Optimizing Scientific Innovation by Learning on Knowledge Graph Dynamics

by

James Woodward Weis

Sc.B., Brown University (2012)
S.M., Massachusetts Institute of Technology (2017)

Submitted to the Computational & Systems Biology Program
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2020

Author . . . .
Computational & Systems Biology Program
August 24, 2020

Certified by . . . .
Joseph M. Jacobson
Associate Professor
Thesis Supervisor

Accepted by . . . .
Christopher Burge
Director, Computational & Systems Biology Graduate Program

Optimizing Scientific Innovation by Learning on Knowledge Graph Dynamics

by

James Woodward Weis

Submitted to the Computational & Systems Biology Program
on August 24, 2020, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

The integration of data-driven methodologies, including techniques from artificial intelligence and network science, into the research process and funding ecosystem is an exciting, potentially paradigm-changing opportunity to augment the effective intelligence of the scientific community—potentially increasing the efficiency, fairness, and overall impact of the scientific enterprise.

In this thesis, we explore the development of new technologies to extract actionable insights from large-scale data corpora through the design and deployment of machine learning approaches. Specifically, we describe (1) the creation of new algorithms that compute on simulations of complex biophysical processes to generate novel scientific insights, (2) artificial intelligence-based improvements to the academic publishing system, (3) a study of institutional barriers bottle-necking the development of large-scale algorithmic approaches to scientific knowledge analysis, and (4) a new algorithmic framework that, by learning from the history of biotechnology innovation as models by dynamic knowledge graphs, is able to identify with high-fidelity new technologies of likely high future impact.

We also develop tools to facilitate the real-world utilization of these quantitative approaches, effectively demonstrating how theses "intelligence-augmenting" algorithms could be used to more efficiently navigate the scientific literature and design scientifically impactful collaborations. Finally, we conclude by discussing the potential deployment of these technologies in the future—with a focus on potential applications in the funding of scientific research and commercialization, and the potential design of diversified, impact-optimized funding portfolios.

Collectively, our results demonstrate that machine learning approaches can be used to extract meaningful insight from existing data corpora, and that these signals can be used synergistically with human intuition to increase the rate at which we collectively generate breakthrough scientific insights and transformative new technologies.

Thesis Supervisor: Joseph M. Jacobson
Title: Associate Professor

Acknowledgments

I would like to express my thanks and gratitude to those that have made this work—as well as the related conversations, reflections, challenges, and adventures that constitute critical, if less tangible, components of the journey and education this thesis represents—not only possible but also fruitful and enjoyable.

Firstly, my advisor, Professor Joseph Jacobson, whose deep and wide-ranging knowledge is coupled with a rare intuition for innovation. Without our dynamic, thought-provoking conversations as context, this work would not have been possible. Joe's unique combination of intellectual rigor, ambitious rumination, and principled execution will undoubtedly serve as an invaluable guiding constellation as I navigate future ventures.

Professor Joi Ito, whose support and advice was critical to the ideation and exploration of the subjects treated in this thesis. From Joi I learned how to iterativly explore new ideas, how to identify and learn from global thought leaders in areas of interest, and how to fearlesslessly build real, impactful organizations and projects around new ideas and insights.

My thesis committee, including Professors Shuguang Zhang, Timothy Lu, and Sandy Pentland. Shuguang's expert and nuanced advise was vital to the development of the projects contained herein, as well as to this thesis, and my overall development as a researcher. Tim provided continual insight and motivation over the length of my time at MIT, and his cross-disciplinary expertise and interests are a constant source of inspiration and motivation. Sandy's targeted feedback, as well as academic and commercialization intuition and multidimensional historical perspective, lent an important dimension to this work. All gave freely and generously of their time, for which I am immensely grateful—and I could not imagine a better set of mentors from which to learn.

My family, including my father, mother, sister, and brother, and my grandmother, grandfather, aunts, uncles, and cousins. Their steadfast encouragement and trust enabled (and at times even encouraged) a level of risk tolerance that, I think, is

necessary for “harvesting from existence the greatest fruitfulness and the greatest enjoyment.”

Trista, who has been my irreplaceable companion on this journey—as well as many others. On her shoulders rests a great deal of responsibility for the completion of this work. Her unyielding support formed the foundation, the “endless summer,” upon which this work, and so much else, depends.

My friends, including Raja, Brian, Jeremy, and Frederik. “A good writer possesses not only his own spirit but also the spirit of his friends.” This thesis—and much else—is deeply imbued with their perspectives. Without the lessons, trials, and tribulations we have faced together, my graduate school experience would be unrecognizable.

Pranam, Noah, and the rest of the Molecular Machines group at the MIT Media Lab. From enthusiastically welcoming me into the group, to providing on-the-grounds guidance, their kindness and interdisciplinary expertise is much appreciated.

Finally, Chris, Jacquie, and the CSBi community. The deep commitment you have towards fostering successful graduate school experiences, even when the paths are complex and divergent, has continually impressed me—and I am deeply thankful for your advocation.

Thank you.

Contents

List of Figures

List of Tables

Onwards. And so onwards along the path of wisdom, with a hearty tread, a hearty confidence! However you may be, be your own source of experience! Throw off your discontent about your nature; forgive yourself your own self, for you have in it a ladder with a hundred rungs, on which you can climb to knowledge. The age into which you feel yourself thrown with sorrow calls you blessed because of this stroke of fortune; it calls to you so that you may share in experiences that men of a later time will perhaps have to forego...

Stroll backwards, treading in the footprints in which humanity made its great and sorrowful passage through the desert of the past; then you have been instructed most surely about the places where all later humanity cannot or may not go again. And by wanting with all your strength to detect in advance how the knot of the future will be tied, your own life takes on the value of a tool and means to knowledge. You have it in your power to merge everything you have lived through–attempts, false starts, errors, delusions, passions, your love and your hope–into your goal, with nothing left over: you are to become an inevitable chain of culture-rings, and on the basis of this inevitability, to deduce the inevitable course of culture in general.

When your sight has become good enough to see the bottom in the dark well of your being and knowing, you may also see in its mirror the distant constellations of future cultures.

Human, All Too Human

1

Augmenting Human Scientific Capacity with Machine Learning

1.1 Introduction & Perspective

The past introduction of computation over the past few decades has arguably led to perhaps the greatest paradigm-shift in human interaction, and thus the structure of society generally, since the agricultural revolution over 12,000 years ago [8]. In this period, spanning from the development of the first general-purpose Turing-complete computer in 1945 to the present day, computational processing has become ubiquitous [100]; and the development of data-driven approaches has made the extraction of increasingly nuanced, super-human insights possible in fields ranging from social media to radiology.

Recent advances in machine learning and artificial intelligence can now extract valuable signals from large corpora of data; in effect, by learning patterns on a high-dimensional space from input data, machine learning algorithms are able to make inferences about the likely properties of the new data based on their relationship to these learned patterns. Because these algorithms are not limited by human cognitive capacity (e.g. memory and dimensionality constraints), super-human performance becomes possible in certain tasks. To-date, many of the successful applications of machine learning algorithms has been within a specific class of problem: the complete replacement of simple tasks with specific and targeted algorithms. This includes, for example, classifying the contents of an image, translating speech to text, identifying tumors, and predicting short-term stock price movements [103, 4, 29, 69]. In many cases, the desired outcome is performing such a highly-targeted task at super-human scale, speed, or accuracy levels.

1.2 An Opportunity To Scale Science

There exists another area of artificial intelligence research, wherein algorithms are designed to support humans dynamically with tasks. Such approaches, which combine the scale and speed of artificial intelligence algorithms with human-level intuition, interpretation, and prioritization, have led to innovations in fields like Human Computer Interaction (HCI), visualization, and music [37]. However, for a variety of reasons, the application of these approaches socially-critical problems like resource allocation, has not yet been deeply explored–despite the enormous benefits that even slight improvements in efficiency could bring.

The potential implied by the application of such intelligence-augmenting algorithms in the scientific ecosystem is extraordinary. In the mid-17th century, the breadth of scientific communications was captured largely by personal letters between a small population of researchers; it was not until 1665 that the world's first scientific journal, *Philosophical Transactions of the Royal Society*, was established. Since then, the universe of scientific communication has grown enormously–accompanied by an explosion of academic journals, research institutions, and new fields of study. As the scope of science and technology continues to grow–with high-quality interactions now taking place not only in academic journals, but via online platforms and "preprint" servers–the ability of any individual, or group of individuals, to understand the nuances, drivers, and potentially promising areas of research of even a single field continues to decline. In a world of rapidly expanding information, and yet limited resources, the decisions of what to study, who to study with, and what research to fund become increasingly difficult. This dynamic has led to increasing evidence that we, as a world population, are getting collectively less scientific impact for each incremental hour (or dollar) deployed [49, 16]

Thus, we may stand on the precipice of potentially another paradigm-change in scientific society; that is, will the increasing breadth of scientific knowledge inevitable dilute our ability to generate new scientific breakthroughs, resulting in a continually decreasing scientific return-on-investment (ROI)—or will we be able to optimize our scientific and technological processes by leveraging this wealth of information, continually refining and re-allocating our resources to generate increasing amounts of breakthrough insights and transformative technologies?

1.3 Thesis Structure

It is the goal of this thesis to explore this question—and to propose new insights and technologies that, collectively, constitute a step towards a future where artificial intelligence algorithms are used not to replace humans, but to augment our effective intelligence, enabling us to chart a path towards greater insights, more technological breakthroughs, and higher scientific efficiency.

I describe herein the creation of new machine learning-based methods to generate new scientific insights from biophysical simulation data (Chapter 2), artificial intelligence-based improvements to the academic publishing system (Chapter 3), institutional barriers to data-driven approaches in science and their potential solutions (Chapter 3), and a new framework that, by learning from scientific history, is able to identify with high-fidelity new technologies likely to be highly-impactful in the future (Chapter 5). I also design tools to demonstrate how such intelligence-augmenting algorithms could be utilized (Chapter 6), and conclude by discussing the myriad avenues through which the technology described herein could be deployed in the future—with a focus on the philanthropic and for-profit funding of science and technology research and commercialization (Chapter 7).

Out there – thus I will; so doing
trust myself now and my grip.
Open lies the sea, its blueing
swallows my Genoese ship.

All things now are new and beaming,
space and time their noon degree –:
Only your eye, monstrous, gleaming
stares at me, infinity!

Toward New Seas

2

Learning on Biomolecular Dynamics to Optimize Enzyme Catalysis

The work in this chapter is adapted with permission from the manuscript *Machine Learning Identifies Chemical Characteristics that Promote Enzyme Catalysis* written by Brian M. Bonk, James W. Weis, and Bruce Tidor and published in the *Journal of the American Chemical Society (JACS)* on February 14, 2019.

B.M.B., J.W.W., and B.T. conceived of the overall project and developed the approach and plan. B.M.B. performed the simulations, generated the data sets, led the data analysis, and wrote the initial manuscript draft. J.W.W. explored methods for feature selection and implemented the method used here. B.M.B., J.W.W., and B.T. contributed to the analysis of the data and developed the final manuscript.

2.1 Abstract

Despite tremendous progress in understanding and engineering enzymes, knowledge of how enzyme structures and their dynamics induce observed catalytic properties is incomplete, and capabilities to engineer enzymes fall far short of industrial needs. Here we investigate the structural and dynamic drivers of enzyme catalysis for the rate-limiting step of the industrially important enzyme ketol-acid reductoisomerase (KARI) and identify a portion of the conformational space of the bound enzyme–substrate complex that, when populated, leads to large increases in reactivity. We apply computational statistical mechanical methods that implement transition interface sampling to simulate the kinetics of the reaction and combine this with machine learning techniques from artificial intelligence to select features relevant to reactivity and to build predictive models for reactive trajectories. We find that conformational descriptors alone, without the need for dynamic ones, are sufficient to predict reactivity with greater than 85% accuracy (90% AUC). Key descriptors distinguishing reactive from almost-reactive trajectories quantify substrate conformation, substrate bond polarization, and metal coordination geometry and suggest their role in promoting substrate reactivity. Moreover, trajectories constrained to visit a portion of the reactant well, separated from the rest by a simple hyperplane defined by ten conformational parameters, show increases in computed reactivity by many orders of magnitude. This study provides evidence for the existence of reactivity promoting regions within the conformational space of the enzyme–substrate complex and develops methodology for identifying and validating these particularly reactive regions of phase space. We suggest that identification of reactivity promoting regions and re-engineering enzymes to preferentially populate them, may lead to significant rate enhancements.

2.2 Introduction

Enzymes are remarkable catalysts that produce substantial rate enhancements, often accompanied by high substrate and product selectivity. They are increasingly important for industrial-scale applications, because of the chemistry they can accomplish sustainably in mild, aqueous conditions. Despite substantial progress made, more is still required along two principal avenues in order to advance enzyme engineering to meet industrial needs. We need a better understanding of the drivers of reactivity promoted by enzymes, some of which have been hypothesized to be dynamic[9, 95, 61] rather than structural, along with a richer set of tools to probe and manipulate the active-site catalytic environment.

Current approaches include directed evolution[89, 53, 81], catalytic antibodies[67, 84, 71] and computational enzyme design [64, 7], the latter two of which focus on tight-binding of transition states. While these approaches have produced tremendous successes, they have not yet become general-purpose tools. The need for directed evolution to improve designs obtained by other methods, and our inability to fully understand the improvements accumulated through evolution, suggest that our understanding may be incomplete, perhaps in some fundamental way, and may require us to incorporate other factors beyond transition-state binding and transition-state stabilization (relative to the bound or unbound ground state).

Here we investigate two fundamental questions of enzyme function motivated by the larger goal of enzyme engineering; note that our focus is on the enzyme–substrate complex without specific reference to the transition state. First, can we gain insight into the nature of the drivers of chemical reactivity, and to what extent are these drivers apparent in the behavior of the bound enzyme–substrate complex, well before the transition state? And second, based on previous work of ourselves and others [101, 58, 96, 124, 110] can we identify regions of the conformational space of the enzyme–substrate complex that are inherently more reactive than others? These questions are addressed using a new approach that combines machine learning with path sampling, applied to the rate-limiting step for the industrially important enzyme ketol-acid reductoisomerase (KARI).

There are a number of approaches for studying enzyme reactivity that do not focus on the transition state per se, although it may enter implicitly. These include the literature investigating near-attack conformations, which has suggested that lowering the energetic barrier to facilitate selective formation of subsets of ground-state conformations that lie on the path to the transition state, can be just as important as lowering the energetic barrier to the transition state itself [96, 66, 23, 22] and the computational path sampling methods [33, 111], which are statistical mechanical techniques for directly computing the rate of a chemical reaction without reliance on transition-state theory or knowledge of either the transition state or a valid reaction coordinate connecting the reactant well with the product well on the free energy surface.

Here we use transition interface sampling [111] (TIS), for its computational efficiency. TIS uses Monte Carlo sampling to construct an ensemble of trajectories that start in the reactant well and pass through an interface on the way toward the product well. Appropriate statistical methods exist to compute the progressive probability that a trajectory starting in the reactant well will reach each interface, a rapidly diminishing cumulative probability, and to convert the probability into a reaction rate, corresponding to the specific activity, k_{cat}, for enzymes. While a valid reaction coordinate is not a requirement, the method uses an order parameter that cleanly distinguishes reactant from product to track progress between the two wells [111]. (The placement of interfaces is shown schematically in Figure A-1A and their progression in Figure S1, with λ representing the order parameter.)

The model system for this study, KARI, is a natural enzyme required for branched-chain amino-acid synthesis, found broadly across plant and microbial species[35]. It carries out two reactions in sequence, first an isomerization, which is generally rate limiting, consisting of an alkyl migration and then a faster reduction carried out by a nucleotide cofactor. It also has an important role in industrial processes for the production of isobutanol, and, due to its role as the rate-limiting step, improvements in its specific activity would improve processes for large-scale isobutanol production [26]. Our studies have focused on the homodimeric enzyme from Spinacia oleracea, due largely to the availability of appropriate crystal structures, and we have studied the industrially relevant, rate-limiting reaction step involving isomerization of (2S)-acetolactate (AL) to (2R)-2,3-dihyroxy-3-isovalerate through methyl migration [26, 10, 102] (Figure 2-1B).

The natural spinach enzyme exhibits a strong preference for NADPH as a cofactor and has two divalent magnesium cations bound at the active-site, in intimate contact with substrate[15], which are each hexacoordinate with oxygen atoms from the substrate, active-site water molecules, and residues Asp315, Glu319, and Glu496 (Figure 2-1C). Note that the C5 represents the methyl group that migrates from C4 to C7.

The current study is based on previous work we carried out on KARI, which identified a “pump-and-push” mechanism for the rate-limiting isomerization reaction, whereby the local environment vibrationally excites the breaking C4–C5 bond and the side chain of Glu319 helps direct and potentially stabilize the migrating methyl group towards its destination, bound to C7 12. Moreover, the work suggested that some portions of the conformational and motional space of the bound enzyme–substrate complex (the reactant well) led to trajectories that have a greater probability of reacting than those that do not pass through or spend as much time in those same portions of the reactant well. The term “more reactive” portions of the reactant well is used to represent this idea.

Here we carried out TIS simulations of wild-type spinach KARI and performed comparative analysis on two sets of ensembles of trajectories–one set that reacted and another set that approached the barrier but did not react (termed “almost-reactive”). We tabulated data on 68 different geometric measurements (Table B.2 and Figure 2-4) in the active site that represent elements of the local conformation in the form of distances between pairs of atoms, planar angles across triplets of atoms, and dihedral angles across quadruplets of atoms. The set was selected based on mechanistic hypotheses of others and ourselves, and includes internal metrics within the substrate; measures of the position and orientation of substrate relative to the environment, particularly for groups that might stabilize the bound substrate or transition state; and measures of conformation of the environment.

Machine learning techniques were applied to identify subsets of this feature list and build predictive models that accurately distinguished reactive from almost-reactive trajectories, based only on data tabulated from before trajectories departed the reactant well. We reasoned that these reduced feature sets and models might indicate key features sufficient to drive reactivity. We analyzed these features in the context of the reactive and almost-reactive trajectories to understand in more detail these drivers and to gain insight into mechanism. We found that key descriptors capable of identifying reactive conformations included those that quantify substrate conformation, substrate bond polarization, and metal coordination geometry and suggest their role in promoting substrate reactivity. To test the notion that these descriptors are sufficient and that they define inherently reactive portions of the reactant well, we compared the computed specific activity of the wild-type enzyme when trajectories were constrained to visit these regions with those that were not. We found that ten features alone were sufficient to describe a portion of the reactant well that led to very large rate increases, demonstrating it as a highly reactive portion of the well.

2.3 Methods

2.3.1 Structure Preparation

The crystal structure of *Spinacia oleracea* KARI was obtained from the Protein Data Bank[13, 12] with the accession code 1YVE [15] and prepared as described previously by Silver[101]. Only the chain A monomer was used for all simulations in order to improve computational efficiency, justified by the significant separation between the active sites of the two monomers [15] (Figure 2-5).

A model of the substrate-bound enzyme was then constructed by running an in vacuo QM ground-state minimization of the substrate, two magnesium centers, five magnesium-coordinating water molecules, and the side chains of three surrounding active-site residues, Asp315, Glu319, and Glu496. Glu496 was protonated, consistent with previous studies indicating its importance in stabilizing the transition and product state by forming a hydrogen bond with the substrate O8 [90]. The GAUSSIAN03 computer program [43] was used to perform in vacuo QM calculations at the rhf/3-21g* level of theory[88, 87].

2.3.2 Simulation Methodology

CHARMM version 41 [21, 20] compiled with the SQUANTUM option was used to perform all molecular dynamics simulations. The QM portion of the energy function was calculated with the AM1 semi-empirical quantum mechanical force field [34]; the MM portion of the energy function was computed using the CHARMM36 all-atom force field [57]. Additional AM1 parameters were used for the magnesium ions 38. The following atoms made up the QM region: substrate (acetolactate), both magnesium centers, five magnesium-coordinating active site water molecules, the side chains of Asp315, Glu319, and Glu496, and the nicotinamide group of NADPH (Figure 2-1C). The Generalized Hybrid Orbital method 39 was used to treat the QM/MM boundary atoms. The substrate O6 was deprotonated and the coordinating Glu496 was protonated, paralleling previous QM/MM studies of KARI [90].

2.3.3 Seed Trajectory Generation

The initial reactive trajectories used to bootstrap the TIS simulations were found by computing a potential of mean force (PMF) along the order parameter λ, defined as the difference of the distance between the substrate breaking bond (C4–C5) and the forming bond (C5–C7), in units of ångstroms. This PMF was computed using umbrella sampling and the weighted histogram analysis method [65]. The umbrella sampling was performed in CHARMM41 using the RXNCOR module with windows 0.05-Å in width and harmonic constraints of 300 kcal/(mol·Å). Candidate seed trajectories were then generated by integrating forward and backward for 2,000 fs without constraints starting from a randomly chosen frame from the umbrella sampling window ensembles centered at λ values of –0.05, 0.00, and +0.05. Trajectories were selected as successful seed trajectories if they connected the reactant basin ($\lambda <$ ˘1) and product basin ($\lambda > +1$).

2.3.4 Training Data Set Generation And Time Point Selection

Three randomly-selected connecting seed trajectories from the collection described above were used as starting trajectories for the generation of a larger ensemble of reactive and almost-reactive trajectories. Each seed was used to generate 9 reactive ensembles and 9 almost-reactive ensembles of 20,000 trajectories each. The combined data set contained 461,422 almost-reactive and 618,578 reactive trajectories. When the almost-reactive process produced a reactive trajectory, it was removed from that set and added to the reactive data set. To ensure a balanced number of reactive and almost-reactive trajectories in each training and testing data set, the reactive trajectories were randomly sampled without replacement to produce a set of 461,422 reactive trajectories.

For the reactive ensembles, the product interface was defined as $\lambda_R = +1.00$, and for the almost-reactive ensembles, the product interface was defined as $\lambda_{AR} =$ ˘0.20 (Figure 2-1A). The TIS methodology was applied in parallel to produce statistical mechanical ensembles containing reactive and to almost-reactive trajectories that could be compared to one another. In both ensembles, the reactant interface was defined as $\lambda =$ ˘1.00. To collect time points early in the reactant basin for analysis, integration was not stopped once a trajectory reached the reactant and product interface (and had been accepted into the Markov chain), but continued forward and backward for a total of 200 fs in each direction.

To ensure that candidate features (see below) were computed at analogous time points between reactive and almost-reactive trajectory ensembles, in a post-processing step, all almost-reactive and reactive trajectories from all 27 pairs of ensembles were time-shifted such that the 0-fs time point corresponded to the bottom of the last "trough" in λ (when plotted vs. time) before the prospective alkyl migration event, a geometric feature that all the collected trajectories shared (Figure 2-1D). This trough was found by first finding the point in the trajectory closest to the transition region at $\lambda = 0$, then scanning along the trajectory backward from this point until the first change in sign of the derivative of λ with respect to time was found with a value of λ

less than 0 (i.e., was located in the reactant basin). All other time points were defined relative to this first trough at time 0. Cartesian coordinate frames of atomic positions were collected in 5-fs increments from the 0-fs time point, going backward to –150 fs and forward to +35 fs from the t=0-fs point, for a total of 38 total time points. This collection of sub-sampled time points was used for all subsequent analysis.

2.3.5 Feature Computation

At each of the 38 time points between –150 and +35 fs, the set of 68 structural features in Table B.2 were computed for each of the trajectories in each of the 27 reactive and 27 almost-reactive ensembles. The 68 features are illustrated structurally in Figure 2-4A (distances), Figure 2-4B (angles), and Figure 2-4C (dihedrals). These data were pooled across ensembles to produce one combined reactive and one combined almost-reactive data set at each of the 38 time points, which were used in machine learning and subsequent analysis described below and stored as a row in a data matrix. For model training, the data matrix at each time point was randomly sampled without replacement to produce 5 equal partitions containing 73,827 trajectories each, and for model testing, the remaining trajectories were randomly sampled to produce five equal partitions containing 18,456 trajectories each.

2.3.6 Machine Learning

For feature regularization and discovery, LASSO[106] was used with the *lassoglm* implementation in MATLAB. In order to select a given number of features with LASSO, the regularization parameter λ was adjusted until a specific number m (1, 5, 10, 15, 20, 25, or 30) of non-zero coefficients β_j remained (using a tolerance of $10^{\breve{}4}$). These m LASSO-selected predictor features with non-zero coefficients were then fit using the *fitglm* function in MATLAB to a logistic classifier. After fitting predictor coefficients, the area under the curve of the receiver operating characteristic (AUC) was computed for each logistic classifier using the *perfcurve* function in MATLAB.

2.3.7 Cluster Assignment

Reactive clusters were assigned by k-means clustering, with the *kmeans* function in MATLAB using $k = 5$ applied to the matrix of consensus feature Z-scores weighted by their corresponding logistic coefficient β_j for all correctly classified reactive trajectories. The number of clusters (5) was chosen based on a hierarchical clustering analysis also performed in MATLAB (data not shown). The Euclidian distance of the consensus feature set from each almost-reactive trajectory to each of the five k-means centers was computed, and each almost-reactive trajectory was then assigned to the cluster with the shortest Euclidian distance to its respective centroid.

2.3.8 Rate Constant Computations

For the TIS flux factor calculations, a total of 10 independent 1-nanosecond molecular dynamics simulations were performed starting from reactant structures derived from each of 6 randomly selected seed trajectories generated as described above. The λ_A interface was set equal to the λ_1 interface at $\lambda =$ ˘0.8. For the control flux factor computations (Figure A-1A), the effective positive flux was computed as the number of times the trajectory crossed the $\lambda_A =$ ˘0.8 interface, having come from the region below the interface, divided by the total amount of time spent below the λ_A interface. For the constrained test flux factor computations (Figure A-1C), the top 10 LASSO-selected features at the $t = 0$ time point were written out during the dynamics run, and the effective positive flux was computed as the number of times the trajectory crossed the $\lambda_1 =$ ˘0.8 interface, having come from the region A', where region A' refers to all points in phase space which lie at the last trough (i.e., the first point at which $\frac{d\lambda}{dt} = 0$ and $\frac{d^2\lambda}{dt^2} > 0$) before crossing $\lambda_A =$ ˘0.8, having first crossed $\lambda_0 =$ ˘1, and for which the logistic classifier with coefficients and features listed in Table 2.2S2 evaluated to true.

For the probability factor calculations, a total of 29 $P(\lambda_{i+1} \mid \lambda_i)$ interface ensembles from each of the six seed trajectories were computed, with the λ_i interfaces spaced between $\lambda =$ ˘0.8 and $\lambda = 0$. The placement of these interfaces relative to the

PMF surface used to generate initial seed is shown in Figure S4. For each interface ensemble, a total of 5000 shooting moves was attempted. In each λ_i ensemble, candidate trajectories were generated using full shooting moves and accepted if they both crossed the $\lambda_A = \breve{}0.8$ interface and crossed the $\lambda = \lambda_i$ interface having first come from crossing interface λ_A. For the unconstrained control ensembles (Figure A-1B), no further acceptance rules were applied.

2.4 Results

2.4.1 Machine Learning

Data sets consisting of 27 ensembles each of reactive and almost-reactive trajectories generated using a combined QM/MM TIS approach, were analyzed with machine learning to identify features with the ability to distinguish reactive from almost-reactive trajectories. At each of 38 time points between –150 and +35 fs (5-fs spacing and shown in Figure 2-1D), the 68 features listed in Table B.2 and illustrated structurally in Figure 2-4 were computed for both sets of reactive and almost-reactive ensembles. To assess individual feature performance, AUC (area under the curve of the receiver operating characteristic) was computed for all single features at the 0-fs time point (Figure 2-4A). The single feature with the maximum AUC performance was the distance between Glu319 $O_{\epsilon 1}$ and substrate C5 (AUC of 0.73). Only two features (distance Glu319/Oe1–AC6/C5 and distance AC6/C4–AC6/C5) produced models with individual AUCs above 0.70, and 18 features produced models with AUCs above 0.60.

To find highly predictive groups of features, LASSO[106] was applied iteratively with different penalty strengths to identify an ordered set of features for each trajectory time point, optimized to distinguish reactive from almost-reactive conformations (see Section 2.3). That is, for each time point a collection of separate classifiers was built, trained, and tested, enabling comparisons of the useful sets of features across time points as well as the performance benefits for increased numbers of features at each time point. Figure 2-2B shows the machine learning results for four classifier performance statistics (AUC, accuracy, sensitivity, and specificity) computed from each model constructed from data at each time point. Results for models constructed with optimized sets of 1, 5, 10, 15, and 20 features selected by LASSO are shown. The results show progressively improved performance as the number of features was increased, with not insignificant performance with just one feature (generally 0.65–0.75 AUC) that rose to excellent performance with 10, 15, and 20 features (generally 0.85–0.95 AUC). Note that the performance of the LASSO-selected 1-feature models, being the "best" feature for each time point, was significantly better than the average AUC of all possible 1-feature models shown in Figure2-2A, which was 57.18%. The similarity in performance between 15- and 20-feature models suggests near convergence with this number of features. The models developed were well balanced between false positives and false negatives as judged by similar values for the sensitivity and specificity metrics of individual classifiers, as well as the AUC values. Models performed similarly (for the same number of features) for time points between –150 and +20 fs, and then became substantially better (approaching an AUC of 1.00) for time points after +20 fs, which corresponds to times when the reactive and almost-reactive trajectories began to separate based on the order parameter λ (Figure 2-1D).

To assess the effect of LASSO-optimized feature selection for use in machine learning models, a control was carried out in which a classifier was trained similarly but using feature sets randomly chosen from the original 68 features. That is, each control classifier was optimally trained for the best performance possible with the random (and not optimized) features it was assigned. Analogous performance statistics for these control classifiers are shown in Figure 2-2C. The results showed improved performance with additional features randomly selected from a chemically plausible set, together with large error bars, which is consistent with the notion that at any given time point some features or combinations of features were much better able than others to create predictive models, and the performance of models depended greatly on the features making up that model. Models with any given number of features performed much better on average when those features were selected by LASSO based on predictive ability than when selected randomly, demonstrating the value of the LASSO-selected features in distinguishing reactive from almost-reactive trajectories; for example, many of the one-feature models with LASSO-selected features had AUCs of about 0.70, whereas the random models had average AUCs of 0.57. The random models showed improved average performance after $t = +20$ fs, consistent with the notion that many features report on the fact that the reaction had largely begun by that time.

