

*Digital Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Science and Technology 2665*

Robust Learning from Distributed and Heterogeneous Data

LI JU



ACTA UNIVERSITATIS
UPSALIENSIS
2026



UPPSALA
UNIVERSITET

Dissertation presented at Uppsala University to be publicly examined in 101195, Heinz-Otto Kreiss, Ångströmlaboratoriet, Uppsala, Thursday, 4 June 2026 at 09:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Professor of Numerical Analysis Desmond Higham.

Abstract

Ju, L. 2026. Robust Learning from Distributed and Heterogeneous Data. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 2665. 62 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-2816-4.

Modern machine learning is increasingly expanding beyond centralized, mono-modal training toward systems that must also learn from data across distributed edge devices and heterogeneous data modalities. This transition breaks the foundational identical and independent distribution (i.i.d.) assumptions of traditional models, making robustness a first-class requirement for real-world applications. This thesis studies the mechanisms and methodologies necessary to achieve algorithmic robustness across three intersecting dimensions: distributed optimization, geometry-aware uncertainty quantification, and simulation-based inference.

The first dimension addresses statistical heterogeneity in Federated Learning (FL), a distributed training framework in which multiple participants collaboratively train a shared model without exchanging their local data. In FL, the non-i.i.d. nature of distributed data often induces performance degradation, convergence issues and fairness problems. Through an empirical study on drug discovery and the development of new algorithms, this work demonstrates that adaptive optimization and dynamic hyperparameter adjustment can mitigate training instabilities. These methods ensure equitable performance across diverse data silos, preventing the global model from favoring specific participants.

The second dimension explores the structural challenges of multi-modal language models, which map data of heterogeneous modalities onto complex, non-Euclidean manifolds. This research models aleatoric and epistemic uncertainty with directional distributions via parametric models and Riemannian Flow Matching. This geometry-aware approach allows models to respect the intrinsic geometric structure of the embedding space, providing a mathematically grounded framework for models to quantify their ignorance when confronted with ambiguous or out-of-distribution inputs.

The final dimension addresses the robustness of a unified framework which supports both forward and inverse processes for Bayesian inference. The proposed framework utilizes a unified Flow Matching model to learn the joint distribution of parameters and observations. By employing randomized masking, this architecture robustly handles partially observed or noisy data, integrating forward and inverse processes into a single cohesive neural network without the need for specialized retraining. Collectively, this thesis contributes theoretical analyses, novel algorithms, and empirical validations that advance the robustness of machine learning across federated optimization, multi-modal uncertainty quantification, and simulation-based inference, bridging the gap between idealized training assumptions and the demands of real-world applications.

Keywords: Machine Learning, Distributed Optimization, Federated Learning, Probabilistic Modeling, Multi-modal Learning

Li Ju, Department of Information Technology, Division of Scientific Computing, Box 337, Uppsala University, SE-751 05 Uppsala, Sweden. Department of Information Technology, Computational Science, Box 337, Uppsala University, SE-75105 Uppsala, Sweden. Science for Life Laboratory, SciLifeLab, Box 256, Uppsala University, SE-75105 Uppsala, Sweden.

© Li Ju 2026

ISSN 1651-6214

ISBN 978-91-513-2816-4

URN urn:nbn:se:uu:diva-583490 (<http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-583490>)

Dedicated to those who refuse to settle for local minima

List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

- I **Li Ju**, Andreas Hellander, and Ola Spjuth. "Federated learning for predicting compound mechanism of action based on image-data from cell painting." *Artificial Intelligence in the Life Sciences* 5 (2024): 100098.
- II **Li Ju**, Tianru Zhang, Salman Toor, and Andreas Hellander. "Accelerating fair federated learning: Adaptive federated adam." *IEEE Transactions on Machine Learning in Communications and Networking* 2 (2024): 1017-1032.
- III **Li Ju**, Max Andersson, Stina Fredriksson, Edward Glöckner, Andreas Hellander, Ekta Vats, and Prashant Singh. "Exploiting the Asymmetric Uncertainty Structure of Pre-trained VLMs on the Unit Hypersphere." *Advances in Neural Information Processing Systems*, vol. 38, 2025.
- IV **Li Ju**, Mayank Nautiyal, Andreas Hellander, Ekta Vats, and Prashant Singh. "Epistemic Uncertainty Quantification for Pre-trained VLMs via Riemannian Flow Matching." *arXiv preprint arXiv:2601.21662* (2026).
- V Mayank Nautiyal, **Li Ju**, Melker Ernfors, Klara Hagland, Ville Holma, Maximilian Werkö Söderholm, Andreas Hellander, and Prashant Singh. "OneFlowSBI: One Model, Many Queries for Simulation-Based Inference." *arXiv preprint arXiv:2601.22951* (2026).

Reprints were made with permission from the publishers.

List of additional papers

In addition to the papers listed before, the author also authored & co-authored the following papers during their PhD studies.

- I **Li Ju**, Prashant Singh, and Salman Toor. "Proactive autoscaling for edge computing systems with kubernetes." In Proceedings of the 14th IEEE/ACM International Conference on Utility and Cloud Computing Companion, pp. 1-8. 2021.
- II Shenghui Li, Edith C-H. Ngai, Fanghua Ye, **Li Ju**, Tianru Zhang, and Thiemo Voigt. "Blades: A Unified Benchmark Suite for Byzantine Attacks and Defenses in Federated Learning." In 2024 IEEE/ACM Ninth International Conference on Internet-of-Things Design and Implementation (IoTDI), pp. 158-169. IEEE, 2024.
- III Mayank Nautiyal, Stela Arranz Gheorghe, Kristiana Stefa, **Li Ju**, Ida-Maria Sintorn, and Prashant Singh. "PARIC: Probabilistic Attention Regularization for Language Guided Image Classification from Pre-trained Vision Language Models." arXiv preprint arXiv:2503.11360 (2025).
- IV Tianru Zhang, **Li Ju**, Prashant Singh, and Salman Toor. "Infohier: Hierarchical information extraction via encoding and embedding." arXiv preprint arXiv:2501.08717 (2025).

Reprints were made with permission from the publishers.

Contents

1	Introduction	13
1.1	The Evolution of Machine Learning	13
1.2	Distributed and Heterogeneous Data	14
1.2.1	Distributed Data	14
1.2.2	Heterogeneous Modalities	15
1.3	The Necessity of Robust Learning	15
1.4	Outline	16
2	Foundations	17
2.1	Learning from Distributed Data	17
2.1.1	Stochastic Optimization	17
2.1.2	Privacy-Preserving Distributed Learning	19
2.1.3	Federated Learning Formulation	19
2.1.4	Federated Optimization	20
2.2	Learning from Heterogeneous Modalities	21
2.2.1	Shared Representation	21
2.2.2	Contrastive Learning	23
2.2.3	Geometric Constraint	24
2.3	Probabilistic Frameworks	25
2.3.1	Uncertainty Quantification	25
2.3.2	Modelling Probability Distributions	25
2.3.3	Simulation-Based Inference	28
3	Robustness in Federated Learning	30
3.1	Statistical Heterogeneity	30
3.1.1	Optimization Instability	31
3.1.2	Fairness Problem	31
3.2	Empirical Study: Federated Learning for Drug Discovery	32
3.3	Accelerating Fair Federated Learning	33
4	Robustness in Vision-Language Models	35
4.1	Aleatoric Uncertainty and Modality Asymmetry	35
4.2	Epistemic Uncertainty via Density Estimation	37
5	Robust Simulation-Based Inference	39
5.1	Challenges for Robust Inference	39
5.2	Unified Joint Inference with OneFlowSBI	40
6	Summary of Papers	42

7	Conclusion and Outlook	45
7.1	Conclusion	45
7.2	Outlook	45
	Author's Contributions	47
	Sammanfattning på svenska	49
	Statement on the Use of Generative AI	51
	Acknowledgement	52
	References	54

List of Abbreviations

ABC	approximate Bayesian computation.
AdaFedAdam	adaptive federated adam.
Adam	adaptive moment estimation.
AsymVLM	asymmetric probabilistic adaptation for vision-language models.
CCPA	California Consumer Privacy Act.
CLIP	contrastive language-image pre-training.
DMOO	dynamic multi-objective optimization.
ELBO	evidence lower bound.
FedAvg	federated averaging.
FedOpt	federated optimization.
FL	federated learning.
FM	flow matching.
GAN	generative adversarial network.
GD	gradient descent.
GDPR	General Data Protection Regulation.
GPU	graphics processing unit.
HPC	high-performance computing.
i.i.d.	independent and identically distributed.
ML	machine learning.
MoA	mechanism of action.
NPE	neural posterior estimation.
ODE	ordinary differential equation.
OOD	out-of-distribution.
RepVLM	epistemic uncertainty quantification for pre-trained VLMs via Riemannian flow matching.
SBI	simulation-based inference.
SDE	stochastic differential equation.
SGD	stochastic gradient descent.
SigLIP	sigmoid loss for image pre-training.
SMC	sequential Monte Carlo.
SNPE	sequential neural posterior estimation.
TPU	tensor processing unit.
UQ	uncertainty quantification.
VAE	variational autoencoder.
VLM	vision-language model.
vMF	von Mises-Fisher.

1. Introduction

This introduction positions the thesis at the intersection of two defining realities of modern machine learning: data is increasingly distributed across edge devices and silos, and it is intrinsically multi-modal and heterogeneous. These shifts break the assumptions behind centralized, independent and identically distributed (i.i.d.) training and make robustness a first-class requirement. The chapter frames robustness along statistical heterogeneity in federated learning (FL), geometry-aware uncertainty for vision-language models (VLMs), and principled Bayesian inference for simulation-based inference (SBI), before closing with a roadmap of the enclosed contributions.

1.1 The Evolution of Machine Learning

Machine learning (ML) has been witnessing a paradigm shift over the past decade, evolving from small-scale statistical models to large-scale contemporary deep learning systems [1, 2].

The deep learning revolution, which started in the early 2010s, was propelled by three key factors:

- **Computational Power:** The growth in computational power afforded by specialized hardware such as graphics processing units (GPUs) and tensor processing units (TPUs).
- **Improved Algorithms:** The development of highly expressive and scalable algorithmic architectures in deep neural networks, such as residual connections [3, 4] and attention mechanisms [5].
- **Data Accessibility:** The widespread availability of massive curated datasets like ImageNet [6].

The underlying assumption of traditional machine learning is that data can be collected and aggregated in a single massive data center, and shuffled uniformly. Furthermore, early successes predominantly relied on mono-modal data, focusing exclusively on standardized image benchmarks or standalone text corpora. This centralized setup strictly adheres to the foundational statistical i.i.d. assumption. Under these idealized conditions of centralized, single-modality data, traditional ML methods are highly effective and scalable [7].

However, as ML systems permeate real-world applications, from healthcare diagnostics [8] to autonomous driving [9], the fundamental limitations of centralized, mono-modal data have become increasingly apparent.

First, centralized data collection inherently faces logistical and legal barriers. Data generation has largely shifted to the edge of the network, with millions of devices, from smartphones and wearable health monitors to autonomous vehicles and industrial IoT sensors, generating petabytes of distributed data daily [10]. Transmitting this vast amount of data to a central server is infeasible due to constraints of network bandwidth, latency boundaries, and privacy regulations, such as the General Data Protection Regulation (GDPR) [11] or California Consumer Privacy Act (CCPA) [12].

Second, mono-modal data presents inherent limitations when attempting to parse the complex reality of real-world environments. By restricting observations exclusively to a single channel, such as only images or solely texts, machine learning models become structurally blind to contextual information from other modalities [13, 14]. This artificial restriction results in models that frequently fail when confronted with ambiguity or noise that could easily be resolved through data of other modalities.

1.2 Distributed and Heterogeneous Data

To overcome these bottlenecks, it is necessary to design and build ML systems which natively rely on *distributed data* residing at the edge, and integrate *heterogeneous modalities* combining visual, textual, and sensor-based signals. Accordingly, the theoretical and practical challenges addressed in this thesis are anchored in these two aspects of modern real-world data environments: learning from distributed data and learning from heterogeneous data.

1.2.1 Distributed Data

Learning from distributed data, most notably through the framework of FL, represents a shift from "data-to-model" toward "model-to-data". In FL, rather than collecting raw data into a central repository, the learning process is decentralized: raw data remains locally on its originating device or silo, and only model updates are iteratively collected and aggregated by a central coordinator [15]. This architecture natively addresses the privacy and logistical concerns of centralized ML by enabling collaborative training while sensitive information stays localized and private.

While shifting computation directly to edge devices or isolated data silos bypasses the logistical and privacy bottlenecks of centralized learning, it introduces a fundamental statistical challenge [16, 17].

Unlike centralized data, data stored across these separate physical locations is naturally imbalanced and lacks the i.i.d. guarantee. This statistical heterogeneity, driven by differing user habits, demographics, or geographic locations, can destabilize training and create performance disparities across the network. Consequently, the resulting global model often favors a subset of participants,

performing highly accurately for them, while yielding degraded, unreliable results for others [18, 19]. For example, a medical diagnostic model trained collaboratively across a network of hospitals might achieve exceptional accuracy for urban facilities with abundant data, yet systematically fail on a rural clinic that contributes fewer samples or serves a different demographic.

1.2.2 Heterogeneous Modalities

The second focused area of this thesis addresses the shift from mono-modal systems to multimodal learning. In this paradigm, models are designed to align and reason across diverse data types, ranging from visual and textual inputs to temporal sensor logs and audio signals. By leveraging the complementary information present in different modalities, these models can achieve a more holistic understanding of a given context than any single data source could provide [13]. This move toward heterogeneity is driven by the realization that real-world intelligence emerges from the integration of correlated, yet distinct, types of information [14].

However, integrating diverse signals, such as high-dimensional images, discrete texts, and continuous sensor readings, introduces a profound algorithmic dilemma [13, 14]. Models must constantly balance cross-modal alignment with the preservation of modality-specific differences. Because modalities vary vastly in dimensionality and semantic density, forcing a strict shared representation risks erasing their unique, complementary features [20]. Conversely, failing to adequately align them prevents the model from synthesizing a unified understanding of the environment. This challenge necessitates rethinking shared representation spaces that ensure the reliability and robustness of multimodal models [21].

1.3 The Necessity of Robust Learning

As ML transitions into highly dynamic environments governed by distributed and heterogeneous data, traditional metrics like average-case accuracy become poor proxies for genuine performance [22, 23]. It is more important to ensure that a model can perform when confronting distribution shifts, uncertainty, bias and other forms of adversity. Therefore, *robustness* is a necessary property.

