

Science Beyond Human Scale: Generative, Agentic, and Interpretable AI for Scientific Discovery

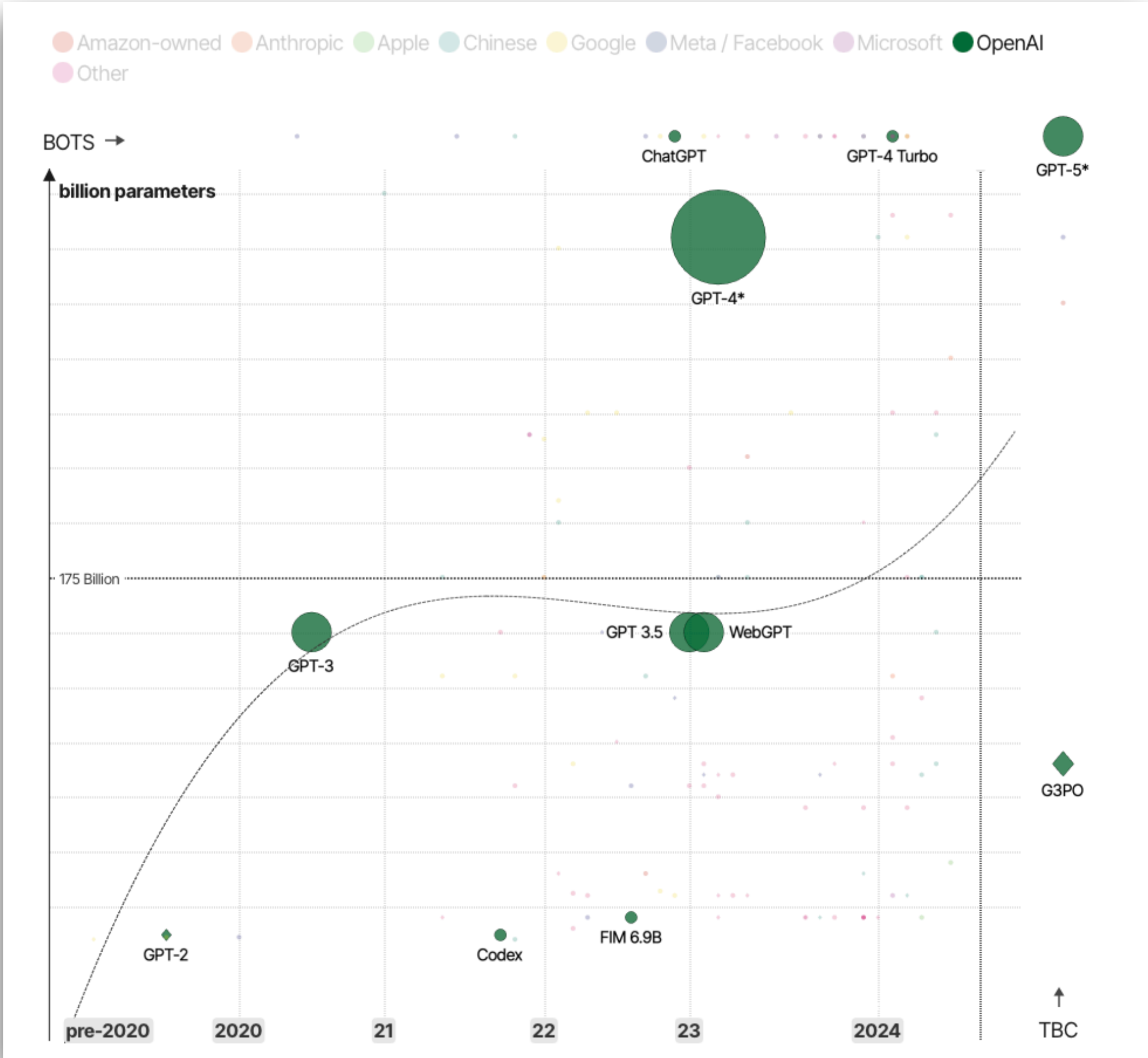


Daniel Saeedi

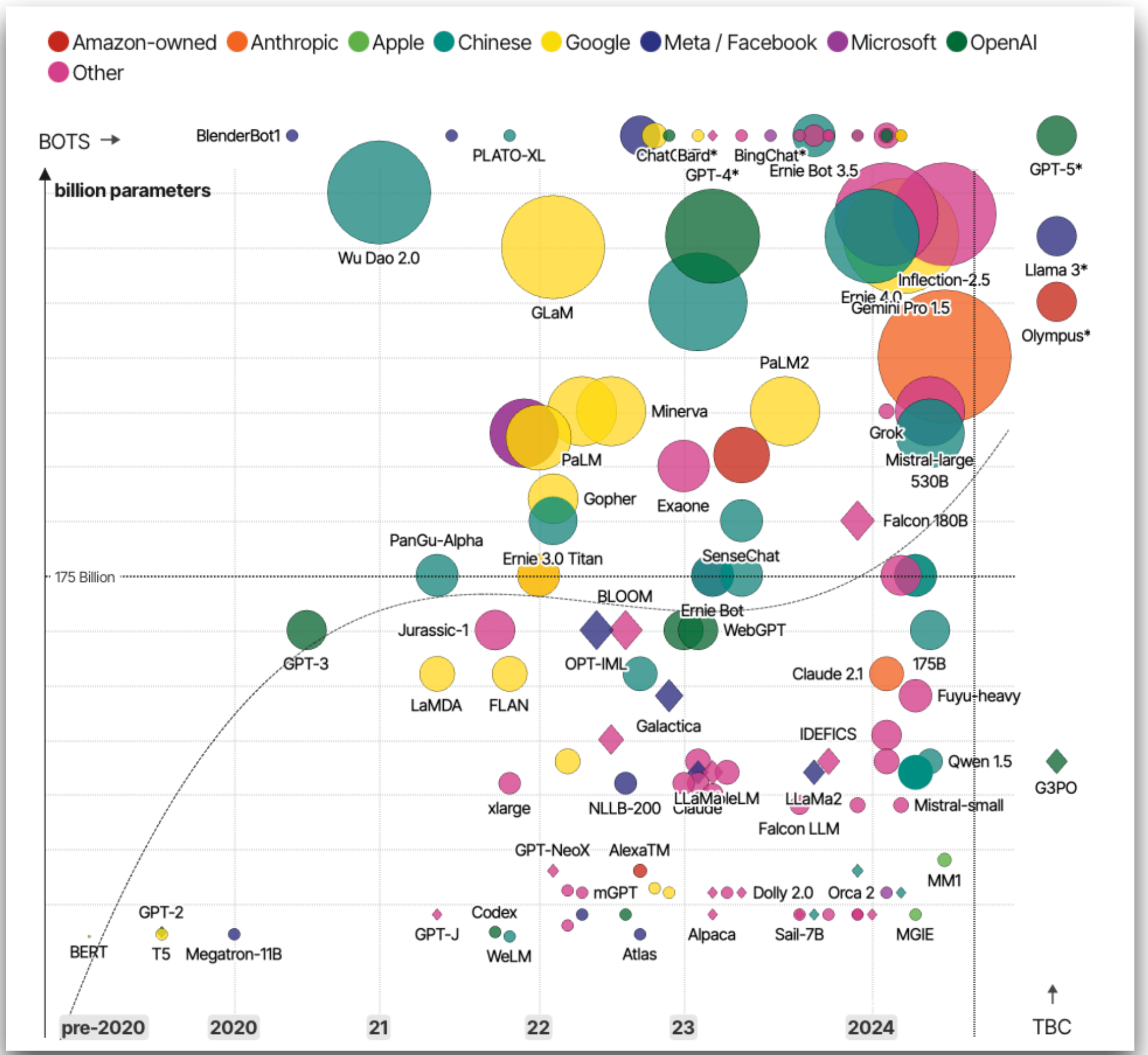
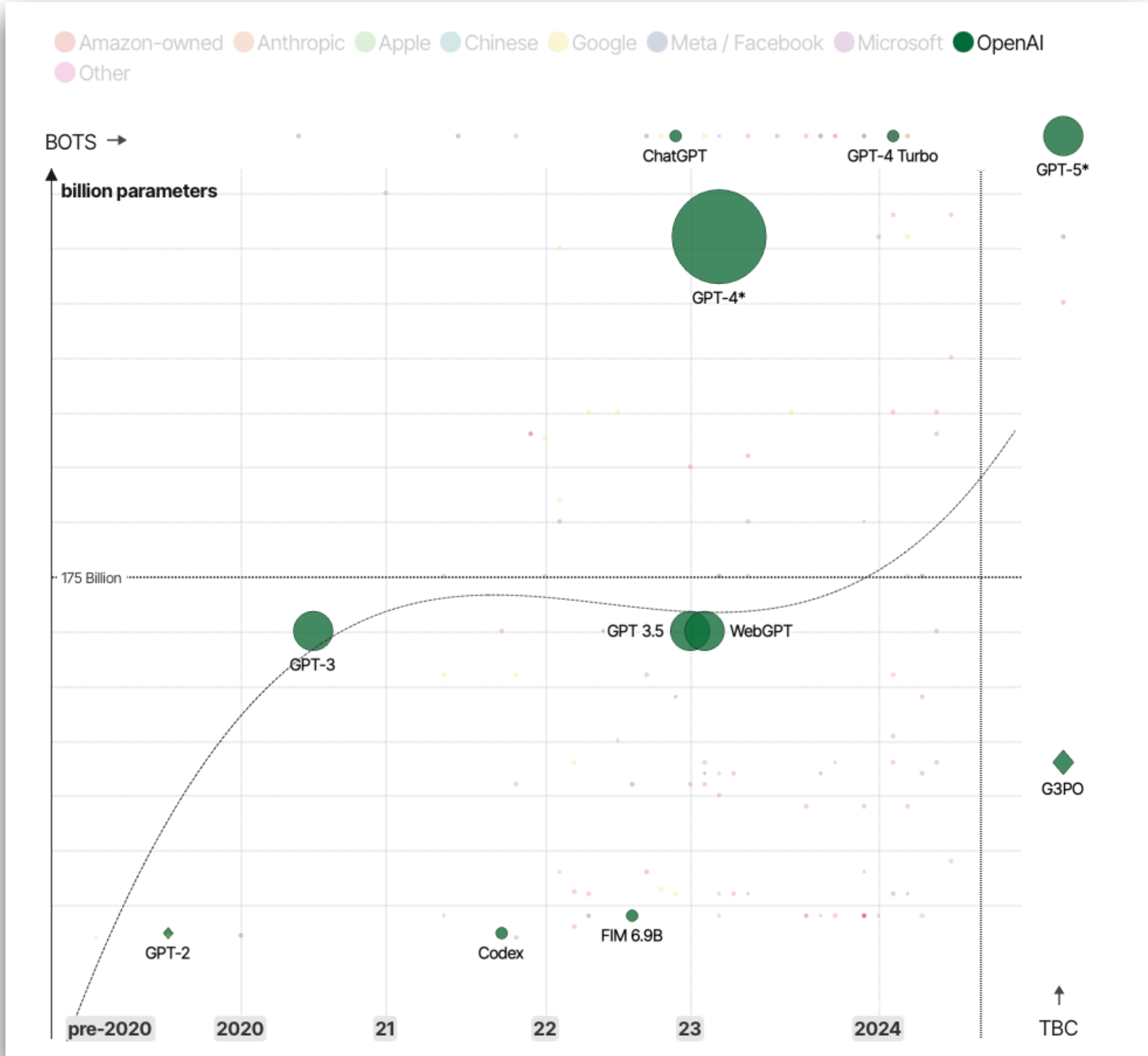
PhD Student



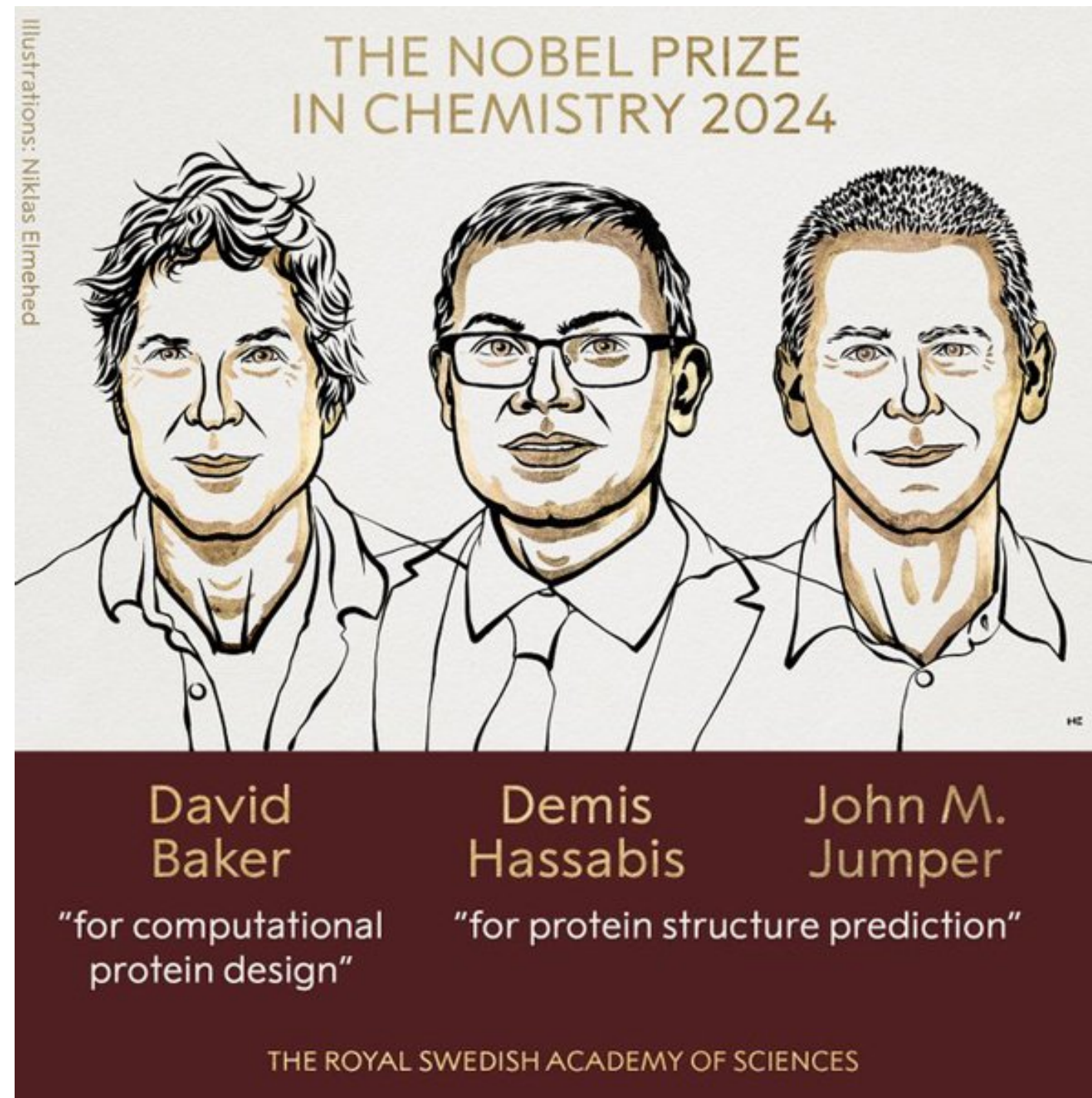
the rise and rise of AI



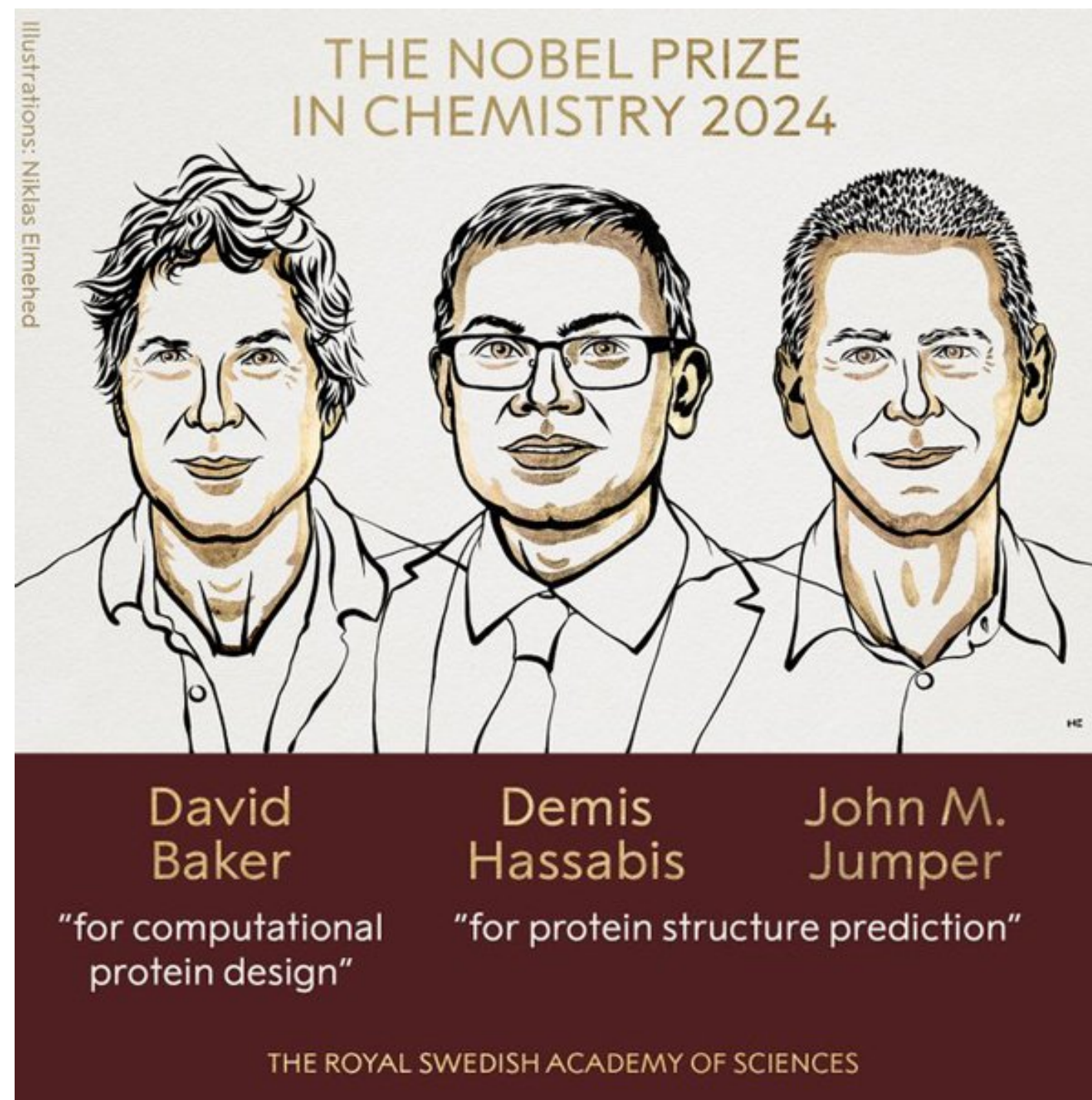
the rise and rise of AI



AI breakthroughs in science



AI breakthroughs in science



JUNE 25, 2024//ANNOUNCEMENT

ESM3: Simulating 500 million years of evolution with a language model

[Preview our paper](#) ➤

Trained on **2.8 billion** proteins
96 billion parameters

AI breakthroughs in science

Introducing ProGen3

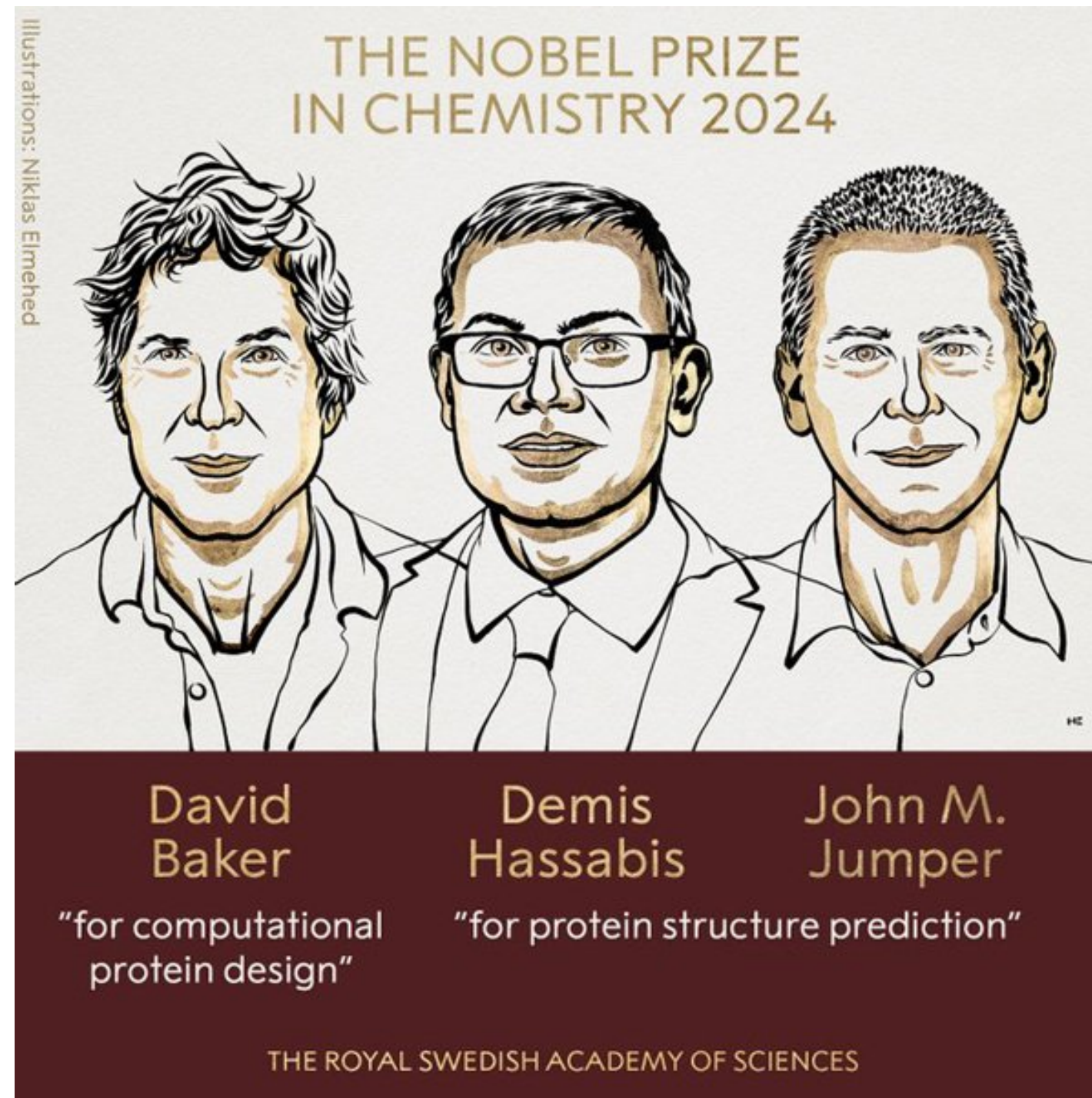
We're announcing ProGen3, our next foundation models for protein generation. Partners can license our molecule assets or sign up for early access to our API.

April 16, 2025

[Get early access to our API →](#)

[License our molecules →](#)

Trained on **3.4 billion** proteins
46 billion parameters



JUNE 25, 2024//ANNOUNCEMENT

ESM3: Simulating 500 million years of evolution with a language model

[Preview our paper ↗](#)

Trained on **2.8 billion** proteins
96 billion parameters

AI breakthroughs in science

Introducing ProGen3

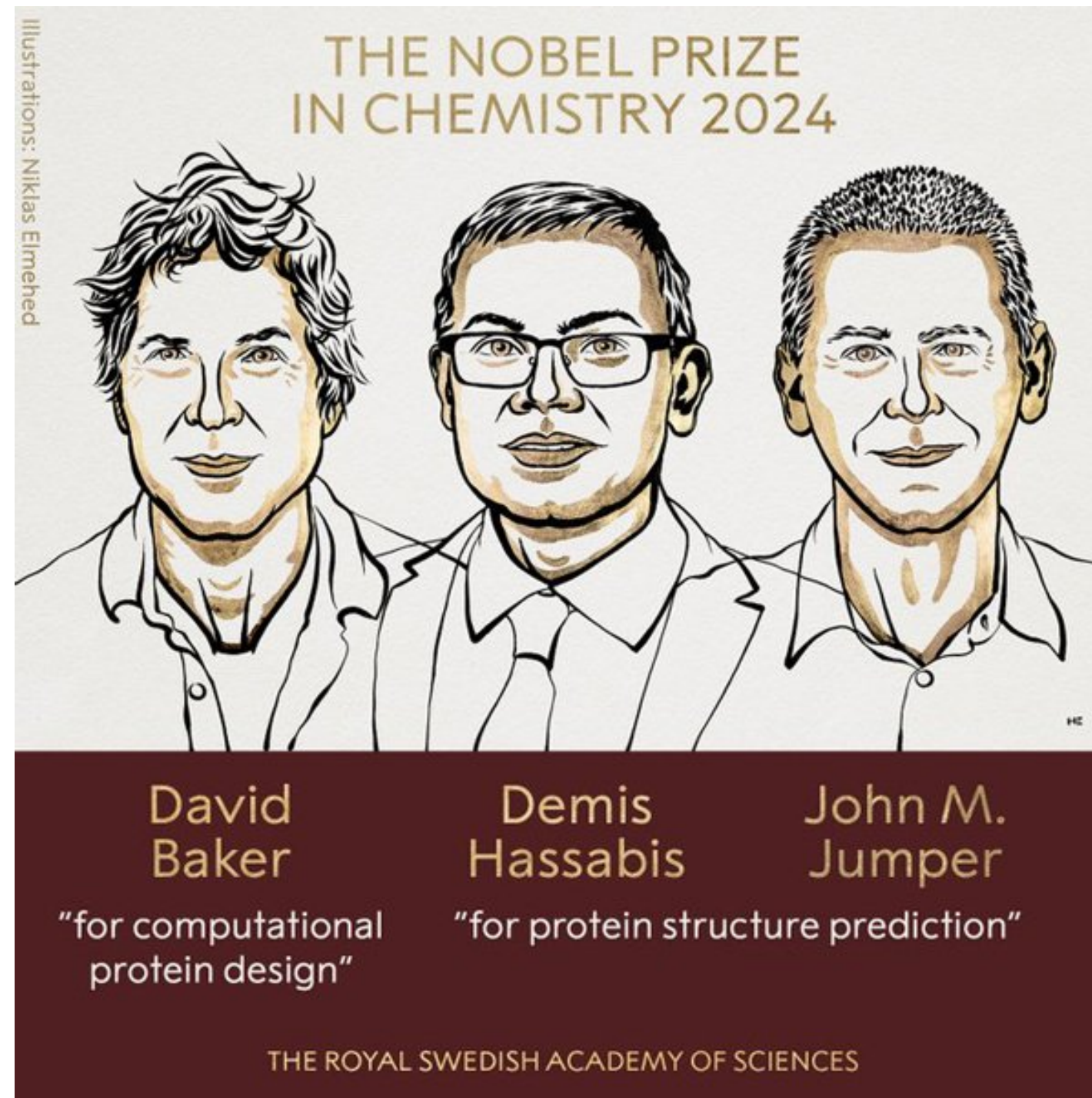
We're announcing ProGen3, our next foundation models for protein generation. Partners can license our molecule assets or sign up for early access to our API.

April 16, 2025

[Get early access to our API →](#)

[License our molecules →](#)

Trained on **3.4 billion** proteins
46 billion parameters



JUNE 25, 2024//ANNOUNCEMENT

ESM3: Simulating 500 million years of evolution with a language model

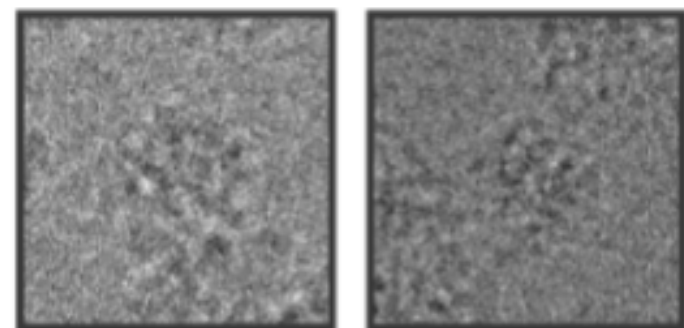
[Preview our paper ↗](#)

Trained on **2.8 billion** proteins
96 billion parameters

But does our power to **explain** them also **scale**?

Every stage of scientific discovery has outgrown human scale

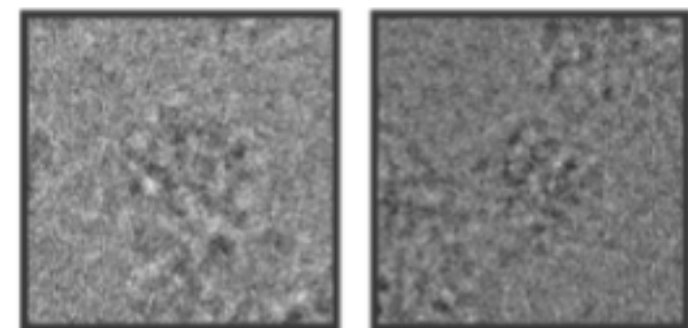
Cryo-electron microscope



0.5 - 5 TB per dataset

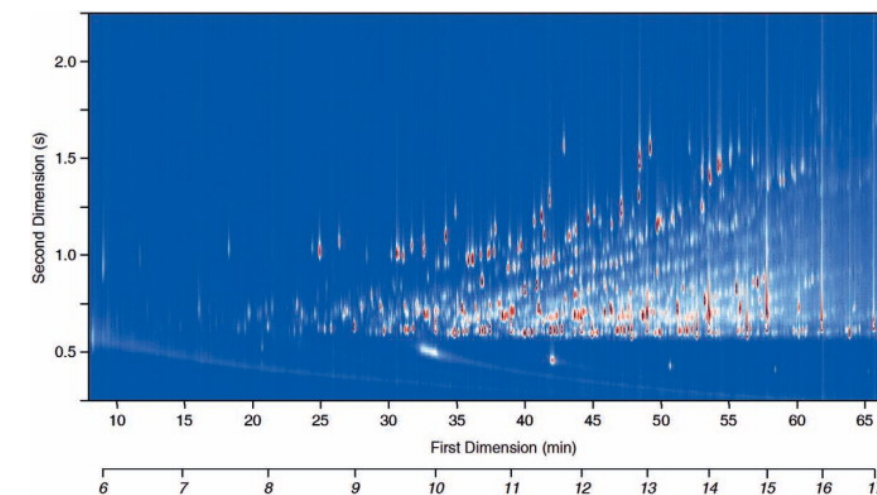
Every stage of scientific discovery has outgrown human scale

Cryo-electron microscope



0.5 - 5 TB per dataset

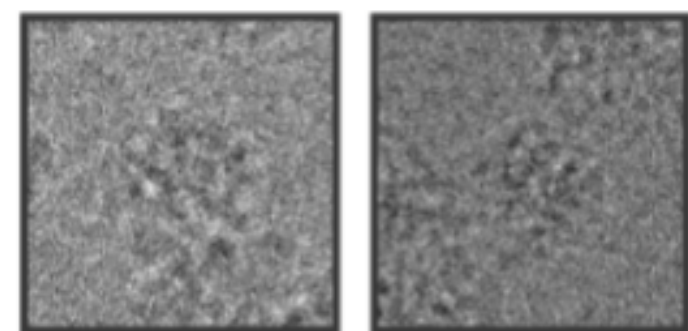
2D gas chromatography



10 - 200 GB per sample

Every stage of scientific discovery has outgrown human scale

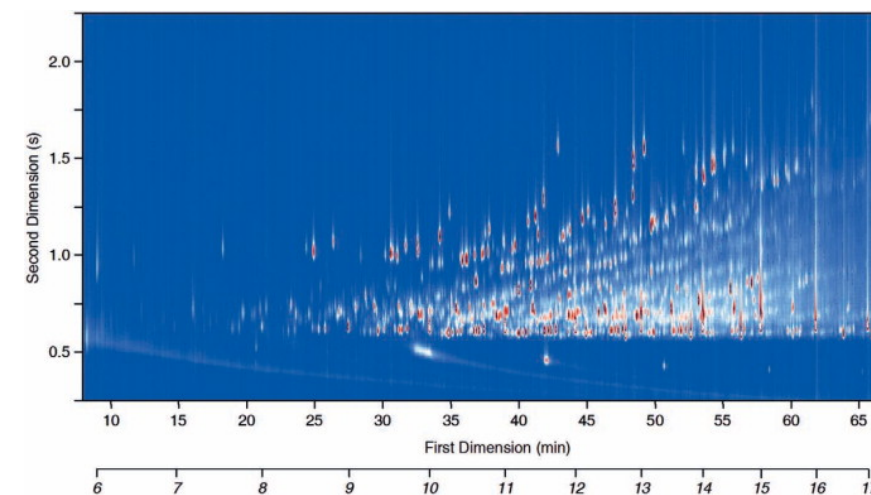
Cryo-electron microscope



0.5 - 5 TB per dataset

<https://www.bnl.gov/cryo-em/instrumentation.php>

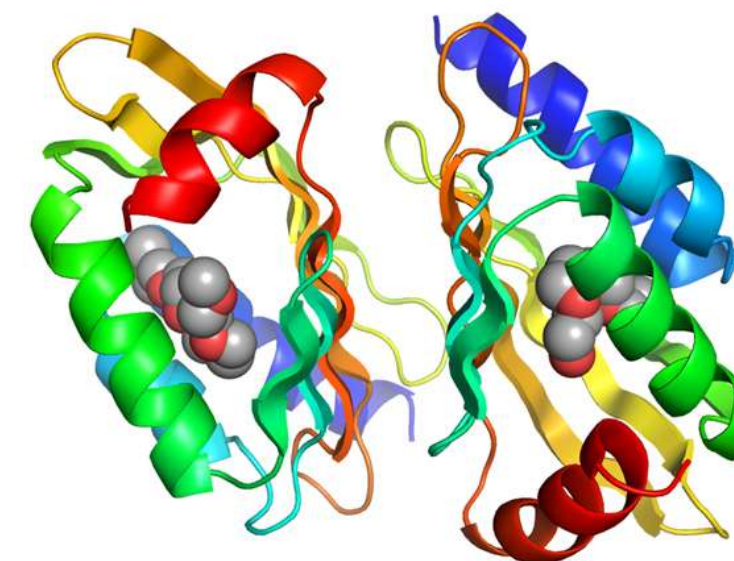
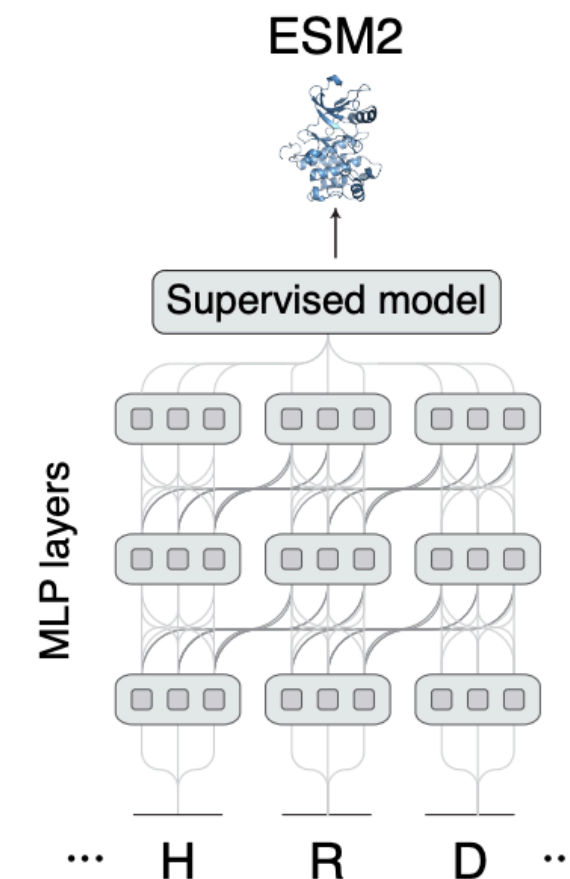
2D gas chromatography



10 - 200 GB per sample

<https://chemistry-matters.com/chemicals/comprehensive-two-dimensional-gas-chromatography-gcxcg/>

Interpreting pLMs

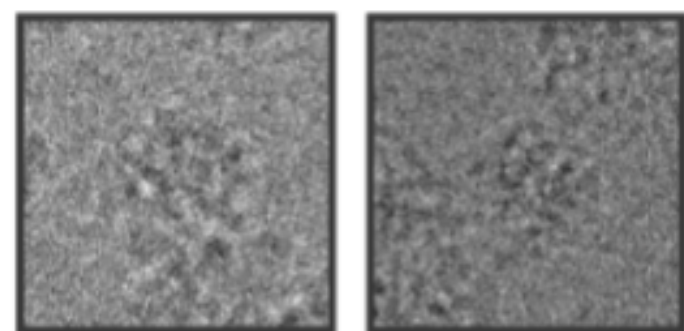


0.1 - 100 B parameters

<https://www.sciencedirect.com/topics/materials-science/two-dimensional-gas-chromatography>

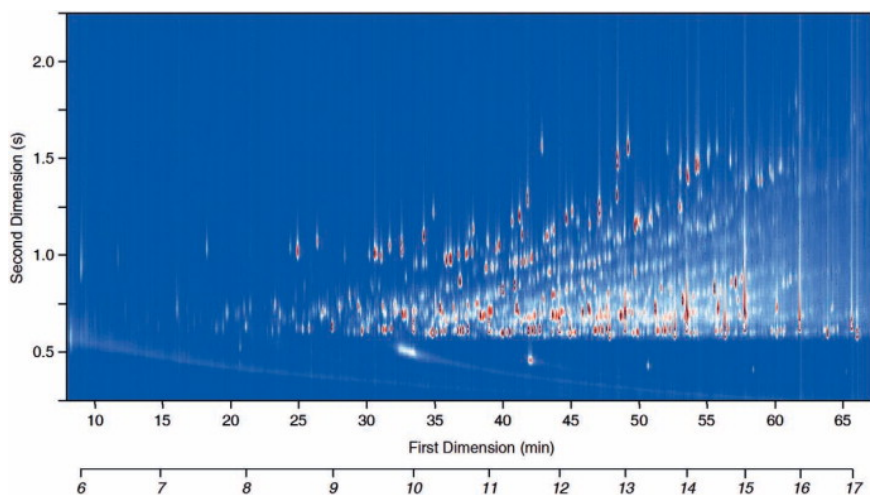
Every stage of scientific discovery has outgrown human scale

Cryo-electron microscope



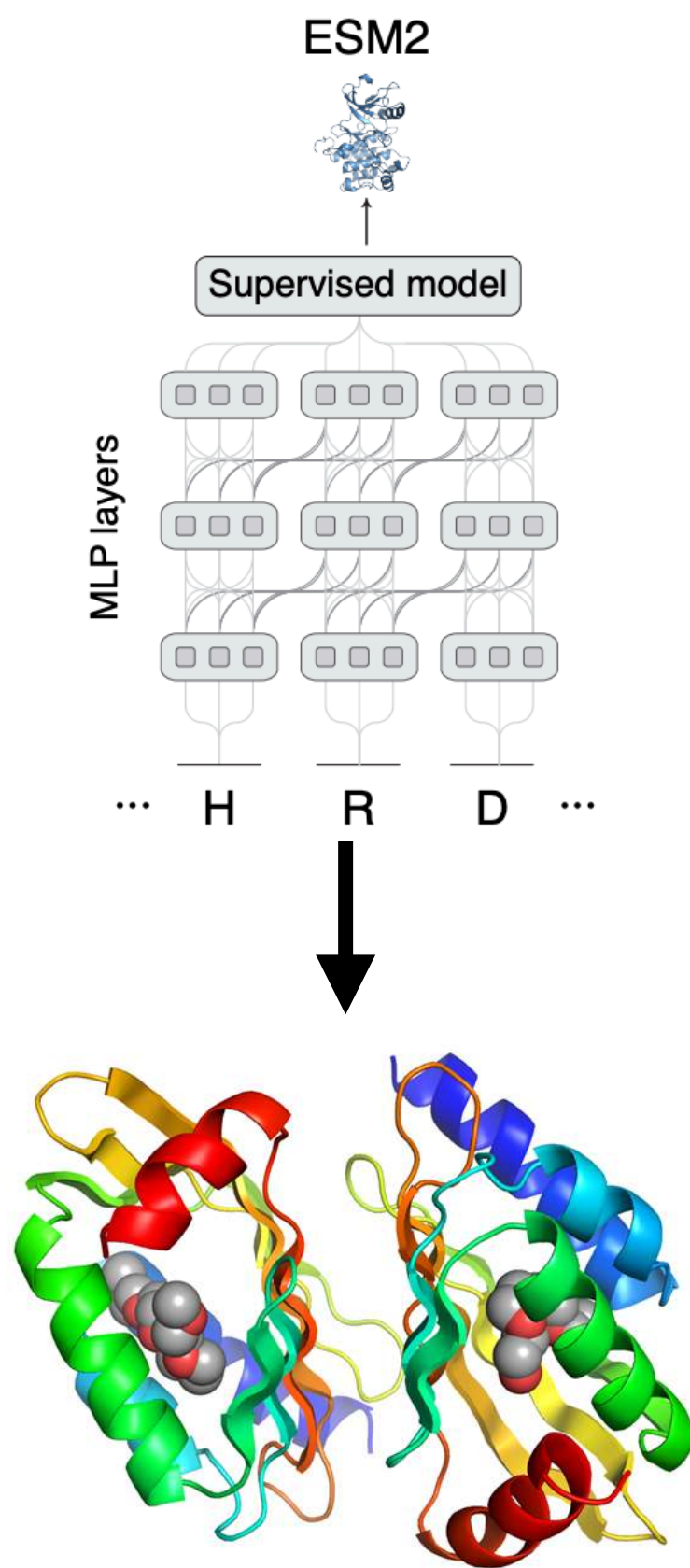
0.5 - 5 TB per dataset

2D gas chromatography



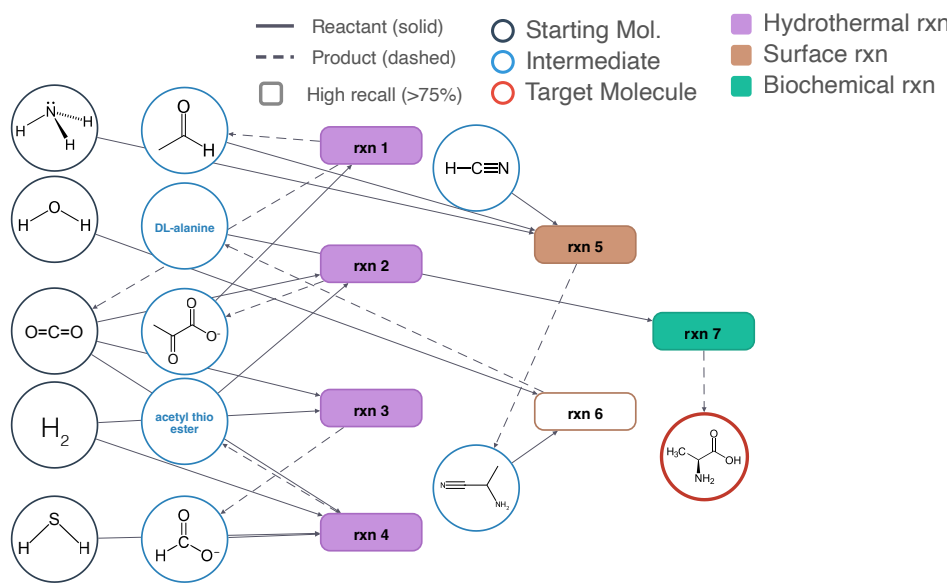
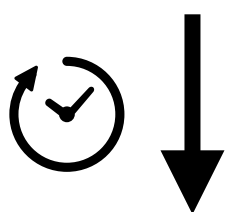
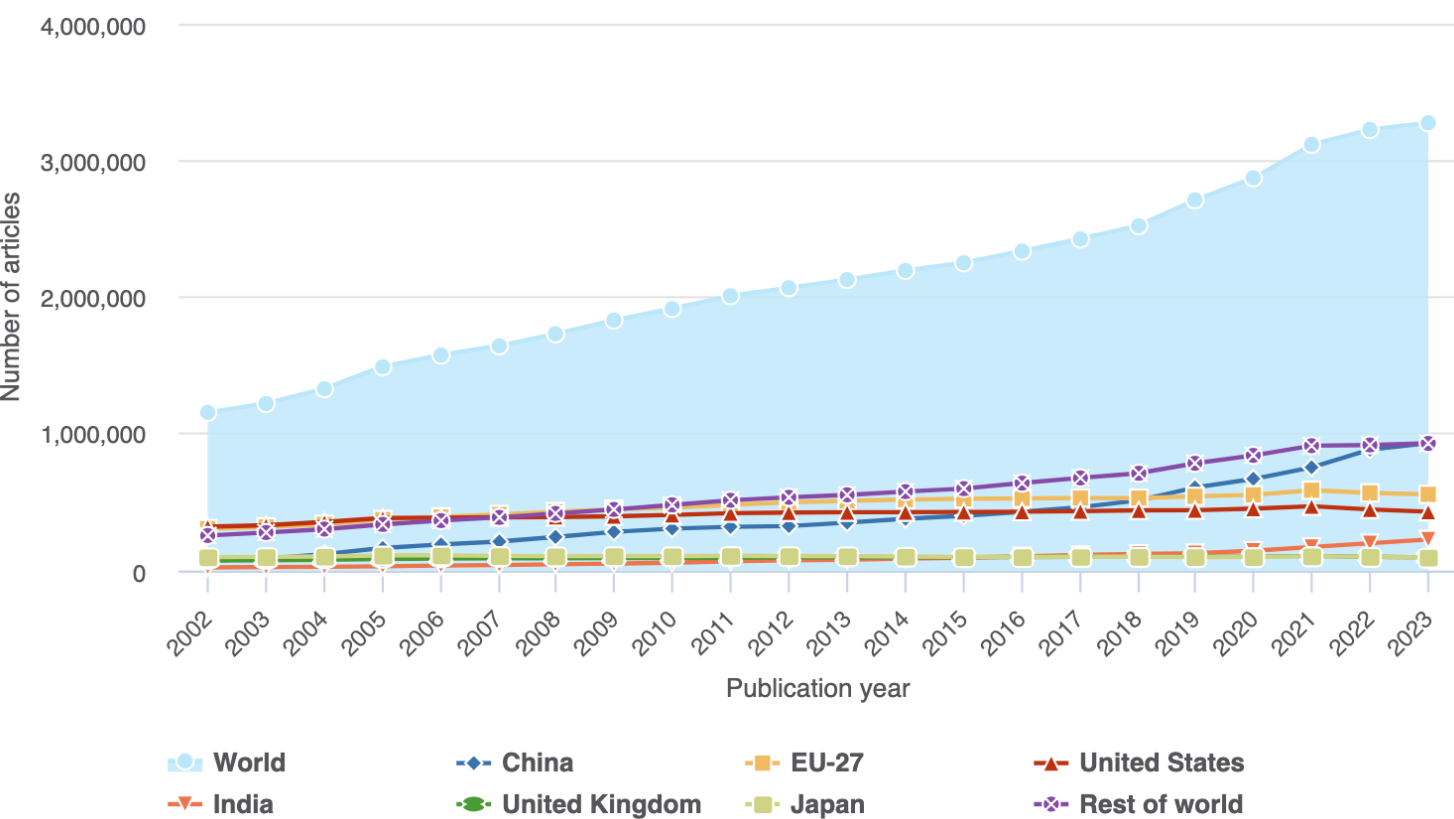
10 - 200 GB per sample

