UNIVERSITY OF CALIFORNIA SAN DIEGO

Multi-target pharmacology of small molecule drugs and first-in-class inhibitors discovery

A dissertation submitted in partial satisfaction of the requirements for the degree

Doctor of Philosophy

in

Chemistry

by

Da Shi

Committee in charge:

Professor Ruben Abagyan, Chair
Professor Andrew McCammon, Co-Chair
Professor Rommie Amaro
Professor Katja Lindenberg
Professor Sanjay Nigam
Professor Emmanuel Theodorakis

2020

The dissertation of Da Shi is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Chair

University of California San Diego

2020

DEDICATION

I dedicate this dissertation and all work therein to my wife Yingyan Zhuang and family.

Without you, none of this would be possible.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my thesis advisor, Ruben Abagyan, for his guidance and being a wonderful person to work for. I didn't have any experience in computational chemistry and drug discovery when I first joined the lab. Ruben showed great patience to me and taught me a lot in nearly all aspect of computational chemistry. He is the real guidance who introduce the amazing field to me. I would also like to acknowledge Drs. Irina Kufareva, Andrey Ilatovskiy, and Tony Ngo, and my lab mates for their help in teaching me various techniques and providing valuable discussions for my research.

I would also thank my co-advisor, Andrew McCammon, and my thesis committee members, Profs. Rommie Amaro, Katja Lindenberg, Sanjay Nigam, and Emmanuel Theodorakis, for their support and insightful discussion throughout my research. Of course, I must thank my wife, Yingyan Zhuang, without her love and support, none of this would be possible.

The material in Chapter 1, in part, was adapted from **Shi Da**, Feroz Khan, and Ruben Abagyan. "Extended Multitarget Pharmacology of Anticancer Drugs." Journal of chemical information and modeling (2019). The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, was adapted from **Shi Da**, Dmitri Svetlov, Ruben Abagyan, and Irina Artsimovitch. "Flipping states: a few key residues decide the winning conformation of the only universally conserved transcription factor." Nucleic acids research (2017). The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, was adapted from Svetlov Dmitri, **Shi Da**, Joy Twentyman, Yuri Nedialkov, David A Rosen, Ruben Abagyan, and Irina Artsimovitch. "In silico discovery

of small molecules that inhibit RfaH recruitment to RNA polymerase." Molecular Microbiology (2018). The dissertation author was the primary investigator and author of these papers.

Chapter 4, in full, was adapted from **Shi Da**, Kirti Kandhwal Chahal, Patricia Oto, Louis-Felix Nothias, Anjan Debnath, James H McKerrow, Larissa M Podust, and Ruben Abagyan. "Identification of four amoebicidal non-toxic compounds by a molecular docking screen of Naegleria fowleri sterol Δ8− Δ7-isomerase and phenotypic assays." ACS infectious diseases (2019). The dissertation author was the primary investigator and author of this paper.

VITA

2011-2015	Bachelor of Science, Dalian University of Technology
2015-2020	Doctor of Philosophy, University of California San Diego

PUBLICATIONS

Nigam, Anisha K., Julia G. Li, Kaustubh Lall, **Da Shi**, Kevin T. Bush, Vibha Bhatnagar, Ruben Abagyan, and Sanjay K. Nigam. "Unique metabolite preferences of the drug transporters OAT1 and OAT3 analyzed by machine learning." Journal of Biological Chemistry, 2020.

Engelhart, Darcy C., Jeffry C. Granados, **Da Shi**, Milton H. Saier Jr, Michael E. Baker, Ruben Abagyan, and Sanjay K. Nigam. "Systems Biology Analysis Reveals Eight SLC22 Transporter Subgroups, Including OATs, OCTs, and OCTNs." International Journal of Molecular Sciences, 2020.

Shi D†, Chahal K†, Oto P†, et al. “Identification of four amoebicidal non-toxic compounds by a molecular docking screen of Naegleria fowleri sterol Δ8−Δ7-isomerase and phenotypic assays”. ACS Infectious Disease, 2019. († equal contribution)

Shi D, Khan F, Abagyan R, “Expended multi-target pharmacology of anti-cancer drugs”. Journal of Chemical Information and Modeling, 2019.

Zhang Y†, **Shi D**†, Abagyan R, “Population Scale Retrospective Analysis Reveals Potential Risk of Cholestasis in Pregnant Women Taking Omeprazole, Lansoprazole, and Amoxicillin”. Interdisciplinary Sciences: Computational Life Sciences, 2019. († equal contribution)

Steinbrenner AD, Muñoz-Amatriaín M, Venegas JMA, Lo S, **Shi D**, et al. “A receptor for herbivore-associated molecular patterns mediates plant immunity”. bioRxiv, 2019.

Svetlov D†, **Shi D**†, et al. “In silico discovery of small molecules that inhibit RfaH recruitment to RNA polymerase”. Molecular Microbiology, 2018 († equal contribution)

Long T, Rojo-Arreola L, **Shi D**, et al. “Phenotypic, chemical and functional characterization of cyclic nucleotide phosphodiesterase 4 (PDE4) as a potential anthelmintic drug target”. PLoS neglected tropical diseases, 2017.

Shi D†, Svetlov D†, et al. “Flipping states: a few key residues decide the winning conformation of the only universally conserved transcription factor”. Nucleic Acids Research, 2017. († equal contribution)

FIELDS OF STUDY

Major Field: Chemistry

Studies in Computational Chemistry
Professors Ruben Abagyan and Andrew McCammon

ABSTRACT OF THE DISSERTATION

Multi-target pharmacology of small molecule drugs and first-in-class inhibitors discovery

by

Da Shi

Doctor of Philosophy in Chemistry

University of California San Diego, 2020

Professor Ruben Abagyan, Chair

Professor Andrew McCammon, Co-Chair

This dissertation describes studies into a new drug candidate discovery philosophy and its application to diseases. Over the last several decades, drug discovery has been focusing on one single target modulation. However, due to the high failure rate in drug development, a new philosophy that focusing on multiple drug-target interactions has gained escalating attention in both academia and industry. In this philosophy, drugs, especially small-molecule drugs interact with multiple protein targets in their therapeutic concentrations.

In chapter 1, we explored the multi-target pharmacology of cancer drugs. We collected information about multiple targets for each cancer drug along with their experimental effective concentrations or binding activities from multiple sources. We showed that the majority of the cancer drugs had substantial multi-target pharmacology based on our current knowledge. The target subset can further be accentuated and personalized by patient sample-specific expression data. Besides analysis, we also built a web database for the public to access and easily explore the multi-target pharmacology of cancer drugs.

To gain a comprehensive multi-target pharmacology, we still need to study new protein targets to further extend the collection of known targets. In chapters 2 and 3, we studied a bacterial transcription factor, RfaH, which may be developed into a new antibacterial target. RfaH is a transcription processivity factor, belonging to a universally conserved transcriptional regulator, NusG/Stp5 family. Unlike other family members, RfaH exerts a distinct structural transformation during its binding to RNA polymerase. We first identified two key residues for the structural transformation through a combined structural and phylogenetic analysis. Then we screened a compound library and identified 3 first-in-class inhibitors for RfaH binding to RNA polymerase.

In chapter 4, we developed a novel target for Naegleria fowleri infection, primary amoebic meningoencephalitis (PAM). PAM is a rapid-onset brain infection in humans with over 97% mortality rate. Despite some progress in the treatment of the disease, there is no single, proven, evidence-based treatment with a high probability of cure. The target we

developed was ERG2, an isomerase in the ergosterol synthetic pathway. We built a homology model of ERG2 and identified 4 amoebicidal chemicals with low human cell toxicity.

Chapter 1 Multi-target pharmacology of cancer drugs

According to the Food and Drug Administration (FDA), drugs are defined as articles intended for use in the diagnosis, cure, mitigation, treatment, or prevention of disease and as articles intended to affect the structure or any function of the body of man or animals[1]. In the past, drugs were identified and used long before their detailed mechanism of action (MOA), including the biological targets, was characterized and understood[2]. Nowadays, with the advance of molecular biology, gene technology, and protein science, drug development paradigm has shifted to rational target-orientated design. The process usually contains: (i) identify a target of suitable function; (ii) find the best chemical modulator of the target with good physicochemical properties; (iii) test the chemical in animal experiments for efficacy and toxicity; (iv) test the chemical in clinical research with human volunteers. However, no matter how careful the study was undertaken, the success rate of drug development didn't increase appreciably[3]. One possible reason is that despite the efforts put in drug optimization for the target, the drug would almost always hit multiple biological targets and exert unforeseen effects. Therefore, the drug development paradigm has been further modified to focusing on the interaction between a drug and multiple targets, known as multi-target pharmacology. In this chapter, we will analyze the multi-target pharmacology of cancer drugs.

1-1. Introduction

Cancer remains an unsolved healthcare challenge which involves multiple hallmarks, pathways, and individual targets[4]. Despite the significant progress in drug discovery in recent years, the problem remains unsolved due to the diversity of cancer types/subtypes, limited efficacy, excessive toxicity, and acquired treatment resistance[5].

Further complications come from the apparent failure of the “one gene, one drug, one disease” paradigm when applied to cancer[6]. For example, cancer cells may have salvage or compensatory pathways counteracting the intended drug mechanism[6-7]. In addition, even though a small molecule drug may be designed to be specific to the “primary” target, in reality the drug and its metabolites will typically manifest multiple “off-target” activities which can be beneficial, adverse, or neutral[8-9]. Therefore, multi-target pharmacology of drugs needs to be taken into consideration and characterized and quantified. On one hand, exploring multi-target pharmacology of existing drugs can help to identify the potential side effects of drugs and repurpose existing drugs for new cancer types. On the other hand, relevant and efficacious drug combinations can be proposed if multi-target pharmacology is taken into consideration.

Currently, there are several databases, such DrugBank[10], Therapeutic Target Database (TTD)[11], ChEMBL[12], PubChem[13], BindingDB[14], and SuperTarget[15], that contain the drug-target information. The data on multiple target activities of each drug is a big step forward. However, none of those databases alone is both complete and quantitative. These data, whether complete or not, can be transformed into networks of drugs and their targets. Building network maps of drugs and targets based on the complex interaction of multiple drugs and multiple targets is quite challenging. Several attempts have been made to show the complex interaction by: 1) connecting drugs with shared targets to form drug maps; 2) connecting targets with shared drugs to form target maps; 3) connecting drugs and targets to form drug-target bipartite maps[16-20]. However, these maps are both too complex and highly variable because the data sources may be inconsistent, the list of drug-target interactions keeps growing, and the considerations for quantitative contribution or threshold for each edge may be missing or oversimplified.

Here we analyzed the multi-target properties of cancer drugs and generated comprehensive network pharmacology maps of cancer drugs and targets. We extended and updated target lists for all cancer drugs from various sources and quantified them according to the drug-target binding activity values. The resulting network pharmacology maps of cancer drugs and targets, CancerDrugMap (http://ruben.ucsd.edu/dnet/maps/drugnet.html), revealed a higher than expected level of multi-target pharmacology of small molecule drugs, most of which even have targets from different classes. The compiled dataset and maps may be helpful to understand the complexity and difference of pharmacological effects of related drugs, repurpose the drugs for specific patient profiles or develop better drug combinations.

1-2. Materials and Methods

1-2-1. Data collection

Drugs in CancerDrugMap were taken from the following sources: 1) drugs with WHO Anatomical Therapeutic Chemical (ATC) code starting with L01, namely antineoplastic agents; 2) drugs included in the NCI cancer drug list (accessed August 24th, 2018) which have direct antineoplastic effects[21]. Drugs taken from the NCI website may have ATC codes other than L01. For each anti-cancer drug, its target-interaction data were extracted and combined mainly from ChEMBL and PubChem, where the data annotated as “inactive” were excluded. For 101 out of 237 cancer drugs, the drug-target pairs were further extended from the following sources: 1) research publications (67 drugs); 2) FDA drug related documents (30 drugs); 3) European Medicines Agency (EMA) drug reviews (3 drugs); 4) product monograph from the manufacturers (e.g. Dabrafenib); 5) books (e.g. Catumaxomab and Daunorubicin); and 6) US Patents from Google Patent Database (e.g. Alemtuzumab). The drug-target interaction values were transformed into a unified value like pChEMBL (referred to as pAct) which is the logarithmic value of binding/inhibition affinities (Kd/Ki) or half maximal effective/inhibition concentration (EC50/IC50), shown by the following equation [22]; the maximal pAct was taken if multiple pAct values were found.

$$pAct = -\,log_{10}(min(Kd, Ki, EC50, IC50))$$

Equation 1-1

Some drug-target interaction data in ChEMBL and PubChem were annotated as "inconclusive", meaning that more experiments might be needed to validate those interactions. These annotations were kept and used if no "active" interactions were reported. For the DNA/RNA targeting drugs, their targets were named based on the mode of action. For example, NA_ALK denotes the target of alkylating agents. NA_TEM denotes the target of nucleoside analog cancer drugs which act as terminators of DNA replication or RNA transcription. NA_ICL denotes the target of drugs that can intercalate into DNA/RNA and inhibit the replication or transcription. NA_NCB denotes the target of drugs that bind to DNA/RNA through non-covalent interaction. For each cancer drug, apart from target binding activities, we also estimated its number of occurrences from the FDA adverse event reporting system (FAERS)[23] and collected the first FDA approval dates. The FAERS database was pre-processed to standardize the data structure, to homogenize the field names and contents, and to translate diverse set of alternative drug names into their generic names as described previously[24]. In addition, we incorporated the endogenous transporter, carrier, and enzyme information for each cancer drug from DrugBank. We also extracted the RNAseq gene expression data of each target from 1019 cancer cell lines from Cancer Cell Line Encyclopedia (CCLE)[25]. The target identifiers were translated into gene names, and their expression values from individual cell lines were additionally averaged per tissue type.

1-2-2. Distance calculation between cancer drugs

Based on the target binding activities of all cancer drugs, we built a distance function to calculate the dissimilarities between drugs. First, a 237 (drugs) by 783 (targets) matrix was built. A row in the matrix represents a drug, while a column represents a target. The matrix element M_{ik} corresponds to the binding activity of drug i and target k, shown as the pAct form (e.g. 8 in the matrix means the binding activity is 10 nM). We subtracted the baseline of 5 from all non-zero matrix elements, and set negative elements to zero. Therefore, a zero element M_{ik} in Matrix 1 means that drug i is not known to bind to target k or the interaction between drug i and target k is too weak.

Second, based on the matrix M values, distances between drugs (i and j) were calculated according to equations 1-2 – 1-4. The overall distance between two drugs is comprised of two parts: distance calculated from target's binding similarity ($D^{binding}$) and distance calculated from ATC codes of drugs (D^{ATC}) as the length of minimal dendrogram path between two drugs divided by the maximal dendrogram path length (e.g. distance between "L-01-A-A-01" and "L-01-A-B-02" is 0.4). The distances between drugs range from 0 to 1, where 0 means identical and 1 means totally different.

$$D_{ij}^{drug} = 0.5 \times D_{ij}^{binding} + 0.5 \times D_{ij}^{ATC}$$

Equation 1-2

$$\mathrm{D_{ij}^{binding}} = 1 - \frac{\overrightarrow{\mathrm{M_\imath}} \cdot \overrightarrow{\mathrm{M_\jmath}}}{\|\mathrm{M_i}\|\|\mathrm{M_j}\|}$$

Equation 1-3

1-2-3. Distance calculation between drug targets

Similar to the drug distance definition, distances between drug targets were calculated from two parts, distance calculated based on drug binding similarity ($\mathrm{D^{binding}}$), and distance calculated based on Gene Ontology (GO)[26]similarity ($\mathrm{D^{GO}}$), see equation 3. To calculate the $\mathrm{D^{binding}}$, the previous 237 by 783 drug target matrix was transposed, so that a row $\mathrm{M_i}$ represents a target *i*. The matrix element $\mathrm{M_{ik}}$ corresponds to the binding activity of target *i* and drug *k*. Similarly, the baseline of five was subtracted from all non-zero matrix elements.

Following the established un-weighted vector-based distance function[27-30], the GO distance was calculated for 783 targets on the basis of a GO terms matrix (783 by 5938). The GO terms of each drug target were extracted from the UniProt database[31]. A binary GO term matrix was built, in which a row in the matrix denotes a target, while a column denotes a GO term. Matrix element $\mathrm{M_{ik}}$ shows whether the target *i* contain the GO term *j* or not by 1 and 0, respectively.

With those two matrices and the following equations 1-4 – 1-6, distances between pairs of targets were calculated. The distance ranges from 0 to 1, where small distance means that two targets are similar in terms of drug binding and GO annotation.

$$\mathrm{D_{ij}^{target}} = 0.5 \times \mathrm{D_{ij}^{binding}} + 0.5 \times \mathrm{D_{ij}^{GO}}$$

Equation 1-4

$$\mathrm{D_{ij}^{binding}} = 1 - \frac{\overrightarrow{\mathrm{M_\imath}} \cdot \overrightarrow{\mathrm{M_\jmath}}}{\|\mathrm{M_i}\|\|\mathrm{M_j}\|}$$

Equation 1-5

$$D_{ij}^{GO} = 1 - \frac{\overrightarrow{M_\imath} \cdot \overrightarrow{M_\jmath}}{\|M_i\| \|M_j\|}$$

Equation 1-6

1-2-4. Network map generation

The drug network, target network, and target expression network were generated with the Graphviz package, including the Neato tool[32] based on the calculated distances between drugs and targets. The maps are comprised of nodes and edges. Nodes in drug and target maps represent cancer drugs and drug targets respectively. The node sizes and node outline thicknesses in drug and target maps were calculated from drug-target interaction data with equations shown in Table 1. Edges were generated to connect drugs or targets within the distance thresholds (0.28 for cancer drug map, 0.35 for drug target map), which were selected to make the maps compact, visible, and clear. To compare the different distance functions, we also generated the drug and target maps solely based on drug-target interaction data to stress the pharmacological similarities between drugs and targets. We also generated network target maps with their expression levels in each cancer cell lines and tissues. In the target expression network maps, the sizes and opacity values of nodes corresponded to their expression level in the cell line or tissue, which were calculated according to the equations in Table 1.

Table 1-1. Equations to calculate the node size, opacity, and outline thickness for drug maps, target maps, and target expression maps. RPKM, reads per kilo base per million mapped reads, is the target expression value in cell lines and tissues.
* Node opacity in the drug focused target maps was calculated according to the (eq T).

	Drug maps	Target maps	Target expression maps
Node size	$\max(7 \times \left(\sum_{t \in \mathrm{tar}} (1 - \mathrm{e}^{4-\mathrm{pAct}_t})\right)^{0.36}, 8)$	$\max(8 \times \left(\sum_{d \in \mathrm{drug}} (1 - \mathrm{e}^{4-\mathrm{pAct}_d})\right)^{0.5}, 8)$	$\max(8 \times \ln(\mathrm{RPKM}_t + 1), 8)$
Node opacity	1	1 or (eq T) *	$\frac{\ln(\mathrm{RPKM}_t + 1)}{\max_{t \in \mathrm{tar}}(\ln(\mathrm{RPKM}_t + 1)}$ (eq T)
Node thickness	$\max(\max_{t \in \mathrm{tar}}(\mathrm{pAct}_t - 5), 0.5)$	$\max(\max_{d \in \mathrm{drug}}(\mathrm{pAct}_d - 5), 0.5)$	$\max(\max_{d \in \mathrm{drug}}(\mathrm{pAct}_d - 5), 0.5)$

1-2-5. Drug and target classification and statistics

All cancer drugs were classified into nine classes based on their ATC codes, as L01A, L01B, L01C, L01D, L01X, L02, L03, L04, and other codes (A, B…). For the drug target statistics, the L01X class was further divided into L01XC, L01XE, and other L01X (L01XA, L01XX…). All targets were classified into nine classes. The first six classes include kinases, other enzymes, nuclear receptors, G-protein coupled receptors (GPCR), transporters, and nucleic acids. If a target doesn't belong to any of those classes, it was classified based on its principal location: membrane, nucleus, or other. The distributions of targets per drug and drugs per target in each class were generated as box-whisker plots with GraphPad Prism 7.01.

1-2-6. Cross-class targeting statistics

For each class of targets, cancer drugs binding to a member of the target class with activities pAct higher than 5 were considered as drugs binding to the target class. For two target classes, the overlaps of their drug sets were calculated as the fraction of the size of the intersection of two drug sets over the size of the smaller set.

1-3. Results

1-3-1. Multi-target pharmacology has been found for most of the cancer drugs

A list of 237 cancer drugs was obtained from the drugs with Anatomical Therapeutic Chemical (ATC) code L01 and in the NCI cancer drug list[33]. The drug-target interaction data were collected from eight different sources, quantified, and converted to the uniform pAct (-Log(molar concentration)) values (See Methods). To reduce the noise of low activity drug-target interaction, targets with binding activity (pAct) lower than 5 were not included. Almost half of the cancer drugs fall into the category L01X, which contains antibodies (L01XC), kinase inhibitors (L01XE), etc. In addition, about 17% of cancer drugs are not classified into the L01 (antineoplastic) category, meaning that those drugs are mainly used for some other diseases, but they also show anticancer effects (Fig. 1-1a). Target-wise, only 27% of the drugs, typically biologics/antibodies, have only one characterized target. The majority of cancer drugs are known to have multiple targets. Nearly half of the known cancer drug targets are kinases, due to the fact that most kinase inhibitors (ATC code: L01XE) have been tested against the other kinases.

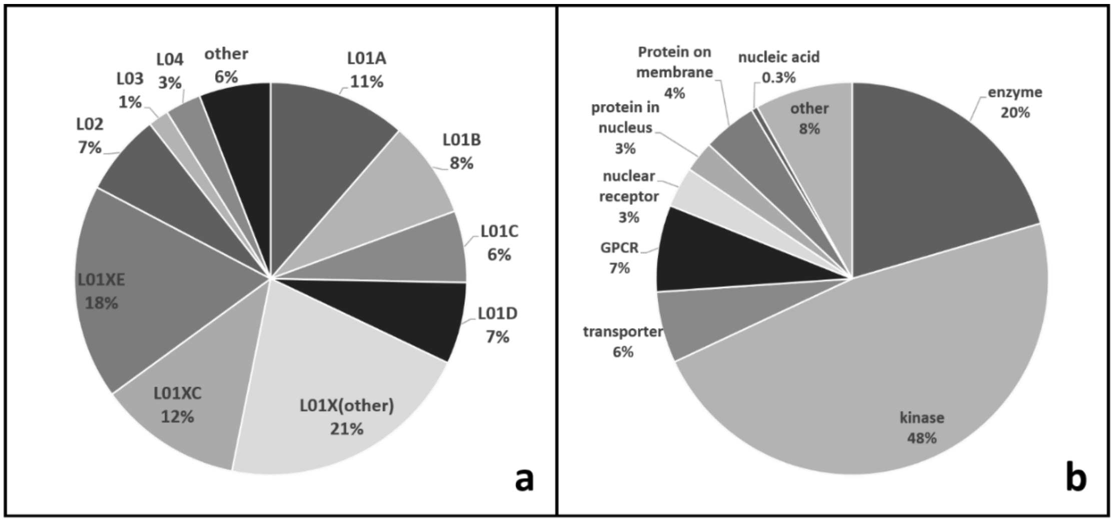


Figure 1-1. Diagrams of (a) ATC code distributions of 237 cancer drugs and (b) drug target classes.

The number of targets per drug in each class is shown in Fig 2. To focus on significant drug-target interaction, we only counted the drug-target pairs with binding activities better than 10μM (pAct higher than 5). The majority of the cancer drugs have two or more known targets, in particular after the antibodies (L01XC) and biologics are excluded. Drugs in the L01XE class (protein kinase inhibitors) have many more known targets than other classes. Most of the drugs in the L01XC class (monoclonal antibodies) have only one target, unless the antibodies are conjugated with a small molecule drug, such as inotuzumab ozogamicin.

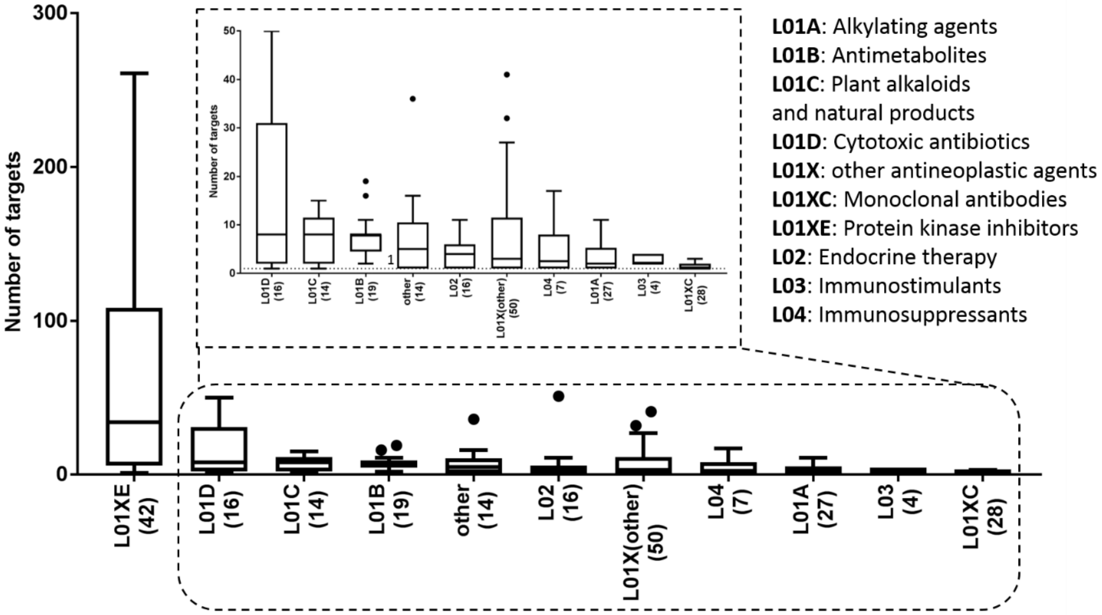


Figure 1-2. Distribution of numbers of targets per drug in each class of cancer drugs, boxes were sorted by the median values.

The number of drugs per target, with pAct greater than 5, in each target class is shown in Fig 3, which could indicate the common target classes of cancer drugs. There are only four targets in the nucleic acid class, leading to the largest number of drugs per target. These nucleic acid targets are special because the drug actions are not as specific as protein targeting drugs. The second most popular target type is protein kinase, despite the large number of kinases, because protein kinase inhibitors (L01XE drug class) are known to act on many kinases concurrently and with significant target overlap.

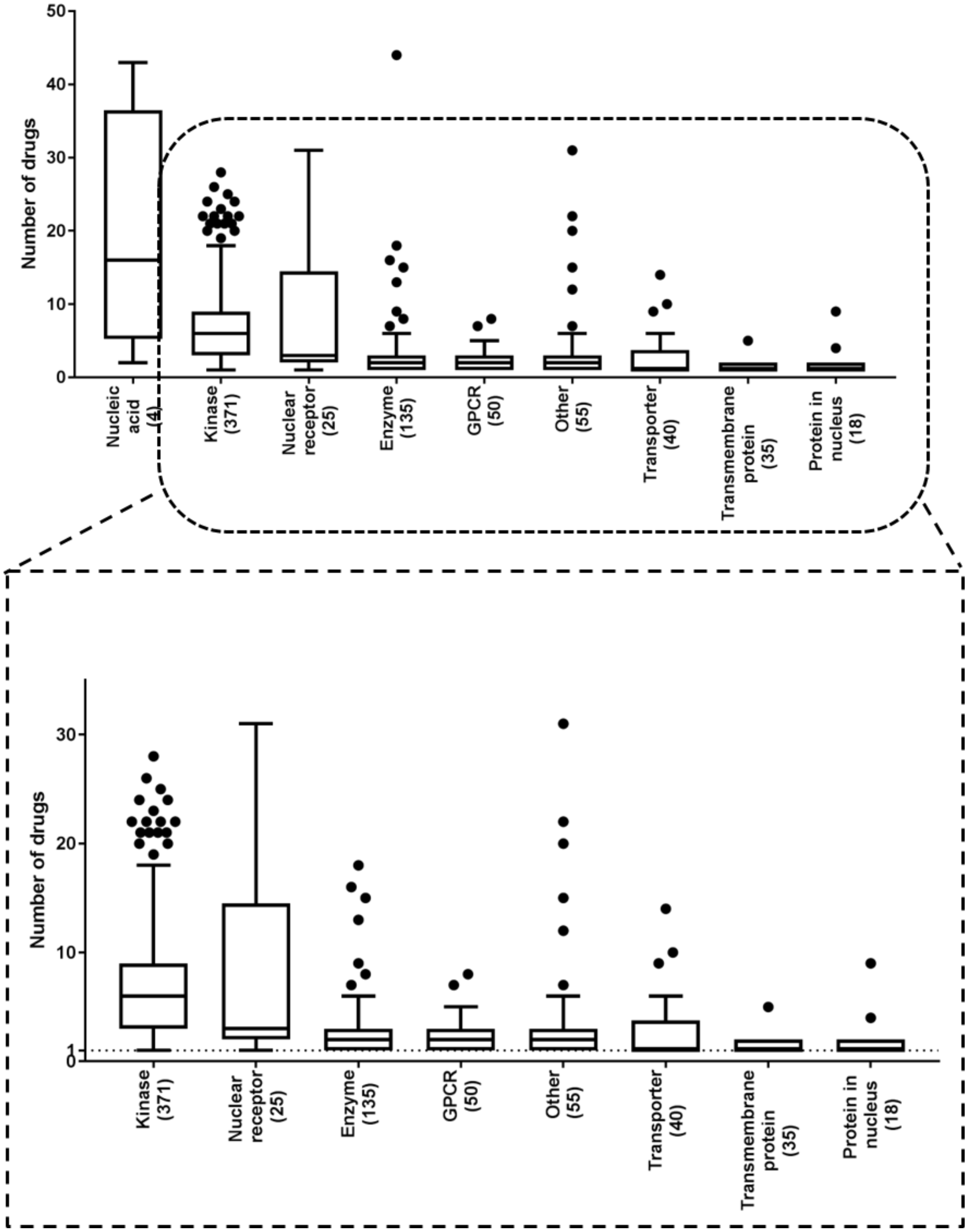


Figure 1-3. Distribution of numbers of drugs per target in each class of drug targets, boxes were sorted by the median values.

1-3-2. New cancer drug development focuses on antibodies and kinase inhibitor

To study the trends of preferred anticancer drug types and target classes, we collected the first FDA approval date of each cancer drug. Most of the newly approved (after 2010) cancer drugs belong to the monoclonal antibody (L01XC) and protein kinase inhibitor (L01XE) classes (Fig 4a). Since the drugs in L01XE class mainly target various protein kinases, targets in the kinase class have the most newly approved drugs (Fig. 1-4b). Similarly, Fig. 1-4b shows that many transmembrane proteins are targeted by newly approved antibodies (L01XC). In addition, the L01XX class has many newly approved drugs targeting newly discovered mechanisms and targets, such as Smoothened receptor (vismodegib, sonidegib), histone deacetylase (vorinostat, belinostat…), proteasome (botezomib, ixazomib…), etc. Discovering and targeting new pathways and proteins continues to lead to new cancer drug developments.

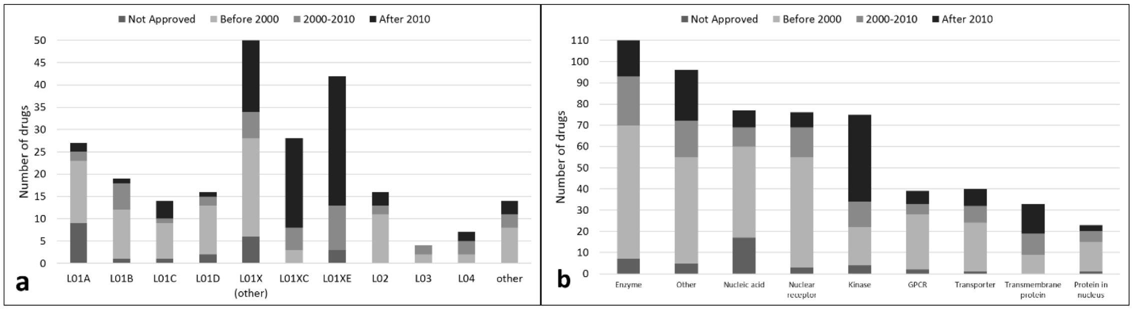


Figure 1-4. Statistics of drugs based on the first approval date. (a) Number of drugs in each drug class, (b) number of drugs binding to each class of targets.

1-3-3. Single cancer drug against multiple target classes

To further explore the multi-target pharmacology of cancer drugs, we studied whether the multiple targets of a drug belong to the same target class or not. We calculated the number of drugs that have known targets belonging to each pair of two target classes (see Methods and Materials), shown in Fig. 1-5a. Most of the target class pairs have overlaps in their drug list, which corresponds to the fact that over half of the drugs have targets belonging to two or more classes, which we named as "cross-class targeting" (Fig. 1-5b). Therefore, the cross-class targeting is a relatively common phenomenon for cancer drugs, and we expect an even higher fraction of cross-class targeting drugs after further research is carried out. One target class that has a significantly low overlap with other classes is the "transmembrane protein". Because many of the targets in this class are receptors of antibodies which usually have single-target pharmacology.

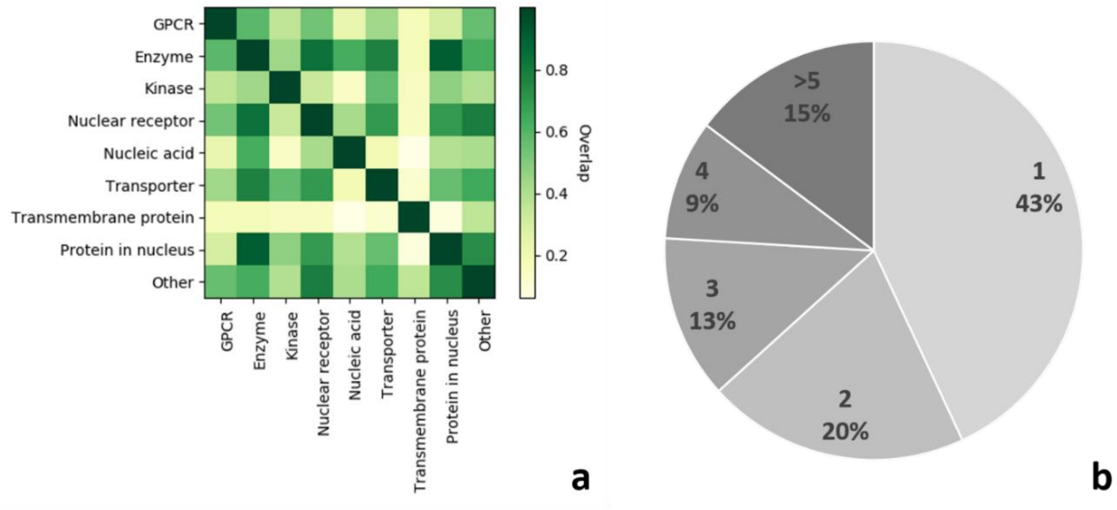


Figure 1-5. Cross-class targeting of cancer drugs. (a) Heatmap showing the overlap of drug sets of two classes of targets, the dark green cells show the drug sets of two classes are highly overlapped. (b) Pie graph of percentage of drugs hitting a given number of target classes.

1-3-4. Drug approval date and number of known targets

It is likely that there is a correlation between the number of known drug-target activities and the time of the drug on the market. We compared the number of known targets for drugs with different approval date ranges. The first approval date was used if the drug has been approved for multiple indications and/or formulations. Drugs were classified into four date ranges, drugs approved before 2000, between 2000 and 2010, after 2010, and not approved (see Fig. 1-6). As expected, drugs approved before 2000 had a larger median number of known targets, while the other three ranges had smaller but comparable values. However, some newly approved drugs, especially protein kinase inhibitors (L01XE), had many more known targets than other drugs (see Fig. 1-6).

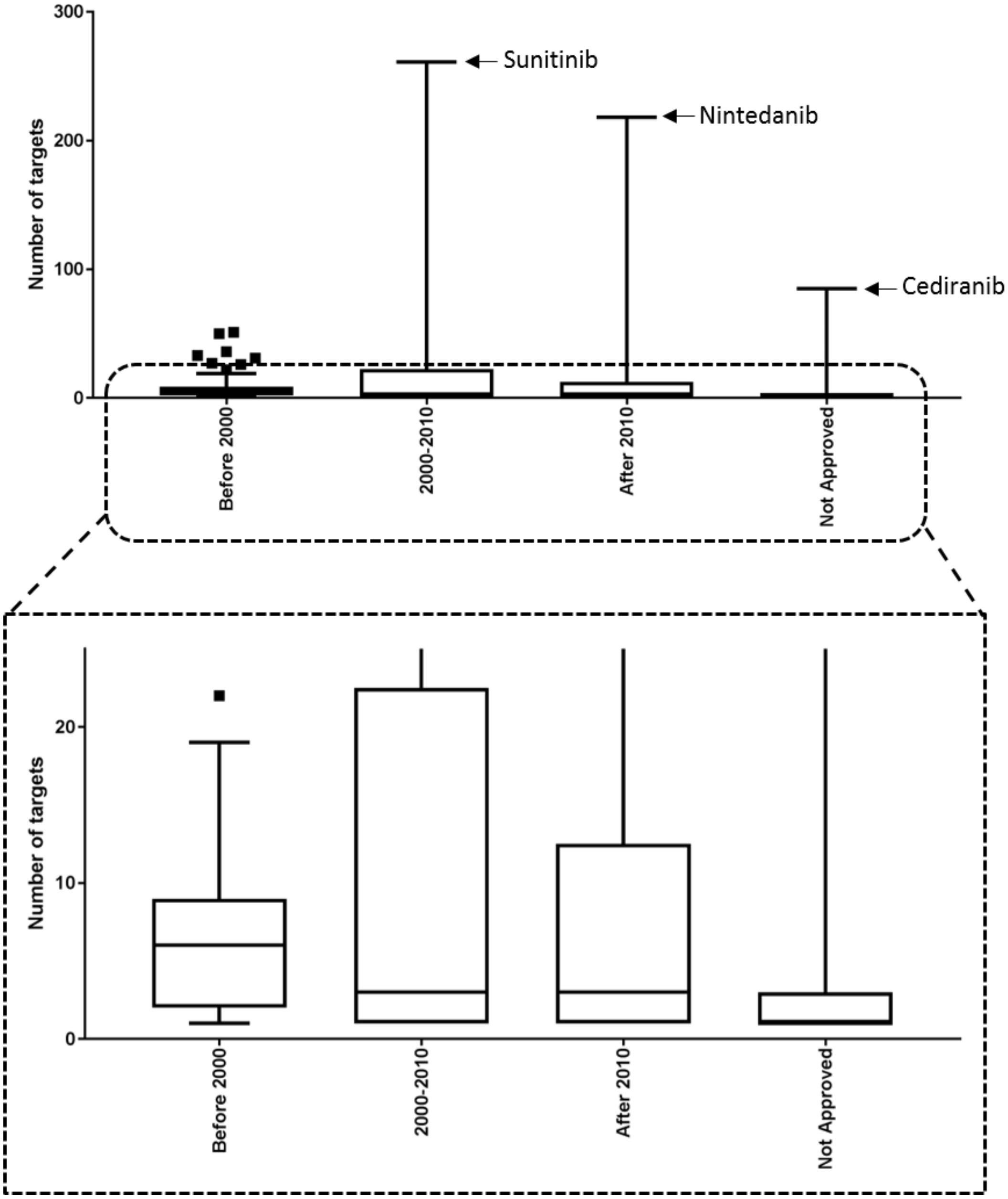


Figure 1-6. Distribution of number of target per drug in each category based on the approval date.

1-3-5. Network pharmacology map and web database layout

To directly show the pharmacology network of cancer drugs, we built sets of two proximity maps, drug-drug map and target-target map. In the drug-drug proximity map (Fig. 1-7), drugs are organized based on fraction of shared targets and ATC code similarities. Drugs are shown as nodes with different colors, and edges connect drugs with the highest target similarities and ATC code similarities. All anti-cancer drugs in the map were classified into nine classes based on their ATC codes, which correspond to different colors of nodes. Inside the node for each drug we incorporated several features of the drug. The Yin-Yang symbol marked covalent drugs which act through covalently binding to their targets, including covalent enzyme inhibitors, alkylating agents, and some nucleoside analogs. The black box symbol marked drugs which have black box warnings in their FDA labels, which are usually more toxic and need special precautions. We also estimated the approximate number of occurrences of each drug from the FAERS database, shown as the black crosses under the drug name. Maps of each of the nine classes of drugs were also generated and can be accessed through the menu. A set of alternative drug maps was also generated in which the distances between drugs were calculated only from drug-target interaction values to emphasize the pharmacological similarity of cancer drugs.

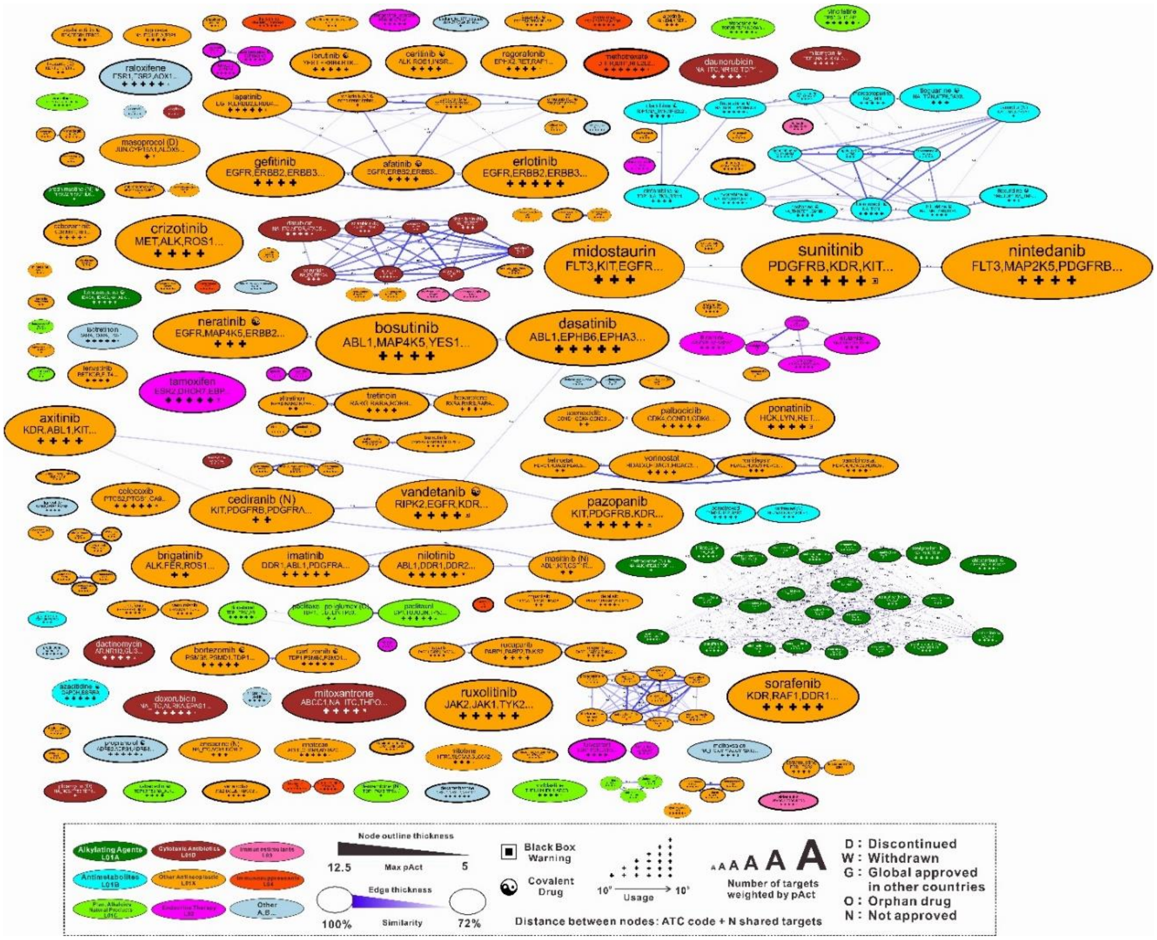


Figure 1-7. Network map of all cancer drugs. Drugs are classified into nine classes based on their ATC codes and colored differently. Drugs within the highest target similarities and ATC code similarities are connected with edges. Size of each node represents the activities weighted sum of number of targets of each drug.

