







SiteMine: Large-scale binding site similarity searching in protein structure databases

Thorben Reim¹  | Christiane Ehrt¹  | Joel Graef¹  | Sebastian Günther²  |
Alke Meents²  | Matthias Rarey¹ 

¹ZBH - Center for Bioinformatics, Universität Hamburg, Hamburg, Germany

²Center for Free-Electron Laser Science CFEL, Deutsches Elektronen-Synchrotron DESY, Hamburg, Germany

Correspondence

Matthias Rarey, Universität Hamburg, ZBH - Center for Bioinformatics, Albert-Einstein-Ring 8-10, Hamburg 22761, Germany.
Email: matthias.rarey@uni-hamburg.de

Funding information

Data Science in Hamburg - Helmholtz Graduate School for the Structure of Matter, Grant/Award Number: HIDSS-0002; Helmholtz Association, Grant/Award Numbers: FISCOV, SFRagX, HIR3X(InternLabs-0011)

Abstract

Drug discovery and design challenges, such as drug repurposing, analyzing protein–ligand and protein–protein complexes, ligand promiscuity studies, or function prediction, can be addressed by protein binding site similarity analysis. Although numerous tools exist, they all have individual strengths and drawbacks with regard to run time, provision of structure superpositions, and applicability to diverse application domains. Here, we introduce SiteMine, an all-in-one database-driven, alignment-providing binding site similarity search tool to tackle the most pressing challenges of binding site comparison. The performance of SiteMine is evaluated on the ProSPECCTs benchmark, showing a promising performance on most of the data sets. The method performs convincingly regarding all quality criteria for reliable binding site comparison, offering a novel state-of-the-art approach for structure-based molecular design based on binding site comparisons. In a SiteMine showcase, we discuss the high structural similarity between cathepsin L and calpain 1 binding sites and give an outlook on the impact of this finding on structure-based drug design. SiteMine is available at <https://uhh.de/naomi>.

KEYWORDS

binding site comparison, cathepsin L, drug repurposing, off-target prediction, structure-based drug design

1 | INTRODUCTION

The steadily growing number of experimentally solved and predicted protein structures and their availability in the Protein Data Bank (PDB),^[1] SWISS-MODEL,^[2] and AlphaFold Protein Structure Database^[3] lay the grounds for data-driven approaches in structure-based drug design (SBDD).^[4,5] There are many application areas for searching for similar protein binding sites. These include function and off-target prediction, protein classification, drug repurposing, and polypharmacology prediction (one drug addressing multiple targets).

Many tools have already been developed to predict and compare binding sites. Some overviews are provided elsewhere (protein pocket prediction,^[6–10] binding site comparisons^[11,12]). A recent extensive review by Eguida and Rognan^[13] gives insights into state-of-the-art binding site analyses. Further binding site comparison tools were published recently (Table 1).

Binding site comparison tools can be divided into different groups by the output they produce (only similarity scores or also structure alignments), the binding site modeling (residue-, surface-, or interaction-based), and the used data structure for similarity calculation (e.g., graphs,

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Authors. *Archiv der Pharmazie* published by Wiley-VCH GmbH on behalf of Deutsche Pharmazeutische Gesellschaft.

TABLE 1 An overview of additional binding site comparison tools published since the ProSPECCTs benchmark study from 2018.¹⁴

Year	Method	Benchmark	Identification	Modeling	Data structure	Alignment
2019	DeepDrug3D ^[15]	TOUGH-C1 ^[15]	Ligand	Interactions	3D points	No
2020	DeeplyTough ^[16]	Vertex ^[17] and ProSPECCTs ^[14] without the ROCS Structures data set ^[18]	Ligand or fpocket 2.0 ^[19]	Interactions	3D image	No
2020	ProCare ^[20]	Balanced Vertex ^[20]	VolSite ^[21]	Interactions	3D points	Yes
2021	PocketShape ^[22]	sc-PDB data set for method evaluation ^[23]	Ligand	Residues	Matrix	Yes
2021	Site2Vec ^[24]	ProSPECCTs without the ROCS Structures data set, APOCS3, ^[25] PLIC data set, ^[26] TOUGH-C1	Ligand	Residues	Histograms	No
2022	BindSiteS-CNN ^{[27]a}	TOUGH-C1, ProSPECCTs without the ROCS Structures data set	Ligand or SURFNET ^[28]	Surface	Graph	No
2023	TWN-RENCOD ^{[29]b}	Developer-built kinase data set	Ligand	Residues	Matrix	No

Note: Explanation of the column headers: Benchmark—data sets used for method benchmarking; Identification—binding site detection method; Modeling—binding site representation; Data Structure—data structure used for similarity calculation; Alignment—the ability to provide binding site alignments.

^aThe GitHub repository for this method is still under preparation: <https://github.com/Jing9558/BindSiteS-CNN>.

^bThis method is not publicly available.

grids, fingerprints, 3D points).^[14] Due to the large number of tools and the different noncomparable evaluations based on differing benchmark data sets, a collection of data sets was developed to comprehensively and comparably assess binding site comparison performance in various application areas (ProSPECCTs).^[18] Thereby, it aims to facilitate choosing a suitable tool for a specific application, which is a nontrivial task given the individual limitations of the tools. Benchmarking binding site comparison tools revealed that none of the evaluated tools showed an overall superiority. In consequence, choosing the appropriate tool depends on the application.

Many recently developed tools use machine learning (ML) and deep learning methods. Three of them, DeeplyTough,^[16] BindSiteS-CNN,^[27] and DeepDrug3D,^[15] use convolutional neural networks. Site2Vec,^[24] a mathematical enhancement of PocketMatch,^[30] considers pairwise distances between representative points of binding sites, showing good overall performance for the ProSPECCTs^[14,18] data sets. However, the applicability of these methods suffers from the lack of corresponding binding site alignments, which are fundamental for evaluating the results in the context of SBDD.

PocketShape^[22] provides structure alignments by calculating residue assignments based on the Hungarian algorithm.^[31] Since the authors did not evaluate the method's performance on standard benchmark data sets, we cannot compare it to the state of the art or assess its suitability for SBDD applications. In addition, its run time is much higher than that of the best-performing tools analyzed in earlier benchmark studies^[14] (seconds vs. nano- and microseconds scale).

To find an optimal superposition of points sharing similar pharmacophoric and topological neighborhoods, ProCare^[20] uses the point cloud registration concept. It was benchmarked on a balanced version^[20] of the Vertex data set.^[17] The authors show pairwise comparison run times in the seconds time scale. Its performance with an area under the receiver operating

characteristics (ROC) curve (AUC) of 0.811 was significantly lower than that of the best-performing tool, ProBiS, with 0.896.

TWN-RENCOD uses topological water networks obtained by short molecular dynamics simulations.^[29] The aqueous environment in binding sites from these simulations is compared. The method was evaluated on a kinase data set comprising only 36 binding site pairs. Alignments were not reported.

To overcome the drawbacks of missing structure alignments, insufficiently comparable evaluations, a method choice dependent on the application field, and restricted availability, we present SiteMine, a new database- and structure-based alignment-providing binding site similarity search tool based on GeoMine.^[10,32,33] SiteMine builds on the NAOMI library with numerous methods for handling and analyzing biomolecular structures^[34] and small organic molecules.^[35,36] GeoMine is a tool for searching user-defined 3D geometric patterns enhanced with textual and numerical filters within predicted and ligand-based small molecule binding sites. A fully automated workflow calculates all data, populating a classical relational database. Every binding site atom is converted into a search point and stored with its properties (source molecule type, element, atom type, protein residue type, solvent-exposed surface area, functional group, and secondary structure type), allowing to customize the search for 3D geometric patterns. The distances between all search point pairs below 15 Å are stored to create 3D geometric search patterns.

SiteMine relies on the TetraScan approach we designed for complex 3D shape-matching applications. In short, tetrahedral search patterns are processed by GeoMine to retrieve binding site matches. In addition to SiteMine, TetraScan is also used for searching geometrically similar protein-protein interfaces (submitted for publication). All proposed matches are subsequently scored and ranked. The best-scored match is superimposed by SiteMine.

Here, we present SiteMine in detail and assess its performance on the ProSPECCTs benchmark. We applied SiteMine on cathepsin L, a promising drug target for SARS-CoV-2 inhibition,^[37] highlighting the applicability of SiteMine for drug repurposing.

2 | RESULTS AND DISCUSSION

All results on the ProSPECCTs data set will be summarized in the following. Furthermore, we evaluate SiteMine's performance on the *Balanced Vertex* data set and show the impact of using predicted pockets compared to ligand radius-defined binding sites. Finally, we showcase the applicability of SiteMine for the drug target cathepsin L.

2.1 | Run time

SiteMine's run times are determined for all-against-all comparisons (10,000) of the *Kahraman* data set.^[38] The test system had Postgres 14.3 (default settings, `max_parallel_workers` = 8), an Intel(R) Core™ i5-8500 CPU with 3.00 GHz and 16 GB RAM installed (workstation). SiteMine was used as a single-core application with a multi-threaded database search. Table 2 shows the run time of SiteMine compared with other binding site comparison tools for the same data set. SiteMine Fast has the lowest average run time per comparison within the tools that model sites as 3D points (18 ms). In contrast, the Precise setting with a run time of 122 ms is still faster than the well-performing 3D point-based methods.

