Combining structural and bioactivity-based fingerprints improves prediction performance and scaffold-hopping capability

Oliver Laufkötter, Noé Sturm, Jürgen Bajorath, Ola Engkvist, Hongming Chen

Submitted date: 15/02/2019 • Posted date: 15/02/2019
Licence: CC BY-NC-ND 4.0

Citation information: Laufkötter, Oliver; Sturm, Noé; Bajorath, Jürgen; Engkvist, Ola; Chen, Hongming (2019): Combining structural and bioactivity-based fingerprints improves prediction performance and scaffold-hopping capability. ChemRxiv. Preprint.

This study aims at improving upon existing activity predictions methods by augmenting chemical structure fingerprints with bio-activity based fingerprints derived from high-throughput screening (HTS) data (HTSFPs). The HTSFPs were generated from HTS data obtained from PubChem and combined with an ECFP4 structural fingerprint. The combined experimental and structural fingerprint (CESFP) was benchmarked against the individual ECFP4 and HTSFP fingerprints. Results showed that the CESFP has improved predictive performance as well as scaffold hopping capability. The CESFP identified unique compounds compared to both the ECFP4 and the HTSFP fingerprint indicating synergistic effects between the two fingerprints. A feature importance analysis showed that a small subset of the HTSFP features contribute most to the overall performance of the CESFP. This combined approach allows for activity prediction of compounds with only sparse HTSFPs due to the supporting effect from the structural fingerprint.

File list (3)

| File name                                      | Size     | Download link     |
|------------------------------------------------|----------|-------------------|
| CESFP_Manuscript_Laufkotter.docx               | 1.87 MiB | view on ChemRxiv  |
| Supplementary_Data.docx                        | 3.18 MiB | download file     |
| Pubchem_assay_list.txt                         | 3.81 KiB | view on ChemRxiv  |
Combining structural and bioactivity-based fingerprints improves prediction performance and scaffold-hopping capability

Oliver Laufkötter\textsuperscript{1,2}, Noé Sturm\textsuperscript{3}, Jürgen Bajorath\textsuperscript{2}, Hongming Chen\textsuperscript{1}, and Ola Engkvist\textsuperscript{1}.

\textsuperscript{1}Hit Discovery, Discovery Sciences, IMED Biotech Unit, AstraZeneca, Gothenburg, Sweden
\textsuperscript{2}BIT, University of Bonn, Germany

Corresponding authors: Oliver Laufkötter: oliver.laufkotter@gmail.com, Ola Engkvist: ola.engkvist@astrazeneca.com

Abstract
This study aims at improving upon existing activity predictions methods by augmenting chemical structure fingerprints with bio-activity based fingerprints derived from high-throughput screening (HTS) data (HTSFPs). The HTSFPs were generated from HTS data obtained from PubChem and combined with an ECFP4 structural fingerprint. The combined experimental and structural fingerprint (CESFP) was benchmarked against the individual ECFP4 and HTSFP fingerprints. Results showed that the CESFP has improved predictive performance as well as scaffold hopping capability. The CESFP identified unique compounds compared to both the ECFP4 and the HTSFP fingerprint indicating synergistic effects between the two fingerprints. A feature importance analysis showed that a small subset of the HTSFP features contribute most to the overall performance of the CESFP. This combined approach allows for activity prediction of compounds with only sparse HTSFPs due to the supporting effect from the structural fingerprint.

Keywords
Machine learning, random forest, high throughput screening, activity prediction, HTSFP, ECFP, circular fingerprints, scaffold hopping.
Introduction

The traditional and most intuitive method of predicting compound activity is through the use of structure activity relationship (SAR) models. Logically, compounds with similar structural features or scaffolds would express similar activities. While SAR-based activity predictions are a practical and often effective method, the predictions made are based on structural similarity and therefore are inherently limited in structural diversity. This limits the scaffold-hopping potential or exploration of chemical space and impedes the identification of novel active compounds. To bypass the drawbacks of SAR models, historical experimental data can be used to build models and make compound property predictions independent of chemical structural information. Such predictive models have been built using experimental data obtained from various sources, e.g. transcriptomics, cell imaging, or high throughput screening (HTS). HTS is a method used for large scale testing of compound libraries, containing up to five million compounds, against a single target. HTS has become feasible during the last three decades due to advances in process automation along with the development of new technologies. HTS is a resource-intensive process, which usually only identifies a very small portion of active compounds. To reduce resource costs in HTS, compound activity prediction methods can be employed. Using machine learning together with either structural or experimental descriptors, predictive models can be built. The limiting factor in using experimental methods is the lack of data, meaning this method can only be applied to existing compounds which have sufficient experimental data. Structural descriptors can be useful for predicting a variety of compound properties. Structure based descriptors such as ECFP/Morgan circular fingerprints are an effective and established method for predicting compound activity, although the structural diversity of predictions can be limited by the training data. To overcome this potential drawback Petrone et al. introduced a bioactivity based descriptor derived from historical HTS data i.e. the HTS Fingerprint (HTSFP). The HTSFP has the advantage of not containing any structural information and thereby can be used to make activity predictions independent of any structural features. Moreover, HTS fingerprints may detect active compounds with distinct MoAs, such as alternate binding sites. Unfortunately, the HTSFP has one major drawback, which is that predictions cannot be made for all compounds but only for compounds that have been previously tested in HTS assays, compounds without any HTS data cannot have an HTSFP. Furthermore, compounds with very sparse HTSFPs i.e. compounds having only been tested in very few assays, have limited practicality in such predictive models. These compounds are often not useful as they introduce noise into the data and reduce the predictive performance of models and therefore are removed from the dataset. A fingerprint density cutoff is commonly used to exclude these compounds. This method of data processing leads to the loss of a significant amount of potentially valuable information. Despite these problems HTSFPs
have proven to be an effective and robust tool for compound activity predictions in a number of retrospective studies.

