REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
20-12-2018	Final Report	13-Feb-2017 - 1-Sep-2018

4. TITLE AND SUBTITLE
Final Report: Thermodynamics of Learning

5a. CONTRACT NUMBER
W911NF-17-1-0107

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES

Columbia University
615 West 131st Street
Room 254, MC8725
New York, NY 10027 -7922

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
70187-EG.6

12. DISTRIBUTION AVAILIBILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Henry Hess
UU	UU	UU	UU		19b. TELEPHONE NUMBER 212-854-7749

RPPR Final Report
as of 28-Dec-2018

Agency Code:

Proposal Number: 70187EG **Agreement Number: W911NF-17-1-0107**
INVESTIGATOR(S):

Name: Ph.D Henry Hess
Email: hhess@columbia.edu
Phone Number: 2128547749
Principal: Y

Organization: **Columbia University**
Address: 615 West 131st Street, New York, NY 100277922
Country: USA
DUNS Number: 049179401 EIN: 135598093
Report Date: 01-Dec-2018 Date Received: 20-Dec-2018
Final Report for Period Beginning 13-Feb-2017 and Ending 01-Sep-2018
Title: Thermodynamics of Learning
Begin Performance Period: 13-Feb-2017 **End Performance Period:** 01-Sep-2018
Report Term: 0-Other
Submitted By: Ph.D Henry Hess Email: hhess@columbia.edu
Phone: (212) 854-7749
Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: 1 **STEM Participants:** 1

Major Goals: The recent exponential increase in the applications of machine learning is based on algo- rithms that were already well known in the second half of the 20th century. These recent successes became possible due to the increased availability of computing resources, which allowed for a new level of complexity in the algorithms, as well as the increased availability of large datasets, which allowed these algorithms to be fit in very high dimensional parameter spaces without overfitting. While these methods have been very successful, two fundamen- tal challenges remain. The first challenge lies in evaluating how well an algorithm works a priori, and in providing bounds on the predictions emanating from the algorithm. We aim to present research directions that may address these ideas at the algorithmic level (Task 1), then show how information theory can help address this constraint at the abstract learning level, independently of the algorithm (Task 2). The second challenge is to overcome the energetic constraints that are currently the principal limits on the size of the computational tasks required by the training of these algorithms. We will outline how information ther- modynamics may help the emerging approximate computation paradigms produce energy efficient frameworks for learning (Task 3).
Statistical learning is the acquisition of knowledge about a previously unknown system or concept from data. Classification, regression, clustering, and density estimation are classic statistical learning problems. The most successful machine learning applications all have in common the very high dimensionality of either the data, the parameter space, or both. An adverse consequence of the availability of large amounts of data is the presence of noise. These two structural features of statistical learning, high dimensionality and presence of randomness, are also fundamental in statistical mechanics, suggesting that the tools of statistical mechanics would provide a solid theoretical grounding for learning algorithms. Indeed, deep ties between inference, machine learning, and statistical mechanics have already been investigated.
Simultaneously, recent breakthroughs at the interface of information theory and statisti- cal physics have shed light on the nature of irreversibility far from equilibrium, extended those results to processes featuring information exchange, provided an explicit framework for non-equilibrium process using master equation formalism, and finally extended those concepts to any process that can be described using a Bayesian network. This line of research, often referred to as information thermodynamics, yields tighter bounds than traditional approaches for processes that operate far from equilibrium in environments where thermal fluctuations are relevant.
Statistical mechanics and machine learning share common theoretical ground; recent breakthroughs in non-equilibrium statistical mechanics have successfully investigated the thermodynamics of processes featuring information processing. We propose to investigate the existence of theoretical bridges between these two lines of research through an information theoretic approach. Information theory is at the base of information thermodynamics and can be used as a general framework for learning.

In summary, although the progress of machine learning is undeniable, the field faces several challenges that we

seek to address:

Task 1: We aim to use the tools of statistical physics and information theory to provide bounds on the efficiency of popular learning algorithms in a practical setting, while also gaining insight on the confidence levels of predictions. These theoretical bounds should also enable the machine learning practitioner to determine which action is the most efficient to improve the prediction confidence level, for example by determining the amount of data needed to achieve a certain confidence level.

Task 2: We aim to generalize the results of Task 1 to an abstract learning process that is independent of the chosen learning algorithm. This will require the abstraction of different types of learning as information theoretic processes to exploit the framework of information thermodynamics.

Task 3: We aim to investigate the possibility of relaxing the deterministic constraints in processors in order to harness the randomness of the processes underlying computation, while improving the efficiency of these processors by reducing the dissipation caused by irreversible computations.

Accomplishments: During the reporting period, we completed the Tasks set out above and established a solid framework for the exploration of statistical mechanics concepts as applied to modern Deep Learning algorithms.

In terms of literature review and formulation of more specific sub-problems, we have identified one area as being the most promising: A learning problem is almost always expressible as the estimation of a probability density function relying on several parameters. A significant issue in function estimation is the exponential increase in complexity as the number of parameters grow. The most common way of dealing with this exponential growth is the factorization of the function; a simple example is the family of Naïve Bayes algorithms, where – conditional on the allocated class – the function to estimate is factorized as a product of single parameter functions. However, this is often too simple and fails to capture many features of more complex density functions. In these cases, factorization is often expressed as a directed or undirected graph, where dependencies between parameters are expressed using edges and parameters are nodes of the graph. Then, optimization techniques, such as the Gibbs algorithm or the message-passing algorithms are used on such graphs. The convergence of such algorithms bears strong resemblance to the convergence of thermodynamic process to equilibrium; and we are focused on the application of stochastic thermodynamics to the analysis of such processes.
To facilitate the investigation of Gibbs and message passing algorithm on graphs, many open-source libraries already exist. These libraries are reasonably efficient to solve real world problems, However, we noted in our initial experiments that we needed an order of magnitude more computational time, as we are not trying to solve a single real world problem, but trying to simulate a class of real world problems. We therefore decided to reimplement an easily extendable, simple network graph library in Julia, a language whose features make it as fast as C++, but as easy to read, modify and write as high level scripting languages such as Python or Matlab. We collaborated on this project with the Stanford Intelligent Systems Laboratory, who had an initial version of a software package entitled BayesNets. We extended that package to deal with undirected graphs or Markov Random Fields, in a package called MarkovNets.jl. This software package is already available online at the following address:
https://github.com/henripal/MarkovNets.jl/blob/master/doc/MarkovNets.ipynb

Our main contribution was to reframe Stochastic Gradient Descent in Bayesian Neural Networks as a thermodynamic relaxation from an initial non-equilibrium state. These results are described in the Ph.D. thesis “Application of Modern Statistical Mechanics: Molecular Transport and Statistical Learning”. The thesis focuses on building an understanding of statistical learning as a thermodynamic relaxation process in a high-dimensional space: in the same way that a statistical mechanical system is composed of a large number of particles relaxing to their equilibrium distribution, a statistical learning system is a parametric function whose optimal parameters minimize an empirical loss. We present this process as a trajectory in a high-dimensional probability Riemannian manifold, and show how this conceptual framework can lead to practical improvements in learning algorithms for large scale neural networks.

We then evaluated the accuracy, performance, and practical use cases of current methods for the reframing of state of the art algorithms in a statistical mechanics framework. These methods transform deterministic neural netwoks into Bayesian, or probabilistic neural networks. This review was conducted using concepts rooted in information thermodynamics and concluded that simple constant rate SGD was the best performing Bayesian neural network method. This method has direct applications in “critical areas” for the applications of the algorithm for which it is important for the network to be able to determine when it is not sure of its output. These results are described in the conference paper “Scalable Natural Gradient Langevin Dynamics in Practice”.

RPPR Final Report
as of 28-Dec-2018

Training Opportunities: Henri Palacci completed a thesis entitled "Applications of Modern Statistical Mechanics: Molecular Transport and Statistical Learning" and was awarded a Ph.D. by the Department of Biomedical Engineering at Columbia University.

Results Dissemination: Henri Palacci “Applications of Modern Statistical Mechanics: Molecular Transport and Statistical Learning”, Ph.D. Thesis submitted to Columbia University 2018

H. Palacci, H. Hess: “Scalable Natural Gradient Langevin Dynamics in Practice”, arXiv:1806.02855 (2018)

H. Palacci, H. Hess: “Scalable Natural Gradient Langevin Dynamics in Practice”, accepted by the International Conference on Machine Learning 2018 Workshop “Modern Trends in Nonconvex Optimization for Machine Learning”, Stockholm Sweden (7/14/2018)

H. Palacci: “Physics: A Gateway to Bayesian Deep Learning”
Scientific Computing with Python (SciPy) Conference 2018, Austin Texas, 7/9-15/2018 accessible at https://www.youtube.com/watch?v=WUs0u2PJ2UU

MarkovNets.jl. This software package is publicly available online at the following address: https://github.com/henripal/MarkovNets.jl/blob/master/doc/MarkovNets.ipynb

Honors and Awards: Nothing to Report

Protocol Activity Status:

Technology Transfer: Nothing to Report

PARTICIPANTS:

Participant Type: PD/PI
Participant: Henry S Hess
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Henri Palacci
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

CONFERENCE PAPERS:

RPPR Final Report
as of 28-Dec-2018

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: the International Conference on Machine Learning 2018 Workshop “Modern Trends in Nonconvex Optimization for Machine Learning”
Date Received: 20-Dec-2018 Conference Date: 14-Jul-2018 Date Published: 07-Jun-2018
Conference Location: Stockholm, Sweden
Paper Title: Scalable Natural Gradient Langevin Dynamics in Practice
Authors: Henri Palacci, Henry Hess
Acknowledged Federal Support: **Y**

DISSERTATIONS:

Publication Type: Thesis or Dissertation
Institution: Columbia University
Date Received: Completion Date:
Title: Applications of Modern Statistical Mechanics: Molecular Transport and Statistical Learning
Authors: Henri Palacci
Acknowledged Federal Support: **N**

WEBSITES:

URL: github.com/henripal/MarkovNets.jl/blob/master/doc/MarkovNets.ipynb
Date Received: 07-Aug-2017
Title: MarkovNets.jl
Description: an easily extendable, simple network graph library in Julia

ARO

Final Report for W911NF-17-1-0107

Thermodynamics of Learning

PI: Dr. Henry Hess,

Department of Biological Engineering, Columbia University, New York, NY
hh2374@columbia.edu

Period of Performance: 2/13/17 – 9/1/18

Major Goals

The recent exponential increase in the applications of machine learning is based on algorithms that were already well known in the second half of the 20th century. These recent successes became possible due to the increased availability of computing resources, which allowed for a new level of complexity in the algorithms, as well as the increased availability of large datasets, which allowed these algorithms to be fit in very high dimensional parameter spaces without overfitting. While these methods have been very successful, two fundamental challenges remain. The first challenge lies in evaluating how well an algorithm works a priori, and in providing bounds on the predictions emanating from the algorithm. We aim to present research directions that may address these ideas at the algorithmic level (Task 1), then show how information theory can help address this constraint at the abstract learning level, independently of the algorithm (Task 2). The second challenge is to overcome the energetic constraints that are currently the principal limits on the size of the computational tasks required by the training of these algorithms. We will outline how information thermodynamics may help the emerging approximate computation paradigms produce energy efficient frameworks for learning (Task 3).

Statistical learning is the acquisition of knowledge about a previously unknown system or concept from data. Classification, regression, clustering, and density estimation are classic statistical learning problems. The most successful machine learning applications all have in common the very high dimensionality of either the data, the parameter space, or both. An adverse consequence of the availability of large amounts of data is the presence of noise. These two structural features of statistical learning, high dimensionality and presence of randomness, are also fundamental in statistical mechanics, suggesting that the tools of statistical mechanics would provide a solid theoretical grounding for learning algorithms. Indeed, deep ties between inference, machine learning, and statistical mechanics have already been investigated.

Simultaneously, recent breakthroughs at the interface of information theory and statistical physics have shed light on the nature of irreversibility far from equilibrium, extended those results to processes featuring information exchange, provided an explicit framework for non-equilibrium process using master equation formalism, and finally extended those concepts to any process that can be described using a Bayesian network. This line of research, often referred to as information thermodynamics, yields tighter bounds than traditional approaches for processes that operate far from equilibrium in environments where thermal fluctuations are relevant.

Statistical mechanics and machine learning share common theoretical ground; recent breakthroughs in non-equilibrium statistical mechanics have successfully investigated the

thermodynamics of processes featuring information processing. We propose to investigate the existence of theoretical bridges between these two lines of research through an information theoretic approach. Information theory is at the base of information thermodynamics and can be used as a general framework for learning.

In summary, although the progress of machine learning is undeniable, the field faces several challenges that we seek to address:

Task 1: We aim to use the tools of statistical physics and information theory to provide bounds on the efficiency of popular learning algorithms in a practical setting, while also gaining insight on the confidence levels of predictions. These theoretical bounds should also enable the machine learning practitioner to determine which action is the most efficient to improve the prediction confidence level, for example by determining the amount of data needed to achieve a certain confidence level.

Task 2: We aim to generalize the results of Task 1 to an abstract learning process that is independent of the chosen learning algorithm. This will require the abstraction of different types of learning as information theoretic processes to exploit the framework of information thermodynamics.

Task 3: We aim to investigate the possibility of relaxing the deterministic constraints in processors in order to harness the randomness of the processes underlying computation, while improving the efficiency of these processors by reducing the dissipation caused by irreversible computations.

Accomplished:

During the reporting period, we completed the Tasks set out above and established a solid framework for the exploration of statistical mechanics concepts as applied to modern Deep Learning algorithms.

In terms of literature review and formulation of more specific sub-problems, we have identified one area as being the most promising: A learning problem is almost always expressible as the estimation of a probability density function relying on several parameters. A significant issue in function estimation is the exponential increase in complexity as the number of parameters grow. The most common way of dealing with this exponential growth is the factorization of the function; a simple example is the family of Naïve Bayes algorithms, where – conditional on the allocated class – the function to estimate is factorized as a product of single parameter functions. However, this is often too simple and fails to capture many features of more complex density functions. In these cases, factorization is often expressed as a directed or undirected graph, where dependencies between parameters are expressed using edges and parameters are nodes of the graph. Then, optimization techniques, such as the Gibbs algorithm or the message-passing algorithms are used on such graphs. The convergence of such algorithms bears strong resemblance to the convergence of thermodynamic process to equilibrium; and we are focused on the application of stochastic thermodynamics to the analysis of such processes.

To facilitate the investigation of Gibbs and message passing algorithm on graphs, many open-source libraries already exist. These libraries are reasonably efficient to solve real world problems, However, we noted in our initial experiments that we needed an order of magnitude more computational time, as we are not trying to solve a single real world problem, but trying to simulate a class of real world problems. We therefore decided to reimplement an easily extendable, simple network graph library in Julia, a language whose features make it as fast as C++, but as easy to read, modify and write as high level scripting languages such as Python or Matlab. We collaborated on this project with the Stanford Intelligent Systems Laboratory, who had an initial version of a software package entitled BayesNets. We extended that package to deal with undirected graphs or Markov Random Fields, in a package called MarkovNets.jl. This software package is already available online at the following address:
https://github.com/henripal/MarkovNets.jl/blob/master/doc/MarkovNets.ipynb

Our main contribution was to reframe Stochastic Gradient Descent in Bayesian Neural Networks as a thermodynamic relaxation from an initial non-equilibrium state. These results are described in the Ph.D. thesis "Application of Modern Statistical Mechanics: Molecular Transport and Statistical Learning". The thesis focuses on building an understanding of statistical learning as a thermodynamic relaxation process in a high-dimensional space: in the same way that a statistical mechanical system is composed of a large number of particles relaxing to their equilibrium distribution, a statistical learning system is a parametric function whose optimal parameters minimize an empirical loss. We present this process as a trajectory in a high-dimensional probability Riemannian manifold, and show how this conceptual framework can lead to practical improvements in learning algorithms for large scale neural networks.

We then evaluated the accuracy, performance, and practical use cases of current methods for the reframing of state of the art algorithms in a statistical mechanics framework. These methods transform deterministic neural netwoks into Bayesian, or probabilistic neural networks. This review was conducted using concepts rooted in information thermodynamics and concluded that simple constant rate SGD was the best performing Bayesian neural network method. This method has direct applications in "critical areas" for the applications of the algorithm for which it is important for the network to be able to determine when it is not sure of its output. These results are described in the conference paper "Scalable Natural Gradient Langevin Dynamics in Practice".

Results Dissemination

Henri Palacci "Applications of Modern Statistical Mechanics: Molecular Transport and Statistical Learning", Ph.D. Thesis submitted to Columbia University 2018

H. Palacci, H. Hess: "Scalable Natural Gradient Langevin Dynamics in Practice", arXiv:1806.02855 (2018)

H. Palacci, H. Hess: "Scalable Natural Gradient Langevin Dynamics in Practice", accepted by the International Conference on Machine Learning 2018 Workshop "Modern Trends in Nonconvex Optimization for Machine Learning", Stockholm Sweden (7/14/2018)

H. Palacci: “Physics: A Gateway to Bayesian Deep Learning”
Scientific Computing with Python (SciPy) Conference 2018, Austin Texas, 7/9-15/2018
accessible at https://www.youtube.com/watch?v=WUs0u2PJ2UU

MarkovNets.jl. This software package is publicly available online at the following address: https://github.com/henripal/MarkovNets.jl/blob/master/doc/MarkovNets.ipynb

Honors and Awards

n/a

Technology Transfer (patent applications, inventions, licenses, interaction with DoD laboratories)

n/a

Students
Number of students receiving STEM degrees during the reporting period: 1 Ph.D.
Number of undergraduate and graduate STEM participants during the reporting period : 1

Scalable Natural Gradient Langevin Dynamics in Practice

Henri Palacci [1] **Henry Hess** [1]

Abstract

Stochastic Gradient Langevin Dynamics (SGLD) is a sampling scheme for Bayesian modeling adapted to large datasets and models. SGLD relies on the injection of Gaussian Noise at each step of a Stochastic Gradient Descent (SGD) update. In this scheme, every component in the noise vector is independent and has the same scale, whereas the parameters we seek to estimate exhibit strong variations in scale and significant correlation structures, leading to poor convergence and mixing times. We compare different preconditioning approaches to the normalization of the noise vector and benchmark these approaches on the following criteria: 1) mixing times of the multivariate parameter vector, 2) regularizing effect on small dataset where it is easy to overfit, 3) covariate shift detection and 4) resistance to adversarial examples.

1. Introduction

Deep Learning is moving into fields for which errors are potentially lethal, such as self-driving cars, healthcare, and biomedical imaging. For these applications, being able to estimate errors is essential. Bayesian methods provide a way to expand scalar predictions to full posterior probabilities (Gelman et al., 2014). Stochastic Gradient Langevin Dynamics (SGLD), is one of the solutions to the issue of probabilistic modeling on large datasets. Gaussian noise is added to the SGD updates (Welling & Teh, 2011). It was proposed to pre-condition the Gaussian noise with a diagonal matrix to adapt to the changing curvature of the parameter space (Li et al., 2016a). Using a full preconditioning matrix corresponding to the metric tensor of the parameter space was previously proposed (Girolami Mark & Calderhead Ben, 2011), but the computation of this tensor is impossible for large-scale neural networks. It was further proposed to use the Kronecker-factored block diagonal approximation of this tensor, first introduced in (Martens & Grosse, 2015a) and (Grosse & Martens, 2016) as the preconditioning tensor for the Langevin noise (Nado et al., 2018). Fixed learning rate vanilla gradient descent also introduces noise in the learning process. Hence, fixed learning rate SGD can also be seen as a variant on the same method (Mandt et al., 2017).

In this paper, we conduct a comparison of all these approaches in a practical setting with a fixed hyperparameter optimization budget. We compare these approaches using traditional Markov Chain Monte Carlo (MCMC) diagnostic tools, but will also evaluate the: performance of models in recognizing data points that are not in the sample distribution, the reduction of overfitting in small data settings, and the robustness to adversarial attacks. We find that Langevin approaches, with a reasonable computing budget for hyperparameter tuning, do not improve overfitting or help with adversarial attacks. However, we do find a significant improvement in the detection of out-of-sample data using Langevin methods.

2. Related Work

SGLD was introduced in (Welling & Teh, 2011) and was further refined using a diagonal preconditioning matrix (pSGLD) in (Li et al., 2016a). The natural gradient method was introduced by (Amari, 1998). Girolami and Calderhead proposed to extend the natural gradient method to neural networks in (Girolami Mark & Calderhead Ben, 2011), and a practical application to probability simplices was presented in (Patterson & Teh, 2013). Finally, the interpretation of fixed rate SGD (FSGD) as a Bayesian approximation was shown in (Mandt et al., 2017). The Kronecker-Factored block-diagonal approximation of the inverse Fisher information matrix was presented for dense layers in (Martens & Grosse, 2015b), then extended to convolutional layers in (Grosse & Martens, 2016). This was used as a preconditioning matrix in SGLD (KSGLD) for smaller scale experiments in (Nado et al., 2018).

3. Preliminaries

3.1. Probabilistic Neural Networks

We consider a supervised learning problem, where we have data $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbb{R}^p$, and labels $y_1, ..., y_n$ drawn from a

[1]Department of Biomedical Engineering, Columbia University. Correspondence to: Henri Palacci <hp2393@columbia.edu>.

distribution $\mathcal{P}$. Our goal is to approximate the distribution $p(y|\mathbf{x})$ by empirical risk minimization of a family of distributions parametrized by a vector $\boldsymbol{\theta}$.

In the non-probabilistic setting, this is done by defining an appropriate loss function $\mathcal{L}(y_i|\mathbf{x}_i;\boldsymbol{\theta_i})$ and minimizing it with respect to $\boldsymbol{\theta}$. Optionally, a regularizing term $\mathcal{R}(\boldsymbol{\theta})$ is added to the minimization problem which can therefore be written as: $\hat{\boldsymbol{\theta}} = \text{argmax} \sum_i -\mathcal{L}(y_i, x_i; \boldsymbol{\theta}) + \mathcal{R}(\boldsymbol{\theta})$. This can be understood as the MAP estimate of the probabilistic model $p(\boldsymbol{\theta}|\mathbf{x}) = p(\boldsymbol{\theta}) \prod_i p(y_i, x_i|\boldsymbol{\theta})$, where $p(\boldsymbol{\theta}|\mathbf{x})$ is the posterior probability of the parameters, $\ln p(\boldsymbol{\theta}) = \mathcal{R}(\boldsymbol{\theta})$ is the log-prior, and $\ln p(y_i, \mathbf{x}_i|\boldsymbol{\theta}) = \mathcal{L}(y_i, x_i; \boldsymbol{\theta})$ is the log-likelihood.

3.2. Stochastic Gradient Langevin Dynamics

The workhorse algorithm for loss minimization for neural networks is mini-batch stochastic gradient descent (SGD). The data $\mathbf{x}_1, ...\mathbf{x}_n$ is grouped into mini batches $B_1, ..., B_j, ...$ of size J such that $(\mathbf{x}_1, ...\mathbf{x}_J) \in B_1, (\mathbf{x}_{J+1}, ..., \mathbf{x}_{2J}) \in B_2, ...$

Stochastic Gradient Langevin Dynamics (SGLD) (Welling & Teh, 2011) updates modifies SGD by adding Gaussian noise at each update step: $\Delta\boldsymbol{\theta}_t = \lambda_t \nabla_\theta \left(\log p(\boldsymbol{\theta}) + \sum_j \log p(B_j, \boldsymbol{\theta}) \right) + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \lambda_t \mathbf{I})$.

3.3. Riemaniann Manifold Langevin Dynamics

The space formed by the parameters of a probability distribution is a Riemaniann manifold (Amari, 1998). Its Riemaniann metric is the Fisher information matrix. This means that the parameter space is curved, and that a local measure of curvature is the Fisher information matrix: $F(\theta) = \mathbb{E}\left[\partial_\theta p(y|x;\theta) \partial_\theta p(y|x;\theta)^T\right]$. Riemaniann Manifold Langevin Dynamics (Marceau-Caron & Ollivier, 2017) preconditions the SGD update with the inverse of the Fisher information matrix: $\Delta\boldsymbol{\theta}_t = F^{-1} \lambda_t \nabla_\theta \left(\log p(\boldsymbol{\theta}) + \sum_j \log p(B_j, \boldsymbol{\theta}) \right) + F^{-1}\boldsymbol{\epsilon}$. Unfortunately, the computation of the inverse Fisher information matrix is impossible in very high dimensional spaces.

3.4. Kronecker-Factored Approximate Curvature

The Kronecker-Factored Appoximate Curvature (KFAC) is a compact and efficiently invertible block-diagonal approximation of the Fisher information matrix proposed in (Martens & Grosse, 2015a) for dense layers of neural networks and in (Grosse & Martens, 2016) for convolutional layers. Each block corresponds to a layer of the neural network, hence this approximation correctly takes into account within-layer geometric structure. Each layer i's activations a_i can be computed from the previous layer's activations by a matrix product $s_i = \mathbf{W} a_{i-1}$. A non-linear activation function ϕ such that $a_i = \phi(s_i)$ is applied. The K-FAC approximation can then be written using the Kronecker product $\otimes$: $\widetilde{F} = \text{diag}\left(A_1 \otimes G_1, ..., A_i \otimes G_i, ..., A_l \otimes G_l\right)$, where $A_i = \mathbb{E}\left[a_i a_i^T\right]$ is the estimated covariance matrix of activations for layer i, and $G_i = \mathbb{E}\left[g_i g_i^T\right]$ where $g_i = \nabla_s \mathcal{L}(y, x; \theta)$. We can invert the Kronecker product of two matrices by $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$, and can therefore compute the approximate inverse Fisher information matrix as $\widetilde{F}^{-1} = \text{diag}\left(\{A_i^{-1} \otimes G_i^{-1}\}_{i=1...l}\right)$.

3.5. Scalable Natural Gradient Langevin Dynamics

To implement a tractable preconditioning inverse matrix, (Li et al., 2016a) used a diagonal preconditioning matrix rescaling the noise by the inverse of its estimated variance (pSGLD). Although this improves on SGLD, it still neglects the off-diagonal terms of the metric. A quasi-diagonal approximation was proposed in (Marceau-Caron & Ollivier, 2017). Here, we follow the results presented in (Nado et al., 2018) and use the K-FAC approximation to the inverse Fisher information matrix as our preconditioning matrix:

$$\Delta\boldsymbol{\theta}_t = \widetilde{F}^{-1} \lambda_t \nabla_\theta \left(\log p(\boldsymbol{\theta}) + \sum_j \log p(B_j, \boldsymbol{\theta}) \right) + \widetilde{F}^{-1}\boldsymbol{\epsilon} \tag{1}$$

Notice that when changing preconditioning matrices in practice, it is unclear if any improvement in convergence of the algorithms comes from preconditioning the gradient term above, or from preconditioning the noise. It is one of the questions that we aim to answer with our experiments.

3.6. Fixed Learning Rate Stochastic Gradient Descent

It has been suggested that traditional SGD, using a decreasing schedule for the learning rate and early stopping performs Bayesian updates (Mandt et al., 2017). The noise introduced by the variability in the data also prevents the posterior from collapsing to the MAP.

4. Experiments

In order for the model comparisons to be fair, we used the same neural network architecture for all experiments: two convolutional layers with 32 and 64 layers and max-pooling, followed by one dense layer with 1024 units. All nonlinearities are ReLU. The hyperparameter optimization was run using grid search, and the computational time for hyperparameter optimization was limited to 5 times that of the standard SGD algorithm for all other algorithms. Batch size for all experiments was 512.

Note that we did not apply the preconditioning matrix to

the gradient term. It is otherwise impossible to tell if the performance improvements come from better gradient updates in the initial, non-Langevin part of training or from the improvement of the latter, steady-state part of training. Our SGD updates are therefore:

$$\Delta\boldsymbol{\theta}_t = \lambda_t \nabla_\theta \left(\log p(\boldsymbol{\theta}) + \sum_j \log p(B_j, \boldsymbol{\theta}) \right) + \widetilde{G}\boldsymbol{\epsilon} \quad (2)$$

Where $G = \mathbf{0}$ for SGD, $G = \mathbf{I}$ for SGLD, G is the diagonal RMSprop matrix for pSGD, $G = \widetilde{F}^{-1}$ for KSGD, and $\lambda_t = \lambda$ for fixed learning rate SGD (FSGD).

4.1. Test Set Accuracy

We first compare the test set accuracy for all methods on 10 epochs of training on the MNIST dataset (LeCun et al., 2010). The results are shown in Figure 1; accuracies for all models are very close and, for a reasonable hyperparameter tuning budget, Bayesian averaging of models does not seem to improve test set accuracy.

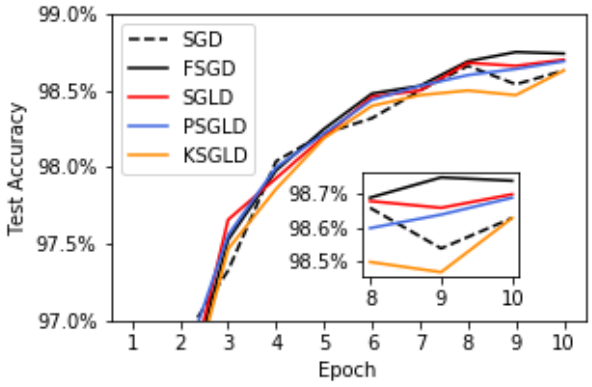

Figure 1. Test set accuracy over ten epochs on the MNIST dataset. SGD: Stochastic Gradient Descent, SGLD: Stochastic Gradient Langevin Dynamics, pSGLD: preconditioned SGLD, KSGLD: K-FAC preconditioned SGLD, FSGD: Fixed rate SGD. Inset: Test set accuracy for the last three epochs.

For the SGLD, pSGLD, and KSGLD methods, the results were very sensitive to the learning rate schedule decrease and most of the hyperparameter optimization computation time was spent on the optimizing it. A longer time spent optimizing the learning rate schedule improved the test rate accuracies slightly.

4.2. Mixing Performance

We approximate (Vats et al., 2015) and estimate the effective sample size as: $\text{mESS} = n \left(\frac{|\Lambda|}{|\Sigma|} \right)^{1/p}$, with n the number of samples in the chain, p the parameter space dimension, $|\Sigma|$ is the covariance matrix of the chain, and $|\Lambda|$ the covariance of matrix of samples. We approximate this by the diagonal approximation of both these matrices, where the ratio of the diagonal terms ess_i is computed as follows $\text{ess}_i = \frac{n}{1+2\sum_k \rho_k}$, where ρ_k is the autocorrelation at lag k truncated to the highest lag with positive autocorrelation (Gelman et al., 2014).

