BioSpectrum
the business of Bio & Health Sciences
ASIA EDITION

# Gero’s ProtoBind?Diff redefines drug discovery by designing molecules from sequence alone

02 July 2025 | Opinion | By Ankit Kankar | ankit.kankar@mmactiv.com

**Peter Fedichev discusses how the AI-driven platform bypasses structural bottlenecks to target elusive proteins in cancer, ageing, and pandemics**


**Peter Fedichev, Ph.D, Gero’s CEO**

***BioSpectrum Asia* spoke to Dr Peter Fedichev, CEO of Gero,** who explains how ProtoBind-Diff is transforming drug discovery by eliminating the need for protein structural data. Built on a vast training set of over a million protein–ligand interactions, the platform leverages masked diffusion and language modelling to generate novel compounds directly from amino acid sequences. Dr Fedichev shares how this approach expands the druggable universe, accelerates discovery against challenging disease targets, and positions Gero to respond rapidly to emerging health threats.

## How does ProtoBind-Diff tackle the challenge of targeting proteins without 3D structural data?

ProtoBind-Diff was designed from the ground up to overcome a fundamental bottleneck in drug discovery: the limited availability of high-quality 3D structural data for protein–ligand complexes. While structure-based approaches like docking or AlphaFold-guided generative models rely on resolved protein structures or predicted pockets, these are not available or reliable for a significant fraction of biologically relevant targets—especially novel, disordered, or poorly characterized proteins. This scarcity of structural data restricts the druggable target space, especially in challenging therapeutic areas such as cancer, aging, and neurodegeneration.

ProtoBind-Diff sidesteps this dependency entirely by using only the linear amino acid sequence of a protein as its input. The model is a masked diffusion language model that learns the joint distribution between protein sequences and chemically valid small molecules that bind to them. This allows ProtoBind-Diff to operate effectively in sequence space, bypassing the need for any structural input.

To achieve this, we trained the model on more than a million known protein–ligand interactions, representing a vastly larger training dataset than any structure-based system can typically access. These interactions were derived from publicly available activity data (e.g., IC50, Ki, EC50) across multiple assay types and conditions. This abundance of sequence-conditioned activity data allows the model to generalize effectively, learning latent representations of protein binding sites and their chemical preferences directly from sequence-level patterns. In other words, ProtoBind-Diff implicitly

learns "what a binding site looks like" from amino acid sequence motifs, without ever needing to "see" the binding site in 3D.

This approach opens the door to rational drug design for previously undruggable or structurally intractable targets, using the most universally available biological feature: the protein sequence.

**What makes ProtoBind-Diff perform better than models like Pocket2Mol on harder targets?**

Pocket2Mol and similar models have made impressive strides in structure-based generation. However, these models are fundamentally limited by the narrow set of protein–ligand co-crystal structures available in public databases like the PDB. These structures tend to represent well-behaved targets—mostly kinases, GPCRs, and other classical drug targets with highly conserved binding pockets.

ProtoBind-Diff differs in both training data scale and model architecture. First, we trained on a dataset more than an order of magnitude larger than what Pocket2Mol or traditional structure-based generative models typically use. Our million-plus protein–ligand pair dataset includes diverse protein families, assay types, and chemical scaffolds. This breadth gives ProtoBind-Diff much more generalizable predictive power, especially for novel or low-data targets.

Second, the model architecture itself—based on masked diffusion and language modeling—enables more flexible generation. Rather than relying on precise geometric constraints from a binding pocket, ProtoBind-Diff generates molecules conditioned on learned sequence motifs, protein family context, and prior examples of active compounds. This sequence-centric strategy proves especially powerful for "hard" targets—those lacking resolved structures, exhibiting high flexibility or disorder, or belonging to poorly annotated protein families.

In internal benchmarks, ProtoBind-Diff outperformed Pocket2Mol on multiple fronts: (1) success rate in generating active-like compounds for challenging targets, (2) chemical diversity of the outputs, and (3) predicted binding strength using orthogonal bioactivity predictors. Importantly, ProtoBind-Diff also exhibited stronger scaffold novelty, suggesting a greater capacity to explore untapped chemical space beyond template-based approaches.

**How could this sequence-only approach impact drug discovery for diseases like cancer and aging?**

The sequence-only approach of ProtoBind-Diff is especially impactful in therapeutic areas like cancer and aging, where many relevant targets fall outside the conventional druggable genome.

In cancer, for example, oncogenic drivers such as transcription factors, intrinsically disordered proteins, and non-canonical protein–protein interactions have historically been difficult to target due to their lack of well-defined binding pockets. Structural disorder, low expression levels, and poor solubility have made many of these proteins resistant to crystallography or AlphaFold modeling. Yet these are precisely the kinds of targets that could dramatically shift the treatment paradigm if made accessible to drug design.

