GENERATIVE HETERO-ENCODER MODEL FOR DE NOVO DESIGN OF SMALL-MOLECULE COMPOUNDS AS POTENTIAL INHIBITORS OF BCR-ABL TYROSINE KINASE

Karpenko Anna, Vaitko Timofey, Tuzikov Alexander, Andrianov Alexander

National Academy of Sciences of Belarus

Problem

Chronic myeloid leukemia (CML) is a clonal hematopoietic stem cell disorder and accounts for approximately 30% of the incidence of adult leukemias

Currently available drugs have high toxicity and resistance

The incidence of CML increases with age

Miranda-Filho A, Piñeros M, Ferlay J, Soerjomataram I, Monnereau A, Bray F. Epidemiological patterns of leukaemia in 184 countries: a population-based study. *Lancet Haematol* (2018) 5(1):e14–24. 10.1016/S2352-3026(17)30232-6

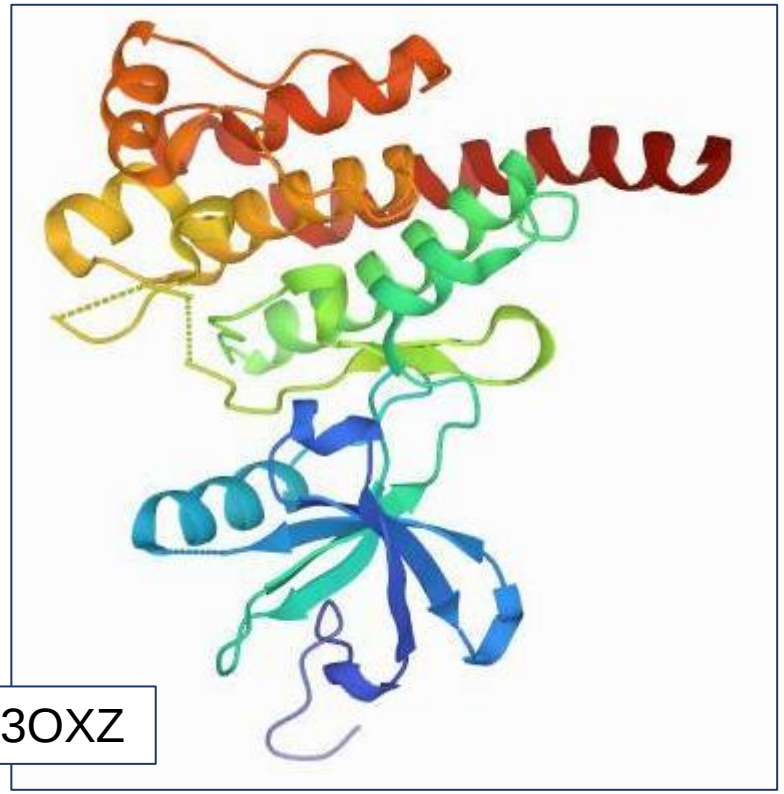
de la Fuente J, Baruchel A, Biondi A, de Bont E, Dresse MF, Suttorp M, et al. Managing children with chronic myeloid leukaemia (CML): recommendations for the management of CML in children and young people up to the age of 18 years. Br J Haematol (2014) 167(1):33–47. 10.1111/bjh.12977

Pipeline of solution

Phase	1. Target discovery	2. Screening	3. Lead generation	4. Validation
Goal	Find all targets from literature and Protein Data Bank	Create a molecular libraries Molecular docking	Selection and development of neural network architecture Generate molecules	Molecular docking Properties prediction

