# Supporting information

# Do Fragments and Crystallization Additives Bind Similarly to Drug-like Ligands?

Malgorzata N. Drwal[∞], Célien Jaquemard[∞], Carlos Perez[§], Jérémy Desaphy[#], Esther Kellenberger[∞*]

[∞]Laboratoire d'Innovation Thérapeutique, UMR 7200 CNRS-Université de Strasbourg, 74 Route du Rhin, 67400 Illkirch, France;

[§]Eli Lilly Research Laboratories, Avenida de la Industria, 30, 28108, Alcobendas, Madrid, Spain;

[#]Lilly Research Laboratories, Eli Lilly and Company, Lilly Corporate Center, Indianapolis, IN 46285, USA;

* To whom correspondence should be addressed: Prof. Esther Kellenberger, Laboratoire d'Innovation Thérapeutique, UMR 7200 CNRS-Université de Strasbourg, 74 Route du Rhin, 67400 Illkirch, France; Phone: +33 3 688 54 221; Email: ekellen@unistra.fr.

**Supplementary Table S1. Unique ligand and PDB file counts in the dataset**

| Set | | Additives | | Fragments | | Drug-like ligands | |
|---|---|---|---|---|---|---|---|
| | **Uniprot AC** | **Unique HET codes** | **PDB entries** | **Unique HET codes** | **PDB entries** | **Unique HET codes** | **PDB entries** |
| **Filtered PDB set[1]** | _ | 362 | 8967 | 1961 | 4809 | 3528 | 5000 |
| Including working dataset | | | | | | | |
| **Entire working dataset[2]** | _ | 31 | 587 | 232 | 291 | 377 | 422 |
| Including the protein targets | | | | | | | |
| **BACE1** | **P56817** | 10 | 126 | 22 | 22 | 124 | 130 |
| **CAH2** | **P00918** | 15 | 168 | 106 | 114 | 49 | 49 |
| **CDK2** | **P29241** | 9 | 120 | 48 | 54 | 147 | 164 |
| **TRY1** | **P00760** | 18 | 173 | 46 | 101 | 57 | 79 |

The numbers of PDB files for each class are given in brackets. [1]Filtered PDB dataset: crystal structures of protein-ligand complexes with a resolution ≤ 3Å, deposited after 01/01/2000, for targets co-crystallized with where both additives/fragments and drug-like ligands (see Methods for more details). [2]Joined counts for BACE1, CAH2, CDK2 and TRY1 (duplicate HET codes are possible).

**Supplementary Table S2. Crystallization additives in the dataset**

| HET code | Additive name | Function | Number of files (dataset) | Number of files (PDB)[1] |
|---|---|---|---|---|
| 1PE | Pentaethylene glycol | Precipitant | 1 | 926 |
| ACE | Acetyl group | Acetyl | 44 | 385 |
| ACT | Acetate ion | Crystallographic buffer | 32 | 6823* |
| AML | Amylamine | Additive | 1 | 2 |
| BCN | Bicine | Crystallographic buffer | 16 | 72 |
| BEZ | Benzoic acid | Additive | 5 | 139 |
| BME | Beta-mercaptoethanol | Reducing agent | 4 | 1107 |
| CIT | Citric acid | Crystallographic buffer | 1 | 955 |
| CO3 | Carbonate ion | Other buffer | 1 | 446 |
| DMF | Dimethylformamide | Solvent | 4 | 371 |
| DMS | Dimethyl sulfoxide | Solvent | 240 | 5942* |
| DTT | 2,4-dihydroxy-1,4-dithiobutane | Reducing agent | 1 | 270 |
| EDO | 1,2-ethanediol | Cryoprotection agent | 281 | 23504* |
| FMT | Formic acid | Precipitant | 6 | 2805* |
| GOL | Glycerol | Cryoprotection agent | 450 | 30514* |
| IMD | Imidazole | Crystallographic buffer | 9 | 906 |
| IPA | Isopropyl alcohol | Solvent and precipitant | 3 | 795 |
| MES | 2-(N-morpholino)-ethansulfonic acid | Crystallographic buffer | 17 | 1074 |
| MLA | Malonic acid | Crystallographic buffer | 1 | 182 |
| MRD | (4R)-2-methylpentane-2,4-diol | Precipitant | 1 | 713 |
| PEG | 2-(2-hydroxyethyloxy)ethanol | Precipitant | 9 | 3339* |
| PGE | Triethylene glycol | Precipitant | 1 | 978 |
| PO4 | Phosphate ion | Crystallographic buffer | 17 | 8567* |
| SAR | Sarcosine | Additive | 2 | 11 |
| SGM | Monothioglycerol | Reducing agent | 1 | 56 |
| SO4 | Sulfate ion | Precipitant | 285 | 42094* |
| SPK | Spermine | Additive | 1 | 2 |
| TAR | D(-)-tartaric acid | Precipitant | 24 | 176 |
| TLA | L(+)-tartaric acid | Precipitant | 45 | 399 |
| TMO | Trimethylamine oxide | Additive | 15 | 28 |
| TRS | 2-amino-2-hydroxymethyl-propane-1,3-diol | Crystallographic buffer | 3 | 1130 |

