# SURFLEMENTARY MATERIAL

Sereina Riniker, Nikolas Fechner, and Gregory A. Landrum*

*Novartis Institutes for BioMedical Research, Basel, Switzerland*

E-mail: gregory.landrum@novartis.com

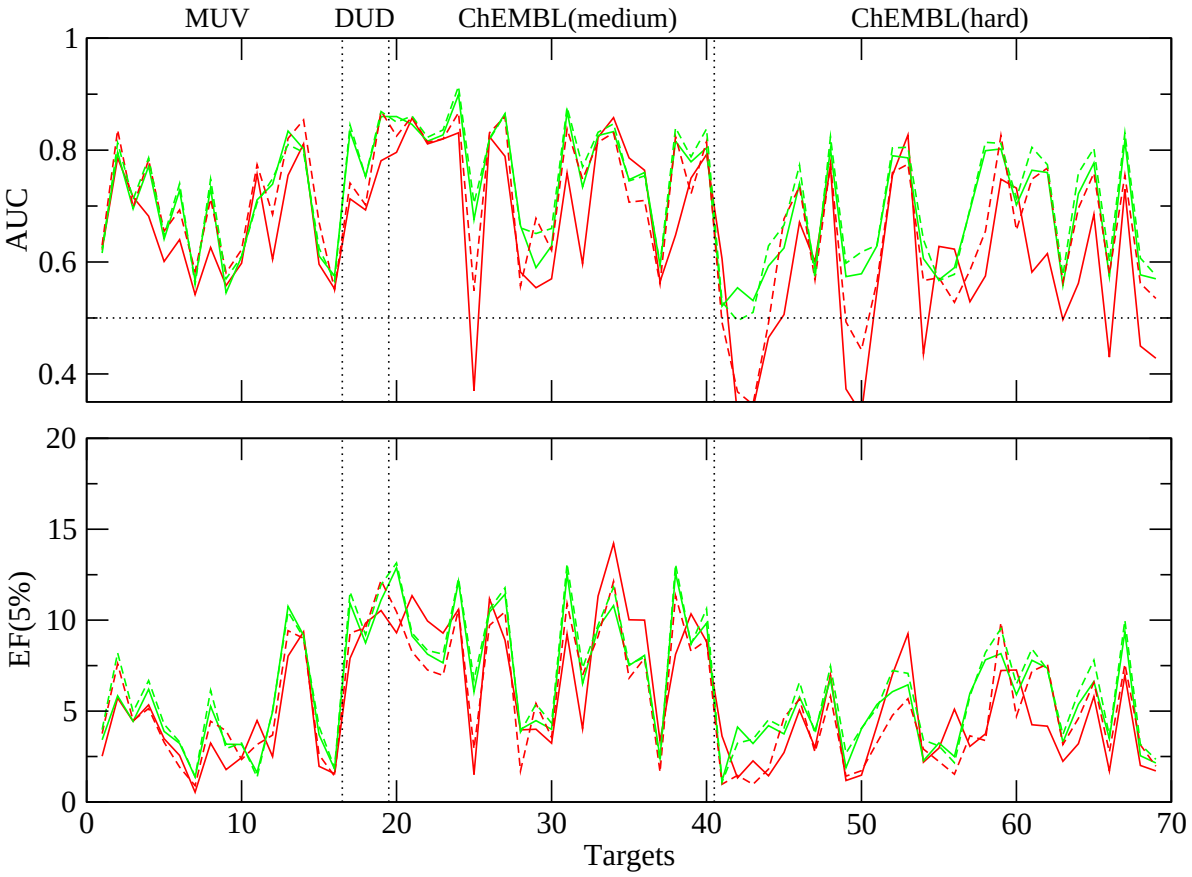


Figure S1: Average performance of AP (red) and TT (green) measured with AUC (top) and EF(5%) (bottom). The count version of the fingerprint is shown with a dashed line, the hashed bit-vector version with a solid line. The horizontal, dotted line indicates random distribution.

