Deep Hidden Physics Modeling of Cell Signaling Networks

Martin Seeger¹,², James Longden², Edda Klipp¹,² and Rune Linding¹,²,*

¹Humboldt-Universität zu Berlin, Theoretical Biophysics, Invalidenstr. 42, 10115 Berlin, Germany; ²Rewire Tx, Humboldt-Universität zu Berlin, Invalidenstr. 42, 10115 Berlin, Germany

Received: October 19, 2020
Revised: March 23, 2021
Accepted: April 07, 2021

DOI: 10.2174/1389202922666210614131236

Abstract: According to the WHO, cancer is the second most common cause of death worldwide. The social and economic damage caused by cancer is high and rising. In Europe, the annual direct medical expenses alone amount to more than €129 billion. This results in an urgent need for new and sustainable therapeutics, which has currently not been met by the pharmaceutical industry; only 3.4% of cancer drugs entering Phase 1 clinical trials get to market. Phosphorylation sites are parts of the core machinery of kinase signaling networks, which are known to be dysfunctional in all types of cancer. Indeed, kinases are the second most common drug target yet. However, these inhibitors block all functions of a protein, and they commonly lead to the development of resistance and increased toxicity. To facilitate global and mechanistic modeling of cancer and clinically relevant cell signaling networks, the community will have to develop sophisticated data-driven deep-learning and mechanistic computational models that generate in silico probabilistic predictions of molecular signaling network rearrangements causally implicated in cancer.

1. INTRODUCTION

The protein kinases are a family of enzymes that transfer phosphate groups to proteins [1-3]. This phosphate can then be recognized and bound by another protein or removed by a protein phosphatase [4]. Together these three simple processes represent one of the most important regulatory mechanisms of eukaryotic cells, directing nearly every aspect of cell biology. They are arranged in complex, kinase/substrate signaling networks (in which kinase substrates can themselves be kinases) that are used by cells, tissues, and multi-cellular organisms to interpret and respond to environmental cues and challenges. The consequences of breakdowns in these signaling networks underlie all cancers, diabetes, and disorders of the immune, neurological and cardiovascular systems in humans, and also some rare diseases.

The causality of cell signaling networks in cancer is prominent across multiple hallmarks of the disease [5]. For example, it is already well known that specific phosphorylation sites can have specific cellular functions; the well-characterized extracellular signal-regulated kinases 1/2 (ERK) are activated by dual phosphorylation of specific residues on the kinase. ERK is initially phosphorylated by the MAPK/ERK kinase (MEK) on the Thr-202/Glu-204/Tyr-204 (TEY) motif [6], with subsequent phosphorylation on the Ser-244/Pro-245/Ser-246 (SPS) nuclear translocation sequence (NTS) mainly by casein kinase 2 (CK2) to generate pSPS-pERK [7]. These events lead to the nuclear translocation of ERK and are essential for cell cycle progression [8]. Another example is the inhibition of cyclin-dependent kinase 1 (CDK1) activity by Tyr-15 phosphorylation, which directly regulates entry into mitosis and is an important element in the control of the unperturbed cell cycle [9].

Kinases are the second most common drug target [10]. Typically, kinase inhibitors are designed to target one specific kinase, completely perturbing its activity. This strategy has resulted in the approval of 52 kinase inhibitors (as of February 2020) overwhelmingly for the treatment of cancer [11]. However, despite this success, issues remain, such as kinase inhibitors are often toxic, and many cancers develop resistance to kinase inhibitors, either mutating in such a way that the inhibitor can no longer bind or bypassing the inhibited kinase by rewiring the signaling network. A first step to understand and, later on, to work around this mechanism by targeted therapeutics consists of combining machine learning-based algorithms with data from proteomic, genomic, and phenotypic screens in order to understand the signaling network in its entirety.

This challenge rests at the border between recent developments in deep machine learning and traditional computational modeling based on ordinary differential equations (ODE), Boolean formalism, or comparable approaches that assume an underlying molecular interaction network.