[1]The filtered PDB dataset contains crystal structures of protein-ligand complexes with resolution ≤3 Å, deposited after 01/01/2000. *Molecules among the top 10 most frequent additives in the filtered PDB dataset. The numbers reflect the count of all copies of the additive in all PDB structures. It should be noted that multiple copies of an additive can be present in one PDB file.

**Supplementary Table S3. Counts of substructure pairs and their components**

| | BACE1 | CAH2 | CDK2 | TRY1 |
|---|---|---|---|---|
| **Additives** | 4 | 4 | 4 | 5 |
| **Fragments** | 1 | 11 | 12 | 7 |
| **Drug-like ligands** | 70 | 24 | 64 | 53 |
| **Substructure pairs** | 82 | 31 | 85 | 150 |

**Supplementary Table S4. Binding mode conservation of substructure pairs**

See separate file: TableS4.txt

**Supplementary Table S5. Fragment docking rescoring performance**

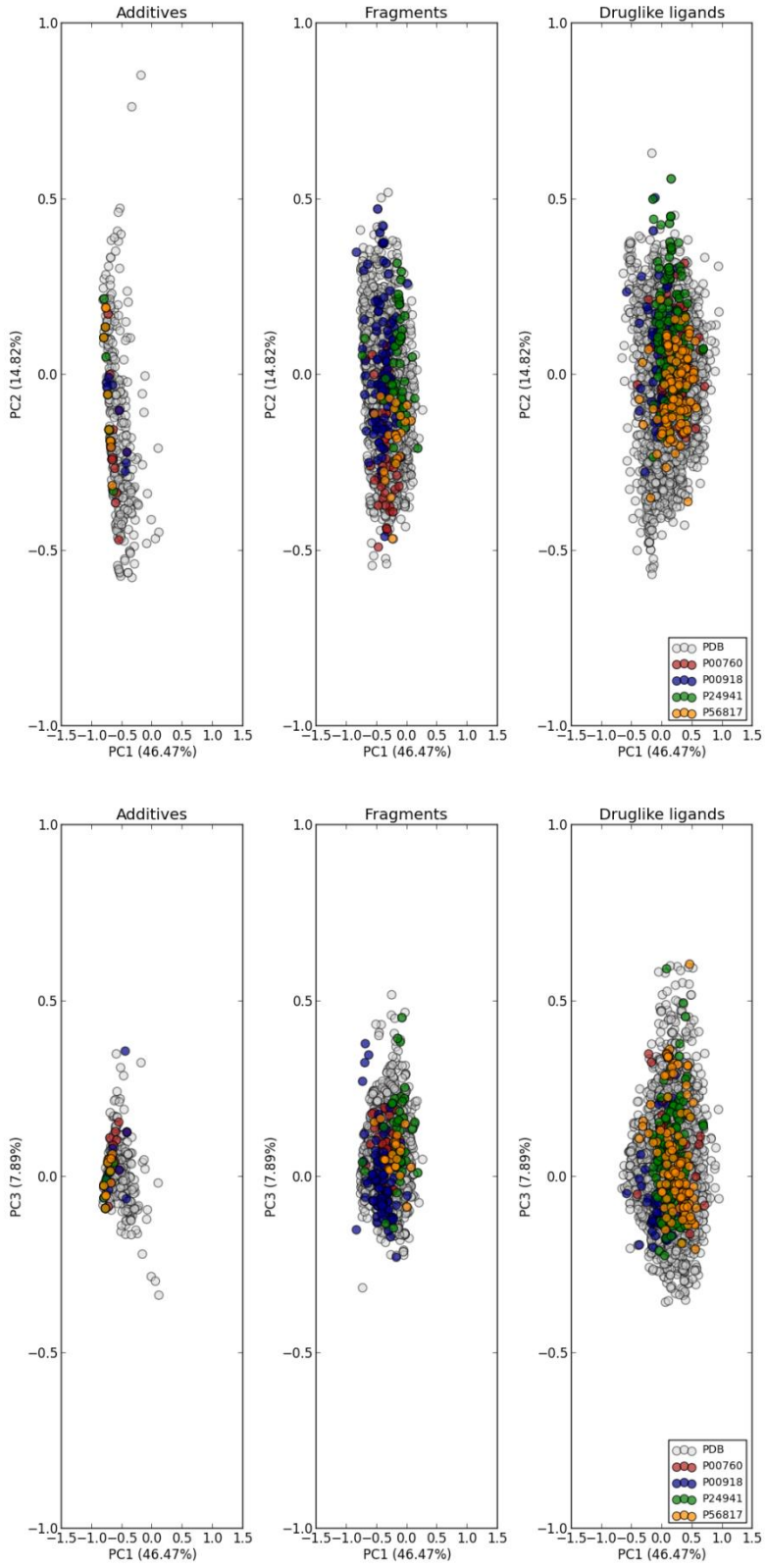
| Scoring | BACE1 | | CAH2 | | CDK2 | | TRY1 | | All targets | |
|---|---|---|---|---|---|---|---|---|---|---|
| | M[1] | P[2] | M[1] | P[2] | M[1] | P[2] | M[1] | P[2] | M[1] | P[2] |
| **ChemPLP** | 4.00 | 11.11 | 1.99 | 50.88 | 2.84 | 25.69 | 0.68 | 84.69 | 2.52 | 43.08 |
| **Max. IFP Tanimoto** | | | | | | | | | | |
| • Additives | 4.90 | 7.41 | 4.70 | 10.53 | 3.45 | 16.51 | 1.7 | 56.12 | 3.92 | 22.64 |
| • Fragments | 3.36 | 22.22 | 1.35 | 59.65 | 2.04 | 46.79 | 0.77 | 72.45 | 1.98 | 50.28 |
| • Drug-like | 3.96 | 18.52 | 3.57 | 33.33 | 2.03 | 47.71 | 0.97 | 84.69 | 2.35 | 46.06 |
| • Fragments and drug-like | 3.12 | 25.93 | 1.50 | 56.14 | 2.05 | 45.87 | 0.77 | 74.49 | 1.97 | 50.60 |
| • All | 3.12 | 25.93 | 1.50 | 56.14 | 2.05 | 45.87 | 0.88 | 72.45 | 1.99 | 50.09 |
