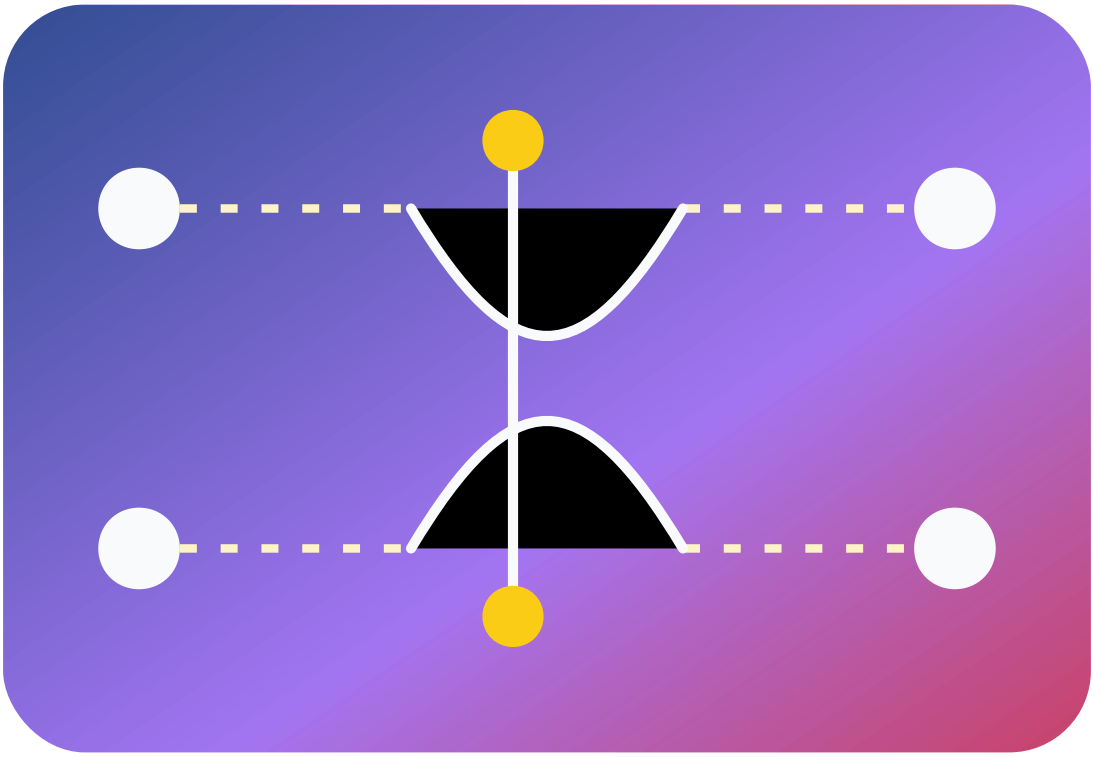
Integrating Multimodal Data Through Interpretable Heterogeneous Ensembles



```
<h3>Integrating Multimodal Data Through Interpretable
Heterogeneous Ensembles · Concept Sketch</h3>
<p>Late-fusion inspired view showing coordinated yet modality-
specific streams.</p>
```

Integrating Multimodal Data Through Interpretable Heterogeneous Ensembles

Authors: Yan Chak Li, Linhua Wang, Jeffrey N. Law, T. M. Murali, Gaurav Pandey
Year: 2022
Venue: Bioinformatics Advances

1. Classification

- **Domain Category:**
  - Multimodal / Integration. The paper focuses on integrating heterogeneous biomedical data types (STRING multi-omic networks and clinical EHR modalities) to improve prediction of protein function and COVID-19 mortality, and on making these integrations interpretable.
- **FM Usage Type:**
  - Multimodal FM or cross-modal integration (conceptual). While the work does not introduce a neural "foundation model" in the modern sense, it develops a general-purpose, reusable late-integration framework (Ensemble Integration, EI) for combining modality-specific models, which fills a similar role for heterogeneous biomedical data.
- **Key Modalities:**
  - Protein function prediction: multi-omic STRING network modalities (protein–protein interaction, curated databases, co-expression, genomic neighborhood, co-occurrence, fusion networks).
  - Clinical prediction: EHR-derived admission features, comorbidities, vital signs, laboratory test measurements for COVID-19 patients.