In the context of this thesis, robustness refers to the ability of a learning system to maintain reliable, fair, and well-calibrated performance under conditions that deviate from the idealized assumptions of standard training, whether those deviations arise from heterogeneous data distributions, non-Euclidean representational structures, or incomplete and noisy observations. The core contribution of the thesis is reflected along the three intersecting dimensions:

- **Robustness to Statistical Heterogeneity:** When data is permanently dispersed across isolated sources, it exhibits non-i.i.d. characteristics,

such as batch effects or demographic skews [16, 17]. A robust learning framework must stabilize the optimization process regardless of these divergent local distributions, mitigating performance drift and ensuring that performance remains stable, fair, and equitable across all participating data silos, rather than biasing toward certain groups [24].

- **Robustness via Geometry-Aware Uncertainty:** Integrating complex, heterogeneous data of different modalities requires mapping them into shared representational spaces. However, treating these inherently non-Euclidean topological spaces with standard Euclidean assumptions inevitably yields unreliable predictions [13, 14]. A robust system must mathematically respect the intrinsic geometry of the data manifold to provide highly calibrated uncertainty. This is critical for reliably detecting out-of-distribution (OOD) anomalies or inherently ambiguous inputs [21].
- **Robustness in Bayesian Inference:** In SBI, ML models act as flexible surrogates for Bayesian posterior estimation [25]. However, empirical observations rarely match idealized simulator outputs; they are frequently corrupted, partially observed, or suffer from missing data. A robust inference system must handle these empirical imperfections, producing statistically valid posterior distributions rather than confidently biased parameter estimates [26].

This thesis addresses these challenges by providing empirical studies, theoretical analysis, and algorithmic insights to improve the robustness of machine learning systems in distributed and heterogeneous environments.

1.4 Outline

The remainder of this thesis is organized as follows.

Chapter 2 establishes the theoretical foundations, covering stochastic and federated optimization, contrastive multi-modal representation learning on hyperspherical embeddings, and the probabilistic tools of uncertainty quantification, flow matching, and simulation-based inference.

Chapter 3 addresses robustness of federated learning to statistical heterogeneity, presenting an empirical study in drug discovery (Paper I) and the adaptive federated adam (AdaFedAdam) optimizer for fair federated optimization (Paper II).

Chapter 4 tackles uncertainty quantification for vision-language models on non-Euclidean manifolds, introducing geometry-aware aleatoric modeling via directional distributions (Paper III) and epistemic density estimation via Riemannian flow matching (Paper IV).

Chapter 5 discusses the robustness challenges of simulation-based inference and introduces a unified flow matching framework for robust SBI (Paper V).

Chapter 6 summarizes each of the five papers individually, and Chapter 7 concludes with a discussion of results and future directions.

2. Foundations

This chapter provides the theoretical foundations necessary to understand the subsequent challenges of learning in distributed and heterogeneous environments. We begin with large-scale optimization and extend it to secure, decentralized learning, i.e., federated learning. We then explore the algorithmic shift towards multi-modal learning via contrastive learning, and focus on the geometrical properties of their embedding spaces. Finally, we introduce the core principles of uncertainty quantification (UQ), generative modeling, and Bayesian inference, which are mathematical frameworks essential for determining model reliability and improving robustness of machine learning models.

2.1 Learning from Distributed Data

Machine learning traditionally assumes data are centralized. However, as established in the introduction, physical, latency, and legal constraints often isolate data across geographically distributed silos or edge devices [27]. Developing algorithms that can efficiently and robustly train on distributed data requires a foundational understanding of both general stochastic optimization and the communication-constrained distributed optimization in the context of decentralized learning.

2.1.1 Stochastic Optimization

The engine driving the successful training of modern deep neural networks is stochastic optimization. The problem of training a supervised neural network $f_{\theta}(\mathbf{x})$ is typically formulated as an optimization problem w.r.t parameters $\theta \in \mathbb{R}^d$ that minimizes an empirical risk $R(\theta)$ over a given dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$:

$$\min_{\theta} R(\theta) = \frac{1}{N} \sum_{i=1}^N \ell(f_{\theta}(\mathbf{x}_i), y_i) \quad (2.1)$$

where \mathbf{x}_i is the input data and y_i is the corresponding label, and $\ell(\cdot)$ acts as a differentiable loss function measuring the discrepancy between the model output $f_{\theta}(\mathbf{x}_i)$ and the true target y_i .

While standard gradient descent (GD) computes the exact gradient of the loss function over the entire dataset $\nabla R(\theta)$, the massive scale of modern high-dimensional data makes this computationally prohibitive [28]. The algorithm

of choice therefore has shifted to stochastic gradient descent (SGD), which efficiently estimates the true gradient using small, randomly sampled mini-batches of data $\mathcal{B} \subset \mathcal{D}$:

$$\nabla R_{\mathcal{B}}(\boldsymbol{\theta}) = \frac{1}{|\mathcal{B}|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{B}} \nabla \ell(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) \approx \nabla R(\boldsymbol{\theta}) \quad (2.2)$$

Then the update iteration at step t is given by

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \nabla R_{\mathcal{B}}(\boldsymbol{\theta}_t) \quad (2.3)$$

where η denotes the learning rate (step size).

Although foundational to the field, vanilla SGD faces significant challenges in contemporary deep learning. It is often plagued by the ravine problem, where ill-conditioned curvature causes the optimizer to oscillate inefficiently [29]. Moreover, its "one size fits all" approach, applying a single learning rate to every parameter, fails to account for varying gradient scales. Additionally, the high variance of its stochastic updates leads to noisy trajectories that struggle to converge in large-scale optimization landscapes [30].

To effectively address these optimization bottlenecks, the community has developed adaptive moment estimation methods. These momentum-based algorithms utilize moving averages of gradient estimates to accelerate convergence along relevant dimensions while dampening optimization oscillations [31, 32, 33]. A premier example is the adaptive moment estimation (Adam) optimizer. Instead of utilizing a single learning rate η across all parameters, Adam leverages the first-order moment to provide directional persistence, while utilizing the second-order moment to scale the learning rate adaptively for each individual parameter. The update rule of Adam at step t is shown as follows:

$$\mathbf{g}_t = \nabla R_{\mathcal{B}}(\boldsymbol{\theta}_t) \quad (2.4)$$

$$\mathbf{m}_t = \beta_1 \mathbf{m}_{t-1} + (1 - \beta_1) \mathbf{g}_t \quad (2.5)$$

$$\mathbf{v}_t = \beta_2 \mathbf{v}_{t-1} + (1 - \beta_2) \mathbf{g}_t^2 \quad (2.6)$$

$$\hat{\mathbf{m}}_t = \frac{\mathbf{m}_t}{1 - \beta_1^t} \quad (2.7)$$

$$\hat{\mathbf{v}}_t = \frac{\mathbf{v}_t}{1 - \beta_2^t} \quad (2.8)$$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \frac{\eta}{\sqrt{\hat{\mathbf{v}}_t} + \epsilon} \odot \hat{\mathbf{m}}_t \quad (2.9)$$

where \mathbf{g}_t is the stochastic gradient evaluated on mini-batch \mathcal{B} at step t . The variables \mathbf{m}_t and \mathbf{v}_t are estimates of the first and second raw moments of the gradients, respectively, while $\hat{\mathbf{m}}_t$ and $\hat{\mathbf{v}}_t$ are their bias-corrected counterparts. The exponential decay rates β_1 and β_2 are hyperparameters that control the smoothness of these moving averages, and \odot denotes element-wise multiplication, with the square and square root operations also applied element-wise.

These adaptive, moment-driven optimizers are essential for training heavily parameterized models, where the high-dimensional loss landscape is highly non-convex, rugged, and complex.

2.1.2 Privacy-Preserving Distributed Learning

In many critical domains, data is permanently isolated across geographic silos or edge devices due to severe bandwidth constraints, unacceptable latency, and stringent legislative frameworks such as the GDPR or CCPA [27]. Consequently, the classical approach of collecting raw data into a single massive data center is often legally and logistically impossible.

FL has emerged as a specialized implementation designed to train models across a network of decentralized edge devices (cross-device FL) or organizational centers (cross-silo FL) without the raw data ever leaving its source [15].



Figure 2.1. An example application of FL for the task of next-word prediction on mobile phones [27]. To preserve the privacy of the text data and to reduce strain on the network, we seek to train a predictor in a distributed fashion, rather than sending the raw local data to a central server.

The primary innovation of FL is the inversion of the traditional training pipeline: it moves the computation to the data, rather than moving the data to the computation. This ensures that sensitive information, ranging from proprietary pharmaceutical structures to personal communications, remains under the local control of the client, thereby bypassing the physical and/or legal bottlenecks inherent in centralized data collection. This structural privacy is the defining characteristic that necessitates the development of federated optimization methods. Figure 2.1 illustrates the architecture of FL on an example application of next-word prediction for text input on mobile phones.

2.1.3 Federated Learning Formulation

To transition from centralized optimization to federated learning, we first establish a mathematical formulation. We define the following:

- **The Client Set:** Let \mathcal{K} denote the set of $|\mathcal{K}|$ participating decentralized entities, such as edge devices or organizational silos.
- **Local Datasets:** Each client $k \in \mathcal{K}$ possesses a private dataset \mathcal{D}_k of size $|\mathcal{D}_k|$, where the data is drawn from a local distribution $p_k(x, y)$.
- **Total Population:** The total number of samples across the entire network is given by $|\mathcal{D}_{\text{total}}| = |\bigcup_{k \in \mathcal{K}} \mathcal{D}_k|$.
- **Aggregation Weights:** We define $w_k = \frac{|\mathcal{D}_k|}{|\mathcal{D}_{\text{total}}|}$ as the relative weight of client k , such that $\sum_{k \in \mathcal{K}} w_k = 1$.

The fundamental objective in FL is to solve a *distributed optimization* problem where the goal is to find a set of model parameters $\theta \in \mathbb{R}^d$ that minimize the aggregate global risk $R(\theta)$. This is formulated as a *finite-sum optimization* problem:

$$\min_{\theta \in \mathbb{R}^d} R(\theta) = \sum_{k \in \mathcal{K}} w_k R_k(\theta) \quad (2.10)$$

where $R_k(\theta)$ represents the local objective function (empirical risk) for client k :

$$R_k(\theta) = \frac{1}{|\mathcal{D}_k|} \sum_{i \in \mathcal{D}_k} \ell(f_\theta(\mathbf{x}_i), y_i). \quad (2.11)$$

Here, $f_\theta(\cdot)$ denotes the predictive model parameterized by θ , and $\ell(\cdot)$ is a differentiable loss function.

2.1.4 Federated Optimization

Federated averaging (FedAvg) [15] is the canonical baseline for federated optimization. It combines local SGD on clients with a simple weighted averaging step on the server, and serves as the reference point for most subsequent algorithmic variants. With standard FedAvg, a coordinating server persistently maintains a global model parameterization θ_t , and training proceeds via synchronous communication rounds. FedAvg is summarized in Algorithm 1.

Algorithm 1 FedAvg

Require: Initial global parameters θ_0 , total rounds T , local steps E

for round $t = 0, \dots, T - 1$ **do**

 Server broadcasts θ_t to all clients $k \in \mathcal{K}$

for client $k \in \mathcal{K}$ **in parallel do**

$\theta_{t,k}^{(0)} \leftarrow \theta_t$ ▷ Initialize local model

$\theta_{t,k}^{(E)} \leftarrow \text{LocalSGD}(\theta_{t,k}^{(0)}, \mathcal{D}_k, E)$ ▷ Perform E local steps

end for

$\theta_{t+1} \leftarrow \sum_{k \in \mathcal{K}} \frac{|\mathcal{D}_k|}{|\mathcal{D}_{\text{total}}|} \theta_{t,k}^{(E)}$ ▷ Weighted aggregation

end for

Federated Optimization Framework

However, FedAvg restricts the server to a simple averaging operation and the clients to SGD. FedAvg can be generalized to the federated optimization (FedOpt) framework by explicitly decoupling the client-side local updates from the server-side global model update [34]. By conceptualizing the aggregated local model updates as a global pseudo-gradient, FedOpt allows for the application of advanced adaptive optimizers directly at both the server and the client level.

Algorithm 2 FedOpt

Require: Initial global parameters θ_0 , total rounds T , other hyperparameters

for round $t = 0, \dots, T - 1$ **do**

for client $k \in \mathcal{K}$ **in parallel do**

$\theta_{t,k} \leftarrow \text{ClientOpt}(\theta_t, \mathcal{D}_k)$ ▷ Local client optimization

$\Delta_{t,k} \leftarrow \theta_{t,k} - \theta_t$ ▷ Compute local pseudo-gradient

end for

$\Delta_t \leftarrow \text{Aggregate}(\{\Delta_{t,k}\}_{k \in \mathcal{K}})$ ▷ Compute the global pseudo-gradient

$\theta_{t+1} \leftarrow \text{ServerOpt}(\theta_t, \Delta_t)$ ▷ Global server optimization

end for

The FedOpt framework provides a generalized mathematical perspective beyond FedAvg. Specifically, if we define $\text{ClientOpt}(\cdot)$ as performing E epochs of local SGD with a learning rate η_l , $\text{Aggregate}(\cdot)$ as the weighted average of local pseudo-gradients, and $\text{ServerOpt}(\cdot)$ as 1 step GD with a learning rate of $\eta_g = 1.0$, FedOpt directly recovers the exact FedAvg algorithm detailed in Algorithm 1. Furthermore, by applying adaptive methods like Adam [33] or Yogi [35] as ServerOpt , creating algorithms like FedAdam or FedYogi, the framework can converge much faster than plain FedAvg without altering the fundamental distributed communication loop [34].

2.2 Learning from Heterogeneous Modalities

Beyond the challenge of physical data distribution, deep learning is increasingly tasked with synthesizing heterogeneous data structures simultaneously. The traditional intersection of isolated computer vision architectures and natural language processing techniques has recently given rise to complex, multi-modal foundation models, most prominently embodied by VLMs [36, 37]. These models radically shift the data paradigm by learning to inherently associate structural images and complex raw text.

