ANYO LABS

White paper

Bringing virtual screening to new levels:
The benchmarks behind ANYO Labs' i-TripleD's novel scoring method.

This whitepaper introduces i-TripleD developed by ANYO Labs, harnessing machine learning and proprietary technologies. Addressing challenges in traditional drug discovery and surpassing current AI-based approaches for virtual screening, these tools prioritize efficiency and accuracy for generating and retrieving biologically active small organic molecules. The accuracy is highlighted through a look at the benchmarks CSAR NRC-HiQ Set1, and CSAR NRC-HiQ Set2 and CASF-2016. Engineered to overcome longstanding bottlenecks in the drug discovery process, they mark a significant advancement in the field.

Executive summary

- The largest High-Throughput Screening (HTS) facilities can manage compounds in the single digit millions, limiting them to both known structures as well as iterative optimizations downstream while being time and cost intensive [1]. More accurate *in silico* methods could eliminate the need for such resource-intense screenings.

- **i-Triple D** is a tool that builds on a unique patent pending method that revolutionizes traditional drug discovery through state-of-the-art AI-based solutions. The understanding of the chemical composition allows for speeds of around 9000 screenings per second, which equates to screening 777 600 000 compounds in one day on a single A100 GPU compute node.

- The performance of our AI tools is predicated on the achieved accuracy, benchmarked using the external test sets CASF and CSAR, highlighting the Scoring, Ranking, and Screening Power, of which ANYO Labs achieved outstanding results compared to a wide range of traditional and AI-based scoring functions.

- In an *in vitro* case study, hit-identification was successful in a fraction of the time and costs when compared to the industry average, with over 39% of the compounds showing inhibitory effect and 6% were identifiable hits with single-digit micromolar potency.

Background

The opportunity in small molecular drug discovery is immense despite increasing new therapeutic modalities [2]. Due to specificity and target-based sensitivity, AI-based approaches have the potential to accelerate the drug development time to within 6-8 years as opposed to the average 10-15 years [3]. In conventional drug discovery, three identifiable bottlenecks impede *in silico* virtual screenings (VS). The first relates to the need for a full picture of atomic contacts and interactions between the receptor and drug molecule through 3D conformational sampling. 3D conformational sampling is very time consuming, computationally expensive, and inefficient. The second bottleneck concerns the lack of fast and accurate scoring functions that can identify and rank promising candidates without the need for 3D conformational sampling. I contrast to the traditional scoring functions, the advanced AI-derived scoring functions (such as Vina XGB [4] and KDeep [5]) have successfully tackled the lack of accuracy in predicting the experimental binding affinities. Nonetheless, all scoring functions developed so far require the expensive and inefficient conformational sampling. The third is related to the size and diversity of the molecular libraries available for VS. Whereas the available drug-like chemical space is estimated to contain up to ~10^{63} unique individual compounds, the largest databases available to date (*e.g.* ZINC22 [6] and similar) only include up to a few billion entities, many of which are furthermore of high structural similarity, making it a challenge for medicinal and organic chemists in selecting, designing, and synthesizing novel molecular structures [7]. Therefore, to improve the potency of current drugs or to develop novel drugs for upcoming diseases, there is inevitably a need to explore unseen regions of the drug-like chemical space in an intelligent way. Generative AI methods, such as REINVENT by AstraZeneca [8, 1, 9] have been highly successful in generating *de novo* drug molecules by sampling the drug-like chemical space, albeit still with significant room for improvement.

i-TripleD

To overcome all these hurdles, ANYO Labs has chosen a completely different strategy when developing our comprehensive AI-based drug discovery pipeline called "intelligent *de novo* drug discoverer", **i-TripleD**. It is a combination of interconnected AI-modules that have been trained on the largest and most reliable publicly available datasets. We have integrated a *de novo* drug-like molecular generator (**iGen**) in the screening workflow and thereby generate and filter compounds of relevance to the selected target on the fly. In the alternate mode, the iGen module can be switched off, allowing the introduction of external molecular libraries for screening. Unlike other structure-based VS tools, ANYO Labs has developed an AI-based scoring function (**iScore**) that bypasses 3D conformational sampling, making the process extremely less computationally expensive, allowing for computations that with other methods takes months to perform, to now be accomplished in a few hours. The computational efficiency results in screening speeds that can be applied to *de novo* drug generation; the infinitesimal task of filtering through the drug-like chemical space to find totally novel molecules that have never been seen before. All this is done with the accuracy that in the recent case studies and benchmarks outperforms all available drug discovery tools.

