"Fall into ML 2024"

25–26 октября 2024 г.,
г. Москва, НИУ ВШЭ и online

Steklov International Mathematical Center

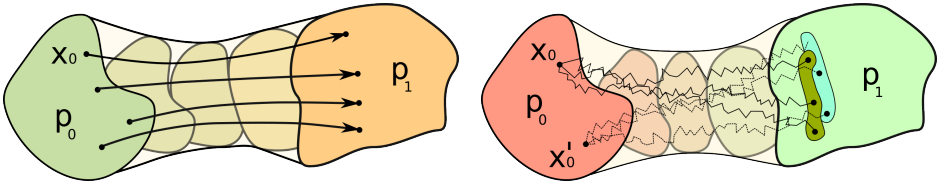
Организации

Национальный исследовательский университет “Высшая школа экономики” г. Москва

Математический институт им. В. А. Стеклова Российской академии наук, г. Москва

Математический центр мирового уровня
“Математический институт им. В. А. Стеклова Российской академии наук”
(МЦМУ МИАН), г. Москва

Конференция проводится при финансовой поддержке
Минобрнауки России (грант на создание и развитие МЦМУ МИАН,
соглашение № 075-15-2022-265).



New Perspective Methods of Generative AI

[based on flows and diffusion bridges]

Alexander Korotin

Skoltech
Skolkovo Institute of Science and Technology

AIRI

Moscow, 2024

Modern Generative Models for Images

Text prompt: woman's transparent futuristic inspired sneakers, glitter, depth of field

KANDINSKY

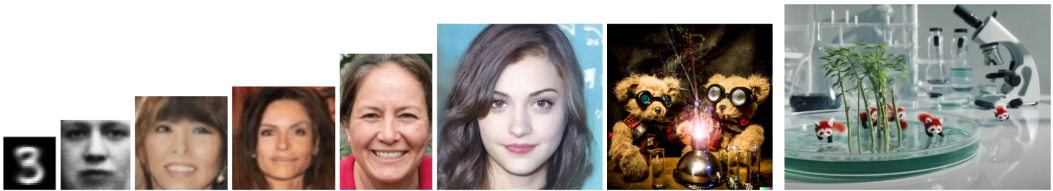
Text prompt: Chicken with potatoes baked in mayonnaise-sour cream sauce

SHEDEVRUM

Text prompt: 1967 Dodge Charger, moody lighting, side view, black, front view, lobby of the Louvre ...

MIDJOURNEY

Evolution of Generative Models



2013-2014
Initial attempts to learn generative models; VAE/GAN algorithms

2015-2017
Active development of (primarily GAN-based) algorithms, first convincingly scalable image generation models appear.

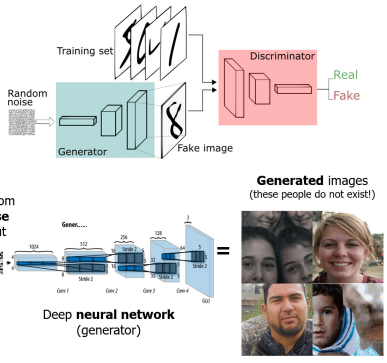
2019-2021
Diffusion-based models are introduced and actively developed.

2022
First inspiring large **text-to-image** models appear (based on diffusion models).

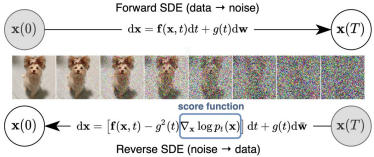
2024
Powerful **text-to-video** models appear based on diffusion models.

Principal Approaches to Generative Modeling[1][2]

Adversarial models (GANs, 2014)



Diffusion Models (DM, 2019)



[1] Ian Goodfellow et al. (2014). "Generative adversarial nets". In: *Advances in neural information processing systems*, pp. 2672–2680.

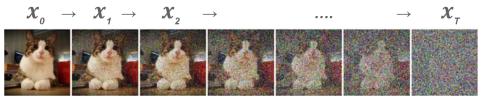
[2] Jascha Sohl-Dickstein et al. (2015). "Deep unsupervised learning using nonequilibrium thermodynamics". In: *International conference on machine learning*. PMLR, pp. 2256–2265.

Diffusion Models[3][4]

MAIN IDEA: reverse the data noising process.

Forward diffusion (noising SDE)

Take a data distribution $x_0 \sim p_0$ and gradually turn it to noise distribution $x_T \sim p_T = \mathcal{N}(0, \sigma^2 I)$.

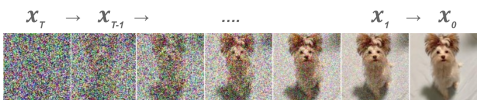


$$dx_t = f(x_t, t)dt + g(t)dW_t$$

(e.g., $dx_t = -\frac{1}{2}\beta_t dt + \sqrt{\beta_t} dW_t$)

Reverse diffusion (denoising SDE)

Sample from noise distribution $x_T \sim p_T$ and reverse the diffusion to get $x_0 \sim p_0$:



$$dx_t = [f(x_t, t) - g^2(t)\nabla_x \log p(x_t, t)]dt + g(t)d\overline{W_t}$$

(or $dx_t = [f(x_t, t) - \frac{1}{2}g^2(t)\nabla_x \log p(x_t, t)]dt$)

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel (2020). "Denoising diffusion probabilistic models". In: *Advances in neural information processing systems* 33, pp. 6840–6851.

[4] Yang Song et al. (2020). "Score-Based Generative Modeling through Stochastic Differential Equations". In: *International Conference on Learning Representations*.

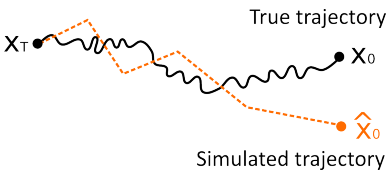
The Key Limitation of Diffusion Models: Time-Consuming Inference

To simulate the denoising process:

$$\mathrm{d}x_t = \left[f(x_t, t) - g^2(t)\nabla_x \log p(x_t, t)\right]\mathrm{d}t + g(t)\mathrm{d}\overline{W}_t$$

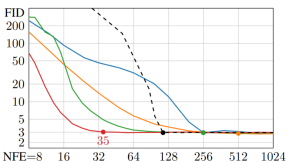
one uses the discretization (e.g., Euler-Maruyama simulation):

$$x_{t-\Delta t} = x_t - \left[f(x_t, t) - g^2(t)\nabla_x \log p(x_t, t)\right]\Delta t + g(t)\sqrt{\Delta t}\xi_t \quad , \xi_t \sim \mathcal{N}(0, I).$$



Remark:

NFE (# function evaluations) ≡ (# discretization steps)



Diff. models performance, CIFAR-10. FID w.r.t NFE.

Straightening The Trajectories of Diffusion Models

What we have

Not straight (deterministic or stochastic) trajectories, which are HARD to simulate.

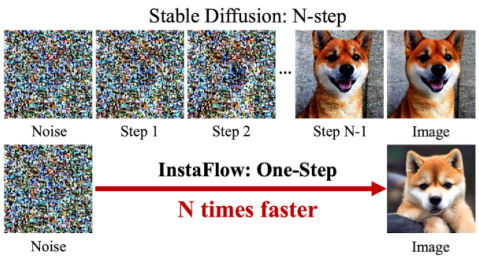
What we want

Straight (deterministic?) trajectories, which are EASY to simulate.

Part I: Flow Matching Framework and Rectified Flows

Teaser: Flow Matching Capabilities[5][6]

Insta**Flow**: 1 Step is Enough for HQ Diffusion-based Text-to-image Synthesis



Scaling Rectified **Flow** Transformers for High-Resolution Image Synthesis

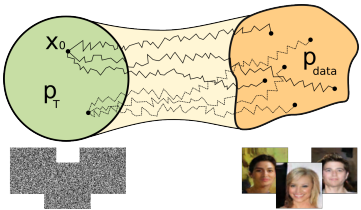
[5]Xingchao Liu, Xiwen Zhang, et al. (2023). "Instaflow: One step is enough for high-quality diffusion-based text-to-image generation". In: *The Twelfth International Conference on Learning Representations*.

[6]Patrick Esser et al. (2024). "Scaling rectified flow transformers for high-resolution image synthesis". In: *Forty-first International Conference on Machine Learning*.

Flow matching vs. Diffusion Models: Key Differences

Diffusion models framework (2019)

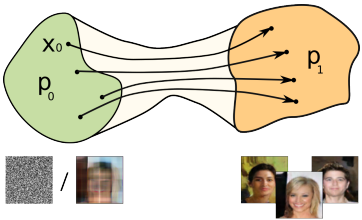
- maps given complex data distribution to the normal distribution.



- uses pre-defined noising process.
- (theoretically) requires infinite time horizon $[0, T]$.
- based on SDEs ($\Rightarrow$ complex stuff).

Flow matching framework (2023)

- maps arbitrary distribution p_0 to arbitrary distribution p_1.

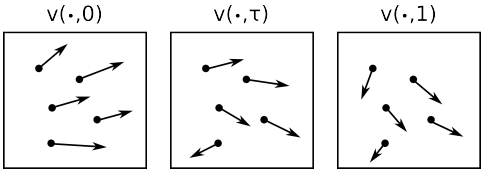


- no pre-defined process.
- finite time horizon $[0, 1]$. $\| T$
- based on ODEs.

Main Flow Matching-related papers

1. **Flow Matching (FM):** Yaron Lipman et al. (2022). "Flow Matching for Generative Modeling". In: *The Eleventh International Conference on Learning Representations*
2. **Rectified Flows (RF):** Xingchao Liu, Chengyue Gong, et al. (2022). "Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow". In: *The Eleventh International Conference on Learning Representations*
3. **Conditional FM (OT-CFM):** Alexander Tong et al. (2024). "Improving and generalizing flow-based generative models with minibatch optimal transport". In: *Transactions on Machine Learning Research*. Expert Certification. ISSN: 2835-8856. URL: https://openreview.net/forum?id=CD9Snc73AW
4. **Straightening FM:** Aram-Alexandre Pooladian et al. (2023). "Multisample Flow Matching: Straightening Flows with Minibatch Couplings". In: *International Conference on Machine Learning*. PMLR, pp. 28100–28127
5. **Optimal FM (OFM):** Nikita Kornilov et al. (2024). "Optimal Flow Matching: Learning Straight Trajectories in Just One Step". In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. URL: https://openreview.net/forum?id=kqmucDKVcU

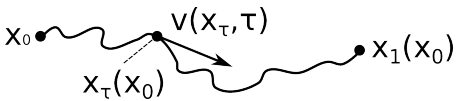
Vector field $v : \mathbb{R}^D \times [0,1] \to \mathbb{R}^D$.



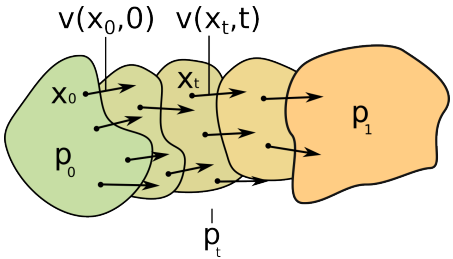
Movement of a point along the field.

Let $x_t(x_0)$ be the solution to $\mathrm{d}x_t = v(x_t, t)\mathrm{d}t$ with initial condition $x_{t=0} = x_0$, i.e.:

$$x_t(x_0) = x_0 + \int_0^t v(x_\tau(x_0), \tau)\mathrm{d}\tau.$$

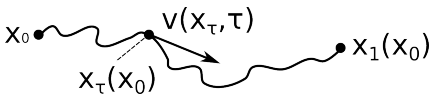


Flow Transport: Key Idea



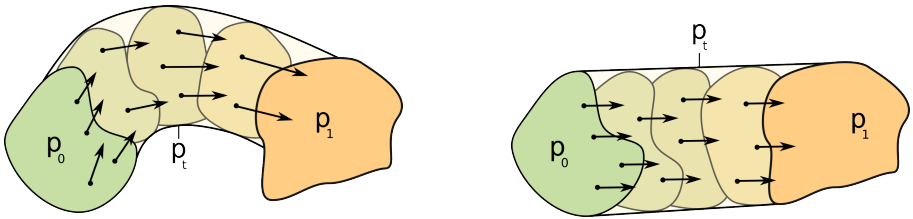
Find a (time-dependent) vector field $v(x_t, t)$ which transports the probability mass of distribution p_0 to distribution p_1, i.e.:

If $x_0 \sim p_0$ then $x_1(x_0) \sim p_1$



Flow Transport: Non-uniqueness

There may exist multiple vector fields v transporting p_0 to p_1.



Let p_t denote the distribution of $x_t(x_0)$ (for $x_0 \sim p_0$) obtained from p_0 by transporting its mass along the vector field $v(x_t, t)$.

How to construct at least one sequence of distributions p_t which transports p_0 to p_1?

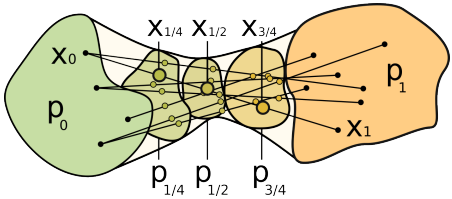
Remark: This should be much easier than constructing vector field v itself.

Flow Transport: Creating Interpolating Curve

Simple interpolation. Pick $x_0 \sim p_0$, $x_1 \sim p_1$ and set:

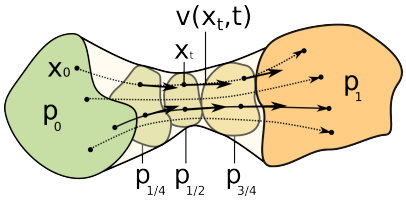
$$x_t = x_0 \cdot (1 - t) + x_1 \cdot t$$

Let p_t be the distribution of points x_t obtained with the procedure above.



How to find some vector field $v(x_t, t)$ which produces this sequence of distributions p_t by transporting p_0?

Flow Matching: Basic Algorithm[7]

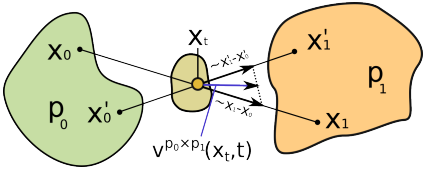


$$\begin{cases} \mathrm{d}x_t = v(x_t, t)\mathrm{d}t \\ x_0 \sim p_0 \end{cases}$$

One of such vector fields could be found as a solution to flow matching (FM) objective:

$$\min_{v} \mathop{\mathbb{E}}_{\substack{x_0 \sim p_0 \\ x_1 \sim p_1}} \left\{ \mathop{\mathbb{E}}_{t \sim [0,1]} \Big\| v(x_t, t) - (x_1 - x_0) \Big\|^2 \right\}.$$

$$x_0 \cdot (1-t) + x_1 \cdot t$$



Let $v^{p_0 \times p_1}(x_t, t)$ denote the minimizer to this problem.

[7] Xingchao Liu, Chengyue Gong, et al. (2022). "Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow". In: *The Eleventh International Conference on Learning Representations*.

Flow Matching: Practical Aspects[8]

$$v_{\theta^*} = \arg\min_{v_\theta} \mathop{\mathbb{E}}_{\substack{x_0 \sim p_0 \\ x_1 \sim p_1}} \left\{ \mathop{\mathbb{E}}_{t \sim [0,1]} \left\| v_\theta(x_t, t) - (x_1 - x_0) \right\|^2 \right\}$$

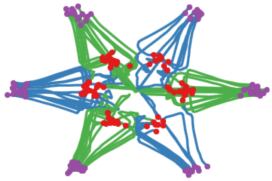
where $x_t = x_0 \cdot (1-t) + x_1 \cdot t$

- $\mathbb{E}_{x_0 \sim p_0, x_1 \sim p_1}(\cdot)$ is estimated using train samples $x_0 \sim p_0$ and $x_1 \sim p_1$ (datasets).
- t is sampled at random from Uniform$[0, 1]$.
- $v = v_\theta$ is a (large) Neural Network which takes data point x_t and time (time embedding) t as the input.

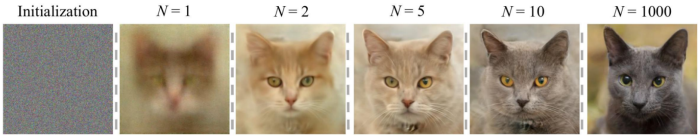
[8] Yaron Lipman et al. (2022). "Flow Matching for Generative Modeling". In: *The Eleventh International Conference on Learning Representations*.

Flow Matching: Preliminary Examples and Some Issues

Toy example:
Gaussians -> Gaussians



Example of image generation:
(different number of trajectory integration steps is shown)



Remark: if the trajectories were really straight, generation with different numbers of steps would give the same result, because there would be no integration errors.

Problem: trajectories are not straight enough.

Rectified Flows[9]

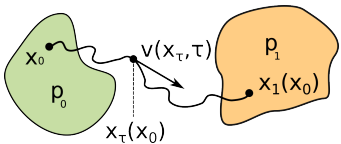
Key idea: iteratively repeat flow matching (FM) procedure to get straighter trajectories.

Step 1: pure flow matching.

$$v^1 = \arg\min_v \mathop{\mathbb{E}}_{\substack{x_0 \sim p_0 \\ x_1 \sim p_1}} \left\{ \mathop{\mathbb{E}}_{t \sim [0,1]} \Big\| v(x_t, t) - \underbrace{(x_1 - x_0)}_{x_0 \cdot (1-t) + x_1 \cdot t} \Big\|^2 \right\}.$$

Result: vector field v^1 moving p_0 to p_1:

$$x_0 \sim p_0 \wedge \mathrm{d}x_t = v^1(x_t, t)\mathrm{d}t \quad \Longrightarrow \quad x_1^{v^1}(x_0) \sim p_1.$$



[9] Xingchao Liu, Chengyue Gong, et al. (2022). "Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow". In: *The Eleventh International Conference on Learning Representations*.

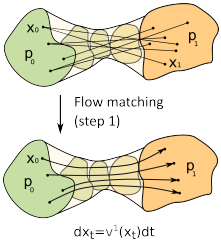
Rectified Flows

Step k+1: flow matching using samples $(x_0, x_1^{v^k}(x_0))$:

$$v^{k+1} = \arg\min_{v} \mathbb{E}_{x_0 \sim p_0} \left\{ \mathbb{E}_{t\sim[0,1]} \left\| v(x_t, t) - (x_1^{v^k}(x_0) - x_0) \right\|^2 \right\}.$$

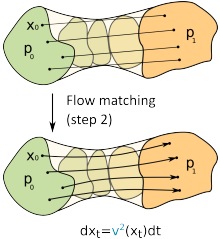
$x_0 \cdot (1-t) + x_1^{v^k}(x_0) \cdot t$

Step 1



Learn some vector field v^1 transporting p_0 to p_1.

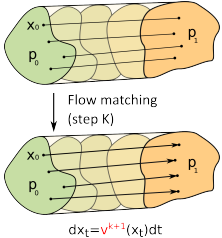
Step 2



Use the previously obtained field v^1 to get samples $(x_0, x_1^{v^1}(x_0))$ to train the new field v^2.

...

Step k+1

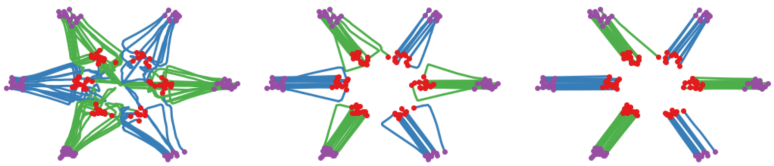


Use the previously obtained field v^k to get samples $(x_0, x_1^{v^k}(x_0))$ to train the new field v^{k+1}.

Rectified Flows: The Straightening Effect

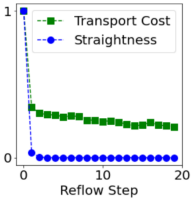
Theorem (*Informal*[10])

Vector field v^K *produces more straight trajectories* ($\mathrm{d}x_t = v^K(x_t, t)\mathrm{d}t \wedge x_0 \sim p_0$) *as* $K \to \infty$.



(a)The 1st rectified flow Z^1 (b) Reflow Z^2 (c) Reflow Z^3

Rectified Flows performance between source and target, $K = 1, 2, 3$.

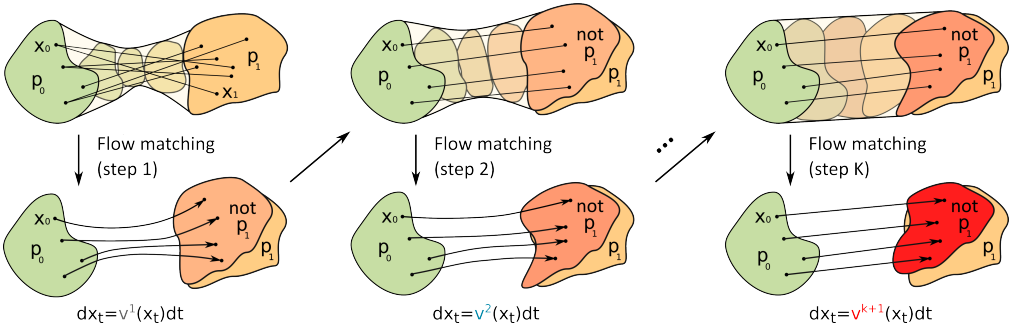


Transport cost, Straightness.

[10]Xingchao Liu, Chengyue Gong, et al. (2022). "Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow". In: *The Eleventh International Conference on Learning Representations*.

Rectified Flows: Practical Issues

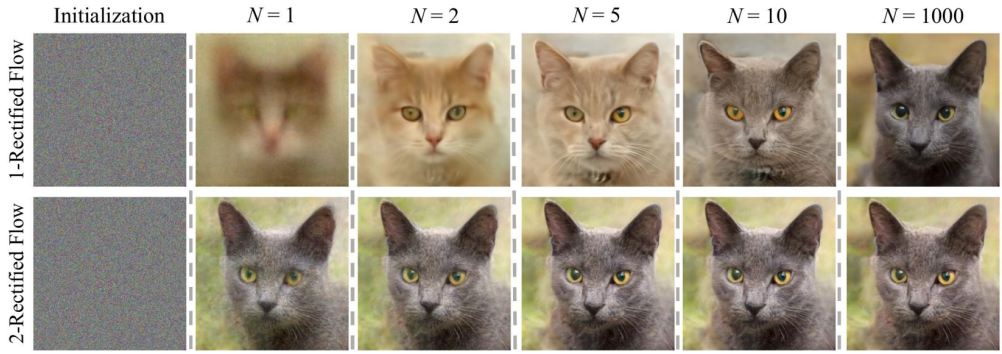
After the first flow matching step, the iterative approach **never** utilizes the true data $x_1 \sim p_1$ but uses learned samples $x_1^{v^1} p_1$, i.e., those obtained via $x_0 \sim p_0 \wedge \mathrm{d}x_t = v^1(x_t, t)\mathrm{d}t$.



This issue leads to the target matching error, i.e., the model never learns the true p_1 due to statistical, approximation, optimization errors, which accumulates with FM iterations.

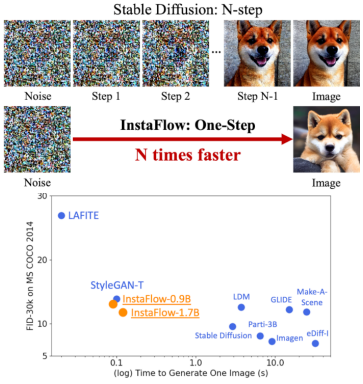
Rectified Flows: Image Generation Examples

Columns - different NFE, **rows** – different FM iterations.



For the 2-step FM the trajectories are rather straight which can be deduced from the fact that the model works reasonable even for $NFE = 1$ (2-Rectified flow, $N = 1$).

Instaflow: Rectified Flow as a Tool for Model Distillation[11]



[11]Xingchao Liu, Xiwen Zhang, et al. (2023). "Instaflow: One step is enough for high-quality diffusion-based text-to-image generation". In: *The Twelfth International Conference on Learning Representations*.

Stable Diffusion: Flow-based Large Text-to-Image Model[12]

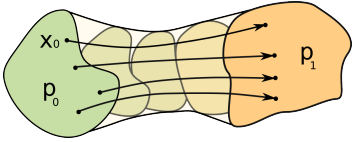
[12]Patrick Esser et al. (2024). "Scaling rectified flow transformers for high-resolution image synthesis". In: *Forty-first International Conference on Machine Learning*.

Part II: Bridge Matching Framework and Diffusion Schrodinger Bridge Matching

Flow Matching vs. Bridge Matching: a Comparison

Flow Matching

$$\mathrm{d}x_t = v(x_t, t)\mathrm{d}t$$



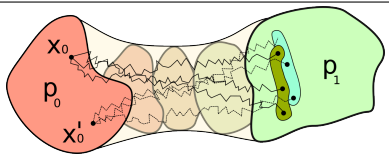
Define interpolation: $x_t \stackrel{\text{def}}{=} x_0 \cdot (1-t) + x_1 \cdot t.$

$$\min_v \mathbb{E}_{\substack{x_0 \sim p_0 \\ x_1 \sim p_1}} \mathbb{E}_{t \sim [0,1]} \Big\| v(x_t, t) - (x_1 - x_0) \Big\|^2.$$

Can be iterated to straighten the flows.
Related to the **Optimal Transport** (OT).

Bridge Matching

$$\mathrm{d}x_t = v(x_t, t)\mathrm{d}t + \sqrt{\epsilon}\mathrm{d}W_t \quad (\epsilon > 0).$$



Define a distribution: $p_t^\epsilon \stackrel{\text{def}}{=} \mathcal{N}\big(x_t, \epsilon t(1-t)\big)$

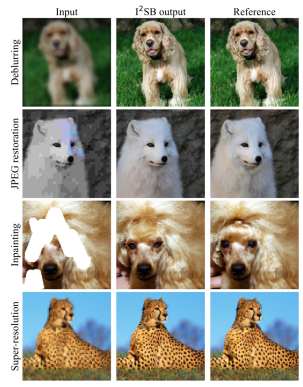
$$\min_v \mathbb{E}_{\substack{x_0 \sim p_0 \\ x_1 \sim p_1}} \mathbb{E}_{t \sim [0,1]} \mathbb{E}_{\tilde{x}_t \sim p_t^\epsilon} \Big\| v(\tilde{x}_t, t) - \frac{x_1 - \tilde{x}_t}{1-t} \Big\|^2.$$

Can be iterated and converges to the
Schrödinger bridge.

Examples of Bridge Matching Models for Images[13][14]

Image-to-image Schrodinger Bridge

for various image restoration problems



Diffusion Schrodinger Bridge Matching

for unpaired image-to-image translation

cat → wild

wild → cat

[13] Guan-Horng Liu et al. (2023). "I2SB: image-to-image Schrödinger bridge". In: *Proceedings of the 40th International Conference on Machine Learning*, pp. 22042–22062.

[14] Yuyang Shi et al. (2024). "Diffusion Schrödinger bridge matching". In: *Advances in Neural Information Processing Systems* 36.

Main Bridge Matching-related papers

1. **Bridge Matching (BM):** Stefano Peluchetti (2022). *Non-Denoising Forward-Time Diffusions*. URL: https://openreview.net/forum?id=oVfIKuhqfC
2. **Bridge Matching (BM):** Xingchao Liu, Lemeng Wu, Mao Ye, et al. (n.d.). "Let us Build Bridges: Understanding and Extending Diffusion Generative Models". In: *NeurIPS 2022 Workshop on Score-Based Methods*
3. **Iterative Bridge Matching (IBM):** Stefano Peluchetti (2022). *Non-Denoising Forward-Time Diffusions*. URL: https://openreview.net/forum?id=oVfIKuhqfC
4. **Iterative Markovian Fitting (IMF):** Yuyang Shi et al. (2024). "Diffusion Schrödinger bridge matching". In: *Advances in Neural Information Processing Systems* 36
5. **Discrete Iterative Markovian Fitting (D-IMF):** Nikita Gushchin, Daniil Selikhanovych, et al. (2024). "Adversarial Schrödinger Bridge Matching". In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. URL: https://openreview.net/forum?id=L3Knnigicu
6. **Optimal Schrödinger Bridge Matching (LightSB-M)** Nikita Gushchin, Sergei Kholkin, et al. (2024). "Light and Optimal Schrödinger Bridge Matching". In: *Forty-first International Conference on Machine Learning*

Part III: Our Results Related to Bridge/Flow Matching Models

Optimal Bridge/Flow Matching & Adversarial Bridge Matching

Our published papers (NeurIPS, ICML 2024):

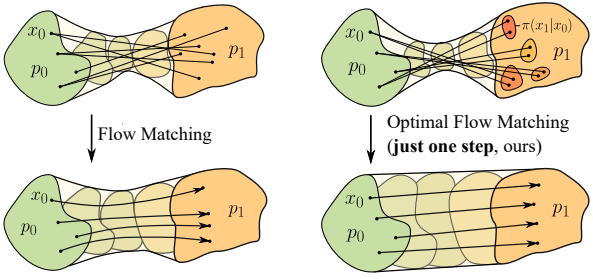
1. Nikita Kornilov et al. (2024). "Optimal Flow Matching: Learning Straight Trajectories in Just One Step". In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. URL: https://openreview.net/forum?id=kqmucDKVcU
2. Nikita Gushchin, Sergei Kholkin, et al. (2024). "Light and Optimal Schrödinger Bridge Matching". In: *Forty-first International Conference on Machine Learning*
3. Nikita Gushchin, Daniil Selikhanovych, et al. (2024). "Adversarial Schrödinger Bridge Matching". In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. URL: https://openreview.net/forum?id=L3Knnigicu

Our related pre-prints:

1. Sergei Kholkin et al. (2024). "Diffusion & Adversarial Schrodinger Bridges via Iterative Proportional Markovian Fitting". In: *arXiv preprint arXiv:2410.02601*

Optimal Flow Matching (OFM)[15]

Main idea: during FM minimization, consider *only specific vector fields* generating exactly straight trajectories. This optimization provably leads to *optimal transport* displacements.

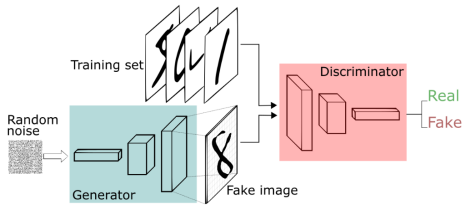


In **just one FM minimization** and **for any initial** π, we get straight trajectories + solve OT.

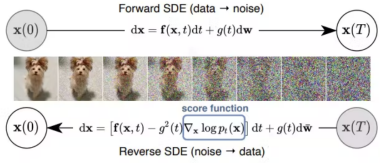
[15]Nikita Kornilov et al. (2024). "Optimal Flow Matching: Learning Straight Trajectories in Just One Step". In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. URL: https://openreview.net/forum?id=kqmucDKVcU.

Conclusion

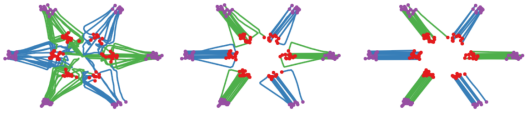
Generative Adversarial Networks (2014-...)



Diffusion models (2019-...)

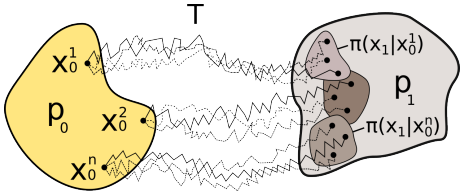


Flow/bridge matching models (2022-...)



What is next?

?



Building Light Schrödinger Bridges

Alexander Korotin
[1]Skolkovo Institute of Science and Technology
[2]Artificial Intelligence Research Institute

Skoltech
Skolkovo Institute of Science and Technology

AIRI

Moscow, 2024

Overview

Modern Diffusion Models and Their Limitations

Optimal Transport and Schrödinger Bridges

Part I. Light Schrödinger Bridge (ICLR 2024)

Part II. Light and Optimal Schrödinger Bridge Matching (ICML 2024)

Other works

Modern Diffusion Models and Their Limitations

Modern Generative Models for Images

Text prompt: woman's transparent futuristic inspired sneakers, glitter, depth of field

KANDINSKY

Text prompt: Chicken with potatoes baked in mayonnaise-sour cream sauce

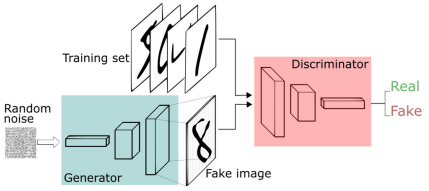
SHEDEVRUM

Text prompt: 1967 Dodge Charger, moody lighting, side view, black, front view, lobby of the Louvre ...

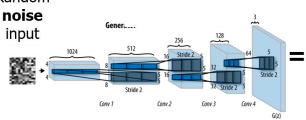
MIDJOURNEY

Principal Approaches to Generative Modeling

Adversarial models (GANs, 2014)



Random **noise** input

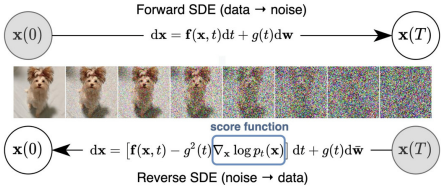


Deep **neural network** (generator)

=

Generated images (these people do not exist!)

Diffusion Models (DM, 2019)

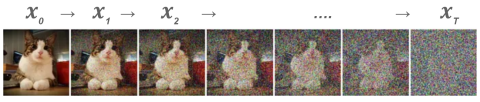


Diffusion Models[1][2]

MAIN IDEA: reverse the data noising process.

Forward diffusion (noising SDE)

Take a data distribution $x_0 \sim p_0$ and gradually turn it to noise distribution $x_T \sim p_T = \mathcal{N}(0, \sigma^2 I)$.

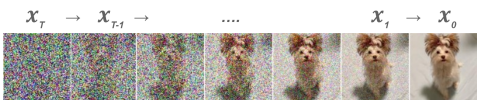


$$dx_t = f(x_t, t)dt + g(t)dW_t$$

(e.g., $dx_t = -\frac{1}{2}\beta_t dt + \sqrt{\beta_t} dW_t$)

Reverse diffusion (denoising SDE)

Sample from noise distribution $x_T \sim p_T$ and reverse the diffusion to get $x_0 \sim p_0$:



$$dx_t = [f(x_t, t) - g^2(t)\nabla_x \log p(x_t, t)]dt + g(t)d\overline{W_t}$$

(or $dx_t = [f(x_t, t) - \frac{1}{2}g^2(t)\nabla_x \log p(x_t, t)]dt$)

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel (2020). "Denoising diffusion probabilistic models". In: *Advances in neural information processing systems* 33, pp. 6840–6851.

[2] Yang Song et al. (2020). "Score-Based Generative Modeling through Stochastic Differential Equations". In: *International Conference on Learning Representations*.

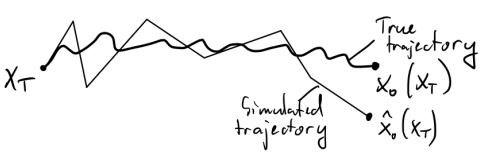
Limitation 1 of Diffusion Models: Time-Consuming Inference

To simulate the denoising process:

$$x_t = \left[f(x_t, t) - g^2(t)\nabla_x \log p(x_t, t)\right] t + g(t)\overline{W}_t$$

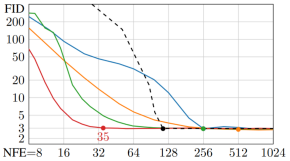
one uses the discretization (e.g., Euler-Maruyama simulation):

$$x_{t-\Delta t} = x_t - \left[f(x_t, t) - g^2(t)\nabla_x \log p(x_t, t)\right]\Delta t + g(t)\sqrt{\Delta t}\xi_t \quad , \xi_t \sim \mathcal{N}(0, I).$$



Remark:

NFE (# function evaluations) $\equiv$ (# discretization steps)



Diff. models performance, CIFAR-10. FID w.r.t NFE.

Desire 1: Straightening The Trajectories of Diffusion Models

What we have

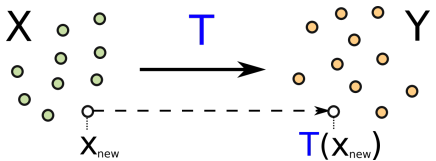
Not straight (deterministic or stochastic) trajectories, which are HARD to simulate.

What we want

Straight (deterministic?) trajectories, which are EASY to simulate.

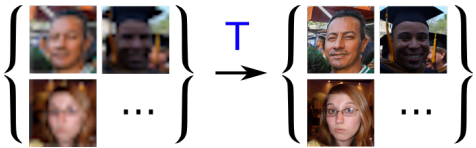
Limitation 2: Inapplicability to (Unpaired) Domain Translation

The task: learn (from samples) a *translation* map between the two given data domains.

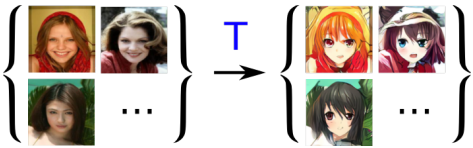


Important: the map should generalize to new data (similar to the train set).

Example 1: Image Super-Resolution



Example 2: Style Translation

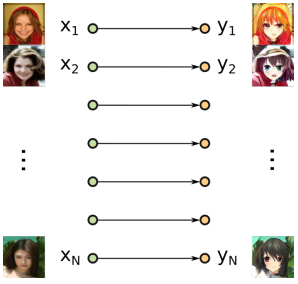


Desire 2: Be able to Solve Unpaired Domain Translation with DMs[3]

Supervised

Paired train samples are available:

$\{(x_1, y_1), \ldots, (x_N, y_N)\}$.

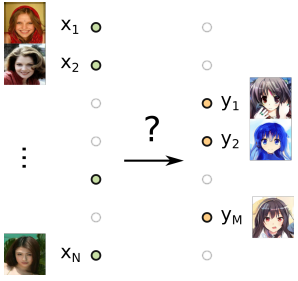


Conditional DMs are applicable.

Unsupervised (our interest)

Only *unpaired* train samples are given:

$\{x_1, \ldots, x_N\}, \{y_1, \ldots, y_M\}$.



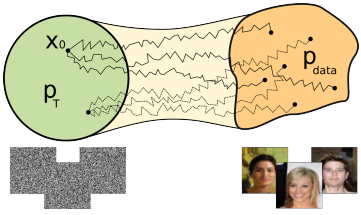
DMs are not applicable.

[3] Jun-Yan Zhu et al. (2017). "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232.

Schrödinger Bridges vs. Diffusion Models: Key Differences

Diffusion models framework (2019)

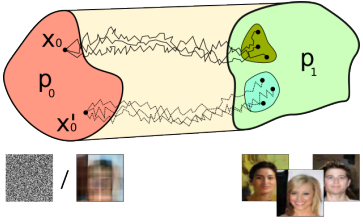
- maps given complex data distribution to the normal distribution.



- uses pre-defined noising process and learns the de-noising process.
- requires infinite time horizon $[0, T]$.

Schrödinger bridge framework (2021)

- maps arbitrary distribution p_0 to arbitrary distribution p_1.



- learns a diffusion that is maximally similar to a given prior process.
- finite time horizon $[0, \underset{\|}{1}]$.
  T

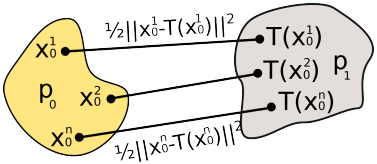
Optimal Transport and Schrödinger Bridges

Monge's Formulation of Optimal Transport[4] (with the Quadratic Cost)

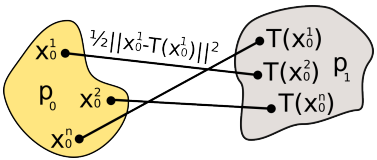
The optimal transport **cost** between distributions $p_0, p_1 \in \mathcal{P}_{2,ac}(\mathbb{R}^D)$ is

$$\text{Cost}(p_0, p_1) = \inf_{T \sharp p_0 = p_1} \int_{\mathcal{X}} \frac{\|x_0 - T(x_0)\|^2}{2} p_0(x_0) dx_0.$$

The map T^* attaining the minimum is called the optimal **transport map**.



Optimal map



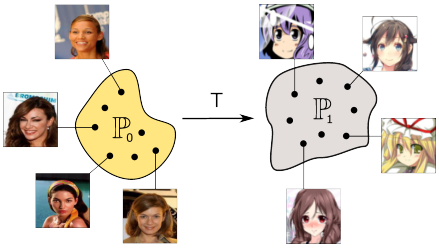
Not optimal map

[4] Cédric Villani (2008). *Optimal transport: old and new.* Vol. 338. Springer Science & Business Media.

Optimal transport applications

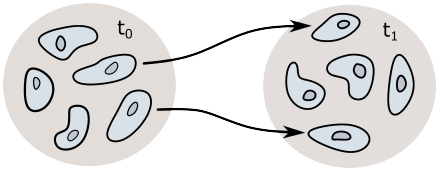
Domain Translation.[2]

By considering two unpaired image datasets as samples from p_0 and p_1, OT learns a map between datasets that preserves content.



Single-Cell (SC) Biological data.[3]

SC technology determines the gene expression profile of each measured cell, but destroys all measured cells. OT learns a map between cell populations before and after the perturbation.



[5]Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev (2022). "Neural Optimal Transport". In: *The Eleventh International Conference on Learning Representations*.

[6]Charlotte Bunne et al. (2023). "Learning single-cell perturbation responses using neural optimal transport". In: *Nature Methods*, pp. 1–10.

Entropic Optimal Transport (OT)[7]

Consider two distributions $p_0, p_1 \in \mathcal{P}_{2,ac}(\mathbb{R}^D)$.

Entropic OT (EOT) is formulated as follows:

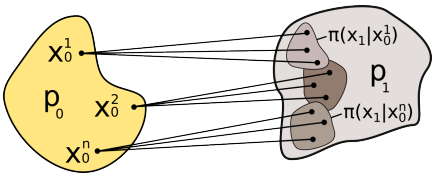
$$\inf_{\pi \in \Pi(p_0,p_1)} \int_{\mathbb{R}^D} C\big(x_0, \pi(\cdot|x_0)\big) p_0(x_0) dx_0.$$

The minimizer π^* is called the Entropic OT plan.

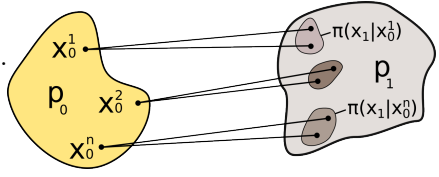
$$C\big(x_0, \pi(\cdot|x_0)\big) \overset{def}{=} \int_{\mathbb{R}^D} \underbrace{\frac{\|x_0 - x_1\|^2}{2}}_{\text{Dissimilarity}} \pi(x_1|x_0) dx_1 - \epsilon \underbrace{H(\pi(\cdot|x_0))}_{\text{Diversity}}.$$

Regularization strength ϵ controls the diversity.

- $\Pi(p_0, p_1)$ are distributions on $\mathbb{R}^D \times \mathbb{R}^D$ with marginals p_0, p_1



Stochastic EOT maps for large ϵ.



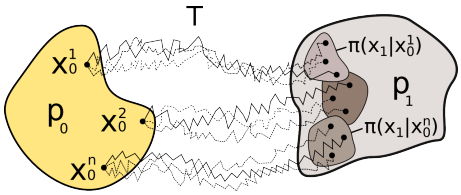
Stochastic EOT maps for small ϵ.

[7] Marco Cuturi (2013). "Sinkhorn distances: Lightspeed computation of optimal transport". In: *Advances in neural information processing systems* 26.

Schrödinger Bridge (SB) problem[8]

Consider two distributions $p_0, p_1 \in \mathcal{P}_{2,ac}(\mathbb{R}^D)$.
The Schrödinger bridge problem is:

$$\inf_{T \in \mathcal{F}(p_0,p_1)} \mathrm{KL}(T \| W^\epsilon),$$



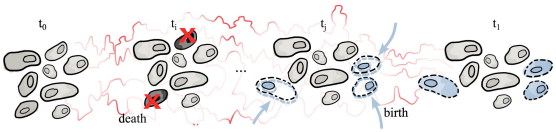
- $\mathcal{F}(p_0, p_1)$ are stochastic processes with marginals p_0, p_1 at $t = 0$ and $t = 1$ respectively.
- W^ϵ is the Wiener process with the variance ϵ.

The process T^* attaining the minimum has joint distribution $\pi^{T^*} = \pi^*$ at time moments $t = 0, 1$ which is the solution to the Entropic OT with regularization parameter ϵ.

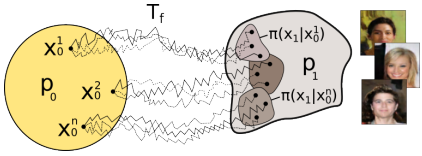
[8] Erwin Schrödinger (1931). *Über die umkehrung der naturgesetze*. Verlag der Akademie der Wissenschaften in Kommission bei Walter De Gruyter u. Company, 1931.

Applications of Schrödinger Bridge

Single-cell biological data.[9]
Solving SB allows to reconstruct the most likely cell trajectories.



Generation and Domain Translation.[10]
Solving SB between noise and data with small ϵ gives diffusion with "straighter" trajectories.



$$T_f : dX_t = f(X_t, t)dt + \sqrt{\epsilon}dW_t, \quad X_0 \sim p_0,$$

[9] Hugo Lavenant et al. (2024). "Toward a mathematical theory of trajectory inference". In: *The Annals of Applied Probability* 34.1A, pp. 428–500. DOI: 10.1214/23-AAP1969.

[10] Valentin De Bortoli et al. (2021). "Diffusion schrödinger bridge with applications to score-based generative modeling". In: *Advances in Neural Information Processing Systems* 34, pp. 17695–17709.

Part I. Light Schrödinger Bridge (ICLR 2024)

Light SB outline.

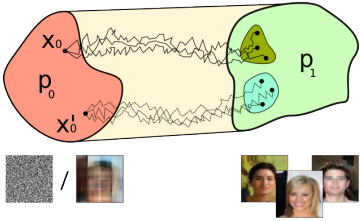
1. Motivation of light SB solvers.
2. Equivalence of SB and EOT problems.
3. Characterisation of SB and EOT solutions.
4. Derivation of the LightSB functional.
5. Gaussian mixture parameterization of Schrödinger Bridges.
6. LightSB training and inference.
7. Experimental Illustrations.

Motivation of light SB solvers

Expectation.

We solve the Scrhödinger Bridge, and it

- maps arbitrary distribution p_0 to arbitrary distribution p_1.



- provides a diffusion that is maximally similar to a given prior process.

Reality.

It is hard to solve the Schrödinger Bridge.

- Many neural-network-based algorithms, almost all of which are poorly scalable and require painful iterative or adversarial learning.
- Absence of simple baseline algorithm, which works fast, provably solves Schrödinger Bridge in moderate dimensions and does not require time-consuming hyperparameter selection.

With this in mind, we started to search for possible solutions.

List of key existing not light SB solvers

See the following **benchmark paper** for a survey of the field in 2023:

- Nikita Gushchin, Alexander Kolesov, Petr Mokrov, et al. (2023). "Building the Bridge of Schr\" odinger: A Continuous Entropic Optimal Transport Benchmark". In: *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. URL: https://openreview.net/forum?id=0HimIaixXk

A (not comprehensive) list of related works is as follows:

1. **MLE-SB:** Francisco Vargas et al. (2021). "Solving schrödinger bridges via maximum likelihood". In: *Entropy* 23.9, p. 1134
2. **DSB**: Valentin De Bortoli et al. (2021). "Diffusion schrödinger bridge with applications to score-based generative modeling". In: *Advances in Neural Information Processing Systems* 34, pp. 17695–17709
3. **ENOT**: Nikita Gushchin, Alexander Kolesov, Alexander Korotin, et al. (2024). "Entropic neural optimal transport via diffusion processes". In: *Advances in Neural Information Processing Systems* 36
4. **FB-SDE:** Tianrong Chen, Guan-Horng Liu, and Evangelos Theodorou (2022). "Likelihood Training of Schrödinger Bridge using Forward-Backward SDEs Theory". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=nioAdKCEdXB
5. **DSBM:** Yuyang Shi et al. (2023). "Diffusion Schrödinger Bridge Matching". In: *Thirty-seventh Conference on Neural Information Processing Systems*. URL: https://openreview.net/forum?id=qy07OHsJT5
6. **ASBM:** Nikita Gushchin, Daniil Selikhanovych, et al. (2024). "Adversarial Schrödinger Bridge Matching". In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. URL: https://openreview.net/forum?id=L3Knnigicu

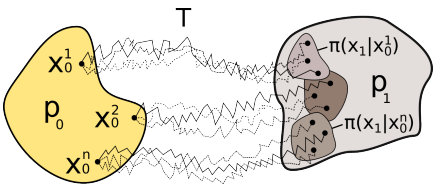
Schrodinger Bridge formulation[11]

The Schrödinger Bridge problem

For two continuous distributions p_0 and p_1 on $\mathbb{R}^D$, the Schrödinger bridge problem is:

$$\inf_{T \in \mathcal{F}(p_0, p_1)} \mathrm{KL}(T \| W^\epsilon).$$

Here $\mathcal{F}(p_0, p_1)$ are stochastic processes with marginals p_0, p_1 at $t = 0$ and $t = 1$.



Here W^ϵ wiener process with the variance ϵ, i.e., it is a stochastic process with the stochastic differential equation (SDE): $dX_t = \sqrt{\epsilon} dW_t$.

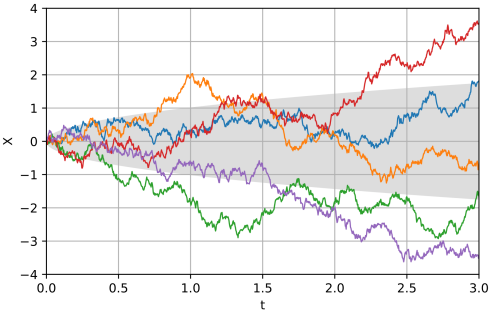


Figure 1: Wiener process with $\epsilon = 1$.

[11] Yongxin Chen, Tryphon T Georgiou, and Michele Pavon (2016). "On the relation between optimal transport and Schrödinger bridges: A stochastic control viewpoint". In: *Journal of Optimization Theory and Applications* 169, pp. 671–691.

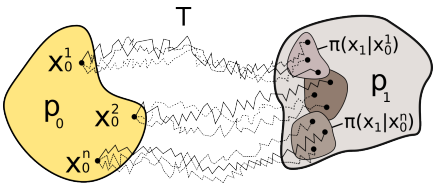
Decomposition of SB on inner and outer parts

Schrödinger Bridge formulation.

For two continuous distributions p_0 and p_1 on $\mathbb{R}^D$, the Schrödinger bridge problem is:

$$\inf_{T \in \mathcal{F}(p_0, p_1)} \mathrm{KL}(T \| W^\epsilon).$$

Here $\mathcal{F}(p_0, p_1)$ are stochastic processes with marginals p_0, p_1 at $t = 0$ and $t = 1$. W^ϵ is a Wiener process with the variance ϵ.



Let π^T denote the joint distribution of a stochastic process T at time moments $t = 0, 1$.

Let $T_{|x,y}$ denote the stochastic processes T conditioned on values x, y at times $t = 0, 1$, respectively.

We can expand the functional as follows:

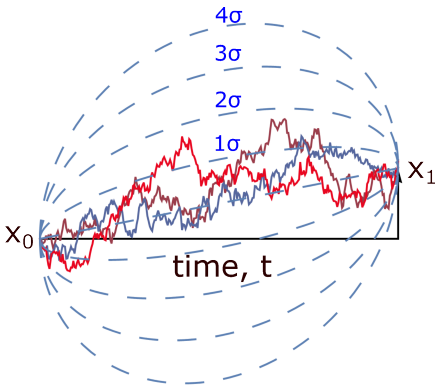
$$\mathrm{KL}(T||W^\epsilon) = \underbrace{\mathrm{KL}(\pi^T||\pi^{W^\epsilon})}_{\text{outer part}} + \underbrace{\int \mathrm{KL}(T_{|x_0,x_1}||W^\epsilon_{|x_0,x_1}) d\pi^T(x_0, x_1)}_{\text{inner part}}.$$

Here $W^\epsilon_{|x_0,x_1}$ is a Wiener processconditioned on its end and start points. It is known as the **Brownian Bridge**.

Schrödinger Bridge is a reciprocal process

Brownian Bridges

The process $W^{\epsilon}_{|x_0,x_1}$ is a Brownian Bridge. It is a Gaussian process starting at x_0 and ending at x_1.



We can set to zero the inner part by searching process in the form of a mixture of Brownian Bridges, i.e. $T = \int W^{\epsilon}_{|x_0,x_1} d\pi^T(x_0, x_1)$.

Such processes form reciprocal class, and for brevity, we just call them reciprocal processes.

In this case:

$$\mathrm{KL}(T||W^{\epsilon}) = \underbrace{\mathrm{KL}(\pi^T||\pi^{W^{\epsilon}})}_{\text{outer part}} + \underbrace{\int \mathrm{KL}(T_{|x_0,x_1}||W^{\epsilon}_{|x_0,x_1}) d\pi^T(x_0, x_1)}_{=0, \text{since } T_{|x_0,x_1} = W^{\epsilon}_{|x_0,x_1}}.$$

Equivalence between EOT and SB

For a reciprocal T, the objective is

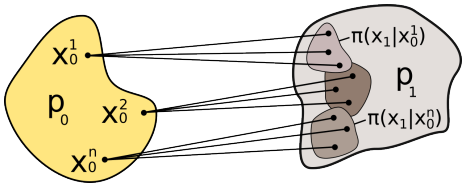
$$\mathrm{KL}(T||W^\epsilon) = \underbrace{\mathrm{KL}(\pi^T||\pi^{W^\epsilon})}_{\text{outer part}}.$$

Hence,

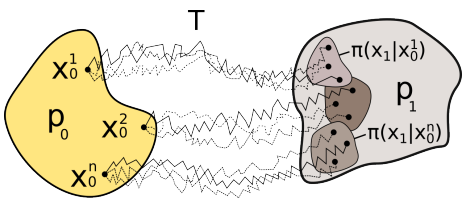
$$\inf_{T\in\mathcal{F}(p_0,p_1)} \mathrm{KL}(T||W^\epsilon) = \inf_{T\in\mathcal{F}(p_0,p_1)} \mathrm{KL}(\pi^T||\pi^{W^\epsilon}).$$

By expanding the outer part, we obtain:

$$\underbrace{\mathrm{KL}(\pi^T||\pi^{W^\epsilon}) = \int_{\mathcal{X}\times\mathcal{Y}} \frac{||x-y||^2}{2\epsilon} d\pi^T(x,y) - H(\pi^T) + C}_{\text{equvivalent to EOT functional.}}.$$



Entropic OT.



Schrödinger Bridge.

Characterization for EOT and SB solutions.

Moreover, both solutions for EOT and SB problems are characterized by the starting distribution p_0 and one scalar-valued function v^*.

EOT solution.

The EOT solution π^* can be represented through the input density p_0 and a function $v^* : \mathbb{R}^D \to \mathbb{R}_+$:

$$\pi^*(x_0, x_1) = \underbrace{p_0(x_0)}_{=\pi^*(x_0)} \cdot \underbrace{\exp\big(\langle x_0, x_1\rangle/\epsilon\big) v^*(x_1) c_{v^*}(x_0)}_{=\pi^*(x_1|x_0)},$$

where $c_{v^*}(x_0) = \int_{\mathbb{R}^D} \exp\big(\langle x_0, x_1\rangle \epsilon\big) v^*(x_1) dy$.

Here v^* is the **adjusted Schrödinger potential.**

SB solution.

The solution T^* for the Schrödinger Bridge is a Markovian process given by the following SDE:

$$dX_t = g^*(X_t, t)dt + \sqrt{\epsilon} dW_t^{\epsilon}, \quad X_0 \sim p_0$$

In turn, the optimal drift $g^*(x_t, t)$ is given by:

$$g^*(x_t, t) = \epsilon \nabla_{x_t} \log \Big(\int_{\mathbb{R}^D} \mathcal{N}(x'|x_t, (1-t)\epsilon I_D) \exp\big(\frac{\|x'\|^2}{2\epsilon}\big) v^*(x') dx' \Big),$$

i.e., is a convolution with the adjusted potential v^*.

Theoretical summary.

Equivalence of EOT and SB problems.

We can solve SB by solving the related EOT problem since:

$$\inf_{T \in \mathcal{F}(p_0, p_1)} \mathrm{KL}(T || W^\epsilon) = \inf_{T \in \mathcal{F}(p_0, p_1)} \mathrm{KL}(\pi^T || \pi^{W^\epsilon}) = \inf_{\pi \in \Pi(p_0, p_1)} \mathrm{KL}(\pi || \pi^{W^\epsilon}),$$

where $\Pi(p_0, p_1)$ is a set of joint distributions on $t = 0$ and $t = 1$ with marginals p_0 and p_1.

Optimal form of the solution.

$$\pi^*(x_0, x_1) = \underbrace{p_0(x_0)}_{=\pi^*(x_0)} \cdot \underbrace{\exp\left(\langle x_0, x_1 \rangle / \epsilon\right) v^*(x_1) c_{v^*}(x_0)}_{=\pi^*(x_1 | x_0)},$$

Still not obvious how to solve. The EOT problem:

$$\inf_{\pi \in \Pi(p_0, p_1)} \mathrm{KL}(\pi || \pi^{W^\epsilon}) = \inf_{\pi \in \Pi(p_0, p_1)} \int_{\mathbb{R}^D \times \mathbb{R}^D} \frac{||x - y||^2}{2\epsilon} d\pi^T(x, y) - H(\pi^T) + C.$$

is a constrained optimization problem, and we do not know how to parametrize a set $\Pi(p_0, p_1)$.

Direct optimization of KL with the solution.

Our new objective:

Instead of trying to solve the constrained optimization problem of EOT, let's just minimize KL with the solution π^*:

$$\underbrace{\underset{\pi \in \Pi(p_0, p_1)}{\arg\min}\ \mathrm{KL}(\pi || \pi^{W^\epsilon})}_{\text{constrained optimization}} \rightarrow \underbrace{\underset{\pi}{\arg\min}\, \mathrm{KL}(\pi^* || \pi)}_{\text{unconstrained optimization}}$$

The problem: we do not know π^*.

Our proposed optimal form parametrization:

It is possible with a proper parametrization of π_θ.

$$\pi_\theta(x_0, x_1) = p_0(x_0)\pi_\theta(x_1|x_0) = p_0(x_0)\frac{\exp\left(\langle x_0, x_1\rangle \epsilon\right) v_\theta(x_1)}{c_\theta(x_0)}.$$

We parameterize v^* as v_θ. In turn, $c_\theta(x_0) = \int_{\mathbb{R}^D} \exp\left(\langle x_0, x_1\rangle / \epsilon\right) v_\theta(x_1) dx_1$ is the normalization.

Deriving the Learning Objective

Magic of KL-divergence.

$$\mathrm{KL}(\pi^*||\pi_\theta) = \int_{\mathbb{R}^D\times\mathbb{R}^D} \pi^*(x_0,x_1)\log\frac{\pi^*(x_0,x_1)}{\pi_\theta(x_0,x_1)}dx_0dx_1 =$$

$$\mathrm{C} - \int_{\mathbb{R}^D\times\mathbb{R}^D} \pi^*(x_0,x_1)\log\underbrace{\frac{\exp\left(\langle x_0,x_1\rangle/\epsilon\right)v_\theta(x_1)}{c_\theta(x_0)}}_{\pi_\theta(x_1|x_0)}dx_0dx_1 = \mathrm{C} - \underbrace{\int_{\mathbb{R}^D\times\mathbb{R}^D}\pi^*(x_0,x_1)\left(\langle x_0,x_1\rangle/\epsilon\right)dx_0dx_1}_{\text{also constant}} +$$

$$\underbrace{\int_{\mathbb{R}^D\times\mathbb{R}^D}\pi^*(x_0,x_1)\log c_\theta(x_0)dx_0dx_1}_{\text{expectation of a function of } x_0} - \underbrace{\int_{\mathbb{R}^D\times\mathbb{R}^D}\pi^*(x_0,x_1)\log v_\theta(x_1)dx_0dx_1}_{\text{expectation of a function of } x_1} =$$

$$\widetilde{C} + \underbrace{\int_{\mathbb{R}^D} p_0(x_0)\log c_\theta(x_0)dx_0 - \int_{\mathbb{R}^D} p_1(x_1)\log v_\theta(x_1)dx_1}_{=\mathcal{L}(\theta)} = \mathrm{Const} + \mathcal{L}(\theta).$$

We can estimate $\mathrm{KL}(\pi^*||\pi_\theta)$ up to a constant, which depends only on π^*. Hence, we can directly optimize $\mathrm{KL}(\pi^*||\pi_\theta)$ knowing nothing about π^* except its marginals p_0 and p_1.

Gaussian parametrization

The functional for optimization.

$$\min_{\theta} \mathrm{KL}(\pi^* || \pi_\theta) - \mathrm{C} = \min_{\theta} \mathcal{L}(\theta) = \min_{\theta} \int_{\mathbb{R}^D} p_0(x_0) \log c_\theta(x_0) dx_0 - \int_{\mathbb{R}^D} p_1(x_1) \log v_\theta(x_1) dx_1.$$

The problem: it is hard to compute normalization constant $c_\theta(x_0)$ for arbitrary potential v_θ.

Gaussian parametrization of adjusted Schrödinger potential.

We recall that:

$$\pi_\theta(x_1|x_0) = \frac{\exp\left(\langle x_0, x_1 \rangle / \epsilon\right) v_\theta(x_1)}{c_\theta(x_0)},$$

For $x = 0$, we have $\pi_\theta(x_1|0) = \frac{v_\theta(x_1)}{c_\theta(x_0)}$, i.e. $v_\theta(x_1)$ is an unnormalized density.

$\Longrightarrow$ Let us approximate v_θ by a Gaussian mixture:

$$v_\theta(x_1) \stackrel{\text{def}}{=} \sum_{k=1}^{K} \alpha_k \mathcal{N}(x_1|r_k, S_k),$$

where $\theta \stackrel{\text{def}}{=} \{\alpha_k, r_k, S_k\}_{k=1}^K$ are the parameters: $\alpha_k \geq 0$, $r_k \in \mathbb{R}^D$ and symmetric $0 \prec S_k \in \mathbb{R}^{D \times D}$.

Gaussian parametrization

Conditional distribution for the Gaussian mixture parametrization.

For a Gaussian mixture approximation $v_\theta(x_1) \stackrel{\text{def}}{=} \sum_{k=1}^K \alpha_k \mathcal{N}(x_1|r_k, S_k)$, it holds that

$$\pi_\theta(x_1|x_0) = \frac{1}{c_\theta(x_0)} \sum_{k=1}^K \widetilde{\alpha}_k(x_0)\mathcal{N}(x_1|r_k(x_0), \epsilon S_k) \qquad \text{where} \qquad r_k(x_0) \stackrel{\text{def}}{=} r_k + S_k x_0,$$

$$\widetilde{\alpha}_k(x_0) \stackrel{\text{def}}{=} \alpha_k \exp\left(\frac{x_0^T S_k x_0 + 2 r_k^T x_0}{2\epsilon}\right), \qquad c_\theta(x_0) \stackrel{\text{def}}{=} \sum_{k=1}^K \widetilde{\alpha}_k(x_0).$$

The functional for optimization.

$$\min_\theta \text{KL}(\pi^* || \pi_\theta) - \text{C} = \min_\theta \mathcal{L}(\theta) = \min_\theta \int_{\mathbb{R}^D} p_0(x_0) \log c_\theta(x_0) dx_0 - \int_{\mathbb{R}^D} p_1(x_1) \log v_\theta(x_1) dx_1.$$

With such parametrization, we can easily estimate and optimize our objective.

Training

The functional for optimization:

$$\min_{\theta} \mathcal{L}(\theta) = \min_{\theta} \int_{\mathbb{R}^D} p_0(x_0) \log c_\theta(x_0) dx_0 - \int_{\mathbb{R}^D} p_1(x_1) \log v_\theta(x_1) dx_1.$$

The empirical functional for optimization.

As the distributions p_0, p_1 are accessible only via samples $X^0 = \{x_0^1, \dots, x_0^N\} \sim p_0$ and $X^1 = \{x_1^1, \dots, x_1^M\} \sim p_1$, we optimize the empirical counterpart of $\mathcal{L}(\theta)$:

$$\widehat{\mathcal{L}}(\theta) \stackrel{\text{def}}{=} \frac{1}{N} \sum_{n=1}^{N} \log c_\theta(x_0^n) - \frac{1}{M} \sum_{m=1}^{M} \log v_\theta(x_1^m) \approx \mathcal{L}(\theta).$$

We use the (minibatch) gradient descent w.r.t. parameters θ.

EOT-based inference

Sampling starting and ending points.
The conditional distributions $\pi_\theta(x_1|x_0)$ are mixtures of Gaussians:

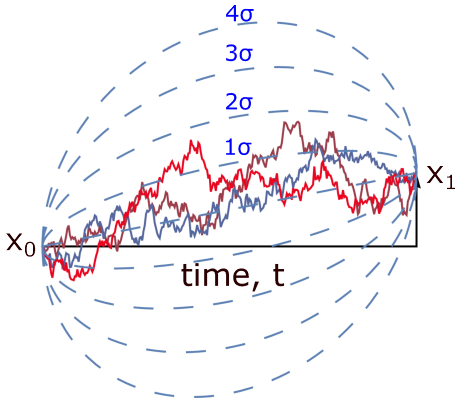
$$\pi_\theta(x_1|x_0) = \frac{1}{c_\theta(x_0)} \sum_{k=1}^{K} \widetilde{\alpha}_k(x_0)\mathcal{N}(x_1|r_k(x_0), \epsilon S_k)$$

Sampling of the pair (x_0, x_1) is straightforward and lightspeed.

Inner trajectory sampling.

To sample trajectory $x_0, x_{t_1}, \ldots, x_{t_L}, x_1$ with $0 < t_1 < \cdots < t_L < 1$ it is enough to sample from the Brownian Bridge $W^{\epsilon}_{|x_0,x_1}$.

Brownian Bridge.
The process $W^{\epsilon}_{|x_0,x_1}$ is a Brownian Bridge. It is a Gaussian process starting at x_0 and ending at x_1.



SDE form of the learned process.

The process T_θ given by the potential v_θ is a diffusion process governed by the following SDE:

$$T_\theta \quad : \quad dX_t = g_\theta(X_t, t)dt + \sqrt{\epsilon}dW_t, \qquad X_0 \sim p_0,$$

$$g_\theta(x,t) \stackrel{\text{def}}{=} \epsilon\nabla_x \log\left(\mathcal{N}(x|0, \epsilon(1-t)I_D)\sum_{k=1}^{K}\left\{\alpha_k \mathcal{N}(r_k|0, \epsilon S_k)\mathcal{N}(h(x,t)|0, A_k^t)\right\}\right),$$

with $A_k^t \stackrel{\text{def}}{=} \frac{t}{\epsilon(1-t)}I_D + \frac{S_k^{-1}}{\epsilon}$ and $h_k(x,t) \stackrel{\text{def}}{=} \frac{1}{\epsilon(1-t)}x + \frac{1}{\epsilon}S_k^{-1}r_k$.

- Any SDE solver can be applied to the sample from this SDE, e.g. Euler-Maruyama.
- EOT-based sampling is always better since it is the analytical solution of this SDE.

Summary

We developed a blazing-fast method for solving the Schrödinger Bridge problem.

The method is based on:

1. New loss function for training the Schrödinger bridge:

$$\mathcal{L}(\theta) = \int_{\mathbb{R}^D} \log c_\theta(x_0) p_0(x_0) dx_0 - \int_{\mathbb{R}^D} \log v_\theta(x_1) p_1(x_1) dx_1, \quad c_\theta(x_0) = \int_{\mathbb{R}^D} \exp\left(\langle x_0, x_1 \rangle / \epsilon\right) v_\theta(x_1) dx_1,$$

where v_θ is an adjusted Schrödinger potential which completely defines the entire Schrödinger Bridge T_θ.

2. Optimal parameterization of the Schrödinger bridge using mixtures of Gaussians:

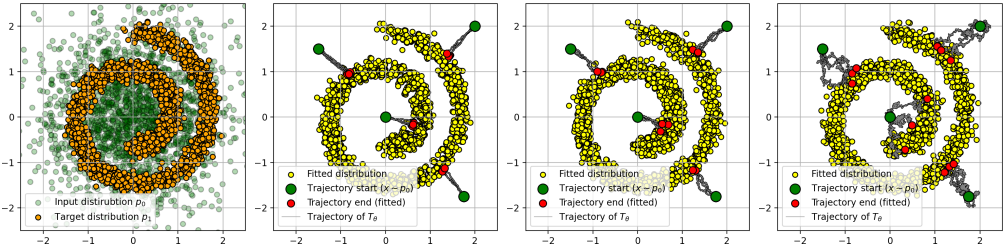
$$v_\theta(x_1) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(x_1 | r_k, S_k), \quad c_\theta(x_0) = \sum_{k=1}^{K} \alpha_k \exp\left(\frac{x_0^T S_k x_0 + 2 r_k^T x_0}{2\epsilon}\right).$$

Our method's advantages:

- Fast training ($<$ 1 minute on 4 CPU cores, not hours of training on GPU, like others).
- Theoretical validity (in this work we prove the guarantees of the method's learning ability from the point of view of statistical learning theory and approximation theory).

Experimental results

1. **Qualitative** results of our algorithm applied to 2D model distributions ("Gaussian" $\rightarrow$ "swiss-roll").



(a) $x \sim p_0$, $y \sim p_1$. **(b)** $\epsilon = 2 \cdot 10^{-3}$. **(c)** $\epsilon = 0.01$. **(d)** $\epsilon = 0.1$.

2. **Quantitative** results of our solver on the standard benchmark for the Schrödinger bridge problem.

Best from the existing methods

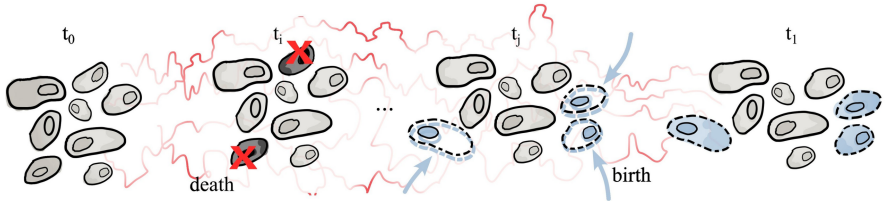
Our method

	$\epsilon = 0.1$				$\epsilon = 1$				$\epsilon = 10$			
	$D=2$	$D=16$	$D=64$	$D=128$	$D=2$	$D=16$	$D=64$	$D=128$	$D=2$	$D=16$	$D=64$	$D=128$
Best solver	1.94	13.67	11.74	11.4	1.04	9.08	18.05	15.23	1.40	1.27	2.36	1.31
LightSB	**0.03**	**0.08**	**0.28**	**0.60**	**0.05**	**0.09**	**0.24**	**0.62**	**0.07**	**0.11**	**0.21**	**0.37**
$\pm std$	± 0.01	± 0.04	± 0.02	± 0.02	± 0.003	± 0.006	± 0.007	± 0.007	± 0.02	± 0.01	± 0.01	± 0.01

*The metric cBW-UVP is used for comparising build schrödinger bridge with ground-truth bridge (lower=better).

Experiments with Single-cell data[12]

3. **Quantitative** results in the problem of predicting single-cell trajectories in the feature space (single-cell trajectory inference).



Our method is superior to analogues in quality of work and speed.

Setup	Solver type	Solver \ DIM	50	100	1000
Discrete EOT	Sinkhorn	(Cuturi, 2013) [1 GPU V100]	2.34 (90 s)	2.24 (2.5 m)	1.864 (9 m)
Continuous EOT	Langevin-based	(Mokrov et al., 2023) [1 GPU V100]	2.39 ± 0.06 (19 m)	2.32 ± 0.15 (19 m)	1.46 ± 0.20 (15 m)
Continuous EOT	Minimax	(Gushchin et al., 2023) [1 GPU V100]	2.44 ± 0.13 (43 m)	2.24 ± 0.13 (45 m)	1.32 ± 0.06 (71 m)
Continuous EOT	IPF	(Vargas et al., 2021) [1 GPU V100]	3.14 ± 0.27 (8 m)	2.86 ± 0.26 (8 m)	2.05 ± 0.19 (11 m)
Continuous EOT	KL minimization	LightSB **(ours)** [4 CPU cores]	2.31 ± 0.27 (65 s)	2.16 ± 0.26 (66 s)	1.27 ± 0.19 (146 s)

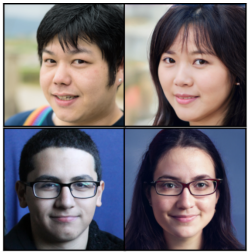
**The Energy distance metric is used to compare the predicted cell position and the observed one (smaller=better).
The operating time of the method in question is indicated in parentheses. 50, 100, 1000 - dimension of the feature space.

[12] Alexander Y Tong et al. (2024). "Simulation-Free Schrödinger Bridges via Score and Flow Matching". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 1279–1287.

Unpaired Image translation in latent space

4. **Qualitative** results of the method for solving the **unpaired** domain translation problem (in the latent space of the ALAE autoencoder[13]).

The latent space size is 512. Images resolution is 1024x1024.



(a) Male → Female.

(b) Female → Male.

(c) Adult → Child.

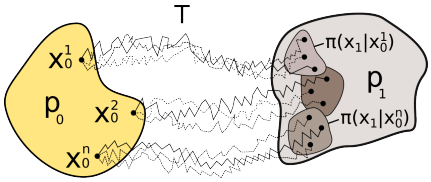
(d) Child → Adult.

[13]Stanislav Pidhorskyi, Donald A Adjeroh, and Gianfranco Doretto (2020). "Adversarial latent autoencoders". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14104–14113.

Thank you

Light Schrödinger Bridge (ICLR 2024)

The novel light and fast algorithm
to solve the Schrödinger Bridge problem.



GitHub

`https://github.com/ngushchin/LightSB`

Part II. Light and Optimal Schrödinger Bridge Matching (ICML 2024)

Reciprocal and Markovian processes

Let $\mathcal{F}$ denote the set of all stochastic processes in $\mathbb{R}^D$ for time interval $[0,1]$ with continuous trajectories $\{x_t\}_{t\in[0,1]}$. Recall that we already use $\mathcal{F}(p_0,p_1)\subset\mathcal{F}$ to denote its subset of processes whose marginals at times $t=0,1$ are p_0 and p_1, respectively.

Reciprocal processes Let $\mathcal{R}\subset\mathcal{F}$ denote the subset of **reciprocal** processes, i.e., those processes can be represented as mixtures of Brownian bridges:

$$T\in\mathcal{R} \qquad \Leftrightarrow \qquad \exists \pi=\pi^T\in\mathcal{P}(\mathbb{R}^D\times\mathbb{R}^D) \text{ s.t. } T=T_\pi \stackrel{\text{def}}{=} \int W^{\epsilon}_{|x_0,x_1}\pi(x_0,x_1)dx_0dx_1.$$

We use $\mathcal{R}(p_0,p_1)$ to denote its subset of processes which satisfy $\pi^T\in\Pi(p_0,p_1)$.

Markov Processes Let $\mathcal{M}\subset\mathcal{F}$ denote the subset of **Markovian** processes, i.e.,

$$T\in\mathcal{M} \qquad \Leftrightarrow \qquad \forall N>1,\ 0\le t_1<\cdots<t_N\le 1:\ p^T(x_{t_N}|x_{t_{N-1}}\ldots,x_1)=p^T(x_{t_N}|x_{t_{N-1}}).$$

In turn, let $\mathcal{M}(p_0,p_1)$ denote its subset of processes which satisfy $\pi^T\in\Pi(p_0,p_1)$.

Schrödinger Bridge is both Markovian and Reciprocal Process

In fact, we already know that $T^* \in \mathcal{M}(p_0, p_1) \cap \mathcal{R}(p_0, p_1)$. Indeed,

- We already derived that SB T^* is a **mixture of Brownian Bridges** $W^{\epsilon}_{|x_0,x_1}$:

$$T^* = \int W^{\epsilon}_{|x_0,x_1} \pi^*(x_0, x_1) dx_0 dx_1 \in \mathcal{M} \cap \mathcal{R},$$

  where $\pi^*(x_0, x_1)$ is the EOT plan. Therefore, $T^* \in \mathcal{R}(p_0, p_1) \subset \mathcal{R}$.

- We have already seen that the solution T^* is a **diffusion** process:

$$dX_t = g^*(X_t, t)dt + \sqrt{\epsilon} dW^{\epsilon}_t, \quad X_0 \sim p_0$$

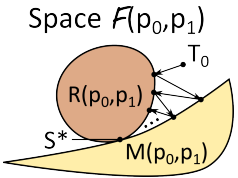
  for some drift g^*. Therefore, T^* is Markovian, i.e., $T^* \in \mathcal{M}(p_0, p_1) \subset \mathcal{M}$.

Iterative Projections onto the Sets of Interest

The Schrödinger Bridge has an **awesome property**[14]: it is the *unique* process (starting at p_0 and ending at p_1) that satisfies both the *markovian* and *reciprocal* property, i.e.,

$$\{T^*\} = \mathcal{M}(p_0, p_1) \cap \mathcal{R}(p_0, p_1).$$

Idea: why not to try to find the process that is both markovian and reciprocal by using some sort of **projections** onto Reciprocal $\mathcal{R}(p_0, p_1)$ and Markovian $\mathcal{M}(p_0, p_1)$ sets of processes?



Note: The subset $R \subset \mathcal{F}$ is convex, while $\mathcal{M} \subset \mathcal{F}$ is, in general, not convex. The latter statement is not obvious and is a good excersize to think about.

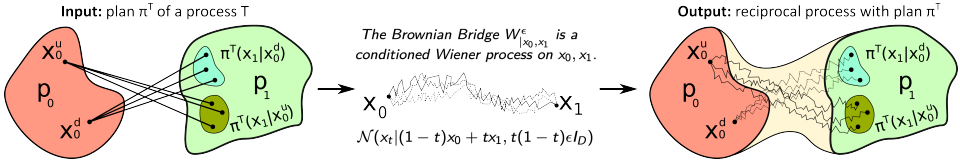
[14] Christian Léonard (2014). "A survey of the Schrödinger problem and some of its connections with optimal transport". In: *Discrete & Continuous Dynamical Systems-A* 34.4, pp. 1533–1574.

Reciprocal Projection

The projection is defined for every $T \in \mathcal{F}$ as follows:

$$\text{proj}_{\mathcal{R}}(T) \stackrel{\text{def}}{=} \text{argmin}_{R \in \mathcal{R}} \text{KL}(T \| R).$$

One may easily prove that the reciprocal projection creates a **mixture** of Brownian Bridges $W^{\epsilon}_{|x_0,x_1}$ with the distribution π^T of a stochastic process $T \in \mathcal{F}$ at times $t = 0, 1$, i.e.,



$$\text{proj}_{\mathcal{R}}(T) = \int W^{\epsilon}_{|x_0,x_1} \pi^T(x_0, x_1) dx_0 dx_1.$$

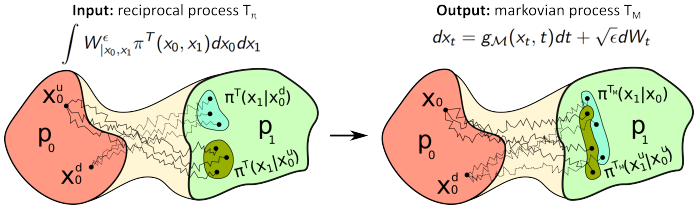
The projection depends only on π^T (transport plan) rather than on the entire process T. Furthermore, $\pi^{\text{proj}_{\mathcal{R}}(T)} = \pi^T$, i.e., this *transport plan is preserved* during the projection.

Markovian Projection

The projection is defined for *reciprical* processes $T = T_\pi \in \mathcal{R}$ as follows:

$$\text{proj}_{\mathcal{M}}(T) \stackrel{\text{def}}{=} \text{argmin}_{M \in \mathcal{M}} \text{KL}(T \| M).$$

It finds the **diffusion** process $T_{\mathcal{M}}$ which is the most similar to T_π:



The drift of the Markovian projection is: $g_{\mathcal{M}} \stackrel{\text{def}}{=} \arg\min_g \int_0^1 \mathbb{E}_{(x_t,x_1)\sim T_\pi} ||g(x_t,t) - \frac{x_1 - x_t}{1-t}||^2 dt$.

The markovian projections *preserves the marginals* of the process at every time t (including $t = 0, 1$), but *alters the transport plan*, i.e., $\pi^T \neq \pi^{T_{\mathcal{M}}}$ (unless T is the Schrodinger Bridge).

Reciprocal and Markovian projections

Reciprocal projection

- Defined for any process $T \in \mathcal{F}$:

$$\text{proj}_{\mathcal{R}}(T) \stackrel{\text{def}}{=} \text{argmin}_{R \in \mathcal{R}} \text{KL}(T \| R)$$

- Yields a mixture of Brownian Bridges:

$$\int W^{\epsilon}_{|x_0,x_1} \pi^T(x_0, x_1) dx_0 dx_1$$

Markovian projection

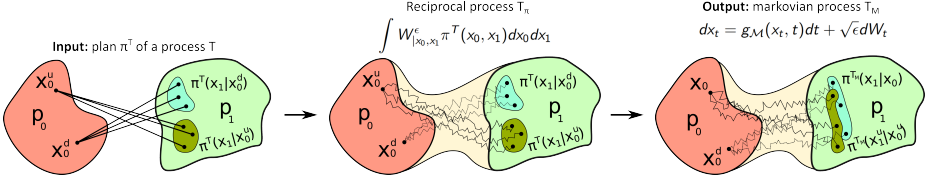
- Defined for a *reciprocal* process $T_\pi \in \mathcal{R}$:

$$\text{proj}_{\mathcal{M}}(T_\pi) \stackrel{\text{def}}{=} \text{argmin}_{M \in \mathcal{M}} \text{KL}(T_\pi \| M)$$

- Yields a **diffusion** with the SDE

$$dx_t = g_{\mathcal{M}}(x_t, t)dt + \sqrt{\epsilon} dW_t, \qquad x_0 \sim p_0.$$

Bridge matching = combination of Reciprocal and Markovian Projections



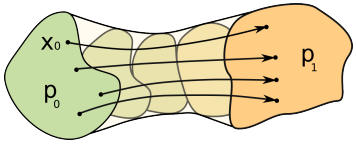
It is a popular way to learn diffusion processes between data distributions p_0, p_1.

Flow Matching and Bridge Matching: a Reminder

Flow matching is a limiting case of the Bridge Matching when $\epsilon \to 0$.

Flow Matching

$$x_t = g(x_t, t)t$$



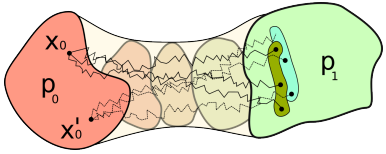
Define interpolation: $x_t \stackrel{\text{def}}{=} x_0 \cdot (1-t) + x_1 \cdot t.$

$$\min_g \mathop{\mathbb{E}}_{\substack{x_0 \sim p_0 \\ x_1 \sim p_1}} \mathop{\mathbb{E}}_{t \sim [0,1]} \Big\| g(x_t, t) - (x_1 - x_0) \Big\|^2.$$

Can be iterated to straighten the flows.
Related to the **Optimal Transport** (OT).

Bridge Matching

$$x_t = g(x_t, t)t + \sqrt{\epsilon} W_t \quad (\epsilon > 0).$$



Define a distribution: $p_t^\epsilon \stackrel{\text{def}}{=} \mathcal{N}\big(x_t, \epsilon t(1-t)\big)$

$$\min_g \mathop{\mathbb{E}}_{\substack{x_0 \sim p_0 \\ x_1 \sim p_1}} \mathop{\mathbb{E}}_{t \sim [0,1]} \mathop{\mathbb{E}}_{\tilde{x}_t \sim p_t^\epsilon} \Big\| g(\tilde{x}_t, t) - \frac{x_1 - \tilde{x}_t}{1-t} \Big\|^2.$$

Can be iterated and converges to the
Schrödinger bridge.

Iterative Markovian Fitting (IMF)[15][16]

Alternating Markovian and Reciprocal projections is called the **Iterative Markovian Fitting** (IMF) procedure, or, alternatively, **Iterative Diffusion Bridge Matching** (IDBM).

Starting from a reciprocal process $T_0 = \int W^{\epsilon}_{|x_0,x_1} d\pi(x_0, x_1)$ induced by some initial plan $\pi(x_0, x_1)$, one performs iterative updates

$$T^{2n+1} = \text{proj}_{\mathcal{M}}(T^{2n}), \qquad T^{2n+2} = \text{proj}_{\mathcal{R}}(T^{2n+1})$$

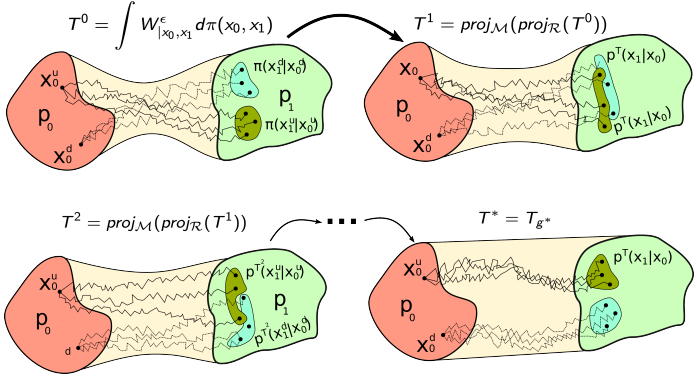
The sequence $\{T^n\}_{n=1}^{\infty}$ converges to the Schrodinger Bridge T^*:

$$\lim_{n\to+\infty} \text{KL}(T^n \| T^*) = 0.$$

[15] Stefano Peluchetti (2023). "Diffusion bridge mixture transports, Schrödinger bridge problems and generative modeling". In: *Journal of Machine Learning Research* 24.374, pp. 1–51.

[16] Yuyang Shi et al. (2023). "Diffusion Schrödinger Bridge Matching". In: *Thirty-seventh Conference on Neural Information Processing Systems*. URL: https://openreview.net/forum?id=qy07OHsJT5.

Iterative Markovian Fitting: An Illustration



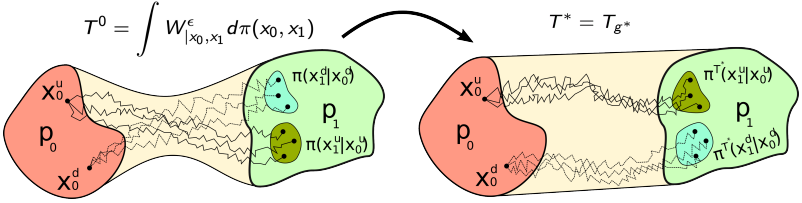
Limitations: The procedure is iterative, i.e., it requires many bridge matching steps.

- Each bridge matching step is a non-trivial drift learning (optimization) procedure.
- Errors in matching the target (p_1) may accumulate during IMF steps.[17]

[17] Rectified flow is a limiting case of the IMF when $\epsilon \to 0$.

Optimal Schrödinger Bridge Matching[18]

While IMF performs iterative Bridge Matching (reciprical and Markovian projections) to recover SB, we propose a novel concept of the **optimal projection**. It Recovers the Schrödinger Bridge T^* is just one iteration of the Bridge Matching.



[18]Nikita Gushchin, Sergei Kholkin, et al. (n.d.). "Light and Optimal Schrödinger Bridge Matching". In: *Forty-first International Conference on Machine Learning*.

Optimality of the "Optimal Projection"

Projection on the set $\mathcal{S}$ of SBs or *"Optimal Projection"* is the foundation of our method.

$$\mathcal{S} \stackrel{\text{def}}{=} \mathcal{R} \cap \mathcal{M}.$$

For **any** reciprocal process T_π with marginals p_0 and p_1, we define the **optimal projection** by

$$\text{proj}_{\mathcal{S}}(T_\pi) = \text{argmin}_{S \in \mathcal{S}} \text{KL}(T_\pi \| S)$$

Theorem (Optimal Projection)

Consider any reciprocal process T_π that has marginals p_0 and p_1 at $t = 0$ and $t = 1$, respectively. Then the Optimal Projection yields the Schrödinger Bridge T^*, i.e.,

$$T^* = proj_{\mathcal{S}}(T_\pi).$$

Looks nice, but how to implement this projection in practice? How to optimize over $S \in \mathcal{S}$?

Characterization for EOT and SB solutions: a Reminder

The solutions for SB problems can be **characterized** by two things:

1. the starting distribution p_0;
2. the scalar-valued function v (potential).

More precisely, the following process (which we denote by S_v)

$$S_v: \qquad dX_t = g_v(X_t, t)dt + \sqrt{\epsilon}dW_t^{\epsilon}, \quad X_0 \sim p_0,$$

$$g_v(x_t, t) \stackrel{\text{def}}{=} \epsilon\nabla_{x_t} \log\Big(\int_{\mathbb{R}^D} \mathcal{N}(x'|x_t, (1-t)\epsilon I_D) \exp\big(\frac{\|x'\|^2}{2\epsilon}\big) v(x')dx'\Big),$$

belongs to $\mathcal{S}(p_0) \subset \mathcal{S}$ and is the Schrodinger bridge between p_0 and its marginal at time $t = 1$. Here $\mathcal{S}(p_0)$ denotes the subset of all Schrodinger Bridges which start at p_0.

Idea: optimize $\arg\min_{S \in \mathcal{S}(p_0)} \mathrm{KL}(T_\pi \| S_v)$ instead of $\arg\min_{S \in \mathcal{S}} \mathrm{KL}(T_\pi \| S)$.[19]

[19] These problems lead to the same solution $T^* \in \mathcal{S}(p_0) \subset \mathcal{S}$.

Tractable Optimization Objective for the Optimal Projection

Optimal projection can be implemented using the *constrained* Bridge matching procedure.

Theorem (Tractable Objective for Optimal Projecton)

Consider set of SBs that start at p_0, i.e., $S_\theta \in \mathcal{S}(p_0)$. Let reciprocal process T_π be a reciprocal process. Then the optimal projection objective satisfies

$$KL(T_\pi \| S_v) = C(\pi) + \underbrace{\frac{1}{2\epsilon}\int_0^1 \mathbb{E}_{(x_t,x_1)\sim T_\pi} \| g_v(x_t,t) - \frac{x_1 - x_t}{1-t}\|^2 dt}_{\text{Bridge Matching}},$$

where

$$g_v(x_t,t) \stackrel{def}{=} \epsilon \nabla_{x_t} \log\Big(\int_{\mathbb{R}^D} \mathcal{N}(x'|x_t,(1-t)\epsilon I_D) \exp\big(\frac{\|x'\|^2}{2\epsilon}\big) v(x') dx'\Big)$$

is the the drift of S_v. Here constant $C(\pi)$ does not depend on S_v.

Nice, but how to compute drift g_v and optimize this objective?