2.2.1 Shared Representation

The fundamental challenge in multi-modal learning is bridging the semantic gap between inherently different data structures, such as a dense matrix of RGB

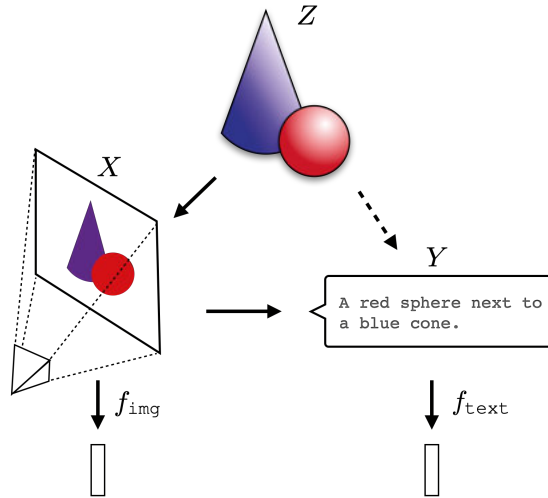


Figure 2.2. The platonic representation hypothesis [39]: Images (X) and text (Y) are projections of a common underlying reality (Z). It is conjectured that representation learning algorithms will converge on a shared representation of Z , and scaling model size, as well as data and task diversity, drives this convergence.

pixels for images and a sparse sequence of discrete language tokens for text. To perform meaningful mathematical comparison or joint reasoning, a system requires a *shared representation space*, a cohesive, continuous latent geometry where both modalities can directly interact. Without this joint representation space, text and images logically exist in completely isolated, incompatible coordinate bases [38].

A compelling theoretical framework for this alignment is the *platonic representation hypothesis* [39]. As illustrated in Figure 2.2, this hypothesis posits that disparate modalities, such as images (X) and text (Y), are merely different projections of a common underlying reality (Z).

Under this view, the goal of representation learning is not merely to "match" data types, but to converge upon a shared representation of the world itself. This convergence is driven by three primary factors:

- **Model Scaling:** Increasing the capacity of neural architectures allows for the discovery of deeper structural invariants.
- **Data Diversity:** Training on vast, uncurated datasets forces the model to ignore modality-specific noise in favor of universal semantic concepts.
- **Task Diversity:** Multi-objective learning ensures the resulting latent space is robust enough to support varied downstream applications.

Historically, researchers approached this bridging problem through heavily constrained or asymmetric algorithms. Early methodologies often relied on mapping one modality directly into the pre-established feature space of the other—for instance, projecting image features into fixed language embedding spaces like those from Word2Vec [40, 41, 42, 43]. Other methods utilized deep

cross-attention mechanisms to facilitate structural interaction deep within the network's internal layers [44, 45, 46].

While these methods proved that token-level fusion was possible, they often failed to achieve the "platonic" ideal. They struggled to extract compact, independently usable global embeddings and frequently required extensive task-specific fine-tuning, limiting their utility as general-purpose foundational models.

2.2.2 Contrastive Learning

In the last few years, *contrastive learning* has become the standard algorithm for building modern foundational VLMs [36, 47, 48]. By training on internet-scale datasets with billions of loosely paired image-caption combinations, this approach efficiently constructs shared representation spaces for visual and textual data. Instead of complex, tangled internal cross-attention, modern foundational architectures like contrastive language-image pre-training (CLIP) and sigmoid loss for image pre-training (SigLIP) [47] train two completely independent neural encoders: an image encoder f_I and a text encoder f_T .

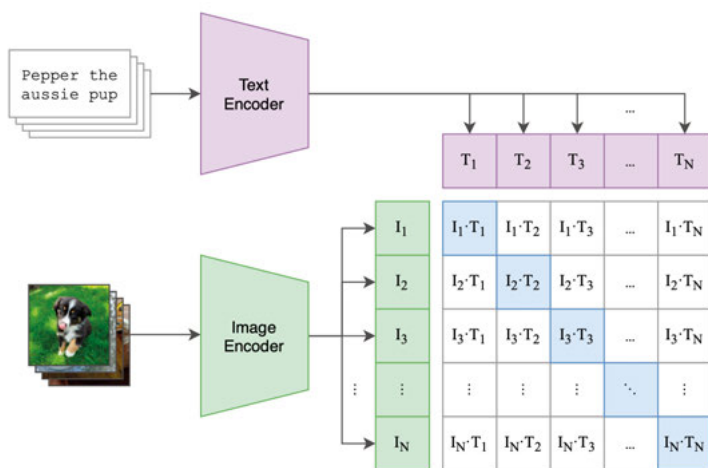


Figure 2.3. Summary of CLIP approach [47]. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of image-text training examples.

As illustrated in Figure 2.3, CLIP forces both of these distinct encoders to project their outputs exactly into a shared continuous latent embedding space. For a paired image-text batch $\mathcal{B} = \{(\mathbf{x}_i^{(I)}, \mathbf{x}_i^{(T)})\}_{i=1}^N$, where $\mathbf{x}_i^{(I)}$ and $\mathbf{x}_i^{(T)}$ are the image and text inputs respectively, let $\mathbf{I}_i = f_I(\mathbf{x}_i^{(I)})$ represent the normalized image embedding and $\mathbf{T}_i = f_T(\mathbf{x}_i^{(T)})$ represent the normalized text

embedding. During the pre-training phase, the model maximizes the geometric angle-based cosine similarity between correctly matched image-text pairs (the "positive" examples) while minimizing the cosine similarity between all other mismatched, incorrect pairings (the "negative" examples). This is achieved via the symmetric InfoNCE loss [49]. For the image-to-text direction, the loss for the i -th image is formulated strictly as a cross-entropy objective over the softmax similarities:

$$\mathcal{L}_{\text{CLIP}}^{(I \rightarrow T)} = -\log \frac{\exp((\mathbf{I}_i^\top \mathbf{T}_i)/\tau)}{\sum_{j=1}^N \exp((\mathbf{I}_i^\top \mathbf{T}_j)/\tau)} \quad (2.12)$$

where τ is a learnable temperature parameter scaling the logits. The total loss $\mathcal{L}_{\text{CLIP}}$ is computed symmetrically as the average of the image-to-text $\mathcal{L}_{\text{CLIP}}^{(I \rightarrow T)}$ and text-to-image $\mathcal{L}_{\text{CLIP}}^{(T \rightarrow I)}$ losses:

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2} \left(\mathcal{L}_{\text{CLIP}}^{(I \rightarrow T)} + \mathcal{L}_{\text{CLIP}}^{(T \rightarrow I)} \right). \quad (2.13)$$

The end goal is that semantically similar concepts are pushed closer to each other in the shared embedding space, while dissimilar concepts are pulled farther apart.

2.2.3 Geometric Constraint

A foundational mathematical aspect dictating both the optimization trajectory and downstream behavior of these contrastive VLMs is the geometric constraint on the learned shared representations. Before computing the contrastive loss, the high-dimensional output vectors from both encoders undergo ℓ_2 -normalization, which ensures training stability, mitigates vanishing gradients, and directs the optimization toward relative semantic alignment rather than vector magnitude.

Therefore, all representations are projected onto the surface of a high-dimensional unit hypersphere, denoted as $\mathbb{S}^{d-1} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 = 1\}$. Within this specific Riemannian geometry, all multi-modal embeddings strictly possess a uniform Euclidean norm exactly equal to 1, and distances between concepts are quantified exclusively in terms of relative angles traversing the curved surface of the sphere.

While this hyperspherical constraint proves effective for aligning dense images and sparse text into a joint space [50], it introduces structural challenges for subsequent probabilistic modeling. Applying standard "flat-space" Euclidean statistical tools, such as evaluating isotropic Gaussian distributions, fails when directly mapped onto \mathbb{S}^{d-1} . To ensure robustness of probabilistic modeling, directional probability distributions [51] are required for these pre-trained VLMs.

2.3 Probabilistic Frameworks

This section introduces the theoretical probabilistic tools utilized throughout this thesis. We first formalize uncertainty quantification as the tool for quantifying model confidence and ambiguity. Subsequently, we detail the core principles of representing complex data distributions, tracing the evolution from parametric forms to modern continuous-time generative models, with a specific focus on flow matching (FM).

2.3.1 Uncertainty Quantification

While standard neural networks act as deterministic functions, mapping inputs directly to uncalibrated point estimates, deploying models in critical domains necessitates a mathematically sound metric of their reliability. UQ provides the structured framework for estimating model confidence [52, 53, 54].

Uncertainty is classically decomposed into two distinct, mathematically separable components: *Aleatoric* uncertainty and *Epistemic* uncertainty [55, 56]. Aleatoric uncertainty captures the irreducible, inherent noise embedded within the observation process itself, such as sensor measurement error or fundamentally ambiguous phrasing in natural language. This uncertainty cannot be resolved by collecting more data. Conversely, epistemic uncertainty quantifies the model’s lack of knowledge regarding the optimal functional parameters, which is a direct result of finite or sparsely sampled training data. Unlike aleatoric uncertainty, epistemic uncertainty is reducible; it systematically vanishes as the model observes infinite training data.

In the context of complex geometries, such as the unit hypersphere governing contrastive VLMs, traditional variance-based UQ must be adapted. Accurately modeling both aleatoric and epistemic uncertainty in these non-Euclidean representation spaces requires defining geometrically valid probability distributions, mapping predictive variance to the angular dispersion of the learned embeddings.

2.3.2 Modelling Probability Distributions

The fundamental mathematical challenge underlying both UQ and generalized generative tasks is accurately learning and representing an underlying, often intractable, probability distribution with only access to i.i.d. samples from it [57].

The traditional baseline is to approximate a target distribution $p(\mathbf{x})$ by fitting a tractable, *parametric distribution* $p_{\theta}(\mathbf{x})$, such as a multivariate Gaussian, where θ represents the learnable parameters (e.g., means and covariance matrices) [58]. While computationally efficient with maximum likelihood estimation for the parameters, these predefined parametric forms severely lack the structural

expressivity to accurately capture highly non-linear and inherently multi-modal probability distributions [59].

To scale beyond restrictive simple parametric forms, the field evolved *deep generative models*. Early breakthroughs in this domain were driven by the variational autoencoders (VAEs) [60] and the generative adversarial networks (GANs) [61]. While revolutionary at their inception, VAEs often struggled with over-smooth outputs due to the explicit likelihood constraints of the evidence lower bound (ELBO) objective [62, 63], and GANs suffered from severe optimization instabilities and mode collapse [64, 65]. Consequently, they largely serve as the historical foundation for modern continuous-time approaches.

Diffusion Models

To overcome the severe structural limitations of VAEs and GANs, modern probability modeling has shifted towards continuous-time dynamical systems, primarily embodied by *diffusion models* [66, 67, 68].

Rather than attempting to map noise to a target distribution in a single, complex, non-linear jump, diffusion models conceptualize generation as a continuous temporal trajectory governed by a stochastic differential equation (SDE) system [69].

Diffusion models explicitly construct a "forward" diffusion process that incrementally corrupts the complex target distribution $p(\mathbf{x})$ into a simple prior distribution (typically a standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$) over a continuous time horizon $t \in [0, 1]$. This forward process is mathematically described by the Itô SDE:

$$d\mathbf{x} = f(\mathbf{x}, t)dt + g(t)d\mathbf{w} \quad (2.14)$$

where \mathbf{w} is the standard Wiener process, $f(\mathbf{x}, t)$ is a vector-valued drift coefficient, and $g(t)$ is a scalar diffusion coefficient [70, 69].

The generative capability arises from a fundamental theorem in stochastic calculus, specifically the Anderson-type reverse-time SDE formulation, which proves that this forward process can be reversed in time by simulating the corresponding reverse SDE [71]:

$$d\mathbf{x} = [f(\mathbf{x}, t) - g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x})] dt + g(t)d\bar{\mathbf{w}} \quad (2.15)$$

where $\bar{\mathbf{w}}$ is a standard Wiener process when time flows backwards from $t = 1$ to $t = 0$, and dt is an infinitesimal negative time step. The only unknown quantity required to simulate this reverse trajectory and generate high-fidelity samples from pure noise is the Stein score vector field $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ [72]. By leveraging deep neural networks $s_{\theta}(\mathbf{x}, t) \approx \nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ trained via denoising score matching [73, 74], diffusion models solve this complex probability modeling task [75].

Flow Matching

While effective, score-based diffusion methods often require hundreds of slow iterative integration steps to accurately track the trajectories induced by standard drift-diffusion formulations [69]. FM resolves this inefficiency by directly matching deterministic vector fields rather than relying on complex stochastic reverse diffusion [76, 77].

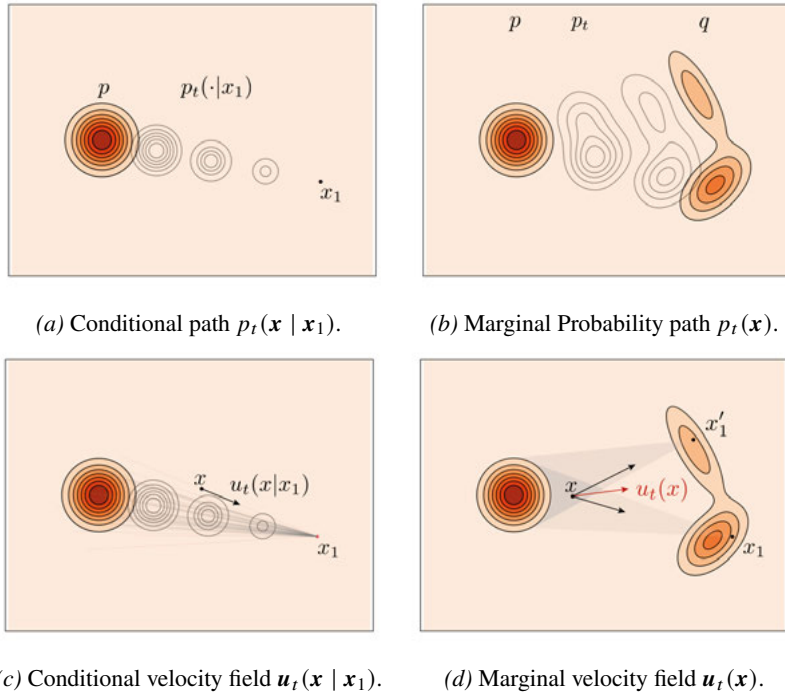


Figure 2.4. Path design in FM [78]. Given a fixed target sample \mathbf{x}_1 , its conditional velocity field $\mathbf{u}_t(\mathbf{x} \mid \mathbf{x}_1)$ generates the conditional probability path $p_t(\mathbf{x} \mid \mathbf{x}_1)$. The (marginal) velocity field $\mathbf{u}_t(\mathbf{x})$ results from the aggregation of all conditional velocity fields—and similarly for the probability path $p_t(\mathbf{x})$.

