Deep multiple-instance learning accurately predicts gene haploinsufficiency and deletion pathogenicity

Zhihan Liu[1,2,3*], Yi-Fei Huang[1,3*]

[1]Department of Biology, Pennsylvania State University,
University Park, PA 16802, USA

[2]Molecular, Cellular, and Integrative Biosciences Program, Pennsylvania State University,
University Park, PA 16802, USA

[3]Huck Institutes of the Life Sciences, Pennsylvania State University,
University Park, PA 16802, USA

*Correspondence: zul327@psu.edu, yuh371@psu.edu

Abstract

Copy number losses (deletions) are a major contributor to the etiology of severe genetic disorders. Although haploinsufficient genes play a critical role in deletion pathogenicity, current methods for deletion pathogenicity prediction fail to integrate multiple lines of evidence for haploinsufficiency at the gene level, limiting their power to pinpoint deleterious deletions associated with genetic disorders. Here we introduce DosaCNV, a deep multiple-instance learning framework that, for the first time, models deletion pathogenicity jointly with gene haploinsufficiency. By integrating over 30 gene-level features potentially predictive of haploinsufficiency, DosaCNV shows unmatched performance in prioritizing pathogenic deletions associated with a broad spectrum of genetic disorders. Furthermore, DosaCNV outperforms existing methods in predicting gene haploinsufficiency even though it is not trained on known haploinsufficient genes. Finally, DosaCNV leverages a state-of-the-art technique to quantify the contributions of individual gene-level features to haploinsufficiency, allowing for human-understandable explanations of model predictions. Altogether, DosaCNV is a powerful computational tool for both fundamental and translational research.

Background

Copy number variations (CNVs) are a major contributor to human genetic diversity[1,2]. These structural variations, which involve deletions or duplications of genomic regions ranging from 50 base pairs to several megabase pairs, result from diverse mutational mechanisms during DNA replication, recombination, and repair[3,4]. CNVs may provide adaptive advantages under certain circumstances[5,6]; however, they can also lead to diseases when critical genomic regions are affected[7]. Both common and rare CNVs (population frequency < 1%) contribute to disease development, with rare CNVs particularly implicated in rare genetic disorders[2,8–13]. In the context of disease-associated CNVs, pathogenicity often arises from alterations affecting dosage-sensitive (DS) driver genes[14,15]. This concept of dosage sensitivity refers to the emergence of deleterious phenotypes when gene copy number changes. Specifically, haploinsufficient (HI) genes are intolerant to copy number loss, whereas triplosensitive (TS) genes are intolerant to copy number gain[16,17]. Thus, identifying DS genes is essential for prioritizing clinically relevant CNVs and understanding the genetic mechanisms of associated diseases.

Efforts to discover DS genes primarily involve three frameworks: machine learning-based methods, case-control data-based statistical methods, and selection pressure-based approaches. Machine learning-based methods utilize gene-level features and various algorithms to differentiate manually curated DS genes from dosage-insensitive genes[18–23]. While these models identify predictive attributes of DS genes, the limited availability of high confidence training data restricts their ability to capture the full spectrum of DS genes. Alternatively, statistical methods can be utilized to analyze case-control CNV data to identify DS genes, assuming that these genes have a higher prevalence among affected individuals[15]. For instance, a recent approach employs a modified GWAS fine-mapping framework to estimate gene-level dosage sensitivity[24]. Although this framework can predict DS genes without labeled training data, it is challenging to pinpoint causal genes because CNVs often span multiple genes. Finally, selection pressure-based methods utilize natural selection against loss-of-function (LOF) mutations or CNVs as indicators of dosage

sensitivity[25–30]. Although dosage sensitivity is strongly correlated with selection pressure at the gene level, exceptions do exist. For instance, alterations in pleiotropic genes critical for fitness may not lead to significant phenotypic changes[31,32], and DS genes associated with late-onset disorders may not be subject to strong selection pressure[33–36]. Additionally, these selection pressure-based methods may be underpowered for detecting short DS genes due to sparse polymorphisms in the human genome[27,28].

Previous research prioritizes pathogenic CNVs using two frameworks: rule-based methods and supervised machine learning methods. Rule-based approaches automate pathogenicity scoring systems based on clinical CNV interpretation guidelines[37,38], such as the one proposed by the American College of Medical Genetics and Genomics (ACMG) and ClinGen[39]. However, these methods face limitations due to rigid score assignments and difficulties in identifying pathogenicity originating from uncatalogued genomic regions. On the other hand, supervised machine learning methods, often trained on clinically curated CNVs, utilize simple summary statistics aggregated across affected genomic elements, such as the mean score of dosage sensitivity across genes, for pathogenicity prediction. However, the optimal summary statistics for deletion pathogenicity prediction are often unknown, limiting the accuracy of supervised machine learning methods. For instance, the mean score of dosage sensitivity across genes might not adequately convey the significance of a pathogenic CNV harboring a single highly penetrant DS driver gene amidst multiple dosage-insensitive genes.

Furthermore, while current methods offer valuable insights into gene dosage sensitivity and CNV pathogenicity, they struggle to connect the two. DS gene identification methods predict gene dosage sensitivity but lack a natural way of aggregating these predictions to infer CNV pathogenicity. Conversely, pathogenic CNV classifiers predict CNV pathogenicity without identifying DS driver genes and evaluating their contributions. To bridge this gap and enhance predictions at both the gene and CNV levels, we propose that multiple-instance learning (MIL) presents a promising, unified solution for the joint predictions of gene dosage sensitivity and CNV pathogenicity[40]. MIL is suited for problems where each

63 sample, or bag, in the dataset contains a varying number of observations, or instances, and is classified as
64 positive if at least one instance is positive[40]. The objective is to predict labels at both the bag and instance
65 levels, assuming that the bag-level label results from the combined effect of individual instances, even
66 when only bag-level labels are available during training. This approach is well-aligned with our problem,
67 as we can view each CNV as a bag and the affected genes within it as instances, with pathogenicity labels
68 given solely at the bag (CNV) level.

69 In this study, we introduce DosaCNV, a deep MIL framework modeling deletion pathogenicity through the
70 joint effect of haploinsufficiency from affected genes. We focus exclusively on deletions that affect
71 protein-coding genes, which have more direct functional consequences compared to deletions that only
72 affect non-coding regions and duplications. DosaCNV leverages over 30 gene-level features from diverse
73 biological processes and is trained on clinically curated deletions. It outperforms existing methods in
74 predicting pathogenic deletions and HI genes across various sources. DosaCNV demonstrates the potential
75 to leverage large-scale deletion data for enhancing the predictions of HI genes and associating these
76 predictions with deletion pathogenicity, establishing it as an accurate and valuable computational tool for
77 both fundamental and translational research.

78 **Results**

79 **Overview of DosaCNV.** DosaCNV is a supervised deep MIL model designed to simultaneously infer the
80 pathogenicity of coding deletions and the haploinsufficiency of genes, based on the assumption that the
81 joint effect of gene haploinsufficiency determines deletion pathogenicity. Utilizing a fully connected
82 neural network, chosen for its strong ability to automatically extract useful features, DosaCNV analyzes
83 coding deletions by first computing the haploinsufficiency of affected genes (Fig. 1). This fully connected
84 neural network processes an input matrix for each deletion, with rows representing affected genes and

85 columns representing gene-level features. Then, it generates a vector of predicted HI probabilities (P_{HI})
86 for each affected gene within the deletion (Fig. 1).

87 Subsequently, DosaCNV utilizes the noisy-or pooling function[41], a probabilistic variant of the logical-or
88 operator commonly used in Bayesian networks, to combine the predicted HI probabilities at the gene level
89 into a scalar output ranging from 0 to 1, representing the joint effect of gene-level haploinsufficiency (Fig.
90 1). This output is also the predicted probability that the given deletion overlaps with at least one HI gene
91 by the formalization of the noisy-or function[41], and serves as a proxy for deletion pathogenicity (Fig. 1).
92 Notably, due to the noisy-or pooling approach, DosaCNV assumes independent effects between affected
93 genes and does not account for gene-gene interactions.

94 DosaCNV's design addresses the challenge of limited access to curated HI genes faced by previous
95 supervised gene classifiers. By training the model's fully connected architecture for HI prediction on all
96 affected genes within the deletion data, DosaCNV potentially accesses a larger and more diverse
97 agglomeration of HI genes compared to currently available sets. This approach may enable DosaCNV to
98 capture distinct HI gene characteristics more effectively, enhancing HI gene prediction relative to other
99 DS gene classifiers. Combining this improvement with the probabilistic integration of HI predictions for all
100 affected genes within a deletion of interest, DosaCNV could potentially yield more accurate deletion-level
101 pathogenicity predictions.

102 **Training DosaCNV on clinically curated deletions.** We retrieved 6,180 pathogenic and 5,822 benign
103 deletions overlapping with gene coding sequences (CDSs) from ClinVar[42], a pan-disorder repository of
104 clinically curated genetic variants. Also, we collected 33 gene-level features potentially predictive of
105 haploinsufficiency. These gene-level features can be grouped into six categories: genomics, epigenomics,
106 evolutionary constraints, functional genomics, deleteriousness upon perturbation, and biological
107 networks (Supplementary Data 1). We used the features of affected protein-coding genes within a

deletion as the input features of DosaCNV, and used the clinically curated pathogenicity of the deletion in ClinVar as the label. Then, we utilized a "leave-one-chromosome-out" approach to partition the data into a training set, a validation set, and a held-out test set (Methods). We trained DosaCNV on the training set and optimized the hyperparameters of DosaCNV on the validation set. To minimize potential bias arising from relying on deletion length for predictions[43,44], we matched pathogenic and benign deletions based on the number of affected protein-coding genes within the training and validation sets. The test set remained unmatched for subsequent analysis. As a result, the final training, validation, and test sets contained 2,308, 1,060, and 2,832 coding deletions, respectively.

Evaluating DosaCNV and alternative methods on ClinVar pathogenic deletions. We assessed DosaCNV's performance in predicting pathogenic deletions on the held-out test set. To mitigate potential biases in performance assessment due to class imbalance and length differences between pathogenic and benign deletions, we benchmarked DosaCNV with five alternative methods (TADA[44], StrVCTVRE[45], HIS-LOD[18], X-CNV[46], and CADD-SV[47]) on the unmatched held-out test set and three subsets matched for the number of affected protein-coding genes, deletion length, and both criteria, respectively.

Within the unmatched test set, 1,211 pathogenic deletions and 1,280 benign deletions had predictions from all five alternative methods (n = 2,491). DosaCNV achieved the highest area under the receiver-operating-characteristic curve (AUC) of 0.968 (Fig. 2A; Supplementary Table 1). To evaluate the performance of various methods for deletions with different lengths, we categorized the unmatched test set into four length groups: small (<50 kb), medium (50-100 kb), large (100-1000 kb), and extra-large (>1000 kb). DosaCNV consistently delivered superior or comparable predictions across all length groups (Fig. 2B; Supplementary Table 2). Although other methods also performed reasonably well, their performance might be attributed to differences in length distribution between pathogenic and benign deletions (Supplementary Fig. 1A & B). Indeed, using either the number of affected genes or deletion

length alone as a predictor for deletion pathogenicity yielded AUC values as high as 0.886 and 0.912, respectively (Fig. 2A).

To account for this potential confounder, we examined the influence of two forms of length information on performance: the number of affected genes and deletion length. We generated two matched test sets with a 1:1 ratio of pathogenic to benign deletions based on the number of affected genes (n=702) and deletion length (n=704), respectively. DosaCNV maintained the best performance on these datasets, achieving AUC values of 0.895 and 0.891, respectively (Supplementary Fig. 2A – 2D; Supplementary Table 1 & 3). Nevertheless, a substantial decrease in performance for all methods was observed when compared to the unmatched test set, suggesting that differences in length distribution contributed to the overly optimistic evaluations in unmatched data.

