Cite this: *Med. Chem. Commun.*,
2018, 9, 606

Approaches to target tractability assessment – a practical perspective†

Kristin K. Brown,^a Michael M. Hann, ^{*b} Ami S. Lakdawala,^c Rita Santos, ^d
Pamela J. Thomas^e and Kieran Todd^e

Received 15th December 2017,
Accepted 13th February 2018

DOI: 10.1039/c7md00633k

rsc.li/medchemcomm

The assessment of the suitability of novel targets to intervention by different modalities, e.g. small molecules or antibodies, is increasingly seen as important in helping to select the most progressable targets at the outset of a drug discovery project. This perspective considers differing aspects of tractability and how it can be assessed using *in silico* and experimental approaches. We also share some of our experiences in using these approaches.

Introduction

When a new protein target is being considered for a potential drug discovery programme, getting an early insight into whether the protein has a binding site that can be exploited for small molecule binding or, alternatively, an accessible epitope for antibody based therapy is increasingly seen as important. With increasing use of genetic, transcriptional and knock out technologies many biologically “validated” targets can be proposed. Which of these is suitable for drug discovery is critically dependent on whether the protein is amenable to interactions with drug like molecules. Historically, especially for small molecule drug discovery, an interesting target may trigger a big screening effort to find starting points for optimisation, however this process can be expensive in both time and resources. This is especially an issue if many targets are available from omics and knock out approaches¹ when prioritization as to suitability for different modalities of intervention or none at all is of critical importance to avoid wasted efforts. Looking for alternative targets in the same pathway should also be considered as an appropriate response for poorly tractable targets.

In this perspective, we will describe our preferred nomenclature for describing different aspects of target quality, target tractability and overall druggability. We will discuss our experiences of using both *in silico* automated high level genome wide evaluation of targets and more individual deep dive target assessments. We will also briefly overview the experimental methods that we and others have utilized for target tractability where such an approach is warranted after initial *in silico* analysis and before committing a target to comprehensive hit identification and downstream drug discovery processes.

Definitions

Since the druggable genome concept was first introduced by Hopkins and Groom in 2002,² terms like druggability, target tractability, ligandability and target quality have been introduced. Sometimes these terms are conflated to mean much the same thing and sometimes they are considered to have distinctive meanings.³ We prefer the latter usage as we believe this helps to delineate separate activities that are critical contributors to target selection (Fig. 1).

Our working definitions

Target quality. Level of confidence that modulation of a target will translate to disease modifying or symptom alleviating effects in humans based on evidence from genetics, expression levels, KO studies, and pathway understanding amongst others.

Target tractability (a.k.a. ligandability). The likelihood of identifying a modulator that interacts effectively with the target/domain (or pathway).

Druggability. The ability of a protein to bind a drug-like modulator (e.g. small molecule or antibody) with a therapeutically useful level of affinity, efficacy and safety. Combines

^a Computational and Modelling Sciences, Platform Technology and Sciences, GlaxoSmithKline, 1250 S. Collegeville Road, Collegeville, Pennsylvania 19426, USA

^b NCE Molecular Discovery, Platform Technology and Sciences, GlaxoSmithKline Medicines Research Centre, Gunnels Wood Road, Stevenage, Hertfordshire, SG1 2NY, UK. E-mail: mike.m.hann@gsk.com

^c In vitro/In vivo Translation Sciences, Platform Technology and Sciences, GlaxoSmithKline, 1250 S. Collegeville Road, Collegeville, Pennsylvania 19426, USA

^d Target Sciences Computational Biology, GlaxoSmithKline Medicines Research Centre, Gunnels Wood Road, Stevenage, Hertfordshire, SG1 2NY, UK

^e Computational and Modelling Sciences, Platform Technology and Sciences, GlaxoSmithKline Medicines Research Centre, Gunnels Wood Road, Stevenage, Hertfordshire, SG1 2NY, UK

† All authors contributed equally to this perspective.

Fig. 1 A schematic representation of the biology (vertical axis) and chemistry (horizontal axis) contributions to the overall concept of druggability.

target tractability and target quality and can be used to help understand the potential of a target to move through the drug discovery process, particularly at the early stages.

Genome wide assessment of tractability/pipelines, data mining & integration

A high throughput *in silico* pipeline is appropriate to process large amounts of data when trying to identify new targets, and obtain a quick initial estimate of their tractability. To address this, within our organization we have developed a knowledge based system to estimate the likelihood that a target will bind a small molecule or an antibody.

