

## 1 Structures Excluded from PDBbind Core Sets

The following structures were excluded from the PDBbind core sets due to parsing errors or the protein or the ligand.

**2007 core:** 1a08, 1a1b, 1d09, 1is0, 1kv5, 1ols, 1olu, 1v16, 1xd1, 2h3e, 4tim, 4tmn, 7cpa, 8cpa.

**2013 core:** 1jyq, 1kel, 1os0, 1vso, 2pq9, 2qft, 2x97, 2xy9, 2zcq, 2zcr, 3fk1, 3i3b, 3muz, 3vd4, 4tmn.

**2016 core:** 1bzc, 1vso, 2zcq, 2zcr, 4tmn, 5tmn.

## 2 Algorithm Selection and RF Hyperparameter Tuning

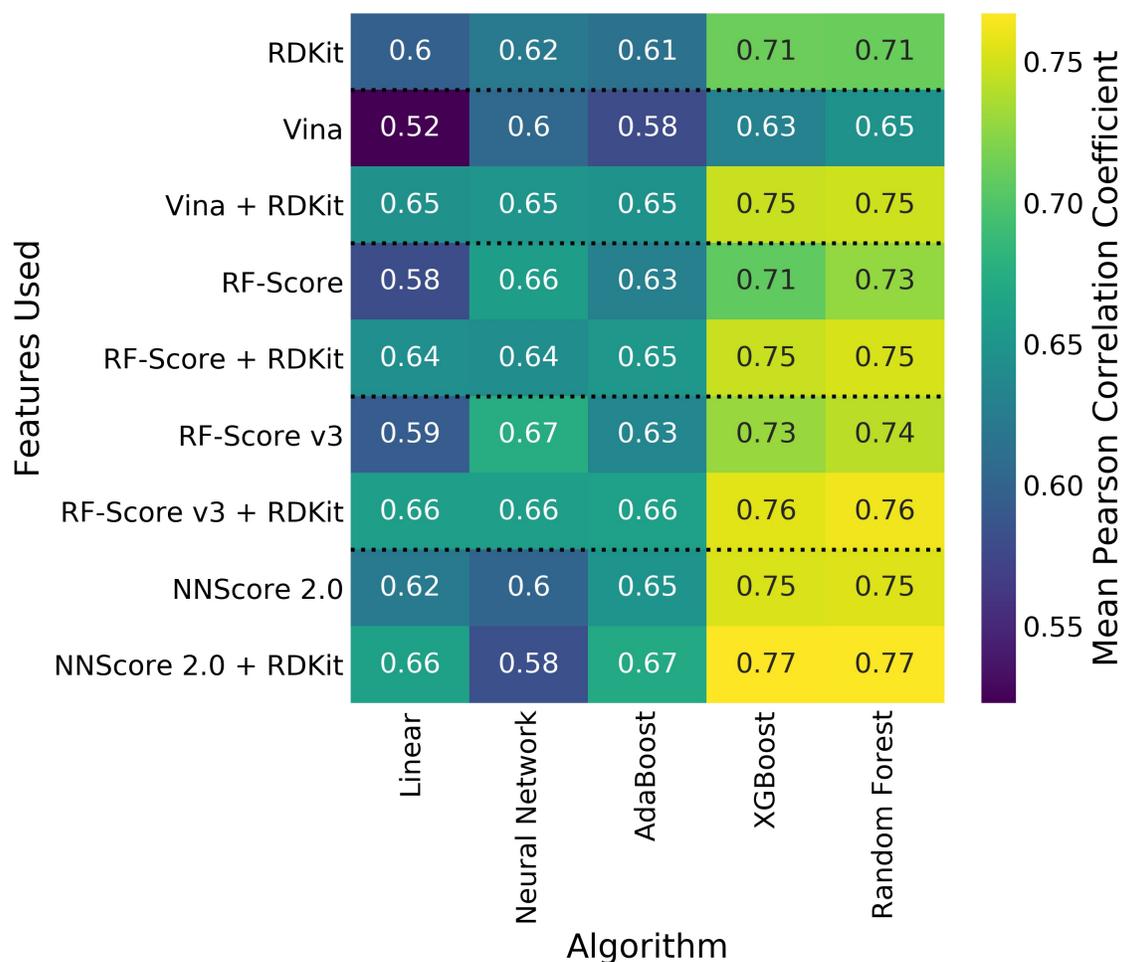


Figure 1: Performance of five regression algorithms using each feature set under stratified five-fold cross-validation using the PDBbind 2018 general set, excluding any structures present in the 2007, 2013, or 2016 core sets. Hyper-parameters for each algorithm were first tuned using randomized search with cross-validation. XGBoost and Random Forest regression consistently outperform a linear model, a single-layer neural network, and AdaBoost using shallow decision trees. For all algorithms except the neural network, the combination of RDKit descriptors and AutoDock Vina terms achieves comparable performance to RF-Score, RF-Score v3, and NNScore 2.0. In the case of the neural network, combining the NNScore 2.0 features with the RDKit descriptors appears to slightly degrade performance. This may be a result of the large size of both of these feature sets; it is possible that allowing the random search to sample larger hidden layer sizes for this feature set could reveal a more optimal neural network architecture.

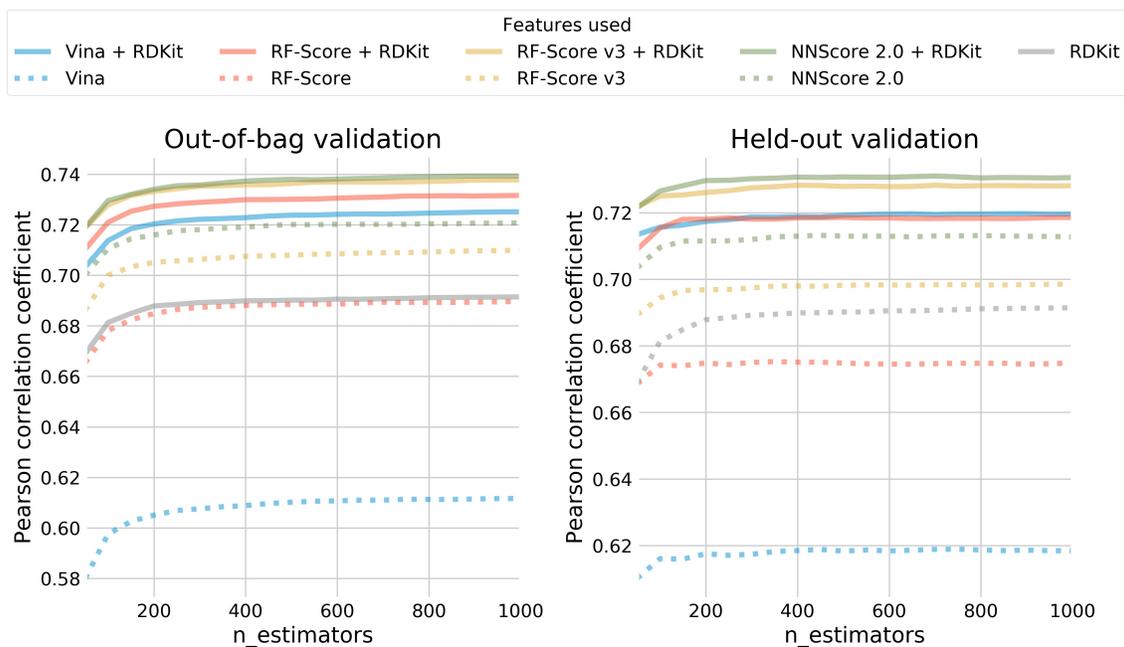


Figure 2: Tuning RF parameter  $n\_estimators$  using the PDBbind 2018 general set. A randomly-chosen subset of 20% of the general set was held out for validation, with the remaining 80% of the general set used for training and out-of-bag validation. Performance of all scoring functions plateaus around  $n\_estimators=500$  for both the out-of-bag and the held-out validation.

