Prospectively validated disease-agnostic predictive medicine with augmented intelligence

Bragi Lovetrue[1] & Idonae Lovetrue[1]

[1]*Demiurge Technologies AG, 14 6300, Baarerstrasse, Zug, Switzerland*

Despite numerous remarkable achievements for a variety of complex challenges,[1,2] standalone artificial intelligence (AI) does not unlock life science from the long-term bottleneck of linearly extracting new knowledge from exponentially growing new data,[3] which has severely limited clinical success rates of drug discovery.[4] Inspired by state-of-the-art AI training methods, we propose a human-centric augmented intelligence[5] (HAI) to learn a foundation model[6] that extracts all-encompassing knowledge of human physiology and pathogenesis. The quality of HAI's extracted knowledge was evaluated using a real-world validation method— the public, prospective prediction of pivotal ongoing clinical trial success outcomes at large scale (PROTOCOLS). This dataset was developed to benchmark HAI performance with an expert baseline for prospective prediction sensitivity of successful clinical programs, using only preclinical data as input. HAI achieved an 11.4-fold improvement over this baseline with a 99% confidence[7] for the PROTOCOLS validation set. The decision-support integration of HAI during the developmental candidate evaluation stage could significantly increase average clinical success rates of investigational new drugs from 7.9%[8] to 90.1% for nearly any disease.[9,10] This result confirms that exponentially extracted knowledge alone may be sufficient to accurately predict clinical success, effecting a complete

reversal of Eroom's law.[4] The trained HAI is also the world's first clinically validated model of human aging that could substantially speed up the discovery of preventive medicine for all age-related diseases.[11] This study demonstrates that disruptive breakthroughs necessitate the smallest team size to attain the largest HAI for optimal knowledge extraction from high-dimensional, low-quality data spaces, thus establishing the first prospective proof of the previous discovery that small teams disrupt[12]. The global adoption of how to train a HAI may provide a new path to mass producing scientific and technological breakthroughs via exponential knowledge extraction and better data label designs for training better-performing AI.[13]

1 INTRODUCTION

Despite recent successes,[1,2] standalone AI has not enabled life science to overcome the bottleneck of linearly extracting new knowledge from exponentially growing new data.[3] This long-term shortage of knowledge in life science has accelerated the decline of productivity in drug discovery, even after the widespread adoption of AI for biomedical data analysis.[4] For example, recent studies have shown that only 20% of biomedical research generated reproducible data[14] and fewer than 10% of animal studies successfully predicted the clinical efficacy of new drugs.[15] It has been proposed that increasing the scale of data collection or the efficiency of knowledge extraction could address these issues.[3]

High dimensionality and low quality are two defining characteristics of biomedical data, as biological systems involve large quantities of entities across multiple scales. These attributes make

it more difficult to reproduce and structure biomedical data. Deep neural networks (DNNs) are well suited for extracting knowledge from the high-dimensional, high-quality data spaces.[2] In contrast, biological neural networks (BNNs) are adept at extracting knowledge from low-dimensional, low-quality data spaces. As a result, neither DNNs nor BNNs alone are suitable for improving the efficiency of knowledge extraction from biological data.

The multi-faceted correspondence between DNNs and BNNs[5] suggests the novel possibility of combining their complementary advantages. As such, we propose adapting best practices for DNN training to a BNN, augmenting the resulting network for learning from high-dimensional data. In principle, an augmented BNN (ABNN) is well suited to learning knowledge from the high-dimensional, low-quality life science data and informing the design of improved data labels (Figure 1a). Despite the potential translatability of ABNNs into black-box DNNs, an ABNN is a quasi-white-box model necessarily hosted in a human brain. Hence, an ABNN is a form of human-centric augmented intelligence (HAI).

There are several key advantages in training an ABNN, compared with a DNN (Figure 1b). First, there is no need to redesign the network architecture or to tune network hyperparameters for the ABNN, since the human brain has been optimized to learn via evolution. Second, there is no need to pre-process data when training an ABNN because all publicly available life science data are interpretable by humans and thus compatible with the ABNN. This is possible even if most of the data are unlabeled, unstructured or uncleaned, which would be inscrutable for a DNN without adequate preprocessing. Third, and more importantly, the ABNN may be immune to overfitting

because it is sufficiently generalizable for learning from data that are categorically heterogeneous with its intended application. In contrast, overfitting is somewhat inevitable with a DNN because it lacks this generalizability and cannot learn from data that are not perfectly homogeneous with its intended application (Figure 1b). In the context of life science and drug discovery, an ABNN could be trained with biological data about normal physiological states, and then tested solely with clinical data about abnormal pathological states. The resulting ABNN could directly extrapolate from the normal states of human physiology to the abnormal states of human disease, even before research data become available on new diseases like COVID-19.

Here we introduce the first ABNN developed by training a BNN to learn like an DNN for over the course of eight years (Figure 2a and 2b). Since larger DNNs typically perform better,[16,17] we propose that the largest ABNN could only be realized by limiting the research team size to a single brain to maximally increase the efficiency of knowledge extraction (Figure 2d). As such, best practices for DNN training were adapted to a single ABNN that learns from all-encompassing life science data in an uninterrupted pass without imposing intermediate milestones (Figure 2e). The trained ABNN achieved exponential knowledge extraction from an exhaustive body of publicly accessible life science publications and databases, thereby constructing a foundation model[6] of human physiology and over 130 specific models of human disease (Figure 2c).

The quality of knowledge extracted with the ABNN was assessed by measuring resulting increases in the productivity of drug discovery. Inspired by the CASP challenge,[18] which served as a benchmark for evaluating DNN performance against expert performance in protein structure

prediction, we designed a real-world validation method—the public prospective prediction of pivotal ongoing clinical trial success outcomes at large scale (PROTOCOLS) challenge. This method provided a benchmark for comparing ABNN and expert performance in clinical success prediction (Figure 3a, 3b and Methods section 'Validation design').

The rationale for the PROTOCOLS challenge is that pivotal clinical trials (phase 2 and phase 3) are statistically powered to assess a drug's clinical success by evaluating its clinical efficacy and safety, which is largely determined by the underlying biological processes that can be distilled into scientific knowledge. Historical clinical success rates for investigational drugs are well-documented[8,19] and provide an accurate baseline for evaluating ABNN performance in the PROTOCOLS challenge.

Clinical success prediction is a critical go/no-go decision step in drug development because only a subset of developable drug candidates, with highest clinical success potential, are selected to transition from non-clinical to clinical development, due to the prohibitive cost of human clinical trials.[20] As a best practice in the biopharmaceutical industry,[21] a group of committed experts reviews both publicly available and privately accessible preclinical data to predict which drug candidates will achieve clinical success. Human clinical trials are initiated only for those drug candidates that are prospectively predicted to be clinically successful (Figure 3a). However, real-world industrial practice has yielded a prospective prediction sensitivity of 7.9%,[19] which is only slightly better than chance.[22]

The PROTOCOLS validation method is designed to mirror the real-world industrial practices

to minimize the performance gap from validation to application. The included ABNN accessed a publicly available subset of expert-accessible preclinical data, used to predict which drug candidates would achieve clinical success. Clinical trial readouts were tracked for all drug candidates that had been prospectively predicted to be either clinically successful or not (Figure 3a). In contrast to the expert baseline performance of 7.9% (Methods section 'Baseline performance'), the ABNN followed identical procedures with fewer data yet yielded a prospective prediction sensitivity of 90.1%, representing an 11.4-fold improvement over the baseline in realistic settings.

The PROTOCOLS validation method was used to evaluate the real-world performance of an ABNN in actual clinical settings with an emphasis on generalizability to novel drug-disease pairs, even with scarce or uninformative prior clinical data (Figure 3b). As such, there is little difference between the validation performance of the ABNN in the PROTOCOLS challenge and the clinical utility of the ABNN in real-world development candidate selection.[23,24]

2 RESULTS

The PROTOCOLS validation set comprised 265 then-ongoing real-world pivotal clinical trials following a predefined set of screening criteria (Figure 3c and Methods section 'Validation data collection').

Validation clinical trials and newly initiated clinical trials since 2020 were shown to follow identical distributions across all major therapeutic areas (two-sample Kolmogorov-Smirnov test, statistic = 0.333; P = 0.73; Figure 4). PROTOCOLS thus provided an unbiased evaluation of

disease-agnostic performance for the ABNN.

Minimum sample size (131) and event size (118) requirements were determined for evaluating ABNN performance with 99% confidence[7] (Figure 3c and Methods section 'Statistical analysis'). Both requirements were met as clinical readouts were available from 157 of 265 predictions by the cutoff date (Methods section 'Clinical readout'). A ground truth was then determined for each readout and was then applied to validation of the corresponding prediction (Methods section 'Prediction validation'), producing a set of confusion matrices for varying subgroups of the validation set (Figure 7).

It should be noted that positive instances (true positives and false negatives) are far more important than negative instances (true negatives and false positives) in real-world drug discovery. True positives are also more interesting than true negatives because the historical clinical success rate for investigational drugs entering human studies is only 7.9% across all diseases[8]. As such, false positives have only a marginal clinical impact because human clinical trials can still fail without doing harm to patients. In contrast, false negatives have a substantial clinical impact because they not only deny clinical benefits to patients, but also eliminate sizable revenue opportunities.

Accordingly, we chose sensitivity and F1 score as appropriate performance metrics for the PROTOCOLS validation.[25] Our choice of sensitivity realistically reflects best practices in actual drug development. Every investigational new drug greenlit to enter first-in-human studies is then genuinely believed by qualified experts to be capable of achieving clinical success. As such, the clinical success rate from phase 1 to approval is construed as a statistical measure of sensitivity.[19]

We also report other standard measures (accuracy, specificity, etc.) to provide a balanced view of ABNN performance (Data Table 2).

Since the ABNN was trained to functionally reconstruct human physiology and human pathogenesis (Methods section 'Model training'), its disease-agnostic performance was first assessed for the full validation set (157 predictions; Supplementary Table 1). The ABNN achieved a 90.1% prospective prediction sensitivity (99% CI 80.0%, 96.9%; $P < 0.001$; Figure 5a;), a 11.4-fold improvement over the realistic baseline sensitivity of 7.9% and a 6.0-fold improvement over the conservative baseline sensitivity of 15.1% (Figure 5a; Methods section 'Baseline performance'). It also achieved 0.86 F1 score (99% CI 0.79, 0.92; $P < 0.0001$; Figure 6c; Data Table 1), 82.8% accuracy (99% CI 74.2%, 89.8%; $P < 0.0001$; Figure 6c; Data Table 1), and 72.7% specificity (99% CI 57.5%, 85.3%; $P < 0.0001$; Figure 6b; Data Table 1). These data clearly demonstrate that the ABNN has extracted large-scale, high-quality knowledge, enabling its excellent disease-agnostic performance in the real-world settings of the PROTOCOLS validation.

Further analysis of subgroups was conducted for the validation set, with varying degrees of clinical data availability. In theory, the ABNN should not require access to any clinical data prior to making predictions of clinical outcomes for investigational drug candidates (Figure 3a).

First-in-class clinical trials for novel drug-disease pairs represent a fairly challenging subset of the PROTOCOLS validation process because no prior clinical data are available for those drug-disease pairs across all therapeutic areas. Since PROTOCOLS included 128 first-in-class clinical trials with clinical readouts, the minimum sample and event size requirements for a 99% confidence

interval (CI) have been met. The ABNN achieved 92.4% prospective prediction sensitivity (99% CI 80.8%, 99.2%; $P < 0.009$; Figure 5b, 6b; Data Table 1), 0.85 F1 score (99% CI 0.76, 0.92; $P < 0.0001$;Figure 6c; Data Table 1), 82.5% accuracy (99% CI 72.8%, 90.3%; $P < 0.0001$; Figure 6c; Data Table 1), and 71.7% specificity (99% CI 55.6%, 85.0%; $P < 0.0001$; Figure 6b; Data Table 1).

Notably, the ABNN exhibited superior disease-agnostic prospective prediction sensitivity for first-in-class clinical trials, compared with all clinical trials ($\Delta = 2.3\%$; 99% CI -0.92%, 5.52%; $P < 0.002$ for superiority at a 2% margin; Figure 5b; Methods section 'Statistical analysis'). This accuracy was also non-inferior to its performance for all clinical trials ($\Delta = 0.3\%$; 99% CI -0.87%, 1.47%; $P < 0.0001$ for non-inferiority at a 2% margin; Methods section 'Statistical analysis'). These results clearly demonstrate that the outstanding performance of the ABNN for PROTOCOLS validation was independent of clinical data availability for arbitrary drug-disease pairs.

Clinical trials in oncology represent a unique challenge for the ABNN, as they conventionally suffer from the lowest clinical success rates. The realistic baseline sensitivity for oncology is 5.3%, in contrast to the highest sensitivity of 23.9% for hematology[8] (Methods section 'Baseline performance'). Consequentially, the importance of preferred positive instances is even higher in oncology than in other therapeutic areas, to the reasonable extent that negative instances are no longer interesting. As such, a clinically meaningful assessment of ABNN performance in oncology should primarily focus on sensitivity, F1 score, and accuracy.

In addition, clinical trials in oncology often include quite complex trial designs involving

active controls and combination therapies as a standard design for pivotal clinical trials. Human cancer patients are further stratified by certain biomarkers,[26] stages of cancer, and prior treatment history, all of which add to the complexity of scientific knowledge required to make accurate predictions of drug clinical efficacy and safety.

PROTOCOLS validation set included 48 clinical trials in oncology with clinical readouts, satisfying the minimum sample and event size requirements for a 90% CI have been met (see Methods section 'Statistical analysis'). The ABNN achieved 88.9% prospective prediction sensitivity (90% CI 75.7%, 94.3%; $P < 0.015$; Figure 5c, 6b; Data Table 1), a 16.8-fold improvement over the realistic baseline sensitivity of 5.3%, and an 8.2-fold improvement over the conservative baseline sensitivity of 10.8% for oncology clinical trials (Methods section 'Baseline performance'). Remarkably, the ABNN achieved 0.85 F1 score (90% CI 0.77, 0.90; $P < 0.0004$; Figure 6c; Data Table 1) and 77.1% prospective prediction accuracy (90% CI 65.1%, 84.9%; $P < 0.0004$; Figure 6c; Data Table 1). These data clearly demonstrate the ABNN has extracted sufficiently complex and intricately nuanced knowledge of the human body and cancer etiology to accurately predict the clinical success of cancer drugs in biomarker-differentiated heterogeneous patient populations.

COVID-19 represents the ultimate test of existing knowledge quality for the ABNN because no knowledge of COVID-19 can be extracted directly, due to a lack of prior data shortly after the outbreak. As such, the entire disease model for COVID-19 can only be constructed by generalizing other knowledge to COVID-19 clinical symptoms. In addition, no clinical data were available worldwide when the preprint of the ABNN COVID-19 disease model was published on March 20,

2020[27].

As there are 49 infectious disease clinical trials in the validation set, the minimum sample size and event size requirements for a 90% CI have been met (see Methods section 'Statistical analysis'). In total, 94% (46/49) are pivotal COVID-19 clinical trials (Supplementary Table 4).

The ABNN achieved 88.9% prospective prediction sensitivity (90% CI 68.3%, 95.4%; $P < 0.05$; Figure 5c; Data Table 1) for infectious diseases in the validation set. This represents a 6.7-fold improvement over the realistic baseline sensitivity of 13.2% and a 3.9-fold improvement over the conservative baseline sensitivity of 22.8% for infectious disease clinical trials (Methods section 'Baseline performance'). The ABNN also achieved 0.78 F1 score (90% CI 0.65, 0.86; $P < 0.0012$; Figure 6c; Data Table 1) and 81.6% prospective prediction accuracy (90% CI 70.1%, 88.4%; $P < 0.002$; Figure 6c; Data Table 1). These data clearly indicate the quality of extracted knowledge in the ABNN has reached a critical mass, enabling de novo generation of comprehensive knowledge for a new disease like COVID-19 while maintaining the same quality as that of the new knowledge extracted from prior biological data.

The ABNN achieved 92.4% 7.8% prospective prediction sensitivity (Figure 5c, 6b; Data Table 2) and 0.91 0.10 F1 score (Figure 6d; Data Table 2) across all other therapeutic areas in the PROTOCOLS validation set (60 predictions in total; Supplementary Table 7). Although the minimum requirements for a 90% CI were not met for these data (Methods section 'Statistical analysis'), the results demonstrate that ABNN performance is stable and consistent.