Interpreting pLMs



0.1 - 100 B parameters

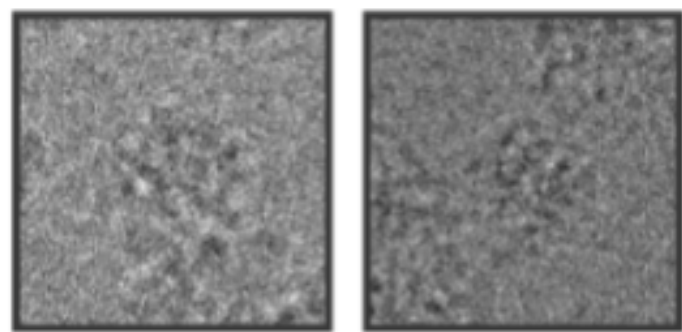
Scientific Hypothesis Generation



3 million new papers per year

High ambient dimension, low intrinsic dimension

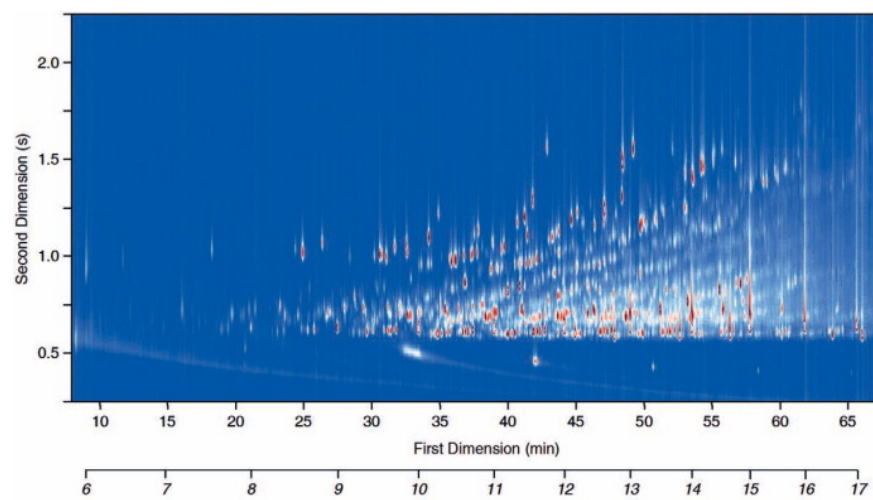
Cryo-electron microscope



0.5 - 5 TB per dataset

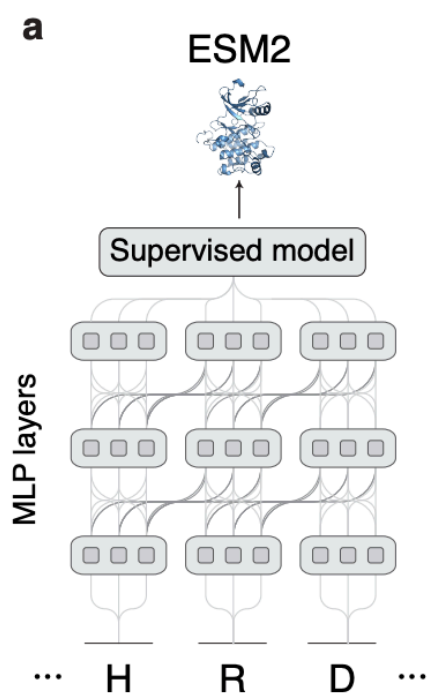
**~5–10 DOF: pose
+ conformation
Low-dim manifold**

2D gas chromatography



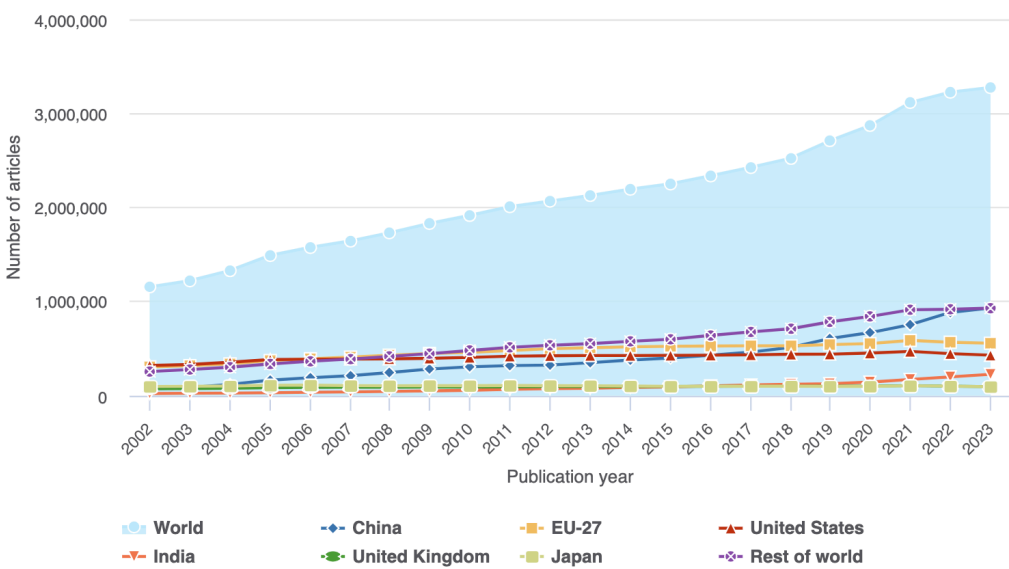
10 - 200 GB per sample

Interpreting pLMs



0.1 - 100 B parameters

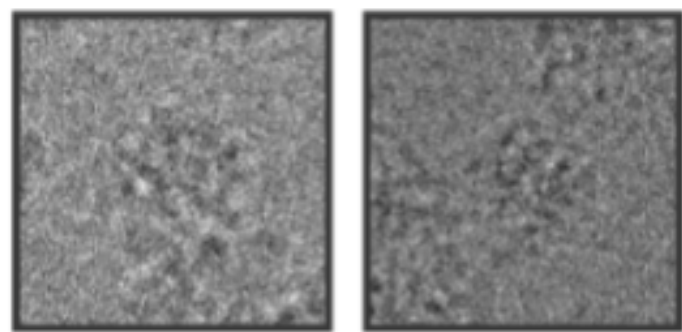
Scientific Hypothesis Generation



3 million new papers per year

High ambient dimension, low intrinsic dimension

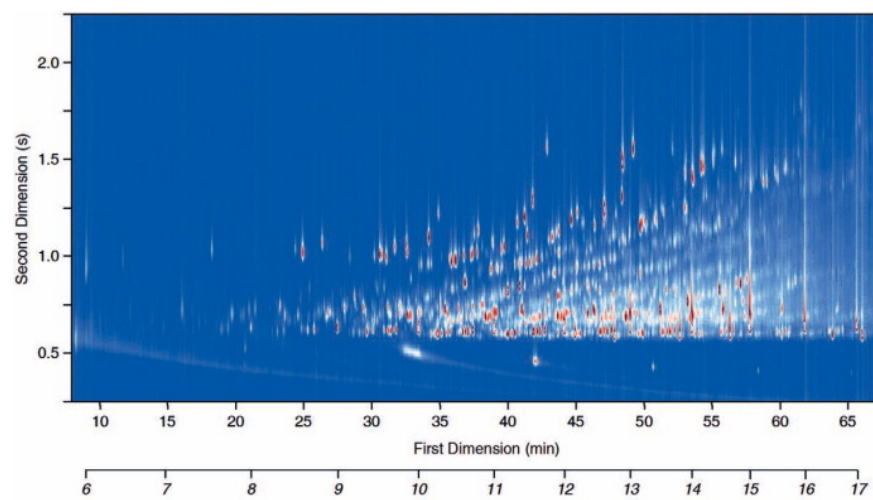
Cryo-electron microscope



0.5 - 5 TB per dataset

**~5–10 DOF: pose
+ conformation
Low-dim manifold**

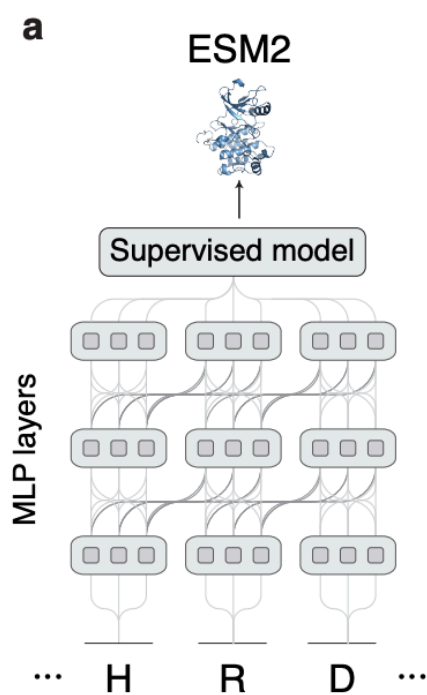
2D gas chromatography



10 - 200 GB per sample

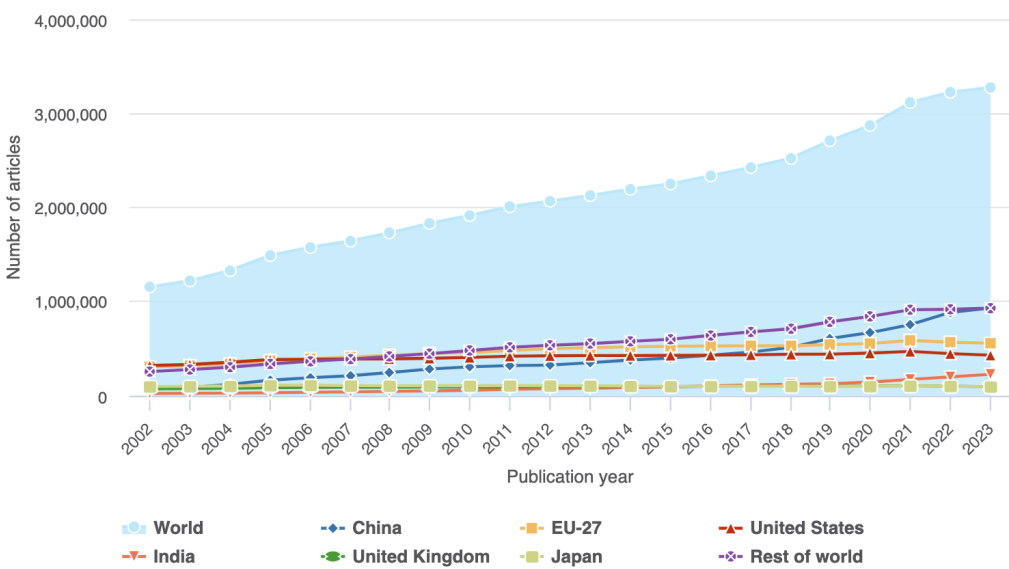
**Sparsity of signal; a
handful discriminate**

Interpreting pLMs



0.1 - 100 B parameters

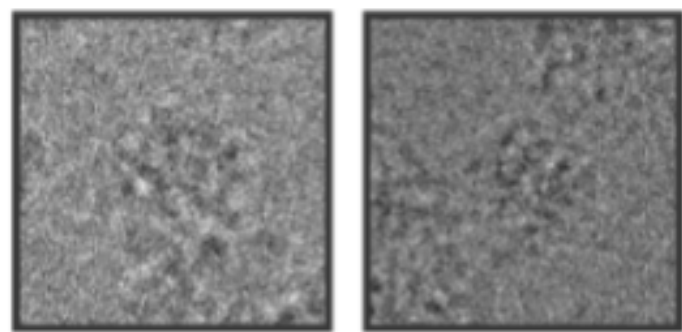
Scientific Hypothesis Generation



3 million new papers per year

High ambient dimension, low intrinsic dimension

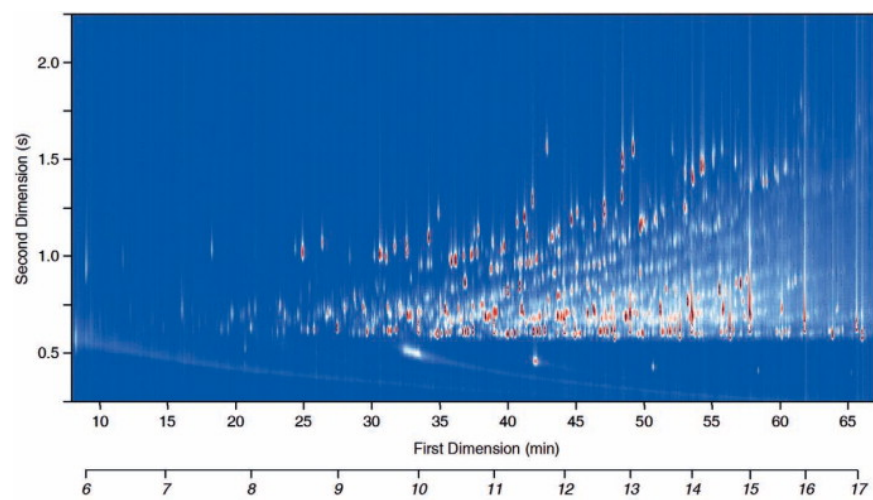
Cryo-electron microscope



0.5 - 5 TB per dataset

**~5–10 DOF: pose
+ conformation
Low-dim manifold**

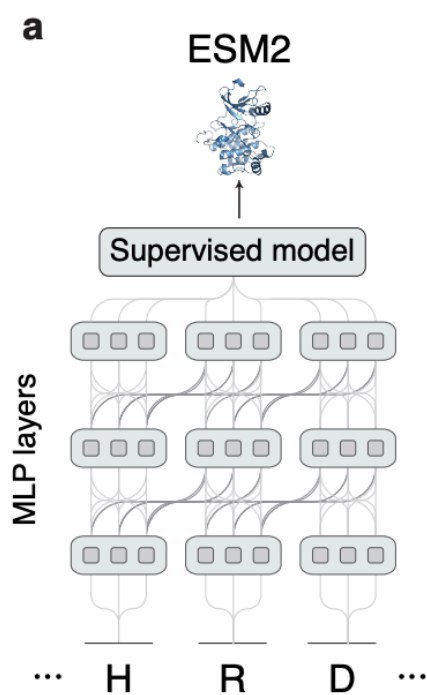
2D gas chromatography



10 - 200 GB per sample

**Sparsity of signal; a
handful discriminate**

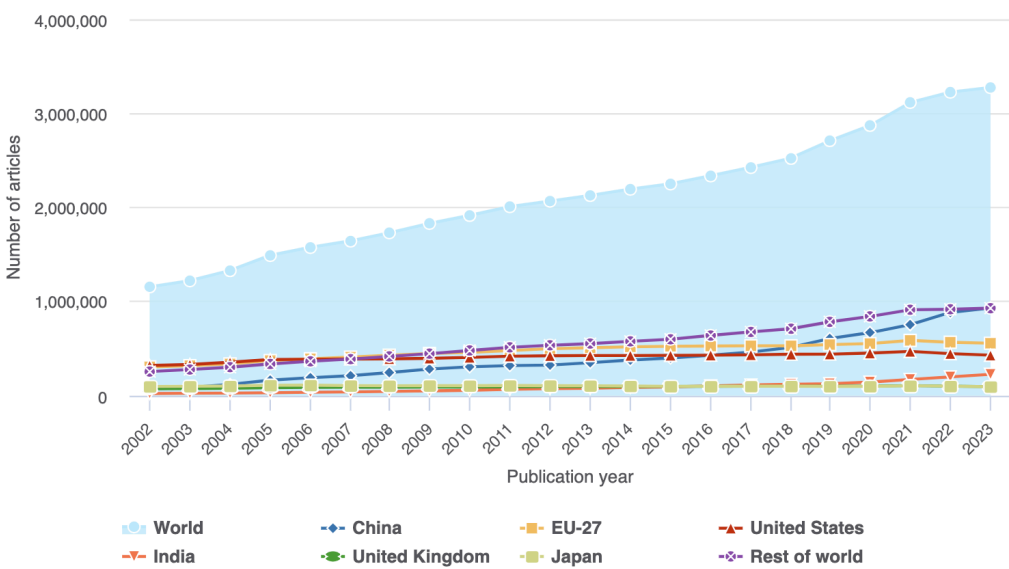
Interpreting pLMs



0.1 - 100 B parameters

**<1% of latents → 79% of
accuracy;
Compressible computation**

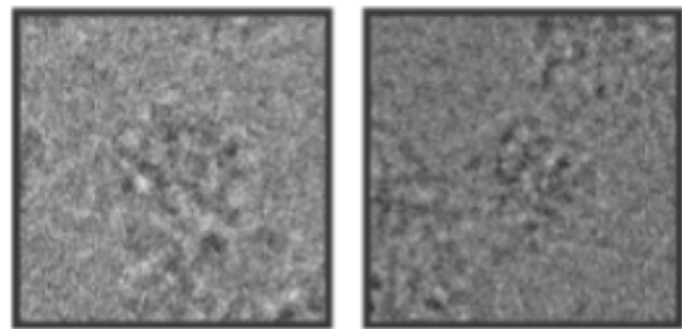
Scientific Hypothesis Generation



3 million new papers per year

High ambient dimension, low intrinsic dimension

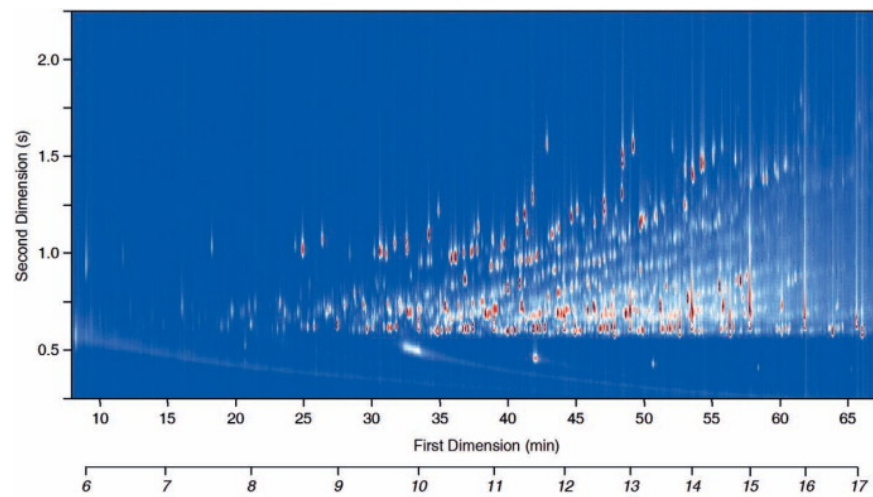
Cryo-electron microscope



0.5 - 5 TB per dataset

**~5–10 DOF: pose
+ conformation
Low-dim manifold**

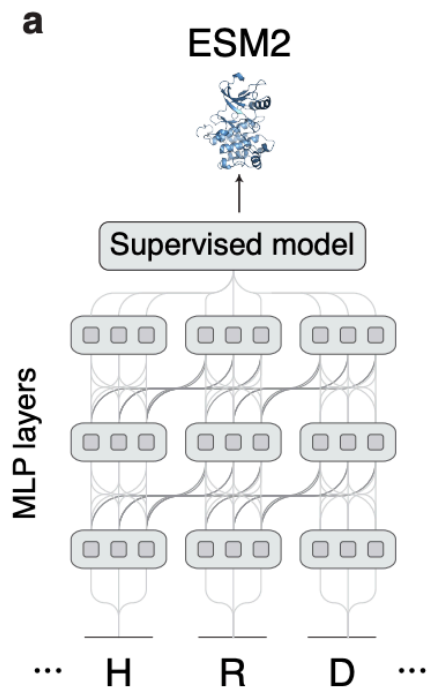
2D gas chromatography



10 - 200 GB per sample

**Sparsity of signal; a
handful discriminate**

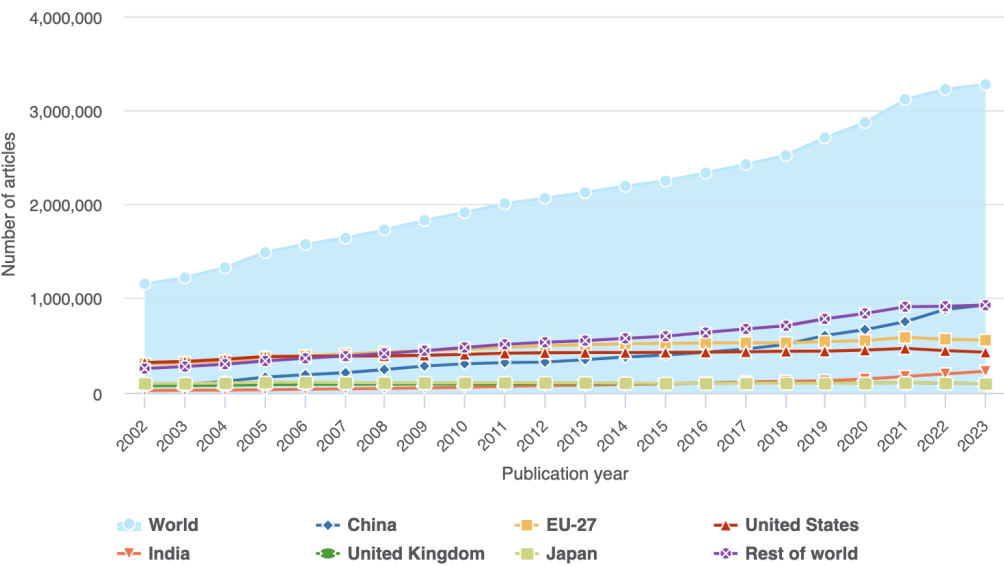
Interpreting pLMs



0.1 - 100 B parameters

**<1% of latents → 79% of
accuracy;
Compressible computation**

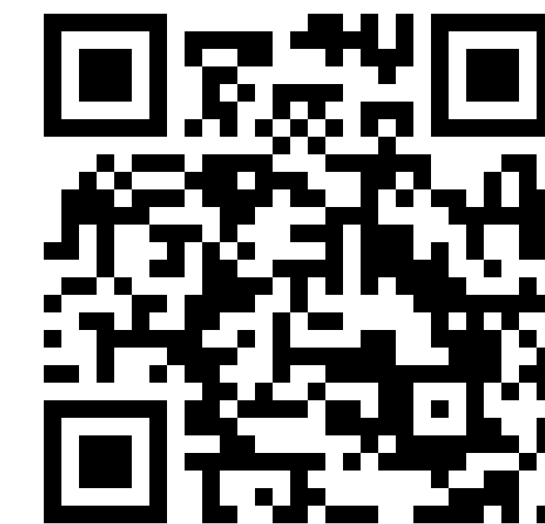
Scientific Hypothesis Generation



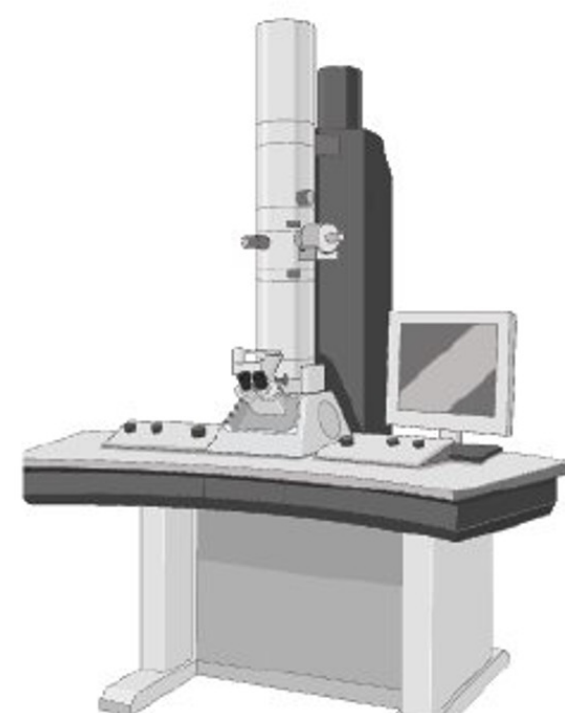
3 million new papers per year

**Constrained by chemistry
+ thermodynamics**

Cryo-EM Workflow



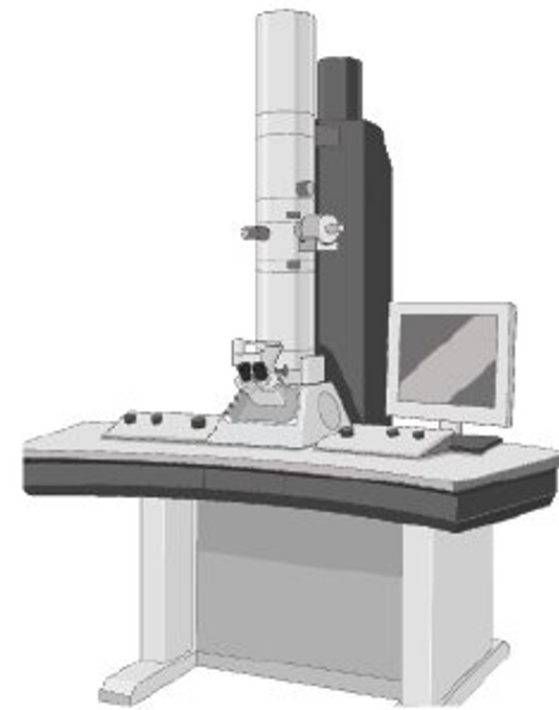
Cryo-electron
microscope



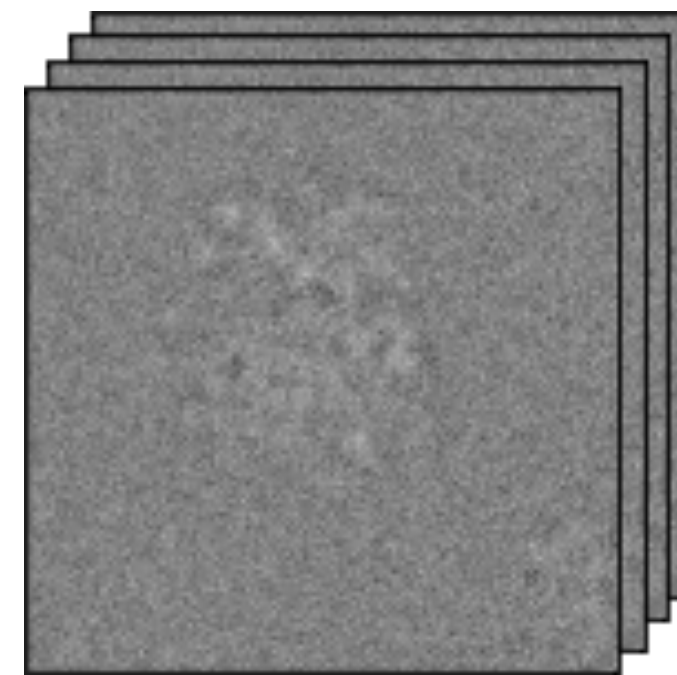
Cryo-EM Workflow



Cryo-electron
microscope



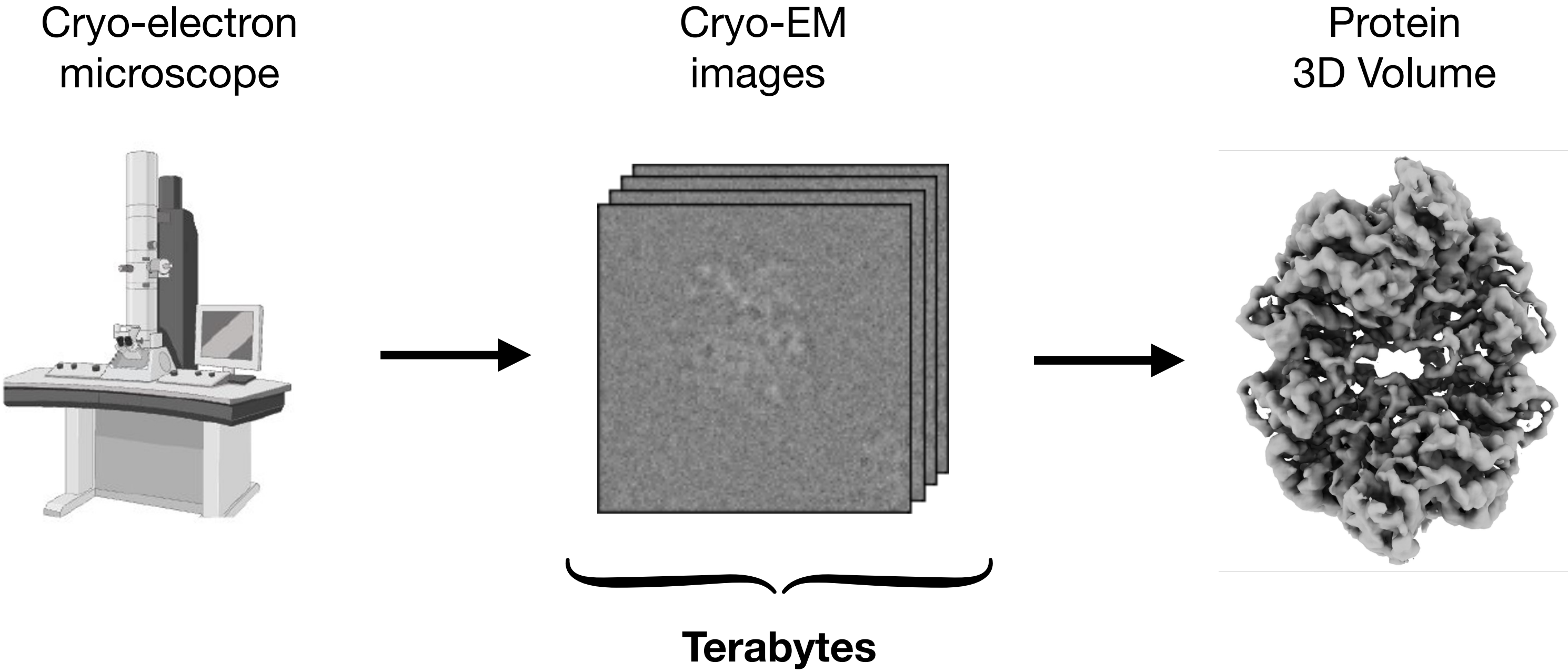
Cryo-EM
images



Terabytes



Cryo-EM Workflow



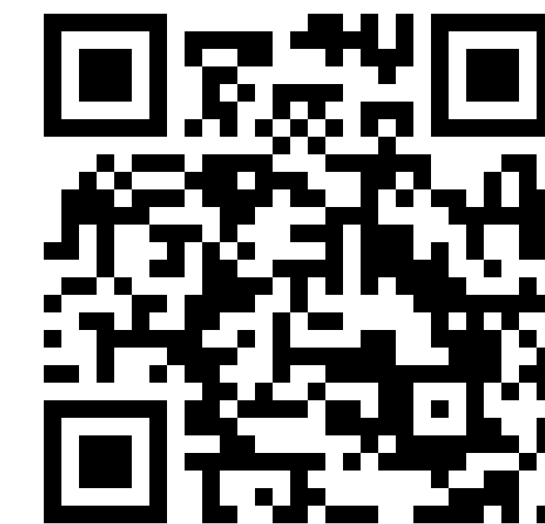
Say what the problem is?

Cryo-EM throughput challenge

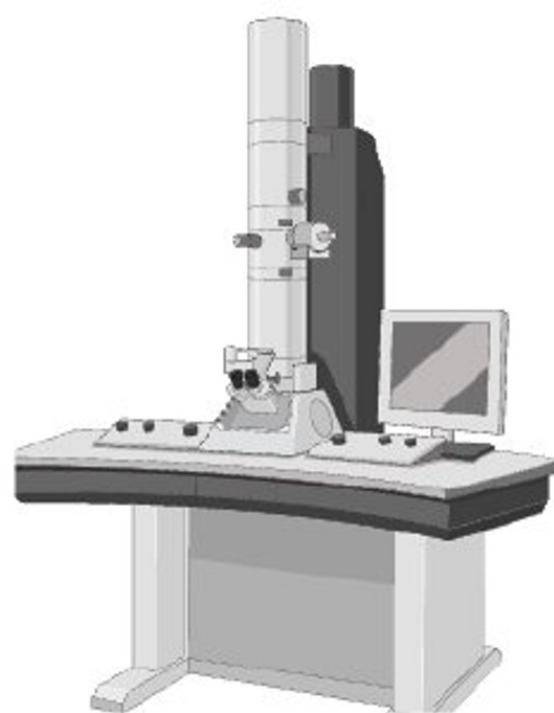


This slide will show the challenge: too much data: maybe some numbers or plots that shows how much data
Is acquired in every cryo run and how much can be processed

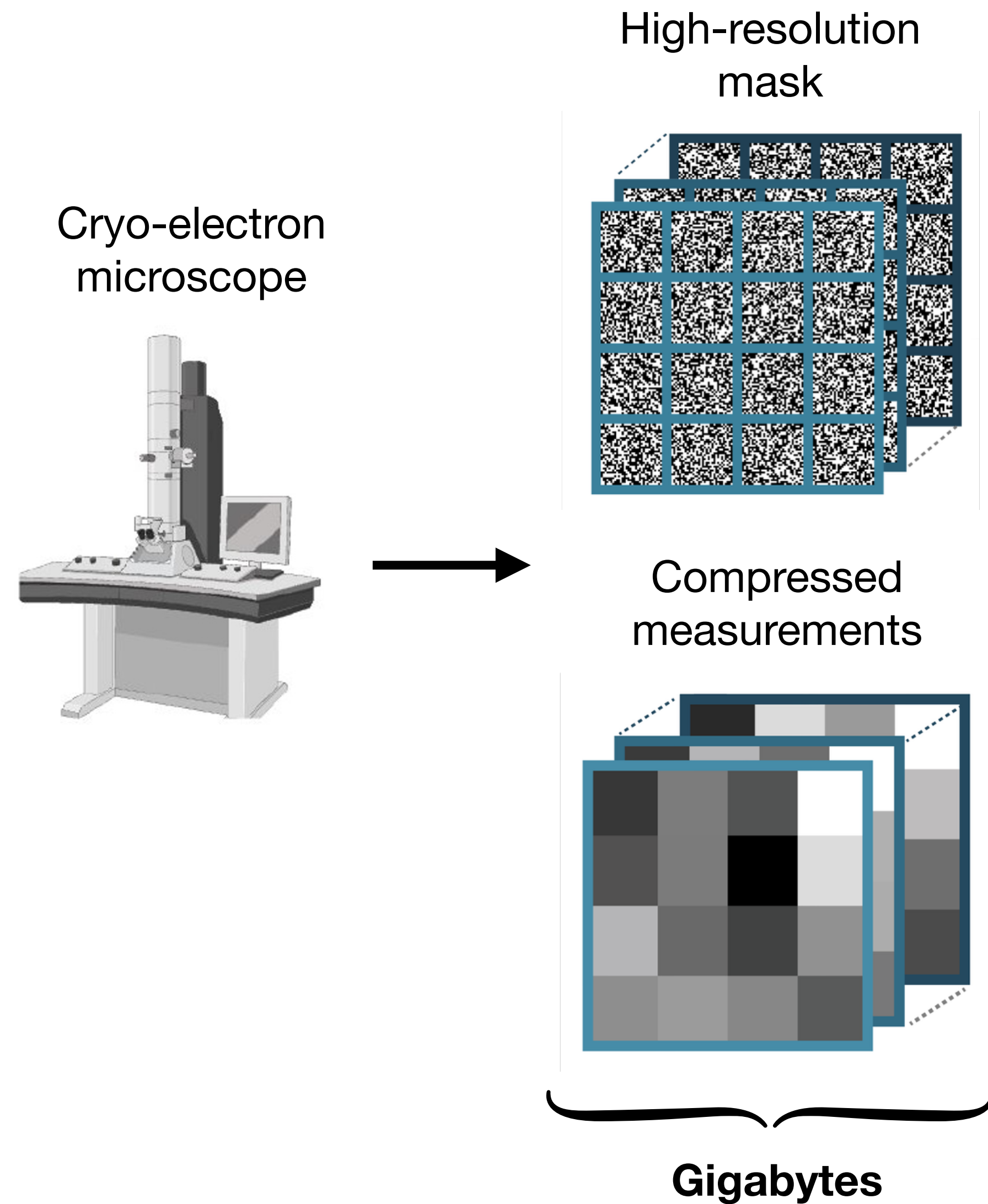
cryoSENSE: Cryo-EM Workflow with Compressive Sensing



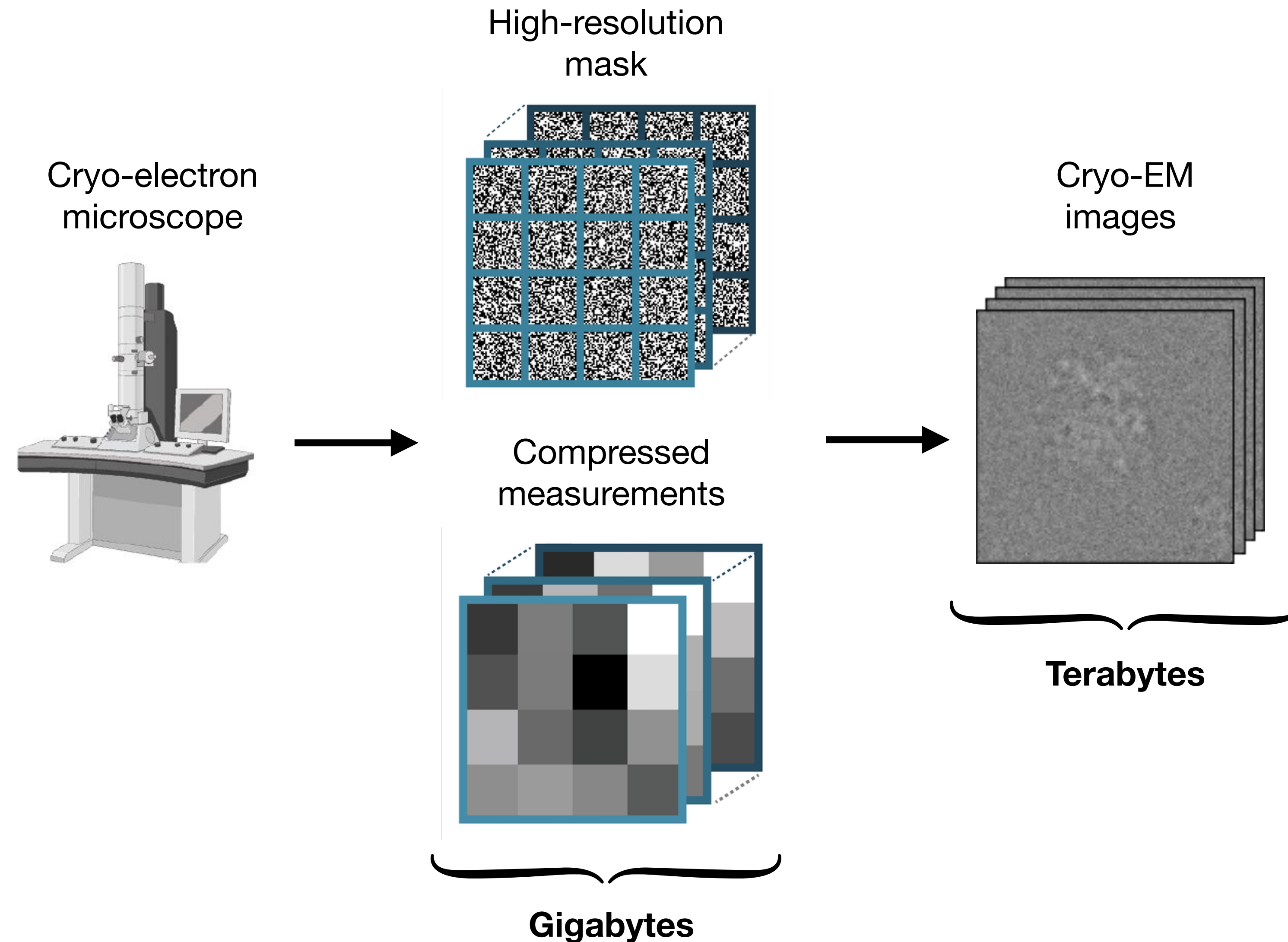
Cryo-electron
microscope



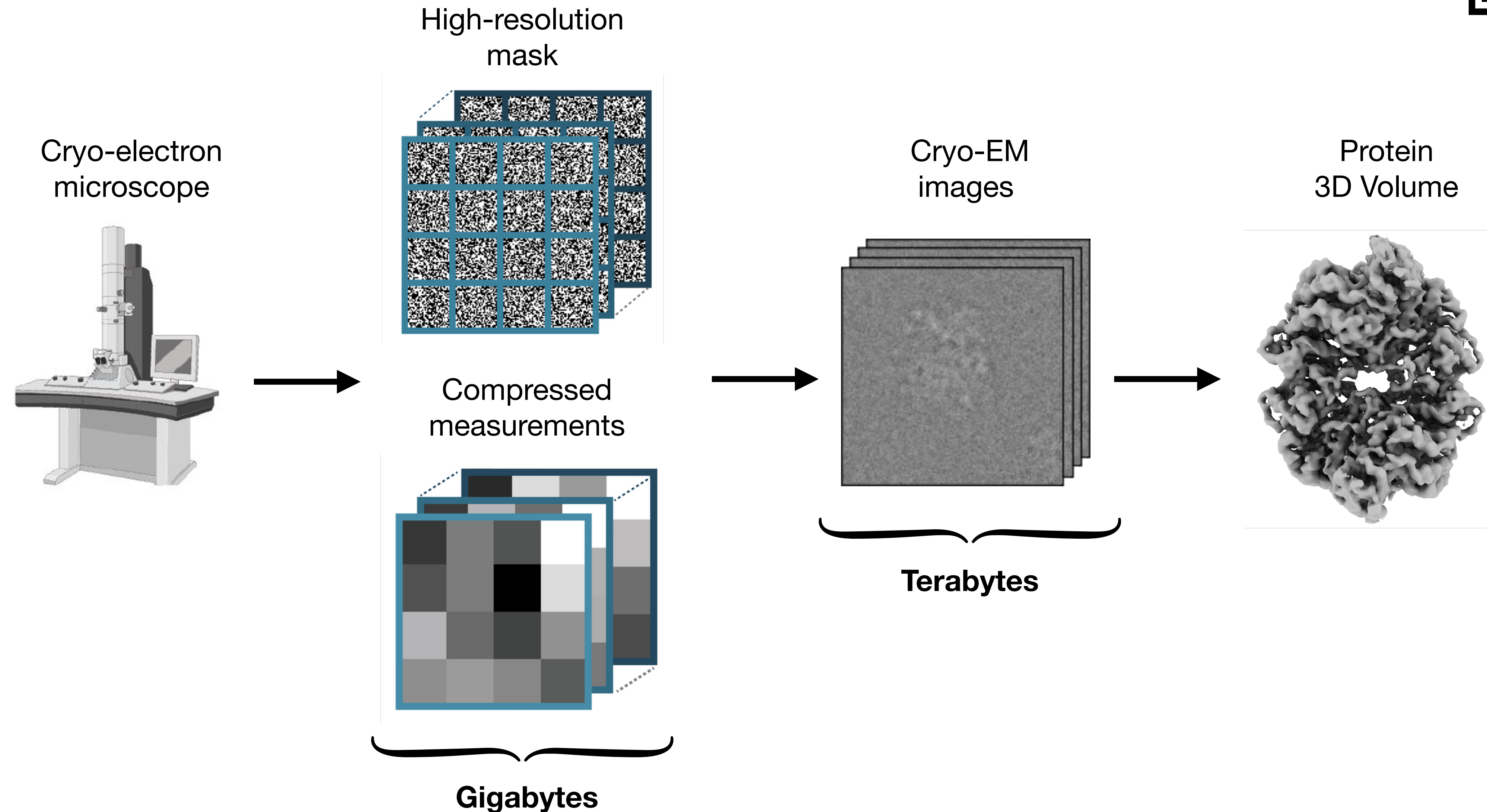
cryoSENSE: Cryo-EM Workflow with Compressive Sensing



