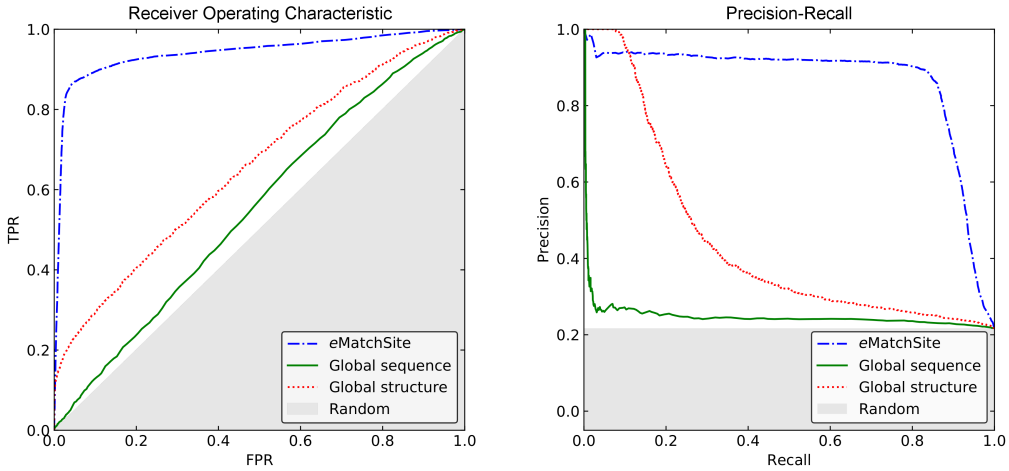
Dataset: SOIPPA

Reference: Xie L, Bourne PE. (2008) Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments. *Proc Natl Acad Sci U S A* 105: 5441-6, DOI: 10.1073/pnas.0704422105.

Description: The SOIPPA dataset comprises 226 adenine-binding proteins as well as 92 control proteins that do not bind ligands containing the adenine moiety. Ligands included in this dataset are adenosine diphosphate (ADP), adenosine triphosphate (ATP), flavin adenine dinucleotide (FAD), nicotinamide adenine dinucleotide (NAD), S-adenosyl-L-homocysteine (SAH), and S-adenosylmethionine (SAM). Control ligands in the SOIPPA dataset form 48 chemically similar clusters at a Tanimoto coefficient threshold of 0.7. This dataset contains 4,562 positive pairs of adenine-binding pockets and 16,458 negative pairs of adenine-binding and control pockets.

Performance: The ability of *e*MatchSite to distinguish between similar and dissimilar binding sites in the SPIPPA dataset is evaluated with the confusion matrix analysis and compared to that obtained by employing the global sequence identity and the global structure similarity.



Algorithm	AUC	ACC	MCC	TPR	FPR	PPV	F1	BM	MK
*e*MatchSite	0.941	0.925	0.789	0.875	0.061	0.799	0.836	0.815	0.764
Global sequence	0.553	0.487	0.068	0.636	0.554	0.241	0.350	0.082	0.057
Global structure	0.656	0.647	0.171	0.516	0.317	0.311	0.388	0.199	0.146

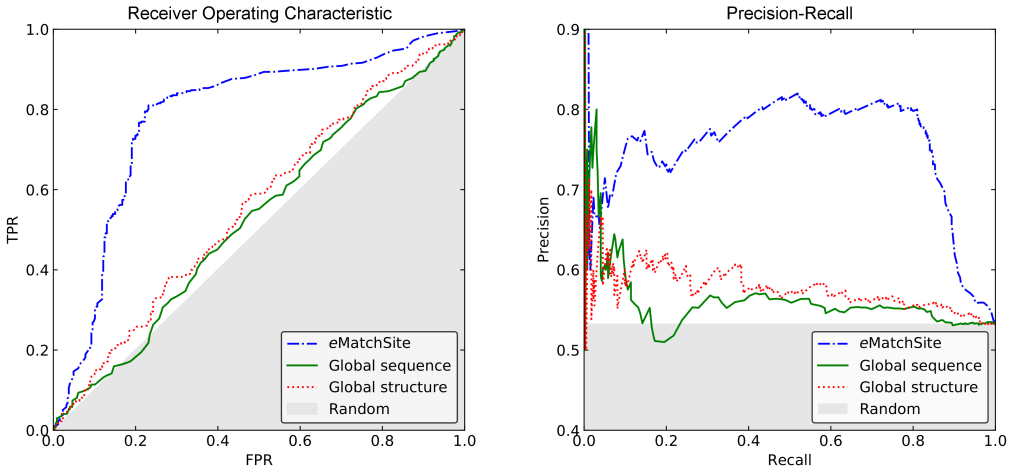
AUC – area under the curve, ACC – accuracy, MCC – Matthews correlation coefficient, TPR – true positive rate, FPR – false positive rate, PPV – precision, BM – bookmarker informedness, MK – markedness.