In the target-target proximity map (Fig. 1-8), targets are organized based on the number of concurrently hitting drugs and gene ontology (GO) similarities. Targets are shown as nodes with different colors, and edges connect the closest targets. Targets were classified into nine classes and colored differently to improve the readability of the map. In each target node, the gene name of protein was used except for nucleic acid and Tubulin which is comprised of various subunits. The number under the gene name illustrates the highest binding activity from all anti-cancer drugs for this target. Maps of each target class were also generated and can be accessed in the menu. A set of alternative target maps was also generated where the distances between targets were derived only from drug-target interaction values to emphasize the pharmacological similarity of targets.

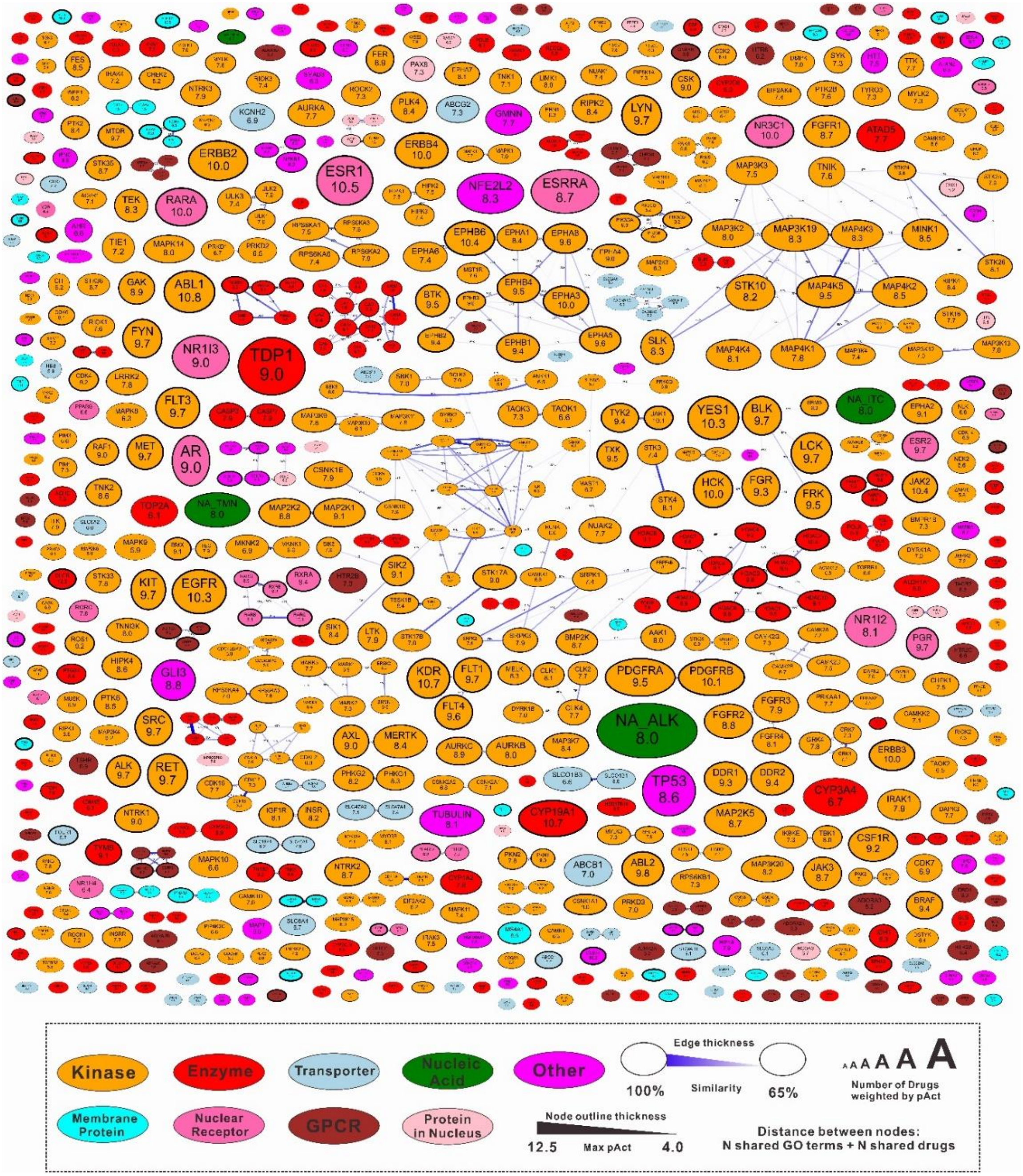


Figure 1-8. Network map of cancer drug targets. Targets are classified into nine classes and colored differently. The closest targets are connected with edges. The size of a node represents the activities weighted sum of number of drugs binding to the target.

A set of expression-value-informed target maps was generated for a various cancer cell lines and tissues by incorporating the expression data of each target into the maps. We extracted the expression level data of each target from the cancer cell line encyclopedia (CCLE)[25], and incorporated the expression data to the map. As shown in Fig. 1-9, the sizes of nodes correspond to the median expression levels of targets in 51 different breast cancer cell lines. The poorly expressed targets were also made pale, so that we can easily identify the highly expressed targets for the cell line of interest.

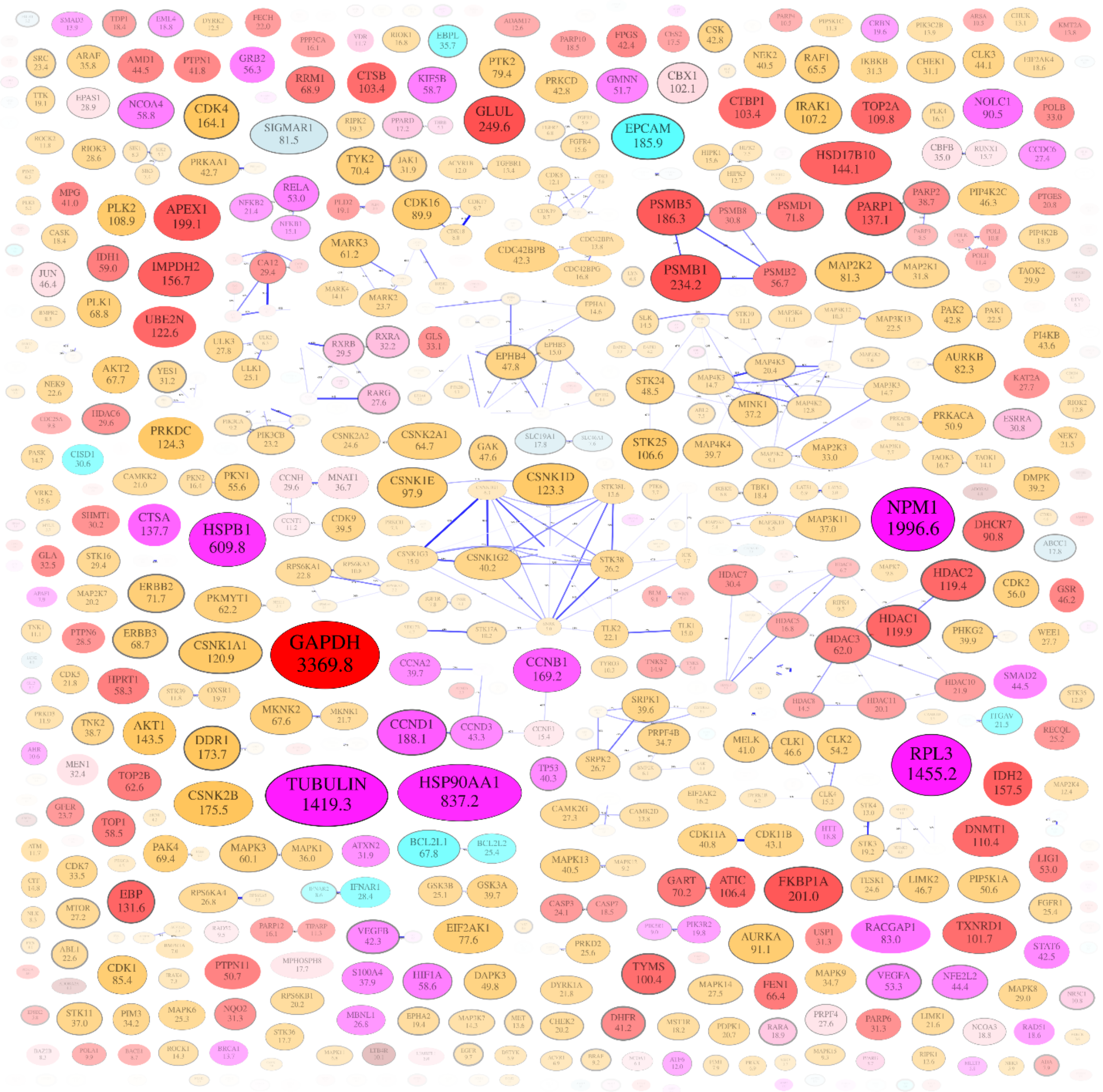


Figure 1-9. Network map of targets of cancer drugs. The median expression level of each target in the breast cancer cell line is incorporated. Size of each node corresponds expression value (Reads per kilo base per million mapped reads, RPKM) of each target which is also shown as the number inside. Nodes with low expression values are pale to highlight the highly expressed proteins.

Besides the network maps, we also generated an information page for each cancer drug and target. These pages can be accessed through clicking the nodes in the drug or target map or searching their names. A drug information page contains some basic information of the drug, such as CAS number, ATC code, and current approval status. The target binding activities of the drug is shown as a table and a bar graph. Using this feature a user can easily figure out the current knowledge about the targets of this drug. In addition, while the drug-drug map only shows connections between drugs within a cutoff distance (0.28), the information page of drug X contains the top-ranking multi-target pharmacology neighbors of drug X. Therefore, we can find the drugs with similar multi-target pharmacology. To study the pharmacokinetics and drug-drug interactions, we included the transporters, carriers and enzymes of each drug in the information page, together with their activity type (substrate, inhibitor, inducer, etc.). To analyze the effects of concurrent usage of cancer drugs, we analyzed the FAERS database and counted the number of records where two cancer drugs were used together. This is displayed as a bar graph in the information page of drug X, and mostly contains drugs which were combined with drug X in FAERS records. However, we should point out that the number of records in FAERS not only means that the drug combination has been used, it also means that one of the drugs or the drug combination is responsible for the reported adverse effects. The information page of target X contains a bar graph with drugs binding to target X, and the likely concurrent targets of target X. For protein targets, we generated a box and whisker graph of the expression levels of each target in different tissues. Additionally, the expression data is displayed in a table.

The complexity of the full drug or target network map can be reduced by focusing on a particular target or drug. The median tissue target expression values may be visualized on the same target map using the opacity property. The focused maps from two drugs or targets can be combined into a single table format to emphasize the differences and overlaps between the two. The filtering feature connects the drug map and target map, and makes it easier for users to explore the multi-target pharmacology of cancer drugs.

1-4. Discussion

With the relatively low success rate of the typical single-target drug discovery paradigm in recent years, multi-target drug and network pharmacology provides a more realistic conceptual framework in both small-molecule cancer therapeutics and other drug development[5-6, 17-18, 34-37]. Our study of cancer drugs revealed that multi-target pharmacology is an expected and inherent property for small-molecule therapeutics. The majority of cancer drugs are already known to hit multiple targets and target classes at therapeutic concentrations. Naturally, these considerations don't extend to monoclonal antibodies that are highly specific to a single target. The cross-class targeting by a single therapeutic is also expected, because receptors and enzymes for the same substrate may differ by backbone topology, yet contain similar binding sites[38]. In addition, some protein targets contain multiple small-molecule binding sites, which may allow chemically diverse drugs to bind[39-41]. For example, it has been shown that protein kinase inhibitors such as, imatinib and nilotinib, are also able to target smoothened receptor of the Hedgehog pathway; celecoxib targets prostaglandin G/H synthase 2, carbonic anhydrases, and several nuclear receptors[42-43].

Even though our study showed that most of the cancer drugs had multiple known targets (the median number of already known concurrent targets of cancer drugs is 5), we believe that the multi-target pharmacology characterizations of cancer drugs is still under-explored. For example, we showed that protein kinase inhibitors (ATC code: L01XE) have a significantly larger number of targets than drugs in other classes. We believe this difference is from the large number of kinases and the experimental availability of the kinase activity panel. However, drugs outside of the L01XE class may not be tested against the whole kinase panel, even if some specific kinase binding activities might be tested. Consequently, we believe the difference of target set sizes of different drug classes results from insufficient experiments. Therefore, our current maps and derivative distributions are built from all experimentally tested and quantified drug-target interactions known today. Naturally, the maps may change as new interactions are discovered and characterized. The current map only represents the presently known sub-group of the full drug-target interaction set. The extent of multi-target pharmacology of small-molecule cancer drugs may expand in the future due to continuous research and improved target identification techniques. As a second tier, computer-based predictions can be performed to identify likely new targets of cancer drugs and prioritize them for experimental validation[44].

This analysis illustrates that the majority of small-molecule cancer drugs have multiple known targets, whereas the biologics (e.g. antibodies) are usually highly specific to one target. The multi-target activities of small-molecule drugs may be both uniquely beneficial and adverse, while the single-target activity of biologics may be insufficient or suboptimal. A recent multi-target drug community challenge, also known as the DREAM challenge, highlighted the emerging appreciation of optimal multi-target profiles: the expected drug candidates were supposed to aim at four different targets simultaneously and avoid three or five other targets[45].

Network maps are an efficient way to visualize and explore the multi-target pharmacology matrix of drugs[46]. In the cancer drug map, we can identify clusters of cancer drugs with similar target activity profiles. Using a quantitative description of multi-target activities of drugs provides a more realistic basis for therapeutic recommendations and new drug development objectives. In addition to the target activity values, we can also project the protein expression data of a specific cancer or patient to this target map, helping to figure out more effective drugs or combination therapies. Furthermore, the network maps can also be used to build predictive models for drug efficacy and drug combination synergy. The distance function for drugs may vary to fit the purpose of the analysis. For example, in a recently published work, the distance between drugs was calculated from target network connectivity counts based on the protein-protein interactome[47]. We adopted the activity-value-weighted distance function together with the ATC-graph-based shortest path distance for drug pairs.

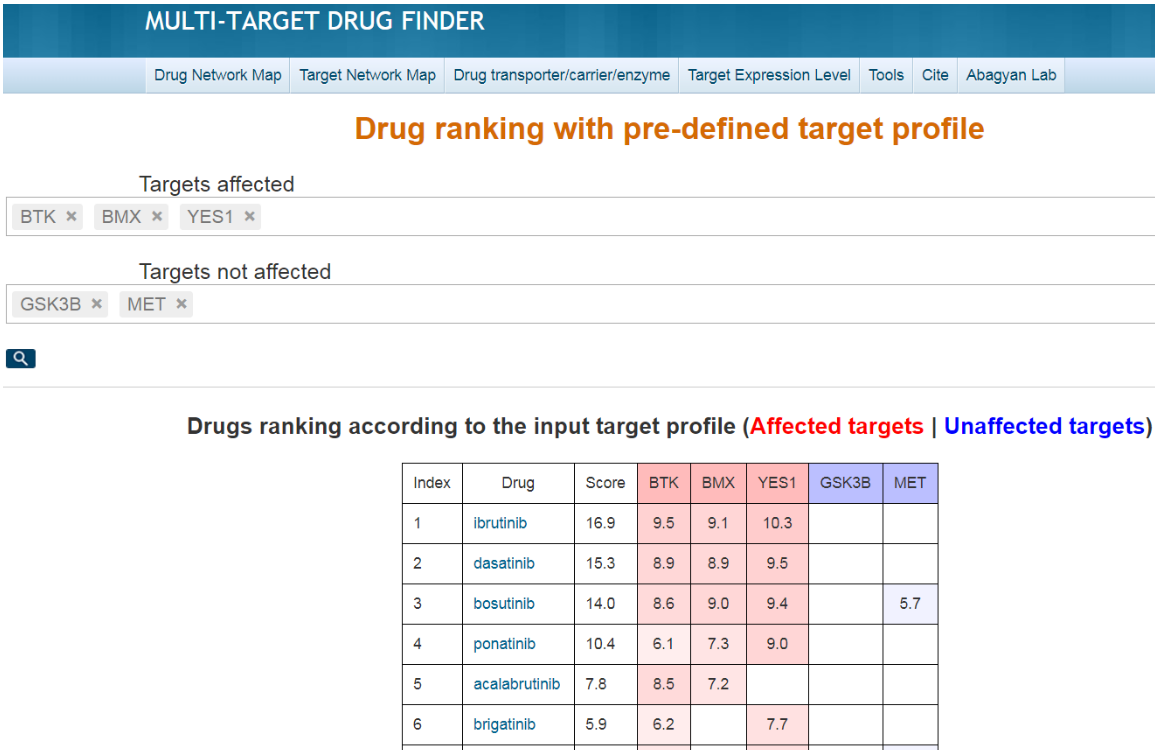
1-5. Conclusion

The substantial and inevitable multi-target pharmacology of small-molecule cancer drugs needs to be incorporated into the mechanism of drug actions, therapeutic strategies and drug discovery objectives[34-36, 48-50]. In this chapter we analyzed the already known multi-target pharmacology properties of cancer drugs. By compiling the drug-target interaction data from various sources, we greatly expanded the number of targets of cancer drugs. We showed that the majority of cancer drugs affect multiple targets at therapeutic concentrations, and over a half of the cancer drugs are known to hit different target classes concurrently. The multi-target pharmacology network of cancer drugs is still not fully explored, and it will grow with the advance of high-throughput experimental binding and activity assays. In addition, based on the expanded drug-target binding activities data, we built cancer drug network maps and target network maps and made them available as a web database, CancerDrugMap. The database contains a comprehensive cancer drug-target interaction data with an emphasis on realistic multi-target pharmacology at therapeutic drug concentrations and target expression levels in different cell lines. This information may be valuable for repurposing drugs to different cancer types, for identifying complementary and synergistic drug combinations, and for customizing prescriptions for patient-specific target profiles.

1-6. Appendix: Multi-target drug query

In this chapter, we analyzed the multi-target pharmacology of cancer drugs by compiling a comprehensive drug-target interaction collections and built a series of proximity network maps. We showed that most of the current cancer drugs had more than one known targets, and even targets from different classes. Therefore, we believe that developing or assessing drugs will benefit from targeting multiple proteins. In the assessment of current drugs, identification of multiple targets of a drug can provide potentials of repurposing the drug for a different indication, or better illustrating the potential adverse effects the drug may cause. In the new drug development, multi-targeted drug may generate better efficacy and potency through regulating multiple proteins/pathways simultaneously. To achieve that, a well-defined target profile is needed which may contain several desired targets to hit and several undesired targets to avoid.

We developed a tool in the CancerDrugMap, Multi-target Drug Finder, to enable querying cancer drugs with predefined target profiles. Shown as the Fig. 1-10, a predefined target profile can be entered as “targets affected” and “targets not affected”. The results rank order of cancer drugs based on profile matching. With that, we can identify drugs with desired target profiles, and some drugs which may be optimized to obtain desired target profiles.



Drug ranking with pre-defined target profile

Targets affected

BTK × BMX × YES1 ×

Targets not affected

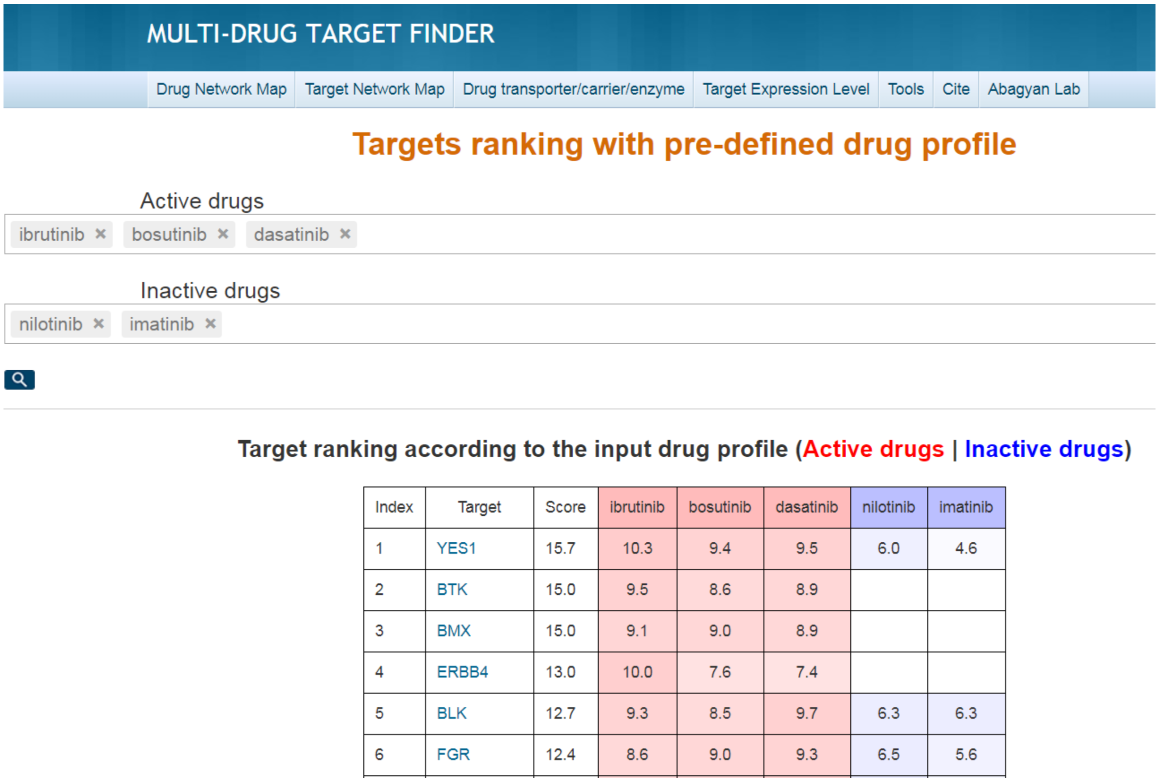
GSK3B × MET ×

Drugs ranking according to the input target profile (Affected targets | Unaffected targets)

Index	Drug	Score	BTK	BMX	YES1	GSK3B	MET
1	ibrutinib	16.9	9.5	9.1	10.3		
2	dasatinib	15.3	8.9	8.9	9.5		
3	bosutinib	14.0	8.6	9.0	9.4		5.7
4	ponatinib	10.4	6.1	7.3	9.0		
5	acalabrutinib	7.8	8.5	7.2			
6	brigatinib	5.9	6.2		7.7		

Figure 1-10. Screenshot of the interface of Multi-target Drug Finder.

In the drug assessment scenario, we may need to compare two groups of drugs sometimes. For example, one group of drugs is active, and the other group is inactive, or one group may cause a specific side effects while the other group not. We may analyze the multi-target pharmacology of each group of drugs and identify the targets that are responsible for the activities or adverse effects. To achieve that, we developed a similar tool, Multi-drug Target Finder, to easily identify target difference between the two groups of drugs. Shown as Fig. 1-11, a user can enter the name of two groups of drugs to “active drugs” and “inactive drugs” boxes, and the results show targets which have high drug binding affinities for active drugs and low/no drug binding affinities for inactive drugs. With this, we can clearly know the target difference between two groups and further investigate the detailed biological mechanism under the phenomena.



Targets ranking with pre-defined drug profile

Active drugs: ibrutinib ×, bosutinib ×, dasatinib ×

Inactive drugs: nilotinib ×, imatinib ×

Target ranking according to the input drug profile (Active drugs | Inactive drugs)

Index	Target	Score	ibrutinib	bosutinib	dasatinib	nilotinib	imatinib
1	YES1	15.7	10.3	9.4	9.5	6.0	4.6
2	BTK	15.0	9.5	8.6	8.9		
3	BMX	15.0	9.1	9.0	8.9		
4	ERBB4	13.0	10.0	7.6	7.4		
5	BLK	12.7	9.3	8.5	9.7	6.3	6.3
6	FGR	12.4	8.6	9.0	9.3	6.5	5.6

Figure 1-11. Screenshot of Multi-drug Target Finder

Furthermore, to apply this multi-target pharmacology philosophy to a broader field, I built two similar tools as Multi-target Drug Finder and Multi-drug Target Finder (http://ruben.ucsd.edu/dnet/dtpro/tar_pro.html). By parsing and compiling the latest ChEMBL database, these tools can run the queries with all known compounds and protein targets, which would be helpful in the early development of drugs.

1-7. Acknowledgement

The material of Chapter 1, before Appendix, has been published by **Shi Da**, Feroz Khan, and Ruben Abagyan. "Extended Multitarget Pharmacology of Anticancer Drugs." Journal of chemical information and modeling (2019). Copyright 2019 American Chemical Society. The dissertation author was the primary investigator and author of this paper.

Chapter 2 Identification of a new target for bacterial infection. Part 1: Identification of key residues for domain interaction

After acknowledging the multi-target pharmacology of drugs, the remaining task would be identifying new targets to expand our known target set which is the fundamental of multi-target drug discovery and assessment. In the following two chapters, we explored and identified a novel target for bacterial infection. The target, RfaH, is essential for the virulence and antibiotic-resistance gene horizontal transfer of Gram-negative pathogens, is a difficult target because of its dynamic structure and rather shallow pocket. In this chapter, we first developed a method to analyze the structure of RfaH and identified two key "switch" residues for the structural transition.

2-1. Introduction

Gene duplication and subsequent functional divergence of paralogs is one of the main sources of evolutionary diversity in all living systems[51]. Two models of functional adaptation are commonly considered: subfunctionalization, wherein the duplicates partition the ancestral function, and neofunctionalization, wherein one duplicate acquires a novel function. The evolution of the NusG family of transcription elongation factors provides a particularly striking example of neofunctionalization accompanied by transformation[52], the ability of one duplicate to undergo an α-to-β fold conversion that bestows a new function.

Proteins from the NusG/Spt5 family are the only known examples of universally conserved transcriptional regulators[53]. NusG-like proteins are composed of an α/β N-terminal domain (NTD) and a β-barrel C-terminal domain (CTD) that contains a Kyprides-

Onzonis-Woese (KOW) motif commonly found in ribosomal proteins[54-55]. The two domains are connected by a flexible linker and together enable uninterrupted synthesis of long RNA molecules in synchrony with ongoing cellular processes, such as translation in prokaryotes and splicing and polyadenylation in eukaryotes. The NTDs bind to the two pincers of elongating RNA polymerase (RNAP), forming processivity clamps around the nucleic-acid chains[56-57]. The location of the RNAP-binding site and the mode of NTD action appear to be ubiquitous among all NusG proteins[58]. In contrast, the CTDs interact with an astonishingly diverse set of cellular partners that include the bacterial ribosome[59] and yeast splicing and capping factors[60].

Escherichia coli NusG and its paralog RfaH are the best characterized transcription elongation factors. RfaH and NusG share binding sites on the transcription elongation complex (TEC) and the ribosome, as well as the molecular mechanism of RNAP modification into a highly processive, pause-resistant state. Strikingly, however, the cellular functions of NusG and RfaH are not only different but opposite (Fig. 2-1). NusG is an essential and abundant (~5,000 copies/cell;[61]) protein that associates with RNAP transcribing almost all genes, displaying no apparent sequence specificity[62]. The NusG CTD binds to the transcription termination factor Rho, stimulating Rho activity *in vitro* and *in vivo*[63]. Together, NusG and Rho silence foreign DNA[64]; NusG becomes largely dispensable in a genome-reduced *E. coli* strain from which the horizontally-acquired regions have been removed[64]. By contrast, RfaH is scarce (50 copies/cell;[61]), does not bind to Rho (at least at physiological conditions/concentrations), and reduces Rho-dependent termination *in vitro*[65], likely by disfavoring the paused RNAP state which is a target for Rho. RfaH is recruited to only those few operons that contain a 12-nt-long *ops* DNA element in their leader regions[66] and strongly activates their expression by abolishing Rho-dependent termination[67] and increasing translation[68]; RfaH excludes NusG through direct competition for the shared binding site on RNAP[66] and is thought to directly recruit the 30S subunit of the ribosome through protein-protein interactions between the CTD and the ribosomal protein S10[68]. Every gene that RfaH controls is horizontally transferred, and many of them are essential for virulence; loss of *rfaH* attenuates virulence in *E. coli*, *Salmonella* and *Klebsiella pneumoniae*[69-71].

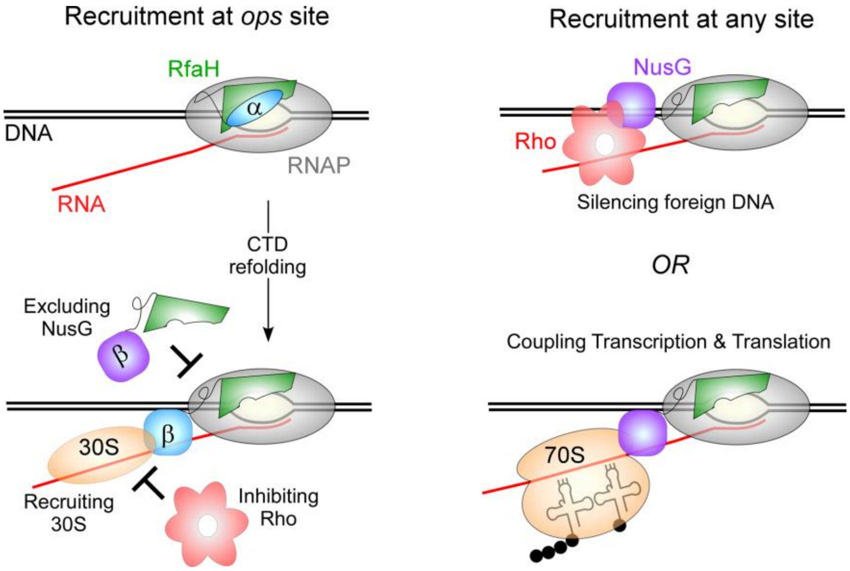


Figure 2-1. Regulatory mechanisms of RfaH and NusG.

Since RfaH directly opposes the action of the essential NusG, RfaH activity needs to be tightly controlled. This is accomplished by a combination of much reduced levels and exquisite specificity of RfaH, which depends absolutely on the *ops* signal for recruitment to the transcription elongation complex (TEC). A basic patch on the RfaH NTD recognizes the *ops* bases[72] on the non-template DNA strand in the transcription bubble exposed on the surface of RNAP paused at the *ops* site[65]. These residues are not conserved in NusG, and this divergence could explain RfaH preference for a specific site. However, the *ops* plays another, more critical role in RfaH recruitment: contacts with *ops* transform a silent, autoinhibited RfaH into an activated state capable of binding to RNAP[73]. In contrast to *E. coli* NusG, in which the freely rotating NTD and CTD are connected by a highly flexible linker[74], the CTD in free RfaH is folded as an α-helical hairpin that forms a large hydrophobic interdomain interface (IDI), masking the RNAP-binding site on the NTD[73]. The domain dissociation is triggered by binding to the *ops* element and is a prerequisite for NTD recruitment to RNAP; similarly to NusG, the isolated RfaH NTD binds to the TEC indiscriminately, bypassing the need for activation[73].

The interconversion between the two different states of the CTD is a signature of RfaH action, with both states playing essential roles. The isolated CTDs of all NusG-like proteins, including RfaH, fold as nearly superimposable β-barrels. The β-CTD of RfaH binds to the ribosomal protein S10 to recruit the ribosome to the nascent mRNA, the most critical activity of RfaH; analogous NusG-S10 contacts are thought to couple transcription to translation. The α-CTD restricts RfaH action to a handful of genes, preserving the essential regulation by NusG. Thus, attainment of the transforming capability that is essential for autoinhibition was the key step in the evolution of dedicated RfaH-like regulators acting alongside NusG. The determinants of the dramatic refolding behavior of RfaH CTD are not yet known, although several molecular dynamics (MD) studies provided insights into this phenomenon. In this work, we carried out an analysis of bacterial NusG and RfaH subfamilies to identify specific residues that may define their different folds and respective properties. We show that substitutions of RfaH residues predicted to play key roles in maintenance of the interdomain contacts, Ile93 and Phe130, for their NusG counterparts relaxes the requirement for *ops*, “converting” RfaH into a non-specific regulator in which the IDI is partially destabilized.

2-2. Materials and Methods

2-2-1. Plasmids and reagents

All general reagents were obtained from Sigma Aldrich (St. Louis, MO) and ThermoFisher Scientific (Pittsburgh, PA); NTPs - from GE Healthcare (Piscataway, NJ); and [α^{32}P]-CTP - from Perkin Elmer (Boston, MA). PCR reagents, restriction and modification enzymes were from NEB and Roche (Indianapolis, IN). Ni-sepharose resin, HiTrap Heparin HP and Resource Q columns were from GE Healthcare. Oligonucleotides were obtained from Sigma Aldrich. DNA purification kits were from Qiagen (Valencia, CA) and Promega (Madison, WI).

2-2-2. Proteins

E. coli RNAP core and σ^{70}, WT RfaH and isolated domains were purified as in[73]. RfaH variants I93E (pIA1253) and F130V (pIA1254) were constructed by site-directed mutagenesis in pIA751; these proteins carry a His$_6$ tag followed by a TEV cleavage site. The mutant proteins were purified from XJb (λDE3) strain as described previously [72]. To remove His tags, His$_6$ tagged TEV protease (100 μg) was incubated with the protein sample (~ 8 mg) at 4 °C for 20 h. The cleaved-off His$_6$ tag, the uncut His$_6$-protein, and (His-tagged) TEV were removed by absorption to Ni-sepharose. Proteins were dialyzed into storage buffer (50% glycerol, 100 mM NaCl, 10 mM Tris-HCl pH 7.9, 0.1 mM EDTA, 0.1 mM DTT) and stored at -20 °C.

2-2-3. Template preparation

Templates for in vitro transcription were generated by PCR amplification from pIA1087 (WT *ops*) and pZL23 (G8C *ops*) reporter plasmids encoding the *rfb* leader region-

lux operon fusion under control of *E. coli* P_{BAD} promoter [68]. To enable efficient transcription and the formation of halted radiolabeled TEC, the first PCR step was performed with a 73-nt long primer adding the T7A1 promoter and a 24-nt long U-less region to the *rfb* leader region (2536; AAAAAGAGTATTGACTTAAAGTCTAACCTATAGGATACTTA CAGCCATCGAGCAGGCAGCGGCAAAGCCATGG) and a reverse primer (2537; AAATAAGCGGCTCTCAGTTT). Following the removal of primers, the second step PCR was performed with primer 2537 and a forward primer 2499 (AAAAAGAGTATTGACTTAAAG). The amplified sequence encompasses -46 through +79 positions relative to the T7A1 transcription start site.

2-2-4. Single-round transcription elongation assays

Linear DNA template (30 nM), holo RNAP (40 nM), ApU (100 µM), and starting NTP subsets (1 µM CTP, 5 µM ATP and UTP, 10 µCi [α^{32}P]-CTP, 3000 Ci/mmol) were mixed in 100 µl of TGA2 (20 mM Tris-acetate, 20 mM Na-acetate, 2 mM Mg-acetate, 5% glycerol, 1 mM DTT, 0.1 mM EDTA, pH 7.9). Reactions were incubated for 15 min at 37 ^{o}C; thus halted TECs were stored on ice. RfaH variants (or an equal volume of storage buffer) were added to the TEC, followed by a 2-min incubation at 37 ^{o}C. Transcription was restarted by addition of nucleotides (10 µM GTP, 150 µM ATP, CTP, and UTP) and rifapentin to 25 µg/ml. Samples were removed at time points indicated in the figures and quenched by addition of an equal volume of STOP buffer (10 M urea, 60 mM EDTA, 45 mM Tris-borate; pH 8.3). Samples were heated for 2 min at 95 ^{o}C and separated by electrophoresis in denaturing 8 % acrylamide (19:1) gels (7 M Urea, 0.5X TBE). The gels were dried and RNA products were visualized and quantified using FLA9000

Phosphorimaging System, ImageQuant Software, and Microsoft Excel.

2-2-5. Chymotrypsin digestion

Chymotrypsin (Sigma Aldrich) was dissolved in 1 mM HCl (as recommended by the manufacturer) at 2 mg/ml and stored at -80 ^{o}C is single-use aliquots. Prior to use, an aliquot was diluted into PBS, pH 7.4 (ThermoFisher Scientific) on ice. 9 µl of chymotrypsin in PBS (0.2 mg enzyme) were mixed with 6 µl of RfaH variants or domains (~2 mg protein) in storage buffer (50% glycerol, 100 mM NaCl, 10 mM Tris-HCl pH 7.9, 0.1 mM EDTA, 0.1 mM DTT). The volume used was dictated by the concentration of the least soluble RfaH variant, I93E; higher glycerol concentrations were found to inhibit chymotrypsin cleavage. To the control samples, only PBS was added. The reactions were incubated at 37 ^{o}C for 10, 20, 40 and 80 min and stopped by the addition of 5 mM PMSF and LDS loading dye (ThermoFisher Scientific). Samples were heated at 75 ^{o}C for 5 min and 8 µl were loaded onto 4-12% Bis-Tris gels, which were run in 1X SDS-MES buffer at 180V. The gels were stained with GelCode Blue (ThermoFisher Scientific). With each RfaH variant, the assay was repeated at least 3 times; the WT protein was assayed in parallel every time.

2-2-6. Calculation of Entropy and Conservation score

RfaH sequences were aligned with implemented tools in ICM[75]. Based on the alignment, we assessed two quantitative characteristics of diversity: Entropy and Conservation score. Entropy was calculated according to Equation 2-1, where P_a^i is the normalized ratio of the observed frequency of amino acid *a* at position *i* divided by the expected frequency for the same amino acid.

$$\mathrm{Entropy\ of\ position\ i} = -\sum_{\mathrm{a}} \mathrm{P_a^i} \ln \mathrm{P_a^i}$$

Equation 2-1

The Conservation score is based on the mean pairwise score between residues *j* and *k* in alignment position *i*. N_{seq} is number of sequences in the alignment, C_j^k is the similarity between residues *k* and residues *j* at position *i* taken from a normalized compare matrix[76].

$$\text{Conservation score of position i} = \frac{1}{2\mathrm{N_{seq}}}\sqrt{8\mathrm{d_i} + 1}$$

$$\mathrm{d_i} = \sum \mathrm{C_j^k}\ (\,\mathrm{j, k} = 1,2,3 \ldots \ldots \mathrm{N_{seq}})$$

Equation 2-2

2-2-7. Calculation of interdomain interface contact area

The IDI contact areas of residues of RfaH were calculated with implemented tools of ICM [77]. Firstly, the solvent accessible areas of each residues were calculated using a water probe with radius of 1.4 Å in the closed state, in which the CTD and the NTD interact. Then solvent accessible areas were calculated upon separation of the two domains. The difference between the two represents the IDI contact areas of residues.

2-2-8. Calculation of domain binding energy contributions (ΔΔG_{bind}) of residues

ΔΔG*bind* of residue k was calculated with implemented tools of ICM by evaluating the effect on the binding free energy upon mutation of residue k to a glycine, using equations 2-3 and 2-4, where

$$\Delta\Delta\mathrm{G_{bind}} = \Delta\mathrm{G_{bind}^{mut}} - \Delta\mathrm{G_{bind}^{wt}}$$

Equation 2-3

$$\Delta\mathrm{G_{bind}} = \left(\mathrm{E_{intra}^{comp}} - \mathrm{E_{intra}^{parts}}\right) + (\Delta\mathrm{G_{solv}^{comp}} - \Delta\mathrm{G_{solv}^{parts}})$$

Equation 2-4

ΔG_{bind}^{wt} represents the binding free energy of NTD and CTD in wildtype RfaH, while ΔG_{bind}^{mut} represents the binding free energy of NTD and CTD in residue mutated RfaH. E_{intra}^{comp} represents the internal energy (van der Waals, electrostatic, hydrogen bonds and torsion components) of NTD-CTD complex, while E_{intra}^{parts} represents the sum of internal energy of NTD and CTD. Similarly, ΔG_{solv}^{comp} represents the solvation energy of NTD-CTD complex, while ΔG_{solv}^{parts} represents sum of solvation energies of NTD and CTD.

2-3. Results

We first performed an *in silico* analysis of RfaH and NusG subfamilies, in the following order: (i) to identify amino acid residues that are conserved in the RfaH subfamily; (ii) to assess their potential to disrupt the closed α-helical state but not the open β-barrel state; (iii) to simulate the structural and energetic effects of a substitution at the IDI in the closed state; and (iv) to identify the equivalent *E. coli* NusG residues that are conserved within the NusG subfamily yet distinct from those in RfaH.

2-3-1. Identifying residues that contribute to the closed-state stabilization in RfaH

1383 sequences of RfaH proteins in different organisms were obtained from InterPro [78], and duplicate identical sequences were removed. Alignment of the remaining 751 sequences built with ICM [79-80] identified ~90% similarity-conservation for 36 positions (Fig. 2-2). To quantitatively assess diversity, we calculated the entropy and the conservation score (Table 2-1) of each RfaH residue (see Materials and Methods). Conserved residues have low entropies and high conservation scores; we set the conservation score > 0.8 and Entropy < 0.9 as filters in this analysis.