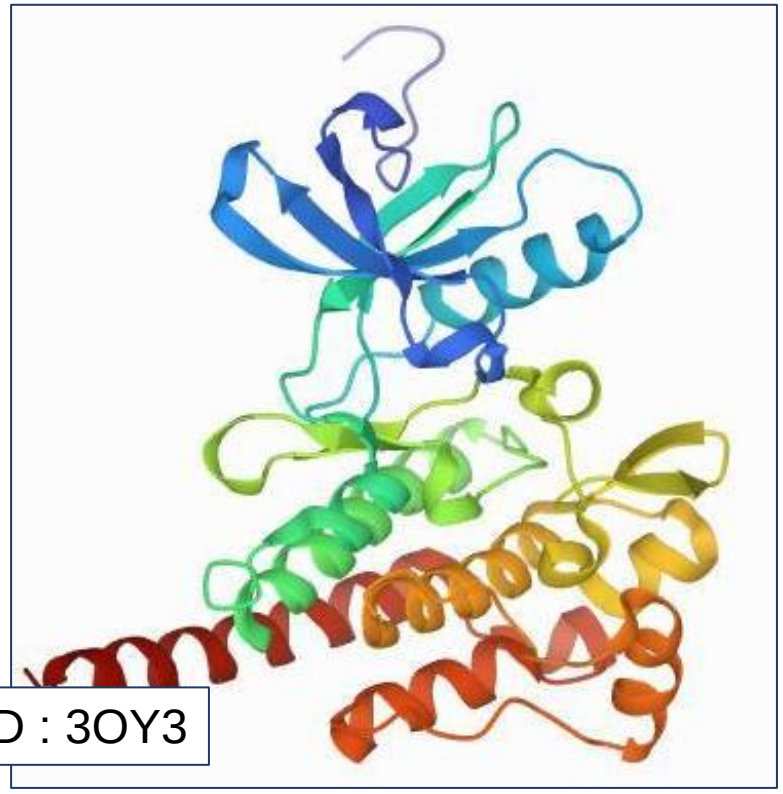
Solution

Phase	1. Target discovery
Goal	Find all targets from literature and Protein Data Bank



PDB ID : 3OXZ

Crystal structure of the ABL kinase domain associated with the DFG-out inhibitor AP24534

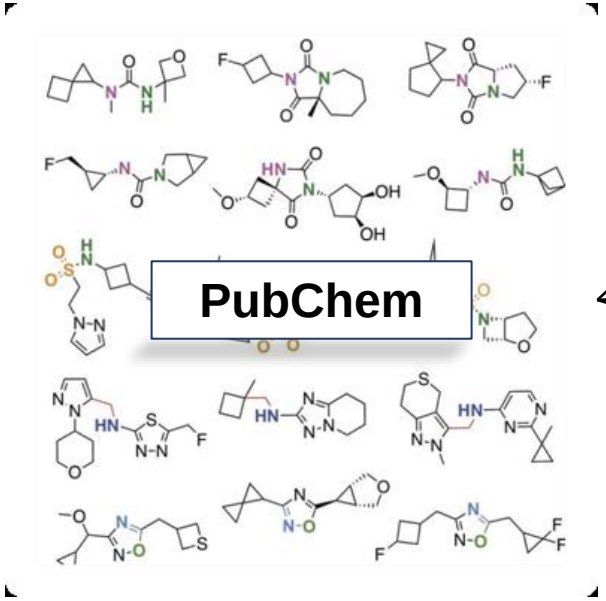


PDB ID : 3OY3

Crystal structure of the domain of the mutant kinase ABL T315I associated with the DFG-out inhibitor AP24589

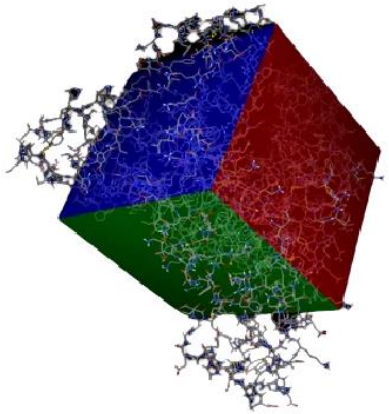
Solution

Phase	1. Target discovery	2. Screening
Goal	Find all targets from literature and Protein Data Bank	Create a molecular library Molecular docking



120.000 compounds containing **aryl-aminopyrimidine**

Molecular docking with rigid receptor and flexible ligand



AutoDock Vina

Ligand-receptor binding energy (Gibbs free energy)

