Inhibitor selectivity: profiling and prediction

Janssen, A.P.A.

Citation

Janssen, A. P. A. (2019, May 1). *Inhibitor selectivity: profiling and prediction*. Retrieved from https://hdl.handle.net/1887/71808

Author: Janssen, A.P.A.
Title: Inhibitor selectivity: profiling and prediction
Issue Date: 2019-05-01

> *A very large part of space-time must be investigated,*
> *if reliable results are to be obtained.*
> Alan Turing

8

Drug Discovery Maps visualizes and predicts kinome-inhibitor interaction landscapes

Part of this research was published in A.P.A. Janssen *et al.*, *J. Chem. Inf. Model.* acs.jcim.8b00640 (2018).

Introduction

Protein kinases are an important class of drug targets due to their key role in intracellular signal transduction processes involved in cancer, auto-immune diseases, and (neuro)inflammation.[1,2] The therapeutic value of the protein kinase family is demonstrated by the 38 kinase inhibitors currently approved by the FDA and the plethora of molecules being tested in clinical trials for this enzyme family.[3] It is anticipated that these clinically approved kinase inhibitors (KI) can serve as starting points to identify novel drug candidates for other kinases.

Most KIs interact with the structurally and functionally conserved ATP-binding site, present in all >500 human protein kinases. It is well established that KIs bind multiple members of the kinase family, and that this may affect their efficacy and toxicity.[4] Detailed investigation of the target-interaction landscape of KIs is, therefore, important to understand their molecular mode of action and offers the opportunity to identify new starting points for other therapeutically interesting kinases. Many complex, high-dimensional datasets with structure-activity relationships (SAR) of KIs over a broad selection of kinases have become available (Table 8.1).[5–10] The wealth of data present in literature has spurred the development of several computational platforms to predict kinase activities of inhibitors.[11,12] These reported highly specialised computational models which were able to predict, to some extent, the inhibition profiles of inhibitors against the kinome.

Given the large amount of data available and the apparent success of previous attempts to use these datasets in a predictive manner it was expected that the strategies introduced in Chapter 7 could be applied to the field of kinase inhibitors. This approach was dubbed Drug Discovery Maps (DDM), a machine learning tool that allows the visualization and prediction of target-ligand interaction landscapes.

Table 8.1 | Summary of large experimental (partially) public kinome screens for sets of kinase inhibitors. Numbers of inhibitors and kinases are as reported in original publications and include mutants, splicing variants and separate kinase domains from the same kinase. Coverage represents the percentage of possible kinase-inhibitor pairs for which a data point is reported.

Set	No. of inhibitors	No. of kinases	Dosage	No. of data points	Coverage
Karaman *et al.*	38	317	Dose response**	12046	100%
Metz *et al.*	3858	172	Dose response	258094	38.9%
Anastassiadis *et al.*	178	300	Single dose	52834	98.9%
Davis *et al.*	72	442	Dose response**	31824	100%
Elkins *et al.*	367*	224	Two doses	81940	99.9%

* contains duplicate molecules

** only if activity is observed at initial 10 μM single dose screen

t-SNE reliably reproduces human chemotype assignments

In Chapter 7 it was demonstrated that t-SNE is able to visualize molecular similarity in a drug-like set of 2774 clinically applied molecules. For the current purposes, it had to be verified that t-SNE is still able to recognize molecular similarity within a smaller set of drug-like molecules that is more homogenous and has higher molecular similarity. To this end, a t-SNE-mediated clustering on the molecules from the Published Kinase Inhibitor Set (PKIS) was performed.[13] The PKIS is a 364-member library of molecules assembled by GlaxoSmithKline that are all classified as inhibitors of protein kinases. The PKIS set represents 31 manually annotated chemotypes and their activity has been measured and published for 200 kinases (Table 8.1, last entry).[9] The resulting map of chemical space representing the KIs (Figure 8.1A) was coloured according to the manual chemotype attribution and showed clear co-localisation of chemotypes. A more in-depth analysis by the unsupervised cluster assignment algorithm DBSCAN confirmed the initial visual inspection and shows high statistical correlation between the autonomously derived clustering and the human annotation.[14] Of the 31 chemotypes annotated, 23 were fully comprised in one computationally assigned cluster. For example, the orange and gold clusters on the left of the map are completely isolated and comprise all compounds of those chemotypes (Figure 8.1B). An enlargement of the cluster dominated by 3-amino-pyrazolopyridines (Figure 8.1B) showed that in fact three different chemotypes were included. Compounds **1**-**3** are examples of each, and they showed that indeed these compounds are highly similar, and perhaps arbitrarily assigned to different chemotypes. The 'stray' sky blue compound **4** is part of a chemotype that is less well defined in t-SNE space, but the similarities with compounds **1**-**3** are obvious. This analysis showed that t-SNE is capable to recognize and cluster molecular entities in a highly specific manner and that it allows the visual inspection of high-dimensional chemical structural data, or chemical space, in an easy and intuitive way.

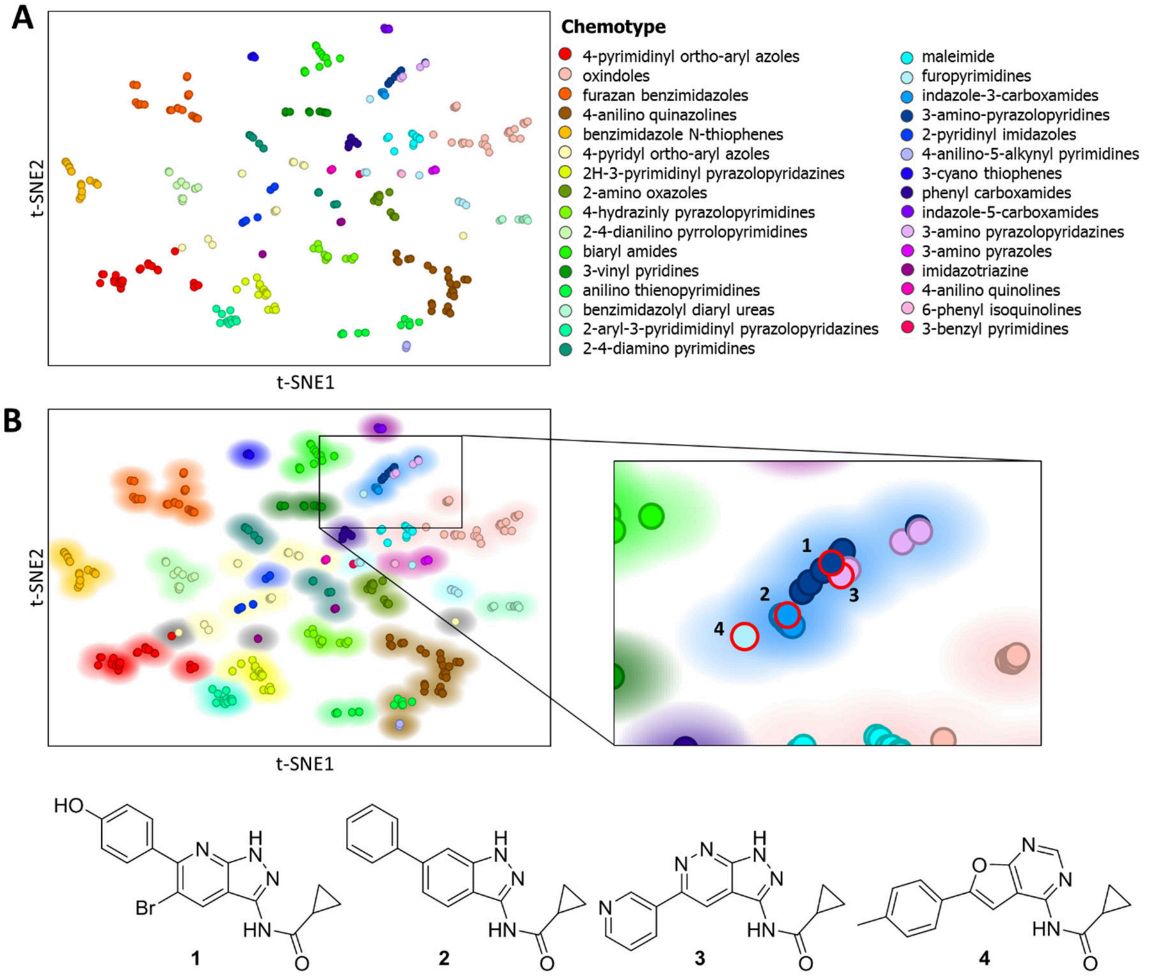


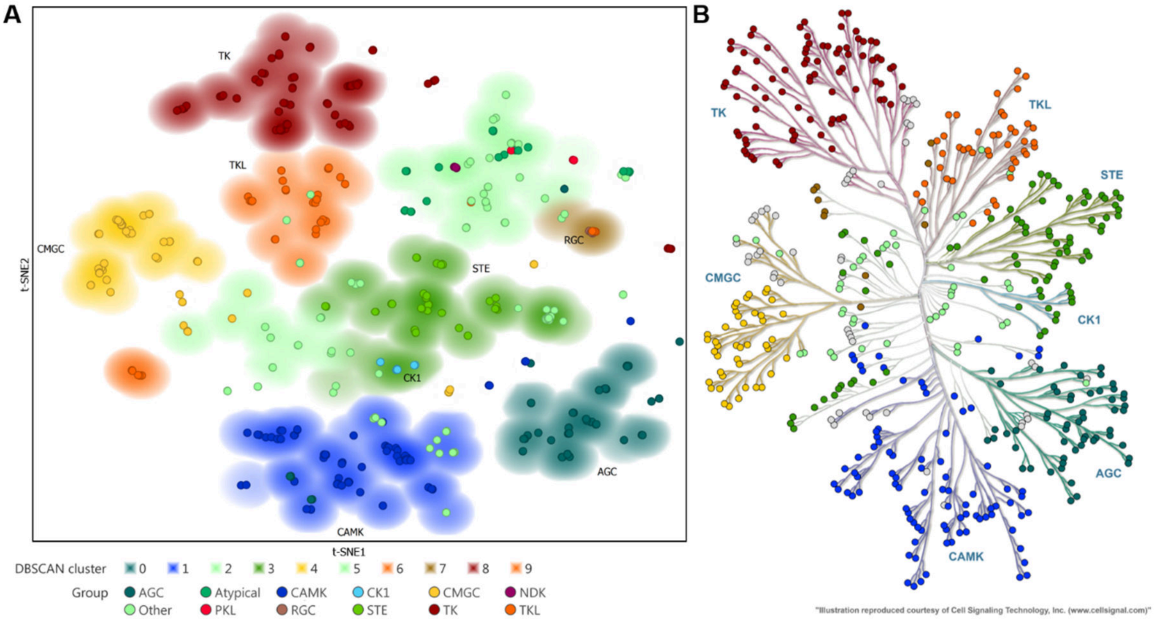
Figure 8.1 | A) t-SNE embedding of the Published Kinase Inhibitor Set. Embedding is based on the 4096-bit Morgan fingerprint. t-SNE settings: perplexity = 50, learning rate = 50, iterations = 10,000. Markers are coloured according to 31 manually attributed chemotypes. B) t-SNE embedding as in A) but including a background of the DBSCAN generated clustering, coloured by the dominant chemotype in that cluster (grey are singletons or duos). Markers are coloured according to the 31 chemotypes defined by Elkins *et al.* as in A).

t-SNE map of the kinase target space

To generate a visualization of the target space, and to quantify the similarity of kinases, first an alignment of the amino acid sequences of the whole kinase domains containing the ATP-binding pocket was created. This was translated to a fingerprint based on physico-chemical properties of the amino acids, analogous to the methods used in Chapter 7.[15] These fingerprints were used to create a two-dimensional map of the target space by the t-SNE algorithm. The resulting embedding (Figure 8.2A) is remarkable, as it almost seamlessly recreates the phylogenetic tree published by Manning *et al*. in 2002.[16] To assign the kinases to clusters, the coordinates of the t-SNE embedding were fed into the DBSCAN algorithm. All ten assigned clusters were significantly ($P < 0.0001$, Hypergeometric test) enriched for a specific kinase group as assigned by Manning *et al*. (Figure 8.2A). Closer inspection of some of the kinases unassigned by DBSCAN reveals that they belong to distinct branches of the phylogenetic tree, corresponding with their separation from the main clusters. As an example, the four tyrosine kinases (TK) in the far right of the embedding (burgundy) all belong to the JAK family (JAK1, 2, 3 and Tyk2), but only represent their second kinase domain. The first kinase domain is more closely associated with the rest of the TK group, and lies just outside the DBSCAN assigned cluster. The close association of the second kinase domains with the RGC cluster (coloured brown) is especially striking, as these domains, just as the RGC kinases, are considered pseudokinases. The same holds true for MLKL, IRAK2 and IRAK3. Intriguingly, the IRAK family of TKL kinases has four members, of which IRAK1 and IRAK4 are catalytically active, whereas IRAK2 and IRAK3 are not.[17] In the t-SNE embedding, the former are located in the major TKL cluster (orange) whereas the latter are actually assigned to the RGC-dominated cluster. MLKL has also been shown to indeed lack catalytic activity in at least one report.[18]

Another interesting feature is the separation of a group (left of plot) of TKL kinases from the major cluster. This subset features all but one of the STKR family of cell surface bound receptor kinases. Upon closer inspection, even the subfamilies of STRK1 and -2 are discernible. Strikingly, the MISR2 (AMHR2) kinase receptor is located with kinases categorised as ‘Other’. This receptor kinase has an atypical DFG motif (DLG), and as such can indeed be classified as a pseudokinase, although phosphorylating activity has experimentally been shown.[19] The other members of the STKR family do all share the conserved DFG motif. Finally, on the lower side of the t-SNE plot several AGC coloured kinases have been clustered with the CAMK kinases. These actually represent the second kinase domains of the RSK family, which were also attributed to the CAMK group by Manning *et al*.[16]

In summary, this analysis of target space of the binding site of protein kinase domains indicated that this embedding is able to recognize overall similarity but also detect subtle differences between the different binding domains of most kinase inhibitors.