---

2. Executive Summary

This paper introduces Ensemble Integration (EI), a framework for integrating multimodal biomedical data by first learning separate models for each modality and then combining those models using heterogeneous ensembles. The authors argue that common "early" and "intermediate" integration methods, which merge data into a single representation, often lose modality-specific signals and therefore underperform when modalities have very different structures and semantics. EI instead treats each modality as its own prediction problem, trains strong local classifiers per modality, and then fuses their predictions via ensemble methods such as mean aggregation, ensemble selection, and stacking. The framework is applied to two challenging tasks: predicting protein functions from multimodal STRING network data and predicting COVID-19 mortality from multimodal EHR data. Across both applications, EI consistently outperforms single-modality models and strong early-integration baselines like Mashup, deepNF, and XGBoost. The paper also introduces an interpretation method that ranks features across all modalities by their contribution to the final ensemble, revealing biologically and clinically meaningful predictors. A new grad student should care because EI offers a practical, interpretable baseline strategy for multimodal integration that respects modality-specific structure and can be combined with more modern foundation-model embeddings.

---

3. Problem Setup and Motivation

- **Scientific / practical problem:**
  - How to integrate heterogeneous biomedical data modalities to predict important outcomes such as protein function and patient mortality.

    - For protein function prediction (PFP), the goal is to predict Gene Ontology (GO) annotations for human proteins using multiple STRING-derived networks that capture different biological relationships.
    - For COVID-19 mortality prediction, the goal is to predict whether a hospitalized patient will die from COVID-19 using multimodal EHR data (admission characteristics, comorbidities, vital signs, lab tests).
- **Why existing approaches are insufficient:**
    - **Heterogeneous semantics:** Different modalities have different structures (e.g., dense gene expression matrices vs. graph-structured protein–protein interactions vs. time-series vital signs), making a single joint representation hard to design.
    - **Early / intermediate integration limitations:** Methods that first create a single integrated network or embedding tend to emphasize agreement between modalities but can suppress modality-specific signals that are important for prediction.
    - **Lack of systematic late integration:** While late integration (combining predictions from modality-specific models) has been discussed, it has not been systematically developed or evaluated for multimodal biomedical prediction tasks.
    - **Interpretability challenges:** Complex integrated models are often "black boxes," making it difficult to understand which modalities and features drive predictions, which is especially problematic in biomedical and clinical settings.
- **Motivating idea:**
    - Treat each modality as a first-class citizen by training specialized models that best match its structure, and then combine these local models with flexible ensemble methods.
    - Provide a principled way to interpret the resulting ensembles by quantifying how much each feature and each local model contributes to the final predictions.

---

4. Data and Modalities

- **Datasets used:**
    - **STRING protein function prediction:**
        - Version 11.5 of STRING, focusing on human proteins and their pairwise functional associations.
        - Predicts GO Molecular Function and Biological Process annotations for 18,866 human proteins and 2,139 GO terms (each with at least 50 annotated proteins).
    - **COVID-19 mortality prediction (EHR):**
        - EHR data from 4,783 COVID-19 patients treated at Mount Sinai (March 15–October 1, 2020).
        - Outcome: in-hospital mortality (27.7% positive cases).
- **Modalities:**
    - **STRING multimodal networks (for PFP):**
        - Protein–protein interactions (PPI).
        - Curated database interactions.
        - Co-expression networks.
        - Genomic neighborhood.
        - Co-occurrence of orthologs across genomes.
        - Fusion events of orthologs.

    - **EHR modalities (for COVID-19):**
        - Admission: demographic and baseline clinical variables (e.g., age, sex, race/ethnicity, some vital signs at admission).
        - Comorbidities: presence of other conditions (e.g., asthma, obesity).
        - Vital signs: max/min heart rate, body temperature, respiratory rate, blood pressure, oxygen saturation over the first 36 hours.
        - Laboratory tests: various lab measurements (e.g., BUN, calcium, sodium, white blood cell count), with 44 features retained.
- **Preprocessing / representation:**
    - **STRING:** Each protein is represented by its adjacency vector in each network modality; missing proteins receive all-zero vectors.
    - **PFP labels:** Positive examples are proteins manually annotated (non-IEA evidence) to a GO term; negatives are proteins with other annotations in the same ontology but not to that term or its relatives.
    - **EHR:**
        - Only features with <30% missingness are retained.
        - Remaining missing values within each modality are imputed using KNNImpute with (K=5).
        - Categorical variables are one-hot encoded; continuous variables are standardized to z-scores.
        - For repeated vital signs and labs, the first values within 36 hours of hospitalization are used to enable early prediction.