Alzheimer's disease (AD) was also included in the PROTOCOLS validation and served to further demonstrate the robustness of extracted knowledge in the ABNN. AD represents an extreme case because the historical clinical success rate of AD-modifying drugs is nearly zero.[28] A single true positive then becomes an appropriate performance metric for AD because both false positive and false negative instances are equally uninteresting in real-world settings. As such, AD represents a significant challenge for the ABNN because of a consensus that it is overly complex, the quantity of existing data is too low, and the quality of existing data is poor.[28]

Despite these limitations, the ABNN achieved 100% accuracy and 100% precision with non-zero true positives and no false positives (Supplementary Table 6) for three prospective predictions of AD pivotal clinical trials during PROTOCOLS validation. The first true positive was Masitinib, a first-in-class c-Kit inhibitor that met the efficacy primary endpoint in a pivotal phase 3 trial (NCT01872598). These results demonstrate the potential of the ABNN for extracting large-scale, high-quality knowledge of AD from existing data.

Aging is an evolutionarily conserved effect across all mammalian species[29] and is clinically recognized as the primary risk factor for numerous diseases in multiple therapeutic areas.[30] The quality of extracted aging-specific knowledge in the ABNN was evaluated by identifying all age-related pivotal clinical trials in the PROTOCOLS validation set (Methods section 'Validation data collection') and corresponding ABNN performance (Figure 8).

As there are 119 age-related clinical trials of age-related diseases in the validation set, the minimum sample and event size requirements for a 99% CI have been met. The ABNN achieved

91.0% prospective prediction sensitivity (99% CI 79.1%, 98.4%; P < 0.005; Data Table 1; Figure 9a), 0.85 F1 score (99% CI 0.76, 0.92; P < 0.0001; Data Table 1; Figure 6c), 81.5% prospective prediction accuracy (99% CI 71.3%, 89.7%; P < 0.0001; Data Table 1; Figure 6c), and 69.2% specificity (99% CI 51.8%, 83.9%; P < 0.0001; Data Table 1; Figure 6b).

Remarkably, ABNN prospective prediction sensitivity for all age-related clinical trials was non-inferior to its performance for all first-in-class clinical trials (Δ = 1.4%; 99% CI -1.12%, 3.92%; P < 0.003 for superiority at a 2% margin; Methods section 'Statistical analysis'). ABNN prospective predictive accuracy for age-related clinical trials was also shown to be non-inferior to its performance for first-in-class clinical trials (Δ = 1.0%; 99% CI -1.14%, 3.14%; P < 0.0001 for non-inferiority at a 2% margin; Methods section 'Statistical analysis').

ABNN prospective predictive sensitivity for all age-related clinical trials was statistically significantly higher than its performance for all clinical trials in the validation set (Δ = 0.9%; 99% CI -1.13%, 2.93%; P < 0.0001 for superiority at a 2% margin; Figure 9a; Methods section 'Statistical analysis'). These results clearly demonstrate the ABNN has extracted sufficient knowledge of aging, which contributed significantly to its consistent performance across all the age-related indications in the PROTOCOLS validation set (Figure 9b).

We also investigated the relationship between the quality of extracted knowledge in the ABNN and the difficulty of drug discovery per therapeutic area, by calculating correlations between F1 scores and realistic baseline sensitivities (or LOA: historical clinical success rates for investigational new drugs since phase 1. See Methods section 'baseline performance') for all val-

idation subgroups (Figure 6d). F1 scores and LOA were negatively and weakly correlated (Exact Pearson r = -0.260), indicating that the quality of ABNN knowledge concerning human pathogenesis is orthogonal to the quality of collective knowledge for the corresponding therapeutic areas. Surprisingly, ABNN performance was higher for increasingly difficult therapeutic areas (Figure 6d; Data Table 2).

Finally, we investigated the relationship between the quality of extracted knowledge in the ABNN and drug modality (Figure 10). The ABNN achieved 91.5% prospective prediction sensitivity (90% CI 80.8%, 95.7%; $P < 0.015$; Figure 11a; Data Table 1) for biologics (68 biological trials; Methods section 'Drug classification') and 85.7% prospective prediction sensitivity (90% CI 71.9%, 92.2%; $P < 0.009$; Figure 11a; Data Table 1) for small molecule drugs (77 NME trials; Methods section 'Drug classification'). This performance was consistently high across drug modalities and even higher for biologics than for small molecules ($\Delta = 5.8\%$; 90% CI 0.78%, 10.82%; $P < 0.0001$ for superiority at a 2% margin; Figure 11a; Methods section 'Statistical analysis'). We further calculated correlations between F1 scores and realistic baseline sensitivities (or LOA: historical clinical success rates for investigational new drugs since phase 1. See Methods section 'baseline performance') across drug modalities. F1 scores and LOA were positively and strongly correlated (Exact Pearson r = 0.975; Figure 11b), indicating that the quality of extracted knowledge in the ABNN may be a natural fit for drug modalities with a higher target specificity.

3 DISCUSSION

In this study, the possibility of a deep-learning-augmented human intelligence (HAI) has been presented for constructing a foundation model of human physiology, potentially enabling highly accurate data-free predictions of clinical success for investigational new drugs applied to almost any disease. This achievement is intractable for data-driven machine intelligence, on which the demands for clinical data are unrealistically high and the efficiency of knowledge extraction is sufficiently low that gaps between promise and proof have become nearly unbridgeable (Figure 1).

Humans have designed machine learning methods to train deep neural networks that have successfully learned black-box models, outperforming human experts in playing games[1] and predicting protein structures.[2] Our work shows that a similar approach could be adapted to training augmented biological neural networks (ABNNs) in learning quasi-white-box models[27] and predicting the clinical success of investigational new drugs (Figure 1b).

The resulting ABNN could safely deliver AI-surpassing clinical utility while averting the privacy and trustworthiness challenges faced by medical AI systems.[24] The ABNN could also inform the better design of labels for training better-performing DNNs with fewer data.[13]

Furthermore, in contrast to domain experts who have privileged access to a large body of private preclinical data, the presented ABNN was trained solely using public preclinical data yet extracted more knowledge than previous attempts by experts (Figure 3a).

PROTOCOLS represents the most rigorous validation design to date for evaluating the real-world utility of extracted knowledge from large-scale, low-quality data spaces (Figure 3 and Figure 4). With a 99% CI, the trained ABNN achieved a prospective prediction sensitivity of more than 90%, an F1 score of more than 0.85, and a prospective prediction accuracy of more than 82% across all therapeutic areas, independent of the availability of prior clinical data (Figure 5 and Figure 6; Data Table 2).

In addition, the drug mechanism of action and clinical trial design protocols were the only inputs used by the ABNN to evaluate the clinical success of novel drugs (Figure 3a). ABNN performance in PROTOCOLS confirmed that large-scale, high-quality scientific knowledge alone may be sufficient to accurately predict drug clinical success.

Given that both ABNN inputs are available prior to first-in-human studies and positive instances matter far more than negative instances in actual drug discovery and development, positive predictions of drug clinical success alone could serve as a decision-support intelligence for the selection of drug candidates for clinical development, significantly increasing the clinical success rates for investigational new drugs from 7.9%[8] to 90.1% and ushering a permanent reversal of Eroom's law[4] (Figure 3a).

The trained ABNN is also the world's first clinically validated model of human aging (Figure 8 and Figure 9). Its application to the development of cellular rejuvenation programming could substantially speed up the discovery of preventive medicines for age-related diseases.[11]

The ABNN also leaves room for improvement as 17.2% (27/157) of predictions were either false positives (18/157) or false negatives (9/157) in the PROTOCOLS validation. As such, further work is planned to investigate why the ABNN performs better for harder diseases (Figure 6d). The exponential efficiency of knowledge extraction enabled the ABNN to translate each false prediction into substantial improvements in the corresponding disease model.

Another potential limitation of our work is that the ABNN was validated in PROTOCOLS using ongoing pivotal clinical trials whose drug candidates had already been pre-selected by experts to enter clinical development. ABNN performance in development candidate selection may thus be conditioned upon prior expert performance, the extent of which could be determined in future rounds of PROTOCOLS validation enabled by an expanded access to private preclinical data, allowing for the ABNN to predict the clinical success for the drug candidates that are crossed out by experts yet still enter clinical-stage development based on predictions by the ABNN.

Nevertheless, the possible conditioning of the ABNN's performance upon prior expert performance in development candidate selection should be seen as an application advantage. The ABNN enables a zero-change integration with any standard workflow of clinical development as an add-on screening of expert-shortlisted development candidates, without making any modifications of current best practices. The ABNN serves as an augmented intelligence in collaboration with experts, in contrast to artificial intelligence systems that are often positioned in competition with or even in substitution for experts.[24]

While this work could significantly reshape medicine, its broader impact lies in the pre-

sented training methodology, which provides a new path for drastically increasing the efficiency of knowledge extraction in every discipline that is characterized by large-scale, low-quality data spaces (e.g., synthetic biology, material science, and climate science). These results also provide the first prospective proof that the smallest research team is necessary for delivering the most disruptive breakthroughs.[12]

The advent of the ABNN represents a new era of human-centric augmented intelligence (HAI). The global adoption of ABNN training could shift research culture and organizational structure towards a potential mass production of scientific and technological breakthroughs. ABNNs could give rise to general purpose technologies that could potentially overcome the productivity J-curve and paradoxes recurrently observed in the past.[31]

4 METHODS

TRAINING DATA COLLECTION

Two types of training data were collected for different stages of training the ABNN.

In stage 1, the ABNN was trained to learn a foundation model of the normal states of human body to capture the latent complexity of human physiology. The stage 1 training data covered all publicly accessible sources of biological data that could inform normal physiological processes in human body, including cross-species data about evolutionarily conserved biological building blocks (Methods section ‘Model training’).

In stage 2, the ABNN was trained to learn disease-specific models of the abnormal states of human body that should match the clinical scope of human pathogenesis. The stage 2 training data covered all publicly accessible sources of biological data that investigated disease etiology in cell cultures and animal models, as well as medical data that reported disease-specific co-morbidities and symptoms on human patients. Nevertheless, no clinical trial data were collected for training the ABNN because they were designed for drawing causal inferences about drug-target interactions under the assumption that human body is a black-box system, thus uninformative for modelling the underlying mechanisms of human pathogenesis (Methods section 'Model training').

MODEL TRAINING

Inspired by the multi-faceted functional correspondence between deep neural networks (DNNs) and biological neural networks (BNNs),[5] we adapt the best practice of training DNNs into a standard protocol of training BNNs to learn representations from large-scale data spaces like DNNs, resulting in augmented biological neural networks (ABNNs) as a form of replicable augmented intelligence. We use our validated ABNN in life science as an example for illustrating each step in the training protocol:

Training Data: DNNs learn better representations of data with more levels of abstraction than with fewer levels, and BNNs share the hierarchical architecture of DNNs.[5,32] Accordingly, ABNNs must be exposed to the full breadth and depth of multi-omics data (e.g., genome, proteome, transcriptome, epigenome, metabolome, and microbiome) to capture the latent complexity of human physiology matching the clinical scope of human pathogenesis (Figure 2a). More specifically,

ABNNs extract right knowledge by being exposed to the entire English corpus of non-clinical-trial biomedical publications in journals whose impact factors were stably greater than two for the most recent three consecutive years (including but not limited to: Science, Nature, Cell, Nature Neuroscience, Nature Metabolism, Cell Stem Cell, Neuron, PNAS, eLife, Nature Communication, Scientific Reports, etc.), as well as public repositories of curated databases (including but not limited to: Human Protein Atlas, Allen Human Brain Atlas, The Cancer Genome Atlas, etc.).

Learn ing Rule: DNNs use backpropagation to improve learned representations of data by sequentially updating its internal parameters from the highest to the lowest level of abstraction. The ubiquitous feedback connections in BNNs may implement backpropagation-like learning rules (Figure 2b). ABNNs extract fast knowledge by iteratively updating internal representations in the strict order from the most macroscopic level (phenotype) through the intermediate level (endophenotype) to the most microscopic level (genotype). ABNNs were trained to zero in on a single area of interest across all levels of abstractions until they completed the backpropagation-like learning for that area before moving on to the next. For example, when cortical columns for motor control were the area of interest for training, our ABNN translated every piece of training data into a sequential update of internal representations in the ascending order of granularity from the phenotype level (e.g., muscle contractions), to the endophenotype level (neuronal circuits responsible for muscle contractions), to the cell level (e.g., Layer 5 pyramidal neurons (L5PT) and Layer 2/3 inhibitory interneurons (L2/3IN) responsible for muscle contractions), to the protein level (e.g., NMDA receptor subtypes and GABA receptor subtypes specifically expressed in L5PT and L2/3IN for muscle contractions), to the epigenotype level (e.g., DNA methylation profiles in connection

with subtype-specific NMDAR and GABAR activities in L5PT and L2/3IN for muscle contractions), and finally to the genotype level (e.g., the potential role of the CLOCK gene expression on subtype-specific NMDAR and GABAR activities in L5PT and L2/3IN for muscle contractions).

Learning Rate and Loss: DNNs learn general-purpose representations of all-encompassing data to deliver maximal performance in a wide range of special-purpose tasks.[6] Both BNNs and DNNs could learn to command general linguistic abilities for numerous tasks (Figure 2c). ABNNs extract full-scope knowledge by first learning a general-purpose model of human physiology from the all-encompassing non-disease biological data (Stage 1; Methods section 'Training data collection') before starting to learn numerous human diseases models from disease-specific data (Stage 2; Methods section 'Training data collection'). In Stage 1, we set the learning rate low and the loss small so that the ABNN was incentivized to do multiple rounds of learning without being sensitive to the learning performance of general human physiology in each round. In Stage 2, we set the learning rate high and the loss large so that the ABNN was incentivized to learn a complete model of a specific human disease in a single round. The ABNN was thus incentivized to re-run Stage 1whenever the ABNN's performance was suboptimal in Stage 2. As a result, the ABNN optimized the global extrema of the disease-agnostic physiology model before starting to optimize the local extrema of any disease-specific pathogenesis model.

Network Size: DNNs maximize the robustness of learned representations of data by increasing the sheer size of network parameters.[17] ABNNs extract novel knowledge if a single ABNN of the largest size learns from the entirety of the all-encompassing data, rather than multiple ABNNs

of smaller sizes learn from siloed sub-datasets (Figure 2d). In practice, this means that we should minimize the number of ABNNs in a research team to one, to maximize the number of ABNN neuronal nodes interconnected by always-on high-bandwidth biological synapses in a single brain, rather than by the intermittently-on low-bandwidth human languages between several brains.

Training Environment : DNNs learn the best representations of data given sufficient time and ample resources to allow for an uninterrupted pass of the full training data before which no meaningful performance milestones can be defined. ABNNs extract testable knowledge if a milestone-free supply of time and fund is guaranteed to ensure uninterrupted learning so that ABNNs won't be evaluated at all before the completion of learning both human physiology and human diseases (Figure 2e). Securing the optimal environment was the most challenging aspect of training ABNNs because it would be extremely difficult to build a research culture and implement terms and protocols such that ABNNs are willing to receive uninterrupted training that may last for a decade without intermediate milestones.

Overfitting Prevention: The ABNN may be immune to overfitting because it is sufficiently generalizable for learning from data that are categorically heterogeneous with its intended application. In contrast, overfitting is somewhat inevitable with a DNN because it lacks this generalizability and cannot learn from data that are not perfectly homogeneous with its intended application (Figure 1b). In the context of life science and drug discovery, the ABNN was first trained with biological data about normal physiological states, and then tested solely with clinical data about abnormal pathological states. The ABNN had no access to disease-specific clinical data until and

unless the etiology of a specific disease was entirely generated de novo from the foundation model of human physiology in the first place. The resulting ABNN should be able to directly extrapolate from the normal states of human physiology to the abnormal states of human disease, which would otherwise be impossible if the ABNN overfits the training data.

VALIDATION DESIGN AND METHODOLOGY

Validation is critical for an objective assessment of ABNN's capabilities in exponentially extracting high-quality knowledge from life science data (Figure 3a and 3b). The simplest approach is to determine the extent to which the ABNN could reverse Eroom's law. However, it would be unrealistically time-consuming and non-scalable to directly begin new drug development. As such, running virtual clinical trials is an effective alternative. Drug mechanisms of action and clinical trial design protocols would serve as the only inputs and the ABNN could be used to predict the clinical efficacy and safety of new drugs solely using extracted knowledge, without access to patient data or confidential third-party information. Highly accurate virtual clinical trials could be run before first-in-human studies to prevent drug developers from initiating pivotal real-world trials that are doomed to fail.[33]

A rigorous validation of the ABNN, including the stringent criteria from the PROTOCOLS challenge as shown in Figure 3b, would involve publishing prospective predictions for pivotal clinical trial outcomes across all human diseases. Tamper-proof timestamps would be included and prediction sensitivity could be determined with a 99% CI, based on a ground truth determined by actual clinical readouts. No selection of lead over non-lead indications would be included,

although non-lead indications have a far lower success rate.[19]

There are a few hundred new pivotal clinical trials each year whose protocols are publicly accessible in a centralized database (as required by the FDA[34]) and whose clinical readouts are typically published by sponsors in a standardized format. Public, prospective predictions of pivotal ongoing clinical trial outcomes could serve as a validation benchmark for the ABNN, similar to the way in which ImageNet provided a benchmark validation for visual object recognition using DNNs.[35]

BASELINE PERFORMANCE

Prospective prediction sensitivity (PPS) is defined as the percentage of clinical programs that have achieved actual clinical success by demonstrating sufficient clinical efficacy and safety in pivotal clinical trials that warrant an FDA approval, given that clinical programs are initiated only for those drug candidates that are prospectively predicted to be clinically successful by experts.

