Artificial Intelligence & Computational Biochemistry: New Frontiers in Drug Discovery

May 2017

nest.bio
RESEARCH

Nest.Bio Ventures is a venture creation and venture capital firm leveraging technological advancements to create next-generation therapeutics. By combining deep scientific expertise, a data-driven approach, and broad, international academic and industrial networks, we help transform scientific breakthroughs into revolutionary companies.

Artificial Intelligence & Computational Biochemistry: New Frontiers in Drug Discovery

May 2017

AUTHORS

Ashvin Bashyam, Jaideep S. Dudani, Karthik Murugadoss, Varesh Prasad

EDITORS

James W. Weis, Xi Chen, Cheryl Cui

Artificial Intelligence & Computational Biochemistry: New Frontiers in Drug Discovery

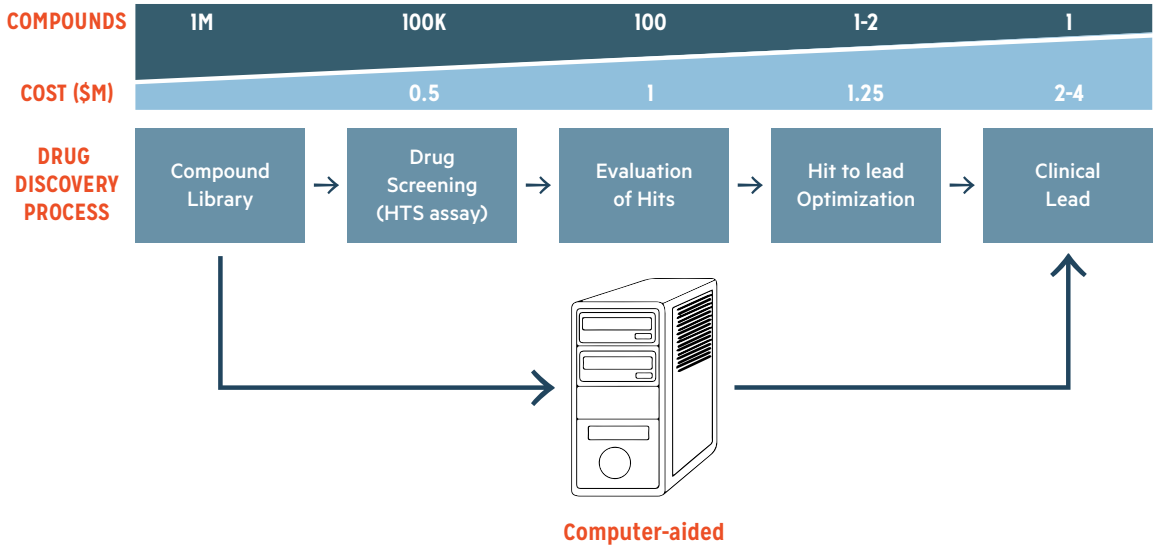
SUMMARY

Drug discovery is an exceptionally challenging process, requiring millions of dollars over many years. However, rapidly emerging computational techniques from artificial intelligence and biochemistry are poised to transform every step of this process—potentially lowering costs, improving throughput, and ultimately bringing more medicines to market. This report describes the current computational drug discovery landscape and presents promising areas for future research, development, and commercialization.

HIGHLIGHTS

1. The unique convergence of drug discovery and computational power has enables several unique development strategies, IP considerations, and business models.
2. Over 50 biomedical technology startups with core computational expertise have raised first equity financing since January 2015. Participations from academic partners, strategic industry partners, traditional venture capital, and government research institutions indicates significant interest in the field.
3. Through an analysis of key opinion leaders in academia and industry, we identify exciting opportunities for innovation and development at the intersection of computation and drug development.
4. Strong integration of virtual screening and experimental drug discovery teams within a program creates the most value for the pharmaceutical industry. While capital intensive, this approach aligns incentives and brings best-in-class talent together with a shared goal.
5. Significant opportunity remains by understanding conformational dynamics and long-range interactions. Both technologies are poised to revolutionize computational drug discovery in larger, more complex biological drugs.

GRAPHICAL ABSTRACT



CONTENTS

CHAPTERS

FIGURES

TABLES

SUPPLEMENTARY TABLES

EXECUTIVE SUMMARY

BACKGROUND AND SCOPE

Successful discovery and translation of new drugs can transform patient lives, but the process is exceptionally challenging and costly. This is due to our limited ability to understand and control the highly complex nature of biological systems. Drug discovery typically suffers from high attrition rates, starting with large screening efforts and followed by several rounds of optimization to produce a clinical lead. This can take many years and cost millions of dollars. Thus, innovations that improve the drug discovery process fill a significant unmet need and could become immensely valuable. Rapidly emerging computational techniques are poised to transform every stage of drug discovery. These platforms can leverage the increasing availability of high-quality data, in conjunction with advances in computational power and sophistication, to perform complex tasks, such as virtual screening of drug libraries or identification of more potent lead compounds. This report provides an overview of the state of computational approaches to drug discovery, which may serve as a guide to students, entrepreneurs, and investors.

OVERVIEW OF REPORT

We examined the drug discovery pipeline to analyze the existing and potential impact of computational and machine learning methods at each stage for both small molecule and biologic therapeutics. Here we found that virtual screening of small molecule libraries is the most common application of computational methods, but there are many opportunities for innovation spanning the entire drug development pipeline and across other therapeutic modalities.

Next, we identified the strategies employed in the field of virtual drug discovery and provided an array of business models and partnerships—ranging from discovering and commercializing potential therapies entirely in-house, to collaborating with more experienced organizations. Outsourcing to contract research organizations at various stages in the process offers a promising alternative. On the other hand, a more technology-focused company could produce software for use by independent drug discovery teams. Each model has its own potential benefits and drawbacks, with tradeoffs in their effectiveness, costs, incentives, and ownership of value. The various business strategies and approaches also result in unique considerations for intellectual property.

We explored the diverse challenges that virtual drug discovery companies focus on, such as structure-based drug design and molecular network-driven discovery. We categorized these companies based on their disease areas and the nature of their computational methods—computational biochemistry or data-driven machine learning. In general, we found that companies tend to focus on one, as opposed to developing hybrid approaches. We, additionally, explored the contributions of key opinion leaders that are shaping the field of computational drug discovery.

NEW OPPORTUNITIES

By establishing a framework to analyze the technical challenges, business models and opportunities, and competitive landscape, we identified gaps in the field and areas for innovation. Our analysis suggests that improving structure-based drug development may enable de novo ligand identification, which is powerful for targets with no known ligands. New advances in statistical machine learning and the availability of large datasets make structure-based approaches feasible.

Despite the significant range of business models, companies with a strong integration of their virtual screening and experimental drug discovery teams have created the most value for the pharmaceutical industry. We believe that,

while these are the most capital intensive, the focus and alignment of incentives of this model allows companies to attract best-in-class talent and advance development pipelines efficiently.

A critical bottleneck slowing the creation of new virtual screening methods is the rate of prospective validation of results. The current paradigm of housing virtual screening and experimental validation separately significantly cuts the bandwidth and feedback response rate of experimental validation. An organization that places high value on reducing the barriers to experimental validation, potentially through technological solutions, would quickly outpace others in advancing their internal software development programs.

Finally, the majority of the field is exploring computational approaches for identifying new small molecules. The past few decades have seen the advent of new classes of therapeutic modalities, such as proteins and nucleic acids, each of which stands to benefit from computational drug discovery and design. While unique challenges exist for each, the general problem is a significant increase in dimensionality when moving from small molecule to biological compounds. We anticipate that technological advances, both in software and hardware, will begin to surmount these barriers.

INTRODUCTION

Successful development and clinical translation of therapeutics can be a richly rewarding endeavor, but the process is filled with tremendous risk, uncertainty, and capital requirements. Difficulties in making early-stage predictions about which compounds will be successful contribute to high attrition rates in drug discovery programs. Even where biological targets have been successfully identified and validated, the subsequent development of small molecule and biologic therapeutics is fraught with challenges. Accurate and comprehensive predictive models would substantially alleviate these problems, but biological systems are complex and highly networked, making modeling a historically challenging endeavor, and experimentation the main—and extremely costly—vehicle for discovering and developing drugs. Recent years, however, have seen ongoing exponential growth of computational power enable application of new computational methods to drug discovery.

The drug discovery process typically follows a well-defined path **(Figure 1A)**. The first step in drug discovery for small molecules is typically performed using high-throughput screening (HTS) assays, which enable the evaluation of hundreds of thousands to millions of compounds for activity against a particular target. Following identification of candidate hits, further evaluation (e.g., for bioactivity in cells) and optimization (e.g., by medicinal chemistry) occurs for properties such as pharmacokinetics and bioavailability. Given the required number of experimentally evaluated compounds, drug development pipelines often use $2-4M to generate a single clinical lead [1, 2]. For biologics, the process is even more challenging due to their increased complexity **(Figure 1B)**. Consequently, the cost for a clinical biologic lead is typically $5-10M. Clearly, the process of drug development is difficult and innovations that can dramatically improve this process will be immensely valuable.