Technology

Our scoring function (**iScore**) has been trained against the largest manually curated dataset available (PDBBind 2020 refined set). The ML-method is based on new levels of sophistication, leading to unprecedented accuracy in scoring, ranking, and screening powers. With our novel Ultra-Fast Screening approach (**UFS**), we can furthermore screen compounds several orders of magnitude faster than any current tool. Coupled to this, all molecules are filtered against a state-of-the-art ML-based tool for ADMET predictions containing 41 different assessments/assays, with a total average prediction accuracy of >86%. An additional Synthetic feasibility module provides a good overview regarding the synthetic feasibility of the hit compounds. i-TripleD is not just another tool but a pioneering force in the field, rooted in highly data-efficient technologies that enhance VS speed while maintaining competitive precision, effectively pushing the boundaries of what is possible with today's hardware.

The discrimination of true binders from false positive decoy nonbinders still remains the major challenge for VS protocols [9, 2]. In i-TripleD, this bottleneck is solved by a newly developed filtration module (**iClass**), resulting in significantly lower number of false positives. In benchmark assessments, i-TripleD displays exceptional promise. Not only do these innovative tools exhibit unprecedented efficiency, but they also retain an unmatched level of accuracy in predictions, a testament to the potential of artificial intelligence in the field of computer aided drug discovery.

Benchmarks:

Molecular docking is undoubtedly the most widely used technique in structure-based computer aided drug discovery, that aims to predict the binding mode and binding affinity of small organic molecules toward a target protein. The performance (speed and accuracy) of a molecular docking program strongly depends on its two main components, sampling and scoring. Sampling refers to a search algorithm that evaluates a finite number of ligand conformations within and around the binding site of a target protein to elucidate the ligand binding mode. Scoring refers to a class of computational methods, called *scoring functions,* that are formulated to predict the binding affinity of each ligand conformation within the protein binding site. The performance of a scoring function can be determined by three evaluation metrics: "*scoring power*" that indicates the degree of correlation in the predicted versus experimentally determined binding affinity values, "*ranking power*" that is the capability of the scoring function to accurately rank the order of a given set of active binders with respect to their predicted binding affinity values, toward a particular target protein, and "*screening power*" that refers to the ability of a scoring function to identify the true binder with the highest affinity against a given target protein among a set of random decoy molecules.

The scoring (in terms of Pearson correlation coefficient and RMSE metrics, Figures 1a-1d), ranking (in terms of average Spearman correlation coefficient, Figure 1e), and screening power (in terms of the success rate of identifying/placing the highest affinity compound among the top 1%, 5% and 10% ranked predictions, Figure 1f) performances of our scoring function (iScore) has been extensively benchmarked towards three gold standard test sets: PDBbind 2016 core set, CSAR NRC-HiQ Set1, and CSAR NRC-HiQ Set2 and compared with the scoring functions examined in CASF-2016.