| **Consensus IFP Tanimoto** | | | | | | | | | | |
| • Additives | 5.3 | 7.41 | 4.95 | 8.77 | 4.43 | 8.26 | 3.56 | 27.55 | 4.74 | 12.99 |
| • Fragments | 3.12 | 11.11 | 2.27 | 45.61 | 2.02 | 48.62 | 0.77 | 78.57 | 2.25 | 45.98 |
| • Drug-like | 4.12 | 7.41 | 2.27 | 43.86 | 1.95 | 52.29 | 0.68 | 85.71 | 2.21 | 47.32 |
| • Fragments and drug-like | 3.98 | 7.41 | 2.27 | 43.86 | 2.03 | 47.71 | 0.72 | 80.61 | 2.26 | 44.90 |
| • All | 3.98 | 7.41 | 2.27 | 45.61 | 1.97 | 53.21 | 0.82 | 71.43 | 2.29 | 44.42 |
| **Max. ROCS Tanimoto** | | | | | | | | | | |
| • Additives | 5.06 | 14.81 | 5.34 | 12.28 | 1.95 | 52.29 | 0.68 | 82.65 | 2.88 | 40.51 |
| • Fragments | 2.88 | 33.33 | 1.23 | 63.16 | 1.83 | 57.80 | 0.68 | 82.65 | 1.55 | 59.23 |
| • Drug-like | 3.12 | 25.93 | 1.35 | 63.16 | 1.98 | 50.46 | 0.94 | 66.33 | 1.94 | 51.47 |
| • Fragments and drug-like | 2.88 | 33.33 | 1.24 | 64.91 | 1.83 | 57.80 | 0.68 | 82.65 | 1.62 | 59.67 |
| • All | 2.88 | 33.33 | 1.24 | 64.91 | 1.84 | 56.88 | 0.68 | 84.69 | 1.62 | 59.95 |
| **Consensus ROCS Tanimoto** | | | | | | | | | | |
| • Additives | 5.24 | 7.41 | 5.57 | 10.53 | 4.50 | 2.75 | 1.82 | 52.04 | 4.98 | 18.17 |
| • Fragments | 2.73 | 33.33 | 1.13 | 68.42 | 1.83 | 58.72 | 0.67 | 82.65 | 1.59 | 60.78 |
| • Drug-like | 3.08 | 25.93 | 1.97 | 52.63 | 1.92 | 54.13 | 0.92 | 72.45 | 1.97 | 51.28 |
| • Fragments and drug-like | 2.73 | 33.33 | 1.13 | 68.42 | 1.83 | 58.72 | 0.67 | 82.65 | 1.59 | 60.78 |
| • All | 2.73 | 33.33 | 1.13 | 68.42 | 1.83 | 58.72 | 1.82 | 52.04 | 1.89 | 53.13 |
| **Best (control)** | 2.15 | 44.44 | 1.02 | 85.96 | 1.22 | 86.24 | 0.64 | 96.94 | 1.11 | 78.40 |

