Published in final edited form as:

Anal Chem. 2021 June 08; 93(22): 7763–7773. doi:10.1021/acs.analchem.0c04850.

Multiomic Big Data Analysis Challenges: Increasing Confidence in the Interpretation of Artificial Intelligence Assessments

Melanie T. Odenkirk,

Department of Chemistry, North Carolina State University, Raleigh, North Carolina 27606, United States

David M. Reif,

Department of Biological Sciences and Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina 27606, United States;

Erin S. Baker

Department of Chemistry, North Carolina State University, Raleigh, North Carolina 27606, United States;

Abstract

The need for holistic molecular measurements to better understand disease initiation, development, diagnosis, and therapy has led to an increasing number of multiomic analyses. The wealth of information available from multiomic assessments, however, requires both the evaluation and interpretation of extremely large data sets, limiting analysis throughput and ease of adoption. Computational methods utilizing artificial intelligence (AI) provide the most promising way to address these challenges, yet despite the conceptual benefits of AI and its successful application in singular omic studies, the widespread use of AI in multiomic studies remains limited. Here, we discuss present and future capabilities of AI techniques in multiomic studies while introducing analytical checks and balances to validate the computational conclusions.

Graphical Abstract

Corresponding Authors: David M. Reif – Department of Biological Sciences and Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina 27606, United States; dmreif@ncsu.edu, **Erin S. Baker** – Department of Chemistry, North Carolina State University, Raleigh, North Carolina 27606, United States; ebaker@ncsu.edu.

Complete contact information is available at: <https://pubs.acs.org/10.1021/acs.analchem.0c04850>

The authors declare no competing financial interest.

Extensive knowledge about disease etiology is essential for accurate and effective patient diagnosis, prognosis, and treatment.¹ Therefore, a thorough assessment of the molecular alterations occurring prior to symptoms and through treatment provides vital information about the best preventative and intervention strategies. To date, the characterization of molecular perturbation and interactions in diseases has mainly focused on an in-depth interrogation of individual biomolecule classes (*e.g.* proteins, genes, metabolites, etc.). Defining disease-related perturbations through the lens of a singular molecule type or one, however, blinds the researcher to the rest of the molecular changes and interactions taking place in the system. Because interactions between biological classes are ubiquitous, gaining a holistic view of the disease is thus impossible in a singular omic study.² Furthermore, some mechanisms such as inflammation occur in multiple disease states and elude definitive diagnosis in even the most comprehensive singular omic analysis.³ Thus, extended molecular approaches are needed to address these limitations.

Over the past decade, initiatives to increase the understanding of molecular perturbations within a given system have moved toward multiomics, an information-rich approach that examines several biological classes simultaneously. Multiomic analyses extend the insights obtained from singular omic studies (*e.g.* genomics, proteomics, metabolomics, etc.) by measuring and correlating those from multiple biomolecular classes to gain a greater understanding of the phenotype expressed. Therefore, multiomics enable a holistic view of what could happen (genomics and transcriptomics) and how it is happening (proteomics, metabolomics, lipidomics, exposomics, and glycomics).⁴ Additionally, evaluating the variation within and between biomolecule classes overcomes the limitations associated with the limited view of singular omic measurements. Thus, multiomics is quickly becoming essential for disease discovery analyses. Numerous caveats to this approach, however, arise from the increased number of experimental extractions and instrument runs, such as the immense data complexity and time required to assess all the relevant associations often being extremely long and at times impossible. These challenges make multiomic data analysis and interpretation a huge bottleneck for practical, routine application. Furthermore, the intricacy of all the molecular associations present in the data is impossible to manually annotate and can be easily biased by experimental collection and analysis protocols.⁵ Powerful computational methods are therefore necessary to elucidate relationships in a

biological system and attain the greatest payoff from multiomic analyses. This necessity, however, requires researchers to move away from their manual annotation comfort zone and into approaches that are highly computational.

Artificial intelligence (AI) is a toolbox of computer-based techniques able to accomplish tasks as well as or better than humans if trained properly (Figure 1). With so many avenues such as deep learning, clustering, and dimensionality reduction, AI is exceptionally suitable to address the challenges surrounding multiomic data analyses. The development and application of AI in multiomics has thus been actively explored for this purpose. Specific examples of its use include strengthening existing data extraction and interpretation capabilities through chemometrics and both supervised (outcome established) and unsupervised (outcome unknown to model) learning approaches.⁶ Despite these capabilities and potential advancements, the extreme biological diversity in humans and complexity of multiomic data sets has left scientists weary of trusting AI-generated conclusions without a means of validation.⁷ Therefore, the aim of this perspective is to (1) detail how and where AI approaches can extend existing multiomic workflows and (2) explore computational and experimental validation strategies of AI-based results to explain limitations, overcome existing skepticism, and increase the widespread application of AI in multiomic evaluations.

KNOWN LIMITATIONS IN MULTIOMIC ANALYSES

Decades of instrumental advancements in each omic area have enabled deeper coverage, enhanced automation, and increased throughput. These advancements have in turn contributed dramatically to the feasibility of multiomic experiments as many of the innovations can be directly applied to other measurement types. Mass spectrometry (MS), for example, is a popular platform routinely applied for the analysis of biomolecules. Despite innovations in MS technology and its great depth of coverage possible, MS remains challenged by molecules with varying ionization efficiencies, in-source fragmentation, numerous isomeric species, and sample dynamic ranges exceeding 10^5 . These limitations result in only a subset of analytes being confidently observed and quantified. Experimentally, the implementation of orthogonal separation techniques, such as chromatography separations (LC and GC) and ion mobility spectrometry (IMS), provide an additional dimension for distinguishing species and increasing the confidence of identifications.⁸ However, even with these additional dimensions and fragmentation information from both data dependent acquisitions and data independent acquisition, numerous features are unable to be annotated or identified. These undetected or unidentified molecular species are typically referred to as “dark matter” with the resulting gaps of information significantly impairing the biological context annotated within a given system. For example, the structural diversity of metabolomics results in only 1.8% of untargeted metabolomics spectra being annotated using MS.⁹ Incomplete coverage is, however, not unique to metabolomics. All omic techniques suffer from partial coverage, albeit at different degrees of severity. Genomics, for example, has been extensively characterized for protein-coding regions of DNA, but these only represent a portion of the genome.¹⁰ As a majority of DNA is noncoding, efforts to understand how these regions are carried out downstream during transcription to RNA and ultimately translation to proteins is still being actively

investigated.¹¹ These gaps in coverage therefore inhibit the full understanding of phenotype dysregulation. Thus, multiomics data interpretation is inherently limited by the quality, depth, and biological knowledge possible within singular omic studies.