Numerous inefficiencies and inadequacies account for these costs. For example, the breadth of potential targets and off-targets that need to be evaluated against libraries with million compounds far outweighs what is feasible [3]. Assays themselves are subject to design bias, leading to inefficiencies in experimental selection. Again, these issues are further exacerbated for biologics and their poorly understood design rules. Several innovations are being explored to address

Figure 1A

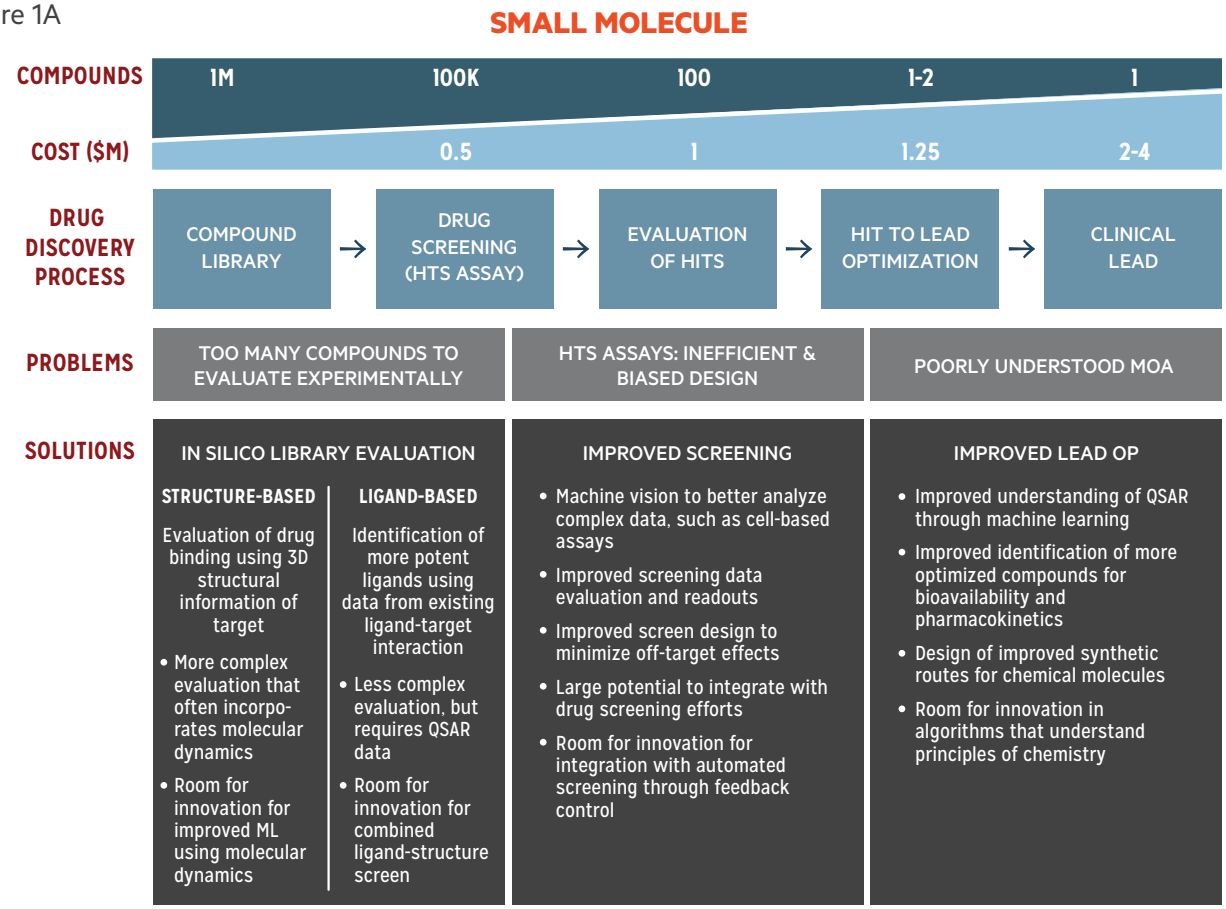


Figure 1B

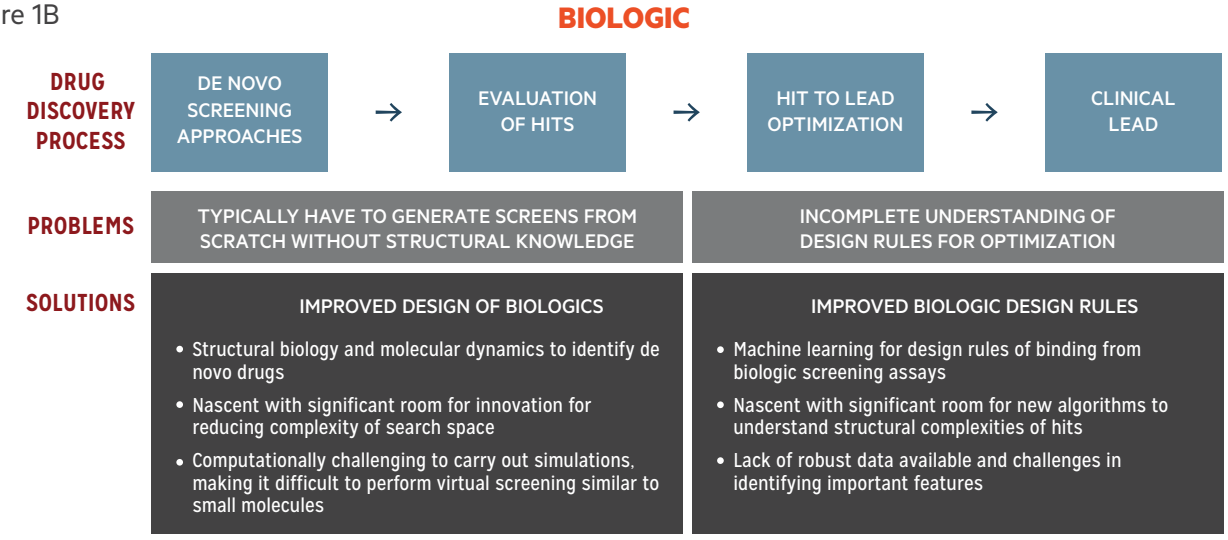


Figure 1. Current drug discovery and development processes, challenges, potential solutions. (a) Small molecule drugs are typically tested in high-throughput fashion and subsequently developed into a hit compound. The process has numerous challenges that can be addressed through innovations in virtual drug discovery and machine learning. (b) Biologics follow a similar approach but are significantly more complex molecules. The available data for these drugs is limited, but computational approaches may provide both improved design of biologics as well as define design rules for such drugs.

these challenges. These include more advanced robotics to minimize operator error, microfluidic systems (such as organs-on-chip) to provide more robust and rapid testing platforms [4], and efforts to leverage the vast data currently amassed from past drug development campaigns to inform future development [5]. The first two approaches, while powerful, only improve upon existing capabilities, rather than shifting the paradigm away from the true source of complexity: experimental screening of each compound.

Rapidly emerging computational techniques, such as statistical machine learning, as well as advances in more traditional computational techniques, such as biochemical modeling, are demonstrating potential to transform each stage of drug development. In particular, these methods have shown promise in the early stages, such as in virtual compound screening. Virtual screening uses computational modeling and/or machine learning to estimate activity against a target of interest, thus selecting only a small subset of hits for experimental evaluation. This nascent field has yet to be truly defined, allowing significant opportunity for new innovation. Additionally, the low-capital investments necessary to build software centric platforms for drug discovery combined with increasing availability of high-quality data make it an exciting time for development of approaches that could become widely adopted.

We first review technological and business strategies being implemented and note intellectual property considerations. We further describe companies undertaking these approaches, their business models, and key opinion leaders in the field. The emphasis of this report is on virtual screening, as it appears to be the part of the process most impacted by computational methods at this time. However, this overall framework can be applied to any stage of drug development, and we will reference other stages where relevant. This report should serve as a baseline upon which to build further analysis of potential opportunities and areas to be explored, such as: new approaches for structure-based (rather than ligand-based) drug discovery, models of innovation for integrated virtual screening and drug development teams, innovation that enables rapid iteration of virtual screening alongside experimentation, and further extension of these approaches to biologics. While we provide preliminary thoughts for additional exploration, there is a need to expand upon these frameworks to fully identify gaps. Lastly, we hope that this report and associated documents serve as an open-access resource for interested parties to rapidly learn about the state of this field for their own explorations.

BROAD DEFINITIONS AND APPROACHES

In **Table 1**, we provide some key terminology used throughout this report. Computational drug development can be broadly categorized by which aspect of the drug discovery pipeline addressed. The first can be viewed as in silico screening, where compounds are evaluated computationally for their potential affinity for a target. This can be applied to both small molecules and biologics, though the computation is more challenging for biologics and consequently poorly explored. There are two primary modes of in silico screening of virtual libraries, structure-based and ligand-based [5]. Structure-based screening requires knowledge of the 3D structure of the target of interest onto which potential compounds are mapped. This is more computationally intensive relative to ligand-based screening and thus far has produced poorer results, perhaps due to incomplete molecular simulations of ligand-target interactions. Further development may enable robust structure-based screening with molecular dynamic simulations to dramatically

TERM	DEFINITION
Drug discovery (DD)	The process for identifying and developing candidate compounds into clinical leads
High-throughput screening (HTS)	A drug discovery process typically leveraging automation to assay biological and biochemical activity of numerous compounds simultaneously
Lead optimization	Diverse chemical and biological processes to optimize lead compounds for improved activity, solubility, pharmacokinetics and pharmacodynamics
Mechanism of action (MOA)	Biochemical pathway through which a drug elicits pharmacological effect
Virtual screening (VS)	Broad term referencing computational techniques in drug discovery to identify drugs against a target of interest; analogous to performing HTS in silico
Machine learning (ML)	A general type of artificial intelligence algorithm that learns from existing data to predict relationships among unknown or unmeasured quantities without having those relationships explicitly programmed
Molecular dynamics (MD)	Computational simulations of the physical movements and interactions of atoms and molecules
Structure-activity relationship (SAR)	Defines the association between the chemical or three dimensional structure of a molecule and its biological activity with the aim of identifying the chemical group responsible for a particular biological effect
Quantitative structure-activity relationship (QSAR)	Subset of structure-activity relationship that identifies mathematical relationships between chemical moiety or structure and biological activity
Absorption, distribution, metabolism, and excretion (ADME)	Pharmacokinetics term for describing the deposition of a pharmaceutical compound within an organism
In silico library	A range of compounds and drugs that can evaluated using virtual screening
Contract Research Organization (CRO)	An organization that provides support to life sciences-based industries through outsourced research services
Structure-based screening	Based on mapping ligands onto the 3D structure of a target
Ligand-based screening	Based on QSAR of ligand-target interaction to identify more effective, safer drugs
Biochemistry-driven computational drug development	Methods that simulate biochemical and physical interactions between ligand and target pair
Data-driven computational drug development	Methods that identify predict drugs through evaluation of experimental data

Table 1. Key terminology.

improve the success of virtual drug discovery. Ligand-based screening requires quantitative structure activity relationship (QSAR) data of ligand-target pairs to evaluate a large number of compounds based on their ability to affect the target by a similar mechanism.

Computational approaches are also being deployed to improve the process of experimental screening assays themselves [3]. These include machine vision approaches for analysis of results from cell-based assays, as well as machine learning approaches to define more optimal screening experiments. Further implementation of integrated systems of HTS and feedback control loops, where computationally-guided experiments are carried using industry-standard automation could be transformative in drug discovery assays.