[1] median of RMSDs; [2] percentage of poses with RMSD below 2 Å; rescoring schemes with the same or a better performance as compared to the docking scoring function are highlighted in green

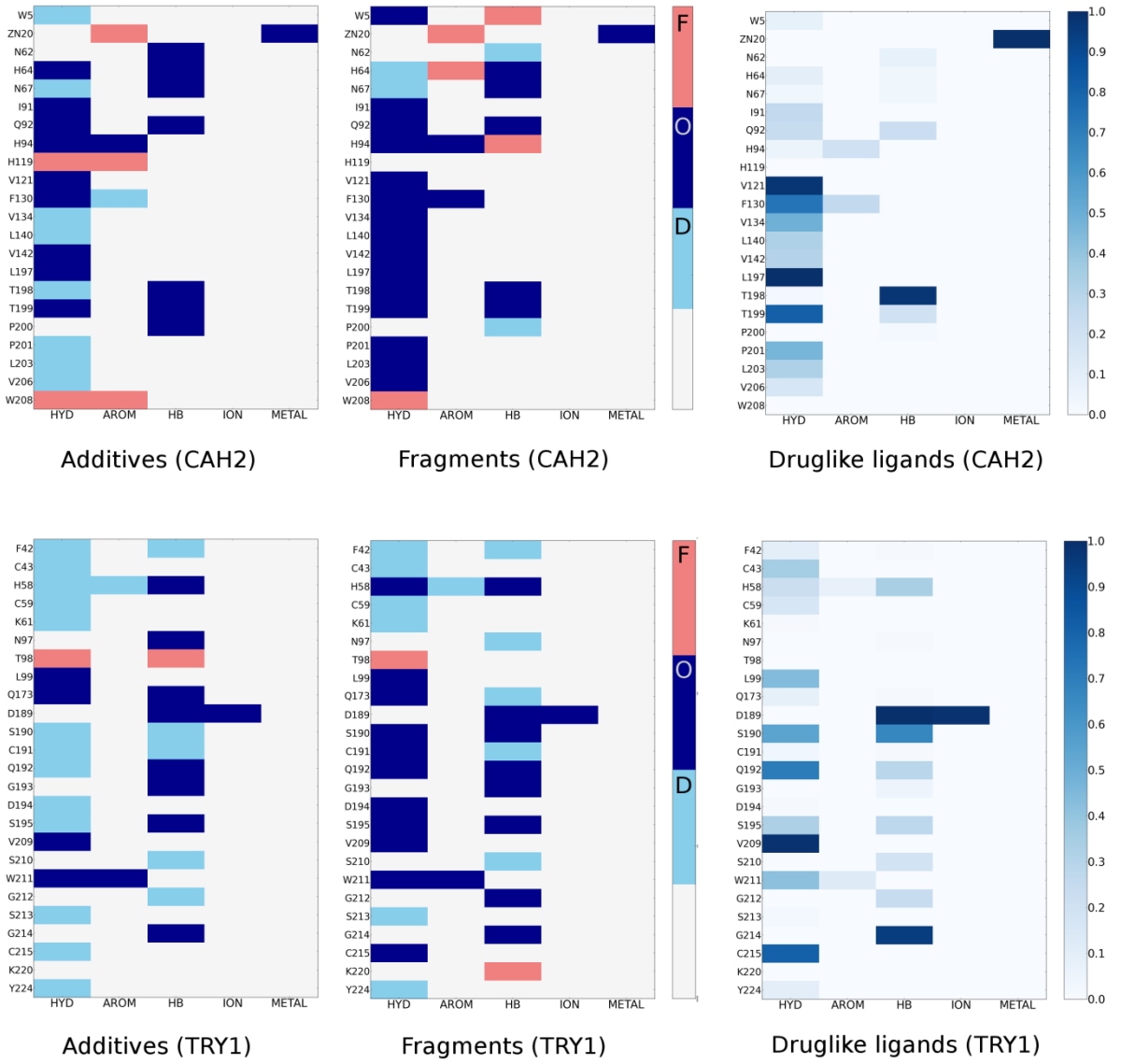
**Supplementary Table S6. Drug-like ligand docking rescoring performance**