Petrone et al. compared the performance of HTSFP and ECFP4 and showed that the HTSFP had better performance for certain targets. The most prominent aspect of this study was the increased structural diversity of the HTSFP predictions. Paricharak et al. showed that HTSFPs are effective tools for iterative screening approaches in HTS to provide more targeted and efficient screening, saving costs and resources. More recently, HTSFPs have been employed for multitask machine learning methods. The study by Noé et al. compared HTSFP and ECFP4, again showing that the predictions returned from HTSFP models have little overlap with those of the ECFP, concluding that HTSFPs are valuable tools for scaffold hopping. A study by Wassermann et al. in 2013 showed the first step in the direction of combining structural and bioactivity descriptors. Their study focused on generating HTSFPs for compounds which had no available HTS data. This was performed by calculating an untested compounds structural similarity to compounds with existing HTSFPs. The HTSFP of compounds with high similarity were substituted onto the untested compounds. A different study by Riniker et al. went a step further and described a method of using both ECFP4 and HTSFPs for activity prediction by building machine learning models on each of the two descriptor types individually and subsequently combining the two trained models using heterogeneous classifier fusion for the final activity predictions.

In this study, a novel fingerprint was designed by combining bioactivity descriptors (HTSFPs) with structural descriptors. The aim was to improve compound activity predictions and scaffold hopping potential of structural fingerprints. The underlying idea was that combining the fingerprints fortifies the HTSFP with structural data, thereby removing the necessity of having to make a HTSFP density cutoff and allowing for a more efficient use of available HTS data. The fingerprint introduced herein was designed by concatenating the HTSFP with an ECFP4 to make a combined experimental and structural fingerprint (CESFP). This CESFP was applied in a machine learning model to make compound activity predictions. A random forest classifier was used to build the predictive model which was based on HTS data from PubChem. The results were validated via a retrospective analysis on a set of HTS assays which had been excluded from the training data. The results were benchmarked against the individual HTSFP and ECFP4.

**Results and Discussion**

The HTS data was obtained from PubChem bioassays and post-refinement, contained a total of 715,000 unique compounds and 561 HTS assays. Eight test assays where chosen at random from this dataset for the retrospective analysis, a detailed overview of these assays is shown in table 1. The
results from the combined experimental and structural fingerprint (CESFP) were benchmarked against the un-concatenated HTSFP and ECFP4. Furthermore, the scaffold hopping potential of the CESFPs was investigated by comparing topological scaffolds and performing a nearest neighbor comparison. The random forest classifier models built on the ECFP4, HTSFP, and CESFP were used to make predictions for each assay. The results of the random forest analysis were investigated for each of the three fingerprint types using a variety of different performance metrics most of which are derived from values of the confusion matrix. Each metric was averaged using the results of a tenfold cross validation and are discussed in detail in the following paragraphs.

| Assay ID | Target                                                                 | Compound Count | % Actives |
|---------|------------------------------------------------------------------------|----------------|-----------|
| 798     | Coagulation factor XI                                                   | 218,716        | 0.14%     |
| 1515    | Retinoblastoma Binding Protein 9 (RBBP9)                                | 217,964        | 0.20%     |
| 2553    | Transient Receptor Potential Cation Channel C6 (TRPC6)                 | 305,614        | 1.06%     |
| 463104  | Activators of the adaptive arm of the Unfolded Protein response        | 331,676        | 0.33%     |
| 504454  | Beta-2 adrenergic receptor                                              | 339,285        | 0.43%     |
| 588497  | Botulinum neurotoxin light chain F protease                            | 340,322        | 0.23%     |
| 624414  | Human Mucolipin Transient Receptor Potential 1 (TRPML1)                | 400,339        | 0.12%     |
| 686964  | Methyl-CpG binding domain protein 2                                    | 369,939        | 0.31%     |

Table 1. Overview of the eight test assays. Shown are their PubChem AID, assay target, total number of compounds tested in assay, and the proportion of actives compounds.

Classification performance

Receiver operator characteristic

Receiver-Operator-Characteristic (ROC) curves for each of the three fingerprint types and each of the eight test assays are displayed in Figure 1. The ROC Area-Under-Curve (ROC-AUC), shown in the Figure 2A bar plot, were calculated to compare the relative performances between the three fingerprint models. The ROC curve compares the true positive rate (TPR) against the false positive rate (FPR), while varying the threshold of the classification confidence scores, this provides an indication of the early enrichment and gives a rough idea of the overall performance. Analysis of these curves and AUC values indicates that prediction performance was poorest in six of the eight assays for the ECFP4 and two of the eight assays for the HTSFP. The relative performance between the ECFP and HTSFP varied from assay to assay, which is likely dependent on the assay target types and also on the density of the HTSFPs for the compounds tested in each assay. The CESFP predictions showed increased ROC for all the eight test assays, although in assays 588497 and 686964 the CESFP showed only minor improvement. Noticeably the ROC curves showed that the early enrichment appeared to be improved in all test assays.
Figure 1. Receiver operator characteristic (ROC) curves comparing the combined fingerprint (CESFP) with the HTSFP and ECFP4, green, orange, and blue respectively. The shaded area either side of the ROC curve represents one standard deviation. Plots A to H correspond to each of the eight test assays.