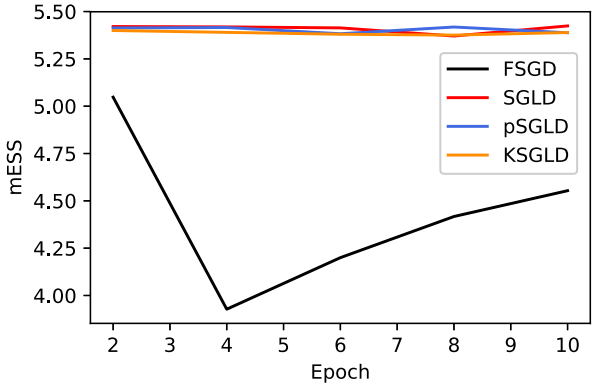

Figure 2. Multivariate Sample Size over epochs for each model over 10 epochs of MNIST training.

The results, shown in Figure 2, all indicate that the MCMC chain mixes poorly in practical settings. Further inspection of the traces shows that almost none of the parameters are stationary. Increasing the run length, or increasing the rate of decrease of the step λ_t, did not improve the aspect of the traces or the effective sample size. These results are consistent with the theoretical analysis of (Betancourt, 2015), who shows that data subsampling is incompatible with any HMC procedure. This is also consistent with (Vollmer et al., 2015) highlighting the problem of stopping while step sizes are still finite.

4.3. Reduction of Overfitting

To test the implicit regularization for the Langevin dynamic models, we truncated the MNIST train set to 5,000 examples (from 60,000). The CNN overfits to the small training set promptly, resulting in decreases in the test set accuracy.

The results, shown in Figure 3, show that the dynamic models underperform SGD on smallMNIST. The only dynamic Bayesian method that matches SGD is SGDA. We hypothesize that adding Gaussian noise on such a small amount of data dramatically deteriorates the initial period of convergence, thus forcing the dynamic Langevin methods to settle for the Langevin period in a local minimum of the loss surface.

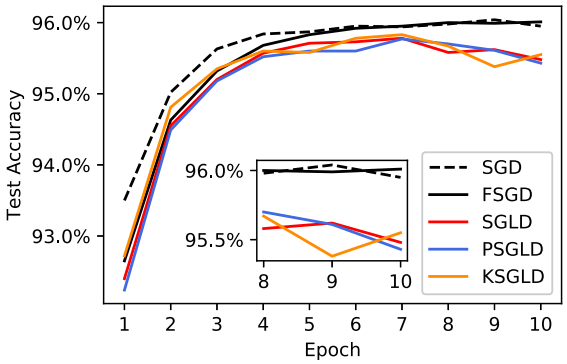

Figure 3. Test set accuracy for all models on ten epochs of training on the reduced MNIST dataset, smallMNIST

4.4. Resistance to Adversarial Attacks

Adversarial attacks are imperceptible modifications to data that cause a model to fail (Goodfellow et al., 2014). We compute adversarial modifications to the test set using the Fast Gradient Sign Method from (Goodfellow et al., 2014). It has previously been shown in (Rawat et al., 2017) that other Bayesian deep learning methods such as Monte Carlo dropout,(Gal & Ghahramani, 2015), Bayes by Backprop (Blundell et al., 2015), matrix variational gaussian (Louizos & Welling, 2016), and probabilistic backpropagation (Hernández-Lobato & Adams, 2015) are vulnerable to adversarial attacks. Our results, presented in Table 1, show that all Langevin dynamic methods also fail to detect adversarial attacks.

Table 1. Classification accuracies for naive Bayes and flexible Bayes on various data sets.

MODEL	TEST ACCURACY	ACCURACY ON ADVERSARIAL EXAMPLES
SGD	96.0	2.9
FSGD	96.5	2.0
SGLD	97.2	1.8
PSGLD	97.1	1.9
KSGLD	97.0	2.0

4.5. Detection of Out of Sample Examples

We assess the epistemic uncertainty inherent in our Bayesian deep neural networks by training it on MNIST but evaluating the network on a completely different dataset, notMNIST (Bulatov). The notMNIST dataset is similar in format to the MNIST dataset, but consists of letters from different fonts.

We expect a network trained on MNIST to give relatively low class probabilities when given examples from the notMNIST dataset. Figure 4 shows the distribution of the highest probability for each example. Vanilla SGD gives very confident predictions for this dataset, whereas all other methods present a similar distribution of uncertainties. This suggests that Langevin dynamics and fixed learning rate SGD are a relatively straightforward way to detect covariate shift in practical classification tasks.

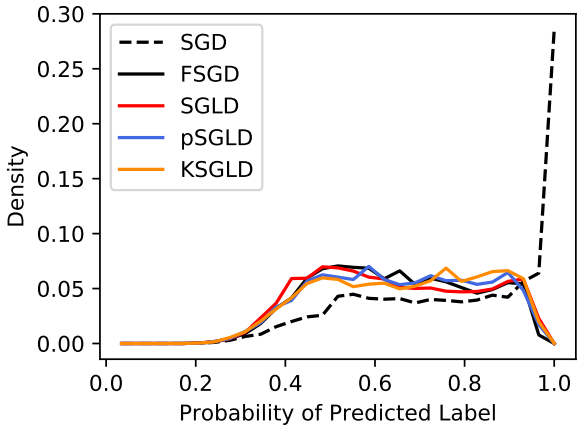

Figure 4. Probability distribution for the most likely class on the notMNIST dataset for all models trained on the MNIST dataset.

5. Discussion

Langevin Stochastic Dynamics provide a scalable way to compute Bayesian posteriors on deep neural network architectures. The noise in stochastic gradient Langevin dynamics is not isotropic due to the geometry of the parameter space. To render the Gaussian noise isotropic, diagonal (Li et al., 2016b), quasi-diagonal (Marceau-Caron & Ollivier, 2017), and block-diagonal (Martens & Grosse, 2015a) approximations have been used. These preconditioning matrices have been proven to work very well as preconditioners for the gradient term, but their use as preconditioners for the Gaussian term in SGLD is subject to significant convergence issues, especially in the transition from the learning phase, where the mini-batch noise dominates.

By contrast, leveraging the mini-batch noise by a constant learning rate to prevent posterior collapse seems to work just as well as the Langevin methods for the experiments described above. This could suggest that the 'data noise' is already appropriately scaled to the manifold structure of the parameter space. This will be the subject of future research.

In practice, our experiments suggest to use Bayesian averaging with a fixed learning rate; this doesn't require any modification to the standard training workflows used by practitioners, and provides implicit protection against covariate shift.

Acknowledgements

This work was supported by the ARO grant "Thermodynamics of Statistical Learning", PI: H. Hess, ARO W911-NF-17-1-0107.

References

Amari, Shun-Ichi. Natural gradient works efficiently in learning. *Neural Comput.*, 10(2):251–276, February 1998.

Betancourt, Michael. The fundamental incompatibility of scalable hamiltonian monte carlo and naive data subsampling. In *International Conference on Machine Learning*, pp. 533–540. jmlr.org, June 2015.

Blundell, Charles, Cornebise, Julien, Kavukcuoglu, Koray, and Wierstra, Daan. Weight uncertainty in neural networks. May 2015.

Bulatov, Yaroslav. notMNIST dataset. http://yaroslavvb.blogspot.com/2011/09/notmnist-dataset.html. Accessed: 2018-4-24.

Gal, Yarin and Ghahramani, Zoubin. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. June 2015.

Gelman, Andrew, Carlin, John B, Stern, Hal S, Dunson, David B, Vehtari, Aki, and Rubin, Donald B. *Bayesian data analysis*, volume 2. CRC press Boca Raton, FL, 2014.

Girolami Mark and Calderhead Ben. Riemann manifold langevin and hamiltonian monte carlo methods. *J. R. Stat. Soc. Series B Stat. Methodol.*, 73(2):123–214, March 2011.

Goodfellow, Ian J, Shlens, Jonathon, and Szegedy, Christian. Explaining and harnessing adversarial examples. December 2014.

Grosse, Roger and Martens, James. A kronecker-factored approximate fisher matrix for convolution layers. *arXiv:1602.01407 [cs, stat]*, February 2016.

Hernández-Lobato, José Miguel and Adams, Ryan P. Probabilistic backpropagation for scalable learning of bayesian neural networks. February 2015.

LeCun, Yann, Cortes, Corinna, and Burges, C J. MNIST handwritten digit database. *AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist*, 2, 2010.

Li, Chunyuan, Chen, Changyou, Carlson, David E, and Carin, Lawrence. Preconditioned stochastic gradient langevin dynamics for deep neural networks. In *AAAI*, volume 2, pp. 4, 2016a.

Li, Chunyuan, Chen, Changyou, Carlson, David E, and Carin, Lawrence. Preconditioned stochastic gradient langevin dynamics for deep neural networks. In *AAAI*, volume 2, pp. 4, 2016b.

Louizos, Christos and Welling, Max. Structured and efficient variational deep learning with matrix gaussian posteriors. March 2016.

Mandt, Stephan, Hoffman, Matthew D, and Blei, David M. Stochastic gradient descent as approximate bayesian inference. *arXiv:1704.04289 [cs, stat]*, April 2017.

Marceau-Caron, Gaétan and Ollivier, Yann. Natural langevin dynamics for neural networks. *arXiv:1712.01076 [cs, stat]*, December 2017.

Martens, James and Grosse, Roger. Optimizing neural networks with kronecker-factored approximate curvature. In *International Conference on Machine Learning*, pp. 2408–2417, 2015a.

Martens, James and Grosse, Roger. Optimizing neural networks with kronecker-factored approximate curvature. March 2015b.

Nado, Zachary, Snoek, Jasper, Grosse, Roger, Duvenaud, David, Xu, Bowen, and Martens, James. STOCHASTIC GRADIENT LANGEVIN DYNAMICS THAT EXPLOIT NEURAL NETWORK STRUCTURE. February 2018.

Patterson, Sam and Teh, Yee Whye. Stochastic gradient riemannian langevin dynamics on the probability simplex. In *Advances in Neural Information Processing Systems*, pp. 3102–3110, 2013.

Rawat, Ambrish, Wistuba, Martin, and Nicolae, Maria-Irina. Adversarial phenomenon in the eyes of bayesian deep learning. November 2017.

Vats, Dootika, Flegal, James M, and Jones, Galin L. Multivariate output analysis for markov chain monte carlo. *arXiv:1512.07713 [math, stat]*, December 2015.

Vollmer, Sebastian J, Zygalakis, Konstantinos C, and Teh, Yee W. (non-) asymptotic properties of stochastic gradient langevin dynamics. January 2015.

Welling, Max and Teh, Yee W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 681–688, 2011.

Applications of Modern Statistical Mechanics: Molecular Transport and Statistical Learning

Henri Palacci

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2018

ABSTRACT

Applications of Modern Statistical Mechanics:

Molecular Transport and Statistical Learning

Henri Palacci

Statistical Mechanics describes the macroscopic behavior of a system through the analysis of its microscopic components. It is therefore a framework to move from a probabilistic, high-dimensional description of a system to its macroscopic description through averaging. This approach, now one of the pillars of physics, has seen successes in other fields, such as statistics or mathematical finance. This broadening of the applications of statistical physics has opened new avenues of research in the field itself. Ideas from information theory, differential geometry, and approximate computation are making their way into modern statistical physics. The research presented in this dissertation straddles this boundary: we start by showing how concepts from statistical physics can be applied to statistical learning, then show how modern statistical physics can provide insights into molecular transport phenomena.

The first three chapters focus on building an understanding of statistical learning as a thermodynamic relaxation process in a high-dimensional space: in the same way that a statistical mechanical system is composed of a large number of particles relaxing to their equilibrium distribution, a statistical learning system is a parametric function whose optimal parameters minimize an empirical loss. We present this process as a trajectory in a high-dimensional probability Riemannian manifold, and show how this conceptual framework can lead to practical improvements in learning algorithms for

large scale neural networks.

The second part of this thesis focuses on two applications of modern statistical mechanics to molecular transport. First, I propose a statistical mechanical interpretation of metabolon formation through cross-diffusion, a generalization of the reaction-diffusion framework to multiple reacting species with non-diagonal terms in the diffusion matrix. These theoretical results are validated by experimental results obtained using a microfluidic system. Second, I demonstrate how fluctuation analysis in motility assays can allow us to infer nanoscale properties from microscale measurements. I accomplish this using computational Langevin dynamics simulations and show how this setup can be used to simplify the testing of theoretical non-equilibrium statistical mechanics hypotheses.

Contents

Acknowledgements

My deepest gratitude goes to my advisor, Professor Henry Hess, for giving me the opportunity to pursue a Ph.D. at Columbia in his group. His decision to accept an atypical candidate like me has had a profound impact on my life and career.

I would like to thank Dr. Samuel Stanton, for helping to create a space for interdisciplinary research aligned with my interests.

I am also very grateful to the current and former members of the Hess Lab, especially to those who held my hand in the first years: Megan Armstrong, Professor Parag Katira, Dr. Amy Lu, and Professor Amit Sitt.

I would also like to thank my Mom, Brigitte. Everything I accomplish is because of all you've done for me.

Finally, thank you to the two most important people in my life: Pierre, for being the best and helping me make the last slide of all my talks, and Annie, for her unwavering support and out of control love. Nothing would make sense if I didn't have you.

Introduction

The methods of statistical mechanics have permeated fields such as statistics, mathematical finance, and approximate computation. In parallel, the discipline itself has seen great breakthroughs, particulary in the understanding of dynamic non-equilibrium processes. In this dissertation, we leverage the tools of modern statistical mechanics to reframe statistical learning as a dynamical process. We then show applications of modern statistical mechanics to directed enzyme motion and fluctuation analysis of microtubule motion in motility assays.

Statistical Physics of Learning

Over the past ten years, the exponential increase in the availability of computational resources and large datasets has allowed industry and academia to leverage machine learning in increasingly productive ways. The parallels between statistical mechanics and machine learning algorithms were already well-established in the nineties [1]. However, the development of machine learning as an applied engineering discipline combined with recent breakthroughs in the understanding of non-equilibrium thermodynamic states through information theory concepts [2] calls for a reexamination of the bridges between the two disciplines. We will demonstrate the analogies between

statistical mechanical systems and statistical learning and propose ways in which the theoretical results from statistical mechanics can allow us to improve the convergence of learning algorithms.

We first present the notation and concepts for both statistical learning and statistical mechanics in Chapter 1. In Chapter 2, we discuss historical approaches of statistical learning inspired from statistical mechanics and sketch some promising but ultimately unsuccessful approaches we attempted. Finally, chapter 3 is an in-depth exploration of Stochastic Gradient Langevin Dynamics (SGLD) and its variants.

Background and Significance

Statistical learning is the acquisition of knowledge about a previously unknown system or concept from data [3]. Classification, regression, clustering, and density estimation are classic statistical learning problems [4]. The most successful machine learning applications have in common the very high dimensionality of either the data, the parameter space, or both. An adverse consequence of the availability of large amounts of data is the presence of noise. These two structural features of statistical learning, high dimensionality and presence of randomness, are also fundamental in statistical mechanics, suggesting that the tools of statistical mechanics would provide a solid theoretical grounding for learning algorithms. Indeed, deep ties between inference, machine learning, and statistical mechanics have already been investigated, see for example [1], [5]–[7] and references therein.

Simultaneously, recent breakthroughs at the interface of information theory and

statistical physics have shed light on the nature of irreversibility far from equilibrium [8], [9] and extended those results to processes featuring information exchange [2]. They have provided an explicit framework for non equilibrium processes using master equation formalism [10], and extended those concepts to any process that can be described using a Bayesian network [11]. This line of research, often referred to as information thermodynamics, yields tighter bounds than traditional approaches for processes that operate far from equilibrium in environments where thermal fluctuations are relevant.

Learning as a Thermodynamic Relaxation Process and Stochastic Gradient Langevin Dynamics

The connections between statistical learning and statistical mechanics can be made at several different levels, which we review in Chapter 2. In this dissertation, we choose to model the learning process as the relaxation from a non-equilibrium initial state, to an equilibrium state. An initial, or prior probability distribution is chosen on the parameters of our model. Then, at each step, this probability distribution is modified to minimize the emrpirical loss (or energy). The final parameter distribution is the one that minimizes the empirical loss, equivalent to the equilibrium distribution in statistical mechanics.

We frame this dynamic learning process as a path on the high dimensional statistical manifold, a Riemannian manifold whose metric is shown to be the Fisher information matrix [12].

Most tractable algorithms for optimization collapse to a single or maximum a posteriori (MAP) solution. To avoid the collapse of the posterior, we need to inject thermal noise. This class of methods, that we will refer to as Stochastic Gradient Langevin Dynamics (SGLD) was initially proposed in [13]. We evaluate this method and variants and discuss potential pitfalls and improvements.

Chemotaxis in Enzyme Cascades

Enzymes that participate in reaction cascades have been shown to assemble into multi-enzyme structures, or metabolons, in response to the presence of the first enzyme's substrate [14]. We will show experimental evidence for directed chemotactic movement of enzymes towards their substrate gradients, and propose a theoretical diffusion model to explain this phenomenon for the purine synthesis cascade. The resulting metabolon, or purinosome, has been experimentally observed previously [14]. A better understanding of its assembly mechanism could potentially allow for new treatments of purine synthesis disorders that target the purinosomes or factors triggering its assembly. I present experimental results from my collaborators showing directional movement of enzymes up their substrate gradient, as well as a theoretical statistical physics model for this directed migration in Chapter 4.

Background and Significance

The interaction between enzymes in living cells is an area of active research. The formation of metabolons in response to the presence of the initial substrate is believed

to facilitate substrate channeling [15]–[18] . Substrate channeling promotes sequential reactions with high yield and high selectivity by directing reaction intermediates along a specific pathway from one enzyme to the next. The diffusive motion of enzymes has been shown to increase as a function of substrate concentration and reaction rate [19]–[22]. Here, we present evidence that suggests that enzymes along the purine synthesis metabolic pathway in which the product of one is the substrate for the next tend to associate through a process of sequential, directed chemotactic movement. Such a process may contribute to the formation of metabolons in living cells co-localized around mitochondria that serve as sources of ATP [23].

We show experimental evidence for the diffusion of an enzyme up the gradient of its substrate, resulting in the formation of regions of higher enzyme concentrations. This phenomenon, called self-diffusiophoresis, or cross-diffusion, has been investigated in both theoretical [24] and experimental [25]–[27] studies.

The mechanisms underlying purinosome formation are likely to also explain metabolon formation, and could provide an understanding of how the cell uses spatial control to regulate enzymes and enzyme complexes and increase metabolic efficiency. A better understanding of these mechanisms might allow for the development of better drug targets for metabolic diseases[28].

Experimental Design

In order to study the movement of enzymes in a cascade in response to a substrate gradient, my collaborators fabricated a microfluidic flow device through photolithog-

raphy. The first and fourth enzymes of the glycolitic cascade, hexokinase (HK) and aldolase (Ald) were fluorescently labeled with distinct amine-reactive and thiol-reactive Dylight dyes. The migration of these enzymes across the channel was measured using confocal microscopy.

Cross diffusion model

We propose that the chemotactic aggregation of enzymes in regions of high substrate concentrations is due to cross-diffusion effects. The substrate-gradient induced aggregation by cross-diffusion counteracts Fickian diffusion of enzymes, which transfers enzymes from regions with high enzyme concentration to those with low enzyme concentration. Cross-diffusion is different from the enhanced diffusion of an enzyme in presence of its substrate, which is also observed for uniform substrate concentrations and accelerates the equilibration of the enzyme concentration by Fickian diffusion. The complete theoretical description of diffusion in a multicomponent system combines the flow of the same species in proportion to its concentration gradient (Fick's law) and the flow of the same species in response to the concentration gradients of other species in solution. The diffusive flow for the concentration c_e of unbound enzyme E in the presence of its substrate S can then be written as:

$$J_e = -D\Delta c_e - D_{XD}\Delta c_s \tag{1}$$

where D is the Fick's law diffusion coefficient, D_{XD} is the "cross-diffusion" coefficient, and Δc_e and Δc_s are gradients in enzyme and substrate concentrations, respectively.

We will show that this model explains the migration of enzymes up their substrate gradient and can promote the formation of metabolons in enzyme cascades.

Recover nanoscale information through microscale measurements using motility assays

Quantifying the behavior of coupled molecular motors is critical to our understanding of systems such as cellular cargo transport, muscle contraction, cell division, and engineered nanodevices. Using the gliding motility assay as model system, we will show through the combination of experimental data and Brownian dynamics simulations that quantitative results about the behavior of the nanoscale motors can be obtained from fluctuation analysis of the microscale microtubule motion. More specifically, we are interested in a factor α quantifying the heterogeneity in motor force production. The theoretical output of the model will be compared to experimental data. These results on the efficiency of collective motor protein action, presented in Chapter 5, will also serve as proof of concept of an experimental methodology to quantify nanoscale dynamics using observed microscale fluctuations.

The recent theoretical breakthroughs in non-equilibrium thermodynamics hold for energies of the order of several kT, which are readily available at the nanoscale, but whose contributions become rapidly undetectable for larger length scales. Nanoscale measurement techniques often require the observed object to be still, or in a specific medium; making the experimental verification of these results more difficult. Building

a system which reflects nanoscale dynamics at the microscale, such as the motility assay, could help with the design of other experiments that yield nanoscale insight from microscale observations.

Chapter 1

Introduction to Statistical Physics and Statistical Learning

Statistical physics aims to characterize the probabilistic laws governing the equilibrium state of a random system with many degrees of freedom. The aim of parametric statistical learning is to learn a rule by minimizing an objective function with respect to a high dimensional parameter vector.

A mechanical system will evolve towards the state minimizing its energy. At fixed temperature, and for that minimal energy, the system's microscopic configuration will evolve into its equilibrium distribution, which will have the maximum entropy for the given energy.

In this chapter, I provide an introductory overview of the bridges between statistical learning and statistical physics, with a special emphasis on simulation methods and geometric approaches.

1.1 Statistical Mechanics

In this section I briefly introduce the notations and terms I will use to present the links between statistical mechanics and statistical learning throughout the next two chapters.

Equilibrium Statistical Mechanics

Equilibrium Thermodynamics I will write the differential first law as:

$$dE = -PdV + TdS - \sum_i \mu_i dNi \tag{1.1}$$

where E is the energy of the system under consideration, P the pressure, V the volume, T the temperature, S the entropy, μ_i and N_i the chemical potential and number of species i.

Choice of an ensemble In thermodynamics, the choice of the ensemble is dictated by the experimental conditions. In statistical learning, the definition of the appropriate ensemble is more problematic. We will suppose the "correct" ensemble to be the canonical ensemble and will return to this point later.

Boltzmann distribution We follow the derivation from [29]:

We suppose the system under consideration to be in the canonical ensemble, thus to be maintainted at a constant temperature T by its contact with a large heat bath R. Let's consider two distinct states s_1 and s_2 for the system. The number of configurations available to the system and reservoir in these two states is equal to the number of configurations $\Omega_R(s_1)$ and $\Omega_R(s_2)$ available to the reservoir only (since we select a specific state for the system).

The probability of the system being in the given state is proportional to the number of microstates of the reservoir for that given state, so we can write:

$$\frac{p^*(s_1)}{p*(s_2)} = \frac{\Omega_R(s_1)}{\Omega_R(s_2)} \tag{1.2}$$

We also have $S(s_i) = -k \ln \Omega_R(s_i)$, so:

$$\frac{p^*(s_1)}{p*(s_2)} = \exp \frac{1}{k} \left(S_R(s_1) - S_R(s_2)\right) \tag{1.3}$$

Finally, since s_1 and s_2 are both equilibrium states, we can integrate Eq. 1.1 at constant temperature and volume between these two states: $\Delta S_R = 1/T \Delta E_R$. We also know that the total energy is conserved so $\Delta E_R = -\Delta E_s$:

$$\frac{p^*(s_1)}{p*(s_2)} = \frac{e^{-E(S_1)/kT}}{e^{-E(S_2)/kT}} \tag{1.4}$$

Since this is valid for any two states, we can now write the Boltzmann probability distribution p^* in the canonical ensemble:

$$p^*(s) = \frac{e^{-E(s)/kT}}{Z} \tag{1.5}$$

where $Z = \sum_s \exp\{-E(s)/kT\}$ is the partition function.

Notice that in the typical case of a system with a well-defined energy function, but a very high number of degrees of freedom, the numerator is easy to compute, but the partition function is often intractable.

Microscopic thermodynamic quantities, and another view of the Boltzmann distribution

In macroscopic thermodynamics, the energy E is a well-defined scalar. Since we are now considering a probabilistic system, we can understand this quantity as an average of energies under the canonical distribution p^*:

$$E = \int_s \epsilon(s) p^*(s) ds \tag{1.6}$$

where $\epsilon(s)$ is the energy of the system in a given microstate s. We will consider this function constant, and linked to the conditions of the experiment, but in a driven system with driving parameter λ, the function ϵ will depend on λ and change over the course of the experiment.

In the microscopic setting, entropy can be written:

$$S(p^*) = -\int_s p^*(s) \ln p^*(s) ds \tag{1.7}$$

Note that in the above we assume the Boltzmann constant k to be unity. We will keep that assumption throughout the rest of this chapter. It can easily be shown that p^* is the probability distribution that maximizes S for the given E. This fact is integral to our understanding of non-equilibrium dynamics of learning.

1.2 Non-equilibrium statistical physics

Over the past twenty years, significant inroads have been made in the theoretical understanding of non-equilibrium processes in statistical physics. The Crooks [9] and Jarzynski [8] fluctuation theorems link equilibrium free-energy differences to the fluctuations during a non-equilibrium process. These theorems rekindled interest in non-equilibrium statistical physics, both experimental [30]–[33] and theoretical [34]–[38].

The key to the theoretical understanding of non-equilibrium states lies in how much more (or less) information the non-equilibrium state has compared to the corresponding equilibrium state. We will make this notion more precise in the following sections.

Shannon Information

In information theory, the term *information* refers to how much we learn about a random variable when we sample from it [39]. For example, sampling from a constant random variable never provides any information. By contrast, observing the outcome of a coin toss provides a certain amount of information. The outcome of this coin toss can be represented using one bit (0 for heads, and 1 for tails).

If we now consider a fair 8-sided die, we now need $\log_2 8 = 3$ bits of information to represent the outcome. If the die was not fair, we could shorten the average description length by using shorter representations for the more likely outcomes.

This concept of information as the average number of bits that we need to repre-

sent the outcome of a random variable is central to non-equilibrium thermodynamics. This average quantity of information, or Shannon information closely resembles Eq. 1.7:

$$H(p) = -\sum_i p(i) \log p(i) \tag{1.8}$$

Non-equilibrium thermodynamic quantities

In this approach, the system has a well-defined Hamiltonian. It also has a well-defined equilibrium probability distribution p^*. We now consider the non-equilibrium case when the distribution of the system is initially out of equilibrium and has a probability distribution $p \neq p^*$.

We define our non equilibrium energy:

$$E(p) = \int p(s)\epsilon(s)ds \tag{1.9}$$

And extend the definition of entropy to this non-equilibrium probability:

$$S(p) = \int p(s) \log p(s)ds \tag{1.10}$$

Relaxation from an initial non-equilibrium state

We will restrict our discussion of non-equilibrium statistical mechanics to the case in which a system is initially not in equilibrium with respect to its Hamiltonian, then

relaxes into its equilibrium state. We will see that this corresponds to the statistical learning process.

We have seen in the previous sections that what characterizes the non-equilibrium state, is its non-equilibrium probability distribution p. In information theory, the Kullback-Leibler Divergence $D_{KL}(p, p^*)$ is the additional number of bits needed to encode the outcome of a random variable following p when the encoding method was optimized for a random variable following p^* [39]:

$$D_{KL}(p, p^*) = \int p(x) \ln \frac{p(x)}{p^*(x)} ds \tag{1.11}$$

It can be shown that, although D_{KL} is not a distance (it is not symmetric), it is always positive (by Jensen's inequality) and is zero only when $p = p^*$.

We now define the relaxation dynamics as the dynamics of a time dependent probability distribution p_t from an initial non-equilibrium state p_0 to the equilibrium state $p_\infty = p^*$. We follow [40] and define a weakly relaxing dynamics as a sequence such that

$$\lim_{t \to \infty} D_{KL}(p, p^*) = 0 \tag{1.12}$$

By contrast, a *strongly* relaxing dynamics is always discarding information with respect to the encoding defined by the equilibrium probability distribution [40]:

$$\frac{\partial D_{KL}(p_t, p^*)}{\partial t} \leq 0 \tag{1.13}$$

It can be shown [40] that any Markovian memoryless dynamics converging to an

equilibrium distribution is strongly relaxing. Therefore Langevin dynamics describe a strongly relaxing process, and we can assume our relaxation dynamics to be strongly relaxing.

Entropy, entropy production, and non-equilibrium second law

The second law states that the entropy of an isolated system can only increase. For our system s in contact with a reservoir, we can write: $\Delta S_{tot} = \Delta S_R + \Delta S_s \geq 0$. Furthermore, the only changes in entropy associated with the reservoir are heat exchanges at constant T, so $\Delta S_R = Q/T$, where Q is the heat from the system to the reservoir. We can therefore write:

$$\Delta S_s = \Delta S_s^{\text{exchange}} + \Delta S_s^{\text{irr}} \tag{1.14}$$

with $\Delta S_s^{\text{exchange}} = -Q/T$. The second law $\Delta S_{tot} \geq 0$ can then be rewritten as:

$$\Delta S_s^{irr} \geq 0 \tag{1.15}$$

It can be shown that under the assumption of strongly relaxing dynamics [40], the irreversible entropy production is also the time derivative of the KL divergence between the non-equilibrium probability distribution and the corresponding equilibrium distribution:

$$\frac{\partial S^{\text{irr}}}{\partial t} = -\frac{\partial D(p_t, p^*)}{\partial t} \leq 0 \tag{1.16}$$

We can see that a strongly relaxing dynamics constantly discards information (information here can be understood as the average description length of the state when the encoding was optimized for the equilibrium distribution).

From this notion of trajectory in phase space where the KL divergence is reduced at every step, we can now naturally formulate the question: What is the *optimal* trajectory in phase space that will lead us to the equilibrium distribution? To answer this we need to develop basic notions of differential geometry.

1.3 Basic Differential Geometry

We have seen in the previous section that the relaxation from a non-equilibrium distribution to an equilibrium distribution can be seen as a trajectory in a space of parametrized probability distributions p_t where the parameters are the positions (and velocities) of all the particles in the system.

In this section, we will place ourselves in the slightly more general framework of a space of parametrized probability distributions p_t parametrized by *any* parameter set $\boldsymbol{\theta}$.

Motivating example

To understand why we need notions from differential geometry to analyze a space of probability distributions, consider the space of Gaussian probability distributions, parametrized by their mean μ and standard deviation σ. Our first intuition would be to consider that space to be like $\mathbb{R}^2$ equipped with the traditional Euclidian distance. With this metric, the distance between $\mathcal{N}(0, 0.1)$ and $\mathcal{N}(1, 0.1)$ is simply $(0-1)^2 + (.1-0.1)^2 = 1$. The distance between $\mathcal{N}(0, 1000)$ and $\mathcal{N}(1, 1000)$ is identical, $(0-1)^2 + (1000-1000)^2 = 1$.