RDKit

Feathers

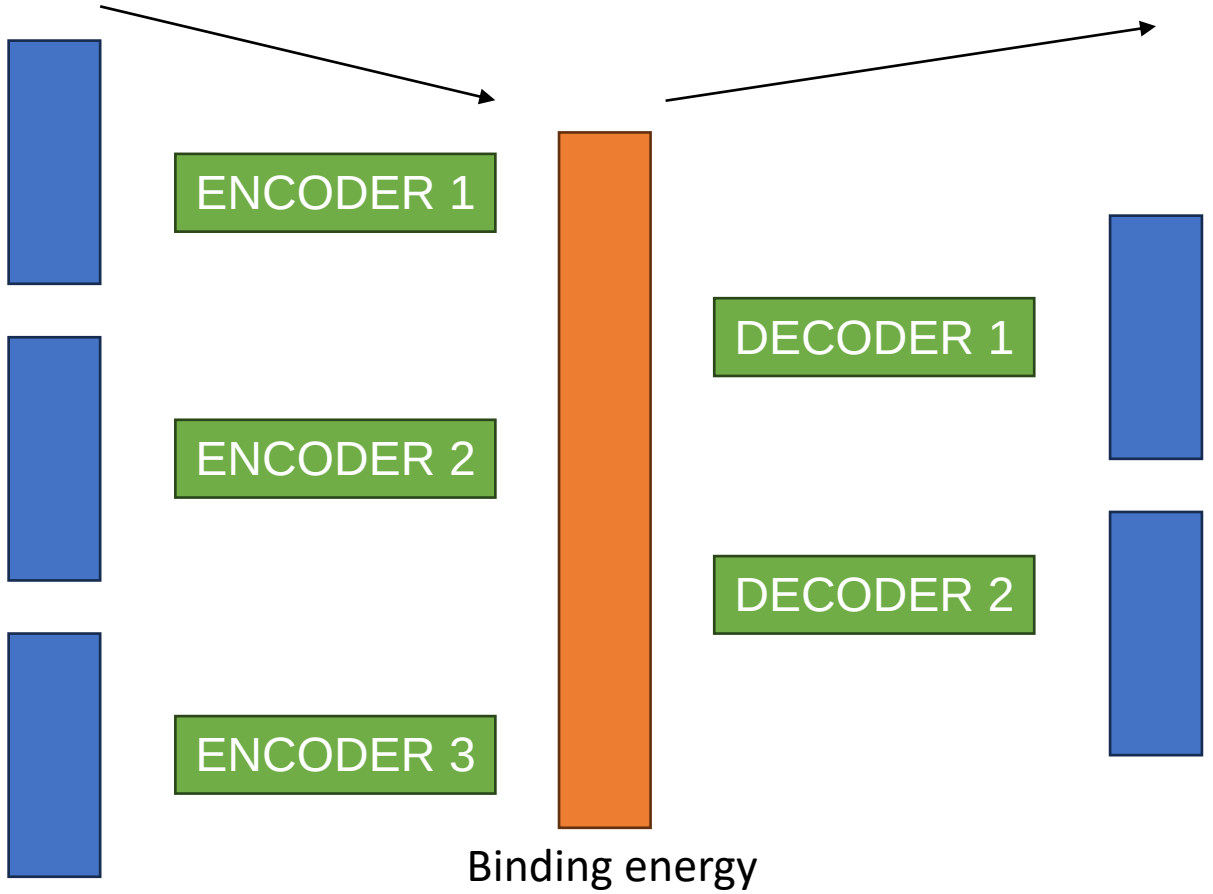
SMILES

CANONICAL SMILES

String-based approach

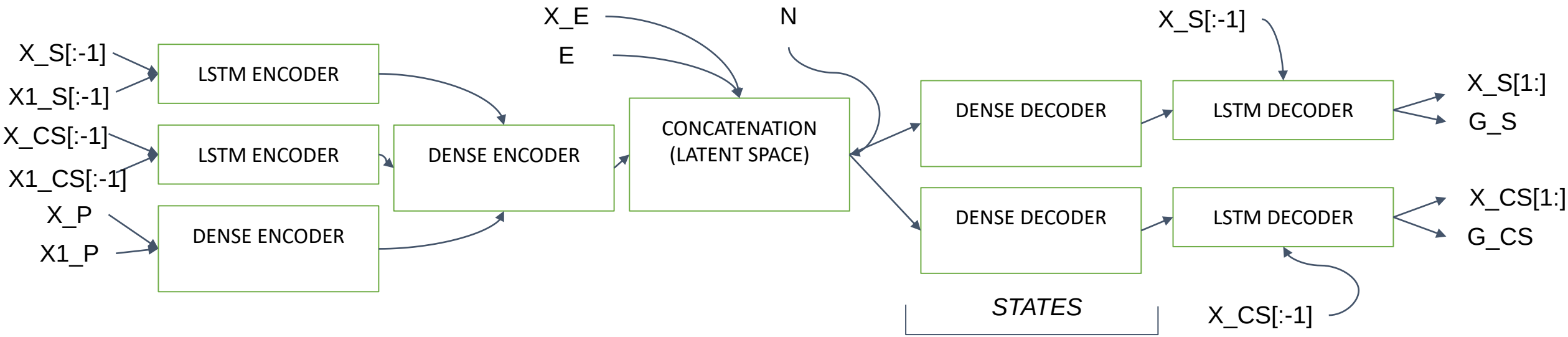
Solution

Phase	1. Target discovery	2. Screening	3. Lead generation
Goal	Find all targets from literature and Protein Data Bank	Create a molecular library Molecular docking	Selection and development of neural network architecture Generate molecules



Solution. ML WORKFLOW

Phase	1. Target discovery	2. Screening	3. Lead generation