Instead of learning to reverse a noisy diffusion process, FM constructs a time-dependent, continuous probability path $p_t(\mathbf{x})$ linking a simple base distribution $p_0(\mathbf{x})$ at $t = 0$ to the complex target distribution $p_1(\mathbf{x})$ at $t = 1$, as shown in Figure 2.4. The goal is to directly train a vector field $\mathbf{v}_\theta(\mathbf{x}, t)$ parameterized by a neural network to regress against the true, yet computationally intractable, target generating vector field $\mathbf{u}_t(\mathbf{x})$ [79]. The time evolution of samples $\mathbf{x}(t)$ follows an ordinary differential equation (ODE):

$$\frac{d\mathbf{x}}{dt} = \mathbf{u}_t(\mathbf{x}) \quad (2.16)$$

Because the marginal target vector field $\mathbf{u}_t(\mathbf{x})$ is generally intractable to compute over the entire true distribution, FM bypasses this via *conditional flow*

matching [77]. Given a sampled point $\mathbf{x}_1 \sim p_1(\mathbf{x}_1)$ and a noise sample $\mathbf{x}_0 \sim p_0(\mathbf{x}_0)$, one specifies a tractable conditional probability path $p_t(\mathbf{x} | \mathbf{x}_1)$ and its corresponding conditional vector field $\mathbf{u}_t(\mathbf{x} | \mathbf{x}_1)$. The neural network simply minimizes the expected mean-squared error against this specific conditional vector field:

$$\mathcal{L}_{\text{CFM}}(\boldsymbol{\theta}) = \mathbb{E}_{t, \mathbf{x}_1, \mathbf{x}_0} [\|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) - \mathbf{u}_t(\mathbf{x}_t | \mathbf{x}_1)\|^2] \quad (2.17)$$

A theoretical advantage of FM is that it naturally permits the utilization of straight-line, deterministic *optimal transport* trajectories [80, 81]. For example, by defining the conditional path as a simple linear interpolation $\mathbf{x}_t = (1 - t)\mathbf{x}_0 + t\mathbf{x}_1$, the corresponding conditional vector field simplifies to the constant velocity $\mathbf{u}_t(\mathbf{x}_t | \mathbf{x}_1) = \mathbf{x}_1 - \mathbf{x}_0$. These straight-line ODE dynamics simplify the underlying numerical landscape, enabling orders-of-magnitude faster, highly accurate sampling using only a fraction of the discretization steps required by its SDE counterparts, i.e., diffusion models.

Furthermore, FM extends generative modeling beyond standard Euclidean parameter spaces [82]. By constructing the stochastic interpolants [76] following geodesics on Riemannian manifolds, FM inherently unlocks continuous-time generative capabilities natively atop complex non-Euclidean topologies, such as the hyperspherical representation spaces defining modern contrastive VLMs.

2.3.3 Simulation-Based Inference

The goal of scientific modeling is often not just to generate predictions, but to reason backwards from observations to the underlying causes that produced them. Bayesian inference provides the principled mathematical framework for this inverse reasoning. Given observed data \mathbf{x}_{obs} and a parametric model governed by parameters $\boldsymbol{\theta}$, Bayes' theorem combines prior beliefs $p(\boldsymbol{\theta})$ with the likelihood of the data $p(\mathbf{x} | \boldsymbol{\theta})$ to yield the posterior distribution:

$$p(\boldsymbol{\theta} | \mathbf{x}_{\text{obs}}) = \frac{p(\mathbf{x}_{\text{obs}} | \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathbf{x}_{\text{obs}})} \quad (2.18)$$

where $p(\mathbf{x}_{\text{obs}}) = \int p(\mathbf{x}_{\text{obs}} | \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$ is the marginal likelihood, or evidence.

In practice, however, many scientific domains rely on complex mechanistic simulators to model physical, biological, or social phenomena. These simulators define a forward process from parameters $\boldsymbol{\theta}$ to synthetic observations \mathbf{x} . However, because these simulators are often non-differentiable or computationally expensive, the likelihood $p(\mathbf{x} | \boldsymbol{\theta})$ is mathematically intractable, rendering exact Bayesian inference impossible for most complex real-world systems.

Likelihood-Free Origins: ABC and SMC-ABC

Historically, the problem of intractable likelihoods was addressed via approximate Bayesian computation (ABC). In its simplest form, ABC relies on a

rejection-based mechanism: parameters are sampled from the prior, $\theta \sim p(\theta)$, and used to simulate synthetic data \mathbf{x}_{sim} [83]. The parameter θ is accepted only if the discrepancy between \mathbf{x}_{sim} and the empirical observation \mathbf{x}_{obs} is below a small threshold ϵ . To improve the efficiency of this rejection process, sequential Monte Carlo (SMC)-ABC was developed. SMC-ABC utilizes a population of particles that evolve through a sequence of intermediate distributions with decreasing ϵ values, using importance sampling to maintain a representative sample of the posterior without the computational waste of vanilla rejection sampling [84].

The Neural Regime: Amortized Inference and NPE

The modern regime of SBI shifts away from these simulation-intensive sampling methods toward *amortized inference* using deep neural networks. neural posterior estimation (NPE) utilizes a neural density estimator (typically a normalizing flow) to directly learn the mapping from observations to the posterior distribution $p(\theta | \mathbf{x})$. Once trained on a dataset of simulated (θ, \mathbf{x}) pairs, these models can produce posterior estimates for new observations instantaneously, bypassing the need for new simulations during inference [85].

To further focus computational resources on a specific observation \mathbf{x}_{obs} , sequential extensions of NPE, sequential neural posterior estimation (SNPE), have been developed:

- **SNPE-A:** The earliest iteration, which used a proposal distribution to focus simulations but required a post-hoc correction to handle the resulting biased posterior [85].
- **SNPE-B:** Improved upon this by utilizing an importance-weighting scheme within the loss function to directly target the true posterior [86].
- **SNPE-C (Automatic Posterior Transformation):** Re-parameterizing the problem as a conditional density estimation task, allowing for more stable and efficient training across multiple rounds of simulations [87].

Generative Scaling: Flow Matching and Diffusion

While normalizing flows provide a robust baseline, recent advancements have introduced more expressive generative paradigms to SBI including variational autoencoders, diffusion models and flow matching [88, 89, 90]. By reframing SBI as a conditional generation task, these methods effectively capture the multi-modalities and complex dependencies inherent in high-dimensional scientific data.

3. Robustness in Federated Learning

Having established the theoretical foundations in Chapter 2, we now turn to the first dimension of robustness. This chapter builds on the distributed optimization framework introduced in Section 2.1 and examines how statistical heterogeneity in federated learning challenges the convergence and fairness guarantees of standard algorithms.

3.1 Statistical Heterogeneity

Although federated learning resembles traditional distributed training at a structural level, the two frameworks differ in the core assumption regarding the underlying data distribution. In conventional distributed learning within centralized data centers, a consolidated dataset is intentionally shuffled before being partitioned across computing nodes, so that every shard is a statistically unbiased, i.i.d. representation of the global population.

FL offers no such guarantee. Because data is generated by distinct local entities, such as individual mobile phones, hospitals, or proprietary research laboratories, the data residing on any given client is tethered to that client's specific context, behavior, and environment [91]. This *statistical heterogeneity*, rather than being an incidental complication, is the defining characteristic of the federated setting [92]. Achieving robustness therefore requires a fundamental redesign of how global optimization handles divergent local model updates.

To characterize the nature of this heterogeneity, the joint distribution $p(\mathbf{x}, y)$ can be decomposed in two primary ways: $p(\mathbf{x}, y) = p(\mathbf{x})p(y | \mathbf{x})$ and $p(\mathbf{x}, y) = p(y)p(\mathbf{x} | y)$. By analyzing these decompositions, we can categorize the statistical heterogeneity into distinct shifts [19, 93]:

- **Covariate Shift:** The marginal feature distribution $p(\mathbf{x})$ varies across clients, but the conditional distribution $p(y | \mathbf{x})$ remains the same. This often occurs due to different sensors or environmental conditions capturing the same underlying phenomena [94].
- **Prior Probability Shift:** The marginal label distribution $p(y)$ changes across clients, while the conditional distribution $p(\mathbf{x} | y)$ remains consistent. This is common when different clients encounter classes at non-uniform frequencies [17].
- **Concept Shift:** The conditional distributions $p(y | \mathbf{x})$ or $p(\mathbf{x} | y)$ differ across clients, meaning the relationship between features and labels changes from one node to another [95].

In real-world applications, such as healthcare or scientific research, data is generally acquired across different environments or institutions. In such scenarios, localized "batch effects" or demographic variances dominate the data distributions. These systematic variations introduce covariate and concept shifts across the distributed network, violating the i.i.d. assumption.

3.1.1 Optimization Instability

Because these local loss landscapes differ, each client's model often moves toward a different local minimum during local training [96]. This creates two major issues for the optimization instability of FL:

- **Incompatible Updates:** When the server averages these diverging weight vectors, the resulting "global" model often lands in a high-loss region of the global landscape. Essentially, the average of well-performing local models does not yield a model that performs well on the global dataset or individual local datasets [97].
- **Client Drift:** As this process repeats, the global model's optimization path begins to oscillate or move away from the true global optimum. This phenomenon is known as client drift [93].

In the presence of high statistical heterogeneity, the gradient diversity inherent in non-i.i.d. data introduces a tension between communication efficiency and global optimality [98]. This mismatch results in a suboptimality gap, where standard FedAvg often exhibits a performance plateau that fails to recover the empirical risk minimization's performance achievable in a centralized training environment [99].

3.1.2 Fairness Problem

Beyond this optimization instability, the non-i.i.d. nature of federated networks inherently induces a *fairness* problem. A "fair" global model is one that delivers consistent and equitable performance across all participating clients, regardless of their individual dataset sizes or specific distributions. However, without structural adaptivity, models trained by FedAvg overfit to a specific subset of the empirical data distribution, resulting in high average accuracy that masks severe performance inconsistencies at the local level [19, 91, 18, 100]. To achieve algorithmic robustness and fairness simultaneously, the optimization method must prevent any subset of clients from dictating the global path through structural adaptivity at the server aggregation level.

The remainder of this chapter presents two complementary contributions that address these challenges: Paper I provides an empirical investigation of FL robustness in a real-world drug discovery application, while Paper II develops the algorithmic tools to resolve the fairness-convergence trade-off.

3.2 Empirical Study: Federated Learning for Drug Discovery

The pharmaceutical industry exemplifies the data silo problem that motivates federated learning. Due to intellectual property protections, regulatory constraints, and the competitive value of proprietary chemical datasets, organizations are rarely able to centralize their data [101]. Paper I investigates whether FL can bridge this gap by collaboratively training deep neural networks on distributed Cell Painting image data to predict a compound’s mechanism of action (MoA) from fluorescence microscopy images [102].

The dataset contains 13,878 images of U2OS cells exposed to 231 distinct compounds, spanning 10 different MoA classes. To systematically evaluate how the heterogeneity issues described in Section 3.1 manifest in practice, we designed three partitioning scenarios of increasing severity: Uniform (i.i.d.), Unbalanced (varying local dataset sizes), and Non-i.i.d. (severe prior probability shifts). Each setup partitioned the data strictly at the compound level to prevent target leakage, creating an environment where a subset of public data was shared among clients, augmented by strictly private local partitions. The models, specifically adapted AlexNet [103] and VGG13 [104] architectures, were trained using the FedAvg algorithm within FEDn [105], a production-grade FL framework deployed across multiple Swedish high-performance computing (HPC) clusters, as shown in Figure 3.1.

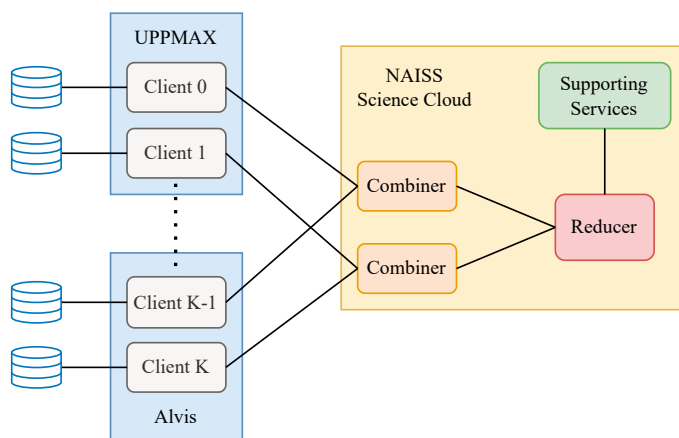


Figure 3.1. Training framework architecture [106]. The training framework, FEDn, is geographically distributed across three HPC clusters: NAISS Science Cloud, UPPMAX, and Alvis, located in Umeå, Uppsala, and Gothenburg, Sweden, respectively. Within this framework, components in the NAISS Science Cloud are responsible for aggregating local models in a hierarchical manner, supported by essential services. Local training occurs on HPC clusters equipped with GPU resources, specifically UPPMAX Snowy and Alvis. The partitioned data are stored on the respective HPC clusters and can only be accessed by their assigned owners.

In both the Uniform and Unbalanced scenarios, federated models outperformed models trained on isolated local datasets and did not significantly underperform their centralized counterparts, confirming that FL bridges the performance gap caused by data fragmentation. More importantly, we examined the fairness problem at the MoA class level. Local models frequently exhibited biases driven by their limited data distributions, while federated training effectively mitigated this localized bias, achieving per-class precision and recall nearly identical to centralized models.

The Non-i.i.d. scenario introduced label skew by simulating a "specialized" client holding exclusive data for specific MoA classes, which is a form of prior probability shift described in Section 3.1. Contrary to the theoretical expectation of severe client drift, our results showed that including such specialized participants is beneficial: the global model's accuracy for the rare MoA increased by up to 40% without catastrophic forgetting, and average accuracy across all other classes improved simultaneously. This demonstrates that federated learning can robustly extract knowledge from specialized local data without sacrificing global stability.

In summary, Paper I establishes the practical viability of FL for drug discovery under real-world heterogeneity. However, the study also reveals the limitations of standard FedAvg: while it handles moderate heterogeneity, it provides no principled mechanism to control the fairness-convergence trade-off. Paper II addresses this gap.

3.3 Accelerating Fair Federated Learning

Existing fairness-aware methods, such as q -FedAvg [19], attempt to mitigate the performance bias described in Section 3.1 by exponentially over-weighting clients with higher losses. However, this introduces gradient scaling issues, forcing a trade-off where improved fairness comes only at the cost of convergence degradation. Paper II resolves this bottleneck both at the formulation and solver levels.