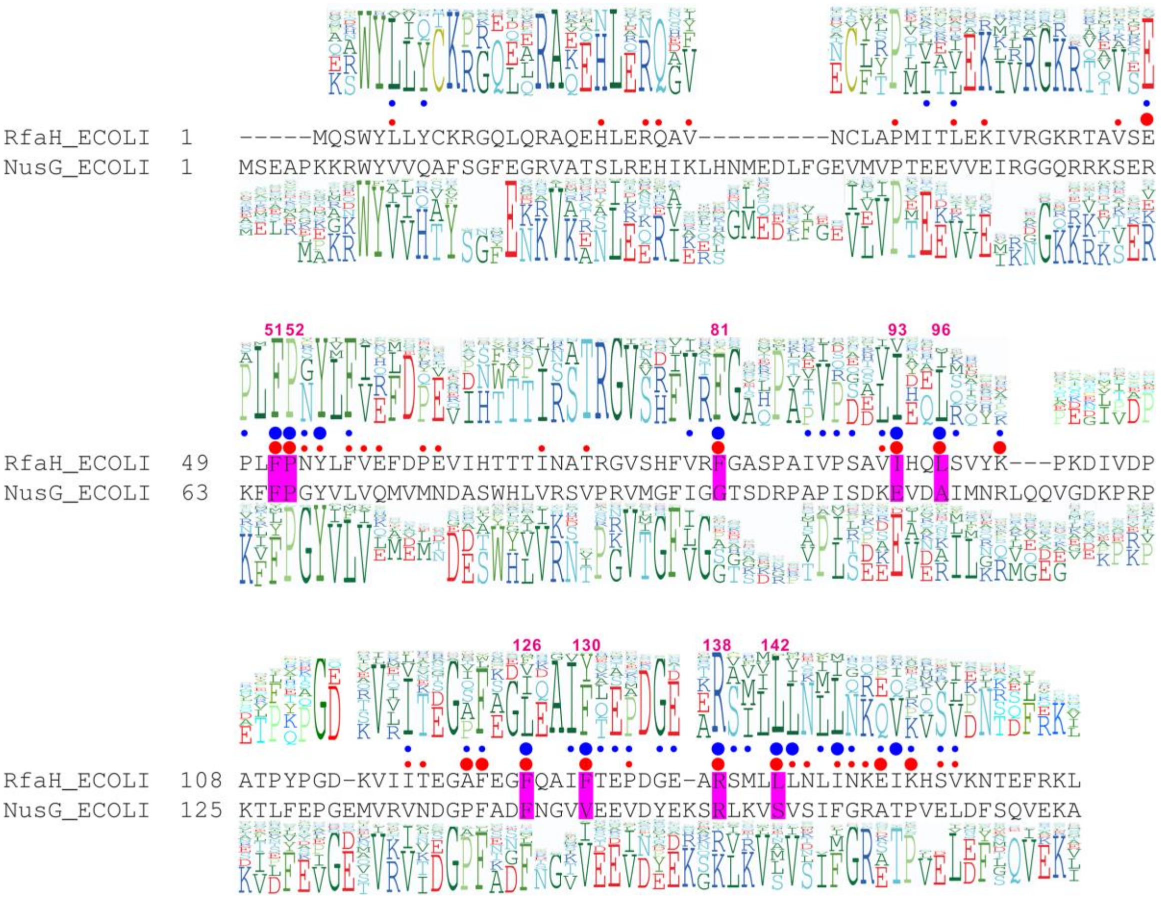

51
52
81
93
96
126
130
138
142
RfaH_ECOLI 1 -----MQSWYLLYCKRGQLQRAQEHLERQAV---------NCLAPMITLEKIVRGKRTAVSE
NusG_ECOLI 1 MSEAPKKRWYVVQAFSGFEGRVATSLREHIKLHNMEDLFGEVMVPTEEVVEIRGGQRRKSER
RfaH_ECOLI 49 PLFPNYLFVEFDPEVIHTTTINATRGVSHFVRFGASPAIVPSAVIHQLSVYK---PKDIVDP
NusG_ECOLI 63 KFFPGYVLVQMVMNDASWHLVRSVPRVMGFIGGTSDRPAPISDKEVDAIMNRLQQVGDKPRP
RfaH_ECOLI 108 ATPYPGD-KVIITEGAFEGFQAIFTEPDGE-ARSMLLLNLINKEIKHSVKNTEFRKL
NusG_ECOLI 125 KTLFEPGEMVRVNDGPFADFNGVVEEVDYEKSRLKVSVSIFGRATPVELDFSQVEKA

Figure 2-2. The key features of the RfaH and NusG families. Structural alignment of E. coli RfaH and NusG is shown in the middle. The NTD alignment was derived from the superposition of PDBs 2OUG (RfaH) and 2K06 (NusG), and the CTD alignment was derived from the superposition of PDBs 2LCL (RfaH) and 2KVQ (NusG). A profile above the alignment was generated from the sequence alignment of 751 RfaH sequences, while profile underneath was generated from the sequence alignment of 9204 NusG sequences. Red circles in the middle indicate the $\Delta\Delta G_{bind}$ value; large, >1.5 kcal/mol, small, 1–1.5 kcal/mol. IDI contact areas are shown as blue circles; large, IDI contact areas >50 Å^2, small, <50 Å^2. The residues with large IDI contact areas and $\Delta\Delta G_{bind}$ are shaded in magenta and labeled with the residue number in RfaH.

The unique closed state of RfaH is stabilized by interactions between the NTD and the α-helical CTD. To identify the residues that make key contributions to the closed-state stabilization, their IDI contact areas were calculated (see Materials and Methods). Residues with larger IDI contact areas are more likely to be directly involved in stabilizing the α-state of CTD and thus the closed state of RfaH. The IDI contact areas of each residue are shown as blue circles in Figure 2A; large circles indicate IDI contact areas larger than 50 Å^2, small circles, IDI contact areas between 0 and 50 Å^2. A contact area of 50 Å^2 was chosen as a filter.

To assess the energetic contribution of individual residues to the closed-state stabilization, we calculated the binding free energy change upon in silico substitution of each residue with glycine [81]. Substitution of a residue important for domain interface stability will be characterized by a positive $\Delta\Delta G_{bind}$ value, indicated with a red dot in Figure 2-2. Large dots correspond to residues with $\Delta\Delta G_{bind}$ larger than 1.5 kcal/mol (chosen as a filter), while small dots correspond to residues with $\Delta\Delta G_{bind}$ between 1 and 1.5 kcal/mol. This analysis identified nine RfaH residues that display large IDI contact areas and $\Delta\Delta G_{bind}$: Phe51, Pro52, Phe81, Ile93, Leu96, Phe126, Phe130, Arg138, Leu142 (shown in magenta boxes in Fig. 2-2). Leu96 and Phe126 residues were filtered out because their Entropy (1 and 1.6, respectively) exceeded 0.9 (Table 2-1).

In summary, seven RfaH residues passed through the selected filters (Conservation score > 0.8; Entropy < 0.9, IDI contact area > 50 Å^2, $\Delta\Delta G_{\mathrm{bind}}$ > 1.5 kcal/mol). Among these residues, Ile93, Phe130, Arg138, and Leu142 have been proposed to play key roles in the stabilization of the IDI, based on computational and experimental evidence [68, 82-87].

Table 2-1. Properties of RfaH residues

position	Residue name	Conservation score	IDI contact area (Å^2)	Entropy	$\Delta\Delta G_{\mathrm{bind}}$ (kcal/mol)
2	gln	0.55	0.00	2.13	-8.61
3	ser	0.52	0.00	2.18	0.97
4	trp	0.97	0.00	0.00	0.88
5	tyr	0.94	0.00	0.27	0.80
6	leu	0.92	31.10	0.44	1.06
7	leu	0.87	0.00	1.18	0.98
8	tyr	0.71	7.11	0.90	0.74
9	cys	0.79	0.00	0.49	0.82
10	lys	0.98	0.00	0.14	0.86

Table 2-1. Properties of RfaH residues, continued.

position	Residue name	Conservation score	IDI contact area (Å^2)	Entropy	$\Delta\Delta G_{bind}$ (kcal/mol)
11	arg	0.65	0.00	1.00	0.89
12	gly	0.64	0.00	1.35	0.00
13	gln	0.86	0.00	0.68	0.78
14	leu	0.59	0.00	1.83	0.98
15	gln	0.48	0.00	2.31	0.98
16	arg	0.95	0.00	0.25	0.98
17	ala	0.98	0.00	0.01	0.99
18	gln	0.64	0.00	1.86	0.98
19	glu	0.73	0.00	1.63	0.93
20	his	0.83	0.00	0.46	1.01
21	leu	0.99	0.00	0.06	0.90
22	glu	0.65	0.00	1.93	0.92
23	arg	0.72	0.00	1.05	1.04
24	gln	0.99	0.00	0.01	1.02
25	ala	0.52	0.00	2.07	0.94
26	val	0.72	0.00	1.37	1.06
27	asn	0.59	0.00	2.06	0.92
28	cys	0.79	0.00	0.50	0.89
29	leu	0.70	0.00	1.55	0.96
30	ala	0.47	0.00	2.01	0.99
31	pro	0.98	0.00	0.05	1.06
32	met	0.57	0.00	1.76	0.91
33	ile	0.67	40.43	2.05	-0.21
34	thr	0.50	0.00	2.33	0.95
35	leu	0.60	38.35	2.34	1.02
36	glu	0.79	0.00	1.31	0.97
37	lys	0.84	0.00	0.81	1.41
38	ile	0.69	0.00	1.72	0.90
39	val	0.58	0.00	2.03	0.82
40	arg	0.74	0.00	1.23	0.80
41	gly	0.81	0.00	1.00	0.00
42	lys	0.83	0.00	1.17	0.89
43	arg	0.60	0.00	0.77	0.86
44	thr	0.54	0.00	1.94	0.68
45	ala	0.44	0.00	2.53	0.83
46	val	0.63	0.00	1.87	1.01
47	ser	0.49	0.00	2.56	0.95
48	glu	0.91	12.97	0.57	1.61
49	pro	0.79	5.07	0.71	0.91
50	leu	0.98	57.80	0.45	1.37
51	phe	1.00	79.82	0.01	3.06

Table 2-1. Properties of RfaH residues, continued.

position	Residue name	Conservation score	IDI contact area (Å^2)	Entropy	$\Delta\Delta G_{bind}$ (kcal/mol)
52	pro	0.99	55.46	0.02	1.71
53	asn	0.63	1.81	1.42	1.02
54	tyr	1.00	57.46	0.01	1.49
55	leu	0.92	0.00	1.07	0.97
56	phe	1.00	12.44	0.01	1.12
57	val	0.90	0.00	0.93	1.09
58	glu	0.59	0.00	2.01	1.13
59	phe	0.79	0.00	1.05	0.85
60	asp	0.87	0.00	0.80	0.97
61	pro	0.54	0.00	1.88	1.02
62	glu	0.62	0.00	1.76	1.00
63	val	0.31	0.00	2.40	0.97
64	ile	0.54	0.00	1.71	0.88
65	his	0.65	0.00	1.15	0.94
66	thr	0.59	0.00	1.16	0.94
67	thr	0.66	0.00	1.60	0.92
68	thr	0.67	0.00	1.29	0.97
69	ile	0.93	0.00	0.79	1.11
70	asn	0.65	0.00	1.48	0.86
71	ala	0.86	0.00	0.82	0.88
72	thr	0.99	0.00	0.05	1.02
73	arg	0.97	0.00	0.22	0.79
74	gly	1.00	0.00	0.01	0.00
75	val	0.98	0.00	0.31	0.96
76	ser	0.69	0.00	1.62	0.77
77	his	0.58	0.00	1.28	0.89
78	phe	0.77	0.00	0.97	0.78
79	val	0.97	16.59	0.53	0.75
80	arg	0.65	0.00	1.45	0.96
81	phe	0.88	96.40	0.75	3.88
82	gly	0.90	0.00	0.66	0.00
83	ala	0.41	0.00	2.25	0.98
84	ser	0.47	0.00	2.15	0.98
85	pro	0.87	0.00	0.79	0.96
86	ala	0.56	0.00	2.04	0.82
87	ile	0.45	36.80	2.15	0.07
88	val	0.94	8.46	0.74	0.92
89	pro	0.59	15.55	1.53	0.48
90	ser	0.54	14.15	2.18	0.29
91	ala	0.49	0.00	2.42	0.59
92	val	0.86	22.81	1.10	1.20

Table 2-1. Properties of RfaH residues, continued.

position	Residue name	Conservation score	IDI contact area (Å^2)	Entropy	$\Delta\Delta G_{bind}$ (kcal/mol)
93	ile	0.97	90.36	0.48	2.26
94	his	0.52	0.00	2.04	0.79
95	gln	0.64	0.00	1.78	0.91
96	leu	0.83	72.29	1.00	1.83
97	ser	0.50	22.94	1.87	0.25
98	val	0.37	0.00	2.46	0.98
99	tyr	0.26	0.00	2.16	0.88
100	lys	0.24	21.65	2.42	1.55
115	lys	0.48	0.00	2.51	-8.43
116	val	0.95	0.00	0.51	0.95
117	ile	0.51	0.00	2.41	0.95
118	ile	0.92	20.21	0.83	1.41
119	thr	0.62	0.00	1.81	1.09
120	glu	0.61	0.00	2.14	0.98
121	gly	0.91	0.00	0.54	0.00
122	ala	0.47	27.99	2.16	1.64
123	phe	0.89	20.67	0.75	1.57
124	glu	0.54	0.00	2.13	0.98
125	gly	0.73	0.00	1.12	0.00
126	phe	0.82	80.34	1.63	3.11
127	gln	0.75	0.00	1.67	0.97
128	ala	0.92	0.00	0.40	0.95
129	ile	0.97	15.36	0.37	0.29
130	phe	0.90	99.58	0.81	2.77
131	thr	0.48	1.46	2.21	0.95
132	glu	0.69	2.73	1.56	0.83
133	pro	0.55	22.76	1.92	1.15
134	asp	0.88	0.00	0.83	0.95
135	gly	0.96	21.35	0.20	0.00
136	glu	0.81	39.86	1.23	0.91
137	ala	0.54	0.00	2.02	0.99
138	arg	0.98	78.08	0.01	4.59
139	ser	0.75	44.41	1.18	0.43
140	met	0.78	0.55	1.41	0.97
141	leu	0.86	0.00	1.31	0.93
142	leu	0.97	60.64	0.33	1.63
143	leu	0.86	57.53	1.33	1.28
144	asn	0.76	0.00	1.21	1.00
145	leu	0.84	6.74	1.28	0.67
146	ile	0.90	81.28	1.04	1.01
147	asn	0.66	16.85	1.35	1.22

Table 2-1. Properties of RfaH residues, continued.

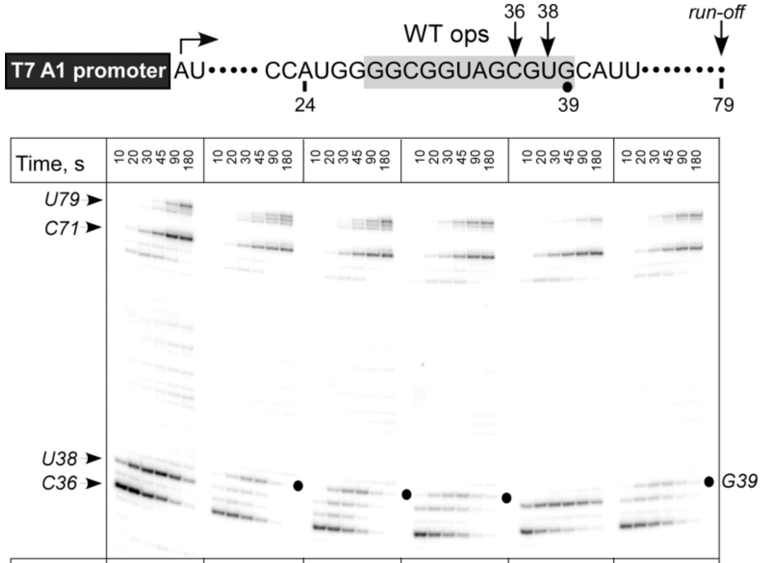
position	Residue name	Conservation score	IDI contact area (Å^2)	Entropy	$\Delta\Delta G_{bind}$ (kcal/mol)
148	lys	0.77	0.00	1.35	1.06
149	glu	0.67	23.32	1.50	1.89
150	ile	0.64	63.75	1.53	0.05
151	lys	0.46	19.18	2.39	2.18
152	his	0.53	0.00	1.97	0.73
153	ser	0.63	1.04	1.61	1.22
154	val	0.75	15.03	1.30	1.12
155	lys	0.50	0.00	1.68	0.83
156	asn	0.63	0.00	1.07	0.89

2-3-2. Identifying key residues that define RfaH and NusG subfamilies

Next, we sought to determine which of the seven selected residues are likely to be required for the formation of the RfaH-like closed state, and are thus different in NusG, in which the NTD and CTD do not interact [74]. To identify NusG residues at the positions corresponding to Phe51, Pro52, Ile93, Phe130, Arg138 and Leu142 in RfaH, we performed structural alignment of *E. coli* RfaH and NusG [88]. This analysis (Fig. 2-2) revealed that Phe51, Pro52, and Arg138 residues are identical between RfaH and NusG, and are therefore unlikely to make specific contributions to the autoinhibitory state of RfaH. By contrast, the remaining four residues differ between the two proteins. We next performed sequence alignment of 9204 bacterial NusG proteins (Fig. 2-2) to determine which of these residues should be selected for experimental validation. We found that NusG residues corresponding to RfaH Ile93 and Phe130 (Glu107 and Val148) are conserved in the alignment of NusG sequences (with Val or homologous Ile at position 148), whereas residues corresponding to Phe81 and Leu142 are not. Thus, we focused our functional analysis on Ile93 and Phe130, substituting these residues with Glu and Val, respectively, and testing the altered proteins in vitro. We expected that thus-altered RfaH proteins will have a weakened IDI and therefore sequence-independent, NusG-like recruitment to the TEC.

2-3-3. NusG-like RfaH variants are fully functional on an *ops*-containing template

We first tested the altered proteins during transcription in vitro. Because the affected residues are not involved in interactions with DNA or RNAP, the mutant proteins should be recruited to RNAP paused at the *ops* site similarly to the wild-type (WT) RfaH, as long as their structure is not altered. To test this, we carried out single-round elongation assays on a template that contains the WT *ops* element (Fig. 2-3). On this template, RNAP can be stalled at position A24 in the absence of UTP and restarted upon the addition of all NTPs. In the absence of transcription factors, RNAP pauses at C36 and U38 within the *ops* element, before making the full-length RNA of 79 nt; a strong arrest is observed at C71, likely because RNAP progression is hindered in the absence of the downstream duplex DNA [89]; pausing at these sites is accentuated at low [GTP], the incoming substrate, as used in this assay. Addition of wild-type RfaH or the isolated NTD reduces pausing at U38 ~3-fold, but delays RNAP 1 nt downstream, presumably via RfaH NTD-DNA interactions that must be broken to allow RNAP escape [72]; this delay is not sensitive to NTP concentrations. I93E and F130V RfaH variants exhibit similar behavior at U38 and G39 sites, whereas NusG does not. These results indicate that I93E and F130V substitutions do not interfere with RfaH recruitment to the TEC and antipausing modification of RNAP.



factor	none	RfaH^{WT}	RfaH^{I93E}	RfaH^{F130V}	NusG	RfaH^{NTD}
U38^{30s}	32.7	10.9	13.1	14.2	34.8	11.8
C71^{30s}	18.3	35.5	23.8	28.0	35.3	32.7

Figure 2-3. Effects of RfaH variants on pausing at the *ops* site. (Top) Transcript generated from the T7A1 promoter on a linear DNA template; transcription start site (a bent arrow), *ops* element (gray box), and transcript end are indicated on top. (Bottom) Halted A24 TECs were formed as described in Materials and Methods. Elongation was restarted upon addition of NTPs and rifapentin in the presence of the indicated transcription factor. Aliquots were withdrawn at times indicated above each lane (in s) and analyzed on an 8% denaturing gel. Positions of the paused and run-off transcripts are indicated with arrows; the position of the RfaH-induced pause at G39, with a circle. Pausing at *ops* (U38; fraction of total RNA) and arrival at the C71 position (fraction of final at 180 sec) were quantified to assess the anti-pausing effects of elongation factors; 30-s values are shown below each panel. The experiment was repeated three times; errors were <15%.

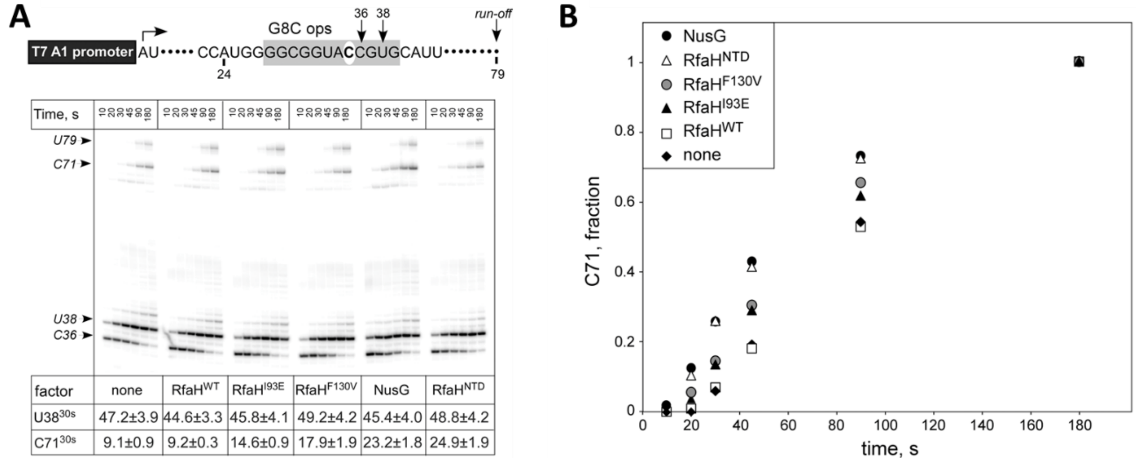
2-3-4. NusG-like RfaH variants can be recruited to TEC in the absence of *ops*

Our analysis suggested that Glu93 and Val130 could disfavor the autoinhibited state of RfaH, thereby facilitating sequence-independent (NusG-like) recruitment to RNAP. To test this hypothesis, we used a template in which an invariant *ops* residue G8 was substituted with C (Fig. 2-4). This substitution preserves the pausing pattern but abolishes recruitment to *ops*, and thus anti-pausing activity, of WT RfaH. By contrast, the isolated NTD and NusG increase the rate of RNAP elongation, leading to faster arrival at C71, a ~2.5-fold effect at the 30-sec timepoint (Fig. 2-4). In support of our prediction, I93E and

F130V RfaH variants exhibit intermediate phenotypes, speeding arrival at C71 1.6- and 2-fold, respectively. These results indicate that a single substitution of a key RfaH residue for its NusG counterpart is sufficient to allow for *ops*-independent recruitment. Conversely, this suggests that a single mutation in the nascent NusG duplicate could enable the formation of the silenced, autoinhibited state.

2-3-5. Probing RNAP-binding site accessibility by proteolysis

Our observations that RfaH I93E and F130V variants facilitate RNA synthesis on the mutant *ops* template (Fig. 2-4) are consistent with the hypothesis that these substitutions destabilize the domain interface, leading to spontaneous, *ops*-independent exposure of the RNAP-binding surface on the NTD. Similarly to the isolated NTD [73], these variants are prone to aggregation and precipitate at concentrations above 10 µM. The limited solubility of altered RfaH variants does not interfere with *in vitro* transcription analysis but hinders their structural characterization. Furthermore, the conformational transitions that accompany RfaH domain dissociation are complex, involving CTD refolding that may proceed via at least one intermediate [85].



factor	none	RfaH^{WT}	RfaH^{I93E}	RfaH^{F130V}	NusG	RfaH^{NTD}
U38^{30s}	47.2±3.9	44.6±3.3	45.8±4.1	49.2±4.2	45.4±4.0	48.8±4.2
C71^{30s}	9.1±0.9	9.2±0.3	14.6±0.9	17.9±1.9	23.2±1.8	24.9±1.9

Figure 2-4. Effects of RfaH variants on pausing in the absence of *ops*. (**A**) The experiment was performed as in Figure 3, except that a mutant *ops* element, with G8 substituted for a C (white oval), was used. (**B**) Arrival at the C71 position was quantified; the error bars are omitted for clarity. A representative example (30-s) is shown below in panel A, along with the fraction of U38 RNA; errors are standard deviations calculated from three repeats.

We therefore sought an approach to directly probe the accessibility of the RNAP-binding site on the NTD at low protein concentrations. The β’ clamp helices (CH) domain interacts with a cluster of aromatic residues in the NTD [73]; substitutions of these residues abolish RfaH recruitment [72]. To directly probe the solvent accessibility of this site, we used chymotrypsin, a serine protease that preferentially binds to and cleaves the C-termini of aromatic residues [90]. In full-length RfaH, all aromatic residues except Tyr99 are buried, whereas upon domain separation, the residues that comprise the RNAP-binding site on the NTD and at least two Phe residues on the CTD should become exposed and thus accessible to chymotrypsin (Fig. 2-5A).

The full-length WT RfaH was highly resistant to chymotrypsin, requiring large concentrations of protease for cleavage (visible on the gel; Fig. 2-5B). By contrast, the isolated domains were rapidly cleaved, confirming the utility of this approach. The I93E and F130V substitutions conferred increased susceptibility to chymotrypsin cleavage as compared to the WT RfaH (Fig. 2-5B). These results indicate that these substitutions weaken the domain interface, promoting CTD dissociation and subsequent RNAP binding. We note that while we cannot identify which form of the CTD is being cleaved (since Phe123 and Phe126 could be accessible in either the α- or β-state; Fig. 2-5A) by gel analysis, this approach could be adapted to monitor CTD folding by measuring the exposure of Phe130, which is part of the hydrophobic core of the β-barrel CTD [68].

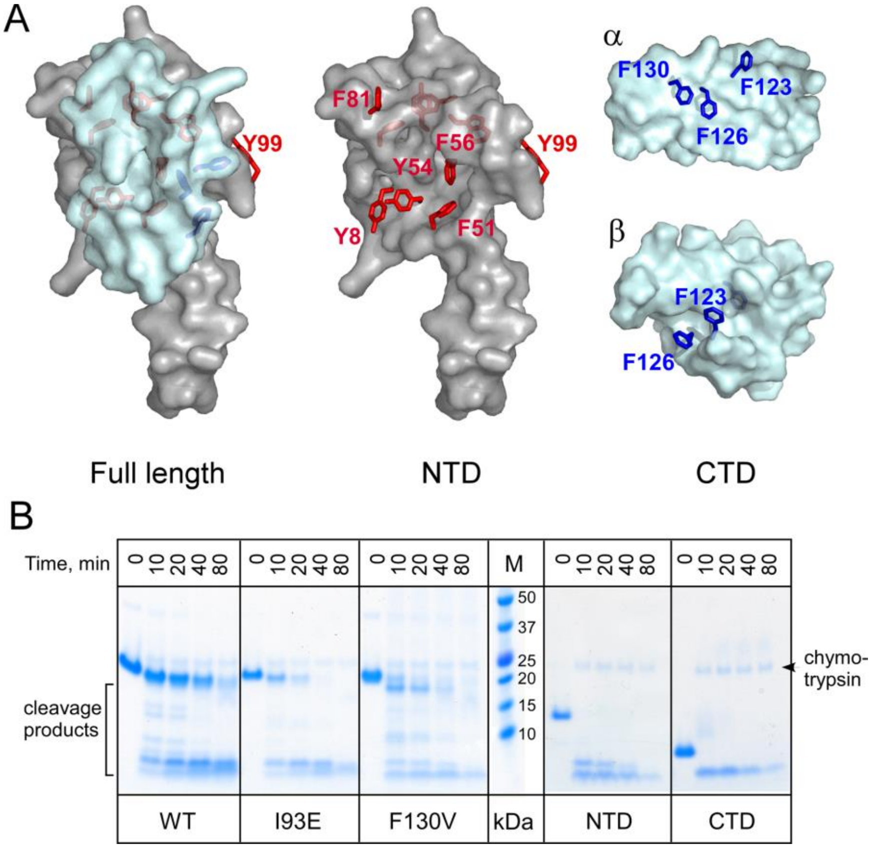


Figure 2-5. Probing the RfaH domain dissociation by chymotrypsin digestion. (**A**) Accessibility of aromatic residues in the full length RfaH and the isolated domains. The NTD is shown in gray and the CTD in cyan; both states of the CTD are shown. The aromatic residues are shown as sticks (red in the NTD; blue in the CTD), with their surfaces hidden. This figure was prepared with Pymol 1.8.2.3 (Schrödinger, LLC) using PDB IDs 2OUG and 2LCL. (**B**) Chymotrypsin cleavage of selected protein variants. The assays were performed as described in Materials and Methods; the samples were analyzed on 17-well 4–12% Bis–Tris gels. The WT, 193E and F130V samples were analyzed on one gel, and the isolated domains (along with the full-length protein, not shown) on another. Chymotrypsin is visible above the uncut proteins.

We argue that proteolytic enzymes are better suited for probing the accessibility of protein-binding interfaces than small molecules, *e.g.* hydrophobic dyes used in differential scanning fluorimetry [91]. Enzymatic probing can be carried out under conditions that mimic those used for functional assays (concentrations, temperature, etc.) and allows for a more realistic assessment of binding-site exposure to a large protein ligand.

2-4. Discussion

Autoinhibition is a widespread phenomenon that links protein activity to the presence of a cognate signal. During autoinhibition, intramolecular interactions between separate regions of a polypeptide negatively regulate its function, ensuring that activation is achieved only in response to proper physiological signals. Inhibition of ligand binding is the most common class of autoinhibition [92], where nucleic acid or protein interaction sites on a functional domain (FD) are masked by an inhibitory module (IM). Autoinhibition frequently modulates binding to DNA in transcription factors, such as σ^{70} [93] and Ets factors [94-95]. Evolution of an autoinhibited state was essential for the diversification of a nascent paralog of NusG, a housekeeping transcription elongation factor that regulates the synthesis of most cellular RNAs, into a dedicated regulator that controls just a handful of genes. In this study, we sought to identify the determinants of autoinhibition using *E. coli* RfaH, a highly specialized NusG paralog in which the relief of autoinhibition is achieved via interactions with a specific target DNA sequence presented on the surface of the elongating RNAP.

2-4-1. Structural determinants of RfaH autoinhibition

E. coli RfaH is a transformer protein that exists in two alternative states [52]. In the closed, autoinhibited state, the α-helical CTD masks the RNAP-binding site on the NTD. Interactions with the *ops* DNA induce opening of the RfaH IDI, releasing the CTD that subsequently refolds into a β-barrel. Our research has demonstrated that the stability of the RfaH IDI is responsible for the maintenance of the alternative α-helical CTD fold, autoinhibition, and resulting sequence specificity all lacked by its NusG-like ancestor [68, 73, 96]. Here, we show that the primary determinants of this increased stability can be identified through a synergistic approach unifying phylogenetic, structural, and biochemical evidence. This suggests that such an approach might prove useful in studying other examples of protein autoinhibition thought to be involved in many fundamental cellular signaling mechanisms [97], virulence [98-99], and disease states [100-103].

Here we have identified two RfaH residues, Ile93 and Phe130, predicted to be uniquely important for the IDI stability. We show that substitution of either residue for its NusG counterpart (I93E and F130V) alters the stability of the RfaH IDI so drastically as to convert the protein into a NusG-like regulator, with the loss of the sequence-dependent recruitment to the TEC characteristic of the former. It should also be acknowledged that many researchers, including ourselves, have studied the two native-state conformations of RfaH and potential mechanisms of interconversion between them using a variety of MD simulations. These simulations, to our knowledge, have only probed the thermodynamics and kinetics of RfaH (re)folding in the absence of DNA, the ligand that triggers the relief of autoinhibition. Nonetheless, they have yielded several testable predictions that our study has been able to validate and place within a broader context.

Chapagain and colleagues devised targeted and steered MD simulations showing that the breaking of contacts in the IDI presents the major thermodynamic barrier to the conversion of the RfaH CTD from α-helix to β-barrel, and also that Phe130 plays an important role in weakening of these contacts [83]. Our group reached the same conclusions independently using a dual-basin structure-based simulation [85]. Chapagain and colleagues also found that a nascent interdomain contact between Ile93 and Phe126 exposes an otherwise buried hydrophobic core in the NTD that prevents its binding to the β' CH domain [83]. These findings are supported by our demonstration of the importance of the Phe130 and Ile93 residues for IDI stability (Fig. 5) and autoinhibition (Fig. 4).

Still other studies explain not only why the Phe130 residue is so vital for RfaH-style functionality, but also why its substitution for valine proves so destructive. Valine and isoleucine residues strongly favor a β secondary structure to an α one [104], and F130V possesses a new valine residue adjacent to an isoleucine (at 129), increasing the propensity of the RfaH CTD to fold as a β-structure (the only one that the NusG CTD forms). Moreover, while three MD simulations using different methodologies, dual-basin structure based [85], Markov State Model and transition path theory [84], and coarse-grained off-lattice MD modeling [86], identified multiple candidate mechanisms for the α→β conversion of RfaH, all of these mechanisms had as their first step the formation of a β-sheet involving Phe130.

Our results also verify and build upon broader findings regarding the fundamental properties and regulation of autoinhibited proteins generally. A study by Gsponer and colleagues [97] found that when an interface exists between the FD and at least one IM (i) residues in the IM-FD IDI are conserved regardless of their diversity across homologs in the IM and (ii) intrinsically disordered IMs are preferable to structured ones since greater variation in intrinsic disorder should allow for fine-tuning of the equilibrium between active and inactive states on which the regulation depends. If we define the RfaH IM to include both its transformable CTD and the flexible linker (the NTD is of course the FD, as it confers the desired sequence-specific recruitment to the TEC), then our validation of (i) is apparent from the phylogenetic analysis (**Fig. 2A**) and the relief of autoinhibition resulting from changes of the IDI residues (**Fig. 4**). The recent μs-timescale MD simulation by Xun *et al*. demonstrated that two intrinsically disordered regions (IDRs) are necessary to stabilize the α-form of the CTD [87], with Phe130 making a contact with IDR1. The status of the linker as an IDR is supported by its tolerance to deletions and insertions and its absence from X-ray and NMR structures [68, 73], implying its flexibility. Thus, the available data validate (ii) as a key feature of IMs, exemplified by RfaH.

2-4-2. Autoinhibition in regulation of NusG-like proteins

While we have focused on converting RfaH into NusG, it is also interesting to ask the reverse question: could NusG be converted into RfaH, conferring autoinhibition in the process? Our results would indicate that if the IDI contacts can be made sufficiently strong, then the reverse conversion should be possible. Indeed, a recent report by Rösch and co-authors showed that *Thermotoga maritima* NusG is autoinhibited due to particularly strong IDI interactions absent from all other NusG variants yet found [105]. Interactions between the NTD and the β-barrel CTD of *T. maritima* NusG mask the binding sites for Rho, S10, and RNAP and must be broken to attain the active state. This autoinhibited state is argued to thermally stabilize the protein, rather than tune its regulatory properties, a function that may be critically important in the hyperthermophilic niche of *T. maritima* [105].

By contrast, autoinhibition is critical for delineating RfaH targets and conferring the dramatic activation of gene expression by RfaH. The closed state of RfaH masks the binding sites for both its cellular protein targets, RNAP and the ribosome. While the contact site with RNAP is merely masked by the IM, and can be exposed upon proteolytic removal of the CTD and part of the linker [73], the ribosome binding site is simply missing in the α-helical CTD. A complete refolding of the RfaH CTD into the β-barrel creates the interaction surface for S10 [68], with the resulting CTD-S10 complex closely resembling that formed by NusG [59]. This transformation is critical for RfaH function as it enables recruitment of the 30S ribosomal subunit to mRNAs that lack ribosome-binding sequences [68]; in fact, expression of a reporter gene can be made dependent on RfaH by adding the *ops* sequence and removing the ribosome binding sequence in front of heterologous reporter genes [68]. Dramatic activation of translation by RfaH is thought to insulate its target RNAs from premature termination by Rho [67], which silences these and other foreign genes in *E. coli* [64]. Curiously, *Clostridium botulinum* Rho has been recently reported to undergo a prion-like transformation that inhibits its function [106], highlighting the widespread role of dramatic conformational changes in regulation of bacterial gene expression.

Specialized NusG paralogs present in diverse bacterial phyla regulate expression of genes encoding biosynthesis of capsules in *K. pneumoniae* [69] and *Bacteroides fragilis* [107], toxins in *E. coli* [108] and *Serratia entomophila* [109], and antibiotics in *Myxococcus xanthus* [110] and *Bacillus amyloliquefaciens* [111]. Some of RfaH homologs are encoded on large conjugative multidrug-resistance plasmids and have been proposed to activate the pilus biosynthesis operons [53], by analogy to RfaH-mediated activation of the *tra* operon on F plasmid [108]. Thus, in addition to their well-established roles in virulence [69-71], RfaH-like regulators may also be essential for the spread of antibiotic-resistant genes. While these factors must function alongside ubiquitous NusG, it is not yet known if their recruitment to RNAP is regulated by autoinhibiton and if they can undergo transformation similarly to RfaH.

2-4-3. Broader impacts

The presence of autoinhibited proteins in key cellular signaling and virulence pathways and their association with a plethora of pathological conditions underlies the importance of better understanding their evolution, diversification, and regulation. Here we have combined experimental and computational techniques into an approach that can quantitatively and directly assess IDI stability and the primary determinants thereof, allowing unification/synthesis of disparate lines of evidence and showing a path towards the rational alteration or disruption of autoinhibited proteins for anti-virulent and other therapeutic ends.

2-5. Conclusion

In this chapter, we studied a bacterial transcription processivity factor, RfaH. RfaH is essential for the virulence and resistance gene horizontal transfer by enabling the transcription of some long genes. Besides the function essentiality, RfaH also has a unique dynamic structure in which one domain can adopt two distinct structures during its work cycle. These features make the RfaH indispensable for the virulence of bacteria to some extent and a promising target for new antibiotics. To study the structural transition of RfaH, we developed a method to dissect the inter-domain interaction based on phylogenetic and structural analysis. We showed that two residues E93 and F130 play key roles in the interaction.

2-6. Acknowledgement

Chapter 2, in full, has been published in **Shi Da**, Dmitri Svetlov, Ruben Abagyan, and Irina Artsimovitch. "Flipping states: a few key residues decide the winning conformation of the only universally conserved transcription factor." Nucleic acids

research (2017). The dissertation author was the primary investigator and author of this paper.

Chapter 3 Identification of a new target for bacterial infection. Part 2: Identification first-in-class inhibitor of RfaH

In the previous chapter, we introduced a bacterial transcription factor, RfaH. RfaH has been shown to be essential for the virulence and antibiotic resistance of Gram-negative bacteria by enabling the production of some long transcripts. RfaH adopts a unique dynamic structure during its work cycle where the C-terminal domain shows a dramatic structural transformation. We better understand this feature of RfaH, we developed a method to identify key residues for the domain interaction. Two residues E93 and F130 were successfully identified as the key players of this flipping structural states. In this chapter, we will further dive into the structure of RfaH and try to modulate its function with small molecules.

3-1. Introduction

Klebsiella pneumoniae, a leading cause of pneumonia in hospitalized patients, has been identified as an urgent public health threat by the World Health Organization and the US Centers for Disease Control and Prevention[112-113]. Rapidly spreading resistance to carbapenems, antibiotics of last resort, necessitate development of novel therapeutics effective against *K. pneumoniae* and other multidrug-resistant Gram-negative pathogens. In an attempt to identify novel *K. pneumoniae* targets for intervention, one study utilized transposon insertion mutagenesis to identify genes required for *K. pneumoniae* strain KPPR1 fitness in a murine model of pneumonia[69]. Among these genes, a null mutant of *rfaH* displayed a greater than 10,000-fold fitness defect in the lung, an effect surpassed only by the disruption of *wzi*, a gene from the capsular biosynthesis operon which is likely activated by RfaH[114]. Bachman *et al*. demonstrated that RfaH was required for capsule production and resistance to complement-mediated serum killing in KPPR1[69]. The contribution of RfaH to pathogen virulence is widespread, as it is known to be required in *Escherichia coli*[71], *Salmonella enterica serovar Typhimurium*[70], and possibly *Vibrio vulnificus*[115]. Additionally, RfaH paralogs encoded on conjugative plasmids could directly activate the spread of antibiotic-resistance genes encoded on these plasmids[116], and RfaH itself is essential for the antibiotic-resistance gene horizontal transfer in Cephalosporin Resistant *Escherichia coli*[117]. Mounting evidence identifies RfaH as a promising, wide-ranging target for drug discovery.

In addition to its essential role in virulence, RfaH utilizes a unique mechanism to activate both transcription and translation of its target genes. RfaH belongs to a ubiquitous NusG/Spt5 family that are the only known examples of universally conserved transcriptional regulators, in contrast to NusG which plays housekeeping roles, RfaH-like proteins are highly specialized[116]. RfaH homologs are required for expression of long operons that encode biosynthesis of capsules, LPS core, antibiotics, toxins and pili in diverse bacterial species ranging from *E. coli* to *Bacillus amyloliquefaciens*[111]. While limited functional information exists for most RfaHs, including that from *K. pneumoniae*, *E. coli* (Eco) RfaH is one of the best-characterized transcription factors. RfaH is recruited to the transcribing RNA polymerase (RNAP) through specific interactions with the single-stranded *ops* element in the non-template DNA strand within the transcription bubble[65]. Following recruitment, RfaH interacts with RNAP and the ribosome to activate expression of horizontally-acquired target genes, which are inefficiently translated and thus silenced by the transcription termination factor Rho. RfaH abrogates Rho-dependent termination by three mechanisms. First, RfaH inhibits RNAP pausing, which is a prerequisite to termination[65, 67]. Second, RfaH excludes NusG, which is required for the efficient RNA release at suboptimal Rho sites[118], via competition for a shared binding site[119]. Third, RfaH activates translation by recruiting the 30S ribosome subunit to mRNA via protein-protein contacts with S10[68]. This alternative recruitment mechanism is essential for expression of Eco RfaH-controlled operons which lack Shine-Dalgarno (SD) mRNA elements[68] and thus cannot recruit 30S via basepairing with a complementary sequence in the 16S rRNA[120].

The available high-resolution X-ray, NMR and cryo-EM structures of Eco RfaH alone or bound to the *ops* DNA, the transcription elongation complex (TEC) or the ribosomal protein S10[68, 73, 119, 121] could be used to guide the design of RfaH inhibitors. Many mutants of key amino acid residues of RfaH[72] and *ops* bases[121] as well as RNAP residues that make contacts to RfaH[67] are also available. Reporter assays and *in vitro* transcription assays of RfaH mechanism have been extensively validated[65, 68, 73]. In this work, we wanted to assess whether *K. pneumoniae* (Kpn) and Eco RfaH work similarly enough to justify the use of Eco RfaH as a template for the development of Kpn RfaH inhibitors. In particular, we wanted to find out whether all functional interactions characterized for Eco RfaH are essential in *K. pneumoniae*, thus representing shared targets for potential inhibitors.

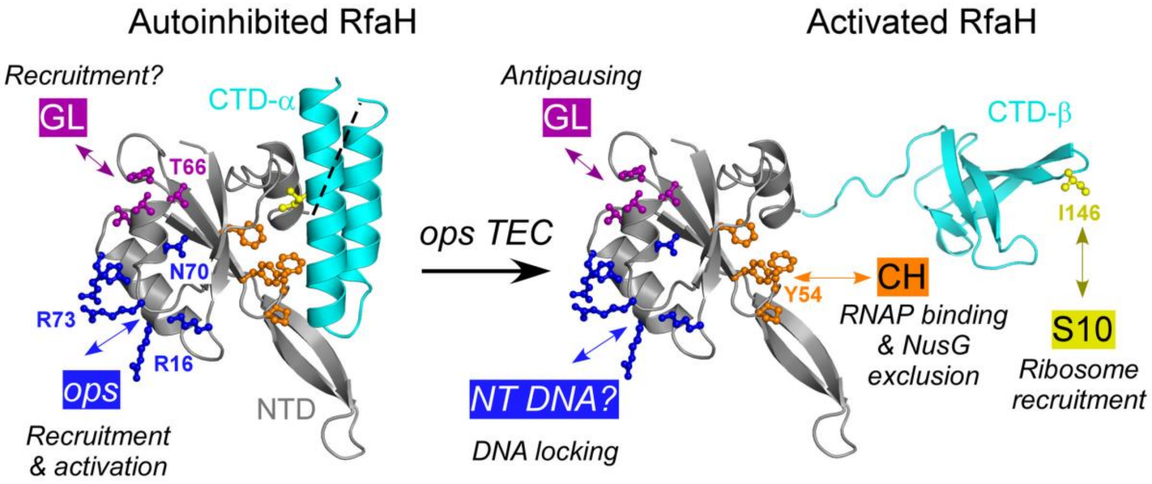


Figure 3-1. Key residues in Eco RfaH. In the autoinhibited state, the NTD (gray) and CTD (α-helices; cyan) interact to bury the NTD residues that bind to the β'CH domain (orange). In the active state, the domains are connected by a (modeled) flexible linker; the NTD and the β-barrel CTD bind to RNAP and S10, respectively. In both states, the NTD can interact with the nontemplate (NT) ops DNA element (blue) and the βGL (dark magenta).