In essence, we have incorporated data from internal, commercial and public resources, such as PharmaProjects,⁴ Uniprot,⁵ HPA,⁶ PDBe,⁷ DrugEBility,⁸ ChEMBL⁹ and SureChEMBL¹⁰ to create a system of hierarchical qualitative buckets of tractability, one for small molecules (SM) and one for antibodies (mAB). Complementary information was also retrieved from Pfam,¹¹ InterPro,¹² Complex Portal,¹³ DrugBank,¹⁴ GO¹⁵ and BioModels,¹⁶ to help further assess targets. Depending on the evidence available, these systems rank and assign human genes to buckets, which represent different levels of tractability, ranging from high confidence to uncertain tractability (Fig. 2). For example, a target that has been co-crystallised with a small molecule, or a target for which there is published bioactivity data is assigned to a higher confidence tractability bucket for SM than one for which there is only a binding prediction. Similarly, a target for which there is high confidence experimental evidence that the target is localised in the plasma membrane, and thus accessible to an antibody, ranks higher in the mAB trac-

Fig. 2 The gene buckets of tractability for small molecules (SM). The main differences between the tractability buckets for SM and antibodies (mAB) are highlighted in the boxes on the right. Human genome defined as all protein coding genes, and extracted from NCBI in November 2016. For more detailed rules see Appendix.

tability system, than a target that only has a GO annotation suggesting it may be localised in the membrane.

The idea behind this knowledge based system is, however, not entirely novel.¹⁷ Campbell *et al.*, described a very similar approach developed at Pfizer *circa* 2010 for SM. In our work we have explored the added value of newer resources as they emerged in the public domain, such as patents data from SureChEMBL, as well as developing a parallel system to assess the tractability for antibodies, as this modality has become more prominent in drug discovery. In addition, we are working with the Open Targets consortium¹⁸ in this arena, and aim to make available non-proprietary versions of these pipelines, so that the collaboration can build on them, improve them, and ultimately release them into the public domain.

For example, one area that might bear improvement is the simplistic approach we took to score and rank targets. In practice, we're using a matrix of presence/absence of certain traits with different degrees of confidence to say that a target is more likely to be tractable than another, but we have not explored the power of an assessment where a combination of traits is considered. For example, should a target for which there is published bioactivity data and evidence of co-crystallisation with a small molecule, rank higher than one for which there is only evidence of published bioactivity data, or for which there is only poor evidence that a small molecule is in a pre-clinical phase? Also, we have applied a generalised activity cut-off on published bioactivity data to discriminate between binders and non-binders (if pXC50 is defined then a compound is classified as binder if pXC50 $\geq$ 5.5). But would an activity cut off per protein family be more robust and accurate?

Many other challenges could be listed when dealing with the challenges of high throughput and thus automated data mining and data integration. Therefore, a deeper dive into the data is always crucial for the final judgement. It is always

important to understand the type and quality of data behind the evidence of 'yes, this target is likely to be tractable'.

Furthermore, the pipeline approach described above is looking at a single protein sequence and all its known biological data. An additional approach to expanding the scope of knowledge around a particular protein is to look at other proteins with homologous domains. When comparing the protein sequences of known drug targets, many cluster within homologous families; GPCRs (12% of known drug targets), kinases (10%), and nuclear hormone receptors (3%) to name a few.¹⁹ Based on these observations, it has been hypothesized that any protein that is homologous to a known drug target has the potential to be "druggable".² Over the years, this trend has held up for most protein families, within the caveat of understanding how the particular protein in question needs to be regulated in order to be efficacious. Branching out to look across a protein family may enrich the knowledge around a protein by identifying possible starting points for hit discovery, selectivity concerns, and the success rate of the protein family in achieving drugs on the market.

There is incredible complexity in predicting if a target is tractable, and more importantly, ultimately druggable. A number of researchers^{20–26} have used machine learning algorithms to predict the tractability of a protein utilizing knowledge of known drug targets and various biophysical and biochemical descriptors. Some approaches use simple sequence properties, such as amino acid and dipeptide content and/or frequency. Since these algorithms are not dependent upon structural information being available, predictions can be made across the complete human proteome. Others focus on using structural information to predict binding pockets and efficiently identify the ones most suitable in terms of tractability. These methods are limited to the 29% of the human proteome which have 3D structural representation, although this can be expanded to 50% if one builds models based on mammalian templates.²⁷ Many of the above mentioned algorithms are reporting 63–93% accuracy of predicting known drug targets. While this is encouraging, much uncertainty around what makes a target tractable persists.

Structure based tractability assessment

The pipelines and sequence based methods described above enable the tractability assessment of large pools of potential targets. However, as already alluded to, when the choice of targets has been narrowed down, a "deep dive" assessment of each target becomes feasible and desirable. Such deeper assessments involve gaining an understanding of factors such as:

1. The location and nature of sites able to bind small molecules or antibodies
2. The function of the target and its integration into pathways
3. Endogenous regulation mechanisms
4. Spatial and temporal location
5. Tractability of homologous targets
6. Properties of known modulators

7. Status of progression of known modulators

The "deep dive" assessment methods used for antibodies are described in more detail in a later section of this paper. For small molecules, several structure based assessment methods, which address point 1 above, have emerged which can be used to analyse both experimentally determined structures and homology models.