2.4.2 Analysis Of Consensus Feature Set Predictive Throughout Pre-Launch Time Window

The union of the complete 20-feature sets predictive at all 31 time points between –150 and 0 fs is depicted in Figure 2-2E. Features are listed in decreasing order of frequency of appearance, and the colored bars indicate the time points for which each feature appears as one of the 20 LASSO-selected features. (The time range –150 to 0 fs will be called the "pre-launch time window" for shorthand, as the 0-fs time point represents the last compression before the ultimate expansion of the putative breaking bond.) The results show that 17 of the features were used throughout at least half the window, 31 features were used at 10 or more time points, nearly all of the original features were used at least once (54 from the collection of 68), and 8 were used at five or fewer time points. The results suggest a commonality amongst the geometric descriptors that were broadly predictive across the pre-launch window. The names and feature types of the top 30 consistently predictive, consensus features are presented in Table B.1 along with the number of occurrences in the top 20 LASSO-selected sets within the pre-launch window. Figure 2-2D shows the classification performance of models trained using the top 1, 5, 10, 15, 20, 25, and 30 consensus features across the 31 time points between –150 and 0 fs. With the 30 consensus features, classification performance was nearly equivalent to or better throughout the pre-launch window (approximately 0.90 AUC) than the performance obtained from 20 LASSO-selected features optimized for each of the individual time points. That is, 30 shared features performed as well as 20 custom features across the range, which is strong evidence that the fundamental determinants of re activity are relatively consistent across the pre-launch window. Because the classifiers were each trained separately at each time point to produce models with different learned coefficients, these fundamental determinants of reactivity can (and do) play different roles at different times.

A structural representation of the set of 30 predictive consensus features is shown in Figure 2-7A (17 distances) and Figure 2-7B (12 planar angles and 1 dihedral angle). Half of the features (15) represent interactions between the substrate and its environment (nearby water molecules, the two magnesium ions, and the side chain of Glu319), 7 represent intra-substrate conformational metrics, 7 represent water–metal interactions, 1 represents an intra-co-factor orientation, and 2 represent other intra-environment interactions. A full third of the features (10) represent distances or angles describing the relationship of a single atom, the substrate hydroxyl oxygen (O6), to its environment–the coordinating magnesium ions and water molecules interacting with the metal ions. The largest number of intermolecular features involving any other substrate atom is 2, for both a substrate carboxylate oxygen (O3) and the substrate carbonyl oxygen (O8), whose carbon receives the migrating methyl group. Only one intermolecular interaction involves the migrating methyl itself. We note two additional characteristics of the feature set: (1) the substrate intramolecular features involve the geometry local to the C4–C7 covalent bond, which is parallel to the path of the migrating methyl group, and (2) 8 of the 10 intermolecular angle features describe the orientation of groups coordinating the metal ions–either their ligated water molecules or oxygen atoms of the substrate. We acknowledge that the composition of the initial 68 features had some effect on the composition of the selected features; nevertheless, the resulting consensus feature set suggests important roles for substrate conformation, substrate bond polarization, and metal coordination in the reaction mechanism.

Average reactive and almost-reactive time traces for the consensus feature set are presented in Figure 2-3A. The closely overlapping distributions of most features in

Figure 2-3A suggest the need for multiple features in combination to make usefully accurate predictions. 2D and 3D histograms of reactive and almost-reactive trajectories for feature pairs and triplets (data not shown) show somewhat greater separation than that seen in Figure 2-3A, but still considerable overlap between reactive and almost-reactive distributions at individual time points, consistent with the relatively poor classification performance of models with fewer than 10 features.

2.4.3 Variations Distinguish Multiple Reactive Channels

We examined the question of whether the reaction proceeded along multiple channels. Clustering was used to organize the correctly predicted reactive and almost-reactive trajectories into related sets, and the magnitude of the differences between the sets was examined, allowing a more fine-grained analysis of the determinants of reactivity as identified by the machine learning. Specifically, all correctly predicted reactive trajectories were clustered based on the 0-fs time point using the 30 consensus features, each weighted by its β_j value (we refer to this as the feature weight, which is listed in Table B.3 for the -150,-100,-50 and 0 time points; see Section 2.3; results for five clusters are shown in Figure 2-3B). The results show at least five different modes of reacting, with each cluster distinguished by which features contribute most and least to the classifier outcome. In Figure 2-3B, the thirty columns represent the contribution from each of the 30 consensus features and the rows each represent one trajectory. Figure 2-3B shows that at the 0-fs time point, roughly half of the 30 features contribute very little to the decision as indicated by white bands in each cluster. Further confirmation is seen by the observation that features that appear as white bands usually do not occur in the top 20 LASSO selected set at this time point (see Figure 2-2E; distance AC6/O6–MG6/M16 and distance MG6/O19–MG6/M17 are exceptions and rank 15 and 18, respectively, in the top 20 LASSO selected set).

Grouping the weighted features into reactive clusters and corresponding almost-reactive clusters allows the subtle differences that define reactivity for each of these subgroups to be more closely examined. To this end, the mean feature contribution for each almost-reactive cluster in Figure 2-3C was subtracted from each of the weighted features from the corresponding cluster of reactive trajectories from Figure 2-3B to obtain a mapping of how each feature in each reactive trajectory differs from its mean in the corresponding almost-reactive cluster (Figure 2-3D); the results show several common features that distinguish correctly predicted reactive from correctly predicted almost-reactive clusters. For example, across all five clusters shown in Figure 2-3D, the darkest red bands appear for distances AC6/C5–AC6/C4 and MG6/M16–AC6/O3 (features 10 and 27, respectively), indicating that these features are critical in driving the reactive/almost-reactive decision. However, there are other cluster-specific differences; for example, the distance AC6/O8–Glu496/$H_{\epsilon 2}$ (feature 6) is responsible for distinguishing reactive from nearly-reactive more for cluster 3 than for any of the others, on average.

Distributions of feature values with the strongest contributions to differences in reactivity amongst the clusters (i.e., the darkest bands in Figure 2-3D), are shown, per cluster, in Figure 2-3E. Although there is often considerable overlap in the individual feature distributions between each reactive and almost-reactive cluster, the set of 5 features alone, when re-trained on each cluster alone, achieved AUCs of 1.00, 1.00, 0.94, 0.91 and 1.00, in classifying trajectories from clusters 1 though 5, respectively, as reactive or almost-reactive. These very high scores suggest that the more general classifiers presented earlier somehow carry out the dual tasks of determining which reaction channel the trajectory is headed toward, as well as whether the trajectory will successfully react through that channel. The high AUCs for the second task above suggest that determining which channel is being approached may be the harder portion of the two, although this effect is convolved with the fact that these clusters are composed of trajectories that were correctly classified previously. When all (including incorrectly classified) data points are used, the intracluster AUCs using the same set of features are 0.92, 0.93, 0.80, 0.88 and 1.00 respectively, supporting the interpretation that predicting reactivity within a cluster is easier than in the absence of knowledge of the cluster for most of the clusters.

Figure 2-3E shows that across all five clusters, some general trends exist for the five features and their relative distribution between reactive and almost-reactive trajectories. The strongest observation is that in almost every instance, each significant feature has a much narrower distribution in the reactive than the almost-reactive set of trajectories. This is consistent with the notion that there are many ways of not reacting, but fewer modalities for successfully traversing the reaction barrier. Across most of the five clusters, in general, reactivity is associated with a shorter AC6/C5–AC6/C4 bond length (column 2; feature 9; clusters 1, 2, 4, and 5), a longer AC6/C1–AC6/C4 bond length (column 4; feature 25; clusters 2, 3, and 5), a longer Glu319/O_ϵ1–AC6/C5 distance (column 1; feature 1; clusters 1, 2, 4, and 5), and a shorter MG6/M16–AC6/O3 distance (column 5; feature 27; clusters 1, 2, and 5). The value of the MG6/H29–MG6/O18–MG6/M17 angle (column 3; feature 20) is associated with reactivity for small values in cluster 1 but large values in cluster 5. Nevertheless, the absolute values associated with reactivity for some of the features varies greatly between clusters (column 3 for clusters 1 and 5, and column 5 for clusters 1 and 2, for example). Taken together, these results reinforce the notion that a common set of fundamental reaction-promoting mechanisms are deployed in somewhat different combinations in the different clusters.

An illustration and further discussion of representative structures corresponding to the feature histograms in Figure 2-3E can be found in Figure 2-8. In summary, a comparison of these histograms and representative structures shows that features distinguishing reactive from almost-reactive trajectories include internal conformational degrees of freedom of the substrate, which may provide distortion toward the transition state and ground-state destabilization; subtle changes to polar interactions of the two magnesium ions with the substrate and with their ligating water molecules and side chains, which could have important effects in polarizing the substrate toward reactivity; and interactions of the side chain of Glu319 with the migrating methyl group, which could be important for steric, kinetic, and electronic reasons. It is anticipated that more detailed molecular orbital analyses will contribute to an understanding of how these structural differences are responsible for changes in relative reactivity.

2.4.4 Predictive Features Direct Reactivity

Machine learning was used to develop predictive models capable of distinguishing reactive from nearly reactive trajectories. Predictions of reactivity were successful, even when applied to trajectories not used in training the models, further supporting the notion that model features represent characteristics of reactivity. We reasoned that these characteristics could be useful not only to predict reactivity, but also to direct it. That is, if the features identify characteristics that are largely sufficient for reactivity, rather than just indicative of it, then trajectories constrained to possess reactive characteristics should show markedly increased reactivity. We tested that notion, described below, and our findings confirm the directive power of the machine learning features and their associated models.

The LASSO-selected, ten-feature model at the 0-fs time point was used, with testing performance AUC of 89.03% and accuracy of 81.57%. Model features and the corresponding logistic-regression coefficients are listed in Table 2.2. Eight of the ten features occur in the 30-feature consensus set, with the exceptions being distance AC6/C4–AC6/O6 and distance AC6/O8–MG6/M17. The ten-feature model achieves very good predictive performance and is composed of many of the consensus features found to be important at other time points.

The logistic regression models used here effectively create a dividing surface in the reactant well (the hyperplane defined by the β_j coefficients; see Section 2.3), and make successful predictions of reactivity based on whether the trajectory is in the "reactive portion" of the well at the appropriate time. We modified the statistical mechanical TIS sampling procedure used here to compute reaction rates, so that we could require all trajectories to be on the reactive side of the hyperplane encoded in the ten-feature model (Table 2-4) during a rate calculation (see Section 2.3). Calculations of the reaction rate were performed with ("test") and without ("control") this constraint applied only at the 0-fs time point from five different starting seeds (three were used previously to train the model, and two were new). The expectation was that the test simulations would show greater reactivity (larger computed k_{cat}) than the

Table 2.1: Computed rate constants, probability factors and flux factors for each seed studied.

Seed	Experiment	Mean P	Mean Flux (1/fs)	Mean Rate Constant (1/s)	Test/Control Fold Increase
1	Control	6.7x10-23	1.0 x10-03	6.7 x10-11	8.7 x10+19
1	Test	1.4 x10-08	4.2 x1002	5.8 x10+09	
2	Control	1.2 x10-22	9.0 x10-04	1.1 x10-10	1.3 x10+17
2	Test	1.1 x10-10	1.2 x1002	1.4 x10+07	
3	Control	2.7 x10-22	1.0 x10-03	2.7 x10-10	1.2 x10+18
3	Test	3.5x10-09	9.6 x10+01	3.4 x10+08	
4	Control	1.6 x10-22	7.0 x10-04	1.1 x10-10	7.8 x10+17
4	Test	1.0 x10-09	8.7 x10+01	8.7 x10+07	
5	Control	3.2 x10-21	1.3 x10-03	4.2 x10-09	2.0 x10+16
5	Test	3.0 x10-10	2.7 x10+02	8.2 x10+07	

controls, as the test simulations satisfied the reactivity conditions in every trajectory (by constraint), whereas on average only 8.03% of control trajectories satisfied them through ordinary sampling.

The observed relative differences in rate constants in all five sets of simulations was consistent with this expectation and quite large, on the order of 1016 to 1019, depending on the initial seed trajectory (Table 2.1). The computed rate is a product of a factor representing the rate of reactant starting toward the barrier and a probability factor representing the cumulative likelihood of progress toward and over the barrier. Here the rate enhancement was driven by both factors, but with a significantly larger effect from the probability factor and with contributions across much of the approach to the barrier, which suggests that greater reactivity was due to increased productivity at multiple stages of the reaction, including those after leaving the reactant well.

Contributions to the probability factor were further examined. Figure A-2A shows the cumulative logarithm of the probability factor as a function of reaction progress for test (red) and control (blue) simulations (essentially the probability that a trajectory that started toward the barrier will reach this value of λ). Figure A-2B shows the individual multiplicative contribution to the probability factor at each progress window (essentially the probability that a trajectory that made it through the previous window will continue through this window). The test simulations show much smaller decreases in reaction probability (Figure A-2A) and much larger contributions to reactivity (Figure A-2B) than the control simulations earlier in the reaction (below $\lambda =$ ˘0.4) but show similar behavior beyond that point (between $\lambda =$ ˘0.4 and 0.0). These data indicate a strong reactivity advantage of the constrained simulations (which was applied at the 0-fs time point, corresponding to a λ value of approximately –0.9 and well before the barrier) across the whole region from $\lambda =$ ˘0.9 through –0.4 but not past this point, noting that by $\lambda =$ ˘0.2 the reaction has essentially already occurred. This is consistent with a picture in which the constraint achieved its large gains in reactivity not by giving those simulations a local, near-term boost in reaction progress, but by directing them into channels that retained a continuous reactivity advantage.

2.5 Discussion

In this work, we find that features evident in the enzyme–substrate complex before it departs the reactant well are highly predictive of reactivity through the identification of relatively subtle conformational effects. These structural characteristics include internal substrate conformation, interactions of substrate with its environment, and details of the electronic environment of the two magnesium ions that coordinate the substrate. A consensus set of 30 features are predictive across the pre-launch window, although the detailed roles of some descriptors change across the window.

Interestingly, velocities are not needed to reliably distinguish reactive from nonreactive trajectories. This does not mean that velocities cannot also be useful or important, but only that conformations alone are sufficient. In fact, in preliminary work leading up to this study, we saw that velocities alone, without direct conformational measures, were also sufficient to distinguish reactive from almost-reactive trajectories. The top 20 velocity descriptors at the 0-fs time point are listed in Table B.4, together with their individual predictive performance. Five of these velocities are for atoms involved in the consensus geometry feature set, and thus may be indicating the same or similar drivers of reactivity. Furthermore, Figure A-3 compares the AUCs of the top 5, 10, 15, and 20 LASSO-selected features from (a) the set consisting of the 68 structural descriptors only (b) the velocity magnitudes of the 341 atoms within within 5 ångstroms of the migrating methyl, and c) the combined structural-velocity set of (a) and (b), showing that the combined set performs better than the structural or velocity set alone, but only by a very small margin. Together, Table B.4 and Figure A-3 suggest an effective overlap of information, in that different descriptors can be equally useful in understanding and predicting reactivity, perhaps through the same or similar explanations. The involvement of some of the same atoms, although possibly the result of there being a small reactive center, suggests that different classes of descriptors may be indicating the same fundamental chemical effects. Although the analysis in the current work appears static, relying on conformations evident at fixed points in time, this may implicitly contain dynamic information. For example, the 0-fs time point corresponds to the maximum compression of the breaking bond before the trajectory launches toward the activation barrier, and so a shorter bond distance, indicating greater potential energy stored in the bond, may signify greater kinetic energy available to surmount the barrier (and, indeed, the velocity of one atom in this bond, C4, was the second most predictive velocity feature).

We acknowledge that a more thorough description may be necessary to truly understand reactivity than to predict it. Whenever any fitting procedure is performed (as in this study), there is a danger of overfitting, but a number of lines of evidence suggest that overfitting is not responsible for the conclusions here. These include the vast overabundance of data points relative to number of parameters included in the fitting procedure, with data samples on the order of 500,000 reactive and almost-reactive samples each used to fit 31 parameters; the use of cross-validation, in which reported results are for testing data that is explicitly excluded from the fitting procedure (that used only training data); and the observation that the highly predictive features are also controlling, i.e. enforcing them enhances reactivity, even for ensembles seeded from trajectories not included in the training/testing data set.

This study presents evidence that there are multiple channels of reactivity, some of which are more productive than others. The existence of multiple reactive channels suggests that there are identifiably different reaction sub-pathways. Results further suggest that within each channel there could be more ways of not reacting than reacting, consistent with the notion that there are many conditions that must be met in order to produce a reactive trajectory, and failing to achieve any of multiple combinations of those features can be detrimental to reactivity. This study also highlights the important role that early active-site conformational effects play in driving chemical catalysis, an idea that underlies existing theories of the importance of early conformational effects such as electrostatic preorganization [115, 61] and enzyme-stabilized "near-attack conformations" in certain catalytic systems [7, 66]. That machine learning methods were able to identify early conformations predictive of reactivity lends additional support to the preorganization and near-attack conformation hypotheses of enzymatic activity, although further research would be necessary to determine whether electrostatic preorganization or stabilization of near-attack conformations is a primary driver of catalysis in the KARI isomerization reaction studied.

Although this study was purely computational, the results are supported by data available from the literature. A number of KARI mutants made characterized experimentally 44. In the closest correspondence with the present work, mutations have been made in the E. coli KARI variant (which exhibits 100% conservation of the 8 polar active site residues with the S. oleracea variant studied here), finding that mutations in positions corresponding to Asp315, Glu319, and Glu496 all reduce specific activity against 2-acetolactate by more than 200-fold 44. The relationship between these experimental results and our computational features is striking–two of the three instances of a debilitating mutation correspond to a residue involved in a feature in the top 30 consensus feature set. For example, Glu319 is involved in the top ranked feature, the Glu319/O$_{\epsilon1} - -AC6/C5$ distance, and Glu496 is involved in the 6th ranked feature, the Glu496/H$_{\epsilon2}$–AC6/O8 distance. Asp315 was not included in the features included for training and so could not appear in our consensus set.

In this work we also showed that that path-sampling techniques combined with QM/MM simulations can be used to generate valuable data sets that allow the question of reactivity to be phrased as a binary classification problem well suited for machine learning. We believe this represents both an exciting and promising application, but also a productive strategy for elucidating subtle yet meaningful drivers of catalysis in enzymatic systems. While this work utilized features selected through human intuition and a linear classification model (LASSO), the application of unsupervised learning techniques to identify perhaps better features combined with non-linear classification models represents an opportunity to understand further the early events that lead to enzymatic catalysis. Although this work utilized TIS to generate only two types of data sets, reactive and almost-reactive, TIS can also be used to generate many more types of data (for example, to generate sets of trajectories that reach progressively higher points along the barrier). Applying machine learning to trajectory outcomes representing more than two states of reactivity can potentially yield new insights as to precisely when and how reactive and non-reactive trajectories diverge. Although this study identified features indicative of reactivity, an understanding of how those structural and potentially electronic effects cooperate to facilitate the reaction is not obvious from structures alone. It is possible that more detailed quantum chemical analysis, perhaps with a focus on orbital behaviors, will lend more insight.

A difference between this work and prior studies of near-attack conformations is that we have defined reactivity at time points relative to the temporal progress of the prospective catalytic event rather than purely configurational states [96, 66, 58]. Although the sampling constraints during the TIS simulations were enforced at specific time points relative to the progress of the prospective catalytic event, e.g. the “last trough” that we have defined as the 0-fs time point in the reaction, future work is needed to test how critical the time point is on the effectiveness of the constraint in leading to more reactive trajectories. Initial results (unpublished) for sets of constrained TIS simulations in which a classifier was learned that was predictive of reactivity across the entire pre-launch window, suggests that reactive trajectories spend significantly more time in the reactive sub-region of the reactant well than almost-reactive trajectories. This result implies that constraints broadly applied across multiple early time points may be just as effective, if not more effective at enhancing reactivity than constraints applied at one specific time point.

The features identified are more than indicators that the reaction will likely occur; they are control levers that can guide and enhance reactivity. Our studies demonstrate that enforcing these indicators of reactivity leads to dramatic computed rate enhancements, largely by increasing the probability of trajectories reaching the product state. This enormous enhancement directly suggests an approach to re-engineering enzymes for enhanced specific activity. The results of this study suggest that the identification of mutants whose predominant effect is to selectively populate regimes identified as promoting reactivity for a set of geometric features could be a useful method of enhancing activity, by causing the enzyme-substrate complex to spend more time in highly-reactive conformations. Such mutations could be especially useful if they have minimal effects elsewhere on the reactive energy surface. For example, the first feature in Table 2.2 indicates that longer distances between Glu319/$O_{\epsilon 1}$ and substrate C5 lead to enhanced reactivity, and so identifying mutations that enlarge or pull back the pocket in which Glu319 sits could be useful. Likewise, the second feature indicates that shorter distances between magnesium M16 and substrate O3 lead to enhanced reactivity, and thus identifying mutations that alter the packing of the magnesium ligands to place it closer to the substrate might also be useful. Indeed, in other ways, several recent studies have attempted to leverage insights from path-sampling simulations in order to design enzyme variants [126, 54], which represents a promising and novel framework for biocatalyst design.

Table 2.2: Top 10 LASSO selected features at 0-fs time point and coefficients β_j used to define reactive region A' in constrained TIS simulations. Note that classification was performed on the fly through the TIS Markov chain and thus features were not normalized by Z-scores, so non-standardized coefficients β_j are reported. The bias β_0 used was -18.603.

j	Feature	β_j
1	Distance GLU‘319/O$_{\epsilon 1}$,AC6/C5	2.1944
2	Distance MG6/M16,AC6/O3	-12.093
3	Distance AC6/C1,AC6/C4	13.447
4	Distance AC6/C4,AC6/O6	20.561
5	Angle NDP/C4N,NDP/N1N,NDP/C1NQ	-2.8234
6	Distance AC6/O8,GLU‘496/H$_{\epsilon 2}$	-3.4298
7	Distance AC6/C5,AC6/C4	-8.8403
8	Distance AC6/O8,MG6/M17	8.8193
9	Dihedral AC6/C5,AC6/C4,AC6/C7,AC6/C9	-3.7307
10	Distance MG6/H28,AC6/O6	-0.5615

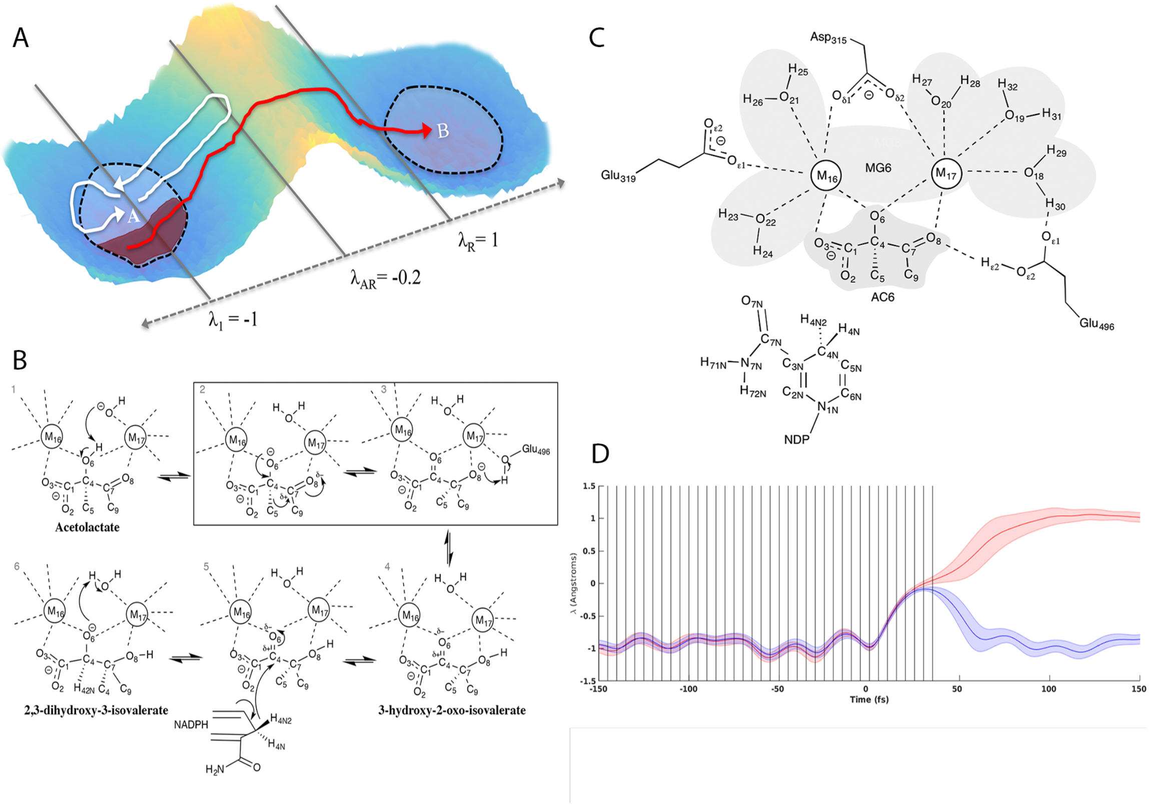


Figure 2-1: (A) Interface placements used to generate reactive and almost-reactive trajectories, where λA denotes the reactant interface, λAR indicates the product interface used to generate the almost-reactive trajectory ensembles and λR indicates the product interface used to generate the reactive trajectory ensembles. (B) Reaction catalyzed by KARI with states 2 and 3 indicating initial and final states used for the specific rate-limiting step of the isomerization studied (C) Atoms and residues included in QM region (non-polar hydrogens not shown) Note that the residue name AC6 is used in this study to refer to the reactant state of the substrate shown in Figure 1C. The residue name NDP refers to the NADPH cofactor and the residue name MG6 refers to the five quantum mechanically-treated waters and two magnesium ions in the active site. (D) Distribution of λ values for reactive (red) and almost-reactive (blue) trajectories time-shifted such that last trough before prospective catalytic event occurs at the 0 fs time point. Vertical lines indicate time points where features were computed.

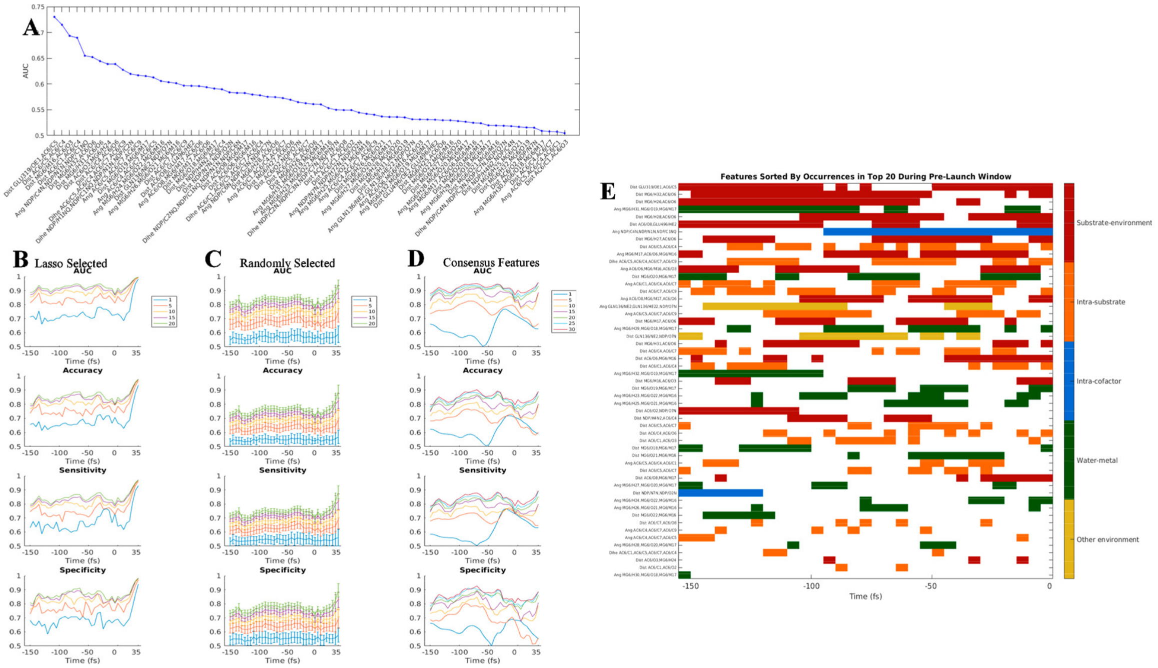


Figure 2-2: (A) AUC performance for all 68 individual features at the o-fs time point. Values of AUC shown represent the mean computed across 5 equal cross-validation training and testing partitions. (B) AUC, accuracy, sensitivity, and specificity for models with LASSO-selected features (C) AUC, accuracy, sensitivity, and specificity are plotted for models with randomly-selected features. (D) AUC, accuracy, sensitivity, and specificity are plotted for models with 30 consensus features. Error bars in (C) correspond to standard error of the mean across 100 randomly-selected feature sets. (E) Top 20 features selected by LASSO at each time point. Features are colored by feature type and sorted by the total number of occurrences in the top 20 between -150 and 0 fs.