Optimal Drift Computation Problem

For general parameterization of the potential v, e.g., with a neural network v_θ, the computation of the drift $g_v = g_{v_\theta}$ is tricky, so is the computation of the loss. It requires tricky Monte Carlo Markov Chain techniques (MCMC), see the appendices of the paper.[20]

Fortunately, for a Gaussian mixture parameterization (as in **LightSB**) $v_\theta(x_1) \stackrel{\text{def}}{=} \sum_{k=1}^{K} \alpha_k \mathcal{N}(x_1|r_k, S_k)$, the drift g_{v_θ} is available in the closed form

$$g_{v_\theta}(x_t, t) = \epsilon \nabla_x (\mathcal{N}(x|0, \epsilon(1-t)I_D) \sum_{k=1}^{K} \{\alpha_k \mathcal{N}(r_k|0, \epsilon S_k) \mathcal{N}(h_k(x,t)|o, A_k^t)\}.$$

Then we can implement the optimal Projection by optimizing

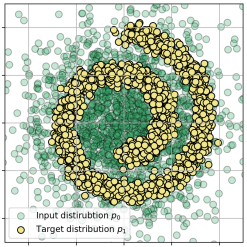
$$\theta^* = \arg\min_\theta \mathrm{KL}(T_\pi \| S_{v_\theta}) = \arg\min_\theta \frac{1}{2\epsilon} \int_0^1 \mathbb{E}_{(x_t, x_1) \sim T_\pi} \| g_\theta(x_t, t) - \frac{x_1 - x_t}{1-t} \|^2 dt.$$

We call the approach by **LightSB-M**. Here **M** stands for **m**atching.

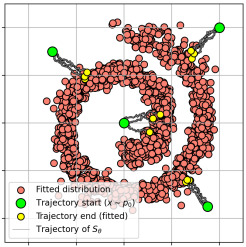
[20] Nikita Gushchin, Sergei Kholkin, et al. (n.d.). "Light and Optimal Schrödinger Bridge Matching". In: *Forty-first International Conference on Machine Learning*.

Qualitative Experiments. 2D Swiss Roll

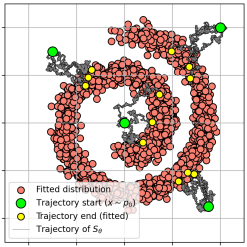
The process $S_\theta = S_{v_\theta}$ learned with LightSB-M in *Gaussian* $\rightarrow$ *Swiss roll* example.



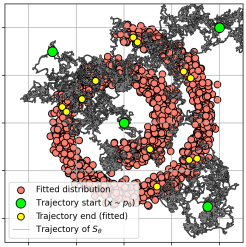
(a) $x \sim p_0$, $y \sim p_1$.



(b) $\epsilon = 0.01$.



(c) $\epsilon = 0.1$.



(d) $\epsilon = 1$.

Experiments. Quantitative SB Benchmark[21]

LightSB-M is the best Bridge Matching method on the SB benchmark. It has comparable performance to LightSB. Also, it yields the same solution for different starting plans $\pi(x_0, x_1)$: independent (**ID**), mini-batch OT (**MB**), ground truth (**GT**).

		$\epsilon = 0.1$				$\epsilon = 1$				$\epsilon = 10$			
	Solver Type	$D=2$	$D=16$	$D=64$	$D=128$	$D=2$	$D=16$	$D=64$	$D=128$	$D=2$	$D=16$	$D=64$	$D=128$
Best solver on SB bench†	Varies	1.94	13.67	11.74	11.4	1.04	9.08	18.05	15.23	1.40	1.27	2.36	1.31
LightSB†	KL minimization	0.03	0.08	0.28	0.60	0.05	0.09	0.24	0.62	0.07	0.11	0.21	0.37
DSBM		5.2	16.8	37.3	35	0.3	1.1	9.7	31	3.7	105	3557	15000
SF^2M-Sink		0.54	3.7	9.5	10.9	0.2	1.1	9	23	0.31	4.9	319	819
LightSB-M (ID, **ours**)	Bridge matching	0.04	0.18	0.77	1.66	0.09	**0.18**	0.47	1.2	**0.12**	0.19	**0.36**	0.71
LightSB-M (MB, **ours**)		**0.02**	**0.1**	0.56	1.32	0.09	**0.18**	**0.46**	**1.2**	0.13	**0.18**	**0.36**	0.71
LightSB-M (GT, **ours**)		**0.02**	**0.1**	**0.49**	**1.16**	**0.09**	**0.18**	0.47	**1.2**	0.13	**0.18**	**0.36**	**0.69**

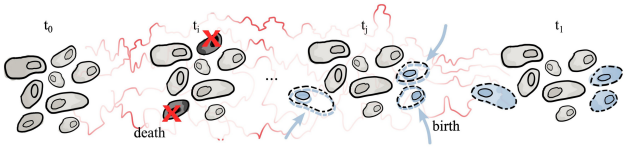
Comparisons of $cB\mathbb{W}_2^2$-UVP $\downarrow$ (%) between the optimal plan π^* and the learned plan π_θ on the EOT/SB benchmark.

The best metric over *bridge matching* solvers is **bolded**. Results marked with † are taken from LightSB paper.

[21] Nikita Gushchin, Alexander Kolesov, Petr Mokrov, et al. (2023). "Building the Bridge of Schr\" odinger: A Continuous Entropic Optimal Transport Benchmark". In: *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. URL: https://openreview.net/forum?id=0HimIaixXk.

Experiments. Quantitative Evaluation on Biological Data

Predicting single-cell trajectories in the feature space.



Solver type	SolverDIM	50	100	1000
Langevin-based	EgNOT† [1 GPU V100]	2.39 $\pm$ 0.06 (19 m)	2.32 $\pm$ 0.15 (19 m)	1.46 $\pm$ 0.20 (15 m)
Minimax	ENOT† [1 GPU V100]	2.44 $\pm$ 0.13 (43 m)	2.24 $\pm$ 0.13 (45 m)	1.32 $\pm$ 0.06 (71 m)
IPF	DSB† [1 GPU V100]	3.14 $\pm$ 0.27 (8 m)	2.86 $\pm$ 0.26 (8 m)	2.05 $\pm$ 0.19 (11 m)
KL minimization	LightSB† [4 CPU cores]	2.31 $\pm$ 0.27 (65 s)	2.16 $\pm$ 0.26 (66 s)	1.27 $\pm$ 0.19 (146 s)
Bridge matching	DSBM [1 GPU V100]	2.46 $\pm$ 0.1 (6.6 m)	2.35 $\pm$ 0.1 (6.6 m)	1.36 $\pm$ 0.04 (8.9 m)
	SF^{2}M-Sink [1 GPU V100]	2.66 $\pm$ 0.18 (8.4 m)	2.52 $\pm$ 0.17 (8.4 m)	1.38 $\pm$ 0.05 (13.8 m)
	LightSB-M (ID, **ours**) [4 CPU cores]	2.347 $\pm$ 0.11 (58 s)	2.174 $\pm$ 0.08 (60 s)	1.35 $\pm$ 0.05 (147 s)
	LightSB-M (MB, **ours**) [4 CPU cores]	**2.33** $\pm$ 0.09 (80 s)	**2.172** $\pm$ 0.08 (80 s)	**1.33** $\pm$ 0.05 (176 s)

Table 1: Energy distance (averaged for two setups and 5 random seeds) on the MSCI dataset

LightSB-M is the best **Bridge Matching** method in thie experiment with Biological data. It provides comparable performance to LightSB that is based on the **KL** minimization principle.

Experiments. Comparison on Unpaired Image-to-image Transfer

Adult to Child Unpaired Translation in the latent space of ALAE[22], 1024x1024 images.

[22]Stanislav Pidhorskyi, Donald A Adjeroh, and Gianfranco Doretto (2020). "Adversarial latent autoencoders". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14104–14113.

Summary

LightSB-M is a method to solve the SB problem in a single Bridge Matching step.
The solver is based on:

- The "Optimal Projection" that translates **any** π with marginals p_0 and p_1 to SB
- Novel Bridge Matching-like optimization objective

$$L_\theta(\pi) = \int_0^1 \mathbb{E}_{(x_t, x_1)\sim T_\pi} \| g_\theta(x_t, t) - \frac{x_1 - x_t}{1-t} \|^2 dt$$

$$g_\theta(x_t, t) = \epsilon \nabla_{x_t} \log \int_{\mathbb{R}^D} \mathcal{N}(x' | x_t, (1-t)\epsilon I_D) \exp\left(\frac{\|x'\|^2}{2\epsilon}\right) v_\theta(x') dx'$$

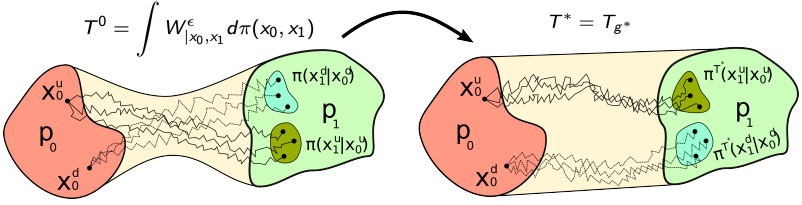
- Parameterization of the SB using mixtures of Gaussians $v_\theta(x) = \sum_{k=1}^K \alpha_k \mathcal{N}(x | r_k, S_k)$. In this case, g_θ admits closed form expression.

LightSB-M's advantages:

- Theoretical novelty (first method solving SB in one Bridge Matching iteration).
- Fast training ($<$ 1 minute on 4 CPU cores, not hours of training on GPU, like others).

Light and Optimal Schrödinger Bridge Matching (ICML 2024)

The novel light and fast algorithm based on the bridge matching
to solve the Schrödinger Bridge problem.



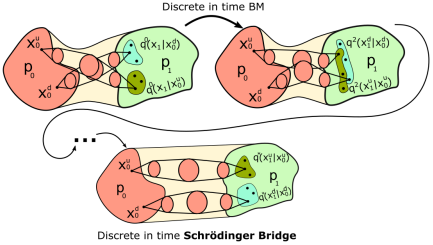
GitHub

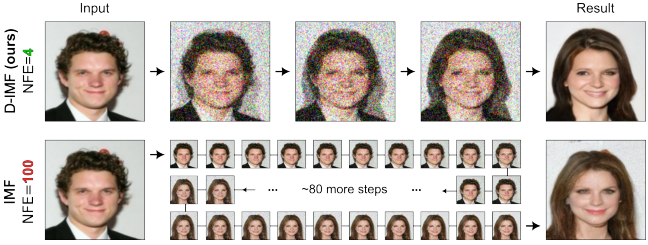
https://github.com/SKholkin/LightSB-Matching

Other works

Adversarial Schrödinger Bridge Matching[23]

We present Discrete in time Bridge Matching and prove that Iterative Discrete in time Bridge Matching (**D-IMF**) converges to discrete in time **Schrödinger Bridge**.



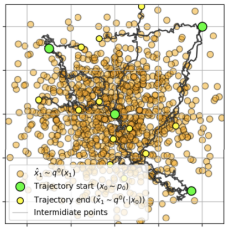


Idea: Substitute the Bridge Matching Diffusion by the **Denoising Diffusion GAN** (DD-GAN). That allows to **speed up** the generation **x25** times while having even better quality.

[23] Nikita Gushchin, Daniil Selikhanovych, et al. (2024). "Adversarial Schrödinger Bridge Matching". In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. URL: https://openreview.net/forum?id=L3Knnigicu.

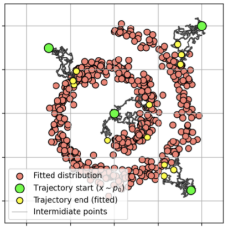
Iterative Proportional Markovian Fitting[24]

Practical implementation of **IMF** algorithm secretly utilizes another popular algorithm **IPF**. We propose Iterative Proportional Markovian Fitting (**IPMF**) algorithm, argue that IMF used in practice and IPF algorithms are a particular cases of IPMF.

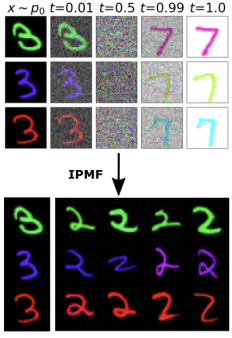


Independent coupling
Gaussian -> Gaussian

IPMF



Schrödinger Bridge
Gaussian -> Swiss Roll



We show empirically and in some cases theoretically that IPMF converges to the Schrödinger Bridge.

[24] Sergei Kholkin et al. (2024). "Diffusion & Adversarial Schr\" odinger Bridges via Iterative Proportional Markovian Fitting". In: *arXiv preprint arXiv:2410.02601*.

3rd conference on machine learning
FALL into ML 2024

Exploring the Potential of AI in Ligand-Based Drug Design

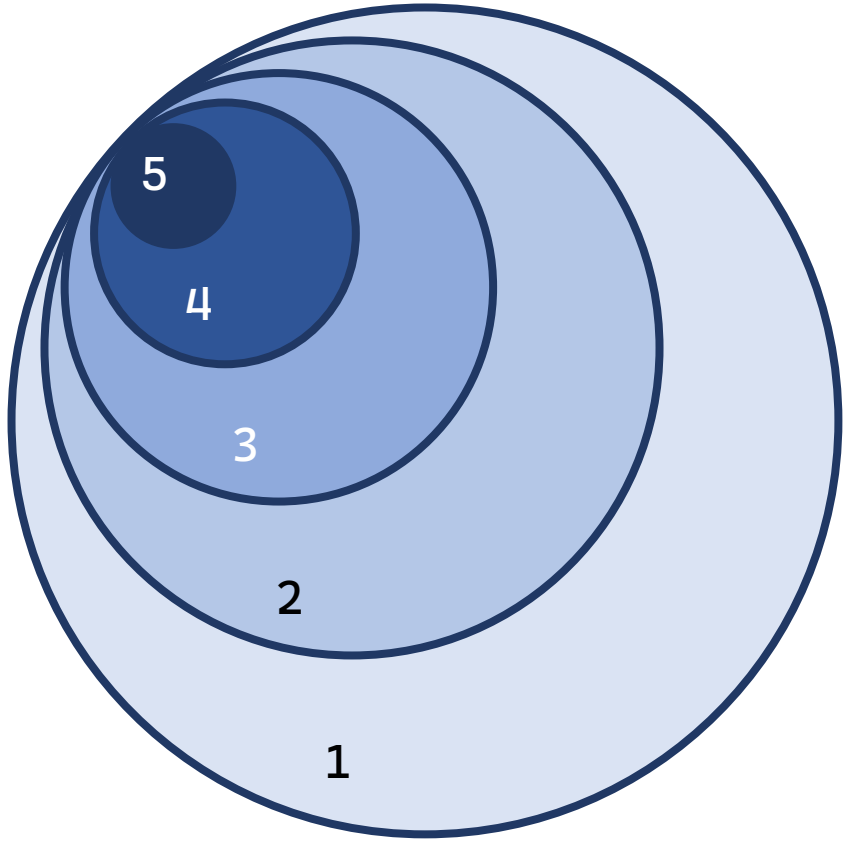
Medicinal Chemistry Toolbox

26.10.2024

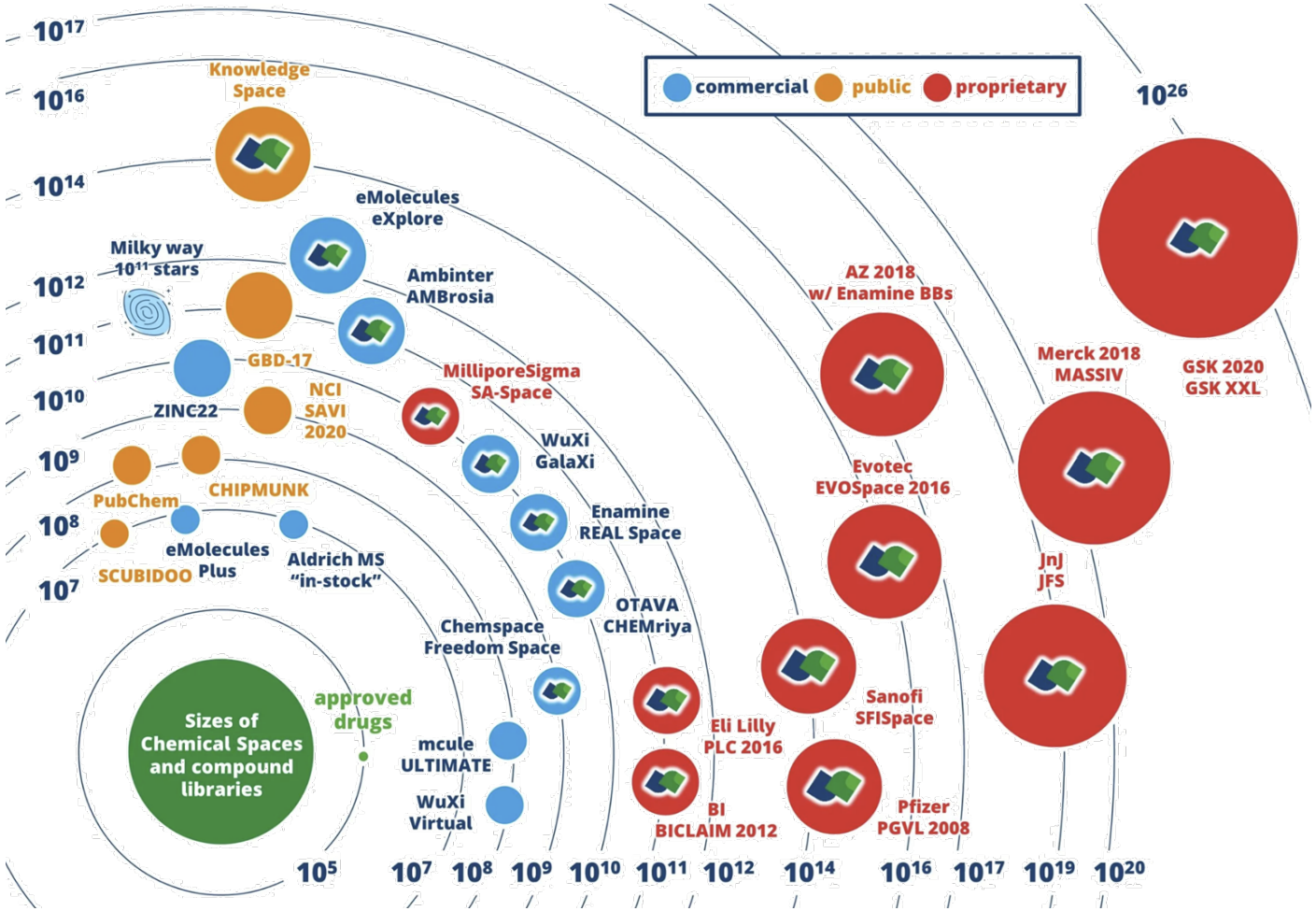
Alex Malyshev

Chemical space

1. **Theoretically possible molecules**
   ~10^{100}-10^{200}
2. **Drug-like molecules**
   ~10^{50}-10^{60}
3. **Synthetically accessible drug-like molecules**
   ~10^{10}-10^{11}
4. **Synthesized drug-like molecules**
   ~10^{7}-10^{8}
5. **Diverse hits**
   ~10^{4}-10^{5}



Drug-like libraries. Chemical databases



Cortellis Drug Discovery Intelligence

CAS SciFinder

Reaxys

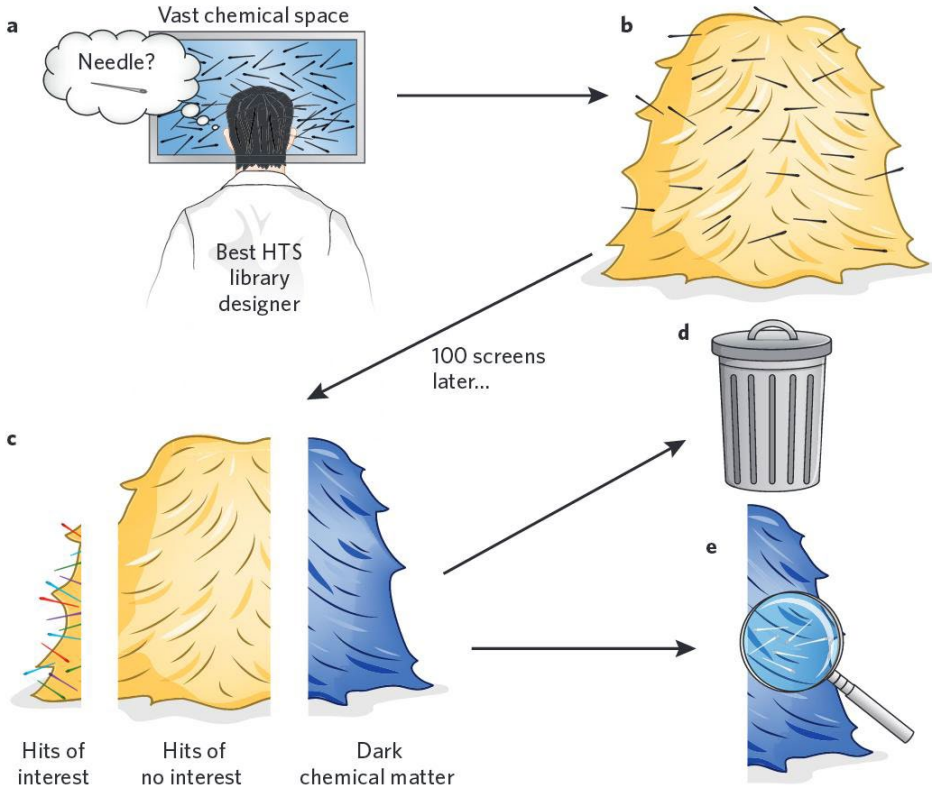
GOSTAR

ChEMBL

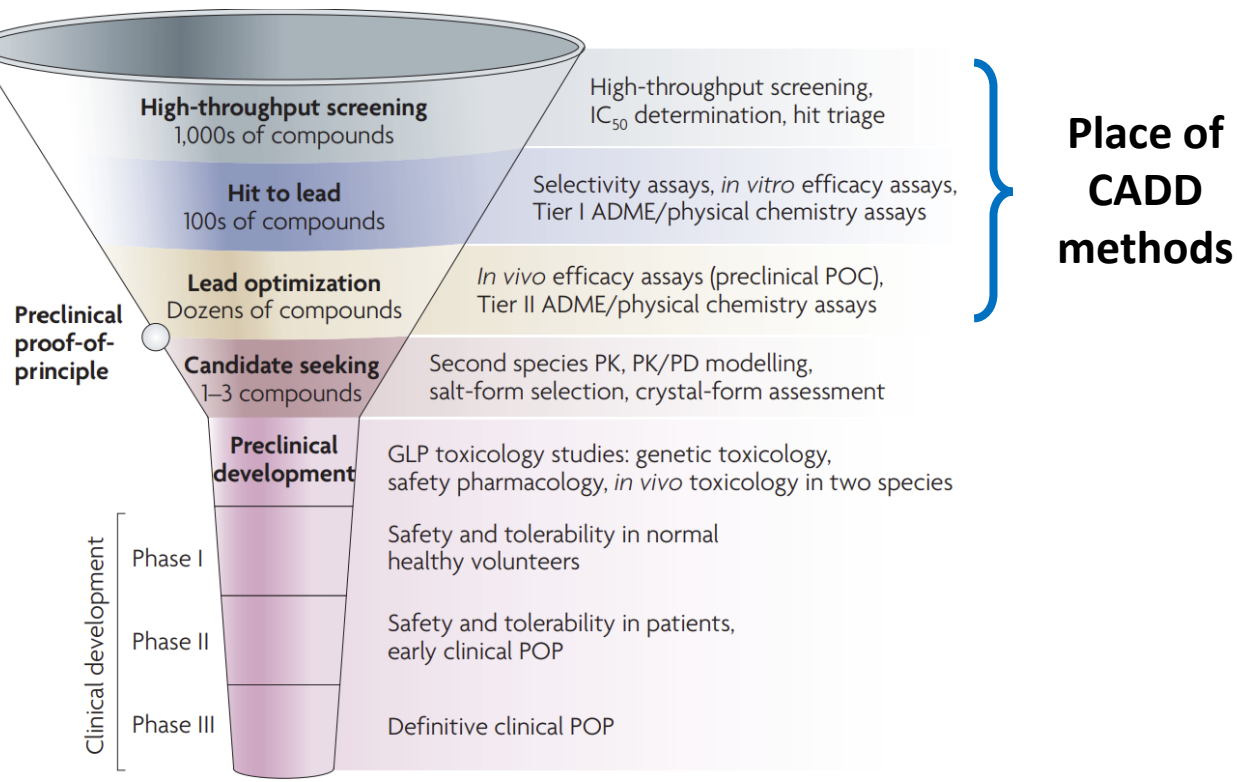
Cambridge Structural Database

https://www.biosolveit.de/chemical-spaces/

Drug discovery



10.1021/cb100420r

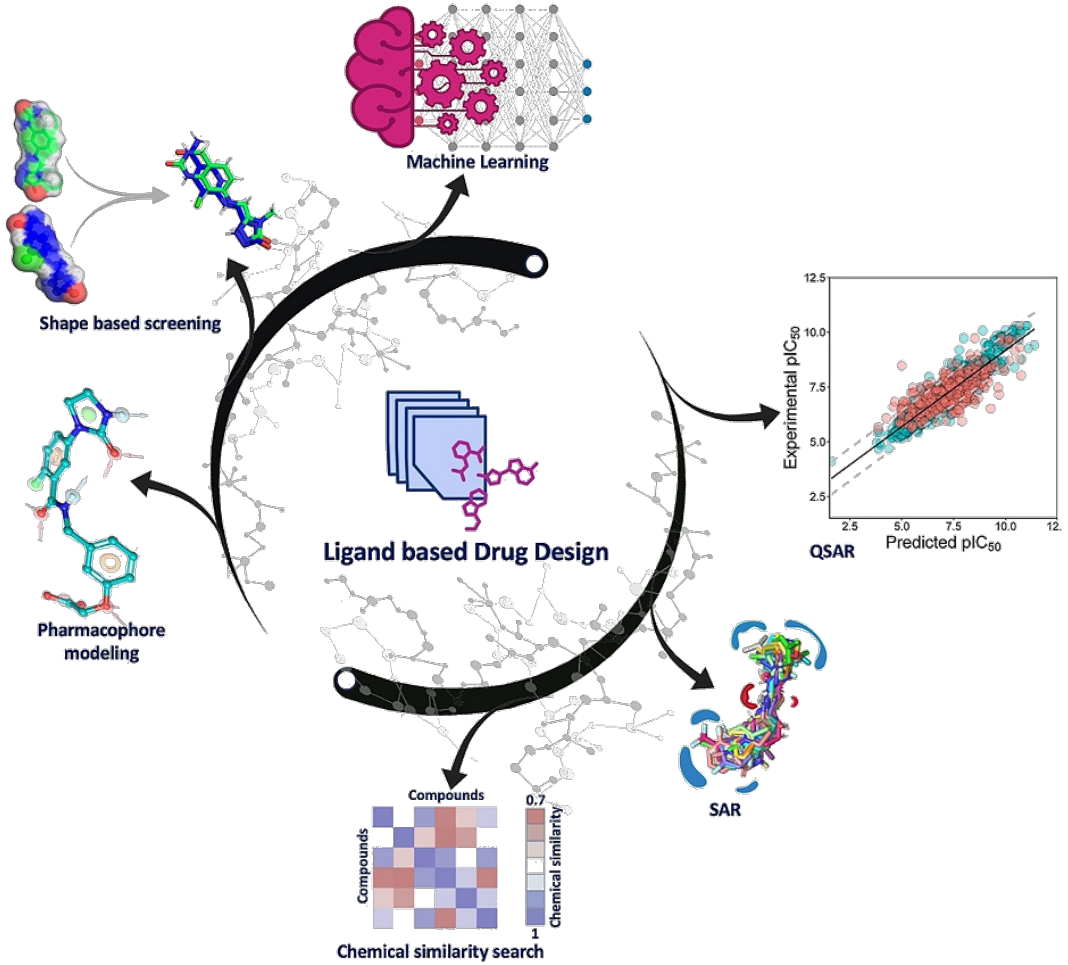


Place of CADD methods

10.1038/nrd2378

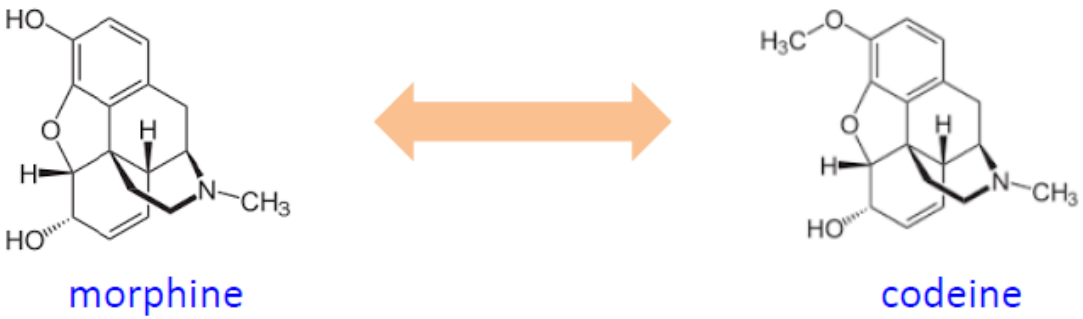
Rational drug design techniques

	Known ligand	Unknown ligand
Known target structure	**Structure-based drug design (SBDD)** Docking	***De novo* design**
Unknown target structure	**Ligand-based drug design (LBDD)** *1 or more ligands* • Similarity search *Several ligands* • Pharmacophore *Large number of ligands (20+)* • Quantitative Structure-Activity Relationships (QSAR)	**CADD not possible** some experimental data needed ADMET filtering



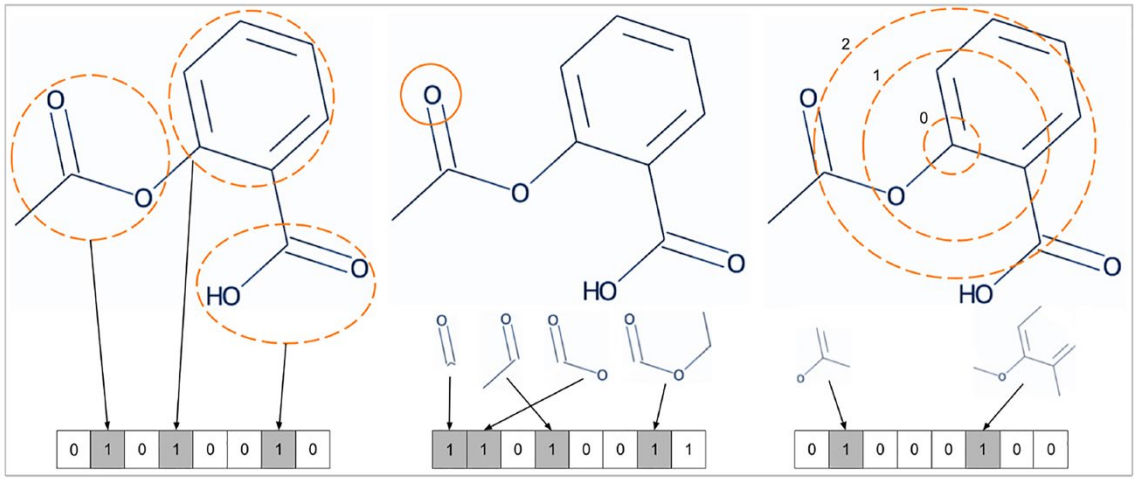
2D similarity

Similar compounds have similar properties



morphine codeine

Molecular Fingerprints

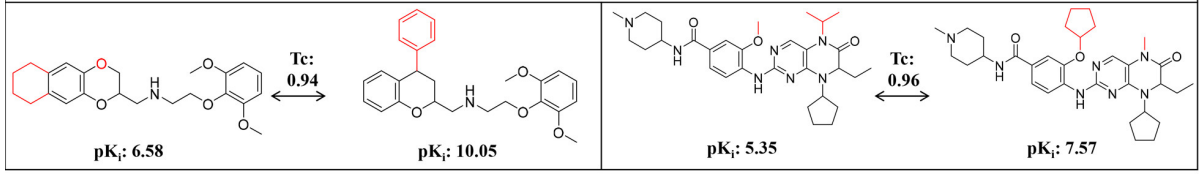


Structural Keys Path-based Circular

Name	Continuous	Dichotomous
Tanimoto coefficient	$T(x_a,x_b)=\frac{\sum_{i=0}^{N}x_{ai}\cdot x_{bi}}{\sum_{i=0}^{N}x_{ai}^{\ 2}+\sum_{i=0}^{N}x_{bi}^{\ 2}-\sum_{i=0}^{N}x_{ai}\cdot x_{bi}}$	$T(x_a,x_b)=\frac{c}{a+b-c}$
Euclidean distance	$D(x_a,x_b)=\sqrt{\sum_{i=0}^{N}(x_{ai}\cdot x_{bi})^2}$	$D(x_a,x_b)=\sqrt{a+b-2c}$
Hamming distance	$D(x_a,x_b)=\sum_{i=0}^{N}\lvert x_{ai}-x_{bi}\rvert$	$D(x_a,x_b)=a+b-2c$
Cosine coefficient	$C(x_a,x_b)=\frac{\sum_{i=0}^{N}x_{ai}\cdot x_{bi}}{\sqrt{\sum_{i=0}^{N}x_{ai}^{\ 2}}\cdot\sqrt{\sum_{i=0}^{N}x_{bi}^{\ 2}}}$	$C(x_a,x_b)=\frac{c}{\sqrt{ab}}$
Dice coefficient	$D(x_a,x_b)=\frac{2\sum_{i=0}^{N}x_{ai}\cdot x_{bi}}{\sqrt{\sum_{i=0}^{N}x_{ai}^{\ 2}}\cdot\sqrt{\sum_{i=0}^{N}x_{bi}^{\ 2}}}$	$D(x_a,x_b)=\frac{2c}{a+b}$
Soergel	$S(x_a,x_b)=\frac{\sum_{i=0}^{N}\lvert x_{ai}-x_{bi}\rvert}{\sum_{i=0}^{N}\max(x_{ai},x_{bi})}$	$S(x_a,x_b)=\frac{a+b-2c}{a+b-c}$

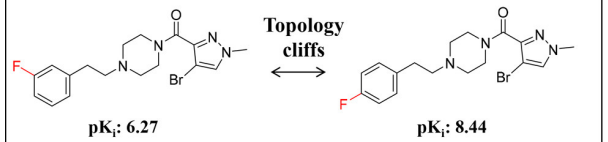
Activity and property cliffs

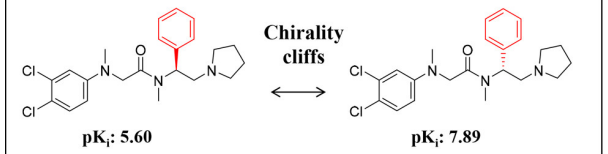
Fingerprint-based



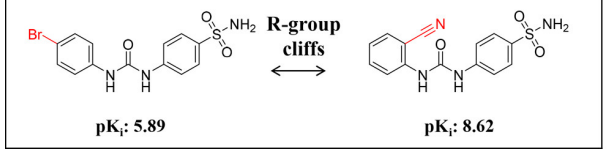
Substructure-based

Category 1: same scaffold, same R-groups

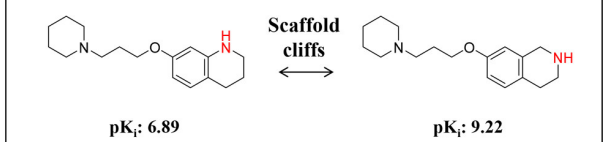


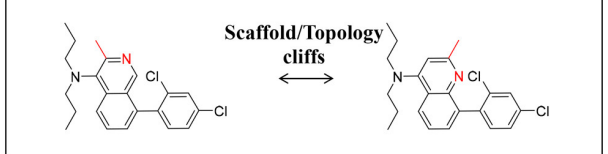


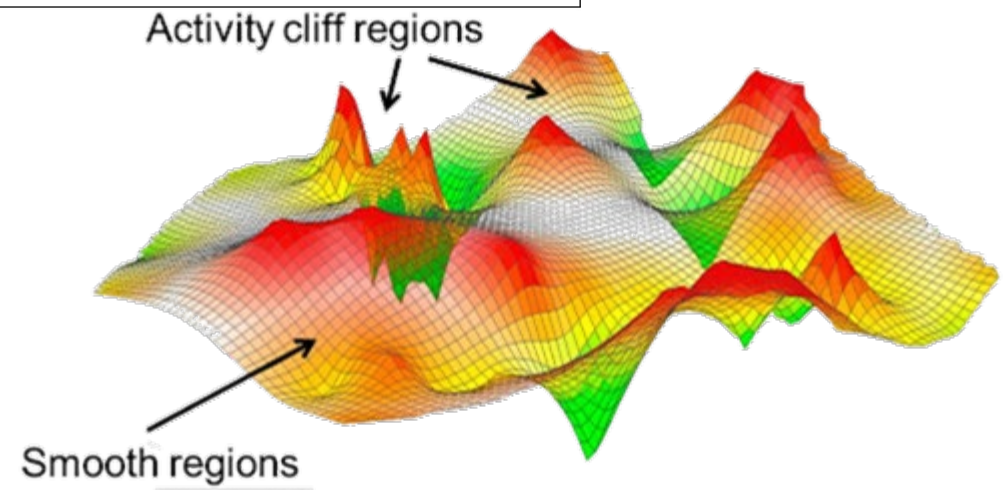
Category 2: same scaffold, different R-groups

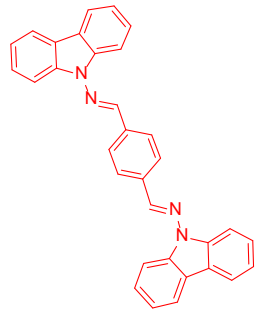
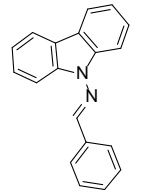
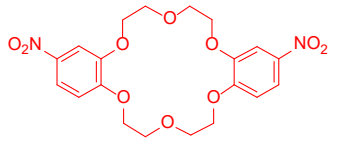
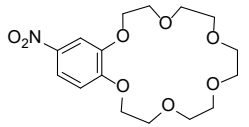
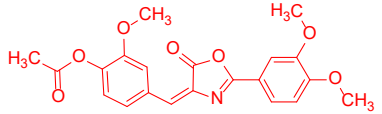
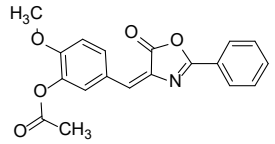
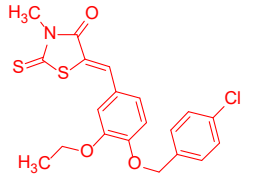
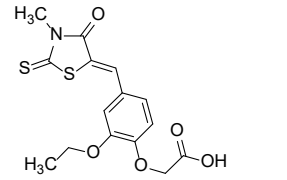
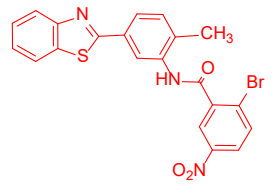
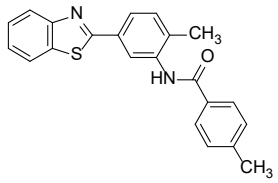


Category 3: different scaffolds, same R-groups





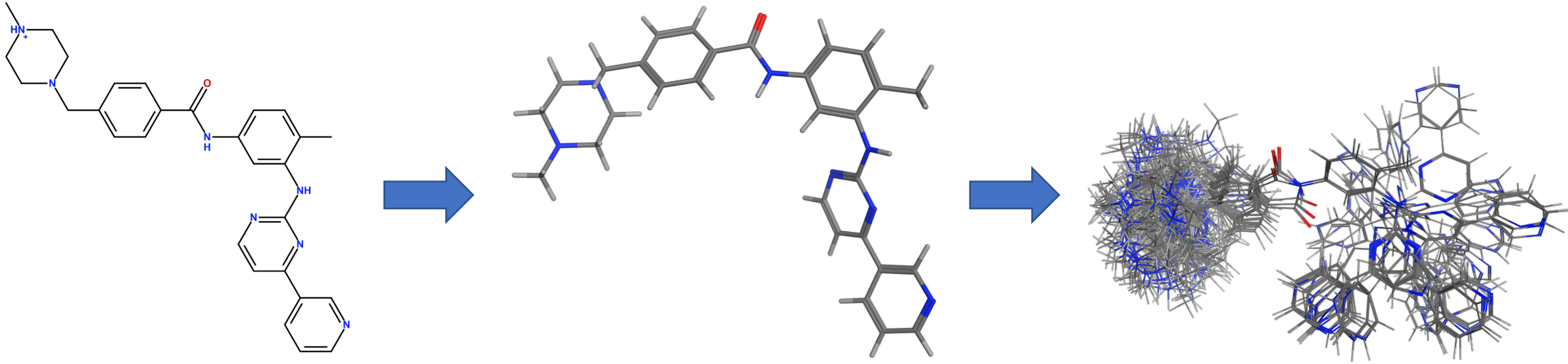


DMSO(-)	Similarity	DMSO(+)
	0.99	
	0.99	
	0.98	
	0.83	
	0.81	

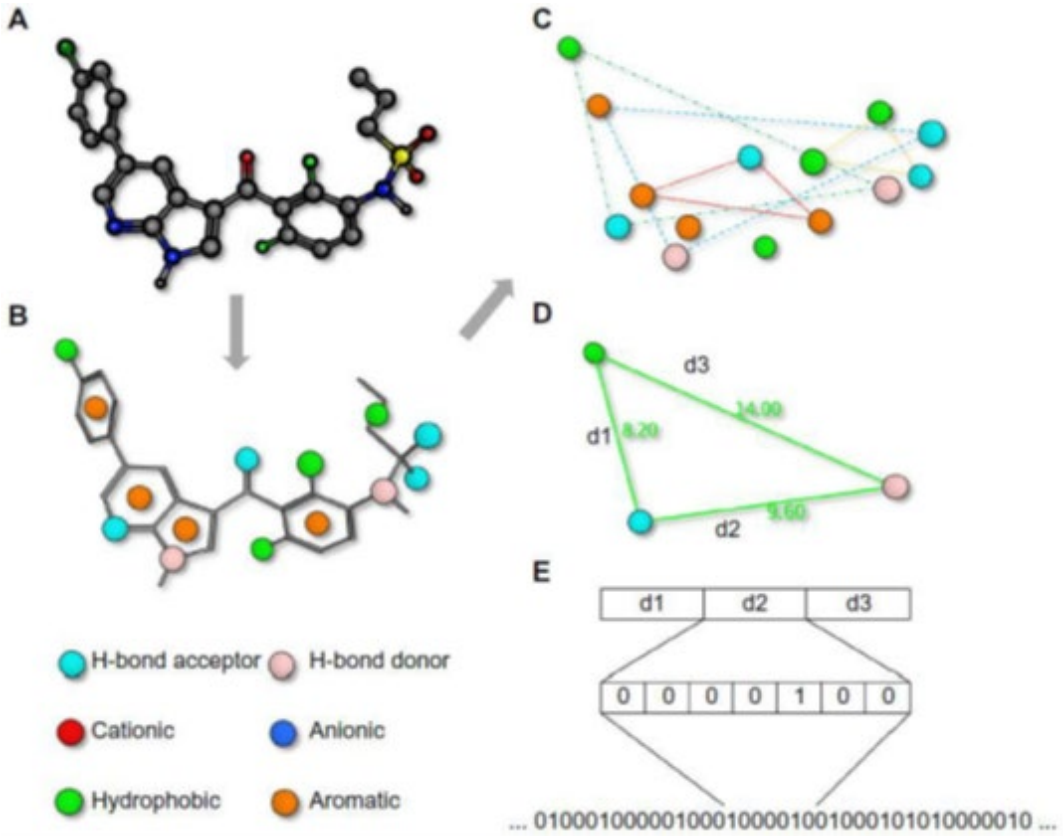
3D approaches. Conformational search

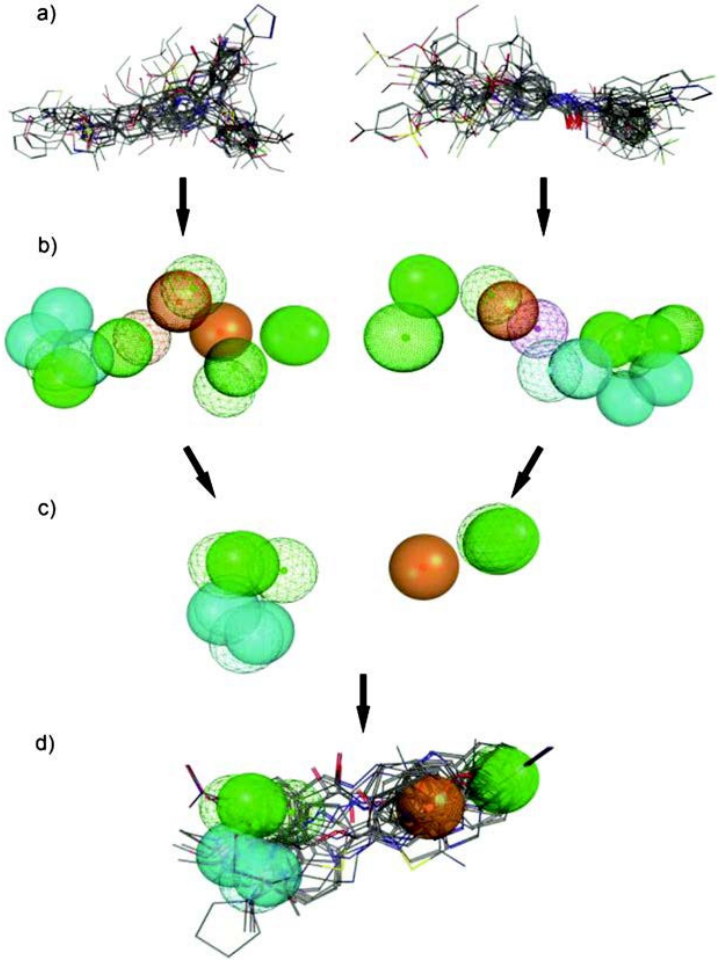
Methods

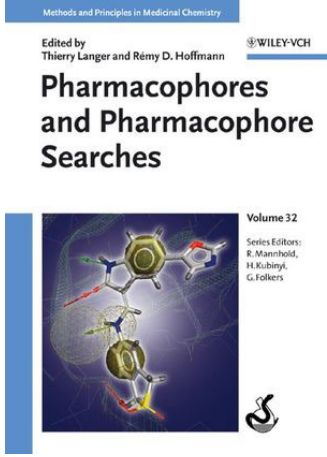
- Fragment-based (Corina)
- MM/MD-based
- QM-based



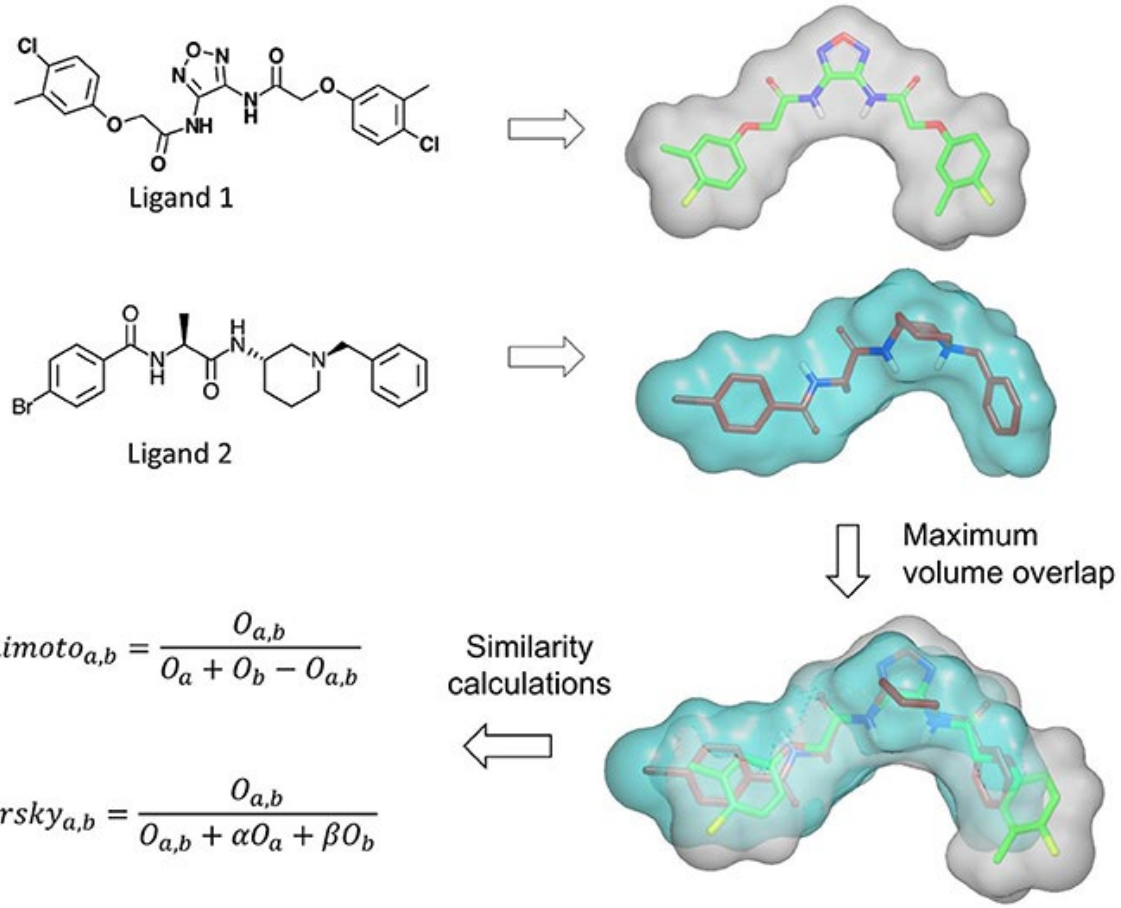
Pharmacophore modeling





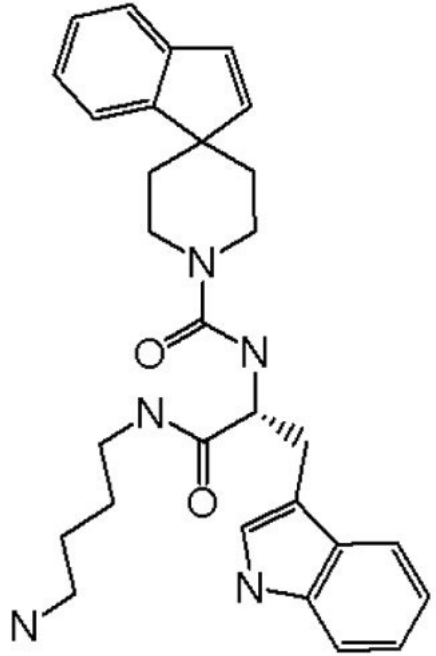


Shape similarity

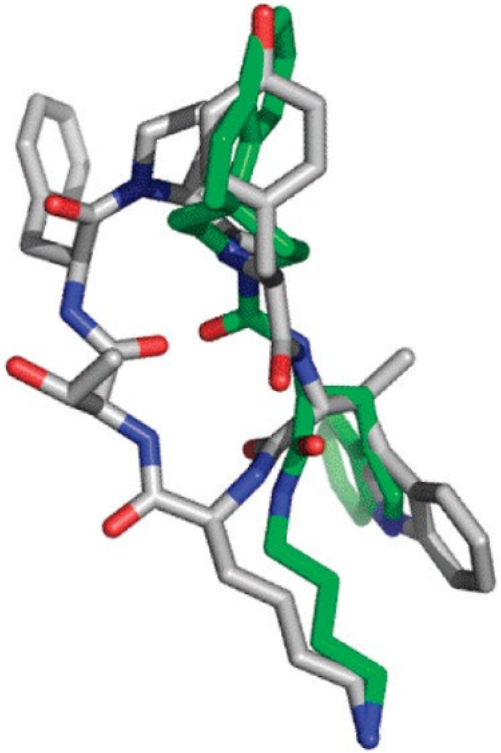


$$Tanimoto_{a,b} = \frac{O_{a,b}}{O_a + O_b - O_{a,b}}$$

$$Tversky_{a,b} = \frac{O_{a,b}}{O_{a,b} + \alpha O_a + \beta O_b}$$



K_i (SST_2) = 300 nM
Top-41 out of 1 M
(Merck database)



10.1021/jm900818s

ML and AI techniques

Input data

Bioassays

Databases

Preprocessing

Data normalization & curation

Feature extraction

Feature engineering

$$x_i' = \frac{x_i - \bar{x}}{\sum_j z_j}$$

Feature selection

Feature combination

Model training

Classification

Regression

Clustering

Model validation

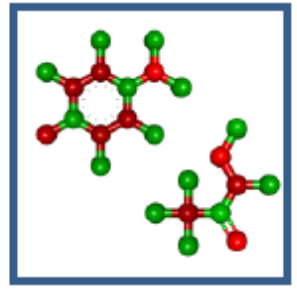
Cross-validation

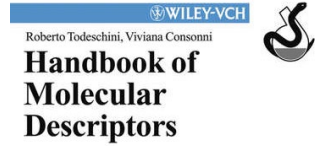
Bootstrap

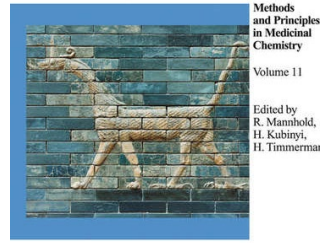
Test set

Applicability Domain

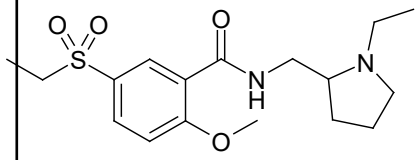
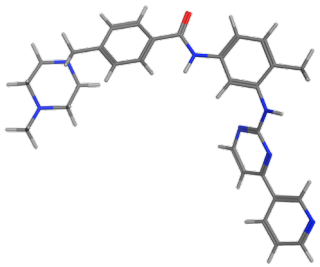
Interpretation







Descriptors as molecular features

Category	Input	Examples
1D	$C_{17}H_{26}N_2O_4S$	MW, N of atoms (by types), etc.
2D	2D structure 	Topological indices, logP, logS, structural fragments, topological pharmacophores, etc.
3D	3D structure 	VdW surface/volume, polar surface area (PSA), moment of inertia, CATS, MoRSE, etc.

Descriptors calculation

RDKit

Alvadesc

MOE

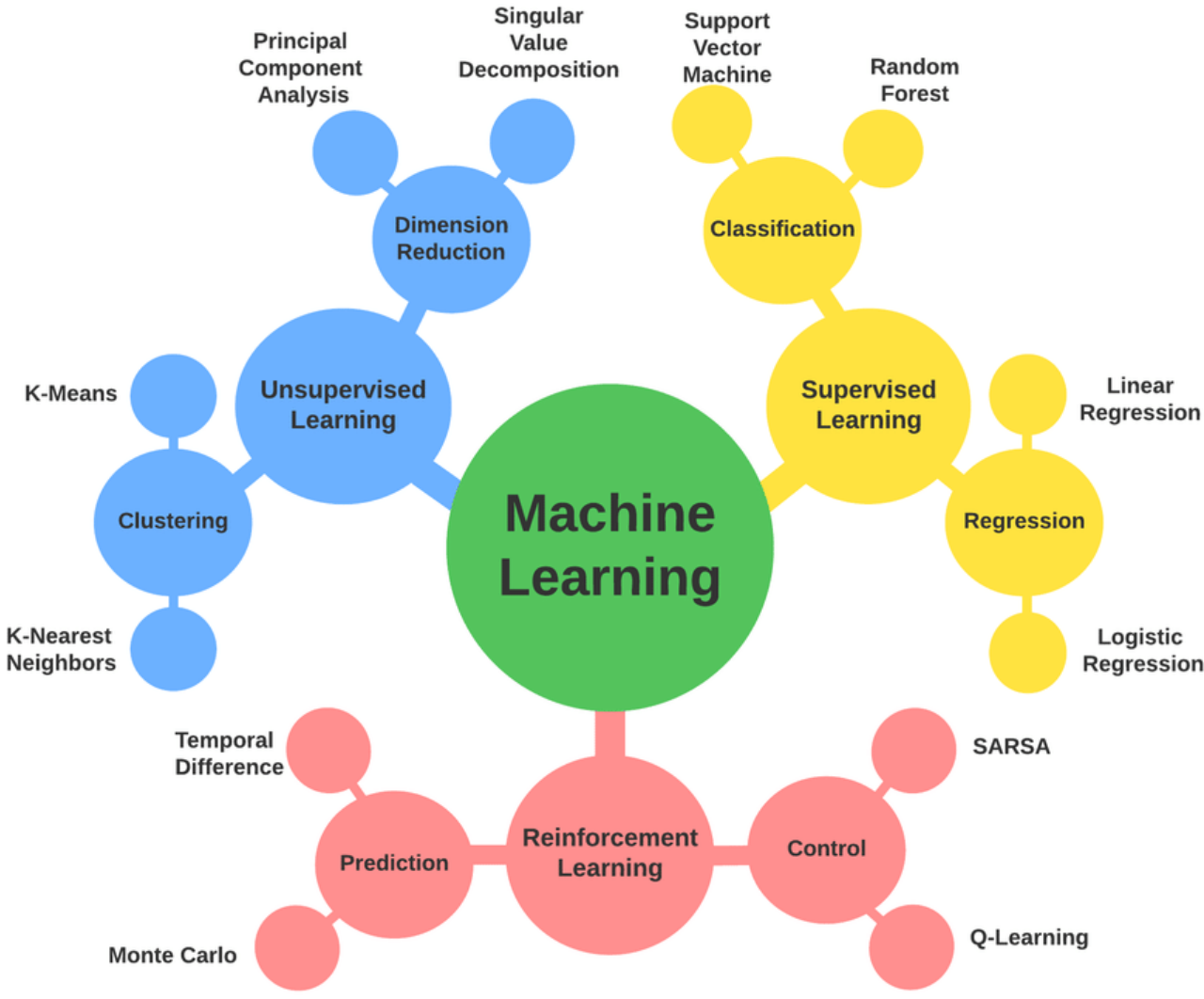
Dragon

SmartMining

AI impacting medicinal chemistry and drug design

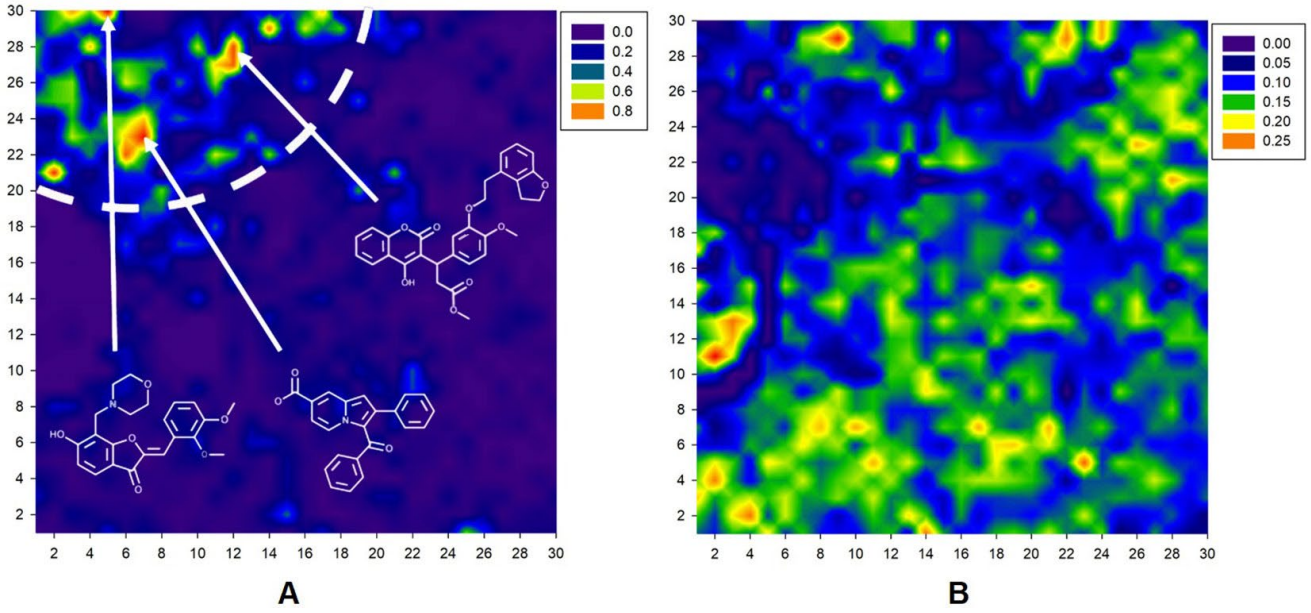
- ML/DL ADMET and other models
- Synthetic accessibility prediction
- *De novo* design:
  - Hit ID and hit-to-lead optimization (H2L)
  - Lead optimization (LeadOpt)

and many other implementations…

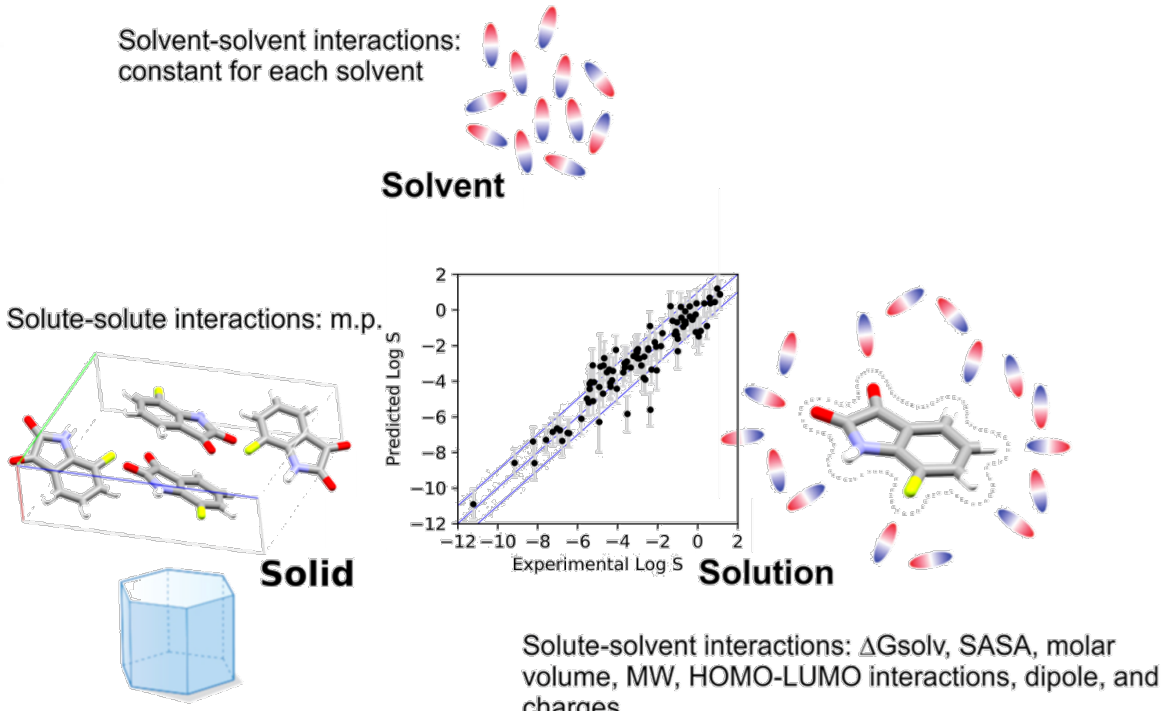


QSAR and QSPR

Antibacterial activity prediction

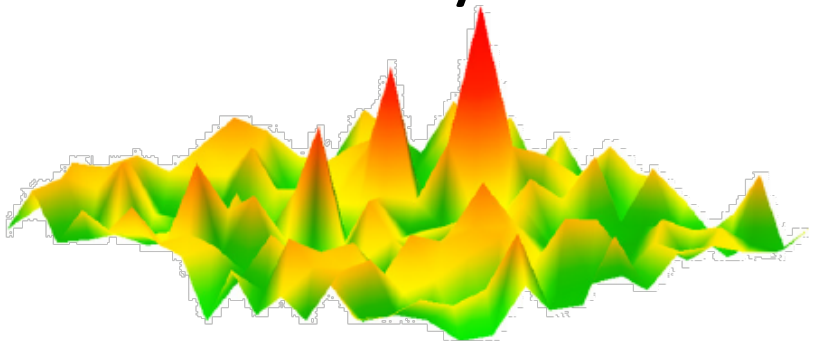


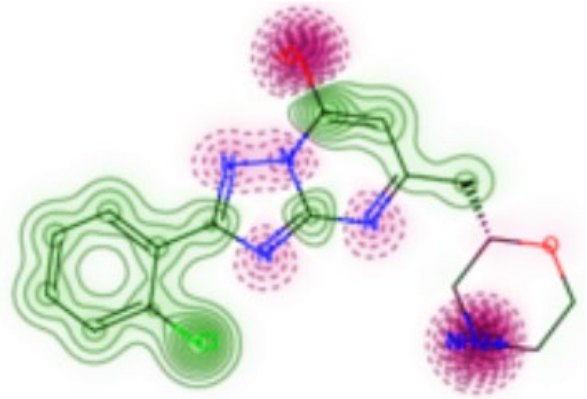
Solubility prediction



10.3389/fphar.2019.00913

Activity Cliffs





- **Molecule and fragment scoring**
- **Rapid prognosis**
- **Fully automatic platform**

✓ *In silico* PoC study (ABL1 ligands)

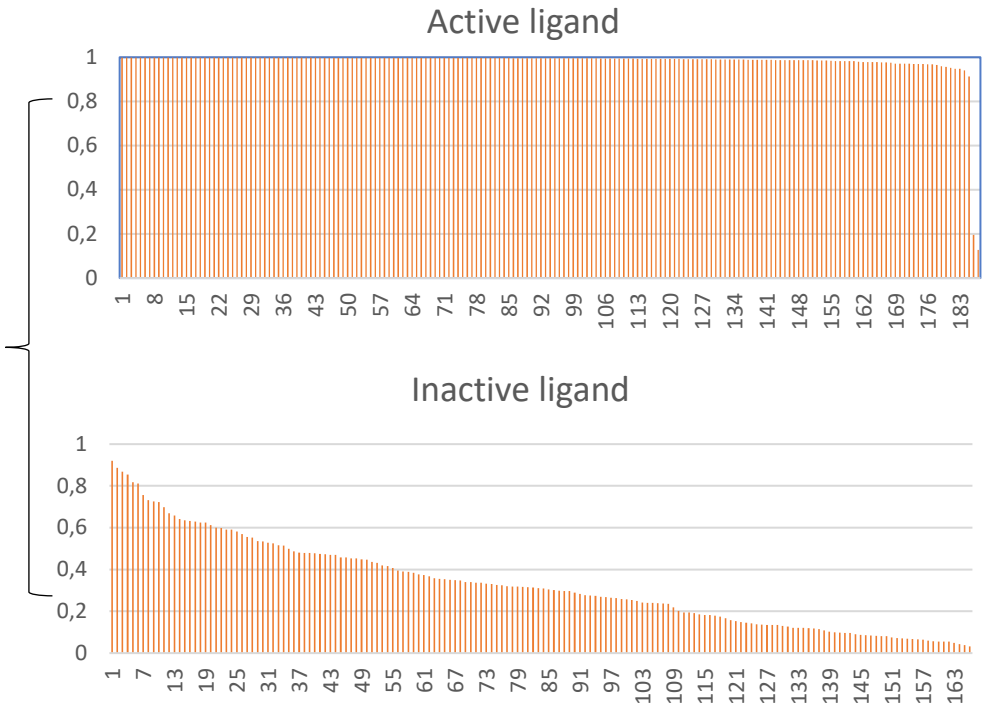
F1-metrics for both classes

Class	F1-score
Active	0.93
Inactive	0.82

Retrospective validation on active and inactive molecules

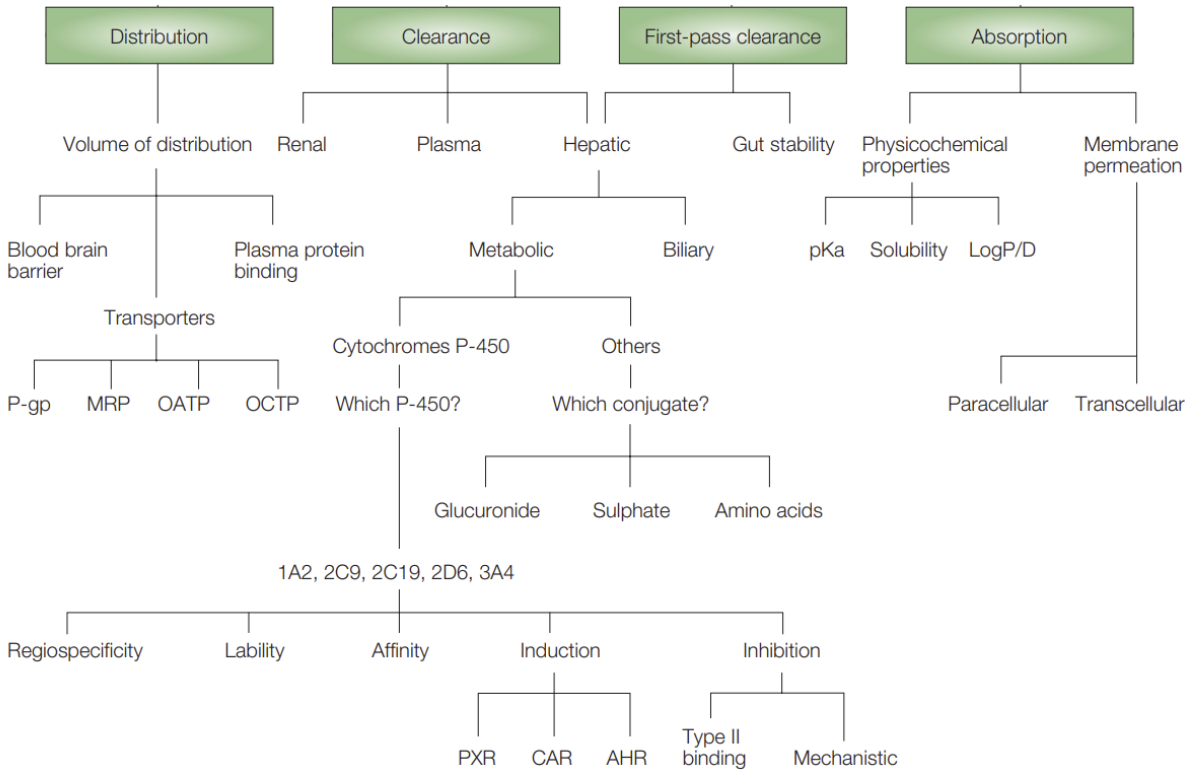
Class	Probability of being active
Active ligand	0.98
Inactive ligand	0.33

Fragment scores

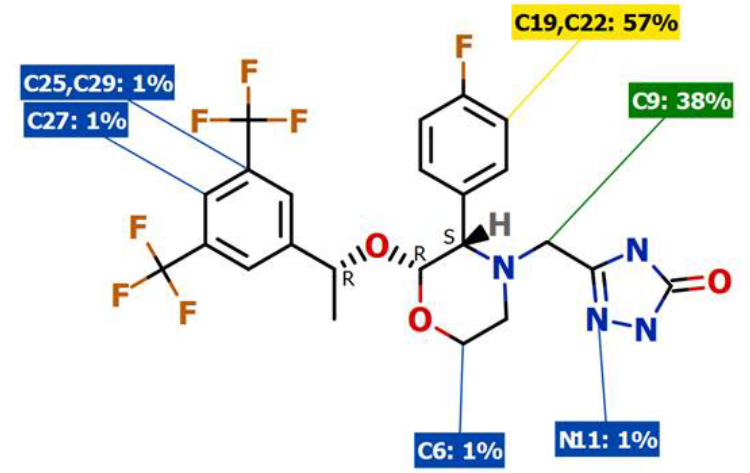


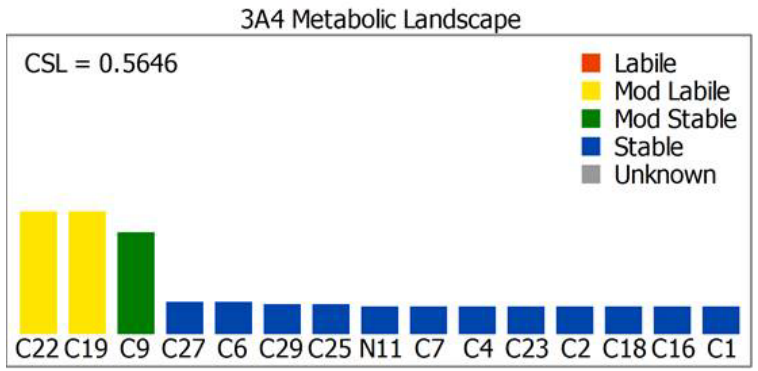
ADMET properties prediction

Classification of ADME properties

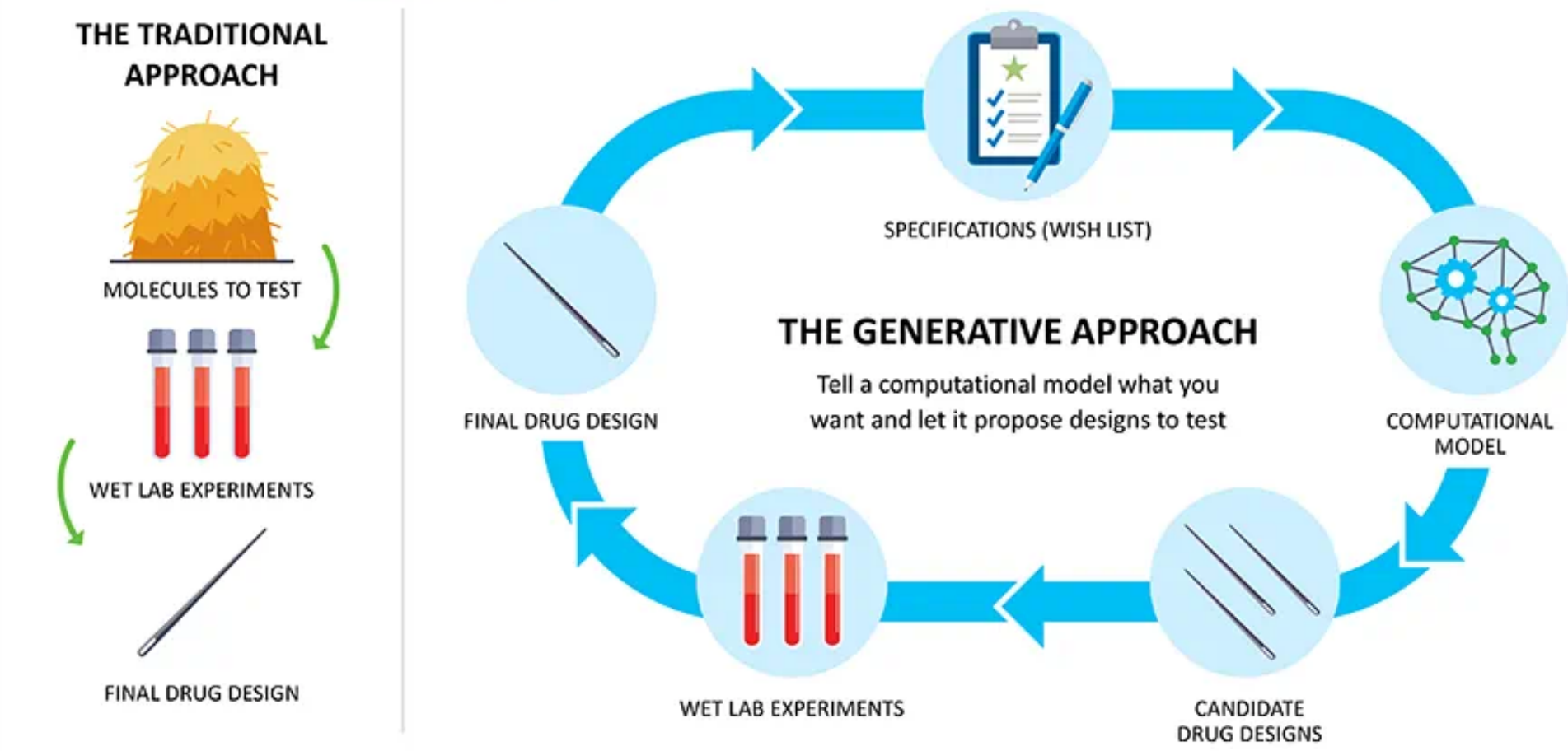


CYP3A4 metabolism prediction



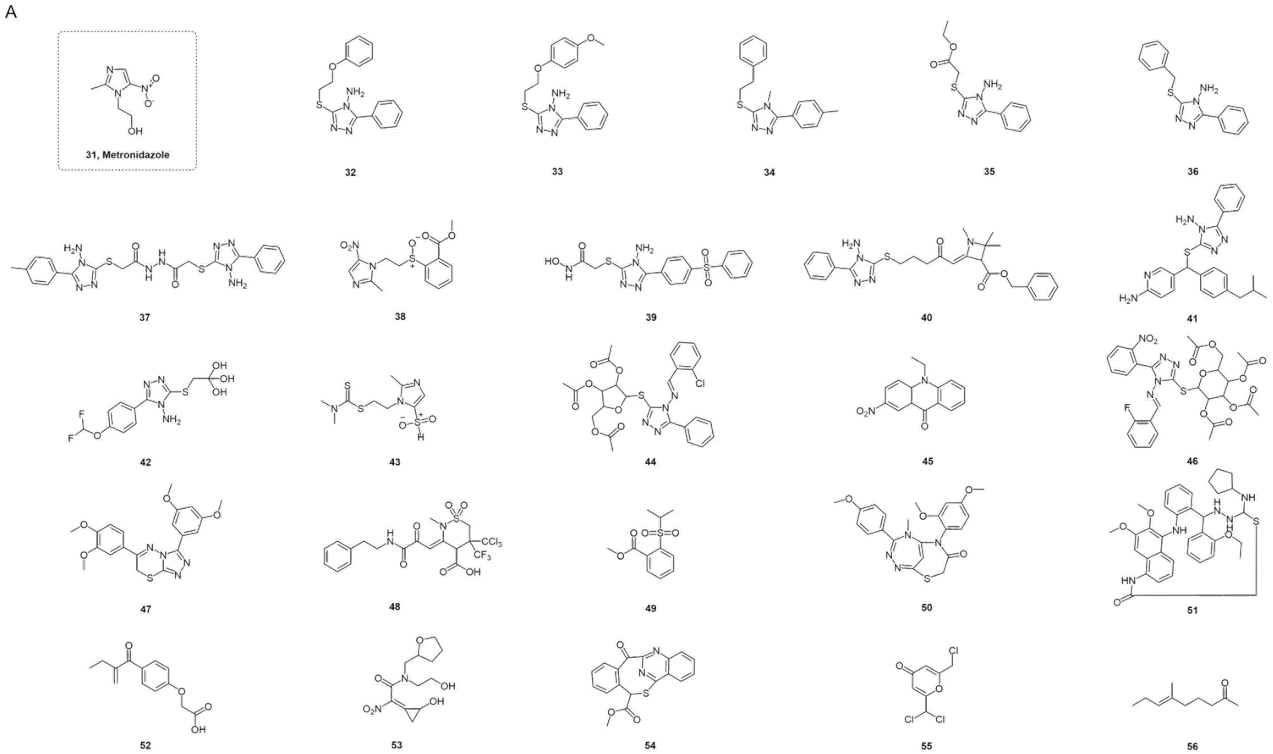


Generation of novel structures



The Hitchhiker's Guide to Deep Learning Driven Generative Chemistry

Output of RNN-based model



Cliff's Notes

IP position and Novelty

Filter out structural alerts

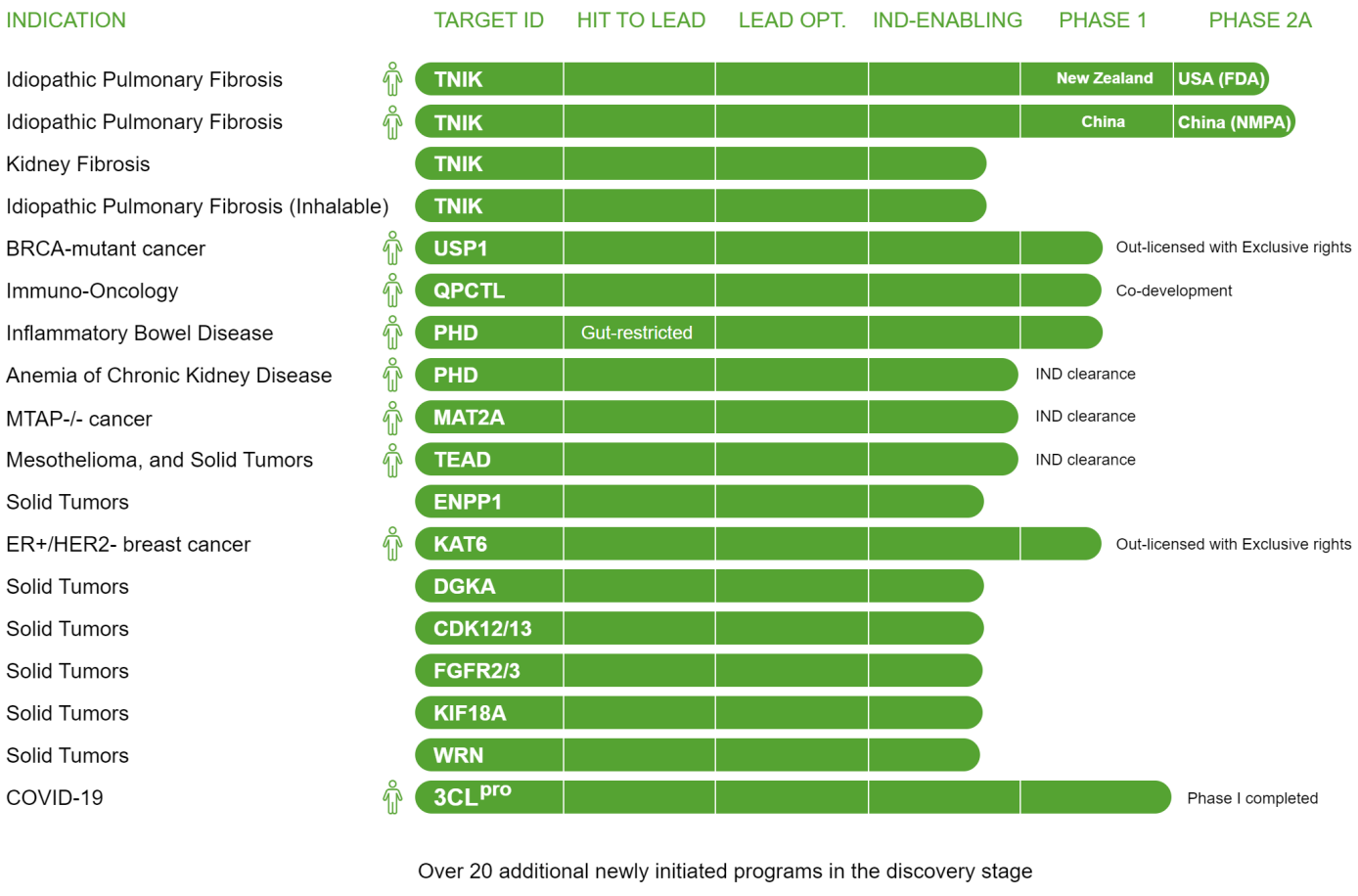
Generate molecules targeting other than kinases (e.g. PPI or GPCRs)

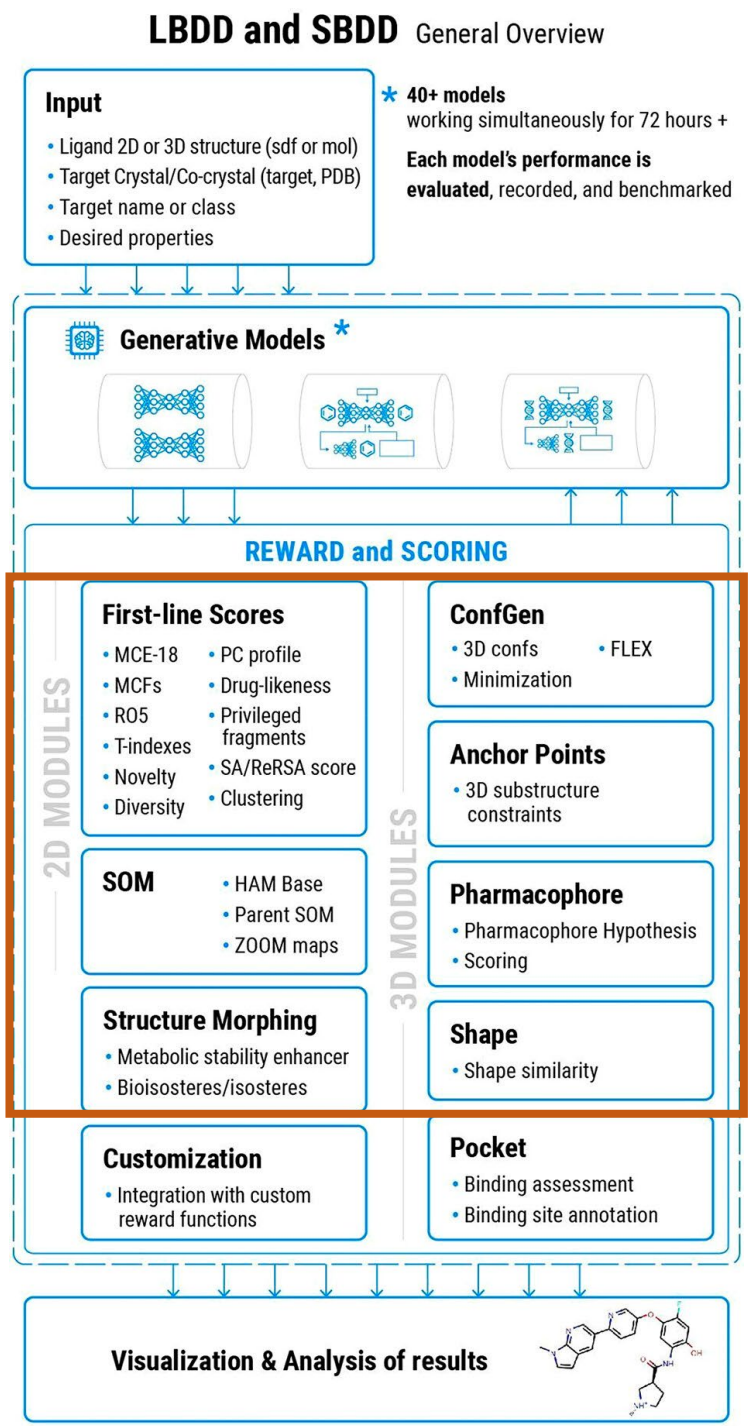
Target-specific profiling before testing

Synthetic accessibility

10.1021/acsmedchemlett.3c00041

Chemistry42 platform



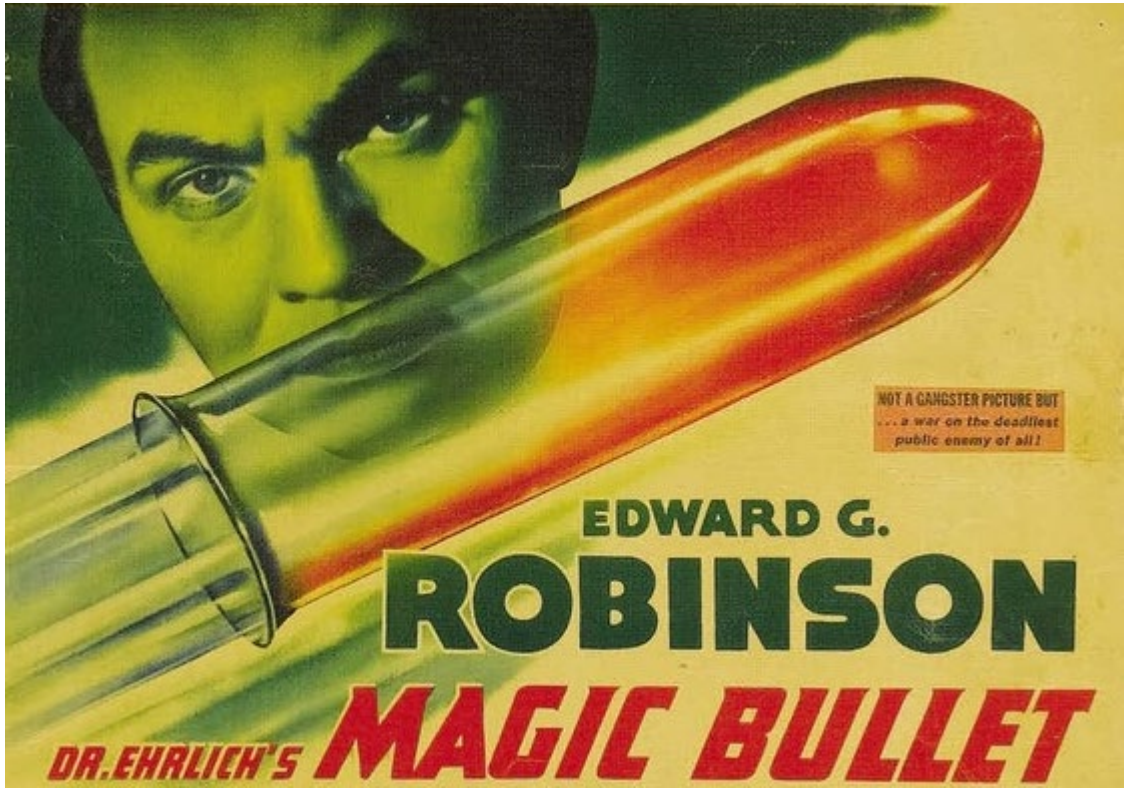


10.1021/acs.jcim.2c01191

Multimodality of drugs

PAST

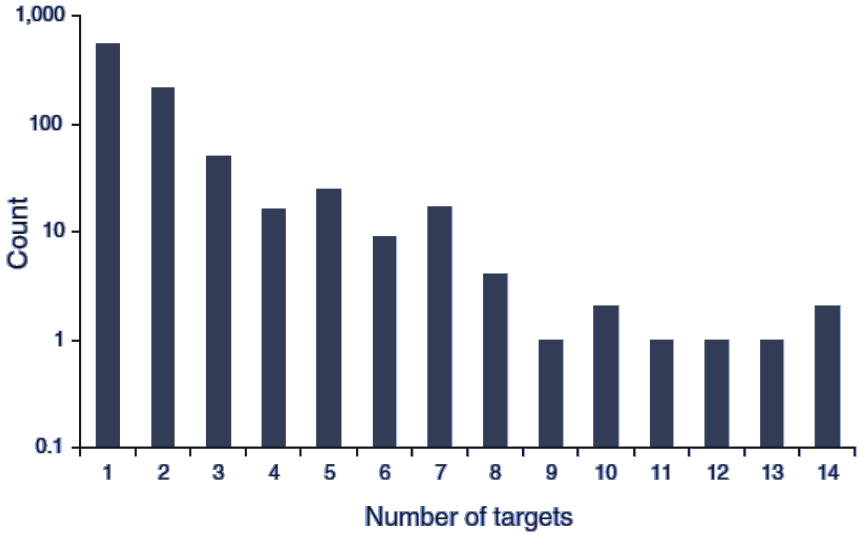
MONOpharmacology – drug is a «magic bullet»