2.2 | Benchmark studies

The ProSPECCTs data sets were used to benchmark the Fast and Precise settings (see Section 4.8) of SiteMine (Table 3). To ensure a fair comparison for the *ROCS Structures* data set, pairs used for optimization (*Optimization Structures* data sets) were excluded. Therefore, all AUC and enrichment factor (EF) values were recalculated for the reduced data set with scoring tables produced for earlier publications.^[14,18]

SiteMine Precise achieved the highest mean AUC with 0.835. With a difference of only 0.029 in the average AUC, SiteMine Fast is the third-best method. This difference can be attributed to performance disparity on the *Barelier*, *Decoy Structures Rational*, *Kahraman*, and *ROCS Structures* data sets. This performance loss regarding the AUC is less than 0.2 with the remaining data sets (*Decoy Structures Shape*, *Structures with Identical Sequences*, *NMR Structures*, and *Successful Applications*). Due to their high sequence similarity, these structures have strongly conserved structural binding sites. As a result, atom mappings are obtained even with a lower distance tolerance. Comparing both methods regarding the EFs (Supporting Information S1: Tables S1–S8), the trend of the observed AUC drop does not occur. A considerable difference was only found

for the *Barelier* data set for which SiteMine achieves the lowest AUC values. Compared to all other methods, it becomes clear that this is not a weakness of SiteMine but can be attributed to the design of the benchmark. The outcomes for this data set should not be over-interpreted, as the number of considered pairs is relatively small (62 pairs), in contrast to the other data sets (Table 8).

The performance differences between the SiteMine parameter sets are due to the optimization procedure. In particular, a higher filter number and more permissive tolerances increase the probability of finding atom mappings for binding site superposing and similarity score calculation. Going from Fast to Precise increases the computational cost (Table 2) but leads to significantly more structural superpositions, making it more likely to find weakly associated matches. This trend is more evident when examining the *NMR Structures* data set results (Supporting Information S1: Figure S1). Comparing SiteMine's score distributions for similar pairs, the lower whisker is 0.3 higher with the Precise setting.

Regarding the AUC (Table 3), two other tools stand out compared to SiteMine: SiteHopper^[46] and KRIPO.^[39] Although SiteEngine^[49] and SMAP^[48] show a high average AUC as well, their performance is already significantly lower, especially on data sets of similar binding pockets in unrelated proteins (*Kahraman* and *ROCS Structures* data set). Therefore, we focused on a comparison of SiteMine with SiteHopper and KRIPO.

SiteHopper is a surface-based binding site similarity tool. SiteHopper defines the binding site via a ligand-based radius. It calculates residue-based chemical properties annotated in a 3D shape calculated by two OpenEye toolkits: Shape^[52] and Spicoli.^[53] Alignments and similarities are calculated based on the physico-chemical and surface shape similarity. For the *Successful Applications* data set, the AUC values achieved with SiteMine are 0.12 and 0.14 higher in Fast and Precise mode, respectively. Also, the early enrichment for SiteHopper is inferior to that for both SiteMine settings. However, SiteHopper performs similarly to SiteMine on the *ROCS Structures* data set. Although there are no considerable differences in early enrichment (Supporting Information S1: Table S8), the AUC of SiteMine Fast is 0.05 lower. SiteMine Precise shows comparable performance to SiteHopper in terms of AUC and early enrichment. The performance similarities of both methods are remarkable, given that SiteHopper builds on the *ROCS*^[54] 3D shape- and chemical feature-based ligand comparison application used to generate the *ROCS Structures* data set. For the *Barelier* data set, SiteHopper and SiteMine show almost identical performance. Only SiteMine Precise is superior regarding AUC and early enrichment (Supporting Information S1: Table S1). For both *Decoy Structures* data sets, SiteHopper's AUC is slightly higher than SiteMine's AUC. Regarding early EF for the *Decoy Structures Rational* data set, the SiteMine methods are equally well-performing (Supporting Information S1: Table S2).

For detecting minor differences, Spearman's Rho correlation coefficients were calculated using the *Decoy Structures* data sets by ranking the scores for different numbers (1–5) of introduced binding site mutations (Supporting Information S1: Table S9). While

TABLE 2 Run times of binding site comparison methods.

Method	Data basis	Preparation run time (s) (number of structures)	Comparison run time (s) (number of comparisons)	Total run time (s)	Average pairwise run time (s)
PocketMatch ^[30]	Distance lists	28.97*	0.28	29.25	0.000028
KRIPO ^[39]	Fingerprint	446.50	0.92	447.42	0.000092
RAPMAD ^[40]	Histogram	71.42 (100)	2.36 (8,281)	73.78	0.000285
FuzCav ^[41]	Fingerprint	399.88 (96)	5.59 (9,216)	405.47	0.000607
FuzCav (PDB)	Fingerprint	236.73 (96)	5.64 (9,216)	242.37	0.000612
TM-align ^[42]	Matrix	25.72*	65.96	91.68	0.006596
SiteMine Fast	3D points	169.56 (100)	186.51	369.09	0.018651
Shaper (PDB) ^[21]	3D points (grid)	181.16 (96)	364.42 (9,216)	545.58	0.039542
Shaper	3D points (grid)	384.21 (96)	367.21 (9,216)	751.42	0.039845
VolSite/Shaper	3D points (grid)	537.00 (76)	248.77 (5,776)	785.77	0.043070
ProBiS ^[43]	Graph	6.95	479.32	486.27	0.047932
VolSite/Shaper (PDB)	3D points (grid)	259.54 (57)	162.26 (3,249)	421.80	0.049942
TIFP ^[44]	Fingerprint	228.30 (77)	550.88 (5,929)	779.18	0.092913
TIFP (PDB)	Fingerprint	194.36 (47)	205.56 (2,209)	399.92	0.093056
SiteMine Precise	3D points	169.56 (100)	1,215.30	1,400.08	0.121530
Grim (PDB) ^[45]	Graph	169.33 (96)	1,714.49 (9,216)	1,883.82	0.186034
Grim	Graph	220.17 (95)	2,104.99 (9,025)	2,325.16	0.233240
IsoMIF ^[11]	Graph	752.83	2,561.44	3,314.27	0.256144
SiteHopper ^[46]	3D points	154.01	3,828.61	3,982.62	0.382861
Cavbase ^[47]	Graph	67.89 (100)	21,823.71 (8,281)	21,891.60	2.635396
SMAp ^[48]	Graph	1.69	42,346.74	42,348.43	4.234674
SiteEngine ^[49]	3D points	328.81	81,193.54	81,522.35	8.119354
SiteAlign ^[50]	Fingerprint	28.97*	286,326.41	286,355.38	28.632641

Note: The star (*) denotes exemplary run times for separate preprocessing steps. The SiteMine rows are highlighted in light gray. The table and run times of the other methods are extracted from previous benchmark studies.^[14] Note that computing times for SiteMine were recorded on different hardware.

SiteHopper's Combo score correlates better with the number of residue substitutions by similarly sized physicochemically diverse residues, SiteMine's score shows a better correlation with the number of residues substituted by differently sized residues. In this context, SiteMine is also one of the most sensitive tools regarding minor differences in the binding site. Regarding the *Structures with Identical Sequences* and *NMR Structures* data set, the early enrichment differences are negligible (Supporting Information S1: Tables S4 and S6). On the *Kahraman* data set, the SiteMine settings perform better than SiteHopper (Supporting Information S1: Table S5). In summary, we can show that SiteMine's performance is comparable and, for some application domains, even superior to SiteHopper.

KRIPO^[39] defines binding sites via a ligand atom radius of 6 Å. An interaction fingerprint represents the binding site. It encodes residue interaction features and binned residue distances. A modified Tanimoto coefficient^[55] is the similarity measure.

Concerning the *Kahraman* data set, the performance regarding the AUC of KRIPO is similar to both SiteMine settings. The early enrichment is slightly lower for SiteMine (Supporting Information S1: Table S5). For the *Decoy Structures*, *NMR Structures*, *Structures with Identical Sequences*, and the *Successful Applications* data set, SiteMine's performance is superior regarding AUC and EFs (Supporting Information S1: Tables S2–S4, S6, S7). For the *ROCS Structures* data set, the AUCs of SiteMine Fast and KRIPO are similar. A performance difference regarding the EFs is apparent: in contrast to SiteMine, the enrichment of similar pairs by KRIPO decreases with increasing percentages of screened data (Supporting Information S1: Table S8). KRIPO's fingerprint-based approach is faster than SiteMine, but binding site superpositions are not computed on the fly. Instead, they are calculated using a clique algorithm. Consequently, KRIPO harbors the disadvantage that the superposition does not necessarily correspond to the fingerprint-based similarity.