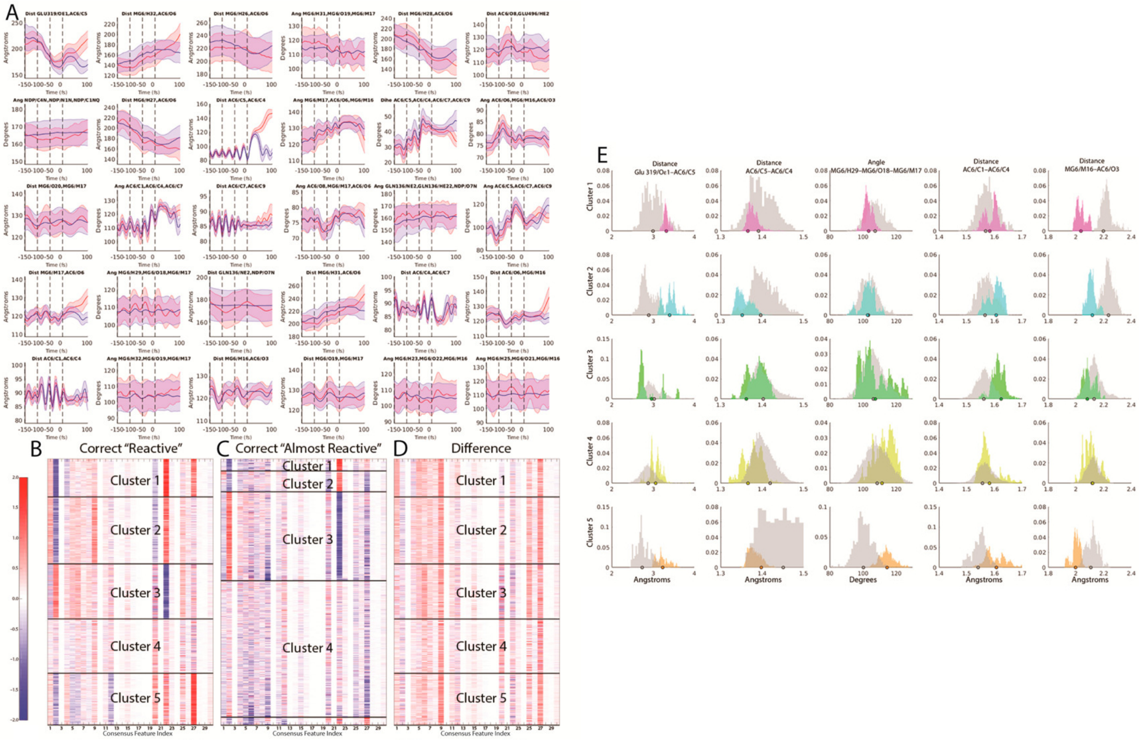


Figure 2-3: (A) Average time traces of consensus features across -150 to +100 fs time points with red indicating average reactive traces and blue indicating average almost-reactive traces. Error bars indicate 2 standard errors of the mean at each time point. Vertical black lines indicate time points at -150,-100,-50 and 0 fs where coefficients listed in Table 2.2 were fit. (B) Z-scores for consensus features (listed in Table B.1 and illustrated structurally in Figure 2-7) evaluated at the 0 fs time point and weighted by their corresponding standardized logistic regression coefficient for all correctly classified reactive trajectories in data set. Dark lines indicate cluster boundaries assigned using k-means clustering with $k = 5$. Within each cluster, features are sorted by distance from the centroid of the respective cluster (closest to centroid at top). (C) Z-scores for the consensus features evaluated at the 0 fs time point and multiplied by their corresponding standardized logistic regression coefficient for all correctly classified almost-reactive trajectories in data set. Dark lines indicate cluster assignments, based on the closest centroid to the five centroids learned on the reactive features shown in (B). (D) Z-scores differences between reactive features in each cluster and the mean almost-reactive feature set of the corresponding almost-reactive cluster. (E) Histograms of weighted feature weight differences across each of the five reactive / almost-reactive cluster sets. The set of five features shown was determined by computing the top three weighted feature differences by absolute value for each cluster shown in Figure 2-3D, then taking the union of the resulting set. Magenta corresponds to cluster 1, cyan corresponds to cluster 2, green corresponds to cluster 3, yellow corresponds to cluster 4, orange corresponds to cluster 5 and gray corresponds to the corresponding almost-reactive cluster for the reactive cluster shown in each histogram. Dots indicate representative structures (the reactive or almost-reactive structures closest to the mean of the centroid for each respective cluster) which are shown in Figure 2-8.

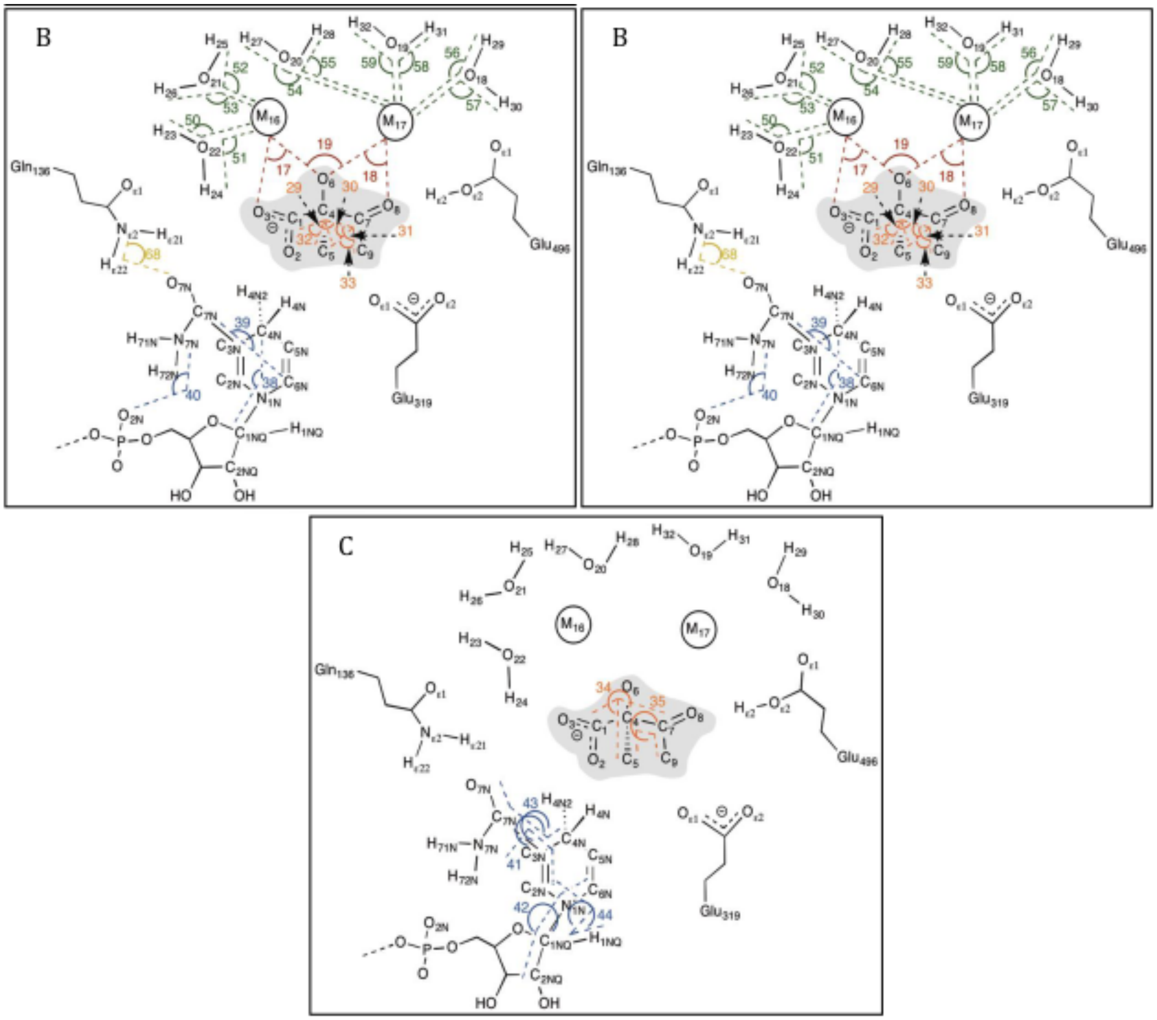


Figure 2-4: Structural representation of (A) distances computed, (B) angles computed, and (C) dihedrals computed at each time point. Numbering of features corresponds to that of Table S1. Coloring of features corresponds to the feature type with red indicating substrate-environment interactions, orange indicating intra-substrate conformations, blue indicating intra-cofactor conformations, green indicating water-metal interactions and gold indicating other environment interactions.

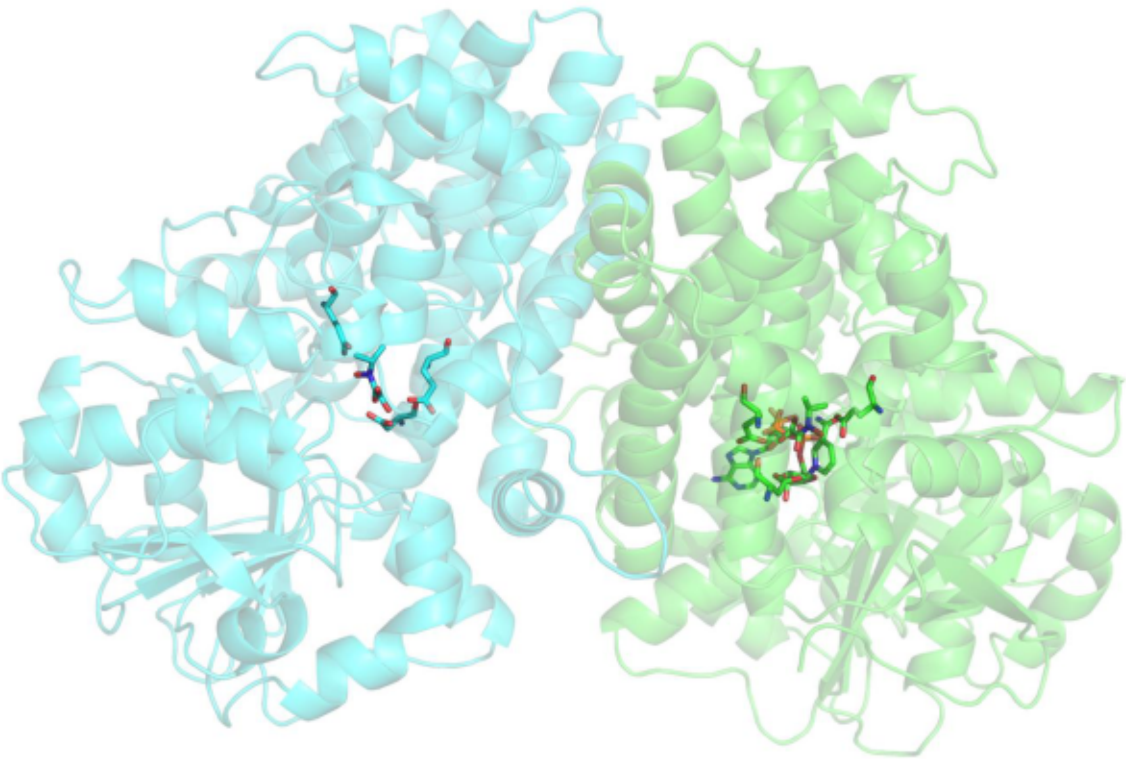


Figure 2-5: Illustration of both KARI homodimer subunits (PDB ID: 1YVE), with active site residues Asp315, Glu319, Glu496, bound transition state analog N-hydroxy-N-isopropyloxamate and NADPH cofactor shown as sticks to indicate active-site separation and to support the choice of using a single subunit in simulations.

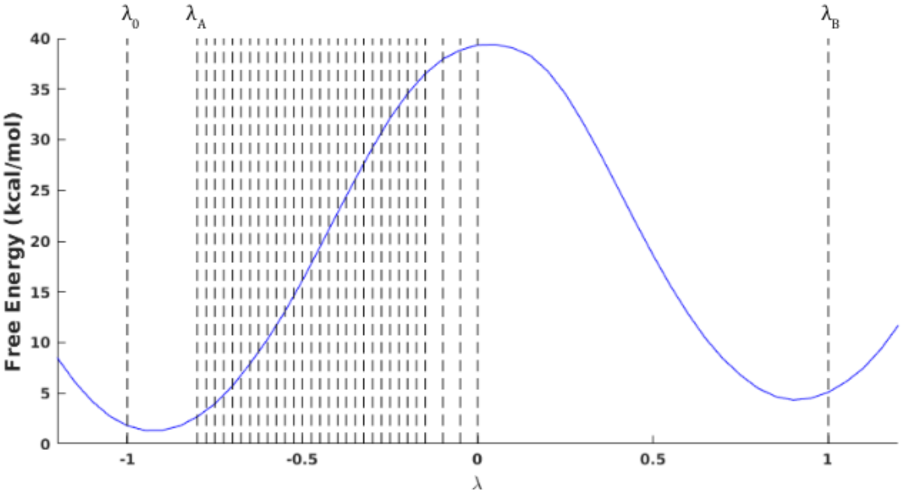


Figure 2-6: Placement of interfaces used in TIS probability factor calculations superimposed onto the potential of mean force surface used to generate initial seed trajectories. Key interfaces $\lambda_0 = -1$, $\lambda_A = -0.8$ and $\lambda_B = 1$ are labeled.

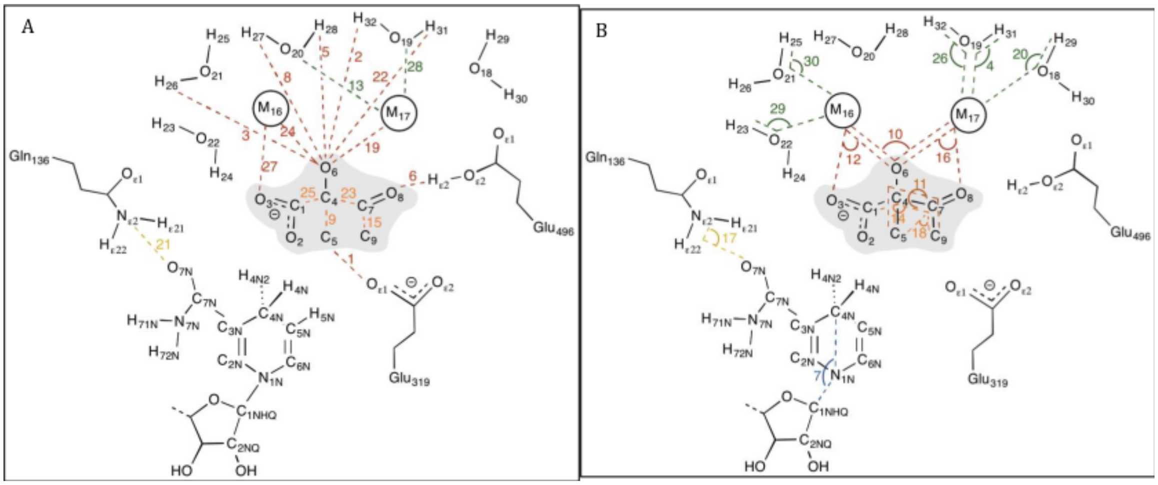


Figure 2-7: Structural representations of top 30 most consistently predictive (A) distances and (B) angles and dihedrals during the -150 to 0 fs time window. Labeling of features corresponds to ranking in Table B.1. Coloring of features corresponds to the feature type with red indicating substrate-environment interactions, orange indicating intra-substrate conformations, blue indicating intra-cofactor conformations, green indicating water-metal interactions and gold indicating other environment interactions.

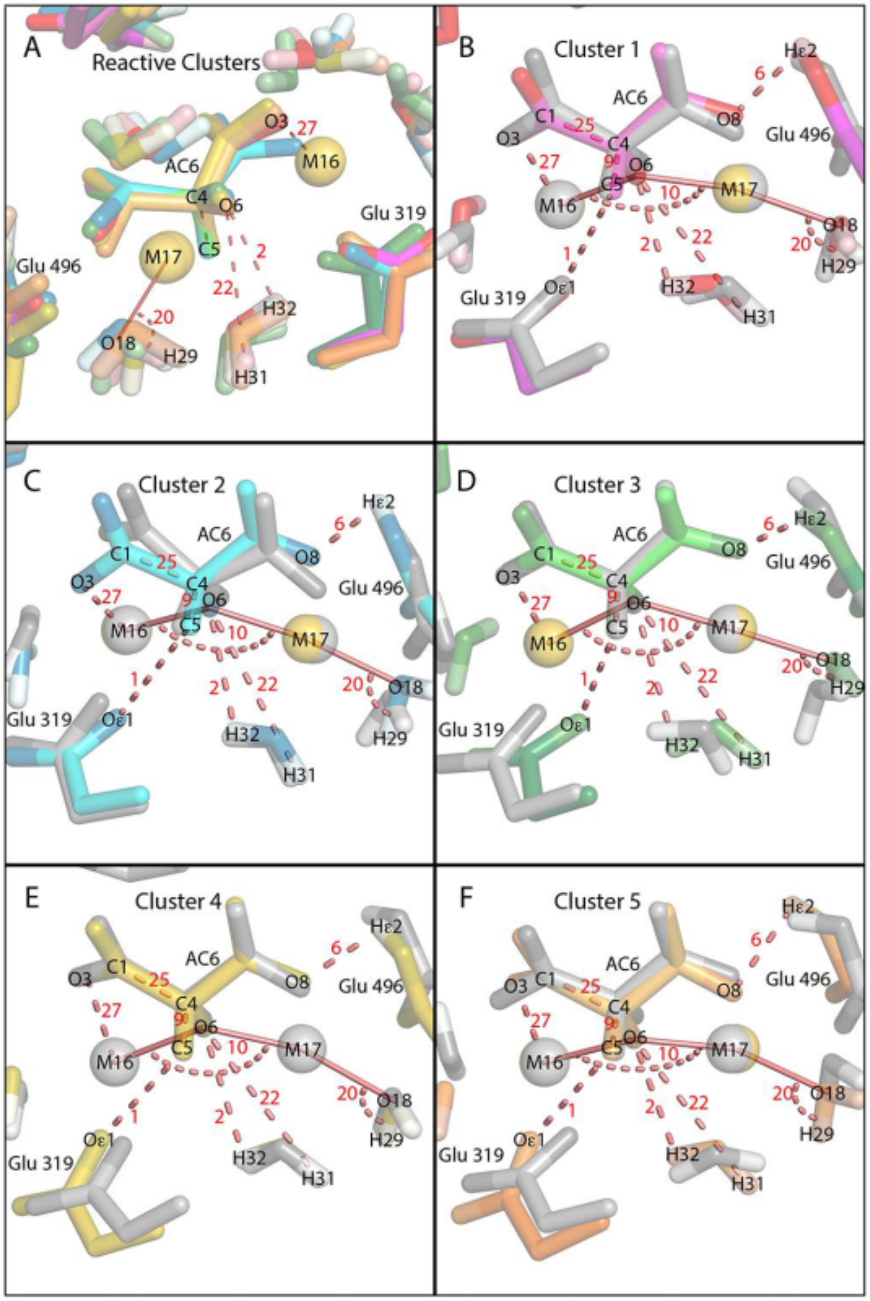


Figure 2-8: Representative structures for the reactive cluster and corresponding almost-reactive clusters described in Figure 2-3B-E. Feature numbering corresponds to that of Table B.1. (A) Representative structures from all five reactive clusters. Representative structures from (B) cluster 1, (C) cluster 2, (D) cluster 3, (E) cluster 4, (F) cluster (5) and their corresponding almost-reactive clusters, respectively. In all panels, magenta corresponds to cluster 1, cyan corresponds to cluster 2, green corresponds to cluster 3, yellow corresponds to cluster 4, orange corresponds to cluster 5 and gray corresponds to the corresponding almost-reactive cluster for the reactive cluster shown in each histogram. In all panels, structures were aligned to minimize the root mean square difference between the two magnesium centers.

He is a man of intelligence, but to act sensibly, intelligence is not enough.

Crime and Punishment

3

Integrating Artificial Intelligence Methodologies into the Academic Publishing Ecosystem

The work in this chapter is adapted with permission from the chapter *Artificial Intelligence in the Academic Publishing Ecosystem* written by James W. Weis and Amy Brand which has been accepted for publication by the *Charleston Briefings*.

J.W.W. and A.B. conceived of the project. J.W.W. conducted the relevant literature review. J.W.W. and A.B. wrote the manuscript,

3.1 Abstract

Technology underpins all aspects of today's academic publishing ecosystem. But only recently have AI-based platforms—systems that employ learning and other human-like or rule-based behavior—begun to affect scholarly publishing. However, the vast amounts of digital content and data output by the publishing ecosystem provide fertile ground for the application of artificial intelligence (AI) and machine learning (ML) methodologies to the production and consumption of research content. On the one hand, data-driven algorithms can provide consumers of academic literature with more efficient ways of identifying and—subsequently—assessing the value and relevance of research content via smarter search, better disambiguation, and improved metrics. On the other hand, academic publishers can use AI techniques to assist in intelligent targeting and curation of research, such as post-publication impact measures, sentiment analysis, and improved methods of plagiarism detection and reproducibility testing.

3.2 Introduction

Those of us who have chosen to devote our working lives to scholarly communication are driven by a desire to accelerate the path from research breakthrough to application and societal benefit. Yet despite the huge advances in digital publishing and research technologies of recent years, how academics produce and consume peer-reviewed scholarship is unchanged from the print era in fundamental ways. A key reason for this is the interdependence of publishing and career advancement in academia, and the ways in which the customs of the latter stifle change in the former. As a result, academic publishing practices have so far failed to take robust advantage of today's information technologies, let alone AI-based computational methods. But, little by little, publishers and aggregators are embracing new metrics, new navigation tools, and smarter approaches to content review and curation. In this chapter, we survey AI-informed developments and opportunities in each of these areas of scholarly publishing, taking care to distinguish true AI-driven approaches—systems that employ learning and other humanlike or rule-based behavior—from other computational methods.

3.3 Smarter Metrics

Prior to the mid-17th century, scientific communication was comprised largely of personal letters between practitioners. *Philosophical Transactions of the Royal Society*, established in 1665, is widely considered to be the world's first scientific journal, underpinning a step-function of ongoing refinements in scientific publishing that continues to the present. Peer and editorial review, as well as specialized journals, emerged early on as filters that allowed the growing scientific community to extract meaningful insights more easily from the increasingly broad and rapid pace of scientific research.

In this new paradigm, research impact was roughly quantified by publication output and citation-based metrics [30]. For individual papers, the total number of citations was (and continues to be) the most frequently used quantification of importance. For journals, Eugene Garfield proposed the Journal Impact Factor, a journal-level measure of per-article citation rates, which became widely adopted [46, 47]. For researchers, Hirsch's h-index attempted to "quantify the cumulative impact and relevance of an individual's scientific research output" [56]. As the dimensions of the scientific literature continued to grow, making personal digestion of all the literature in any field near impossible, administrators began to rely on these metrics for assessment of scientific impact–which, consequently, made them targets for researchers, who need to demonstrate scientific impact to be hired or promoted.

The transition from print-based to web-based scientific correspondences has contributed to further explosive growth in the breadth and scope of academic communication. This expansion includes not only web-replicates of traditional journal-based, peer reviewed research articles, but also a variety of new media, including preprint servers, blogs, and even social media conversations. As a consequence, traditional citation-based impact metrics have become increasingly poor (and manipulated) proxies for actual academic impact [117, 42].

While the large amount of data produced by the modern research ecosystem may diminish the value of the metrics in most widespread use today, it also provides an exciting area for future innovation. Scientific communications contain valuable information in the form of text, images, and data files, and are also are linked to one other via citations and on social media. At the same time, publishing in web-native formats and in open access journals (where the results are not locked behind a paywall, and thus are freely available) is becoming increasingly standard. The combination of these trends makes it possible to develop, track, and potentially even predict, new, nuanced, and targeted metrics of scientific impact.

It is easy to imagine a near-future in which scientific impact is quantified by algorithms that, rather than simply counting citations, traverse the full body of scientific literature to extract more nuanced measures. Google's PageRank algorithm, for example, rose to dominance in part because it ranks web pages not by the simple count of references, but by weighting each reference by the relative importance of the corresponding webpage—references from highly-ranked websites, like trusted news outlets, therefore score significantly higher than references from lesser-known pages, such as personal blogs. Similar methods can be deployed on the scientific literature, and would thus compute not only the number of citations, but also the contribution of individual co-authors and the authority of each citing body. This kind of approach, which has recently been shown to most other impact metrics in publication ranking challenges, gives greater weight to citations that come from important or impactful sources, calculated by the same algorithm in a recursive fashion [104, 76, 63, 121].

Similar computational methods could then be extended to calculate different dimensions of scientific impact, such as measures of novelty, collaboration, diversity, or interdisciplinarity. Further, as discussed in the next section, the algorithms could be adjusted based on an individual's publication history or stated interests, to suggest the literature, or collaborators, in a tunable way–even recommending the work of highest relevance or optimizing for cross-field insights. In fact, such work is already well underway. Network-based approaches have been used not only to quantify long-term scientific impact, but also to identify scientific "gems," measure technological innovation, and quantify the disruptiveness of new work [114, 45, 27]. These methods enable more nuanced, granular exploration of the scientific literature, and also facilitate new types of research; for example, Wu, Wang, and Evans recently found that there are observable differences in the types of innovations produced by large and small teams, with larger teams tending to build on previous work, and smaller teams more likely to develop disruptive ideas and technologies [119].

The new metrics resulting from the application of AI-based methods to the academic publishing ecosystem are transformative. By incorporating more data in the construction of impact scores, and by defining impact in a more nuanced and multifaceted manner, we will be far more accurate in judging the relevance of research, leading to more efficient hiring and promotion decisions, and thus a more meritocratic scientific ecosystem. Furthermore, the development of these methods will catalyze new research in the science of scientific research and development, with broad implications not only for academic career advancement, but also for scientific funding and the optimization of scientific resource allocation more generally.

3.4 Smarter Search

The changing landscape of academic publishing, and especially the increasing scale and speed of research output, has thrust search and information retrieval into a position of unique importance. Just as increasing scale and complexity in the nascent World Wide Web led to a transition from the Yahoo! index-driven portal page, which was organized around human-curated keywords, to the search-centric Google model, which leveraged computational techniques to identify the most relevant results, so are academic literature searchers increasingly relying on complex information retrieval algorithms. These algorithms currently allow rapid keyword-based retrieval of academic articles, ranked by different parameters. And future models will expand beyond keywords to include smarter metrics and personalized algorithms. In the future, these models could capture enough patterns from the history of scientific research and development to potentially even assist in the creation and assessment of more unique, impactful, and testable scientific hypothesis [?].

Currently, many researchers rely on indexes like PubMed and Google Scholar to find articles of interest. However, such methodologies are inexact, and keywords alone are often insufficient to balance the sensitivity and specificity necessary for a successful search[51]. Methods borrowed from branches of AI that deal with both information retrieval and natural language processing (NLP) will allow more granular and customized searches. These algorithms could improve keyword search by, for example, combining it with the searcher's publications, academic co-authorship network and social media connections, and previous search history to rank results in a personalized manner [6, 60]. In combination, topic model-based approaches could use similar papers, rather than keywords, as the inputs to search queries. Companies like Yewno are already exploring the real-world implementation commercialization of such approaches, which augments users ability to navigate search results by providing them with computationally-augmented search and filtering mechanisms—for example, allowing search by concept or topic, rather than searching for specific words [51, 36].

Another promising area is the incorporation of computational reasoning into the search process. Future search engines could be designed to return publications most likely to spur disruptive thinking or impactful collaborations, rather than the academic papers most similar to the input keywords. These search algorithms could then be fine-tuned, for example by broadening or narrowing scope, by filtering out content with specific attributes, or by incorporating work that researchers with similar profiles have found valuable. Even more powerfully, computational parsing of natural language present in scientific articles could be used to auto-construct ontologies of scientific thought which, when used in combination with social networks, would identify the specific parts of research articles that support (or refute) specific queries, and potentially even encourage exposure to alternative viewpoints, account for cultural biases, or identify social dependencies in reasoning [38, 123].

3.5 Smarter Curation

The move towards faster publication cycles and increasingly interdisciplinary research puts strain on the time-tested model of academic peer review prior to publication. At the same time, the growing adoption of preprint servers like arXiv and bioRxiv better aligns with a post-publication peer review model, which many have argued results in more efficient and less bias-prone curation [**?**]. While some fields, like mathematics and physics, have relied on public discourse around research findings for some time, other disciplines are finding themselves thrust into a form of online, crowdsourced peer review spanning many sites and applications, from blog posts to Twitter conversations [72].

AI-based methods can be used to help organize, structure, and extract value from the many diverse conversations circulating online around academic research. They can also help us make the best possible use of high-quality peer review, which should be considered a sparse, valuable resource. For example, sentiment analysis, a method from NLP that allows the positive or negative valence of a comment to be quantified, could be used to extract more nuanced meaning from Tweets, blogs, and comments. Then, prediction models that take the quantity and content of online conversations into account could be used to determine when a preprint article has garnered sufficient positive attention to merit peer review. Learning algorithms could even be used to quantify the quality of reviews, so that the peer reviewers with the best "track record" or history of correctly identifying positive or negative indicators in specific fields are allocated to the papers in most need of their skills.

Finally, AI algorithms can also be deployed to increase reproducibility in science. Existing NLP methods are increasingly being leveraged to compare texts for similarities with proceeding work in an author-specific manner, and further research in this vein is current and ongoing [23–25]. Machine vision algorithms, including Convolutional Neural Networks (CNNs), could be used to detect image and figure reuse at broad scale [26], and network-based approaches could be leveraged in combination with large databases of retracted papers, such as Retraction Watch Database, to identify warning signs of irreproducibility [94].

3.6 Conclusion

Artificial intelligence and machine learning promise huge near-term advances in our ability to intelligize and accelerate the academic research and communication process. In this brief chapter, we have described a subset of these application domains in order to illustrate this tremendous potential.

You must be ready to burn yourself in your own flame;
how could you rise anew if you have not first become ashes?

Thus Spoke Zarathustra

4

New Strategies for Designing Knowledge Infrastructure

The work in this chapter is adapted with permission from the manuscript *The Case for an Institutionally Owned Knowledge Infrastructure* written by James W. Weis, Amy Brand, and Joi Ito and published in *Inside Higher Ed* on January 7, 2020.

J.I. and A.B. conceived of the project. J.W.W. conducted the relevant literature review. J.W.W., J.I., and A.B. wrote the manuscript.

4.1 Abstract

Science and technology are propelled forward by the sharing of knowledge. Yet, despite its critical importance in today's innovation-driven economy, our knowledge infrastructures have failed to scale with today's rapid pace of research and discovery. For example, academic journals, the dominant scientific knowledge dissemination platform, have not been able to take advantage of the linking, transparency, dynamic communication, and decentralized authority and review enabled by the Internet. Many other knowledge-driven sectors, from journalism to law, suffer from a similar bottleneck—caused not by a lack of technological capacity, but by an inability to design and implement efficient, open, and trustworthy mechanisms of information dissemination. Fortunately, growing dissatisfaction with current knowledge sharing infrastructures have led to a more nuanced understanding of the requisite features such platforms must provide. These lessons can be leveraged by organizations around the world to begin recapturing control, and increasing the utility, of the knowledge they produce.

4.2 Introduction

When the World Wide Web emerged in the 1990's, an era of robust scholarship based on open sharing of scientific advancements appeared inevitable. The Internet—initially a research network—promised a democratization of science, universal access to the academic literature, and a new form of open publishing that supported the discovery and re-use of knowledge artifacts on a global scale. Unfortunately, this promise was never realized. Universities, researchers, and funding agencies failed to organize and secure the investment necessary to build scalable knowledge infrastructures, and publishing corporations moved in to solidify their position as the purveyors of knowledge.