| Scoring | BACE1 | | CAH2 | | CDK2 | | TRY1 | | All targets | |
|---|---|---|---|---|---|---|---|---|---|---|
| | M[1] | P[2] | M[1] | P[2] | M[1] | P[2] | M[1] | P[2] | M[1] | P[2] |
| **ChemPLP** | 4.58 | 25.23 | 1.63 | 58.82 | 3.64 | 18.37 | 1.04 | 65.82 | 2.71 | 42.04 |
| **Max. IFP Tanimoto** | | | | | | | | | | |
| • Additives | 6.74 | 7.01 | 5.84 | 6.95 | 4.06 | 8.16 | 4.45 | 12.66 | 5.01 | 8.69 |
| • Fragments | 5.20 | 8.88 | 2.75 | 39.04 | 3.56 | 28.57 | 4.16 | 10.13 | 2.93 | 36.91 |
| • Drug-like | 3.88 | 27.10 | 1.81 | 55.08 | 3.28 | 32.65 | 2.42 | 40.51 | 2.74 | 38.82 |
| • Fragments and drug-like | 4.00 | 27.10 | 1.85 | 53.48 | 3.08 | 36.73 | 2.94 | 30.38 | 3.88 | 21.64 |
| • All | 4.00 | 27.10 | 1.85 | 53.48 | 3.08 | 36.73 | 3.03 | 29.11 | 2.94 | 36.60 |
| **Consensus IFP Tanimoto** | | | | | | | | | | |
| • Additives | 6.37 | 1.87 | 5.90 | 5.35 | 4.40 | 4.08 | 4.28 | 17.72 | 5.22 | 7.25 |
| • Fragments | 4.53 | 10.28 | 2.58 | 41.18 | 3.28 | 20.41 | 3.98 | 10.13 | 3.55 | 25.08 |
| • Drug-like | 4.09 | 23.36 | 2.01 | 49.73 | 3.55 | 28.57 | 3.67 | 11.39 | 3.39 | 28.26 |
| • Fragments and drug-like | 4.09 | 22.90 | 2.02 | 49.20 | 3.36 | 18.37 | 3.92 | 7.59 | 3.75 | 20.49 |
| • All | 4.09 | 22.90 | 2.01 | 49.73 | 3.63 | 16.33 | 3.60 | 12.66 | 3.40 | 24.82 |
| **Max. ROCS Tanimoto** | | | | | | | | | | |