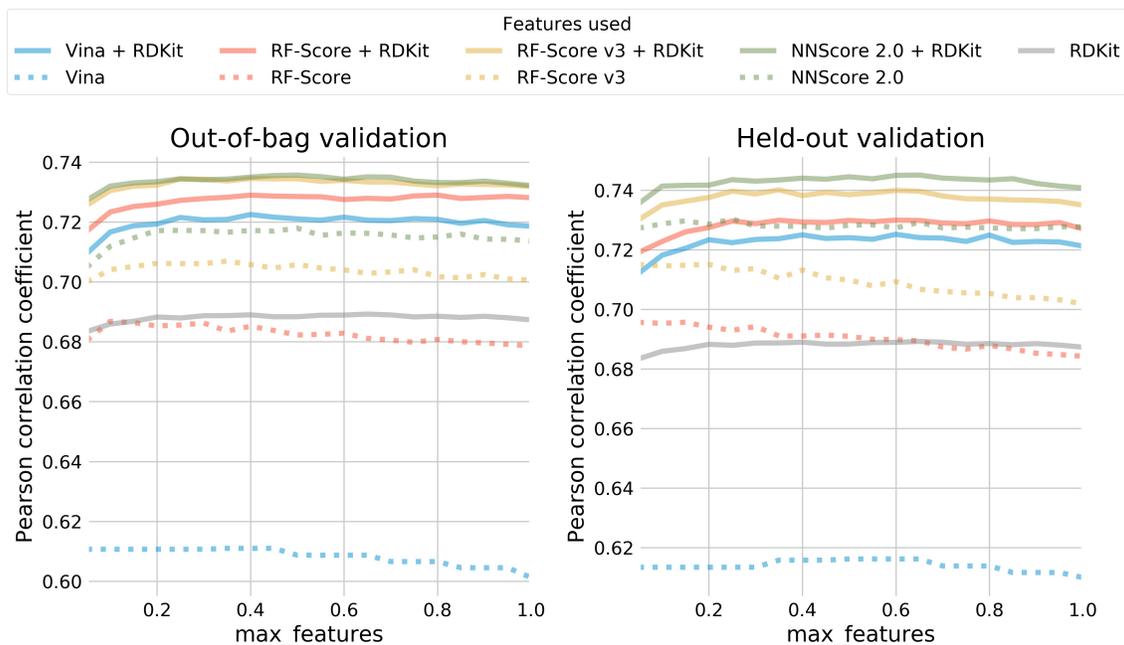


Figure 3: Tuning RF parameter *max\_features* using the PDBbind 2018 general set. A randomly-chosen subset of 20% of the general set was held out for validation, with the remaining 80% of the general set used for training and out-of-bag validation. Optimal performance on out-of-bag validation is obtained for values of *max\_features* between 0.2 and 0.4. RF-Score and RF-Score v3 are notable outliers in held-out validation, with performance beginning to drop off by *max\_features*=0.2.

### 3 Feature Importance

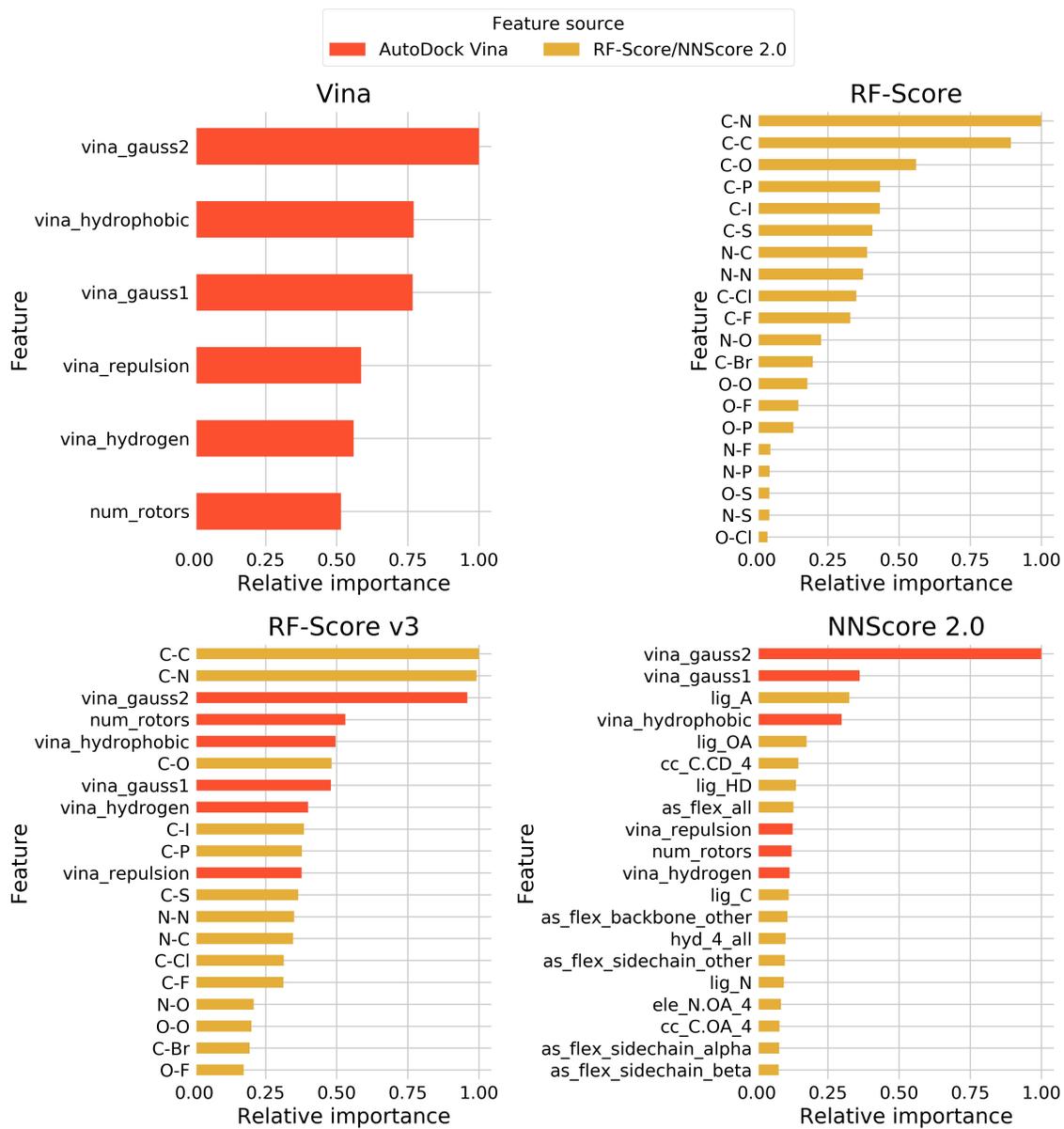


Figure 4: Relative feature importance of the highest-ranking features in RF scoring functions using the AutoDock Vina, RF-Score, RF-Score v3, and NNScore 2.0 features. Terms from AutoDock Vina are shown in red, other features from RF-Score or NNScore 2.0 are shown in yellow.