Realistic baseline sensitivity (RBS), as a realistic estimate of PPS, is defined as the probability of FDA approval for drugs in phase 1 development (Phase1 likelihood of approval, or P1LOA). It was calculated using corresponding P1LOAs in a 2011-2020 survey.[8] This baseline sensitivity is realistic as it acknowledges that an ABNN would make identical predictions at earlier stages of development with drug mechanisms of action and clinical trial design protocols serving as the only inputs, both of which are available as early as the preclinical stages.

Conservative baseline sensitivity (CBS), as a conservative estimate of PPS, is defined as the probability of FDA approval for drugs in phase 2 development (Phase2 likelihood of approval, or P2LOA). It is calculated as the corresponding P2LOA in a 2011-2020 survey.[8] This baseline sensitivity is conservative because it does not consider that an ABNN only requires two inputs to generate predictions (drug mechanisms of action and clinical trial design protocols), both of which are readily available in the real world even before first-in-human studies (phase 1). Regardless, this baseline sensitivity is more balanced in consideration of both the phase distribution for clinical trials in the validation set (157 clinical readouts (3.18% (5/157) phase 1 trials, 43.9% (69/157) phase 2 trials, and 54.1% (85/157) phase 3 trials, and the irrelevance of phase transitions in making predictions.

RBS was not directly available from previous survey studies[8,19] for two subgroups of the PROTOCOLS validation set: all first-in-class indications (128 trials) and all age-related indications (119 trials). As diseases in these two subgroups originate from therapeutic areas featuring lower-than-average historical clinical success rates (oncology, neurology, cardiovascular, etc.), the real RBSs for both subgroups should be lower than those of all indications (7.9%[8]). We therefore assumed that a reasonable estimate of RBS for the subgroup containing all first-in-class indications (128 trials) would be 6.32%, equal to 80% of the RBS for all indications (7.9%[8]). Since no anti-aging therapeutics have ever been clinically approved, we assumed that a reasonable estimate of RBS for the subgroup of all age-related indications (119 trials) would be 5.69%, equal to 72% of the RBS for all indications (7.9%[8]).

VALIDATION DATA COLLECTION

After completing the two-stage training, the ABNN turns out to be a 'virtual patient' that functionally reconstruct both human physiology and human pathogenesis with sufficient fidelity. The trained ABNN is thus clinically equivalent to a real human body such that 'viritual clinical trials' could be run on the ABNN to evaluate drug-body interactions and predict clinical success in human patients. Clinical trial data were thus all reserved for testing the fully trained ABNN in the real-world PROTOCOLS validation.

The validation set comprised the registration information for human clinical trials on www .clinicaltrials.gov, with the screening criteria as shown in Figure 3c (phase 1/2 is categorized as phase 1, and phase 2/3 is categorized as phase 2). We estimated that there are approximately 200-300 clinical trials each year that would meet these screening criteria, which is consistent with the 424 annual clinical trials exhibiting a much looser set of screening criteria (no limit on primary completion date, no preference of first-in-class trials, etc.) in a comprehensive survey study.[19] See the supplementary tables for the full validation data.

A clinical trial in the validation set was designated as "first-in-class" if the drug-indication pair had not been approved for marketing worldwide.

A clinical trial in the validation set was designated as "age-related" if age was a clinically identified risk factor for the indication thereof.

The ABNN is principally disease-agnostic yet leaves room for improvement in identifying hematological disorders, rare neurodevelopmental disorders, and rare musculoskeletal disorders. These unfinished disorders were excluded from the performance validation step yet were included in validation data collection because prospective predictions can provide additional insights to accelerate model building for such disorders.

PREDICTION GENERATION

Per industry-wise best practices,[36] the ABNN predicted SUCCESS for a pivotal clinical trial if at least one of its pre-specified efficacy primary endpoint(s) was believed to be met with predefined statistical significance. In contrast, the ABNN predicted FAILURE for a pivotal clinical trial if none of its pre-specified efficacy primary endpoint(s) were believed to be met with predefined statistical significance (Figure 3d).

If a pivotal clinical trial had multiple co-primary endpoints, PARTIAL SUCCESS or PARTIAL FAILURE was predicted if at least one of its pre-specified efficacy primary endpoint(s) was believed to be met with predefined statistical significance and at least one of its prespecified efficacy primary endpoint(s) was believed to not be met with predefined statistical significance.

PREDICTION PUBLICATION

A total of 265 predictions were published as timestamped tweets @DemiurgeTech to establish a publicly accessible track record for prospective predictions (Supplementary Figure 1). The

immutability of tweets ensures that each published prediction could not be post hoc modified. We also indexed each tweet to prevent any intentional deletions of false predictions that go unnoticed, to ensure a reliable performance evaluation. However, index gaps do exist for several legitimate cases, which does not compromise the rigor of validation for the following reasons. 1) An index was unintentionally skipped and remained unused from that point forward. 2) An indexed prediction was made but later became disqualified as a prospective prediction, because the corresponding result had been announced after the indexed prediction was made but before the indexed prediction was published. We used DocuSign to distinguish these two cases of index gaps by electronically signing every indexed prediction before publishing it as a tweet. As such, no DocuSign certificates are available for unused indices. The indices 001 and 286 denote the indices of the first and the last published prospective predictions, respectively. A total of 21 index gaps were detected, 5 of which were unused and 5 of which were for disqualified predictions lacking clinical readouts as of the cutoff date (March 1, 2022). Nine predictions were made for disqualified predictions with clinical readouts as of the cutoff date (3 false predictions and 6 true predictions). Two predictions were for qualified unpublished predictions whose results are still not available. DocuSign certificates for disqualified and qualified unpublished predictions are available upon written request (Figure 3d).

CLINICAL READOUT

It is customary for private companies and mandatory for public companies in the biopharmaceutical industry to forecast the estimated date of clinical readouts and to publish the actual results from human clinical trials. We tracked the availability of clinical readouts for every pub-

lished prospective prediction from multiple online sources, including https://www.globenewswire.com/ for press releases, and https://www.biospace.com/ for curated news, and financial reports from public companies. We published links to available clinical readouts as timestamped tweets by replying to the original timestamped tweet for the corresponding prediction, thus establishing a clear timeline of predictions followed by readouts (Figure 3d).

GROUND-TRUTH DETERMINATION

Under the FDA guidelines,[37] a clinical trial with available readouts is assigned the ground-truth label success if and only if at least one efficacy primary endpoint is met according to interim/topline analyses or final data published by the trial sponsors (Figure 3d).

Similarly, a clinical trial with available readouts is assigned the ground-truth label failure if and only if (1) none of the efficacy primary endpoint(s) are met according to topline analyses or final data or none are deemed likely to be met according to interim analyses; (2) the trial is terminated for non-recruitment or non-business reasons; (3) the trial is terminated after the failure of other clinical trials for identical drug-disease pairs; (4) the post-trial development is discontinued for non-business or unspecified reasons; (5) the trial is dropped out from the pipeline without specific reasons, all based on the official websites or announcements by the trial sponsors.

Given that the reliability of ground-truth determination is capped by an 85.1% theoretical maximum prediction accuracy for drug clinical success in human clinical trials,[38] we refrained from directly labeling missed predictions as false without further consideration of subsequent de-

velopments. Specifically, we considered missed failure predictions for phase 2b trials that were less statistically powered than phase 3 trials. Missed failure predictions should thus be more likely to be true rather than false if the trial sponsor decided not to conduct a phase 3 trial for non-business or unspecified reasons.

PREDICTION VALIDATION

A prospective prediction of success was validated as TRUE if its corresponding ground-truth label was also success or invalidated as FALSE if its corresponding ground-truth label was failure. Similarly, a prospective prediction of failure was validated as TRUE if its corresponding ground-truth label was also failure or invalidated as FALSE if its corresponding ground-truth label was success. Confusion matrices were produced to assess ABNN validation performance on a disease-agnostic basis and are shown per therapeutic-area basis in Figure 7.

DRUG CLASSIFICATION

Drugs in the PROTOCOLS validation set were generally categornized using the FDA classification codes for new drug applications: new molecular entity (NME), biologic, and non-NME codes compromising vaccines and other modalities.

NMEs are novel small molecule drugs or novel combinations of multiple old or new small molecule drugs. Biologics comprise a broad range of drug types such as antibodies, peptides, RNAi therapies, cell therapies, gene therapies, and microbiome therapies. Non-NMEs include

vaccines, cytokines, enzymes, insulins and other reformulation of approved drugs.

STATISTICAL ANALYSIS

ABNN performance in clinical success prediction was evaluated using the public, prospective prediction of pivotal ongoing clinical trial success outcomes at large scale (PROTOCOLS) challenge, a rigorous external validation of a clinical prediction model with a binary outcome.[7]

Multiple reasonable assumptions were included to calculate minimal sample and event sizes, for the determination of prospective prediction sensitivity, F1 score, accuracy, and specificity at a 99% CI inspired by the previously reported method.[7]

The anticipated clinical readout event proportion (φ) was set to 0.9 since, as of the cutoff date, an average of 75% of clinical trials on clinicaltrials.gov have reported clinical trial results in compliance with the FDAAA 801 and the Final Rule tracked by fdaaa.trialstracker.net. More than 90% of big-pharma-sponsored clinical trials report results.[39]

The ratio of total observed clinical readout events and total expected clinical readout events (O/E) was set to 1 with a width of 0.15 to account for the 12% of clinical trials that are prematurely terminated without clinical trial results.[40]

As such, the minimum sample size requirement was set to 131 predictions, and the minimal event size requirement was 118 readouts for a 99% CI. From February 11, 2020 to December 22, 2020, 265 prospective predictions were published to meet this minimal sample size requirement.

By the time of the study cutoff date (March 1, 2022), 157 clinical readouts (5 phase 1 trials, 68 phase 2 trials, and 85 phase 3 trials) were available (Supplementary Table 1). Thus, the minimum event size requirement has been met.

All other factors were held constant and performance measures were detected at a 90% CI for validation subgroups with smaller event sizes, using a minimum requirement of 52 predictions and a required event size of 47 readouts. Only simple means and standard deviations were calculated for validation subgroups with even smaller event sizes over corresponding proportions.

The direct translatability of prospective validations to clinical ABNN applications was evaluated using a two-sample Kolmogorov-Smirnov test, the standard approach for comparing validation clinical trials and initiated clinical trials since 2020. These results followed an identical distribution (see Figure 4) and the oncology data in the validation set exhibited the same proportion as in the benchmark.[19]

Statistical significance in real-world scenarios was ensured for actual clinical settings, using Agresti-Coull Intervals[41] for the proportions and Wald Intervals[41] for the differences. Two-sided p-values[42] were calculated for every proportion and difference and a 2% absolute margin was selected for non-inferiority and superiority comparisons, using a statistical significance threshold of 0.01. The SciPy and NumPy libraries in python were used for all function commands.

ROLE OF THE FUNDING SOURCE

The funder participated in study design, data collection, data analysis, data interpretation, and manuscript writing. The funder managed the twitter account for the public disclosure of the data to which all authors and the public have equal and full access. All authors maintained control over the final content of this manuscript and had final responsibility for the decision to submit for publication.

5 DATA AVAILABILITY

The PROTOCOLS validation raw data are publicly available at twitter.com/demiurgetech. The PROTOCCOLS validation summary data are publicly downloadable as supplementary tables.

6 ACKNOWLEDGMENTS

TBA

7 AUTHOR CONTRIBUTIONS

B.L. and I.L. contributed to the study conception. B.L. and I.L. contributed to the study design. B.L. and I.L. contributed to data collection and interpretation. B.L. performed the statistical analysis. B.L. and I.L. contributed to the interpretation of the results and writing of the manuscript. All authors read and approved the final version of the paper.

8 COMPETING INTERESTS

This study was funded by Demiurge Technologies AG ('Demiurge'). B.L. and I.L. are employees, board members, and shareholders of Demiurge.

9 SUPPLEMENTAL INFORMATION

Detailed summary information of the PROTOCOLS validation data is available in the Excel spreadsheets.

References

REFERENCES

1. Silver, D. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).

2. Tunyasuvunakool, K. Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).

3. Nurse, P. Biology must generate ideas as well as data. *Nature* **597**, 305–305 (2021).

4. Scannell, J. W., Blanckley, A., Boldon, H. & Warrington, B. Diagnosing the decline in pharmaceutical R&D efficiency. *Nat. Rev. Drug Discov* **11**, 191–200 (2012).

5. Richards, B. A. A deep learning framework for neuroscience. *Nat. Neurosci* **22**, 1761–1770 (2019).

6. Bommasani, R. On the opportunities and risks of foundation models. Preprint at https://arxiv.org/abs/2108.07258 (2021).

7. Riley, R. D. Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat. Med* **40**, 4230–4251 (2021).

8. BIO, Informa, QLS. Clinical Development Success Rates and Contributing Factors 2011–2020. URL: https://www.bio.org/clinical-development-success-rates-and-contributing-factors-2011-2020. (2021) (Accessed: 28th February 2022)

9. High-Accuracy MoA-based Clinical Efficacy Prediction Validation. Demiurge Technologies AG (2022). URL: https://www.demiurge.technology/validation. (Accessed: 25th February 2022)

10. Lovetrue, B. The AI-discovered aetiology of COVID-19 and rationale of the irinotecan+ etoposide combination therapy for critically ill COVID-19 patients. *Med. Hypotheses* **144**, 110180–110180 (2020).

11. Browder, K. C. In vivo partial reprogramming alters age-associated molecular changes during physiological aging in mice. *Nat. Aging* (2022).

12. Wu, L., Wang, D. & Evans, J. A. Large teams develop and small teams disrupt science and technology. *Nature* **566**, 378–382 (2019).

13. Jiang, L. *et al.* Learning data-driven curriculum for very deep neural networks on corrupted labels. *International Conference on Machine Learning* 2304–2313 (2018).

14. Mullard, A. Cancer reproducibility project yields first results. *Nat. Rev. Drug Discov* **16**, 77–77 (2017).

15. Perrin, S. Preclinical research: Make mouse studies work. *Nature* **507**, 423–425 (2014).

16. Brown, T. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst* **33**, 1877–1901 (2020).

17. Kaplan, J. Scaling laws for neural language models. Preprint at https://arxiv.org/abs/2001.08361 (2020).

18. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)-Round XIV. *Proteins Struct. Funct. Bioinforma* **89**, 1607–1617 (2021).

19. Hay, M., Thomas, D. W., Craighead, J. L., Economides, C. & Rosenthal, J. Clinical development success rates for investigational drugs. *Nat. Biotechnol* **32**, 40–51 (2014).

20. Pritchard, J. F. Making Better Drugs: Decision Gates in Non-Clinical Drug Development. *Nat. Rev. Drug Discov* **2**, 542–553 (2003).

21. Seyhan, A. A. Lost in translation: the valley of death across preclinical and clinical divide - identification of problems and overcoming obstacles. *Transl. Med. Commun* **4**, 18–18 (2019).

22. Hingorani, A. D. Improving the odds of drug development success through human genomics: modelling study. *Sci. Rep* **9**, 18911–18911 (2019).

23. Wu, E. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat. Med* **27**, 582–584 (2021).

24. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. AI in health and medicine. *Nat. Med* **28**, 31–38 (2022).

25. Chicco, D., Tötsch, N. & Jurman, G. The matthews correlation coefficient (Mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Min* **14**, 1–22 (2021).

26. Liu, R. Evaluating eligibility criteria of oncology trials using real-world data and AI. *Nature*

592, 629–633 (2021).

27. Lovetrue, B. The AI-Discovered Aetiology of COVID-19 and Rationale of the Irinotecan+Etoposide Combination Therapy for Critically Ill COVID-19 Patients. Preprint at https://www.preprints.org/manuscript/202003.0341/v1 (2020).

28. E, T, G., T, Douglas, D. S. & G. Alzheimer's disease: The right drug, the right time. *Science* **362**, 1250–1251 (2018).