TABLE 3 Overview of the SiteMine results and the tools of the benchmark studies.^{14,18}

Method	Mean	Barelier ^[51]	Decoy Structures Rational	Decoy Structures Shape	Structures with Identical Sequences	Kahraman ^[38]	NMR Structures	Successful Applications	ROCS Structures
SiteMine Precise	0.835	0.61	0.69	0.72	1.00	0.78	1.00	0.91	0.97
SiteHopper	0.813	0.56	0.75	0.75	0.98	0.72	1.00	0.77	0.97
SiteMine Fast	0.806	0.56	0.65	0.71	1.00	0.74	0.98	0.89	0.92
KRIPO	0.794	0.73	0.60	0.61	0.91	0.76	0.96	0.85	0.93
SiteEngine	0.771	0.55	0.82	0.79	0.96	0.64	1.00	0.86	0.55
SMAP	0.766	0.68	0.76	0.65	1.00	0.62	1.00	0.86	0.56
SiteAlign	0.759	0.44	0.85	0.80	0.97	0.59	1.00	0.87	0.55
Shaper (PDB)	0.749	0.54	0.71	0.76	0.96	0.66	0.93	0.75	0.68
Shaper	0.746	0.54	0.71	0.76	0.96	0.65	0.93	0.75	0.67
TM-align	0.738	0.59	0.49	0.49	1.00	0.66	1.00	0.88	0.79
VolSite/Shaper	0.734	0.71	0.68	0.76	0.93	0.56	0.78	0.77	0.68
IsoMIF	0.733	0.62	0.59	0.59	0.77	0.75	0.70	0.87	0.97
FuzCav	0.720	0.67	0.69	0.58	0.94	0.55	0.99	0.77	0.57
FuzCav (PDB)	0.718	0.65	0.69	0.58	0.94	0.56	0.98	0.77	0.57
PocketMatch	0.714	0.51	0.59	0.57	0.82	0.66	0.96	0.82	0.78
Cavbase	0.711	0.55	0.65	0.64	0.98	0.60	0.87	0.82	0.58
VolSite/Shaper (PDB)	0.698	0.50	0.68	0.76	0.94	0.57	0.76	0.72	0.65
ProBiS	0.686	0.50	0.47	0.46	1.00	0.54	1.00	0.85	0.67
TIFP	0.680	0.55	0.66	0.66	0.66	0.71	0.91	0.71	0.58
Grim	0.665	0.45	0.55	0.56	0.69	0.69	0.92	0.70	0.76
RAPMAD	0.649	0.60	0.61	0.63	0.85	0.55	0.82	0.74	0.39
Grim (PDB)	0.633	0.45	0.57	0.56	0.62	0.61	0.85	0.64	0.76
TIFP (PDB)	0.598	0.56	0.56	0.57	0.55	0.54	0.78	0.66	0.56

Note: For each tool and data set, the area under curve (AUC) is given. The table is sorted according to the mean AUC for all data sets. The SiteMine methods are highlighted in light gray. The *Optimization Structures* data set pairs are excluded from the *ROCS Structures* data set for benchmarking all methods on this data set.

SiteMine is also applicable to compare predicted binding sites, representing a considerable advantage over SiteHopper and KRIPO. The *ROCS Structures* data set corresponds to common use cases, finding similar binding sites in unrelated proteins (e.g., off-target prediction). Also, IsoMIF^[11] demonstrates remarkable performance but is slower than SiteMine. Considering the total of all data sets (AUC, EF, Spearman's Rho), SiteMine shows promising performance (Table 4).

2.3 | Comparison to ML tools evaluated on ProSPECCTs

Since their publication, the ProSPECCTs data sets have been used to evaluate three ML-based methods: DeeplyTough, Site2Vec, and

BindSiteS-CNN. Therefore, we can readily compare them to SiteMine (Table 5). Although the methods were not evaluated on the *ROCS Structures* data set, the second closest to a real-world application scenario, we can assess their general applicability to realistic use cases employing the *Successful Applications* data set.

Site2Vec has the highest mean AUC (0.87), while BindSiteS-CNN and DeeplyTough rank below SiteMine. The AUC values for the *Decoy Structures* data sets are lower for SiteMine. In contrast, SiteMine's AUC values are significantly higher for the *Successful Applications* data set than for the three ML-based methods. As this data set represents the most meaningful data set regarding SBDD studies, the poor performance of Site2Vec for this data set questions its applicability in SBDD.

Moreover, recent tools do not provide binding site alignments for proper visual inspections, restricting their usefulness for structure-based

TABLE 4 Criteria of importance for choosing a suitable binding site comparison method.

Method	Preparation (ease)	Preparation (completeness)	Application to predicted sites	Run time ^a	Definition ^b	Definition (ranking) ^b	Flexibility ^c	Properties (ranking) ^d	ROCS structures	Successful applications	Visualization
SiteMine	+	+	+	/	+	+	+	+	+	+	+
Cavbase	+	–	+	–	+	+	+	+	–	+	+
FuzCav	/	+	+	+	+	/	+	+	–	+	–
Grim	/	–	–	/	–	–	+	–	/	–	+
IsoMIF	+	+	+	/	–	–	–	–	+	+	+
KRIPO	+	+	–	+	–	/	+	+	/	+	+
PocketMatch	–	–	(+)	+	–	/	+	–	+	+	–
ProBiS	+	+	(+)	+	+	+	+	–	–	+	+
RAPMAD	+	–	+	+	–	–	–	+	–	–	–
VolSite/Shaper	/	–	+	/	+	/	–	+	–	+	+
SiteAlign	–	+	(+)	–	+	+	+	+	(+)	+	+
SiteEngine	+	+	–	–	+	/	+	+	–	+	+
SiteHopper	+	/	(+)	/	+	+	+	+	+	+	+
SMAP	+	+	(+)	–	+	+	+	+	–	+	+
TIFP	/	–	–	/	–	–	+	–	–	–	–
TM-align	–	+	(+)	/	+	+	+	n/a	+	+	+

Note: Besides its intermediate run time, SiteMine is superior to other tools investigated previously.^[14,18] With respect to run time evaluation, “+,” “/,” “–” denote comparison algorithms that require several ns, μs, or s per comparison, respectively. With respect to the scoring, a “+” was assigned to tools if the intervals of upper and lower whiskers of active and inactive pairs do not overlap. A “/” was assigned to tools whose upper and lower quartiles for the pairs do not overlap. With respect to other factors, tools that were clearly outperformed by many other tools were assigned a “–.” The table was adapted based on earlier benchmark studies.^[56]

^aKahraman data set.
^bStructures with Identical Sequences data set.
^cNMR Structures data set.
^dDecoy Structures Rational & Shape data set.

TABLE 5 Performance comparison of SiteMine with ML-based methods for the ProSPECCTs data sets.

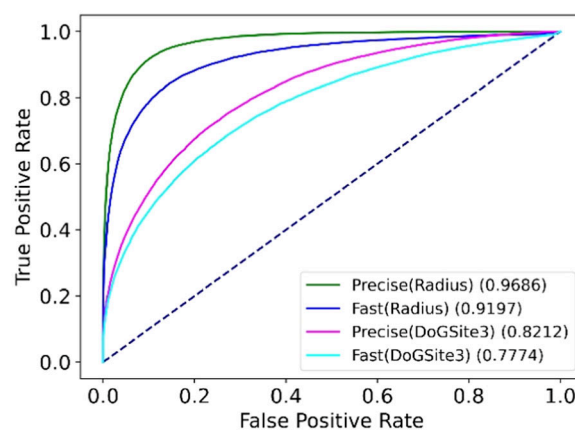
Method	Structures with Identical Sequences	NMR Structures	Decoy Structures Rational	Decoy Structures Shape	Kahraman	Barelier	Successful Applications	Mean
SiteMine Precise	1.00	1.00	0.69	0.72	0.78	0.61	0.91	0.82
SiteMine Fast	1.00	0.98	0.65	0.71	0.74	0.56	0.89	0.79
Site2Vec	1.00	1.00	0.99	0.99	0.86	0.53	0.66	0.87
BindSiteS-CNN	0.94	0.83	0.91	0.79	0.66	0.62	0.78	0.79
DeeplyTough	0.95	0.90	0.76	0.75	0.63	0.54	0.83	0.77

Note: The table shows the area under curve (AUC) values for the individual data sets. The results of the other methods were extracted from previous studies.^[16,24,27]

studies. The number of applications that solely rely on similarity scores is considerably low,^[12] highlighting that superpositions are indispensable for evaluating the results to assess their impact on SBDD projects. As shown in Table 5, the performances of ML-based comparison methods are inferior for data sets of active site pairs without sequential relationships (*Kahraman*, *Barelier*, *Successful Applications*). While detecting sequentially related binding sites can be regarded as solved, further developments of ML-based methods should focus on a good performance on structurally similar binding sites with low sequence similarity. Future efforts should also focus on reporting reliable binding site alignments.

2.4 | Evaluation of the impact of ligand radius-defined and predicted binding sites

Binding sites defined by ligands introduce a bias since the binding site description includes the ligand's exact location and size. Therefore, we evaluated the impact of predicted binding sites on SiteMine's performance on the ProSPECCTs benchmark data sets. To ensure the correctness of the compared binding site, DoGSite3's^[57] new mode for detecting difficult ligand-occupied pockets was used. Here, ligand fragments are only used to bias the binding site grid if the ratio of the bound solvent-accessible surface area and the unbound solvent-accessible surface area is below 0.35 (maxSASLigandRatio). Using this feature, we ensure that we use the ligand-occupied pocket without biasing the binding site dimensions based on the ligand alone. The Supporting Information includes all AUC values and EFs for SiteMine on all ProSPECCTs data sets (Supporting Information S1: Table S10). SiteMine with predicted pockets performs slightly worse in terms of AUC for the *Successful Applications* and *NMR Structures* data sets. For both *Decoy Structures* data sets, a similar performance is observed. In contrast, the mean Spearman's Rho correlation coefficients decrease for the *Decoy Structures* data sets (Supporting Information S1: Tables S11 and S12) as mutations might lead to different binding site dimensions. SiteMine performs similarly with both types of binding site definition for the *Structures with Identical Sequences* data set.

**FIGURE 1** Receiver operating characteristics (ROC) curves for the ROCS Structures data set for SiteMine settings with both types of binding site definitions: ligand radius-based (Radius) and DoGSite3-defined (DoGSite3).

