Emerging structure-based computational methods to screen the exploding accessible chemical space.

Corentin Bedart, Conrad Veranso Simoben, Matthieu Schapira

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Emerging structure-based computational methods to screen the exploding accessible chemical space

Corentin Bedart[1], Conrad Veranso Simoben[2] and Matthieu Schapira[2,3]

Abstract

Structure-based virtual screening can be a valuable approach to computationally select hit candidates based on their predicted interaction with a protein of interest. The recent explosion in the size of chemical libraries increases the chances of hitting high-quality compounds during virtual screening exercises but also poses new challenges as the number of chemically accessible molecules grows faster than the computing power necessary to screen them. We review here two novel approaches rapidly gaining in popularity to address this problem: machine learning-accelerated and synthon-based library screening. We summarize the results from seminal proof-of-concept studies, highlight the latest developments, and discuss limitations and future directions.

Addresses

[1] Univ. Lille, Inserm, CHU Lille, U1286 - INFINITE - Institute for Translational Research in Inflammation, F-59000, Lille, France
[2] Structural Genomics Consortium, University of Toronto, 101 College Street, MaRS South Tower, Suite 700, Toronto, Ontario M5G 1L7, Canada
[3] Department of Pharmacology and Toxicology, University of Toronto, 1 King's College Circle, Toronto, Ontario M5S 1A8, Canada

Corresponding author: Schapira, Matthieu (matthieu.schapira@utoronto.ca)
(Schapira M.)

Current Opinion in Structural Biology 2024, **86**:102812

This review comes from a themed issue on **New Concepts in Drug Discovery (2024)**

Edited by **Andrea Cavalli** and **Alessio Ciulli**

For complete overview of the section, please refer the article collection - New Concepts in Drug Discovery (2024)

Available online 10 April 2024

https://doi.org/10.1016/j.sbi.2024.102812

Introduction

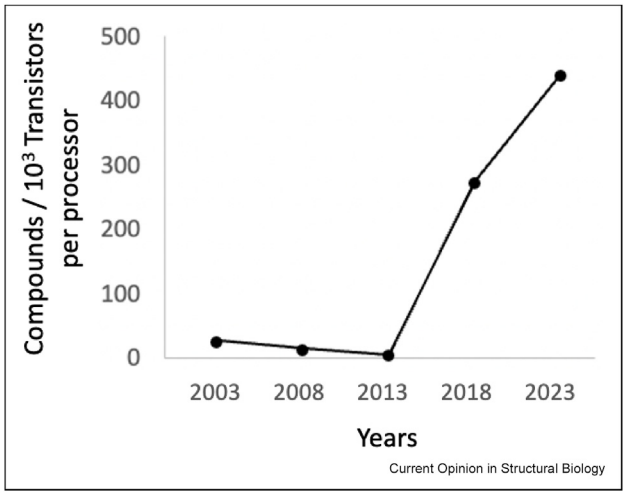
Structure-based virtual screening is a common technique where drug-like molecules are docked to the structure of a protein target to predict which compounds out of a large chemical library bind to the target and should therefore be tested experimentally [1]. Performances vary with the target, the computational tools, and the scientist, but in favorable cases, they can be significantly better than random and have slowly but steadily improved over the years. While the emergence of deep learning is expected to have a positive and potentially profound impact in the field, the two main engines of progress until now have arguably been (1) the rapid increase in computing power and (2) the recent explosion in the accessible chemical space [2]. Indeed, new chemotypes found only in ultra-large chemical libraries sometimes lead to more potent ligands that better emerge from the noise inherent to virtual screening [3]. The number of compounds available from the three leading commercial sources (Enamine REAL, WuXi GalaXi, and Otava CHEMriya) has grown from about 25 million to 50 billion in the past ten years, and even larger libraries are reported in pharmaceutical companies [4]. As the growth rate of chemical libraries surpasses that of computing power (Figure 1), more efficient virtual screening methods and tools are emerging to explore the ever-expanding universe of drug-like molecules. In particular, machine learning (ML)-accelerated virtual screening and synthon-based library screening (SBLS) are the centers of increasing attention in the field, which we review here.

