WHITE PAPER

cool clinical

Consortium for Clinical Science and AI

Computational Methods in Drug Discovery

September 2020

Felix Beacher, PhD

About Cool Clinical

Cool Clinical is a non-profit consortium of clinical and computational scientists. Cool Clinical publishes original articles and reports for businesses, the public sector, NGOs and the general public. Our goal is to advance the conversation on AI applications to clinical science.

To learn more about the research of Cool Clinical, visit www.coolclinical.com

Contents

Computational Methods in Drug Discovery

'Drug discovery' refers to the processes by which new drug molecules are discovered. 'Drug development' follows the drug discovery stage and refers to the process of bringing the new drug to the market, via the processes of preclinical research, regulatory filing, and clinical trials in humans.

Traditional Forward Pharmacology

The traditional method of drug discovery, known as 'forward pharmacology', involved screening chemical libraries of drug candidates, to identify candidate molecules with a possible therapeutic effect, as measured with cell or animal models. Once a therapeutic effect had been established, attempts were then made to infer the biological mechanisms and the drug target. The principle of forward pharmacology was greatly boosted by the development of high-throughput screening (HTS), which allowed the quick and automated screening of a very large number of compounds against a molecular target (screening up to a million compounds is typical in HTS).

Despite the advances made in the process of new drug development, the traditional method is time consuming and expensive. From target discovery to market entry the cost of new drug development is now between 1 and 2 billion USD and the process can take up to 17 years (Wouters et al, 2020; Leelananda and Lindert, 2016). Also, there are very high attrition rates in drug development, with around 90 to 95% of potential drugs abandoned at some point in development (Hay et al 2014; Arrowsmith et al, 2012).

Rational Drug Discovery

It remains that most drug discovery campaigns are performed on large compound libraries using HTS. However, pharmaceutical companies have, in recent years, increasingly invested in new and more efficient methods of drug discovery, which involve searching in a more targeted manner (Hillisch et al., 2004). This is known as 'rational drug discovery' or 'reverse pharmacology'. Fundamentally, this involves the opposite process as forward pharmacology: it begins with a hypothesis about a biological mechanism and drug target, and only considers compounds which, based on various calculations, have a reasonable chance of modulating the target.

The candidates are not only pre-selected on the basis of desired therapeutic effect, they also have to be 'druggable', e.g. have oral bioavailability, be chemically stable, and have tolerable levels of toxicity. Rules of thumb such as Lipinski's Rule of Five are often used to increase likelihood that the preselected molecules will be druggable. Rational drug discovery typically involves the screening of fewer candidate molecules and is therefore often perceived as more cost-efficient (although this has been contested, see for example Kotz 2012).

