Topic: Structural Similarity and Drug Candidate Selection
Submitted by: Avirup Guha Neogi
Bonn-Aachen International Center for Information Technology, Bonn
Course: Molecular Modelling and Drug design Lab
Supervisor: Dr. Dagmar Stumpfe
Course Coordinator: Professor Dr. Jürgen Bajorath

Introduction

Discovery of potential drug molecules has been one of the popular and hot topics in the field of biological sciences for decades. Traditionally, it involved disciplines like Chemistry, biology and pharmacology. However, with the advent of the informatics era, new computational approaches are being developed which further gained momentum with the introduction of Artificial Intelligence and Quantum computation. Along with the wet lab legacy techniques, modern science offers a myriad of in-silico approaches for molecular modelling and subsequent identification and designing of drug molecules. This involves rigorous statistical data analysis to gain insights into the various structural and functional aspects of chemical objects that are represented in computers in convenient formats which are then processed by leveraging modern intelligent algorithms.

In this report, the potential drug candidate for Elastase dataset [1] is identified using MOEsaic application of the MOE (Molecular Operating Environment). MOE is a computer-aided molecular design software which is developed by the Chemical Computing Group [2]. It bundles a wide range of products that integrates modelling and simulations and visualizations along with various other design tools. At the very outset, the dataset is analysed to reveal its potencies against the targets and anti-target proteins which would be further considered to evaluate the compounds' specificity, toxicity, or stability. Subsequently, the Structure Activity Relationship (SAR) analysis is performed to understand the correlation among the candidate molecules and account for selectivity profile or the polypharmacological behaviour. The SAR analysis also involves exploring various similarity methods like MMPs (Matched Molecular Pairs), substructure search, similarity search, etc. Finally, one drug candidate is chosen based on the selectivity analysis and similar compounds with similarity search is obtained.

Analysis of the Dataset

There are overall 42 compounds present in the Elastase dataset. The compounds are tested against two biological targets namely, Human Neutrophil Elastase (abbreviated as HNE), Rat Neutrophil Elastase (abbreviated as RNE). Here, HNE is observed to be the primary target and RNE is observed to be the secondary target. Two anti-targets were also identified against whom these compounds are also tested. These are Cytochrome P450 3A4 (abbreviated as CYP3A4) and Cytochrome P450 2C9 (abbreviated as CYP2C9).

HNE is found in humans that belong to the Neutrophil Elastase protein family which is encoded by the ELANE gene that resides on chromosome 19. It is an important regulator of the immune response and plays a significant role in host defence mechanisms and further physiological processes [1]. The uncontrolled activity of this serine protease may cause severe tissue alterations and impair inflammatory states [1]. RNE also belongs to the same

protein family, the only difference being that in this case it is found in a different organism i.e. rats. CYP3A4 and CYP2C9 are well-known anti-targets that belong to the CYP40 family that can cause undesirable side effects.

The following table summarizes some of the basic counts that are observed in the given dataset.

Compound Characteristics	Count
Number of compounds with annotations for all four targets	10
Number of compounds with only HNE annotation	9
Number of compounds with only RNE annotation	0
Number of compounds with HNE annotation	42
Number of compounds with RNE annotation	12
Maximum Potency for HNE	10.62
Minimum Potency for HNE	7.13
Maximum Potency for RNE	7.82
Minimum Potency for RNE	5.99
Maximum Potency for CYP3A4	8.40
Minimum Potency for CYP3A4	7.30
Maximum Potency for CYP2C9	9.30
Minimum Potency for CYP2C9	7.30

Table 1. Statistics of the dataset: This table shows some of the relevant characteristics of the Elastase dataset.

In the domain of pharmacology, the potency values [3] are expressed as either IC_{50} (Half maximal inhibitory concentration), EC_{50} (Half maximal effective Concentration), Ki (Inhibitor constant) or Kd (Dissociation constant) values. IC_{50} is a measure of the potency of a substance in inhibiting a specific biological or biochemical function. It refers to the amount of the inhibitory substance required to inhibit a given biological process by half. On a parallel note, EC_{50} is defined as the amount of agonist or excitatory drugs required to trigger a response halfway between baseline and maximum value after interaction with the target. For the sake of simplicity, the discussion about the Ki and Kd values are skipped in this report.

In this dataset, the potency values are represented as IC_{50} values, more specifically as pIC_{50}. pIC_{50} takes the negative logarithm of IC_{50}. This gives the advantage to interpret the values more conveniently. The more the value of pIC_{50}, the more potent the compound is against the target.

Fig 1. shows the distribution of the potency values for the targets as well as the anti-targets.

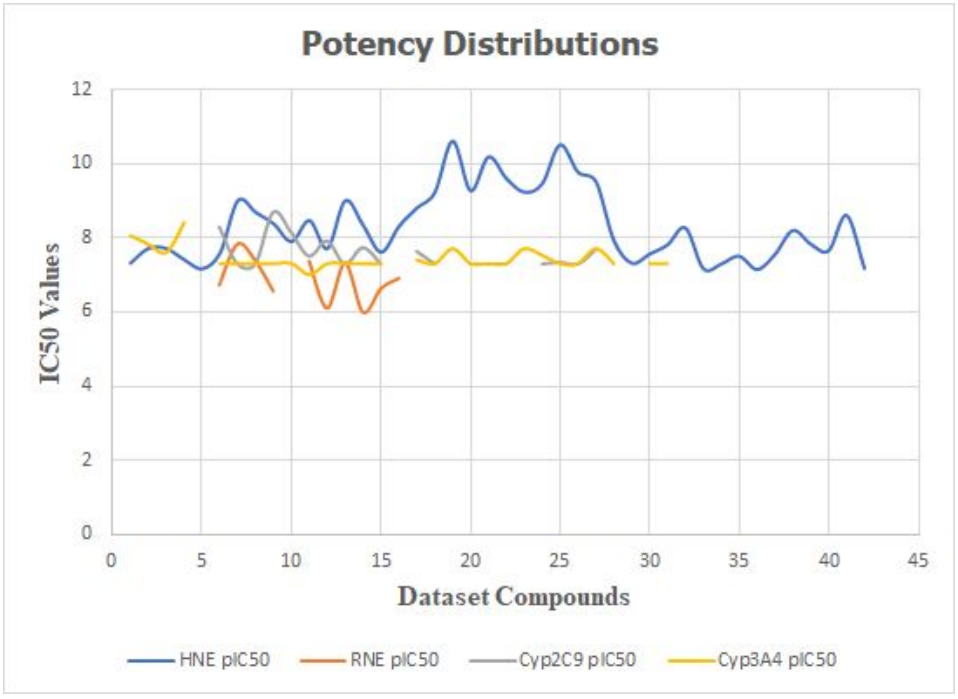

Fig 1. Potency value distributions: The potency value distributions for both the targets and both the anti-targets are given in the above line graph. It is observed that all the compounds are tested against the primary target and for very few compounds the potency values are available for secondary targets.