The first roadblock for increasing multiomics capabilities is to address the missing coverage of singular omic analyses. To overcome these impediments, two challenges must be addressed, including how to deal with (1) unknown signals and (2) partial coverage. While partial coverage is prevalent across all omics, the combined infancy, breadth, and size of chemical space covered by metabolomics puts this field at perhaps the greatest disadvantage for confident and comprehensive annotation. Unlike genomics, transcriptomics, and proteomics, metabolomics lacks structural building blocks and instead the estimated well over 100 000 unique metabolites coexist within a limited mass range of 60–2000 m/z with a majority falling below 600 m/z . For MS applications which rely on mass to make identifications, discriminating between the potential molecular matches within this small, populous space is critical. Confidence in analyte identifications has been defined in metabolomics as a tiered function based on experimental information. The levels thus range from the observation of a unique feature (level 5), assignment of molecular formula (level 4), tentative structure annotation (level 3), putative identification (level 2), and, finally, validation with a reference standard (level 1).¹² Unsurprisingly, as identification requirements increase, the number of species meeting each confidence level drastically decreases due to the additional expenses of buying all the standards and also whether purchasing is even possible. Matching MS/MS spectra for a level 2 annotation is also limited by the number of experimental MS/MS spectra available. Currently, 351 754 of the molecules in the human metabolome database (HMDB) have either experimental or theoretical MS/MS, GC-MS, or NMR reference spectra, but less than 10% of these have been experimentally authenticated.¹³ This limitation also adds to the greatest hurdle in metabolomics of assigning a unique feature to a molecular structure. However, within each omics discipline, the biomolecule classes being investigated diverge in their approaches to exploring dark matter. For the more established omics such as proteomics, routine proteomic workflows require identification from MS/MS spectra or presence in the genome as a prerequisite for further data analysis leading to an estimated 50% of dark proteome matter being neglected.¹⁴ Sub-branches of proteomic analyses (e.g., glycomics and phosphoproteomics) have become a prominent means of exploring post-translational modifications (PTMs) which are hypothesized to be the predominant entities in the proteome undiscovered space.¹⁵ Genomic and transcriptomic fields have benefitted from decades of thorough characterization for regions with established biological implications. Preliminary investigation into the central dogma roles of DNA and RNA, however, only reflects a subset of the biological roles of these species as evidenced by the human genome project finding that <10% of diseases were associated with coding regions.¹⁶ Recent efforts have focused on elucidating the remaining genome and transcriptome space through epigenetics and annotation of noncoding DNA and RNA biological implications. Thus, the complexity across all omics is still being actively investigated and understood but ambiguity and unknowns will continue to exist. This then begs the question of why the association between unknowns is so commonly emphasized in metabolomics and not all other omic areas? While other omics have shifted emphasis onto subareas, metabolomic workflows often focus on

both known and unknown features, including potential artifacts such as those from in-source fragments, adducts, or heterodimers that might misguide data interpretation and further complicate the annotation of disease perturbations.⁹ Therefore, confident analyte annotation is a substantial obstacle for multiomic correlation.

AI USE IN UNKNOWN ANALYSES

Assigning unknowns with confidence is still extremely difficult in all omic studies, but freely available software and databases that interface AI technologies to enhance analyte identification coverage and confidence are proving to be extremely promising for proteomics and metabolomics alike.^{17–19} In metabolomic studies, a majority of the existing databases, such as HMDB, METLIN, and NIST, among others, identify metabolite monoisotopic masses and isotopic ratio fingerprints.^{13,20,21} This information allows for the prediction and prioritization of chemical formulas and candidate structures based on similarity searches with computationally or experimentally generated MS/MS spectra. However, overlapping structural elements are often unresolved from this information alone, resulting in uncertainty in assigning the identification. Filtering databases by sample type, organism, and other factors can diminish the number of unlikely matches, but for a significant number of features, this ambiguity still remains. Database complexity is also often presented in lengthy tables that make visualizing all potential annotations difficult, further precluding how metabolome coverage compares to other experiments. Developments in this area, however, include the global natural product social (GNPS) and Reanalysis of Data User (ReDU) from the Dorrestein laboratory. These networking tools enable the visualization of public repositories and user data simultaneously, illustrating structural associations of the MS and MS/MS data, as well as coverage across different instruments and biological matrixes.^{22,23} Text mining by the Fiehn Laboratory has also been applied to expand *in vivo* databases from metabolites previously measured in published studies, resulting in the largest *in vivo* database of human blood small molecules.²⁴ *In silico* permutation of MS/MS spectra in databases such as METLIN, MoNA, and others have also created the ability to make level 1 and 2 identifications that lack *in vivo* fragmentation information.^{21,25–27} Extrapolation of existing metabolite databases traditionally centered on primary metabolites to include secondary and xenobiotic metabolites, through tools like BioTransformer and SyGMA, is also critical to increase the space and relevance of databases.^{28,29}

Machine learning (ML) within organic synthesis has also been utilized to extrapolate known reaction mechanisms.³⁰ While still in development, this approach has shown a lot of promise for advancing fields like drug discovery and pharmaceutical development.^{31–33} The prediction of MS/MS spectra has also been utilized greatly for matching unknown spectra. These predictions have increased the size of databases exponentially to provide more information to researchers, however this has also heightened concern for false identifications due to the validity of theoretical MS/MS spectra prediction for diverse molecular species from singular fragmentation approaches. To date, comparable MS/MS *in silico* reference spectra have been found across numerous algorithms (i.e., MAGMA, CFM-ID, MS-FINDER, Met-FragCL), thereby increasing confidence in this approach.^{25,34–41} The IMS CCS compendium by the McLean lab also provides ~3800 CCS values of small molecules generated experimentally to facilitate unknown feature comparisons.⁴² These

values and those obtained from chromatography measurements can then be utilized to predict retention time (RT) and CCS for matching to experimental analyses when standards are unavailable.^{43–46} AI approaches are utilizing this information to provide the ability to generate spectra, predict metabolism and mine text, in addition to experimental validation (Figure 2). Thus, many researchers are beginning to integrate these into discovery workflows into their studies to provide a complementary means of extending databases and shedding light on the dark matter of metabolomics.

Thus far, our discussion of unknowns has focused on metabolomics and expanding what is known about the metabolome with AI approaches. Unknown features and the associated challenges discussed herein extend to all forms of omics. Proteomics databases, for example, often stem from the sequencing of an organisms genome and the ultimate translation of that information to proteins likely present in the system.⁴⁷ In actuality, this simplistic outline can be influenced by potential factors such as noncoding DNA and alternative splicing patterns, among many others.^{48,49} Furthermore, the importance of evaluating how proteins in the microbiome effect health cause even further challenges since many are not sequenced. Unknowns surrounding genomics and transcriptomics also hinder the accuracy of proteomic databases with an estimated error rate of 0.1% for most genome assemblies and the implications of coding versus noncoding regions of DNA and RNA presently being undetermined.^{50,51} Additionally, post-translational modifications (PTMs) also contribute to a significant percentage of the human proteome and since new PTMs are being uncovered yearly, there are still many that are unknown.^{11,48} Thus, the biological complexity across all omics is still being actively investigated and understood, but ambiguity and unknowns will continue to exist across all omics.

While the progress possible with AI-generated approaches for unknown identification is extremely exciting, its validation can be extraordinarily time-consuming and difficult. Test data sets are the easiest way to check AI approaches are working correctly by assessing their performance on a limited set of known identifications not used in the prior learning steps. This approach can, however, only go so far. Then, the use of existing analytical techniques is needed to confirm the AI findings to provide another checkpoint for assessing the number of accurate analyte annotations within databases. However, based on the molecule type, challenges such as half-lives, poor ionization, and low abundance can impede this evaluation. While optimizing method development can surmount obstacles facing experimental validation, this strategy can quickly become lengthy and expensive. *In silico* prediction of properties and AI method develop offer promising ways to expedite this process following sufficient model training with structurally similar species. For example, metabolism prediction is readily amenable to verification through labeling strategies.^{52,53} However, the depth of coverage from analytical assays will always be less than its computational counterpart and, as a result, a portion of database matches will be left unvalidated. The faith in AI approaches therefore relies on experimental validation of a subset of diverse experimentally generated metabolites to address concerns of AI capabilities and provide scientists with greater confidence in the benefits of AI for assigning the tentative identifications.