Machine learning-accelerated virtual screening

Mechanisms to use ML as a tool to accelerate the virtual screening of ultra-large chemical libraries are actively explored [5]. A strategy that is rapidly gaining popularity is to apply conventional—generally physics-based—computational techniques to screen a small subset of the library and then use the results to train ML models that can quickly screen billions of molecules. The process can be repeated in multiple active-learning cycles. During each cycle, the hits predicted by ML are evaluated using physics-based methods to improve the ML model [6,7] (Figure 2).

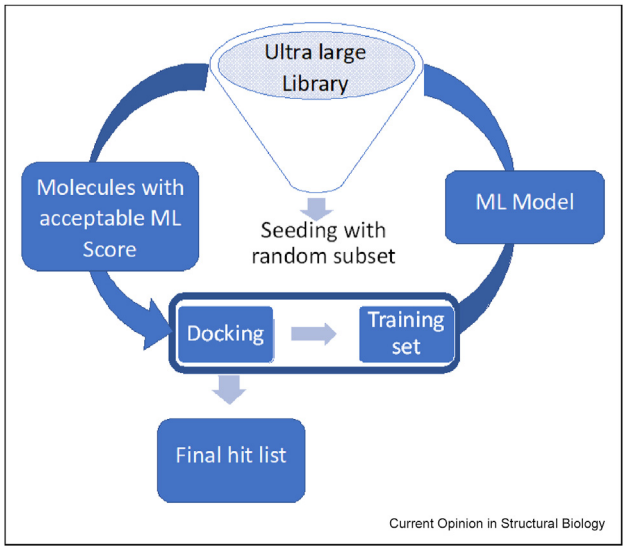
In a seminal work, Gentile et al. introduced in 2020 the open-source Deep Docking approach, where they screened over 1 billion molecules against twelve targets [8]. The strategy included 11 cycles where, at each cycle, 1 million compounds predicted by a deep neural network were docked and scored with OpenEye's FRED

Figure 1



The growth of chemical libraries outpaces the growth of computing power. The aggregated number of commercial compounds available from Enamine REAL, WuXi GalaXi, and Otava CHEMriya libraries is divided by the average number of transistors per processor. The number of compounds were obtained from the vendors' websites and the Enamine Webinar (URL: https://www.youtube.com/watch?v=Vn5Z2nFhXL4). The number of transistors per processor was adapted from (URL: https://github.com/karlrupp/microprocessor-trend-data).

Figure 2



Machine learning-accelerated virtual screening. Throughout multiple active learning cycles, docking scores are slowly calculated with conventional methods for small subsets of a large library and rapidly predicted with gradually improved ML models for the rest of the library.

[9]. The neural network was iteratively re-trained on the augmented docking/scoring data to improve the hit rate in the next accelerated ML screen of the full library. Despite an overall 100-fold reduction in the number of compounds docked, hit enrichment rates of 6 to 660-fold in the top 100,000 molecules and 240 to 6000-fold in the top ten were observed. It is important to note that these enrichment rates were calculated based on virtual hits, i.e. compounds predicted by FRED to be active. The enrichment rate in bioactive molecules would depend on the accuracy of docking scores generated by FRED.

Another implementation of the same strategy published the following year used Glide or DOCK3.7 [10] to screen 0.1% of the library and Schrödinger's AutoQSAR/DeepChem ML engine for score prediction [11]. The study successfully recovered 80% of experimentally confirmed hits while saving 14-fold in computing costs. Interestingly, the authors investigated various strategies for selecting ML training sets and found that re-training the ML models on compounds predicted with low confidence to have good scores enabled a good balance between exploration and exploitation and yielded the best performance. The same dataset was used by another group to compare the performance of random forest with neural network architectures using MolPAL [12]. Although differences were not always significant, a message-passing neural network [13] consistently outperformed the other methods.