ProtoBind-Diff can address these targets because it relies only on the primary sequence—something available for nearly every human protein. This allows us to systematically generate small molecules against long-overlooked or "undruggable" targets, such as MYC, FOXO, or IDPs implicated in cellular senescence and age-related inflammation.

In the context of aging, the opportunity is even greater. Gero has developed a physics-informed large model of human health based on 50 million longitudinal patient records. This model allows us to identify the biological root causes of aging and the earliest upstream regulators of disease progression. These upstream targets, often unrelated to canonical drug targets, tend to be low-expression, non-enzyme proteins with little structural information available. With ProtoBind-Diff, we can now rapidly design compounds that engage these regulators based on sequence-level insight alone.

In short, ProtoBind-Diff expands the druggable universe—especially for age-related and oncology targets that have remained beyond reach for structure-reliant platforms.

**Why was training on a million protein–ligand pairs crucial to the model's success?**

Training on such a large and diverse dataset was essential to ensure ProtoBind-Diff's broad generalizability and real-world applicability. Unlike structure-based generative models, which are typically trained on a few hundred thousand resolved protein–ligand structures, ProtoBind-Diff leverages a vastly larger set of experimental activity data, much of which is tied to sequence but lacks corresponding structural information.

This large training corpus allows the model to learn nuanced relationships between sequence motifs and chemical features—essentially capturing the statistical co-occurrence of specific residues or domains with ligand scaffolds,

functional groups, and pharmacophores. These associations would be impossible to infer reliably from smaller or more structurally constrained datasets.

Additionally, this scale supports ProtoBind-Diff's ability to perform well in low-data regimes. For example, even when presented with a novel protein from an underrepresented family, the model has likely seen similar sequence motifs or related interaction patterns during training. This gives it a "prior" over likely binding chemotypes, enabling it to generate meaningful candidates even when no structural or ligand data exists for the target of interest.

Finally, the breadth of chemical space covered by our dataset ensures that the model does not overfit to a narrow set of well-explored compounds. Instead, ProtoBind-Diff learns a rich chemical language that allows for scaffold diversity, novel linker formation, and the generation of truly first-in-class molecules.

### What role could ProtoBind-Diff play in fast-tracking treatments during future pandemics

One of the critical lessons of the COVID-19 pandemic was the need for rapid drug discovery platforms that can respond to emerging pathogens without waiting for structural biology or wet lab screening to catch up. ProtoBind-Diff is uniquely suited to address this challenge.

Because it requires only the amino acid sequence of a protein to begin molecular generation, ProtoBind-Diff can be deployed immediately after sequencing a novel viral genome. There's no need to wait for expression, purification, crystallography, or cryo-EM data. This ability to "go from genome to drug candidate" in a matter of days could shave critical months off the therapeutic development timeline in a pandemic setting.

Moreover, ProtoBind-Diff's flexible architecture allows it to generate diverse compounds against multiple viral targets in parallel—such as proteases, polymerases, or host interaction factors—enabling a multi-pronged response strategy. These candidates can then be triaged using high-throughput virtual screening and prioritized for synthesis and testing based on binding predictions, ADMET properties, and chemical novelty.

We believe ProtoBind-Diff represents a key enabling technology for real-time response drug development, with the potential to dramatically compress the discovery-to-clinic timeline in future health crises.

### What's next for Gero in bringing ProtoBind-Diff into real-world drug development?

Our focus now is on operationalizing ProtoBind-Diff into Gero's internal discovery engine and partnering ecosystem. Internally, we've already integrated the model into our AI-driven drug discovery pipeline, targeting age-related diseases with high unmet need—such as fibrotic disorders, immune aging, and neurodegeneration.

We are also actively validating ProtoBind-Diff-designed compounds in vitro and in vivo. Several hits generated by the model have shown promising activity in primary screens, with optimization underway. These programs represent the first fully sequence-driven, AI-generated molecules for complex aging-related targets to enter experimental validation.

Externally, we are engaging with potential partners in pharma and biotech to co-develop drugs against challenging or novel targets. ProtoBind-Diff offers particular value to companies with proprietary protein targets but limited ligand data or no structures available. Our aim is to collaborate with these partners to shorten discovery timelines, expand target portfolios, and bring novel therapeutics to clinic faster.

Ultimately, we view ProtoBind-Diff as the foundation of a new paradigm in drug discovery—one that treats protein sequences as the universal design language of molecular therapeutics. By removing structural dependence, we unlock the full potential of AI for drug discovery—making previously intractable biology accessible and accelerating the path to new medicines.