (3.69-43.08)
C	Common variant	3.19 (2.53-4.03)	3.14 (2.23-4.42)
	Rare variant	8.75 (5.17-14.80)	17.24 (7.74-38.37)
	Ultra-rare variant	4.02 (2.35-6.88)	7.01 (2.97-16.52)
Additional coverage offered by B and C			
B-P	Common variant	5.21 (3.44-7.90)	6.10 (3.30-11.29)
	Rare variant	15.25 (5.71-40.68)	23.21 (1.45-371.53)
	Ultra-rare variant	10.86 (4.49-26.26)	0.96 (0.00-342.99)
C-P	Common variant	2.62 (2.01-3.41)	2.20 (1.45-3.34)
	Rare variant	7.49 (4.18-13.40)	20.73 (8.30-51.80)
	Ultra-rare variant	3.96 (2.24-6.99)	3.96 (1.16-13.49)

Abbreviations: L2G (locus-to-gene); P (observed case/control); B (binarized model probabilities/predicted case control); C (continuous model probabilities).

Supplementary Table 4: Results of permutation tests comparing AUPRC of different GPS models.

Model comparison	p-value (Open Targets)	p-value (SIDER)
Model architecture comparison		
LR versus GB	0.019	< 0.001
GB versus GB (CE)	< 0.001	< 0.001
GB (CE) versus GB (CE, number weight)	0.887	0.348
GB (CE) versus GB (CE, phase weight)	0.798	0.179
GB (CE, number weight) vs GB (CE, phase weight)	0.885	0.667
Package comparison sensitivity analysis		
LightGBM versus XGBoost	0.001	< 0.001
LightGBM versus random forest	< 0.001	< 0.001
XGBoost versus random forest	< 0.001	0.03
Feature comparison		
L2G versus Clinical	< 0.001	< 0.001
Clinical versus L2G + Clinical	< 0.001	< 0.001
L2G + Clinical versus L2G + Clinical + P	< 0.001	0.008
L2G + Clinical + P versus L2G + Clinical + PBC	< 0.001	< 0.001
Feature comparison (activator drug indications)		
L2G versus Clinical	0.02	0.565
L2G versus L2G + Clinical	0.033	0.041
Clinical versus L2G + Clinical	< 0.001	0.106
L2G + Clinical versus L2G + Clinical + P	0.007	0.008
L2G + Clinical versus L2G + Clinical + PBC	< 0.001	< 0.001
L2G + Clinical + P versus L2G + Clinical + PBC	< 0.001	< 0.001
Feature comparison (inhibitor drug indications)		
L2G versus Clinical	< 0.001	< 0.001
Clinical versus L2G + Clinical	< 0.001	< 0.001
L2G + Clinical versus L2G + Clinical + P	< 0.001	< 0.001
L2G + Clinical + P versus L2G + Clinical + PBC	< 0.001	< 0.001

p-values were calculated using two-sided paired permutation tests with 1,000 permutations, with each permutation entailing random shuffling of predictions for each gene-phecode pair from two different models (see Methods). Values of 0 were recorded as < 0.001. No adjustments were made for multiple comparisons as these tests were hypothesis driven. Abbreviations: LR (logistic regression); GB (gradient boosting); CE (continuous encoding); L2G (locus-to-gene); P (observed case/control); B (binarized model probabilities/predicted case control); C (continuous model probabilities).

Supplementary Table 5: Tractability characteristics of highest-scoring gene-phecode pairs.

Variable	No value	< 0 (unfavorable)	0 (neutral)	> 0 (favorable)
Membrane protein	990 [10.0%]	0 [0.0%]	7080 [71.4%]	1851 [18.7%]
Secreted protein	990 [10.0%]	0 [0.0%]	8045 [81.1%]	886 [8.9%]
Safety event	9554 [96.3%]	367 [3.7%]	0 [0.0%]	0 [0.0%]
Has pocket	838 [8.4%]	0 [0.0%]	8465 [85.3%]	618 [6.2%]
Has ligand	838 [8.4%]	0 [0.0%]	7625 [76.9%]	1458 [14.7%]
Has small molecule binder	838 [8.4%]	0 [0.0%]	7232 [72.9%]	1851 [18.7%]
Is cancer driver gene	9688 [97.7%]	233 [2.3%]	0 [0.0%]	0 [0.0%]
Has TEP	9893 [99.7%]	0 [0.0%]	0 [0.0%]	28 [0.3%]
Has high quality chemical probes	9352 [94.3%]	0 [0.0%]	223 [2.2%]	346 [3.5%]
Tissue specificity	908 [9.2%]	3999 [40.3%]	0 [0.0%]	5014 [50.5%]
Tissue distribution	908 [9.2%]	4536 [45.7%]	2943 [29.7%]	1534 [15.5%]

Tractability information for 9,916 distinct genes represented among the top 23,626 pairs. We obtained this data from Open Targets Platform version 23.12. Abbreviations: TEP (target enabling package).