At the formulation level, we reformulate fairness-aware federated learning as a dynamic multi-objective optimization (DMOO) problem. Rather than relying on rigid scaling factors that diminish the scale of gradients over time, the DMOO formulation introduces a dynamic inverse training rate, $I_k(t)$, which quantifies the training progress of each client k at round t . The trade-off between fairness and optimization efficiency is controlled by a tunable fairness hyperparameter α . We theoretically prove that the DMOO formulation shares the exact same (p, α) -proportional fairness guarantees as the standard q -fairness formulations in [19], but constructs a loss landscape that is more compatible with first-order distributed optimization methods by construction.

At the solver level, we analyze the limitations of standard FedAdam in non-i.i.d. settings and demonstrate that it relies on biased pseudo-gradients. Because

local objective functions across heterogeneous clients possess wildly varying Lipschitz smoothness constants, simply aggregating local updates implicitly over-weights updates from clients with flatter local loss landscapes, pulling the global model away from the true optimum.

To fix these flaws, we propose AdaFedAdam, which introduces three corrective steps. First, it normalizes local updates based on the ℓ_2 -norm of local gradient estimations, neutralizing biases from varying local Lipschitz constants. Second, it calculates a confidence score for each local update, reflecting directional reliability. Third, it uses the aggregated confidence scores to dynamically adjust the server-side learning rate (η) and momentum decay factors (β_1, β_2). We formally prove the convergence guarantee for AdaFedAdam on non-convex functions with Lipschitz gradients, matching the optimal $O(1/T)$ rate while mitigating the bias inherent to standard FedAdam.

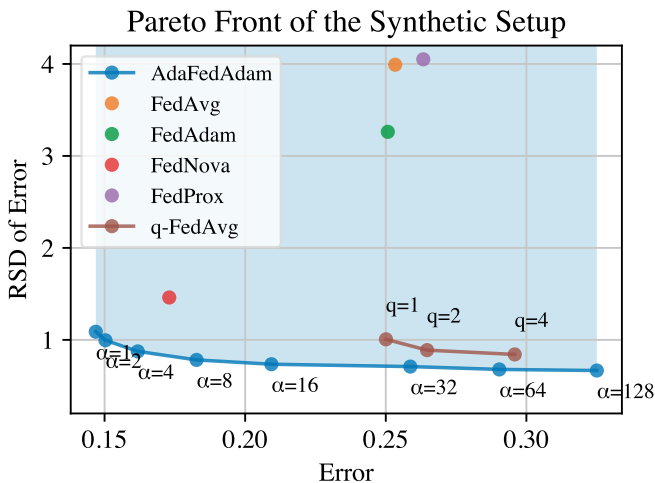


Figure 3.2. Pareto front formed by AdaFedAdam with different α [107]. The x -axis represents the average error of clients as a measure of convergence, and the y -axis represents the relative standard deviation of client error as a measure of fairness. By adjusting values of α , the trade-off between the convergence and the fairness can be observed together with the suboptimality of other federated algorithms.

The empirical results demonstrate that AdaFedAdam achieves up to $6.58\times$ speedup over FedAvg in reaching target accuracy, while simultaneously outperforming all baselines in fairness metrics (smaller standard deviation of client accuracy, higher accuracy among the 30% worst-performing clients). Across all settings, as shown in Figure 3.2, AdaFedAdam consistently established a new state-of-the-art Pareto frontier, breaking the traditional fairness-convergence trade-off. The algorithm also maintained its superiority under resource and data heterogeneity of different levels (varying local steps and Dirichlet concentration parameters), confirming its structural robustness.

4. Robustness in Vision-Language Models

The second dimension of robustness concerns how a pre-trained multi-modal model interprets unfamiliar inputs and quantifies its own ignorance [47, 48, 108]. VLMs present new challenges in this domain because they map heterogeneous modalities structurally onto complex, non-Euclidean topologies [109]. As introduced in Chapter 2, robustness for these models requires uncertainty quantification, the capacity to decompose and measure the distinct forms of uncertainty inherent within their high-dimensional latent representations [110, 111]. Without it, VLMs produce overconfident predictions, and an inability to reliably detect OOD data leads to silent, undetected failures [112, 113, 114]. This chapter addresses both components of UQ introduced in Section 2.3.1: Section 4.1 develops geometry-aware aleatoric uncertainty quantification via directional distributions, and Section 4.2 tackles epistemic uncertainty quantification through manifold-native density estimation.

4.1 Aleatoric Uncertainty and Modality Asymmetry

As discussed in Section 2.3.1, aleatoric uncertainty represents the inherent, irreducible ambiguity within the data itself. Traditional approaches for modeling this uncertainty assume flat Euclidean latent spaces. However, as introduced in Section 2.2.3, contrastive VLMs project all embeddings onto the unit hypersphere \mathbb{S}^{d-1} via ℓ_2 -normalization, rendering standard Euclidean Gaussian distributions geometrically misaligned and producing uncalibrated predictions [115, 116].

Constructing a robust probabilistic framework for these VLMs requires adhering to the underlying non-Euclidean topology and addressing the structural *asymmetry* of aleatoric uncertainty between different modalities [20]. A discrete text caption (e.g., "a dog") is a semantic abstraction. It carries inherent aleatoric ambiguity, mapping logically to thousands of highly distinct, diverse, valid instantiations, occupying a large area on the hypersphere in the representation space [116]. In contrast, a specific, dense input image is a deterministic, concrete physical instantiation possessing virtually zero inherent structural ambiguity, occupying a highly concentrated, discrete point in the representation space. While recent post-hoc probabilistic adaptation methods [110, 111] have attempted to quantify uncertainty of pre-trained models, they overlook both this structural asymmetry and the non-Euclidean geometry constraint, applying symmetric Euclidean Gaussian distributions to both modalities.

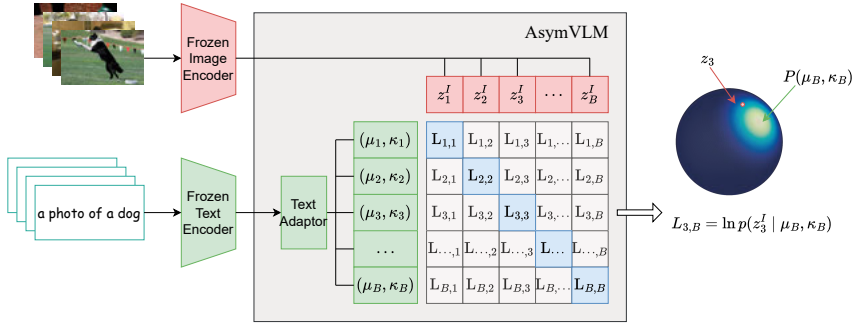


Figure 4.1. Overview of the AsymVLM framework [117]. Texts are encoded with a frozen text encoder and adaptor to produce probabilistic hyperspherical embeddings (e.g., vMF distribution), while images are deterministically encoded via a frozen image encoder. The log likelihood matrix L , with element $L_{m,n}$ representing the log likelihood of image vector z_n^I given text embedding $z_m^T \sim P(\mu_m, \kappa_m)$, is optimized using InfoNCE to maximize diagonals and minimize off-diagonals.

Paper III formalizes this asymmetric uncertainty structure by analyzing the mapping relations between text, image, object, and representation spaces. The analysis reveals that the uncertainty in text-to-image retrieval, rooted in semantic abstraction and visual variability, is fundamentally different from the uncertainty in image-to-text retrieval, rooted in descriptive variance: the former is dominated by aleatoric uncertainty, while the latter by epistemic uncertainty. Recognizing this structural asymmetry motivates an asymmetric probabilistic framework.

Further, we propose AsymVLM, a lightweight post-hoc adaptation framework that explicitly leverages modality asymmetry. As shown in Figure 4.1, AsymVLM models text embeddings as random variables governed by directional distributions, specifically, the vMF and power spherical distribution parameterized by an orientational mean and a measurable concentration parameter [51, 118], residing on the unit hypersphere. Conversely, image embeddings are maintained as deterministic point estimates on the same manifold. An adaptor projects frozen text features into the parameters of the directional distribution, and the loss function is constructed by integrating maximum likelihood estimation into an InfoNCE-based objective. This ensures that AsymVLM is a natural probabilistic extension of the standard CLIP and SigLIP loss, utilizing the concentration parameter to explicitly capture text ambiguity without drifting from the pre-trained semantic space.

Empirical validation across various established benchmarks demonstrated that AsymVLM achieves superior UQ compared to baseline methods. Ablation studies confirmed that both the asymmetric structure and the geometric constraint are essential for the performance. To intuitively verify what the learned uncertainty represents, we conducted an analysis using the HierarCap dataset,

which pairs images with captions across four hierarchical levels of abstraction. AsymVLM accurately captured this semantic hierarchy without any explicit supervision, confirming that the model’s statistical outputs faithfully mirror linguistic ambiguity.

Finally, when deployed for a "none-of-the-above" zero-shot classification task, AsymVLM outperformed deterministic threshold-based and margin-based rejection baselines. By mapping a dummy prompt like "a photo" to a wide-variance spherical distribution, the model rejected negative samples while maintaining high accuracy on positive in-distribution samples.

However, the vMF distributions used to model aleatoric uncertainty are inherently limited by their unimodal structural shapes. They cannot capture the complex, multi-modal distribution of internet-scale data across the hypersphere, nor can they identify structurally sparse regions where the model possesses little training data. Addressing this complementary form of uncertainty, epistemic uncertainty, requires a different approach, which is the subject of the next section.

4.2 Epistemic Uncertainty via Density Estimation

While Paper III addresses aleatoric uncertainty through parametric directional distributions, a complete UQ framework must also address *epistemic uncertainty*—the model’s lack of knowledge about inputs outside its training distribution (Section 2.3.1). Standard Bayesian estimators such as Deep Ensembles [119] or Monte Carlo Dropout [120] are impractical for foundation-scale VLMs due to the prohibitive cost of running multiple forward passes through backbones. Paper IV tackles this bottleneck by reformulating epistemic uncertainty as a problem of manifold-native density estimation.

The core theoretical contribution is establishing a link between the probability density of an embedding and the model’s confidence for the input. We demonstrate that the epistemic uncertainty of a Bayesian encoder, defined as the trace of the covariance matrix of its embedding over the parameter posterior, statistically increases with the norm of the Parameter-Jacobian. In contrastive learning setups where the encoder achieves dynamical isometry, the training objective implicitly minimizes this Jacobian norm in high-density training regions. Consequently, an embedding in a sparse, low-density region exhibits a high Parameter-Jacobian norm, directly signaling model ignorance. Through this theoretical chain, we prove that the negative log-density $-\log p(z)$ serves as a principled, scalable proxy for epistemic uncertainty.

Based on this analysis, we propose RepVLM, a framework that leverages conditional Riemannian flow matching to estimate the probability density of pre-trained VLM embeddings directly on the hypersphere, as shown in Figure 4.2. RepVLM learns a time-dependent, modality-conditioned vector field that defines a continuous probability flow, transporting a uniform base

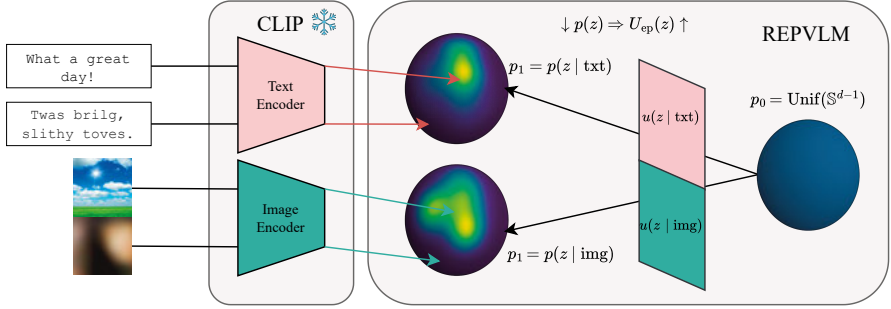


Figure 4.2. Overview of RepVLM. The framework estimates the probability density $p(z)$ of pre-trained VLM embeddings on the hypersphere \mathbb{S}^{d-1} . A unified model learns a vector field v_t that transports a simple uniform base distribution $P_0 = \text{Unif}(\mathbb{S}^{d-1})$ to the empirical modality-specific distributions P_1 (Image and Text). Standard inputs map to high-density regions (yellow), while ambiguous or out-of-distribution inputs such as the distorted image or nonsensical text reside in low-density regions (purple). The negative log-likelihood $-\log p(z | c)$ thus serves as a principled proxy for epistemic uncertainty $U_{\text{ep}}(z)$.

distribution on \mathbb{S}^{d-1} along geodesic paths to match the complex empirical target distributions of image and text embeddings. During inference, the log-likelihood of a new query embedding is efficiently computed by solving a reverse ODE using a Riemannian-adapted simulation method, integrating the divergence of the vector field via a manifold-constrained Hutchinson trace estimator.

Across selective classification tasks on a wide array of benchmarks, RepVLM demonstrated superior uncertainty calibration, outperforming baselines such as [110] and [120]. Crucially, by decoupling density estimation from the heavy VLM backbone, RepVLM requires only a lightweight, three-layer residual fully-connected network operating in the latent space, delivering superior calibration at only 10% of the computational cost of Monte Carlo Dropout. Beyond selective classification, RepVLM’s density scores proved effective for robust OOD detection and automated data curation, reliably filtering out nonsensical and corrupted inputs by flagging samples in low-density regions.

Together with Paper III, this work provides a complete geometry-aware UQ framework for VLMs: AsymVLM captures the inherent ambiguity of language through directional aleatoric distributions, while RepVLM identifies the boundaries of the model’s knowledge through manifold-native epistemic density estimation.

5. Robust Simulation-Based Inference

The final dimension of robustness explored in this thesis concerns the reliability of SBI [25]. As discussed in Section 2.3.3, when the likelihood $p(\mathbf{x} \mid \boldsymbol{\theta})$ of a complex mechanistic simulator is mathematically intractable, SBI provides a family of methods for approximating the posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{x})$ from simulated data [121, 122]. While classical approaches such as ABC achieve this through rejection or regression without learning an explicit density model [83], recent neural SBI methods train deep conditional density estimators on simulated parameter-observation pairs to amortize inference [25]. However, deploying neural density estimators for such an inverse problem introduces distinct challenges that must be resolved for robust real-world SBI [123, 124].

5.1 Challenges for Robust Inference

Standard NPE methods, while effective in idealized settings, rely on two implicit assumptions that frequently break down in practice: that posterior estimation is the only inference task of interest, and that observations are always complete and well-formed. Relaxing these assumptions reveals two fundamental challenges for robust SBI.