---

5. Model / Foundation Model

- **Model Type:**
    - EI is a heterogeneous ensemble framework that combines multiple base classifiers trained on separate modalities.
    - Base models are standard supervised learning algorithms (e.g., decision trees, random forests, SVMs, logistic regression, k-NN, Naive Bayes, boosting methods).
    - Ensemble methods used for integration include simple averaging, ensemble selection, and stacking with various meta-learners.
- **Is it a new FM or an existing one?**
    - The paper does **not** introduce a large-scale neural foundation model.
    - Instead, it proposes a **general-purpose late integration framework** (EI) that can sit "on top of" any modality-specific models, including future foundation-model encoders for genomics or clinical data.
- **Key components and innovations:**

Component	Role in EI
Local models per modality	Capture modality-specific structure and signals
Heterogeneous ensembles	Combine diverse local models into a global predictor
Mean aggregation	Simple averaging of local predictions
Caruana's ensemble selection	Greedy selection of local models to maximize performance
Stacking	Meta-model trained on local-model predictions

Component	Role in EI
Feature-importance interpretation	Ranks features via combined local feature and model importance

- **Training setup (as described):**
  - **Local models (all modalities):**
    - Trained with 10 standard binary classifiers in Weka: AdaBoost, decision tree (J48), gradient boosting, k-NN, SVM, random forest (RF), logistic regression (LR), PART (rule-based), Naive Bayes, Voted Perceptron.
    - Class imbalance handled by random undersampling of the majority (negative) class in training; test sets preserve original class ratios.
  - **Ensemble methods:**
    - Mean aggregation.
    - Caruana-style ensemble selection (CES) that iteratively adds models that most improve validation performance.
    - Stacking with meta-predictors from scikit-learn: AdaBoost, decision tree, gradient boosting, k-NN, SVM (linear kernel), RF, LR, Naive Bayes, plus XGBoost as an additional meta-classifier.
  - **Evaluation protocol:**
    - EI and heterogeneous-ensemble baselines are trained and evaluated using 5-fold nested cross-validation to separate local-model and ensemble training.
    - Mashup, deepNF, and XGBoost baselines use standard 5-fold cross-validation.

---

6. Multimodal / Integration Aspects (If Applicable)

This paper is fundamentally about **multimodal data integration**, and EI is a concrete implementation of a **late integration** strategy.

- **Which modalities are integrated?**
  - For PFP, EI integrates multiple STRING network modalities capturing complementary biological evidence (PPIs, curated databases, co-expression, genomic neighborhood, co-occurrence, fusions).
  - For COVID-19 mortality, EI integrates clinical admission data, comorbidities, vital signs, and lab tests.
- **How are they integrated?**
  - EI follows a strict **late integration** pipeline:
    1. Train multiple local models per modality using algorithms appropriate for that modality.
    2. Obtain prediction scores from each local model.
    3. Combine these scores across modalities using heterogeneous ensemble methods (mean, ensemble selection, stacking).
  - No single “joint embedding” or early-fusion feature vector is constructed; instead, integration happens purely at the level of model outputs.
- **Why this integration is useful / what new capabilities it gives:**
  - Preserves **modality-specific signals** by allowing each modality to use its most suitable modeling approach, rather than forcing all data into a common structure.
  - Flexible enough to incorporate arbitrary new modalities (e.g., new omics assays, imaging-derived features, or embeddings from foundation models)