Lastly, machine learning has the ability to transform the downstream analysis of drug discovery assays [6] and be applied prospectively for new drug targets. As the data from such experiments gets more complex and convoluted, machine learning algorithms can be used to identify meaningful latent relationships and design rules for future implementation. This may be exceptionally useful in the case of biologics, where de novo screening and development is often extremely costly. Understanding design rules using data generated experimentally and evaluated computationally may enable more robust design of biologics.

VIRTUAL DRUG DISCOVERY STRATEGIES

Differences in company strategy will affect capital requirements and the amount of control and/or ownership maintained over the results of these programs **(Figure 2)**. When defining a strategy for pursuing a virtual DD program, the sources of molecular structure/affinity datasets, sources of virtual screening algorithms, and methods used to validate the results of screening must be considered.

STRUCTURE AND AFFINITY DATA

A virtual screening program requires well-annotated datasets to train and validate the models before it can perform new target/compound identification. For data intensive approaches, such as deep neural networks, the dataset often consists of structural and affinity data. For techniques that explicitly incorporate more biochemical relationships, such as SAR based techniques, these datasets may need to be expanded to include more context such as signaling and regulatory pathways, off-target biological effects, or chemical design rules for acceptable ADME. These datasets can be publicly available, generated in collaboration with contractors or partners, or generated internally.

Public datasets, both free and paid, can lay the foundations for an algorithm. While relatively low in cost and requiring little release of ownership, these datasets have significant limitations in quality of documentation and breadth of coverage. Many computational techniques cannot generalize beyond the particular problem domain for which experimental data was originally generated [7]. Aggregating datasets from multiple sources can further exacerbate problems in consistency between experimental conditions and their documentation [8,9]. Some techniques, however, have demonstrated improved classification accuracy when trained across diverse datasets opening up a scalable approach to increasing the effectiveness of virtual screening [10]. Gaining a competitive advantage for a development program may be difficult by relying solely on these public datasets.

As an alternative, working with external organizations, such as CROs or collaborators, offers the ability to produce novel datasets

tailored to the virtual screening algorithm helping to confer a competitive advantage. In general, working with a CRO will be more capital intensive than partnering with a collaborator, but will tend to require giving up less ownership and control of the program. These collaborators can include academic institutions and industrial partners in biopharma.

Lastly, instead of sourcing data externally, generating it internally offers the most flexibility and control. Resources can best be focused on libraries and targets relevant to the development program. This focus can be valuable for methods that generalize poorly across different problem domains. This strategy often requires the most significant capital expenditures as new capabilities must be established.

VIRTUAL SCREENING ALGORITHMS

Virtual screening programs also require software to synthesize these datasets and predict the behavior of new molecules in new

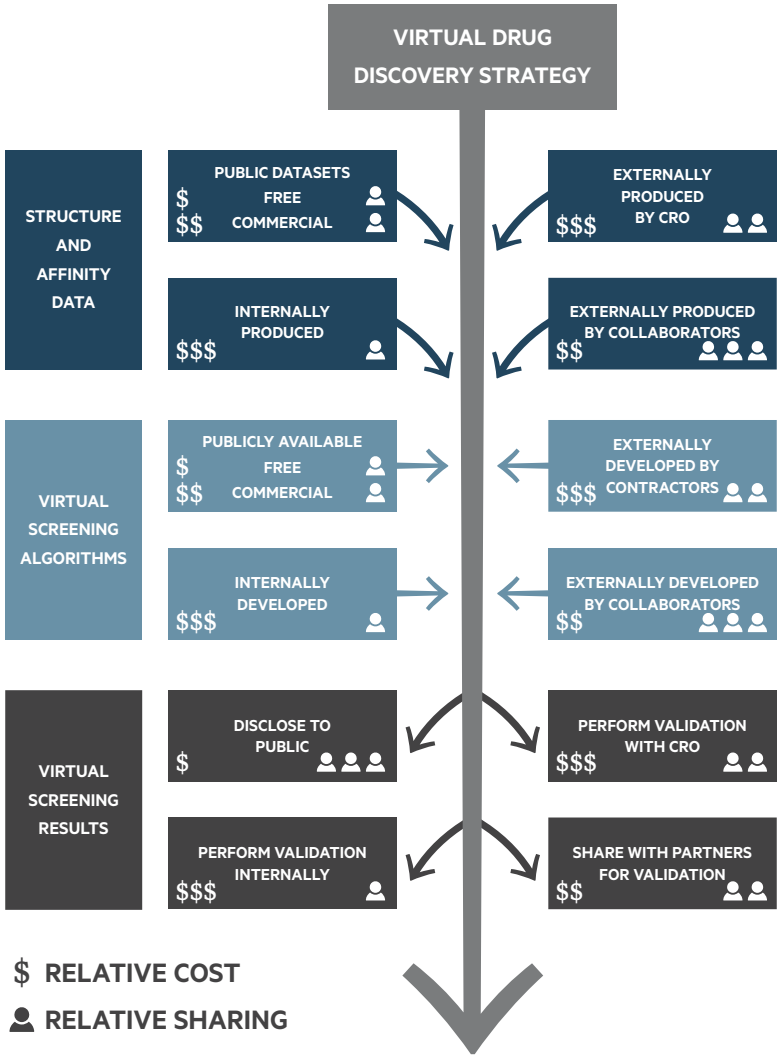


Figure 2. Strategic decision making in virtual drug discovery. Strategic decisions must be made around three distinct stages of a virtual DD program: sourcing structure and affinity data, sourcing and/or developing VS algorithms, and advancing VS results towards experimental validation. These decisions can have material impact on a business as each requires tradeoffs to be made between capital efficiency, development time, and the development of internal expertise.

environments. Once again, while these software tools can be obtained from publicly accessible sources (freely or commercially available) for low cost and minimal loss of ownership, it may be difficult to build a competitive program while relying solely on generic software. While this may not work for a program focused around virtual screening, publicly available tools will play a large role in augmenting traditional screening efforts.

Working with a partner, either a non-stakeholder with expertise in computational approaches who can develop a specialized solution or a close collaborator, is likely to improve the results of the program. Working with a stakeholder partner helps conserve capital compared to a contractor, but at the expense of ownership or control.

Internal development of VS algorithms enables fine-grained design of the tools for the specific DD problem. While best for retaining control, this may involve a significant investment of both time and capital to realize these advantages. This approach is best suited for focused efforts in particular around particular molecular classes, therapeutic areas, or when leveraging a particular algorithmic insight.

VIRTUAL SCREENING RESULTS

Finally, if the value of a VS-DD program is to be realized, the results must be validated using experimental methods before advancing to clinical development. If the goal is to build knowledge and advance the field as quickly as possible, without regard for ownership, then public disclosure of the results can allow a program to focus on building their screening platform and generating results, while encouraging others to continue with experimental validation. Unfortunately, unless proper control over IP is maintained, the commercial value of new discoveries can easily be eroded through public disclosure. For newer, less established teams seeking to demonstrate a capability, this approach can quickly increase visibility and help form partnership for future development efforts.

In order to retain more control, working with an external organization, such as a CRO or a stakeholder partner, can provide moderate capital efficiency while still advancing a program towards the clinic by validating the computational results. This approach is very common in the virtual DD industry and, as such, provides a straightforward route for experimental validation with significant capital expenditure into internal facilities.

The most capital and time intensive approach, internal validation, allows complete retention of ownership, but the risks related to raising capital and building expertise in-house must be weighed against pursuing a lower cost, more flexible option. This makes sense as a large organization leveraging its internal capabilities to retain control over a development program, but otherwise is an inefficient use of time and money.

Another important consideration is that, since raising large amounts of capital often requires equity investors, highly resource-intensive efforts, such as internal development programs, can indirectly lead to some loss of control and ownership. Considerations between exchanging ownership at the organizational level (equity investors) versus at the development program level (CROs, stakeholder partners) may influence which strategy is most suitable.

CASE STUDY

Nimbus Therapeutics provides an insightful case study into strategically building an integrated company leveraging advances in virtual screening with experimental validation [11]. Nimbus works with external organizations, both CROs and collaborators, to provide validation on which to train and validate their models. By providing cash and equity incentives to partners, Nimbus can ensure

an alignment of motivations; their success in advancing candidates through the pipeline speaks to the strength of this approach [12]. Nimbus works with a stakeholder collaborator, Schrodinger, to provide both the software and some expertise to perform in silico screening. This enables capital efficiency while sharing in the rewards of their success. Finally, as these screening programs yield candidates, CROs and stakeholder partners both validate these compounds. Nimbus consistently demonstrates a willingness to work with external organizations to efficiently and rapidly advance programs into clinical development. While their shareholders do not retain complete ownership, much more value is ultimately created by leveraging best-in-class resources at each stage of the pipeline [13].

MODELS OF INTEGRATING VIRTUAL SCREENING & EXPERIMENTAL DRUG DISCOVERY PROGRAMS

The strategies described above naturally lend themselves to different business models and partnership structures for VS programs. These discovery programs first require software to build predictive models of ligand/receptor interactions based on structural and/or experimental information. These software tools are then used, often in combination with experimental methods, to identify targets and/or compounds. Finally, these targets/compounds must be validated through well-defined experimental methods before further clinical development can proceed. **Figure 3** presents an overview of possible relationships between the VS and experimental DD teams, housed either within the same organization or independent entities working together. By understanding where and how the two teams interface, we can analyze the relative merits of the various structures. The following business models differ significantly in their effectiveness, costs, incentives, and ownership over value-creation events.

BUSINESS MODEL 1: STRONG INTEGRATION OF INTERNAL DD AND VS TEAMS

When housed within the same organization, the integration of the DD and VS teams has the potential to unlock new scientific breakthroughs and realize sizable returns. The ability of these teams to specialize in targeting particular ligand/receptor classes, biological phenomena, and/or disease areas opens opportunity for high-efficiency discovery programs spanning multiple targets (e.g., integrin targeting: autoimmunity, fibrosis, vascular disorders and immuno-oncology at Morphic Therapeutic). The expertise that these teams build spans both experimental and computational approaches. These programs can be especially effective since all team members are incentivized to work towards the same goal.