Dataset: Kahraman

Reference: Kahraman A, Morris RJ, Laskowski RA, Thornton JM. (2007) Shape variation in protein binding pockets and their ligands. *J Mol Biol* 368: 283-301, DOI: 10.1016/j.jmb.2007.01.086.

Description: The Kahraman dataset comprises proteins bound to adenosine monophosphate (AMP), 3-β-hydroxy-5-androsten-17-one (AND), adenosine triphosphate (ATP), estradiol (EST), flavin-adenine dinucleotide (FAD), flavin mononucleotide (FMN), α-D-glucose (GLC), heme (HEM), and nicotinamide adenine dinucleotide (NAD). Positives are defined as pairs of proteins that bind exactly the same ligand, whereas those proteins binding different ligands are considered negatives. This dataset contains 395 positive pairs (similar pockets) and 346 negative pairs (dissimilar pockets).

Performance: The ability of *e*MatchSite to distinguish between similar and dissimilar binding sites in the Kahraman dataset is evaluated with the confusion matrix analysis and compared to that obtained by employing the global sequence identity and the global structure similarity.



Algorithm	AUC	ACC	MCC	TPR	FPR	PPV	F1	BM	MK
*e*MatchSite	0.784	0.791	0.579	0.810	0.231	0.800	0.805	0.579	0.580
Global sequence	0.531	0.530	0.061	0.527	0.465	0.564	0.545	0.061	0.061
Global structure	0.559	0.544	0.080	0.603	0.523	0.568	0.585	0.079	0.080

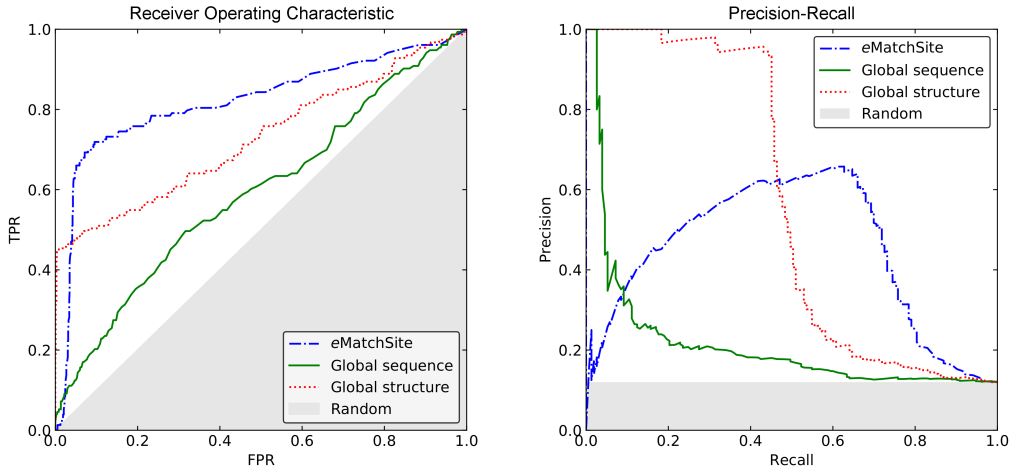
AUC – area under the curve, ACC – accuracy, MCC – Matthews correlation coefficient, TPR – true positive rate, FPR – false positive rate, PPV – precision, BM – bookmarker informedness, MK – markedness.