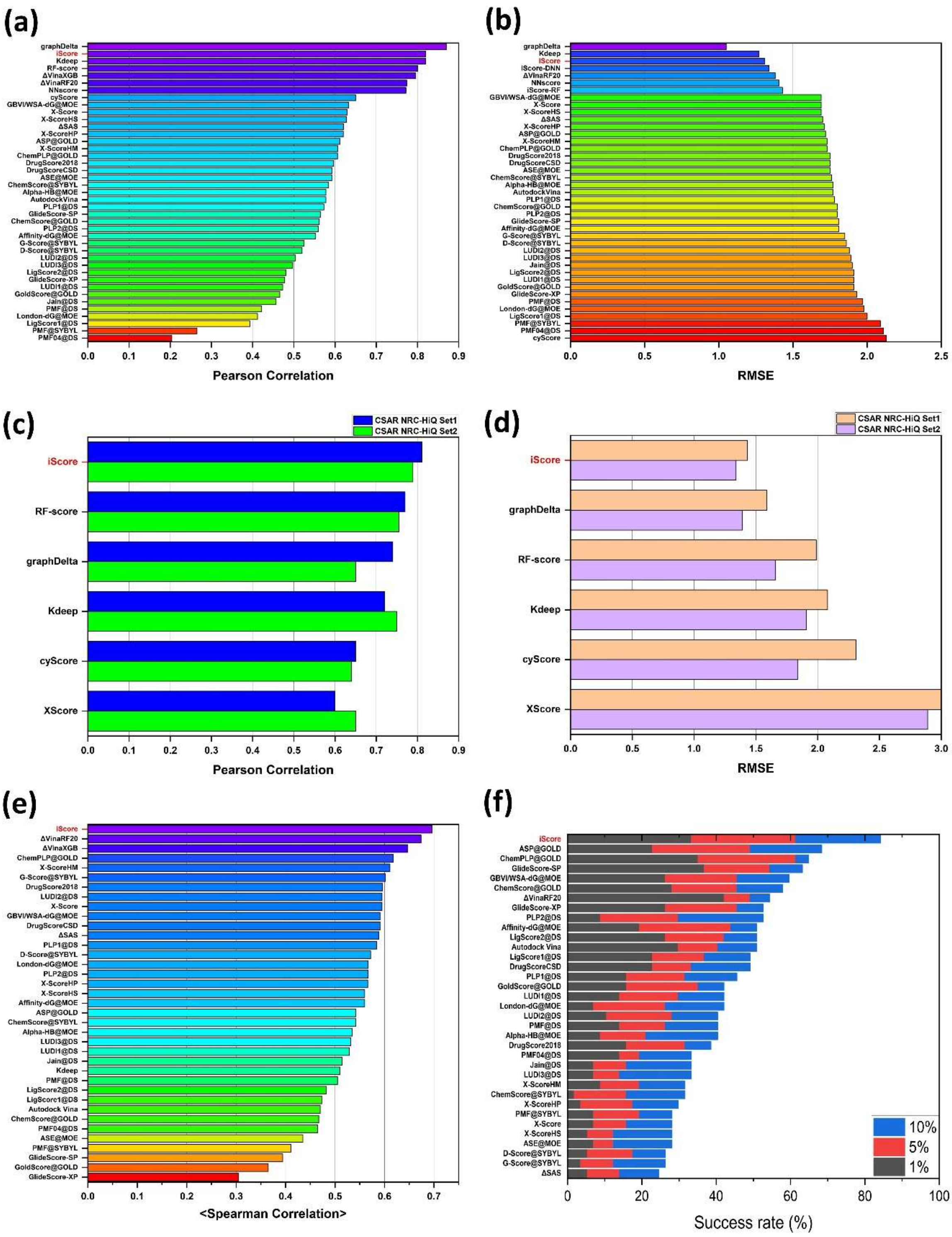
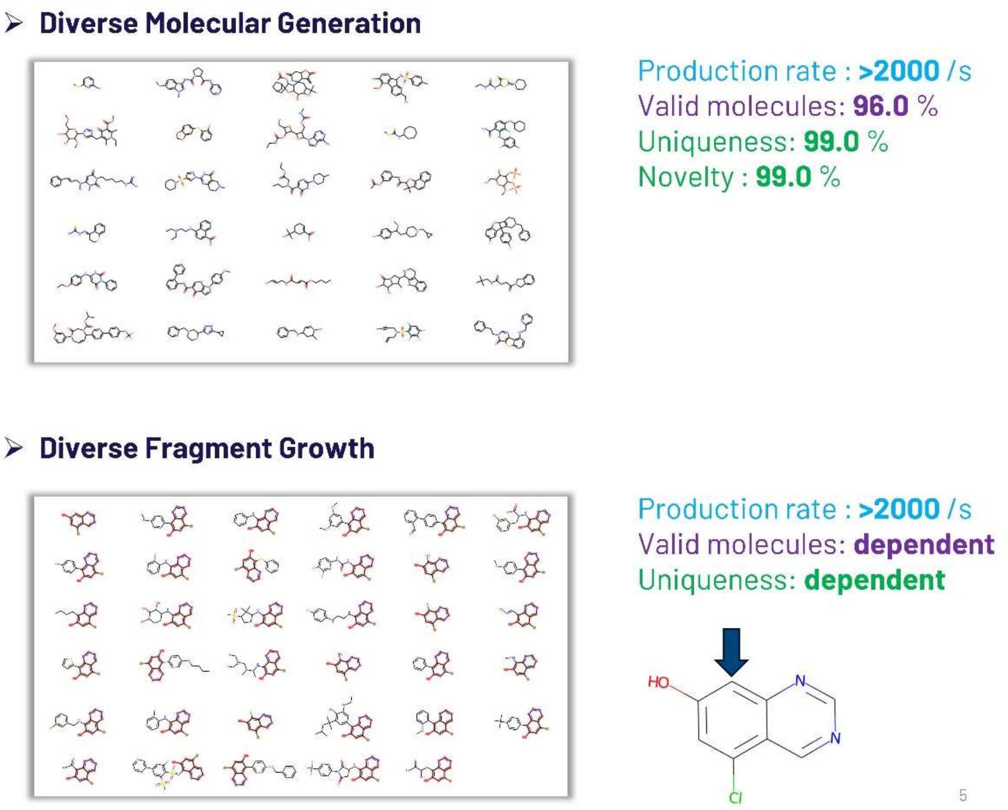