*To whom correspondence should be addressed

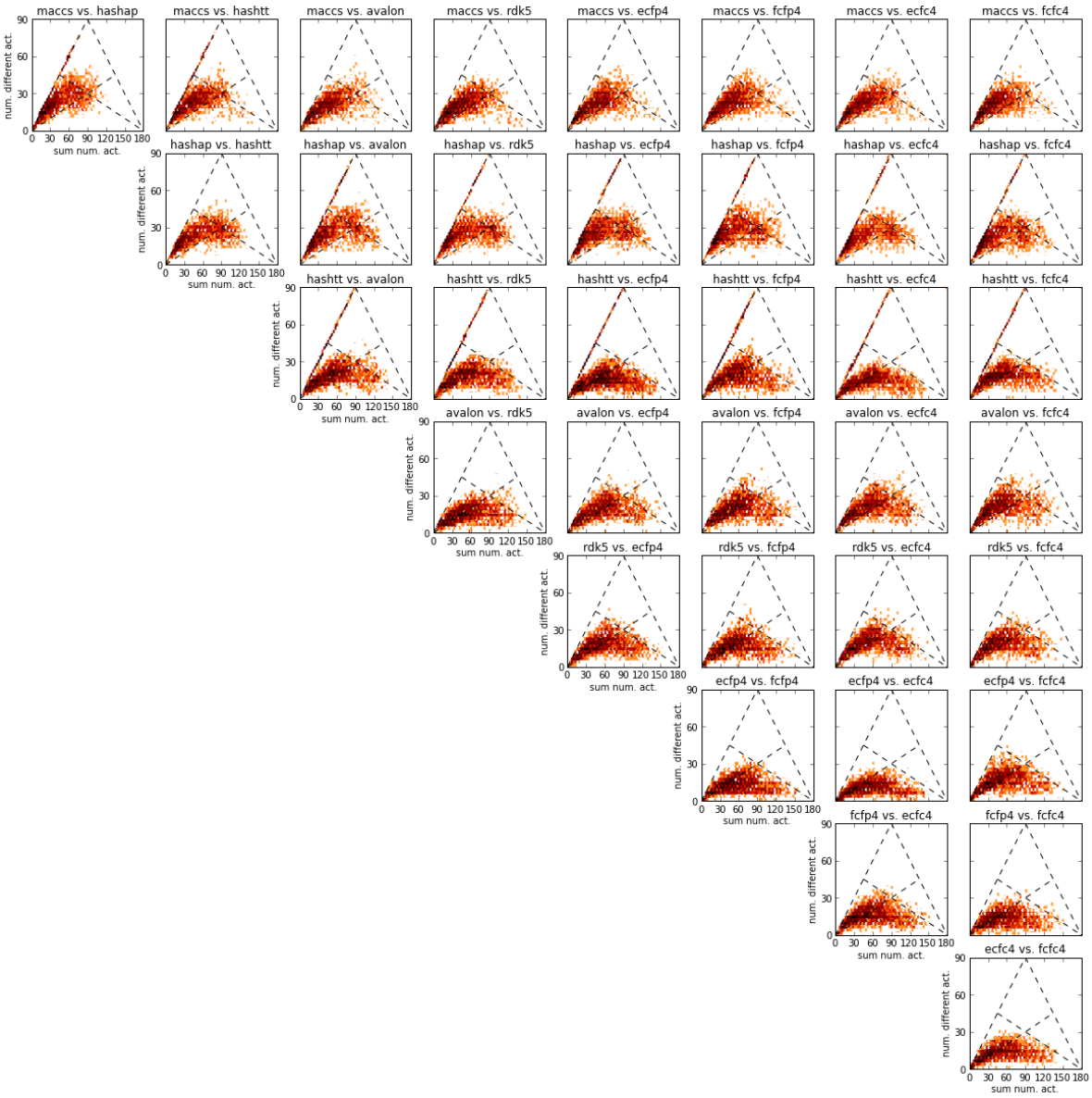


Figure S2: Heat map of the number of actives that were only found by one fingerprint in the first 5% of the ranked list as a function of the sum of the actives found by the two fingerprints in the first 5% of the ranked list for ten fingerprints: MACCS, AP, TT, Avalon, RDK5, ECFP4, FCFP4, ECFC4 and FCFC4. Ranked lists of 50 ChEMBL targets in data sets I and 50 repetitions per target were used.