Using predicted pockets, SiteMine performs significantly poorer for the *Kahraman* and ROCS Structures data sets (see Figure 1). Both data sets have a substantial similarity: the similar binding site pairs contain similar or identical ligands. The similarity classification for the other data sets except *Barelier* relies on the protein instead of the ligand. Due to the ligand radius-based binding site definition, per se, the comparison is biased, focusing on the probably most similar parts of the sites: Similar or identical ligands in the ROCS Structures or the *Kahraman* data set show some similarity when evaluating size alone. The pockets are considerably larger when not selecting site residues based on the ligand. The increased binding site size leads to poor scoring performance, as the score is normalized by the larger binding site in terms of the number of solvent-exposed heavy atoms. In addition, SiteMine was optimized on a set of ligand-derived sites with similarly sized ligands.

However, one of the most common applications of binding site comparison is screening a database of ligand-based pockets against a predicted site to find potential ligands or off-targets, necessitating a high early enrichment rather than a promising overall performance. Given the comparison of the EFs for ligand radius-based and predicted pockets, SiteMine performs equally well in both scenarios.

2.5 | The Balanced Vertex data set

ProCare,^[20] which is not ML-based, was not evaluated on the ProSPECCTs^[14,18] data sets. Instead, the authors modified the Vertex^[17] data set to create a balanced version with 676 pairs (338 similar and dissimilar). The Vertex data set, as the ROCS Structures data set, was developed based on the hypothesis that similar ligands bind to similar sites. Therefore, the similar site pairs were derived from their ligand-based similarity. Both data sets differ in the considered ligands for data set generation. In contrast to the ROCS Structures data set, the Vertex data set considers binding affinities. Due to the ambiguity of some binding site ligands for the Balanced Vertex^[20] data set, we have revised the ligand identifiers of the binding site pairs (see Supporting Information S1: Table S13). Note that it is undocumented how the ambiguity was resolved by the other tools. ProCare compares VolSite-predicted^[21] binding sites. For the Balanced Vertex^[20] data set, SiteMine was benchmarked using DoGSite3-defined (maxSASLigandRatio = 0.35) and ligand radius-defined binding sites to allow a fair comparison (Table 6). Here, SiteMine with the Precise setting performs better than the Fast setting, as observed for the ProSPECCTs data sets regarding the AUC. Also, ligand radius-defined binding sites result in a higher AUC than DoGSite3-defined pockets. SiteMine performs better than ProCare except for the Fast setting with predicted pockets. ProCare's average pairwise run time of 2 s is several orders of magnitude slower than SiteMine's (see Table 2).

Looking at the performance of the analyzed tools for this data set, it meets the eye that methods performing only mediocly in previous studies perform best on this data set (PocketMatch and

TABLE 6 Performance comparison for the *Balanced Vertex* data set.²⁰

Method	AUC	Completeness (%)
SiteMine Precise (Radius)	0.906	98.5
SiteMine Precise (DoGSite3)	0.898	98.1
ProBiS	0.896	64.2
PocketMatch	0.895	99.4
SiteMine Fast (Radius)	0.874	98.5
KRIPO	0.862	95.2
SiteAlign	0.859	100.0
SiteMine Fast (DoGSite3)	0.846	98.1
FuzCav	0.831	100.0
ProCare	0.811	99.7
Shaper	0.774	99.7

Note: The results of the other methods were extracted from previous studies.^[20] For SiteMine, the revised version was used (Supporting Information S1: Table S13).

Abbreviation: AUC, area under the receiver operating characteristics curve.

ProBiS), while more reliable tools (KRIPO and SiteAlign) show a poorer performance. This finding can be partially attributed to the high overall similarity of the pairs classified as similar, leading to a good performance of approaches relying solely on the pockets' residues sequence in a 7 Å radius and using sequence identity as the scoring measure (AUC of 0.9).^[17] Furthermore, the Vertex data set includes pairs of structurally and functionally related protein pairs (e.g., protein kinases or phosphodiesterase enzymes) in both the active and inactive pairs, which might be caused by their selection relying solely on data available in the ChEMBL database.^[58] In summary, we can conclude that the *Balanced Vertex* data set cannot reflect realistic scenarios for which elaborate binding site comparison tools are indispensable.

2.6 | Cathepsin L—searching for similar sites in the PDB

Cathepsins belong to the peptidase C1 family and play a role in the hydrolytic degradation of the extracellular matrix.^[59] They also participate in apoptosis and antigen processing, as well as lysosomal recycling of cellular proteins. Cathepsin L, in particular, plays a pivotal role in the infection of human coronaviruses such as SARS-CoV and SARS-CoV-2 by facilitating their entry into the cell through proteolysis of the spike protein.^[60] Inhibition of this protease can thus prevent infection, making it a target of interest for SBDD.^[37,61]

A sequence-culled PDB subset was created with PISCES^[62] (see Supporting Information S1: Table S14 for parameters) to speed up the search. For this subset of 40,207 PDB entries, a database was built using ligand radius-defined binding sites. For searching, SiteMine's Fast setting was used. The query binding site was defined via the bound inhibitor (radius of 6.5 Å, ligand identifier: 424) of PDB entry 2xu1. Searching this database with 63,106 binding sites took 230 s single-threaded on a single desktop computer, corresponding to an average run time per comparison of about 3 ms.

We inspected the 30 top-scored (normalized by the larger binding site, Supporting Information S1: Table S15) binding site superpositions and mostly found papain-like proteases of a similar fold. Searching for nonobvious similarities in differently folded structures, we also inspected the top 30 superpositions of non-normalized scores (raw scores, Supporting Information S1: Table S16). On rank 27, we found the active site of human calpain 1 (PDB entry 1zcm). Interestingly, the rank within the normalized scores is much higher at 573 but still within the top 85% (Supporting Information S1: Figure S2).

The sequence identity is low (14.5%, EMBOSS Needle^[63]), but both proteins belong to the same family of cysteine proteases. This similarity becomes more evident when examining the binding site superposition (Figure 2). It shows eight superimposed identical residues with similar side chain orientations. These residues are near the catalytic center of the reactive cysteines (cathepsin L—Cys25, calpain 1—Cys115).

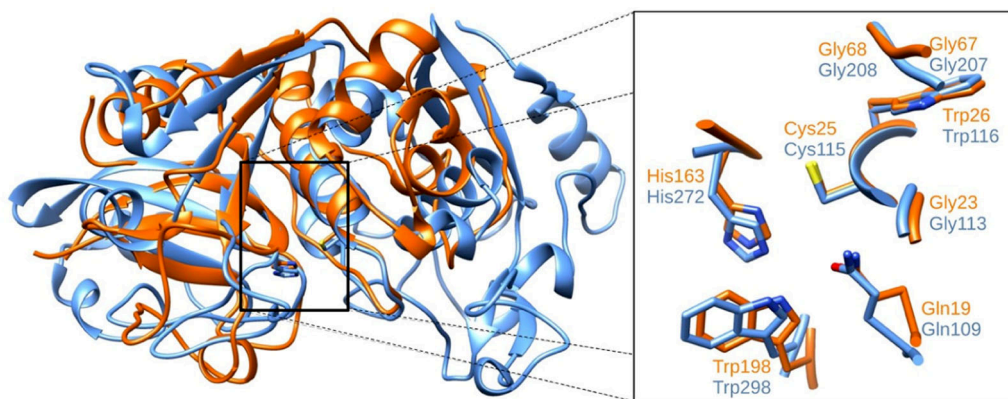


FIGURE 2 The SiteMine binding site superposition of human cathepsin L (orange, PDB entry 2xu1) and human calpain protease (blue, PDB entry 1zcm). Identical residues of chain A of the binding sites are shown. The image was created with UCSF Chimera.^[64]