29. Fei, Z., Raj, K., Horvath, S. & Lu, A. *Universal DNA Methylation Age Across Mammalian Tissues. Innov. Aging* **5**, 412–412 (2021).

30. Niccoli, T. & Partridge, L. Ageing as a Risk Factor for Disease. *Curr. Biol* **22**, 741–752 (2012).

31. Brynjolfsson, E., Rock, D. & Syverson, C. The productivity J-curve: How intangibles complement general purpose technologies. *Am. Econ. J. Macroecon* **13**, 333–372 (2021).

32. Woo, M. An AI boost for clinical trials. *Nature* **573**, 100–100 (2019).

33. Medicine, N. L., Of, Clinicaltrials & Gov. *National Library of Medicine (US) Available* 27–27 (2022).

34. Russakovsky, O. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis* **115**, 211–252 (2015).

35. Mcleod, C. Choosing primary endpoints for clinical trials of health care interventions. *Contemp. Clin. Trials Commun* **16**, 100486–100486 (2019).

36. U.S.Food and Drug Administration, (2005). URL http//www.fda.gov/RegulatoryInformation/ Guidances/ucm127069.htm.

37. Burt, T., Button, K. S., Thom, H. H. Z., Noveck, R. J. & Munafò, M. R. The Burden of the "False-Negatives" in Clinical Development: Analyses of Current and Alternative Scenarios and Corrective Measures. *Clin. Transl. Sci* **10**, 470–479(2017).

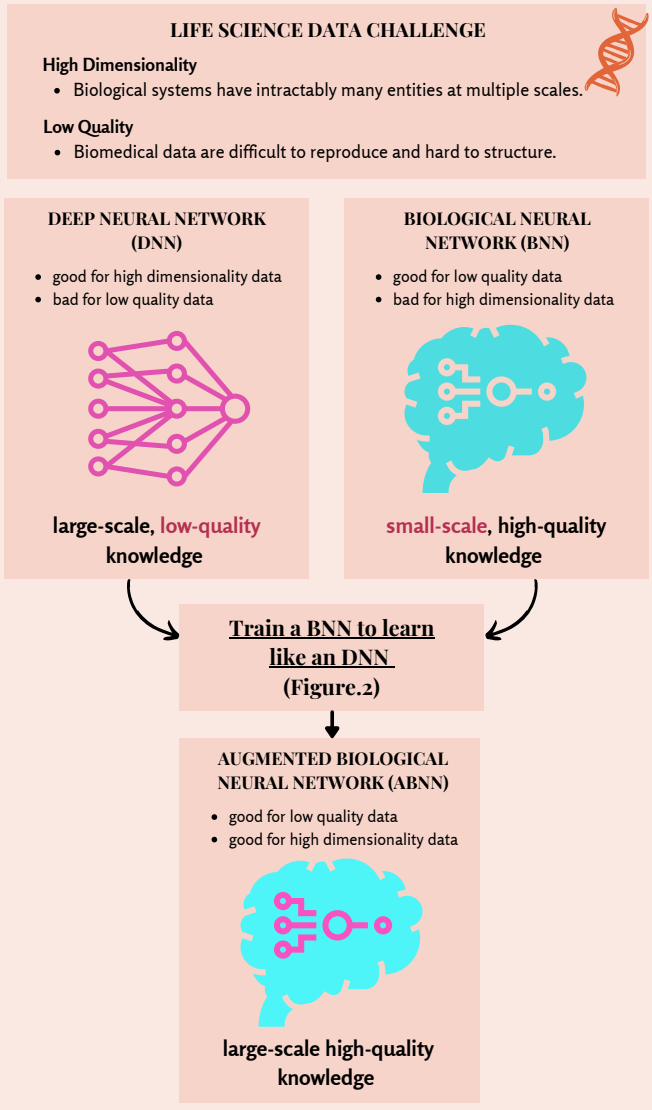
38. Devito, N. J., Bacon, S. & Goldacre, B. Compliance with legal requirement to report clinical trial results on ClinicalTrials. gov: a cohort study. *Lancet* **395**, 361–369(2020).

39. Williams, R. J., Tse, T., Dipiazza, K. & Zarin, D. A. Terminated trials in the ClinicalTrials. gov results database: evaluation of availability of primary outcome data and reasons for termination. *PLoS One* **10**, 127242–127242(2015).

40. Fagerland, M. W., Lydersen, S. & Laake, P. Recommended tests and confidence intervals for paired binomial proportions. *Stat. Med* **33**, 2850–2875(2014).

41. Altman, D. G. & Bland, J. M. How to obtain the P value from a confidence interval. *BMJ* **343**, 2304–2304 (2011).

Figure 1. The challenge and solution of extracting knowledge from life science data



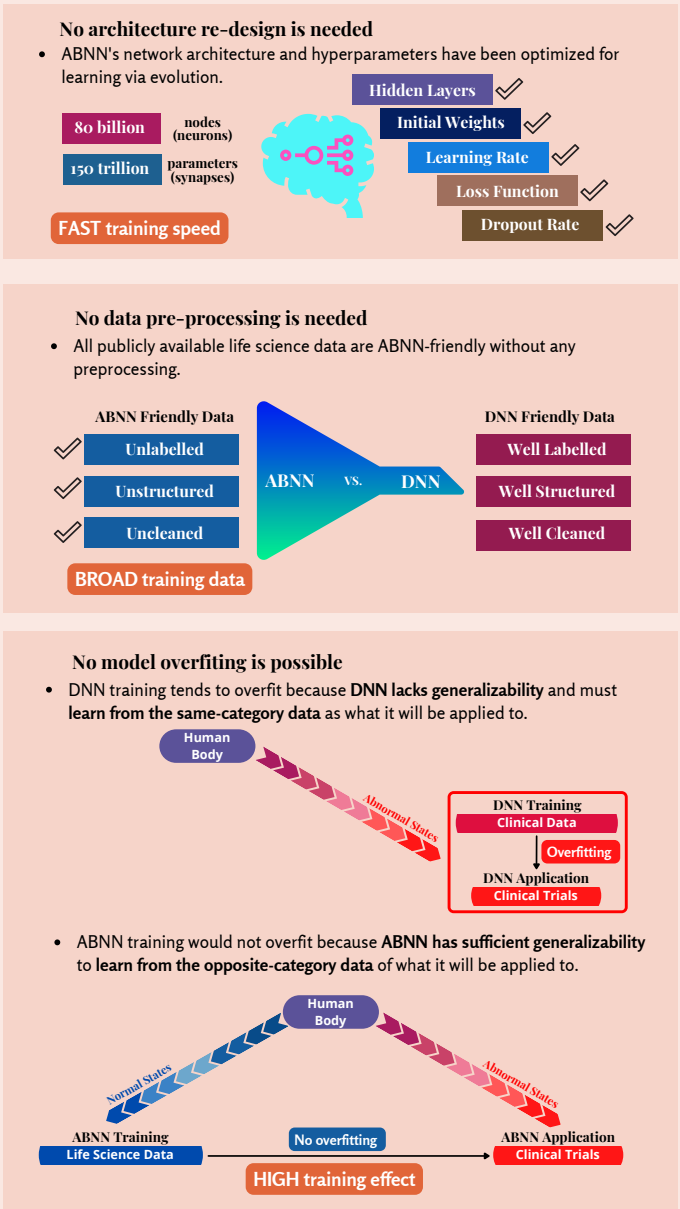


Figure 1a. The challenge and solution of extracting knowledge from life science data. High dimensionality and low quality are two defining characteristics of life science data. Deep neural networks (DNNs) are well suited for learning knowledge from the high-dimensional, high-quality data spaces, whereas biological neural networks (BNNs) are good at learning knowledge from the low-dimensional, ow-quality data spaces. The proposed augmented biological neural network (ABNN) combines the complementary advantages of DNNs and BNNs for knowledge acquisition from high-dimensional, low-quality life science data.

Figure 1b. The training advantages of ABNNs compared with DNNs. ABNN training is far more efficient than that of DNN because no time or cost is required for designing the network architecture, tuning network hyperparameters, or cleaning the training data. Furthermore, the application of an ABNN to clinical trials is also more effective than a DNN because it offers sufficient generalizability for learning from large-scale life science data, while DNNs suffer from limited generalizability for small-scale clinical data and are prone to overfitting.

Figure 2. Train a BNN to learn like an DNN

a Learn from Multiple Levels of Abstraction

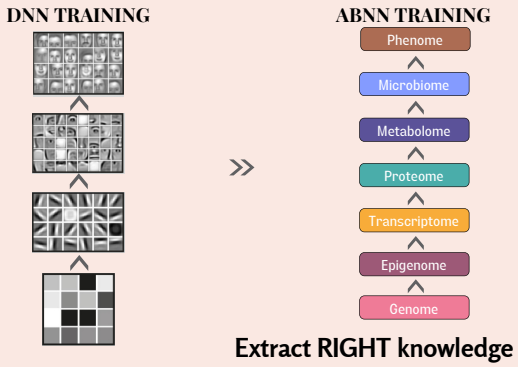


Figure 2a Left. DNNs learn better representations of data with more levels of abstraction than with fewer levels and share a hierarchical architecture with BNNs. **Figure 2a Right.** ABNNs extract right knowledge by integrating more multiomics data (e.g., genome, proteome, transcriptome, epigenome, metabolome and microbiome).

b Learn via the Backpropagation Algorithm

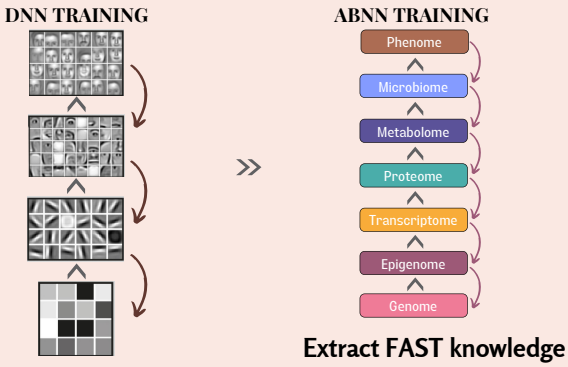


Figure 2b Left. DNNs use backpropagation to improve learned representations of data by sequentially updating its internal parameters from the highest to the lowest levels of abstraction. The ubiquitous feedback connections in BNNs may be used to implement backpropagation-like learning rules. **Figure 2b Right.** ABNNs extract fast knowledge by iteratively updating internal representations in a strict order from the most macroscopic level (phenotype), through the intermediate level (endophenotype), to the most microscopic level (genotype).

c Learn for a Foundation Model

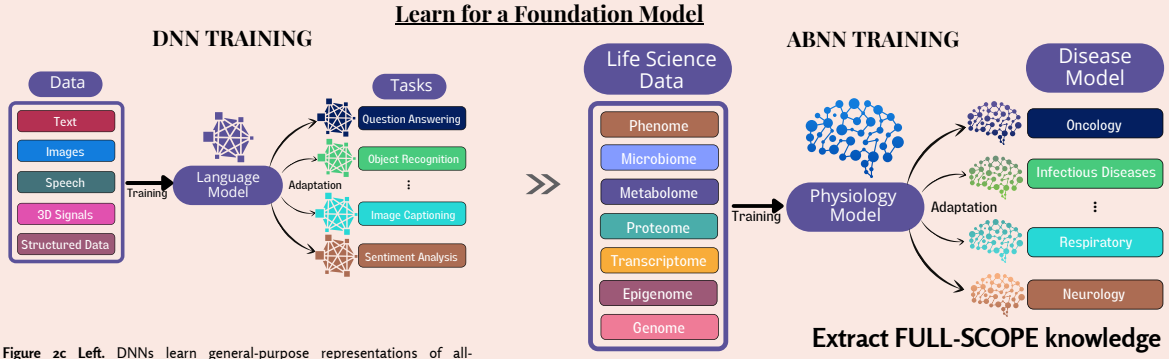


Figure 2c Left. DNNs learn general-purpose representations of all-encompassing data to deliver optimal performance for a wide range of special-purpose tasks. Both BNNs and DNNs could learn to command general linguistic capabilities for multiple tasks.

Figure 2c Right. ABNNs extract full-scope knowledge by first learning a general-purpose model of human physiology from the all-encompassing data, prior to learning human diseases models from disease-specific data.

d Learn in the Biggest Network

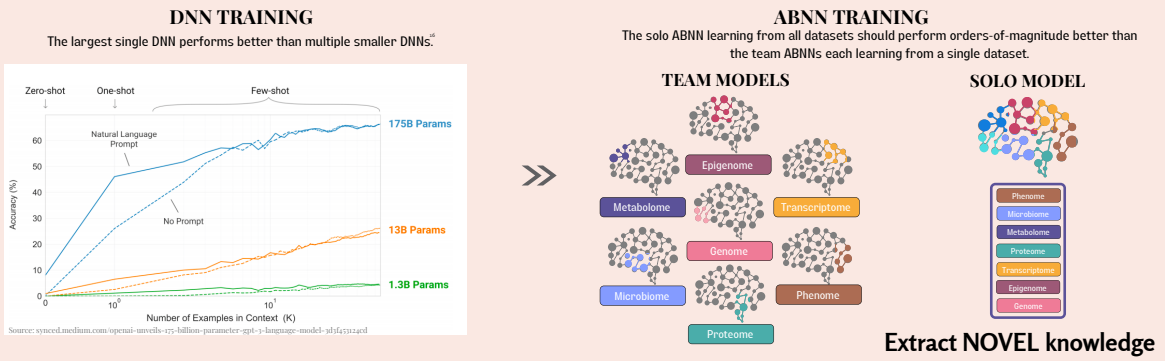


Figure 2d Left. DNNs maximize the robustness of learned representations of data by increasing the sheer size of network parameters.

Figure 2d Right. ABNNs extract novel knowledge if a single ABNN of the largest size learns from the entirety of the all-encompassing data, rather than multiple ABNNs of smaller sizes learn from siloed sub-datasets.

e Learn with an Uninterrupted Pass

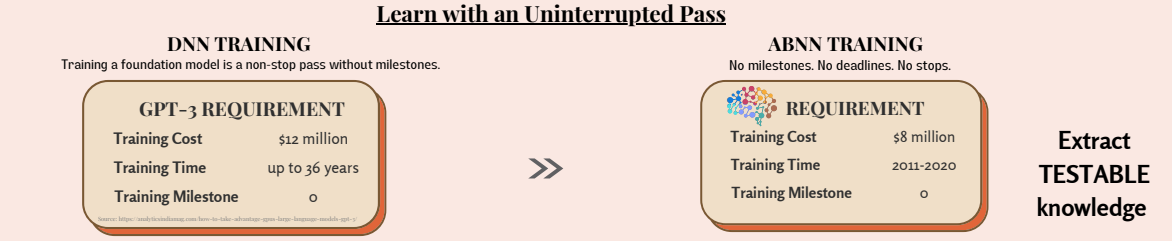


Figure 2e Left. DNNs learn the best representations of data when provided with sufficient time and ample resources to allow for an uninterrupted pass of a full training set, prior to which no meaningful performance milestones can be defined.

Figure 2e Right. ABNNs extract testable knowledge if a milestone-free resources are provided for an uninterrupted learning pass, ensuring the ABNN won't be evaluated prior to the completion of learning both human physiology and human diseases.