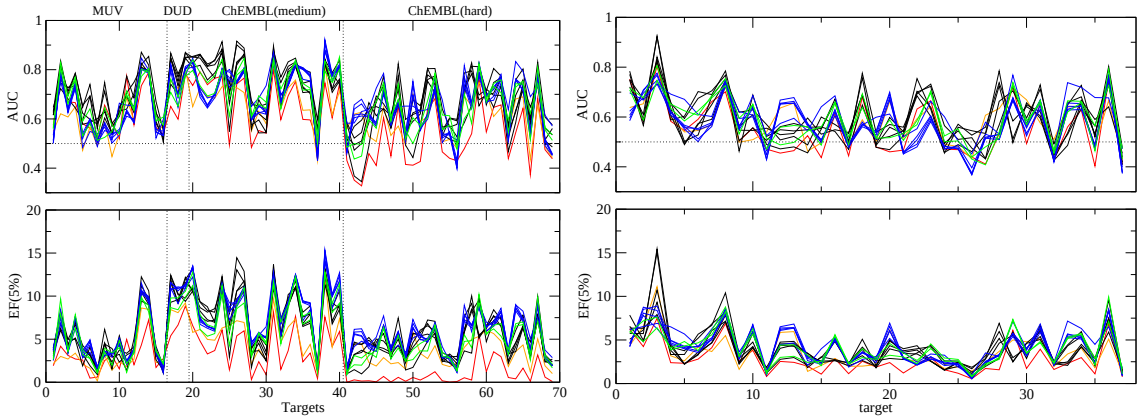


Figure S3: Average performance of 14 2D fingerprints measured with AUC (top) and EF(5%) (bottom). The two baseline fingerprints are shown in red (ECFC0) and orange (MACCS). Path-based fingerprints (AP, TT, Avalon, long Avalon, RDK5) are shown in black, circular fingerprints with bit string (ECFP4, long ECFP4, ECFP6, long ECFP6, FCFP4) in blue, and circular fingerprints with count vector (ECFC4, FCFC4) in green. The horizontal, dotted line indicates random distribution.