Precision, Recall, and F1 score
The precision, recall and F1 scores were calculated for each of fingerprints, these were based on the averages from the tenfold cross validation and are plotted in Figure 2E, 2F, 2G respectively. The HTSFP had low precision for all test assays but high sensitivity. The CESFP performed best in seven of eight assays for precision but was out performed by the HTSFP in the sensitivity scores, although the CESFP
sensitivity was consistently significantly above that of the ECFP4. This shows that the HTSFP is predicting a very large number of compounds to be active, thereby recovering a large portion of the true positives, i.e. the high sensitivity, but at the same time a large number of false positives are predicted giving the HTSFP a low precision. To further analyze these results the F1 scores were compared, which gave the harmonic mean of the precision and sensitivity. The plot of the F1 scores in Figure 2G resolves the previously unclear results, showing that the CESFP is consistently outperforming the other two fingerprints in all test assays particularly in cases of assays 624414 and 686964.

**Mathews correlation coefficient**

Results were also compared using the Mathews correlation coefficient (MCC). This is a well-suited metric for measuring the predictive quality on very unbalanced datasets, which is the case for this data, having an average active compound rate of 0.80% across the 561 assays used. The average proportion of actives across the eight tested assays was 0.35% which can distort the quality of other measurement techniques. The bar plot in Figure 2C compares the average MCC scores from the tenfold cross validation for each of the three fingerprint types. In six of the eight test assays the ECFP4 outperformed the HTSFP by a clear margin, in the remaining two assays the HTSFP scored higher than the ECFP4. Again, the varying relative performances between these two fingerprints is likely highly dependent on the amount of information in the assay compounds’ HTS fingerprints, therefore test assays with large portions of mostly sparse HTSFPs will be expected to perform poorer. The MCC scores for the CESFP are consistently and significantly higher than those of the other two fingerprints for all eight of the test assays.

**Cohen’s Kappa score**

The Cohen’s kappa scores were also calculated from the cross-validation results and are plotted in Figure 2B. The plot again shows the improved performance of the CESFP compared with the other two fingerprints. The Kappa score shows an almost identical trend to that seen in the MCC plot with two exceptions, assays 463104 and 686964. In these two cases, HTSFP performed worse than ECFP4 on the basis of Kappa scores, in contrast to MCC analysis.

**Enrichment**

To further investigate the relative performances of the three different fingerprints, the top scoring 1% of compounds from each prediction run of the cross validation were compared. The top 1% represented between 2000-4000 compounds, depending on the assay. The enrichment factor (EF₁%) for active compounds was determined. The average enrichment factor for each fingerprint type in each assay is shown in Figure 2D. The ECFP4 showed the poorest enrichment in five of eight assays, in assays 1515 and 624414 the ECFP4 and HTSFP had similar performance, only in assay 2553 did the
ECFP perform noticeably better than the HTSFP. Overall, the CESFP produced the best enrichment factor for all test assays, but in some cases i.e. 686964, 588497, and 798, EF\textsubscript{1}\% differences were only marginal.

![Performance metrics for the eight test assays comparing the combined fingerprint (CESFP) with the HTSFP and the ECFP4.](image)

Figure 2. Performance metrics for the eight test assays comparing the combined fingerprint (CESFP) with the HTSFP and the ECFP4, green, orange, and blue respectively. A: ROC-AUC, B: Cohens Kappa score, C: Mathews correlation coefficient, D: Enrichment factor, E: Precision, F: Recall, G: F1 score. The errors bars in grey represent one standard deviation.

Scaffold hopping analysis
Scaffold overlap
The second goal of the study was to determine the scaffold hopping potential of the CESFP compared with the ECFP4 and the HTSFP. The scaffold hopping capabilities of the HTSFP is well known and has been demonstrated in a number of studies.9, 11 To compare the chemical diversity of the predicted compounds, the topological scaffolds of each of the true positive predicted compounds in the top
scoring 1% of predictions were compared. The topological scaffold is created by removing all side chains and converting all atoms in the structure to sp3 carbons. As expected the scaffolds predicted using the HTSFP had only a limited overlap with the scaffolds predicted using the ECFP4. On average, 61% of the scaffolds from the ECFP4 were also detected by HTSFP. Venn diagrams were constructed for the three fingerprint types and are shown in Figure 3. Combining the structural (ECFP4) and experimental (HTSFP) fingerprints into one fingerprint (CESFP), one would expect the therefrom predicted scaffolds to reflect some form of overlap from the predictions of both the other two fingerprint types. Assays 463104 and 588497 are representatives of the two extremes within the eight test assays and are shown in Figure 3. In the case of assay 463104, a very wide separation between the three scaffold groups can be seen, whereas in assay 588497 the CESFP overlaps with almost all the scaffolds of both the ECFP4 and the HTSFP. Interestingly, CESFP also predicted an additional completely unique set of topological scaffolds that did not overlap with either of the ECFP4 or the HTSFP predictions in all test assays (green shaded area). This effect was most pronounced in assay 463104 showing 23% unique scaffolds predicted only by the CESFP. On average, the CESFP predicted 10% unique scaffolds across the eight test assays. These results indicated synergistic effects when combining the two fingerprints, leading to the detection of additional scaffolds. The overall count of true positive scaffolds predicted within the top scoring 1% of compounds was also highest for CESFP in all test assays. This suggested that the CESFP was a more effective fingerprint for scaffold hopping than its precursors. The count of unique topological scaffolds for the predictions of each fingerprint type is also noted next to the fingerprint name in Figure 3. Venn diagrams of all test assays can be found in the supplementary data.