| • Additives | 5.48 | 7.48 | 5.81 | 7.49 | 3.42 | 28.57 | 2.31 | 40.51 | 4.25 | 21.00 |
| • Fragments | 4.93 | 13.55 | 1.68 | 63.10 | 2.75 | 38.78 | 3.97 | 25.32 | 2.84 | 35.17 |
| • Drug-like | 3.45 | 33.18 | 1.47 | 71.66 | 2.33 | 42.86 | 1.05 | 74.68 | 1.77 | 55.57 |
| • Fragments and drug-like | 3.45 | 33.18 | 1.45 | 71.12 | 2.10 | 46.94 | 1.05 | 74.68 | 1.72 | 56.46 |
| • All | 3.45 | 33.18 | 1.45 | 71.12 | 2.10 | 46.94 | 1.05 | 74.68 | 1.72 | 56.46 |
| **Consensus ROCS Tanimoto** | | | | | | | | | | |
| • Additives | 5.90 | 2.80 | 5.75 | 6.95 | 4.37 | 8.16 | 3.25 | 35.44 | 4.86 | 13.33 |
| • Fragments | 5.07 | 15.89 | 1.56 | 62.57 | 2.88 | 32.65 | 3.72 | 29.11 | 2.93 | 35.03 |
| • Drug-like | 3.59 | 28.97 | 1.69 | 63.10 | 2.42 | 36.73 | 1.05 | 64.56 | 2.06 | 48.32 |
| • Fragments and drug-like | 5.07 | 15.89 | 1.56 | 62.57 | 2.88 | 32.65 | 3.72 | 29.11 | 2.93 | 35.03 |
| • All | 5.07 | 15.89 | 1.56 | 62.57 | 2.88 | 32.65 | 3.25 | 35.44 | 2.87 | 36.62 |
| **Best (control)** | 2.49 | 44.86 | 1.18 | 83.96 | 1.49 | 67.35 | 0.89 | 77.22 | 1.36 | 68.32 |

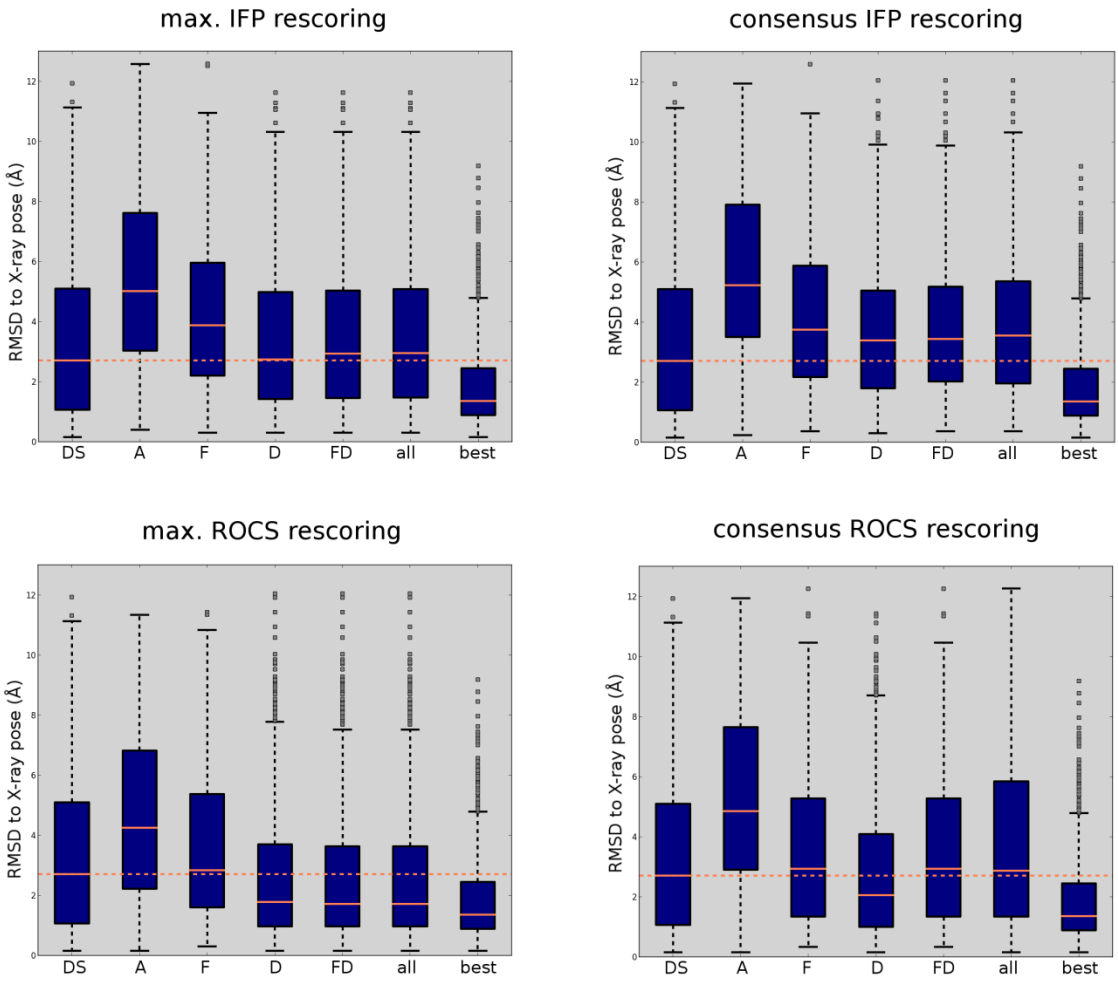
[1] median of RMSDs; [2] percentage of poses with RMSD below 2 Å; rescoring schemes with the same or a better performance as compared to the docking scoring function are highlighted in green