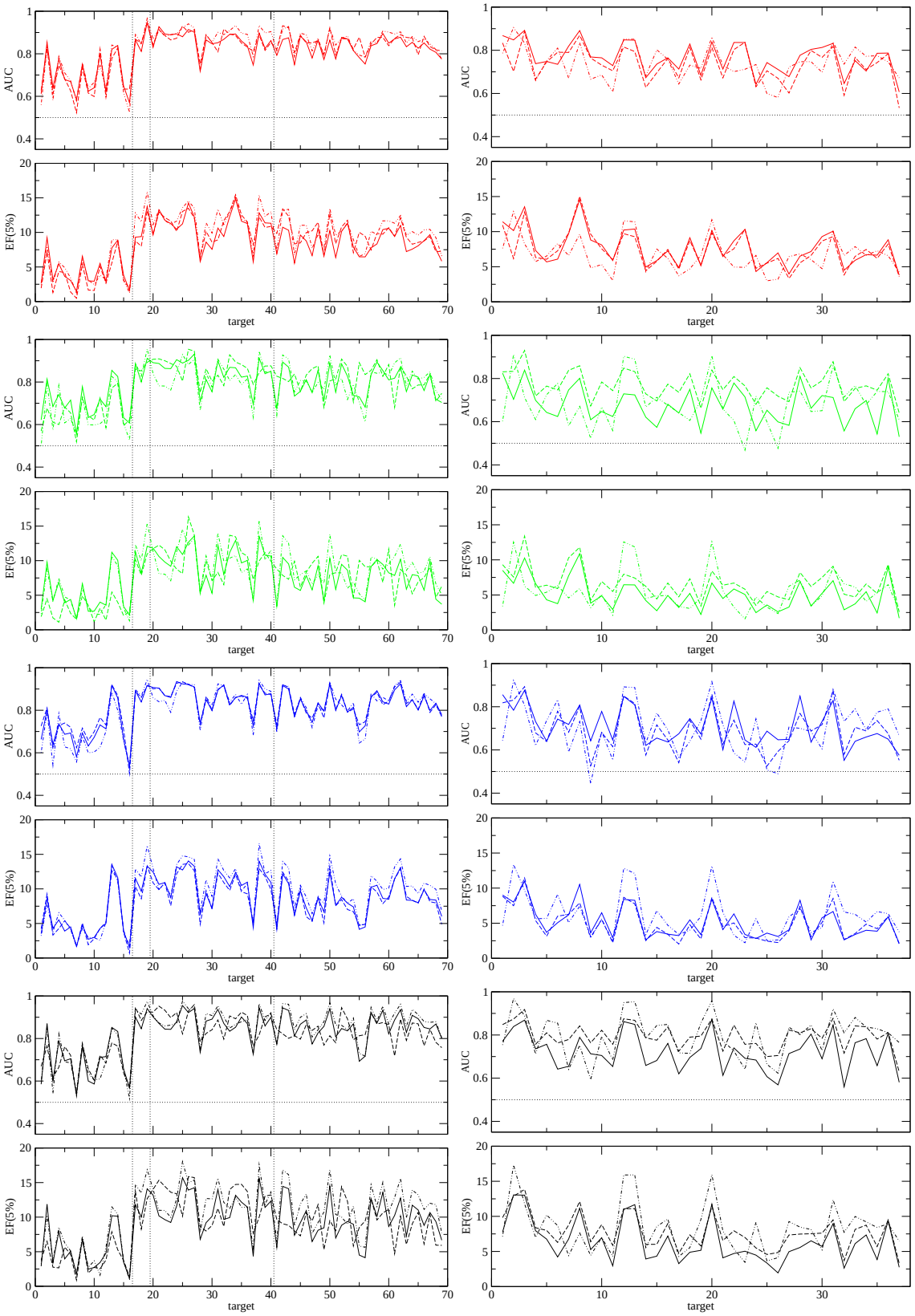


Figure S4: Average performance of random forest (RF) (solid lines), naïve Bayes (NB) (dashed lines) or logistic regression (LR) (dashed-dotted lines) measured with AUC and EF(5%) for data sets I (left) and data sets II (right). Four standard fingerprints are compared: AP (red), TT (green), RDK5 (blue), and Morgan2 (black).