by simply adding new local models.
    - Provides an **interpretation mechanism** that attributes importance to features across modalities via a unified ranking, which is crucial for biomedical trust and discovery.
- **Relation to the Integration Baseline Plan:**
    - The paper operationalizes the principle: **"Prefer late integration first under heterogeneous semantics."** EI explicitly avoids premature joint spaces, instead building strong per-modality models and then combining them, which directly aligns with the plan's recommendation to preserve modality-specific signals and avoid over-aggressive early fusion.
    - EI's approach resembles the plan's suggestion of **concatenating compact per-modality features and training robust tabular models**—but at a higher level: EI concatenates prediction scores (rather than raw features) and uses ensembles like LR, RF, and boosting as meta-models.
    - Regarding **robustness and evaluation discipline**, EI uses nested cross-validation, imbalance-aware metrics (AUPRC, Fmax), and formal statistical tests (Friedman–Nemenyi, Wilcoxon tests), which is very much in the spirit of the plan's emphasis on disciplined evaluation and uncertainty quantification.
    - The plan's emphasis on **CCA and permutation tests** is conceptually related to EI's interpretation method: both aim to understand cross-modal relationships and contributions. EI's permutation-based local model ranks (LMRs) and feature ranks (LFRs) serve a similar diagnostic role for model-level integration.
    - Overall, EI provides a concrete, well-validated example of **late fusion via heterogeneous ensembles**, which can be adopted as the default multimodal baseline before moving to more complex contrastive or joint-embedding approaches described in the integration baseline.

---

7. Experiments and Results

- **Tasks / benchmarks:**
    - **Protein function prediction (PFP):** Predict GO Molecular Function and Biological Process annotations for 2,139 GO terms using multimodal STRING networks.
    - **COVID-19 mortality prediction:** Predict in-hospital death for COVID-19 patients using multimodal EHR features.
- **Baselines:**
    - **For PFP:**
        - Early integration methods: Mashup and deepNF, which fuse multiple networks into a single integrated network before classification.
        - Single-modality baselines: heterogeneous ensembles trained separately on each STRING modality.
    - **For COVID-19 mortality:**
        - Early integration baseline: XGBoost trained on the concatenated feature vector across all EHR modalities (a strong tabular-data baseline).
        - Single-modality baselines: heterogeneous ensembles trained on each EHR modality individually.
- **Key findings (PFP):**
    - EI achieves significantly higher Fmax scores than Mashup, deepNF, and any single STRING modality across 2,139 GO terms.

    - This performance advantage persists across GO terms with varying depth, information content, and number of annotated proteins, although performance understandably decreases for terms with very few annotations.
    - Stacking with RF and LR as meta-learners tends to perform best among EI variants, consistent with previous work on heterogeneous ensembles.
    - The only setting where EI underperforms is for GO terms with very few annotations (50–100), where Mashup slightly outperforms EI.
- **Key findings (COVID-19 mortality):**
    - EI outperforms ensembles trained on individual EHR modalities and slightly surpasses the early-integration XGBoost baseline in Fmax and AUROC.
    - The best EI variant (stacking with LR) achieves a modest but meaningful improvement in Fmax over XGBoost, with a slightly better balance of precision and recall.
    - EI-based predictions confirm that laboratory test features are particularly informative, but admission and comorbidity features also contribute.
- **Interpretation results:**
    - The proposed interpretation method identifies the top-contributing features for the best EI model in the COVID-19 task.
    - Top features include age at admission, minimum oxygen saturation, blood urea nitrogen (BUN), calcium, chloride, sodium, venous (_2), respiratory rate, and atrial fibrillation—variables known to be clinically relevant to COVID-19 severity and mortality.
    - There is statistically significant overlap between EI's top features and those highlighted by XGBoost's SHAP-based importance, supporting the validity of EI's explanations.

---

8. Strengths, Limitations, and Open Questions

- **Strengths:**
    - Provides a **clear, general framework** for late integration that can be applied to many multimodal biomedical problems.
    - Preserves **modality-specific structure and signals** by training specialized local models instead of forcing all data into a common representation.
    - Demonstrates strong **empirical performance** on two very different tasks (PFP and clinical mortality prediction) relative to established early-integration baselines.
    - Introduces a **model-agnostic interpretation method** that yields clinically and biologically meaningful feature rankings, increasing trust and enabling scientific insight.
    - Uses **rigorous evaluation practices** (nested CV, appropriate metrics, statistical tests), aligning well with best-practice recommendations like DOME.
- **Limitations:**
    - The framework is evaluated mainly on **structured data** and classical machine learning models; it does not yet incorporate modern deep or foundation-model encoders for unstructured inputs (e.g., sequences, images, free text).
    - Comparisons focus on early integration (Mashup, deepNF, XGBoost); **intermediate integration** methods are not systematically evaluated.
    - Interpretation is only fully explored for the COVID-19 mortality task; PFP interpretations for thousands of GO terms are not examined in depth.