A protein structure can be used to identify pockets into which a small molecule can be bound and to analyse the properties of that pocket to determine the likelihood that a drug-like small molecule can be bound. Such pockets can be obvious parts of a structure *e.g.* an active site, an allosteric site or a site of interaction with another protein. However, the identification of potential new allosteric sites only seen in the presence of particular small molecules, which are sometimes called cryptic sites, are only seen in the presence of particular small molecules, has also become an area of increasing research effort.²⁸

The presence of a pocket in a protein is a pre-requisite for small molecule binding but in order to distinguish those pockets that can bind drug-like molecules from the pool that bind any type of small molecule the properties of a pocket must be analysed. Key properties are size, shape (concavity) and the nature and mix of the protein atoms that are exposed to interactions with a bound ligand in terms of lipophilicity and hydrophilicity. A number of methods have been developed for the analysis of the ligandability of protein pockets. An overview of these can be found in Hussein *et al.*³

Analysis of water molecules and water networks in binding sites is now also beginning to be considered in the assessment of tractability *e.g.* Mason *et al.*^{29,30} A number of methods have been developed to assess the thermodynamic properties of waters which might contribute to ligand binding³¹ and which, therefore, might also contribute to the assessment of the ligandability of a pocket. In any case, pocket detection and analysis should account for the positions of well conserved or highly hydrogen bound waters.

Comparison of real world tractability and predictions

To understand the link between structure based tractability calculations and real world tractability, we conducted a retrospective analysis of tractability for 70 diverse targets for which structural information was available and for which we also had evidence from screening (usually HTS) of real world tractability or not. The set is composed of over ten major target classes of receptors, enzymes (including kinases) and various signalling systems (including bromodomains). For each of these targets, screening outcomes and expert scientist opinion on the tractability resulted in the assignment of a tractability classification of high, medium or low. For each target, calculations were carried out using SiteFinder,³² SiteMap,^{33,34} Fpocket³⁵ and mixed probe molecular dynamics.³⁶ The ability of the output from these programs to correctly reproduce the classification assigned to each target was

	Predicted Tractability	
Real world tractability	Low	High
Low	15	8
High	7	40

Fig. 3 Confusion matrix for two class model of tractability prediction based on SiteMap DScore values obtained for 70 unique proteins. Accuracy = 0.79, CI = (0.67, 0.87).

assessed. The results are illustrated in the confusion matrix in Fig. 3. The results of the computational methods showed that, for all methods, a clear separation between high and low tractability targets is achieved. A significant correlation between the DScore results from SiteMap and a high/low classification scheme was found. With a 2-class model, 'low' vs. 'moderate-high' tractability, SiteMap DScore demonstrated a 85% specificity in identifying moderate-high tractability targets, with an overall 80% accuracy rate. These results are obtained with a straightforward cut-off SiteMap DScore, where DScore ≤ 0.85 describes a 'low' tractability target.

We made a number of observations on the use of structural information in tractability assessment. As part of our experiment we looked at the effects on the SiteMap DScore of using apo v holo co-ordinates (to account for changes induced by ligand binding), structure resolution, the use of multiple structures for a single target *versus* one and the effect of rotating the protein co-ordinates. The effect on the values obtained for the SiteMap DScore were minimal leading to a consistent prediction. In some cases, the calculations involved can be sensitive to the absolute values of atomic co-ordinates so that resolution, disorder and translation/rotation of co-ordinates can influence the outcomes of the calculations and should be checked for robustness.

Any influence of variation in coordinates is likely to be magnified where a homology model of a protein is used instead of experimentally determined structures. The expectation would be that this would be exacerbated as the percentage identity between a target and the structure on which it is modelled decreases. In our experiments using two targets; one high tractability and one of low tractability we found that the predictions were highly variable when homology models were used.

Individual target assessment of biopharmaceuticals

A subset of tractable targets that can be predicted from sequence are those that are amenable to an antibody or antibody derivative drug.

Whilst a number of new technologies such as ImmTACs (immune mobilising monoclonal T-cell receptors against cancer)³⁷ make all proteins potential biopharmaceutical targets, the vast majority of biopharmaceutical approaches require the target to be cell surface exposed or secreted. A 'Potential Biopharmable Genome' therefore requires the identification

of all proteins and biologically active peptides that are annotated as, or predicted to be, completely or partially extracellular and therefore secreted or spanning the plasma membrane.