Figure 1. *The overall benchmarking performance of iScore (indicated in red) in (a-d) scoring power, (e) ranking power, and (f) screening power campaigns.*

iGen

is the AI-derived generative module of i-TripleD which generates novel drug-like molecules by sampling unseen regions of the drug-like chemical space with the production rate of >2000 molecules per second and high validity, uniqueness, and novelty (Figure 2). It is also capable of generating a big library of derivatives of a certain scaffold (Figure 2). iGen has been trained of ChEMBL-30 dataset with around 20,000,000 data points.



➢ **Diverse Molecular Generation**

Production rate : **>2000** /s
Valid molecules: **96.0** %
Uniqueness: **99.0** %
Novelty : **99.0** %

➢ **Diverse Fragment Growth**

Production rate : **>2000** /s
Valid molecules: **dependent**
Uniqueness: **dependent**

Figure 2. *iGen can be used in two different modes to either generate diverse molecules by sampling the entire drug-like chemical space or diverse derivatives for a given scaffold.*

Example Use Case:

“Three novel inhibitors of the Dengue virus protease have already been identified using ANYO Labs novel in silico AI-driven screening program that allows for rapid and cost-efficient exploration of an extremely large portion of chemical space, which cannot be matched by any other in silico drug discovery tool.”

– Prof. Johan Lennerstrand, Uppsala University.

Background:

Despite the great need, there is currently no antiviral prophylaxis/treatments available against the Dengue virus. About half of the world’s population is at risk of infection, and the endemic areas are continually growing. Viral proteases have constituted targets for the development of antivirals against HIV and HCV infections [10] as well as SARS-CoV-2 virus; as exemplified by the approval of Pfizer’s Paxlovid, which is a highly potent, first-in-line main protease inhibitor with 19 billion USD generated in 2022 [11]. Aiming to identify and optimize potent inhibitors of the Dengue virus NS2B-NS3 protease, i-TripleD was employed to identify novel hits that would then be validated using enzymatic and cell-based assays.

After the initial preparations, ANYO Labs implemented its generative AI (iGen) to generate 300,000 potential hits in a 24-hour screening (a total of ~170 million compounds were generated and screened). From these, the top ranked binders were manually chosen based on the synthetic feasibility score. To lower costs, the top predicted hits were then compared with highly similar structures available from the MolPort molecular supplier. Based on this analysis, 33 compounds were purchased and tested experimentally. 13 compounds demonstrated inhibitory effect and were tested further in dose-response measurements. These *in vitro* results confirmed the inhibitory effect of 3 compounds in micro molar range (Figure 3).