Furthermore, we noted that controlling for one form of length information may not necessarily account for the other (Supplementary Fig. 1A & B), as pathogenic and benign deletions with similar nucleotide spans might not overlap with a comparable number of genes, and vice versa. To further minimize the impact of length information, we generated a test set matched for both deletion length and the number of affected genes (n=618). This strategy resulted in AUC values of approximately 0.5 for both the length features (deletion length and gene number) in the matched test set and across different length groups (Fig. 2C & D; Supplementary Table 1 & 2), indicating that these two length features could no longer be used for prediction. Even though the overall performance of almost all methods was further reduced, DosaCNV remained the top-performing method (Fig. 2C). DosaCNV outperformed other methods in all length groups, except for the small group (Fig. 2D), where XCNV's performance was marginally better but not statistically significant (DeLong test p-value = 0.582). By comparing the AUC values of this matched test set to the unmatched test set, we estimated each method's degree of dependency on length information. CADD-SV demonstrated the greatest dependency with a 0.227 decrease in AUC (unmatched: 0.888, matched: 0.661), while TADA exhibited the smallest dependency with a 0.080 decrease in AUC

155 (unmatched: 0.923, matched: 0.843). DosaCNV had the second smallest dependency, with a 0.091
156 decrease in AUC (unmatched: 0.968, matched: 0.877).

157 **DosaCNV enhanced prioritization of putative pathogenic deletions in neurodevelopmental disorder**
158 **case-control studies.** To further assess DosaCNV's capacity to discriminate between pathogenic and
159 benign deletions, we compared its performance to alternative methods in prioritizing potential
160 pathogenic deletions from case-control studies of neurodevelopmental disorders (NDDs). While ground
161 truth labels for deletions in case-control datasets are unknown, prior research indicates that CNVs may
162 contribute to 15-25% of NDD cases[48,49]. As a result, an effective pathogenic CNV classifier should prioritize
163 these pathogenic CNVs, placing them within the top predictions among all CNVs from cases and controls,
164 with the top-ranked CNVs predominantly depleted in controls. Thus, we assessed the performance of each
165 method based on the enrichment of case-associated deletions within their top predictions, using this as a
166 measure of their ability to prioritize putative pathogenic deletions. Specifically, for the analysis of each
167 dataset, we established the number of top predictions for evaluation as 5% of the total case count
168 (referred to as the top 5% predictions). Based on the conservative assumption that coding CNVs contribute
169 to 10% of cases, with half of these CNVs being deletions.

170 We examined 16,048 coding deletions from Coe et al., 2014 (dbVar accession number: nstd100)[50],
171 comprising 29,085 intellectual disability, developmental delay, and autism spectrum disorder (ID/DD/ASD)
172 cases, alongside 19,584 controls. Subsequently, we evaluated 4,230 coding deletions from Zarrei et al.,
173 2019 (dbVar accession number: nstd173)[51], which included 2,691 NDD cases and 1,769 controls. Within
174 the nstd100 dataset, DosaCNV had the highest enrichment of case-associated deletions among its top 5%
175 predictions (odds ratio = 102.3), which significantly outperforms the second-best method HIS-LOD (odds
176 ratio = 31.1; z-test for log odds ratios p-value = 1.026×10^{-10}) (Fig. 3A; Supplementary Table 4). Specifically,
177 DosaCNV prioritized 84 more case-associated deletions among top predictions compared to HIS-LOD,
178 suggesting a potential to detect 5.8% (84/1,454) more putative pathogenic deletions. For the nstd173

179 dataset, DosaCNV again had the highest enrichment of case-associated deletions among its top 5%
180 predictions (odds ratio = 2.12) (Fig. 3B; Supplementary Table 4). However, this did not significantly differ
181 from the next best method TADA (odds ratio = 1.79; z-test for log odds ratios p-value = 0.562), likely
182 attributable to the smaller sample size.

183 Additionally, we investigated the overlap between the two case-control datasets and the ClinVar test set
184 to provide further evidence for DosaCNV's capacity to prioritize pathogenic deletions. We posited that
185 deletions with over 95% overlap should exhibit similar functional consequences and, therefore,
186 comparable pathogenicity. We identified deletions from the case-control datasets with over 95% overlap
187 with true pathogenic deletions from the ClinVar test set. Then, we assessed whether they were primarily
188 found in cases and prioritized as top predictions by DosaCNV. In the nstd100 dataset, we uncovered 164
189 overlapping deletions, with 154 observed in cases, substantiating the depletion of most likely pathogenic
190 deletions within controls. Furthermore, 148 of these 164 deletions were included in DosaCNV's nstd100
191 top predictions, demonstrating its robust prioritization capability. For the nstd173 dataset, we observed
192 13 deletions overlapping with true pathogenic deletions, with 11 of them present in cases. Only 4 of the
193 13 overlapping deletions were among the top 5% predictions of DosaCNV, possibly due to the smaller
194 sample size of the overlapping deletions. Notably, DosaCNV did not prioritize any deletions that
195 overlapped with benign ClinVar deletions as top predictions in either dataset.

196 Finally, we assessed DosaCNV's capacity to distinguish pathogenic deletions from benign deletions in
197 patients, a prevalent and challenging clinical scenario when identifying causal variants. We obtained 740
198 clinically curated coding deletions within 140 DD cases from Foong et al., 2015[52], which included 33
199 pathogenic deletions. DosaCNV effectively prioritized pathogenic deletions among benign deletions (Fig.
200 3C), attaining an AUC of 0.997, which was the highest among all methods (Supplementary Fig. 3). However,
201 the difference in AUC was not statistically significant compared to the second-best method StrVCTVRE
202 (AUC = 0.981; DeLong test p-value = 0.115). With a DosaCNV score threshold of 0.8, the true positive rate

203 reached 0.970, and the false positive rate was 0.028, indicating the accurate classification of almost all
204 pathogenic deletions with only a small fraction of benign deletions misclassified. The precision of DosaCNV
205 was also high (0.604) at this threshold. These results suggest that DosaCNV can accurately distinguish
206 pathogenic deletions from benign deletions in patients, and 0.8 is a reasonable threshold in the practice
207 of clinical variant interpretation.

208 **DosaCNV-HI outperforms alternative methods in predicting HI genes from multiple sources.** To
209 investigate whether the superior performance of DosaCNV at the deletion level is due to more accurate
210 gene haploinsufficiency estimation, we evaluated our model's ability to distinguish HI and likely HI genes
211 from haplosufficient (HS) genes. First, we assembled six positive sets of HI and likely HI genes, including:
212 1) 340 ClinGen "gold standard" HI genes[53], 2) 397 human orthologs of mouse heterozygous knockout
213 lethal (MHKO) genes[54], 3) 214 high-confidence SFARI ASD genes (SFARI gene score 1)[55], 4) 432 DDG2P
214 genes sensitive to heterozygous LOF mutations[56], 5) 72 eyeG2P genes sensitive to heterozygous LOF
215 mutations[56], and 6) 82 skinG2P genes sensitive to heterozygous LOF mutations[56]. We defined all other
216 genes as the HS gene set. To balance the number of genes and mitigate the impact of CDS length on model
217 performance, we used MatchIt[57] to match HI and HS genes by CDS length, retaining only matched genes
218 for downstream performance analysis.

219 To circumvent circularity, we retrained DosaCNV by excluding DDG2P and EDS scores from gene-level
220 input features, as a subset of DDG2P genes was used for gene-level prediction assessment, and the EDS
221 score was trained on a portion of our assembled HI genes. After training, the DosaCNV gene-level model,
222 DosaCNV-HI (Methods), generated HI probability predictions for 20,268 protein-coding genes.

223 We assessed DosaCNV-HI's performance against 12 alternative methods across six datasets,
224 encompassing two metrics for LOF mutation intolerance (DeepLOF[29], and LOEUF[28]), one for missense
225 mutation intolerance (Mis. OEUF[28]), two for both LOF and missense mutation intolerance (RVIS[26],

226 VIRLoF[58]), and seven for haploinsufficiency (HIS[18], IS[23], GHIS[19], HIPred[20], Episcore[21], EDS[22], pHaplo[24]). The
227 comparison included genes with predictions available from all methods. To evaluate performance robustly
228 within imbalanced sets, we computed the AUC for each method by sampling an equal number of HS genes
229 of matching CDS length as the HI or likely HI gene sets.

230 DosaCNV-HI significantly surpassed alternative methods in predicting ClinGen HI genes, achieving an AUC
231 of 0.901, which is substantially higher than the second-best method, Episcore, with an AUC of 0.854
232 (Delong test p-value = 0.019) (Fig. 4A; Supplementary Table 5). DosaCNV-HI consistently achieved high
233 rankings in the remaining five sets of likely HI genes. Notably, even in cases where DosaCNV-HI was not
234 the top performer (ranking second or third), the differences in the AUCs were not statistically significant
235 compared to those of the best methods (Fig. 4B-F; Supplementary Table 5). It is worth noting that the
236 performance of DosaCNV-HI could still be overly optimistic when analyzing all available genes without
237 considering overlap between HI/HS genes and deletions in the ClinVar training set. To address this, we
238 conducted a separate analysis excluding genes overlapped by deletions in the ClinVar training set, and
239 DosaCNV-HI maintained its high-ranking performance across all six gene sets (Supplementary Fig. 4;
240 Supplementary Table 6).

241 **Evaluating the robustness of DosaCNV-HI and alternative methods in identifying challenging HI and HS**
242 **genes.** Metrics based on evolutionary constraints, such as LOEUF and DeepLOF, demonstrated reasonable
243 performance, indicating that many HI genes are indeed subject to strong purifying selection in human
244 populations. However, these metrics, along with supervised HI gene scores dependent on them, may face
245 challenges in identifying HI genes in three particular cases: 1) short HI genes with limited statistical power
246 for constraint inference[27,28], 2) HI genes under low evolutionary constraint, including those involved in
247 late-onset disorders (e.g., *BRCA2*, *ATM*, and *ELN*)[33–36], and 3) falsely predicted HS genes under high
248 evolutionary constraint, such as pleiotropic genes with minor effects on single phenotypes but significant

249 cumulative fitness impacts[31,32]. To evaluate the robustness of our method and alternative approaches in
250 these scenarios, we generated three challenging gene sets.

251 First, we combined all six HI and likely HI gene sets to form a unified set of 988 genes, aiming to enhance
252 statistical power. From this combined set, we focused on the 758 genes that have predictions across all
253 methods for further analysis. Next, we filtered this set for genes with fewer than 10 expected LOF
254 mutations in human populations[29], creating a "short, likely HI" gene set (n=73). Second, using an
255 established LOEUF cutoff (<0.35 for constrained genes) and a similar cutoff for DeepLOF (>0.835 for
256 constrained genes), we generated a "low constraint, likely HI" gene set (n=278). Finally, we applied the
257 same cutoffs to identify a set of highly evolutionarily constrained HS genes not implicated in diseases
258 across G2P, Human Phenotype Ontology[59], ClinVar[42], and OMIM databases[60], yielding a "high constraint,
259 unlikely HI" gene set (n=209).

260 We identified the 2,402 top-ranked genes for each method as predicted HI genes, comparable to the
261 LOEUF cutoff of 0.35, and evaluated performance based on the percentage of each challenging gene set
262 included within the top-ranked genes. An ideal HI gene classifier should maximize the inclusion of genes
263 from the two likely HI gene sets while minimizing the inclusion of genes from the unlikely HI gene set.
264 Among all evaluated methods for the "short, likely HI" gene set, HIS exhibited the highest inclusion
265 percentage (0.466, 34/73), followed by Episcore (0.411, 30/73), and then our method, DosaCNV-HI (0.397,
266 29/73) (Fig. 4G; Supplementary Table 7). Notably, almost all methods showed a higher inclusion
267 percentage compared to LOEUF. This suggests that these methods, potentially due to the integration of
268 diverse complementary gene-level features, are less affected by the limited statistical power in detecting
269 short, constrained genes. For the "low constraint, likely HI" gene set, DosaCNV-HI and HIS achieved the
270 highest (0.439, 122/278) and second highest (0.414, 115/278) inclusion percentages, respectively (Fig. 4H;
271 Supplementary Table 7). This outcome may be attributed to the fact that these two methods directly
272 model haploinsufficiency instead of evolutionary constraints.

273 Remarkably, using gene CDS length alone as a predictor resulted in the lowest inclusion percentage (0.072,
274 15/209) for the "high constraint, unlikely HI" gene set (Fig. 4I; Supplementary Table 7). In contrast,
275 DosaCNV-HI exhibited the second lowest inclusion percentage (0.148, 31/209), with CDS length excluding
276 7.6% more unlikely HI genes (Fig. 4I). This observation emphasizes a positive correlation between gene
277 haploinsufficiency and CDS length for evolutionarily constrained genes, aligning with prior research that
278 suggests longer genes tend to be more crucial during early developmental stages[61], with
279 haploinsufficiency potentially serving as an inherent property of these genes[62]. In summary, while
280 DosaCNV-HI may not always be the top-performing method, it consistently ranks highly for predicting
281 challenging HI genes and difficult HS genes.