The problem with the Euclidian distance is now apparent. The two low-variance Gaussians are very different from each other. Their mean is over ten standard deviations apart, and it would be extremely rare to mistake a sample from one of the low-variance Gaussians with a sample from the other. By contrast, the high variance Gaussians are very similar, and it would be very hard to distinguish samples from the two distributions.

We've already introduced a better notion of similarity between distributions in equation 1.11: the KL divergence. The KL-divergence, however, is not symmetric, so cannot be used as a distance. This problem can be solved by introducing the *symmetrized* KL divergence, $\widetilde{D}$ [12] :

$$\widetilde{D}_{KL}(p, p^*) = \frac{D_{KL}(p, p^*) + D_{KL}(p^*, p)}{2} \tag{1.17}$$

For our 2-dimensional space of 1-dimensional gaussians, it can be shown that this

is:

$$\widetilde{D}_{KL}(\mathcal{N}(\mu,\sigma),\mathcal{N}(\mu^*,\sigma^*)) = \frac{1}{4\sigma^2\sigma^{*2}}\left[(\mu^*-\mu)^2(\sigma^2+\sigma^{*2})+(\sigma^{*2}-\sigma^2)^2\right] \tag{1.18}$$

and the distance between the two low-variance Gaussian of our motivating example can be computed to be 50, while the distance between the two high-variance Gaussians can be computed to be $0.5.10^{-6}$, which is much more in line with intuition.

We now see that the space of one-dimensional Gaussian distributions is *curved*: the distance between a distributions with two means separated by the same interval depends on the level of the standard deviation. A natural way to define distance in such a space is by using the symmetrized KL divergence 1.17. We will make one step in the direction of formalization of these concepts, to see how paths between distributions can be optimized in such spaces.

Fisher Information

In classical statistics, the Fisher Information is used to compute the Cramer-Rao bound on the variance of any unbiased maximum likelihood estimator $\hat{\theta}$ [41]:

$$\text{Var } \hat{\theta} \geq \frac{1}{\mathcal{I}(\theta)} \tag{1.19}$$

where $\mathcal{I}(\theta)$ is the Fisher Information, usually defined as:

$$\mathcal{I}(\theta) = \mathbb{E}\left[\left(\frac{\partial.}{\partial\theta}\log f(X;\theta)\right)^2\middle|\,\theta\right] \tag{1.20}$$

where f is the likelihood of the data X given parameters θ.

It can be shown that the Fisher Information matrix is also the Hessian of the symmetrized KL divergence. In other words, the Fisher Information is a way to quantify *curvature* in the space of parametrized probability distributions:

$$\mathcal{I}(\theta) = \nabla_\theta^2 \widetilde{D}_{KL}\left(p_\theta, p_{\theta^*}\right)|_{\theta=\theta^*} \tag{1.21}$$

Distance on a Riemannian Manifold

So far, we have seen that the space of probability distributions is a curved space. It can be shown [12] that it is a Riemannian manifold, which can be informally defined as a curved space that locally resembles $\mathbb{R}^n$. For example, a sphere in $\mathbb{R}^3$ is curved, but at each point locally resembles $\mathbb{R}^2$.

On such a space, notions of angles, dot products, and distances are all local: distances in a patch with higher curvature are not the same as distances in patches with lower curvature.

In Euclidean spaces, all notions of angles and distances depend on the dot product of two vectors $\langle \mathbf{u}, \mathbf{v} \rangle$. In a Riemannian manifold, this vector product at a point θ is corrected for the curvature using the metric tensor $\mathbf{F}$: $\langle \mathbf{u}, \mathbf{F}\mathbf{v} \rangle$, and a length is locally defined as $||\mathbf{u}|| = \sqrt{\langle \mathbf{u}, \mathbf{F}\mathbf{u} \rangle}$.

The distance between two points p_θ and p_{θ^*} is therefore a *geodesic*, the minimum curve length of paths between these two points, where the curve length along a path

$\lambda(t)$ such that $\lambda(0) = p_\theta$ and $\lambda(T) = p_{\theta^*}$ is calculated as:

$$l_\lambda(\theta, \theta^*) = \int_0^T ||\lambda'(t)||dt \tag{1.22}$$

where the norm $||.||$ is defined using the Riemannian metric tensor seen above. Therefore the distance is:

$$d(p_\theta, p_{\theta^*}) = \min_\lambda l_\lambda(\theta, \theta^*) \tag{1.23}$$

For our space of probability distributions, the appropriate metric $\mathbf{F}$ is the Fisher information matrix:

$$\mathbf{F}_\theta = \mathcal{I}(\theta) \tag{1.24}$$

It is important to note that even if I have introduced the notion of distance on a parametric probability space using the symmetrized KL divergence, and that the *local* curvature of the space is the second derivative of this distance, the distance between two points on our probability manifold is *not* exactly equal to the symmetrized KL divergence (as it is a result of the minimization program in 1.23).

The Natural Gradient

The Natural Gradient method, first proposed as applied to neural networks in [12], is an efficient method to find minima of a function of parametrized probability distributions. Going back to the example of a non-equilibrium relaxation process, we start with an initial probability distribution p and relax into the the equilibrium probabil-

ity distribution p^* that minimizes the energy of our system. The distance between p and p^* is the minimum path length between the two distributions. Determining the optimal path p_t would be computationally expensive, especially for a large number of particles.

The standard way to compute an optimal path locally would be at each step to use the steepest descent: compute the local gradient, then make a step in that direction:

$$\theta_{t+\Delta t} = \theta_t + \lambda \nabla_\theta E(\theta) \tag{1.25}$$

where λ is a small step size.

However, the equation above only describes the steepest descent method in an Euclidean space. The probability space is curved, so we need to modify the above update to take the curvature into account. What is the direction that will maximize the change in E? We can formalize this as a maximization problem:

$$\max E(\theta + \delta\theta) \text{ u.c. } ||\delta\theta|| < \epsilon \tag{1.26}$$

which can be rewritten:

$$\max E(\theta) + \epsilon \mathbf{v}^T \nabla E(\theta) \text{ u.c. } \langle \mathbf{v}, \mathbf{F}_\theta \mathbf{v} \rangle = 1 \tag{1.27}$$

Solving this equation with the standard method of Lagrange multipliers yields the Natural Gradient method:

$$\theta_{t+\Delta t} = \theta_t + \lambda \mathbf{F}_\theta^{-1} \nabla_\theta E(\theta) \tag{1.28}$$

Note that in the case of a simple Euclidean space, the Riemannian metric tensor is the identity and we recover the classic steepest descent method.

Geodesics on the Probability Manifold and Thermodynamic Length

In [36], Crooks makes the connection between fluctuations and thermodynamic length, and shows that the distance between two equilibrium distributions, which can also be understood as the number of natural fluctuations along the finite-time path between the two distributions is a geodesic on a Riemannian manifold. He also computes the corresponding Riemannian metric and shows that it is equivalent to the Fisher information metric - therefore the natural Riemaniann manifold for thermodynamics is the statistical Riemannian manifold we have described above.

1.4 Bayesian Machine Learning and Connections to Statistical Physics

We will now introduce a subset of machine learning: parametrized supervised learning. This encompasses a vast subset of machine learning, but to fix ideas and show the limitations of computational approaches, we will show examples in deep neural networks.

Learning formalism

We choose this formalism as it can readily be generalized to most forms of supervised learning. In the most general formalism possible, we have a matrix of observed variables $\mathbf{X} \in \mathbf{R}^{n \times p}$, where n is the number of observations and p is the dimension of the observed variables. For example, $\mathbf{X}$ could be a set of n images representing animals, and the $1 \times p$ row vectors would be some encoding of the pixel values of the image. We also have a vector $\mathbf{y} \in \mathbf{R}^n$ containing the labels for each example (for example, a number identifying which animal is in which picture).

The goal of statistical learning is to find a function $f(\mathbf{X}, \boldsymbol{\theta})$ of the inputs $\mathbf{X}$ parametrized by a vector $\boldsymbol{\theta}$ whose output approximates $\mathbf{y}$. Note that we do not specify the dimension of $\boldsymbol{\theta}$, as the infinite dimensional case corresponds to non-parametric estimation.

In probabilistic terms, each example (X, y) is a random variable generated from an unknown probability distribution. Our goal is then to minimize the risk function:

$$R(f, X, y, \boldsymbol{\theta}) = \mathbf{E}\left[L\left(f\left(X, \boldsymbol{\theta}\right), y\right)\right] \tag{1.29}$$

where L is some predefined loss function (for example the L_2 norm of $f(X) - y$) and $\mathbf{E}$ is the expectation with respect to the true joint distribution of (X, y). Unfortunately, we do not have acess to this joint distribution. Therefore, the main focus of statistical learning becomes the minimization with respect to $\boldsymbol{\theta}$ of the empirical risk $\hat{R}$; this is

the ERM, *Empirical Risk Minimization principle* [42]:

$$\hat{R}(f, \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}) = \langle L\left(f\left(\mathbf{X}, \boldsymbol{\theta}\right), \mathbf{y}\right)\rangle \tag{1.30}$$

where the brackets represent the average over all examples. Minimizing the empirical risk, or learning from examples, is typically done using gradient descent (computing the gradient over all examples at the same time, then adjusting the parameters $\boldsymbol{\theta}$ accordingly) or more commonly stochastic gradient descent (computing the gradient and updating the parameters successively for each example).

We note that the ERM minimization principle is not perfect, even for large numbers of examples. If the number of parameters is very large, typically larger than the number of examples, the learning function can simply memorize the outcomes. This would lead to an empirical risk of zero, but does not guarantee good performance on out of sample examples (overfitting) [4].

Neural networks are a special class of functions f representable by layers of "neurons", which are simply a linear transformation of the previous output layer followed by a non-linear "activation function" σ. Common examples of activation functions are the hyperbolic tangent or the sigmoid. For layers $i = 0, 1, ..., I$, we can write:

$$\mathbf{O}_{i+1} = \boldsymbol{\sigma}(\mathbf{W}_i \mathbf{O}_i) \tag{1.31}$$

Where $\mathbf{O}_i$ is the output of layer i, $\mathbf{O}_0 = \mathbf{X}$, and $\mathbf{O}_I$ is the output, to be compared to $\mathbf{y}$. The matrices $\mathbf{W}_i$ are the weights of the neural network and are of dimension

$h_{i-1} \times h_i$, where h_i is the number of neurons on layer i. Note that it is common to add an additive bias to each layer, that we omit here for conciseness.

Stochastic Gradient Descent for Empirical Risk Minimization

We define the loss function $\mathcal{L}(y_i|\mathbf{x}_i; \boldsymbol{\theta_i})$ and minimize it with respect to $\boldsymbol{\theta}$. Optionally, a regularizing term $\mathcal{R}(\boldsymbol{\theta})$ is added to the minimization problem. The minimization problem can therefore be written as:

$$\hat{\boldsymbol{\theta}} = \arg\max \sum_i -\mathcal{L}(y_i, x_i; \boldsymbol{\theta}) + \mathcal{R}(\boldsymbol{\theta}) \tag{1.32}$$

For deep neural networks, this minimization is done using Minibatch Stochastic Gradient Descent.

$$\theta_{t+\Delta t} = \theta_t + \lambda_t \nabla L_{MB(t)} \tag{1.33}$$

where $L_{MB(t)}$ is the loss computed only on a random subset (or minibatch) of the data, and λ_t is a decreasing function such that $\sum_\infty \lambda_t = +\infty$ and $\sum_\infty \lambda_t^2 < +\infty$ [43].

This method has proven to work very well to minimize the loss of deep neural networks. In addition, it scales very well, as it does not require to fit the entire dataset in memory.

Probabilistic or Bayesian Neural Networks

In the section above, we outlined a way to estimate the vector of parameters θ that approximately minimizes the loss $\mathcal{L}$ on our data $(\mathbf{X}, \mathbf{y})$. In most applications, having a point estimate is fine. However, it is often desirable to estimate a probability distribution on our parameter vectore θ, as it would provide us with 1) a means to estimate the error on our predictions 2) an implicit regularization scheme and 3) potentially a way to identify out-of-sample data.

Bayesian Estimation

If we can describe our model for the data as a parametric conditional probability function $p((y, X)|\theta)$, we can use Bayes' theorem to write:

$$p((y, X)|\theta)p(\theta) = p(\theta|(y, X))p((y, X)) \tag{1.34}$$

where $p((y, X)|\theta)$, seen as a function of θ is the likelihood, $p(\theta)$ is the prior probability of the parameters (often chosen to be uninformative), and $p(\theta|(y, X))$ is the posterior probability of the parameters given the data.

A specification of the priors on the parameters along with the probabilistic model allows us to compute the likelihood, and therefore gives us the posterior on θ. The posterior is our target distribution as it gives us the ability to compute errors over the parameters, but also to use the model for prediction by taking expectations of the likelihood over the parameters using the posterior.

Loss as Energy

I will now hypothesize a form for the probabilistic model attached to our parameters as following the Boltzmann distribution as defined in equation 1.5, where the energy is replaced by the loss function L:

$$p(\theta|(X,y)) = \left(\exp[-L(\theta)]\right)/Z \tag{1.35}$$

where $Z = \int \exp[-L(\theta)]d\theta$ normalizes the expression to a probability.

If we now recast both the loss term and regularizing term as energies, with Gibbs probability distributions $p(E) = \left(\exp -E\right)/Z$, the above equation can be understood as the MAP estimate of the following probabilistic model:

$$p(\boldsymbol{\theta}|\mathbf{x}) = p(\boldsymbol{\theta}) \prod_i p(y_i, x_i|\boldsymbol{\theta}) \tag{1.36}$$

where $p(\boldsymbol{\theta}|\mathbf{x})$ is the posterior probability of the parameters, $\log p(\boldsymbol{\theta}) = \mathcal{R}(\boldsymbol{\theta})$ is the log-prior, and $\log p(y_i, \mathbf{x}_i|\boldsymbol{\theta}) = \mathcal{L}(y_i, x_i; \boldsymbol{\theta})$ is the log-likelihood.

In a traditional Bayesian probabilistic approach, the likelihood is computed using the MCMC (Markov Chain Monte-Carlo) method. For deep neural networks, the computation of $p(y_i, \mathbf{x}_i|\boldsymbol{\theta})$ is intractable, as the dimension of the parameter vector $\boldsymbol{\theta}$ is very high (often more than millions), and each MCMC sample would require iterating through the entire datase. We will show in Chapter 3 how Langevin dynamics on a Riemannian manifold can help us compute posteriors for deep neural networks with very large datasets.

Chapter 2

Statistical Physics of Learning

In thermodynamics, a non-equilibrium relaxation process can be described as a transformation from an initial probability distribution p to a final Boltzmann equilibrium probability distribution p^*. A strongly relaxing transformation discards information continuously, thus reducing the distance between p and p^* continuously. In statistical learning, the prior probability on the parameters θ define an initial distribution p on these parameters. Through a sequential optimization scheme, these parameters are changed progressively towards an approximation of the minimum empirical loss distribution p^*.

These two dynamical process are very similar in their dynamics. This analogy between statistical learning and statistical physics is but one of the possible connections between the two fields that have been made over the past twenty years. In this chapter, we will first outline the different historical approaches to the connections between statistical mechanics and statistical learning. We will then make more precise our approach to this connection, and show some unsucessful attempts at leveraging statistical mechanic results to provide theoretical insights or better practical methodologies in statistical learning. Succesful experiments are presented in the next chapter.

2.1 Previous Work

The past twenty years have seen many succesful attempts to use statistical mechanical tools and concepts in machine learning. We outline these connections below.

Monte Carlo Evaluation of High Dimensional Integrals

The partition function $Z = \int \exp\left[-E(s)/T\right] ds$ is the central quantity in equilibrium statistical physics. The state vector s is always high-dimensional, and the evaluation of this integral is non-trivial. In simple cases, saddle point approximations are used [29], but in most practical settings, these integrals are evaluated using Monte Carlo approximations. This evaluation problem is similar to the estimation of posteriors in probabilistic models. Indeed, given a probabilistic model, it is easy to sample from the posterior but difficult to evaluate integrals, since the parameter space is often ten- to a hundred-dimensional [44].

Markov Chain Monte Carlo approximations attempt to efficiently sample from the the state space by prioritizing sampling in high-probability regions. This is done using a Markov "walker" whose next step is proposed at random in the state space, and is more likely to be accepted if the proposed state's energy is lower. This algorithm, the Metropolis-Hastings [45] sampling scheme, can get stuck in local energy minima and sometimes fail to effectively sample the state space. Moreover, its mixing times are typically slow as the walker only steps locally.

Hamiltonian Monte-Carlo, a physics-inspired, significantly more efficient sampling scheme was initially devised by physicists as "Hybrid Monte Carlo" then later redis-

covered and popularized in the mathematical statistics community by Radford Neal [46]. The main idea behind Hamiltonian Monte Carlo (HMC) is to sample from the state space, but also to sample momenta, and to use the momenta to propose the next step, allowing for bigger jumps and a more efficient exploration of state space. A very thorough exploration of HMC can be found in [47].

However, these sampling methods still fail for integrals in millions of dimensions, which would be necessary to compute for modern deep learning models. They also fail if the dataset is large, since a single MCMC step requires a pass through the entire dataset. Alternatives to these MCMC sample schemes in very high dimensions for large datasets will be discussed in the next chapter.

Combinatorial Statistical Physics of Learning

The combinatorial approach to statistical physics of learning [48] attempts to compute generalization curves for machine learning problems given a model, such as the perceptron, and a learning rule, such as the Hebbian rule [49].

The generalization curve is defined as the validation error (error on examples not in the training set) as a function of the size of the training set. In the combinatorial approach, the rule to be learned is parametrized as a "teacher" vector in state space. Initially, the candidate space for admissible "student" vectors spans the entire state space. As more examples from the training set are added, the candidate space is restricted to student vectors that are compatible with these examples. When training is finished, the volume of admissible vectors is postulated to be proportional to the

generalization errors. This problem is treated as a statistical mechanical problem, and the entropy, as the logarithm of the size of phase space, gives a theoretical value for the generalization error.

Spin Glasses and Learning on Networks

This approach linking statistical physics and statistical learning has been the most fruitful and is still actively pursued today; see the textbook [50] or a more recent review [51].

In this approach, connections are drawn by likening learning problems to the description of spin glasses in statistical physics. We informally describe the approach below.

The Ising Model

In statistical physics, the Ising Model describes a symmetrical model of interacting spins on a lattice [29]. The spin at site i, $\sigma_i = \pm 1$ interacts with neighboring spins j according to the Hamiltonian:

$$H(\sigma) = \sum_{\substack{i,j \\ \text{i,j neighbors}}} J\sigma_i\sigma_j + h\sum_i \sigma_i \tag{2.1}$$

where $J \in \mathbb{R}$ is the interaction energy constant and h is an external field. If $J > 0$ the interaction is ferromagnetic, if $J < 0$ the interaction is antiferromagnetic. It can be shown that in two dimensions or more, this system exhibits a phase transition between an ordered phase at low temperature and a disordered phase at higher temperatures.

Spin Glasses

Spin Glasses are a more general case of the Ising model, where the interaction energy constant $J = J_{ij}$ is now a random variable, whose values are fixed at the beginning of the experiment. There are now two levels of randomness: the quenched randomness due to the coefficients J_{ij} and the randomness due to the rapid fluctuations of the spins σ_i.

We are interested in the *general* behavior of the system, that is, we are interested in the values of thermodynamic quantities averaged over both levels of fluctuations [52].

Connection with statistical learning

The connection between Spin Glass models and statistical learning can be formalized following several different approaches. Mezard's book [50] is an in-depth review of these connections. Here, we will outline the connection between probabilistic graphical models and spin glasses.

Probabilistic Graphical Models: Probabilistic Graphical Models are a subset of probabilistic models. The model is represented by a graph specifying the dependency structure between variables. Some standard graphical structures are Bayesian networks, directed graphs whose dependencies are hierarchical, or Conditional Random Fields (CRFs), whose graphs are undirected [53]. For CRFs, the probability of a given state is given by an energy function defined as a function of the values of the nodes of each maximum clique. Probabilistic graphical models are a way to express

a complex probabilistic model of many variables in a *tractable* form. Indeed, specifying the complete joint distribution for all variables in the model is often impossible. Probabilistic graphical models (PGMs) allow for the factorization of the joint, thus greatly reducing the dimensonality of the problem.

Looking back at our description of the Ising model, it is easy to see that the specification of interaction energies for neighboring nodes on the lattice is equivalent to specifying a graphical model on a lattice. The Ising model is homogeneous, but actual probabilistic models will not be. Therefore probabilistic models will have random interaction energies that depend on the specification of the problem.

The two levels of randomness in learning problems (thermal fluctuations of the weights, quenched fluctuations of the examples) also have their counterpart in the two levels of randomness in spin glass systems: Spins fluctuate thermally, while there is an experiment-specific, quenched randomness in the interaction parameters, linked to the distribution of impurities in the system. Methods that have proven successful in spin glass analysis, such as the replica method or cavity method [50], [54], have also given rise to interesting results in optimization and learning settings.

MarkovNets.jl: To investigate further the links between statistical mechanics and statistical learning on conditional random fields, I developed, in collaboration with the Stanford Artificial Intelligence Laboratory, the library `MarkovNets.jl`. This library is written in the high level language Julia, so it is very easy to extend and modify. Moreover, Julia is one of the only high-level scripting scientific languages to perform type-inference and just in time compilation, which allows it to run just as fast as

lower level languages, such as `C++`. The package and its documentation are available at `https://github.com/henripal/markovnets.jl`.

2.2 Learning as a quenched thermodynamic relaxation

As we have seen in the previous chapter, learning and thermodynamic relaxation can be seen as very similar processes. Both start with an initial probability distribution, the prior on the parameters. Then, at each step, both processes converge to the equilibrium, or Boltzmann, distribution, by minimizing the KL divergence between the equilibrium distribution and the current distribution.

We propose to use a fundamental result of information thermodynamics, and to apply it to statistical learning. The fundamental result of information thermodynamics, best explained in [2], is that the relaxation from a non-equilibrium state to an equilibrium state produces work (or negative work). The maximum absolute amount of work produced by this relaxation will be when the Hamiltonian is quenched to match the initial, non-equilibrium work distribution (transforming this distribution to an equilibrium distribution), then deformed quasi-statically to the target equilibrium distribution. We propose to investigate the performance of this procedure as applied to the learning procedure of a neural network.

As we will see, this method does not improve the convergence or optimality of the learning process in our experiments, simply slowing down the convergence process

compared to traditional SGD.

Methodology and Results

To be able to test ou hypotheses, we unfortunately cannot use pre-existing frameworks such as TensorFlow, PyTorch, or Keras, as we need to monitor and access every step of the internals and radically change the way traditional convergence is achieved (we cannot use built-in loss functions). Therefore, we developed our own simplified neural network library, toyNN. It is an open source, modular, neural network library. This package, implemented in Python, is available at `https://github.com/henripal/toynn`.

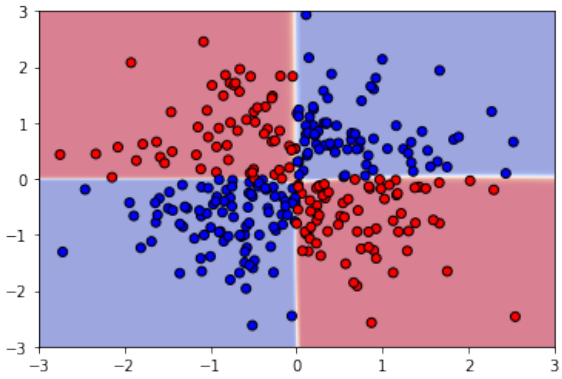

Figure 2.1: The generated examples (dots) follow a random normal bivariate distribution. The classes for each point are generated through the XOR rule. The background colors show how points would be classified by the neural network after training. The neural network has three layers with 20 hidden neurons and was trained for 40,000 epochs using SGD.

We tried out this methodology on a simple XOR classification problem (see 2.1). We used a three layer neural network with 20 units each (multilayer perceptron). To emulate the quenching behavior, we used a weighted average of the current output of the MLP with the real data labels to calculate the quenched loss function. Unfor-

tunately, this procedure did not have the desired effect: it did not allow us to avoid overfitting or local minima - it essentially reproduced the results of the non-quenched procedure, albeit at a linearly slower rate.

In the next chapter, we will see that the thermodynamic view of machine learning as a dynamic process on a Riemannian manifold can however produce positive results.

Chapter 3

Stochastic Langevin Dynamics

Stochastic Gradient Langevin Dynamics (SGLD) is a statistical-physics-inspired sampling scheme for Bayesian modeling adapted to large datasets and models. SGLD relies on the injection of Gaussian Noise at each step of a traditional Stochastic Gradient Descent (SGD) optimization algorithm. In this scheme, every component in the multi-dimensional noise vector is independent and has the same scale, whereas the parameters we seek to estimate exhibit strong variations in scale and significant correlation structures, leading to poor convergence and mixing times. In this chapter, we compare different preconditioning approaches to the normalization of the noise vector and benchmark the viability of SGLD approaches on the following criteria: 1) mixing times of the multivariate parameter vector, 2) regularizing effect on small dataset where it is easy to overfit, 3) covariate shift detection and 4) resistance to adversarial examples.

3.1 Introduction

Deep Learning is moving into fields for which errors are potentially lethal, such as self-driving cars, healthcare, and biomedical imaging. For these particular applications, being able to estimate errors is essential. Bayesian methods provide a way to

expand scalar predictions to full posterior probabilities [44]. Standard techniques for the estimation of intractable integrals in Bayesian statistics can be simulation-based Monte-Carlo [46] or variational [55]. Standard Monte-Carlo techniques break down for larger datasets, as one sample from the posterior distribution requires a pass over the entire dataset. By contrast, Stochastic Gradient Descent (SGD) allows to split the optimization into mini-batches over the dataset [43].

Stochastic Gradient Langevin Dynamics (SGLD), is one of the proposed solutions to the issue of large datasets. In SGLD, Gaussian noise is added to the SGD updates [13]. This method closely resembles Langevin dynamics, originally an approximation of the random dynamics of microscopic particles. Given an appropriate learning rate schedule, and under assumptions of normality, it has been shown that this method converges towards the correct posterior.

To improve convergence and speed up mixing, it was proposed to pre-condition the Gaussian noise with a diagonal matrix to adapt to the changing curvature of the parameter space [56]. This only captures a fraction of the curvature as all multivariate effects are neglected. Using a full preconditioning matrix corresponding to the metric tensor of the underlying parameter space was previously proposed [57], but the computation of the metric tensor is impossible for large-scale neural networks with millions of parameters.

It was further we proposed to use the Kronecker-factored block diagonal approximation of the metric tensor, first introduced in [58] and [59] as the preconditioning tensor for the Langevin noise [60]. In these approaches, the step size or learning rate is reduced according to a Robbins-Monro schedule such that the noise introduced by

the varying batches becomes dominated by the Gaussian noise.

These methods all prevent the posterior from collapsing to the maximum a-priori (MAP) estimate by adding Gaussian noise. However, it was recently argued that fixed learning rate vanilla gradient descent also introduces noise in the learning process. Hence, fixed learning rate SGD can also be seen as a variant on the same method [61].

In this chapter, we conduct a reasoned comparison of all these approaches in a practical setting, with a fixed hyperparameter optimization budget. We compare these approaches using traditional Markov Chain Monte Carlo (MCMC) diagnostic tools, but we will also present the results of experiments for which Bayesian posteriors are desired and evalutate the:

1. Performance of models in recognizing data points that are not in the sample distribution,
2. Reduction of overfitting in small data settings,
3. Robustness to adversarial attacks.

We find that Langevin approaches, with a reasonable computing budget for hyperparameter tuning, do not improve overfitting or help with adversarial attacks. However, we do find a significant improvement in the detection of out-of-sample data using Langevin methods. Within the Langevin methods, we find that the simple fixed learning rate approach [61] performs the best, with no need for modification of existing neural network libraries.

3.2 Related Work

The computation of Bayesian posteriors for Neural Networks was pioneered in [62], using a Monte Carlo approach, by MacKay [63] through a Gaussian approximation of the posterior, and Le Cun [64]. Neal also introduced a Physics-inspired Markov Chain Monte-Carlo (MCMC) method, Hamiltonian MCMC ([46]).

Welling and Teh introduced SGLD in [13] and this method was further refined using a diagonal preconditioning matrix (pSGLD) in [56]. The natural gradient method and relationship between the geometry of the parameter space and the Fisher information matrix was introduced by [65]. Girolami and Calderhead proposed to extend the natural gradient method to Riemannian manifolds in [57], and a practical application to probability simplices was presented in [66]. A quasi-diagonal approximation of the Fisher information matrix applicable to neural networks was proposed in [67]. Finally, the interpretation of fixed rate SGD (FSGD) as a Bayesian approximation was shown in [61].

The Kronecker-Factored block-diagonal approximation of the inverse Fisher information matrix was presented for dense layers in [68], then extended to convolutional layers in [59]. This was used as a preconditioning matrix in SGLD (KSGLD) for smaller scale experiments in [60].

3.3 Preliminaries

Probabilistic Neural Networks

We consider a supervised learning problem, where we have data $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbb{R}^p$, and labels $y_1, ..., y_n$ drawn from a distribution $\mathcal{P}$. Our goal is to approximate the distribution $p(y|\mathbf{x})$ by empirical risk minimization of a family of distributions parametrized by a vector $\boldsymbol{\theta}$.

In the non-probabilistic setting, this is done by defining an appropriate loss function $\mathcal{L}(y_i|\mathbf{x}_i; \boldsymbol{\theta_i})$ and minimizing it with respect to $\boldsymbol{\theta}$. Optionally, a regularizing term $\mathcal{R}(\boldsymbol{\theta})$ is added to the minimization problem which can therefore be written as:

$$\hat{\boldsymbol{\theta}} = \arg\max \sum_i -\mathcal{L}(y_i, x_i; \boldsymbol{\theta}) + \mathcal{R}(\boldsymbol{\theta}) \tag{3.1}$$

If we now recast both the loss term and regularizing term as energies, with Gibbs probability distributions $p(E) = \exp -E/Z$, the above equation can be understood as the MAP estimate of the following probabilistic model:

$$p(\boldsymbol{\theta}|\mathbf{x}) = p(\boldsymbol{\theta}) \prod_i p(y_i, x_i|\boldsymbol{\theta}) \tag{3.2}$$

where $p(\boldsymbol{\theta}|\mathbf{x})$ is the posterior probability of the parameters, $\ln p(\boldsymbol{\theta}) = \mathcal{R}(\boldsymbol{\theta})$ is the log-prior, and $\ln p(y_i, \mathbf{x}_i|\boldsymbol{\theta}) = \mathcal{L}(y_i, x_i; \boldsymbol{\theta})$ is the log-likelihood.