Figure 3. Venn diagrams showing the number of unique topological scaffolds in the top scoring 1% of predictions. Each circle represents one of the three predictive models: CESFP, HTSFP, and ECFP4 (green, orange, blue respectively). Plot A refers to test assay 463104 and B to test assay 588497.
Nearest neighbor Tanimoto similarity
To further investigate the results shown in the Venn diagrams a nearest neighbor analysis was performed. The Venn diagrams revealed the presence of different scaffolds but did not reveal how different these scaffolds were to one another structurally. By plotting the nearest neighbor for each compound, the overall structural diversity of the compound set could be visualized. Figure 4 shows the plot of the nearest neighbor Tanimoto similarity of the top scoring 1000 compounds of assay 463104 using each of the three prediction models. The plot shows that the compounds predicted using the ECFP4 share a larger degree of structural similarity relative to the predictions made using the HTSFP. The majority of the compounds predicted using ECFP4 have a Tanimoto similarity between 0.7-0.9 whereas the majority of compounds predicted using the HTSFP have a Tanimoto similarity around 0.3. The compounds predicted using the CESFP had similarity values in between those of ECFP4 and the HTSFP. This distribution provides evidence that the ECFP4 is not as well suited for scaffold hopping as the HTSFP or the CESFP. Although the compounds predicted with the CESFP exhibit a lower degree of structural diversity than the HTSFP, the predictive accuracy of the CESFP is better and is therefore the favored model. All eight test assays followed the same trend as the seen in Figure 4, plots for all test assays are shown in the supplementary figures.

Figure 4. Compound diversity of top scoring 1000 compounds. The nearest neighbor Tanimoto similarity was calculated for each of the 1000 compounds and plotted as a fitted histogram. The nearest neighbor similarity was calculated for each of the 3 predictive models CESFP: green, HTSFP: orange, and ECFP4: blue.

Compound ranking comparison
The top ranked 1000 compounds predicted using CESFP were selected. The rankings of these compounds in the ECFP4 predictions were plotted against the rankings from the HTSFP predictions. These plots for assays 463104 and 624414 are shown in Figure 5. The green dots represent active compounds and the orange dots represent inactive compounds. Compounds above the diagonal black line were ranked higher in the ECFP4 model and compounds below the line were ranked higher in the HTSFP model. The dashed lines represent the boundary for rankings not in the top 1000 for
either the ECFP4 or HTSFP. It was expected that these compounds would be within the top ranking 1000 compounds of either the ECFP4 or the HTSFP i.e. not in the upper right quadrant of the plot. This expectation would give a rise to a ‘L’ shaped clustering. This ‘L’ shaped clustering was only clearly visible in the plot of assay 624414, but even here a small number of the compounds were located outside the expected rankings. In the remaining seven assays larger portions of the 1000 CESFP predicted compounds appeared in the upper right quadrant (see supplementary data). For example, assay 463104 showed a large portion of compounds ranked outside the top 1000 for both the ECFP4 and the HTSFP. The fact that the CESFP predicts many active compounds outside the top 1000 rankings of ECFP4 and HTSFP demonstrates a synergistic effect between structural and experimental descriptors. This synergistic effect allows for improved predictive performance and scaffold hopping capability. The scatter plots for all eight test assays are show in the supplementary information.

Figure 5. Comparison of compound rankings for the three prediction models. The top scoring 1000 compounds predicted using the CESFP are shown. The rankings of the same compounds in the HTSFP model (Y-axis) and the ECFP4 model (X-axis) are compared. The green and orange dots represent active and inactive compounds, respectively. The dashed line boarders the upper right quadrant, which refers to rankings outside the top 1000 rankings for the HTSFP and EFCP4. Results from two test assays are shown in the plots left: AID 463104 and right: AID 624414.

Feature Importance
The feature importances of each of the models for the CESFP tenfold cross validation were analyzed using a feature importance function. The feature importance for assay 463104 is plotted in Figure 6. Features 0-553 refer to the HTSFP (orange) while features 553-1577 refer to the ECFP4 (blue). This plot shows the average and the maximum importance (light and dark color respectively) calculated from the tenfold cross validation. The ECFP4 does not show any features that are significantly more important than others and displays an overall constant basal level of importance, i.e. almost every ECFP4 feature has some importance. For some of the assays certain features in the ECFP4 show higher importance but due to the way the ECFP4 is folded into a 1024 binary vector it is impossible to determine precisely which structural features each bit corresponds to. The HTSFP portion of the CESFP shows much greater variability in feature importance from assay to assay. Overall the basal
level of feature importance in the HTSFP is lower than in the ECFP4, although a small number of the HTSFP features show highly pronounced importance values. This trend of pronounced HTSFP features could be seen across all eight test assays (see plots in supplementary data). The assays corresponding to these pronounced features were investigated in more detail. Discussed here are three representative test assays i.e. AID 798, AID 463104, and AID 504454. The assay biological targets corresponding to the top 5 most important HTSFP features were determined and are shown in Table 2.