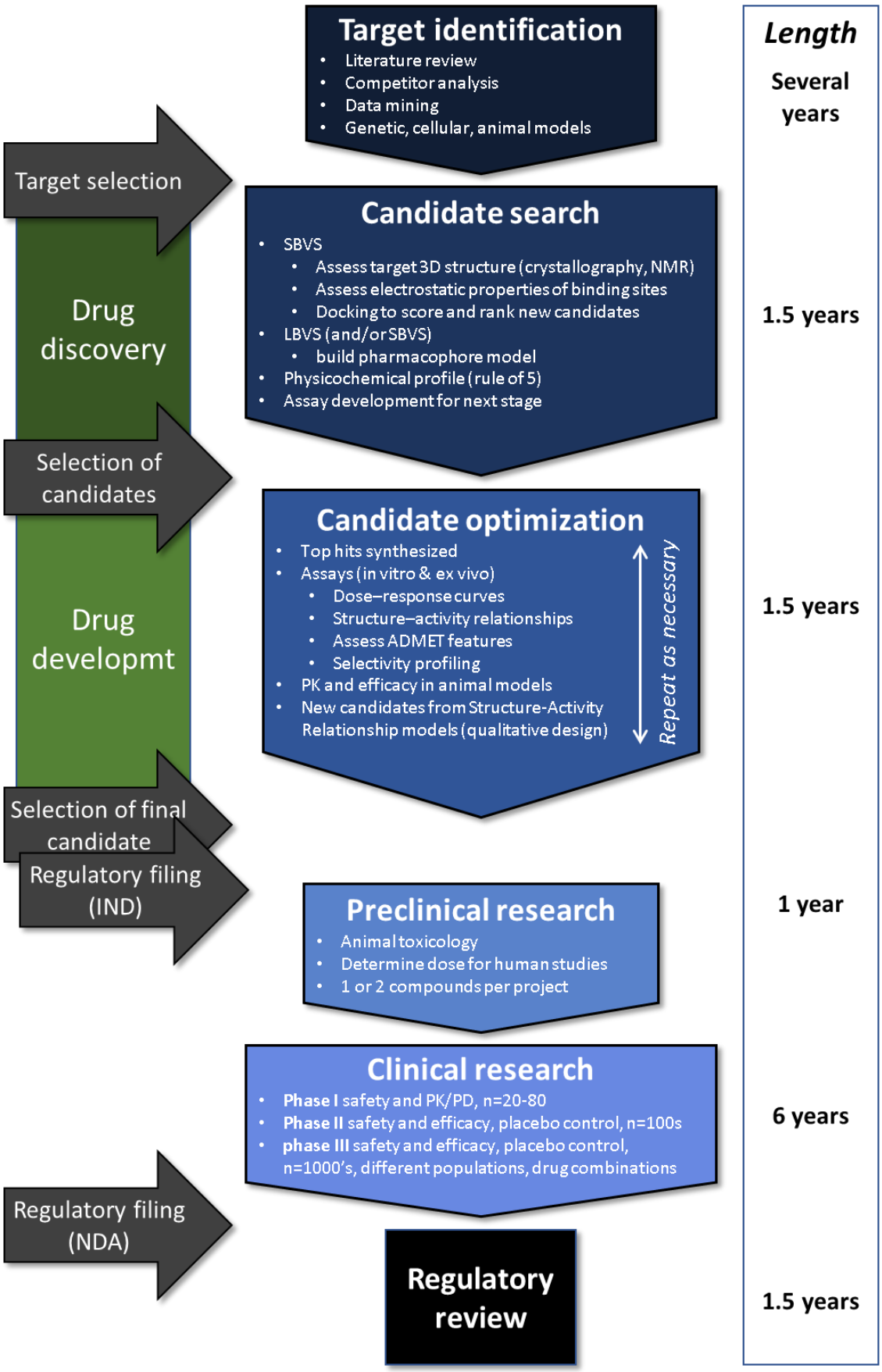

Figure 1: Summary of the drug development process

Virtual screening

The process of rational drug discovery has been enhanced by the techniques of 'virtual screening'. This refers to a computational search of small molecules to identify those most likely to bind to a drug target. Virtual screening, often used in conjunction with HTS, has become an important part of drug discovery and drug development. The first key advantage of virtual screening is that it can radically reduce the number of small molecules which are screened, thus saving time and reducing costs. Secondly, the virtual nature of the screening means that the molecules do not have to be physically synthesized for testing, further saving time and reducing costs.

There are two major types of virtual screening: Structure-Based Virtual Screening and Ligand-Based Virtual Screening.

Structure-Based Virtual Screening (SBVS)

Structure-based virtual screening (SBVS; also known as 'molecular docking-based virtual screening') aims to establish the 3D structure of the target, using a variety of techniques including X-ray crystallography, NMR spectroscopy and homology modeling (the construction of an atomic-resolution model of a protein based on its amino acid sequence). Once the 3D structure of the target has been established, cavities in the structure can be identified. Molecule databases can then be scanned for molecules which are complementary to the target cavities in terms of shape and charge ('docking'). This complementarity indicates that the new molecule should bind with the target structure in a stable and efficient way and could therefore form the basis of a viable new drug (Reynolds et al 2010). The 'hits' are then ranked on a scoring system based on their electrostatic and steric interactions with the target binding site. The highest scoring hits are taken forward for further development.

SBVS can not only save the time and costs involved in drug discovery, it can also identify compounds which have greater specificity in terms of the therapeutic effect, because it is focused on a known binding site or receptor conformation (Bruning et al 2010).

Perhaps the most prominent example of the success of SBVS is the development of amprenavir, a protease inhibitor used to treat HIV which was approved by the FDA in 1999 (Clark, 2006; Wlodawer and Vondrasek 1998).

SBVS typically takes data from databases such as Protein Data Bank (www.rcsb.org) and screening is performed using software such as AutoDock Vina (www.vina.scripps.edu) and SwissDock (www.swissdock.ch). A comprehensive list of such software is given in the Appendix.