For a traditional Bayesian probabilistic approach, the likelihood is computed using the MCMC method. For deep neural networks, the computation of $p(y_i, \mathbf{x}_i|\boldsymbol{\theta})$ is intractable, as the dimension of the parameter vector $\boldsymbol{\theta}$ is very high (often more than

millions), and each MC sample would require iterating through the entire dataset. We will now tackle some proposed methods to compute the posterior in this high-dimensional large-data setting.

Stochastic Gradient Langevin Dynamics

The workhorse algorithm for loss minimization for neural networks is mini-batch stochastic gradient descent (SGD). The data $\mathbf{x}_1, ...\mathbf{x}_n$ is grouped into mini batches $B_1, ..., B_j, ...$ of size J such that $(\mathbf{x}_1, ...\mathbf{x}_J) \in B_1, (\mathbf{x}_{J+1}, ..., \mathbf{x}_{2J}) \in B_2, ...$

Stochastic Gradient Descent updates are then computed as follows:

$$\Delta\boldsymbol{\theta}_t = \lambda_t \nabla_\theta \left(\mathcal{R}(\boldsymbol{\theta}) + \sum_j \mathcal{L}(B_j, \boldsymbol{\theta}) \right) \tag{3.3}$$

where λ_t is a decreasing learning rate. The gradient is computed by backpropagation [69]. Using the probabilistic formulation from Eq. 3.2, this becomes:

$$\Delta\boldsymbol{\theta}_t = \lambda_t \nabla_\theta \left(\log p(\boldsymbol{\theta}) + \sum_j \log p(B_j, \boldsymbol{\theta}) \right) \tag{3.4}$$

Stochastic Gradient Langevin Dynamics (SGLD) [13] slightly modifies this update by adding Gaussian noise at each update step:

$$\Delta\boldsymbol{\theta}_t = \lambda_t \nabla_\theta \left(\log p(\boldsymbol{\theta}) + \sum_j \log p(B_j, \boldsymbol{\theta}) \right) + \boldsymbol{\epsilon} \tag{3.5}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \lambda_t \mathbf{I})$. It has been shown that if λ_t decreases according to a Robbins-Monro [43] schedule ($\sum_i^\infty \lambda_t = +\infty$ and $\sum_i^\infty \lambda_t^2 < +\infty$), then $\boldsymbol{\theta}$ converges

to its true posterior [70].

Intuitively, randomness in these updates comes from both the Gaussian noise term and the randomness in the gradient term coming from the stochasticity of the mini-batches. However, the variance of the gradient term decreases with λ_t^2, while the variance in the Gaussian term decreases in λ_t. Therefore the Gaussian noise will dominate for large t.

SGLD therefore provides a method to compute a posterior distribution for the parameters, while not requiring computation of gradients over the whole dataset. This method is therefore tractable and applicable to deep neural networks. However, both in physics and machine learning, an important requisite is that the noise be isotropic. There is no guarantee of isotropicity in SGLD, as the parameter vector can have a complex dependence structure (correlations, or order of magnitude variations in individual variances).

Riemaniann Manifold Langevin Dynamics

The space formed by the parameters of a probability distribution is a Riemaniann manifold [65]. Its Riemaniann metric is the Fisher information matrix. This means that the parameter space is curved, and that a local measure of curvature is the Fisher information matrix:

$$F(\theta) = \mathbb{E}\left[\partial_\theta p(y|x;\theta)\partial_\theta p(y|x;\theta)^T\right] \tag{3.6}$$

Amari's natural gradient method proposes to pre-condition gradient updates by

the inverse of the Fisher information matrix. This ensures that the gradients are scaled appropriately. For example, if one of the parameters was changing with a larger order of magnitude than the others, the curvature of the manifold would be higher in that direction and the updates should be scaled down for this parameter to maintain the locality of updates. Riemaniann Manifold Langevin Dynamics [67] follows this principle and preconditions the SGD update with the inverse of the Fisher information matrix:

$$\Delta\boldsymbol{\theta}_t = F^{-1}\lambda_t \nabla_\theta \left(\log p(\boldsymbol{\theta}) + \sum_j \log p(B_j, \boldsymbol{\theta}) \right) + F^{-1}\boldsymbol{\epsilon} \tag{3.7}$$

Unfortunately, the computation of the inverse Fisher information matrix is impossible in very high dimensional spaces. In our example neural network, it would require the storage and evaluation of a three million by three millon matrix, which is not feasible.

Kronecker-Factored Approximate Curvature

The Kronecker-Factored Appoximate Curvature (KFAC) is a compact and efficiently invertible block-diagonal approximation of the Fisher information matrix proposed in [58] for dense layers of neural networks and in [59] for convolutional layers. Each block corresponds to a layer of the neural network, hence this approximation correctly takes into account within-layer geometric structure. Each layer i's activations a_i can be computed from the previous layer's activations by a matrix product $s_i = \mathbf{W}a_{i-1}$. A non-linear activation function ϕ such that $a_i = \phi(s_i)$ is applied. The K-FAC

approximation can then be written using the kronecker product $\otimes$:

$$\widetilde{F} = \text{diag}\left(A_1 \otimes G_1, ..., A_i \otimes G_i, ..., A_l \otimes G_l\right) \tag{3.8}$$

where $A_i = \mathbb{E}\left[a_i a_i^T\right]$ is the estimated covariance matrix of activations for layer i, and $G_i = \mathbb{E}\left[g_i g_i^T\right]$ where $g_i = \nabla_s \mathcal{L}(y, x; \theta)$. We can invert the Kronecker product of two matrices by $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$, and can therefore compute the approximate inverse Fisher information matrix as:

$$\widetilde{F}^{-1} = \text{diag}\left(\{A_i^{-1} \otimes G_i^{-1}\}_{i=1...l}\right) \tag{3.9}$$

Scalable Natural Gradient Langevin Dynamics

To implement a tractable preconditioning inverse matrix in equation 3.7, [56] used a diagonal preconditioning matrix rescaling the noise by the inverse of its estimated variance (pSGLD). Although this improves on SGLD, it still neglects the off-diagonal terms of the metric. A quasi-diagonal approximation was proposed in [67]. Here, we follow the results presented in [60] and use the K-FAC approximation to the inverse Fisher information matrix as our preconditioning matrix in equation 3.7:

$$\Delta\boldsymbol{\theta}_t = \widetilde{F}^{-1}\lambda_t \nabla_\theta \left(\log p(\boldsymbol{\theta}) + \sum_j \log p(B_j, \boldsymbol{\theta})\right) + \widetilde{F}^{-1}\boldsymbol{\epsilon} \tag{3.10}$$

Notice that when changing preconditioning matrices in practice, it is unclear if any improvement in convergence of the algorithms comes from preconditioning the gradient term above, or from preconditioning the noise. It is one of the questions

that we aim to answer with our experiments.

Fixed Learning Rate Stochastic Gradient Descent

We have described in 3.3 how the Robbins-Monro learning rate schedule ensures that the Gaussian noise dominates the mini-batch noise in the gradient updates for large t. Recently, it has been suggested that traditional SGD, using a decreasing schedule for the learning rate and early stopping, also indirectly performs Bayesian updates [61]. Indeed, the noise introduced by the variability in the data also prevents the posterior from collapsing to the MAP. This indirect Bayesian updating would be responsible for the effective regularization of modern neural networks through simple SGD.

3.4 Experiments

Our objective is to evaluate the performance of these scalable dynamic methods for Bayesian updating in deep neural networks. Our criteria for evaluation are closely aligned with the well-known benefits of Bayesian statistics:

1. Mixing times. We evaluate each method's mixing time using a multivariate estimated sample size criteria (mESS) following [71].
2. Implicit Regularization. We evaluate each method's resistance to overfitting using a subset of MNIST, smallMNIST.
3. Resistance to adversarial attacks. Bayesian averaging of parameters could, in theory, provide resistance to adversarial attacks [72]. We evaluate each method's resistance to adversarial attacks.

4. Ability of model to detect changes in data distribution. If new examples do not match the training distribution, it would be desirable for the model to output lower prediction probabilities. We test this using the notMNIST dataset, which has the same format as MNIST but with letters instead of numbers [73].

In order for the model comparisons to be fair, we used the same neural network architecture for all experiments: two convolutional layers with 32 and 64 layers and max-pooling, followed by one dense layer with 1024 units. All nonlinearities are ReLU. The hyperparameter optimization was run using grid search, and the computational time for hyperparameter optimization was limited to 5 times that of the standard SGD algorithm for all other algorithms. Batch size for all experiments was 512.

Note that we did not apply the preconditioning matrix to the gradient term in Equation 3.7. It is otherwise impossible to tell if the performance improvements come from better gradient updates in the initial, non-Langevin part of training or from the improvement of the latter, steady-state part of training. Our SGD updates are therefore:

$$\Delta\boldsymbol{\theta}_t = \lambda_t \nabla_\theta \left(\log p(\boldsymbol{\theta}) + \sum_j \log p(B_j, \boldsymbol{\theta}) \right) + \widetilde{G}\boldsymbol{\epsilon} \tag{3.11}$$

Where $G = \mathbf{0}$ for SGD, $G = \mathbf{I}$ for SGLD, G is the diagonal RMSprop matrix for pSGD, $G = \widetilde{F}^{-1}$ for KSGD, and $\lambda_t = \lambda$ for fixed learning rate SGD (FSGD).

Test Set Accuracy

We first compare the test set accuracy for all methods on 10 epochs of training on the MNIST dataset [74]. The results are shown in Figure 3.1; accuracies for all models are very close and, for a reasonable hyperparameter tuning budget, Bayesian averaging of models does not seem to improve test set accuracy.

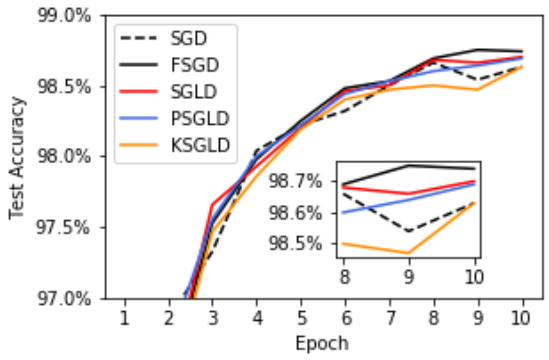

Figure 3.1: Test set accuracy over ten epochs on the MNIST dataset. SGD: Stochastic Gradient Descent, SGLD: Stochastic Gradient Langevin Dynamics, pSGLD: preconditioned SGLD, KSGLD: K-FAC preconditioned SGLD, FSGD: Fixed rate SGD. Inset: Test set accuracy for the last three epochs.

For the SGLD, pSGLD, and KSGLD methods, the results were very sensitive to the learning rate schedule decrease, and most of the hyperparameter optimization computation time was spent on optimizing it. A longer time spent optimizing the learning rate schedule improved the test rate accuracies slightly.

Mixing Performance

The very high dimension of the parameter space and impossibility of computing over the entire dataset at once are specific to Bayesian Neural Networks. Therefore,

the measures of Monte-Carlo convergence usually used for probabilistic models, such as [75], which compares within and between chains, are not applicable. Here, we approximate [76] and estimate the effective sample size as:

$$\text{mESS} = n \left(\frac{|\Lambda|}{|\Sigma|} \right)^{1/p} \tag{3.12}$$

with n the number of samples in the chain, p the parameter space dimension, $|\Sigma|$ the covariance matrix of the chain, and $|\Lambda|$ the covariance of matrix of samples. We approximate this by the diagonal approximation of both these matrices, where the ratio of the diagonal terms ess_i is computed as follows:

$$\text{ess}_i = \frac{n}{1 + 2\sum_k \rho_k} \tag{3.13}$$

where ρ_k is the autocorrelation at lag k truncated to the highest lag with positive autocorrelation [44].

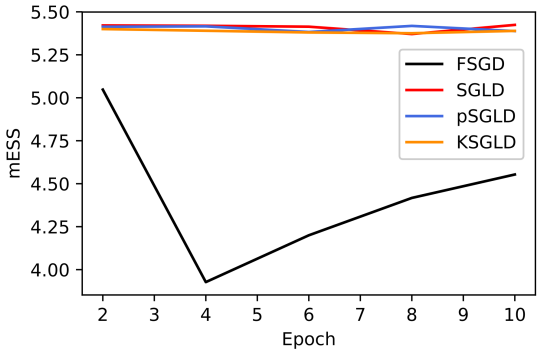

Figure 3.2: Multivariate Sample Size over epochs for each model over 10 epochs of MNIST training.

The results, shown in Figure 3.2, all indicate that the MCMC chain mixes poorly

in practical settings. Further inspection of the traces shows that almost none of the parameters are stationary. Increasing the run length, or increasing the rate of decrease of the step λ_t, did not improve the aspect of the traces or the effective sample size. These results are consistent with the theoretical analysis of [77], who shows that data subsampling is incompatible with any HMC procedure. This is also consistent with [78] highlighting the problem of stopping while step sizes are still finite.

Overfitting Reduction

Bayesian models implicitly regularize the objective, which helps models avoid overfitting. This is especially true for smaller datasets. In order to test this for the Langevin dynamic models, we truncated the MNIST train set to 5,000 examples (from 60,000). The CNN overfits to the small training set promptly, resulting in decreases in the test set accuracy.

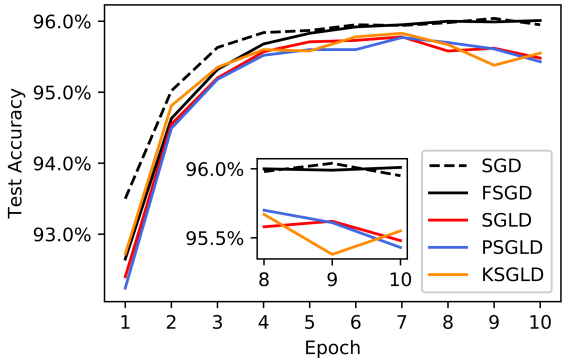

Figure 3.3: Test set accuracy for all models on ten epochs of training on the reduced MNIST dataset, smallMNIST

The results, shown in Figure 3.3, show that the dynamic models dramatically un-

derperform SGD on smallMNIST. The only dynamic Bayesian method that matches SGD is SGDA. We hypothesize that adding Gaussian noise on such a small amount of data dramatically deteriorates the initial period of convergence, thus forcing the dynamic Langevin methods to settle for the Langevin, or stationary period in a local minimum of the loss surface.

Resistance to Adversarial Attacks

Adversarial attacks are imperceptible modifications to data that cause a model to fail [72]. After training our CNN, we compute adversarial modifications to the test set using the Fast Gradient Sign Method from [72]. It has previously been shown in [79] that other Bayesian deep learning methods such as Monte Carlo dropout [80], Bayes by Backprop [81], matrix variational Gaussian [82], and probabilistic backpropagation [83] are vulnerable to adversarial attacks. Our results, presented in Table 3.1 show that all Langevin dynamic methods also fail to detect adversarial attacks.

Table 3.1: Classification accuracies for naive Bayes and flexible Bayes on various data sets.

Model	Test Accuracy	Accuracy on Adversarial Examples
SGD	96.0	2.9
FSGD	96.5	2.0
SGLD	97.2	1.8
pSGLD	97.1	1.9
KSGLD	97.0	2.0

Detection of Out of Sample Examples

A key advantage of Bayesian probabilistic modeling over simple MAP estimation is the better representation of uncertainty. The distinction between *epistemic* uncertainty, the uncertainty in our model's parameters and *aleatoric* uncertainty, the uncertainty linked to noise, and their relationship to Bayesian uncertainty was presented in [84].

Here, we assess the epistemic uncertainty inherent in our Bayesian deep neural networks by training it on MNIST but evaluating the network on a completely different dataset, notMNIST [73]. The notMNIST dataset is similar in format to the MNIST dataset, but consists of letters from different fonts (see Figure 3.4.)

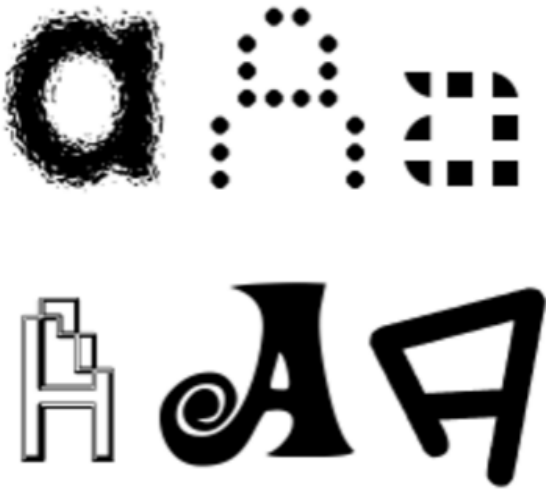

Figure 3.4: Some example images from the notMNIST dataset.

We expect a network trained on MNIST to give relatively low class probabilities when given examples from the notMNIST dataset. Figure 3.5 shows the distribu-

tion of the highest probability for each example. Vanilla SGD gives very confident predictions for this dataset, whereas all other methods present a similar distribution of uncertainties. This suggests that Langevin dynamics and fixed learning rate SGD are relatively straightforward ways to detect covariate shift in practical classification tasks.

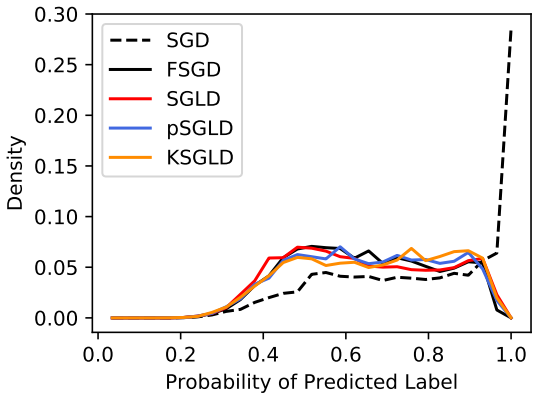

Figure 3.5: Distribution of probabilities for the most likely class on the notMNIST dataset for all models trained on the MNIST dataset.

3.5 Discussion

Langevin Stochastic Dynamics provide a scalable way to compute Bayesian posteriors on deep neural network architectures. Langevin dynamics in physics describe the random motion of a particle in a thermal bath. The particle is subject to *isotropic* forces, typically from the impacts of the water molecules. By contrast, the noise in stochastic gradient Langevin dynamics is not isotropic due to the geometry of the parameter space. The geometric nature of the phase space in thermodynamics and

the common expression of the local curvature matrix between the space of probability distributions and the phase space has been described in [36].

To render the Gaussian noise isotropic, diagonal [56], quasi-diagonal [67], and block-diagonal [58] approximations have been used. These preconditioning matrices have been proven to work very well as preconditioners for the gradient term, but their use as preconditioners for the Gaussian term in SGLD is subject to significant convergence issues, especially in the transition from the learning phase, where the mini-batch noise dominates.

By contrast, leveraging the mini-batch noise by a constant learning rate to prevent posterior collapse seems to work just as well as the Langevin methods for the experiments described above. This suggests that the 'data noise' is already appropriately scaled to the manifold structure of the parameter space.

In practice, our experiments suggest to use Bayesian averaging with a fixed learning rate; this doesn't require any modification to the standard training workflows used by practitioners, and provides implicit protection against covariate shift.

Chapter 4

Chemotaxis in Enzyme Cascades

Catalysis is essential to cell survival. In many instances, enzymes that participate in reaction cascades have been shown to assemble into metabolons in response to the presence of the substrate for the first enzyme. However, what triggers metabolon formation has remained an open question. Through a combination of theory and experiments, we show that enzymes in a cascade can assemble via chemotaxis. Each enzyme independently follows its own specific substrate gradient, which in turn is produced as a product of the preceding reaction.

This chapter follows closely *Xi Zhao, Henri Palacci, Vinita Yadav, Michelle M. Spiering, Michael K. Gilson, Peter J. Butler, Henry Hess, Stephen J. Benkovic, and Ayusman Sen. 2018. "Substrate-Driven Chemotactic Assembly in an Enzyme Cascade." Nature Chemistry 10 (3): 311.*, with the addition of section 4.4 describing the background theory and modeling.

4.1 Introduction

The interaction between enzymes in living cells is an area of active research. In many instances, enzymes that participate in reaction cascades have been shown to assemble into metabolons in response to the presence of the initial substrate to facilitate

substrate channeling [85]–[87]. Substrate channeling promotes sequential reactions with high yield and high selectivity by directing reaction intermediates along a specific pathway from one enzyme to the next. Inspired by these biological cascade reactions, multicatalyst nanostructures have been fabricated for efficient chemical synthesis [88]–[90].

There are several suggested mechanisms for biological metabolon formation and substrate channeling. Some involve stable protein/protein interactions [17] [91], while others invoke electrostatic guidance [92] [15] or spatial organization and clustering [93] [18]. In other cases metabolon formation through reversible and/or post-translational modifications has been suggested, but such transient complexes have eluded isolation [94], [95]. Recently, the diffusive motion of enzymes has been shown to increase as a function of substrate concentration and reaction rate; furthermore, active enzymes migrate up the substrate gradient, an example of molecular chemotaxis [96]–[98]. Here we present evidence that suggests that enzymes along a metabolic pathway in which the product of one is the substrate for the next tend to associate through a process of sequential, directed chemotactic movement. Such a process may contribute to the formation of metabolons in living cells co-localized around mitochondria that serve as sources of ATP [23].

Our experimental study applies microfluidic and fluorescence spectroscopy techniques to study the coordinated movement of hexokinase (HK) and aldolase (Ald), the first and fourth enzymes of the glycolysis cascade, which are connected by the intermediate enzymes phosphoglucose isomerase (Iso) and phosphofructokinase (PFK) (Figure 4.1A.). Metabolon formation by glycolytic enzymes has been suggested in

the literature [99]. In order to monitor the movement of HK and Ald by confocal microscopy, we fluorescently labeled them with distinct amine-reactive (ex/em: 493/518) and thiol-reactive (ex/em: 638/658) Dylight dyes, respectively. The use of different dyes enables simultaneous measurement of both enzymes in microfluidic experiments. For both HK and Ald, a linear relationship is known to be observed between fluorescence intensity and concentration. This allows us to estimate the amount of enzyme that migrated into a specific substrate channel.

4.2 Catalysis-Induced Enhanced Diffusion of Hexokinase and Aldolase.

Before examining the effect of an imposed substrate gradient on the movement of HK and Ald, we measured their diffusion coefficients in uniform substrate concentrations by fluorescence correlation spectroscopy (FCS). The diffusion constants of both enzymes rise with increasing substrate concentration, saturating at increases of 38% for HK and 35% for Ald (Figure 4.2). As previously reported [21], [100], the substrate-induced increase in the diffusion constant is proportional to the catalytic velocity computed from known Michaelis-Menten parameters.

4.3 Individual Enzyme Chemotaxis

To examine the chemotactic movement of enzymes in response to a substrate gradient, a three inlet and one outlet microfluidic flow device was fabricated through

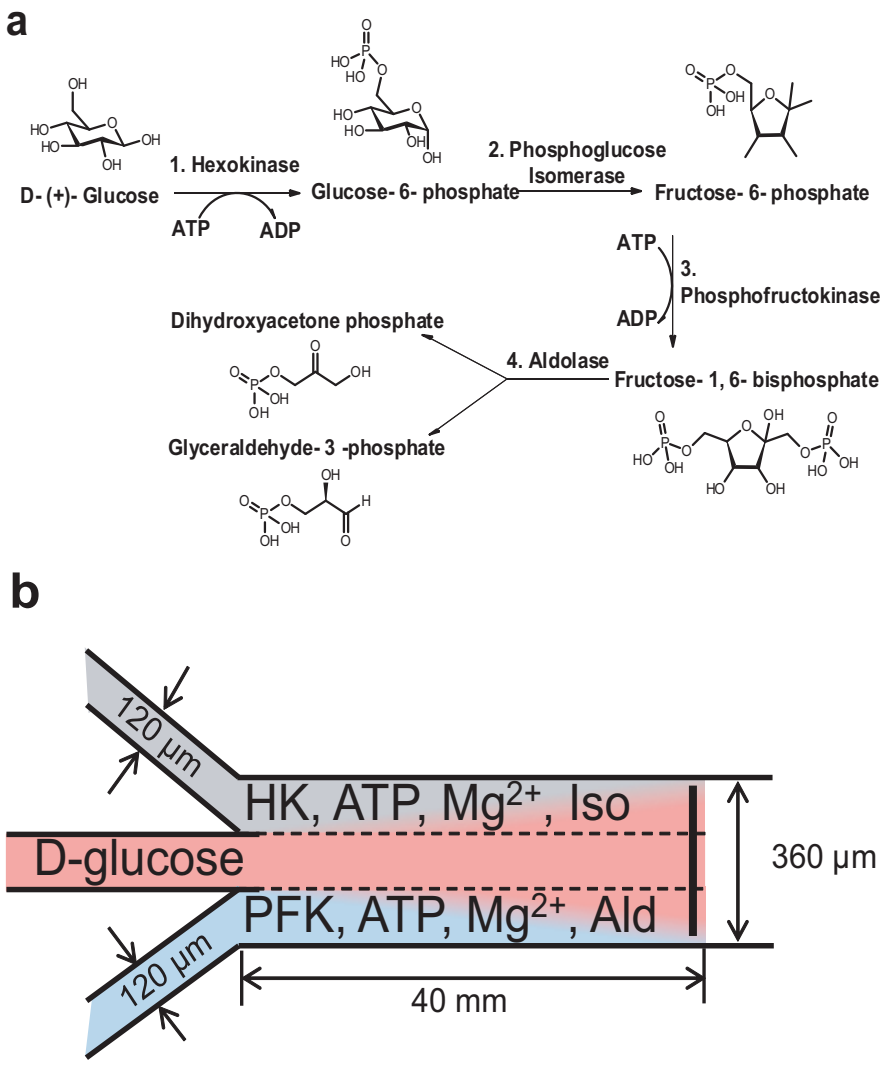

Figure 4.1: a, First four steps of glycolysis and their associated enzymes: hexokinase (HK), phosphoglucose isomerase (Iso), phosphofructokinase (PFK) and aldolase (Ald). b, Photolithographically fabricated flow-based microfluidic channel (channel length, 40 mm; width, 360 µm; depth, 100 µm). Due to laminar flow, the effective width of each flow channel is 120 µm. Fluorescence intensities were analysed across the combined channel where indicated by the vertical black line, except for 20 µm next to the sidewalls.

photolithography (Figure 4.1b.). With known fluid flow rates and channel geometries, the distance from the input points to the measurement line can be converted into the time available for the enzymes to react and diffuse. As a control experiment, HK (200 nM), D-glucose (50 mM) and MgCl2 (100 mM) were passed through all three

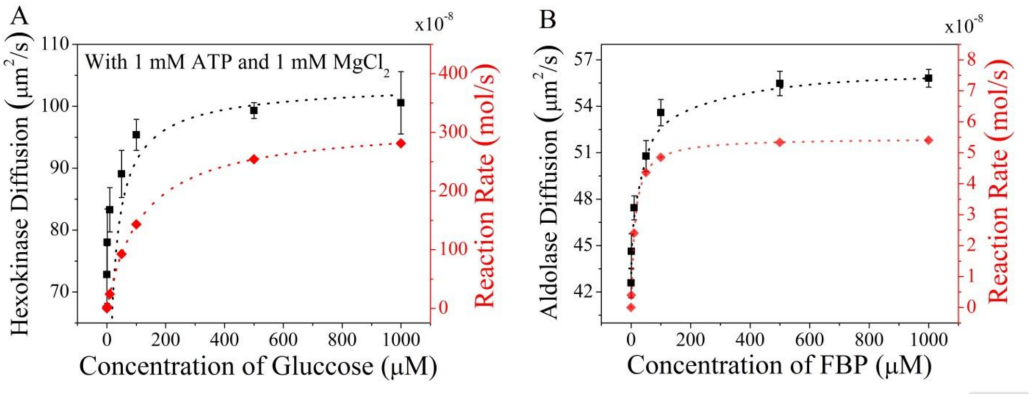

Figure 4.2: Fluorescence correlation spectroscopy (FCS) results showing an enhanced diffusion coefficient for HK (A) and Ald (B) in the presence of their respective substrates, D-glucose and fructose 1,6-bisphosphate. In each case, the enzyme diffusivity increases with increasing reaction rate. The error bars represent standard deviations calculated for 15 different measurements under identical conditions.

channels. Then, the solution in the central channel was changed to HK (200 nM), D-glucose (50 mM), MgCl2 (100 mM), and ATP (50 mM). 150 mM NaCl was added into the two flanking channels to balance the ionic strength of the ATP disodium salt added to the center channel. As shown in Figure 4.3, we observed significant enzyme focusing in the central channel following an interaction time of 34.6 s in the microchannel, compared to when ATP was absent.

The total fluorescence intensity in all the experiments was normalized to 1 for comparison and representation on a common scale. We repeated the experiment substituting mannose, a substrate which binds more strongly but is turned over more slowly by HK, and found that the enzyme focused less than in the presence of D-glucose. We also repeated the experiment substituting L-glucose, the enantiomer of D-glucose that is not a substrate, and observed no focusing. Similarly, the substitution of ATP by its analog, adenosine 5'-(β,γ-methylene) triphosphate (AMP-PCP) at the same concentration, in the central channel resulted in no focusing. Note that

both ATP and AMP-PCP bind to HK but that the latter cannot turnover and phosphorylate glucose[101].