The Need for Forward Surrogates.

Focusing exclusively on posterior estimation (mapping from data to parameters) leaves researchers entirely dependent on the slow, expensive original simulator for any forward modeling. In practice, approximating the likelihood $p(\mathbf{x} \mid \boldsymbol{\theta})$ is frequently as important as predicting the posterior: fast access to an accurate forward surrogate enables rapid iteration, validation of parameter estimates, and deeper understanding of the underlying physical problem [125, 121]. Traditional approaches necessitate training entirely separate architectures for posterior estimation and forward simulation, a disjointed strategy that is computationally wasteful and mathematically fragile [125].

Inference with Partially Observed Data.

The second, and perhaps more severe, challenge arises from the reality of empirical data collection. In many real-world scenarios, collected observations are incomplete, with the specific subset of missing variables varying across observations due to sensor failures, differing experimental setups, or corrupted data pipelines [126, 127]. Standard neural density estimators, trained to ingest

a fixed-size vector, fundamentally break when confronted with variable-length or masked inputs [127]. It is structurally impossible for these architectures to perform mathematically sound inference on dynamic, partially observed data without resorting to ad-hoc imputation techniques or training an exponentially large ensemble of specialized models for every possible combination of missing features.

To overcome both the fragility of disjointed architectures and the rigidity of fixed-input models, robust SBI requires transitioning to an *all-in-one* methodology that unifies forward simulation and inverse inference within a single architecture, while natively handling incomplete observations [128, 129].

5.2 Unified Joint Inference with OneFlowSBI

Paper V resolves both challenges identified in Section 5.1 by introducing OneFlowSBI, an all-in-one framework that abandons single-task estimators and instead learns the complete joint distribution of parameters and observations, $p(\boldsymbol{\theta}, \mathbf{y})$. By capturing the holistic dependencies of the joint space, a single trained model can autonomously answer a vast spectrum of Bayesian queries, including posterior sampling, likelihood estimation, marginal generation, and arbitrary mixed conditionals, without requiring any task-specific retraining.

OneFlowSBI leverages continuous normalizing flows trained via flow matching [77]. We define a joint state $\mathbf{z} = [\boldsymbol{\theta}; \mathbf{y}] \in \mathbb{R}^{d_{\boldsymbol{\theta}}+d_{\mathbf{y}}}$ and introduce a dynamic binary mask $\mathbf{m} \in \{0, 1\}^{d_{\boldsymbol{\theta}}+d_{\mathbf{y}}}$. During training, this mask explicitly defines which variables are treated as fixed observations (conditioned) and which are generated. The model learns a time-dependent, masked vector field $\mathbf{v}_t^\phi(\mathbf{z}_t, \mathbf{m})$ that transports a standard Gaussian base distribution along straight-line optimal transport paths exclusively on the unobserved coordinates. The training objective minimizes the conditional flow matching loss with randomized masking. We prove via a Conditional Continuity Equation that integrating this masked ODE yields a valid, mass-conserving conditional probability flow, theoretically ensuring that altering the binary mask \mathbf{m} at inference time allows the model to instantly pivot between different tasks.

The framework is subjected to empirical evaluation, including ten canonical tasks from the SBI Benchmark [130] and two real-world inverse problems. OneFlowSBI demonstrates competitive or superior performance compared to both specialized NPE baselines and recent generalized Transformer-based architectures like Simformer. As illustrated in Figure 5.1, it can perform inference for arbitrary conditional or marginal distributions by simply varying the inference mask.

Stress-testing confirms the framework’s robustness to the challenges outlined in Section 5.1. Under additive Gaussian noise, OneFlowSBI maintains stable posterior inference, degrading only smoothly when noise severely dominates the signal. In the missing data ablation, OneFlowSBI produces plausible posteriors

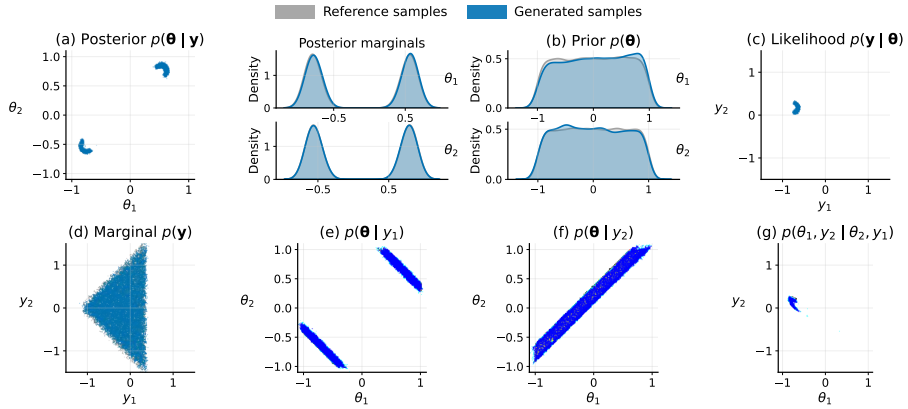


Figure 5.1. Multi-query inference on Two Moons using OneFlowSBI. We target diverse conditional distributions $p(\cdot|\cdot)$ and marginal distributions $p(\cdot)$ by varying the inference mask. Panels show the (a) posterior, (b) prior, (c) likelihood, and (d) evidence, alongside (e–g) arbitrary partial conditionals.

even when up to 50% of observation coordinates were entirely masked at test time, directly addressing the partial observability challenge without any ad-hoc imputation or retraining. Furthermore, the near-straight transport geometry induced by the masked linear interpolants enables highly accurate posterior sampling with merely 2 to 3 ODE integration steps, making the framework efficient for rapid, multi-query exploratory analysis.

In conclusion, by unifying forward simulation and inverse inference into a single, dynamically masked generative architecture, OneFlowSBI eliminates the fragility of traditional specialized estimators, providing a scalable and robust tool for navigating the incomplete, noisy realities of real-world scientific data.

6. Summary of Papers

Paper I

Title: Federated learning for predicting compound mechanism of action based on image-data from cell painting
Authors: **Li Ju**, Andreas Hellander, and Ola Spjuth
Venue: Artificial Intelligence in the Life Sciences 5 (2024): 100098

Paper I provides an empirical study into the practicality of federated learning for drug discovery, specifically focusing on predicting the MoA from cell painting image data. While theoretical studies often suggest that data heterogeneity in FL is detrimental to model convergence and performance, this work demonstrates that such pessimism may be overstated in real-world applications. By simulating diverse collaborative scenarios including uniform, unbalanced, and non-i.i.d. data distributions, the study shows that FL models not only significantly outperform models trained on isolated local datasets but also achieve predictive accuracy comparable to centralized, data-sharing baselines. Notably, the inclusion of "specialized" participants who own data for unique MoA classes was shown to enhance the global model's capabilities for those specific categories without introducing significant fairness or convergence issues. In summary, with empirical evidence, this paper proves that effective model training via FL is achievable even in the face of significant real-world data heterogeneity, for large-scale collaborative drug discovery.

Paper II

Title: Accelerating fair federated learning: Adaptive federated adam
Authors: **Li Ju**, Tianru Zhang, Salman Toor, and Andreas Hellander
Venue: IEEE Transactions on Machine Learning in Communications and Networking 2 (2024): 1017-1032

Paper II addresses the dual challenges of convergence loss and the "fairness problem" in FL when dealing with non-i.i.d. data. This work reformulates fairness-aware federated learning as a DMOO problem, which has a theoretical guarantee of fairness. To solve the DMOO problem, the paper proposes AdaFedAdam, which incorporates the normalization of local updates and the dynamic adjustment of hyperparameters based on the estimated certainty

of pseudo-gradients. Theoretical analysis provides a convergence guarantee for the proposed method in general non-convex settings. The effectiveness of AdaFedAdam was validated through extensive experiments on diverse benchmarks. The results demonstrate that AdaFedAdam consistently achieves Pareto frontiers of fairness and convergence compared to existing methods.

Paper III

Title: Exploiting the Asymmetric Uncertainty Structure of Pre-trained VLMs on the Unit Hypersphere
Authors: **Li Ju**, Max Andersson, Stina Fredriksson, Edward Glöckner, Andreas Hellander, Ekta Vats, and Prashant Singh
Venue: Advances in Neural Information Processing Systems, vol. 38, 2025

Paper III addresses the limitations of existing post-hoc uncertainty quantification methods for pre-trained VLMs, which overlook the inherent structural asymmetry between textual and visual modalities and the geometric constraints of embeddings residing on a unit hypersphere. By proposing AsymVLM, we model the aleatoric uncertainty in text via directional distributions on the hypersphere while keeping image embeddings deterministic, respecting the underlying non-Euclidean topology. We demonstrate that AsymVLM serves as a natural probabilistic extension of standard CLIP and SigLIP losses, providing a statistically grounded framework for modeling the inherent ambiguity of embeddings. Experimental results across multiple benchmarks and ablation studies validate the effectiveness of our approach, showing improved cross-modal retrieval performance and more reliable uncertainty estimates compared to existing methods.

Paper IV

Title: Epistemic Uncertainty Quantification for Pre-trained VLMs via Riemannian Flow Matching
Authors: **Li Ju**, Mayank Nautiyal, Andreas Hellander, Ekta Vats, and Prashant Singh
Venue: arXiv:2601.21662 (2026)

Paper IV proposes RepVLM for epistemic uncertainty quantification in pre-trained VLMs, by establishing that embedding density can serve as a theoretically justified proxy for epistemic uncertainty of a pre-trained VLM. Unlike prior works restricted to simple parametric shapes, this approach utilizes flow matching to represent the complex probability distributions in

the high-dimensional embedding space. By implementing this framework via Riemannian flow matching, the method respects the natural geometric constraints of the unit hypersphere, enabling the calculation of exact likelihoods for input queries. This allows a robust measure of epistemic uncertainty that surpasses the limitations of traditional parametric modeling, and provides reliable identification of out-of-distribution anomalies in structurally low-density regions.

Paper V

Title: OneFlowSBI: One Model, Many Queries for Simulation-Based Inference
Authors: Mayank Nautiyal, **Li Ju**, Melker Ernfors, Klara Hagland, Ville Holma, Maximilian Werkö Söderholm, Andreas Hellander, and Prashant Singh
Venue: arXiv preprint arXiv:2601.22951 (2026)

Paper V introduces OneFlowSBI, a unified framework for simulation-based inference designed to overcome the structural rigidity of traditional, task-specific SBI methods. Instead of training separate models for posterior estimation and forward modeling, OneFlowSBI learns a single flow matching model over the joint distribution of parameters and observations. By employing randomized masking on the input during training, the model can support a wide range of inference tasks, including posterior sampling, likelihood estimation, and arbitrary conditional distributions, without requiring task-specific retraining. Evaluated on various benchmarks and high-dimensional real-world inverse problems, the framework demonstrates competitive performance and significant robustness when handling noisy or partially observed data.

7. Conclusion and Outlook

7.1 Conclusion

This thesis has investigated the shift from centralized, mono-modal machine learning toward the complex realities of distributed and heterogeneous machine learning systems. We have demonstrated that the traditional reliance on i.i.d. data is increasingly incompatible with real-world constraints such as data privacy regulations, the physical isolation of edge devices, and the inherent structural differences between modalities like text and images.

Through the five research papers presented, we have addressed robustness across three primary dimensions:

- **Robustness in Distributed Optimization:** We established that while statistical heterogeneity in Federated Learning can destabilize training and induce "client drift," it can be mitigated through adaptive optimization. Specifically, the introduction of AdaFedAdam provided an optimization method that balances global convergence with local fairness, ensuring that the resulting models do not disproportionately favor specific data silos.
- **Geometry-Aware Multi-modal Robustness:** Our work on VLMs highlighted the limitations of applying Euclidean statistical tools to the curved manifolds of hyperspherical embedding spaces. By modeling aleatoric and epistemic uncertainty through directional distributions and Riemannian Flow Matching, we developed mechanisms for models to accurately quantify their own ignorance when confronted with ambiguous or out-of-distribution inputs.
- **Robustness in Simulation-Based Inference:** In the context of Bayesian inference, we moved away from task-specific, rigid architectures. The development of OneFlowSBI demonstrated that a single, unified architecture can robustly handle partially observed or noisy data, providing both forward and inverse modeling capabilities without the need for specialized retraining.

7.2 Outlook

While this thesis provides a foundation for robust learning in modern learning systems, several promising directions for future research remain:

- **Decentralized Foundation Models:** A natural intersection of the robustness in distributed optimization and multi-modal models is the robust training of large-scale multi-modal models with federated learning. Managing the computational overhead while maintaining communication efficiency and geometric robustness remains a significant open challenge.
- **Uncertainty Quantification for Generative VLMs:** While this work focused on UQ for embedding-based vision-language models like CLIP, extending these principles to generative VLMs (e.g., GPT and Qwen family [131, 132]) is a critical frontier. "White-box" approaches that leverage internal mechanisms to quantify hallucination risks should be a built-in feature of robust generative systems. Transitioning from post-hoc UQ to "by-design" uncertainty awareness is essential for moving toward truly self-aware models.
- **AI-Driven Scientific Discovery:** Integrating OneFlowSBI into active learning loops could allow for autonomous experimental design, where the model identifies which physical parameters or observations would most effectively reduce posterior uncertainty.

Ultimately, the goal of this research is to build machine learning systems that are not only high-performing on controlled benchmarks but are also resilient, fair, and self-aware when deployed in the complex, uncurated world.

Author's Contributions

Paper I

The author of this thesis conceptualized the study in collaboration with other authors. The author designed the experiments, implemented the federated learning framework, performed the computations, and wrote the original draft of the manuscript. The analysis and interpretation of the results were performed in close collaboration with other authors.

Paper II

The author of this thesis identified the limitations of existing fair federated learning methods and proposed the DMOO formulation. The author performed the mathematical analysis, designed the algorithm, and conducted numerical experiments. The author wrote the original draft of the manuscript, which was further refined through collaborative iterations with the other authors.

Paper III

The author of this thesis realized the structural and geometric limitations of existing post-hoc uncertainty quantification methods for pre-trained VLMs and, in collaboration with other authors, proposed AsymVLM. The author was responsible for the full implementation of the algorithm and other baseline methods. The experiment design was developed in close collaboration with the co-authors. Finally, the author drafted the original manuscript, which was subsequently refined and revised by the collaborators.