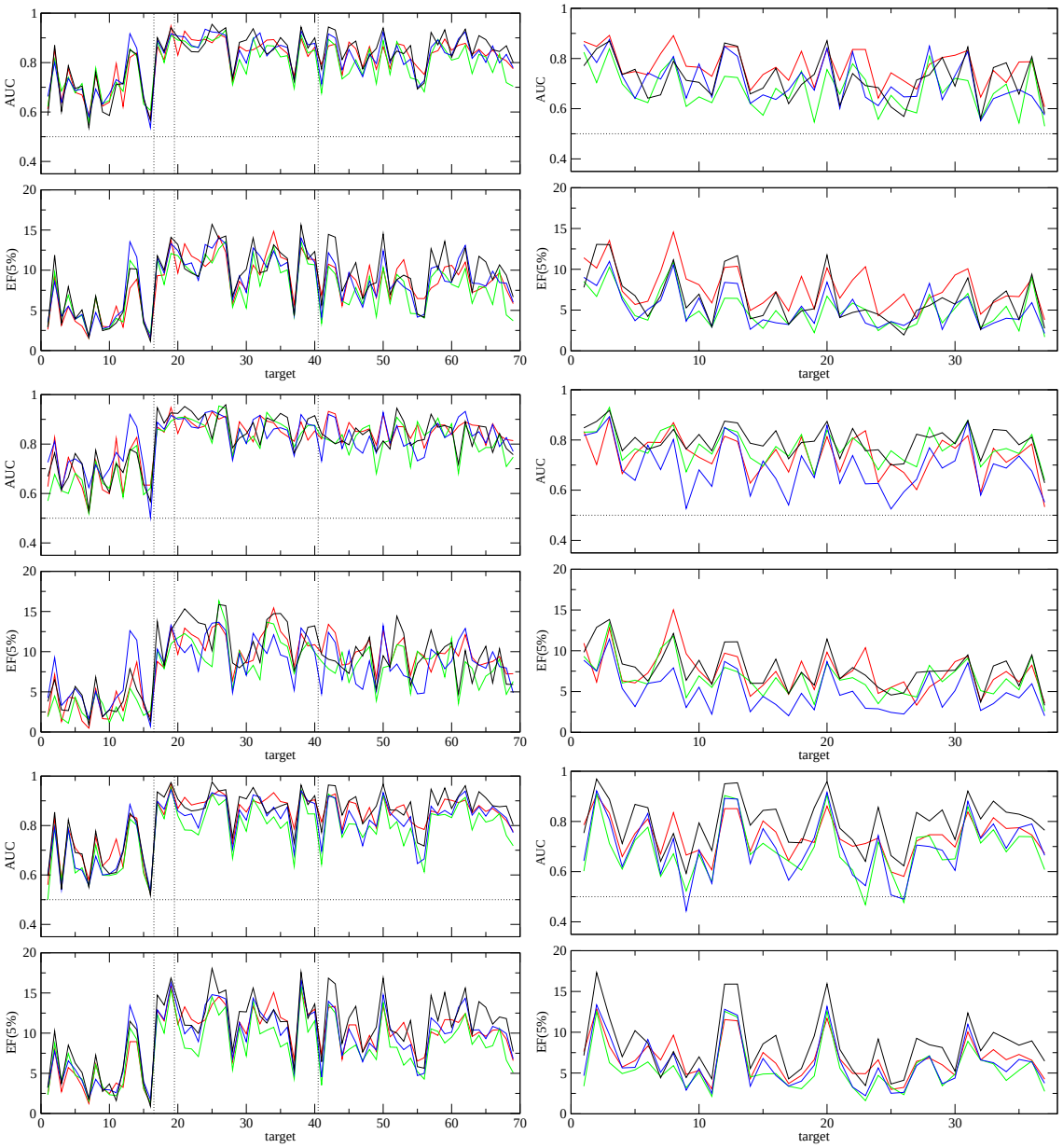


Figure S5: Average performance of random forest (RF) (top), naïve Bayes (NB) (middle) or logistic regression (LR) (bottom) measured with AUC and EF(5%) for data sets I (left) and data sets II (right). Four standard fingerprints are compared: AP (red), TT (green), RDK5 (blue), and Morgan2 (black).

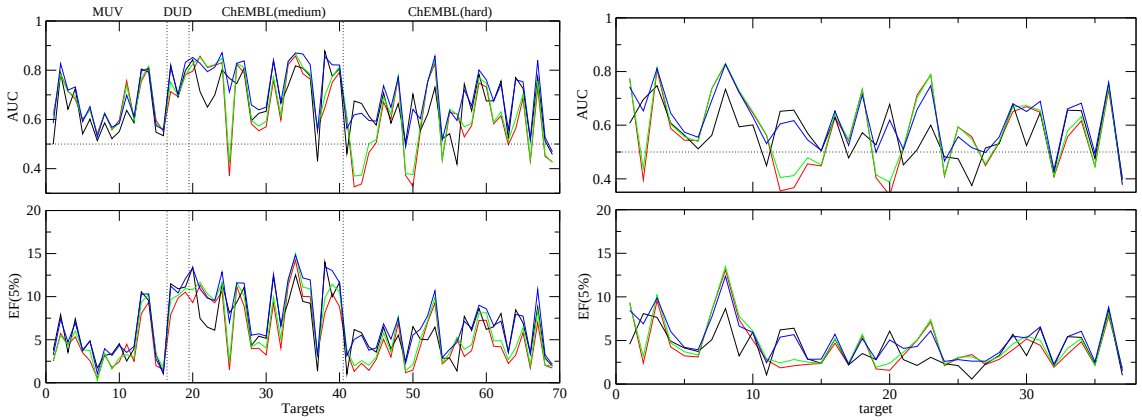


Figure S6: Average performance of AP (red), Morgan2 (black), hybrid fingerprint (green), and *similarity-fused* AP and Morgan2 (blue) measured with AUC (top) and EF(5%) (bottom) for data sets I (left) and data sets II (right).

Table S1: List of 69 targets with data set origin, target ID, target description, number of actives (A), number of decoys (D), and ratio actives/decoys (A/D).