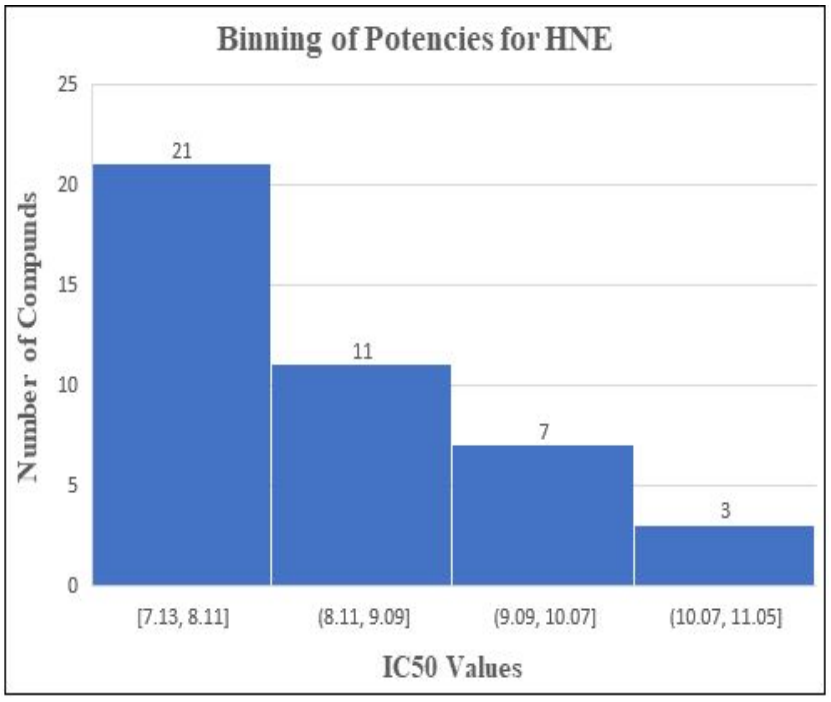

Fig 2. Binning of potencies for HNE: The histogram depicts the ranges of the potency values against the primary target i.e. HNE. The binning is done only for the HNE since the potential drug candidate is to be chosen for HNE as it is the primary drug target.

Hence, it is observed that the potency distribution is fairly uniform for 50% of the data values for HNE with nearly 20% of the compounds being highly potent against it. Only 12 compounds are tested against RNE and the distribution is fairly uniform with a low average potency value of 6.75. A total of 31 compounds have been tested against CYP3A4 with an average medium high potency of 7.4, having a good uniform distribution close to the mean. Finally, for CYP2C9 there are 24 targets which are tested and the distribution is again fairly uniform with a medium high average potency of 7.6.

Additionally, there are some core properties of the compounds [4] which are part of the dataset. They are *molecular weight* (MW), *hydrogen bond acceptor* (HBA), *hydrogen bond donor* (HBD), *hydrophobicity* (h_logD), enhanced atomic partition coefficient (SlogP) for *lipophilicity* and *topological polar surface area* (tPSA). Each of these terms are explained in the following. Molecular weight is simply the molecular mass of the chemical compound. Generally measured in daltons(Da). Molecular masses would differ for different isotopes of the same compound. The atom, ion or molecule component of a hydrogen bond which does not supply the bridging hydrogen atom is known as HBA. On the contrary, the bond or molecule that supplies the hydrogen atom of a hydrogen bond is known as HBD. H_logD is a log of partitions of a chemical compound between the lipid and aqueous phases (octanol in general). This measure is used to account for hydrophobicity. SLogP is the partition coefficient of a molecule between an aqueous and lipophilic phase (Usually octanol and water). It is used to account for lipophilicity of the compound. TPSA of a compound is defined as the surface sum over all polar atoms or molecules, mainly oxygen and nitrogen including their attached hydrogen atoms. This accounts for the drug's permeability though the cells.

Based on the properties of the compounds that are discussed in the above section, there is rule to evaluate the drug likeness of the chemical compound. This rule is popularly known as the *Lipinski's rule of five* [5]. These rules help choose probable drug candidates based on its defined criteria. It is just a rule of thumb and doesn't require to strictly satisfy all the defined criteria under the rule. But it does give a promising drug like biologically active candidates which are likely to have the chemical and physical properties to be orally bioavailable. The rules are as follows:

i) No more than 5 hydrogen bond donors
ii) No more than 10 hydrogen bond acceptors
iii) Molecular mass less than 500 Da
iv) Partition coefficient not greater than 5

If any two conditions are violated, it is predicted that the candidate is a non-orally available drug.

Exploring similarity methods in the dataset

Molecular and chemical similarity are being studied extensively in the domain of Chemoinformatics and Medical Chemistry. Similarity assessment is one of the main exercises which is performed to predict similar biological activities of a target against similar compounds. The idea stems from the *Similarity Property Principle (SPP)* [6] which states that similar compounds should have similar properties and the most significant property being biological activity. This gives rise to a concept which is known as *Structure Activity Relationship (SAR)*. SAR can be formally defined as the relationship between the chemical structure of a molecule with its biological activity [7]. SAR is a key concept in drug discovery which is applied in various aspects of it ranging from virtual screening to lead optimization. There are mainly following types of SARs viz., Continuous SAR, Discontinuous SAR, heterogeneous-relaxed, and heterogeneous-constrained SARs [8]. However, in this analysis, only Continuous SARs and Discontinuous SARs would be introduced. *Continuous SARs* are observed when a small change in the structures of analogues produces very little or no change in their activity profiles. On the other hand, when a small change in the structures among the analogues produces a significant change in the activity of one or more analogues in the similarity set, a *discontinuous SAR* is observed.