This approach may not apply to all discovery programs, as some may not benefit from VS efforts. These may include investigations into broad disease areas, poorly understood structural biology, or targets with delivery challenges. Further, significant commitment of capital and human resources require that these programs offer strategic value to the organization.

Examples of companies: Morphic Therapeutic, Relay Therapeutics, Schrodinger, Berg Health.

BUSINESS MODEL 2: VS TEAM OUTSOURCES VALIDATION

While not as well integrated as the former business model, an organization that only outsources the execution of target/compound validation studies can still retain many of the former's most significant benefits. Most importantly, since expertise related to both computational and experimental approaches

is maintained within a single organization, the strengths and efficiencies gained from integration are still captured. The primary advantages of this approach are reduced costs and expedited timelines, since significant opportunity exists to partner with industry to perform late-stage validation and clinical development. It is still critical that this team understands the motivation and design of these validation studies, as they will play a key role in guiding further clinical development.

There are two main approaches to outsourcing experimental validation: working with a stakeholder or a non-stakeholder, which offer tradeoffs regarding control, quality of execution, and capital efficiency. Working with a non-stakeholder (e.g., contract research organizations) offers potentially complete retention of intellectual property rights and, oftentimes, the ability to execute quickly, but at significant expense. However, given the possible loss of control over experimental design, special care must be taken to ensure high-quality results. Alternatively, bringing in a stakeholder, such as a larger biopharmaceutical company, can offer significant benefits such as alignment of incentives, integration into an existing clinical development pipeline, and the expertise brought by a team dedicated to a particular treatment modality or disease area. Although some control and/or ownership may be ceded at this stage, the improved probability of success of the development program often justifies these partnerships.

Examples of companies: Nimbus Therapeutics, TwoXAR, Atomwise.

BUSINESS MODEL 3: STRONG COLLABORATION BETWEEN INDEPENDENT VS AND DD TEAMS

In this model, one team brings expertise in computational screening, while the other brings capabilities in experimental drug discovery. A strong collaboration, typically focused on a particular ligand/target class, biological phenomenon, or disease area, allows the development program to build unique expertise. While this proficiency may not run as deep as that of an integrated team, the flexibility it offers around deal structure and optionality to engage with multiple partners can be a significant advantage. A VS team with a broadly applicable platform may consider this structure to simultaneously pursue many development programs with minimal capital expenditure. With these deals, rapid feedback and shared value creation helps align development efforts and incentives.

Examples of companies: TwoXAR (partners: Stanford Medicine, Michigan State University, Mount Sinai), Atomwise (partners: Stanford University, Scripps Research Institute, University of Toronto, Merck, UC San Diego, Notable Labs), Cloud Pharmaceuticals (partners: Genomeon, University of Florida).

BUSINESS MODEL 4: WEAK COLLABORATION BETWEEN INDEPENDENT VS AND DD TEAMS

To contrast the strong collaboration model, a similar structure with weaker bonds between those on the computational and experimental sides allow for shared value creation during the target/compound identification and validation process while allowing for more flexibility on both sides. A less structured collaboration may result from a VS team pursuing a large number of programs with several partners. Alternatively, a deal structured to incentivize the process of virtual screening, as opposed to its results (identified targets/compounds), may create a misalignment of goals between the teams. A nebulous collaboration, slower feedback cycles between computational and experimental efforts, and inability to share in long-term value creation can prevent even

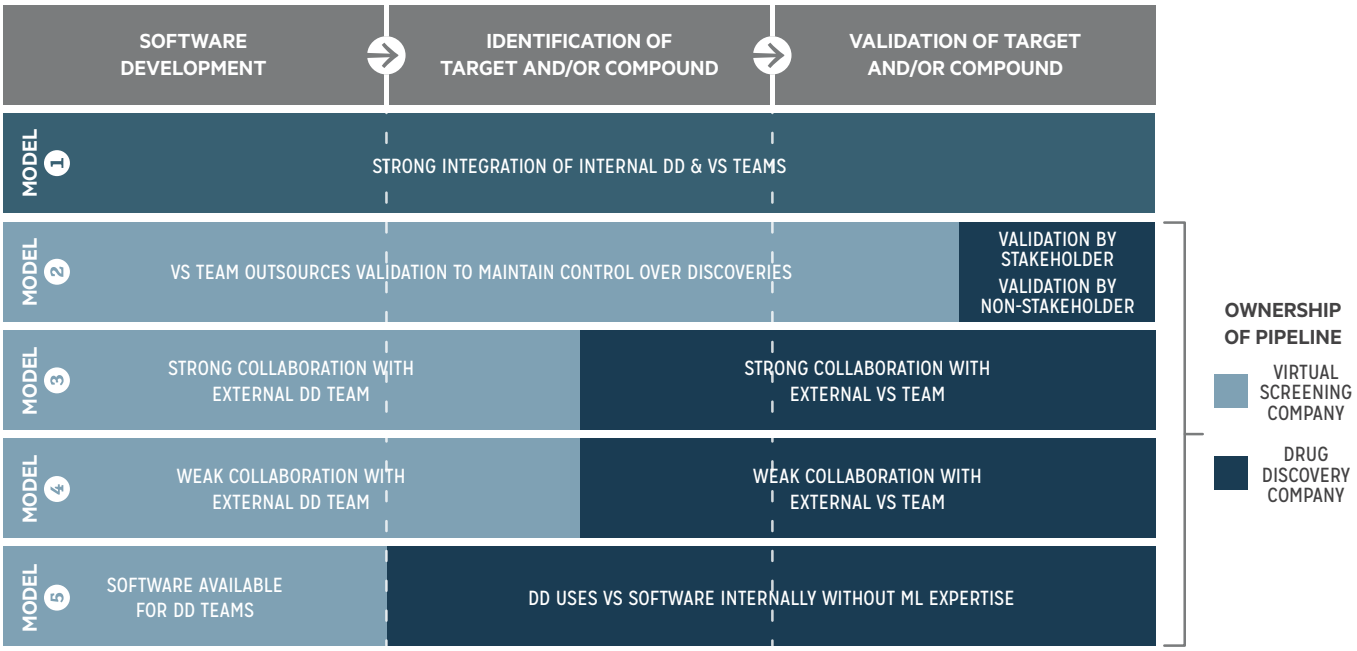


Figure 3. Overview of business models for virtual screening drug development programs. A VS drug development program involves three core functions: software development, identification of target and/or compound, and validation of target and/or compound. Many business models exist that allow these functions to be performed internally or outsourced to a partner. Strong integration of each function enables specialization and control of the entire pipeline though requires significant investments. In contrast, an organization may choose to focus on a core competency, such as VS tools made available to partners. This allows for capital efficiency at the expense of control.

a truly innovative computational technology from directly influencing a DD program in this model. This structure makes sense when a VS team is very interested in preserving optionality among partnerships, operating with a focus on technology development, or simply new to the field.

Another reason for limited collaboration between virtual screening and experimental teams would be if one partner were publishing or otherwise making their results non-proprietary. This may be beneficial for a newer team working to establish credibility; an academic group focused on achieving non-commercial goals (e.g., publications), or a team trying to gain visibility as a means of seeking partners for future development. Finally, a low commitment, less resource intensive collaboration allows for more independence for the virtual screening company, as it is easier to end a partnership because of reduced dependencies and invested resources.

Examples of companies: Google Research (partner: Stanford University), D.E. Shaw Research (partner: Relay Therapeutics).

BUSINESS MODEL 5: VS SOFTWARE AVAILABLE FOR DD TEAMS TO USE

Breaking from the previous models where a VS team is engaged in some target/compound identification and possibly validation, many companies choose to release their software for DD teams to use independently. This software can be open source (DeepChem), freely available (MLViS), commercially available (AutoQSAR,

INTELLECTUAL PROPERTY	LEGAL CLASSIFICATION
Existence, extent of, and terms of partnerships	Trade secret
Sources of data	Trade secret
Methods of data processing	Trade secret
Software/algorithm enabling VS	Utility: method/process or trade secret
Result of VS: target identification	Trade secret
Result of VS: compound identification	Utility: composition of matter

Table 2. Innovation and intellectual property classification across the landscape of virtual drug discovery. The most valuable form of intellectual property that can be generated during drug discovery is a utility patent that protects the structure of a compound. Many of the other most valuable forms of intellectual property are best retained as trade secrets as they are difficult to protect and enforce through the patent system.

Desmond, PIPER), and/or a cloud service. Drug discovery teams utilizing this software give up no ownership in their development programs by simply paying for the right to use these methods. However, without significant expertise in integrating VS and experimental DD methods, these tools may not be deployed effectively. The inefficiency both slows the growth of the VS field and limits the potential upside within the DD programs hoping to benefit from VS techniques. The benefit of this approach for the software developer is that it allows them to focus their resources on improved methods for virtual screening, without linking the success of their company to the unpredictability of the DD process. A drug discovery team may prefer this approach if there is only nascent interest in VS methods, limited available resources to invest in the project, or the benefit of the program is uncertain.

Examples of companies: Schrodinger, D.E. Shaw Research, Stanford (Vijay Pande lab).

INTELLECTUAL PROPERTY

The numerous business models that can be implemented naturally raises interesting questions about how intellectual property is handled in computational approaches to drug discovery **(Table 2)**. The unique convergence of two fields, machine learning and drug discovery, creates new opportunities for developing and protecting IP. Some forms of IP will be best protected as patents while for others patenting may be inappropriate. For the majority of the IP generated in this field, retaining trade secrets and using them to gain a competitive advantage will likely be the most effective strategy.

The most valuable IP that is likely to be generated during the DD process is related to the identification of a new compound. A utility patent (composition of matter) filed to protect the structure of this compound will

provide a highly defensible and enforceable IP position. A patent demonstrating a new use for a known compound can be much more difficult to enforce, but still provide some value to both maintain freedom to operate and block competitors [14]. The datasets used to drive the VS algorithms, and preprocessing steps on these datasets represent IP best maintained as trade secrets. These are essential components for successful virtual drug discovery, as the quality of the dataset will directly impact the quality of the algorithms.