AI FOR MISSING ANALYTES

Another challenge for multiomic studies is partial coverage as some analytes go completely undetected or others are only observed in a subset of samples. Addressing missing analytes first requires a scientist to assign reasoning for why a value is missing. Is the analyte not present in the sample? Does the analyte signal fall below the limit of detection for the analysis used? Is this missing signal due to a specific sample preparation or instrumental error? Unfortunately, the true reason for a missing value is often not something that can be confidently pinpointed, but the effects of missing values, including either imputation, removal, or no adjustment, are dictated by missing value origins, and when falsely corrected for, can have lasting effects on the statistical analysis and data interpretation (Figure 3a).⁵⁴

Single cell transcriptomic analyses have become of interest for understanding cellular homogeneity. However, with a significant reduction of starting material, these studies have seen a missing value rate of as much as 30% from low capture efficiency, technical variation, or stochastic gene expression.^{55,56} This has created an overwhelming number of strategies for model-^{55,57–60} and smooth-based^{61,62} imputation and then data reconstruction with either deep learning^{63–68} or matrix-based methods^{69,70} that can be difficult for experimentalists to select from. Investigations of imputation method variability have also shown fluctuation across methods to further complicate making a choice for how to best handle missing values in RNA-seq single cell analyses.⁷¹ This same challenge of missing values is abundantly present in proteomics data sets as well, where missing values make up a substantial portion of DDA experiments and often require these same imputations.⁷² While imputation is a plausible workaround, it is not always the correct means for handling methods and also includes caveats such as increasing the false positive rate that has lead others to explore strategies for analysis that overlook missing values altogether. Statistical analyses, for example, have been completed using sequential projection and qualitative independence tests for groupwise comparisons.⁷³ Generative models are another alternative solution for overlooking missing data that have been applied in genomic data analysis.⁷⁴

From these examples, it is clear AI offers a variety of options for missing values, but these strategies still rely on our initial question of why a value is missing at all. Fortunately for us, *in silico* methods and AI have the potential to aid in differentiating low abundance signals from random error or absent values. For single-cell-level analyses, cellular heterogeneity has the potential to be a significant contributor to missing values as only a select portion of the genome is likely expressed, influencing the downstream molecules present.⁷⁵ On a larger scale, AI has been used to explain heterogeneity of complex diseases through subgroup identification in unsupervised workflows.⁷⁶ However, the success of this approach has been sparsely validated through follow-up studies. Repeated single-cell analysis is another means where these same AI workflows could be applied for identifying cell types and determining what should be expressed from this designation to allow for a better assessment of why signals may be missing.^{77,78} Another AI strategy to provide insight into the origin of missing values is assessing the theoretical likelihoods for molecule half-life. Predicting biologically relevant species serves as a good starting point for suggesting species likely present in samples that have previously gone unannotated.^{28,29,79–82} Molecular docking (MD) simulations is an extension

of pathway predictions that has shown promising insight for predicting interactions in drug discovery pipelines.⁸³ These capabilities allow for an assessment of interactions and binding efficiencies between small molecules and enzymes.^{83,84} Extrapolation of these results can provide information to justify missing value origins through computational measurements.^{83,84} This ability is significant because validating the presence of metabolites with short half-lives solely with experimental approaches has complicated their detection. However, promiscuous species such as acyl-CoA thioesterase 4, an enzyme that is well-recognized to hydrolyze all acyl-CoA derivatives, challenge the application of molecular docking, which typically only focuses on a singular interaction and not inhibitory effects of other metabolites or fluxes in metabolite abundance from other enzymes.⁸⁵

Quantitative structure–activity relationships (QSAR) is another computational approach that can be used to provide similar structures and predict activities from physiochemical properties.⁸⁶ Together, molecular docking and QSAR workflows are commonly implemented to develop drugs for specific enzymes through their complementary validation of the biological relevance of potential inhibitors. This approach has shown to be incredibly efficient in drug discovery pipelines, where QSAR modeling results aid in prioritizing targeted assays.⁸⁷ Cryogenic electron microscopy (cryo-EM), X-ray diffraction (XRD), and nuclear magnetic resonance (NMR) are experimental techniques that allow for *in vitro* assessment of structural biology and ligand binding.^{88,89} Gene knockout experiments have also become a common procedure for clarifying the biological significance of interactions with other biomolecules and better informing the understanding of protein–ligand interactions.^{90,91} This insight is also vital for informing therapeutic intervention strategies and developing methods for earlier detection. The rigorous experimental characterization of knockout genes has also allowed for metabolite profiling databases such as the one developed for *Arabidopsis* (*Arabidopsis thaliana*) to predict the molecular changes likely to occur from a genetic mutation.⁹² The success of this workflow could be greatly expounded by the complementary integration of MD, QSAR, and other computational efforts (Figure 3b).^{93–96}

USING AI TO INTEGRATE MULTIOMIC ANALYSES

Once highly confident singular omic analyses are attained, an additional challenge for multiomic studies is how to combine these results. Multiomic analyses of biological systems result in many data sets with an overwhelming amount of information from each omic study. Manually elucidating relationships between biomolecule classes is an exhaustive and infeasible task, requiring expertise across multiple disciplines and an ability to decode complex relationships not obvious to the human eye. AI technologies thrive in identifying complex patterns, making them well-suited for investigating associations even experienced scientists fail to recognize. These techniques have been applied for both singular and multiomic analyses with enhanced performance in sensitivity, specificity, and efficiency.⁹⁷ For example, networks of disease and gene associations have elucidated similarities in the biological functions of comparable disease mutations that have been extended to other biomolecule classes.^{98–101}

AI data interpretation falls within two categories: supervised, where the model is told the outcome of interest (i.e., disease versus control), or unsupervised, where distinctions are made without any knowledge of an outcome. Supervised approaches are united by the classification of data into categories which is typically accomplished first by parsing data into training and tests sets. To begin, a model is built from a training data set and then evaluated on the test data.¹⁰² Algorithms are then compared by their accuracy, comprehensibility, learning time, and speed metrics.¹⁰² ML, discriminant analysis, and k-nearest neighbors are all examples of supervised approaches with their own advantages and challenges.¹⁰² ML algorithms, for example, often consist of a trade-off between interpretability and performance. Neural network models are commonly referred to as black boxes that produce an output with little information on discriminating factors. This can be troublesome for experimentalists who seek tangible outcomes that can be validated in targeted follow-up analyses for further pursuit as clinical biomarkers. Thus, uncovering the mechanisms behind the black boxes is a priority.^{103,104} One approach for this is mimic learning which combines the transparency of simpler ML approaches, while maintaining the complex association of more advanced algorithms. In mimic learning, training is performed on simpler models from deep learning outputs, which sheds light on the black box algorithms and improves the simpler model performance.¹⁰⁵ However, uncovering the black box of complex computational algorithms through workflows like mimic learning are relatively new, and their performance is still being evaluated. Others have taken to understanding more complex ML techniques through back-propagation and dropout which omits model terms to assess individual variable significance.^{106,107} Decisions trees, another ML approach, are simpler algorithms that provide more user-friendly conclusions. Building and evaluating the appropriate model is an imperative step to leveraging supervised AI techniques, with most methods providing the best performance for data sets with thousands of entries to avoid overfitting.¹⁰⁸ For all omic workflows, the high sample count is a challenge, but for those with analytical techniques that are consistent, public repositories offer a means of further model validation leveraging experimental data. This again results in a handicap for metabolomics where different sample preparation and data analysis techniques can produce publicly available data sets with greater disparity.