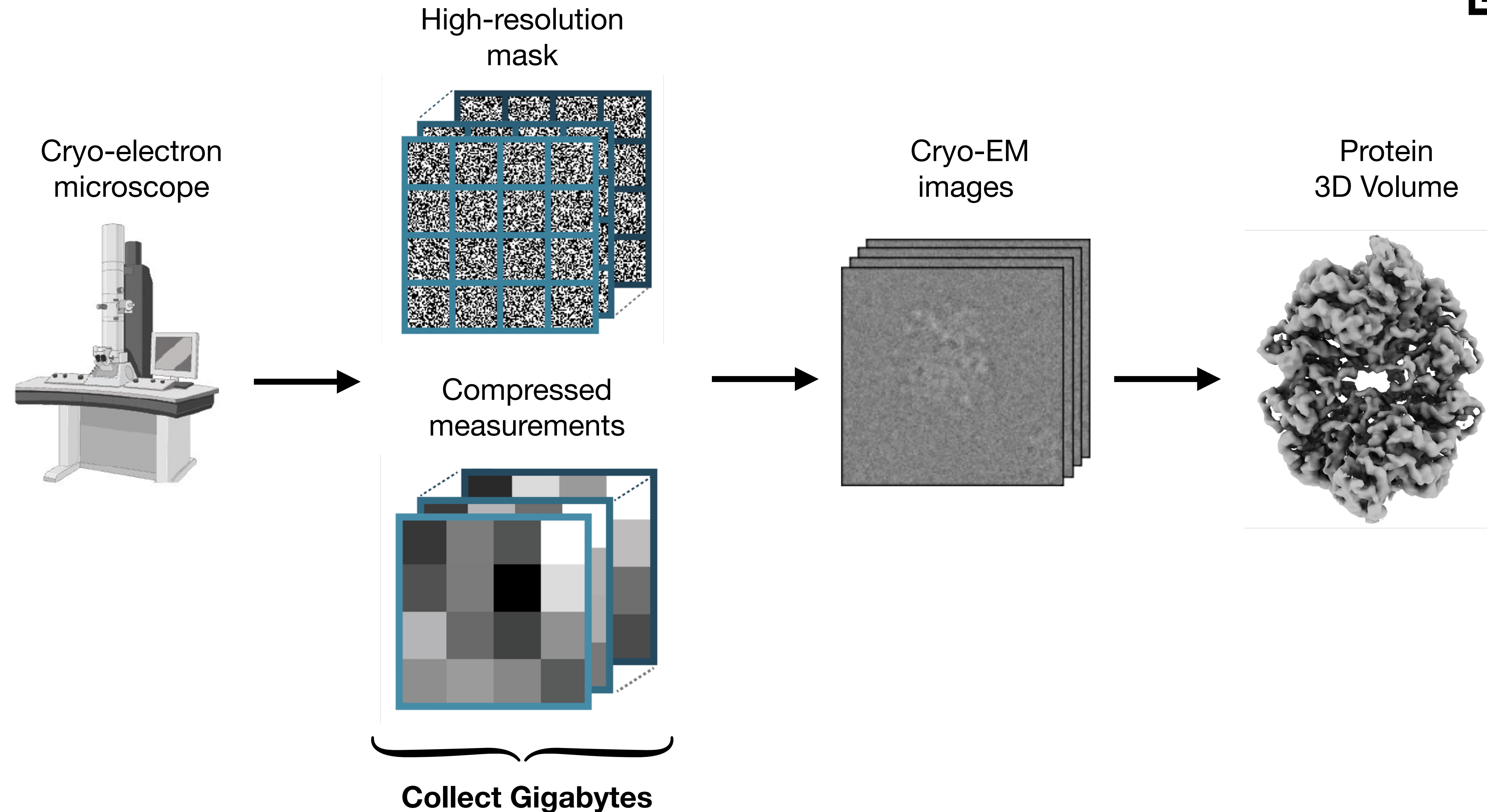
cryoSENSE: Cryo-EM Workflow with Compressive Sensing



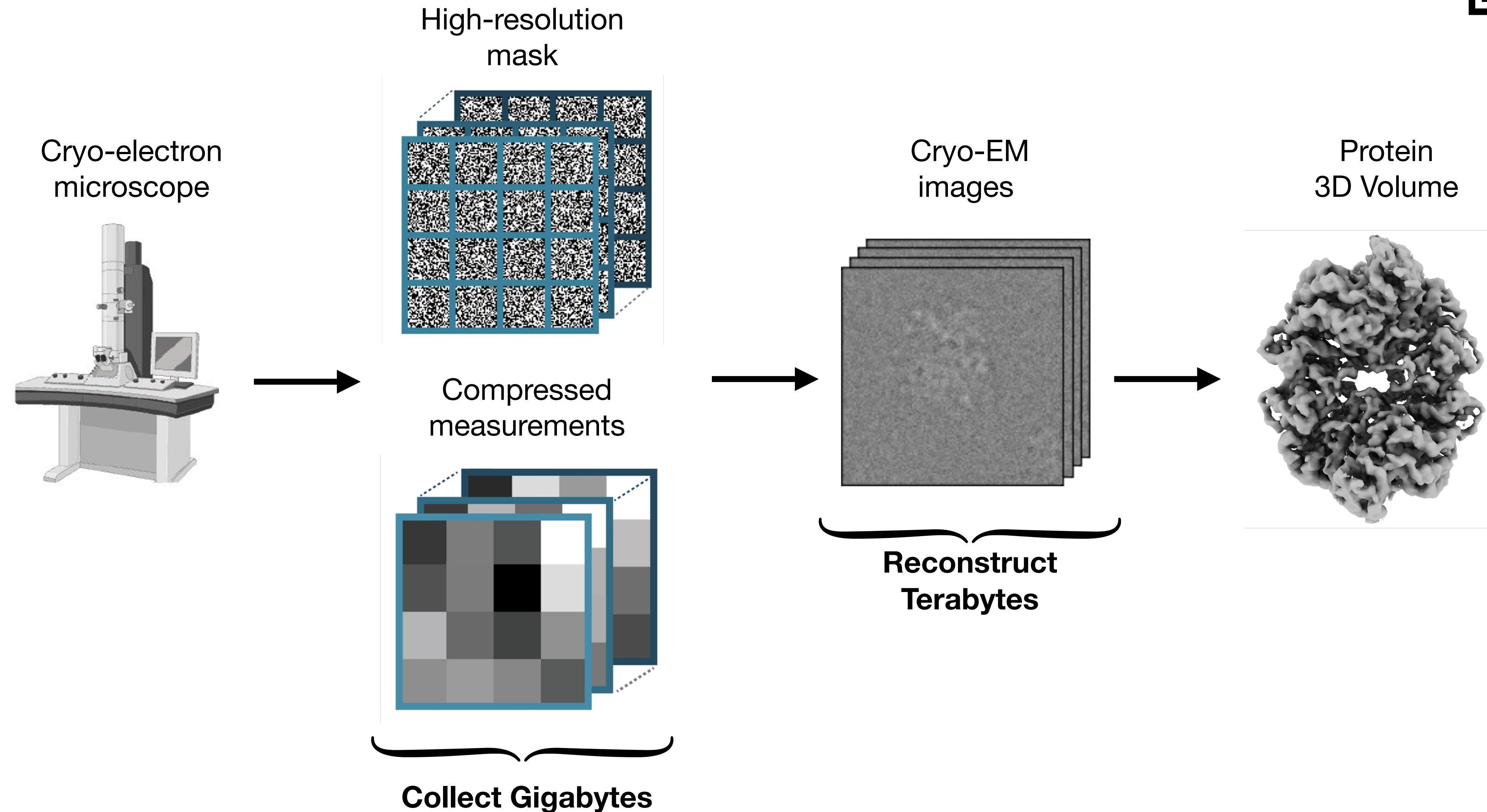
cryoSENSE: Cryo-EM Workflow with Compressive Sensing



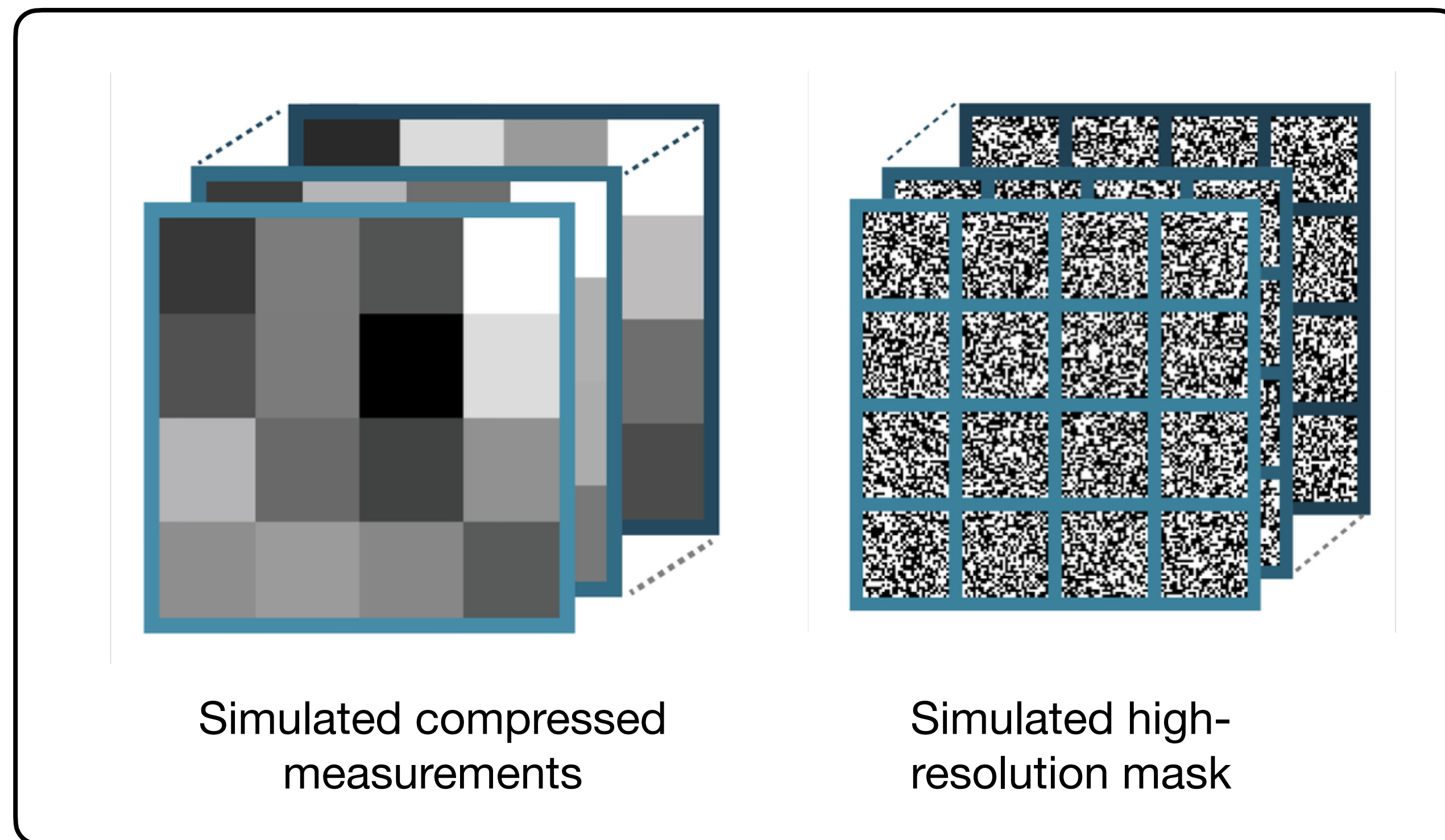
cryoSENSE: Cryo-EM Workflow with Compressive Sensing



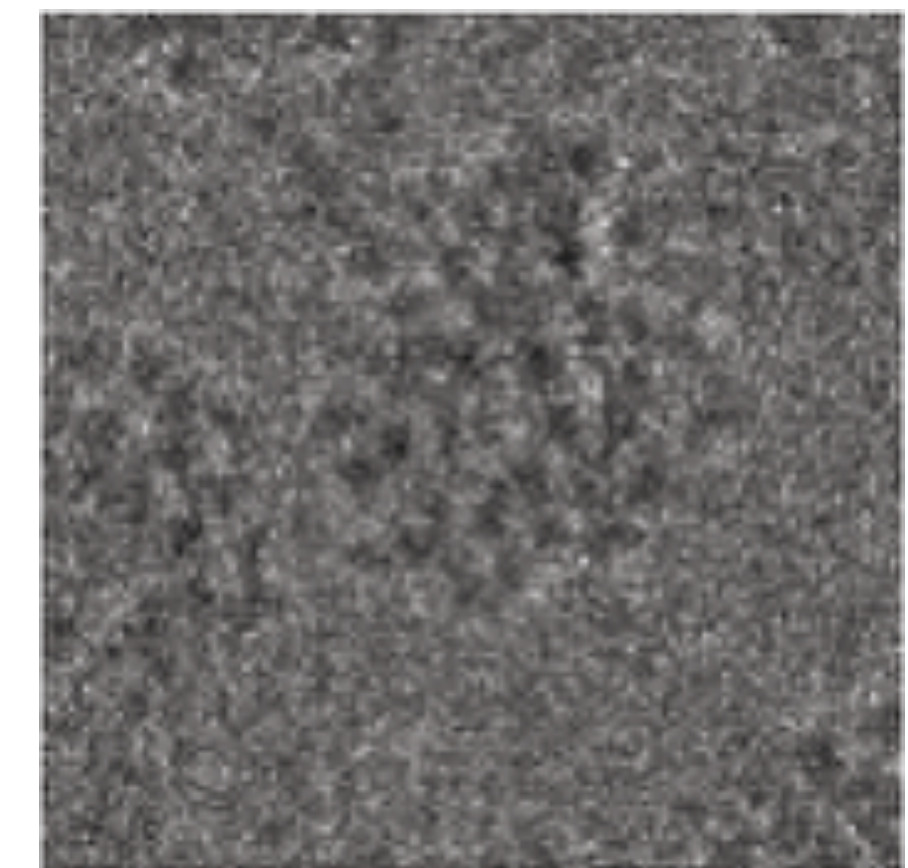
cryoSENSE: Cryo-EM Workflow with Compressive Sensing



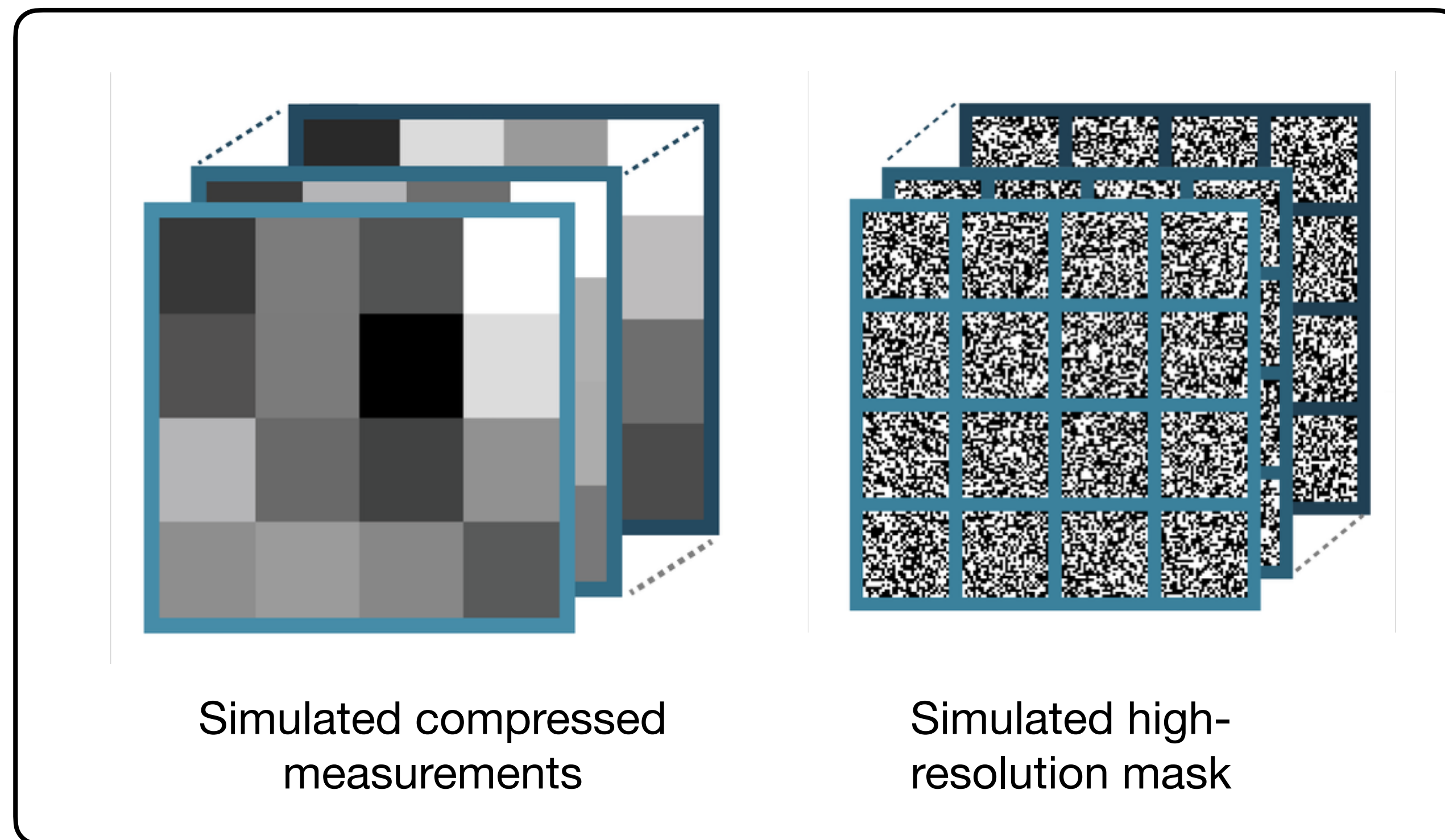
Simulating compressed measurements



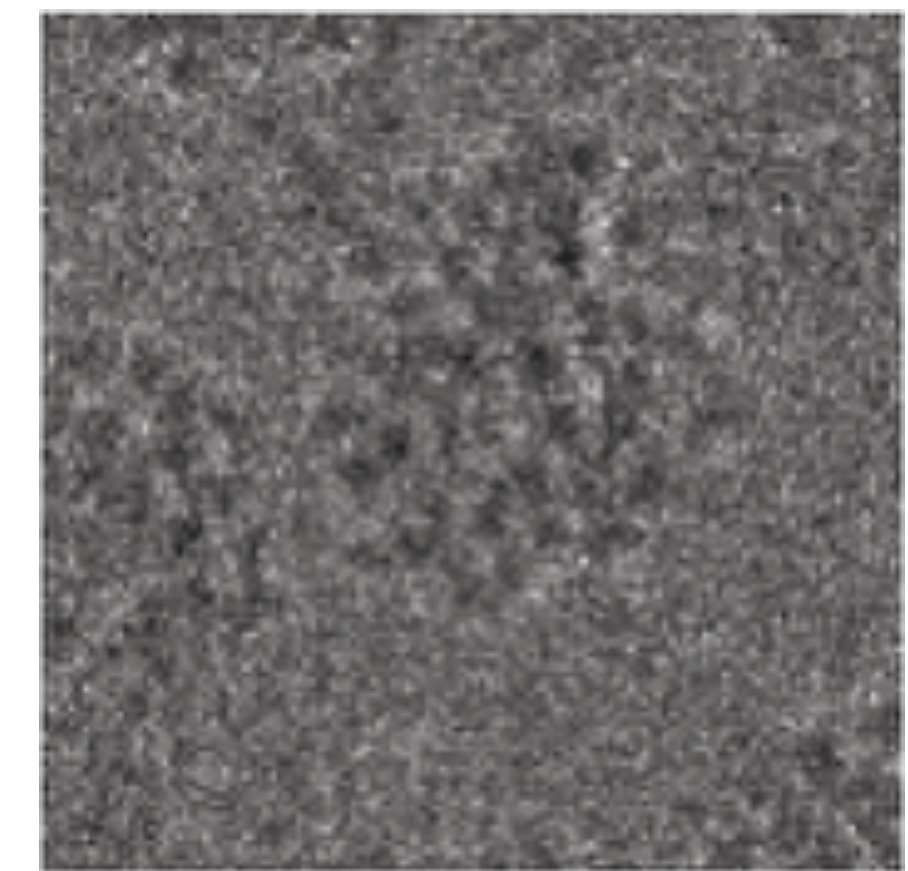
Applied to



Simulating compressed measurements



Given compressed
measurements,
reconstruct high-res
image





Compressive sensing in cryo-EM

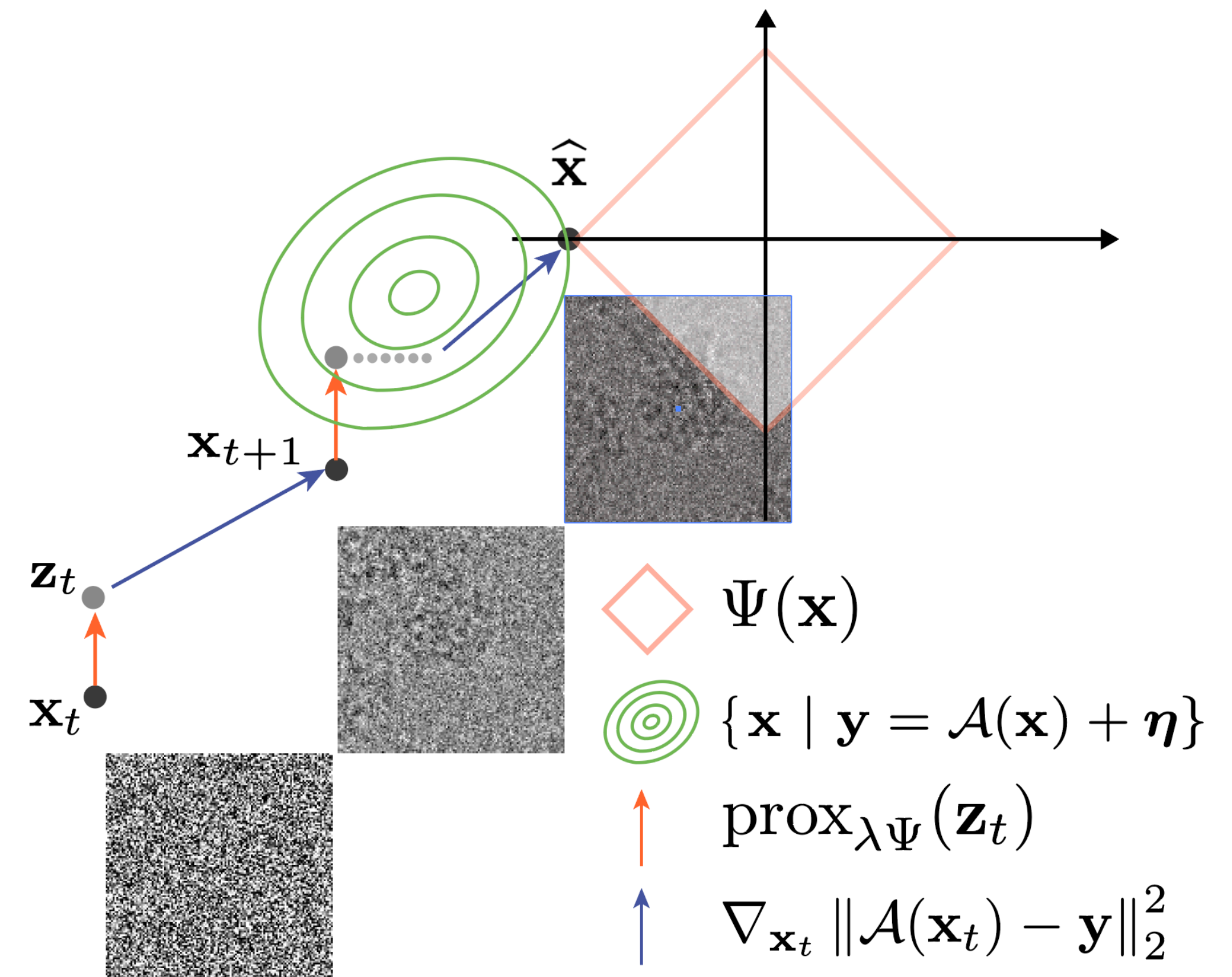
- Reconstructing high-resolution images (x) from compressed measurements (y) with measurement operator ($\mathcal{A}$) and measurement noise (η): $y = \mathcal{A}x + \eta$
- ill-posed: many x match y
- Need a prior on cryo-EM images:
 - Sparse prior: proximal gradient descent
 - Generative prior: posterior sampling



- Solve the convex objective:

$$\min_x \|\mathcal{A}(x) - y\|_2^2 + \lambda \Psi(x),$$

- ψ enforces sparsity in a chosen basis
- Bases tested: DCT, wavelets, and total variation.
- Solved using proximal gradient descent
- Strongest under Fourier-space measurements and moderate real-space compression.
- No need for any external dataset or model





Prior sampling in diffusion models

- A clean cryo-EM image is x_0 . The forward process gradually adds noise: $dx = -\frac{\beta_t}{2}x dt + \sqrt{\beta_t} dw$. $x_T \sim N(0, I)$
- To move from noise x_T back toward a clean image, the corresponding reverse time SDE:

$$d\mathbf{x} = \left[-\frac{\beta_t}{2}\mathbf{x} - \underbrace{\beta_t \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)}_{s_\theta(\mathbf{x}_t, t)} \right] dt + \sqrt{\beta_t} d\bar{\mathbf{w}}.$$

- The network $s_\theta(x_t, t)$ is trained to approximate: $s_\theta(x_t, t) \approx \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$
- Using Tweedie's formula: $\hat{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}} [x_t + (1 - \bar{\alpha}_t)s_\theta(x_t, t)]$
- Unconditional reverse sample: $x'_{t-1} = \frac{\sqrt{\bar{\alpha}_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t}\hat{x}_0 + \tilde{\sigma}_t z.$



From prior sampling to posterior sampling

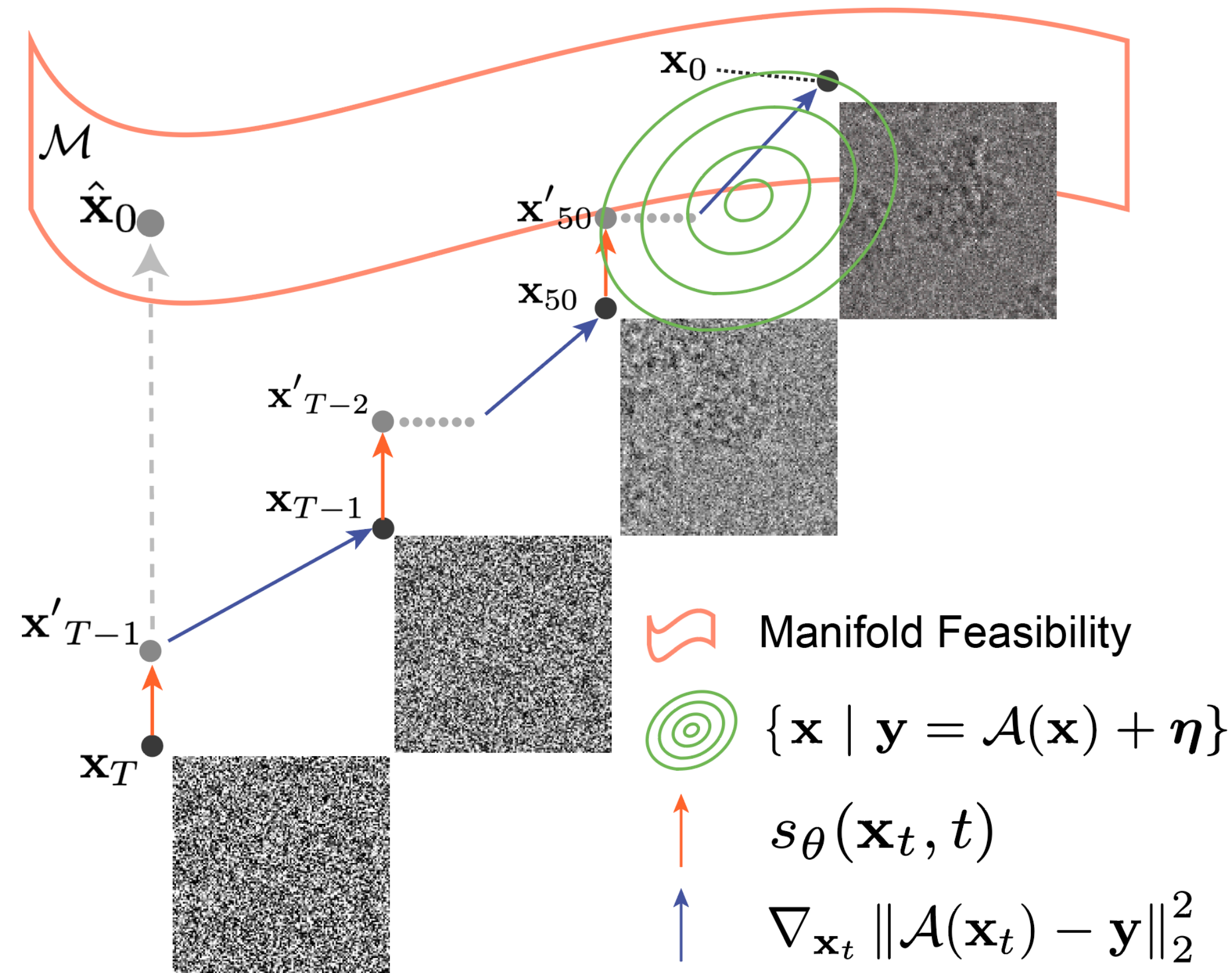
- The reconstruction must satisfy $y = \mathcal{A}(x) + \eta$, therefore, the desired distribution is not merely $p(x_t)$, but $p(x_t | y)$.
- Using Bayes' rule, posterior score is:

$$\nabla_{x_t} \log p(x_t | y) = \underbrace{\nabla_{x_t} \log p(x_t)}_{\text{image prior}} + \underbrace{\nabla_{x_t} \log p(y | x_t)}_{\text{measurement consistency}}$$

- $dx = \left[-\frac{\beta_t}{2} x - \beta_t \left(\nabla_{x_t} \log p(x_t) + \nabla_{x_t} \log p(y | x_t) \right) \right] dt + \sqrt{\beta_t} d\bar{w}.$
- Assuming Gaussian measurement noise: $p(y | x_0) \propto \exp \left(-\frac{\|\mathcal{A}(x_0) - y\|_2^2}{2\sigma^2} \right).$
- Using Bayes' rule:

$$\nabla_{x_t} \log p(y | x_t) \approx \nabla_{x_t} \log p(y | \hat{x}_0(x_t)) = -\frac{1}{\sigma^2} \nabla_{x_t} \|\mathcal{A}(\hat{x}_0) - y\|_2^2.$$

- The measurement loss: $L(x) = \|\mathcal{A}(x) - y\|_2^2$
- Measurement correction: $x_{t-1} = x'_{t-1} - \zeta_t \nabla_{\hat{x}_0} L(\hat{x}_0).$



Algorithm 1 cryoSENSE sampling with Nesterov momentum.

Require: $\mathbf{y}, \mathcal{A}, s_{\theta}(\mathbf{x}_t, t), \{\alpha_t\}_{t=1}^T, \{\kappa_t\}_{t=1}^T, \{\zeta_t\}_{t=1}^T$

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \mathbf{m}_T \leftarrow \mathbf{0}$

2: **for** $t = T, \dots, 1$ **do**

$$\alpha_t = 1 - \beta_t$$

3: $\mathbf{s} \leftarrow s_{\theta}(\mathbf{x}_t, t)$

$$\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$$

4: $\hat{\mathbf{x}}_0 \leftarrow \frac{1}{\sqrt{\bar{\alpha}_t}} (\mathbf{x}_t + (1 - \bar{\alpha}_t) \mathbf{s})$

5: $\mathbf{x}'_{t-1} \leftarrow \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t} \hat{\mathbf{x}}_0 + \tilde{\sigma}_t \mathbf{z}$

6: $\mathbf{p}_t \leftarrow \hat{\mathbf{x}}_0 - \kappa_t \mathbf{m}_t$

7: $\mathbf{g}_t \leftarrow \nabla_{\mathbf{p}_t} \|\mathbf{y} - \mathcal{A}(\mathbf{p}_t)\|_2^2$

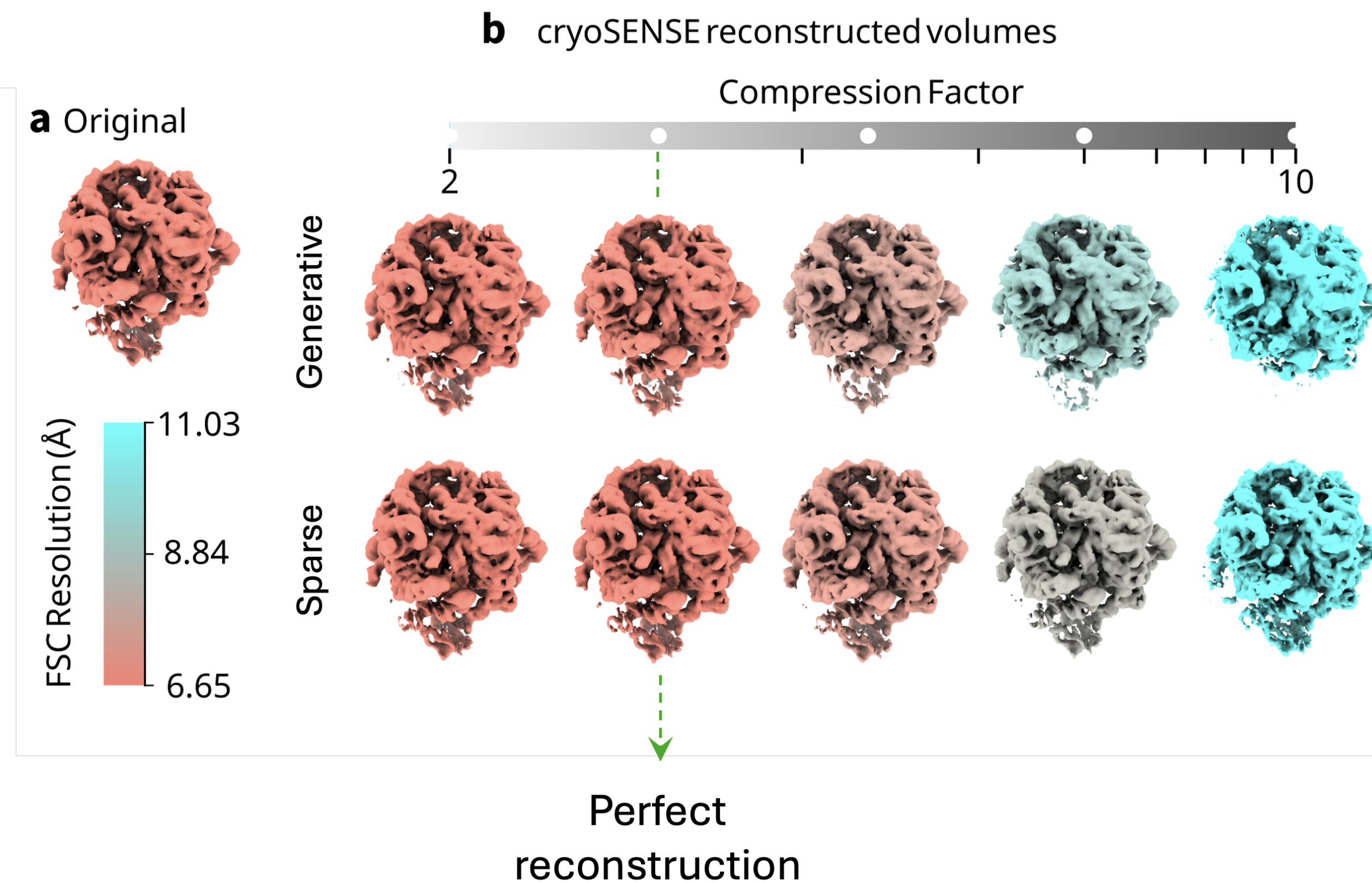
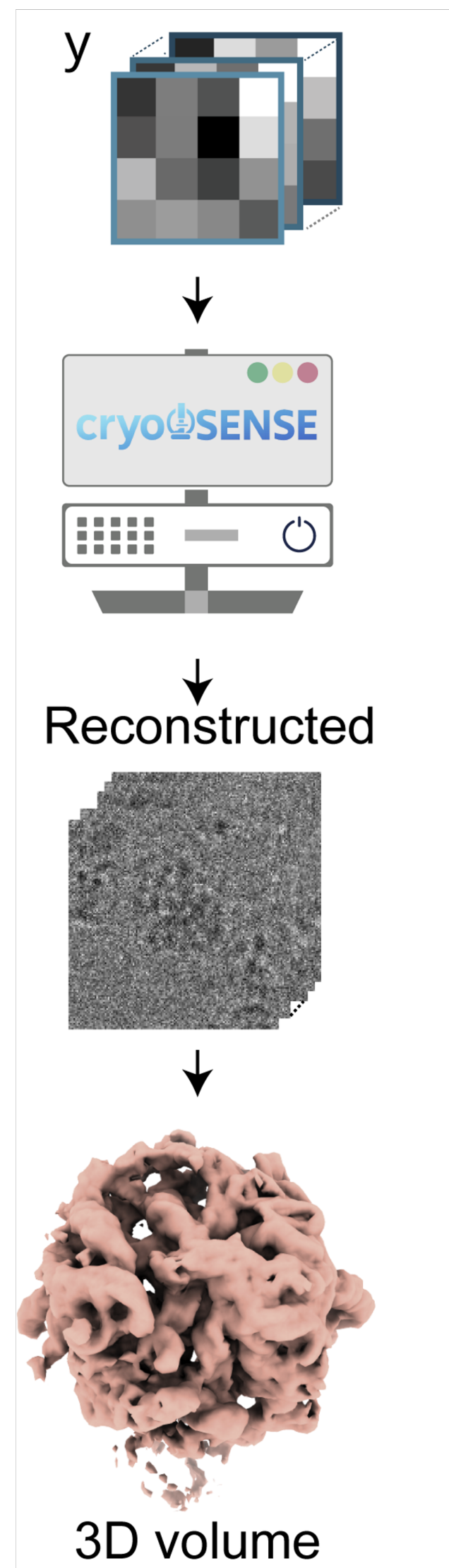
8: $\mathbf{m}_{t-1} \leftarrow \kappa_t \mathbf{m}_t + \zeta_t \mathbf{g}_t$

