ELSEVIER

Reaxys®
DRUG DISCOVERY & DEVELOPMENT

Case Study: Dr. Jonny Wray, Head of Discovery Informatics at e-Therapeutics PLC

Clean compound and bioactivity data are essential to successful modeling of the impact of a compound on a biological system

SUMMARY

e-Therapeutics hunts for novel, impactful medicines in a very different way. Their workspace is filled with computers, not lab benches. Their research subjects are models of biological networks, which they probe to find compounds that cause a desired change. The results are promising—and data from Reaxys are instrumental to their work.

ELSEVIER

"Reaxys is the largest source of compound and bioactivity data that feed into our library"

- Dr. Jonny Wray, Head of Discovery Informatics, e-Therapeutics

Biography

e-Therapeutics is a drug development group in Oxford, UK, that uses a network biology approach to discover novel drug candidates for biologically complex diseases. Dr. Jonny Wray joined e-Therapeutics in 2011. Trained as a computational neuroscientist, he saw this switch from academia to corporate as an opportunity to apply his strengths in computational theory towards solving a very real problem in pharmaceutical development. Over the last five years, he has built up the informatics underpinning the innovative work that distinguishes e-Therapeutics from other drug discovery companies. Dr. Wray tells us about the practical aspects of using network models to identify drug candidates and the way in which Reaxys supports this goal.

HOW DOES A COMPUTATIONAL NEUROSCIENTIST FIT INTO THE PICTURE OF DRUG DEVELOPMENT?

Neuroscience was one of the first biological disciplines to embrace the idea that functions and properties can emerge from a system, rather than its individual components. A single neuron has a pretty straightforward, well-understood job; but it is the collection of neurons interacting in a network that give rise to perception and thought. In a similar vein, we understand a great deal about the function and interactions of an individual protein, but cellular phenotype, whether diseased or normal, emerges from several proteins interacting in complex networks. The problem with standard drug discovery approaches is that they attempt to perturb a single molecule in hopes to manipulate a phenotype. For biologically complex diseases, this constitutes a significant, and often blind, gap between the action—inhibiting a protein—and the effect—changing a phenotype. Where my training contributes to our work is in enabling a way to perturb the network. Manipulating networks bridges that gap between molecules and phenotypes.

We perturb networks in silico; that is, we build network models of disease mechanisms and manipulate them computationally to observe the outcomes of these perturbations. One can envision techniques that would allow examining the interactions between a handful of proteins in the lab, but that is still quite limited. The in silico models we construct reflect the involvement of 500–1,500 proteins. Furthermore, we are also able to test the impact of millions of compounds on these network models. So we have a far greater analysis scope.

When I started at e-Therapeutics, the data to make such models even possible were just beginning to surface. Over the course of four years, I applied my knowledge from constructing neural networks to create the informatics framework that allows us to portray and test complex protein–protein networks relevant to a disease. Now we have a number of ongoing projects, two of which are at lead optimization stages.

WHAT IS YOUR PROCESS FOR BUILDING THESE NETWORK MODELS, AND HOW DO YOU CAPTURE THE FULL COMPLEXITY OF EVENTS UNDERLYING A DISEASE?

We start by looking at gaps in available treatments and selecting diseases that are particularly amenable to a network approach. At this stage, our disease biologists define an "intervention strategy", which means how we want to approach the problem, based on exhaustive information about known mechanisms of the disease, as well as clinical data on disease progression, diagnostic methods and diagnosis time point.

This intervention strategy guides the de novo construction of an initial model, tailored to the strategy. We have an integrated database on protein-protein interactions and signaling pathways of the cell curated with data from the primary literature, which we use to subsequently enhance the model with networks neighboring and/or interacting with those identified in the intervention strategy. The addition of these "guilty by association" networks is key to our approach. All too often in drug development, focus resides on one or two pathways and ignores alternative paths that can lead to the robustness responsible for decreasing the efficacy of a drug. Our analyses do not only consider the canonicals, but rather look for synergistic effects.

Our models do not cover the entire complexity of a given system. With 500–1,500 proteins in a model, we cover approximately 10% of what is happening in a cell. One reason is that we have to focus on functions that are present and disease-specific to make the model meaningful. Another issue is that biological data are noisy and incomplete. That is, we don't really have the knowledge to depict the entire complexity. Nevertheless, we construct multiple models for every intervention strategy. These models differ in their ruling assumptions or starting data. If we test these models and get similar answers from all of them, we can confidently pursue the resulting insights.

MODELS IN HAND, HOW DOES REAXYS SUPPORT USING THEM TO GAIN INSIGHTS FOR DRUG DISCOVERY?

