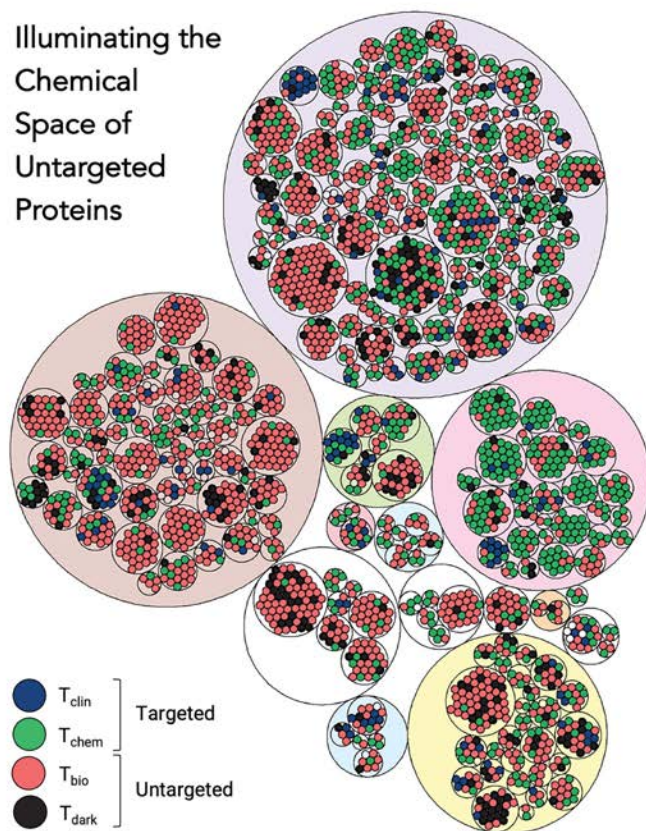


### Illuminating the Chemical Space of Untargeted Proteins



## Illuminating the Chemical Space of Untargeted Proteins

Maria J. Falaguera and Jordi Mestres\*

Cite This: *J. Chem. Inf. Model.* 2023, 63, 2689–2698

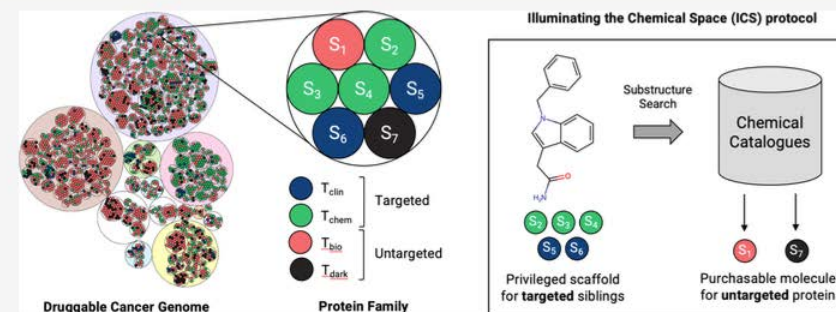
Read Online

ACCESS |

Metrics & More

Article Recommendations

Supporting Information







**ABSTRACT:** According to the Illuminating the Druggable Genome (IDG) initiative, 90% of the proteins encoded by the human genome still lack an identified active ligand, that is, a small molecule with biologically relevant binding potency or functional activity in an *in vitro* assay. Under this scenario, there is an urgent need for new approaches to chemically address these yet untargeted proteins. It is widely recognized that the best starting point for generating novel small molecules for proteins is to exploit the expected polypharmacology of known active ligands across phylogenetically related proteins following the paradigm that similar proteins are likely to interact with similar ligands. Here, we introduce a computational strategy to identify privileged structures that, when chemically expanded, are highly probable to contain active small molecules for untargeted proteins. The protocol was first tested on a set of 576 currently targeted proteins having at least one protein family sibling the year before their first active ligand was reported. A privileged structure contained in active ligands that were identified in the following years was correctly anticipated for 214 (37%) of those targeted proteins, a lower-bound recall estimate when considering data completeness issues. When applied to a set of 1184 untargeted potential druggable genes in cancer, the identification of privileged structures from known bioactive ligands of protein family siblings allowed for extracting a priority list of diverse commercially available small molecules for 960 of them. Assuming a minimum success rate of 37%, the chemical library selections should be able to deliver active ligands for at least 355 currently untargeted proteins associated with cancer.