Unsupervised approaches such as clustering, multistep analyses, Bayesian methods, and networking have no prior knowledge of groups and instead use data to build associations.¹⁰⁹ While unsupervised approaches appear unbiased, these algorithms are also subject to being influenced by noise not related to the condition of interest. However, the caveat of solely relying on data to make distinctions can also allow for the detection of disease subgroups, thereby providing vital insight for designing therapeutics. Correlation has become a popular tool for unsupervised investigation of interactions; however, relating unknown signals across other biomolecule classes can hinder the elucidation of the biological relationships.^{110,111} Efforts to provide associations of unknown species in pathways have thus been ongoing to overcome existing limitations of correlation networks.¹¹² Semisupervised approaches to develop learning algorithms with and without known outcomes have benefitted from the advantages of both unsupervised and supervised workflows.¹¹³ Even with the variety of data interpretation techniques for omic analyses, as the number of data sets integrated in multiomics grows, algorithms capable of preserving both global and local features with

a sufficient level of detail are not viable.¹¹⁴ Therefore, the application of multiple AI approaches is necessary to provide thorough annotations of a biological system.^{111,115,116} Analyses of this type are currently taking place in many studies. For example, the pathway annotation of metabolite and protein relationships is made possible through KEGG and Wikipath-ways.^{117,118} Other workflows like DIABLO have also been used for identifying driving factors in disease development with discriminant analysis.¹⁰² Additionally, iOmicsPASS has been developed and applied for identifying disease subtypes through network analysis.¹¹⁶ Where these workflows often fall short is instances where singular omics are in disagreement when combined, which is often the case for proteomic and transcriptomic integration.^{119,120} The disparity between these biomolecules, however, is not surprising as variation in the timeline of protein versus RNA turnover can contribute significantly to discrepancies between the analyses.¹²⁰ Furthermore, variations in analysis methods, data processing, and statistics can also influence the data. Currently, discordance of multiomic studies typically forces scientists to choose what they consider to be the correct picture of the system perturbation. Unsurprisingly, this choice is heavily biased by user experience, but extending computational modeling of interactions and molecular turnover rate can aid in informing omic differences resulting from this fluctuation.

Differences between AI and experimental conclusions have seen similar distrust from experimentalists even when disagreement is minimal. However, AI offers another strategy to learn from when calculations agree with experimental findings and when they do not. Deep learning of the deviations of experimental and computational abundances can be logically related to analyte physiochemical properties in hopes of pulling out specific molecular classes with similar prediction accuracy. Deep reinforcement learning is a reward-driven AI approach that can be used to compensate for the agreement between computational and experimental approaches, thereby providing classes of molecules most likely to be accurately defined from computational methods.¹²¹ Trends in failed computations are also incredibly informative as these have the potential to elucidate unknown interactions and roles. However, deep learning mechanisms are hindered by their reliance on exceptionally large databases for training as well as in the decision-making processes that are difficult for users to fully understand and trust. The agreement of experimental and computational data for known data sets is therefore essential to help alleviate concerns of uncertainty.

While docking and QSAR modeling have traditionally focused on small molecule-enzyme interactions, other biomolecule associations are also of interest for assessing missing values. Because the central dogma of DNA to RNA to proteins allows for the abundance of RNA and protein species to be predicted from the genome sequence, deep learning algorithms utilize *a priori* knowledge of the intricate biological factors that converge to influence protein and RNA abundances from DNA. A variety of AI applications have also worked to adapt algorithms to existing biological complexity for predicting molecular abundances.^{122–125} However, the biological mechanisms influencing transcription and translation including alternative splicing, remain somewhat elusive.⁴⁸ To better understand these processes and their implications on molecular concentrations across omics, computational efforts have focused on modeling approaches.¹²⁶ However, experimental validation again serves as a crucial benchmark for the *in silico* explanation, where the results serve to refine the accuracy of computations. Altogether, the application of these

in silico approaches and AI strategies aid in elucidating missing value origins, thereby enabling more accurate imputation strategies to improve statistical analyses across all omic measurements.¹²⁷ Deep learning has been applied in transcriptomics for predicting cryptic splicing patterns and DNA/RNA–protein binding sites, a remarkable achievement to aid in explaining RNA variation for comparing proteomic and transcriptomic results.^{128–131} Experimental validation with RNA-seq analysis verified 75% of cryptic splicing variants, further proposing AI as a viable approach for elucidating biological mechanisms we are still aiming to comprehend.¹³⁰

Longitudinal multiomic studies are the most comprehensive characterization of disease perturbations possible to date with a majority focused on defining disease progression. However, biological variability arising from factors like age, sex, diet, and lifestyle can cause great fluctuations across molecules observed, greatly obscuring disease and healthy diagnoses (Figure 4).^{132,133} The addition of phenotype, which is the expression of certain characteristics based on interactions between genetics and the environment, is therefore another component of omics required to rigorously evaluate disease etiology and enhance patient prognosis.¹³² Traditionally, metadata effects on subject variation are commonly reduced by experimentally matching patient variables such as age and sex to avoid potential factors that eclipse comparisons.^{134,135} While this approach illuminates biological perturbations related to the comparisons, it also neglects how variability affects subjects across a population and results in a biased understanding of a disease. Unfortunately, this often becomes apparent in the validation stages of biomarker assessment, resulting in great losses of time and financial investment. Therefore, advancements in current ML to correlate analytes and metadata are essential but will remain challenging due to the biodiversity of the human population and their surrounding environments.¹³⁶

While we have only focused thus far on how direct molecular interactions limit data integration and interpretation, indirect interactions are also essential and provide an even greater challenge in multiomic analyses. Weak, indirect interactions within a cellular system are incredibly difficult to verify as they are often destroyed in experimental preparation procedures and their binding is absent upon analysis. Methods for elucidating indirect associations must therefore turn to statistical patterns for detailing these relationships.¹³⁷ The most obvious means of exploring indirect interactions is through correlation networks, which result from hundreds of identified biomolecules and quickly overwhelm human interpretation capabilities. To address this challenge coexpression networking techniques have been developed with popular methods including Search Tool for Recurring Instances of Neighboring Genes (STRING) and Weighted Gene Correlation Network Analysis (WGCNA) methods for proteomic and genomic data.^{111,138} These tools have opened the door for interrogating relationships between species with WGCNA networking providing another means of visualizing node associations beyond direct, established interactions. In general, these networks serve as viable AI approaches for connecting species by their relative abundances and extending these relationships across biomolecule classes.^{116,139} However, the propensity of correlations to have high type I and type II errors requires that these associations be scrutinized and validated through experimental strategies like assays, gene knockout, and binding simulations.¹⁴⁰ Factor analysis has also been used to elucidate the major variables driving dysregulation across a variety of omic layers.¹⁴¹ While

these methods offer a means for AI to investigate indirect interactions, experimentally these require system manipulations which are difficult to implement for all potential associations. Thus, this showcases how deep learning has the potential to understand trends and dive deeper into our existing knowledge of biology. However, validation of indirect interactions is, to our knowledge, impossible to accomplish experimentally with well-established interactions being the only benchmark for assessing performance. Improving our singular-omic knowledge, however, offers us the greatest chance of increasing our understanding of what currently looks like incomplete, indirect associations across biology.

CONCLUDING REMARKS

Multiomic analyses are quickly becoming imperative to probe molecular changes that might be the easiest for diagnosis or prognosis. However, caveats to multiomic studies arise from the increased data complexity and gaps during singular omic assessment where missing analyte annotations and unknown values significantly impede their inclusion or cause incorrect conclusions. Disease characterization and environmental perturbations are, however, moving toward more multiomic studies, so we must find ways to fill these knowledge gaps and obtain a more holistic understanding of the patients and systems being evaluated. This challenge must first be addressed in the singular omic studies as these errors propagate to multiomic evaluations. To date, AI strategies have already proven monumental in advancing many current omic capabilities through the expansion of chemical space, missing value estimation, complex pattern recognition, and integration of singular omic information for comprehensive inquiry into biomarker identification and subgrouping (Figure 5). However, current skepticism due to difficulties in validating the conclusions generated by AI techniques has resulted in less acceptance than expected. While analytical approaches and manual validation can only go so far as data is being generated on the subminute time scale at times and the combined data required by multiomic studies needs comprehensive evaluations to attain novel biological insights. Therefore, to address this doubt, the integration of AI, analytical chemistry, and visualization techniques on subsets of known analyses is extremely important. These validations must be rigorous, comprehensive, and capable of being visualized so they can be thoroughly understood and trusted. Furthermore, to accomplish this goal, developers must work with analytical chemists to optimize AI performance first on the singular omic scale and then build up to the complex associations across these studies. When these two worlds fully collide, the appropriate checks and balances will allow AI to establish new molecular pathways and associations and provide a holistic understanding of each system being studied for greatly improved disease diagnosis, prognosis, and treatment.