9: $\mathbf{x}_{t-1} \leftarrow \mathbf{x}'_{t-1} - \mathbf{m}_{t-1}$

10: **end for**

11: **return** $\mathbf{x}_0$

cryoSENSE Results

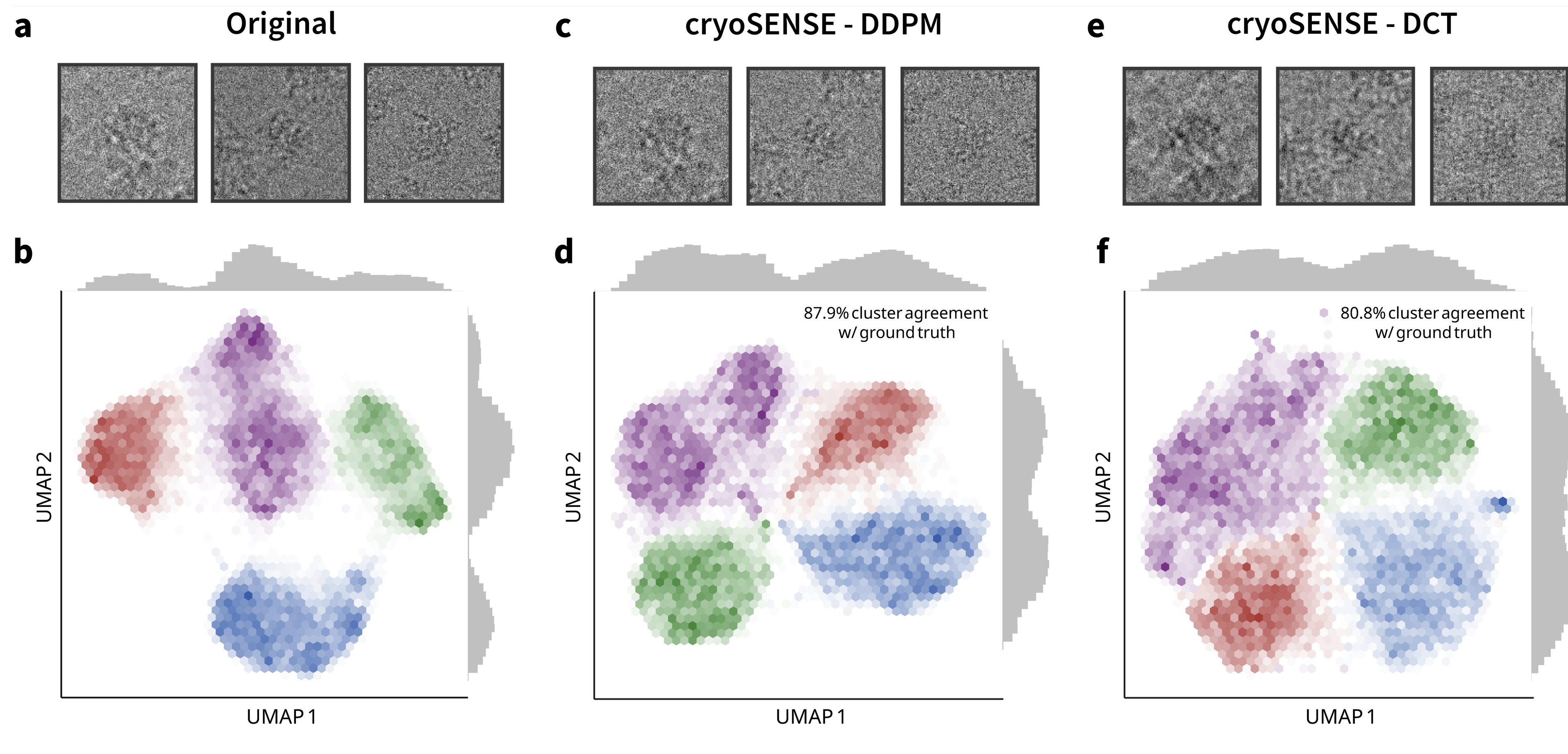


- cryoSENSE enables high throughput acquisition of cryo-EM images with a tunable compression factor C
- We are able to increase throughput by 2.5X

cryoSENSE Results



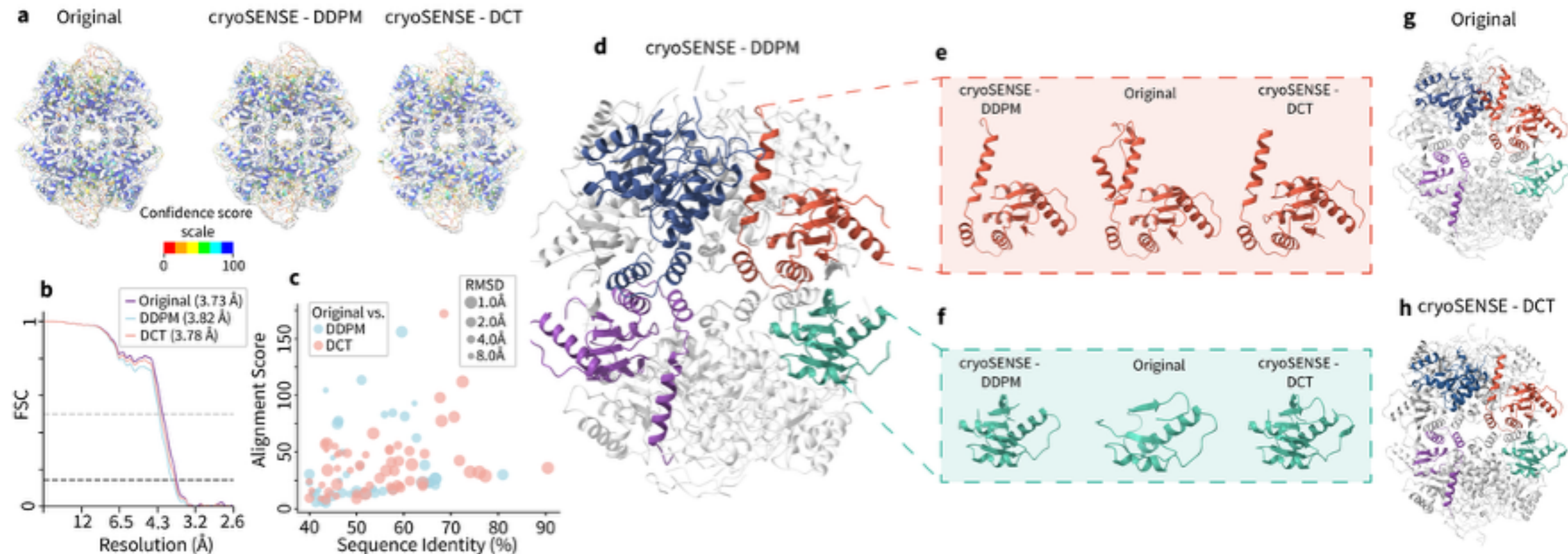
- cryoSENSE preserves structural heterogeneity of 50S ribosome with 16X down-sampled measurements ($C=1.25$)
- Generative prior outperforms sparse prior at high down-sampling levels



cryoSENSE Results



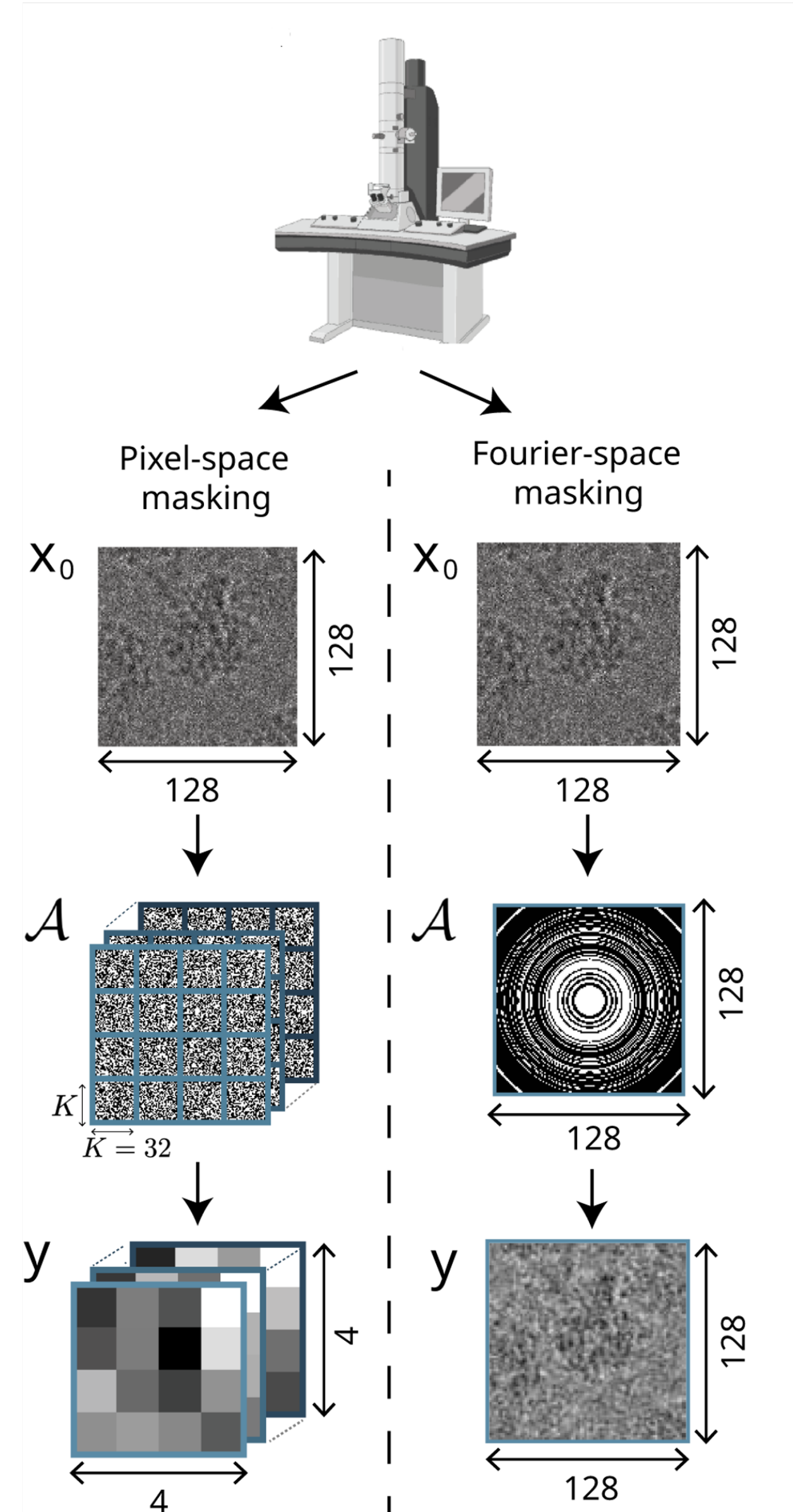
- cryoSENSE preserves protein atomic structure with 2X down-sampled images ($C=1.33$)
- Sparse prior slightly outperforms generative prior



cryo**SENSE** Proposed methods for collecting measurements



Pixel space masking collects multiple masked down-sampled measurements in real space

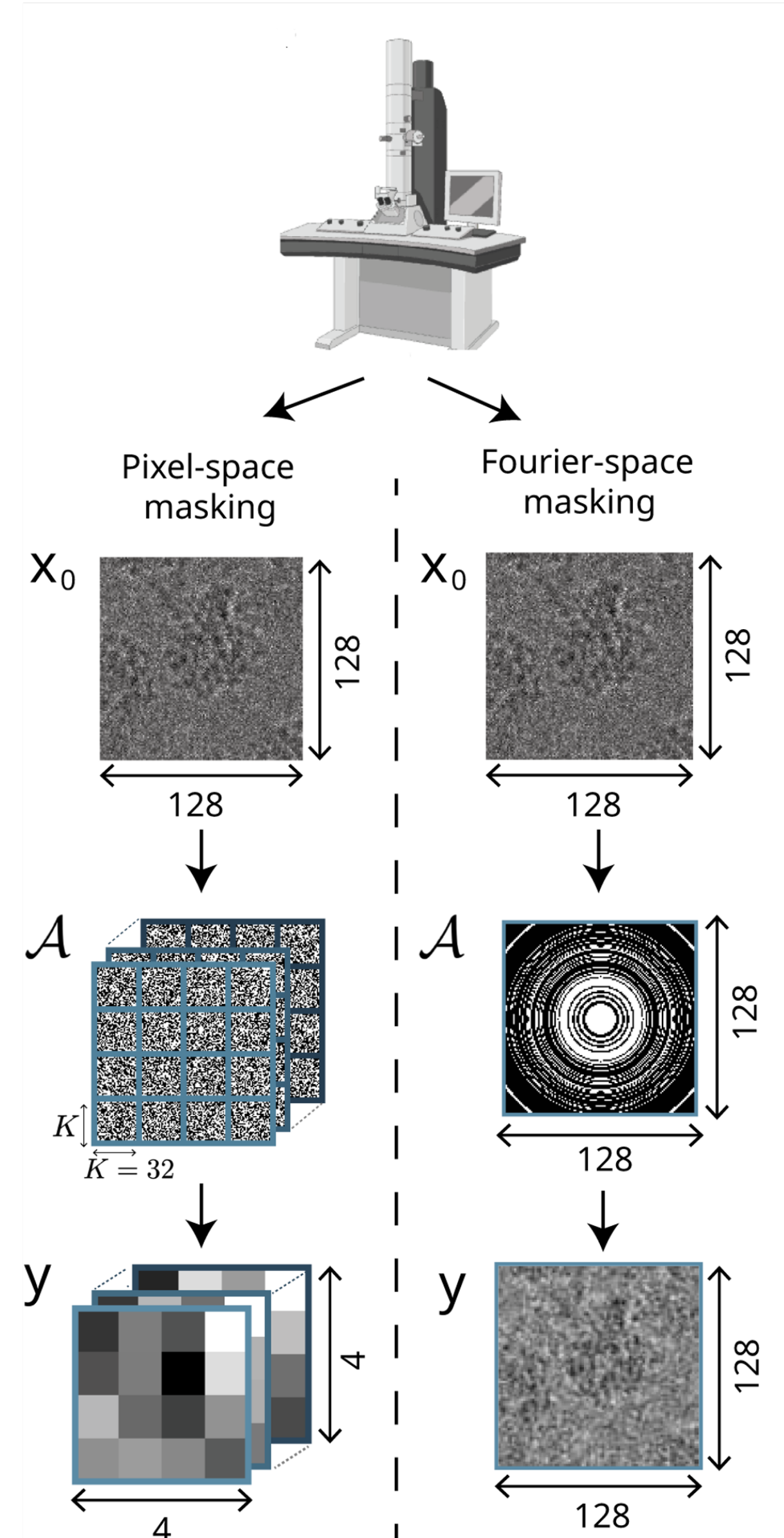


cryo**SENSE** Proposed methods for collecting measurements



Pixel space masking collects multiple masked down-sampled measurements in real space

Physical pattern (mask) that modulates transmission before electrons reach detector



Fourier space masking collects measurements by keeping only a subset of fourier coefficients

MEMS-based phase plates to modulate phase and amplitude in the back focal plane

Inverse-fourier transformed to collect compressed measurement

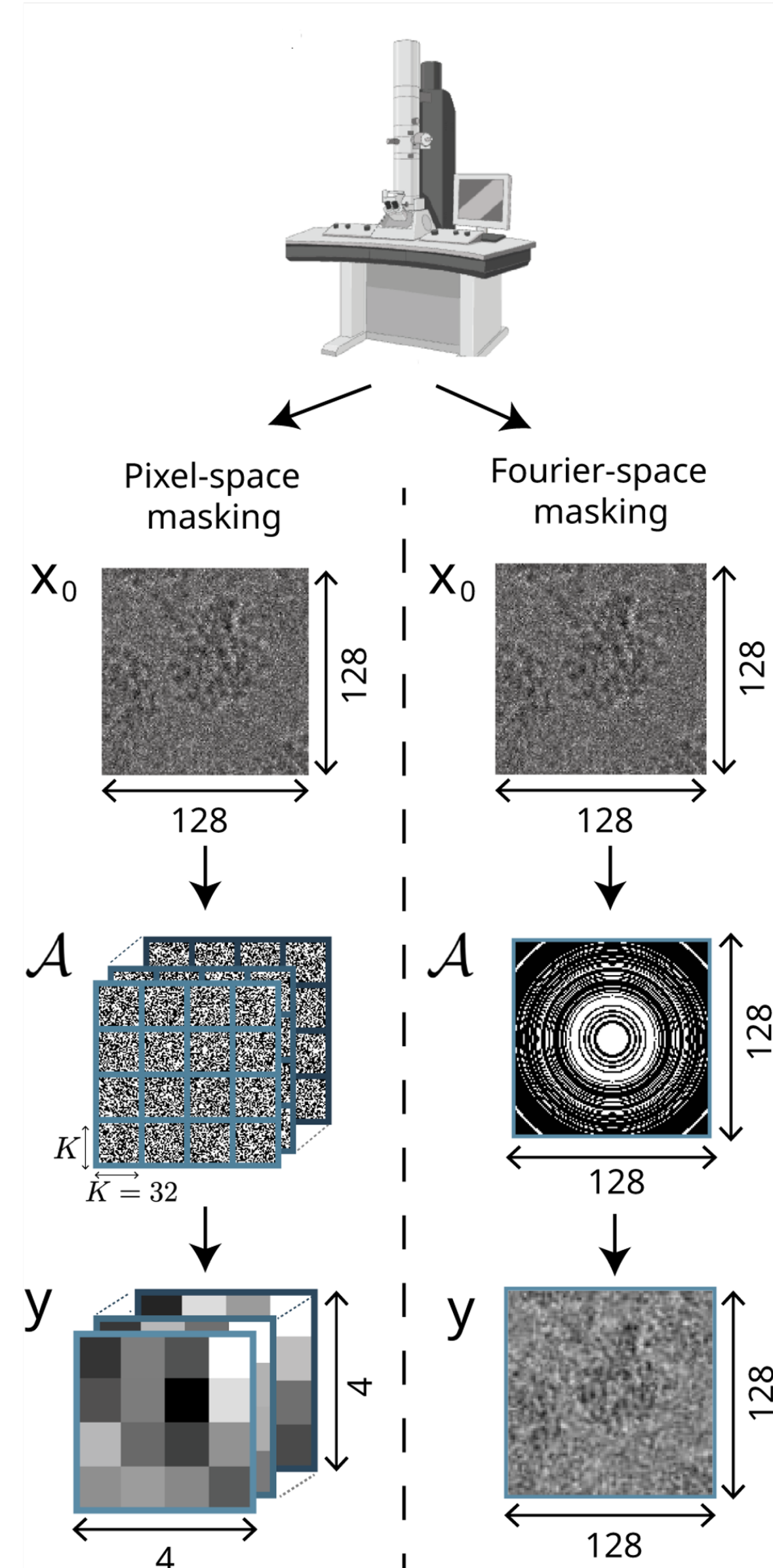
cryo**SENSE** Proposed methods for collecting measurements



Pixel space masking collects multiple masked down-sampled measurements in real space

Physical pattern (mask) that modulates transmission before electrons reach detector

Binning on the detector to collect compressed measurement



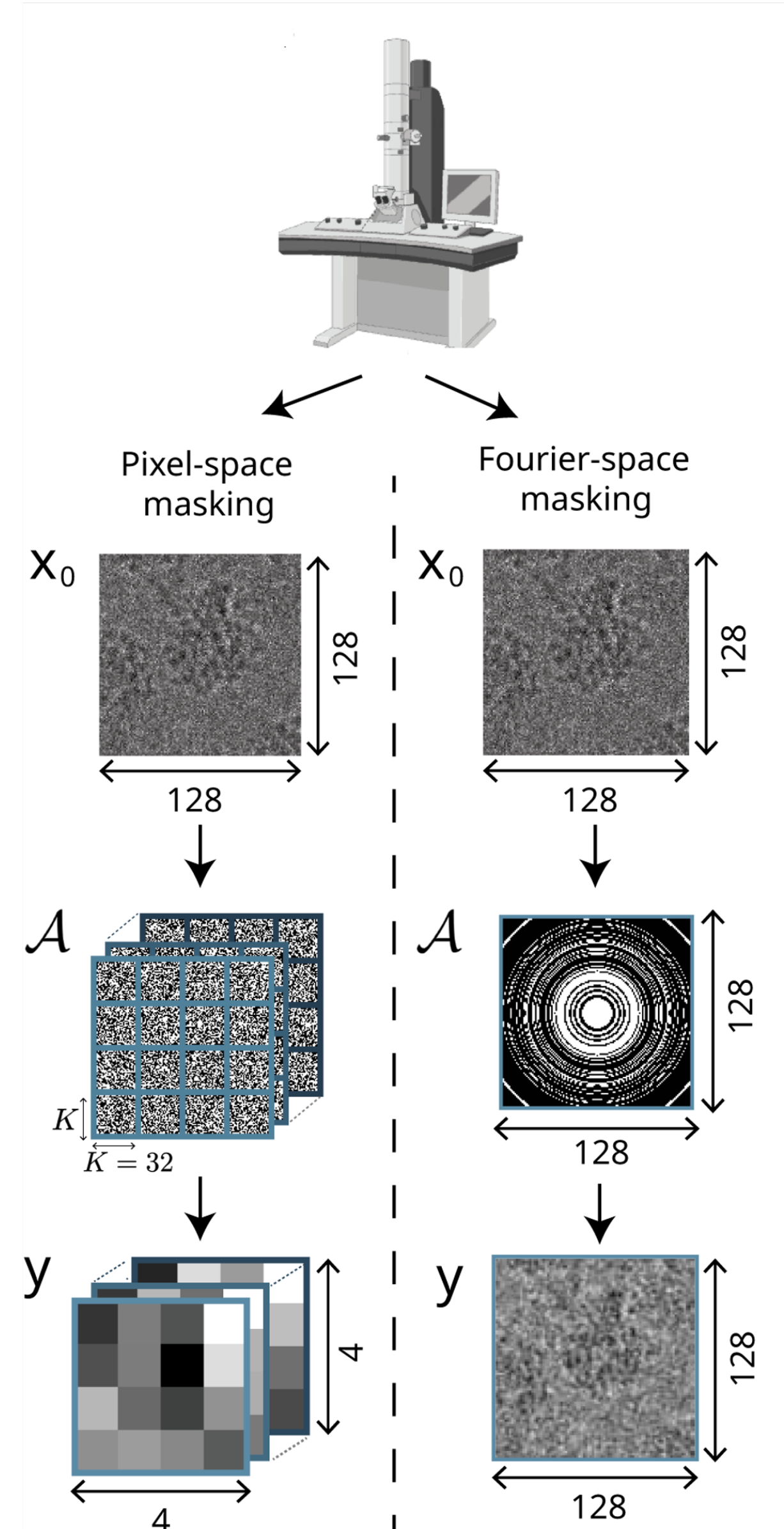
cryoSENSE Proposed methods for collecting measurements



Pixel space masking collects multiple masked down-sampled measurements in real space

Physical pattern (mask) that modulates transmission before electrons reach detector

Binning on the detector to collect compressed measurement



Fourier space masking collects measurements by keeping only a subset of fourier coefficients

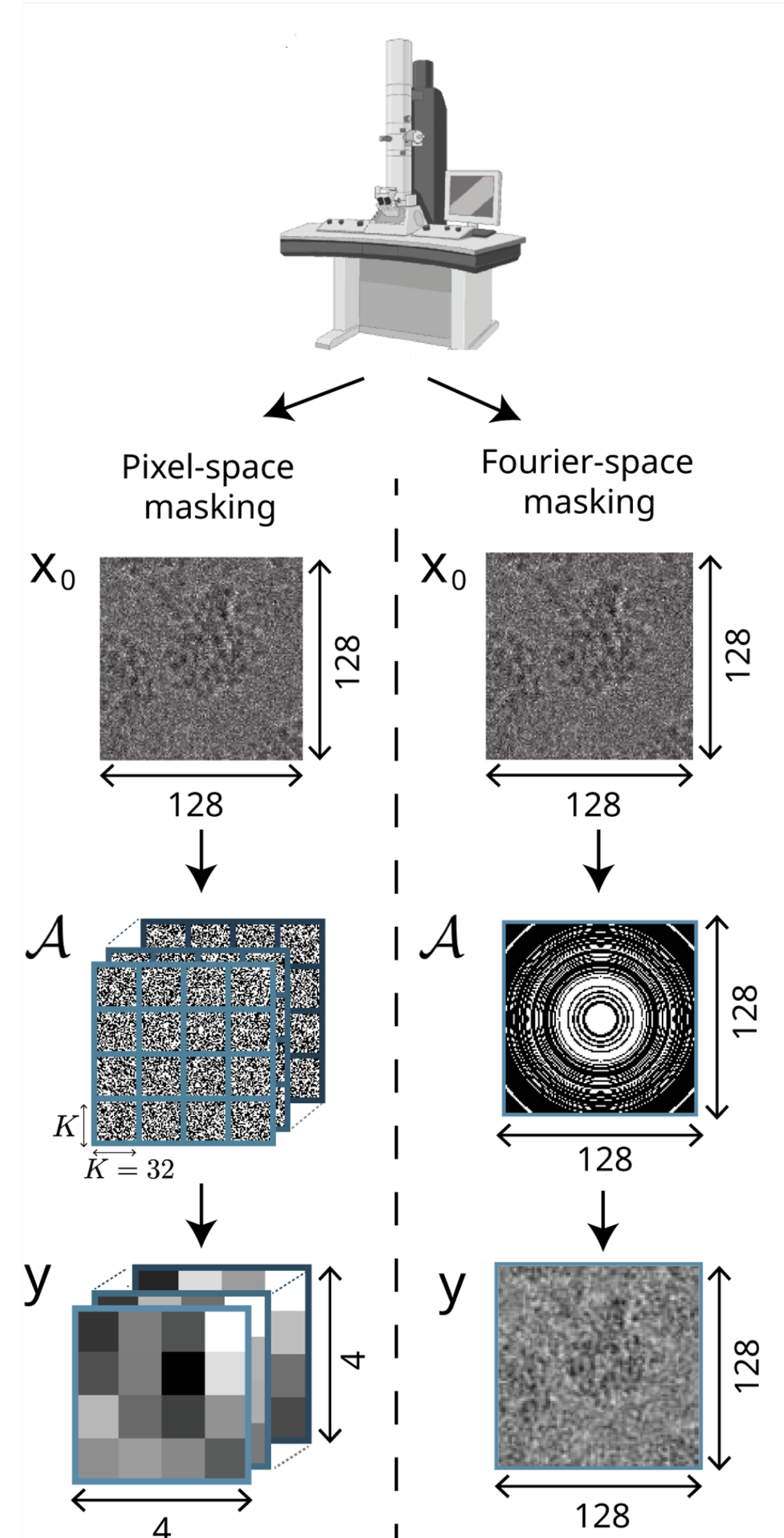
cryoSENSE Proposed methods for collecting measurements



Pixel space masking collects multiple masked down-sampled measurements in real space

Physical pattern (mask) that modulates transmission before electrons reach detector

Binning on the detector to collect compressed measurement



Fourier space masking collects measurements by keeping only a subset of fourier coefficients

MEMS-based phase plates to modulate phase and amplitude in the back focal plane

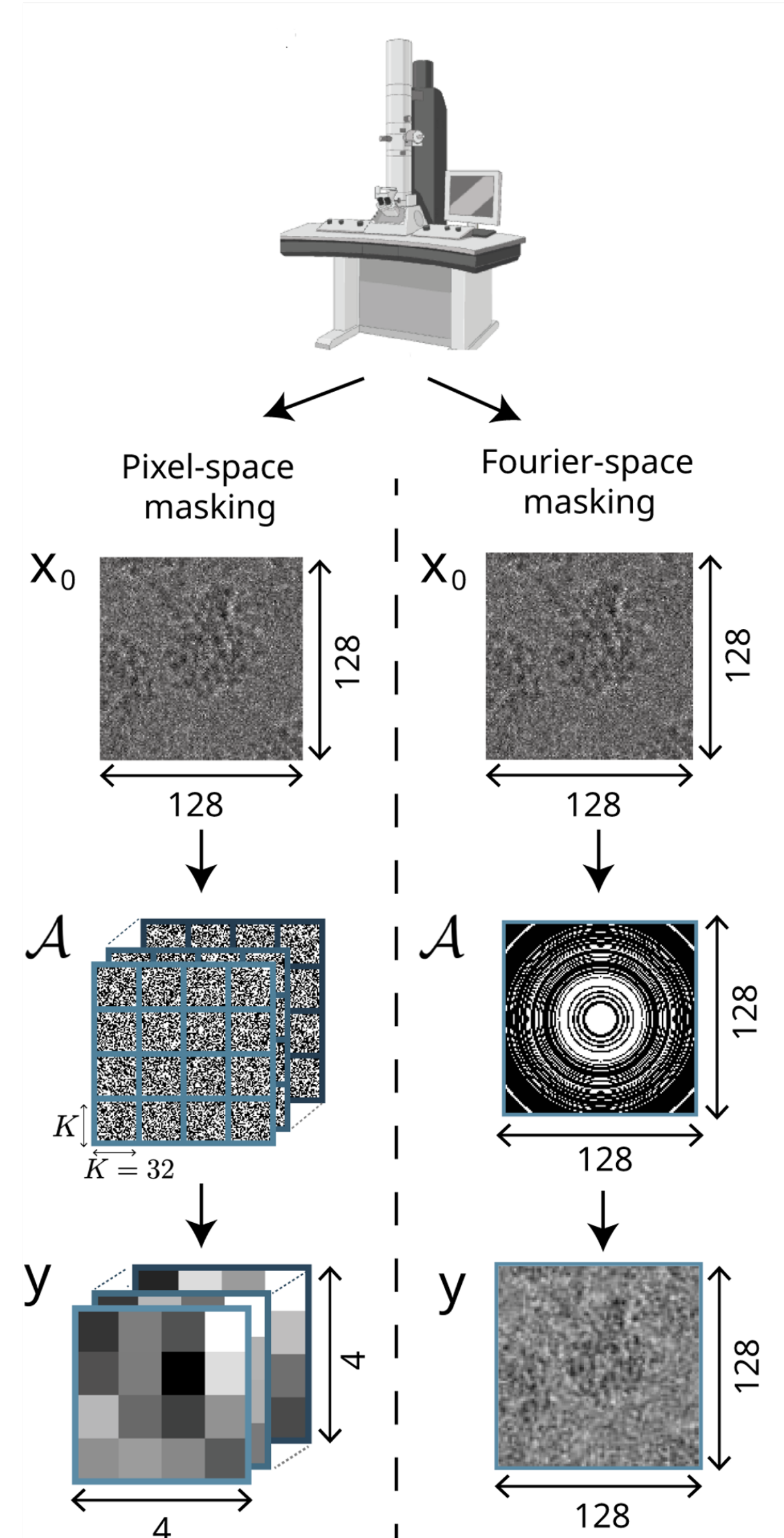
cryo**SENSE** Proposed methods for collecting measurements



Pixel space masking collects multiple masked down-sampled measurements in real space

Physical pattern (mask) that modulates transmission before electrons reach detector

Binning on the detector to collect compressed measurement

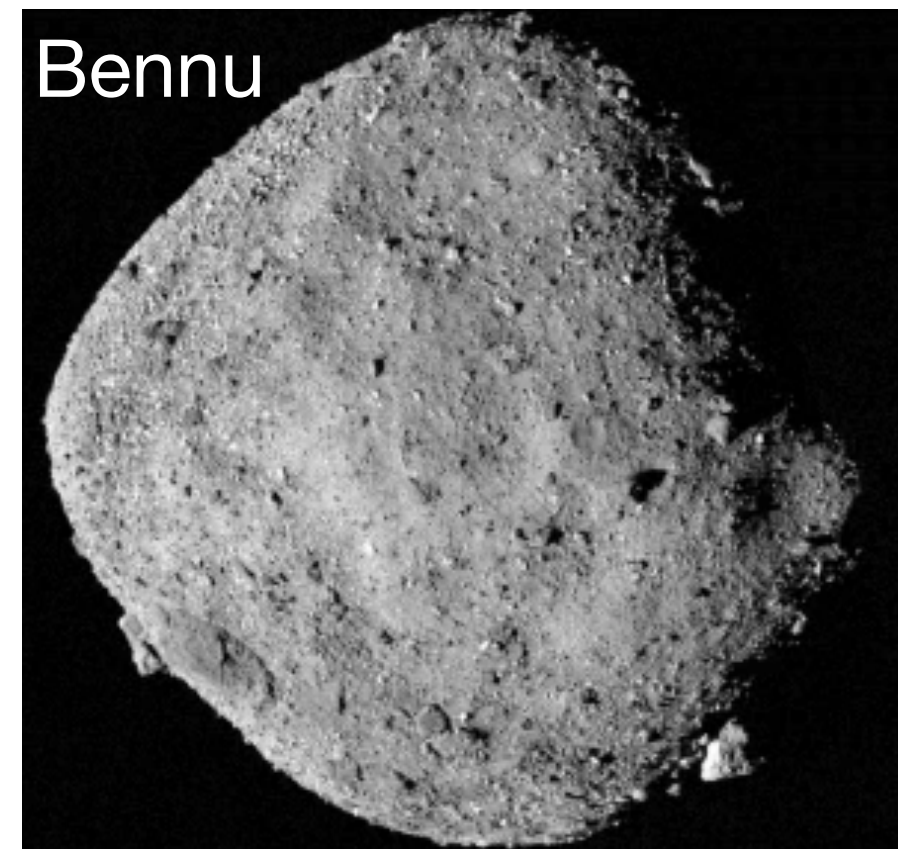


Fourier space masking collects measurements by keeping only a subset of fourier coefficients

MEMS-based phase plates to modulate phase and amplitude in the back focal plane

Inverse-fourier transformed to collect compressed measurement

Sample return missions: need for tools to detect alien life

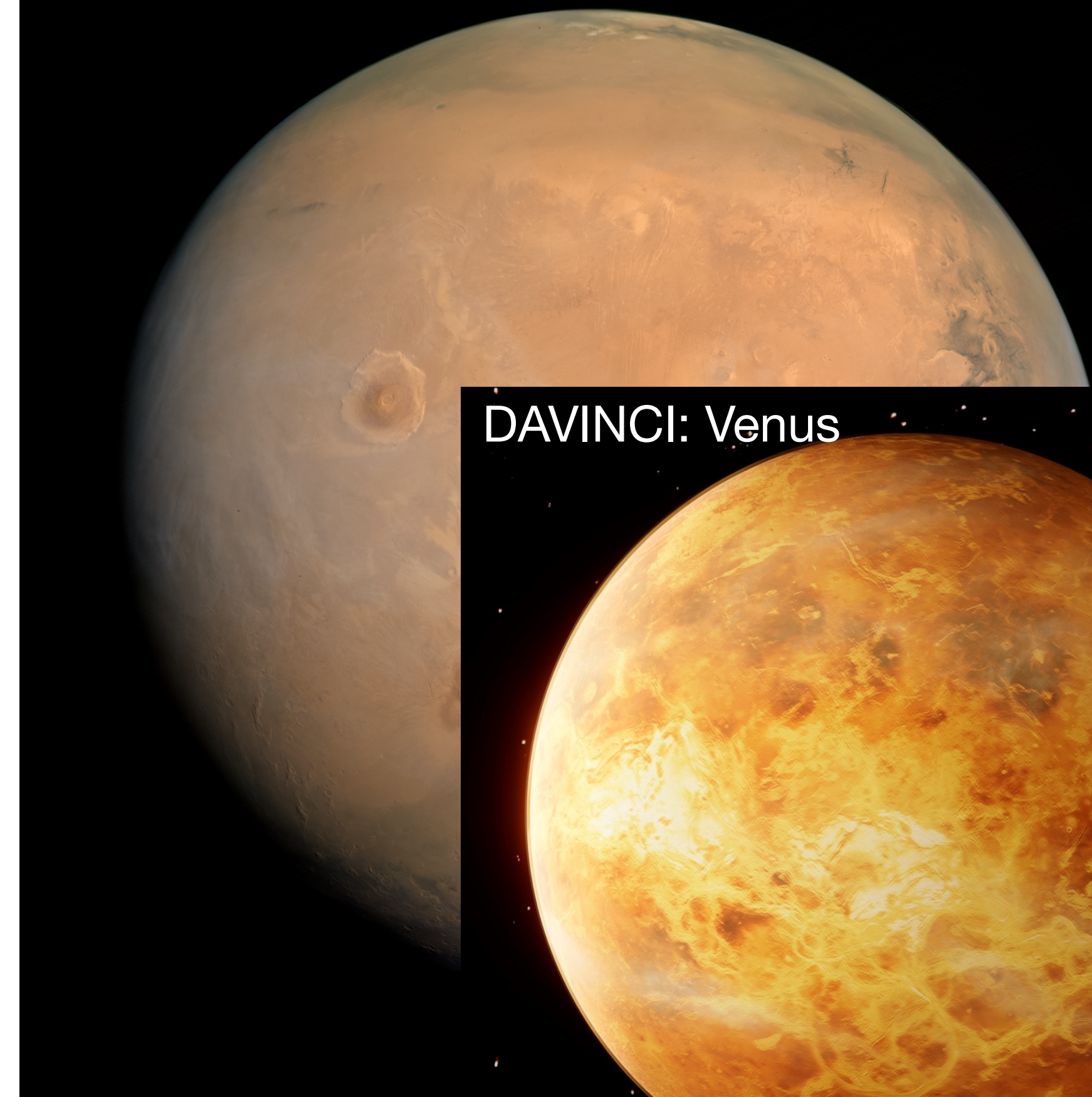


The New York Times

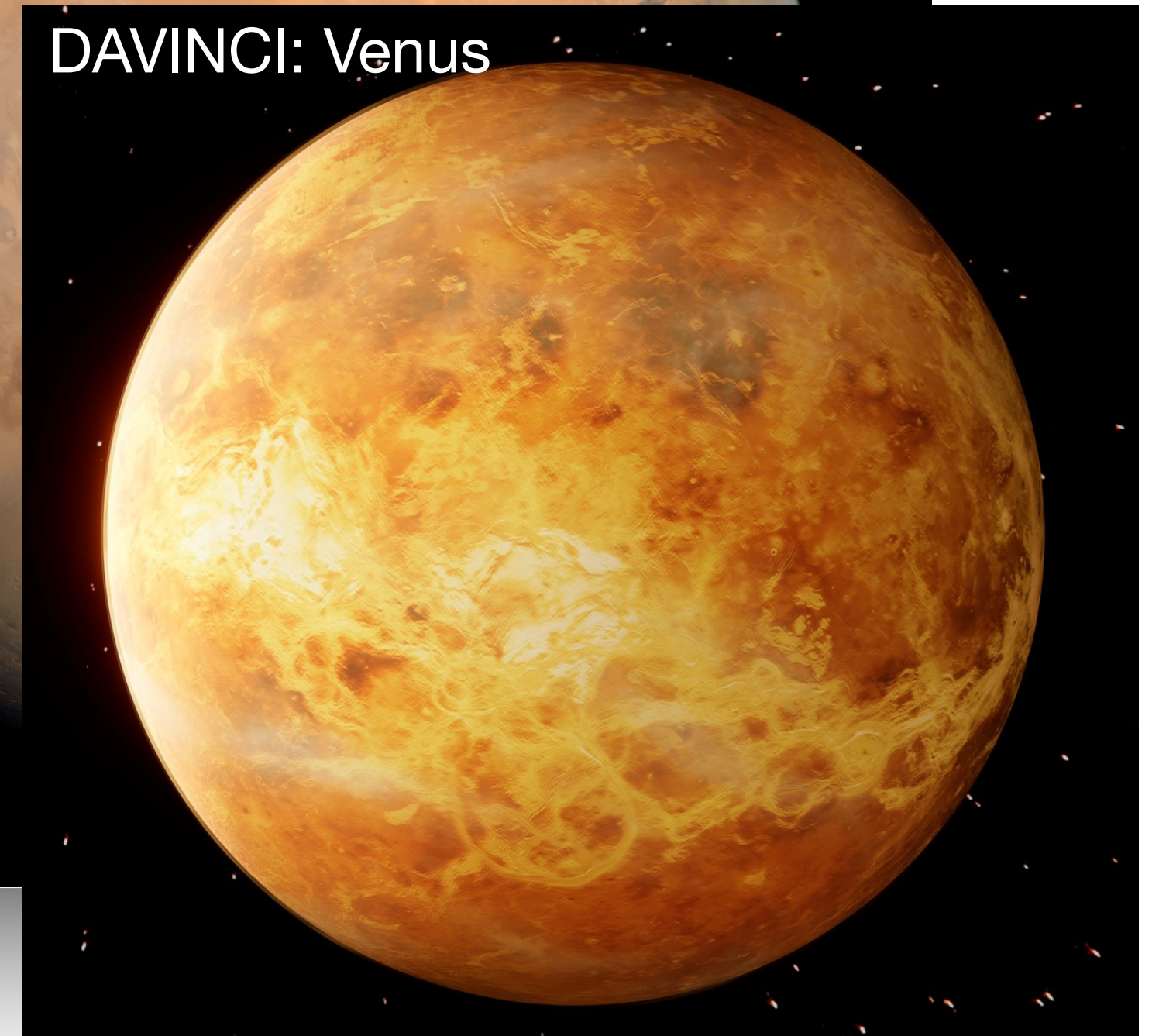
Lurking Inside an Asteroid: Life's Ingredients

Scientists studying samples that NASA collected from the asteroid Bennu found a wide assortment of organic molecules that shed light on how life arose.

Future NASA mission: MARS



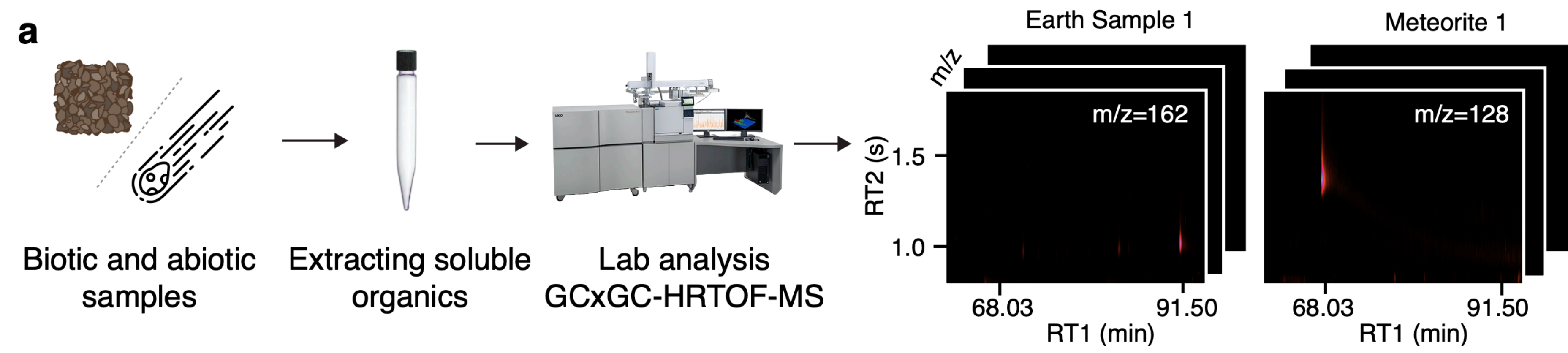
DAVINCI: Venus



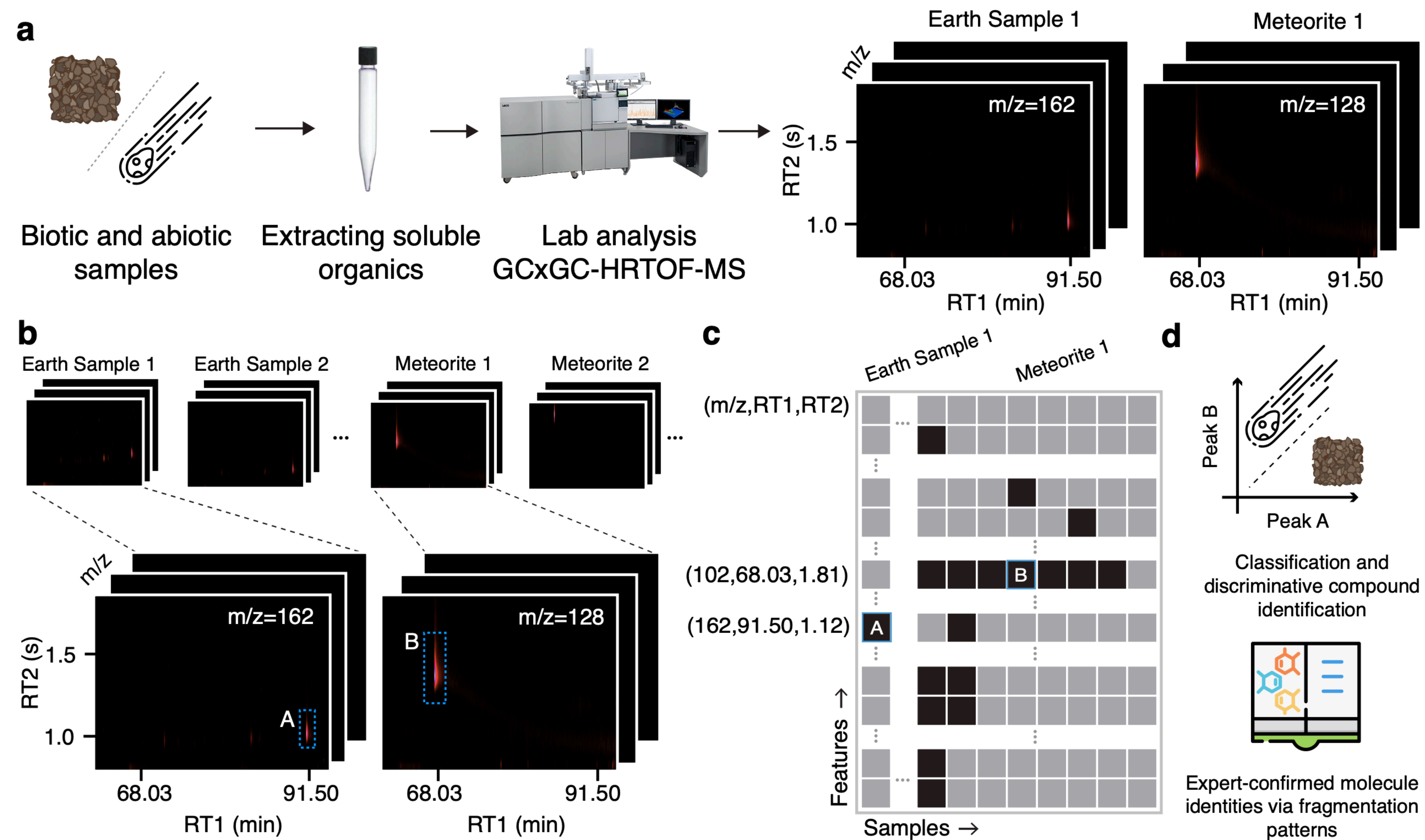
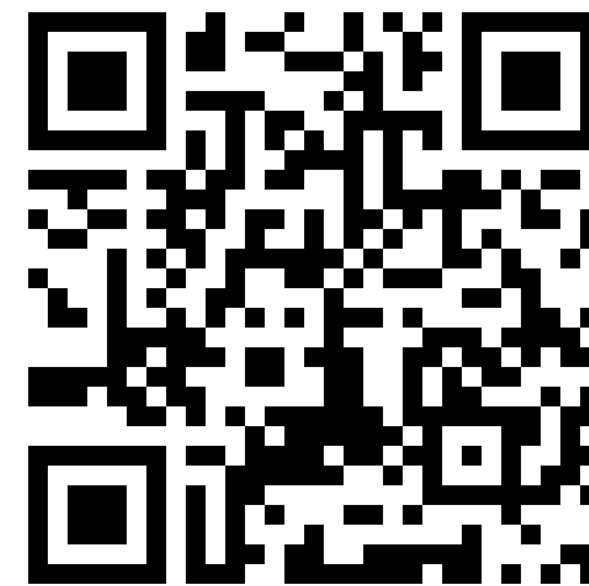
Say what the problem is?