# Shedding light on the pharmacological and biological knowledge of **understudied proteins**



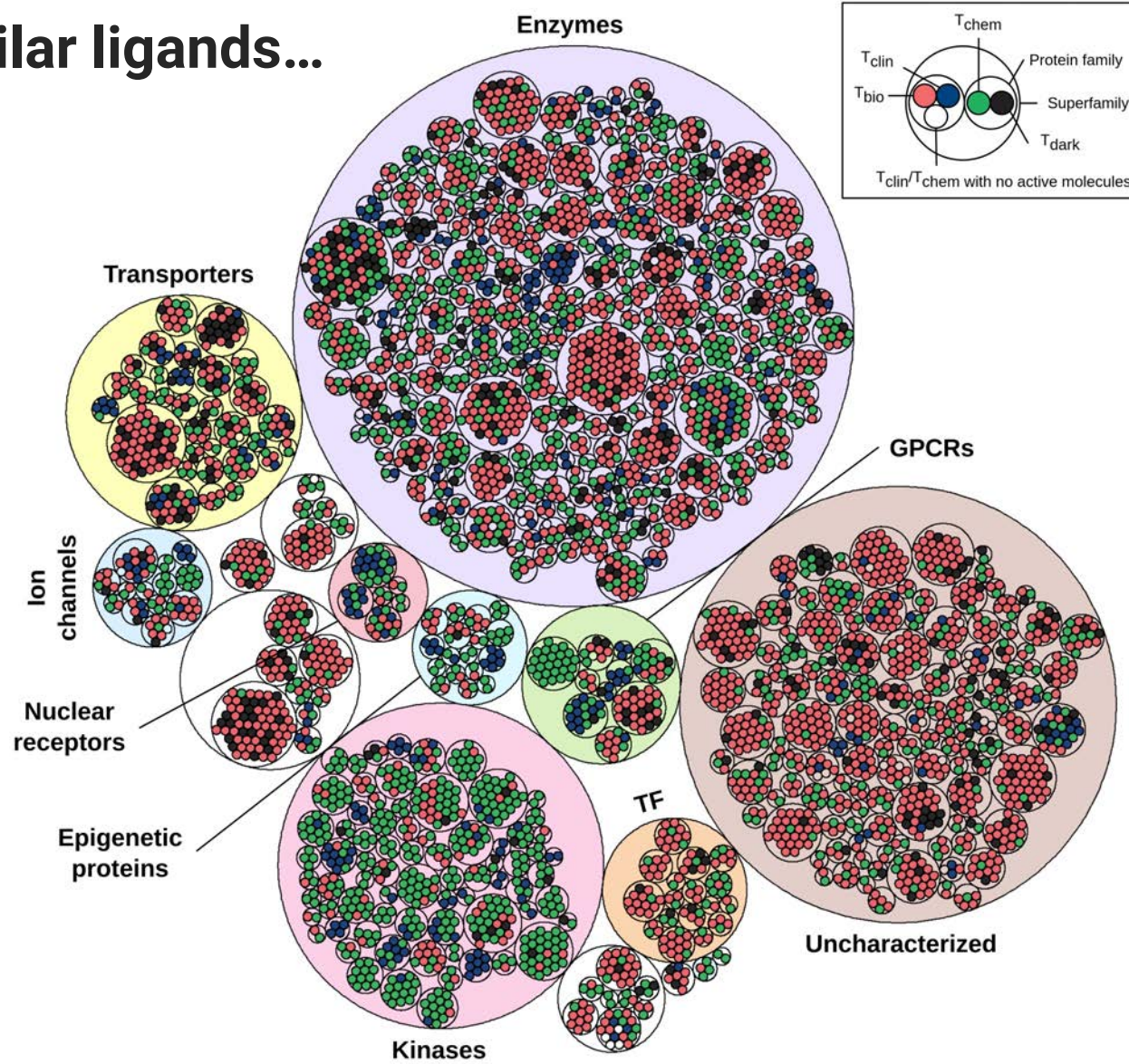
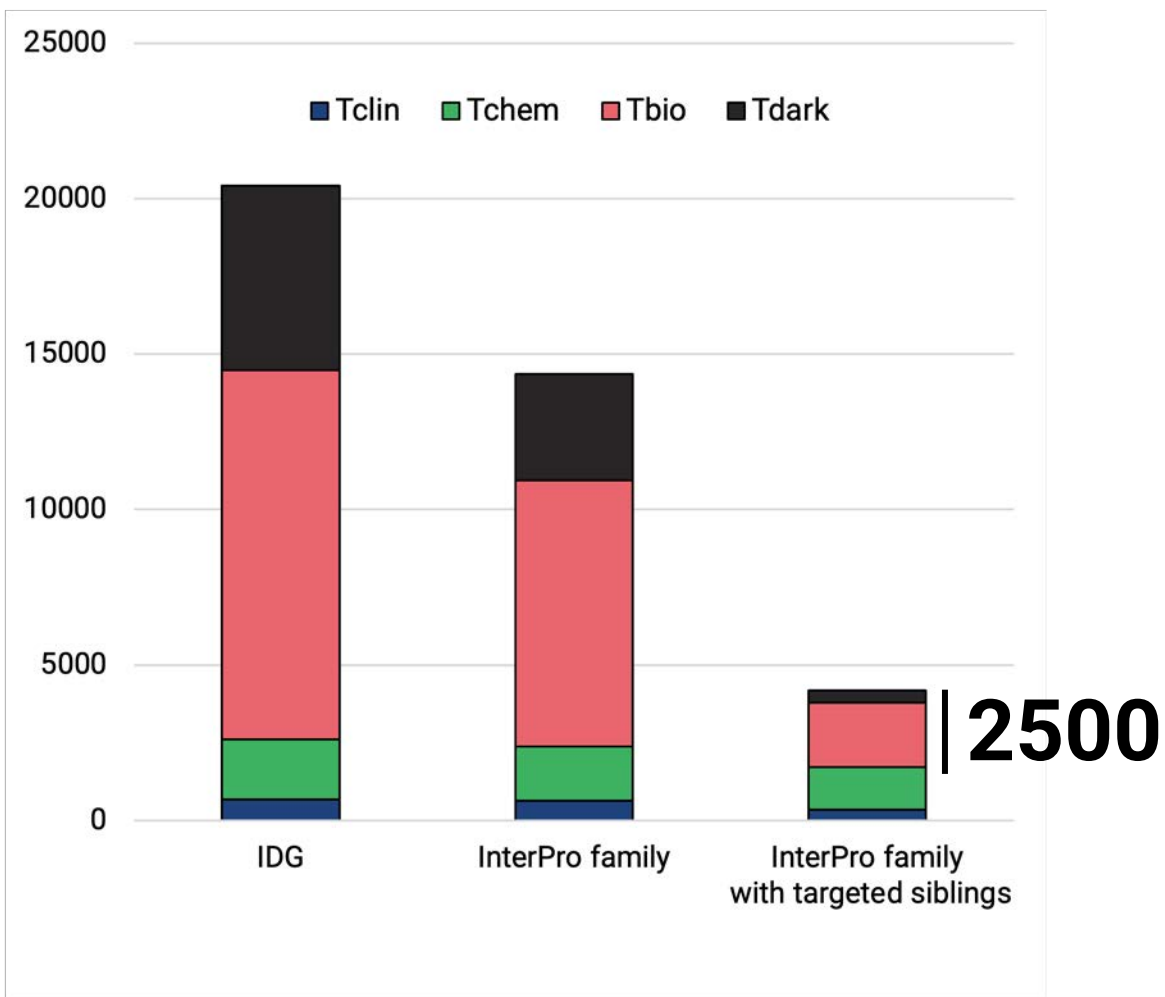
**IDG**  
ILLUMINATING the  
DRUGGABLE GENOME