ACKNOWLEDGMENTS

This work was funded, in part, by grants from the National Institutes of Health (P30 ES025128, P42 ES027704, and P42 ES031009) and a cooperative agreement with the United States Environmental Protection Agency (STAR RD 84003201). The views expressed in this manuscript do not reflect those of the funding agencies. The use of specific commercial products in this work does not constitute endorsement by the authors or the funding agencies.

REFERENCES

- (1). Baerheim AFam Pr. 2001, 18 (3), 243–245.

- (2). O'Donnell ST; Ross RP; Stanton C *Front. Microbiol*2020, 10, 3084. [PubMed: 32047482]
- (3). Rotroff DM; Motsinger-Reif AA *Int. J. Genomics*2016, 2016, 1715985. [PubMed: 27689071]
- (4). Schmidt CW *Environ. Health Perspect*2004, 112 (7), A410–5. [PubMed: 15159214]
- (5). Evans AM; O'Donovan C; Playdon M; Beecher C; Beger RD; Bowden JA; Broadhurst D; Clish CB; Dasari S; Dunn WB; Griffin JL; Hartung T; Hsu PC; Huan T; Jans J; Jones CM; Kachman M; Kleensang A; Lewis MR; Monge ME; Mosley JD; Taylor E; Tayyari F; Theodoridis G; Torta F; Ubhi BK; Vuckovic D *Metabolomics*2020, 16 (10), 113. [PubMed: 33044703]
- (6). Huang M; Wang J; Torre E; Dueck H; Shaffer S; Bonasio R; Murray JI; Raj A; Li M; Zhang NR *Nat. Methods*2018, 15 (7), 539–542. [PubMed: 29941873]
- (7). Why the Metabolism Field Risks Missing out on the AI Revolution. *Nat. Metab*2019, 1 (10) 929–930. [PubMed: 32694838]
- (8). Zheng X; Aly NA; Zhou Y; Dupuis KT; Bilbao A; Paurus VL; Orton DJ; Wilson R; Payne SH; Smith RD; Baker ES *Chem. Sci*2017, 8 (11), 7724–7736. [PubMed: 29568436]
- (9). da Silva RR; Dorrestein PC; Quinn RA *Proc. Natl. Acad. Sci. U. S. A*2015, 112 (41), 12549–12550. [PubMed: 26430243]
- (10). Spielmann M; Mundlos S *Hum. Mol. Genet*2016, 25 (R2), R157–R165. [PubMed: 27354350]
- (11). Li J; Liu C *Front. Genet*2019, 10, 496. [PubMed: 31178900]
- (12). Schrimpe-Rutledge AC; Codreanu SG; Sherrod SD; McLean JAJ. *Am. Soc. Mass Spectrom*2016, 27 (12), 1897–1905. [PubMed: 27624161]
- (13). Wishart DS; Feunang YD; Marcu A; Guo AC; Liang K; Vazquez-Fresno R; Sajed T; Johnson D; Li C; Karu N; Sayeeda Z; Lo E; Assempour N; Berjanskii M; Singhal S; Arndt D; Liang Y; Badran H; Grant J; Serra-Cayuela A; Liu Y; Mandal R; Neveu V; Pon A; Knox C; Wilson M; Manach C; Scalbert A *Nucleic Acids Res.* 2018, 46 (D1), D608–D617. [PubMed: 29140435]
- (14). Perdigão N; Heinrich J; Stolte C; Sabir KS; Buckley MJ; Tabor B; Signal B; Gloss BS; Hammang CJ; Rost B; Schafferhans A; O'Donoghue SI *Proc. Natl. Acad. Sci. U. S. A*2015, 112 (52), 15898–15903. [PubMed: 26578815]
- (15). Skinner OS; Kelleher NL *Nat. Biotechnol*2015, 33 (7), 717–718. [PubMed: 26154010]
- (16). Lander ES; Linton LM; Birren B; Nusbaum C; Zody MC; Baldwin J; Devon K; Dewar K; Doyle M; FitzHugh W; Funke R; Gage D; Harris K; Heaford A; Howland J; Kann L; Lehoczky J; LeVine R; McEwan P; McKernan K; Meldrim J; Mesirov JP; Miranda C; Morris W; Naylor J; Raymond C; Rosetti M; Santos R; Sheridan A; Sougnez C; Stange-Thomann Y; Stojanovic N; Subramanian A; Wyman D; Rogers J; Sulston J; Ainscough R; Beck S; Bentley D; Burton J; Clee C; Carter N; Coulson A; Deadman R; Deloukas P; Dunham A; Dunham I; Durbin R; French L; Grafham D; Gregory S; Hubbard T; Humphray S; Hunt A; Jones M; Lloyd C; McMurray A; Matthews L; Mercer S; Milne S; Mullikin JC; Mungall A; Plumb R; Ross M; Shownkeen R; Sims S; Waterston RH; Wilson RK; Hillier LW; McPherson JD; Marra MA; Mardis ER; Fulton LA; Chinwalla AT; Pepin KH; Gish WR; Chissole SL; Wendl MC; Delehaunty KD; Miner TL; Delehaunty A; Kramer JB; Cook LL; Fulton RS; Johnson DL; Minx PJ; Clifton SW; Hawkins T; Branscomb E; Predki P; Richardson P; Wenning S; Slezak T; Doggett N; Cheng JF; Olsen A; Lucas S; Elkin C; Uberbacher E; Frazier M *Nature*2001, 409 (6822), 860–921. [PubMed: 11237011]
- (17). Li Y; Kuhn M; Gavin AC; Bork P *Bioinformatics*2019, 36 (4), 1213–1218.
- (18). Lin YM; Chen CT; Chang J *BMC Genomics*2019, 20 (9), 906. [PubMed: 31874640]
- (19). Zhou XX; Zeng WF; Chi H; Luo C; Liu C; Zhan J; He SM; Zhang Z *Anal. Chem*2017, 89 (23), 12690–12697. [PubMed: 29125736]
- (20). Simón-Manso Y; Lowenthal MS; Kilpatrick LE; Sampson ML; Telu KH; Rudnick PA; Mallard WG; Bearden DW; Schock TB; Tchekhovskoi DV; Blonder N; Yan X; Liang Y; Zheng Y; Wallace WE; Neta P; Phinney KW; Remaley AT; Stein SE *Anal. Chem*2013, 85 (24), 11725–11731. [PubMed: 24147600]
- (21). Smith CA; O'Maille G; Want EJ; Qin C; Trauger SA; Brandon TR; Custodio DE; Abagyan R; Siuzdak G *Ther. Drug Monit*2005, 27 (6), 747–751. [PubMed: 16404815]
- (22). Aron AT; Gentry EC; McPhail KL; Nothias LF; Nothias-Esposito M; Bouslimani A; Petras D; Gauglitz JM; Sikora N; Vargas F; van der Hooft JJJ; Ernst M; Kang KB; Aceves CM; Caraballo-Rodriguez AM; Koester I; Weldon KC; Bertrand S; Roullier C; Sun K; Tehan RM; Boya PC;

Christian MH; Gutierrez M; Ulloa AM; Tejeda Mora JA; Mojica-Flores R; Lakey-Beitia J; Vasquez-Chaves V; Zhang Y; Calderon AI; Tayler N; Keyzers RA; Tugizimana F; Ndlovu N; Aksenov AA; Jarmusch AK; Schmid R; Truman AW; Bandeira N; Wang M; Dorrestein PC*Nat. Protoc*2020, 15 (6), 1954–1991. [PubMed: 32405051]