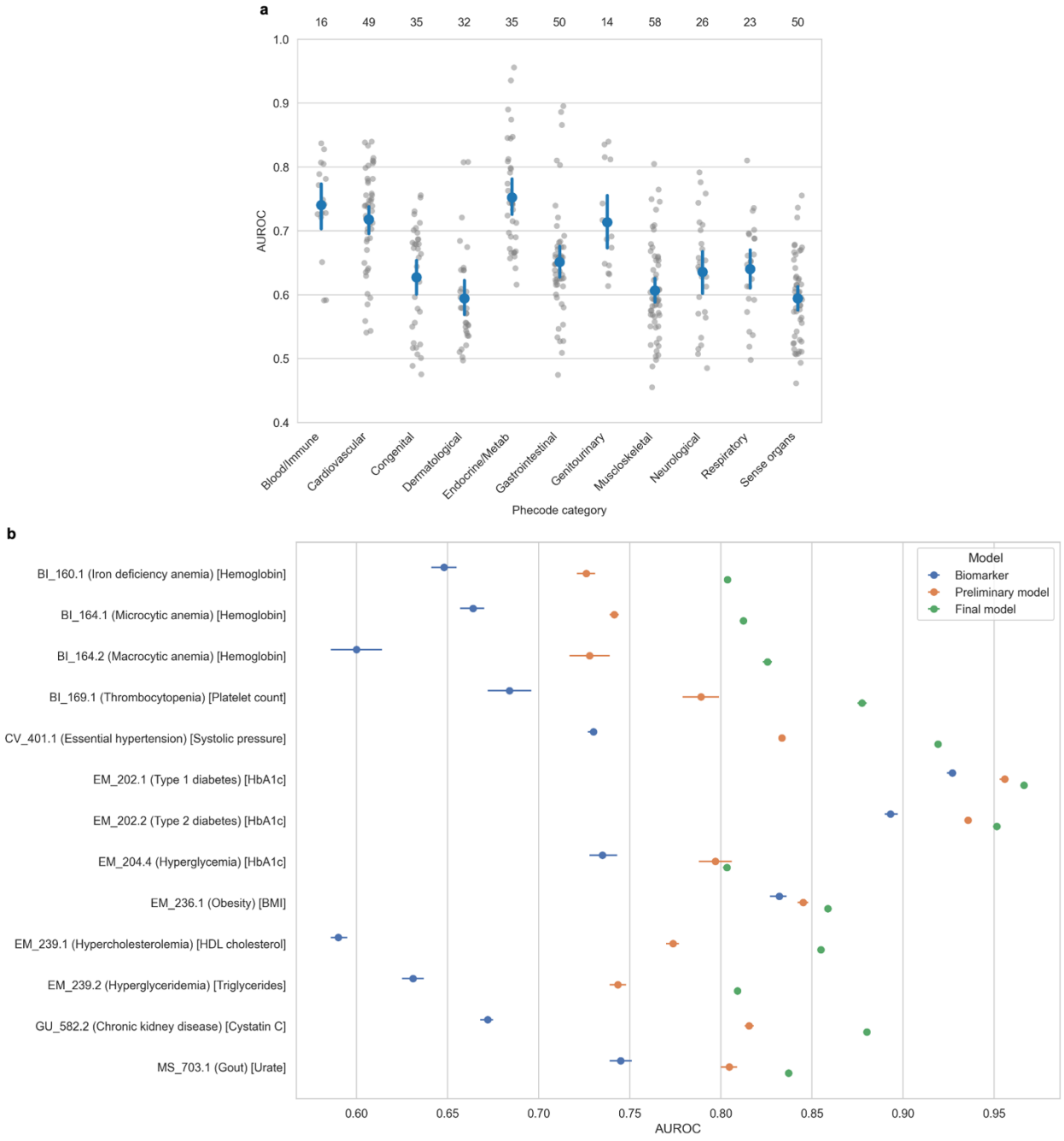
Supplementary Table 6: Coefficients from the ElasticNet non-directional genetic priority score.