Dataset: Homogeneous

Reference: Hoffmann B, Zaslavskiy M, Vert JP, Stoven V. (2010) A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3D: application to ligand prediction. *BMC Bioinformatics* 11: 99, DOI: 10.1186/1471-2105-11-99.

Description: The Homogeneous dataset comprises proteins complexed with the following ligands: pentaethylene glycol (1PE), β-octylglucoside (BOG), glutathione (GSH), lauryl dimethylamine-N-oxide (LDA), palmitic acid (PLM), 4'-deoxy-4'-aminopyridoxal-5'-phosphate (PMP), S-adenosylmethionine (SAM), sucrose (SUC), and uridine monophosphate (U5P). Positives are defined as pairs of proteins that bind exactly the same ligand, whereas those proteins that bind different ligands are considered negatives. This dataset contains 153 positive pairs (similar pockets) and 1,122 negative pairs (dissimilar pockets).

Performance: The ability of *e*MatchSite to distinguish between similar and dissimilar binding sites in the Homogeneous dataset is evaluated with the confusion matrix analysis and compared to that obtained by employing the global sequence identity and the global structure similarity.



Algorithm	AUC	ACC	MCC	TPR	FPR	PPV	F1	BM	MK
*e*MatchSite	0.820	0.858	0.496	0.732	0.125	0.444	0.553	0.607	0.404
Global sequence	0.593	0.626	0.109	0.523	0.360	0.165	0.251	0.163	0.073
Global structure	0.729	0.776	0.273	0.549	0.193	0.280	0.371	0.357	0.209

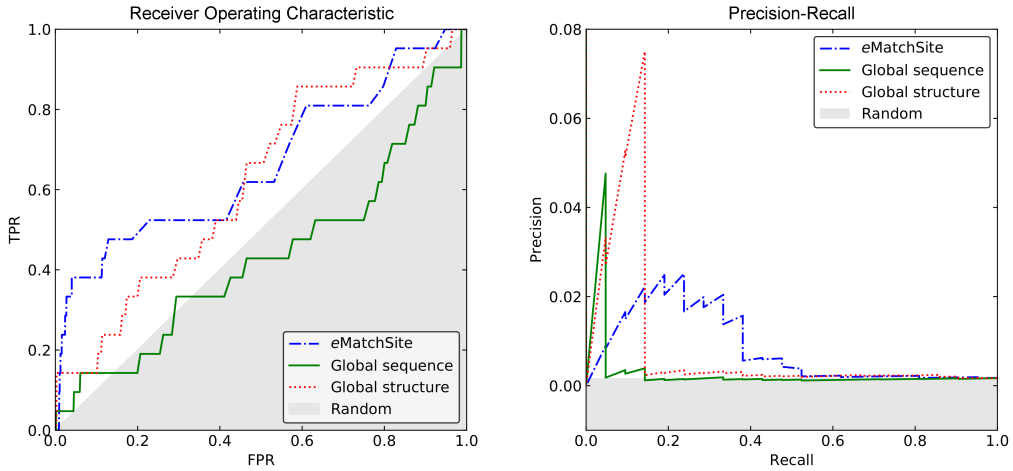
AUC – area under the curve, ACC – accuracy, MCC – Matthews correlation coefficient, TPR – true positive rate, FPR – false positive rate, PPV – precision, BM – bookmarker informedness, MK – markedness.

Dataset: Steroid

Reference: Brylinski M. (2014) *e*MatchSite: Sequence order-independent structure alignments of ligand binding pockets in protein models. *PLoS Comput Biol* 10: e1003829, DOI: 10.1371/journal.pcbi.1003829.

Description: The Steroid dataset comprises 8 pharmacologically relevant steroid-binding proteins complexed with 17β-estradiol (EST), estradiol-17β-hemisuccinate (HE7), and equilenin (EQU), as well as 1,854 control proteins binding small molecules whose size is comparable to that of steroids. Control ligands have different chemical structures with a Tanimoto coefficient vs. EST of ≤0.1, and form 334 diverse clusters at a Tanimoto coefficient threshold of 0.7. The Steroid dataset contains 21 positive pairs of steroid-binding pockets and 12,563 negative pairs of steroid-binding and control pockets.