Figure 3. A real-world, large-scale, public, prospective, external validation of the trained ABNN

a Real-world ABNN validation and Real-world Industrial Practice in Prospective Predictions of Clinically Successful Drug Candidates Share Identical Process and Measure

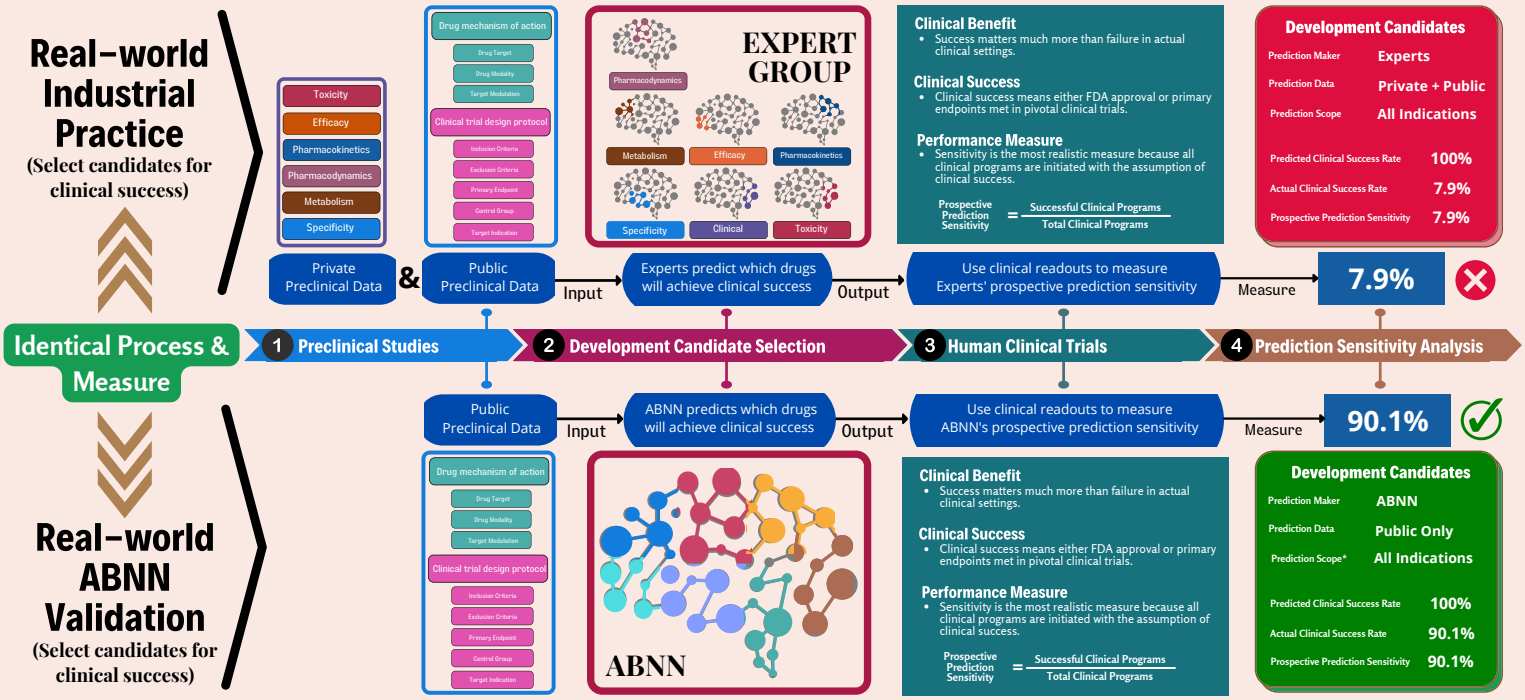


Figure 3a. Real-world validation is critical to the objective assessment of ABNN's potential for improving the efficiency of global drug research and development. As the foundation model is intended to functionally reconstruct nearly all major diseases that could occur to a single human patient, clinical success prediction is the most effective validation method for the ABNN. The clinical efficacy and safety of novel drugs were evaluated solely using mechanisms of action, to prevent drug developers from initiating pivotal real-world trials that are doomed to failure. Several key go/no-go decision steps are involved in drug development, as illustrated in Steps 1-4 in the figure. In Step 1, preclinical studies aim to determine which drug candidates are developable and generate sufficient data indicative of efficiency and safety in humans for each developable candidate. Biopharmaceutical companies have a considerable data advantage over the ABNN in this step as they have exclusive access to internally generated private data in addition to public data. In contrast, the ABNN only has access to public data of drug mechanism of action (i.e., drug targets, drug modality, and target modulation) and clinical trial design protocol (i.e., inclusion criteria, exclusion criteria, primary endpoint, control groups, and target indication).

Step 2 is a critical go/no-go decision step in drug development because only a subset of developable drug candidates (with highest clinical success potential) can be selected to transition from non-clinical to clinical development, due to the prohibitive costs of human clinical trials. Best practices in the biopharmaceutical industry involve a group of committed experts using private and public preclinical data to predict which drug candidates will achieve clinical success (Step 2). Clinical trials are then initiated only for those drug candidates that are prospectively predicted to be clinically successful (Step 3). However, the resulting prospective prediction sensitivity is only 7.9% (i.e., for every 100 clinical programs that experts prospectively predict to achieve clinical success, only 7.9 are actual successful) (Step 4). The real-world PROTOCOLS validation set was designed to mirror real-world industrial practices (Steps 2-4), to ensure the transferability of ABNN performance from validation to application. The ABNN solely uses public preclinical data only to predict which drug candidates will achieve clinical success (Step 2). Clinical readouts were tracked and recorded for those drug candidates that are prospectively predicted to be clinically successful (Step 3). The resulting prospective prediction sensitivity was 90.1% (i.e., for every 100 clinical programs that were prospectively predicted to achieve clinical success by the ABNN, 90.1 were actually successful) (Step 4). *These indications cover all diseases except for hematological disorders, rare neurodevelopmental disorders, and rare musculoskeletal disorders.

Figure 3. A real-world, large-scale, public, prospective, external validation of the trained ABNN

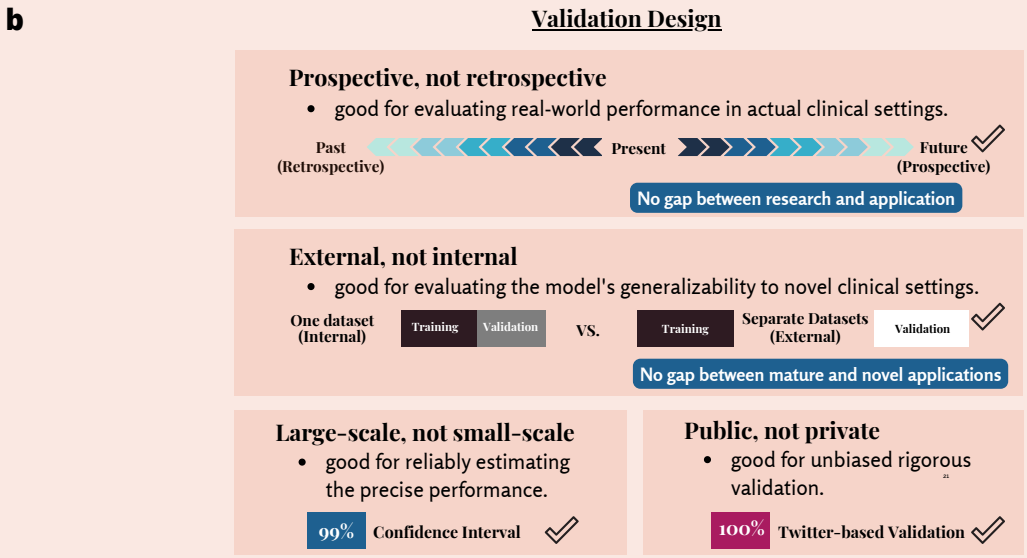


Figure 3b. Prospective validation is far more rigorous than retrospective validation for examining the real-world performance of AI-based systems in actual clinical practice. Large-scale validation is also more rigorous than small-scale validation for enabling an unbiased estimation of the actual predictive power of AI-based systems. External validation using separate datasets (independent of the training datasets) is more rigorous than internal validation using held-out portions of the training data, in evaluating the generalizability of AI-based systems for targeted clinical applications. Public validation is more rigorous than private validation for ruling out all the possibility of biasing performance estimates.

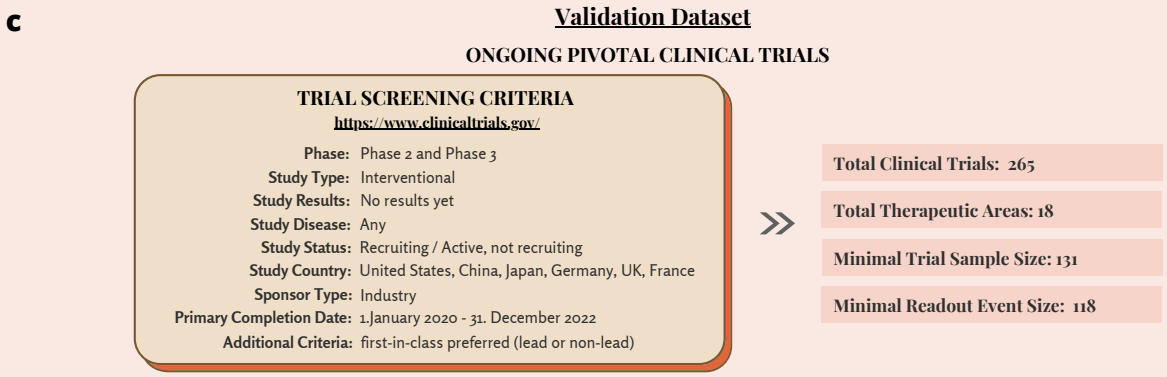


Figure 3c. The list of screening criteria used to identify pivotal clinical trials for ABNN validation. The standalone performance of the foundation model was evaluated using the public prospective predictions of pivotal ongoing clinical trial success outcomes at large scale (PROTOCOLS), the most rigorous external validation to date of a clinical prediction model with a binary outcome. A total of 265 prospective predictions were published to satisfy the minimum sample size requirement of 131 and evaluate prospective prediction sensitivity, F1 score, accuracy, and specificity at 99% confidence, Similarly, 158 readouts were accumulated to meet the minimum event size requirement of 118 (see Methods section 'Statistical analysis').

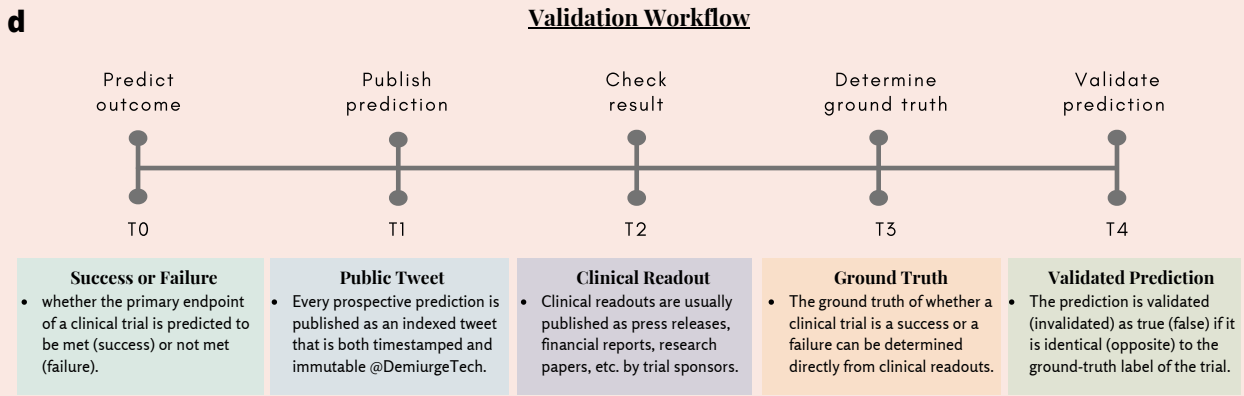


Figure 3d. The workflow for every validation step from beginning to end (see Methods section for details). Twitter was used to meet the stringent criteria of the PROTOCOLS validation, providing immutable timestamped tweets to construct a publicly verifiable track record on the publicly accessible account of @DemiurgeTech. The ABNN was used to made prospective predictions of pivotal clinical trial outcomes for all human diseases, excluding rare neurodevelopmental , rare skeletomuscular, and hematological disorders. There was no selection of lead over non-lead indications, though non-lead indications have far lower success rates. There are a few hundred new pivotal clinical trials whose protocols are publicly accessible in a centralized database, as required by FDA, and whose clinical readouts are typically published by sponsors in a standardized format. We calculated prospective prediction sensitivity with a 99% confidence interval using a ground truth determined by actual clinical readouts.

Figure 4. Proportions of Clinical Trials Per Key Therapeutic Area

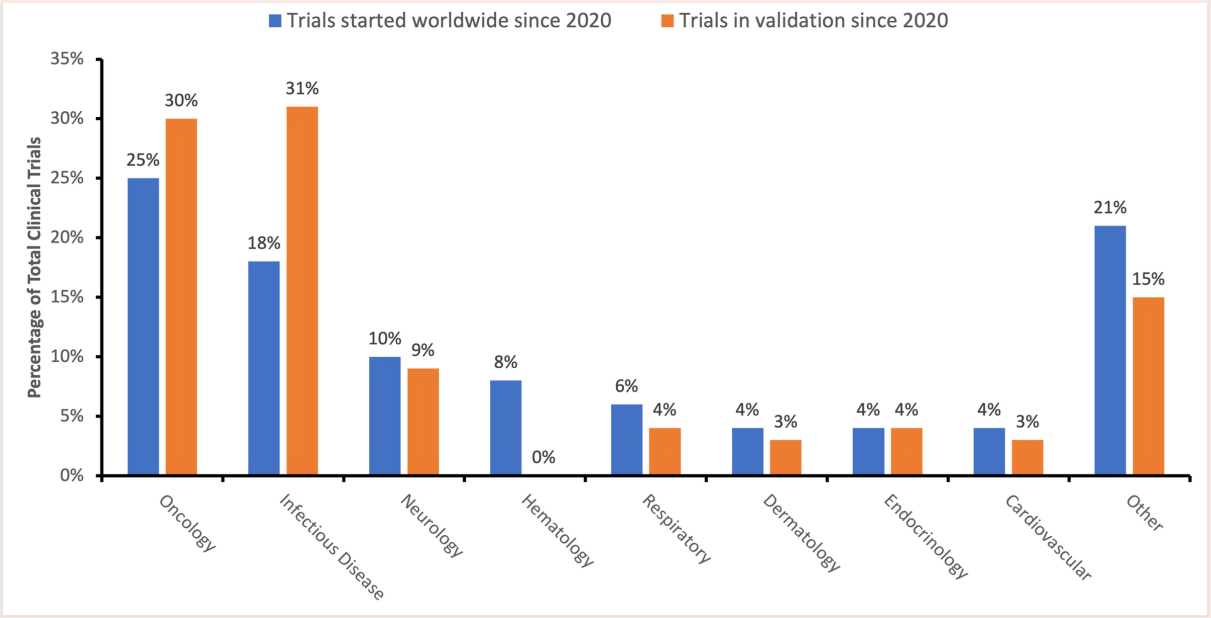


Figure 4. A two-sample Kolmogorov-Smirnov test was used as the standard method to evaluate the direct translatability between prospective validation and clinical applications. Validation clinical trials and all clinical trials initiated since 2020 followed identical distributions (statistic = 0.333 p-value = 0.73).

Figure 5a. Prospective Prediction Sensitivity of Clinical Trials for All Therapeutic Areas

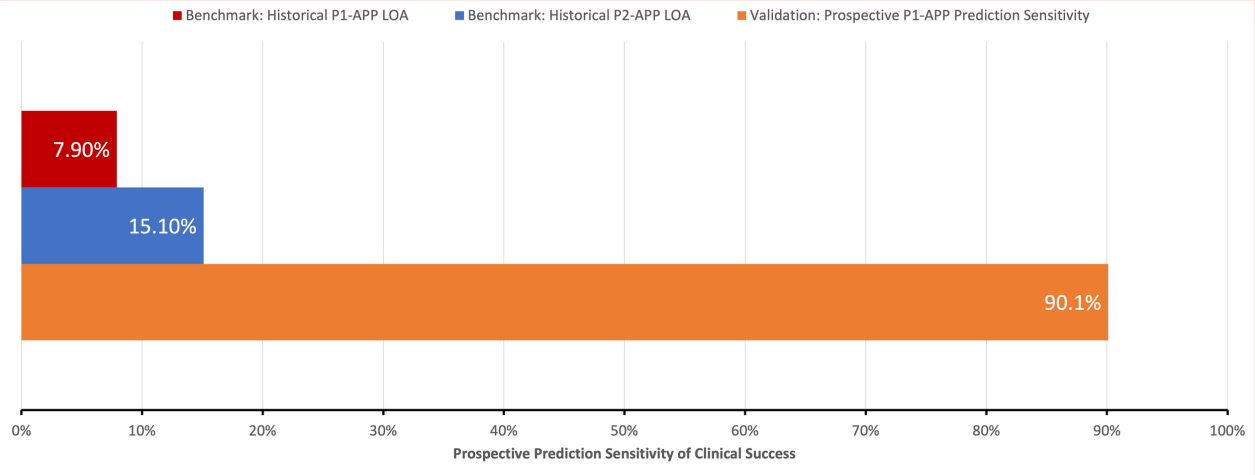


Figure 5a. The PROTOCOLS validation involved a realistic baseline sensitivity of 7.90% (Historical P1-APP LOA: historical likelihood of success from phase 1 to approval) and a conservative baseline sensitivity of 15.1% (Historical P2-APP LOA: historical likelihood of success from phase 2 to approval) to benchmark ABNN performance (see Methods section 'Baseline performance' for details). The overall prospective prediction sensitivity for pivotal clinical trials in the PROTOCOLS validation was 90.1% (99% CI 80.0%, 96.9%; P < 0.001) .