Ligand-Based Virtual Screening (LBVS)

Sometimes the 3D structure of the target is not available and therefore SBVS is not viable. In this case a widely used method for drug discovery is Ligand-based virtual screening (LBVS; also known as 'pharmacophore-based virtual screening'). LBVS uses information about compounds known to activate a particular target and compounds known not to activate that target. This reference knowledge is used to build a 'pharmacophore model' which describes the set of common chemical features which underly the bioactivity. This model can then be used to screen new compounds, predicting those that will be active with respect to the same target.

LBVS typically takes data from protein–ligand interaction and binding affinity databases such as Drug2Gene (www.Drug2Gene.com) and BindingDB (www.BindingDB.com) and computes the pharmacophore model using software such as AutoDock Vina and SwissDock.

Current virtual screening techniques have limitations, including a low hit rate, a low enrichment factor (the degree of advantage over a random search), and a high rate of false positives (e.g. Ren et al 2011). Two approaches have emerged to address these limitations: consensus virtual screening (CVS) and hybrid virtual screening.

Consensus Virtual Screening (CVS)

Consensus virtual screening (CVS) aims to improve the accuracy of techniques such as SBVS and LVVS, by combining the results of different algorithms (Kukol 2011) implemented by software packages such as AutoDock Vina and SwissDock. Alternatively, software packages such as VoteDock have been specifically developed to implement CVS.

Hybrid Virtual Screening (HVS)

Hybrid virtual screening (HVS) uses a combination of data from both structure-based and ligand-based analyses, for example by using the combined dataset to train machine learning models (Mize et al 2011; Sanam et al 2010). This is reported to improve the performance of virtual screening (Di-Wu et al 2012).

Limitations of Virtual Screening

As noted above, virtual screening still requires further development and refinement. It remains often difficult to accurately predict the binding positions of target proteins, due to the complexity of ligand-receptor interactions. This limitation of virtual screening has contributed to some real-world failures in drug development. For example, RPX00023 was developed using SBVS as an antidepressant with agonistic action at the 5-HT1A receptor, but it actually inhibited the receptor (De Paulis 2007).

Another fundamental limitation of virtual screening is that while SBVS is reasonably successful in predicting binding affinity, and LBVS is reasonably successful in predicting active compounds, there are other properties important for a viable new drug (e.g. bioavailability, metabolic half-life, side effects) which must also be optimized. Thus, while SBVS is an important tool for drug discovery it must be supplemented with other types of analysis.

Finally, virtual drug screening, and rational drug design in general, is only as good as the understanding of the molecular processes underlying the disease. It is very often the case that disease mechanisms are poorly or incompletely understood. In this case rational drug design may overly restrict the search for new therapeutic molecules.

Quantitative Structure-Activity Relationship (QSAR) models

Quantitative structure-activity relationship (QSAR) modelling takes a set of molecules whose structure and biological activity is known, and constructs a model of how these relate to one another. QSAR uses the basic assumption that similar molecular structures have similar biological activities. Classification QSAR methods characterize molecules in terms of categories such as 'active' and 'inactive'. Regression QSAR methods go beyond this and predict the *level* of biological activity.

In the context of drug discovery and development, QSAR models can predict which chemical groups play an important role in evoking a target effect in the organism and can therefore be used. They can help discover or design new candidate molecules with greater potency and can predict toxicity of new compounds.

Machine Learning Methods for Drug Discovery

The application of ML has been touted as a coming revolution in drug discovery and development (Mak and Pichika, 2019). The adoption of ML in drug discovery and development is still in its infancy, however, there have already been tangible signs of its potential. Notably, Pfizer formed a collaboration with IBM in 2016 on ML approaches to drug discovery for immuno-oncology.

ML has been applied to many areas of drug discovery and development. Typically, ML exploits relationships between various features of ultimate interest (such as biological activity, efficacy and/or toxicity) and the physical features of the molecules tested (such as their chemical structure outputs of molecular docking analysis or pharmacophore models). ML can potentially be applied in various ways to support drug discovery and development. Examples include:

- Identifying new drug targets
- Predicting which molecules will activate a drug target
- Predicting the degree to which molecules will activate a drug target
- Predicting efficacy of new molecules
- Predicting the ADMET features of new molecules
- Predicting adverse effects of new molecules
- Predicting drug-drug interactions

These are all examples of 'supervised' ML, where the models are trained on data with known labels, e.g. 'active' or 'inactive'. ML also includes 'unsupervised' learning, where models are used to find clusters of unlabeled data. Unsupervised ML can also potentially support various aspects of clinical science related to new drug discovery and development, for example the discovery of disease subtypes, which may be treated with different molecules. This is an example of how ML can potentially support the development of 'personalized medicine'.