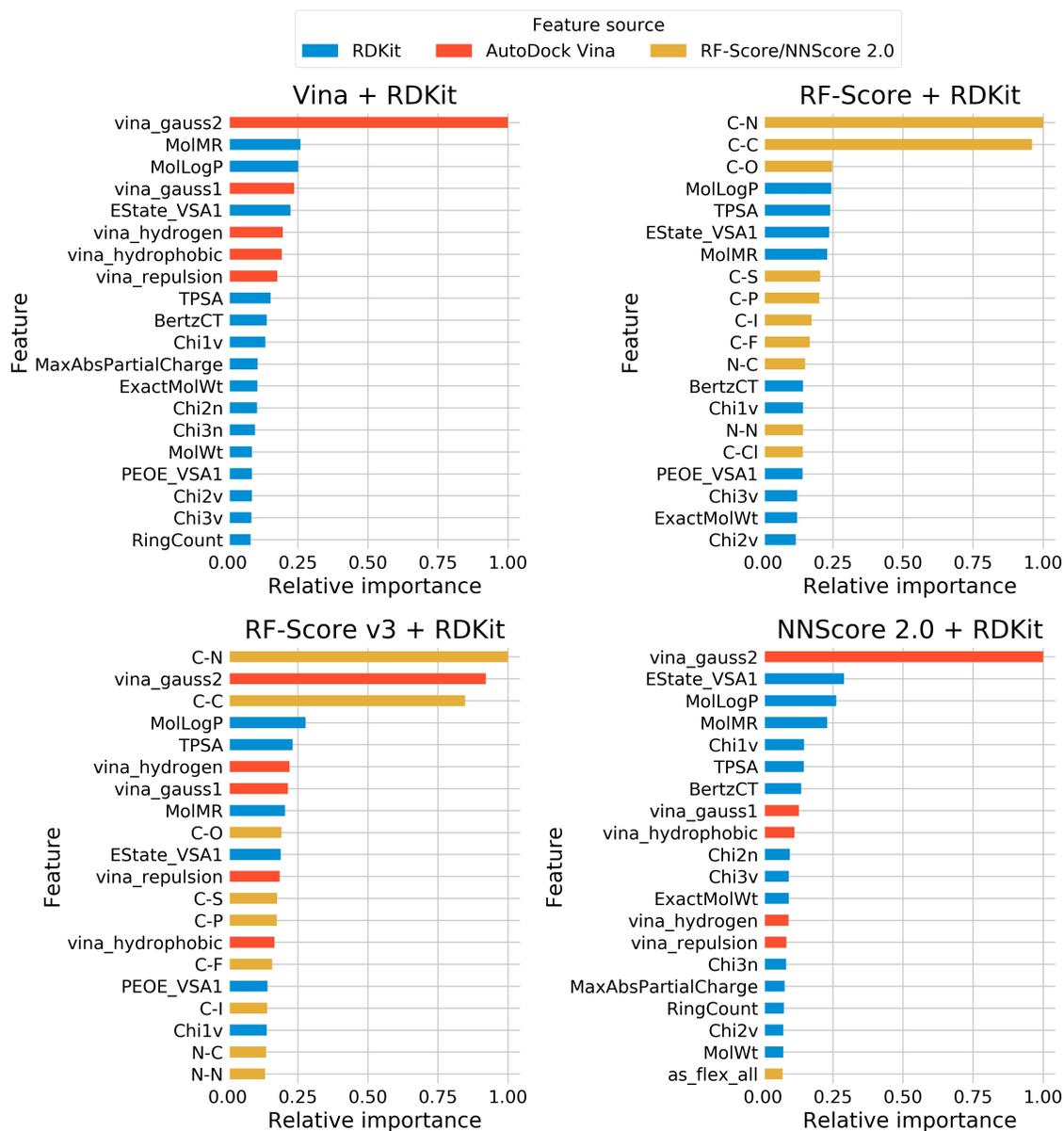


Figure 5: Relative feature importance of the highest-ranking features in RF scoring functions function the AutoDock Vina, RF-Score, RF-Score v3, and NNScore 2.0 features, augmented by the RDKit descriptors. The RDKit descriptors are shown in blue, terms from AutoDock Vina are shown in red, other features from RF-Score or NNScore 2.0 are shown in yellow. In all four scoring functions, both structure-based (red/yellow) and ligand-based (blue) features are ranked highly. Furthermore, the same RDKit descriptors (including MolLogP, MolMR, TPSA, BertzCT) are consistently ranked highly regardless of which structure-based features they are combined with.