In the subsequent decade, these publishers consolidated their hold. By controlling the most prestigious journals, they were able to charge for access—extracting billions of dollars in subscription fees while barring much of the world from the academic literature. Publishers like Elsevier (the science, technology, and medicine-focused branch of the RELX Group conglomerate) have 36.7% profit margins—higher than Apple, Google/Alphabet, or Microsoft [79, 93, 5, 3]. This structure has reached such fantastic proportions that some of the world's wealthiest academic institutions are no longer able or willing to pay the subscription costs required [5]. Further, by controlling many of the most prestigious journals, publishers are also able to position themselves between the creation and consumption of research, and so wield enormous power over peer review and metrics of scientific impact—and thus academic reputation, hirings, promotions, career progressions, and ultimately, the direction of science itself.

However, there are signs that the bright future envisioned in the early days of the Internet is still within reach. Increasing awareness of, and dissatisfaction with, the many bottlenecks imposed by the commercial monopoly on research information is fermenting interest in, and development of, new strategies for the development of the future's knowledge infrastructures. One of the most promising developments is the shift towards infrastructures developed and supported by academic institutions—the original creators of the information being shared—and aligned non-profit consortia such as the Collaborative Knowledge Foundation and the Center for Open Science.

In this article, we first review three components of today's knowledge infrastructure that we believe are in critical need of modernization. We then propose a new, general model for the development, and scaling of academically-owned knowledge infrastructure with wide applicability. Finally, we present a partnership between the MIT Media Lab and the MIT Press, which constitutes an instantiation of this model and briefly review the projects this partnership has enabled.

4.3 Critical Features Of A Knowledge Ecosystem

We believe an institutionally-owned knowledge infrastructure should fully exploit the technological capabilities of the web to accelerate discovery, research funding, and the structuring and transmission of knowledge. By aligning academic incentives with socially beneficial outcomes, such a system could enrich the public while also amplifying the technological and societal impact of investment in research and innovation. The three areas in which we believe a shift to an academically-owned platforms would yield the highest impact on investment are (1) truly open access to the academic literature, (2) meaningful article-level impact metrics and (3) a trustworthy and bias-free system peer review platform.

4.3.1 Truly Open Access

The movement to a digital medium enabled a decomposition of the previously black-box academic publication process into its component parts–including peer review, copy editing, and design. The Open Access (OA) movement, which aims to make scholarly literature freely available online, began as a response to this potential. Initially focused on self-archiving, or "Green OA," researchers began making their results easily and freely accessible by uploading pre-publication manuscripts to university based institutional repositories and services like arXiv.org. The repository movement began gaining steam in earnest when Harvard established the first U.S. self-archiving policy in 2008 [48, **?**, 18]. Other research universities around the world quickly followed, along with bioRxiv and other field-specific preprint servers [25, 14].

However, OA and institutional repositories never realized their potential to transform research communication. Not only did investment fall short of the funds necessary to support the development of scalable platforms, but commercial publishers successfully circumvented the movement by revising licenses to block or delay self-archiving, creating pay-to-publish or "Gold OA" journals, and launching analytics and research workflow services. This clever divide-and-conquer strategy successfully stymied collaboration on OA and academic infrastructure development.

One possible reaction to this subversion of the OA movement is to pressure publishers to lower fees. So far, such collective bargaining has been successfully blocked through confidentiality agreements and other legal means [109, 107]. On the one hand, there are signs that this is changing: Plan S, for example, is an OA-focused initiative supported by a coalition of roughly a dozen leading European research funders responsible for €17.6 billion (US$8.8 billion) of funding a year that was launched in 2018 and will go live in 2020 [39, 2, 19]. On the other hand, as long as the underlying infrastructure, including the key journals, remain in the control of the publishers, they can always simply extract fees elsewhere or monetize other parts of the research pipeline. Making matters worse, the OA movement has been further undermined by the emergence of predatory OA journals, which have little-to-no quality control or peer review, and often target scholars from developing countries [73, 62, 120].

One route to lower publishing costs is the unbundling of publisher services, and charges that they accurately reflect value-added work, while also compressing margins to universally-accessible levels. In some ways, this solution could resemble the transition from online publishing to blogging. Before blogging platforms, large software companies charged millions of dollars for Content Management Systems (CMSs), which are still used in complicated professional settings. However, it turned out that simple scripts, free and open source software, and open standards to interoperate between services allowed the creation of simple and extremely low-cost publishing platforms–leading to the emergence of "user generated content" and what has now become social media. While academic publishing is much more complicated, a refactoring and an overhaul of the software, protocols, processes and business underlying academic publishing could revolutionize it both financially as well as structurally—allowing sustainable, universal open access publishing without paywalls.

4.3.2 Meaningful Impact Metrics

Typically, researchers must publish impactful work to further their careers. While research should ideally be judged on its individual merits, the current paradigm relies heavily on the prestige of the journal in which the research was accepted as a heuristic for importance. Because a handful of commercial entities control these "highly impactful" journals, they are able to leverage the academic reputation systems in the reinforcement of this journal-based status quo.

A consequence of this system is the often-referenced "Impact Factor" of a journal, which is supposed to indicate the impact or quality of the research that journal accepts for publication. The impact factor of a journal in year y is the average number of citations C per article A the journal received over the past two years:

$$IF_y = \frac{\sum_{i=y-2}^{y-1} C_i}{\sum_{i=y-2}^{y-1} A_i}$$

The pervasive Impact Factor is known to not only be a poor proxy for research quality, but also to be easily gamed by "citation cartels," coercive self-citations, and other well documented strategies [32, 41, 117, 42]. Despite this, Impact Factor has a significant, and self-fulfilling, impact on the hiring and promotion of researchers. The committee members making these decisions often evaluate a candidate based on the prestige of the journal in which their research has been published. So, young researchers on the tenure path are generally forced to prioritize publishing in journals with high Impact Factors, faulty as the metric is. As a result, the corporate grip on our knowledge infrastructure strengthens, and important work ends up behind paywalls and largely inaccessible to anyone outside a major university or research laboratory.

4.3.3 Trustworthy Peer Review

Peer review, wherein area experts evaluate new research, is a critical component of the academic ecosystem–and one that is currently managed largely by publishers as part of the process of accepting or rejecting a new finding for publication. Currently, most articles are reviewed by an anonymous panel of peer reviewers, and some journals require double-blind review in an attempt to combat bias.

This process is broken in many ways. For one thing, many papers are already published on archive and preprint servers, so in the case of double-blind review it is trivial for reviewers to find the authors of a paper on the Internet. This not only obviates double-blind review, but serves to reinforce tribal biases and affiliations. Further, there is evidence that reviewers are not able to consistently and accurately judge the quality of new ideas, and typically discount the value of novel ideas, as well as work in their own fields. For example, while publishing in the top five economics journals is very important for tenure decisions in the same field, those journals tend to be conservative and resistant to publishing novel ideas [17, 55]. Also, it is not clear that reviewers are sufficiently incentivized to pay attention. While conducting peer review was traditionally viewed as part of one's academic obligations, busy researchers are increasingly less willing to devote significant time to reviewing research for publishers that benefit from their donated time. As an extreme example, in 2014, academic publishers Springer and Institute of Electrical and Electronic Engineers (IEEE) had to remove more than 120 computer-generated papers with meaningless content from their subscription services [112].

Perhaps we can improve peer review, as with other aspects of publishing, by taking inspiration from technological developments outside the traditional academic publishing domain. In fact, emergent review via social media and blogs are already a force in many scientific disciplines, with open, unsolicited reviews often appearing within hours of publication. Efforts to capture, organize, and structure these diverse sources of feedback, such as Faculty of 1000, are motivational, but have a limited audience [72, 39]. Constructing schemas that provide academic credit to reviewers, such as the CRediT taxonomy, is one promising way of incentivizing review and thus scaling these alternative yet valuable sources of post-publication review [2, 19].

As a related example, the streaming music website thesixtyone (t61) encouraged musicians to upload their compositions, and allowed listeners to distribute a limited number of "hearts" to songs they liked [97]. Users that gave positive reviews to songs that later became hits were awarded with more hearts–allowing the best judges to have increasing weight in future assessments. Given the millions of active graduate students, post-docs, and other regular consumers of academic literature, creating a system that rewards people for looking for, finding, and betting some reputation on unverified new works–similar to a sport scout or early-stage venture capitalist–seems promising. Additional layers could then be added; for example, resources could be added to incentivize reviews of traditionally overlooked research to combat biases in the current system.

If we can indeed make the work of peer review more about looking for and rewarding new and novel ideas, instead of a system that reinforces the tribal networks and biases of academia, we can substantially improve the progress of scholarship while making it more equitable and available to the world at large.

4.4 Towards New Knowledge Infrastructures

4.4.1 Institutional Incubation Of New Knowledge Platforms

Successful revamping of the current ecosystem, including not only the features discussed above but also the multitude of other components of an efficient knowledge sharing pipeline, is likely not possible with a handful of highly profitable commercial entities in control. For such a paradigm shift to occur, we believe universities need to assert some ownership over the mechanisms for knowledge production sharing.

Such an effort presupposes expertise from both the production and communication perspectives, and could be achieved through partnerships between knowledge-producing organizations, such as research laboratories, and mission-aligned information-disseminating institutions, such as university publishing houses. These partnerships can build on existing resources, brand recognition, trust, and networks of talent and capital, to facilitate the incubation of new knowledge infrastructures and related projects. Further, the new organizations formed by these partnerships can work together to create inter-organizational consortia—via which information can be exchanged and the most successful incubated projects and frameworks can organically grow. This model is both general and scalable, and could be deployed and replicated at scale.

4.4.2 The Knowledge Futures Group

To make this vision a reality, the MIT Press and MIT Media Lab have recently launched a collaboration called The Knowledge Futures Group (KFG) [118, 105]. The KFG is the first partnership between a pedigreed publisher and a world-class research lab with a focus on developing and deploying next-generation technologies.

By serving as an incubator for publication-related projects, the KFG aims to (1) support projects that enrich the knowledge infrastructure, and (2) spark a movement towards greater institutional investment in, and ownership of, that infrastructure.

For example, the KFG is developing a new, open source publishing platform called PubPub, which uses a simple graphical format and supports both programmatic illustrations and text as well as static PDFs. The goal of the PubPub project is to create an author-driven alternative to academic journals that is tuned to the dynamic nature of many of our modern experiments and discoveries. Also being developed with the KFG is Underlay, a global, distributed method of linking and understanding public knowledge, which will make the data and content hosted on PubPub available to other platforms [91, 108]. These platforms can be used to experiment with transparent, bias-adjusted peer review and to build on previous work in the implementation of credit allocation frameworks [2, 19, 80].

Additionally, the KFG is supporting the development of new platforms for the calculation and sharing of more rigorous, article- and researcher-level metrics of scientific impact. By combining these metrics with machine learning, it is possible to gain insight into the trends and features that lead to impactful ideas. These predictions can also be used in the construction of quantitative, data-driven frameworks for the allocation of resources to research projects in an impact-maximizing way–and we are exploring opportunities to pilot these new funding mechanisms in the real world.

4.5 Conclusion

If we are to realize the transformative promise of the Web for science and scholarship, the control of knowledge infrastructure needs to transition from a commercial oligopoly to academically-owned and managed partnerships. For this to occur, it's essential that universities continue to assert greater control over systems for knowledge representation, dissemination, and preservation. This will require not only building new open source tools and protocols, but also collectively aiming to construct new platforms for peer review, attribution, and impact tracking that actively reward new, novel, and high-quality ideas.

Through the construction of these academically-driven partnerships, we can leverage the continually growing ecosystem of open source tools to develop, test, and deploy new, open, transparent, and cost-effective systems and processes that will help researchers and organizations. This will enable a shift towards greater institutional and public ownership of the platforms underlying the dissemination of knowledge–and the recapturing of the territory lost to publishers and commercial technology providers in the past decades.

What constitutes knowledge, the use of knowledge, and the funding of knowledge is integrally intertwined with the future of our planet and our species, and it must be actively protected from purely market-driven incentives and other corrupting forces. The transformation will require a movement involving a global network of collaborators, and we hope to contribute to catalyzing it.

So use all that is called Fortune. Most men gamble with her, and gain all, and lose all, as her wheel rolls. But do thou leave as unlawful these winnings, and deal with Cause and Effect. . . In the Will work and acquire, and thou hast chained the wheel of Chance, and shalt sit hereafter out of fear from her rotations.

Self-Reliance

5

Learning on Knowledge Graph Dynamics Provides Early Warning of Impactful Research

The work in this chapter is adapted with permission from the manuscript *Learning on knowledge graph dynamics provides early warning of impactful research* written by James W. Weis and Joseph M. Jacobson and currently under review in *Nature Biotechnology.*

J.W.W. and J.M.J. conceived the study. J.W.W. performed the data structuring, algorithm design, and computational implementation. J.W.W. and J.M.J. drafted the manuscript and figures. J.M.J. supported and supervised the project.

5.1 Abstract

In a world of rapidly expanding science and limited resources, the identification of promising scientific research is critically important—and forms the foundation for much of the scientific enterprise, from grant funding to faculty hiring. Here, we present DELPHI, a machine learning framework that provides an early-warning signal for impactful new research by learning high-dimensional, time-based patterns from the historical biotechnology publication network. DELPHI identifies high-impact publications with superior precision and recall than is possible using citations or other recently described metrics—including identifying twice as many high-impact publications for relevant precisions than is possible using citations alone—and successfully predicts seminal biotechnology innovations in a blind back-testing. We prospectively identify a list of recent publications that we expect to be of high future impact, and discuss the utilization of DELPHI in the machine-augmented design of scientific funding portfolios—with the overarching goal of increasing the scientific return on resources deployed.

5.2 Introduction

The efficient progression of the scientific enterprise depends largely on our collective ability to optimally allocate resources across a set of promising researchers and projects. This process relies, in turn, on the identification of valuable potential research contributions–both directly, via the direct distribution of government, philanthropic and for-profit capital, and indirectly, via hiring choices, promotion decisions, and research publication. The proliferation of the digital scientific corpus, both in terms of size and medium, is enabling the development of new, data-driven methods to assist us in the optimization of project identification, funding, and commercialization–with the ultimate goal of producing with a higher scientific return on the resources deployed.

Currently, the scientific ecosystem relies heavily on citation-based metrics: citation count, h-index, and Journal Impact Factor. However, as has been discussed extensively in previous literature, these metrics are imperfect, inconsistent, and manipulatable measures of quality, and their utilization can lead to suboptimal decisions in academic hiring, promotion, and funding[77, 86, 117, 98]. For-profit funding for scientific entrepreneurs is analogously susceptible to biases[31, 50, 82]. In both cases, the application of methods from artificial intelligence to the vast amount of data produced by the modern scientific enterprise could provide new, deeper signals of scientific impact and innovation–allowing us to, in a machine-assisted manner, learn from the history of science to proactively design improved research and funding strategies.

Such data-driven algorithms would digest the broad range of high-dimensional digital scientific information currently available to produce meaningful, lower-dimensional signals that can then be combined with human expertise and intuition. Further, such approaches could incorporate multiple objective functions, and thus be extensible across a range of desired outcomes–for example, while granting agencies may want to maximize scientific innovation in a field of interest, venture capitalists may prefer to find a single innovation with maximum market value. In either case, given the current reality of expanding science and limited resources, these learned algorithmic signals could help us progress beyond simple citation-based measures of impact and better guide attention, funding, and investment to the right places

Recent work has demonstrated the value of extracting valuable signals from knowledge graphs. Adjusted graph centrality measures outperform citations-based measures in their ability to quantify the relative impact of scientific work, and network science-based methods enable the quantification of innovation and technological novelty[45, 74]. Recently, the application of these methods has enabled insights into team design, scientific prize networks, and the dynamics of scientific careers[119, 70, 11]. Similarly, earlier work has attempted to predict citation-based metrics from academic publication history[1, 44, 116]. However, to-date there does not exist a framework unifying these approaches with methods from artificial intelligence–allowing us to learn from the past to improve our ability to identify and fund the most impactful science and technology of the future.

Here, we describe DELPHI (**D**ynamic **E**arly-warning by **L**earning to **P**redict **H**igh **I**mpact), a framework that provides an early-warning signal for impactful new research and technology by autonomously learning high-dimensional relationships between a range of features calculated across time from the scientific literature. We prototype this framework and deduce its performance and scaling properties on time-series graphs of biotechnology-related publications, including over 7.8 million individual nodes, 201 million inter-edge relationships, and over 3.8 billion calculated metrics. We demonstrate the performance of our framework across a range of scenarios, show the correct identification of seminal biotechnologies via a blind retrospective study, and exhibit it's utility in proactively identifying important research. Finally, we conclude by discussing the potential incorporation of DELPHI into the design of diversified, impact-optimized scientific funding portfolios.

5.3 Results

5.3.1 Structuring And Computing On Knowledge Graph Dynamics

As the size of the scientific corpus continues to grow rapidly, the identification of important research becomes an increasingly important problem. While the field of bibliometrics has produced a plethora of ranking methodologies, more recent work has shown that the integration of both network structure and temporal dynamics are important in discovering work that produces lasting impact[113, 75]. To capture the relationships between temporal patterns in both knowledge network structure and content, we collected metadata for articles published in 42 of the top biotechnology journals (Table 5.1). We designed a heterogenous graph database schema (Figure 5-1b) to allow for the aggregation of this information across multiple entity types (including publications, authors, and journals) in a time-dependent manner, and calculated a variety of metrics (Table 5.2) for each entity over time. To better capture the diverse array of interactions underlying the production of impactful

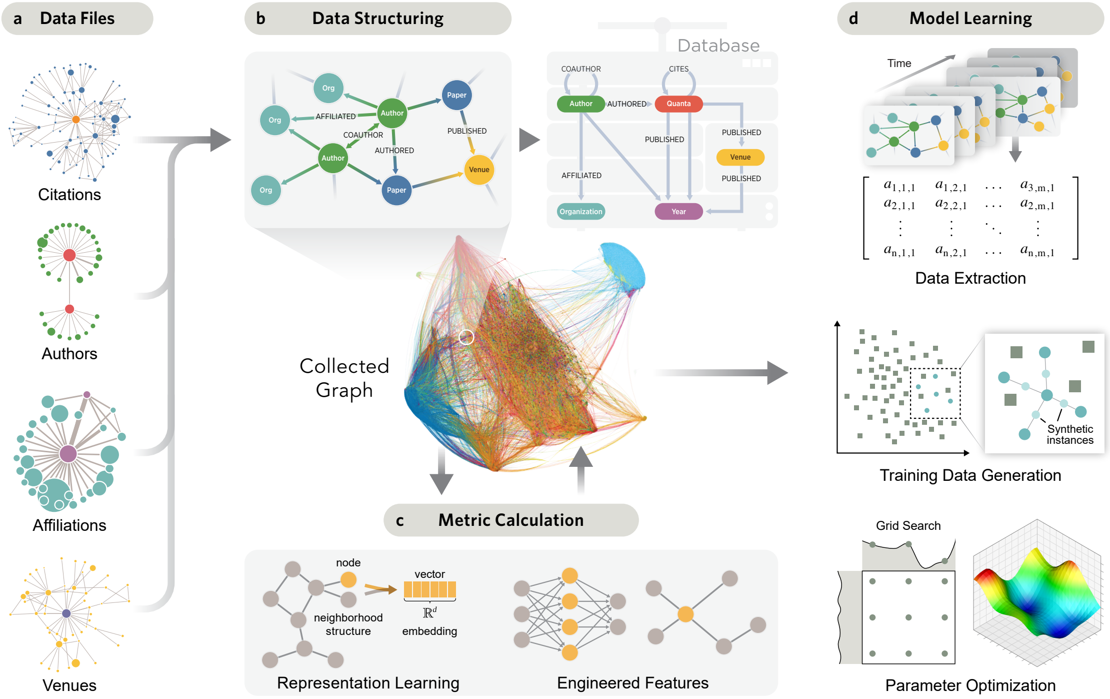


Figure 5-1: **Collecting, structuring, computing on, and learning an early-warning signal of scientific impact from dynamic knowledge graphs.** **a** Multiple data types are collected, merged, and disambiguated. **b** The resulting datasets are structured as a dynamic, heterogenous graph, wherein the temporal dynamics and time-based structural changes of the graph are also encoded. **c** A diverse set of metrics are calculated from, and stored in, the graph database in an iterative fashion. **d** A machine learning pipeline is constructed to learn differential patterns in structure between the calculated metrics over time by predicting the future impact of papers using balanced datasets extracted from the graph database. This pipeline is then optimized over a grid of possible parameters. The consequent DELPHI model can be used to produce an early-warning of impactful research.

scientific research, we included not only simple citation metrics and other metrics suggested as important by recent literature, but also learned in an unsupervised manner a low-dimensional representation of the local graph structure surrounding each node (Figure 5-1c)[116, 52]. We note that the inclusion of these metrics in our dataset does not mean they are assumed to positively correlate with scientific impact (and as such do not represent a reliance on reputation-based signals) because the DELPHI model learns to use only those metrics that truly contain information about future impact. Importantly, we used a time-rescaled measure of node centrality as our targeted impact metric due to its demonstrated state-of-the-art performance in ranking milestone technologies, as well as it's removal of age-bias, which allows for meaningful comparisons across years[75, 121]. All metrics were calculated for all years between 1980 to 2019.

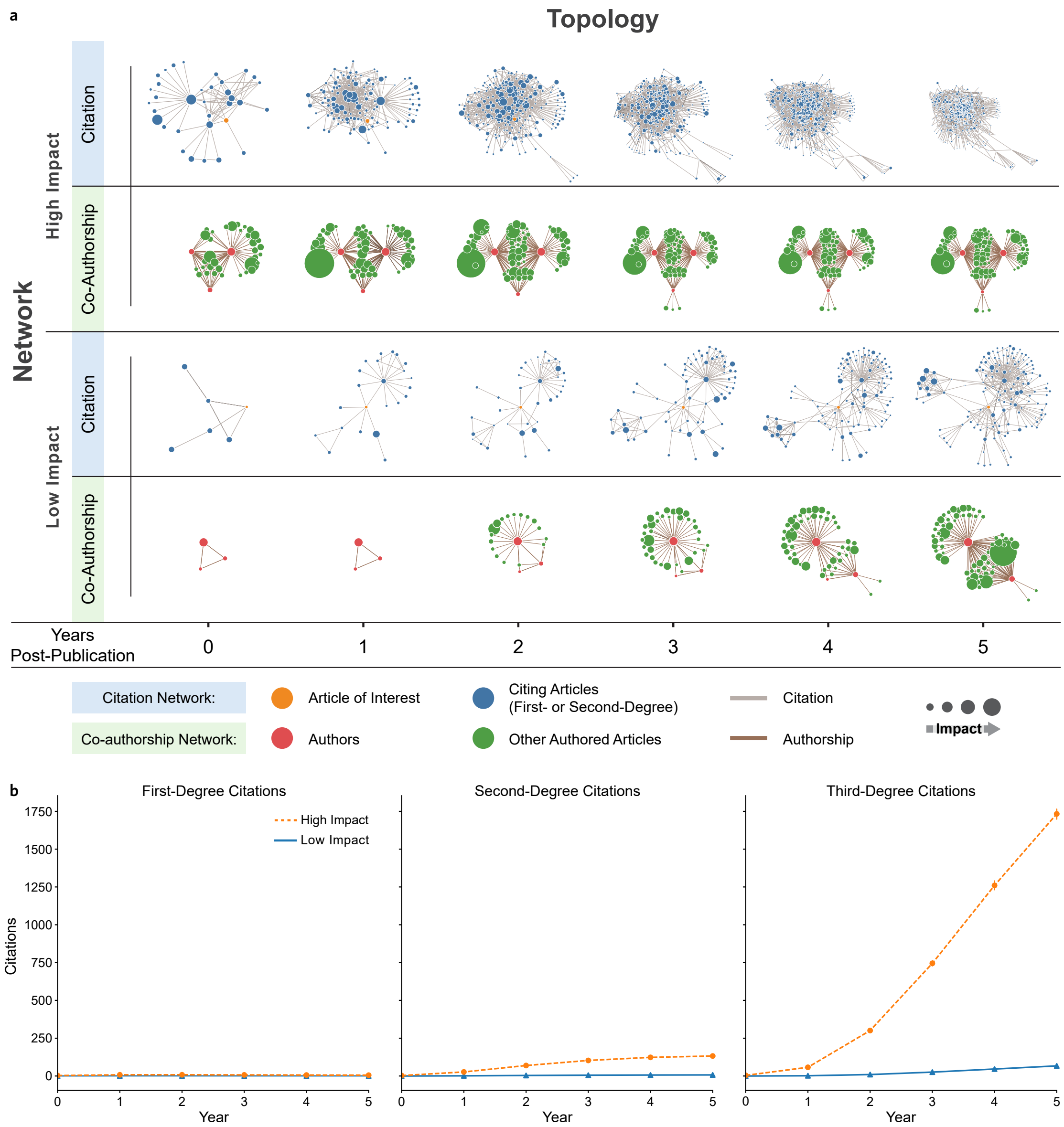


Figure 5-2: **Dynamics of knowledge graph structure contain information about future scientific impact. a** Visualization of the comparative evolution of coauthorship and citation network structures for canonical low- and high-impact publications. These papers were sampled from a set with an equal number of authors and similar citation counts. The citation network topology shown contains both first-order (directly citing) and second-order (papers that cite direct citations). Despite the similar citation count, the extended citation community of the high-impact article is significantly larger. Similarly, the coauthorship network of the high-impact publication, despite having the same number of authors, exhibits differential dynamics over time. **b** The mean number of first-, second- and third-degree citations were calculated for the high- and low-impact publication groupings for multiple years post-publication. The high-order subset exhibits significant and increasing differential enrichment for higher-order citations, as expected, indicating increased citations by highly-cited work.

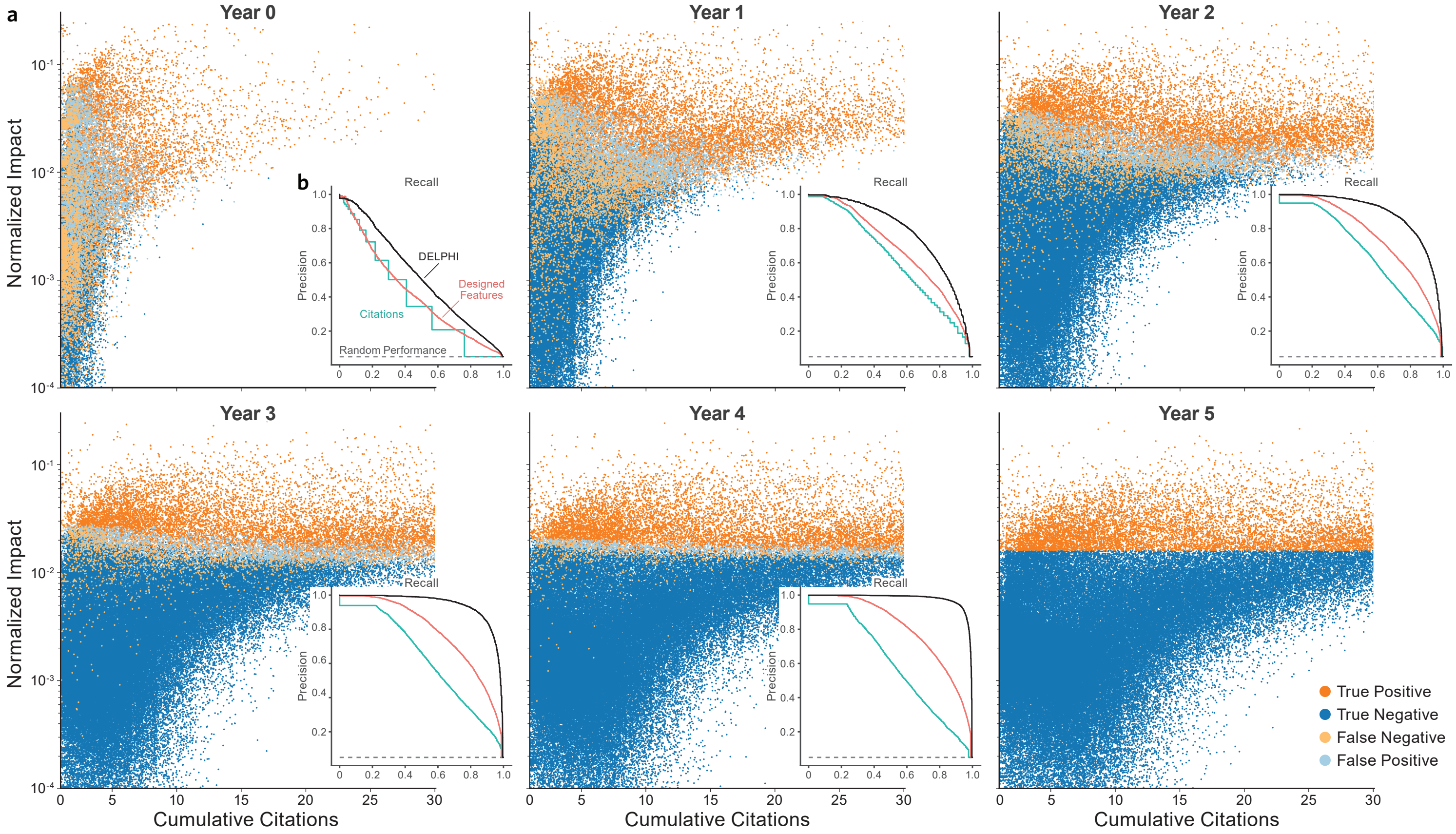


Figure 5-3: **DELPHI leverages temporal dynamics to identify high-impact research early and with state-of-the-art performance characteristics.** **a** By ingesting the dynamics of a broad range of metrics over time, DELPHI is able to provide an early-warning signal of highly-impactful research as early as the year of publication–even when that research does not have high citation count. Here, all articles within our biotechnology-focused sub-graph are shown as points, plotted by their cumulative citation count (x-axis) and scaled impact score (y-axis) for multiple years post-publication (columns and rows). Each point corresponds to an individual publication in the given year, and is colored according to the prediction made by DELPHI. Incorrect predictions are highlighted by being plotted on top of correct predictions for clarity. No predictions are made for Year 5, as that is our target year. **b** The precision and recall of DELPHI was compared to similar models that were only allowed to access citation data ("Citation-only Model") or metrics used in recent literature ("Designed Features"). DELPHI shows significantly improved performance characteristics as early as the year of publication, with the DELPHI early-warning signal increasing in fidelity as more data is consumed over time.

5.3.2 Learning On Biotechnology Knowledge Graph Dynamics Provides an Early-Warning Signal for High Impact Research

Scientific innovation is a highly complex process that evolves dynamically, and as such we model it in a high-dimensional, time-structured manner. In contrast to previous work, DELPHI operates across multiple data types, integrates both engineered metrics and automatically learned graph representations, and leverages valuable temporal information, such as changes in network structure and related features over time, to produce an early warning score for future scientific impact. Due to this graph-native structure, DELPHI is also easily extensible to new data types.