282 **Elucidating key features and their contributions to gene haploinsufficiency prediction.** We applied Deep
283 SHAP[63] (SHapley Additive exPlanations) alongside DosaCNV-HI to identify key features determining a
284 gene's probability of being HI. Based on coalition game theory, Deep SHAP fairly and uniquely allocates
285 an input's prediction to its features (SHAP values) by computing a weighted average of marginal
286 contributions across all feature combinations. We obtained SHAP values for each feature across 988 HI
287 and likely HI genes using Deep SHAP and sampled 1,000 HS genes as a background set. We then visualized
288 these values using a default SHAP summary plot[63], offering a global explanation for the entire HI and likely
289 HI gene set (Fig. 5A). The global feature importance is derived from the mean absolute SHAP values over
290 all HI and likely HI genes. Pathogenic SNV presence emerged as the most important feature, as genes
291 sensitive to disrupted SNVs may also be sensitive to dosage changes. Features linked to evolutionary
292 constraints and conservation, such as PLI, LOEUF, and PhastCons, also ranked highly, emphasizing the
293 importance of evolutionary constraints and conservation as common HI gene characteristics (Fig. 5A).

294 Notably, we observed a positive correlation between the second most important feature, the number of
295 transcript isoforms, and the DosaCNV-HI prediction (Fig. 5A). This finding is in line with a previous study
296 showing that genes with a greater number of isoforms are more likely to be essential genes[64]. The sixth

297 most important feature, tissue specificity (tau), also exhibited a positive correlation with the DosaCNV-HI
298 prediction (Fig. 5A), which implies that HI genes may have greater tissue specificity in expression than HS
299 genes[18]. Epigenetically, higher DosaCNV-HI predictions were associated with stronger signals for
300 repressive histone modifications (H3K27me3) and weaker signals for active modifications (H3K4me3 and
301 H3K9ac) (Fig. 5A). Considering that the majority of epigenetic data were obtained from adult
302 tissues/organs and most HI genes are developmental genes[15], this result suggests that many HI genes may
303 be epigenetically silenced outside early developmental stages. Intriguingly, we observed a positive
304 correlation between the H2A.Z signal and DosaCNV-HI predictions, although this histone variant is known
305 to facilitate both gene activation and repression[65]. Additionally, several features, such as a greater number
306 of enhancers, increased promoter CpG density, more extensive protein-protein interactions, elevated
307 mean expression levels, and fewer paralogs, contributed to higher DosaCNV-HI predictions (Fig. 5A),
308 aligning with previous studies[18,21,66].

309 Deep SHAP also offers local explanations by attributing the HI prediction to individual feature
310 contributions for a specific gene. To demonstrate this, we retrieved coordinates of five recurrent
311 pathogenic microdeletions from ClinGen[67], and obtained DosaCNV-HI predictions and corresponding
312 SHAP values for 30 protein-coding genes overlapping with these microdeletions (Fig. 5B). DosaCNV-HI
313 identified candidate disease genes with plausible haploinsufficiency mechanisms, consistent with previous
314 findings, such as *CHD1L* and *GJA8* for the 1q21.1 deletion[68,69], *CHRNA7* for the 15q13.3 deletion[70,71],
315 *PMP22* for the 17p12 deletion[72], *CRHR1* and *KANSL1* for the 17q21.31 deletion[73,74], and *SHANK3* for the
316 22q13 deletion[75] (Fig. 5B). Comparing the SHAP values of all features for these candidate genes allows us
317 to understand why they are predicted to have a higher chance of being HI. For example, in the well-
318 documented HI gene *PMP22*, encoding the essential peripheral myelin protein 22 for myelin formation
319 and stability[72], key contributors to its HI prediction revealed by SHAP values include pathogenic SNVs
320 presence, numerous transcript isoforms, evolutionary constraints, enhancer abundance, and extensive

321 protein interactions (Fig. 5B). These features highlight *PMP22*'s regulatory complexity within the myelin
322 gene network, which implicates the necessity of precise dosage control for its role in proper neural
323 development and function[72]. Similarly, *KANSL1*, another established HI gene involved in chromatin
324 modification and gene regulation[74], demonstrated a comparable pattern of feature importance, though
325 adapted to its unique biological functions (Fig. 5B). These insights elucidate the rationale behind
326 DosaCNV's predictions for each gene and deletion, ultimately improving our capacity to explore the
327 biological mechanisms underpinning gene haploinsufficiency and deletion pathogenicity.

328 **Discussion**

329 In this study, we present DosaCNV, a novel deep MIL model that jointly predicts coding deletion
330 pathogenicity and gene haploinsufficiency. Prior research has examined these predictions independently,
331 without effectively connecting the two. Our approach, for the first time, bridges this gap in a biologically
332 coherent manner, which significantly enhances predictions at both the gene and deletion levels. Moreover,
333 this method enables the evaluation of HI genes' contribution to deletion pathogenicity and expands the
334 spectrum of learned HI gene features. Additionally, we employ rigorous training and data preprocessing
335 strategies, such as size matching and leave-one-chromosome-out, to strengthen model generalizability
336 and mitigate biases arising from information leakage[76].

337 While DosaCNV demonstrated superior performance to other prediction tools in ClinVar data, the results
338 also revealed a notable challenge shared by all methods, including DosaCNV: the dependence of
339 predictions on deletion length. While a strong positive association exists between deletion length and
340 pathogenicity, it is not the primary determinant of pathogenicity. Overreliance on this non-causal
341 association may lead to biased predictions for long benign deletions and short pathogenic deletions[43].
342 Furthermore, as machine-learning models can easily capture this strong correlation and maintain high
343 accuracy within unmatched datasets, the impact of other features directly involved in the underlying

344 mechanisms of pathogenic deletions might be overshadowed. This emphasizes the importance of taking
345 the length information into account during both model development and performance evaluation.

346 A robust pathogenic CNV classifier must be generalizable to effectively prioritize putative pathogenic
347 deletions across various disease contexts, addressing the growing accumulation of unlabeled CNV data.
348 DosaCNV, trained on a limited set of pan-disorder deletions, exhibited superior performance in prioritizing
349 putative pathogenic deletions in ID/DD/ASD cases, as evidenced by the highest enrichment of case-
350 associated deletions in predicted pathogenic deletions and its favorable sensitivity in prioritizing clinically
351 curated pathogenic deletions in a real clinical scenario. This generalizability supports DosaCNV's
352 underlying hypothesis that the pathogenicity of coding deletions predominantly results from affected HI
353 genes, which may share cross-disease commonalities captured by the gene-level features used in the
354 model. However, the performance may also be influenced by the enrichment of ID/DD/ASD patients in
355 our pan-disorder training data. While HI genes often share certain characteristics, their effects may also
356 differ in a disease-dependent manner based on their involvement in distinct biological processes. To
357 account for the variation of gene effects across diseases, DosaCNV is designed for adaptability, enabling
358 easy training with diverse datasets and gene-level features tailored to the specific disease context under
359 investigation.

360 Data scarcity in predicting gene haploinsufficiency poses a significant challenge for supervised machine-
361 learning methods, with fewer than 1,000 high-confidence HI and likely HI genes available out of an
362 estimated ~2,900[27,28]. DosaCNV adopts an innovative model design, learning predictive features for
363 haploinsufficiency from affected genes within curated deletions, without requiring gene-level labels. This
364 approach enables a more comprehensive capture of the HI gene landscape, allowing DosaCNV-HI to
365 outperform alternative methods in six HI and likely HI gene sets. However, our model remains dependent
366 on a limited number of clinically curated deletions, which may be subject to ascertainment bias, such as
367 the enrichment of certain types of genes and deletions[45]. Consequently, developing larger and less biased

368 datasets of curated CNVs is crucial for improving novel HI gene discovery and pathogenic deletion
369 prediction.

370 Evolutionary constraint metrics, such as pLI and LOEUF, have effectively prioritized disease-associated
371 genes based on the idea that deleterious mutations in disease genes are often depleted through purifying
372 selection. While these methods primarily focus on fitness effects, they can also suggest haploinsufficiency,
373 a phenotypic effect, especially for HI genes involved in early developmental processes where carrying a
374 null allele can result in reduced fitness. However, exceptions exist where reduced fitness does not align
375 with haploinsufficiency. For example, mutations in late-onset HI genes may not significantly reduce fitness,
376 and mutations in pleiotropic genes may lead to reduced fitness without causing discernible phenotypic
377 changes[32]. Moreover, evolutionary constraint metrics based on population genetics data are known to be
378 underpowered for short genes with sparse polymorphisms[27–29]. To address these challenges,
379 incorporating gene features orthogonal to evolutionary constraint metrics is crucial. By integrating over
380 30 gene-level features from diverse biological processes, DosaCNV-HI has demonstrated promising
381 performance in scenarios that pose challenges for traditional evolutionary constraint metrics (Fig. 4G, H,
382 I), underscoring the importance of considering a broader array of gene features to enhance the accuracy
383 and reliability of haploinsufficiency predictions.

384 To address the limited interpretability of deep learning models, we integrated DosaCNV-HI with a model-
385 specific explanation tool, Deep SHAP[63], to elucidate key gene features that highlighted the biological
386 characteristics of HI genes (Fig. 5A). Several metrics for evolutionary constraints, such as pLI and LOEUF,
387 emerged as top features, reinforcing the idea that HI gene disruption frequently leads to developmental
388 disorders and is thus subject to strong purifying selection[15,62]. Epigenetically, higher DosaCNV-HI
389 predictions correlated with weaker signals for active histone modification marks (H3K4me3, H3K9ac) and
390 stronger signals for repressive marks (H3K27me3). This supports the hypothesis that a substantial number
391 of HI genes function as developmental transcription factors, which are critical during specific

392 developmental stages and become epigenetically silenced outside these periods[77,78]. Additionally, our
393 results revealed a positive correlation between DosaCNV-HI predictions and H2A.Z signal, which can be
394 attributed to H2A.Z's prevalent occurrence in bivalent domains within promoters of developmental
395 genes[65], where HI genes are notably enriched. Regarding regulatory elements, higher enhancer numbers
396 and promoter CpG densities were associated with increased DosaCNV-HI predictions. These
397 characteristics have been associated with essential developmental genes[21,66,79,80], enabling intricate
398 spatiotemporal expression patterns associated with hypothesized mechanisms of haploinsufficiency[81].

399 Certain limitations in our study highlight the need for further refinement to improve the accuracy and
400 comprehensiveness of our method. First, we focused on coding regions in the current study, but the
401 potential contributions of non-coding regions to CNV pathogenicity should not be overlooked[82,83]. Ignoring
402 non-coding regions constrains the interpretation of non-coding CNVs and may impair coding CNV
403 predictions by erroneously assigning pathogenicity to seemingly innocuous coding elements. Second,
404 DosaCNV's assumption that independent effects of HI genes are the sole contributors to deletion
405 pathogenicity oversimplifies the complex interplay between genomic elements. This simplification fails to
406 account for the epistasis between affected genomic elements and variable penetrance/expressivity
407 arising from other genetic variants, such as SNVs[84–86]. Lastly, we assumed heterozygosity for all deletions
408 in the training and testing sets due to the absence of zygosity information in ClinVar. This assumption is
409 partially supported by the overrepresentation of dominant disease genes and limited overlap with
410 gnomAD benign SVs in ClinVar pathogenic deletions[45]. However, we observed a few pathogenic deletions
411 affecting only single recessive disease genes, suggesting the potential existence of homozygous deletions.
412 Consequently, the accuracy of our HI gene and pathogenic deletion predictions could be impacted, given
413 that DosaCNV was specifically designed to predict haploinsufficiency which is a dominant effect. The
414 future availability of fully phased CNV data will be instrumental in facilitating a more comprehensive

415 understanding of the relationships between the mode of inheritance of disease genes and the
416 pathogenicity of CNVs.