- (23). Jarmusch AK; Wang M; Aceves CM; Advani RS; Aguire S; Aksenov AA; Aleti G; Aron AT; Bauermeister A; Bolleddu S; Bouslimani A; Caraballo Rodrigues AM; Chaar R; Coras R; Elijah EO; Ernst M; Gauglitz JM; Gentry EC; Husband M; Jarmusch SA; Jones II KL; Kamenik Z; Gouellec AL; McCall LI; McPhail KL; Meehan MJ; Melnik AV; Menezes RC; Montoya Giraldo YANNH; Mothian LF; Nothias-Esposito M; Panitchpakdi M; Petras D; Quinn R; Sikora N; van der Hooft JJ; Vargas F; Verbanac A; Weldon K; Knight RBN; Dorrestain PC*Repository-Scale Co- and Re-Analysis of Tandem Mass Spectrometry Data. bioRxiv*2019. DOI: 10.1101/750471.
- (24). Barupal DK; Fiehn O*Env. Heal. Perspect*2019, 127 (9), 97008.
- (25). Tsugawa H; Kind T; Nakabayashi R; Yukihira D; Tanaka W; Cajka T; Saito K; Fiehn O; Arita M*Anal. Chem*2016, 88 (16), 7946–7958. [PubMed: 27419259]
- (26). Montenegro-Burke JR; Guigas C; Siuzdak G*Methods Mol. Biol*2020, 2104, 149–163. [PubMed: 31953817]
- (27). Kong AT; Leprevost FV; Avtonomov DM; Mellacheruvu D; Nesvizhskii AI*Nat. Methods*2017, 14 (5), 513–520. [PubMed: 28394336]
- (28). Djoumbou-Feunang Y; Fiamoncini J; Gil-De-La-Fuente A; Greiner R; Manach C; Wishart DSJ. *Cheminf*2019, 11 (1), 2.
- (29). Ridder L; Wagener M*ChemMedChem*2008, 3 (5), 821–832. [PubMed: 18311745]
- (30). Gajewska EP; Szymkuc S; Dittwald P; Startek M; Popik O; Mlynarski J; Grzybowski B*Chem.* 2020, 6 (1), 280–293.
- (31). Maryasin B; Marquetand P; Maulide N*Angew. Chem., Int. Ed*2018, 57 (24), 6978–6980.
- (32). Granda JM; Donina L; Dragone V; Long DL; Cronin L*Nature*2018, 559 (7714), 377–381. [PubMed: 30022133]
- (33). de Almeida AF; Moreira R; Rodrigues T*Nat. Rev. Chem*2019, 3, 589–604.
- (34). Blazenovic I; Kind T; Torbasinovic H; Obrenovic S; Mehta SS; Tsugawa H; Wermuth T; Schauer N; Jahn M; Biedendieck R; Jahn D; Fiehn O*JCheminf.* 2017, 9 (1), 32.
- (35). Ridder L; van der Hooft JJ; Verhoeven S; de Vos RC; van Schaik R; Vervoort J*Rapid Commun. Mass Spectrom*2012, 26 (20), 2461–2471. [PubMed: 22976213]
- (36). Ridder L; van der Hooft JJ; Verhoeven S; de Vos RC; van Schaik R; Vervoort J*Rapid Commun. Mass Spectrom*2012, 26 (20), 2461–2471. [PubMed: 22976213]
- (37). Wolf S; Schmidt S; Muller-Hannemann M; Neumann SB*Mol. Cell Proteom.* 2010, 9 (1), 148.
- (38). Tsugawa H; Kind T; Nakabayashi R; Yukihira D; Tanaka W; Cajka T; Saito K; Fiehn O; Arita M*Anal. Chem*2016, 88 (16), 7946–7958. [PubMed: 27419259]
- (39). Verdegem D; Lambrechts D; Carmeliet P; Ghesquiere B*Metabolomics*2016, 12, No. 98.
- (40). Allen F; Greiner R; Wishart D*Metabolomics*2015, 11 (1), 98–110. Allen FGR; Wishart D*Metabolomics*2014, No. 11, 98–110.
- (41). Chao A; Al-Ghoul H; McEachran AD; Balabin I; Transue T; Cathey T; Grossman JN; Singh RR; Ulrich EM; Williams AJ; Sobus J*Anal. Bioanal. Chem*2020, 412 (6), 1303–1315. [PubMed: 31965249]
- (42). Picache JA; Rose BS; Balinski A; Leaptrot KL; Sherrod SD; May JC; McLean J*Chem. Sci*2019, 10 (4), 983–993. [PubMed: 30774892]
- (43). Colby SM; Thomas DG; Nunez JR; Baxter DJ; Glaesemann KR; Brown JM; Pirrung MA; Govind N; Teeguarden JG; Metz TO; Renslow RS*Anal. Chem*2019, 91 (7), 4346–4356. [PubMed: 30741529]
- (44). Domingo-Almenara X; Guigas C; Billings E; Montenegro-Burke JR; Uritboonthai W; Aisporna AE; Chen E; Benton HP; Siuzdak G*Nat. Commun*2019, 10 (1), 5811. [PubMed: 31862874]
- (45). Zhou Z; Xiong X; Zhu ZJ*Bioinformatics*2017, 33 (14), 2235–2237. [PubMed: 28334295]
- (46). Zhou Z; Tu J; Xiong X; Shen X; Zhu ZJ*Anal. Chem*2017, 89 (17), 9559–9566. [PubMed: 28764323]

- (47). Wilhelm M; Schlegl J; Hahne H; Gholami AM; Lieberenz M; Savitski MM; Ziegler E; Butzmann L; Gessulat S; Marx H; Mathieson T; Lemeer S; Schnatbaum K; Reimer U; Wenschuh H; Mollenhauer M; Slotta-Huspenina J; Boese JH; Bantscheff M; Gerstmair A; Faerber F; Kuster BNature2014, 509 (7502), 582–587. [PubMed: 24870543]
- (48). Liu Y; Gonzalez-Porta M; Santos S; Brazma A; Marioni JC; Aebersold R; Venkitaraman AR; Wickramasinghe VCell Rep. 2017, 20 (5), 1229–1241. [PubMed: 28768205]
- (49). Timp W; Timp GSci. Adv2020, 6 (2), No. eaax8978. [PubMed: 31950079]
- (50). Gloss BS; Dinger MExp. Mol. Med2018, 50 (8), 97.
- (51). Apweiler R; Hermjakob H; Sharon NBiochim. Biophys. Acta, Gen. Subj1999, 1473 (1), 4–8.
- (52). Mahieu NG; Huang X; Chen YJ; Patti GJAnal. Chem2014, 86 (19), 9583–9589. [PubMed: 25160088]
- (53). Piqueras MDC; Myer C; Junk A; Bhattacharya SKMethods Mol. Biol2019, 1996, 179–185. [PubMed: 31127556]
- (54). Lazar C; Gatto L; Ferro M; Bruley C; Burger TJProteome Res. 2016, 15 (4), 1116–1125.
- (55). Yang MQ; Weissman SM; Yang W; Zhang J; Canaann A; Guan RBMC Syst. Biol2018, 12 (S7), 114. [PubMed: 30547798]
- (56). Hicks SC; Townes FW; Teng M; Irizarry RABiostatistics2018, 19 (4), 562–578. [PubMed: 29121214]
- (57). Tang W; Bertaux F; Thomas P; Stefanelli C; Saint M; Marguerat S; Shahrezaei VBioinformatics2019, 36 (4), 1174–1181.
- (58). Huang M; Wang J; Torre E; Dueck H; Shaffer S; Bonasio R; Murray JI; Raj A; Li M; Zhang NRNat. Methods2018, 15 (7), 539–542. [PubMed: 29941873]
- (59). Wang J; Agarwal D; Huang M; Hu G; Zhou Z; Ye C; Zhang NRNat. Methods2019, 16 (9), 875–878. [PubMed: 31471617]
- (60). Chen M; Zhou XGenome Biol. 2018, 19 (1), 196. [PubMed: 30419955]
- (61). Gong W; Kwak IY; Pota P; Koyano-Nakagawa N; Garry DJBMC Bioinf. 2018, 19 (1), 220.
- (62). van Dijk D; Sharma R; Nainys J; Yim K; Kathail P; Carr AJ; Burdziak C; Moon KR; Chaffer CL; Pattabiraman D; Bierie B; Mazutis L; Wolf G; Krishnaswamy S; Pe'er DCell2018, 174 (3), 716–729. [PubMed: 29961576]
- (63). Talwar D; Mongia A; Sengupta D; Majumdar ASci. Rep2018, 8 (1), 16329. [PubMed: 30397240]
- (64). Eraslan G; Simon LM; Mircea M; Mueller NS; Theis FJNat. Commun2019, 10 (1), 390. [PubMed: 30674886]
- (65). Arisdakessian C; Poirion O; Yunits B; Zhu X; Garmire LXGenome Biol. 2019, 20 (1), 211. [PubMed: 31627739]
- (66). Amodio M; van Dijk D; Srinivasan K; Chen WS; Mohsen H; Moon KR; Campbell A; Zhao Y; Wang X; Venkataswamy M; Desai A; Ravi V; Kumar P; Montgomery R; Wolf G; Krishnaswamy SNat. Methods2019, 16 (11), 1139–1145. [PubMed: 31591579]
- (67). Deng Y; Bao F; Dai Q; Wu LF; Altschuler SJNat. Methods2019, 16 (4), 311–314. [PubMed: 30886411]
- (68). Lopez R; Regier J; Cole MB; Jordan MI; Yosef NNat. Methods2018, 15 (12), 1053–1058. [PubMed: 30504886]
- (69). Mongia A; Sengupta D; Majumdar AFront. Genet2019, 10, 9. [PubMed: 30761179]
- (70). Zhang L; Zhang SJMol. Cell Biol2021, 13 (1), 29–40.
- (71). Hou W; Ji Z; Ji H; Hicks SCGenome Biol. 2020, 21 (1), 218. [PubMed: 32854757]
- (72). Liu M; Dongre AProper Imputation of Missing Values in Proteomics Datasets for Differential Expression Analysis. Briefings Bioinf. 2020. DOI: 10.1093/bib/bbaa112.
- (73). Stratton KG; Webb-Robertson BM; McCue LA; Stanfill B; Claborne D; Godinez I; Johansen T; Thompson AM; Burnum-Johnson KE; Waters KM; Bramer LMJ. Proteome Res2019, 18 (3), 1418–1425. [PubMed: 30638385]
- (74). Libbrecht MW; Noble WSNat. Rev. Genet2015, 16 (6), 321–332. [PubMed: 25948244]
- (75). Altschuler SJ; Wu LFCell2010, 141 (4), 559–563. [PubMed: 20478246]
- (76). Helal SJ. Comput. Sci. Technol2016, 31 (3), 561–576.

