A unified benchmarking framework for gene prioritization across feature representations and integration strategies

Ziyu Zhang[1], Fatemeh Ghorbani[2], Ole Christian Lingjærde[1], Pooya Zakeri[1] [3]
[1]Department of Informatics, University of Oslo, Norway [2]School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Iran
[3]Integreat, Norwegian Centre for knowledge-driven machine learning, University of Oslo, Norway

Gene prioritization methods integrate heterogeneous biological data, but comparisons are complicated by differences in features, evaluation protocols, and leakage control. We present a unified, time-aware benchmark across 48 diseases and 11 ICD-10 categories, comparing gene representations, machine-learning models, and fusion strategies. Classical methods such as SVMs remain competitive with deep neural networks, while feature choice strongly affects performance. Intermediate fusion, especially geometric kernel fusion, provides the most consistent gains in ranking and early retrieval, whereas late fusion can improve biological coherence. We also introduce BioSim, an enrichment-based metric for assessing functional similarity between predicted and known disease genes. Our benchmark provides practical guidance for leakage-aware evaluation, feature integration, and model selection in gene prioritization.

Benchmark Design

Current limitations

- **Leakage control:** Disease-related signals may be embedded in features hidden, leading to overly optimistic performance.
- **Evaluation metrics:** Current metrics e.g. AUROC does not reflect early retrieval, Recall@k depends on arbitrary cutoffs. Neither directly evaluates the biological relevance of top-ranked candidates.
- **Disease dependence:** Related diseases may share genes and mechanisms, so treating all disease tasks as independent can bias average performance estimates.

Benchmark pipeline

- Uses a time-based data split to better reflect realistic disease-gene discovery.
- Combines ranking- and biology-oriented metrics for fairer and more reproducible comparisons.
- Evaluates feature representations, two base learning models, and multi-level integration strategies within a controlled setting.

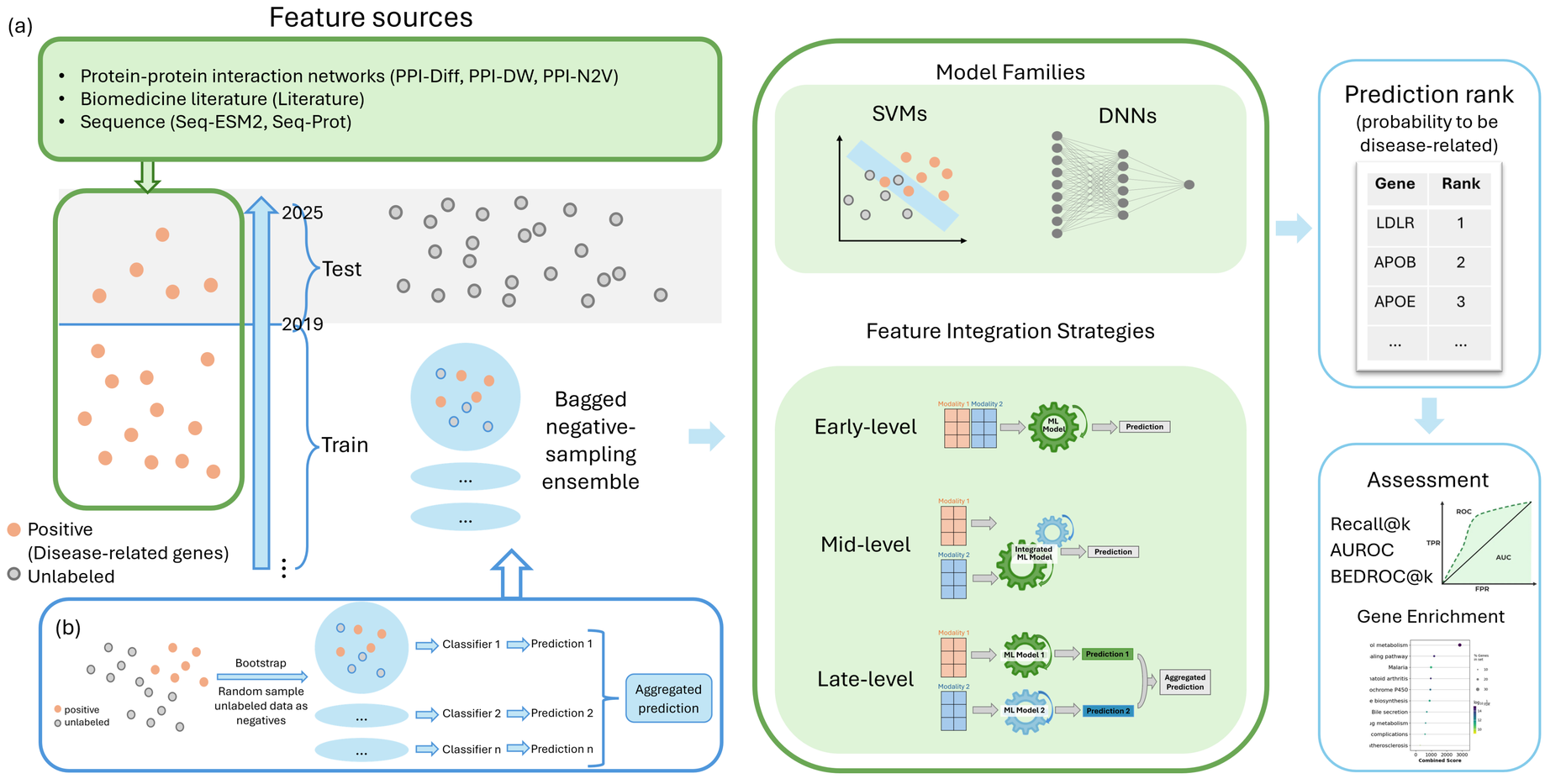
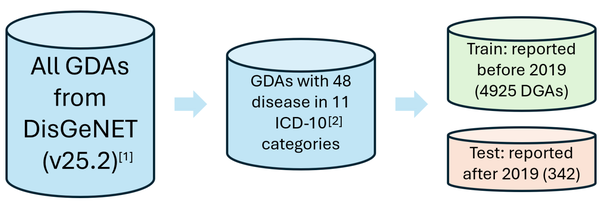