Figure 5b. Prospective Prediction Sensitivity of Clinical Trials for All Trials and First-in-class Trials

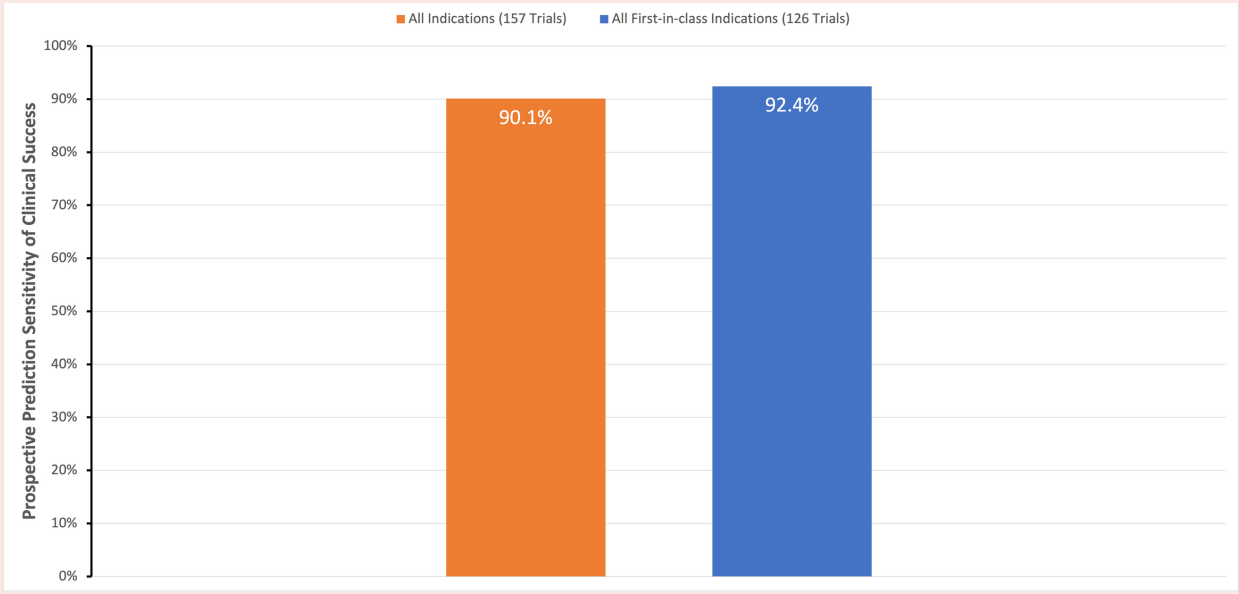


Figure 5b. Prospective prediction accuracy for first-in-class pivotal clinical trials was superior to that of all pivotal clinical trials (Δ = 2.3%; 99% CI -0.92%, 5.52%; P < 0.002; 2% margin for non-inferiority) for all therapeutic areas, excluding rare neurodevelopmental, rare skeletomuscular, and hematological disorders (see Methods section 'Model training' for details).

Figure 5c. Prospective Prediction Sensitivity of Clinical Trials Per Therapeutic Area

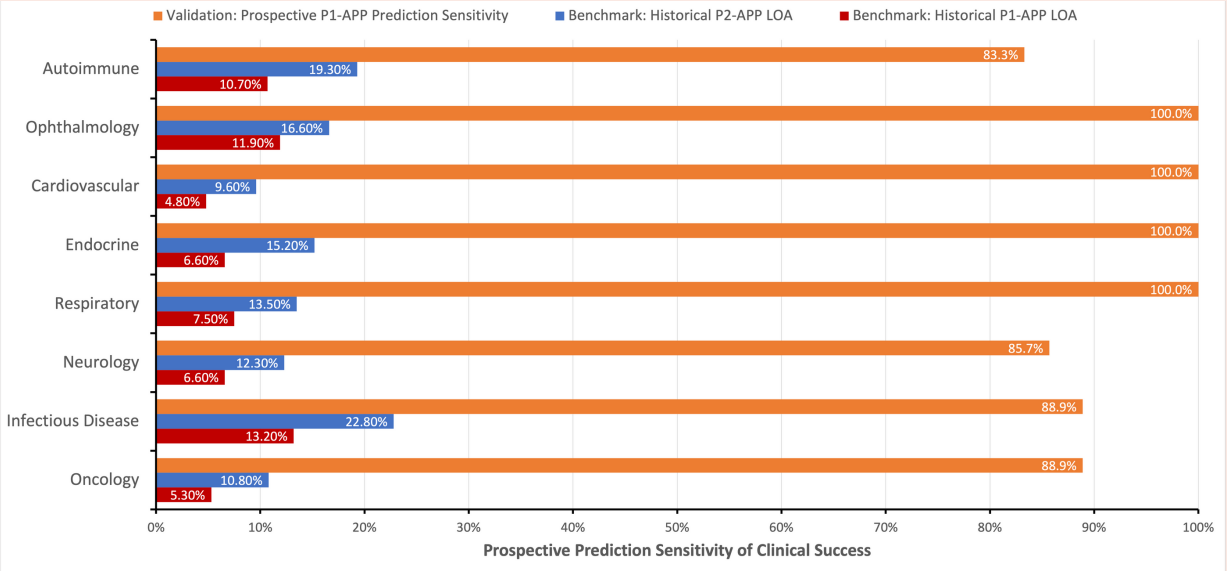


Figure 5c. The PROTOCOLS validation utilized a realistic baseline sensitivity (Historical P1-APP LOA: historical likelihood of success from phase 1 to approval) and a conservative baseline sensitivity (Historical P2-APP LOA: historical likelihood of success from phase 2 to approval) to benchmark ABNN performance (see Methods section 'Baseline performance' for details). Results were as follows: Oncology: 77.1% (90% CI 65.1%, 84.9%; P < 0.0004); Infectious Disease: 81.6% (90% CI 70.1%, 88.4%; P < 0.002); *The minimum sample and event size requirements for a 90% confidence have been met (see Methods section 'Statistical analysis'). **The minimum sample and event size requirements for 90% confidence have not been met (see Methods section 'Statistical analysis').

Figure 6a. Performance of ABNN: Precision and Recall

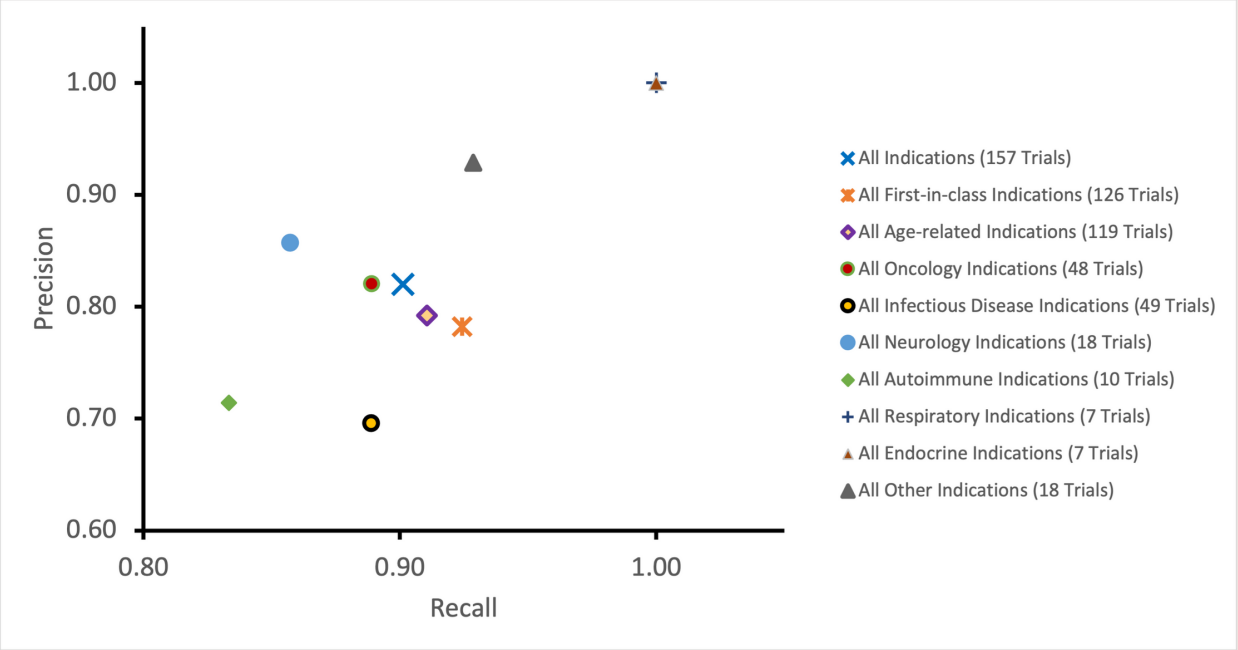


Figure 6a. ABNN performance as measured by precision and recall (see Data Table 1 for more details).

Figure 6b. Performance of ABNN: Sensitivity and 1 - Specificity

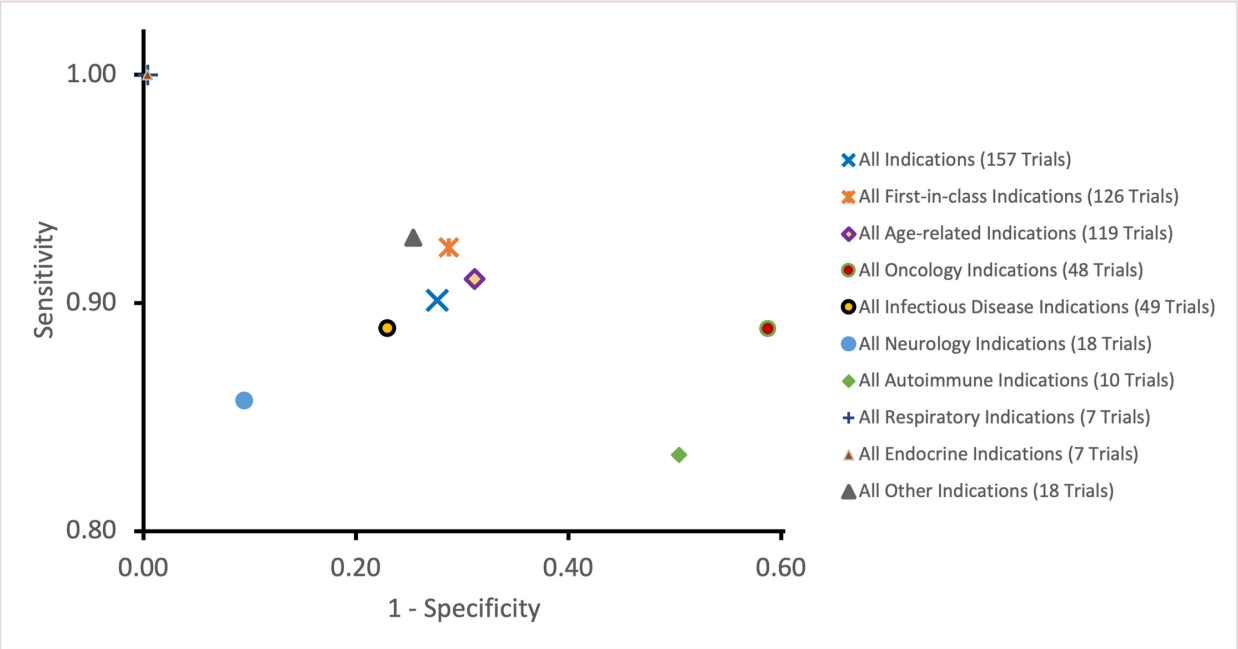


Figure 6b. ABNN performance as measured by sensitivity and 1 - specificity (see Data Table 1 for more details).

Figure 6c. Performance of ABNN: Accuracy and F1 Score

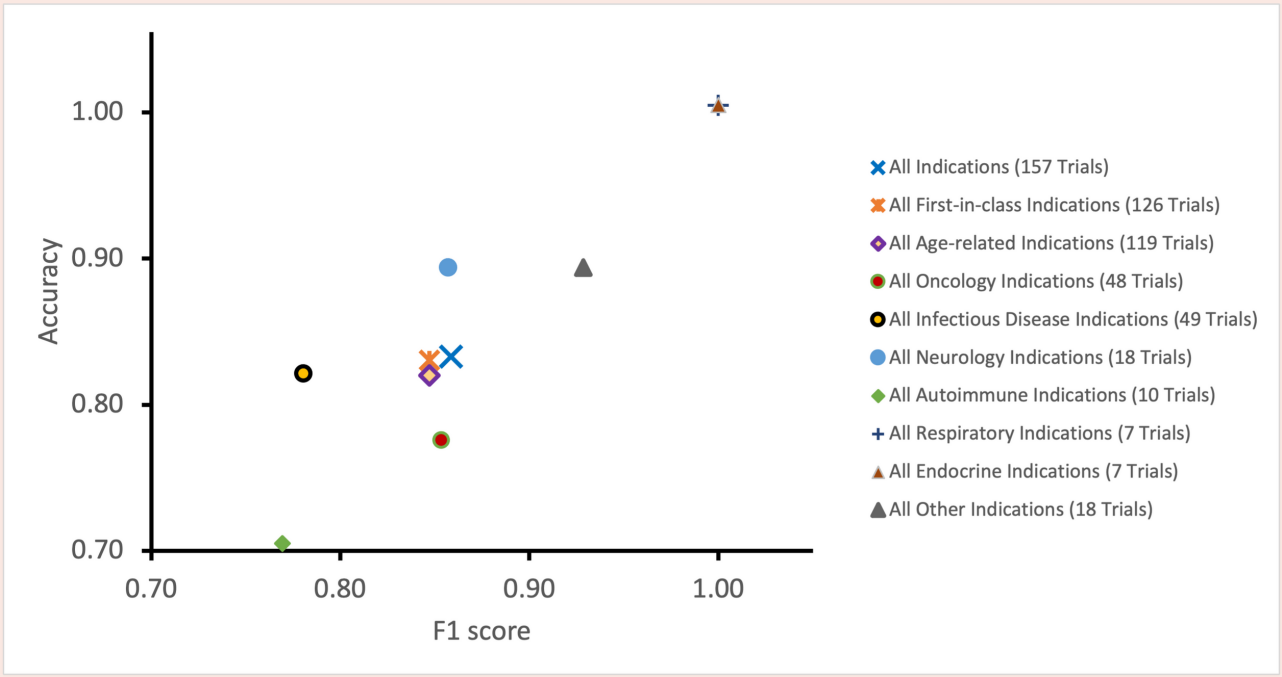


Figure 6c. ABNN performance as measured by accuracy and F1 score (see Data Table 1 for more details).