Life is complex; current features are biased, need for AI

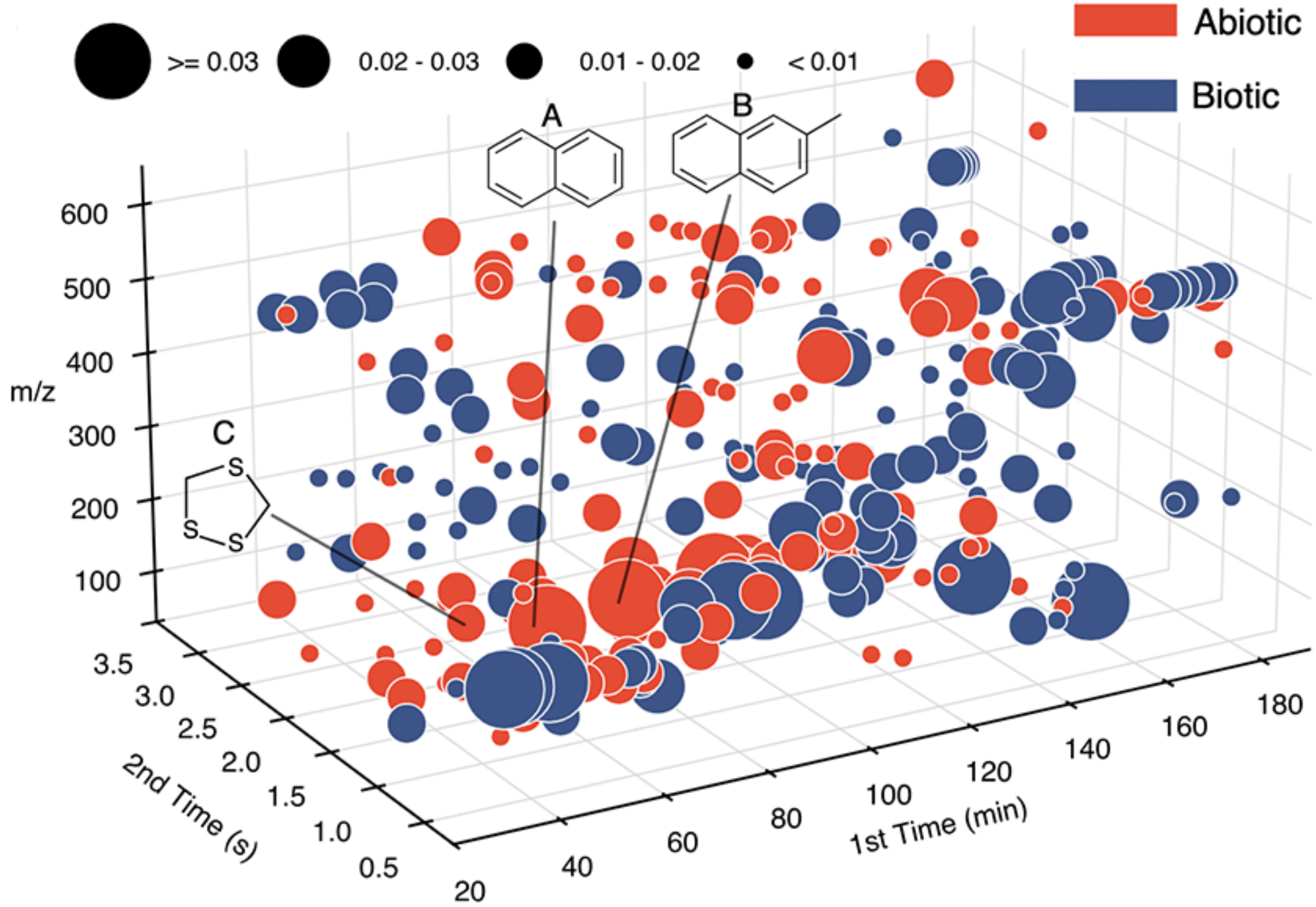
LifeTracer



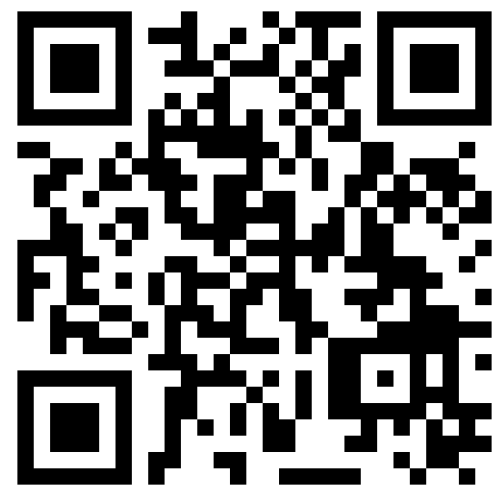
LifeTracer

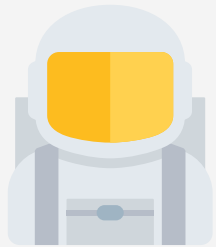


Discovered biosignatures by LifeTracer



AstroAgents



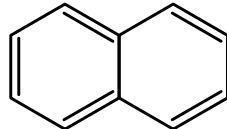


AstroAgents



Add relevant research papers

Mass Spectrometry Data:

ID	M/Z	RTs	Compound	Samples
1	128.17	(68 min, 1.2s)	 Naphthalene	Orgueil Lignite

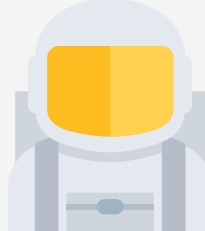
Instructions:

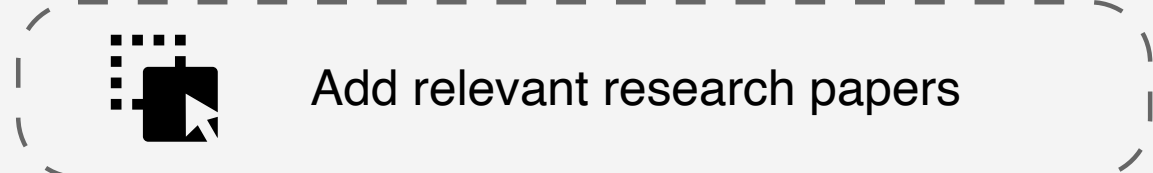
I have conducted an experiment using GCxGC to analyze biotic (terrestrial) and abiotic (meteorite) samples.

I want you to analyze the table and come up with interesting conclusions and hypotheses. Note that your hypotheses should be novel and specific.

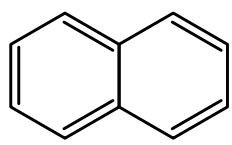
AstroAgents v1



**AstroAgents**



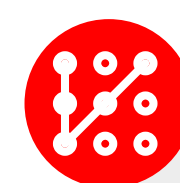
Mass Spectrometry Data:

ID	M/Z	RTs	Compound	Samples
1	128.17	(68 min, 1.2s)	 Naphthalene	Orgueil Lignite
⋮				

Instructions:

I have conducted an experiment using GCxGC to analyze biotic (terrestrial) and abiotic (meteorite) samples.

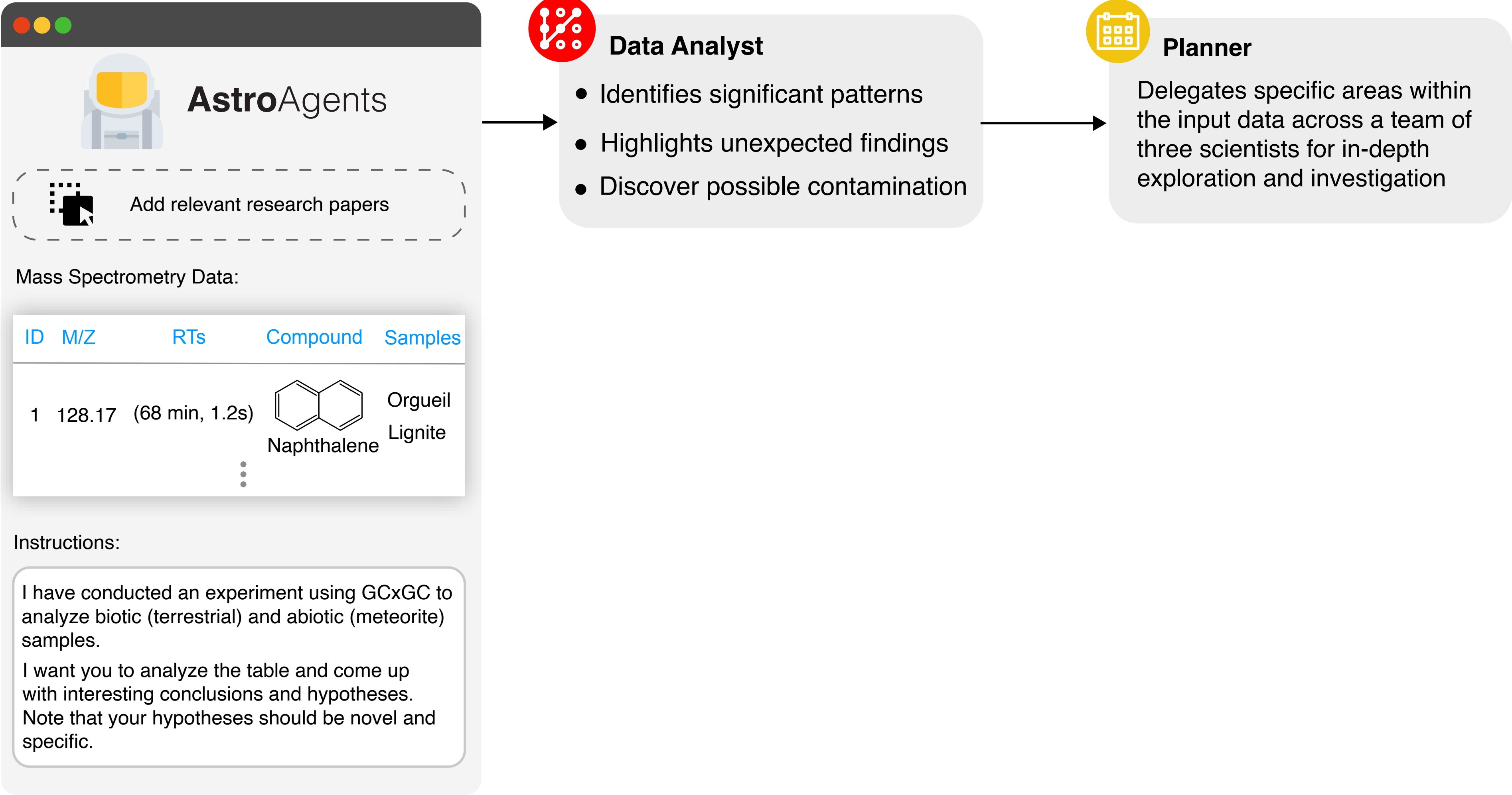
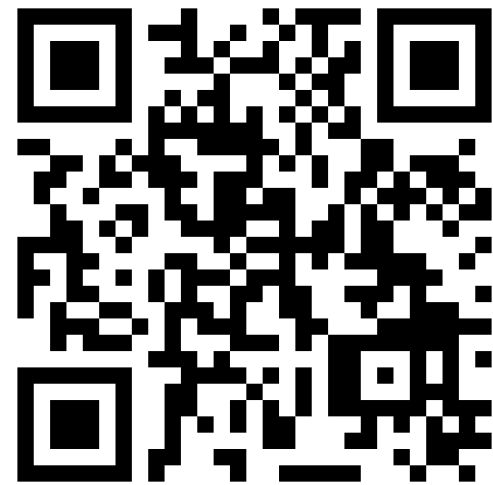
I want you to analyze the table and come up with interesting conclusions and hypotheses. Note that your hypotheses should be novel and specific.



Data Analyst

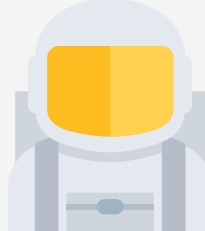
- Identifies significant patterns
- Highlights unexpected findings
- Discover possible contamination

AstroAgents v1



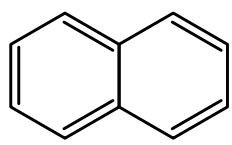
AstroAgents v1



**AstroAgents**

Add relevant research papers

Mass Spectrometry Data:

ID	M/Z	RTs	Compound	Samples
1	128.17	(68 min, 1.2s)	 Naphthalene	Orgueil Lignite
⋮				

Instructions:

I have conducted an experiment using GCxGC to analyze biotic (terrestrial) and abiotic (meteorite) samples.
I want you to analyze the table and come up with interesting conclusions and hypotheses.
Note that your hypotheses should be novel and specific.



Data Analyst

- Identifies significant patterns
- Highlights unexpected findings
- Discover possible contamination



Planner

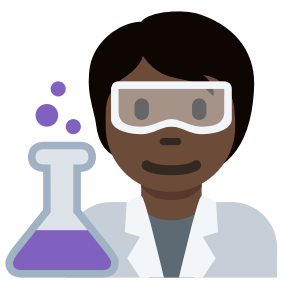
Delegates specific areas within the input data across a team of three scientists for in-depth exploration and investigation

Analyze PAH distribution patterns in meteorites, focus on IDs 1, 2 ...



Scientist 1

Investigate sulfur-containing and oxygen-containing compounds (IDs 14, 27, 30, 44)...



Scientist 2

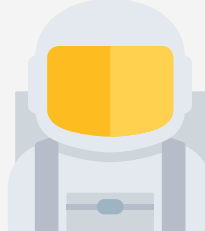
Study the distribution patterns of complex alkylated aromatics and their derivatives (IDs 15, 23) ...



Scientist 3

AstroAgents



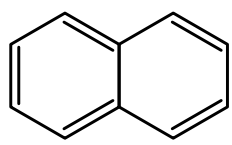


AstroAgents



Add relevant research papers

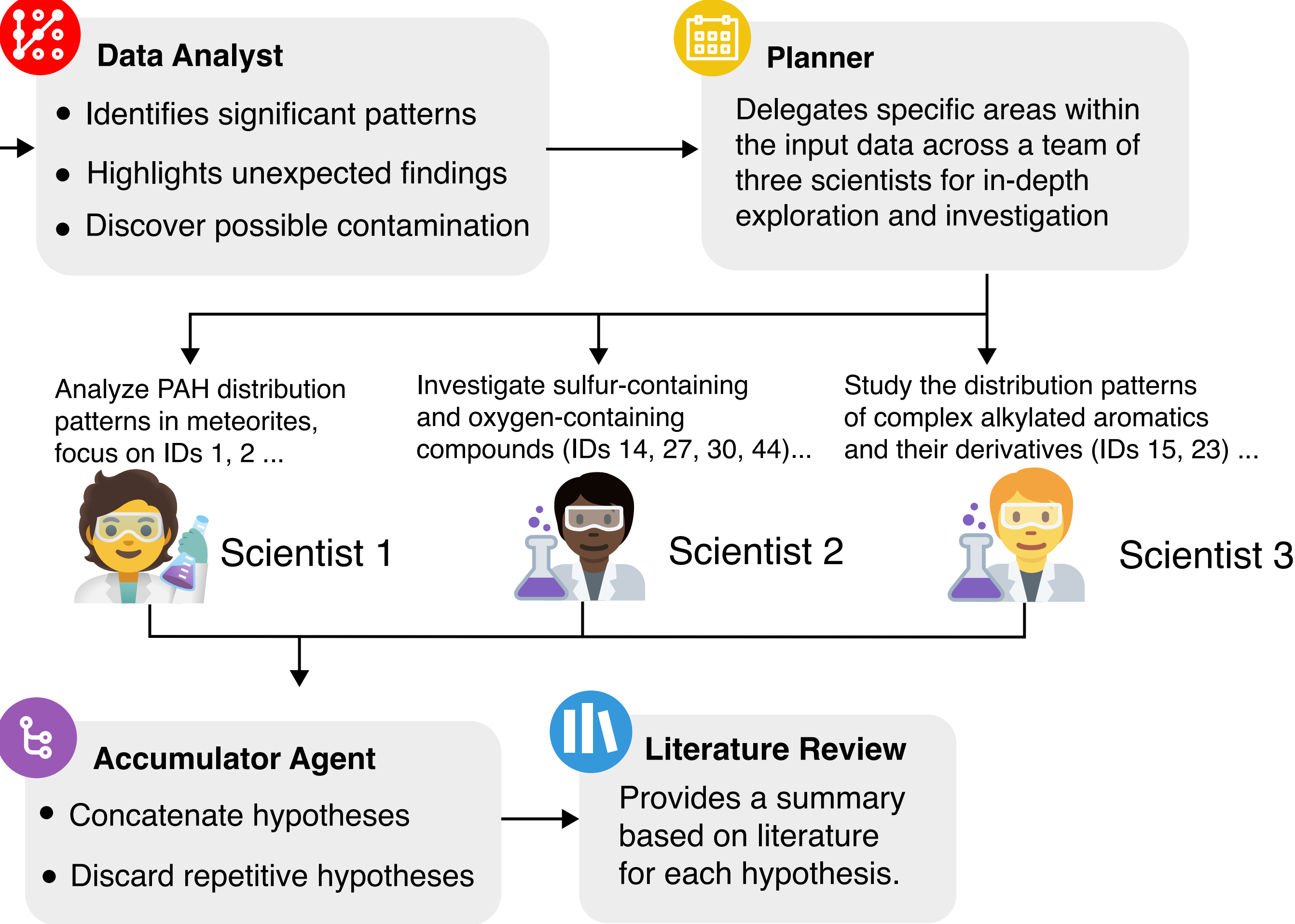
Mass Spectrometry Data:

ID	M/Z	RTs	Compound	Samples
1	128.17	(68 min, 1.2s)	 Naphthalene	Orgueil Lignite
⋮				

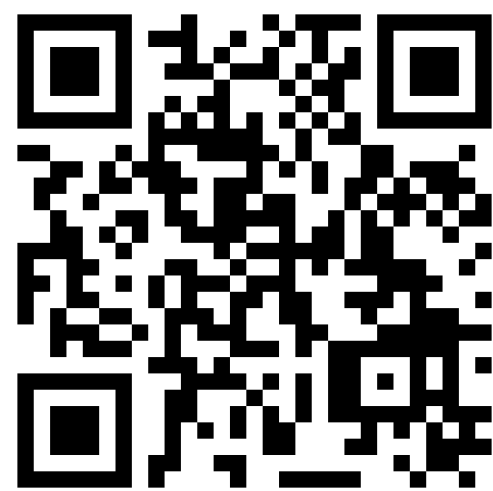
Instructions:

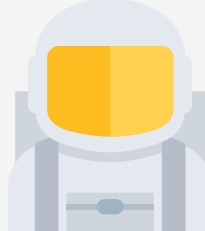
I have conducted an experiment using GCxGC to analyze biotic (terrestrial) and abiotic (meteorite) samples.

I want you to analyze the table and come up with interesting conclusions and hypotheses. Note that your hypotheses should be novel and specific.



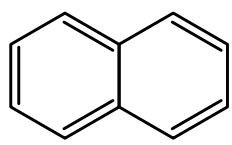
AstroAgents v1



**AstroAgents**

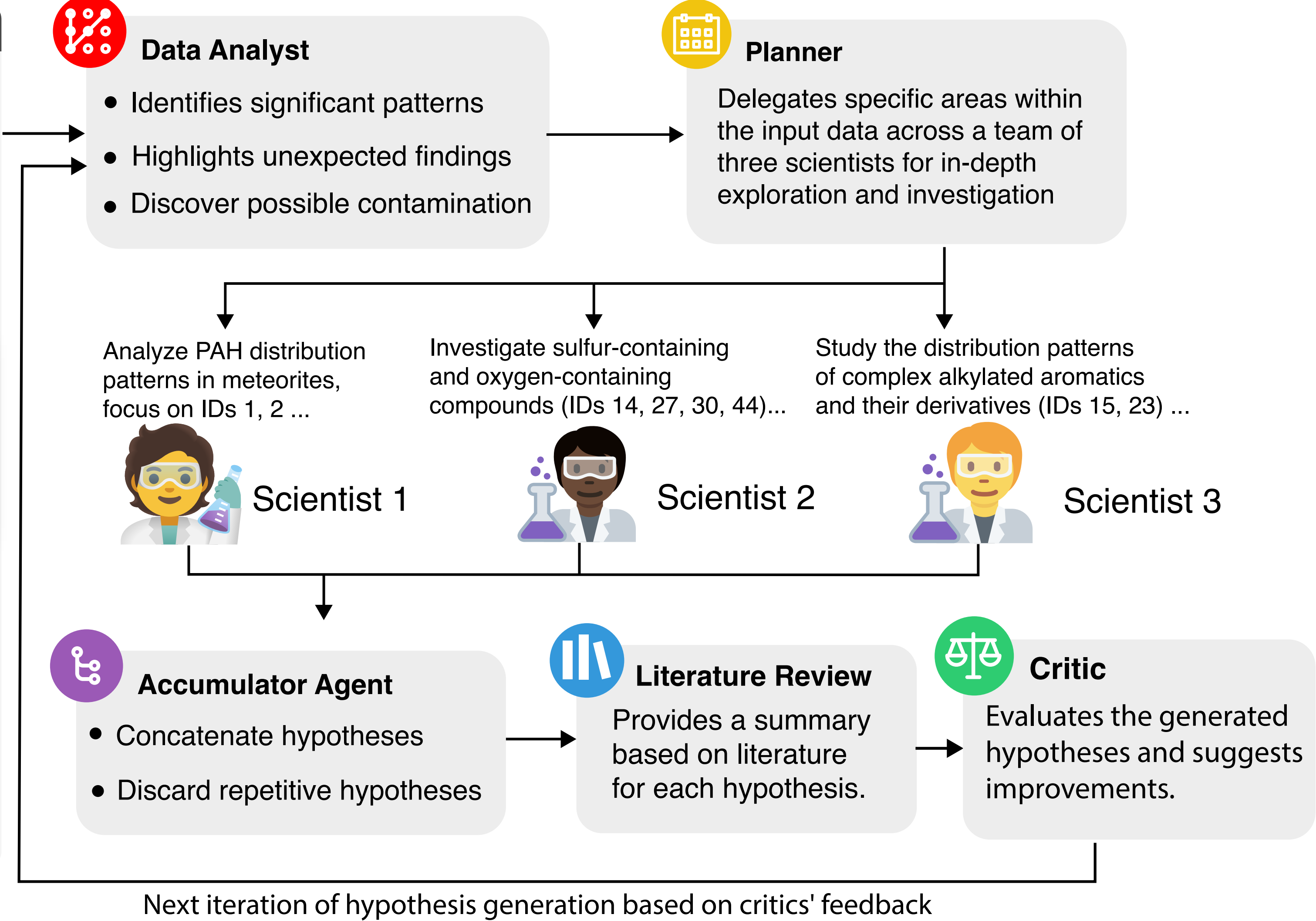
Add relevant research papers

Mass Spectrometry Data:

ID	M/Z	RTs	Compound	Samples
1	128.17	(68 min, 1.2s)	 Naphthalene	Orgueil Lignite
⋮				

Instructions:

I have conducted an experiment using GCxGC to analyze biotic (terrestrial) and abiotic (meteorite) samples.
I want you to analyze the table and come up with interesting conclusions and hypotheses. Note that your hypotheses should be novel and specific.



Human expert evaluations



- **Claude Sonnet 3.5** (48 hypotheses)
 - fewer logical errors
 - stronger consistency with the literature
 - scored higher on **clarity**
- **Gemini 2.0 Flash** (101 hypotheses)
 - higher **novelty** scores
 - but more logical errors

Criteria	Claude Sonnet 3.5	Gemini 2.0 Flash
Novelty	2.75 ± 0.75	4.26 ± 1.87
Consistency with the literature	7.60 ± 1.91	6.19 ± 2.88
Clarity and precision	7.20 ± 2.30	5.92 ± 2.86
Empirical Support	6.75 ± 2.63	5.79 ± 2.86
Scope & Generalizability	7.60 ± 1.91	6.01 ± 2.80
Predictive Power	7.60 ± 1.91	5.86 ± 2.68
Overall Average	6.58 ± 1.74	5.67 ± 0.64

#	Statement	Key Datapoints	Evaluation Score
1	Gemini 2.0 Flash: The presence of 1H-Phenalen-1-one or 9H-Fluoren-9-one (ID 44, MW 180) exclusively in Orgueil and LEW 85311, and the presence of Biphenyl (ID 43, MW 154) also in the same meteorites, suggests a unique chemical environment or alteration history shared by these samples, potentially indicating a similar formation region within the early solar system. Given their related structures, this may indicate a similar source.	1H-Phenalen-1-one or 9H-Fluoren-9-one (ID 44, MW 180): Orgueil, LEW 85311; Biphenyl (ID 43, MW 154): Orgueil, LEW 85311.	Novelty: 7/10 Literature: 9/10 Clarity/Precision: 9/10 Empirical Support: 9/10 Generalizability: 9/10 Predictive Power: 8/10

5	Claude 3.5 Sonnet: The detection of possible terpenes exclusively in soil samples indicates that complex branched isoprenoid structures require enzymatic biosynthesis and are not readily formed through abiotic processes in space, making them reliable biomarkers.	IDs 4, 17, and 18 (possible terpenes) were found only in soil samples (Iceland Soil, Atacama, Utah soil, GSFC soil)	Novelty: 3/10 Literature: 10/10 Clarity/Precision: 10/10 Empirical Support: 10/10 Generalizability: 10/10 Predictive Power: 10/10
---	---	---	---

ASTRA

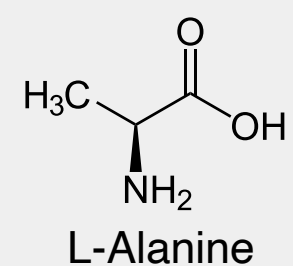


User task

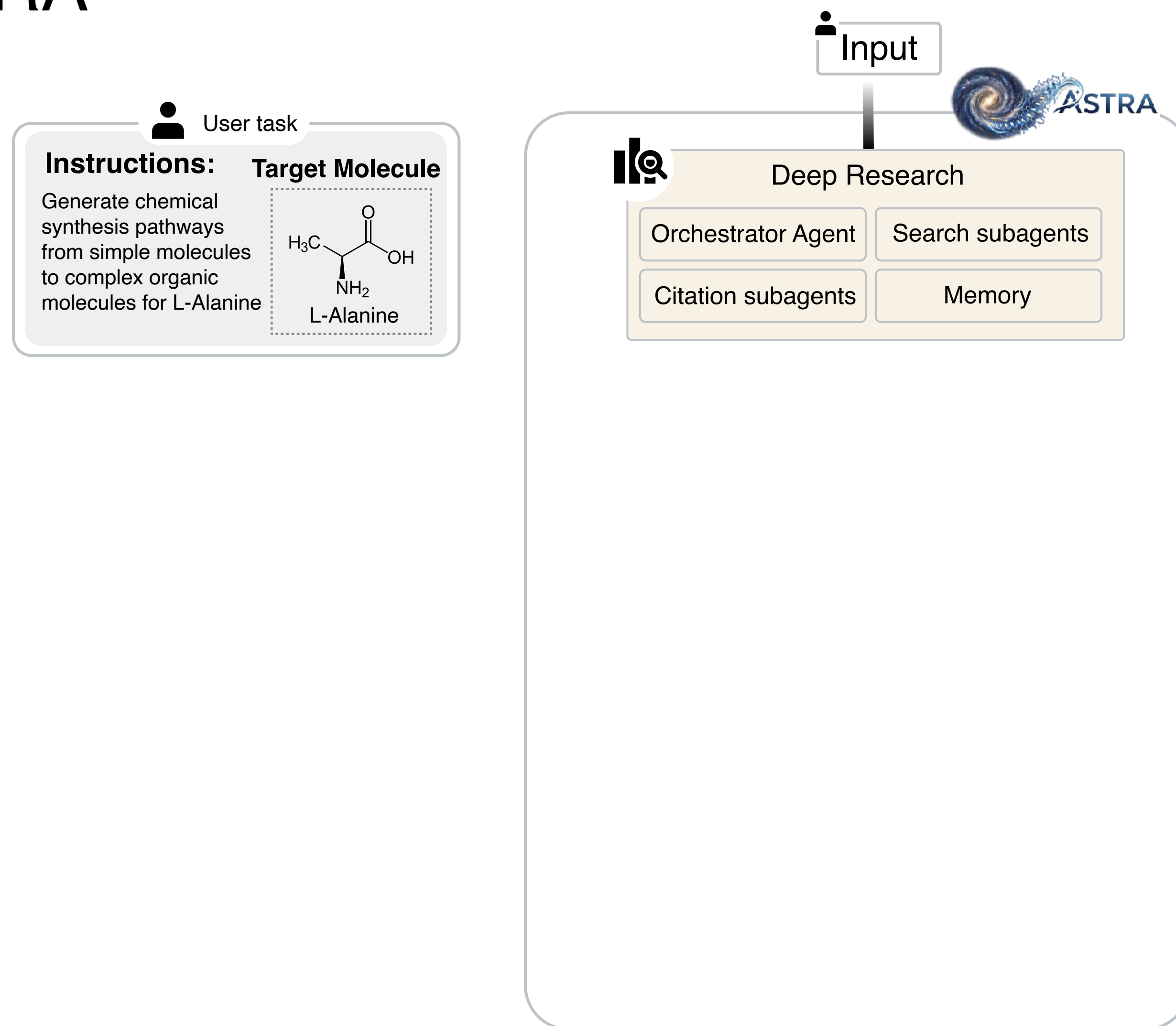
Instructions:

Generate chemical synthesis pathways from simple molecules to complex organic molecules for L-Alanine

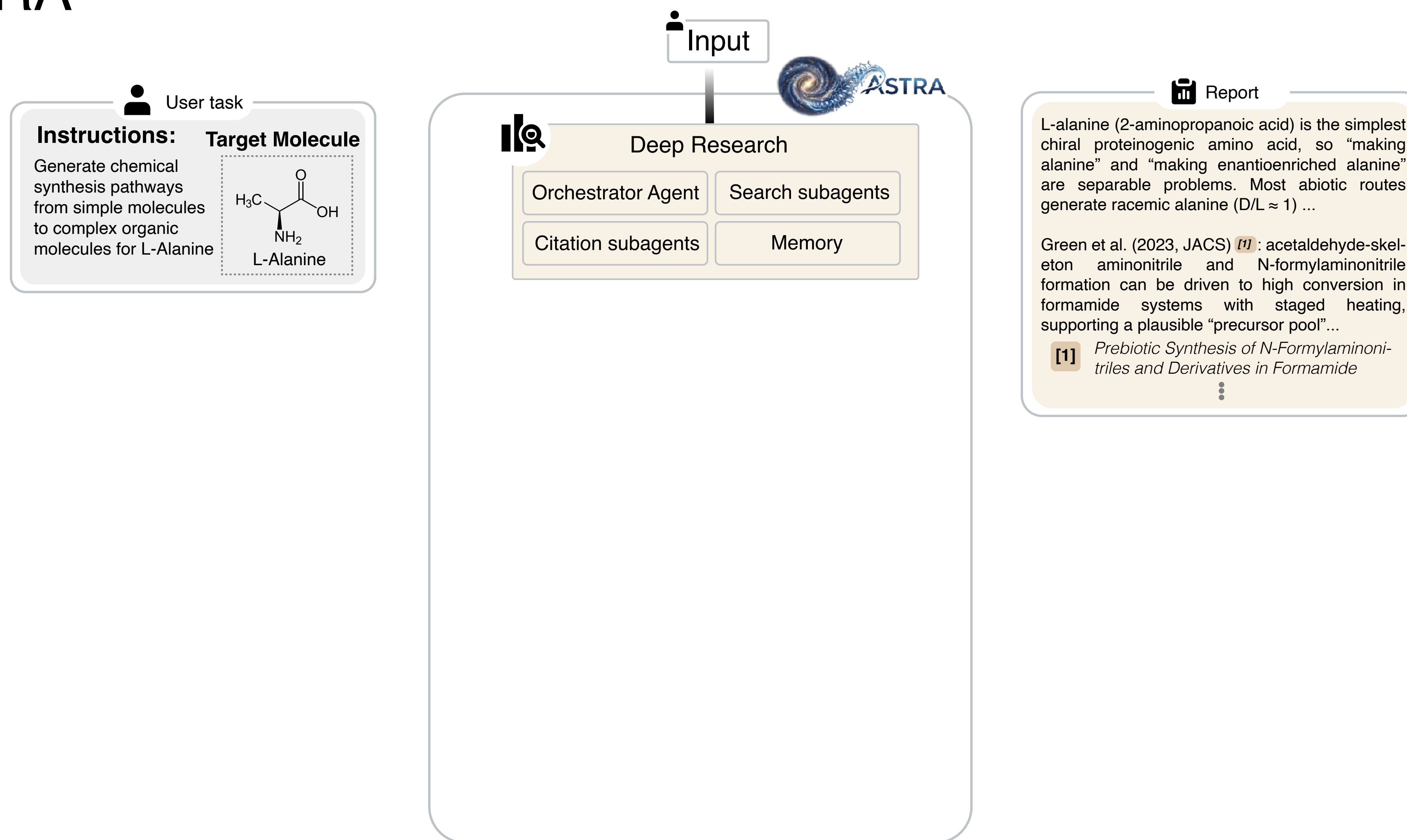
Target Molecule



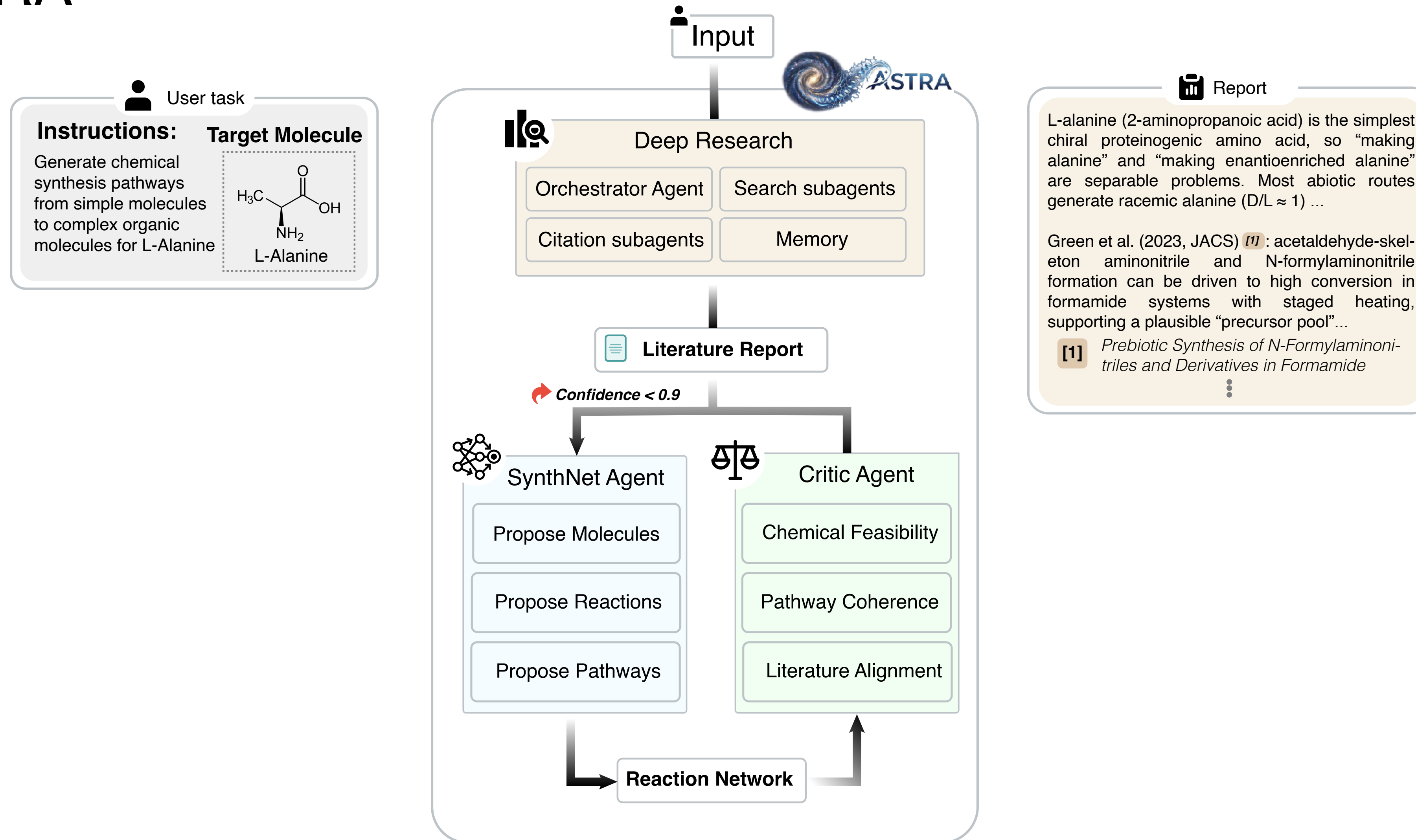
ASTRA



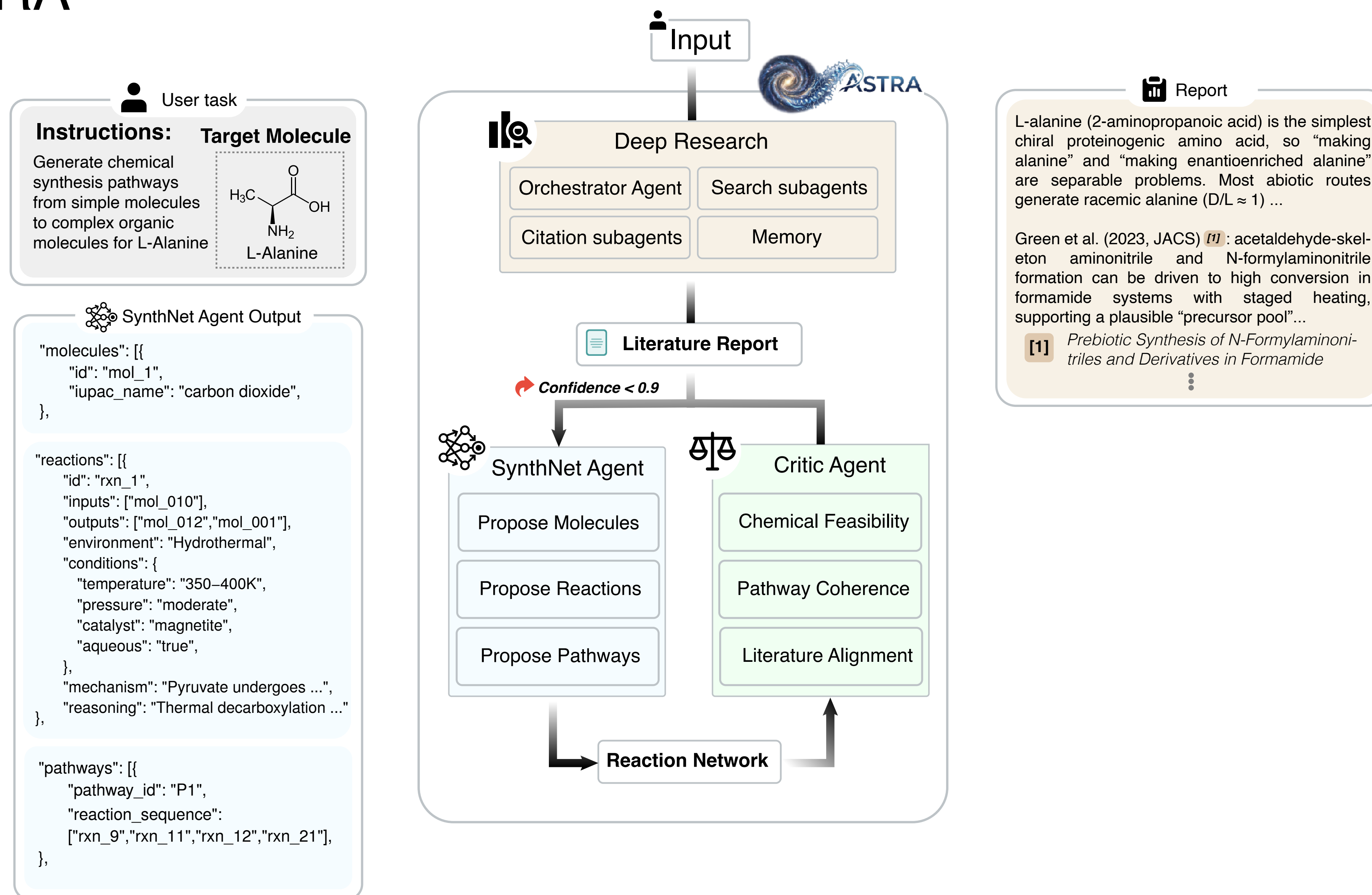
ASTRA



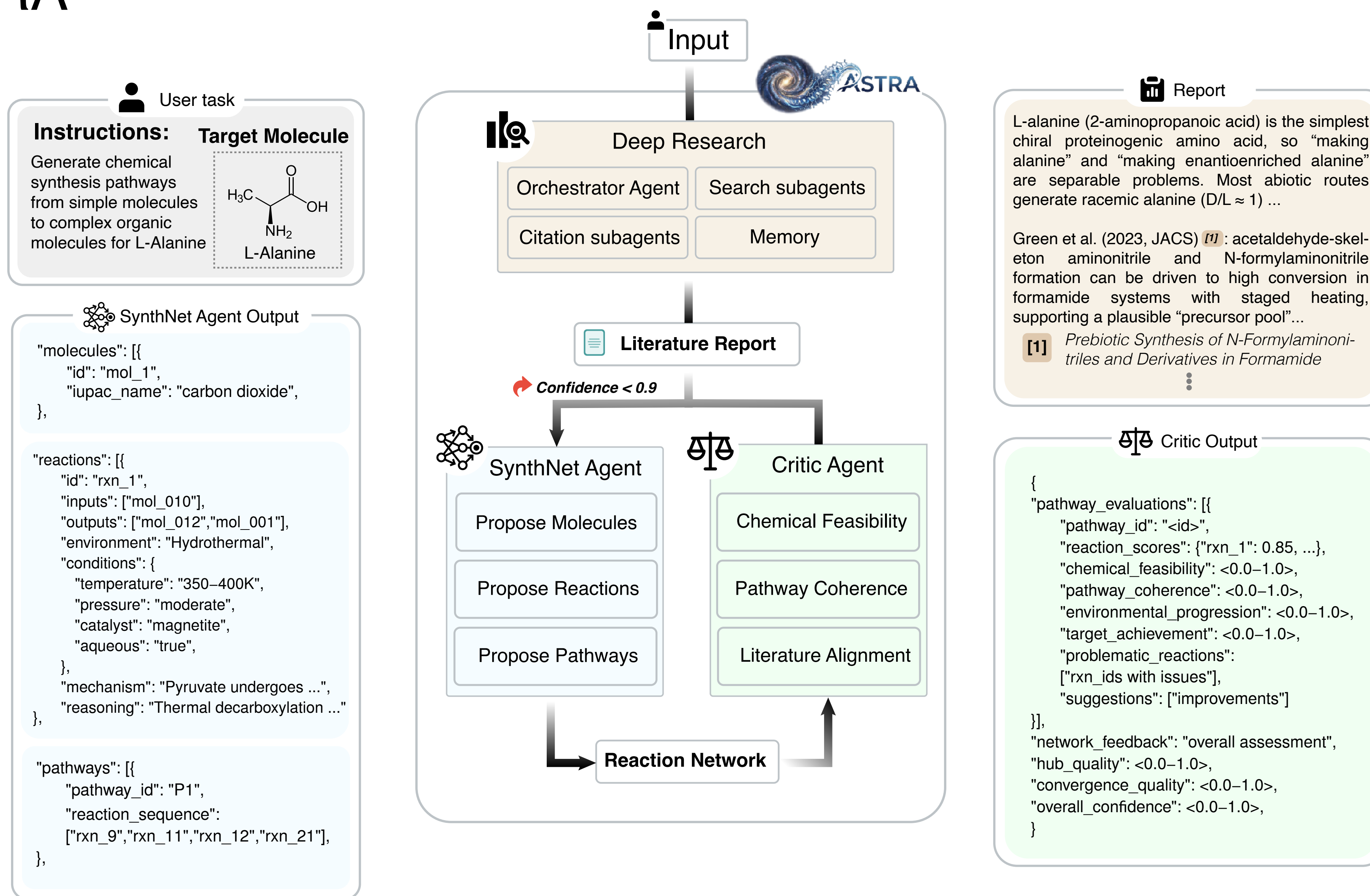
ASTRA



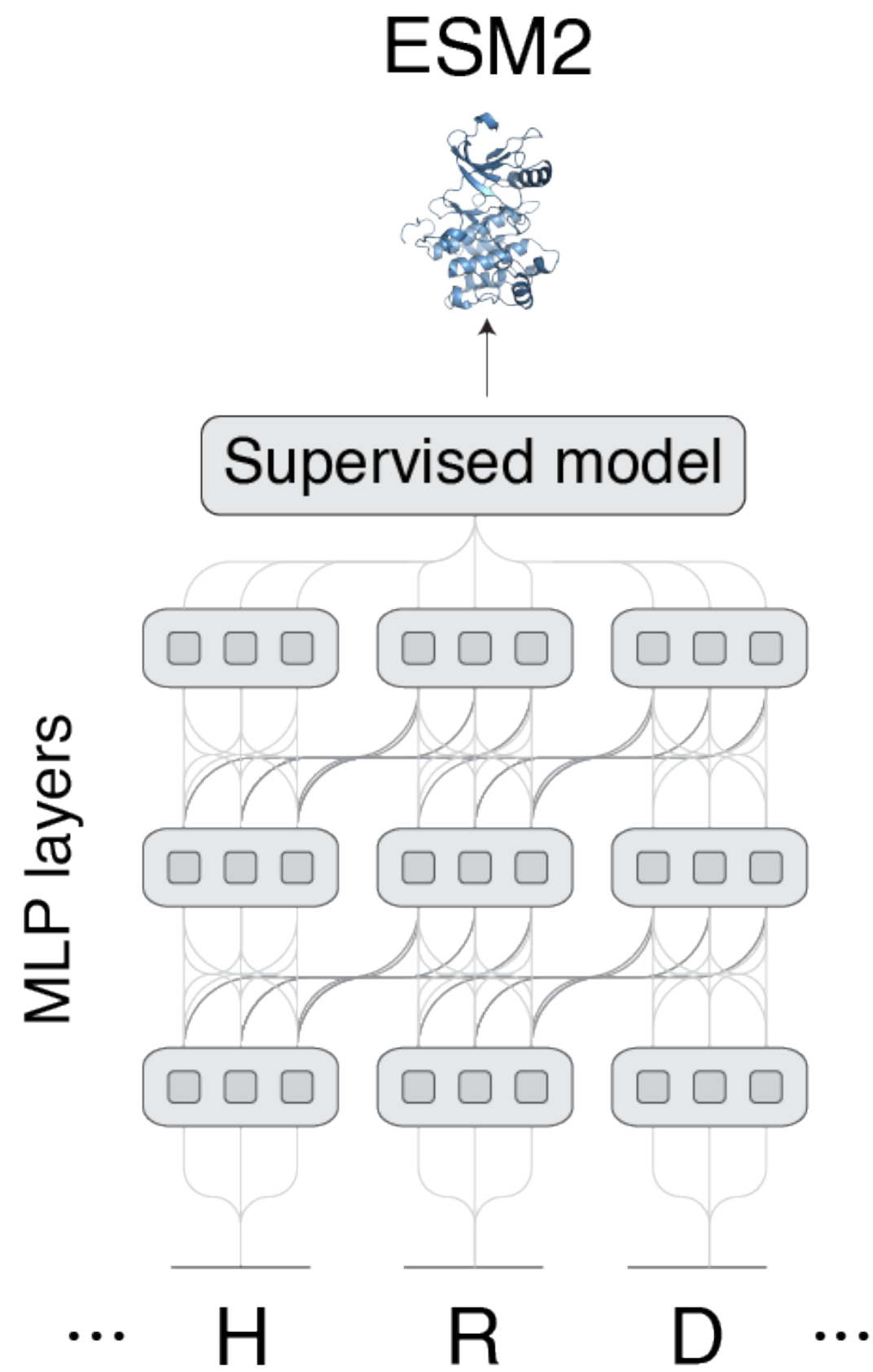
ASTRA



ASTRA



per-layer transcoders (PLTs) as a replacement model in ESM2



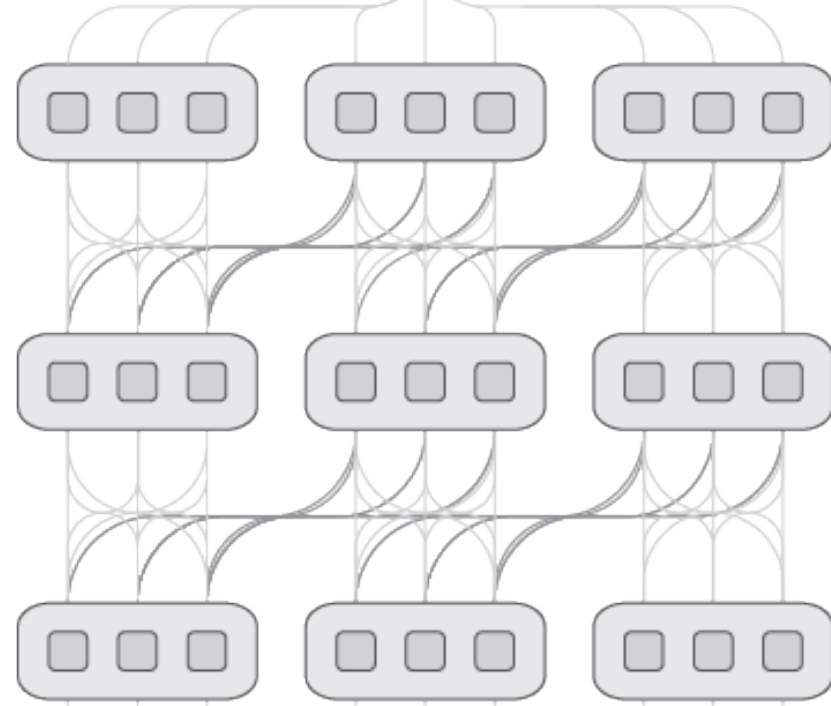
per-layer transcoders (PLTs) as a replacement model in ESM2