Goal	Find all targets from literature and Protein Data Bank	Create a molecular library Molecular docking	Selection and development of neural network architecture Generate molecules



Training
X_S - Smiles One Hot Encoded
X_CS - Canonical Smiles One Hot Encoded
X_P - Characteristics of molecules
X_E - Binding Energy

Test
X1_S - Smiles One Hot Encoded
X1_CS - Canonical Smiles One Hot Encoded
X1_P - Characteristics of molecules

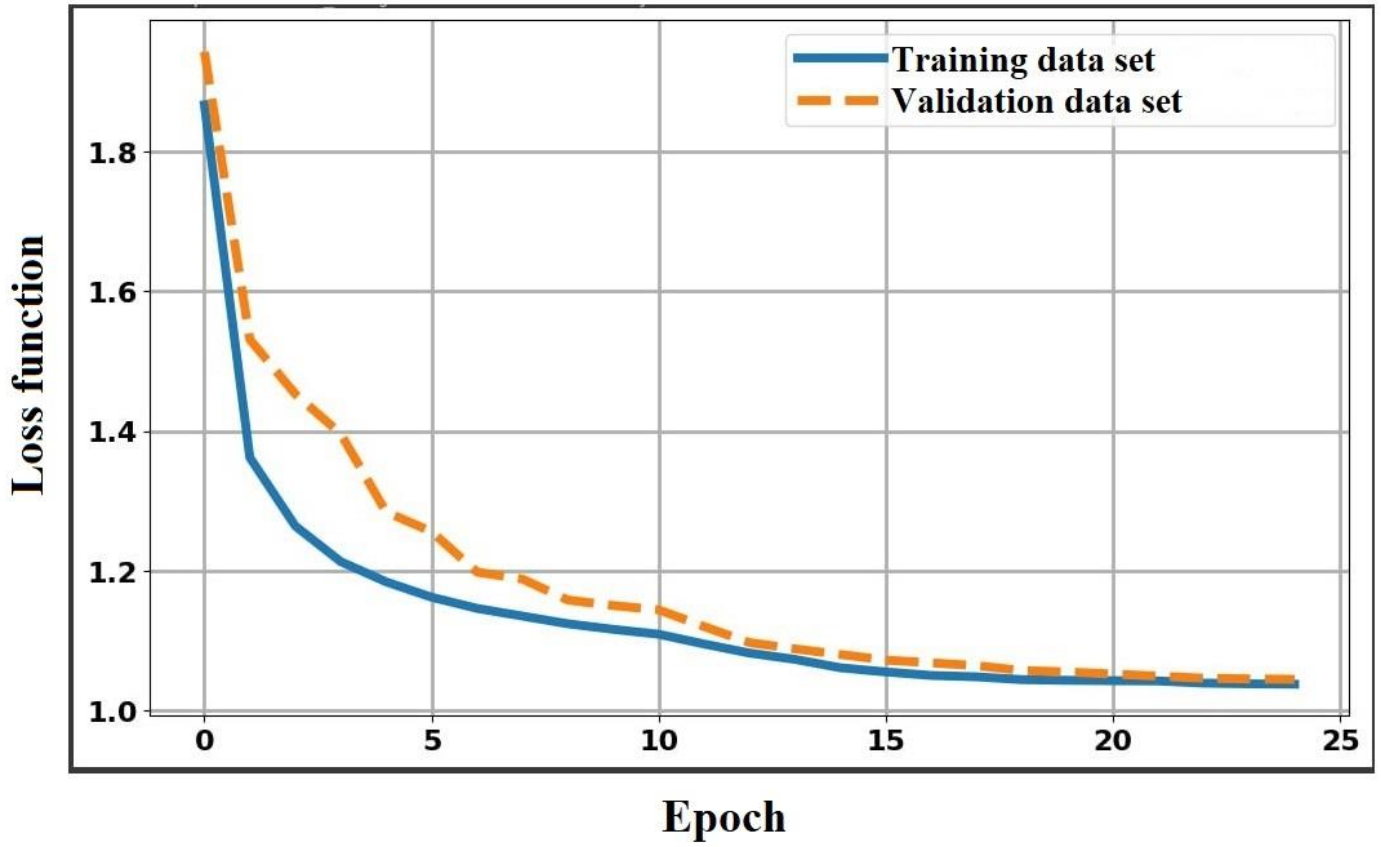
Generated
G_S- Smiles One Hot Encoded
G_CS - Canonical Smiles One Hot Encoded