In contrast to small molecules a number of relevant predictions can be based on sequence information, in particular the presence of signal peptides and transmembrane regions.

Annotation about protein function, subcellular location and protein family members can also be used to help verify the predictions or identify potential targets that would otherwise have been missed.

Once identified as biopharmable it is then useful to know the portion of the protein that is amenable to the drug. In the case of biopharmable targets this requires identifying the best epitopes. These are the regions that are extracellular, surface exposed and have the best biophysical, biological and cross reactivity properties. Again, the sequence can be used to predict extracellular regions, hydrophobicity and antigenicity. Where available characteristics such as protein surface residues, active sites and binding sites can be derived from 3-dimensional structure information, (see previous section on SM) and annotated experimental evidence (for example antibody binding, SNPS and tissue expression) can all be used to identify the best sites within the available regions of the target.

Experimental approaches

While *in silico* tractability assessments are the major thrust of this perspective, it should be remembered that the more novel the target is, the less likely is the quality of the prediction. In any case the ultimate test of the tractability of a target is through physical experiments and validated measurement approaches.

For the small molecule modality of intervention, the core purpose of experimentally based ligandability assessments is to get an idea, at the earliest possible stage, as to whether the target(s) in question can be effectively modulated by a chemical entity and with the minimum of commitment of resources. The context of the question of target tractability is also important as the depth of experimental analysis needed will be situation dependent. Thus, if presented with a number of possible targets do you need to know the best and least tractable for prioritization purposes, or is a more in depth analysis needed for one or two targets? Such questions will determine the resource (time and money) that should be deployed to do experimental assessments.

One key consideration is how many and what type of compounds need to be screened to get the rapid insights needed. Label free methods to detect binding are usually easier and quicker to set up than an assay which measures functional inhibition. However not all ligands that bind will have a functional effect and so subsequent proof of effective target engagement through functional assays may be needed as follow up. The limits of what can reasonably be done as target tractability assessment rather than lead generation may however preclude this at this stage. High throughput screening (HTS)

assays, which will usually be of this functional assay type, can require many months of set up before screening the wide diversity of compounds in an HTS deck, typically 1–2 million compounds. As such HTS is not normally considered a method for rapid target tractability assessment. By contrast either direct binding biophysical assays (*e.g.* SPR, NMR, MS, X-ray) in conjunction with a fragment screen of <5k compounds or an affinity selection from a DNA encoded library (DEL) of billions of compounds are more rapid and thus appropriate. The power of fragment based approaches is that they can cover large chemical diversity in a small (*ca.* 1–5k) number of molecules, although the resulting binding will be intrinsically weak.⁴⁷ The reason why a library of only a few thousand fragments can be effective is that fragments are intrinsically less complex molecules and thus are easier to fit to a pocket as they do not have to match so many features, as is the case for more complex molecules.⁴¹ The concept of using fragment screening for ligandability assessment was first proposed by Hajduk and colleagues at Abbott using NMR methods³⁸ and was later exemplified in a broader context by Breeze and colleagues at AZ³⁹ and by Hubbard and colleagues at Vernalis.⁴⁰ However, there is a requirement for very sensitive biophysical assays to detect the intrinsically weak binding. If evidence of functional binding is required, then the potency may not be sufficient without some chemical optimisation which will normally require insights from an X-ray structure. Availability of this information and the speed of the optimisation may be considered a weakness of the fragments approach for ligandability assessment.

Other methods that have been exploited for ligandability assessments include protein NMR observed measurements^{44,45} and high throughput thermal shift analysis.⁴⁶ Recent progress in using covalently trapped electrophilic fragments⁴⁸ and photoactivatable “nucleophilic” fragments⁴⁹ open up new ways to survey targets in a cellular context enabling the concept of the exploration of the tractability of targets *in situ*.

By contrast the DEL approach exposes the immobilised protein to potentially billions of larger and more complex molecules. Those with affinity will be retained by the protein and can be subsequently decoded from the associated DNA barcode to identify potential binders. This can be a very rapid process, although again subsequent chemistry will be needed, initially to reproduce the ligand off DNA to confirm binding and then subsequently in functional assays. Despite the fact that the chemical space covered by DEL libraries can be considered to be biased by available aqueous chemistry for core structures, their use is becoming a powerful technique for assessing single and multiple targets for tractability.⁴²

Another approach which is gaining in applicability for tractability assessment is affinity selection mass spectrometry (ASMS) which is again an affinity selection process but relies on MS to directly identify retained compounds rather than using DNA encoding to identify the structural formula of the ligand. This enables screening of a wider range of compounds than can be made with DELs and thus allows access to the

types of diversity of compounds that are typically used for HTS. One limitation (and possibly also strength in that it leads to low false positives) is that the method will only detect compounds with a slow off-rate as compounds which are fast off will immediately be absorbed by the matrix and thus not be detected as passing through the column with the protein.⁴³