Paper IV

The author of this thesis realized the connection between embedding density and epistemic uncertainty in pre-trained VLMs and recognized that Riemannian flow matching can be utilized for this purpose while respecting natural geometric constraints. The core implementation of the algorithm was performed by the author, with further implementation details and the design of the experiments conducted in collaboration with the co-authors. Finally, the author drafted the original manuscript, which was subsequently refined and revised by the collaborators.

Paper V

The author of this thesis initiated the research by identifying the conceptual link between image inpainting with flow matching and the handling of partially observed data for Bayesian inference. Working with other collaborators, the methodology was further developed for randomized masking within the joint distribution. While not responsible for the core implementation, the author contributed significantly to the code review and the analysis of the experimental results. Finally, the author participated in the revision of the manuscript with the other authors.

Sammanfattning på svenska

Denna avhandling har undersökt skiftet från centraliserad, monomodal maskininlärning till distribuerade och heterogena maskininlärningssystem. Vi har visat att det traditionella antagandet om oberoende och likformigt fördelad data i allt högre grad är oförenligt med verkliga begränsningar såsom dataskyddsregleringar, fysisk isolering av edge-enheter, och de inneboende strukturella skillnaderna mellan modaliteter som exempelvis text och bilder.

Genom de fem forskningsartiklar som presenteras häri har vi adresserat robusthet utifrån tre primära dimensioner:

- **Robusthet i federerad inlärning:** Vi visar att destabiliserad träning och “klientdrift” orsakad av statistisk heterogenitet mellan distribuerade dataset, kan mildras genom adaptiv optimering. Vi utvecklade AdaFedAdam – en optimeringsmetod som balanserar global konvergens med lokal rättvisa mellan klienter. Detta säkerställer att de resulterande modellerna inte gynnar specifika datasilor oproportionerligt mycket.
- **Geometrimedveten multimodal robusthet:** Vårt arbete med Vision-Language Models belyser begränsningarna med att förlita sig på euklidiska statistiska verktyg för de krökta mångfalderna i hypersfäriska inbäddningar. Genom att modellera aleatorisk och epistemisk osäkerhet via riktningsfördelningar och Riemannsk Flow Matching, utvecklade vi mekanismer för modeller att kvantifiera sin egen okunskap när de ställs inför tvetydiga eller out-of-distribution-indata.
- **Tillförlitlighet i simuleringsbaserad inferens:** I kontexten av Bayesiansk inferens gick vi från uppgiftsspecifika, rigida modellarkitekturer till arkitekturer som är robusta för flera olika uppgifter. Utvecklingen av OneFlowSBI visade att en enda, enhetlig arkitektur kan hantera delvis observerade eller brusiga data, vilket möjliggör både framåt- och inversmodellering utan behov av specialiserad omträning av modellen.

Medan denna avhandling utgör en grund för robust inlärning i moderna inlärningssystem, återstår flera lovande riktningar för framtida forskning:

- **Decentraliserade basmodeller:** En naturlig skärningspunkt mellan robusthet i distribuerad optimering och multimodala modeller är den robusta träningen av storskaliga multimodala modeller med federerad inlärning. Att reducera beräkningsoverhead samtidigt som kommunikationseffektivitet och geometrisk robusthet bibehålls är en betydande utmaning.
- **Osäkerhetskvantifiering för generativa VLM:er:** Medan detta arbete fokuserade på osäkerhetskvantifiering för inbäddningsbaserade vision-language-modeller som CLIP, är utvidgningen av dessa principer till

generativa VLM:er (t.ex. Qwen och GPT-familjen) en viktig utvidgning. “White-box”-metoder som utnyttjar interna mekanismer för att kvantifiera hallucinationsrisker bör vara en inbyggd funktion i robusta generativa system. Att gå från post-hoc UQ till “by-design” osäkerhetsmedvetenhet är nödvändigt för att röra oss mot genuint självmedvetna modeller.

- **AI-driven vetenskaplig upptäckt:** Integrering av OneFlowSBI i aktiva inlärningsloopar skulle kunna möjliggöra autonom experimentell design, där modellen själv identifierar vilka fysiska parametrar eller observationer som mest effektivt skulle minska osäkerheten i posteriorfördelningen.

Slutligen är målet med denna forskning att bygga maskininläringssystem som inte bara presterar väl på kontrollerade mätvärden, utan som också är motståndskraftiga, rättvisa och självmedvetna när de driftsätts i den komplexa, okurerade verkligheten.

Statement on the Use of Generative AI

This thesis was originally written by me as the doctoral student. Generative AI has been used for linguistic improvements and proofreading using Gemini 3.1 Pro. This usage follows the guidelines for the use of generative AI in the Faculty of Science and Technology's general study syllabus for doctoral education at Uppsala University.

Acknowledgement

Personal Acknowledgement

I would like to express my sincere gratitude to my supervisor, Andreas Hellander. I am grateful for the intellectual freedom he gave me and his openness in allowing me to explore diverse facets of machine learning. His unique ability to balance academic rigor with an industrial perspective offered me a broader view of the machine learning landscape, and I deeply appreciate his career guidance and his efforts in fostering an environment where I could grow into an independent researcher.

My deepest thanks go to my co-supervisor, Prashant Singh, who has been the cornerstone of this journey. Prashant, thank you for always being "behind me" with unwavering support; this thesis would not have reached its final form without your constant presence and dedicated mentorship. I also thank Salman Toor for his contributions to my growth as a researcher. Our deep discussions on the "philosophy of research" challenged me to look beyond the code and consider the broader implications of scientific inquiry.

I have been fortunate to collaborate with and learn from an exceptional group of researchers. My thanks to senior collaborators Ekta Vats and Ola Spjuth for their invaluable insights and the opportunities we shared. A special place in these acknowledgements is reserved for my close friends and peer collaborators: Tianru Zhang, Mayank Nautiyal, and Shenghui Li. You made the "PhD grind" not just bearable, but truly enjoyable; whether we were dissecting the latest papers or sharing laughs in real life, your friendship has been my reliable support. To my group members at SciML, Aleksandr Karakulev, Zhenlu Sun, Usama Zafar, Csongor Horváth and Dhanushki Mapitigama, thank you for the countless group meetings. I deeply appreciated the pedagogical and low-level details you brought to the weekly talks, which remained a highlight of my time here.

Beyond my immediate research, I am grateful to my colleagues at TDB and IT department. Those conversations during TDB lunches, ranging from random trivia to shifts in research paradigms, were a constant source of inspiration. In particular, I would like to thank my office mates, Erik Blom and Léo Bechet, for the great company and fun discussions about life and everything in between.

Finally, my heart goes out to my family. To my Mom and Dad, thank you for your endless love and support all along. Most importantly, to my wife, Aijia Zhang, thank you for being my partner in everything. Your "all-in-one" support makes my world function.

Infrastructure & Funding Acknowledgement

I would like to thank the Center for Interdisciplinary Mathematics (CIM) at Uppsala University for supporting my PhD funding. I am also grateful to eSSENCE for their support of this research.

The computations were supported by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), with computational resources provided through the Chalmers Centre for Computational Science and Engineering (Alvis HPC).

References

- [1] M. I. Jordan and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects,” *Science*, vol. 349, no. 6245, pp. 255–260, 2015.
- [2] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [3] K. He, X. Zhang, S. Ren, and J. Sun, “Identity mappings in deep residual networks,” in *European conference on computer vision*. Springer, 2016, pp. 630–645.
- [4] R. K. Srivastava, K. Greff, and J. Schmidhuber, “Highway networks,” *arXiv preprint arXiv:1505.00387*, 2015.
- [5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
- [7] K. Hsieh, A. Phanishayee, O. Mutlu, and P. Gibbons, “The non-iid data quagmire of decentralized machine learning,” in *International Conference on Machine Learning*. PMLR, 2020, pp. 4387–4398.
- [8] M. Bakator and D. Radosav, “Deep learning and medical diagnosis: A review of literature,” *Multimodal Technologies and Interaction*, vol. 2, no. 3, p. 47, 2018.
- [9] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, “A survey of deep learning techniques for autonomous driving,” *Journal of field robotics*, vol. 37, no. 3, pp. 362–386, 2020.
- [10] H. Hua, Y. Li, T. Wang, N. Dong, W. Li, and J. Cao, “Edge computing with artificial intelligence: A machine learning perspective,” *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.
- [11] P. Voigt and A. Von dem Bussche, “The eu general data protection regulation (gdpr),” *A practical guide, 1st ed.*, Cham: Springer International Publishing, vol. 10, no. 3152676, pp. 10–5555, 2017.
- [12] R. Bonta, “California consumer privacy act (ccpa),” *Retrieved from State of California Department of Justice: <https://oag.ca.gov/privacy/ccpa>*, pp. 4–40, 2022.
- [13] Y. Huang, C. Du, Z. Xue, X. Chen, H. Zhao, and L. Huang, “What makes multi-modal learning better than single (provably),” *Advances in Neural Information Processing Systems*, vol. 34, pp. 10944–10956, 2021.
- [14] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.
- [15] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in *Artificial intelligence and statistics*. Pmlr, 2017, pp. 1273–1282.

- [16] Q. Li, Y. Diao, Q. Chen, and B. He, “Federated learning on non-iid data silos: An experimental study,” in *2022 IEEE 38th international conference on data engineering (ICDE)*. IEEE, 2022, pp. 965–978.
- [17] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, “Federated learning with non-iid data,” *arXiv preprint arXiv:1806.00582*, 2018.
- [18] M. Mohri, G. Sivek, and A. T. Suresh, “Agnostic federated learning,” in *International conference on machine learning*. PMLR, 2019, pp. 4615–4625.
- [19] T. Li, M. Sanjabi, A. Beirami, and V. Smith, “Fair resource allocation in federated learning,” *arXiv preprint arXiv:1905.10497*, 2019.
- [20] S. Chun, W. Kim, S. Park, and S. Yun, “Probabilistic language-image pre-training,” in *The Thirteenth International Conference on Learning Representations*, 2024.
- [21] Y. Ji, J. Wang, Y. Gong, L. Zhang, Y. Zhu, H. Wang, J. Zhang, T. Sakai, and Y. Yang, “Map: Multimodal uncertainty-aware vision-language pre-training model,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 23 262–23 271.
- [22] X.-Y. Zhang, C.-L. Liu, and C. Y. Suen, “Towards robust pattern recognition: A review,” *Proceedings of the IEEE*, vol. 108, no. 6, pp. 894–922, 2020.
- [23] A. Robey, L. Chamon, G. J. Pappas, and H. Hassani, “Probabilistically robust learning: Balancing average and worst-case performance,” in *International Conference on Machine Learning*. PMLR, 2022, pp. 18 667–18 686.
- [24] X. Ma, J. Zhu, Z. Lin, S. Chen, and Y. Qin, “A state-of-the-art survey on solving non-iid data in federated learning,” *Future Generation Computer Systems*, vol. 135, pp. 244–258, 2022.
- [25] K. Cranmer, J. Brehmer, and G. Louppe, “The frontier of simulation-based inference,” *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, pp. 30 055–30 062, 2020.
- [26] M. Gloeckler, M. Deistler, C. D. Weilbach, F. Wood, and J. H. Macke, “All-in-one simulation-based inference,” in *International Conference on Machine Learning*. PMLR, 2024, pp. 15 735–15 766.
- [27] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, “Federated learning: Challenges, methods, and future directions,” *IEEE signal processing magazine*, vol. 37, no. 3, pp. 50–60, 2020.
- [28] S.-i. Amari, “Backpropagation and stochastic gradient descent method,” *Neurocomputing*, vol. 5, no. 4-5, pp. 185–196, 1993.
- [29] C. Xing, D. Arpit, C. Tsirigotis, and Y. Bengio, “A walk with sgd,” *arXiv preprint arXiv:1802.08770*, 2018.
- [30] R. Abdulkadimov, P. Lyakhov, and N. Nagornov, “Survey of optimization algorithms in modern neural networks,” *Mathematics*, vol. 11, no. 11, p. 2466, 2023.
- [31] J. Duchi, E. Hazan, and Y. Singer, “Adaptive subgradient methods for online learning and stochastic optimization.” *Journal of machine learning research*, vol. 12, no. 7, 2011.
- [32] T. Tieleman and G. Hinton, “Rmsprop: Divide the gradient by a running average of its recent magnitude. coursera,” *Neural Networks for Machine Learning*, p. 31, 2012.
- [33] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv*

- preprint *arXiv:1412.6980*, 2014.
- [34] S. J. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečný, S. Kumar, and H. B. McMahan, “Adaptive federated optimization,” in *International Conference on Learning Representations*, 2021.
 - [35] M. Zaheer, S. Reddi, D. Sachan, S. Kale, and S. Kumar, “Adaptive methods for nonconvex optimization,” *Advances in neural information processing systems*, vol. 31, 2018.
 - [36] J. Zhang, J. Huang, S. Jin, and S. Lu, “Vision-language models for vision tasks: A survey,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 46, no. 8, pp. 5625–5644, 2024.
 - [37] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, “Minigpt-4: Enhancing vision-language understanding with advanced large language models,” in *The Twelfth International Conference on Learning Representations*, 2023.
 - [38] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton, “Similarity of neural network representations revisited,” in *International conference on machine learning*. PMIR, 2019, pp. 3519–3529.
 - [39] M. Huh, B. Cheung, T. Wang, and P. Isola, “Position: The platonic representation hypothesis,” in *Forty-first International Conference on Machine Learning*, 2024.
 - [40] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, “Devise: A deep visual-semantic embedding model,” *Advances in neural information processing systems*, vol. 26, 2013.
 - [41] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, “Zero-shot learning through cross-modal transfer,” *Advances in neural information processing systems*, vol. 26, 2013.
 - [42] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, “Vse++: Improving visual-semantic embeddings with hard negatives,” *arXiv preprint arXiv:1707.05612*, 2017.
 - [43] A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.
 - [44] J. Lu, D. Batra, D. Parikh, and S. Lee, “Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks,” *Advances in neural information processing systems*, vol. 32, 2019.
 - [45] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, “Uniter: Universal image-text representation learning,” in *European conference on computer vision*. Springer, 2020, pp. 104–120.
 - [46] H. Tan and M. Bansal, “Lxmert: Learning cross-modality encoder representations from transformers,” in *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, 2019, pp. 5100–5111.
 - [47] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
 - [48] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, “Sigmoid loss for language