Additional for generation:
E - desired energy
N - random noise

Result activation function	**SOFTMAX**
Optimizer	ADAM
Train size	Epochs = 25

Solution

Phase	1. Target discovery	2. Screening	3. Lead generation	4. Validation
Goal	Find all targets from literature and Protein Data Bank	Create a molecular library Molecular docking	Selection and development of neural network architecture Generate molecules	Molecular docking Properties prediction



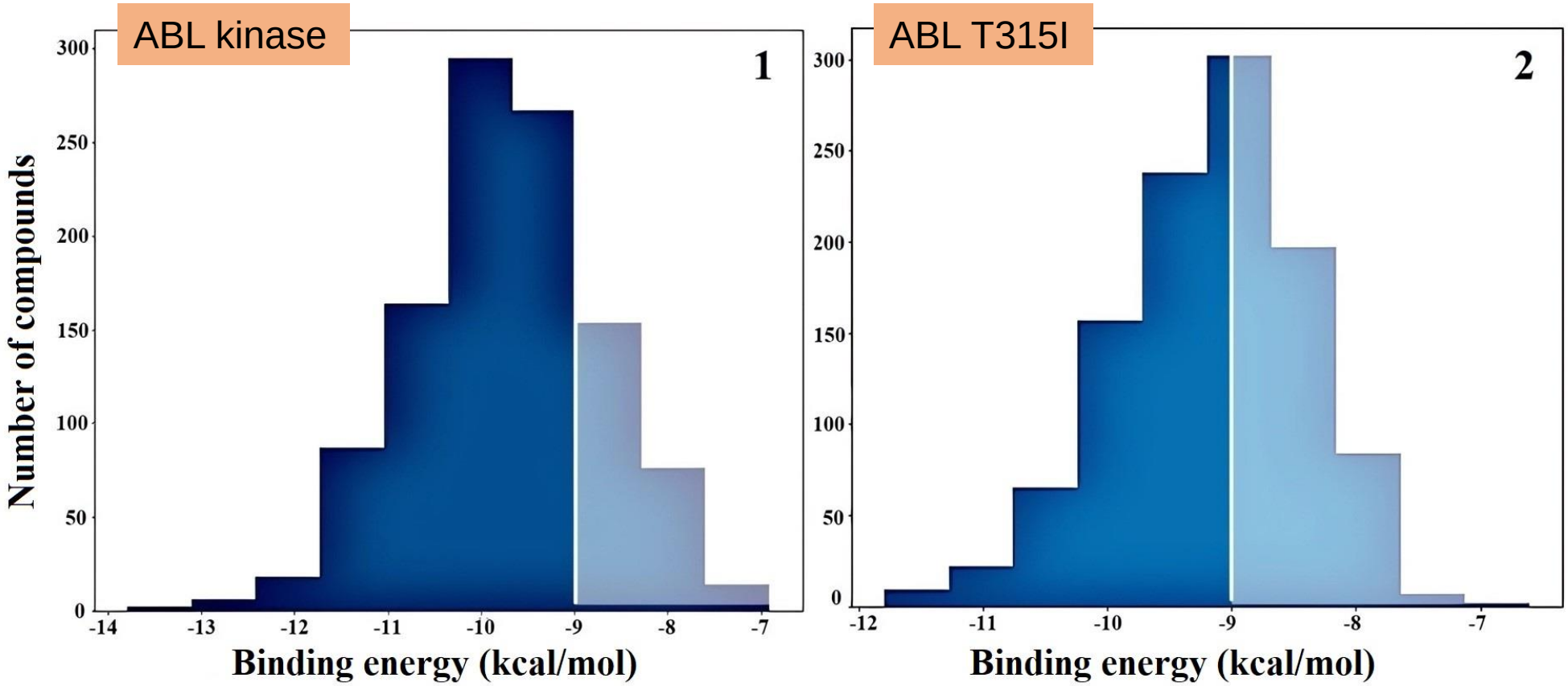
$$LF(s) = CCE(s) + 0.1 \cdot CCL(s),$$

- **CCE(s)** is the categorical cross entropy,