    - The interpretation method uses AUPRC-based ranks and percentile normalization, which may slightly bias importance toward modalities with more features.
    - Computational cost may increase with many modalities and large libraries of local models, since both ensemble selection and permutation-based importance are relatively heavy.
- **Open Questions and Future Directions:**
    - How does EI perform when **local models are foundation-model encoders** (e.g., protein language models, EHR transformers, imaging FMs) instead of classical classifiers?
    - Can we design **hybrid integration schemes** that combine EI-style late fusion with intermediate or contrastive objectives (e.g., CCA, cross-modal contrastive learning) to better exploit cross-modal structure?
    - How can the **interpretation framework** be extended to handle thousands of labels (as in PFP) in a scalable way, perhaps by grouping labels or focusing on biologically meaningful subsets?
    - Can we mitigate the bias of the current ranking scheme toward feature-rich modalities, for example by normalizing importance across modalities or using multi-task feature selection?
    - What are the best practices for choosing and tuning the **library of local models** and meta-learners in EI when applying it to new domains or datasets?

---

9. Context and Broader Impact

- **Position in the landscape:**
    - Within the broader space of foundation models and multimodal integration, EI provides a **strong, interpretable late-integration baseline** for combining outputs from many models or modalities.
    - For genomics and proteomics, it complements work on network integration (Mashup, deepNF) by showing that keeping network modalities separate and integrating predictions can outperform aggressive early fusion.
    - For clinical applications, EI offers an alternative to monolithic models like XGBoost or single neural networks, emphasizing modularity and interpretability.
- **Relation to well-known ideas:**
    - Conceptually, EI is like building a **committee of experts**, where each expert specializes in one modality, and a supervisor aggregates their opinions via ensemble methods.
    - It aligns with ensemble learning ideas such as stacking and ensemble selection, and can be viewed as a **late-fusion wrapper** around any set of modality-specific models, including deep FMs.
    - From the perspective of modern FMs, you can imagine replacing local Weka classifiers with **frozen or fine-tuned foundation models** that output per-modality predictions or embeddings, which EI then integrates.
- **Relevance to the integration baseline plan:**
    - EI is a concrete instantiation of the plan's recommendation to **start with late fusion under heterogeneous semantics**, using robust tabular models and careful evaluation before moving to more complex multimodal architectures.
    - The paper's nested CV, imbalance-aware metrics, and statistical comparisons mirror the plan's emphasis on **robustness and evaluation**

discipline, making it a strong methodological template.
    - For future multimodal FM work (e.g., combining brain FMs, genomic FMs, and clinical data), EI can serve both as a **baseline pipeline** and as a reusable integration layer atop pre-trained encoders.

---

10. Key Takeaways (Bullet Summary)

- **Problem:**
    - Integrating heterogeneous multimodal biomedical data (e.g., STRING networks, EHR features) is crucial for accurate prediction of protein function and clinical outcomes but is challenging due to differing data structures and semantics.
    - Early and intermediate integration approaches that force data into a single representation can lose modality-specific information and underperform.
- **Method / model:**
    - Ensemble Integration (EI) is a **late integration framework** that first trains multiple local models per modality and then combines their prediction scores using heterogeneous ensemble methods (mean aggregation, ensemble selection, stacking).
    - EI is model-agnostic and can work with any base classifier, enabling flexible integration of diverse modalities and algorithms.
    - A novel interpretation method combines **local feature ranks** and **local model ranks** to produce a global ranking of features across all modalities.
- **Results:**
    - On protein function prediction with multimodal STRING data, EI significantly outperforms early integration methods (Mashup, deepNF) and single-modality baselines across thousands of GO terms.
    - On COVID-19 mortality prediction from EHR data, EI slightly but consistently outperforms a strong early-integration XGBoost baseline and individual-modality ensembles.
    - The interpretation framework highlights clinically and biologically meaningful features (e.g., age, oxygen saturation, BUN, calcium), with significant overlap with SHAP-based importances from XGBoost.
- **Why it matters:**
    - EI offers a **practical, interpretable, and extensible baseline** for multimodal integration that respects modality-specific signals and uses rigorous evaluation.
    - It provides a natural way to integrate outputs from future **foundation models for genomics, proteomics, and clinical data**, making it highly relevant for students and researchers planning multimodal FM systems.

Generated via custom pipeline · 2025-11-18