*Address correspondence to this author at the Humboldt-Universität zu Berlin, Theoretical Biophysics, Invalidenstr. 42, 10115 Berlin, Germany; E-mail: rune.linding@hu-berlin.de
Currently, ODE models for biological processes and especially signaling networks have typically formulated ad hoc, based on the biological understanding of the instance, specifically created data, and the aim of the modeling exercise. Then, data are used to estimate the model parameters [12, 13]. In the best case, an ensemble of plausible models is formulated and data are also used to evaluate which model out of the whole ensemble explains the data best [14]. Given that our biological knowledge is still limited, this traditional approach also has its limitations. Thus, theoretically speaking, it should be possible to use the (big amount of) data also to teach the model itself and, thereby, overcoming the bias/limitation in our current knowledge.

It is evident that such efforts would be timely and of great importance in enabling mechanistic models to be derived from complex data sets. This, in turn, will enable us to challenge our concepts about the dynamics of signaling networks relevant for cancer development and their connection to various cellular properties including gene expression and e.g. cell morphology.

Cellular signaling pathways have attracted quite some attention in order to understand the wiring of the network, conceptualize the modes of signal transmission, and integrate and analyze different types of data because of their importance in cancer development. Few examples of network models are EGFR, WNT, TGFbeta, and/or MAP kinase signaling networks [15-18]. Other models consider the interaction of signaling with metabolism [19]. Despite the vast amount of data, networks are mainly reconstructed manually, while data are used to estimate parameters and to decide which model to favor.

Recently, there have also been attempts to understand how the cell shape influences the way cells perceive and transmit the signal [20, 21]. Furthermore, Heinonen et al. published a landmark paper [22] in applied mathematics and machine learning (Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018, https://arxiv.org/abs/1803.04303) in which they presented a novel paradigm for non-parametric ODE modeling.

In recent years, we have focused our research on big data network biology, exploring biological systems by developing and deploying algorithms aimed to predict cell behavior, particularly looking at cellular signal processing and decision making [23-26]. Our laboratories have developed computational tools and deployed these on genome-scale quantitative data obtained by, for example, mass spectrometry, genomics, and phenotypic screens to understand the principles of how spatial and temporal assembly of mammalian signaling networks transmit and process information at a systems level in order to alter cell behavior.

Our overarching research aims to advance network medicine by identifying and targeting signaling networks associated with complex diseases. To this end, we have developed algorithms such as KINomeXplorer and NetworkKIN that model kinase-substrate interactions from the linear motif (classified by NetPhorest), a sequence of amino acids on a substrate that is recognized by a specific kinase [27]. KIN-spect predicts the impact on mutations in those residues in a kinase catalytic domain that are responsible for its specificity and activity [28]. ReKINect predicts the effect of non-synonymous point mutations on kinase signaling networks, which can lead to rewiring of the network, constitutive activation, or inactivation of the loss or generation of a phosphorylation site [29].

In parallel, the group based in Berlin has created a series of mathematical models and modeling approaches with relevance to signaling in eukaryotic cells and with relevance to various processes in cancer cells. The mathematical methods include ordinary and partial differential equations, Boolean and Bayesian networks, agent-based models, deterministic and stochastic simulations as well as various aspects of dynamic systems theory. Efforts to develop computational tools for the processing of such models, including parameter estimation, sensitivity analysis, and semantic assignments, have proven very useful in developing mechanistic and stochastic models of signaling networks [30]; this includes reaction-contingency based bipartite Boolean modelling, an approach for qualitative modeling of signaling networks taking into account their precise chemical modification, e.g. which residue has to be phosphorylated to change activity. This approach can then be extended to a quantitative, stochastic model using stochastic simulation of Boolean models [31] by assigning probabilities to each event in the signaling cascade. A major research focus has previously been given to demonstrate the feasibility of large-scale reconstruction of regulatory networks for compiling large numbers of compounds and reactions into an executable network allowing for Boolean simulations [32], that stochastic modeling of signaling networks can be used to calculate information transfer [33] and that kinetic parameters can be determined using Bayesian statistics [34, 35].

To facilitate the creation of predictive models algorithmically derived from comprehensive data sets, we propose that it is vital that the community adopts SOPs (Standard Operating Procedures) and best practices for the basic quality control of the deployed machine learning (these are a generalization of the principles first published in another study [36]):

1. **Data Organization**: It is evident that organizing big data according to a graph (e.g. a phylogenetic tree or network) can be of great help for the subsequent handling of data in a machine learning pipeline.