Feature	Coefficient
EVA-ClinVar	0.35
HGMD	1.062
OMIM	1.32
L2G	1.187
P (common)	0.18
P (rare)	0.797
P (ultra-rare)	0.143
C (common)	0.562
C (rare)	1.058
C (ultra-rare)	0.277
B (common)	0.526
B (rare)	0.84
B (ultra-rare)	0

Coefficients are averages of five holdout predictions during five-fold cross-validation. Abbreviations: L2G (locus-to-gene); P (observed case/control); B (binarized model probabilities/predicted case control); C (continuous model probabilities).

Supplemental Figures

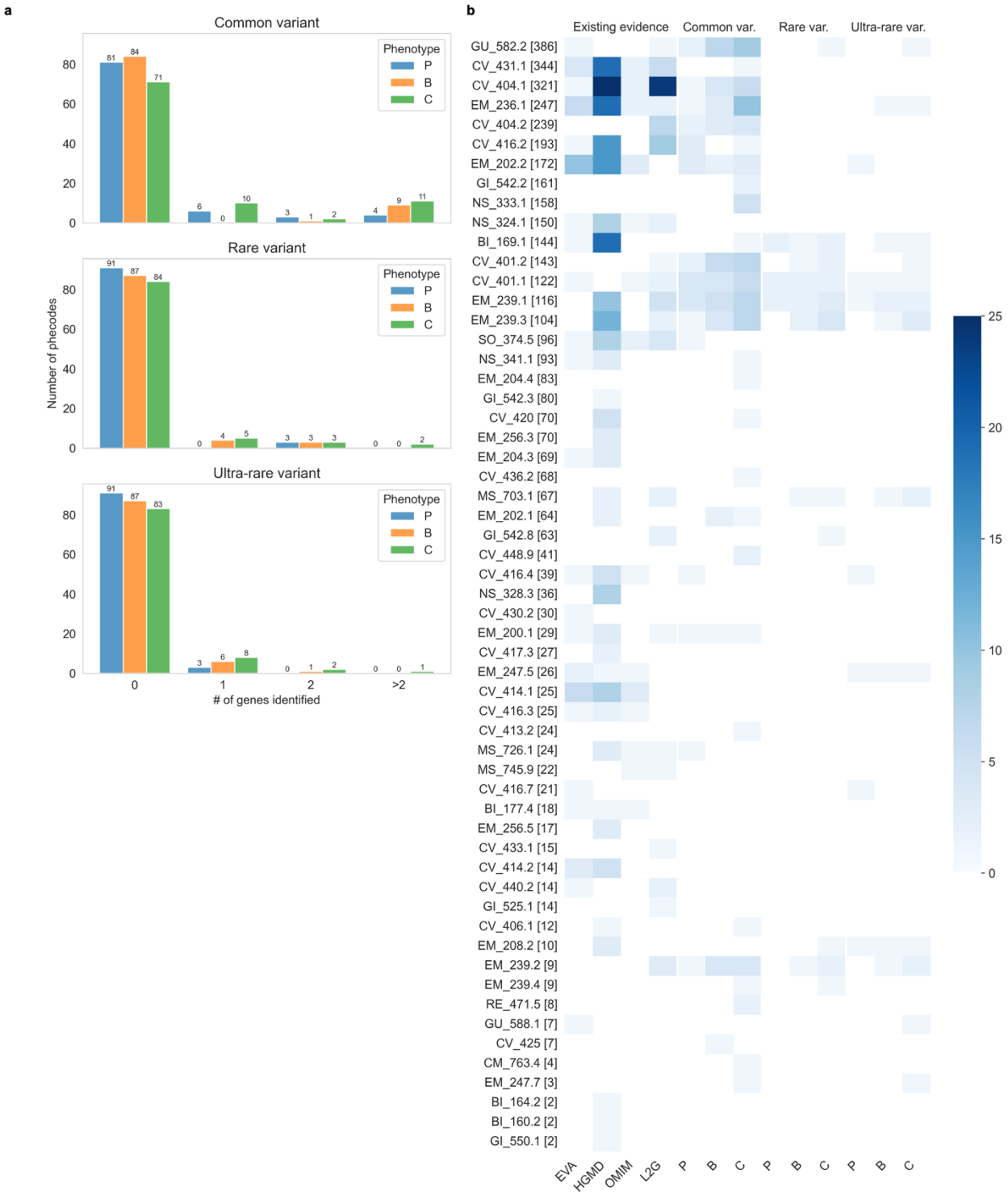
Supplementary Figure 1: Performance of preliminary models and single biomarkers for predicting phecode diagnoses.