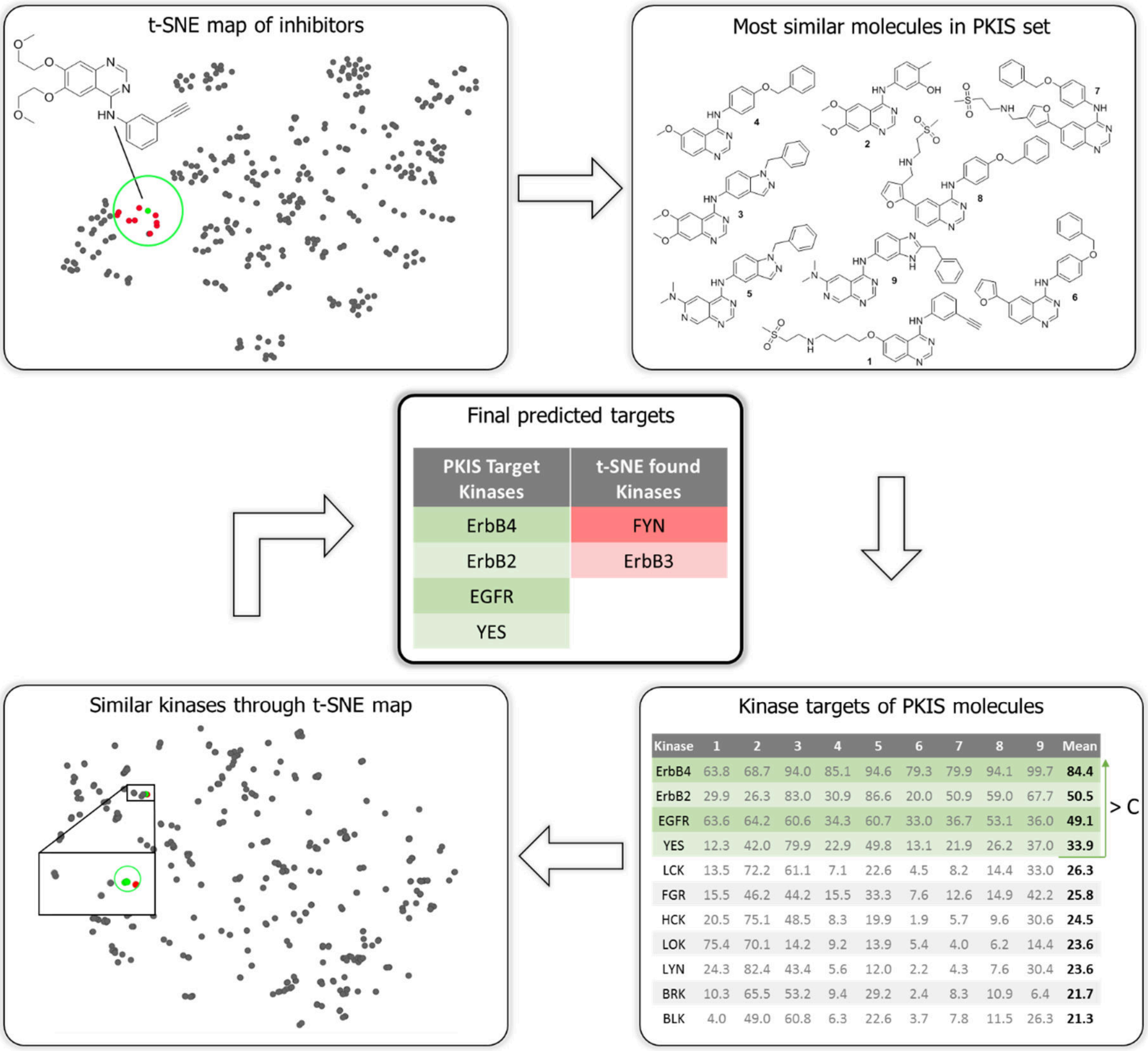


"Illustration reproduced courtesy of Cell Signaling Technology, Inc. (www.cellsignal.com)"

Figure 8.2 | t-SNE visualization of kinase domains reveals phylogenetic information: A) t-SNE embedding of physicochemical fingerprint of the kinase domains of 535 human kinase domains. t-SNE settings: perplexity = 50, learning rate = 50, iterations = 25,000. Arbitrary t-SNE coordinates are rotated to match dendrogram orientation of Manning *et al*. Markers are coloured according to the 12 groups defined by Manning *et al*., background is coloured based on the DBSCAN generated clustering, coloured by the dominant kinase group in that cluster (blanks are unclustered kinases). B) Manning *et al*. manually curated kinome dendrogram, overlaid with circles coloured according to the background colouring from the t-SNE map in A), based on the unsupervised DBSCAN clustering.[20]

DDM can predict target-ligand interaction landscapes

Based on the successful analysis of the chemical and target space of kinases and their inhibitors, a workflow was set up to predict the activity of novel compounds for the entire kinome. The bioactivity data measured by Elkins *et al.* for the Published Kinase Inhibitor Set was used as training set, as it contains the most unique interactions of all open datasets (Table 8.1).[9] The optimisation of the workflow with all parameters is described in more detail in the Methods. The final architecture of the algorithm is depicted in Figure 8.3, illustrated for the EGFR inhibitor erlotinib. At first, a t-SNE embedding is generated where erlotinib is mapped onto the chemical space of the PKIS (top left). This information is used to find the 9 most similar molecules (top right). Of these, the inhibition data measured by Elkins *et al*. is averaged, and all kinases above a threshold value *C* are considered targets (lower right). These kinases are then looked up in the target space map (Figure 8.2A), and the most similar kinases are appended (lower left) to yield the final prediction (centre). As the molecular t-SNE embedding is slightly stochastic, the described process is repeated several times (*R*) and the number of times a kinase is predicted is tracked.



PKIS Target Kinases	t-SNE found Kinases
ErbB4	FYN
ErbB2	ErbB3
EGFR	
YES	

Kinase	1	2	3	4	5	6	7	8	9	Mean
ErbB4	63.8	68.7	94.0	85.1	94.6	79.3	79.9	94.1	99.7	**84.4**
ErbB2	29.9	26.3	83.0	30.9	86.6	20.0	50.9	59.0	67.7	**50.5**
EGFR	63.6	64.2	60.6	34.3	60.7	33.0	36.7	53.1	36.0	**49.1**
YES	12.3	42.0	79.9	22.9	49.8	13.1	21.9	26.2	37.0	**33.9**
LCK	13.5	72.2	61.1	7.1	22.6	4.5	8.2	14.4	33.0	**26.3**
FGR	15.5	46.2	44.2	15.5	33.3	7.6	12.6	14.9	42.2	**25.8**
HCK	20.5	75.1	48.5	8.3	19.9	1.9	5.7	9.6	30.6	**24.5**
LOK	75.4	70.1	14.2	9.2	13.9	5.4	4.0	6.2	14.4	**23.6**
LYN	24.3	82.4	43.4	5.6	12.0	2.2	4.3	7.6	30.4	**23.6**
BRK	10.3	65.5	53.2	9.4	29.2	2.4	8.3	10.9	6.4	**21.7**
BLK	4.0	49.0	60.8	6.3	22.6	3.7	7.8	11.5	26.3	**21.3**

Figure 8.3 | Schematic overview of the DDM workflow. In this example, the targets of erlotinib are predicted. Based on a t-SNE embedding (top left) the PKIS inhibitors nearest to erlotinib are found, depicted in the top right panel. For these, the inhibition data as measured by Elkins *et al*. is averaged and used as initial prediction. These targeted kinases are then looked up in the t-SNE embedding in the lower left panel, where the most similar kinases are added to yield the final prediction (centre).

The DDM model was validated using an independent dataset generated by Karaman *et al*.[5] The resulting prediction statistics for each of the 38 compounds in this test set are summarised in Table S8.1. The average positive prediction value (PPV) was 40%, with a Matthews correlation coefficient (MCC) of 0.21. These statistics were compared with previously published methods, and it was found that DDM performed better than QSAR-models and was equal in performance to random forest based proteochemometric models (Figure 8.4). A receiver operating characteristic (ROC) analysis of the performance of DDM on this test set shows an area under the curve (AUC) of 0.76 (Figure 8.5). Taken all together, a novel machine learning model to predict kinome-inhibitor landscapes was developed and validated.

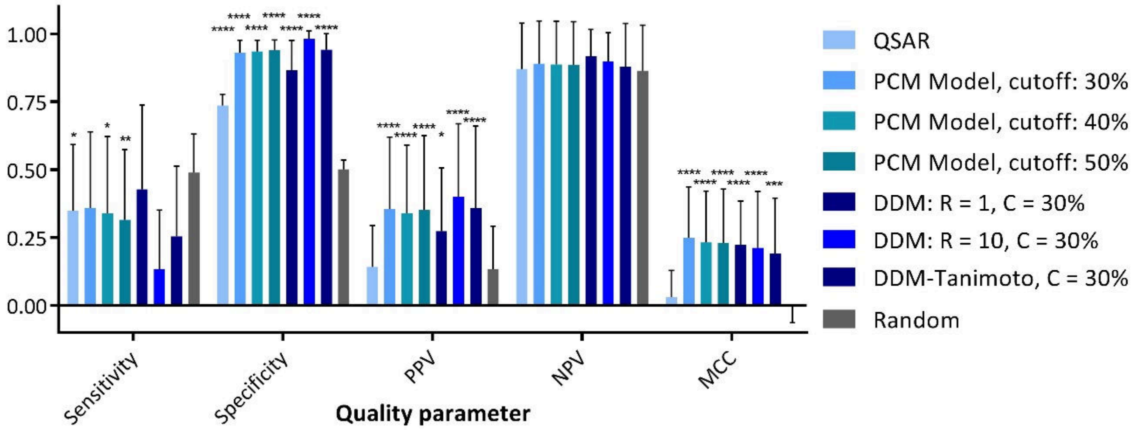


Figure 8.4 | Statistical comparison of the new t-SNE based model with four models based on published procedures. Bars denote mean ± SD of the 38 inhibitors. Significant improvements over random have been highlighted, * $p < 0.0332$, ** $p < 0.0021$, *** $p < 0.0002$, **** $p < 0.0001$ (Two-way ANOVA, Tukey's multiple comparison test). The relatively low PPV and high NPV for the random predictions are caused by the activity bias in the data set (13% active vs 87% inactive).

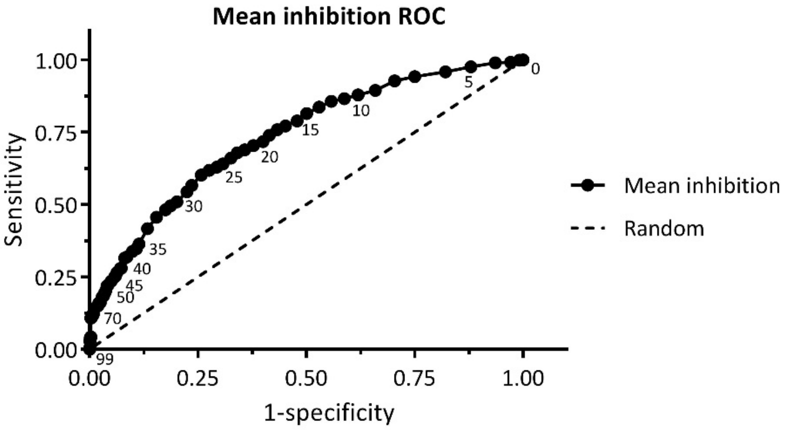


Figure 8.5 | Receiver operating characteristic curve of the test set predictions with varying cut-off C values. Marker labels represent *C* values. The area under the curve is 0.76, with the maximum deviation from random at C values between 25 and 35%. Axes have been normalized to take into account the inability to predict all kinases based on the PKIS coverage.