Figure 6d. Performance of ABNN: LOA and F1 Score

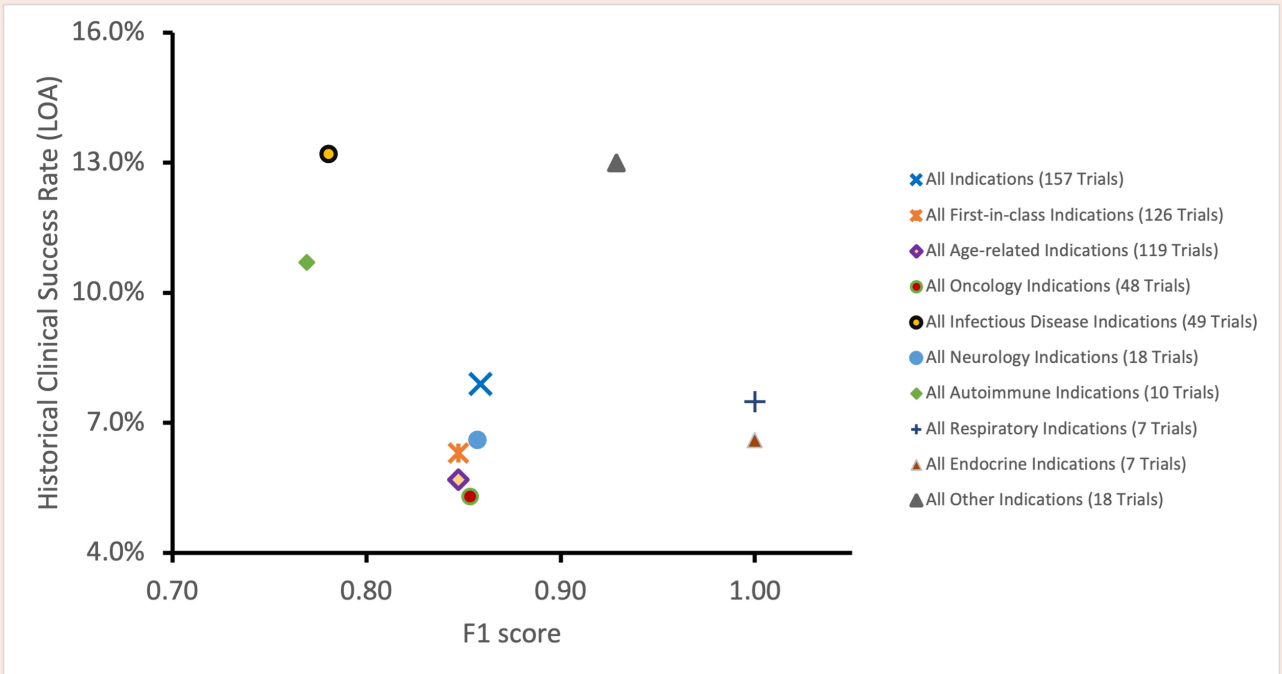


Figure 6d. ABNN performance as measured by historical clinical success rates for investigation new drugs (LOA) and F1 scores (see Data Table 1 for details). LOA is identical to the realistic baseline accuracy (see Methods section 'Baseline performance'). F1 scores and LOAs across overall validation subgroups were negatively and weakly correlated (exact Pearson coefficient = -0.260)

Figure 7a. Performance of ABNN: Confusion Matrices for All Therapeutic Areas

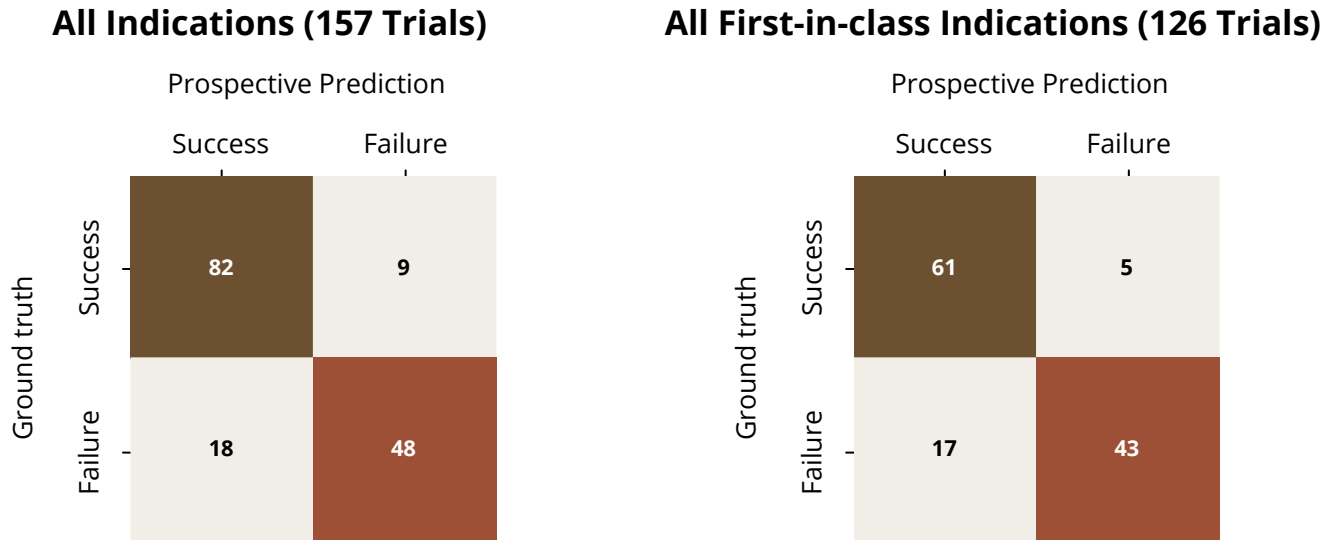


Figure 7a. ABNN performance as recorded by confusion matrices for all therapeutic areas.

Figure 7b. Performance of ABNN: Confusion Matrices Per Therapeutic Area

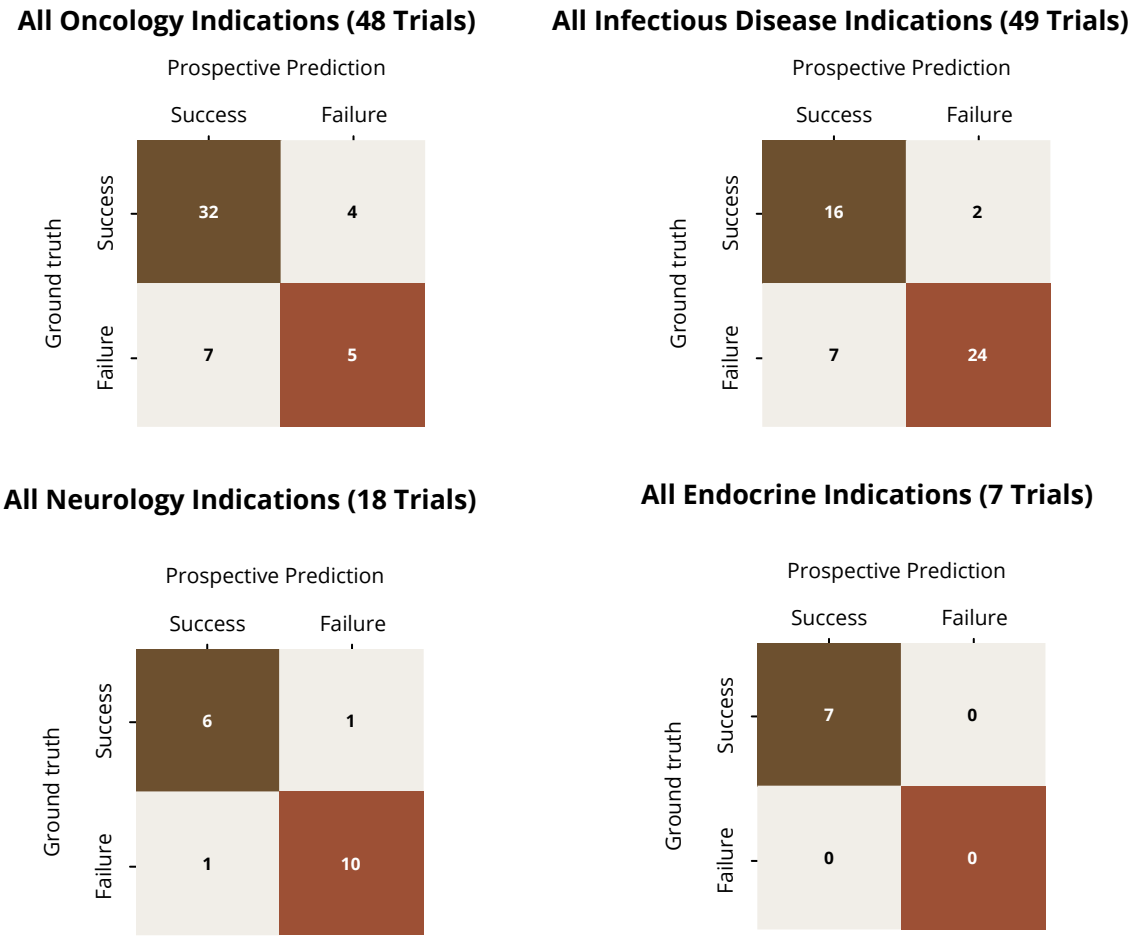


Figure 7b. ABNN performance as recorded by confusion matrices per therapeutic area.

Figure 7c. Performance of ABNN: Confusion Matrices Per Therapeutic Area

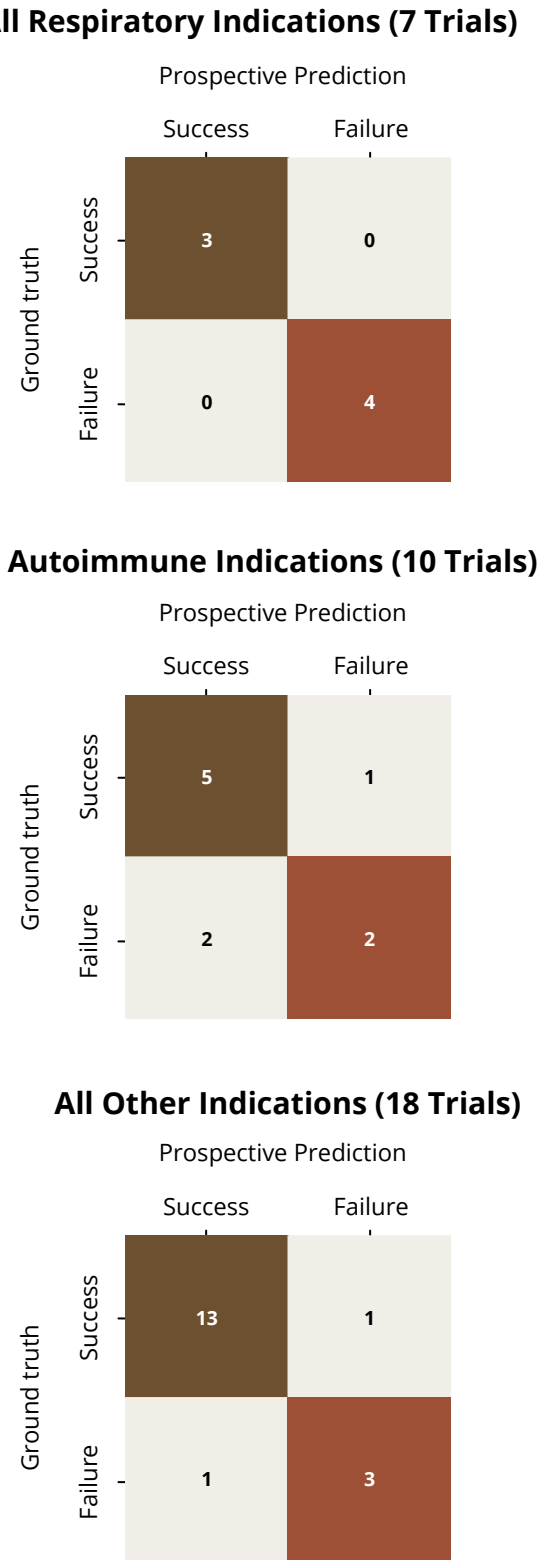


Figure 7c. ABNN performance as recorded by confusion matrices per therapeutic area.

Figure 8. Performance of ABNN: Confusion Matrices for Age-Related Diseases

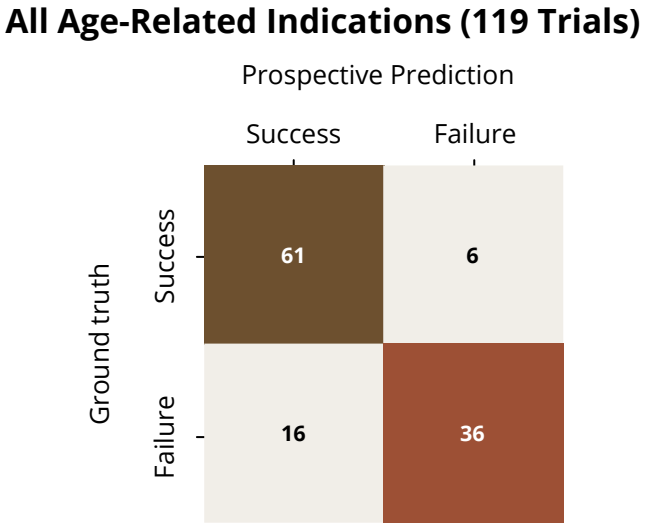


Figure 8. ABNN performance as recorded by confusion matrices for all age-related diseases.

Figure 9a. Prospective Prediction Sensitivity for All Age-related Indications

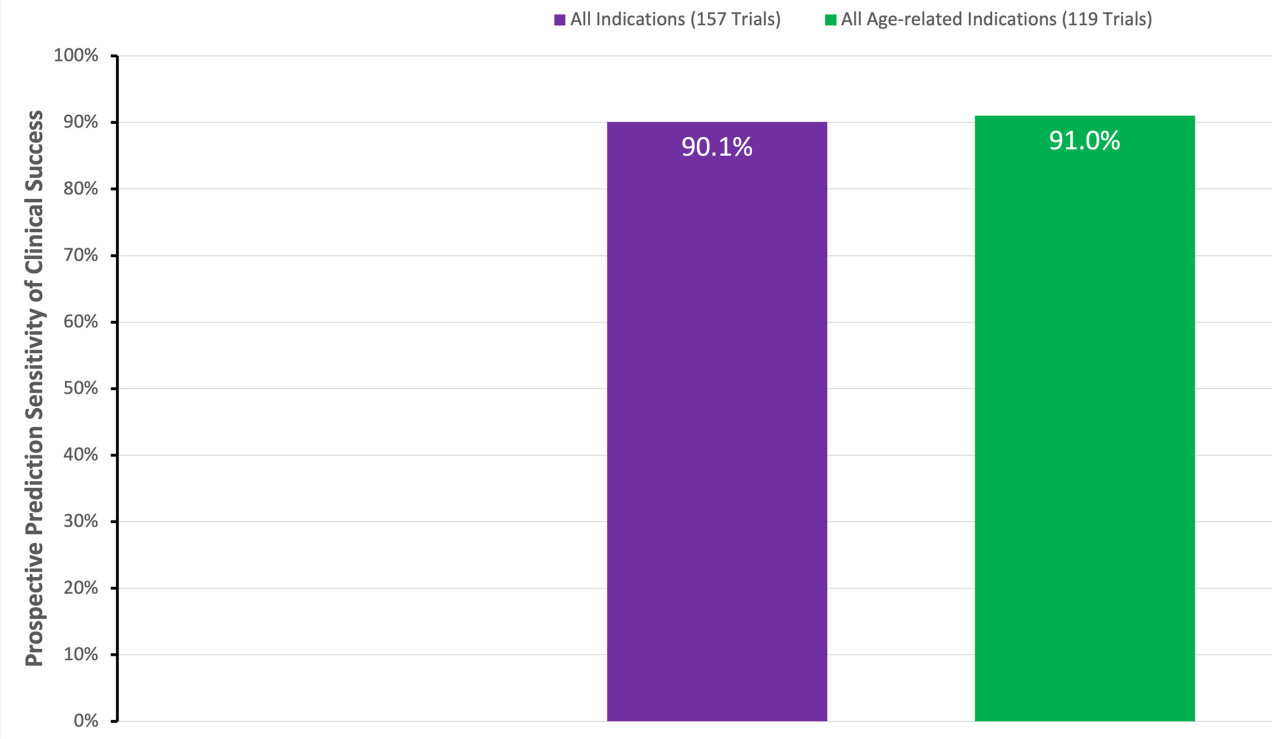


Figure 9a. Prospective prediction sensitivity of all age-related pivotal clinical trials was 91.0% (99% CI 79.1%, 98.4%; P < 0.005) in the PROTOCOLS validation set (Methods section 'Statistical analysis').

Figure 9b. Distribution of Age-related Indications per Therapeutic Area

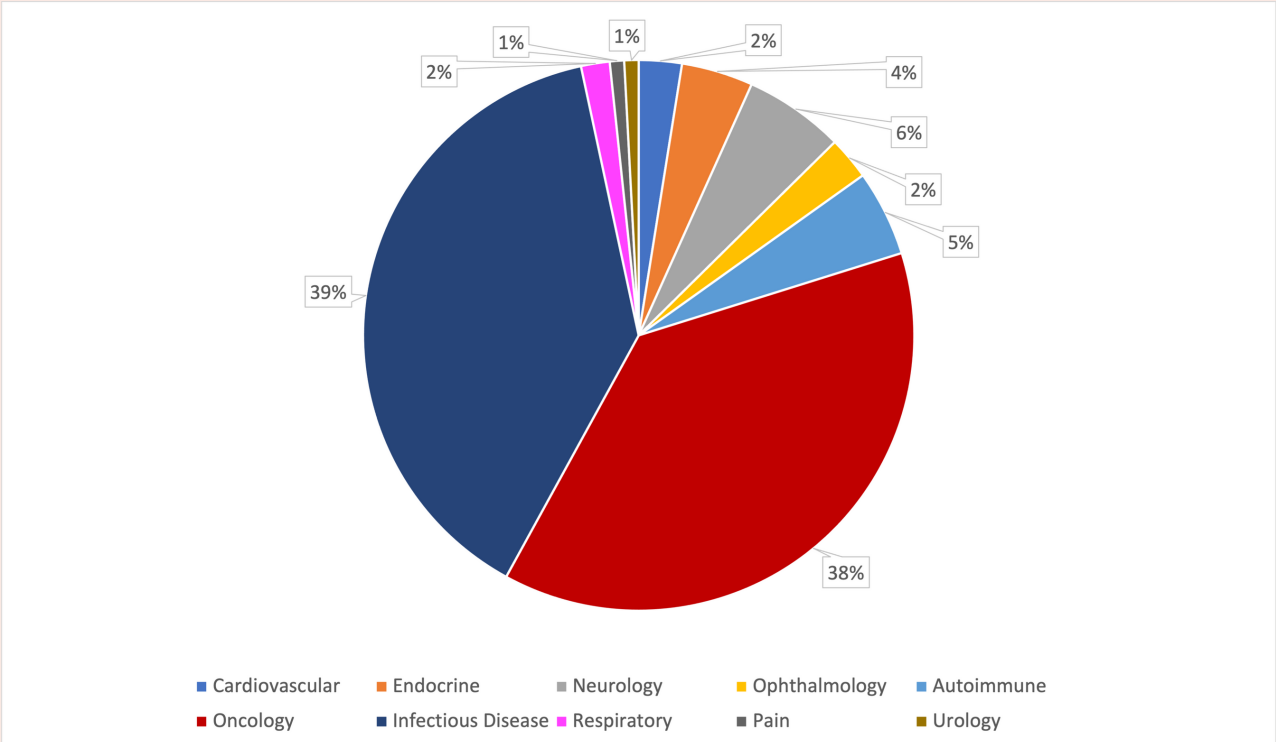


Figure 9b. The distribution of all age-related pivotal clinical trials per therapeutic area in the PROTOCOLS validation set(Methods section 'Validation data collection').