There are instances where certain compounds which might not be considered similar exhibit similar activity profiles. This type of relationship is known as *Horizontal Compound Relationship* [9]. In contrast, compounds that should be considered similar, differ largely in their activity profiles. This type of relationship is known as *Vertical Compound Relationship* [9]. This gives rise to the concept of *Activity Cliffs (AC)* which are nothing but structurally similar active compounds having large potency differences. In other words, ACs are the extreme form of SAR discontinuity [10].

Many times, the terms ‘Molecular Similarity’ and ‘Chemical Similarity’ are used interchangeably. However, there are differences in their assessment criteria. Chemical similarity mainly refers to the physicochemical properties of the compounds like molecular weight, solubility, boiling point, log P, etc. However, molecular similarity primarily focuses on the structural features of the compounds like the shared substructure, topologies, ring systems, etc.

In the molecular graph-based similarity approach, there are various techniques which are employed to compute the same. All these fall under the category of 2D similarity assessment. Broadly they can be classified as *Fingerprint-Based* similarity assessment and *Substructure-Based* similarity assessment. In the fingerprint-based approach, the AC is encoded as a bit string of chemical structure and properties and Tanimoto Coefficient (Tc) is used to assess the similarity criterion. But this approach is subjective in nature and different fingerprints would produce different Tc values. Also, since Tc values are calculated as whole-molecule similarity measures, it is difficult to interpret it from a chemical perspective. Alternatively, a second-generation similarity method which is based on shared substructure is developed. *Matched Molecular Pair* (MMP) [11] is one such popular similarity criteria to determine molecular similarity. An MMP is a pair of compounds that differ only at a single

site. Matched Molecular Series (MMS) is an extension of MMP that comprises a set of analogous compounds differing only at a single site. In a recent development, a computational approach is proposed to systematically identify analogue series with single and multiple substitution sites in compound datasets. This novel approach is termed as *Compound-Core Relationship* (CCR). This method elevates the similarity assessment techniques to a next level and hence known as a third-generation method. ACs can also be extended to be defined as 3D structures and hence give rise to various approaches to assess similarity at 3D.

MMP Search:

The following section explores the MMPs and MMS in the dataset.

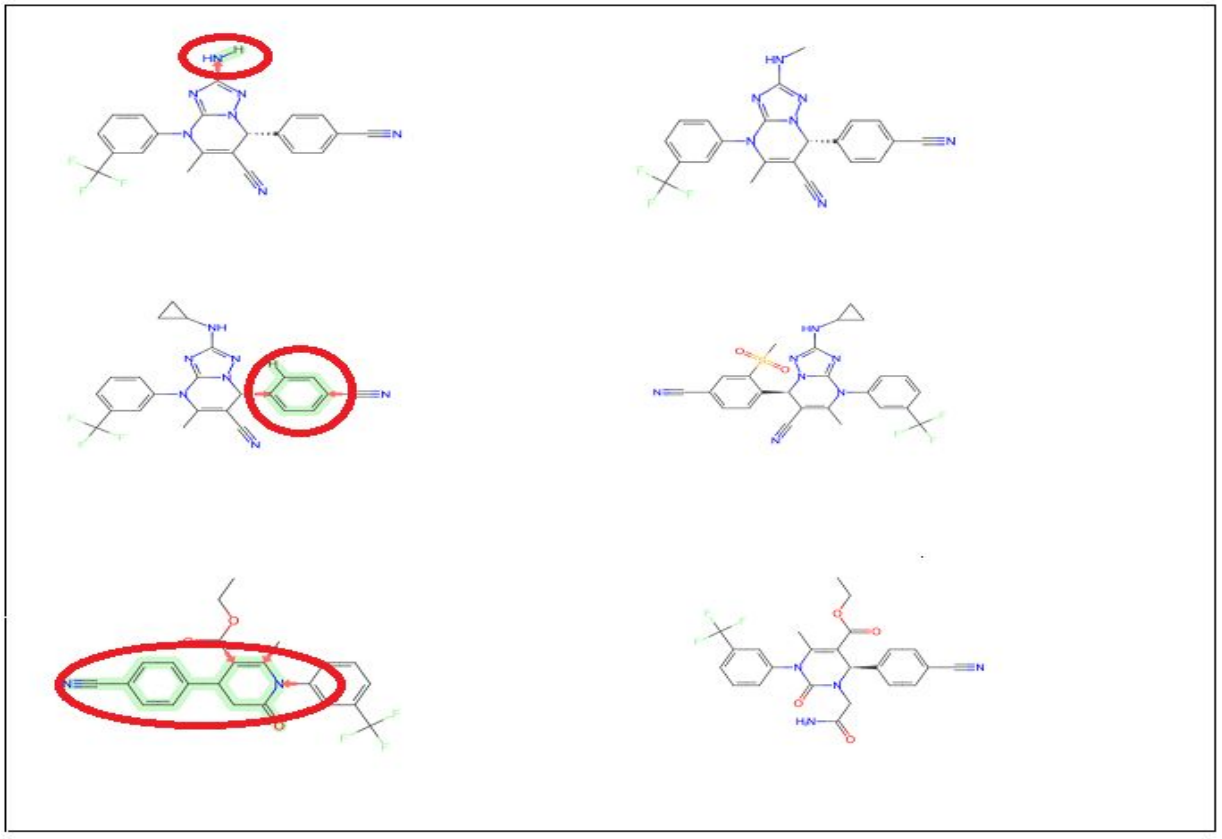

Fig 3. Single Cut: The green part of the left picture of the first row shows the variable part and the one red arrow shows the point of cut. Double Cut: The green part of the left picture of the second row shows the variable part and the two red arrows show the points of cuts. Triple Cut: The green part of the left picture shows the variable part and the three red arrows show the points of cuts.