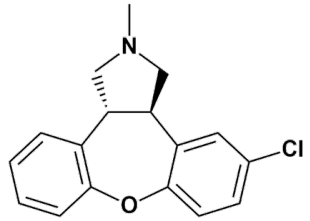
Multimodality of drugs

PAST — MONOpharmacology – drug is a «magic bullet»

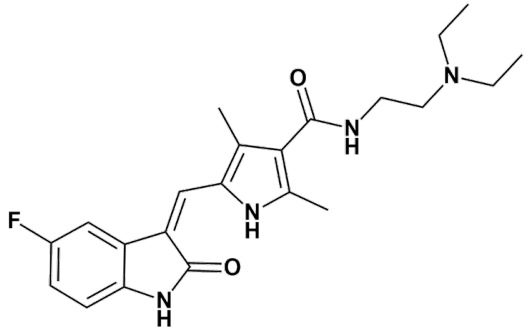
NOW — **POLYpharmacology – drug is a «magic shrapnel»**



Distribution of drugs & number of their targets



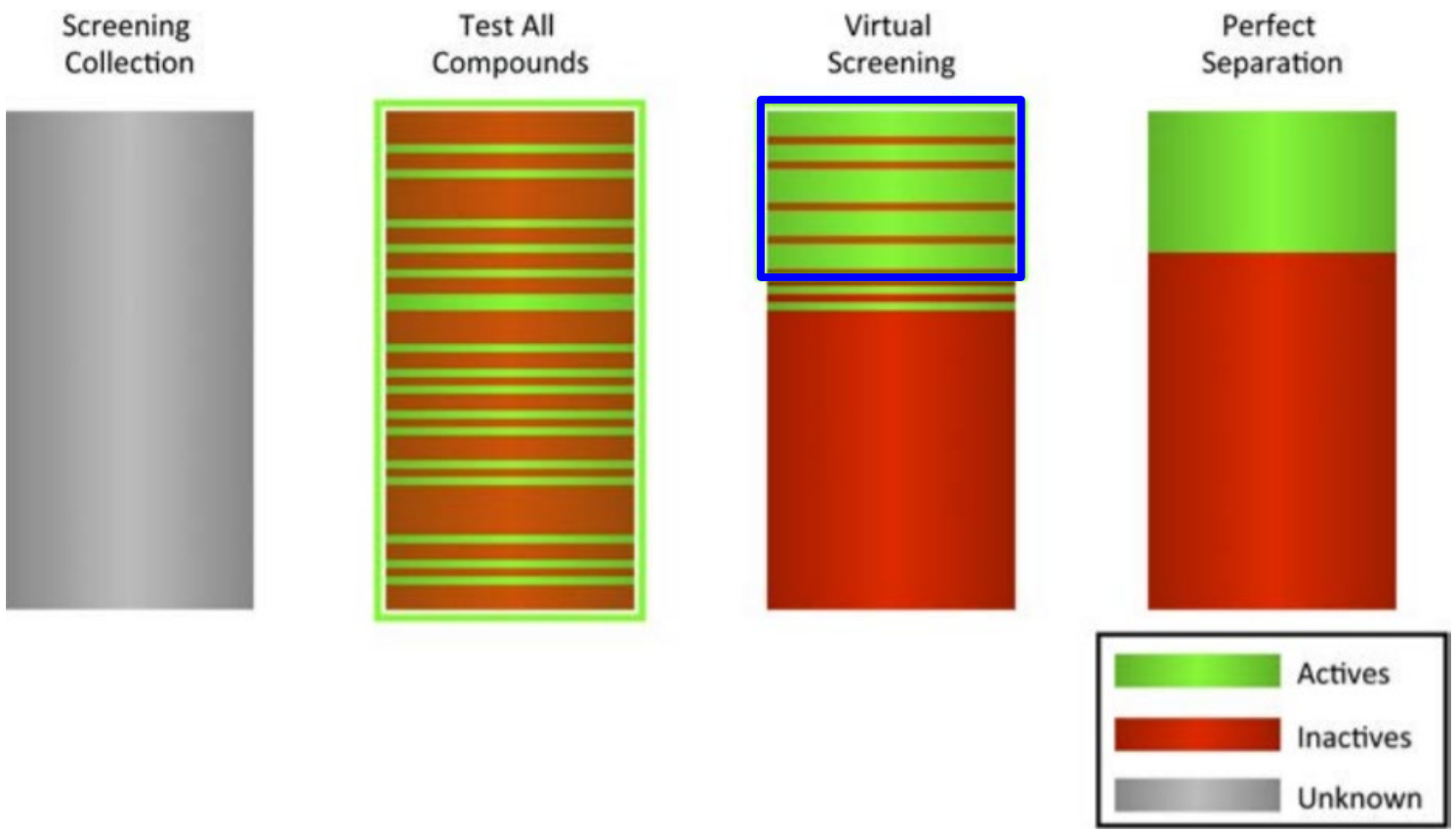
Asenapine
Schering-Plough (2009)
Schizophrenia
Low nM affinity for at least 18 GPCRs



Sunitinib
Pfizer (2006)
Cancer
Inhibition of 79 kinases (K_d < 10 µM)

Nat Biotechnol. 2007 25:1119–1126

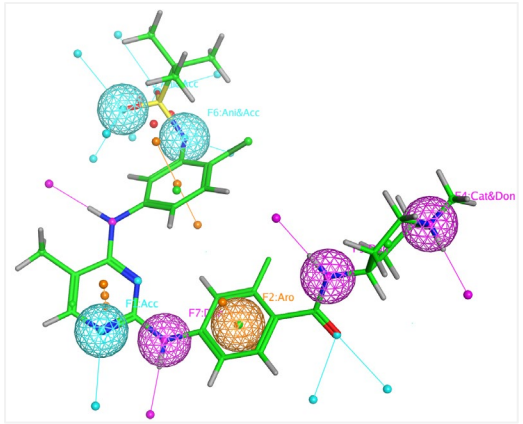
Strategies for ligand-target profiling

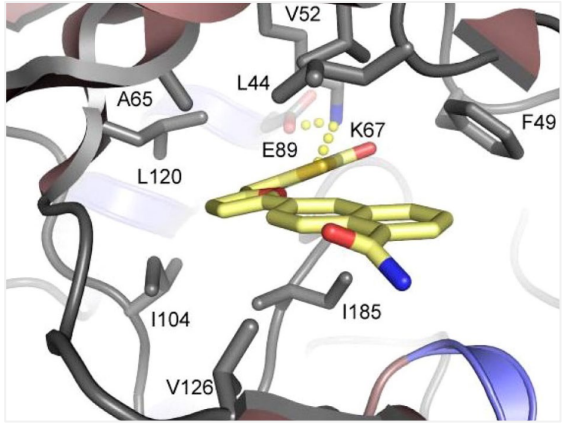


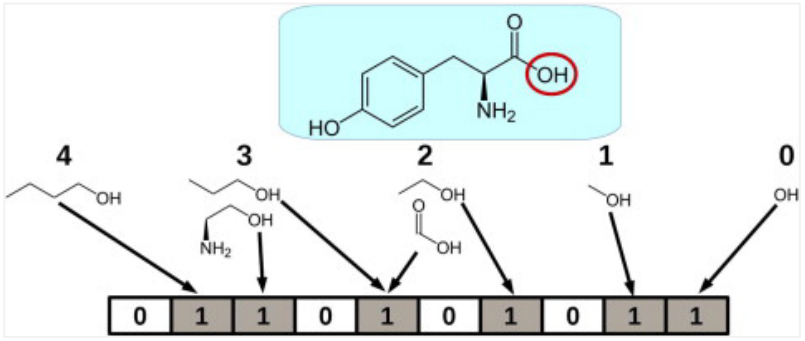
In Silico Medicinal Chemistry: Computational Methods to Support Drug Design, 2016, N. Brown

Strategies for ligand-target profiling

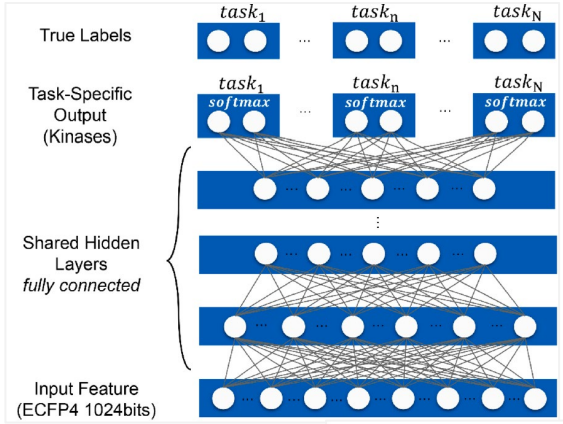
Classical approaches

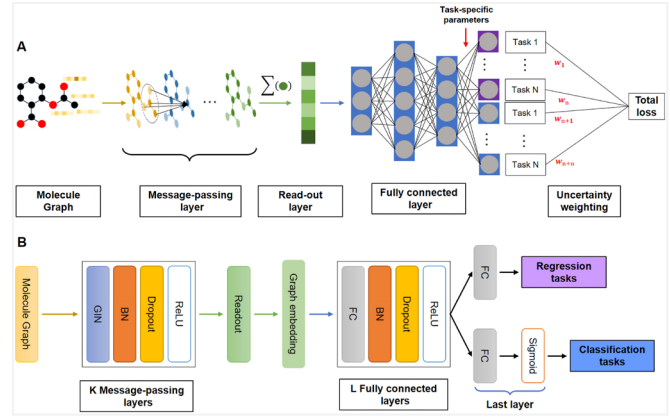


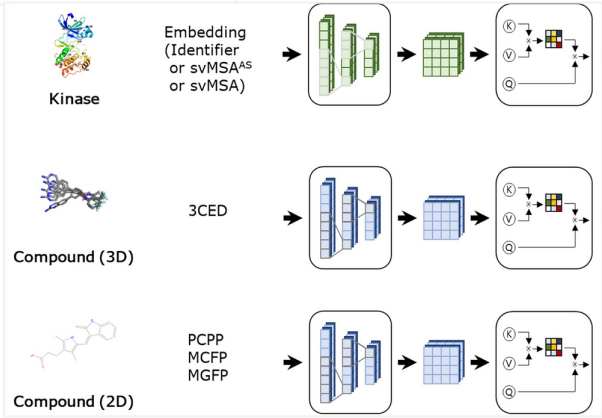




AI-based approaches







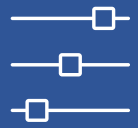
Insilico Medicine

GoldenCubes

Use cases

Predict Kinase Selectivity

Identify potential on- and off-targets for compounds during hit and lead optimization stages



Design Multimodal Compounds

Manage the kinome promiscuity of the molecules depending on your needs

Drug Repurposing

Discover novel targets for existing compounds and reduce costs for their development timeline

Core features

Big Data-Driven Approach

Large-scale and precisely annotated molecular datasets

>500K 2D structures

>2K 3D ligands

Descriptor-Based Engine

Evaluate molecules beyond the explored kinase chemical space

Interpretable descriptors

Diverse Kinase Target Set

Annotate structures across all existing kinase families

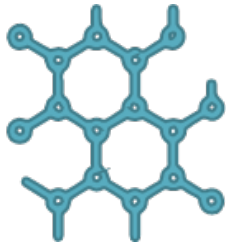
100 kinases

Golden Cubes scoring workflow

Input data

SMILES

2D/3D SDF

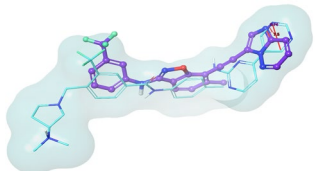


Preprocessing

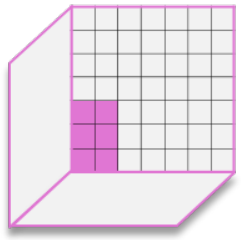
Scaffold selection

ConfGen

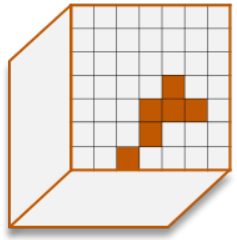
Descriptors calculation



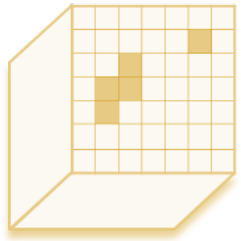
ANN-based HTVS engine



General 3D SOM

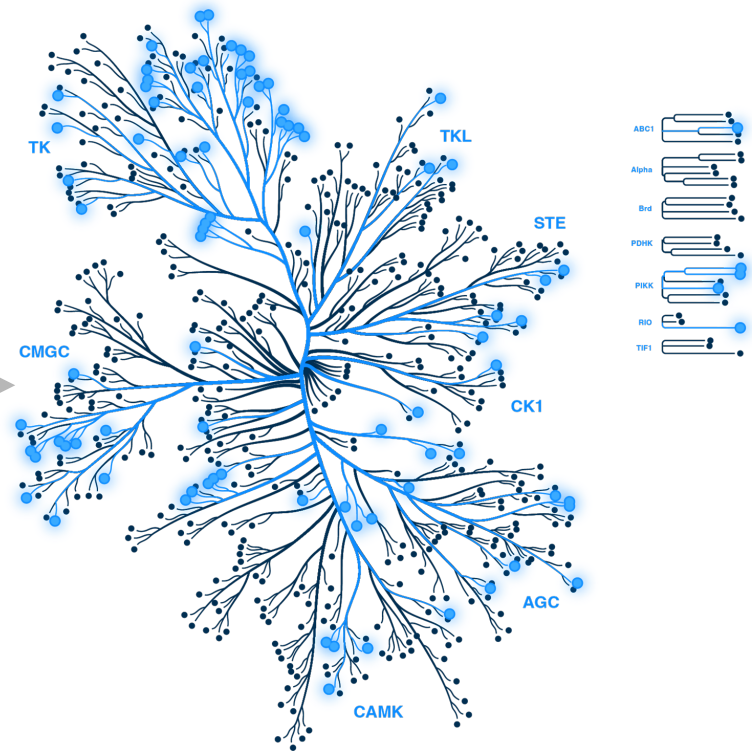


Group 3D SOMs



Atomic 3D SOMs

Scoring function



Retrospective validation

Test set

103 molecules with full-kinome profiles

LINCS KINOME scan

Full-kinome metrics

Metrics	Golden Cubes	Benchmark (QSAR)
BA	0.72	0.71
Precision	0.52	0.43
Specificity	0.9	0.85
ROC-AUC	0.72	0.75

Heatmap of predictions

Prospective validation

Test set

5 clinical candidates – kinase inhibitors with unreported profiles
TanSim **< 0.7**
Available for purchase

KINOME scan

Full-kinome metrics

Metrics	Golden Cubes
BA	0.57
Precision	0.33
Specificity	0.98
ROC-AUC	0.57

Heatmap of predictions

PATENT BUSTERS

Inside the patents tangle

US 20220119419A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2022/0119419 A1**

Zavoronkovs et al. (43) **Pub. Date: Apr. 21, 2022**

(54) **BYCYCLIC JAK INHIBITORS AND USES THEREOF**

(71) Applicant: **Insilico Medicine IP Limited**, Hong Kong (HK)

(72) Inventors: **Aleksandrs Zavoronkovs**, Pak Shek Kok (HK); **Yan Ivanenkov**, Moscow (RU); **Aleksandr Aliper**, Moscow (RU); **Anton S. Vantskul**, Moscow (RU)

(21) Appl. No.: **17/478,152**

(22) Filed: **Sep. 17, 2021**

Related U.S. Application Data

(63) Continuation of application No. PCT/US2020/025206, filed on Mar. 27, 2020.

(60) Provisional application No. 62/824,485, filed on Mar. 27, 2019.

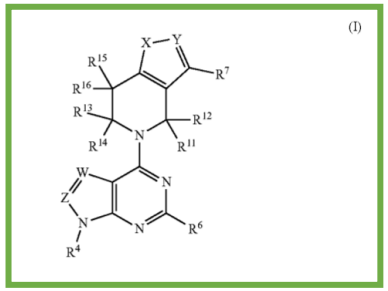
Publication Classification

(51) **Int. Cl.**
C07D 519/00 (2006.01)

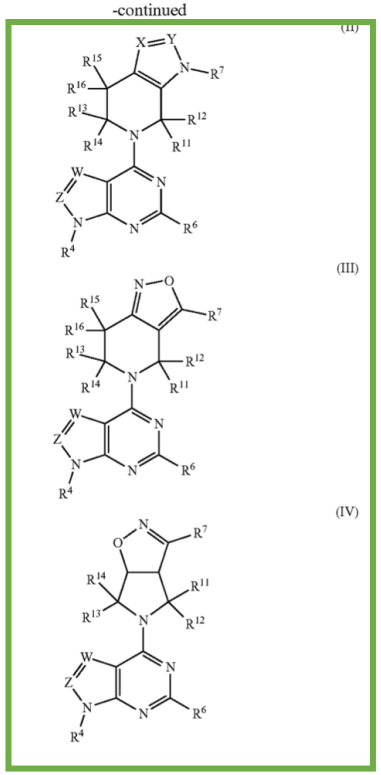
(52) **U.S. Cl.**
CPC ***C07D 519/00*** (2013.01)

(57) **ABSTRACT**

Provided herein are compounds of Formulas (I), (II), (III), and (IV)



-continued



and subformulas thereof, wherein the variables are defined herein. Also provided herein are pharmaceutical compositions comprising a compound of Formula (I), (II), (III), or (IV) and methods of using the compounds, e.g., in the treatment of immune disorders, inflammatory disorders, and cancer.

For formula I, formula II, formula III, and formula (IV):

W is N or CR^5;

Y is N or CR^2;

Z is N or CR^3;

wherein W and Z are not both N;

R^1 is selected from the group consisting of cyano, hydroxyl, NR^aR^b, C_{1-6}alkoxy, and -A-L^1-R^9;

R^2, R^3, R^4, and R^6 are each independently selected from the group consisting of hydrogen, deuterium, halogen, cyano, hydroxyl, —NR^aR^b, C_{1-6}alkyl, C_{2-6}alkenyl, C_{2-6}alkynyl, C_{1-6}haloalkyl, and C_{1-6}alkoxy;

R^5 is selected from the group consisting of hydrogen, C_{1-6}alkyl, C_{2-6}alkenyl, C_{2-6}alkynyl, C_{1-6}haloalkyl, aryl, heteroaryl, cycloalkyl, heterocyclyl, -aryl-C_{1-6}alkyl, -heteroaryl-C_{1-6}alkyl, -heterocyclyl-C_{1-6}alkyl, halogen, cyano, hydroxyl, C_{1-6}alkoxy, C_{1-6}haloalkoxy, amino, carboxy, aminocarbonyl, —C_{1-6}alkyl-aminocarbonylamino, C_{1-6}alkyl-aminocarbonyl, —S(O)—R^8, —$S(O)_2$—R^8, —NR^8—$S(O)_2$—R^8, —$S(O)_2$—NR^aR^b, —NR^8—$S(O)_2$—NR^aR^b, —C_{1-6}alkyl-aryl, —C_{1-6}alkyl-heteroaryl, —C_{1-6}alkyl-heterocycle, and —C_{1-6}alkyl-cycloalkyl, wherein said alkyl, aryl, and heteroaryl is optionally substituted with one or substituents independently selected from the group consisting of halo, hydroxyl, methoxy, amino, cyano, alkylamino, dialkylamino, CF_3, aminocarbonyl, —C_{1-6}alkyl-aminocarbonylamino, and C_{3-6}cycloalkyl;

R^7 is B-L^2-R^{10}, or R^7 is aryl or heteroaryl, wherein the aryl or heteroaryl is optionally substituted with one to four R^{17};

each R^8 is independently selected from the group consisting of hydrogen, C_{1-6}alkyl, C_{1-6} haloalkyl, hydroxy, C_{1-6}alkoxy, and —O—C_{1-6}haloalkyl;

R^9 is selected from the group consisting of hydrogen, cycloalkyl, heterocycloalkyl, aryl, and heteroaryl, wherein any non-hydrogen R^9 is optionally substituted with one to four R^{17};

R^{10} is selected from the group consisting of hydrogen, substituted by 1-3 substituents independently selected from the group consisting of halogen, C_{1-6}alkyl, and C_{1-6}haloalkyl;

R^{17} is, independently for each occurrence, selected from the group consisting of halogen, cyano, hydroxyl, —NR^aR^b, C_{1-6}alkyl, C_{2-6}alkenyl, C_{2-6}alkynyl, C_{1-6}haloalkyl, C_{1-6}alkoxy, CF_3, —SH, —S—C_{1-6}alkyl, —COOH, —CO_2—C_{1-6}alkyl, —C_{1-6}alkyl-CN, —C(O)NR^aR^b, —C(O)—C_{1-6} alkyl-NR^aR^b, —C(O)—NR^a—$S(O)_2$—C_{1-6}alkyl, —$S(O)_2$—C_{1-6}alkyl, —$S(O)_2$—NR^aR^b, —$S(O)_2$—C_{1-6} alkyl-NR^aR^b;

A is selected from the group consisting of —C(O)—, —S(O)—, and —$S(O)_2$—, or A is absent;

B is selected from the group consisting of —C(O)—, —$S(O)_2$—NR^8—, —CH_2—NR—, and —C(O)NR^8—;

L^1 is selected from the group consisting of a bond, C_{1-6}alkylene, C_{1-6}heteroalkylene, C_{2-6}alkenylene, and C_{2-6}alkynylene, wherein L^1 is optionally substituted with one to four R^{17} groups;

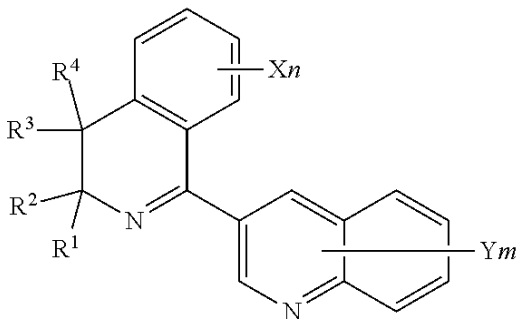
L^2 is selected from the group consisting of a bond, C_{1-6}alkylene, C_{2-6}alkenylene, and C_{2-6}alkynylene, wherein any CH_2 group of C_{1-6}alkylene can be replaced with a moeity selected from the group consisting of —O—, —NR^a—, and —$S(O)_2$—, and one CH_2 group of C_{1-6} alkylene can be replaced with a moiety selected from the group consisting of cycloalkylene, heterocycloalkylene, arylene, and heteroarylene, and wherein L^2 is optionally substituted with one to four R^{17} groups; or

when B is —$S(O)_2$—NR^8—, —CH_2—NR^8—, or —C(O)NR^8—, R^8 and L^2 can be taken together including the nitrogen atom to which they are attached to form a 3-7-membered heterocycloalkyl optionally substituted with one to four R^{17} groups; and

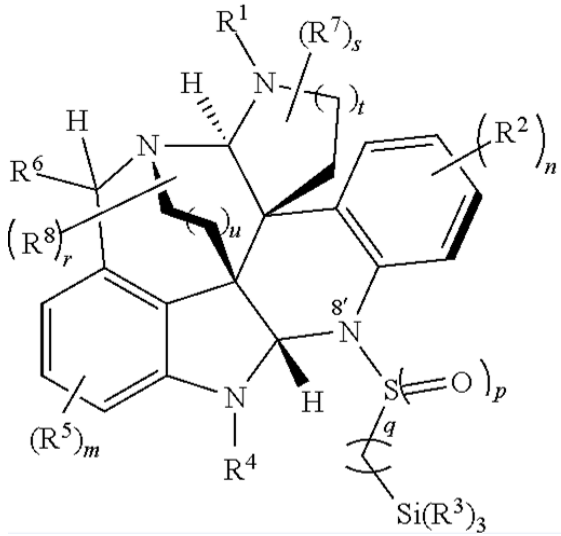
each of R^a and R^b are, independently for each occurrence, selected from the group consisting of hydrogen, C_{1-6}alkyl, and C_{1-6}haloalkyl, or R^a and R^b are taken together, including the nitrogen to which they are attached, to form a heterocycloalkyl ring.

Markush structures: the art of chemistry

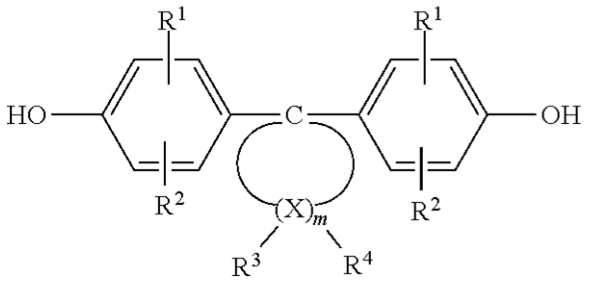
US20210000112A1



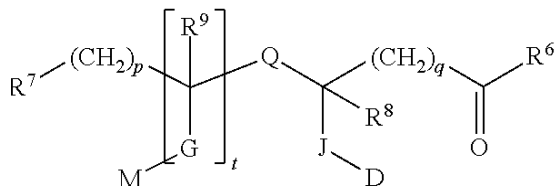
US20210000119A1



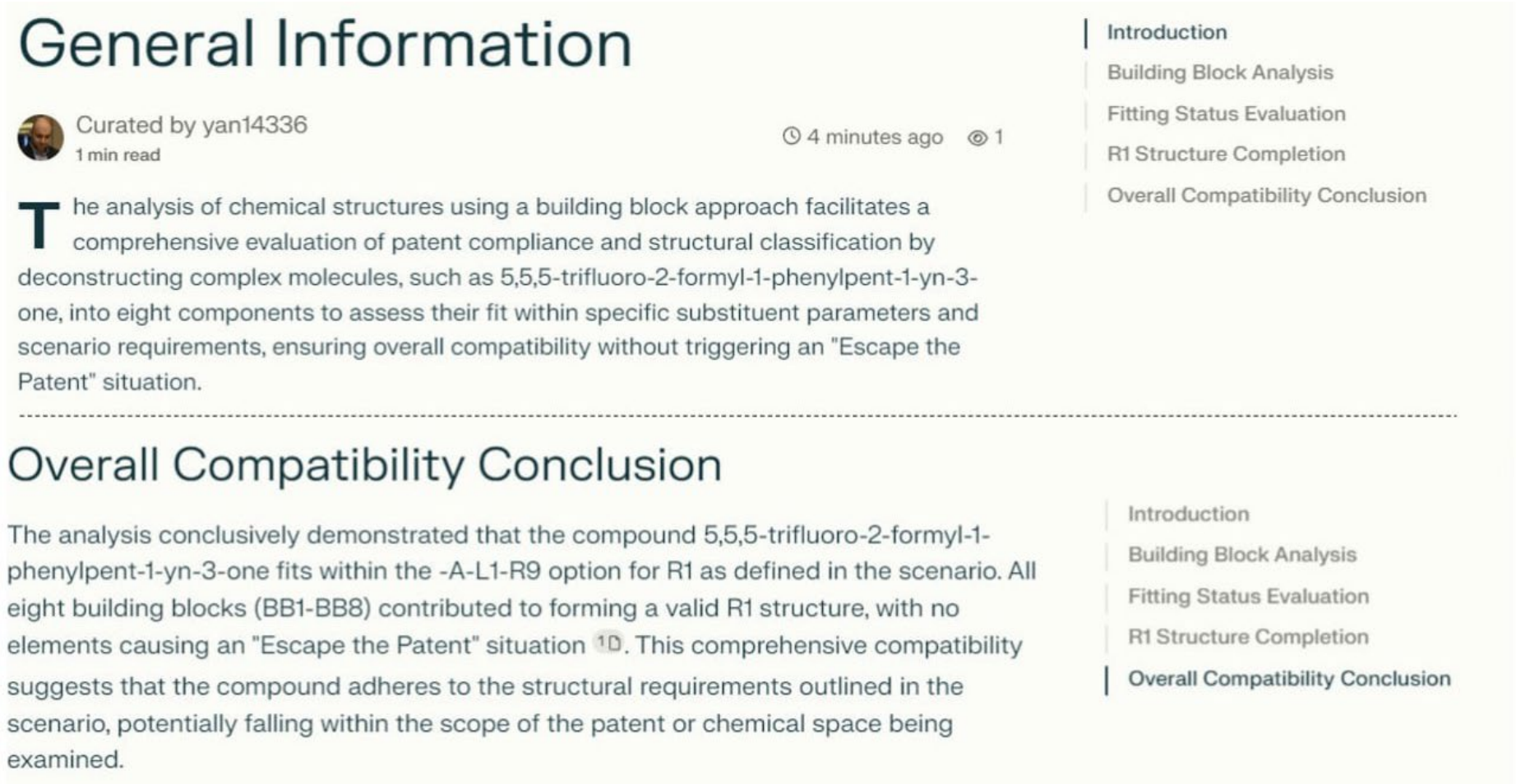
US20210001614A1



US20200360258A1



Output of the Patent Busters



General Information

Curated by yan14336
1 min read

4 minutes ago 1

The analysis of chemical structures using a building block approach facilitates a comprehensive evaluation of patent compliance and structural classification by deconstructing complex molecules, such as 5,5,5-trifluoro-2-formyl-1-phenylpent-1-yn-3-one, into eight components to assess their fit within specific substituent parameters and scenario requirements, ensuring overall compatibility without triggering an "Escape the Patent" situation.

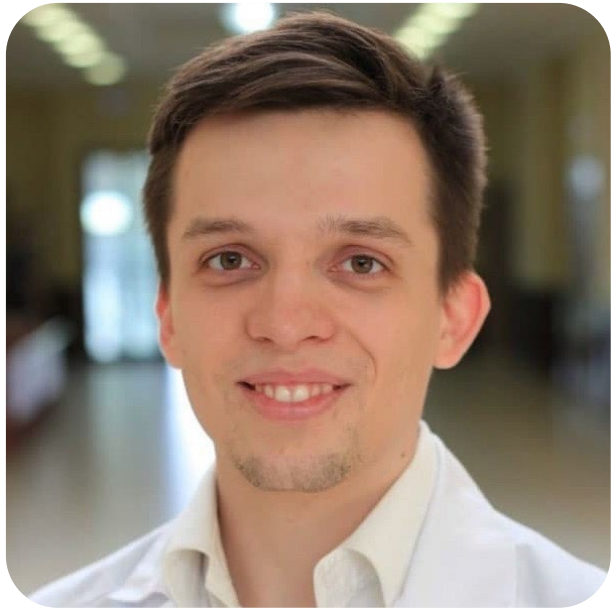
Introduction
Building Block Analysis
Fitting Status Evaluation
R1 Structure Completion
Overall Compatibility Conclusion

Overall Compatibility Conclusion

The analysis conclusively demonstrated that the compound 5,5,5-trifluoro-2-formyl-1-phenylpent-1-yn-3-one fits within the -A-L1-R9 option for R1 as defined in the scenario. All eight building blocks (BB1-BB8) contributed to forming a valid R1 structure, with no elements causing an "Escape the Patent" situation 1D. This comprehensive compatibility suggests that the compound adheres to the structural requirements outlined in the scenario, potentially falling within the scope of the patent or chemical space being examined.

Introduction
Building Block Analysis
Fitting Status Evaluation
R1 Structure Completion
Overall Compatibility Conclusion

Thank you!



Alex Malyshev

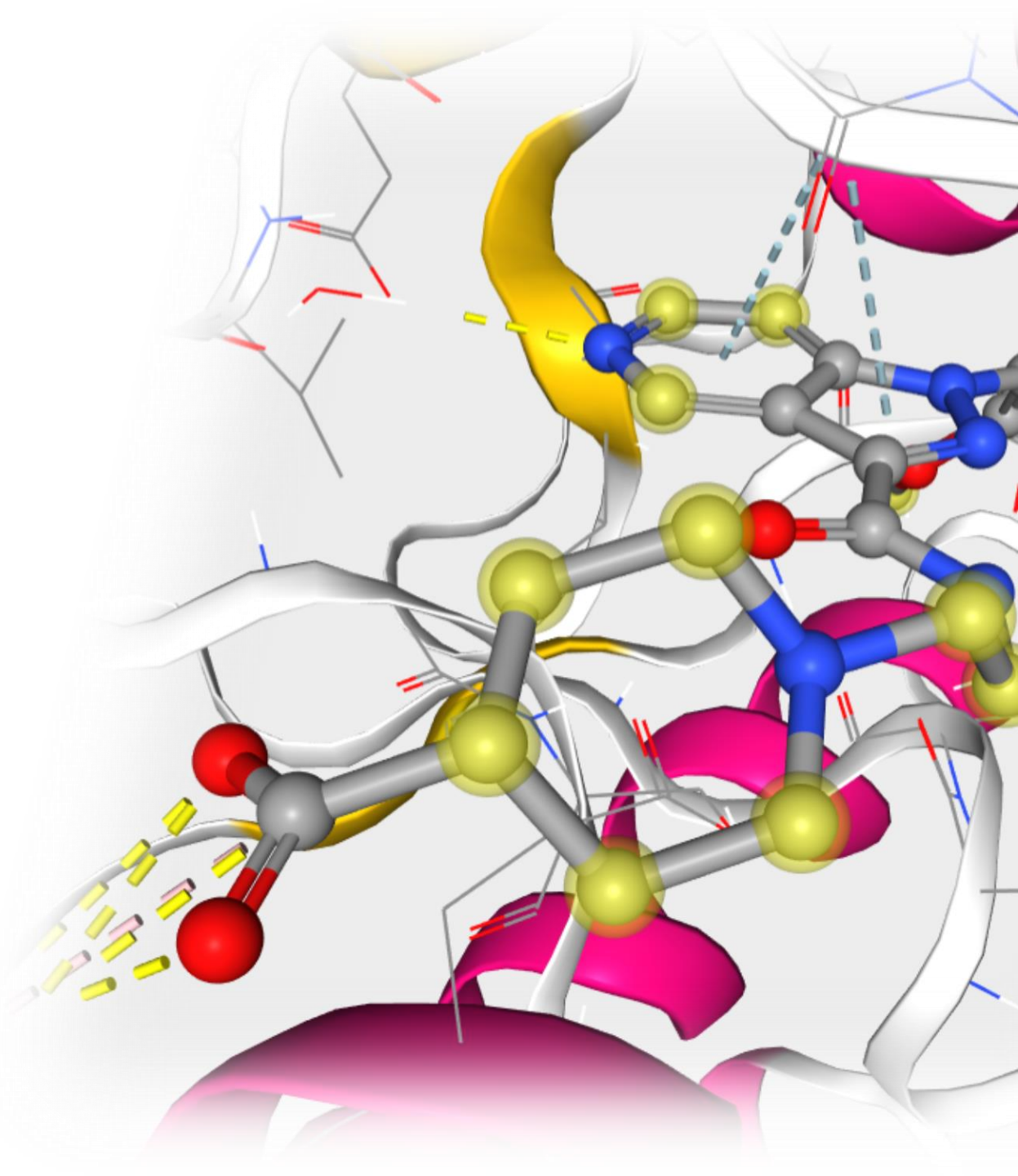
@ALEXMALYSHEV95

alexmalyshev95@gmail.com

Structure-based Drug Design

Sergei Evteev
Lead Scientist
FSUE VNIIA

Fall into ML 2024

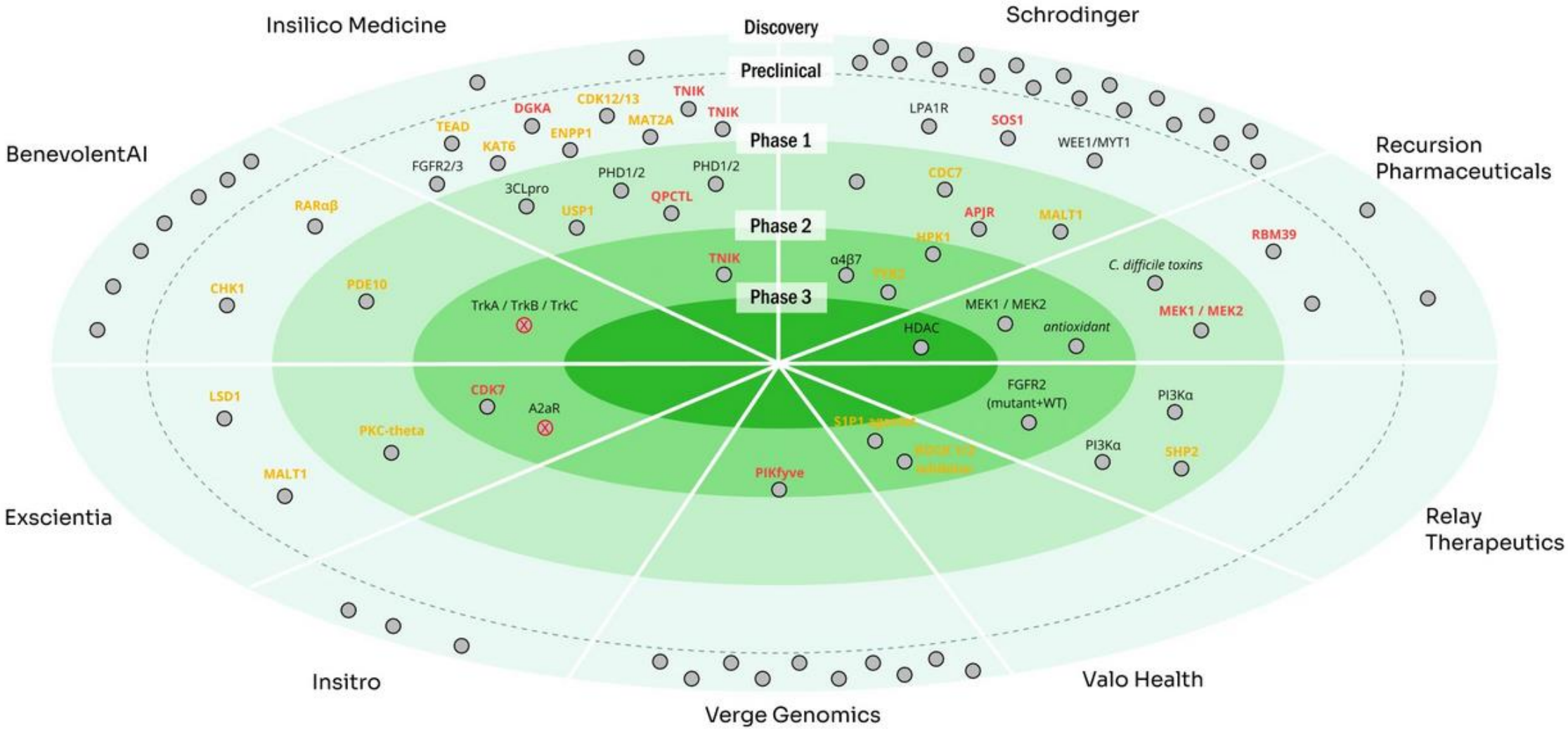


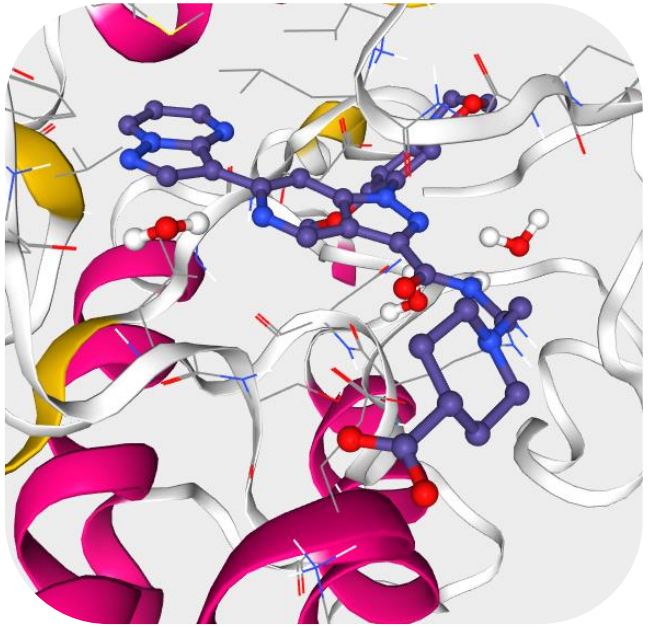
The Landscape of AI-discovered Drug Candidates and Targets

The indicated data are for 2023

Legend

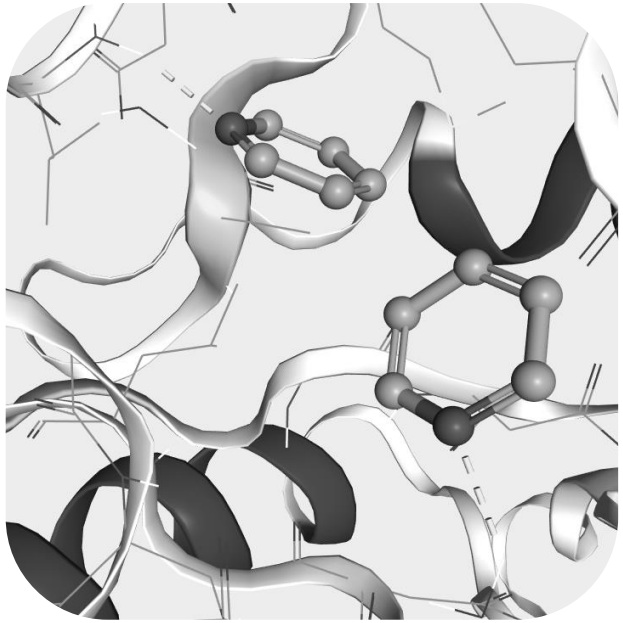
Target Novelty	Low	Moderate	High
Pipeline	Active	molecule failed or discontinued	



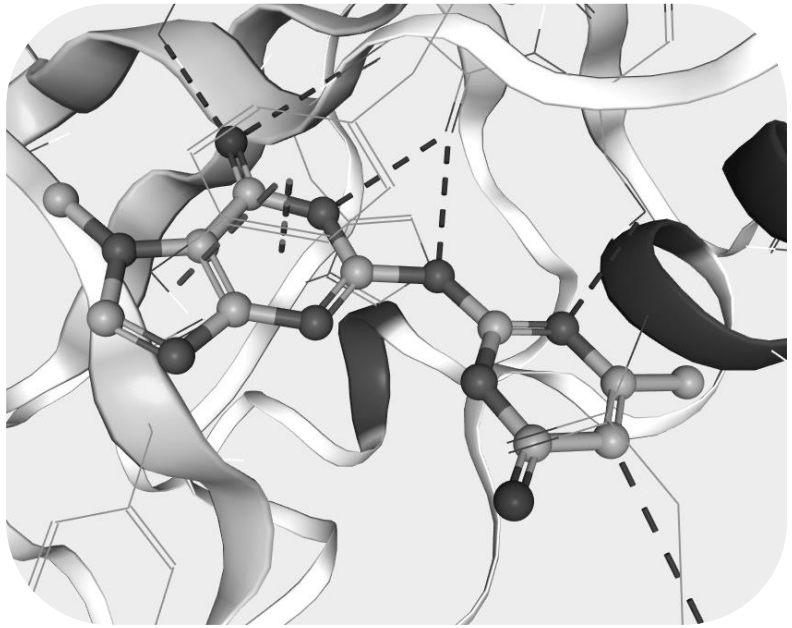


Model preparation

Binding site identification



Hot spots identification



de novo generation

Russian Chemical Reviews, 2024, Volume 93, Issue 3, RCR5107
DOI: https://doi.org/10.59761/RCR5107

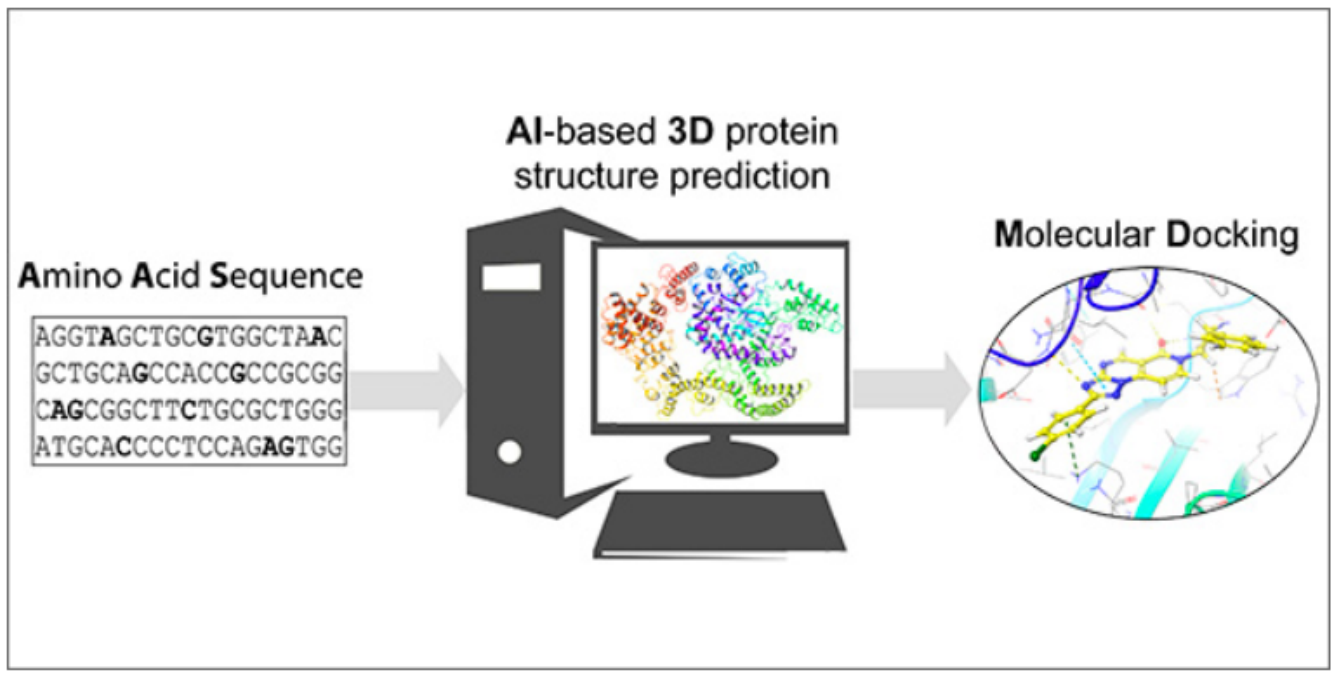
AlphaFold for a medicinal chemist: tool or toy?

Ya. A. Ivanenkov[ab], S. A. Evteev[ab], A. S. Malyshev[ab], V. A. Terentiev[ab], D. S. Bezrukov[c], A. V. Ereshchenko[ab], A. A. Korzhenevskaya[a], B. A. Zagribelnyy[c], P. V. Shegai[a], A. D. Kaprin[da]

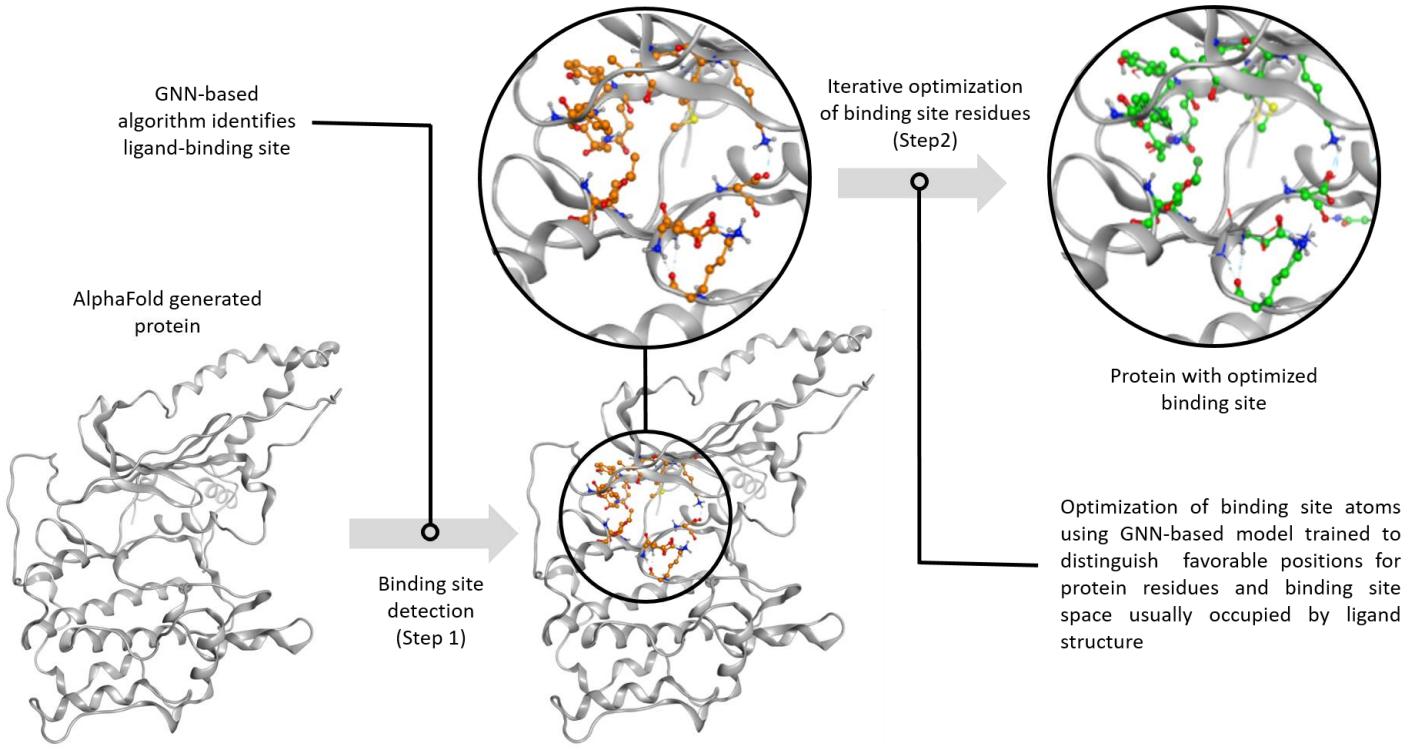
[a] *P.Hertsen Moscow Oncology Research Institute, Moscow, Russian Federation*
[b] *Dukhov Automatics Research Institute (VNIIA), Moscow, Russian Federation*
[c] *Department of Chemistry, Lomonosov Moscow State University, Moscow, Russian Federation*
[d] *Peoples' Friendship University of Russia (RUDN), Moscow, Russian Federation*

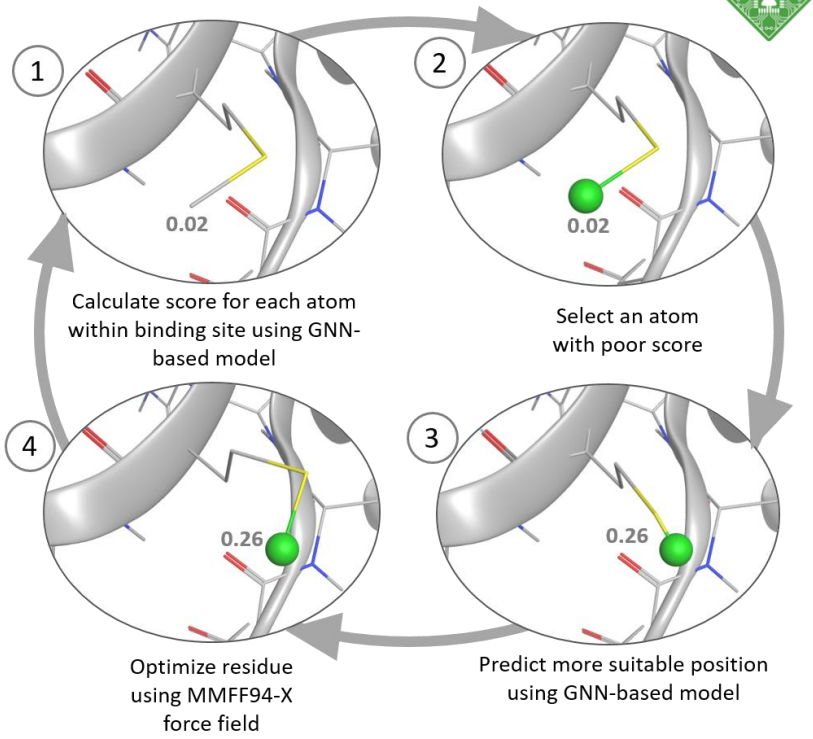
English full-text

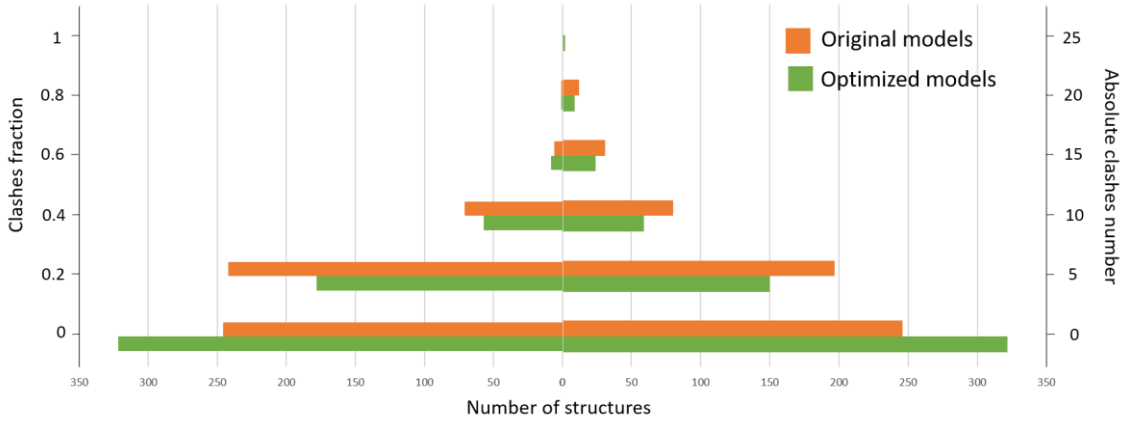
DOI: https://doi.org/10.59761/RCR5107

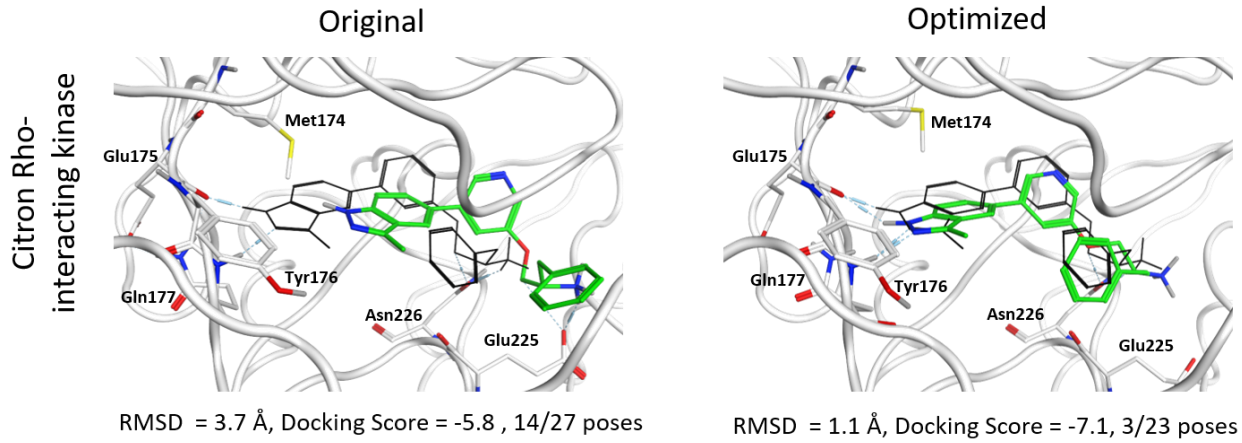


AlphaFold Optimizer

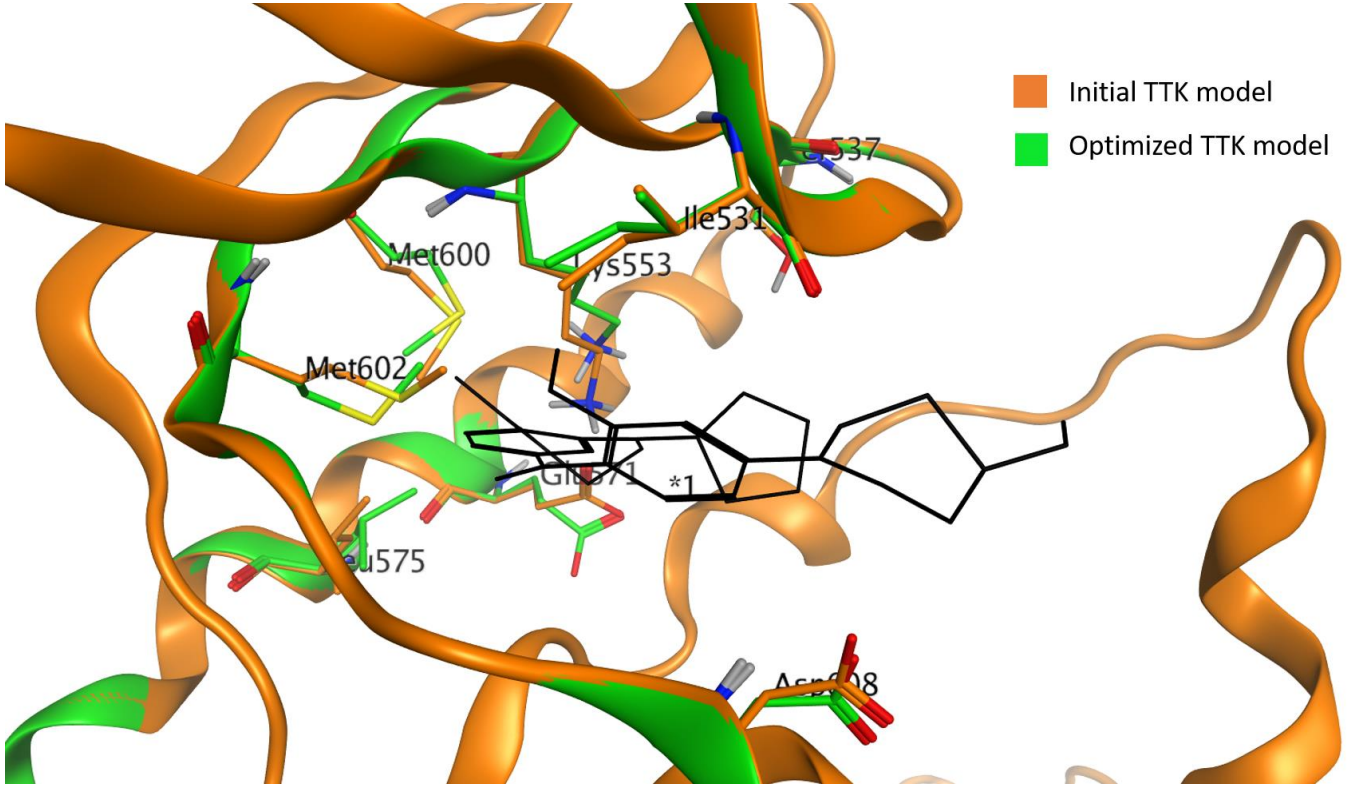




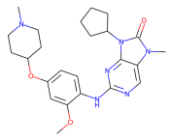
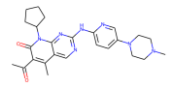
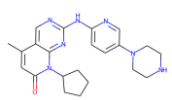
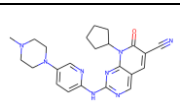
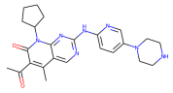
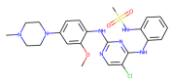
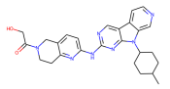
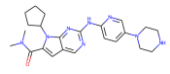
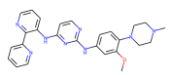
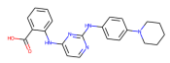


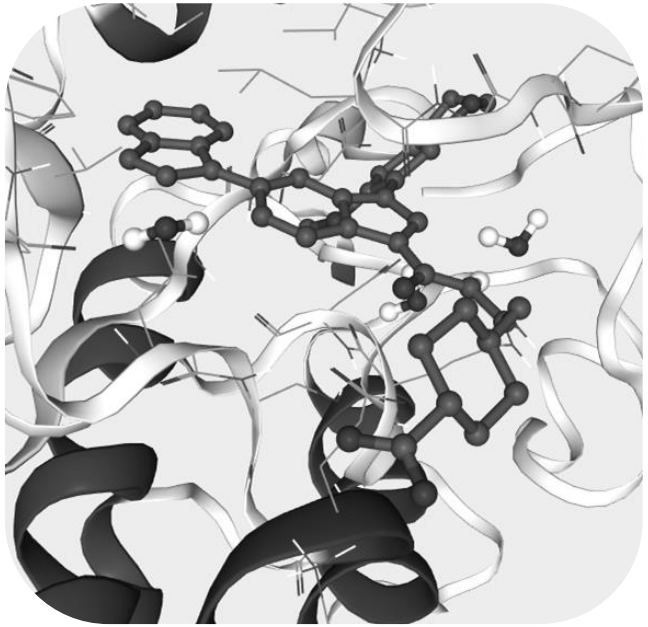


AlphaFold Optimizer

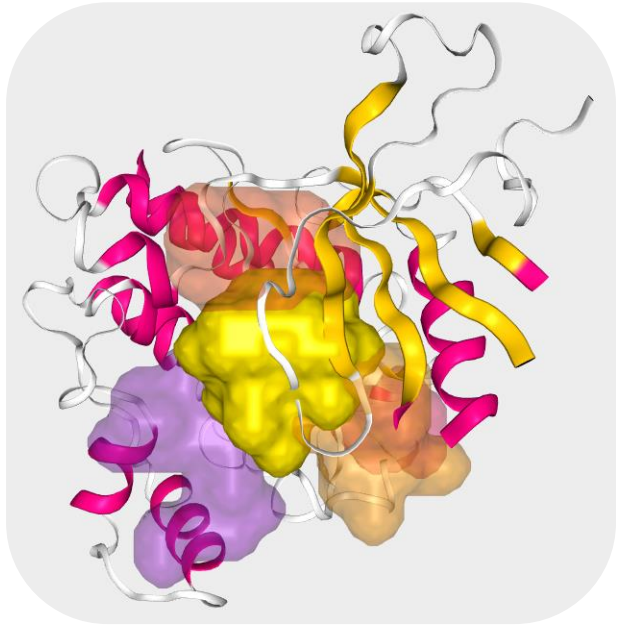


- Virtual screening was performed using both original and optimized Alphafold models
- Among 39 compounds tested, 10 hits were found
- All 10 hits were detected using optimized model while only 7 of them were detected using original AlphaFold structure

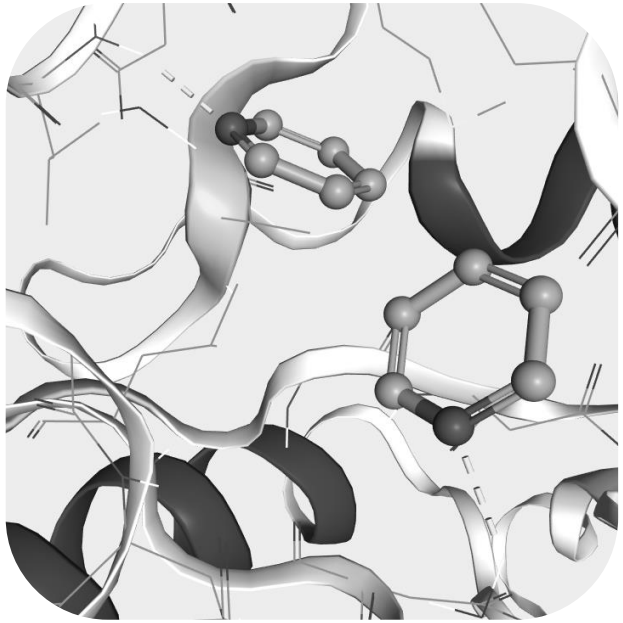
Compound ID	Structure	Inhibition, %	Predicted by initial model	Predicted refined model
1		94	+	+
2		84	-	+
3		93	+	+
4		92	-	+
5		86	+	+
6		84	+	+
7		100	+	+
8		51	-	+
9		54	+	+
10		70	+	+



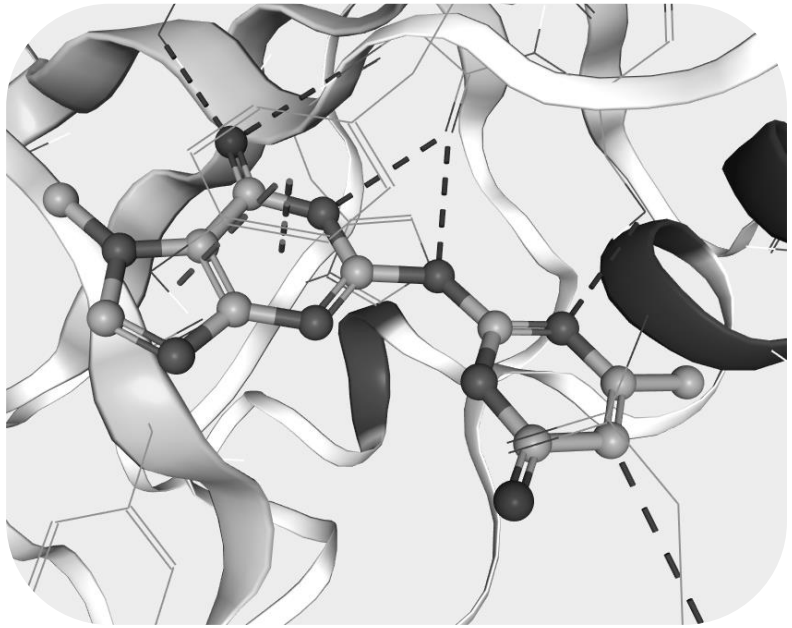
Model preparation



Binding site identification



Hot spots identification



de novo generation

Journal of Chemical Information and Modeling > Vol 63/Issue 4 > Article

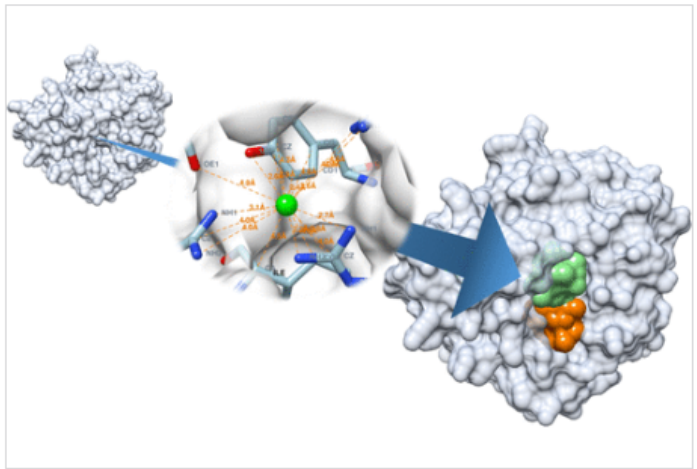
MACHINE LEARNING AND DEEP LEARNING | February 6, 2023

SiteRadar: Utilizing Graph Machine Learning for Precise Mapping of Protein–Ligand-Binding Sites

Sergei A. Evteev*, Alexey V. Ereshchenko, and Yan A. Ivanenkov

Abstract

Identifying ligand-binding sites on the protein surface is a crucial step in the structure-based drug design. Although multiple techniques have been proposed, including those using machine learning algorithms, the existing solutions do not provide significant advantages over nonmachine learning approaches and there is still a big room for improvement. The low ability to identify protein–ligand-binding sites makes available approaches inapplicable to automated drug design. Here, we present SiteRadar, a new algorithm for mapping cavities that are likely to bind a small-molecule ligand. SiteRadar shows higher accuracy in binding site identification compared with FPocket and PUResNet. SiteRadar demonstrates an ability to detect up to 74% of true ligand-binding sites according to the top N + 2 metric and usually covers approximately 80% of ligand atoms. Therefore, SiteRadar can be regarded as a promising solution for implementation into algorithms for automated drug design.



Journal of Chemical Information and Modeling
Cite this: *J. Chem. Inf. Model.* 2023, 63, 4, 1124–1132

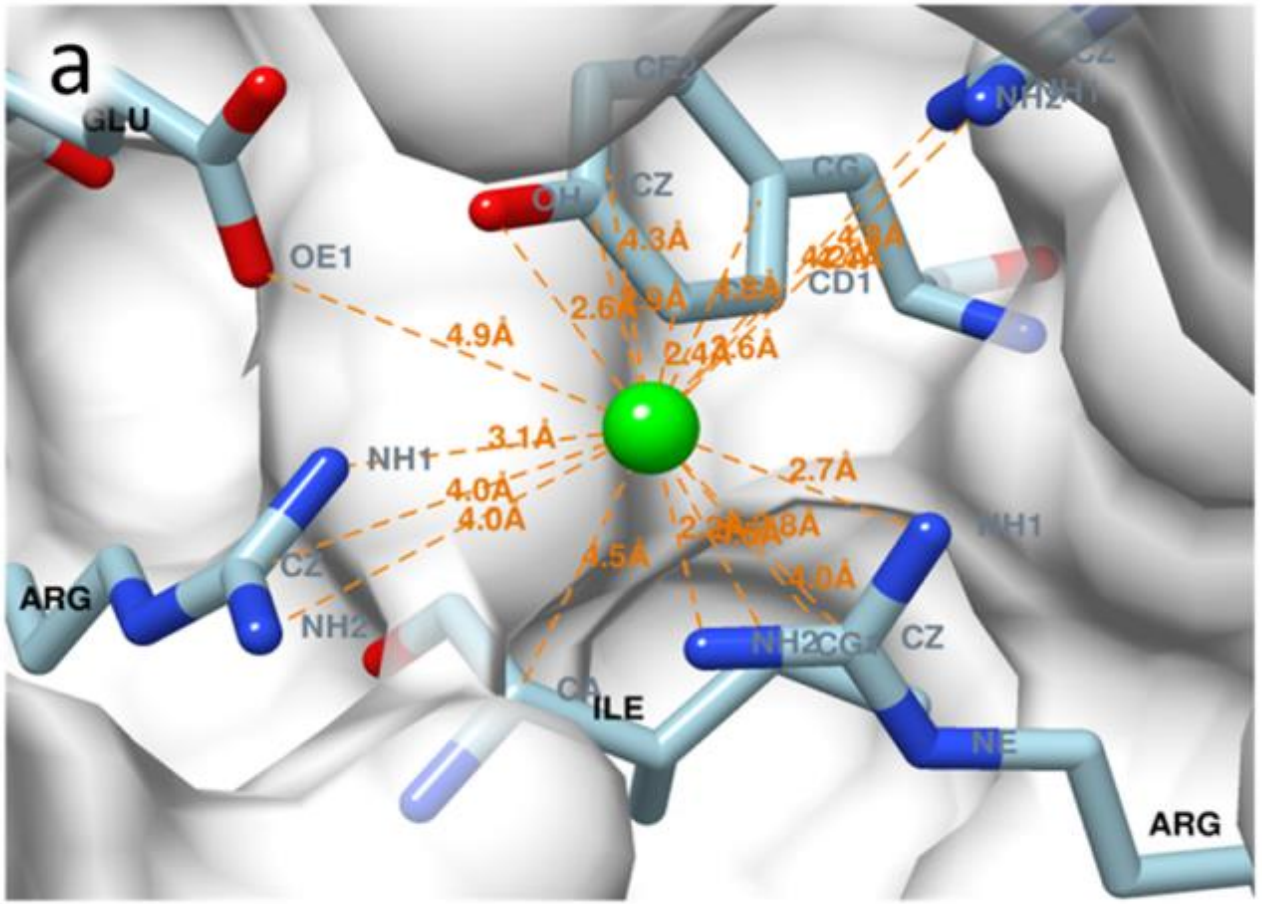
https://doi.org/10.1021/acs.jcim.2c01413
Published February 6, 2023

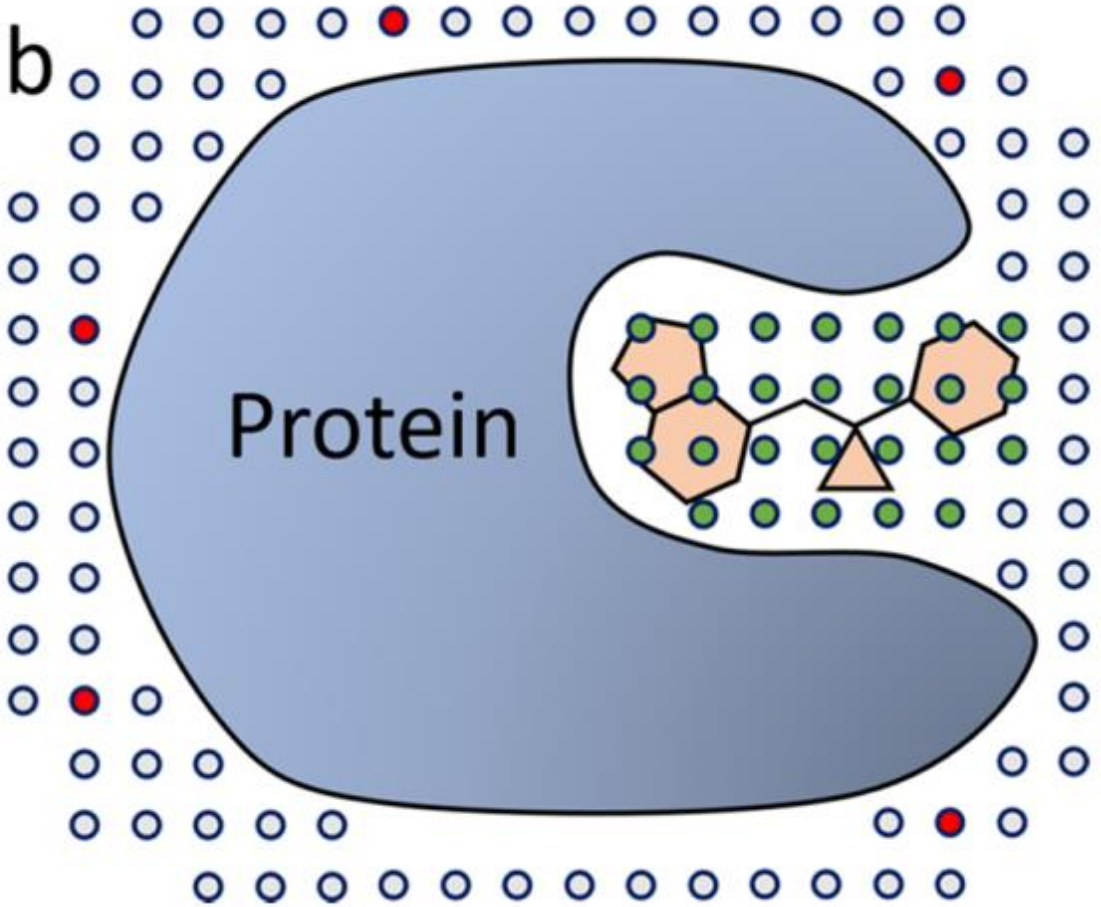
Recommended Articles

DeepPocket: Ligand Binding Site Detection and Segmentation using 3D Convolutional Neural Networks
August 10, 2021 | *Journal of Chemical Information and Modeling*
Rishal Aggarwal, Akash Gupta, Vineeth Chelur, C. V. Jawahar, and U. Deva...

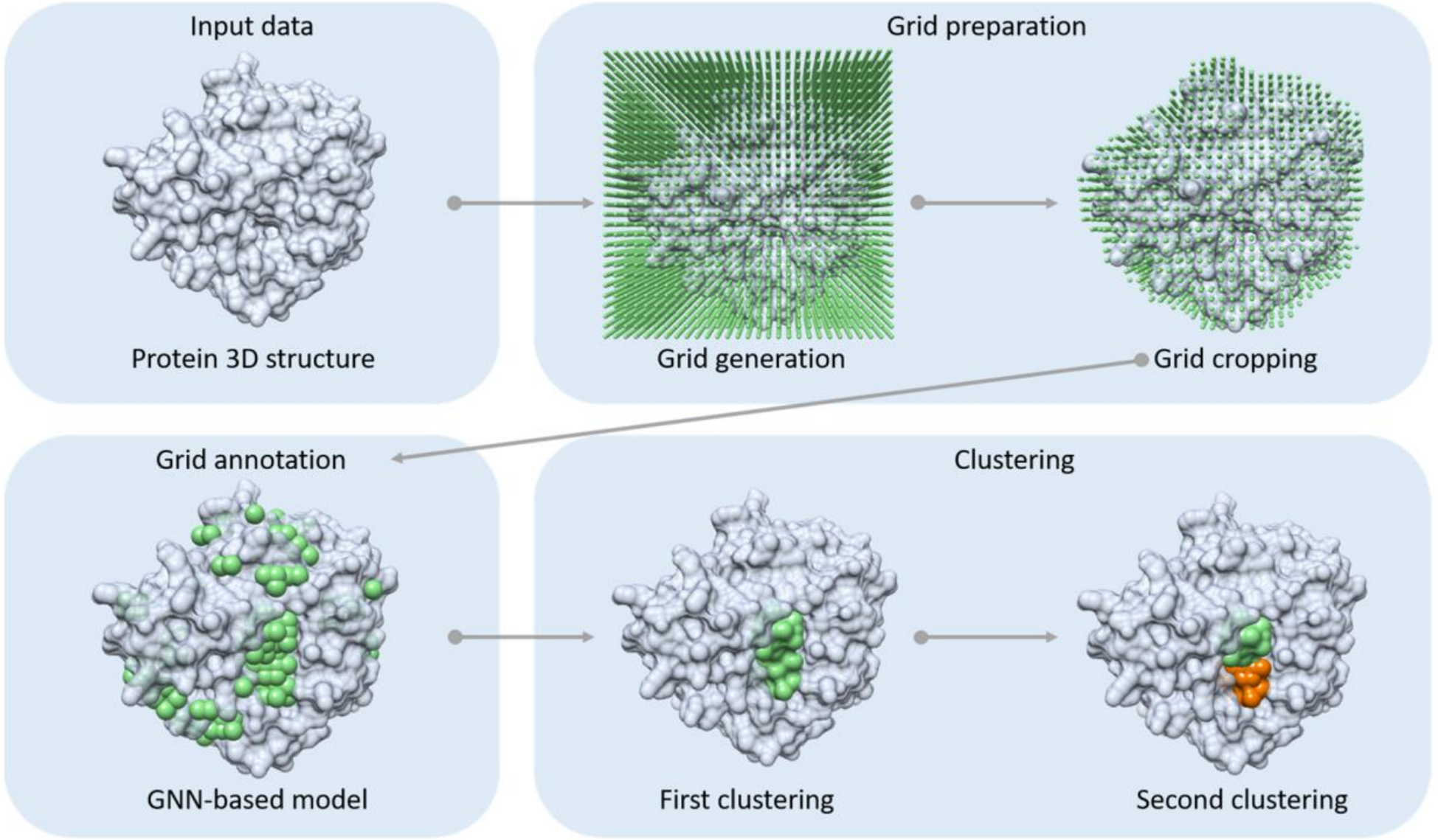
PLANET: A Multi-objective Graph Neural Network Model for Protein–Ligand Binding Affinity Prediction
June 15, 2023 | *Journal of Chemical Information and Modeling*
Xiangying Zhang, Haotian Gao, Haojie Wang, Zhihang Chen, Zhe Zhang, ...