Fig 1. Gene prioritization benchmark pipeline. (a) Features are collected from three source types: PPI networks, biomedical literature, and sequence; disease-gene associations are split by a 2019 cutoff into training and test sets. During training, unlabeled genes are handled by a bagged negative-sampling ensemble, which repeatedly samples unlabeled genes as negatives. Two model families are evaluated: SVMs and DNNs. For multi-source integration, three fusion strategies are compared: early-level fusion concatenates features before training, intermediate fusion integrates features within the model, and late-level fusion aggregates predictions from separately trained models. Gene rankings are assessed using ranking-based metrics (Recall@k, AUROC, BEDROC@k) and gene enrichment metric BioSim@k. (b) The bagged negative-sampling ensemble in detail: unlabeled genes are repeatedly sampled as negatives across bootstrap iterations, producing one classifier per iteration, and predictions are aggregated into a final ranking score.

Data and processing

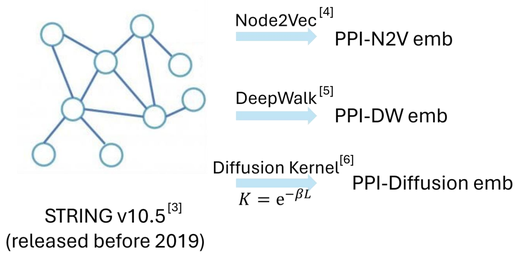
Gene-disease associations



Disease inclusion criteria
≥5 training genes & ≥1 testing gene

Gene-level features

Time dependent features



Time independent features

Feature analysis

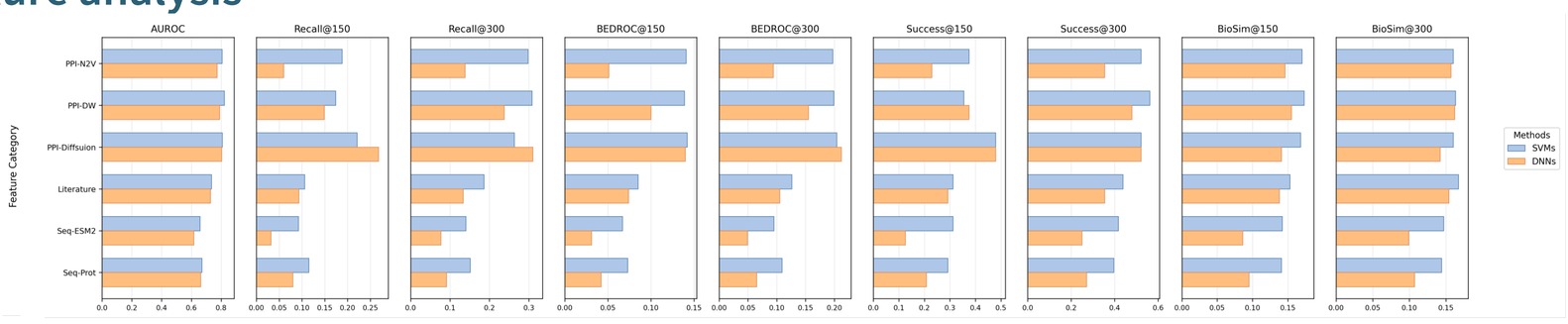


Fig 2. Feature performance across kernel and DNN models. AUROC measures overall ranking performance by estimating how often a disease-associated gene is ranked above a non-associated gene. Recall@k reports the proportion of test disease genes recovered within the top k candidates, while Success@k indicates whether