Discovery of novel FLT3 inhibitors using DDM

To investigate the utility of the model in early drug development, it was applied to the identification of new inhibitors for FMS-like Tyrosine kinase 3 (FLT3). FLT3 is implicated in advanced myeloid leukaemia, where approximately 30% of patients carry an internal tandem duplication (ITD) in their FLT3 gene that activates the kinase and acts as a driver mutation.[21] Recently, midostaurin has been FDA approved for the treatment of AML patients and several other inhibitors are currently being tested in clinical trials. However, fast adaptive mutations in the FLT3 gene quickly result in drug-induced resistance of AML, warranting the search for novel chemotypes to inhibit this kinase. To this end, the DDM model was used to predict the kinome-ligand interaction landscape of a small kinase-focused library of 1152 molecules. They were analyzed using various values for the activity cut-off *C*, and were ultimately filtered using *C* = 40%, and a prediction count of at least 9 out of 10 runs, to have a balanced number of molecules to be tested. These stringent cut-offs yielded a workable set of 44 compounds predicted to be active against FLT3.

To validate our virtual DDM screen, a time-resolved fluorescence-resonance energy transfer (FRET)-based biochemical assay was performed with all 1152 compounds against FLT3, at an initial concentration of 10 µM. This screen yielded 184 actives, with >50% loss of activity (16% of all compounds). Of these compounds, the pIC_{50} was measured, resulting in 135 compounds with a $pIC_{50} > 5$, with a mean of 6.7 ± 0.9. 18 of the 184 compounds were part of the 44 compounds identified by the DDM screen, which results in a positive predicted value (PPV; or hit rate) of 41% (Figure 8.6A, $P < 0.0001$ (Hypergeometric test)), which is almost three-fold higher compared to the hit rate of the biochemical assay. Of note, 15 of the predicted compounds demonstrated an IC_{50} of less than 2 µM (34%, $P < 0.0001$ (Hypergeometric test)) with an average pIC_{50} 7.3 ± 1.1, and also included the most active compound found in the screen (crenolanib, pIC_{50} of 9.0). The hit rate was nearly identical to the validation statistics for the test set (Figure 8.4), where an overall PPV of 40% was achieved. The same holds for the negative predictive value (89%) and the sensitivity (11%). The successful application of the DDM model for the FLT3 screen may partially be attributed to the high coverage for the TK family of kinases. Of note, the relatively low sensitivity (11%) is a balanced choice between minimising the number of compounds to screen and finding more actual hits. This can easily be tuned by varying the cut-off parameter.

Two of the predicted compounds, 5 and 6 (Figure 8.6B), were selected based on their chemical properties, novelty regarding FLT3 inhibition, and their predicted interaction profile (*vide infra*). These compounds were resynthesized using established methods (see Methods). The activity of the compounds was confirmed in a FRET assay using recombinant human FLT3 (Figure 8.6C). Compounds 5 and 6 showed a concentration-dependent activity with pIC_{50} values of 7.3 ± 0.1 and 8.8 ± 0.1, respectively. To determine the cellular activity of both compounds, a cell proliferation assay using the FLT3-dependent AML cell line MV4:11 was performed. Both compound 5 and 6 showed clear cellular activity with pEC_{50} values of 6.3 ± 0.1 and 8.5 ± 0.1, respectively. In summary, the experimental validation of the hits illustrated the power of the DDM workflow for compound selection in the lab.

Finally, to explain the potential binding mode of compound 5 and 6 both compounds were docked using a DFG-in model (for 5) and the DFG-out structure 4RT7 (for 6) (Figure 8.6E). Compound 5 binds to the hinge region with the aminopyrimidine moiety in a fashion typical for Type 1 kinase inhibitors. Compound 6 binds FLT3 in the DFG-out conformation analogous to RIPK2 (5AR7) by forming hydrogen bonds to the DFG-motif using the urea functionality and to the hinge region with the pyridine nitrogen.[22]

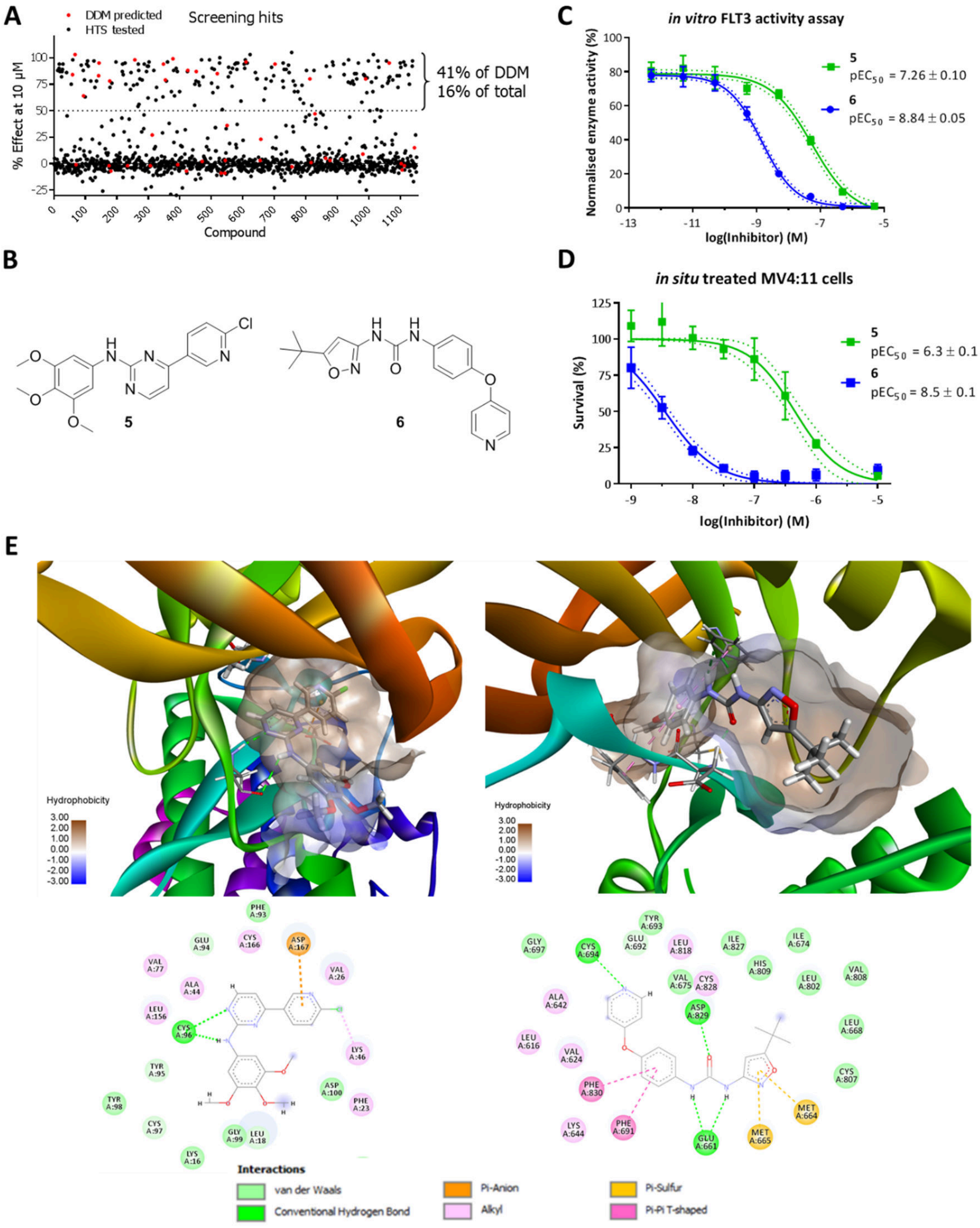


Figure 8.6 | Discovery of novel FLT3 inhibitors using DDM. A) Scatter plot of all compounds and their inhibitory effect at 10 μM as measured in the high-througput screen. DDM predicted molecules are marked red. B) Structures of the two compounds re-synthesized and tested *in situ* against MV4:11 cells. C) Dose response curves of compounds 5 and 6 against recombinant FLT3 in a FRET-based activity assay. Markers denote mean ± SD (N=4). Dotted lines denote 95% confidence interval of EC_{50} fit. D) Dose response curves of compounds 5 and 6 against MV4:11 leukaemia cells. Markers denote mean ± SD (N=3). Dotted lines denote 95% confidence interval of EC_{50} fit. E) Docking poses of 5 and 6 in the 3D models of FLT3 and the corresponding 2D interaction plots.

Kinome activity spectrum prediction using DDM

To reduce potential toxic side effects, kinase cross reactivity is ideally minimized. DDM enables rapid assessment of the predicted cross reactivity, because by default, DDM predicts the interactions with the entire kinome. Thus far, however, only the FLT3 prediction has been taken into account. As final validation, the activity of the two inhibitors on the predicted off-targets was tested in biochemical assays. In addition to FLT3, compounds **5** and **6** were predicted to be active against 35 and 33 kinases, respectively (C = 40%, $R > 0.5$). The off-targets were validated using KinaseProfiler™ by Eurofins at 10 µM. The inhibition data per compound are shown in Table S8.2. For compound **5** the predictions were 69% accurate (24 of the 35 off-targets confirmed (<50% remaining activity) with two additional off-targets in the low 50% residual activity range). For compound **6** the prediction was exceedingly accurate as 26 of the 33 targets (79%) were indeed inhibited >50%. To conclude, DDM was able to predict the kinome-inhibitor interaction landscape at a relatively high accuracy.

Predicting the activity profile of a large commercially available hinge binder set

The numerous recent approvals of small molecule kinase inhibitors by the FDA (6 in 2017, 13 since 2014) conceals the fact that none of the last 13 approvals targeted novel kinases or mechanisms. A lot of therapeutic potential may still lie in currently untargeted pathways. To aid in the elucidation of signalling pathways and validation thereof, easy access to KIs is crucial to allow biochemical evaluation by acute inhibition of regulating kinases. To facilitate this, and in the spirit of the Published Kinase Inhibitor Set goals, the model was used to predict the interaction landscape of a large commercially available hinge binder set. To generate a large dataset like this, the Enamine Kinase Hinge Region Directed Library of 18,020 molecules was downloaded and all targets were predicted for these molecules (Figure 8.7A).[23] After filtering for unpredicted kinases and inert molecules all 18,020 inhibitors remain, with 290 kinases as targets (confidence ≥ 0.1, C = 20%). The distribution of the targets across the kinome tree is quite homogenous, as shown in Figure 8.7B. The only major group that is underrepresented are the TKL kinases, but this follows directly from the t-SNE map depicted in Figure S8.1, as the biochemical data generated for the PKIS set did not include many TKL kinases. In total 1,888,418 interactions are predicted, with FLT3 as most inhibited kinase. For all compounds a value similar to the Selectivity Entropy (S_{sel}) could also be calculated, when the confidence parameter is assumed as the association constant ($1/K_d$).[24] This parameter can be used to judge the overall promiscuity of a selected inhibitor, to directly avoid pursuing the least selective scaffold.

As an illustration of the use of this large dataset, two kinase targets were chosen based on recent literature, for which few molecular modulators are known and which were not included in the PKIS dataset. Homeodomain-interacting protein kinase 3 (HIPK3, also known as FIST) is a protein kinase involved in transcription regulation, and is thought to negatively regulate apoptosis in certain triple negative breast cancer cell lines. Inhibition of this kinase may thus prove useful in treatment of these breast cancers.[25] p21-activated