417 Despite the exclusive focus on deletions in this study, our proposed framework can be suitably adapted
418 to predict the pathogenicity of duplications. This adaptation essentially involves shifting the focus of gene-
419 level prediction from haploinsufficiency to triplosensitivity, achievable by simply utilizing curated
420 duplications as training data. Despite maintaining the same framework structure, a reassessment of
421 predictive features is needed, given the potential discrepancies between HI and TS genes. The integration
422 of appropriate input features will further elucidate the unique characteristics of TS genes.

423 **Conclusions**

424 In conclusion, DosaCNV is an accurate and robust deep multiple-instance learning framework that
425 simultaneously predicts deletion pathogenicity and gene haploinsufficiency. Traditional methods have
426 often treated these domains in isolation. DosaCNV, however, offers a unified approach to bridge this gap,
427 enhancing predictions at both the deletion and gene levels evidenced in multiple datasets. Furthermore,
428 by integrating the Deep SHAP technique to quantify gene-level feature contributions, DosaCNV not only
429 provides transparent and human-understandable explanations for its predictions but also holds promise
430 in further elucidating the intricate landscape of gene haploinsufficiency. However, challenges persist,
431 notably potential biases in clinically curated CNV datasets, with the ClinVar dataset being an example.
432 Despite our rigorous efforts during model development and evaluation, these biases inevitably limit our
433 model's precision and overall applicability. As our understanding of previously underexplored genomic
434 regions, such as non-coding regions, expands in tandem with rapid advancements in sequencing
435 technology, we anticipate the emergence of more comprehensive and unbiased CNV catalogs and
436 supporting resources. Such developments will be pivotal in improving the accuracy and reliability of tools
437 like DosaCNV, while also expanding their range of applications.

Methods

Details of DosaCNV. We denote $\mathbf{X}_i$ as the input for the i-th deletion, which is a matrix of dimensions M x L, where M represents the number of genes and L corresponds to the number of gene-level features. Given that the number of genes varies for each deletion, each $\mathbf{X}_i$ is automatically padded to accommodate 100 genes if M < 100 by adding dummy genes with all feature values set to zero. Consequently, $\mathbf{X}_i$ assumes a shape of 100 x L. In the fully connected-layer, we aim to model HI probabilities for each of the M genes within the i-th deletion, utilizing their respective gene-level feature vectors (rows of $\mathbf{X}_i$). To achieve this, we employ two hidden layers for feature representation learning from $\mathbf{X}_i$. The calculations for these layers can be represented by the following equations:

$$\mathbf{Z}_i^{(1)} = \mathrm{ReLU}(\mathbf{X}_i\mathbf{W}^{(1)} + \mathbf{b}^{(1)}),$$

$$\mathbf{Z}_i^{(2)} = \mathrm{Dropout}(\mathrm{ReLU}(\mathbf{Z}_i^{(1)}\mathbf{W}^{(2)} + \mathbf{b}^{(2)})),$$

where superscripts (1) and (2) denote the 1st and 2nd hidden layers, respectively. ReLU and Dropout refer to the rectified linear layer[87] and dropout layer[88], while $\mathbf{W}$ and $\mathbf{b}$ represent the weight matrix and bias vector, respectively. $\mathbf{Z}_i$ serves as the matrix of hidden units for the i-th deletion, with each row constituting the hidden unit vector for a gene. Subsequently, we incorporate an additional layer to convert $\mathbf{Z}^{(2)}$ into HI probabilities ($\mathbf{P}_{\mathrm{HI}}$) for the genes:

$$\mathbf{P}_{\mathrm{HI},i} = \sigma(\mathbf{Z}_i^{(2)}\mathbf{W}^{(3)} + \mathrm{b}^{(3)}),$$

where σ denotes the logistic function, which ensures that the HI probabilities remain within the range of 0 and 1. $\mathbf{W}^{(3)}$ and $\mathrm{b}^{(3)}$ represent the weight vector and bias term, respectively. $\mathbf{P}_{\mathrm{HI},i}$ is a vector of length 100, wherein each element serves as the gene-level prediction of haploinsufficiency (DosaCNV-HI score) for a gene or dummy gene within the i-th deletion. To naturally aggregate probabilities within $\mathbf{P}_{\mathrm{HI},i}$ and infer deletion pathogenicity, we introduce a pooling layer that employs the noisy-or function:

460 $$\hat{y}_i = 1 - \prod_{j=1}^{100} (1 - \mathrm{P}_{\mathrm{HI},i,j} \times \mathrm{Mask}_{i,j}),$$

461 where $\hat{y}_i$ is the predicted probability of the i-th deletion harboring at least one HI gene and serves as a
462 proxy for deletion pathogenicity. $\mathrm{P}_{\mathrm{HI},i,j}$ denotes the P_{HI} for the j-th gene or dummy gene within the i-th
463 deletion. $\mathrm{Mask}_{i,j}$ is a binary indicator, specifying whether the j-th gene is an actual gene ($\mathrm{Mask}_{i,j} = 1$) or
464 a dummy gene ($\mathrm{Mask}_{i,j} = 0$). This mechanism prevents the propagation of information from dummy
465 genes by setting $\mathrm{P}_{\mathrm{HI},i,j}$ to zero.

466 We use the mean binary cross entropy loss to measure the discrepancy between the labels of deletion
467 pathogenicity and predicted pathogenicity probabilities:

468 $$\mathrm{loss}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)],$$

469 where $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ represent the vectors of the observed pathogenicity labels and predicted pathogenicity
470 probabilities in a mini-batch of data of size N, respectively.

471 For model training, we initialize all parameters using He Normal initialization[89] and employ the Adam
472 algorithm[90] for mini-batch gradient descent to update model parameters. To mitigate overfitting, we
473 implement early stopping and incorporate an L2 regularization term.

474 Following the training process, the full model is utilized for variant-level prediction of deletion
475 pathogenicity. Concurrently, all layers of the full model are extracted, excluding the noisy-or pooling layer,
476 to establish a model for gene-level prediction of haploinsufficiency, referred to as DosaCNV-HI.

477 **ClinVar training, validation, and test set.** We obtained the clinically curated rare SVs dataset (GRCh37;
478 dbVar accession number: nstd102) from ClinVar[42] on May 25, 2021. We filtered for copy number
479 losses/deletions meeting specific criteria: (1) >49 bp in length, (2) not somatic origin, (3) at least 1 bp
480 overlap with a CDS, (4) <101 affected protein-coding genes, and (5) clinical interpretations as pathogenic,
481 likely pathogenic, pathogenic/likely pathogenic, benign, likely benign, or benign/likely benign. Inner start

482 and end coordinates were also required for inclusion. We retrieved coordinates for 20,268 protein-coding
483 genes and their corresponding CDSs from the GENCODE[91] v38 human basic gene annotation file in GRCh37
484 and used BEDTools[92] to determine overlaps with deletions.

485 To minimize overfitting due to potential ascertainment bias, we identified redundant deletion groups with
486 95% reciprocal overlap using BEDTools and retained only the shortest deletion within each group. This
487 process yielded 12,002 deletions, which were labeled as pathogenic (1) if they had clinical interpretations
488 as pathogenic, likely pathogenic, or pathogenic/likely pathogenic and benign (0) if they had clinical
489 interpretations as benign, likely benign, or benign/likely benign (6,180 pathogenic and 5,822 benign).

490 To further minimize overfitting and avoid overly optimistic evaluations due to partial overlaps between
491 deletions in the same chromosome, we adopted a "leave-one-chromosome-out" strategy to partition the
492 deletions into training, validation, and testing sets. The test set included deletions from chromosomes 1,
493 3, 5, and 7, while the validation set comprised deletions from chromosomes 2, 4, 6, and 8, and the training
494 set contained deletions from the remaining chromosomes. This approach produced 6,678, 2,492, and
495 2,832 deletions for the training, validation, and testing sets.

496 We used R package MatchIt[57] to match pathogenic and benign deletions by the number of affected
497 protein-coding genes in the training and validation sets; only matched pairs were kept. We left the test
498 set unmatched for further analysis of the length impact on performance evaluation. This approach
499 resulted in 2,308 and 1,060 deletions in the final training and validation sets, respectively.

500 **DosaCNV gene-level features.** DosaCNV leverages 33 gene-level features of 20,268 protein-coding genes
501 to predict gene haploinsufficiency and subsequently infer deletion pathogenicity. The gene-level features
502 can be grouped into six main categories: genomics, epigenomics, evolutionary constraints, functional
503 genomics, deleteriousness upon perturbation, and biological networks (Supplementary Data 1). (1)
504 Genomics: We used the Ensembl biomaRt R package[93] to extract the number of coding exons and

505 transcript isoforms for each gene. We sourced promoter CpG density from a previous study[66]. We
506 retrieved a list of 3,543 human duplicated genes, organized into 945 groups, from the Duplicated Gene
507 Database (DGD)[94]. We determined the number of paralogs for each gene based on the group size of the
508 corresponding gene group and assigned a paralog count of zero for genes absent in the list. (2)
509 Epigenomics: We obtained ChIP-seq broad peaks for H2A.Z, H3K4me3, H3K9ac, and H3K27me3 in
510 promoter regions across cell types, as well as promoter-enhancer interactions predicted by EpiTensor[95]
511 from a previous study[21]. We calculated the average length of the corresponding ChIP-seq peak in a gene's
512 promoter region across all cell types to define input features for these epigenomic marks. Additionally,
513 we determined the number of enhancers associated with a gene by averaging the number of predicted
514 promoter-enhancer interactions across cell types. (3) Evolutionary constraint: We extracted lof_z[25],
515 mis_z[25], pLI[27], LOEUF[28], and Mis. OEUF[28] from the gnomAD database[28] v2.1.1. We retrieved SNP-level
516 PhastCons100way vertebrate and PhastCons30way mammalian scores[96] from the dbNSFP database[97] v4,
517 and averaged these two metrics across the entire span of each gene to obtain gene-level scores. We
518 sourced promoter and exonic PhastCons100way vertebrate scores from a previous study[66] and obtained
519 sHet values from the original publication[98]. (4) Functional genomics: We sourced GTEx mean expression[99],
520 GO terms for embryo development[100], and tissue specificity (tau)[101] values from a recent study[29]. We
521 retrieved median gene-level transcripts per million (TPMs) across 54 tissues from the v7 release of the
522 GTEx portal[99]. We counted the number of non-zero expression tissues to determine the number of tissues
523 expressed for each gene. (5) Deleteriousness upon perturbation: We retrieved SNP-level MetaLR[102], CADD
524 raw[103], and DANN[104] scores from the dbNSFP database[97], and calculated the average scores across all
525 coding exons of each gene. We downloaded a list of 2,545 DDG2P genes from the G2P database[56] and
526 generated a binary feature to indicate whether a gene is included in this list. We sourced EDS values from
527 the original publication[22], obtained CRISPR HAP1 cell knockout q-values from a previous study[105], and
528 downloaded SNV data from the ClinVar[42]. We created a binary feature indicating the presence or absence

529 of pathogenic SNVs in a gene, based on ClinVar SNV data. Additionally, for each gene, we calculated the
530 percentage of pathogenic SNVs relative to the total number of submitted SNVs in ClinVar. (6) Biological
531 networks: We downloaded a dataset of 64,006 binary protein-protein interactions involving 9,094
532 proteins from the Human Reference Protein Interactome Mapping Project[106]. We counted the number of
533 interaction partners for each gene included in the dataset and assigned a count of zero for genes absent
534 in the dataset. We obtained PPI degree values from a recent study[29].

535 We performed mean imputation for missing values and z-standardized all features, excluding binary ones,
536 based on the means and standard deviations derived from genes not overlapping deletions in the ClinVar
537 held-out test set.

538 **Model training.** We mapped 33 gene-level features to each affected gene within deletions, and padded
539 deletions with fewer than 100 genes using dummy genes with all zero feature values within the ClinVar
540 training, validation, and test sets. This process resulted in an input matrix for each deletion with a shape
541 of 100 x 33. We then trained DosaCNV using the ClinVar training set and tuned hyperparameters via grid
542 search based on the ClinVar validation set. The search space included: the number of nodes in each hidden
543 ReLU layer (32, 64, 128, 256); lambda for the L2 penalty (0.006, 0.02, 0.06, 0.2); batch size (16, 32, 64);
544 and Adam learning rate ($3x10^{-5}$, 10^{-4}, $3x10^{-4}$, 10^{-3}). The optimal hyperparameters based on the lowest
545 validation loss were 128 and 64 nodes for the first and second hidden ReLU layers, 0.02 for the L2 penalty,
546 a batch size of 32, and a learning rate of $3x10^{-4}$. Additionally, we set the dropout rate to 0.5[88].

