Supporting Information

Profiling Prediction of Kinase Inhibitors: Toward the Virtual Assay

Benjamin Merget, Samo Turk, Sameh Eid, Friedrich Rippmann, and Simone Fulle

Contents

Supporting Figures

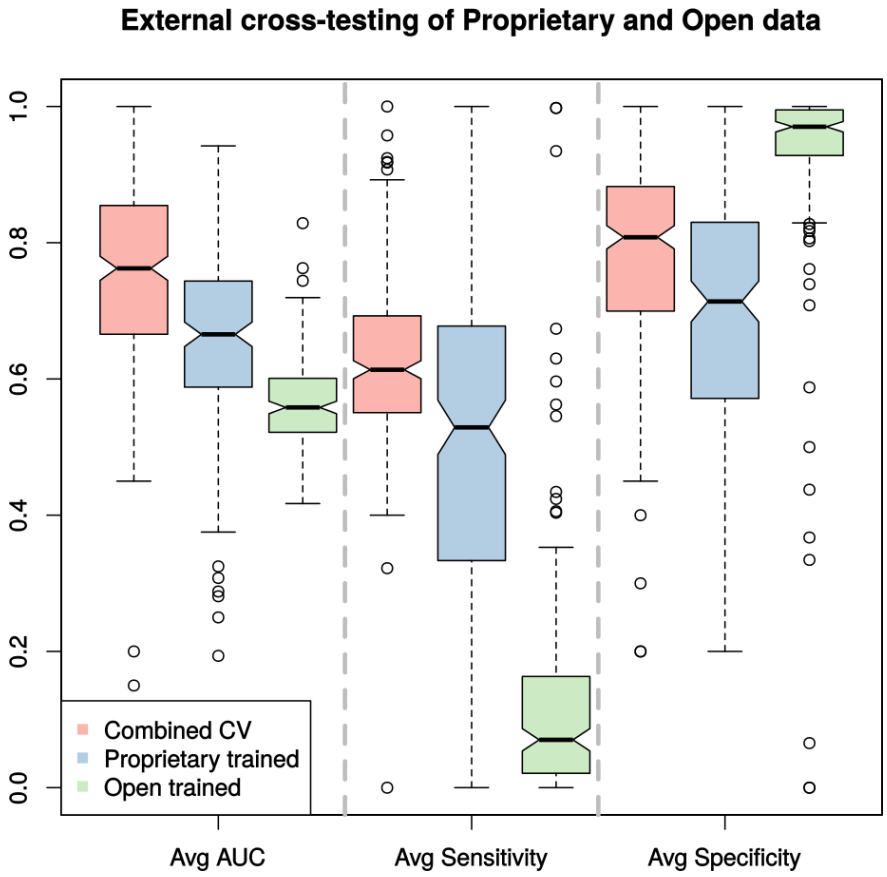

Figure S1: External validation of models trained on the *Proprietary* data set and tested on the *Open* data and vice versa ("*Proprietary* trained" and "*Open* trained", respectively). The performance of the "*Open* trained" models is very low when tested on the *Proprietary* data (average AUC of 0.56±0.06). However, models trained on *Proprietary* data show reasonable prediction results when tested on the *Open* data (average AUC of 0.65±0.13).