a Mean AUROCs (blue) of preliminary models for 386 chronic disease phecodes. Numbers at the top of the graph indicate the number of phecodes in each phecode category; each phecode is represented as a grey dot in the background. **b** Mean AUROCs of biomarkers (blue), preliminary models (orange), and final models (green) for 13 phecodes definable using single

biomarkers. Labels on the y-axis are as follows: phecode (phecode name) [biomarker name]. For EM_239.1 (Hyperglyceridemia), HDL cholesterol achieved a higher AUROC than LDL cholesterol (Table S7). AUROCs in **a** and **b** were calculated among 183,021 UK Biobank participants with GP records (see “Study sample” in the Methods section). Plot **b** shows means with 95% confidence intervals. Source data are provided in Supplementary Data 2 and Supplementary Data 5. Abbreviations: AUROC (area under the receiver operating characteristic curve); HbA1c (hemoglobin A1c); BMI (body mass index).

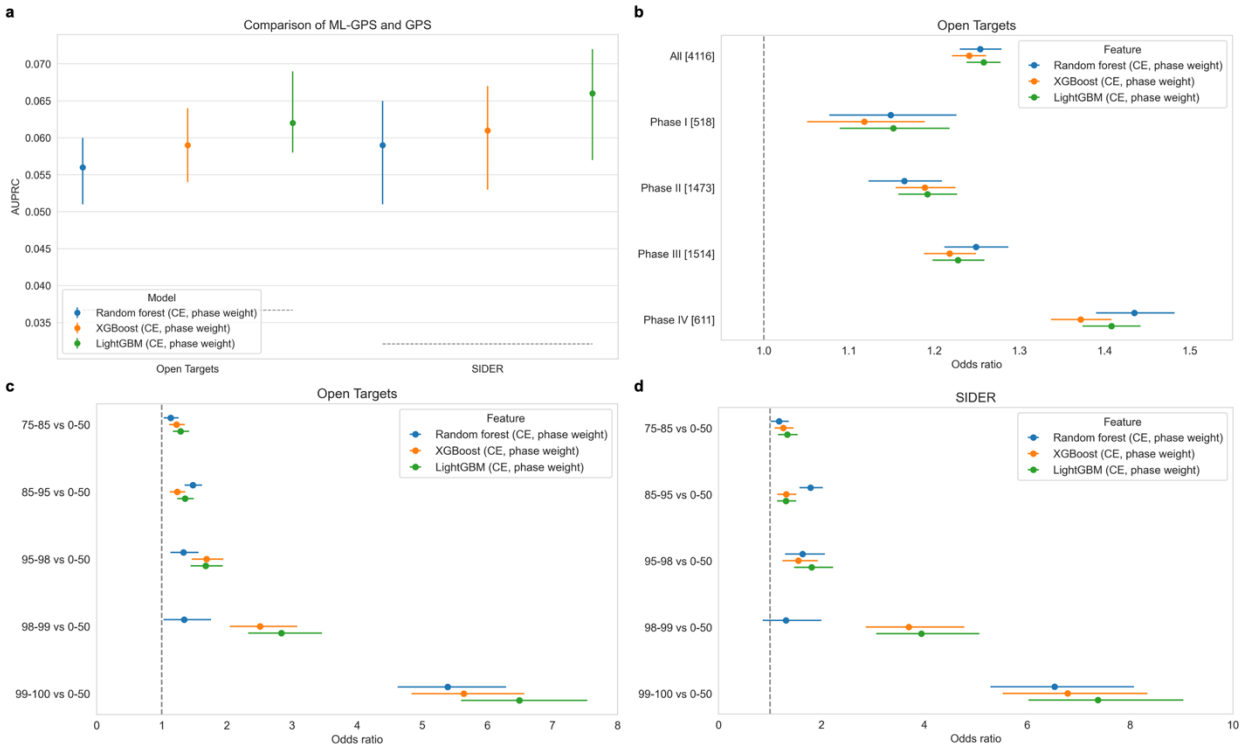
Supplementary Figure 2: Identification of genes with drug indications by individual variables.



a Number of genes with drug indications identified per phecode by P (blue), B (orange), and C (green). **b** Number of genes with drug indications for each phecode identified by each variable,