ESM2



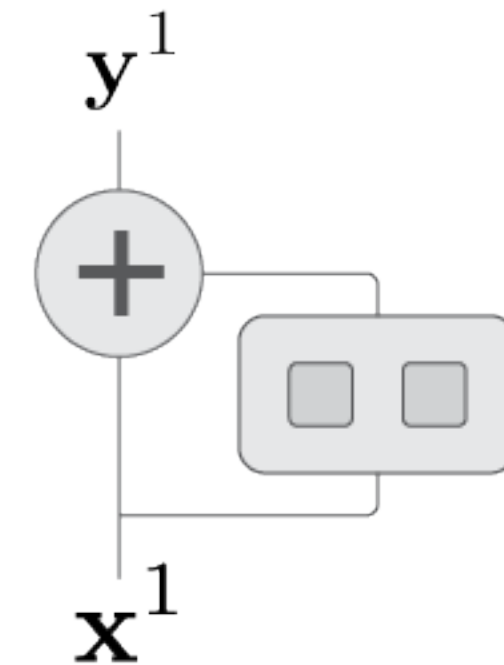
Supervised model

MLP layers



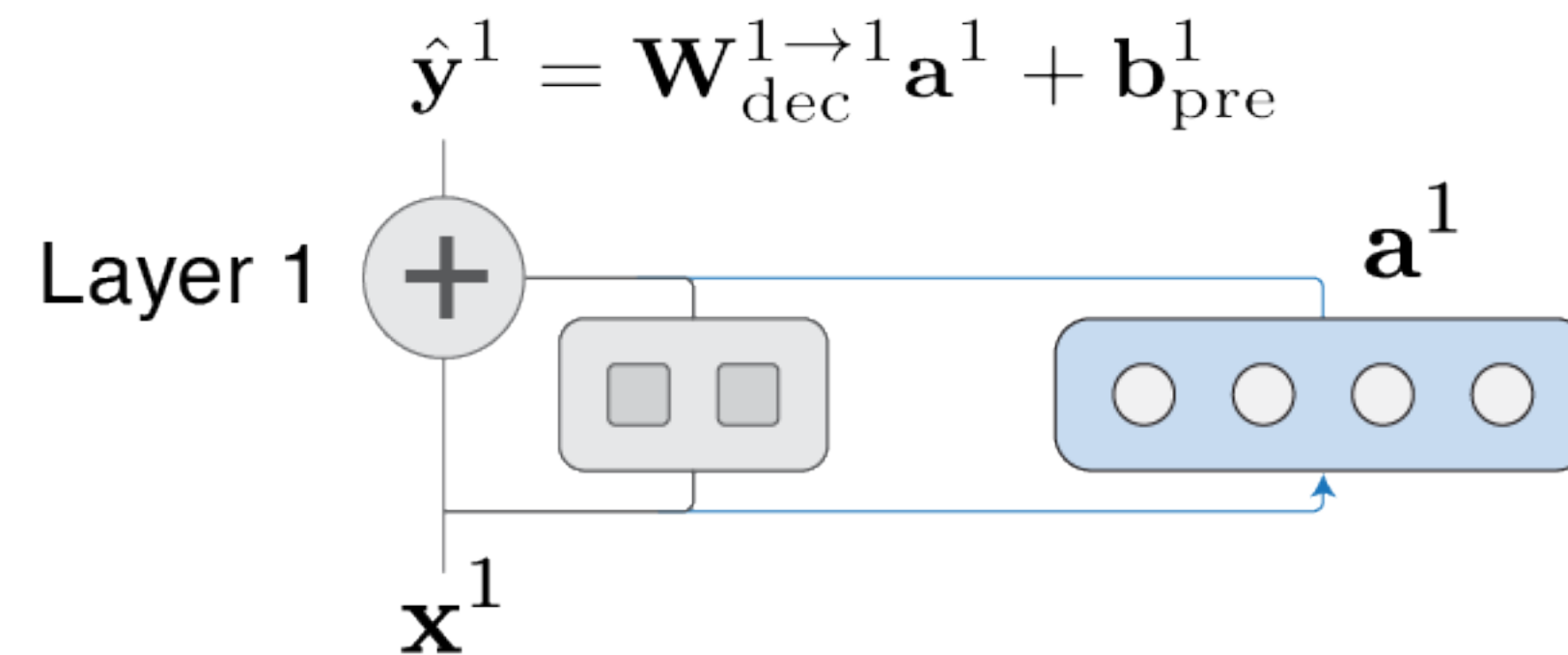
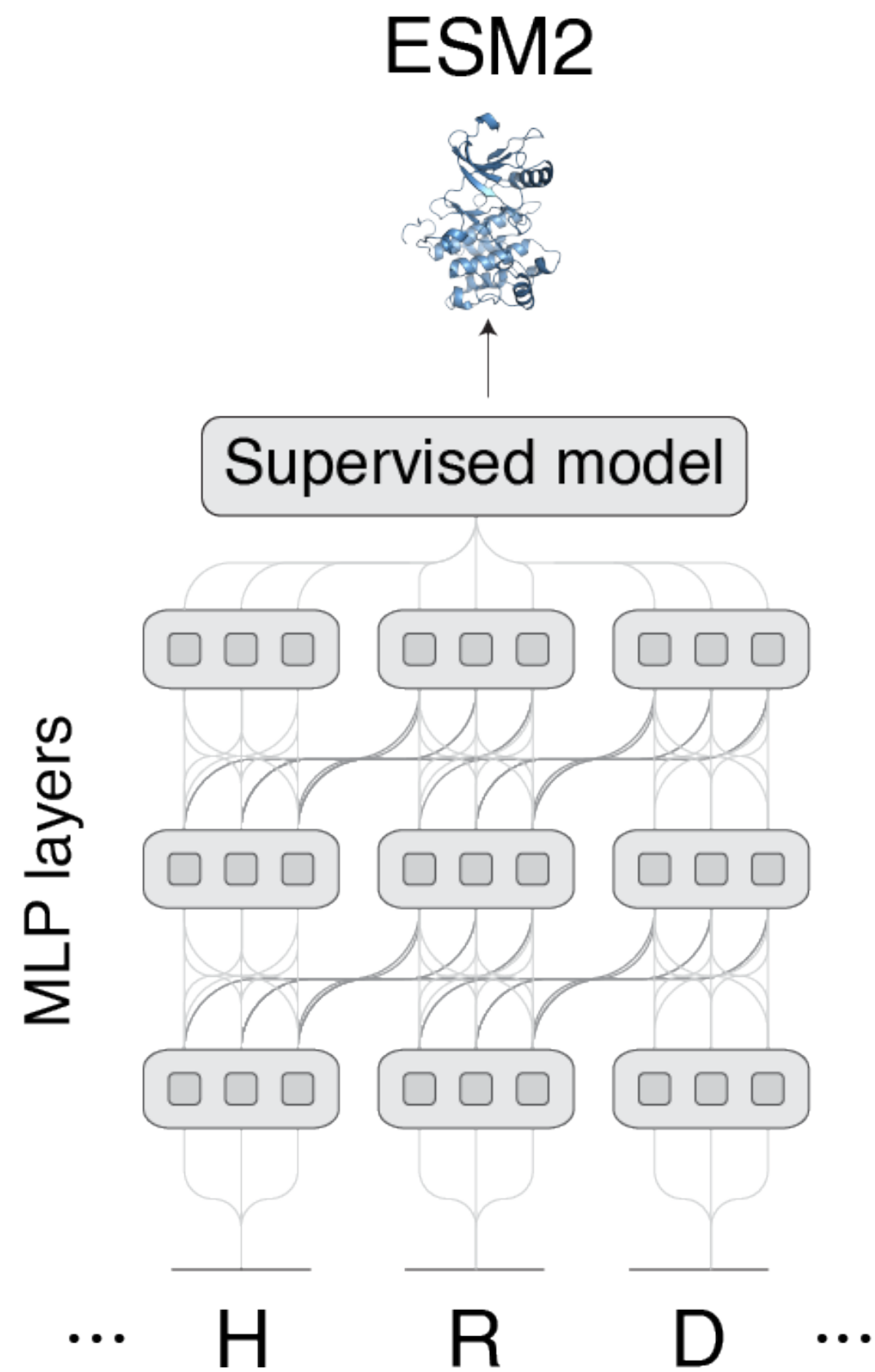
... H R D ...

Layer 1



per-layer transcoders (PLTs) as a replacement model in ESM2

$$\mathbf{a}^1 = \text{TopK}(\mathbf{W}_{\text{enc}}^1(\mathbf{x}^1 - \mathbf{b}_{\text{pre}}^1) + \mathbf{b}_{\text{enc}}^1)$$



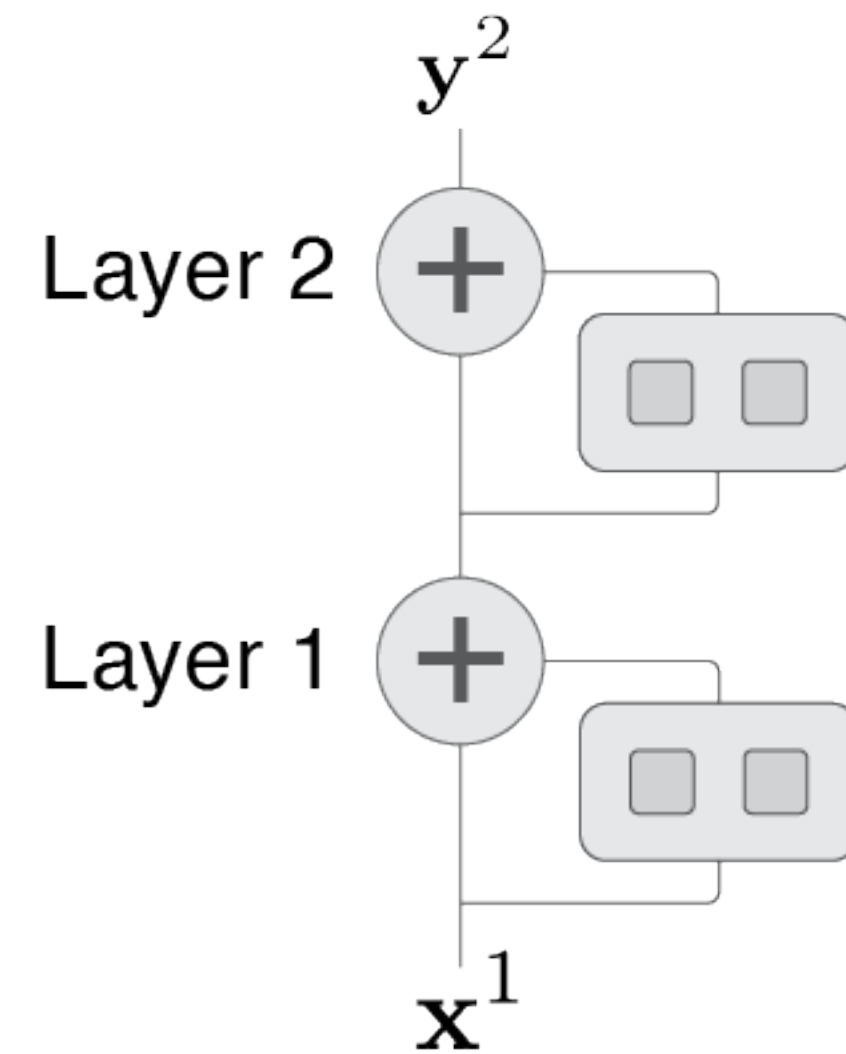
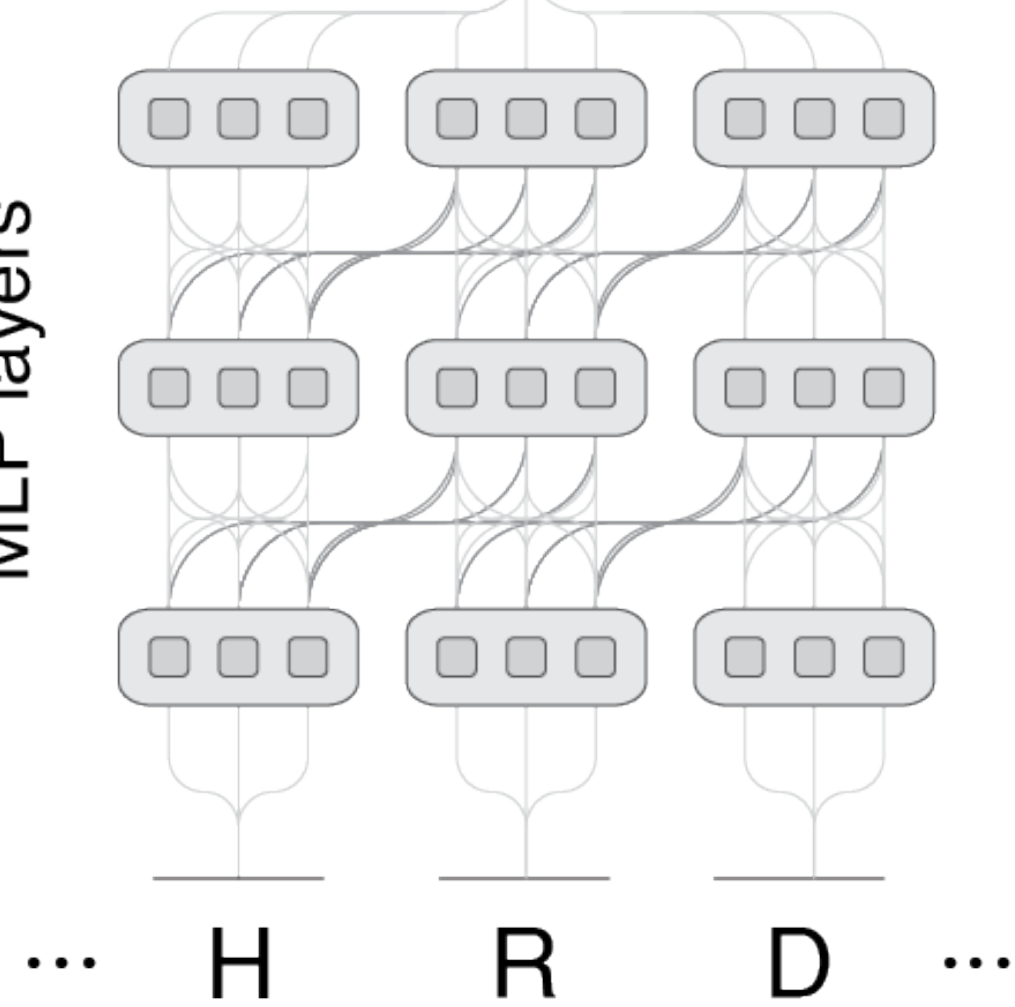
per-layer transcoders (PLTs) as a replacement model in ESM2

ESM2



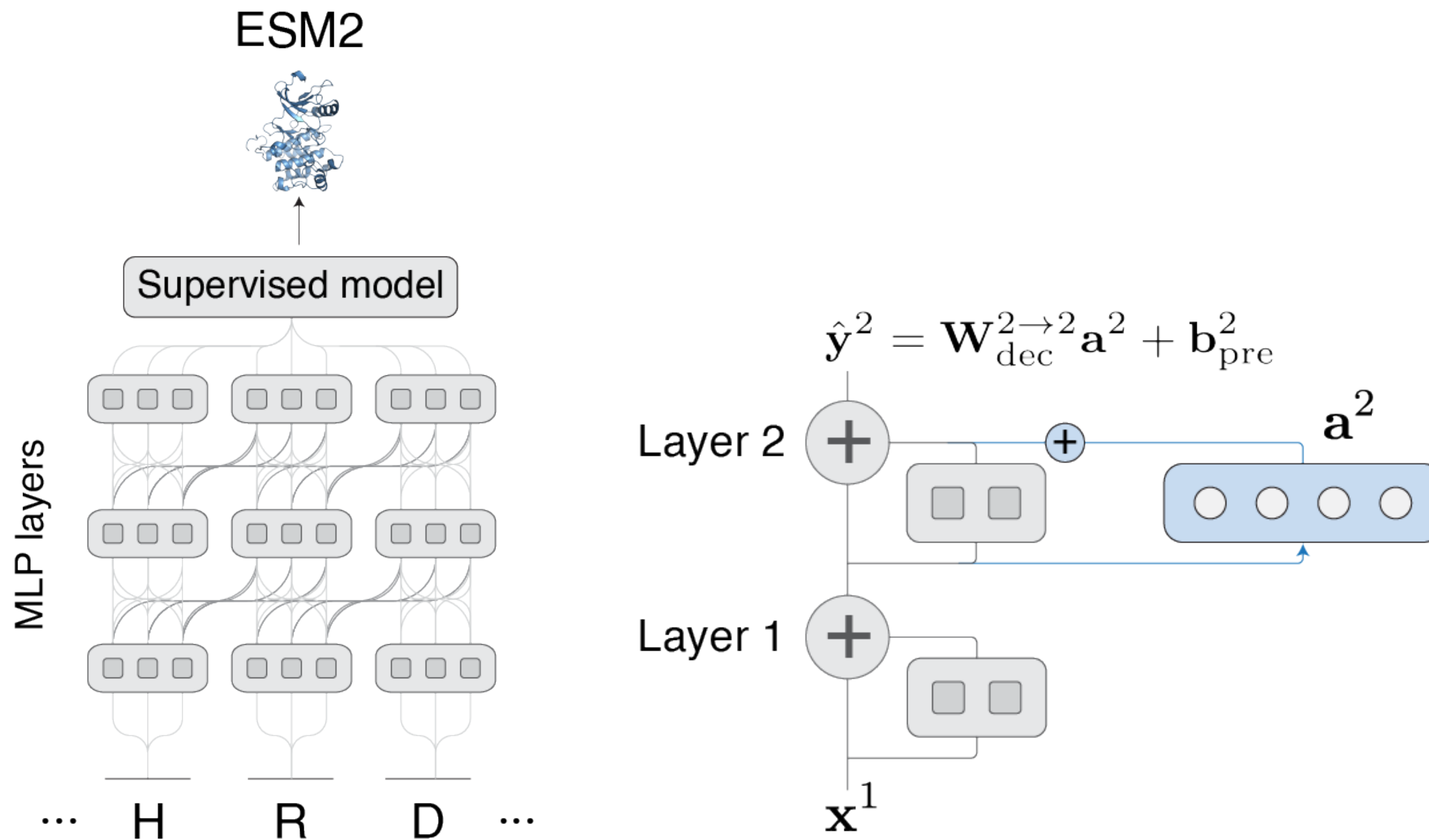
Supervised model

MLP layers

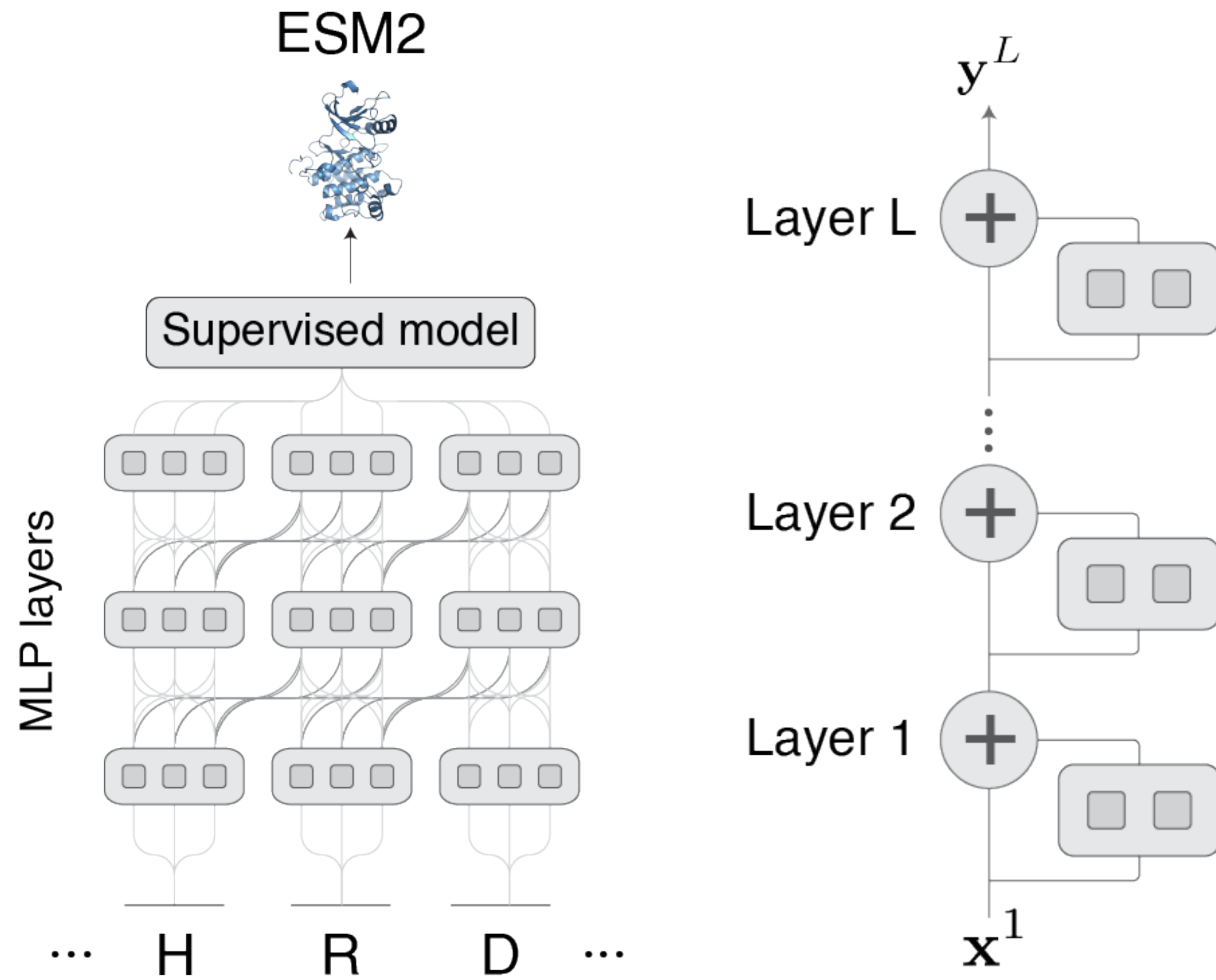


per-layer transcoders (PLTs) as a replacement model in ESM2

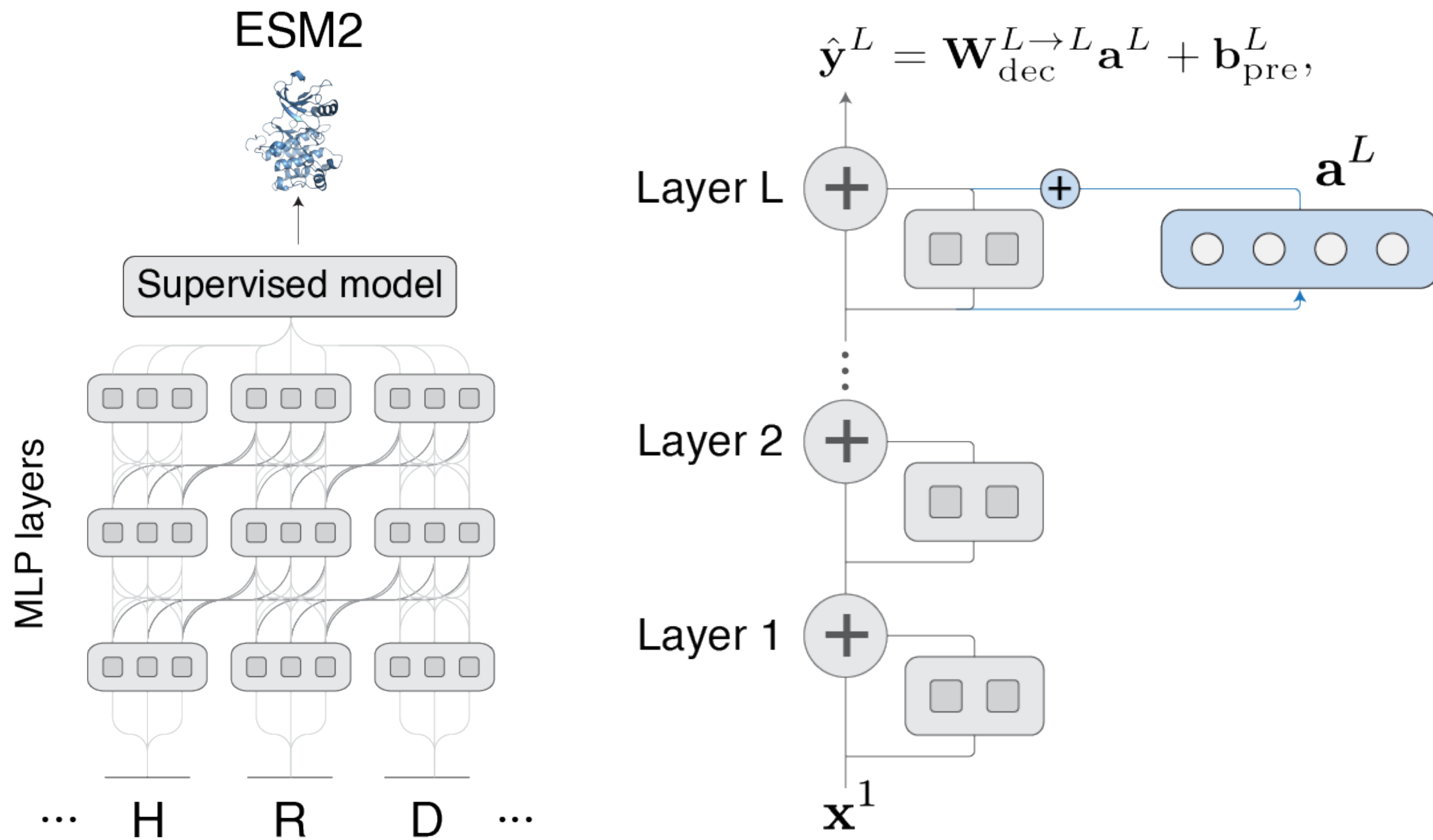
$$\mathbf{a}^2 = \text{TopK}(\mathbf{W}_{\text{enc}}^2(\mathbf{x}^2 - \mathbf{b}_{\text{pre}}^2) + \mathbf{b}_{\text{enc}}^2)$$



per-layer transcoders (PLTs) as a replacement model in ESM2



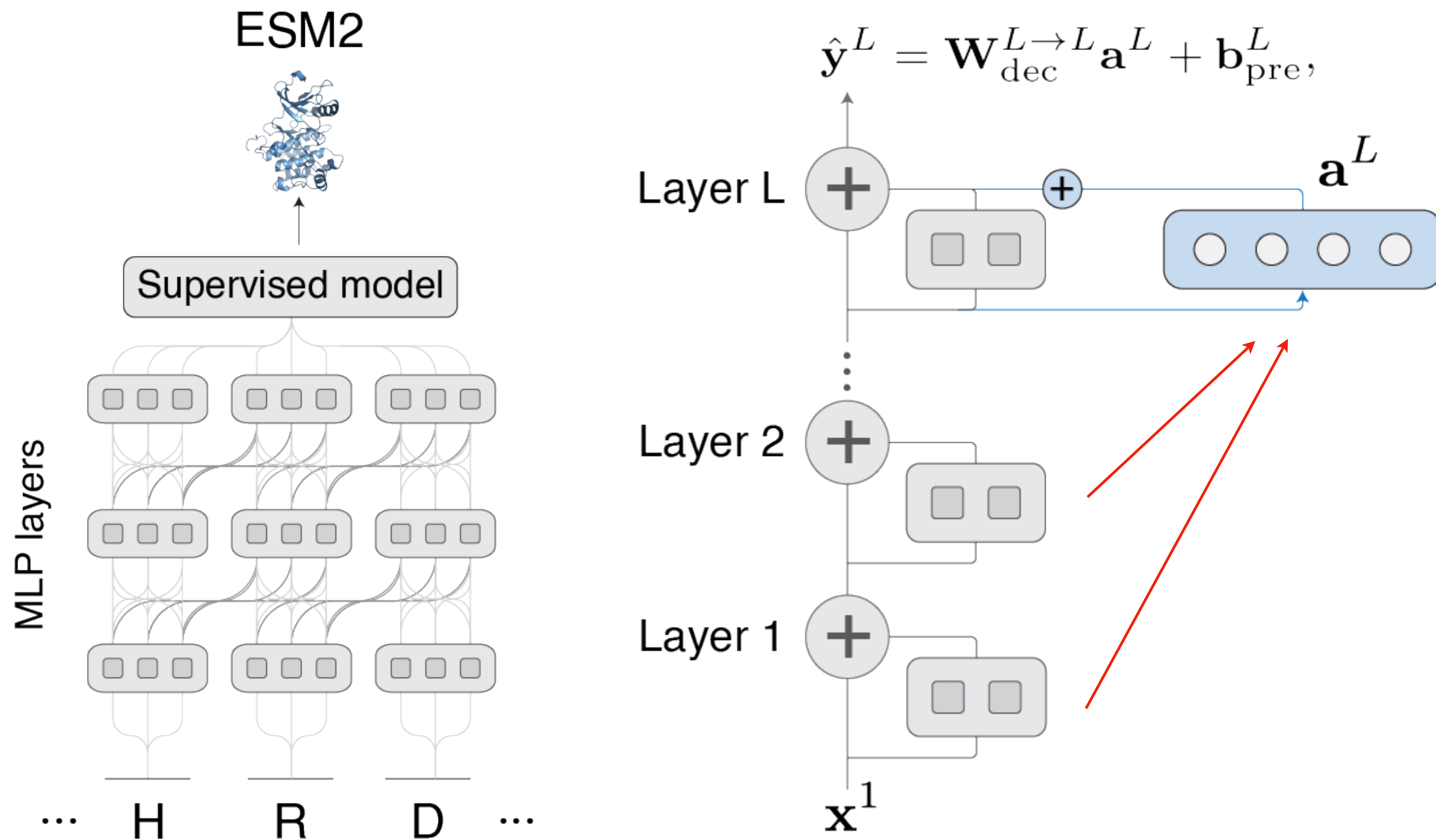
per-layer transcoders (PLTs) as a replacement model in ESM2



$$\mathbf{a}^L = \text{TopK}(\mathbf{W}_{\text{enc}}^L(\mathbf{x}^L - \mathbf{b}_{\text{pre}}^L) + \mathbf{b}_{\text{enc}}^L)$$



per-layer transcoders (PLTs) as a replacement model in ESM2



$$\mathbf{a}^L = \text{TopK}(\mathbf{W}_{\text{enc}}^L(\mathbf{x}^L - \mathbf{b}_{\text{pre}}^L) + \mathbf{b}_{\text{enc}}^L)$$

Approximation is
inherently **local**



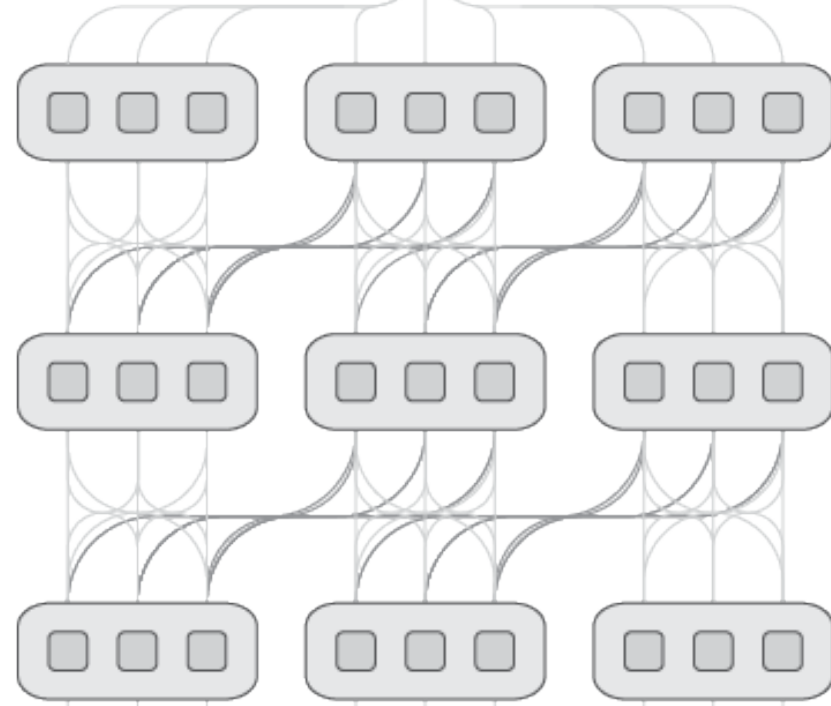
per-layer transcoders (PLTs) as a replacement model in ESM2

ESM2



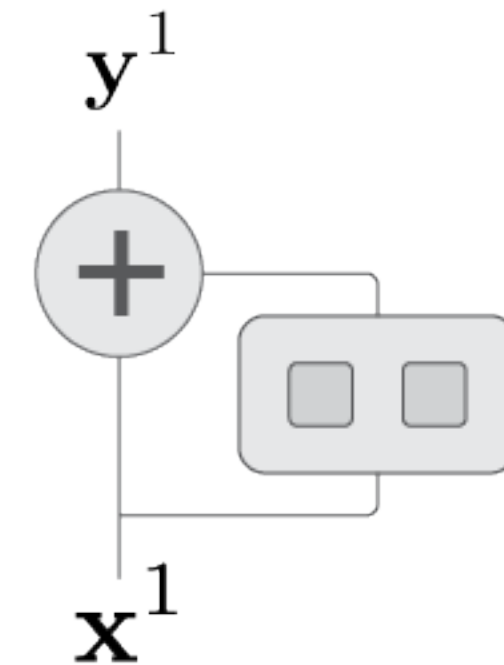
Supervised model

MLP layers



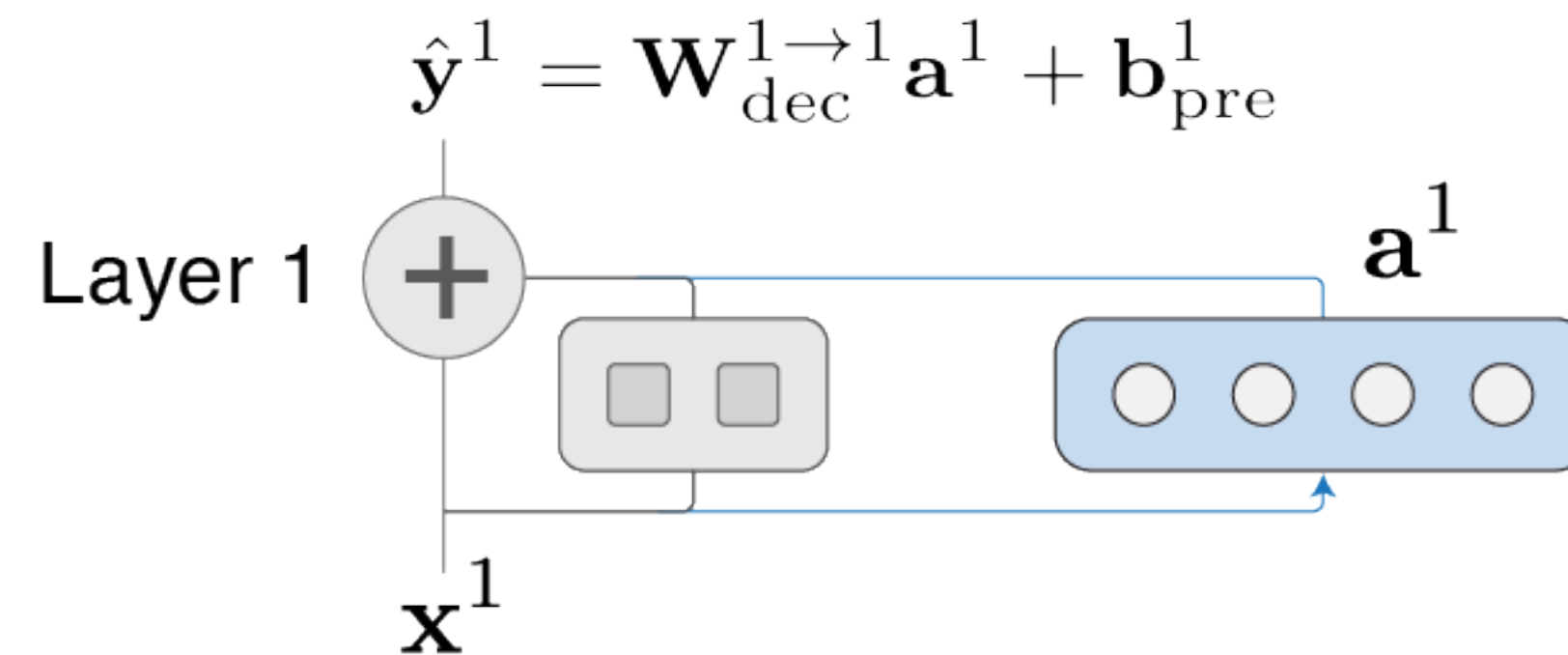
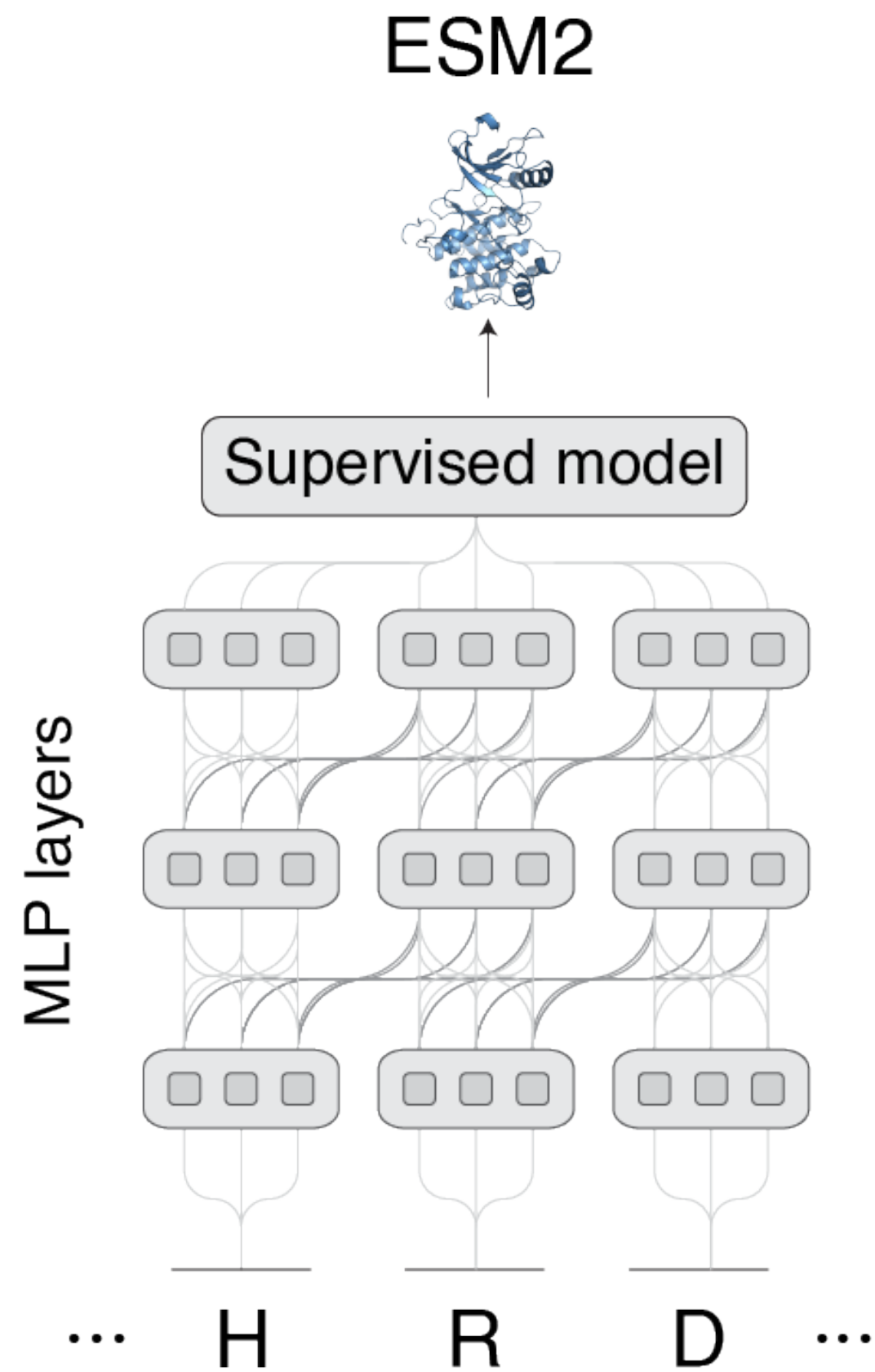
... H R D ...