## 4 Additional Results

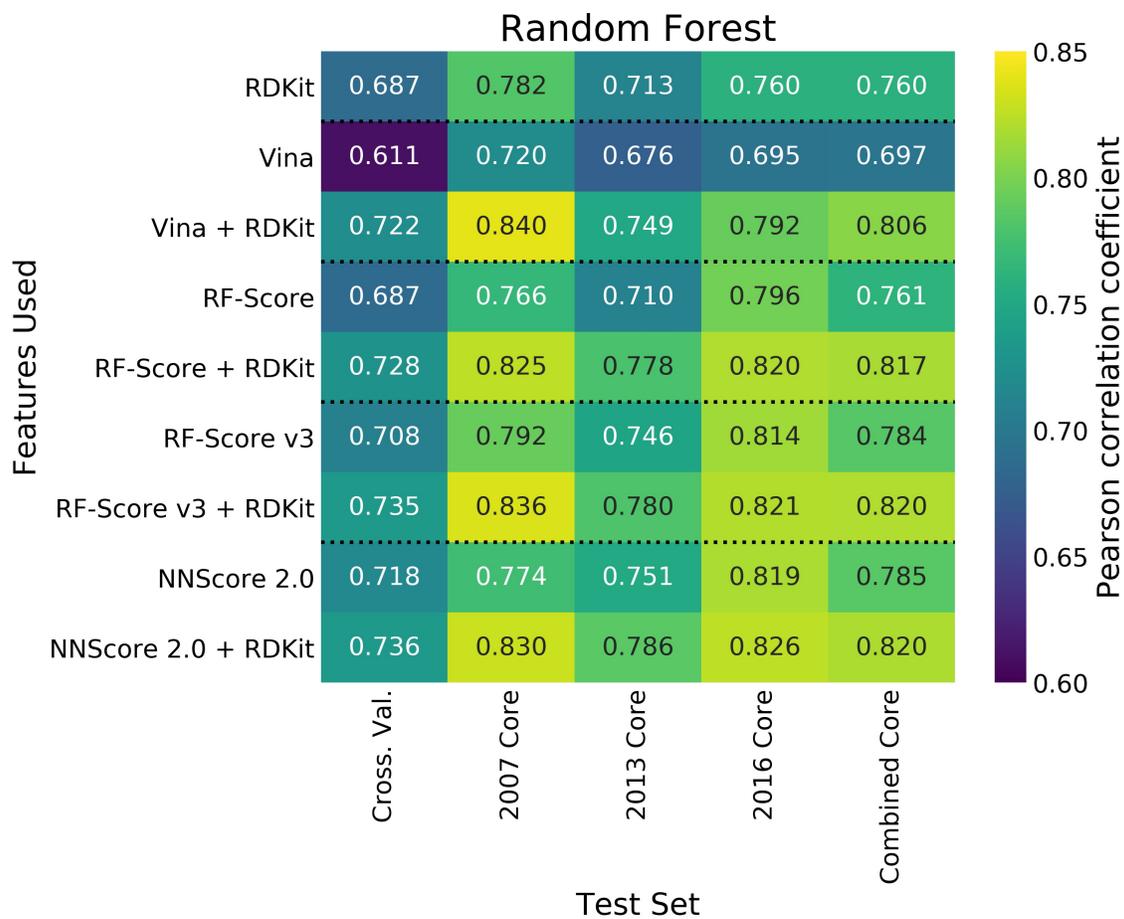


Figure 6: Pearson correlation coefficient on PDBbind core sets, and average Pearson correlation coefficient on five-fold cross-validation, for a Random Forest using each set of features.

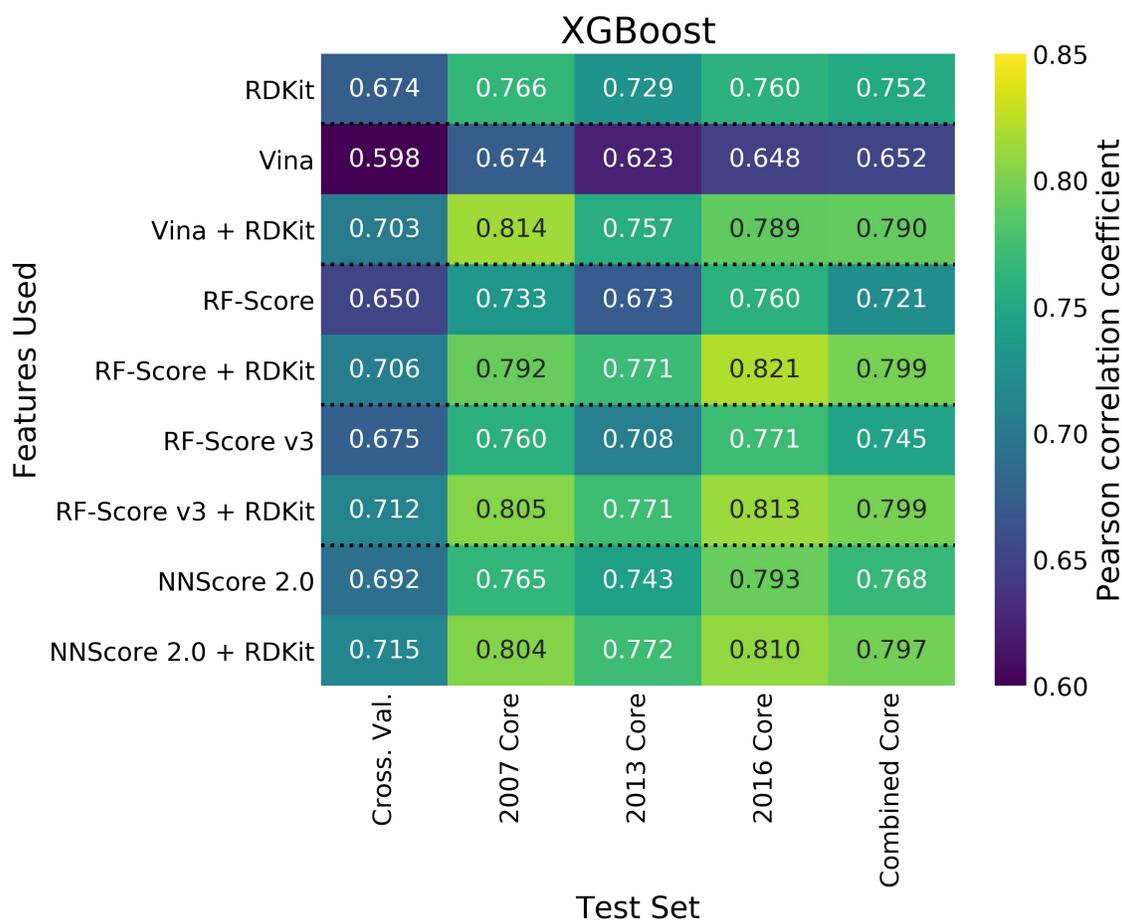


Figure 7: Pearson correlation coefficient on PDBbind core sets, and average Pearson correlation coefficient on five-fold cross-validation, for an XGBoost regression model using each set of features. XGBoost hyper-parameters were tuned using randomized search with five-fold cross-validation on the PDBbind 2018 general set, stratified by experimental  $pK$  value.