547 For gene-level prediction using DosaCNV-HI, we excluded EDS and DDG2P from the gene-level features to
548 avoid potential circularity. DosaCNV was trained on the same ClinVar training set using the same set of
549 optimal hyperparameters from the DosaCNV model for deletion-level prediction.

550 **Deletion-level evaluation.** We obtained deletion-level predictions for comparison from five alternative
551 methods: (1) TADA v1.0.2, available at (https://github.com/jakob-he/TADA), using default settings on the

552 GRCh37 assembly; (2) StrVCTVRE v1.7, accessed from (https://github.com/andrewSharo/StrVCTVRE),
553 employing default settings with all analyzed deletions lifted to GRCh38 using the UCSC liftOver tool; (3)
554 HIS-LOD, in which we retrieved gene-level predictions (HIS) for all genes from the original publication and
555 aggregated them to deletions-level predictions based on the method described therein[18]; (4) X-CNV,
556 downloaded from (https://github.com/kbvstmd/XCNV), with default settings on the GRCh37 assembly;
557 and (5) CADD-SV v1.1 webtool (https://cadd-sv.bihealth.org/score), using default settings on the GRCh37
558 assembly.

559 To evaluate performance on the ClinVar held-out test set (n=2,832) and assess the impact of length
560 distribution differences, we compiled a joint set of 2,491 predictions from all five alternative methods,
561 denoting it as the unmatched test set. Using MatchIt, we matched pathogenic and benign deletions in the
562 unmatched test set based on the number of affected protein-coding genes (n=702), deletion length in
563 base pairs (n=704), and both criteria combined (n=618), retaining only matched pairs. To examine
564 performance on deletions of varying lengths, we further divided each of the four sets into four length
565 groups: small (<50 kb), medium (50-100 kb), large (100-1000 kb), and extra-large (>1000 kb). We obtained
566 AUC values and corresponding standard errors for each method on each set using the R package pROC[107].
567 We visualized the AUC values for all methods using R package ggplot2[108] and tested the statistical
568 significance for the differences between AUC values using the DeLong test, via the roc.test function in
569 pROC. To estimate each method's dependency on length information, we subtracted the AUC value of the
570 matched test set from that of the unmatched test set, considering both deletion length and the number
571 of affected genes.

572 We evaluated the prioritization ability for putative pathogenic deletions by obtaining rare SVs from case-
573 control studies by Coe et al., 2014[50] (nstd100) with 29,085 ID/DD/ASD cases and 19,584 controls, and
574 Zarrei et al., 2019[51] (nstd173) with 2,691 NDD cases and 1,769 family member controls. We applied the
575 same filtering conditions with the ClinVar data, excluding the condition for clinical interpretation. The

576 resulting nstd100 and nstd173 sets contained 16,528 and 4,500 coding deletions, respectively, with
577 16,048 and 4,230 deletions having predictions from all methods. Since these datasets lack pathogenicity
578 labels, we calculated the enrichment of case-associated deletions in each method's top predictions using
579 odds ratio as the performance measure. For each dataset, we defined the number of top predictions for
580 evaluation as 5% of the total case count (referred to as the top 5% predictions), assuming that coding
581 CNVs contribute to 10% of cases, half of which are deletions. This approach resulted in 1,454 and 134 top
582 predictions for nstd100 and nstd173, respectively. We visualized the odds ratios by ggplot2 and tested
583 the statistical significance for the differences between odds ratios by transforming them to log odds ratios
584 and then using the z-test.

585 To further assess our method, we determined the inclusion percentage of the nstd100 and nstd173
586 deletions that had over 95% reciprocal overlap with curated pathogenic deletions from ClinVar, within our
587 method's top predictions. Specifically, we found 164 nstd100 deletions and 13 nstd173 deletions had over
588 95% reciprocal overlap with pathogenic deletions in the ClinVar held-out test set.

589 To evaluate performance in a clinical scenario, we acquired clinically curated rare SVs from 140
590 developmental delay cases reported by Foong et al., 2015[52]. After filtering, 740 coding deletions, including
591 33 pathogenic ones, remained. We generated predictions for these deletions and visualized them in a
592 scatter plot using ggplot2. Additionally, 663 of the 740 deletions had predictions from all methods. Based on
593 the joint set of 663 deletions, we determined the AUC values for our method and the alternative methods
594 using pROC.

595 **Gene-level evaluation.** To evaluate the performance of DosaCNV-HI and 12 alternative methods—
596 including DeepLOF[29], LOEUF[28], Mis.OEUF[28], RVIS[26], VIRLoF[58], HIS[18], IS[23], GHIS[19], HIPred[20], Episcore[21], EDS[22],
597 and pHaplo[24]—in predicting HI and likely HI genes, we obtained LOEUF and Mis.OEUF from the gnomAD
598 database v2.1.1 and gene scores for all other methods from their respective original publications.

599 We retrieved six HI and likely HI gene sets for evaluation: (1) 340 "gold standard" HI genes with sufficient
600 evidence for haploinsufficiency (ClinGen HI score 3) from the ClinGen gene dosage sensitivity map[53]; (2)
601 397 human orthologs of mouse heterozygous knockout lethal genes from the GitHub repository for
602 gnomAD (https://github.com/macarthur-lab/gnomAD_lof/); (3) 214 high-confidence ASD genes (SFARI
603 gene score 1) from the SFARI GENE database[55]; (4) 432 DDG2P genes with the monoallelic autosomal
604 allelic requirement and absent gene product mutation consequence from the G2P database[56]; (5) 72
605 eyeG2P genes with the same requirements as DDG2P genes from the G2P database; and (6) 82 skinG2P
606 genes with the same requirements as DDG2P genes from the G2P database. We conducted analyses using
607 genes with predictions from all methods, resulting in 210, 315, 176, 360, 59, and 76 HI/likely HI genes
608 remaining within the six gene lists, respectively. All other protein-coding genes were considered as the HS
609 gene set.

610 To mitigate the imbalance between the six HI or likely HI gene sets and the HS gene set, and to control
611 the influence of gene length on performance evaluation, we adopted a matching approach. For each HI
612 or likely HI gene set, we used MatchIt to pair each HI gene with an HS gene of a similar CDS length to
613 obtain a matched gene set. Subsequently, we utilized pROC to compute the AUC values and corresponding
614 standard errors for each method within the matched gene sets. We visualized the AUC values for all
615 methods using ggplot2 and tested the statistical significance for the differences between AUC values using
616 the DeLong test.

617 For the evaluation of challenging HI and HS genes, we aggregated the six gene sets into a single HI and
618 likely HI gene list (n=988). From genes with predictions across all methods (n=758), we generated a "short,
619 likely HI" gene set (n=73) by filtering genes with fewer than 10 expected LOF mutations from the gnomAD
620 database. Using established LOEUF (<0.35 for constrained genes)[28] and DeepLOF (>0.835 for constrained
621 genes)[29] cutoffs for constrained genes, we created a "low constraint, likely HI" gene set (n=278). Applying
622 the same cutoffs, we generated a "high constraint, unlikely HI" gene set (n=209) for constrained HS genes

not implicated in diseases across the G2P, Human Phenotype Ontology[59], ClinVar, and OMIM databases[60]. Subsequently, we defined the top 2,402 genes predicted by each method as the top predictions for each method (based on cutoff: LOEUF <0.35), representing the predicted HI genes. We calculated the inclusion percentages with corresponding standard errors of the three challenging gene sets in each method's top predictions as the performance measure and visualized these inclusion percentages using ggplot2.

Feature importance. To elucidate the key features driving DosaCNV-HI gene-level predictions, we employed the DeepExplainer function from the SHAP v0.41.0 Python library[63]. We generated a background set, consisting of 31 gene-level features for 1,000 randomly selected HS genes, resulting in a 1,000 x 31 matrix where each row represents a gene-level feature vector for a specific gene. The explainer was initialized with the DosaCNV-HI model and the defined background set, allowing the calculation of conditional expectations for each gene-level feature and the corresponding conditional expectation of HI predictions for the 1,000 randomly sampled HS genes. Subsequently, we generated an input set containing 31 gene-level features for 988 HI and likely HI genes, which we fed into the explainer to obtain the SHAP values. The output was a 988 x 31 SHAP value matrix, where each row represents the SHAP values for corresponding gene-level feature values of a particular gene. We used the default bee swarm summary plot from the package to visualize the global importance and impact distributions of gene-level features. The features on the y-axis were ranked in descending order of global importance, which the package calculated as the mean absolute value for the corresponding column in the output SHAP value matrix.

We employed the same approach to generate SHAP values for 30 protein-coding genes overlapping pathogenic microdeletions at 1q21.1 (distal, BP3-BP4), 15q13.3 (BP4-BP5), 17p12 (HNPP), 17q21.31, and 22q13, with coordinates retrieved from the ClinGen database in the GRCh37 assembly. We visualized the SHAP values and percentiles of DosaCNV-HI predictions for these affected genes as a heatmap using ggplot2.

647 **Availability of data and materials**

648 The DosaCNV program and the precomputed DosaCNV-HI gene scores can be accessed at
649 https://github.com/Zhihan-Leo-Liu/DosaCNV. Datasets used in this study, along with the source code for
650 analysis can be found at https://github.com/Zhihan-Leo-Liu/DosaCNV/analysis/. Additionally, the sources
651 and download links for all gene-level features are provided in Supplementary Data 1.

652 **Acknowledgments**

653 The authors thank Mingfu Shao and Xinru Zhang for helpful discussions.

654 **Funding**

655 Research reported in this publication was supported by the National Institute of General Medical Sciences
656 of the National Institutes of Health under Award Number R35GM142560 and by startup funds from
657 Pennsylvania State University. The content is solely the responsibility of the author and does not
658 necessarily represent the official views of the National Institutes of Health.

659 **Author information**

660 **Authors and Affiliations**

661 Department of Biology, Pennsylvania State University, University Park, PA 16802, USA
662 Zhihan Liu & Yi-Fei Huang

663 Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, PA 16802, USA
664 Zhihan Liu & Yi-Fei Huang

665 Molecular, Cellular, and Integrative Biosciences Program, Pennsylvania State University, University Park,
666 PA 16802, USA
667 Zhihan Liu

668 **Contributions**

669 ZL implemented the DosaCNV program and conducted all analyses. YH conceived and supervised the
670 project. ZL and YH prepared and reviewed the manuscript.

671 **Corresponding authors**

672 Correspondence to Zhihan Liu or Yi-Fei Huang.

673 **Ethics declarations**

674 **Ethics approval and consent to participate**

675 Not applicable

676 **Consent for publication**

677 Not applicable

678 **Competing interests**

679 The authors declare no competing interests.

680 **Additional materials**

681 **Additional file 1**

682 Additional file 1. docx

683 Supplementary Figures 1-4 and Supplementary Tables 1-7

684 **Additional file 2**

685 Additional file 2.xls

686 The sources and download links for all gene-level features used in this study. Referred to as
687 Supplementary Data 1 in the main text.

688 **References**

689 1. Zhang, F., Gu, W., Hurles, M. E. & Lupski, J. R. Copy Number Variation in Human Health, Disease, and
690 Evolution. *Annu. Rev. Genomics Hum. Genet.* **10**, 451–481 (2009).
691 2. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome.
692 *Nature* **464**, 704–712 (2010).
693 3. Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number.
694 *Nat. Rev. Genet.* **10**, 551–564 (2009).
695 4. Liu, P., Carvalho, C. M. B., Hastings, P. J. & Lupski, J. R. Mechanisms for recurrent and complex
696 human genomic rearrangements. *Curr. Opin. Genet. Dev.* **22**, 211–220 (2012).
697 5. Perry, G. H. *et al.* Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.*
698 **39**, 1256–1260 (2007).
699 6. Leffler, E. M. *et al.* Resistance to malaria through structural variation of red blood cell invasion
700 receptors. *Science* **356**, eaam6393 (2017).
701 7. Malhotra, D. & Sebat, J. CNVs: Harbingers of a Rare Variant Revolution in Psychiatric Genetics. *Cell*
702 **148**, 1223–1241 (2012).
703 8. McCarroll, S. A. *et al.* Deletion polymorphism upstream of IRGM associated with altered IRGM
704 expression and Crohn's disease. *Nat. Genet.* **40**, 1107–1112 (2008).
705 9. McCarroll, S. A. *et al.* Donor-recipient mismatch for common gene deletion polymorphisms in graft-
706 versus-host disease. *Nat. Genet.* **41**, 1341–1344 (2009).
707 10. Cooper, G. M. *et al.* A copy number variation morbidity map of developmental delay. *Nat. Genet.*
708 **43**, 838–846 (2011).
709 11. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders.
710 *Nature* **466**, 368–372 (2010).