In a single cut MMP, there should be one bond at which the structures must be cut so that it becomes equal to the invariant part of the core substructure. In a double cut MMP, there should be two bonds at which the structures must be cut so that it becomes equal to the invariant part of the core substructure. In a triple cut MMP, there should be three bonds at which the structures must be cut so that it becomes equal to the invariant part of the core substructure.

It is quite clear that the single cut MMP is chemically intuitive. The variable part also plays a major role for the derivation of a good quality MMP. The larger the variable part of the structure, the larger is the number of matches between the pairs but the lower is the quality of the MMP. The bigger the shared substructure, the higher is the similarity of the two compounds forming the MMP. It should be noted that the invariant part should be large enough to be counted as a valid invariant for an MMP.

In the dataset, the following MMS is found where discontinuous SAR containing AC is observed. The AC is considered when in an MMS or an MMP or in any other similarity, there is an approximately 100-fold or more difference in potency values between the compounds. From the drug designing perspective, discontinuous SAR is of great importance.

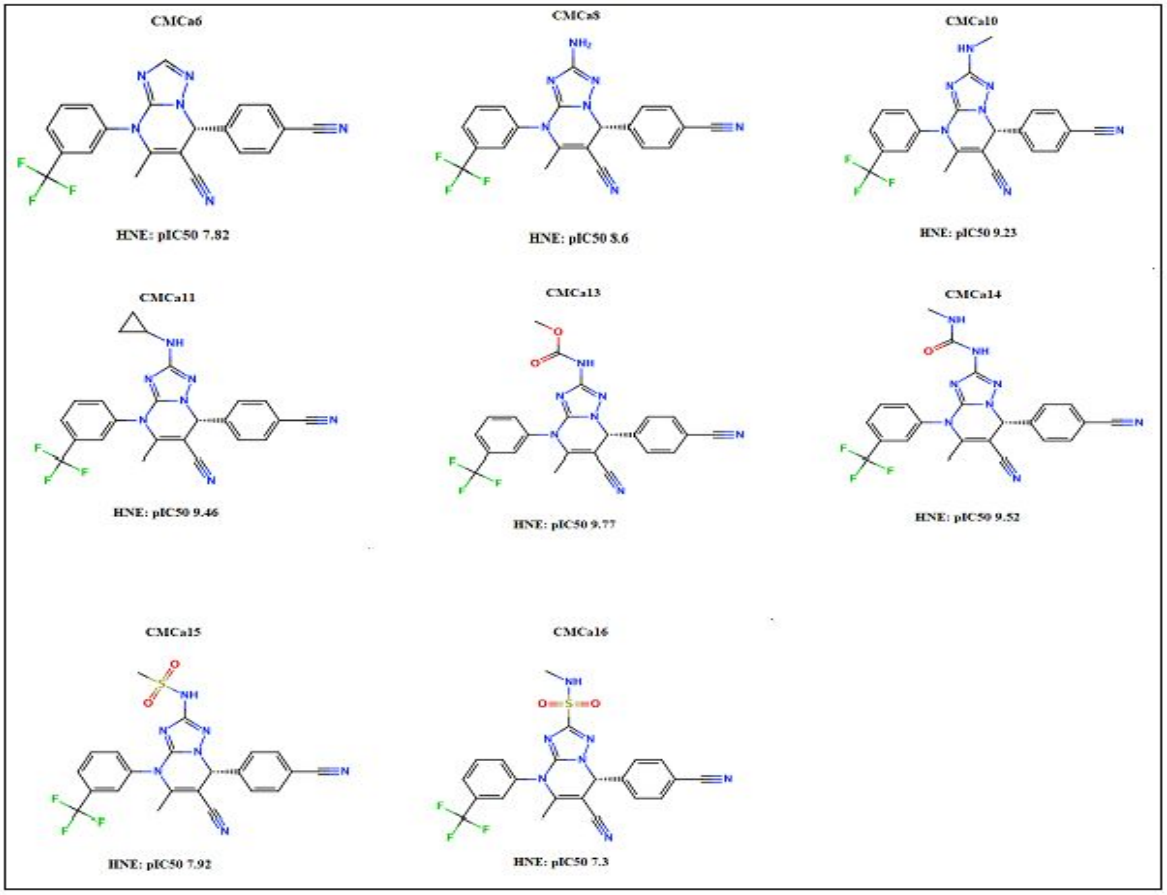

Fig 4. MMS showing discontinuous SAR: The figure shows an activity cliff in the discontinuous SAR. There is an AC between the set of compounds { CMCa6, CMCa8, CMCa15, CMCa16} and {CMCa10, CMCa11, CMCa13, CMCa14}.

In the dataset, there are other examples of MMS which are observed. In case of continuous SAR, the relative difference among the potencies of the datasets lie within a similar range.

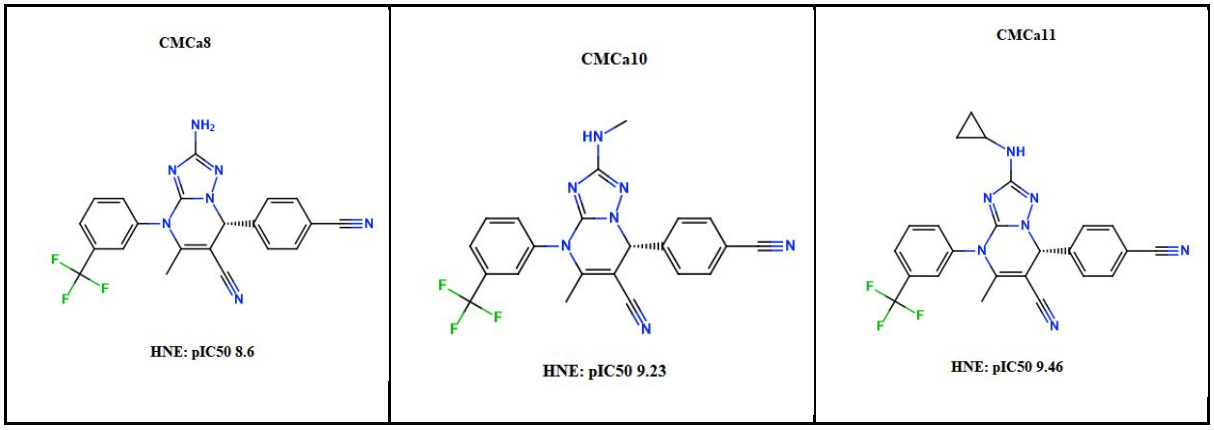

Fig 5. MMS shows continuous SAR among the compounds CMCa8, CMCa10 and CMCa11. The potency difference in this case is not more than 10 fold between any pair.