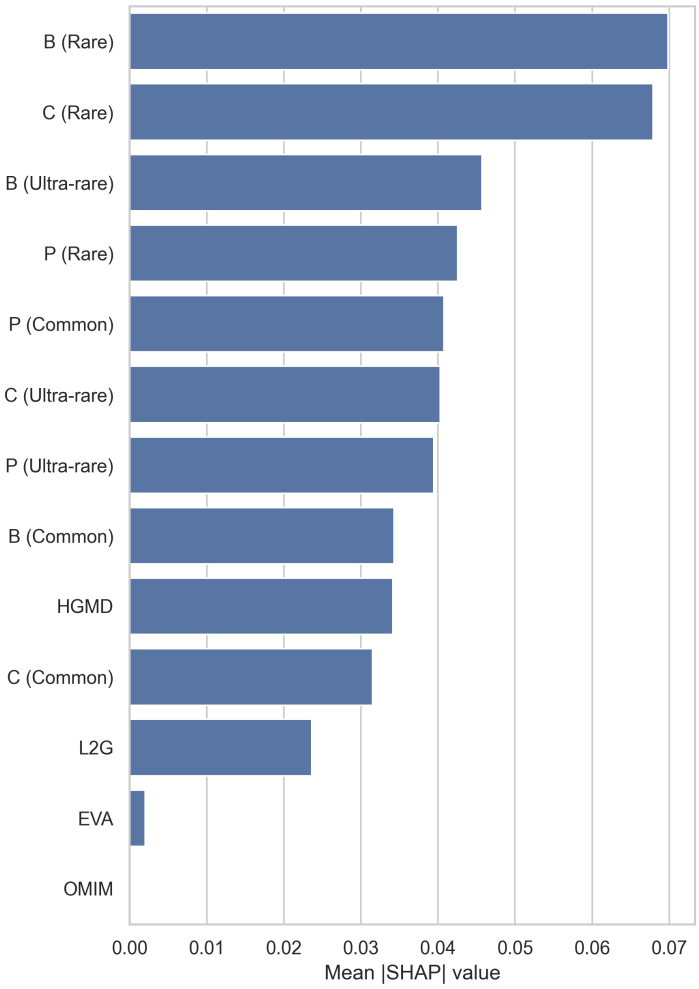
with darker colors representing more genes (see color bar on right side). Labels on the y-axis are as follows: phecode [total number of genes with drug indications]. Note that both **a** and **b** represent binary-encoded variables, whereas ML-GPS used continuously-encoded variables as features. Source data are provided as a Source Data file. Abbreviations: P (observed case/control); B (binarized model probabilities/predicted case control); C (continuous model probabilities).

Supplementary Figure 3: Comparison of random forest, XGBoost, and LightGBM models.



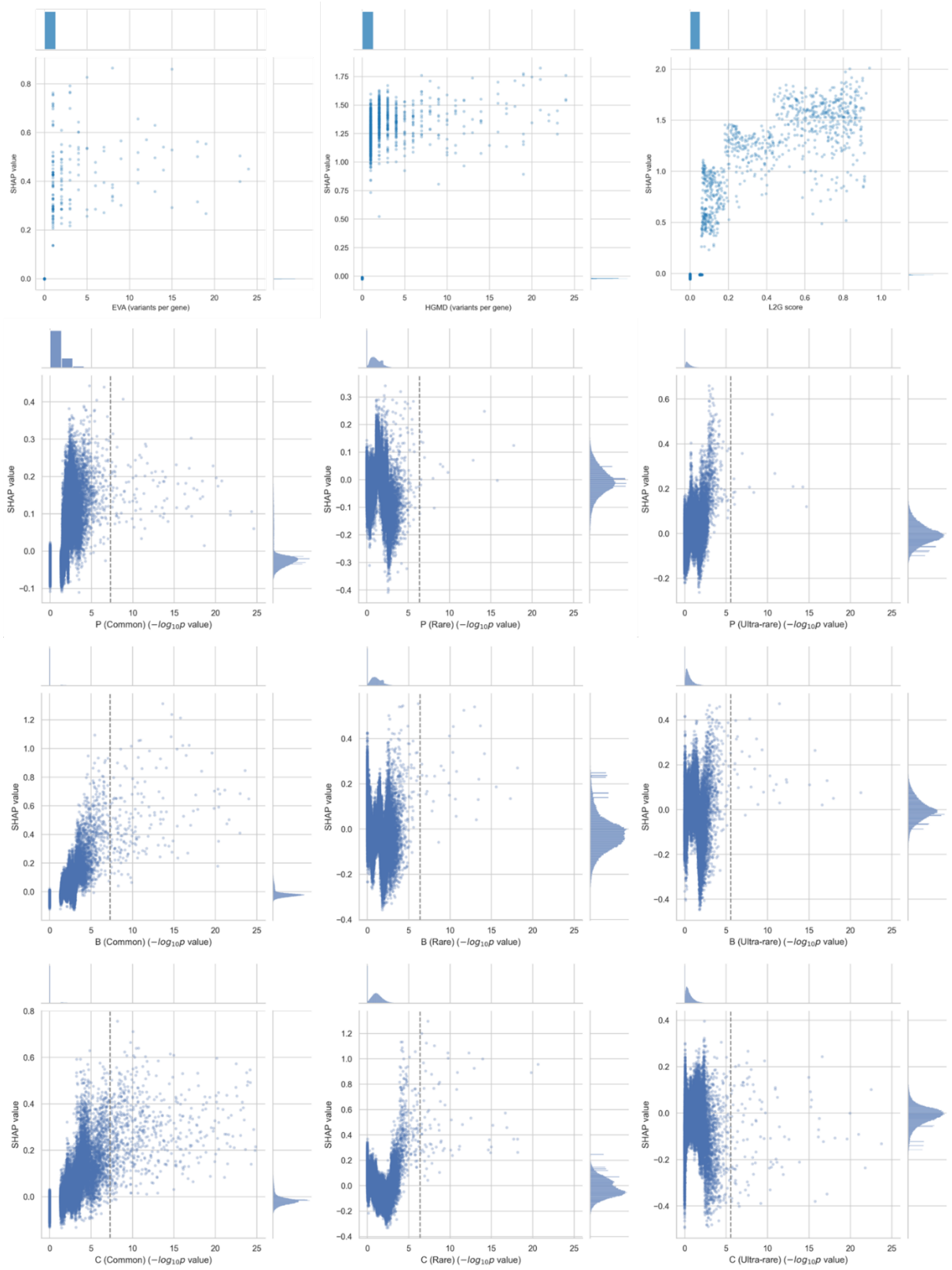
a AUPRC for drug indication in Open Targets (holdout testing) and SIDER (external testing) for random forest (blue), XGBoost (orange), or LightGBM (green) models. Grey dotted lines show the proportion of gene-phecode pairs with indications in each dataset. **b** Odds ratios per standard deviation increase in score for any drug indication and separately for drug indications in specific clinical trial phases in Open Targets. Brackets indicate number of indications in each phase. **c-d** Odds ratios for drug indication for gene-phecode pairs in the top X score percentiles compared to pairs in the 0-50 percentiles in Open Targets **(c)** and SIDER **(d)**. Plots **a-c** and represent analyses of 112,274 gene-phecode pairs in Open Targets, of which 4,116 had a drug indication. Plots **a** and **d** represent analyses of 58,674 gene-phecode pairs in SIDER, of which 1,883 had a drug indication. All plots show means with 95% confidence intervals. Source data are provided as a Source Data file. Abbreviations: AUPRC (area under the precision-recall curve).