711 12. McCarthy, S. E. *et al.* Microduplications of 16p11.2 are associated with schizophrenia. *Nat. Genet.*
712 **41**, 1223–1227 (2009).
713 13. Shao, X. *et al.* Copy number variation is highly correlated with differential gene expression: a pan-
714 cancer study. *BMC Med. Genet.* **20**, 175 (2019).
715 14. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural
716 variation: insights from and for human disease. *Nat. Rev. Genet.* **14**, 125–138 (2013).
717 15. Rice, A. M. & McLysaght, A. Dosage sensitivity is a major determinant of human copy number
718 variant pathogenicity. *Nat. Commun.* **8**, 14366 (2017).
719 16. Seidman, J. G. & Seidman, C. Transcription factor haploinsufficiency: when half a loaf is not enough.
720 *J. Clin. Invest.* **109**, 451–455 (2002).
721 17. Veitia, R. A. & Potier, M. C. Gene dosage imbalances: action, reaction, and models. *Trends Biochem.*
722 *Sci.* **40**, 309–317 (2015).
723 18. Huang, N., Lee, I., Marcotte, E. M. & Hurles, M. E. Characterising and Predicting Haploinsufficiency in
724 the Human Genome. *PLOS Genet.* **6**, e1001154 (2010).
725 19. Steinberg, J., Honti, F., Meader, S. & Webber, C. Haploinsufficiency predictions without study bias.
726 *Nucleic Acids Res.* **43**, e101 (2015).
727 20. Shihab, H. A., Rogers, M. F., Campbell, C. & Gaunt, T. R. HIPred: an integrative approach to
728 predicting haploinsufficient genes. *Bioinformatics* **33**, 1751–1757 (2017).
729 21. Han, X. *et al.* Distinct epigenomic patterns are associated with haploinsufficiency and predict risk
730 genes of developmental disorders. *Nat. Commun.* **9**, 2138 (2018).
731 22. Wang, X. & Goldstein, D. B. Enhancer Domains Predict Gene Pathogenicity and Inform Gene
732 Discovery in Complex Disease. *Am. J. Hum. Genet.* **106**, 215–233 (2020).
733 23. Khurana, E., Fu, Y., Chen, J. & Gerstein, M. Interpretation of Genomic Variants Using a Unified
734 Biological Network Approach. *PLOS Comput. Biol.* **9**, e1002886 (2013).

735 24. Collins, R. L. *et al.* A cross-disorder dosage sensitivity map of the human genome. *Cell* **185**, 3041-
736 3055.e25 (2022).
737 25. Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat.*
738 *Genet.* **46**, 944–950 (2014).
739 26. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic Intolerance to Functional
740 Variation and the Interpretation of Personal Genomes. *PLoS Genet.* **9**, e1003709 (2013).
741 27. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291
742 (2016).
743 28. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456
744 humans. *Nature* **581**, 434–443 (2020).
745 29. LaPolice, T. M. & Huang, Y.-F. *A deep learning framework for predicting human essential genes from*
746 *population and functional genomic data*. http://biorxiv.org/lookup/doi/10.1101/2021.12.21.473690
747 (2021) doi:10.1101/2021.12.21.473690.
748 30. Ruderfer, D. M. *et al.* Patterns of genic intolerance of rare copy number variation in 59,898 human
749 exomes. *Nat. Genet.* **48**, 1107–1111 (2016).
750 31. Simons, Y. B., Bullaughey, K., Hudson, R. R. & Sella, G. A population genetic interpretation of GWAS
751 findings for human quantitative traits. *PLOS Biol.* **16**, e2002985 (2018).
752 32. Fuller, Z. L., Berg, J. J., Mostafavi, H., Sella, G. & Przeworski, M. Measuring intolerance to mutation in
753 human genetics. *Nat. Genet.* **51**, 772–776 (2019).
754 33. Wooster, R. *et al.* Identification of the breast cancer susceptibility gene BRCA2. *Nature* **378**, 789–792
755 (1995).
756 34. Roy, R., Chun, J. & Powell, S. N. BRCA1 and BRCA2: different roles in a common pathway of genome
757 protection. *Nat. Rev. Cancer* **12**, 68–78 (2012).

758 35. Raybould, M. C., Birley, A. J. & Hultén, M. Molecular variation of the human elastin (ELN) gene in a
759 normal human population. *Ann. Hum. Genet.* **59**, 149–161 (1995).
760 36. Lu, S. *et al.* Atm-haploinsufficiency enhances susceptibility to carcinogen-induced mammary tumors.
761 *Carcinogenesis* **27**, 848–855 (2006).
762 37. Ganel, L., Abel, H. J., FinMetSeq Consortium & Hall, I. M. SVScore: an impact prediction tool for
763 structural variation. *Bioinformatics* **33**, 1083–1085 (2017).
764 38. Geoffroy, V. *et al.* AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics*
765 **34**, 3572–3574 (2018).
766 39. Riggs, E. R. *et al.* Technical standards for the interpretation and reporting of constitutional copy
767 number variants: a joint consensus recommendation of the American College of Medical Genetics
768 and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). *Genet. Med. Off. J. Am. Coll.*
769 *Med. Genet.* **22**, 245–257 (2020).
770 40. Amores, J. Multiple instance classification: Review, taxonomy and comparative study. *Artif. Intell.*
771 **201**, 81–105 (2013).
772 41. Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. (Morgan
773 Kaufmann, 1988).
774 42. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence.
775 *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
776 43. Skelly, A. C., Dettori, J. R. & Brodt, E. D. Assessing bias: the importance of considering confounding.
777 *Evid.-Based Spine-Care J.* **3**, 9–12 (2012).
778 44. Hertzberg, J., Mundlos, S., Vingron, M. & Gallone, G. TADA—a machine learning tool for functional
779 annotation-based prioritisation of pathogenic CNVs. *Genome Biol.* **23**, 67 (2022).

780 45. Sharo, A. G., Hu, Z., Sunyaev, S. R. & Brenner, S. E. StrVCTVRE: A supervised learning method to
781 predict the pathogenicity of human genome structural variants. *Am. J. Hum. Genet.* **109**, 195–209
782 (2022).
783 46. Zhang, L. *et al.* X-CNV: genome-wide prediction of the pathogenicity of copy number variations.
784 *Genome Med.* **13**, 132 (2021).
785 47. Kleinert, P. & Kircher, M. A framework to score the effects of structural variants in health and
786 disease. *Genome Res.* gr.275995.121 (2022) doi:10.1101/gr.275995.121.
787 48. Girirajan, S. & Eichler, E. E. Phenotypic variability and genetic susceptibility to genomic disorders.
788 *Hum. Mol. Genet.* **19**, R176–R187 (2010).
789 49. Miller, D. T. *et al.* Consensus Statement: Chromosomal Microarray Is a First-Tier Clinical Diagnostic
790 Test for Individuals with Developmental Disabilities or Congenital Anomalies. *Am. J. Hum. Genet.* **86**,
791 749–764 (2010).
792 50. Coe, B. P. *et al.* Refining analyses of copy number variation identifies specific genes associated with
793 developmental delay. *Nat. Genet.* **46**, 1063–1071 (2014).
794 51. Zarrei, M. *et al.* A large data resource of genomic copy number variation across
795 neurodevelopmental disorders. *NPJ Genomic Med.* **4**, 26 (2019).
796 52. Foong, J., Girdea, M., Stavropoulos, J. & Brudno, M. Prioritizing Clinically Relevant Copy Number
797 Variation from Genetic Interactions and Gene Function Data. *PLOS ONE* **10**, e0139656 (2015).
798 53. Riggs, E. R. *et al.* Copy number variant discrepancy resolution using the ClinGen dosage sensitivity
799 map results in updated clinical interpretations in ClinVar. *Hum. Mutat.* **39**, 1650–1659 (2018).
800 54. Bult, C. J., Blake, J. A., Smith, C. L., Kadin, J. A. & Richardson, J. E. Mouse Genome Database (MGD)
801 2019. *Nucleic Acids Res.* **47**, D801–D806 (2019).
802 55. Abrahams, B. S. *et al.* SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum
803 disorders (ASDs). *Mol. Autism* **4**, 36 (2013).

804 56. Thormann, A. *et al.* Flexible and scalable diagnostic filtering of genomic variants using G2P with
805 Ensembl VEP. *Nat. Commun.* **10**, 2373 (2019).
806 57. Ho, D., Imai, K., King, G. & Stuart, E. A. MatchIt: Nonparametric Preprocessing for Parametric Causal
807 Inference. *J. Stat. Softw.* **42**, 1–28 (2011).
808 58. Abramovs, N., Brass, A. & Tassabehji, M. GeVIR is a continuous gene-level metric that uses variant
809 distribution patterns to prioritize disease candidate genes. *Nat. Genet.* **52**, 35–39 (2020).
810 59. Köhler, S. *et al.* The Human Phenotype Ontology in 2021. *Nucleic Acids Res.* **49**, D1207–D1217
811 (2020).
812 60. Amberger, J., Bocchini, C. A., Scott, A. F. & Hamosh, A. McKusick's Online Mendelian Inheritance in
813 Man (OMIM®). *Nucleic Acids Res.* **37**, D793–D796 (2009).
814 61. Lopes, I., Altab, G., Raina, P. & de Magalhães, J. P. Gene Size Matters: An Analysis of Gene Length in
815 the Human Genome. *Front. Genet.* **12**, (2021).
816 62. Zug, R. Developmental disorders caused by haploinsufficiency of transcriptional regulators: a
817 perspective based on cell fate determination. *Biol. Open* **11**, bio058896 (2022).
818 63. Lundberg, S. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. Preprint at
819 https://doi.org/10.48550/arXiv.1705.07874 (2017).
820 64. Yong Ryu, J., Uk Kim, H. & Yup Lee, S. Human genes with a greater number of transcript variants
821 tend to show biological features of housekeeping and essential genes. *Mol. Biosyst.* **11**, 2798–2807
822 (2015).
823 65. Ku, M. *et al.* H2A.Z landscapes and dual modifications in pluripotent and multipotent stem cells
824 underlie complex genome regulatory functions. *Genome Biol.* **13**, R85 (2012).
825 66. Boukas, L., Bjornsson, H. T. & Hansen, K. D. Promoter CpG Density Predicts Downstream Gene Loss-
826 of-Function Intolerance. *Am. J. Hum. Genet.* **107**, 487–498 (2020).
827 67. Rehm, H. L. *et al.* ClinGen — The Clinical Genome Resource. *N. Engl. J. Med.* **372**, 2235–2242 (2015).

828 68. Harvard, C. *et al.* Understanding the impact of 1q21.1 copy number variant. *Orphanet J. Rare Dis.* **6**,
829 54 (2011).
830 69. Prokudin, I. *et al.* Exome sequencing in developmental eye disease leads to identification of causal
831 variants in GJA8, CRYGC, PAX6 and CYP1B1. *Eur. J. Hum. Genet.* **22**, 907–915 (2014).
832 70. Hoppman-Chaney, N., Wain, K., Seger, P., Superneau, D. & Hodge, J. Identification of single gene
833 deletions at 15q13.3: further evidence that CHRNA7 causes the 15q13.3 microdeletion syndrome
834 phenotype. *Clin. Genet.* **83**, 345–351 (2013).
835 71. Gillentine, M. & Schaaf, C. P. The Human Clinical Phenotypes of Altered CHRNA7 Copy Number.
836 *Biochem. Pharmacol.* **97**, 352–362 (2015).
837 72. Li, J., Parker, B., Martyn, C., Natarajan, C. & Guo, J. The PMP22 Gene and Its Related Diseases. *Mol.*
838 *Neurobiol.* **47**, 673–698 (2013).
839 73. Wray, C. D. 17q21.31 Microdeletion associated with infantile spasms. *Eur. J. Med. Genet.* **56**, 59–61
840 (2013).
841 74. Koolen, D. A. *et al.* Mutations in the chromatin modifier gene KANSL1 cause the 17q21.31
842 microdeletion syndrome. *Nat. Genet.* **44**, 639–641 (2012).
843 75. Bozdagi, O. *et al.* Haploinsufficiency of the autism-associated Shank3 gene leads to deficits in
844 synaptic function, social interaction, and social communication. *Mol. Autism* **1**, 15 (2010).
845 76. Whalen, S., Schreiber, J., Noble, W. S. & Pollard, K. S. Navigating the pitfalls of applying machine
846 learning in genomics. *Nat. Rev. Genet.* **23**, 169–181 (2022).
847 77. Creyghton, M. P. *et al.* H2AZ Is Enriched at Polycomb Complex Target Genes in ES Cells and Is
848 Necessary for Lineage Commitment. *Cell* **135**, 649–661 (2008).
849 78. Yang, X. *et al.* Silencing of developmental genes by H3K27me3 and DNA methylation reflects the
850 discrepant plasticity of embryonic and extraembryonic lineages. *Cell Res.* **28**, 593–596 (2018).