In a related contemporaneous work, a Lean Docking strategy was developed [14]. This approach included a single (passive) ML-training step instead of multiple active learning cycles, trained on docking scores generated by one of five different docking tools to screen fifteen targets. The training data set was generated using Molecular Operating Environment from Chemical Computing Group, Montreal (CCG's MOE), Schrödinger's Glide [15], OpenEye's FRED [9], CCDC's Gold [16], and AutoDock-Vina [17]. Instead of a ML classifier, the authors applied a regression model, resulting in nearly 6000 predicted docking scores per second and per CPU. This led to a 75% reduction in the number of compounds docked without any significant loss in virtual screening performance. However, the results varied significantly depending on the docking/scoring software used.

While ML-accelerating engines typically use a string or two-dimensional representation of compounds, Geometry Enhanced Molecular screen (GEM-screen) incorporates the docked pose of the compounds in the training set [18]. Although the performances did not appear superior to some of the other work reviewed above, this effort exemplifies yet another possible vector of optimization for ML-accelerated virtual screening.

More recently reported, PyRMD2Dock [19] is an open-source implementation of the approach. It uses PyRMD [20], a random matrix determinant algorithm previously designed for ligand-based virtual screening, as the ML acceleration engine coupled with the docking software AutoDock-GPU [21], with encouraging results.

HASTEN is another open-source tool that uses Chemprop, a directed-message-passing neural network, to recover 90% of top-scoring hits while docking only 1% of the library [22,23].

Given the fast progress in the field of deep learning and the massive expansion of chemical libraries, ML-accelerated virtual screening is bound to become an increasingly popular approach in computational hit finding, and commercial tools such as Molsoft's GigaScreen (https://www.molsoft.com/GigaScreen.html), OpenEye's Gigadock Warp (https://docs.eyesopen.com/floe/modules/large-scale-floes/docs/source/explanations/gigadock_warp_explanation.html) or Schrodinger's Active Learning Glide (https://newsite.schrodinger.com/platform/products/active-learning-applications/) were released in the past few months. To guide and democratize the use of ML-accelerated methods, Tran-Nguyen et al. are now providing a set of protocols, scripts, and datasets, which can be helpful for both experts and nonexperts [24].

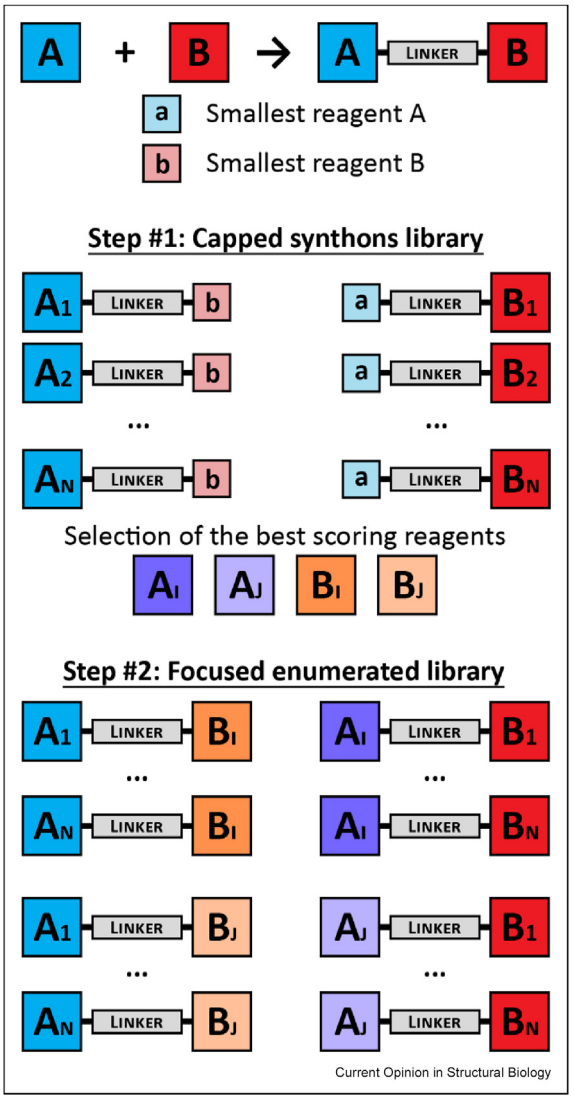
Although this strategy is becoming a leading approach to address the explosion of the accessible chemistry space, it still presents practical challenges due to the storage and management of billions of compounds, which will only increase in the near future. One viable solution that has gained momentum in recent years is SBLS.