- (77). Mund A; Coscia F; Hollandi R; Kovacs F; Kriston A; Brunner A-D; Bzorek M; Naimy S; Gjerdem LMR; Dyring-Andersen B; Bulkescher J; Lukas C; Gnann C; Lundberg E; Horvath P; Mann M; Profile VO, AI-driven Deep Visual Proteomics defines cell identity and heterogeneity. bioRxiv2021. DOI: 10.1101/2021.01.25.427969.
- (78). Gupta RK; Kuznicki JCells2020, 9 (8), 1751.
- (79). Rydberg P; Gloriam DE; Olsen LBioinformatics2010, 26 (23), 2988–2989. [PubMed: 20947523]
- (80). Terfloeth L; Bienfait B; Gasteiger JJ. Chem. Inf. Model2007, 47 (4), 1688–1701. [PubMed: 17608404]
- (81). van de Waterbeemd H; Gifford ENat. Rev. Drug Discovery2003, 2 (3), 192–204. [PubMed: 12612645]
- (82). Zaretski J; Matlock M; Swamidass SJJ. Chem. Inf. Model2013, 53 (12), 3373–3383. [PubMed: 24224933]
- (83). Meng XY; Zhang HX; Mezei M; Cui MCurr. Comput.-Aided Drug Des2011, 7 (2), 146–157. [PubMed: 21534921]
- (84). Attique SA; Hassan M; Usman M; Atif RM; Mahboob S; Al-Ghanim KA; Bilal M; Nawaz MZInt. J. Environ. Res. Public Health2019, 16 (6), 923.
- (85). Hunt MC; Siponen MI; Alexson SEBiochim. Biophys. Acta, Mol. Basis Dis2012, 1822 (9), 1397–1410.
- (86). Cherkasov A; Muratov EN; Fourches D; Varnek A; Baskin II; Cronin M; Dearden J; Gramatica P; Martin YC; Todeschini R; Consonni V; Kuz'min VE; Cramer R; Benigni R; Yang C; Rathman J; Terfloeth L; Gasteiger J; Richard A; Tropsha AJMed. Chem2014, 57 (12), 4977–5010.
- (87). Sheridan RP; McMasters DR; Voigt JH; Wildey MJJ. Chem. Inf. Model2015, 55 (2), 231–8. [PubMed: 25551659]
- (88). Venien-Bryan C; Li Z; Vuillard L; Boutin JAActa Crystallogr., Sect. F: Struct. Biol. Commun2017, 73 (4), 174–183. [PubMed: 28368275]
- (89). Maveyraud L; Mourey LMolecules2020, 25 (5), 1030.
- (90). Campos AI; Zampieri MMol. Cell2019, 74 (6), 1291–1303. [PubMed: 31047795]
- (91). Hui DYJ. Nutr1998, 128 (11), 2052–2057. [PubMed: 9808665]
- (92). Fukushima A; Kusano M; Mejia RF; Iwasa M; Kobayashi M; Hayashi N; Watanabe-Takahashi A; Narisawa T; Tohge T; Hur M; Wurtele ES; Nikolau BJ; Saito KPlant Physiol2014, 165 (3), 948–961. [PubMed: 24828308]
- (93). Schwaller P; Laino T; Gaudin T; Bolgar P; Hunter CA; Bekas C; Lee AAACS Cent. Sci2019, 5 (9), 1572–1583. [PubMed: 31572784]
- (94). Bakker BM; van Eunen KPerspect. Sci2014, 4 (1–6), 126–130.
- (95). Verma R; Mitchell-Koch KCatalysts2017, 7 (7), 212. [PubMed: 30464857]
- (96). Lo R; Chandar NB; Kesharwani MK; Jain A; Ganguly BPLoS One2013, 8 (12), No. e79591. [PubMed: 24312449]
- (97). Martorell-Marugan J; Tabik S; Benhammou Y; del Val C; Zwir I; Herrera F; Carmona-Saez PDeep Learning in Omics Data Analysis and Precision Medicine. In Computational Biology; Husi H, Ed.; Brisbane, Australia, 2019. DOI: 10.15586/computationalbiology.2019.ch3.
- (98). Goh KI; Cusick ME; Valle D; Childs B; Vidal M; Barabasi ALProc. Natl. Acad. Sci. U. S. A2007, 104 (21), 8685–8690. [PubMed: 17502601]
- (99). Ideker T; Sharan RGenome Res. 2008, 18 (4), 644–652. [PubMed: 18381899]
- (100). Zhang L; Lv C; Jin Y; Cheng G; Fu Y; Yuan D; Tao Y; Guo Y; Ni X; Shi TFront. Genet2018, 9, 477. [PubMed: 30405689]
- (101). Sharifi-Noghabi H; Zolotareva O; Collins CC; Ester MBioinformatics2019, 35 (14), i501–i509. [PubMed: 31510700]
- (102). Singh A; Shannon CP; Gautier B; Rohart F; Vacher M; Tebbutt SJ; Le Cao KABioinformatics2019, 35 (17), 3055–3062. [PubMed: 30657866]
- (103). Zhang Z; Beck MW; Winkler DA; Huang B; Sibanda W; Goyal HAnn. Transl Med2018, 6, 216. [PubMed: 30023379]
- (104). Castelvechi DNature2016, 538 (7623), 20–23. [PubMed: 27708329]