Test assay 798, from the PubChem dataset, aimed at identifying compounds which inhibit coagulation factor XI. The random forest model used to make the retrospective predictions on this assay was analyzed and the feature importances were determined. The five most important features all correlate to compounds which were active against targets involved in secondary hemostasis which all have a serine protease function. The 4th and 5th most important features referred to assays 687 and 680 which also targeted coagulation factor XI but had surprisingly low importance. Closer investigation of these two assays revealed the reason for their relatively low importance. One point was that the two assays only had tested compound sets with a test assay overlap of 32,511 and 59,853 respectively, which is relatively small compared to the 798-test assay (718,716 compounds). Another point was that the agreement between the assays was limited, only 23/94 and 21/120 actives were in agreement with the 798-test assay for assays 687 and 680 respectively. These results prove that the random forest model can successfully identify and correlate compounds which have similar activities against similar targets.

Test assay 463104 is targeted at identifying promoters of the unfolded protein response (UPR), specifically the adaptive arm. UPR is involved in protein degradation as well as apoptosis related processes. The top 5 most important features of the random forest model were determined for this assay and their corresponding assay biological targets are listed in Table 2. The first most important feature corresponds to an assay targeting E3 ubiquitin protein ligase. The E3 ubiquitin protein ligase is involved in Ubiquitination processes, which are directly involved with protein degradation, and are a vital element of the UPR. The 2nd and 3rd most important features both refer to assays also targeting different domains of the UPR. The 4th most important feature corresponds to an assay targeting ‘Protein phosphatase 1 regulatory subunit 15A’. This target is involved with regulation of protein synthesis and plays a role in the UPR, its relatively high importance suggests that this target is likely also present or closely related to the targets in the 463104-test assay. Again, the fact that the four most important features all correspond to compounds which were active in the UPR process, shows the predictive model’s ability to make sensible correlations within the HTSFP. The 5th most important feature corresponds to an assay targeted at melanocortin receptor 4 (MC4R). MC4R is a GPCR which
has no known association to the UPR. This result suggests that the predictive model has the ability to draw correlations from unrelated features of the HTSFP, thereby supporting a wider applicability domain. Considering this lack of correlation between the two targets, it must also be mentioned that the relative importance of this feature is much lower (0.013) as can be inferred from figure 6.

Assay 504454 is aimed at identifying inhibitors of the beta-2 adrenergic receptor (b2AR) which is a member of the GPCR family. The biological target of test assay 504454 did not have any known relation to the biological targets of the top 5 features. The PubChem assays and their associated biological targets corresponding to these five features are listed in table 2. These five assays target a variety of different functional proteins, none of which are members of the GPCR family. The targets types include regulatory subunits, inflammasomes, protein ligases, and two transcription factors. This result shows activity predictions for a given assay are not dependent on the HTSFP containing assays with related or similar biological targets. In other words, valid activity predictions can be made for compounds which are being tested on previously unexplored targets.

| PubChem AID | Feature number | Importance value | Assay biological target |
|-------------|----------------|------------------|-------------------------|
| 798         | Test Assay     | Coagulation factor XI |
| 800         | 525            | 0.048            | Coagulation factor XIIa light chain |
| 873         | 544            | 0.043            | Human kallikrein 5 (hK5) serine protease |
| 1046        | 17             | 0.019            | Prothrombin |
| 687         | 459            | 0.005            | Coagulation factor XI |

Figure 6. Feature importance of the combined fingerprint (CESFP). Features 0-553 correspond to the HTSFP portion (orange) and features 553-1577 correspond to the ECFP4 portion (blue) of the combined fingerprint. The light and dark shades of each feature refer to the mean and max values from the tenfold cross validation, respectively.
## Conclusion

From analysis of the various metrics used to assess the prediction quality of the CESFP it can be concluded that the CESFP yields a significant improvement in prediction performance relative to the individual ECFP4 and HTSFP. The MCC, F1 score, enrichment factor, ROC-AUC and Cohen's kappa score all show evidence of the combined fingerprint's enhanced performance. The results indicate that this combined fingerprint is a useful tool for scaffold hopping, detecting not only a more diverse set of active compounds with different scaffolds but also identifying novel scaffolds that were not identified with either the ECFP4 or the HTSFP. The improved scaffold hopping ability of the CESFP was further supported by the nearest neighbor analysis. A comparison of the compound rankings provided evidence of the synergistic effects between the structural and bioactivity-based fingerprints. Feature importance analysis quantified the relative contributions of ECFP4 and HTSFP to the CESFP predictions, revealing that a small subset of the HTSFP features contribute most to the overall performance. This subset of features often corresponded to assays with targets biologically related to the test assays, however, this was not necessary for the HTSFP’s increased contribution. Overall the CESFP appears to be a promising tool for activity prediction.