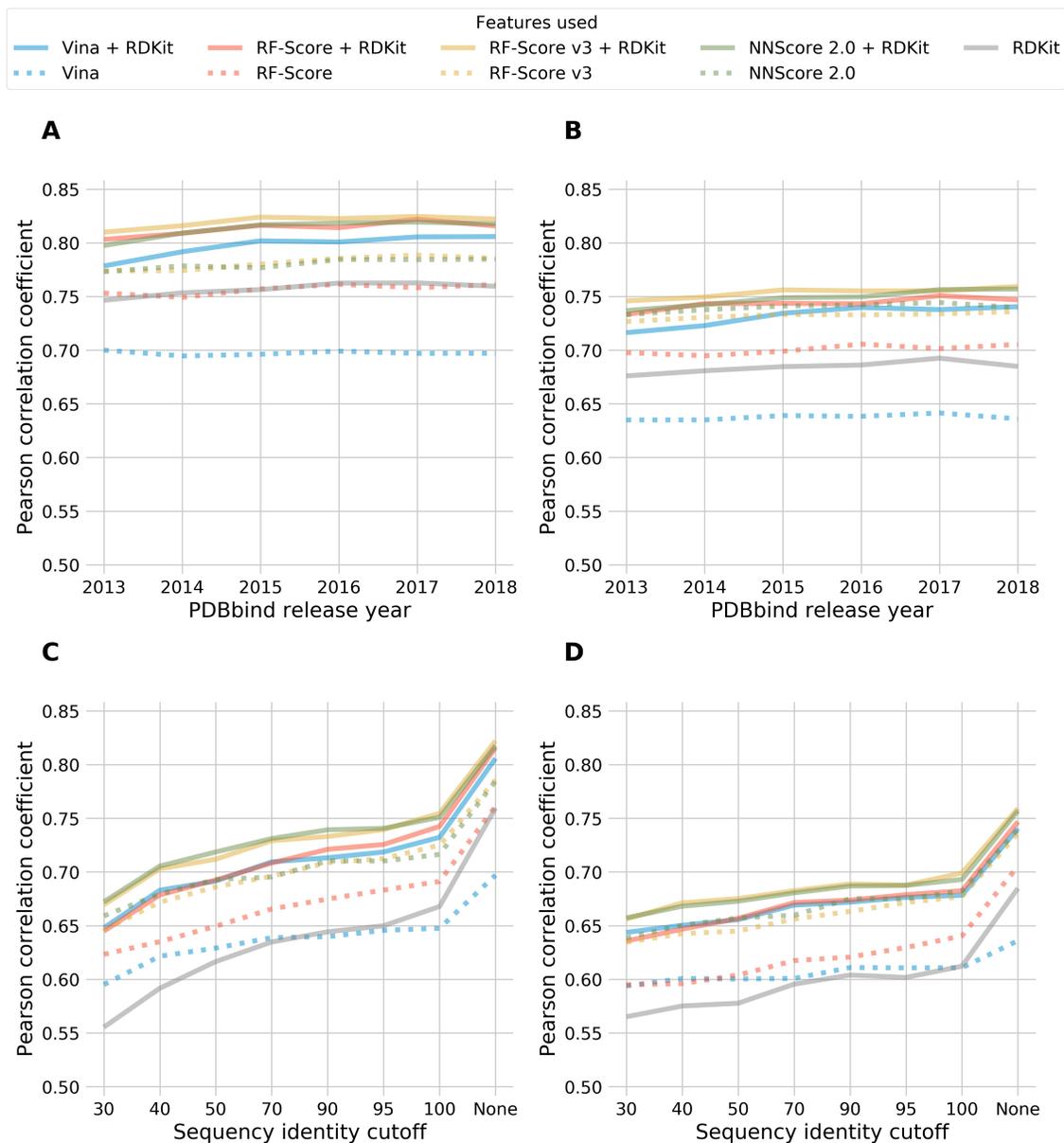


Figure 8: Pearson correlation coefficient for scoring function predictions against experimental  $pK$  values for the combined core set. Scoring functions trained on data drawn from the PDBbind general set. (A, B): Varying training set size chronologically; the version of the PDBbind general set is indicated on the  $x$ -axis. (C, D): Varying training set composition by removing structures with high protein sequence identity to any test set proteins from the PDBbind 2018 general set. The sequence identity threshold above which proteins were excluded is indicated on the  $x$ -axis. (A, C): No data excluded on the basis of ligand similarity. (B, D): Structures whose ligand has a Tanimoto similarity of 0.9 or greater to any ligand in the test set were excluded.

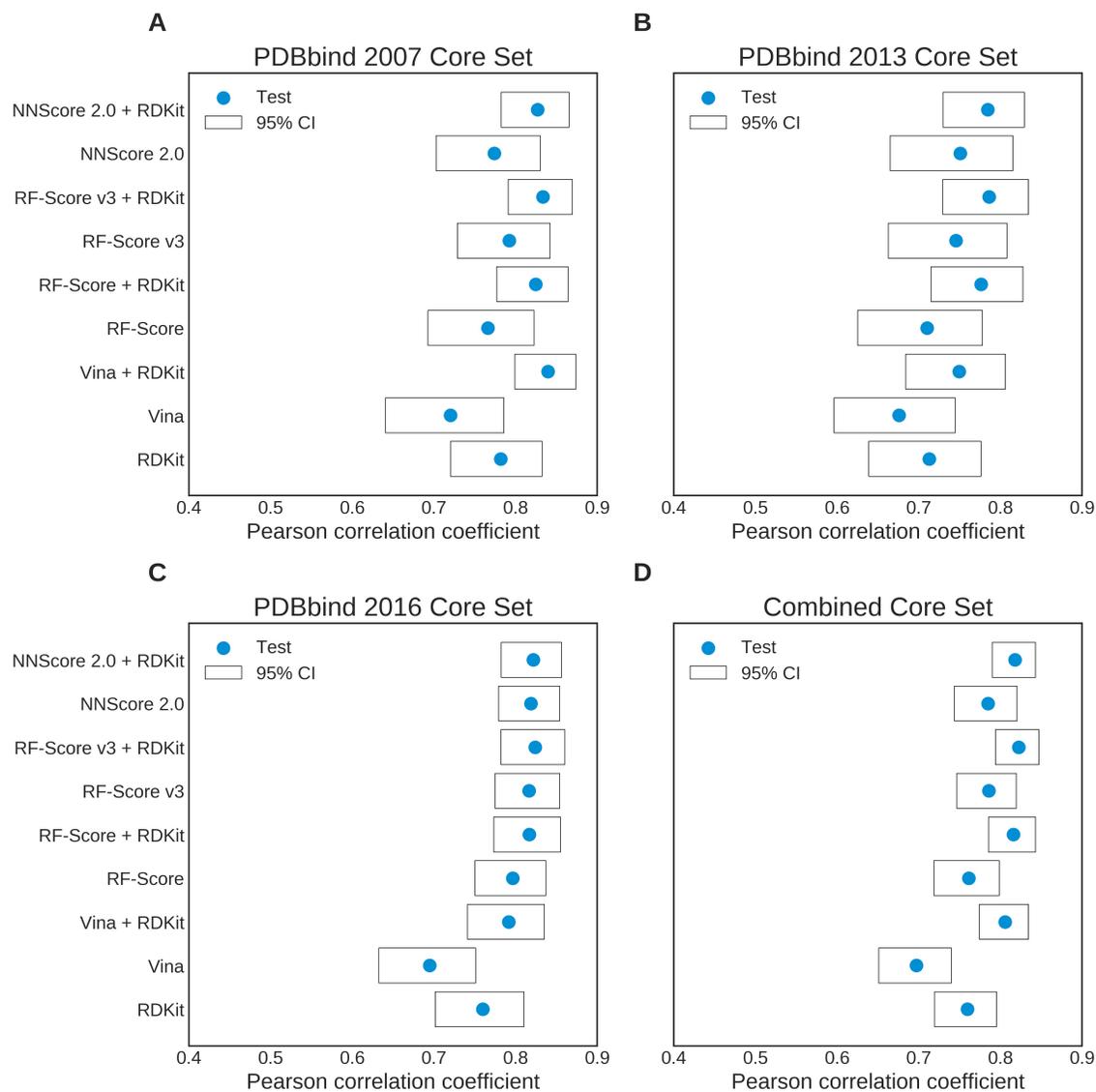


Figure 9: Bootstrapped 95% confidence interval on the Pearson correlation coefficient achieved by each RF scoring function on each test set when trained on the PDBbind 2018 general set. Confidence intervals are wider on the 2007 and 2013 core sets ( $n = 200$ ), and are most narrow on the combined core set ( $n=525$ ).