On the other hand, the process of target identification can generate highly valuable information, but these findings cannot be protected through the patent process. The methods developed to identify compounds and/or targets, such as VS algorithms, may be patentable as a methods utility patent, but enforcement of these patents can be difficult, as algorithms themselves are not patentable in the US system [15]. Including a technological advance in the associated hardware, such as an integration of virtual screening and experimental validation may produce IP that is much more readily protectable through patents. Unless there is a strategic reason to patent these methods, keeping them internal as a trade secret may be preferable.

DISEASE-AREA FOCUS OF VIRTUAL DRUG DISCOVERY COMPANIES

The number of biomedical technology startups with core expertise in artificial intelligence and machine learning has increased in recent years. Over 50 companies have raised first equity financing since January 2015. In the drug discovery space, startups are implementing computationally intensive methods and machine learning for prediction of ligand activity and target structure as well as hit discovery, optimization, and prioritization.

We compiled a list of key players that implement machine learning and computation for drug discovery **(Supplementary Table 1)**. We identified 15 companies that focus on a number of different problems ranging from structure-based drug design and molecular network-driven discovery to construction of knowledge graphs for inferring novel drug-target relationships. The identified set of companies targeted a wide range of disease types **(Figure 4A)** of which a majority of them had a program in the oncology space. The methods implemented by these companies were classified into machine learning-based or computational biochemistry-based **(Figure 4B)** to explore the relationship between the number of disease areas focus and the computational approach adopted. The resultant software products developed at these companies range from molecular dynamics simulation tools and VS-based target prioritization methods to deep learning approaches **(Supplementary Table 2)**. Information about these companies were extracted from their websites, associated press releases, and news articles. In certain cases, the availability of a supporting publication allowed us to better characterize a computational platform or machine learning framework. Below we provide examples of innovative companies:

Nimbus Therapeutics is a Cambridge, Massachusetts-based biotech company that applies expertise in computational chemistry to design medicines for human diseases falling broadly within the domains of metabolism, oncology and immunology. With the goal of capturing the mechanistic relationship between these disorders, Nimbus has four disclosed targets in its pipeline: TYK2, STING, ACC, and IRAK4. In 2016, Gilead acquired Nimbus' lead program targeting NASH (nonalcoholic fatty liver disease) and related metabolic disorders for $400M upfront and up to another $800M in development milestones. On termination of this program, Nimbus Apollo, the subsidiary that owns the acetyl-coA carboxylase (ACC)

Figure 4A

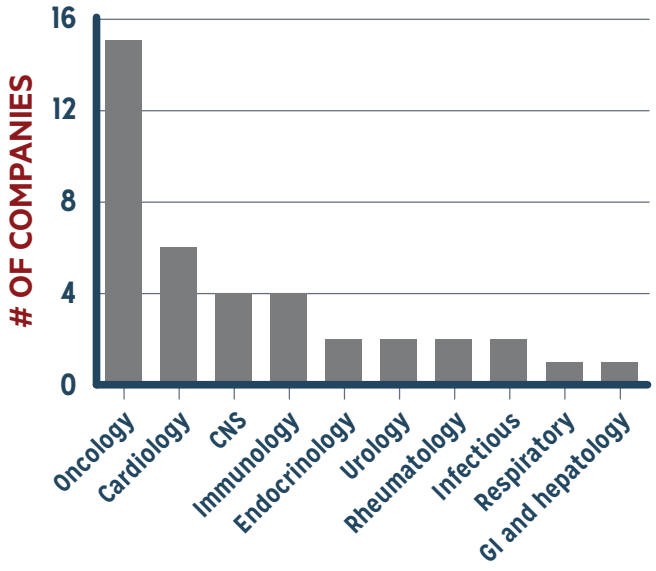


Figure 4B

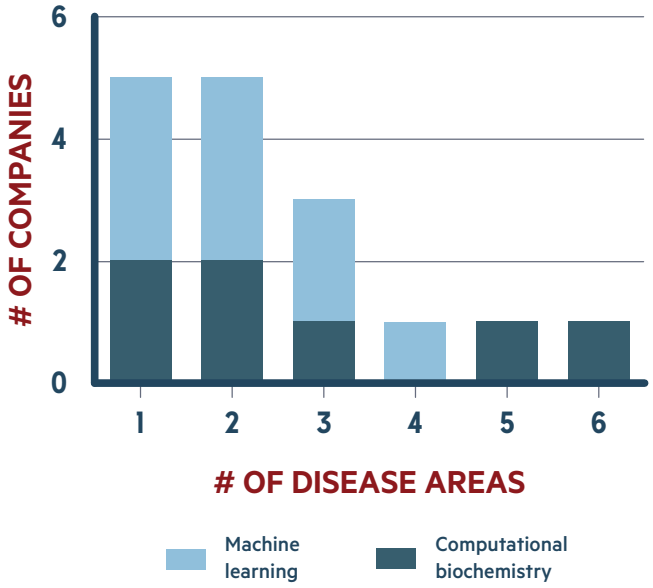


Figure 4. Disease area focus among companies. (a) We identified the specific disease areas of focus of 15 companies and categorized them into ten groups, based on the ten highest ranked therapeutic areas17. Cloud Pharmaceuticals has the most number of disease areas of focus (6) – CNS, oncology, rheumatology, urology, infectious, and respiratory. (b) The distribution of companies pursuing multiple disease areas. Also annotated in this plot are stacked bars representing a binary classification of these companies into being computational biochemistry-focused versus being data-driven and machine learning-focused. The high data requirements of machine learning algorithms could limit companies that focus on this class of approaches to limit themselves to fewer disease types.

franchise was acquired by Gilead. Furthermore, ACC can be potentially relevant for numerous disease areas such as cancer metabolism, lipid-related disorders, and inflammatory disease.

Relay Therapeutics is engineering a DD engine to detect and characterize the dynamic interactions and motion of proteins. The Cambridge, Massachusetts-based startup was launched by a $57 million Series A financing by Third Rock Ventures last September. Relay Therapeutics leverages advances in bio-chemistry and computation to rationally design allosteric modulators and detect conformational changes in disease-causing proteins upon binding by a small molecule. Molecular simulations enabled by powerful hardware will aid the exploration of the relationships between structure, stability and interactions to drive the development of breakthrough medicines. Relay's breakthrough drug-candidate research can synthesize more informative digital movies of proteins inside cells as opposed to traditionally generated static images. This technology is enabled by integrating imaging, such as magnetic resonance and X-ray, with powerful analytics algorithms and allows drug developers to see whether a particular protein target is still viable after a conformational change to the protein. The company's initial programs revolve around developing therapeutics in oncology.

Atomwise uses a proprietary deep learning algorithms to search across millions of potential compounds for novel small molecule discovery.

Deep learning methods utilize multiple nonlinear processing units for feature extraction and inference of hierarchical representations of the input data. The San Francisco-based company received seed funding worth $6.3 million from technology investors such as Khosla Ventures, Draper Fisher Jurvetson, and OSFund. The first scientific paper published by Atomwise describes AtomNet [16], which is a deep convolutional neural network focused on structure-based rational drug design by predicting the bioactivity of small molecules. Atomwise's most dramatic work involves the Ebola epidemic in 2014. After coordinating with a structural virologist to identify the mechanism that Ebola employs to invade healthy cells, scientists at Atomwise ran analyses and compiled a short list of potential Ebola inhibitors. Their framework effectively short-circuited the usual 14-year cycle required for the full DD process. More recently, Atomwise gained access to IBM technologies and expertise, enabling them to accelerate drug discovery for diseases such as malaria and cancer in addition to working with Autodesk and Merck on confidential projects.

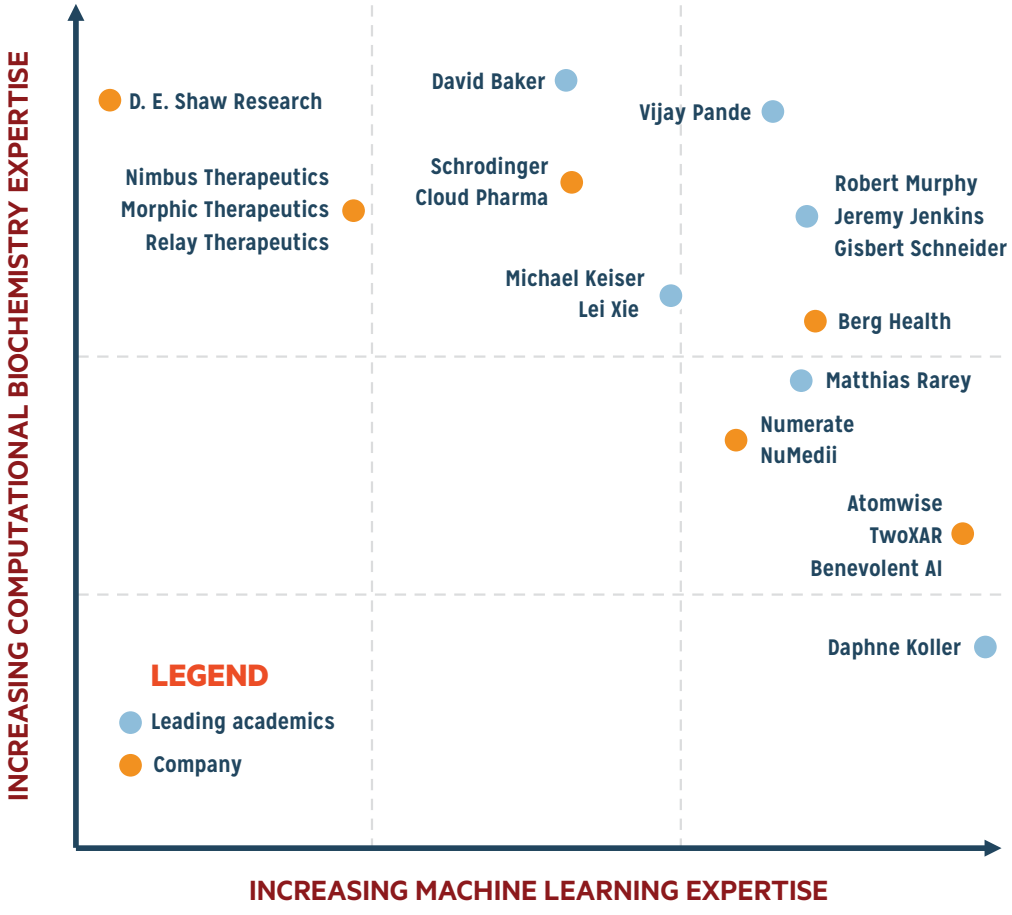


Figure 5. Diversity in computational drug discovery approaches in academia and industry. Expertise in computation-based approaches to drug discovery can be split along two axes: biochemistry-driven (vertical axis) and data-driven (horizontal axis). While most academic experts have expertise in both approaches, companies generally describe their unique strengths by emphasizing one method or the other, rather than synthesizing them.