## Methods and data

### Dataset

For this research eight HTS assays obtained from PubChem were investigated retrospectively, they contained diverse ratios of active to inactive compounds as well as varying target types and a range of assay sizes ranging from approximately 200,000 – 400,000 compounds per assay. An overview of the eight test assays is shown in table 1.

| AID  | Test Assay | Feature Position | Importance Value | Assay Target                                      |
|------|------------|------------------|------------------|--------------------------------------------------|
| 680  | 452        | 0.003            | Coagulation factor XI | Activators of the adaptive arm of the unfolded protein response |
| 463104 | Test Assay | Activators of the adaptive arm of the unfolded protein response | 485346 | 214 | 0.088 | E3 ubiquitin-protein ligase Mdm2/MdmX |
| 449763 | 188 | 0.038 | Activators of the apoptotic arm of the unfolded protein response |
| 2732 | 163 | 0.037 | Inhibitors of CHOP to regulate the unfolded protein response |
| 588405 | 312 | 0.025 | Protein phosphatase 1 regulatory subunit 15A |
| 540308 | 290 | 0.013 | Melanocortin receptor 4 (MC4R) |
| 504454 | Test Assay | Bet-2 adrenergic receptor | 588405 | 312 | 0.035 | Protein phosphatase 1 regulatory subunit 15A |
| 743279 | 499 | 0.029 | Inhibitors of inflammasome signaling: IL-1-beta |
| 485346 | 214 | 0.020 | E3 ubiquitin-protein ligase Mdm2/MdmX |
| 488899 | 223 | 0.017 | MITF microphthalmia-associated transcription factor |
| 624352 | 390 | 0.014 | Endothelial PAS domain-containing protein 1 |

Table 2. The PubChem assays corresponding to the five highest importance features as seen in Figure 6. Column one refers to the PubChem AID, column two refers to feature position with the combined fingerprint, column 3 indicates the importance value, and column 4 gives information on the assay target.
Descriptors for models
Generation of HTS fingerprints. A list of 582 HTS assays were downloaded from the PubChem database. Assays containing fewer than 20,000 compounds were discarded, leaving a total of 561 assays. This cut off was made to reduce the size and sparsity of the HTS fingerprint. If any compounds were tested multiple times with mixed activity outcomes, the most common activity flag was used. In the case where there were equal numbers of active and inactive flags, the active flag was used. All compound's activity flags were collated into a matrix of 'compound ID' versus 'Assay ID', with dimensions 715,328(compounds) x 561(assays). The fingerprint was subsequently binarized by converting all active labels to '1' and inactive labels to '0'. All missing data was also set to '0', the reasoning for this was that the HTS data is very unbalanced and a compound with unknown activity has a much higher probability of being inactive and is therefore given the label of an inactive bit.

Structural descriptors. For the same list of 715,328 compounds as in the HTSFP, ECFP4 fingerprints were created. The PubChem HTS data contained only the CID for the compounds and to make the ECFP fingerprints the smiles for each compound was required. Using the list of CIDs, the Smiles for each compound were downloaded from the PubChem database. The Morgan circular fingerprint (an analogue of ECFP) implemented in RDKit was used. After removal of compounds with invalid or unreadable smiles for RDKit, a compound set of 715,327 was obtained. The bit length was set to 1024 bits and the fragment radius was set to 2 (diameter 4). Tests were run comparing 1024-bit ECFP4 with 1024-bit ECFP6 for one of the test assays. Only minor differences could be seen in predictive performance but the ECFP6 appeared to be slightly weaker, therefore the ECFP4 was chosen for the full analysis.

Generation of the CESFP. The combined experimental and structural fingerprint (CESFP) was created by concatenating the ECFP4 to the HTSFP, giving a new fingerprint of length 1585 (561+1024). These fingerprints were created using the same compound set (715,327) as output from the ECFP4.

Modelling methods
Due the nature of the random forest learning method, where specific features within a fingerprint are identified and not the entire fingerprint, it was theorized that RF would be the best suited technique to deal with the large portion of majorly sparse HTSFPs in the dataset. A test run was performed comparing random forest with support vector machine models of the Scikit-learn package. The two models were tested on one of the eight test assays, the random forest showed better performance according to the ROC AUC values and also ran significantly faster.

The random forest classifier machine learning package from Scikit-learn was used for building models of three different descriptor types, i.e. ECFP4, HTSFP, and CESFP. Here the ECFP4 and HTSFP
were used for comparative and benchmarking purposes in all performance evaluations. The hyperparameters used for the Random forest classifier are: \( n\text{\_estimators}=150, \) \( \text{class\_weight='balanced', max\_features='sqrt', min\_samples\_leaf=10, n\_jobs=-1.} \) The number of trees \( (n\text{\_estimators}) \) was set to 150 as above this threshold model performance did not appear to improve. A ‘balanced’ class weighting was used due to the imbalanced nature of the data, the ‘balanced’ setting of this hyperparameter was vital for adequate performance of the models. For model validation a ten-fold cross-validation was performed, averages and standard deviations were calculated across the ten folds for each of the test assays. Every fold contains between 0.09-1.15% active labeled compounds. The metrics for each test assay were calculated using the mean values and standard deviations calculated across the ten folds.

For the scaffold hopping analysis the true positives in the top ranked 1% of predictions were extracted for each cross-validation fold and their compound IDs (CIDs) were mapped to smiles. Using RDKit each compound was converted to a topological Bemis-Murcko scaffold (generic scaffold) i.e. all side chains were removed, all heteroatoms converted to carbons, and all bond orders set to 1 (all C = sp3). The number of unique topological scaffolds were then counted and averaged across the ten folds. The unique scaffolds predicted from each of the 3 tested fingerprints were compared using Venn diagrams made from the matplotlib-venn add-on. Venn diagrams were made for each cross-validation fold and the average for each region in the diagram was taken to make the final diagram.