Synthon-based approaches

SBLS aims to identify new hits by virtual screening without assembling, storing, and screening the full enumeration of the screened library [25]. Here, a library is represented as the sum of its building blocks (or synthons). The concept is related to fragment-based drug discovery (FBDD), where fragments are linked within the boundaries of a pre-determined library for rapid chemical accessibility.

In a SBLS screen, fully enumerated hit candidates are progressively built through multiple synthon selection stages (Figure 3). In the case of a single-step synthetic route, if two synthons A and B with reactive groups x and y react to form an enumerated compound A-B (A-x + B-y -> A-B), a first step is generally to identify and remove reactive groups from synthons that may mislead the docking/scoring process. For instance, in A-x, the reactive group is either removed (A-x -> A), replaced with a methyl group (A-x -> A-m), or replaced with the smallest compatible B (A-x -> A-b). The same operation is repeated for all synthons, followed by a virtual screen of the modified synthon library. In this way, the chemical diversity of the library is explored additively (A_1, …, A_n, B_1, …, B_n) instead of combinatorially (A_1B_1, A_1B_2, A_2B_1, A_2B_2, …, A_nB_n), enabling a dramatic decrease in computational cost, especially for large libraries or multi-step chemical reactions including 3 or more

Figure 3



Synthon-based library screening. Synthons are capped, docked, and scored. The top predicted synthons are then used to enumerate a smaller, focused library for virtual screening.

synthons. After selecting the best synthons based on their docking score and/or other criteria, such as the orientation of their attachment point in the docked pose, they are combined with all other compatible synthons to create a focused, enumerated library that can be subjected to another round of virtual screening. This ensures that only compounds containing one of the best synthons are considered, considerably reducing the combinatorial chemical space to be explored.

The origins of SBLS can be traced back to the 1990s, with the CombiDOCK strategy to efficiently dock a large combinatorial library into a target receptor. CombiDOCK was initially tested on a single chemical

reaction using a 10 × 10 × 10 combinatorial library, with results deemed encouraging at the time [26]. The concept evolved further with the introduction of PRO_SELECT in 2002 [27] and de novo methods based on FlexX in 2006 [28,29]. These developments ultimately led to the creation of the Basis Products method in 2009 [25], which laid the foundations for SBLS algorithms currently in use. Using a known 2-reagent chemical reaction in the form of A+B→AB, capping reactants are selected by identifying the smallest a and b and Basis Products are formed by capping all synthons (A_1b, A_2b, …, A_nb; aB_1, aB_2, …, aB_n). This provides a representative subset of the entire library, which is screened virtually to identify the best reactants (A_i, A_j, B_i, and B_j). These are finally combined with all synthons (A_1B_i, …, A_nB_i; A_iB_1, …, A_iB_n) to identify hits for further computational or experimental evaluation.

In 2021, the basis products concept was improved and successfully applied with V-SYNTHES (Virtual SYNThon Hierarchical Enumeration Screening) [30], used to efficiently screen computationally 11 billion compounds from Enamine REAL spanning multiple chemical reactions. Reactants are typically capped with a methyl or phenyl group and screened virtually with Molsoft's ICM (but any other virtual screening software can be used instead). The cap of the best-scoring compounds is replaced with all chemically compatible building blocks, and the resulting library is screened again to select candidate hits for further experimental testing. In a proof-of-concept study aimed at identifying new cannabinoid antagonists that selectively target the CB2 receptor, a significantly improved hit rate was observed compared to conventional screening methods. V-SYNTHES achieved a hit rate of 33%, while a standard virtual screen required 100 times more computational resources to achieve a hit rate of only 15%.