We train our model using the metrics and learned representations for all publications in our biotechnology-focused database for a time-window that spans from the year of publication to five-years post-publication (Figure 5-1d). Because we are interested in the subset of papers that have high influence on the biotechnology graph, we label as highly-impactful those papers that are in the top 5% of time-rescaled node centrality five-years post-publication (as these papers account for over 35% of total aggregate impact in our dataset). We find that high-impact articles have distinct time-based patterns of adoption that manifest in our heterogeneous knowledge graph model, and that DELPHI is able to correctly identify research of future high-impact as early as the year of publication (Figure 5-2a). DELPHI further significantly outperforms similar models trained with only citation data or metrics used in recent literature (Figure 5-2b). Interestingly, many of the correctly-identified high-impact publications have low early-year citation count, and as such represent important research "gems" that would not be found using standard citation-based metrics.

The strength of DELPHI's early-warning signal increases with the amount of time-series data utilized; for example, DELPHI correctly predicts which research will be highly-impactful with 77% balanced accuracy with less than one-year of data, but with 87% balanced accuracy using less than two-years of data (Figure 5-4a). While we initially chose a 5% threshold for our definition of high-impact, we find that DELPHI is robust to this definition, and exhibits comparable performance across a full range of tested thresholds (Figure 5-4a).

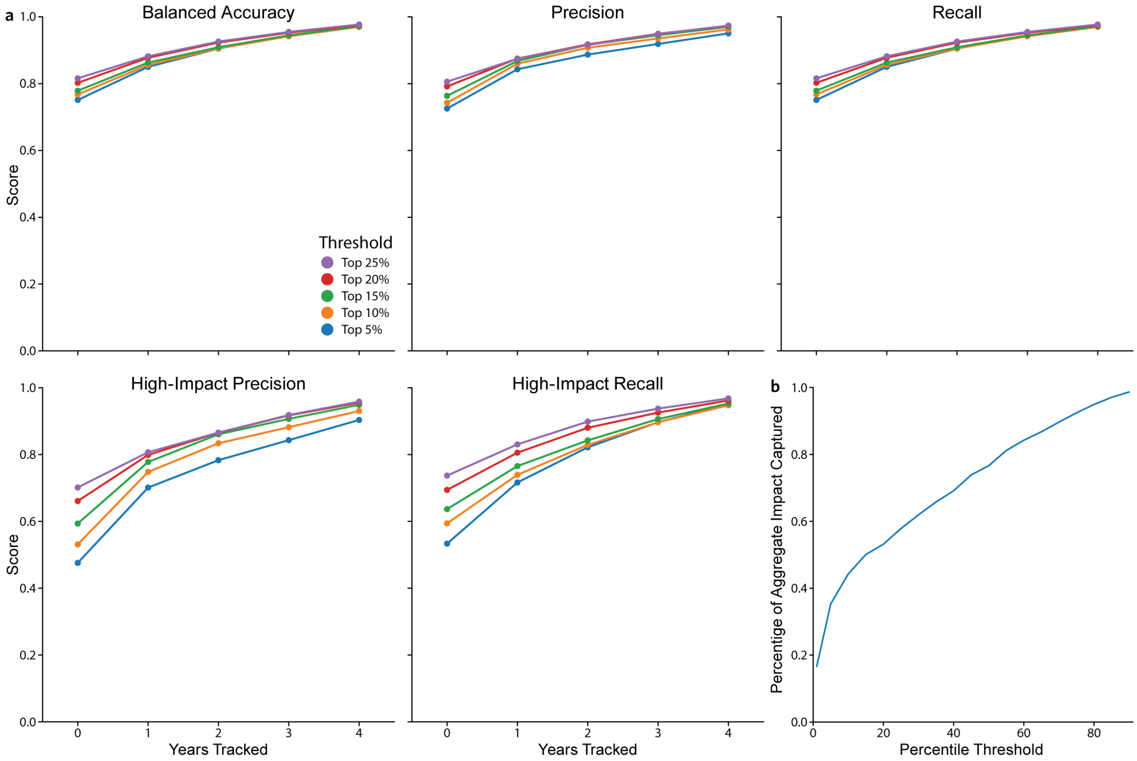


Figure 5-4: **The DELPHI approach exhibits strong performance characteristics across a range of definitions of high-impact and and model evaluation criteria. a** The DELPHI framework is based on a user-defined definition of high-impact. However, the performance of the framework is robust to the specific parameters of that definition. DELPHI models were constructed with a range of threshold definitions between 5% and 25%, and evaluated across a range of criteria to demonstrate this robustness. **b** Those papers in the top 5% of our impact metric, time-rescaled node centrality, contain over 35% of total aggregate impact. As such, the high-impact threshold of 5% was chosen for this study.

5.3.3 DELPHI Early-Warning Signal Correctly Identifies Seminal Biotechnology Breakthroughs And Prospectively Flags Interesting Research

To validate that DELPHI correctly classifies known breakthroughs, we collected a list of seminal biotechnology innovations–including both technical breakthroughs and therapeutic modalities–along with their corresponding publications[83]. Retraining our framework on a dataset with the target technology removed, we find that DELPHI correctly identifies all technologies adequately represented in our database–usually doing so within the same year as publication (Figure 5-5).

Next, we used DELPHI to predict the recently-published papers of likely future importance. Using a re-trained model blinded to papers from 2018 onwards, we calculated early-warning scores for those articles in our database published in 2019. The results highlight the potential of DELPHI to autonomously highlight interesting research in a variety of fields and areas of application (Table **??**).

5.4 Discussion

In this paper we have described DELPHI, a system that, by using machine learning to compute on large, heterogenous, time-structured network data, can produce a quantitative early-warning score for high-impact scientific research in biotechnology able to significantly outperform previous citation and handcrafted systems for impact prediction. Although this manuscript represents an initial demonstration, we believe that such a system offers some intriguing possibilities.

Across a diverse set of network types, from the physical network fabric of the internet to social networks (e.g. Facebook), the value of a network scales in proportion to the number of connections (edges) between individual nodes of the network—therefore those networks for which a linear expenditure on nodes generates a super-linear number of edges have the potential to be both very quickly scaling and extremely valuable[78, 125]. We hypothesize that the same is true of the scientific enterprise. Specifically, the value created by the ensemble of scientific research may scale in proportion to the number of connections (edges) between projects—and, therefore, research resources could be allocated so as to maximize the number of connections (edges) within the scientific research graph (as opposed to optimizing the number of citations). Although such beneficial allocations have hitherto been difficult to implement, we can now begin to think about using DELPHI to aid in designing funding strategies of this type–for example, by identifying identifying holes in the scientific graph that could be filled with new research program opportunities designed to optimize for connections of predicted high impact.

Another intriguing possibility is the quantitative diversification of research programs. Funding many research programs with similar scientific approaches, or composed of teams that lack diversity, constitutes an approach which is often suboptimal in the discovery phase of scientific endeavor. Borrowing from finance theory, we can think about reducing this risk by using DELPHI to construct a diversified portfolio that maximizes risk-adjusted scientific impact–that is, a group of research programs that, collectively, have optimized risk-reward characteristics. In this new model of grant allocation, the risk of a funding portfolio can be quantified empirically, such as via the historical correlation between the corresponding researcher's publication records (Figure 4a). In this context, portfolio optimization strategies will automatically identify diversified baskets of researchers (Figure 4b).

As motivation for the exploration of such computationally-assisted funding strategies, consider the substantial capital deployed in support of scientific research annually, with the NIH alone, for example, allocating some $40B USD each year to biomedical research. Despite a decline in the percentage of grant applications funded, as well as evidence that grant study sessions are unable to meaningfully outperform random selection in the identification of applications with high productivity, the primary scientific funding mechanisms remain largely unchanged [40, 85]. This is further complicated by the complexity of the modern scientific ecosystem, which contains more high-dimensional interactions than we can reasonably expect to be processed fully and in a bias-free manner without computational assistance. As an example, if we choose to fund the top 5% of the papers in our dataset on the basis of the citation count available two-years post-publication, we would have funded about 59% of high-impact research, but with over a 41% false positive rate—whereas, using DELPHI, we could have identified and funded over 81% of high-impact research with a false positive rate of only 20%. We believe such methods, therefore, hold great promise for improved aggregate productivity—including not only identifying the most promising projects, but also more completely exploring the tail-ends of scientific research, where revolutionary innovations can disproportionately occur.

While the results described herein are both exciting and tantalizing, we emphasize that DELPHI represents only a first-step towards the real-world application of machine-augmented analysis of the scientific literature. As such, DELPHI should be understood as part of a broader scientific analysis toolkit, to be utilized in combination with human experience and intuition—augmenting, not replacing, human-level understanding.

As with all machine learning-based systems, care must be taken to ensure these methods reduce (and do not unintentionally aggravate) latent systemic biases, and also do not provide opportunities for malicious actors to manipulate the system for their own gain. By considering a broad range of features and only utilizing those that hold real signal about future impact, we believe DELPHI holds the potential to reduce bias by obviating reliance on simpler (and often reputation-related) metrics. For the same reason, it is possible that DELPHI scores will be more difficult for authors or journals to manipulate than, for example, simple citation counts (upon which h-Index and Journal Impact Factor are based). However, additional studies, as well as careful human examination of anly calculations, are critical to more fully understand these possibilities.

Future research should also focus on the definition of impact and innovation utilized in the DELPHI objective function. As described above, we chose time-rescaled node centrality due to its demonstrated best-in-class ability to identify milestone research, as well as its removal of age-bias, which facilitates comparisons across the time domain [121, 75]. However, as the decision of impact metric is of critical importance, and also because DELPHI could naturally be extended to provide early-warning signals for other user-supplied metrics of interest (such as measures of scientific disruption), the properties and performance of various impact metrics should be explored in detail[45].

While this study is focused on scientific impact that materializes within a five-year post-publication window, there are longer-term scientific trends that are not captured by this temporal window—for example, the discovery of monoclonal antibodies occurred in the mid-1970s, but do not accelerate as a field until the early 1990s. DELPHI's combination of time-series analysis and network-level metrics may contain the expressivity necessary to capture and help understand these trends. Similarly, while we aggregated metrics in this study by paper, author, and journal, similar methods could be applied to the study of scientific ideas or concepts. As one possibility, the application of existing Natural Language Processing methods could help aggregate DELPHI scores by scientific concepts (e.g. "artificial organs"), or even different applications of the same concept (e.g. "artificial kidney" and 'artificial liver"). Such analyses may require significantly more data and computational resources, but could help us differentiate between fashionable and truly revolutionary ideas—or uncover promising but under-appreciated scientific "gems".

Future research could also explore technical ways to improve the DELPHI framework. As our interest in this study was the identification of the most promising 5% of biotechnology research, we adopted a classification-based approach. However, the DELPHI framework can also be used with regression-based methods, which may be able to improve performance by extracting additional signals or better identifying outliers. In addition, the direct application of graph-based machine learning methods could be used to explore community-based feature development or eliminate the need for data export from a graph-structured database.

Continual growth of the scientific corpus, as well as increasing importance of non-traditional literature such as academic pre-prints, is expected. DELPHI, due to it's inherently heterogeneous design, is capable of incorporating additional data sources in a straightforward manner, and the integration of commercialization data (such as patents and startups) could provide insight intthe drivers of translational success. Our discussed portfolio-theoretic approach could provide a mechanism via-which to incorporate such diverse signals—and requires additional theoretical development, analyses, and exploration to identify valuable quantifications of scientific risk. With such further developments, DELPHI-like approaches could be used to improve the modeling of financial returns for investors—potentially facilitating the deployment of more capital into scientific ventures or the creation of new classes of financial products.

Given the critical acknowledgement and awareness of the diverse caveats and potential stumbling blocks outlined above, and with the benefit of careful reflection and methodical deployment, we believe the careful development and deployment AI-assisted approaches such as DELPHI could unlock a wealth of existing but currently untapped resources. By computationally digesting, at scale, the vast amount of information contained in the scientific enterprise, we may be able to allocate our collectively limited resources in a more efficient, fair, and productive manner—and thus increase our return on the resources we collectively deploy into science and technology.

5.5 Methods

5.5.1 Collection Of Dataset

Our database is constructed using metadata retrieved from the Lens Labs API, which includes integrated and disambiguated data from PubMed, Crossref, Microsoft Academic, CORE, and PubMed Central. The retrieved data includes all publications available, as well as their associated meta-data, including inter-paper citations, from the 42 biotechnology-oriented journals outlined in Table 5.1. As of the retrieval date (April, 2020), this constituted 3,078,897 unique publications, where uniqueness was judged by the provided *LensID* field. As the rate of academic publishing prior to 1980 was substantially lower than in the following decades, resulting in a sparser network and different dynamics, and the data after 2020 is not fully populated, we narrowed the scope of our analysis to papers with viable metadata published between 1980 and 2019. This next filter resulted in 1,687,850 unique publications.

5.5.2 Construction Of Knowledge Graph

Using our filtered dataset, we construct a heterogeneous knowledge graph network. In this network, papers are represented as nodes, and citations are represented as directed edges between paper nodes. Similarly, journals and years are represented as nodes, with edges from each article representing the venue and year of publication, respectively. Authors are also represented as nodes, and authorship is expressed as a directed edge from an author to a paper. Additional entity-level metadata is stored as properties in the corresponding node. To store time-series of calculated metrics of interest for use in later training and evaluation of the machine learning model, relationships are added from each entity to the year of the metric calculation.

5.5.3 Disambiguation Of Knowledge Graph

Because our knowledge graph is constructed from data retrieved from the Lens Labs, it is already unified and heavily disambiguated, especially with respect to the publication entities and inter-publication citations. However, we apply a second-stage disambiguation algorithm: For each node, if that node is an author, then we hash that node using, separately and in each case if it is available, that author's Microsoft Academic ID and ORCID iD. Then, using this hashmap, identify and merge all author nodes with overlapping IDs. We also ensure that there are no nonsensical citation edges: for every edge, we ensure that the citing paper was published no earlier than the year of publication of the cited paper. We also reviewed all journal nodes by hand, to ensure that no duplicates existed.

5.5.4 Calculation Of Metrics

We use our constructed and disambiguated biotechnology knowledge graph to calculate time-series of publication and author metrics, which we use as features in our machine learning pipeline. While the DELPHI framework is easily extensible to additional metrics, the specific metrics utilized in this study were collected from commonly used metrics, as well as features used in the related work highlighted earlier, and are outlined in Table 5.2. Specifically, the metrics fall into four primary categories: (1) paper-level, metrics, (2) journal-level metrics, (3) author-level metrics, and (4) network-level metrics (all of which are described in more detail below).

For each metric, we implemented a separate algorithm that uses the structure of our knowledge graph to calculate the desired value in a computationally efficient manner. Then, we calculated the full set of metrics for each paper, author, and journal in our knowledge graph, for every year from the year of publication to 2019.

Paper-level: For each paper, for each year, we calculate the number of incoming citations from all other papers in the dataset that were published prior to the year under consideration. These calculations are then used to calculate a range of other citation-based properties for each paper, e.g. the total number of papers, total number of citations, the number of citations per year, etc.

Author-level: For each author, the previously calculated paper-level metrics were aggregated to derive additional author-level metrics of interest. These include the author's *h*-index (calculated, as with all metrics, using only the papers from the dataset in question), years since publication of the first paper, total number of papers, citation count, etc.

Journal-level: We then used both the paper- and author-level metrics to calculate aggregated metrics for each journal. This was completed in a similar process to the author-level metrics, and included metrics like the journal's paper count and maximum number of citations. Then, for each author, for each journal-level metric, the maximum, mean, and minimum values for all journals that author has published in were aggregated back to the author.

Network-level: Because of the complexity of the scientific community, the simple citation- and literature-based metrics described may not be sufficiently expressive to capture the diversity of structure and relationships in our knowledge graph. Thus, we also learn, in an unsupervised manner, continuous feature representations for each paper based on it's local citation network structure using the node2vec algorithm with settings for directed edges, and an 80-step random walk, for each year the paper exists in the dataset [52].

5.5.5 Machine Learning Pipeline

The metrics calculated above are used in the development of the machine learning pipeline, which uses these inputs, to produce a quantitative early-warning signal. First, the data is extracted from the network database, aggregated to a per-paper level. For each paper, we calculate all of the metrics described above, for every paper in our database with at least one citation, from the year of publication to the year that is five years after the year of publication. Next, we remove all those papers for which authorship information is missing, leaving 1,540,798 unique publications. Then, as discussed in the main text, we follow recent literature in the leverage of network-based metrics for scientific impact. Specifically, we define a paper as "high-impact" if, five-years after publication, it has a time-rescaled PageRank score in the top 5% of all scores for that year (although our approach is robust to this selection, as shown in Figure 5-4). Given a specific number of years to track (ranging just the year of publication, to all data up to five-years after the year of publication), we first separate the data into training (75%) and test (25%) data sets. Because low-impact works are significantly more frequent than high-impact works according to our definition, we apply synthetic minority over-sampling (SMOTE) to generate balanced training sets[?]. We then perform robust feature preprocessing to the calculated metrics, to allow our values, which are measured at different scales, to be more directly comparable. Next, given our prepared training data, we train a machine learning model, when we optimize by conducting cross-validation on the training data across a grid of possible model parameters. While we have explored a variety of machine learning models, we chose to use random forest classifiers in the current study because of computational and memory efficiency, the potentially decreased risk of over fitting, the potential reduction in prediction variance, and the ability to capture non-linear relationships natively. Because the network-level features add significant dimensionality to the model, we inspect whether their inclusion improves or degrades the model's performance, and choose the best-performing model. Finally, we evaluate the performance of our model by comparing our impact predictions with the previously calculated true impact label on the held-out test data.

5.5.6 Retrospective Analysis Of Biotechnology Breakthroughs

A list of seminal technical breakthroughs and therapeutic modalities, along with their corresponding papers, was collected from [83]. Then, we selected those technology/paper pairs that (a) were published in one of the 42 journals included in our database, (b) were published between 1980 and 2014, to allow for a full five-years of tracking, and (c) contained more than one citation for each of the first five-years post publication from another paper in our dataset, to ensure sufficient representative signal to make a realistic prediction. We constructed alternative DELPHI models that were blinded to the technology/papers under analysis. Then, for each technology, we found the mean DELPHI early-warning score for each year between the year of publication up to five-years post-publication. We also analogously calculated DELPHI early-warning signals for all articles in our test set published in the journal Nature for the same years, for comparison.

5.5.7 Prospective Study Predicting Recent Publications Of High Future Impact

To calculate the DELPHI early-warning signal for recently published papers, we trained a DELPHI models on papers from 1980-2017. Then, we deployed these models on features calculated in 2018-2019 for papers published in 2018. The resulting top-ranked 100 papers were sampled so as to identify non-repetitive trends, with the results shown in Table **??**.

Table 5.1: The 42 life sciences journals from which research was identified and included in the biotechnology-graph analysis

Angewandte Chemie	The Lancet
Blood	Nature Cell Biology
Cancer Cell	Nature Chemical Biology
Cancer Discovery	Nature Chemistry
Cancer Research	Nature Medicine
Cell	Nature Methods
Cell Host & Microbe	Nature
Cell Metabolism	Nature Biotechnology
Cell Stem Cell	The New England Journal of Medicine
Cell Chemical Biology	Neuron
The EMBO Journal	Nature Genetics
Genes & Development	Nature Immunology
Immunity	Nature Neuroscience
Journal of Neurology	Nature Structural & Molecular Biology
Journal of the American Chemical Society	PLOS Biology
JAMA	PLOS Genetics
Journal of Biological Chemistry	PLOS Pathogens
Journal of Cell Biology	Proceedings of the National Academy of Sciences
Journal of Clinical Investigation	Science Signaling
Journal of Experimental Medicine	Science Translational Medicine
Journal of Medicinal Chemistry	Science

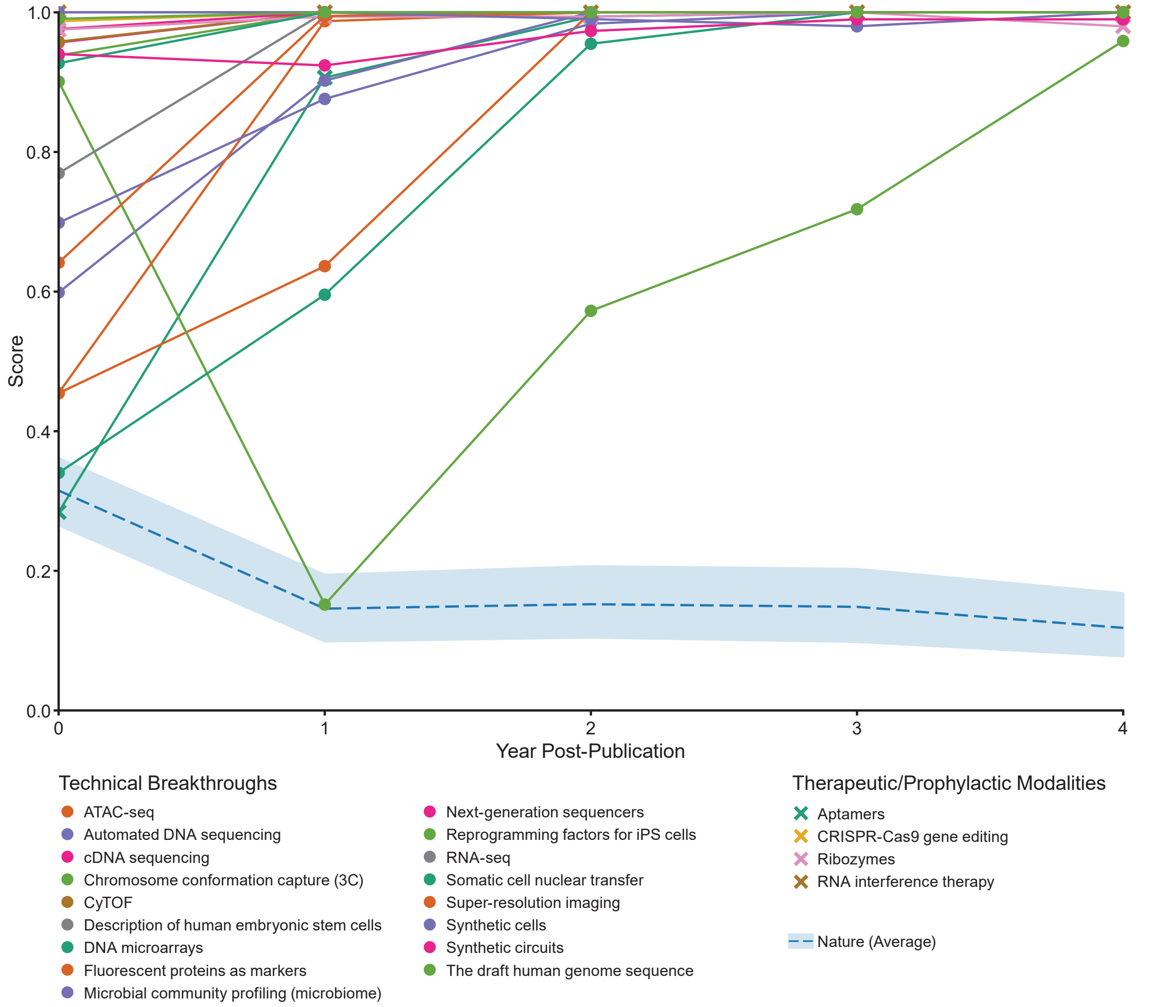


Figure 5-5: **DELPHI correctly identifies historical biotechnology breakthroughs in a blind back-testing.** DELPHI Or were trained on datasets that did not include the seminal papers corresponding to the technical breakthroughs or therapeutic modalities indicated. These "blinded" DELPHI models rapidly and accurately identified these pioneering papers as likely to be highly-impactful. For comparison, the average and 95% confidence interval for a sampling of articles from the journal Nature, a group that is already considered enriched for impactful research, are also shown for comparison. The early-warning signals for some of these technologies, most notably Chromosome conformation capture (3C), is reduced because a large number of their real-world citations come from articles that are not in our 42 journal dataset, and as such these citations are not present in our graph database and thus not available for consideration by DELPHI.

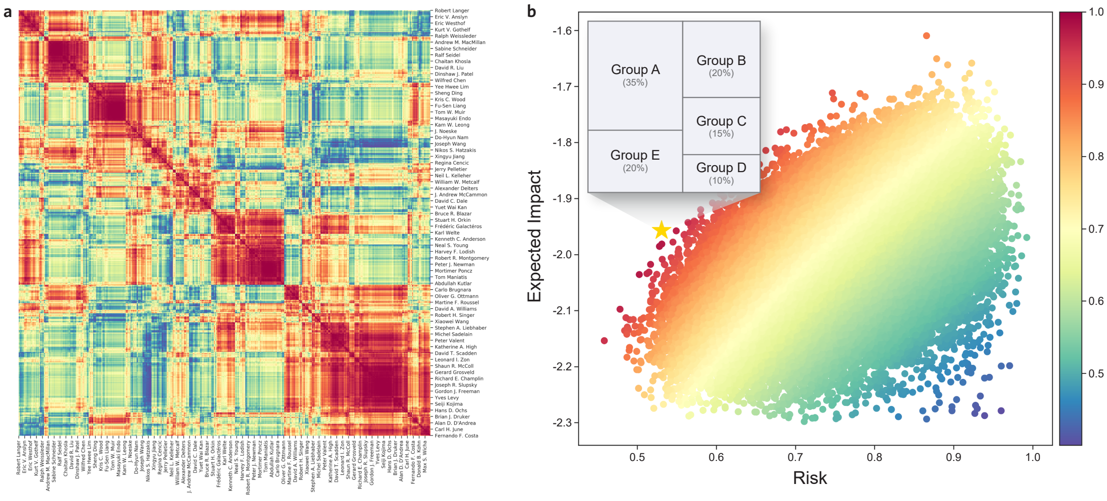


Figure 5-6: **In a world of expanding science and limited resources, quantitative approaches like DELPHI can be used to help guide research funding allocations to maximize scientific return on investment. a** Measures of risk can be calculated using quantitative metrics. Shown here is a clustered correlation matrix between researchers with demonstrated interest in genome editing. Clusters identify groups of researchers with potentially related research interests, as judged from the dynamics of their historical publication record. **b** An example Monte Carlo simulation exploring the risk-reward tradeoff for various potential five-person grant decisions. Here, portfolios with researchers whose previous publication histories are highly-correlated represent potentially higher-risk portfolios from a funding perspective, and researchers with established scientific contribution records have higher expected return. There is a trade-off between risk and return, with higher-risk portfolios also have higher expected returns.

Table 5.2: Features used in machine learning-based early warning system. These features are extracted from the graph databases in question, and their calculation constitute the first step of the DELPHI framework. Note that each feature is also paramaterized by the year in question—that is, it is calculated, in all cases, for all relevant years.

Category	Variable	Description
Paper	Citations Per Paper	Mean number of citations per paper for papers the author has published.
	Delta Citations Per Paper	Change in the mean number of citations per paper for the author over the preceding two years.
	Citations Per Year	Average number of citations per year for papers the author has published.
	Maximum Citations	Maximum number of citations a paper has received out of all the papers the author has currently published.
	Rank Citations Per Year	Rank of the author among all other authors in terms of mean citations per year.
	Total Citations	Number of citations author has received.
	Delta Total Citations	Change in the total number of citations for the author over the preceding two years.
	Total Papers	Total number of papers published by the author.
	Delta Total Papers	Change in the total number of papers over the proceeding two years.
	Citations	Citations collected in the current year.
Author	Adopters	Number of unique citing authors in the current year.
	Author Age	Number of years since the year of publication of the author's first paper
	h-Index	Author's h-Index
	Delta h-Index	Change in the author's h-index over the past two years.

	Recent Coauthors	Number of coauthors the author has had in the current and immediately preceding year.
Journal	Delta Mean Journal Citations Per Paper	Two-year change in the mean number of citations per paper of the journals the author has published in.
	Mean Journal Citations Per Paper	Mean number of citations per paper for the journals the author has published in.
	Delta Mean Journal *h*-Index	Two-year change in the mean *h*-index for the journals the author has published in.
	Mean Journal *h*-Index	Mean *h*-index for the journals the author has published in.
	Mean Journal Maximum Citations	Mean of the maximum number of citations any paper published in a journal has received for each journal the author has published in.
	Mean Journal Rank Citations Per Paper	Rank of journal in which the author has published, as determined by the mean number of citations per paper.
	Mean Delta Journal Total Papers	Change in the mean of the total number of papers published in journals the author has published in.
	Total Journals	Total number of journals published in by the author.
	Mean Journal Total Papers	Mean of the total number of papers published in journals the author has published in.
Network	Learned Network Embedding	Unsupervised embedding of local network structure calculated via application of the *node2vec* algorithm on the citation graph.
	Time-Scaled Impact	Time-balanced network centrality calculated using the full citation network.
	Unweighted PageRank	PageRank score of author, calculated on the unweighted coauthorship network.
	Weighted PageRank	PageRank score of author, calculated on the weighted coauthorship network.

Table 5.3: A sampling of recent publications predicted to be of future high-impact by DELPHI. A DELPHI model was trained on a restricted version of the biotechnology-focused dataset which only contained data from 2017 and earlier. Metrics were then extracted for articles in our dataset published in 2018 for 2018 and 2019, and the new DELPHI model was used to rank the publications by predicted probability of high-impact. Despite the fact that this analysis is limited to features extracted from the sub-graph of 42 journals, the identified publications including research topics spanning beyond biotechnology. Citation counts as of 2019 from within our 42 journal dataset, as well as the full academic graph, are shown.