Understanding tractability from the perspective of either small molecules or antibodies as potential ligands is being greatly enhanced by continuing advances in structural elucidation techniques. Current improvements with methods such as cryo-electron microscopy⁵⁰ and free electron laser (X-FEL) based crystallography⁵¹ will help continue to expand the number of targets where we have insights into binding pockets and interacting surfaces. Other rapidly improving methods that can give insights for tractability studies include the use of hydrogen–deuterium eXchange (HDX)⁵² and other mass spectrometry dependent methods for chemoproteomic studies.⁵³

Consideration of specificity in tractability assessments

Returning to the theme of earlier sections, it is important to consider how similar a target's paralogs and homologs are to help assess drug cross-reactivity risk and the likelihood of binding to equivalent model organism proteins. This is true whether working from an *in silico* or experimental point of view.

We often refer to a target as a protein and this usually refers to a canonical representation of a gene product, however most genes code a number of isoforms (including splice variants, differential processing, *etc.*) and each isoform can potentially have a different function and/or location. It is therefore important to know which isoforms may be of interest, can be targeted and whether it is possible to differentiate between them.

In addition to the presence of isoforms, the tractability of any particular protein can change over its lifetime due to post-translational modifications (such as phosphorylation, glycosylation, acetylation and lipidation) which can act like switches for their function. Such changes can, for the same protein, vary with cell type, cell status (*e.g.* a cell reacting to an external stimulus compared to a resting state) and with the age of a protein. These different states of a protein may have different physiological roles and not all may be the desired target of any therapy. Thus, the effect of post-translational modification on physiological function, location and/or disease relevance will need to be taken into consideration.

Protein structures are often determined using recombinant proteins which may be truncated or mutated relative to their wild type sequences to facilitate structure determination or assay set up. This manipulation of a protein may affect the presence or structure of potential ligand binding pockets. Furthermore, in a physiological context, proteins are exposed to other large and small molecules in a spatio-temporal sense that can influence their structure and function. Therefore, an

experimentally determined structure may not always represent the physiological state or states accessible to a protein and this can result in false positive or negative predictions of structure based tractability. Furthermore, the quality of a structure can also influence tractability assessments, for example, low resolution and/or disorder can result in uncertainties in atomic positions which may not be taken in to account in structure based methods.

Conclusion

There is a great deal of interest in target tractability assessment from both industry and academia as a means for improving the selection of targets for the drug discovery process (EMBL-EBI, tractability workshop, May 2017 (ref. 54). The assessment of target tractability is still a challenge at all levels of assessment, whether it be in high throughput pipelines or deep assessments by *in silico* or experimental methods. However, by implementing and continuously evaluating a range of tractability assessment methods we believe we can make better choices of targets and modalities now and gather information to improve the predictions for the future.

One of the key gaps in the area is the availability of data to benchmark new and existing assessment methods, as well as the lack of common terminology, and standards to report the data. Improvements in *in silico* methods could be made if more and larger collections of data linking targets and their modulators to *in vitro* and *in vivo* screening outcomes were available to the whole research community. We will be working through the Open Targets consortium to help alleviate this situation. Further development of methods that allow the assessment of targets for which experimentally determined structural information is not available or methods for detecting cryptic pockets also have the potential to have significant impact on target choices.

Where possible, carrying out expedient and informative experiments to assess tractability will increasingly facilitate a data driven decision early in the life time of a program. Where the preferred target is not tractable we should have the confidence to look at the pathways associated with the target, develop phenotypic⁵⁵ as opposed to target based screens or be prepared to embrace newer modalities such as PROTACs, using a binding ligand that does not need to be a functional inhibitor *per se*,⁵⁶ or cell and gene therapy.⁵⁷ We are beginning to formulate rules to assess these emerging modalities, as they develop. It is fair to say that now more than at any time in the past we should be optimistic about the ability of tractability assessments to have a real impact on the drug discovery process.