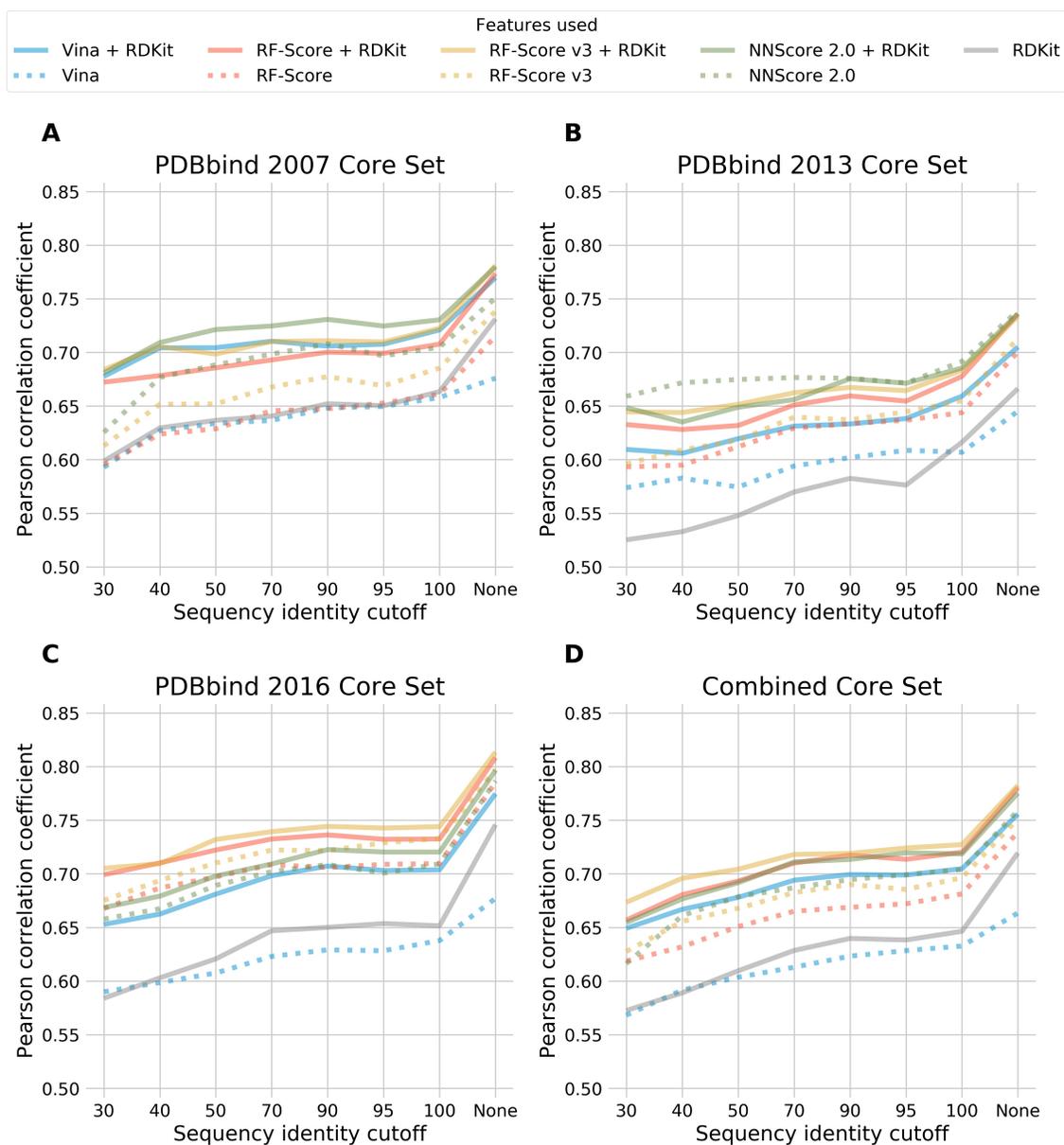


Figure 10: Scoring function performance on PDBbind core sets. Random forest models trained on PDBbind 2018 refined set; structures with protein sequence identity to any test set protein above a defined cutoff (shown on  $x$ -axis) were excluded from the training set.