To compare the compound diversity for the predictions made using each of 3 fingerprint types (HTSFP, ECFP4, CESFP) a nearest neighbor comparison was performed. The nearest neighbor is calculated by performing a Tanimoto similarity comparison of the ECFP4s for each compound in the prediction set. A Tanimoto similarity score of 1.0 is obtained for two compounds whose fingerprints are identical, whereas a score of 0.0 means that the fingerprints have no overlap. The similarity scores for all compounds in the top 1000 predictions were calculated and their distribution plotted (Figure 4).

**Calculation of metrics**

Receiver operator characteristic curves were constructed using the false positive rate (FPR) and true positive rate (TPR) while changing the classification threshold according to the prediction probability scores. The two equations in (1) show how the FPR and TPR are calculated.

\[
\begin{align*}
\text{FPR} &= \frac{FP}{FP + TN} \\
\text{TPR} &= \frac{TP}{TP + FN}
\end{align*}
\]  

\[ (1) \]
The precision and recall were calculated using the formulas shown in (2). The F1 score is the harmonic mean of the precision and recall and the calculation formula is also shown in (2).

\[
\text{Precision} = \frac{TP}{TP + FP} \quad \text{Recall} = \frac{TP}{TP + FN} \quad F1\text{score} = \frac{2TP}{2TP + FP + FN}
\]

(2)

The Matthews correlation coefficient (MCC) is a performance metric optimized for imbalanced datasets. The equation to calculate the MCC is shown in (3). The MCC covers a range from -1 to 1, where a value of 1 indicates a perfect prediction, -1 a perfect inverse prediction and 0 indicating prediction no better than random.

\[
MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}
\]

(3)

The equation in (4) shows how the Cohen’s Kappa score is calculated, where \( p_o \) is the relative observed agreement of a class (accuracy) and \( p_e \) is the hypothetical probability of chance agreement. A kappa score of 0 reflects a performance no better than random chance, the more positive the score the better.

\[
\kappa = \frac{p_o - p_e}{1 - p_e}
\]

(4)

The Enrichment factor provides a measure of how much the model performance improves compared to random screening. The resulting score refers to a factor of improvement, where a score of 1.0 is equivalent to random. The formula to calculate the enrichment factor for the top scoring 1% of compounds is shown in (5). The Hitrate\(^{1\%}\) refers to the count of true positives in the top scoring 1%, and the Hitrate\(^{100\%}\) refers to the hit rate for the overall screen.

\[
EF_{1\%} = \frac{\text{Hitrate}^{1\%}}{\text{Hitrate}^{100\%}}
\]

(5)

Software used: Python 3.6.5, SKLearn 0.19.1, SciPy 1.1.0, RDKit 2018.03.1.0.

**Abbreviations**

HTS: high throughput screening, CESPF: combined experimental and structural fingerprint, ECFP: extended connectivity fingerprint, HTSFP: high throughput screening fingerprint, GPCR: G-protein coupled receptor, MCC: Mathews correlation coefficient, EF: enrichment factor
Declarations

Availability of data and material
A .txt file with the list of PubChem assays used and all supplementary figures are provided in additional files. Assay data can be downloaded from PubChem at https://pubchem.ncbi.nlm.nih.gov/ All source code is available upon request.

Competing interests
The authors declare no competing financial interest.

Funding
Oliver Laufkötter has received funding from the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska Curie grant agreement No 676434, “Big Data in Chemistry” (“BIGCHEM”, http://bigchem.eu). Noé Sturm received funding from European Union’s Horizon 2020 research and innovation program under the grant agreement no. 671555 “ExCAPE”, http://www.excape-h2020.eu/. The article reflects only the authors view and neither the European Commission nor the Research Executive Agency (REA) are responsible for any use that may be made of the information it contains.

Authors’ contributions
Oliver Laufkötter carried out this study and wrote the manuscript. Noé Sturm supervised and assisted throughout the project. Jürgen Bajorath, Hongming Chen, and Ola Engkvist supervised the project and gave project guidance. All authors participated in manuscript proofreading and approved the final manuscript.

Acknowledgements
Michael Withnall helped with debugging, implementing code, and proofreading. Laurianne David, Josep Arus-Pous and Thomas Blaschke provided valuable insight and discussions.

References
1. Wu, Y.; Wang, G., Machine Learning Based Toxicity Prediction: From Chemical Structural Description to Transcriptome Analysis. Int J Mol Sci 2018, 19 (8).
2. Simm, J.; Klambauer, G.; Arany, A.; Steijaert, M.; Wegner, J. K.; Gustin, E.; Chupakhin, V.; Chong, Y. T.; Vialard, J.; Buijsters, P.; Velter, I.; Vapirev, A.; Singh, S.; Carpenter, A. E.; Wuyts, R.; Hochreiter, S.; Moreau, Y.; Ceulemans, H., Repurposing High-Throughput Image Assays Enables Biological Activity Prediction for Drug Discovery. Cell Chemical Biology 2018, 25 (5), 611-618.e3.
3. Wassermann, A. M.; Lounkine, E.; Urban, L.; Whitebread, S.; Chen, S.; Hughes, K.; Guo, H.; Kutlina, E.; Fekete, A.; Klumpp, M.; Glick, M., A screening pattern recognition method finds new and divergent targets for drugs and natural products. ACS Chem Biol 2014, 9 (7), 1622-31.
4. Mayr, L. M.; Bojanic, D., Novel trends in high-throughput screening. Curr Opin Pharmacol 2009, 9 (5), 580-8.
5. Battersby, B. J.; Trau, M., Novel miniaturized systems in high-throughput screening. *Trends Biotechnol* **2002**, 20 (4), 167-73.