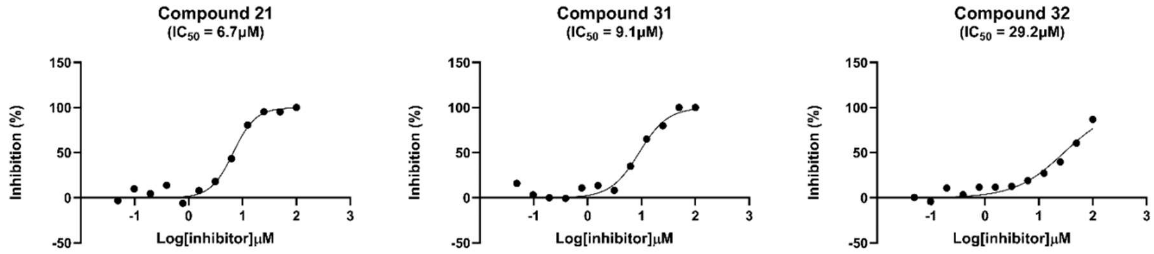


Figure 3: Dose response curves of compounds 21, 31, and 32 against Dengue virus NS2B-NS3 protease.

The preparation, screening and selection of hit compounds took under 2 weeks. The ordering of the compounds and following *in vitro* testing thereafter took an additional 5 weeks. This equates to less than two months’ time from starting the project until single digit micromolar hits had been confirmed in *in vitro* experiments. Further analog variant generation, ADME&T property predictions and original *de novo* compound synthesis is being provided for rapid generation of lead candidates.

Impact:

The efficiency and performance of i-TripleD's workflow allows for a unique value to be achieved. By providing much more accurate predicative tools, the combination provided by ANYO Labs allows for effectively identifying significantly more reliable candidates *in silico* early in the drug development process, and reduce time, cost, and resource needs. Ultimately, this enables high precision AI predictive tools at scale as well as the following unique advantages:

Uncovering new block-buster drugs

By radically increasing the speed of virtual screening, new areas of the drug-like chemical space can be explored with the help of our *de novo* generator (iGen). Shining light on molecules never seen before by the human eye enhances the creativity of chemists to find the next block-buster drug without the need for iterative scaffolding.

Reducing time & costs in early DD

Savings in the hit-identification stage by our scoring methods have been seen in practice. With the aim to use all supporting modules, the resulting ADMET filtered lead-like hits results in fewer iterations downstream in hit-to-lead as well as lead optimization stages. These stages end up being the costliest when adjusted to capitalized costs per launch [12]

Lowering the cost threshold for neglected targets

Neglected diseases tend to produce less revenue than the most profitable therapeutic areas such as oncology or immunotherapy [13]. By lowering the time and costs of early drug discovery, the ANYO Labs' workflow enables huge savings in the most cost inefficient stages of lead optimization (normally reaching 400 million dollars due to failure and iterations), and thus make neglected diseases a viable target.

The urgency of advancing drug discovery methodologies cannot be overstated. The Inflation Reduction Act (IRA), enacted in the United States, has placed substantial constraints on the profit margins that pharmaceutical companies can derive from their products [14]. This legislation has significantly intensified the economic pressures associated with drug discovery and development. In an era characterized by evolving health threats, from emerging infectious diseases to a growing spectrum of cancer types, there is an undeniable demand for new therapies that can tackle these challenges. Moreover, the cost and time required to bring a new drug to market are astronomical, with a high attrition rate at each stage of development. In this context, the ability to harness the power of machine learning and artificial intelligence to streamline the identification and optimization of potential drug candidates is a transformative opportunity. Not only can it expedite the drug discovery process, but it can also significantly reduce the financial burdens and risks associated with bringing new treatments to patients [15]. The efficiency of the tools developed by ANYO Labs have the potential to address these issues and accelerate drug discovery.

Conclusion:

ANYO Labs stands at the forefront of a paradigm shift in drug discovery, strategically harnessing state-of-the-art machine learning technologies to overcome longstanding challenges in the field. The conventional narrative of drug discovery, entangled in laborious and time-consuming processes, encounters a formidable disruptor in ANYO Labs' innovative tool suite.