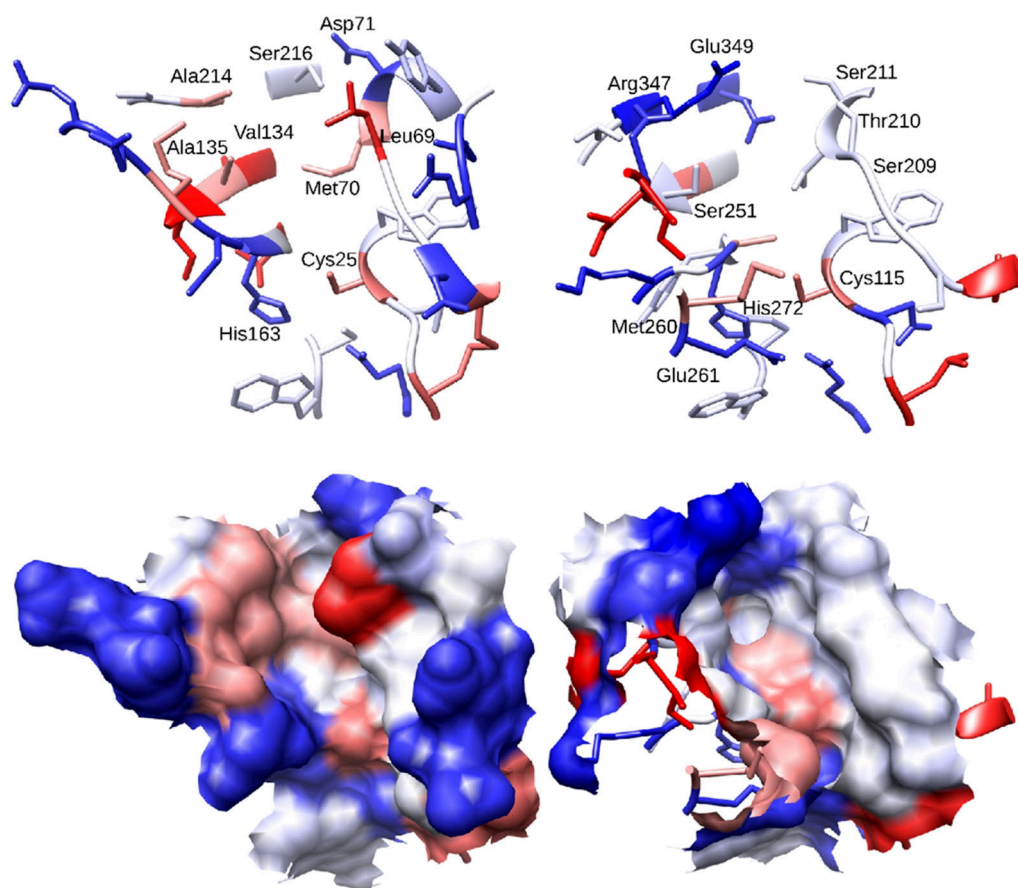


FIGURE 3 SiteMine binding site alignment of cathepsin L (left, PDB entry 2xu1) and human calpain protease (right, PDB entry 1zcm). Top: residue arrangement. Bottom: the surface representation. The residues are color-coded according to the hydrophobicity scale of Kyte and Doolittle^[65] in UCSF Chimera^[64] and UCSF ChimeraX^[66] (low hydrophobicity—blue, high hydrophobicity—red). The catalytic residues (His163 and Cys25, His272 and Cys115) in their hydrophobicity scale and further site residues are labeled.

Also, the pockets differ in several aspects (Figure 3). Regarding the binding site shape, in calpain 1, residues Met260 and Glu261 narrow the pocket moderately and cause a slight closure in the front part of the pocket (S1 pocket^[67]). Furthermore, the

properties of some residues in the posterior (S2, S3 pocket^[67]) part of the binding site differ. While cathepsin L is predominantly lipophilic (lipophilic: Ala135, Ala214, Leu69, Met70; hydrophilic: Ser216, and Asp71), calpain 1 has predominantly hydrophilic

residues (hydrophilic: Ser251, Arg347, Glu349, Ser209, Thr210, and Ser211).

The identified binding site similarity assists in SBDD. On the one hand, the two binding sites have a high global similarity, meaning that inhibitors binding to calpain 1 could also bind to cathepsin L (drug repurposing, off-target prediction). On the other hand, selectivity is achievable by exploring the identified differences. One possibility would be to superpose available structures with bound inhibitors and derive the core and specificity-mediating fragments to design new, potentially more specific binders.

3 | CONCLUSION

Searching for similar protein binding sites can support several SBDD challenges, such as drug repurposing, analyzing protein–ligand and protein–protein complexes, and off-target or function prediction. According to the review by Eguida and Rognan,^[13] almost 40 software tools have been developed in the past 20 years. However, only a few were evaluated based on unique benchmark sets^[14,18] to determine strengths and weaknesses and, thus, their application domains. None of the comprehensively benchmarked tools showed a promising

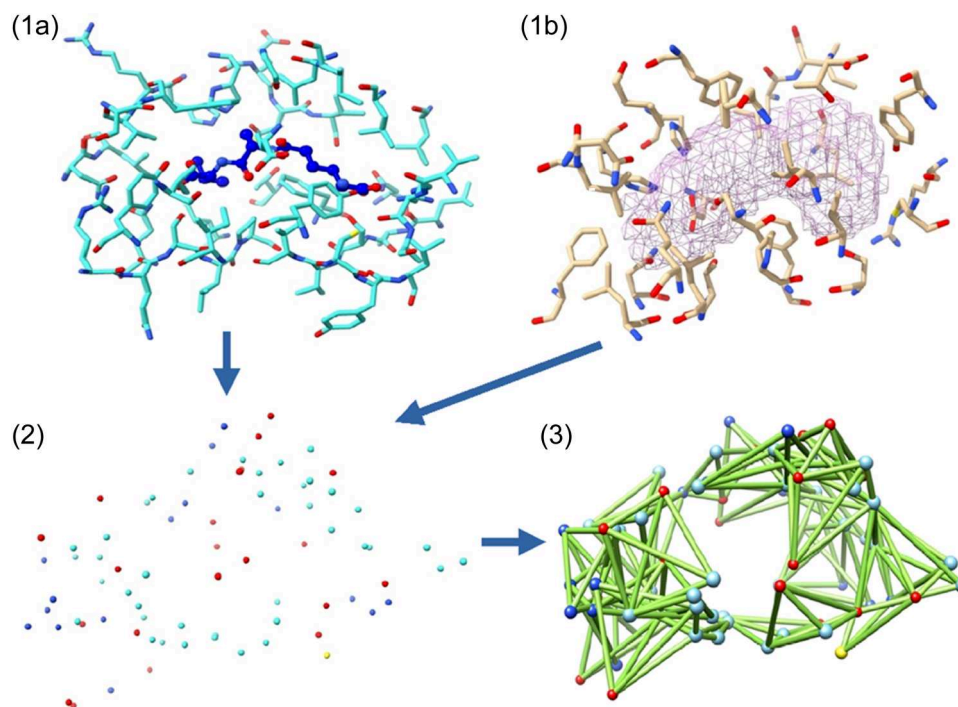


FIGURE 4 Binding site modeling with SiteMine. SiteMine supports using (1a) ligand radius-defined or (1b) predicted binding sites. The solvent-accessible atom selection results in an atom subset (2). Tetrahedra are built and selected to represent the query-binding site (3). Element-specific atom coloring: cyan/beige—carbon, red—oxygen, blue—nitrogen, yellow—sulfur. The image was created with UCSF ChimeraX.^[66]

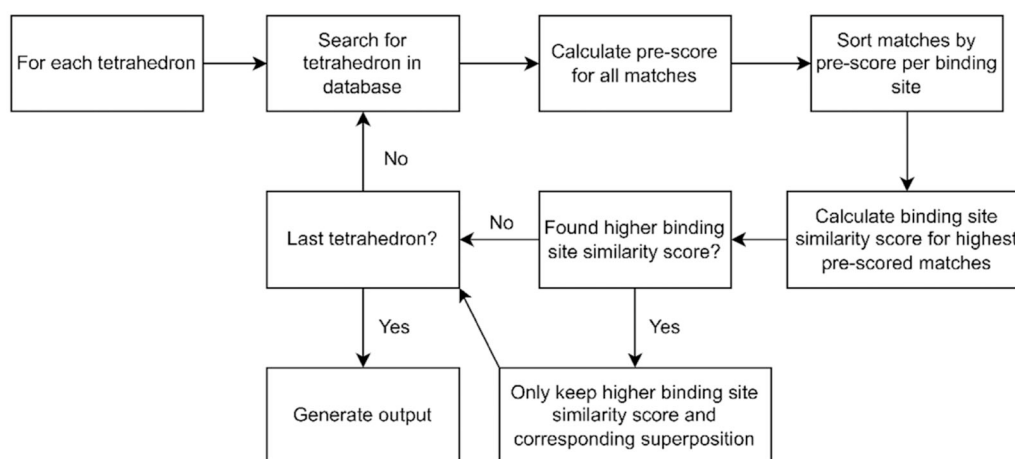


FIGURE 5 Binding site comparison with SiteMine.

performance regarding critical criteria for reliable SBDD, that is, good performance in terms of AUC and early enrichment for all data sets, the possibility of comparing predicted binding sites, and reasonable run time to screen extensive collections of protein binding sites.

In this work, we introduced SiteMine, a new database-driven binding site comparison method providing similarity scores and the corresponding alignments. We evaluated it using the ProSPECTs benchmark data sets, also comparing it to published tools. SiteMine is one of the best-performing tools for all data sets, demonstrating its broad applicability. In the run time comparison, SiteMine is slower than fingerprint-based methods but among the fastest for tools with comparable features regarding binding site modeling and the possibility of providing alignments. SiteMine is available for Linux, macOS, and Windows as part of the NAOMI ChemBio Suite (<https://uhh.de/naomi>) and is free for academic use and evaluation purposes. To enable screenings for similarity searches in huge databases, we established a second parameter set (Fast) in addition to the Precise setting. Therefore, we recommend the Fast setting for a large-scale similarity screening with subsequent Precise setting runs to improve scores and superposition on a promising selection.

Potential binders of novel detected binding sites can be predicted by screening for similar ligand-bound pockets, representing a frequent use case of automated binding site comparisons. These comparisons are particularly useful for binding sites in proteins with a low overall structural similarity to already known structures. We showed that our method performs reliably using both predicted and ligand-defined binding sites. We also realized a significant performance loss upon comparing differently sized binding sites, indicating the importance of adjusting the score normalization in screenings for similar sites with one query.

Given the rising number of developed binding site comparison tools, the scientific community might further benefit from even better-performing methods. However, it renders choosing a suitable tool infeasible without commonly used benchmark sets and unique evaluation pipelines. Current ML-based approaches gain attention, but their applicability suffers from the lack of binding site alignments. With SiteMine, we present a novel tool to the SBDD community that is easy to use, applicable to predicted sites, and shows promising performance regarding the most crucial quality criteria.

Despite SiteMine's comparably good run time within alignment-providing methods, a similarity search within the AlphaFold database with over 200 million structures seems challenging. This task becomes even more complex when using structure ensembles to consider protein flexibility.