The success of any ML approach for drug discovery or drug development, like any ML problem, depends on the factors such as the separability of the data, the size of the dataset and the degree to which the dataset is balanced in terms of its categories. However, properly developed ML models can handle huge libraries of compounds, and are able to predict a good number of active compounds with few false positives.

References

Arrowsmith (2012). A decade of change. Nat. Rev. Drug Discov. 11, 17–18.

Bruning et al 2010. Coupling of receptor conformation and ligand orientation determine graded activity Nature Chemical Biology, 6, 837-843.

Clark, 2006. What has computer-aided molecular design ever done for drug discovery? Expert Opin. Drug Discov.;1:103–110.

De Paulis 2007. Drug evaluation: Prx-00023, a selective 5-ht1a receptor agonist for depression. Curr. Opin. Investig. Drugs. 8:78–86.

Di-wu et al 2012. Identification of CK2 inhibitors with new scaffolds by a hybrid virtual screening approach based on Bayesian model; pharmacophore hypothesis and molecular docking. J Mol Graph Model. 36:42-7.

Hay et al 2014. Clinical development success rates for investigational drugs. Nature Biotechnology 32, 40-51

Hillisch et al, 2004. Utility of homology models in the drug discovery process. Drug Discov. Today 9, 659–669.

Kukol 2011 Consensus virtual screening approaches to predict protein ligands. European Journal of Medicinal Chemistry, Volume 46, Issue 9, Pages 4661-4664

Kotz 2012. Phenotypic screening, take two. Science-Business eXchange. 5 (15): 380.

Leelananda and Lindert, 2016. Computational methods in drug discovery. Beilstein J. Org. Chem. 12, 2694–2718.

Maia et al 2020. Structure-Based Virtual Screening: From Classical to Artificial Intelligence. Front. Chem. 28 April.

Mak and Pichika, 2019. Artificial intelligence in drug development: Present status and future prospects. Drug Discov. Today.;24:773–780.

Mize et al 2011. Ligand-based autotaxin pharmacophore models reflect structure-based docking results J. Mol. Graphics Modell., 31, pp. 76-86

Ren et al 2011. Discovery of novel Pim-1 kinase inhibitors by a hierarchical multistage virtual screening approach based on SVM model pharmacophore, and molecular docking. J. Chem. Inf. Model., 51, pp. 1364-1375.

Reynolds et al eds. 2010. Drug Design: Structure- and Ligand-Based Approaches (1 ed.). Cambridge, UK: Cambridge University Press.

Sanam et al 2010. Combined pharmacophore and structure-guided studies to identify diverse HSP90 inhibitors. J. Mol. Graphics Modell., 28, pp. 472-477

Wlodawer and Vondrasek 1998. Inhibitors of HIV-1 protease: A major success of structure-assisted drug design. Annu Rev. Biophys Biomol. Struct;27:249–284.

Appendices

Appendix 1: Software for Structure-Based Virtual Screening (Credit: Maia et al 2020)