Substructure Search (SSS):

A substructure search in a set of chemicals is employed when the requirement is getting all the available starting materials containing a certain structure fragment (the substructure). Following are the working principles of the substructure search:

- The substructure query is defined by the user.
- The substructure query is superimposed with each other available structures in the library.
- The superimposition of the substructure query and available starting materials consider the overlap of atoms and bonds.
- A hit occurs when there is a total overlap among all atoms and bonds of the substructure query with a set of atoms and bonds of the compound from the catalog of chemicals.

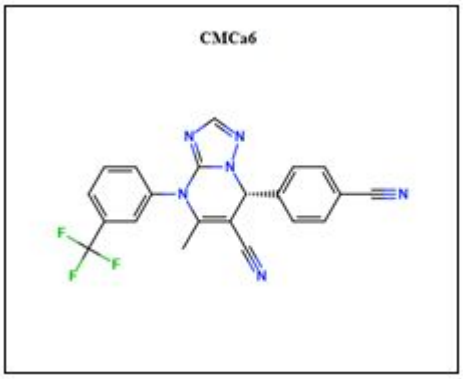

Fig 6. The selected molecule against whom the substructure search is run.

The following results are obtained for the SSS. The result set contains CMCa6, CMCa8, CMCa9, CMCa10, CMCa11, CMCa12, CMCa13, CMCa14, CMCa15 and CMCa16. Hence, it is observed that in case of SSS, the number of hits are two more than that of its corresponding MMP search matches because of the working principles of SSS mentioned above.

Similarity Search (SIM) :

The similarity search is a fingerprint-based technique that involves searching structures similar to the selected scaffold. It plays a crucial role during the design phase of a synthesis for a single target structure. A similarity search considers only the largest fragment which results from the application of a similarity criterion to the compound from the catalog of chemicals. The similarity search for the compound CMCa6 yields five similar compounds, namely, CMCa6, CMCa7, CMCa8, CMCa10 and CMCa11. Another similarity search is run for the final drug candidate which is described later in this report.

Choosing the Promising Drug Candidate

At the very outset of the drug candidate selection process, it is essential to follow one of the two popular approaches in drug design viz., the *polypharmacological* approach [12] or the *selectivity profile* analysis approach [13]. The choice of the approach is best determined by the initial data analysis of the compound set.

Polypharmacology is one of the emerging trends in the drug discovery spectrum which aims to make a paradigm shift in the philosophy of the traditional drug designing approach of "one drug one target" to "one drug multiple targets" by analysing the whole disease picture holistically. Hence, it involves drug acting on multiple targets in a disease pathway which can also aid in drug repurposing. Complex diseases involve complex therapeutic interventions and hence, creates the necessity for a single drug to act upon multiple targets thereby creating a more efficient next-generation rational drug design approach.

Selectivity on the other hand, follows a completely different approach which upholds the popular idea of chemical biology being a reductionist discipline that relies on chemical probes which are highly selective in nature and affect only one particular target. In this context, whenever there are significantly large differences in potencies between one or two targets of a pair, a *selectivity cliff* [14] is formed. Compounds forming a selectivity cliff may or may not form an AC.

In the Elastase dataset, at first the possibility of applying the polypharmacology approach is determined. It is observed that only for 10 compounds out of a total of 42, it has multiple annotations for all the targets including the anti-targets. The following table enumerates the list of such compounds along with its corresponding potency values for all the targets.

ID	HNE	RNE	Cyp2C9	Cyp3A4
CMC14	7.57	6.72	8.30	7.30
CMC15	9.00	7.82	7.30	7.30
CMC16	8.68	7.37	7.30	7.30
CMC17	8.37	6.56	8.70	7.30
CMC19	8.46	7.36	7.51	7.00
CMC20	7.70	6.10	7.92	7.30
CMC21	9.00	7.31	7.30	7.30
CMC22	8.33	5.99	7.74	7.30
CMC23	7.60	6.63	7.30	7.30
CMCa24	7.55	6.25	7.30	7.30

Table 2. Potency Values: This table shows the potency values of only those compounds having multiple annotations for all the targets.

It has been observed that although the compounds CMC23 and CMCa24 may be promising drug candidates by the polypharmacological approach, they also have a fairly good potency against the anti-targets which would unnecessarily lead to the undesirable increase in toxicity and other adverse effects. Hence, this approach is not considered as a method of drug selection for this dataset. The selectivity approach is hence adopted for the drug candidate selection.

In the selectivity approach, the dataset is refined in such a way that the potency values for HNE is as high as possible along with RNE but with a 10 fold or more approximate difference between them. Also, the potencies for the anti-targets are also considered to be as low as possible. With this filtering criteria the following compounds are selected as the promising candidates.