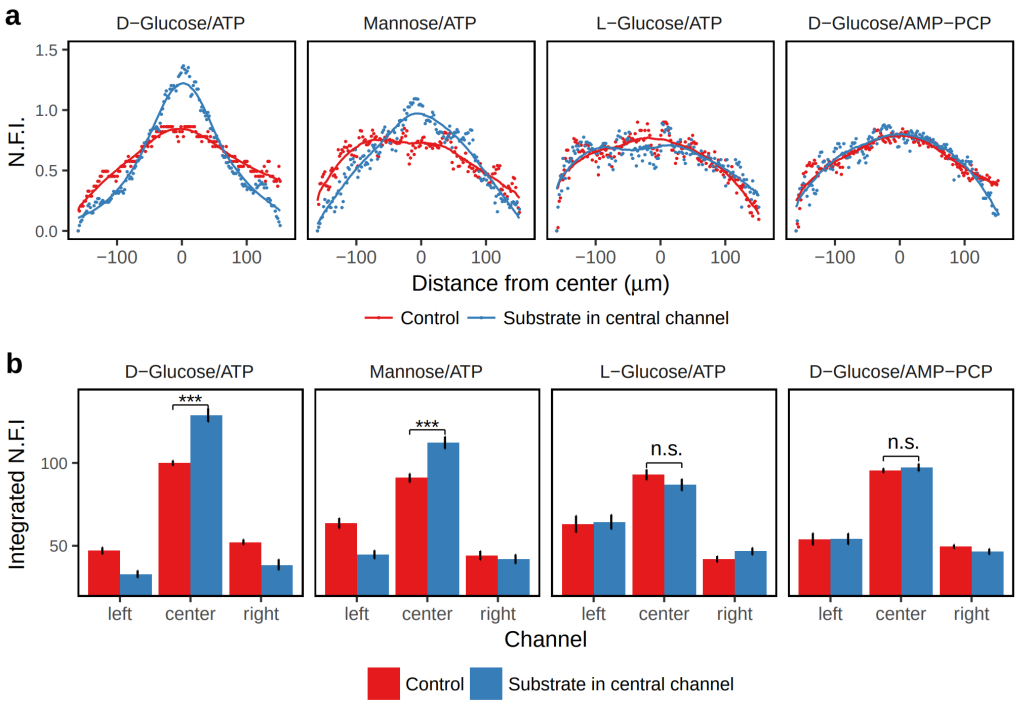

Figure 4.3: A starting equilibrium distribution of HK (200 nM), D-glucose (50 mM) and MgCl2 (100 mM) shows focusing towards the middle channel when ATP (50 mM) is introduced into it. Note that catalysis does not occur in the absence of ATP (control). Experimental conditions: flow rate, 50 µLh^{-1}; distance, 38 mm; interaction time, 34.6 s. The general concave shape of the curves is indicative of the wall effect. a, Normalized fluorescence intensity (NFI) averaged across three experiments as a function of distance from the centre of the channel. Fluorescence intensities are normalized across all channels such that the total fluorescence intensity across all channels is fixed for all experiments and rescaled such that the central channel for the D-glucose control experiment sums to 100. Side channels are shaded in grey. Data points are locally fitted to a second degree polynomial b, Integrated NFI per channel averaged over three experiments. Error bars are 95% confidence intervals obtained from 500 bootstrap iterations of the fitting process. A pairwise t-test with Holm adjustment was conducted to test for significant differences in the intensities across channels. The pairwise t-test for the sum of the left and right channels would give the same results because the total fluorescence across the three channels is normalized for each experiment. ***: $P < 0.001$; NS, not significant.

We propose that the chemotactic aggregation of enzymes in regions of high substrate concentrations is due to cross-diffusion effects [102]. The substrate gradient-

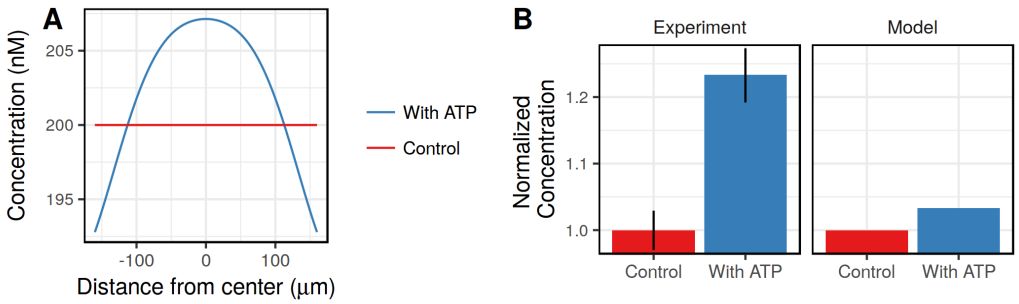

Figure 4.4: **Catalysis-induced enzyme focusing and computed profiles of total enzyme concentration replicating experimental conditions from Fig. 4.3** A, Modelled chemotactic response of HK in the presence of ATP, which is consistent with the experimental result. Parameters were chosen to replicate the conditions of the experiment described in Fig. 4.3A. B, Comparison between experimental results and computational results for the integrated NFI of the enzyme in the central channel: experimental (left) and modelled (right) enzyme focusing in the presence and absence of ATP. The experimental figure is the average of three experimental trials. Error bars are 95% confidence intervals obtained from 500 bootstrap iterations of the fitting process

induced aggregation by cross-diffusion counteracts Fickian diffusion of enzymes, which transfers enzymes from regions with high enzyme concentration to regions with low enzyme concentration. Cross-diffusion is different from the enhanced diffusion of an enzyme in the presence of its substrate [21], [96], [97], [100], which is also observed for uniform substrate concentrations and accelerates the equilibration of the enzyme concentration by Fickian diffusion. The complete theoretical description of diffusion in a multicomponent system combines the flow of a species in proportion to its concentration gradient (Fick's law) and the flow of the same species in response to the concentration gradients of other species in solution. The diffusive flow for the concentration c_e of unbound enzyme E in the presence of its substrate S can then be written as:

$$J_e = -D\nabla c_e - D_{XD}\nabla c_s \tag{4.1}$$

where D is the Fick's law diffusion coefficient, D_{XD} is the "cross-diffusion" coefficient, and ∇c_e and ∇c_s are gradients in enzyme and substrate concentrations, respectively. "Cross-diffusive" effects have been experimentally measured in ternary reaction-diffusion systems [103], protein-electrolyte solutions [25], protein-polymer solutions [26], and in many other systems [24]. We followed the theory of chemotaxis originating from short-range ligand binding proposed by Schurr et al. [104] to obtain the cross diffusion coefficient, D_{XD}, as a function of the local substrate concentration, c_s, the diffusion coefficient, D, computed from the Einstein relation ($70\,\mu m^2/s$ for the HK-glucose complex), and the equilibrium constant K of ATP binding to the enzyme ($5 \times 10^3\,M^{-1}$ for the binding of ATP to HK-glucose [105]):

$$D_{XD} = -Dc_e\frac{K}{1+Kc_s} \tag{4.2}$$

Inserting Equation 4.2 into Equation 4.1 shows the factors driving cross diffusion flow:

$$J_e = -D\left(\nabla c_e - c_e\frac{K}{1+Kc_s}\nabla c_s\right) \tag{4.3}$$

The first term inside the parenthesis is traditional diffusion towards lower concentrations of enzyme. The second term's sign is opposite, showing that this flow is towards higher concentration of substrate. In addition to the substrate gradient, this term's magnitude is determined by three factors: the diffusion coefficient D, the

enzyme concentration c_e, and a factor proportional to the fraction of binding sites occupied by substrate at a given time. As with Fickian diffusion, the cross-diffusion drift arises from a thermodynamic driving force that lowers the chemical potential of the system due to favorable enzyme-substrate binding.

The system of partial differential equations corresponding to the HK-glucose catalysis reaction diffusion system has been solved numerically. The initial presence of ATP in the central channel gives rise to strong ATP gradients at the boundaries between the central channel and the left and right channels. D-Glucose, present in all channels, converts HK to the HK-DG complex, which is the cross-diffusing entity described by Eq. 4.3. Without any adjustable parameters and without accounting for catalysis-induced enhanced diffusion, the model predicts focusing lower than that seen in experiments, but of the same direction and order of magnitude (Figure 4.4). Thus, hexokinase will chemotax up an ATP gradient due to the cross-diffusion phenomenon. One reason for the difference between experiment and theory is enhanced diffusion of the enzymes in the presence of catalysis; increased D will increase the amount of focusing, as predicted by the model. However, since there is no established theoretical framework for the determination of D as a function of position across the microfluidic channel, we have not included it in our model.

We also modeled the focusing experiment in the presence of the non-hydrolyzable ATP analog, AMP-PCP, and found that the model predicts reduced focusing compared to the ATP-induced focusing (around 1% increase in the concentration in the central channel). The significantly stronger binding of AMP-PCP reduces the concentration c_e of unbound enzyme [106], and thereby the cross-diffusion effect. This

suggests that the model is also compatible with the results for the AMP-PCP experiment in which little focusing was observed.

4.4 Cross-Diffusion Model

In this section, we show how cross-diffusion coefficents can be caculated for a ternary system with interacting species. We first outline the background theory, Kirkwood-Buff's statistical mechanics of solutions [107]. We then rederive the theoretical value of cross-diffusion coeffcients for enzymes following [104]. Finally, we make explicit the modeling choices for the glycolitic cascade.

Kirkwood-Buff Theory

We follow closely Newman's [108] reasoning to present the main ideas underlying Kirkwood-Buff theory. We will not rederive all results, but rather will expose the main ideas concisely.

Radial distribution functions

For a solution with a central molecule i surrounded by a solution of molecules j, we take r to be the radial distance from the central molecule i. We assume that the spatial distribution of j is somewhat affected by the presence of i. The radial distribution function $g_{ij}(r)$ is the ratio of the probability density of finding a molecule j at a distance r of the central molecule to the probability density of finding the molecule j at that point if i was not present. These radial distribution functions

can be measured, and present "peaks" relative to 1 when there i creates a layer of j. When $r \to \infty$, $g_{ij}(r) \to 1$. At great distances from the central molecules, their spatial distribution is unaffected. The aim of KB theory is to relate the KB integrals:

$$G_{ij} = \int_0^\infty 4\pi r^2 (g_{ij}(r) - 1) dr \tag{4.4}$$

to thermodynamic properties of the solvent.

Relationship between average properties and thermodynamic properties

For the grand canonical ensemble (ensemble of system with constant T, V, μ), the free energy can be written $E(j) - \sum \mu_i N_i$, where the $E(j)$ are the energy levels, and μ_i are the chemical potentials (partial free energies) of the N_i molecules i. For a two particle solution, the energy levels $E(j)$ and the number of particles N_1 and N_2 can vary between states. We get the partition function summing over all the Boltzmann coefficients:

$$\Xi = \sum_j \sum_{N_1} \sum_{N_2} \mathbf{P}(j, N_1, N_2) \tag{4.5}$$

$$\Xi = \sum_j \sum_{N_1} \sum_{N_2} e^{-\beta(E(j) - \mu_1 N_1 - \mu_2 N_2))} \tag{4.6}$$

Differentiating both these equations with respect to chemical potentials lead to averages of the number of particles. For example,

$$\frac{\partial \Xi}{\partial \mu_1} = \sum_{j, N_1, N_2} \beta N_1 e^{(-\beta(E(j) - \mu_1 N_1 - \mu_2 N_2))} \tag{4.7}$$

Dividing both sides by $\beta\Xi$ yields:

$$\frac{1}{\beta\Xi}\frac{\partial\Xi}{\partial\mu_1} = \sum_{j,N_1,N_2} N_1 \frac{e^{-\beta(E(j)-\mu_1 N_1-\mu_2 N_2)}}{\Xi} \tag{4.8}$$

$$\frac{1}{\beta\Xi}\frac{\partial\Xi}{\partial\mu_1} = \langle N_1\rangle \tag{4.9}$$

Differentiating the equation above again and using the product rule yields [108]:

$$\langle N_1 N_2\rangle - \langle N_1\rangle\langle N_2\rangle = kT\frac{\partial N1}{\partial\mu_2} = kT\frac{\partial N2}{\partial\mu_1} \tag{4.10}$$

We have linked thermodynamic properties with average properties of the system. We now introduce the KB integrals.

Relationship between KB integrals and average properties

Let's take a central molecule 1, in a solution of 2. Then, in a sphere of large radius R, and volume $V = 4/3\pi R^3$, with densities of molecules ρ_i:

$$\langle N_i\rangle = V\rho_i \tag{4.11}$$

If $\rho_{12}(r)$ is the density of particles 2 at a distance r of our central particle, we can write the number of 2 particles in volume V around 1 as:

$$N_{12} = \int_0^R 4\pi r^2\rho_{12}(r)dr \tag{4.12}$$

Then the product of the number of 1 and 2 molecules averages out to

$$\langle N_1 N_2 \rangle = V \rho_1 \int_0^R 4\pi r^2 \rho_{12}(r) dr \tag{4.13}$$

Finally, we note that $g_{12} = \rho_{12}(r)/\rho_2$. Rearranging all above equations gives us:

$$\langle N_1 N_2 \rangle - \langle N_1 \rangle \langle N_2 \rangle = V \rho_1 \rho_2 G_{12} \tag{4.14}$$

Which links the average properties with the KB integrals.

Kirkwood-Buff Theory and Chemotaxis

As in the previous section, this is a summary of the relevant results in Schurr et al. [104] as applied to our system. We use the subscripts W for the solvent, E for the enzyme or macromolecule, and S for the solute.

Chemical potential and standard diffusion.

For the simple case of one solute in solution, the chemotactic force exerted on the enzyme in the direction x is the variation of its chemical potential along that direction. It can be written:

$$F_{ch} = -(d\mu_E/dx) \tag{4.15}$$

In the "normal" case, the chemical potential is

$$\mu_E = \mu_E^0 + kT \ln(c_E) \tag{4.16}$$

where μ_E^0 is the standard chemical potential, and does not vary with x (we will see later that in the three component solution, it does vary with x, whereby the chemotactic force arises).

$$F_{ch} = -(kT/c_E)dc_E/dx \tag{4.17}$$

We want to see that this force, if inserted in the Smoluchowski equation, gives back the traditional diffusion term. The Smoluchowski equation without the diffusive term is:

$$\frac{\partial c_E}{\partial t} = -\frac{D_E}{kT}\frac{\partial F_{ch} c_E}{\partial x} \tag{4.18}$$

Inserting the above chemotactic force in this, we get:

$$\frac{\partial c_E}{\partial t} = D\frac{\partial^2 c_E}{\partial x^2} \tag{4.19}$$

This confirms that the normal diffusion equation can be rederived through the expression of the chemotactic force. Now, there remains to see how $\partial\mu_E/\partial x$ varies in the case where μ_E^0 depends on the solute S.

Theoretical results for chemical potential dependence

We are now back to the three component case. We are interested in how equation 4.17 changes when μ_E^0 changes with x. Differentiating equation 4.16 in the general case, we get:

$$F_{ch} = -d\mu_E^0/dx - (kT/c_E)dc_E/dx \tag{4.20}$$

Where the second term on the RHS has already been discussed above and seen to be the traditional diffusive term. We now focus only on the first term: it is our "cross-diffusion" effect. Re-arranging, we can express it as follows:

$$\frac{\mathrm{d}\mu_E^0}{\mathrm{d}x} = \left(\frac{\partial \mu_E^0}{\partial \mu_S}\right)_{T,p,c_E^\infty} \left(\frac{\partial \mu_S}{\partial c_S}\right)_{T,p,c_E^\infty} \frac{\mathrm{d}c_S}{\mathrm{d}x} \tag{4.21}$$

The second term of the product RHS can be obtained by differentiating equation 4.16 applied to S. We then get:

$$\frac{\mathrm{d}\mu_E^0}{\mathrm{d}x} = \left(\frac{\partial \mu_E^0}{\partial \mu_S}\right)_{T,p,c_E^\infty} \frac{kT}{c_S} \frac{\mathrm{d}c_S}{\mathrm{d}x} \tag{4.22}$$

We now want to express the first term in this product. Kirkwood-Buff theory states that:

$$\left(\frac{\partial \mu_E^0}{\partial \mu_S}\right)_{T,p,c_E^\infty} = -c_S(G_{S,E} - G_{W,E}) \tag{4.23}$$

Then Schurr et al. [104], making use of a geometric argument, show that the RHS of the equation can be redefined as a function of the fraction f of occupied binding sites.

$$f = \frac{Kc_S}{1 + Kc_S} \tag{4.24}$$

where K is the equilibrium constant of the binding. This equation is only valid when $K \gg v_S$, with v_S the partial molar volume of the substrate. In this case, the final

expression for the chemotactic force is then:

$$F_{ch} = kT \frac{K c_S}{1 + K c_S} \left(\frac{\partial \ln c_S}{\partial x} \right)_{T,P,c_E^\infty} \tag{4.25}$$

Application to Hexokinase Focusing

Validity of approximations

The partial molar volume of D-Glucose is of the order of 20 ml/mol, whereas the equilibrium constant K is approximately equal to: $4.3 \times 10^4\,\mathrm{M}^{-1}$ so $K \gg v_3$, and we can apply the above equations.

We are using, contrary to the article [104], $M = 1$. The binding constant K is defined as an equilibrium between the empty (full of water) binding sites, and the full (of glucose) binding sites. The chemical equlilibrium definition of K is a good approximation for this constant.

Reaction diffusion equation with chemotactic force

The chemotactic flow can be rewritten as: $j_{ch} = c_E v_{ch}$, and $v_c h = \frac{F_{ch}}{\gamma}$, where γ is the friction coefficient of E in the solution. Using the Einstein relation, the chemotactic flow is then:

$$j_{ch} = \frac{D c_E F_{ch}}{kT} \tag{4.26}$$

$$j_{ch} = D_{E,E} \frac{K c_E}{1 + K c_S} \left(\frac{\partial c_S}{\partial x} \right)_{T,P,c_E^\infty} \tag{4.27}$$

Hence the final expression for the cross-diffusion coefficient:

$$D_{E,S} = D_{E,E}\frac{Kc_E}{1+Kc_S} \tag{4.28}$$

The general form of the reaction-diffusion equation for our enzyme then becomes:

$$\begin{aligned}\frac{\partial c_E}{\partial t} &= D\frac{\partial^2 c_E}{\partial x^2} + \frac{\partial j_{ch}}{\partial x} + u(\mathbf{c}) &(4.29)\\ \frac{\partial c_E}{\partial t} &= D\frac{\partial^2 c_E}{\partial x^2} + \frac{\partial}{\partial x}\left(\frac{DKc_E}{1+Kc_S}\frac{\partial c_S}{\partial x}\right)\\ &\quad +u(\mathbf{c}) &(4.30)\end{aligned}$$

where $u(\mathbf{c})$ represents the reaction term as a function of the concentration vector $\mathbf{c}$, and D is the diffusion coefficient of the enzyme $D_{E,E}$. The reaction-diffusion equation for other components are similar, save for the cross-diffusion term specific to the enzyme.

Modeling the microfluidic experiment

We followed Wilkinson and Rose [105] for the kinetics of the reaction as follows:

$$\mathrm{E} + \mathrm{G} \underset{60\,\mathrm{s}^{-1}}{\overset{2\,10^{6}\,\mathrm{M}^{-1}\mathrm{s}^{-1}}{\rightleftharpoons}} \mathrm{EG}$$

$$\mathrm{EG} + \mathrm{A} \underset{200\,\mathrm{s}^{-1}}{\overset{1\,10^{6}\,\mathrm{M}^{-1}\mathrm{s}^{-1}}{\rightleftharpoons}} \mathrm{EGA}$$

$$\mathrm{EGA} \underset{8000\,\mathrm{s}^{-1}}{\overset{8000\,\mathrm{s}^{-1}}{\rightleftharpoons}} \mathrm{EG_6D}$$

$$\mathrm{EGA} \xrightarrow{1\,\mathrm{s}^{-1}} \mathrm{G} + \mathrm{EA}$$

$$\mathrm{EG_6D} \underset{0.2\,10^{6}\,\mathrm{M}^{-1}\mathrm{s}^{-1}}{\overset{400\,\mathrm{s}^{-1}}{\rightleftharpoons}} \mathrm{EG_6} + \mathrm{D}$$

$$\mathrm{EG_6} \underset{2\,10^{6}\,\mathrm{M}^{-1}\mathrm{s}^{-1}}{\overset{6\,10^{3}\,\mathrm{s}^{-1}}{\rightleftharpoons}} \mathrm{E} + \mathrm{G_6}$$

$\mathrm{EG_6D} \xrightarrow{4\,\mathrm{s}^{-1}} \mathrm{ED} + \mathrm{G_6}$

$\mathrm{EA} \xrightarrow{500\,\mathrm{s}^{-1}} \mathrm{E} + \mathrm{A}$

$\mathrm{ED} \xrightarrow{100\,\mathrm{s}^{-1}} \mathrm{E} + \mathrm{D}$

where E is the Enzyme, A is ATP, G is D-glucose, G6 is glucose-6-phosphate, and D is ADP. We then modeled each reaction-diffusion equation for every component Q as follows:

$$\frac{\partial Q}{\partial t} = D_Q \nabla^2 Q + R(Q) + XD(Q, S) \tag{4.31}$$

where D_Q is the diffusion coefficient of Q, $R(Q)$ are the reactions as shown above, and $XD(Q, S)$ is the cross diffusion term of Q towards the gradient of S. We modeled the cross diffusion term following equation 4.3:

$$XD(Q, S) = \nabla(C_Q \frac{K}{1 + KC_S} \nabla C_S) \tag{4.32}$$

We then obtained a set of eleven partial differential equations, for which we discretized the spatial derivatives. We then solved this system of discretized equations by numerically integrating in time using the odeint function from the Python Scipy package [109]. The numerical scheme for discretizing in space and then numerically integrating over time is taken from [110].

For the modeling of the AMP-PCP experiments, we modified the kinetic constants as follows:

$$\mathrm{EG} + \mathrm{A} \underset{200\,\mathrm{s}^{-1}}{\overset{10\,10^{\,6}\,\mathrm{M}^{-1}\mathrm{s}^{-1}}{\rightleftharpoons}} \mathrm{EGA}$$

$$\mathrm{EGA} \xleftarrow[8000\,\mathrm{s}^{-1}]{} \mathrm{EG_6D}$$

This is compatible with the 10-fold reduction in dissociation constant of AMP-PCP determined previously [106], and with the absence of glucose phosphorylation. The model, using these parameters, predicted significant reduction in the amount of focusing in the central channel (1.6% for the AMP-PCP focusing).

4.5 Role of Catalysis in Chemotaxis

To confirm the role of substrate turnover in the observed chemotaxis, HK was exposed to its usual substrate, D-glucose, and mannose, a competitive substrate, as well as L-glucose, which is not a substrate. HK shows a higher binding affinity towards mannose compared to D-glucose (Km = 40 M versus 120 M); on the other hand, pyruvate kinase/lactate dehydrogenase coupled assays for HK activity confirmed mannose phosphorylation to be half as fast as D-glucose phosphorylation under similar reaction conditions. In the experiments, the flow rate in each port was set to 200 L/h, and the fluorescence was measured 30 mm down the channel allowing for a total diffusion/interaction time of 6.5 s. Buffer containing 200 nM HK, 10 mM ATP and 10 mM MgCl2 was flowed through the middle channel while one flanking channel contained buffer with 10 mM D-glucose, buffer with 10 mM mannose or buffer with 10 mM L-glucose, and the other channel contained buffer only, as a control.

A significantly higher chemotactic shift was observed towards the D-glucose channel compared to the mannose channel (Figure 4.5) suggesting that catalysis, rather than simple substrate binding, is important for the observed enzyme transport [111].

Although cross-diffusion itself only requires substrate binding, the diffusion coefficient controlling the magnitude of this effect will be significantly affected by catalysis through the enhanced diffusion mechanism. Equations 4.2 and 4.3 show that the cross diffusion coefficient D_{XD} is directly proportional to the enzymes diffusion coefficient D. The magnitude of the enzymes diffusion coefficient is therefore one of the determining factors of enzyme chemotaxis and focusing.

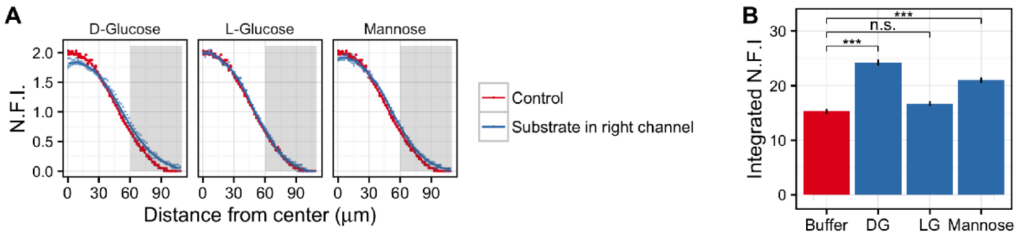

Figure 4.5: **Chemotactic shifts observed for HK in response to gradients of different substrates.** HK shows a greater chemotactic shift towards its preferred substrate D-glucose compared to mannose, which it phosphorylates at a significantly lower rate. No chemotactic shift was observed with L-glucose. A, Experimental NFI in the central and right channels. All fluorescence intensities are normalized to a total of 1 across all channels, corresponding to a fixed total amount of enzyme in each experiment. B, Integrated NFI in the right channel. Experimental conditions: starting enzyme concentration: 200 nM (100%); flow rate: 200 µLh^1; distance: 30 mm; interaction time: 6.5 s. The percentage of enzyme migration into the D-glucose channel is 7.3 ± 2.0% and into the mannose channel is 2.5 ± 1.2%, relative to buffer channel. Error bars are 95% confidence intervals. A pairwise t-test with Holm adjustment was conducted to test for significant differences in the intensities across channels. ***: P < 0.001; NS, not significant.

4.6 Chemotaxis as a Factor in Metabolon Formation

Having demonstrated that HK undergoes chemotaxis up its substrate gradient, we then probed the chemotactic behavior of the entire four-enzyme cascade. The first experiment was designed to examine the response of Ald towards its substrate, fructose 1,6-bisphosphate, generated from D-glucose by the successive actions of the first three enzymes. In the microfluidic device, the Ald was flowed through the middle channel. The first three enzymes, HK, Iso and PFK, with Mg2+ and ATP (required by the kinases) were passed through one of the flanking channels along with 10 mM D-glucose, while buffer was passed through the flanking channel on the opposite side. The volumetric flow rate per inlet was fixed at 50 µL/h allowing for a total interaction time of 17.3 seconds in a 20 mm long channel. 11.9 ± 3.0 % of the Ald moved into the channel where its substrate was being formed in situ (Figure 4.6A). When the interaction time was reduced to 8.6 s, the chemotactic migration correspondingly reduced to 4.9 ± 2.4 % of the Ald.

We then sought to examine whether there was a sequential spreading of HK and Ald when exposed to D-glucose. This is expected since D-glucose is the substrate for HK, while the substrate for Ald, fructose 1,6-bisphosphate, is only formed from D-glucose through three successive enzymatic steps. The components of the cascade were now separated into two batches consisting of the first two and the last two enzymes, respectively. HK, ATP, Mg2+, and Iso were flowed through one flanking channel, while PFK, ATP, Mg2+, and Ald were flowed through the other flanking

channel. A solution of D-glucose passed through the middle channel. The flow rate was reduced to 30 µL/h and the channel length was increased to 40 mm allowing for a total interaction time of 57.6 s within the channel. As discussed, we hypothesized that HK should respond first to its substrate gradient by moving into the D-glucose channel, thereby producing the substrate for enzyme 2, Iso. The cascade would continue with PFK participation, finally producing fructose 1,6-bisphosphate that in turn should prompt Ald to chemotax towards the central channel. The fluorescence profiles for enzymes HK and Ald were noted at different interaction times, 14.4 s, 28.8 s, 43.2 s and 57.6 s, and their chemotactic behavior is summarized in Figure 4.6B. For HK, our results indicate that, in 57.6 s, 37.0 ± 0.3% of the starting 200 nM enzyme moves into the central channel containing D-glucose (10 mM) compared to 6.7 ± 1.3% of the enzyme moving into the same channel when flowing only buffer. The corresponding numbers for Ald are 8.9 ± 0.7% and 5.9 ± 1.0%, respectively. Thus, a sequential movement of HK, followed by Ald towards the central channel was observed. We also observed a sequential movement of the two enzymes when we added mannose to D-glucose. Since mannose binds more strongly to HK but turns over more slowly, a smaller chemotactic shift is observed.

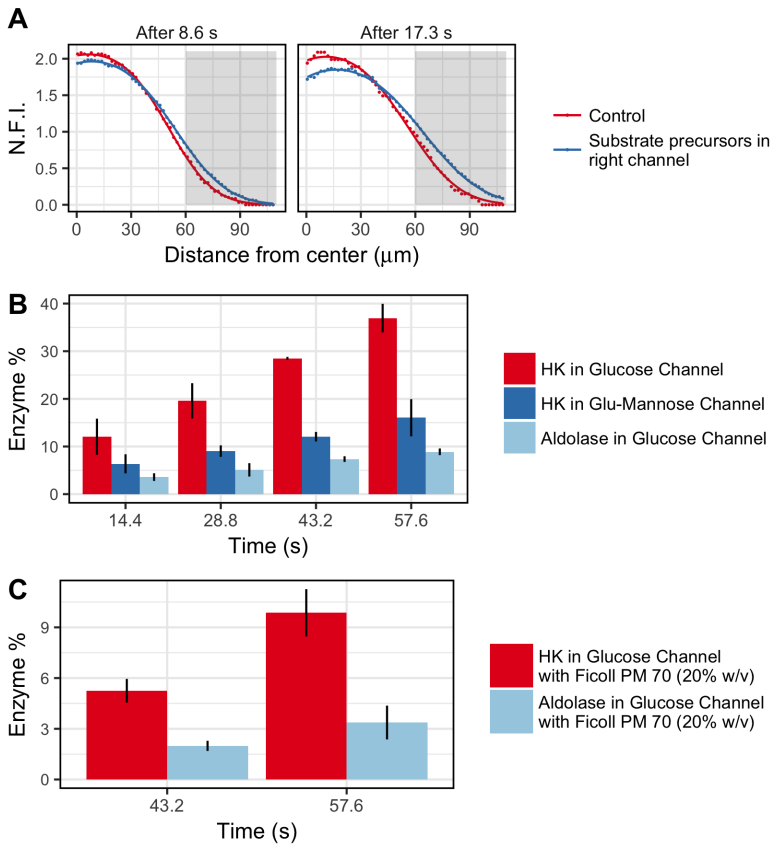

Figure 4.6: **Chemotactic assembly of enzymes in the microfluidic channel under different reaction conditions.** A, Fluorescence intensity measured across the channels, plotted against width of the channels for the centre and right channels. The grey background represents the approximate right channel. When compared to the movement towards buffer, Ald shows enhanced migration into the channel that generates fructose-1,6-bisphosphate in situ. B, Ald shows a time-delayed chemotactic response compared to HK, as expected based on the sequence of reactions. When 10 mM mannose is introduced along with 10 mM D-glucose, HK shows a reduced chemotaxis corresponding to the slower rate of mannose phosphorylation by HK. Error bars are 95% confidence intervals. C, D-glucose gradient-driven sequential movement of HK and Ald for the entire enzymatic reaction cascade was also observed in the presence of Ficoll PM 70 (20% wt/vol), an induced crowded environment mimicking cytosolic conditions in a cell. Error bars are 95% confidence intervals.

4.7 Chemotactic Co-localization of Hexokinase and Aldolase

With the same crowding conditions and enzymes used in the microfluidic experiments, we also observed the co-localization of HK and Ald (metabolon formation) in a sealed

hybridization chamber starting with a uniform distribution of all the four enzymes in the cascade, as well as the substrates for HK. In the presence of D-glucose and ATP, both the fluorescently labeled HK and Ald form bright moving spots. When the spots of HK and Ald with diameters ranging from 600 to 1000 nm were tracked, the trajectories of the two enzymes were found to be highly correlated, suggesting metabolon assembly during enzyme cascade reactions (Figure 4.7, Table 4.1). Similar experiments were also performed either with D-glucose but no Iso and PFK present, or substituting D-glucose with L-glucose, or with no glucose. As shown in Table 4.1, there were far fewer HK spots and fewer Ald trajectories that correlated with HK trajectories.