Our next step is to perturb the models, observe the impact of those perturbations, and gather information that helps select optimal drug candidates for the particular intervention strategy. One way to do that is to identify a subset of proteins from our models that upon perturbation maximize the desired outcome defined by the intervention strategy. For example, you can remove combinations of proteins and measure to what extent the network becomes fragmented. Needless to say, with 1500 proteins in a model, it would take an inordinate amount of time to go through each combinatorial possibility. Instead, we use stochastic optimization algorithms that streamline the selection of combinations most likely to generate the desired effect. In the end, armed with this selected subset of proteins, we turn to our virtual compound library to identify drug candidates known or predicted to target those proteins. Alternatively, we may screen the virtual compound library directly in our models, assessing the impact of a compound's bioactivity against a null hypothesis of random perturbation.

The outcome is a set of compounds with a desired chemical biology "footprint;" that is, their predicted activity for a given disease-intervention constellation based on our models. This enriched set may be narrowed down further based on drug likeness, manufacturability, patent conflicts and known safety profile. Ultimately, roughly 1000 strong candidates are then tested in phenotypic assays, which is a good match because by their very nature, phenotypic assays do not accommodate the high throughput needed to blindly screen large libraries of compounds.

Whichever approach is used, our virtual compound library is a vital component of our workflow. It contains roughly 10 million compounds, approximately half with known bioactivity data and the other half with activity information predicted by machine learning. Reaxys is the largest source of compound and bioactivity data that feed into our library and as such, a cornerstone of our ability to connect insights from our models with interesting compounds, many of which are proving to be novel drugs.

> **“We have empirically proven that our network-based approach to drug discovery works.”**
>
> - Dr. Jonny Wray, Head of Discovery Informatics, e-Therapeutics

HOW DO YOU AUGMENT THE CONTENT OF YOUR VIRTUAL COMPOUND LIBRARY WITH REAXYS CONTENT? WHAT DATA DO YOU USE AND WHAT CHALLENGES DO YOU FACE?

We use the Reaxys Flat Files to feed our virtual library. Those files are updated more often than we can practically deal with, but we update our content from Reaxys approximately twice a year. Some of the data we use are bioactivity values, assay and measurement type, and associated references—basically anything that serves as evidence of compound-target interaction. We also use the web interface of Reaxys to do manual searches on smaller compound sets whenever we need additional information.

The hardest part in building our compound library was establishing the pipeline to extract, transform and load data from multiple sources. Each source has its own structure and format, so data must be transformed to meet our needs for machine learning and statistical pattern recognition. By the time we incorporated Reaxys data, that architecture was already solid. Now our biggest issue is cleaning up source data, and this is no simple task. Things as simple as spelling mistakes can impair our pipeline. The data from Reaxys are the cleanest we have access to. The effort put into standardizing and normalizing Reaxys data makes them easy to map to our internal database structure.

What surprised me in the implementation of data integration from all our sources is that there is really minimal overlap in content. The data we are collecting is to large extent complementary. This means that the contribution of Reaxys is quite substantial and my overall impression is that Reaxys excels not only in number of compounds, but also in terms of the comprehensiveness of bioactivity and target information.