851 79. Frankel, N. *et al.* Phenotypic robustness conferred by apparently redundant transcriptional
852 enhancers. *Nature* **466**, 490–493 (2010).
853 80. Norris, M., Lovell, S. & Delneri, D. Characterization and Prediction of Haploinsufficiency Using
854 Systems-Level Gene Properties in Yeast. *G3 GenesGenomesGenetics* **3**, 1965–1977 (2013).
855 81. Johnson, A. F., Nguyen, H. T. & Veitia, R. A. Causes and effects of haploinsufficiency. *Biol. Rev.* **94**,
856 1774–1785 (2019).
857 82. Spielmann, M., Lupiáñez, D. G. & Mundlos, S. Structural variation in the 3D genome. *Nat. Rev.*
858 *Genet.* **19**, 453–467 (2018).
859 83. Bhartiya, D. & Scaria, V. Genomic variations in non-coding RNAs: Structure, function and regulation.
860 *Genomics* **107**, 59–68 (2016).
861 84. Carvalho, C. M. B. *et al.* Dosage Changes of a Segment at 17p13.1 Lead to Intellectual Disability and
862 Microcephaly as a Result of Complex Genetic Interaction of Multiple Genes. *Am. J. Hum. Genet.* **95**,
863 565–578 (2014).
864 85. Girirajan, S. *et al.* Phenotypic Heterogeneity of Genomic Disorders and Rare Copy-Number Variants.
865 *N. Engl. J. Med.* **367**, 1321–1331 (2012).
866 86. Davies, R. W. *et al.* Using common genetic variation to examine phenotypic expression and risk
867 prediction in 22q11.2 deletion syndrome. *Nat. Med.* **26**, 1912–1918 (2020).
868 87. Nair, V. & Hinton, G. E. Rectified Linear Units Improve Restricted Boltzmann Machines.
869 88. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A Simple Way to
870 Prevent Neural Networks from Overfitting.
871 89. He, K., Zhang, X., Ren, S. & Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level
872 Performance on ImageNet Classification. Preprint at http://arxiv.org/abs/1502.01852 (2015).
873 90. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. Preprint at
874 http://arxiv.org/abs/1412.6980 (2017).

875 91. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids*
876 *Res.* **47**, D766–D773 (2019).
877 92. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features.
878 *Bioinformatics* **26**, 841–842 (2010).
879 93. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of
880 genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).
881 94. Ouedraogo, M. *et al.* The Duplicated Genes Database: Identification and Functional Annotation of
882 Co-Localised Duplicated Genes across Genomes. *PLoS ONE* **7**, e50653 (2012).
883 95. Zhu, Y. *et al.* Constructing 3D interaction maps from 1D epigenomes. *Nat. Commun.* **7**, 10812 (2016).
884 96. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.
885 *Genome Res.* **15**, 1034–1050 (2005).
886 97. Liu, X., Li, C., Mou, C., Dong, Y. & Tu, Y. dbNSFP v4: a comprehensive database of transcript-specific
887 functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome*
888 *Med.* **12**, 103 (2020).
889 98. Cassa, C. A. *et al.* Estimating the Selective Effects of Heterozygous Protein Truncating Variants from
890 Human Exome Data. *Nat. Genet.* **49**, 806–810 (2017).
891 99. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–
892 1330 (2020).
893 100. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338
894 (2019).
895 101. Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships
896 in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).
897 102. Dong, C. *et al.* Comparison and integration of deleteriousness prediction methods for
898 nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.* **24**, 2125–2137 (2015).

899 103. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human
900 genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
901 104. Quang, D., Chen, Y. & Xie, X. DANN: a deep learning approach for annotating the pathogenicity
902 of genetic variants. *Bioinformatics* **31**, 761–763 (2015).
903 105. Blomen, V. A. *et al.* Gene essentiality and synthetic lethality in haploid human cells. *Science* **350**,
904 1092–1096 (2015).
905 106. Luck, K. *et al.* A reference map of the human binary protein interactome. *Nature* **580**, 402–408
906 (2020).
907 107. Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves.
908 *BMC Bioinformatics* **12**, 77 (2011).
909 108. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer, 2009). doi:10.1007/978-0-
910 387-98141-3.

Figures

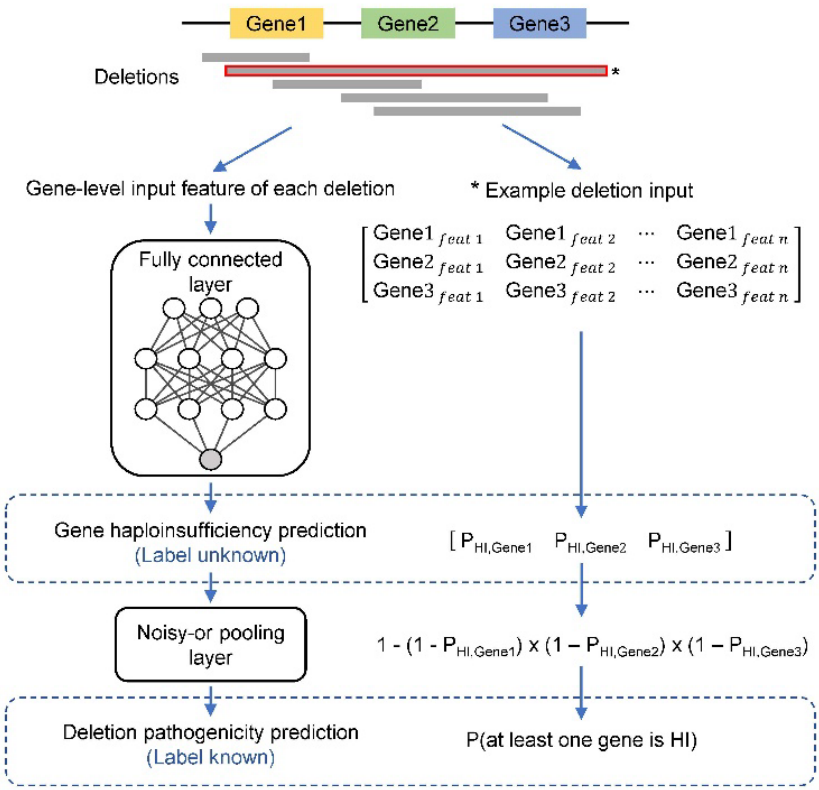

Figure 1. Schematic representation of the DosaCNV model. DosaCNV utilizes a fully connected architecture in conjunction with a noisy-or pooling layer to jointly infer gene haploinsufficiency and coding deletion pathogenicity. Deletions are first annotated with gene-level features of affected genes as input. For a given deletion, the fully connected architecture processes its gene-level input features and predicts the HI probability (P_{HI}) for each affected gene. Subsequently, the noisy-or pooling layer integrates these gene-level predictions, assuming independent gene effects, to compute the probability of the deletion harboring at least one HI gene, which serves as a proxy for deletion-level pathogenicity. The right side of the panel illustrates an example deletion input, along with its intermediate gene-level predictions and the noisy-or pooling calculation.

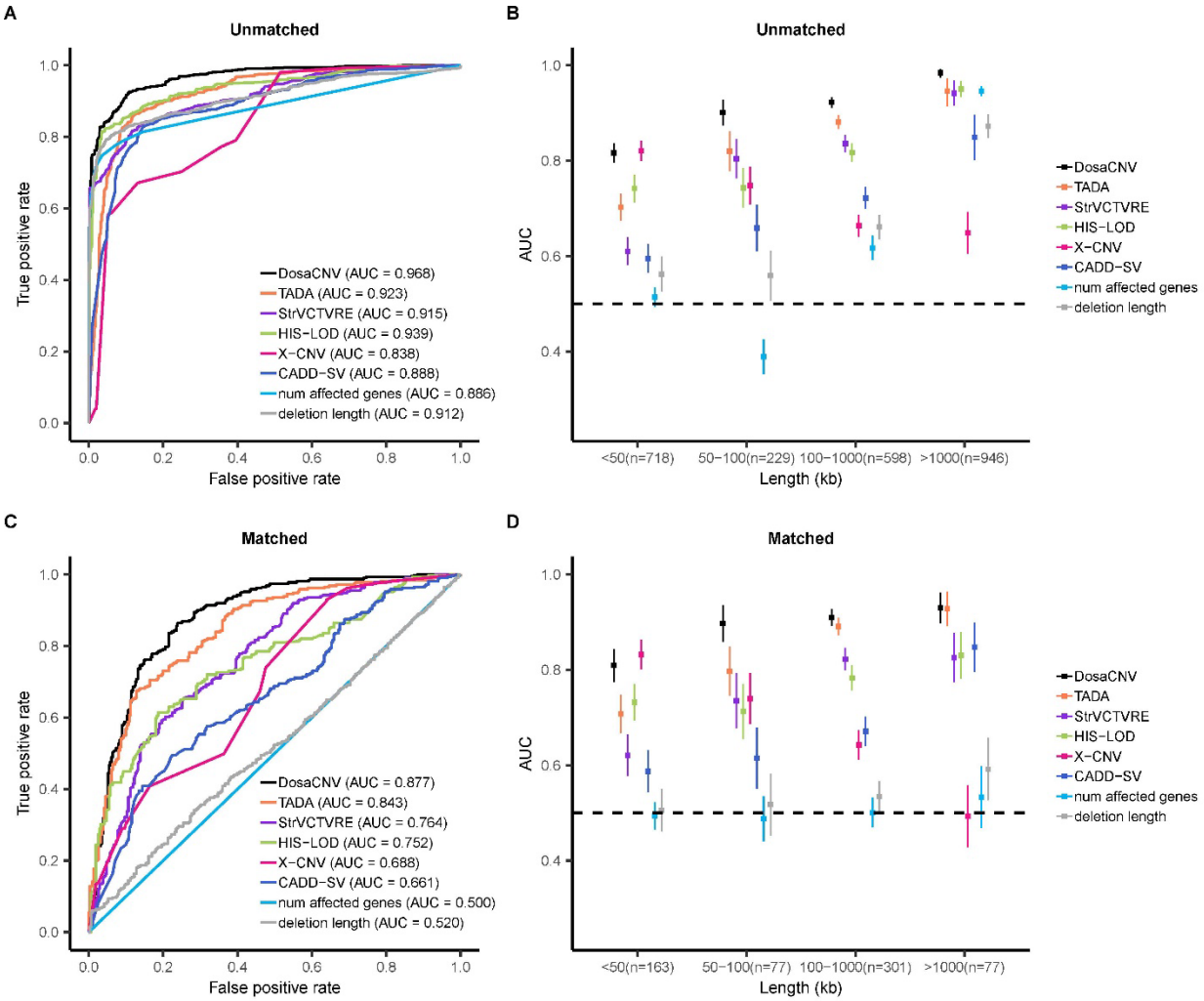

Figure 2. Performance comparison of DosaCNV and alternative methods in predicting pathogenicity of deletions from the unmatched and matched ClinVar held-out test sets. For each set, comparisons are presented for the entire set and four deletion subgroups based on length: small (<50 kb), medium (50-100 kb), large (100-1000 kb), and extra-large (>1000 kb). (A)(B) Unmatched test set (n=2,491). (C)(D) Test set matched by both the number of affected genes and deletion length (n=618). Vertical lines denote ± one standard error. Horizontal dashed lines indicate an AUC of 0.5.