Table 4.1: Statistic of HK and Ald trajectories

Experiment	Total HK trajectories	HK trajectories with high Ald correlation
D-Glucose with all four enzymes	48 ± 3 (s.e.m.)	32 ± 2 (s.e.m.)
D-Glucose without Iso and PFK	12 ± 2	5 ± 1
L-Glucose with all four enzymes	1	0
No glucose with all four enzymes	1	0

Analysis of HK and Ald Aggregates Trajectories

To test for the presence of aggregates, we ran a series of computational operations to detect punctates with sub-pixel positional accuracy using a customized version of the Python package Trackpy [112]. First, the parameters of the object detection algorithm were calibrated using visual inspection of two randomly chosen time points in the videos, both of the Aldolase channel and of the Hexokinase channel. The

parameters chosen are shown in Table 4.2. The detection algorithm thresholds the image, keeping only grayscale values above an adaptively determined value. A band pass filter is applied to the Fourier-transformed grayscale image (to remove artifacts and background noise). A two-dimensional Gaussian is then fit to the brightness peaks to locate the object. Then, the identified objects were assigned to trajectories by a linking algorithm. The linking algorithm determines potential trajectories by tracking objects from frame to frame, specifying a maximum distance the object can travel from frame to frame and a "memory" allowing the object to disappear for up to five frames. We then removed trajectories where the object was present for less than ten frames.

The parameters were calibrated very conservatively to only identify obvious trajectories. The parameters were then kept constant across all trajectories, channels, and experiments. Once the trajectories were identified, we defined a rectangular region around each HK trajectory adding a margin of 10 pixels on each side. Within this region, the closest Ald trajectory was chosen as potentially originating from the same HK-Ald aggregate as the HK trajectory. If there was no candidate trajectory, we deemed that there was no HK-Ald aggregate corresponding to this HK trajectory. The above analysis was run on five experiments in which D-glucose was present (Figure 4.7), three experiments with L-Glucose, and two experiments without glucose. We identified HK trajectories and Ald trajectories (punctates which persist for at least 10 frames), and the subset of HK trajectories for which there is a corresponding ALD trajectory for which the spatial correlation is higher than 95% (Table 4.1). Over half of the detected HK trajectories have an overlapping Ald trajectory. The presence

of correlated HK and ALD trajectories suggests that multi-enzyme aggregates are formed. We note that although a significant fraction of the HK trajectories do not appear to have overlapping Ald trajectories, the quality and duration of the overlap between the a priori independent HK and Ald trajectories provides overwhelming evidence that HK-Ald aggregation occurs, which is reflected in the p-value calculated below.

We ran a Pearson correlation analysis between the coordinates of the HK trajectory and that of the corresponding Ald trajectory. For each pair of trajectories, we tested for the null hypothesis that the correlation between trajectories was zero, using a Ljung-Box test. If we found no Aldolase trajectories near the hexokinase trajectory, the null hypothesis was accepted. The p-values were then aggregated using Fisher's method. The aggregated p-value for the D-glucose experiments was lower than 10-80. We can therefore reject the null hypothesis for the glucose experiments. By contrast, the p-value for the 'no-glucose' experiment was 0.24 and the p-value for the 'l-glucose' experiments was greater than 0.8, leading us to accept the null for these experiments.

Table 4.2: Statistic of HK and Ald trajectories

Parameter	Value
Minimum Distance between two aggregates	20 pixels
Minimum number of bright pixels	5 pixels
Number of pixels around the HK trajectory	10 pixels
Maximum number of pixels a particle moves between two frames	10 pixels
Number of frames a particle can disappear during trajectory	5 frames
Number of frames a trajectory has to last to be considered	10 frames

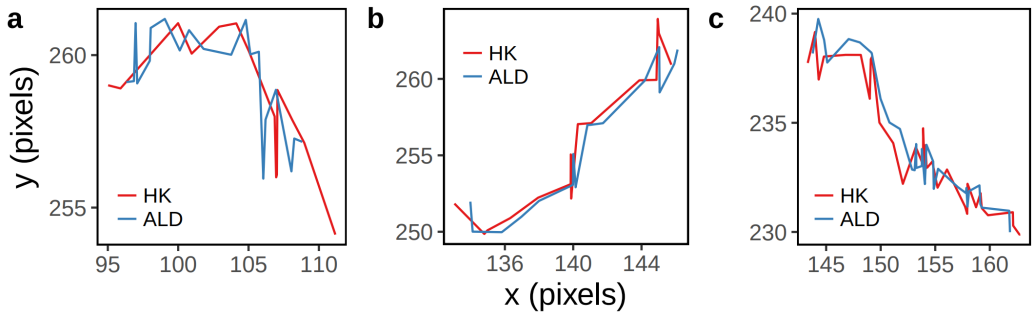

Figure 4.7: **Examples of HK and Ald trajectories from an experiment in which D-glucose and all four enzymes were present, for which the corresponding Ald trajectory was highly correlated.** Experimental conditions: 200 nM HK labelled with amine-reactive (excitation/emission: 493/518 nm) Dylight dye; 200 nM Iso; 200 nM FPK; 200 nM Ald conjugated with thiol-reactive (excitation/emission: 638/658 nm) Dylight dye; 10 mM ATP; 20 mM MgCl2; and 10 mM D-glucose in 20% wt/vol 70 M Ficoll mixed and injected into a sealed hybridization chamber. A pixel is 0.46 × 0.46 µm and the frame rate is 1 frame every 1.29 s. Trajectories are recorded for 10 frames, or 13 s.

4.8 Conclusion

Our results suggest that the observed assembly of enzymes participating in a cascade in response to the presence of the initial substrate can be a result of individual enzymes undergoing chemotaxis in response to their specific substrate gradients. We identified and quantified the two major effects explaining chemotaxis: first, in the case of HK cross-diffusion up the ATP and glucose gradients is the main mechanism causing localization. It is dependent on ATP and glucose binding. Second, the magnitude of the effect is increased by the enhanced diffusion effect, which we have shown to be dependent on catalysis, when both ATP and glucose are present. The extent of enzyme migration is proportional to the exposure time to the substrate gradient. The reduced chemotaxis with mannose, a less active substrate for HK, emphasizes the contribution of catalysis to the phenomenon. This phenomenon, chemotaxis, does

not require the need for direct interaction between the enzymes to form complexes that promote substrate channeling; metabolon formation could simply be triggered by the presence of an initial substrate gradient, for example ATP gradients near mitochondria in the case of the transient metabolon, the purinosome. Furthermore, the enzymes should revert to their equilibrium distribution once the initial substrate is completely reacted and the substrate gradients for the individual enzymes disappear. Presuming this phenomenon to be general [106], chemotaxis may be a basis for the organization of metabolic networks in the cytosol of the cell.

Chapter 5

Velocity Fluctuations in Kinesin-1 Gliding Motility Assays Originate in Motor Attachment Geometry Variations

Motor proteins such as myosin or kinesin play a major role in cellular cargo transport, muscle contraction, cell division, and also in engineered nanodevices. Quantifying the collective behavior of coupled motors is critical for our understanding of these systems. An excellent model system is the gliding motility assay, where hundreds of surface-adhered motors propel one cytoskeletal filament such as an actin filament or a microtubule. The filament motion can be observed using fluorescence microscopy, revealing fluctuations in gliding velocity. These velocity fluctuations have been previously quantified by a motional diffusion coefficient, and have been explained by the addition and removal of motors from the linear array of motors propelling the filament as it advances, assuming that different motors are not equally efficient in their force generation. A computational model of kinesin head diffusion and binding to the microtubule allowed us to quantify the heterogeneity of motor efficiency arising from the combination of anharmonic tail stiffness and varying attachment geometries assuming random motor locations on the surface and an absence of coordination between motors. We then experimentally measured the diffusion coefficient and obtained agreement with the model. This allowed us to quantify the loss in efficiency

of coupled molecular motors arising from heterogeneity in the attachment geometry.

This chapter is an abbreviated version of: *Palacci, Henri, Ofer Idan, Megan J. Armstrong, Ashutosh Agarwal, Takahiro Nitta, and Henry Hess. 2016. "Velocity Fluctuations in Kinesin-1 Gliding Motility Assays Originate in Motor Attachment Geometry Variations." Langmuir: The ACS Journal of Surfaces and Colloids 32 (31): 794350.*

5.1 Introduction

Ensembles of molecular motors are fascinating objects of study in the field of complex dynamical systems because they combine mechanical complexity with chemical stochasticity [113]. The collective behavior of biomolecular motor proteins from the kinesin, dynein and myosin families and their associated cytoskeletal filaments (microtubules and actin filaments) can be investigated through the construction of in vitro model systems called gliding motility assays where the system dynamics are readily observable by fluorescence microscopy [114].

In a gliding motility assay, the motor proteins tails are attached to a surface, and their heads bind to a cytoskeletal filament. As the motors step along the filament, the filament is propelled forward. As a result, motors are binding to the tip of the filament and unbinding from its end, thus changing the linear array of motors upon each binding and unbinding event. The elucidation of the dynamics of molecular motors in gliding motility assays has been a goal of theoretical [115]–[124] and experimental efforts [125]–[127] for 20 years.

Kinesin-1 motor proteins, prominent examples of processive motors [128], bind to microtubules and execute force-producing steps of constant length $d = 8\text{nm}$ [129]. The number of steps in a given time interval is Poisson-distributed [130], so that the movement of a microtubule propelled by a single kinesin can be characterized by an average velocity v and a diffusive term according to:

$$\left\langle (\Delta X(t) - v\Delta t)^2 \right\rangle = 2D_m \Delta t \tag{5.1}$$

where $X(t)$ is the position of the filament along its trajectory, $dX = X(t + \Delta t) - X(t)$ is the displacement of the filament during time Δt, and D_m is the motional diffusion coefficient. The motional diffusion coefficient characterizes the fluctuations around the linear velocity of the filament. Measurements with optical tweezers have shown the motional diffusion coefficient D_m to be equal to 1400 nm^2/s for movement driven by individual kinesin-1 motors at saturating ATP concentrations ($v = 670\text{nm/s}$), which is about half of the value of $D_m = vd/2 = 2700\text{nm}^2/\text{s}$ expected for a Poisson stepper with a step size $d = 8\text{nm}$ [130].

When two surface-adhered kinesins are bound to the same microtubule, tracking of the microtubule position with nanometer accuracy has revealed that the microtubule shifts position in increments of one half of a kinesin step because the low viscous drag on the microtubule leads to a near-instantaneous sharing of the displacement between the attached motors [127]. However, the steps of the two motors are uncorrelated, and already a third motor modifies the microtubule dynamics so that distinct steps cannot be detected anymore.

As the density of surface-adhered kinesins is increased, the average spacing between motors can be reduced all the way to 10 nm [131], which implies that hundreds of motors interact simultaneously with a microtubule. The mean microtubule velocity has been found to be largely unaffected by the density of kinesin-1 motors, which is expected when the drag on the microtubule is small compared to the force generated by the motors, and the motors do not hinder each other so that they can step at the same velocity as a single unencumbered motor [132]. When a large number of motors is attached, individual steps in the microtubule motion cannot be distinguished [127]. In the limit of many motors, the reduced individual step sizes should lead to a reduction in velocity fluctuations, and for a given motor density, the motional diffusion coefficient should be reduced with microtubule length: indeed, as more motors are attached to the microtubule, the fluctuations should be reduced according to the central limit theorem. However, it has been observed that this is not the case for high kinesin densities [125]. This apparent paradox is resolved by a theoretical analysis by Sekimoto and Tawada [120], which incorporates the heterogeneity of motor force generation into the analysis of the motion. They conclude that the reduction of velocity fluctuations by the addition of independently acting motors is balanced by an increase in the velocity fluctuations due to the addition and removal of heterogeneous motors from the linear array propelling the filament as it advances [120]. As a result, the motional diffusion coefficient D_m is independent of the length of the microtubule for high kinesin densities [125]. Sekimoto and Tawada's model attributes the heterogeneity of the motors to heterogeneity in their step sizes. However, for kinesin gliding assays, while filament displacement can vary, step sizes are constant at $d = 8$nm due

to the spacing of the tubulin dimers, and their model cannot be applied.

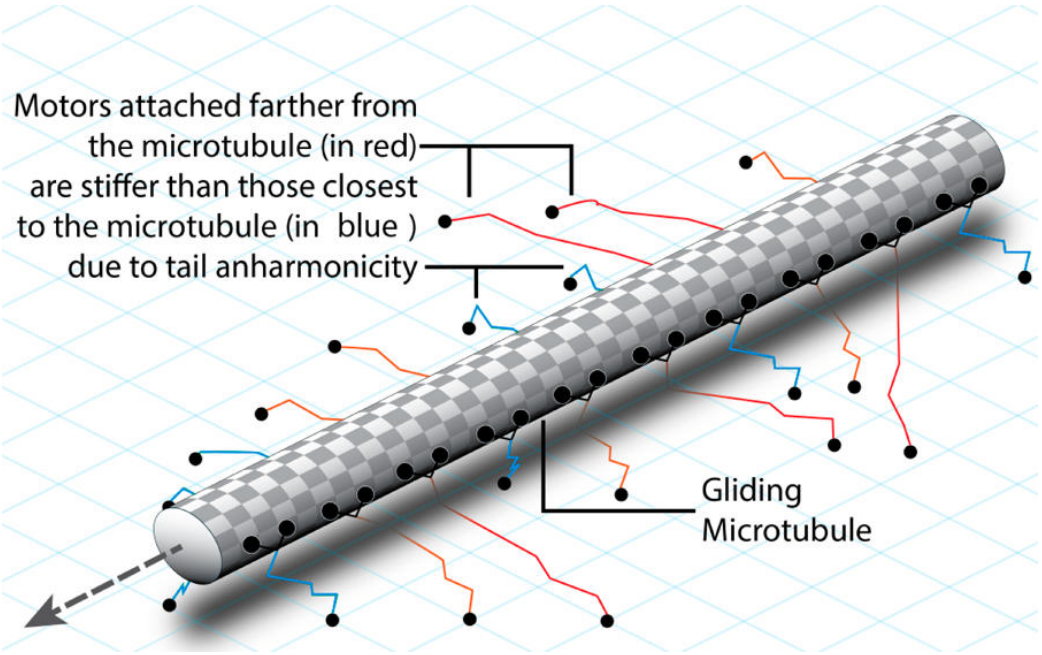

Figure 5.1: Motors step hand over hand on the microtubules surface, thus stretching their tail and propelling the microtubule forward. The motors bound farther from the microtubules axis are stretched more than the bound motors close to the axis. Because the kinesin tail is an anharmonic spring, the stretched motors have a higher force contribution to the forward movement.

Our extension of Sekimoto and Tawada's model to motors with constant step sizes relies on the attribution of the heterogeneity in force production to the anharmonic stiffness of the kinesin tail [133]. The tails of bound motors attached to the surface at a greater distance from the microtubules axis will tend to have greater stiffnesses, and therefore larger force contributions to the shifting force balance between motors as a result of a step (see Fig. 5.1). The diffusion coefficient, restating Sekimoto and Tawada's equality in terms of force adjusted step sizes, becomes:

$$D_m = \frac{\left\langle \left(k_i - \langle k_i \rangle\right)^2 \right\rangle}{\langle k_i \rangle^2} \frac{v}{2\rho} = \alpha_k \frac{v}{2\rho} \tag{5.2}$$

where k_i is the local stiffness of motor i, ρ is the linear density of motors, and the constant α_k quantifies the heterogeneity of the motors stiffnesses.

Computational models have previously been used to determine attachment geometry under harmonic potentials [134], to provide insight into the mechanisms of myosin-coated tracks [135], or to predict myosin ensemble processivity [136]. Here we modeled the diffusion and binding of the kinesin head under an anharmonic spring potential using Brownian dynamics [137]. This allowed us to determine the spatial distribution of the bound motors as well as the distribution of their tail extension. The force-extension relationship for the kinesin tail then yielded the distribution of bound motor stiffnesses and allowed us to determine a value for the heterogeneity factor α_k of 0.3 (SEM $\pm$ 0.002).

The spatial distribution of attached motors given by our model also allowed us to determine an effective surface width around the microtubule from which motors bind. This effective width was then used to determine a linear density of bound motors from kinesin surface densities derived from landing rate measurements. Landing rate measurements, in contrast to ellipsometry of quartz crystal microbalance measurements, determine the density of functional motors on the surface (not just the total protein adsorbed) [132], [138], [139]. Combined experimental measurements of the motor surface density, velocity and motional diffusion coefficient enabled us to determine, for the first time, the constant α in a kinesin/microtubule motility assay for high kinesin densities. A constant α of 0.43 ± 0.3 SEM (Standard Error of the Mean) was measured, therefore theoretical and experimental results are in agreement within experimental error. The relatively large error in the experimental determination of α

is primarily a result of the uncertainty in the method to measure the motor density and cannot be noticeably reduced by analyzing a larger number of gliding filaments. While this agreement between the model and the experiment does not validate the microscopic model, it implies that the experimental observations are consistent with non-interacting motors randomly distributed on the surface.

Determining the origin of velocity fluctuations is a critical step in designing efficient molecular motor-based nanodevices. We have shown statistical agreement between experimental velocity fluctuation data and our model. Our model links the heterogeneity of force production to the heterogeneity of the attachment geometry. Therefore heterogeneity of attachment geometry is a main factor in limiting the energy efficiency of the motor array. Our model can be used in the design of optimized devices, such as motility assays with microfabricated tracks or muscle-like actuators with well-aligned motors.

5.2 Methods

Gliding motility assays

The experiments were performed at a temperature of 25 °C in approximately 100 µm high and 1 cm wide flow cells assembled from two coverslips and double-stick tape [140]. A kinesin construct consisting of the wild-type, full-length Drosophila melanogaster kinesin heavy chain and a C-terminal His-tag was expressed in Escherichia coli and purified using a Ni-NTA column [141]. Microtubules were prepared by poly-

merizing 20 µg of rhodamine-labeled tubulin (Cytoskeleton Inc., Denver, CO) in 6.5 µL of growth solution containing 4 mM of MgCl2, 1 mM of GTP, and 5% DMSO (Dimethyl Sulfoxide) (v/v) in BRB80 buffer (80 mM of PIPES, 1 mM of MgCl2, 1 mM of Ethylene Glycol Tetraacetic Acid, and pH of 6.9) for 30 min at 37 °C. The microtubules were then 100-fold diluted and stabilized in 10 M paclitaxel (Sigma, Saint Louis MO). The microtubule lengths are Schulz-distributed [142] with an average length of 10.5 µm, a standard deviation of 7 µm and a minimum length of 3 µm. The same microtubule preparation was used for all kinesin surface densities. The flow cells were first filled with a solution of casein (0.5 mg/mL, Sigma) dissolved in BRB80. After 5 min, it was exchanged with a kinesin solution of concentrations corresponding to motor surface densities of 310 ± 100, 620 ± 200, 1250 ± 400, 2500 ± 790, 3100 ± 1180 μm^{-2} (all errors are SEM), obtained from landing rate measurements described previously [138], in BRB80 with 0.5 mg/mL of casein and 1 mM of ATP. After another 5 minutes, this was exchanged against a motility solution (10 µM of paclitaxel, an antifade system made up of 20 mM of D-glucose, 20 µg/mL of glucose oxidase, 8 µg/mL of catalase, 10 mM of dithiothreitol, and 1 mM of ATP in BRB80) containing 6.4 µg/mL microtubules, and was injected for 5 minutes, followed by two washes of motility solution (without microtubules) to remove excess tubulin. Each flow cell was immediately moved to an epifluorescence microscope (Nikon TE2000), and movies of 5 different fields of view were taken using a 40x oil objective. The flow cell was imaged every two seconds for 200 seconds per movie with an exposure time of 200 ms, leading to 100 observations per microtubule. Therefore a total of 2000 instantaneous velocities for each kinesin densities was obtained. The camera used was

an iXON DU885LC (Andor Technology Ltd.) electron-multiplying charge-coupled device (EMCCD). The pixel size on the EMCCD was 8 x 8 µm corresponding to 200 x 200 nm in the object plane.

For each kinesin density, 20 smoothly gliding microtubules were tracked using ImageJ software (NIH), and the tip location was manually determined at every frame. While automated tracking software has made great progress in the last few years [143]–[145], here the expected gain in accuracy is small since a reduced position measurement error mainly affects the offset in the fluctuation analysis (Figure 5.4c).

Using MATLAB (Mathworks, Inc.), the distance between two consecutive tip locations, r_j was measured. The cumulative time interval after i image acquisitions is defined as $\Delta t(i) = i\delta t$, where $\delta t = 2s$ is the time between image acquisitions. The cumulative distance travelled over the cumulative time interval $\Delta t(i)$, starting at time j is the sum of single steps, $x_j(i) = \sum_j^{j+i} r_k$, where r_k is the k-th distance between tip locations. We therefore obtain, for each trajectory, 100 cumulative distances travelled for time interval of $\Delta t(1) = \delta t$, 50 cumulative distances travelled for a time interval of $\Delta t(2) = 2\delta t$, etc. For each microtubule and time interval, the deviation from the mean cumulative distance travelled, $\Delta x_j(i)$ was calculated as $\Delta x_j(i) = x_j(i) - x_0(i)$, where $x_0(i)$ is the mean cumulative distance travelled over time interval $\Delta t(i)$. The mean square deviation (MSD) for time interval $\Delta t(i)$, $\langle(\Delta x_j(i))^2\rangle = \langle(\Delta x(i))^2\rangle$ was then calculated as an average over all j of the square deviations over time interval $\Delta t(i)$. We then performed a linear fit of the MSD as a function of the time interval.

The diffusion coefficient is related to the slope of this linear fit through the equation:

$$\left\langle (\Delta x)^2 \right\rangle = 2D_m \Delta t + \sigma_{err}^2 \tag{5.3}$$

where σ_{err}^2 is the variance of the distance measurement errors [124].

The diffusion coefficient for each kinesin concentration was then calculated by averaging the slopes of the linear fits over the 20 microtubules.

To test for the potential length dependence of the velocity fluctuations, we followed Imafuku et al. [125] and fitted the following equation to our experimental data (SI6):

$$D_m(L, \rho) = \frac{kT}{L\zeta} + D_m(\rho) \tag{5.4}$$

where L is the length of the filament, and ζ is the friction coefficient per unit length, k is the Boltzmann constant, T is the temperature, and $D_m(\rho)$ is a length-independent diffusion term. In accordance with Imafuku et al. [125], we find that the length dependent term is negligible compared to $D_m(\rho)$ for high kinesin concentrations (kinesin surface densities above $310\,\mu m^2$). We therefore restricted our analysis to the higher kinesin densities, and focused our analysis on $D_m(\rho)$.

To calculate the heterogeneity factor, the linear density of kinesins was calculated based on the surface density:

$$\rho = \sigma w \tag{5.5}$$

where w is the effective width of the region on the surface from which kinesins can attach to the microtubule. This width is an output of the computational model of

kinesin head diffusion and binding to the microtubule (see Results section). We find $w = 88$ nm. The diffusion coefficient was plotted as a function of the inverse of the linear density, and mean squares regression was used to fit the experimental values to a linear function. The constant α was then calculated according to Eq. 5.2 using this linear fit.

Simulating kinesin head diffusion to determine attachment geometry

In order to compute the tail stiffness distribution and linear density of motors from Eq. 5.2, we use a Brownian Dynamics model of diffusion of the kinesin tail to determine the tail extension distribution and effective binding width. We denote the position of the kinesin head at time t by $r(t)$ in Cartesian coordinates. The motion of the kinesin head in each dimension i is given by:

$$\frac{\mathrm{d}x_i(t)}{\mathrm{d}t} = \frac{1}{\zeta}(f_i^k + f^r) \tag{5.6}$$

where f^k is the elastic force exerted on the diffusing head by the kinesin tail, f^r is the random Brownian force of mean 0 and variance given by:

$$\langle f^r(t_1) f^r(t_2) \rangle = 2kT\zeta\delta(t_1 - t_2) \tag{5.7}$$

where the friction coefficient ζ is given by the Einstein relation for the diffusing tethered kinesin head $D^{kinesin} = kT/\zeta$, δ is the Dirac delta function, k is Boltz-

mann's constant and T is the temperature. The value for this diffusion coefficient has been found to be $20\,\mu m^2/s$ including hydrodynamic effects [146]. We therefore set $D^{kinesin} = 20\ \mu m\, s^{-2}$ in the simulations.

The elastic force f^k as a function of tail length at time t, l, was determined based on a numerical inversion of the freely jointed chain (FJC) force-extension relation:

$$r(t) = \sum_{i=1}^{n} \coth\left(\frac{f^k l_i}{kT}\right) - \frac{kT}{f^k l_i} \tag{5.8}$$

where n is the total number of segments i of Kuhn length l_i. The kinesin head has been found to be linked, through a series of 5 stiffer coiled-coil segments, to its globular tail segment, with a total contour length of 57 nm including head and tail, thus justifying the freely jointed chain approximation. We here used for the entropic spring n = 6 segments, corresponding to the head and freely moving segments between head and tail with the lengths l_i = (8 nm, 15 nm, 10 nm, 5 nm, 6 nm, 8 nm) as specified in [147], and consider the initial tail segment immobilized on the glass surface. In this model the tail stiffness exhibits significant non-linearity, increasing from an initial stiffness on the order of 20 fN/nm to the pN/nm range (see Fig. 5.2). The FJC model has been previously used to model tether stiffness [148], [149], and fits previous determinations of tether stiffness for low extension [113], [150].

We then discretize these equations for a time step $\Delta t = 0.1\,\mu s$ and run 1000 simulations for a given distance d_{attach} between the microtubules axis projection on the surface and the kinesin tails attachment point. We repeat this for d_{attach} taking all integer values between 0 and 50 nm, for a total of 5×10^4 simulations. Each

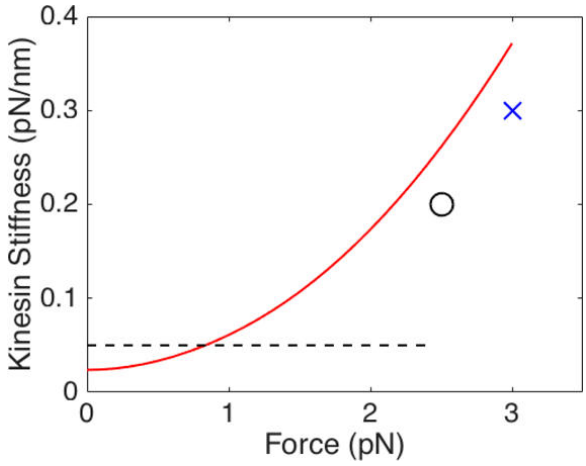

Figure 5.2: Kinesin-1 tail stiffness as a function of applied force. Red curve: Freely jointed chain model. Black dashed line and circle: approximation of kinesin tail stiffness used by Driver et al. [113] based on their experimental data. Blue cross: approximation of kinesin tail stiffness used by Coppin et al. [150] based on their experimental data. A force of 3 pN is exerted by a fully extended kinesin tail.

simulation stops upon the binding event when the kinesin head reaches the surface of the microtubule, modeled as a 25 nm diameter cylinder held 17 nm above the surface [147], or after $t_{max} = 5$s if no binding event has occurred. We choose this value for t_{max} because initial simulations with longer binding windows showed that only 0.01% of motors have not bonded to the microtubule after 2.5 s. We do not model the unbinding of motors, as under our assumptions a completely unbound motor will rebind immediately to the microtubule close to the initial binding site. This also ensures that the initial out of equilibrium extension energy distribution is preserved over the binding time scale for one motor.

Modeling the heterogeneity in motor efficiency

In Sekimoto and Tawada's approach [120], the filament is initially in mechanical equilibrium: the bound motor elastic forces on the microtubule are balanced. Motor

i then steps a distance a_i at time t, resulting in a displacement ΔX of the filament to restore mechanical equilibrium. Note that this approximation, and therefore our model, is only valid when the filament is unloaded and in the limit of many motors attached to the filament. Sekimoto and Tawada's original expression for the motional diffusion coefficient was then:

$$D_m = \frac{\left\langle (a_i - \langle a_i \rangle)^2 \right\rangle}{\langle a_i \rangle^2} \frac{v}{2\rho} = \alpha \frac{v}{2\rho} \tag{5.9}$$

where v is the gliding velocity and ρ is the linear density of motors (the inverse of the average spacing), a_i are the effective step sizes of motor i, and the constant α the heterogeneity of the motors. A central assumption is that the resistance of the motor to stretching can be described by a spring constant k in the harmonic approximation.

In our approach, the motor step size is fixed at $d = 8\,\mathrm{nm}$. However the mechanical equilibrium, and thus the filament displacement will depend on motor i's stiffness k_i.

Assuming motor i steps at time t, and that the microtubules position at time t is $X(t)$, Sekimoto and Tawada's equilibrium condition at t^- (right before the step) and t^+ (right after the step) is:

$$\Delta X = X(t^+) - X(t^-) = \frac{a_i}{N} \tag{5.10}$$

where N is the total number of attached motors and a_i is the step size of motor i. In our model with fixed step sizes and heterogeneous kinesin stiffnesses, the

equilibrium condition becomes:

$$\Delta X = X(t^+) - X(t^-) = \frac{k_i d}{N\langle k_j \rangle} \tag{5.11}$$

where d is the 8 nm step size for the kinesin motor and k_i is the stiffness of motor i. We can therefore, by analogy between Equations 5.10 and 5.11, define $\tilde{a}_i = k_i d/\langle k_j \rangle$. Combining this definition of force-adjusted step sizes with Eq. 5.9 yields the following expression for the motional diffusion coefficient:

$$D_m = \frac{\left\langle (k_i - \langle k_i \rangle)^2 \right\rangle}{\langle k_i \rangle^2} \frac{v}{2\rho} = \alpha_k \frac{v}{2\rho} \tag{5.12}$$

Using the simulation results for the distribution of the stiffnesses k_i then allows us to determine the theoretical value of α_k.