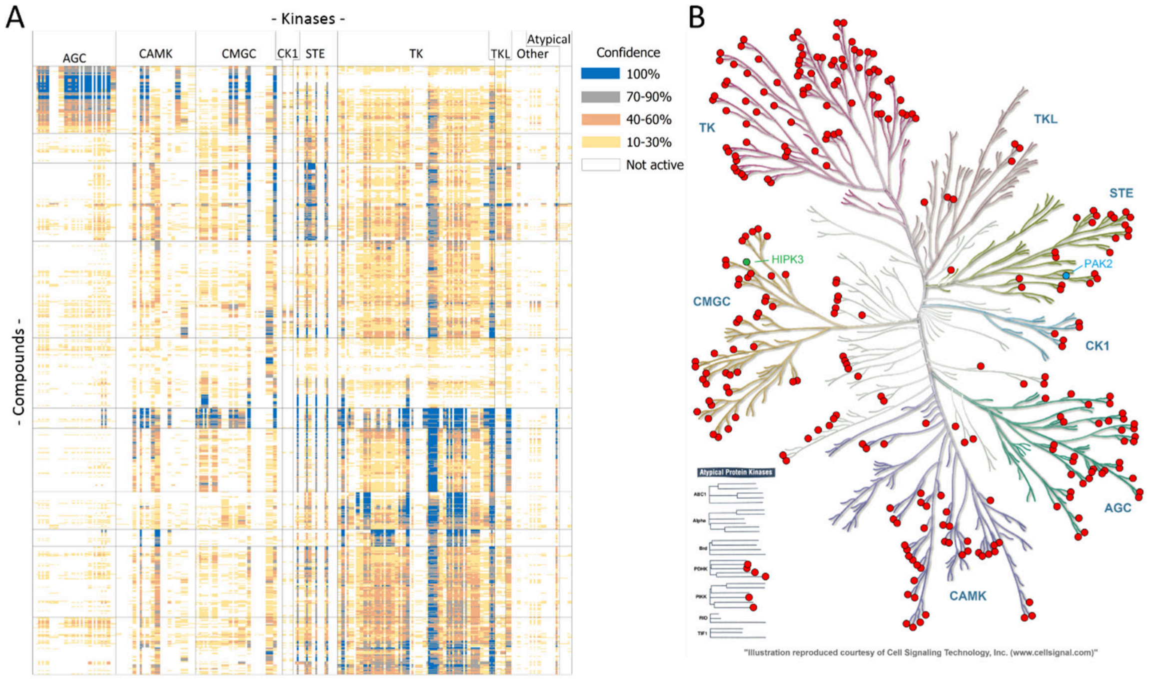


Figure 8.7 | A) Heatmap of the enamine inhibitor set. Compounds are hierarchically sorted on the predicted activity profile, kinases are sorted by group, then family. Cut-off value of 20% is used to increase sensitivity. B) Traditional kinase dendrogram showing the predicted kinases in the Enamine heatmap in red, PAK2 and HIPK3 are highlighted in blue/green. Picture was generated using KinMap.[20]

kinase 2 (PAK2) is activated by CDC42 and RAC1. When active it stimulates cell survival and growth. Recently it was implicated in oestrogen receptor positive (ER^+) breast cancer as tamoxifen resistance inducer.[26] These two kinases belong to two distinct groups (CMGC and STE, respectively), also distinct from FLT3 (TK) treated above. HIPK3 was predicted as target quite frequently, leading to a set of 172 potential inhibitors (C = 30%, confidence = 1). PAK2 on the other hand is only rarely predicted, and only 57 molecules were found, with poor confidence (C = 20%, confidence = 0.1). The chosen targets are thus also at the far ends of the potentially usable spectrum of the dataset. For both targets 10 molecules were selected based on availability, price, and selectivity entropy. These 20 compounds were purchased and screened in the KinaseProfiler™ platform against both PAK2 and HIPK3. As there was no overlap between the two sets of predicted molecules, the molecules chosen for HIPK3 were used as negative controls for PAK2 and *vice versa* to check for specific enrichment by the model. The activity data, together with the molecular structures, is summarized in Table S8.3. These data showed no identified actives for HIPK3 or PAK2. The high positive prediction value reached for FLT3 above is thus shown not to hold for all targets.

Discussion

Drug discovery is still largely an empirical process, which is challenging, time consuming and hard.[27] The multi-parameter optimization of chemical structures, which is needed to balance the activity and selectivity of a drug candidate, requires the understanding of high-dimensional datasets. Machine learning algorithms have been employed to analyse and predict compound activity using large datasets with varying success.[11,12,28] Some of the major drawbacks of most computational models are the complexity of the algorithm and the ‘black box’ nature of the systems. Implementation and interpretation of such systems is not trivial and, consequently, has not been widely adopted by the drug discovery community.

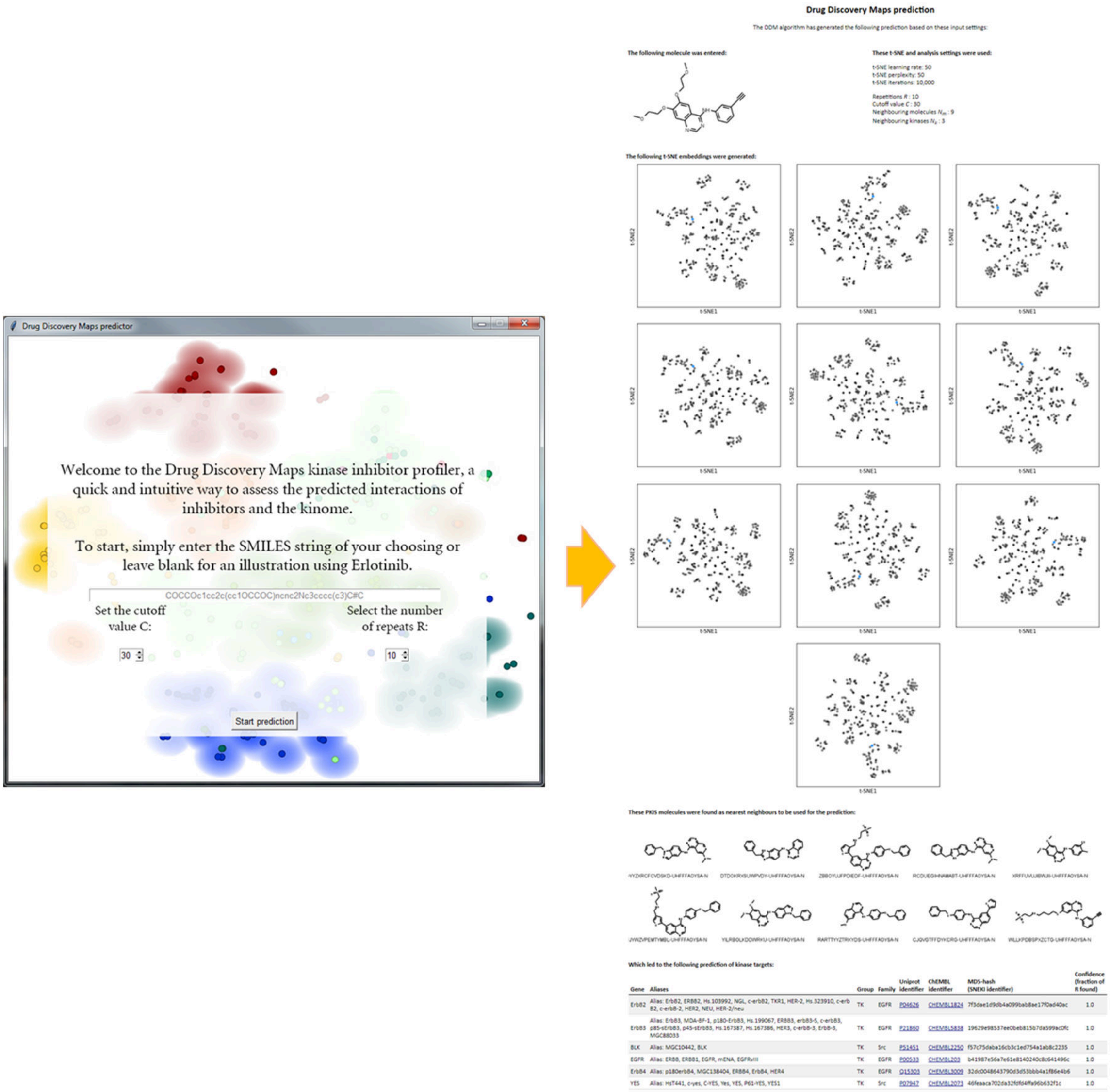
Drug Discovery Maps provides an intuitive, data-driven (bio)molecule similarity clustering procedure using state-of-the-art machine learning techniques. The model is based on the t-Distributed Stochastic Neighbour Embedding algorithm to generate a visualization of molecular similarity in two dimensions.[29,30] Colour is used as a third dimension to interactively visualize the biological activity or compound class (chemotype). DDM combines two different maps. The first map depicts the chemical space in which compounds are clustered based on their molecular similarity, whereas in the second map protein targets are clustered based on the chemical similarity of the amino acids making up the kinase domain. By combining both maps, DDM was able to predict bio-activities of small molecules across a protein family. DDM was applied to visualize the chemical space of the Published Kinase Inhibitor Set (PKIS) and the target space of the protein kinase family (kinome). It was able to predict the kinome activity profile of another independent set of kinase inhibitors with comparable or better scores than the currently available machine learning techniques. DDM was then used to identify new hits for the oncogene FMS-like tyrosine kinase 3 (FLT3), a validated therapeutic target for the treatment of acute myeloid leukaemia.[31] The hits were resynthesised and their biological activity was validated in biochemical and cellular assays. Finally, the off-target profile of the hits as predicted by DDM was validated in a panel of kinase assays. The method was then applied to a large commercially available screening collection to predict the interaction landscape of over 18,000 inhibitors. 20 inhibitors, predicted to be active for either HIPK3 or PAK2, were purchased and tested *in vitro*. Unfortunately, none proved to be active, which was an illustration of the limits of our approach. The average positive predictive value of 40% found in the validation and the FLT3 screen is apparently not universally achieved. This could be attributed to the sparse data density around the two selected targets. Another factor to take into account is plain luck of the draw (or lack thereof). If the PPV was 20%, there is an 11% chance of choosing 10 inactive inhibitors at random from a set of predicted actives.

Although the model performs equally well or better than the current computational drug discovery tools, it is envisioned that it can be further improved when more comprehensive datasets become available in the public domain. In the PKIS training set, 364 inhibitors were only tested at two concentrations on approximately 200 unique wild-type kinases. A more expansive dataset of a broader set of more diverse compounds tested on a larger number of kinases in a concentration-response fashion would inherently improve the predictions generated over the entire kinome.

The fact that a complete interaction profile is predicted uniquely enables the ranking of predicted hits by selectivity, for example by calculation of the predicted selectivity entropy.[24] This allows medicinal chemists in theory to rank scaffolds based on promiscuity and also to select for accepted off-targets, that depend on the biological questions or medical indication. This was demonstrated by the KinaseProfiler™ screen of predicted off-targets for the two FLT3 inhibitors.

To aid in the implementation of the tool as it is presented here, a Python based executable including a Graphical User Interface has been made available online[32] (Figure 8.8). The unpackaged Python script with a list of dependencies is also available. Also included is a fully annotated KNIME workflow, to allow step by step execution and analysis. This set of tools should enable the integration of this data-driven approach into any project without any need of investments *a priori*.

To conclude, the machine learning algorithm Barnes-Hut t-SNE was successfully implemented in a drug discovery setting to predict ligand-protein interaction landscapes. The concept of DDM is applicable to a multitude of drug discovery challenges, which, given the proper dataset, can be used to design a small molecule with a balanced set of physico-chemical and biological properties as required for drug candidates. It is envisioned that DDM may make the drug discovery process more efficient.



Welcome to the Drug Discovery Maps kinase inhibitor profiler, a quick and intuitive way to assess the predicted interactions of inhibitors and the kinome.

To start, simply enter the SMILES string of your choosing or leave blank for an illustration using Erlotinib.

COCCOc1cc2c(cc1OCCOC)ncnc2Nc3cccc(c3)C#C

Set the cutoff value C: 30

Select the number of repeats R: 10

Drug Discovery Maps prediction

The DDM algorithm has generated the following prediction based on these input settings:

The following molecule was entered:

These t-SNE and analysis settings were used:

t-SNE learning rate: 50
t-SNE perplexity: 50
t-SNE iterations: 10,000

Repetitions R : 10
Cutoff value C : 30
Neighbouring molecules N_m : 9
Neighbouring kinases N_k : 3

The following t-SNE embeddings were generated:

These PKIS molecules were found as nearest neighbours to be used for the prediction:

Which led to the following prediction of kinase targets:

Gene	Aliases	Group	Family	Uniprot identifier	ChEMBL identifier	MD5-hash (SNEKI identifier)	Confidence (fraction of R found)
ErbB2	Alias: ErbB2, ERBB2, Hs.103992, NGL, c-erbB2, TKR1, HER-2, Hs.323910, c-erb B2, c-erbB-2, HER2, NEU, HER-2/neu	TK	EGFR	P04626	CHEMBL1824	7f3dae1d9db4a099bab8ae17f0ad40ac	1.0
ErbB3	Alias: ErbB3, MDA-BF-1, p180-ErbB3, Hs.199067, ERBB3, erbB3-S, c-erbB3, p85-sErbB3, p45-sErbB3, Hs.167387, Hs.167386, HER3, c-erbB-3, ErbB-3, MGC88033	TK	EGFR	P21860	CHEMBL5838	19629e98537ee0beb815b7da599ac0fc	1.0
BLK	Alias: MGC10442, BLK	TK	Src	P51451	CHEMBL2250	f57c75daba16cb3c1ed754a1ab8c2235	1.0
EGFR	Alias: ERBB, ERBB1, EGFR, mENA, EGFRvIII	TK	EGFR	P00533	CHEMBL203	b41987e56a7e61e8140240c8c641496c	1.0
ErbB4	Alias: p180erbB4, MGC138404, ERBB4, ErbB4, HER4	TK	EGFR	Q15303	CHEMBL3009	32dc0048643790d3d53bbb4a1f86e4b6	1.0
YES	Alias: HsT441, c-yes, C-YES, Yes, YES, P61-YES, YES1	TK	Src	P07947	CHEMBL2073	46feaaca702da32fdfd4ffa96b632f1c	1.0

Figure 8.8 | Graphical User Interface (left) and generated output (right) of the Python implementation of the DDM algorithm presented here. Only a SMILES string is required as input, output is provided as depicted on the right. The packaged executable as well as the original Python script have been made available online.