Software	License	Platform	Protein flexibility	Docking algorithm	Scoring function
AutoDock4 (Morris et al., 2009)	Free for academic use	Windows, Linux and Mac	Yes	Genetic algorithm Simulated annealing	Hybrid (Force-field and empirical)
Autodock Vina (Trott and Olson, 2009)	Open- source	Windows, Linux and Mac	Yes	Genetic algorithm Simulated annealing Local search Particle swarm optimization	Hybrid (Empirical and knowledge-based)
DOCK 6 (Allen et al., 2015)	Free for academic use	Windows, Linux and Mac	Yes	Shape fitting (sphere sets) Lowest energy binding	Force-Field Empirical
SwissDock/EADock DSS (Grosdidier et al., 2011)	Free for academic use	Web	No	Stochastic (Tabu search based) Local search Combination of broad and local search of the conformational space	Force-field
eHiTS (Zsoldos et al., 2007)	Freeware for academic use	Unix	No	Exhaustive search	Hybrid (Empirical and knowledge-based)
FITTED (Corbeil et al., 2007, 2008)	Commercial	Linux, Windows and Mac	Yes	Genetic algorithm	Force-field
FlexX (Rarey et al., 1996)	Commercial	Windows and Linux	No	Incremental construction	Empirical
FLIPDock (Zhao and Sanner, 2007)	Freeware for academic Use	Linux e Windows	Yes	Genetic algorithm	Force-field
Fred (McGann, 2011)	Free for academic use	Windows, Linux and Mac	No	Exhaustive search algorithm	Hybrid
GalaxyDock2 (Shin et al., 2013)	Freeware	Linux	Yes	Conformational analysis Genetic algorithm	Force-field
GeauxDock (Fang et al., 2016)	Open-source	Linux	Yes	Monte Carlo	Hybrid (Empirical and knowledge-based)
GlamDock (Tietze and Apostolakis, 2007)	Freeware	Windows, Linux and Mac	No	Monte Carlo Simulated annealing Local search Conformational analysis	Empirical
Glide (Friesner et al., 2004)	Commercial	Windows, Linux	Yes	Conformational analysis Monte Carlo sampling	Empirical
GOLD (Verdonk et al., 2003)	Commercial	Linux and Windows	Yes	Genetic algorithm	Force-field
ICM (Abagyan et al., 1994)	Commercial	Windows, Linux and Mac	Yes	Monte Carlo minimization	Force-field
iGEMDOCK/GEMDOCK (Hsu et al., 2011)	Freeware	Windows and Linux	Yes	Genetic algorithm	Empirical
LigandFit (Montes et al., 2007)	Commercial	Linux	Yes	Monte Carlo	Force-field
LigDockCSA (Shin et al., 2011)	–	–	Yes	Conformational space annealing Global optimization	Hybrid (Empirical and Force-field)
MOE (Vilar et al., 2008)	Commercial	Windows, Linux and Mac	Yes	Conformational analysis	Empirical, Force-field
ParaDockS (Meier et al., 2010)	Freeware	Linux	No	Genetic algorithm	Hybrid (Knowledge-based and empirical)
rDOCK (Ruiz-Carmona et al., 2014)	Open-source	Linux	Yes	Genetic algorithm, Monte CarloSimplex minimization	Hybrid (Empirical and force-field)
SLIDE (Schnecke and Kuhn, 2000)	Free for academic use	Linux	Yes	Conformational analysis	Empirical
Surflex (Spitzer and Jain, 2012)	Commercial	Windows, Linux and Mac	Yes	Incremental xonstruction	Empirical
Sybyl-X (Certara, 2016)	Commercial	Windows	Yes	Incremental construction	Force field
vLifeDock (Chopade, 2015)	Commercial	Windows, Linux and Mac	Yes	Genetic algorithm	Empirical

Appendix 2: Software Resources for Calculating Molecular Descriptors

Software is needed to calculate molecular descriptors. Once this is done, QSAR can be done using any statistical software.

qsartoolbox.org
www.vcclab.org/
www.chemosophia.com
http://padel.nus.edu.sg/software/padeldescriptor/
http://jcompoundmapper.sourceforge.net/
https://sourceforge.net/projects/pydpicao/
http://www.way2drug.com/Projects.php
https://www.click2drug.org/
http://crdd.osdd.net/qsar.php
https://www.cgl.ucsf.edu/chimera/
http://lqta.iqm.unicamp.br/portugues/siteLQTA/LQTAgrid.html
https://discover.3ds.com/discovery-studio-visualizer-download
http://www.vlifesciences.com/products/VLifeMDS/VLifeQSAR.php
https://www.schrodinger.com/autoqsar
http://open3dqsar.sourceforge.net/

Python toolkits

https://cinfony.github.io/
https://pypi.org/project/pydpi/

About the author

Felix Beacher | felix@coolclinical.com

Felix Beacher is the founder of Cool Clinical. Felix has a PhD in neuroscience and has worked in drug development and various therapy areas, including neurodegeneration and cancer.

Legal disclaimer

This publication has been written in general terms and is not intended to be relied on to cover specific situations. Any application of the information given in this publication will depend upon the particular circumstances involved, As such, we recommend that professional advice is sought before acting or refraining from acting on any of the contents of this publication. This publication and the information contained herein is provided “as is”. Cool Clinical makes no express or implied warranties that this publication is error-free or meet any particular criterion of performance or quality.

Cool Clinical accepts no duty of care or liability for any loss to any person acting or refraining from action as a result of any material in this publication.