BenevolentBio is a subsidiary of BenevolentAI and applies machine learning to massive medical databases. The mission of the company primarily revolves around rehashing highly fragmented scientific information into novel insight to accelerate drug discovery. BenevolentAI has raised around $100 million, backed by Upsher-Smith Laboratories, Neil Woodford's fund, Lundbeck and Lansdowne Partners amongst others. BenevolentBio experienced initial success in discovering potential treatments for ALS by identifying two compounds that performed as well as the gold standard while another two compounds performed even better. In November 2016, BenevolentAI acquired an exclusive license for a series of novel clinical stage drug candidates from Janssen Pharmaceuticals. The license agreement confers upon BenevolentAI the sole right to develop, manufacture, and commercialize these small molecules. The company aims to leverage its AI platform to evaluate the potential of these drug candidates towards a goal of beginning late-stage phase II clinical trials in mid-2017. The company is optimistic about the scope of its AI framework and aims to remain as versatile as possible with regard to target diseases of interest.

EXPERTISE OF COMPANIES & KEY OPINION LEADERS

Academic research has played a significant role in conceiving and developing computational approaches for drug discovery. We sought to understand and highlight the contributions from the academic community, which includes key opinion leaders in both drug discovery in general and computational drug discovery.

Expertise in both academia and industry can be divided into two general computation-based approaches **(Figure 5)**. The first is to begin from scientific first principles and computationally simulate the biochemical and physical interactions between a potential ligand and target pair. The second method is to use known experimental data related to a potential molecule (as well as related molecules) and corresponding targets to train machine learning tools to predict drug quality. These divergent approaches also mirror the dichotomy between structure-based and ligand-based screening, with the former typically done through explicit simulations while the latter sees more use of both techniques. Although the two computational methods are not mutually exclusive and a single project can synthesize both simulation and learning based approaches, they represent two fundamentally different scientific philosophies.

In the "biochemistry-driven" approach, a typical end-goal is to compute a physical quantity that describes the quality of a specific ligand-target interaction, such as a binding affinity. This general concept is older and more established. However, it has gained popularity in recent years as computational power has increasingly become available, allowing for higher quality simulations and screening of larger numbers of molecule-target pairs. Access to high-performance, cloud-based parallelized computation, such as Amazon Web Services and Google Compute Cloud, has further accelerated adoption of process- and memory-intensive analyses.

The more data-driven approaches, which leverage machine learning techniques, can similarly predict the quality of a specific ligand-target interaction, but they are frequently used to also estimate other drug properties relevant to lead optimization and ADME. For example, machine learning methods can simultaneously predict off-target effects and can be easily extended to identify new indications for existing drugs. These ideas are newer, but machine learning is quickly demonstrating potential, especially as more advanced methods, such as deep neural networks, gain power and accessibility.

Academic research has played a significant role in advancing both approaches. Biochemical modeling draws primarily on basic science knowledge; hence, computational simulation has been a topic of academic interest in biochemistry. One prominent academic in this field is David Baker, Professor of Biochemistry at the University of Washington, who focuses on the design and prediction of protein structures and folding along with analysis of interactions between proteins and other macro- and small molecules. His lab is more interested in designing proteins themselves rather than finding small molecules (and in a twist on traditional drug discovery, has actually designed a protein to bind a naturally occurring small molecule [18]). Baker cofounded a company, Virvio, which uses molecular simulation techniques to discover novel biotherapeutics.

Academic research is also driving significant advances in machine learning as a field on its own, making it natural that applications in drug discovery and development would present themselves to academics. Largely, many of the academic thought-leaders for machine learning in drug discovery also have expertise in more traditional computational approaches. An overview of this distribution is shown in **Figure 5**. Most prominent academics have strong expertise in both machine learning and biochemistry-driven approaches to drug discovery or other parts of the therapeutics development pipeline.

Several examples of academics can illustrate the dual nature of their expertise **(Supplementary Table 3)**. One such expert is Stanford University professor and Andreessen Horowitz general partner Vijay Pande, most notable for his Folding@Home project [19], which allows individuals to donate computation time on their own personal computers towards the goal of simulating protein folding dynamics from biophysical principles. In general, his research lab uses biophysical knowledge to simulate interactions between small molecules and biological pathways and macromolecules for the purpose of drug discovery. Recently, Pande partnered with Google in these endeavors to describe shifting from a supercomputing platform to a cloud-computing platform [20]. He furthered this partnership with Google to develop a deep learning-based method for synthesizing data across publicly accessible experimental datasets from multiple biological sources to virtually screen molecules for drug discovery [10].

Another example of an academic leader with broad expertise in both biochemistry and computational methods is Robert Murphy, Professor of Computational Biology at Carnegie Mellon University. The focus of Murphy's lab has been computational analysis of microscopy, particularly fluorescence, images to understand intracellular dynamics and subcellular protein localization. Though he has not been involved in computational methods for molecular modeling, he has made heavy use of biochemistry knowledge and computation for image analysis. Murphy began applying machine learning techniques to these problems before further transitioning into machine learning for drug discovery. He gained influence for developing active machine learning methods to efficiently predict which experiments would yield the most useful data for virtual screens. Murphy's lab showed that similar performance in virtual screens for predicting both on- and off-target effects can be achieved by using only a small but focused subset of experimental data [21].

Pande and Murphy exhibit a similar characteristic in that both began their academic careers in biochemistry before machine learning became a common computational method. Newer academic faculty members with biochemistry experience may be more likely to directly begin developing machine learning techniques for drug discovery. One example is UCSF Assistant Professor Michael

Keiser whose similarity ensemble approach combines machine learning and biochemistry knowledge to predict interactions between new ligand-target pairs using existing knowledge of interactions [22]. Similarly, CUNY Associate Professor Lei Xie has begun his career with key papers on the use of machine learning to predict in vivo binding affinity and side effects [23,24].

Even these researchers, however, tend to ensure that their work is informed by biochemical and biophysical modeling and simulation. In contrast, companies that use machine learning for drug discovery and development primarily describe their expertise as singularly in machine learning or artificial intelligence. TwoXAR and Atomwise, for instance, market their technology as "big data" and "deep learning," respectively, with few references to any expertise in the basic science needed for modeling and simulation. Similarly, companies that focus on modeling and simulation (e.g., D. E. Shaw Research or Relay Therapeutics) describe their superior abilities to compute physical quantities such as free energies, binding affinities, and structures without claiming artificial intelligence or machine learning expertise. In a nascent field, it is perhaps unsurprising that companies have not yet synthesized expertise in both approaches. Combining these strategies is a potential advantage for a new entrant.