- **s** is a molecule in the SMILES format,
- **CCL(s)** (CustomChemLoss) is the function that imposes penalties for violations of a molecule stereochemistry and the absence of 2-arylaminopyrimidine in its chemical structure.

Solution

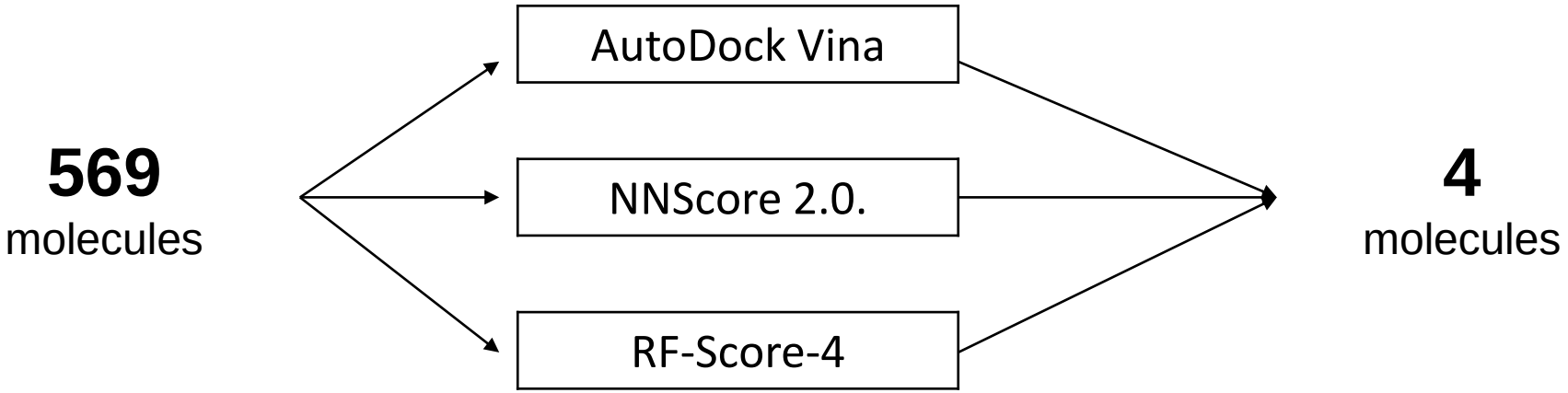
Phase	1. Target discovery	2. Screening	3. Lead generation	4. Validation
Goal	Find all targets from literature and Protein Data Bank	Create a molecular library Molecular docking	Selection and development of neural network architecture Generate molecules	Molecular docking Properties prediction