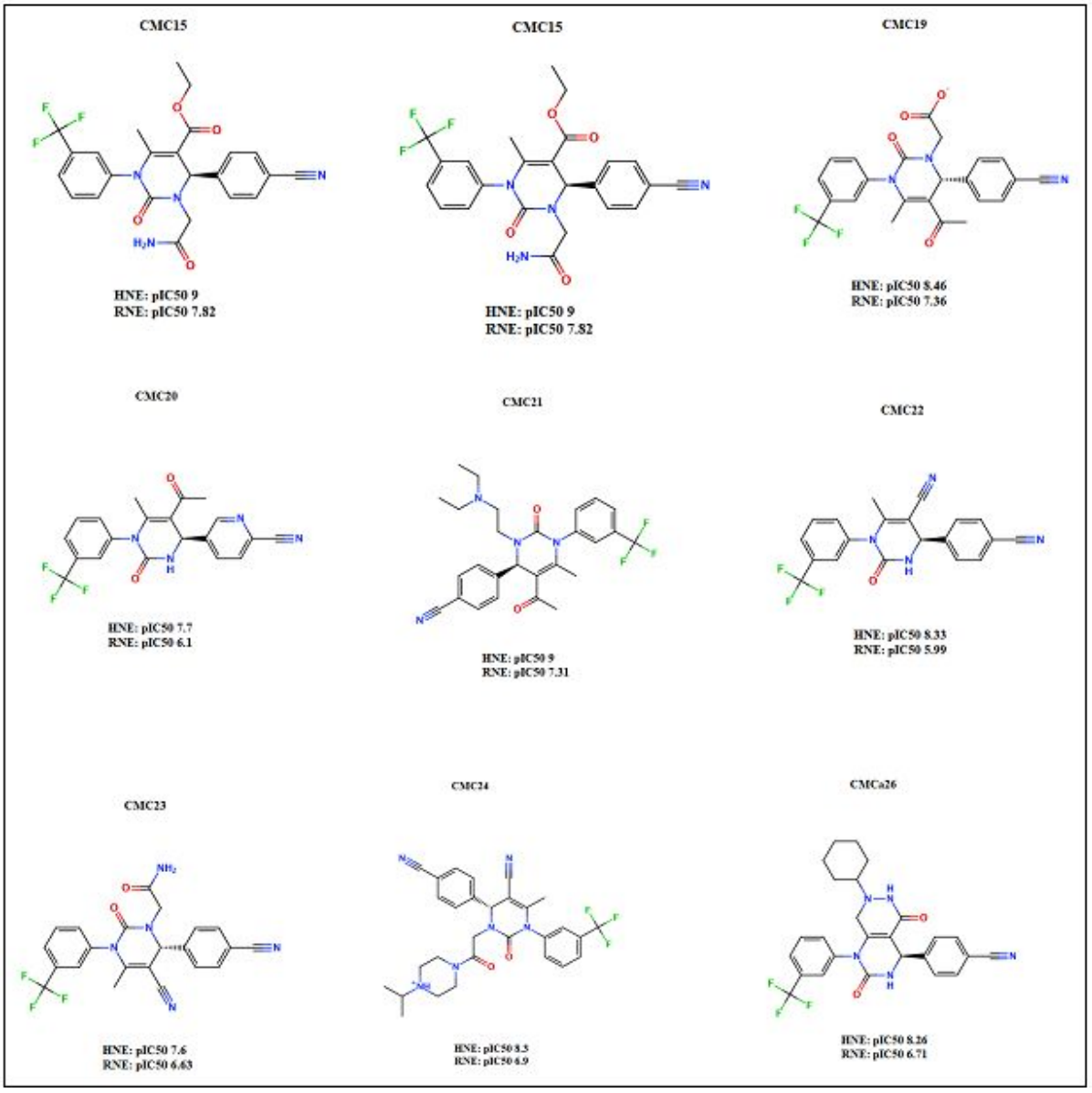

Fig. 7. Selected compounds: The compounds are selected based on the high potency values for HNE and low potency values for anti-targets.

In order to be a successful drug candidate, the compounds must satisfy the rule of five laid down by Lipinski which is discussed in the previous section. Accordingly, the filtering criteria is applied and the following refined set is obtained. The following five compounds would then be further scrutinized to select the best candidate among them.

	Properties			
Compounds	**MW**	**HBD**	**HBA**	**SlogP**
CMC15	486.45	1	4	4.290
CMC16	486.426	0	3	3.554
CMC19	456.4	0	3	3.580
CMC23	439.397	1	4	4.250
CMCa24	427.386	2	4	3.486

Table 3. Refined compound set: This set of five refined sets satisfies all the conditions of Lipinski's rule of five.

Out of the five compounds, CMC15 has the highest potency for HNE followed by CMC16, CMC19, CMC23 and CMCa24. However, the molecular weight plays an important role in the drug selection process as well along with potencies. The drug candidate should have a low molecular weight. There are compounds like CMC19, CMC23 and CMCa24 with low molecular weight. But, CMC23 and CMCa24 have very less selectivity profile. So considering all these aspects from the refined set, CMC19 is finally chosen as the potential drug candidate for the dataset.

The following paragraphs presents the CMC19 molecule using various model plots like *Oral Drugs*, *Sweet Spot* and *Golden Triangle*.

In the Oral Drug model the box plot represents the Z-score of standard deviations around the mean for the properties like MW, SlogP, tPSA, HBA and HBD. The median value is represented by the line in the box and the box covers the +25% to -25% quartiles and the line shows the minimum and maximum Z-scores.

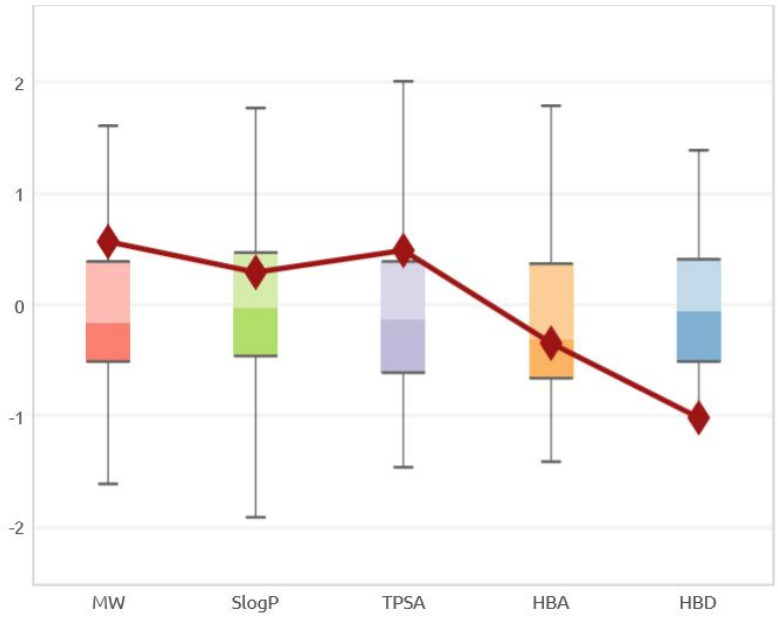

Fig. 8. The Oral Drug Model for CMC19: The X-axis represents the properties of the compounds and Y-Axis represents some statistics as box plots.