Layer 1



ProtoMech models cross-layer dependencies in ESM2

$$\mathbf{a}^1 = \text{TopK}(\mathbf{W}_{\text{enc}}^1(\mathbf{x}^1 - \mathbf{b}_{\text{pre}}^1) + \mathbf{b}_{\text{enc}}^1)$$



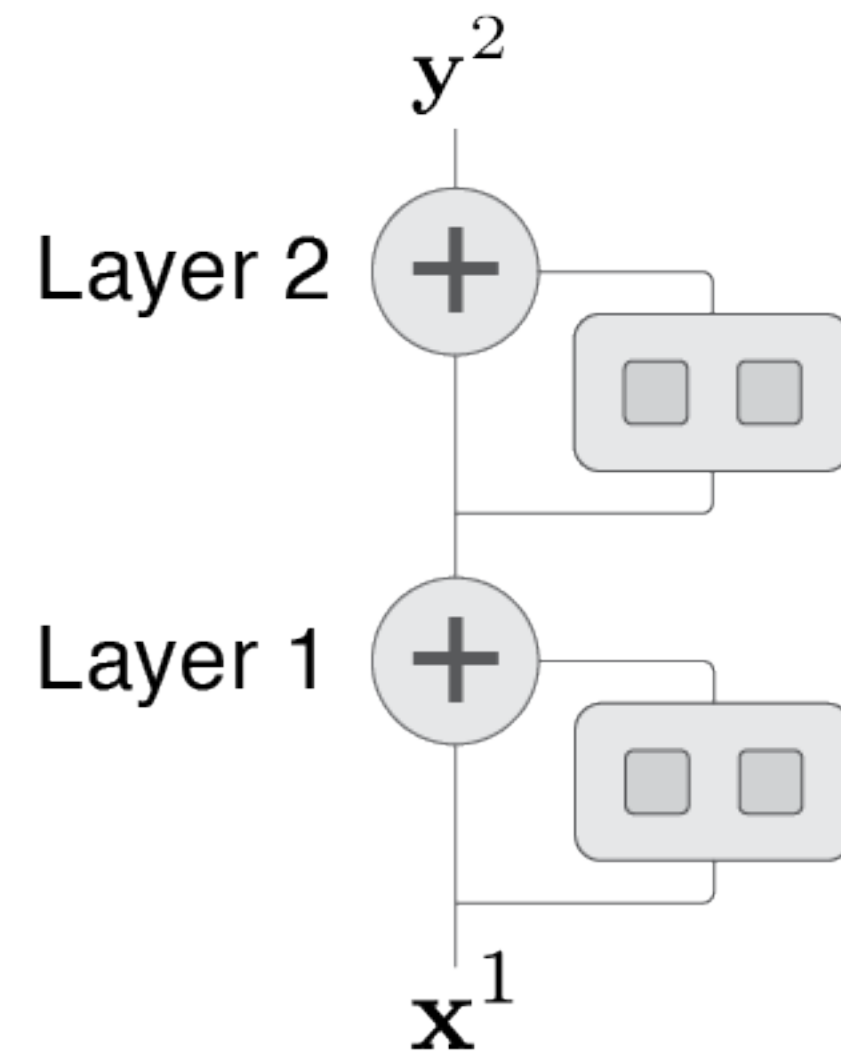
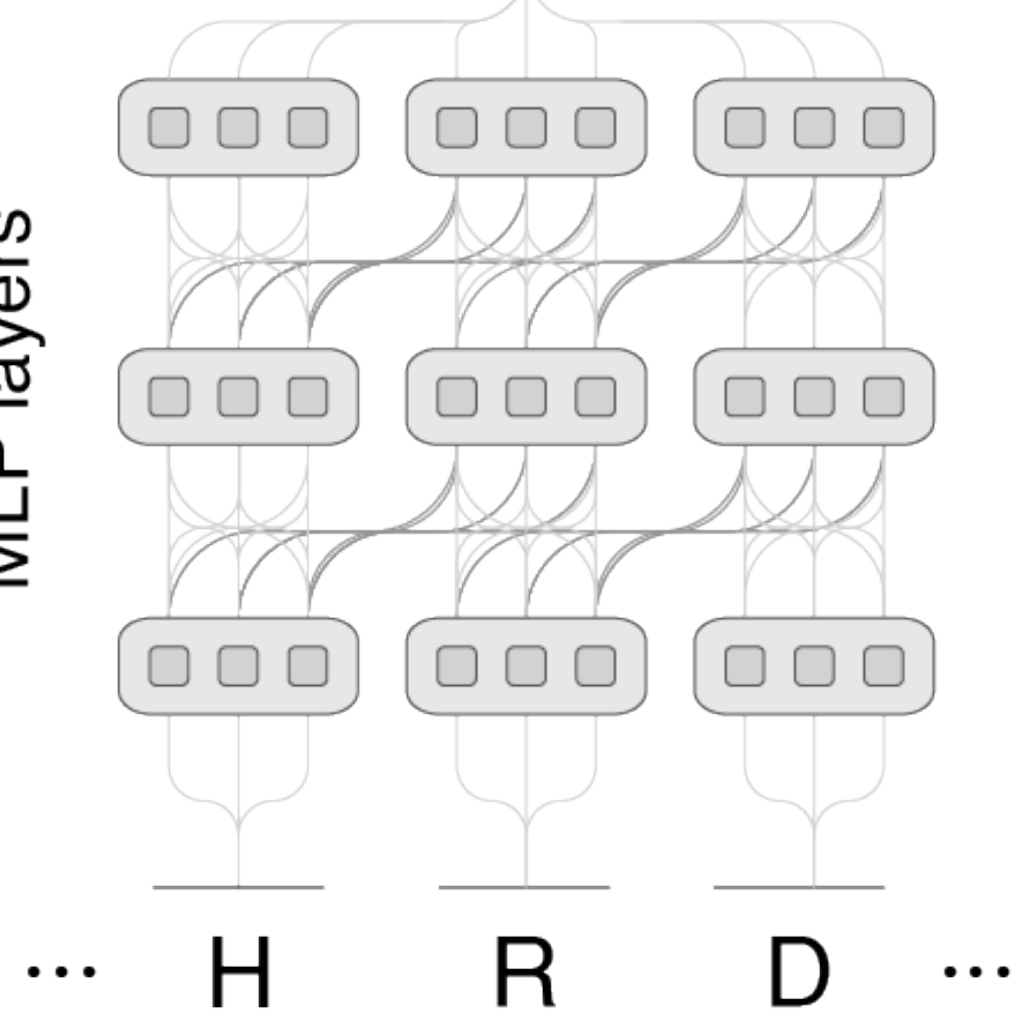
ProtoMech models cross-layer dependencies in ESM2

ESM2



Supervised model

MLP layers



ProtoMech models cross-layer dependencies in ESM2

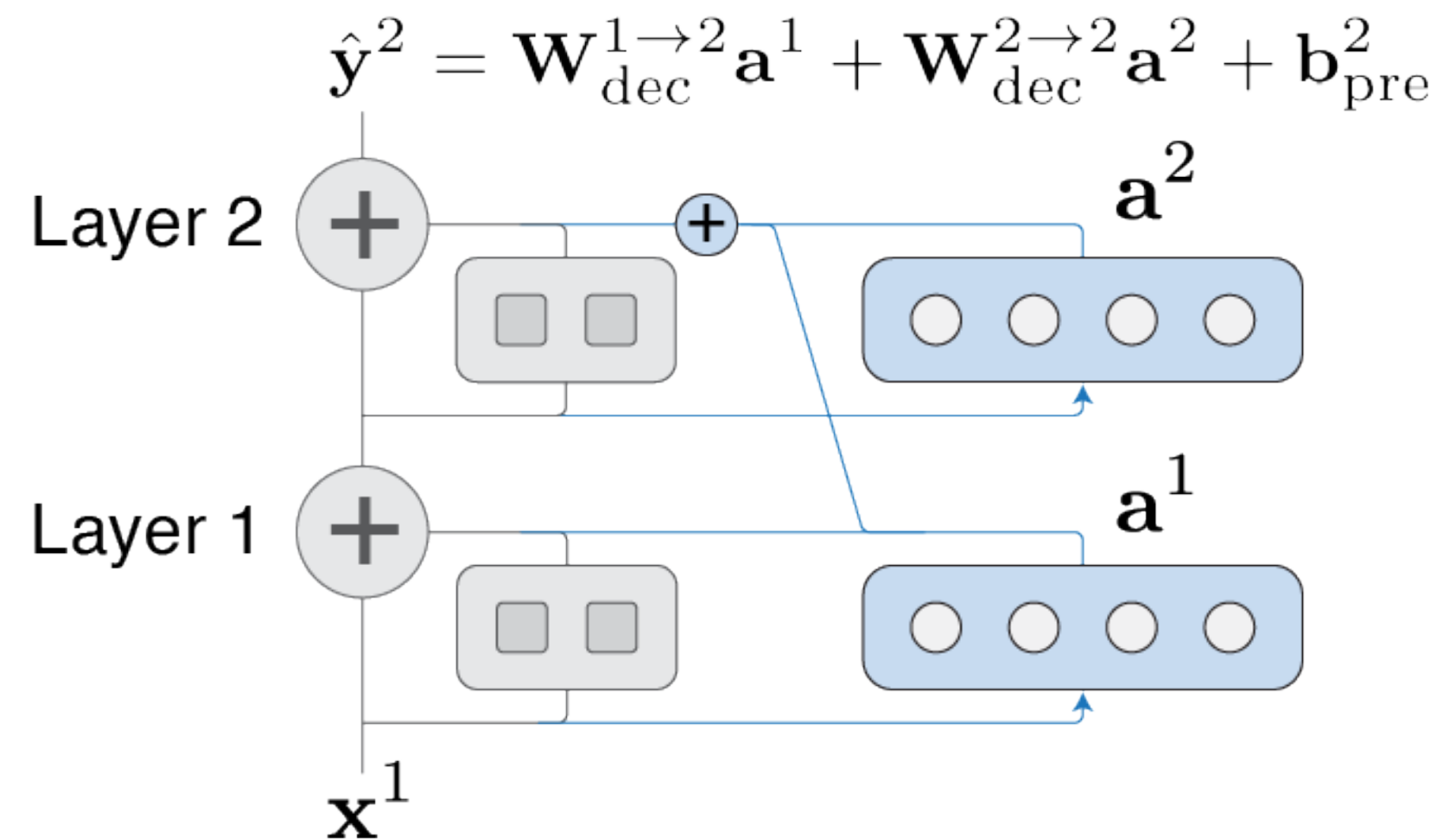
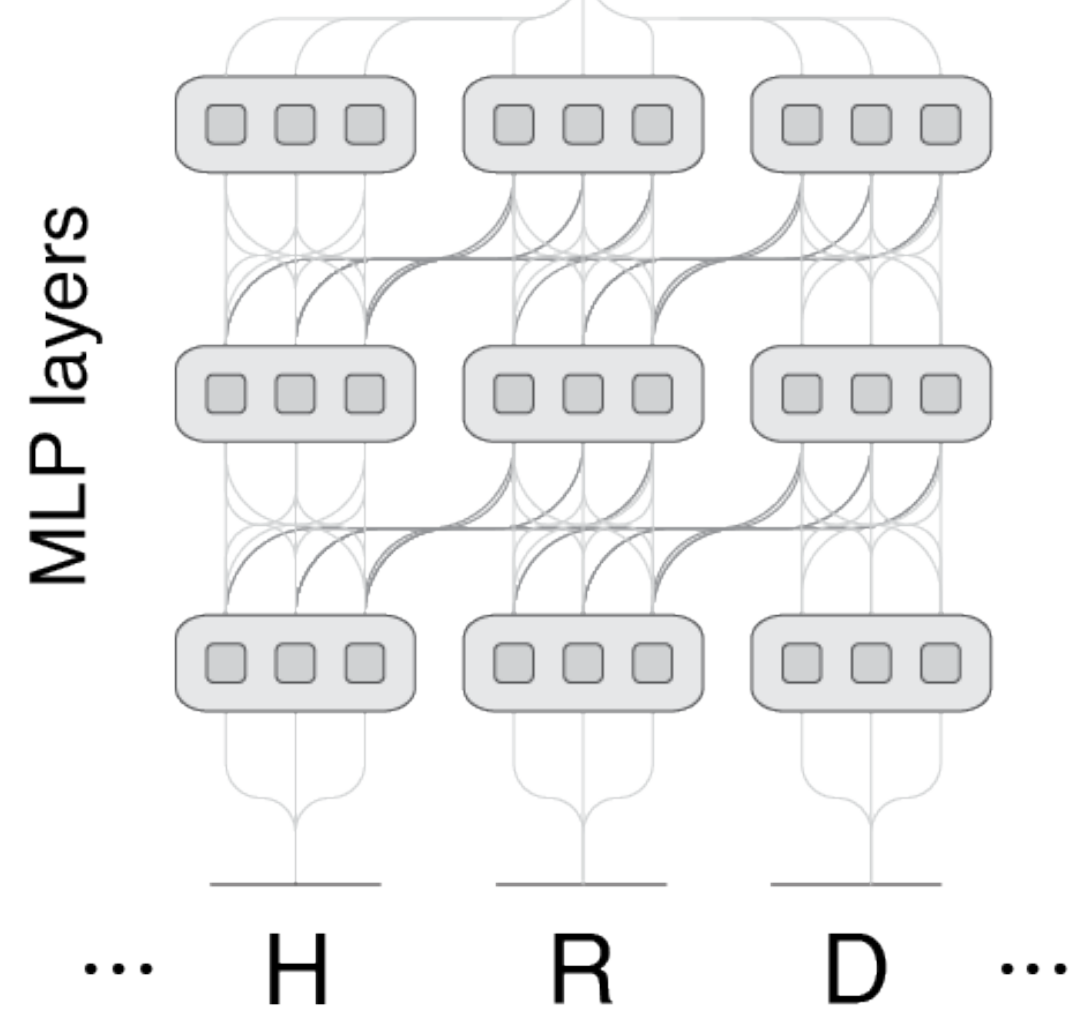
$$\mathbf{a}^1 = \text{TopK}(\mathbf{W}_{\text{enc}}^1(\mathbf{x}^1 - \mathbf{b}_{\text{pre}}^1) + \mathbf{b}_{\text{enc}}^1)$$

$$\mathbf{a}^2 = \text{TopK}(\mathbf{W}_{\text{enc}}^2(\mathbf{x}^2 - \mathbf{b}_{\text{pre}}^2) + \mathbf{b}_{\text{enc}}^2)$$

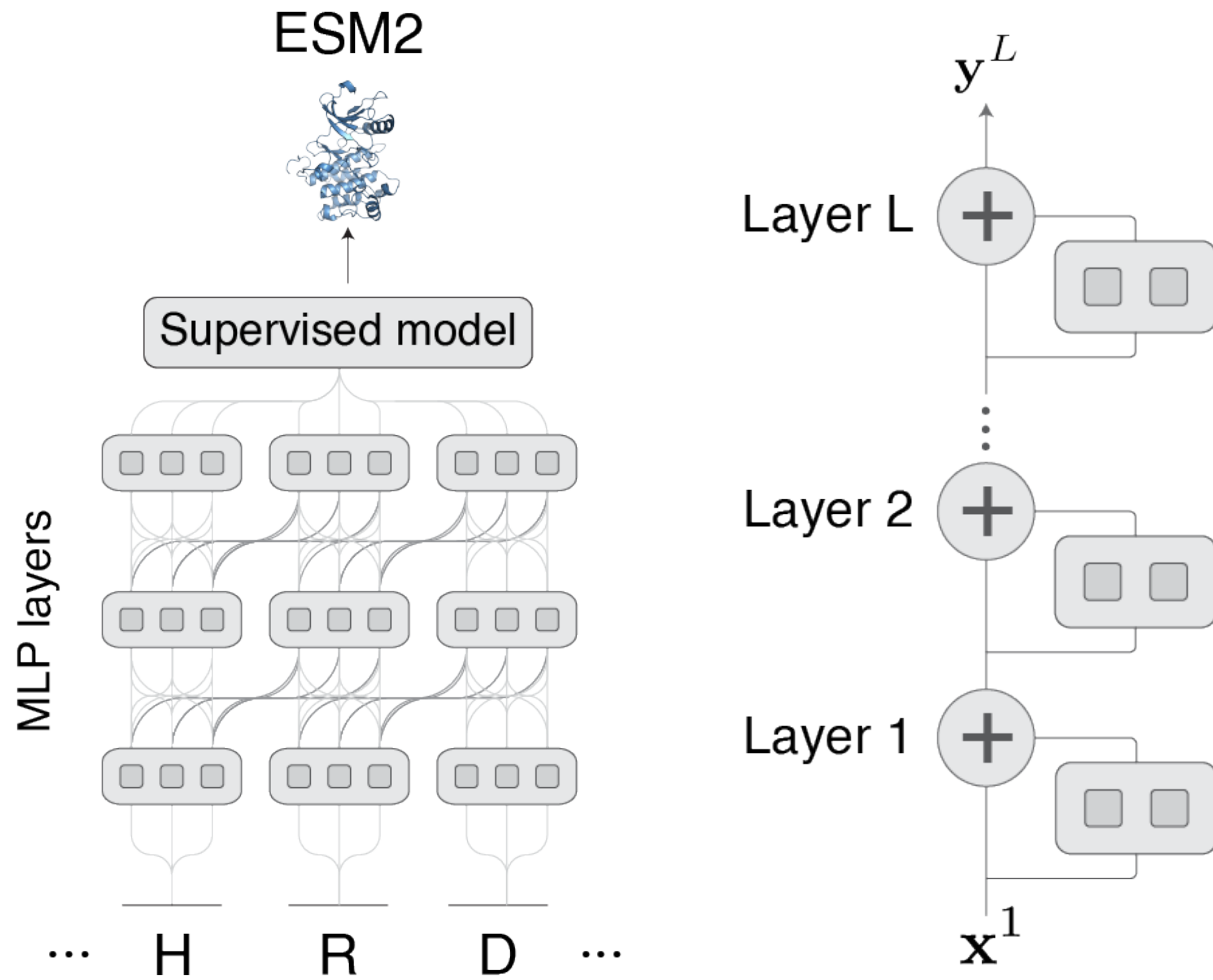
ESM2



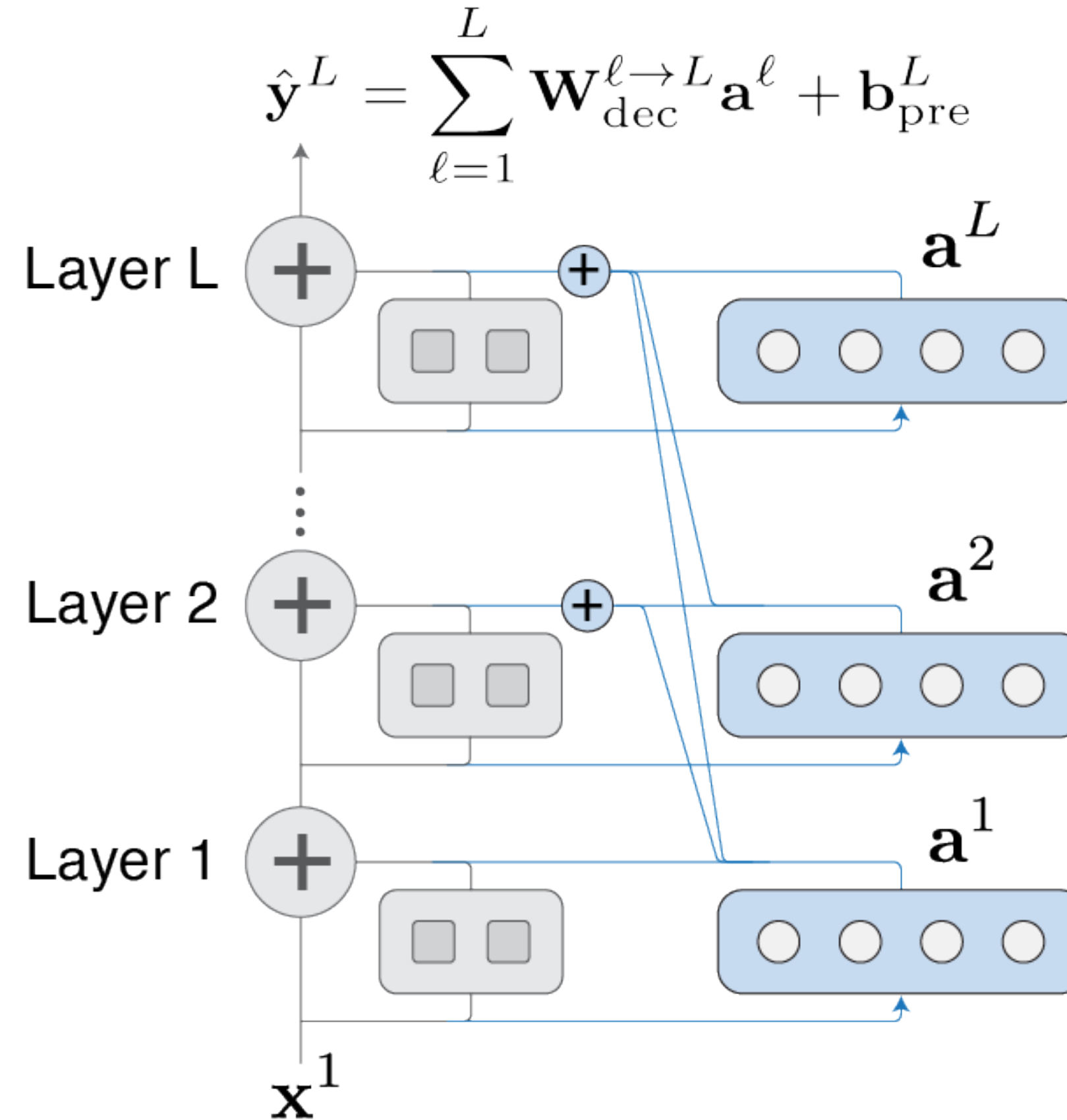
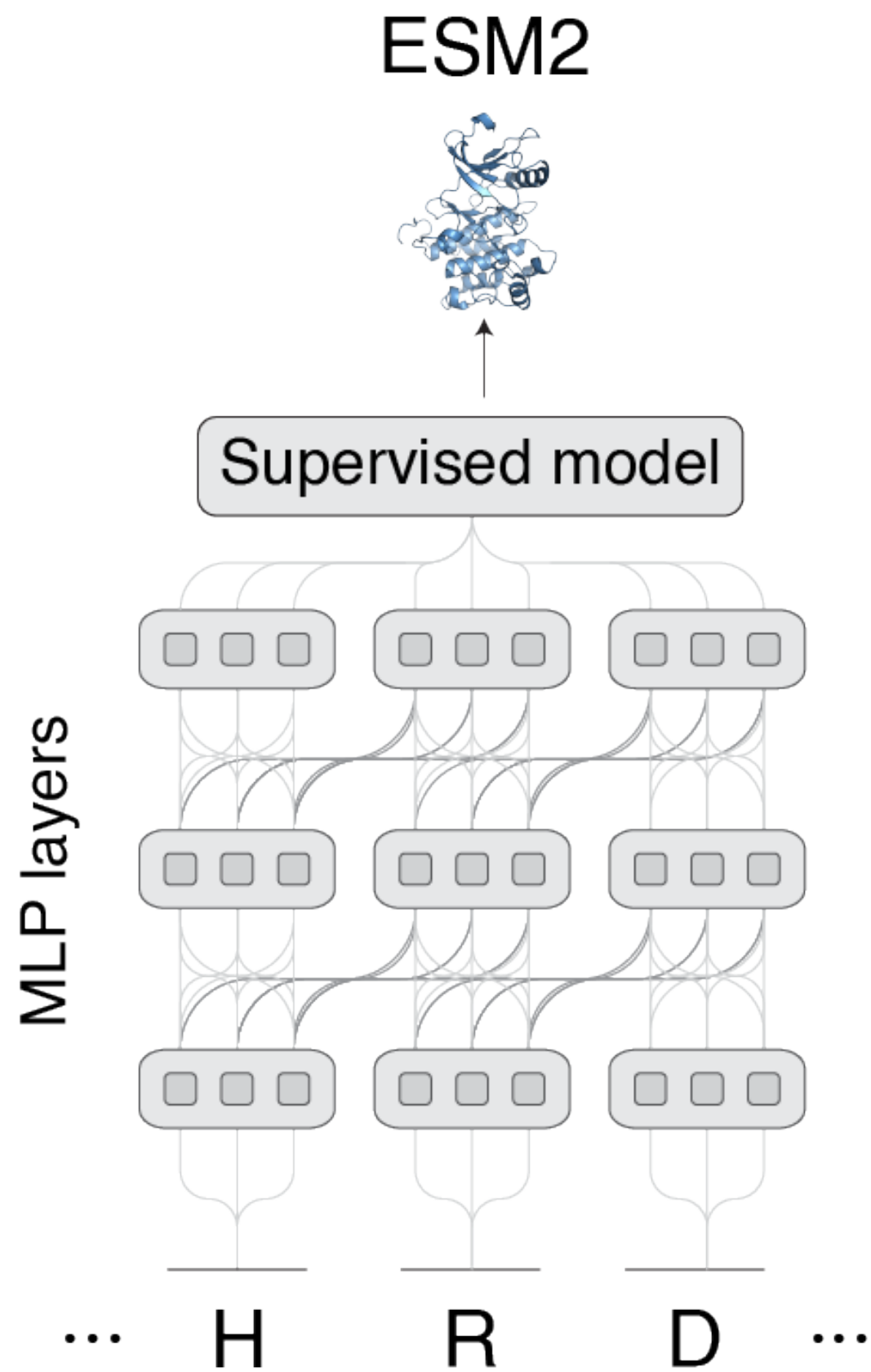
Supervised model



ProtoMech models cross-layer dependencies in ESM2



ProtoMech models cross-layer dependencies in ESM2



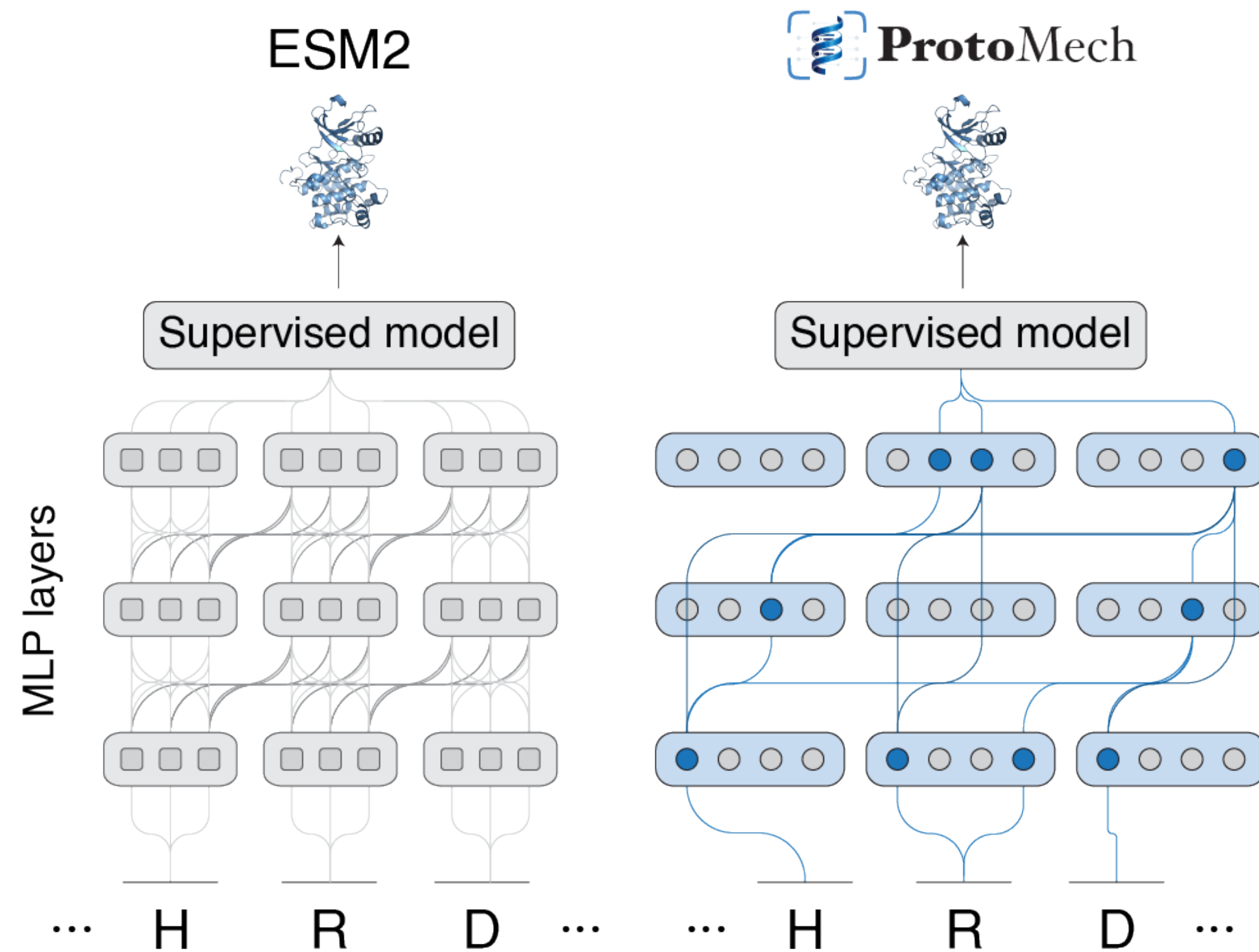
$$\mathbf{a}^1 = \text{TopK}(\mathbf{W}_{\text{enc}}^1(\mathbf{x}^1 - \mathbf{b}_{\text{pre}}^1) + \mathbf{b}_{\text{enc}}^1)$$

$$\mathbf{a}^2 = \text{TopK}(\mathbf{W}_{\text{enc}}^2(\mathbf{x}^2 - \mathbf{b}_{\text{pre}}^2) + \mathbf{b}_{\text{enc}}^2)$$

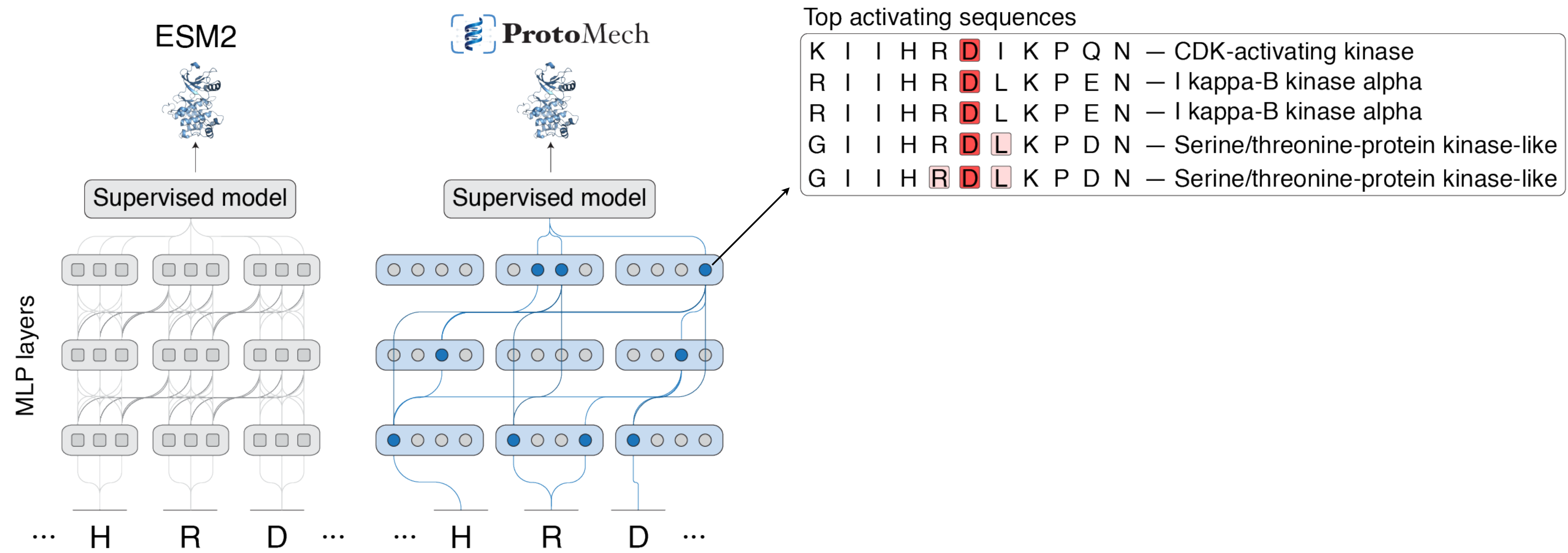
$$\mathbf{a}^L = \text{TopK}(\mathbf{W}_{\text{enc}}^L(\mathbf{x}^L - \mathbf{b}_{\text{pre}}^L) + \mathbf{b}_{\text{enc}}^L)$$



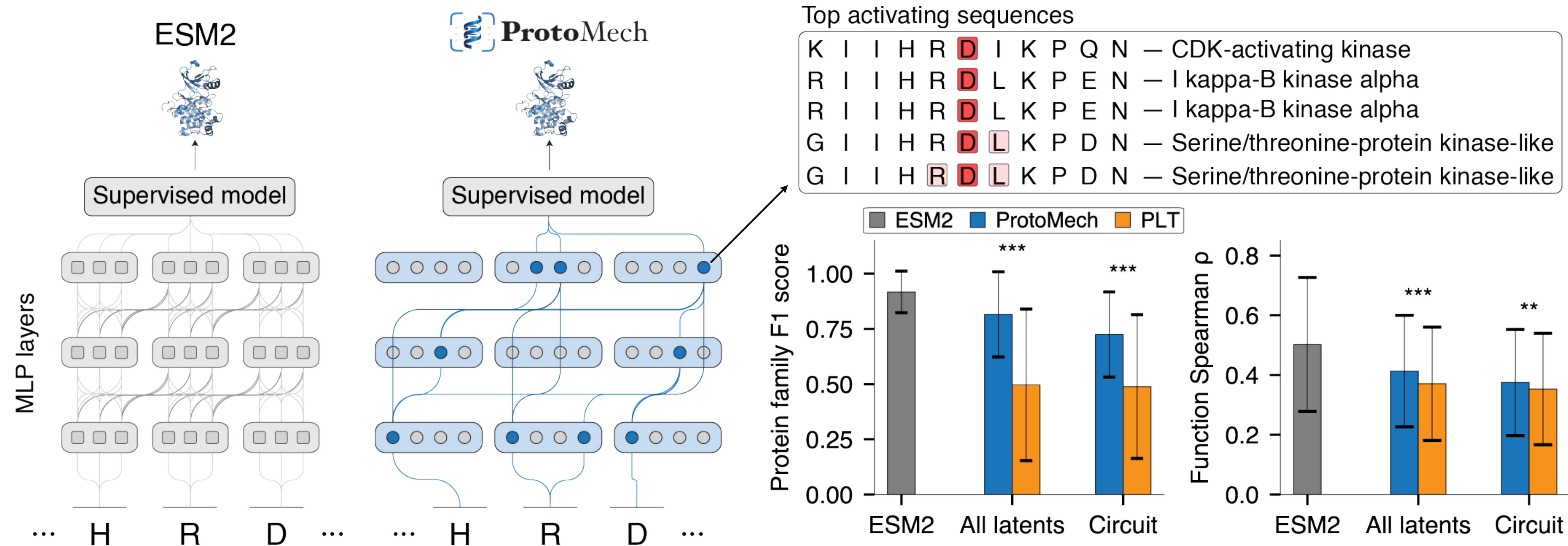
ProtoMech functions as a replacement model for ESM2



ProtoMech functions as a replacement model for ESM2

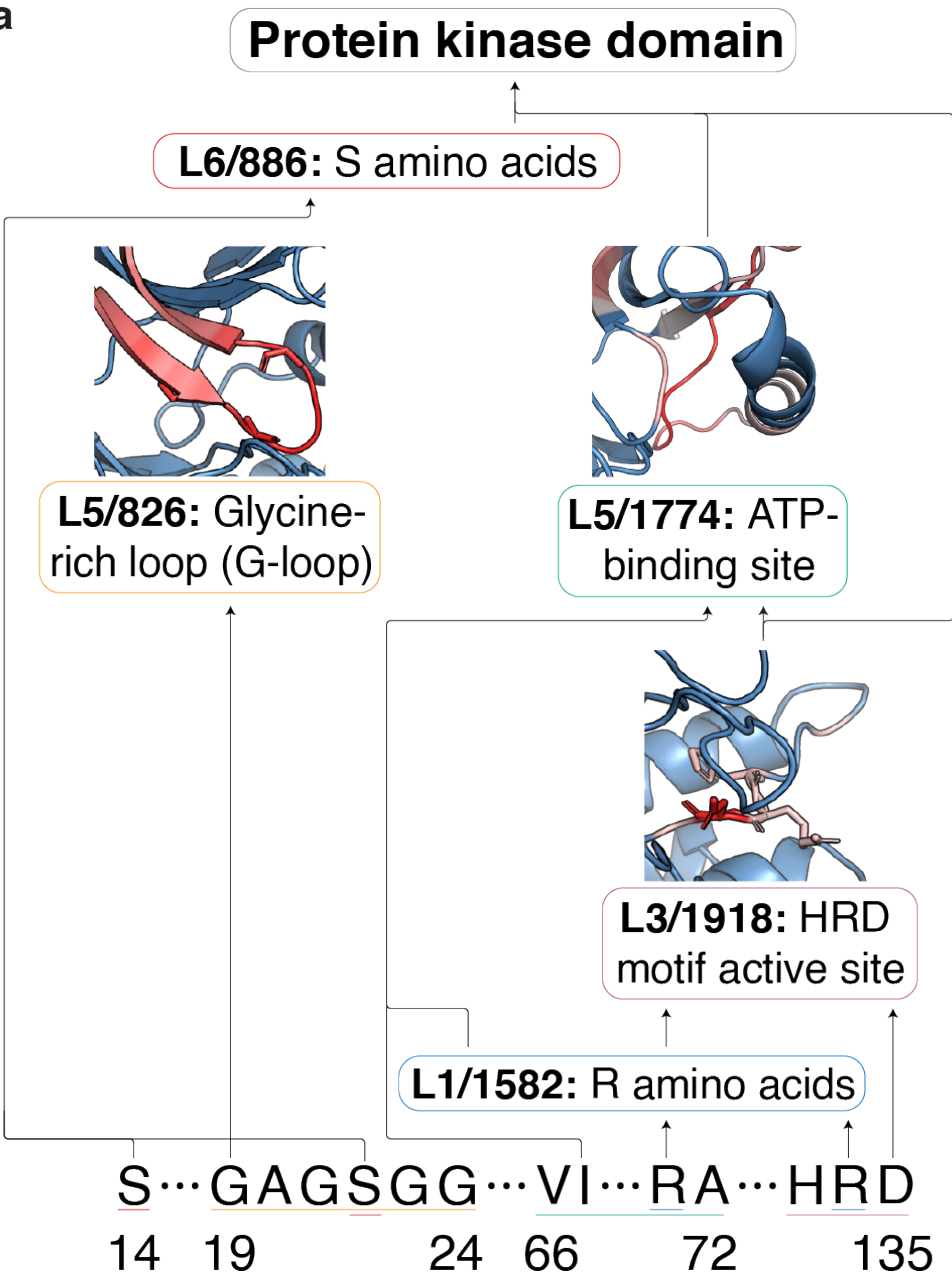


ProtoMech functions as a replacement model for ESM2

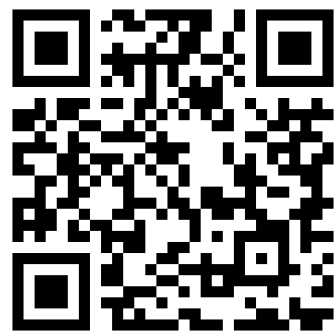
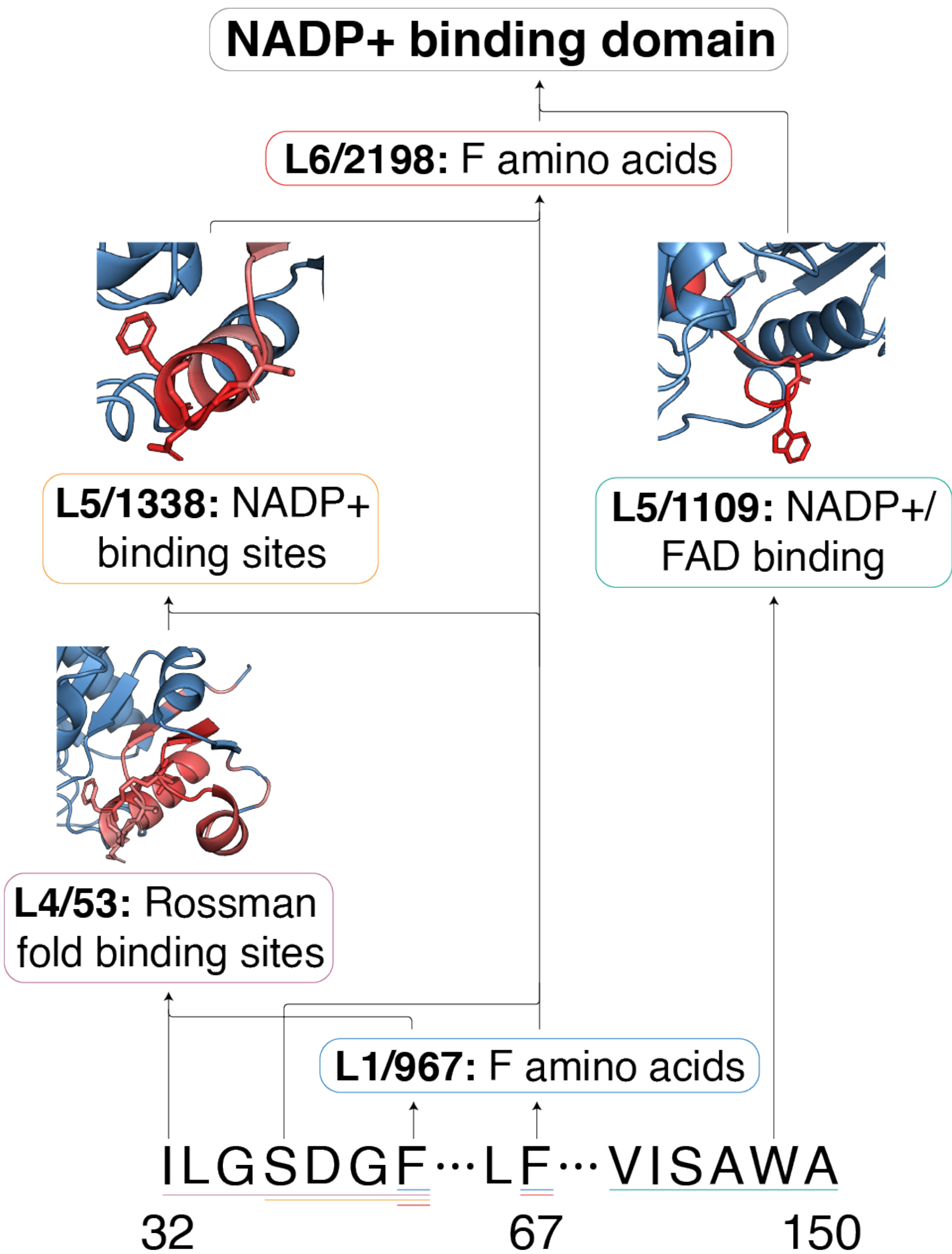