2. **Data Compilation**: Extracting from the data graph positive and negative examples for the different objects and classes that are to be ‘perceived’ by the machine learning algorithms can easily be automated because of the graph.

3. **Redundancy Reduction**: A critical step, which is often forgotten, is to eliminate examples from training data that are highly similar. In sequence analysis, this can be done based on similarity but across domains, different metrics can be used. It is vital in ensuring generalization of the algorithm but also to get realistic performance metrics.

4. **Data Partitioning**: Following redundancy reduction, the data can be automatically split into parts that are used for training, test, and validation of the machine learning model.
5. Selecting a Nonredundant set of Classifiers/Models: The selection of the best classifiers/models should also be automated and transparent. Often, it is desirable to have general/broad classifiers but also not too diverse sets of classifiers as this may render it more difficult to extract detailed information.

6. Benchmarking: Validation and benchmarking are the most critical step of any deployment of machine learning and modeling in biology. It must be conducted with the utmost rigor. Historically, most bioinformatics predictors are overestimated in performance due to lack of redundancy reduction (this is also a major reason for the rejection of publications). However, the principles of benchmarking are important to consider. Using a standard metric like, for binary classification problems, AUC (Area under the receiver-operating curve) with an independent test/validation data set is imperative, but a thorough discussion of systematic and statistical sources of error is also important.

7. Significance Testing: What is less common but just as important is that a classifier should be proven to be better in performance than other classifiers, while at the same time minimizing bias and variance in a controlled manner. This can be achieved by computing p-values (e.g. resampling of scores of positive and negative samples to construct a bootstrapped AUC distribution) or conducting more complex, e.g. Bayesian analysis. Of course, a goal is still to optimize the performance for classifiers, but given the steps taken to balance and render the data more realistic (redundancy), it is possible to expect that the models have been challenged enough to make this a safe and sound selection process.

8. Dimensionality Reduction: For years, most efforts in reducing the dimensionality of big data biology relied on PCA (Principal Component Analysis). However, this is problematic because this method transmits errors and can even boost these in combination with certain regression type models more importantly it is a linear method, and biology and underlying casualties and mechanisms are highly non-linear. We propose that moving forward the community deploys machine learning (deep learning in particular) itself to reduce dimensionality, this can be done by automating the funneling of data spaces through a lower number of dimensions at the deepest levels of a model. For example, deep autoencoders are highly nonlinear, and powerful generalizations of PCA, and t-SNE, UMAP, and related techniques have proven especially helpful in particular in the visualization of high-dimensional datasets.

9. Integration of Heterogeneous Models: Previously, we have shown how machine learning models (e.g. ANNs) can be integrated with other models (e.g. PSSMs) by designing a scoring scheme in which the primary scores from classifiers are calibrated by benchmarking the models to a common reference data set. By converting the benchmarking scores into probabilistic scores, it is possible to directly compare classifiers even from very distinct regimes of machine learning algorithms. In recent years, it has become more streamlined using the tools in Google’s TensorFlow and alike.

10. Gaining Insight: A leading principle in any life-science project is to gain insight and understand new biology. This is often forgotten in machine learning projects, however, if the new algorithm/classifier/model cannot project or make predictions for globally descriptive data (e.g. cell behavior) then a model does not have much real value. The take-home message is that fully utilizing machine/deep learning tools in biology require deep biological understanding and the ability to challenge such models with very specific questions/knowledge/unsolved problems which only can happen if the people involved are highly trained in biology/bio-chemistry/genetics and, thus, such projects should always be conducted in an interdisciplinary environment.

CONCLUSION

Moving forward, our work using machine learning and mechanistic modelling aims to enable further developments through the collaboration of platforms to cover and target other cancers and other complex or orphan diseases. We find it plausible that the tools we aim to develop will enable both basic research in cell biology as well as industrial drug discovery research and clinical research. This is indeed a great motivation to undertake such development work. In addition, we will also continue our efforts to establish multi-scale models of cell signaling networks involved in cancer metastasis, which is the primary (>90%) cause for cancer deaths.

CONSENT FOR PUBLICATION

Not applicable.