In the Sweet Spot model [15], the molecular weight is represented against logP, with property value guidelines. The radial gradient background provides a visual guide for tracking the "ideal" drug-likeness of a molecule in terms of log P and molecular weight.

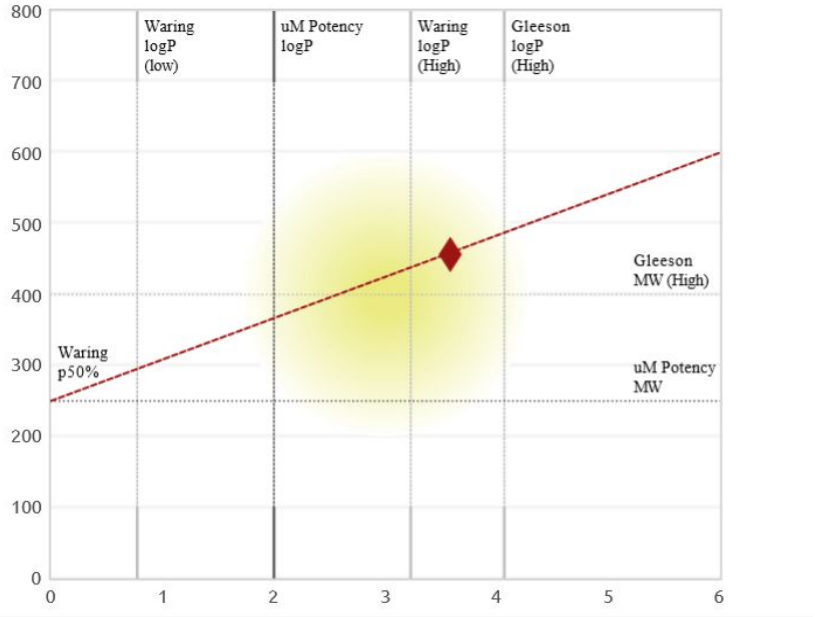

Fig. 9. Sweet Spot model for CMC19: It is observed that our candidate falls within the sweet spot.

In the Golden Triangle model [16], the molecular weight is plotted against the log D value. It helps to visually optimize the clearance and absorption of drug molecules Compounds that fall within the golden triangle are supposed to have a good balance between clearance and absorption.

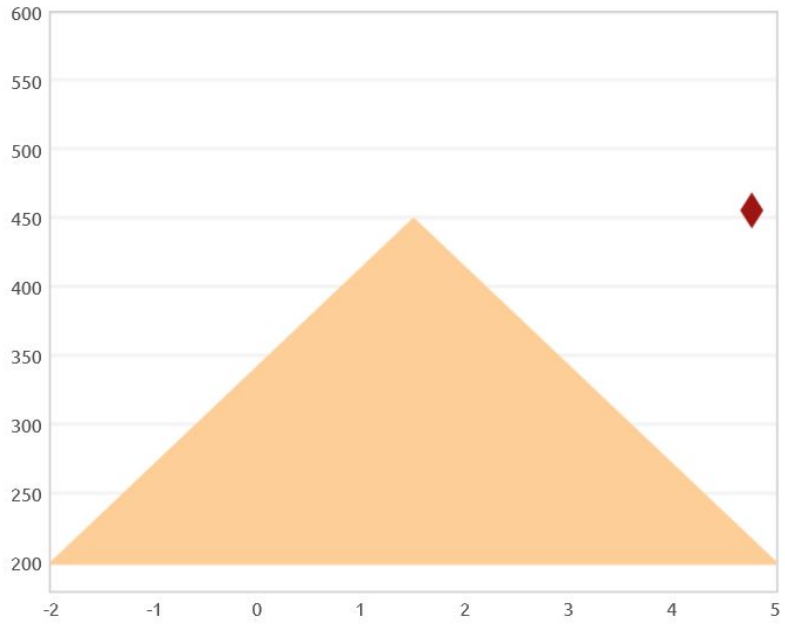

Fig. 10. The Golden triangle model for CMC19: The log D values are plotted on the X-axis and molecular weights are plotted on the y-axis. Although the chosen compound does to full fall within the triangle, however, it is visible in the graph and is close to the ideal case.

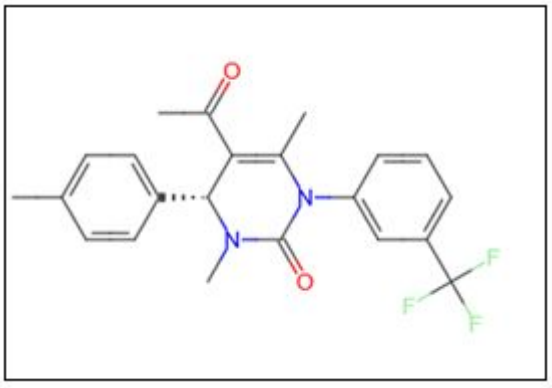

Fig. 11. The structure of CMC19 chosen to perform a similarity search with a Tanimoto similarity of 0.85.

There are four compounds which are obtained after the search is performed. They are CMC18, CMC20, CMC25 and CMC27. They are displayed below.