at least one test gene is retrieved. BEDROC emphasizes early recognition by assigning greater weight to positives near the top of the ranked list. BioSim@k complements these label-based metrics by measuring the functional similarity between top-ranked unlabeled genes and known disease genes using the Jaccard overlap of enriched GO and KEGG terms. @150 ≈1%, @300 ≈ 2%

Integration strategy evaluation

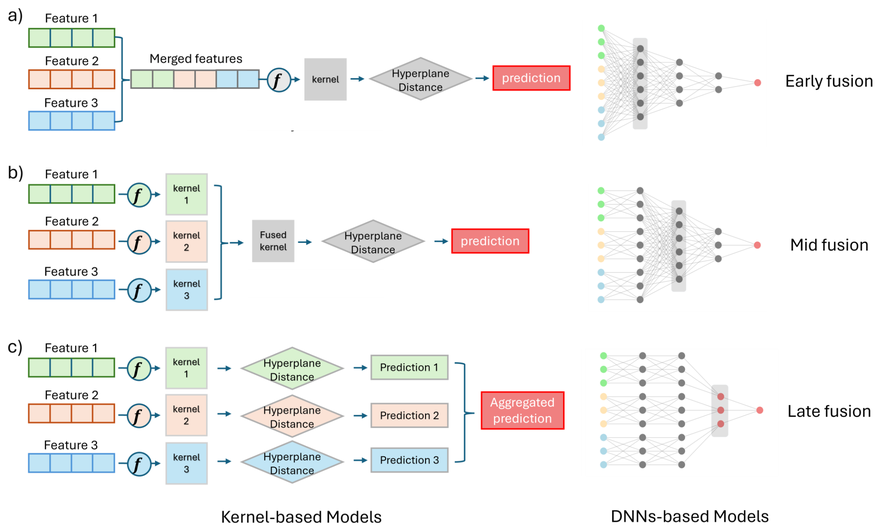


Fig 3. Kernel- and DNN-based feature fusion strategies. Inputs in different colors represent distinct feature groups. The left panels illustrate Kernel-based fusion, where square boxes denote computed kernels, diamond-shaped boxes denote distances to the learned hyperplane, and red boxes indicate final prediction scores. The right panels illustrate the corresponding DNN-based architectures, where grey nodes represent neural-network units, red nodes represent prediction outputs, and shaded regions indicate the stage at which fusion occurs. (a) Early fusion: feature groups are concatenated before model training and processed jointly, either by computing a single kernel or through a shared neural network. (b) Intermediate fusion: each feature group is first transformed separately; kernels are combined using a linear or geometric mean in the SVM setting, while learned representations are merged at an intermediate network layer in the DNN setting. (c) Late fusion: separate models generate feature-specific prediction scores, which are subsequently aggregated to produce the final prediction.

Table 1. Performance Comparison of Fusion Strategies for Kernel and DNN-based Models.