1083
unique molecules have been generated



Solution

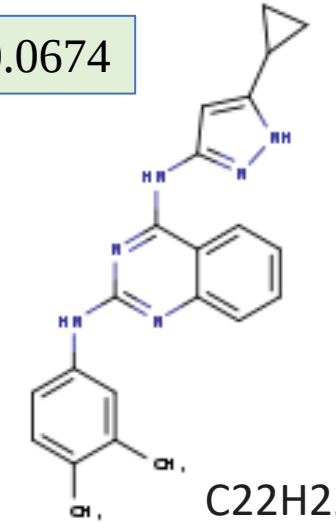
Phase	1. Target discovery	2. Screening	3. Lead generation	4. Validation
Goal	Find all targets from literature and Protein Data Bank	Create a molecular library Molecular docking	Selection and development of neural network architecture Generate molecules	Molecular docking Properties prediction



$$ECR = \sum_{sf} \frac{1}{\sigma_{sf}} \cdot \exp\left(-\frac{rank_{sf}}{\sigma_{sf}}\right),$$

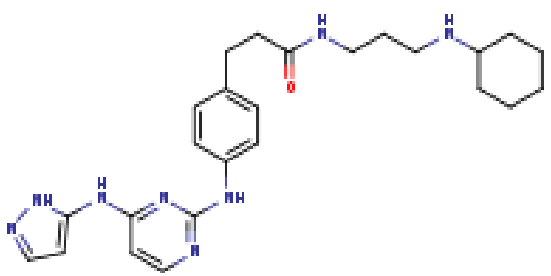
Results

CrossECR = 0.0674

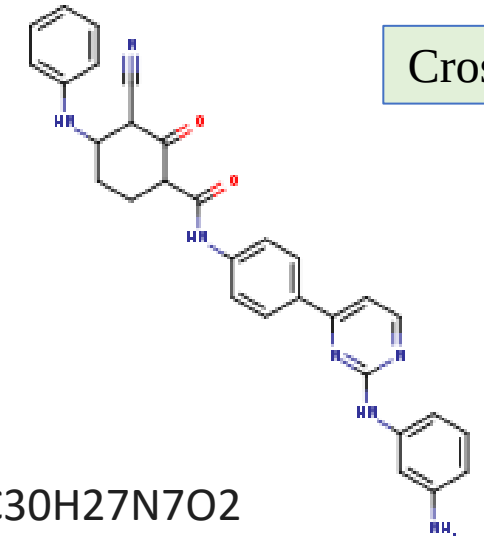


C22H22N6

CrossECR = 0.0835



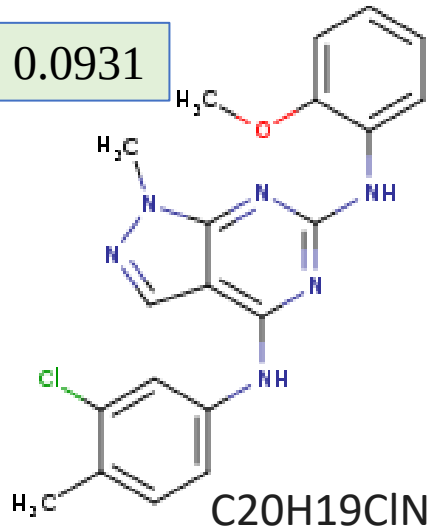
C25H34N8O



C30H27N7O2

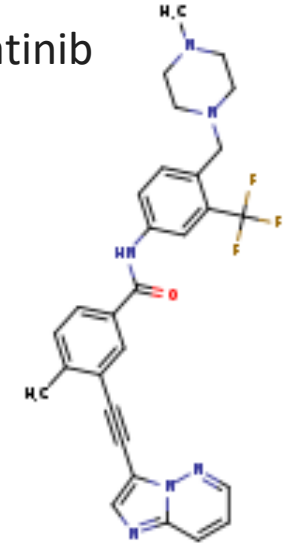
CrossECR = 0.0674

CrossECR = 0.0931



C20H19ClN6O

Ponatinib



CrossECR = 0.0399

$$crossECR(i) = \frac{ECR_1(i)}{max_i\{ECR_1(i)\}} + \frac{ECR_2(i)}{max_i\{ECR_2(i)\}},$$

THANK YOU

This study was supported by the State Program of Scientific Research “Convergence 2025” (subprogram “Interdisciplinary research and emerging technologies”, project 3.4.1).

Hanna Karpenka

rfe.Karpenko@gmail.com

https://www.linkedin.com/in/anna-karpenko-by/