Methods

Training dataset selection

Several large datasets screening dozens of small molecule inhibitors against hundreds of kinases have appeared in literature. Five datasets stand out when judged by size, these are summarized in Table 8.1. The first large dataset (>10,000 data points) published was that of Karaman *et al.* in 2008.[5] 38 commercially available kinase inhibitors were screened *in vitro* against 287 distinct kinases. In 2011, three large datasets were reported. Metz *et al*.[7] published a brief communication describing their analysis of more than 250,000 data points. However, a large portion of these interactions have not been made public. Later that same year, two papers were simultaneously published by Anastassiadis *et al.* and Davis *et al.*[6,8] Both studies comprise known kinase inhibitors and are screened against large portions of the kinome. The latest addition to the field is that by Elkins *et al*., who screened the Published Kinase Inhibitor Set (PKIS) by GlaxoSmithKline in an *in vitro* assay against 224 kinases, including a number of mutants.[9,13] This dataset is particularly interesting, as the PKIS molecules are available free of charge for academic research, and the set contains the most drug like molecules at the largest coverage. For these reasons, the dataset by Elkins *et al.* was chosen as a starting point for our data-driven approach.

t-SNE for the Published Kinase Inhibitor Set

The curated dataset from ChEMBL 23[33] was used, as the original publication contained some duplicate molecules, and using the open source software package KNIME, the Morgan fingerprints (RD-Kit, 4096-bits, radius = 2) were generated.[33–35] These fingerprints were then clustered using the Python implementation of Barnes-Hut t-SNE.[36] Visual inspection of the embedding shows striking co-localization of the pre-defined chemotypes when the chemotype annotation by Elkins *et al.* is used post-hoc to colour the markers (Figure 8.1). To quantify the clustering quality, the embedding was analyzed using the DBSCAN algorithm, where the eps-parameter was optimized by maximizing the Silhouette Coefficient.[14] The best clustering was found for eps 0.9, for which the Adjusted Rand Index was 0.774. This unbiased clustering produced 33 clusters (Figure 8.1), 29 of which were significantly ($P < 0.0001$, hypergeometric test) enriched for a manually attributed chemotype. Of the 31 chemotypes annotated, 23 were fully comprised in one computationally assigned cluster. To inspect where it diverges from the human annotation, we inspected the cluster dominated by the 3-amino-pyrazolopyridines closer (Figure 8.1B). Inspection of the co-clustering of chemotypes shows that this cluster also contains all 3-amino-pyrazolopyridazines and the indazole-3-carboxamides, structurally very similar classes, as compounds **1**, **2** and **3** illustrate.

Nearest neighbour selection of molecules using t-SNE and initial target predictions

By appending a molecule to the PKIS dataset and regenerating the t-SNE mapping, simple Euclidian distances could be used to find its nearest neighbours. Selecting a set of neighbouring molecules should account for most chemical variation, leading to an 'average molecule' highly similar to our (new) molecule of interest.

For this set of neighbouring molecules, the interaction landscapes against 200 unique non-mutant kinases have been experimentally measured, which can be averaged to yield a predicted inhibition value for the new molecule for all these kinases. For the PKIS dataset inhibition was measured at 100 nM and 1 µM. The inhibition values measured at 1 µM were used as these have the highest information density and in many clinically relevant experiments low micromolar concentrations can (locally) be reached, thus warranting a screen for off-targets at this concentration.

Expanding target prediction using t-SNE to find most similar kinases

To be able to extrapolate from the kinases measured in the PKIS set to incorporate more of the kinome, again a t-SNE based approach was used. Of the entire kinome, the amino acid sequences of the kinase domains containing the ATP-binding pocket, were aligned and expressed using a fingerprint based on physicochemical properties, derived by Heil *et al*.[15] The resulting t-SNE embedding is shown in Figure 8.2 of the main text, and reproduced with annotation of the inclusion in the PKIS set (Figure S8.3). With this similarity mapping in hand, the prediction based on the training dataset could be expanded to theoretically include the whole human kinome, by considering neighbouring kinases as plausible additional targets. The distribution of the measured kinases in the PKIS set is visualized in Figure S8.3 and is rather well, but certainly not homogeneous. This means that for fair parts of the kinome no truly reliable prediction can be made based only on this set, as there are no close neighbouring kinases measured. In the workflow all kinases are however still included, as this will allow any dataset to be loaded as training set, without large adaptations.

The DDM workflow

The workflow envisaged was briefly described in Figure 8.3, and is depicted in more detail in Figure S8.1: a molecule of interest is converted, via its SMILES representation, into a 4096-bit fingerprint. This fingerprint is appended to the PKIS molecules and a 2D t-SNE embedding is generated, as in Figure 8.1A. Based on this embedding, the closest molecules are selected and their measured bioactivity is averaged and considered as the activity prediction. Then, using the t-SNE embedding of the kinase domains (Figure 8.2), the kinases most similar to the predicted kinases are appended to this prediction, which is then the output of the model. Since the t-SNE algorithm is inherently stochastic, this whole process is repeated *R* times, after which the initial outputs are weighed and finally returned as overall output. The model thus accepts any molecular string representation and returns a list of predicted kinase targets, with a confidence parameter based on the number of repetitions in which a specific target has been found. The PKIS molecules the prediction is based upon can also be viewed, to assess with a chemical eye whether the prediction is to be trusted.

Several optimizable parameters naturally arise in this workflow: the number of considered neighbouring molecules N_m, the number of considered neighbouring kinases N_k, the number of repetitions of the whole process *R*, and the cut-off value *C* of the mean inhibition above which an inhibitor is deemed active against that kinase.

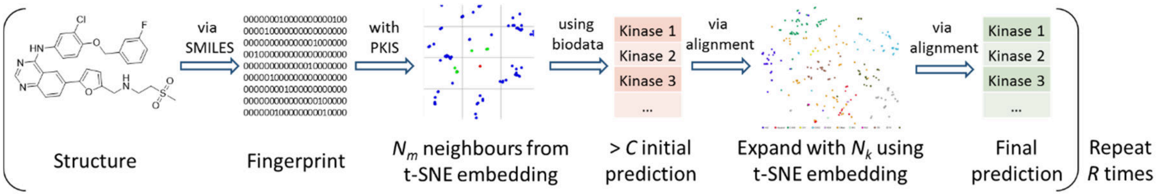


Figure S8.1 | Schematic overview of the DDM workflow. A chemical structure is converted into a Morgan fingerprint using the SMILES-string, and appended to the PKIS molecules set. After t-SNE embedding, the closest neighbour molecules (N_m) are identified and their biodata is averaged. A minimal average inhibition (*C*) is used as criterion to yield the initial prediction. The N_k kinases with the most similar kinase domain are then found in the kinase t-SNE embedding and appended to form the final prediction. This is repeated *R* times, weighing the final predictions to return a weighed final prediction.

Optimization of the model parameters

To optimize the model parameters and validate our model, the dataset generated by Karaman *et al*. was used as test set as this provided a comprehensive but diverse set of molecules, biochemically tested against a large set of kinases. For this test set, all K_d values below 1 µM were considered as actives, to mirror the training set, also measured at 1 µM. A multidimensional optimization was performed, initially varying the cut-off *C* and the two neighbouring values, N_m and N_k (Figure S8.2A and S8.2B). The number of repetitions was also optimized, as depicted in Figure S8.2C. The best predictions were found for 9 neighbouring molecules, 3 neighbouring kinases, and 10 repetitions. The cut-off percentage was found to be a valuable tuning parameter to either have a high sensitivity (low *C*) or rather a high positive predictive value (high *C*), which can be chosen depending on the projects specific demands. This is further illustrated by the ROC-curve in Figure 8.5. The mid-way value of *C* = 30% is recommended. The area under the ROC-curve is 0.76, scoring it as ‘fair’ according to standard criteria.

Comparison with state of the art methods

With this optimization of the model completed, the next step was to compare the new t-SNE based model with the state of the art in literature. To this end, three proteochemometric (PCM) models and a QSAR model were trained according to published procedures, and the DDM method was compared with those and a random model.[12,37] The result of this comparison is summarized in Figure 8.4, which shows that the DDM model is significantly better than both the random model and the QSAR, and performs similarly when compared to the PCM variants. A variant of the DDM model where the standard Tanimoto distances were used to find the nearest molecular neighbours in the first step of the workflow was also included in the comparison. t-SNE was used in the kinase lookup step. This DDM-Tanimoto variant performed rather well, and not significantly worse than our t-SNE based approach, for this test set.

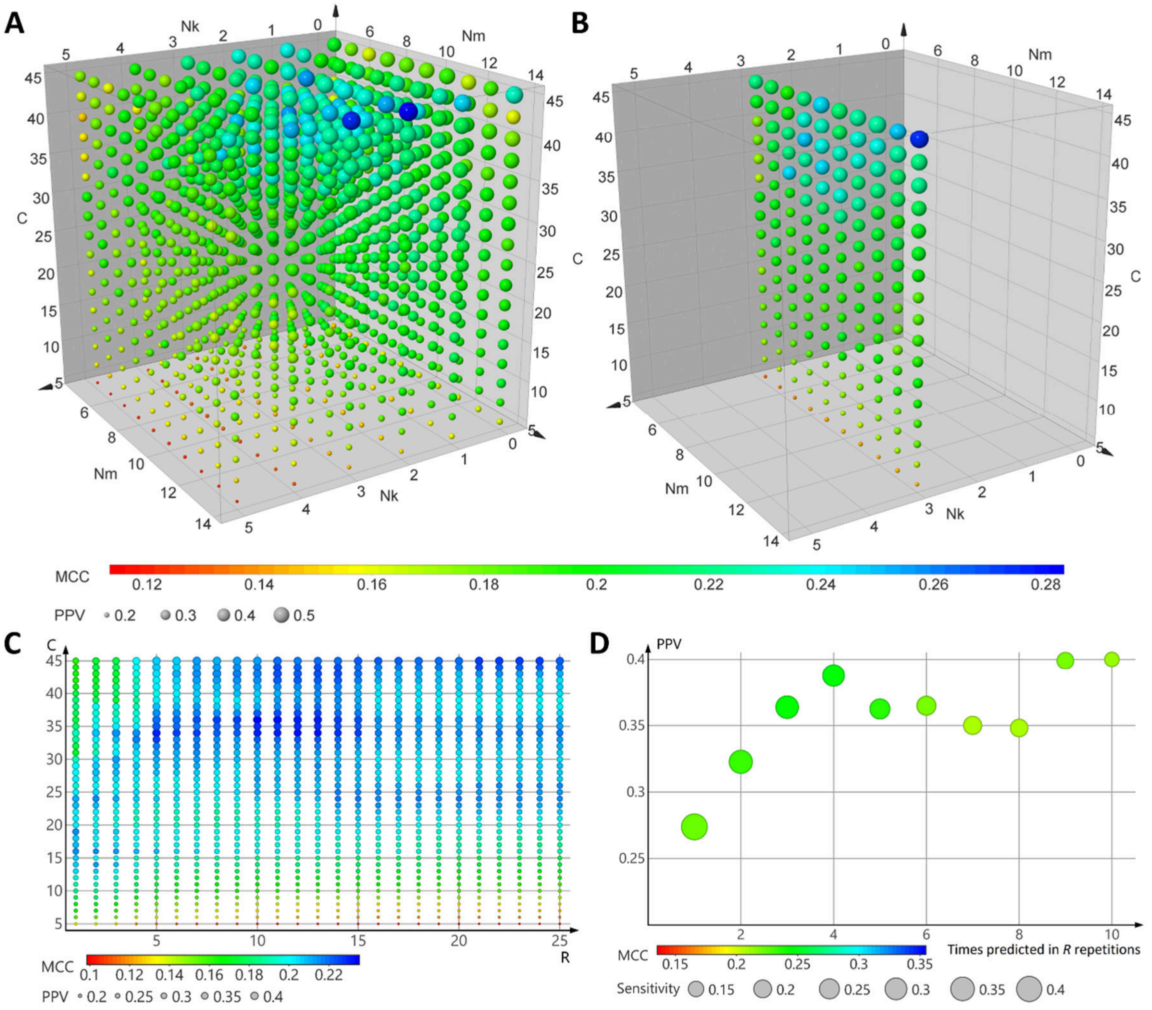


Figure S8.2 | Optimization of the model parameters using the Karaman *et al.* dataset as test set. Optimized parameters are the number of neighbouring molecules used (N_m), the number of neighbouring kinases used, N_k, and the cut-off inhibition percentage deemed active (*C*). Markers are sized based on the positive predictive value (PPV) and coloured based on the Matthews correlation coefficient (MCC). In A), all possible combinations are shown. In B), only the series for N_k = 3 is shown for clarity.