Method	AUROC	Recall		BEDROC		Success		BioSim	
		@150	@300	@150	@300	@150	@300	@150	@300
Kernel-Early fused	0.772	0.121	0.207	0.090	0.136	0.312	0.479	0.165	0.171
Kernel-Mid linear fused	0.814	**0.220**	0.286	**0.151**	0.205	**0.438**	0.500	0.179	0.166
Kernel-Mid geo fused	**0.821**	0.190	0.226	0.135	0.183	0.396	0.521	**0.182**	0.171
Kernel-Late fused	0.819	0.177	0.269	0.130	0.185	0.396	0.521	0.181	**0.176**
Kernel-Early PPI fused	0.817	0.214	0.253	0.144	0.198	0.417	0.479	**0.177**	**0.162**
Kernel-Mid PPI linear fused	0.820	0.219	0.299	0.143	0.206	**0.438**	0.521	0.170	0.159
Kernel-Mid PPI geo fused	**0.824**	**0.239**	**0.321**	**0.145**	**0.208**	**0.438**	**0.542**	0.161	0.149
Kernel-Late PPI fused	0.818	0.193	0.306	0.143	0.204	0.375	**0.542**	0.172	0.161
DNN-Early fused	0.706	0.066	0.114	0.055	0.085	0.188	0.375	0.133	0.145
DNN-Mid fused	0.802	**0.231**	**0.300**	**0.157**	**0.217**	**0.500**	**0.562**	0.131	0.126
DNN-Late fused	**0.805**	0.187	0.258	0.114	0.171	0.417	0.479	**0.157**	**0.160**
DNN-Early PPI fused	**0.805**	**0.193**	0.274	0.123	0.181	0.396	0.521	0.151	0.159
DNN-Mid PPI fused	0.803	0.192	**0.283**	**0.150**	**0.209**	**0.417**	**0.562**	0.130	0.128
DNN-Late PPI fused	0.803	0.179	0.269	0.107	0.167	**0.417**	0.521	**0.155**	**0.160**

BEDROC@k: 80% of the score is determined by the top-k ranked genes

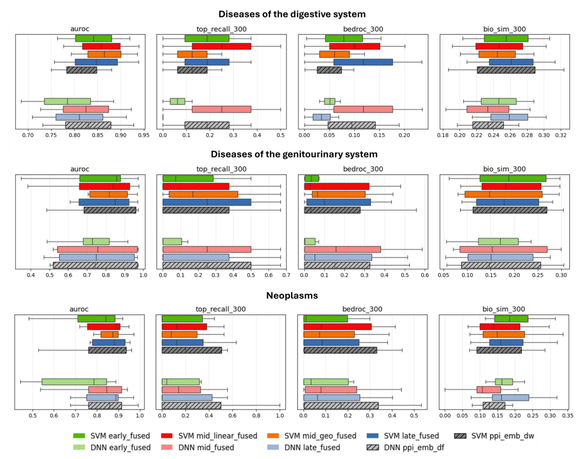


Fig. 4. Fusion strategy performance across three example disease categories using all features. Subplots show individual metrics; dark and light bars represent kernel- and DNN-based fusion, respectively, and grey bars show the best single-feature baseline for each model.

References

[1] Piñero, Janet, et al. "DISGENET: accelerating data-driven discovery in disease genomics and therapeutic development." BioRxiv (2026): 2026-01.
[2] World Health Organization. ICD-10: international statistical classification of diseases and related health problems: tenth revision. World Health Organization, 2004.
[3] Szklarczyk, Damian, et al. "STRING v10: protein–protein interaction networks, integrated over the tree of life." Nucleic acids research 43.D1 (2015): D447-D452.
[4] Grover, Aditya, and Jure Leskovec. "node2vec: Scalable feature learning for networks." Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. 2016.
[5] Perozzi, Bryan, Rami Al-Rfou, and Steven Skiena. "Deepwalk: Online learning of social representations." Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014.
[6] Kondor, Risi, and Jean-Philippe Vert. "Diffusion Kernels." Kernel Methods in Computational Biology, 2004, pp. 171–92.
[7] Park, Junseok, et al. "Concept embedding to measure semantic relatedness for biomedical information ontologies." Journal of biomedical informatics 94 (2019): 103182.
[8] Lin, Zeming, et al. "Evolutionary-scale prediction of atomic-level protein structure with a language model." Science 379.6637 (2023): 1123-1130.
[9] Dallago, Christian, et al. "Learned embeddings from deep learning to visualize and predict protein sets." Current Protocols 1.5 (2021): e113.