Eco RfaH consists of two domains connected by a flexible linker. The N-terminal domain (NTD) interacts with RNAP in all characterized RfaH homologs. The C-terminal domain (CTD) has an unprecedented ability to reversibly switch folds between the α-helical hairpin that masks the RNAP-binding site on the NTD in the absence of the *ops* signal[73] and the β-barrel that interacts with S10[68]. The available data demonstrate that Eco RfaH activation of gene expression relies on four sets of interactions (Fig. 3-1). Binding of a cluster of charged NTD residues (that includes Arg73) to the *ops* DNA is necessary to relieve autoinhibition by triggering the dissociation of the CTD to expose the RNAP- and the ribosome-binding sites[73]. Binding of a hydrophobic surface (that includes Tyr54) to the β' clamp helices (β' CH) is thought to make high-affinity interactions that persist throughout transcription[66]. Contacts between the HTTT motif of RfaH (residues His65 through Thr68; Fig. 3-2) and the β subunit gate loop (GL) motif are required for antipausing effects[67]. Finally, contacts between the Leu145 and Ile146 CTD residues and the ribosomal protein S10 enable ribosome recruitment in the absence of an SD element[68]. While one could assume that these interactions are preserved in all RfaH orthologs, these proteins are quite divergent[122] and studies of the highly-conserved housekeeping paralog NusG revealed significant differences amongst bacterial species[63, 123-125].

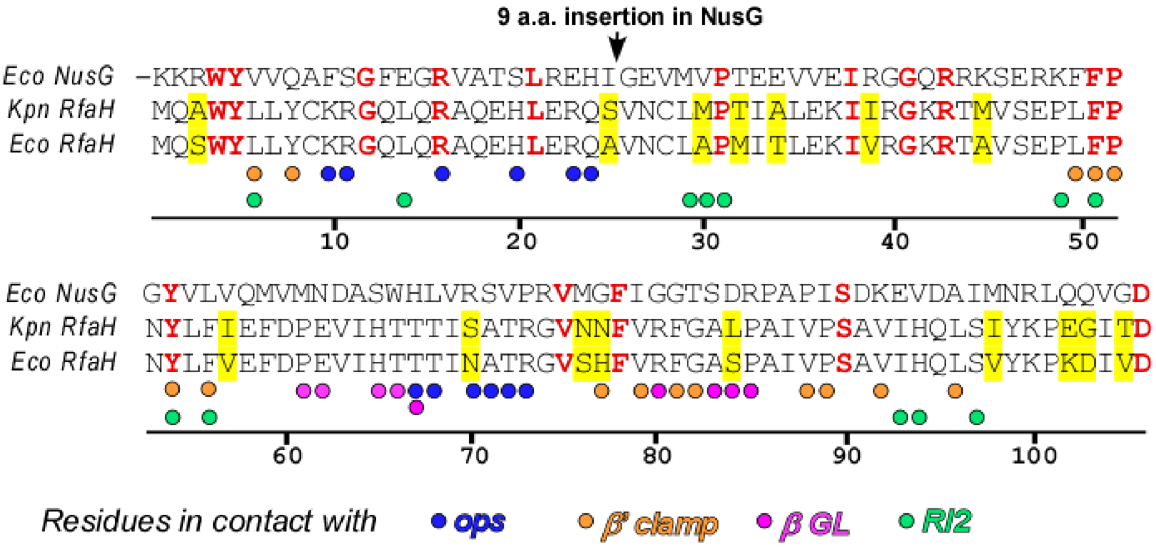

9 a.a. insertion in NusG
Eco NusG -KKRWYVVQAFSGFEGRVATSLREHIGEVMVPTEEVVEIRGGQRRKSERKFFP
Kpn RfaH MQAWYLLYCKRGQLQRAQEHLERQSVNCIMPTIALEKIIRGKRTMVSEPLFP
Eco RfaH MQSWYLLYCKRGQLQRAQEHLERQAVNCLAPMITLEKIVRGKRTAVSEPLFP
10
20
30
40
50
Eco NusG GYVLVQMVMNDASWHLVRSVPRVMGFIGGTSDRPAPISDKEVDAIMNRLQQVGD
Kpn RfaH NYLFIEFDPEVIHTTTISATRGVNNFVRFGALPAIVPSAVIHQLSIYKPEGITD
Eco RfaH NYLFVEFDPEVIHTTTINATRGVSHFVRFGASPAIVPSAVIHQLSVYKPKDIVD
60
70
80
90
100
Residues in contact with
● ops
● β' clamp
● β GL
● RI2

Figure 3-2. Structure-based alignment of the Eco RfaH, Eco NusG, and Kpn RfaH NTDs. The numbers below indicate RfaH residues; NusG NTD contains an N-terminal extension and a nine-residue insertion at the indicated position. Residues that are identical in all proteins are shown in red. Residues that differ between Eco and Kpn RfaHs (16 total in the NTD) are highlighted in yellow. Green dots indicate hypothetical contacts between the RfaH-NTD and RI2. Blue, orange, and magenta dots indicate RfaH contacts to the ops DNA, the β'CH, and βGL, respectively, observed in the cryoEM structure of RfaH bound to the E. coli ops TEC [119].

In this chapter, we found that the *ops* element and contacts to the β'CH and S10 are required for RfaH function in *K. pneumoniae in vivo*, whereas contacts to GL appear to be partially dispensable. This conclusion is supported by footprinting and *in vitro* transcription analyses. Based on these findings, we carried out an *in silico* search for small molecules that could interfere with RfaH interactions with RNAP. We successfully identify a lead molecule predicted to bind at the NTD-β'CH interface and demonstrate inhibition of Eco and Kpn RfaH recruitment to RNAP *in vitro*.

3-2. Methods and Materials

3-2-1. Tentative pocket identification and analysis

Crystal structure of RfaH (PDB code: 2OUG) was used to identify tentative binding pockets for virtual ligand screening. First, the CTD of RfaH (residues 115-156) was removed to unmask the β'CH interface of RfaH-NTD. Tentative pockets were identified using the ICMPocketFinder tool [126-127] of ICM-Pro [128], with default values of input parameter (tolerance = 4.6). Predicted pockets were analyzed by their volumes, position, and conservation of surrounding residues, which were used to prioritize pockets for docking. The conservation of each residue was evaluated as residue composition Entropy at each position in an alignment of RfaH sequences as reported earlier [129]. Briefly, Entropy was calculated according to Equation 3-1, where P_a^i is calculated from ratios of the observed frequency of amino acid a at position i in the sequence alignment over the expected frequency for the same amino acid, followed by normalizing the probabilities to the total value of one.

$$\text{Entropy of position i} = -\sum_{\mathrm{a}} \mathrm{P}_{\mathrm{a}}^{\mathrm{i}} \ln \mathrm{P}_{\mathrm{a}}^{\mathrm{i}}$$

Equation 3-1

3-2-2. Generation of RfaH ligand pocket models using SCARE

From identified tentative pockets, the pocket that has a large volume and conserved surrounding residues was selected for ligand screening. This binding site was used to generate an ensemble of modified conformations with the Scan Alanines and Refine (SCARE) approach in ICM [130]. SCARE generates 20 conformations where pairs of residues in the binding site are systematically masked out and restored around the bound ligand to mimic the induced fit effect. All conformations were combined to generate potential map ensembles for docking screening without further modifications. The potential map ensembles are calculated on a 0.5 Å 3D grid, containing: (i) van der Waals interaction; (ii) electrostatic interaction; (iii) hydrogen bond; and (iv) hydrophobic potential grids.

3-2-3. *In silico* screening for RfaH inhibitors

In silico screening was conducted by docking the ZINC library of 20 million small molecules to the pre-defined pockets on RfaH-NTD and ranking them based on their docking scores. The docking and scoring of one molecule was conducted using a stochastic global energy optimization procedure in internal coordinates [79] implemented in the ICM-Pro v3.8-6a. Ligand docking started with generating multiple starting conformations of the ligand by sampling it *in vacuo* and placing each sampled conformations to the binding pocket with four principal orientations. Then the ligand was sampled in the pre-calculated potential map ensembles through biased probability Monte Carlo method to optimize the position and internal variables of the ligand. For each ligand, 10 top ranking conformations were re-scored with ICM full atom scoring function [131], and conformation with the best docking score was kept for comparison. ZINC molecules were pre-filtered with the molecular weight between 100 to 1000, and ICM ToxScore smaller than 1.5. ICM ToxScore of a compound was calculated based on the number of present bioactive chemical fragments that were identified as structural alerts [132-133]. Then the compounds were docked to the selected pockets on RfaH-NTD following the procedures described above with a computing cluster containing 128 cores. After docking, chemicals were ranked by their docking scores and the top ten hits were tested in *in vitro* RfaH assay.

3-2-4. Identification of RfaH residues interacting with RI2 in the predicted model

After the docking screen, the model of RI2 binding to RfaH was refined by restrained energy optimization of the full atom model. To identify residues of RfaH interacting with RI2, contact areas of residues with RI2 were calculated as the difference of areas of molecular surface of each residue with and without RI2. Big contact areas implied stronger interactions between residues and RI2. The contact area threshold of 15 Å^2 identified 12 residues on RfaH-NTD potentially interacting with RI2.

3-2-5. Reagents and proteins

All general reagents were obtained from Sigma Aldrich (St. Louis, MO) and Fisher (Pittsburgh, PA); NTPs, from GE Healthcare (Piscataway, NJ); [γ^{32}P]-ATP and [α^{32}P]-GTP, from Perkin Elmer (Boston, MA); PCR reagents and modification enzymes, from NEB (Ipswich, MA). Oligonucleotides were obtained from Integrated DNA Technologies (Coralville, IA) and Sigma Aldrich. DNA purification kits were from Qiagen (Valencia, CA). *E. coli* RNAP [134], RfaH [73], NusG [66], Rho [66], and GreB [135] were purified as described previously. Kpn RfaH was purified as in [122]. Plasmids and oligonucleotides are listed in Table S1.

3-2-6. Structural probing of the ops TEC

Scaffolds were assembled as described previously [136]. The template DNA strand was end-labeled with [γ^{32}P]-ATP using T4 polynucleotide kinase (PNK; NEB). The assembled TEC were resuspended in TB40 (20 mM Tris-Cl, 5% Glycerol, 40 mM KCl, 5 mM $MgCl_2$, 10 mM β-mercaptoethanol, pH 7.9). For Exo III probing, was divided in two aliquots; one was incubated with 100 nM RfaH and the other – with storage buffer for 3 min at 37 °C. For each time point, 5 µl EC were mixed with 5 µl of Exo III (NEB, 40 U) and incubated at 21 °C. At times indicated in figure legends, the reactions were quenched with an equal volume of Stop buffer (8 M Urea, 20 mM EDTA, 1 x TBE, 0.5 % Brilliant Blue R, 0.5 % Xylene Cyanol FF). For psoralen crosslinking, the TECs were supplemented with 6.3 % DMSO and 0.92 mM 8-MP and incubated for 2 min at 37 °C, followed by addition of 100 nM Eco RfaH, 250 nM Kpn RfaH, or storage buffer and a 3-min incubation at 37 °C. Complexes were then exposed to 365 nm UV light (8W Model UVLMS-38; UVP, LLC) for 20 min on ice. The reactions were quenched as above.

3-2-7. In vitro transcription assays

RfaH recruitment assays were performed as described previously [121]. Templates were made by a two-step PCR on pIA1087 plasmid that encodes the wild-type *ops* signal. Linear DNA template (30 nM), holo RNAP (40 nM), ApU (100 μM), and starting NTP subsets (1 μM GTP, 5 μM ATP and CTP, 10 μCi [α^{32}P]-GTP, 3000 Ci/mmol) were mixed in 100 μl of TGA2 (20 mM Tris-acetate, 20 mM Na-acetate, 2 mM Mg-acetate, 5% glycerol, 1 mM DTT, 0.1 mM EDTA, pH 7.9). Reactions were incubated for 14 min at 37 °C; thus halted TECs were stored on ice. RfaH (100 nM final concentration or an equal volume of storage buffer) was mixed with RIs (at concentrations indicated in figures, or DMSO), chase NTPs (20 μM GTP, 300 μM ATP, CTP, and UTP) and rifapentin (50 μg/ml) in TGA2, followed by a 3-min incubation at 37 °C. Equal volumes of prewarmed at 37 °C halted A24 TEC and RfaH/RI2/NTP mix were combined, followed by incubation at 37 °C. Samples were removed at time points indicated in the figures and quenched by addition of an equal volum*e* of 10 M urea, 60 mM EDTA, 45 mM Tris-borate; pH 8.3. GreB-mediated cleavage [135] and Rho-dependent termination [66] assays were performed as described previously, with modifications indicated in Figure 3-10 legend. Samples were heated for 2 min at 95 °C and separated by electrophoresis in denaturing acrylamide (19:1) gels (7 M

Urea, 0.5X TBE). The gels were dried and the products were visualized and quantified using a FLA9000 Phosphorimaging System (GE Healthcare), ImageQuant Software, and Microsoft Excel.

3-2-8. Capsule quantification assays

Capsule extraction and uronic acid quantification were performed using a modified protocol [137-138]. 20 ml cultures of *K. pneumoniae* TOP52 or TOP52Δ*rfaH* [139] transformed with pIA947 (empty vector), pIA957 (Eco RfaH) or pIA1282 (Kpn RfaH) (Table 3-1) were grown shaking for 16 h in Luria–Bertani (LB) broth. Cultures were titered to determine colony forming units (CFU)/ml for normalization. 500 µl of each culture was mixed with 100 µl of 1% Zwittergent 3–14 in 100 mM citric acid and incubated at 50 °C for 20 min. Samples were centrifuged at 10,000 x *g* for 5 min, and 300 µl of each supernatant was precipitated with 1 mL cold ethanol for 20 min at 4°C. After centrifugation at 10,000 x *g* for 5 min, the pellet was dissolved in 200 µL water, and 1.2 mL of 12.5 mM sodium tetraborate in concentrated sulfuric acid was added. Samples were vortexed, boiled at 95°C for 5 min, and mixed with 20 µL of 0.15% 3-phenylphenol in 0.5% NaOH. Absorbance was measured at 520 nm and divided by bacterial titer to determine and absorbance/10^8 CFU; assays were performed in triplicate. Relative capsule production was determined by dividing all absorbance/10^8 CFU values by that of TOP52Δ*rfaH*.

Table 3-1. Strains, plasmids and oligonucleotides

Strains		
Name	Features	Reference
MG1655	Wild-type *E. coli*	N. Ruiz
IA228	MG1655 *ΔrfaH*	[140]
TOP52	Wild-type *K. pneumoniae*	[139]
TOP52*ΔrfaH*	*rfaH* knockout in TOP52	[139]
Plasmids		
Name	Key features	Reference
pIA947	P_{trc}–no insert; control plasmid; P_{lacIQ1}-*lacI*. P15A origin, Cm^R	[72]
pIA957	P_{trc}–*E. coli* RfaH; P15A origin, Cm^R	[72]

Table 3-1. Strains, plasmids and oligonucleotides, continued.

Plasmids		
Name	Key features	Reference
pIA1001	P_{trc}–*E. coli* RfaH T66A; P15A origin, Cm^R	[72]
pIA1003	P_{trc}–*E. coli* RfaH R73D; P15A origin, Cm^R	[72]
pIA1005	P_{trc}–*E. coli* RfaH R16A*;* P15A origin, Cm^R	[72]
pIA1006	P_{trc}–*E. coli* RfaH Y54F*;* P15A origin, Cm^R	[72]
pIA1094	P_{trc}–*E. coli* RfaH I146D; P15A origin, Cm^R	[68]
pIA1282	P_{trc}–*K. pneumoniae* RfaH*;* P15A origin, Cm^R	This work
pHK2	P_{BAD}–*ops*–TC_{15}–*luxCDABE;* pSC101 origin, Sp^R	[140]
pIA1283	P_{BAD}–*ops*–*luxCDABE;* pSC101 origin, Sp^R	This work
pIA1297	P_{BAD}–*ops* G8C–*luxCDABE;* pSC101 origin, Sp^R	This work

Oligonucleotides		
Name	Features and sequence	Reference
2536	T7A1 promoter primer; -35, -10 and transcription start site are underlined AAAAAGAGTATTGACTTAAAGTCTAACCTATAGGATACTTACAGCCATCGAGCAGGCAGCGGCAAAGCCATGG	[121]
2537	Downstream primer: AAATAAGCGGCTCTCAGTTT	[121]
2499	Upstream primer step 2: AAAAAGAGTATTGACTTAAAG	[121]
NT43	Scaffold assembly, nontemplate DNA strand GAAACACCACCAGTAGGCGGTAGCGTGCGTTTTTCGTTCTTCC	[136]
T43	Scaffold assembly, template DNA strand GGAAGAACGAAAAACGCACGCTACCGCCTACTGGTGGTGTTTC	[136]
R43	Scaffold assembly, RNA strand UUAUUCGGUAGCGU	[136]
2141	Upstream Rho terminator primer GTGATAATGGTTGCATGTAGTAAGGAGGTTGTATGGAAGACCGGTAACATTAATCAACGCGTT	This work
2142	Downstream Rho terminator primer GCGCCTGCAACCGCTGAAATTTG	This work
2143	Upstream λ P_R promoter primer; -35, -10 and start site are underlined CTAACACCGTGCGTGTTGACTATTTTACCTCTGGCGGTGATAATGGTTGCATGTAG	This work

3-2-9. In vivo lux assays

Plasmids carrying RfaH variants were co-transformed with a *lux* reporter vector (pHK2, pIA1297, or pIA1293) into *K. pneumoniae* strain TOP52Δ*rfaH* and plated on LB agar containing 50 μg/ml spectinomycin and 20 μg/ml chloramphenicol. Strains containing both a *lux* reporter and an RfaH variant were grown overnight at 37°C shaking in LB broth. Cultures were subcultured 1:100 into 10 ml of LB broth containing antibiotics and incubated at 37°C with agitation for 6 hours. Neither construct required induction (with IPTG or arabinose), since background expression of *rfaH* and *lux* operon was sufficient to generate signal. 200 µl of each culture was added in triplicate to a black polysterene 96 well plate with clear bottoms (Corning 3904). A Biotek Synergy 2 plate reader was used to measure luminescence, with an integration time of 1 sec and a vertical offset of 5 mm. Luminescence was corrected for the cell densities of individual cultures.

3-3. Results

3-3-1. Eco RfaH and Kpn RfaH substitute for each other in vivo

In contrast to Eco RfaH, which has been extensively studied, little mechanistic information is available for its orthologs, even in closely related species. In our early work, we demonstrated that Eco and Kpn RfaHs had similar stimulatory effects on the expression of the plasmid-borne hemolysin (*hly*) operon, the best characterized Eco RfaH target at the time, in *E. coli* [122]. In this operon, RfaH appears to reduce termination at an unusual weak hairpin-dependent terminator between the *hlyA* and *hlyB* genes [141], an effect that is distinct from other characterized RfaH-dependent operons in which RfaH counteracts Rho-mediated polarity [116]. We wanted to test whether Eco and Kpn RfaH proteins have similar effects on the expression of chromosomal operons activated by RfaH. In *E. coli*, deletion of *rfaH* confers dramatic sensitivity to SDS, an effect that is phenocopied by an early polar mutation in the RfaH-activated *waa* LPS biosynthesis operon [142] and suppressed by mutations in *rho* [140]. In *K. pneumoniae*, the deletion of *rfaH* leads to decreases in capsule production [69]; similar effects, attributed to a significant similarity between the capsule biosynthesis clusters, were observed in *E. coli* [143]. Eco RfaH has been shown to inhibit Rho-dependent termination within capsule operons [144].

We expressed Eco and Kpn RfaH from an IPTG-inducible P_{trc} promoter on a plasmid and tested whether they complemented the SDS sensitivity and abrogated capsule production phenotypes in Δ*rfaH E. coli* and *K. pneumoniae* strains, respectively. We observed that both proteins behaved indistinguishably in these assays (Fig. 3-3). The *E. coli* MG1655 strain lacking *rfaH* was unable to grow at 0.5 % SDS, whereas the induction of either Eco or Kpn RfaH restored growth to the levels observed with the wild-type MG1655 (Fig. 3-3A); no growth was observed with an empty vector or in the absence of IPTG (not shown). Similarly, expression of either Eco or Kpn RfaH complemented the loss of the chromosomal gene, restoring capsule production in *K. pneumoniae* TOP52 (Fig. 3-3B). We conclude that Eco and Kpn RfaH proteins act similarly in both species.

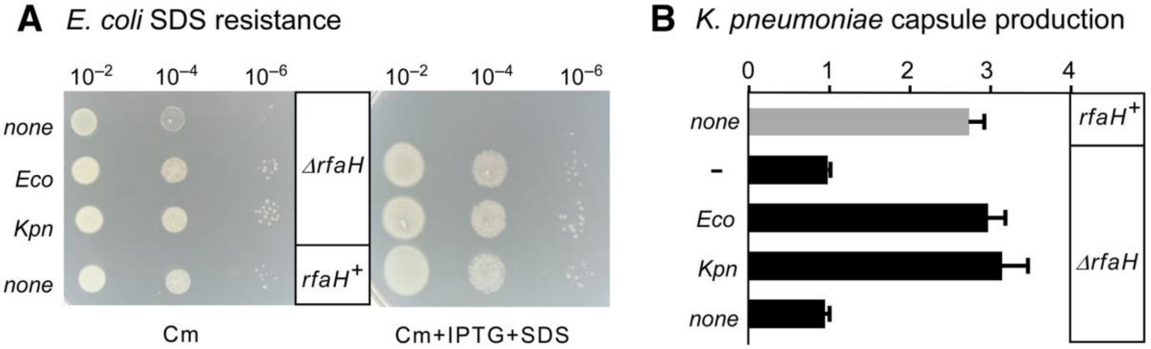


Figure 3-3. Plasmid-encoded Eco and Kpn RfaH complement rfaH deletions in E. coli and K. pneumoniae. A. Dilutions of exponentially growing cultures of MG1655ΔrfaH strain transformed with plasmids expressing Eco RfaH, Kpn RfaH, or a control vector were plated on LB-chloramphenicol (left) or LB-Cm supplemented with 0.5% SDS and 0.2 mM IPTG (right) and incubated at 37°C overnight. A representative set from three independent experiments is shown. B. Relative capsule production in K. pneumoniae TOP52 or TOP52ΔrfaH strains transformed with plasmids containing Eco RfaH, Kpn RfaH, or no insert. Data are combined from three independent experiments, normalized to TOP52ΔrfaH without an RfaH plasmid, and error bars represent standard deviation.

3-3-2. Contributions of key RfaH regions to in vivo activity in K. pneumoniae

We next wanted to determine whether all Eco RfaH regions identified previously as critical for gene activation in *E. coli* are also necessary for its activity in *K. pneumoniae*.

We first tested the ability of plasmid-borne wild-type Eco and Kpn RfaHs to activate expression of a *lux* reporter in a Δ*rfaH K. pneumoniae* TOP52 strain. In this reporter (Fig. 3-4A), the *Photorhabdus luminescen*s *lux* operon is positioned downstream from an *ops* element, which is known to recruit both Eco and Kpn RfaH [114, 122]. We have used a nearly identical reporter to identify the key functional residues of Eco RfaH [72]; in this work, we switched the antibiotic-resistance determinants to enable experiments in ampicillin-resistant *K. pneumoniae*. We observed that, similarly to their effects on activation of LPS and capsule biosynthesis operons (Fig. 3-3), the wild-type Eco or Kpn RfaH led to similar increases in *lux* expression (Fig. 3-4A).

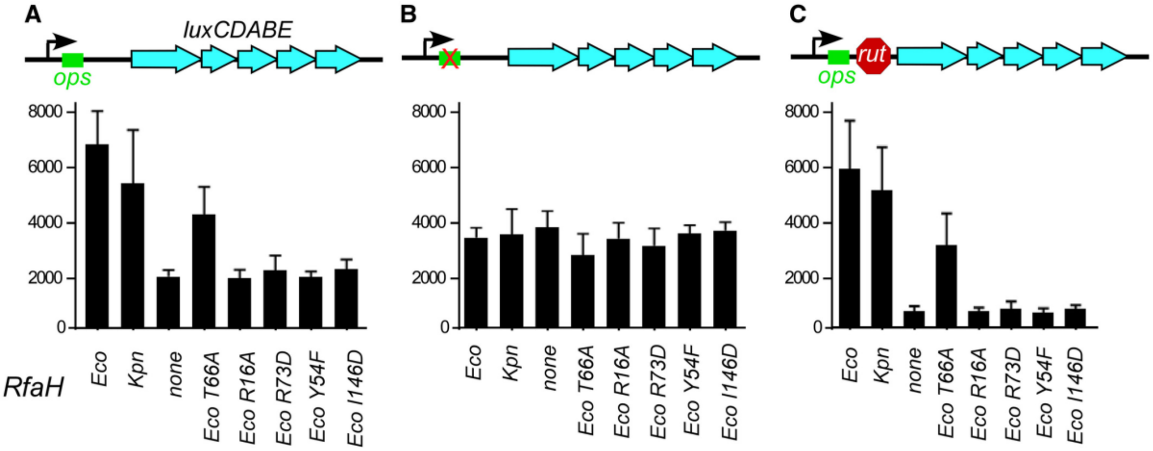


Figure 3-4. Reporter assays in K. pneumoniae. Plasmids encoding wild-type RfaH proteins or Eco RfaH variants with single-residue substitutions under the control of P_{trc} promoter were co-transformed into TOP52ΔrfaH strain with reporter vectors containing the Photorhabdus luminescens luxCDABE operon under the control of P_{BAD} promoter, with ops and rut elements in the leader region as indicated in the schematics. The results are expressed as luminescence corrected for the cell densities of individual cultures. Data are combined from three independent experiments and error bars represent standard deviation.

We have shown that substitutions of RfaH residues that mediate contacts with β'CH (Tyr54), *ops* DNA (Arg73), βGL (Thr66) and S10 (Ile146) (Fig. 3-1) abolish RfaH-dependent activation in *E. coli* [68, 72]. Here we tested whether these RfaH regions contribute similarly to its activity in *K. pneumoniae*. As could be expected, disruptions of contacts with β'CH, S10 and *ops* DNA reduced *lux* activity to background levels observed in the absence of RfaH (Fig. 3-4A). In contrast, the loss of contacts with βGL led to only a small decrease in the *lux* expression (Fig. 3-4A).

RfaH orthologs from a variety of bacteria, including those from *V. cholerae* and *E. coli* which are only 43% identical, bind to the *ops* element *in vitro* [122]. Together with conservation of *ops* sequences in diverse bacteria [145], this suggests that even phylogenetically diverse RfaHs make functionally important contacts to *ops*. To confirm this conclusion, we tested whether a G8C substitution in the *ops* element, which eliminates Eco RfaH recruitment to the TEC *in vitro* [129], interferes with RfaH function in *K. pneumoniae in vivo*. We found that the *lux* expression from a reporter in which the *ops* element contained a G8C substitution was reduced with all RfaH variants to the levels observed with an empty vector (Fig. 3-4B).

Eco RfaH has been shown to reduce Rho-dependent termination *in vitro* and *in vivo* [65, 67]. To test whether RfaH could overcome the effects of Rho in *K. pneumoniae*, we used a reporter in which $(TC)_{15}$, a synthetic Rho-utilization (*rut*) signal, was placed between the *ops* site and the *lux* operon (Fig. 3-4C). Consistent with $(TC)_{15}$-induced Rho-dependent termination, *lux* expression from this reporter was reduced ~five-fold in the absence of RfaH. Wild-type Eco or Kpn RfaH restored *lux* expression to a level observed in the absence of $(TC)_{15}$. Substitutions of Eco RfaH residues interacting with *ops*, β'CH and S10 failed to activate expression, whereas the T66A RfaH variant restored *lux* activity to ~50% of the levels obtained with the wild-type Eco RfaH (Fig. 3-4C). We conclude that Eco and

Kpn RfaHs inhibit Rho-dependent termination in *K. pneumoniae*, with three out of four key functional regions of Eco RfaH being essential for antitermination.

These results suggest that while Eco and Kpn RfaH proteins require the *ops* element for recruitment and utilize similar mechanisms to activate gene expression, they display one significant mechanistic difference: the disruption of RfaH-βGL contacts is less detrimental for RfaH-mediated activation of gene expression in *K. pneumoniae*, as compared to *E. coli* [72].

3-3-3. Kpn RfaH fails to lock the non-template DNA in the TEC

In the *ops*-paused TEC, the single-stranded non-template DNA in the transcription bubble simultaneously interacts with Eco RfaH and the βGL; these contacts are further stabilized by interactions between the βGL and RfaH HTTT motif [119]. Disruption of Eco RfaH-βGL interactions do not abolish RfaH recruitment to RNAP but eliminate antipausing activity [67]. We proposed that RfaH and βGL act together to constrain the non-template DNA strand to prevent it from assuming nonproductive conformations during elongation, thereby reducing pausing and facilitating transcription [146]. By locking the non-template DNA, RfaH and βGL restrict the mobility of the upstream duplex DNA, inhibiting digestion by Exo III, a double-strand DNA specific exonuclease that digests DNA in a 3'→5' direction. When bound to the TEC, Eco RfaH confers protection of 12 bp of DNA upstream of the transcription bubble (Fig. 3-5A), as compared to 5 bp in a free TEC [136]. Disruption of RfaH-βGL contacts by the deletion of βGL or substitutions of the HTTT residues weaken the RfaH-dependent Exo III protection [136]. If Kpn RfaH makes fewer stable contacts with the βGL, it would be unable to hinder Exo III digestion of the upstream DNA duplex.

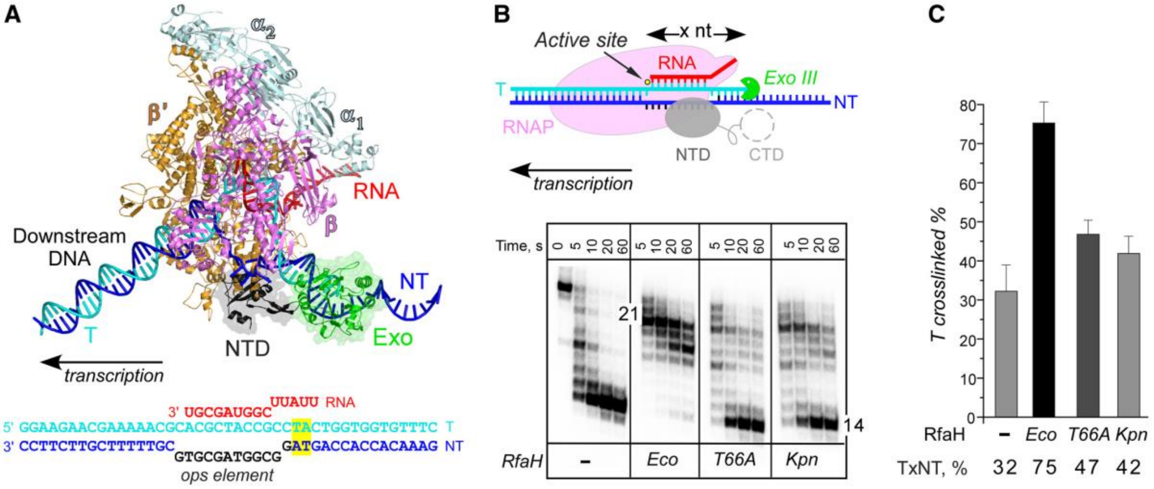


Figure 3-5. Probing RfaH-DNA interactions. A. A model of the RfaH-bound TEC**.** RNAP α (pale cyan), β (magenta) and β' (orange) subunits, RfaH-NTD (gray), nucleic acids and Exo (green) are shown as cartoons. To provide an unobstructed view of the RfaH-NTD and the exposed nontemplate DNA, the EC is shown in an orientation that is opposite to the conventional left-to-right direction of transcription. The ops TEC scaffold used in these experiments is shown below, with nucleic acid chains colored and oriented as in the model; the ops element is in black. The upstream TA cross-linking motif is highlighted in yellow. B. Footprinting of the upstream RNAP boundary. The template strand DNA was 5'-end labeled with [γ^{32}P]-ATP. After the addition of Exo III, aliquots were quenched at the indicated times (0 represents an untreated DNA control) and analyzed on a 12% denaturing gel; a representative of three independent experiments is shown. Numbers indicate the distance from the RNAP active site (yellow circle). C. Probing the upstream fork junction by cross-linking with 8-MP. TECs were supplemented with 100 nM Eco or 250 nM Kpn RfaH (where indicated) and illuminated with the 365 nm UV light. Fractions of the cross-linked DNA were determined after analysis on denaturing gels. Error bars indicate the SDs of triplicate measurements. See also Fig. 3-6.

To test this idea, we assembled *ops* TECs on a nucleic-acid scaffold containing the *ops* element (Fig. 3-5A). The template strand and the nascent RNA were end-labeled with γ^{32}P-ATP. We incubated the assembled TECs with RfaH variants (or storage buffer) and then added Exo III. Samples were quenched following incubation for indicated times and analyzed on denaturing urea-acrylamide gels. As expected, in the absence of RfaH, RNAP protected 14 nts of the template DNA strand upstream from the RNAP active site from Exo III digestion (Fig. 3-5B). When added, Eco RfaH strongly protected the upstream DNA from digestion; the footprint boundary was extended by 7 nt, whereas Eco RfaH with the

T66A substitution failed to confer protection, as observed previously [136]. Kpn RfaH displayed an intermediate phenotype even though it was recruited to RNAP as well as the Eco RfaH at the same concentration (see below, **Fig. 6**).

The results of Exo III footprinting suggest that the DNA-lock mechanism is partially disabled in Kpn RfaH. By constraining the upstream DNA duplex, RfaH-βGL contacts could also stabilize the upstream edge of the transcription bubble. To test whether Kpn RfaH is less efficient in stabilizing the upstream fork junction, we used crosslinking with 8-methoxypsoralen (8-MP). 8-MP specifically intercalates into double-stranded 5'-TA-3' motifs and introduces a T-T inter-strand crosslink upon exposure to UV light. Crosslinking is strongly increased by Eco RfaH [136]. We assembled *ops* TEC on a scaffold with a TA motif positioned 12 nucleotides upstream of the RNA 3' end (Fig. 3-5A and Fig. 3-6), with 5'-labeled template DNA and RNA (the latter is used as a loading control). We induced crosslinking upon addition of 8-MP and exposure to 365 nM UV light (Fig. 3-5C) and monitored the inter-strand crosslinking efficiency by gel electrophoresis in denaturing urea-acrylamide gels. Upon addition of Eco RfaH, crosslinking increased more than two fold; 100 nM RfaH (Fig. 3-5C) gave the same effect as 25 or 50 nM [136]. By contrast, T66A RfaH and Kpn RfaH (at 100 nM) increased crosslinking only modestly (Fig. 3-5C); the efficiency did not increase even at 1000 nM Kpn RfaH (Fig. 3-6). The observed small effects are similar to that of Eco NusG, which acts independently of βGL [147].

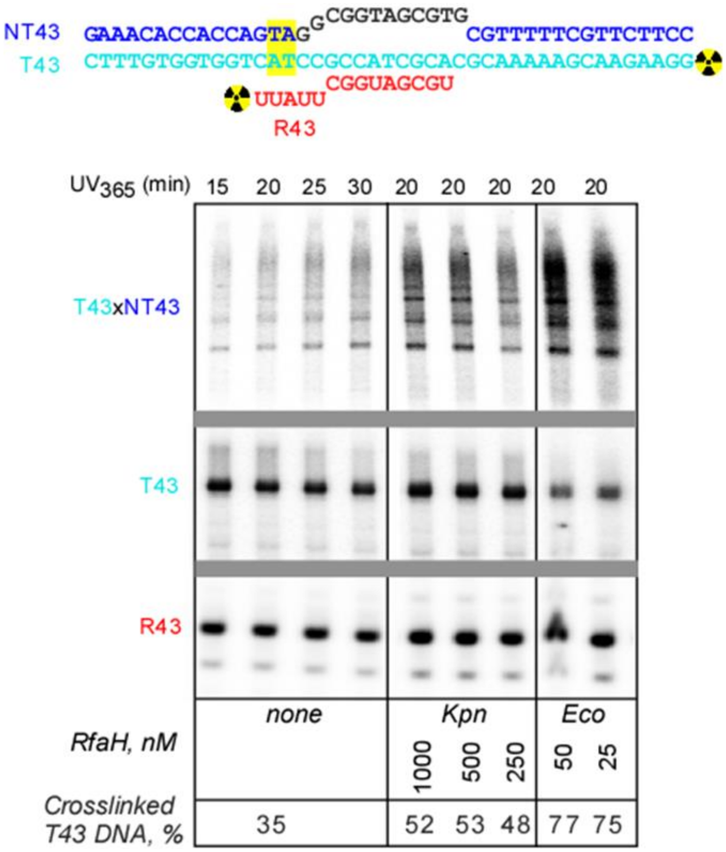


Figure 3-6. Assaying the upstream fork junction by psoralen crosslinking. TECs were assembled on the scaffold shown on top, with the template DNA and RNA strands labeled with [γ^{32}P]-ATP. TECs were supplemented with Eco or Kpn RfaH (at indicated concentrations) or storage buffer and illuminated with the 365 nm UV light on ice for 15-30 min (as shown). Samples were mixed with an equal volume of stop buffer and analyzed on a denaturing 12% acrylamide-urea gel. The positions of the RNA, free template DNA (T43) and crosslinked species are indicated. Fractions of the template strand DNA crosslinked after 20 min incubation are indicated below.

Together, these results are consistent with less stable interactions between the βGL and Kpn RfaH. Observations that Kpn RfaH is recruited to Eco RNAP at 100 nM (**Fig. 6**) excludes the binding defect as an explanation for weak Exo III protection and upstream duplex stabilization. At present, we do not know the basis for the observed differences. The βGLs are identical between Eco and Kpn RNAPs, but these elements are flexible and their positions, rather than sequence, may determine how they interact with RfaH. Interestingly, Ser84, which interacts with the βGL in Eco RfaH-bound TEC [119], is replaced by Leu in

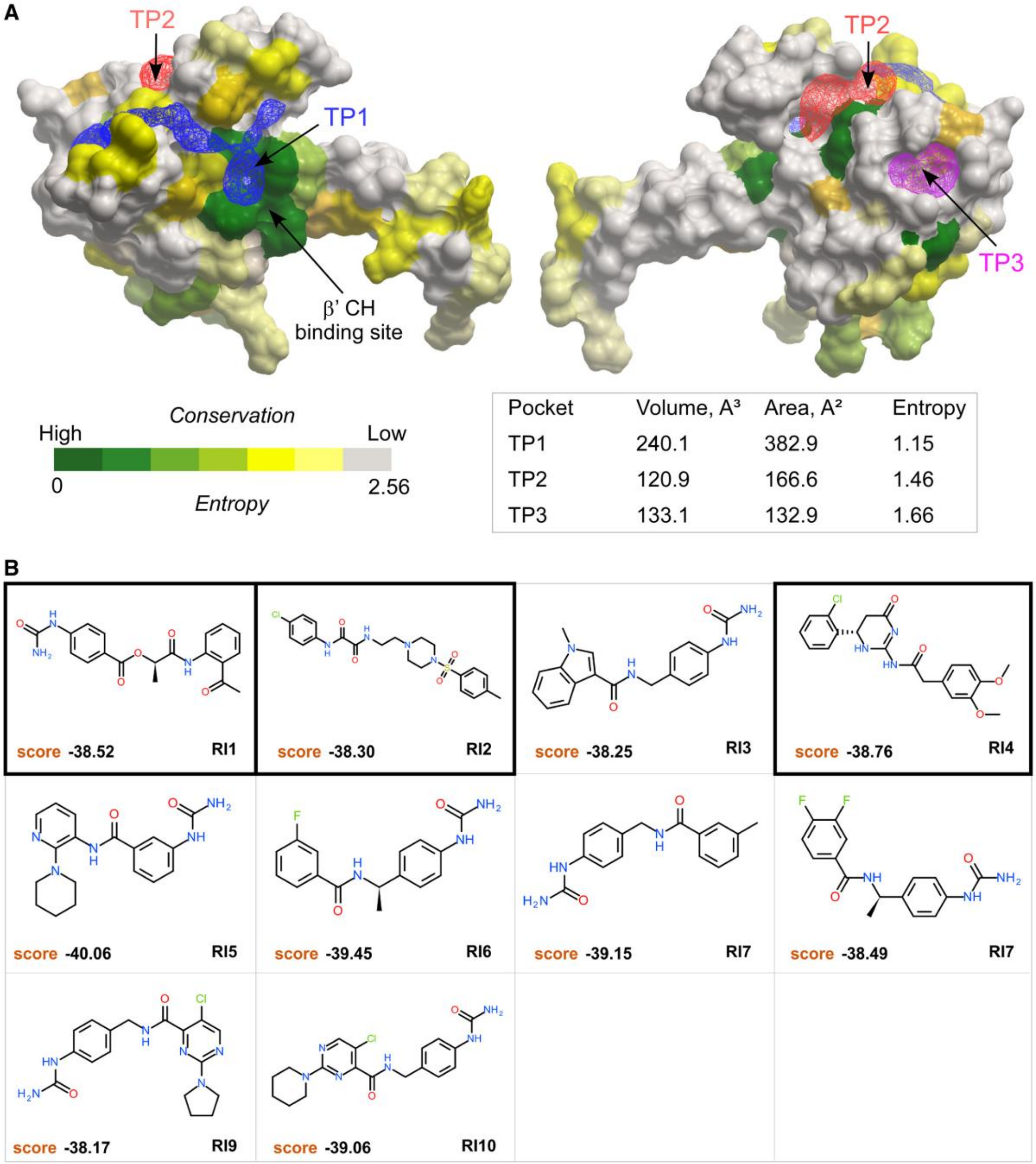
Kpn RfaH (Fig. 3-2). In addition, Asn70, which interacts with the *ops* DNA [119, 121], is substituted by Ser in Kpn RfaH. These changes may destabilize the tripartite network of RfaH/non-template DNA/βGL contacts. In contrast, 14 out of 15 residues that make contacts to the β' subunit are identical between the two proteins (Fig. 3-2).