Graph-based approach





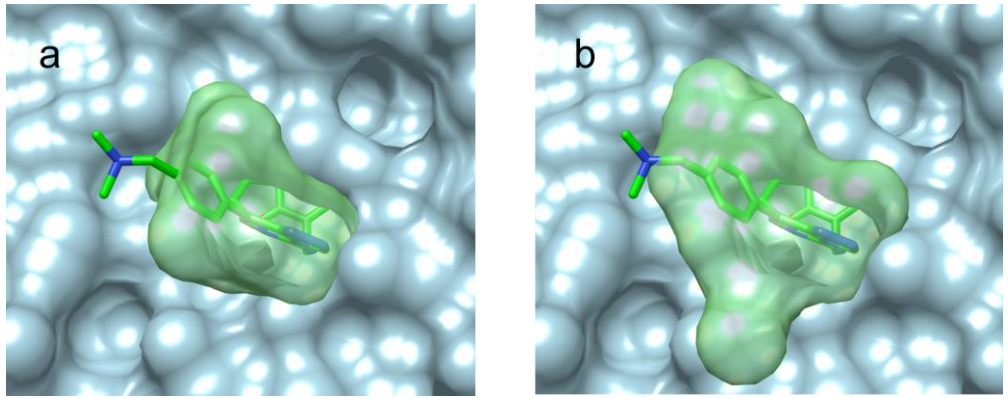
SiteRadar Pipeline



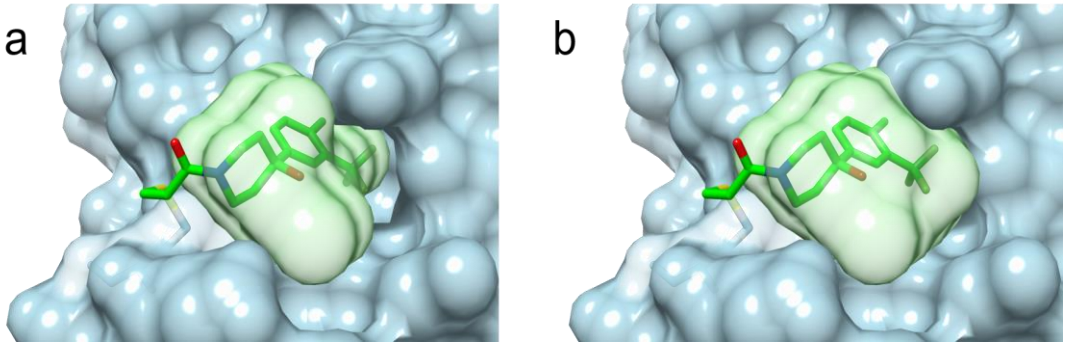
Evteev S.A. et al. SiteRadar: Utilizing Graph Machine Learning for Precise Mapping of Protein-Ligand-Binding Sites. J. Chem. Inf. Model. 2023

Case studies

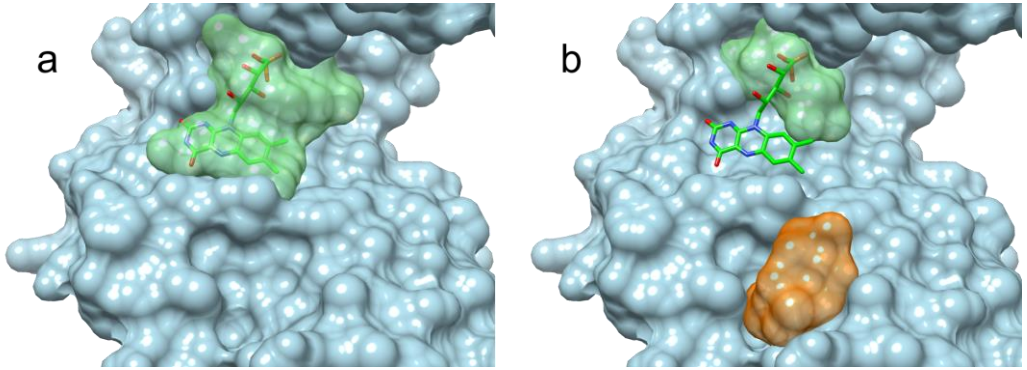
Allosteric binding site



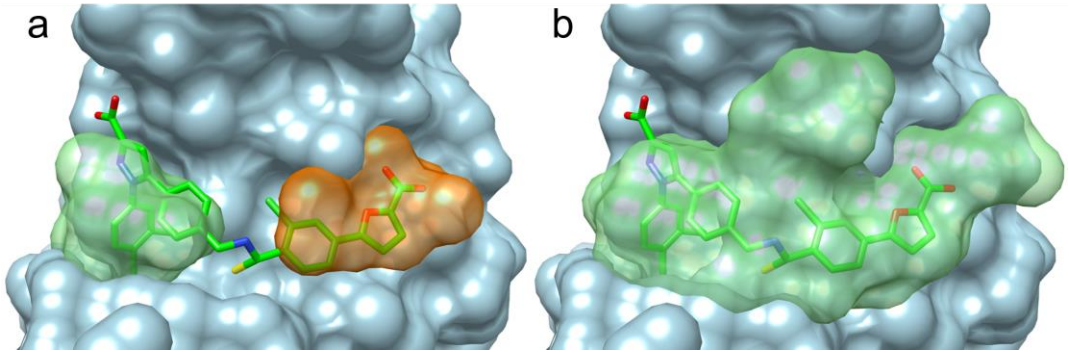
Site for covalent ligand binding



Solvent-exposed binding site



Protein-protein interaction



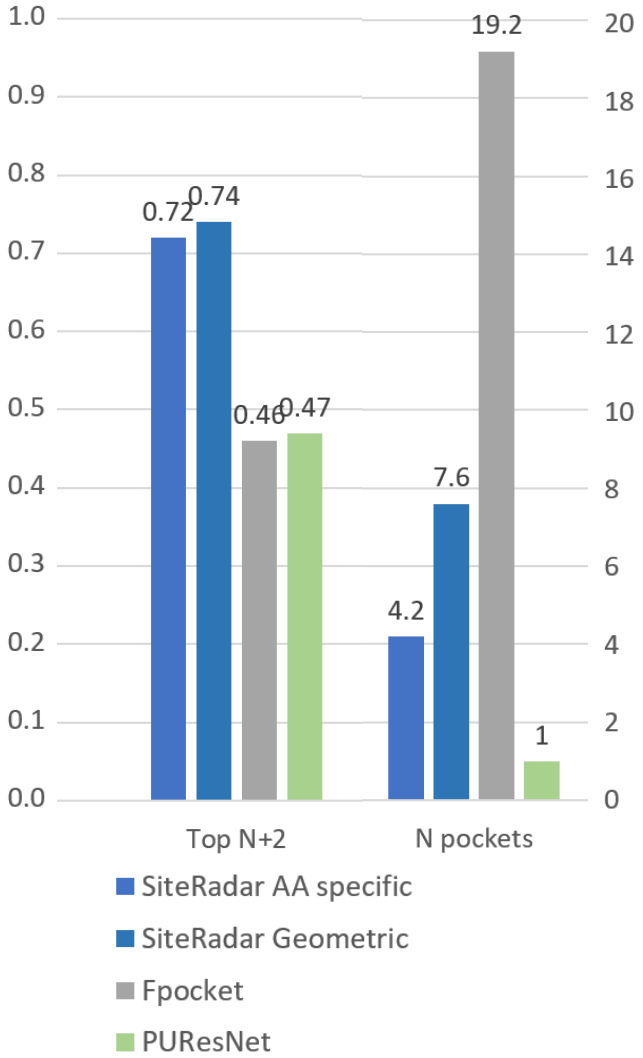
a - AA specific

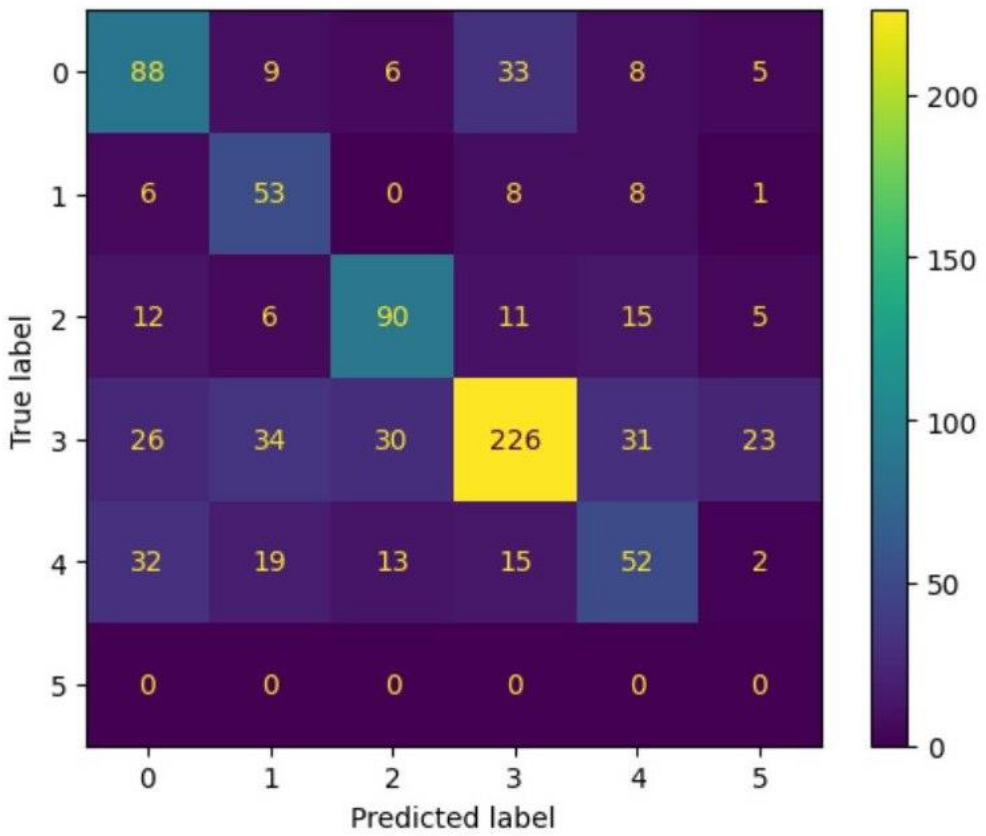
b - Geometric

Evteev S.A. et al. SiteRadar: Utilizing Graph Machine Learning for Precise Mapping of Protein-Ligand-Binding Sites. J. Chem. Inf. Model. 2023


Elements
Cartoon
Input parameters
View
Pockets
Merge mode
Extend size, Å
0 2 4 6
Pocket characteristic
Druggability 0.89
0.01 Nucleic Acid
Allosteric 0.04
0.12 Peptide
Covalent 0.57
0.88 Nucleotide/ATP
Carbohydrate 0.28
Heme 0

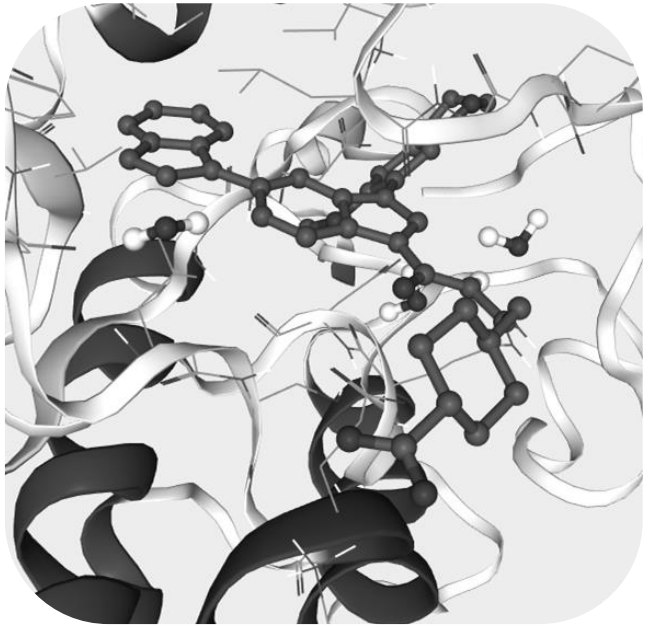
in silico validation





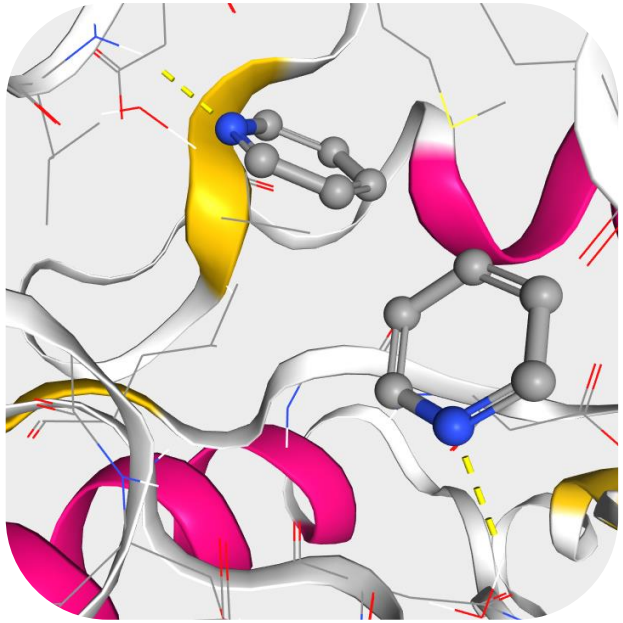
0 – Nucleic acids
1 – Carbohydrates
2 – Heme
3 – Nucleotides / ATP
4 – Peptides
5 – Undefined

Evteev S.A. et al. SiteRadar: Utilizing Graph Machine Learning for Precise Mapping of Protein-Ligand-Binding Sites. J. Chem. Inf. Model. 2023

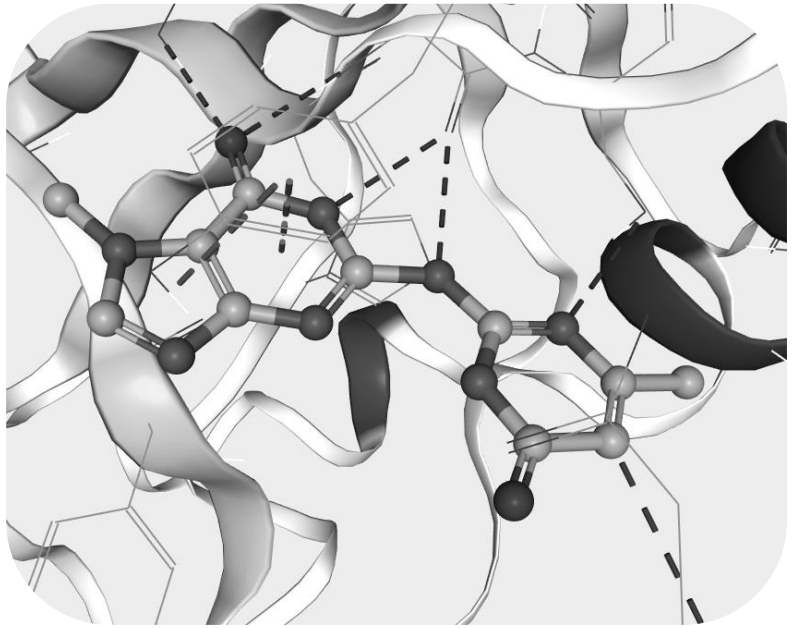


Model preparation

Binding site identification



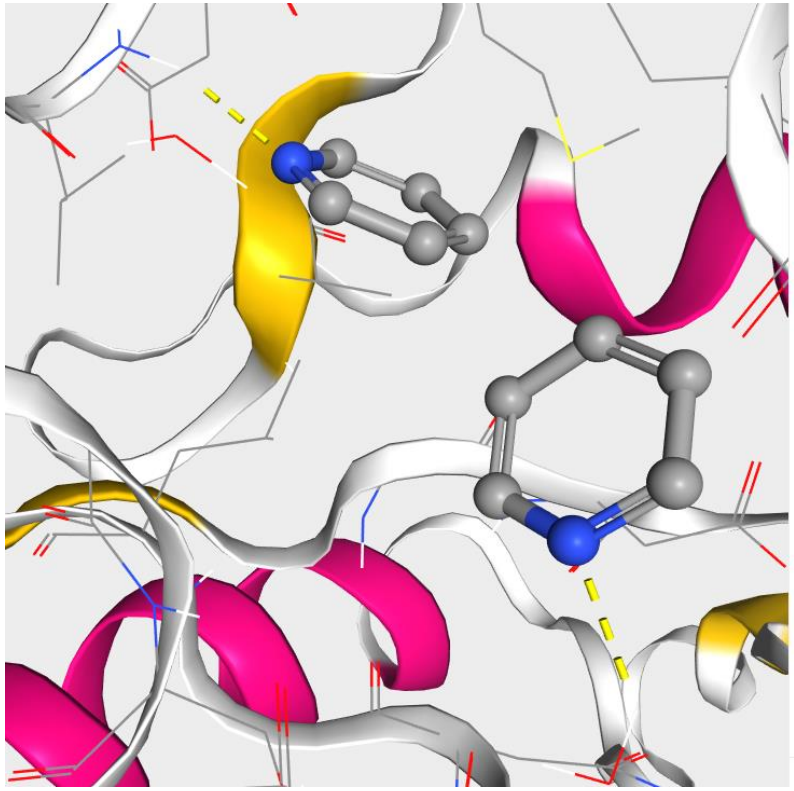
Hot spots identification



de novo generation

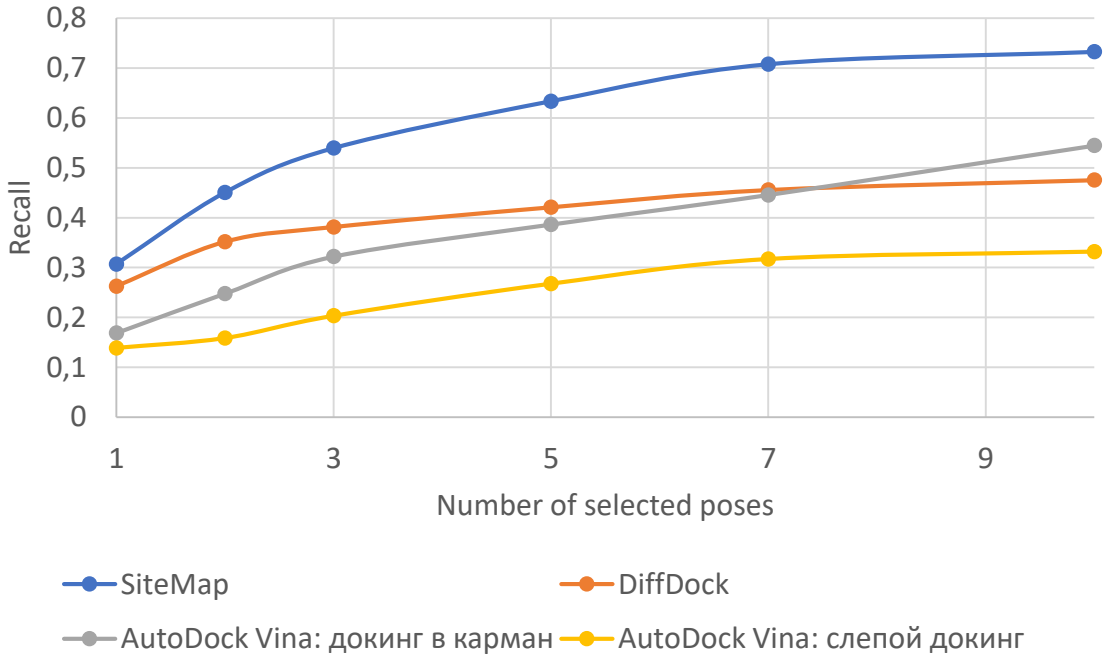
SiteMap

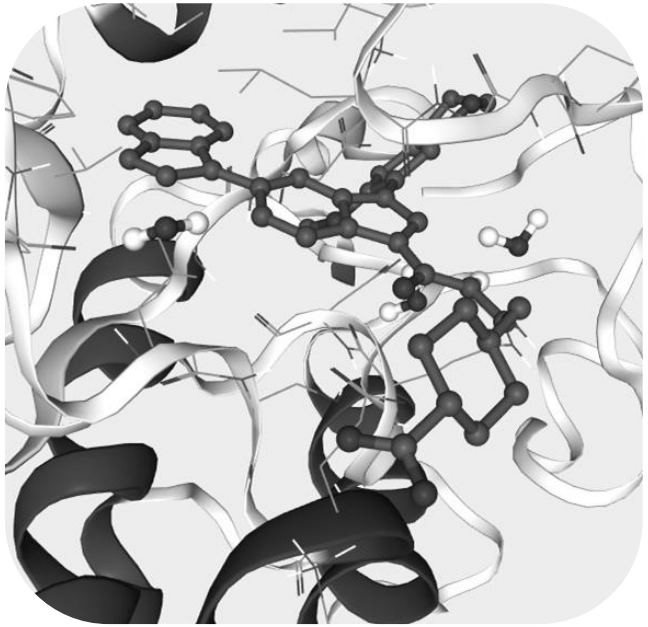
Utilizes traditional docking, diffusion and positional filters



Scaffolds Selected 0 / 3

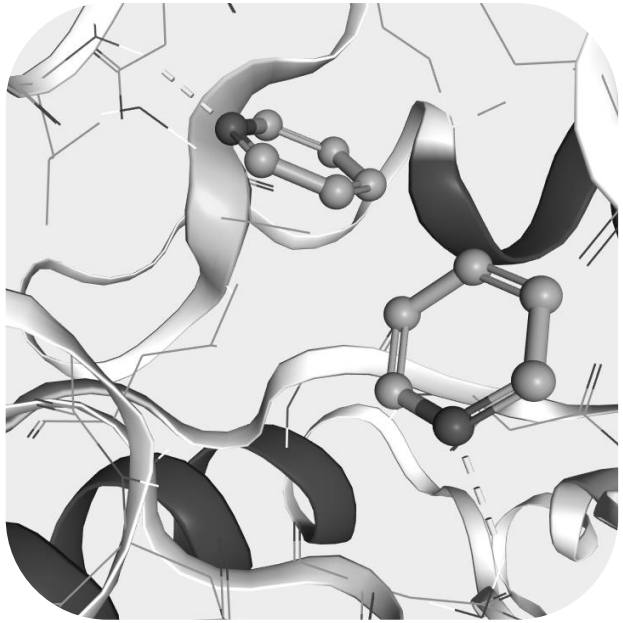
ID	SMILES	Docking score
Mol-0.1	c1ccncc1	-3.95
Mol-0.2	c1ccncc1	-3.70
Mol-0.3	c1ccncc1	-3.69



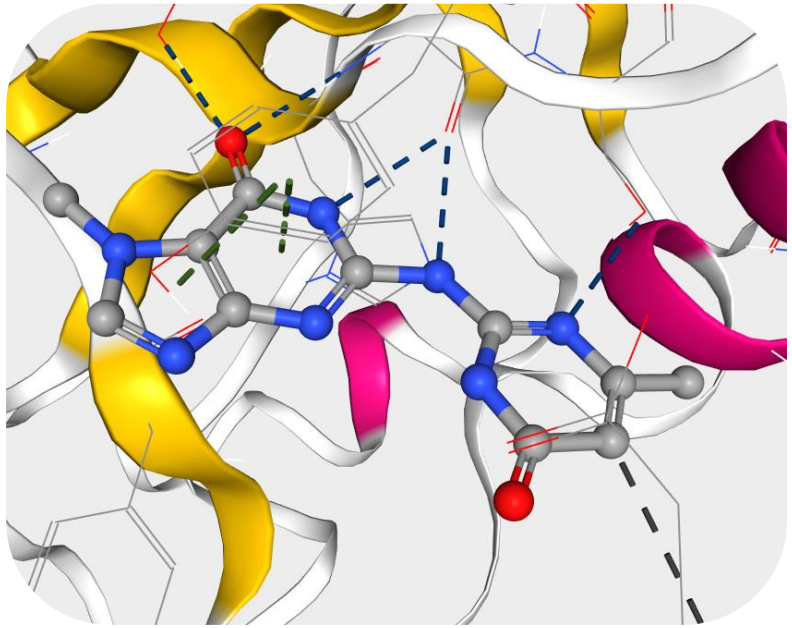


Model preparation

Binding site identification



Hot spots identification



de novo generation

ACS Medicinal Chemistry Letters > Vol 14/Issue 7 > Article

Open Access

Cite Share Jump to Expand

MICROPERSPECTIVES | June 29, 2023

The Hitchhiker's Guide to Deep Learning Driven Generative Chemistry

Yan Ivanenkov, Bogdan Zagribelnyy, Alex Malyshev, Sergei Evteev, Victor Terentiev, Petrina Kamya, Dmitry Bezrukov, Alex Aliper, Feng Ren, and Alex Zhavoronkov*

Open PDF

Abstract

This microperspective covers the most recent research outcomes of artificial intelligence (AI) generated molecular structures from the point of view of the medicinal chemist. The main focus is on studies that include synthesis and experimental *in vitro* validation in biochemical assays of the generated molecular structures, where we analyze the reported structures' relevance in modern medicinal chemistry and their novelty. The authors believe that this review would be appreciated by medicinal chemistry and AI-driven drug design (AIDD) communities and can be adopted as a comprehensive approach for qualifying different research outcomes in AIDD.

ACS Medicinal Chemistry Letters
Cite this: *ACS Med. Chem. Lett.* 2023, 14, 7, 901–915
https://doi.org/10.1021/acsmedchemlett.3c00041
Published June 29, 2023

Get e-Alerts

Article Views	Altmetric	Citations
15k	24	4

Learn about these metrics

Recommended Articles

Chemistry42: An AI-Driven Platform for Molecular Design and Optimization
February 2, 2023 | *Journal of Chemical Information and Modeling*
Yan A. Ivanenkov, Daniil Polykovskiy, Dmitry Bezrukov, Bogdan Zagribelnyy, ...

Generative Models as an Emerging Paradigm in the Chemical Sciences
April 13, 2023 | *Journal of the American Chemical Society*

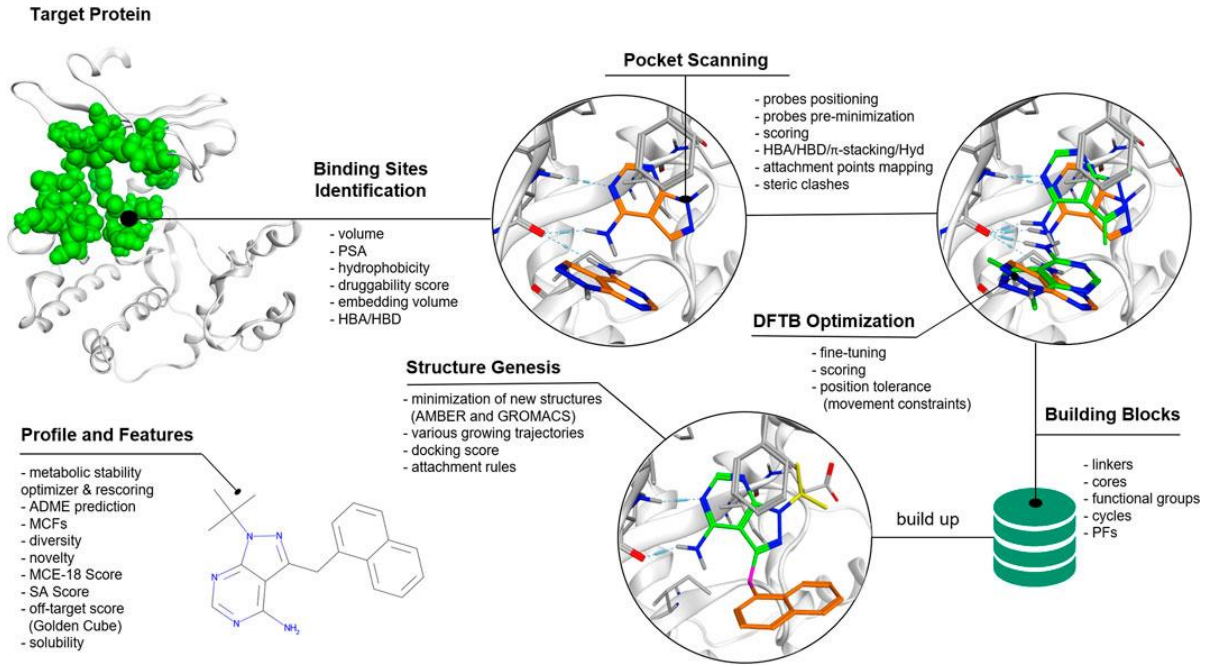
Quantum-assisted fragment-based automated structure generator (QFASG) for small molecule design: an *in vitro* study

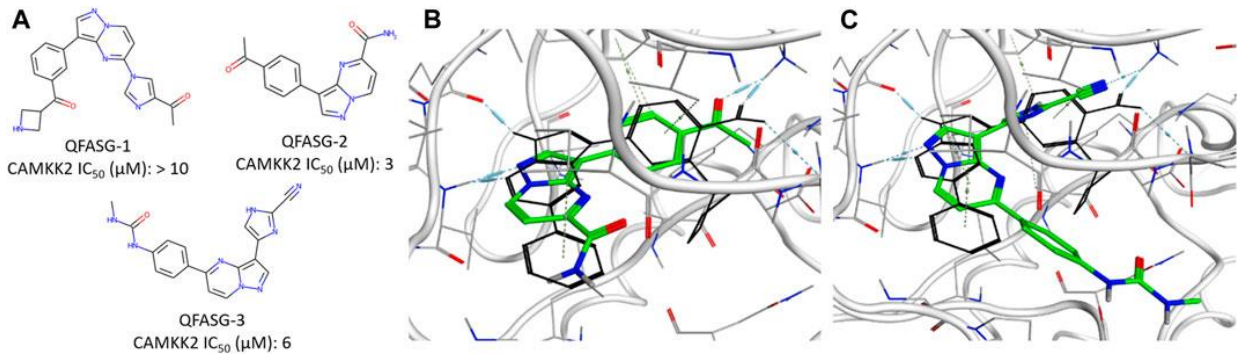
Sergei Evteev[1*] Yan Ivanenkov[1] Ivan Semenov[1] Maxim Malkov[2] Olga Mazaleva[1]

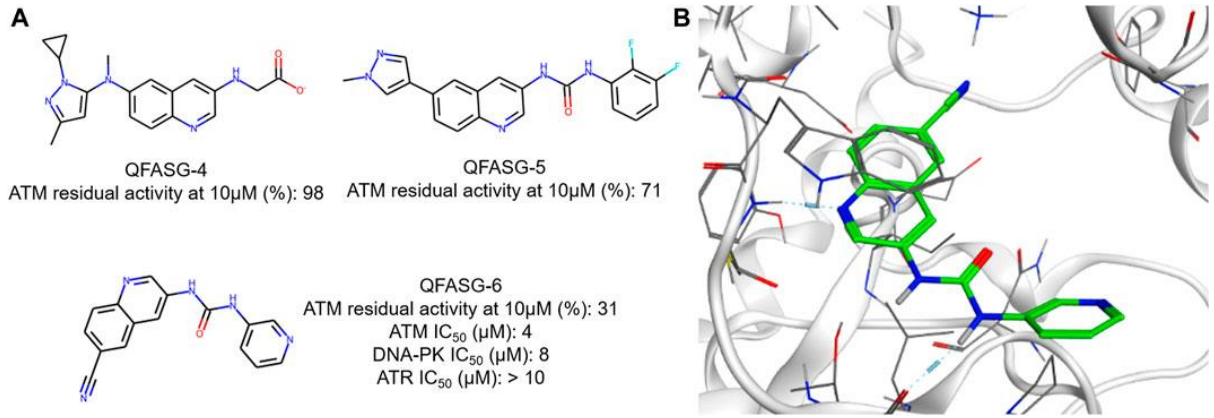
Artem Bodunov[1] Dmitry Bezrukov[2] Denis Sidorenko[1] Victor Terentiev[1] Alex Malyshev[1]

Bogdan Zagribelnyy[2] Anastasia Korzhenevskaya[1] Alex Aliper[2] Alex Zhavoronkov[2]

[1] Insilico Medicine Hong Kong Ltd., Hong Kong, Hong Kong SAR, China
[2] Insilico Medicine AI Limited, Abu Dhabi, United Arab Emirates

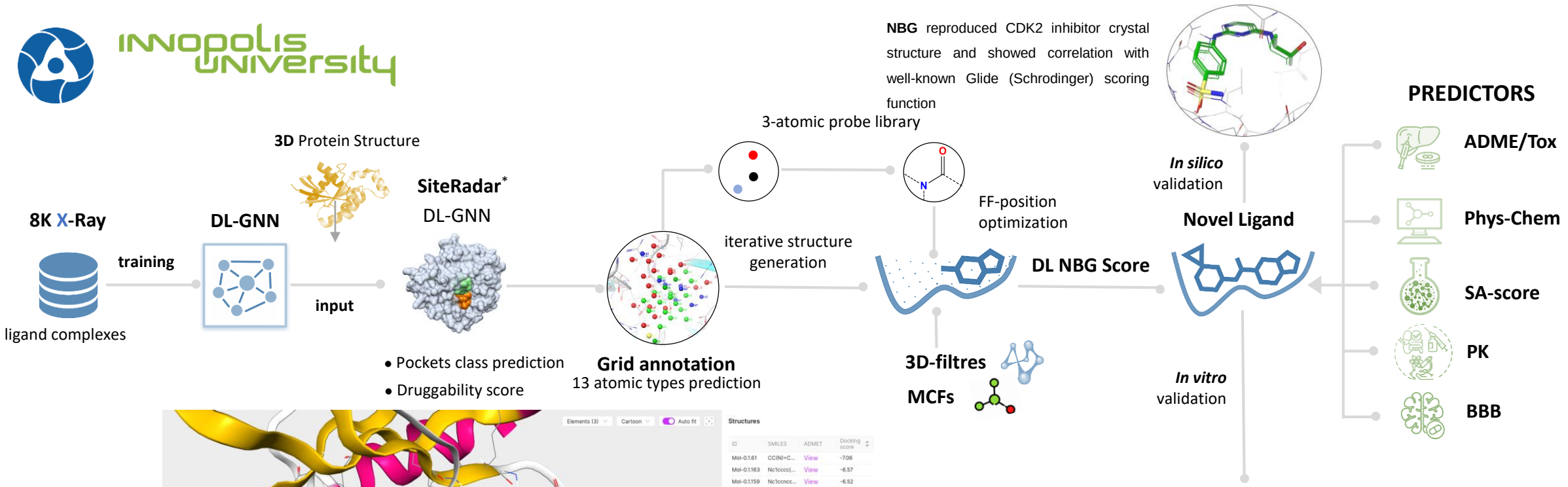






NATURE-BASED GENERATOR (NBG)

Atom-wise Generation of Ligand Structures Complementary to Macromolecular Environment



Target X inhibitors (Presude Lifesciences)

- Synthesis is ongoing

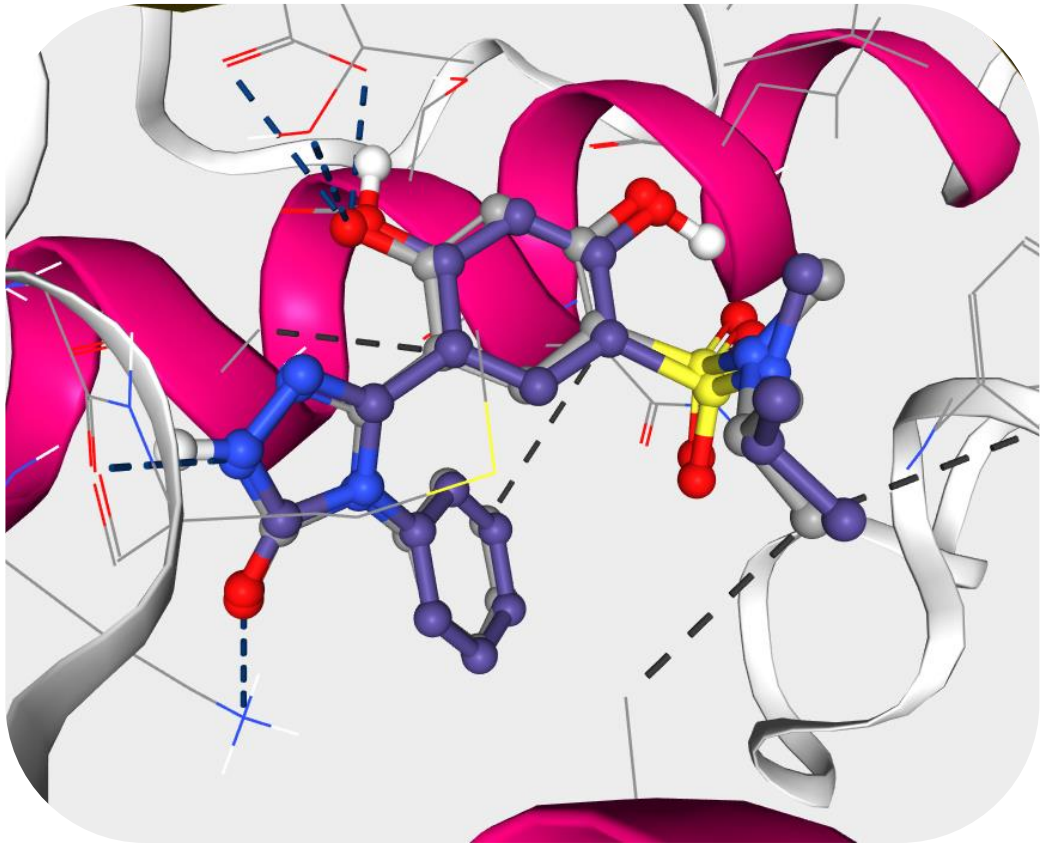
Kinase inhibitors (*in*-house program)

- Novel kinase with no ligand reported
- Hit-compound was developed with good selectivity

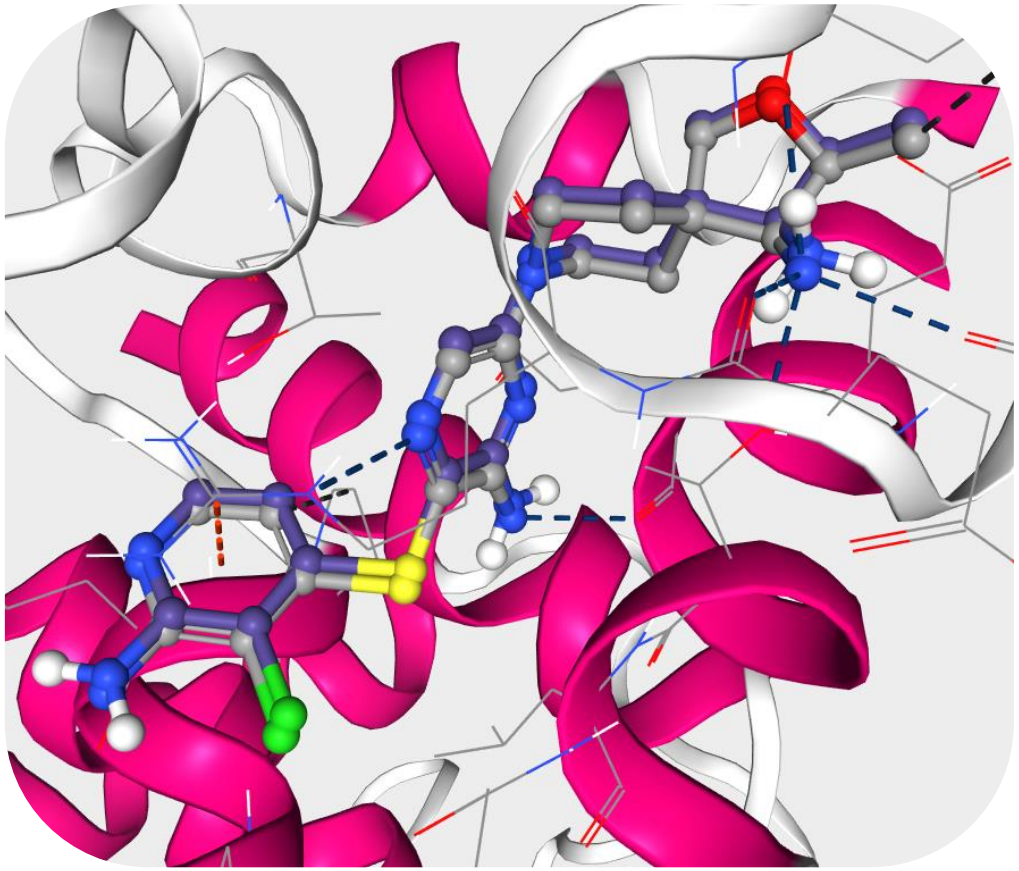
*Evteev SA, Ereshchenko AV, Ivanenkov YA. SiteRadar: *Utilizing Graph Machine Learning for Precise Mapping of Protein–Ligand-Binding Sites*. J. Chem. Inf. Model. **2023**, 63, 4, 1124–1132

Structure generation

Known crystals reproduction



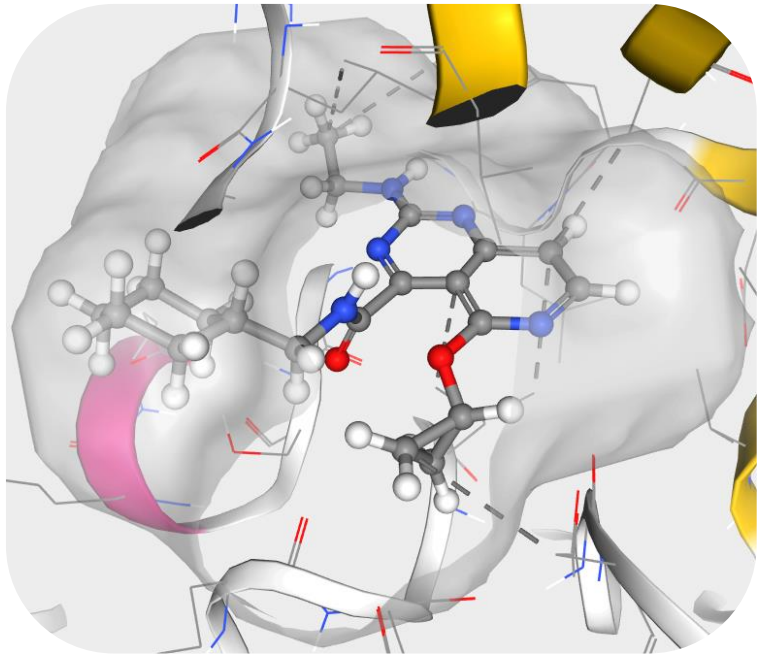
Heat shock protein 90-α
PDB ID 5J82



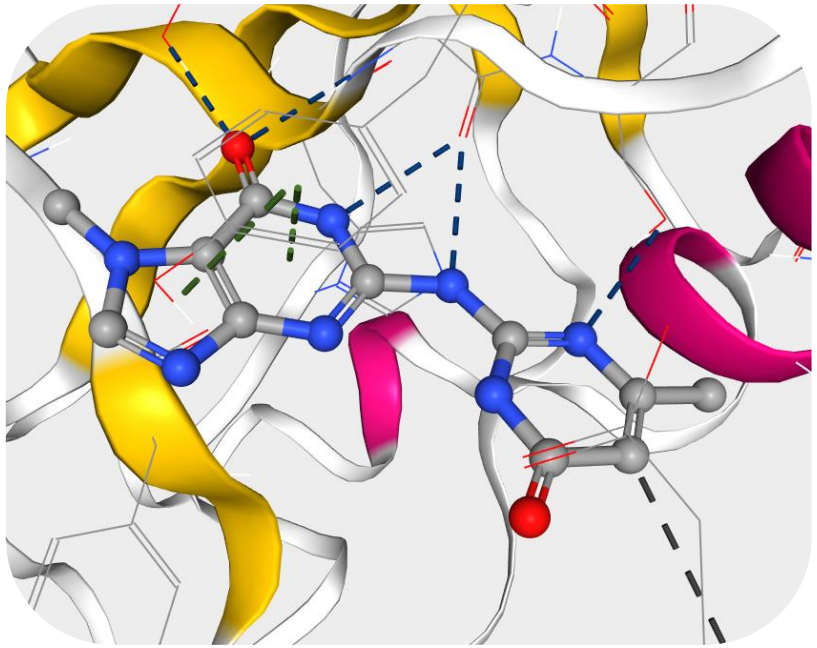
Phosphatase SHP2
PDB ID 7JVM

Structure generation

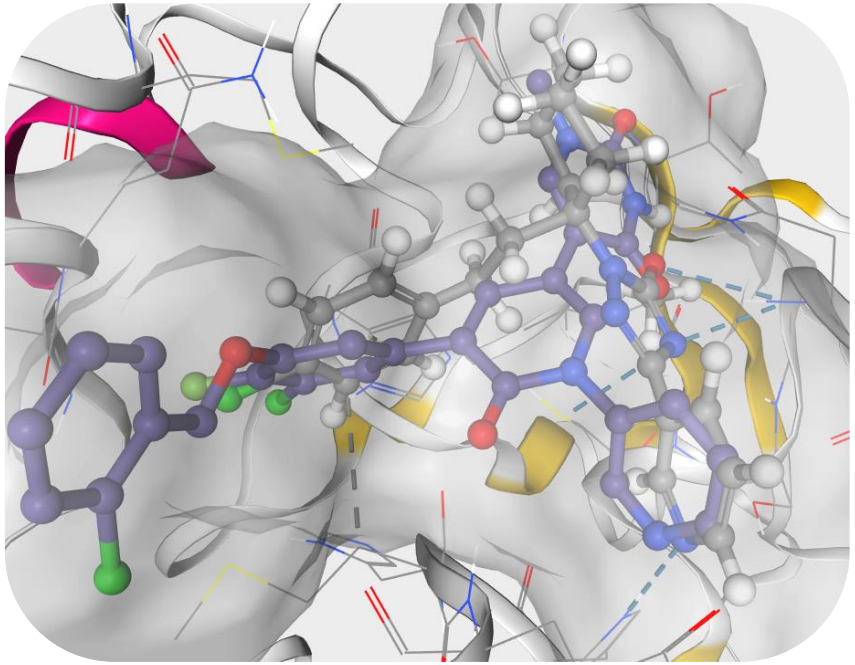
Design of new compounds



JAK1
PDB ID 6ELR



PARP1
PDB ID 4ZZZ



SARS Cov2 main protease
PDB ID 8UR9

Target X (Presude Lifesciences)

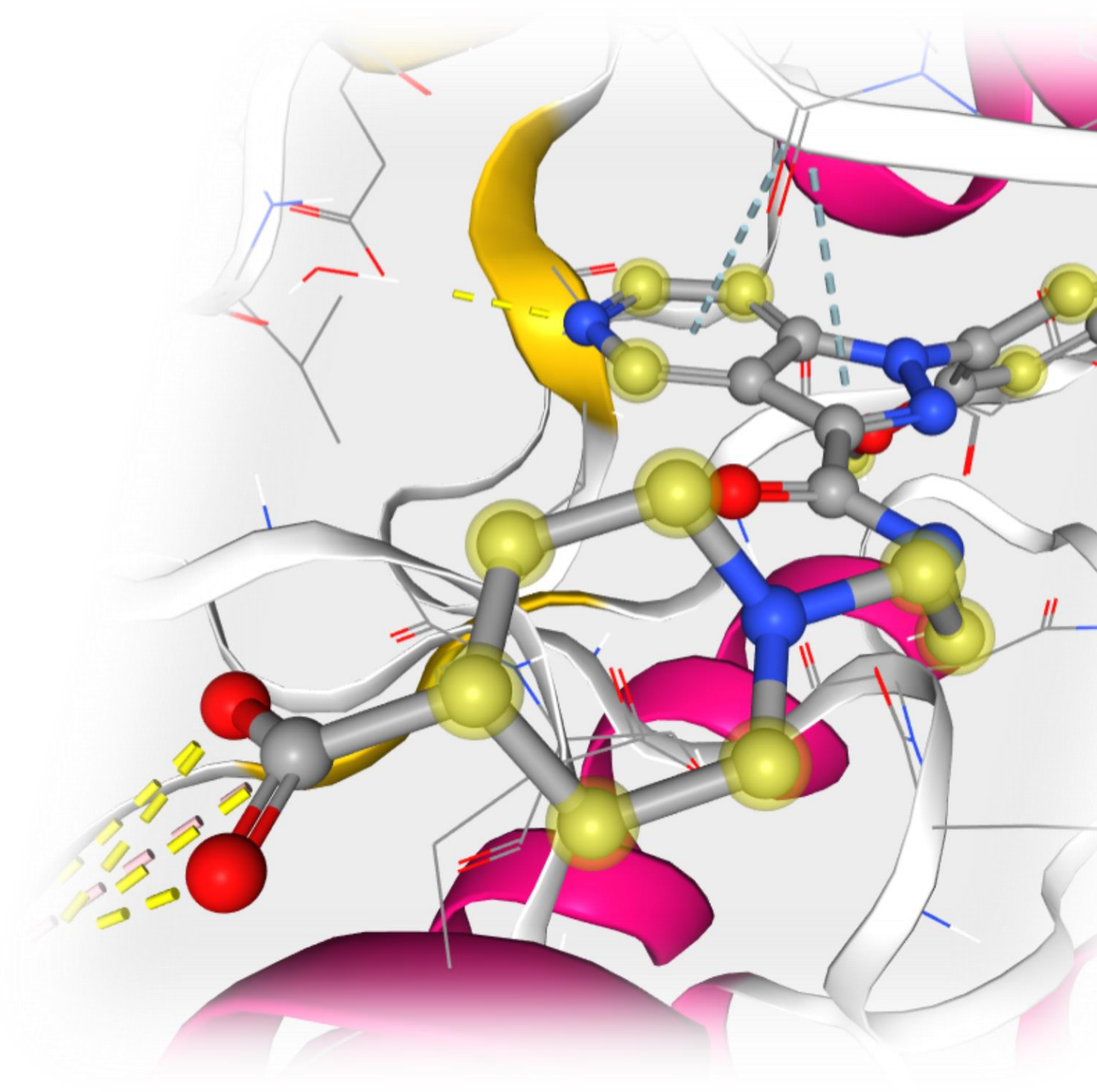
- Hit compound with IC_{50} = 3µM was obtained

Thank you

Sergei Evteev

saevteev@vniia.ru

@SERGEI_EVTEEV



Sequence- and structure-based prediction of protein stability change due to single mutations

Skoltech

Dmitry Ivankov

Protein design and redesign

Skoltech

Nobel prize 2024

Protein design

Daniela Röthlisberger[1]*, Olga Khersonsky[4]*, Andrew M. Wollacott[1]*, Lin Jiang[1,2], Jason DeChancie[6], Jamie Betker[3], Jasmine L. Gallaher[3], Eric A. Althoff[1], Alexandre Zanghellini[1,2], Orly Dym[5], Shira Albeck[5], Kendall N. Houk[6], Dan S. Tawfik[4] & David Baker[1,2,3]

Kemp elimination catalysts by computational enzyme design

doi:10.1038/nature06879

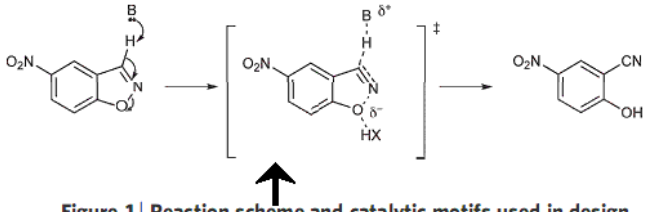
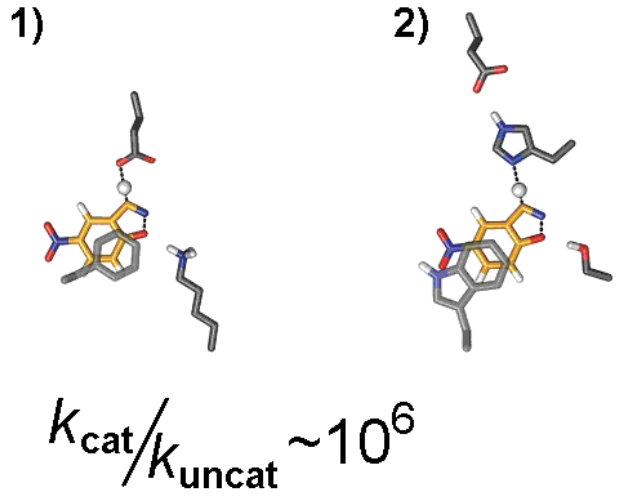


Figure 1 | Reaction scheme and catalytic motifs used in design.



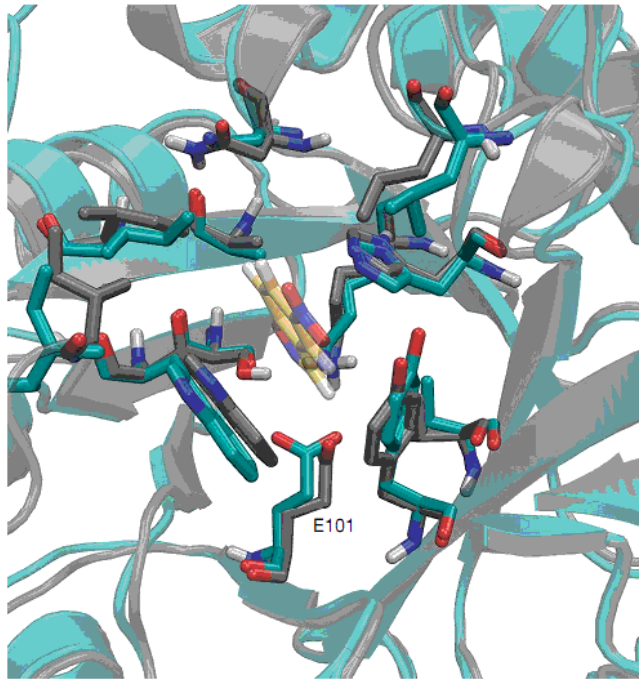
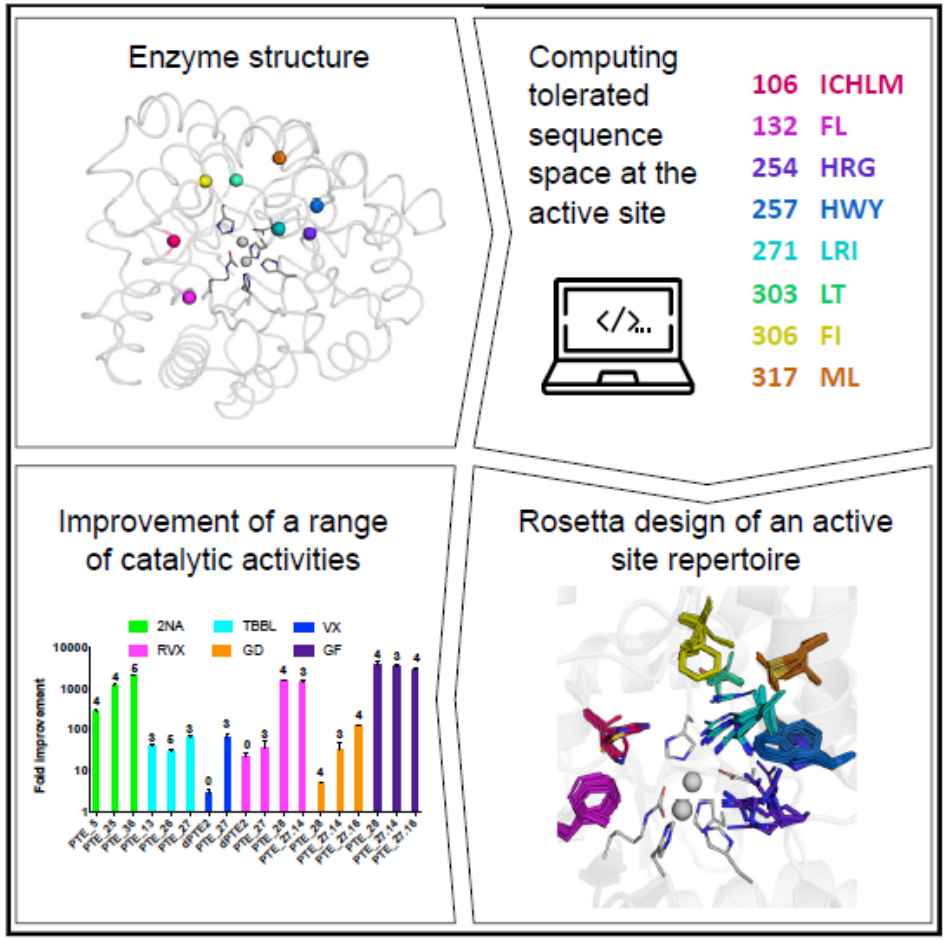
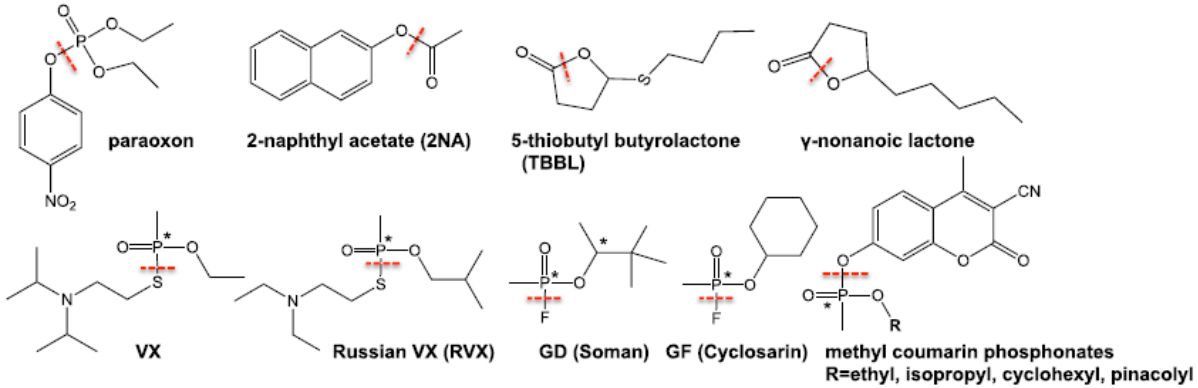


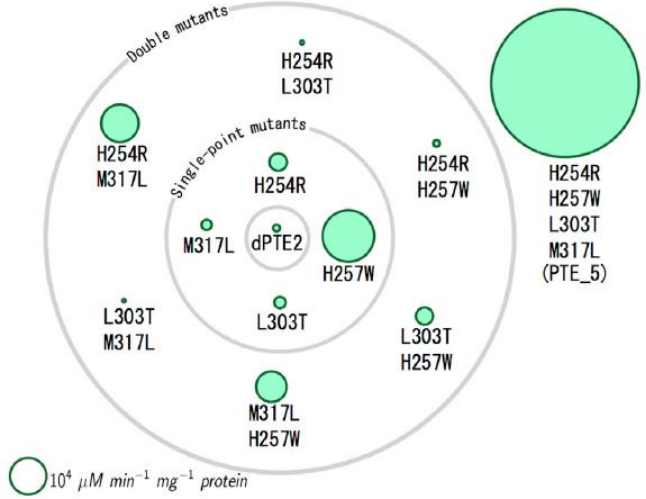
Figure 4 | Comparison of the designed model of KE07 and the crystal structure.

Skoltech

Protein redesign: phosphotriesterase







Skoltech

Khersonsky et al., Automated Design of Efficient and Functionally Diverse Enzyme Repertoires. Mol. Cell, 2018.

Why protein redesign: enzymes in washing powder

- Enzymes added to washing powder:
  - Proteases – break down protein chains from stains;
  - Lipases – break down fats and oils in stains;
  - Amylases – break down starch;
  - Cellullases – break down cellulose;
  - Mannanases – break down mannans.
- Enzymes work at normal temperatures
- We need to increase their thermostability to allow for washing at higher temperatures

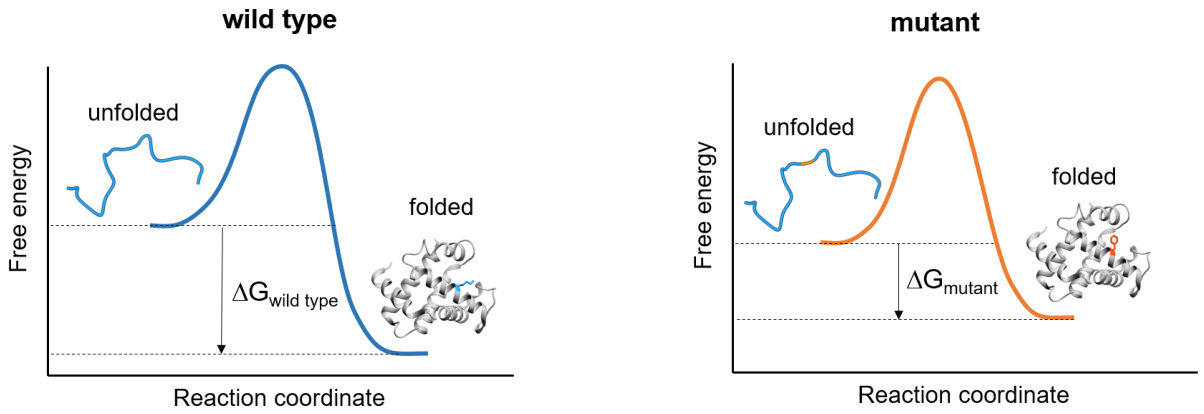
Skoltech

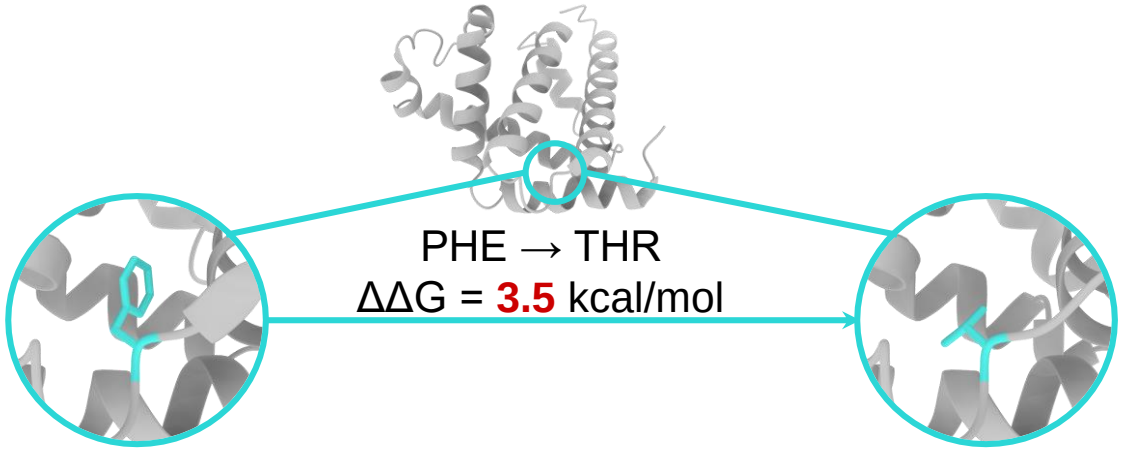
https://www.persil.com/uk/laundry/detergent/capsules.html

Change of protein stability on mutation

$$\Delta G = G_{folded} - G_{unfolded}$$

$$\Delta\Delta G = \Delta G_{mutant} - \Delta G_{wild\ type}$$



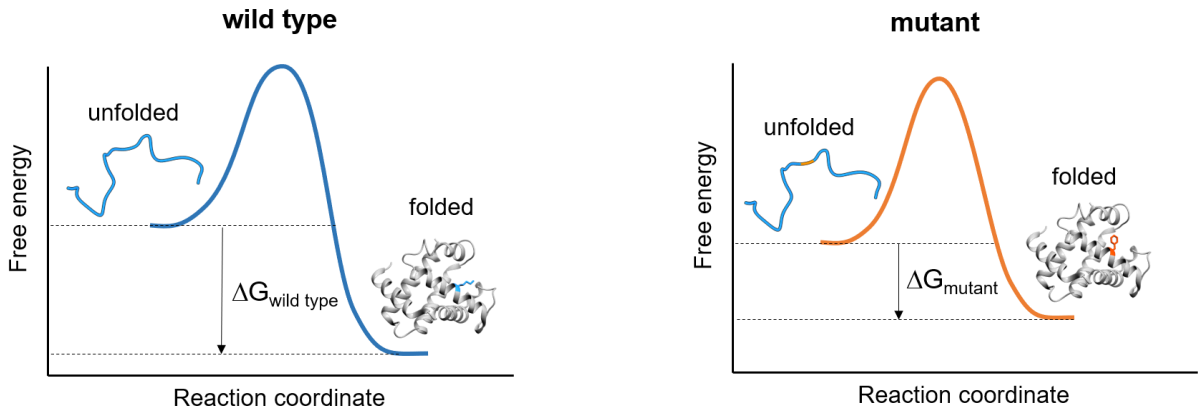


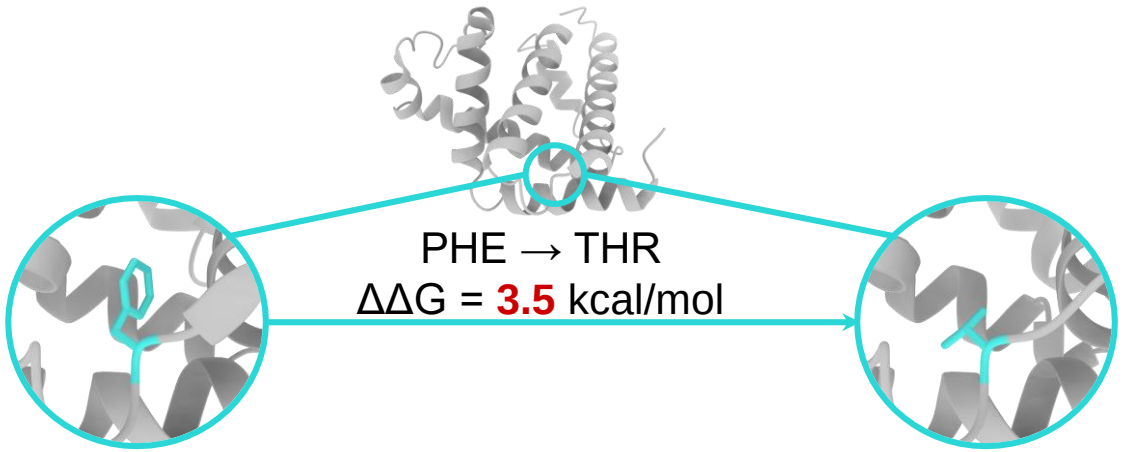
Skoltech

ΔΔG prediction: simplest task of protein design

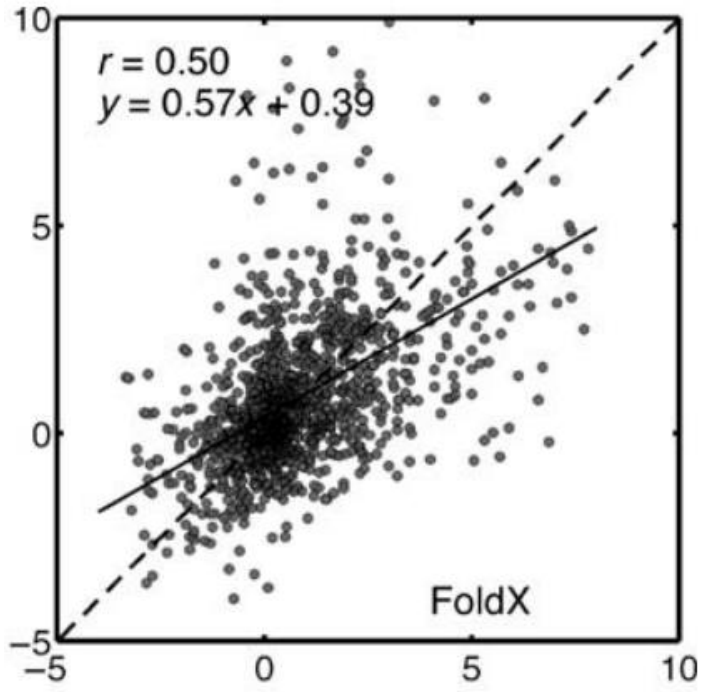
$$\Delta G = G_{folded} - G_{unfolded}$$

$$\Delta\Delta G = \Delta G_{mutant} - \Delta G_{wild\ type}$$





- Important for protein engineering
- Performance is ~ 50-60% (Pearson correlation)

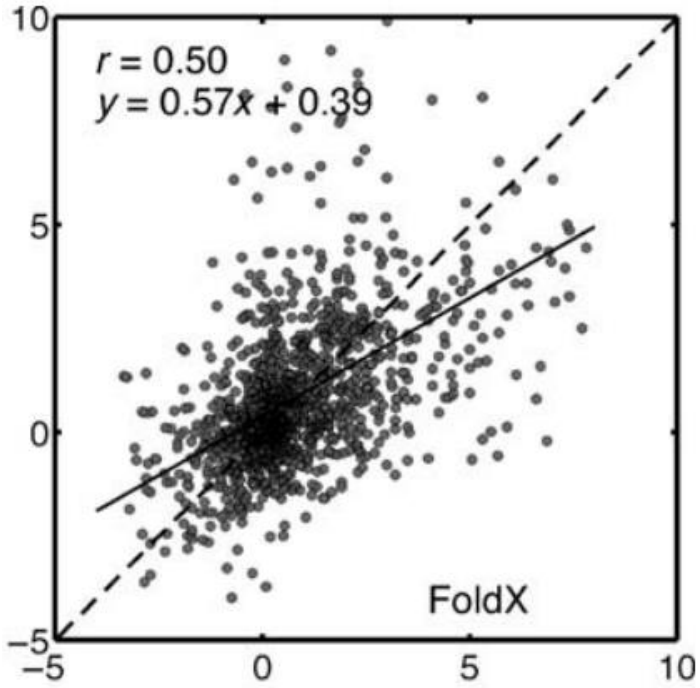


Skoltech

The number of predictors is 40+

PremPS
BoostDDG
I-Mutant ACDC
FoldX Maestro
HoTMuSiC INPS
SDM DynaMut mCSM
AUTO-MUTE DUET
STRUM PoPMuSiC
ThermoNet
DDGun EASE-MM
Eris Rosetta

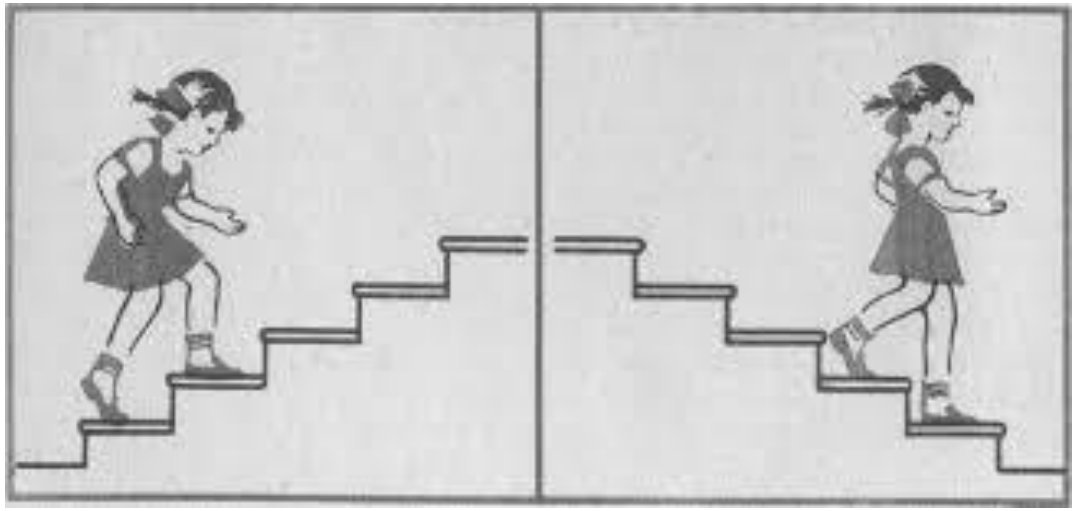
- Correlation ~ 50-60%



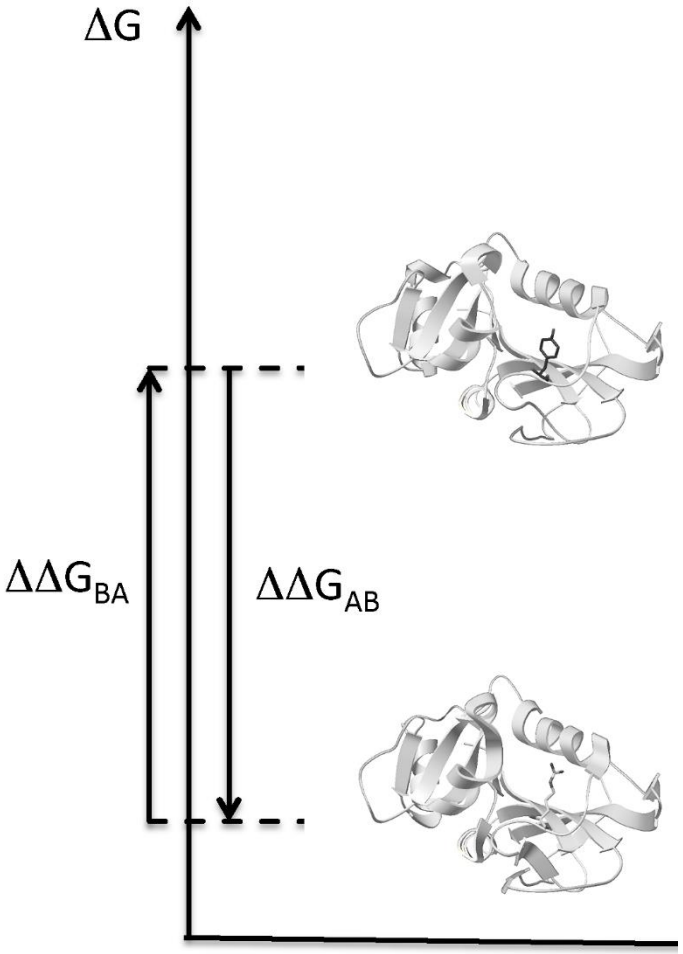
Skoltech

Predictors overestimate ΔΔG

- How to measure the overestimation?



Δh = 1m Δh = -1m

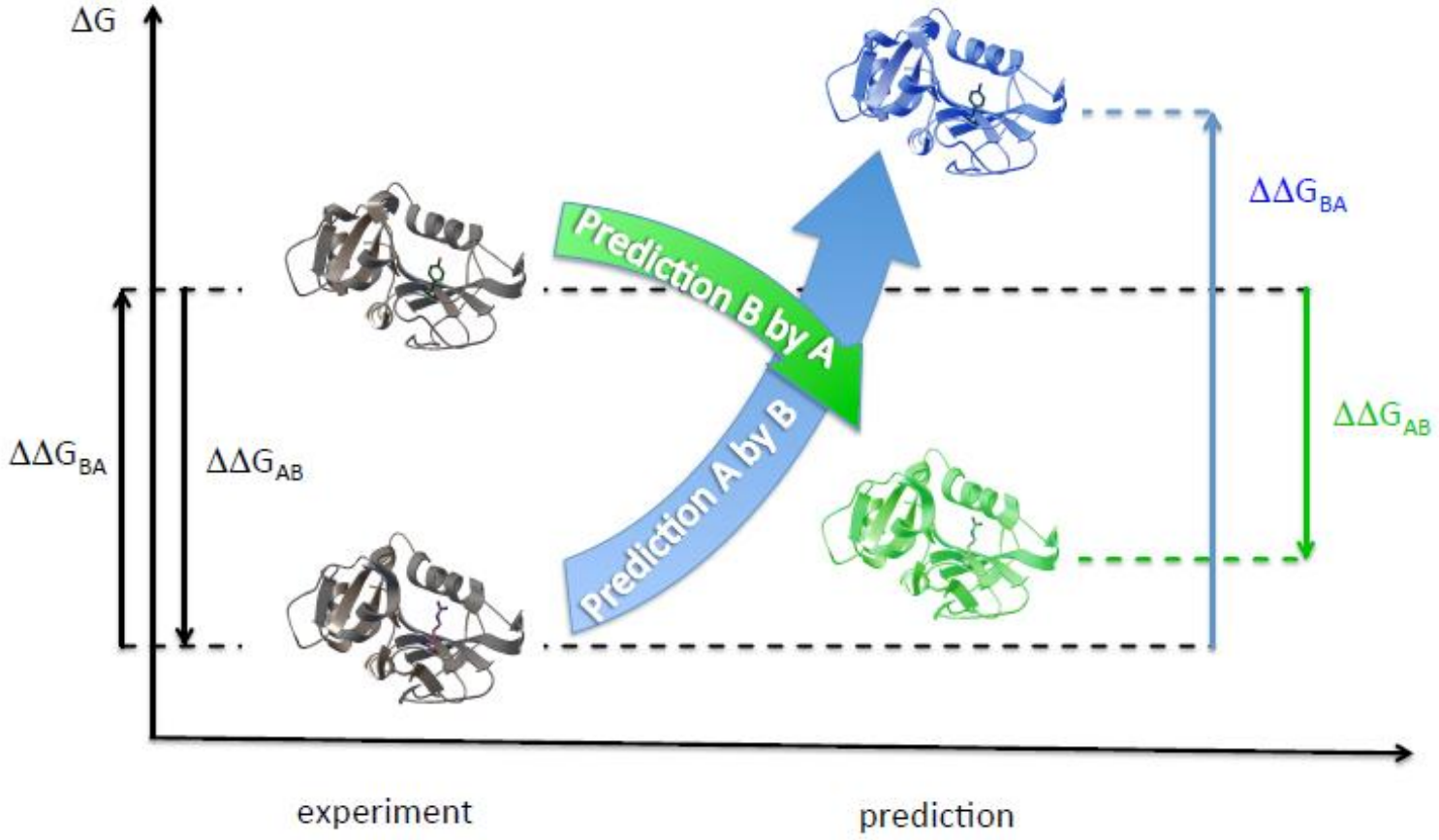


Skoltech

Usmanova et al., 2018, Bioinformatics

Self-consistency test

- We do not need experimental ΔΔG data!

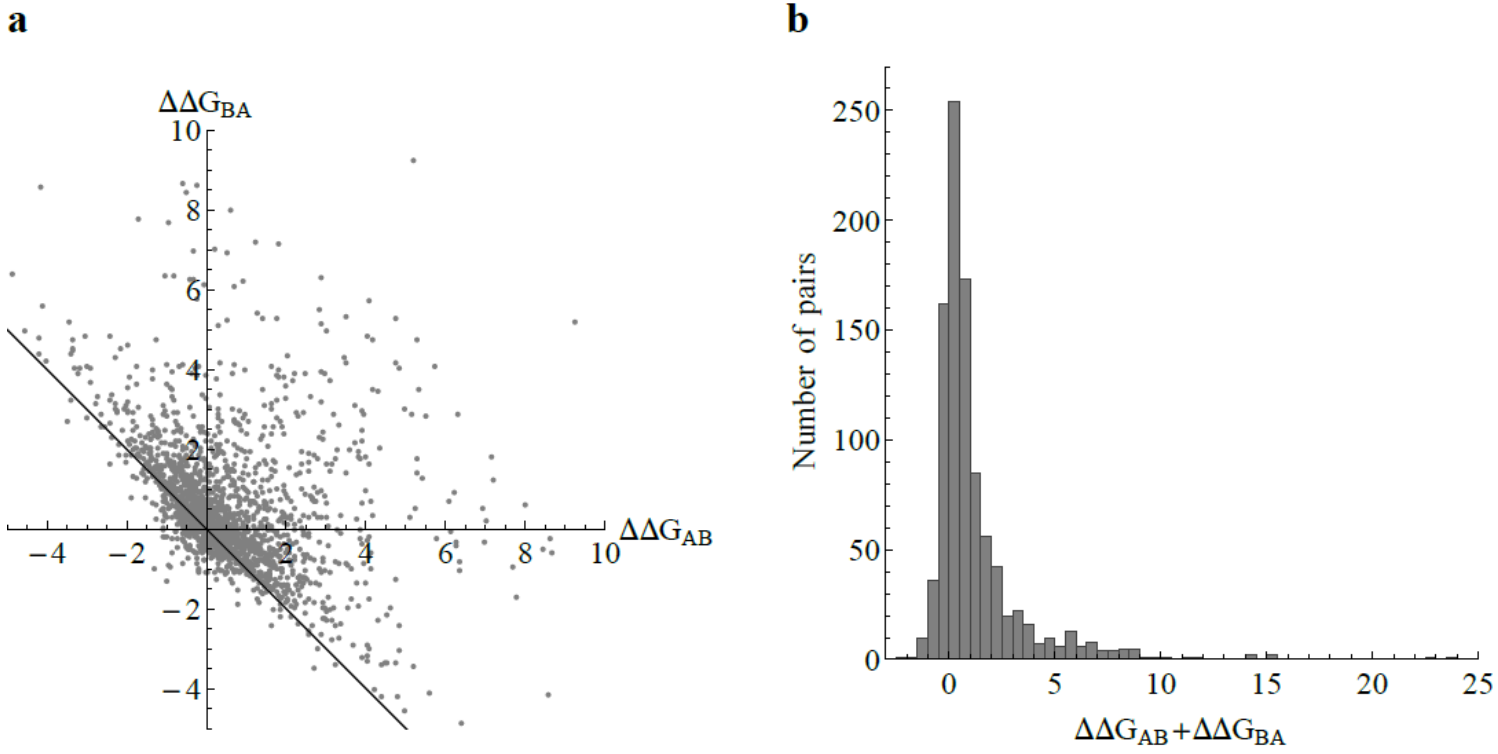


Skoltech

Usmanova et al., 2018, Bioinformatics

Bias for FoldX

- Equals 0.72 kcal/mol per single mutation
- Structure A is not optimal for new amino acid residue

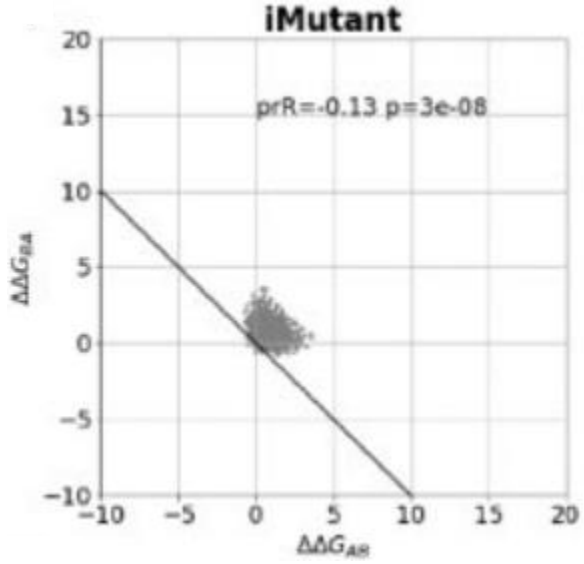


Skoltech

Usmanova et al., 2018, Bioinformatics

Bias for iMutant

- Equals 0.80 kcal/mol per single mutation
- Reflects the trend of the training dataset: most mutations are deleterious



Skoltech

Usmanova et al., 2018, Bioinformatics

How to exclude the bias? (1/2)

- Data symmetrization:

Myoglobin1	A13M	2kcal/mol
Myoglobin2	M13A	-2kcal/mol

- All new predictors after 2018 are symmetrized

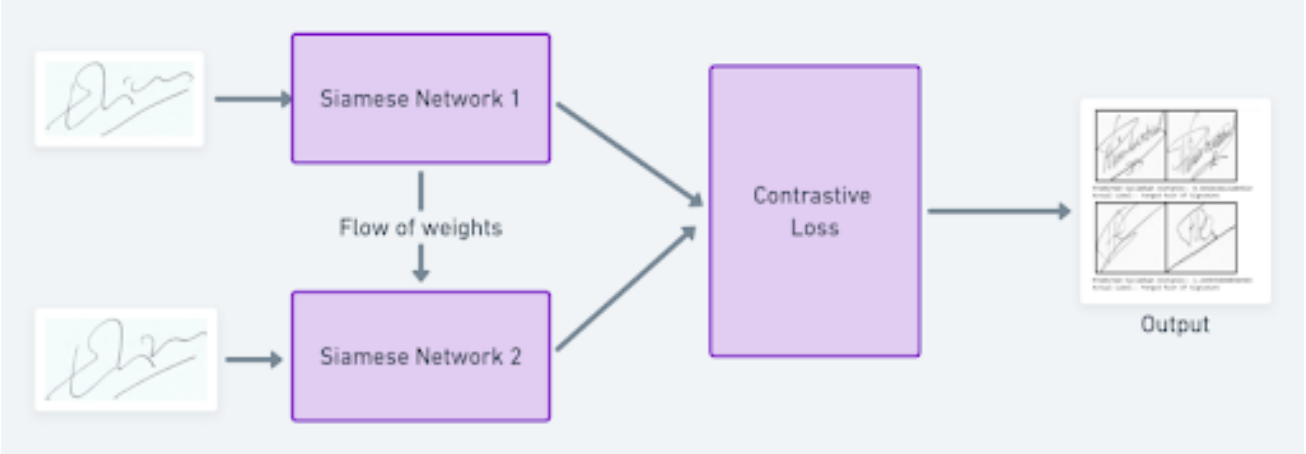
Skoltech

How to exclude the bias? (2/2)

- Predictor symmetrization during learning:

ADHase1	S123T	Xkcal/mol
ADHase2	T123S	-Xkcal/mol

- Siamese neural network architecture



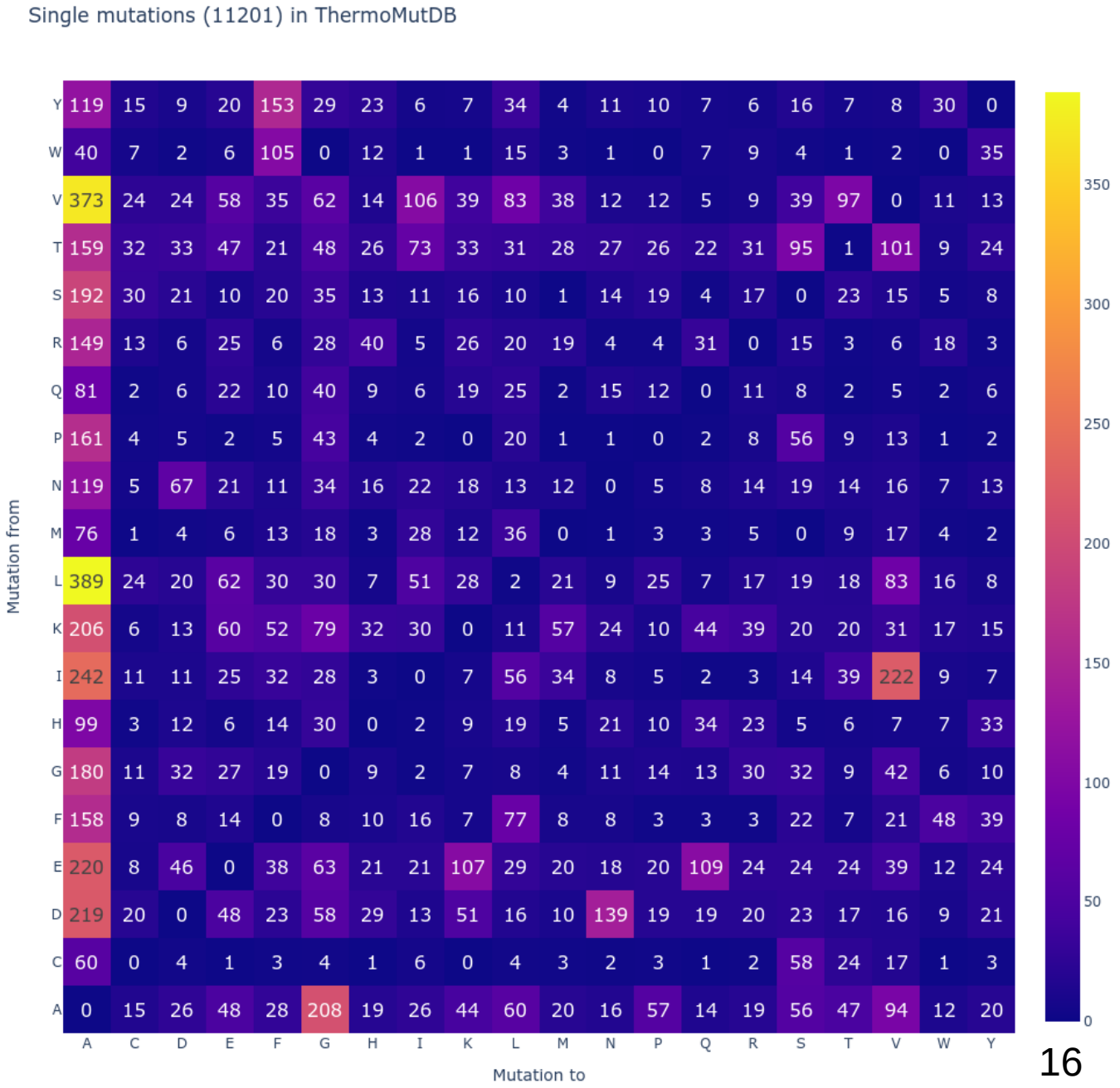
https://builtin.com/machine-learning/siamese-network

Skoltech

Experimental dataset is unbalanced

- ThermoMutDB

11 201 single mutations

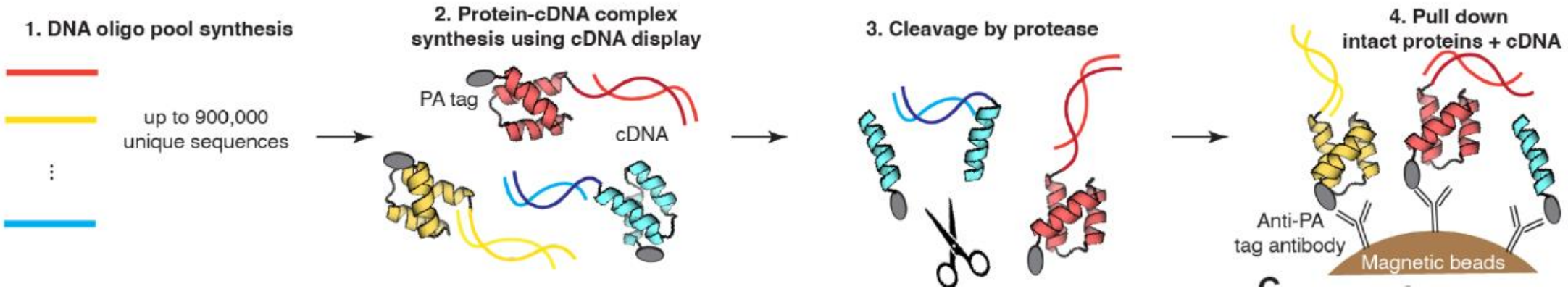


Skoltech

851 552 new mutations / 376 918 single

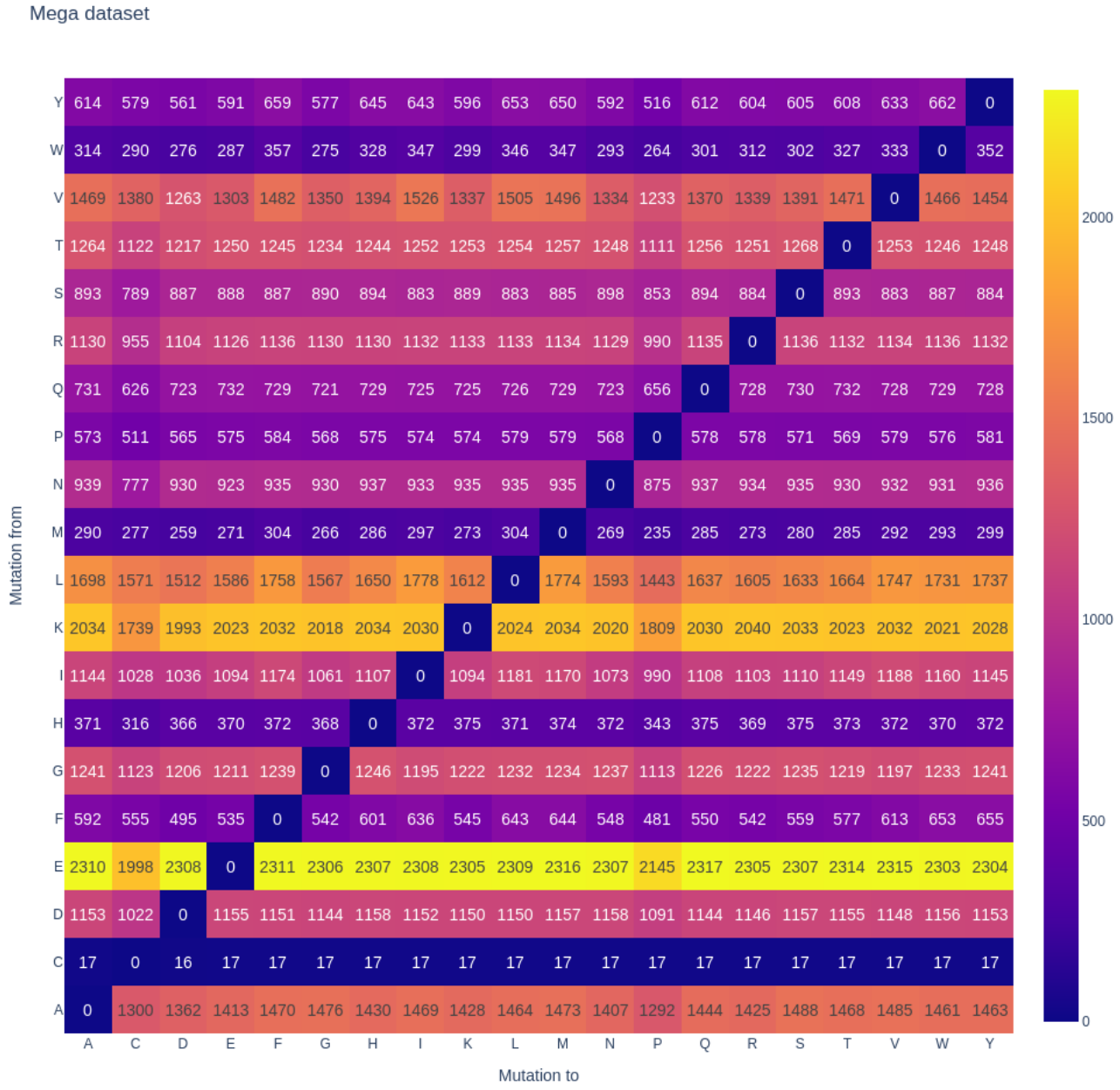
Mega-scale experimental analysis of protein folding stability in biology and protein design

Authors: Kotaro Tsuboyama[1,2,3], Justas Dauparas[4,5], Jonathan Chen[1,2], Niall M. Mangan[2,7], Sergey Ovchinnikov[8], Gabriel J. Rocklin[1,2] *



Skoltech

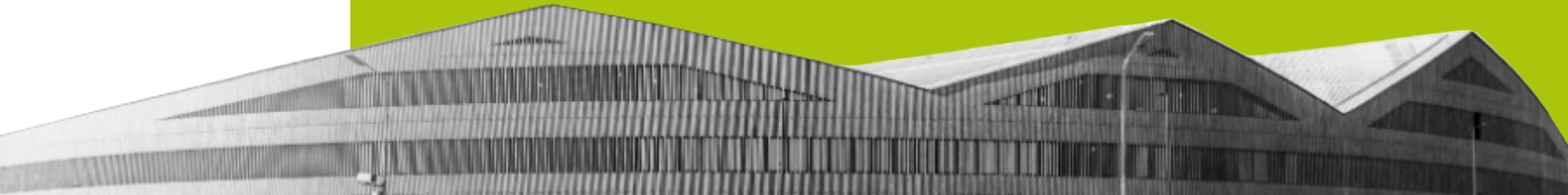
Statistics of single mutations for Mega-dataset



Skoltech

Sequence-based ΔΔG prediction

Skoltech



Dataset and Design

Data:

Mega dataset: All possible single-point mutations in 396 proteins

Tsuboyama et al. (2023). Nature, 620, 434.

Protein representation:

ESM-2 embeddings

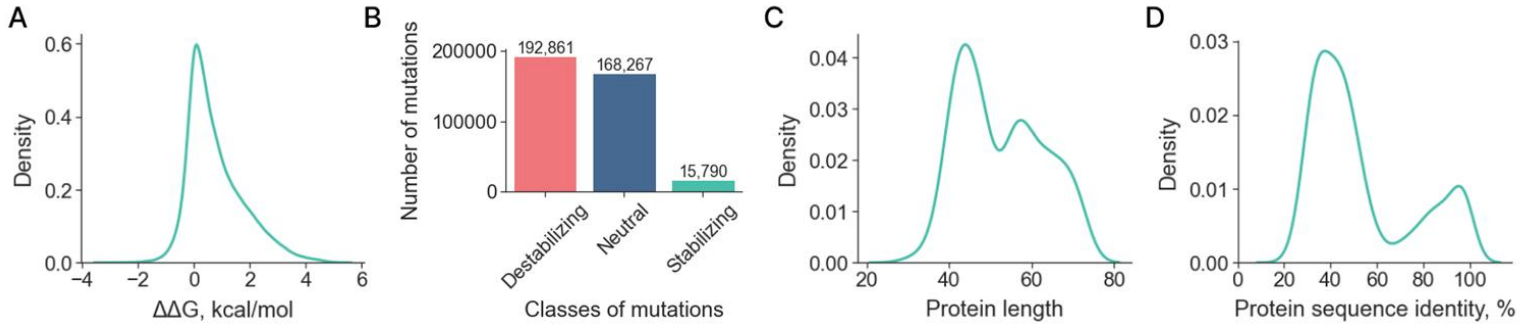
Lin et al. (2023). Science, 379, 1123.

Antisymmetry of ΔΔG prediction:

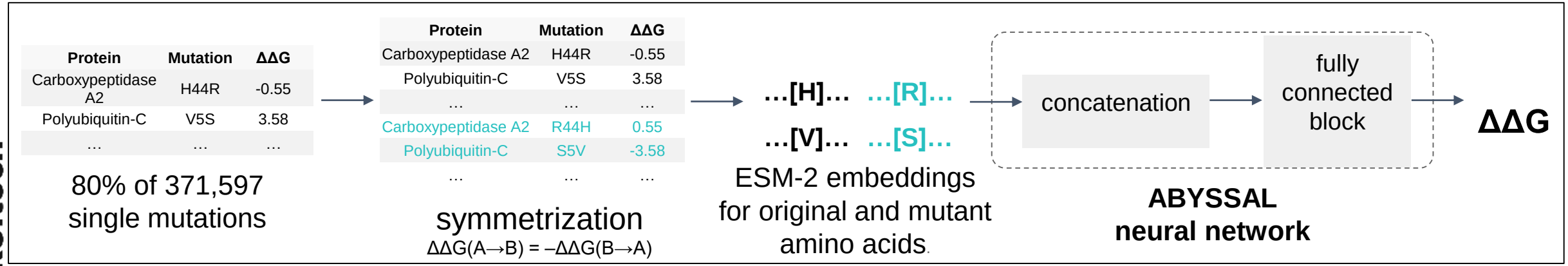
Dataset symmetrization

Siamese network

Bromley et al. (1993). International Journal of Pattern Recognition and Artificial Intelligence. 7, 669.



Description of the filtered Mega dataset



Design of the ΔΔG predictor

ABYSSAL performance

ABYSSAL outperformed other predictors on unseen subset of Mega dataset.

Predictor	PCC	SCC	MSE, kcal/mol	Accuracy
ABYSSAL	**0.76±0.01**	**0.71±0.01**	**0.67**	**0.75**
DeepDDG	0.70±0.01	0.58±0.01	1.01	0.72
INPS 3D	0.69±0.01	0.61±0.01	0.78	0.73
DDGun 3D	0.66±0.01	0.51±0.01	1.00	0.67
INPS	0.61±0.01	0.56±0.01	0.88	0.72

Performance of predictors on new data: Mega Holdout dataset (5321 mutations in 5 proteins)

Tsuboyama et al. (2023). Nature, 620, 434.

On old data ABYSSAL is comparable with top-performing predictors implying the ceiling of 50% PCC on this type of data.

Predictor	Symmetric data				PCC (f-r)	<δ>
	PCC	SCC	MSE, kcal/mol	Accuracy		
INPS-Seq	0.50±0.03	0.51±0.03	1.74	0.66	-0.99	0.00
ABYSSAL	**0.49±0.03**	**0.48±0.03**	**1.74**	**0.63**	**-0.98**	**0.02**
PremPS	0.49±0.03	0.48±0.03	1.75	0.67	-0.84	0.06
ACDC-NN3D	0.49±0.03	0.47±0.03	1.74	0.65	-0.98	-0.02
ACDC-NN	0.47±0.03	0.45±0.03	1.76	0.64	-1.00	0.00

Performance of predictors on old data: S669 dataset (420 mutations in 86 proteins)

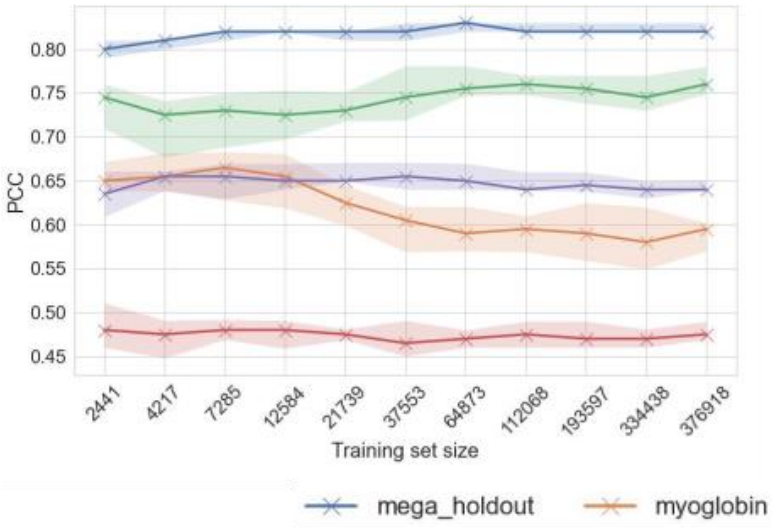
Pancotti et al. (2022). Briefings in Bioinformatics, 23(2).