## Target Development Level (TDL):

|   |                         |  |                       |
|---|-------------------------|--|-----------------------|
|  | <b>T<sub>clin</sub></b> | Mechanism of action of an approved drug    | } Targeted proteins   |
|  | <b>T<sub>chem</sub></b> | Non-drugged target with bioactive compound |                       |
|  | <b>T<sub>bio</sub></b>  | Relevant in physiological processes        | } Untargeted proteins |
|  | <b>T<sub>dark</sub></b> | Only its primary sequence is known         |                       |



# Similar proteins are likely to bind to similar ligands...

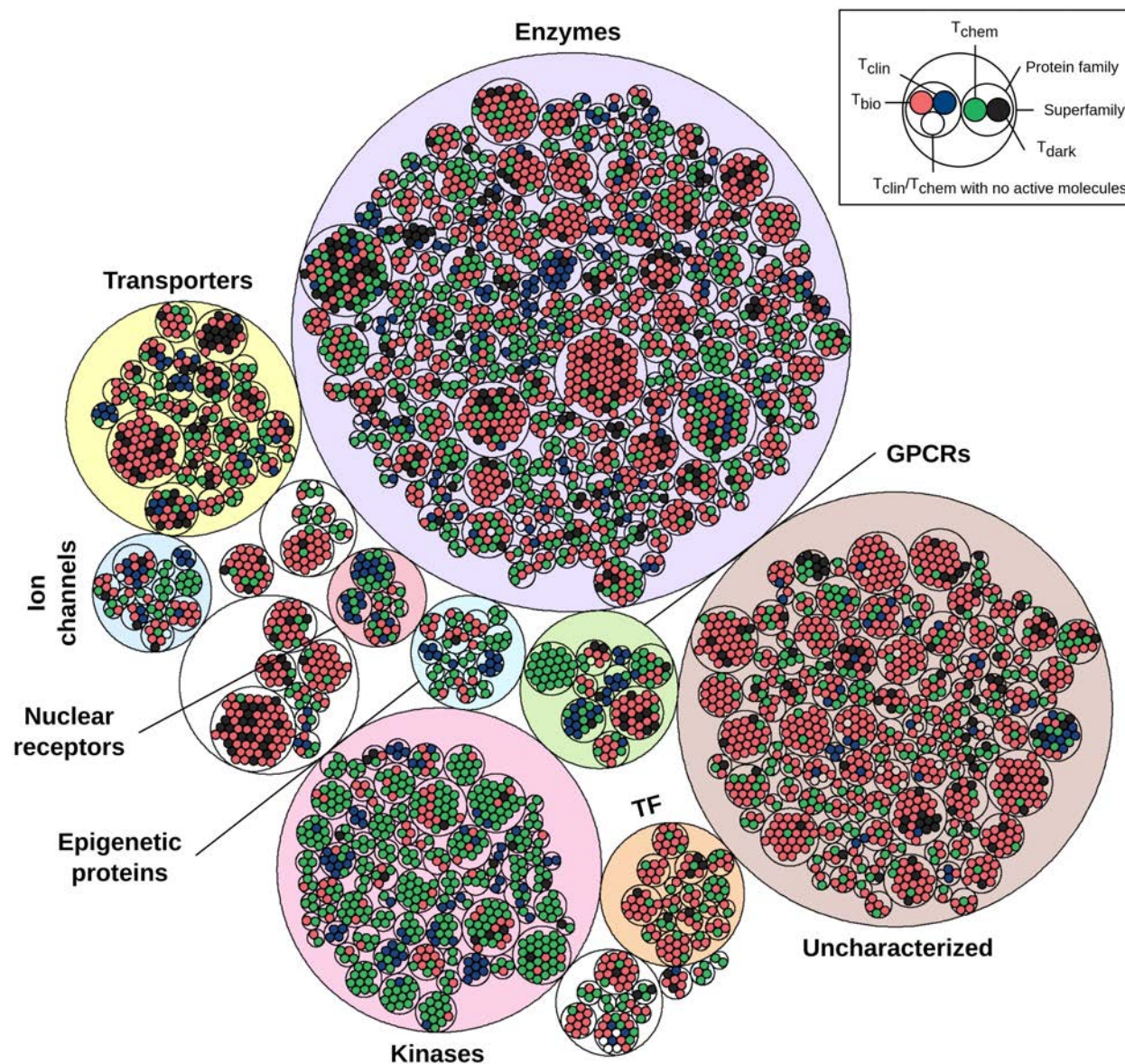




# Illuminating the Chemical Space

## ICS pipeline

- Collect **molecules** in the protein neighbourhood surrounding an untargeted protein.
- Identify most **shared chemical series** enriched with potent molecules.
- Scan repository of ***in-stock* compounds** to obtain a priority list for screening.



# Illuminating the Chemical Space

## ICS pipeline

- Collect **molecules** in the protein neighbourhood surrounding an untargeted protein.
- Identify most **shared chemical series** enriched with potent molecules.
- Scan repository of ***in-stock* compounds** to obtain a priority list for screening.

### Identification of the Core Chemical Structure in SureChEMBL Patents

Maria J. Falaguera and Jordi Mestres\*

Cite This: *J. Chem. Inf. Model.* 2021, 61, 2241–2247

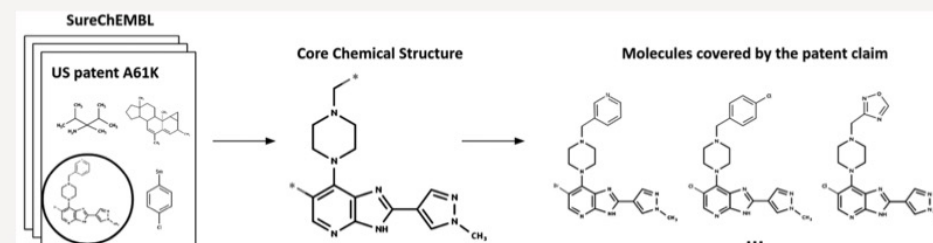
Read Online

ACCESS |

Metrics & More

Article Recommendations

Supporting Information



**ABSTRACT:** The SureChEMBL database provides open access to 17 million chemical entities mentioned in 14 million patents published since 1970. However, alongside with molecules covered by patent claims, the database is full of starting materials and intermediate products of little pharmacological relevance. Herein, we introduce a new filtering protocol to automatically select the core chemical structures best representing a congeneric series of pharmacologically relevant molecules in patents. The protocol is first validated against a selection of 890 SureChEMBL patents for which a total of 51,738 manually curated molecules are deposited in ChEMBL. Our protocol was able to select 92.5% of the molecules in ChEMBL from all 270,968 molecules in SureChEMBL for those patents. Subsequently, the protocol was applied to all 240,988 US pharmacological patents for which 9,111,706 molecules are available in SureChEMBL. The unsupervised filtering process selected 5,949,214 molecules (65.3% of the total number of molecules) that form highly congeneric chemical series in 188,795 of those patents (78.3% of the total number of patents). A SureChEMBL version enriched with molecules of pharmacological relevance is available for download at <https://ftp.ebi.ac.uk/pub/databases/chembl/SureChEMBLccs>.