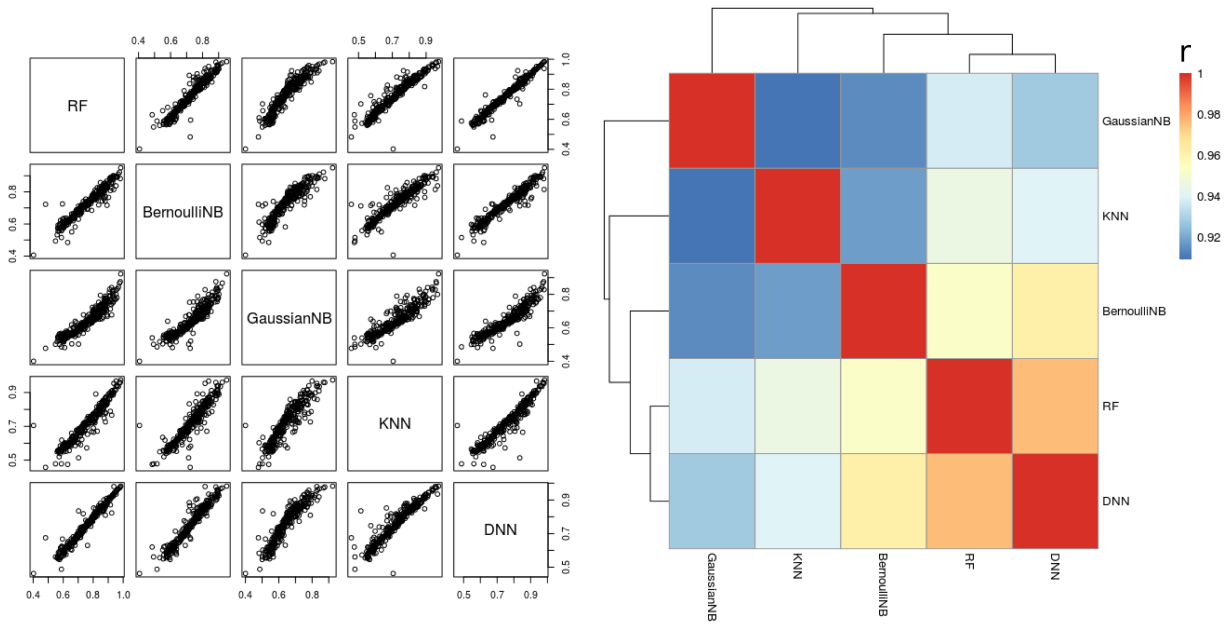

Figure S2: Pairs plot and heatmap of AUC values of various classifiers. The AUC values are highly correlated as indicated by Pearson r-values above 0.9.

The impact of the balancing technique

The class imbalance problem of Machine Learning can constitute complications, including poor performance.[1–3] Generally, it can be tackled on the data level by undersampling of the majority class (inactives) or by oversampling of the minority class (actives). Besides *random* under- and oversampling, other algorithms have been proposed in the literature. In this study, undersampling was also performed based on centroid clustering and Nearest Neighbor (NearMiss-3) search.[4] For oversampling of the minority class, the Synthetic Minority Oversampling Technique (SMOTE) was used,[5] which generates new synthetic data points within the feature space of the existing samples. These methods are implemented in the python library *unbalanced dataset* (https://github.com/fmfn/UnbalancedDataset). Additionally, we introduce a to our knowledge novel undersampling technique: *PCA-Centroids*. Here, a Principal Component Analysis (PCA) is performed on the fingerprint data of the majority class. Subsequently, a K-Medoids clustering is employed on a selected number of principal components with the number of clusters being the number of positive compounds (i.e., active against a given kinase) and the original fingerprints of the cluster centroids (medoids) are extracted. To ensure maximum fairness in model evaluation for advanced undersampling algorithms, the training set was undersampled using the described technique, while the test set was undersampled randomly in every run of the 5-fold CV. Again, 10 external test sets were evaluated additionally in every run. The performances of the different undersampling methods are summarized in Figure S3. As an alternative approach for data balancing, the active molecules were oversampled using two different algorithms (Figure S4).

In summary, the two fundamentally different approaches for data balancing might serve different purposes. Whereas undersampled models might be useful to capture a large number of potentially active molecules (high sensitivity), oversampling can be used to create models which can reliably detect true negative compounds (high specificity), leading to a less polluted set of positively classified molecules (Figures S3 and S4).