5.3 Results

Theoretical determination of the heterogeneity factor

The frequency of occurrence of binding at a specific tail extension as a function of the distance d_{attach} between kinesin surface attachment point and the projection of the microtubule axis on the surface (Fig. 5.3A) was determined using our computational model. We used this data to estimate the probability of a kinesin binding as a function of its horizontal distance from the microtubule axis d_{attach} (Fig. 5.3B). Due to the dramatic increase in the tail stiffness for extensions above 40 nm, the

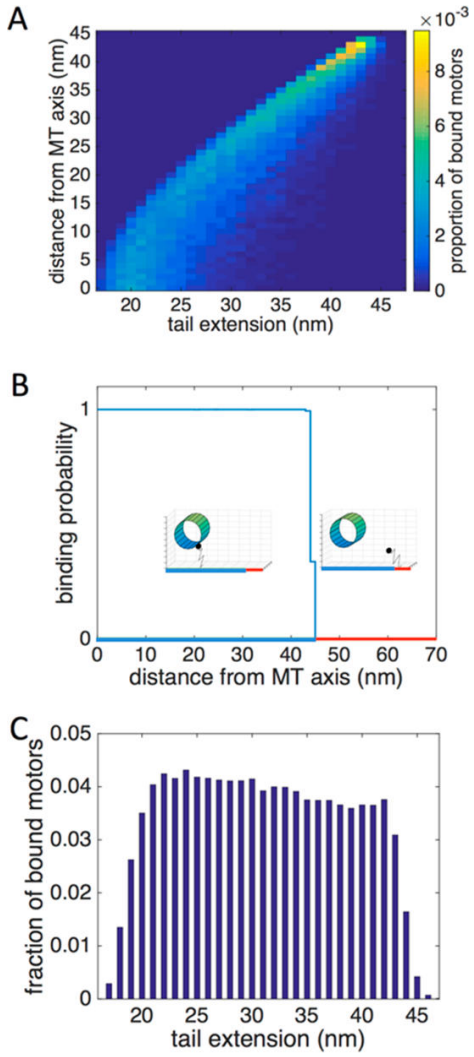

Figure 5.3: Simulation results. (A) Heat map of motor binding frequency as a function of tail extension and distance d_{attach} from the microtubule axis. (B) Binding probability of surface-adhered kinesins as a function of the distance between the microtubule axis and the kinesin attachment point. (C) Distribution of tail extension over all bound kinesins.

probability of binding within 10 s falls sharply from 100% for $d_{attach} = 43$ nm to 0% for $d_{attach} = 45$ nm. This allows us to determine a well-defined effective width w = 88 nm to compute the linear density according to Equation 5.5.

We then used the frequency distribution shown in Fig. 5.3A to compute the extension probability distribution $P(r_i) = P(\text{extension of kinesin } i = r_i)$ over all N

bound kinesins (Fig. 5.3C). The force-extension relation for the FJC model in Eq. 5.8 links the force f^k to the extension r_i. We can numerically invert this relationship to determine the force distribution $\{f_i^k\}_{i=1\ldots j}$. Numerically differentiating Eq. 5.8 with respect to f^k yields a relationship between the force and the local stiffness k_i. We then combine the force distribution with the force-stiffness relationship to compute the distribution of kinesin stiffnesses. Calculating the stiffness distributions mean and variance allows us to use Eq. 5.12 to compute a theoretical value for α_k of 0.3 (SEM $\pm$ 0.002).

Experimental measurement of the heterogeneity factor

The manual tracking of microtubule tip positions from the fluorescence microscopy images yielded microtubule position trajectories and time series of velocity fluctuations.

The measurements were conducted at saturating ATP concentrations (1 mM) for 5 different kinesin motor densities ranging from 310 to 3100 μm^{-2}. The lowest kinesin density measurements (310 μm^{-2}) were then excluded from our fit due to the length dependence of the motional diffusion coefficient at this kinesin density. The motional diffusion coefficient averaged over 20 microtubule trajectories is shown in Figure 5.4 first as a function of the motor density (Fig. 5.4A) and as a function of the calculated inverse of the linear density (Fig. 5.4B). We then used the slope of this fit and obtained a value of the heterogeneity factor α of 0.43 $\pm$ 0.3 (SEM).

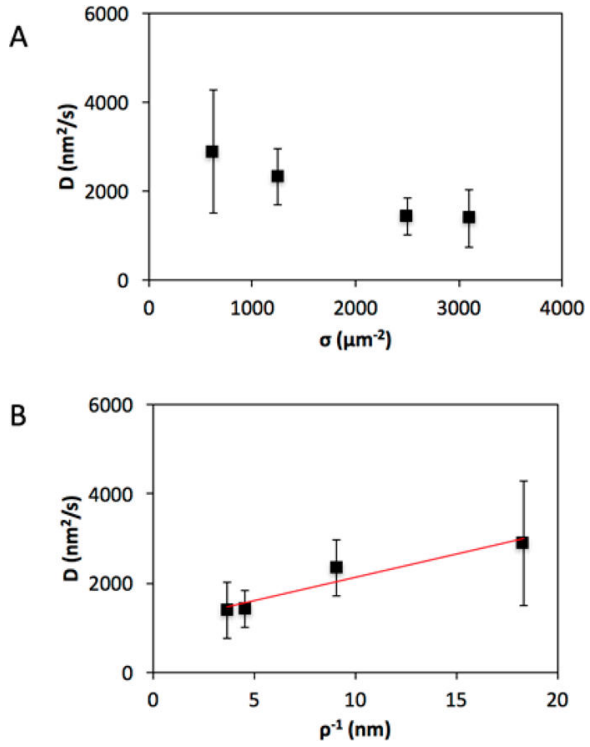

Figure 5.4: Motional diffusion coefficient as a function of motor density. (A) Motional diffusion coefficient D as a function of kinesin surface density σ. (B) Motional diffusion coefficient D as a function of the inverse of the calculated linear density, ρ^{-1}

5.4 Discussion

Through a computational model of kinesin head diffusion, we were able to estimate the distribution of kinesin tail extensions for kinesins uniformly distributed on a surface and bound to a microtubule. We then combined this distribution and the calculated anharmonic force-extension relation to quantify the theoretical heterogeneity of motor force production. Our model yields a value for the heterogeneity factor α_k of 0.3 (SEM $\leq$ 0.002).

By combining measurements of the kinesin surface density and of the motional diffusion coefficient, we were also able, for the first time, to determine this heterogeneity factor experimentally. While Sekimoto and Tawada proposed that the constant is

about one, in our assay a value of 0.43 $\pm$ 0.3 (SEM) was found, in good agreement with our theoretical value.

Under our assumptions, we have shown that the variability in the displacement of the microtubule after each motor step can be explained by the variable force contribution of each motor. In our model, each motor has an approximately constant stiffness during its attachment period to the microtubule. This stiffness increases with the distance between the microtubules axis and the kinesins attachment point on the surface. This variability leads to heterogeneity in motor force production originating in the heterogeneity in attachment geometry. The asymmetric, highly heterogeneous force production profile is shown in Figure 5.5.

One of the goals of this study was to observe deviations from the linear dependence of the motional diffusion coefficient on the motor spacing predicted by the model of Sekimoto and Tawada. We expected deviations especially at high motor densities (small spacings) where increases in the motional diffusion coefficient would indicate increasing correlations between steps. Such stepping cascades have been observed in the gliding of actin filaments on myosin [143]. Kinesin-kinesin cooperation, although relevant when two motors are under load [133], has not been observed in when a large number of kinesins propelled a microtubule whose position was tracked with nanometer resolution [127]. Our measurements, covering a wider range of motor densities also do not give any indication of a deviation from the expected evolution of the magnitude of the fluctuations. Thus the Sekimoto-Tawada model (modified to account for the motor distribution on the surface) seems to describe the fluctuations in microtubule gliding on a large number of kinesins in the absence of a large external

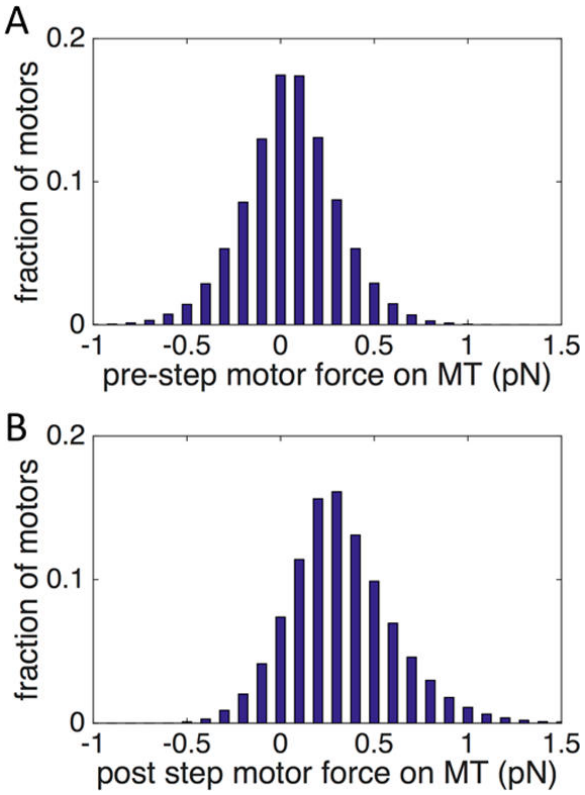

Figure 5.5: Distribution of forces exerted on the microtubule by the motors along the microtubule axis. (A) Distribution of forces exerted on the microtubule before the step. The average force exerted is 0 pN (by construction). (B) Distribution of the force one motor would exert on the microtubule immediately after it stepped. The average force exerted is 0.3 pN.

load concisely. In addition to the heterogeneity in the attachment geometry, which is present with certainty, there are potentially other sources of fluctuations, such as defective motors, or motor orientation [151]. However, theories accounting for these sources of fluctuations have not yet been formulated and therefore cannot be falsified by the present experiments.

Although the original model by Sekimoto and Tawada was formulated for kinesin/microtubule [143] and myosin/actin gliding assays, the results presented here do not translate to the actin/myosin II gliding assay, since myosin II is not processive and motor-motor coupling plays a major role [143]. Nevertheless, the impact of

heterogeneity in the attachment geometry and non-Hookean tail stiffness [152] may be worth further examination also in the actin/myosin gliding assay.

Here, we studied gliding microtubules whose movement is only opposed by viscous drag forces. Velocity fluctuations in a viscous medium will lead to a loss of efficiency on the order of our heterogeneity coefficient (see S4). If loads increase, e.g. due to the presence of cargo [153], the heterogeneity in force production will prevent homogeneous distribution of load among motors, and thereby prevent uniform loading with optimal force [154]. This situation is well understood for cargo transport in vivo, where a small number of kinesins collectively pull cargo along microtubules [133], [155], [156].

An implication of the above considerations is that obtaining a more uniform attachment geometry via a method to position the motors directly beneath the microtubule48, such as that described by Hariadi et al. [157], would aid the propulsion of the microtubule (Fig. 5.6). A reduction of the track width from 88 nm to 44 nm would reduce the heterogeneity factor tenfold. Indeed, muscle, one of the most efficient arrays of molecular motors [158], features precise alignment of these motors through the arrangement of thick and thin filaments.

These lessons are instructive for the design of future nanoactuators and molecular motor-based devices. Although individual components such as kinesin motors may be able to operate with high energy efficiency [159], the efficiency of arrays and systems may suffer if these components are not appropriately integrated.

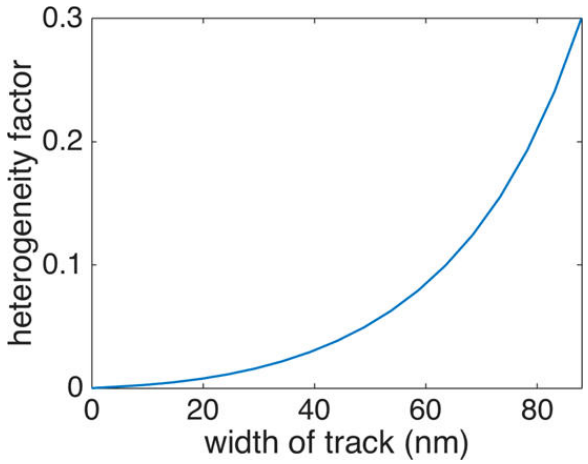

Figure 5.6: Predicted heterogeneity coefficient as a function of the width of the track the filament is gliding on. As the width of the track goes to 0, the heterogeneity factor is reduced. All track widths greater than 88 nm will display the same heterogeneity coefficient.

Appendix A

Supplementary Information for Enzyme Chemotaxis

Materials and Methods

Fluorescent Labeling of HK and Ald: Hexokinase (from Saccharomyces cerevisiae; Sigma-Aldrich) was tagged with an amine-reactive dye, Dylight 488 (ex/em: 493/518; Thermo Fisher Scientific). Hexokinase (44 M) was reacted with a threefold excess of the fluorescent probe and 10 mM mannose in 50 mM Hepes (pH 7.0) at 4C for 24 h on a rotator. Aldolase (from rabbit muscle; Sigma-Aldrich) was labeled with a thiol-reactive dye, Dylight 633 (ex/em: 638/658; Thermo Fisher Scientific). Labeling of Aldolase (75 M) was carried out with two fold excess of the fluorescent dye and 1 mM EDTA on a rotator at 4 °C for 23 h in 50 mM Hepes buffer (pH 7.4). The enzymedye conjugates were purified using a Sephadex G-25 (GE Healthcare) size exclusion column with 50 mM HEPES buffer (pH 7.4) to reduce the free-dye concentration. For FCS measurements, all enzymes were tagged with Alexa Fluor 532 dye (ex/em: 532/ 553; Thermo Fisher Scientific) by using of Alexa Fluor 532 protein labeling kit. The number of dye molecules per HK or Ald enzyme molecule was 0.4 or 0.6, respectively, as quantified using UV-vis spectroscopy. All solutions for experiments were prepared in 50 mM HEPES, pH 7.4 buffer.

Enzyme activity assays: Hexokinase activity before and after attachment of the fluorophore was measured spectrophotometrically by coupling with glucose-6-phosphate dehydrogenase (Sigma-Aldrich) and following the reduction of $NADP^+$ at 340 nm. An assay mixture, 1 mL in total volume contained 1 mM glucose, 2 mM ATP, 10 mM MgCl2, 50 mM HEPES (pH 7.4), 0.5 mM $NADP^+$, 2 units glucose-6-phosphate dehydrogenase, and 5 nM hexokinase. All assays were performed at 25 °C. The enzymatic activity was not significantly altered by the attachment of the fluorophore.

Aldolase activity before and after attachment of the fluorophore was measured spectrophotometrically by coupling with -glycerophosphate dehydrogenase/triosephosphate isomerase (Sigma-Aldrich) and following the oxidation of NADH at 340 nm. An assay mixture, 1 mL in total volume contained 2 mM fructose-1,6-disphosphate, 50 mM HEPES (pH 7.4), 0.1 mM NADH, 1.5 units -glycerophosphate dehydrogenase/triosephosphate isomerase (based on GDH units), and 50 nM aldolase. All assays were performed at 25 °C. The enzymatic activity was not significantly altered by the attachment of the fluorophore.

The difference in hexokinase activity using glucose or mannose as the substrate was measured spectrophotometrically by coupling with pyruvate kinase/lactate dehydrogenase (Sigma-Aldrich) and following the oxidation of NADH at 340 nm. An assay mixture, 1 mL in total volume contained 1 mM glucose or mannose, 2 mM ATP, 10 mM MgCl2, 3.3 mM phosphoenolpyruvate, 50 mM HEPES (pH 7.4), 0.2 mM NADH, 2 units pyruvate kinase/lactate dehydrogenase (based on PK units), and 5 nM hexokinase. All assays were performed at 25 C. The enzymatic activity

of hexokinase with mannose as the substrate was approximately half the enzymatic rate with D-glucose as the substrate under these conditions.

Progress Curve Simulation: The substrate depletion and product formation through the first four enzymes in the glycolytic cascade were simulated using Global Kinetic Explorer software (version 4.0, KinTek Corporation). The steady-state reaction scheme assumed: 1) substrate binding rates at the diffusion limit for glucose binding to hexokinase since the initial glucose concentration was sufficient to saturate the enzyme binding sites, and at k_{cat}/K_m for the subsequent enzyme reactions because the substrates were the product of the previous enzyme reaction and their concentrations did not reach the level of saturation; 2) irreversible reaction rates fixed at kcat for each enzyme since the product of each reaction would be pulled through the cascade by the presence of the downstream enzymes preventing the reverse reaction or product inhibition; and 3) that product release was not rate limiting for any individual reaction. The simulation input values were 10 mM for the starting glucose concentration; 74 nM for each starting enzyme concentration; $k_1 = 120\,\mu M^{-1}s^{-1}$ and $k_{cat} = 315\,s^{-1}$ for hexokinase; $k_{cat} = 408\,s^{-1}$ and $K_m = 700\,\mu M$ for isomerase; $k_{cat} = 113\,s^{-1}$ and $K_m = 30\,\mu M$ for PFK; and $k_{cat} = 5\,S^{-1}$ and $K_m = 60\,\mu M$ for Aldolase (all values were obtained from Sigma-Aldrich Product Information sheets: cat. # H6380 for HK; cat. # P5381 for Iso; cat. # F0137 for PFK; and cat. # A2714 for Ald). The simulation assumes that all the enzymes and glucose are combined in one reaction mixture; an enzyme concentration of 74 nM was chosen because that is the amount of hexokinase determined to migrate into a channel containing 10 mM

D-glucose (Figure A.1.).

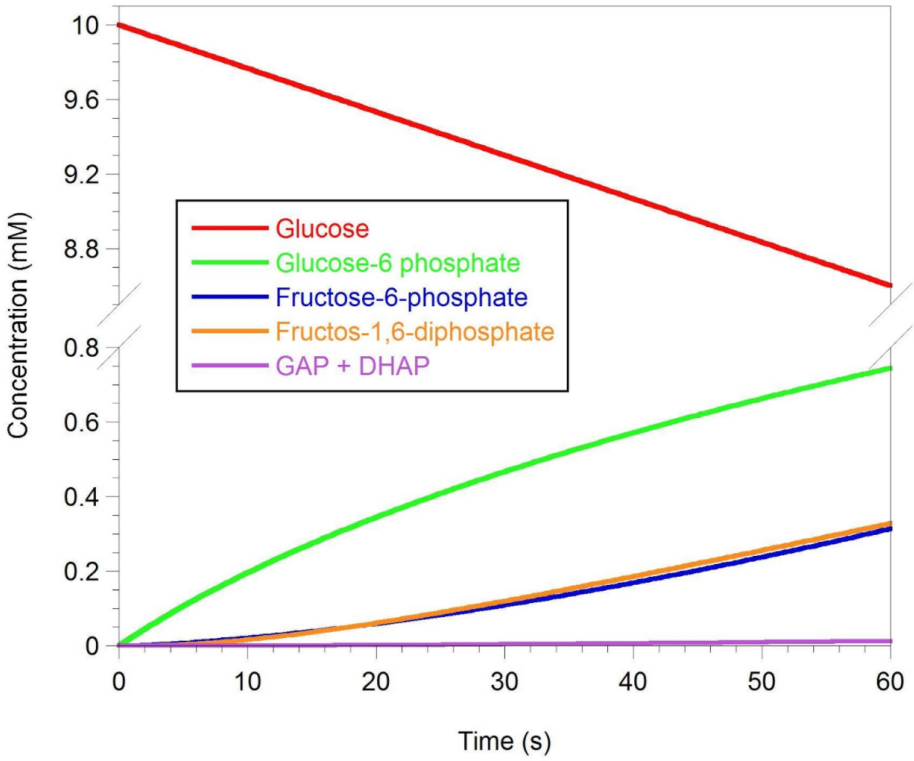

Figure A.1: The simulated substrate and product progression curves through the first four enzymes in the glycolytic cascade.

Microfluidic Device Fabrication: The microfluidic device was cast in polydimethylsiloxane (PDMS, Sylgard 184, Dow Corning) using standard soft lithography protocols. A 100 µm deep master pattern was created on a silicon wafer (Silicon Quest) using SPR-955 resist (Microposit) and deep reactive ion etching (Alcatel). The master was exposed to 1H,1H,2H,2H-perfluorooctyl-trichlorosilane (Sigma Aldrich) to minimize adhesion of PDMS during the peeling step. After the PDMS was peeled off, the inlet and outlet regions were opened by drilling, and the device was sealed to a No. 1 glass coverslip (VWR). Fluid flow through the channel was controlled by syringe

pumps (KDS 200 and 220, KD Scientific) connected by polyethylene tubing to the device.

Confocal Microscope Imaging: Confocal images were acquired using a Leica TCS SP5 laser scanning confocal inverted microscope (LSCM, Leica Microsystems) with a 10x objective (HCX PL APO CS, 0.70 NA) incorporated in it. The plane of interest (along the z-axis) for confocal imaging was chosen such that fluorescence intensity was captured from the plane that is half of the height into the channel. Videos were recorded and analyzed using Image J software. In each experiment, the mean fluorescence intensity was calculated from videos from three independent experiments. Each video is a collection of 667 images over a period of 5 min. A region of interest (ROI) was selected along the channel (as indicated by the vertical line near the end of the channel in Figure 1B), and the stack-averaged fluorescence intensity was plotted as a function of distance along the width of the channel.

Fluorescence Correlation Spectroscopy: Spectroscopy measurements were performed on a custom-built microscope- based optical setup. Briey, a PicoTRAIN laser (High-Q Laser) delivered 5.4 ps pulses of 532 nm light at 80 MHz frequency. This light was guided through a fiber optic cable, expanded and directed through an IX-71 microscope (Olympus) with an Olympus 60x/1.2-NA water-immersion objective. Emitted fluorescent light from the sample was passed through a dichroic beam splitter (Z520RDC-SP-POL, Chroma Technology) and focused onto a 50 µm, 0.22-NA optical fiber (Thorlabs), which acted as a confocal pinhole. The signal from the photomul-

tiplier tube was routed to a preamplier (HFAC-26) and then to a time-correlated single-photon counting (TCSPC) board (SPC-630, Becker and Hickl). The sample was positioned with a high-resolution 3-D piezoelectric stage (NanoView, Mad City Laboratories). Fluorescent molecules moving into and out of the diraction-limited observation volume induce bursts in uorescence collected in first-in, first-out mode by the TCSPC board, which was incorporated in the instrument. Fluctuations in fluorescence intensity from the diusion of molecules were auto-correlated and fit by a single component 3D model to determine the diffusion coefficients of individual species. Contributions to the autocorrelation curve from fluctuations in molecular fluorescence intensity due to fast processes such as triplet state excitation were minimal. Nevertheless, when the shape of the autocorrelation curve indicated the need to include the triplet state in the fit and the alternative Eq. A.2 was used [160].

FCS measurements were performed with 30 W excitation power, and the optical system (r and w of confocal volume) was calibrated before each experiment using free 50 nm polystyrene fluorescent beads in deionized water. Autocorrelation curves were t to eq. A.1 or A.2 using the Levenberg-Marquardt nonlinear least- squares regression algorithm with Origin software to determine N, T, and τ_D.

$$G(\tau) = \frac{1}{N}[1 + (\frac{1}{w})^{-1}][1 + (\frac{1}{w})^2(\frac{\tau}{\tau_D})]^{-1/2} \tag{A.1}$$

$$G(\tau) = (1 + \frac{T}{1-T}e^{-\tau/\tau_T})\frac{1}{N}[1 + (\frac{1}{w})^{-1}][1 + (\frac{1}{w})^2(\frac{\tau}{\tau_D})]^{-1/2} \tag{A.2}$$

where N is the number of molecules in the confocal volume, w is the structure

factor (radius, r, of the confocal volume over its half height), τ is the correlation time, τ_D is the characteristic diffusion time (where $\tau_D = r^2/4D$ (D is the diffusion coefficient), and T is the triplet fraction, τ_T .

Statistical Significance Analysis of FCS data: Diffusion coefficients of each enzyme for each substrate concentration were entered into a table in Graphpad Prism software. Means and standard deviations were calculated. After this, an analysis of variance (ANOVA) test was performed followed by Tukey's multiple comparisons of means. For HK in Figure 4.2A, all means except for 1 M were statistically significantly greater than the values at 0 M substrate. For Ald (Figure 4.2B) all values were statistically significantly greater than the value at 0 M.

Chemotactic Co-localization of Hexokinase and Aldolase: We observed the co-localization of HK and Ald (metabolon formation) in a sealed hybridization chamber starting with a uniform distribution of all the four enzymes in the cascade, as well as the substrates for HK. 200 nM HK labeled with amine-reactive (ex/em: 493/518) Dylight dye, 200 nM Iso, 200 nM FPK, 200 nM Ald conjugated with thiol-reactive (ex/em: 638/658) Dylight dye, 10 mM ATP, 20 mM MgCl2, and 10 mM D-glucose in 20% w/v 70 M Ficoll was mixed and injected into a hybridization chamber, which was sealed on the surface of a glass slide.

Simplified Model for Cross-Diffusion Illustrating the Role of Catalysis: To show that catalysis plays a crucial part in the cross-diffusion process, we modeled a simple, generic enzyme reaction as follows:

$$\mathrm{E} + \mathrm{S} \underset{\mathrm{k_{off}} = 60\,\mathrm{s}^{-1}}{\overset{\mathrm{k_{on}} = 2\,10^{\,6}\,\mathrm{M}^{-1}\mathrm{s}^{-1}}{\rightleftharpoons}} \mathrm{ES}$$

$$\mathrm{ES} \xrightarrow{\mathrm{k_{cat}} = 8000\,\mathrm{s}^{-1}} \mathrm{E} + \mathrm{P}$$

We then modeled a cross diffusion experiment with initial conditions: [E] = 200 nM in all channels, [S] = 50 mM in central channel only. The differential equations were modeled as described in the supplementary information section Computational Modeling of Cross-Diffusion above. We then compared the focusing amplitudes with catalysis ($k_{cat} = 8000\,\mathrm{s}^{-1}$) and without catalysis (k_{cat}=0). The results are plotted in Fig. A.2A: there is significant enzyme focusing towards the central channel in the case with catalysis, and almost no focusing in the case without catalysis. In the case with catalysis, ES is consumed leading to several orders of magnitude more forward binding events than in the case without catalysis (Figure A.2B). Indeed, in the case without catalysis, the equilibrium shifts very rapidly towards the enzyme complex ES and no more forward binding events are observed (Figure A.2C and A.2D). After approximately twenty seconds, all the substrate is consumed, but enzyme continues to spread diffusively, thus explaining the peak in Figure A.2A and the flattening in Figure A.2B-D. This shows that (1) the cross-diffusion phenomenon is dependent on the number of forward binding events in the region where the substrate gradient has been established and (2) the turnover induced by enzyme catalysis allows for orders of magnitude more binding events to take place. Therefore, the catalytic step is crucial in the observation of enzyme focusing.

Statistical analysis of experimental microfluidic results: Catalysis induced focusing (Fig. 4.3) the data points were fitted with a smoothing function using the R

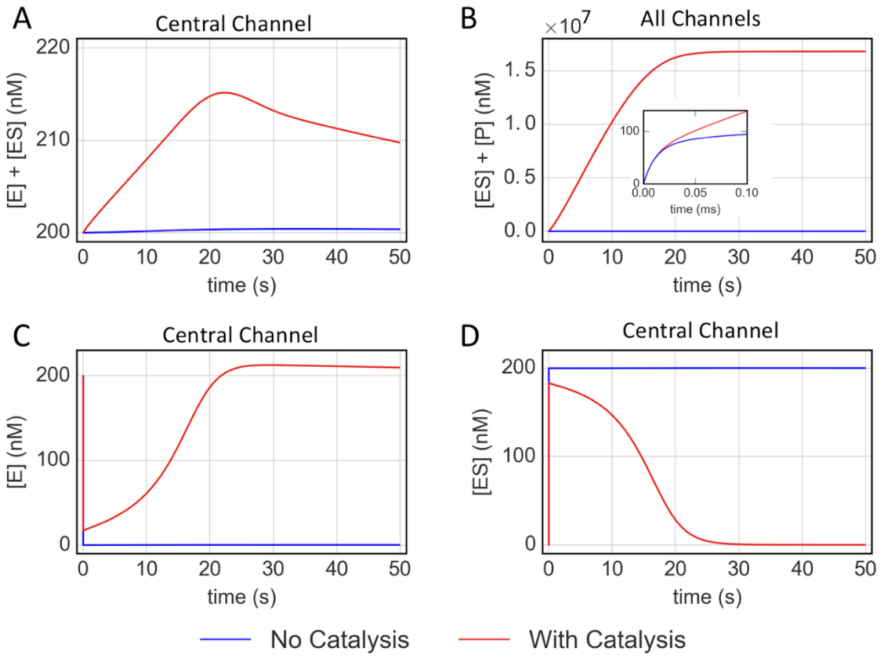

Figure A.2: Modeling results for the simplified enzyme cross diffusion. (A) Total enzyme concentration in central channel. With catalysis we observe significant enzyme focusing. Without catalysis, focusing is not noticeable. (B) The sum of product concentration and enzyme complex concentration is used to estimate the number of forward binding events. The number of binding events in the case with catalysis is several orders of magnitude greater than in the case with no catalysis due to turnover. Insert: close up of data at beginning of the reaction. (C) Enzyme concentration in central channel over time. Without catalysis, the enzyme-complex equilibrium shifts almost immediately towards the complex ES, and [E] drops to zero. With catalysis, ES is turned over and the equilibrium shifts towards E. (D) Enzyme-Substrate complex concentration in the central channel over time.

programming language, as no closed form formula for the focusing curves exists. The smoothing function chosen is a second degree polynomial, and the weighing chosen is the default tricubic weighing for polynomial fitting. The error bars in Figure 4.6 are 95% confidence intervals derived from the three replications of the experimental process. The statistical significance was derived from a pairwise t-test with Holm adjustment.

Bibliography

[1] A. Engel and C. Van den Broeck, *Statistical mechanics of learning*. Cambridge University Press, 2001.

[2] J. M. R. Parrondo, J. M. Horowitz, and T. Sagawa, "Thermodynamics of information," en, *Nat. Phys.*, vol. 11, no. 2, pp. 131–139, 2015.

[3] V. Kecman, *Learning and soft computing: support vector machines, neural networks, and fuzzy logic models*. MIT Press, 2001.

[4] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer Series in Statistics, Springer, Berlin, 2001, vol. 1.

[5] T. L. H. Watkin, A. Rau, and M. Biehl, "The statistical mechanics of learning a rule," *Rev. Mod. Phys.*, vol. 65, no. 2, p. 499, 1993.

[6] M. Biehl and N. Caticha, "Statistical mechanics of on–line learning and generalization," in *The Handbook of Brain Theory and Neural Networks*, MIT Press, 2003.

[7] L. Zdeborova and F. Krzakala, "Statistical physics of inference: thresholds and algorithms," *arXiv:1511.02476*, 2015.

[8] C. Jarzynski, "Nonequilibrium equality for free energy differences," *Phys. Rev. Lett.*, vol. 78, no. 14, pp. 2690–2693, 1997.

[9] G. E. Crooks, "Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences," *Phys. Rev. E*, vol. 60, no. 3, pp. 2721–2726, 1999.

[10] B. Altaner, "Foundations of Stochastic Thermodynamics," *arXiv:1410.3983 [cond-mat]*, 2014.

[11] S. Ito and T. Sagawa, "Information flow and entropy production on Bayesian networks," *arXiv:1506.08519 [cond-mat]*, 2015.

[12] S.-I. Amari and H. Nagaoka, *Methods of information geometry*. American Mathematical Soc., 2007, vol. 191.

[13] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient Langevin dynamics," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 681–688.