ProtoMech reveals known biological motifs

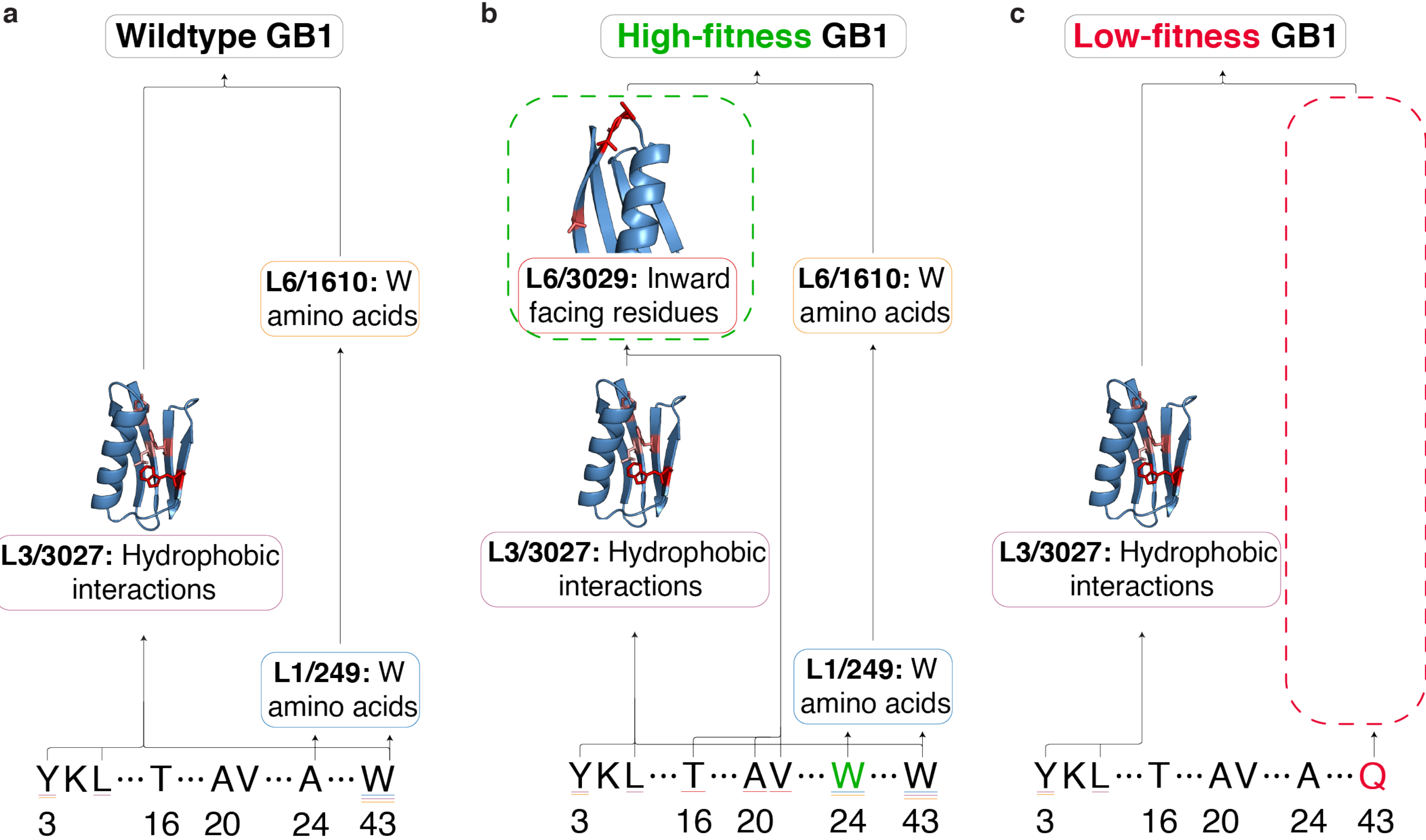
a



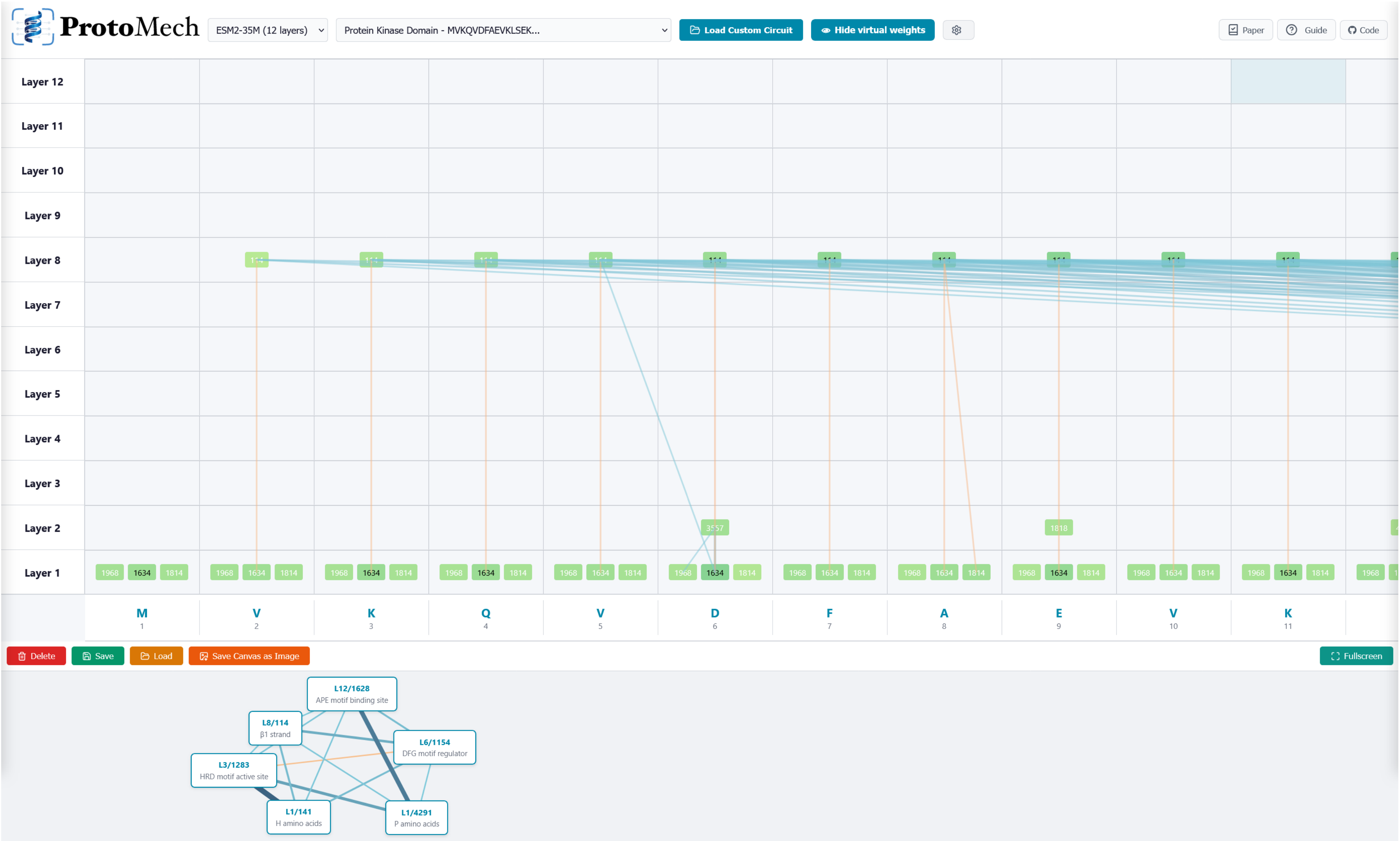
b



ProtoMech reveals known biological motifs

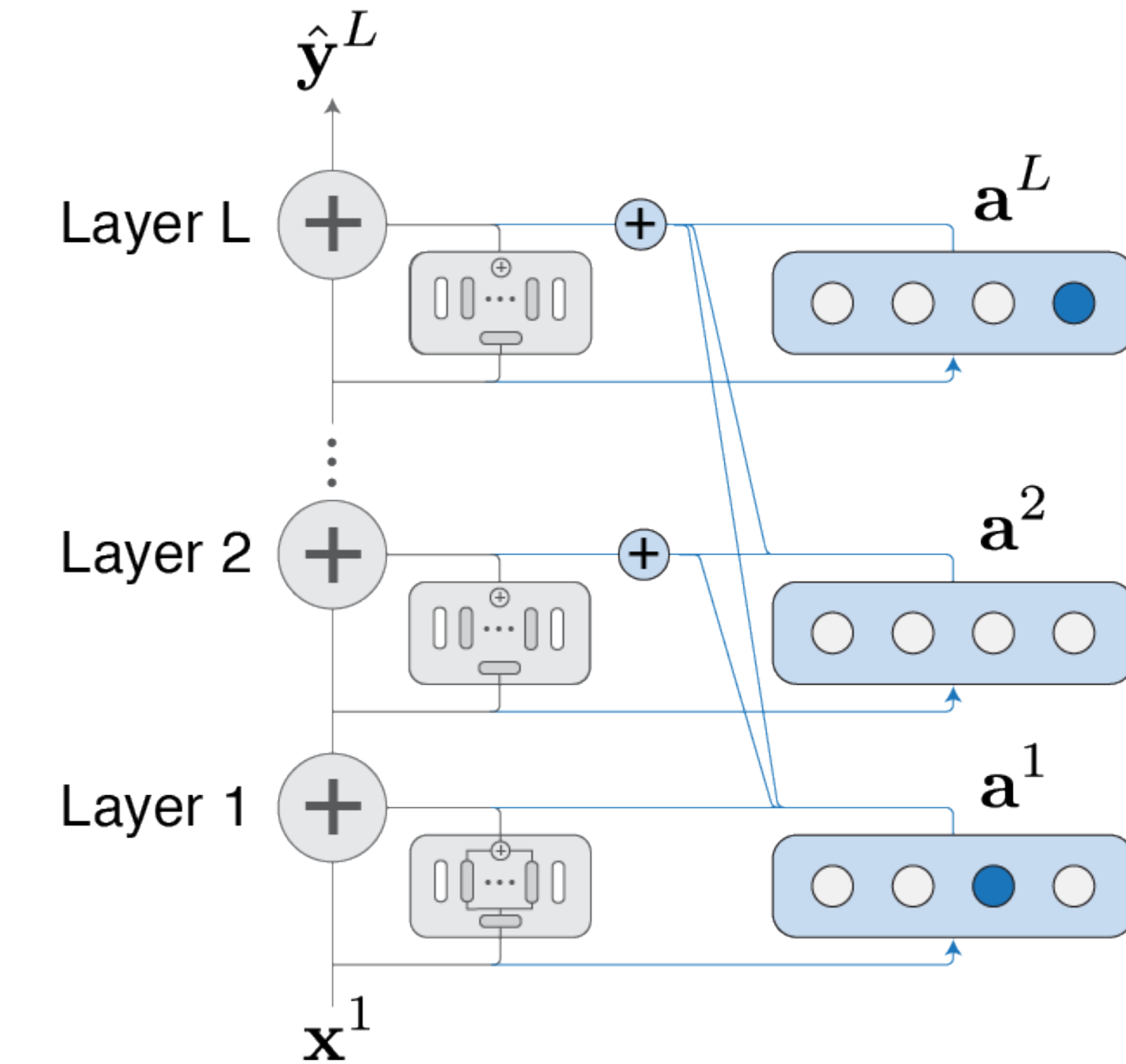


<https://protmech.github.io/>: democratizing circuit tracing



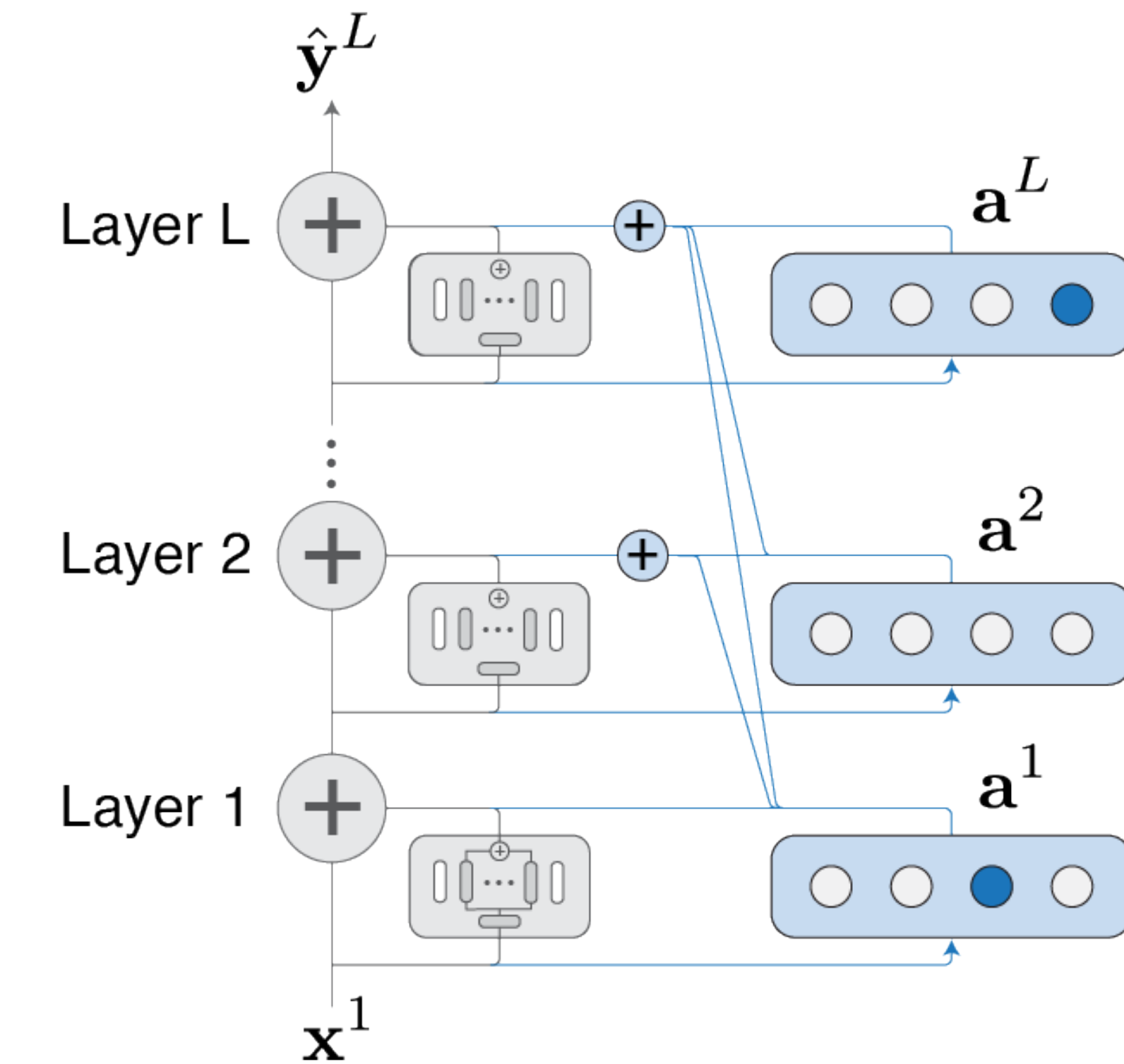
adapting ProtoMech for ProGen3

a Mixture of experts architecture

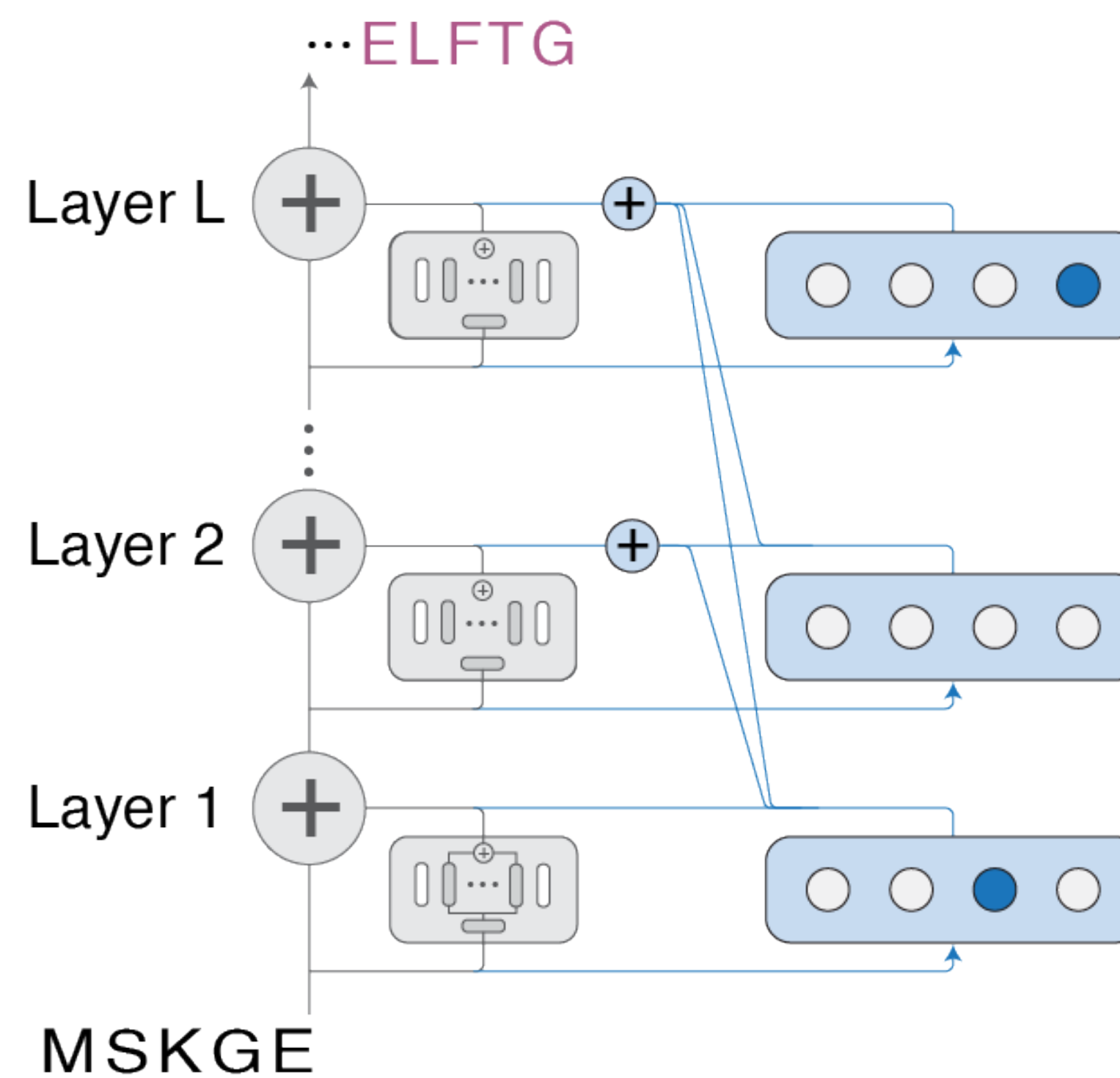


adapting ProtoMech for ProGen3

a Mixture of experts architecture

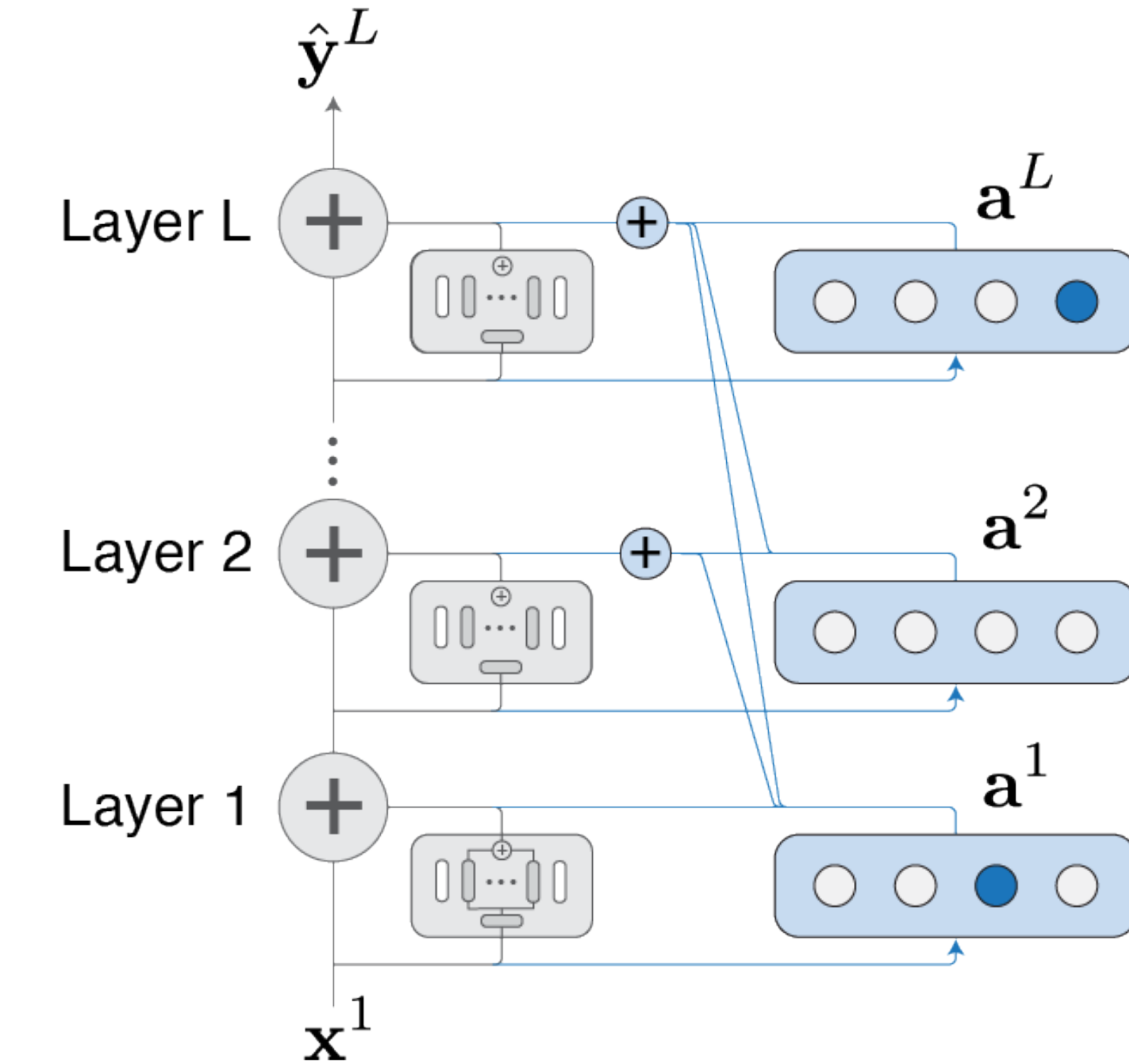


b Causal language modeling (CLM)

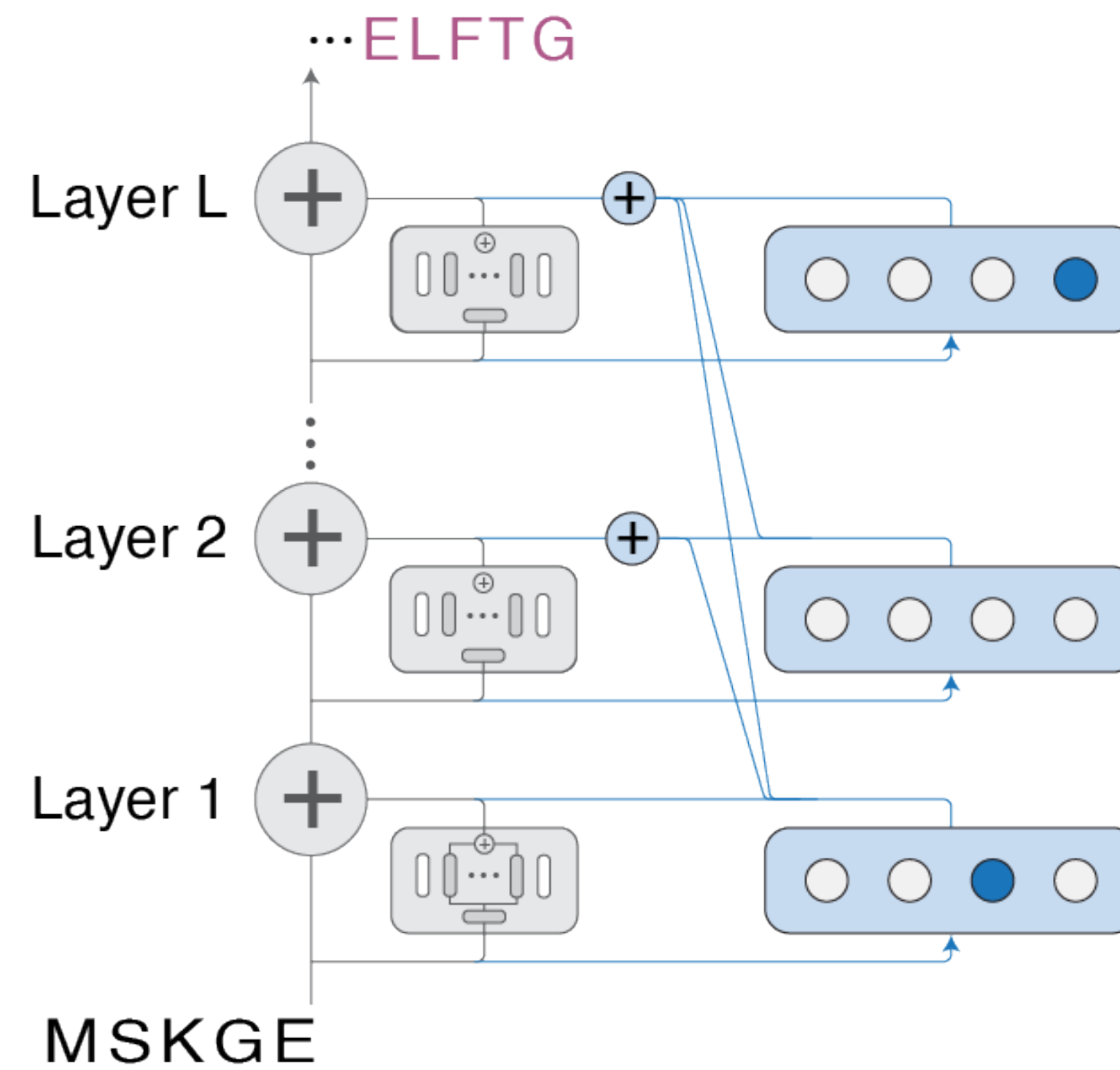


adapting ProtoMech for ProGen3

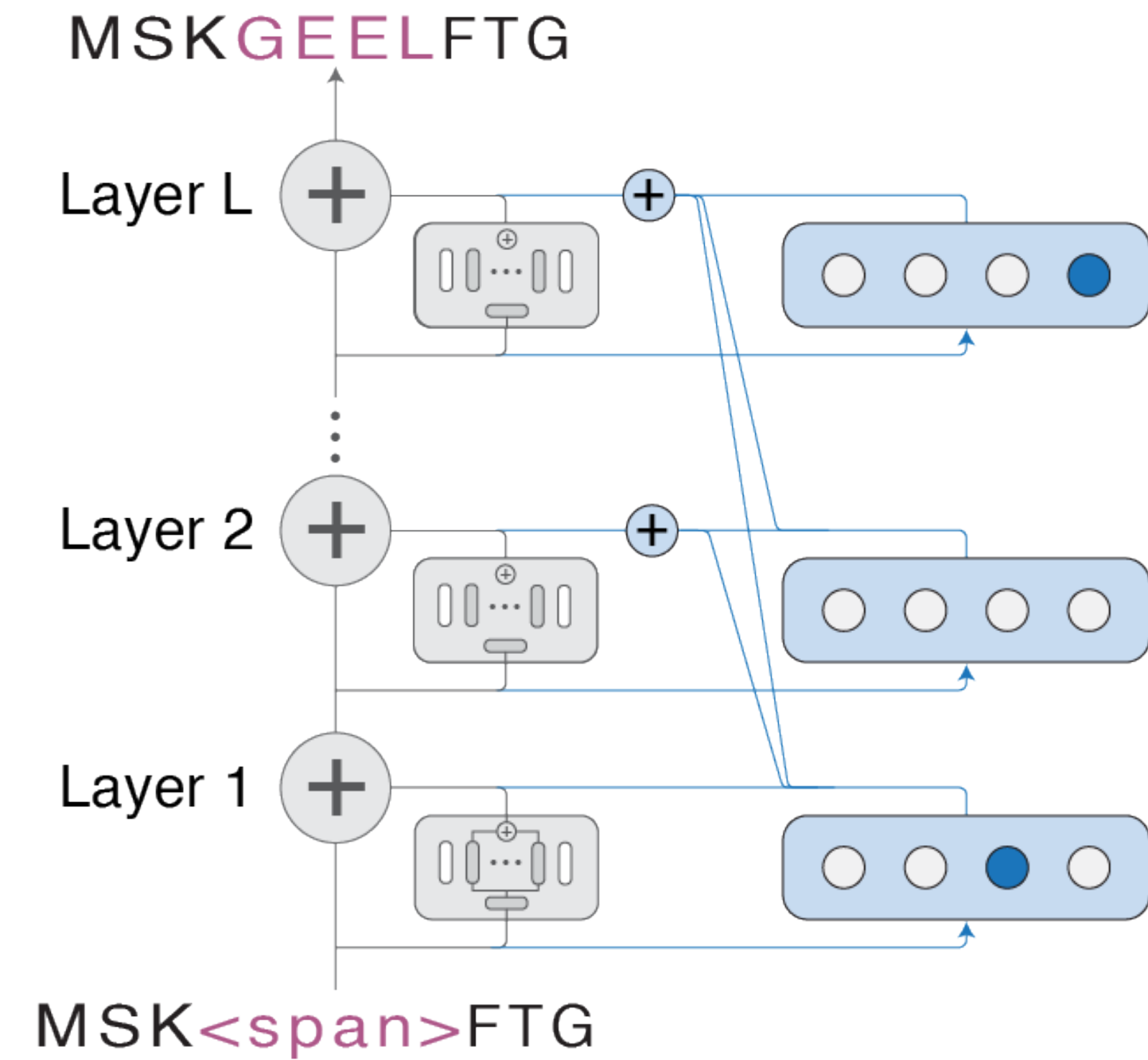
a Mixture of experts architecture



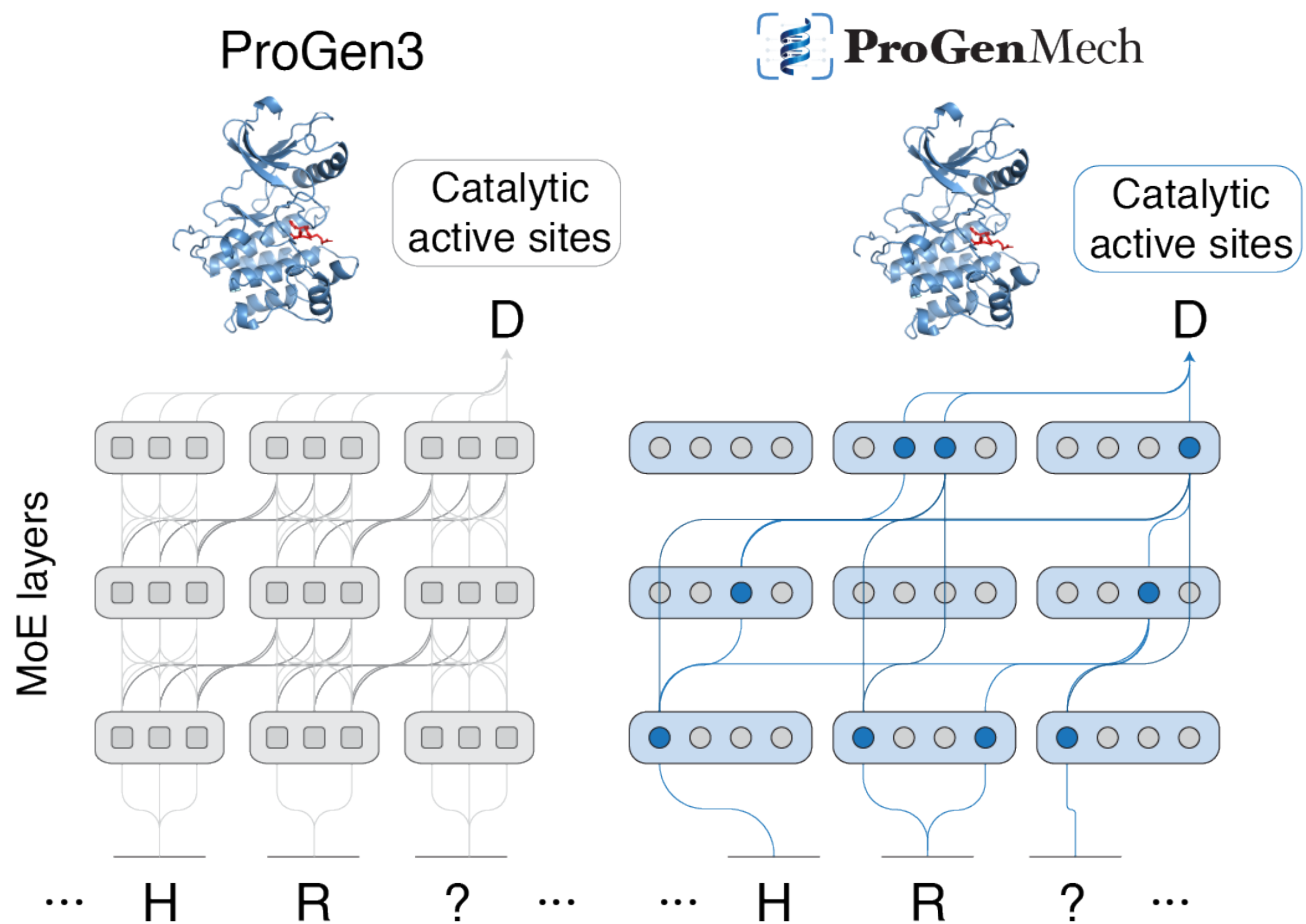
b Causal language modeling (CLM)



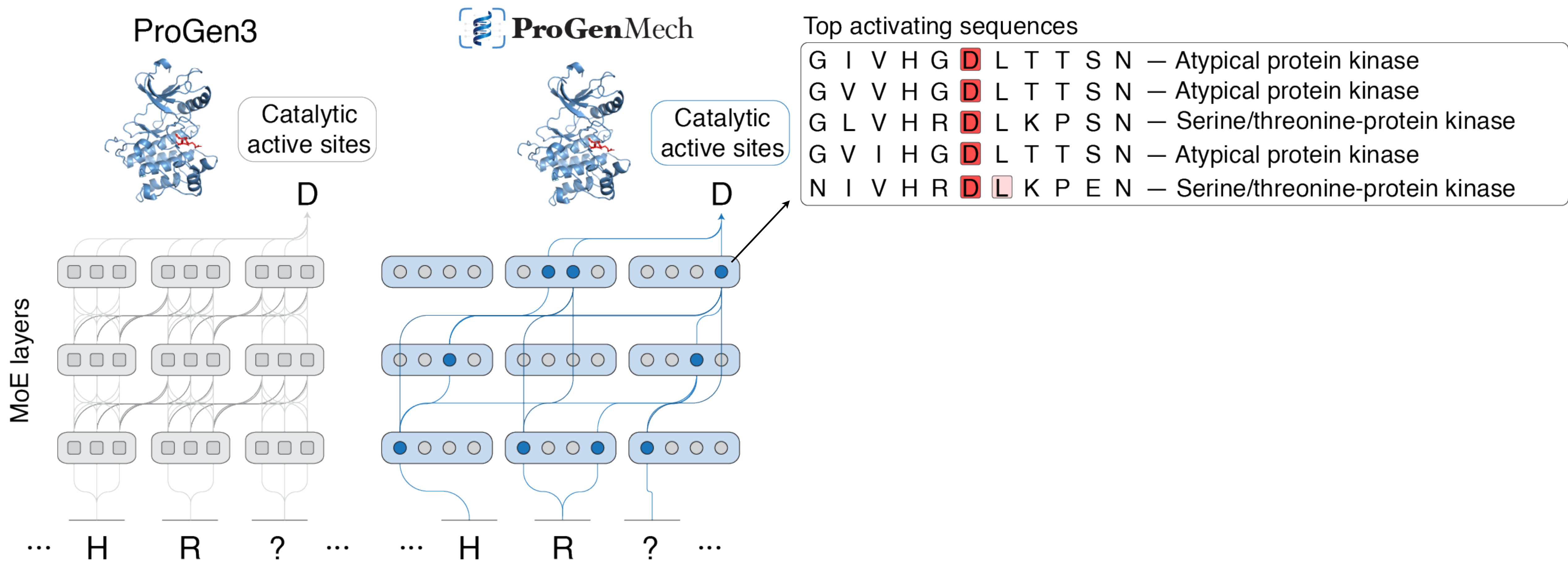
c Generalized language modeling (GLM)



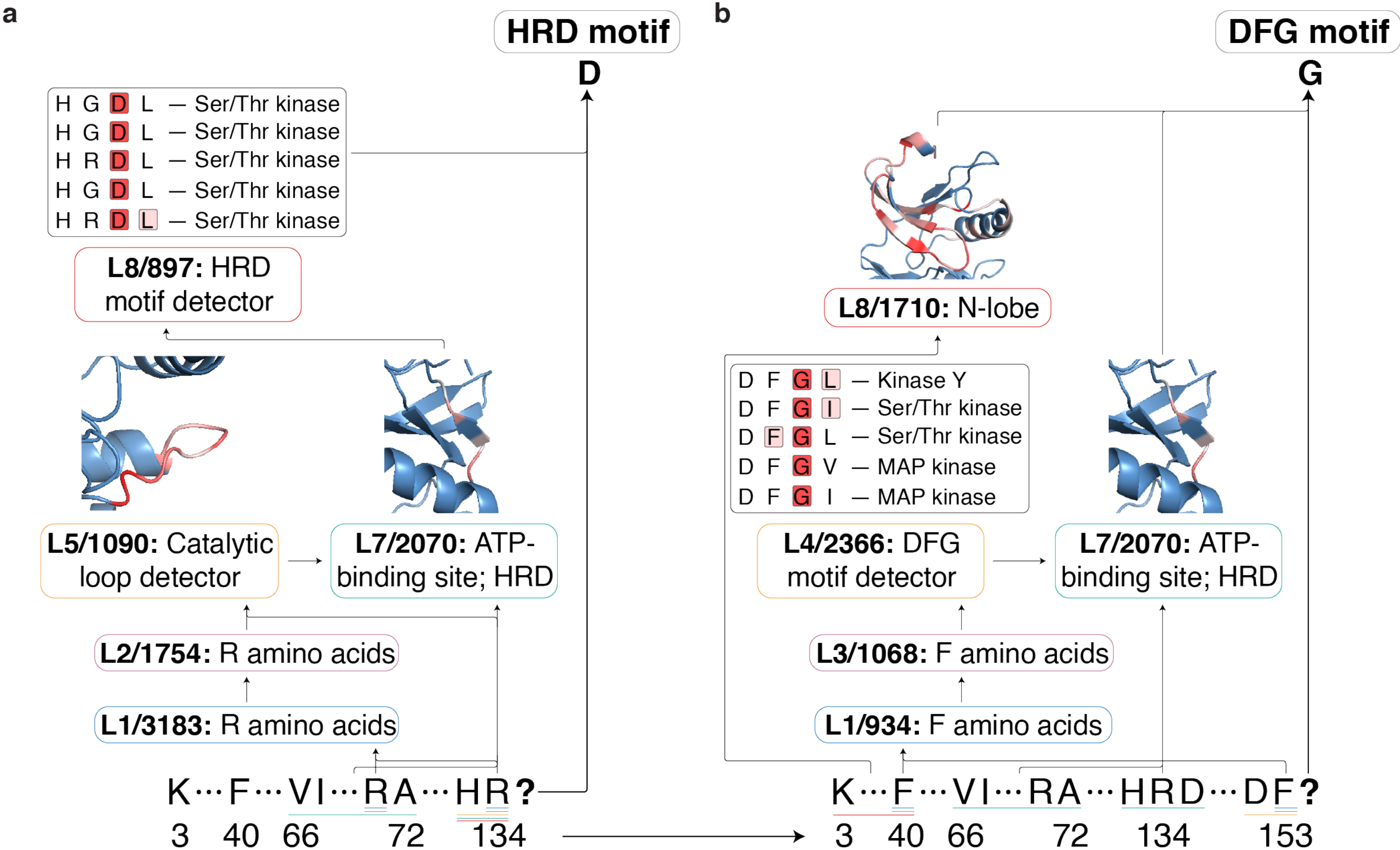
ProGenMech functions as a replacement model for ProGen3



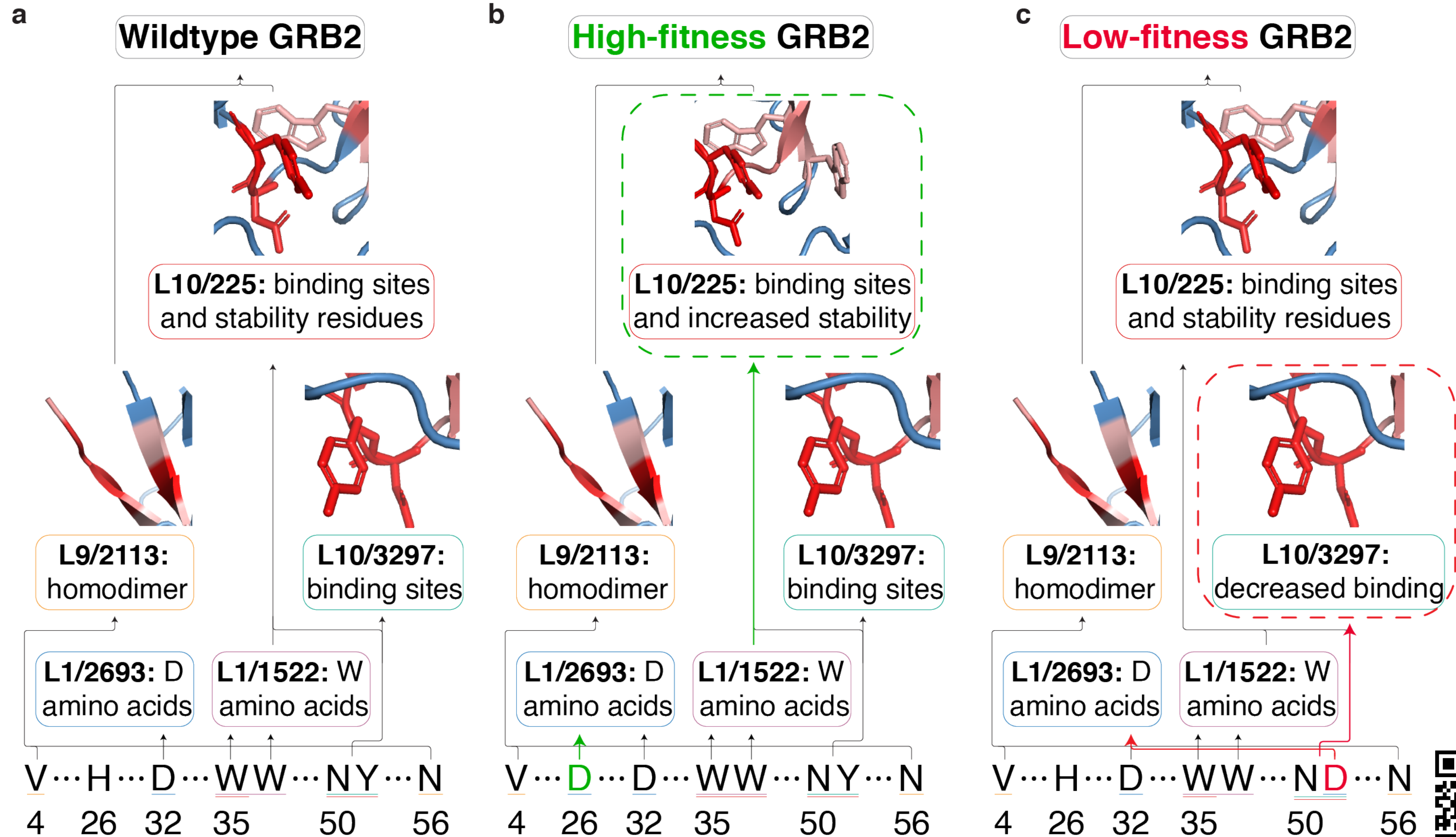
ProGenMech functions as a replacement model for ProGen3



ProGenMech reveals known biological motifs



ProGenMech reveals known biological motifs



discussion and limitations

- ProGenMech is only applied to ProGen3-112M
 - what new insights will we gain by scaling CLTs?
- we're missing a lot of performance from CLM/GLM generation
 - abstracting away the MoE layer may leave performance on the table
 - GLM performance in the original ProGen3 model is particularly weak
- ProGenMech can only explain what the model understands
 - incorrect/uncertain predictions won't reveal anything meaningful

future outlook

- will scaling ProGenMech lead to uncovering novel biological motifs?
- can we automate the interpretation process of latents with agentic AI?
- what potential use cases exists for mechanistic interpretability in biology?
- can we use ProGenMech's latents to propose hypotheses and design new experiments?

Thank you for listening!



AMIR Group



Daniel Saeedi
PhD Student



Amirali Aghazadeh
Assistant Professor of ECE