Appendix

Detailed rules for SM bucket assignments

Bucket 1 (ChEMBL)	Targets with approved SM drugs (phase4)
-------------------	---

Bucket 2 (ChEMBL)	Targets with SM in $\geq$ phase2
Bucket 3 (ChEMBL)	Pre-clinical targets with SM
Bucket 4 (PDB)	Targets with crystal structures with ligands
Bucket 5 (DrugEBility)	Targets with Ensembl score ≥ 0.7
Bucket 5 (DrugEBility)	Targets with $0 < \text{Ensembl score} < 0.7$
Bucket 7 (ChEMBL)	Targets with ligands (PFI ≤ 7 , SMART hits ≤ 2 , scaffolds ≥ 2)
Bucket 8 (Hopkins & Groom 2002)	Targets with a predicted 'Ro5 druggable' domain
Bucket 9 (SureChEMBL)	Targets with 'chemical' patents in the last 5 years

Detailed rules for mAB

Bucket 1 (ChEMBL)	Targets with approved mAB drugs (phase4)
Bucket 2 (ChEMBL)	Targets with mAB in $\geq$ phase2
Bucket 3 (ChEMBL)	Pre-clinical targets with mAB
Bucket 4 (HPA)	Targets in "Plasma membrane", high confidence
Bucket 5 (Uniprot loc)	Targets in "Cell membrane" or "Secreted", high confidence
Bucket 6 (Uniprot loc)	Targets in "Cell membrane" or "Secreted" or "Membrane", low or unknown confidence
Bucket 7 (SigP + TMHMM)	Targets with predicted signal peptide or trans-membrane regions, and not destined to organelles
Bucket 8 (GO CC)	Targets with the parent term GO: 0005576 (extracellular region) or GO: 0031012 (extracellular matrix) or GO: 0005886 (plasma membrane) or their child terms.

Conflicts of interest

The authors declare no competing interest.

Acknowledgements

The authors gratefully acknowledge all colleagues at GSK and beyond who have helped in the development and application of the methods described in this perspective.

References

- O. Shalem, N. E. Sanjana, E. Hartenian, X. Shi, D. A. Scott, T. Mikkelsen, D. Heckl, B. L. Ebert, D. E. Root, J. G. Doench and F. Zhang, *Science*, 2014, **343**, 84–87.