Developed simultaneously, Chemical Space Docking follows a nearly identical strategy [31]. Reactive groups in synthons are replaced with a dummy atom to generate fragments docked with FlexX [32] into a binding site with pharmacophoric constraints. The best fragments are selected based on various criteria, such as chemical diversity, ligand efficiency, torsion energy, and/or physicochemical properties, and expanded using other compatible synthons from a list of chemical reactions. In a proof-of-concept study aimed at discovering novel ROCK1 kinase inhibitors, this SBLS method once again demonstrated an impressive hit rate of 39% after selecting 69 out of one billion compounds.

SpaceDock, a variation on the theme, was recently applied to discover novel ligands for the dopamine D3 receptor. Here, reagents are directly docked (with no preliminary capping) and linked inside the binding pocket [33]. Shape-Aware Synthon Search (SASS) is another variation where shape similarity to a query molecule instead of docking to a binding pocket is used to select synthons of interest [34]. Commercial solutions such as Cresset's Ignite™ (https://www.cresset-group.com/discovery/specific/virtual-screening/) are also becoming available.

Following these recent and encouraging results, open-source software is being developed to facilitate synthon-based virtual library screening. For example, SATELLiTES (Synthon-based Approach for the Targeted Enumeration of Ligand Libraries and Expeditious Screening) [35] uses a chemical reaction and libraries of compatible synthons as input. Users have the option to select their own capping group, such as the smallest reactant, a specific reagent from the dataset, or a dummy reagent that best mimics pharmacophoric expectations. Once the best capped synthon candidates have been selected with a user's preferred virtual screening tool, SATELLiTES generates a focused library for further screening.

Concluding remarks and outlook

Clearly, recent successes in ML-accelerated or synthon-driven screening of ultra large chemical libraries propelled these two rapidly growing trends to efficiently explore a chemical space that has grown beyond the reach of standard virtual screening techniques. Successful application of these methods in the prospective discovery of novel ligands underscores their real-world potential [30,31,33,36]. But important challenges remain to be addressed before either approach or their combination is firmly established as the gold standard for virtual screening. First and foremost, both approaches rely on one or multiple docking/scoring steps and are bound to fail if docking poses are inaccurate or scoring functions are insufficiently robust. Continued efforts to improve docking poses and, especially, scoring functions are therefore critical. A promising area of investigation here is to train neural networks on quantum chemistry data to dramatically increase prediction accuracy while maintaining high throughput [37]. Synthon-based approaches also rely on the assumption that building blocks adopt the same binding pose in isolation and in the context of an enumerated molecule, which may not always be the case and can depend on the selection of appropriate capping groups to represent reactive sites. An elegant solution is the synthon-based Thompson sampling approach reported by Klarich et al., which learns through iterative cycles the best molecules to enumerate and score [38]. The authors found that over 50% of the top 100 scoring compounds were retrieved while docking/scoring less than 1% of the full library. A more pragmatic challenge facing ML-accelerated screening of fully enumerated libraries is the ever-increasing internet resources necessary to

download billions and soon trillions of molecules and the storage capacity required to manage 10 to 50 times more 3D conformers generated locally or publicly available [39], though the latter is not necessary if 3D conformers are generated on the fly (for instance, with Auto3D [40]) for compounds actually docked. Limiting computational screening campaigns to drug-like molecules does not alleviate this challenge; for instance, the Enamine Real Database contains over six billion compounds satisfying Lipinski and Weber rules [41,42]. Altogether, after decades of slow and incremental progress, the field is suddenly benefiting from an explosion in the size and diversity of the accessible chemical space and from rapid developments in machine learning. Benchmarking exercises such as CACHE may set the stage for a future breakthrough, as was experienced with AlphaFold for protein structure prediction. Interestingly, we note that one of the two best-performing virtual screening pipelines in the first CACHE challenge [43] used deep docking as a primary screening step, and fragment-based strategies were employed by three of the seven top-performing workflows (https://cache-challenge.org/). Another successful avenue was the use of ML-driven generative design techniques such as REINVENT [44] to generate customized molecules for a specific binding site. While this approach does not directly explore the commercial chemical space, it is becoming increasingly relevant as the chances of finding a close commercial analog of a computationally invented molecule increase with the rapid growth of on-demand libraries. Only one thing is certain: it will be captivating to see where the field is going in the upcoming months and years.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgements

M.S. gratefully acknowledges support from the Canadian Strategic Innovation Fund (SIF Stream 5), the University of Toronto's Data Sciences Institute and EPIC Consortium, MITACS accelerate (IT13051), NSERC (RGPIN-2019-04416), and CIHR (202309PJT). The Structural Genomics Consortium is a registered charity (no: 1097737) that receives funds from Bayer AG, Boehringer Ingelheim, Bristol-Myers Squibb, Genentech, Genome Canada through Ontario Genomics Institute [OGI-196], EU/EFPIA/OICR/McGill/KTH/Diamond Innovative Medicines Initiative 2 Joint Undertaking [EUbOPEN grant 875510], Janssen, Merck KGaA (aka EMD in Canada and US), Pfizer and Takeda.

References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest
** of outstanding interest

1. * Sadybekov AV, Katritch V: **Computational approaches streamlining drug discovery**. *Nature* 2023, **616**:673–685.
An outstanding survey of the state-of-the-art and challenges in virtual ligand screening for drug discovery.

2. Cherkasov A: **The 'Big Bang'of the chemical universe**. *Nat Chem Biol* 2023:1–2.

3. Lyu J, Wang S, Balius TE, Singh I, Levit A, Moroz YS, O'Meara MJ, Che T, Algaa E, Tolmachova K: **Ultra-large library docking for discovering new chemotypes**. *Nature* 2019, **566**:224–229.

4. * Warr WA, Nicklaus MC, Nicolaou CA, Rarey M: **Exploration of ultralarge compound collections for drug discovery**. *J Chem Inf Model* 2022, **62**:2021–2034.
This article provides some insight on the rapid expansion of chemical libraries, not only in commercial vendors but also in large pharmaceutical companies.

5. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, Li B, Madabhushi A, Shah P, Spitzer M: **Applications of machine learning in drug discovery and development**. *Nat Rev Drug Discov* 2019, **18**:463–477.

6. Cavasotto CN, Di Filippo JI: **The impact of supervised learning methods in ultralarge high-throughput docking**. *J Chem Inf Model* 2023, **63**:2267–2280.

7. Kuan J, Radaeva M, Avenido A, Cherkasov A, Gentile F: **Keeping pace with the explosive growth of chemical libraries with structure-based virtual screening**. *Wiley Interdiscip Rev Comput Mol Sci* 2023, e1678.

8. Gentile F, Agrawal V, Hsing M, Ton A-T, Ban F, Norinder U, Gleave ME, Cherkasov A: **Deep docking: a deep learning platform for augmentation of structure based drug discovery**. *ACS Cent Sci* 2020, **6**:939–949.
Seminal proof-of-concept use of deep learning and active learning to accelerate structure-based virtual screening.

9. McGann M: **FRED and HYBRID docking performance on standardized datasets**. *J Comput Aided Mol Des* 2012, **26**:897–906.

10. Mysinger MM, Shoichet BK: **Rapid context-dependent ligand desolvation in molecular docking**. *J Chem Inf Model* 2010, **50**:1561–1573.

11. ** Yang Y, Yao K, Repasky MP, Leswing K, Abel R, Shoichet BK, Jerome SV: **Efficient exploration of chemical space with docking and deep learning**. *J Chem Theor Comput* 2021, **17**:7106–7119.
Seminal use of deep learning and active learning to accelerate structure-based virtual screening.

12. Graff DE, Shakhnovich EI, Coley CW: **Accelerating high-throughput virtual screening through molecular pool-based active learning**. *Chem Sci* 2021, **12**:7866–7881.

13. Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, Guzman-Perez A, Hopper T, Kelley B, Mathea M: **Analyzing learned molecular representations for property prediction**. *J Chem Inf Model* 2019, **59**:3370–3388.

14. Berenger F, Kumar A, Zhang KY, Yamanishi Y: **Lean-docking: exploiting ligands' predicted docking scores to accelerate molecular docking**. *J Chem Inf Model* 2021, **61**:2341–2352.

15. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK: **Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy**. *J Med Chem* 2004, **47**:1739–1749.

16. Jones G, Willett P, Glen RC, Leach AR, Taylor R: **Development and validation of a genetic algorithm for flexible docking**. *J Mol Biol* 1997, **267**:727–748.

17. Trott O, Olson AJ: **AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading**. *J Comput Chem* 2010, **31**:455–461.

18. Yu L, He X, Fang X, Liu L, Liu J: **Deep learning with geometry-enhanced molecular representation for augmentation of large-scale docking-based virtual screening**. *J Chem Inf Model* 2023, **63**:6501–6514.

19. Roggia M, Natale B, Amendola G, Di Maro S, Cosconati S: **Streamlining large chemical library docking with artificial intelligence: the PyRMD2Dock approach**. *J Chem Inf Model* 2023.

20. Amendola G, Cosconati S: **PyRMD: a new fully automated ai-powered ligand-based virtual screening tool**. *J Chem Inf Model* 2021, **61**:3835–3845.

21. Santos-Martins D, Solis-Vasquez L, Tillack AF, Sanner MF, Koch A, Forli S: **Accelerating AutoDock4 with GPUs and gradient-based local search**. *J Chem Theor Comput* 2021, **17**: 1060–1073.

22. Kalliokoski T: **Machine learning boosted docking (HASTEN): an open-source tool to accelerate structure-based virtual screening campaigns**. *Molecular Informatics* 2021, **40**, 2100089.

23. Sivula T, Yetukuri L, Kalliokoski T, Käsnänen H, Poso A, Pöhner I: **Machine learning-boosted docking enables the efficient structure-based virtual screening of giga-scale enumerated chemical libraries**. *J Chem Inf Model* 2023, **63**:5773–5783.

24. Tran-Nguyen V-K, Junaid M, Simeon S, Ballester PJ: **A practical guide to machine-learning scoring for structure-based virtual screening**. *Nat Protoc* 2023, **18**:3460–3511.

25. Zhou JZ, Shi S, Na J, Peng Z, Thacher T: **Combinatorial library-based design with Basis Products**. *J Comput Aided Mol Des* 2009, **23**:725–736.

26. Sun Y, Ewing T, Skillman A, Kuntz I: **CombiDOCK: structure-based combinatorial docking and library design**. *J Comput Aided Mol Des* 1998, **12**:597–604.

27. Liebeschuetz JW, Jones SD, Morgan PJ, Murray CW, Rimmer AD, Roscoe JM, Waszkowycz B, Welsh PM, Wylie WA, Young SC: **PRO_SELECT: combining structure-based drug design and array-based chemistry for rapid lead discovery. 2. The development of a series of highly potent and selective factor Xa inhibitors**. *J Med Chem* 2002, **45**:1221–1232.

28. Gastreich M, Lilienthal M, Briem H, Claussen H: **Ultrafast de novo docking combining pharmacophores and combinatorics**. *J Comput Aided Mol Des* 2006, **20**:717–734.

29. Degen J, Rarey M: **FlexNovo: structure-based searching in large fragment spaces**. *ChemMedChem: Chemistry Enabling Drug Discovery* 2006, **1**:854–868.

30. ** Sadybekov AA, Sadybekov AV, Liu Y, Iliopoulos-Tsoutsouvas C, Huang X-P, Pickett J, Houser B, Patel N, Tran NK, Tong F: **Synthon-based ligand discovery in virtual libraries of over 11 billion compounds**. *Nature* 2022, **601**:452–459.

Seminal proof-of-concept study demonstrating the use of SBLS to efficiently screen billions of compounds.

31. ** Beroza P, Crawford JJ, Ganichkin O, Gendelev L, Harris SF, Klein R, Miu A, Steinbacher S, Klingler F-M, Lemmen C: **Chemical space docking enables large-scale structure-based virtual screening to discover ROCK1 kinase inhibitors**. *Nat Commun* 2022, **13**:6447.