Skoltech

Factors influencing performance

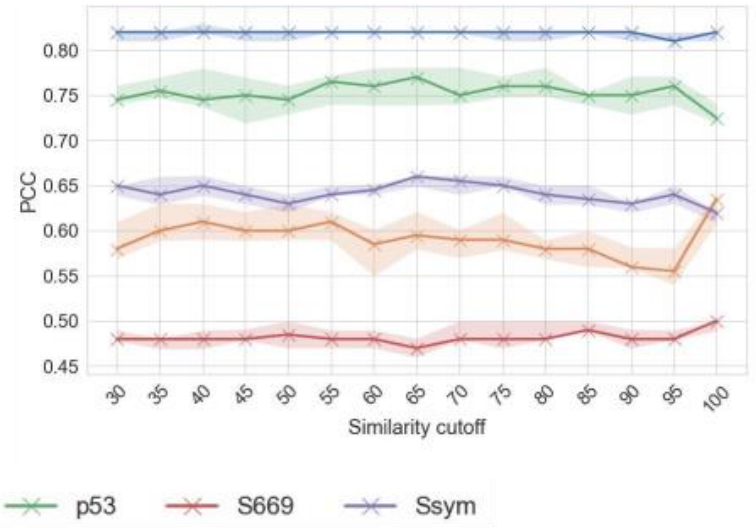
Data quality is the key factor influencing performance.

No significant change in performance when trained on a subset of Mega dataset as low as 2441 mutations.



Influence of training set size

Protein sequence identity cutoff for train-test split does not influence performance. Naive random split approach shows the same performance.



Influence of train-test splits by protein sequence identity

ABYSSAL ranks in the top-5 on Mega dataset when trained on old data of S2648.

	New data (Mega train)	Old data (S2648)
New data (Mega Holdout)	0.84±0.01	0.75±0.01
Old data (S669)	0.49±0.03	0.50±0.03

Dehouck, Y. et al. (2009). Bioinformatics, 25, 2537.

Influence of type of training data

Skoltech

Conclusion #1

- Transformer-based siamese network trained on symmetrized ESM-2 embeddings achieves top performance in ΔΔG prediction.

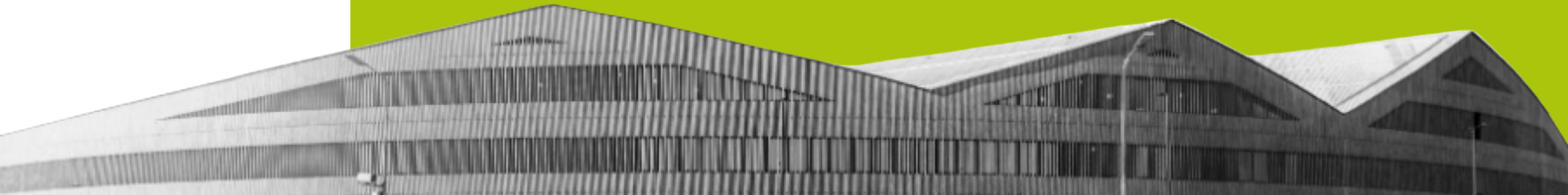
- Training set size and splitting strategy do not influence the performance much, while dataset quality is the key factor.

Skoltech

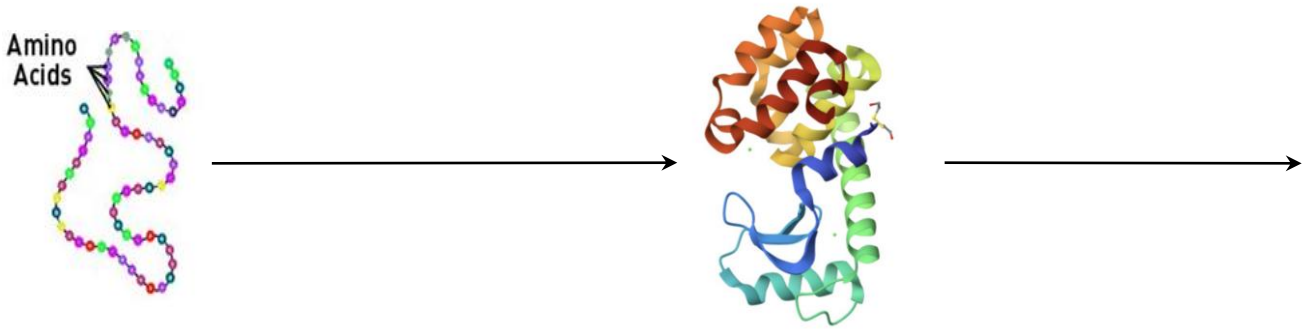
https://github.com/ivankovlab/abyssal

Structure-based ΔΔG prediction

Skoltech



Protein representation learning task



Protein properties
Protein interactions

Protein sequence
[Available]

Protein structure
[Available with AlphaFold]

Protein language models (PLM), eg. ESM-2 [1]

- Graph Neural Networks [3]
- Structure-aware PLMs
  - trained from scratch (SaProt [4])
  - add large number of parameters

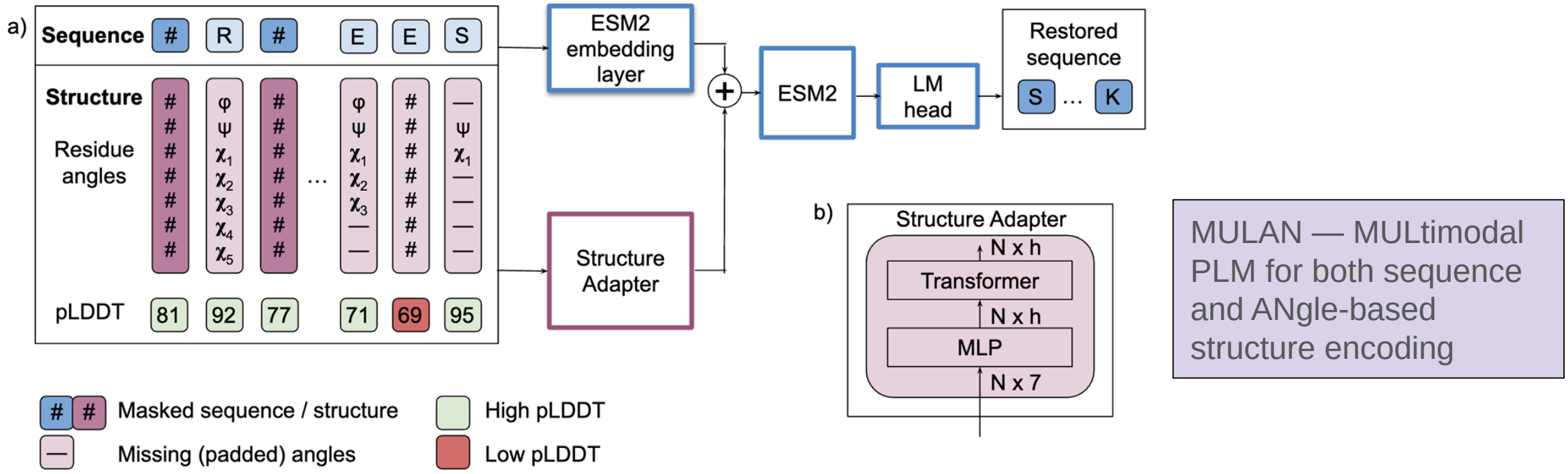
Goals:

- improve protein representation quality with the use of the novel structure-aware PLM;
- add protein structural information to a transformer model with a lightweight finetuning of existing PLMs.

Image credits: https://byjus.com/biology/proteins-structure-and-functions/

[1] Language models of protein sequences at the scale of evolution enable accurate structure prediction, Lin Z. et al., 2022
[2] Highly accurate protein structure prediction with AlphaFold, Jumper J. et al., 2021
[3] Diffdock: Diffusion steps, twists, and turns for molecular docking, Corso G. et al., 2022
[4] Saprot: Protein language modeling with structure-aware vocabulary, Su J. et al., 2023

MULAN architecture



MULAN — MULtimodal PLM for both sequence and ANgle-based structure encoding

Figure 2: The architecture of MULAN. a) MULAN processes sequence inputs with the ESM2 embeddings module, while structure inputs are passed to the Structure Adapter. Both sequence and structure embeddings are summed up and passed to the ESM2 model, which is then finetuned. Sequence-only ESM2 modules (blue) are initialized from the pre-trained ESM2 checkpoint. Structure processing modules are shown in pink. b) The architecture of the Structure Adapter.

Skoltech

Experimental setup

- Train on top of existing PLMs:
  - sequence-only ESM-2 8M, 35M, 650M
  - structure-aware SaProt 35M, 650M
- Only finetune base PLM together with the Structure Adapter
- Use dataset with 17M AlphaFold structures for training
- Evaluate protein embeddings on 7 downstream tasks
  - eg. protein property prediction and protein interaction prediction
  - protein embedding = average of all residue embeddings
  - train small downstream model on protein embeddings for each downstream task independently

Skoltech

Results

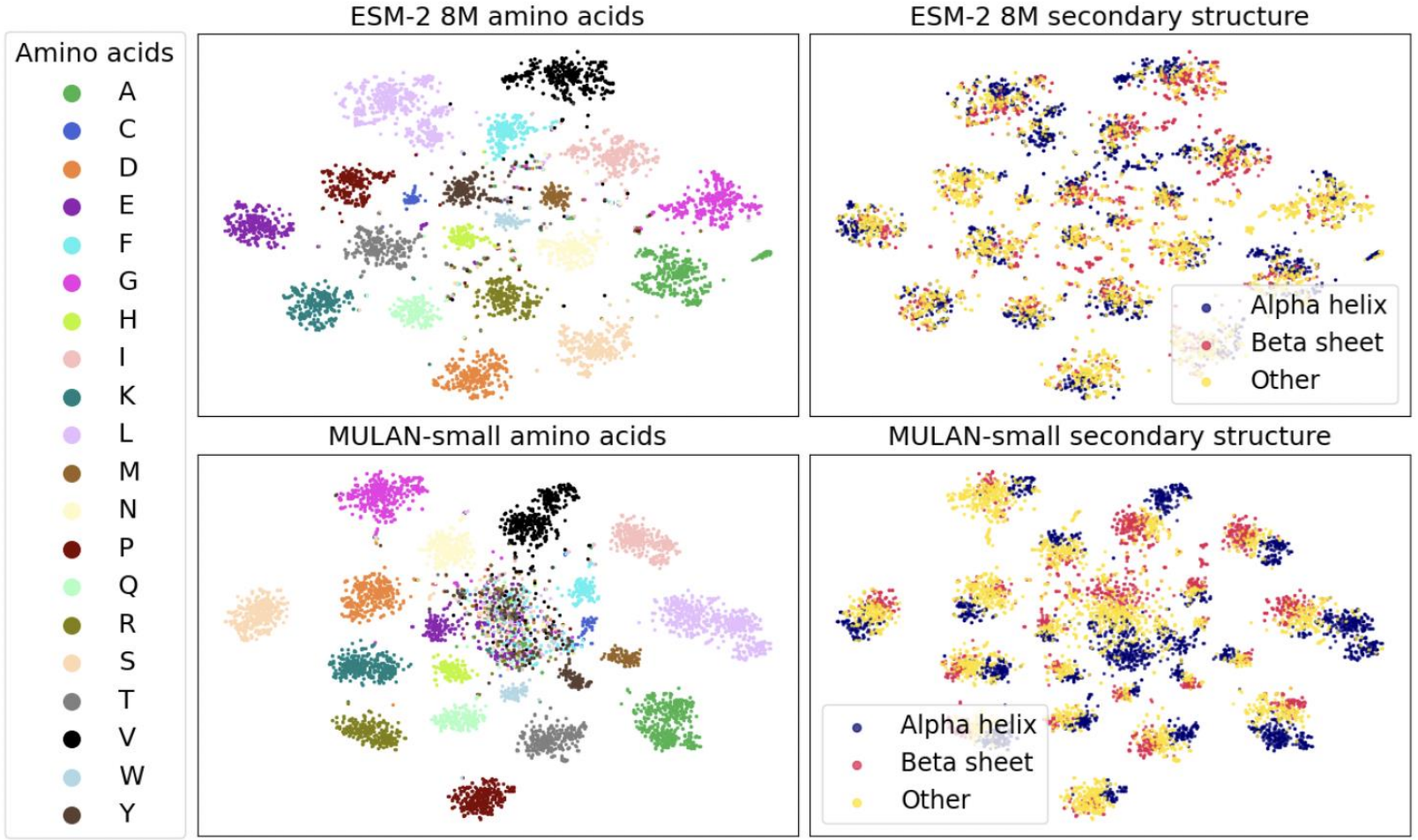
Table 6: The improvement shown by adding MULAN to various PLMs and SPLMs on all downstream tasks. The best results for each base model are shown in bold

Model name	Thermo-stability	Fluore-scence	Metal Ion Binding	Human PPI	GO		
					CC	MF	BP
	SCC ↑	SCC ↑	AUC ↑	AUC ↑	F_{max} ↑	F_{max} ↑	F_{max} ↑
Small models							
ESM-2 8M	.666	.579	.731	.698	.490	.529	.400
Δ MULAN-small 8M	**.006**	**.017**	**.047**	**.055**	**.002**	**.058**	**.026**
Medium models							
ESM-2 35M	.689	.592	.793	.751	.489	.621	.443
Δ MULAN-ESM2 35M	**.012**	**.017**	**.001**	**.031**	**.027**	**.015**	**.004**
Saprot AF 35M	.699	.639	.783	.731	.501	**.632**	.440
Δ MULAN-SaProt 35M	**.005**	**.003**	**.017**	**.048**	**.004**	−.001	**.002**
Large models							
ESM-2 650M	.694	.601	.781	.754	**.523**	**.678**	**.479**
Δ MULAN-ESM 650M	**.009**	**.007**	**.013**	**.117**	−.004	−.001	−.004
SaProt AF 650M	**.711**	.668	.776	.720	.540	.658	.464
Δ MULAN-SaProt 650M	−.008	**.001**	**.026**	**.048**	**.005**	**.005**	**.006**

MULAN generally improves the quality of base PLMs (and even structure-aware PLMs) of various sizes

Skoltech

Visualization of structural awareness



MULAN produces structure-aware protein representations

T-SNE visualization of residue embeddings of MULAN-small and ESM-2 8M on CASP12 dataset.

We use different colors for amino acid residue types (left) and for the 3 states of secondary structure (right)

Skoltech

Conclusion #2

- Proposed **MULAN** — MULtimodal PLM for both sequence and ANgle-based structure encoding.
- Evaluated the obtained structure-aware protein representations on a wide range of downstream tasks. We show that **MULAN improves over any base PLM it is applied to**.
- MULAN requires finetuning of the underlying base PLM together with the Structure Adapter → MULAN offers a **cheap increase in performance**.
- Demonstrated the **structural awareness of MULAN** embeddings.

Skoltech

https://github.com/DFrolova/MULAN

Acknowledgements

Marina Pak

Nikita Dovidchenko

Darya Frolova

Marina Pak

Ilya Sharov

Satyarth Mishra Sharma

Anna Litvin

Ivan Oseledets

thx.

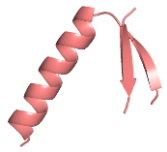
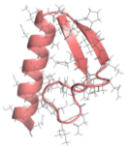
Skoltech

Robust Evaluation Strategies for Protein Design

Andrey Shevtsov

Research engineer, AIRI

Why protein generation

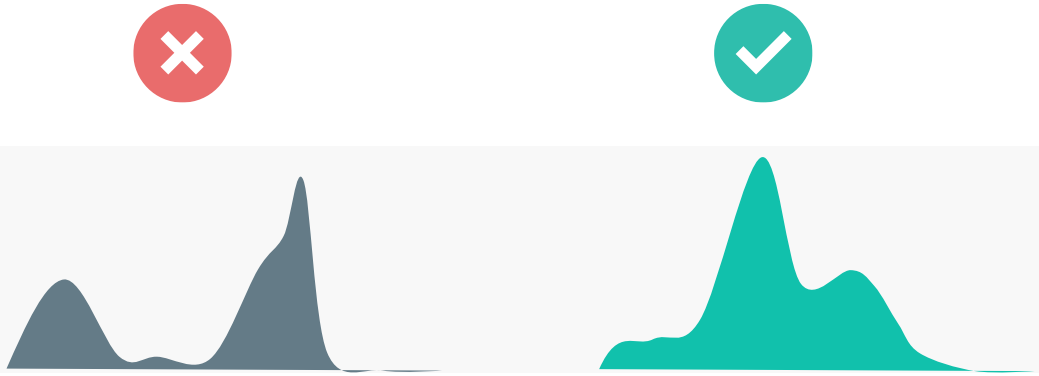
Scaffolding			Create efficient surrounding	→ vaccines → enzymes
Sequence design	 ??????????????????????????????	 NAILAPVNLASQNSQGGVLNGFYSAN	New synthetic protein generation	→ antibiotic properties → enzymes → new active peptides
Protein fragment design	 GTNAILAPVNLASQ????????FYSAN	 GTNAILAPVNLASQGGVLNGFYSAN	Completing a protein region	→ specific antibodies → proteins with improved properties
De novo generation			Create new proteins	→ vaccines → new receptor binders

Generative models

- Autoregressive models: ProGen, RITA, ProtGPT2
- Diffusion models: Evodiff, DPLM, Rfdiffusion
- Flow matching models: FoldFlow, AlphaFlow, FrameFlow, MultiFlow
- GAN: ProteinGAN

What makes a generative model great?

- Generated distribution is similar to the training set



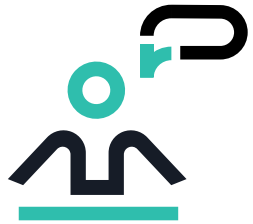
- High quality samples

- Diverse samples

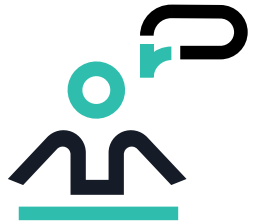
Challenges

- Limited human feedback capability
- Lack of standardized metrics
- Insufficient metric validation

We created new SOTA model.
It generates the GREENEST proteins



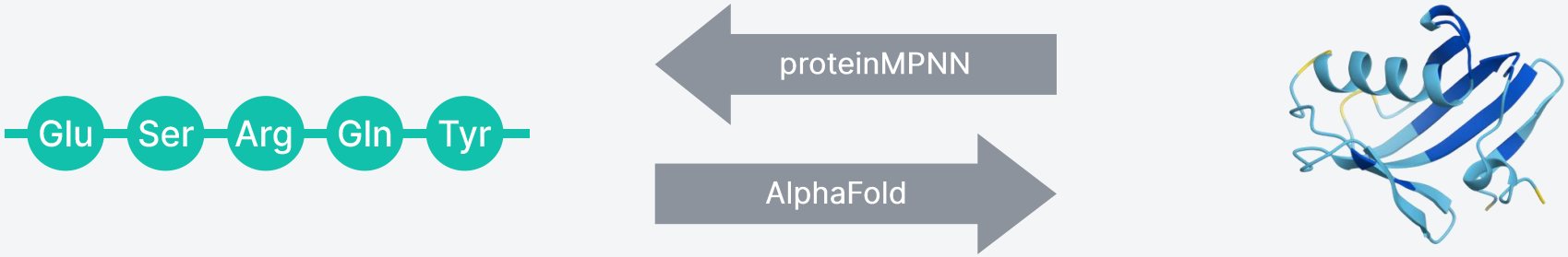
We created new SOTA model.
It generates the HARDEST proteins



We created new SOTA model.
It generates the SMARTEST proteins

The bright side of protein GenAI evaluation

Protein modalities



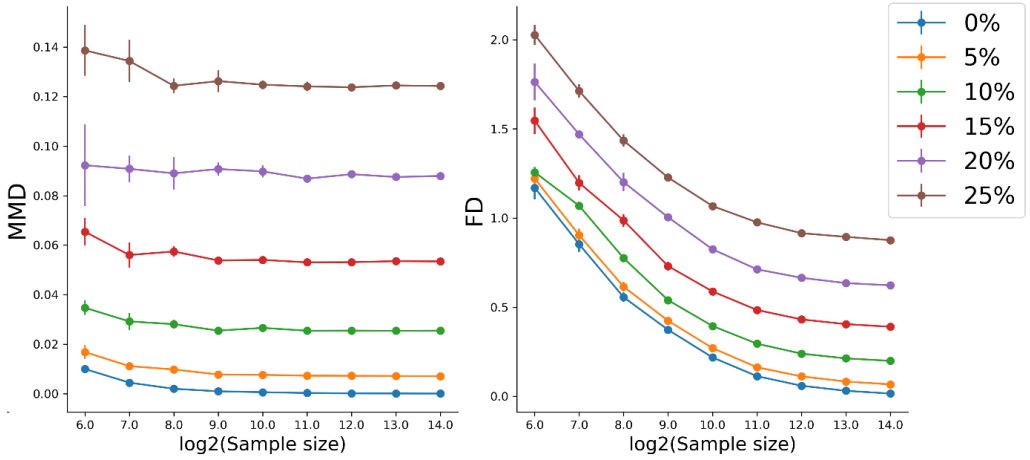
Experience from other fields

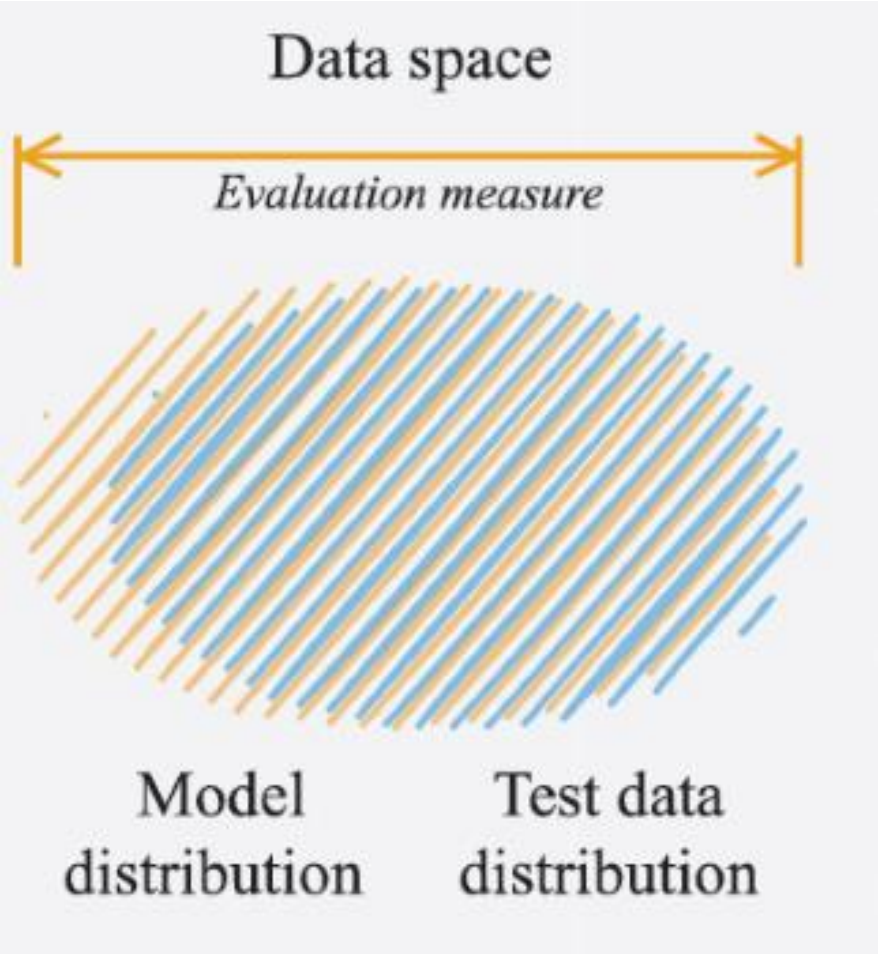
NLP

Image generation

Distribution similarity metrics

- Fréchet distance (1000+ samples)
- MMD (500+ samples)
- MMD kernel: RBF (sigma=10)
- Latent space: ProtT5 sequence embeddings

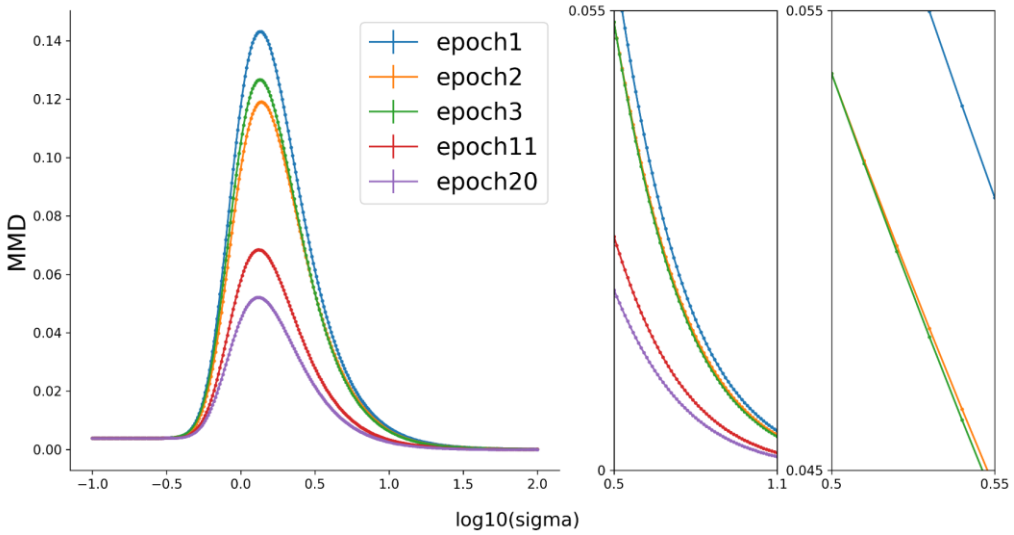


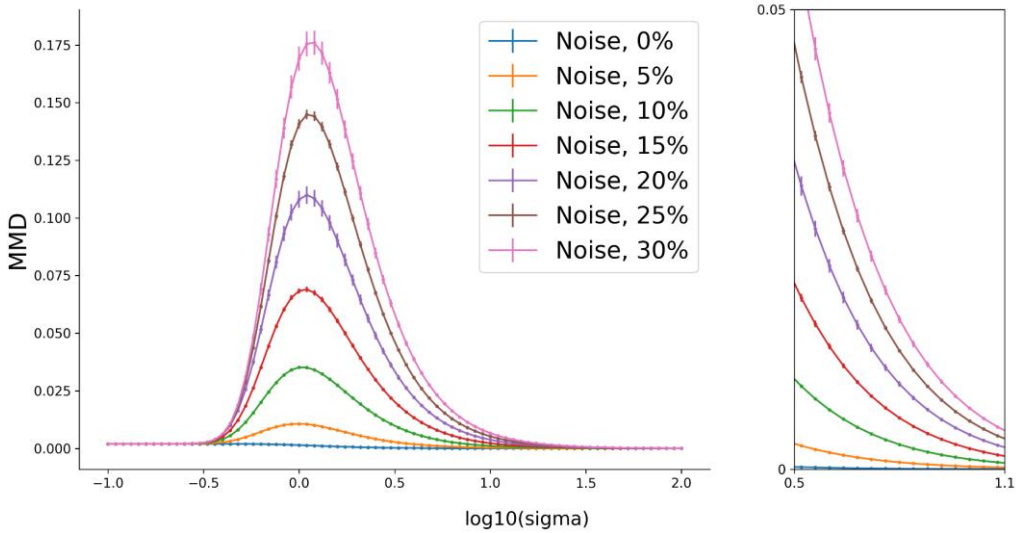


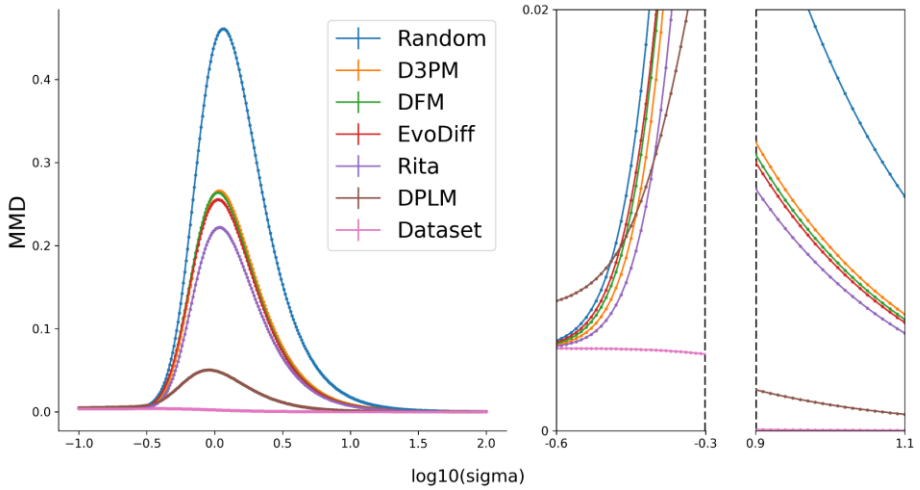
MMD Sigma choice

$$MMD^2(P,Q) = \mathbb{E}_{x\sim P}\left[k(x,x')\right] + \\ + \mathbb{E}_{y\sim Q}\left[k(y,y')\right] - 2\mathbb{E}_{x,y\sim P,Q}\left[k(x,y)\right]$$

$$K(X_1, X_2) = \exp(-\frac{\|X_1 - X_2\|^2}{2\sigma^2})$$

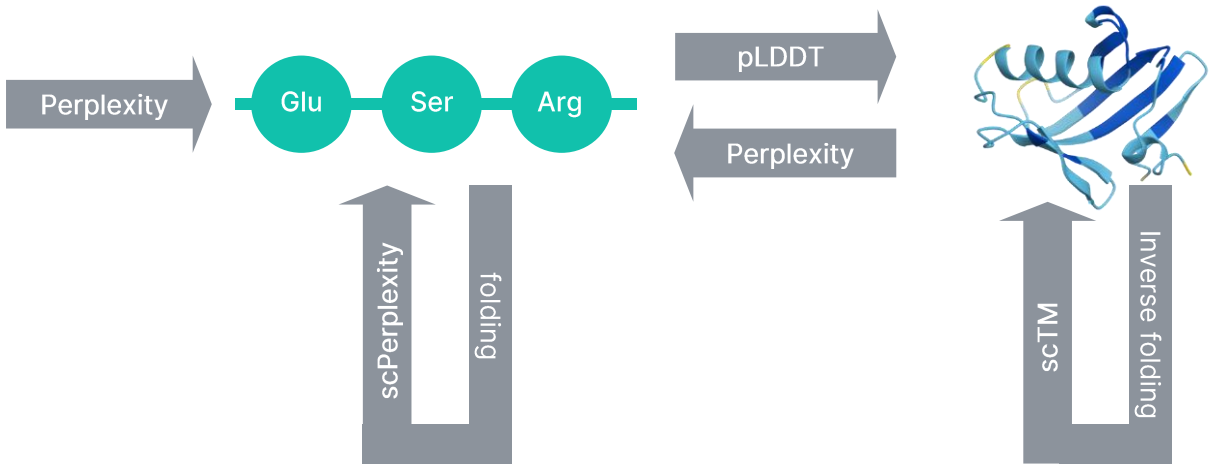


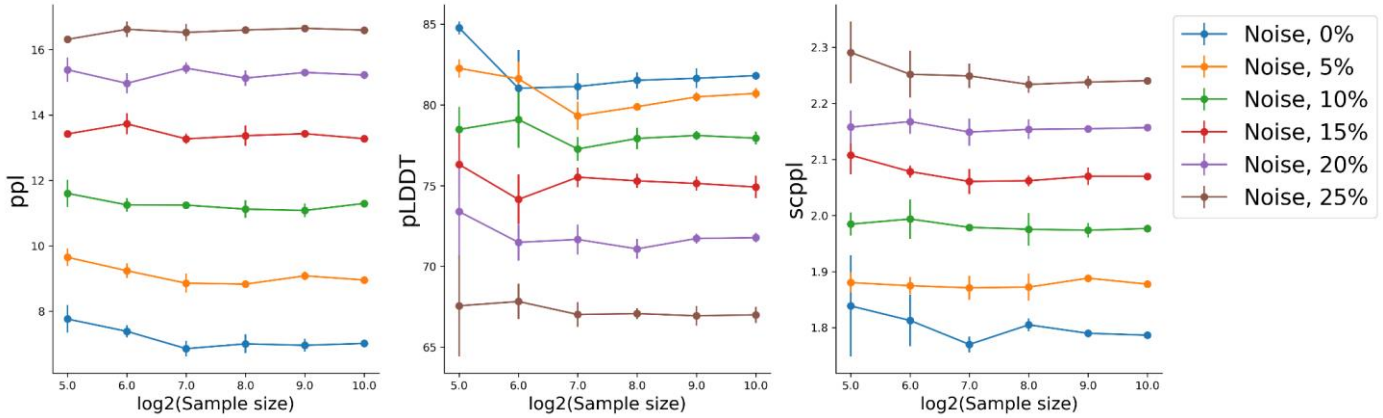




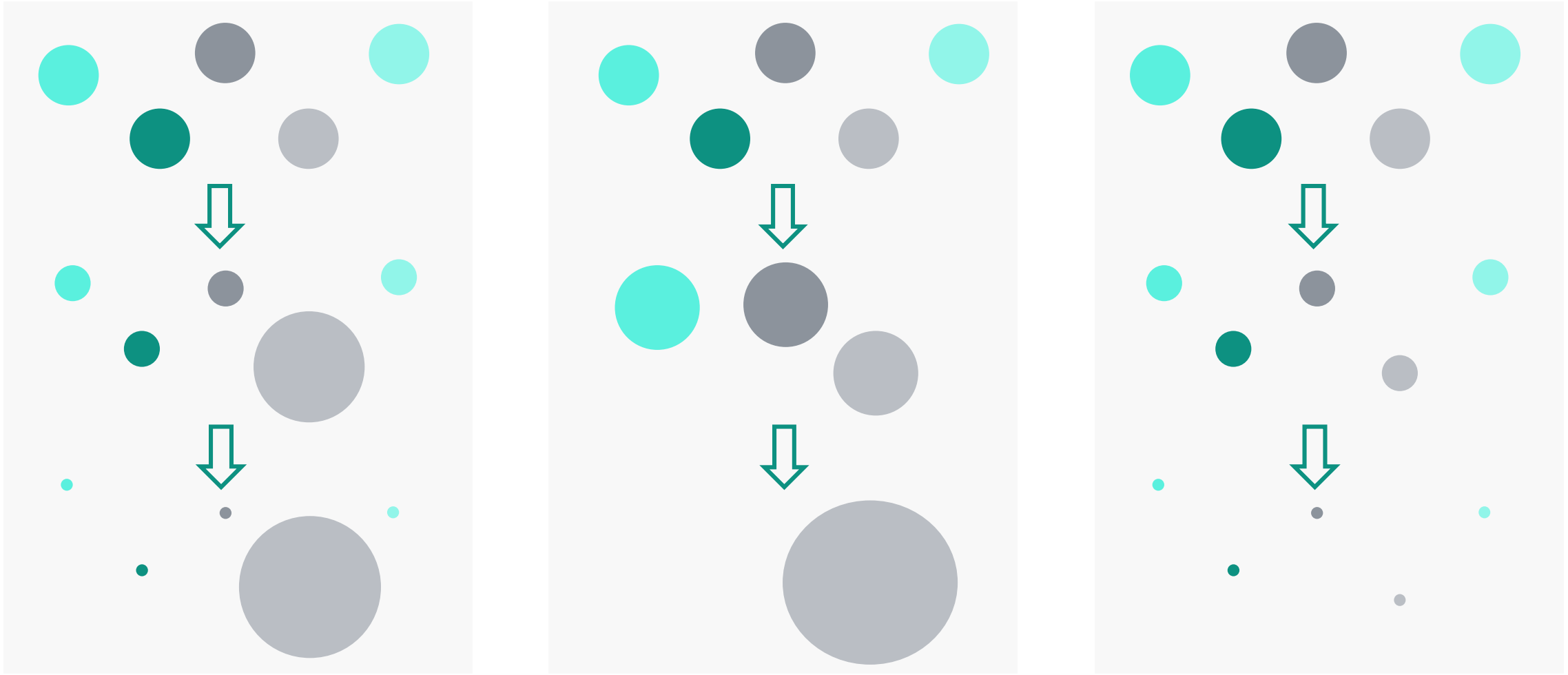
Quality metrics

- Do not use only pLDDT or perplexity
- Use both pLDDT and Perplexity
- Use scPerplexity/ scTM
- OmegaFold, ESMFold, ProteinMPNN





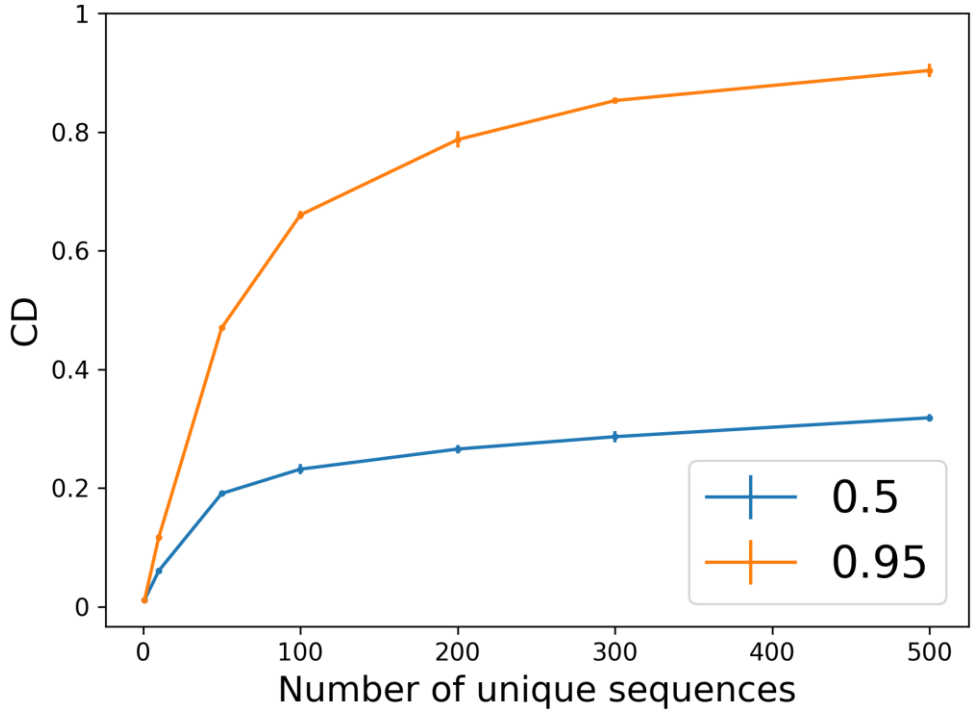
How to identify diverse generation?

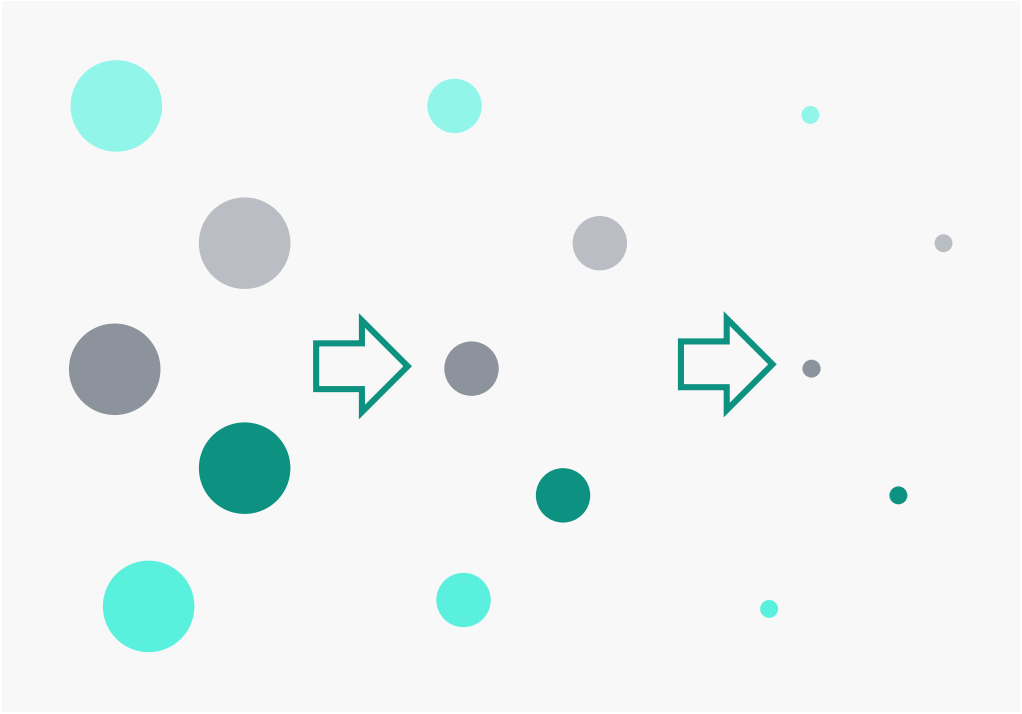


Inner diversity

- Clustering method: MMseqs2
- 2 thresholds: 0.5 and 0.95

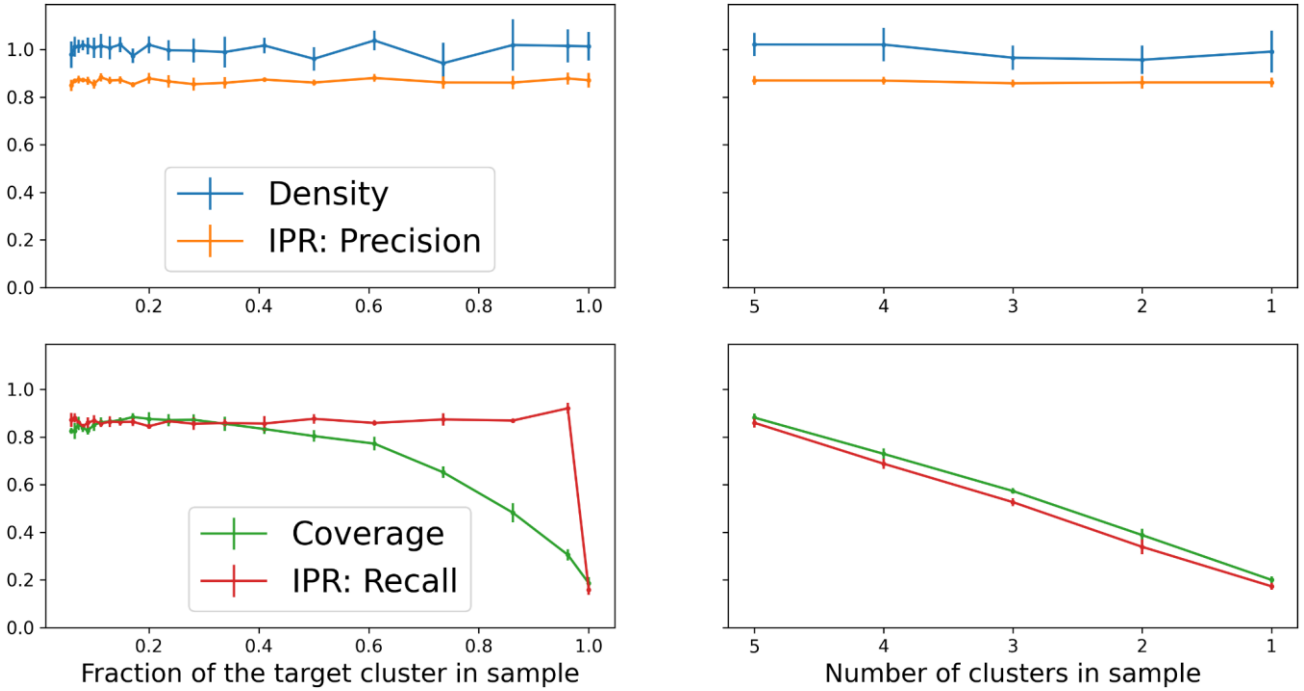
$$CD = \frac{\#Clusters}{\#Seqs}$$

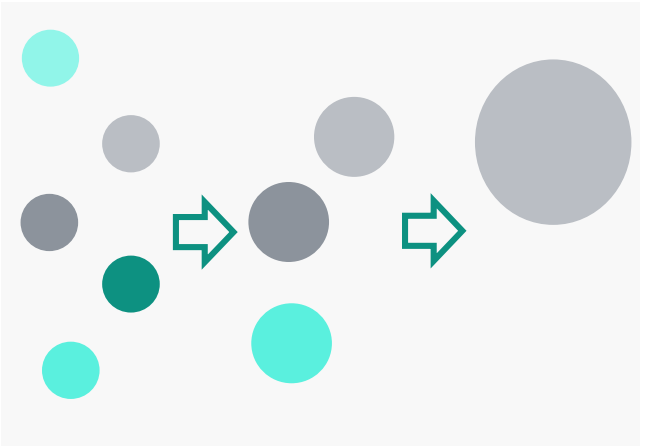


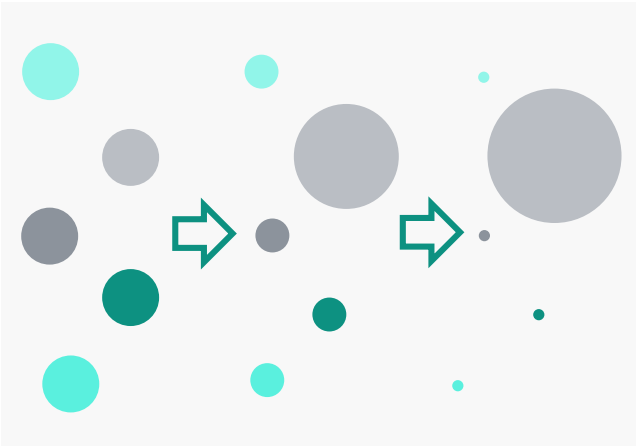


Mixed metrics: quality+diversity

- DC is better than IPR



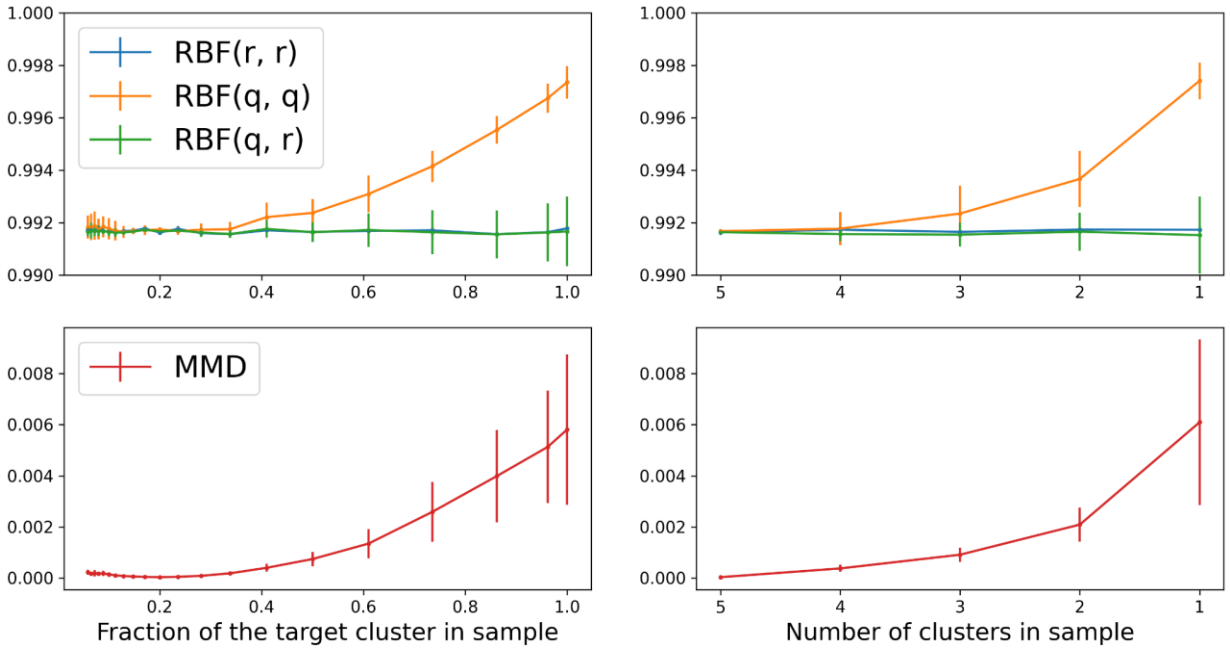


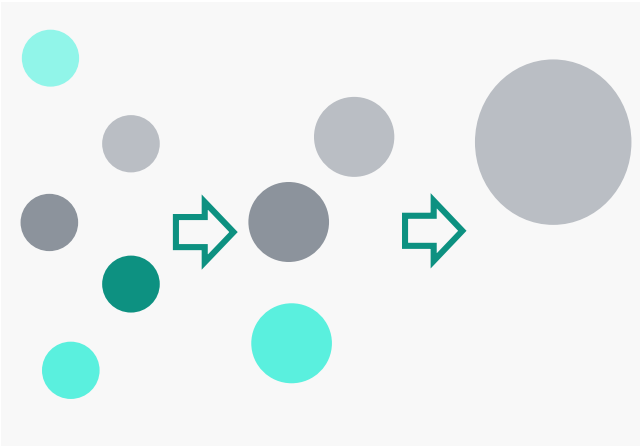


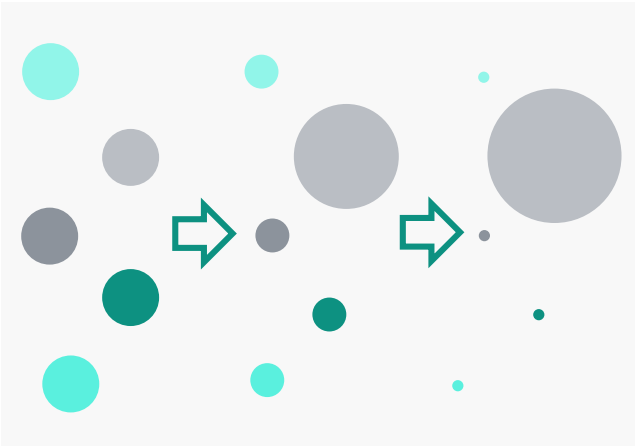
https://doi.org/10.48550/arXiv.2002.09797
https://doi.org/10.48550/arXiv.1904.06991

MMD can be used this way

$$MMD^2(P,Q) = \mathbb{E}_{x\sim P}\left[k(x,x')\right] +$$
$$+ \mathbb{E}_{y\sim Q}\left[k(y,y')\right] - 2\mathbb{E}_{x,y\sim P,Q}\left[k(x,y)\right]$$

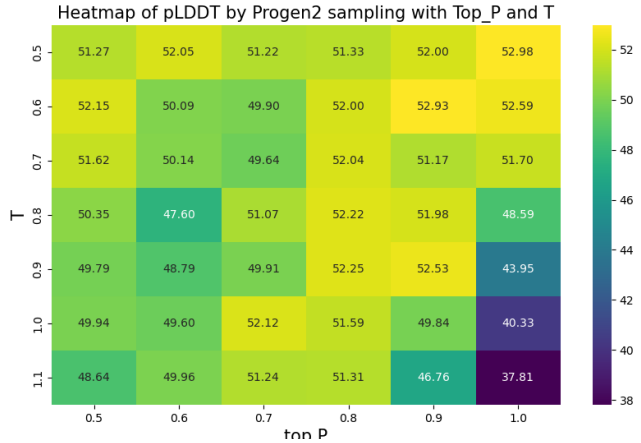


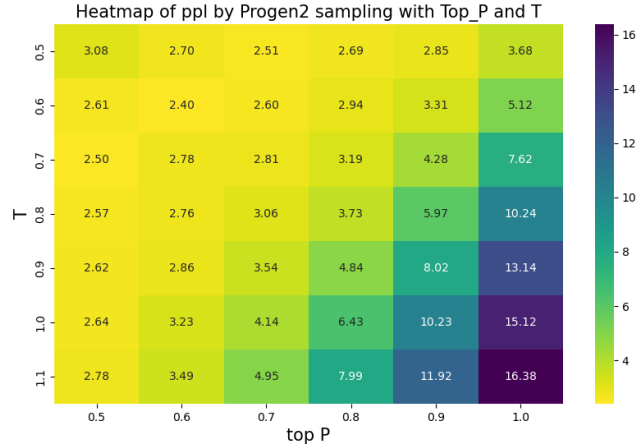




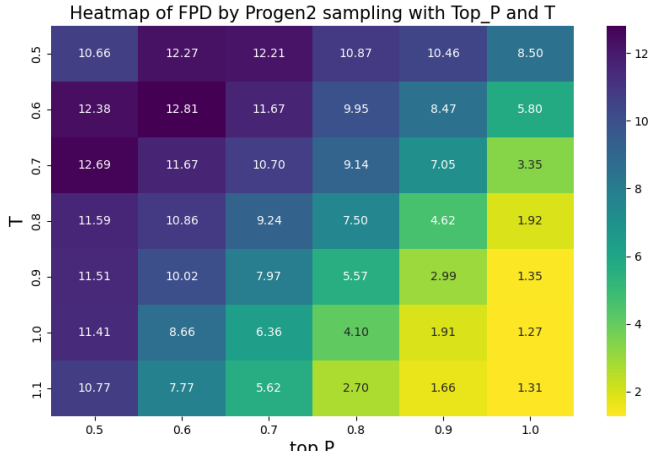
Quality vs Diversity tradeoff

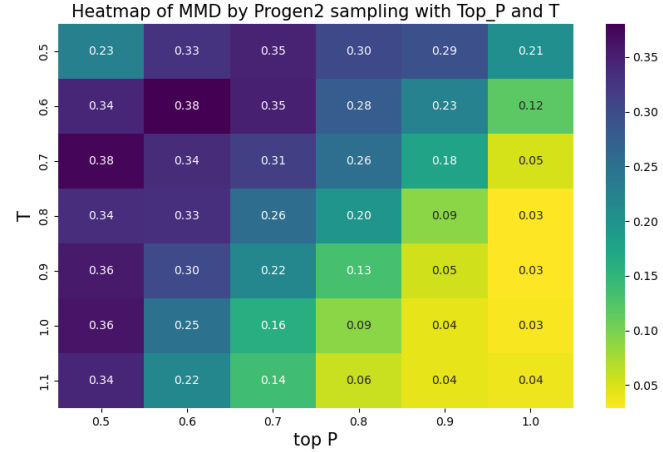
Quality



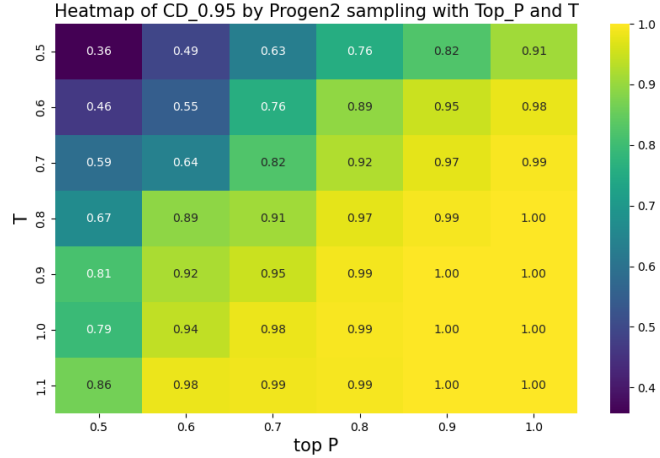


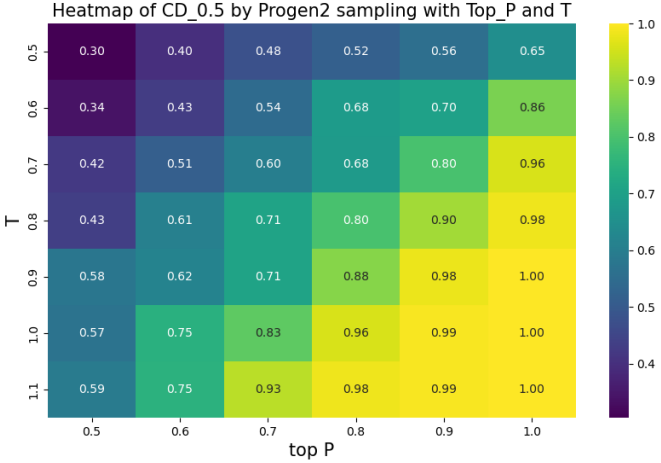
Distribution





Diversity

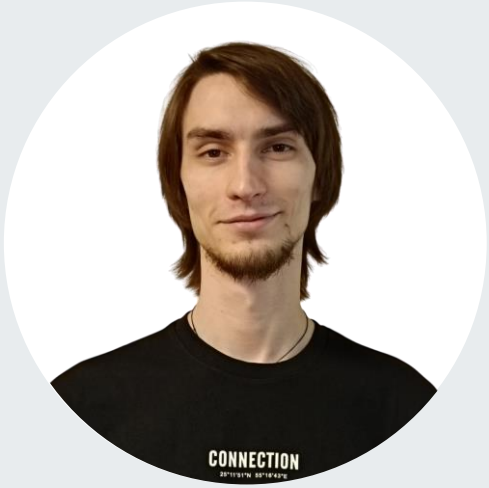




Models comparison

Generative model	pLDDT (↑)	ppl (↓)	$CD_{0.95}$ (↑)
RFDiffusion-80M	76.7	12.07	1.0
ProtGPT2-738M	63.0	7.79	1.0
ProGen2-151M	46.2	12.78	1.0
ProGen2-2.7B	52.2	11.78	0.994
ProGen2-6.4B	57.2	9.71	1.0
EvoDiff-38M	40.2	17.46	1.0
EvoDiff-640M	40.5	17.35	1.0
ProLLAMA-7B	53.1	10.50	1.0
RITA-85M	40.3	18.34	1.0
RITA-300M	41.5	19.10	0.990
RITA-680M	42.5	20.48	0.958
RITA-1.2B	42.6	19.39	0.966
DPLM-150M	81.8	3.90	0.917
DPLM-650M	81.7	4.36	0.943
DPLM-3B	83.1	4.16	0.732
DiMA-33M	83.3	5.07	0.992

Contacts



Andrey Shevtsov

Research engineer, AIRI

Mail: Shevtsov@airi.net @Andr_Shevtsov

AIRI

airi.net

airi_research_institute

AIRI Institute

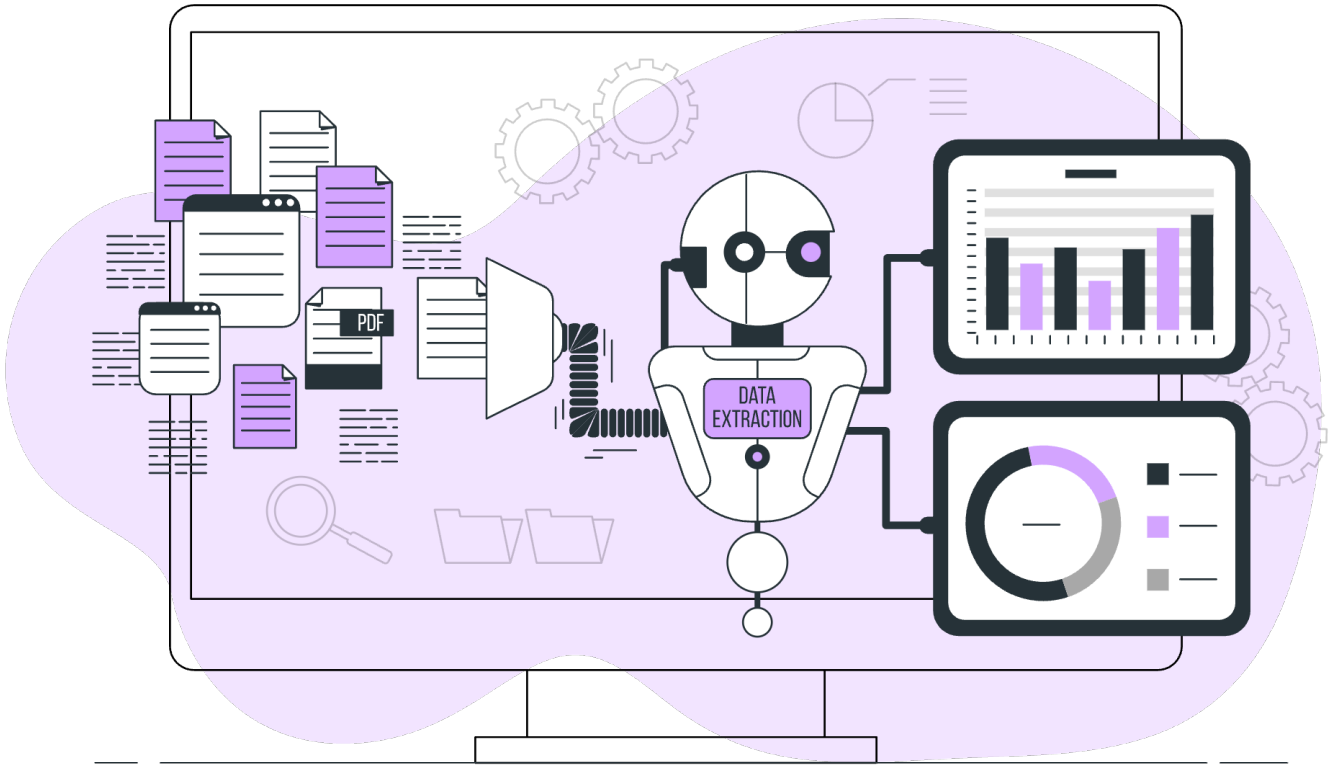
Telegram

AIRI

Управление данными и разработкой как основа для применения ИИ

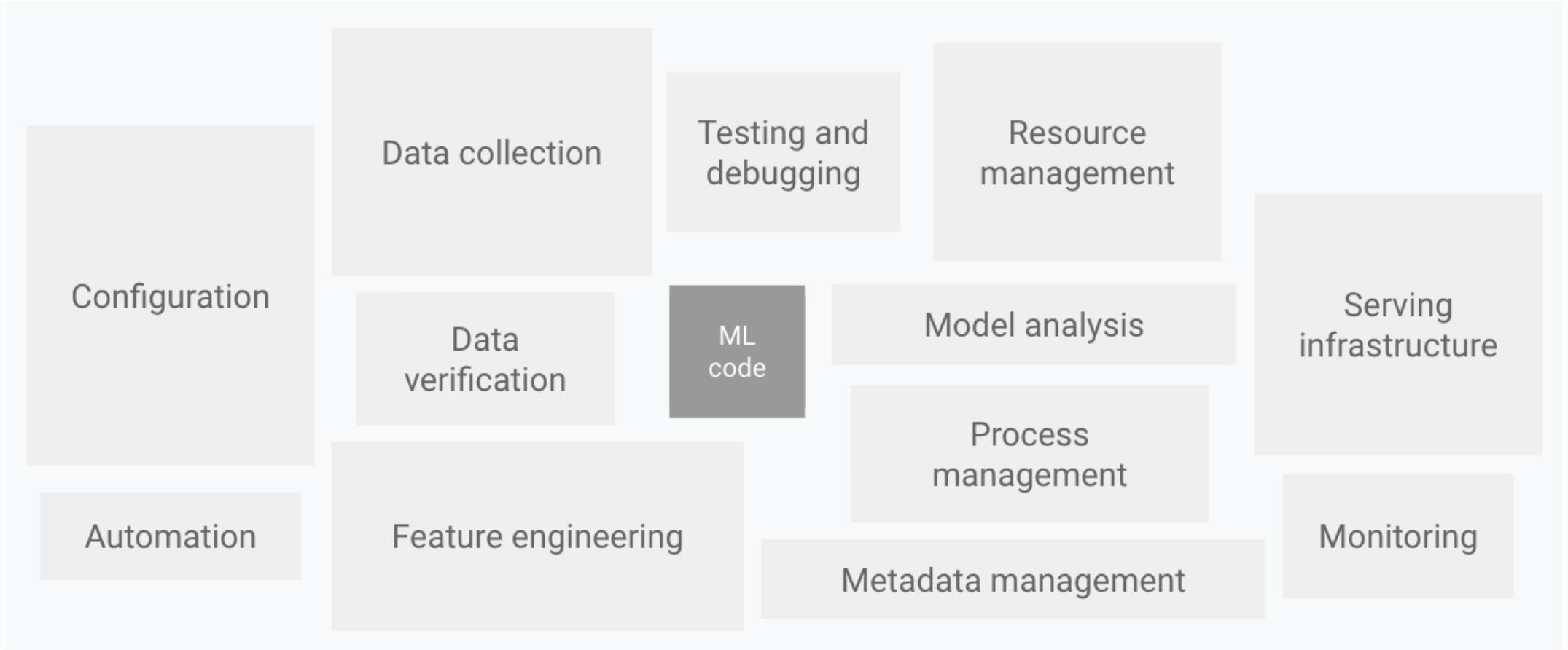
Загорулькин Дмитрий Эдуардович

Заместитель директора Центра стратегической аналитики и больших данных
dzagorulkin@hse.ru



Fall into ML
’24

Составные части ML системы



https://proceedings.neurips.cc/paper_files/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf

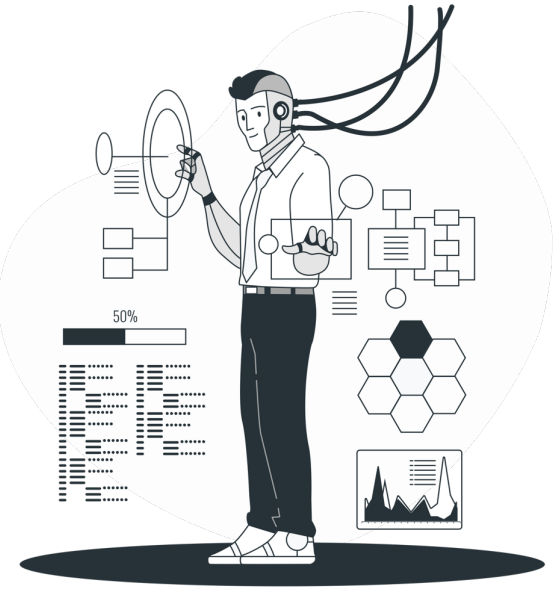
Создание AI проектов

01 Процессы

02 Инфраструктура

03 Данные

04 Моделирование



Понятная постановка задачи

Нам нужно сделать модель, которая определяет уровень технологии.

ПЛОХАЯ ПОСТАНОВКА ЗАДАЧИ

Мы собираем данные об упоминании различных технологий в СМИ. Данные собираются регулярно. Нам необходимо улучшить точность оценки минимум на 20%. Также нужно учитывать, что технологии могут относиться к разным отраслям и иметь разный жизненный цикл. Уровни нужно агрегировать с 10 до 3 классов.

ХОРОШАЯ ПОСТАНОВКА ЗАДАЧИ

Метрики — они для всех разные!

01

Бизнес сфокусирован на получении прибыли!

Бизнес оперирует понятными метриками, такими как процент оттока клиентов, время, проведенное в продукте, количество пользовательских входов в месяц и т.д.

02

Дата-сайентист оперирует модельными метриками (P/R, Accuracy, F1 и другие)

03

Для получения профита от внедрения ML решения важно связать модельные метрики и метрики бизнеса

Важно найти и протестировать это влияние как можно скорее

Высокие модельные метрики не гарантируют улучшение бизнес-метрик!

Управление AI проектом

SCRUM

Подходит под ИТ-проекты

CRISP-DM (Cross-Industry Standard Process for Data Mining)

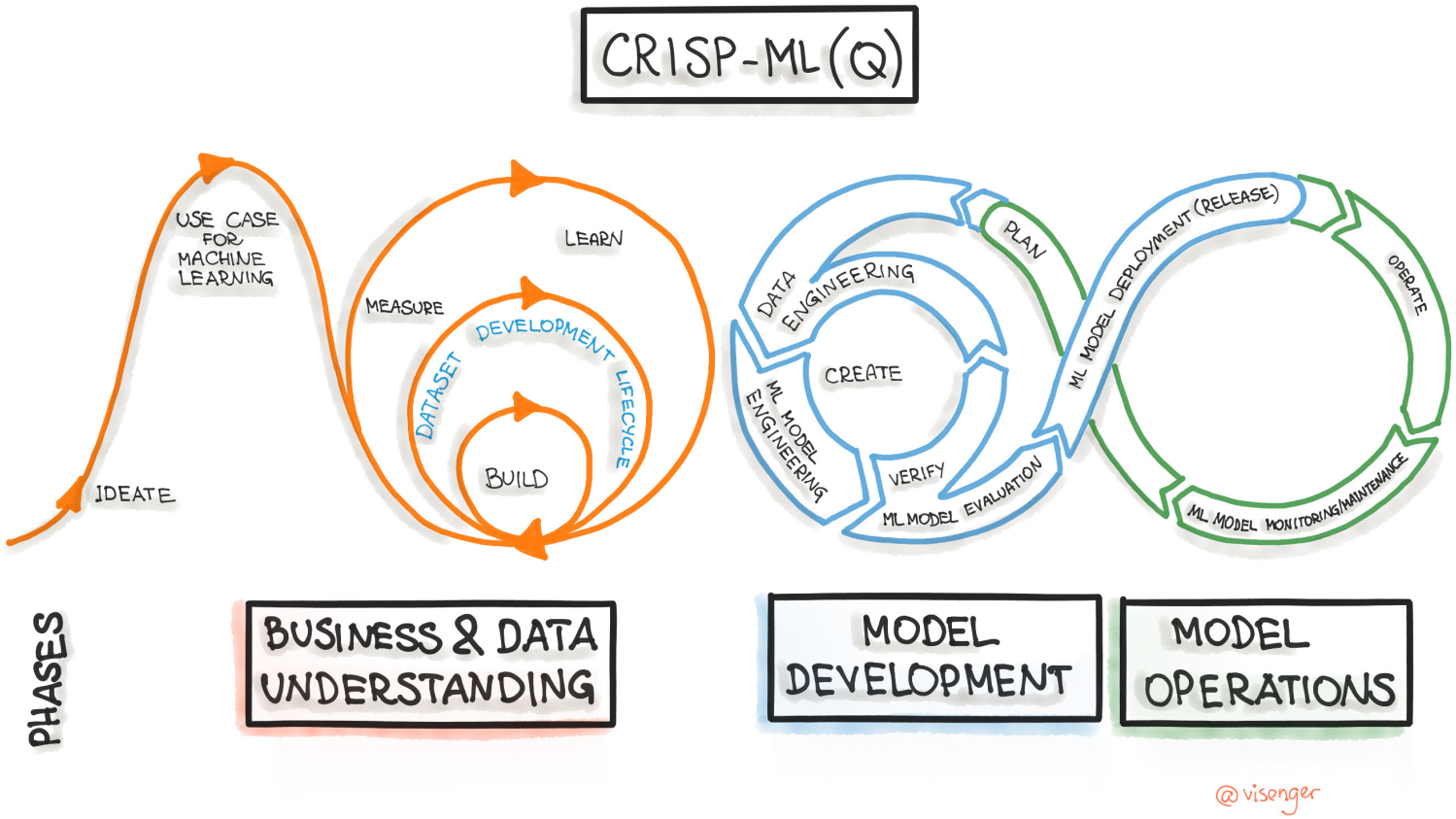
1. Понимание бизнес-целей (Business Understanding) – необходимо привлечение всех заинтересованных сторон
2. Понимание данных (Data Understanding) – проведение разведочного анализа и другие проверки данных
3. Подготовка данных (Data Preparation) – консолидация, агрегация
4. Моделирование (Modeling) – выбор методологии и построения модели
5. Оценка результата (Evaluation)
6. Внедрение модели/процесса (Deployment)

Проблемы: Не предназначена под итеративный процесс улучшения. Отсутствует QA.

CRISP-ML(Q)

Расширение методологии

CRISP-DM сравнение с CRISP-ML(Q)



CRISP-ML(Q)	CRISP-DM
Business & Data Understanding	Business Understanding
	Data Understanding
Data Preparation	Data Preparation
Modeling	Modeling
Evaluation	Evaluation
Deployment	Deployment
Monitoring & Maintenance	-

https://ml-ops.org/content/crisp-ml | https://arxiv.org/pdf/2003.05155

Данные

Основа построения систем машинного обучения

Существуют разные типы данных. Для одних моделей нужны сильно структурированные данные, для других — нет (картинки, видео, книги и таблицы в БД)

Отсутствие культуры работы с данными зачастую является тормозом внедрения систем машинного обучения в бизнес-процессы

Необходимо развивать культуру работы с данными в части автоматической проверки качества данных и других data governance подходов (ведение дата-каталогов и управление метаданными и др.)

Не все данные подходят для моделирования!

Управление данными

DWH (Data Warehouse)

Сложность внесения изменений • Долго строить • Нужно хорошо понимать данные • В основном для структурированных данных

DataLake

Единая точка получения данных • Нужно следить за качеством и метой • Нет ACID транзакций, эволюции схем

DataLakehouse = DataLake + DWH

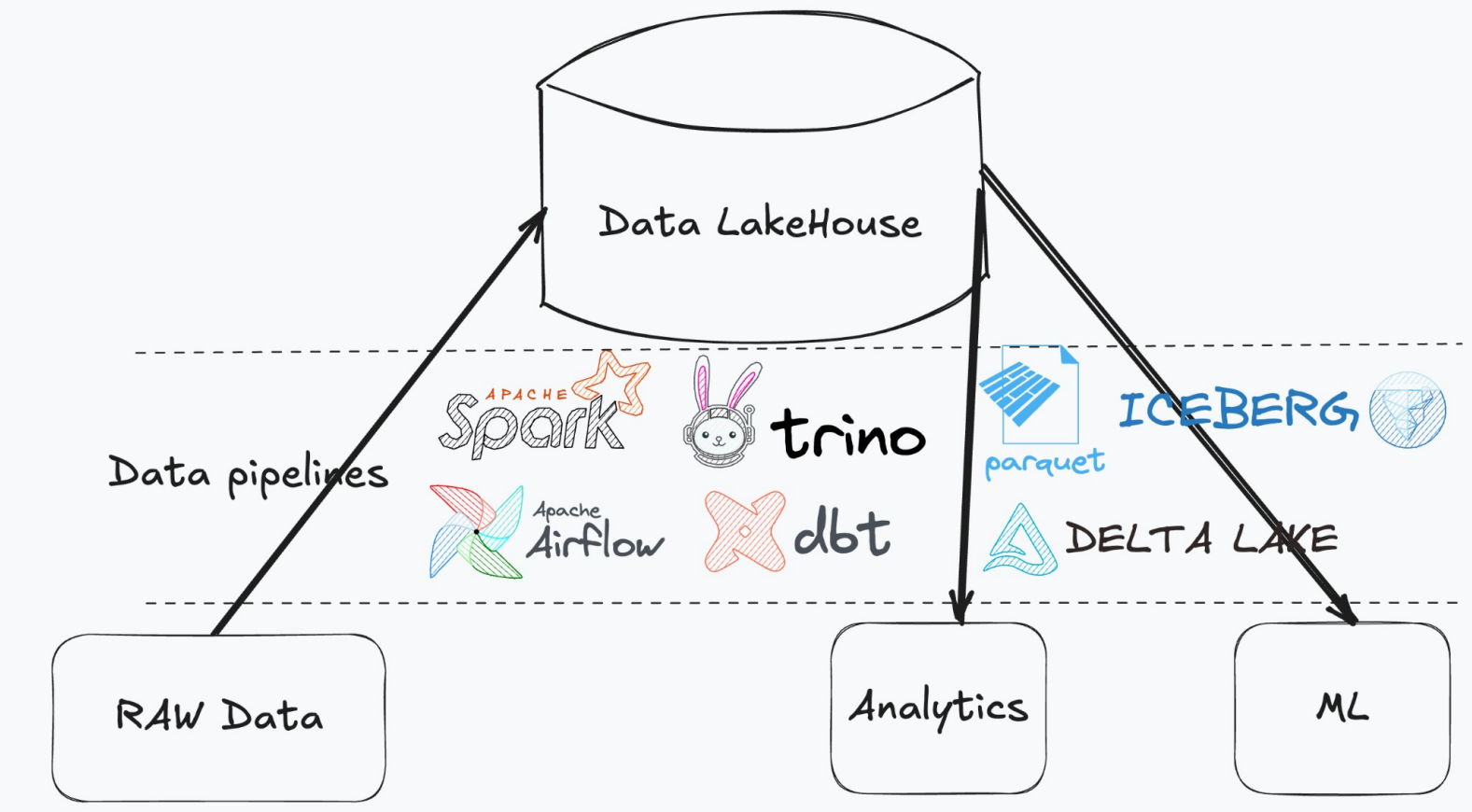
DataMesh

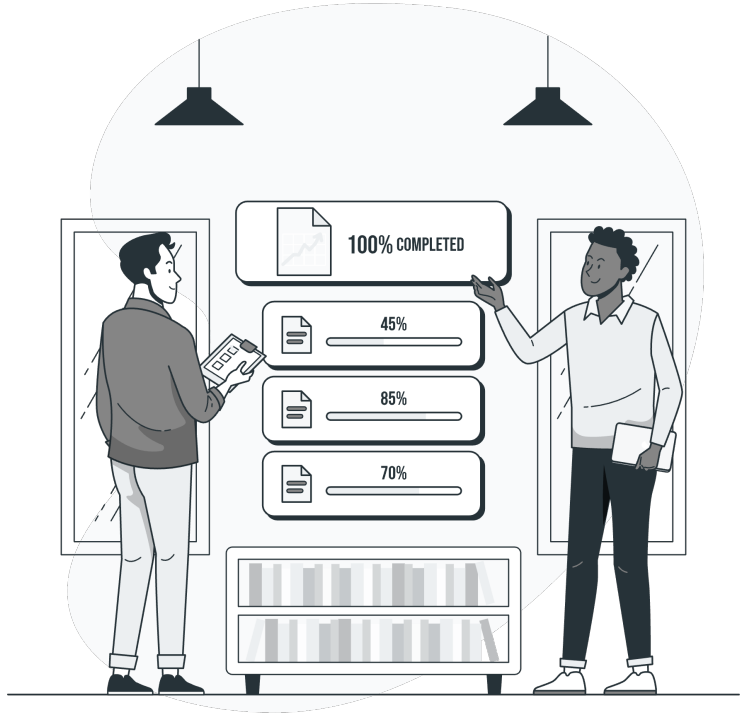
Организационный подход к управлению данными • Полностью децентрализованный • За отдельные данные отвечает владелец, который передает их при необходимости потребителям

DataFabric

Включает в себя элементы DataLakehouse и DataMesh и вносит дополнительно элементы data governance

Процесс работы с данными





Инфраструктура для ML (DL)

Критерий	Своя	Облачная
Обслуживание инфраструктуры	Нужны квалифицированные инженеры (высокие операционные издержки)	–
Инженеры DevOps/MLOps	+	+
Готовность крупного бизнеса передавать чувствительные данные (безопасность)	Нужны специалисты по безопасности	Нет
Расширяемость (Managed решения)	- / +	Ограничена предоставляемыми сервисами
Vendor lock	Нет	Да
Гибкая масштабируемость	Ограничена физическими серверами и используемыми инструментами	Да
Отказоустойчивость	Нужно несколько ЦОДов	+
Цена (особенно при использовании GPU ускорителей)	Получается на порядок дешевле на длинной дистанции	Pay as you go

Вывод: использование только облачной инфраструктуры для DL выходит дорого, идеально использовать гибрид, если это возможно для бизнеса. Разработка внутри, инференс для клиентов снаружи.

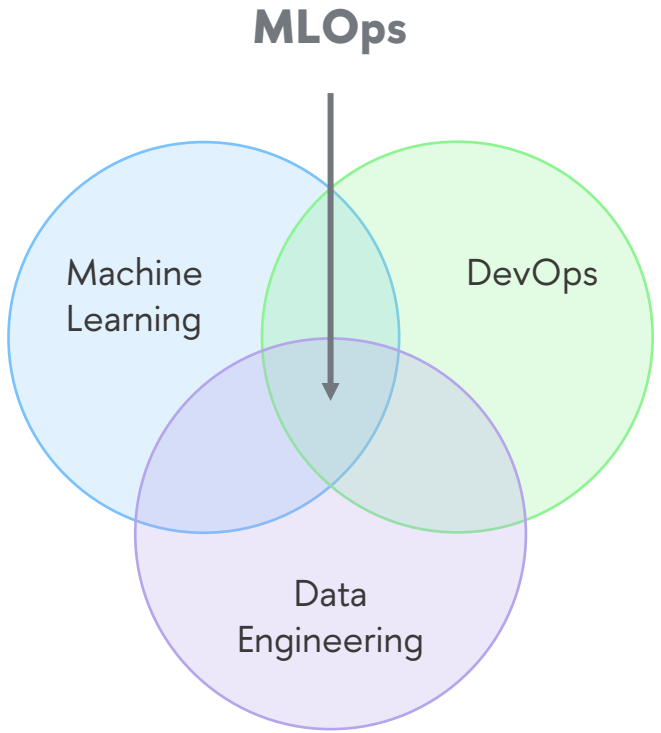
Всегда проводите оценку, какие ресурсы потребуются под вашу систему и будет ли готов бизнес на такие расходы!

MLOps

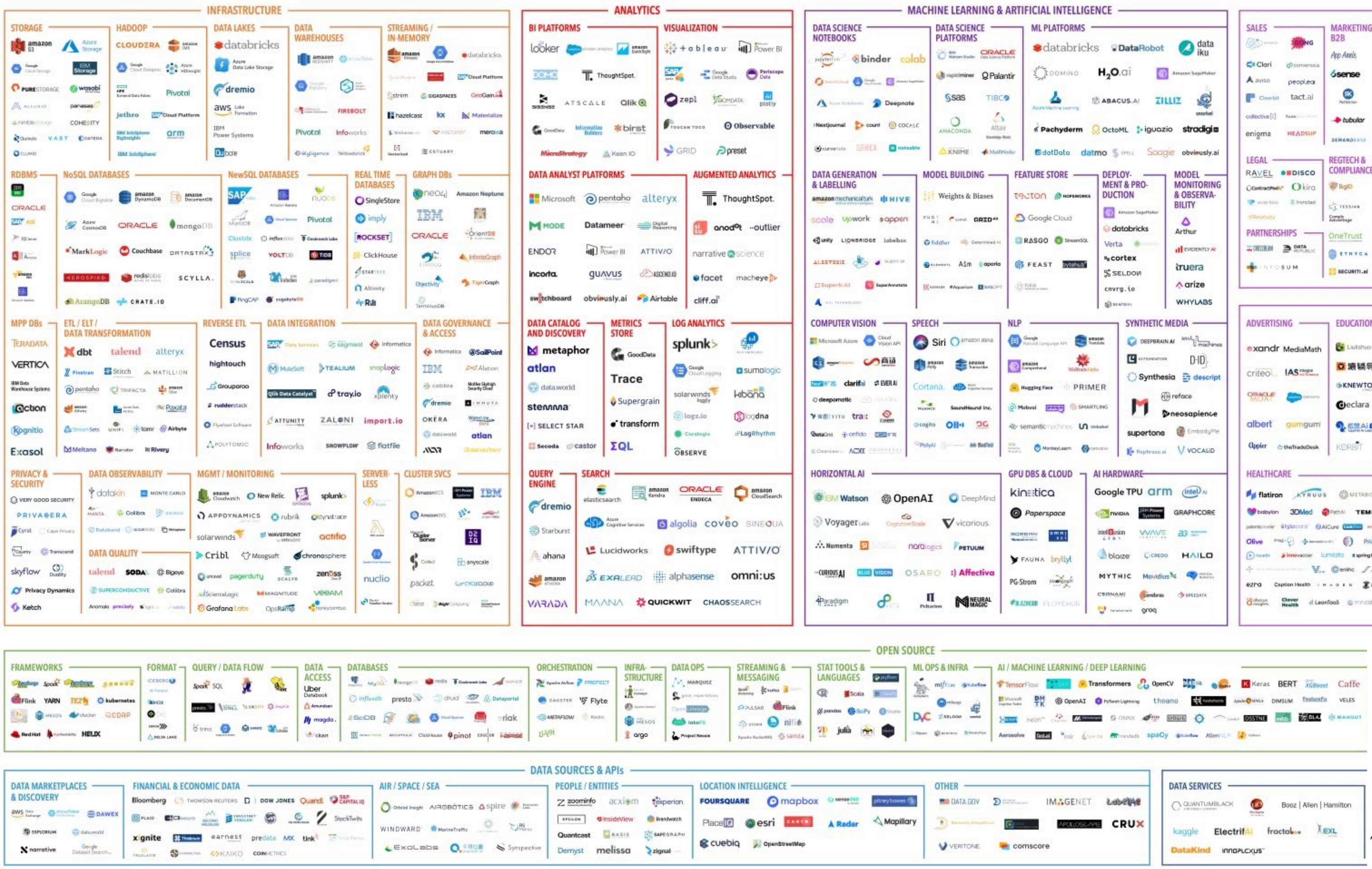
Основные плюшки:

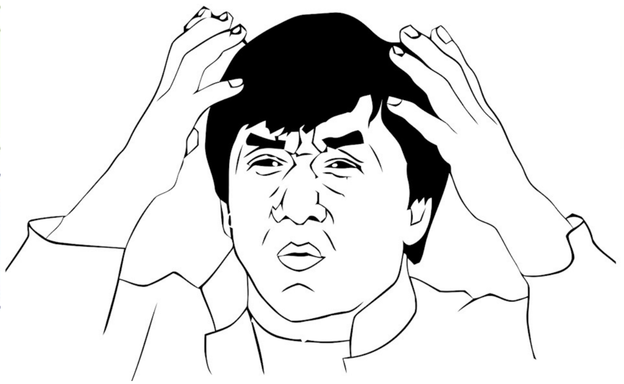
- Автоматизация рутинных процессов
- Масштабируемость
- CI/CD для моделей (canary deploy)
- Воспроизводимость
- Улучшения взаимодействия между командами (DS,OPS,DEV)
- Мониторинг решений

Большое количество инструментов разного уровня зрелости.
Может быть сложно интегрировать их между собой и в бизнес-процессы компании.



MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, AND DATA (MAD) LANDSCAPE 2021



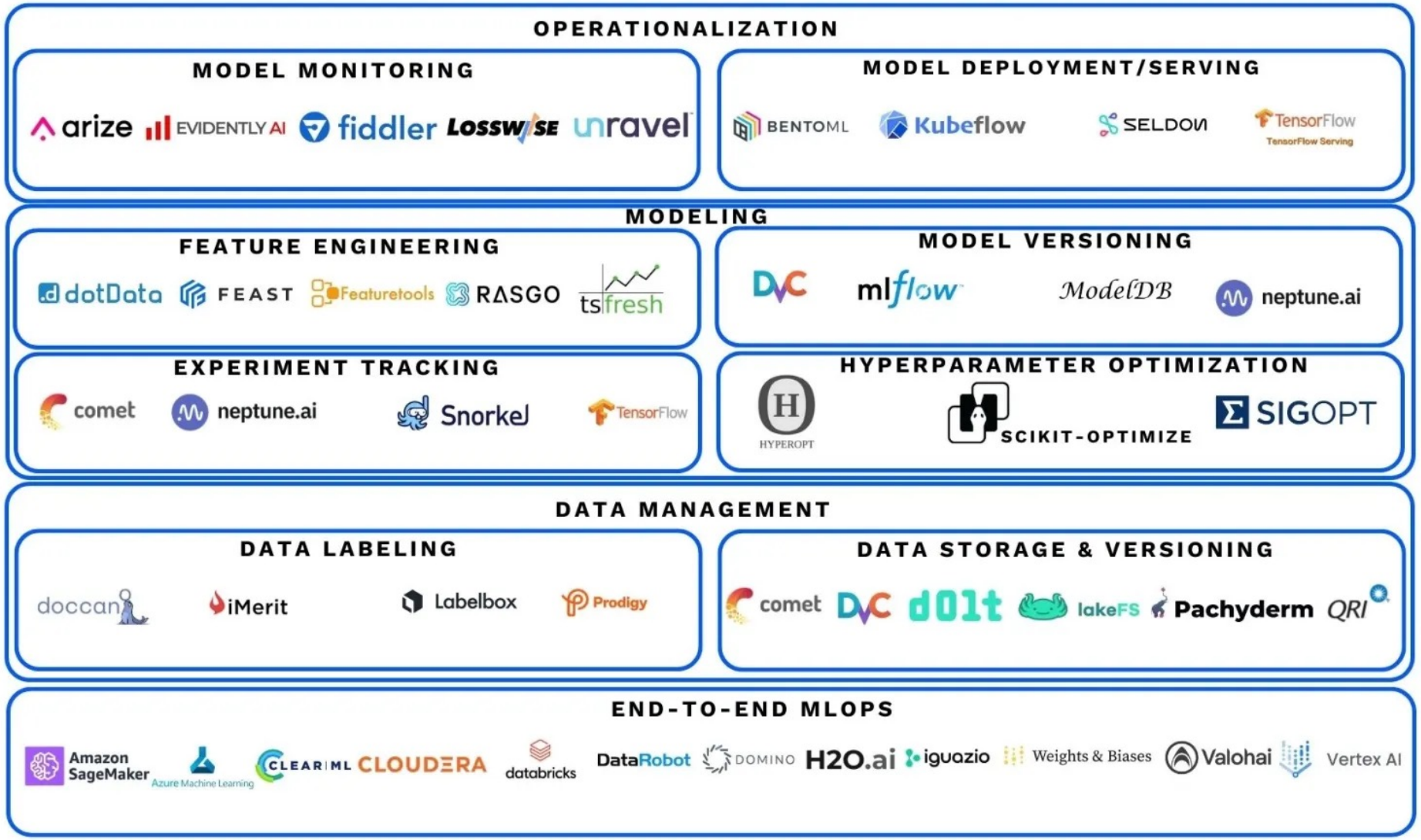


Version 3.0 - November 2021

© Matt Turck (@mattturck), John Wu (@john_d_wu) & FirstMark (@firstmarkcap)

mattturck.com/data2021

Распространенные MLOps-инструменты



Основные инструменты

Версионирование и отслеживание экспериментов: MLflow, DVC (Data Version Control), WANDB

CI/CD пайплайны: Jenkins, GitLab CI, CircleCI, Kubeflow Pipelines

Контейнеризация и оркестрация: Docker, Kubernetes

Мониторинг: Prometheus, Grafana, AWS CloudWatch, Azure Monitor

DataOps: Apache Airflow, Prefect, Kafka, DBT

ML Платформы: Weights & Biases, Neptune.ai, ClearML.