- image pre-training,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 11 975–11 986.
- [49] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint arXiv:1807.03748*, 2018.
- [50] T. Wang and P. Isola, “Understanding contrastive representation learning through alignment and uniformity on the hypersphere,” in *International conference on machine learning*. PMLR, 2020, pp. 9929–9939.
- [51] C. Ley and T. Verdebout, *Modern directional statistics*. Chapman and Hall/CRC, 2017.
- [52] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher *et al.*, “A survey of uncertainty in deep neural networks,” *Artificial intelligence review*, vol. 56, no. Suppl 1, pp. 1513–1589, 2023.
- [53] W. He, Z. Jiang, T. Xiao, Z. Xu, and Y. Li, “A survey on uncertainty quantification methods for deep learning,” *ACM Computing Surveys*, vol. 58, no. 7, pp. 1–35, 2026.
- [54] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya *et al.*, “A review of uncertainty quantification in deep learning: Techniques, applications and challenges,” *Information fusion*, vol. 76, pp. 243–297, 2021.
- [55] T. J. Sullivan, *Introduction to uncertainty quantification*. Springer, 2015, vol. 63.
- [56] A. Der Kiureghian and O. Ditlevsen, “Aleatory or epistemic? does it matter?” *Structural safety*, vol. 31, no. 2, pp. 105–112, 2009.
- [57] D. Rezende and S. Mohamed, “Variational inference with normalizing flows,” in *International conference on machine learning*. PMLR, 2015, pp. 1530–1538.
- [58] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.
- [59] M. Callahan, D. Calvetti, and E. Somersalo, “Beyond the model limit: Parameter inference across scales,” *SIAM/ASA Journal on Uncertainty Quantification*, vol. 5, no. 1, pp. 665–693, 2017.
- [60] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.
- [61] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” *Advances in neural information processing systems*, vol. 27, 2014.
- [62] G. Bredell, K. Flouris, K. Chaitanya, E. Erdil, and E. Konukoglu, “Explicitly minimizing the blur error of variational autoencoders,” in *The Eleventh International Conference on Learning Representations*, 2023.
- [63] G. Loaiza-Ganem and J. P. Cunningham, “The continuous bernoulli: fixing a pervasive error in variational autoencoders,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [64] V. Nagarajan and J. Z. Kolter, “Gradient descent gan optimization is locally stable,” *Advances in neural information processing systems*, vol. 30, 2017.
- [65] Y. Zou, Y. Wang, and X. Lu, “Auto-encoding generative adversarial networks towards mode collapse reduction and feature representation enhancement,” *Entropy*, vol. 25, no. 12, p. 1657, 2023.
- [66] P. Dhariwal and A. Nichol, “Diffusion models beat gans on image synthesis,”

- Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.
- [67] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
- [68] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, “Deep unsupervised learning using nonequilibrium thermodynamics,” in *International conference on machine learning*. PMLR, 2015, pp. 2256–2265.
- [69] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, “Score-based generative modeling through stochastic differential equations,” in *International Conference on Learning Representations*, 2021.
- [70] S. Särkkä and A. Solin, *Applied stochastic differential equations*. Cambridge University Press, 2019, vol. 10.
- [71] B. D. Anderson, “Reverse-time diffusion equation models,” *Stochastic Processes and their Applications*, vol. 12, no. 3, pp. 313–326, 1982.
- [72] Q. Liu, J. Lee, and M. Jordan, “A kernelized stein discrepancy for goodness-of-fit tests,” in *International conference on machine learning*. PMLR, 2016, pp. 276–284.
- [73] A. Hyvärinen and P. Dayan, “Estimation of non-normalized statistical models by score matching,” *Journal of Machine Learning Research*, vol. 6, no. 4, 2005.
- [74] P. Vincent, “A connection between score matching and denoising autoencoders,” *Neural computation*, vol. 23, no. 7, pp. 1661–1674, 2011.
- [75] Y. Song and S. Ermon, “Generative modeling by estimating gradients of the data distribution,” *Advances in neural information processing systems*, vol. 32, 2019.
- [76] M. Albergo, N. M. Boffi, and E. Vanden-Eijnden, “Stochastic interpolants: A unifying framework for flows and diffusions,” *Journal of Machine Learning Research*, vol. 26, no. 209, pp. 1–80, 2025.
- [77] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, “Flow matching for generative modeling,” in *The Eleventh International Conference on Learning Representations*, 2023.
- [78] Y. Lipman, M. Havasi, P. Holderrieth, N. Shaul, M. Le, B. Karrer, R. T. Chen, D. Lopez-Paz, H. Ben-Hamu, and I. Gat, “Flow matching guide and code,” *arXiv preprint arXiv:2412.06264*, 2024.
- [79] X. Liu, C. Gong *et al.*, “Flow straight and fast: Learning to generate and transfer data with rectified flow,” in *The Eleventh International Conference on Learning Representations*, 2023.
- [80] A.-A. Pooladian, H. Ben-Hamu, C. Domingo-Enrich, B. Amos, Y. Lipman, and R. T. Chen, “Multisample flow matching: Straightening flows with minibatch couplings,” in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 100–28 127.
- [81] A. Y. Tong, N. Malkin, K. Fatras, L. Atanackovic, Y. Zhang, G. Hugué, G. Wolf, and Y. Bengio, “Simulation-free schrödinger bridges via score and flow matching,” in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2024, pp. 1279–1287.
- [82] R. T. Chen and Y. Lipman, “Flow matching on general geometries,” in *The Twelfth International Conference on Learning Representations*, 2024.
- [83] J.-M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder, “Approximate bayesian

- computational methods,” *Statistics and computing*, vol. 22, no. 6, pp. 1167–1180, 2012.
- [84] T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. Stumpf, “Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems,” *Journal of the Royal Society Interface*, vol. 6, no. 31, pp. 187–202, 2009.
- [85] G. Papamakarios and I. Murray, “Fast ε -free inference of simulation models with bayesian conditional density estimation,” *Advances in neural information processing systems*, vol. 29, 2016.
- [86] J.-M. Lueckmann, P. J. Goncalves, G. Bassetto, K. Öcal, M. Nonnenmacher, and J. H. Macke, “Flexible statistical inference for mechanistic models of neural dynamics,” *Advances in neural information processing systems*, vol. 30, 2017.
- [87] D. Greenberg, M. Nonnenmacher, and J. Macke, “Automatic posterior transformation for likelihood-free inference,” in *International conference on machine learning*. PMLR, 2019, pp. 2404–2414.
- [88] M. Nautiyal, A. Shternshis, A. Hellander, and P. Singh, “Accelerating simulation-based inference with variational autoencoders,” in *2025 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2025, pp. 1–10.
- [89] M. Nautiyal, A. Hellander, and P. Singh, “Condisim: Conditional diffusion models for simulation based inference,” *arXiv preprint arXiv:2505.08403*, 2025.
- [90] J. Wildberger, M. Dax, S. Buchholz, S. Green, J. H. Macke, and B. Schölkopf, “Flow matching for scalable simulation-based inference,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 16 837–16 864, 2023.
- [91] P. Kairouz and H. B. McMahan, “Advances and open problems in federated learning,” *Foundations and trends in machine learning*, vol. 14, no. 1-2, pp. 1–210, 2021.
- [92] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, “Byzantine-robust distributed learning: Towards optimal statistical rates,” in *International conference on machine learning*. PMLR, 2018, pp. 5650–5659.
- [93] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, “Scaffold: Stochastic controlled averaging for federated learning,” in *International conference on machine learning*. PMLR, 2020, pp. 5132–5143.
- [94] H. Shimodaira, “Improving predictive inference under covariate shift by weighting the log-likelihood function,” *Journal of statistical planning and inference*, vol. 90, no. 2, pp. 227–244, 2000.
- [95] G. Widmer and M. Kubat, “Learning in the presence of concept drift and hidden contexts,” *Machine learning*, vol. 23, no. 1, pp. 69–101, 1996.
- [96] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, “Visualizing the loss landscape of neural nets,” *Advances in neural information processing systems*, vol. 31, 2018.
- [97] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, “Tackling the objective inconsistency problem in heterogeneous federated optimization,” *Advances in neural information processing systems*, vol. 33, pp. 7611–7623, 2020.
- [98] T.-M. H. Hsu, H. Qi, and M. Brown, “Measuring the effects of non-identical data distribution for federated visual classification,” *arXiv preprint arXiv:1909.06335*, 2019.
- [99] R. Pathak and M. J. Wainwright, “Fedsplit: An algorithmic framework for fast

- federated optimization,” *Advances in neural information processing systems*, vol. 33, pp. 7057–7066, 2020.
- [100] W. Du, D. Xu, X. Wu, and H. Tong, “Fairness-aware agnostic federated learning,” in *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*. SIAM, 2021, pp. 181–189.
- [101] J. Jiménez-Luna, F. Grisoni, N. Weskamp, and G. Schneider, “Artificial intelligence in drug discovery: recent advances and future perspectives,” *Expert opinion on drug discovery*, vol. 16, no. 9, pp. 949–959, 2021.
- [102] S. M. Ivanov, A. A. Lagunin, and V. V. Poroikov, “In silico assessment of adverse drug reactions and associated mechanisms,” *Drug Discovery Today*, vol. 21, no. 1, pp. 58–71, 2016.
- [103] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in neural information processing systems*, vol. 25, 2012.
- [104] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [105] M. Ekmefjord, A. Ait-Mlouk, S. Alawadi, M. Åkesson, P. Singh, O. Spjuth, S. Toor, and A. Hellander, “Scalable federated machine learning with fedn,” in *2022 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*. IEEE, 2022, pp. 555–564.
- [106] L. Ju, A. Hellander, and O. Spjuth, “Federated learning for predicting compound mechanism of action based on image-data from cell painting,” *Artificial Intelligence in the Life Sciences*, vol. 5, p. 100098, 2024.
- [107] L. Ju, T. Zhang, S. Toor, and A. Hellander, “Accelerating fair federated learning: Adaptive federated adam,” *IEEE Transactions on Machine Learning in Communications and Networking*, vol. 2, pp. 1017–1032, 2024.
- [108] A. Fang, G. Ilharco, M. Wortsman, Y. Wan, V. Shankar, A. Dave, and L. Schmidt, “Data determines distributional robustness in contrastive language image pre-training (clip),” in *International conference on machine learning*. PMLR, 2022, pp. 6216–6234.
- [109] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković, “Geometric deep learning: Grids, groups, graphs, geodesics, and gauges,” *arXiv preprint arXiv:2104.13478*, 2021.
- [110] U. Upadhyay, S. Karthik, M. Mancini, and Z. Akata, “Problm: Probabilistic adapter for frozen vision-language models,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 1899–1910.
- [111] A. Baumann, R. Li, M. Klasson, S. Mentu, S. Karthik, Z. Akata, A. Solin, and M. Trapp, “Post-hoc probabilistic vision-language models,” *arXiv preprint arXiv:2412.06014*, 2024.
- [112] D. Samuel and G. Chechik, “Distributional robustness loss for long-tail learning,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9495–9504.
- [113] T. Groot and M. Valdenegro-Toro, “Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models,” in *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, 2024, pp. 145–171.
- [114] X. Wang and E. Nalisnick, “Are vision language models robust to uncertain

- inputs?” *arXiv preprint arXiv:2505.11804*, 2025.
- [115] Z. Gao, X. Jiang, X. Xu, F. Shen, Y. Li, and H. T. Shen, “Embracing unimodal aleatoric uncertainty for robust multimodal fusion,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 26 876–26 885.
- [116] M. Kirchhof, E. Kasneci, and S. J. Oh, “Probabilistic contrastive learning recovers the correct aleatoric uncertainty of ambiguous inputs,” in *International Conference on Machine Learning*. PMLR, 2023, pp. 17 085–17 104.
- [117] L. Ju, M. Andersson, S. Fredriksson, E. Glöckner, A. Hellander, E. Vats, and P. Singh, “Exploiting the asymmetric uncertainty structure of pre-trained vlms on the unit hypersphere,” vol. 38, 2025.
- [118] N. De Cao and W. Aziz, “The power spherical distribution,” *arXiv preprint arXiv:2006.04437*, 2020.
- [119] B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” *Advances in neural information processing systems*, vol. 30, 2017.
- [120] Y. Gal and Z. Ghahramani, “Dropout as a bayesian approximation: Representing model uncertainty in deep learning,” in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.
- [121] G. Papamakarios, D. Sterratt, and I. Murray, “Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows,” in *The 22nd international conference on artificial intelligence and statistics*. PMLR, 2019, pp. 837–848.
- [122] D. Greenberg, M. Nonnenmacher, and J. Macke, “Automatic posterior transformation for likelihood-free inference,” in *International conference on machine learning*. PMLR, 2019, pp. 2404–2414.
- [123] D. Ward, P. Cannon, M. Beaumont, M. Fasiolo, and S. Schmon, “Robust neural posterior estimation and statistical model criticism,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 33 845–33 859, 2022.
- [124] M. Deistler, J. Boelts, P. Steinbach, G. Moss, T. Moreau, M. Gloeckler, P. L. Rodrigues, J. Linhart, J. K. Lappalainen, B. K. Miller *et al.*, “Simulation-based inference: A practical guide,” *arXiv preprint arXiv:2508.12939*, 2025.
- [125] J. Boelts, J.-M. Lueckmann, R. Gao, and J. H. Macke, “Flexible and efficient simulation-based inference for models of decision-making,” *Elife*, vol. 11, p. e77220, 2022.
- [126] Y. Verma, A. Bharti, and V. Garg, “Robust simulation-based inference under missing data via neural processes,” in *The Thirteenth International Conference on Learning Representations*, 2025.
- [127] Z. Wang, J. Hasenauer, and Y. Schälte, “Missing data in amortized simulation-based neural posterior estimation,” *PLOS Computational Biology*, vol. 20, no. 6, p. e1012184, 2024.
- [128] M. Gloeckler, M. Deistler, C. D. Weilbach, F. Wood, and J. H. Macke, “All-in-one simulation-based inference,” in *International Conference on Machine Learning*. PMLR, 2024, pp. 15 735–15 766.
- [129] S. T. Radev, M. Schmitt, V. Pratz, U. Picchini, U. Köthe, and P.-C. Bürkner, “Jana: Jointly amortized neural approximation of complex bayesian models,” in *Uncertainty in Artificial Intelligence*. PMLR, 2023, pp. 1695–1706.
- [130] J.-M. Lueckmann, J. Boelts, D. Greenberg, P. Goncalves, and J. Macke,

- “Benchmarking simulation-based inference,” in *International conference on artificial intelligence and statistics*. PMLR, 2021, pp. 343–351.
- [131] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [132] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang *et al.*, “Qwen technical report,” *arXiv preprint arXiv:2309.16609*, 2023.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 2665

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology”.)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-583490



ACTA UNIVERSITATIS
UPSALIENSIS
2026