| Origin | Target ID | Description | A | D | A/D |
|---|---|---|---|---|---|
| MUV | 466 | Sphingosine 1-phosphate (S1P1) receptor | 30 | 15000 | 0.002 |
| | 548 | Protein kinase A (PKA) | 30 | 15000 | 0.002 |
| | 600 | Steroidogenic factor 1 (SF1): inhibitors | 30 | 15000 | 0.002 |
| | 644 | Rho-kinase 2 | 30 | 15000 | 0.002 |
| | 652 | HIV-1 RT-Rnase H | 30 | 15000 | 0.002 |
| | 689 | Ephrin receptor A4 | 30 | 15000 | 0.002 |
| | 692 | Steroidogenic factor 1 (SF1): agonists | 30 | 15000 | 0.002 |
| | 712 | Heat shock protein 90 (HSP90) | 30 | 15000 | 0.002 |
| | 713 | Estrogen receptor (ER) $\alpha$: inhibitors | 30 | 15000 | 0.002 |
| | 733 | Estrogen receptor (ER) $\beta$ | 30 | 15000 | 0.002 |
| | 737 | Estrogen receptor (ER) $\alpha$: potentiators | 30 | 15000 | 0.002 |
| | 810 | Focal adhesion kinase (FAK) | 30 | 15000 | 0.002 |
| | 832 | Cathepsin G | 30 | 15000 | 0.002 |
| | 852 | Factor XIIa (FXIIa) | 30 | 15000 | 0.002 |
| | 858 | Dopamine receptor D1 | 30 | 15000 | 0.002 |
| | 859 | Muscarinic receptor M1 | 30 | 15000 | 0.002 |
| DUD | cdk2 | Cyclin-dependent kinase 2 | 47 | 2070 | 0.023 |
| | hivrt | HIV-1 RT-Rnase | 34 | 1494 | 0.023 |
| | vegfr2 | Vascular endothelial growth factor receptor 2 | 48 | 2712 | 0.011 |
| ChEMBL | 11359 | Phosphodiesterase 4D | 100 | 10000 | 0.010 |
| | 8 | Tyrosine-protein kinase ABL | 100 | 10000 | 0.010 |
| | 10434 | Tyrosine-protein kinase SRC | 100 | 10000 | 0.010 |
| | 12670 | Tyrosine-protein kinase receptor FLT3 | 100 | 10000 | 0.010 |
| | 12261 | c-Jun N-terminal kinase I | 100 | 10000 | 0.010 |
| | 12209 | Carbonic anhydrase XII | 100 | 10000 | 0.010 |
| | 25 | Glucocorticoid receptor | 100 | 10000 | 0.010 |
| | 36 | Progesterone receptor | 100 | 10000 | 0.010 |
| | 43 | $\beta$-2 adrenergic receptor | 100 | 10000 | 0.010 |
| | 219 | Muscarinic acetylcholine receptor M3 | 100 | 10000 | 0.010 |
| | 130 | Dopamine receptor D3 | 100 | 10000 | 0.010 |
| | 105 | Serotonin 1d (5-HT1d) receptor | 100 | 10000 | 0.010 |
| | 126 | Cyclooxygenase-2 | 100 | 10000 | 0.010 |
| | 12252 | $\beta$-secretase 1 | 100 | 10000 | 0.010 |
| | 11575 | C-C chemokine receptor type 2 | 100 | 10000 | 0.010 |
| | 11534 | Cathepsin S | 100 | 10000 | 0.010 |
| | 10498 | Cathepsin L | 100 | 10000 | 0.010 |
| | 12911 | Cytochrome P450 2C9 | 100 | 10000 | 0.010 |
| | 100579 | Nicotinic acid receptor 1 | 100 | 10000 | 0.010 |
| | 10378 | Cathepsin B | 100 | 10000 | 0.010 |
| | 11631 | Sphingosine 1-phosphate receptor Edg-1 | 100 | 10000 | 0.010 |
| | 165 | hERG | 100 | 10000 | 0.010 |
| | 10193 | Carbonic anhydrase I | 100 | 10000 | 0.010 |
| | 15 | Carbonic anhydrase II | 100 | 10000 | 0.010 |
| | 11489 | 11-$\beta$-hydroxysteroid dehydrogenase 1 | 100 | 10000 | 0.010 |
| | 121 | Serotonin transporter | 100 | 10000 | 0.010 |
| | 72 | Dopamine D2 receptor | 100 | 10000 | 0.010 |
| | 259 | Cannabinoid CB2 receptor | 100 | 10000 | 0.010 |
| | 10188 | MAP kinase p38 $\alpha$ | 100 | 10000 | 0.010 |
| | 108 | Serotonin 2c (5-HT2c) receptor | 100 | 10000 | 0.010 |
| | 12952 | Carbonic anhydrase IX | 100 | 10000 | 0.010 |

| | | | | | |
|---|---|---|---|---|---|
| | 93 | Acetylcholinesterase | 100 | 10000 | 0.010 |
| | 10980 | Vascular endothelial growth factor receptor 2 | 100 | 10000 | 0.010 |
| | 19905 | Melanin-concentrating hormone receptor 1 | 100 | 10000 | 0.010 |
| | 107 | Serotonin 2a (5-HT2a) receptor | 100 | 10000 | 0.010 |
| | 87 | Cannabinoid CB1 receptor | 100 | 10000 | 0.010 |
| | 17045 | Cytochrome P450 3A4 | 100 | 10000 | 0.010 |
| | 11140 | Dipeptidyl peptidase IV | 100 | 10000 | 0.010 |
| | 114 | Adenosine A1 receptor | 100 | 10000 | 0.010 |
| | 90 | Dopamine D4 receptor | 100 | 10000 | 0.010 |
| | 13001 | Matrix metalloproteinase-2 | 100 | 10000 | 0.010 |
| | 104 | Monoamine oxidase B | 100 | 10000 | 0.010 |
| | 65 | Cytochrome P450 19A1 | 100 | 10000 | 0.010 |
| | 61 | Muscarinic acetylcholine receptor M1 | 100 | 10000 | 0.010 |
| | 10280 | Histamine H3 receptor | 100 | 10000 | 0.010 |
| | 51 | Serotonin 1a (5-HT1a) receptor | 100 | 10000 | 0.010 |
| | 100 | Norepinephrine transporter | 100 | 10000 | 0.010 |
| | 10260 | Vanilloid receptor | 100 | 10000 | 0.010 |
| | 52 | $\alpha$-2a adrenergic receptor | 100 | 10000 | 0.010 |
| | 11365 | Cytochrome P450 2D6 | 100 | 10000 | 0.010 |