Сервинг моделей: TensorFlow Serving, TorchServe, Flask, FastAPI, KServe, Nvidia Triton, Nuclio (Serveless)

Различие в инфраструктуре при обучении модели и для инференса

	Обучение	Инференс
Цель	Улучшение качества модели	Меньшая задержка на ответ
Вычислительные ресурсы	Необходимы большие ресурсы, в т.ч. GPU	Ресурсов требуется меньше, GPU тоже меньше объема
Загрузка	Батч загрузка, много длинных задач	Запрос/Ответ, может быть батч. Оптимизировано под быстрый ответ
Необходимость масштабирования	Большие кластеры для обучения крупных моделей	Автомасштабирование при увеличении нагрузки
Задержка	-	Адекватное для человека время на ответ
Дисковое пространство	Большие данные, озера данных	Минимальные объемы (модель, конфиг, логи, дополнительный код)
Отказоустойчивость	-	Высокая избыточность, важность безотказной работы
Мониторинг	Mlflow и подобные для трекинга экспериментов	Постоянно следить за качеством модели, оценивать data drift и performance degradation

Деплоймент моделей (инференс)

По типу выполнения:

Онлайн инференс:

- Специализированные облачные решения (AWS SageMaker, GCP Cloud Run, Selectel Inference Platform и др.)
- Или развернутые там же opensource инструменты

ОБЛАКО

- Существует множество решений, подходящих под разные технологии, но нет единого стандартного
- Необходимо подбирать инструмент под конкретную технологию
- С развитием LLM начали появляться специализированные решения для инференса с возможностью запуска (квантизованных) и обычных моделей ollama, vllm и т.д.

СВОЯ ИНФРАСТРУКТУРА

По типу:

- REST/GRPC
- Model as a Service (MAS)
- Model-on-Demand
- ...

Оффлайн инференс:

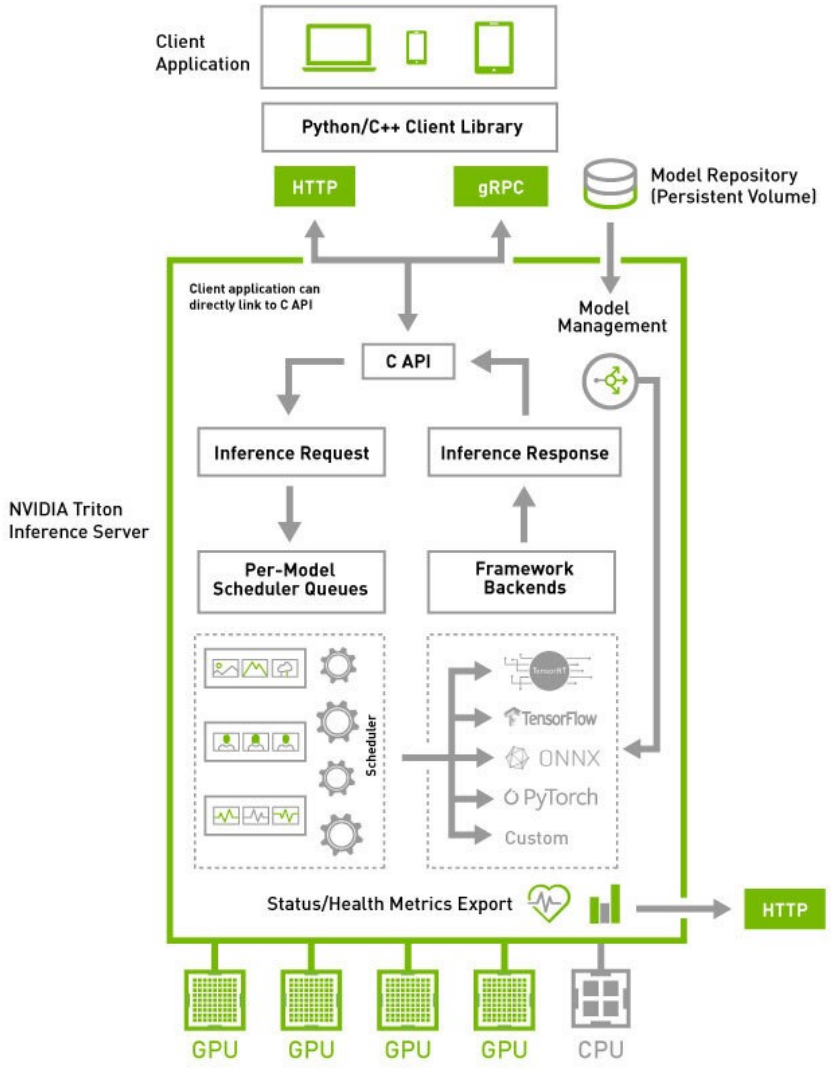
- Батч обработка
- REST/GRPC

Деплоймент Model as a Service (REST/GRPC)


```
from fastapi import FastAPI
from pydantic import BaseModel

class ModelRequest(BaseModel):
    name: str
    price: float

app = FastAPI()

@app.post("/predict/")
async def next_best_offer(request: ModelRequest):
    return modelService.getModel().predict(request)
```

+	Ds-way подойдет для быстрого прототипа решения
–	Нет поддержки cloud native и интеграции с k8s, как следствие, решение не масштабируется
–	Непонятно, как будет работать под нагрузкой?
–	Быстрее будет собирать запросы в батчи или отправлять в модель только по одному?
–	Если потребуется добавить новый функционал (например, авторизацию и аутентификацию), нужно внести изменения в код. Изменения могут быть довольно частыми
–	Нет мониторинга качества и непонятно, когда модель станет деградировать
–	Может меняться как код, так и сама модель, придется все переписывать заново

Inference Servers



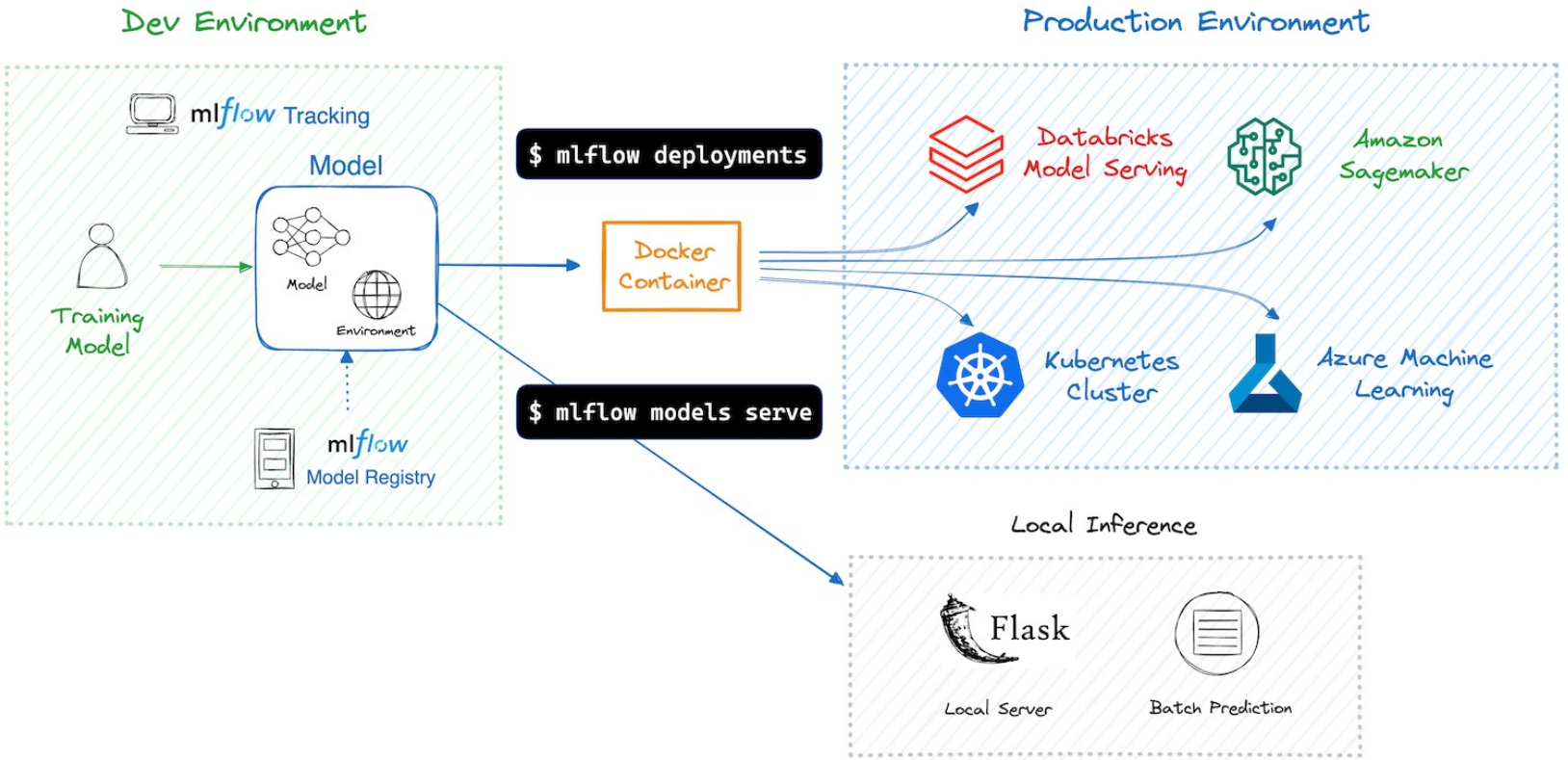
Triton Inference Server

+	Одновременное выполнение моделей
+	GPU/CPU
+	Адаптивный (динамический) батчинг
+	Поддержка практически всех известных бэкендов (TensorRT, TensorFlow, PyTorch, ONNX, OpenVINO, Python и другие)
+	Горячая замена моделей
+	Много дополнительных функций
–	Весьма сложен

Другие:

- TFServing - tensorflow
- TorchServe - pytorch

MLFlow Serving



Локальный инференс

–	Для тестирования
–	Не подходит под прод нагрузку
+	Быстро и просто развернуть

Прод инференс

+	Лишен недостатков локального
+	Асинхронный, отдельный пул обработчиков, можно одновременно сервить несколько моделей

Model Serving. Kserve

- Поддерживает большинство известных райнтаймов TFServing, TorchServe, Triton Inference Server

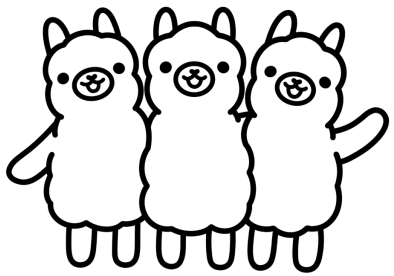
- Cloud native
  - Масштабирование на CPU/GPU
  - Управление версиями
  - Батчинг
  - Логирование (запрос/ответ)
  - Управление трафиком
  - Метрики

и др.

Специализированные решения

LLM - vLLM, Ollama, llama.cpp

- llama3.2 1b, 3b
- llama3.1 8b,70b,450b
- gemma2 2b,9b,27b
- qwen2.5 0.5...72b
- mistral-nemo 12b
- mistral
- mixtral
- llava 7b,13b,34b



https://ollama.com/library

ЗАГОРУЛЬКИН ДМИТРИЙ ЭДУАРДОВИЧ

Заместитель директора Центра стратегической аналитики и больших данных ИСИЭЗ НИУ ВШЭ

Заместитель руководителя системы интеллектуального анализа больших данных iFORA

dzagorulkin@hse.ru

Сайт iFORA

iFORA в Telegram

iFORA-экспрессы

Хотите у нас работать или пройти стажировку?
Сканируйте QR-код

Кейсы решения бизнес-задач с помощью автоматической обработки больших данных

Вишневский Константин Олегович

Директор Центра стратегической аналитики и больших данных, PhD
kvishnevsky@hse.ru

Fall into ML
’24

Центр стратегической аналитики и больших данных – ведущий российский think tank в сфере новых технологий

Отдел разработки интеллектуальных систем	Отдел информационно-аналитических систем	Отдел исследований больших данных	Отдел исследований цифровых технологий

Активности:

- **Консалтинг** для государства и бизнеса (проектная деятельность)
- **Разработка** системы/платформы, ИТ-продуктов и моделей
- **Научные исследования** (в т.ч. публикации в топовых журналах и международные конференции)
- **Обучение**, ДПО, мастер-классы, тренинги, стажировки

Инструменты:

- **Система iFORA** (intelligent FOResight Analytics)
- **Форсайт** и technology roadmapping
- **Технологическая аналитика**
- **Мониторинги и обследования** компаний
- …

Стремительный рост объема и сложности данных трансформирует сферу аналитики

Традиционная ручная аналитика → **Развитие систем автоматизированного анализа больших данных** → **Аналитика на основе новейших технологий NLP**

Смещенная выборка источников

- Огромный объем информации, который невозможно обработать вручную
- Выбраны случайно
- Общедоступны
- Не всегда высокого качества
- Устаревшие

Аналитик

- Слишком узкая специализация, консерватизм, ограниченное знание мировой повестки
- Торопится и делает ошибки
- Лоббирует определенные интересы

Недостоверная информация

- Из-за повсеместного внедрения технологий генеративного ИИ возникают риски распространения недостоверной информации

Все доступные источники

- Многие миллионы документов
- Полные тексты
- Разнообразные форматы данных
- Отбор по единым объективным критериям качества
- Постоянное пополнение

Автоматический анализ

- Прозрачная, воспроизводимая, валидированная методика
- Снижение рисков «человеческого фактора»
- Высокая скорость выдачи аналитических результатов

Надежные выводы

- Высокое качество и достоверность данных
- Снижение рисков распространения фейков

Fall into ML '24

Мировой рынок ИИ-решений для аналитики больших данных вырастет в 3 раза к 2032 г.

530%
рост глобального объема данных с 2018 г.

10,9 млн
число специалистов по обработке данных в ЕС

50 млрд долл.
глобальные инвестиции в развитие генеративного ИИ за 2020-2023 гг.

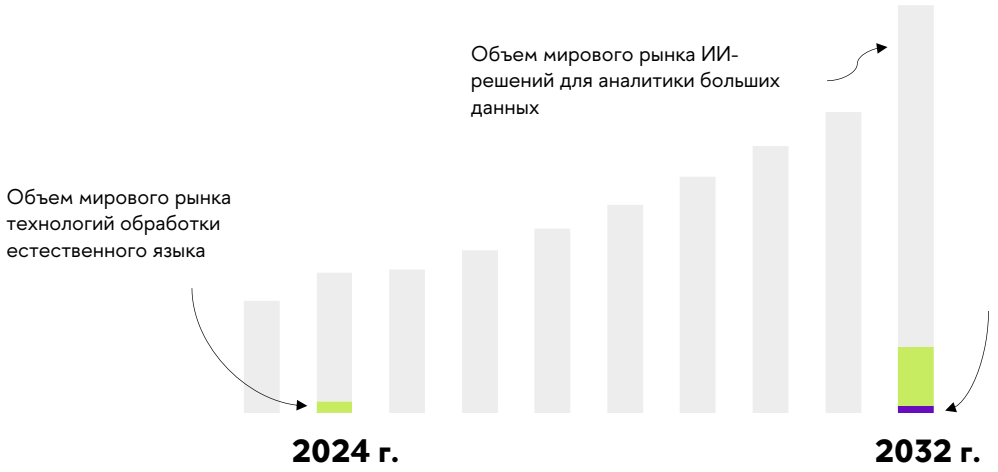
10 новых моделей
генеративного ИИ появляются в мире каждый месяц

51 %
российских организаций используют технологии обработки естественного языка

К 2032 г. в **5 раз** увеличатся объемы мировых рынков:

систем интеллектуального анализа текстов
до **$24,8** млрд

технологий обработки естественного языка
до **$158,1** млрд



Тренды рынка аналитики больших данных

- **технологический прогресс** в IT, динамичное развитие технологий ИИ, ML-алгоритмов, NLP-решений, LLM
- **оперативная аналитика** больших данных для органов власти и компаний
- **продвинутая аналитика** с использованием постоянно совершенствующихся ИИ-инструментов
- **клиентоориентированный** анализ данных
- **предиктивная аналитика** для поддержки принятия решений
- автоматизация функций и использование сложных ИИ-решений непрофессиональными пользователями (**демократизация ИИ**)
- цифровые аналитические продукты с **пользовательским интерфейсом** и др.

Fall into ML
'24

**из числа российских организаций, использующих технологии искусственного интеллекта*

Источники: European Commission, Fortune Business Insights, Reports and Data, Департамент предпринимательства и инновационного развития г. Москвы, Агентство инноваций г. Москвы, НИУ ВШЭ

Система iFORA позволяет воспользоваться преимуществами аналитики на основе ИИ

iFORA Intelligent Foresight Analytics

уникальная пополняемая **мультилингвальная** база данных

>800 млн документов

+30 тыс. документов ежедневно

Языки

Русский | Английский | Китайский | Кириллические | Латинские

Аналитика

- стратегическая
- технологическая
- операционная

>390 млн
Научные публикации

>50 млн
Рыночная аналитика и профессиональные СМИ

>3.5 млн
Клинические исследования

>1 млн
Документы международных организаций, консалтинговых компаний

>150 млн
Патенты

>4 млн
Научные проекты / гранты международных и национальных программ / фондов

>3.5 млн
Социальные сети

>300 тыс.
Отчеты о НИР

+ Конфиденциальные кастомизированные датасеты под конкретные задачи Заказчиков

>75 млн
Научно-популярные медиа

>3.5 млн
Данные государственных закупок

>2 млн
Вакансии

>5 тыс.
Образовательные программы

>100 тыс.
Научные конференции

Система iFORA позволяет воспользоваться преимуществами аналитики на основе ИИ

Комплекс количественных и качественных методов

- собственные алгоритмы универсальной обработки текстов
- 20+ обновляемых ML-моделей
- 10+ функциональных витрин данных и микросервисов

Аналитика

- стратегии
- прогнозы
- приоритеты
- долгосрочные программы развития
- программы инновационного развития
- технологические дорожные карты и др.

NLP-решения и сервисы

- автоматическая суммаризация текстов
- диалоговая система на основе генеративного ИИ и метода RAG
- интерактивные интерфейсы и витрины данных

Российская апробация:

>100 проектов по заданиям Аппарата Правительства РФ и заказам ФОИВ и крупнейших компаний

Международная апробация:

- OECD, Париж
- NISTEP, Токио
- Innovation Forum, Шанхай
- Forum on STI, Претория
- Joanneum Research, Вена
- University of Manchester и др.

iFORA™ отмечена в журнале Nature в качестве эффективного инструмента поддержки принятия решений (Nature, 2020, Vol. 583)

iFORA™ отмечена ОЭСР в качестве успешной инициативы в области цифровизации науки (OECD Science, Technology and Innovation Outlook 2018)

Суперкомпьютер cHARISMa ВШЭ получил премию «Приоритет-2020» в области эффективного применения передовых технологий. *Пиковая производительность составляет 2 петафлопса на 2023 г.*

iFORA™ включена в каталог цифровых решений ICT.Moscow (2020)

iFORA™ экспонировалась на Международной выставке-форуме «Россия» среди передовых отечественных достижений в научно-технологической сфере (2023)

iFORA отмечена в сборнике основных результатов научно-исследовательской деятельности Сбера «Наука в Сбере 2023»

Более 40 выпусков оперативной технологической аналитики («iFORA-экспрессов»)

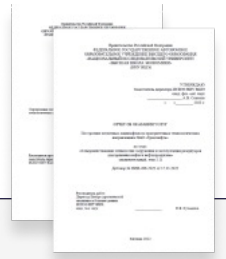
ВЕДОМОСТИ

02 октября, 02:09 / Общество

В России создают систему прогнозирования для разработки инновационных лекарств

В 2024 г. совместно с Сеченовским Университетом началась разработка системы раннего выявления перспективных технологий на основе iFORA

iFORA позволяет получать результаты в разнообразных форматах

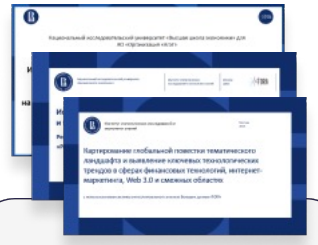


Аналитика

PDF

На бумажном носителе

- Аналитические отчеты
- Краткие аналитические записки и др.



Презентационные материалы

PDF

На бумажном носителе

Результаты исследования в кратком формате, в т. ч.:

- Визуальные представления
- Основные выводы и др.

Таблицы

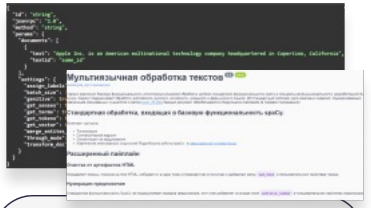
CSV

- Перечни организаций, экспертов, тематик и т.д.
- Статистические данные и др.

Семантические базы

JSON CSV

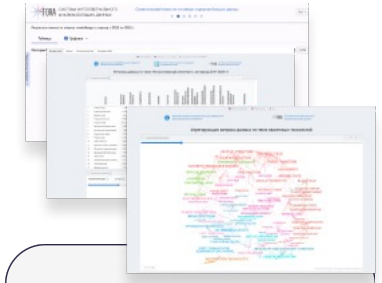
- Полные тексты публикаций СМИ, аннотации научных статей и т.д.
- Метаданные публикаций
- Основные тематики публикаций и др.



Модели машинного обучения

jupyter JSON

- Программный код для решения задач классификации, кластеризации и т.д. на основе обработки больших массивов информации



Пользовательские интерфейсы

HTML

- Инструменты для удобного взаимодействия пользователей с системой iFORA, получения визуализаций, таблиц и проч. по настраиваемым параметрам

Система iFORA основана на модульном подходе и позволяет комбинировать специализированные блоки для конкретных задач

Тренды	Анализ технологического развития	Оценка технологической независимости	Оценки рынков	Прогнозы	Оценка рисков	Анализ правового поля	Региональный анализ	Выявление сетей и центров компетенций	Анализ и прогнозирование профессиональных компетенций	Новейшие NLP-решения / сервисы
Выявление трендов	Картирование научно-технологического ландшафта	Расчет значимости и динамичности технологий в России и мире	Количественные оценки рынков	Формирование консенсус-прогнозов	Анализ конкуренто-способности	Анализ нормативно-правовой базы, стандартов	Выявление барьеров развития регионального бизнеса	Выявление сетей связей организаций	Выявление перспективных профессий, связанных с возникающими технологиями	Автоматическая суммаризация текстов
Оценка значимости и динамичности трендов	Анализ жизненного цикла технологий	Выявление различий в уровне развития отдельных технологий в России и мире	Оценка зрелости рынков	Построение таймлайнов событий будущего	Репутационный анализ	Выявление приоритетов	Репутационный анализ в медиа-пространстве	Определение специализации организаций	Определение наиболее перспективных компетенций	Профильный анализ документов на основе NER-моделей
Анализ структурных изменений	Анализ влияния технологий на сектора	Выбор мер поддержки	Анализ закупок	Выбор направлений развития продуктов	Определение направлений стратегического развития и угроз	Сопоставление российской и международной повесток	Построение независимых рейтингов	Анализ образовательных программ	Формирование проектных команд, подбор специалистов	Разработка интерактивных интерфейсов и витрин данных
Выявление хайпов	Определение уровня готовности технологий	Выявление возможных точек роста	Формирование технологических и продуктовых портфелей	Выявление возможных точек роста	Систематизация и картирование рисков	Анализ пробелов в нормативно-правовой базе	Выявление ключевых направлений для развития и «белых пятен»	Анализ экспертного ландшафта	Сопоставление трендов и спроса на компетенции кадров	Разработка кастомизированных моделей машинного обучения
Определение зарождающихся трендов	...	...	...	...	Определение индикаторов воздействия СМИ и рекламы	...	...	Выявление лидеров проф. сообщества	...	...
...					...			...		

Fall into ML
’24

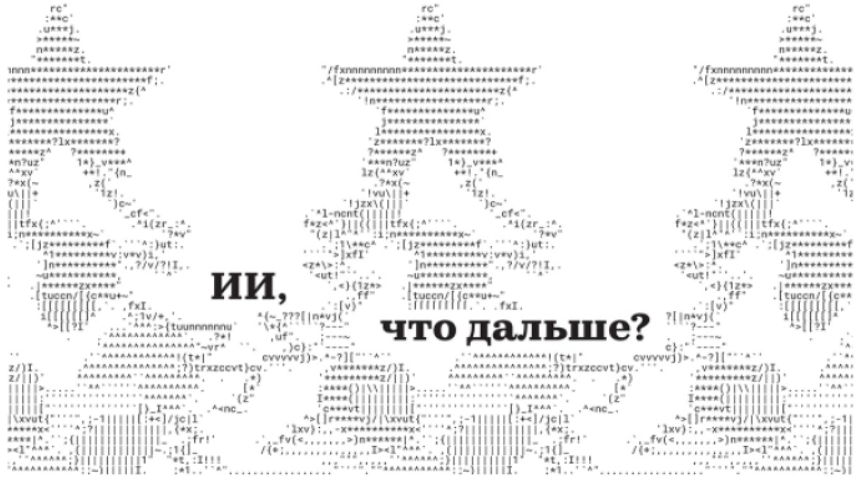
Новогодний выпуск газеты РБК за 2023 г.

Итоги 2023-го, 29 дек, 10:15 | 4 021 | Поделиться

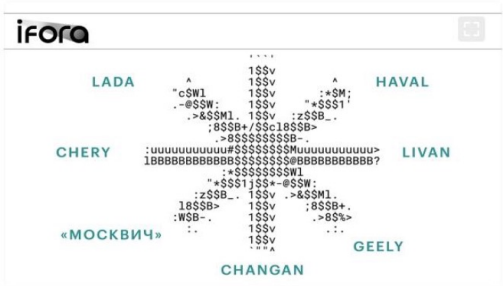
ИИ, что дальше? Итоговый номер газеты РБК

Сюжет
Итоги 2023-го

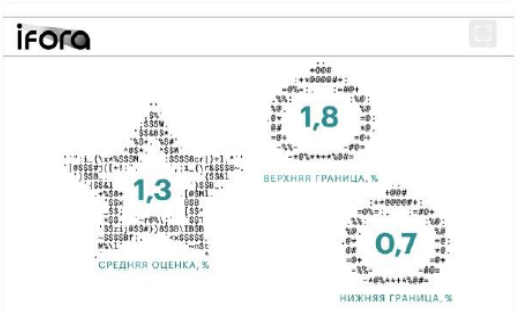
Перед вами финальный выпуск газеты РБК за 2023 год. В отличие от прошлых лет мы решили посвятить его не итогам уходящего года, а прогнозам на ближайшее будущее, причем сформированным российскими системами искусственного интеллекта



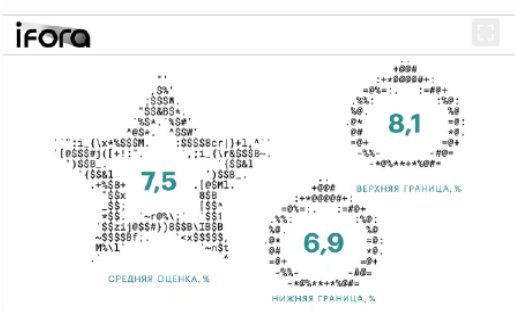
Какие марки машин будут самыми популярными среди россиян в 2024 году?



Насколько вырастет ВВП в 2024 году?



Какая будет инфляция в 2024 году?

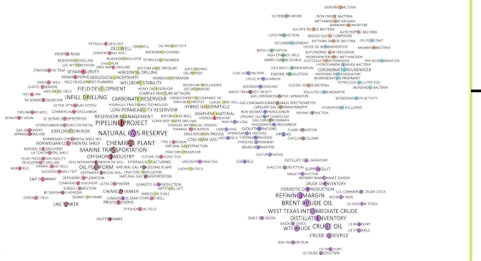


Сколько будет стоить нефть в 2024 году?

https://www.rbc.ru/technology_and_media/29/12/2023/658d6c149a79478079562474

Возможности комплексного кастомизированного анализа технологических разработок

Картирование технологического ландшафта



Семантическая карта

Анализ структурных изменений технологического ландшафта

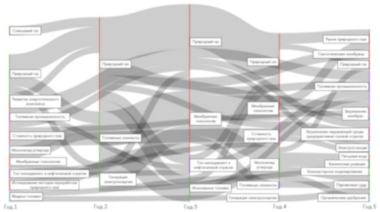
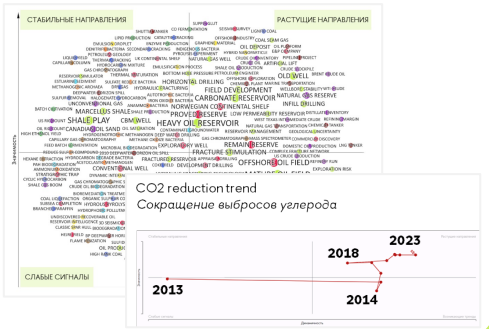


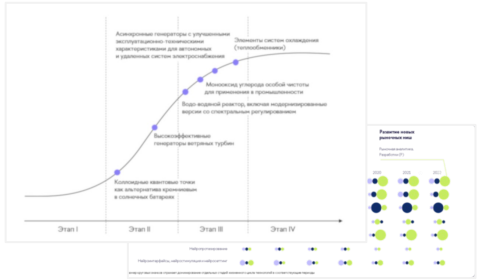
Диаграмма структурной динамики

Выявление наиболее перспективных технологий



Тренд-карта

Анализ уровня готовности технологий



Кривая технологической готовности

Выявление рисков завышенных ожиданий

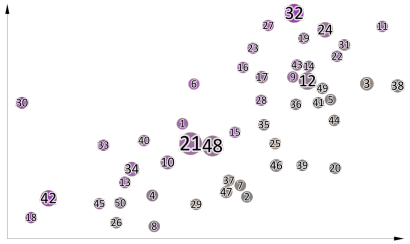
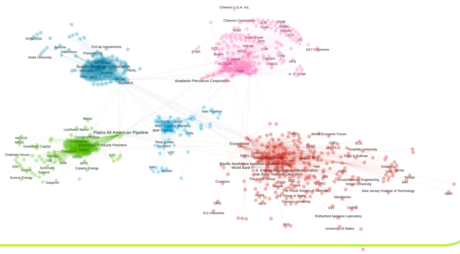


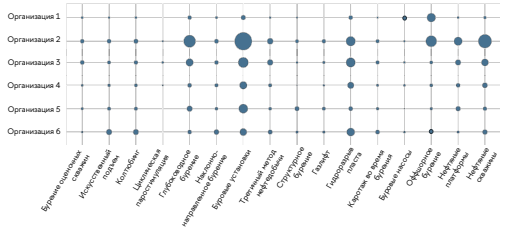
Диаграмма сопоставления актуальности разработок

Выявление сетей центров компетенций



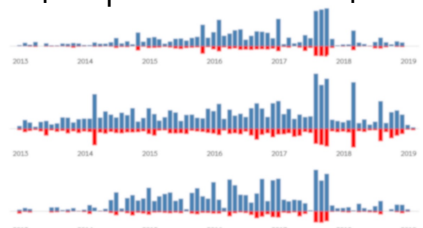
Граф связей

Определение специализации центров компетенций



Матрица взаимосвязей

Репутационный анализ центров компетенций

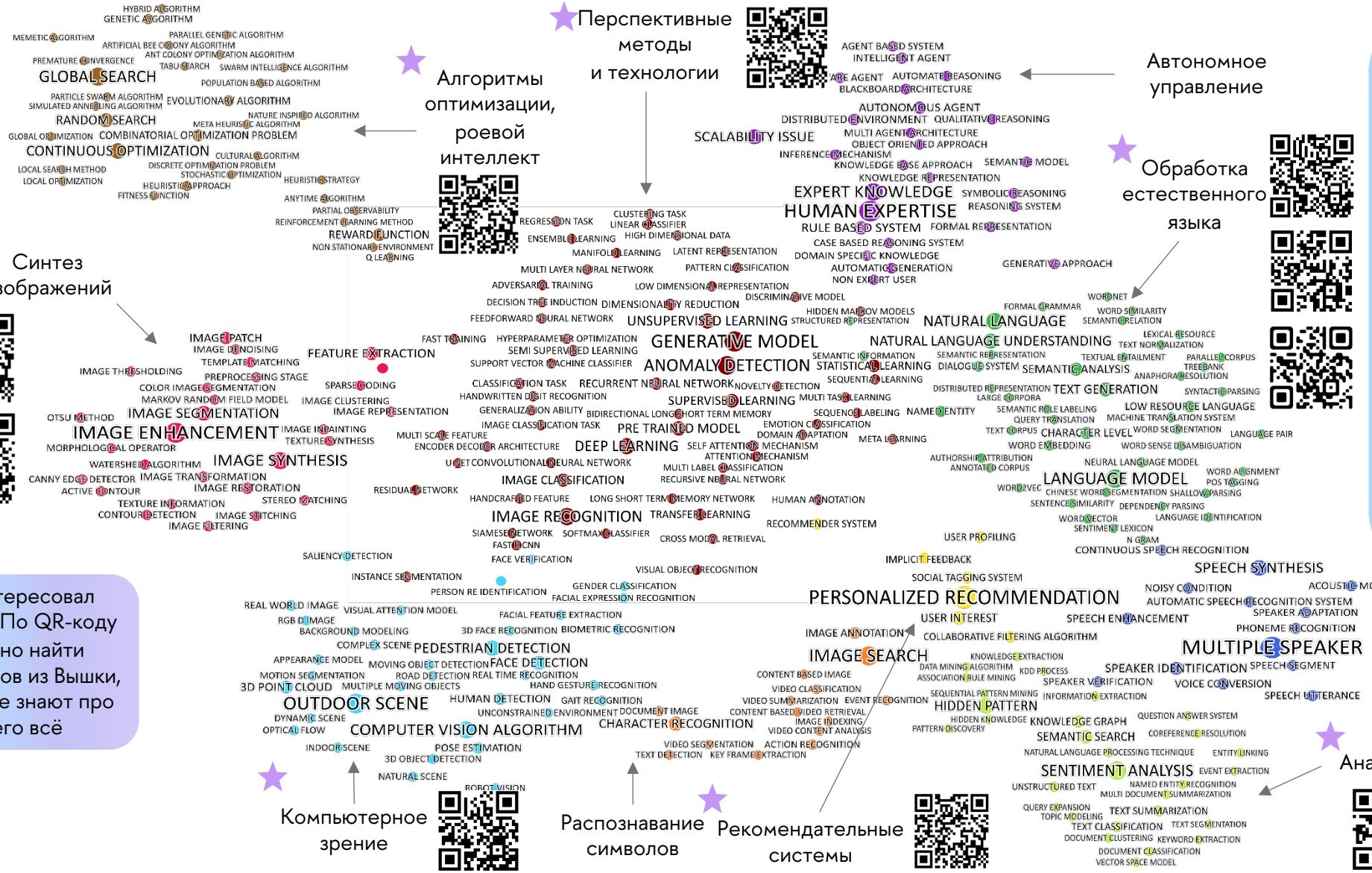


Диаграммы репутационного анализа

Fall into ML
'24

Картирование тематического ландшафта

Кейс: выявление направлений развития ИИ



- **Ключевые технологические направления** связаны с анализом визуальных (3/10 направлений), текстовых данных (4/10) и сквозными методами, составляющими основу для развития ИИ (3/10)
- Большинство выделившихся тематик связаны с методами ИИ. Отдельно можно выделить **Рекомендательные системы**
- Направление «**Алгоритмы оптимизации, роевой интеллект**» развивается обособленно от остальных направлений. Это одна из наиболее перспективных тематик ИИ следующего поколения

Заинтересовал тренд? По QR-коду можно найти экспертов из Вышки, которые знают про него всё

Fall into ML ’24

Выявление и оценка технологических трендов

Intelligent Foresight Analytics

Кейс: выявление наиболее перспективных цифровых технологий

Задача: Отбор наиболее перспективных трендов и технологических направлений

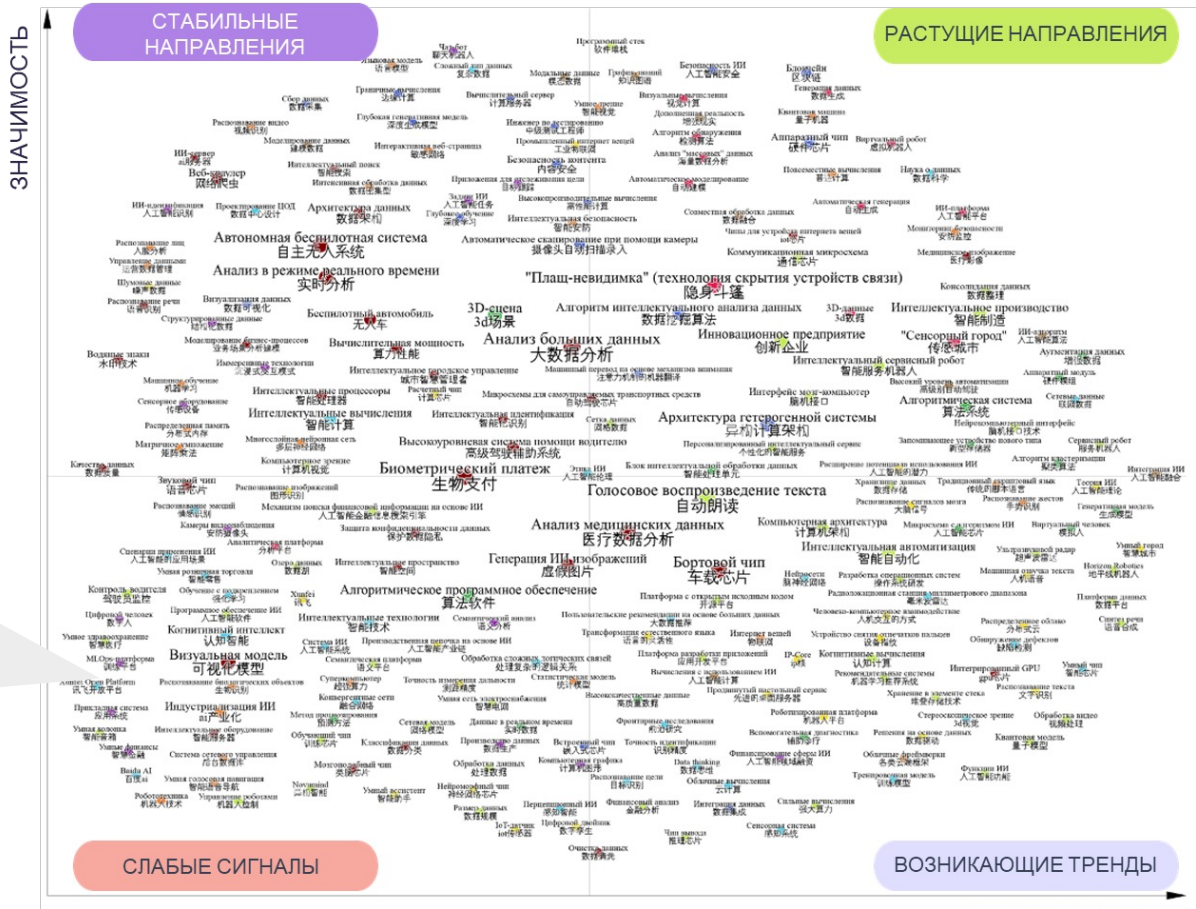
Решение: Извлечение из текстов на китайском языке ключевых тематик по семантическим показателям, оценка их значимости и динамичности на тренд-карте

Эффект: Определение потенциальных «точек роста» для компании на основе объективной оценки развития трендов, выбор стратегии

В зависимости от комбинации значимости и динамичности технологические разработки могут быть классифицированы по **четырем типам тренда**

Приоритетами для дальнейшего анализа являются разработки с **высокой значимостью и динамичностью**

Тренд-карта

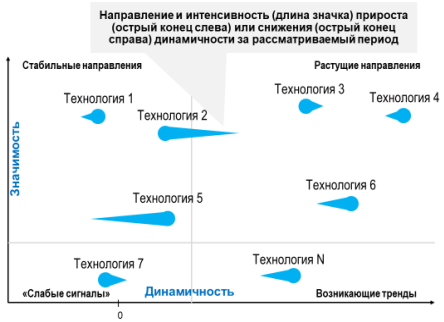


Другие варианты визуализации

Ранг	Технологии	Индекс значимости	Уровень динамичности	Сроки массового внедрения
1	Интернет медицинских вещей (IoMT)	1.00		3-5 лет
2	Туманные вычисления и облачный Интернет вещей	0.97		1-2 года
3	Мобильный Интернет вещей	0.81		1-2 года
4	Искусственный интеллект вещей (AIoT)	0.70		3-5 лет
5	Интернет вещей для умного города / дома	0.58		1-2 года
6	Интернет робототехнических вещей (IoRT)	0.23		3-5 лет
7	Спутниковый Интернет вещей	0.21		4-6 лет
8	Носимый Интернет вещей	0.16		3-5 лет
9	Интеграция Интернета вещей и периферийных устройств	0.12		1-2 года
10	Интернет вещей на транспорте	0.09		1-2 года

Легенда: Стабильные · Растущие · Быстрорастущие

Дополнительные возможности: картирование направления динамики трендов



Углубленный анализ жизненного цикла трендов

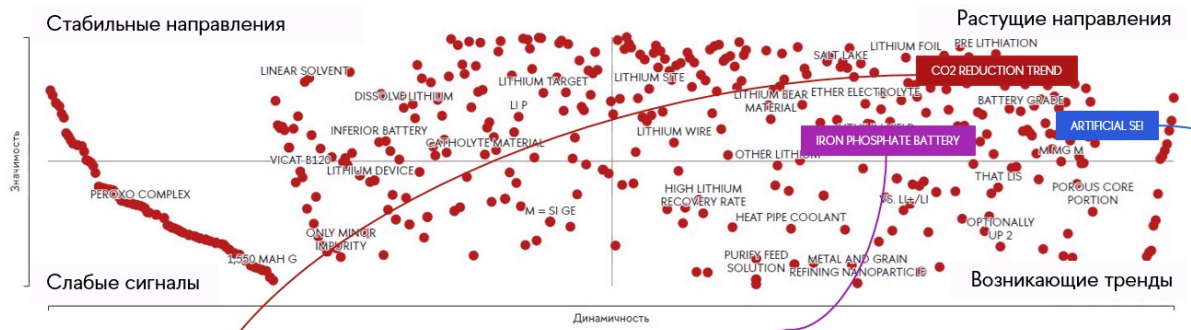
Кейс: формирование динамических траекторий трендов (на примере литий-ионных аккумуляторов)

Задача: Оценка и прогнозирование жизненного цикла технологий и продуктов на основе их упоминания в текстах

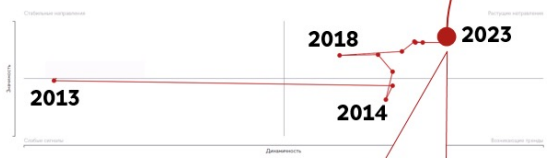
Решение: Ретроспективный анализ значимости и динамичности тематик по годам, формирование гипотез по дальнейшим траекториям

Эффект: Оценка готовности технологий и их ожидаемого потенциала

Тренд-карты с траекториями трендов

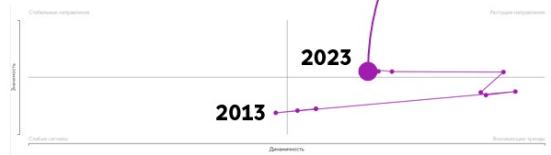


CO2 reduction trend
Сокращение выбросов углерода



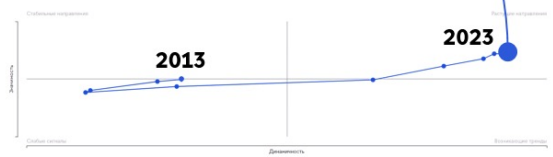
Источник данных: патенты

Iron phosphate battery
Литий-железо-фосфатный аккумулятор



Источник данных: рыночная аналитика и профессиональные СМИ

Artificial solid electrolyte interphase (SEI)
Искусственная межфазная граница твердого электролита



Источник данных: научные публикации

Тренд характеризуется высокой значимостью, динамичным **движением в лидеры роста** за счет расширения приложения и **роста зрелости технологии**

Fall into ML
'24

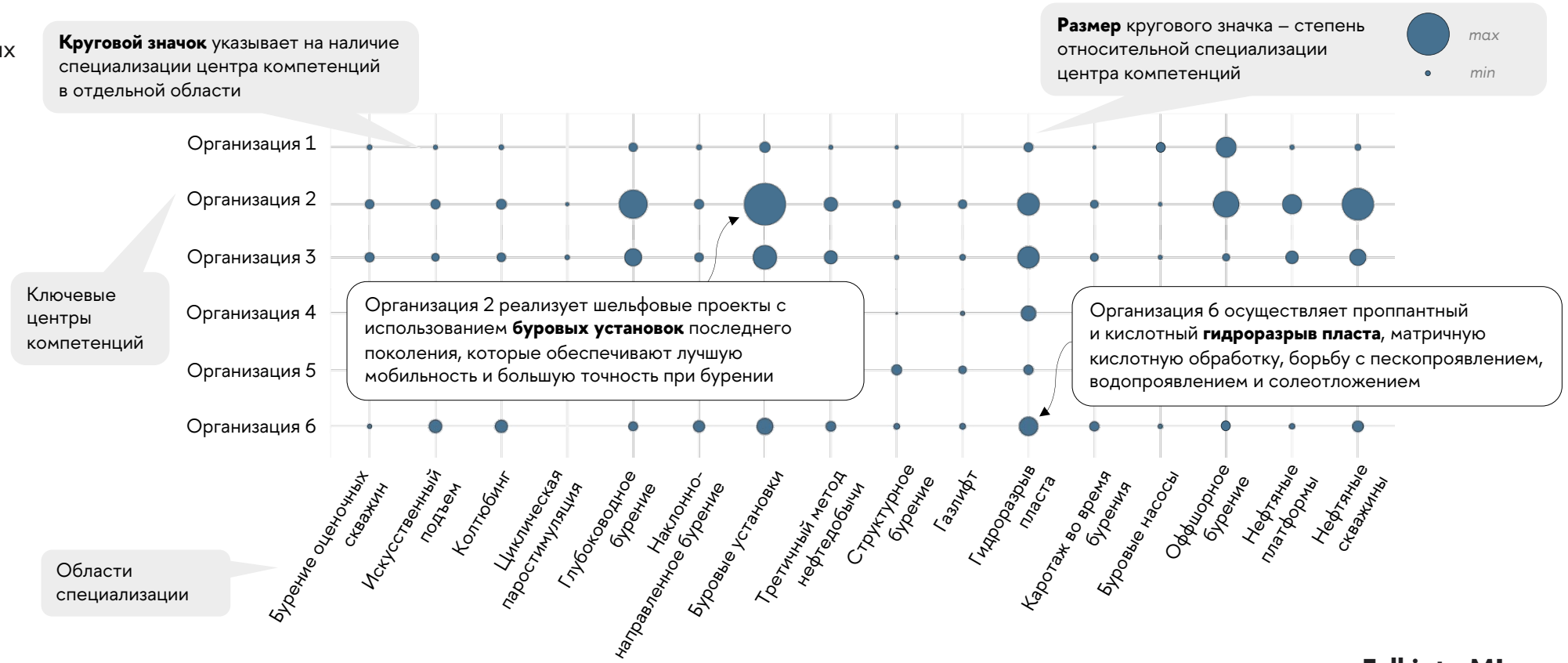
Определение специализации центров компетенций

Кейс: определение специализации центров компетенций в сфере нефтедобычи

Задача: Выявление высококонкурентных рыночных сегментов и свободных рыночных ниш

Решение: Диаграмма специализации организаций

Эффект: Бенчмаркинг конкурентов, выявление угроз и возможностей



Тестирование и выбор LLM для iFORA-ассистента

Использование RAG и LLM

Техническое тестирование

8 открытых LLM, в т.ч. GigaChat

Отбор LLM с лучшими метриками качества

Отобраны 4 открытых LLM для обработки текста

1. Research_qwen_2_7b
2. Llama3.1:8b
3. Research_llama3_8
4. Mistral

+ GigaChat

Отобраны 2 открытые LLM для обработки визуализаций

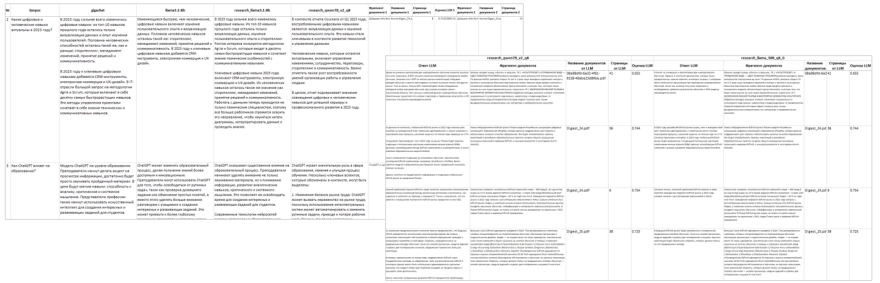
1. Qwen VL
2. Llava

	bleu	ref_len	resp_len	rouge_1	rouge_2	rouge_l	bert_score_f1	bert_score_r	bert_score_p	sem_sim
research_qwen_2_7B	0,02	288	188	0,15	0,02	0,14	0,66	0,66	0,66	0,86
llama3.1:8b	0,02	288	194	0,15	0,02	0,14	0,66	0,66	0,67	0,84
research_llama3_8	0,02	288	286	0,14	0,02	0,13	0,66	0,65	0,68	0,84
mistral	0,02	288	193	0,14	0,02	0,13	0,66	0,66	0,67	0,83
saiga_llama3	0,02	288	182	0,15	0,02	0,14	0,66	0,66	0,67	0,85
t-lite	0,02	288	385	0,14	0,01	0,13	0,67	0,68	0,66	0,84
qwen2:7b	0,02	288	177	0,14	0,02	0,13	0,66	0,66	0,66	0,84
research_llama_38B	0,02	288	166	0,15	0,02	0,14	0,66	0,65	0,67	0,83
llama3:8b	0,02	288	165	0,15	0,02	0,14	0,66	0,65	0,67	0,82
gigachat	0,03	288	130	0,13	0,02	0,12	0,66	0,65	0,67	0,81
qwen:14b	0,01	288	74	0,09	0,01	0,09	0,64	0,62	0,67	0,73

Экспертное тестирование LLM для обработки текстов

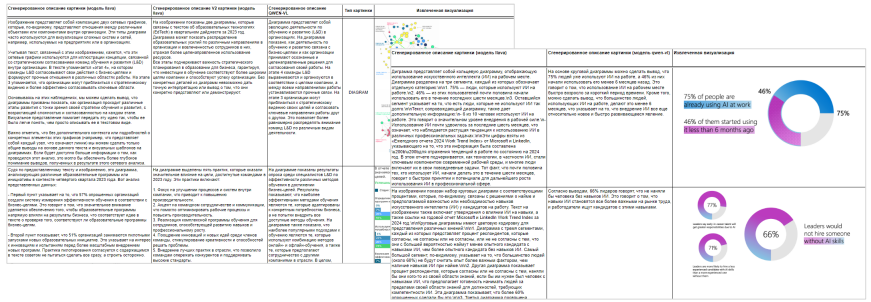
Выбрана 1 открытая LLM - Research_qwen_2_7B

+ GigaChat



Экспертное тестирование LLM для обработки визуализаций

Выбрана 1 открытая LLM - Qwen VL



Модели взаимодействия с бизнес-заказчиком

1 **Бутиковая аналитика**

2 **Подписка на услуги**

3 **Коробочное решение**

4 **ИТ-платформа**

- **Кастомизация** под потребности заказчика
- Доступ через **оператора** и аналитика
- Сравнительно **высокая цена**

Примеры запросов:

1) Определить в какие технологии инвестировать в ближайшие 3 года

2) Выявить факторы влияющие на счастье и психофизическое состояние жителей мегаполисов

Модели взаимодействия с бизнес-заказчиком

1 **Бутиковая аналитика**

2 **Подписка на услуги**

3 **Коробочное решение**

4 **ИТ-платформа**

- Продажа **типовых услуг**
- Заказчик **выбирает из пула** решаемых задач
- **Возможно дополнение** углубленной аналитикой
- **Ценник ниже**, чем в кастомизированной аналитике

Пример запроса:

В течение квартала сформировать комплект аналитики по 10 технологиям:

- Картирование ландшафта
- Выявление и отбор
- Анализ цифрового следа
- Определение решений-аналогов

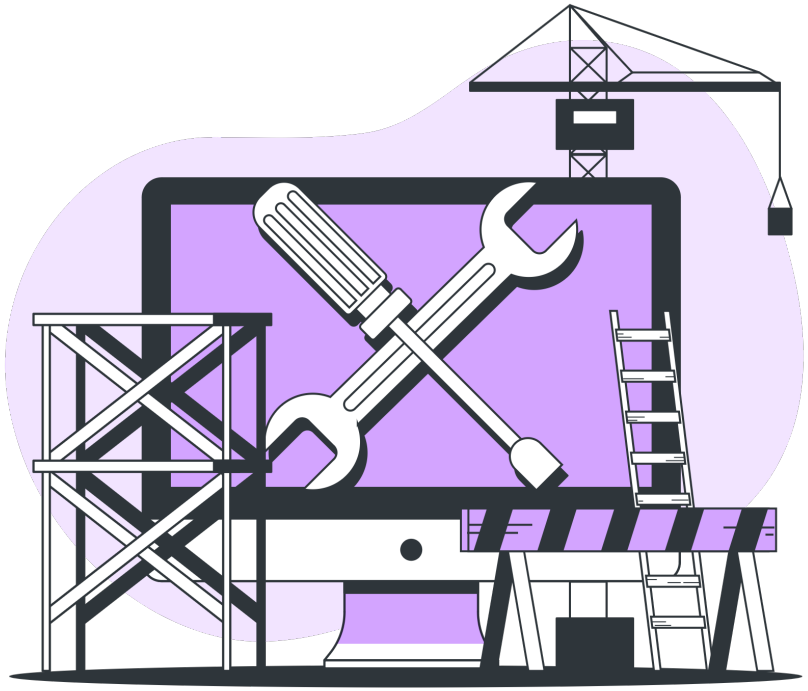
Модели взаимодействия с бизнес-заказчиком

1 **Бутиковая аналитика**

2 **Подписка на услуги**

3 **Коробочное решение**

4 **ИТ-платформа**



- **Кастомизированное ИТ-решение**
- Возможность использовать **локально**
- Доступ **напрямую**
- Необходимо **обучение пользователей**
- Может требовать **собственную инфраструктуру**

Пример запроса:

Разработать рекомендательный сервис для автоматизации аналитических процессов научно-технической деятельности в отрасли X

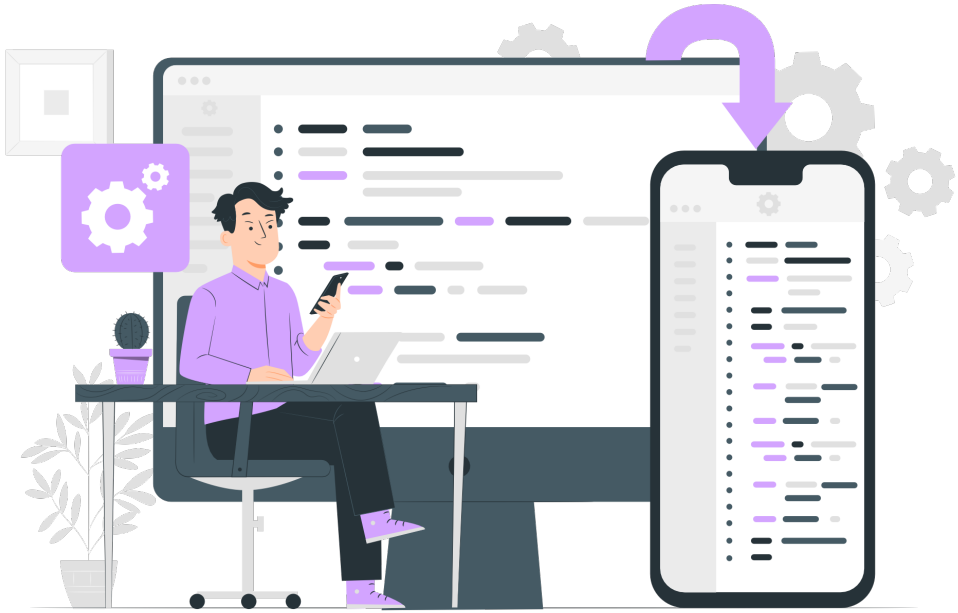
Модели взаимодействия с бизнес-заказчиком

1 **Бутиковая аналитика**

2 **Подписка на услуги**

3 **Коробочное решение**

4 **ИТ-платформа**



- **Предоставляется доступ** к платформе
- Возможности для **глобального бизнеса**
- **Гибкие** тарифные планы
- Необходима солидная **инфраструктура**
- **Расширение пула** возможных заказчиков

Пример запроса:

Обеспечить доступ всех подразделений к оперативной аналитике о развитии науки и технологии, трендах и рынках

Факапы в проектной деятельности...

Инициирование и планирование проекта

- Неконкретные цели
- Отсутствие количественных критериев успеха
- Будущие исполнители не участвуют в планировании
- Методология не прописана
- Неправильная оценка сроков реализации
- Изолированность от багажа других проектов и знаний
- Отсутствие оценки рисков
- Выбор неподходящего менеджера проекта
- Отсутствие мотивации команды
- ...

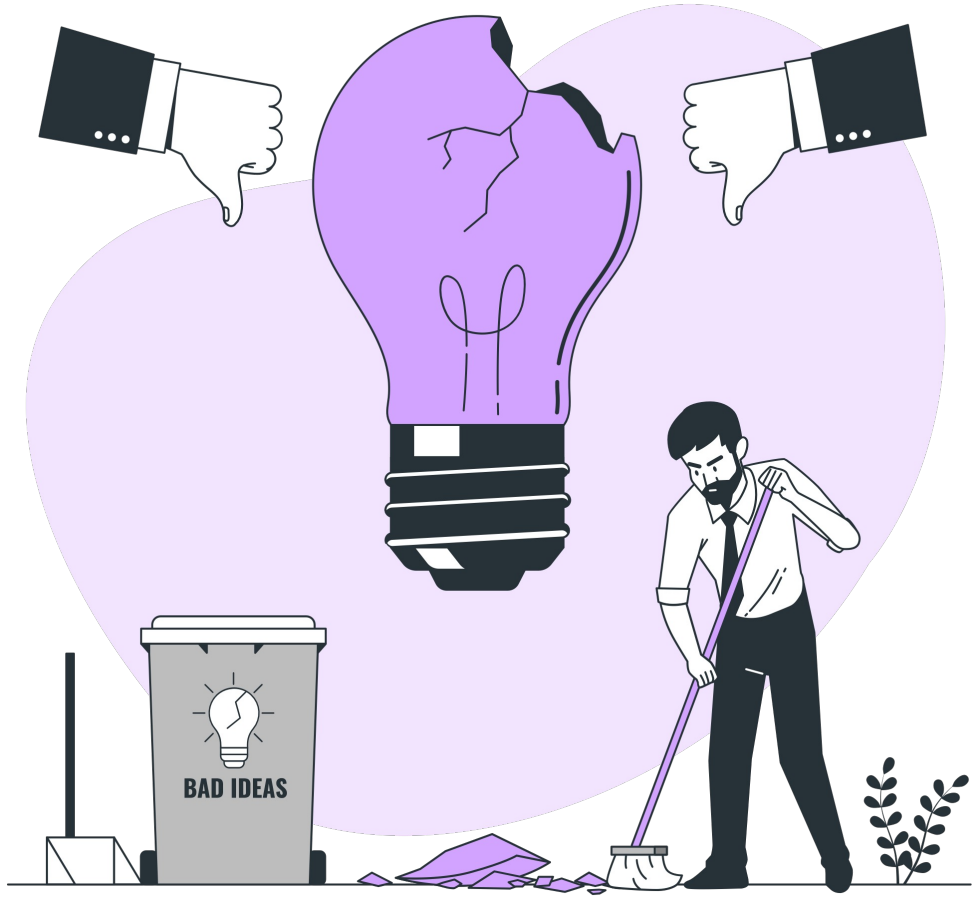
Выполнение проекта и мониторинг

- Отсутствует единый ответственный
- Плохая коммуникация в команде
- Непонимание стейкхолдеров и их целей
- Не отслеживаются изменения в процессе работы над проектом
- Отсутствие механизмов управления рисками
- Административные проблемы (закупки, оборудование, согласования....)
- Крайности в формализации
  - Недостаточный оперативный контроль
  - Микроменеджмент
- ...

Завершение проекта

- Заказчик хотел чего-то другого
- Результаты не используются заказчиком в достаточной мере
- Результаты не масштабируются, следующий проект «как в первый раз»
- Ошибки проекта не документируются и не исправляются в дальнейшем
- Команде не дается обратная связь
- ...

Факапы в проектной деятельности...



Fall into ML
’24

… и как их избежать (ну или с чего начать)

- Заручиться поддержкой спонсора/инвестора/руководителя
- Сформулировать цели проекта и критерии успеха
- Определить всех стейкхолдеров проекта и их реальные интересы
- Начинать планирование от общего видения к более мелким задачам
- Оценить ресурсы и риски
- Сформировать и декомпозировать структуру работы
- Четко определить роли и ответственность за работы в проекте
- Синхронизировать понимание ЧТО, ЗАЧЕМ и КАК делаем для команды
- Выстроить коммуникации в команде
- Практики регулярного менеджмента
- Управлять изменениями
- Получить обратную связь от заказчика
- Отфиксировать «уроки» проекта

ВИШНЕВСКИЙ КОНСТАНТИН ОЛЕГОВИЧ

Директор Центра стратегической аналитики
и больших данных
ИСИЭЗ НИУ ВШЭ, PhD

Руководитель системы интеллектуального
анализа больших данных iFORA

kvishnevsky@hse.ru

Сайт iFORA

iFORA в Telegram

iFORA-экспрессы

Хотите у нас работать
или пройти
стажировку?
Сканируйте QR-код

СБЕР

Multimodal banking data and event sequences

Ivan Kireev
Sber AI Lab

26.10.2024

Ivan Kireev

7 years in Machine Learning

5 years in Deep Learning (Sber AI Lab)

2 years as Head of Deep Learning Center

7 scientific publications

Website
Sber AI Lab

Research Interests:

- event sequences
- representation learning
- Matching
- LLM
- adversarial methods
- dynamic graphs

Deep Learning Center

Event Sequences

Neural Network Potential:

- Large data volume
- Complex data structure

Open Source Datasets

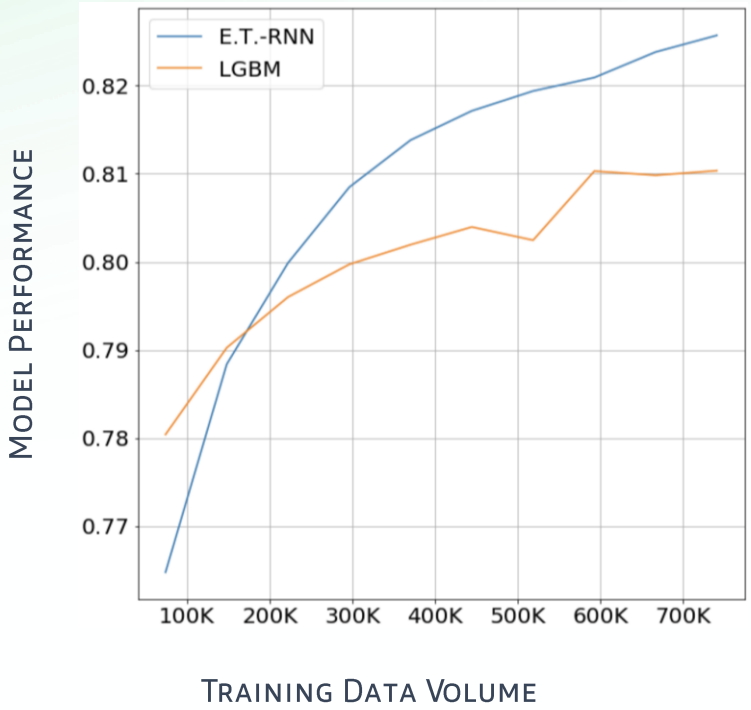
Sequence type	Target	Dataset
Financial history	Client's Gender and Age	SberAgePred
		SberGender
	Churn Prediction	Rosbank
		DataFusion
	Default Indicator	Alphabattle
Retail Receipt History	Uplift	X5RetailHero
Learning App Logs	Exam Score	Bowl2019
URL Visit History	Client's Gender and Age	MtsMLCup
Music Listening Logs	Genre	YandexMLCup

Sequential Data – Example

cl_id int64 2 10.2k	MCC int64 742 9.41k	channel_type string · *classes* 5 values	currency int64 32 986	TRDATETIME string · *lengths* 16 16	amount float64 0.01 18.3M	trx_category string · *classes* 10 values
3	6,011	null	810	09AUG17:20:08:44	2,000	WD_ATM_ROS
3	5,814	null	810	06JUL17:00:00:00	695	POS
3	5,999	null	810	21JUL17:11:20:12	100	POS
3	5,912	null	810	01JUL17:00:00:00	1,966	POS
3	5,411	null	810	27JUL17:00:00:00	360	POS
3	5,977	null	810	12JUN17:00:00:00	1,064	POS
3	6,011	null	810	14JUL17:00:00:00	5,500	WD_ATM_OTHER
3	5,814	null	810	12JUN17:00:00:00	187	POS
4	5,541	null	810	04FEB18:00:00:00	304	POS
4	6,012	null	810	21FEB18:13:03:19	700	C2C_OUT
4	5,631	null	810	18FEB18:00:00:00	373.5	POS
4	5,921	null	810	23MAR18:00:00:00	212.98	POS
4	5,814	null	810	09MAR18:00:00:00	622	POS

Creating a Universal Embedding on a Large Volume of Unlabeled Data

Example: End-to-end training for scoring tasks
ET-RNN outperforms boosting in quality as the training data volume increases



Problem:

Neural networks require large amounts of labeled data for training, which are not always available for specific tasks.

Solution:

- Train a universal model using large volumes of unlabeled data.
- Adapt this model for individual tasks.

Technology:

Embedding as the output of a universal model

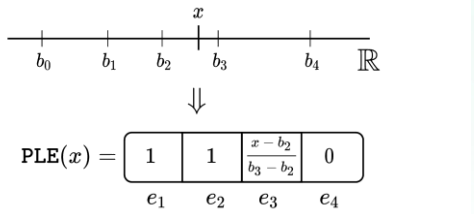
ET-RNN: Applying Deep Learning to Credit Loan Applications [KDD '19]

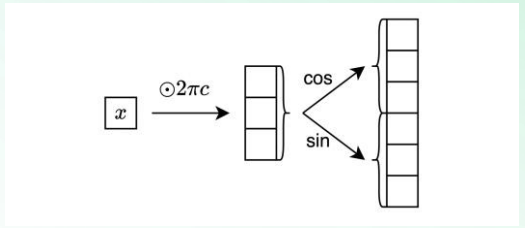
6

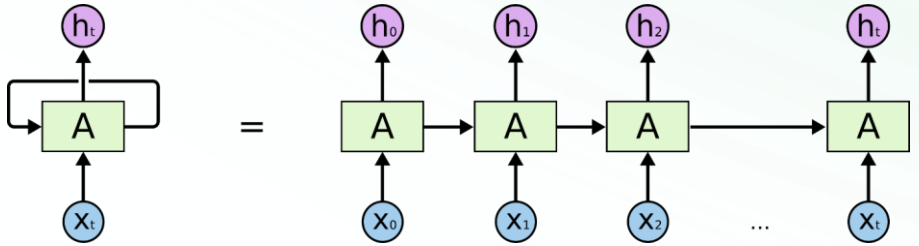
Architectures and Algorithms

Architectures:

- Transaction Encoders
- Sequence Encoders

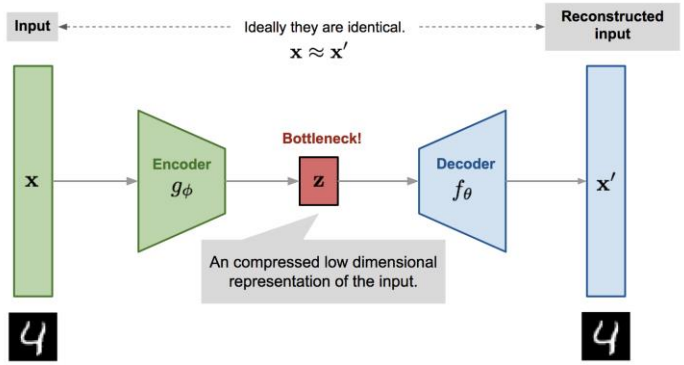




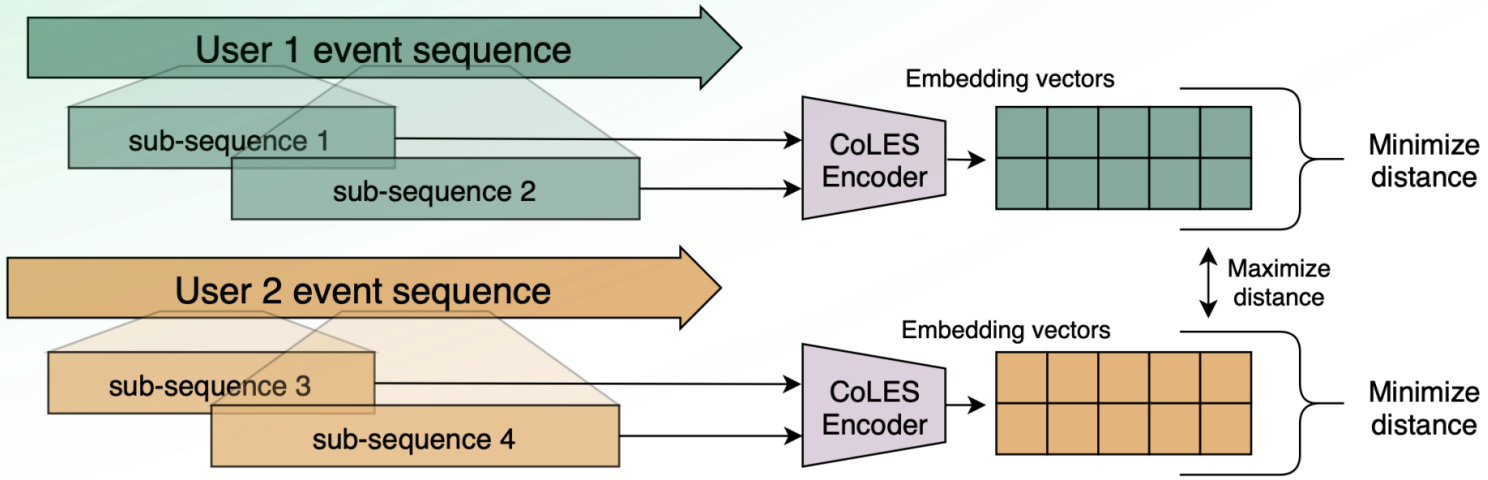


Learning Methods:

- Unsupervised / Self Supervised
- Contrastive



CoLES



CoLES: Contrastive Learning for Event Sequences with Self-Supervision
[SIGMOD'22]

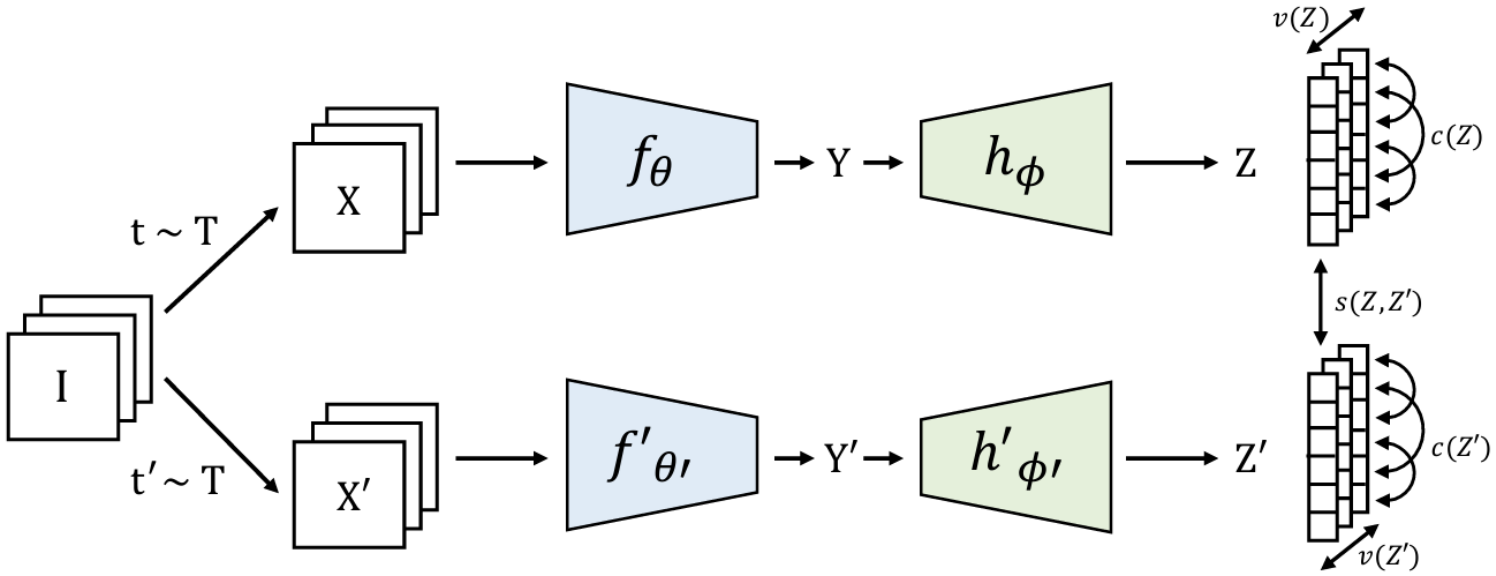
Advantages:

- Representation of the entire object
- Proximity is explicitly defined
- Splits resemble the entire sequence

Disadvantages:

- Negative examples are required
- Object dynamics are not taken into account

VICReg



VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning
https://arxiv.org/abs/2105.04906

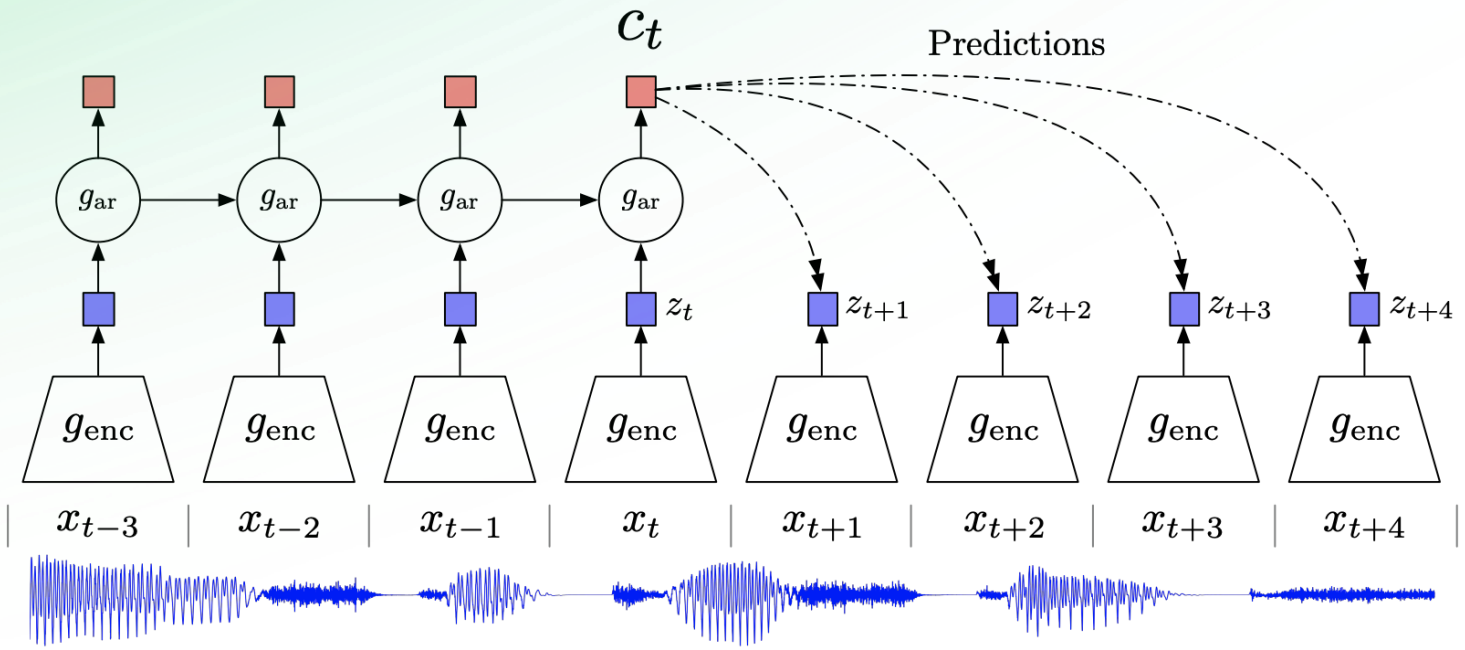
Advantages :

- Representation of the entire object
- Proximity is explicitly defined
- Negative examples are not required

Disadvantages :

- Sensitivity to hyperparameters

Contrastive Predictive Coding



Representation Learning with Contrastive Predictive Coding

https://arxiv.org/abs/1807.03748

Advantages :

- The hidden state contains all information about the object
- Predictive tasks are addressed

Disadvantages :

- Negative examples are required
- Mandatory splits are needed
- Predictive tasks are more complex

Multimodality for Embeddings

The use of additional data (modalities) enhances the quality of customer embeddings

Multimodal embeddings can be applied to the same tasks as traditional embeddings but perform better

Examples of Modalities

- Purchase History
- Financial operations
- Transfers
- Customer Communications
- Website and App Activity
- Receipts

Improvements for Individual Sources

- Raw, noisy data
- Large category dictionaries
- Rare events with limited coverage

New Types of Data

- Geolocation Data
- Graphs
- Text

Multimodal Banking Dataset

The largest open-source multimodal banking dataset

Data from 2 million clients has been collected and anonymized

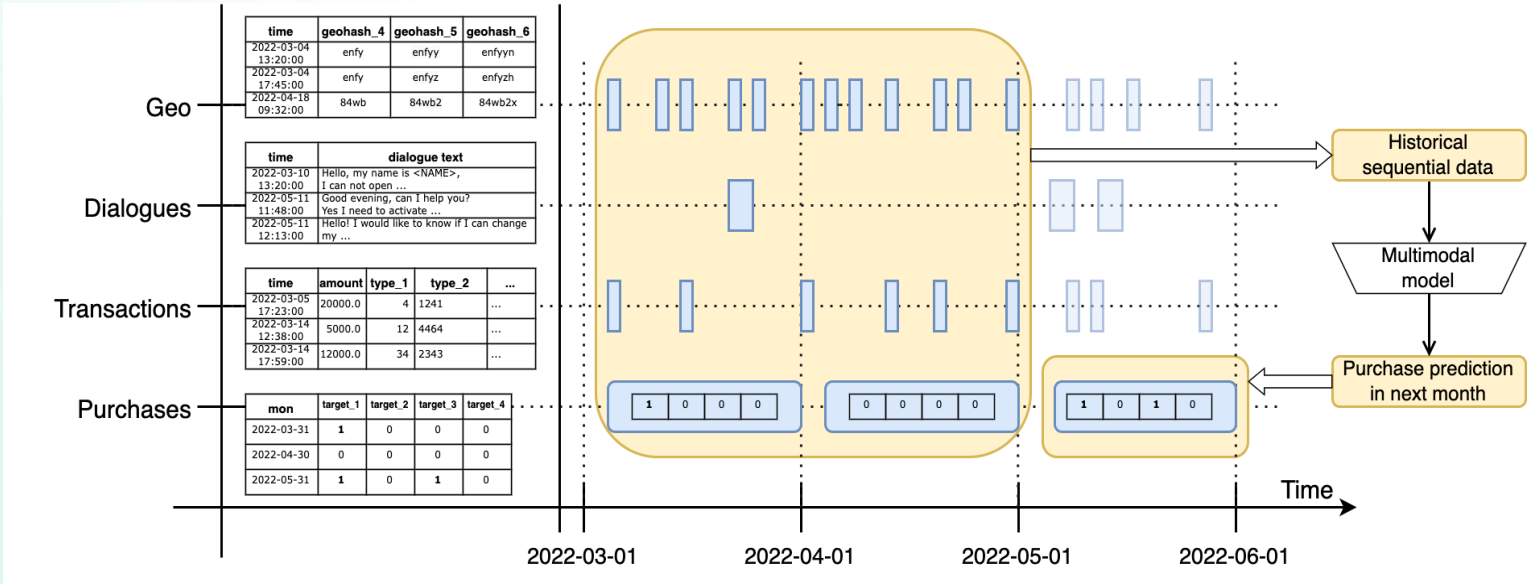
Modalities:

- Transactions
- Dialogues
- Geostream

Task:

Predicting the purchase of 4 products for the next month

Link on Hugging Face:



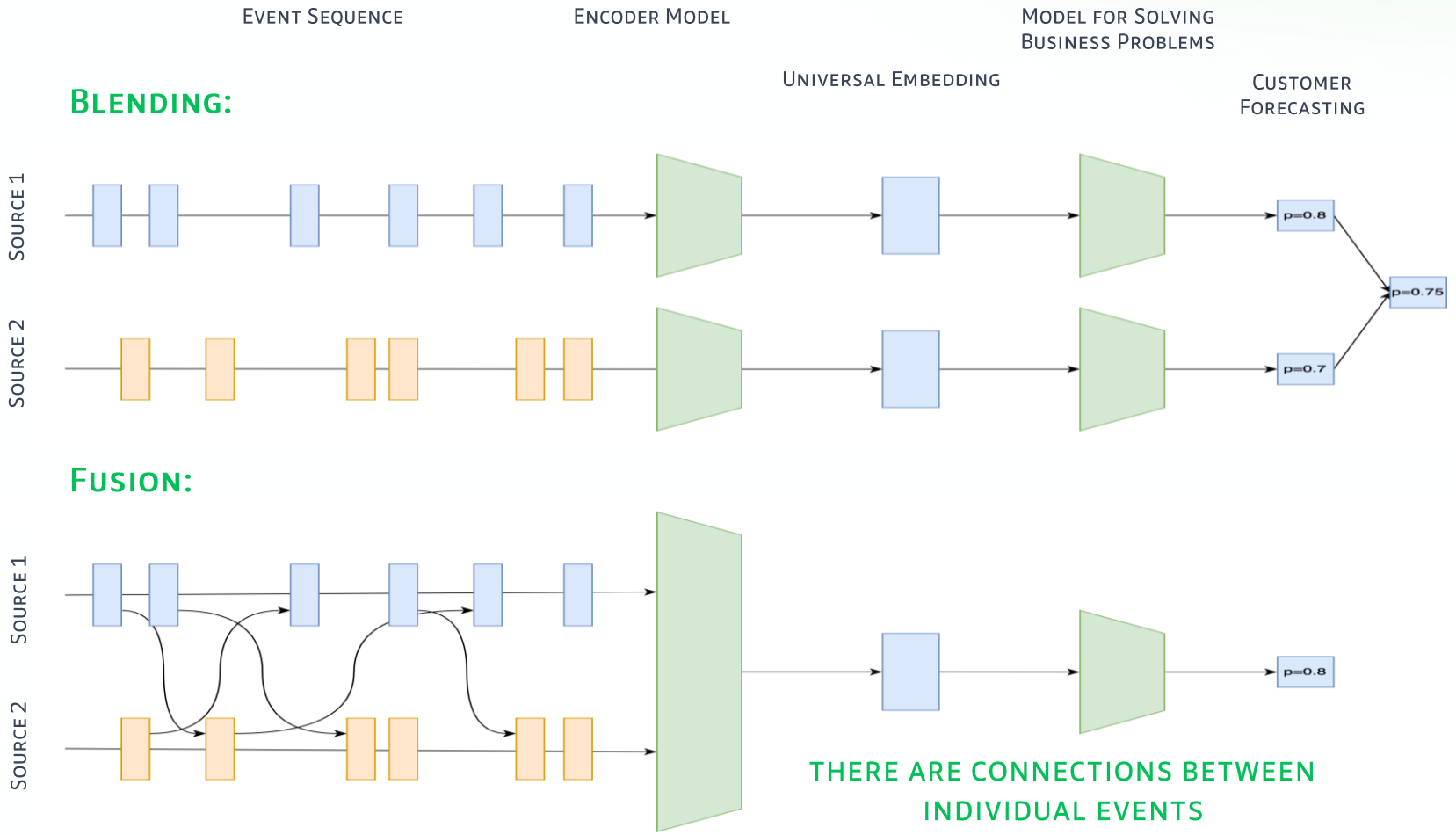
https://arxiv.org/abs/2409.17587

Fusion of Modalities – Accounting for Deep Relationships

Deeper utilization of additional data results in higher quality

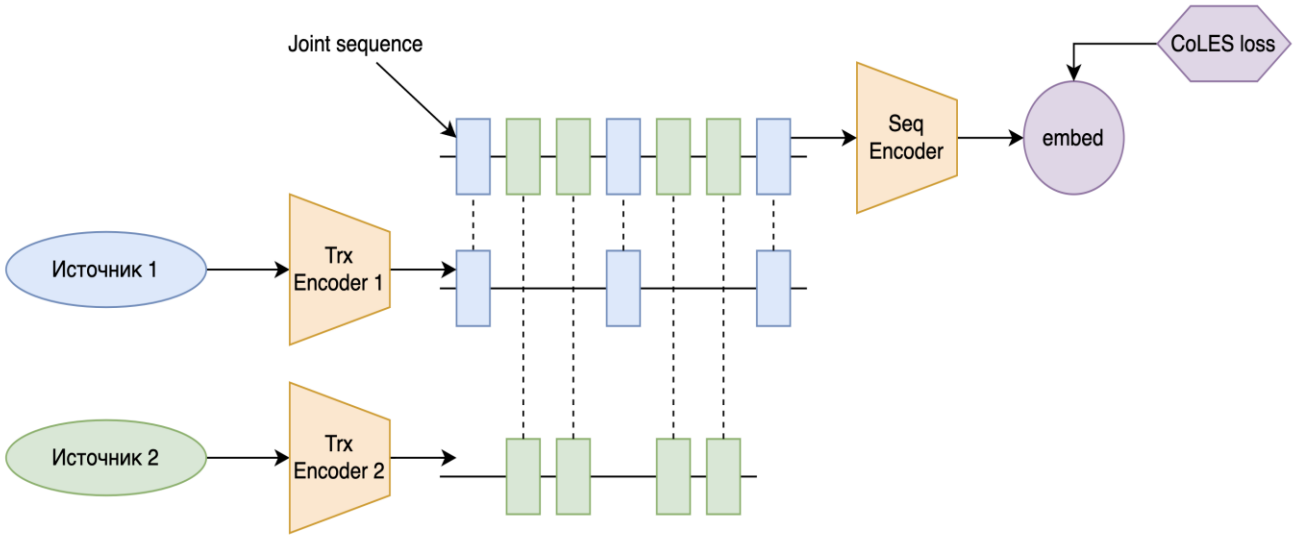
Options for Combining Modalities:

- Without using additional data
- Blending
- Late Fusion
- Early Fusion
- Mid Fusion



Early Fusion

Объединение событий в одну цепочку



Description:

Events from each modality are mixed into a single chain

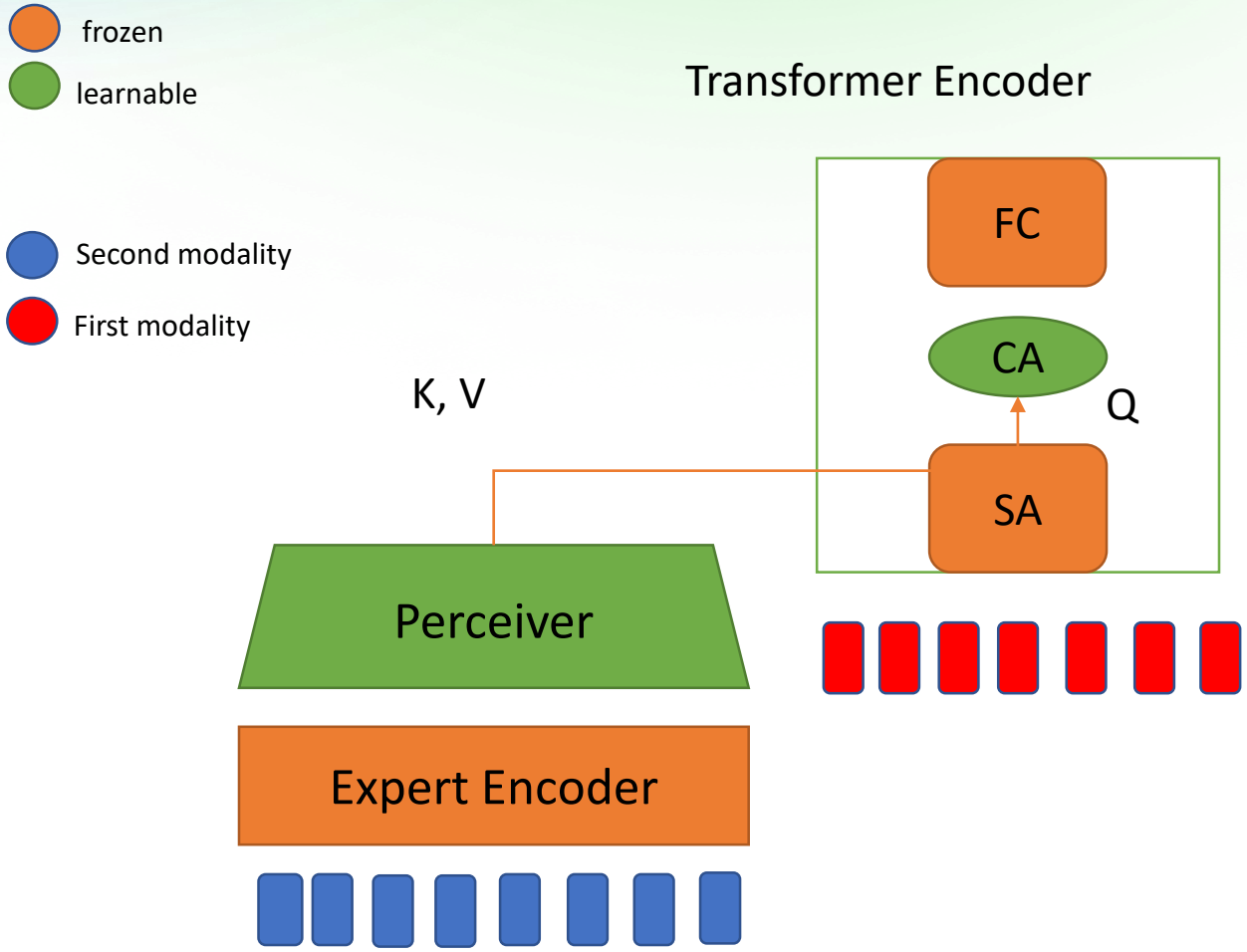
Advantages :

Allows for more detailed information about modalities than Late Fusion

Disadvantages:

Events with low frequency may get lost in the overall flow of events

Early Fusion Flamingo



Flamingo: a Visual Language Model for Few-Shot Learning, NeurIPS 2022

Description:

Incorporating modalities into the Transformer using cross-attention

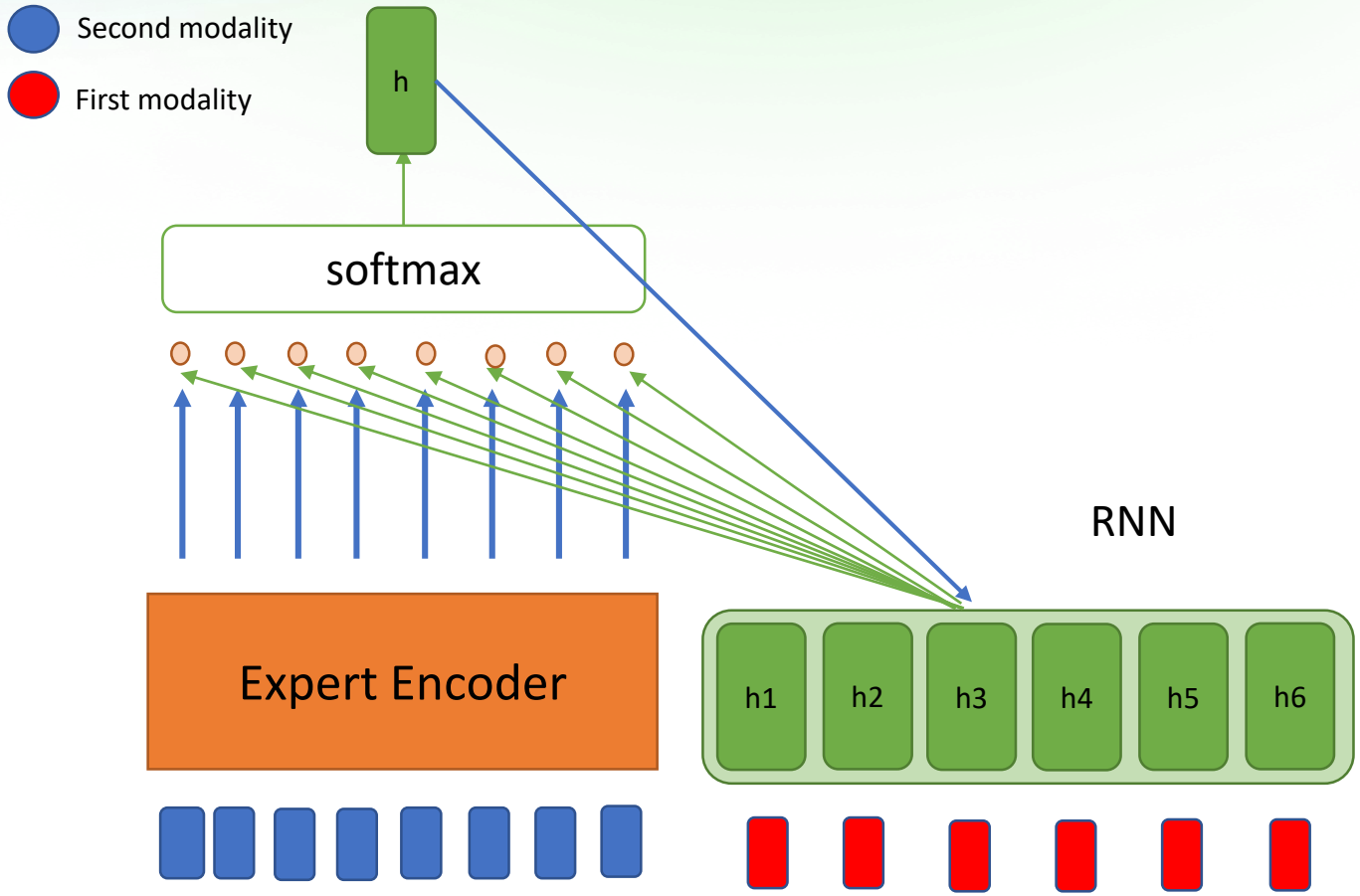
Advantages:

- Pre-training expert encoders for modalities allows for extracting more detailed information about individual modalities
- Trainable cross-attention in the Transformer layers helps address the issue of low-frequency modalities

Disadvantages :

- Scalability issues with a larger number of modalities
- Complex tuning of the Transformer for the event sequence domain.