Supplementary Figure 4: SHAP analysis of L2G + Clinical + PBC model holdout predictions in Open Targets.



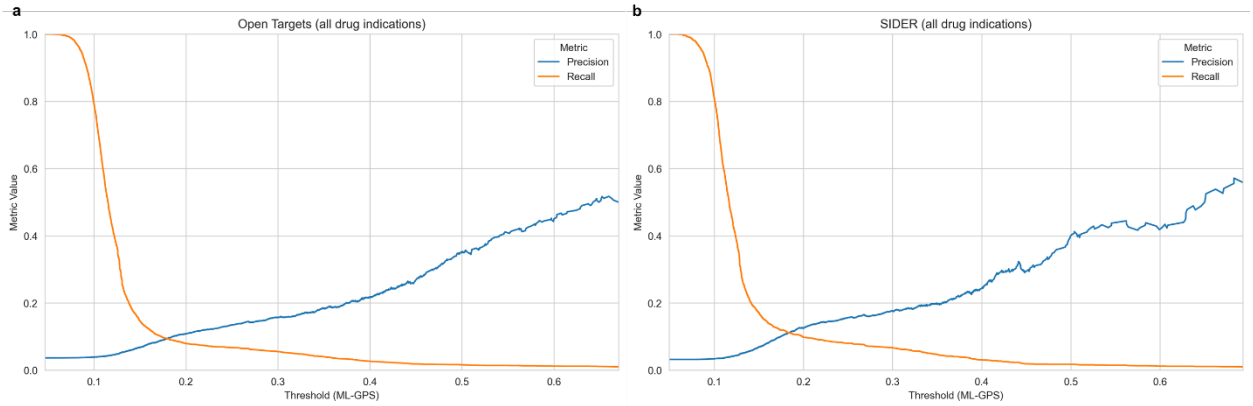
Relative feature importance, as represented by the mean of absolute SHAP values for each feature. SHAP analysis was performed for holdout predictions for 122,274 gene-phecode pairs in the Open Targets dataset. Abbreviations: SHAP (SHapley Additive exPlanations); L2G (locus-to-gene); P (observed case/control); B (binarized model probabilities/predicted case control); C (continuous model probabilities).