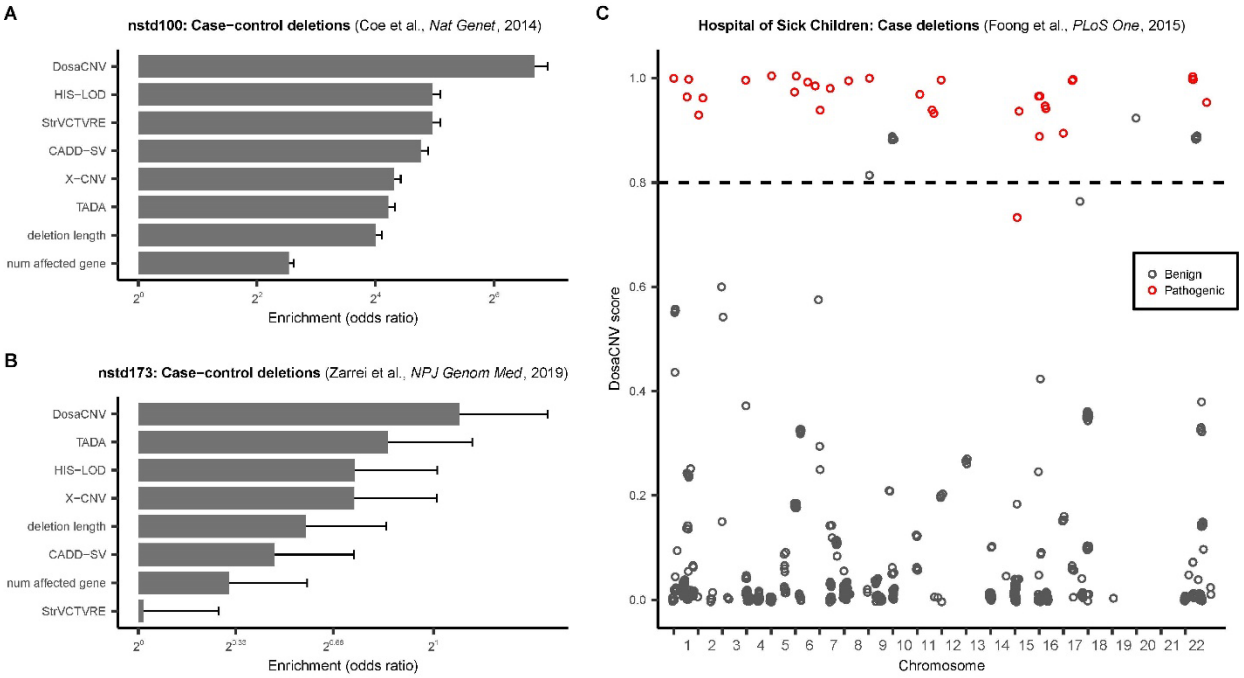

Figure 3. Comparative assessment of DosaCNV and alternative methods for prioritizing potential pathogenic deletions in case-control and curated case-only datasets. (A)(B) Evaluation of prioritization performance in case-control nstd100 (16,048 deletions; 29,085 ID/DD/ASD cases and 19,584 population controls) and nstd173 (4,230 deletions; 2,691 NDD cases and 1,769 family member controls) datasets. The enrichment of case-associated deletions within the top 5% predictions (1,454 for nstd100 and 134 for nstd173) for each method was used as the performance metric. Error bars represent one standard error. (C) DosaCNV scores for 740 clinically curated deletions in a case-only dataset from Foong et al., 2015. A DosaCNV score threshold of 0.8 (indicated by the horizontal dashed line) yields a true positive rate of 0.970 and a false positive rate of 0.028.

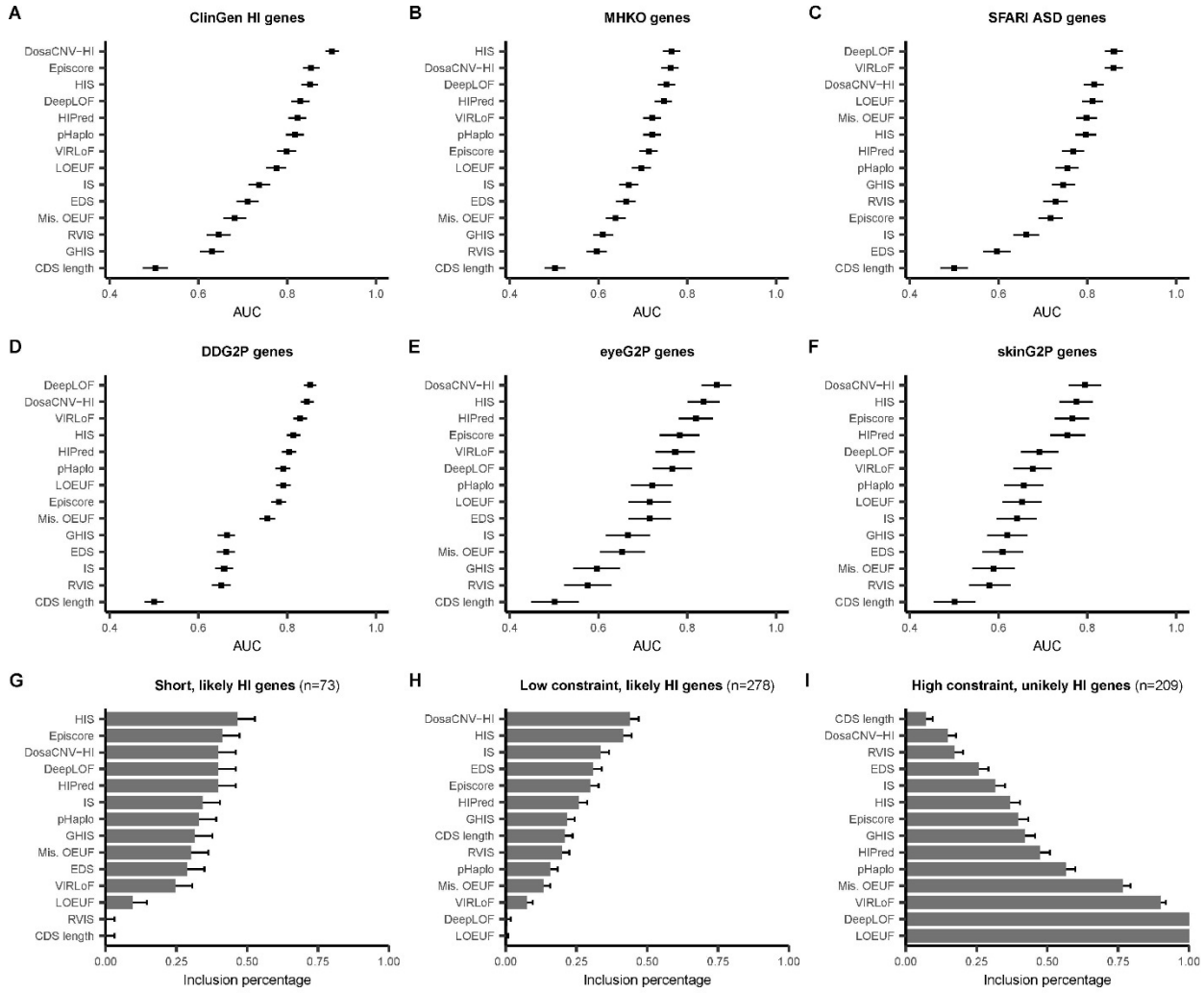

Figure 4. Predictive power of DosaCNV-HI and 12 alternative methods in discriminating HI and likely HI genes from HS genes. AUC values for DosaCNV-HI and alternatives are shown in predicting (A) ClinGen HI genes (n=210/340; 210 out of 340 genes had predictions from all methods), (B) human orthologs of mouse heterozygous knock-out lethal genes (n=315/397), (C) high confidence SFARI ASD genes (n=176/214), (D) DDG2P genes sensitive to heterozygous LOF mutations (n=366/432), (E) eyeG2P genes sensitive to heterozygous LOF mutations (n=59/72), (F) skinG2P genes sensitive to heterozygous LOF mutations (n=76/82). Horizontal black lines denote ± one standard error. Inclusion percentages for (G) "short, likely HI", (H) "low constraint, likely HI", and (I) "high constraint, unlikely HI" gene sets within the top 2,402 genes ranked by each method; a higher percentage for "likely HI" sets or a lower percentage for "unlikely HI" sets suggests superior performance. Error bars depict one standard error.

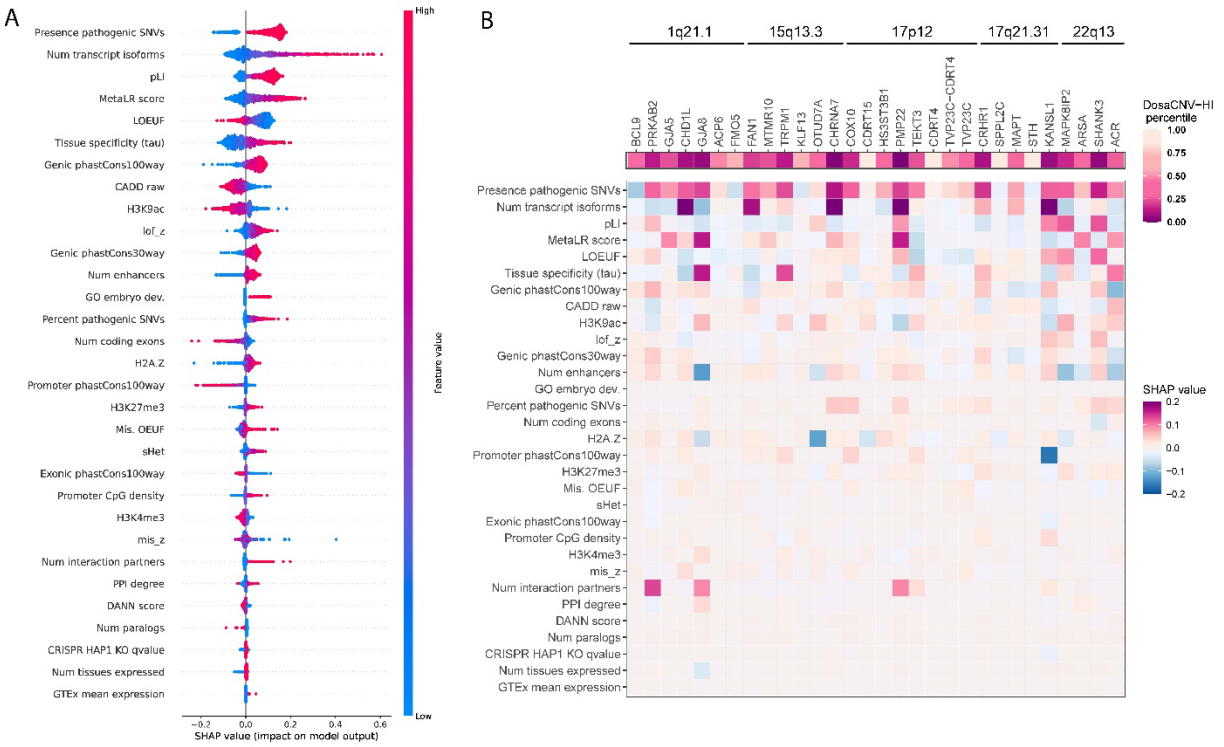

Figure 5. Elucidating gene-level feature contributions to haploinsufficiency predictions via SHAP values. (A) The SHAP summary plot is presented with gene-level features on the y-axis, arranged by global importance (mean absolute SHAP values) in descending order. Each dot represents a specific feature value for a gene within the HI or likely HI gene set (n=988), and the color denotes the actual feature value (red: high; blue: low). The x-axis displays SHAP values, which indicate the impact of each feature on the deviation between a gene's DosaCNV-HI prediction and the average DosaCNV-HI prediction for background HS genes (n=1000). Positive SHAP values enhance the DosaCNV-HI prediction (genes more likely to be HI), whereas negative values diminish it (genes less likely to be HI). The sum of SHAP values across all features for each gene equals the difference between its predicted HI probability and the expected HI probability of a designated background gene set (HS genes in our case). (B) DosaCNV-HI percentiles are shown for genes overlapping five pathogenic microdeletions, along with their corresponding SHAP values. A gene with a lower DosaCNV-HI percentile is more likely to be HI.