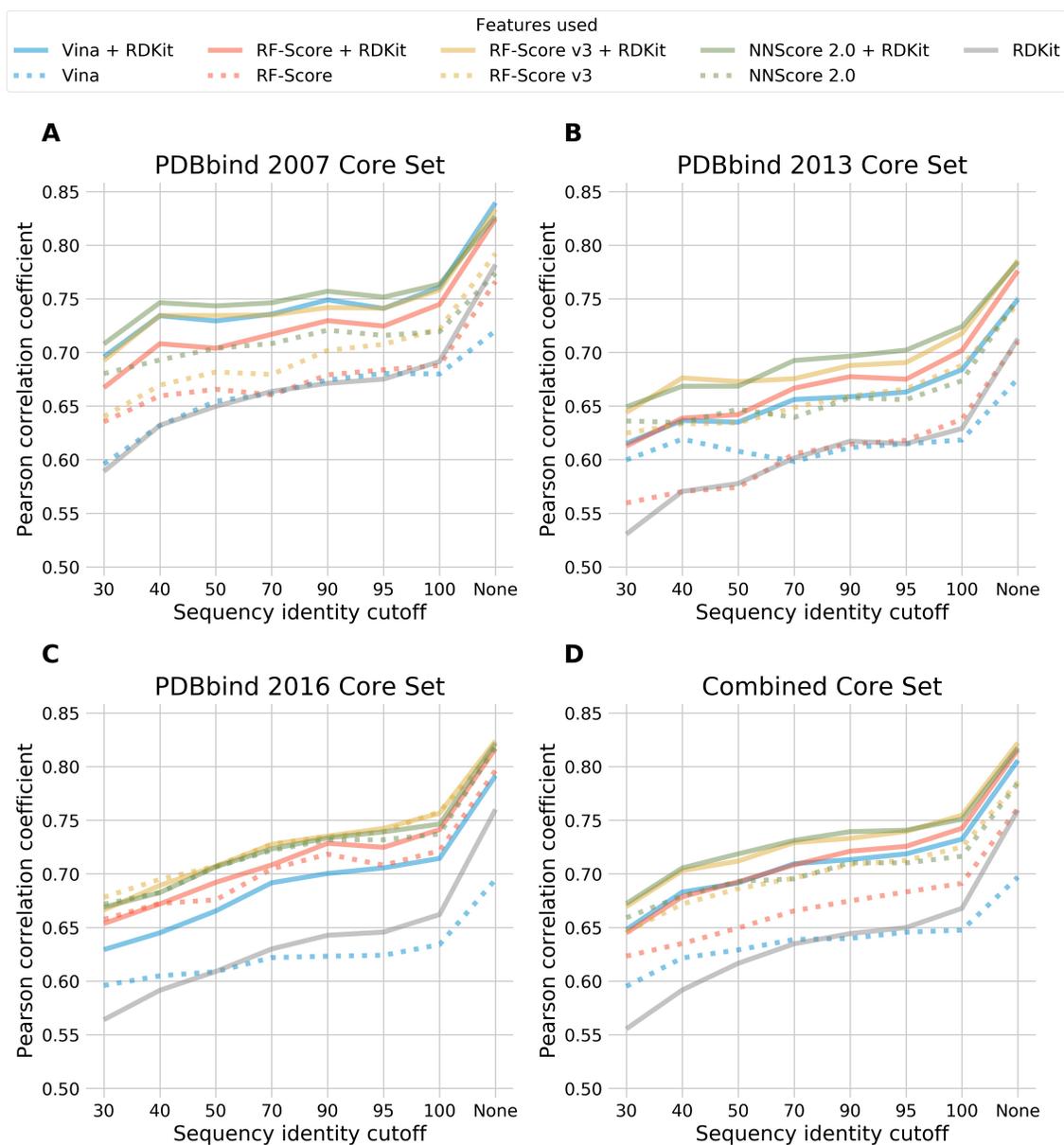


Figure 11: Scoring function performance on PDBbind core sets. Random forest models trained on PDBbind 2018 general set; structures with protein sequence identity to any test set protein above a defined cutoff (shown on  $x$ -axis) were excluded from the training set.

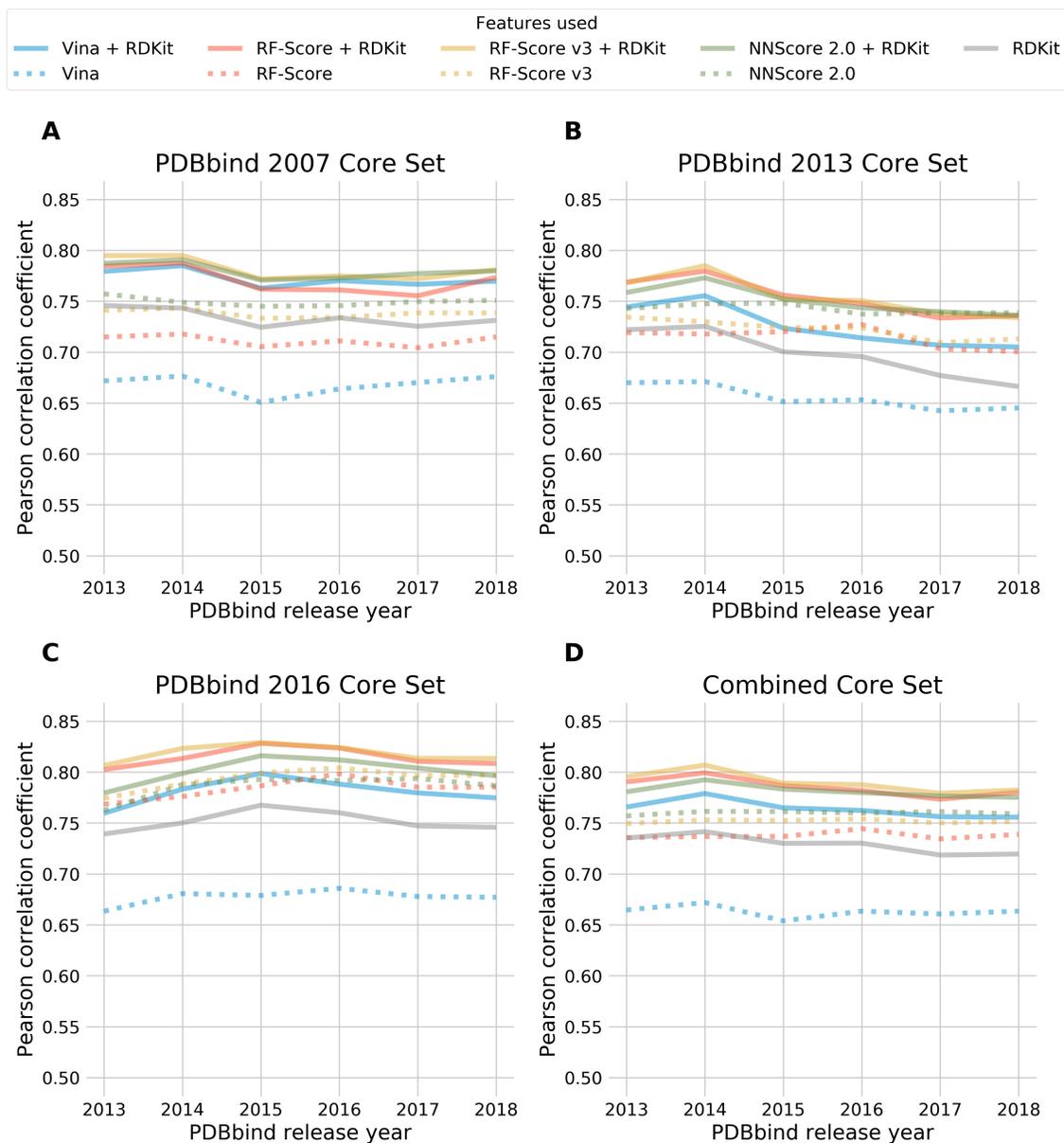


Figure 12: Scoring function performance on PDBbind core sets. Random forest models trained on different released of the PDBbind 2018 refined set. The release year used for each training set is shown on the  $x$ -axis. No data were excluded on the basis of protein or ligand similarity.