Title	Journal	Dataset Citations	Full-graph Citations
A DNA nanorobot functions as a cancer therapeutic in response to a molecular trigger in vivo	Nature Biotechnology	24	212
A high-energy-density lithium-oxygen battery based on a reversible four-electron conversion to lithium oxide	Science	13	60
A living biobank of breast cancer organoids captures disease heterogeneity	Cell	24	213
An APOBEC3A-Cas9 base editor with minimized bystander and off-target activities	Nature Biotechnology	11	88
An in vivo model of functional and vascularized human brain organoids	Nature Biotechnology	18	155
Analysis of shared heritability in common disorders of the brain	Science	20	329
Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors	Nature Biotechnology	42	239

Table 5.3 continued on the following page

Table 5.3 continued from the previous page

Title	Journal	Dataset Citations	Full-graph Citations
Bystander CD8+ T cells are abundant and phenotypically distinct in human tumour infiltrates.	Nature	26	140
Coactivator condensation at super-enhancers links phase separation and gene control	Science	53	267
CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity	Science	24	208
De novo DNA synthesis using polymerase-nucleotide conjugates.	Nature Biotechnology	4	49
Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study	The Lancet	9	67
Design of Single-Atom Co-N$_5$ Catalytic Site: A Robust Electrocatalyst for CO2 Reduction with Nearly 100% CO Selectivity and Remarkable Stability	JACS	19	133
Development of a synthetic live bacterial therapeutic for the human metabolic disease phenylketonuria	Nature Biotechnology	10	53
Enhancing the potential of enantioselective organocatalysis with light	Nature	55	97
Evolved Cas9 variants with broad PAM compatibility and high DNA specificity	Nature	36	367

Table 5.3 continued on the following page

Table 5.3 continued from the previous page

Title	Journal	Dataset Citations	Full-graph Citations
Genetic identification of brain cell types underlying schizophrenia	Nature genetics	16	108
Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma	NEJM	24	251
Genomic correlates of response to immune checkpoint therapies in clear cell renal cell carcinoma	Science	30	199
Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980-2017: a systematic analysis for the Global Burden of Disease Study 2017	The Lancet	28	423
H3B-8800, an orally available small-molecule splicing modulator, induces lethality in spliceosome-mutant cancers	Nature Medicine	18	93
Health Care Spending in the United States and Other High-Income Countries	JAMA	29	347
Human ADAR1 Prevents Endogenous RNA from Triggering Translational Shutdown	Cell	16	79
Hypothalamic Circuits for Predation and Evasion	Neuron	15	48
Imaging dynamic and selective low-complexity domain interactions that control gene transcription	Science	27	145

Table 5.3 continued on the following page

Table 5.3 continued from the previous page

Title	Journal	Dataset Citations	Full-graph Citations
Integrating single-cell transcriptomic data across different conditions, technologies, and species	Nature Biotechnology	205	1040
Itaconate is an anti-inflammatory metabolite that activates Nrf2 via alkylation of KEAP1	Nature	34	179
Mapping the Mouse Cell Atlas by Microwell-Seq.	Cell	37	279
Miniaturized neural system for chronic, local intracerebral drug delivery	Science Translational Medicine	8	28
Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain	Cell	29	168
mTORC1 Controls Phase Separation and the Biophysical Properties of the Cytoplasm by Tuning Crowding.	Cell	11	49
NK Cells Stimulate Recruitment of cDC1 into the Tumor Microenvironment Promoting Cancer Immune Control	Cell	21	177
Planning chemical syntheses with deep neural networks and symbolic AI	Nature	22	274
Regulation of age-associated B cells by IRF5 in systemic autoimmunity	Nature Immunology	5	24

Table 5.3 continued on the following page

Table 5.3 continued from the previous page

Title	Journal	Dataset Citations	Full-graph Citations
RNA velocity of single cells	Nature	37	219
Semiconducting Polymer Nanoenzymes with Photothermic Activity for Enhanced Cancer Therapy	Angewandte Chemie	16	82
Shared and distinct transcriptomic cell types across neocortical areas.	Nature	26	180
Simultaneous lineage tracing and cell-type identification using CRISPR–Cas9-induced genetic scars	Nature Biotechnology	21	103
Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain	Nature Biotechnology	22	135
Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment	Cell	42	204
Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease	Science	18	175
STK11/LKB1 Mutations and PD-1 Inhibitor Resistance in KRAS-Mutant Lung Adenocarcinoma	Cancer Discovery	14	133
The human transcription factors	Cell	24	289

Table 5.3 continued on the following page

Table 5.3 continued from the previous page

Title	Journal	Dataset	Full-graph
		Citations	Citations
The long tail of oncogenic drivers in prostate cancer	Nature genetics	24	142
Thermal proximity coaggregation for system-wide profiling of protein complex dynamics in cells	Science	7	29
Ultrafast neuronal imaging of dopamine dynamics with designed genetically encoded sensors	Science	21	127
Unexpected reversal of C3 versus C4 grass response to elevated CO2 during a 20-year field experiment	Science	8	73

It is not enough to know; we must also apply.
It is not enough to will; we must also do.
The Maxims and Reflections of Goeth

6 Machine-Augmented Scientific Exploration and Collaboration

6.1 Introduction

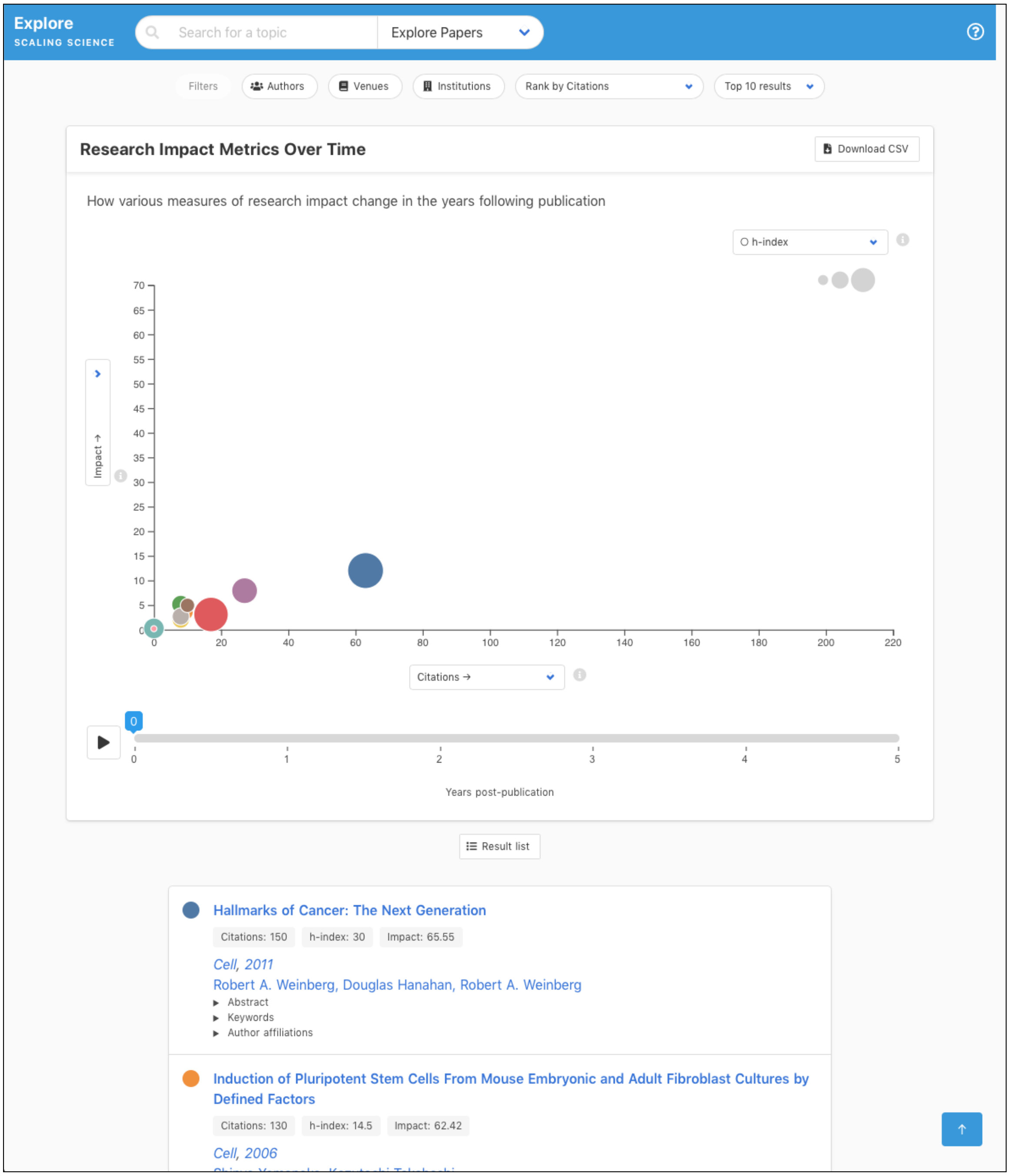
It has been the work of the proceeding chapters, especially Chapter 5, to demonstrate that extracting valuable signals via broad computational analysis of the scientific ecosystem is not only possible, but imminently so. Further, I have demonstrated that these signals can be utilized to the presumptive benefit of society–for example, via the less-biased identification of promising research talent or projects, or the construction of diversified funding portfolios. However, there exists another "activation barrier" in the long series of hurdles that span from research identification to commercialization–the development of used-focused tools to facilitate the real-world utilization of these learned signals.

In this Chapter, I take the first-steps towards applications that, effectively, ingest the signals produced by our proposed learning-based systems, and collect, condense, and present them in a user-friendly manner. While an innovation of a different sort than those presented earlier in this thesis, the development of such tools is critical on multiple levels: First, and most obviously, the influence of any innovation is limited by the capacity of the end-user to leverage it. Second, the iterative-development of these platforms, with the goal of augmenting their real-world utility, is only possible via empirical studies.

As such, herein I describe two tools: *Scaling Science Explore* (Section 6.2), which is designed to enable end-user analysis and navigation of the scientific landscape using the metrics and analyses outlined in Chapter 5, and *TitleGen* (Section 6.3), which is a first-step towards the utilization of scientific metadata in the construction of promising scientific collaborations. In both cases, the development and especially the features of the corresponding tool are explained and demonstrated, with avenues for future development highlighted.

6.2 Scaling Science Explore

Chapter 5 describes the calculation, both via network theory and machine learning, of a variety of metrics and signals of interest (See Table 5.2). One of the primary underlying motivations for the development of this platform was that the consequent signals, as well as the metrics calculated in the process, could be leveraged by the scientific community (e.g. researchers, program managers, government entities, investors) to facilitate the optimized allocation of a limited amount of resources. However, as is obvious, not all members of these communities are data scientists–and of those who are, their ability to explore large and nuanced datasets is naturally limited by the correspondingly large computational resources required. As such, it is necessary to explore the development of tools to make the utilization of our methods and metrics–as well as the critical examination of them–easier. *Scaling Science Explore* was developed with this goal in mind.



Research Impact Metrics Over Time

How various measures of research impact change in the years following publication

Hallmarks of Cancer: The Next Generation

Citations: 150 h-index: 30 Impact: 65.55

Cell, 2011

Robert A. Weinberg, Douglas Hanahan, Robert A. Weinberg

- Abstract
- Keywords
- Author affiliations

Induction of Pluripotent Stem Cells From Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors

Citations: 130 h-index: 14.5 Impact: 62.42

Cell, 2006

Shinya Yamanaka, Kazutoshi Takahashi

Figure 6-1: Screenshot of the *Scaling Science Explore* application. The primary features of interest: (A) a full-dataset search functionality, (B) filtering mechanisms, (C) a ranking mechanism, (D) a link for dataset download, (E) a four-dimensional visualization (three metrics and time) (D) a results list.

6.2.1 Collection Of Data And Calculation Of Metrics

As *Scaling Science Explore* was designed to facilitate the analysis of the metrics utilized in the study described in Chapter 5, the underlying data source is highly similar. Specifically, *Scaling Science Explore* takes as input a large CSV file structured in paper-level record format, with authors, institutional affiliations, and metrics of interest included as distinct columns. This CSV is then processed into a JSON format using an associated configuration file, which specifies the metrics to use, as well as how to describe them in the front-end of the application. The consequent datafile is then used for the on-the-fly display of the subset of research described by the user, and the corresponding metrics indicated by the user are aggregated and displayed across the time domain (Figure 6-1 and Figure A-5).

6.2.2 Application Features

The primary goal of *Scaling Science Explore* is to allow the (1) rapid identification of subsets of research of interest, (2) rapidly calculate features of interest to the user for that research subset, (3) display the relationships between these metrics, within the subgroup selected, over time, and (4) allow the user to use the results to quickly adjust the subgroup of interest.

Specifically, the features developed (which are enumerated in Figure 6-1 and described in more detail in Figures A-5, A-4, and 6-2) are as follows:

Full-dataset search: Allows users to rapidly identify papers, authors, venues (journals) or institutions of interest via keyword search.

Filtering mechanisms: Via which users can identify targeted subgroups of research of interest by selecting specific authors, venues, or institutions (See Figure A-4 for detail).

Ranking mechanism: Provides the ability to rank by all of the metrics calculated (including standard metrics like citation count, h-Index, as well as alternative metrics like the aggregated PageRank of the authors based on their coauthorship network (Figure A-4).

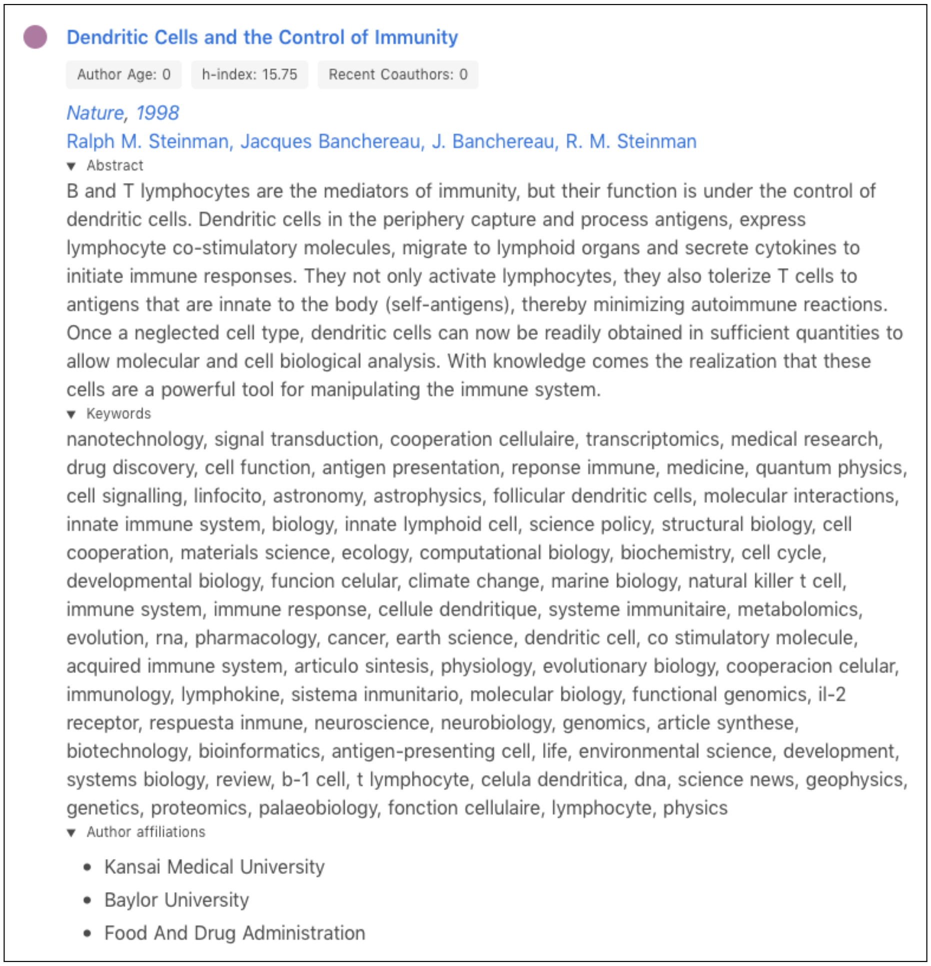
Dataset download: Provides the ability to download the selected dataset to raw CSV files for further analysis.

Four-dimensional visualization: A time-series scatter plot, wherein up to three distinct metrics can be selected and their relationship for the work of interest can be viewed over time (Figure A-5. This visualization can be fully customized by the user, with the metric displayed on the x-axis, y-axis, or as the "size" of the data point user-selected from a list of options. The relationships between these can then be "played" on a per-year basis from the year of publication to five-years after the year of publication.

Linked results list: Shows details about the publications selected via the search and filtering mechanisms described above. These results have hyperlinks embedded, which allows the user to quickly re-orient/update the visualization around other areas of interest (Figure 6-2).

6.2.3 Discussion

I anticipate that *Scaling Science Explore* can provide a foundational visualization platform, upon which more sophisticated visualizations can be developed. Further, *Scaling Science Explore* was designed to be naturally extensible to additional data. Thus, researchers can freely download the core functionality, update the data and metrics utilized with their own new metrics or analyses of interest, and host the visualization (either locally or globally) for visualization. To that end, I anticipate continued development of *Scaling Science Explore* core functionality, including both flexibility (to enable new visualizations) and speed (to enable the deployment on larger data sources).



Dendritic Cells and the Control of Immunity

Author Age: 0 | h-index: 15.75 | Recent Coauthors: 0

Nature, *1998*

Ralph M. Steinman, Jacques Banchereau, J. Banchereau, R. M. Steinman

▼ Abstract

B and T lymphocytes are the mediators of immunity, but their function is under the control of dendritic cells. Dendritic cells in the periphery capture and process antigens, express lymphocyte co-stimulatory molecules, migrate to lymphoid organs and secrete cytokines to initiate immune responses. They not only activate lymphocytes, they also tolerize T cells to antigens that are innate to the body (self-antigens), thereby minimizing autoimmune reactions. Once a neglected cell type, dendritic cells can now be readily obtained in sufficient quantities to allow molecular and cell biological analysis. With knowledge comes the realization that these cells are a powerful tool for manipulating the immune system.

▼ Keywords

nanotechnology, signal transduction, cooperation cellulaire, transcriptomics, medical research, drug discovery, cell function, antigen presentation, reponse immune, medicine, quantum physics, cell signalling, linfocito, astronomy, astrophysics, follicular dendritic cells, molecular interactions, innate immune system, biology, innate lymphoid cell, science policy, structural biology, cell cooperation, materials science, ecology, computational biology, biochemistry, cell cycle, developmental biology, funcion celular, climate change, marine biology, natural killer t cell, immune system, immune response, cellule dendritique, systeme immunitaire, metabolomics, evolution, rna, pharmacology, cancer, earth science, dendritic cell, co stimulatory molecule, acquired immune system, articulo sintesis, physiology, evolutionary biology, cooperacion celular, immunology, lymphokine, sistema inmunitario, molecular biology, functional genomics, il-2 receptor, respuesta inmune, neuroscience, neurobiology, genomics, article synthese, biotechnology, bioinformatics, antigen-presenting cell, life, environmental science, development, systems biology, review, b-1 cell, t lymphocyte, celula dendritica, dna, science news, geophysics, genetics, proteomics, palaeobiology, fonction cellulaire, lymphocyte, physics

▼ Author affiliations

- Kansai Medical University
- Baylor University
- Food And Drug Administration

Figure 6-2: Screenshot of the results list (Figure 6-1D). The top-ranked publications are shown, in accordance with the user-supplied ranking methodology. For each publication, when available, the venue (journal) and year of publication is shown. Also shown are three key metrics, the abstract of the paper, affiliated keywords, and the affiliations of the authors (not necessarily at the time of publication). Importantly, clicking on the title links to the paper of interest, clicking on the venue re-filters to publications from that journal, and clicking on any of the authors re-orients around that authors publications.

6.3 TitleGen

In Chapter 5 I explored data-driven solutions to start answering questions like "What should I work on?" and "What should I fund?". In this section, I explore data-driven approaches to the natural follow-on question: "With whom should I collaborate?".

Currently, collaboration decisions are made largely on the basis of pre-existing social connections or research publicity. For example, a researcher may launch a collaboration with someone in the same department, with someone identified at a conference, or, perhaps more rarely, via “cold calling” on the basis of recently published research. As such, there are network-based biases that causally affect the structure of collaborations–which are important because the structure of these collaborations and corresponding networks is known to be important in downstream scientific productivity [17, 99, 68, 28].

One can imagine designing tools that aid in the collaboration identification process. By digesting historical publication and even collaboration information, along with related measures of success, we could imagine building machine learning platforms that aid in the generation of collaborations that are most-likely to generate imapactful research. Alternatively, these same models could be used to optimize on different objective functions; for example, one model could generate collaborations to maximize general impact, while other models could maximize the probability of generating a commercializable technology of future market high value. Interestingly, we already have broad evidence that data-driven platforms can modulate human behavior from online social media networks and matchmaking applications. *TitleGen* was developed as a first-step towards exploring these intriguing possibilities.

6.3.1 Data Collections And Model Construction

The area of focus was chosen to be biotechnology and related sub-disciplines. Thus, publication data was collected from Microsoft Academic Graph [79] with related keyword tags. This data was structured to form $\mathcal{D}$, the data set comprised of the individual publication pairs $(a, \{p_i\})$, where a is an individual author, and $\{p_i\}$ is the set publications by that author. I denote $\mathcal{A} = \{a \in \mathcal{D}\}$ is the set of unique authors in $\mathcal{D}$

For each author, I then generate a corpus of titles, a_c, which I use to train a character-based Markov process in the standard way. Specifically, I model each character in a_c as a random variable, and learn the probabilities of transitioning between distinct characters states empirically from a_c. It is then straightforward to generate synthetic titles for each author by conducting a random walk across the resulting Markov chain. I denote the set of such trained author-level Markov models as

$\{m_a \mid a \in \mathcal{A}\}$

For any set of authors of interest queried, $\mathcal{A}^c = \{a_0, .., a_n\}$ with associated weightings $w_0, ..., w_n$ where $\sum_i w_i = 1$, it is then possible to construct a new Markov model $m_{\mathcal{A}^c}$ by joining the distinct elements of each corresponding model $m_a \mid a \in \mathcal{A}^c$ with new transition probability edges according to the corresponding weightings w_a. A random walk, then, can be undertaken to traverse this joined model $m_{\mathcal{A}^c}$, resulting in synthetically generated titles. These consequent sampled walks/titles represent one manner of drawing from an course estimation of the joined empirical character-level Markov process.

6.3.2 Application Features

This functionality was deployed for real-world experimentation and utilization via the *TitleGen* application (Figure 6-3. Specifically, the *TitleGen* application includes the following functionality:

Author search: Allows the rapid identification of authors using a keyword search across *name*, *institutional*, and *keyword* fields. These authors can then be easily added to $\mathcal{A}^c$ for $\mid \mathcal{A}^c \mid \leq 6$ (See Figure 6-3 for detail).

Author weighting adjustment: The weighting w_i for each author i selected can be adjusted graphically. The fact that these weightings are normalized such that $\sum_i w_i = 1$ is made apparent via the visualization design.

Prompt Generation: Given $\mathcal{A}^c$ and weightings $w_i, ..., w_n$, model $m_{\mathcal{A}^c}$ is created dynamically, and new prompts (sampled from $m_{\mathcal{A}^c}$) are generated on-demand.

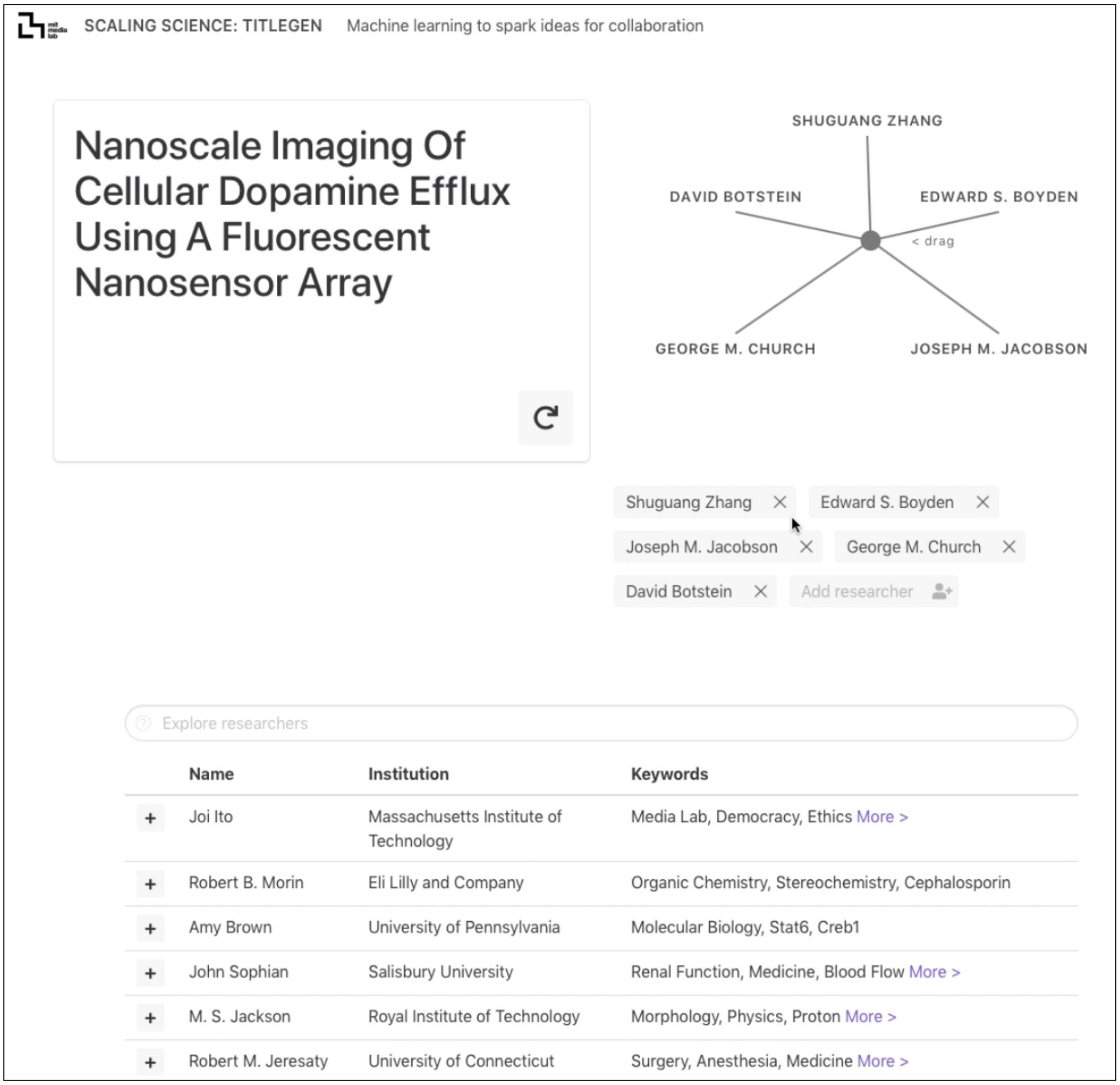
Author Removal: Authors can also be easily removed from $\mathcal{A}^c$ graphically via the author selection list underneath the weighting diagram.

6.3.3 Discussion

Importantly, and perhaps obviously given the inability of the described model to capture more nuanced relationships between concepts, the goal of *TitleGen* is to generate ideas or prompts for discussion, upon which collaboration ideas may be stimulated. Because $m_{\mathcal{A}^c}$ samples from a historical dataset using only publication titles, it is unlikely that the generated titles will be themselves innovative–but the combination of prompts generated can spark collaboration.

Future development of *TitleGen* could focus on constructing more nuanced models that combine the historical publication record with related metric of impact and disruption, such as those outlined in Chapter 5. Additionally, methods from NLP could be deployed to extract more semantically meaningful concepts. Both of these improvements could be used to generate collaborations that not only sample from previous publications, but that generate titles or abstracts such that the future expected impact (or disruption) of the resulting publication is maximized. Further, constraints (like geographic location, institutional affiliation, or language) could be used to maximize the relevancy of the results, making them actually actionable.

Such a model could be deployed in combination with a new funding model, such as that outlined in Section 5.4. This combined platform could learn from the historical publication record to identify those collaborations (or "edges" between authors) that would be most likely to generate innovations in specific areas of interest. Then, funding could be deployed as a "carrot" to incentivize the coresponding researchers to explore the joint project–closing the loop between publication, impact metrics, new project identification, and research funding.

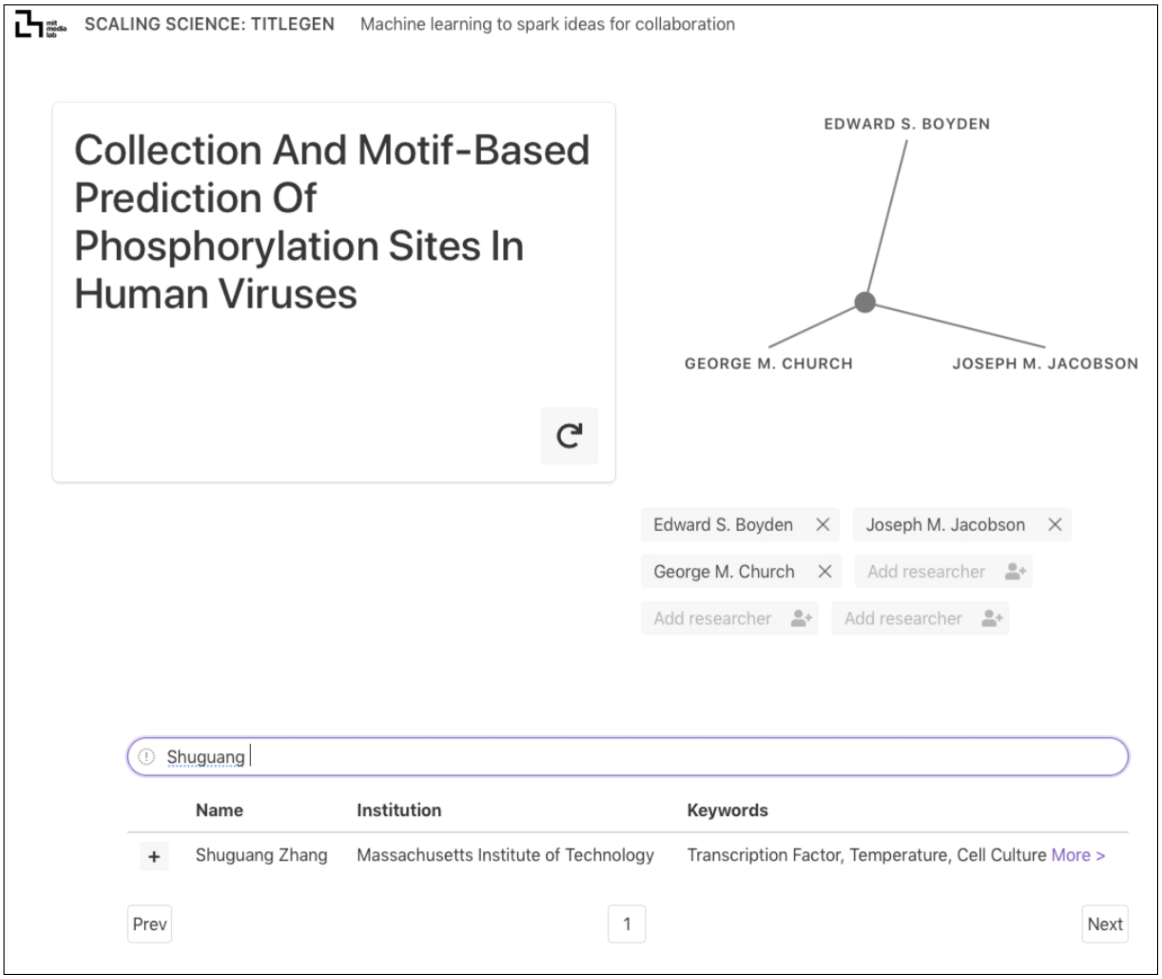


Nanoscale Imaging Of Cellular Dopamine Efflux Using A Fluorescent Nanosensor Array

Shuguang Zhang ✕ Edward S. Boyden ✕
Joseph M. Jacobson ✕ George M. Church ✕
David Botstein ✕

	Name	Institution	Keywords
+	Joi Ito	Massachusetts Institute of Technology	Media Lab, Democracy, Ethics More >
+	Robert B. Morin	Eli Lilly and Company	Organic Chemistry, Stereochemistry, Cephalosporin
+	Amy Brown	University of Pennsylvania	Molecular Biology, Stat6, Creb1
+	John Sophian	Salisbury University	Renal Function, Medicine, Blood Flow More >
+	M. S. Jackson	Royal Institute of Technology	Morphology, Physics, Proton More >
+	Robert M. Jeresaty	University of Connecticut	Surgery, Anesthesia, Medicine More >

Figure 6-3: A screenshot of the *TitleGen* application. *TitleGen* provides functionality to (A) search the dataset for up to six unique authors, comprising the set $\mathcal{A}^c$ (B) adjust the weighting w_i for each author i selected, while maintaining the constraint $\sum_i w_i = 1$, and (D) repeatedly sample from the dynamically created Markov model $m_{\mathcal{A}^c}$ to generate potential collaboration-stimulating ideas.