SiteMine can be successfully applied for selectivity analyses and the discovery of novel targets for known drugs. In an application showcase, we used SiteMine to search for similar binding sites for cathepsin L. We found a high similarity to the active site of calpain 1. Thus, some inhibitors derived from calpain 1 might also bind cathepsin L, opening a potential avenue for drug repurposing. Similarly, calpain 1 should be considered a potential off-target when profiling cathepsin L binders for selectivity.

In summary, we hope the scientific community will benefit from using SiteMine in various SBDD projects and find the depicted similarity for cathepsin L and calpain 1 inspiring for searching for new inhibitors.

4 | EXPERIMENTAL

In the following, we describe and visualize the SiteMine comparison algorithm and outline tailor-made benchmarking data sets and the performance assessment (see also Figures 4 and 5). Subsequently, we provide details regarding the parameter optimization of SiteMine.

4.1 | Binding site definition

Predefined binding sites can easily be fetched from the GeoMine database. SiteMine also parses DoGSite3-predicted^[57] binding sites or site atoms in a 6.5 Å radius of the ligand's heavy atoms. Alternatively, a custom binding site can be specified using residue IDs. The identification of interactions, protonation and tautomeric states, and hydrogen orientations generated by Protoss is described elsewhere.^[68]

4.2 | Selecting search atoms

For 3D geometrical query generation, all solvent-accessible heavy atoms of all site residues, each aromatic ring center (His, Phe, Trp, and Tyr), and all side chain carbon atoms of hydrophobic residues (Ala, Ile, Leu, Lys, Met, Pro, and Val) are selected.

4.3 | Building and selecting tetrahedra

A series of search tetrahedra is constructed using the selected atoms as corners (Figure 4). As searching for all possible tetrahedra (counting to N^4 , where N is the number of selected atoms) is prohibitive, we introduced an algorithm for tetrahedra selection. The algorithm aims for an equal distribution of tetrahedra across the binding site according to atom usage.

In the first step (Algorithm 1), tetrahedra fulfilling distance and properties constraints are created (L.2). Distances between search atoms corresponding to tetrahedron edge lengths have to be between 1 and 8 Å. The property constraints are the number of atom types representing the tetrahedron corners (see Section 4.8 for details). The user can adapt the distance values for specific application scenarios. Property constraints can be turned on or off.

All resulting tetrahedra are sorted in descending order by the sum of their edge lengths (L.3) to ensure that large tetrahedra with most atoms of minimum atom usage are preferred (L.9). Initially and during tetrahedra selection, the occurrence of every site atom as tetrahedron corner in selected tetrahedra is counted (atom usage count, L.4). The algorithm ensures that we always select tetrahedra with the maximum number of atoms as corners with minimal occurrence so far (L.9). Adding a tetrahedron to the selection list implies its deletion in the list of all tetrahedra and the update of the atom usage counts (L.8–12). The process terminates once the selected number of tetrahedra exceeds a user-defined

Algorithm 1 Procedure of the tetrahedra selection algorithm

```

1: procedure SELECTTETRAHEDRA(atoms, max, nofTetrahedra)
2:   allTetrahedra = createAllPossibleTetrahedra(atoms, max)
3:   sortTetrahedra(allTetrahedra)
4:   atomTetrahedraUsage = {atom1 = 0, atom2 = 0,..., atomn = 0}
5:   selectedTetrahedra = ∅
6:   while | selectedTetrahedra | < nofTetrahedra and min(atomTetrahedraUsage) == 0 do
7:     for i = 4 down to 1 do
8:       for each tetrahedron ∈ allTetrahedra do
9:         if tetrahedron has i atoms with min(atomTetrahedraUsage) then
10:           selectedTetrahedra.add(tetrahedron)
11:           allTetrahedra.delete(tetrahedron)
12:           increase(atomTetrahedraUsage, tetrahedron)
13:         end if
14:       end for
15:     end for
16:   end while
17:   return selectedTetrahedra
18: end procedure

```

count (default: 30) if each atom already occurs in at least one tetrahedron (L.6), ensuring a complete binding site representation.

4.4 | Filter building

The selected tetrahedra represent the 3D geometrical queries (called filters in the following) to search within the GeoMine database for atom mappings in binding sites. The atoms constitute search points annotated by their coordinates, atom types (acceptor, donor, acceptor & donor, aromatic, hydrophobic, anion, or cation), and solvent accessibility. The coordinates are only used to calculate superpositions in the match-processing step. Tetrahedra edges are translated to distance ranges (search point distances including relative tolerances, default: 20%).

4.5 | Match processing

Found filter matches in the target binding sites result in atom mappings. Binding site superpositions are calculated by the C++ Eigen library^[69] implementation of the Umeyama algorithm.^[70] A prescore is computed by counting the query Cα atom occupancy. An atom within a 6 Å radius (rounded average amino acid diameter 10.6 Å^[71]) of a target Cα atom is considered occupied. This prescore serves as superposition quality estimation. The highest prescored \sqrt{N} superpositions per binding site are chosen, where N is the number of query hits found for the binding site. This heuristic limitation does not influence the quality of the result while simultaneously reducing compute resources otherwise spent for more expensive similarity score calculations.

4.6 | Binding site similarity scoring

For each target binding site superimposed on the query, a so-called SP-score consisting of a shape and a pharmacophore component is calculated. For each solvent-exposed atom (solvent-accessible surface > 0 Å²) in the query-binding site, neighboring solvent-exposed target atoms in a predefined radius of 1.5 Å are searched. If at least one atom is found, an atom pair is formed, and the shape score is increased by one. If more than one atom is found, the closest one is chosen to build the atom pair. The atom pair's similarity is evaluated by a pharmacophore-based scoring matrix (pharmacophore score, Table 7) and added up to the pharmacophore score. The shape and pharmacophore scores are equally weighted, summed up, and normalized to form the binding site similarity SP-score. For normalization, the score is divided by the maximum number of solvent-exposed atoms of the two compared binding sites. Among all possible superpositions, the one maximizing the SP-score is finally selected. The complete comparison process with SiteMine is summarized in Figure 5.

4.7 | Benchmark data sets

The ProSPECCTs^[14,18] data sets and the *Balanced Vertex*^[20] data set are used for method evaluation (Table 8). ProSPECCTs aims to reveal the strengths and weaknesses of binding site similarity search tools.

We compared the performance of SiteMine and other binding site comparison methods evaluated in earlier benchmark studies.^[14,18] Hence, the same evaluation metrics, that is, the AUC and the EF, are applied.

TABLE 7 Scoring scheme for an atom pair according to its pharmacophore similarity.

	Acc/Don	Acc	Don	Aro	HyPhob	Ca	Pos/Don	Neg/Acc
Acc/Don	1	0.6	0.6	0	0	0	0.6	0.6
Acc		1	0	0	0	0	0	0.8
Don			1	0	0	0	0.8	0
Aro				1	0.8	0	0	0
HyPhob					1	0	0	0
Ca						1	0	0
Pos/Don							1	0
Neg/Acc								1

Abbreviations: Acc/Don, hydrogen bond acceptor and donor; Acc, hydrogen bond acceptor; Don, hydrogen bond donor; Aro, atom is part of an aromatic system; HyPhob, hydrophobic atom; Ca, alpha carbon atom; Pos/Don, positively charged hydrogen bond donor; Neg/Acc, negatively charged hydrogen bond acceptor.

TABLE 8 Brief overview of the used data sets.

Data set name	Number of similar pairs	Number of dissimilar pairs	Evaluation purpose
Structures with Identical Sequences ^[14,18]	13,430	92,846	Influence of the binding site definition
NMR Structures ^[14,18]	7,729	100,512	Influence of the binding site flexibility
Decoy Structures Rational ^[14,18]	13,430	13,430	Differentiation of minor physicochemical changes
Decoy Structures Shape ^[14,18]	13,430	13,430	Differentiation of minor shape changes
Barelier ^[51]	19	43	Identification of unrelated binding site pairs with identical ligands in similar environments
Kahraman ^[38]	1,320	8,680	Recovery of sites with identical ligands and cofactors
Successful Applications ^[14,18]	115	56,284	Recovery of known similar binding site pairs
ROCS Structures ^[18]	15,339	56,179	Recovery of similar sites with similar ligands in similar conformations in sequentially unrelated site pairs
Optimization Structures	150	450	SiteMine's parameter optimization (subset of ROCS Structures)
Balanced Vertex ^[20] data set	338	338	Recovery of similar sites with ligands with similar binding affinities

4.8 | Parameter optimization

For filter and parameter optimization of SiteMine, a subset of the ROCS Structures data set^[18] was created, named the *Optimization Structures* data set (Supporting Information S1: Table S17).

All ligands of the similar binding site pairs of the ROCS Structures data set were extracted as SD files and loaded in KNIME 4.3.3.^[72] Their ECFP4 fingerprints were calculated using the CDK Fingerprints node. Next, a Tanimoto coefficient-based distance matrix was calculated for a k-Medoids clustering with a partition count of 150. This procedure was also applied to the ligands of the dissimilar binding site pairs using a partition count of 450. To compile a data set of 150 “active” and 450 “inactive” site pairs for parameter optimization, we extracted all respective pairs per clustered ligand (this ligand had to be in at least one binding site).

Finally, we randomly selected one pair not already chosen to represent a previously chosen ligand.

The search time of SiteMine is mainly influenced by the distances (tetrahedra edge lengths), their tolerance, the number of filters, and the search point properties (atom types).