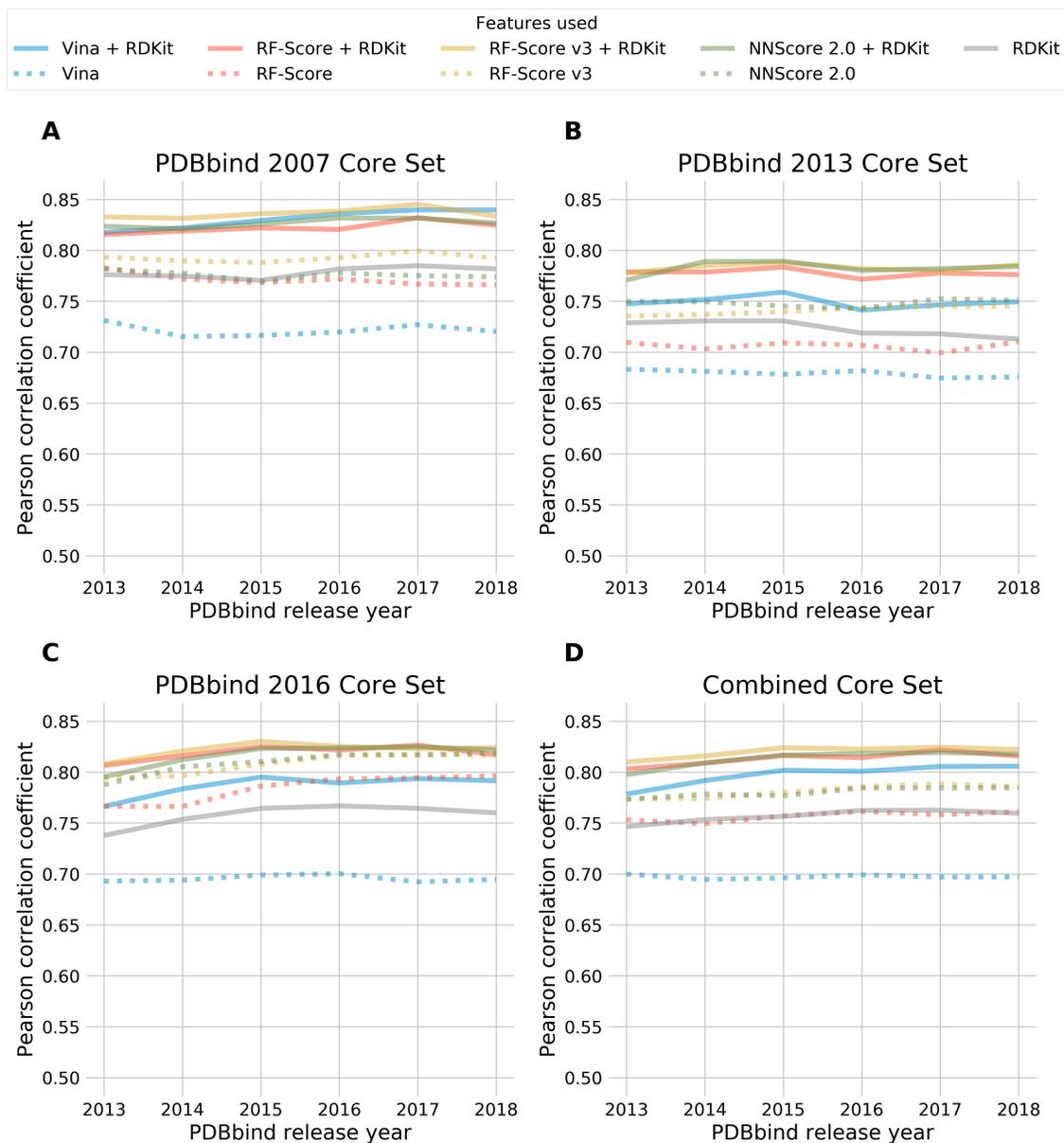


Figure 13: Scoring function performance on PDBbind core sets. Random forest models trained on different released of the PDBbind 2018 general set. The release year used for each training set is shown on the  $x$ -axis. No data were excluded on the basis of protein or ligand similarity.

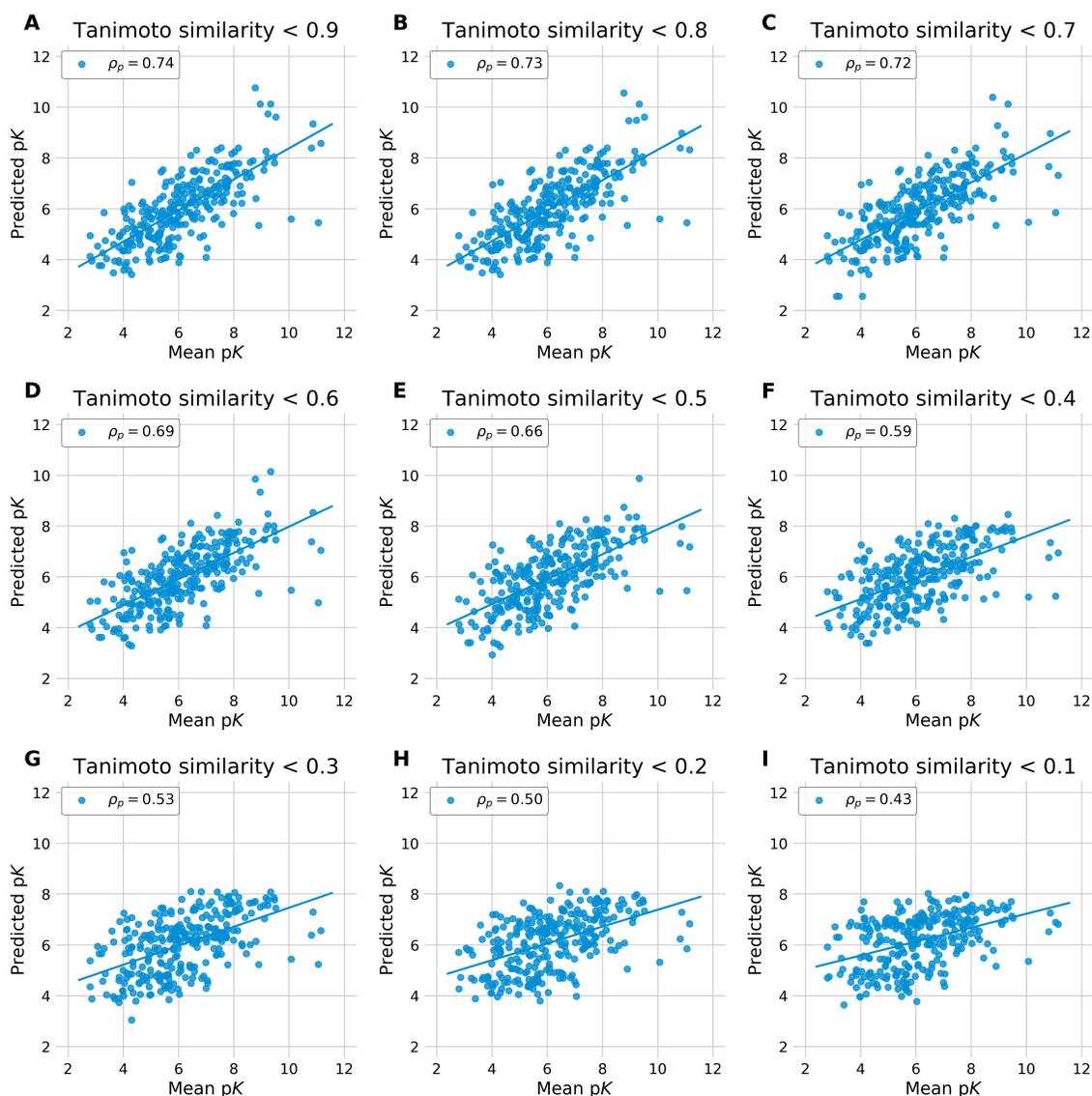


Figure 14: RDKit RF model pK predictions against mean experimental pK for ligands found in multiple structures in the PDBbind 2018 general set, for varying values of the Tanimoto similarity cutoff above which similar ligands were excluded from the training set. Each marker represents a single ligand; the Pearson correlation coefficient for the predictions is indicated in the legend of each plot, with the line indicating a linear regression fit through the points. The Tanimoto similarity cutoff ranges from 0.9 (A) to 0.1 (I) in increments of 0.1. The correlation remains strong ( $\rho_p$ ) when ligands with Tanimoto similarity greater than 0.7 to the test ligand are excluded from the training data, while a moderate correlation ( $\rho_p \approx 0.6$ ) remains when ligands with Tanimoto similarity greater than 0.4 to the test ligand are excluded from the training data (F). When ligands with Tanimoto similarity greater than 0.1 to the test ligand are excluded from the training data, there is little correlation between the predictions of the RDKit RF model and the mean experimental pK of the ligand (I).