Bioactivity datasets used

Activity data for Karaman *et al.* and the Published Kinase Inhibitor Set by Elkins *et al.* was retrieved from the ChEMBL database version 23.[33,38] Data was retrieved from a local SQL install, directly from the website or through the KNIME extensions provided by the EMBL-EBI.

Kinase sequence information and bitstring

Sequence information of the kinase domains was retrieved from the KinBase situated on kinase.com, based on the paper by Manning *et al.*[16] Missing (pseudo)kinases were appended using data from the ChEMBL database version 23 and Uniprot.[39] The mapping of ChEMBL, Uniprot and KinBase identifiers was performed based on the KinBase website and extensive manual curation. The kinase domains were aligned using the online Clustal Omega tool provided by the EMBL-EBI.[40] The standard “Clustal w/o numbers” output generated was transformed to a bitstring using the amino acid fingerprints as provided in Heil *et al.*[15] with the following additions: alignment dashes (-), stops (*) and blanks (X) were all considered empty, represented by 23 0’s.

t-SNE algorithm

All t-SNE embeddings were generated with the Python Scikit-learn (v. 0.19) implementation of the Barnes-Hut t-SNE algorithm, either implemented in a ‘Python for KNIME’ node or as part of a Python script.[36]

DBSCAN algorithm

The Python implementation of DBSCAN, available through the Scikit-learn module Cluster (v. 0.19) was used.[36] The clustering quality metrics Silhouette Coefficient and Adjusted Rand Index were calculated with the Metrics module and the former was optimised by tuning the eps-parameter in steps of 0.05. For the PKIS inhibitors a minimal cluster size of 1 was chosen, for the more disperse plotted kinases the minimal cluster size was set to 10. Comparisons to manual attributions was done manually. Statistical evaluation was performed using Microsoft Excel 2013.

Cheminformatics tools

All molecular descriptors, molecular representations (SMILES, InChIKey) and fingerprints were generated using the RDKit software, either using the KNIME extensions or as the Python implementation.[34]

QSAR and PCM models

QSAR and PCM models were trained as has been described before but for PCM using the fully aligned sequences rather than the binding site and using classification rather than regression.[12,37] PCM models were trained at an activity cut-off of 30%, 40%, and 50%. Pipeline Pilot 2016 (version 16.2.0.58 by BioVia) was used to process the data and random forests were trained in R (version 3.3, package randomforest) using 500 trees, equalizing class sizes, and randomly sampling the square root of the total present descriptors at individual splits.[41] Protein descriptors used were the first three z-scales with a mean value for each sequence for each z-scale.[42] Chemical descriptors used were circular fingerprints with a diameter of 6 bonds (FCFP_6) and physicochemical descriptors as was done previously.[37,43]

PCM models were trained on the full set of kinases and PKIs. QSAR models were trained per kinase but only if at least 5 active and 10 inactive PKIs per kinase where present (at a cut off of 30%). If no QSAR was trained missing predictions were completed to avoid bias and have the ability to compare identical prediction counts. This was obtained using a random number generator in the range of 0-1 where > 0.5 was deemed 'active' as was done previously.[37]

Out of Bag error estimates of PCM models trained on the full set were 8.47% (30% cut off), 8.15 % (40% cut off), and 7.83 % (50% cut off).

Docking of 5 and 6 in FLT3 crystal structure

Docking and structure based modeling was performed in the Schrödinger suite.[44] For 5 a DFG-in model was constructed on the basis of 4RT7 and 3LCD, in a similar fashion as has been done before,[45] using the knowledge-based potential in Prime.[46,47] Subsequently 5 was docked into this model using induced fit docking.[48] For the induced fit docking a hydrogen bond constraint on the backbone of Cys96 was used. Since 6 shares the same substructure as quizartinib, the crystal structure of FLT3 co-crystalized with quizartinib (4RT7)[49] was used as a starting point. 6 was docked using Glide SP.[50] Ligand surfaces, structure renderings and 2D interaction plots were generated using Discovery Studio Visualizer v16.1.

Statistical methods

Clustering enrichment was analysed using a hypergeometric test and was calculated using Microsoft Excel 2013. Significance was attributed only if $P < 0.0001$.

In the comparison of the quality of prediction of the various prediction models a regular 2-way ANOVA with Tukey's multiple comparison test was performed using GraphPad Prism 7. Significance is attributed according to the standard GraphPad style: * $p < 0.0332$, ** $p < 0.0021$, *** $p < 0.0002$, **** $p < 0.0001$.

High Troughput Screening FLT3

20 nL of 2 mM inhibitor solution in DMSO (row 9-48) or control (DMSO, row 1-8) was dispensed in a 1536-well plate. 2 µL assay buffer (50 mM HEPES (pH 7.5), 1 mM EGTA, 10 mM $MgCl_2$, 0.01% Tween-20, 2 mM DTT) without protein was loaded in row 5-6 as negative controls. Rows 1-4 and 7-48 were charged with 2 µL 0.75 ng/µL FLT3 in assay buffer. The plates were spun down for 30 seconds at 187.5x *g* and incubated for 30 minutes in the dark. Subsequently, 2 µL of substrate solution was added to all wells (50 mM HEPES (pH 7.5), 1 mM EGTA, 10 mM $MgCl_2$, 0.01% Tween-20, 2 mM DTT, 600 µM ATP, 12.5 nM Lance TK-peptide, 4 nM Lance anti-phosphotyrosine). The plates were spun down for 30 seconds at 187.5x *g* and incubated for 90 minutes in the dark at RT. Plates were then read on the Envision plate reader (Excitation 337 nm (laser), Emission first filter 615 nm, second filter 665 nm). Data was analysed using ActivityBase. Final assay concentrations: 10 µM

inhibitor, 300 µM ATP, 0.365 ng/µL FLT3, 6.25 nM (0.5 K_M) Lance TK-peptide, 2 nM Lance anti-phosphotyrosine, 0.5% DMSO.

In vitro FRET based FLT3 assay

In a 384-wells plate, 5 µL kinase+petide mix (0.06 ng/µL FLT3, 200 nM Lance TK-peptide) in assay buffer (50 mM HEPES (pH 7.5), 1 mM EGTA, 10 mM $MgCl_2$, 0.01% Tween-20, 2 mM DTT) was dispensed. Seperately inhibitor solutions (10 µM – 0.1 pM) were prepared in assay buffer containing 400 µM ATP and 1% DMSO. 5 µL of these solutions was dispensed and the plate was incubated for 90 minutes in the dark. After 90 minutes the reaction was quenched by the addition of 10 µL of 20 mM EDTA containing 4 nM Lance anti-phosphotyrosine. After thorough mixing it was incubated for 60 minutes in the dark. The FRET fluorescence was measured on a Tecan Infinite M1000 Pro plate reader (excitation 320 nm, emission first filter 615 nm, second filter 665 nm). Data was processed using Microsoft Excel 2013, pIC_{50} values were fitted using GraphPad Prism 7.0. Final assay concentrations during phosphorylation: 200 µM ATP, 0.03 ng/µL FLT3, 100 nM Lance TK-peptide, 0.5% DMSO)

In situ testing of kinase inhibitors

MV4:11 cells were grown in IMDM with 10% fetal bovine serum at 37 °C under 5% CO_2. For viability assays, 10,000 cells were seeded per well in a 96-wells plate and inhibitors were added at the indicated concentration. Three days later, cell viability was measured using the Cell Titer Blue viability assay (Promega), fluorescence was measured using the Clariostar (BMG Labtech). Relative survival was normalized to the untreated control and corrected for background signal.

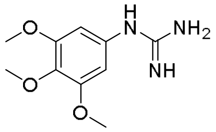
Synthesis of *in situ* tested kinase inhibitors

General remarks

All reactions were performed using oven- or flame-dried glassware and dry solvents. Reagents were purchased from Sigma-Aldrich, Acros, and Merck and used without further purification unless noted otherwise. All moisture sensitive reactions were performed under a nitrogen atmosphere.

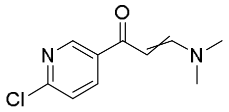
^1H and ^{13}C NMR spectra were recorded on a Bruker AV-400 (400 MHz). Used software for interpretation of NMR-data was Bruker TopSpin 1.3 and MestreNova 11.0. Chemical shift values are reported in ppm with tetramethylsilane or solvent resonance as the internal standard ($CDCl_3$: δ 7.26 for ^1H, δ 77.16 for ^{13}C; DMSO-*d6*: δ 2.50 for ^1H, δ 39.52 for ^{13}C). Data are reported as follows: chemical shifts (δ), multiplicity (s = singlet, d = doublet, dd = double doublet, td = triple doublet, t = triplet, q = quartet, bs = broad singlet, m = multiplet), coupling constants *J* (Hz), and integration. Liquid chromatography was performed on a Finnigan Surveyor LC/MS system, equipped with a C18 column. Flash chromatography was performed using SiliCycle silica gel type SiliaFlash P60 (230−400 mesh). TLC analysis was performed on Merck silica gel 60/Kieselguhr F254, 0.25 mm. Compounds were visualized using $KMnO_4$ stain (K_2CO_3 (40 g), $KMnO_4$ (6 g), and water (600 mL)) or CAM stain ($Ce(NH_4)_4(SO_4)_4{\cdot}2H_2O$ (ceric ammonium sulfate: 10.0 g); ammonium molybdate (25 g); conc. H_2SO_4 (100 mL); H_2O (900 mL)). Preparative HPLC (Waters, 515 HPLC pump M; Waters, 515 HPLC pump L; Waters, 2767 sample manager; Waters SFO System Fluidics Organizer; Waters Acquity Ultra Performance LC, SQ Detector; Waters Binary Gradient Module) was performed on a Waters XBridgeTM column (5 µM C18, 150 x 19 mm). Diode detection was done between 210 and 600 nm. Gradient: ACN in (H_2O + 0.2% TFA).

1-(3,4,5-Trimethoxyphenyl)guanidine (7)



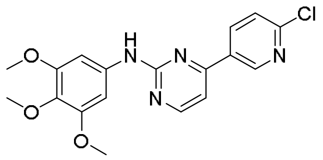
3,4,5-trimethoxyaniline (500 mg, 2.73 mmol) and cyanamide (574 mg, 13.6 mmol) were dissolved in ethanol (15 mL) before nitric acid (69%$_{wt}$, 0.20 mL, 3.0 mmol) was added. The mixture was refluxed for 40 h and concentrated *in vacuo*. The resulting residue was suspended in diethylether (10 mL) and kept at 4 °C overnight. It was then filtered and air dried. The nitrate salt of the title compound was obtained as a dark purple solid (0.60 g, 2.1 mmol, 76%). ^1H NMR (400 MHz, DMSO) δ 9.51 (bs, 1H), 8.73 – 8.12 (m, 2H), 6.55 (s, 2H), 5.44 (s, 2H), 3.77 (s, 6H), 3.65 (s, 3H). ^{13}C NMR (101 MHz, DMSO) δ 161.53, 155.86, 153.38, 130.67, 102.91, 60.06, 56.01.

1-(6-Chloropyridin-3-yl)-3-(dimethylamino)prop-2-en-1-one (8)



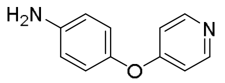
1-(6-chloropyridin-3-yl)ethan-1-one (1.00 g, 6.43 mmol) was dissolved in toluene (20 mL) and dimethylformamide diethylacetal (1.65 mL, 9.64 mmol) was added. The mixture was stirred for 16 h at 80 °C. TLC showed near complete conversion. The mixture was cooled to 50 °C and slowly concentrated at reduced pressure. The resulting crude residue was dissolved in warm toluene and pentane was slowly added. A yellow precipitate formed which was filtered off and rinsed with pentane. After drying this yielded the title compound (1.08 g, 5.12 mmol, 80%) as a yellow solid. ^{1}H NMR (400 MHz, CDCl$_3$) δ 8.85 (dd, *J* = 0.7, 2.5 Hz, 1H), 8.17 (dd, *J* = 2.5, 8.3 Hz, 1H), 7.86 (d, *J* = 12.2 Hz, 1H), 7.38 (dd, *J* = 0.8, 8.2 Hz, 1H), 5.63 (d, *J* = 12.1 Hz, 1H), 3.20 (s, 3H), 2.97 (s, 3H). ^{13}C NMR (101 MHz, CDCl$_3$) δ 184.90, 155.07, 153.38, 149.08, 138.10, 134.63, 124.10, 91.54, 45.46, 37.60.