- 2 A. L. Hopkins and C. R. Groom, *Nat. Rev. Drug Discovery*, 2002, 1, 727–730.
- 3 H. A. Hussein, C. Geneix, M. Petitjean, A. Borrel, D. Flatters and A.-C. Camproux, *Drug Discovery Today*, 2017, 22, 404–415.
- 4 Informa, *PharmaProjects*, <http://www.citeline.com/products/pharmaprojects/>.
- 5 UniProt Consortium, *Nucleic Acids Res.*, 2015, 43, D204–D212.
- 6 M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A.-K. Szigartyo, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P.-H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen and F. Pontén, *Science*, 2015, 347, 1260419.
- 7 A. Gutmanas, Y. Alhroub, G. M. Battle, J. M. Berrisford, E. Bochet, M. J. Conroy, J. M. Dana, M. A. Fernandez Montecelo, G. van Ginkel, S. P. Gore, P. Haslam, P. Haslam, R. Hatherley, P. M. S. Hendrickx, M. Hirshberg, I. Lagerstedt, S. Mir, A. Mukhopadhyay, T. J. Oldfield, A. Patwardhan, L. Rinaldi, G. Sahni, E. Sanz-García, E. Sanz-García, S. Sen, R. A. Slowley, S. Velankar, M. E. Wainwright and G. J. Kleywegt, *Nucleic Acids Res.*, 2014, 42, D285–D291.
- 8 ChEMBL, *DrugEBLity*, <https://www.ebi.ac.uk/chembl/drugability/structure>.
- 9 A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrián-Uhalte, M. Davies, N. Dedman, A. Karlsson, M. P. Magariños, J. P. Overington, G. Papadatos, I. Smit and A. R. Leach, *Nucleic Acids Res.*, 2017, 45, D945–D954.
- 10 G. Papadatos, M. Davies, N. Dedman, J. Chambers, A. Gaulton, J. Siddle, R. Koks, S. A. Irvine, J. Pettersson, N. Goncharoff, A. Hersey and J. P. Overington, *Nucleic Acids Res.*, 2016, 44, D1220–D1228.
- 11 R. D. Finn, P. Coghill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate and A. Bateman, *Nucleic Acids Res.*, 2016, 44, D279–D285.
- 12 A. Mitchell, H. Y. Chang, L. Daugherty, M. Fraser, S. Hunter, R. Lopez, C. McAnulla, C. McMenamin, G. Nuka, S. Pesseat, A. Sangrador-Vegas, M. Scheremetjew, C. Rato, S.-Y. Yong, A. Bateman, M. Punta, T. K. Attwood, C. J. A. Sigrist, N. Redaschi, C. Rivoire, I. Xenarios, D. Kahn, D. Guyot, P. Bork, I. Letunic, J. Gough, M. Oates, D. Haft, H. Huang, D. A. Natale, C. H. Wu, C. Orengo, I. Sillitoe, H. Mi, P. D. Thomas and R. D. Finn, *Nucleic Acids Res.*, 2015, 43, D213–D221.
- 13 B. H. M. Meldal, O. Forner-Martinez, M. C. Costanzo, J. Dana, J. Demeter, M. Dumousseau, S. S. Dwight, A. Gaulton, L. Licata, A. N. Melidoni, S. Ricard-Blum, B. Roechert, M. S. Skyzypek, M. Tiwari, S. Velankar, E. D. Wong, H. Hermjakob and S. Orchard, *Nucleic Acids Res.*, 2015, 43, D479–D484.
- 14 D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang and J. Woolsey, *Nucleic Acids Res.*, 2006, 34, D668–D672.
- 15 Gene Ontology Consortium, *Nucleic Acids Res.*, 2015, 43, D1049–D1056.
- 16 V. Chelliah, N. Juty, I. Ajmera, R. Ali, M. Dumousseau, M. Glont, M. Hucka, G. Jalowicki, S. Keating, V. Knight-Schrijver, A. Lloret-Villas, K. N. Natarajan, J.-B. Pettit, N. Rodriguez, M. Schubert, S. M. Wimalaratne, Y. Zhao, H. Hermjakob, N. Le Novère and C. Laibe, *Nucleic Acids Res.*, 2015, 43, D542–D548.
- 17 S. J. Campbell, A. Gaulton, J. Marshall, D. Bichko, S. Martin, C. Brouwer and L. Harland, *Drug Discovery Today*, 2010, 15, 3–15.
- 18 G. Koscielny, P. An, D. Carvalho-Silva, J. A. Cham, L. Fumis, R. Gasparyan, S. Hasan, N. Karamanis, M. Maguire, E. Papa, A. Pierleoni, M. Pignatelli, T. Platt, F. Rowland, P. Wankar, A. P. Bento, T. Burdett, A. Fabregat, S. Forbes, A. Gaulton, C. Y. Gonzalez, H. Hermjakob, A. Hersey, S. Jupe, Ş. Kafkas, M. Keays, C. Leroy, F.-J. Lopez, M. P. Magarinos, J. Malone, J. McEntyre, A. Munoz-Pomer Fuentes, C. O'Donovan, I. Papatheodorou, H. Parkinson, B. Palka, J. Paschall, R. Petryszak, N. Pratanwanich, S. Sarntivijal, G. Saunders, K. Sidiropoulos, T. Smith, Z. Sondka, O. Stegle, Y. A. Tang, E. Turner, B. Vaughan, O. Vrousseau, X. Watkins, M.-J. Martin, P. Sanseau, J. Vamathevan, E. Birney, J. Barrett and I. Dunham, *Nucleic Acids Res.*, 2017, 45, D985–D994.
- 19 R. Santos, O. Ursu, A. Gaulton, A. P. Bento, R. S. Donadi, C. G. Bologa, A. Karlsson, B. Al-Lazikani, A. Hersey, T. I. Oprea and J. P. Overington, *Nat. Rev. Drug Discovery*, 2017, 16, 19–34.
- 20 L. Yao and A. Rzhetsky, *Genome Res.*, 2008, 206–213.
- 21 M. A. Yildirim, K.-I. Goh, M. E. Cusick, A.-L. Barabási and M. Vidal, *Nat. Biotechnol.*, 2007, 25, 1119–1126.
- 22 P. R. Costa, M. L. Acencio and N. Lemke, *BMC Genomics*, 2010(11 Suppl 5), S9.
- 23 D. Emig, A. Ivliev, O. Pustovalova, L. Lancashire, S. Bureeva, Y. Nikolsky and M. Bessarabova, *PLoS One*, 2013, 8, e60618.
- 24 G. Laenen, L. Thorrez, D. Börnigen and Y. Moreau, *Mol. Biosyst.*, 2013, 9, 1676–1685.
- 25 J. Jeon, S. Nim, J. Teyra, A. Datti, J. L. Wrana, S. S. Sidhu, J. Moffat and P. M. Kim, *Genome Med.*, 2014, 6, 57.
- 26 Z.-C. Li, W.-Q. Zhong, Z.-Q. Liu, M.-H. Huang, Y. Xie, Z. Dai and X.-Y. Zou, *Anal. Chim. Acta*, 2015, 871, 18–27.
- 27 J. C. Somody, S. S. MacKinnon and A. Windemuth, *Drug Discovery Today*, 2017, 22, 1792–1799.
- 28 S. R. Kimura, H. P. Hu, A. M. Ruvinsky, W. Sherman and A. D. Favia, *J. Chem. Inf. Model.*, 2017, 57, 1388–1401.
- 29 J. S. Mason, A. Bortolato, M. Congreve and F. H. Marshall, *Trends Pharmacol. Sci.*, 2012, 33, 249–260.
- 30 M. S. Bodnarchuk, *Drug Discovery Today*, 2016, 21, 1139–1146.
- 31 A. P. Graves, I. D. Wall, C. M. Edge, J. M. Woolven, G. Cui, A. Le Gall, X. Hong, K. Raha and E. S. Manas, *Curr. Top. Med. Chem.*, 2017, 17, 2599–2616.
- 32 *Sitfinder from Molecular Operating Environment (MOE)*, 2013.08, Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2018.
- 33 T. Halgren, *Chem. Biol. Drug Des.*, 2007, 69, 146–148.