# Illuminating the Chemical Space

## ICS pipeline

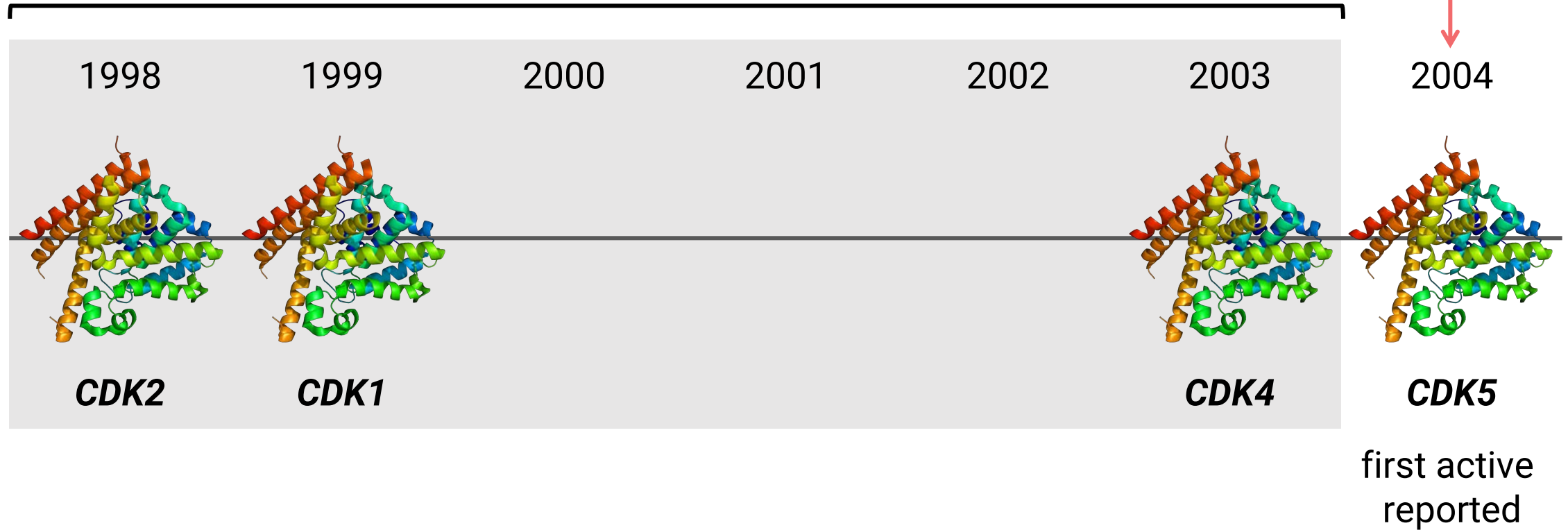
- Collect **molecules** in the protein neighbourhood surrounding an untargeted protein.
- Identify most **shared chemical series** enriched with potent molecules.
  1. Extraction of **Maximum Common Substructures (MCS)**
  2. MCS **ranking** according to siblings/molecules coverage
  3. Selection of **top ranked** substructures
- Scan repository of ***in-stock* compounds** to obtain a priority list for screening.

# ICS validation by virtual de-targeting exercise

Capacity to anticipate active compounds for untargeted proteins

Do privileged structures found for **CDK1/2/4** in 2003 anticipate active ligands for **CDK5**?

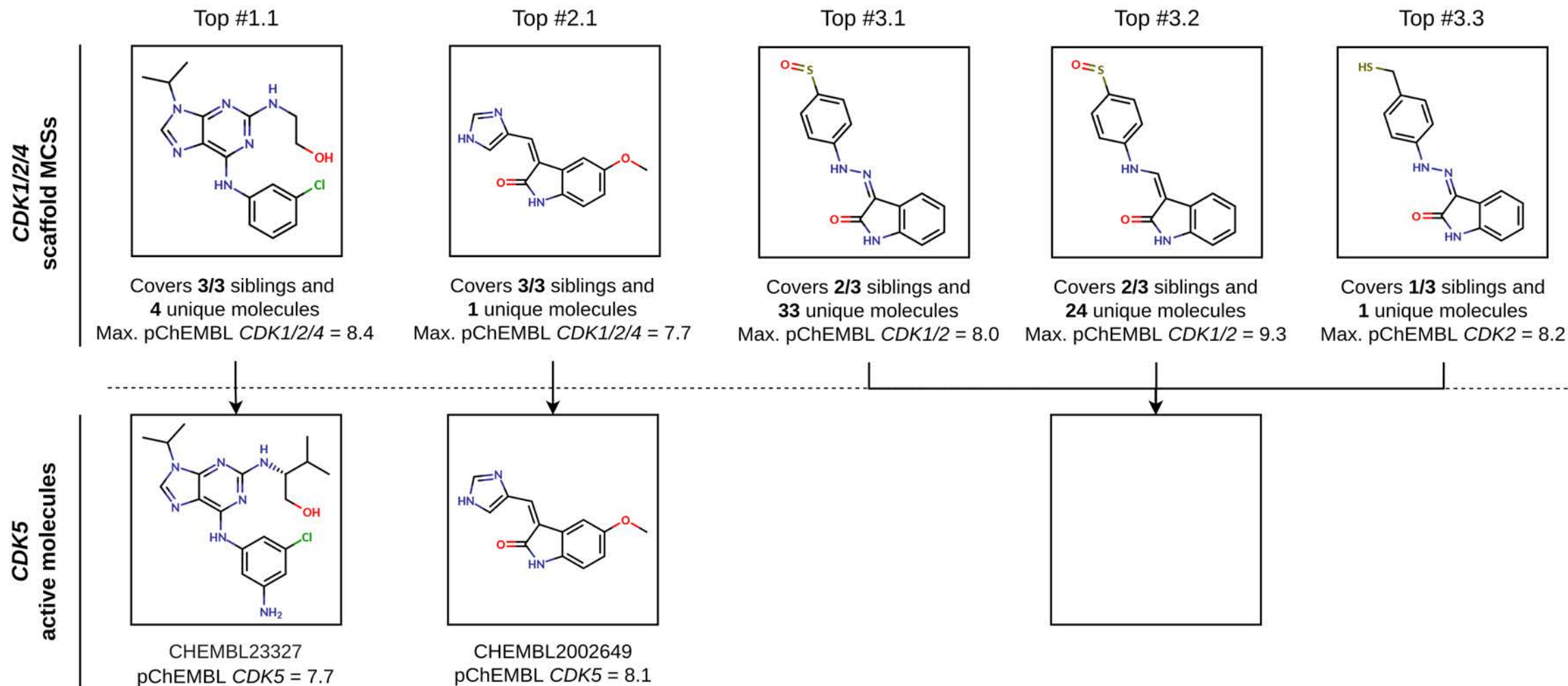
Run ICS pipeline





# ICS validation by virtual de-targeting exercise

## Capacity to anticipate active compounds for untargeted proteins





# ICS validation by virtual de-targeting exercise

Capacity to anticipate active compounds for untargeted proteins

Overall recall for  $T_{\text{clin}}$  = 61%

For 61% of the  $T_{\text{clin}}$ , the **PSs** identified for the siblings **one year before** the first bioactive molecule was reported for it are **contained in** at least one of the **molecules known today**