Supplementary Figure 5: Correlations between SHAP values and feature values for L2G + Clinical + PBC model holdout predictions in Open Targets.



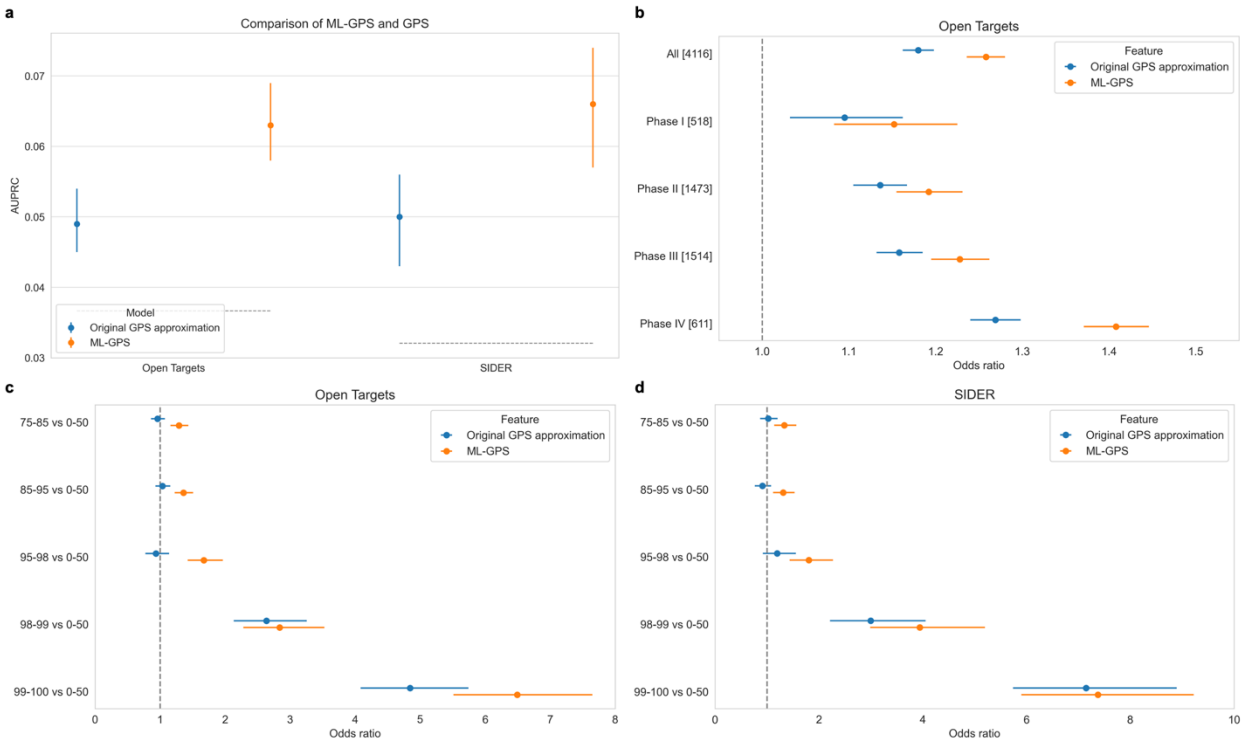
Scatterplots between feature values and SHAP values. Each point on the scatterplot represents one of 122,274 gene-phecode pairs in the Open Targets dataset. Abbreviations: SHAP (SHapley Additive exPlanations); L2G (locus-to-gene); P (observed case/control); B (binarized model probabilities/predicted case control); C (continuous model probabilities).

Supplementary Figure 6: Performance of ML-GPS at different score thresholds.