To investigate the run time behavior of the filters composed of different point properties, we created filters with all possible property combinations and uniform edge lengths (4.5 Å with a tolerance of 3.5 Å representing a distance range between 1 and 8 Å). We found that filters became faster with increasing numbers of aromatic, anion, and cation points. The opposite was observed with increasing numbers of acceptor, donor, acceptor and donor, and hydrophobic points. The number of matches is inversely proportional to the run time (see Supporting Information S1: Table S18 for details). To find a compromise between optimum run time and performance, we derived the following rules:

TABLE 9 Final parameter combinations resulting from the optimization.

Name	Minimum number of filters	Distance tolerance (%)	Maximum edge length (Å)
SiteMine Fast	30	20	8
SiteMine Precise	40	25	9

Filters must include at least one aromatic, anion, or cation point and two hydrophobic points at maximum. The latter rule limits the maximum number of hydrophobic points since these considerably contribute to the run time costs compared to acceptor, donor, and acceptor and donor points.

Using these rules, the remaining three parameters were optimized in a brute-force approach (Supporting Information S1: Table S19). The results of this parameter optimization can be found in Supporting Information S1: Table S20. We selected two parameter combinations based on the AUC, EFs, and run time: Fast and Precise (Table 9).

ACKNOWLEDGMENTS

The authors thank the whole development team of the NAOMI library for forming the basis of this work, as well as the members of our research group, Computational Molecular Design for code reviewing. This work was supported by DASHH (Data Science in Hamburg - Helmholtz Graduate School for the Structure of Matter) with the Grant-No. HIDSS-0002. Christiane Ehrt and Thorben Reim are funded by Data Science in Hamburg - Helmholtz Graduate School for the Structure of Matter (Grant-No. HIDSS-0002). Sebastian Günther and Alke Meents acknowledge financial support obtained from the Helmholtz Association through the projects FISCOV, SFragX and the Helmholtz Association Impulse and Networking funds InternLabs-0011 "HIR3X". Open Access funding enabled and organized by Projekt DEAL.

CONFLICTS OF INTEREST STATEMENT

ProteinsPlus and the NAOMI ChemBioSuite use some methods jointly owned by and/or licensed to BioSolveIT GmbH, Germany. Matthias Rarey is a shareholder of BioSolveIT GmbH. The other authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that supports the findings of this study are available in the Supporting Information of this article.

ORCID

Thorben Reim  <http://orcid.org/0009-0002-7712-8515>

Christiane Ehrt  <http://orcid.org/0000-0003-1428-0042>

Joel Graef  <http://orcid.org/0000-0001-8327-4936>

Sebastian Günther  <https://orcid.org/0000-0002-7329-6653>

Alke Meents  <https://orcid.org/0000-0001-6078-4095>

Matthias Rarey  <http://orcid.org/0000-0002-9553-6531>

REFERENCES

- [1] H. M. Berman, *Nucleic Acids Res.* **2000**, 28(1), 235. <https://doi.org/10.1093/nar/28.1.235>
- [2] A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, F. T. Heer, T. A. P. de Beer, C. Rempfer, L. Bordoli, R. Lepore, T. Schwede, *Nucleic Acids Res.* **2018**, 46(W1), W296. <https://doi.org/10.1093/nar/gky427>
- [3] M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Židek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis, S. Velankar, *Nucleic Acids Res.* **2022**, 50(D1), D439. <https://doi.org/10.1093/nar/gkab1061>
- [4] A. Volkamer, M. Rarey, *Future Med. Chem.* **2014**, 6(3), 319. <https://doi.org/10.4155/fmc.14.3>
- [5] A. V. Sadybekov, V. Katritch, *Nature* **2023**, 616(7958), 673. <https://doi.org/10.1038/s41586-023-05905-z>
- [6] A. Volkamer, A. Griewel, T. Grombacher, M. Rarey, *J. Chem. Inf. Model.* **2010**, 50(11), 2041. <https://doi.org/10.1021/ci100241y>
- [7] X. Zheng, L. Gan, E. Wang, J. Wang, *AAPS. J.* **2013**, 15(1), 228. <https://doi.org/10.1208/s12248-012-9426-6>
- [8] N. K. Broomhead, M. E. Soliman, *Cell Biochem. Biophys.* **2017**, 75(1), 15. <https://doi.org/10.1007/s12013-016-0769-y>
- [9] J. Liao, Q. Wang, F. Wu, Z. Huang, *Molecules* **2022**, 27(20), 7103. <https://doi.org/10.3390/molecules27207103>
- [10] J. Graef, C. Ehrt, K. Diedrich, M. Poppinga, N. Ritter, M. Rarey, *J. Med. Chem.* **2022**, 65(2), 1384. <https://doi.org/10.1021/acs.jmedchem.1c01046>
- [11] M. Chartier, R. Najmanovich, *J. Chem. Inf. Model.* **2015**, 55(8), 1600. <https://doi.org/10.1021/acs.jcim.5b00333>
- [12] M. Naderi, J. M. Lemoine, R. G. Govindaraj, O. Z. Kana, W. P. Feinstein, M. Brylinski, *Briefings Bioinf.* **2019**, 20(6), 2167. <https://doi.org/10.1093/bib/bby078>
- [13] M. Eguida, D. Rognan, *Int. J. Mol. Sci.* **2022**, 23(20), 12462. <https://doi.org/10.3390/ijms232012462>
- [14] C. Ehrt, T. Brinkjost, O. Koch, *PLoS Comput. Biol.* **2018**, 14(11), e1006483. <https://doi.org/10.1371/journal.pcbi.1006483>
- [15] L. Pu, R. G. Govindaraj, J. M. Lemoine, H. C. Wu, M. Brylinski, *PLoS Comput. Biol.* **2019**, 15(2), e1006718. <https://doi.org/10.1371/journal.pcbi.1006718>
- [16] M. Simonovsky, J. Meyers, *J. Chem. Inf. Model.* **2020**, 60(4), 2356. <https://doi.org/10.1021/acs.jcim.9b00554>
- [17] Y. C. Chen, R. Tolbert, A. M. Aronov, G. McGaughey, W. P. Walters, L. Meireles, *J. Chem. Inf. Model.* **2016**, 56(9), 1734. <https://doi.org/10.1021/acs.jcim.6b00118>
- [18] C. Ehrt, T. Brinkjost, O. Koch, *MedChemComm* **2019**, 10(7), 1145. <https://doi.org/10.1039/c9md00102f>
- [19] V. Le Guilloux, P. Schmidtke, P. Tuffery, *BMC Bioinformatics* **2009**, 10, 168. <https://doi.org/10.1186/1471-2105-10-168>
- [20] M. Eguida, D. Rognan, *J. Med. Chem.* **2020**, 63(13), 7127. <https://doi.org/10.1021/acs.jmedchem.0c00422>
- [21] J. Desaphy, K. Azdimousa, E. Kellenberger, D. Rognan, *J. Chem. Inf. Model.* **2012**, 52(8), 2287. <https://doi.org/10.1021/ci300184x>
- [22] S. Li, C. Cai, J. Gong, X. Liu, H. Li, *Proteins: Struct. Funct. Bioinf.* **2021**, 89(11), 1541. <https://doi.org/10.1002/prot.26176>
- [23] J. Desaphy, G. Bret, D. Rognan, E. Kellenberger, *Nucleic Acids Res.* **2015**, 43(Database issue), D399. <https://doi.org/10.1093/nar/gku928>
- [24] A. Bhadra, K. Yeturu, *Mach. Learn. Sci. Technol.* **2020**, 2(1), 015005. <https://doi.org/10.1088/2632-2153/abad88>
- [25] M. Gao, J. Skolnick, *Bioinformatics* **2013**, 29(5), 597. <https://doi.org/10.1093/bioinformatics/btt024>
- [26] P. Anand, D. Nagarajan, S. Mukherjee, N. Chandra, *Database* **2014**, 2014, bau029. <https://doi.org/10.1093/database/bau029>
- [27] O. B. Scott, J. Gu, A. W. E. Chan, *J. Chem. Inf. Model.* **2022**, 62(22), 5383. <https://doi.org/10.1021/acs.jcim.2c00832>