3-3-4. In silico design of inhibitors targeting RfaH NTD/CH interactions

We sought to identify druggable pockets on RfaH and use structure-based screening to find small molecule modulators that bind directly to RfaH. This task is extremely challenging since the two domains of RfaH are small and flexible and RfaH biological function does not include binding to a small molecule cofactor or substrate; not surprisingly, small molecules that bind to RfaH have not been identified. Our results show that RfaH contacts with the *ops* DNA, the β' subunit of RNAP, and the ribosome appear to play important roles in both *E. coli* and *K. pneumoniae*, at least during the activation of the *lux* reporter operon (Fig. 3-4). We reasoned that small molecule ligands of the RfaH-NTD that bind at the interface with DNA and β'CH would interfere with RfaH function and that molecules designed to bind to Eco RfaH, for which the structural data are available, could be similarly effective with Kpn RfaH. While ligands that bind to the conserved CTD could abolish its interactions with ribosome and thus compromise function (similarly to I146D substitution; Fig. 3-4), we currently do not have a suitable *in vitro* assay for RfaH-dependent activation of translation.

The X-ray structure of Eco RfaH (PDB: 2OUG) was used to identify hypothetical pockets for potential RfaH inhibitors via *in silico* ligand screening. The RfaH-CTD was removed to unmask the β'CH binding interface of the NTD and a pocket-finding algorithm based on the mathematical transformation of the surface attraction fields called

ICMPocketFinder [126-127] was applied to the NTD model. Three tentative pockets (TP) on the RfaH-NTD were identified (Fig. 3-7A). The largest pocket, TP1, located near the interface of the NTD and β'CH, was chosen for further analysis (Fig. 3-7A).



Pocket	Volume, A³	Area, A²	Entropy
TP1	240.1	382.9	1.15
TP2	120.9	166.6	1.46
TP3	133.1	132.9	1.66

Figure 3-7. Tentative pockets on the RfaH-NTD and structures of potential inhibitors. A. Three tentative pockets (TP1, blue; TP2, red; and TP3, magenta) identified by ICMPocketFinder tool in ICM-Pro v3.8-6a are shown as transparent meshes; the volume and area data are shown in the table below. The RfaH-NTD is shown as a molecular surface where residues are colored by the alignment conservation Entropies (see Methods and Materials), with highly conserved (low Entropy) residues shown in green. The Entropy of each pocket was calculated as the average Entropy of residues around the pocket. B. Structures and docking scores of the top 10 hits from virtual ligand screening predicted to bind to TP1. Three molecules that show inhibitory activity against RfaH are indicated by thick borders.

To evaluate how residues around TP1 are conserved among all RfaH sequences, we aligned 751 RfaH sequences from various bacteria and quantified diversity of each position (see Materials and Methods). This analysis reveals that residues around TP1 are quite conserved (Fig. 3-7A), indicating the structural or functional importance of those residues and the feasibility of modulating diverse RfaHs by targeting TP1.

The RfaH-NTD structure needed to be optimized to achieve better docking results. To take the induced fit effect of protein pocket into account, we applied the SCARE method that generates a set of conformers with systematic omissions of pairs of interacting flexible residues [130], an approach that partially takes the induced fit effect into consideration. The ZINC database containing over 20 million (potentially commercially available) small molecules was chosen as the small molecule library for virtual ligand screening [148]. After docking and scoring 20 million compounds, 10 putative RfaH inhibitors (RI 1-10) were selected based on docking score and availability for further experimental validation (Fig. 3-7B).

3-3-5. Three small molecules inhibit Eco RfaH recruitment

In initial experiments, we tested whether RIs inhibited Eco RfaH effects on transcription when present at 1 and 2 mM. We used single-round *in vitro* transcription assays on a template that contains the 12-nt *ops* element downstream from a strong T7 A1 promoter (Fig. 3-8A). On this template, RNAP can be stalled at position A24 in the absence of UTP; the inclusion of α^{32}P-labeled NTP allows for the formation of radiolabeled halted TEC. The synchronized halted A24 TECs are restarted upon the addition of all NTPs. Rifapentin, which blocks re-initiation, is added to restrict transcription to a single round. In the absence of RfaH, RNAP pauses after the addition of C9 and U11 within the *ops* element (Figs. 3-8B and 3-9C), before making the full-length RNA of 79 nt. Addition of Eco RfaH reduces pausing at U11 ~3-fold, a reflection of RfaH antipausing activity, but not at U9 because RfaH is not yet recruited to the TEC [136]. In contrast, Eco RfaH delays RNAP escape from the G12 position, a well-documented consequence of RfaH recruitment which is presumably due to RfaH NTD-DNA interactions that must be broken to allow RNAP escape [72]. This delay is commonly used as a reporter of RfaH binding to the TEC [72]. Here, we used a one-point assay (Fig. 3-8) and a six-point time course (Fig. 3-9C) to assay RfaH-dependent delay at G12.

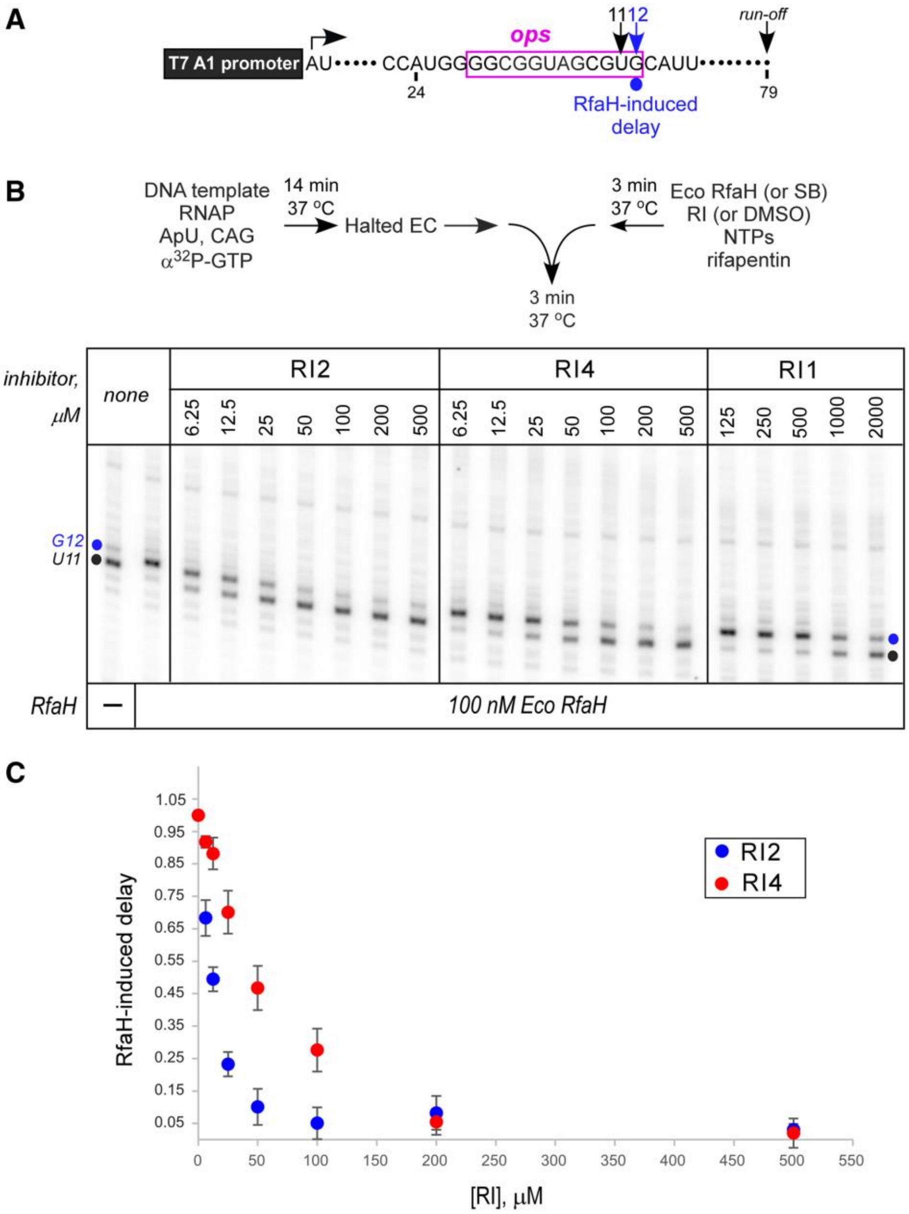
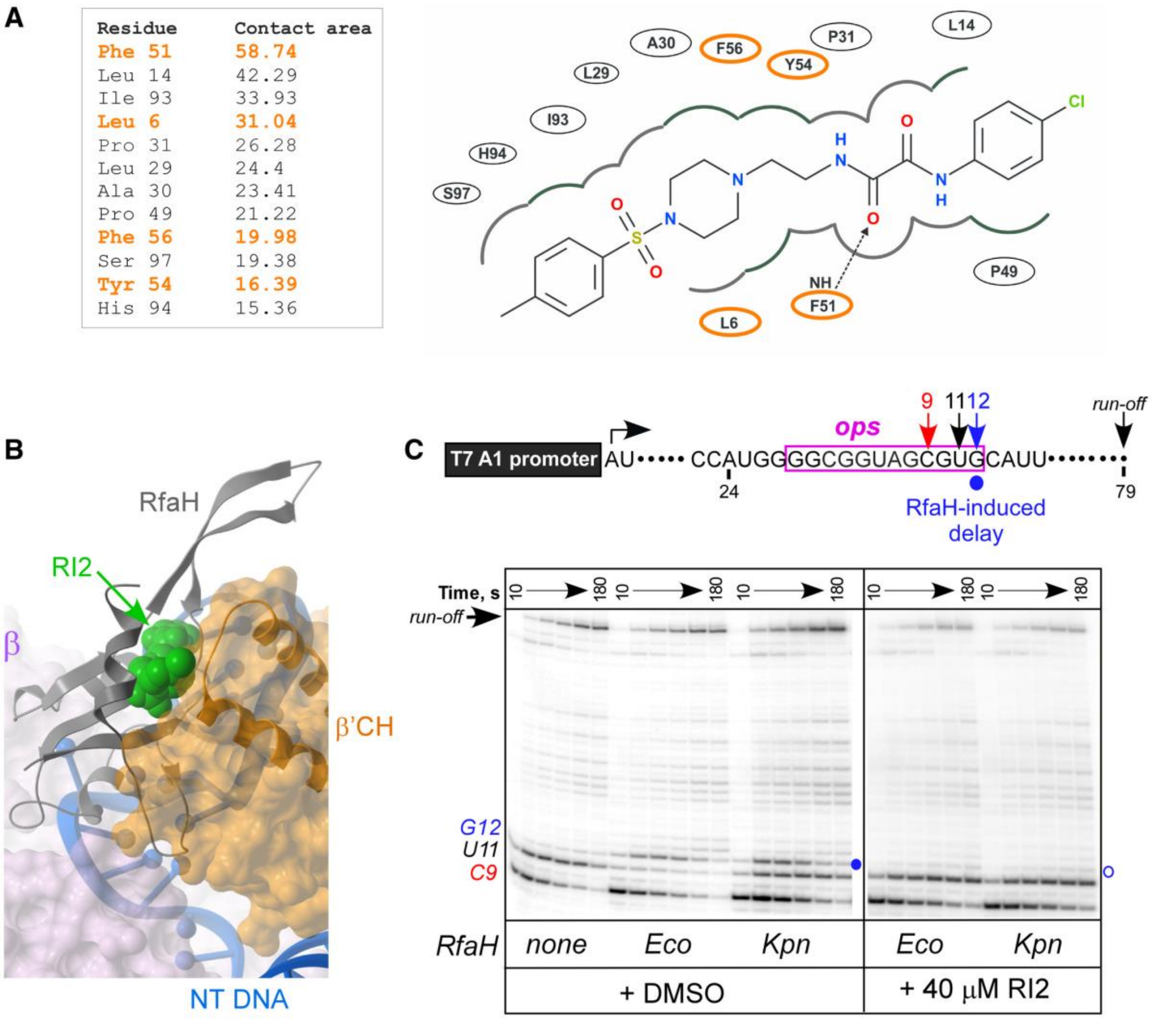


Figure 3-8. Inhibition of Eco RfaH recruitment by RIs. A. Transcript generated from the T7A1 promoter on a linear DNA template; transcription start site (a bent arrow), ops element (magenta box), pause sites and transcript end are indicated on top. B. Halted A24 TECs were formed as described in Materials and Methods. Elongation was restarted upon addition of NTPs and rifapentin in the presence of Eco RfaH (100 nM) preincubated with increasing concentrations of RI 1, 2 or 4. Aliquots were withdrawn at selected times and analyzed on a 10% denaturing gel. Positions of the paused and run-off transcripts are indicated; the position of the RfaH-induced RNAP pause at G12 is indicated with a circle. C. The fraction of G12 RNA was quantified as a function of RI concentration and corrected for levels observed in the absence of RfaH; the G12 RNA in the absence of RI (DMSO control) was defined as 1. The results of triplicate measurements for RI2 and RI4 are shown; errors are ± SD. Assays with RI1 were also performed in triplicates, but the observed inhibition was too weak to accurately determine the apparent IC_{50}.

We found that three compounds (RI1, RI2, and RI4) inhibited Eco RfaH recruitment to the transcribing RNAP; the remaining compounds did not exhibit any effect. RI2 and RI4 exhibited apparent IC_{50} of ~12 and 50 μM, respectively, whereas RI1 was only marginally active with IC_{50} of ~1 mM. For subsequent experiments, we focused on RI2, the most potent among the three ligands.

3-3-6. RI2 is predicted to block RfaH binding to the β' clamp helices

The hypothetical mode of RI2 binding to the RfaH-NTD suggests that it would sterically occlude the β'CH-binding site (Fig. 3-9AB). RI2 has two amide groups connected through the two carbonyl groups, which forms a large conjugation system with the benzene ring connected. The delocalization of electrons in the π bonds in two carbonyl groups and lone pairs of secondary amines makes the molecule rigid to some degree, potentially aiding its binding to RfaH-NTD. Several RfaH residues may be involved in the interaction with RI2 (Fig. 3-9A), and a subset of these residues (highlighted in orange) interact with the β'CH (Fig. 3-2 and [119]). Substitutions of RfaH residues that interact with the β'CH, including Tyr54 and Phe56, abolish RfaH activity [72-73]. In a model, RI2 is positioned between the NTD and the tip of the β'CH (Fig. 3-9B), the high-affinity RfaH binding on the TEC [119].



Residue	Contact area
Phe 51	58.74
Leu 14	42.29
Ile 93	33.93
Leu 6	31.04
Pro 31	26.28
Leu 29	24.4
Ala 30	23.41
Pro 49	21.22
Phe 56	19.98
Ser 97	19.38
Tyr 54	16.39
His 94	15.36

Figure 3-9. RI2 is predicted to block RfaH interactions with the β'CH domain. A. RfaH residues interacting with RI2 in the predicted binding pose. Residues indicated in orange interact with β'CH of RNAP. Left, contact areas (Å^2) of the RfaH-NTD residues interacting with RI2. Right, a 2D interaction diagram of RfaH-NTD and RI2 in the predicted model. The dashed line with an arrow represents a hydrogen bond between residues and RI2. B. Superposition of the modeled RfaH-NTD/RI2 complex and the cryo-electron microscopy structure (PDB ID: 6C6T) of the RfaH/TEC complex using the RfaH-NTD backbone. Binding of RI2 (green) is incompatible with the β'CH (orange). C. Effects of RI2 on RfaH recruitment at the ops site. Halted A24 TECs were formed as described in Materials and Methods. Elongation was restarted upon addition of NTPs and rifapentin in the presence of Eco or Kpn RfaH (100 nM) and RI2 (or DMSO). Aliquots were withdrawn at selected times and analyzed on a 10% denaturing gel. Positions of the paused and run-off transcripts are indicated; the position of the RfaH-induced RNAP pause at G12 is indicated with a circle.

All but one residue that interact with the β' clamp domain are identical between Eco and Kpn RfaH, suggesting that RI2 may also inhibit Kpn RfaH. We found that, similarly to Eco RfaH, Kpn RfaH was recruited to *E. coli* RNAP, delaying its escape from G12. However, in contrast to Eco RfaH, Kpn RfaH did not reduce pausing at U11 (Fig. 3-9C). This observation is consistent with the lack of productive interactions with GL inferred from Exo III and crosslinking experiments (Fig. 3-5) because the loss of GL contacts abolishes antipausing activity of Eco RfaH [67]. The addition of RI2 (at 40 μM) completely abolished recruitment of either RfaH (Fig. 3-9C).

To confirm that RI2 is a specific inhibitor of RfaH interactions with RNAP, and not a promiscuous inhibitor that could nonspecifically "stick" to nonpolar surfaces of diverse proteins and block their functional interactions, we tested if RI2 inhibits GreB-assisted transcript cleavage. Like RfaH, which binds to the N-terminal β' subunit coiled-coil domain (CH), GreB is a similarly-sized (22 kDa), two-domain protein that uses hydrophobic interactions with the C-terminal β' subunit coiled-coil domain (a.k.a rim helices) to bind to RNAP [135]. We found that, despite similar modes of binding of GreB and RfaH, RI2 did not inhibit GreB cleavage (Fig. 3-10A).

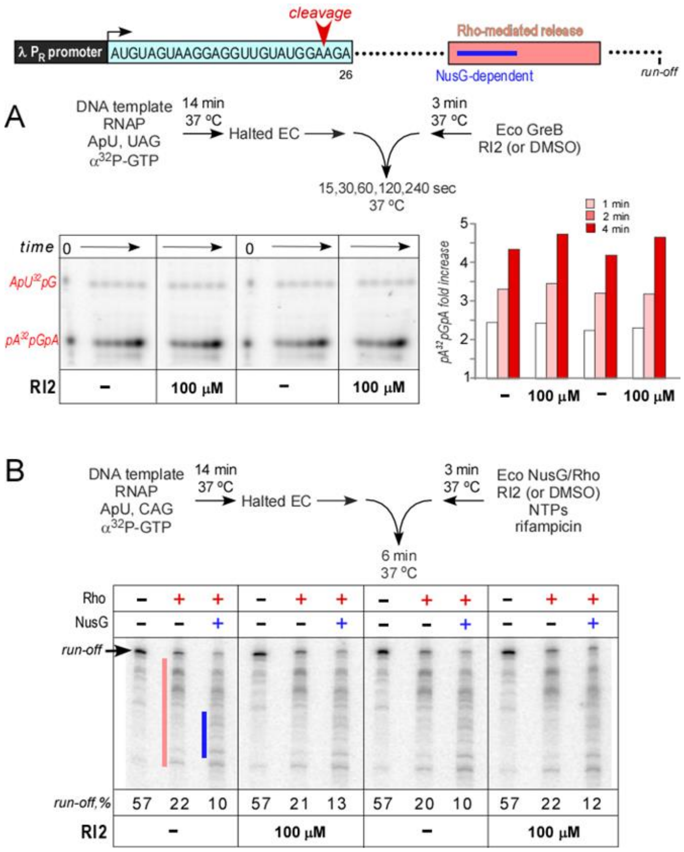


Figure 3-10. Analysis of RI effects on Eco GreB, NusG, and Rho transcription factors. A linear PCR-generated DNA template encoding a phage λP_R promoter, an initial 26-nt transcribed region lacking C residues, and a Rho-dependent *yhjG* terminator is shown on top. NusG induces an early RNA release, indicated by a blue bar within the red Rho-dependent release window. A. GreB-assisted RNA cleavage assays. Halted α^{32}P-GMP labeled A26 TECs were prepared by withholding CTP. In A26 TECs, the backtracked RNA is susceptible to cleavage, which is greatly enhanced by GreB [135]. Eco GreB (200 nM) was preincubated with RI2 (200 µM) or DMSO and mixed with an equal volume of A26 TEC to initiate the reaction; the final concentrations of GreB and RI2 were 100 nM and 100 µM, respectively. Following incubation at 37 ^{o}C for the indicated times, the reactions were quenched and analyzed on a 12 % urea-acrylamide (19:1) gel. The pA^{32}pGpA RNA, generated during repeated cleavage and re-synthesis of A26 RNA, was quantified as a function of time using ApU^{32}pG abortive RNA product as a loading control. Note that ApU^{32}pG migrates slower due to the absence of the 5' phosphate group. A graph on the right shows an increase of pA^{32}pGpA RNA over intrinsic, GreB-independent RNA cleavage (at time 0) as a function of time. Duplicate experiments are shown. B. Single-round Rho-dependent termination assays. Eco Rho (40 nM), Eco NusG (300 nM) and RI2 (200 µM) [or DMSO/storage buffers] were preincubated with rifampicin and chase NTPs and mixed with an equal volume of α^{32}P-GMP labeled A26 TECs; the final concentrations of Rho, NusG, and RI2 were 20 nM, 150 nM and 100 µM, respectively. Following a 6-min incubation at 37 ^{o}C, the reactions were quenched and analyzed on a 7 % urea-acrylamide (19:1) gel. Fractions of the full-length run-off RNA determined from duplicate experiments are shown below each panel.

We next tested whether RI2 inhibits Rho-dependent termination and NusG-stimulation of Rho (Fig. 3-10B). While Rho-binding site on RNAP remains to be identified, NusG binds to the same region of β'CH as does RfaH, yet most residues that make contacts are different (Fig. 3-2). NusG potentiates early Rho-mediated RNA release, shifting the termination window upstream [66]. We found that RI2 had no effect on Rho-dependent termination and marginally inhibited NusG (Fig. 3-10B). These observations support a model in which RI2 binds to the CH-docking site on RfaH, which is also nonpolar but rather different in NusG [119]. Neither RI had any effect on RNA synthesis by RNAP (Fig. 3-10 and data not shown).

These compounds are only modestly active, and their chemical characteristics suggest that they will not accumulate in Gram-negative bacteria [149]. Nonetheless, we tested RI2 for the ability to inhibit RfaH function in *E. coli.* RfaH is not essential, but its deletion in MG1655 confers extreme sensitivity to SDS (Fig. 3-3A). However, RI2 did not sensitize *E. coli* cells to SDS in a disk-diffusion assay, even when a *tolC* derivative of MG1655 strain was used (data not shown), suggesting that these compounds are not able to cross the cell wall to gain access to the cytoplasmic location of transcription and translation.

3-4. Discussion

RfaH-like regulators are dedicated activators of long operons which depend on antitermination mechanisms for complete synthesis of unusually long, up to 80,000 nts RNAs (reviewed in [116]). These operons encode proteins required for biosynthesis of a plethora of factors including capsule polysaccharides, LPS, and toxins. RfaH is critical for virulence in several pathogenic *Enterobacteriaceae* and its plasmid-encoded homologs are proposed to activate conjugative transfer of antibiotic-resistance genes [116].

Despite being easily recognizable as orthologs, RfaH proteins are unusually diverse, with identity as low as 43 % between Eco and *V. cholerae* RfaH [122]. Indeed, our results demonstrate that while mutations in regions that interact with the β'CH, the nontemplate DNA, and ribosome compromise Eco RfaH activity in *K. pneumoniae*, a substitution in the βGL contact site has only a modest effect (Fig. 3-4). This result indirectly suggests that RfaH-GL contacts may be dispensable in *Klebsiella*, an idea supported by *in vitro* transcription, footprinting, and crosslinking assays (Figs. 3-5 and 3-9). Consistently, two residues involved in the RfaH-GL-DNA network of interactions differ between Eco and Kpn RfaHs (Fig. 3-2). Notably, Eco NusG function is also independent of the βGL [146-147], in part due to differences in the NTD region that contacts the βGL [119].

In contrast to differences in DNA and β subunit contacts, all NusG homologs utilize the β'CH as a high-affinity site on the TEC [119]. We reasoned that molecules that bind to the β'CH binding site on the RfaH-NTD could give rise to broad-spectrum RfaH inhibitors. In this work, we identified a druggable pocket at the β'CH interface and used structure-based screening to find small molecule modulators that bind this surface and alter RfaH's function. This task is compounded by a small size of the interface and the lack of a natural or obvious small molecule binding site on RfaH. Furthermore, identifying the very first small molecule modulator for a conformationally flexible protein presents a significant challenge. These obstacles notwithstanding, two small molecules among the ten top hits were able to inhibit Eco RfaH *in vitro,* and the most potent lead, RI2, blocked recruitment of both Eco and Kpn RfaH to RNAP (Fig. 3-9). The predicted location of the RI2 binding site on the RfaH-NTD is consistent with a competition with the β'CH (Fig. 3-9B). However, we note that the proposed interactions are tenuous and need extensive validation.

Given the uncertainty of the RfaH-RI2 contacts, a systematic analysis using substitutions in RfaH and derivatives of RI2 would be necessary, a goal for future studies.

An obvious question is how RIs gain access to their putative binding site hidden at the RfaH domain interface (**Fig. 3-1**). We note that interactions between the NTD and CTD are relatively weak, and single substitutions in either domain have been shown to destabilize the domain interface to expose the β'CH binding site on the NTD, thereby bypassing a need for the *ops* element for RfaH recruitment to the TEC [68, 129]. We hypothesize that equilibrium between the closed, autoinhibited and an open, activated states of RfaH enables RI recruitment to free RfaH *in vitro*. Notably, RfaH inhibition by RI is observed only when the inhibitor is preincubated with RfaH in the absence of TEC; the RfaH-bound TEC is insensitive to inhibition, consistent with extensive interactions between the β'CH and the tentative RI docking site. Equilibrium between the two states could explain a lone example of RfaH association with an operon in the absence of the *ops* site [66]. Rapid and reversible transitions between these states are necessary for RfaH recruitment to the transcribing RNAP and refolding into the autoinhibited state upon dissociation from RNAP at the terminator. We are currently testing this hypothesis using NMR analysis.

Even though the identified leads have modest apparent affinities and do not inhibit Eco RfaH *in vivo*, these results are encouraging. Several key physicochemical characteristics favor compound accumulation in Gram-negative cells: low globularity, amphiphilic nature, rigidity, and the presence of a primary amine [149]. Extensive modifications would be required to turn our leads, which meet only the first criterion, into promising bioactive molecules. However, recent insights into the structural basis of the

RfaH action [119, 121] and into the rules that govern small molecule accumulation in Gram-negative cells [149] could be leveraged to design more potent RfaH inhibitors. It is also worth noting that we have developed a sensitive whole-cell reporter assay that can be used to screen libraries of drug-like molecules for potential RfaH inhibitors. We hope that our efforts to rationally re-design the leads described in this work and find new RfaH ligands by *in silico* and high-throughput screening will lead to identification of potent RIs. Since RfaH is required for LPS core biogenesis, RIs would be expected to act synergistically with existing antibiotics. Given the urgent concern of mounting Gram-negative antimicrobial resistance, new therapeutics are desperately needed. Inhibitors of Eco, Kpn, and related RfaHs could serve as novel antivirulence compounds to inhibit pathogenesis of organisms we are unable to kill with our failing armamentarium of antibiotics.

3-5. Conclusion

In this chapter, we further studied the interaction between RfaH and TEC. RfaH bind to the ops strand of NT DNA, βGL and β'CH of RNAP, and S10 subunit of ribosome. Our experiment showed that the interaction between RfaH and βGL of RNAP is to some degree dispensable, while the binding of RfaH to β'CH is necessary for all RfaH homologs. Therefore, we performed an *in silico* screen to identify chemicals that can disrupt the interaction between RfaH and β'CH. From the *in silico* screen, we successfully identified three first-in-class compounds that can inhibit the interaction between RfaH and RNAP and its processivity effect. The lead compound is active against both Eco RfaH and Kpn RfaH. Even though the lead compound cannot accumulate inside the cell well and doesn't work in vivo, with further modifications, they may be converted into a novel antibacterial drug candidate.

3-6. Acknowledgement

Chapter 3, in full, has been published in Svetlov Dmitri, **Shi Da**, Joy Twentyman, Yuri Nedialkov, David A Rosen, Ruben Abagyan, and Irina Artsimovitch. "In silico discovery of small molecules that inhibit RfaH recruitment to RNA polymerase." Molecular Microbiology (2018). The dissertation author was the primary investigator and author of these papers.

Chapter 4 Identification of amoebicidal non-toxic compounds based on a distant homology model of a target

In previous chapter, we introduced a promising novel target for bacterial infections, RfaH. We also identified 3 first-in-class inhibitors of RfaH through a large scale docking screen of 2 million compounds. However, we cannot ignore the fact that all the above screening and analysis cannot be done if we don't have structures of RfaH. In the same words, structures of target protein is the fundament of the structure-based drug discovery. In reality, only a small fraction of proteins have their structures solved. It is a common situation that we need to find compounds for targets without any structure. In this chapter, we will show that this problem can be "solved" to some degree by building structure models of the target based on homology. We will explain this idea by our work on identifying active compounds for a highly deadly parasitic disease, primary amoebic meningoencephalitis.

4-1. Introduction

Primary amoebic meningoencephalitis (PAM) is a rare but deadly disease caused by the opportunistic pathogen, *Naegleria fowleri*. *N. fowleri* is a free-living amoeba found in warm freshwater and soil habitats in all continents, except Antarctica[150-154]. *N. fowleri* mostly feeds on bacteria but after infecting humans it switches to human brain cells and causes severe brain inflammation and irreversible brain damage, leading to death [155-157].

Since the first report of PAM in 1965 in Australia, several hundred PAM cases have been reported worldwide. PAM is considered rare in the US with less than 10 reported cases per year [33]. However, this number is likely to be underestimated, because of diagnostic limitations and fast disease progression. Symptoms begin 1-9 days after the infection and disease results in nearly 97% mortality within the following two weeks [33].

High mortality associated with PAM results from the rapid onset, delayed diagnosis and lack of effective treatment. Until 2018, only 4 people in the U.S. out of 145 well documented cases had survived infection [158-160]. All survivors were treated with anti-fungal drug amphotericin B (AmpB), rifampicin, dexamethasone and one or more drugs from the following list: miconazole [160], fluconazole, miltefosine, and azithromycin [158-159]. No treatment regimen with consistent survival outcome has been established so far.

AmpB is used for serious fungal infections and leishmaniasis [161]. It acts through binding to ergosterol in the pathogen cell membranes, causing rapid leakage of monovalent ions, such as K^+ and Na^+ [162]. Other anti-fungal drugs, e.g. fluconazole and miconazole, act by inhibiting the sterol 14-demethylase (CYP51) and disrupting the ergosterol biosynthetic pathway [163]. Fluconazole, and some other "conazole" drugs (posaconazole, ketoconazole, voriconazole and itraconazole) were reported to kill *N. fowleri* in *in vitro* assays [164-166].

Drug discovery for PAM has been hindered by the lack of validated molecular targets and drug candidates that readily cross the blood-brain barrier (BBB). BBB permeability of drugs is a prerequisite for the treatment of brain infections. For example, only a small amount (~ 3% as compared to plasma concentration) of AmpB crosses the BBB [167], which may explain its limited efficacy against PAM. This limitation motivated us to look for new therapeutic targets and their inhibitors with strong amoebicidal activities and BBB permeability.

The sterol biosynthesis pathway has pathogen-specific enzymes in fungi and some protozoa, such as kinetoplastids and free-living amoeba [166, 168-170]. Successful development of the "conazole" group of antifungal drugs (miconazole, ketoconazole, voriconazole, etc.)

demonstrated the feasibility of targeting this pathway[165, 171]. Furthermore, *N. fowleri* CYP51 and two other enzymes in this pathway, 24-sterol methyl transferase (24-SMT) and sterol Δ8−Δ7 isomerase (yeast ERG2 equivalent referred here as NfERG2), are validated as potentially druggable targets for anti-PAM drug development [164, 166].

Due to notable sequence similarity between the catalytic domain of ERG2 and human σ_1 non-opioid brain receptor, ERG2 is of particular interest for PAM drug discovery. NfERG2 catalyzes sterol $\Delta^8 \rightarrow \Delta^7$ double-bond isomerization in the pathway for biosynthesis of ergosterol, an essential component of *N. fowleri* plasma membranes (Fig. 4-1). Inhibition of NfERG2 depletes the intracellular ergosterol pool, disrupts cell and organelle membranes and induces autophagocytosis leading to *N. fowleri* death [166].

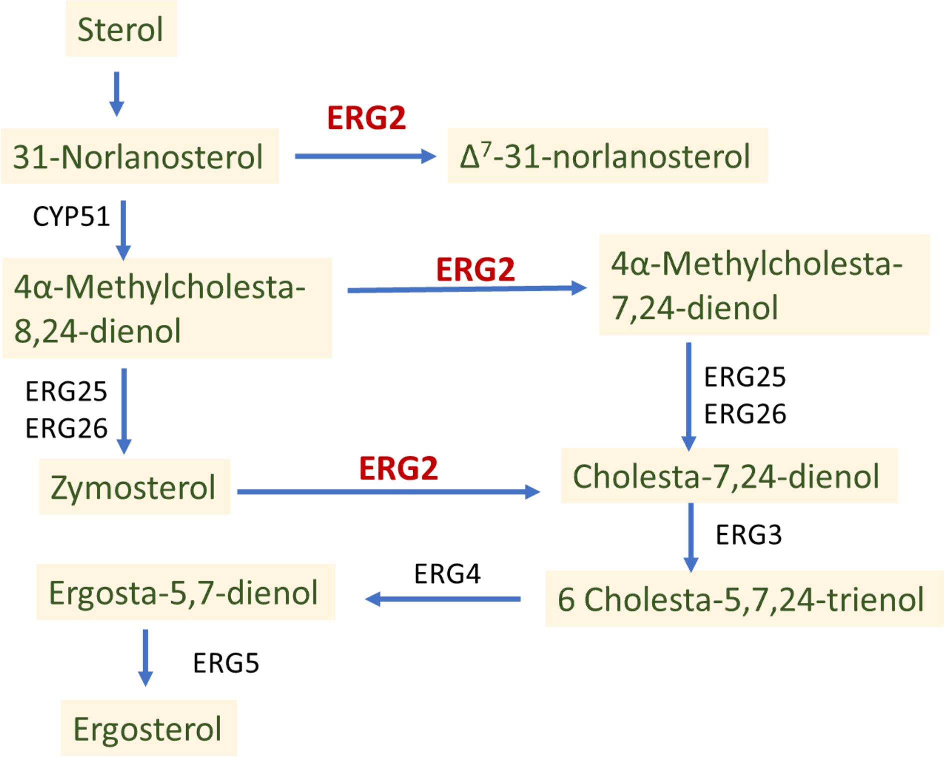


Figure 4-1. Ergosterol Biosynthesis in N. fowleri. Biosynthetic steps catalyzed by sterol Δ8−Δ7 isomerase (ERG2) in ergosterol biosynthesis in N. fowleri as reported elsewhere [166].

In this work, we applied a structure-based docking screen against a homology model of NfERG2 (AmoebaDB NF0056720), followed by experimental validation of hits in cell-based assays. First, based on the x-ray structure of human receptor (PDB ID: 5HK1) [172], we built a homology model of NfERG2 and identified a potentially druggable pocket using a pocket finding algorithm [173]. Then virtual ligand screening was performed by docking a library of 26,000 small molecules to the predicted pocket. Based on the *in silico* screening results, we tested 30 top ranking hits in a cell-based assay for efficacy against *N. fowleri* trophozoites. Out of eight experimentally active compounds, four compounds had high amoebicidal potencies and low human cell toxicity.

4-2. Methods and Materials

4-2-1. Software

Sequence alignment, homology modeling of ERG2, pocket identification and virtual ligand screening were performed using the inbuilt tools of ICM-Pro (version v3.8-6a)[75].

4-2-2. Sequence alignment of ERG2 in various organisms

369 ERG2 sequences from different organisms were downloaded from UniProt via searching “C-8 sterol isomerase” in the protein name field, including ERG2 of *N. fowleri*. Sequence alignment was generated using ICM-Pro using the zero end-gap global alignment method [174]. The comparison matrix was introduced by Gonnet et. al. [76]. A residue conservation profile was generated to show the amino acids essential for the structure and function of the protein. The amino acid counts were normalized by the same factor (1/369) in all alignment positions.

4-2-3. Homology modeling of ERG2 of *N. fowleri*

The crystal structure of the human σ1 receptor (PDB ID: 5HK1) was used as a template. Sequence alignment of ERG2 of *N. fowleri* and human σ1 receptor for homology modeling was built through the zero end-gap global alignment method with the Gonnet comparison matrix [76, 175]. The gap opening and extension penalty were set as 2.4 and 0.15, respectively. A pP value was calculated to show the probability that the alignment was random, shown as equation 1. Based on the alignment and structure template, homology model of ERG2 of *N. fowleri* was built with the default parameters in ICM-Pro, with all side chains and insertions/deletions sampled and refined via a biased probability Monte Carlo method[176].

$$\mathrm{pP} = -\log_{10}(\mathrm{P\ value})$$

Equation 4-1

4-2-4. Tentative pocket identification and potential maps generation

In the homology model of ERG2 of *N. fowleri*, a tentative pocket was identified using the ICMPocketFinder tool of ICM-Pro [173], with default values of input parameter (tolerance = 4.6). The calculated volume of the predicted pocket was used to pre-filter compounds for docking with a 20% margin. Based on the identified pocket, the docking region was defined. The potential maps for the docking screen were calculated on a 0.5 Å 3D grid, containing: (i) van der Waals interactions; (ii) electrostatic interactions; (iii) hydrogen bonds; and (iv) hydrophobic potentials.

4-2-5. Virtual ligand screening for *N. fowleri* inhibitors

Virtual ligand screening was conducted by docking a digitized in-house chemical diversity library of the Center for Discovery and Innovation in Parasitic Diseases (CDIPD) containing over 26,000 small molecules to the pre-defined pocket on ERG2 of *N. fowleri* and ranking them by docking scores. Prior to the docking screen, chemicals in the library were filtered by their volumes to fit the predicted pocket volume with a 20% margin. The docking and scoring of each chemical was conducted using a stochastic global energy optimization procedure in internal coordinates [177] implemented in the ICM-Pro v3.8-6a, described as the following steps. 1) A ligand was sampled in an implicit solvent model to generate a series of starting conformations, and each starting conformation was placed into the binding pocket with four principal orientations. 2) The ligand was sampled in the pre-calculated potential maps through biased probability Monte Carlo sampling to optimize the position and internal variables of the ligand. 3) For each ligand, 10 top ranking conformations were optimized and re-scored with ICM full atom scoring function [131], and conformations with the best docking score were kept for comparison. 4) All filtered chemicals were docked to the selected pocket on ERG2 of *N. fowleri* following the above procedures with a computing cluster containing 128 cores. After docking, all chemicals were ranked by their docking scores and the top 30 hits were tested experimentally [178].

4-2-6. Blood-brain barrier permeability score calculation

The BBB permeability scores were evaluated by the BBB-MPO method for the eight active compounds. The BBB-MPO score is calculated by transforming five physicochemical properties of a compound, calculated partition coefficient (CLogP), molecular weight (MW), topological polar surface area (PSA), number of hydrogen bond donors (HBD), and pKa of the most basic center into a number ranging from zero to five. Detailed description of BBB-MPO score calculation can be found in the ICM manual and the original publications [75, 179-180].

4-2-7. Chemicals and reagents

White, solid flat-bottom 96-well microplates (GREINER BIO-ONE). A CellTiter-Glo luminescence-based cell viability assay kit (Promega Corporation). Dimethyl Sulfoxide (DMSO) and amphotericin B, both purchased from Sigma-Aldrich. The 30 compounds selected for *in vitro* testing were taken from an in-house chemical diversity library of the CDIPD containing over 26,000 compounds (average molecular weight of 445±115, CLogP of 4.1±2 and PSA of 75±33) donated by Biosero Inc. Each compound was prepared as a 10 mM DMSO solution.

4-2-8. Proliferation Inhibition assay for N. fowleri

N. fowleri strain KUL, originally isolated from human cerebrospinal fluid in Belgium in 1973 [181], was obtained from ATCC. KUL is type 3 strain based on the length of the internal transcribed spacers 1 (ITS1), with the T at position 31 in the 5.8S rDNA sequence [151]. The trophozoites were cultured axenically in Nelson medium and 10% fetal bovine serum at 37°C. All experiments were performed using cells harvested during the logarithmic phase of growth. All experiments were conducted in a biosafety cabinet following the BSL-2 procedures as specified in the UCSD Biosafety Practice Guidelines.

Primary screening: negative control wells in the screening plates contained 0.5% DMSO, and positive-control wells contained 50 μM amphotericin B (Sigma-Aldrich). The assay was performed in triplicate. *N. fowleri* trophozoites (10,000 amebae per well) were plated in 96-well plates with Nelson medium. The test compounds (diluted in Nelson medium) were added to the wells to achieve a final concentration of 50 μM in each well. Total volume in each well was 100 μL. Assay plates were incubated for 48 hours at 37°C.

Secondary screen for potency determination: For confirmatory screens of the best hits from the primary screen, serial dilutions of test compounds were prepared from 10 mM stock. For determination of half-maximal effective concentration (EC_{50}), the stocks (10 mM) were diluted with DMSO to yield a 2X serial dilution with a concentration range of 0.39-50 μM. The serially diluted compounds were added to the wells of 96-well plates and *N. fowleri* trophozoits (10,000 amebae per well) were added in each well. The assays were performed in triplicate. Assay plates were incubated at 37°C for 48 hours.

Estimation of bioluminescence: At the end of incubation period, 50 μL of Cell Titer-Glo luminescent cell viability reagent was added to each well of the plate. The plates were then placed on an orbital shaker at room temperature for 10 min to induce cell lysis. After lysis, the plates were equilibrated at room temperature for 10 min to stabilize the luminescent signal. The resulting ATP bioluminescence was measured at room temperature by use of a Perkin Elmer Envision plate reader. Data were expressed as mean ± standard deviation for all experiments. The results were analyzed using a non-linear regression in Prism 7.0.1.

4-2-9. Cell toxicity assay

HEK293 cells (ATCC) were cultured in DMEM+10% FBS culture media in tissue culture treated (Corning, 430641U) flasks at 37 °C, 5% CO_2. The cells were plated at 5000 cells per well in 90 µL of respective media in tissue culture treated (Falcon, 353219) 96-well plates. 10 μL of test compound dilution prepared in culture media was added at different serially diluted concentrations (100 μM and lower) and incubated the plate for at least 72 hours at 37 °C. After the incubation period, 8 µL of Alamar blue dye (Invitrogen, DAL1100) was added to each well, the cells were incubated for 2-4 hours, and then analyzed using SpectraMax fluorescence reader using excitation and emission wavelengths of 544 nm and 590 nm, respectively. Data were expressed as mean ± standard error of mean for all experiments. The results were analyzed using a non-linear regression in Prism 7.0.1.

4-2-10. HPLC-MS/MS analysis

The purity and identity of eight active compounds was confirmed by the HPLC-MS/MS method with the following steps. The compound stock solutions (DMSO at 10 mM) were dissolved in HPLC-MS grade methanol to obtain a concentration of 0.05 mg/mL. These samples were analyzed with an ultra-high-performance liquid chromatography device (Vanquish, Thermo Scientific) coupled to a quadrupole-Orbitrap mass spectrometer (Q Exactive, Thermo Scientific). Chromatographic separation was done using a Kinetex C18 1.7 µm (Phenomenex, Torrance, USA), 100 Å pore size, 2.1 mm (internal diameter) x 150 mm (length) column with a C18 guard cartridge (Phenomenex). The column was maintained at 40°C. The mobile phases used were 0.1% formic acid in water (A) and 0.1% formic acid in acetonitrile (B), and the flow rate was set to 0.5 mL/min. Chromatographic elution gradient was: 0.00-1.00 min, 5% B; 1.00 - 15.00 min, 5% to 100% B; 15.00 -16.9 min, 100% B; 17.0 - 19.0 min, 5% B. The injection volume was set to 1 µL.