Table S2: List of 37 ChEMBL targets with target ID, target description, and number of papers ($N_p$).

| **Target ID** | **Description** | $\mathbf{N}_p$ |
|---|---|---|
| 10434 | Tyrosine-protein kinase SRC | 8 |
| 12209 | Carbonic anhydrase XII | 8 |
| 25 | Glucocorticoid receptor | 6 |
| 43 | $\beta$-2 adrenergic receptor | 14 |
| 130 | Dopamine receptor D3 | 25 |
| 126 | Cyclooxygenase-2 | 8 |
| 12252 | $\beta$-secretase 1 | 6 |
| 11575 | C-C chemokine receptor type 2 | 8 |
| 11534 | Cathepsin S | 11 |
| 11631 | Sphingosine 1-phosphate receptor Edg-1 | 7 |
| 165 | hERG | 32 |
| 10193 | Carbonic anhydrase I | 33 |
| 15 | Carbonic anhydrase II | 33 |
| 11489 | 11-$\beta$-hydroxysteroid dehydrogenase 1 | 15 |
| 121 | Serotonin transporter | 37 |
| 72 | Dopamine D2 receptor | 21 |
| 259 | Cannabinoid CB2 receptor | 33 |
| 10188 | MAP kinase p38 $\alpha$ | 18 |
| 108 | Serotonin 2c (5-HT2c) receptor | 15 |
| 12952 | Carbonic anhydrase IX | 24 |
| 93 | Acetylcholinesterase | 8 |
| 10980 | Vascular endothelial growth factor receptor 2 | 24 |
| 19905 | Melanin-concentrating hormone receptor 1 | 33 |
| 107 | Serotonin 2a (5-HT2a) receptor | 20 |
| 87 | Cannabinoid CB1 receptor | 32 |
| 17045 | Cytochrome P450 3A4 | 8 |
| 11140 | Dipeptidyl peptidase IV | 23 |

| | | |
|---|---|---|
| 114 | Adenosine A1 receptor | 33 |
| 90 | Dopamine D4 receptor | 4 |
| 13001 | Matrix metalloproteinase-2 | 10 |
| 65 | Cytochrome P450 19A1 | 6 |
| 61 | Muscarinic acetylcholine receptor M1 | 7 |
| 10280 | Histamine H3 receptor | 23 |
| 51 | Serotonin 1a (5-HT1a) receptor | 24 |
| 100 | Norepinephrine transporter | 22 |
| 10260 | Vanilloid receptor | 4 |
| 11365 | Cytochrome P450 2D6 | 5 |

## References

(1) Noronha, G. et al. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 602–608.

Table S3: Results for data sets I: pairwise post-hoc Friedman tests of the average rank between the three machine-learning (ML) methods, random forest (RF), naïve Bayes (NB) and logistic regression (LR), trained with four 2D fingerprints, atom pairs (AP), topological torsions (TT), RDKit fingerprint (RDK5) and Morgan fingerprint (Morgan2), for the evaluation methods AUC (top) and EF(5%) (bottom). Pairs of ML methods with no statistically significant difference are marked with "X", pairs with an adjusted p-value distribution around the confidence level $\alpha$ with "o", and pairs with a statistically significant difference with "-". ML models are ordered according to ascending average rank.

| **AUC** | **LR(Morgan2)** | **LR(AP)** | **RF(Morgan2)** | **NB(Morgan2)** | **NB(RDK5)** | **NB(AP)** | **RF(RDK5)** | **RF(AP)** | **LR(RDK5)** | **NB(TT)** | **RF(TT)** | **LR(TT)** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **LR(Morgan2)** | | X | X | X | o | o | - | - | - | - | - | - |
| **LR(AP)** | | | X | X | X | X | o | o | o | - | - | - |
| **RF(Morgan2)** | | | | X | X | X | X | X | X | o | - | - |
| **NB(Morgan2)** | | | | | X | X | X | X | X | X | - | - |
| **NB(RDK5)** | | | | | | X | X | X | X | X | o | - |
| **NB(AP)** | | | | | | | X | X | X | X | o | - |
| **RF(RDK5)** | | | | | | | | X | X | X | X | - |
| **RF(AP)** | | | | | | | | | X | X | X | - |
| **LR(RDK5)** | | | | | | | | | | X | X | - |
| **NB(TT)** | | | | | | | | | | | X | o |
| **RF(TT)** | | | | | | | | | | | | X |
| **LR(TT)** | | | | | | | | | | | | |