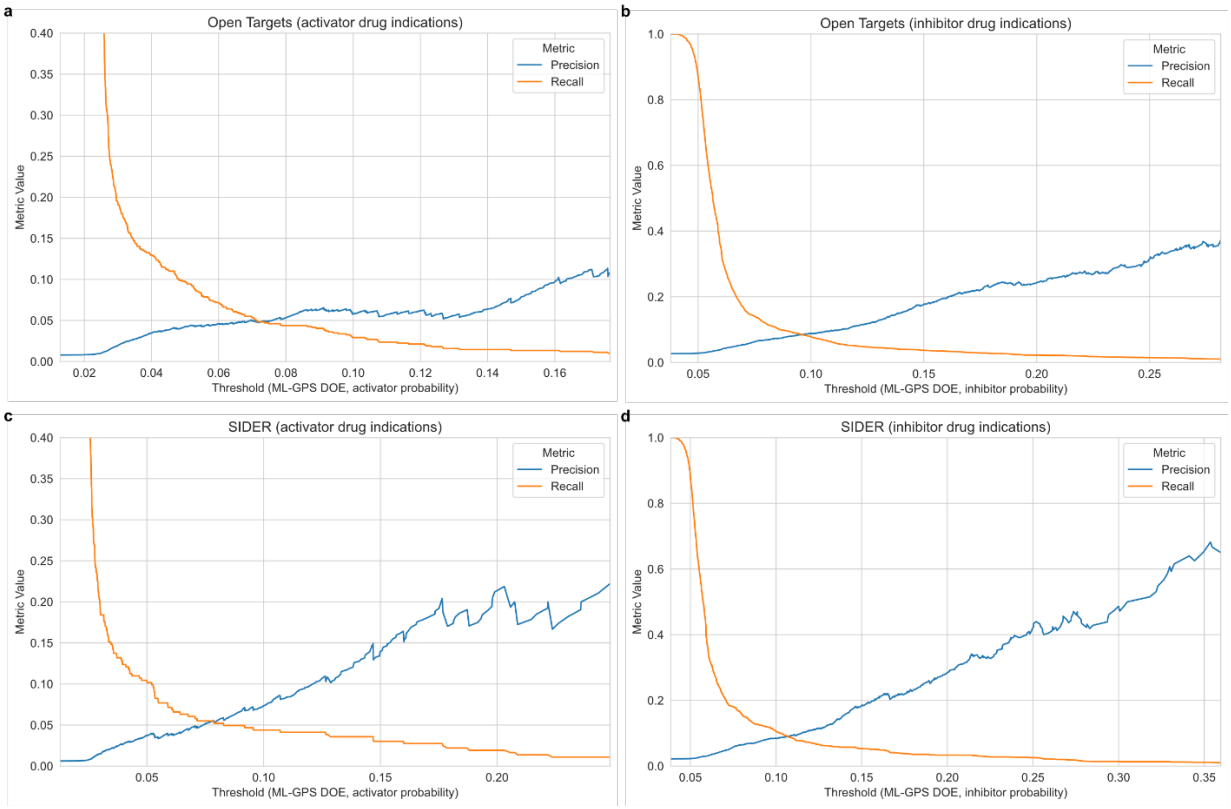
Plots of precision (blue) and recall (orange) for different thresholds of ML-GPS score in either Open Targets **(a)** or SIDER **(b)**. Plot **a** represents 122,274 gene-phecode pairs in Open Targets, of which 4,116 had a drug indication. Plot **b** represents 58,674 gene-phecode pairs in SIDER, of which 1,883 had a drug indication. Source data are provided as a Source Data file.

Supplementary Figure 7: Comparison of ML-GPS to an approximation of the original GPS (logistic regression architecture with L2G + Clinical + P features).



a AUPRC for drug indication in Open Targets (holdout testing) and SIDER (external testing) for an approximation of the original GPS (blue) and ML-GPS (orange). Grey dotted lines show the proportion of gene-phecode pairs with indications in each dataset. **b** Odds ratios per standard deviation increase in score for any drug indication and separately for drug indications in specific clinical trial phases in Open Targets. Brackets indicate number of indications in each phase. **c-d** Odds ratios for drug indication for gene-phecode pairs in the top X score percentiles compared to pairs in the 0-50 percentiles in Open Targets **(c)** and SIDER **(d)**. Plots **a-c** and represent analyses of 112,274 gene-phecode pairs in Open Targets, of which 4,116 had a drug indication. Plots **a** and **d** represent analyses of 58,674 gene-phecode pairs in SIDER, of which 1,883 had a drug indication. All plots show means with 95% confidence intervals. Source data are provided as a Source Data file. Abbreviations: AUPRC (area under the precision-recall curve).

Supplementary Figure 8: Performance of ML-GPS DOE at different score thresholds.



Plots of precision (blue) and recall (orange) for different thresholds of each score in either Open Targets **(a-b)** or SIDER **(c-d)**: ML-GPS DOE activator predictions **(a, c)**, and ML-GPS DOE inhibitor predictions **(b, d)**. In analyses of activator drug indications, inhibitor drug indications were set to 0 (no drug indication), and vice versa. Plots **a** and **c** represent 122,274 gene-phecode pairs in Open Targets, of which 890 had an activator drug indication and 3,019 had an inhibitor drug indication. Plots **b** and **d** represent 58,674 gene-phecode pairs in SIDER, of which 364 had an activator drug indication and 1,288 had an inhibitor drug indication. Source data are provided as a Source Data file. Abbreviations: DOE (direction-of-effect).