COMPANY NAME	KEY PEOPLE (SCIENCE)	BUSINESS MODEL (MAKE LESS GRANULAR)	DISEASES AREAS	PROBLEM DOMAIN	APPROACH	FUNDING AND INVESTORS	PARTNERS
Nimbus Therapeutics	Ramy Farid PhD, Rosan Kapellera	Strong collaboration with Schrodinger, internal development of new molecules for targets that span disease areas. Nimbus is a holding company for subsidiaries which run discovery programs	Oncology, Immunology	Deep computational chemistry throughout the drug discovery and development process to create novel small molecules for human diseases. Targets: TYK2, STING, ACC, IRAK4	Computational biochemistry	$72M, Atlas Venture, Lightstone Ventures, Schrodinger, SR One, Lilly Ventures, Pfizer Venture Investments	Gilead, Monsanto, Genentech
Morphic Therapeutics	Tim Springer PhD, Ramy Farid PhD, Mark Murcko PhD	Strong collaboration with Schrodinger, internal development of new molecules for targets that span disease areas.	Oncology, Cardiology, Immunology	Develop first orally administered Integrin inhibitor. Performs conformational modeling of integrins using computational and experimental approaches.	Computational biochemistry	$51M, Polaris Partners, T. A. Springer, Schrodinger, SR one, Pfizer Venture Investments, Omega Funds, Abbie Ventures	Schrodinger
Relay Therapeutics	Pascal Fortin PhD, Jon Weiss, Pat Walters PhD	Strong collaboration with DesRes, internal development of new molecules for targets that span disease areas.	Oncology	Deploy allosteric drug discovery platform to develop breakthrough medicines	Computational biochemistry	$57M, Third Rock Ventures, DE Shaw Research	DE Shaw Research, Schrodinger, X-Chem Pharmaceuticals, Biodesy
Schrodinger	Ramy Farid PhD, Robert Abel PhD, Leah Frye PhD	Partner with Biotech/pharma; exclusive/specialized collaborations. Selling packaged software to drug discovery groups.	N/A	Molecular modeling, drug design and materials science software.	Computational biochemistry	$52M, Bill Gates, Cascade Investment, Scott Becker	Altoris, AWS, ChemAxon, Cycle Computing, DE Shaw research, eMolecules, Exxact, John McNeil & Co.
D.E. Shaw Research	David Shaw, many scientists	Partner with Biotech/pharma; exclusive/specialized collaborations. Selling packaged software to drug discovery groups. Sells some molecular dynamics software via Schrodinger collab	N/A	Computational biochemistry, high-speed MD simulations of proteins/macromolecules. Structural changes underlying biological phenomena.	Computational biochemistry	The D. E. Shaw Group	National Resource for Biomedical Supercomputing (NRBSC), Schrodinger, Relay Therapeutics
TwoXAR	Andrew A. Radin, Andrew M. Radin	Aggregate/synthesize diverse biological datasets to predict drug/target for a disease, internal software development, outsourced in vivo testing, assist drug discovery efforts with partners (prioritize existing candidates, perform targeted searches, identify new targets)	Oncology, Urology, Rheumatology, Cardiology, Endocrinology	Empirical, statistical prediction of drug/target/disease relation. Does not seem to be using mechanistic information/MD/chemoinformatics	Data-driven machine learning	$3.4M, Andreessen Horowitz, CLI Ventures, StartX	Stanford Medicine, Michigan State, Mount Sinai, University of Chicago, Vium, Parkinson's Progression Markers Initiative, MIT Startup Exchange, Startx
Atomwise	Izhar Wallach, Michael Dzamba, Abraham Heifets	Internal software deveopment for candidate identification, library pruning ; external partnership with pharma/biotech/academia animal models/clinical studies	Infectious, Oncology, Immunology	Deep convolutional neural networks for structure based drug design	Data-driven machine learning	$6.4M, AME Cloud Ventures, Data Collective, Draper Associates, Draper Fisher Jurvetson, Khosla Ventures + 2 more	Stanford University, Scripps Research Institute, University of Toronto, Merck, UC San Diego, Notable Labs, Princess Margaret Hospital, CAMH, SickKids, Autodesk, IIT Bombay, University of the Philippines, Merck, IBM
In Silico Medicine	Alex Zhavoronkov, Michael Levitt PhD, Dr. Donald Small, Dr. Bud Mishra, Dr. Alexey Moskalev, Dr. Kristen Fortney, Dr. Yuri Nikolsky, Dr. Charles Cantor	Internal software development, collaborates and provides services to academia/drug discovery teams/cosmetics industry; drug discovery and biomarker discovery; tools for aging research. Also performs independent internal research: ID new targets/candidates	Oncology	(see business model/specific approaches)	Data-driven machine learning	Deep Knowledge Ventures	Nvidia, Asus, BioTime, JHU, IBM Watson + ~ 150 total
Berg Health	Carl Berg, Mitch Gray, Niven R Narain	Proprietary internally developed software platform to perform patient stratification, biomarker/compound discovery; internal drug discovery/development. Also sells services to perform analytics for patient monitoring/management.	Oncology, Cardiology, CNS	Prediction for patient response / toxicity and drug discovery through deep learning from biological datasets, patient samples and mass spec inputs.	Data-driven machine learning	Carl Berg, Pathfinder Management	DoD – Breast Cancer+Prostate Disease Research, Pancreatic Cancer Research Team, HMS, Cancer Research and Biostatistics, Genomics England, MD Anderson, Mount Sinai, Miller School of Medicine, MSKCC, Weill Cornell, MedUniv SC
Cloud Pharmaceuticals	Shahar Keinan PhD	Work with organizations to jointly design new drugs and then partner with later-stage developers. Partnerships to jointly develop drugs. Also wholly owned, independent drug development pipeline efforts	CNS, Oncology, Rheumatology, Urology, Infectious, Respiratory	Quantum molecular design and cloud computing to improve the accuracy and success probability of drug discovery	Computational biochemistry	$1.4M, NSF SBIR Grant	Genomeon, University of Florida
Numerate	Brandon Allgood	Internal development programs; internal software; external validation	CNS, Cardiology	Drug design platform that can deliver novel leads with no need for crystal structure and very little SAR data. Model the phenomena that are critical to the success of hits, leads and candidates	Data-driven machine learning	$17.4M, Atlas Ventures, Lilly Ventures, LanzaTech Ventures	J. David Gladstoe Institutes, Google Compute engine, Sage Bionetworks, IBM, Merck
NuMedii	Gini Deshpande PhD, Craig Webb PhD, Asim Siddiqui	Internal software development; partners for clinical development and commercialization of its de-risked drug candidates; Partners provide data to discover new uses for discontinued or de-prioritized drug candidates.	GI and hepatology, Cardiology, Oncology, Immunology	Use human clinical and molecular network data to drive discovery. Use expertise to select and de-risk commercially viable drug candidates.	Data-driven machine learningy	$5.5M, Claremont Creek Ventures, Lightspeed Venture Partners	Allergan, California Kids Cancer Comparison project, Astellas + Some big pharma

Supp. Table 1. Sampling of notable companies.

COMPANY NAME	KEY PEOPLE (SCIENCE)	BUSINESS MODEL (MAKE LESS GRANULAR)	DISEASES AREAS	PROBLEM DOMAIN	APPROACH	FUNDING AND INVESTORS	PARTNERS
Envisagenics	Maria Luisa Pineda PhD, Martin Akerman PhD	Software as a service (SaaS) to extract biologically relevant RNA isoforms from raw RNA-seq data, external pharma partners	Oncology	In-silico RNA therapeutic and Biomarker Discovery	Computational biochemistry	<$1M, NIH, Accelerate Long Island, Long Island Emerging Technologies Fund, topspin, Jove Equity Investors	Cold Spring Harbor Laboratory
Benevolent AI	Jackie Hunter, Kenneth Mulvany, Jerome Pesenti	Internal software development, identifying new compounds	CNS, Oncology	Harness AI to enhance scientific discovery through integration of highly fragmented information. Company aims to remain as flexible as possible should opportunities present themselves in the future	Data-driven machine learning	$100M, Lansdowne Partners, Lundbeck, Upsher Smith Laboratories, Woodford Investment Management	Janssen Pharmaceutica (Exclusive license for novel clinical stage drug candidates)
Lantern Pharma	Arun Asaithambi PhD, Peter Nara PhD, Jeff Keyser PhD	Partners with, licenses or acquires abandoned late-stage clinical drugs that are selectively effective. Performs 'rescue' and 'repurposing' of these drugs. Drug along with biomarker diagnostic will be partnered or licensed out.	Oncology	Tailors promising precision drug programs to the right cancer patients through genomics and AI. Identifies patient biomarkers for in-licensed, de-risked drugs that are discontinued for development. Conducts focussed phase 2 clinical trials – create exit points for the drug	Data-driven machine learning	$1M, Green Park & Golf Ventures, Health Wildcatters	
Micar21	Filip Fratev	Internal software development utilizing Schrodinger packages; provide virtual screening services for biopharma	Oncology, Cardiology	Genetic screen analysis, In silico drug design	Data-driven machine learning		
Google Research	Patrick Riley, Dale Webster, Bharath Ramsundar, Vijay Pande	None yet. Academia/industry collaboration to develop software	N/A	Massively multitask neural architectures provide a learning framework for drug discovery that synthesizes information from many distinct biological sources	Data-driven machine learning	N/A	Stanford (Vijay Pande lab)

Supp. Table 1 (continued). Sampling of notable companies.

COMPANY NAME	PRODUCT NAME	PRODUCT DESCRIPTION
D.E. Shaw Research	Desmond	High-speed molecular dynamics simulations of biological systems on conventional commodity clusters
	Random123	Parallelizable counter-based random number generators (CBRNGs) that were originally developed for Anton but may be useful for a wide range of applicationsus compounds simultaneously.
	Cascade	C++ library for efficient cycle-based simulation of hardware architectures
	HiMach	Parallel analysis framework that allows users to write trajectory analysis programs sequentially
	TimeScapes	Analysis package that can be used to efficiently detect and characterize significant conformational changes in simulated biomolecular systems
Schrodinger	Small molecule drug-discovery suite	Virtual screening, rank order compounds, target preparation, ligand preparation, workflow automation
	Biologics suite	Comprehensive protein modeling, protein engineering, antibody modeling, protein-protein docking
	Materials science suite	Atomic-scale chemical simulation, predictive capabilities, in silico discovery and automation
	Discovery informatics suite	Data visualization, analysis, compound registration
TwoXAR	DUMA drug discovery platform	A secure cloud-based solution that uses a patent-pending algorithm to find unanticipated associations between drug and disease.
Atomwise	AtomNet	Deep convolutional neural network for bioactivity prediction in structure-based drug discovery
Berg Health	Interrogative Biology Platform	A combination of adaptive-omic biological data and advanced AI machine learning algorithms allow us to stratify patient populations by phenotype to build predictive models
	bAIcis	Analytics engine that empowers health plans, providers, EHR vendors, and clinical decision support organizations to plan, implement and measure interventions that benefit individuals
Envisagenics	SpliceCore	Software as a service (SaaS) to extract biologically relevant RNA isoforms from raw RNA-seq data

Supp. Table 2. Sampling of notable software products.

PERSON	BACKGROUND	KEY APPROACHES / INNOVATIONS / CONTRIBUTIONS	COMPANIES / OTHER INDUSTRY WORK
Robert Murphy (Carnegie Mellon University)	Biochemistry, fluorescence imaging	Active learning to simultaneously predict effects of many compounds on many targets and suggest high-value experiments [21, 25]	Quantitative Medicine, LLC
Vijay Prande (Stanford University)	Physics, structural biology	Markov state models for molecular dynamics [26]; Distributed and cloud-computing for molecular simulations [19, 20]; Massively multitask networks for virtual screening [10]	General partner at Andreessen Horowitz; Scientific advisor at Schrodinger
David Baker (University of Washington)	Biochemistry, biophysics	De novo design of proteins for binding to specific ligands [18]	Virvio
Jeremy Jenkins (Novartis Institues for Biomedical Research)	Molecular genetics, molecular simulation	Machine learning with HTS fingerprinting for hit expansion, especially with new molecules [27, 28]	Director of Research Informatics, Novartis
Michael Kaiser (University of California-SF)	Bioinformatics	Similarity ensemble approach – use of knowledge of known ligands and targets, as well as their similarities, to predict off-target effects of both candidate compounds and existing drugs [22, 28]	
Daphne Koller (Stanford University)	Computer science, machine learning	Machine learning analysis of gene co-fitness data to predict drug targets [29]; machine learning for identification of protein-protein binding sites [30]	Chief Computing Officer, Calico
Gisbert Schneider (ETH Zurich)	Biochemistry, computer science	Machine learning identification of targets for new, computationally designed, compounds [31]; machine learning to speed up molecular dynamics simulations [32]; fragment-based *de novo* design of small molecules [33]	inSili.com Gmbh, AlloCyte Pharmaceuticals
Lei Xie (City University of New York)	Chemistry, computer science	Machine learning for prediction of binding kinetics and *in vivo* drug activity [24]; machine learning for drug repurposing and prediction of side effects [23]	
Matthias Rarey (University of Hamburg)	Computer science, bio/ cheminformatics	Efficient searching of virtual ligand libraries [34]; active learning to reduce required number of training compounds [35]	BioSolveIT

Supp. Table 3. Sampling of key thought leaders and areas of expertise.