Mass spectrometry experiments were performed in electrospray ionization, operating in positive ionization mode with a heated electrospray ionization source. The following source parameters were used: spray voltage, +3000 V; heater temperature, 370°C; capillary temperature, 350°C; S-lens RF, 55 (arb. units); sheath gas flow rate, 55 (arb. units); and auxiliary gas flow rate, 20 (arb. units). The MS^1 scans were acquired at a resolution of 35,000 (at m/z 200) for the 100-1500 m/z range, and the MS^2 scans at a resolution of 17,500 from 0.48 to 16.0 min. The automatic gain control (AGC) target and maximum injection time were set at 5 x 10^5 and 150 ms for MS1 and MS^2 scans. Up to four MS^2 scans in the data-dependent mode were acquired for most abundant ions per duty cycle, with a starting value of m/z 70. Higher-energy collision-induced dissociation was performed with a normalized collision energy of 20, 35, 50 eV. The apex trigger mode was used (2-7 sec) and the isotopes were excluded. The dynamic exclusion parameters were to 6 sec. Compound purity was estimated by integration of the HPLC-MS peak area.

4-2-11. LC-MS/MS data conversion, analysis and deposition

Thermo raw data were converted to m/z extensible markup language (mzML) in centroid mode using MSConvert (part of ProteoWizard) [182]. The data were visualized with the TOPPView OpenMS software [183]. The mass spectrometry data have been deposited on the MassIVE public repository under the accession number MSV000083490. The reference spectra were deposited to GNPS spectral library (CCMSLIB00004752955 - CCMSLIB00004752981) [184].

4-3. Results

4-3-1. Homology modeling of ERG2 of *N. fowleri*

Given that no experimental ERG2 structure is yet available, we built a homology model of NfERG2 based on the crystal structure of the human σ1 non-opioid receptor (PDB ID 5HK1) [172]. The human σ1 receptor is implicated in various CNS diseases such as addiction, amnesia, pain and depression [185]. However, despite homology to ERG2, human σ1 receptor lacks Δ^8-Δ^7 isomerase activity [186]. Fig. 4-2a shows sequence alignment of the catalytic domain of NfERG2 and human σ1 receptor sharing 30% sequence identity and 60% sequence similarity over 177-amino acids length, which implies statistically significant structural similarity. Furthermore, it is known that the yeast ERG2 and mammalian σ1 receptor can be targeted with the same compounds. Thus, Moebius and co-authors tested 11 chemicals against the σ1 non-opioid receptor of guinea-pig and ERG2 of yeast, and found 10 chemicals out of 11 tested had similar binding affinities in both proteins [187-188]. To study the conservation and relative importance of each position in ERG2, 369 sequences from different organisms, annotated in UniProt database as sterol $\Delta^8 \rightarrow \Delta^7$ isomerases, were aligned. The alignment conservation profile was added to the pairwise alignment of NfERG2 and human σ1 receptor (Fig. 4-2a).

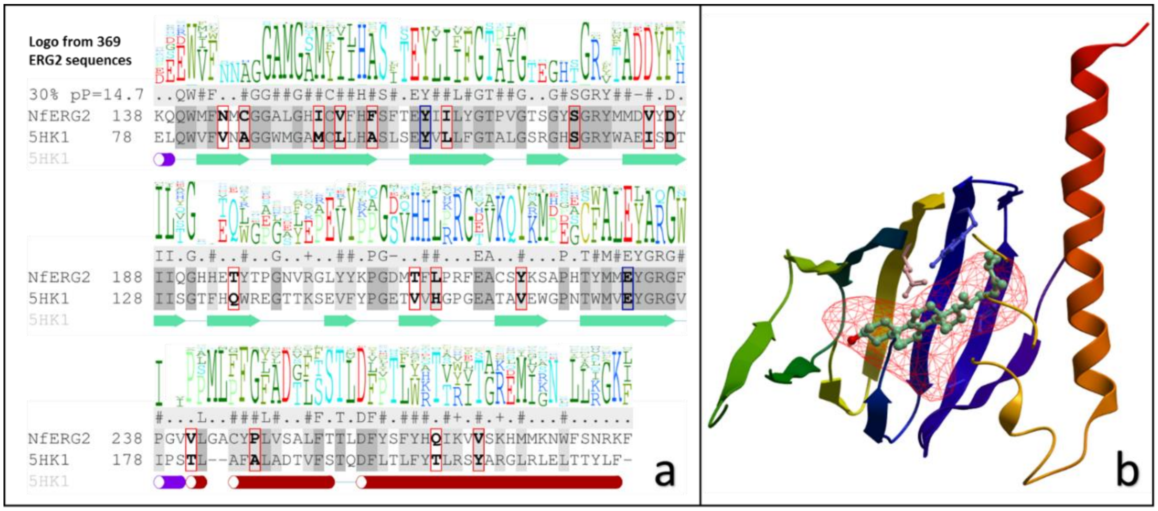

Logo from 369 ERG2 sequences
30% pP=14.7 ..QW#F..#GG##G##C##H#S#.EY##L#GT##G..G#SGRY##-#.D.
NfERG2 138 KQQWMFNMCGGALGHICVFHFSFTEYIILYGTPVGTSGYSGRYMMDVYDY
5HK1 78 ELQWVFVNAGGWMGAMCLLHASLSEYVLLFGTALGSRGHSGRYWAEISDT
5HK1
II.G.#..#..G..+..##.PG-..##...EA..#...P.T#M#EYGRG#
NfERG2 188 IIQGHHETYTPGNVRGLYYKPGDMTFLPRFEACSYKSAPHTYMMEYGRGF
5HK1 128 IISGTFHQWREGTTKSEVFYPGETVVHGPGEATAVEWGPNTWMVEYGRGV
5HK1
#...L..###L#..#F.T.DF#.###.#+.#.+.#...#......
NfERG2 238 PGVVLGACYPLVSALFTTLDFYSFYHQIKVVSKHMMKNWFSNRKF
5HK1 178 IPSTL--AFALADTVFSTQDFLTLFYTLRSYARGLRLELTTYLF-
5HK1
a
b

Figure 4-2. Sequence alignments and homology model of NfERG2. a) Sequence alignments of NfERG2 and human σ1 receptor ligand binding cavity. Positions with red and blue boxes correspond to residues of the binding pocket in the NfERG2 model. Conservation profile above the NfERG2 sequence was generated from sequence alignment of ERG2 in 369 different organisms. The secondary structure elements of the 5HK1 σ1 receptor are marked by different colors and shapes as following: red cylinder = alpha helix, green arrow = beta sheet, blue cylinder = pi helix, magenta cylinder = 3/10 helix. b) Structure of NfERG2 model in complex with zymosterol. NfERG2 model is colored in rainbow colors to emphasize the N (purple) and C (red) termini. Two residues (Y163 and E232) that are potentially critical for the enzymatic reaction of NfERG2 are shown as stick representations. The red mesh represents the binding pocket of NfERG2. The predicted binding pose of substrate zymosterol in the homology model of NfERG2 is shown in (b).

The homology model was constructed using the computational tools implemented in ICM-Pro v3.8-6a software suite [75]. The structure of the NfERG2 model is shown in Fig.

2b. The high quality of alignment and lack of long insertions or deletions resulted in a low-energy model. The stereochemical quality of the model was checked through Ramachandran plot generated with ProCheck [189]. 82.6% of the residues have Phi and Psi angles in the most favored regions, 16.5% of residues in the allowed regions, and 0.8% of residues in the generously allowed regions (see Fig 4-3). None of the residues are in the disallowed regions, indicating the overall satisfactory quality. A well-defined 505 Å^3 pocket/cavity was identified (Fig. 4-2b) using ICMPocketFinder utilizing mathematical transformation of the surface attraction fields [173]. The shape and size of the identified pocket are compatible with small molecule inhibitors. The binding pocket residues of NfERG2 are similar to corresponding residues of human σ1 receptor (20% identity and 74% similarity). However, the larger size of the pocket and residue differences may be sufficient to identify NfERG2-specific inhibitors.

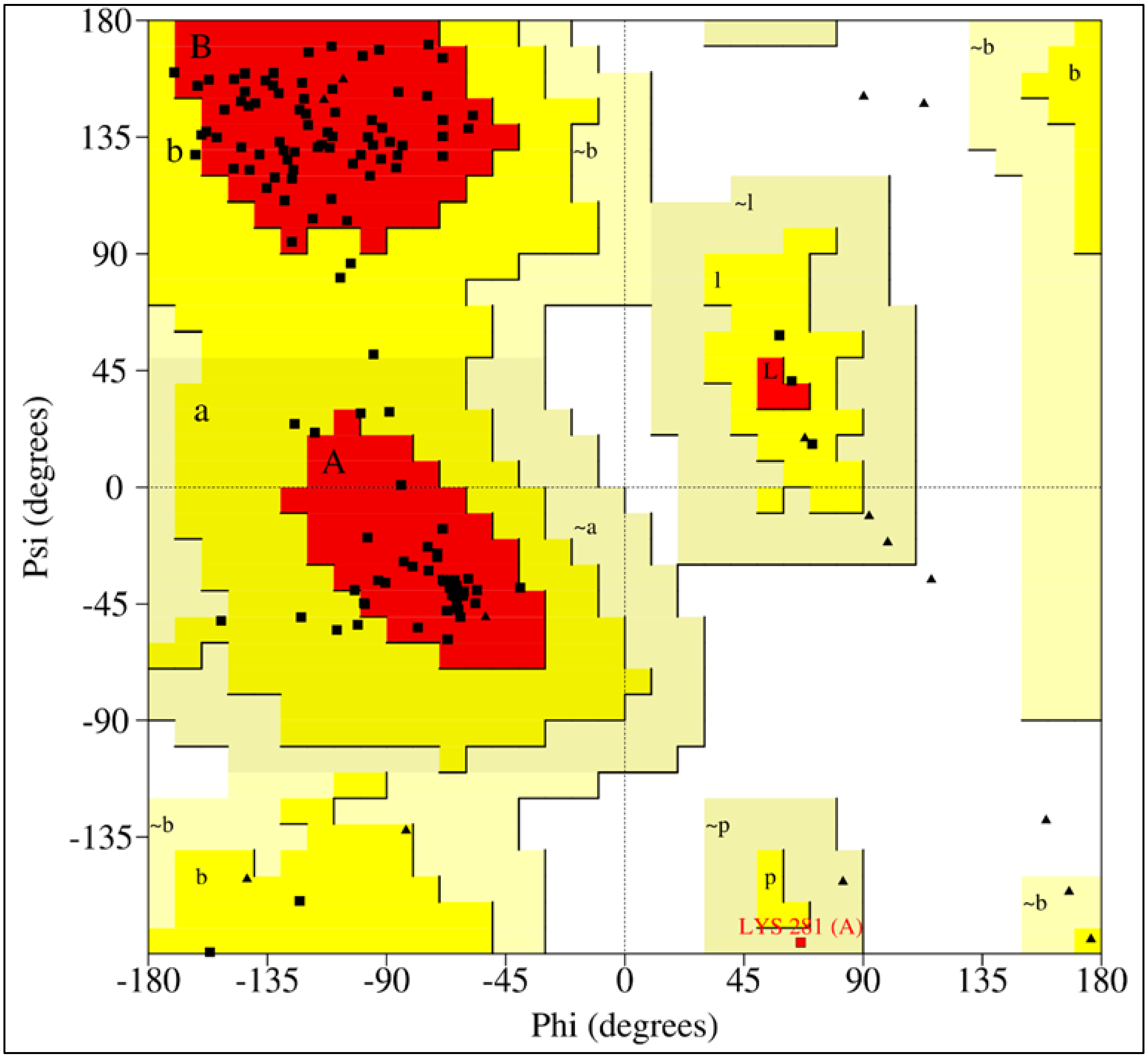


Figure 4-3. Ramachandran plot of the homology model of NfERG2.

Two residues of the binding site, Y163 and E232, are believed to be important for the enzymatic reaction of NfERG2, as they are highly conserved among ERG2 enzymes of different organisms [174]. Furthermore, the predicted binding pose of ERG2 substrate, zymosterol, shows that the negatively charged E232 is in proximity of the C8 carbon of the substrate and may stabilize the transition state carbocation formed at C8 during isomerization reaction. Y163 can form a hydrogen bond with E232 and further stabilize the transition state of substrates by interacting with the carbocation [190]. In addition, mutagenesis experiments and crystal structure confirmed the essentiality of the corresponding residues, Y103 and E172, for the ligand binding to human σ1 receptor [172, 191]. The pocket around those two residues in its lowest energy conformation was used as the binding site for further *in silico* screening of large compound library.

4-3-2. *In-silico* screening of ERG2 inhibitors

An in-house chemical library at the Center for Discovery and Innovation in Parasitic Diseases (CDIPD) was donated by Biosero Inc. It contains over 26,000 compounds with an average molecular weight of 445±115, cLogP of 4.1±2 and polar surface area (PSA) of 75±33. The library was digitized and pre-filtered for *in silico* screening. Approximately 16,000 compounds with the volume ranging from 400 $Å^3$ to 600 $Å^3$ were docked to the pre-defined binding site of the NfERG2 model. Thirty computationally predicted hits with the highest scores were experimentally tested against proliferating *N. fowleri* trophozoites. The chemical structures of top scoring hits and their docking scores are shown in Fig. 4-4.

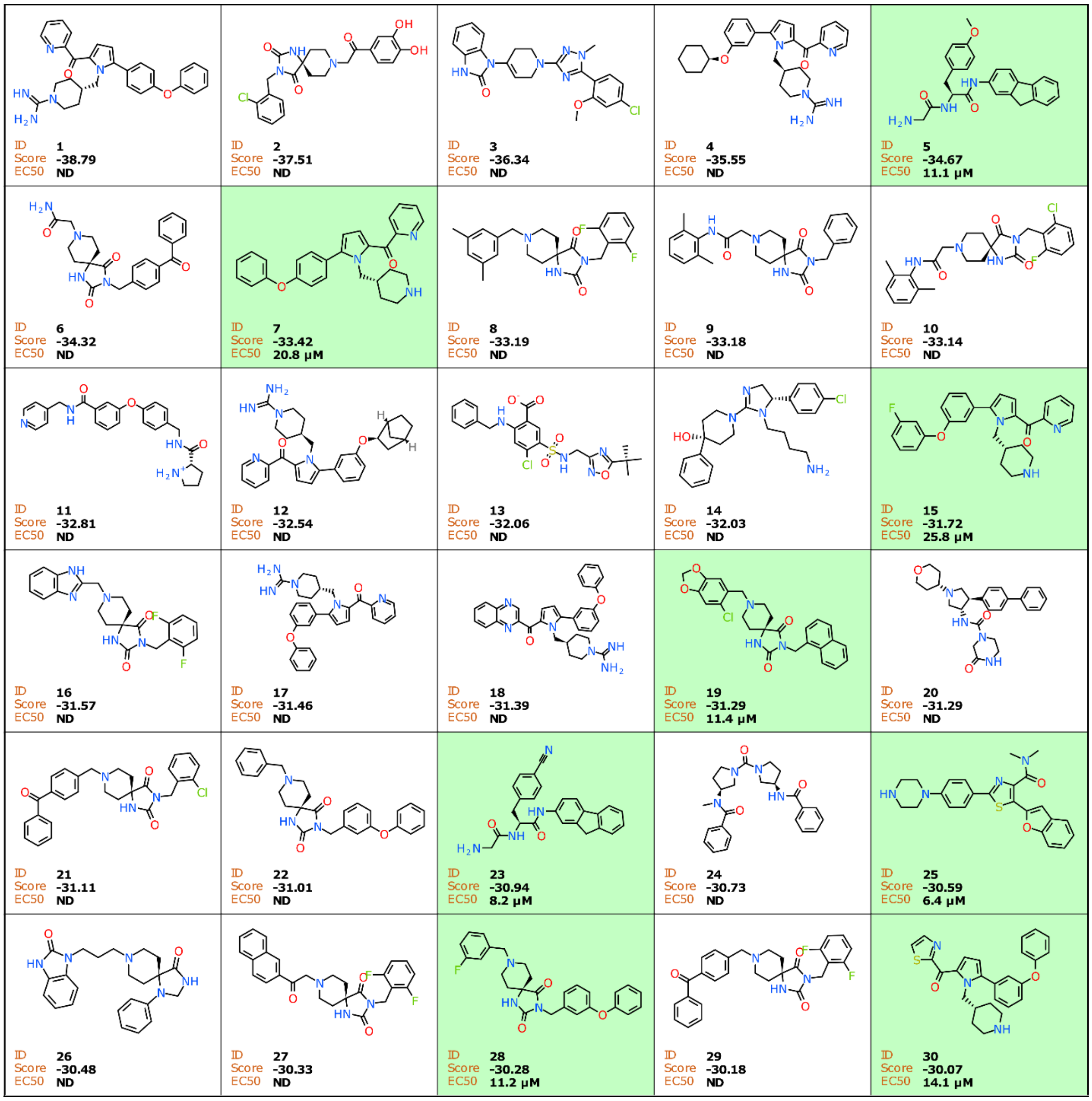

ID 1 Score -38.79 EC50 ND
ID 2 Score -37.51 EC50 ND
ID 3 Score -36.34 EC50 ND
ID 4 Score -35.55 EC50 ND
ID 5 Score -34.67 EC50 11.1 µM
ID 6 Score -34.32 EC50 ND
ID 7 Score -33.42 EC50 20.8 µM
ID 8 Score -33.19 EC50 ND
ID 9 Score -33.18 EC50 ND
ID 10 Score -33.14 EC50 ND
ID 11 Score -32.81 EC50 ND
ID 12 Score -32.54 EC50 ND
ID 13 Score -32.06 EC50 ND
ID 14 Score -32.03 EC50 ND
ID 15 Score -31.72 EC50 25.8 µM
ID 16 Score -31.57 EC50 ND
ID 17 Score -31.46 EC50 ND
ID 18 Score -31.39 EC50 ND
ID 19 Score -31.29 EC50 11.4 µM
ID 20 Score -31.29 EC50 ND
ID 21 Score -31.11 EC50 ND
ID 22 Score -31.01 EC50 ND
ID 23 Score -30.94 EC50 8.2 µM
ID 24 Score -30.73 EC50 ND
ID 25 Score -30.59 EC50 6.4 µM
ID 26 Score -30.48 EC50 ND
ID 27 Score -30.33 EC50 ND
ID 28 Score -30.28 EC50 11.2 µM
ID 29 Score -30.18 EC50 ND
ID 30 Score -30.07 EC50 14.1 µM

Figure 4-4. Chemical structures of the top 30 docking hits from the in-silico screening. The compounds are identified by arbitrary numbers as per the docking list generated. Docking score is shown for each chemical structure. All 30 compounds were tested for anti-N. fowleri activity in cell-based assay and eight experimentally active compounds are highlighted in green. The EC_{50} values observed for active compounds are also provided with the corresponding structures.

4-3-3. Anti-proliferative activity of compounds in cell-based assay.

An *in vitro* assay used to test the anti-proliferative activities of top scoring compounds was developed and validated previously [192]. *N. fowleri* KUL strain in trophozoite form, used in the assay, is highly pathogenic and causes mortality in mice within seven days, and the strain relevance and applicability was validated previously [164, 166, 193-195]. Thirty compounds were first tested at 50 μM concentration (EC_{50} of miltefosine); eight compounds (highlighted in green in Fig. 4-4) showed 100% inhibition in this assay and were further evaluated for purity, identity and dose-response. The purity and identity of 8 hits were confirmed by HPLC-MS analysis. Based on LC-MS peak area in the total ion current (TIC) chromatograms, the purity of all compounds was >98%, except for compound 7 which was >95% pure (see Fig 4-9 – 4-16 in appendix). The half- maximal effective concentrations (EC_{50}) for *N. fowleri* proliferation were estimated using Prism 7.0.1. Fig. 4-5 shows the dose-response curves (blue) for all eight active compounds along with the observed EC_{50} values ranging from 6.4 μM to 25.8 μM. The compounds 23 and 25 were observed to be most potent among 8 active compounds. The observed EC_{50} values for the compounds 23 and 25 were 8.2 μM (95% CI: 4.6 - 15.5) and 6.4 μM (95% CI: 4.3 - 9.2), respectively. The compounds 5, 19 and 28 showed approximately the same EC_{50} values of 11.1 μM (95% CI: 10.1 - 12.6), 11.4 μM (95% CI: 11.3 - 11.5) and 11.2 μM (95% CI: 6.5 - 15.2), respectively. However, the compound 7 and 15 were less potent in inhibiting *N. fowleri* proliferation and had EC_{50} of 20.8 μM (95% CI: 18.9 - 22.8) and 25.8 μM (95% CI: 17.0 - 31.3), respectively in this assay. Activities of these compounds against the cyst stage of *N. fowleri* were not evaluated, since, according to the CDC, cysts are not found in the brain, and are unlikely to be involved in the acute phase of the disease in human [33].

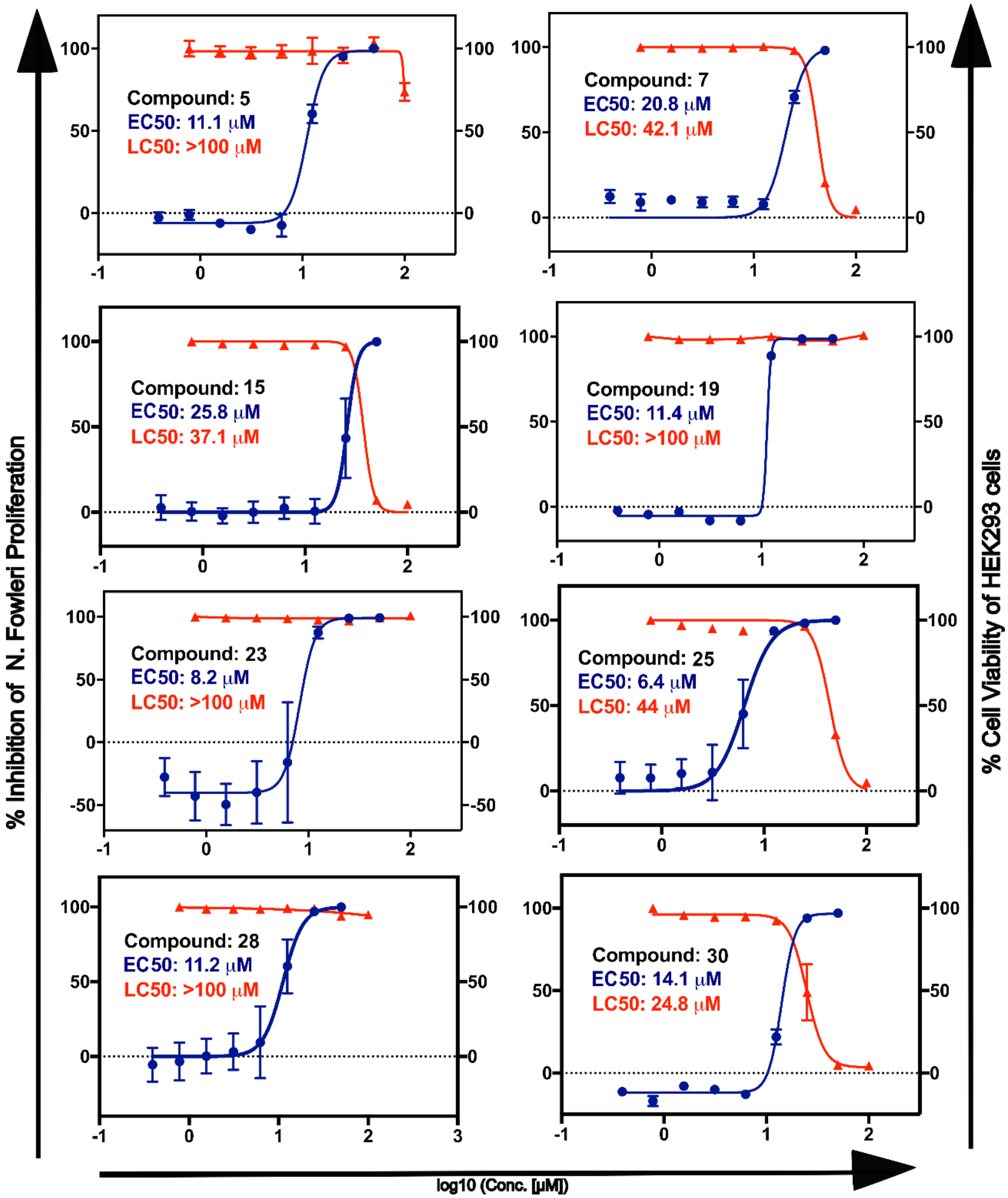


Figure 4-5. Dose-response curves of active compounds for the N. fowleri inhibition (blue) and cell toxicity in human HEK-293 cells (red). The observed dose-response curve with derived EC_{50} value for N. fowleri proliferation inhibition assay, and human cell viability with derived LC_{50} value for HEK293 cell assay are provided for each compound. Image represents the mean and standard error of mean of at least three experiments.

4-3-4. Cytotoxicity assay in HEK-293 cells

To address cytotoxicity of the eight validated compounds, we performed a cell viability assay using HEK-293 cells at serially diluted concentrations of eight active compounds, with a highest concentration of 100 μM. The half-maximal lethal concentrations (LC_{50}s) were estimated from the concentration-response curves (shown as red in Fig. 4-5). The cell viability was assessed after 72-hour incubation of different concentrations of test compounds with HEK-293 cells. Cell viability was determined using Alamar Blue assay [33]. The compounds 7, 15, 25 and 30 were cytotoxic to HEK293 cells, with LC_{50} values of 42.1, 37.1, 44 and 24.8 μM, respectively (95% CI in Table 4-1). Whereas, the compounds 5, 19, 23 and 28 have not showed cytotoxicity at amoebicidal concentrations and did not kill HEK293 cells at concentrations higher than 50 μM (observed LC_{50} more than 100 μM).

Selectivity index (SI) was calculated as the ratio of observed LC_{50} (for HEK293 cells) to EC_{50} (for *N. fowleri* cells). A compound with SI of 10 or more is considered selective according to Quispe and collaborators [196]. Compounds 5, 19, 23 and 28 had a selectivity index (SI) greater than 10 showing low or no cytotoxicity to mammalian HEK-293 cells. These four compounds with strong amoebicidal properties and low human cell toxicities fall into two distinct and novel chemical scaffolds with drug like properties.

4-3-5. Analysis of the docking poses of eight experimentally active compounds

The optimal binding poses of all eight compounds were well-defined and are consistent with the asymmetric shape of the binding pocket (see Fig. 4-6). Shown in Fig. 4-6, all eight compounds (represented as yellow sticks) showed good space and surface property fits in NfERG2 binding cavity (presented as red mesh). The detailed ligand-target interactions are shown in the appendix figures 4-17 – 4-24 as 2D interaction diagrams. In all predicted binding poses, residues Y163 and E232 are involved in the ligand-target interactions, which is consistent with the conserved nature of these residues in the ERG2 protein family (Fig. 4-2a). In addition to residues Y163 and E232, the docked compounds also form hydrogen bonds with N144, C146, and/or Y167 of NfERG2, which confers the polar group complementarity.

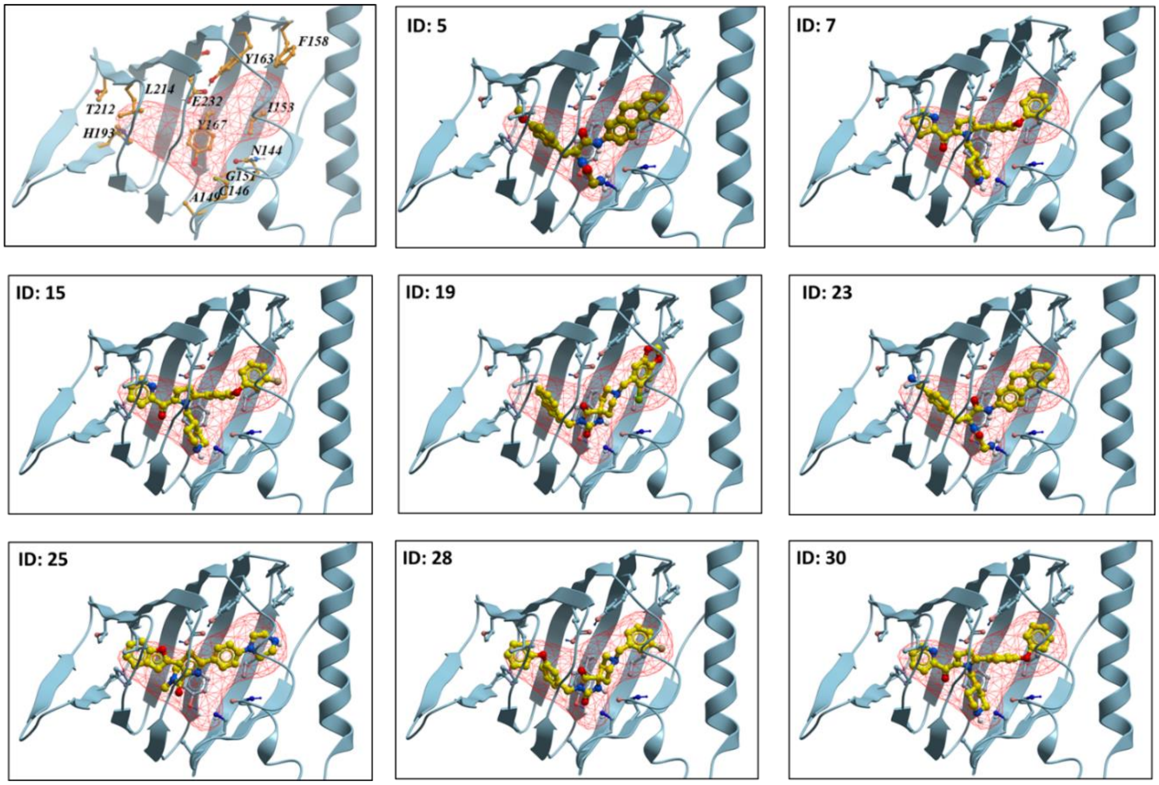
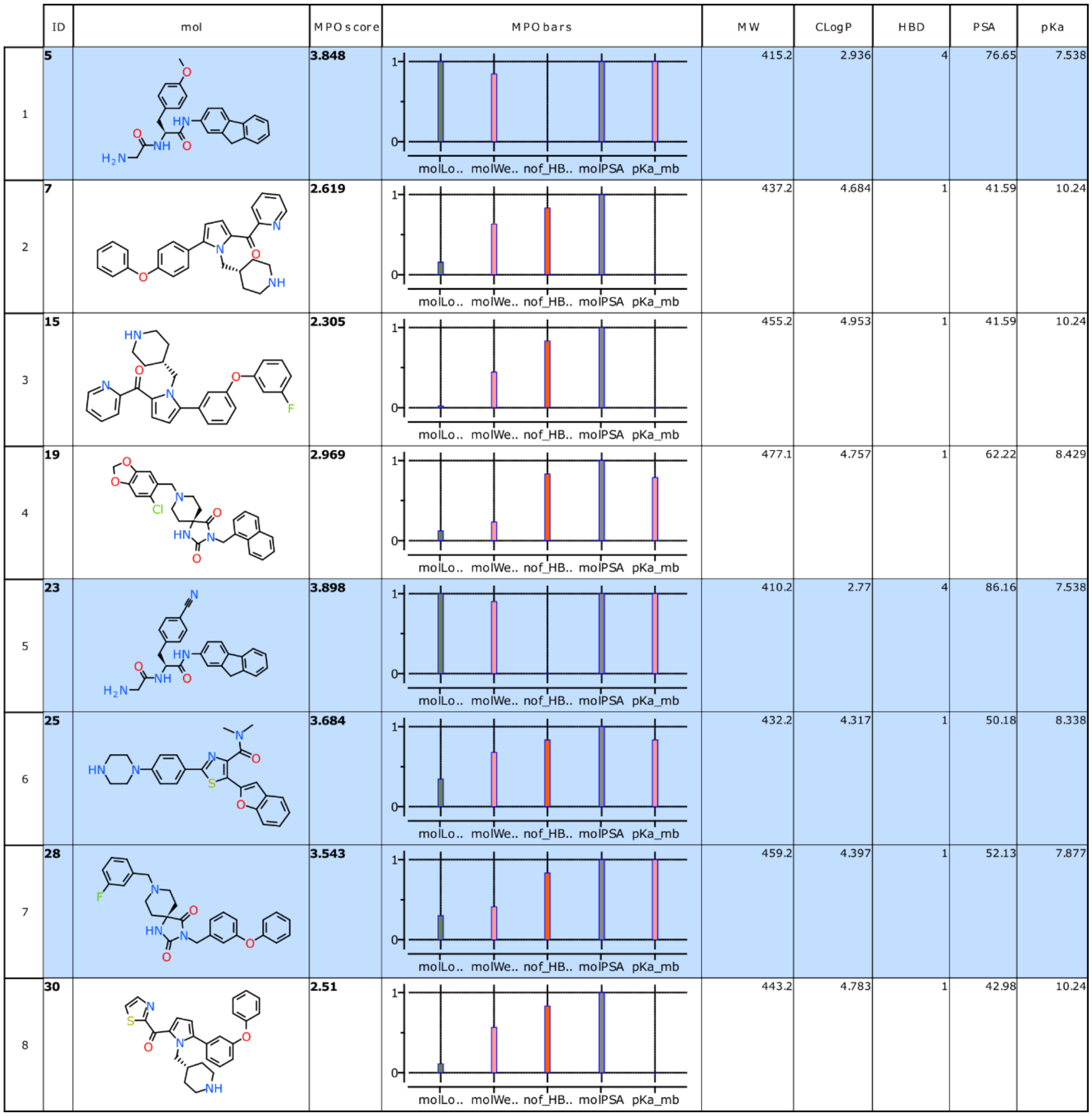


Figure 4-6. Predicted docking poses of the active compounds in the binding cavity of NfERG2. NfERG2 is shown as a ribbon, compounds are shown in ball-and-stick mode, and binding pocket of NfERG2 predicted by homology modelling is shown in red mesh.

4-3-6. Brain permeability assessment of active compounds

To be active against *N. fowleri* residing in the CNS, the drugs must be able to cross the blood-brain barrier (BBB). To analyze the BBB permeability properties of the active compounds, we calculated the BBB permeability multi-parameter score (abbreviated as BBB-MPO) for each compound based on five physicochemical parameters (calculated partition coefficient cLogP, molecular weight MW, topological polar surface area PSA, number of hydrogen bond donors HBD, and pKa of the most basic center). The calculated physicochemical parameters used to estimate the BBB-MPO scores of the active compounds are shown in Fig 4-7.


1
0
molLo.. molWe.. nof_HB.. molPSA pKa_mb

	ID	mol	MPO score	MPO bars	MW	CLogP	HBD	PSA	pKa
1	**5**		**3.848**		415.2	2.936	4	76.65	7.538
2	**7**		**2.619**		437.2	4.684	1	41.59	10.24
3	**15**		**2.305**		455.2	4.953	1	41.59	10.24
4	**19**		**2.969**		477.1	4.757	1	62.22	8.429
5	**23**		**3.898**		410.2	2.77	4	86.16	7.538
6	**25**		**3.684**		432.2	4.317	1	50.18	8.338
7	**28**		**3.543**		459.2	4.397	1	52.13	7.877
8	**30**		**2.51**		443.2	4.783	1	42.98	10.24

Figure 4-7. Physicochemical properties for the MPO scores of active chemicals.

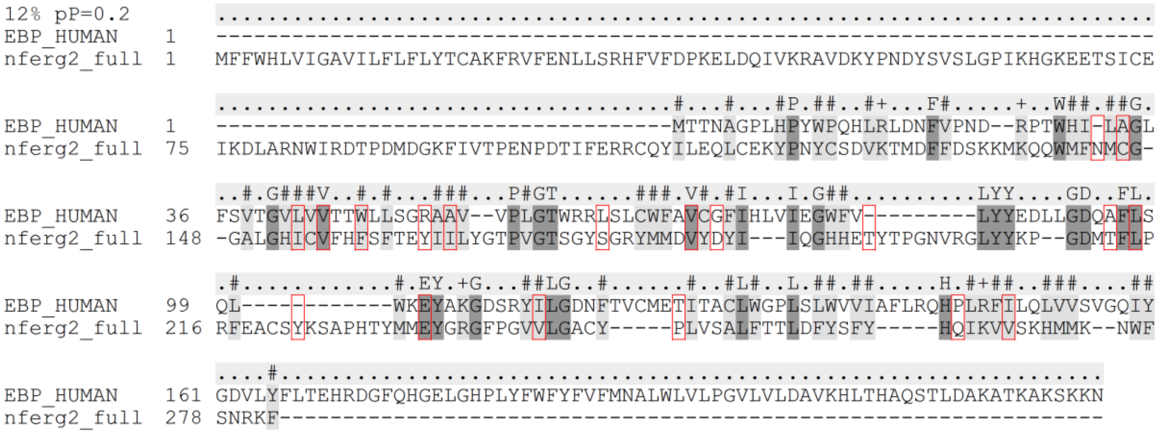
The range of BBB-MPO score, originally introduced by Pfizer scientists in 2010 [179-180], is from zero to five, and the majority of known BBB permeable drugs have BBB-MPO scores higher than or equal to 3. Four compounds 5, 23, 25 and 28 (out of eight active compounds) have shown BBB-MPO scores greater than 3, indicating their potential for BBB permeability (Table 4-1). Compound 19 showed a calculated BBB-MPO score of 2.97, which is very close to 3, and is likely to cross BBB to some extent. For the other three compounds, further modifications may be needed to increase their BBB permeability. Furthermore, the compounds with desired BBB-MPO scores also had lower EC_{50} values in the cell-based assay, making them promising candidates for further evaluation and development.

Table 4-1. BBB multi-parameter optimization (BBB-MPO) score of active compounds. Compounds with BBB-MPO scores >3 are highlighted in blue.

Compound ID	Docking Score	BBB-MPO Score	EC_{50} (μM)	EC_{50} 95% CI	LC_{50} (μM)	LC_{50} 95% CI
5	-34.67	3.85	11.1	10.1 - 12.6	> 100	N.A.
7	-33.42	2.62	20.8	18.9 - 22.8	42.1	40.5 - 43.4
15	-31.72	2.31	25.8	17.0 - 31.3	37.1	36.3 - 37.9
19	-31.29	2.97	11.4	11.3 - 11.5	> 100	N.A.
23	-30.94	3.90	8.2	4.6 - 15.5	> 100	N.A.
25	-30.59	3.68	6.4	4.3 - 9.2	44	47.9 - 48.8
28	-30.28	3.54	11.2	6.5 - 15.2	> 100	N.A.
30	-30.07	2.51	14.1	13 - 15.3	24.8	19.8 - 29.8

4-4. Discussion

Sterol biosynthesis is a basic metabolic pathway of eukaryotes giving rise to essential membrane components. For the purpose of drug discovery, it offers an array of druggable molecular targets accessible for homology modeling and structure-based screening. Targeting sterol Δ8–Δ7-isomerase (ERG2) in *N. fowleri*, an enzyme without an ortholog in human proteome, opens up a possibility of identifying novel drug candidates for the treatment of PAM [164, 166]. Emopamil-binding protein (EBP), a functional counterpart of ERG2 in humans, shares no sequence similarity with ERG2 (Fig. 4-8). Recently released crystal structure of human EBP (PDB IDs: 6OHT, 6OHU) is topologically different from ERG2 [190]. While no ERG2 enzymes from any species have been characterized crystallographically, we found a receptor with sufficient sequence similarity and 2.5Å-resolution structure. Based on the structural similarity between the catalytic domain of ERG2 and the human σ$_1$ non-opioid receptor, we built a homology model for NfERG2. This model was sufficient to identify novel inhibitory scaffolds with good efficacy.



```
12% pP=0.2         ..........................................................................
EBP_HUMAN     1    --------------------------------------------------------------------------
nferg2_full   1    MFFWHLVIGAVILFLFLYTCAKFRVFENLLSRHFVFDPKELDQIVKRAVDKYPNDYSVSLGPIKHGKEETSICE

                   ....................................#...#...#P.##..#+...F#.....+..W##.##G.
EBP_HUMAN     1    ------------------------------------MTTNAGPLHPYWPQHLRLDNFVPND--RPTWHI-LAGL
nferg2_full   75   IKDLARNWIRDTPDMDGKFIVTPENPDTIFERRCQYILEQLCEKYPNYCSDVKTMDFFDSKKMKQQWMFNMCG-

                   ..#.G###V..#.#...###...P#GT......###.V#.#I...I.G##..........LYY....GD..FL.
EBP_HUMAN     36   FSVTGVLVVTTWLLSGRAAV--VPLGTWRRLSLCWFAVCGFIHLVIEGWFV---------LYYEDLLGDQAFLS
nferg2_full   148  -GALGHICVFHFSFTEYIILYGTPVGTSGYSGRYMMDVYDYI---IQGHHETYTPGNVRGLYYKP--GDMTFLP

                   .#............#.EY.+G...##LG..#......#..#L#..L.##.##.....H.#+##..###....##
EBP_HUMAN     99   QL------------WKEYAKGDSRYILGDNFTVCMETITACLWGPLSLWVVIAFLRQHPLRFILQLVVSVGQIY
nferg2_full   216  RFEACSYKSAPHTYMMEYGRGFPGVVLGACY-----PLVSALFTTLDFYSFY-----HQIKVVSKHMMK--NWF

                   ....#.................................................................
EBP_HUMAN     161  GDVLYFLTEHRDGFQHGELGHPLYFWFYFVFMNALWLVLPGVLVLDAVKHLTHAQSTLDAKATKAKSKKN
nferg2_full   278  SNRKF-----------------------------------------------------------------
```

Figure 4-8. Sequence alignment of full length NfERG2 and human homolog EBP. Positions with red boxes correspond to residues of the binding pocket in the NfERG2 model. The low identity percentage and pP value illustrate the lack of topological similarity between the two proteins.

The predictive power of the computational methods depends on the reliability of the model used for *in silico* screening. Even though the sequence identity between the template and NfERG2 was moderate, the chosen strategy worked well due to the strong conservation of the backbone topology evidenced by lack of the insertions or deletions in the protein scaffold around the binding pocket. A relatively high hit rate (8 active hits out of 30 experimentally tested predictions) further implies reasonably good conformation accuracy of the binding pocket. The EC$_{50}$ of these eight hits was confirmed to be in a low micromolar range. As an added bonus, four compounds had selectivity index greater than 10 against human HEK-293 cells and high blood-brain barrier permeability scores.

The top four validated compounds fall into two novel chemical scaffolds.

Compound 5 and 23 both contain a fluorene moiety which fits well into the binding pocket and also contains a nucleophilic carbon in the center feasible for further chemical modifications. Compound 19 and 28 share the same piperidine-spirohydantoin core that doesn't exist in approved drugs, and have two different attachments that can be modified to optimize the efficacy, specificity, or ADMET properties.

The results also emphasize that homology modeling of essential targets in rare pathogen (followed by docking screen of a large chemical library and experimental testing of top hits) is a useful initial strategy even in absence of crystal structures of those targets [197-198]. The chemical diversity of identified hits is a further evidence of this approach.

As far as target specificity is concerned, it is most likely that the identified compounds are not uniquely specific to NfERG2. The ergosterol biosynthesis pathway includes multiple reaction steps catalyzed by different enzymes utilizing structurally similar substrates. It is quite common for compounds to act on more than one enzyme in the sterol biosynthetic pathway [164, 199]. Further studies are needed to characterize the full profile of affected molecular targets responsible for the amoebicidal mechanism of identified hits.