**Supplementary Figure S1.** The first (PC1) and second (PC2, upper panel) as well as the first and third (PC3, lower panel) components of a PCA are shown for additives, fragments and drug-like ligands from the dataset. The variance of the data explained is indicated in brackets. The datasets of this study are shown as colorful circles: BACE1 (orange), CDK2 (green), CAH2 (blue) and TRY1 (red). The grey circles indicate all additives, fragments and ligands present in the filtered PDB database. Each circle represents a specific compound.




**Supplementary Figure S2. Interaction heatmaps for CAH2 and TRY1.** The interaction heatmaps for additives (left), fragments (middle) and drug-like ligands (right) are shown for the targets CAH2 (top panel) and TRY1 (bottom panel). Pocket residues taking part in ligand interactions are shown on the y-axis while different interaction types are shown on the x-axis. Five different interaction types are distinguished: hydrophobic interactions (HYD), aromatic interactions (AROM) including face-to-face and edge-to-face π-π interactions as well as π-cation interactions, hydrogen bonding interactions (HB), ionic interactions (ION) and interactions with metal ions (METAL). The heatmap for drug-like ligand shows the frequency of each interaction in all protein complexes containing drug-like molecules, with dark blue encoding a high frequency. On the other hand, the heatmaps for additives and fragments show the differences in interactions between these sets and the drug-like ligands. Interactions occurring only with drug-like ligands (D) are shown in light blue, those occurring only with additives/fragments (F) in red and the interactions occurring in both sets (overlap, O) in dark blue.




**Supplementary Figure S3. Ligand docking rescoring.** Boxplots of the RMSD of the selected docking pose to the original X-ray structure pose (Å). Four rescoring schemes are evaluated: maximal IFP similarity (upper left), consensus IFP similarity (upper right), maximal ROCS similarity (lower left) and consensus ROCS similarity (lower right). In each panel, different molecule sets are used for rescoring: (A) additives, (F) fragments, (D) drug-like ligands, (FD) fragments and drug-like ligands, (all) additives, fragments and drug-like ligands. DS indicates the use of the original docking score (ChemPLP) and best is a control, indicating the best solution among the docking poses for each molecule. The horizontal dotted orange line indicates the median of the RMSDs of poses selected by the docking scoring function.