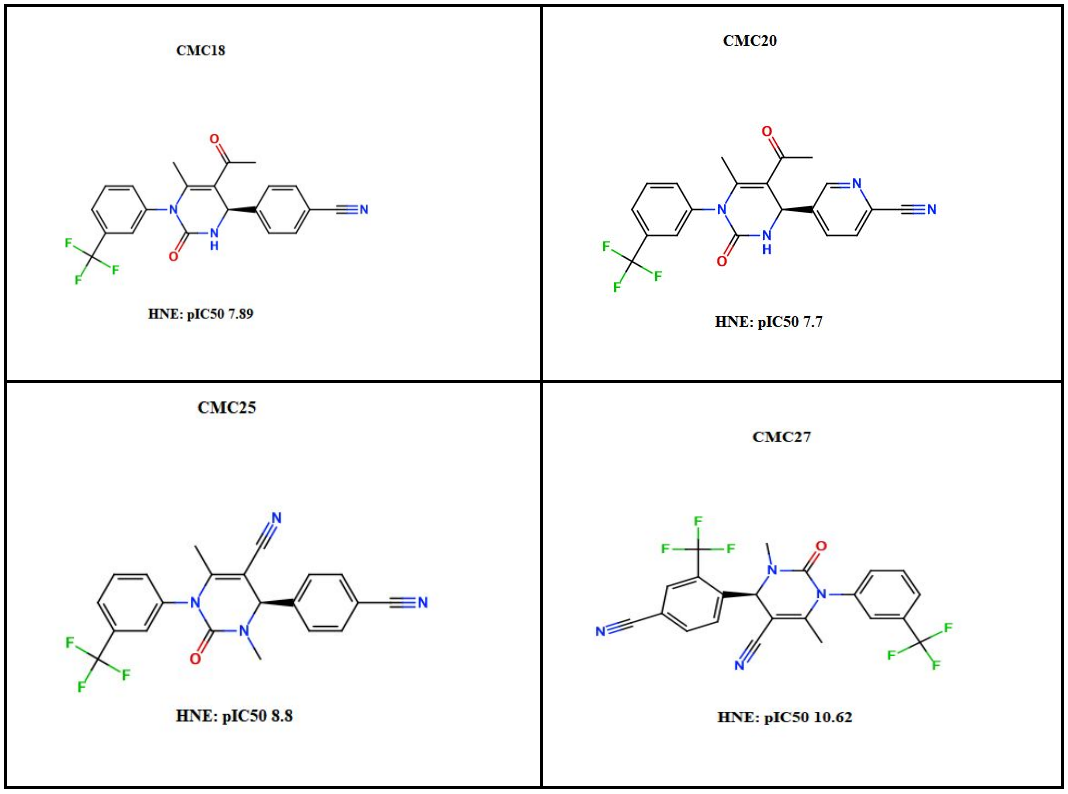

Fig. 12. The Similarity Search results of CMC19: These compounds all have the basic structure which is preserved and due to minor differences between the groups which are attached to that structure, the selectivity of the profile varies greatly.

Conclusion and Discussions

The report discussed different angles of analyzing a dataset of compounds and identifying the relationship among its primary target, off targets and anti-targets. The potency distribution reveals a great deal of information regarding the same. The dataset is then investigated for similar compounds based on various techniques like MMS, SSS and SIM. Finally, a subset of compounds are chosen exhibiting high potency for the primary target and satisfying the Lipiski's rule of five. Further consideration of other aspects like molecular weight, toxicity, absorption, etc. ultimately led to the choice of one drug candidate for the Elastase dataset. Most lead optimization projects will try and improve potency, reduce toxicity and ensure sufficient bioavailability, amongst other properties. As a future work, the development of virtual compounds and a comparative analysis with a marketed drug can be considered.

References

1. Schulz-Fincke, A. C., Tikhomirov, A. S., Braune, A., Girbl, T., Gilberg, E., Bajorath, J., ... & Gütschow, M. (2018). Design of an Activity-Based Probe for Human Neutrophil Elastase: Implementation of the Lossen Rearrangement To Induce Förster Resonance Energy Transfers. *Biochemistry*, *57*(5), 742-752.2.

2. CCGI, M. (2016). Molecular Operating Environment (MOE), 2013.08. *Chemical Computing Group Inc., Montreal*.

3. Stewart, M. J., & Watson, I. D. (1983). Standard units for expressing drug concentrations in biological fluids. *British journal of clinical pharmacology*, *16*(1), 3.

4. Mills, I. (1993). *Quantities, units and symbols in physical chemistry/prepared for publication by Ian Mills...[et al.]*. Oxford; Boston: Blackwell Science; Boca Raton, Fla.: CRC Press [distributor]

5. Lipinski, CA; Lombardo, F; Dominy, BW; Feeney, PJ (March 2001). *"Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings". Advanced Drug Delivery Reviews.*

6. Johnson, Mark A., and Gerald M. Maggiora. *Concepts and applications of molecular similarity*. Wiley, 1990.

7. Crum-Brown, A., & Fraser, T. R. (1865). The connection of chemical constitution and physiological action. *Trans R Soc Edinb*, *25*(1968-1969), 257.

.

8. Peltason, L., & Bajorath, J. (2007). SAR index: quantifying the nature of structure− activity relationships. Journal of medicinal chemistry, 50(23), 5571-5578.

9. Maggiora, G., Vogt, M., Stumpfe, D., & Bajorath, J. (2013). Molecular similarity in medicinal chemistry: miniperspective. *Journal of medicinal chemistry*, *57*(8), 3186-3204.

10. Stumpfe, D., Hu, H., & Bajorath, J. (2019). Evolving Concept of Activity Cliffs. *ACS omega*, *4*(11), 14360-14368.

11. Hussain, J., & Rea, C. (2010). Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *Journal of chemical information and modeling*, *50*(3), 339-348.

12. Anighoro, A., Bajorath, J., & Rastelli, G. (2014). Polypharmacology: challenges and opportunities in drug discovery: miniperspective. *Journal of medicinal chemistry*, *57*(19), 7874-7887.

13. Kawasaki, Y., & Freire, E. (2011). Finding a better path to drug selectivity. *Drug discovery today*, *16*(21-22), 985-990.

14. Peltason, L., Hu, Y., & Bajorath, J. (2009). From structure–activity to structure–selectivity relationships: quantitative assessment, selectivity cliffs, and key compounds. *ChemMedChem: Chemistry Enabling Drug Discovery*, *4*(11), 1864-1873.

.

15. Hann, M. M., & Keserü, G. M. (2012). Finding the sweet spot: the role of nature and nurture
in medicinal chemistry. *Nature reviews Drug discovery*, *11*(5), 355.

16. Johnson, T. W., Dress, K. R., & Edwards, M. (2009). Using the Golden Triangle to optimize
clearance and oral absorption. *Bioorganic & medicinal chemistry letters*, *19*(19), 5560-5564.