4-5. Conclusion

In this chapter, we identified four novel compounds as potential drug candidates for a highly deadly parasitic disease, Primary amoebic meningoencephalitis. The target we focused on is ERG2, an enzyme in the ergosterol biosynthetic pathway. The first obstacle we encountered was that there was no experimentally solved structure of ERG2. Fortunately, we were able to find a human protein, σ1 non-opioid receptor, having a strikingly high homology with ERG2. Therefore, we built a homology model of ERG2 based on the human σ1 non-opioid receptor. Even though the template we use for homology modeling doesn't share a very high sequence identity with ERG2, we were still able to build a decent model because of lack of a relatively conserved backbone topology without big insertions and deletions. Our models was proved effective as we achieved a quite high hit rate of the screening compounds, as 8 out of 30 tested. Among the active compounds, four candidates also showed low HEK-293 cell toxicities and acceptable brain permeability scores. The next steps would include compound optimization, formulation and testing in an animal model of PAM, either alone or in combination with other treatments.

4-6. Appendix

4-6-1. LC-MS/MS chromatograms of the active compounds

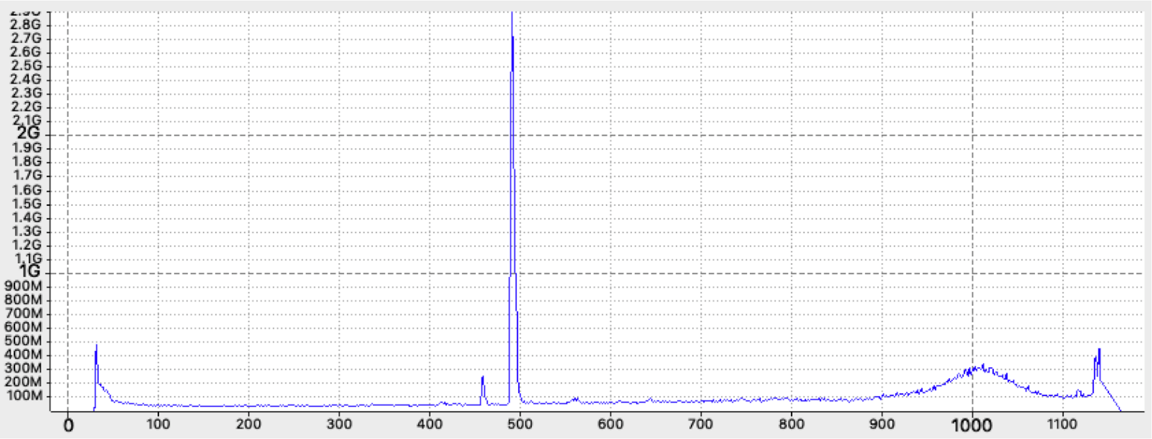


Figure 4-9. Chromatogram from the LC-MS/MS analysis of chemical 5 (total ion current chromatogram). (x axis = retention time (sec), y axis = ion intensity (count).

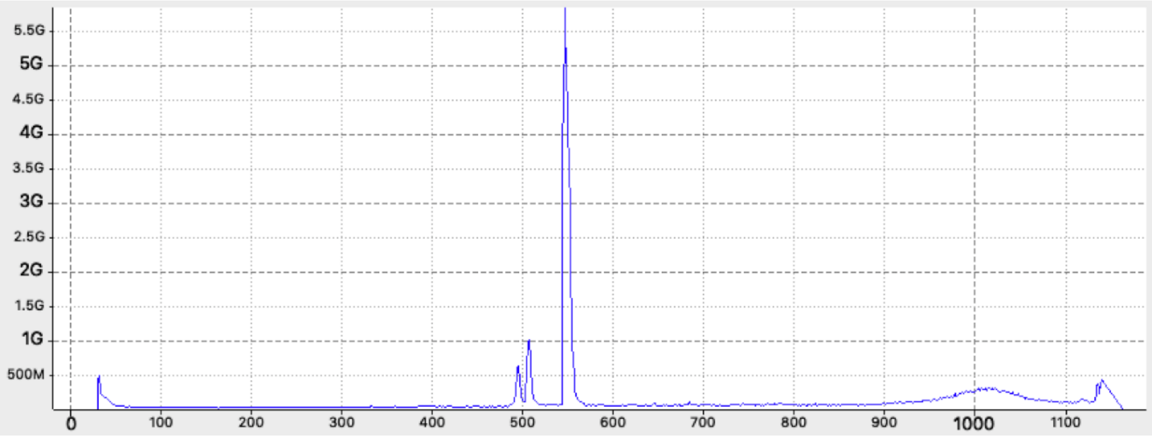


Figure 4-10. Chromatogram from the LC-MS/MS analysis of chemical 7 (total ion current chromatogram). (x axis = retention time (sec), y axis = ion intensity (count).

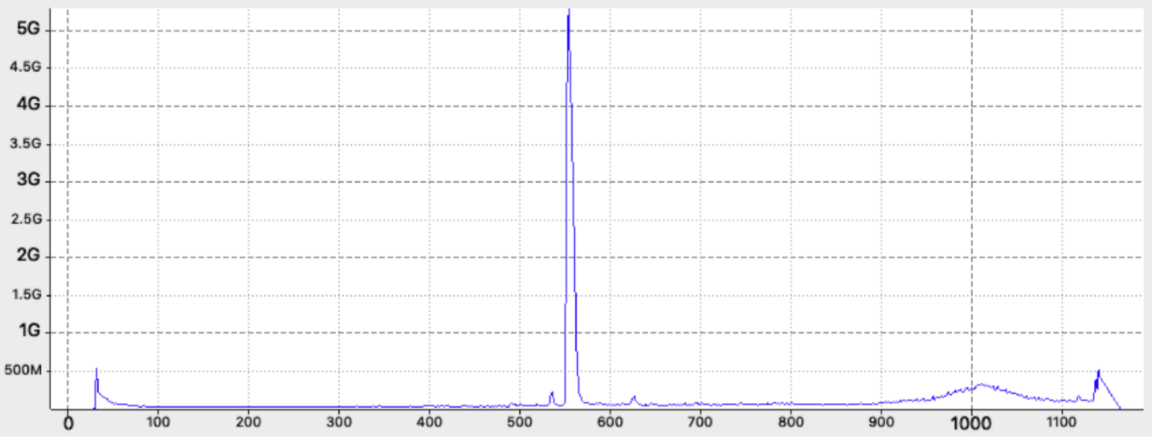


Figure 4-11. Chromatogram from the LC-MS/MS analysis of chemical 15 (total ion current chromatogram). (x axis = retention time (sec), y axis = ion intensity (count).

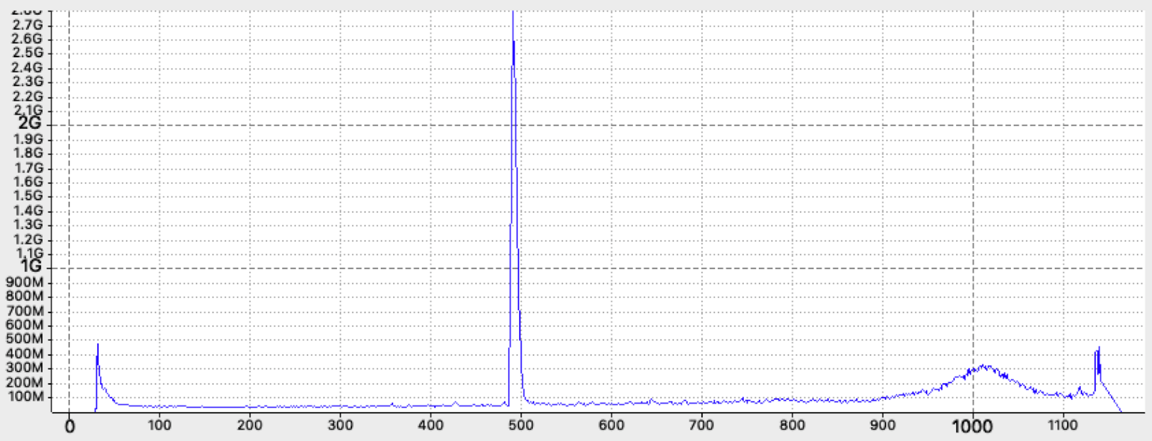


Figure 4-12. Chromatogram from the LC-MS/MS analysis of chemical 19 (total ion current chromatogram). (x axis = retention time (sec), y axis = ion intensity (count).

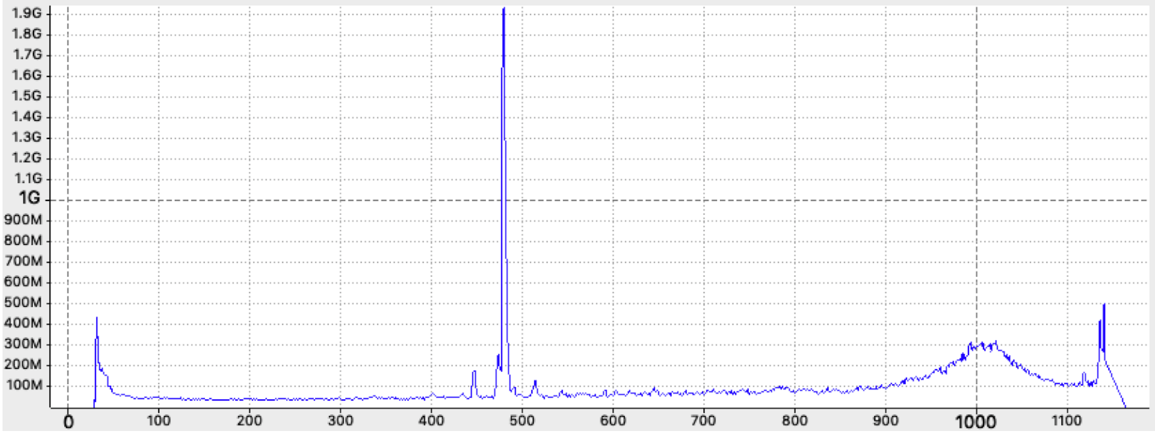


Figure 4-13. Chromatogram from the LC-MS/MS analysis of chemical 23 (total ion current chromatogram). (x axis = retention time (sec), y axis = ion intensity (count).

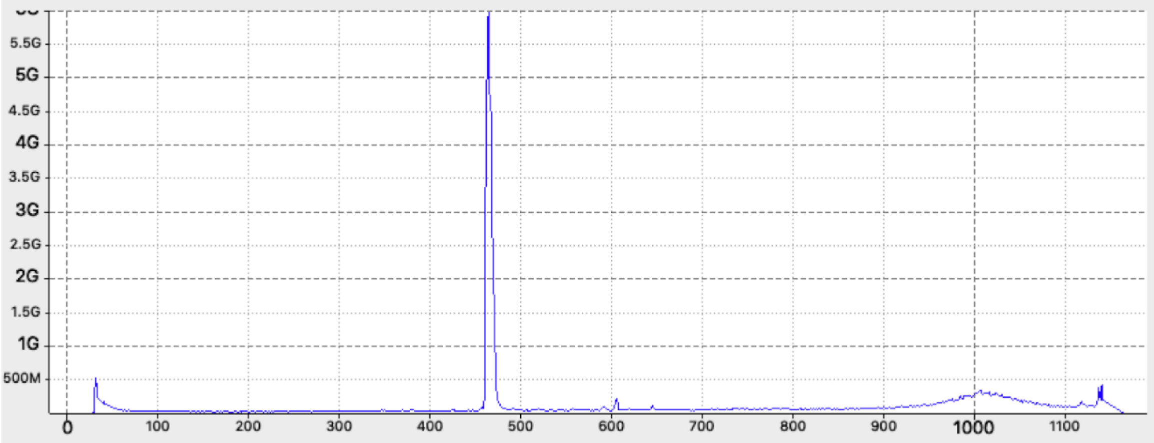


Figure 4-14. Chromatogram from the LC-MS/MS analysis of chemical 25 (total ion current chromatogram). (x axis = retention time (sec), y axis = ion intensity (count).

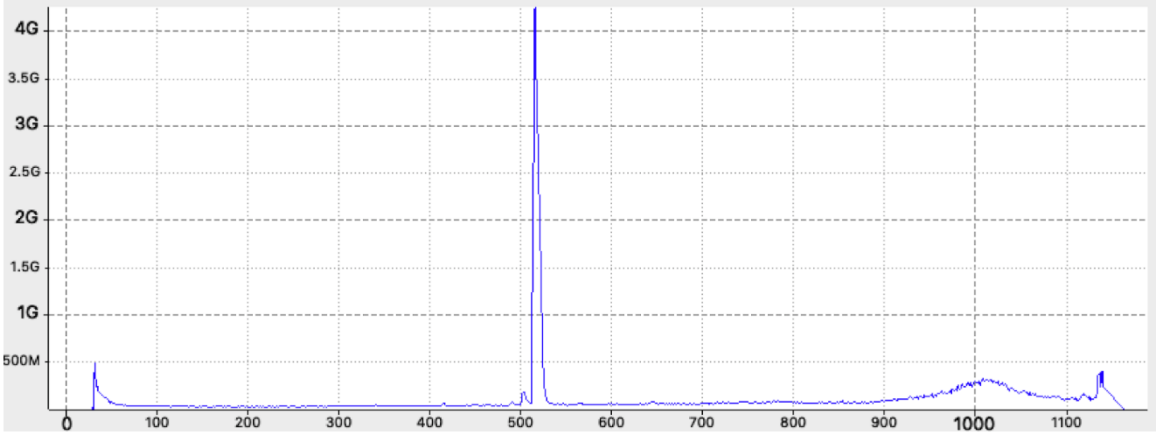


Figure 4-15. Chromatogram from the LC-MS/MS analysis of chemical 28 (total ion current chromatogram). (x axis = retention time (sec), y axis = ion intensity (count).

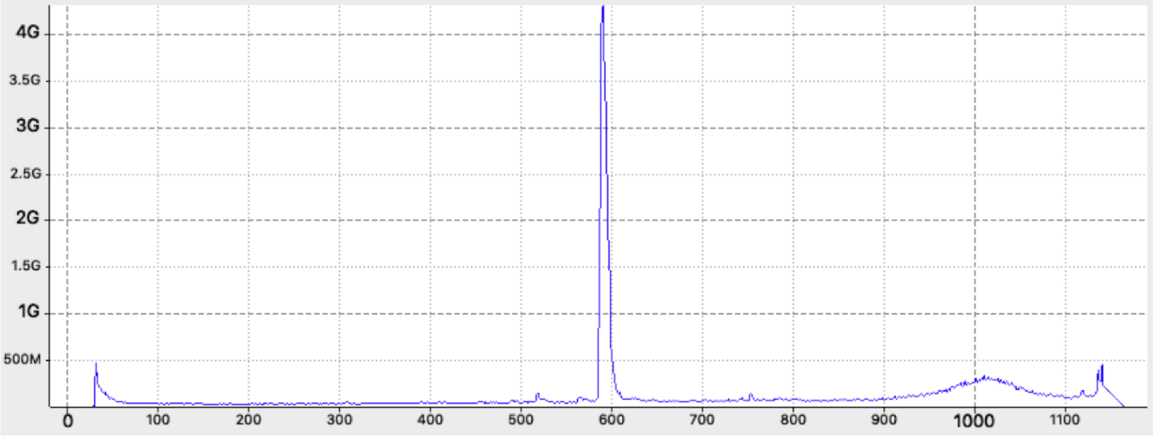


Figure 4-16. Chromatogram from the LC-MS/MS analysis of chemical 30 (total ion current chromatogram). (x axis = retention time (sec), y axis = ion intensity (count).

4-6-2. 2D diagrams of the interactions between active compounds and ERG2

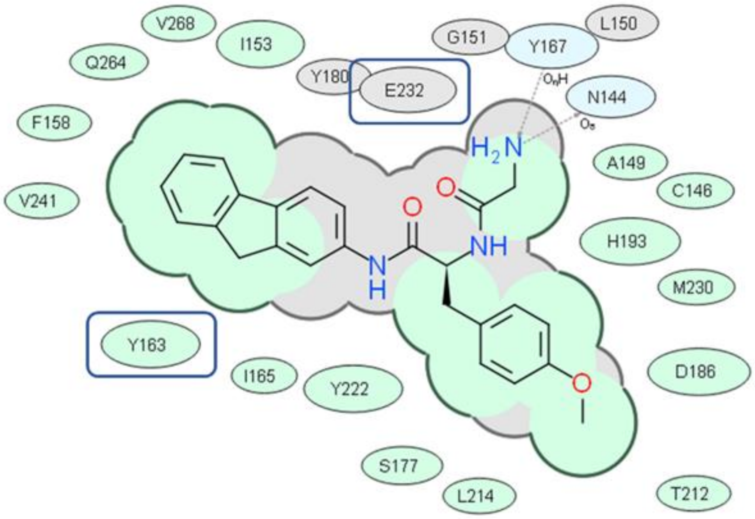


Figure 4-17. 2D diagram of interactions between chemical 5 and NfERG2 model. The cutoff distance of hydrophobic interaction between ligand and protein sidechains was set to 4.5 Å. Green shading represents hydrophobic region. White dashed arrows represent hydrogen bonds. Grey parabolas represent accessible surface for large areas. Broken thick line around ligand shape indicates accessible surface. Size of residue ellipse represents the strength of the contact. Distance between residue label and ligand represents proximity. Two potentially critical residues (Y163 and E232) are marked with blue squares.

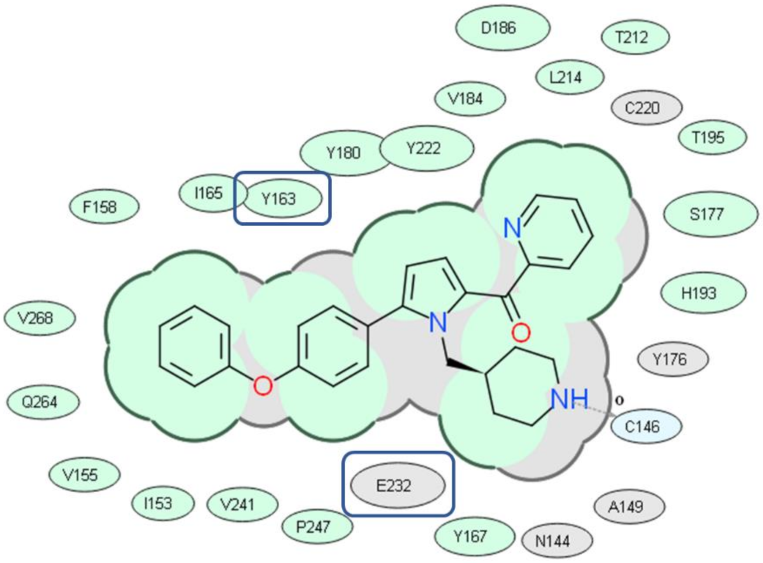


Figure 4-18. 2D diagram of interactions between chemical 7 and NfERG2 model. The cutoff distance of hydrophobic interaction between ligand and protein sidechains was set to 4.5 Å. Green shading represents hydrophobic region. White dashed arrows represent hydrogen bonds. Grey parabolas represent accessible surface for large areas. Broken thick line around ligand shape indicates accessible surface. Size of residue ellipse represents the strength of the contact. Distance between residue label and ligand represents proximity. Two potentially critical residues (Y163 and E232) are marked with blue squares.

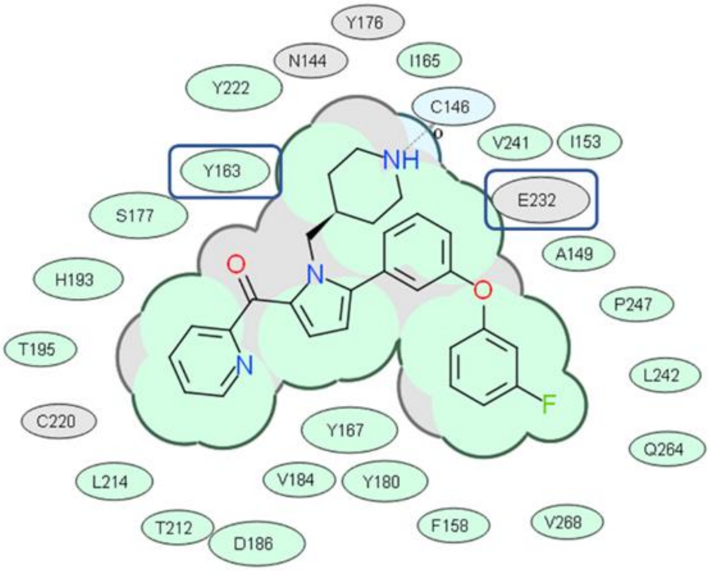


Figure 4-19. 2D diagram of interactions between chemical 15 and NfERG2 model. The cutoff distance of hydrophobic interaction between ligand and protein sidechains was set to 4.5 Å. Green shading represents hydrophobic region. White dashed arrows represent hydrogen bonds. Grey parabolas represent accessible surface for large areas. Broken thick line around ligand shape indicates accessible surface. Size of residue ellipse represents the strength of the contact. Distance between residue label and ligand represents proximity. Two potentially critical residues (Y163 and E232) are marked with blue squares.

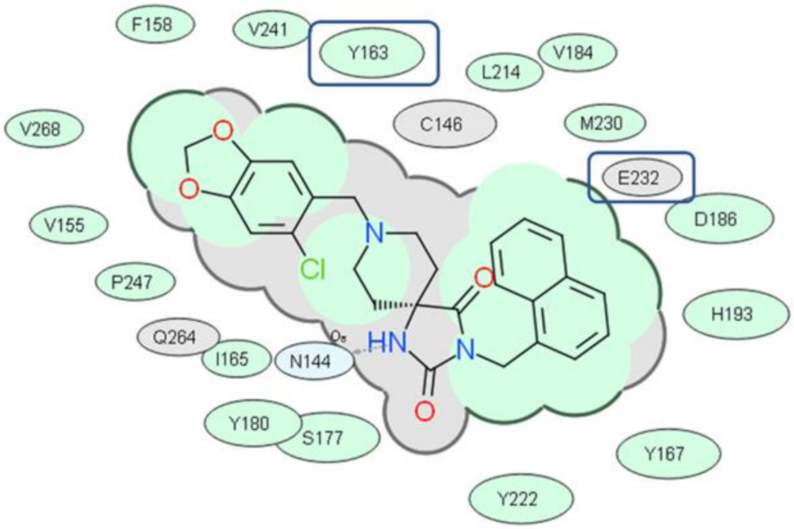


Figure 4-20. 2D diagram of interactions between chemical 19 and NfERG2 model. The cutoff distance of hydrophobic interaction between ligand and protein sidechains was set to 4.5 Å. Green shading represents hydrophobic region. White dashed arrows represent hydrogen bonds. Grey parabolas represent accessible surface for large areas. Broken thick line around ligand shape indicates accessible surface. Size of residue ellipse represents the strength of the contact. Distance between residue label and ligand represents proximity. Two potentially critical residues (Y163 and E232) are marked with blue squares.

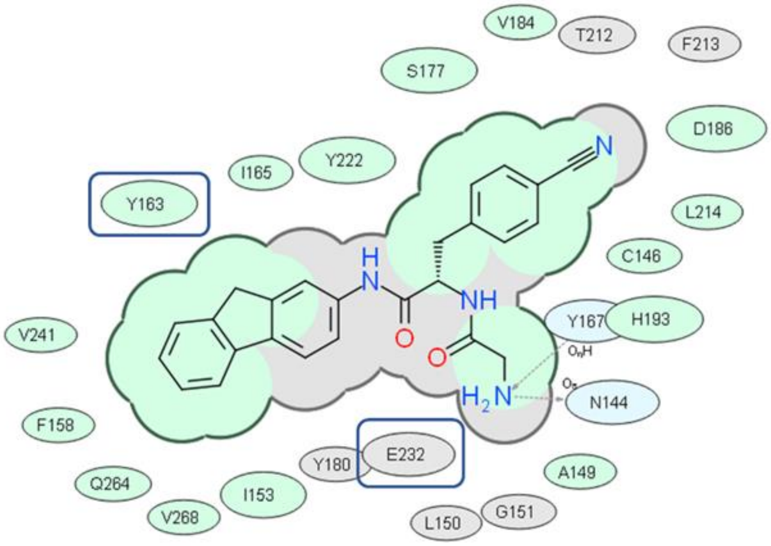


Figure 4-21. 2D diagram of interactions between chemical 23 and NfERG2 model. The cutoff distance of hydrophobic interaction between ligand and protein sidechains was set to 4.5 Å. Green shading represents hydrophobic region. White dashed arrows represent hydrogen bonds. Grey parabolas represent accessible surface for large areas. Broken thick line around ligand shape indicates accessible surface. Size of residue ellipse represents the strength of the contact. Distance between residue label and ligand represents proximity. Two potentially critical residues (Y163 and E232) are marked with blue squares.

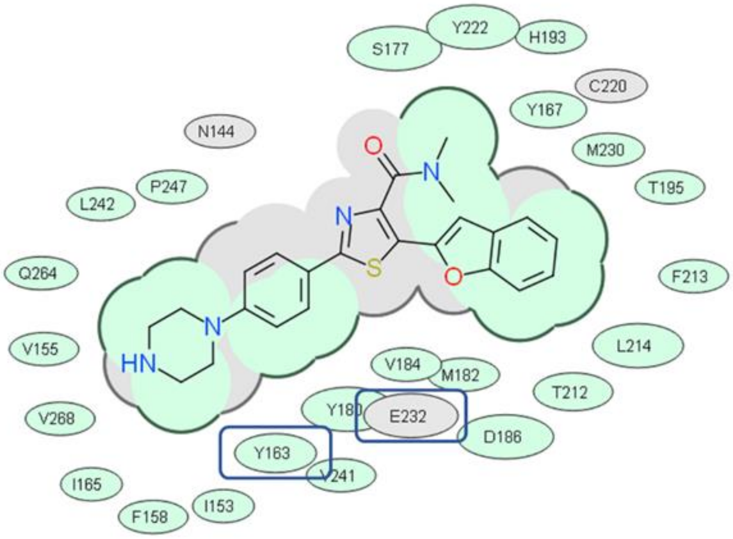


Figure 4-22. 2D diagram of interactions between chemical 25 and NfERG2 model. The cutoff distance of hydrophobic interaction between ligand and protein sidechains was set to 4.5 Å. Green shading represents hydrophobic region. White dashed arrows represent hydrogen bonds. Grey parabolas represent accessible surface for large areas. Broken thick line around ligand shape indicates accessible surface. Size of residue ellipse represents the strength of the contact. Distance between residue label and ligand represents proximity. Two potentially critical residues (Y163 and E232) are marked with blue squares.

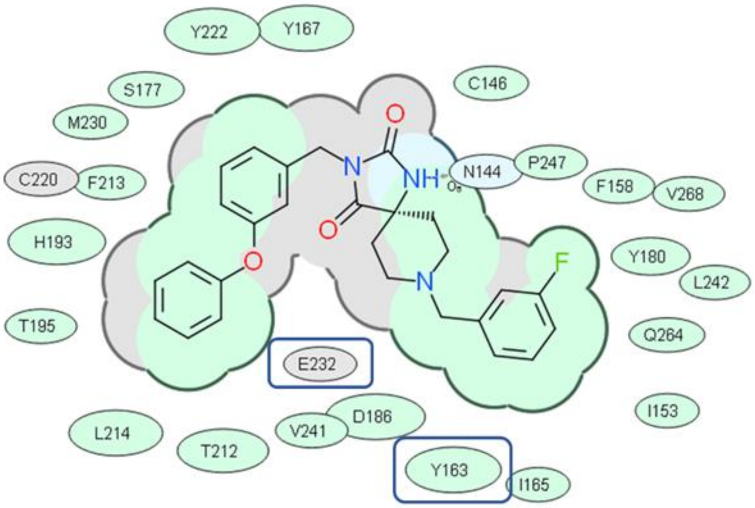


Figure 4-23. 2D diagram of interactions between chemical 28 and NfERG2 model. The cutoff distance of hydrophobic interaction between ligand and protein sidechains was set to 4.5 Å. Green shading represents hydrophobic region. White dashed arrows represent hydrogen bonds. Grey parabolas represent accessible surface for large areas. Broken thick line around ligand shape indicates accessible surface. Size of residue ellipse represents the strength of the contact. Distance between residue label and ligand represents proximity. Two potentially critical residues (Y163 and E232) are marked with blue squares.

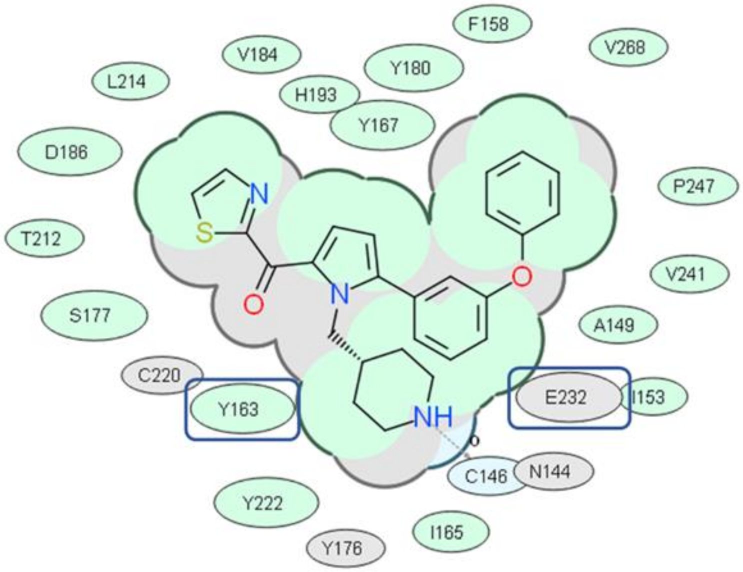


Figure 4-24. 2D diagram of interactions between chemical 30 and NfERG2 model. The cutoff distance of hydrophobic interaction between ligand and protein sidechains was set to 4.5 Å. Green shading represents hydrophobic region. White dashed arrows represent hydrogen bonds. Grey parabolas represent accessible surface for large areas. Broken thick line around ligand shape indicates accessible surface. Size of residue ellipse represents the strength of the contact. Distance between residue label and ligand represents proximity. Two potentially critical residues (Y163 and E232) are marked with blue squares.

4-7. Acknowledgement

Chapter 4, in full, has been published in **Shi Da**, Kirti Kandhwal Chahal, Patricia Oto, Louis-Felix Nothias, Anjan Debnath, James H McKerrow, Larissa M Podust, and Ruben Abagyan. "Identification of four amoebicidal non-toxic compounds by a molecular docking screen of Naegleria fowleri sterol Δ8− Δ7-isomerase and phenotypic assays." ACS infectious diseases (2019). The dissertation author was the primary investigator and author of this paper.

REFERENCES

1. Definition of a Drug (April 2017). https://www.fda.gov/drugs/information-healthcare-professionals-drugs/definition-drug-april-2017.

2. Murray, G. R., The Life-History of the First Case of Myxoedema Treated by Thyroid Extract. *British Medical Journal* **1920,** *1* (3089), 359.

3. Szuromi, P.; Vinson, V.; Marshall, E., Rethinking Drug Discovery. *Science* **2004,** *303* (5665), 1795.

4. Hanahan, D.; Weinberg, Robert A., Hallmarks of Cancer: The Next Generation. *Cell* **2011,** *144* (5), 646-674.

5. Hutchinson, L.; Kirk, R., High Drug Attrition Rates—Where Are We Going Wrong? *Nat. Rev. Clin. Oncol.* **2011,** *8*, 189-190.

6. Hopkins, A. L., Network Pharmacology: The Next Paradigm in Drug Discovery. *Nat. Chem. Biol.* **2008,** *4*, 682-690.

7. Kaelin Jr, W. G., The Concept of Synthetic Lethality in the Context of Anticancer Therapy. *Nat. Rev. Cancer* **2005,** *5*, 689-698.

8. Campillos, M.; Kuhn, M.; Gavin, A.-C.; Jensen, L. J.; Bork, P., Drug Target Identification Using Side-Effect Similarity. *Science* **2008,** *321* (5886), 263-266.

9. Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L., How Many Drug Targets Are There? *Nat. Rev. Drug Discovery* **2006,** *5*, 993-996.

10. Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; Assempour, N.; Iynkkaran, I.; Liu, Y.; Maciejewski, A.; Gale, N.; Wilson, A.; Chin, L.; Cummings, R.; Le, D.; Pon, A.; Knox, C.; Wilson, M., DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* **2018,** *46* (D1), D1074-D1082.

11. Li, Y. H.; Yu, C. Y.; Li, X. X.; Zhang, P.; Tang, J.; Yang, Q.; Fu, T.; Zhang, X.; Cui, X.; Tu, G.; Zhang, Y.; Li, S.; Yang, F.; Sun, Q.; Qin, C.; Zeng, X.; Chen, Z.; Chen, Y. Z.; Zhu, F., Therapeutic Target Database Update 2018: Enriched Resource for Facilitating Bench-to-Clinic Research of Targeted Therapeutics. *Nucleic Acids Res.* **2018,** *46* (D1), D1121-D1127.

12. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R., The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017,** *45* (D1), D945-D954.

13. Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H., PubChem

Substance and Compound Databases. *Nucleic Acids Res.* **2016,** *44* (D1), D1202-D1213.

14. Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J., BindingDB in 2015: A Public Database for Medicinal Chemistry, Computational Chemistry and Systems Pharmacology. *Nucleic Acids Res.* **2016,** *44* (D1), D1045-D1053.

15. Günther, S.; Kuhn, M.; Dunkel, M.; Campillos, M.; Senger, C.; Petsalaki, E.; Ahmed, J.; Urdiales, E. G.; Gewiess, A.; Jensen, L. J.; Schneider, R.; Skoblo, R.; Russell, R. B.; Bourne, P. E.; Bork, P.; Preissner, R., SuperTarget and Matador: Resources for Exploring Drug-Target Relationships. *Nucleic Acids Res.* **2008,** *36* (suppl_1), D919-D922.

16. Yıldırım, M. A.; Goh, K.-I.; Cusick, M. E.; Barabási, A.-L.; Vidal, M., Drug—Target Network. *Nat. Biotechnol.* **2007,** *25*, 1119-1126.

17. Cheng, F.; Liu, C.; Jiang, J.; Lu, W.; Li, W.; Liu, G.; Zhou, W.; Huang, J.; Tang, Y., Prediction of Drug-Target Interactions and Drug Repositioning via Network-Based Inference. *PLoS Comput. Biol.* **2012,** *8* (5), e1002503.

18. Lu, J.-J.; Pan, W.; Hu, Y.-J.; Wang, Y.-T., Multi-Target Drugs: The Trend of Drug Research and Development. *PLoS One* **2012,** *7* (6), e40262.

19. Santos, R.; Ursu, O.; Gaulton, A.; Bento, A. P.; Donadi, R. S.; Bologa, C. G.; Karlsson, A.; Al-Lazikani, B.; Hersey, A.; Oprea, T. I.; Overington, J. P., A Comprehensive Map of Molecular Drug Targets. *Nat. Rev. Drug Discovery* **2016,** *16*, 19-34.

20. Hert, J.; Keiser, M. J.; Irwin, J. J.; Oprea, T. I.; Shoichet, B. K., Quantifying the Relationships among Drug Classes. *J. Chem. Inf. Model.* **2008,** *48* (4), 755-765.

21. Cancer Drugs - National Cancer Institute. https://www.cancer.gov/about-cancer/treatment/drugs (accessed 2018/08/24).

22. Bento, A. P.; Gaulton, A.; Hersey, A.; Krüger, F. A.; Papadatos, G.; Chambers, J.; Mak, L.; Bellis, L. J.; Davies, M.; Nowotka, M.; Santos, R.; McGlinchey, S.; Light, Y.; Overington, J. P., The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.* **2013,** *42* (D1), D1083-D1090.

23. Fang, H.; Su, Z.; Wang, Y.; Miller, A.; Liu, Z.; Howard, P. C.; Tong, W.; Lin, S. M., Exploring the FDA Adverse Event Reporting System to Generate Hypotheses for Monitoring of Disease Characteristics. *Clin. Pharmacol. Ther.* **2014,** *95* (5), 496-498.

24. Makunts, T.; Cohen, I. V.; Awdishu, L.; Abagyan, R., Analysis of Postmarketing Safety Data for Proton-Pump Inhibitors Reveals Increased Propensity for Renal Injury, Electrolyte Abnormalities, and Nephrolithiasis. *Sci. Rep.* **2019,** *9* (1), 2282.

25. Barretina, J.; Caponigro, G.; Stransky, N.; Venkatesan, K.; Margolin, A. A.; Kim, S.; Wilson, C. J.; Lehár, J.; Kryukov, G. V.; Sonkin, D.; Reddy, A.; Liu, M.; Murray, L.; Berger, M. F.; Monahan, J. E.; Morais, P.; Meltzer, J.; Korejwa, A.; Jané-Valbuena, J.;

Mapa, F. A.; Thibault, J.; Bric-Furlong, E.; Raman, P.; Shipway, A.; Engels, I. H.; Cheng, J.; Yu, G. K.; Yu, J.; Aspesi, P.; de Silva, M.; Jagtap, K.; Jones, M. D.; Wang, L.; Hatton, C.; Palescandolo, E.; Gupta, S.; Mahan, S.; Sougnez, C.; Onofrio, R. C.; Liefeld, T.; MacConaill, L.; Winckler, W.; Reich, M.; Li, N.; Mesirov, J. P.; Gabriel, S. B.; Getz, G.; Ardlie, K.; Chan, V.; Myer, V. E.; Weber, B. L.; Porter, J.; Warmuth, M.; Finan, P.; Harris, J. L.; Meyerson, M.; Golub, T. R.; Morrissey, M. P.; Sellers, W. R.; Schlegel, R.; Garraway, L. A., The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity. *Nature* **2012,** *483*, 603-607.

26. Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T., Gene Ontology: Tool for the Unification of Biology. *Nat. Genet.* **2000,** *25* (1), 25-29.

27. Popescu, M.; Keller, J. M.; Mitchell, J. A., Fuzzy Measures on the Gene Ontology for Gene Product Similarity. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2006,** *3* (3), 263-274.

28. Mistry, M.; Pavlidis, P., Gene Ontology Term Overlap as a Measure of Gene Functional Similarity. *BMC Bioinf.* **2008,** *9* (1), 327.

29. Pesquita, C.; Faria, D.; Falcão, A. O.; Lord, P.; Couto, F. M., Semantic Similarity in Biomedical Ontologies. *PLoS Comput. Biol.* **2009,** *5* (7), e1000443.

30. Mamano, N.; Hayes, W. B., SANA NetGO: A Combinatorial Approach to Using Gene Ontology (GO) Terms to Score Network Alignments. *Bioinformatics* **2017,** *34* (8), 1345-1352.

31. The UniProt Consortium, UniProt: The Universal Protein Knowledgebase. *Nucleic Acids Res.* **2017,** *45* (D1), D158-D169.

32. Ellson, J.; Gansner, E. R.; Koutsofios, E.; North, S. C.; Woodhull, G., Graphviz and Dynagraph — Static and Dynamic Graph Drawing Tools. In *Graph Drawing Software*, Jünger, M.; Mutzel, P., Eds. Springer Berlin Heidelberg: Berlin, Heidelberg, 2004; pp 127-148.

33. O'Brien, J.; Wilson, I.; Orton, T.; Pognan, F., Investigation of the Alamar Blue (resazurin) fluorescent dye for the assessment of mammalian cell cytotoxicity. *European Journal of Biochemistry* **2000,** *267* (17), 5421-5426.

34. Giordano, S.; Petrelli, A., From Single-To Multi-Target Drugs in Cancer Therapy: When Aspecificity Becomes an Advantage. *Curr. Med. Chem.* **2008,** *15* (5), 422-432.

35. Poornima, P.; Kumar, J. D.; Zhao, Q.; Blunder, M.; Efferth, T., Network Pharmacology of Cancer: From Understanding of Complex Interactomes to the Design of Multi-Target Specific Therapeutics From Nature. *Pharmacol. Res.* **2016,** *111*, 290-302.

36. Tang, J.; Aittokallio, T., Network Pharmacology Strategies Toward Multi-Target Anticancer Therapies: From Computational Models to Experimental Design Principles. *Curr. Pharm. Des.* **2014,** *20* (1), 23-36.

37. Hopkins, A. L., Network Pharmacology. *Nat. Biotechnol.* **2007,** *25*, 1110-1111.

38. Duran-Frigola, M.; Siragusa, L.; Ruppin, E.; Barril, X.; Cruciani, G.; Aloy, P., Detecting Similar Binding Pockets to Enable Systems Polypharmacology. *PLoS Comput. Biol.* **2017,** *13* (6), e1005522.

39. Kufareva, I.; Ilatovskiy, A. V.; Abagyan, R., Pocketome: An Encyclopedia of Small-Molecule Binding Sites in 4D. *Nucleic Acids Res.* **2011,** *40* (D1), D535-D540.

40. Lallous, N.; Dalal, K.; Cherkasov, A.; Rennie, P., Targeting Alternative Sites on the Androgen Receptor to Treat Castration-Resistant Prostate Cancer. *Int. J. Mol. Sci.* **2013,** *14* (6), 12496-12519.

41. Siarheyeva, A.; Senisterra, G.; Allali-Hassani, A.; Dong, A.; Dobrovetsky, E.; Wasney, Gregory A.; Chau, I.; Marcellus, R.; Hajian, T.; Liu, F.; Korboukh, I.; Smil, D.; Bolshan, Y.; Min, J.; Wu, H.; Zeng, H.; Loppnau, P.; Poda, G.; Griffin, C.; Aman, A.; Brown, Peter J.; Jin, J.; Al-awar, R.; Arrowsmith, Cheryl H.; Schapira, M.; Vedadi, M., An Allosteric Inhibitor of Protein Arginine Methyltransferase 3. *Structure* **2012,** *20* (8), 1425-1435.

42. Morphy, R.; Rankovic, Z., Designed Multiple Ligands. An Emerging Drug Discovery Paradigm. *J. Med. Chem.* **2005,** *48* (21), 6523-6543.

43. Gupta, A. K.; Gupta, R. A.; Soni, L. K.; Kaskhedikar, S. G., Exploration of Physicochemical Properties and Molecular Modelling Studies of 2-Sulfonyl-Phenyl-3-Phenyl-Indole Analogs As Cyclooxygenase-2 Inhibitors. *Eur. J. Med. Chem.* **2008,** *43* (6), 1297-1303.

44. Bottegoni, G.; Favia, A. D.; Recanatini, M.; Cavalli, A., The Role of Fragment-Based and Computational Methods in Polypharmacology. *Drug Discovery Today* **2012,** *17* (1), 23-34.

45. Schlessinger, A.; Abagyan, R.; Carlson, H. A.; Dang, K. K.; Guinney, J.; Cagan, R. L., Multi-targeting Drug Community Challenge. *Cell Chem. Biol.* **2017,** *24* (12), 1434-1435.

46. Lamb, J., The Connectivity Map: A New Tool for Biomedical Research. *Nat. Rev. Cancer* **2007,** *7*, 54-60.

47. Cheng, F.; Kovács, I. A.; Barabási, A.-L., Network-Based Prediction of Drug Combinations. *Nat. Commun.* **2019,** *10* (1), 1197.

48. Cavalli, A.; Bolognesi, M. L.; Minarini, A.; Rosini, M.; Tumiatti, V.; Recanatini, M.; Melchiorre, C., Multi-target-Directed Ligands To Combat Neurodegenerative Diseases. *J. Med. Chem.* **2008,** *51* (3), 347-372.