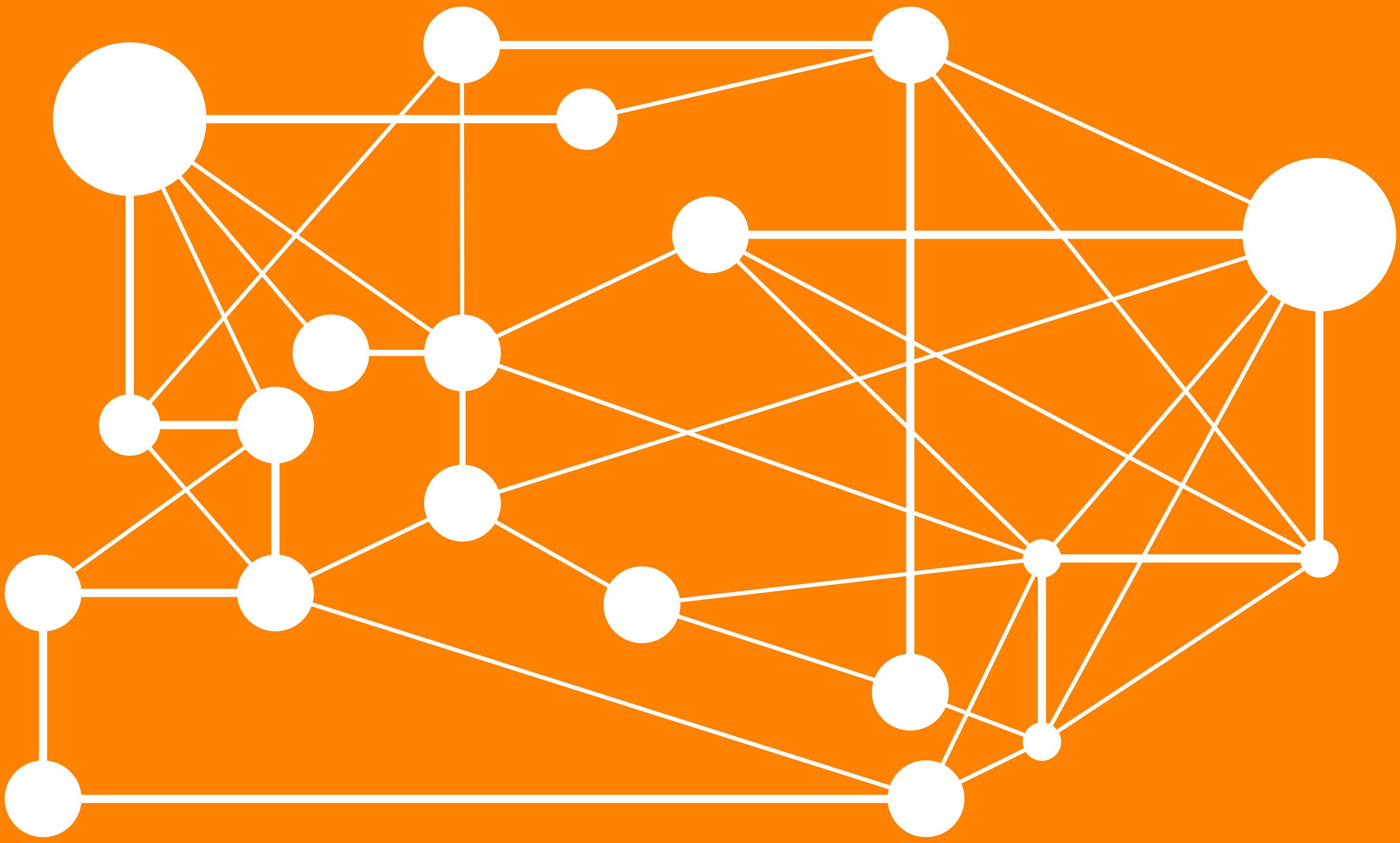
Thousands of players. Visualizations of the complex network models that e-Therapeutics explores to identify compounds with a desired chemical biology footprint.

E-THERAPEUTICS SEEMS TO BE AT THE FOREFRONT OF A PARADIGM SHIFT IN THE METHODOLOGIES USED TO IDENTIFY AND OPTIMIZE NEW DRUGS. WHAT TRENDS TO YOU ANTICIPATE FOR THE FUTURE?

There is a clear signal in the drug industry that network pharmacology is landing on fertile ground. We understand that compounds often affect more than one target and that future drug development must accommodate and even capitalize on that promiscuity. Phenotypic screening is also becoming more popular again. Independent of the idea of networks, the growing appreciation that no one gene can be responsible for complex phenotypes, and that it is necessary to measure multiple factors to truly assess the value of a compound as a drug, is heartening to me. It means that we are moving in the right direction!

But we are by no means at the end of the journey. There are a number of areas that need improvement and I anticipate that the future will bring these. For example, as inclusive as our models are, they are still based only on protein-protein interactions. A next step would be to include other molecules that play a role in the complex theatre of cellular function. Also, at this point we account for only binary effects—what happens when I knock out a protein in a given network—but we know that compounds can also activate proteins and that their effect can be graded, rather than all-or-none. We are working on addressing some of these challenges both internally and with external partners.

Another question that always plays in the back of my mind is that we have no concrete measure of whether our approach to drug discovery is lending benefits. This is simply because it takes on average 15 years for a new drug to emerge on the market and only then do we really know if we had a positive impact. Nevertheless, we have empirically proven that our network-based approach to drug discovery works: 10–20% of the candidates we send to phenotypic testing demonstrate the desired activity profile. And that is the reason why I go to work every day.



Reaxys

Reaxys helps customers drive successful early drug discovery by providing chemists the shortest path to relevant literature, patent information, valid compound properties, and experimental procedures.

LEARN MORE

To request information or a product demonstration, please contact us at elsevier.com/reaxys/contact-sales.

Visit elsevier.com/rd-solutions
or contact your nearest Elsevier office.

ASIA AND AUSTRALIA
Tel: + 65 6349 0222
Email: sginfo@elsevier.com

JAPAN
Tel: + 81 3 5561 5034
Email: jpinfo@elsevier.com

KOREA AND TAIWAN
Tel: +82 2 6714 3000
Email: krinfo.corp@elsevier.com

EUROPE, MIDDLE EAST AND AFRICA
Tel: +31 20 485 3767
Email: nlinfo@elsevier.com

NORTH AMERICA, CENTRAL AMERICA AND CANADA
Tel: +1 888 615 4500
Email: usinfo@elsevier.com

SOUTH AMERICA
Tel: +55 21 3970 9300
Email: brinfo@elsevier.com

February 2017