Traditional high-throughput screening (HTS) facilities, grappling with the management of compounds in the single-digit millions, confront obsolescence as i-TripleD emerges as a powerhouse, seamlessly screening an unprecedented 777,600,000 molecules within a single day on a single A100 GPU compute node (173,000,000 *de novo* molecules generated by iGen module). The key to ANYO Labs' ingenuity resides in i-TripleD, an intelligent *de novo* drug discovery tool engineered to bypass the computational intricacies of 3D conformational sampling inherent in traditional methods.

Benchmark assessments meticulously illustrate i-TripleD's ability to amplify virtual screening speed without compromising precision, showcasing the transformative potential of artificial intelligence and machine learning in drug discovery. While further projects are being conducted for a full scale process of lead optimization, these tools have undergone *in silico* and *in vitro* testing for hit identification, demonstrating readiness for integration into the scientific community.

Works Cited

[1] M. Butkiewicz, Y. Wang, S. Bryant, E. J. Lowe, W. D. and M. J, "High-Throughput Screening Assay Datasets from the PubChem Database," *Chem. Inform. (Wilmington Del.),* 2017.

[2] L. Howes, "Why small-molecule drug discovery is having a moment," 2023. [Online]. Available: https://cen.acs.org/pharmaceuticals/drug-discovery/small-molecule-drug-discovery-having/101/i36.

[3] M. Lohani, M. Chaturvedi, M. Negi and S. Yadav, "Artificial Intelligence: A New Trend in Drug Designing Against SARS Cov-2.," *International Journal of Molecular Sciences,* p. (6):3261, 2022.

[4] J. Lu, X. Hou, C. Wang and Y. Zhang, "Incorporating explicit water molecules and ligand conformation stability in machine-learning scoring functions.," *Journal of chemical information and modeling,* pp. 59(11), 4540-4549, 2019.

[5] J. Jiménez, M. Skalic, G. Martinez-Rosell and G. De Fabritiis, "K deep: protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks," *Journal of chemical information and modeling,* pp. 58(2), 287-296., 2018.

[6] B. I. Tingle, K. G. Tang, M. Castanon, J. J. Gutierrez, M. Khurelbaatar, C. Dandarchuluun and J. J. Irwin, "ZINC-22— A free multi-billion-scale database of tangible compounds for ligand discovery.," *Journal of Chemical Information and Modeling,* pp. 63(4), 1166-1176, 2023.

[7] V. Mouchlis, A. Afantitis, A. Serra, M. Fratello, A. Papadiamantis, V. Aidinis, I. Lynch, D. Greco and G. Melagraki, "Advances in de Novo Drug Design: From Conventional to Machine Learning Methods.," *Int J Mol Sci.,* p. 7;22(4):1676, 2021.

[8] T. Blaschke, J. Arús-Pous, H. Chen, C. Margreitter, C. Tyrchan, O. Engkvist and A. & Patronov, "REINVENT 2.0: an AI tool for de novo drug design.," *Journal of chemical information and modeling,* pp. 60(12), 5918-5922., 2020.

[9] A. Graves, R. Brenk and B. Shoichet, "Decoys for docking.," *Journal of medicinal chemistry,* vol. 48(11), pp. pp.3714-3728., 2005.

[10] T. Majerová and J. Konvalinka, "Viral proteases as therapeutic targets," *Molecular Aspects of Medicine,* no. 88:101159, 2022.

[11] Pfizer, "PFIZER REPORTS RECORD FULL-YEAR 2022 RESULTS AND PROVIDES FULL-YEAR 2023 FINANCIAL GUIDANCE," Pfizer, 2023.

[12] S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg and A. L. Schacht, "How to improve R&D productivity: the pharmaceutical industry's grand challenge.," *Nature reviews Drug discovery,* pp. 203-214.´, 2010.

[13] M. Mikulic, "Leading 10 therapeutic areas worldwide by sales in 2019," 2021. [Online]. Available: https://www.statista.com/statistics/407971/projected-revenue-of-top-therapeutic-areas-worldwide/.

[14] H. R. Bill 5376, *To provide for reconciliation pursuant to title II of S. Con. Res. 14,* Washington DC, 2022.

[15] Biostock, "Inflation Reduction Act and AI set the tone for biotech," 4 10 2023. [Online]. Available: https://www.biostock.se/en/2023/10/inflation-reduction-act-and-ai-set-the-tone-for-biotech/.