Early Fusion Attention-Rnn



NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE, ICLR 2015

Description:

Embedding modalities in RNNs using attention

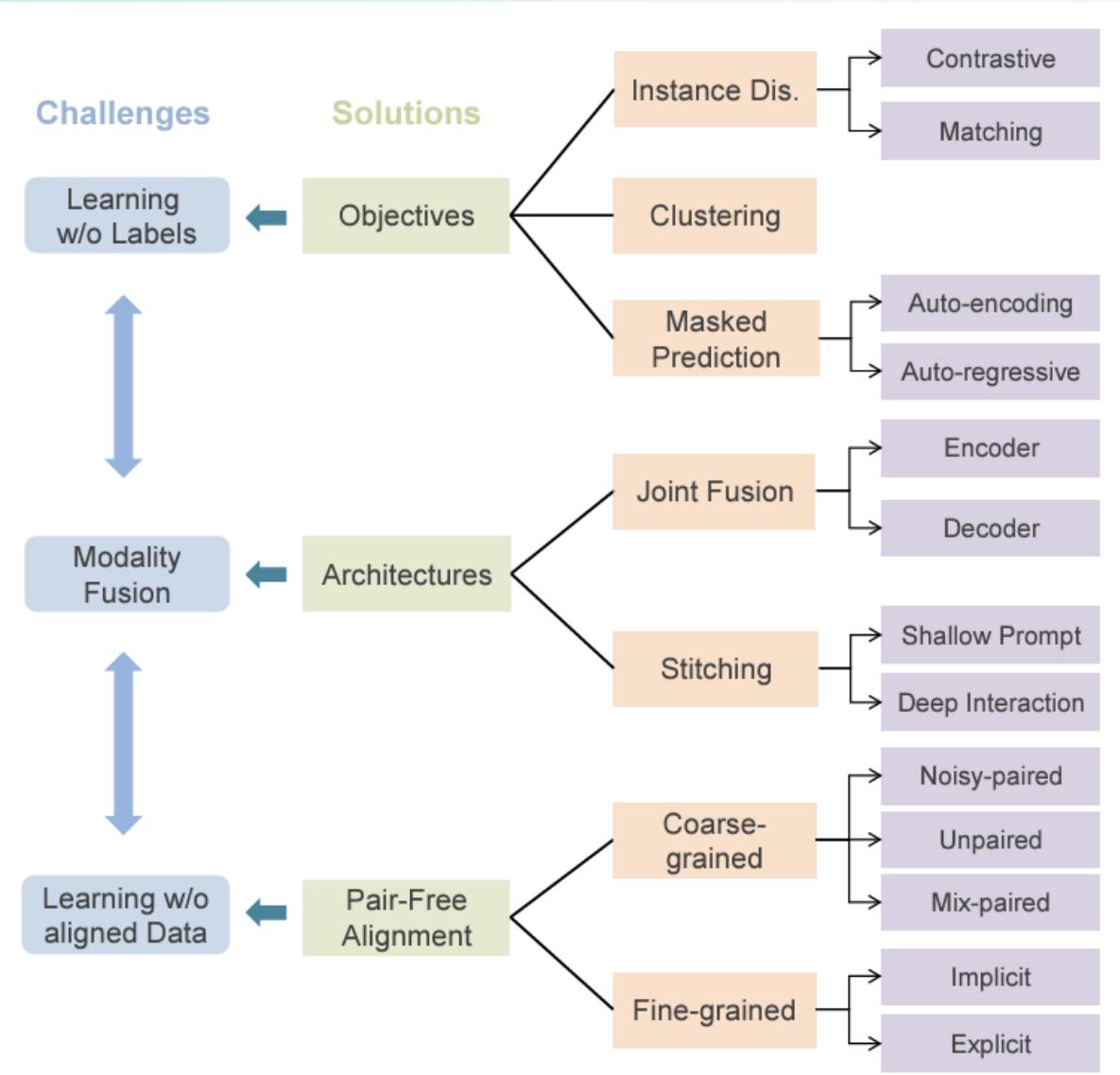
Advantages :

- Fewer parameters to tune
- RNN architecture performs better than Transformer for event sequence domains

Disadvantages :

Scalability issues with a larger number of modalities

Self-Supervised Multimodal Learning



Self-Supervised Multimodal Learning: A Survey

https://arxiv.org/pdf/2304.01008

Thank you for your attention!

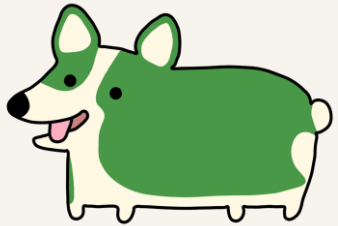
GitHub
sb-ai-lab

Website
Sber AI Lab

@ivkir8

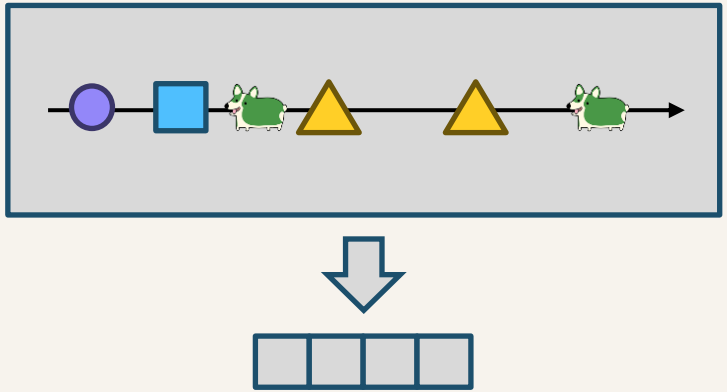
Ivan Kireev

Universal representations for event sequences: financial transactional data and beyond



Alexey Zaytsev

Assistant professor
LARSS laboratory,
Skoltech



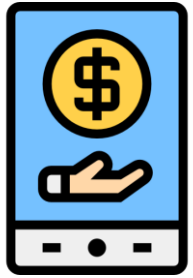
Skoltech

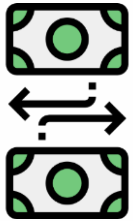
HSE
October 2024

Event sequences: intro

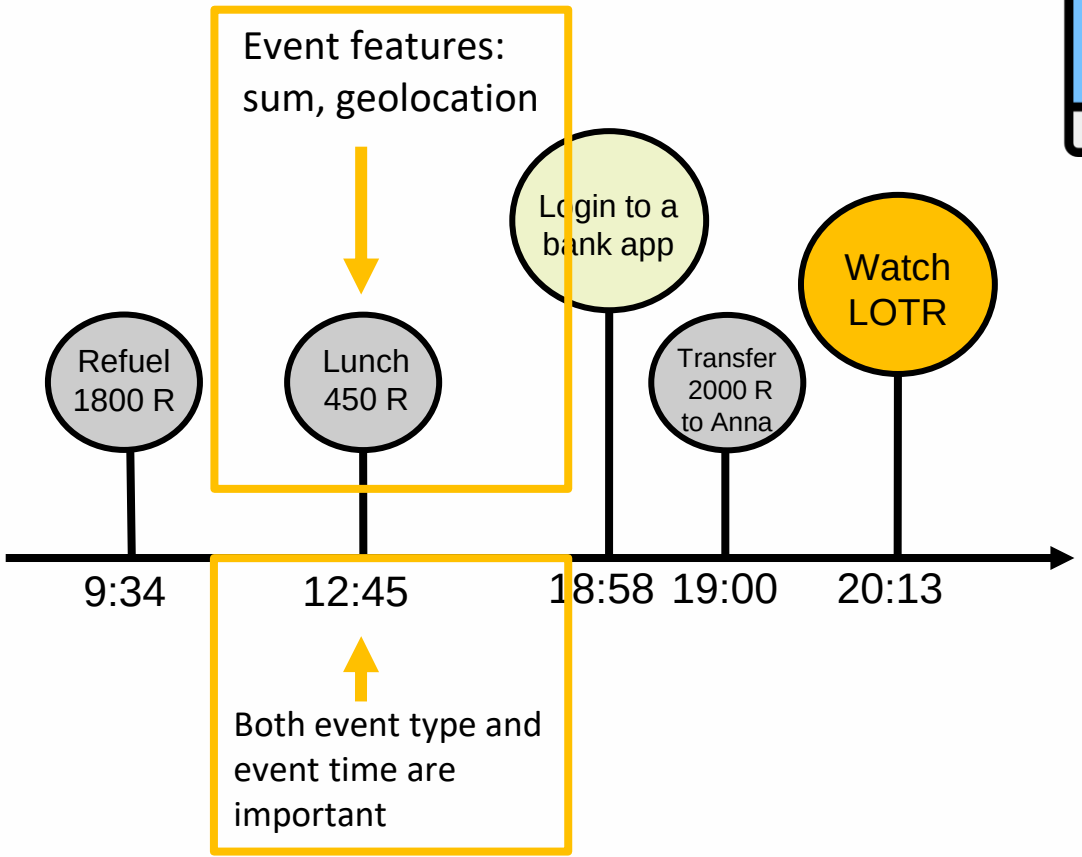
Skoltech

Event sequences data





Alex, 35, man
works in academia



Skoltech

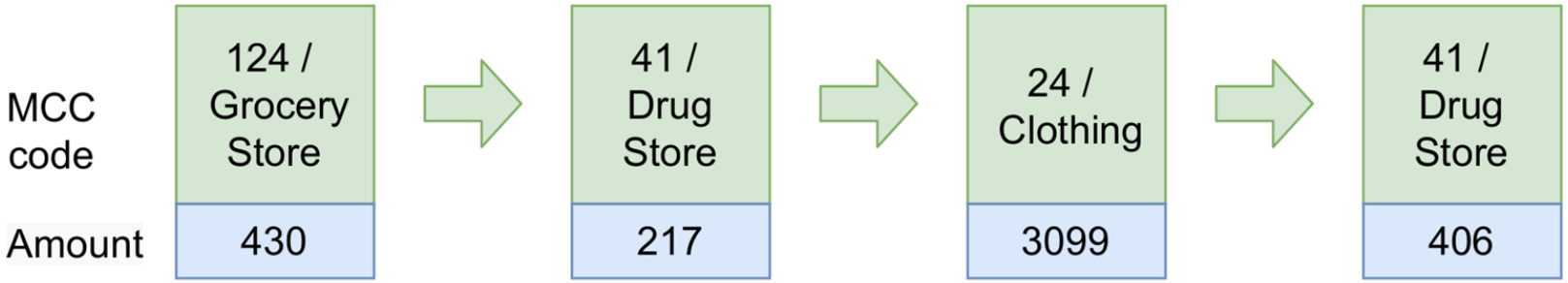
Discrete sequential transactional data

Transaction records data sequence includes:

- MCC (Merchant Category Codes)
- Purchase amount
- Time values
- Transaction location
- …

Data characteristics:

- Heterogeneous features
- Non-regularity of observations
- Varying lengths of sequences



Skoltech

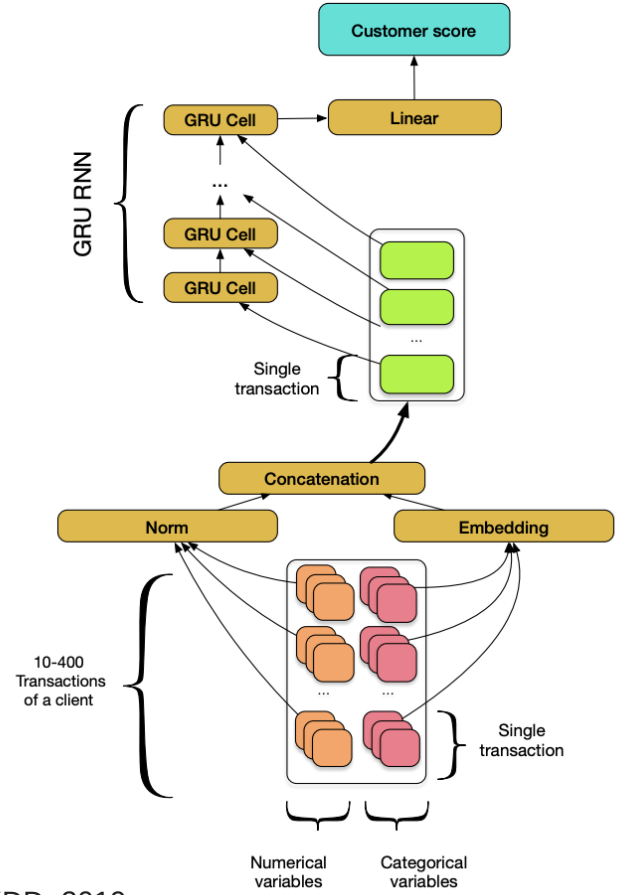
One way: Supervised approach

Library pytorch-lifestream

Recurrent (or Transformer) Neural Network with self-supervised contrastive learning

	ROC AUC	N Features
Logistic regression	0.78	~ 400
LGBM	0.81	~ 7000
E.T.-RNN	0.83	12

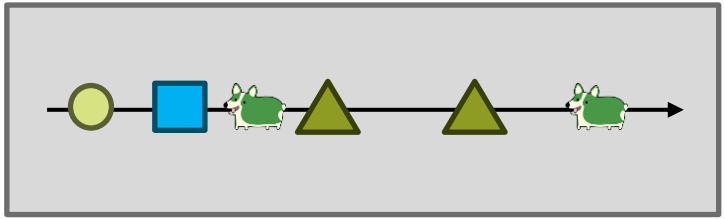
- Requires labeled data!



D. Babaev et al. E.T.-RNN: Applying Deep Learning to Credit Loan Applications. KDD. 2019

Skoltech

Our way: self-supervised approach



Encoder

Trained without labels

Lightweight heads

Trained using small number of labels

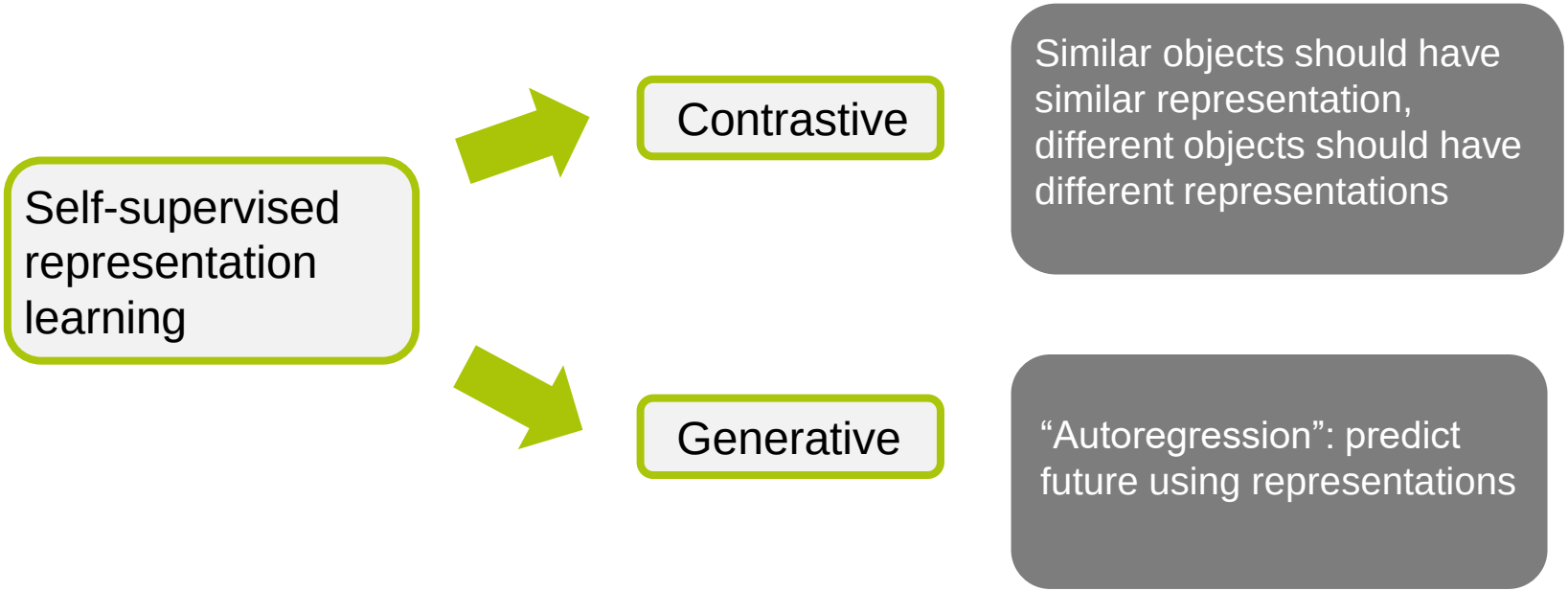
Skoltech

D. Babaev et al. E.T.-RNN: Applying Deep Learning to Credit Loan Applications. KDD. 2019

Types of self-supervised learning

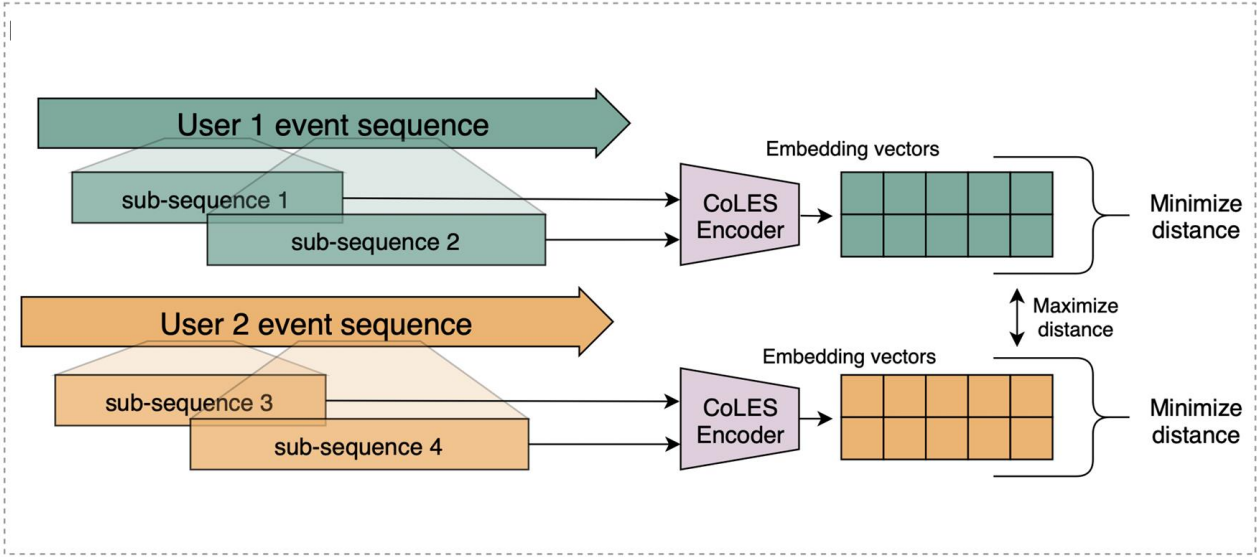
Skoltech

Two main ideas of self-supervised learning: generative and contrastive



Liu, Xiao, et al. "Self-supervised learning: Generative or contrastive." *IEEE Transactions on Knowledge and Data Engineering* (2021).

CoLES contrastive learning



Skoltech

Babaev, Dmitrii, et al. "CoLES: Contrastive learning for event sequences with self-supervision." *SIGMOD.* 2022.

Contrastive learning for sequential data

Weak augmentation: jitter-and-scale strategy
Strong augmentation: permutation-and-jitter strategy

$$\mathcal{L} = \lambda_1 \cdot \boxed{(\mathcal{L}_{TC}^{s} + \mathcal{L}_{TC}^{w})} + \lambda_2 \cdot \boxed{\mathcal{L}_{CC}}$$

Predict future representation from the current context

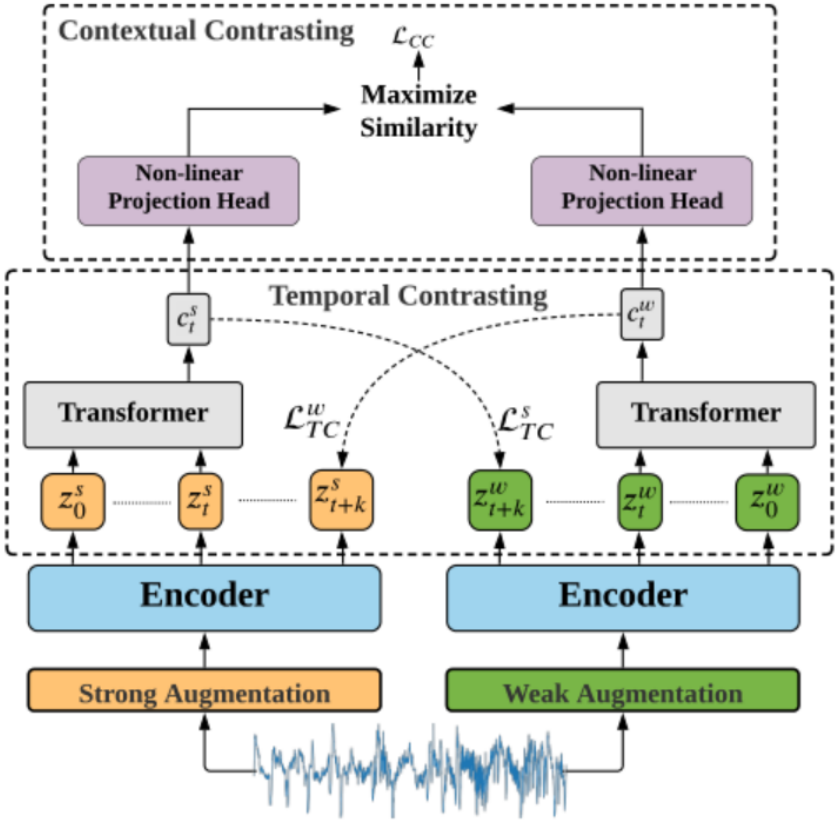
Compare contexts

Context:

$$\boldsymbol{c}_t = f_{AR}(Z_{\leq t}),$$

Representation:

$$\boldsymbol{z}_t = f_{enc}(\boldsymbol{x}_t)$$



Skoltech

Eldele, Emadeldeen, et al. Time-series representation learning via temporal and contextual contrasting. *IJCAI. 2021*.

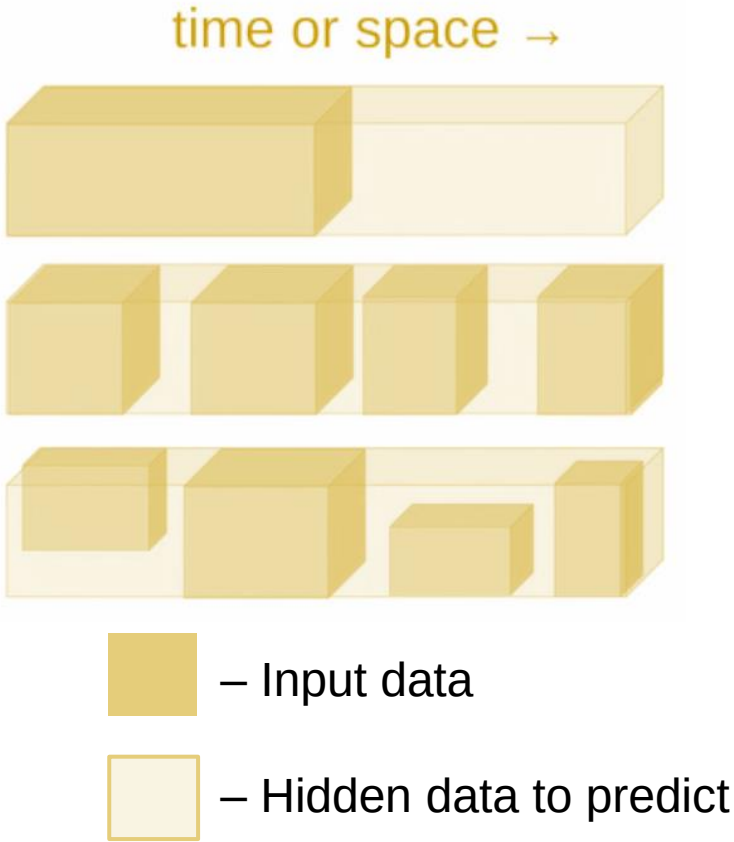
Generative models: masking

Training steps:
1. Hide some part of the data
2. Try to recover it via representation learning

A. Predict the future from the past

B. Predict the invisible from the visible

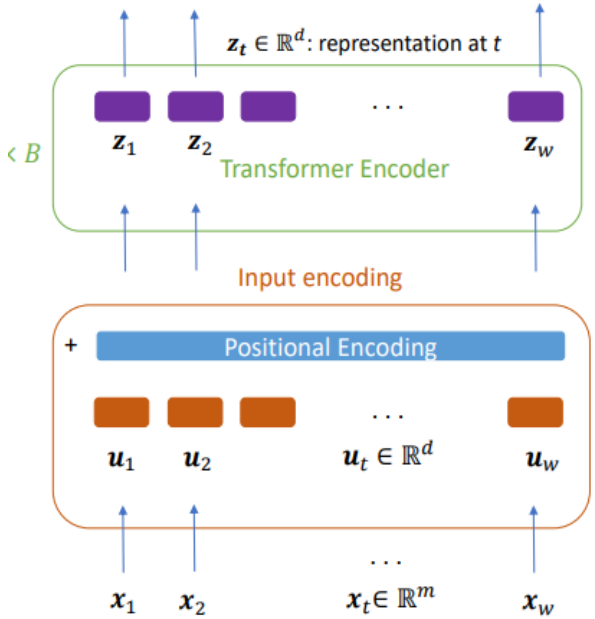
C. Predict occluded, masked or corrupted part



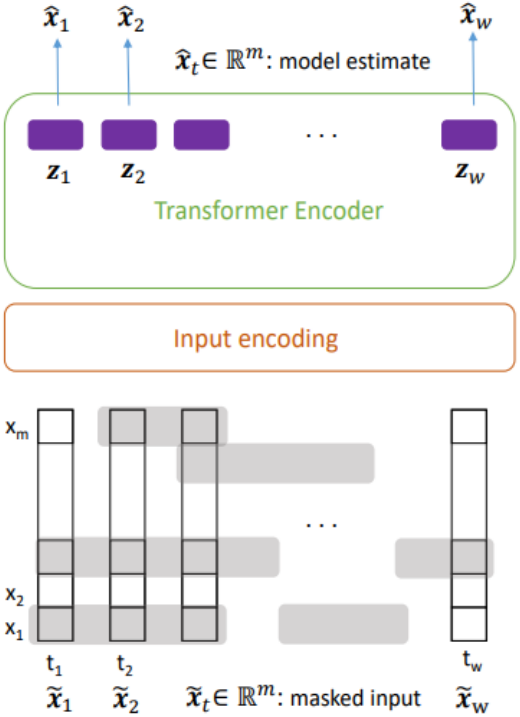
Skoltech

Yoshua Bengio, Yann LeCun Reflections (Self-Supervised Learning). ICLR 2020

Time-series unsupervised representations



Time series encoding via Transformers



Masking for model training

Zerveas, George, et al. A transformer-based framework for multivariate time series representation learning. KDD. 2021.

Skoltech

Desired properties of embeddings

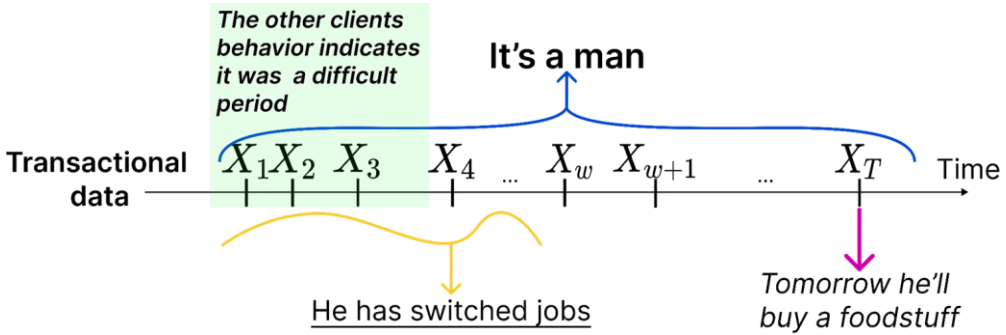
Skoltech

Properties of event sequence embeddings

Goal: to obtain a good encoder for transactional data

Three main properties of local embedding for transactional data:

1. **Global property** - describe a client in general;
2. Local property – describe a client's state at a particular moment in time;
3. Dynamic property - the embeddings should change with time, reflecting the changes in the client's behavior.



Three main properties of embedding and corresponding business information: global (blue), local (pink) and dynamic (orange)

Bazarova, Alexandra, et al. "Universal representations for financial transactional data: embracing local, global, and external contexts." *arXiv preprint arXiv:2404.02047* (2024).

Skoltech

Global and local quality of the models

- *Global validation* – solve a downstream task via a boosting model, get ROC AUC;

- *Local validation* – two approaches:
  a. predict the next event type (MCC) via MLP, get ROC AUC instead of likelihood**;**
  b. predict a local downstream target (churn/default state at the moment) via MLP, estimate ROC AUC**.**

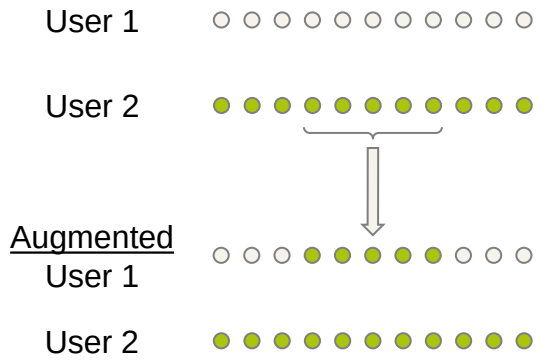
Skoltech

Bazarova, Alexandra, et al. "Universal representations for financial transactional data: embracing local, global, and external contexts." *arXiv preprint arXiv:2404.02047* (2024).

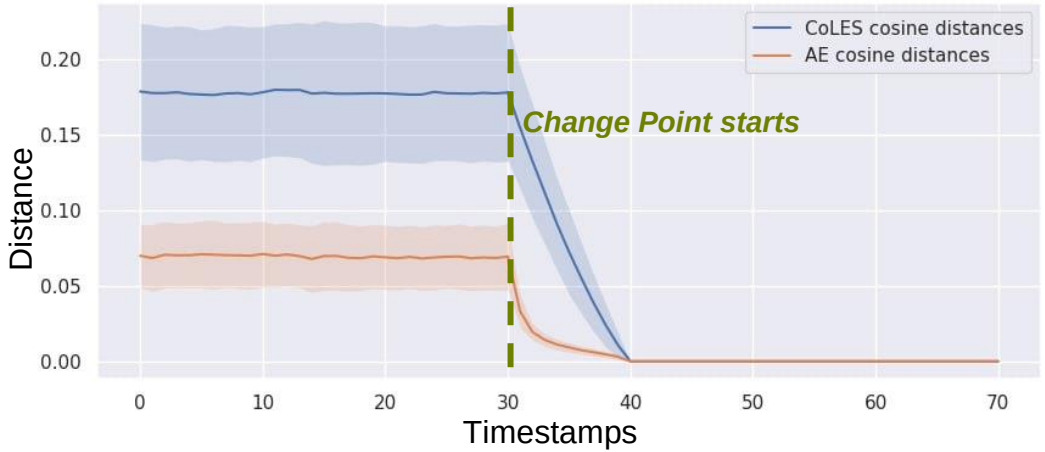
CoLES (contrastive) vs AE (generative): reaction to change

We also evaluate the models' **ability to detect user behavior change.** See an artificial change.

Experiment: "A poor man won a lottery".



Augmentation procedure. User 1 transactions were replaced with User 2 transactions. We compare User 2 to the augmented User 1.



Cosine distance between embeddings obtained from raw users and augmented ones. Snapshot near the Change Point

We expect embedding during the "augmented" area will be close to each other and far during other timestamps.

Skoltech

Global properties of models

Ranks for a local problem

	Age	Churn	Default	HSBC	Mean
AR§	1	1	1	1	1.00
MLM§	3	1	2	1	1.75
CoLES ext.†	3	2	2	3	2.50
AE§	2	3	4	2	2.75
CoLES†	3	4	3	3	3.25
Best baseline	4	4	5	3	4.00
TS2Vec†	6	4	5	4	4.75
A-NHP‡	5	5	6	4	5.00
NHP‡	5	5	6	4	5.00
COTIC‡	6	6	7	5	6.00

Ranks for a global problem

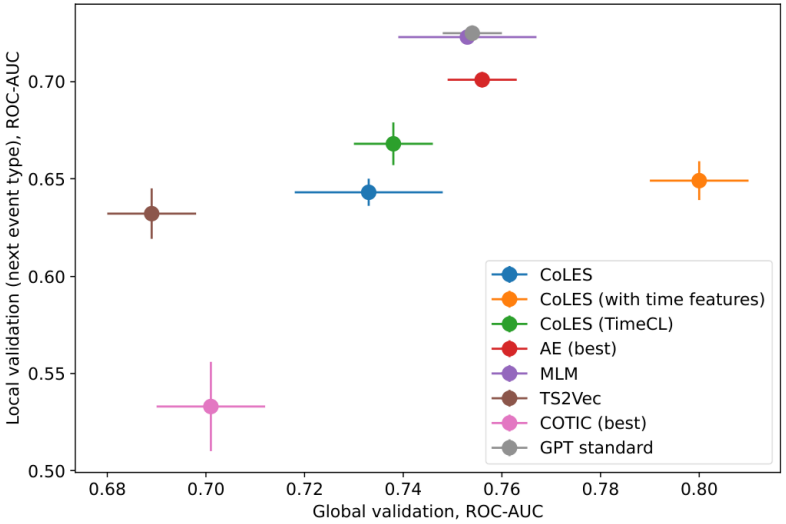
	Age	Churn	Default	HSBC	Mean
CoLES ext.†	1	1	1	1	1.00
CoLES†	1	3	1	1	1.50
MLM§	2	2	2	2	2.00
Best baseline	1	4	2	2	2.25
AR§	3	2	1	3	2.25
AE§	4	2	2	2	2.50
NHP‡	5	2	2	3	3.00
COTIC‡	6	3	1	4	3.50
TS2Vec†	2	5	2	5	3.50
A-NHP‡	5	3	3	4	3.75

Models are colour-coded: blue for generative, green for contrastive and fuchsia for TPP.

Skoltech

Bazarova, Alexandra, et al. "Universal representations for financial transactional data: embracing local, global, and external contexts." *arXiv preprint arXiv:2404.02047* (2024).

Comparison of local and global properties of models



Main conclusions:

- GPT is better in local task.
- CoLES with time features is a clear leader in global validation.

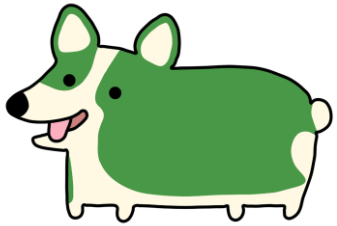
Skoltech

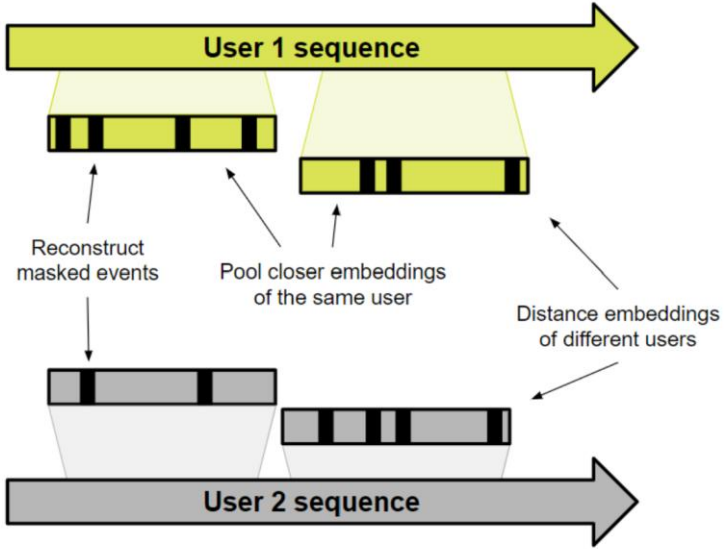
Bazarova, Alexandra, et al. "Universal representations for financial transactional data: embracing local, global, and external contexts." *arXiv preprint arXiv:2404.02047* (2024).

Combining contrastive learning and autoregression

Skoltech

Combining local and global properties





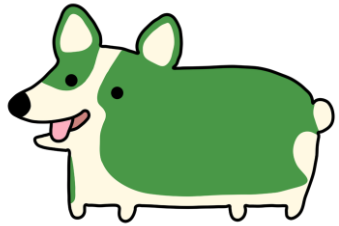
1. Generative reconstruction embeddings of masked events
2. Contrastive comparison of embeddings from different users

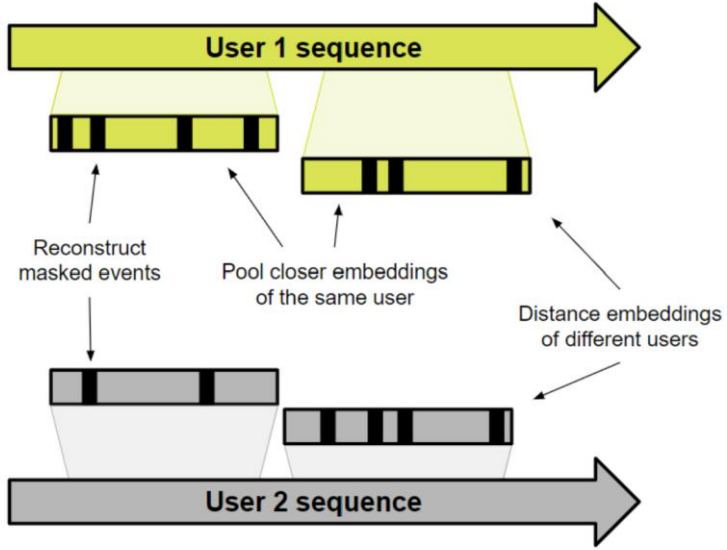
We simultaneously reconstruct embeddings with our CMLM and contrast in CoLES style

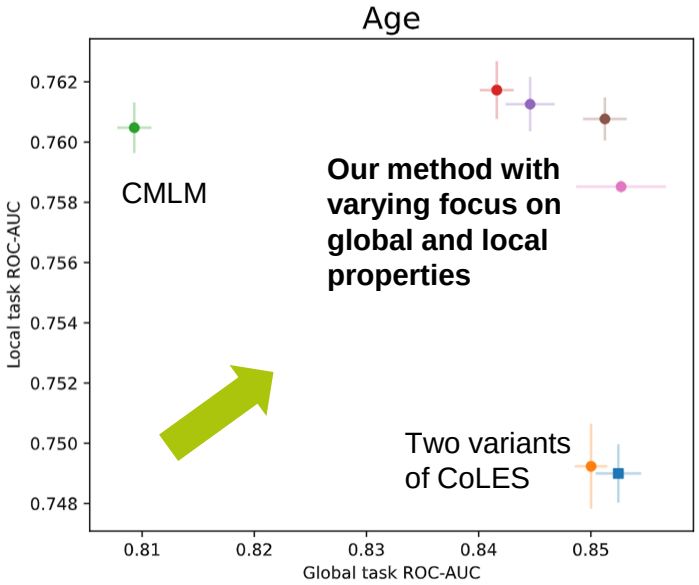
Skoltech

Yugay, Aleksandr, and Alexey Zaytsev. Uniting contrastive and generative learning for event sequences models. AIST. 2024.

Combining local and global properties



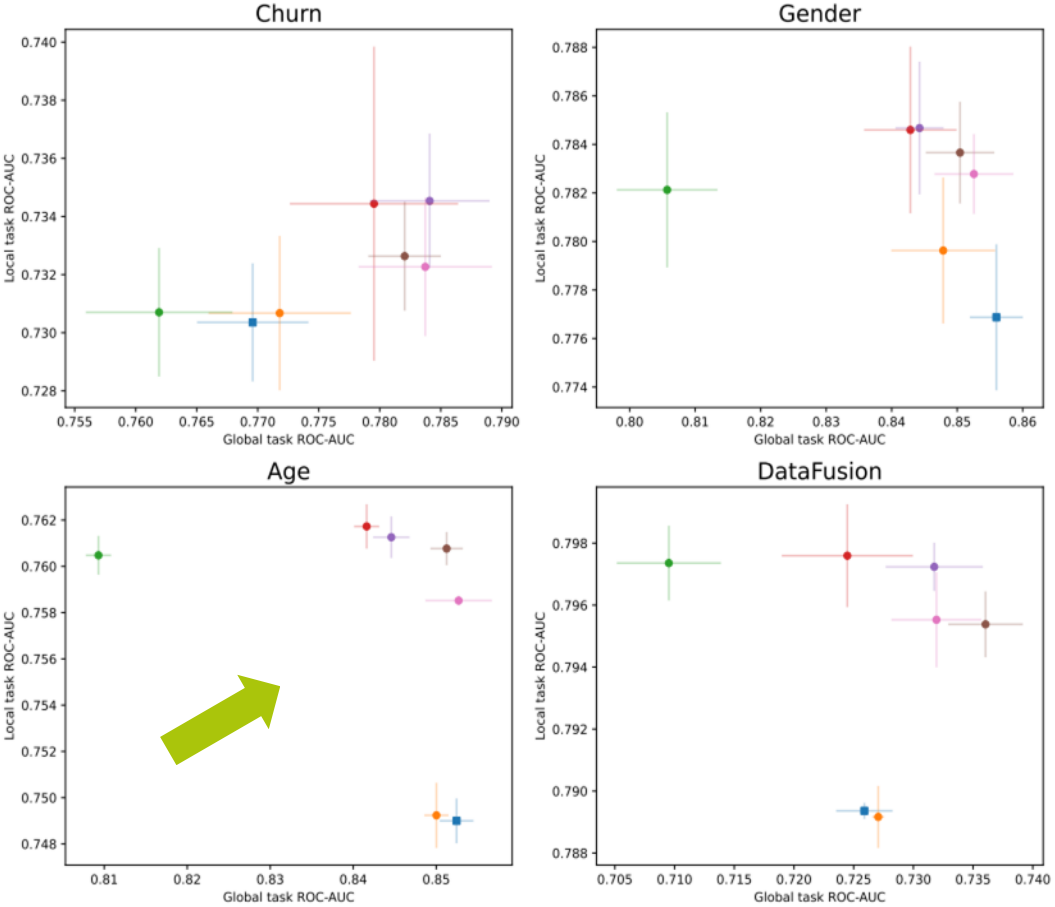




We simultaneously reconstruct embeddings with our CMLM and contrast in CoLES style

Skoltech

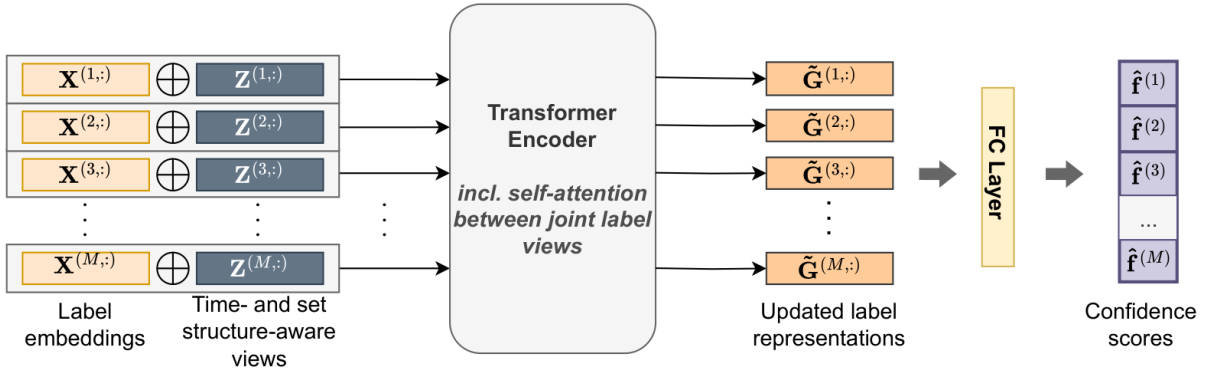
Results



Method	*Global Task ROC-AUC*	*Local Task ROC-AUC*
	Churn	
CoLES	0.770±0.007	0.730±0.003
CoLES (masking)	0.772±0.007	0.731±0.003
CMLM	0.762±0.010	0.731±0.004
CMLM+CoLES ($\lambda = 0.1$)	0.780±0.008	0.734±0.006
CMLM+CoLES ($\lambda = 0.05$)	**0.784±0.008**	**0.735±0.004**
CMLM+CoLES ($\lambda = 0.01$)	0.782±0.005	0.733±0.003
CMLM+CoLES ($\lambda = 0.005$)	**0.784±0.009**	0.732±0.004
	Gender	
CoLES	**0.856±0.005**	0.777±0.004
CoLES (masking)	0.848±0.009	0.780±0.003
CMLM	0.806±0.009	0.782±0.004
CMLM+CoLES ($\lambda = 0.1$)	0.843±0.007	**0.785±0.004**
CMLM+CoLES ($\lambda = 0.05$)	0.844±0.004	**0.785±0.003**
CMLM+CoLES ($\lambda = 0.01$)	0.850±0.005	0.784±0.002
CMLM+CoLES ($\lambda = 0.005$)	0.853±0.008	0.783±0.002
	Age	
CoLES	0.852±0.002	0.749±0.001
CoLES (masking)	0.850±0.001	0.749±0.001
CMLM	0.809±0.002	0.760±0.001
CMLM+CoLES ($\lambda = 0.1$)	0.842±0.002	**0.762±0.001**
CMLM+CoLES ($\lambda = 0.05$)	0.845±0.002	0.761±0.001
CMLM+CoLES ($\lambda = 0.01$)	0.851±0.002	0.761±0.001
CMLM+CoLES ($\lambda = 0.005$)	**0.853±0.005**	0.759±0.000
	DataFusion	
CoLES	0.726±0.003	0.789±0.000
CoLES (masking)	0.727±0.001	0.789±0.001
CMLM	0.710±0.005	0.797±0.001
CMLM+CoLES ($\lambda = 0.1$)	0.724±0.005	**0.798±0.001**
CMLM+CoLES ($\lambda = 0.05$)	0.732±0.005	0.797±0.001
CMLM+CoLES ($\lambda = 0.01$)	**0.736±0.003**	0.795±0.001
CMLM+CoLES ($\lambda = 0.005$)	0.734±0.005	0.795±0.001

Skoltech

You were looking at a wrong self-attention?



We compute self-attention over event types and get prediction of next event type, imposing simple aggregation of temporal encodings.

Our LaNET model is now SOTA for the next basked prediction

Paper

GitHub

Kovtun, Elizaveta, et al. Label attention network for sequential multi-label classification: you were looking at a wrong self-attention. ECAI. 2024.

Skoltech

Few final words

Skoltech

Conclusion

- Typical SSL approaches focus on different aspects of embedding properties, also demonstrating generative capabilities
- We propose an SSL hybrid approach CMLM+CoLES that achieve notable improvements in both local and global properties of learned representations.
- Generative models for event sequences data are on their way!

Thanks my lab for help with these slides and you for your attention!

Alexandra Bazarova
Maria Kovaleva
Ilya Kuleshov
Evgenia Romanenkova
Alexander Stepikin
Alexandr Yugay
Elizaveta Kovtun
Galina Boeva
Andrey Shulga
Alexey Zaytsev

Skoltech

Thanks for your attention!

Backslides

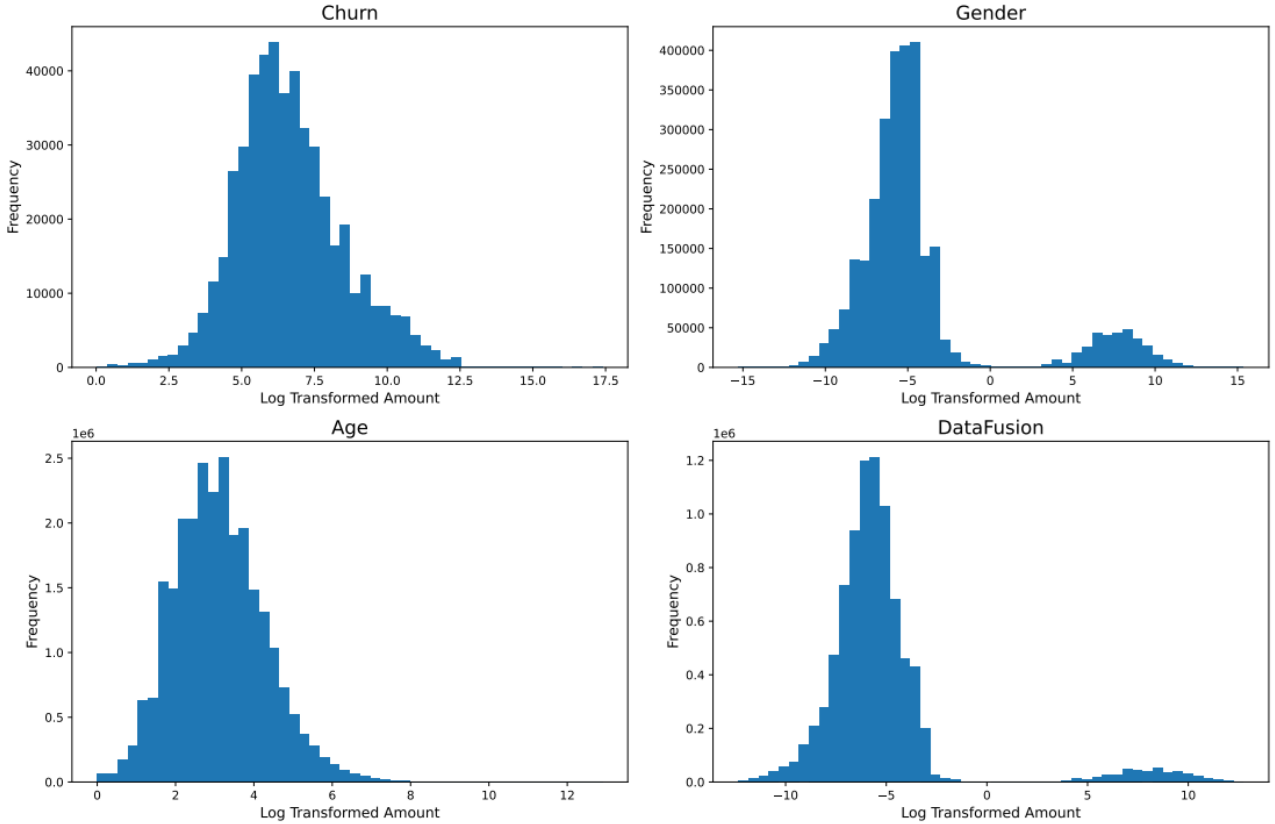
Experiment design

1. Pretrain models in a self-supervised regime
2. Use the obtained encoder as feature extractor
3. Train another model in a supervised regime on extracted features to solve downstream tasks:
   - Sequence classification
   - Next event type prediction

	Churn	Gender	Age	DataFusion
Num Transactions	490K	2.9M	26M	8.7M
Num Sequences	5K	7.4K	30K	64K
Mean Sequence Length	98.1	388.2	881.7	136.5
Std. Sequence Length	78.1	309.4	124.8	148.9
Num Unique MCC	344	184	202	323

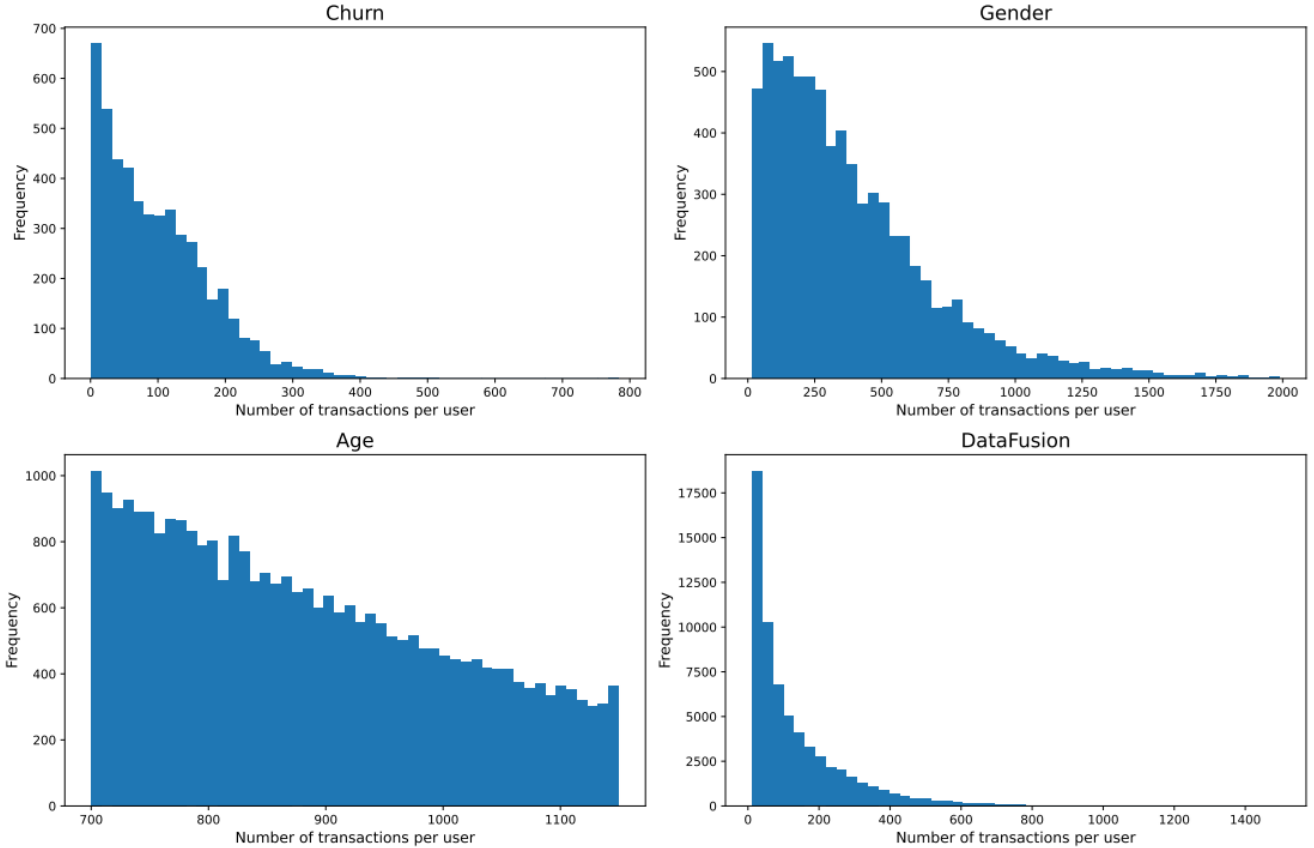
Skoltech

EDA: Amount



Skoltech

EDA: Sequence length



Skoltech

Stock market meets external events: can we predict the reaction?

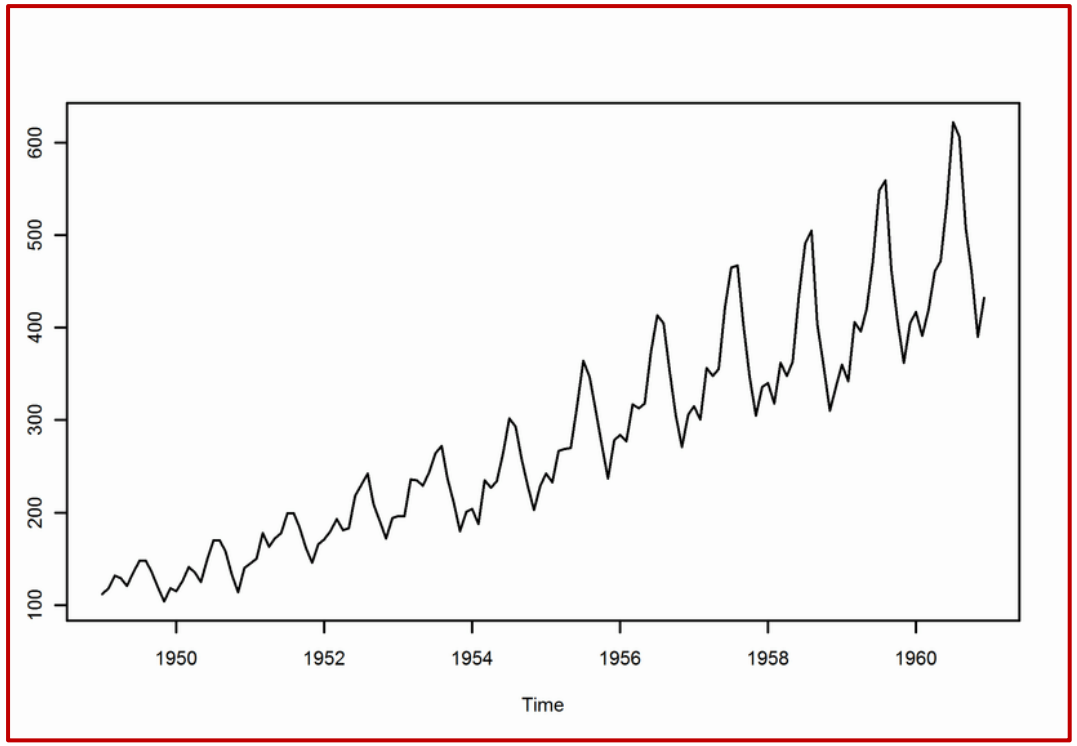
Elizaveta Kovtun
Sber, Skoltech

Stock price

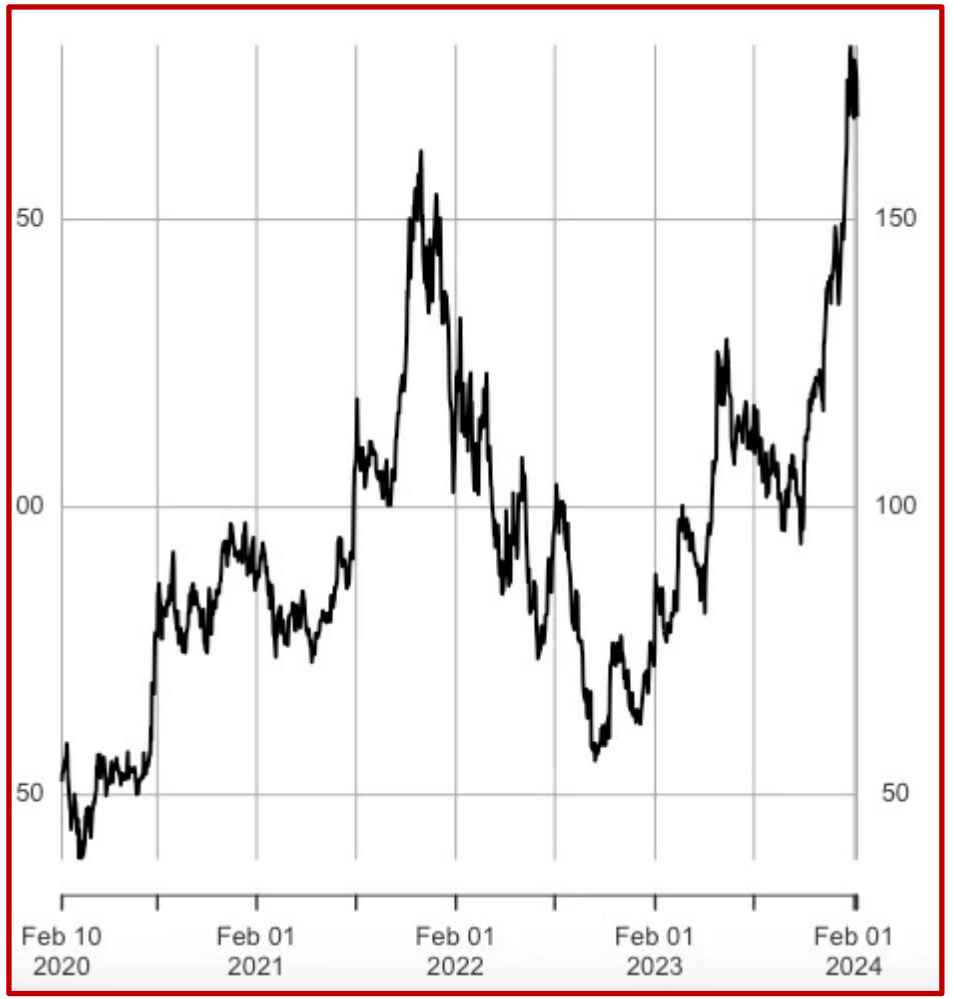
Stock price


What next?
2.84%
145.00
142.50
141.72
140.00
137.50
10/22
10/23
10/24
10/25
Current moment

Time series of different nature



VS



Stock price and external events

Problem statement

Our goal is to predict price change after influential events

Published paper

Scientific Reports (Nature)

New drugs and stock market: a machine learning framework for predicting pharma market reaction to clinical trial announcements

Semen Budennyy, Alexey Kazakov, Elizaveta Kovtun, and Leonid Zhukov

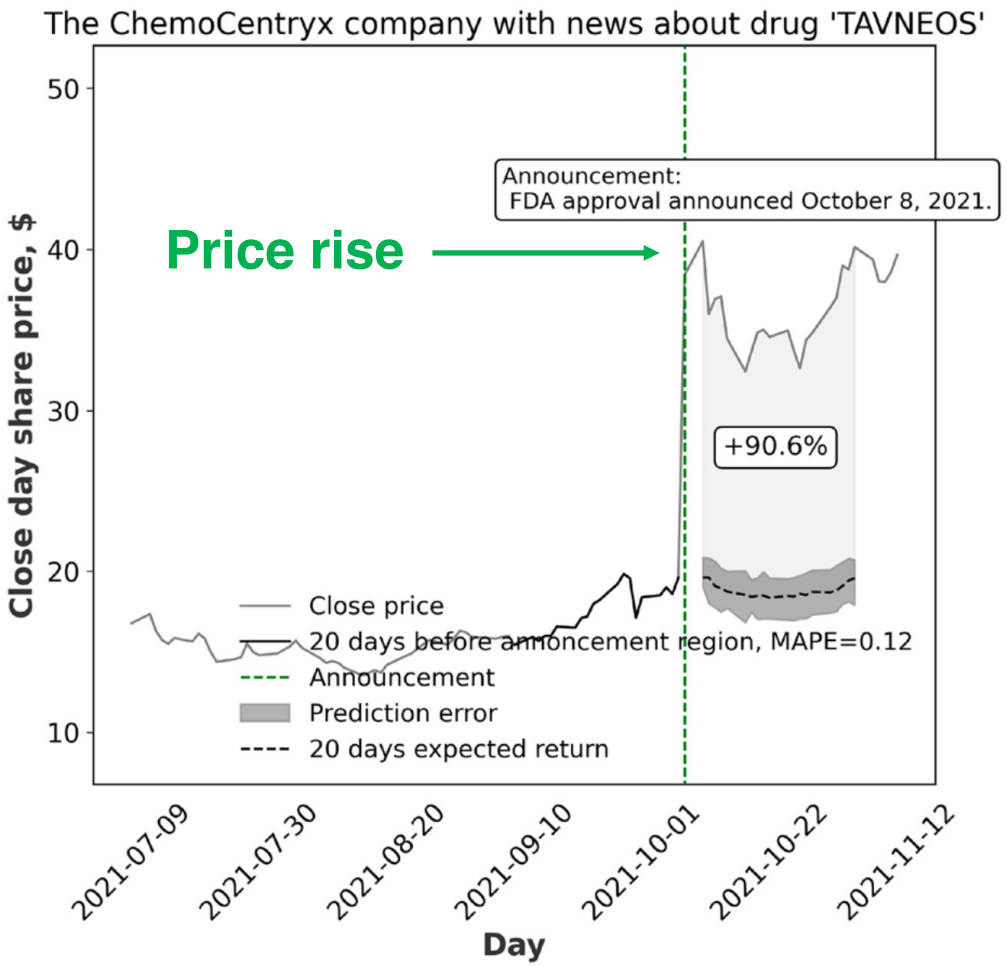
Published paper

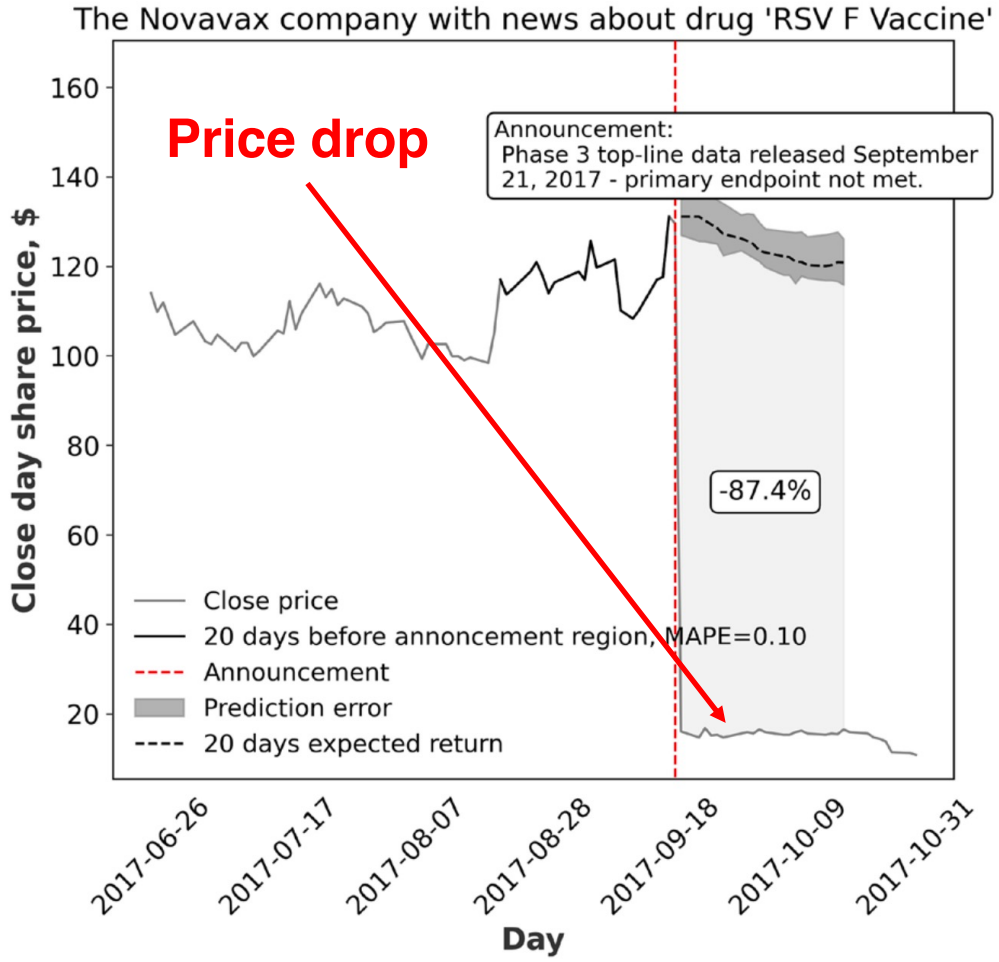
New drugs and stock market: a machine learning framework for predicting pharma market reaction to clinical trial announcements

Time Series

Events

Why pharma market and clinical trial announcements?





1. Sentiment polarity extraction from clinical announcements

Positive Announcements

Keywords: “approve”, “meets”, “show”

Negative Announcements

Keywords: “failed”, “halted”, “did not reach”

Neutral Announcements

1. Compose dictionaries with keywords that reflect the announcement polarity
2. Train BERT on announcement texts and a rule-based markup
3. Complement dictionaries with keyword from mistakenly classified texts

1. Sentiment polarity extraction from clinical announcements

Positive Announcements

Keywords: “approve”, “meets”, “show”

+

“demonstrate”, “potential”, “accepted”, “encouraging”

Negative Announcements

Keywords: “failed”, “halted”, “did not reach”

+

“terminated”, “discontinued”, “insufficient”, “paused”

Neutral Announcements

1. Sentiment polarity extraction from clinical announcements

Positive Announcements

Negative Announcements

Neutral Announcements

Rule-based markup is reasonable since a message of announcement is straightforward

2. Construction of feature space

Market features

- NASDAQ biotechnology index
- Mean number of trading volume peaks per year
- Stock price trend for the last 30 days before the event

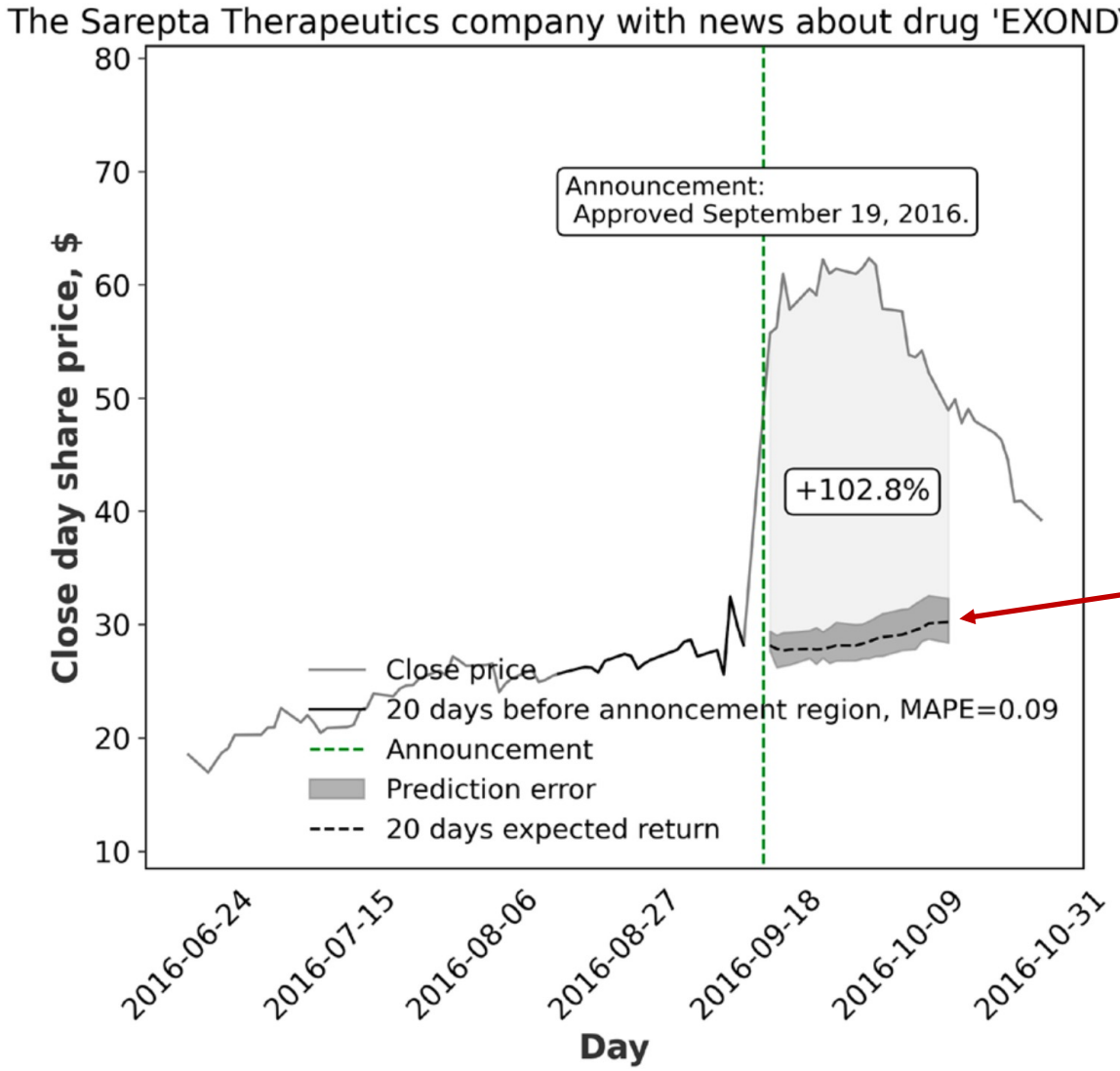
Company features

- Income statement
- Full-Time Employees
- Cash flow

Announcement features

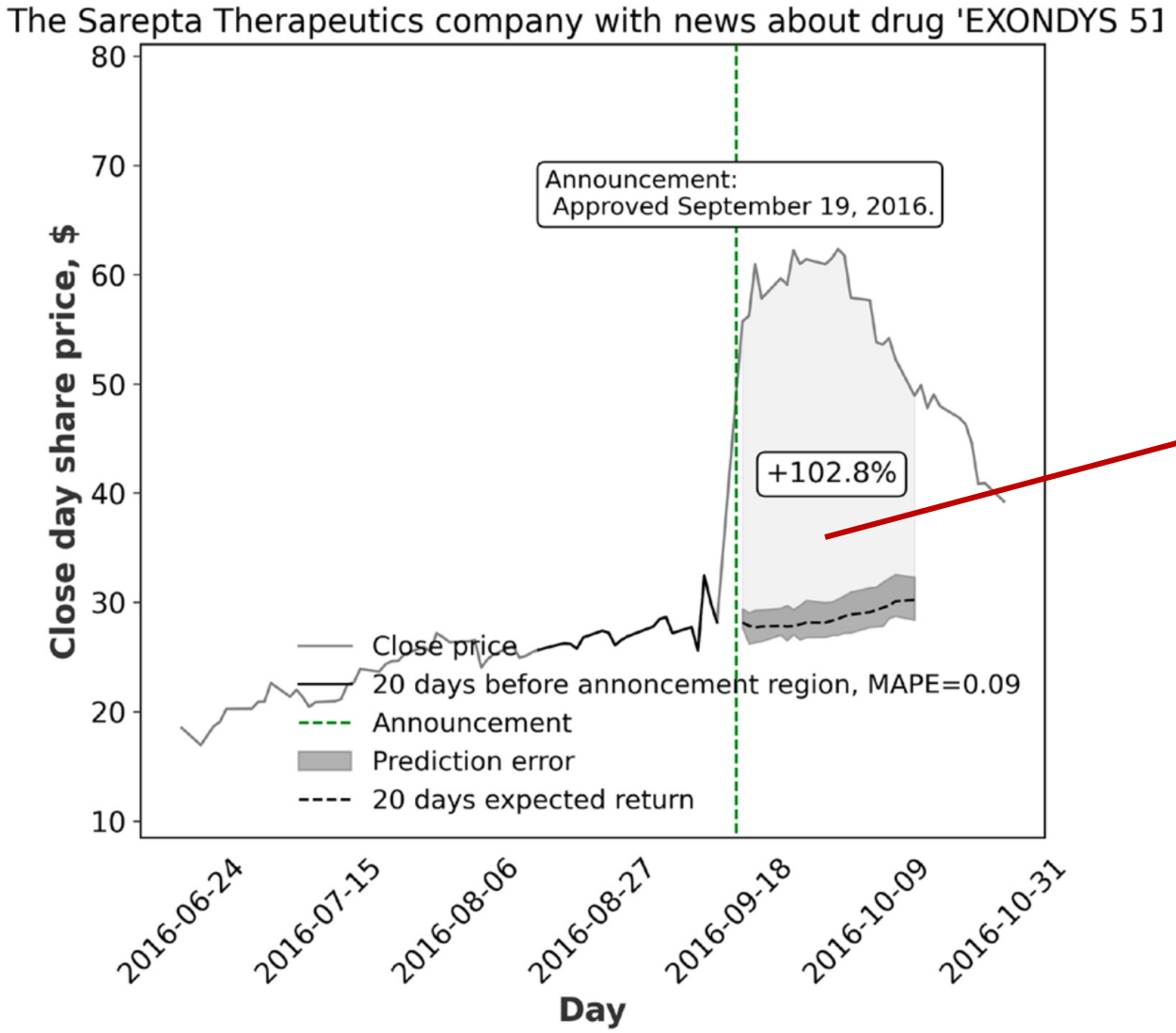
- Announcement sentiment polarity
- ICD-10 codes

3. Evaluation of expected return



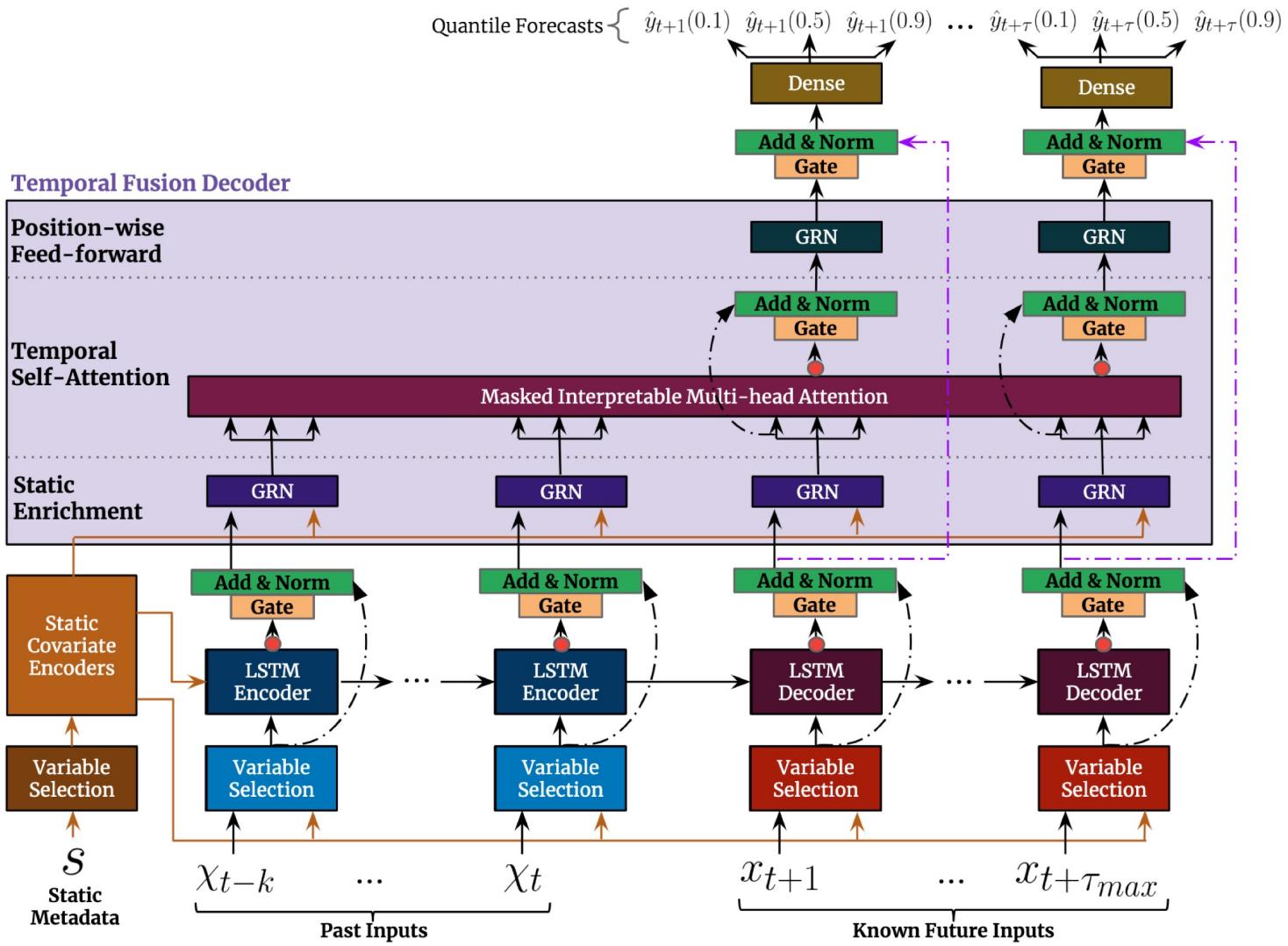
Expected return

3. Evaluation of expected return

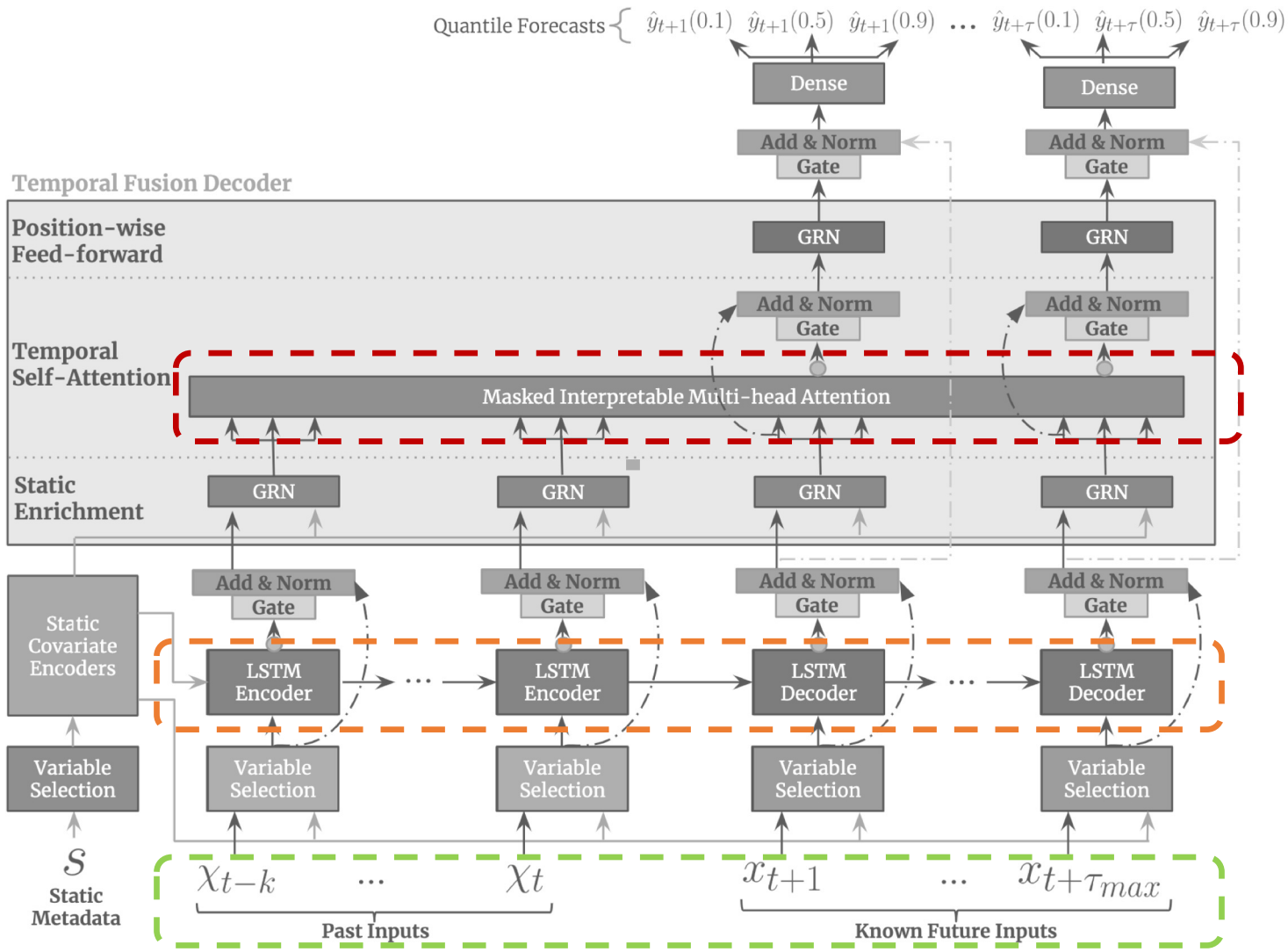


Target measure:
NCAR_20

3. Evaluation of expected return



3. Evaluation of expected return



Attention mechanism

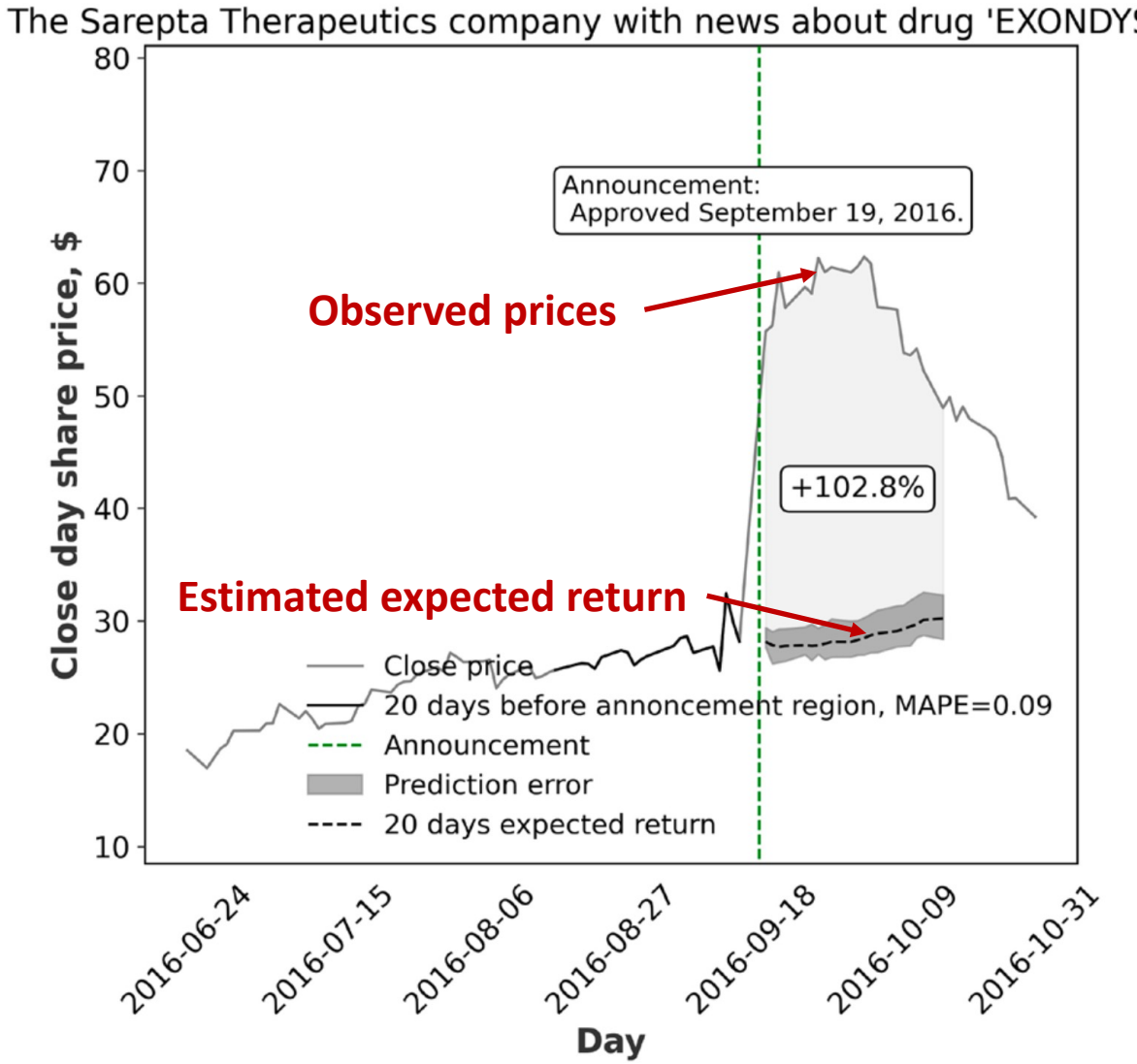
Recurrent module

Covariates processing

3. Evaluation of expected return

Estimation of expected return:

3. Evaluation of expected return



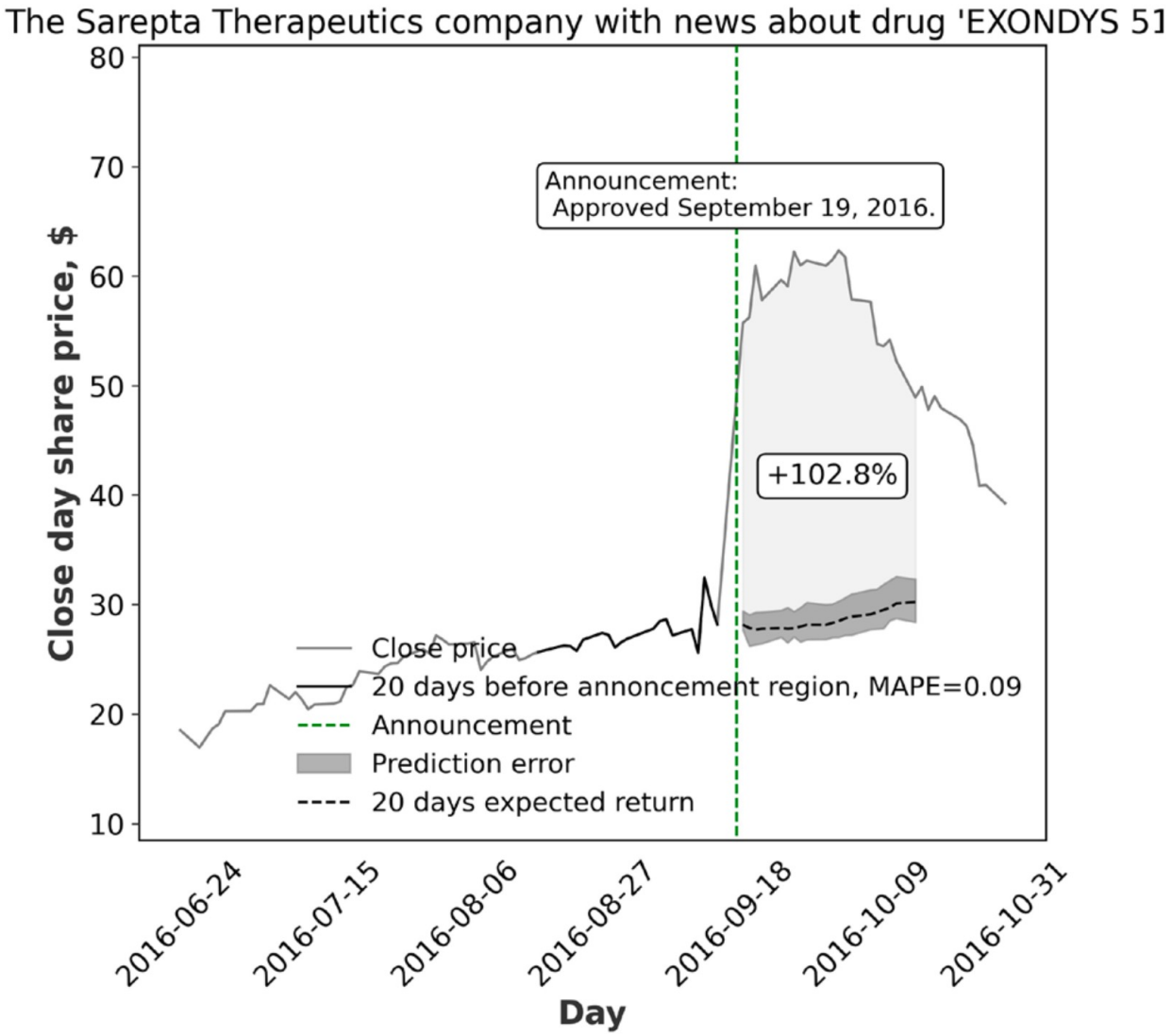
Expected return estimation allows calculating a target value, NCAR_20

From regression to classification

Regression setting:
Prediction of NCAR_20

Multi-class classification setting:
Prediction of NCAR_20 change **range**

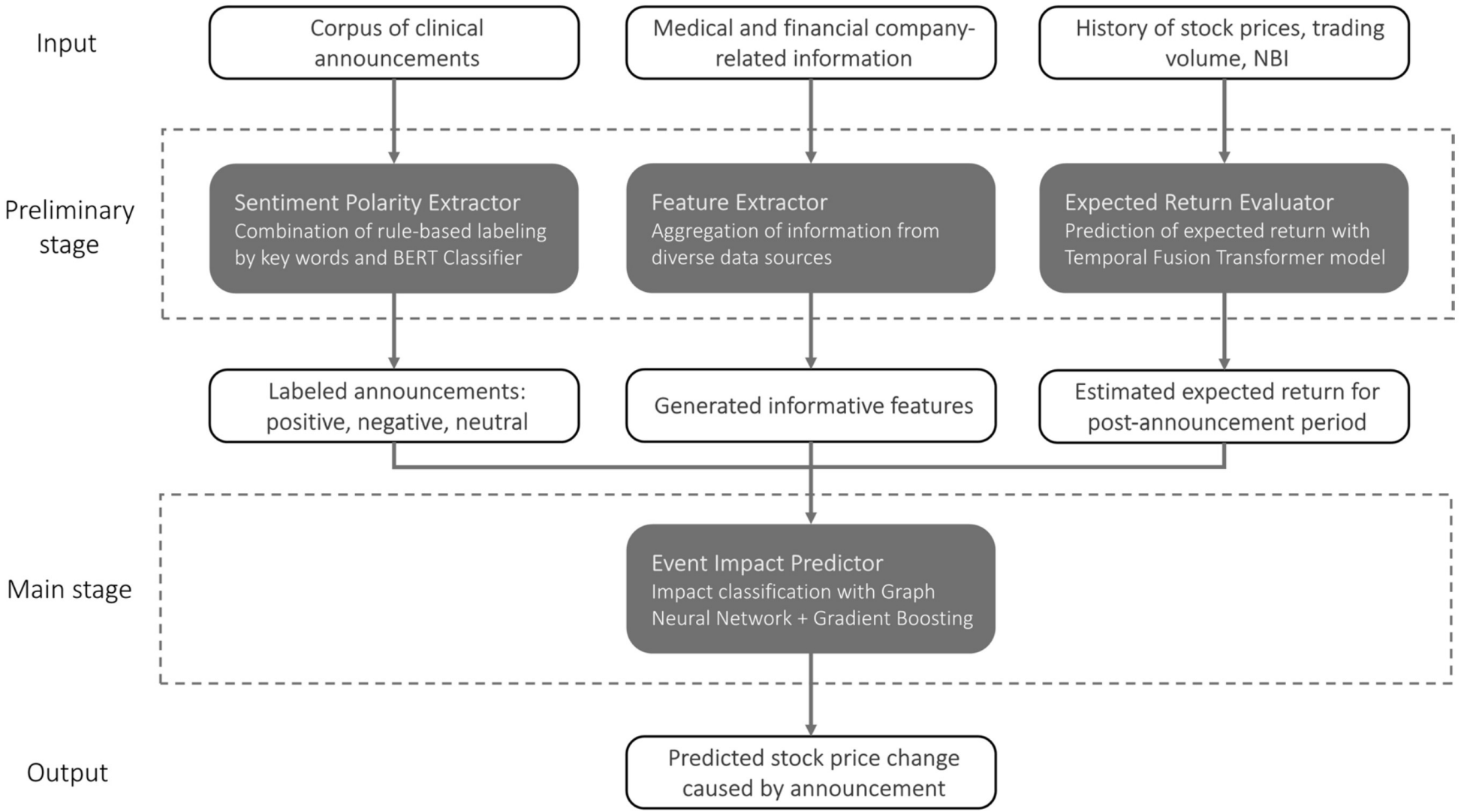
General pipeline



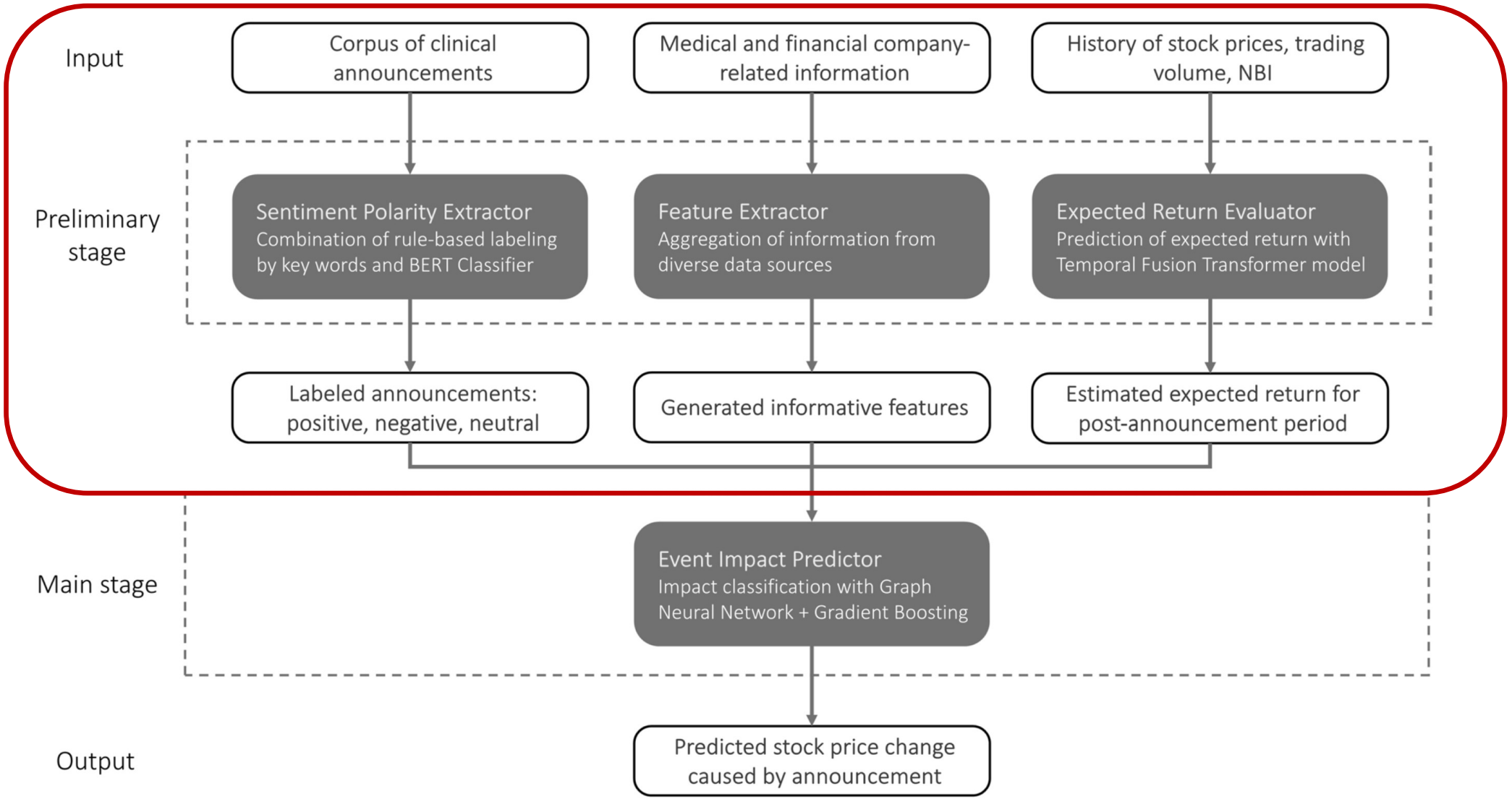
Input: time series + event

Output: prediction of event influence

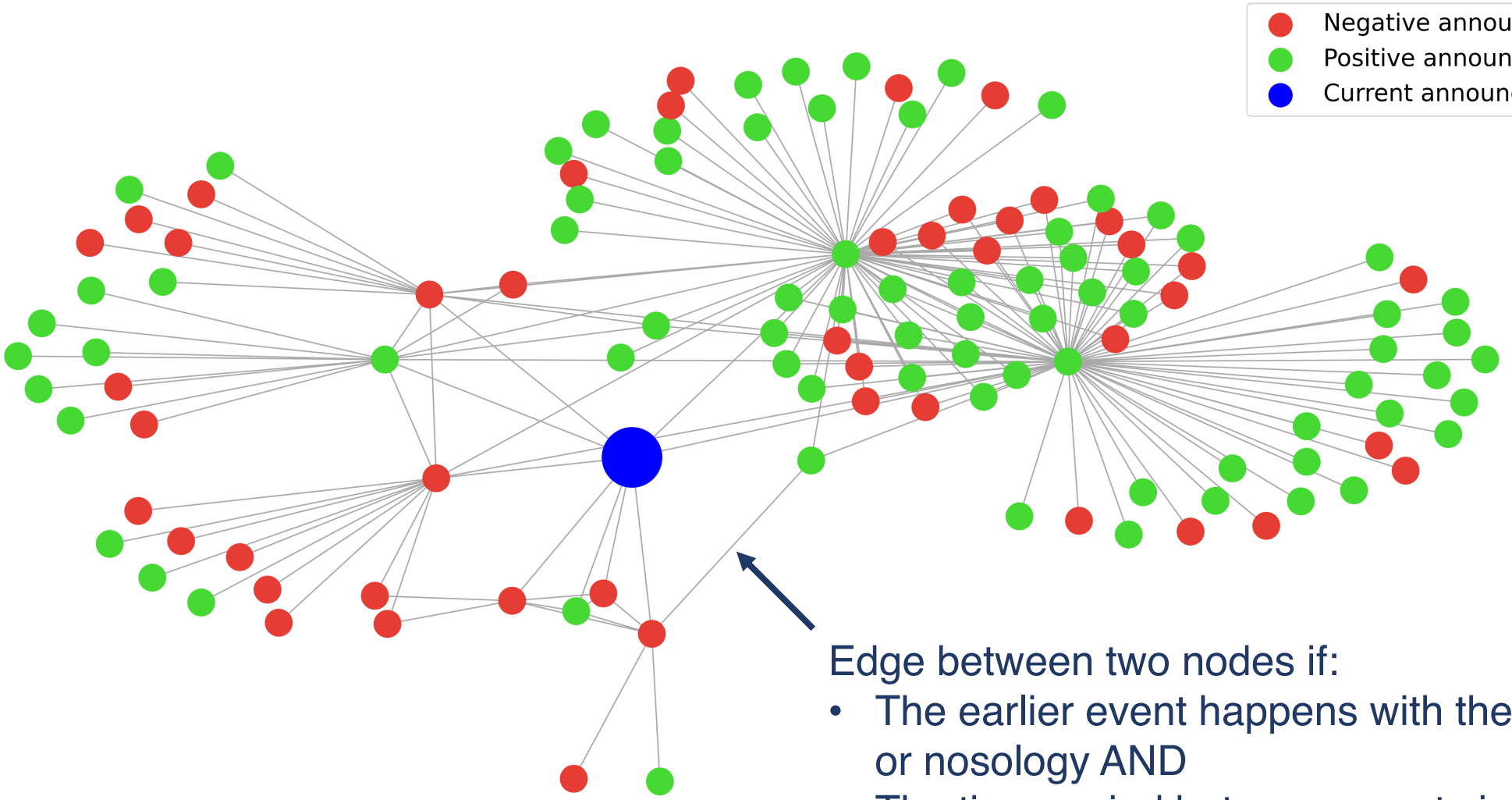
General pipeline



General pipeline



Adoption of GCN



Edge between two nodes if:

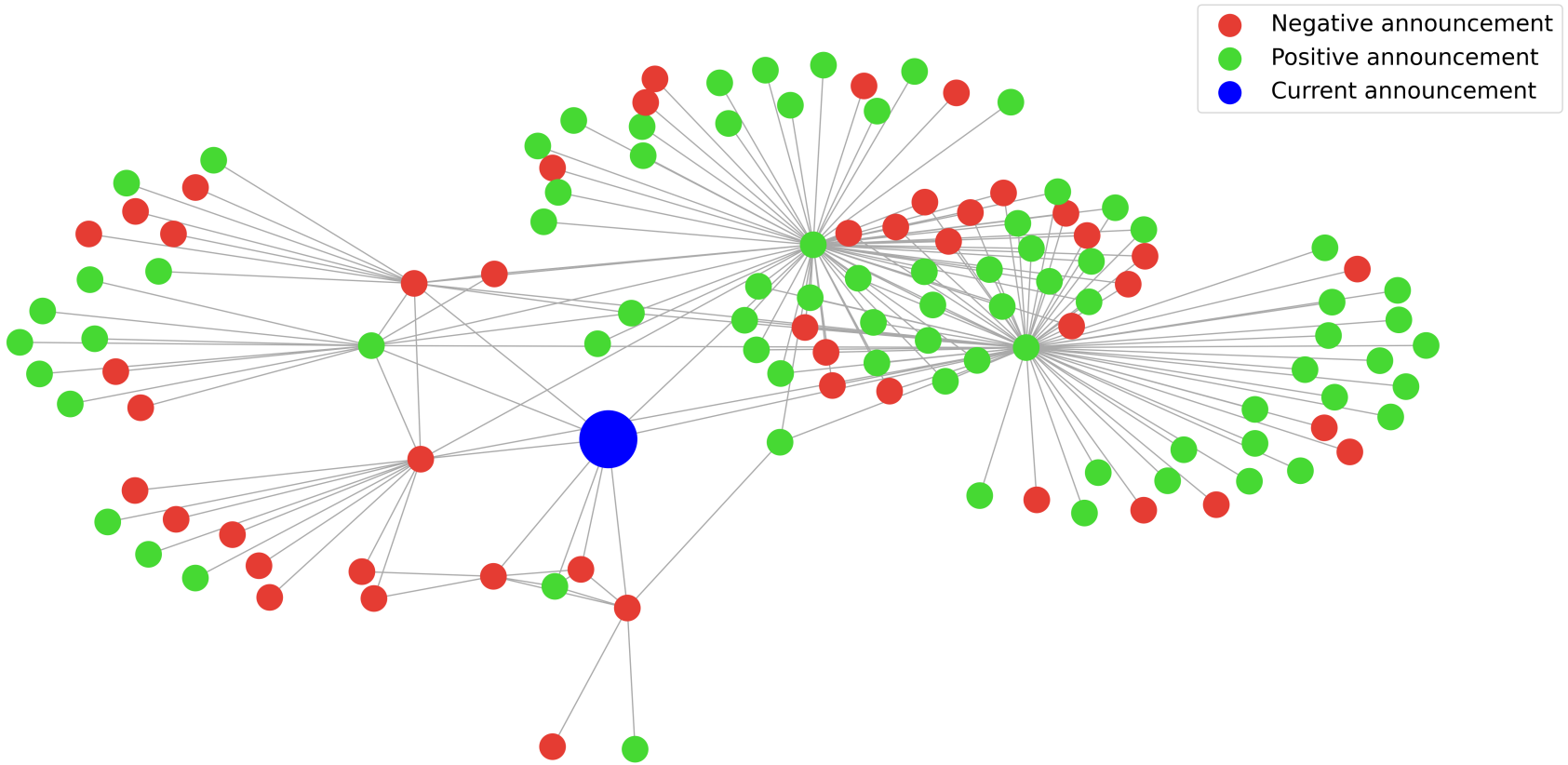
- The earlier event happens with the same company or nosology AND
- The time period between events is less than 1 year

Adoption of GCN

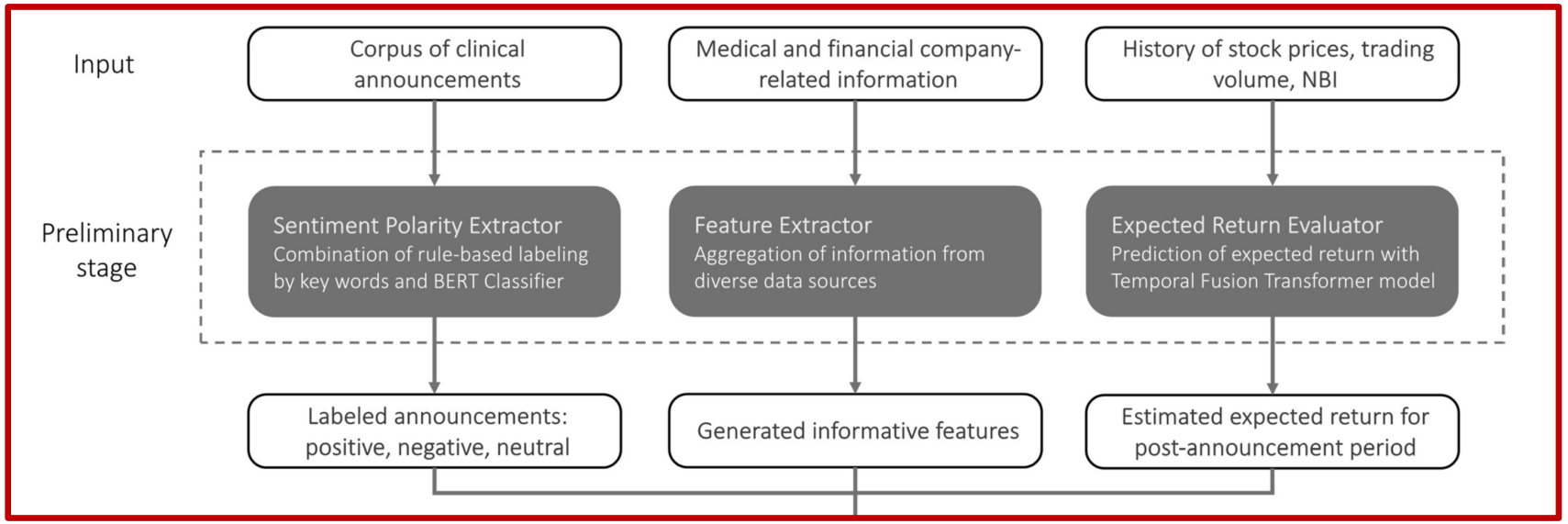
- **Announcement information**
- Graph

→ **GCN** →

Class probabilities of price change range



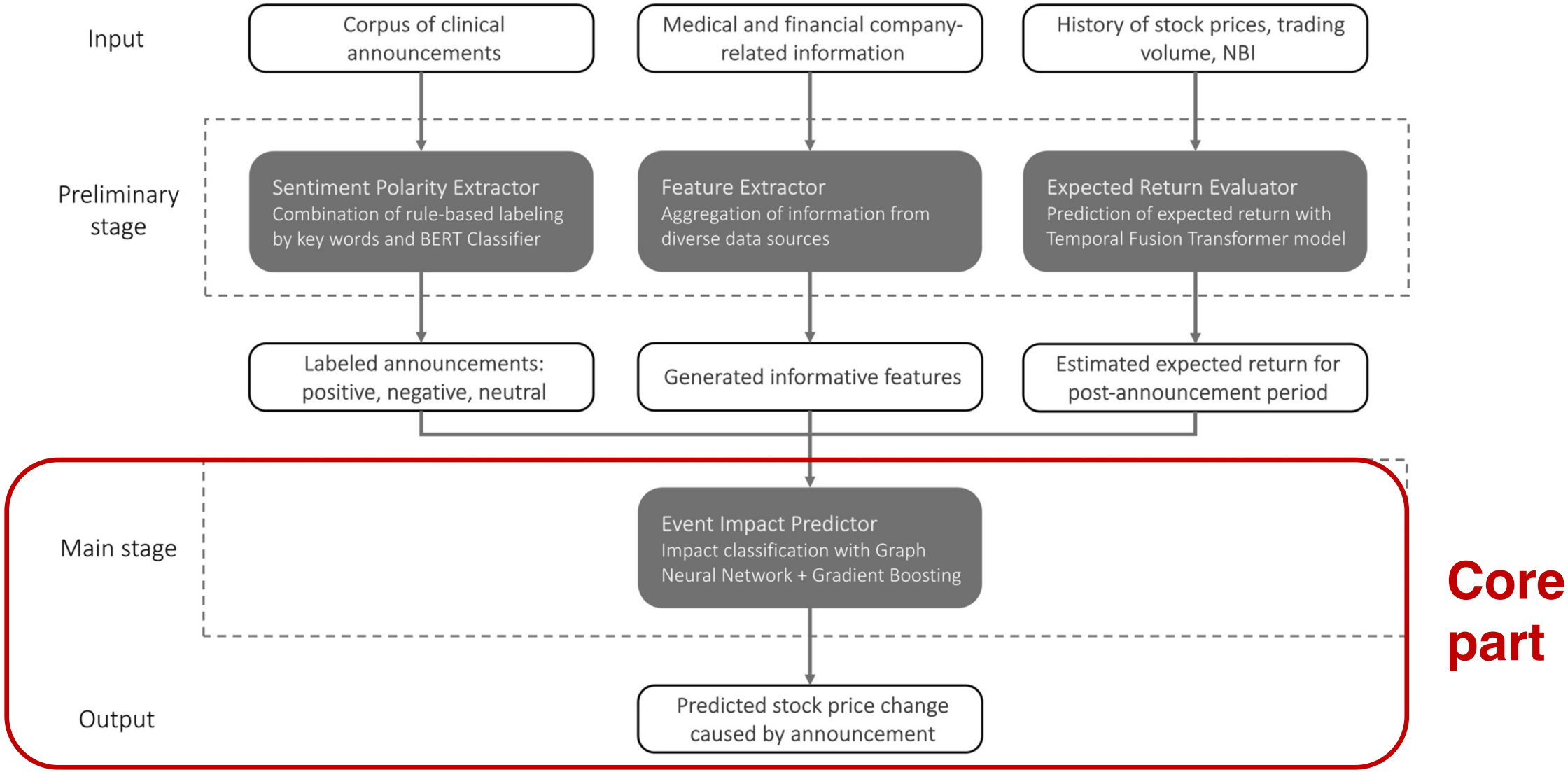
Core classifier



GCN-induced probabilities

Input to Gradient Boosting

Core classifier



Results

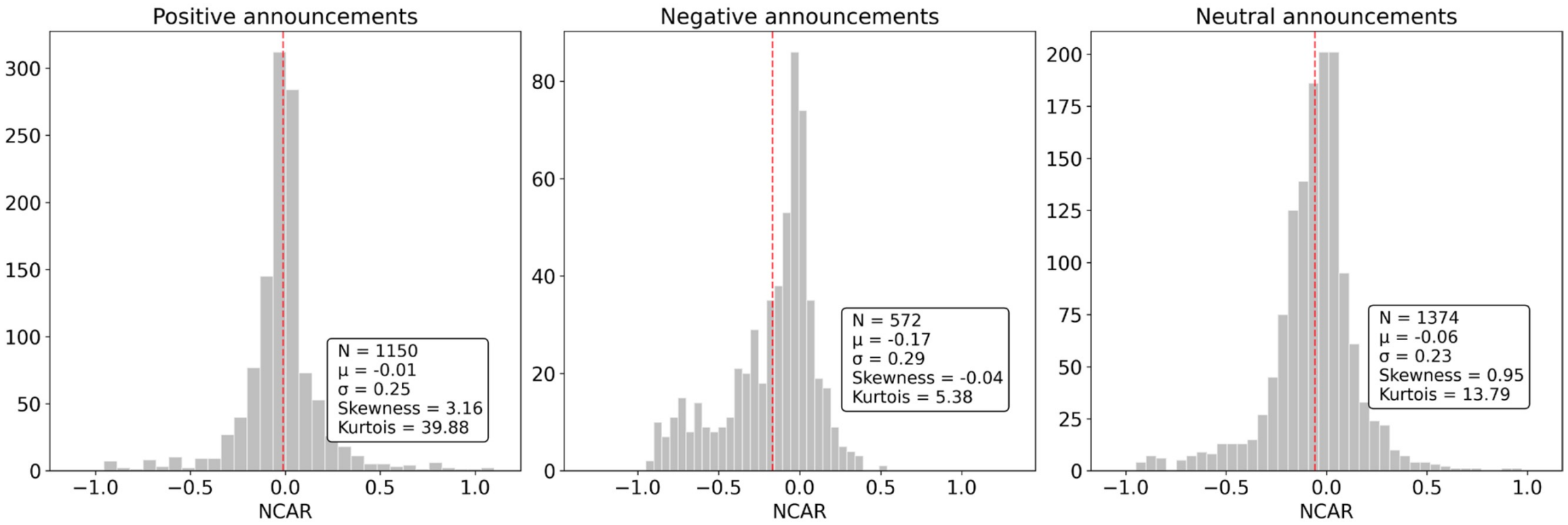
Data

- 5436 clinical trial announcements
- 681 companies
- years 2018-2022

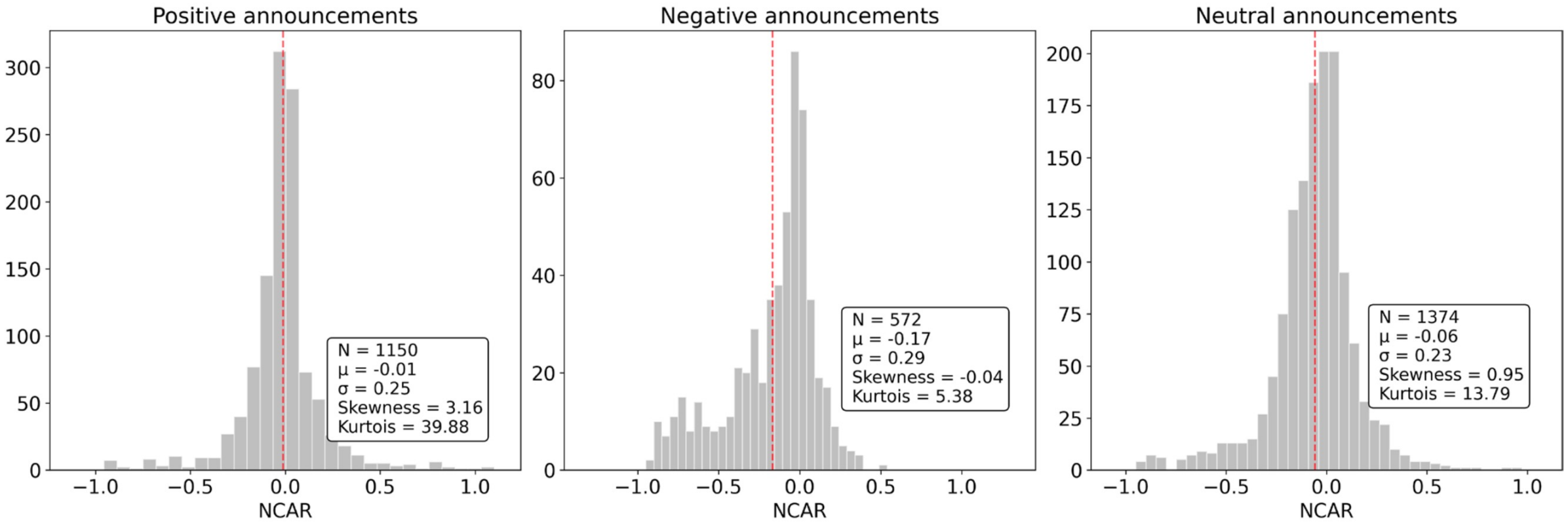
Sentiment polarity evaluation

	# Divergences	# Coinciding positives	# Coinciding negatives	# Coinciding neutrals
With the initial keywords	207	1447	445	337
With the updated keywords	66	1562	765	304

Announcement impact analysis

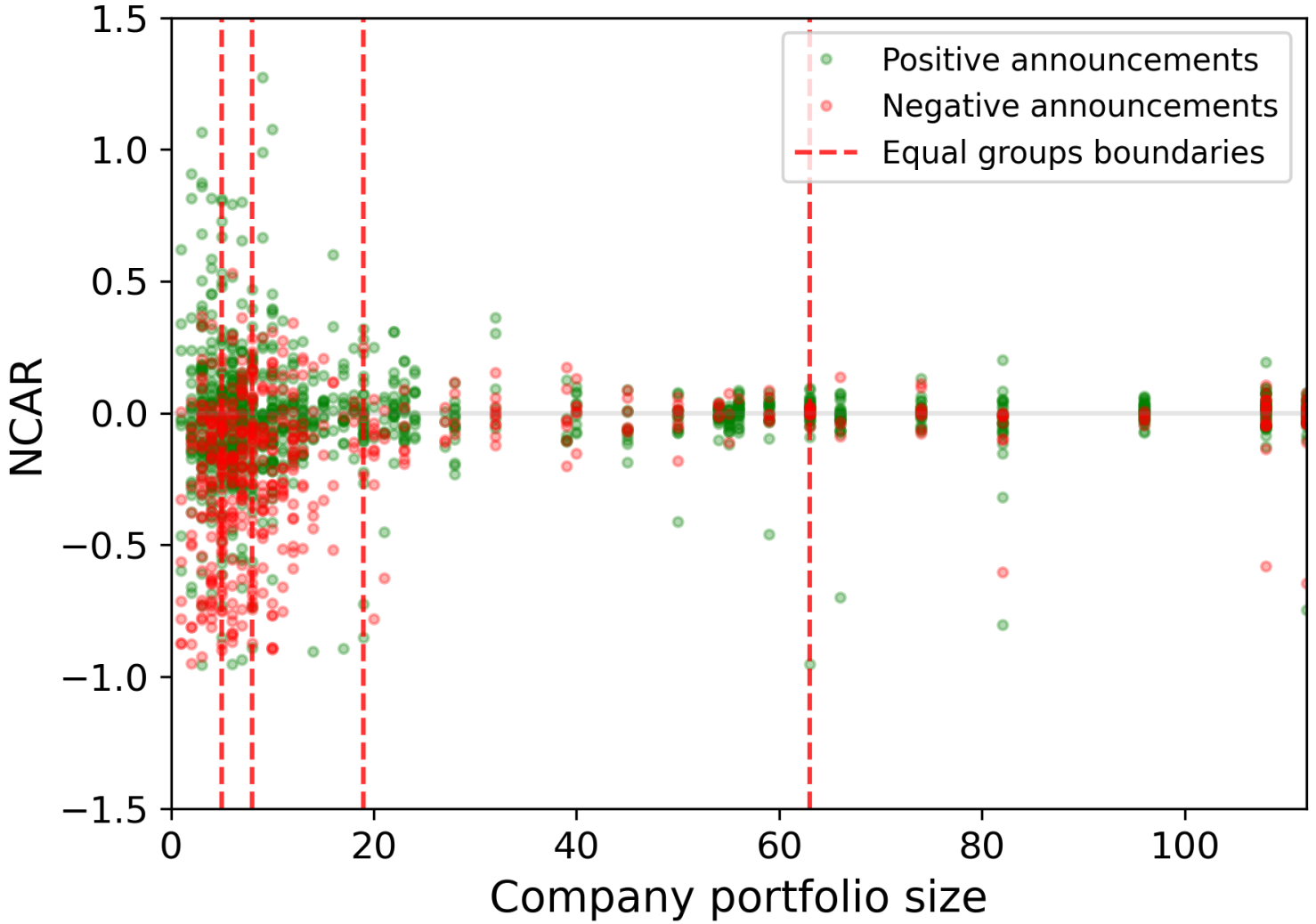


Announcement impact analysis

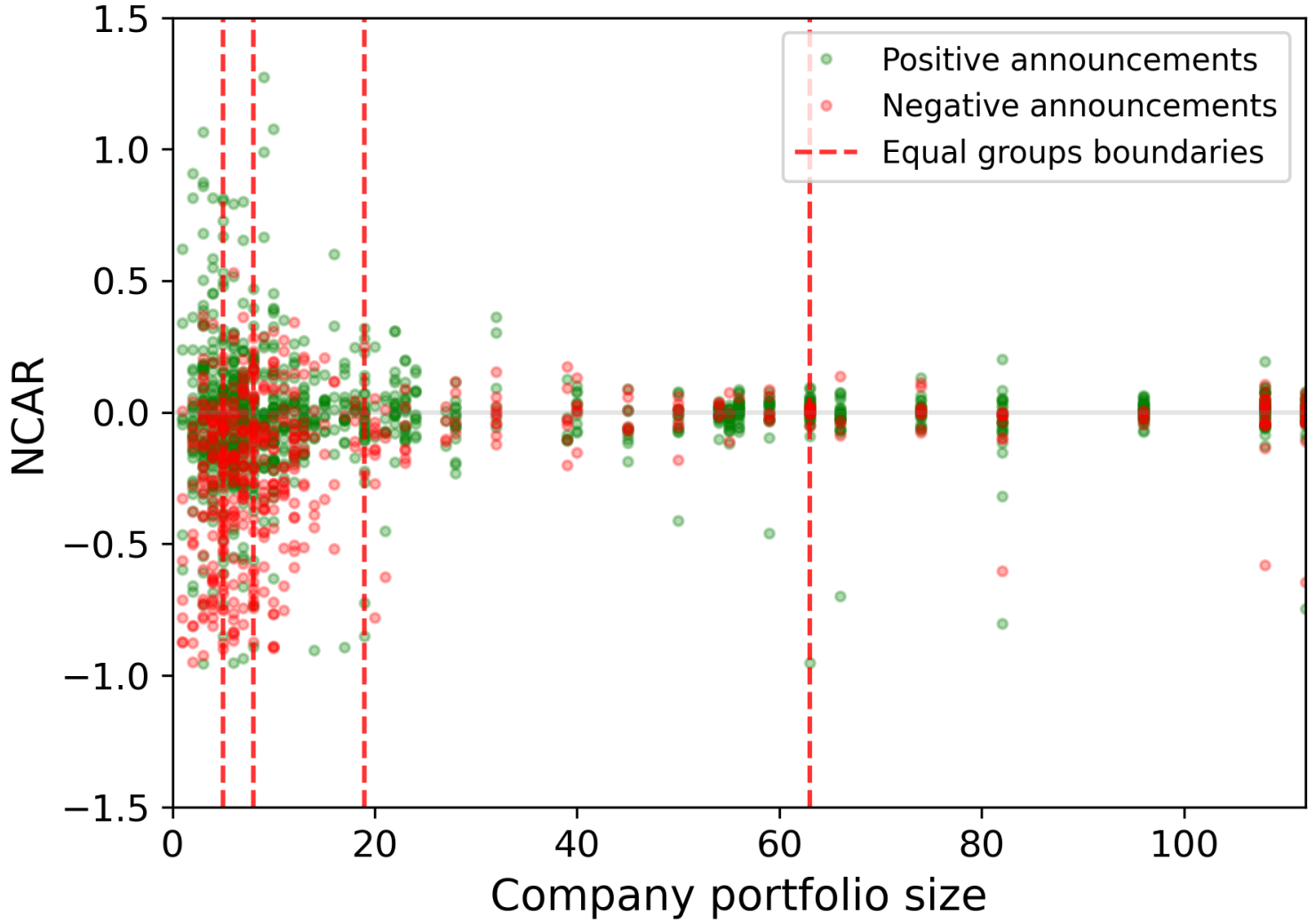


Mann-Whitney U test: negative and non-announcement distributions are different

Impact of company background on stock prices

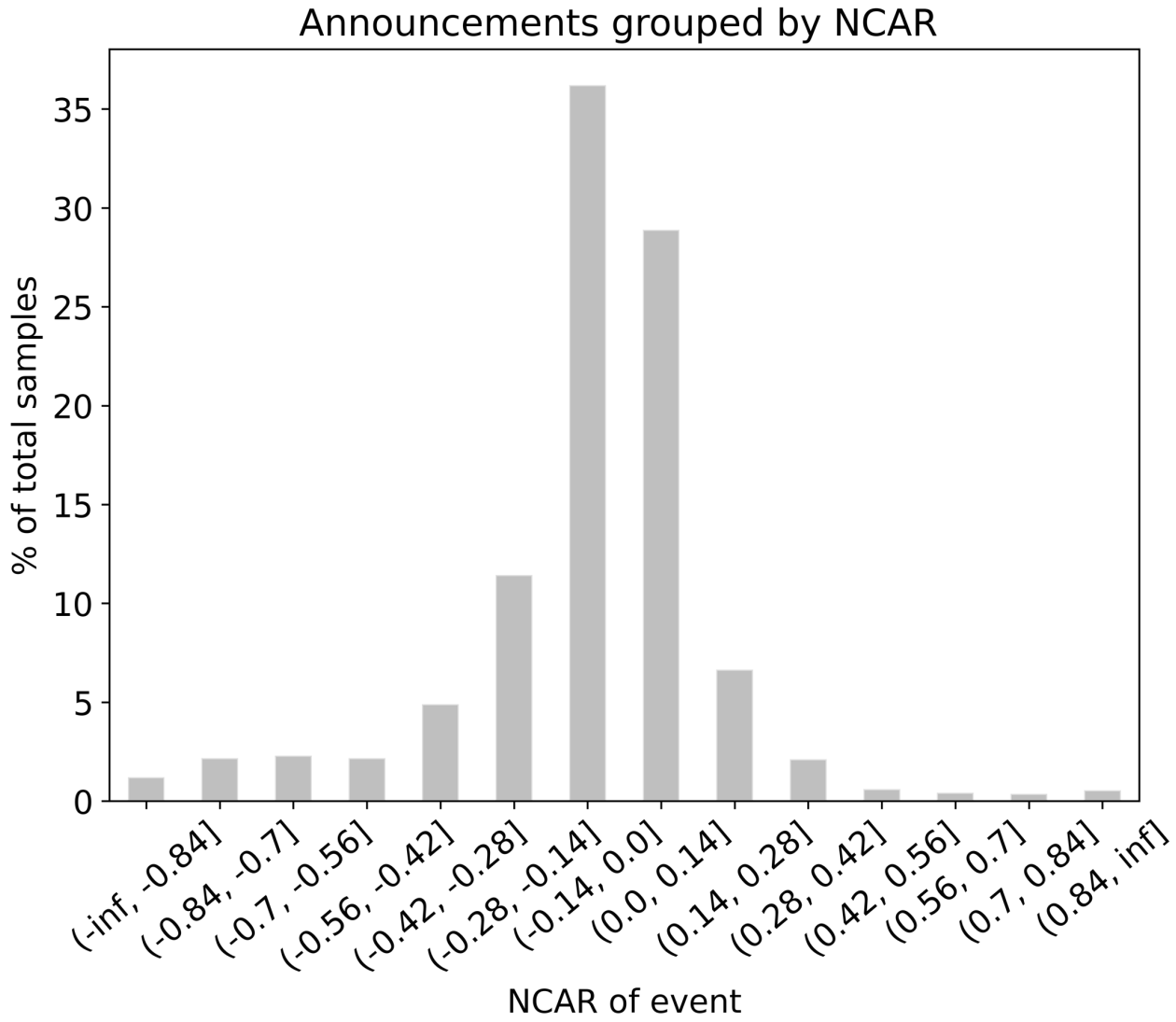


Impact of company background on stock prices

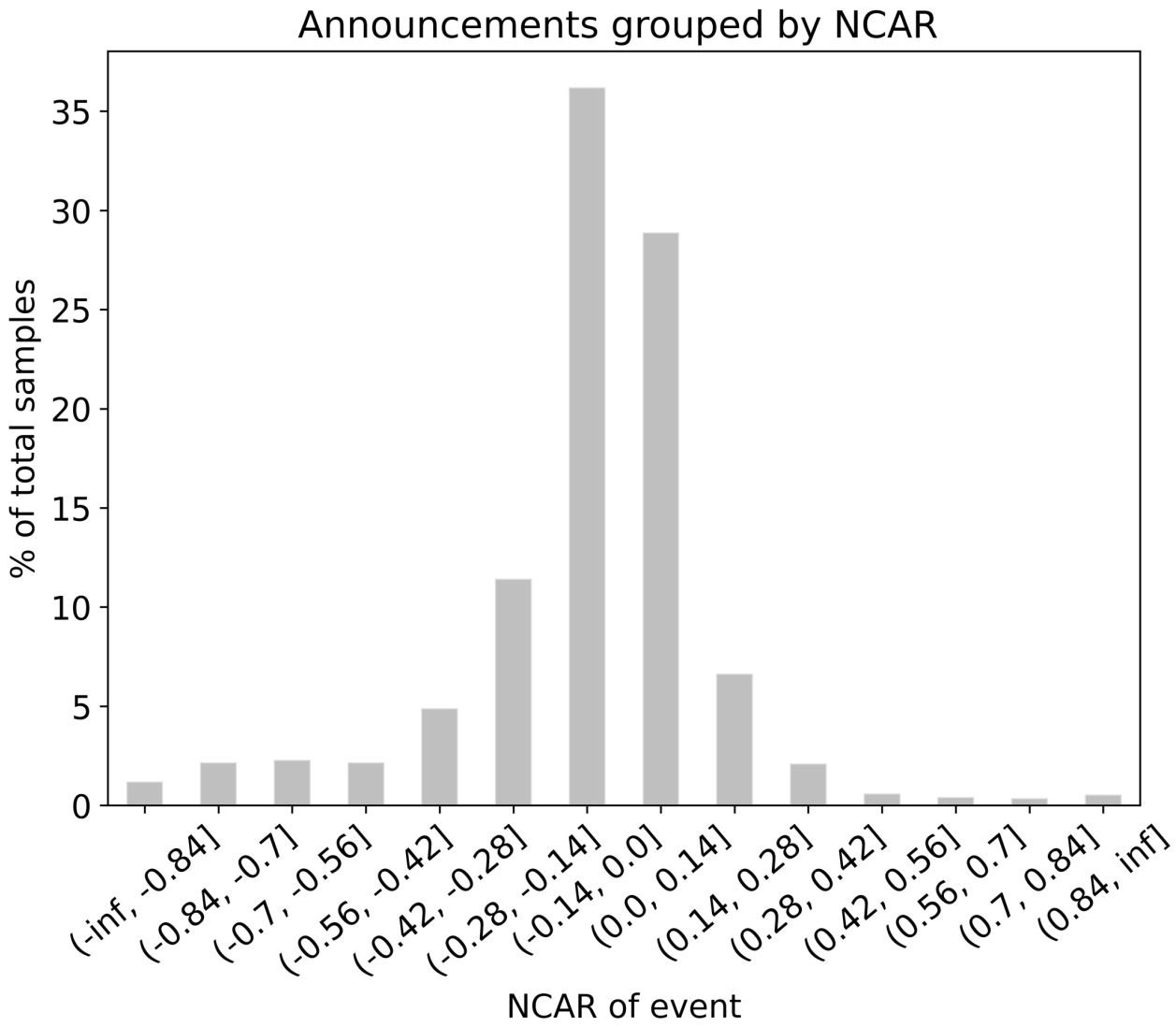


Generation of diverse feature space is important

Class distribution



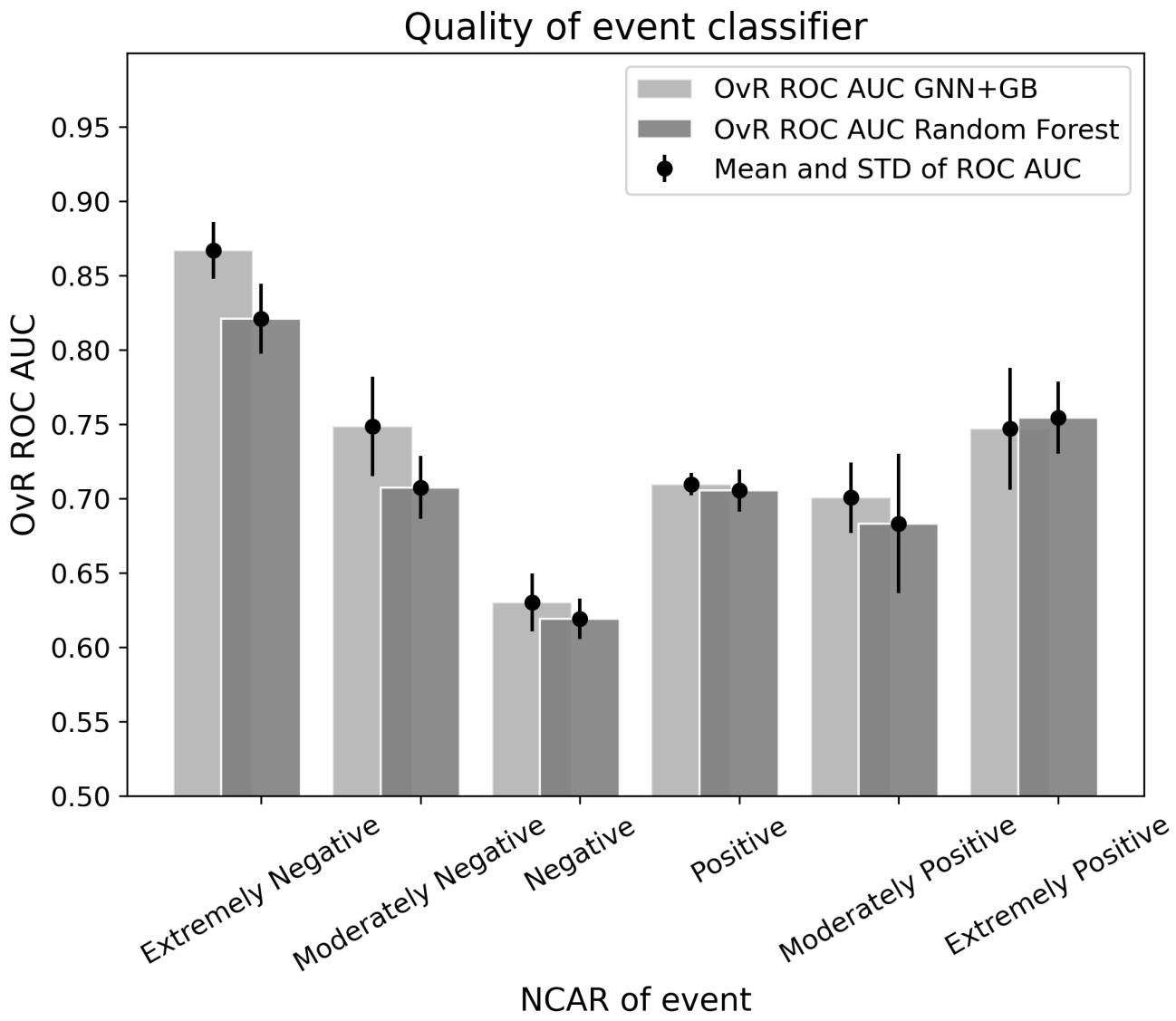
Class distribution



- Each class must be representative
- MAPE for expected return evaluation is 7%

Categorize stock price change in 6 classes

Classification results



Classification results

Class name*	Extremely Negative	Moderately Negative	Negative	Positive	Moderately Positive	Extremely Positive
Stock price change range	**(− ∞, − 0.28]**	**(− 0.28, − 0.14]**	**(− 0.14, 0]**	**(0, 0.14]**	**(0.14, 0.28]**	**(0.28, + ∞)**
Number of events	211	189	599	478	110	67
Positive events**	72	106	421	366	83	57
Negative events**	139	83	178	112	27	10
OvR ROC AUC for GCN+GB	0.87 ± 0.02	0.77 ± 0.03	0.63 ± 0.02	0.71 ± 0.01	0.70 ± 0.02	0.75 ± 0.04
OvR ROC AUC for GB	0.85 ± 0.02	0.72 ± 0.02	0.60 ± 0.02	0.67 ± 0.02	0.66 ± 0.04	0.74 ± 0.05
Welch's t-test p-value***	0.09	0.05	0.002	5.4×10^{-5}	0.02	0.65

Classification results

Class name*	Extremely Negative	Moderately Negative	Negative	Positive	Moderately Positive	Extremely Positive
Stock price change range	**(− ∞, − 0.28]**	**(− 0.28, − 0.14]**	**(− 0.14, 0]**	**(0, 0.14]**	**(0.14, 0.28]**	**(0.28, + ∞)**
Number of events	211	189	599	478	110	67
Positive events**	72	106	421	366	83	57
Negative events**	139	83	178	112	27	10
OvR ROC AUC for GCN+GB	0.87 ± 0.02	0.77 ± 0.03	0.63 ± 0.02	0.71 ± 0.01	0.70 ± 0.02	0.75 ± 0.04
OvR ROC AUC for GB	0.85 ± 0.02	0.72 ± 0.02	0.60 ± 0.02	0.67 ± 0.02	0.66 ± 0.04	0.74 ± 0.05
Welch's t-test p-value***	0.09	0.05	0.002	5.4×10^{-5}	0.02	0.65

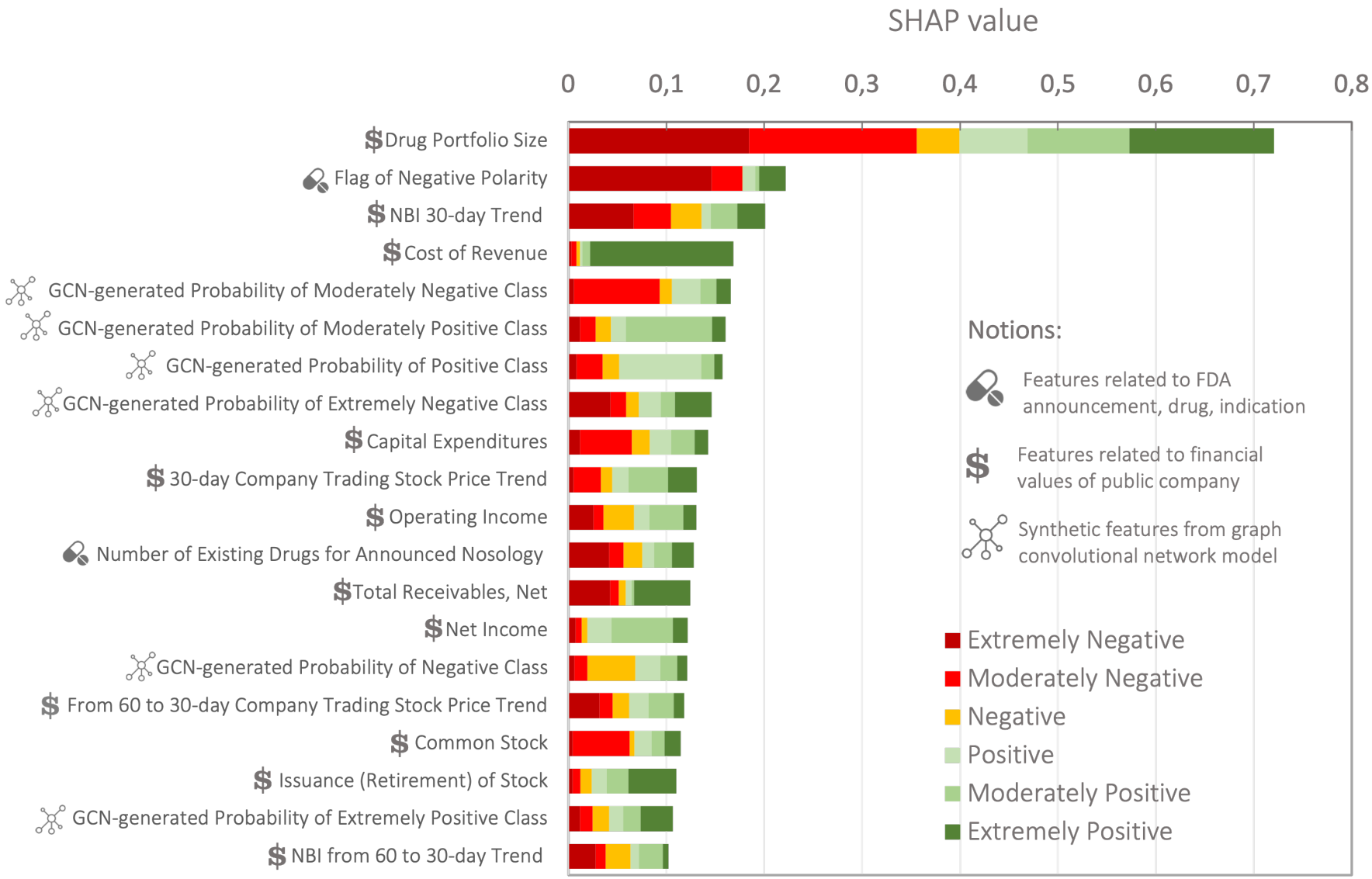
GCN-based features improve quality

Classification results

Class name*	Extremely Negative	Moderately Negative	Negative	Positive	Moderately Positive	Extremely Positive
Stock price change range	**$(-\infty, -0.28]$**	**$(-0.28, -0.14]$**	**$(-0.14, 0]$**	**$(0, 0.14]$**	**$(0.14, 0.28]$**	**$(0.28, +\infty)$**
Number of events	211	189	599	478	110	67
Positive events**	72	106	421	366	83	57
Negative events**	139	83	178	112	27	10
OvR ROC AUC for GCN+GB	0.87 ± 0.02	0.77 ± 0.03	0.63 ± 0.02	0.71 ± 0.01	0.70 ± 0.02	0.75 ± 0.04
OvR ROC AUC for GB	0.85 ± 0.02	0.72 ± 0.02	0.60 ± 0.02	0.67 ± 0.02	0.66 ± 0.04	0.74 ± 0.05
Welch's t-test p-value***	0.09	0.05	0.002	5.4×10^{-5}	0.02	0.65

Extremely Negative class is the easiest to predict, while Negative class is the hardest one

Feature importance analysis

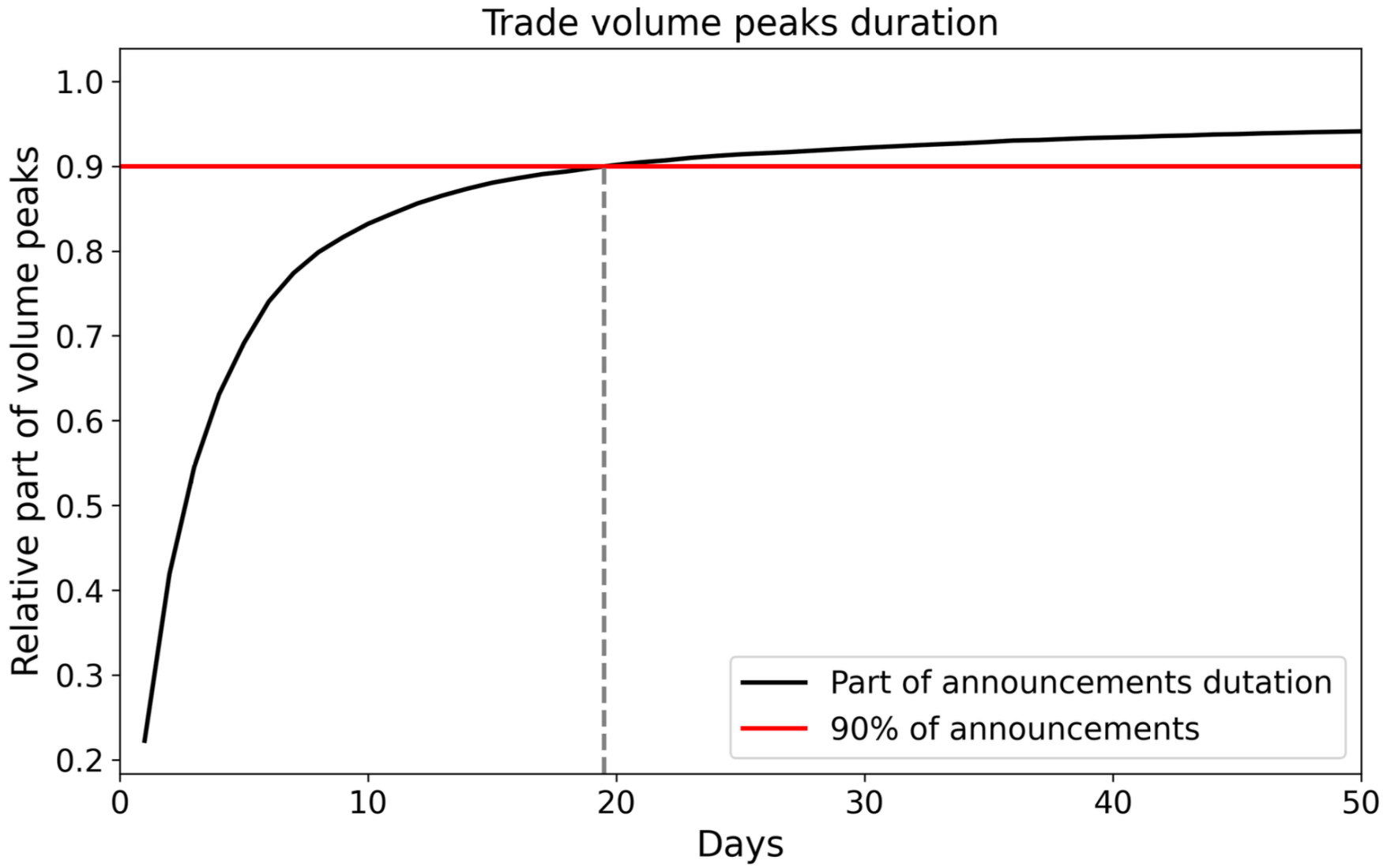


Take away

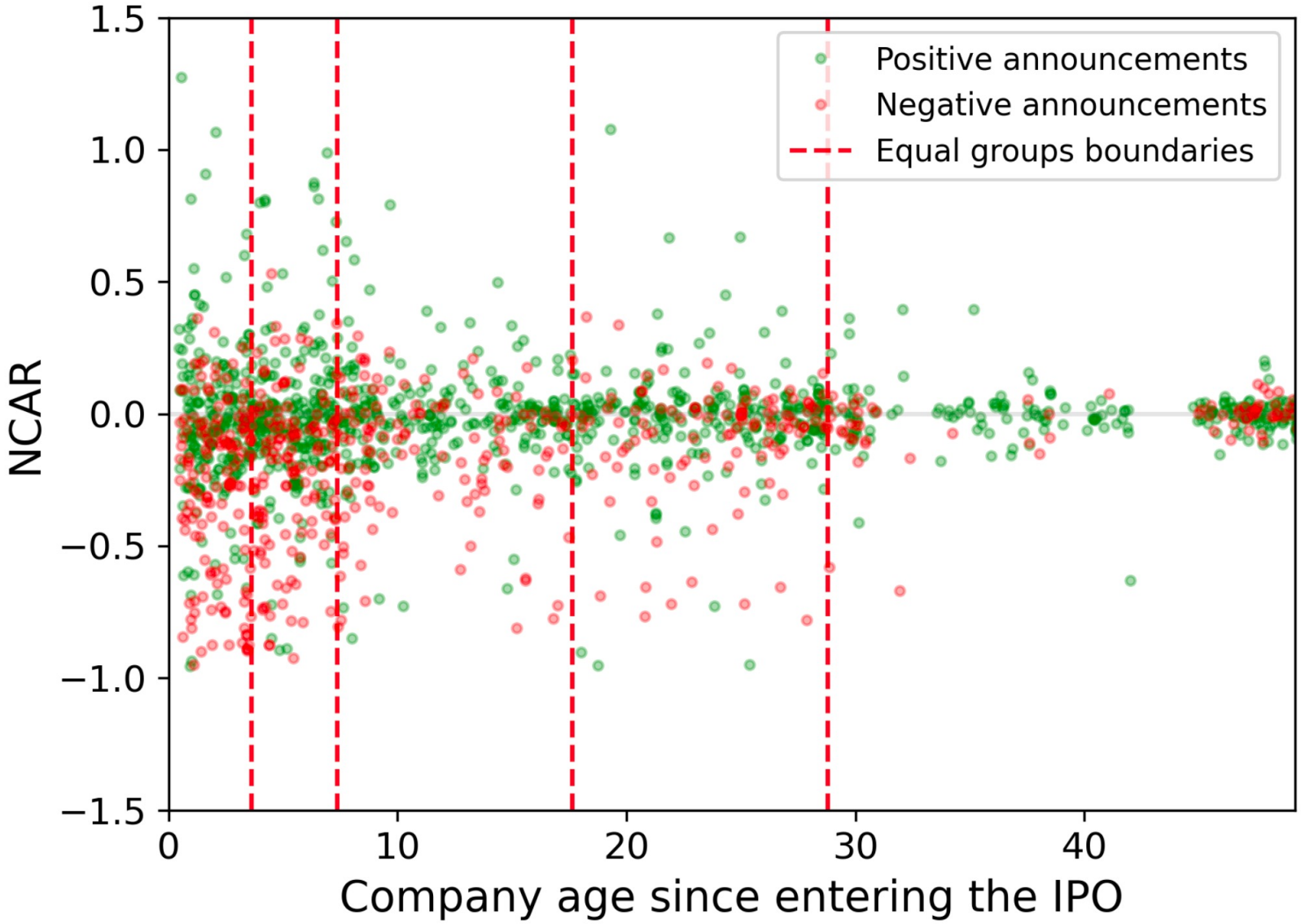
- Prove the existence of event impact on time series before solving the prediction task
- Look at a relationship between different characteristics and a target variable
- Generate comprehensive feature space

Appendix

Post-announcement period duration



Impact dependence on company age



Zero-shot Time Series Forecasting in Financial Data Analysis: Prospects and Challenges

Kostromina Alina
Sber AI Lab, MLTools
October 2024

СБЕР

Data Characteristics and Problem Statements

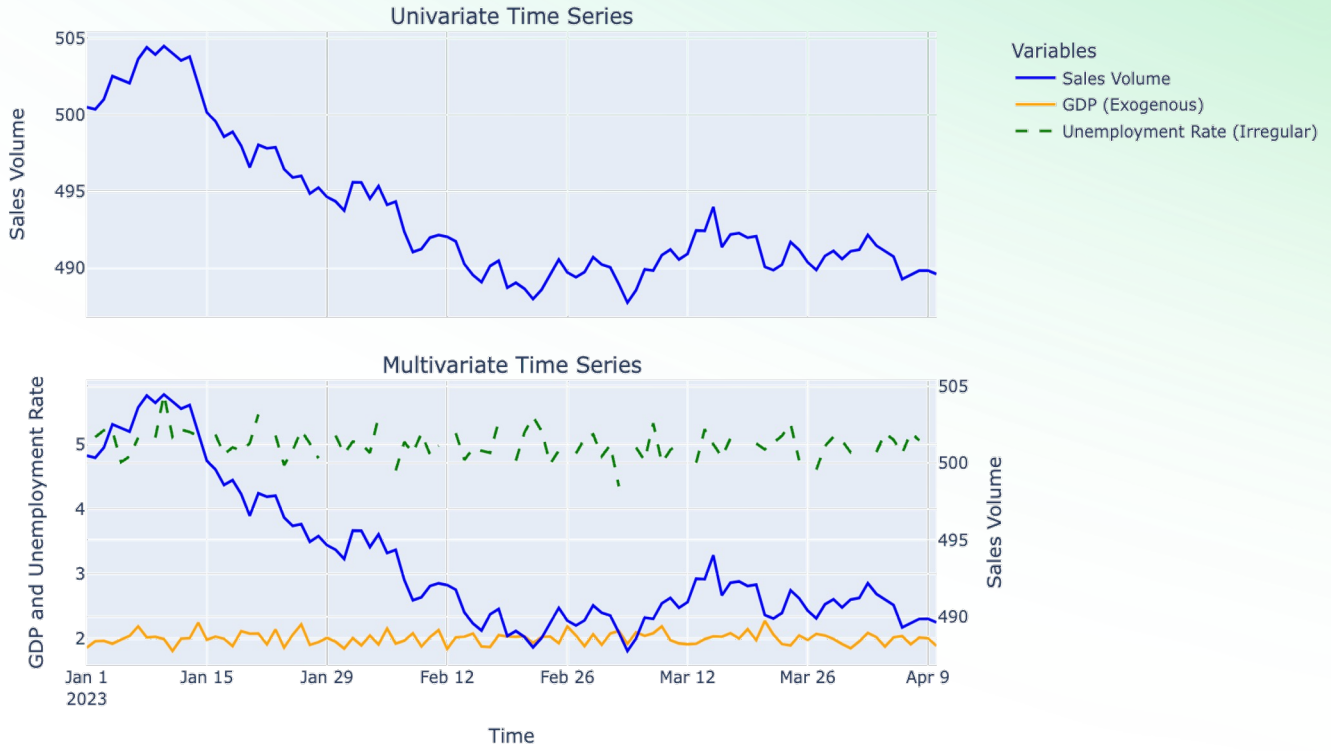
Forecasting instruments

Zero-shot models

Time Series Forecasting

Let’s define the terminology

- **A time series** is a sequence of values ordered by time.
- **A multivariate time series** is a structure where multiple individual time series are considered simultaneously**.**
- **A regular time series** is a time series with evenly spaced time intervals between data points.
- **Additional features (exogenous variables, covariates)** are external factors not generated by the system but influencing the target variables.

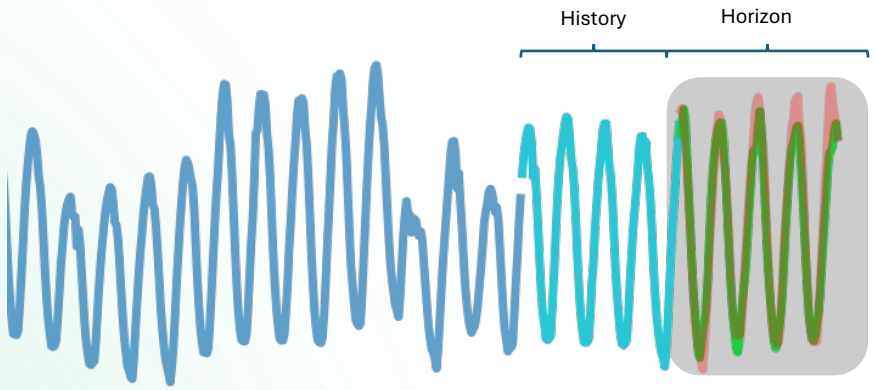


*The top subplot represents the **univariate** TS, while the bottom subplot — the **multivariate** TS. Both `Sales Volume` and `GDP` are **regular**, whereas the `Unemployment Rate` is **irregular**. `GDP` and `Unemployment Rate` are **exogenous features** relative to `Sales Volume`.*

Time Series Forecasting

Let's define the terminology

- **Forecasting Task:** given a history, the goal is to forecast over a horizon.



The history is not all the available data points, but rather the context length currently considered for generating predictions.

- **Forecasting Metrics** (between the prediction and the actual series):
  - Mean Squared Error (MSE)

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2$$

  - Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|$$

  - Mean Absolute Percentage Error (MAPE)

$$MAPE = 100 \times \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right|$$

  - Etc.

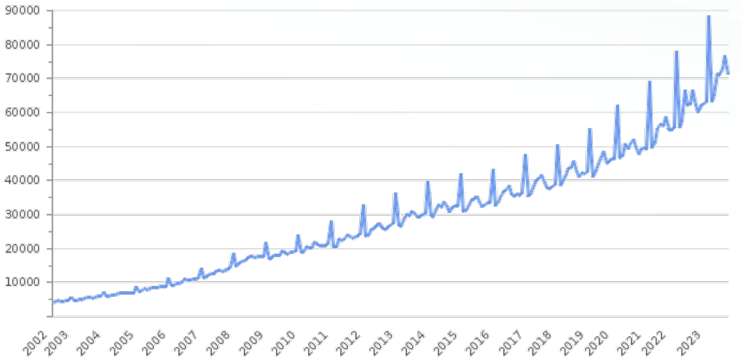
Characteristics of Financial Data and Tasks

which affect how they need to be analyzed

- We are not considering stock market data (**asset prices, market indices**, etc.).
- We are also not considering **irregular** time series.

Interest rate in Russia (%)

- We are focusing on macro- and microeconomic indicators, as well as financial data that reflect the activities of agents (such as **data on loans and deposits, labor market indicators**, and so on).



Average wage in Russia (Rub / Month)

Characteristics of Financial Data and Tasks

which affect how they need to be analyzed

What properties of the data do we face?

1. Short time series
2. Data Instabilities and External Disruptions
3. Numerous exogenous variables and contextual information

What do we expect from the model?

1. The model should work efficiently under data scarcity.
2. The model should be flexible and adapt to changing macro-level distributions.
3. The model should be able to select relevant factors and account for their dynamics.

СБЕР

Data Characteristics and Problem Statements

Forecasting instruments

Zero-shot models

How to generate a forecast?

Groups of methods commonly used in forecasting tasks

Naive methods

- (Seasonal) Naive
- Mean, Median

Statistical methods

- ETS
- Theta
- ARIMA

ML methods

- ElasticNet
- GBMs

DL methods

- DLinear
- NBEATS
- PatchTST
- GPT2

How these methods meet our requirements?

Methods Group	Data scarcity	Flexibility and adaptability	Exogenous variables
Naive methods	+	-	-
Statistical methods	+	-	+-
ML methods	+-	+-	+
DL methods	-	+	+

СБЕР

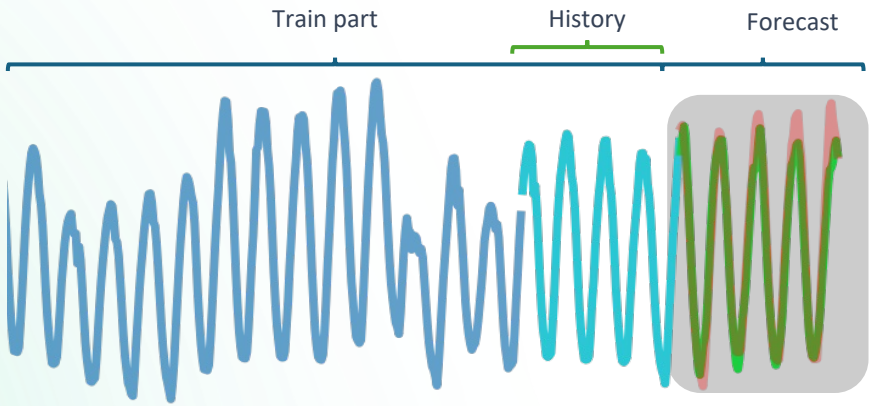
Data Characteristics and Problem Statements

Forecasting instruments

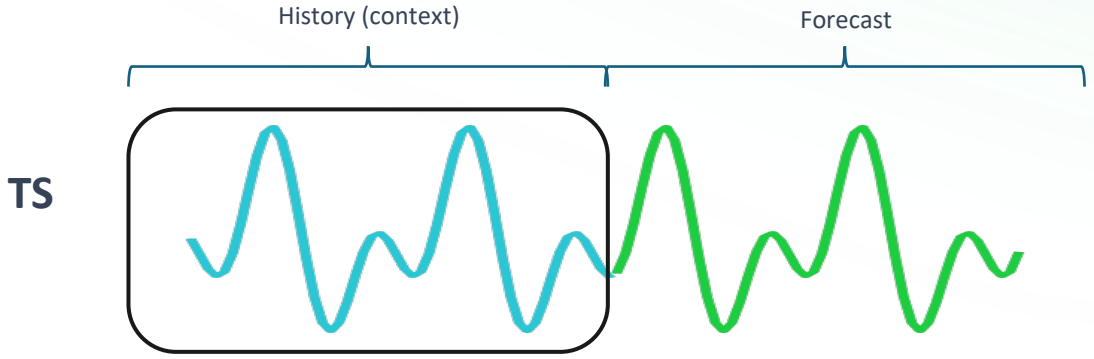
Zero-shot models

Supervised vs Zero-shot models

Supervised: Train the model to predict labels for new data based on patterns identified in the **training data**.



Zero-shot: Train the model to predict labels for new data **without training on the target dataset**, based on patterns identified in the unrelated data.



The variety of Zero-shot models

Zero-shot models for time series are an actively developing area

LLM

Non-adapted LLM

- LSTPrompt
- PromptCast
- LLMTime

Adapted LLM

- Time-LLM
- FPT
- Chronos
- UniTS
- DAM

Specialized

Using synthetic data

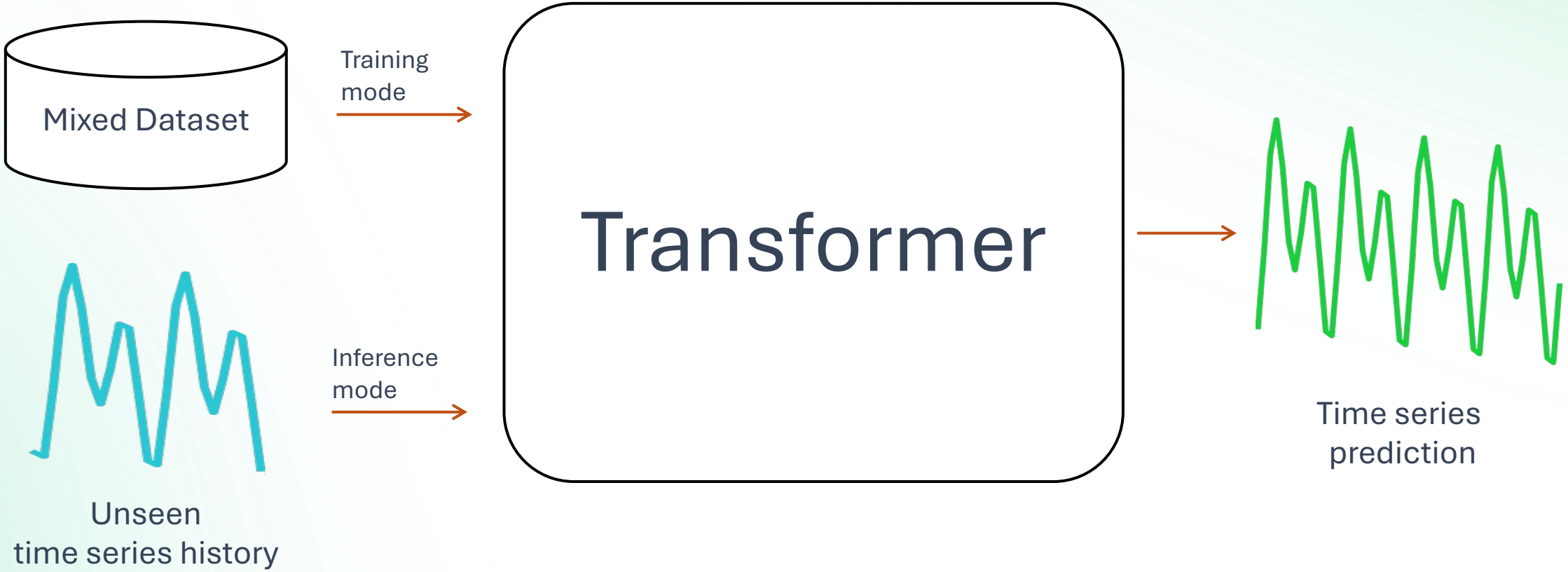
- ForecastPFN

Using real data

- GPHT
- MOIRAI
- Moment

Zero-shot models

- Zero-shot models for time series are mainly Transformers
- They require a large and diverse dataset for pretraining



Kuvshinova, K., Tsymboi, O., Kostromina, A., Simakov, D., & Kovtun, E. (2024). Towards Foundation Time Series Model: To Synthesize Or Not To Synthesize?. *arXiv preprint arXiv:2403.02534*.

Zero-shot for business case

We can adapt the training data properties to a specific task.

- In this case, we are forecasting **macroeconomic** time series
- Dozens to hundreds of **short** time series, which are mostly **independent**.

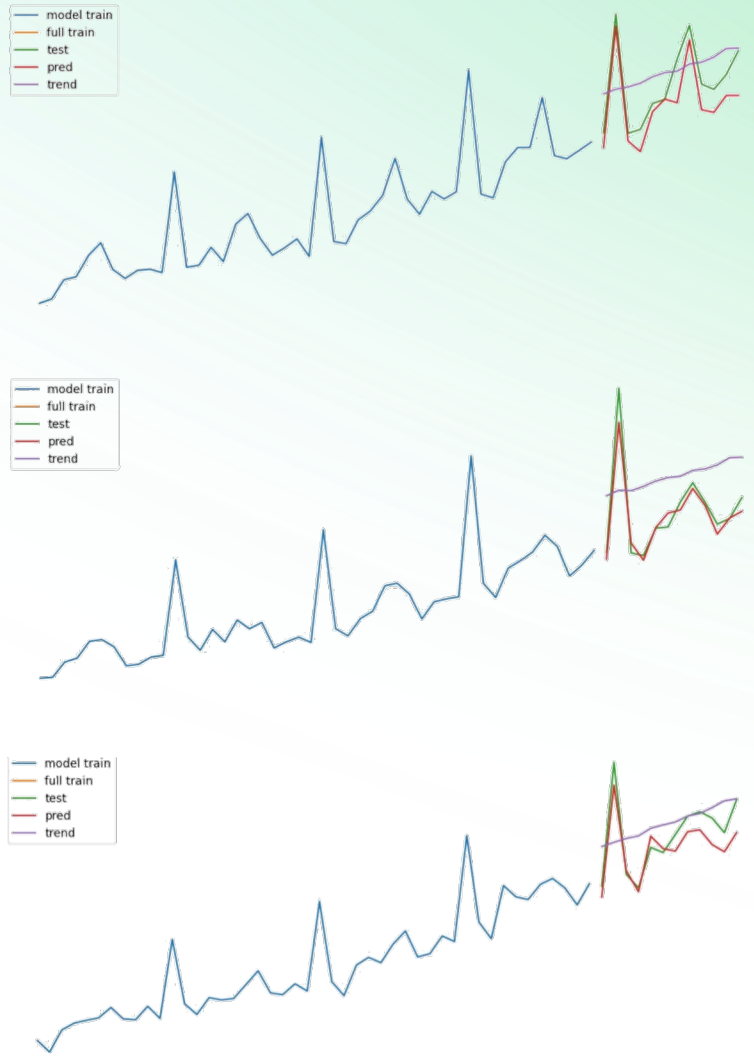
MAPE:

- Zero-shot – x%
- GBDT – **↑0.7%**
- Prophet x~100 – **↓0.2%**

Time (train + inference):

- Zero-shot – **< 1sec**
- GBDT – **30 sec**
- Prophet x~100 – **5 min**

Great quality with a speed advantage!



Predictions of a Zero-shot model trained not only to forecast the continuation of the time series but also to decompose it into trend and seasonality.

Zero-shot with exogenous data

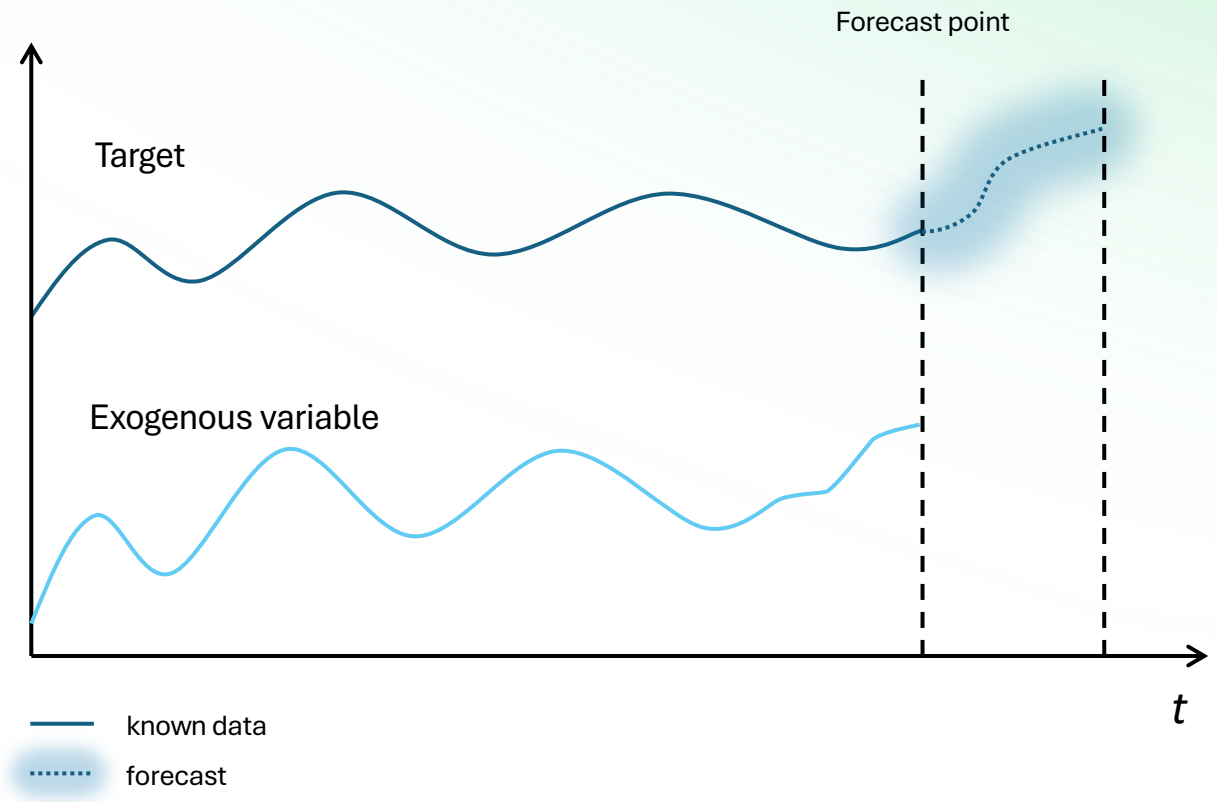
The functionality of Zero-shot model can be extended to handle additional information.

Zero-shot with additional features is an unexplored scientific topic.

Possible solution:

- There are additional time series. The task is to transform these series into a forecast.
- Use TabPFN (Tabular Prior-Data Fitted Network).

- There is no open-source TabPFN for regression tasks yet, but there is a paper from the NeurIPS 2024 Workshop*.



* Hoo, S. B., Müller, S., Salinas, D., & Hutter, F. The Tabular Foundation Model TabPFN Outperforms Specialized Time Series Forecasting Models Based on Simple Features. In *NeurIPS 2024 Third Table Representation Learning Workshop*.

How does Zero-shot fit the requirements from data?

Let’s return to our table

Methods Group	Data scarcity	Flexibility and adaptability	Exogenous variables
Naive methods	+	-	-
Statistical methods	+	-	+-
ML methods	+-	+-	+
DL methods	-	+	+
Zero-shot	+	+	+

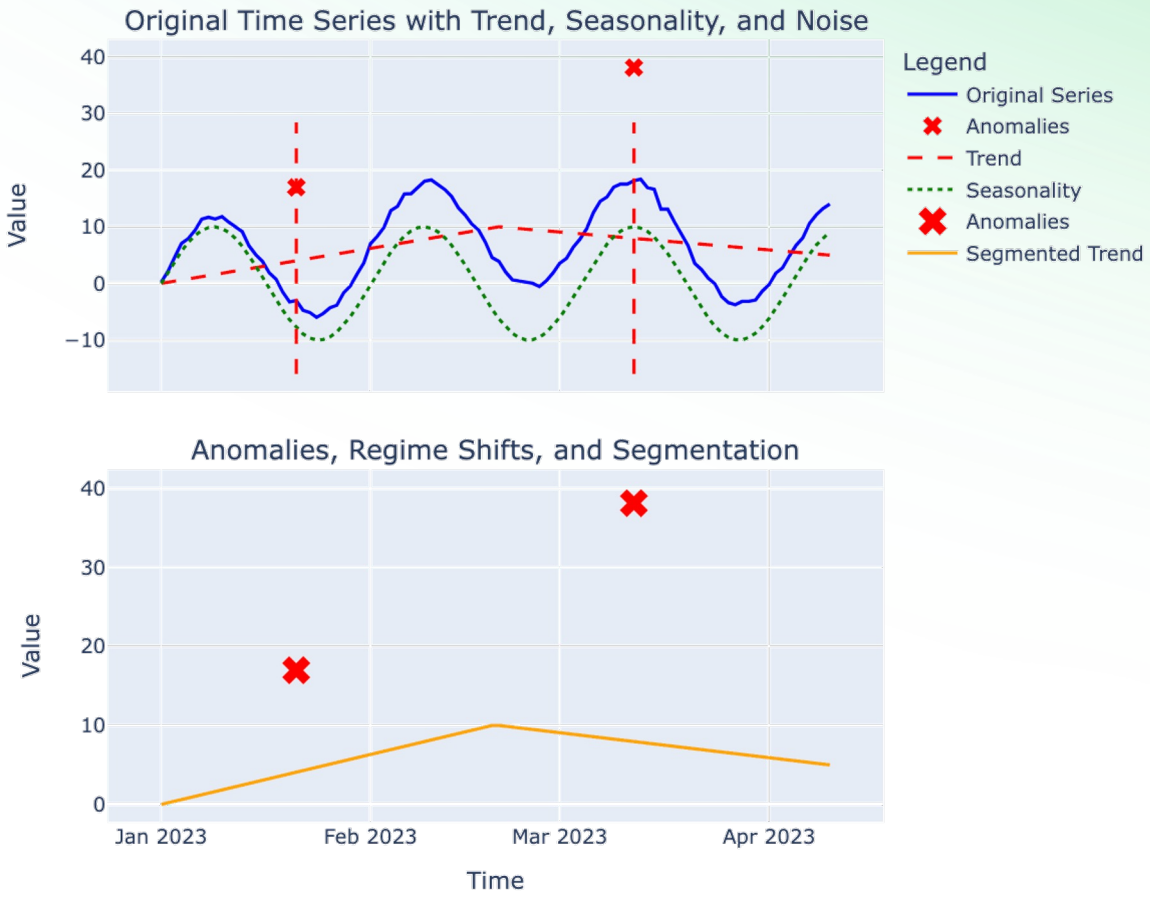
Possible directions for the development and challenges to overcome

1. Development of Multitasking:

- **Decomposition** into trend, seasonality, and noise, with tasks such as:
  - removal of trend, seasonality, or noise
  - **segmentation** based on trend
  - predicting the **trend type**
  - forecasting Fourier coefficients; **seasonality** period and **type**
- Detection of **anomalies** and **regime shifts**
- Generative tasks such as predicting stochastic differential equation **(SDE) parameters, interval forecasting, and filling in missing values**

2. Challenges:

- Integrating forecasting and regression tasks in one model
- Developing effective methods for generating synthetic data
- Finding efficient approaches to utilize exogenous features



Representation of different tasks: decomposition into trend, seasonality, and noise, along with detection of anomalies and regime shifts.

Thank you for your attention!

Kostromina Alina (tg: @elineii)

Sber AI Lab, MLTools