[14] H. Zhao, J. B. French, Y. Fang, and S. J. Benkovic, "The purinosome, a multi-protein complex involved in the de novo biosynthesis of purines in humans," en, *Chem. Commun.*, vol. 49, no. 40, p. 4444, 2013.

[15] F. Wu and S. Minteer, "Krebs cycle metabolon: structural evidence of substrate channeling revealed by cross-linking and mass spectrometry," en, *Angew. Chem. Int. Ed Engl.*, vol. 54, no. 6, pp. 1851–1854, 2015.

[16] M. Castellana, M. Z. Wilson, Y. Xu, P. Joshi, I. M. Cristea, J. D. Rabinowitz, Z. Gitai, and N. S. Wingreen, "Enzyme clustering accelerates processing of intermediates through metabolic channeling," *Nat. Biotechnol.*, vol. 32, no. 10, pp. 1011–1018, 2014.

[17] S. An, R. Kumar, E. D. Sheets, and S. J. Benkovic, "Reversible Compartmentalization of de Novo Purine Biosynthetic Complexes in Living Cells," en, *Science*, vol. 320, no. 5872, pp. 103–106, 2008.

[18] K. Jorgensen, A. V. Rasmussen, M. Morant, A. H. Nielsen, N. Bjarnholt, M. Zagrobelny, S. Bak, and B. L. Moller, "Metabolon formation and metabolic channeling in the biosynthesis of plant natural products," *Curr. Opin. Plant Biol.*, vol. 8, no. 3, pp. 280–291, 2005.

[19] C. Riedel, R. Gabizon, C. A. M. Wilson, K. Hamadani, K. Tsekouras, S. Marqusee, S. Press, and C. Bustamante, "The heat released during catalytic turnover enhances the diffusion of an enzyme," *Nature*, vol. 517, no. 7533, pp. 227–230, 2014.

[20] S. Sengupta, K. K. Dey, H. S. Muddana, T. Tabouillot, M. E. Ibele, P. J. Butler, and A. Sen, "Enzyme Molecules as Nanomotors," *J. Am. Chem. Soc.*, vol. 135, no. 4, pp. 1406–1414, 2013.

[21] H. S. Muddana, S. Sengupta, T. E. Mallouk, A. Sen, and P. J. Butler, "Substrate catalysis enhances single-enzyme diffusion," *J. Am. Chem. Soc.*, vol. 132, no. 7, pp. 2110–2111, 2010.

[22] P. J. Butler, K. K. Dey, and A. Sen, "Impulsive enzymes: a new force in mechanobiology," *Cell. Mol. Bioeng.*, pp. 106–118, 2015.

[23] J. B. French, S. A. Jones, H. Deng, A. M. Pedley, D. Kim, C. Y. Chan, H. Hu, R. J. Pugh, H. Zhao, Y. Zhang, T. J. Huang, Y. Fang, X. Zhuang, and S. J. Benkovic, "Spatial colocalization and functional link of purinosomes with mitochondria," en, *Science*, vol. 351, no. 6274, pp. 733–737, 2016.

[24] V. K. Vanag and I. R. Epstein, "Cross-diffusion and pattern formation in reaction–diffusion systems," en, *Phys. Chem. Chem. Phys.*, vol. 11, no. 6, pp. 897–912, 2009.

[25] O. Annunziata, A. Vergara, L. Paduano, R. Sartorio, D. G. Miller, and J. G. Albright, "Quaternary diffusion coefficients in a protein-polymer-salt-water system determined by Rayleigh interferometry," *J. Phys. Chem. B*, vol. 113, no. 40, pp. 13 446–13 453, 2009.

[26] A. Vergara, L. Paduano, and R. Sartorio, "Mechanism of protein-poly(ethylene glycol) interaction from a diffusive point of view," *Macromolecules*, vol. 35, no. 4, pp. 1389–1398, 2002.

[27] L. Paduano, R. Sartorio, and V. Vitagliano, "Diffusion coefficients of the ternary system α-cyclodextrin-sodium benzenesulfonate-water at 25 C: the effect of chemical equilibrium and complex formation on the diffusion coefficients of a ternary system," *J. Phys. Chem. B*, vol. 102, no. 25, pp. 5023–5028, 1998.

[28] R. Fu, D. Sutcliffe, H. Zhao, X. Huang, D. J. Schretlen, S. Benkovic, and H. A. Jinnah, "Clinical severity in Lesch–Nyhan disease: the role of residual enzyme and compensatory pathways," *Mol. Genet. Metab.*, vol. 114, no. 1, pp. 55–61, 2015.

[29] D. Chandler, *Introduction to Modern Statistical Mechanics*, en. Oxford University Press, 1987.

[30] A. Berut, A. Arakelyan, A. Petrosyan, S. Ciliberto, R. Dillenschneider, and E. Lutz, "Experimental verification of Landauer's principle linking information and thermodynamics," *Nature*, vol. 483, no. 7388, pp. 187–189, 2012.

[31] D. Collin, F. Ritort, C. Jarzynski, S. B. Smith, I. Tinoco, and C. Bustamante, "Verification of the Crooks fluctuation theorem and recovery of RNA folding free energies," *Nature*, vol. 437, no. 7056, pp. 231–234, 2005.

[32] Z. Lu, D. Mandal, and C. Jarzynski, "Engineering Maxwell's demon," en, *Phys. Today*, vol. 67, no. 8, pp. 60–61, 2014.

[33] S. Toyabe, T. Sagawa, M. Ueda, E. Muneyuki, and M. Sano, "Experimental demonstration of information-to-energy conversion and validation of the generalized Jarzynski equality," *Nat. Phys.*, vol. 6, no. 12, pp. 988–992, 2010.

[34] E. A. Calzetta, "Kinesin and the Crooks fluctuation theorem," en, *Eur. Phys. J. B*, vol. 68, no. 4, pp. 601–605, 2009.

[35] J. L. England, "Statistical physics of self-replication," *J. Chem. Phys.*, vol. 139, no. 12, p. 121 923, 2013.

[36] G. E. Crooks, "Measuring thermodynamic length," *Phys. Rev. Lett.*, vol. 99, no. 10, p. 100 602, 2007.

[37] T. Sagawa and M. Ueda, "Generalized Jarzynski equality under nonequilibrium feedback control," en, *Phys. Rev. Lett.*, vol. 104, no. 9, 2010.

[38] S. Deffner and E. Lutz, "Information free energy for nonequilibrium states," *arXiv:1201.3888 [cond-mat]*, 2012.

[39] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 2012.

[40] B. Altaner, "Nonequilibrium thermodynamics and information theory: Foundations and relaxing dynamics," *arXiv:1702.07906 [cond-mat]*, 2017.

[41] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference*, en. Springer Science & Business Media, 2013.

[42] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 988–999, 1999.

[43] H. Robbins and S. Monro, "A stochastic approximation method," en, *Ann. Math. Stat.*, vol. 22, no. 3, pp. 400–407, 1951.

[44] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis*. CRC press Boca Raton, FL, 2014, vol. 2.

[45] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *J. Chem. Phys.*, vol. 21, no. 6, pp. 1087–1092, 1953.

[46] R. M. Neal, "MCMC using Hamiltonian dynamics," *arXiv:1206.1901 [physics, stat]*, 2012.

[47] M. Betancourt, "A conceptual introduction to Hamiltonian Monte Carlo," 2017. arXiv: `1701.02434 [stat.ME]`.

[48] C. Van den Broeck, "Stochastic thermodynamics: A brief introduction," in *Proceedings of the International School of Physics 'Enrico Fermi*, vol. 184, 2013, pp. 155–93.

[49] R. Linsker, "Self-organization in a perceptual network," *Computer*, vol. 21, no. 3, pp. 105–117, 1988.

[50] M. Mezard and A. Montanari, *Information, physics, and computation*. Oxford University Press, 2009.

[51] M. Advani, S. Lahiri, and S. Ganguli, "Statistical mechanics of complex neural systems and high dimensional data," *J. Stat. Mech: Theory Exp.*, vol. 2013, no. 03, pp. 3014–3080, 2013.

[52] M. Mezard, *Spin glass theory and beyond*. Teaneck, NJ, USA: World Scientific, 1987.

[53] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*, en. MIT Press, 2009.

[54] C. Moore, "The computer science and physics of community detection: landscapes, phase transitions, and hardness," *arXiv:1702.00467 [cond-mat, physics:physics]*, 2017.

[55] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: a review for statisticians," *arXiv:1601.00670 [cs, stat]*, 2016.

[56] C. Li, C. Chen, D. E. Carlson, and L. Carin, "Preconditioned stochastic gradient Langevin dynamics for deep neural networks," in *AAAI*, vol. 2, 2016, p. 4.

[57] Girolami Mark and Calderhead Ben, "Riemann manifold Langevin and Hamiltonian Monte Carlo methods," *J. R. Stat. Soc. Series B Stat. Methodol.*, vol. 73, no. 2, pp. 123–214, 2011.

[58] J. Martens and R. Grosse, "Optimizing neural networks with Kronecker-factored approximate curvature," in *International Conference on Machine Learning*, 2015, pp. 2408–2417.

[59] R. Grosse and J. Martens, "A Kronecker-factored approximate Fisher matrix for convolution layers," *arXiv:1602.01407 [cs, stat]*, 2016.

[60] Z. Nado, J. Snoek, R. Grosse, D. Duvenaud, B. Xu, and J. Martens, "Stochastic gradient Langevin dynamics that exploit neural network structure," 2018.

[61] S. Mandt, M. D. Hoffman, and D. M. Blei, "Stochastic gradient descent as approximate Bayesian inference," *arXiv:1704.04289 [cs, stat]*, 2017.

[62] R. M. Neal, "Bayesian learning via stochastic dynamics," in *Advances in Neural Information Processing Systems*, 1993, pp. 475–482.

[63] D. J. C. MacKay, "A practical Bayesian framework for backpropagation networks," *Neural Comput.*, vol. 4, no. 3, pp. 448–472, 1992.

[64] J. Denker and Y. Lecun, "Transforming neural-net output levels to probability distributions," in *Advances in Neural Information Processing Systems 3*, Morgan Kaufmann, 1991, pp. 853–859.

[65] S.-I. Amari, "Natural gradient works efficiently in learning," *Neural Comput.*, vol. 10, no. 2, pp. 251–276, 1998.

[66] S. Patterson and Y. W. Teh, "Stochastic gradient Riemannian Langevin dynamics on the probability simplex," in *Advances in Neural Information Processing Systems*, 2013, pp. 3102–3110.

[67] G. Marceau-Caron and Y. Ollivier, "Natural Langevin dynamics for neural networks," *arXiv:1712.01076 [cs, stat]*, 2017.

[68] J. Martens and R. Grosse, "Optimizing neural networks with Kronecker-factored approximate curvature," 2015.

[69] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, p. 533, 1986.

[70] Y. W. Teh, A. H. Thiery, and S. J. Vollmer, "Consistency and fluctuations for stochastic gradient Langevin dynamics," *J. Mach. Learn. Res.*, vol. 17, pp. 1–33, 2016.

[71] G. Jones, M. Haran, B. Caffo, and R. Neath, "Fixed-width output analysis for Markov chain Monte Carlo," *arXiv:math/0601446*, 2006.

[72] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014. arXiv: `1412.6572 [stat.ML]`.

[73] Y. Bulatov, *notMNIST dataset*, `http://yaroslavvb.blogspot.com/2011/09/notmnist-dataset.html`, Accessed: 2018-4-24.

[74] Y. LeCun, C. Cortes, and C. J. Burges, "MNIST handwritten digit database," *AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist*, vol. 2, 2010.

[75] A. Gelman and D. B. Rubin, "Inference from iterative simulation using multiple sequences," EN, *Stat. Sci.*, vol. 7, no. 4, pp. 457–472, 1992.

[76] D. Vats, J. M. Flegal, and G. L. Jones, "Multivariate output analysis for Markov chain Monte Carlo," *arXiv:1512.07713 [math, stat]*, 2015.

[77] M. Betancourt, "The fundamental incompatibility of scalable Hamiltonian Monte Carlo and naive data subsampling," en, in *International Conference on Machine Learning*, jmlr.org, 2015, pp. 533–540.

[78] S. J. Vollmer, K. C. Zygalakis, Teh, and Y. Whye, "(Non-) asymptotic properties of Stochastic Gradient Langevin Dynamics," 2015. arXiv: `1501.00438 [stat.ME]`.

[79] A. Rawat, M. Wistuba, and M.-I. Nicolae, "Adversarial phenomenon in the eyes of Bayesian deep learning," 2017. arXiv: `1711.08244 [stat.ML]`.

[80] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: representing model uncertainty in deep learning," 2015. arXiv: `1506.02142 [stat.ML]`.

[81] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural networks," 2015. arXiv: `1505.05424 [stat.ML]`.

[82] C. Louizos and M. Welling, "Structured and efficient variational deep learning with matrix Gaussian posteriors," 2016. arXiv: `1603.04733 [stat.ML]`.

[83] J. M. Hernandez-Lobato and R. P. Adams, "Probabilistic backpropagation for scalable learning of Bayesian neural networks," 2015. arXiv: `1502.05336 [stat.ML]`.

[84] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" *arXiv:1703.04977 [cs]*, 2017.

[85] E. W. Miles, S. Rhee, and D. R. Davies, "The molecular basis of substrate channeling," en, *J. Biol. Chem.*, vol. 274, no. 18, pp. 12 193–12 196, 1999.

[86] I. Wheeldon, S. D. Minteer, S. Banta, S. C. Barton, P. Atanassov, and M. Sigman, "Substrate channelling as an approach to cascade reactions," *Nat. Chem.*, vol. 8, p. 299, 2016.

[87] J.-L. Lin, L. Palomec, and I. Wheeldon, "Design and analysis of enhanced catalysis in scaffolded multienzyme cascade reactions," *ACS Catal.*, vol. 4, no. 2, pp. 505–511, 2014.

[88] Y. Yamada, C.-K. Tsung, W. Huang, Z. Huo, S. E. Habas, T. Soejima, C. E. Aliaga, G. A. Somorjai, and P. Yang, "Nanocrystal bilayer for tandem catalysis," *Nat. Chem.*, vol. 3, p. 372, 2011.

[89] M. Zhao, K. Deng, L. He, Y. Liu, G. Li, H. Zhao, and Z. Tang, "Core–shell palladium nanoparticle@metal–organic frameworks as multifunctional catalysts for cascade reactions," *J. Am. Chem. Soc.*, vol. 136, no. 5, pp. 1738–1741, 2014.

[90] Y.-H. P. Zhang, "Substrate channeling and enzyme complexes for biotechnological applications," en, *Biotechnol. Adv.*, vol. 29, no. 6, pp. 715–725, 2011.

[91] H. Nishi, K. Hashimoto, and A. R. Panchenko, "Phosphorylation in protein-protein binding: effect on stability and function," en, *Structure*, vol. 19, no. 12, pp. 1807–1815, 2011.

[92] C. Lindbladh, M. Rault, C. Hagglund, W. C. Small, K. Mosbach, L. Buelow, C. Evans, and P. A. Srere, "Preparation and kinetic characterization of a fusion protein of yeast mitochondrial citrate synthase and malate dehydrogenase," *Biochemistry*, vol. 33, no. 39, pp. 11 692–11 698, 1994.

[93] B. Winkel-Shirley, "Evidence for enzyme complexes in the phenylpropanoid and flavonoid pathways," *Physiol. Plant.*, vol. 107, no. 1, pp. 142–149, 1999.

[94] J. W. A. Graham, T. C. R. Williams, M. Morgan, A. R. Fernie, R. G. Ratcliffe, and L. J. Sweetlove, "Glycolytic enzymes associate dynamically with mitochondria in response to respiratory demand and support substrate channeling," *Plant Cell*, vol. 19, no. 11, pp. 3723–3738, 2007.

[95] M. E. Campanella, H. Chu, and P. S. Low, "Assembly and regulation of a glycolytic enzyme complex on the human erythrocyte membrane," en, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 7, pp. 2402–2407, 2005.

[96] S. Sengupta, K. K. Dey, H. S. Muddana, T. Tabouillot, M. E. Ibele, P. J. Butler, and A. Sen, "Enzyme molecules as nanomotors," *J. Am. Chem. Soc.*, vol. 135, no. 4, pp. 1406–1414, 2013.

[97] S. Sengupta, M. M. Spiering, K. K. Dey, W. Duan, D. Patra, P. J. Butler, R. D. Astumian, S. J. Benkovic, and A. Sen, "DNA polymerase as a molecular motor and pump," *ACS Nano*, vol. 8, no. 3, pp. 2410–2418, 2014.

[98] H. Yu, K. Jo, K. L. Kounovsky, J. J. d. Pablo, and D. C. Schwartz, "Molecular propulsion: Chemical sensing and chemotaxis of DNA driven by RNA polymerase," *J. Am. Chem. Soc.*, vol. 131, no. 16, pp. 5722–5723, 2009.

[99] C. L. Kohnhorst, M. Kyoung, M. Jeon, D. L. Schmitt, E. L. Kennedy, J. Ramirez, S. M. Bracey, B. T. Luu, S. J. Russell, and S. An, "Identification of a multienzyme complex for glucose metabolism in living cells," *J. Biol. Chem.*, 2017.

[100] C. Riedel, R. Gabizon, C. A. M. Wilson, K. Hamadani, K. Tsekouras, S. Marqusee, S. Press, and C. Bustamante, "The heat released during catalytic turnover enhances the diffusion of an enzyme," *Nature*, vol. 517, p. 227, 2014.

[101] P. Illien, X. Zhao, K. K. Dey, P. J. Butler, A. Sen, and R. Golestanian, "Exothermicity Is Not a Necessary Condition for Enhanced Diffusion of Enzymes," *Nano Lett.*, vol. 17, no. 7, pp. 4415–4420, 2017.

[102] T. C. Myers, K. Nakamura, and J. W. Flesher, "Phosphonic acid analogs of nucleoside phosphates. I. The synthesis of 5-adenylyl methylenediphosphonate, a phosphonic acid analog of ATP," *J. Am. Chem. Soc.*, vol. 85, no. 20, pp. 3292–3295, 1963.

[103] L. Paduano, R. Sartorio, G. D'Errico, and V. Vitagliano, "Mutual diffusion in aqueous solution of ethylene glycol oligomers at 25 C," en, *J. Chem. Soc. Faraday Trans.*, vol. 94, no. 17, pp. 2571–2576, 1998.

[104] J. M. Schurr, B. S. Fujimoto, L. Huynh, and D. T. Chiu, "A theory of macromolecular chemotaxis," *J. Phys. Chem. B*, vol. 117, no. 25, pp. 7626–7652, 2013.

[105] K. Wilkinson and I. A. Rose, "Isotope trapping studies of yeast hexokinase during steady state catalysis," 1979.

[106] A. Bermingham, J. R. Bottomley, W. U. Primrose, and J. P. Derrick, "Equilibrium and kinetic studies of substrate binding to 6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase from Escherichia coli," en, *J. Biol. Chem.*, vol. 275, no. 24, pp. 17 962–17 967, 2000.

[107] J. G. Kirkwood and F. P. Buff, "The statistical mechanical theory of solutions. I," *J. Chem. Phys.*, vol. 19, no. 6, pp. 774–777, 1951.

[108] K. E. Newman, "Kirkwood–Buff solution theory: derivation and applications," en, *Chem. Soc. Rev.*, vol. 23, no. 1, pp. 31–40, 1994.

[109] E. Jones, T. Oliphant, and P. Peterson, *SciPy: Open source scientific tools for Python*, 2001.

[110] W. E. Schiesser and G. W. Griffiths, *A compendium of partial differential equation models: method of lines analysis with Matlab*. Cambridge University Press, 2009.

[111] F. Wu, L. N. Pelster, and S. D. Minteer, "Krebs cycle metabolon formation: metabolite concentration gradient enhanced compartmentation of sequential enzymes," en, *Chem. Commun.*, vol. 51, no. 7, pp. 1244–1247, 2015.

[112] D. Allan, T. Caswell, N. Keim, and C. van der Wel, *trackpy: Trackpy v0.3.2*, 2016.

[113] J. W. Driver, A. R. Rogers, D. K. Jamison, R. K. Das, A. B. Kolomeisky, and M. R. Diehl, "Coupling between motor proteins determines dynamic behaviors of motor protein assemblies," en, *Phys. Chem. Chem. Phys.*, vol. 12, no. 35, pp. 10 398–10 405, 2010.

[114] T. Yanagida, M. Nakase, K. Nishiyama, and F. Oosawa, "Direct observation of motion of single F-actin filaments in the presence of myosin," *Nature*, vol. 307, p. 58, 1984.

[115] F. Jlicher and J. Prost, "Cooperative molecular motors," en, *Phys. Rev. Lett.*, vol. 75, no. 13, pp. 2618–2621, 1995.

[116] F. Julicher, A. Ajdari, and J. Prost, "Modeling molecular motors," *Rev. Mod. Phys.*, vol. 69, no. 4, pp. 1269–1281, 1997.

[117] F. Julicher and J. Prost, "Molecular motors: From individual to collective behavior," *Progr. Theoret. Phys.*, vol. 130, pp. 9–16, 1998.

[118] K. Tawada and K. Sekimoto, "A physical model of ATP-induced actin-myosin movement in vitro," en, *Biophys. J.*, vol. 59, no. 2, pp. 343–356, 1991.

[119] K. Tawada and K. Sekimoto, "Protein friction exerted by motor enzymes through a weak-binding interaction," *J. Theor. Biol.*, vol. 150, no. 2, pp. 193–200, 1991.

[120] K. Sekimoto and K. Tawada, "Extended time correlation of in vitro motility by motor protein," en, *Phys. Rev. Lett.*, vol. 75, no. 1, pp. 180–183, 1995.

[121] ——, "Fluctuations in sliding motion generated by independent and random actions of protein motors," *Biophys. Chem.*, vol. 89, no. 1, pp. 95–99, 2001.

[122] T. Duke, "Cooperativity of myosin molecules through strain-dependent chemistry," *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences*, vol. 355, no. 1396, pp. 529–538, 2000.

[123] T. A. J. Duke, "Molecular model of muscle contraction," en, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 96, no. 6, pp. 2770–2775, 1999.

[124] Y. Imafuku, N. Mitarai, K. Tawada, and H. Nakanishi, "Anomalous fluctuations in sliding motion of cytoskeletal filaments driven by molecular motors: model simulations," en, *J. Phys. Chem. B*, vol. 112, no. 5, pp. 1487–1493, 2008.

[125] Y. Imafuku, Y. Y. Toyoshima, and K. Tawada, "Fluctuation in the microtubule sliding movement driven by kinesin in vitro," en, *Biophys. J.*, vol. 70, no. 2, pp. 878–886, 1996.

[126] T. Nitta and H. Hess, "Dispersion in active transport by kinesin-powered molecular shuttles," en, *Nano Lett.*, vol. 5, no. 7, pp. 1337–1342, 2005.

[127] C. Leduc, F. Ruhnow, J. Howard, and S. Diez, "Detection of fractional steps in cargo movement by the collective operation of kinesin-1 motors," en, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 104, no. 26, pp. 10 847–10 852, 2007.

[128] J. Howard *et al.*, "Mechanics of motor proteins and the cytoskeleton," 2001.

[129] S. M. Block, "Kinesin motor mechanics: Binding, stepping, tracking, gating, and limping," en, *Biophys. J.*, vol. 92, no. 9, pp. 2986–2995, 2007.

[130] K. Svoboda, P. P. Mitra, and S. M. Block, "Fluctuation analysis of motor protein movement and single enzyme kinetics," en, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 91, no. 25, pp. 11 782–11 786, 1994.

[131] A. Agarwal, E. Luria, X. Deng, J. Lahann, and H. Hess, "Landing rate measurements to detect fibrinogen adsorption to non-fouling surfaces," *Cell. Mol. Bioeng.*, vol. 5, no. 3, pp. 320–326, 2012.

[132] J. Howard, A. J. Hudspeth, and R. D. Vale, "Movement of microtubules by single kinesin molecules," *Nature*, vol. 342, p. 154, 1989.

[133] F. Berger, C. Keller, R. Lipowsky, and S. Klumpp, "Elastic coupling effects in cooperative transport by a pair of molecular motors," en, *Cell. Mol. Bioeng.*, vol. 6, no. 1, pp. 48–64, 2012.

[134] G. Arpa, S. Shastry, W. O. Hancock, and E. Tzel, "Transport by populations of fast and slow Kinesins uncovers novel family-dependent motor characteristics

important for in vivo function," en, *Biophys. J.*, vol. 107, no. 8, pp. 1896–1904, 2014.

[135] Y. Ishigure and T. Nitta, "Simulating an actomyosin in vitro motility assay: Toward the rational design of actomyosin-based microtransporters," en, *IEEE Trans. Nanobioscience*, vol. 14, no. 6, pp. 641–648, 2015.

[136] P. Egan, J. Moore, C. Schunn, J. Cagan, and P. LeDuc, "Emergent systems energy laws for predicting Myosin ensemble processivity," *PLoS Comput. Biol.*, vol. 11, no. 4, e1004177, 2015.

[137] P. S. Grassia, E. J. Hinch, and L. C. Nitsche, "Computer simulations of Brownian motion of complex systems," *J. Fluid Mech.*, vol. 282, pp. 373–403, 1995.

[138] P. Katira, A. Agarwal, T. Fischer, H.-Y. Chen, X. Jiang, J. Lahann, and H. Hess, "Quantifying the performance of protein-resisting surfaces at ultra-low protein coverages using Kinesin motor proteins as probes," *Adv. Mater.*, vol. 19, no. 20, pp. 3171–3176, 2007.

[139] E. L. P. Dumont, H. Belmas, and H. Hess, "Observing the mushroom-to-brush transition for Kinesin proteins," en, *Langmuir*, vol. 29, no. 49, pp. 15 142–15 145, 2013.

[140] J. Howard, A. Hunt, and S. Baek, "Assay of microtubule movement driven by single Kinesin molecules," in *Methods in Cell Biology, Vol 39: Motility Assays for Motor Proteins*, J. M. Scholey, Ed., vol. 39, 1993, pp. 137–147.

[141] D. L. Coy, M. Wagenbach, and J. Howard, "Kinesin takes one 8-nm step for each ATP that it hydrolyzes," en, *J. Biol. Chem.*, vol. 274, no. 6, pp. 3667–3671, 1999.

[142] Y. Jeune-Smith and H. Hess, "Engineering the length distribution of microtubules polymerized in vitro," en, *Soft Matter*, vol. 6, no. 8, pp. 1778–1784, 2010.

[143] L. Hilbert, S. Cumarasamy, N. B. Zitouni, M. C. Mackey, and A.-M. Lauzon, "The kinetics of mechanically coupled Myosins exhibit group size-dependent regimes," *Biophys. J.*, vol. 105, no. 6, pp. 1466–1474, 2013.

[144] S. Marston, "Random walks with thin filaments: application of in vitro motility assay to the study of actomyosin regulation," *J. Muscle Res. Cell Motil.*, vol. 24, no. 2, pp. 149–156, 2003.

[145] L. Scharrel, R. Ma, R. Schneider, F. Jlicher, and S. Diez, "Multimotor transport in a system of active and inactive Kinesin-1 motors," *Biophys. J.*, vol. 107, no. 2, pp. 365–372, 2014.

[146] D. M. Leitner and J. E. Straub, *Proteins: energy, heat and signal flow*. CRC Press, 2009.

[147] J. Kerssemakers, J. Howard, H. Hess, and S. Diez, "The distance that kinesin-1 holds its cargo from the microtubule surface measured by fluorescence interference contrast microscopy," en, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 43, pp. 15 812–15 817, 2006.

[148] J. Kierfeld, K. Frentzel, P. Kraikivski, and R. Lipowsky, "Active dynamics of filaments in motility assays," *Eur. Phys. J. Spec. Top.*, vol. 157, no. 1, pp. 123–133, 2008.

[149] P. Kraikivski, R. Lipowsky, and J. Kierfeld, "Enhanced ordering of interacting filaments by molecular motors," en, *Phys. Rev. Lett.*, vol. 96, no. 25, p. 258 103, 2006.

[150] C. M. Coppin, D. W. Pierce, L. Hsu, and R. D. Vale, "The loud dependence of kinesin's mechanical cycle," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 94, no. 16, pp. 8539–8544, 1997.

[151] H. Tanaka, A. Ishijima, M. Honda, K. Saito, and T. Yanagida, "Orientation dependence of displacements by a single one-headed myosin relative to the actin filament," en, *Biophys. J.*, vol. 75, no. 4, pp. 1886–1894, 1998.

[152] M. Kaya and H. Higuchi, "Nonlinear elasticity and an 8-nm working stroke of single myosin molecules in myofilaments," en, *Science*, vol. 329, no. 5992, pp. 686–689, 2010.

[153] H. Hess, J. Clemmens, D. Qin, J. Howard, and V. Vogel, "Light-controlled molecular shuttles made from motor proteins carrying cargo on engineered surfaces," *Nano Lett.*, vol. 1, no. 5, pp. 235–239, 2001.

[154] H. Hess, "Optimal loading of molecular bonds," *Nano Lett.*, vol. 12, no. 11, pp. 5813–5814, 2012.

[155] D. K. Jamison, J. W. Driver, and M. R. Diehl, "Cooperative responses of multiple kinesins to variable and constant loads," en, *J. Biol. Chem.*, vol. 287, no. 5, pp. 3357–3365, 2012.

[156] K. Uppulury, A. K. Efremov, J. W. Driver, D. K. Jamison, M. R. Diehl, and A. B. Kolomeisky, "How the interplay between mechanical and nonmechanical

interactions affects multiple Kinesin dynamics," *J. Phys. Chem. B*, vol. 116, no. 30, pp. 8846–8855, 2012.

[157] R. F. Hariadi, R. F. Sommese, A. S. Adhikari, R. E. Taylor, S. Sutton, J. A. Spudich, and S. Sivaramakrishnan, "Mechanical coordination in motor ensembles revealed using engineered artificial myosin filaments," *Nat. Nanotechnol.*, vol. 10, p. 696, 2015.

[158] Z. H. He, R. Bottinelli, M. A. Pellegrino, M. A. Ferenczi, and C. Reggiani, "ATP consumption and efficiency of human single muscle fibers with different myosin isoform composition," en, *Biophys. J.*, vol. 79, no. 2, pp. 945–961, 2000.

[159] W. Wang, T.-Y. Chiang, D. Velegol, and T. E. Mallouk, "Understanding the efficiency of autonomous nano- and microscale Motors," en, *J. Am. Chem. Soc.*, vol. 135, no. 28, pp. 10 557–10 565, 2013.

[160] R. R. Gullapalli, T. Tabouillot, R. Mathura, J. H. Dangaria, and P. J. Butler, "Integrated multimodal microscopy, time-resolved fluorescence, and optical-trap rheometry: toward single molecule mechanobiology," en, *J. Biomed. Opt.*, vol. 12, no. 1, p. 014 012, 2007.