- (105). Che Z; Purushothan S; Khemani R; Liu Y Distilling Knowledge from Deep Networks with Application to Healthcare Domain. ArXiv2015.
- (106). Grapov D; Fahrman J; Wanichthanarak K; Khoomrung SOMICS2018, 22 (10), 630–636. [PubMed: 30124358]
- (107). Beck MWJ. Stat Softw2018, 85 (11), 1–20. [PubMed: 30505247]
- (108). Hicks J; Uchida T; Seth A; Rajagopal A; Delp SJ Biomech. Eng2015, 137, No. 020905.
- (109). Tini G; Marchetti L; Priami C; Scott-Boyer MP Briefings Bioinf. 2019, 20 (4), 1269–1279.
- (110). Lewis R; Guha R; Korcsmaros T; Bender AJ Cheminf. 2015, 7, No. 36.
- (111). Langfelder P; Horvath SBMC Bioinf. 2008, 9, 559.
- (112). Hsu YH; Churchhouse C; Pers TH; Mercader JM; Metspalu A; Fischer K; Fortney K; Morgen EK; Gonzalez C; Gonzalez ME; Esko T; Hirschhorn JN PLoS Comput. Biol2019, 15 (1), No. e1006734. [PubMed: 30640898]
- (113). Huang S; Chaudhary K; Garmire LX Front. Genet2017, 8, 84. [PubMed: 28670325]
- (114). Reymond JL; Probst DJ Cheminf. 2020, 12, 12.
- (115). Koh HWL; Fermin D; Vogel C; Choi KP; Ewing RM; Choi HN PJ. Syst. Biol. Appl2019, 5, 22. [PubMed: 31312515]
- (116). Du J; Yuan Z; Ma Z; Song J; Xie X; Chen Y Mol. BioSyst2014, 10 (9), 2441–2447. [PubMed: 24994036]
- (117). Pico AR; Kelder T; van Iersel MP; Hanspers K; Conklin BR; Evelo C PLoS Biol. 2008, 6 (7), No. e184. [PubMed: 18651794]
- (118). Cheng Z; Teo G; Krueger S; Rock TM; Koh HW; Choi H; Vogel C Mol. Syst. Biol2016, 12 (1), 855. [PubMed: 26792871]
- (119). Wang D Comput. Biol. Chem2008, 32 (6), 462–468. [PubMed: 18757239]
- (120). Rahman M; Sadygov RG PLoS One2017, 12 (7), No. e0180428. [PubMed: 28719664]
- (121). Doya KH FSP J. 2007, 1 (1), 30–40. [PubMed: 19404458]
- (122). Washburn JD; Mejia-Guerra MK; Ramstein G; Kremling KA; Valluru R; Buckler ES; Wang H Proc. Natl. Acad. Sci. U. S. A2019, 116 (12), 5542–5549. [PubMed: 30842277]
- (123). Zimmer D; Schneider K; Sommer F; Schroda M; Muhlhaus TF Front. Plant Sci2018, 9, 1559. [PubMed: 30483279]
- (124). Swan AL; Mobasheri A; Allaway D; Liddell S; Bacardit J OMICS2013, 17 (12), 595–610. [PubMed: 24116388]
- (125). Ferreira M; Ventorim R; Almeida E; Silveria S; Silveira W Protein Abundance Prediction Through Machine Learning Methods. bioRxiv2020. DOI: 10.1101/2020.09.17.302182.
- (126). Mathur M; Kim CM; Munro SA; Rudina SS; Sawyer EM; Smolke CD Nat. Commun2019, 10 (1), 2673. [PubMed: 31209208]
- (127). Costello Z; Martin HG NPJ. Syst. Biol. Appl2018, 4, 19. Costello, Z.; Martin, H. G. *NPJ. Syst. Biol. Appl.* **2018**, 4, 19. [PubMed: 29872542]
- (128). Deng L; Liu Y; Shi Y; Zhang W; Yang C; Liu H BMC Genomics2020, 21, 21. [PubMed: 31906847]
- (129). Alam T; Islam MT; Househ M; Bouzerdoum A; Kawsar FA Stud. Health Technol. Inform2019, 262, 236–239. [PubMed: 31349311]
- (130). Jaganathan K; Kyriazopoulou Panagiotopoulou S; McRae JF; Darbandi SF; Knowles D; Li YI; Kosmicki JA; Arbelaez J; Cui W; Schwartz GB; Chow ED; Kanterakis E; Gao H; Kia A; Batzoglou S; Sanders SJ; Farh KK Cell2019, 176 (3), 535–548. [PubMed: 30661751]
- (131). Qu YH; Yu H; Gong XJ; Xu JH; Lee HS PLoS One2017, 12 (12), No. e0188129. [PubMed: 29287069]
- (132). Geiler-Samerotte KA; Bauer CR; Li S; Ziv N; Gresham D; Siegal ML Curr. Opin. Biotechnol2013, 24 (4), 752–759. [PubMed: 23566377]
- (133). Beuchel C; Becker S; Dittrich J; Kirsten H; Toenjes A; Stumvoll M; Loeffler M; Thiele H; Beutner F; Thiery J; Ceglarek U; Scholz MM Mol. Metab2019, 29, 76–85. [PubMed: 31668394]
- (134). Raina SK Neurol. India2015, 63 (6), 1005–1006. [PubMed: 26588657]
- (135). Forshed JJ. Proteome Res2017, 16 (11), 3954–3960. [PubMed: 28965415]

- (136). Badgeley MA; Zech JR; Oakden-Rayner L; Glicksberg BS; Liu M; Gale W; McConnell MV; Percha B; Snyder TM; Dudley JTNPJ. Digit Med2019, 2, 31. [PubMed: 31304378]
- (137). Moore JH; Williams SMBioEssays2005, 27 (6), 637–646. [PubMed: 15892116]
- (138). Szklarczyk D; Gable AL; Lyon D; Junge A; Wyder S; Huerta-Cepas J; Simonovic M; Doncheva NT; Morris JH; Bork P; Jensen LJ; Mering CVNucleic Acids Res. 2019, 47 (D1), D607–D613. [PubMed: 30476243]
- (139). Hawe JS; Theis FJ; Heinig MFront. Genet2019, 10, 535. [PubMed: 31249591]
- (140). Knudson DV; Lindsey CCompr. Psychol2014, 3, 3.
- (141). Argelaguet R; Velten B; Arnol D; Dietrich S; Zenz T; Marioni JC; Buettner F; Huber W; Stegle OMol. Syst. Biol2018, 14, No. e8124. [PubMed: 29925568]

Figure 1.
Types of artificial intelligence strategies suitable for omic data analysis and interpretation.

Figure 2.

Standard validation and AI molecular identification steps. AI is capable of increasing identification confidence by assessing available manuscripts and experimental and biological information so time is not wasted ruling out both false and true annotations that do not meet the requirements for identification levels.

Figure 3.

Missing value estimation in multiomic analyses. (a) Overview for missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) strategies. (b) Workflow for predicting the analyte half-lives of promiscuous enzymes. First, molecular docking is used to probe metabolite–enzyme interactions. QSAR modeling then predicts structurally similar species with corresponding activities to suggest additional enzyme interactions. The presence of inhibitory species can then be confirmed through kinetics experiments.

Figure 4.

A statistically significant molecule and its potential relationship between abundance in a healthy and disease cohort and parameters such as lifestyle, diet, sex, age, ethnicity, etc., which have been linked to influxes across the population. Defining how these factors trigger disparity in molecule abundances is crucial for assessing biomarkers, developing therapeutic strategies, and providing a comprehensive understanding of molecular mechanisms leading to a disease.

Figure 5.
Summary of the experimental (top) and computational (bottom) methods discussed for facilitating multiomic data analysis and interpretation.