| **EF(5%)** | **LR(Morgan2)** | **LR(RDK5)** | **LR(AP)** | **RF(Morgan2)** | **NB(Morgan2)** | **RF(RDK5)** | **NB(AP)** | **RF(AP)** | **NB(RDK5)** | **LR(TT)** | **RF(TT)** | **NB(TT)** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **LR(Morgan2)** | | X | o | o | - | - | - | - | - | - | - | - |
| **LR(RDK5)** | | | X | X | X | X | o | o | - | - | - | - |
| **LR(AP)** | | | | X | X | X | X | X | o | - | - | - |
| **RF(Morgan2)** | | | | | X | X | X | X | o | - | - | - |
| **NB(Morgan2)** | | | | | | X | X | X | X | o | o | - |
| **RF(RDK5)** | | | | | | | X | X | X | o | X | o |
| **NB(AP)** | | | | | | | | X | X | X | X | X |
| **RF(AP)** | | | | | | | | | X | X | X | X |
| **NB(RDK5)** | | | | | | | | | | X | X | X |
| **LR(TT)** | | | | | | | | | | | X | X |
| **RF(TT)** | | | | | | | | | | | | X |
| **NB(TT)** | | | | | | | | | | | | |

Table S4: Results for data sets II: pairwise post-hoc Friedman tests of the average rank between the three machine-learning (ML) methods, random forest (RF), naïve Bayes (NB) and logistic regression (LR), trained with four 2D fingerprints, atom pairs (AP), topological torsions (TT), RDKit fingerprint (RDK5) and Morgan fingerprint (Morgan2), for the evaluation methods AUC (top) and EF(5%) (bottom). Pairs of ML methods with no statistically significant difference are marked with "X", pairs with an adjusted p-value distribution around the confidence level $\alpha$ with "o", and pairs with a statistically significant difference with "-". ML models are ordered according to ascending average rank.

| **AUC** | **LR(Morgan2)** | **NB(Morgan2)** | **NB(TT)** | **RF(AP)** | **NB(AP)** | **LR(AP)** | **RF(Morgan2)** | **RF(RDK5)** | **LR(RDK5)** | **NB(RDK5)** | **LR(TT)** | **RF(TT)** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **LR(Morgan2)** | | X | X | X | o | o | - | - | - | - | - | - |
| **NB(Morgan2)** | | | X | X | - | - | - | - | - | - | - | - |
| **NB(TT)** | | | | X | X | X | o | - | o | - | - | - |
| **RF(AP)** | | | | | X | X | o | o | o | - | - | - |
| **NB(AP)** | | | | | | X | X | X | X | X | X | o |
| **LR(AP)** | | | | | | | X | X | X | X | X | o |
| **RF(Morgan2)** | | | | | | | | X | X | X | X | X |
| **RF(RDK5)** | | | | | | | | | X | X | X | X |
| **LR(RDK5)** | | | | | | | | | | X | X | X |
| **NB(RDK5)** | | | | | | | | | | | X | X |
| **LR(TT)** | | | | | | | | | | | | X |
| **RF(TT)** | | | | | | | | | | | | |

| **EF(5%)** | **LR(Morgan2)** | **NB(Morgan2)** | **RF(AP)** | **NB(AP)** | **LR(AP)** | **NB(TT)** | **RF(Morgan2)** | **LR(RDK5)** | **LR(TT)** | **RF(RDK5)** | **RF(TT)** | **NB(RDK5)** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **LR(Morgan2)** | | X | X | X | o | X | o | - | - | - | - | - |
| **NB(Morgan2)** | | | X | X | o | X | o | - | - | - | - | - |
| **RF(AP)** | | | | X | X | X | o | o | - | - | - | - |
| **NB(AP)** | | | | | X | X | X | o | - | - | - | - |
| **LR(AP)** | | | | | | X | X | X | o | o | o | - |
| **NB(TT)** | | | | | | | X | X | o | o | - | - |
| **RF(Morgan2)** | | | | | | | | X | X | X | o | o |
| **LR(RDK5)** | | | | | | | | | X | X | X | o |
| **LR(TT)** | | | | | | | | | | X | X | X |
| **RF(RDK5)** | | | | | | | | | | | X | X |
| **RF(TT)** | | | | | | | | | | | | X |
| **NB(RDK5)** | | | | | | | | | | | | |