FUNDING

None.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.
[27] Horn, H.; Schoof, E.M.; Kim, J.; Robin, X.; Miller, M.L.; Diella, F.; Palma, A.; Cesareni, G.; Jensen, L.J.; Linding, R. KinomeXplorer: An integrated platform for kinome biology studies. *Nat. Methods*, **2014**, *11*(6), 603-604. http://dx.doi.org/10.1038/nmeth.2968 PMID: 24874572

[28] Creixell, P.; Palmeri, A.; Miller, C.J.; Lou, H.J.; Santini, C.C.; Nielsen, M.; Turk, B.E.; Linding, R. Unmasking determinants of specificity in the human kinome. *Cell*, **2015**, *163*(1), 187-201. http://dx.doi.org/10.1016/j.cell.2015.08.057 PMID: 26388442

[29] Creixell, P.; Schoof, E.M.; Simpson, C.D.; Longden, J.; Miller, C.J.; Lou, H.J.; Perryman, L.; Cox, T.R.; Zivanovic, N.; Palmeri, A.; Wesolowska-Andersen, A.; Helmer-Citterich, M.; Ferkinghoff-Borg, J.; Iamochi, H.; Bodenmiller, B.; Erler, J.T.; Turk, B.E.; Linding, R. Kinome-wide decoding of network-attacking mutations rewiring cancer signaling. *Cell*, **2015**, *163*(1), 202-217. http://dx.doi.org/10.1016/j.cell.2015.08.056 PMID: 26388441

[30] Flöttmann, M.; Scharp, T.; Klipp, E. A stochastic model of epigenetic dynamics in somatic cell reprogramming. *Front. Physiol.*, **2012**, *3*, 216. http://dx.doi.org/10.3389/fphys.2012.00216 PMID: 22754535

[31] Mori, T.; Flöttmann, M.; Knartz, M.; Akutsu, T.; Klipp, E. Stochastic simulation of Boolean reaction models: Towards quantitative analysis of large signaling networks. *BMC Syst. Biol.*, **2015**, *9*, 45. http://dx.doi.org/10.1186/s12918-015-0193-8 PMID: 26259567

[32] Münzner, U.; Klipp, E.; Knartz, M. A comprehensive, mechanistically detailed, and executable model of the cell division cycle in *Saccharomyces cerevisiae*. *Nat. Commun.*, **2019**, *10*(1), 1308. http://dx.doi.org/10.1038/s41467-019-08903-w PMID: 30899900

[33] Stojanovski, K.; Ferrar, T.; Benisty, H.; Uschner, F.; Delgado, J.; Jimenez, J.; Solé, C.; de Nadal, E.; Klipp, E.; Posas, F.; Serrano, L.; Kiel, C. Interaction dynamics determine signaling and output pathway responses. *Cell Rep.*, **2017**, *19*(1), 136-149. http://dx.doi.org/10.1016/j.celrep.2017.03.029 PMID: 28380353

[34] Liebermeister, W.; Klipp, E. Bringing metabolic networks to life: Convenience rate law and thermodynamic constraints. *Theor. Biol. Med. Model.*, **2006**, *3*, 41. http://dx.doi.org/10.1186/1742-4682-3-41 PMID: 17173669

[35] Lubitz, T.; Schulz, M.; Klipp, E.; Liebermeister, W. Parameter balancing in kinetic models of cell metabolism. *J. Phys. Chem. B*, **2010**, *114*(49), 16298-16303. http://dx.doi.org/10.1021/jp108764b PMID: 21038890

[36] Miller, M.L.; Jensen, L.J.; Diella, F.; Jorgensen, C.; Tinti, M.; Li, L.; Hsiung, M.; Parker, S.A.; Bordeaux, J.; Sicheritz-Ponten, T.; Olhovsky, M.; Pascale, A.; Alexander, J.; Knapp, S.; Blom, N.; Bork, P.; Li, S.; Cesareni, G.; Pawson, T.; Turk, B.E.; Yaffe, M.B.; Brunak, S.; Linding, R. Linear motif atlas for phosphorylation-dependent signaling. *Sci. Signal.*, **2008**, *1*(35), ra2. http://dx.doi.org/10.1126/scisignal.1159433 PMID: 18769331