Energy-Based Methods for Interpretable Multimodal Biomedical Decision Making

Context and Motivation. Massive biomedical data lie at the forefront of personalized medicine, but effectively harnessing their complexity remains a major challenge. Despite the rapid growth of multimodal data (genomics, EHR, sensor, biomarker), existing methods fail to capture their intricate interplay. Techniques like t-SNE and UMAP handle high-dimensional data but lack a statistical framework for multimodal fusion and uncertainty quantification.

We propose leveraging energy-based models (EBMs) inspired by statistical physics and Markov random fields, where local pairwise similarities between modalities (e.g., genomics, imaging, clinical data) define a global energy function. Unlike traditional probabilistic models that explicitly define likelihoods, EBMs implicitly model dependencies through an energy function, offering greater flexibility. Compared to deep learning, EBMs naturally encode domain knowledge via pairwise constraints, enhancing interpretability and robustness to missing data while facilitating structured optimization and uncertainty-aware decision-making.

Scientific Objective. Our goal is to develop a scalable, interpretable Energy-Based Model (EBM) framework for biomedical data integration, validated against real-world datasets. This requires addressing three core challenges:

1. Multimodal Data Fusion: Develop an EBM methodology to model interactions across biomedical modalities while preserving dependencies. This enables structured fusion, probabilistic inference, and robust missing data imputation.

2. Scalability and Theoretical Foundations: Ensure that EBMs remain computationally feasible for large-scale biomedical applications by developing efficient inference algorithms. We will analyze the probabilistic structure of energy landscapes, the distribution of critical points, and phase transitions to uncover meaningful multimodal manifolds.

3. Efficiency and Interpretability: Optimize inference techniques for high-dimensional EBMs to ensure that clinical predictions remain computationally scalable, trustworthy, and interpretable. By improving uncertainty quantification and model calibration, we will enhance real-world applicability in disease subtyping, biomarker discovery, and clinical decision support.

Justification of the Approach. Biomedical datasets are large, multimodal, and often incomplete, encompassing high-dimensional genetic data, imaging, and clinical metadata. Existing methods either process modalities separately, losing cross-modality insights, or concatenate them, leading to over-fitting and reduced interpretability.

EBMs enable structured multimodal fusion by modeling patients as nodes in a similarity graph, where pairwise constraints define an energy function. This allows for: (1) Cross-modality integration without naive aggregation, (2) Uncertainty-aware probabilistic inference for clinical applications, and (3) Robust missing data handling through dependency-driven inference. Despite these advantages, several critical barriers limit EBM adoption in biomedical settings, particularly in scalability, reliability, and interpretability:

- **Intractability of the partition function** limits training, inference, and uncertainty estimation, posing challenges for clinical decision-making where reliable probability estimates are crucial. Approximate methods (e.g., CD, Langevin MCMC) lack theoretical guarantees but remain feasible when combined with structured approximations. We will integrate gradient-based MCMC with low-rank tensor approximations and variational inference to improve scalability and theoretical robustness, ensuring accurate probabilistic assessments in applications such as disease risk stratification and treatment planning.

- **Lack of multimodal generalization bounds for EBMs** hinders model reliability across diverse populations. A model trained on UK Biobank may fail in hospitals with different demographics, leading to unreliable predictions. To address this, we will develop approximate generalization bounds for EBMs in multimodal graphs using information-theoretic and variational methods, ensuring stability across datasets.

- **Lack of tools for detecting structural changes** in multimodal data obscures hidden disease subtypes. Cancer subtypes, for example, emerge from complex gene-biomarker-imaging interactions, not single mutations. We hypothesize that phase transitions in the EBM energy landscape correspond to structural changes in the underlying data manifold, helping identify biologically significant patient subgroups. We will thus analyze Hessian eigenspectra and local curvature in multimodal manifolds to characterize these transitions and improve disease subtyping and biomarker discovery.

- **Limited understanding of the geometry of random energy landscapes** weakens EBM robustness, as too many local minima can cause overfitting, leading to spurious associations rather than true disease mechanisms. We will analyze the distribution of critical points and saddle points in high-dimensional loss surfaces to design stable energy functions, ensuring EBMs remain expressive yet robust. This will improve model calibration, uncertainty quantification, and rare disease detection, making EBMs both mathematically sound and clinically actionable.

Our work on these computational and mathematical challenges will enable EBMs to scale, generalize, and provide interpretable biomedical insights, ensuring they are not merely theoretical but practically viable for clinical use.

Three Year Timeline.

Year 1. Development of a baseline EBM framework integrating two biomedical modalities (e.g., SNP and imaging). We will validate small-scale inference (MCMC, mean-field) on synthetic and partially labeled datasets. Initial scalability studies will identify bottlenecks. Start work on generalization bounds for EBMs in multimodal graphs and analysis of random energy landscapes, focusing either on loss surface geometry or phase transition behaviors.

Year 2. Extension to real-world biomedical datasets, improving scalability, robustness, and interpretability. We will refine approximate partition function estimation and benchmark EBM-based fusion against deep learning and probabilistic models. Iterative validation cycles will refine assumptions and ensure real-world applicability. Further mathematical analysis will improve soundness of the developing methods.

Year 3. Optimize HPC scalability for real-world deployment. Establish theoretical guarantees linking empirical results to statistical models and/or learning theory. Assess model interpretability using clinical metrics. Finalize publications, open-source tools, and explore industry collaborations.

Candidate Profile. Candidates should have: (1) Strong math/stats/ML background, (2) programming skills for implementing algorithms, (3) keen interest in energy-based AI models, and (4) willingness to engage across disciplines. The language of study is English.

Team and Location. The candidate must physically study at Sorbonne University Abu Dhabi (UAE) for the duration of their PhD, although their enrollment and degree will be from Sorbonne University (France) in Mathematics[1]. The primary adviser will be Dr. Samuel F. Feng, co-advised by Dr. Omar El-Dakkak. Dr. Gérard Biau will sign as HDR holder.

Selected References

Murphy, K. P. (2023). Ch. 24: Energy-based methods, and Ch 34: Decision making under uncertainty. In ***Probabilistic machine learning: Advanced topics***. MIT Press.

Schröder et al (2023). Energy discrepancies: A score-independent loss for energy-based models. ***NeurIPS 2023***.

Muneeb et al (2022). Transfer learning for genotype–phenotype prediction using deep learning models. In ***BMC Bioinformatics***. doi:10.1186/s12859-022-05036-8

[1]L'école doctorale Sciences Mathématiques de Paris-Centre (ED 386)