TEAM MEMBER BIOGRAPHIES

ASHVIN BASHYAM is a third-year PhD candidate in Electrical Engineering & Computer Science at MIT advised by Michael Cima at the Koch Institute for Integrative Cancer Research. He is a recipient of both the Hertz Foundation Graduate Fellowship and the NSF Graduate Research Fellowship. Broadly, Ashvin is interested in translational biomedical research with a particular focus on diagnostics. His graduate research brings magnetic resonance to bear in assessing fluid dysregulation disorders for clinical and military applications. As President of the MIT Biotechnology Group, Ashvin leads programs aimed at fostering the translational community at MIT. Specifically, he has helped build a life sciences focused angel investor network, published analysis and recommendations to improve university technology transfer, and is currently leading a student-run due diligence network that helps investors assess opportunities. Ashvin received his undergraduate degree in Biomedical Engineering from the University of Texas at Austin where he worked with Stanislav Emelianov on photoacoustic imaging for cancer staging.

JAIDEEP DUDANI is a fourth-year PhD candidate in Biological Engineering at MIT, working in the lab of Sangeeta Bhatia at the Koch Institute for Integrative Cancer Research. He is currently an NSF fellow and Ludwig Center for Molecular Oncology fellow. He is broadly interested in the development and translation of innovative scientific concepts to useful technologies. His research is focused on developing functional biomarkers of disease that can be therapeutically targeted using injectable nanotechnologies that probe and perturb proteolytic enzymes aberrantly expressed in numerous cancers and during bacterial infection. He collaborates closely with industry, clinicians, and academic groups. He holds an undergraduate degree in Bioengineering from UCLA, where he developed microfluidic technologies for high-throughput single-cell phenotyping in the lab of Dino Di Carlo. He has published over ten peer-reviewed papers (including 7 first-author publications) and has several patent filings.

KARTHIK MURUGADOSS is a second-year S.M. candidate in Computational Sciences at MIT advised by Manolis Kellis at the Computer Science and Artificial Intelligence Laboratory. Karthik is interested in regulatory genomics and epigenomics applied to cancer research through the application of machine learning and statistical inference. Specifically, he is focused on the molecular dysregulation within tumor cells that impacts their interaction with the immune system and how this information can inform therapeutic decision-making. After graduation, Karthik will be joining nference, a Cambridge-based startup where he has worked in the past on research projects encompassing natural language processing (NLP) and deep learning neural networks. He independently developed a software platform for "vectorizing" several thousands of academic journal papers (from the PLoS Open Access database) and applying state-of-the-art deep learning methods to learn relevant information from them. Karthik received his undergraduate degree in Mechanical Engineering from the Indian Institute of Technology, Madras.

VARESH PRASAD is a fourth-year PhD candidate in Medical Engineering & Medical Physics in the Harvard-MIT Program for Health Sciences & Technology. He is advised by Thomas Heldt at the Institute for Medical Engineering and Science and is a recipient of a National Defense Science & Engineering Graduate fellowship. Varesh is interested in hardware and software technology for innovations in various aspects of healthcare. His graduate research involves

the use of machine learning and computational techniques to inform clinical management of patients in acute and critical care settings such as in major surgeries, the emergency department, and intensive care. In the past, he has worked with a start-up medical device company in India as a Whitaker International Fellow, where he designed imaging tools appropriate for ophthalmic screening in adults and infants in regions of the world with few eye care specialists. Varesh received his undergraduate degree in Bioengineering from the University of Pennsylvania, where he worked with Peter Davies to show how the design of coronary artery stents interacts with blood flow patterns to influence endothelial cell responses and adverse events in coronary interventions.

REFERENCES

1. Hughes, J. P., Rees, S., Kalindjian, S. B. & Philpott, K. L. Principles of early drug discovery. *Br. J. Pharmacol.* **162**, 1239–49 (2011).
2. Strovel, J. et al. *Early Drug Discovery and Development Guidelines: For Academic Researchers, Collaborators, and Start-up Companies. Assay Guidance Manual* (Eli Lilly & Company and the National Center for Advancing Translational Sciences, 2004).
3. Murphy, R. F. An active role for machine learning in drug development. *Nat. Chem. Biol.* **7**, 327–330 (2011).
4. Esch, E. W., Bahinski, A. & Huh, D. Organs-on-chips at the frontiers of drug discovery. *Nat. Rev. Drug Discov.* **14**, 248–260 (2015).
5. Katsila, T., Spyroulias, G. A., Patrinos, G. P. & Matsoukas, M.-T. Computational approaches in target identification and drug discovery. *Comput. Struct. Biotechnol. J.* **14**, 177–184 (2016).
6. Lavecchia, A. Machine-learning approaches in drug discovery: Methods and applications. *Drug Discovery Today* **20**, 318–331 (2015).
7. Unterthiner, T., Mayr, A., Klambauer, G. & Hochreiter, S. Toxicity Prediction using Deep Learning. *arXiv* (2015).
8. Bodenreider, O. & Stevens, R. Bio-ontologies: current trends and future directions. *Br. Bioinform* **7**, 256–274 (2006).
9. Attwood, T. K. et al. Calling International Rescue: knowledge lost in literature and data landslide! *Biochem. J.* **424**, 317–333 (2009).
10. Ramsundar, B. et al. Massively Multitask Networks for Drug Discovery. (2015).
11. Booth, B. The Nimbus Experiment: Structure-Based Drug Deals. *Life Sci VC* (2013). Available at: https://lifescivc.com/2013/06/the-nimbus-experiment-structure-based-drug-deals/. (Accessed: 19th February 2017)
12. Booth, B. Nimbus Delivers Its Apollo Mission: A $1.2B Gilead Partnership. *Life Sci VC* (2016). Available at: https://lifescivc.com/2016/04/nimbus-delivers-apollo-mission-1-2b-gilead-partnership/. (Accessed: 19th February 2017)
13. Keiper, J. Life In The Trenches Of A Biotech LLC. *Life Sci VC* (2016). Available at: https://lifescivc.com/2016/06/life-trenches-biotech-llc/. (Accessed: 19th February 2017)
14. Oprea, T. I. & Mestres, J. Drug Repurposing: Far Beyond New Targets for Old Drugs. *AAPS J.* **14**, 759–763 (2012).
15. Biewald, L. Why did Google open-source their core machine learning algorithms? CrowdFlower (2015). Available at: https://www.crowdflower.com/why-did-google-open-source-their-core-machine-learning-algorithms/. (Accessed: 30th March 2017)
16. Wallach, I., Dzamba, M. & Heifets, A. AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery. *arXiv*

Prepr. arXiv1510.02855 1–11 (2015). doi:10.1007/s10618-010-0175-9
17. Karlberg, J. P. E. Trends in disease focus of drug development. *Nat. Rev. Drug Discov.* **7**, 639–640 (2008).
18. Tinberg, C. E. et al. Computational design of ligand-binding proteins with high affinity and selectivity. *Nature* **501**, 212–6 (2013).
19. Larson, S. M., Snow, C. D., Shirts, M. & Pande, V. S. Folding@Home and Genome@ Home: Using distributed computing to tackle previously intractable problems in computational biology. 31 (2009).
20. Kohlhoff, K. J. et al. Cloud-based simulations on Google Exacycle reveal ligand modulation of GPCR activation pathways. *Nat. Chem.* **6**, 15–21 (2014).
21. Naik, A. W., Kangas, J. D., Langmead, C. J. & Murphy, R. F. Efficient modeling and active learning discovery of biological responses. *PLoS One* **8**, e83996 (2013).
22. Keiser, M. J. et al. Predicting new molecular targets for known drugs. *Nature* **462**, 175–181 (2009).
23. Lim, H. et al. Large-Scale Off-Target Identification Using Fast and Accurate Dual Regularized One-Class Collaborative Filtering and Its Application to Drug Repurposing. *PLoS Comput. Biol.* **12**, e1005135 (2016).
24. Chiu, S. H. & Xie, L. Toward High-Throughput Predictive Modeling of Protein Binding/Unbinding Kinetics. *J. Chem. Inf. Model.* **56**, 1164–1174 (2016).
25. Kangas, J. D., Naik, A. W. & Murphy, R. F. Efficient discovery of responses of proteins to compounds using active learning. *BMC Bioinformatics* **15**, 143 (2014).
26. Bowman, G. R., Beauchamp, K. A., Boxer, G. & Pande, V. S. Progress and challenges in the automated construction of Markov state models for full protein systems. *J. Chem. Phys.* **131**, 124101 (2009).
27. Riniker, S., Wang, Y., Jenkins, J. L. & Landrum, G. A. Using information from historical high-throughput screens to predict active compounds. J. *Chem. Inf. Model.* **54**, 1880–1891 (2014).
28. Lounkine, E. et al. Large-scale prediction and testing of drug activity on side-effect targets. *Nature* **486**, 361–7 (2012).
29. Lasko, T. A., Denny, J. C. & Levy, M. A. Computational Phenotype Discovery Using Unsupervised Feature Learning over Noisy, Sparse, and Irregular Clinical Data. *PLoS One* **8**, e66341 (2013).
30. Wang, H. et al. InSite: a computational method for identifying protein-protein interaction binding sites on a proteome-wide scale. *Genome Biol.* **8**, R192 (2007).
31. Reker, D., Rodrigues, T., Schneider, P. & Schneider, G. Identifying the macromolecular targets of de novo-designed chemical entities through self-organizing map consensus. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 4067–72 (2014).
32. Rupp, M. et al. Machine Learning Estimates of Natural Product Conformational Energies. *PLoS Comput. Biol.* **10**, e1003400 (2014).
33. Reutlinger, M., Rodrigues, T., Schneider, P. & Schneider, G. Multi-Objective Molecular De Novo Design by Adaptive Fragment Prioritization. *Angew. Chemie Int. Ed.* **53**, 4244–4248 (2014).
34. von Behren, M. M., Bietz, S., Nittinger, E. & Rarey, M. mRAISE: an alternative algorithmic approach to ligand-based virtual screening. J. *Comput. Aided. Mol. Des.* **30**, 583–594 (2016).
35. Lang, T., Flachsenberg, F., Von Luxburg, U. & Rarey, M. Feasibility of Active Machine Learning for Multiclass Compound Classification. *J. Chem. Inf. Model.* **56**, 12–20 (2016).