Performance: The ability of *e*MatchSite to distinguish between similar and dissimilar binding sites in the Steroid dataset is evaluated with the confusion matrix analysis and compared to that obtained by employing the global sequence identity and the global structure similarity.



Algorithm	AUC	ACC	MCC	TPR	FPR	PPV	F1	BM	MK
*e*MatchSite	0.663	0.720	0.022	0.524	0.280	0.003	0.006	0.244	0.002
Global sequence	0.434	0.705	0.004	0.333	0.294	0.002	0.004	0.039	0.000
Global structure	0.621	0.551	0.010	0.571	0.449	0.002	0.004	0.123	0.001

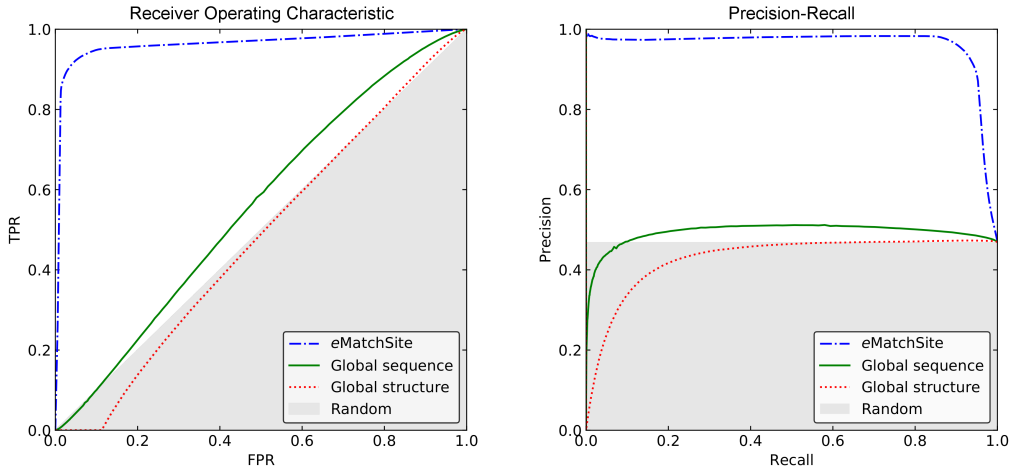
AUC – area under the curve, ACC – accuracy, MCC – Matthews correlation coefficient, TPR – true positive rate, FPR – false positive rate, PPV – precision, BM – bookmarker informedness, MK – markedness.

Dataset: TOUGH-M2

Reference: Govindaraj RG, Brylinski M (2017) DaTaset tO evalUate alGoritHms for binding site Matching. DOI: 10.17605/OSF.IO/2QZBU.

Description: The TOUGH-M2 is a non-redundant and representative set of 5,873 protein-drug complexes with computationally predicted pockets. Target proteins share less than 40% sequence identity and have globally dissimilar structures with a TM-score of <0.4. Bound ligands cluster into 1,266 groups of chemically similar molecules at a Tanimoto coefficient threshold of 0.7. The positive subset of TOUGH-M2 comprises 308,665 protein pairs having different structures, yet binding chemically similar ligands. The negative subset of TOUGH-M2 comprises 348,926 protein pairs that have different structures and bind chemically dissimilar ligands.

Performance: The ability of *e*MatchSite to distinguish between similar and dissimilar binding sites in the TOUGH-M2 dataset is evaluated with the confusion matrix analysis and compared to that obtained by employing the global sequence identity and the global structure similarity.



Algorithm	AUC	ACC	MCC	TPR	FPR	PPV	F1	BM	MK
*e*MatchSite	0.963	0.936	0.872	0.927	0.056	0.936	0.932	0.871	0.872
Global sequence	0.557	0.542	0.085	0.551	0.466	0.511	0.530	0.085	0.085
Global structure	0.479	0.494	-0.008	0.520	0.529	0.465	0.491	-0.008	-0.008

AUC – area under the curve, ACC – accuracy, MCC – Matthews correlation coefficient, TPR – true positive rate, FPR – false positive rate, PPV – precision, BM – bookmarker informedness, MK – markedness.