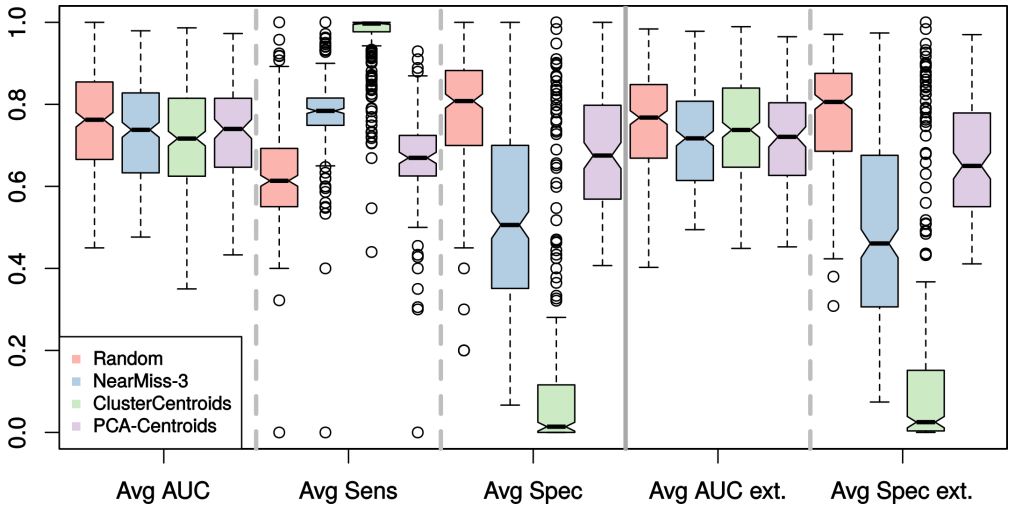

Figure S3: Boxplots of model quality measurements for Random Forest using various *undersampling* methods. Of the tested undersampling methods, none yields a better performance than random undersampling. Although the sensitivity can be raised by advanced undersampling, this approach often results in significantly lower specificity, limiting the applicability of the models, particularly for the NearMiss and the ClusterCentroid approach. In terms of AUC, all methods are significantly worse than random undersampling ($p \ll 0.001$ in pairwise Mann-Whitney-U tests). Good results are also achieved by the PCA-Centroids method with an average AUC of 0.73 ± 0.11 and an average sensitivity significantly higher compared to random undersampling ($p \ll 0.001$). A major drawback of this method, however, is a decreased specificity for internal and external data likewise. In summary, random undersampling followed by PCA-Centroids show the best average performance.

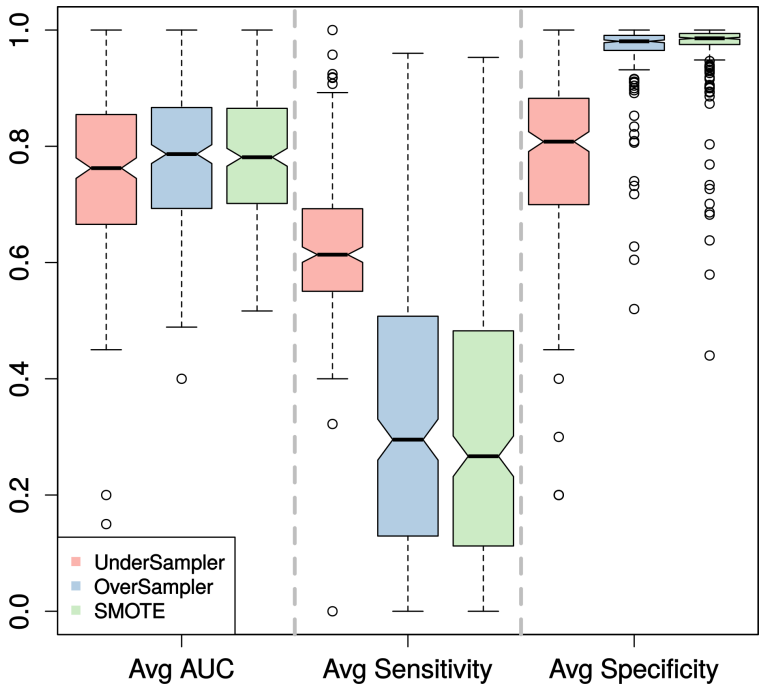

Figure S4: Boxplots of model quality measurements for Random Forest using various *oversampling* methods. Random *undersampling* is added as a baseline. Although the average AUC values are comparable to that of random undersampling, sensitivity and specificity become highly imbalanced by oversampling, i.e., a very high specificity is achieved at the cost of a low sensitivity.

References

(1) Batista, G. E.; Prati, R. C.; Monard, M. C. A study of the behavior of several methods for balancing machine learning training data. *ACM Sigkdd Explorations Newsletter* **2004**, *6*, 20–29.

(2) Chawla, N. V. *Data Mining and Knowledge Discovery Handbook*; Springer, 2009; pp 875–886.

(3) Zakharov, A. V.; Peach, M. L.; Sitzmann, M.; Nicklaus, M. C. QSAR modeling of imbalanced high-throughput screening data in PubChem. *Journal of Chemical Information and Modeling* **2014**, *54*, 705–712.

(4) Mani, I.; Zhang, I. kNN approach to unbalanced data distributions: a case study involving information extraction. Proceedings of Workshop on Learning from Imbalanced Datasets. 2003.

(5) Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* **2002**, 321–357.