Collection And Motif-Based Prediction Of Phosphorylation Sites In Human Viruses

Shuguang

	Name	Institution	Keywords
+	Shuguang Zhang	Massachusetts Institute of Technology	Transcription Factor, Temperature, Cell Culture More >

Figure 6-4: Screenshot demonstrating the usage of *TitleGen*. A set of three researchers has been selected via search, and the distribution has been adjusted to non-default weighting. The combined model $m_{\mathcal{A}^c}$ clearly has learned from the publication history of the associated researchers, as the generated title demonstrates.

There always comes a time when one must choose between contemplation and action. This is called becoming...
Such wrenches are dreadful. But for a proud heart there can be no compromise. There is God or time, that cross or this sword. This world has a higher meaning that transcends its worries, or nothing is true but those worries. One must live with time and die with it, or else elude it for a greater life...
I tell you, tomorrow you will be mobilized.

The Myth of Sisyphus

7

Looking Toward the Future

The explosion of scientific data, as well as other relevant, if not specifically scientific data, is both a challenge and an opportunity: On one hand, the ability to parse the information produced, even in a single field of interest, has become a Sisyphean task– making human-based resource allocation decisions (e.g. choosing new faculty, which faculty to promote, which papers to publish, which technologies to commercialize) ever-more distant from optimal. On the other hand, the vast amount of scientific data produced annually holds, as demonstrated both by recent literature[122, 24, 92] and in this thesis, highly-valuable information that, if correctly parsed and interpreted, could become actionable—helping us understand the context and evolution of science, as well as the possibilities for future development, on a more vast and nuance scale than ever before.

While the potential of the integration of data-driven technologies to our scientific and commercialization workflows is difficult to overstate, it is not currently a leading focus for machine learning researchers. This is due at least in part to structural bottlenecks and existing incentive misalignment in the publishing, discussed in detail in Chapter 3, wherein the existing metrics used to evaluate impact and innovation–despite widespread acknowledgement of their diverse failings [42, 117, 98, 121]–are deeply entrenched into the academic and scientific ecosystems.

Augmenting the difficulty of this research is the related, but distinct, problem of knowledge infrastructure ownership. As addressed in Chapter 4, the current scientific knowledge infrastructure–despite being funded in large part via governmental and philanthropic avenues–has been captured by institutions with incentives that, for a variety of reasons (not the least of which is fiduciary duty), are often directly mis-aligned with those of the scientific community and society generally. These infrastructures not only are required for the development of scientific literacy (which is, over time, changing due to increasing adoption of a variety of Open Access approaches), but especially for truly valuable and widespread research into the data produced by the scientific ecosystem. For example, even if datasets are made openly available, the critical linkages between these diverse data silos (e.g. articles and the corresponding data from journals owned by different publishers, patents, funding agencies, clinical trials, etc.) are kept proprietary. The release of the Microsoft Academic Graph[79] and the development of Lens.org [59] are both indications of increasing acknowledgement of this problem–although neither has yet reached sufficient scale (as measured by the breadth of data captured and size of audience) or quality (as measured by the identification of linkages and disambiguation of entities) to spur widespread adoption.

In this thesis, I have attempted to not only outline the innovation-styming issues touched on in the previous two paragraphs (Chapters 3 and 4), but have also demonstrated the potential of deploying methods from artificial intelligence to non-standard domains. This includes molecular dynamics simulations of biomolecular systems (Chapter 2), and more notably and directly relevant, scientific ecosystem knowledge graphs (Chapter 3). In both cases, there were key structural and organizational hurdles–due to the lack of attention in this area, and/or the cross-incentives underlying the general field, collecting and structuring the data of interest was a key concerns. For example, in Chapter 5, the underlying data–despite being generally available–needed to be retrieved, organized, and then structured into a graph-structured database for the deployment of graph-traversal and network-analysis methodologies. However, when these "activation barriers" were surmounted, the insights derived were of exceptional interest, and of direct relevance to the areas of interest. For example, by generating an early-warning signal for research of likely future impact, I have enabled the downstream utilization of that signal by researchers, research organizations, and funding agencies. Further, the quantitative nature of the result forms the foundation upon-which additional innovations can be build–including portfolio-theoretic platforms for optimizing scientific funding (Section 5.4).

Another crucial concern is that, once such large-scale platforms are built, they will be re-captured by institutions in a repeat of the trends outlined in Chapter 4. This is a failure-mode that has been demonstrated consistently throughout history–and forms, in large part, the motivation for modern anti-trust legislation. On the other hand, there exists another failure mode: If the underlying infrastructure is not developed in an appropriate way, and appropriate usage licenses crafted, the engine of capitalism will not be sufficiently harnessed to develop and deploy applications "on top" of the crucial underlying knowledge. The standards underlying the Internet Protocol Suite (TCP/IP), most notably the Transmission Control Protocol (TSP) and Internet Protocol (IP), along with the subsequent historically unique wave of technology and internet development, can serve as a motivating success case in alignment with which future approaches and standards could be crafted.

Given this, there is a need to develop foundational knowledge infrastructure, upon which the next-generation of machine-augmented scientific tools can be developed. As a first step down this path, and especially to stimulate further large-scale development, I have designed and launched the *Scaling Science Explore* and *TitleGen* tools. *Scaling Science Explore* former provides a visualization front-end through which the exploration of the scientific literature on the basis of more complex metrics is possible (that is, not just on the bases of citations, h-Index, and Journal Impact Factor). Further, *Scaling Science Expore* can be used by researchers to visualize or publicize their own scientific metrics. *TitleGen* provides a first-step towards real-world, machine-agumented stimulation of collaboration–using historical scientific data to train models that can proactively help us explore the deeply complex landscape of science in a more efficient way.

These results hold intriguing implications for the future of scientific funding–both for research and for technology commercialization. If, by building models of the type described in this thesis, we are able to gain insight into the future impact of a new technology, author, or idea–then it is natural for us to leverage this signal when deciding what, or whom, to fund. Further, this "quantification" of the research system could lead to new funding-related projects; for example, if the top research innovations can be identified prospectively, why not create financial products that dynamically allocate capital across the related researchers using the signals provided? Further, why not dynamically adjust these funding models through the application of reinforcement learning? Also interesting is the integration of commercialization-related data into these models; if patterns indicitive not just of future scientific impact, but of future commercial potential, can be identified, then can diversified, perhaps collateralized, funding products be designed to channel for-profit capital into the research ecosystem? Can such projects be structured so that not just the inventor of the technology commercialized, but also the researchers upon whose innovation that technology depended, are also compensated–and thus incentivized to continue more basic research, even if it is several degrees removed from the final product that is commercialized?

The application of such methods, especially if more "black box" machine learning approaches are deployed, must be conducted carefully. The integration of large-scale machine learning with human society in other domains is already well underway–most notably through social media platforms. From these real-world experiences, it is obvious that there is room for malicious actors to leverage such systems for their own gain–and so the deployment of such systems must be conducted carefully, incrementally, and in close collaboration with, and supervision by, humans. However, it is difficult to imagine a scenario in which we are able to effectively operate in a world with both rapidly expanding scientific complexity and limited resources without designing tools to help us–and the potential held by the usage of such tools, including a possible paradigm change in the rate of scientific innovation, in our opinion, makes the challenge worth the risk.

A

Supplementary Figures

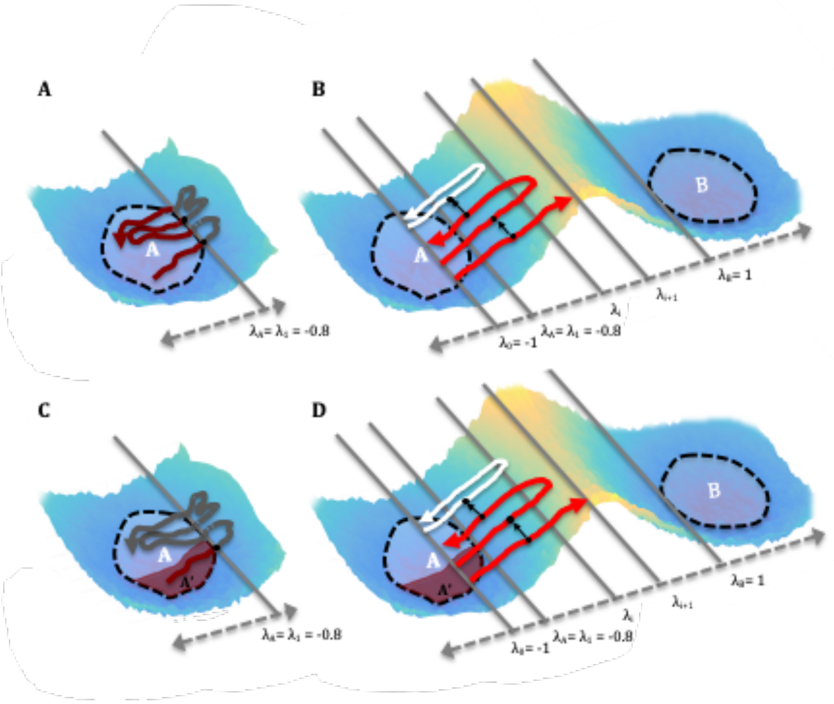


Figure A-1: (A) Illustration of computation of the TIS flux factor. The red and gray line represents a long molecular dynamics trajectory originating in region A. Portions of the trajectory in red indicate the time points in region A used to normalize the flux factor. Black dots represent effective crossings of the λ_A interface. (B) Illustration of computation of a $P(\lambda_{i+1} \mid \lambda_i)$ ensemble. Each red and white line indicates an attempted shooting move. Black dots indicate shooting points. Red lines indicate accepted shooting moves, while white lines indicate rejected shooting moves. (C) Illustration of procedure used to compute the constrained flux factor. The dark red region indicates the reactive subregion A' identified using machine learning. Portions of the trajectory in red indicate the time point in either region A' used to compute the constrained flux factor. Black dots represent effective crossings of the λ_A interface. (D) Illustration of a constrained $P(\lambda_{i+1} \mid \lambda_i)$ ensemble. The dark red region indicates the reactive subregion A' identified using machine learning.

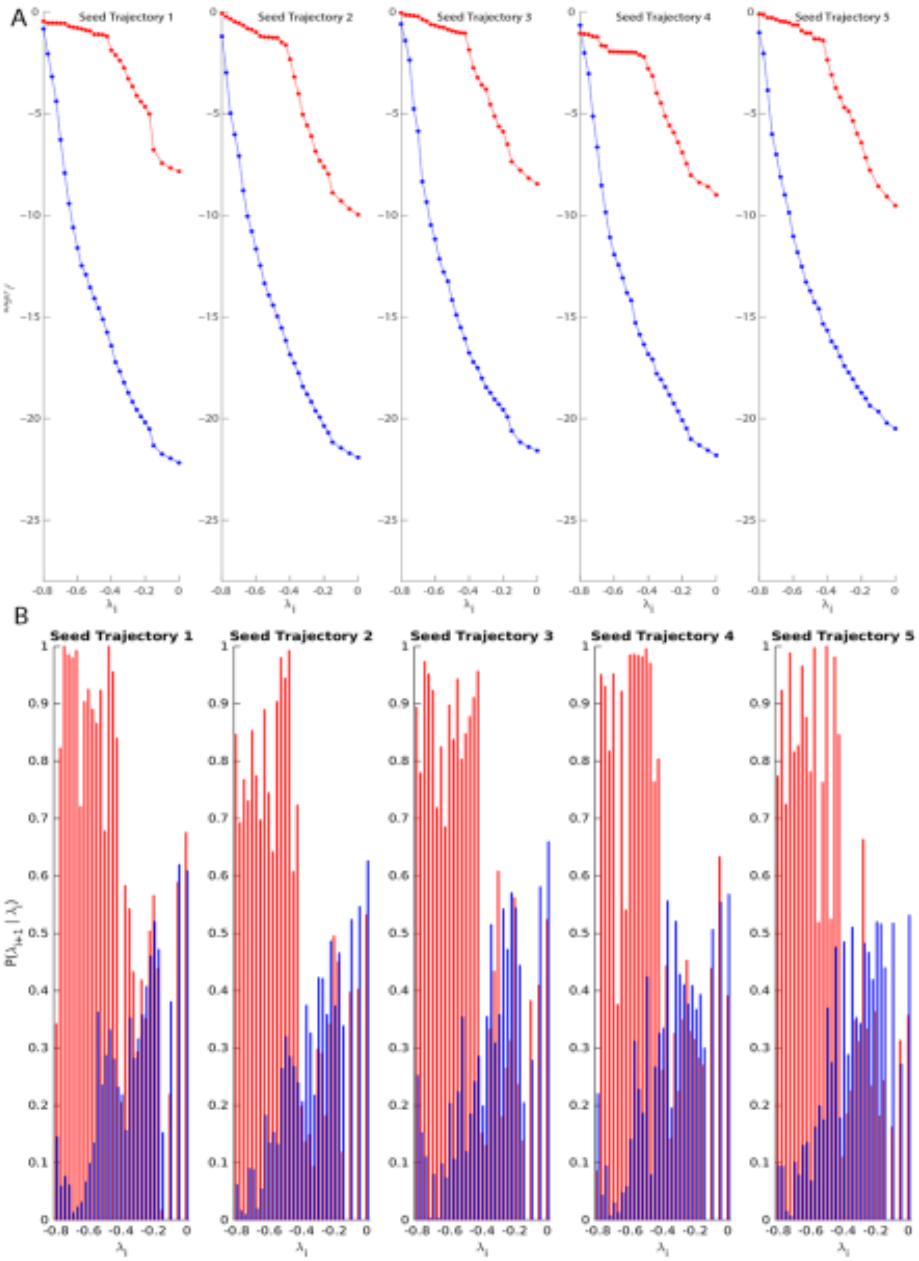


Figure A-2: (A) Cumulative $\log(P)$ for increasing interface placement for each of the 5 seed trajectories tested. Red lines indicate trajectories sampled with the reactant basin constrained to only include the region where the 10 feature classifier evaluated to true. Blue lines indicate unconstrained control simulations. (B) Individual values of $P(\lambda_{i+1} \mid \lambda_i)$ for each λ_i ensemble computed. Error bars correspond to two standard errors of the mean across three independent Markov chains at each λ_i ensemble. Red bars indicate test simulations, while blue bars indicate unconstrained control simulations.

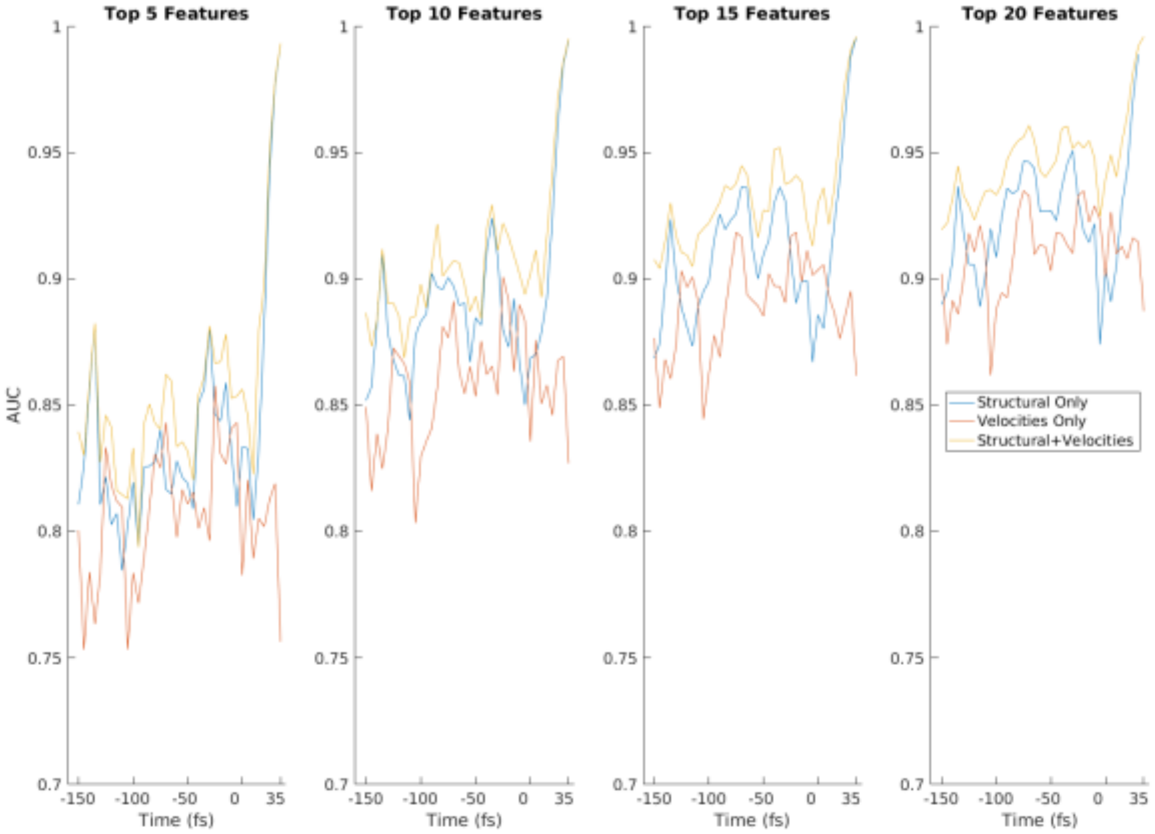
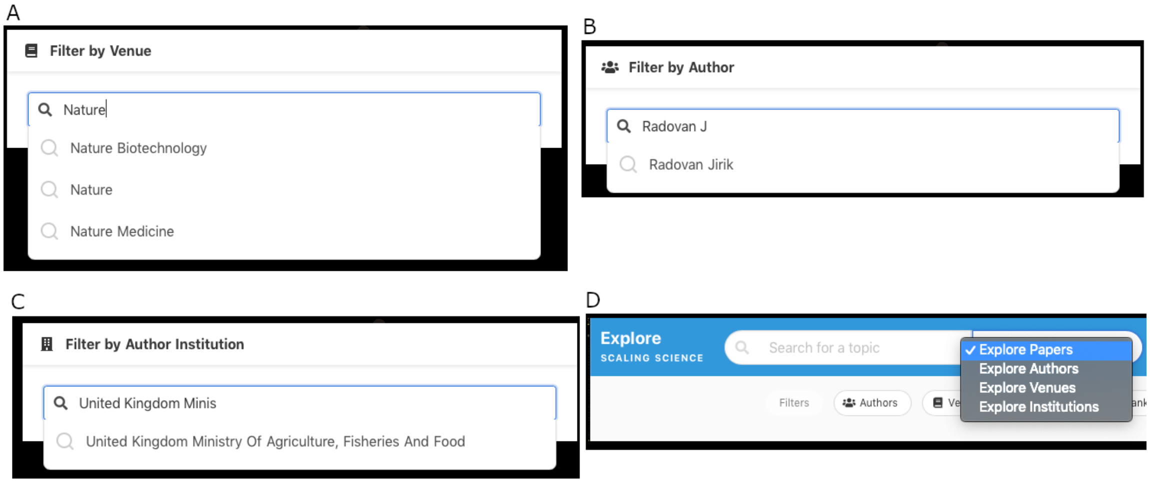


Figure A-3: Representative structures for the reactive cluster and corresponding almost-reactive clusters described in Figure 2-3B-E. Feature numbering corresponds to that of Table B.1. (A) Representative structures from all five reactive clusters. Representative structures from (B) cluster 1, (C) cluster 2, (D) cluster 3, (E) cluster 4, (F) cluster (5) and their corresponding almost-reactive clusters, respectively. In all panels, magenta corresponds to cluster 1, cyan corresponds to cluster 2, green corresponds to cluster 3, yellow corresponds to cluster 4, orange corresponds to cluster 5 and gray corresponds to the corresponding almost-reactive cluster for the reactive cluster shown in each histogram. In all panels, structures were aligned to minimize the root mean square difference between the two magnesium centers.



Nature

Nature Biotechnology

Nature

Nature Medicine

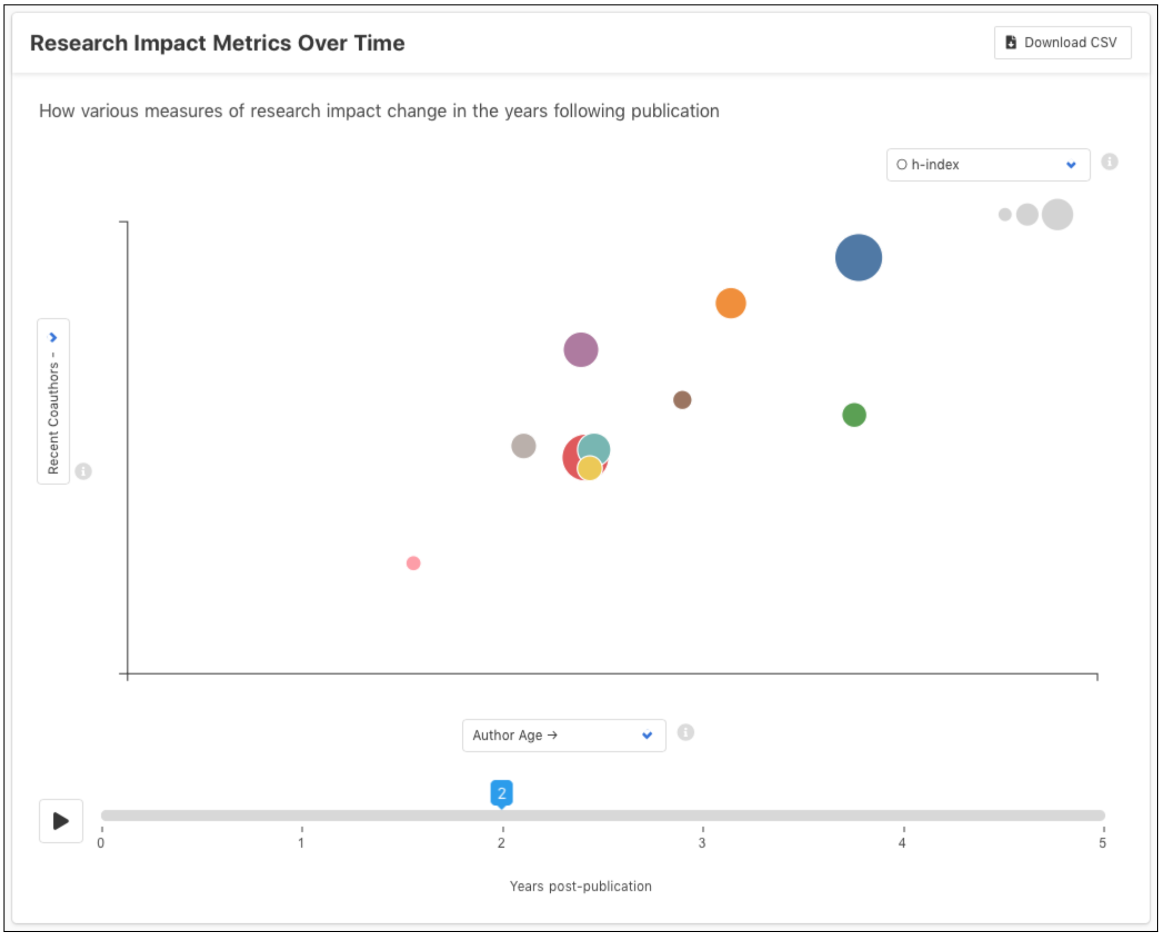
Radovan J

Radovan Jirik

United Kingdom Minis

United Kingdom Ministry Of Agriculture, Fisheries And Food

Figure A-4: Four separate screenshots showing the filtering mechanisms/capabilities of *Scaling Science Explore*, including (A) filtering by venue (journal), (B) by author, (C) by institutional affiliation of the authors, or (D) by keyword-based search across all of those options.



Research Impact Metrics Over Time

How various measures of research impact change in the years following publication

Figure A-5: Detailed screenshot of *Scaling Science Explore*'s four-dimensional visualization. Three dimensions are fully user-supplied (the x-axis, y-axis, and data point size), and the fourth dimension includes the visualization of these three metrics across time, on a per-year basis from the year of publication to five-years post-publication.

B

Supplementary Tables

Table B.1: Top 30 consensus features for the –150 to 0 fs time window. Feature rank indicates ranking according to the number of occurrences in the 20 LASSO-selected feature sets.

Rank	Feature Name	Feature Type	Occ.
1	Dist GLU319/$O_{\epsilon1}$,AC6/C5	Substrate-environment	24
2	Dist MG6/H32,AC6/O6	Substrate-environment	23
3	Dist MG6/H26,AC6/O6	Substrate-environment	22
4	Ang MG6/H31,MG6/O19,MG6/M17	Water-metal	20
5	Dist MG6/H28,AC6/O6	Substrate-environment	19
6	Dist AC6/O8,GLU496/$H_{\epsilon2}$	Substrate-environment	19
7	Ang NDP/C4N,NDP/N1N,NDP/C1NQ	Intra-cofactor	19
8	Dist MG6/H27,AC6/O6	Substrate-environment	18
9	Dist AC6/C5,AC6/C4	Intra-substrate	18
10	Ang MG6/M17,AC6/O6,MG6/M16	Substrate-environment	18
11	Dihe AC6/C5,AC6/C4,AC6/C7,AC6/C9	Intra-substrate	17
12	Ang AC6/O6,MG6/M16,AC6/O3	Substrate-environment	17
13	Dist MG6/O20,MG6/M17	Water-metal	17
14	Ang AC6/C1,AC6/C4,AC6/C7	Intra-substrate	16
15	Dist AC6/C7,AC6/C9	Intra-substrate	16
16	Ang AC6/O8,MG6/M17,AC6/O6	Substrate-environment	16
17	Ang GLN136/$N_{\epsilon2}$,GLN136/$H_{\epsilon22}$,NDP/O7N	Other environment	16
18	Ang AC6/C5,AC6/C7,AC6/C9	Intra-substrate	15
19	Dist MG6/M17,AC6/O6	Substrate-environment	15
20	Ang MG6/H29,MG6/O18,MG6/M17	Water-metal	15
21	Dist GLN136/$N_{\epsilon2}$,NDP/O7N	Other environment	15
22	Dist MG6/H31,AC6/O6	Substrate-environment	13
23	Dist AC6/C4,AC6/C7	Intra-substrate	13
24	Dist AC6/O6,MG6/M16	Substrate-environment	13
25	Dist AC6/C1,AC6/C4	Intra-substrate	12
26	Ang MG6/H32,MG6/O19,MG6/M17	Water-metal	12
27	Dist MG6/M16,AC6/O3	Substrate-environment	10
28	Dist MG6/O19,MG6/M17	Water-metal	10
29	Ang MG6/H23,MG6/O22,MG6/M16	Water-metal	10
30	Ang MG6/H25,MG6/O21,MG6/M16	Water-metal	10

Table B.2: Feature names, feature indices and feature types computed at each time . point. Residue name AC6 refers to the substrate, residue name NDP refers to the NADPH cofactor, and the residue name MG6 refers to the 5 active site waters and two magnesium ions. Structural representations of features are shown in Figure 2-4.