Overall recall for  $T_{\text{chem}}$  = 32%

$T_{\text{clin}}$  are **5x more chemically explored** than  $T_{\text{chem}}$

$T_{\text{clin}}$  have **2.5x more  $T_{\text{clin}}$  siblings** than  $T_{\text{chem}}$

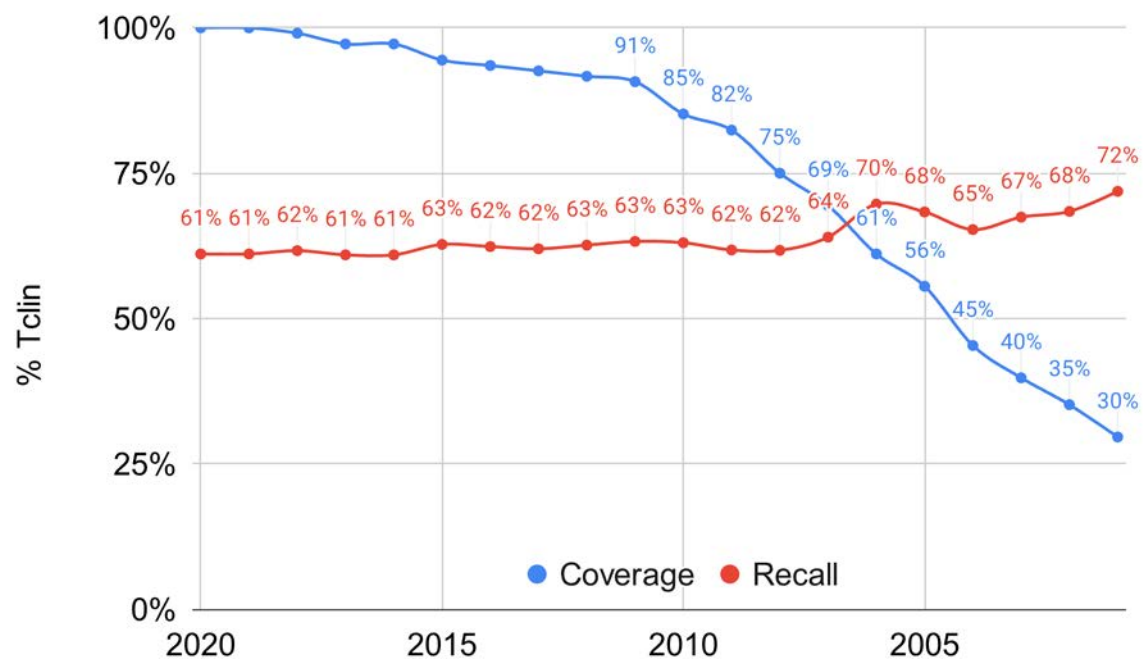
$T_{\text{clin}}$  are **6 years older** than  $T_{\text{chem}}$