Figure 10. Performance of ABNN: Confusion Matrices Per Drug Modality

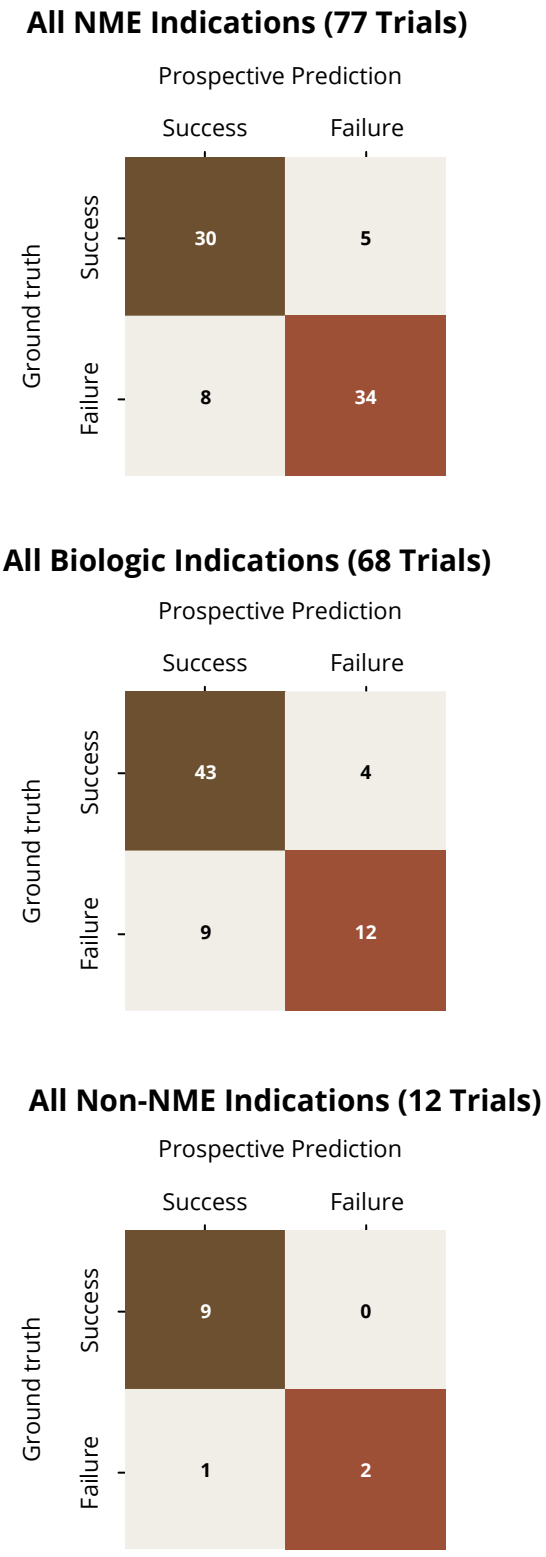


Figure 7d. ABNN performance as recorded by confusion matrices per drug modality.

Figure 11a. Performance of ABNN Per Drug Modality: Prospective Prediction Sensitivity

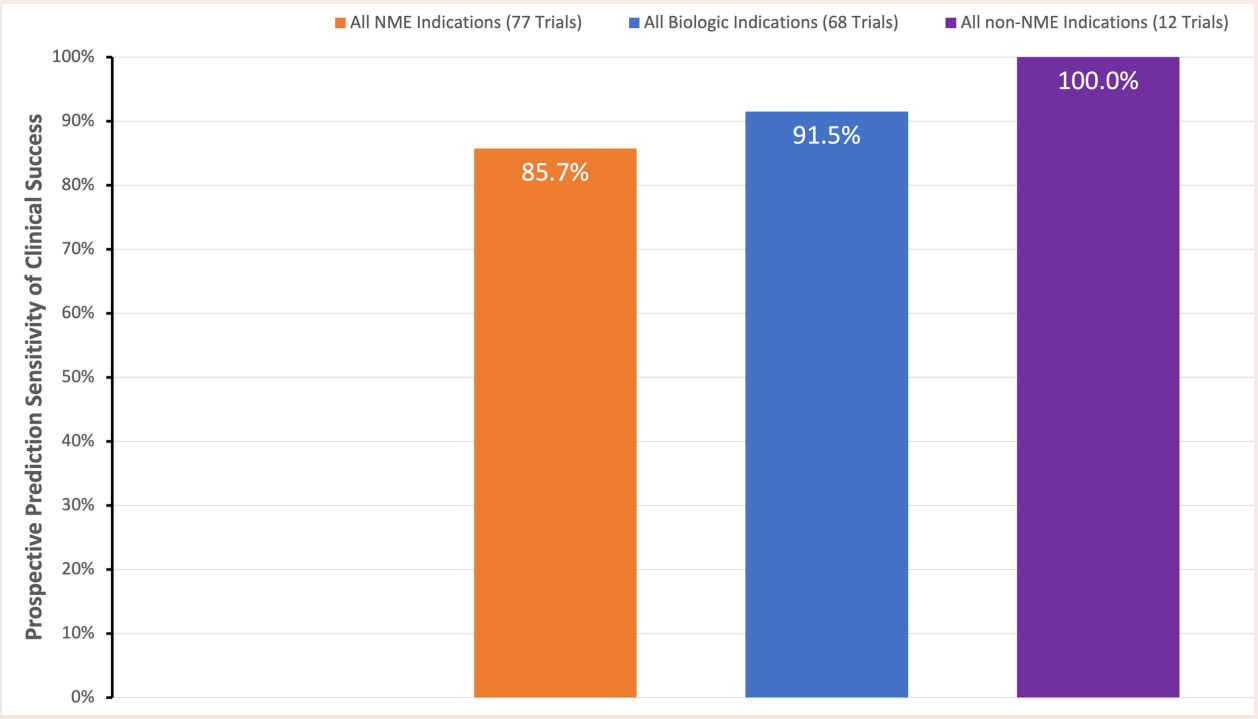


Figure 11a. ABNN performance as measured by prospective prediction sensitivity across drug modalities (see Data Table 1 for more details). ABNN achieved superior performance for biologics than for small molecules ((Δ = 5.8%; 90% CI 0.78%, 10.82%; P < 0.0001 for superiority at a 2% margin; Methods section 'Statistical analysis').

Figure 11b. Performance of ABNN Per Drug Modality: LOA and F1 Score

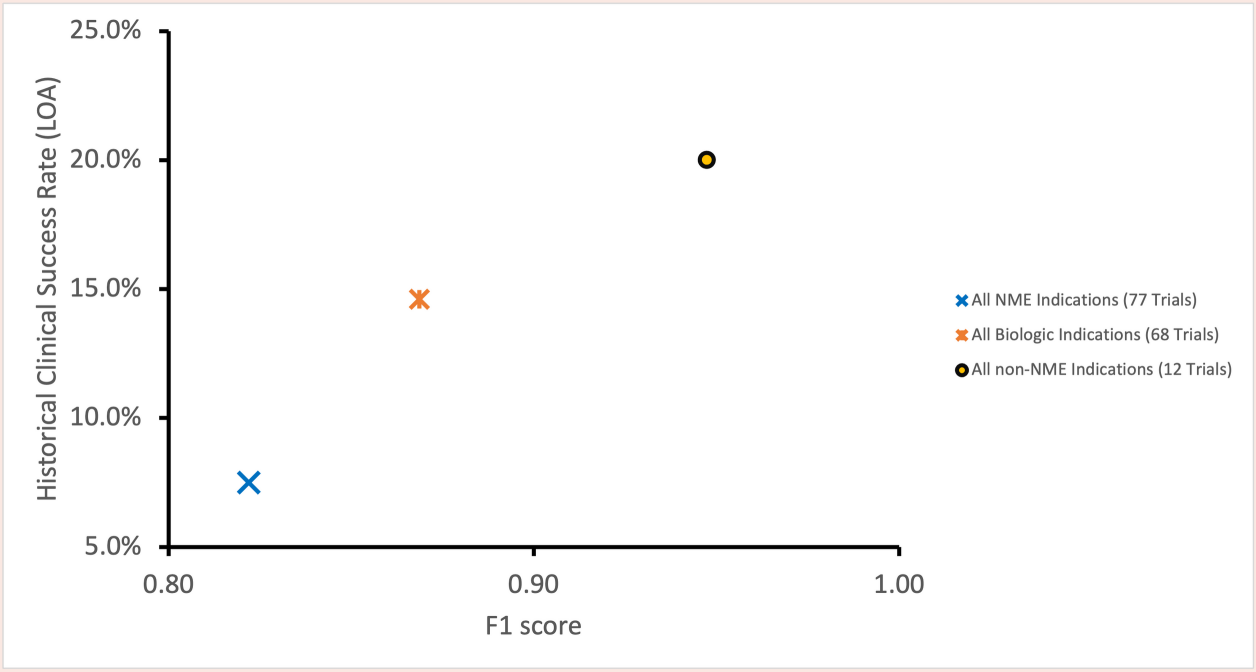


Figure 11b. ABNN performance as measured by historical clinical success rates for investigation new drugs (LOA) and F1 scores (see Data Table 1 for details). LOA is identical to the realistic baseline accuracy (see Methods section 'Baseline performance'). F1 scores and LOAs across overall validation subgroups are strongly positively correlated (exact Pearson coefficient = 0.975)

Data Table 1. ABNN's Performance Measures with Statistical Significance

Validation Subgroups	N	Metric	Performance	Significance	CI (%)	p-value
All Indications	157	Sensitivity	90.1%	99%	(80.0, 96.9)	0.001
All First-in-class Indications	126	Sensitivity	92.4%	99%	(80.8, 99.2)	0.009
All Age-related Indications	119	Sensitivity	91.0%	99%	(79.1, 98.4)	0.005
All Infectious Disease Indications	49	Sensitivity	88.9%	90%	(68.3, 95.4)	0.050
All Oncology Indications	48	Sensitivity	88.9%	90%	(75.7, 94.3)	0.015
All NME Indications	77	Sensitivity	85.7%	90%	(71.9, 92.2)	0.009
All Biologics Indications	68	Sensitivity	91.5%	90%	(80.8, 95.7)	0.015

Validation Subgroups	N	Metric	Performance	Significance	CI (%)	p-value
All Indications	157	F1-Score	0.86	99%	(0.79, 0.92)	0.0001
All First-in-class Indications	126	F1-Score	0.85	99%	(0.76, 0.92)	0.0001
All Age-related Indications	119	F1-Score	0.85	99%	(0.76, 0.92)	0.0001
All Infectious Disease Indications	49	F1-Score	0.78	90%	(0.65, 0.86)	0.0012
All Oncology Indications	48	F1-Score	0.85	90%	(0.77, 0.90)	0.0004
All NME Indications	77	F1-Score	0.82	90%	(0.73, 0.88)	0.0002
All Biologics Indications	68	F1-Score	0.87	90%	(0.80, 0.91)	0.0002

Validation Subgroups	N	Metric	Performance	Significance	CI (%)	p-value
All Indications	157	Accuracy	82.8%	99%	(74.2, 89.8)	0.0001
All First-in-class Indications	126	Accuracy	82.5%	99%	(72.8, 90.3)	0.0001
All Age-related Indications	119	Accuracy	81.5%	99%	(71.3, 89.7)	0.0001
All Infectious Disease Indications	49	Accuracy	81.6%	90%	(70.1, 88.4)	0.0020
All Oncology Indications	48	Accuracy	77.1%	90%	(65.1, 84.9)	0.0004
All NME Indications	77	Accuracy	83.1%	90%	(74.4, 88.6)	0.0002
All Biologics Indications	68	Accuracy	80.9%	90%	(71.3, 87.0)	0.0002

Validation Subgroups	N	Metric	Performance	Significance	CI (%)	p-value
All Indications	157	Specificity	72.7%	99%	(57.5, 85.3)	0.0001
All First-in-class Indications	126	Specificity	71.7%	99%	(55.6, 85.0)	0.0001
All Age-related Indications	119	Specificity	69.2%	99%	(51.8, 83.9)	0.0001
All Infectious Disease Indications	49	Specificity	77.4%	90%	(62.1, 86.4)	0.0040
All Oncology Indications	48	Specificity	41.7%	90%	(23.4, 64.2)	0.0080
All NME Indications	77	Specificity	81.0%	90%	(68.3, 88.3)	0.0018
All Biologics Indications	68	Specificity	57.1%	90%	(39.7, 72.3)	0.0016

Data Table 1. ABNN primary performance measures.

Data Table 2. ABNN's All Performance Measures

Validation Subgroups	N	SEN	F1	ACC	SPE	NPV	PPV	FOR	FDR	FPR	FNR	PLR	NLR	DOR	MCC
All Indications	157	**90.1%**	0.86	82.8%	72.7%	84.2%	82.0%	18.0%	17.8%	27.3%	9.9%	3.30	0.14	24.29	0.65
All First-in-class Indications	126	**92.4%**	0.85	82.5%	71.7%	89.6%	78.2%	10.4%	21.8%	28.3%	7.6%	3.26	0.11	30.86	0.66
All Age-related Indications	119	**91.0%**	0.85	81.5%	69.2%	85.7%	79.2%	14.3%	20.8%	30.8%	9.0%	2.96	0.13	22.88	0.63
Infectious Disease Indications	49	**88.9%**	0.78	81.6%	77.4%	92.3%	69.6%	7.7%	30.4%	22.6%	11.1%	3.94	0.14	27.43	0.64
Oncology Indications	48	**88.9%**	0.85	77.1%	41.7%	55.6%	82.1%	44.4%	17.9%	58.3%	11.1%	1.52	0.27	5.71	0.34
Other Indications	18	**92.9%**	0.93	88.9%	75.0%	75.0%	92.9%	25.0%	7.1%	25.0%	7.1%	3.71	0.10	39.00	0.68
Neurology Indications	18	**85.7%**	0.86	88.9%	90.9%	90.9%	85.7%	9.1%	14.3%	9.1%	14.3%	9.43	0.16	60.00	0.77
Autoimmune Indications	10	**83.3%**	0.77	70.0%	50.0%	66.7%	71.4%	33.3%	28.6%	50.0%	16.7%	1.67	0.33	5.00	0.36
Respiratory Indications	7	**100.0%**	1.00	100.0%	100.0%	100.0%	100.0%	0.0%	0.0%	0.0%	0.0%	-	0.00	-	1.00
Endocrine Indications	7	**100.0%**	1.00	100.0%	100.0%	100.0%	100.0%	0.0%	0.0%	0.0%	0.0%	-	0.00	-	1.00
All NME Indications	77	**85.7%**	0.82	83.1%	81.0%	87.2%	78.9%	12.8%	21.1%	19.0%	0.14	4.50	0.18	25.50	0.66
All Biologic Indications	68	**91.5%**	0.87	80.9%	57.1%	75.0%	82.7%	25.0%	17.3%	42.9%	0.09	2.13	0.15	14.33	0.53

Data Table 2. SEN: Sensitivity; F1: F1-score; ACC: Accuracy; SPE: Specificity; NPV: Negative Predictive Value. PPV: Positive Predictive Value; FOR: False Omission Rate;FDR: False Discovery Rate; FPR: False Positive Rate; FNR: False Negative Rate; PLR: Positive Likelihood Ratio; NLR: Negative Likelihood Ratio; DOR: Diagnostic Odds Ratio; MCC: Matthews Correlation Coefficient; NME: New Molecular Entity (small molecule drugs);