49. Fang, J.; Wu, Z.; Cai, C.; Wang, Q.; Tang, Y.; Cheng, F., Quantitative and Systems Pharmacology. 1. In Silico Prediction of Drug–Target Interactions of Natural Products

Enables New Targeted Cancer Therapy. *J. Chem. Inf. Model.* **2017,** *57* (11), 2657-2671.

50. Cheng, F.; Jia, P.; Wang, Q.; Zhao, Z., Quantitative Network Mapping of the Human Kinome Interactome Reveals New Clues for Rational Kinase Inhibitor Discovery and Individualized Cancer Therapy. *Oncotarget* **2014,** *5* (11), 3697-3710.

51. Espinosa-Cantu, A.; Ascencio, D.; Barona-Gomez, F.; DeLuna, A., Gene duplication and the evolution of moonlighting proteins. *Front Genet* **2015,** *6*, 227.

52. Knauer, S. H.; Rosch, P.; Artsimovitch, I., Transformation: the next level of regulation. *RNA Biol* **2012,** *9* (12), 1418-23.

53. Tomar, S. K.; Artsimovitch, I., NusG-Spt5 proteins-Universal tools for transcription modification and communication. *Chem Rev* **2013,** *113* (11), 8604-19.

54. Kyrpides, N. C.; Woese, C. R.; Ouzounis, C. A., KOW: a novel motif linking a bacterial transcription factor with ribosomal proteins. *Trends Biochem Sci* **1996,** *21* (11), 425-6.

55. Steiner, T.; Kaiser, J. T.; Marinkovic, S.; Huber, R.; Wahl, M. C., Crystal structures of transcription factor NusG in light of its nucleic acid- and protein-binding activities. *Embo J* **2002,** *21* (17), 4641-53.

56. Hartzog, G. A.; Kaplan, C. D., Competing for the clamp: promoting RNA polymerase processivity and managing the transition from initiation to elongation. *Mol Cell* **2011,** *43* (2), 161-3.

57. Svetlov, V.; Nudler, E., Clamping the clamp of RNA polymerase. *Embo J* **2011,** *30* (7), 1190-1.

58. Yakhnin, A. V.; Babitzke, P., NusG/Spt5: are there common functions of this ubiquitous transcription elongation factor? *Curr Opin Microbiol* **2014,** *18*, 68-71.

59. Burmann, B. M.; Schweimer, K.; Luo, X.; Wahl, M. C.; Stitt, B. L.; Gottesman, M. E.; Rosch, P., A NusE:NusG complex links transcription and translation. *Science* **2010,** *328* (5977), 501-4.

60. Hartzog, G. A.; Fu, J., The Spt4-Spt5 complex: a multi-faceted regulator of transcription elongation. *Biochim Biophys Acta* **2013,** *1829* (1), 105-15.

61. Schmidt, A.; Kochanowski, K.; Vedelaar, S.; Ahrne, E.; Volkmer, B.; Callipo, L.; Knoops, K.; Bauer, M.; Aebersold, R.; Heinemann, M., The quantitative and condition-dependent Escherichia coli proteome. *Nat Biotechnol* **2016,** *34* (1), 104-10.

62. Mooney, R. A.; Davis, S. E.; Peters, J. M.; Rowland, J. L.; Ansari, A. Z.; Landick, R., Regulator trafficking on bacterial transcription units in vivo. *Mol Cell* **2009,** *33* (1), 97-108.

63. Mooney, R. A.; Schweimer, K.; Rosch, P.; Gottesman, M.; Landick, R., Two structurally independent domains of E. coli NusG create regulatory plasticity via distinct interactions with RNA polymerase and regulators. *J Mol Biol* **2009,** *391* (2), 341-58.

64. Cardinale, C. J.; Washburn, R. S.; Tadigotla, V. R.; Brown, L. M.; Gottesman, M. E.; Nudler, E., Termination factor Rho and its cofactors NusA and NusG silence foreign DNA in E. coli. *Science* **2008,** *320* (5878), 935-8.

65. Artsimovitch, I.; Landick, R., The transcriptional regulator RfaH stimulates RNA chain synthesis after recruitment to elongation complexes by the exposed nontemplate DNA strand. *Cell* **2002,** *109* (2), 193-203.

66. Belogurov, G. A.; Mooney, R. A.; Svetlov, V.; Landick, R.; Artsimovitch, I., Functional specialization of transcription elongation factors. *Embo J* **2009,** *28* (2), 112-22.

67. Sevostyanova, A.; Belogurov, G. A.; Mooney, R. A.; Landick, R.; Artsimovitch, I., The beta subunit gate loop is required for RNA polymerase modification by RfaH and NusG. *Mol Cell* **2011,** *43* (2), 253-62.

68. Burmann, B. M.; Knauer, S. H.; Sevostyanova, A.; Schweimer, K.; Mooney, R. A.; Landick, R.; Artsimovitch, I.; Rosch, P., An alpha helix to beta barrel domain switch transforms the transcription factor RfaH into a translation factor. *Cell* **2012,** *150* (2), 291-303.

69. Bachman, M. A.; Breen, P.; Deornellas, V.; Mu, Q.; Zhao, L.; Wu, W.; Cavalcoli, J. D.; Mobley, H. L., Genome-Wide Identification of Klebsiella pneumoniae Fitness Genes during Lung Infection. *MBio* **2015,** *6* (3), e00775.

70. Nagy, G.; Danino, V.; Dobrindt, U.; Pallen, M.; Chaudhuri, R.; Emody, L.; Hinton, J. C.; Hacker, J., Down-regulation of key virulence factors makes the Salmonella enterica serovar Typhimurium rfaH mutant a promising live-attenuated vaccine candidate. *Infect Immun* **2006,** *74* (10), 5914-25.

71. Nagy, G.; Dobrindt, U.; Schneider, G.; Khan, A. S.; Hacker, J.; Emody, L., Loss of regulatory protein RfaH attenuates virulence of uropathogenic Escherichia coli. *Infect Immun* **2002,** *70* (8), 4406-13.

72. Belogurov, G. A.; Sevostyanova, A.; Svetlov, V.; Artsimovitch, I., Functional regions of the N-terminal domain of the antiterminator RfaH. *Mol Microbiol* **2010,** *76* (2), 286-301.

73. Belogurov, G. A.; Vassylyeva, M. N.; Svetlov, V.; Klyuyev, S.; Grishin, N. V.; Vassylyev, D. G.; Artsimovitch, I., Structural basis for converting a general transcription factor into an operon-specific virulence regulator. *Mol Cell* **2007,** *26* (1), 117-29.

74. Burmann, B. M.; Scheckenhofer, U.; Schweimer, K.; Rosch, P., Domain interactions of the transcription-translation coupling factor Escherichia coli NusG are intermolecular and transient. *Biochem J* **2011,** *435* (3), 783-9.

75. Abagyan, R.; Raush, E.; Totrov, M.; Orry, A., ICM Manual v3.8-6; Molsoft LCC: San Diego, CA. **2017**.

76. Gonnet, G. H.; Cohen, M. A.; Benner, S. A., Exhaustive Matching of the Entire Protein Sequence Database. *Science* **1992,** *256* (5062), 1443-1445.

77. Abagyan, R. A.; Totrov, M. M., Contact area difference (CAD): a robust measure to evaluate accuracy of protein models. *J Mol Biol* **1997,** *268* (3), 678-85.

78. Finn, R. D.; Attwood, T. K.; Babbitt, P. C.; Bateman, A.; Bork, P.; Bridge, A. J.; Chang, H. Y.; Dosztanyi, Z.; El-Gebali, S.; Fraser, M.; Gough, J.; Haft, D.; Holliday, G. L.; Huang, H.; Huang, X.; Letunic, I.; Lopez, R.; Lu, S.; Marchler-Bauer, A.; Mi, H.; Mistry, J.; Natale, D. A.; Necci, M.; Nuka, G.; Orengo, C. A.; Park, Y.; Pesseat, S.; Piovesan, D.; Potter, S. C.; Rawlings, N. D.; Redaschi, N.; Richardson, L.; Rivoire, C.; Sangrador-Vegas, A.; Sigrist, C.; Sillitoe, I.; Smithers, B.; Squizzato, S.; Sutton, G.; Thanki, N.; Thomas, P. D.; Tosatto, S. C.; Wu, C. H.; Xenarios, I.; Yeh, L. S.; Young, S. Y.; Mitchell, A. L., InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res* **2017,** *45* (D1), D190-D199.

79. Abagyan, R.; Totrov, M.; Kuznetsov, D., ICM — a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. *Journal of computational chemistry* **1994,** *15* (5), 488-506.

80. Abagyan, R. A.; Batalov, S., Do aligned sequences share the same fold? *J Mol Biol* **1997,** *273* (1), 355-68.

81. Bordner, A. J.; Abagyan, R. A., Large-scale prediction of protein geometry and stability changes for arbitrary single point mutations. *Proteins* **2004,** *57* (2), 400-13.

82. Balasco, N.; Barone, D.; Vitagliano, L., Structural conversion of the transformer protein RfaH: new insights derived from protein structure prediction and molecular dynamics simulations. *J Biomol Struct Dyn* **2015,** *33* (10), 2173-9.

83. Gc, J. B.; Gerstman, B. S.; Chapagain, P. P., The Role of the Interdomain Interactions on RfaH Dynamics and Conformational Transformation. *J Phys Chem B* **2015,** *119* (40), 12750-9.

84. Li, S.; Xiong, B.; Xu, Y.; Lu, T.; Luo, X.; Luo, C.; Shen, J.; Chen, K.; Zheng, M.; Jiang, H., Mechanism of the All-alpha to All-beta Conformational Transition of RfaH-CTD: Molecular Dynamics Simulation and Markov State Model. *J Chem Theory Comput* **2014,** *10* (6), 2255-64.

85. Ramirez-Sarmiento, C. A.; Noel, J. K.; Valenzuela, S. L.; Artsimovitch, I., Interdomain Contacts Control Native State Switching of RfaH on a Dual-Funneled Landscape. *PLoS Comput Biol* **2015,** *11* (7), e1004379.

86. Xiong, L.; Liu, Z., Molecular dynamics study on folding and allostery in RfaH. *Proteins* **2015,** *83* (9), 1582-92.

87. Xun, S.; Jiang, F.; Wu, Y. D., Intrinsically disordered regions stabilize the helical form of the C-terminal domain of RfaH: A molecular dynamics study. *Bioorg Med Chem* **2016**.

88. Marsden, B.; Abagyan, R., SAD--a normalized structural alignment database: improving sequence-structure alignments. *Bioinformatics* **2004,** *20* (15), 2333-44.

89. Izban, M. G.; Samkurashvili, I.; Luse, D. S., RNA polymerase II ternary complexes may become arrested after transcribing to within 10 bases of the end of linear templates. *J Biol Chem* **1995,** *270* (5), 2290-7.

90. Appel, W., Chymotrypsin: molecular and catalytic properties. *Clin Biochem* **1986,** *19* (6), 317-22.

91. Lavinder, J. J.; Hari, S. B.; Sullivan, B. J.; Magliery, T. J., High-throughput thermal scanning: a general, rapid dye-binding thermal shift screen for protein engineering. *J Am Chem Soc* **2009,** *131* (11), 3794-5.

92. Pufall, M. A.; Graves, B. J., Autoinhibitory domains: modular effectors of cellular regulation. *Annu Rev Cell Dev Biol* **2002,** *18*, 421-62.

93. Dombroski, A. J.; Walter, W. A.; Record, M. T., Jr.; Siegele, D. A.; Gross, C. A., Polypeptides containing highly conserved regions of transcription initiation factor sigma 70 exhibit specificity of binding to promoter DNA. *Cell* **1992,** *70* (3), 501-12.

94. Samorodnitsky, D.; Szyjka, C.; Koudelka, G. B., A Role for Autoinhibition in Preventing Dimerization of the Transcription Factor ETS1. *J Biol Chem* **2015,** *290* (36), 22101-10.

95. Currie, S. L.; Lau, D. K.; Doane, J. J.; Whitby, F. G.; Okon, M.; McIntosh, L. P.; Graves, B. J., Structured and disordered regions cooperatively mediate DNA-binding autoinhibition of ETS factors ETV1, ETV4 and ETV5. *Nucleic Acids Res* **2017**.

96. Tomar, S. K.; Knauer, S. H.; Nandymazumdar, M.; Rosch, P.; Artsimovitch, I., Interdomain contacts control folding of transcription factor RfaH. *Nucleic Acids Res* **2013,** *41* (22), 10077-85.

97. Trudeau, T.; Nassar, R.; Cumberworth, A.; Wong, E. T.; Woollard, G.; Gsponer, J., Structure and intrinsic disorder in protein autoinhibition. *Structure* **2013,** *21* (3), 332-41.

98. Zouhir, S.; Bernal-Bayard, J.; Cordero-Alba, M.; Cardenal-Munoz, E.; Guimaraes, B.; Lazar, N.; Ramos-Morales, F.; Nessler, S., The structure of the Slrp-Trx1 complex sheds light on the autoinhibition mechanism of the type III secretion system effectors of the NEL family. *Biochem J* **2014,** *464* (1), 135-44.

99. Cai, Z.; Yuan, Z. H.; Zhang, H.; Pan, Y.; Wu, Y.; Tian, X. Q.; Wang, F. F.; Wang, L.; Qian, W., Fatty acid DSF binds and allosterically activates histidine kinase RpfC of

phytopathogenic bacterium Xanthomonas campestris pv. campestris to regulate quorum-sensing and virulence. *PLoS Pathog* **2017,** *13* (4), e1006304.

100. Takemoto-Kimura, S.; Suzuki, K.; Horigane, S. I.; Kamijo, S.; Inoue, M.; Sakamoto, M.; Fujii, H.; Bito, H., Calmodulin kinases: essential regulators in health and disease. *J Neurochem* **2017**.

101. Emptage, R. P.; Lemmon, M. A.; Ferguson, K. M., Molecular determinants of KA1 domain-mediated autoinhibition and phospholipid activation of MARK1 kinase. *Biochem J* **2017,** *474* (3), 385-398.

102. Bianchi, S.; van Riel, W. E.; Kraatz, S. H.; Olieric, N.; Frey, D.; Katrukha, E. A.; Jaussi, R.; Missimer, J.; Grigoriev, I.; Olieric, V.; Benoit, R. M.; Steinmetz, M. O.; Akhmanova, A.; Kammerer, R. A., Structural basis for misregulation of kinesin KIF21A autoinhibition by CFEOM1 disease mutations. *Sci Rep* **2016,** *6*, 30668.

103. Reubold, T. F.; Faelber, K.; Plattner, N.; Posor, Y.; Ketel, K.; Curth, U.; Schlegel, J.; Anand, R.; Manstein, D. J.; Noe, F.; Haucke, V.; Daumke, O.; Eschenburg, S., Crystal structure of the dynamin tetramer. *Nature* **2015,** *525* (7569), 404-8.

104. Myers, J. K.; Pace, C. N.; Scholtz, J. M., A direct comparison of helix propensity in proteins and peptides. *Proc Natl Acad Sci U S A* **1997,** *94* (7), 2833-7.

105. Drogemuller, J.; Schneider, C.; Schweimer, K.; Strauss, M.; Wohrl, B. M.; Rosch, P.; Knauer, S. H., Thermotoga maritima NusG: domain interaction mediates autoinhibition and thermostability. *Nucleic Acids Res* **2017,** *45* (1), 446-460.

106. Yuan, A. H.; Hochschild, A., A bacterial global regulator forms a prion. *Science* **2017,** *355* (6321), 198-201.

107. Chatzidaki-Livanis, M.; Coyne, M. J.; Comstock, L. E., A family of transcriptional antitermination factors necessary for synthesis of the capsular polysaccharides of Bacteroides fragilis. *J Bacteriol* **2009,** *191* (23), 7288-95.

108. Bailey, M. J.; Koronakis, V.; Schmoll, T.; Hughes, C., Escherichia coli HlyT protein, a transcriptional activator of haemolysin synthesis and secretion, is encoded by the rfaH (sfrB) locus required for expression of sex factor and lipopolysaccharide genes. *Mol Microbiol* **1992,** *6* (8), 1003-12.

109. Hurst, M. R.; Beard, S. S.; Jackson, T. A.; Jones, S. M., Isolation and characterization of the Serratia entomophila antifeeding prophage. *FEMS Microbiol Lett* **2007,** *270* (1), 42-8.

110. Paitan, Y.; Orr, E.; Ron, E. Z.; Rosenberg, E., A NusG-like transcription anti-terminator is involved in the biosynthesis of the polyketide antibiotic TA of Myxococcus xanthus. *FEMS Microbiol Lett* **1999,** *170* (1), 221-7.

111. Goodson, J. R.; Klupt, S.; Zhang, C.; Straight, P.; Winkler, W. C., LoaP is a broadly

conserved antiterminator protein that regulates antibiotic gene clusters in Bacillus amyloliquefaciens. *Nat Microbiol* **2017,** *2*, 17003.

112. Viaggi, B.; Sbrana, F.; Malacarne, P.; Tascini, C., Ventilator-associated pneumonia caused by colistin-resistant KPC-producing Klebsiella pneumoniae: a case report and literature review. *Respir Investig* **2015,** *53* (3), 124-8.

113. Chen, L.; Mathema, B.; Chavda, K. D.; DeLeo, F. R.; Bonomo, R. A.; Kreiswirth, B. N., Carbapenemase-producing Klebsiella pneumoniae: molecular and genetic decoding. *Trends Microbiol* **2014,** *22* (12), 686-96.

114. Rahn, A.; Drummelsmith, J.; Whitfield, C., Conserved organization in the cps gene clusters for expression of Escherichia coli group 1 K antigens: relationship to the colanic acid biosynthesis locus and the cps genes from Klebsiella pneumoniae. *J Bacteriol* **1999,** *181* (7), 2307-13.

115. Garrett, S. B.; Garrison-Schilling, K. L.; Cooke, J. T.; Pettis, G. S., Capsular polysaccharide production and serum survival of Vibrio vulnificus are dependent on antitermination control by RfaH. *FEBS Lett* **2016,** *590* (24), 4564-4572.

116. NandyMazumdar, M.; Artsimovitch, I., Ubiquitous transcription factors display structural plasticity and diverse functions: NusG proteins - Shifting shapes and paradigms. *Bioessays* **2015,** *37* (3), 324-34.

117. Liu, G.; Olsen, J. E.; Thomsen, L. E., Identification of Genes Essential for Antibiotic-Induced Up-Regulation of Plasmid-Transfer-Genes in Cephalosporin Resistant Escherichia coli. *Frontiers in Microbiology* **2019,** *10*, 2203.

118. Peters, J. M.; Mooney, R. A.; Grass, J. A.; Jessen, E. D.; Tran, F.; Landick, R., Rho and NusG suppress pervasive antisense transcription in Escherichia coli. *Genes Dev* **2012,** *26* (23), 2621-33.

119. Kang, J. Y.; Mooney, R. A.; Nedialkov, Y.; Saba, J.; Mishanina, T. V.; Artsimovitch, I.; Landick, R.; Darst, S. A., Structural Basis for Transcript Elongation Control by NusG Family Universal Regulators. *Cell* **2018**.

120. Gualerzi, C. O.; Pon, C. L., Initiation of mRNA translation in bacteria: structural and dynamic aspects. *Cell Mol Life Sci* **2015,** *72* (22), 4341-67.

121. Zuber, P. K.; Artsimovitch, I.; NandyMazumdar, M.; Liu, Z.; Nedialkov, Y.; Schweimer, K.; Rosch, P.; Knauer, S. H., The universally-conserved transcription factor RfaH is recruited to a hairpin structure of the non-template DNA strand. *Elife* **2018,** *7*.

122. Carter, H. D.; Svetlov, V.; Artsimovitch, I., Highly divergent RfaH orthologs from pathogenic proteobacteria can substitute for Escherichia coli RfaH both in vivo and in vitro. *J Bacteriol* **2004,** *186* (9), 2829-40.

123. Czyz, A.; Mooney, R. A.; Iaconi, A.; Landick, R., Mycobacterial RNA polymerase

requires a U-tract at intrinsic terminators and is aided by NusG at suboptimal terminators. *MBio* **2014,** *5* (2), e00931.

124. Sevostyanova, A.; Artsimovitch, I., Functional analysis of Thermus thermophilus transcription factor NusG. *Nucleic Acids Res* **2010,** *38* (21), 7432-45.

125. Yakhnin, A. V.; Murakami, K. S.; Babitzke, P., NusG Is a Sequence-specific RNA Polymerase Pause Factor That Binds to the Non-template DNA within the Paused Transcription Bubble. *J Biol Chem* **2016,** *291* (10), 5299-308.

126. Abagyan, R.; Kufareva, I., The flexible pocketome engine for structural chemogenomics. *Methods Mol Biol* **2009,** *575*, 249-79.

127. An, J.; Totrov, M.; Abagyan, R., Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol Cell Proteomics* **2005,** *4* (6), 752-61.

128. Abagyan, R.; Raush, E.; Totrov, M.; Orry, A., *ICM Manual v3.8-6*. Molsoft LCC: San Diego, CA, 2017.

129. Shi, D.; Svetlov, D.; Abagyan, R.; Artsimovitch, I., Flipping states: a few key residues decide the winning conformation of the only universally conserved transcription factor. *Nucleic Acids Res* **2017,** *45* (15), 8835-8843.

130. Bottegoni, G.; Kufareva, I.; Totrov, M.; Abagyan, R., A new method for ligand docking to flexible receptors by dual alanine scanning and refinement (SCARE). *J Comput Aided Mol Des* **2008,** *22* (5), 311-25.

131. Neves, M. A.; Totrov, M.; Abagyan, R., Docking and scoring with ICM: the benchmarking results and strategies for improvement. *J Comput Aided Mol Des* **2012,** *26* (6), 675-86.

132. Barratt, M. D.; Basketter, D. A.; Chamberlain, M.; Admans, G. D.; Langowski, J. J., An expert system rulebase for identifying contact allergens. *Toxicol In Vitro* **1994,** *8* (5), 1053-60.

133. Gerner, I.; Barratt, M. D.; Zinke, S.; Schlegel, K.; Schlede, E., Development and prevalidation of a list of structure-activity relationship rules to be used in expert systems for prediction of the skin-sensitising properties of chemicals. *Altern Lab Anim* **2004,** *32* (5), 487-509.

134. Svetlov, V.; Artsimovitch, I., Purification of bacterial RNA polymerase: tools and protocols. *Methods Mol Biol* **2015,** *1276*, 13-29.

135. Vassylyeva, M. N.; Svetlov, V.; Dearborn, A. D.; Klyuyev, S.; Artsimovitch, I.; Vassylyev, D. G., The carboxy-terminal coiled-coil of the RNA polymerase beta'-subunit is the main binding site for Gre factors. *EMBO Rep* **2007,** *8* (11), 1038-43.

136. Nedialkov, Y.; Svetlov, D.; Belogurov, G. A.; Artsimovitch, I., Locking the non-

template DNA to control transcription. *Mol Microbiol* **2018**.

137. Blumenkrantz, N.; Asboe-Hansen, G., New method for quantitative determination of uronic acids. *Anal Biochem* **1973,** *54* (2), 484-9.

138. Lin, T. L.; Yang, F. L.; Yang, A. S.; Peng, H. P.; Li, T. L.; Tsai, M. D.; Wu, S. H.; Wang, J. T., Amino acid substitutions of MagA in Klebsiella pneumoniae affect the biosynthesis of the capsular polysaccharide. *PLoS One* **2012,** *7* (10), e46783.

139. Rosen, D. A.; Twentyman, J.; Hunstad, D. A., High Levels of Cyclic Di-GMP in Klebsiella pneumoniae Attenuate Virulence in the Lung. *Infect Immun* **2018,** *86* (2).

140. Hu, K.; Artsimovitch, I., A Screen for rfaH Suppressors Reveals a Key Role for a Connector Region of Termination Factor Rho. *MBio* **2017,** *8* (3).

141. Koronakis, V.; Cross, M.; Hughes, C., Expression of the E.coli hemolysin secretion gene hlyB involves transcript anti-termination within the hly operon. *Nucleic Acids Res* **1988,** *16* (11), 4789-800.

142. Moller, A. K.; Leatham, M. P.; Conway, T.; Nuijten, P. J.; de Haan, L. A.; Krogfelt, K. A.; Cohen, P. S., An Escherichia coli MG1655 lipopolysaccharide deep-rough core mutant grows and survives in mouse cecal mucus but fails to colonize the mouse large intestine. *Infect Immun* **2003,** *71* (4), 2142-52.

143. Navasa, N.; Rodriguez-Aparicio, L. B.; Ferrero, M. A.; Monteagudo-Mera, A.; Martinez-Blanco, H., Transcriptional control of RfaH on polysialic and colanic acid synthesis by Escherichia coli K92. *FEBS Lett* **2014,** *588* (6), 922-8.

144. Stevens, M. P.; Clarke, B. R.; Roberts, I. S., Regulation of the Escherichia coli K5 capsule gene cluster by transcription antitermination. *Mol Microbiol* **1997,** *24* (5), 1001-12.

145. Bailey, M. J.; Hughes, C.; Koronakis, V., RfaH and the ops element, components of a novel system controlling bacterial transcription elongation. *Mol Microbiol* **1997,** *26* (5), 845-51.

146. NandyMazumdar, M.; Nedialkov, Y.; Svetlov, D.; Sevostyanova, A.; Belogurov, G. A.; Artsimovitch, I., RNA polymerase gate loop guides the nontemplate DNA strand in transcription complexes. *Proc Natl Acad Sci U S A* **2016,** *113* (52), 14994-14999.

147. Turtola, M.; Belogurov, G. A., NusG inhibits RNA polymerase backtracking by stabilizing the minimal transcription bubble. *Elife* **2016,** *e18096.*

148. Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G., ZINC: a free tool to discover chemistry for biology. *J Chem Inf Model* **2012,** *52* (7), 1757-68.

149. Richter, M. F.; Drown, B. S.; Riley, A. P.; Garcia, A.; Shirai, T.; Svec, R. L.; Hergenrother, P. J., Predictive compound accumulation rules yield a broad-spectrum

antibiotic. *Nature* **2017,** *545* (7654), 299-304.

150. Yoder, J. S.; Eddy, B. A.; Visvesvara, G. S.; Capewell, L.; Beach, M. J., The epidemiology of primary amoebic meningoencephalitis in the USA, 1962–2008. *Epidemiology and Infection* **2009,** *138* (7), 968-975.

151. De Jonckheere, J. F., Origin and evolution of the worldwide distributed pathogenic amoeboflagellate Naegleria fowleri. *Infection, Genetics and Evolution* **2011,** *11* (7), 1520-1528.

152. Siddiqui, R.; Ali, I. K. M.; Cope, J. R.; Khan, N. A., Biology and pathogenesis of Naegleria fowleri. *Acta Tropica* **2016,** *164*, 375-394.

153. Javanmard, E.; Niyyati, M.; Lorenzo-Morales, J.; Lasjerdi, Z.; Behniafar, H.; Mirjalali, H., Molecular identification of waterborne free living amoebae (Acanthamoeba, Naegleria and Vermamoeba) isolated from municipal drinking water and environmental sources, Semnan province, north half of Iran. *Experimental Parasitology* **2017,** *183*, 240-244.

154. Reyes-Batlle, M.; Wagner, C.; López-Arencibia, A.; Sifaoui, I.; Martínez-Carretero, E.; Valladares, B.; Piñero Jose, E.; Lorenzo-Morales, J., Isolation and molecular characterization of a Naegleria strain from a recreational water fountain in Tenerife, Canary Islands, Spain. In *Acta Parasitologica*, 2017; Vol. 62, p 265.

155. John, D. T., Primary Amebic Meningoencephalitis and the Biology of Naegleria fowleri. *Annu. Rev. Microbiol.* **1982,** *36* (1), 101-123.

156. Cervantes-Sandoval, I.; Serrano-Luna, J. d. J.; Meza-Cervantez, P.; Arroyo, R.; Tsutsumi, V.; Shibayama, M., Naegleria fowleri induces MUC5AC and pro-inflammatory cytokines in human epithelial cells via ROS production and EGFR activation. *Microbiology* **2009,** *155* (11), 3739-3747.

157. Cervantes-Sandoval, I.; Serrano-Luna, J. d. J.; García-Latorre, E.; Tsutsumi, V.; Shibayama, M., Characterization of brain inflammation during primary amoebic meningoencephalitis. *Parasitology International* **2008,** *57* (3), 307-313.

158. Linam, W. M.; Ahmed, M.; Cope, J. R.; Chu, C.; Visvesvara, G. S.; da Silva, A. J.; Qvarnstrom, Y.; Green, J., Successful Treatment of an Adolescent With Naegleria fowleri Primary Amebic Meningoencephalitis. *Pediatrics* **2015,** *135* (3), e744.

159. Cope, J. R.; Conrad, D. A.; Cohen, N.; Cotilla, M.; DaSilva, A.; Jackson, J.; Visvesvara, G. S., Use of the Novel Therapeutic Agent Miltefosine for the Treatment of Primary Amebic Meningoencephalitis: Report of 1 Fatal and 1 Surviving Case. *Clinical Infectious Diseases* **2015,** *62* (6), 774-776.

160. Seidel, J. S.; Harmatz, P.; Visvesvara, G. S.; Cohen, A.; Edwards, J.; Turner, J., Successful Treatment of Primary Amebic Meningoencephalitis. *N. Engl. J. Med.* **1982,** *306* (6), 346-348.

161. Mesa-Arango, A. C.; Scorzoni, L.; Zaragoza, O., It only takes one to do many jobs: Amphotericin B as antifungal and immunomodulatory drug. *Frontiers in Microbiology* **2012,** *3* (286).

162. Baginski, M.; Czub, J., Amphotericin B and Its New Derivatives – Mode of Action. *Current Drug Metabolism* **2009,** *10* (5), 459-469.

163. Ghannoum, M. A.; Rice, L. B., Antifungal agents: mode of action, mechanisms of resistance, and correlation of these mechanisms with bacterial resistance. *Clinical microbiology reviews* **1999,** *12* (4), 501-517.

164. Debnath, A.; Calvet, C. M.; Jennings, G.; Zhou, W.; Aksenov, A.; Luth, M. R.; Abagyan, R.; Nes, W. D.; McKerrow, J. H.; Podust, L. M., CYP51 is an essential drug target for the treatment of primary amoebic meningoencephalitis (PAM). *PLoS neglected tropical diseases* **2017,** *11* (12), e0006104.

165. Schuster, F. L.; Guglielmo, B. J.; Visvesvara, G. S., In-Vitro Activity of Miltefosine and Voriconazole on Clinical Isolates of Free-Living Amebas: Balamuthia mandrillaris, Acanthamoeba spp., and Naegleria fowleri. *Journal of Eukaryotic Microbiology* **2006,** *53* (2), 121-126.

166. Zhou, W.; Debnath, A.; Jennings, G.; Hahn, H. J.; Vanderloop, B. H.; Chaudhuri, M.; Nes, W. D.; Podust, L. M., Enzymatic chokepoints and synergistic drug targets in the sterol biosynthesis pathway of Naegleria fowleri. *PLOS Pathogens* **2018,** *14* (9), e1007245.

167. Kethireddy, S.; Andes, D., CNS pharmacokinetics of antifungal agents. *Expert Opinion on Drug Metabolism & Toxicology* **2007,** *3* (4), 573-581.

168. Raederstorff, D.; Rohmer, M., The Action of the Systemic Fungicides Tridemorph and Fenpropimorph on Sterol Biosynthesis by the Soil Amoeba Acanthamoeba Polyphaga. *Eur. J. Biochem.* **1987,** *164* (2), 421-426.

169. Nes, Craigen R.; Singha, Ujjal K.; Liu, J.; Ganapathy, K.; Villalta, F.; Waterman, Michael R.; Lepesheva, Galina I.; Chaudhuri, M.; Nes, W. D., Novel sterol metabolic network of &lt;em&gt;Trypanosoma brucei&lt;/em&gt; procyclic and bloodstream forms. *Biochemical Journal* **2012,** *443* (1), 267.

170. Zhou, W.; Warrilow, A. G. S.; Thomas, C. D.; Ramos, E.; Parker, J. E.; Price, C. L.; Vanderloop, B. H.; Fisher, P. M.; Loftis, M. D.; Kelly, D. E.; Kelly, S. L.; Nes, W. D., Functional importance for developmental regulation of sterol biosynthesis in Acanthamoeba castellanii. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* **2018,** *1863* (10), 1164-1178.

171. Choi, J. Y.; Podust, L. M.; Roush, W. R., Drug Strategies Targeting CYP51 in Neglected Tropical Diseases. *Chemical Reviews* **2014,** *114* (22), 11242-11271.

172. Schmidt, H. R.; Zheng, S.; Gurpinar, E.; Koehl, A.; Manglik, A.; Kruse, A. C., Crystal structure of the human σ1 receptor. *Nature* **2016,** *532*, 527.

173. An, J.; Totrov, M.; Abagyan, R., Pocketome via Comprehensive Identification and Classification of Ligand Binding Envelopes. *Molecular & Cellular Proteomics* **2005,** *4* (6), 752-761.

174. Shi, D.; Abagyan, R.; Svetlov, D.; Artsimovitch, I., Flipping states: a few key residues decide the winning conformation of the only universally conserved transcription factor. *Nucleic Acids Research* **2017,** *45* (15), 8835-8843.

175. Abagyan, R. A.; Batalov, S., Do aligned sequences share the same fold?11Edited by F. E. Cohen. *Journal of Molecular Biology* **1997,** *273* (1), 355-368.

176. Abagyan, R.; Totrov, M., Biased Probability Monte Carlo Conformational Searches and Electrostatic Calculations for Peptides and Proteins. *J. Mol. Biol.* **1994,** *235* (3), 983-1002.

177. Abagyan, R.; Totrov, M.; Kuznetsov, D., Icm—A New Method for Protein Modeling and Design: Applications to Docking and Structure Prediction From the Distorted Native Conformation. *J. Comput. Chem.* **1994,** *15* (5), 488-506.

178. Svetlov, D.; Shi, D.; Twentyman, J.; Nedialkov, Y.; Rosen, D. A.; Abagyan, R.; Artsimovitch, I., In silico discovery of small molecules that inhibit RfaH recruitment to RNA polymerase. *Molecular Microbiology* **2018,** *110* (1), 128-142.

179. Wager, T. T.; Hou, X.; Verhoest, P. R.; Villalobos, A., Moving beyond Rules: The Development of a Central Nervous System Multiparameter Optimization (CNS MPO) Approach To Enable Alignment of Druglike Properties. *ACS Chemical Neuroscience* **2010,** *1* (6), 435-449.

180. Wager, T. T.; Hou, X.; Verhoest, P. R.; Villalobos, A., Central Nervous System Multiparameter Optimization Desirability: Application in Drug Discovery. *ACS Chemical Neuroscience* **2016,** *7* (6), 767-775.

181. Dinauer, M.; Pierre, R., Primary Amoebic Meningoencephalitis After Swimming in Stream Water. *The Lancet* **1973,** *302* (7835), 971.

182. Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. A.; Brusniak, M.-Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P., A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology* **2012,** *30*, 918-920.

183. Röst, H. L.; Sachsenberg, T.; Aiche, S.; Bielow, C.; Weisser, H.; Aicheler, F.; Andreotti, S.; Ehrlich, H.-C.; Gutenbrunner, P.; Kenar, E.; Liang, X.; Nahnsen, S.; Nilse, L.; Pfeuffer, J.; Rosenberger, G.; Rurik, M.; Schmitt, U.; Veit, J.; Walzer, M.; Wojnar, D.; Wolski, W. E.; Schilling, O.; Choudhary, J. S.; Malmström, L.; Aebersold, R.; Reinert, K.;

Kohlbacher, O., OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nature Methods* **2016,** *13*, 741-748.

184. Wang, M.; Carver, J. J.; Phelan, V. V.; Sanchez, L. M.; Garg, N.; Peng, Y.; Nguyen, D. D.; Watrous, J.; Kapono, C. A.; Luzzatto-Knaan, T.; Porto, C.; Bouslimani, A.; Melnik, A. V.; Meehan, M. J.; Liu, W.-T.; Crüsemann, M.; Boudreau, P. D.; Esquenazi, E.; Sandoval-Calderón, M.; Kersten, R. D.; Pace, L. A.; Quinn, R. A.; Duncan, K. R.; Hsu, C.-C.; Floros, D. J.; Gavilan, R. G.; Kleigrewe, K.; Northen, T.; Dutton, R. J.; Parrot, D.; Carlson, E. E.; Aigle, B.; Michelsen, C. F.; Jelsbak, L.; Sohlenkamp, C.; Pevzner, P.; Edlund, A.; McLean, J.; Piel, J.; Murphy, B. T.; Gerwick, L.; Liaw, C.-C.; Yang, Y.-L.; Humpf, H.-U.; Maansson, M.; Keyzers, R. A.; Sims, A. C.; Johnson, A. R.; Sidebottom, A. M.; Sedio, B. E.; Klitgaard, A.; Larson, C. B.; Boya P, C. A.; Torres-Mendoza, D.; Gonzalez, D. J.; Silva, D. B.; Marques, L. M.; Demarque, D. P.; Pociute, E.; O'Neill, E. C.; Briand, E.; Helfrich, E. J. N.; Granatosky, E. A.; Glukhov, E.; Ryffel, F.; Houson, H.; Mohimani, H.; Kharbush, J. J.; Zeng, Y.; Vorholt, J. A.; Kurita, K. L.; Charusanti, P.; McPhail, K. L.; Nielsen, K. F.; Vuong, L.; Elfeki, M.; Traxler, M. F.; Engene, N.; Koyama, N.; Vining, O. B.; Baric, R.; Silva, R. R.; Mascuch, S. J.; Tomasi, S.; Jenkins, S.; Macherla, V.; Hoffman, T.; Agarwal, V.; Williams, P. G.; Dai, J.; Neupane, R.; Gurr, J.; Rodríguez, A. M. C.; Lamsa, A.; Zhang, C.; Dorrestein, K.; Duggan, B. M.; Almaliti, J.; Allard, P.-M.; Phapale, P.; Nothias, L.-F.; Alexandrov, T.; Litaudon, M.; Wolfender, J.-L.; Kyle, J. E.; Metz, T. O.; Peryea, T.; Nguyen, D.-T.; VanLeer, D.; Shinn, P.; Jadhav, A.; Müller, R.; Waters, K. M.; Shi, W.; Liu, X.; Zhang, L.; Knight, R.; Jensen, P. R.; Palsson, B. Ø.; Pogliano, K.; Linington, R. G.; Gutiérrez, M.; Lopes, N. P.; Gerwick, W. H.; Moore, B. S.; Dorrestein, P. C.; Bandeira, N., Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology* **2016,** *34*, 828-837.

185. Maurice, T.; Su, T.-P., The pharmacology of sigma-1 receptors. *Pharmacology & therapeutics* **2009,** *124* (2), 195-206.

186. Hanner, M.; Moebius, F. F.; Flandorfer, A.; Knaus, H. G.; Striessnig, J.; Kempner, E.; Glossmann, H., Purification, molecular cloning, and expression of the mammalian sigma1-binding site. *Proceedings of the National Academy of Sciences* **1996,** *93* (15), 8072.

187. Moebius, F. F.; Reiter, R. J.; Hanner, M.; Glossmann, H., High affinity of sigma1-binding sites for sterol isomerization inhibitors: evidence for a pharmacological relationship with the yeast sterol C8–C7 isomerase. *British journal of pharmacology* **1997,** *121* (1), 1-6.

188. Moebius, F. F.; Bermoser, K.; Reiter, R. J.; Hanner, M.; Glossmann, H., Yeast Sterol C8−C7 Isomerase: Identification and Characterization of a High-Affinity Binding Site for Enzyme Inhibitors. *Biochemistry* **1996,** *35* (51), 16871-16878.

189. Laskowski, R. A.; MacArthur, M. W.; Moss, D. S.; Thornton, J. M., Procheck: A Program to Check the Stereochemical Quality of Protein Structures. *J. Appl. Crystallogr.* **1993,** *26* (2), 283-291.

190. Long, T.; Hassan, A.; Thompson, B. M.; McDonald, J. G.; Wang, J.; Li, X., Structural basis for human sterol isomerase in cholesterol biosynthesis and multidrug recognition. *Nature Communications* **2019,** *10* (1), 2452.

191. Seth, P.; Ganapathy, M. E.; Conway, S. J.; Bridges, C. D.; Smith, S. B.; Casellas, P.; Ganapathy, V., Expression pattern of the type 1 sigma receptor in the brain and identity of critical anionic amino acid residues in the ligand-binding domain of the receptor. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **2001,** *1540* (1), 59-67.

192. Debnath, A.; Tunac, J. B.; Galindo-Gómez, S.; Silva-Olivares, A.; Shibayama, M.; McKerrow, J. H., Corifungin, a New Drug Lead against Naegleria, Identified from a High-Throughput Screen. *Antimicrobial Agents and Chemotherapy* **2012,** *56* (11), 5450-5457.

193. Debnath, A.; Nelson, A. T.; Silva-Olivares, A.; Shibayama, M.; Siegel, D.; McKerrow, J. H., In Vitro Efficacy of Ebselen and BAY 11-7082 Against Naegleria fowleri. *Frontiers in Microbiology* **2018,** *9*, 414.

194. Zyserman, I.; Mondal, D.; Sarabia, F.; McKerrow, J. H.; Roush, W. R.; Debnath, A., Identification of cysteine protease inhibitors as new drug leads against Naegleria fowleri. *Experimental Parasitology* **2018,** *188*, 36-41.

195. Singh, A.; Nisha; Bains, T.; Hahn, H. J.; Liu, N.; Tam, C.; Cheng, L. W.; Kim, J.; Debnath, A.; Land, K. M.; Kumar, V., Design, synthesis and preliminary antimicrobial evaluation of N-alkyl chain-tethered C-5 functionalized bis-isatins. *MedChemComm* **2017,** *8* (10), 1982-1992.

196. Quispe M, A.; Zavala C, D.; Rojas C, J.; Posso R, M.; Vaisberg W, A., Efecto citotóxico selectivo in vitro de muricin H(acetogenina de Annona muricata) en cultivos celulares de cáncer de pulmón. *Revista Peruana de Medicina Experimental y Salud Pública* **2006,** *23* (4), 265-269.

197. Zhong, H.-J.; Wang, W.; Kang, T.-S.; Yan, H.; Yang, Y.; Xu, L.; Wang, Y.; Ma, D.-L.; Leung, C.-H., A Rhodium(III) Complex as an Inhibitor of Neural Precursor Cell Expressed, Developmentally Down-Regulated 8-Activating Enzyme with in Vivo Activity against Inflammatory Bowel Disease. *Journal of Medicinal Chemistry* **2017,** *60* (1), 497-503.

198. Yang, C.; Wang, W.; Chen, L.; Liang, J.; Lin, S.; Lee, M.-Y.; Ma, D.-L.; Leung, C.-H., Discovery of a VHL and HIF1α interaction inhibitor with in vivo angiogenic activity via structure-based virtual screening. *Chemical Communications* **2016,** *52* (87), 12837-12840.

199. Yates, C. M.; Garvey, E. P.; Shaver, S. R.; Schotzinger, R. J.; Hoekstra, W. J., Design and optimization of highly-selective, broad spectrum fungal CYP51 inhibitors. *Bioorganic & Medicinal Chemistry Letters* **2017,** *27* (15), 3243-3248.