4-(6-Chloropyridin-3-yl)-*N*-(3,4,5-trimethoxyphenyl)pyrimidin-2-amine (5)



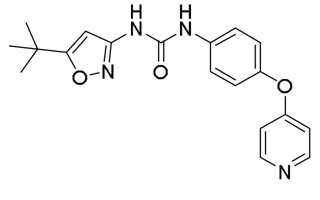
To a solution of 7 (376 mg, 1.30 mmol) and 8 (250 mg, 1.19 mmol) in ethanol (20 mL) K$_2$CO$_3$ (492 mg, 3.56 mmol) was added. The resulting mixture was refluxed for 20 h. The reaction was quenched by the addition of saturated aqueous Na$_2$CO$_3$ (15 mL) and the water layer was extracted with EtOAc (3x 25 mL). The combined organic layers were washed with brine, dried (MgSO$_4$), filtered and concentrated *in vacuo*. The resulting crude was purified using silica column chromatography (50-70% EtOAc/pentane). The resulting yellow solid was crystallised from warm diethylether to yield the title compound (137 mg, 0.37 mmol, 31%) as light yellow crystals. ^{1}H NMR (400 MHz, CDCl$_3$) δ 9.07 (dd, *J* = 0.7, 2.5 Hz, 1H), 8.52 (d, *J* = 5.2 Hz, 1H), 8.32 (dd, *J* = 2.5, 8.3 Hz, 1H), 7.46 (dd, *J* = 0.7, 8.4 Hz, 1H), 7.38 (s, 1H), 7.14 (d, *J* = 5.2 Hz, 1H), 6.99 (s, 2H), 3.90 (s, 6H), 3.85 (s, 3H). ^{13}C NMR (101 MHz, CDCl$_3$) δ 161.53, 160.28, 159.11, 153.74, 153.47, 148.62, 137.11, 135.30, 131.78, 124.52, 108.15, 97.42, 61.17, 56.26.

4-(Pyridin-4-yloxy)aniline (9)



4-aminophenol (2.18 g, 20 mmol), 4-chloropyridine (3.15 g, 21 mmol) and NaOH (2.0 g, 50 mmol) were dissolved in DMSO (50 mL) and heated to 100 °C for 17 h. The mixture was cooled to RT and poured into ice water (300 mL). This was extracted with 10% MeOH in chloroform (4x 150 mL). The combined organic layers were washed with brine (2x 150 mL), dried (Na$_2$SO$_4$), filtered and concentrated *in vacuo*. The residue was flushed over a silica pad with pure EtOAc and concentrated to yield the title compound as a white solid (2.5 g, 13 mmol, 65%). ^{1}H NMR (400 MHz, CDCl$_3$) δ 8.51 – 8.33 (m, 2H), 6.97 – 6.85 (m, 2H), 6.85 – 6.77 (m, 2H), 6.77 – 6.62 (m, 2H), 3.50 (s, 2H). ^{13}C NMR (101 MHz, CDCl$_3$) δ 165.89, 151.25, 145.80, 144.25, 122.12, 116.36, 111.70.

1-(5-(*tert*-Butyl)isoxazol-3-yl)-3-(4-(pyridin-4-yloxy)phenyl)urea (6)



3-amino-5-tert-butylisooxazole (0.20 g, 1.4 mmol) was dissolved in DCM (14 mL) and DIPEA (0.50 mL, 2.9 mmol) was added. The mixture was cooled to 0 °C and triphosgene (0.42 g, 1.4 mmol) was added. After stirring the mixture for 19 h at RT it was refluxed for 1 h before the reaction was quenched by the addition of saturated aqueous NaHCO$_3$ (10 mL). The water layer was separated and extracted with DCM (3x 10 mL). The combined organic layers were washed with brine, dried (Na$_2$SO$_4$) and filtered. The solvent was removed under reduced pressure and the crude isocyanate was dissolved in 1,4-dioxane (14 mL). 9 (0.29 g, 1.6 mmol) was added to the solution and the mixture was heated to 110 °C for 2.5 h. After cooling to RT the mixture was diluted with DCM (30 mL) and saturated aqueous NaHCO$_3$ (30 mL) was added. The water layer was separated and extracted with DCM (3x 30 mL). The combined organic layers were washed with brine, dried (Na$_2$SO$_4$), filtered and concentrated *in vacuo*. The resulting yellow oil was purified by preparative HPLC to yield the title compound as a slightly yellow oil (0.22 g, 0.62 mmol, 44%). ^{1}H NMR (400 MHz, CDCl$_3$) δ 9.41 (s, 1H), 8.67 (s, 1H), 8.47 (dd, *J* = 1.5, 5.1 Hz, 2H), 7.63 – 7.54 (m, 2H), 7.10 – 7.03 (m, 2H), 6.89 – 6.81 (m, 2H), 5.95 (s, 1H), 1.36 (s, 9H). ^{13}C NMR (101 MHz, CDCl$_3$) δ 181.60, 165.38, 158.45, 152.58, 151.22, 149.95, 135.58, 122.15, 121.60, 112.17, 91.94, 33.08, 28.75.

Supplementary Tables and Figures

Table S8.1 | Overview of prediction qualities split per compound. Data corresponds to *R* = 10 and *C* = 30%. PPV = positive predictive value, NPV = negative predictive value, MCC = Matthew's correlation coefficient.

CHEMBL ID	True positives	False positives	True negatives	False negatives	Sensi-tivity	Speci-ficity	PPV	NPV	MCC
CHEMBL10	4	8	254	13	0.24	0.97	0.33	0.95	0.24
CHEMBL101253	7	41	230	1	0.88	0.85	0.15	1.00	0.32
CHEMBL103667	5	21	228	25	0.17	0.92	0.19	0.90	0.09
CHEMBL119385	2	11	265	1	0.67	0.96	0.15	1.00	0.31
CHEMBL124660	9	66	202	2	0.82	0.75	0.12	0.99	0.25
CHEMBL1336	20	28	217	14	0.59	0.89	0.42	0.94	0.41
CHEMBL1421	19	5	207	48	0.28	0.98	0.79	0.81	0.40
CHEMBL14762	0	49	227	3	0.00	0.82	0.00	0.99	-0.05
CHEMBL1721885	35	29	147	69	0.34	0.84	0.55	0.68	0.20
CHEMBL191003	20	43	174	42	0.32	0.80	0.32	0.81	0.12
CHEMBL215152	1	5	258	15	0.06	0.98	0.17	0.95	0.07
CHEMBL221959	1	46	226	6	0.14	0.83	0.02	0.97	-0.01
CHEMBL223360	26	62	187	4	0.87	0.75	0.30	0.98	0.41
CHEMBL24828	20	16	217	26	0.43	0.93	0.56	0.89	0.41
CHEMBL259084	20	65	191	3	0.87	0.75	0.24	0.98	0.37
CHEMBL261849	4	74	201	0	1.00	0.73	0.05	1.00	0.19
CHEMBL278041	2	14	249	14	0.13	0.95	0.13	0.95	0.07
CHEMBL296468	2	3	250	24	0.08	0.99	0.40	0.91	0.14
CHEMBL31965	3	3	261	12	0.20	0.99	0.50	0.96	0.29
CHEMBL388978	57	3	42	179	0.24	0.93	0.95	0.19	0.16
CHEMBL428690	2	19	228	30	0.06	0.92	0.10	0.88	-0.02
CHEMBL440084	0	12	263	4	0.00	0.96	0.00	0.99	-0.03
CHEMBL477772	22	56	185	16	0.58	0.77	0.28	0.92	0.26
CHEMBL483321	2	4	272	1	0.67	0.99	0.33	1.00	0.46
CHEMBL522892	24	31	183	41	0.37	0.86	0.44	0.82	0.24
CHEMBL535	44	30	117	89	0.33	0.80	0.59	0.57	0.14
CHEMBL553	3	3	253	20	0.13	0.99	0.50	0.93	0.23
CHEMBL554	2	0	277	0	1.00	1.00	1.00	1.00	1.00
CHEMBL558752	10	72	185	12	0.45	0.72	0.12	0.94	0.10
CHEMBL572878	2	11	199	67	0.03	0.95	0.15	0.75	-0.05
CHEMBL572881	6	16	242	15	0.29	0.94	0.27	0.94	0.22
CHEMBL573339	0	58	219	2	0.00	0.79	0.00	0.99	-0.04
CHEMBL574738	48	12	165	54	0.47	0.93	0.80	0.75	0.47
CHEMBL607707	11	45	203	20	0.35	0.82	0.20	0.91	0.14
CHEMBL608533	11	10	169	89	0.11	0.94	0.52	0.66	0.10
CHEMBL91829	1	12	240	26	0.04	0.95	0.08	0.90	-0.01
CHEMBL939	3	15	254	7	0.30	0.94	0.17	0.97	0.18
CHEMBL941	10	115	153	1	0.91	0.57	0.08	0.99	0.19
Average	12.05	29.29	211.58	26.18	0.38	0.88	0.32	0.89	0.21

Table S8.2 | Summary of the off-target screen performed for the *in situ* tested compounds **5** and **6**. Data are averages of two duplicates, and are percentages activity compared to vehicle control.

5				6			
Gene Name	Uniprot accession	CHEMBL ID	Remaining activity (%)	Gene Name	Uniprot accession	CHEMBL ID	Remaining activity (%)
ABL1	P00519	CHEMBL1862	26	ABL1	P00519	CHEMBL1862	23
ABL2	P42684	CHEMBL4014	12	ABL2	P42684	CHEMBL4014	28
AurA	O14965	CHEMBL4722	2	AurA	O14965	CHEMBL4722	60
AurB	Q96GD4	CHEMBL2185	2	AurB	Q96GD4	CHEMBL2185	60
AurC	Q9UQB9	CHEMBL3935	8	AurC	Q9UQB9	CHEMBL3935	91
BLK	P51451	CHEMBL2250	34	DDR1	Q08345	CHEMBL5319	7
BRK	Q13882	CHEMBL4601	61	DDR2	Q16832	CHEMBL5122	1
EGFR	P00533	CHEMBL203	96	EphA3	P29320	CHEMBL4954	9
EphA4	P54764	CHEMBL3988	72	EphA4	P54764	CHEMBL3988	28
EphA5	P54756	CHEMBL3987	52	EphA5	P54756	CHEMBL3987	7
EphB2	P29323	CHEMBL3290	14	EphA7	Q15375	CHEMBL4602	-2
EphB3	P54753	CHEMBL4901	100	EphB1	P54762	CHEMBL5072	3
ErbB2	P04626	CHEMBL1824	94	EphB2	P29323	CHEMBL3290	5
ErbB4	Q15303	CHEMBL3009	108	EphB3	P54753	CHEMBL4901	55
FER	P16591	CHEMBL3982	52	FER	P16591	CHEMBL3982	86
FES	P07332	CHEMBL5455	19	FES	P07332	CHEMBL5455	86
FGFR1	P11362	CHEMBL3650	58	FGFR1	P11362	CHEMBL3650	6
FGFR3	P22607	CHEMBL2742	48	FGFR3	P22607	CHEMBL2742	29
FGFR4	P22455	CHEMBL3973	91	FGFR4	P22455	CHEMBL3973	70
FGR	P09769	CHEMBL4454	30	FLT1	P17948	CHEMBL1868	1
FLT1	P17948	CHEMBL1868	1	FLT3	P36888	CHEMBL1974	1
FLT3	P36888	CHEMBL1974	3	FLT4	P35916	CHEMBL1955	-1
FLT4	P35916	CHEMBL1955	1	FMS	P07333	CHEMBL1844	-1
FMS	P07333	CHEMBL1844	-1	KDR	P35968	CHEMBL279	7
FYN	P06241	CHEMBL1841	9	KIT	P10721	CHEMBL1936	7
HCK	P08631	CHEMBL3234	17	MUSK	O15146	CHEMBL5684	13
KDR	P35968	CHEMBL279	5	PDGFRa	P16234	CHEMBL2007	15
KIT	P10721	CHEMBL1936	2	PDGFRb	P09619	CHEMBL1913	9
LCK	P06239	CHEMBL258	15	RET	P07949	CHEMBL2041	-2
LYN	P07948	CHEMBL3905	26	TIE2	Q02763	CHEMBL4128	1
PDGFRa	P16234	CHEMBL2007	47	TRKA	P04629	CHEMBL2815	5
PDGFRb	P09619	CHEMBL1913	46	TRKB	Q16620	CHEMBL4898	0
RET	P07949	CHEMBL2041	53	TRKC	Q16288	CHEMBL5608	-2
SRC	P12931	CHEMBL267	4				
TNK1*	Q13470	CHEMBL5334	-				
YES	P07947	CHEMBL2073	10				

*TNK1 was not available in the KinomeProfiler™

Table S8.3 | Overview of the molecular structures and biological activity for the 20 purchased Enamine small molecules. Biological data are averages of two duplicates, and are percentages activity remaining compared to vehicle control.