- 34 T. A. Halgren, *J. Chem. Inf. Model.*, 2009, **49**, 377–389.
- 35 V. Le Guilloux, P. Schmidtke and P. Tuffery, *BMC Bioinf.*, 2009, **10**, 168.
- 36 A. Bakan, N. Nevins, A. S. Lakdawala and I. Bahar, *J. Chem. Theory Comput.*, 2012, **8**, 2435–2447.
- 37 J. Oates, N. J. Hassan and B. K. Jakobsen, *Mol. Immunol.*, 2015, **67**, 67–74.
- 38 P. J. Hajduk, J. R. Huth and C. Tse, *Drug Discovery Today*, 2005, **10**, 1675–1682.
- 39 F. N. B. Edfeldt, R. H. A. Folmer and A. L. Breeze, *Drug Discovery Today*, 2011, **16**, 284–287.
- 40 I.-J. Chen and R. E. Hubbard, *J. Comput.-Aided Mol. Des.*, 2009, **23**, 603–620.
- 41 M. M. Hann, A. R. Leach and G. Harper, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 856–864.
- 42 C. C. Arico-Muendel, *Med. Chem. Commun.*, 2016, **7**, 1898–1909.
- 43 T. N. O'Connell, J. Ramsay, S. F. Rieth, M. J. Shapiro and J. G. Stroh, *Anal. Chem.*, 2014, **86**, 7413–7420.
- 44 D. M. Dias, I. Van Molle, M. G. J. Baud, C. Galdeano, C. F. G. C. Geraldine and A. Ciulli, *ACS Med. Chem. Lett.*, 2014, **5**, 23–28.
- 45 C. T. Gee, E. J. Koleski and W. C. K. Pomerantz, *Angew. Chem., Int. Ed.*, 2015, **54**, 3735–3739.
- 46 C. Abell and C. Dagostin, in *Fragment-Based Drug Discovery*, The Royal Society of Chemistry, 2015, pp. 1–18.
- 47 A. R. Leach and M. M. Hann, *Curr. Opin. Chem. Biol.*, 2011, **15**, 489–496.
- 48 K. M. Backus, B. E. Correia, K. M. Lum, S. Forli, B. D. Horning, G. E. González-Páez, S. Chatterjee, B. R. Lanning, J. R. Teijaro, A. J. Olson, D. W. Wolan and B. F. Cravatt, *Nature*, 2016, **534**, 570–574.
- 49 C. G. Parker, A. Galmozzi, Y. Wang, B. E. Correia, K. Sasaki, C. M. Joslyn, A. S. Kim, C. L. Cavallaro, R. M. Lawrence, S. R. Johnson, I. Narvaiza, E. Saez and B. F. Cravatt, *Cell*, 2017, **168**, 527–541e29.
- 50 R. Fernandez-Leiro and S. H. W. Scheres, *Nature*, 2016, **537**, 339–346.
- 51 I. Schlichting, W. E. White and M. Yabashi, *J. Synchrotron Radiat.*, 2015, **22**, 471.
- 52 B. Deng, C. Lento and D. J. Wilson, *Anal. Chim. Acta*, 2016, **940**, 8–20.
- 53 M. Bantscheff, *Methods Mol. Biol.*, 2012, **803**, 3–13.
- 54 *EBI industry workshop: Target Tractability Assessment*, May 3rd/5th, 2017, Access for EMBL-EBI Industry Programme members is available at: <https://www.ebi.ac.uk/industry/private/industry-workshop/2017/05/target-tractability-assessment>.
- 55 D. N. Petersen, J. Hawkins, W. Ruangsiriluk, K. A. Stevens, B. A. Maguire, T. N. O'Connell, B. N. Rocke, M. Boehm, R. B. Ruggeri, T. Rolph, D. Hepworth, P. M. Loria and P. A. Carpino, *Cell Chem. Biol.*, 2016, **23**, 1362–1371.
- 56 A. C. Lai and C. M. Crews, *Nat. Rev. Drug Discovery*, 2017, **16**, 101–114.
- 57 C. Fellmann, B. G. Gowen, P.-C. Lin, J. A. Doudna and J. E. Corn, *Nat. Rev. Drug Discovery*, 2016, **16**, 89–100.