Data completeness issues are reduced in  $T_{\text{clin}}$  scenarios

# ICS validation by virtual de-targeting exercise

Capacity to anticipate active compounds for untargeted proteins

Overall recall for  $T_{\text{clin}} = 61\%$

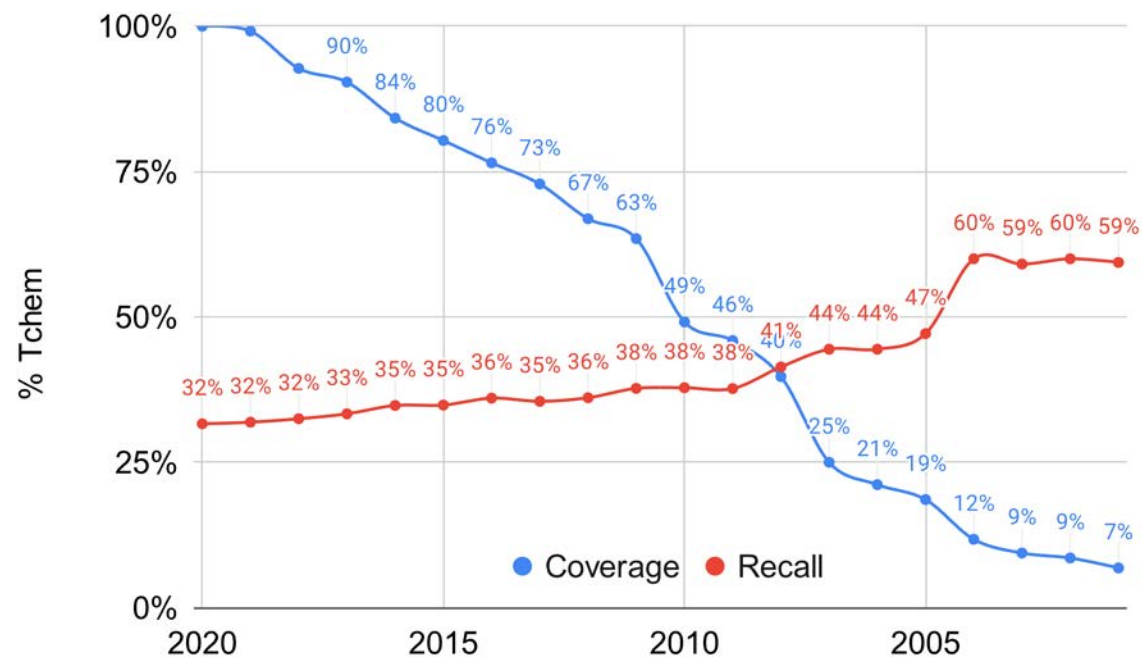


Newer target



Older target

Overall recall for  $T_{\text{chem}} = 32\%$



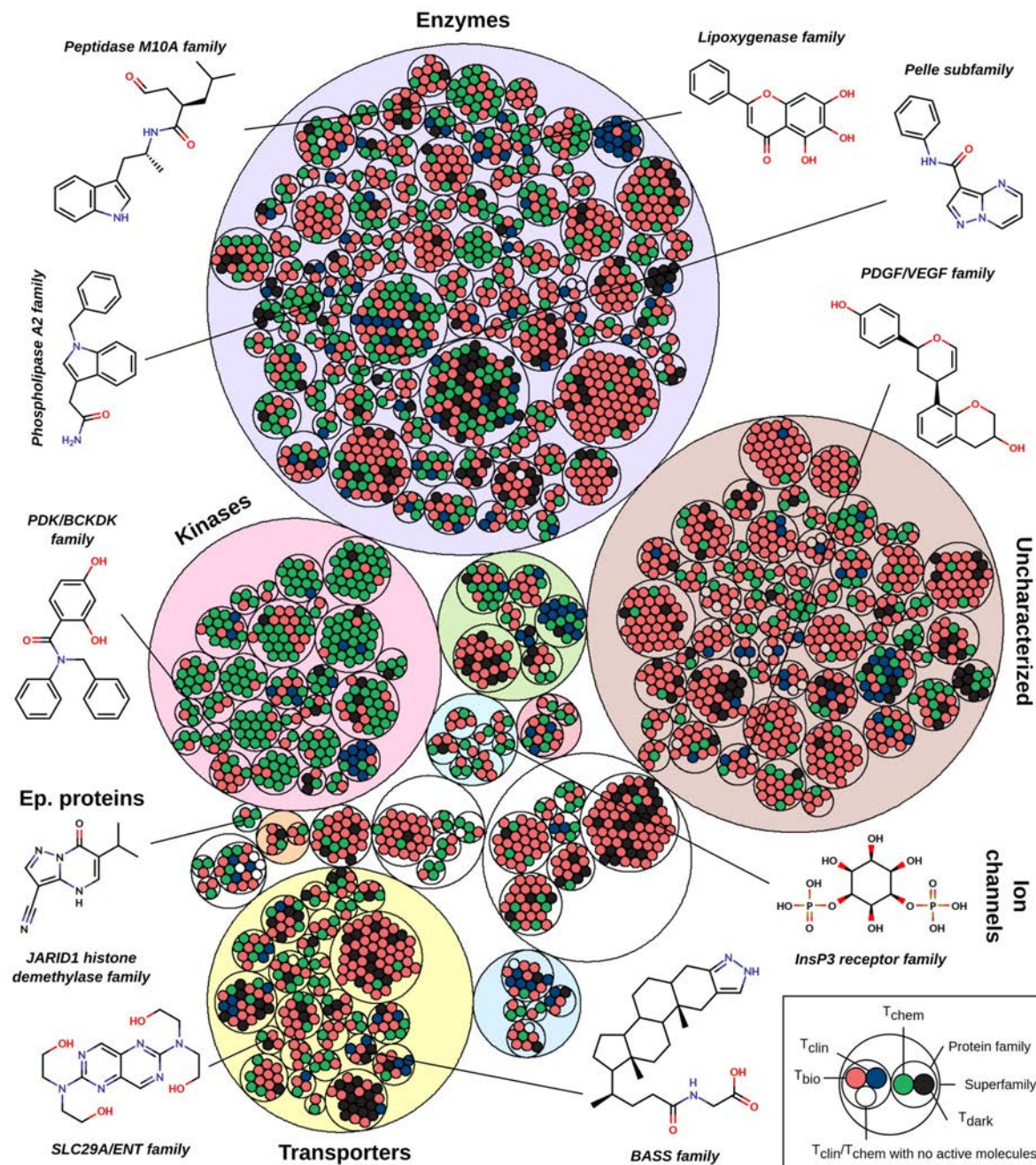
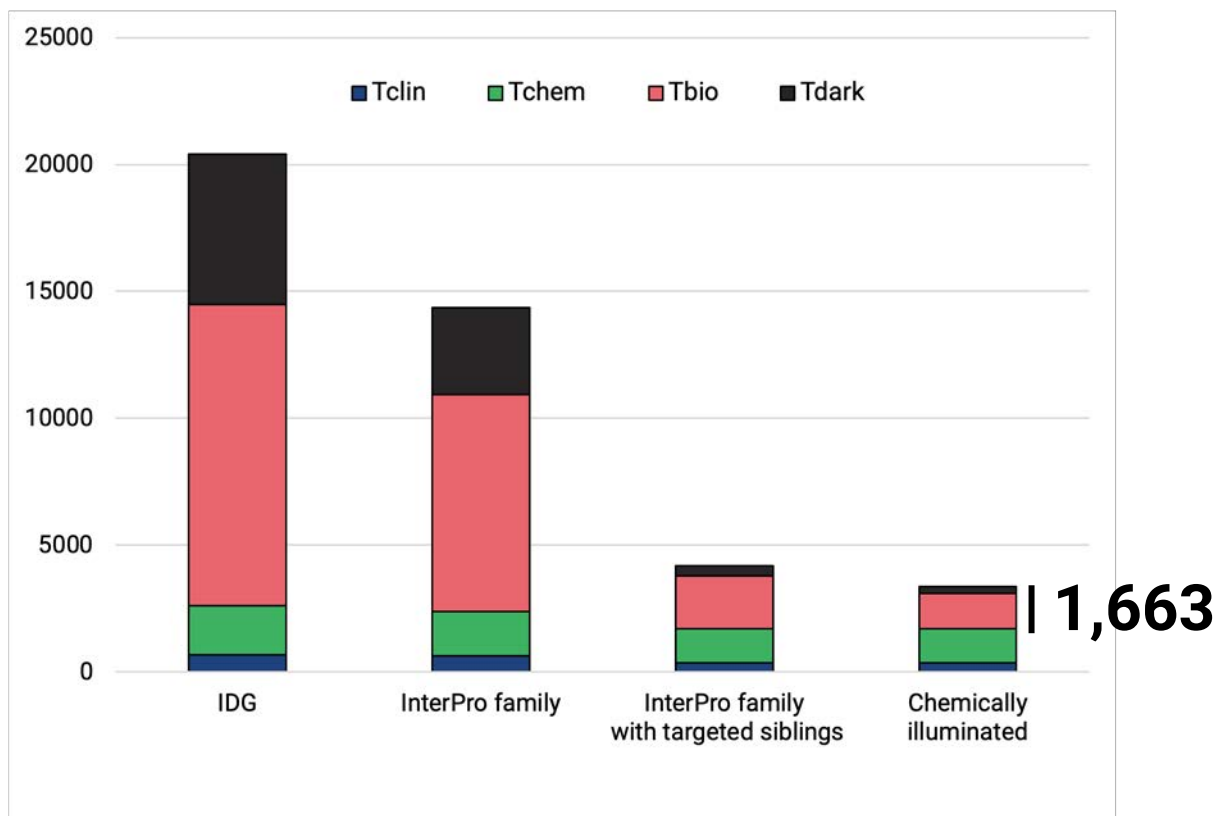
Newer target