- [28] R. A. Laskowski, *J. Mol. Graphics* **1995**, 13(5), 323. [https://doi.org/10.1016/0263-7855\(95\)00073-9](https://doi.org/10.1016/0263-7855(95)00073-9)
- [29] K. E. Choi, A. Balupuri, N. S. Kang, *Comput. Struct. Biotechnol. J.* **2023**, 21, 425. <https://doi.org/10.1016/j.csbj.2022.12.014>
- [30] K. Yeturu, N. Chandra, *BMC Bioinformatics* **2008**, 9, 543. <https://doi.org/10.1186/1471-2105-9-543>
- [31] H. W. Kuhn, *Naval Res. Log. Quar.* **1955**, 2(1–2), 83. <https://doi.org/10.1002/nav.3800020109>
- [32] K. Diedrich, J. Graef, K. Schöning-Stierand, M. Rarey, *Bioinformatics* **2021**, 37(3), 424. <https://doi.org/10.1093/bioinformatics/btaa693>
- [33] T. Inhester, S. Bietz, M. Hilbig, R. Schmidt, M. Rarey, *J. Chem. Inf. Model.* **2017**, 57(2), 148. <https://doi.org/10.1021/acs.jcim.6b00561>
- [34] S. Urbaczek, A. Kolodzik, I. Groth, S. Heuser, M. Rarey, *J. Chem. Inf. Model.* **2013**, 53(1), 76. <https://doi.org/10.1021/ci300358c>
- [35] S. Bietz, T. Inhester, F. Lauck, K. Sommer, M. M. von Behren, R. Fährrolfes, F. Flachsenberg, A. Meyder, E. Nittinger, T. Otto, M. Hilbig, K. T. Schomburg, A. Volkamer, M. Rarey, *J. Biotechnol.* **2017**, 261, 207. <https://doi.org/10.1016/j.jbiotec.2017.06.004>
- [36] S. Urbaczek, A. Kolodzik, J. R. Fischer, T. Lippert, S. Heuser, I. Groth, T. Schulz-Gasch, M. Rarey, *J. Chem. Inf. Model.* **2011**, 51(12), 3199. <https://doi.org/10.1021/ci200324e>
- [37] C. P. Gomes, D. E. Fernandes, F. Casimiro, G. F. da Mata, M. T. Passos, P. Varela, G. Mastroianni-Kirsztajn, J. B. Pesquero, *Front. Cell. Infect. Microbiol.* **2020**, 10, 589505. <https://doi.org/10.3389/fcimb.2020.589505>
- [38] A. Kahraman, R. J. Morris, R. A. Laskowski, J. M. Thornton, *J. Mol. Biol.* **2007**, 368(1), 283. <https://doi.org/10.1016/j.jmb.2007.01.086>
- [39] D. J. Wood, J. Vlieg, M. Wagener, T. Ritschel, *J. Chem. Inf. Model.* **2012**, 52(8), 2031. <https://doi.org/10.1021/ci3000776>
- [40] T. Krotzky, C. Grunwald, U. Egerland, G. Klebe, *J. Chem. Inf. Model.* **2015**, 55(1), 165. <https://doi.org/10.1021/ci5005898>
- [41] N. Weill, D. Rognan, *J. Chem. Inf. Model.* **2010**, 50(1), 123. <https://doi.org/10.1021/ci900349y>
- [42] Y. Zhang, *Nucleic Acids Res.* **2005**, 33(7), 2302. <https://doi.org/10.1093/nar/gki524>
- [43] J. Konc, D. Janežič, *Bioinformatics* **2010**, 26(9), 1160. <https://doi.org/10.1093/bioinformatics/btq100>
- [44] G. Marcou, D. Rognan, *J. Chem. Inf. Model.* **2007**, 47(1), 195. <https://doi.org/10.1021/ci600342e>
- [45] J. Desaphy, E. Raimbaud, P. Ducrot, D. Rognan, *J. Chem. Inf. Model.* **2013**, 53(3), 623. <https://doi.org/10.1021/ci300566n>
- [46] J. Batista, P. C. Hawkins, R. Tolbert, M. T. Geballe, *J. Cheminf.* **2014**, 6(S1), P57. <https://doi.org/10.1186/1758-2946-6-s1-p57>
- [47] S. Schmitt, D. Kuhn, G. Klebe, *J. Mol. Biol.* **2002**, 323(2), 387. [https://doi.org/10.1016/s0022-2836\(02\)00811-2](https://doi.org/10.1016/s0022-2836(02)00811-2)
- [48] L. Xie, L. Xie, P. E. Bourne, *Bioinformatics* **2009**, 25(12), i305. <https://doi.org/10.1093/bioinformatics/btp220>
- [49] A. Shulman-Peleg, R. Nussinov, H. J. Wolfson, *J. Mol. Biol.* **2004**, 339(3), 607. <https://doi.org/10.1016/j.jmb.2004.04.012>
- [50] C. Schalon, J. S. Surgand, E. Kellenberger, D. Rognan, *Proteins Struct. Funct. Bioinf.* **2008**, 71(4), 1755. <https://doi.org/10.1002/prot.21858>
- [51] S. Barelier, T. Sterling, M. J. O'Meara, B. K. Shoichet, *ACS Chem. Biol.* **2015**, 10(12), 2772. <https://doi.org/10.1021/acschembio.5b00683>
- [52] S. F. OpenEye Scientific Software, NM. **2023**. *Shape Toolkit*. <http://www.eyesopen.com>
- [53] S. F. OpenEye Scientific Software, NM. **2023**. *Spicoli Toolkit*. <http://www.eyesopen.com>
- [54] S. F. OpenEye Scientific Software, NM. **2023**. *ROCS*. <http://www.eyesopen.com>
- [55] M. A. Fligner, J. S. Verducci, P. E. Blower, *Technometrics* **2002**, 44(2), 110. <https://doi.org/10.1198/004017002317375064>
- [56] C. Ehrt, *The Impact of Binding Site Similarity on Hit Identification in Early Drug Discovery*, TU Dortmund University, Dortmund, Germany **2019**.
- [57] J. Graef, C. Ehrt, M. Rarey, *J. Chem. Inf. Model.* **2023**, 63(10), 3128. <https://doi.org/10.1021/acs.jcim.3c00336>
- [58] B. Zdrazil, E. Felix, F. Hunter, E. J. Manners, J. Blackshaw, S. Corbett, M. de Veij, H. Ioannidis, D. M. Lopez, J. F. Mosquera, M. P. Magarinos, N. Bosc, R. Arcila, T. Kizilören, A. Gaulton, A. P. Bento, M. F. Adasme, P. Monecke, G. A. Landrum, A. R. Leach, *Nucleic Acids Res.* **2023**, 52, D1180. <https://doi.org/10.1093/nar/gkad1004>
- [59] M. Novinec, B. Lenarčič, *BioMol. Concepts* **2013**, 4(3), 287. <https://doi.org/10.1515/bmc-2012-0054>
- [60] X. Ou, Y. Liu, X. Lei, P. Li, D. Mi, L. Ren, L. Guo, R. Guo, T. Chen, J. Hu, Z. Xiang, Z. Mu, X. Chen, J. Chen, K. Hu, Q. Jin, J. Wang, Z. Qian, *Nat. Commun.* **2020**, 11(1), 1620. <https://doi.org/10.1038/s41467-020-15562-9>
- [61] P. Y. A. Reinke, E. E. de Souza, S. Günther, S. Falke, J. Lieske, W. Ewert, J. Loboda, A. Herrmann, A. Rahmani Mashhour, K. Karničar, A. Usenik, N. Lindič, A. Sekirnik, V. F. Botosso, G. M. M. Santelli, J. Kapronezai, M. V. de Araújo, T. T. Silva-Pereira, A. F. S. Filho, M. S. Tavares, L. Flórez-Álvarez, D. B. L. de Oliveira, E. L. Durigon, P. R. Giaretta, M. B. Heinemann, M. Hauser, B. Seychell, H. Böhrer, W. Rut, M. Drag, T. Beck, R. Cox, H. N. Chapman, C. Betzel, W. Brehm, W. Hinrichs, G. Ebert, S. L. Latham, A. M. S. Guimarães, D. Turk, C. Wrenger, A. Meents, *Commun. Biol.* **2023**, 6(1), 1058. <https://doi.org/10.1038/s42003-023-05317-9>
- [62] G. Wang, R. L. Dunbrack Jr., *Bioinformatics* **2003**, 19(12), 1589. <https://doi.org/10.1093/bioinformatics/btg224>
- [63] F. Madeira, M. Pearce, A. R. N. Tivey, P. Basutkar, J. Lee, O. Edbali, N. Madhusoodanan, A. Kolesnikov, R. Lopez, *Nucleic Acids Res.* **2022**, 50(W1), W276. <https://doi.org/10.1093/nar/gkac240>
- [64] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, T. E. Ferrin, *J. Comput. Chem.* **2004**, 25(13), 1605. <https://doi.org/10.1002/jcc.20084>
- [65] J. Kyte, R. F. Doolittle, *J. Mol. Biol.* **1982**, 157(1), 105. [https://doi.org/10.1016/0022-2836\(82\)90515-0](https://doi.org/10.1016/0022-2836(82)90515-0)
- [66] T. D. Goddard, C. C. Huang, E. C. Meng, E. F. Pettersen, G. S. Couch, J. H. Morris, T. E. Ferrin, *Protein Sci.* **2018**, 27(1), 14. <https://doi.org/10.1002/pro.3235>
- [67] L. A. Hardegger, B. Kuhn, B. Spinnler, L. Anselm, R. Ecabert, M. Stihle, B. Gsell, R. Thoma, J. Diez, J. Benz, J. M. Plancher, G. Hartmann, D. W. Banner, W. Haap, F. Diederich, *Angew. Chem. Int. Ed.* **2011**, 50(1), 314. <https://doi.org/10.1002/anie.201006781>
- [68] S. Bietz, S. Urbaczek, B. Schulz, M. Rarey, *J. Cheminf.* **2014**, 6, 12. <https://doi.org/10.1186/1758-2946-6-12>
- [69] G. Gaël, J. Benoît, **2010**. *Eigen v3 (C++ library)*. <http://eigen.tuxfamily.org>
- [70] S. Umeyama, *IEEE. Trans. Pattern. Anal. Mach. Intell.* **1991**, 13(4), 376. <https://doi.org/10.1109/34.88573>
- [71] E. Calenoff, *ISRN Neurol.* **2012**, 2012, 1. <https://doi.org/10.5402/2012/851541>
- [72] S. Beisken, T. Meinl, B. Wiswedel, L. F. de Figueiredo, M. Berthold, C. Steinbeck, *BMC Bioinformatics* **2013**, 14, 257. <https://doi.org/10.1186/1471-2105-14-257>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: T. Reim, C. Ehrt, J. Graef, S. Günther, A. Meents, M. Rarey, *Arch. Pharm.* **2024**;357:e2300661. <https://doi.org/10.1002/ardp.202300661>