**Benchmarking**

*Core benchmarks*

In the benchmarks performed in this paper, we have had to make decisions that could impact the projected performance of the different methods. To be fair to all methods, we have tried to equalize the input as much as possible. To this end, the core three benchmarks as displayed in Figure 4 are performed on fine-mapping results from L2G1. This ensures that cS2G2, distance, FLAMES, and L2G are run on the same SNPs. To equalize across different annotation methods we have done the following:

i. If a gene is not annotated by both L2G and FLAMES, we discard that gene. This ensures that differences are not due to annotation.

ii. If a gene is not annotated by L2G and FLAMES, we do not assign a cS2G score.

To prevent overrepresentation of a single gene-trait pair, we remove duplicate gene-trait pairs, by only keeping the credible set that comes from the lowest P-value L2G locus mapping to a true positive gene for any given phenotype. The methods describing the core benchmarks can be found in the methods.

*Additional benchmarks – Ei*

Ei is different from L2G, FLAMES, and cS2G as the entire unit of analysis shifts from a single SNP/credible set, to the entire locus. This means that we need to convert FLAMES scores to scores for all the credible sets within each locus. The difficulty here is that FLAMES scores do not incorporate the fine-mapping relevance of the credible sets, only the confidence of a gene being the effector gene of a credible set. With the unit of analysis shifting to larger loci, it matters which genes are annotated to the locus to compare performance in an Area Under the Precision-Recall curve (AUPRC), as this can be strongly skewed by class imbalance. To transform the fine-mapped regions from Forgetta *et al*3*.*, we transform the PIPs by multiplying them by their Bayes factor and dividing them by the summed Bayes factor of the locus, merging into a single credible set. Ei scores for ExWAS-implicated genes were acquired from the authors of the publication benchmarking Ei on ExWAS genes4. Ei scores were made per locus, inputting multiple credible sets. To compare them to FLAMES scores, we averaged all scores for a gene within traits.

*Additional benchmarks – three molecular traits with full fine-mapping*

An additional benchmarking was performed for the FUMA5 loci that map a true positive gene from the three molecular traits benchmark. Given the increased chance of false positives from erroneously generated credible sets by FINEMAP6 we opted to use the more conservative raw-only threshold. In this benchmark, recall is calculated as the proportion of unique causal genes recovered. Given that the chance of spurious signals being present in the loci with multiple fine-mapped credible sets is higher, we used the more conservative minimum raw score threshold (see next section – calibration).

**Calibration**

We calculated a recommended threshold of the raw and scaled FLAMES scores that corresponds with an estimated overall prediction 75% precision in the ExWAS-implicated set (Fig. 4d). This set was chosen because:

1. It is data-driven.
2. It has the closest gene precision most similar to previously established GWAS estimates7.
3. It has the lowest overall AUPRC, therefore resulting in the most conservative threshold estimation.

We recommend prioritizing genes if they have the highest scaled FLAMES score in the locus, this score is higher than 0.248, and they also have a raw FLAMES score higher than 0.136.

Alternatively, we calculated a 75% precision threshold based solely on the raw scores, corresponding to a minimum raw score of 0.307. This ensures that there is always a high baseline level of evidence in a locus and scores are not inflated due to scaling. However, this lowers the estimated recall from 0.364 to 0.273. We therefore recommend using the combined scaled FLAMES score > 0.248 and raw FLAMES score > 0.136 by default. We would like to stress that these confidence thresholds were calculated in genome-wide significant loci and might not translate to non-significant loci.

**References**

1. Mountjoy, E. *et al.* An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nat. Genet.* **53**, 1527–1533 (2021).

2. Gazal, S. *et al.* Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity. *Nat. Genet.* **54**, 827–836 (2022).

3. Forgetta, V. *et al.* An effector index to predict target genes at GWAS loci. *Hum. Genet.* **141**, 1431–1447 (2022).

4. Liang, K. Y. H. *et al.* Predicting ExWAS findings from GWAS data: a shorter path to causal genes. *Hum. Genet.* **142**, 749–758 (2023).

5. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).

6. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).

7. Weeks, E. M. *et al.* Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. *Nat. Genet.* 1–10 (2023) doi:10.1038/s41588-023-01443-6.