Feature Index	Feature Name	Feature Type
1	Dist AC6/O2,NDP/N7N	Substrate-environment
2	Dist AC6/O2,NDP/O7N	Substrate-environment
3	Dist AC6/O3,MG6/H24	Substrate-environment
4	Dist AC6/O6,MG6/M16	Substrate-environment
5	Dist AC6/O8,GLU496/He2	Substrate-environment
6	Dist AC6/O8,MG6/M17	Substrate-environment
7	Dist GLU319/Oe1,AC6/C5	Substrate-environment
8	Dist MG6/H25,AC6/O6	Substrate-environment
9	Dist MG6/H26,AC6/O6	Substrate-environment
10	Dist MG6/H27,AC6/O6	Substrate-environment
11	Dist MG6/H28,AC6/O6	Substrate-environment
12	Dist MG6/H31,AC6/O6	Substrate-environment
13	Dist MG6/H32,AC6/O6	Substrate-environment
14	Dist MG6/M16,AC6/O3	Substrate-environment
15	Dist MG6/M17,AC6/O6	Substrate-environment
16	Dist NDP/H4N2,AC6/C4	Substrate-environment
17	Ang AC6/O6,MG6/M16,AC6/O3	Substrate-environment
18	Ang AC6/O8,MG6/M17,AC6/O6	Substrate-environment
19	Ang MG6/M17,AC6/O6,MG6/M16	Substrate-environment
20	Dist AC6/C1,AC6/C4	Intra-substrate
21	Dist AC6/C1,AC6/O2	Intra-substrate
22	Dist AC6/C1,AC6/O3	Intra-substrate
23	Dist AC6/C4,AC6/C7	Intra-substrate

Table B.2 continued from previous page

Feature Index	Feature Name	Feature Type
24	Dist AC6/C4,AC6/O6	Intra-substrate
25	Dist AC6/C5,AC6/C4	Intra-substrate
26	Dist AC6/C5,AC6/C7	Intra-substrate
27	Dist AC6/C7,AC6/C9	Intra-substrate
28	Dist AC6/C7,AC6/O8	Intra-substrate
29	Ang AC6/C1,AC6/C4,AC6/C7	Intra-substrate
30	Ang AC6/C4,AC6/C7,AC6/C5	Intra-substrate
31	Ang AC6/C4,AC6/C7,AC6/C9	Intra-substrate
32	Ang AC6/C5,AC6/C4,AC6/C1	Intra-substrate
33	Ang AC6/C5,AC6/C7,AC6/C9	Intra-substrate
34	Dihe AC6/C1,AC6/C5,AC6/C7,AC6/C4	Intra-substrate
35	Dihe AC6/C5,AC6/C4,AC6/C7,AC6/C9	Intra-substrate
36	Dist NDP/H4N2,NDP/C4N	Intra-cofactor
37	Dist NDP/N7N,NDP/O2N	Intra-cofactor
38	Ang NDP/C4N,NDP/N1N,NDP/C1NQ	Intra-cofactor
39	Ang NDP/C6N,NDP/C3N,NDP/C7N	Intra-cofactor
40	Ang NDP/N7N,NDP/H72N,NDP/O2N	Intra-cofactor
41	Dihe NDP/C2N,NDP/C3N,NDP/C7N,NDP/N7N	Intra-cofactor
42	Dihe NDP/C2NQ,NDP/C1NQ,NDP/N1N,NDP/C6N	Intra-cofactor
43	Dihe NDP/C4N,NDP/C3N,NDP/C7N,NDP/O7N	Intra-cofactor
44	Dihe NDP/H1NQ,NDP/C1NQ,NDP/N1N,NDP/C2N	Intra-cofactor
45	Dist MG6/O18,MG6/M17	Water-metal
46	Dist MG6/O19,MG6/M17	Water-metal
47	Dist MG6/O20,MG6/M17	Water-metal
48	Dist MG6/O21,MG6/M16	Water-metal
49	Dist MG6/O22,MG6/M16	Water-metal

Table B.2 continued from previous page

Feature Index	Feature Name	Feature Type
50	Ang MG6/H23,MG6/O22,MG6/M16	Water-metal
51	Ang MG6/H24,MG6/O22,MG6/M16	Water-metal
52	Ang MG6/H25,MG6/O21,MG6/M16	Water-metal
53	Ang MG6/H26,MG6/O21,MG6/M16	Water-metal
54	Ang MG6/H27,MG6/O20,MG6/M17	Water-metal
55	Ang MG6/H28,MG6/O20,MG6/M17	Water-metal
56	Ang MG6/H29,MG6/O18,MG6/M17	Water-metal
57	Ang MG6/H30,MG6/O18,MG6/M17	Water-metal
58	Ang MG6/H31,MG6/O19,MG6/M17	Water-metal
59	Ang MG6/H32,MG6/O19,MG6/M17	Water-metal
60	Dist GLU496/Oe2,GLU496/He2	Other environment
61	Dist GLN136/Ne2,NDP/O7N	Other environment
62	Dist MG6/H25,MG6/O21	Other environment
63	Dist MG6/H26,MG6/O21	Other environment
64	Dist MG6/H27,MG6/O20	Other environment
65	Dist MG6/H28,MG6/O20	Other environment
66	Dist MG6/H31,MG6/O19	Other environment
67	Dist MG6/H32,MG6/O19	Other environment
68	Ang GL136/Ne2,GLN136/He22,NDP/O7N	Other environment

Table B.3: Mean standardized logistic regression coefficients fit to classifier trained using the top 30 most consistently predictive features between -150 and 0 fs (listed in Table B.1 and illustrated structurally in Figure 2-7) at the -150, -100, -50 and 0 fs time points relative to the last trough in the order parameter prior to the prospective catalytic event. Coefficients shown represent the mean values across 5 cross-validation partitions.

Standardized Regression Coefficient	Time Before Last Trough			
	-150 fs	-100 fs	-50 fs	0 fs
β_0	-0.059	-0.195	-0.269	-0.094
β_1	-0.361	-0.527	0.354	0.470
β_2	-0.198	-0.569	-0.374	0.874
β_3	-0.303	-0.957	-0.497	-0.035
β_4	0.615	0.706	0.117	0.453
β_5	0.094	0.069	0.401	-0.477
β_6	-0.365	-0.265	-0.147	-0.613
β_7	0.273	-0.251	-0.423	-0.397
β_8	0.293	-0.428	-1.134	-0.356
β_9	0.068	0.446	0.533	-1.030
β_{10}	0.318	-1.060	-0.666	-0.025
β_{11}	-0.307	0.058	-1.379	-0.289
β_{12}	-0.723	0.414	0.179	-0.510
β_{13}	0.236	-0.129	0.610	0.050
β_{14}	-0.256	0.214	-0.348	-0.107
β_{15}	-0.132	-0.460	-0.227	0.269
β_{16}	-0.330	-0.237	1.049	0.106
β_{17}	0.065	0.302	0.137	0.039
β_{18}	0.193	-0.704	0.665	0.026
β_{19}	-0.426	0.252	-0.425	0.007
β_{20}	0.033	0.141	0.477	0.704
β_{21}	0.319	-0.327	-0.471	-0.013
β_{22}	-0.135	-0.630	-0.162	-1.100
β_{23}	0.790	0.281	-0.089	0.200
β_{24}	-0.048	-0.179	-0.127	-0.014
β_{25}	0.083	-0.047	-0.182	0.504
β_{26}	0.592	0.592	0.434	-0.244
β_{27}	0.142	0.093	0.241	-0.944
β_{28}	-0.208	0.477	0.437	-0.083
β_{29}	-0.148	-0.370	-0.327	-0.061
β_{30}	0.183	0.151	0.338	-0.174

Table B.4: Top 20 atomic velocity magnitudes at the 0-fs time point ranked by individual AUC. The feature set was comprised of the velocity magnitudes of the 341 atoms within 5 ångstroms of the migrating methyl, AC6/C5 (including all hydrogens). Note that of these 341 velocities, only 17 exhibited individual AUCs greater than 0.60 at the "last trough" of the pre-launch window, and of these 17, 5 involved atoms included in the "consensus set".

Rank	Atom Name	AUC	Involved in Consensus Feature Set?
1	AC6/O6	0.854	Yes
2	AC6/C4	0.738	Yes
3	Glu319/$O_{\epsilon1}$	0.6713	Yes
4	AC6/H12	0.6687	No
5	NDP/P2A	0.6435	No
6	NDP/H2A	0.6406	No
7	MET254/C	0.637	No
8	THR520/HN	0.6363	No
9	AC6/C7	0.6213	Yes
10	GLU319/C	0.6187	No
11	GLN136/CA	0.6154	No
12	NDP600/O3	0.6082	No
13	GL6319/$O_{\epsilon2}$	0.6077	No
14	GL6319/O	0.6034	No
15	GLU496/$O_{\epsilon2}$	0.6022	No
16	AC6/C1	0.6022	Yes
17	LYS252 /HG1	0.602	No
18	PRO251/O	0.5991	No
19	NDP600/H1NQ	0.5984	No
20	AC6/O8	0.5953	Yes

C
Publications

1. **Learning on Knowledge Graph Dynamics Provides Early Warning of Impactful Research**
In review at Nature Biotechnology.

2. **Artificial Intelligence in Publishing**
Weis JW, Brand A. Artificial Intelligence in Publishing. Charleston Briefings. *Accepted and pending publication.*

3. **The Case for an Institutionally Owned Knowledge Infrastructure**
Weis JW, Brand A, Ito J. The Case for an Institutionally Owned Knowledge Infrastructure. Inside Higher Ed. 2020 Jan 7.

4. **Machine Learning Identifies Chemical Characteristics That Promote Enzyme Catalysis**
Bonk BM, Weis JW, Tidor B. Machine Learning Identifies Chemical Characteristics That Promote Enzyme Catalysis. Journal of the American Chemical Society. 2019 Feb 14; 141(9):4108-18.

5. **Evaluating disparities in the US technology transfer ecosystem to improve bench to business translation**
Weis JW, Bashyam A, Ekchian GJ, Paisner K, Vanderford NL. Evaluating disparities in the US technology transfer ecosystem to improve bench to business translation. F1000Research. 2018;7.

6. **Tackling regional public health issues using mobile health technology**

Pathanasethpong A, Soomlek C, Morley K, Morley M, Polpinit P, Dagan A, Weis JW, Celi LA. Tackling regional public health issues using mobile health technology: event report of an mHealth hackathon in Thailand. JMIR mHealth and uHealth. 2017;5(10):e155.

7. **Genotype specification language**

Wilson EH, Sagawa S, Weis JW, Schubert MG, Bissell M, Hawthorne B, Reeves CD, Dean J, Platt D. Genotype specification language. ACS Synthetic Biology. 2016 Jun 17;5(6):471-8.

Bibliography

[1] Daniel E Acuna, Stefano Allesina, and Konrad P Kording. *Predicting scientific success*, volume 489. 2012.

[2] Liz Allen, Jo Scott, Amy Brand, Marjorie Hlava, and Micah Altman. Publishing: Credit where credit is due. *Nature*, 508(7496):312–313, April 2014.

[3] Alphabet Investor Relations. `https://abc.xyz/`. Accessed: 2020-01-02.

[4] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182, 2016.

[5] Apple - Investor Relations - Financial Information. `https://investor.apple.com/investor-relations/financial-information/`. Accessed: 2020-01-02.

[6] Bahman Bahmani, Abdur Chowdhury, and Ashish Goel. *Fast incremental and personalized PageRank*, volume 4. 2010.

[7] David Baker. An exciting but challenging road ahead for computational enzyme design: Computational Enzyme Design. *Protein Science*, 19(10):1817–1819, October 2010.

[8] Ofer Bar-Yosef. The Natufian culture in the Levant, threshold to the origins of agriculture. *Evolutionary Anthropology: Issues, News, and Reviews*, 6(5):159–177, 1998.

[9] Jodi E. Basner and Steven D. Schwartz. How Enzyme Dynamics Helps Catalyze a Reaction in Atomic Detail: A Transition Path Sampling Study. *Journal of the American Chemical Society*, 127(40):13822–13831, October 2005.

[10] Sabine Bastian, Xiang Liu, Joseph T. Meyerowitz, Christopher D. Snow, Mike M.Y. Chen, and Frances H. Arnold. Engineered ketol-acid reductoisomerase and alcohol dehydrogenase enable anaerobic 2-methylpropan-1-ol production at theoretical yield in Escherichia coli. *Metabolic Engineering*, 13(3):345–352, May 2011.

[11] Federico Battiston, Federico Musciotto, Dashun Wang, Albert-László Barabási, Michael Szell, and Roberta Sinatra. *Taking census of physics*, volume 1. 2019.

[12] H. M. Berman. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, January 2000.

[13] Frances C. Bernstein, Thomas F. Koetzle, Graheme J.B. Williams, Edgar F. Meyer, Michael D. Brice, John R. Rodgers, Olga Kennard, Takehiko Shimanouchi, and Mitsuo Tasumi. The protein data bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, 112(3):535–542, May 1977.

[14] bioRxiv preprints can now be submitted directly to leading research journals. `https://phys.org/news/2016-01-biorxiv-preprints-submitted-journals.html`. Accessed: 2020-01-02.

[15] V. Biou. The crystal structure of plant acetohydroxy acid isomeroreductase complexed with NADPH, two magnesium ions and a herbicidal transition state analog determined at 1.65Aresolution. *The EMBO Journal*, 16(12):3405–3415, June 1997.

[16] Nicholas Bloom, Charles Jones, John Van Reenen, and Michael Webb. Are Ideas Getting Harder to Find? Technical Report w23782, National Bureau of Economic Research, Cambridge, MA, September 2017.

[17] Kevin J Boudreau, Eva C Guinan, Karim R Lakhani, and Christoph Riedl. Looking Across and Looking Beyond the Knowledge Frontier: Intellectual Distance, Novelty, and Resource Allocation in Science. *Manage. Sci.*, 62(10):2765–2783, October 2016.

[18] Amy Brand. Beyond mandate and repository, toward sustainable faculty self-archiving. *Learn. Publ.*, 25(1):29–34, 2012.

[19] Amy Brand, Liz Allen, Micah Altman, Marjorie Hlava, and Jo Scott. Beyond authorship: attribution, contribution, collaboration, and credit. *Learn. Publ.*, 28(2):151–155, 2015.

[20] B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry*, 30(10):1545–1614, July 2009.

[21] Bernard R. Brooks, Robert E. Bruccoleri, Barry D. Olafson, David J. States, S. Swaminathan, and Martin Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4(2):187–217, 1983.

[22] Thomas C. Bruice. A View at the Millennium: the Efficiency of Enzymatic Catalysis. *Accounts of Chemical Research*, 35(3):139–148, March 2002.

[23] Thomas C. Bruice and Felice C. Lightstone. Ground State and Transition State Contributions to the Rates of Intramolecular and Enzymatic Reactions. *Accounts of Chemical Research*, 32(2):127–136, February 1999.

[24] Keith T. Butler, Daniel W. Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555, July 2018.

[25] Ewen Callaway. Preprints come to life. *Nature*, 503(7475):180, November 2013.

[26] Chang-Ting Chen and James C. Liao. Frontiers in microbial 1-butanol and isobutanol production. *FEMS Microbiology Letters*, 363(5):fnw020, March 2016.

[27] P Chen, H Xie, S Maslov, and S Redner. Finding scientific gems with Google's PageRank algorithm. *J. Informetr.*, 1(1):8–15, January 2007.

[28] Aaron Clauset, Samuel Arbesman, and Daniel B. Larremore. Systematic inequality and hierarchy in faculty hiring networks. *Science Advances*, 1(1):e1400005, February 2015.

[29] Angel Cruz-Roa, Hannah Gilmore, Ajay Basavanhally, Michael Feldman, Shridar Ganesan, Natalie NC Shih, John Tomaszewski, Fabio A González, and Anant Madabhushi. Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach for quantifying tumor extent. *Scientific reports*, 7:46450, 2017.

[30] Alex Csiszar. How Lives Became Lists and Scientific Papers Became Data: Cataloguing Authorship during the Nineteenth Century - Corrigendum. *Br. J. Hist. Sci.*, 50(3):567, September 2017.

[31] Douglas J Cumming and Na Dai. *Local Bias in Venture Capital Investments.*

[32] Phil Davis. *The Emergence of a Citation Cartel - The Scholarly Kitchen.* April 2012.

[33] Christoph Dellago, Peter G. Bolhuis, Félix S. Csajka, and David Chandler. Transition path sampling and the calculation of rate constants. *The Journal of Chemical Physics*, 108(5):1964–1977, February 1998.

[34] Michael J. S. Dewar, Eve G. Zoebisch, Eamonn F. Healy, and James J. P. Stewart. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *Journal of the American Chemical Society*, 107(13):3902–3909, June 1985.

[35] Renaud Dumas, Valérie Biou, Frédéric Halgand, Roland Douce, and Ronald G. Duggleby. Enzymology, Structure, and Dynamics of Acetohydroxy Acid Isomeroreductase. *Accounts of Chemical Research*, 34(5):399–408, May 2001.

[36] Khalid El-Arini and Carlos Guestrin. *Beyond keyword search.* 2011.

[37] Douglas C Engelbart. Augmenting human intellect: A conceptual framework. *Menlo Park, CA*, 1962.

[38] James A Evans and Andrey Rzhetsky. Advancing science through mining libraries, ontologies, and communities. *J. Biol. Chem.*, 286(27):23659–23666, July 2011.

[39] Faculty of 1000. `https://f1000.com`. Accessed: 2020-01-02.

[40] Ferric C Fang and Arturo Casadevall. Research Funding: the Case for a Modified Lottery. *MBio*, 7(2):e00422–16, April 2016.

[41] Iztok Fister, Iztok Fister, and Matjaž Perc. *Toward the Discovery of Citation Cartels in Citation Networks*, volume 4. 2016.

[42] Georg Franck. Scientific Communication–A Vanity Fair? *Science*, 286(5437):53–55, October 1999.

[43] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople. *Gaussian 03.*

[44] Lawrence D Fu and Constantin F Aliferis. *Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature*, volume 85. 2010.

[45] Russell J Funk and Jason Owen-Smith. *A Dynamic Network Measure of Technological Change*, volume 63. 2017.

[46] E Garfield. Citation indexes for science; a new dimension in documentation through association of ideas. *Science*, 122(3159):108–111, July 1955.

[47] Eugene Garfield. The history and meaning of the journal impact factor. *JAMA*, 295(1):90–93, January 2006.

[48] Paul Ginsparg. It was twenty years ago today . August 2011.

[49] B. Glass. Science: Endless Horizons or Golden Age? *Science*, 171(3966):23–29, January 1971.

[50] Paul Gompers, William Gornall, Steven Kaplan, and Ilya Strebulaev. *How Do Venture Capitalists Make Decisions?* 2016.

[51] Ruggero Gramatica and Ruth Pickering. *Start-up story Yewno: an AI-driven path to a knowledge-based future*, volume 30. 2017.

[52] Aditya Grover and Jure Leskovec. *node2vec*. 2016.

[53] Stephan C. Hammer, Anders M. Knight, and Frances H. Arnold. Design and evolution of enzymes for non-natural chemistry. *Current Opinion in Green and Sustainable Chemistry*, 7:23–30, October 2017.

[54] Rajesh K. Harijan, Ioanna Zoi, Dimitri Antoniou, Steven D. Schwartz, and Vern L. Schramm. Catalytic-site design for inverse heavy-enzyme isotope effects in human purine nucleoside phosphorylase. *Proceedings of the National Academy of Sciences*, 114(25):6456–6461, June 2017.

[55] James Heckman and Sidharth Moktan. Publishing and Promotion in Economics: The Tyranny of the Top Five. Technical report, 2018.

[56] J E Hirsch. An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci. U. S. A.*, 102(46):16569–16572, November 2005.

[57] Jing Huang and Alexander D. MacKerell. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *Journal of Computational Chemistry*, 34(25):2135–2145, September 2013.

[58] S. Hur and T. C. Bruice. The near attack conformation approach to the study of the chorismate to prephenate reaction. *Proceedings of the National Academy of Sciences*, 100(21):12015–12020, October 2003.

[59] Osmat Azzam Jefferson, Deniz Koellhofer, Ben Warren, and Richard Jefferson. The Lens MetaRecord and LensID: An open identifier system for aggregated metadata and versioning of knowledge artefacts. preprint, LIS Scholarship Archive, November 2019.

[60] Thorsten Joachims and Filip Radlinski. *Search Engines that Learn from Implicit Feedback*, volume 40. 2007.

[61] S. C. L. Kamerlin, P. K. Sharma, Z. T. Chu, and A. Warshel. Ketosteroid isomerase provides further support for the idea that enzymes work by electrostatic preorganization. *Proceedings of the National Academy of Sciences*, 107(9):4075–4080, March 2010.

[62] Margaret H Kearney and INANE Predatory Publishing Practices Collaborative. Predatory publishing: what authors need to know. *Res. Nurs. Health*, 38(1):1–3, February 2015.

[63] Bryan Kelly, Dimitris Papanikolaou, Amit Seru, and Matt Taddy. *Measuring Technological Innovation over the Long Run*. 2018.

[64] Gert Kiss, Nihan Çelebi Ölçüm, Rocco Moretti, David Baker, and K. N. Houk. Computational Enzyme Design. *Angewandte Chemie International Edition*, 52(22):5700–5725, May 2013.

[65] Shankar Kumar, John M. Rosenberg, Djamal Bouzida, Robert H. Swendsen, and Peter A. Kollman. THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry*, 13(8):1011–1021, October 1992.

[66] Edmond Y. Lau and Thomas C. Bruice. Importance of Correlated Motions in Forming Highly Reactive Near Attack Conformations in Catechol *O* -Methyltransferase. *Journal of the American Chemical Society*, 120(48):12387–12394, December 1998.

[67] R. Lerner, S. Benkovic, and P. Schultz. At the crossroads of chemistry and immunology: catalytic antibodies. *Science*, 252(5006):659–667, May 1991.

[68] Jean F. Liénard, Titipat Achakulvisut, Daniel E. Acuna, and Stephen V. David. Intellectual synthesis in mentorship determines success in academic careers. *Nature Communications*, 9(1):4840, November 2018.

[69] Wen Long, Zhichen Lu, and Lingxiao Cui. Deep learning-based feature engineering for stock price movement prediction. *Knowledge-Based Systems*, 164:163–173, 2019.

[70] Yifang Ma and Brian Uzzi. *Scientific prize network predicts who pushes the boundaries of science*, volume 115. 2018.

[71] Yoshiaki Maeda, Olga V. Makhlynets, Hiroshi Matsui, and Ivan V. Korendovych. Design of Catalytic Peptides and Proteins Through Rational and Combinatorial Approaches. *Annual Review of Biomedical Engineering*, 18(1):311–328, July 2016.

[72] Apoorva Mandavilli. Peer review: Trial by Twitter. *Nature News*, 469(7330):286–287, January 2011.

[73] Many Academics Are Eager to Publish in Worthless Journals, October 2017. Accessed: 2020-01-02.

[74] Manuel Sebastian Mariani, Matúš Medo, and François Lafond. *Early identification of important patents: Design and validation of citation network metrics*, volume 146. 2019.

[75] Manuel Sebastian Mariani, Matúš Medo, and Yi-Cheng Zhang. *Identification of milestone papers through time-balanced network centrality*, volume 10. 2016.

[76] Sergei Maslov and Sidney Redner. Promise and pitfalls of extending Google's PageRank algorithm to citation networks. *J. Neurosci.*, 28(44):11103–11105, October 2008.

[77] Marcia McNutt. The measure of research merit. *Science*, 346(6214):1155, December 2014.

[78] Bob Metcalfe. *Metcalfe's Law after 40 Years of Ethernet*, volume 46. 2013.

[79] Microsoft 2017 Annual Report. `https://www.microsoft.com/investor/reports/ar17/index.html`. Accessed: 2020-01-01.

[80] MIT DCI's Cryptocurrency Research Review #1. `https://mitcryptocurrencyresearch.substack.com/p/mit-dcis-cryptocurrency-research`. Accessed: 2020-01-06.

[81] Patricia Molina-Espeja, Javier Viña-Gonzalez, Bernardo J. Gomez-Fernandez, Javier Martin-Diaz, Eva Garcia-Ruiz, and Miguel Alcalde. Beyond the outer limits of nature by directed evolution. *Biotechnology Advances*, 34(5):754–767, September 2016.

[82] Diane Mulcahy, Bill Weeks, and Harold Bradley. *We Have Met the Enemy... and He is Us: Lessons from Twenty Years of the Kauffman Foundation's Investments in Venture Capital Funds and the Triumph of Hope Over Experience.*

[83] Nature and biotechnology. *Nat. Biotechnol.*, 37(12):1383–1383, December 2019.

[84] Georgy A. Nevinsky and Valentina N. Buneva. Natural Catalytic Antibodies - Abzymes. In Ehud Keinan, editor, *Catalytic Antibodies*, pages 505–569. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, FRG, May 2005.

[85] Joshua M Nicholson and John P A Ioannidis. *Conform and be funded*, volume 492. 2012.

[86] Not-so-deep impact. *Nature*, 435(7045):1003–1004, June 2005.

[87] Chunyang Peng, Philippe Y. Ayala, H. Bernhard Schlegel, and Michael J. Frisch. Using redundant internal coordinates to optimize equilibrium geometries and transition states. *Journal of Computational Chemistry*, 17(1):49–56, January 1996.

[88] Chunyang Peng and H. Bernhard Schlegel. Combining Synchronous Transit and Quasi-Newton Methods to Find Transition States. *Israel Journal of Chemistry*, 33(4):449–454, 1993.

[89] Joanne L. Porter, Rukhairul A. Rusli, and David L. Ollis. Directed Evolution of Enzymes for Industrial Biocatalysis. *ChemBioChem*, 17(3):197–203, February 2016.

[90] Flavien Proust-De Martin, Renaud Dumas, and Martin J. Field. A Hybrid-Potential Free-Energy Study of the Isomerization Step of the Acetohydroxy Acid Isomeroreductase Reaction. *Journal of the American Chemical Society*, 122(32):7688–7697, August 2000.

[91] PubPub - Community Publishing. `https://www.pubpub.org`. Accessed: 2020-01-03.

[92] Paul Raccuglia, Katherine C. Elbert, Philip D. F. Adler, Casey Falk, Malia B. Wenny, Aurelio Mollo, Matthias Zeller, Sorelle A. Friedler, Joshua Schrier, and Alexander J. Norquist. Machine-learning-assisted materials discovery using failed experiments. *Nature*, 533(7601):73–76, May 2016.

[93] Relx annual Reports. `https://www.relx.com/investors/annual-reports/2017`. Accessed: 2020-01-01.

[94] Retraction Watch Database. `http://retractiondatabase.org/RetractionSearch.aspx?` Accessed: 2020-01-04.

[95] Jory Z. Ruscio, Jonathan E. Kohn, K. Aurelia Ball, and Teresa Head-Gordon. The Influence of Protein Dynamics on the Success of Computational Enzyme Design. *Journal of the American Chemical Society*, 131(39):14111–14115, October 2009.

[96] S. Kashif Sadiq and Peter V. Coveney. Computing the Role of Near Attack Conformations in an Enzyme-Catalyzed Nucleophilic Bimolecular Reaction. *Journal of Chemical Theory and Computation*, 11(1):316–324, January 2015.

[97] Chris Salmon. *Click to download: How to beat the video ban.* March 2009.

[98] P O Seglen. *Why the impact factor of journals should not be used for evaluating research*, volume 314. 1997.

[99] Vedran Sekara, Pierre Deville, Sebastian E. Ahnert, Albert-László Barabási, Roberta Sinatra, and Sune Lehmann. The chaperone effect in scientific publishing. *Proceedings of the National Academy of Sciences*, 115(50):12603–12607, December 2018.

[100] Joel N. Shurkin. Engines of the mind: the evolution of the computer from mainframes to microprocessors. 1996.

[101] Nathan Silver. *Ensemble methods in computational protein and ligand design: Applications to the Fc[gamma] immunoglobulin, HIV-1 protease, and ketol-acid reductoisomerase system.* PhD thesis, Massachusetts Institute of Technology, 2011.

[102] Sonya Tadrowski, Marcelo M. Pedroso, Volker Sieber, James A. Larrabee, Luke W. Guddat, and Gerhard Schenk. Metal Ions Play an Essential Catalytic Role in the Mechanism of Ketol-Acid Reductoisomerase. *Chemistry - A European Journal*, 22(22):7427–7436, May 2016.

[103] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv:1905.11946 [cs, stat]*, November 2019. arXiv: 1905.11946.

[104] The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, April 1998.

[105] The MIT Press. *Knowledge Futures Group.*

[106] Robert Tibshirani. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, January 1996.

[107] UC and Elsevier - Office of Scholarly Communication. `https://osc.universityofcalifornia.edu/open-access-at-uc/publisher-negotiations/uc-and-elsevier/`. Accessed: 2020-01-02.

[108] Underlay. `https://underlay.mit.edu`. Accessed: 2020-01-02.

[109] University of California and Elsevier Locked in Negotiations. `https://www.the-scientist.com/news-opinion/university-of-california-and-elsevier-locked-in-negotiations-65209`. Accessed: 2020-01-02.

[110] Titus S. van Erp, Mahmoud Moqadam, Enrico Riccardi, and Anders Lervik. Analyzing Complex Reaction Mechanisms Using Path Sampling. *Journal of Chemical Theory and Computation*, 12(11):5398–5410, November 2016.

[111] Titus S. van Erp, Daniele Moroni, and Peter G. Bolhuis. A novel path sampling method for the calculation of rate constants. *The Journal of Chemical Physics*, 118(17):7762–7774, May 2003.

[112] Richard Van Noorden. Publishers withdraw more than 120 gibberish papers. *Nature*, 2014.

[113] Alexandre Vidmer and Matúš Medo. *The essential role of time in network-based recommendation*, volume 116. 2016.

[114] Dashun Wang, Chaoming Song, and Albert-László Barabási. *Quantifying Long-Term Scientific Impact*, volume 342. 2013.

[115] Arieh Warshel. Electrostatic Origin of the Catalytic Power of Enzymes and the Role of Preorganized Active Sites. *Journal of Biological Chemistry*, 273(42):27035–27038, October 1998.

[116] Luca Weihs and Oren Etzioni. *Learning to Predict Citation-Based Impact Measures.* 2017.

[117] Allen W Wilhite and Eric A Fong. Coercive Citation in Academic Publishing. *Science*, 335(6068):542–543, February 2012.

[118] Kate Silverman Wilson, MIT Press, and Mit Media Lab. *MIT Press, Media Lab launch Knowledge Futures Group.*

[119] Lingfei Wu, Dashun Wang, and James A Evans. Large teams develop and small teams disrupt science and technology. *Nature*, 566(7744):378–382, February 2019.

[120] Jingfeng Xia, Jennifer L Harmon, Kevin G Connolly, Ryan M Donnelly, Mary R Anderson, and Heather A Howard. Who publishes in "predatory" journals? *Journal of the Association for Information Science and Technology*, 66(7):1406–1417, 2014.

[121] Shuqi Xu, Manuel Sebastian Mariani, Linyuan Lü, and Matúš Medo. *Unbiased evaluation of ranking metrics reveals consistent performance in science and technology citation data*, volume 14. 2020.

[122] Christopher M. Yeomans, Robin K. Shail, Stephen Grebby, Vesa Nykänen, Maarit Middleton, and Paul A.J. Lusty. A machine learning approach to tungsten prospectivity modelling using knowledge-driven feature extraction and model confidence. *Geoscience Frontiers*, June 2020.

[123] Jia You. Artificial intelligence. DARPA sets out to automate research. *Science*, 347(6221):465, January 2015.

[124] Jun Zhang, Zhen Zhang, Yi Isaac Yang, Sirui Liu, Lijiang Yang, and Yi Qin Gao. Rich Dynamics Underlying Solution Reactions Revealed by Sampling and Data Mining of Reactive Trajectories. *ACS Central Science*, 3(5):407–414, May 2017.

[125] Xing-Zhou Zhang, Jing-Jie Liu, and Zhi-Wei Xu. *Tencent and Facebook Data Validate Metcalfe's Law*, volume 30. 2015.

[126] Ioanna Zoi, Javier Suarez, Dimitri Antoniou, Scott A. Cameron, Vern L. Schramm, and Steven D. Schwartz. Modulating Enzyme Catalysis through Mutations Designed to Alter Rapid Protein Dynamics. *Journal of the American Chemical Society*, 138(10):3403–3409, March 2016.