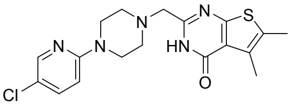
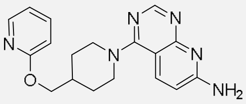
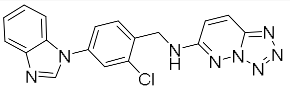
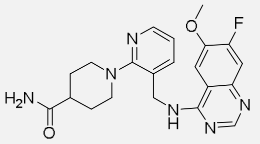
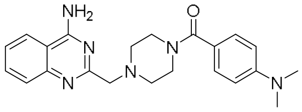
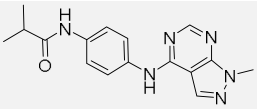
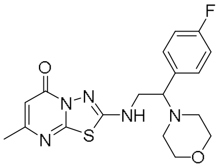
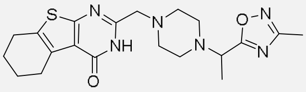
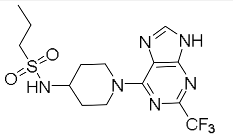
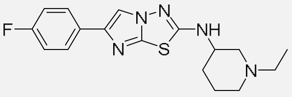
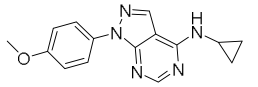
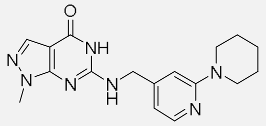
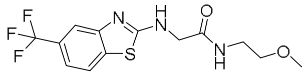
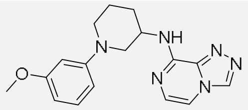
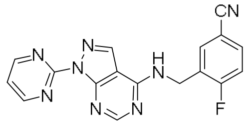
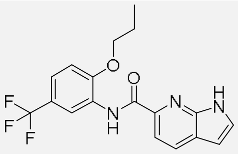
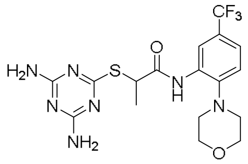
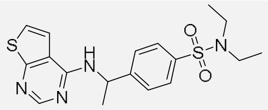
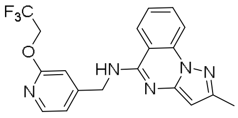
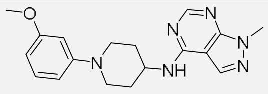
Predicted for PAK2			
Enamine ID	Structure	HIPK3 inhibition (%)	PAK2 inhibition (%)
Z1272746053		-4	-6
Z1537171258		2	-5
Z1557057955		3	1
Z1695667186		-2	8
Z226653560		-2	5
Z228788294		-3	2
Z237652740		2	3
Z316373096		7	5
Z367648718		1	10
Z785216724		-1	8

Table S8.3 (continued) | Overview of the molecular structures and biological activity for the 20 purchased Enamine small molecules. Biological data are averages of two duplicates, and are percentages activity remaining compared to vehicle control.

Predicted for HIPK3			
Enamine ID	**Structure**	**HIPK3 inhibition (%)**	**PAK2 inhibition (%)**
Z1041114058		-6	11
Z1102997288		3	-1
Z1272153794		-4	-8
Z1281728869		-1	2
Z1502786962		-6	9
Z1625107708		1	9
Z17559373		-1	-11
Z31057533		-2	6
Z644918986		2	13
Z909646456		0	-2

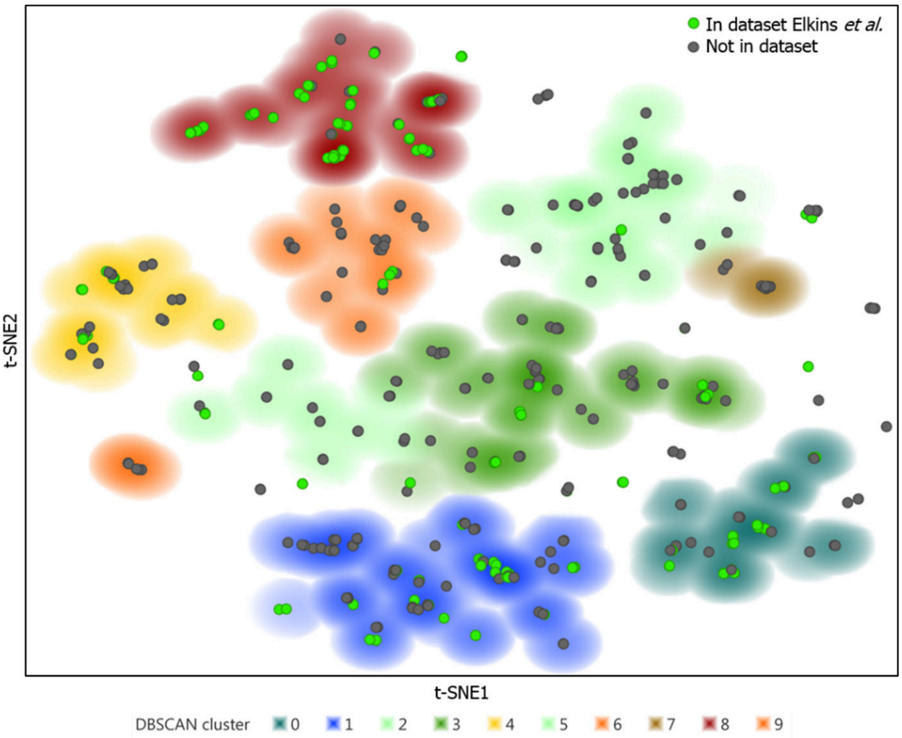


Figure S8.3 | t-SNE embedding of physicochemical fingerprint of the kinase domains of 535 human kinases as in Figure 8.2. Markers are coloured based on the presence (green) or absence (grey) in the Elkins *et al.* PKIS dataset.

References

1. Johnson, L. N. & Lewis, R. J. Structural Basis for Control by Phosphorylation. *Chem. Rev.* **101**, 2209–2242 (2001).
2. Adams, J. A. Kinetic and Catalytic Mechanisms of Protein Kinases. *Chem. Rev.* **101**, 2271–2290 (2001).
3. Klaeger, S. *et al.* The target landscape of clinical kinase drugs. *Science* **358**, (2017).
4. Müller, S., Chaikuad, A., Gray, N. S. & Knapp, S. The ins and outs of selective kinase inhibitor development. *Nat. Chem. Biol.* **11**, 818–821 (2015).
5. Karaman, M. W. *et al.* A quantitative analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **26**, 127–132 (2008).
6. Davis, M. I. *et al.* Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **29**, 1046–1051 (2011).
7. Metz, J. T. *et al.* Navigating the kinome. *Nat. Chem. Biol.* **7**, 200–202 (2011).
8. Anastassiadis, T., Deacon, S. W., Devarajan, K., Ma, H. & Peterson, J. R. Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. *Nat. Biotechnol.* **29**, 1039–1045 (2011).
9. Elkins, J. M. *et al.* Comprehensive characterization of the Published Kinase Inhibitor Set. *Nat. Biotechnol.* **34**, 95–103 (2015).
10. Miduturu, C. V. *et al.* High-Throughput Kinase Profiling: A More Efficient Approach toward the Discovery of New Kinase Inhibitors. *Chem. Biol.* **18**, 868–879 (2011).
11. Merget, B., Turk, S., Eid, S., Rippmann, F. & Fulle, S. Profiling Prediction of Kinase Inhibitors: Toward the Virtual Assay. *J. Med. Chem.* **60**, 474–485 (2017).
12. Christmann-Franck, S. *et al.* Unprecedently Large-Scale Kinase Inhibitor Set Enabling the Accurate Prediction of Compound–Kinase Activities: A Way toward Selective Promiscuity by Design? *J. Chem. Inf. Model.* **56**, 1654–1675 (2016).
13. Drewry, D. H., Willson, T. M. & Zuercher, W. J. Seeding collaborations to advance kinase science with the GSK Published Kinase Inhibitor Set (PKIS). *Curr. Top. Med. Chem.* **14**, 340–2 (2014).
14. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. in *KDD-96 Proceedings* 226–231 (1996).
15. Heil, B., Ludwig, J., Lichtenberg-Frate, H. & Lengauer, T. Computational recognition of potassium channel sequences. *Bioinformatics* **22**, 1562–1568 (2006).
16. Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The Protein Kinase Complement of the Human Genome. *Science* **298**, 1912–1934 (2002).
17. Murphy, J. M. *et al.* A robust methodology to subclassify pseudokinases based on their nucleotide-binding properties. *Biochem. J.* **457**, 323–334 (2014).
18. Murphy, J. M. *et al.* The Pseudokinase MLKL Mediates Necroptosis via a Molecular Switch Mechanism. *Immunity* **39**, 443–453 (2013).
19. Josso, N. & Clemente, N. di. Transduction pathway of anti-Müllerian hormone, a sex-specific member of the TGF-β family. *Trends Endocrinol. Metab.* **14**, 91–97 (2003).
20. Eid, S., Turk, S., Volkamer, A., Rippmann, F. & Fulle, S. KinMap: a web-based tool for interactive navigation through human kinome data. *BMC Bioinformatics* **18**, 16 (2017).
21. Larrosa-Garcia, M. & Baer, M. R. FLT3 Inhibitors in Acute Myeloid Leukemia: Current Status and Future Directions. *Mol. Cancer Ther.* **16**, 991–1001 (2017).
22. Charnley, A. K. *et al.* Crystal structures of human RIP2 kinase catalytic domain complexed with ATP-competitive inhibitors: Foundations for understanding inhibitor selectivity. *Bioorg. Med. Chem.* **23**, 7000–7006 (2015).
23. Enamine Ltd. Enamine. Available at: http://www.enamine.net/.
24. Uitdehaag, J. C. & Zaman, G. J. A theoretical entropy score as a single value to express inhibitor selectivity. *BMC Bioinformatics* **12**, 94 (2011).
25. Fokkelman, M. *et al.* Cellular adhesome screen identifies critical modulators of focal adhesion dynamics, cellular traction forces and cell migration behaviour. *Sci. Rep.* **6**, 31707 (2016).
26. Zhang, Y. *et al.* IGF1R signaling drives antiestrogen resistance through PAK2/PIX activation in luminal breast cancer. *Oncogene* **37**, 1869–1884 (2018).
27. Scannell, J. W. & Bosley, J. When Quality Beats Quantity: Decision Theory, Drug Discovery, and the Reproducibility Crisis. *PLoS One* **11**, e0147215 (2016).
28. Sorgenfrei, F. A., Fulle, S. & Merget, B. Kinome-Wide Profiling Prediction of Small Molecules. *ChemMedChem* **13**, 495–499 (2018).
29. Van Der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
30. Reutlinger, M. & Schneider, G. Nonlinear dimensionality reduction and mapping of compound libraries for drug discovery. *J. Mol. Graph. Model.* **34**, 108–117 (2012).
31. Zarrinkar, P. P. *et al.* AC220 is a uniquely potent and selective inhibitor of FLT3 for the treatment of acute myeloid leukemia (AML). *Blood* **114**, 2984–92 (2009).
32. Janssen, A. P. A. Drug Discovery Maps on GitHub. (2018). Available at: https://github.com/APAJanssen/DrugDiscoveryMaps/.
33. EMBL-EBI. CHEMBL database release 23. (2017). doi:10.6019/CHEMBL.database.23

34. Landrum, G. RDKit: Open-source cheminformatics; http://www.rdkit.org.
35. Berthold, M. R. *et al.* KNIME: The Konstanz Information Miner. in *Data Analysis, Machine Learning and Applications* (ed. Preisach C., Burkhardt H., Schmidt-Thieme L., D. R.) 319–326 (Springer, Berlin, Heidelberg, 2008).
36. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830 (2011).
37. Lenselink, E. B. *et al.* Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *J. Cheminform.* 9, 45 (2017).
38. Gaulton, A. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40, D1100–D1107 (2012).
39. Consortium, T. U. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 45, D158–D169 (2017).
40. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7, 539 (2011).
41. Breiman, L. Random Forests. *Mach. Learn.* 45, 5–32 (2001).
42. van Westen, G. J. *et al.* Benchmarking of protein descriptor sets in proteochemometric modeling (part 2): modeling performance of 13 amino acid descriptor sets. *J. Cheminform.* 5, 42 (2013).
43. Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* 50, 742–754 (2010).
44. Schrödinger Suite: Maestro. (2017).
45. Ke, Y.-Y. *et al.* Homology modeling of DFG-in FMS-like tyrosine kinase 3 (FLT3) and structure-based virtual screening for inhibitor identification. *Sci. Rep.* 5, 11702 (2015).
46. Jacobson, M. P. *et al.* A hierarchical approach to all-atom protein loop prediction. *Proteins Struct. Funct. Bioinforma.* 55, 351–367 (2004).
47. Jacobson, M. P., Friesner, R. A., Xiang, Z. & Honig, B. On the role of the crystal environment in determining protein side-chain conformations. *J. Mol. Biol.* 320, 597–608 (2002).
48. Sherman, W., Day, T., Jacobson, M. P., Friesner, R. A. & Farid, R. Novel Procedure for Modeling Ligand/Receptor Induced Fit Effects. *J. Med. Chem.* 49, 534–553 (2006).
49. Smith, C. C. *et al.* Characterizing and Overriding the Structural Mechanism of the Quizartinib-Resistant FLT3 ‘Gatekeeper’ F691L Mutation with PLX3397. *Cancer Discov.* 5, 668–679 (2015).
50. Friesner, R. A. *et al.* Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* 47, 1739–1749 (2004).