Older target

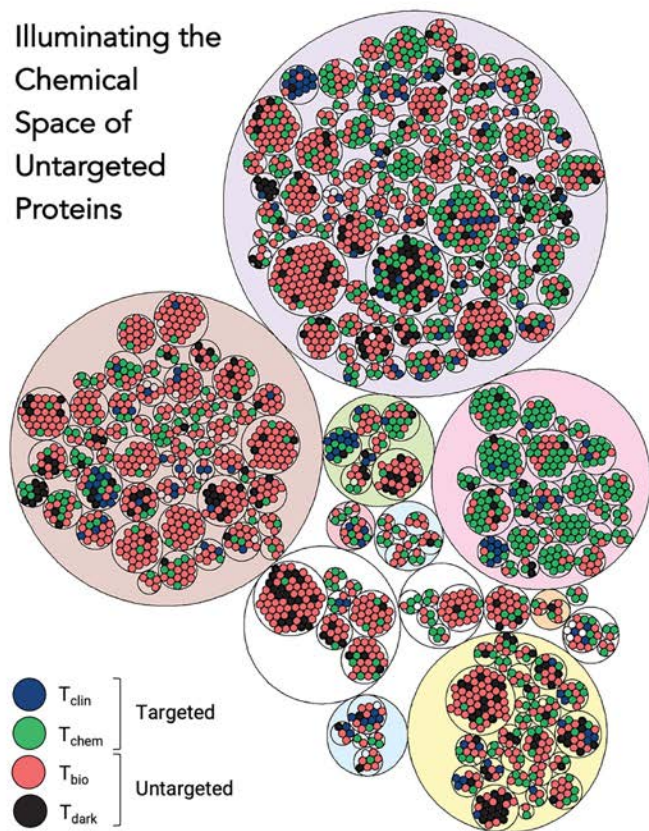
**1,663 untargeted proteins**  
chemically illuminated by the ICS protocol

**5,501 purchasable compounds**





Illuminating the  
Chemical  
Space of  
Untargeted  
Proteins



## Acknowledgements

Systems Pharmacology group (IMIM)  
Chemotargets SL

Questions and feedback welcome



mariaf@ebi.ac.uk



@mjfalaguera