Seminal proof-of-concept study demonstrating the use of SBLS to efficiently screen billions of compounds.

32. Rarey M, Kramer B, Lengauer T, Klebe G: **A fast flexible docking method using an incremental construction algorithm**. *J Mol Biol* 1996, **261**:470–489.

33. * Sindt F, Seyller A, Eguida M, Rognan D: **Protein structure-based organic chemistry-driven ligand design from ultralarge chemical spaces**. *ACS Cent Sci* 2024, https://doi.org/10.1021/acscentsci.3c01521.

A first effort to link fragments in the context of their binding poses, which imposes necessary restraints and accelerates SBLS.

34. * Cheng C, Beroza P: **Shape-Aware Synthon Search (SASS) for virtual screening of synthon-based chemical spaces**. *J Chem Inf Model* 2024, **64**:1251–1260.

A pioneering implementation of SBLS where queries are ligands rather than binding sites.

35. Bedart C, Shimokura G, West FG, Wood TE, Batey RA, Irwin JJ, Schapira M: **A mechanism to open academic chemistry to high-throughput virtual screening**. *Chem* 2023.

36. ** Gentile F, Fernandez M, Ban F, Ton A-T, Mslati H, Perez CF, Leblanc E, Yaacoub JC, Gleave J, Stern A: **Automated discovery of noncovalent inhibitors of SARS-CoV-2 main protease by consensus Deep Docking of 40 billion small molecules**. *Chem Sci* 2021, **12**:15960–15974.

The first reported prospective discovery of novel inhibitors via ML-accelerated virtual screening.

37. * Glick Z, Metcalf D, Sargent C, Spronk S, Koutsoukas A, Cheney D, Sherrill CD: ***A physics-aware neural network for protein-ligand interactions with quantum chemical accuracy***. https://doi.org/10.26434/chemrxiv-2024-5v6gh.

Pionnering effort to train a machine learning potential on quantum chemistry calculations for a dataset composed of paired ligand and protein fragments could lead to improved scoring functions.

38. ** Klarich K, Goldman B, Kramer T, Riley P, Walters WP: **Thompson Sampling– an efficient method for searching ultralarge synthesis on demand databases**. *J Chem Inf Model* 2024.

A novel synthon-based virtual screening approach that docks fully enumerated compounds instead of fragments but still avoids the computational cost of enumerating and docking full libraries.

39. Tingle BI, Tang KG, Castanon M, Gutierrez JJ, Khurelbaatar M, Dandarchuluun C, Moroz YS, Irwin JJ: **ZINC-22– A free multi-billion-scale database of tangible compounds for ligand discovery**. *J Chem Inf Model* 2023, **63**:1166–1176.

40. * Liu Z, Zubatiuk T, Roitberg A, Isayev O: **Auto3d: automatic generation of the low-energy 3d structures with ANI neural network potentials**. *J Chem Inf Model* 2022, **62**:5373–5382.

The first use of neural network potentials is to accelerate conformer generation for virtual screening.

41. Lipinski CA: **Lead-and drug-like compounds: the rule-of-five revolution**. *Drug Discov Today Technol* 2004, **1**:337–341.

42. Veber DF, Johnson SR, Cheng H-Y, Smith BR, Ward KW, Kopple KD: **Molecular properties that influence the oral bioavailability of drug candidates**. *J Med Chem* 2002, **45**:2615–2623.

43. Ackloo S, Al-awar R, Amaro RE, Arrowsmith CH, Azevedo H, Batey RA, Bengio Y, Betz UAK, Bologa CG, Chodera JD, *et al.*: **CACHE (Critical Assessment of Computational Hit-finding Experiments): a public–private partnership benchmarking initiative to enable the development of computational methods for hit-finding**. *Nat Rev Chem* 2022, **6**:287–295.

44. Loeffler HH, He J, Tibo A, Janet JP, Voronov A, Mervin LH, Engkvist O: **Reinvent 4: modern AI–driven generative molecule design**. *J Cheminf* 2024, **16**:20, https://doi.org/10.1186/s13321-024-00812-5.