6. Karnachi, P. S.; Brown, F. K., Practical approaches to efficient screening: information-rich screening protocol. *J Biomol Screen* **2004**, 9 (8), 678-86.

7. Glem, R. C.; Bender, A.; Arnby, C. H.; Carlsson, L.; Boyer, S.; Smith, J., Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME. *iDrugs* **2006**, 9 (3), 199-204.

8. Avram, S.; Bora, A.; Halip, L.; Curpan, R., Modeling Kinase Inhibition Using Highly Confident Data Sets. *J Chem Inf Model* **2018**.

9. Sturm, N.; Sun, J.; Vandriessche, Y.; Mayr, A.; Klambauer, G. U. N.; Carlsson, L.; Engkvist, O.; Chen, H., Application of Bioactivity Profile Based Fingerprints For Building Machine Learning Models. *J Chem Inf Model* **2018**.

10. Riniker, S.; Wang, Y.; Jenkins, J. L.; Landrum, G. A., Using information from historical high-throughput screens to predict active compounds. *J Chem Inf Model* **2014**, 54 (7), 1880-91.

11. Petrone, P. M.; Simms, B.; Nigsch, F.; Lounkine, E.; Kutchukian, P.; Cornett, A.; Deng, Z.; Davies, J. W.; Jenkins, J. L.; Glick, M., Rethinking molecular similarity: comparing compounds on the basis of biological activity. *ACS Chem Biol* **2012**, 7 (8), 1399-409.

12. Helal, K. Y.; Maciejewski, M.; Gregori-Puigjane, E.; Glick, M.; Wassermann, A. M., Public Domain HTS Fingerprints: Design and Evaluation of Compound Bioactivity Profiles from PubChem’s Bioassay Repository. *J Chem Inf Model* **2016**, 56 (2), 390-8.

13. Cortes Cabrera, A.; Petrone, P. M., Optimal HTS Fingerprint Definitions by Using a Desirability Function and a Genetic Algorithm. *J Chem Inf Model* **2018**, 58 (3), 641-646.

14. Paricharak, S.; AP, I. J.; Bender, A.; Nigsch, F., Analysis of Iterative Screening with Stepwise Compound Selection Based on Novartis In-house HTS Data. *ACS Chem Biol* **2016**, 11 (5), 1255-64.

15. Wassermann, A. M.; Lounkine, E.; Glick, M., Bioturbo similarity searching: combining chemical and biological similarity to discover structurally diverse bioactive molecules. *J Chem Inf Model* **2013**, 53 (3), 692-703.

16. Pedregosa, F.; Ga. #235; Varoquaux, l.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; #201; Duchesnay, d., Scikit-learn: Machine Learning in Python %J J. Mach. Learn. Res. **2011**, 12, 2825-2830.
Supplementary Information

Databases

PubChem HTS assays. List of assays used for HTSFP generation attached: Pubchem_assay_list.txt

Figure 1. Overview of PubChem HTS data. 582 unique assays were used, and 715,326 unique compounds were identified. White space refers to the number of compounds not tested for a given assay (missing data), blue space refers to the count of compounds marked as ‘not active’, and orange space refers to the count of compounds marked as ‘active’.
Figure 2. Venn diagrams showing the number of unique topological scaffolds in the top scoring 1% of predictions. Each circle represents one of the three predictive models: CESFP, HTSFP, and ECFP4 (green, orange, blue respectively). Plot A to H correspond to each of the test assays.
Figure 3. Compound diversity of top scoring 1000 compounds for each test assay. The nearest neighbor Tanimoto similarity was calculated for each of the 1000 compounds and plotted as a fitted histogram. The nearest neighbor similarity was calculated for each of the 3 predictive models CESFP: green, HTSFP: orange, and ECFP4: blue.
Figure 4. Comparison of compound rankings for the three prediction models. The top scoring 1000 compounds predicted using the CESFP are shown. The rankings of the same compounds in the HTSFP model (Y-axis) and the ECFP4 model (X-axis) are compared. The green and orange dots represent active and inactive compounds, respectively. The dashed line refers to rankings outside the top 1000 for both HTSFP and EFCP4. Plots A to H refer to each of the eight test assays.
Figure 5. Feature importance analysis of the combined fingerprint (CESFP). Features 0-553 correspond to the HTSFP portion (orange) and features 553-1577 correspond to the ECFP4 portion (blue) of the combined fingerprint. The light and dark shades of each feature refer to the mean and max values from the tenfold cross validation, respectively.
### Other files

| File Name         | Size   | View on ChemRxiv | Download File |
|-------------------|--------|------------------|---------------|
| Pubchem_assay_list.txt | 3.81 KiB | [view on ChemRxiv](#) | [download file](#) |